Contents

Preface **56**

.1 What is in the book . . . 56
.2 What we promise about proof . . . 56
.3 What a section looks like . . . 57
.4 What we assume . . . 57
.5 How to read it . . . 58

0 Chapter 0 **59**

Preliminaries **59**

0.0.1 Roadmap . . . 59
0.0.2 Named moves . . . 60

0.1 Statements, Implications, and Quantifiers **60**

0.1.1 Statements and connectives . . . 60
0.1.2 Implication . . . 61
0.1.3 Quantifiers . . . 64
0.1.4 The order of quantifiers . . . 65
0.1.5 Negation . . . 66
0.1.6 Exercises . . . 70
0.1.6.1 A. Check your understanding . . . 70
0.1.6.2 B. Practice . . . 70
0.1.6.3 C. Going deeper . . . 71

0.2 How Proofs Are Built **72**

0.2.1 Direct proof . . . 73
0.2.2 Proof by contrapositive . . . 74
0.2.3 Proof by contradiction . . . 74
0.2.4 Proof by cases . . . 76
0.2.5 Proving “if and only if” . . . 76
0.2.6 Existence and uniqueness . . . 77
0.2.7 Counterexamples . . . 77
0.2.8 Mathematical induction . . . 78
0.2.9 Strong induction and well-ordering . . . 80
0.2.10 Writing proofs in this book . . . 82
0.2.11 Exercises . . . 82
0.2.11.1 A. Check your understanding . . . 82
0.2.11.2 B. Practice . . . 83
0.2.11.3 C. Going deeper . . . 84

0.3 Sets **85**

0.3.1 Sets, elements, and set-builder notation . . . 86

0.3.2 Subsets . . . 87
0.3.3 Equality of sets and double inclusion . . . 89
0.3.4 Union, intersection, and difference . . . 90
0.3.5 Indexed families of sets . . . 92
0.3.6 Cartesian products and n-tuples . . . 94
0.3.7 Power sets . . . 95
0.3.8 Exercises . . . 95
0.3.8.1 A. Check your understanding . . . 95
0.3.8.2 B. Practice . . . 95
0.3.8.3 C. Going deeper . . . 97

0.4 Functions **98**
0.4.1 What a function is . . . 98
0.4.2 Images and preimages . . . 99
0.4.3 Injective, surjective, bijective . . . 100
0.4.4 Composition . . . 102
0.4.5 Identity and one-sided inverses . . . 105
0.4.6 Inverses . . . 106
0.4.7 Maps between finite sets . . . 108
0.4.8 Exercises . . . 109
0.4.8.1 A. Check your understanding . . . 109
0.4.8.2 B. Practice . . . 110
0.4.8.3 C. Going deeper . . . 111

0.5 Relations, Equivalence Classes, and Quotients **113**
0.5.1 Relations . . . 113
0.5.2 Equivalence relations . . . 114
0.5.3 Equivalence classes and partitions . . . 116
0.5.4 Quotient sets . . . 119
0.5.5 Functions on a quotient . . . 120
0.5.6 Exercises . . . 122
0.5.6.1 A. Check your understanding . . . 122
0.5.6.2 B. Practice . . . 123
0.5.6.3 C. Going deeper . . . 125

0.6 Complex Numbers **127**
0.6.1 The complex numbers as pairs of reals . . . 127
0.6.2 Conjugate and modulus . . . 130
0.6.3 Two things that do not carry over from the reals . . . 132
0.6.4 Polar form: multiplication rotates and scales . . . 133
0.6.5 Roots of unity . . . 135
0.6.6 The Fundamental Theorem of Algebra . . . 136
0.6.7 Exercises . . . 137
0.6.7.1 A. Check your understanding . . . 137
0.6.7.2 B. Practice . . . 137
0.6.7.3 C. Going deeper . . . 139

0.7 Fields **140**
0.7.1 What solving linear equations needs . . . 140
0.7.2 The definition . . . 141
0.7.3 First consequences of the axioms . . . 142
0.7.4 Examples . . . 143

0.7.5 Finite fields . . . 144
0.7.6 Nonexamples . . . 147
0.7.7 Why these axioms . . . 148
0.7.8 The characteristic of a field . . . 148
0.7.9 Exercises . . . 150
0.7.9.1 A. Check your understanding . . . 150
0.7.9.2 B. Practice . . . 150
0.7.9.3 C. Going deeper . . . 152

0.8 Polynomials 154
0.8.1 What a polynomial is . . . 155
0.8.2 Adding and multiplying polynomials . . . 156
0.8.3 Degree . . . 158
0.8.4 Polynomials of bounded degree . . . 161
0.8.5 Evaluation and roots . . . 162
0.8.6 Looking ahead . . . 164
0.8.7 Exercises . . . 164
0.8.7.1 A. Check your understanding . . . 164
0.8.7.2 B. Practice . . . 165
0.8.7.3 C. Going deeper . . . 166

0.9 Matrices 168
0.9.1 Matrices and their entries . . . 168
0.9.2 Special matrices . . . 169
0.9.3 Addition and scalar multiplication . . . 171
0.9.4 Matrix multiplication . . . 172
0.9.5 Three views of a product . . . 174
0.9.6 The rules of multiplication . . . 176
0.9.7 The transpose . . . 178
0.9.8 The trace . . . 180
0.9.9 Inverse matrices . . . 181
0.9.10 Powers and polynomials of a matrix . . . 185
0.9.11 Two equivalence relations on matrices . . . 187
0.9.12 Exercises . . . 188
0.9.12.1 A. Check your understanding . . . 188
0.9.12.2 B. Practice . . . 189
0.9.12.3 C. Going deeper . . . 191

0.10 Groups and Permutations 193
0.10.1 Groups . . . 193
0.10.2 Subgroups . . . 196
0.10.3 Homomorphisms . . . 196
0.10.4 Permutations . . . 197
0.10.5 Exercises . . . 199
0.10.5.1 A. Check your understanding . . . 199
0.10.5.2 B. Practice . . . 199
0.10.5.3 C. Going deeper . . . 201

1 Chapter 1 203

Vector Spaces 203

1.0.1 Roadmap . . . 203
1.0.2 Named moves . . . 204

1.1 Vector Spaces 204
1.1.1 One set of rules in three places . . . 204
1.1.2 The definition . . . 205
1.1.3 Examples . . . 206
1.1.4 Nonexamples . . . 210
1.1.5 Why these axioms . . . 210
1.1.6 First consequences of the axioms . . . 210
1.1.7 Exercises . . . 214
1.1.7.1 A. Check your understanding . . . 214
1.1.7.2 B. Practice . . . 215
1.1.7.3 C. Going deeper . . . 216

1.2 Subspaces 218
1.2.1 Spaces inside spaces . . . 219
1.2.2 The subspace test . . . 220
1.2.3 Examples: the three-check move . . . 221
1.2.4 Nonexamples . . . 225
1.2.5 Why this definition . . . 226
1.2.6 Intersections of subspaces . . . 226
1.2.7 Unions of subspaces . . . 227
1.2.8 Exercises . . . 228
1.2.8.1 A. Check your understanding . . . 228
1.2.8.2 B. Practice . . . 229
1.2.8.3 C. Going deeper . . . 231

1.3 Linear Combinations and Span 233
1.3.1 Linear combinations . . . 234
1.3.2 The span . . . 235
1.3.3 The span is the smallest subspace . . . 236
1.3.4 Spanning lists and finite-dimensional spaces . . . 238
1.3.5 Working with spans . . . 240
1.3.6 Exercises . . . 242
1.3.6.1 A. Check your understanding . . . 242
1.3.6.2 B. Practice . . . 243
1.3.6.3 C. Going deeper . . . 244

1.4 Linear Independence 245
1.4.1 Redundant vectors . . . 245
1.4.2 Definition and examples . . . 246
1.4.3 Why lists, and why this definition . . . 248
1.4.4 The Linear Dependence Lemma . . . 249
1.4.5 Unique representations . . . 252
1.4.6 Polynomials of distinct degrees . . . 252
1.4.7 Independence depends on the field . . . 253
1.4.8 Exercises . . . 253
1.4.8.1 A. Check your understanding . . . 253
1.4.8.2 B. Practice . . . 254
1.4.8.3 C. Going deeper . . . 255

1.5 Bases and Coordinates **257**
1.5.1 What makes coordinates work . . . 257
1.5.2 Bases . . . 258
1.5.3 Unique representation and coordinates . . . 260
1.5.4 Every finite spanning list contains a basis . . . 262
1.5.5 Exercises . . . 266
1.5.5.1 A. Check your understanding . . . 266
1.5.5.2 B. Practice . . . 266
1.5.5.3 C. Going deeper . . . 268

1.6 The Exchange Theorem and Dimension **269**
1.6.1 How big is a vector space? . . . 269
1.6.2 The Steinitz Exchange Theorem . . . 270
1.6.3 Dimension . . . 272
1.6.4 Size bounds . . . 274
1.6.5 Count instead of check . . . 275
1.6.6 Extending to a basis . . . 276
1.6.7 Largest independent lists and smallest spanning lists . . . 277
1.6.8 Dimensions of subspaces . . . 278
1.6.9 Exercises . . . 279
1.6.9.1 A. Check your understanding . . . 279
1.6.9.2 B. Practice . . . 280
1.6.9.3 C. Going deeper . . . 281

1.7 Sums and Direct Sums **283**
1.7.1 The sum of subspaces . . . 283
1.7.2 The dimension formula . . . 285
1.7.3 Direct sums . . . 288
1.7.4 Direct sums of several subspaces . . . 291
1.7.5 Complements . . . 293
1.7.6 Exercises . . . 295
1.7.6.1 A. Check your understanding . . . 295
1.7.6.2 B. Practice . . . 295
1.7.6.3 C. Going deeper . . . 297

1.8 Infinite-Dimensional Spaces **299**
1.8.1 Bases of infinite sets of vectors . . . 299
1.8.2 Zorn's Lemma . . . 301
1.8.3 Every vector space has a basis . . . 302
1.8.4 Existence without construction: $\mathbb{R}$ over $\mathbb{Q}$ . . . 303
1.8.5 What changes in infinite dimension . . . 305
1.8.6 Exercises . . . 306
1.8.6.1 A. Check your understanding . . . 306
1.8.6.2 B. Practice . . . 306
1.8.6.3 C. Going deeper . . . 308

1.9 Changing the Field **309**
1.9.1 Restricting the scalars . . . 309
1.9.2 The tower law . . . 310
1.9.3 Complexification . . . 312
1.9.4 Exercises . . . 314
1.9.4.1 A. Check your understanding . . . 314

1.9.4.2 B. Practice . . . 315
1.9.4.3 C. Going deeper . . . 316

2 Chapter 2 **318**

Linear Maps **318**
2.0.1 Roadmap . . . 318
2.0.2 Named moves . . . 319

2.1 Linear Maps **319**
2.1.1 Functions that behave like matrices . . . 319
2.1.2 The definition . . . 320
2.1.3 Examples . . . 320
2.1.4 Three consequences of linearity . . . 325
2.1.5 Nonexamples . . . 326
2.1.6 A linear map is determined by a basis . . . 328
2.1.7 Exercises . . . 331
2.1.7.1 A. Check your understanding . . . 331
2.1.7.2 B. Practice . . . 331
2.1.7.3 C. Going deeper . . . 333

2.2 Kernel and Image **334**
2.2.1 Two questions, two subspaces . . . 334
2.2.2 Injectivity via the kernel . . . 337
2.2.3 Computing the image . . . 338
2.2.4 Examples . . . 338
2.2.5 What injective and surjective maps do to lists . . . 340
2.2.6 Exercises . . . 341
2.2.6.1 A. Check your understanding . . . 341
2.2.6.2 B. Practice . . . 341
2.2.6.3 C. Going deeper . . . 343

2.3 The Rank–Nullity Theorem **344**
2.3.1 Rank and nullity . . . 344
2.3.2 The theorem . . . 345
2.3.3 Count instead of check . . . 347
2.3.4 Existence with conditions on derivatives . . . 348
2.3.5 Dimensions of subspaces cut out by conditions . . . 349
2.3.6 Exercises . . . 349
2.3.6.1 A. Check your understanding . . . 349
2.3.6.2 B. Practice . . . 350
2.3.6.3 C. Going deeper . . . 351

2.4 The Algebra of Linear Maps **353**
2.4.1 Adding and scaling linear maps . . . 353
2.4.2 Composition . . . 355
2.4.3 Operators, powers, and polynomials . . . 357
2.4.4 Exercises . . . 360
2.4.4.1 A. Check your understanding . . . 360
2.4.4.2 B. Practice . . . 360
2.4.4.3 C. Going deeper . . . 362

2.5 Invertible Maps and Isomorphisms **364**
2.5.1 When are two spaces the same? . . . 364
2.5.2 Classification by dimension . . . 368
2.5.3 One-sided inverses . . . 370
2.5.4 The Invertible Operator Theorem . . . 371
2.5.5 Exercises . . . 373
2.5.5.1 A. Check your understanding . . . 373
2.5.5.2 B. Practice . . . 373
2.5.5.3 C. Going deeper . . . 375

3 Chapter 3 **377**

Linear Systems and Matrices **377**
3.0.1 Roadmap . . . 377
3.0.2 Named moves . . . 378

3.1 Three Views of a Linear System **378**
3.1.1 One system, three pictures . . . 378
3.1.2 Linear systems . . . 379
3.1.3 Consistency is a question about span . . . 382
3.1.4 The shape of the solution set . . . 383
3.1.5 When is the solution unique? . . . 385
3.1.6 Exercises . . . 386
3.1.6.1 A. Check your understanding . . . 386
3.1.6.2 B. Practice . . . 386
3.1.6.3 C. Going deeper . . . 388

3.2 Gaussian Elimination **389**
3.2.1 Three moves that do not change the answer . . . 389
3.2.2 Row operations preserve solutions . . . 391
3.2.3 Echelon forms . . . 393
3.2.4 Gauss–Jordan elimination . . . 394
3.2.5 Reading off the solutions . . . 398
3.2.6 More unknowns than equations . . . 400
3.2.7 Exercises . . . 401
3.2.7.1 A. Check your understanding . . . 401
3.2.7.2 B. Practice . . . 402
3.2.7.3 C. Going deeper . . . 405

3.3 Uniqueness of the Reduced Form **407**
3.3.1 What row operations remember . . . 407
3.3.2 The columns of a reduced matrix . . . 408
3.3.3 The uniqueness theorem . . . 409
3.3.4 Pivot columns are well defined . . . 411
3.3.5 Exercises . . . 413
3.3.5.1 A. Check your understanding . . . 413
3.3.5.2 B. Practice . . . 414
3.3.5.3 C. Going deeper . . . 415

3.4 Elementary Matrices and Inverses **417**
3.4.1 Row operations are matrix multiplications . . . 417
3.4.2 The invertible matrix theorem . . . 420

3.4.3 One-sided inverses are two-sided . . . 423
3.4.4 Computing the inverse . . . 424
3.4.5 Permutation matrices . . . 427
3.4.6 Row equivalence, revisited . . . 428
3.4.7 Exercises . . . 428
3.4.7.1 A. Check your understanding . . . 428
3.4.7.2 B. Practice . . . 429
3.4.7.3 C. Going deeper . . . 432

3.5 Row Space, Column Space, Null Space, and Rank 434
3.5.1 Three subspaces of a matrix . . . 434
3.5.2 Rank and nullity . . . 435
3.5.3 What row operations preserve . . . 436
3.5.4 Bases from the reduced form . . . 437
3.5.5 Row rank equals column rank . . . 441
3.5.6 Rank–nullity and the Rouché–Capelli theorem . . . 442
3.5.7 Rank of a product . . . 444
3.5.8 A toolkit for questions from Chapter 1 . . . 444
3.5.9 Exercises . . . 448
3.5.9.1 A. Check your understanding . . . 448
3.5.9.2 B. Practice . . . 449
3.5.9.3 C. Going deeper . . . 451

3.6 LU and PLU Factorizations 453
3.6.1 Triangular systems are cheap . . . 454
3.6.2 LU factorizations . . . 454
3.6.3 Unit lower triangular matrices . . . 456
3.6.4 Elimination without row swaps . . . 458
3.6.5 Solving with an LU factorization . . . 461
3.6.6 Uniqueness . . . 462
3.6.7 Row swaps and the PLU factorization . . . 463
3.6.8 Symmetric matrices: the $LDL^\top$ factorization . . . 466
3.6.9 Exercises . . . 467
3.6.9.1 A. Check your understanding . . . 467
3.6.9.2 B. Practice . . . 468
3.6.9.3 C. Going deeper . . . 470

3.7 Applications of Linear Systems 472
3.7.1 Polynomial interpolation . . . 472
3.7.2 Interpolation without a matrix . . . 476
3.7.3 Balancing chemical equations . . . 476
3.7.4 Flows in a network . . . 478
3.7.5 Lights out over $\mathbb{F}_2$ . . . 481
3.7.6 Exercises . . . 483
3.7.6.1 A. Check your understanding . . . 483
3.7.6.2 B. Practice . . . 484
3.7.6.3 C. Going deeper . . . 486

4 Chapter 4 489

Matrices of Maps 489

4.0.1 Roadmap . . . 490
4.0.2 Named moves . . . 490

4.1 The Matrix of a Linear Map 490
4.1.1 Writing a map as a matrix . . . 491
4.1.2 The fundamental square . . . 494
4.1.3 Composition is matrix multiplication . . . 496
4.1.4 Linear maps are matrices . . . 498
4.1.5 Reading rank and invertibility from a matrix . . . 499
4.1.6 Exercises . . . 502
4.1.6.1 A. Check your understanding . . . 502
4.1.6.2 B. Practice . . . 502
4.1.6.3 C. Going deeper . . . 504

4.2 Change of Basis and Similarity 505
4.2.1 Change-of-coordinates matrices . . . 506
4.2.2 Change of basis for linear maps . . . 509
4.2.3 Similar matrices . . . 511
4.2.4 Similarity invariants . . . 512
4.2.5 A basis adapted to the operator . . . 514
4.2.6 Exercises . . . 515
4.2.6.1 A. Check your understanding . . . 515
4.2.6.2 B. Practice . . . 516
4.2.6.3 C. Going deeper . . . 518

4.3 Rank, Equivalence, and Factorizations 520
4.3.1 The column–row factorization . . . 520
4.3.2 Equivalent matrices . . . 522
4.3.3 Rank inequalities for compositions . . . 525
4.3.4 Exercises . . . 526
4.3.4.1 A. Check your understanding . . . 526
4.3.4.2 B. Practice . . . 527
4.3.4.3 C. Going deeper . . . 529

4.4 Products and Quotient Spaces 530
4.4.1 Products of vector spaces . . . 531
4.4.2 Cosets . . . 533
4.4.3 The quotient space . . . 535
4.4.4 The First Isomorphism Theorem . . . 539
4.4.5 The Second and Third Isomorphism Theorems . . . 540
4.4.6 The universal property of the quotient . . . 542
4.4.7 Exercises . . . 543
4.4.7.1 A. Check your understanding . . . 543
4.4.7.2 B. Practice . . . 543
4.4.7.3 C. Going deeper . . . 545

4.5 Projections and the Trace of an Operator 547
4.5.1 Projections . . . 547
4.5.2 The trace of an operator . . . 550
4.5.3 Involutions and reflections . . . 553
4.5.4 Exercises . . . 554
4.5.4.1 A. Check your understanding . . . 554

4.5.4.2 B. Practice . . . 554
4.5.4.3 C. Going deeper . . . 556

5 Chapter 5 **558**

Duality **558**
5.0.1 Roadmap . . . 558
5.0.2 Named moves . . . 559

5.1 Linear Functionals and the Dual Space **559**
5.1.1 Measuring a vector by a number . . . 559
5.1.2 The dual basis . . . 562
5.1.3 Testing against functionals . . . 566
5.1.4 When V is infinite-dimensional . . . 567
5.1.5 Exercises . . . 568
5.1.5.1 A. Check your understanding . . . 568
5.1.5.2 B. Practice . . . 569
5.1.5.3 C. Going deeper . . . 570

5.2 Annihilators **572**
5.2.1 The equations a subspace satisfies . . . 573
5.2.2 Counting equations . . . 575
5.2.3 Sums, intersections, and spans . . . 577
5.2.4 Exercises . . . 578
5.2.4.1 A. Check your understanding . . . 578
5.2.4.2 B. Practice . . . 579
5.2.4.3 C. Going deeper . . . 580

5.3 The Double Dual **582**
5.3.1 Measuring the measurements . . . 582
5.3.2 The evaluation map never loses information . . . 584
5.3.3 What “natural” means . . . 586
5.3.4 Consequences in finite dimension . . . 587
5.3.5 Infinite dimension: the evaluation map is not surjective . . . 590
5.3.6 Exercises . . . 590
5.3.6.1 A. Check your understanding . . . 590
5.3.6.2 B. Practice . . . 591
5.3.6.3 C. Going deeper . . . 593

5.4 The Dual Map and the Transpose **594**
5.4.1 Pulling functionals back . . . 594
5.4.2 The matrix of the dual map . . . 597
5.4.3 Kernel and image of the dual map . . . 599
5.4.4 Naturality of the double dual . . . 602
5.4.5 Exercises . . . 603
5.4.5.1 A. Check your understanding . . . 603
5.4.5.2 B. Practice . . . 604
5.4.5.3 C. Going deeper . . . 605

5.5 Duality for Subspaces and Quotients **607**
5.5.1 Restricting functionals to a subspace . . . 607
5.5.2 Functionals on a quotient . . . 608

5.5.3 A line in $\mathbb{R}^3$, in coordinates . . . 610
5.5.4 Why "natural" matters here . . . 611
5.5.5 Codimension and hyperplanes . . . 611
5.5.6 Exercises . . . 614
5.5.6.1 A. Check your understanding . . . 614
5.5.6.2 B. Practice . . . 615
5.5.6.3 C. Going deeper . . . 616

5.6 Hyperplanes and Linear Systems Revisited 617
5.6.1 Row vectors are functionals . . . 618
5.6.2 Every subspace is a solution set . . . 618
5.6.3 From spans to equations and back . . . 620
5.6.4 Affine solution sets . . . 621
5.6.5 The row space annihilates the null space . . . 622
5.6.6 Sums and intersections by duality . . . 623
5.6.7 Independent functionals . . . 624
5.6.8 Summary and transfer . . . 625
5.6.9 Exercises . . . 626
5.6.9.1 A. Check your understanding . . . 626
5.6.9.2 B. Practice . . . 627
5.6.9.3 C. Going deeper . . . 629

6 Chapter 6 631

Polynomials 631
6.0.1 Roadmap . . . 631
6.0.2 Named moves . . . 632

6.1 Division with Remainder 632
6.1.1 Long division by hand . . . 632
6.1.2 The Division Theorem . . . 633
6.1.3 Divisibility . . . 635
6.1.4 The Remainder and Factor Theorems . . . 637
6.1.5 Polynomials and polynomial functions . . . 638
6.1.6 Exercises . . . 640
6.1.6.1 A. Check your understanding . . . 640
6.1.6.2 B. Practice . . . 640
6.1.6.3 C. Going deeper . . . 642

6.2 Ideals and Greatest Common Divisors 644
6.2.1 All the combinations of two polynomials . . . 644
6.2.2 Ideals . . . 644
6.2.3 Every ideal is generated by one polynomial . . . 647
6.2.4 The greatest common divisor . . . 648
6.2.5 The Euclidean algorithm . . . 650
6.2.6 Coprime polynomials . . . 652
6.2.7 The gcd does not depend on the field . . . 654
6.2.8 Exercises . . . 655
6.2.8.1 A. Check your understanding . . . 655
6.2.8.2 B. Practice . . . 655
6.2.8.3 C. Going deeper . . . 657

6.3 Irreducible Polynomials and Unique Factorization 659
6.3.1 Irreducible polynomials . . . 660
6.3.2 Euclid's lemma . . . 662
6.3.3 Unique factorization . . . 663
6.3.4 Tools over $\mathbb{Q}$: rational roots and Eisenstein's criterion . . . 665
6.3.5 Exercises . . . 667
6.3.5.1 A. Check your understanding . . . 667
6.3.5.2 B. Practice . . . 668
6.3.5.3 C. Going deeper . . . 669

6.4 Roots, Multiplicity, and Derivatives 671
6.4.1 Multiplicity of a root . . . 672
6.4.2 Counting roots with multiplicity . . . 674
6.4.3 The formal derivative . . . 675
6.4.4 Repeated roots and the derivative . . . 677
6.4.5 Taylor expansion of a polynomial . . . 679
6.4.6 Exercises . . . 681
6.4.6.1 A. Check your understanding . . . 681
6.4.6.2 B. Practice . . . 682
6.4.6.3 C. Going deeper . . . 683

6.5 The Fundamental Theorem of Algebra 685
6.5.1 The goal and the plan . . . 685
6.5.2 Step 1: growth . . . 686
6.5.3 Step 2: the minimum is attained . . . 687
6.5.4 Step 3: d'Alembert's lemma . . . 688
6.5.5 The proof of the theorem . . . 690
6.5.6 Complex polynomials split . . . 691
6.5.7 Real polynomials . . . 692
6.5.8 Exercises . . . 694
6.5.8.1 A. Check your understanding . . . 694
6.5.8.2 B. Practice . . . 695
6.5.8.3 C. Going deeper . . . 697

6.6 Interpolation and the Chinese Remainder Theorem 699
6.6.1 The Lagrange basis . . . 699
6.6.2 Evaluations are the dual basis . . . 702
6.6.3 Congruences and the Chinese Remainder Theorem . . . 703
6.6.4 The quotient space $F[x]/\langle m \rangle$ . . . 707
6.6.5 Exercises . . . 709
6.6.5.1 A. Check your understanding . . . 709
6.6.5.2 B. Practice . . . 709
6.6.5.3 C. Going deeper . . . 711

6.7 Polynomials of an Operator 713
6.7.1 Evaluating at an operator . . . 713
6.7.2 Annihilating polynomials . . . 715
6.7.3 Kernels of polynomials in T . . . 717
6.7.4 The kernel splitting lemma . . . 718
6.7.5 Exercises . . . 722
6.7.5.1 A. Check your understanding . . . 722
6.7.5.2 B. Practice . . . 722

6.7.5.3 C. Going deeper . . . 724

7 Chapter 7 **727**

Determinants **727**
7.0.1 Roadmap . . . 727
7.0.2 Named moves . . . 728

7.1 Signed Area and Volume **728**
7.1.1 Area as a function of two vectors . . . 728
7.1.2 Why the area needs a sign . . . 729
7.1.3 The rules force $ad - bc$ . . . 730
7.1.4 Volume in three dimensions . . . 732
7.1.5 What we will prove . . . 732
7.1.6 Exercises . . . 733
7.1.6.1 A. Check your understanding . . . 733
7.1.6.2 B. Practice . . . 734
7.1.6.3 C. Going deeper . . . 735

7.2 Multilinear Alternating Forms **736**
7.2.1 Multilinear forms . . . 736
7.2.2 Alternating and skew-symmetric forms . . . 737
7.2.3 Consequences of the alternating condition . . . 739
7.2.4 Alternating versus skew-symmetric . . . 740
7.2.5 The space of alternating forms . . . 741
7.2.6 Exercises . . . 741
7.2.6.1 A. Check your understanding . . . 741
7.2.6.2 B. Practice . . . 742
7.2.6.3 C. Going deeper . . . 743

7.3 Permutations and Sign **745**
7.3.1 Permutations, recalled . . . 745
7.3.2 Inversions and the sign . . . 745
7.3.3 The sign is multiplicative . . . 746
7.3.4 Transpositions and parity . . . 748
7.3.5 The alternating group . . . 749
7.3.6 Back to alternating forms . . . 749
7.3.7 Computing signs in practice . . . 750
7.3.8 Exercises . . . 751
7.3.8.1 A. Check your understanding . . . 751
7.3.8.2 B. Practice . . . 751
7.3.8.3 C. Going deeper . . . 753

7.4 Existence and Uniqueness of the Determinant **753**
7.4.1 The goal and the plan . . . 754
7.4.2 Rearranging the arguments . . . 754
7.4.3 Uniqueness . . . 755
7.4.4 Existence . . . 757
7.4.5 The determinant . . . 758
7.4.6 Transposes and triangular matrices . . . 760
7.4.7 Determinants over a commutative ring . . . 762
7.4.8 Exercises . . . 762

7.4.8.1 A. Check your understanding . . . 762
7.4.8.2 B. Practice . . . 763
7.4.8.3 C. Going deeper . . . 764

7.5 Row Operations and Multiplicativity 766
7.5.1 Row operations . . . 766
7.5.2 Determinant and invertibility . . . 768
7.5.3 Multiplicativity . . . 770
7.5.4 The determinant of an operator . . . 772
7.5.5 The determinant as a homomorphism . . . 773
7.5.6 Which proofs survive over a ring . . . 774
7.5.7 Exercises . . . 774
7.5.7.1 A. Check your understanding . . . 774
7.5.7.2 B. Practice . . . 775
7.5.7.3 C. Going deeper . . . 776

7.6 Cofactor Expansion, the Adjugate, and Cramer's Rule 778
7.6.1 Minors and cofactors . . . 778
7.6.2 The Laplace expansion . . . 779
7.6.3 The adjugate . . . 782
7.6.4 Cramer's rule . . . 785
7.6.5 Integer matrices . . . 786
7.6.6 Exercises . . . 787
7.6.6.1 A. Check your understanding . . . 787
7.6.6.2 B. Practice . . . 788
7.6.6.3 C. Going deeper . . . 789

7.7 Minors, Rank, and Cauchy–Binet 791
7.7.1 Submatrices and minors . . . 792
7.7.2 Rank via minors . . . 793
7.7.3 The Cauchy–Binet formula . . . 795
7.7.4 Sums of squares over the reals . . . 797
7.7.5 Exercises . . . 798
7.7.5.1 A. Check your understanding . . . 798
7.7.5.2 B. Practice . . . 799
7.7.5.3 C. Going deeper . . . 800

7.8 Special Determinants 802
7.8.1 The Vandermonde determinant . . . 802
7.8.2 Tridiagonal determinants . . . 805
7.8.3 Block triangular matrices . . . 806
7.8.4 Rank-one updates . . . 808
7.8.5 Exercises . . . 810
7.8.5.1 A. Check your understanding . . . 810
7.8.5.2 B. Practice . . . 810
7.8.5.3 C. Going deeper . . . 811

7.9 Orientation and Volume 814
7.9.1 What should the volume of a parallelepiped be? . . . 814
7.9.2 Volume is forced by its rules . . . 815
7.9.3 Linear maps scale volume . . . 817
7.9.4 Orientation . . . 818

7.9.5 The right-hand rule . . . 821
7.9.6 Orientation-preserving maps . . . 821
7.9.7 Exercises . . . 822
7.9.7.1 A. Check your understanding . . . 822
7.9.7.2 B. Practice . . . 822
7.9.7.3 C. Going deeper . . . 824

7.10 The Characteristic Polynomial 825
7.10.1 From a number to a variable . . . 825
7.10.2 The shape of the characteristic polynomial . . . 828
7.10.3 Similar matrices, same polynomial . . . 829
7.10.4 Roots and singular matrices . . . 831
7.10.5 A tempting false proof . . . 832
7.10.6 Exercises . . . 832
7.10.6.1 A. Check your understanding . . . 832
7.10.6.2 B. Practice . . . 833
7.10.6.3 C. Going deeper . . . 835

8 Chapter 8 838

Block Matrices 838
8.0.1 Roadmap . . . 838
8.0.2 Named moves . . . 839

8.1 Partitions and Block Multiplication 839
8.1.1 Cutting a matrix into blocks . . . 839
8.1.2 Block multiplication . . . 841
8.1.3 Rows, columns, and outer products . . . 843
8.1.4 Block diagonal matrices . . . 844
8.1.5 Block triangular matrices . . . 846
8.1.6 Matrices of operators in adapted bases . . . 847
8.1.7 Exercises . . . 850
8.1.7.1 A. Check your understanding . . . 850
8.1.7.2 B. Practice . . . 850
8.1.7.3 C. Going deeper . . . 853

8.2 Rank and Block Matrices 854
8.2.1 Three basic inequalities . . . 855
8.2.2 Block elimination . . . 857
8.2.3 Frobenius's inequality by blocks . . . 858
8.2.4 A worked example: $\mathbf{I}-\mathbf{AB}$ and $\mathbf{I}-\mathbf{BA}$ . . . 859
8.2.5 Exercises . . . 861
8.2.5.1 A. Check your understanding . . . 861
8.2.5.2 B. Practice . . . 861
8.2.5.3 C. Going deeper . . . 864

8.3 Schur Complements 866
8.3.1 Eliminating a block . . . 866
8.3.2 The block LDU factorization . . . 868
8.3.3 Determinant and rank . . . 869
8.3.4 Inverting a block matrix . . . 870
8.3.5 The Sherman–Morrison–Woodbury formula . . . 873

8.3.6 Commuting blocks . . . 875
8.3.7 Exercises . . . 877
8.3.7.1 A. Check your understanding . . . 877
8.3.7.2 B. Practice . . . 877
8.3.7.3 C. Going deeper . . . 879

8.4 The Kronecker Product 882
8.4.1 From a matrix equation to a linear system . . . 882
8.4.2 The Kronecker product . . . 882
8.4.3 Algebraic properties . . . 884
8.4.4 Swapping the factors . . . 886
8.4.5 Rank, trace, and determinant . . . 887
8.4.6 Vectorization . . . 888
8.4.7 Exercises . . . 891
8.4.7.1 A. Check your understanding . . . 891
8.4.7.2 B. Practice . . . 891
8.4.7.3 C. Going deeper . . . 894

8.5 Commutators and Trace-Zero Matrices 895
8.5.1 Commutators . . . 896
8.5.2 The identity in characteristic p . . . 897
8.5.3 Trace-zero matrices with zero diagonal . . . 898
8.5.4 Shoda's theorem . . . 901
8.5.5 The span of the commutators . . . 903
8.5.6 Exercises . . . 904
8.5.6.1 A. Check your understanding . . . 904
8.5.6.2 B. Practice . . . 904
8.5.6.3 C. Going deeper . . . 907

9 Chapter 9 910

Eigenvalues and Diagonalization 910
9.0.1 Roadmap . . . 911
9.0.2 Named moves . . . 911

9.1 Invariant Subspaces 911
9.1.1 Subspaces an operator maps into itself . . . 912
9.1.2 Restriction and the induced operator . . . 914
9.1.3 Sums, intersections, and direct sums . . . 916
9.1.4 Exercises . . . 918
9.1.4.1 A. Check your understanding . . . 918
9.1.4.2 B. Practice . . . 918
9.1.4.3 C. Going deeper . . . 920

9.2 Eigenvalues and Eigenvectors 921
9.2.1 When is a matrix diagonal? . . . 921
9.2.2 Finding eigenvalues . . . 923
9.2.3 Distinct eigenvalues give independent eigenvectors . . . 926
9.2.4 Existence over the complex numbers . . . 928
9.2.5 Real matrices and complex eigenvalues . . . 929
9.2.6 The invertible matrix theorem grows . . . 930
9.2.7 Exercises . . . 931

9.2.7.1 A. Check your understanding . . . 931
9.2.7.2 B. Practice . . . 932
9.2.7.3 C. Going deeper . . . 934

9.3 Algebraic and Geometric Multiplicity 935
9.3.1 Two ways to count an eigenvalue . . . 936
9.3.2 The geometric multiplicity never exceeds the algebraic . . . 938
9.3.3 Multiplicities, trace, and determinant . . . 939
9.3.4 Exercises . . . 941
9.3.4.1 A. Check your understanding . . . 941
9.3.4.2 B. Practice . . . 941
9.3.4.3 C. Going deeper . . . 943

9.4 Diagonalization 945
9.4.1 Diagonalizable operators . . . 945
9.4.2 The Diagonalization Theorem . . . 947
9.4.3 Computing **P** and **D** . . . 949
9.4.4 Powers and polynomials . . . 951
9.4.5 When diagonalization fails . . . 953
9.4.6 Exercises . . . 954
9.4.6.1 A. Check your understanding . . . 954
9.4.6.2 B. Practice . . . 954
9.4.6.3 C. Going deeper . . . 957

9.5 The Minimal Polynomial 959
9.5.1 The polynomial that divides all the others . . . 959
9.5.2 The roots are the eigenvalues . . . 962
9.5.3 Computing the minimal polynomial . . . 963
9.5.4 Similar matrices and block diagonal matrices . . . 966
9.5.5 Exercises . . . 967
9.5.5.1 A. Check your understanding . . . 967
9.5.5.2 B. Practice . . . 967
9.5.5.3 C. Going deeper . . . 970

9.6 The Cayley–Hamilton Theorem 971
9.6.1 The statement, and a proof that is not one . . . 971
9.6.2 A proof through the adjugate . . . 972
9.6.3 Consequences . . . 974
9.6.4 Exercises . . . 977
9.6.4.1 A. Check your understanding . . . 977
9.6.4.2 B. Practice . . . 978
9.6.4.3 C. Going deeper . . . 979

9.7 Triangularization 982
9.7.1 Triangular matrices and chains of invariant subspaces . . . 982
9.7.2 The Triangularization Theorem . . . 984
9.7.3 What the diagonal records . . . 986
9.7.4 Triangularizing a matrix in practice . . . 987
9.7.5 Exercises . . . 988
9.7.5.1 A. Check your understanding . . . 988
9.7.5.2 B. Practice . . . 989
9.7.5.3 C. Going deeper . . . 990

9.8 Primary Decomposition **993**
9.8.1 Splitting the space along the minimal polynomial . . . 993
9.8.2 Diagonalizability and the minimal polynomial . . . 997
9.8.3 Restrictions and polynomials of diagonalizable operators . . . 999
9.8.4 When the minimal polynomial splits: a first look at Chapter 10 . . . 999
9.8.5 Exercises . . . 1001
9.8.5.1 A. Check your understanding . . . 1001
9.8.5.2 B. Practice . . . 1001
9.8.5.3 C. Going deeper . . . 1003

9.9 Commuting Operators **1005**
9.9.1 When can two operators share a basis? . . . 1005
9.9.2 Commuting operators preserve eigenspaces . . . 1006
9.9.3 A common eigenvector . . . 1007
9.9.4 Simultaneous triangularization . . . 1009
9.9.5 Simultaneous diagonalization . . . 1010
9.9.6 Exercises . . . 1014
9.9.6.1 A. Check your understanding . . . 1014
9.9.6.2 B. Practice . . . 1014
9.9.6.3 C. Going deeper . . . 1017

9.10 Spectral Mapping, Products in Both Orders, Left Eigenvectors **1018**
9.10.1 Eigenvalues of a polynomial in T . . . 1019
9.10.2 Traces of powers detect nilpotency . . . 1021
9.10.3 **AB** and **BA** . . . 1023
9.10.4 Left eigenvectors . . . 1025
9.10.5 Exercises . . . 1028
9.10.5.1 A. Check your understanding . . . 1028
9.10.5.2 B. Practice . . . 1028
9.10.5.3 C. Going deeper . . . 1030

9.11 Powers, Recurrences, and Markov Chains **1032**
9.11.1 Vector recurrences . . . 1032
9.11.2 Scalar recurrences and companion matrices . . . 1033
9.11.3 Stochastic matrices . . . 1036
9.11.4 Markov chains . . . 1037
9.11.5 A population model . . . 1041
9.11.6 Exercises . . . 1042
9.11.6.1 A. Check your understanding . . . 1042
9.11.6.2 B. Practice . . . 1043
9.11.6.3 C. Going deeper . . . 1045

10 Chapter 10 **1048**

Canonical Forms **1048**
10.0.1 Roadmap . . . 1049
10.0.2 Named moves . . . 1049

10.1 Generalized Eigenspaces **1049**
10.1.1 Enlarging an eigenspace . . . 1049
10.1.2 Generalized eigenvectors and generalized eigenspaces . . . 1050
10.1.3 The kernel chain stabilizes . . . 1052

10.1.4 The generalized eigenspace decomposition . . . 1054
10.1.5 Exercises . . . 1057
10.1.5.1 A. Check your understanding . . . 1057
10.1.5.2 B. Practice . . . 1058
10.1.5.3 C. Going deeper . . . 1060

10.2 Nilpotent Operators 1061
10.2.1 Operators killed by a power . . . 1062
10.2.2 What one chain looks like . . . 1064
10.2.3 The structure theorem . . . 1065
10.2.4 Block sizes from the ranks of the powers . . . 1068
10.2.5 Partitions and Young diagrams . . . 1070
10.2.6 Finding a Jordan basis . . . 1071
10.2.7 Exercises . . . 1073
10.2.7.1 A. Check your understanding . . . 1073
10.2.7.2 B. Practice . . . 1074
10.2.7.3 C. Going deeper . . . 1076

10.3 The Jordan Canonical Form 1077
10.3.1 The shape we are aiming for . . . 1077
10.3.2 The theorem . . . 1079
10.3.3 Reading the invariants off the blocks . . . 1082
10.3.4 Deciding similarity . . . 1083
10.3.5 Computing a Jordan form . . . 1084
10.3.6 Exercises . . . 1087
10.3.6.1 A. Check your understanding . . . 1087
10.3.6.2 B. Practice . . . 1088
10.3.6.3 C. Going deeper . . . 1091

10.4 The Real Jordan Form and the Weyr Form 1092
10.4.1 A block for a conjugate pair . . . 1092
10.4.2 The real Jordan form . . . 1096
10.4.3 The Weyr form . . . 1101
10.4.4 Exercises . . . 1106
10.4.4.1 A. Check your understanding . . . 1106
10.4.4.2 B. Practice . . . 1107
10.4.4.3 C. Going deeper . . . 1109

10.5 Cyclic Subspaces and Companion Matrices 1111
10.5.1 The subspace generated by one vector . . . 1111
10.5.2 The annihilator of a vector . . . 1113
10.5.3 Companion matrices . . . 1115
10.5.4 Cyclic operators . . . 1116
10.5.5 Exercises . . . 1121
10.5.5.1 A. Check your understanding . . . 1121
10.5.5.2 B. Practice . . . 1122
10.5.5.3 C. Going deeper . . . 1124

10.6 The Rational Canonical Form 1126
10.6.1 The plan . . . 1126
10.6.2 One prime at a time . . . 1126
10.6.3 Invariant factors . . . 1129

10.6.4 The rational canonical form . . . 1135
10.6.5 Elementary divisors, and the way back to Jordan . . . 1136
10.6.6 Similarity does not depend on the field . . . 1138
10.6.7 Computing the forms . . . 1139
10.6.8 Exercises . . . 1142
10.6.8.1 A. Check your understanding . . . 1142
10.6.8.2 B. Practice . . . 1142
10.6.8.3 C. Going deeper . . . 1145

10.7 Polynomial Matrices and the Smith Normal Form 1147
10.7.1 The space as a module over the polynomials . . . 1147
10.7.2 Row and column operations with polynomial multipliers . . . 1148
10.7.3 The Smith normal form . . . 1150
10.7.4 Similarity read off from a polynomial matrix . . . 1155
10.7.5 Every matrix is similar to its transpose . . . 1158
10.7.6 Exercises . . . 1159
10.7.6.1 A. Check your understanding . . . 1159
10.7.6.2 B. Practice . . . 1160
10.7.6.3 C. Going deeper . . . 1163

10.8 Semisimple Operators and the Jordan–Chevalley Decomposition 1164
10.8.1 Diagonalizable once the field is big enough . . . 1165
10.8.2 Splitting an operator in two . . . 1168
10.8.3 Exercises . . . 1171
10.8.3.1 A. Check your understanding . . . 1171
10.8.3.2 B. Practice . . . 1172
10.8.3.3 C. Going deeper . . . 1174

10.9 Functions of Matrices, the Exponential, and Linear Differential Equations 1175
10.9.1 What a polynomial does to a Jordan block . . . 1176
10.9.2 Functions of a matrix . . . 1177
10.9.3 The matrix exponential . . . 1180
10.9.4 Linear systems of differential equations . . . 1185
10.9.5 Exercises . . . 1188
10.9.5.1 A. Check your understanding . . . 1188
10.9.5.2 B. Practice . . . 1188
10.9.5.3 C. Going deeper . . . 1191

10.10 Convergent and Power-Bounded Matrices 1193
10.10.1 Entrywise limits and bounds . . . 1193
10.10.2 Powers tending to zero . . . 1195
10.10.3 Powers that stay bounded . . . 1196
10.10.4 Powers that converge . . . 1196
10.10.5 Markov chains, finished . . . 1199
10.10.6 The geometric series of a matrix . . . 1200
10.10.7 Exercises . . . 1202
10.10.7.1 A. Check your understanding . . . 1202
10.10.7.2 B. Practice . . . 1203
10.10.7.3 C. Going deeper . . . 1205

11 Chapter 11 **1208**

Inner Product Spaces **1208**
11.0.1 Roadmap . . . 1209
11.0.2 Named moves . . . 1209

11.1 Inner Products and Norms **1209**
11.1.1 The field for this chapter . . . 1209
11.1.2 From the dot product to the axioms . . . 1210
11.1.3 Examples . . . 1212
11.1.4 Nonexamples, by minimal change . . . 1213
11.1.5 Norms . . . 1214
11.1.6 Orthogonality and Pythagoras . . . 1215
11.1.7 The Cauchy–Schwarz inequality . . . 1216
11.1.8 Exercises . . . 1219
11.1.8.1 A. Check your understanding . . . 1219
11.1.8.2 B. Practice . . . 1220
11.1.8.3 C. Going deeper . . . 1222

11.2 Orthonormal Bases and Gram–Schmidt **1224**
11.2.1 Orthonormal lists . . . 1224
11.2.2 Coordinates in an orthonormal basis . . . 1226
11.2.3 The Gram–Schmidt process . . . 1227
11.2.4 Two worked examples . . . 1229
11.2.5 All inner product spaces of the same dimension look alike . . . 1231
11.2.6 Gram matrices . . . 1231
11.2.7 Exercises . . . 1234
11.2.7.1 A. Check your understanding . . . 1234
11.2.7.2 B. Practice . . . 1234
11.2.7.3 C. Going deeper . . . 1237

11.3 Orthogonal Complements and Projections **1238**
11.3.1 Orthogonal complements . . . 1238
11.3.2 The orthogonal decomposition . . . 1240
11.3.3 Orthogonal projection . . . 1241
11.3.4 Best approximation and distance . . . 1243
11.3.5 Three traps . . . 1245
11.3.6 Exercises . . . 1246
11.3.6.1 A. Check your understanding . . . 1246
11.3.6.2 B. Practice . . . 1246
11.3.6.3 C. Going deeper . . . 1248

11.4 Least Squares and Minimum-Norm Solutions **1250**
11.4.1 The residual, and a piece of notation . . . 1250
11.4.2 The normal equations . . . 1251
11.4.3 When is the least-squares solution unique? . . . 1253
11.4.4 Fitting a line and a parabola . . . 1254
11.4.5 The minimum-norm solution . . . 1256
11.4.6 Exercises . . . 1257
11.4.6.1 A. Check your understanding . . . 1257
11.4.6.2 B. Practice . . . 1258
11.4.6.3 C. Going deeper . . . 1259

11.5 The Riesz Representation Theorem **1262**
11.5.1 Every functional is an inner product . . . 1262
11.5.2 The Riesz map, and the data Chapter 5 was missing . . . 1264
11.5.3 Annihilators become orthogonal complements . . . 1265
11.5.4 What happens in infinite dimension . . . 1266
11.5.5 Exercises . . . 1266
11.5.5.1 A. Check your understanding . . . 1266
11.5.5.2 B. Practice . . . 1267
11.5.5.3 C. Going deeper . . . 1268

11.6 The Adjoint **1270**
11.6.1 Paying to cross the inner product . . . 1270
11.6.2 The adjoint . . . 1271
11.6.3 Arithmetic of the adjoint . . . 1274
11.6.4 The matrix of the adjoint . . . 1275
11.6.5 The adjoint and the dual map . . . 1277
11.6.6 The four fundamental subspaces . . . 1278
11.6.7 Self-adjoint and normal operators . . . 1281
11.6.8 Exercises . . . 1283
11.6.8.1 A. Check your understanding . . . 1283
11.6.8.2 B. Practice . . . 1284
11.6.8.3 C. Going deeper . . . 1286

11.7 Isometries and Unitary Maps **1288**
11.7.1 Maps that do not change length . . . 1288
11.7.2 Five ways to say the same thing . . . 1289
11.7.3 Unitary and orthogonal maps . . . 1292
11.7.4 The Frobenius norm and unitary factors . . . 1293
11.7.5 Rotations and reflections of the plane . . . 1295
11.7.6 Rotations of space . . . 1297
11.7.7 Rigid motions . . . 1299
11.7.8 Exercises . . . 1300
11.7.8.1 A. Check your understanding . . . 1300
11.7.8.2 B. Practice . . . 1301
11.7.8.3 C. Going deeper . . . 1302

11.8 QR, Householder, and Hessenberg **1304**
11.8.1 Gram–Schmidt in matrix form . . . 1305
11.8.2 Least squares through QR . . . 1307
11.8.3 Reflections . . . 1308
11.8.4 QR by reflections . . . 1311
11.8.5 Givens rotations . . . 1314
11.8.6 Hessenberg form . . . 1315
11.8.7 Exercises . . . 1318
11.8.7.1 A. Check your understanding . . . 1318
11.8.7.2 B. Practice . . . 1318
11.8.7.3 C. Going deeper . . . 1321

11.9 Bessel, Parseval, and Fourier Series **1324**
11.9.1 Coefficients against an orthonormal list . . . 1324
11.9.2 The trigonometric system . . . 1326
11.9.3 A worked computation . . . 1329

11.9.4 What best approximation does and does not say . . . 1330
11.9.5 Legendre and Chebyshev . . . 1331
11.9.6 Exercises . . . 1332
11.9.6.1 A. Check your understanding . . . 1332
11.9.6.2 B. Practice . . . 1333
11.9.6.3 C. Going deeper . . . 1335

11.10 Orthogonal Polynomials 1337
11.10.1 Two ways to make an inner product on polynomials . . . 1337
11.10.2 Orthogonal polynomial sequences . . . 1338
11.10.3 The three-term recurrence . . . 1340
11.10.4 Where the roots are . . . 1342
11.10.5 Gaussian quadrature . . . 1343
11.10.6 The classical families . . . 1346
11.10.7 Exercises . . . 1346
11.10.7.1 A. Check your understanding . . . 1346
11.10.7.2 B. Practice . . . 1347
11.10.7.3 C. Going deeper . . . 1349

12 Chapter 12 1351

Spectral Theory 1351
12.0.1 Roadmap . . . 1352
12.0.2 Named moves . . . 1352

12.1 Schur Triangularization 1353
12.1.1 Why the obvious route fails . . . 1353
12.1.2 The theorem . . . 1354
12.1.3 The matrix form . . . 1355
12.1.4 Reading the diagonal . . . 1358
12.1.5 The real version . . . 1359
12.1.6 Exercises . . . 1361
12.1.6.1 A. Check your understanding . . . 1361
12.1.6.2 B. Practice . . . 1362
12.1.6.3 C. Going deeper . . . 1364

12.2 Self-Adjoint Operators 1366
12.2.1 The operators fixed by starring . . . 1366
12.2.2 Eigenvalues are real . . . 1367
12.2.3 Existence over the real numbers . . . 1368
12.2.4 Eigenvectors for different eigenvalues are orthogonal . . . 1369
12.2.5 Invariant complements . . . 1371
12.2.6 Two zero tests, and the gap between the fields . . . 1371
12.2.7 Exercises . . . 1373
12.2.7.1 A. Check your understanding . . . 1373
12.2.7.2 B. Practice . . . 1374
12.2.7.3 C. Going deeper . . . 1376

12.3 Normal Operators 1378
12.3.1 The hypothesis that self-adjointness overshoots . . . 1378
12.3.2 Normality is a statement about lengths . . . 1380
12.3.3 What normality buys . . . 1382

12.3.4 Exercises . . . 1386
12.3.4.1 A. Check your understanding . . . 1386
12.3.4.2 B. Practice . . . 1386
12.3.4.3 C. Going deeper . . . 1388

12.4 The Complex Spectral Theorem 1391
12.4.1 Triangular plus normal forces diagonal . . . 1391
12.4.2 The theorem . . . 1392
12.4.3 Reading the class off the spectrum . . . 1394
12.4.4 Exactly what is unique . . . 1396
12.4.5 Exercises . . . 1397
12.4.5.1 A. Check your understanding . . . 1397
12.4.5.2 B. Practice . . . 1398
12.4.5.3 C. Going deeper . . . 1401

12.5 The Real Spectral Theorem 1402
12.5.1 One eigenvector, then its orthogonal complement . . . 1403
12.5.2 The negative result: normal is not enough over the reals . . . 1404
12.5.3 What a real normal operator does look like . . . 1406
12.5.4 A symmetric matrix as a sum of projections . . . 1408
12.5.5 The principal axis theorem . . . 1409
12.5.6 Exercises . . . 1411
12.5.6.1 A. Check your understanding . . . 1411
12.5.6.2 B. Practice . . . 1412
12.5.6.3 C. Going deeper . . . 1414

12.6 The Shape of an Orthogonal Matrix 1416
12.6.1 From the real normal form to the canonical form . . . 1416
12.6.2 Rotations of space . . . 1418
12.6.3 The cross-product matrix . . . 1418
12.6.4 Rodrigues' formula . . . 1420
12.6.5 Reading off the axis and the angle . . . 1422
12.6.6 Reflections generate . . . 1424
12.6.7 Composing rotations . . . 1425
12.6.8 Exercises . . . 1426
12.6.8.1 A. Check your understanding . . . 1426
12.6.8.2 B. Practice . . . 1427
12.6.8.3 C. Going deeper . . . 1430

12.7 Commuting Families 1432
12.7.1 What Chapter 9 already gives, and what is missing . . . 1432
12.7.2 One orthonormal basis for the whole family . . . 1433
12.7.3 Sums and products stay normal . . . 1437
12.7.4 Without normality: simultaneous triangularization . . . 1438
12.7.5 Exercises . . . 1439
12.7.5.1 A. Check your understanding . . . 1439
12.7.5.2 B. Practice . . . 1439
12.7.5.3 C. Going deeper . . . 1441

12.8 Spectral Resolutions and Functions of a Normal Operator 1443
12.8.1 Grouping an eigenbasis by eigenvalue . . . 1444
12.8.2 Orthogonal projections versus oblique ones . . . 1446

12.8.3 What makes a resolution the resolution . . . 1447
12.8.4 Functions of an operator . . . 1449
12.8.5 Square roots . . . 1452
12.8.6 The exponential and the unitary group . . . 1455
12.8.7 Exercises . . . 1457
12.8.7.1 A. Check your understanding . . . 1457
12.8.7.2 B. Practice . . . 1458
12.8.7.3 C. Going deeper . . . 1460

12.9 Circulants and the Discrete Fourier Transform 1462
12.9.1 Matrices whose rows walk around . . . 1463
12.9.2 The eigenvectors of the shift . . . 1466
12.9.3 One unitary for every circulant . . . 1468
12.9.4 Multiplying circulants is convolution . . . 1470
12.9.5 Solving a circulant system, and a word about speed . . . 1472
12.9.6 Exercises . . . 1473
12.9.6.1 A. Check your understanding . . . 1473
12.9.6.2 B. Practice . . . 1474
12.9.6.3 C. Going deeper . . . 1476

12.10 The Sylvester Equation 1478
12.10.1 The Sylvester operator . . . 1478
12.10.2 The spectrum of a Kronecker sum . . . 1479
12.10.3 When the equation is uniquely solvable . . . 1481
12.10.4 Clearing the corner . . . 1483
12.10.5 Commuting with a matrix of disjoint spectra . . . 1485
12.10.6 Lyapunov's equation . . . 1486
12.10.7 Exercises . . . 1487
12.10.7.1 A. Check your understanding . . . 1487
12.10.7.2 B. Practice . . . 1488
12.10.7.3 C. Going deeper . . . 1490

12.11 How Far from Normal 1492
12.11.1 The yardstick . . . 1493
12.11.2 Schur's inequality . . . 1493
12.11.3 The departure from normality . . . 1495
12.11.4 Eigenvalues that will not sit still . . . 1497
12.11.5 The characterizations collected . . . 1498
12.11.6 Exercises . . . 1500
12.11.6.1 A. Check your understanding . . . 1500
12.11.6.2 B. Practice . . . 1500
12.11.6.3 C. Going deeper . . . 1502

13 Chapter 13 1505

Positive Matrices and the Singular Value Decomposition 1505
13.0.1 Roadmap . . . 1506
13.0.2 Named moves . . . 1507

13.1 Positive Semidefinite and Positive Definite Matrices 1507
13.1.1 Which self-adjoint matrices deserve to be called nonnegative? . . . 1507
13.1.2 Positivity survives a change of variable . . . 1510

13.1.3 The equivalent descriptions . . . 1510
13.1.4 The definite case, and Sylvester's criterion . . . 1513
13.1.5 Two by two, and where these matrices come from . . . 1515
13.1.6 The invertible matrix theorem grows . . . 1516
13.1.7 Exercises . . . 1517
13.1.7.1 A. Check your understanding . . . 1517
13.1.7.2 B. Practice . . . 1518
13.1.7.3 C. Going deeper . . . 1519

13.2 Square Roots and the Cholesky Factorization 1521
13.2.1 The positive square root . . . 1521
13.2.2 The absolute value of a matrix . . . 1524
13.2.3 Cholesky: elimination on a positive definite matrix . . . 1525
13.2.4 Exercises . . . 1529
13.2.4.1 A. Check your understanding . . . 1529
13.2.4.2 B. Practice . . . 1530
13.2.4.3 C. Going deeper . . . 1532

13.3 Gram Matrices 1533
13.3.1 Which matrices are Gram matrices? . . . 1534
13.3.2 The Gram determinant is a squared volume . . . 1536
13.3.3 Distance to a subspace, by determinants . . . 1538
13.3.4 Exercises . . . 1540
13.3.4.1 A. Check your understanding . . . 1540
13.3.4.2 B. Practice . . . 1540
13.3.4.3 C. Going deeper . . . 1542

13.4 Two Forms at Once 1543
13.4.1 Similarity is for operators, congruence is for forms . . . 1544
13.4.2 Diagonalizing two forms at once . . . 1544
13.4.3 The generalized eigenvalue problem . . . 1545
13.4.4 When neither form is definite . . . 1548
13.4.5 Exercises . . . 1549
13.4.5.1 A. Check your understanding . . . 1549
13.4.5.2 B. Practice . . . 1549
13.4.5.3 C. Going deeper . . . 1551

13.5 The Loewner Order 1553
13.5.1 Comparing two self-adjoint matrices . . . 1553
13.5.2 What survives . . . 1554
13.5.3 What fails: squaring . . . 1556
13.5.4 Exercises . . . 1558
13.5.4.1 A. Check your understanding . . . 1558
13.5.4.2 B. Practice . . . 1559
13.5.4.3 C. Going deeper . . . 1560

13.6 Positive Matrices in Blocks 1562
13.6.1 What the blocks of a positive matrix must satisfy . . . 1562
13.6.2 Testing one block at a time . . . 1563
13.6.3 A positive definite certificate for stability . . . 1567
13.6.4 Exercises . . . 1572
13.6.4.1 A. Check your understanding . . . 1572

13.6.4.2 B. Practice . . . 1572
13.6.4.3 C. Going deeper . . . 1574

13.7 Determinant Inequalities for Positive Matrices 1575
13.7.1 Fischer's inequality . . . 1576
13.7.2 Hadamard's inequality . . . 1578
13.7.3 Exercises . . . 1581
13.7.3.1 A. Check your understanding . . . 1581
13.7.3.2 B. Practice . . . 1582
13.7.3.3 C. Going deeper . . . 1583

13.8 Products Entry by Entry 1583
13.8.1 The entrywise product . . . 1584
13.8.2 The Schur product theorem . . . 1585
13.8.3 Oppenheim's inequality . . . 1588
13.8.4 Where this gets used . . . 1590
13.8.5 Exercises . . . 1590
13.8.5.1 A. Check your understanding . . . 1590
13.8.5.2 B. Practice . . . 1591
13.8.5.3 C. Going deeper . . . 1591

13.9 The Singular Value Decomposition 1593
13.9.1 Where to look when there are no eigenvalues . . . 1593
13.9.2 The theorem . . . 1595
13.9.3 What is unique and what is not . . . 1597
13.9.4 The compact form, and the four subspaces . . . 1598
13.9.5 Two computations and one comparison . . . 1600
13.9.6 Exercises . . . 1601
13.9.6.1 A. Check your understanding . . . 1601
13.9.6.2 B. Practice . . . 1602
13.9.6.3 C. Going deeper . . . 1604

13.10 Polar Decomposition and Takagi's Factorization 1606
13.10.1 Rotation times stretch . . . 1606
13.10.2 The nearest unitary matrix . . . 1609
13.10.3 Takagi's factorization . . . 1610
13.10.4 Exercises . . . 1614
13.10.4.1 A. Check your understanding . . . 1614
13.10.4.2 B. Practice . . . 1615
13.10.4.3 C. Going deeper . . . 1616

13.11 Low-Rank Approximation 1618
13.11.1 Cutting the sum short . . . 1618
13.11.2 How much a subspace can hold . . . 1620
13.11.3 The Eckart–Young theorem . . . 1621
13.11.4 A worked example . . . 1623
13.11.5 Best-fitting subspaces . . . 1624
13.11.6 Exercises . . . 1625
13.11.6.1 A. Check your understanding . . . 1625
13.11.6.2 B. Practice . . . 1626
13.11.6.3 C. Going deeper . . . 1627

13.12 The Moore–Penrose Pseudoinverse 1629

13.12.1 One matrix instead of three formulas . . . 1629
13.12.2 Inverting what can be inverted . . . 1630
13.12.3 The four Penrose conditions . . . 1633
13.12.4 What the pseudoinverse does . . . 1634
13.12.5 Least squares and minimum norm, in one matrix . . . 1636
13.12.6 The block criterion, finished . . . 1639
13.12.7 Exercises . . . 1640
13.12.7.1 A. Check your understanding . . . 1640
13.12.7.2 B. Practice . . . 1641
13.12.7.3 C. Going deeper . . . 1642

13.13 Projections Revisited **1644**
13.13.1 Which projections are orthogonal? . . . 1644
13.13.2 The block that measures obliqueness . . . 1647
13.13.2.1 The two-dimensional case, in full . . . 1648
13.13.3 Looking back at Part III . . . 1650
13.13.4 Exercises . . . 1650
13.13.4.1 A. Check your understanding . . . 1650
13.13.4.2 B. Practice . . . 1651
13.13.4.3 C. Going deeper . . . 1653

14 Chapter 14 **1656**

Bilinear and Quadratic Forms **1656**
14.0.1 Roadmap . . . 1657
14.0.2 Named moves . . . 1657

14.1 Bilinear Forms and Their Matrices **1658**
14.1.1 Pairing without measuring . . . 1658
14.1.2 The matrix of a form . . . 1660
14.1.3 The form as a map into the dual space . . . 1663
14.1.4 Exercises . . . 1666
14.1.4.1 A. Check your understanding . . . 1666
14.1.4.2 B. Practice . . . 1666
14.1.4.3 C. Going deeper . . . 1668

14.2 Congruence **1670**
14.2.1 The change-of-basis rule for a form . . . 1670
14.2.2 Congruent matrices . . . 1671
14.2.3 What congruence preserves . . . 1673
14.2.4 What congruence destroys . . . 1675
14.2.5 Where the two relations meet . . . 1676
14.2.6 Exercises . . . 1677
14.2.6.1 A. Check your understanding . . . 1677
14.2.6.2 B. Practice . . . 1677
14.2.6.3 C. Going deeper . . . 1679

14.3 Three Kinds of Form **1681**
14.3.1 Symmetric forms . . . 1681
14.3.2 Alternating and skew-symmetric forms . . . 1682
14.3.3 Where characteristic 2 earns its keep . . . 1683
14.3.4 Splitting a form into two halves . . . 1685

14.3.5 The quadratic form . . . 1685
14.3.6 Hermitian forms . . . 1688
14.3.7 Exercises . . . 1689
14.3.7.1 A. Check your understanding . . . 1689
14.3.7.2 B. Practice . . . 1690
14.3.7.3 C. Going deeper . . . 1691

14.4 Completing the Square 1693
14.4.1 Orthogonality without an inner product . . . 1693
14.4.2 The theorem, and the case that breaks it . . . 1694
14.4.3 The algorithm: symmetric row and column operations . . . 1696
14.4.4 Over the complex numbers, and over a general field . . . 1698
14.4.5 Exercises . . . 1699
14.4.5.1 A. Check your understanding . . . 1699
14.4.5.2 B. Practice . . . 1700
14.4.5.3 C. Going deeper . . . 1701

14.5 Sylvester's Law of Inertia 1703
14.5.1 Normalizing the diagonal . . . 1703
14.5.2 The law of inertia . . . 1704
14.5.3 The classification over the reals . . . 1705
14.5.4 Reconciling congruence with the spectral theorem . . . 1706
14.5.5 Computing a signature . . . 1707
14.5.6 Exercises . . . 1708
14.5.6.1 A. Check your understanding . . . 1708
14.5.6.2 B. Practice . . . 1709
14.5.6.3 C. Going deeper . . . 1710

14.6 Definiteness and Second Derivatives 1713
14.6.1 Five words for the sign of a form . . . 1713
14.6.2 The second-derivative test . . . 1715
14.6.3 Exercises . . . 1719
14.6.3.1 A. Check your understanding . . . 1719
14.6.3.2 B. Practice . . . 1719
14.6.3.3 C. Going deeper . . . 1720

14.7 Where the Form Vanishes 1721
14.7.1 An intuition to give up . . . 1722
14.7.2 Isotropic vectors and totally isotropic subspaces . . . 1722
14.7.3 Splitting off a nondegenerate piece . . . 1725
14.7.4 Hyperbolic planes . . . 1726
14.7.5 The Witt decomposition . . . 1727
14.7.6 The two familiar fields . . . 1730
14.7.7 Exercises . . . 1731
14.7.7.1 A. Check your understanding . . . 1731
14.7.7.2 B. Practice . . . 1731
14.7.7.3 C. Going deeper . . . 1733

14.8 Witt's Extension and Cancellation Theorems 1735
14.8.1 Isometries of forms . . . 1735
14.8.2 Reflections . . . 1736
14.8.3 Witt's extension theorem . . . 1738

14.8.4 Cancellation . . . 1741
14.8.5 The Witt decomposition is unique . . . 1742
14.8.6 Exercises . . . 1743
14.8.6.1 A. Check your understanding . . . 1743
14.8.6.2 B. Practice . . . 1744
14.8.6.3 C. Going deeper . . . 1745

14.9 Alternating Forms and Darboux Bases 1747
14.9.1 The two-dimensional model . . . 1747
14.9.2 Splitting off a hyperbolic pair . . . 1749
14.9.3 The standard form . . . 1750
14.9.4 Running the algorithm . . . 1752
14.9.5 What the standard form classifies . . . 1753
14.9.6 The determinant, and the Pfaffian . . . 1754
14.9.7 Symmetric against alternating . . . 1755
14.9.8 Exercises . . . 1756
14.9.8.1 A. Check your understanding . . . 1756
14.9.8.2 B. Practice . . . 1756
14.9.8.3 C. Going deeper . . . 1758

14.10 Groups Defined by Forms 1760
14.10.1 The isometry group of a form . . . 1761
14.10.2 Congruent forms give conjugate groups . . . 1762
14.10.3 The four families . . . 1763
14.10.4 Determinants . . . 1764
14.10.5 Counting parameters . . . 1765
14.10.6 Boundedness and its absence . . . 1766
14.10.7 Exercises . . . 1767
14.10.7.1 A. Check your understanding . . . 1767
14.10.7.2 B. Practice . . . 1768
14.10.7.3 C. Going deeper . . . 1770

14.11 Minkowski Space 1772
14.11.1 The form of special relativity . . . 1772
14.11.2 Timelike, spacelike, and null . . . 1773
14.11.3 Hyperbolic rotations . . . 1775
14.11.4 Circular and hyperbolic, side by side . . . 1777
14.11.5 The triangle inequality, reversed . . . 1777
14.11.6 Exercises . . . 1779
14.11.6.1 A. Check your understanding . . . 1779
14.11.6.2 B. Practice . . . 1780
14.11.6.3 C. Going deeper . . . 1781

14.12 A First Look at Clifford Algebras 1782
14.12.1 Making a form into a square . . . 1783
14.12.2 Algebras . . . 1783
14.12.3 The definition . . . 1783
14.12.4 Three algebras, worked out . . . 1785
14.12.5 How big is it . . . 1787
14.12.6 What Clifford algebras are for . . . 1789
14.12.7 Exercises . . . 1790
14.12.7.1 A. Check your understanding . . . 1790

14.12.7.2 B. Practice . . . 1790
14.12.7.3 C. Going deeper . . . 1791

15 Chapter 15 **1794**

Tensors and Exterior Algebra **1794**
15.0.1 Roadmap . . . 1795
15.0.2 Named moves . . . 1796

15.1 Multilinear Maps **1796**
15.1.1 From forms to maps . . . 1797
15.1.2 Multilinear is not linear . . . 1799
15.1.3 Values on basis tuples . . . 1799
15.1.4 The space of multilinear maps . . . 1800
15.1.5 The problem this chapter solves . . . 1803
15.1.6 Exercises . . . 1803
15.1.6.1 A. Check your understanding . . . 1803
15.1.6.2 B. Practice . . . 1804
15.1.6.3 C. Going deeper . . . 1806

15.2 The Tensor Product **1808**
15.2.1 What the object has to do . . . 1808
15.2.2 Any two tensor products are the same . . . 1809
15.2.3 Building one: a free space modulo relations . . . 1810
15.2.4 A basis and a dimension . . . 1813
15.2.5 Coordinates: the tensor product of coordinate spaces . . . 1814
15.2.6 Three things a tensor is not . . . 1816
15.2.6.1 Not every tensor is simple . . . 1816
15.2.6.2 A simple tensor vanishes only for an obvious reason . . . 1817
15.2.6.3 A simple tensor does not remember its factors . . . 1817
15.2.7 Exercises . . . 1818
15.2.7.1 A. Check your understanding . . . 1818
15.2.7.2 B. Practice . . . 1819
15.2.7.3 C. Going deeper . . . 1821

15.3 Tensor Products of Maps **1822**
15.3.1 A map on each factor gives a map on the product . . . 1822
15.3.2 The matrix of a tensor product of maps . . . 1826
15.3.3 What Chapter 8 was computing . . . 1828
15.3.4 Maps as tensors . . . 1831
15.3.5 Exercises . . . 1833
15.3.5.1 A. Check your understanding . . . 1833
15.3.5.2 B. Practice . . . 1834
15.3.5.3 C. Going deeper . . . 1836

15.4 Contraction and the Trace **1838**
15.4.1 Evaluation, made linear . . . 1838
15.4.2 Contraction is the trace . . . 1839
15.4.3 Tensors with many slots . . . 1841
15.4.4 Exercises . . . 1843
15.4.4.1 A. Check your understanding . . . 1843
15.4.4.2 B. Practice . . . 1843

15.4.4.3 C. Going deeper . . . 1845

15.5 Index Notation and the Transformation Law 1846
15.5.1 Components, upstairs and downstairs . . . 1846
15.5.2 How the components change . . . 1847
15.5.3 The classical definition . . . 1849
15.5.4 An array is not a tensor . . . 1851
15.5.5 Exercises . . . 1852
15.5.5.1 A. Check your understanding . . . 1852
15.5.5.2 B. Practice . . . 1853
15.5.5.3 C. Going deeper . . . 1854

15.6 The Tensor Algebra 1855
15.6.1 The k-th tensor power . . . 1856
15.6.2 A basis, and the dimension . . . 1857
15.6.3 Stacking the powers: the tensor algebra . . . 1859
15.6.4 The universal property of the tensor algebra . . . 1861
15.6.5 Exercises . . . 1863
15.6.5.1 A. Check your understanding . . . 1863
15.6.5.2 B. Practice . . . 1863
15.6.5.3 C. Going deeper . . . 1865

15.7 Symmetric Powers 1866
15.7.1 Symmetric multilinear maps . . . 1867
15.7.2 The symmetric power . . . 1867
15.7.3 A basis, and the dimension . . . 1869
15.7.4 The symmetric algebra is a polynomial algebra . . . 1872
15.7.5 Symmetric powers versus symmetric tensors . . . 1874
15.7.6 Exercises . . . 1877
15.7.6.1 A. Check your understanding . . . 1877
15.7.6.2 B. Practice . . . 1877
15.7.6.3 C. Going deeper . . . 1878

15.8 Exterior Powers 1880
15.8.1 Alternating multilinear maps . . . 1881
15.8.2 Which relation to divide by . . . 1882
15.8.3 The exterior power . . . 1883
15.8.4 A basis, and the dimension . . . 1885
15.8.5 Independence detects wedges . . . 1887
15.8.6 Exercises . . . 1888
15.8.6.1 A. Check your understanding . . . 1888
15.8.6.2 B. Practice . . . 1889
15.8.6.3 C. Going deeper . . . 1890

15.9 The Determinant as a Top Exterior Power 1892
15.9.1 Wedging with a fixed vector . . . 1892
15.9.2 Induced maps on exterior powers . . . 1893
15.9.3 The top exterior power . . . 1895
15.9.4 What the third construction adds . . . 1896
15.9.5 The coordinates of a wedge . . . 1898
15.9.6 Laplace expansion . . . 1900
15.9.7 Exercises . . . 1901

15.9.7.1 A. Check your understanding . . . 1901
15.9.7.2 B. Practice . . . 1901
15.9.7.3 C. Going deeper . . . 1903

15.10 The Exterior Algebra, and Clifford at Last 1904
15.10.1 Stacking the exterior powers . . . 1904
15.10.2 Decomposable elements . . . 1908
15.10.3 Algebras modulo relations . . . 1910
15.10.4 The Clifford algebra, constructed . . . 1911
15.10.5 The dimension, at last . . . 1912
15.10.6 Exercises . . . 1918
15.10.6.1 A. Check your understanding . . . 1918
15.10.6.2 B. Practice . . . 1919
15.10.6.3 C. Going deeper . . . 1920

15.11 Tensors in Quantum Mechanics 1922
15.11.1 What is quoted and what is proved . . . 1922
15.11.2 Entanglement is failure of simplicity . . . 1923
15.11.3 The Schmidt decomposition . . . 1925
15.11.4 Looking back at Part IV . . . 1929
15.11.5 Exercises . . . 1930
15.11.5.1 A. Check your understanding . . . 1930
15.11.5.2 B. Practice . . . 1930
15.11.5.3 C. Going deeper . . . 1932

16 Chapter 16 1935

Norms and Matrix Analysis 1935
16.0.1 The analysis this chapter imports . . . 1935
16.0.2 What you need . . . 1936
16.0.3 Roadmap . . . 1937
16.0.4 Named moves . . . 1938

16.1 Norms 1938
16.1.1 What measurement actually needs . . . 1938
16.1.2 Unit balls . . . 1941
16.1.3 Which norms come from an inner product . . . 1943
16.1.4 Exercises . . . 1947
16.1.4.1 A. Check your understanding . . . 1947
16.1.4.2 B. Practice . . . 1948
16.1.4.3 C. Going deeper . . . 1949

16.2 All Norms Agree 1951
16.2.1 Trapped between multiples . . . 1951
16.2.2 Compactness enters the book . . . 1952
16.2.3 What the theorem buys . . . 1954
16.2.4 Two warnings, and where the theorem stops . . . 1957
16.2.5 Exercises . . . 1958
16.2.5.1 A. Check your understanding . . . 1958
16.2.5.2 B. Practice . . . 1959
16.2.5.3 C. Going deeper . . . 1960

16.3 Measuring a Matrix **1962**
16.3.1 What a matrix norm has to do . . . 1962
16.3.2 The norm a vector norm induces . . . 1964
16.3.3 The three formulas . . . 1967
16.3.4 What submultiplicativity does not buy . . . 1970
16.3.5 Exercises . . . 1971
16.3.5.1 A. Check your understanding . . . 1971
16.3.5.2 B. Practice . . . 1972
16.3.5.3 C. Going deeper . . . 1973

16.4 The Spectral Radius **1975**
16.4.1 The number Chapter 10 named . . . 1975
16.4.2 Every norm dominates the spectral radius . . . 1976
16.4.3 A norm as close to the spectral radius as we like . . . 1978
16.4.4 Gelfand's formula . . . 1979
16.4.5 What is new, and what is not . . . 1981
16.4.6 Exercises . . . 1983
16.4.6.1 A. Check your understanding . . . 1983
16.4.6.2 B. Practice . . . 1983
16.4.6.3 C. Going deeper . . . 1985

16.5 Inverting a Perturbation of the Identity **1986**
16.5.1 The geometric series, with a number attached . . . 1987
16.5.2 Which theorem is stronger . . . 1989
16.5.3 The invertible matrices form an open set . . . 1990
16.5.4 Perturbing a solve . . . 1992
16.5.5 Exercises . . . 1994
16.5.5.1 A. Check your understanding . . . 1994
16.5.5.2 B. Practice . . . 1994
16.5.5.3 C. Going deeper . . . 1996

16.6 Bounds for the Exponential, and Two Derivatives **1997**
16.6.1 How big is the exponential . . . 1998
16.6.2 Decay in the stable case . . . 2000
16.6.3 Differentiating a matrix-valued function . . . 2003
16.6.4 The derivative of the inverse . . . 2004
16.6.5 Jacobi's formula . . . 2005
16.6.6 Exercises . . . 2007
16.6.6.1 A. Check your understanding . . . 2007
16.6.6.2 B. Practice . . . 2008
16.6.6.3 C. Going deeper . . . 2010

16.7 Eigenvalues Move Continuously **2012**
16.7.1 From matrices to polynomials . . . 2012
16.7.2 Why the answer must be about the multiset . . . 2013
16.7.3 The imported fact . . . 2014
16.7.4 Eigenvalues, and the spectral radius . . . 2015
16.7.5 The limit argument, made routine . . . 2016
16.7.6 The warning: continuous is not Lipschitz . . . 2017
16.7.7 Exercises . . . 2019
16.7.7.1 A. Check your understanding . . . 2019
16.7.7.2 B. Practice . . . 2020

16.7.7.3 C. Going deeper . . . 2021

16.8 How Wrong Can the Answer Be 2023
16.8.1 A recurring product . . . 2024
16.8.2 What the condition number predicts . . . 2027
16.8.3 Why the normal equations are worse . . . 2030
16.8.4 Conditioning is a property of the problem . . . 2031
16.8.5 What this chapter has settled . . . 2032
16.8.6 Exercises . . . 2033
16.8.6.1 A. Check your understanding . . . 2033
16.8.6.2 B. Practice . . . 2033
16.8.6.3 C. Going deeper . . . 2035

17 Chapter 17 2038

Variational Principles and Interlacing 2038
17.0.1 The analysis this chapter imports . . . 2039
17.0.2 What you need . . . 2039
17.0.3 Roadmap . . . 2040
17.0.4 Named moves . . . 2041

17.1 The Rayleigh Quotient 2041
17.1.1 From a test to an object . . . 2041
17.1.2 What values it takes . . . 2044
17.1.3 Where the quotient is flat . . . 2046
17.1.4 Why the second eigenvalue is not a second maximum . . . 2048
17.1.5 What is still wrong . . . 2050
17.1.6 Exercises . . . 2050
17.1.6.1 A. Check your understanding . . . 2050
17.1.6.2 B. Practice . . . 2051
17.1.6.3 C. Going deeper . . . 2052

17.2 The Min–Max Theorem 2055
17.2.1 Two subspaces that have to meet . . . 2055
17.2.2 The quotient on a subspace . . . 2056
17.2.3 Every eigenvalue is a min–max . . . 2057
17.2.4 The two debts this pays at once . . . 2061
17.2.5 Exercises . . . 2063
17.2.5.1 A. Check your understanding . . . 2063
17.2.5.2 B. Practice . . . 2064
17.2.5.3 C. Going deeper . . . 2065

17.3 How Far Eigenvalues Move 2067
17.3.1 Two small facts about eigenvalue lists . . . 2068
17.3.2 The two families of Weyl inequalities . . . 2069
17.3.3 Perturbing one matrix . . . 2073
17.3.4 Perturbations of small rank . . . 2076
17.3.5 Exercises . . . 2077
17.3.5.1 A. Check your understanding . . . 2077
17.3.5.2 B. Practice . . . 2078
17.3.5.3 C. Going deeper . . . 2080

17.4 Deleting a Row and a Column **2081**
17.4.1 A matrix restricted to a coordinate hyperplane . . . 2082
17.4.2 Interlacing . . . 2082
17.4.3 When is the interlacing strict? . . . 2085
17.4.4 A worked example . . . 2090
17.4.5 Inertia under compression . . . 2092
17.4.6 Interlacing is a Hermitian phenomenon . . . 2093
17.4.7 Exercises . . . 2094
17.4.7.1 A. Check your understanding . . . 2094
17.4.7.2 B. Practice . . . 2094
17.4.7.3 C. Going deeper . . . 2096

17.5 Adding a Rank-One Matrix **2098**
17.5.1 One bump, every eigenvalue up . . . 2099
17.5.2 Where the new eigenvalues actually are . . . 2101
17.5.3 The shifts add up . . . 2105
17.5.4 Exercises . . . 2106
17.5.4.1 A. Check your understanding . . . 2106
17.5.4.2 B. Practice . . . 2106
17.5.4.3 C. Going deeper . . . 2108

17.6 Subspaces and Sums of Eigenvalues **2110**
17.6.1 Restricting a matrix to a subspace . . . 2110
17.6.2 Interlacing without coordinates . . . 2111
17.6.3 The top eigenvalues, summed . . . 2113
17.6.4 The promise from Chapter 13 . . . 2116
17.6.5 Sums are convex; individual eigenvalues are not . . . 2117
17.6.6 Exercises . . . 2119
17.6.6.1 A. Check your understanding . . . 2119
17.6.6.2 B. Practice . . . 2119
17.6.6.3 C. Going deeper . . . 2121

17.7 Arbitrary Index Sets **2123**
17.7.1 What adding up the min–max gives . . . 2123
17.7.2 Flags and adapted lists . . . 2124
17.7.3 Wielandt's minimax principle . . . 2125
17.7.4 Lidskii's inequality . . . 2130
17.7.5 What the flags bought . . . 2134
17.7.6 Exercises . . . 2134
17.7.6.1 A. Check your understanding . . . 2134
17.7.6.2 B. Practice . . . 2135
17.7.6.3 C. Going deeper . . . 2137

17.8 The Diagonal and the Spectrum **2139**
17.8.1 What the diagonal already knows . . . 2140
17.8.2 Majorization . . . 2141
17.8.3 Majorization is an order . . . 2144
17.8.4 Schur's theorem . . . 2145
17.8.5 Horn's converse . . . 2146
17.8.6 Hadamard's inequality from the diagonal . . . 2150
17.8.7 Exercises . . . 2153
17.8.7.1 A. Check your understanding . . . 2153

17.8.7.2 B. Practice . . . 2154
17.8.7.3 C. Going deeper . . . 2156

17.9 The Same Theory for Singular Values 2158
17.9.1 The stretch ratio is a Rayleigh quotient . . . 2159
17.9.2 Each singular value as an optimization . . . 2160
17.9.3 Perturbation, with no hypothesis at all . . . 2162
17.9.4 The best approximation of low rank . . . 2165
17.9.5 How far an invertible matrix is from being singular . . . 2167
17.9.6 Deleting a row . . . 2169
17.9.7 Exercises . . . 2171
17.9.7.1 A. Check your understanding . . . 2171
17.9.7.2 B. Practice . . . 2172
17.9.7.3 C. Going deeper . . . 2174

17.10 Inertia, Again and Additively 2175
17.10.1 A symbol for the triple . . . 2175
17.10.2 Sylvester's law, proved again . . . 2176
17.10.3 Inertia under a perturbation of small rank . . . 2179
17.10.4 Additivity across a Schur complement . . . 2181
17.10.5 Exercises . . . 2185
17.10.5.1 A. Check your understanding . . . 2185
17.10.5.2 B. Practice . . . 2185
17.10.5.3 C. Going deeper . . . 2187

17.11 Which Functions Respect the Order 2189
17.11.1 Functions of a Hermitian matrix . . . 2190
17.11.2 Operator monotone and operator convex functions . . . 2190
17.11.3 The examples we can prove . . . 2192
17.11.4 Eigenvalues move monotonically for every increasing function . . . 2195
17.11.5 Operator convexity, and where the two notions part . . . 2197
17.11.6 What is not proved here . . . 2198
17.11.7 Summary and transfer . . . 2199
17.11.8 Exercises . . . 2201
17.11.8.1 A. Check your understanding . . . 2201
17.11.8.2 B. Practice . . . 2202
17.11.8.3 C. Going deeper . . . 2204

18 Chapter 18 2207

Convexity 2207
18.0.1 The analysis this chapter imports . . . 2207
18.0.2 What you need . . . 2208
18.0.3 Roadmap . . . 2209
18.0.4 Named moves . . . 2209

18.1 Convex Sets and Hulls 2210
18.1.1 Segments, and sets that contain them . . . 2210
18.1.2 Operations that preserve convexity . . . 2214
18.1.3 Convex combinations and the convex hull . . . 2215
18.1.4 Affine combinations and the affine hull . . . 2218
18.1.5 Closed and open sets . . . 2219

18.1.6 Exercises . . . 2222
18.1.6.1 A. Check your understanding . . . 2222
18.1.6.2 B. Practice . . . 2223
18.1.6.3 C. Going deeper . . . 2224

18.2 Carathéodory's Theorem 2225
18.2.1 When there are too many points . . . 2225
18.2.2 The theorem . . . 2227
18.2.3 The hull of a compact set . . . 2230
18.2.4 Radon's theorem and Helly's theorem . . . 2232
18.2.5 Exercises . . . 2234
18.2.5.1 A. Check your understanding . . . 2234
18.2.5.2 B. Practice . . . 2235
18.2.5.3 C. Going deeper . . . 2236

18.3 Separating Hyperplanes 2238
18.3.1 The nearest point of a closed convex set . . . 2239
18.3.2 Hyperplanes and half-spaces . . . 2242
18.3.3 Cutting off a point . . . 2242
18.3.4 Supporting hyperplanes . . . 2243
18.3.5 Separating two convex sets . . . 2246
18.3.6 Where strict separation fails . . . 2247
18.3.7 Extending linear functionals . . . 2249
18.3.8 Exercises . . . 2251
18.3.8.1 A. Check your understanding . . . 2251
18.3.8.2 B. Practice . . . 2251
18.3.8.3 C. Going deeper . . . 2252

18.4 Gauges, Support Functions, and Dual Norms 2254
18.4.1 Gauges . . . 2254
18.4.2 Support functions . . . 2257
18.4.3 Polar sets . . . 2259
18.4.4 Dual norms . . . 2260
18.4.5 The dual of the dual . . . 2262
18.4.6 Operator norms as maxima of bilinear values . . . 2264
18.4.7 Exercises . . . 2265
18.4.7.1 A. Check your understanding . . . 2265
18.4.7.2 B. Practice . . . 2265
18.4.7.3 C. Going deeper . . . 2267

18.5 Cones and Farkas's Lemma 2268
18.5.1 Convex cones . . . 2269
18.5.2 The dual cone . . . 2271
18.5.3 Finitely generated cones are closed . . . 2273
18.5.4 Farkas's lemma . . . 2275
18.5.5 The dual of the dual . . . 2278
18.5.6 When the cone is not finitely generated . . . 2278
18.5.7 Exercises . . . 2280
18.5.7.1 A. Check your understanding . . . 2280
18.5.7.2 B. Practice . . . 2280
18.5.7.3 C. Going deeper . . . 2282

18.6 Linear Programming Duality **2283**
18.6.1 Linear programs . . . 2284
18.6.2 The dual program . . . 2286
18.6.3 Weak duality . . . 2287
18.6.4 Strong duality . . . 2288
18.6.5 Complementary slackness . . . 2290
18.6.6 Traffic, with inequalities . . . 2291
18.6.7 Exercises . . . 2292
18.6.7.1 A. Check your understanding . . . 2292
18.6.7.2 B. Practice . . . 2293
18.6.7.3 C. Going deeper . . . 2294

18.7 Matrix Games and the Minimax Theorem **2296**
18.7.1 Matrix games . . . 2296
18.7.2 Mixed strategies . . . 2297
18.7.3 The minimax theorem . . . 2299
18.7.4 Two min–max theorems . . . 2300
18.7.5 Solving a game . . . 2301
18.7.6 Exercises . . . 2303
18.7.6.1 A. Check your understanding . . . 2303
18.7.6.2 B. Practice . . . 2304
18.7.6.3 C. Going deeper . . . 2305

18.8 Extreme Points **2306**
18.8.1 Points that are not in between . . . 2307
18.8.2 Faces . . . 2309
18.8.3 Minkowski's theorem . . . 2310
18.8.4 Three unit balls . . . 2313
18.8.5 Density matrices . . . 2314
18.8.6 Doubly stochastic matrices . . . 2317
18.8.7 When the extreme points are not closed . . . 2318
18.8.8 Exercises . . . 2319
18.8.8.1 A. Check your understanding . . . 2319
18.8.8.2 B. Practice . . . 2320
18.8.8.3 C. Going deeper . . . 2322

18.9 Polytopes **2324**
18.9.1 Two finite descriptions . . . 2324
18.9.2 Vertices are basic feasible points . . . 2326
18.9.3 Fourier–Motzkin elimination . . . 2328
18.9.4 The Minkowski–Weyl theorem . . . 2331
18.9.5 The cube and the cross-polytope . . . 2333
18.9.6 Exercises . . . 2333
18.9.6.1 A. Check your understanding . . . 2333
18.9.6.2 B. Practice . . . 2334
18.9.6.3 C. Going deeper . . . 2336

18.10 Convex Functions **2338**
18.10.1 Convex functions and their epigraphs . . . 2338
18.10.2 Jensen's inequality . . . 2341
18.10.3 Suprema of affine functions . . . 2342
18.10.4 Convex functions of a Hermitian matrix . . . 2343

18.10.5 A second-derivative test . . . 2344
18.10.6 Continuity . . . 2347
18.10.7 Exercises . . . 2349
18.10.7.1 A. Check your understanding . . . 2349
18.10.7.2 B. Practice . . . 2350
18.10.7.3 C. Going deeper . . . 2351

18.11 Three Classical Inequalities 2352
18.11.1 The logarithm and real powers . . . 2353
18.11.2 Arithmetic and geometric means . . . 2354
18.11.3 Hölder's inequality and the dual p-norms . . . 2355
18.11.4 Summary and transfer . . . 2358
18.11.5 Exercises . . . 2359
18.11.5.1 A. Check your understanding . . . 2359
18.11.5.2 B. Practice . . . 2360
18.11.5.3 C. Going deeper . . . 2361

19 Chapter 19 2363

Nonnegative Matrices 2363
19.0.1 The analysis this chapter imports . . . 2364
19.0.2 What you need . . . 2364
19.0.3 Roadmap . . . 2364
19.0.4 Named moves . . . 2365

19.1 Positive Matrices and Perron's Theorem 2365
19.1.1 The entrywise order . . . 2366
19.1.2 The spectral radius is monotone . . . 2369
19.1.3 Perron's theorem . . . 2370
19.1.4 What positivity buys, and what it costs . . . 2376
19.1.5 Exercises . . . 2376
19.1.5.1 A. Check your understanding . . . 2376
19.1.5.2 B. Practice . . . 2377
19.1.5.3 C. Going deeper . . . 2378

19.2 Irreducibility and Graphs 2380
19.2.1 The graph of a matrix . . . 2381
19.2.2 Strong connectivity . . . 2383
19.2.3 Irreducible matrices . . . 2384
19.2.4 A test by one matrix power . . . 2388
19.2.5 The reducible normal form . . . 2389
19.2.6 Exercises . . . 2391
19.2.6.1 A. Check your understanding . . . 2391
19.2.6.2 B. Practice . . . 2391
19.2.6.3 C. Going deeper . . . 2393

19.3 The Perron–Frobenius Theorem 2395
19.3.1 What survives for every nonnegative matrix . . . 2395
19.3.2 The theorem for irreducible matrices . . . 2397
19.3.3 The left Perron vector . . . 2400
19.3.4 Exercises . . . 2402
19.3.4.1 A. Check your understanding . . . 2402

19.3.4.2 B. Practice . . . 2403
19.3.4.3 C. Going deeper . . . 2405

19.4 The Collatz–Wielandt Formula 2407
19.4.1 Bounds from a single test vector . . . 2407
19.4.2 The two Collatz–Wielandt functions . . . 2410
19.4.3 The formula . . . 2411
19.4.4 Row sums and column sums . . . 2412
19.4.5 Estimating the spectral radius . . . 2413
19.4.6 Exercises . . . 2414
19.4.6.1 A. Check your understanding . . . 2414
19.4.6.2 B. Practice . . . 2415
19.4.6.3 C. Going deeper . . . 2417

19.5 Primitive Matrices 2418
19.5.1 Walks in a nonnegative matrix . . . 2418
19.5.2 Primitive matrices . . . 2419
19.5.3 The limit theorem . . . 2421
19.5.4 The peripheral spectrum and the period . . . 2424
19.5.5 The cyclic permutation . . . 2426
19.5.6 How large a power is needed . . . 2427
19.5.7 Exercises . . . 2428
19.5.7.1 A. Check your understanding . . . 2428
19.5.7.2 B. Practice . . . 2428
19.5.7.3 C. Going deeper . . . 2431

19.6 Markov Chains 2432
19.6.1 Steady states and the Perron root . . . 2433
19.6.2 The long-run theorem . . . 2434
19.6.3 Irreducible chains: convergence on average . . . 2436
19.6.4 Reducible chains: many steady states . . . 2438
19.6.5 Absorbing chains . . . 2439
19.6.6 Exercises . . . 2442
19.6.6.1 A. Check your understanding . . . 2442
19.6.6.2 B. Practice . . . 2443
19.6.6.3 C. Going deeper . . . 2445

19.7 Birkhoff's Theorem 2447
19.7.1 What is left to prove . . . 2447
19.7.2 The doubly stochastic matrices form a polyhedron . . . 2447
19.7.3 Birkhoff's theorem . . . 2449
19.7.4 How many permutation matrices . . . 2451
19.7.5 Finding the permutations . . . 2452
19.7.6 Schur's theorem again . . . 2454
19.7.7 Exercises . . . 2456
19.7.7.1 A. Check your understanding . . . 2456
19.7.7.2 B. Practice . . . 2456
19.7.7.3 C. Going deeper . . . 2458

19.8 PageRank 2461
19.8.1 The web as a Markov chain . . . 2461
19.8.2 The Google matrix . . . 2462

19.8.3 A unique positive ranking . . . 2463
19.8.4 The second eigenvalue . . . 2465
19.8.5 How fast the iteration converges . . . 2466
19.8.6 A small web, exactly . . . 2467
19.8.7 Exercises . . . 2469
19.8.7.1 A. Check your understanding . . . 2469
19.8.7.2 B. Practice . . . 2469
19.8.7.3 C. Going deeper . . . 2471

19.9 M-Matrices 2472
19.9.1 Z-matrices and M-matrices . . . 2472
19.9.2 The characterization theorem . . . 2475
19.9.3 The second-difference matrix . . . 2479
19.9.4 The Leontief input–output model . . . 2481
19.9.5 Summary and transfer . . . 2483
19.9.6 Exercises . . . 2484
19.9.6.1 A. Check your understanding . . . 2484
19.9.6.2 B. Practice . . . 2485
19.9.6.3 C. Going deeper . . . 2487

20 Chapter 20 2490

Perturbation Theory 2490
20.0.1 The analysis this chapter imports . . . 2491
20.0.2 What you need . . . 2491
20.0.3 Roadmap . . . 2492
20.0.4 Named moves . . . 2493

20.1 Gershgorin Discs 2493
20.1.1 Discs around the diagonal . . . 2493
20.1.2 The theorem . . . 2494
20.1.3 Diagonal dominance . . . 2497
20.1.4 Hermitian matrices, and intervals . . . 2497
20.1.5 Scaling the discs . . . 2497
20.1.6 Eigenvalues on the boundary . . . 2499
20.1.7 Exercises . . . 2502
20.1.7.1 A. Check your understanding . . . 2502
20.1.7.2 B. Practice . . . 2502
20.1.7.3 C. Going deeper . . . 2504

20.2 Other Inclusion Regions 2505
20.2.1 Using two coordinates: Brauer's ovals . . . 2506
20.2.2 The numerical range . . . 2508
20.2.3 Convexity: the Toeplitz–Hausdorff theorem . . . 2510
20.2.4 Normal matrices . . . 2512
20.2.5 Bendixson's rectangle . . . 2513
20.2.6 What these regions do not do . . . 2514
20.2.7 Exercises . . . 2515
20.2.7.1 A. Check your understanding . . . 2515
20.2.7.2 B. Practice . . . 2515
20.2.7.3 C. Going deeper . . . 2516

20.3 Continuity of Roots **2518**
20.3.1 A bound for the roots . . . 2518
20.3.2 The roots move continuously . . . 2519
20.3.3 How far: Elsner's bound . . . 2521
20.3.4 Counting eigenvalues in a region . . . 2523
20.3.5 From one-sided to matched . . . 2526
20.3.6 Exercises . . . 2527
20.3.6.1 A. Check your understanding . . . 2527
20.3.6.2 B. Practice . . . 2528
20.3.6.3 C. Going deeper . . . 2529

20.4 The Bauer–Fike Theorem **2530**
20.4.1 Factoring out the inverse . . . 2531
20.4.2 The theorem . . . 2532
20.4.3 Normal matrices . . . 2533
20.4.4 Residuals . . . 2534
20.4.5 A worked example . . . 2536
20.4.6 Defective matrices . . . 2537
20.4.7 Exercises . . . 2538
20.4.7.1 A. Check your understanding . . . 2538
20.4.7.2 B. Practice . . . 2539
20.4.7.3 C. Going deeper . . . 2540

20.5 The Hoffman–Wielandt Theorem **2542**
20.5.1 One-sided bounds do not pair eigenvalues . . . 2542
20.5.2 Birkhoff turns weights into a permutation . . . 2543
20.5.3 The theorem . . . 2544
20.5.4 Normality cannot be dropped . . . 2546
20.5.5 Hermitian matrices: the pairing is the order . . . 2547
20.5.6 Exercises . . . 2550
20.5.6.1 A. Check your understanding . . . 2550
20.5.6.2 B. Practice . . . 2550
20.5.6.3 C. Going deeper . . . 2552

20.6 Hermitian Perturbations **2554**
20.6.1 What Weyl's bound leaves open . . . 2554
20.6.2 Residual bounds . . . 2554
20.6.3 The Kato–Temple bound . . . 2555
20.6.4 Relative perturbations and congruence . . . 2558
20.6.5 Exercises . . . 2561
20.6.5.1 A. Check your understanding . . . 2561
20.6.5.2 B. Practice . . . 2561
20.6.5.3 C. Going deeper . . . 2563

20.7 The Sylvester Equation Revisited **2564**
20.7.1 How large can the solution be? . . . 2565
20.7.2 Normal pairs: the separation is the gap . . . 2568
20.7.3 Nonnormal pairs: the separation can collapse . . . 2569
20.7.4 A one-sided bound in the spectral norm . . . 2570
20.7.5 Exercises . . . 2573
20.7.5.1 A. Check your understanding . . . 2573
20.7.5.2 B. Practice . . . 2574

20.7.5.3 C. Going deeper . . . 2575

20.8 Principal Angles 2577
20.8.1 One angle is not enough . . . 2578
20.8.2 Principal angles . . . 2578
20.8.3 Sines and the orthogonal complement . . . 2582
20.8.4 The distance between two subspaces . . . 2584
20.8.5 Exercises . . . 2587
20.8.5.1 A. Check your understanding . . . 2587
20.8.5.2 B. Practice . . . 2588
20.8.5.3 C. Going deeper . . . 2589

20.9 The Davis–Kahan Theorem 2591
20.9.1 Eigenvectors can move a long way . . . 2591
20.9.2 Two norms, three facts . . . 2592
20.9.3 The perturbation of a subspace is a Sylvester equation . . . 2593
20.9.4 The $\sin\Theta$ theorem in the Frobenius norm . . . 2595
20.9.5 A one-sided version in the spectral norm . . . 2596
20.9.6 Single eigenvectors . . . 2597
20.9.7 Clusters: the subspace is stable when the vectors are not . . . 2599
20.9.8 Exercises . . . 2600
20.9.8.1 A. Check your understanding . . . 2600
20.9.8.2 B. Practice . . . 2601
20.9.8.3 C. Going deeper . . . 2603

20.10 Singular Values and Polar Factors 2604
20.10.1 Singular values in the Frobenius norm . . . 2605
20.10.2 The square root is Lipschitz away from zero . . . 2607
20.10.3 The positive polar factor . . . 2609
20.10.4 The unitary polar factor . . . 2610
20.10.5 Exercises . . . 2612
20.10.5.1 A. Check your understanding . . . 2612
20.10.5.2 B. Practice . . . 2613
20.10.5.3 C. Going deeper . . . 2615

20.11 Condition Numbers of Eigenvalues 2617
20.11.1 Left and right eigenvectors . . . 2617
20.11.2 The derivative of a simple eigenvalue . . . 2619
20.11.3 The condition number of a simple eigenvalue . . . 2622
20.11.4 A worked example . . . 2625
20.11.5 When eigenvalues coalesce . . . 2626
20.11.6 Summary and transfer . . . 2627
20.11.7 Exercises . . . 2629
20.11.7.1 A. Check your understanding . . . 2629
20.11.7.2 B. Practice . . . 2630
20.11.7.3 C. Going deeper . . . 2632

21 Chapter 21 2635

Matrix Inequalities 2635
21.0.1 The analysis this chapter imports . . . 2636
21.0.2 What you need . . . 2637

21.0.3 Roadmap . . . 2638
21.0.4 Named moves . . . 2639

21.1 Doubly Stochastic Maps 2639
21.1.1 What is still missing . . . 2639
21.1.2 Averaging two coordinates . . . 2640
21.1.3 One averaging step . . . 2642
21.1.4 The theorem . . . 2644
21.1.5 Majorization as a polytope . . . 2647
21.1.6 Exercises . . . 2648
21.1.6.1 A. Check your understanding . . . 2648
21.1.6.2 B. Practice . . . 2648
21.1.6.3 C. Going deeper . . . 2650

21.2 Convex Functions of a Majorization 2652
21.2.1 Karamata's inequality . . . 2652
21.2.2 The converse, and the weak version . . . 2654
21.2.3 Schur-convex functions . . . 2655
21.2.4 Convex functions of a Hermitian matrix . . . 2659
21.2.5 Exercises . . . 2662
21.2.5.1 A. Check your understanding . . . 2662
21.2.5.2 B. Practice . . . 2663
21.2.5.3 C. Going deeper . . . 2665

21.3 Symmetric Gauge Functions 2667
21.3.1 What a norm must forget . . . 2667
21.3.2 A symmetric gauge is monotone . . . 2670
21.3.3 From weak majorization to a dominating vector . . . 2671
21.3.4 The inequality . . . 2674
21.3.5 The dual gauge . . . 2675
21.3.6 Exercises . . . 2676
21.3.6.1 A. Check your understanding . . . 2676
21.3.6.2 B. Practice . . . 2676
21.3.6.3 C. Going deeper . . . 2678

21.4 Unitarily Invariant Norms 2680
21.4.1 Norms that do not see the unitary factors . . . 2681
21.4.2 A unitarily invariant norm is a function of the singular values . . . 2683
21.4.3 Singular values are subadditive . . . 2684
21.4.4 The correspondence . . . 2685
21.4.5 What every unitarily invariant norm satisfies . . . 2687
21.4.6 Exercises . . . 2689
21.4.6.1 A. Check your understanding . . . 2689
21.4.6.2 B. Practice . . . 2689
21.4.6.3 C. Going deeper . . . 2691

21.5 The Trace Inequality 2693
21.5.1 How large can a pairing be? . . . 2693
21.5.2 Weights that do not add up to more than one . . . 2693
21.5.3 The trace inequality . . . 2695
21.5.4 The dual of a unitarily invariant norm . . . 2697
21.5.5 Exercises . . . 2699

21.5.5.1 A. Check your understanding . . . 2699
21.5.5.2 B. Practice . . . 2700
21.5.5.3 C. Going deeper . . . 2701

21.6 Ky Fan Dominance 2702
21.6.1 One weak majorization, every norm . . . 2702
21.6.2 Low-rank approximation in every unitarily invariant norm . . . 2703
21.6.3 The nearest unitary matrix . . . 2706
21.6.4 Two smaller consequences . . . 2708
21.6.5 Exercises . . . 2709
21.6.5.1 A. Check your understanding . . . 2709
21.6.5.2 B. Practice . . . 2710
21.6.5.3 C. Going deeper . . . 2711

21.7 Lidskii's Theorem in Majorization Form 2712
21.7.1 What Chapter 17 proved, and what it left . . . 2712
21.7.2 Every convex function at once . . . 2713
21.7.3 Every unitarily invariant norm at once . . . 2714
21.7.4 The same theorem for singular values . . . 2716
21.7.5 Exercises . . . 2718
21.7.5.1 A. Check your understanding . . . 2718
21.7.5.2 B. Practice . . . 2718
21.7.5.3 C. Going deeper . . . 2720

21.8 Operator Convex Functions 2722
21.8.1 What operator convexity asks . . . 2722
21.8.2 The examples we can prove . . . 2723
21.8.3 Convex is not operator convex . . . 2725
21.8.4 Contractions and their dilations . . . 2727
21.8.5 Jensen's inequality for operators . . . 2728
21.8.6 Where this sits . . . 2731
21.8.7 Exercises . . . 2731
21.8.7.1 A. Check your understanding . . . 2731
21.8.7.2 B. Practice . . . 2732
21.8.7.3 C. Going deeper . . . 2733

21.9 Loewner Matrices 2735
21.9.1 Slopes between two points . . . 2735
21.9.2 The Loewner matrix . . . 2737
21.9.3 Differentiating a matrix function . . . 2739
21.9.4 The criterion . . . 2743
21.9.5 What this gives, and what it does not . . . 2747
21.9.6 Exercises . . . 2747
21.9.6.1 A. Check your understanding . . . 2747
21.9.6.2 B. Practice . . . 2748
21.9.6.3 C. Going deeper . . . 2749

21.10 Loewner's Theorem 2752
21.10.1 Integrating a matrix-valued function . . . 2752
21.10.2 The one fact this chapter quotes . . . 2755
21.10.3 Loewner's theorem . . . 2756
21.10.4 The powers . . . 2758

21.10.5 The logarithm and the exponential . . . 2763
21.10.6 Exercises . . . 2766
21.10.6.1 A. Check your understanding . . . 2766
21.10.6.2 B. Practice . . . 2767
21.10.6.3 C. Going deeper . . . 2769

21.11 Means and Matrix Cauchy–Schwarz 2771
21.11.1 An off-diagonal block is small . . . 2771
21.11.2 A product against a sum of squares . . . 2772
21.11.3 Cauchy–Schwarz . . . 2773
21.11.4 The geometric mean of two positive definite matrices . . . 2774
21.11.5 What the geometric mean does . . . 2776
21.11.6 The chain of three means . . . 2779
21.11.7 Exercises . . . 2780
21.11.7.1 A. Check your understanding . . . 2780
21.11.7.2 B. Practice . . . 2780
21.11.7.3 C. Going deeper . . . 2782

21.12 Golden–Thompson 2784
21.12.1 Sums in the exponent, products in the answer . . . 2784
21.12.2 The trace of a product of two positive matrices . . . 2786
21.12.3 The inequality of Golden and Thompson . . . 2788
21.12.4 Three matrices are one too many . . . 2790
21.12.5 Klein's inequality and relative entropy . . . 2791
21.12.6 What is stated here and not proved . . . 2794
21.12.7 Summary and transfer . . . 2795
21.12.8 Exercises . . . 2797
21.12.8.1 A. Check your understanding . . . 2797
21.12.8.2 B. Practice . . . 2798
21.12.8.3 C. Going deeper . . . 2800

22 Chapter 22 2802

Affine and Projective Geometry 2802
22.0.1 Conventions . . . 2802
22.0.1.1 Which field, and where the characteristic bites . . . 2803
22.0.2 What the chapter quotes, and what it does not prove . . . 2803
22.0.3 What you need . . . 2804
22.0.4 Roadmap . . . 2806
22.0.5 Named moves . . . 2808

22.1 Affine Spaces and Affine Maps 2808
22.1.1 A set with no zero in it . . . 2809
22.1.2 Choosing an origin . . . 2812
22.1.3 Affine combinations . . . 2813
22.1.4 Affine maps . . . 2814
22.1.5 The affine group, and the matrix that computes it . . . 2817
22.1.6 Exercises . . . 2819
22.1.6.1 A. Check your understanding . . . 2819
22.1.6.2 B. Practice . . . 2820
22.1.6.3 C. Going deeper . . . 2822

22.2 Flats, Independence, and Coordinates 2823
22.2.1 Flats . . . 2824
22.2.2 When two flats meet, and how big their join is . . . 2826
22.2.3 Affine frames and barycentric coordinates . . . 2829
22.2.4 An affine map is fixed by a frame . . . 2833
22.2.5 A flat is a solution set . . . 2834
22.2.6 Exercises . . . 2836
22.2.6.1 A. Check your understanding . . . 2836
22.2.6.2 B. Practice . . . 2836
22.2.6.3 C. Going deeper . . . 2838

22.3 Convex Polyhedra, Geometrically 2840
22.3.1 Faces, cut by a supporting hyperplane . . . 2840
22.3.2 Vertices, edges, facets . . . 2843
22.3.3 Two face lattices, computed . . . 2845
22.3.4 A linear program, seen . . . 2847
22.3.5 What the dual variables are, geometrically . . . 2849
22.3.6 Polyhedra without corners . . . 2850
22.3.7 Exercises . . . 2851
22.3.7.1 A. Check your understanding . . . 2851
22.3.7.2 B. Practice . . . 2852
22.3.7.3 C. Going deeper . . . 2853

22.4 Euclidean Motions 2856
22.4.1 Distance alone . . . 2856
22.4.2 The group of motions . . . 2859
22.4.3 Motions of the plane . . . 2861
22.4.4 Motions of space . . . 2865
22.4.5 Exercises . . . 2868
22.4.5.1 A. Check your understanding . . . 2868
22.4.5.2 B. Practice . . . 2869
22.4.5.3 C. Going deeper . . . 2870

22.5 Affine and Euclidean Quadrics 2873
22.5.1 Where characteristic two bites . . . 2873
22.5.2 The equation of a quadric . . . 2873
22.5.3 The matrix of a quadric . . . 2875
22.5.4 Centers . . . 2876
22.5.5 The affine classification over the reals . . . 2878
22.5.5.1 The nine real conics . . . 2880
22.5.6 Euclidean quadrics and principal axes . . . 2882
22.5.7 Exercises . . . 2885
22.5.7.1 A. Check your understanding . . . 2885
22.5.7.2 B. Practice . . . 2886
22.5.7.3 C. Going deeper . . . 2887

22.6 Projective Space 2890
22.6.1 Two lines that do not meet . . . 2890
22.6.2 The projective space of a vector space . . . 2890
22.6.3 The affine chart, and where the missing points live . . . 2892
22.6.4 Projective subspaces, and a dimension formula with one case . . . 2895
22.6.5 Counting points . . . 2898

22.6.6 Exercises . . . 2900
22.6.6.1 A. Check your understanding . . . 2900
22.6.6.2 B. Practice . . . 2900
22.6.6.3 C. Going deeper . . . 2902

22.7 Projective Transformations 2904
22.7.1 Maps induced by linear isomorphisms . . . 2904
22.7.2 The projective linear group . . . 2906
22.7.3 Frames, and how many points pin a projectivity down . . . 2907
22.7.4 The cross-ratio . . . 2912
22.7.5 Perspectivities . . . 2916
22.7.6 Affine maps inside the projective group . . . 2917
22.7.7 Exercises . . . 2919
22.7.7.1 A. Check your understanding . . . 2919
22.7.7.2 B. Practice . . . 2919
22.7.7.3 C. Going deeper . . . 2921

22.8 Duality 2923
22.8.1 Projective subspaces remember their subspaces . . . 2923
22.8.2 The points of the dual space are the hyperplanes . . . 2924
22.8.3 The duality correspondence . . . 2926
22.8.4 Dualizing twice . . . 2928
22.8.5 The duality principle . . . 2929
22.8.6 Exercises . . . 2931
22.8.6.1 A. Check your understanding . . . 2931
22.8.6.2 B. Practice . . . 2932
22.8.6.3 C. Going deeper . . . 2933

22.9 Desargues and Pappus 2935
22.9.1 Perspective from a point, perspective from a line . . . 2935
22.9.2 Choosing the representatives . . . 2936
22.9.3 Desargues's theorem . . . 2937
22.9.4 Pappus's theorem . . . 2941
22.9.5 What the two theorems measure . . . 2944
22.9.6 Exercises . . . 2944
22.9.6.1 A. Check your understanding . . . 2944
22.9.6.2 B. Practice . . . 2945
22.9.6.3 C. Going deeper . . . 2946

22.10 Conics and Lines in Space 2949
22.10.1 Quadrics in projective space . . . 2949
22.10.2 Five points determine a conic . . . 2952
22.10.3 Projecting a conic onto a line . . . 2955
22.10.4 Pole, polar, and the dual conic . . . 2957
22.10.5 Lines in space: Plücker coordinates . . . 2958
22.10.6 Summary and transfer . . . 2962
22.10.7 Exercises . . . 2963
22.10.7.1 A. Check your understanding . . . 2963
22.10.7.2 B. Practice . . . 2964
22.10.7.3 C. Going deeper . . . 2965

23 Chapter 23 **2968**

Algebras and Representations **2968**
23.0.1 Conventions . . . 2969
23.0.1.1 Which field, and where the hypothesis bites . . . 2969
23.0.2 What the chapter quotes, and what it does not prove . . . 2970
23.0.3 What you need . . . 2971
23.0.4 Roadmap . . . 2973
23.0.5 Named moves . . . 2976

23.1 Associative Algebras **2976**
23.1.1 What an algebra is, and what we assume about it . . . 2977
23.1.2 Subalgebras . . . 2977
23.1.3 Homomorphisms, isomorphisms, and the algebra generated by a set . . . 2979
23.1.4 The group algebra . . . 2982
23.1.5 Every finite-dimensional algebra is an algebra of matrices . . . 2983
23.1.6 Commutants . . . 2985
23.1.7 Exercises . . . 2987
23.1.7.1 A. Check your understanding . . . 2987
23.1.7.2 B. Practice . . . 2988
23.1.7.3 C. Going deeper . . . 2990

23.2 Ideals and Quotients **2991**
23.2.1 Three kinds of ideal . . . 2992
23.2.2 Matrix units multiply in one line . . . 2993
23.2.3 The matrix algebra is simple . . . 2994
23.2.4 The left ideals of the matrix algebra . . . 2995
23.2.5 Quotients, and the first isomorphism theorem . . . 2997
23.2.6 Exercises . . . 2999
23.2.6.1 A. Check your understanding . . . 2999
23.2.6.2 B. Practice . . . 3000
23.2.6.3 C. Going deeper . . . 3001

23.3 Simple and Semisimple Algebras **3003**
23.3.1 Two meanings of the word simple . . . 3004
23.3.2 Algebras acting on spaces . . . 3005
23.3.3 Splitting a space into pieces . . . 3006
23.3.4 How far from semisimple: the radical . . . 3011
23.3.5 Back to a single operator . . . 3012
23.3.6 Exercises . . . 3014
23.3.6.1 A. Check your understanding . . . 3014
23.3.6.2 B. Practice . . . 3015
23.3.6.3 C. Going deeper . . . 3017

23.4 Schur's Lemma **3018**
23.4.1 Maps that respect the action . . . 3018
23.4.2 The lemma . . . 3019
23.4.3 What goes wrong over the reals . . . 3020
23.4.4 Families of operators . . . 3021
23.4.5 Looking back at commuting operators . . . 3023
23.4.6 Exercises . . . 3025
23.4.6.1 A. Check your understanding . . . 3025

23.4.6.2 B. Practice . . . 3026
23.4.6.3 C. Going deeper . . . 3027

23.5 Burnside and the Double Commutant 3029
23.5.1 Commutants, and when an operator is a scalar . . . 3029
23.5.2 Algebras that act irreducibly . . . 3030
23.5.3 Burnside's theorem . . . 3030
23.5.4 The double commutant theorem . . . 3033
23.5.5 The commutant of one operator, revisited . . . 3036
23.5.6 Exercises . . . 3037
23.5.6.1 A. Check your understanding . . . 3037
23.5.6.2 B. Practice . . . 3038
23.5.6.3 C. Going deeper . . . 3039

23.6 The Structure Theorem 3041
23.6.1 Products of matrix algebras . . . 3041
23.6.2 The regular representation . . . 3042
23.6.3 The structure theorem . . . 3042
23.6.4 What is quoted and not proved . . . 3046
23.6.5 Two worked decompositions . . . 3047
23.6.6 Exercises . . . 3048
23.6.6.1 A. Check your understanding . . . 3048
23.6.6.2 B. Practice . . . 3049
23.6.6.3 C. Going deeper . . . 3050

23.7 Group Representations 3052
23.7.1 Letting a group act by operators . . . 3052
23.7.2 Subrepresentations, equivalence, and irreducibility . . . 3054
23.7.3 Representations of G and representations of $F[G]$ . . . 3055
23.7.4 Maschke's theorem . . . 3057
23.7.5 Two families worked out . . . 3060
23.7.6 Exercises . . . 3061
23.7.6.1 A. Check your understanding . . . 3061
23.7.6.2 B. Practice . . . 3062
23.7.6.3 C. Going deeper . . . 3064

23.8 Characters and Orthogonality 3066
23.8.1 Conjugacy classes . . . 3066
23.8.2 The character of a representation . . . 3067
23.8.3 Class functions and their inner product . . . 3069
23.8.4 Two averaging lemmas . . . 3071
23.8.5 Orthogonality of the irreducible characters . . . 3073
23.8.6 The characters span the class functions . . . 3075
23.8.7 Two character tables . . . 3078
23.8.8 Exercises . . . 3080
23.8.8.1 A. Check your understanding . . . 3080
23.8.8.2 B. Practice . . . 3081
23.8.8.3 C. Going deeper . . . 3083

23.9 Finite Abelian Groups and the Fourier Transform 3084
23.9.1 One dimension is all there is . . . 3085
23.9.2 The dual group . . . 3088

23.9.3 The character matrix, the transform, and inversion . . . 3091
23.9.4 The cyclic group is Chapter 12 . . . 3093
23.9.5 Exercises . . . 3096
23.9.5.1 A. Check your understanding . . . 3096
23.9.5.2 B. Practice . . . 3097
23.9.5.3 C. Going deeper . . . 3099

23.10 What the Chapter Buys **3102**
23.10.1 Four moves . . . 3102
23.10.2 Four old theorems, seen from here . . . 3102
23.10.3 Which hypothesis did what . . . 3103
23.10.4 Quoted, not proved . . . 3105
23.10.5 Left outside on purpose . . . 3105
23.10.6 One more look back . . . 3106
23.10.7 Exercises . . . 3106
23.10.7.1 A. Check your understanding . . . 3106
23.10.7.2 B. Practice . . . 3107
23.10.7.3 C. Going deeper . . . 3108

24 Chapter 24 **3109**

Computation and Applications **3109**
24.0.1 Conventions . . . 3109
24.0.2 What the chapter assumes . . . 3110
24.0.3 What is quoted and not proved . . . 3110
24.0.4 What you need . . . 3111
24.0.5 Roadmap . . . 3113
24.0.6 Named moves . . . 3114

24.1 Floating Point and Backward Error **3115**
24.1.1 Two answers to one question . . . 3115
24.1.2 A model of machine arithmetic . . . 3115
24.1.3 Accumulated rounding: the constant γ_k . . . 3117
24.1.4 The first backward error analysis . . . 3118
24.1.5 Forward error, backward error, stability . . . 3120
24.1.6 Multiplying the two together . . . 3121
24.1.7 Cancellation, and where the damage really happens . . . 3123
24.1.8 Exercises . . . 3124
24.1.8.1 A. Check your understanding . . . 3124
24.1.8.2 B. Practice . . . 3125
24.1.8.3 C. Going deeper . . . 3127

24.2 Elimination, Pivoting, and Cost **3128**
24.2.1 Elimination is a sequence of Schur complements . . . 3129
24.2.2 What elimination costs . . . 3132
24.2.3 One row at a time . . . 3133
24.2.4 Solving a triangular system . . . 3134
24.2.5 The factorization itself . . . 3135
24.2.6 Pivoting . . . 3137
24.2.7 The example Chapter 3 promised . . . 3140
24.2.8 Cholesky needs no pivoting . . . 3141

24.2.9 Exercises . . . 3143
24.2.9.1 A. Check your understanding . . . 3143
24.2.9.2 B. Practice . . . 3144
24.2.9.3 C. Going deeper . . . 3146

24.3 Least Squares in Practice 3148
24.3.1 How sensitive is the problem? . . . 3148
24.3.2 Two routes, and the digits they lose . . . 3151
24.3.3 Why the reflection route is the stable one . . . 3154
24.3.4 What goes wrong with Gram–Schmidt . . . 3156
24.3.5 The sign in the reflection, quantified . . . 3157
24.3.6 Exercises . . . 3159
24.3.6.1 A. Check your understanding . . . 3159
24.3.6.2 B. Practice . . . 3160
24.3.6.3 C. Going deeper . . . 3162

24.4 The Power Method 3164
24.4.1 Multiplying by the matrix . . . 3164
24.4.2 Shifting and inverting . . . 3170
24.4.3 The shift that improves itself . . . 3173
24.4.4 Deflation, and what comes next . . . 3176
24.4.5 Exercises . . . 3176
24.4.5.1 A. Check your understanding . . . 3176
24.4.5.2 B. Practice . . . 3177
24.4.5.3 C. Going deeper . . . 3179

24.5 The QR Algorithm 3181
24.5.1 Factor and multiply back . . . 3181
24.5.2 The iteration is the power method in disguise . . . 3183
24.5.3 What converges, and how fast . . . 3185
24.5.4 Hessenberg first . . . 3187
24.5.5 Shifts . . . 3189
24.5.6 Three unshifted steps, watched . . . 3192
24.5.7 Exercises . . . 3193
24.5.7.1 A. Check your understanding . . . 3193
24.5.7.2 B. Practice . . . 3194
24.5.7.3 C. Going deeper . . . 3196

24.6 Krylov Subspaces 3198
24.6.1 The space that k products span . . . 3198
24.6.2 Arnoldi: Gram–Schmidt with a purpose . . . 3200
24.6.3 Lanczos: the recurrence collapses to three terms . . . 3202
24.6.4 Ritz values . . . 3204
24.6.5 How fast the extreme Ritz values converge . . . 3207
24.6.6 Losing orthogonality . . . 3209
24.6.7 Exercises . . . 3210
24.6.7.1 A. Check your understanding . . . 3210
24.6.7.2 B. Practice . . . 3210
24.6.7.3 C. Going deeper . . . 3212

24.7 Steepest Descent and Conjugate Gradients 3215
24.7.1 A linear system is a minimization problem . . . 3215

24.7.2 Steepest descent . . . 3216
24.7.3 Conjugate directions . . . 3220
24.7.4 Chebyshev polynomials . . . 3224
24.7.5 The square root . . . 3226
24.7.6 Two footnotes: Chebyshev iteration and preconditioning . . . 3228
24.7.7 Exercises . . . 3228
24.7.7.1 A. Check your understanding . . . 3228
24.7.7.2 B. Practice . . . 3229
24.7.7.3 C. Going deeper . . . 3231

24.8 Lax Pairs and the Toda Flow 3232
24.8.1 Which matrix differential equations preserve the spectrum? . . . 3233
24.8.2 The isospectral theorem . . . 3234
24.8.3 The Toda flow . . . 3236
24.8.4 The QR algorithm is the Toda flow at integer times . . . 3239
24.8.5 What this buys . . . 3242
24.8.6 Exercises . . . 3242
24.8.6.1 A. Check your understanding . . . 3242
24.8.6.2 B. Practice . . . 3243
24.8.6.3 C. Going deeper . . . 3245

24.9 Vibrations 3246
24.9.1 A chain of masses and springs . . . 3247
24.9.2 Normal modes . . . 3249
24.9.3 Frequencies as a minimum . . . 3253
24.9.4 Stiffer, heavier, pinned . . . 3255
24.9.5 Exercises . . . 3257
24.9.5.1 A. Check your understanding . . . 3257
24.9.5.2 B. Practice . . . 3258
24.9.5.3 C. Going deeper . . . 3260

24.10 Principal Components 3261
24.10.1 Centering . . . 3261
24.10.2 The best-fitting affine subspace . . . 3262
24.10.3 The sample covariance matrix . . . 3265
24.10.4 How much variance a subspace explains . . . 3267
24.10.5 Computing it, and what is not proved . . . 3269
24.10.6 Exercises . . . 3271
24.10.6.1 A. Check your understanding . . . 3271
24.10.6.2 B. Practice . . . 3272
24.10.6.3 C. Going deeper . . . 3273

24.11 Graph Laplacians 3275
24.11.1 The Laplacian of a graph . . . 3275
24.11.2 The kernel counts the components . . . 3277
24.11.3 The Fiedler value . . . 3279
24.11.4 Counting spanning trees . . . 3282
24.11.5 Exercises . . . 3287
24.11.5.1 A. Check your understanding . . . 3287
24.11.5.2 B. Practice . . . 3288
24.11.5.3 C. Going deeper . . . 3290

24.12 The Fast Fourier Transform **3292**
24.12.1 Counting the cost of a transform . . . 3292
24.12.2 Splitting a transform in two . . . 3293
24.12.3 The cost, exactly . . . 3296
24.12.4 What the speed buys . . . 3298
24.12.5 Exercises . . . 3300
24.12.5.1 A. Check your understanding . . . 3300
24.12.5.2 B. Practice . . . 3301
24.12.5.3 C. Going deeper . . . 3303

24.13 Fast Matrix Multiplication **3304**
24.13.1 What a fast algorithm would be . . . 3304
24.13.2 Seven products instead of eight . . . 3306
24.13.3 The recursion, and its exact cost . . . 3308
24.13.4 What is known about the exponent, and what is not . . . 3310
24.13.5 Accuracy: a norm bound, but not an entrywise one . . . 3310
24.13.6 Whether to use it . . . 3314
24.13.7 Summary and transfer . . . 3315
24.13.8 Exercises . . . 3319
24.13.8.1 A. Check your understanding . . . 3319
24.13.8.2 B. Practice . . . 3320
24.13.8.3 C. Going deeper . . . 3322

Notation **3324**
24.0.1 The standing conventions . . . 3324
24.0.2 Numbers, sets, and fields . . . 3325
24.0.3 Spaces . . . 3326
24.0.4 Vectors, matrices, and their operations . . . 3327
24.0.5 Maps and their spaces . . . 3328
24.0.6 Inner products and norms . . . 3329
24.0.7 Bilinear forms, tensors, and exterior algebra . . . 3330
24.0.8 Spectra and canonical forms . . . 3331
24.0.9 Orders, convexity, and nonnegative matrices . . . 3333
24.0.10 Algebras, groups, and representations . . . 3334
24.0.11 Geometry . . . 3335
24.0.12 Computation . . . 3336
24.0.13 Where the book overrides itself . . . 3337

Preface

This is a book about linear algebra, written for someone reading it alone, with no lecture to fill the gaps and nobody to ask.

It has twenty-five chapters and two hundred and thirty-six sections, in six parts. It begins at Chapter 0 with statements, proofs, sets, functions, equivalence relations, fields, polynomials, and matrices: the tools everything else is built from, and where careful readers first stumble. It ends with representation theory and with numerical computation.

What is in the book

Part I, Spaces and Maps. Vector spaces and dimension; linear systems and elimination, proved rather than recited; linear maps, rank, and nullity; and duality, where a subspace is traded for the measurements that vanish on it.

Part II, Determinants and the Structure of Operators. Polynomial arithmetic first, because the structure of an operator is read off polynomials. Then determinants, built from what a signed volume must do and not from a formula; block matrices; eigenvalues; and the canonical forms that finally decide when two matrices are similar.

Part III, Inner Products and Spectral Theory. Length and angle arrive, ten chapters in. Orthonormal bases, projections, and least squares; the spectral theorems; positive matrices and the singular value decomposition.

Part IV, Forms and Multilinear Algebra. The measuring is taken away again and the pairing kept: bilinear and quadratic forms, congruence, inertia, the classical groups. Then tensor, symmetric, and exterior powers, where the determinant reappears as an induced map on a line.

Part V, Matrix Analysis. Six chapters on how big things are and how far they move: norms, variational principles and interlacing, convexity, nonnegative matrices, perturbation theory, matrix inequalities.

Part VI, Geometry, Algebra, Computation. Affine and projective geometry; algebras and their representations; and a last chapter on how any of this is computed, and what a computed answer means.

What we promise about proof

This is the part worth reading before the rest, because it is unusual and because we keep it.

Every result in this book is in one of two states, and we always say which. Either it is **proved here**, from material that came earlier — never from a later section, and never from an exercise — or it is **stated with credit and marked plainly as not proved here**, in which case nothing in the book depends on it. There is no third state. We never leave a gap and then build

on it.

Keeping that promise costs something, and we would rather the cost were visible. Chapter 7 defines the volume of a parallelepiped by the determinant, proves that no other definition obeys the rules volume must obey, and then states without proof that this agrees with Lebesgue measure, and that nothing depends on it. Chapter 24 lists section by section every classical result it quotes and does not prove, and then says that no theorem in the book rests on any of them.

From Part V on, each chapter opens by listing what it borrows. The clearest instance is Chapter 16. It leans on analysis harder than anything before it, so its opening page numbers the analytic facts it quotes — completeness, monotone convergence, compactness of closed bounded sets, the extreme value theorem, the intermediate value theorem, and the mean value theorem — as (A1) to (A6), and every later use names the fact again by its number. Chapters 17 to 20 and 22 to 24 add nothing to that list and say so. Chapter 21 adds exactly one item, (A7), a theorem about functions of a complex variable that linear algebra cannot produce; it states it once and then sets out which of its own results rest on it and which do not.

That accounting matters to us more than any single theorem in the book.

What a section looks like

Sections are all built the same way.

A section opens with two to four sentences: where we are, what question is open, what this section does. A new idea then arrives through a hook — a naive notion shown failing, an expression that keeps recurring and deserves a name, or a theorem we want and cannot yet state. There is a one-line slogan in plain words, then the definition, with the small words that carry the weight set in bold, then the definition read back clause by clause, because a definition is to be parsed and not memorized.

Examples follow, simplest first, each checked against the definition one clause at a time, spread across the families the book cares about — tuples, polynomials, functions, matrices — with one degenerate case among them. Then a nonexample, made by changing an example as little as possible, so that the exact clause that fails can be named. Where there is a trap, a warning names it with a concrete failing case, before you can fall into it.

Theorems come with a sentence saying what need they fill. Most come with an Idea: a few lines of plain prose on how the proof was found. That is where “why would anyone try that” is answered. The proof is kept separate and does mathematics from its first sentence.

Scattered through the text are Quick checks, about three hundred and fifty of them: one short question, with the answer folded away just below. Take them. They are the cheapest way to find out whether reading felt like understanding.

Every section ends with exercises in three groups. **A** is one question in several parts, to be answered before moving on. **B** is practice like the worked examples. **C** goes deeper: a prove-and-deduce pair, a counterexample, a change of field. There are more than fifteen hundred exercises, each with a complete written solution, folded away until you ask for it.

What we assume

The book is algebraic. For most of its length the scalars are an arbitrary field, and the arguments use the field axioms and what is built from them. Chapters 11 to 13 and much of Part V restrict

to the real and complex numbers, and say why each time: a length needs an order so that it can be positive, and a conjugation so that it can be real.

Analysis does enter, because it must — the Fundamental Theorem of Algebra is not a theorem of algebra — but it never enters silently. Chapter 6 isolates the two analytic facts its proof of that theorem uses and proves everything else. Chapter 10 names five facts from analysis before building the matrix exponential. Chapter 14 names the two results its second-derivative test assumes. Chapter 16 gives the standing list, and every chapter after it cites that list by number.

So we do not ask you to know calculus. We ask you to accept, where the book says so, a short list of named facts about the real numbers. If you have met continuity, compactness, and the mean value theorem before, Part V will read faster. If you have not, you can still follow every proof, because each one says which fact it uses and states it.

What we do assume is some comfort with proof: reading one, and writing one. Chapter 0 supplies the vocabulary and the standard moves — contrapositive, contradiction, induction, double inclusion, checking that something is well defined — and it is a chapter rather than an appendix because those are where readers stumble. We assume no previous linear algebra: matrices, elimination, and determinants are built here from nothing. A reader who has taken a first course will move quickly through the early chapters, and should still not skip them, since the vocabulary of the whole book is fixed there.

How to read it

Read Parts I, II, and III in order. They are the spine, and nothing later stands without them. Chapters 0 to 5 fix the language; Chapters 6 to 10 answer one question, how simple the matrix of an operator can be made; Chapters 11 to 13 answer it again with lengths and angles present.

After that the book branches, and every chapter's opening page tells you what it needs. Spend a minute on it first. It gives the chapter's story, the earlier results it uses with a link to each, a roadmap of one sentence per section, and the named moves the chapter introduces.

Part IV can be deferred, though not skipped. Chapter 17 uses its law of inertia; Chapter 22 uses its quadratic forms and, for one section, its exterior powers; Chapter 23 needs its definition of an algebra and its free vector space. Chapter 24 uses neither chapter at all. Part V is best read in order, with Chapter 16 as its gate. Part VI's three chapters are independent of one another, and each names on its opening page the earlier chapters it never cites.

If you came for one theorem, do not take a chapter list from us. The one that stood here was wrong in every entry, and wrong in the same direction each time: it named fewer chapters than the proofs need. The build now follows the proofs backwards instead, section by section, and works out what a given theorem actually rests on. Ten of those paths, one for each kind of reader, are on the reading paths page. Each gives its sections in reading order, says how many of the two hundred and thirty-six they are, and says how many more the exercises cost; and each is checked, every time the book is built, to contain everything its own proofs use. The dependency graph will do the same for any one section you name.

The proofs are written to be read with a pen. Where a move is not visible, a clause says why we make it; where a clause of a definition is delicate, a nonexample fails on that clause and no other. None of that replaces closing the book and trying to state the theorem yourself. When something will not come, the chapter's opening page usually names the earlier result you are missing.

Chapter 0

Preliminaries

Most people meet linear algebra as a way of computing: row reduction, determinants, maybe eigenvalues. This book treats it as a subject you **prove**. Every statement comes with a reason, and those reasons are built from a small set of ideas: sets, functions, equivalence relations, fields, polynomials, and matrices. This chapter collects that toolkit in one place.

You have probably seen some of this before. Some of it, such as negating a statement with several quantifiers, or checking that a rule on equivalence classes is well defined, is exactly where careful readers later stumble. So each section spends its time on the delicate points and moves quickly past the rest.

How to use this chapter. Skim the sections you know, but try the *A. Check your understanding* exercises at the end of each one. If you can do them without looking back, move on. If not, that section is worth a slower read.

What you need. Nothing from this book. School algebra is enough, plus some familiarity with numbers such as $\mathbb{Z}$, $\mathbb{Q}$, and $\mathbb{R}$. One convention to note now: in this book the natural numbers $\mathbb{N}$ start at 0.

0.0.1 Roadmap

- **Statements, implications, and quantifiers.** How mathematical statements are built, and how to negate them correctly.
- **How proofs are built.** Direct proof, contrapositive, contradiction, cases, uniqueness, counterexamples and induction, and the conventions this book's proofs follow.
- **Sets.** Subsets, operations, indexed families, and the habit of proving two sets equal by double inclusion.
- **Functions.** Injective, surjective, and bijective maps; composition; left, right, and two-sided inverses.
- **Relations, equivalence classes, and quotients.** Sorting a set into classes, and defining functions on classes without ambiguity.
- **Complex numbers.** Arithmetic, conjugates, the modulus, polar form, and roots of unity.
- **Fields.** The number systems in which linear algebra works, from $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$ to the finite fields $\mathbb{F}_p$.
- **Polynomials.** Degree, evaluation, and why a polynomial is not the same thing as a function.
- **Matrices.** Addition, the three ways to read a matrix product, transpose, trace, and inverses.
- **Groups and permutations.** One structure shared by invertible matrices, the nonzero

elements of a field, and rearrangements of a set.

0.0.2 Named moves

These proof techniques appear throughout the book. Each is introduced in this chapter.

- **Contrapositive.** To prove "if P then Q", prove "if not Q then not P".
- **Assume two.** To prove something is unique, suppose two objects have the property and show they are equal.
- **Double inclusion.** To prove $A = B$, prove $A \subseteq B$ and $B \subseteq A$.
- **Suppose $f(a) = f(b)$.** The standard first line when proving that a function is injective.
- **Check well-definedness.** A rule given through a representative must not depend on which representative is chosen.
- **Induction on n.** Find a smaller instance of the problem inside the larger one.

0.1 Statements, Implications, and Quantifiers

Linear algebra is a subject of proofs. Every theorem in this book has the shape "if these hypotheses hold, then this conclusion holds", and most definitions hide a "for every" or a "there exists". Before we can prove anything, we need to read such sentences exactly: when they are true, what they claim, and what their opposite says. This section builds that reading skill. It is the grammar that the rest of the book is written in.

0.1.1 Statements and connectives

A **statement** is a sentence that is either true or false, but not both. "7 is prime" is a statement, and it is true. "$2 + 2 = 5$" is a statement, and it is false. "Is 7 prime?" is not a statement, and neither is "$x > 3$" on its own: its truth depends on x, which nobody has chosen yet. Sentences with a free variable are handled by quantifiers later in this section.

We build longer statements from shorter ones with a few fixed words. Everyday language uses these words loosely, so mathematics pins each one down by saying exactly when the combined statement is true.

A connective builds a new statement whose truth depends only on the truth of its parts.

Definition 0.1.1. Negation, conjunction, disjunction

Let P and Q be statements.

- The **negation** $\neg P$ ("not P") is true when P is false, and false when P is true.
- The **conjunction** $P \wedge Q$ ("P and Q") is true when **both** P and Q are true, and false otherwise.
- The **disjunction** $P \vee Q$ ("P or Q") is true when **at least one** of P, Q is true, and false only when both are false.

A **truth table** lists every combination of truth values of the parts, one per row, and records the value of the compound statement. With T for true and F for false:

P	Q	$\neg P$	$P \wedge Q$	$P \vee Q$
T	T	F	T	T
T	F	F	F	T
F	T	T	F	T
F	F	T	F	F

Two statement forms are **logically equivalent**, written $\equiv$, if they have the same truth value in every row of the truth table. For example, $\neg(\neg P) \equiv P$: negating twice returns the original value in both rows.

A few examples, checked against the definition:

- "7 is prime **and** 7 is even" is false, because its second part is false.
- "7 is prime **or** 7 is even" is true, because its first part is true.
- "$2 < 3$ **or** $3 < 4$" is true. Both parts are true, and that is allowed.

Warning. In mathematics "or" is **inclusive**: $P \vee Q$ is true when both parts are true. The menu phrase "soup or salad", which forbids having both, is a different connective. So "6 is even or 6 is divisible by 3" is true, even though both halves hold.

0.1.2 Implication

Almost every theorem is an implication: "if P, then Q". It is worth being completely precise about when such a sentence counts as true, because the answer surprises many readers in one row of the table.

Think of an implication as a **promise**. A parent says: "If you finish your homework, then you may watch a film." When has the parent broken the promise? Only when the homework is finished and the film is still refused. If the homework is not finished, the promise says nothing about what happens, so whatever happens, the promise was not broken.

An implication is false only when the hypothesis holds and the conclusion fails.

Definition 0.1.2. Implication, converse, contrapositive

Let P and Q be statements.

- The **implication** $P \Rightarrow Q$ ("if P then Q") is false when P is true and Q is false, and it is true in **every** other case. We call P the **hypothesis** and Q the **conclusion**.
- The **converse** of $P \Rightarrow Q$ is the implication $Q \Rightarrow P$.
- The **contrapositive** of $P \Rightarrow Q$ is the implication $\neg Q \Rightarrow \neg P$.

In words: the first clause says an implication can only fail in one way, by a true hypothesis leading to a false conclusion. The converse swaps the roles of hypothesis and conclusion. The contrapositive swaps them **and** negates both.

P	Q	$P \Rightarrow Q$
T	T	T

P	Q	$P \Rightarrow Q$
T	F	F
F	T	T
F	F	T

The implication $P \Rightarrow Q$ is read in many ways, and all of them mean the same thing:

- "P implies Q", "Q whenever P", "Q if P";
- "P only if Q";
- "P is **sufficient** for Q", "Q is **necessary** for P".

The phrase "only if" trips people up. "P only if Q" says that P cannot happen without Q, which is exactly $P \Rightarrow Q$.

The last two rows of the table, where the hypothesis is false, are called **vacuous** cases. An implication with a false hypothesis is **vacuously true**.

Example. A vacuously true implication

Let n be an integer. Is the statement "if n is odd and n is even, then $n = 7$" true?

Solution. Yes. No integer is both odd and even, so the hypothesis "n is odd and n is even" is false for every integer n. By Definition 0.1.2, an implication with a false hypothesis is true, whatever the conclusion says. So the statement is true for every integer n, including $n = 7$ and $n = 0$. It is true, but it tells us nothing.

Why this definition. The two vacuous rows may look like a convention picked for convenience. They are forced once we want implications to work inside "for every" statements. Consider the true fact

for every integer n, if $n > 2$ then $n^2 > 4$.

For this to be true, the implication must be true for **every** integer n. Try $n = -3$: the hypothesis $-3 > 2$ is false, and the conclusion $9 > 4$ is true. Try $n = 1$: the hypothesis $1 > 2$ is false, and the conclusion $1 > 4$ is false. If either row were declared false, this perfectly good theorem would become false. So both vacuous rows must be T. In the same way, a theorem with hypotheses says nothing about objects that fail them, and that is exactly how we want to read it.

Remark. The implication $P \Rightarrow Q$ does **not** claim that P causes Q, or that the two are related at all. "If $2 + 2 = 4$, then 7 is prime" is a true implication, because both parts are true. In practice we only ever prove implications where the hypothesis is used, but the truth value does not depend on that.

The contrapositive is the most useful rewriting of an implication, because it is always interchangeable with the original.

Theorem 0.1.1. Contrapositive equivalence

For all statements P and Q, the implication $P \Rightarrow Q$ is logically equivalent to its contrapositive $\neg Q \Rightarrow \neg P$.

Proof. Let P and Q be statements. We compute both sides in each of the four rows, using Definition 0.1.1 for the negations and Definition 0.1.2 for the implications.

P	Q	$\neg Q$	$\neg P$	$P \Rightarrow Q$	$\neg Q \Rightarrow \neg P$
T	T	F	F	T	T
T	F	T	F	F	F
F	T	F	T	T	T
F	F	T	T	T	T

In the second row, $\neg Q$ is true and $\neg P$ is false, so $\neg Q \Rightarrow \neg P$ is false. In every other row the hypothesis $\neg Q$ is false or the conclusion $\neg P$ is true, so the contrapositive is true. The last two columns agree in every row. This proves that the two statements are logically equivalent. ■

So to prove "if P then Q" we may instead prove "if Q fails then P fails". The next section turns this into a proof technique.

The converse is a different matter. It is a separate statement, and it can be false while the original is true.

Warning. **The converse is not the contrapositive.** "If $x = 2$, then $x^2 = 4$" is true for every real number x. Its converse "if $x^2 = 4$, then $x = 2$" is false: $x = -2$ satisfies the hypothesis and fails the conclusion. Proving $P \Rightarrow Q$ never proves $Q \Rightarrow P$.

Later in the book the converse question comes up constantly. In Chapter 3 we will see that if a square matrix $\mathbf{A}$ is invertible, then $\mathbf{Ax} = \mathbf{0}$ has only the solution $\mathbf{x} = \mathbf{0}$. The converse, "if $\mathbf{Ax} = \mathbf{0}$ has only the solution $\mathbf{x} = \mathbf{0}$, then $\mathbf{A}$ is invertible", also turns out to be true for square matrices. But it is a separate theorem, it needs its own proof, and that proof takes real work.

Quick check. Consider "if an integer n is divisible by 6, then n is even". Write its converse and its contrapositive, and decide which of the three statements are true for every integer n.

Solution. The converse is "if n is even, then n is divisible by 6". The contrapositive is "if n is not even, then n is not divisible by 6", that is, "if n is odd, then $6 \nmid n$".

The original is true: if $n = 6k$ then $n = 2(3k)$ is even. The contrapositive is true, by Theorem 0.1.1. The converse is false: $n = 2$ is even but not divisible by 6.

When an implication and its converse are both true, we say the two statements are equivalent.

The **biconditional** $P \Leftrightarrow Q$ ("P if and only if Q", often written "P iff Q") is the statement $(P \Rightarrow Q) \wedge (Q \Rightarrow P)$. It is true exactly when P and Q have the same truth value:

P	Q	$P \Rightarrow Q$	$Q \Rightarrow P$	$P \Leftrightarrow Q$
T	T	T	T	T
T	F	F	T	F
F	T	T	F	F
F	F	T	T	T

The phrase splits in two: "P if Q" is $Q \Rightarrow P$, and "P only if Q" is $P \Rightarrow Q$. Other phrasings are "P is necessary and sufficient for Q" and "P and Q are equivalent". Proving an "iff" therefore always means proving two implications.

0.1.3 Quantifiers

The sentence "$x^2 \geq 0$" is not yet a statement, because x is not specified. A sentence $P(x)$ that becomes a statement once a value of x is substituted is called a **predicate** (or open sentence) in the variable x. There are two standard ways to turn a predicate into a statement: claim it for every value, or claim it for at least one.

"For every" makes a claim about each element; "there exists" asks for a single witness.

Definition 0.1.3. Quantifiers

Let $P(x)$ be a predicate whose variable x ranges over a fixed collection S.

- The **universal** statement $\forall x \in S : P(x)$ ("for every x in S, $P(x)$") is true when $P(x)$ is true for **every** element x of S, and false otherwise.
- The **existential** statement $\exists x \in S : P(x)$ ("there exists x in S such that $P(x)$") is true when $P(x)$ is true for **at least one** element x of S, and false otherwise.

The symbols $\forall$ and $\exists$ are the **universal** and **existential quantifiers**.

In words: a universal statement is a promise about every element, so a single element where P fails makes it false. An existential statement needs one element that works. It does not say which one, and it does not say there is only one.

The collection S is part of the statement. Changing S can change the truth value, so always ask: *ranging over what?*

Example. Quantified statements

Decide whether each statement is true.

(a) $\forall x \in \mathbb{R} : x^2 \geq 0$.
(b) $\exists x \in \mathbb{R} : x^2 = -1$.
(c) $\exists n \in \mathbb{Z} : n^2 = 2$.

(d) $\forall n \in \mathbb{Z} : n^2 \neq 2$.

Solution.

(a) True. Every real number x is either ≥ 0 or < 0, and in both cases $x^2 = x \cdot x \geq 0$.

(b) False. By (a), $x^2 \geq 0 > -1$ for every real x, so no real x is a witness. Over the complex numbers (introduced later in this chapter), $x = i$ is a witness, so the same predicate becomes true when the range changes from $\mathbb{R}$ to $\mathbb{C}$.

(c) False. If $|n| \leq 1$ then $n^2 \leq 1 < 2$, and if $|n| \geq 2$ then $n^2 \geq 4 > 2$. So no integer squares to 2.

(d) True. This is the statement shown in (c), said the other way round: every integer fails to square to 2.

Parts (c) and (d) already show the pattern of the next subsection: "there is no witness" is the same as "everything fails".

A degenerate case. What if the range contains no elements at all? Take "every integer n with $n^2 < 0$ satisfies $n = 7$". There are no such integers, so there is nothing that could fail, and the statement is true. It is the quantifier version of vacuous truth: it reads $\forall n \in \mathbb{Z} : (n^2 < 0 \Rightarrow n = 7)$, and every instance has a false hypothesis. By contrast, "there exists an integer n with $n^2 < 0$" is false, because there is no witness. This case matters later: facts about "every element of an empty list" will be vacuously true, and that is what makes certain edge cases of definitions work.

Remark. We write $\exists!\, x \in S : P(x)$ for "there exists a **unique** x in S with $P(x)$". It means two things: some x satisfies P, and any two elements satisfying P are equal. The next section shows how each half is proved.

0.1.4 The order of quantifiers

Most definitions in this book use several quantifiers in a row. Their order is part of the meaning.

Example. Swapping two quantifiers

Compare the two statements

$$\text{(i)}\ \ \forall x \in \mathbb{R}\ \exists y \in \mathbb{R} : y > x, \qquad \text{(ii)}\ \ \exists y \in \mathbb{R}\ \forall x \in \mathbb{R} : y > x.$$

Decide which is true.

Solution. Statement (i) says: whichever real x you are given, you can find a real y larger than it. This is true. Given x, take $y = x + 1$, and then $y > x$.

Statement (ii) says: there is one real number y that is larger than every real x. This is false. Whatever y is proposed, the choice $x = y + 1$ gives $y > x$ false. So no y is a witness.

The difference is **who chooses first**. In (i), x is chosen first and y may depend on it: $y = x+1$ changes with x. In (ii), y must be fixed before x is known, so it has to work for all x at once. Reading quantifiers as a game helps: $\forall$ is a move by an opponent, $\exists$ is your reply, and you may only use information already on the table.

Warning. Never swap a $\forall$ and an $\exists$ without a proof. The implication "$\exists y\,\forall x \Rightarrow \forall x\,\exists y$" is always valid, since the single y works for each x. The reverse direction fails, as Example (Swapping two quantifiers) shows. Two quantifiers of the **same** kind, such as $\forall x\,\forall y$, may be swapped freely.

Quick check. Over the positive real numbers, statement (A) is "$\forall \varepsilon > 0\ \exists \delta > 0 : \delta < \varepsilon$". Statement (B) is obtained by swapping the quantifiers: "$\exists \delta > 0\ \forall \varepsilon > 0 : \delta < \varepsilon$". Are (A) and (B) both true?

Solution. No. (A) is true: given $\varepsilon > 0$, take $\delta = \varepsilon/2$, which is positive and smaller than ε. (B) is false: whatever $\delta > 0$ is proposed, the choice $\varepsilon = \delta$ makes $\delta < \varepsilon$ false. In (A) the number δ may depend on ε; in (B) it may not.

0.1.5 Negation

Much of mathematics consists of showing that something **fails**: a function is not injective, a list is not independent, a claimed theorem is false. To do this we need the negation of the statement written in a usable form. Putting "it is not the case that" in front is correct but useless; we want the negation with the "not" pushed all the way inside, next to a simple condition we can check.

First, the connectives.

Theorem 0.1.2. De Morgan's laws for statements

For all statements P and Q:

(a) $\neg(P \wedge Q) \equiv \neg P \vee \neg Q$;
(b) $\neg(P \vee Q) \equiv \neg P \wedge \neg Q$;
(c) $\neg(P \Rightarrow Q) \equiv P \wedge \neg Q$.

Proof. Let P and Q be statements. We compare truth tables, using Definition 0.1.1 and Definition 0.1.2.

P	Q	$\neg(P \wedge Q)$	$\neg P \vee \neg Q$	$\neg(P \vee Q)$	$\neg P \wedge \neg Q$	$\neg(P \Rightarrow Q)$	$P \wedge \neg Q$
T	T	F	F	F	F	F	F
T	F	T	T	F	F	T	T
F	T	T	T	F	F	F	F
F	F	T	T	T	T	F	F

For (a), $P \wedge Q$ is true only in the first row, so its negation is false only there. The statement $\neg P \vee \neg Q$ is false only when $\neg P$ and $\neg Q$ are both false, which is again only the first row. For (b), $\neg(P \vee Q)$ is true only in the last row, where both parts are false, and so is $\neg P \wedge \neg Q$. For (c), $P \Rightarrow Q$ is false only in the second row, so $\neg(P \Rightarrow Q)$ is true only there, and $P \wedge \neg Q$ is also true only there. In each case the two columns agree in all four rows, as claimed.

■

Law (c) deserves a sentence. The negation of an implication is **not** another implication. To refute "if P then Q", you must exhibit a situation where P holds **and** Q fails.

Now the quantifiers. A universal statement fails exactly when some element fails it, and an existential statement fails exactly when every element fails it. We record this as the rule we will use every time.

To negate, flip each quantifier and negate what is inside.

Definition 0.1.4. Negation of quantifiers

Let $P(x)$ be a predicate with x ranging over S. Then

$$\neg(\forall x \in S : P(x)) \;\equiv\; \exists x \in S : \neg P(x), \qquad \neg(\exists x \in S : P(x)) \;\equiv\; \forall x \in S : \neg P(x).$$

For a statement with several quantifiers, apply these rules from the outside in: each $\forall$ becomes $\exists$, each $\exists$ becomes $\forall$, the ranges stay the same, and the innermost condition is negated using Theorem 0.1.2.

In words: "not every x works" means "some x fails", and "no x works" means "every x fails". Both rules are just Definition 0.1.3 read carefully. The universal statement is false precisely when it is not the case that every element satisfies P, which is precisely when at least one element satisfies $\neg P$. The **ranges do not change**: the negation of "every real number squared is nonnegative" talks about real numbers, not about some other collection.

Example. Negating simple quantified statements

Negate each statement, and say which of the statement and its negation is true.

(a) $\forall x \in \mathbb{R} : x^2 \geq 0$.

(b) $\exists n \in \mathbb{Z} : n$ is even and $n > 10$.

(c) $\forall x \in \mathbb{R}\ \exists y \in \mathbb{R} : y > x$.

Solution.

(a) The negation is $\exists x \in \mathbb{R} : x^2 < 0$. The original is true, by Example (Quantified statements), so the negation is false.

(b) By Definition 0.1.4 and Theorem 0.1.2 (a), the negation is $\forall n \in \mathbb{Z} : n$ is odd or $n \leq 10$. The original is true, with witness $n = 12$, so the negation is false: $n = 12$ is neither odd nor ≤ 10.

(c) Flip both quantifiers and negate the inside: $\exists x \in \mathbb{R}\ \forall y \in \mathbb{R} : y \leq x$. This says that

some real number is at least as large as every real number. The original is true by Example (Swapping two quantifiers), so the negation is false.

The next two examples negate definitions that will appear over and over. We use the word **function** informally here, as a rule f that assigns to each a in a collection A one value $f(a)$ in a collection B; functions are treated properly later in this chapter.

Example. What "not injective" means

A function f from A to B is called **injective** if

$$\forall a_1 \in A\ \forall a_2 \in A : (f(a_1) = f(a_2) \Rightarrow a_1 = a_2).$$

Write down, with no "not" in front, what it means for f to be **not** injective. Then show that $f(x) = x^2$, from $\mathbb{R}$ to $\mathbb{R}$, is not injective.

Solution. By Definition 0.1.4, both $\forall$ become $\exists$, and the inside implication is negated. By law (c) of Theorem 0.1.2, $\neg(f(a_1) = f(a_2) \Rightarrow a_1 = a_2) \equiv f(a_1) = f(a_2) \wedge a_1 \neq a_2$. So f is not injective if and only if

$$\exists a_1 \in A\ \exists a_2 \in A : f(a_1) = f(a_2) \text{ and } a_1 \neq a_2.$$

In words, two **different** inputs have the same output. For $f(x) = x^2$, take $a_1 = 1$ and $a_2 = -1$. Then $f(1) = 1 = f(-1)$ and $1 \neq -1$, so f is not injective.

Example. What "not surjective" means

A function f from A to B is called **surjective** if

$$\forall b \in B\ \exists a \in A : f(a) = b.$$

Negate this, and show that $f(x) = x^2$, from $\mathbb{R}$ to $\mathbb{R}$, is not surjective.

Solution. Flipping the two quantifiers and negating the equation gives: f is not surjective if and only if

$$\exists b \in B\ \forall a \in A : f(a) \neq b.$$

In words, **one** value in B is never reached. For $f(x) = x^2$, take $b = -1$. For every real a we have $f(a) = a^2 \geq 0 > -1$ by Example (Quantified statements), so $f(a) \neq -1$. Hence f is not surjective.

The most important negation in the early chapters is that of a statement ending in "all zero".

We meet it here without any vector spaces, using pairs of real numbers added and scaled entry by entry.

Example. Negating "only the zero combination"

Consider the pairs $(1,2)$, $(2,4)$, $(0,1)$, and the statement

$$\forall a,b,c \in \mathbb{R} : \Big(a(1,2)+b(2,4)+c(0,1)=(0,0) \;\Rightarrow\; a=0 \wedge b=0 \wedge c=0\Big).$$

Negate it, and decide which of the statement and its negation is true.

Solution. Flip the quantifier, and use law (c) of Theorem 0.1.2 for the implication and law (a) for the conjunction. The negation is

$$\exists a,b,c \in \mathbb{R} : a(1,2)+b(2,4)+c(0,1)=(0,0) \quad \text{and} \quad (a\neq 0 \vee b\neq 0 \vee c\neq 0).$$

The last condition says that a,b,c are **not all zero**: at least one of them is nonzero.

The negation is true. Take $a=2$, $b=-1$, $c=0$. Then

$$2(1,2)-(2,4)+0(0,1)=(2-2+0,\ 4-4+0)=(0,0),$$

and $a=2\neq 0$. So the original statement is false.

In Chapter 1 the original statement is what "the list is linearly independent" means, and its negation is "linearly dependent". The small words in the negation carry the weight.

Warning. **"Not all zero" is not "all nonzero".** The negation of "$a=b=c=0$" is "at least one of a,b,c is nonzero". In Example (Negating "only the zero combination") the witness is $(2,-1,0)$, which has a zero entry. In fact **no** witness has all three entries nonzero: comparing entries, $a+2b=0$ and $2a+4b+c=0$, so $c=-2(a+2b)=0$. Negating with "all nonzero" would wrongly conclude that the original statement is true.

Quick check. Negate "every student passed some exam", with no "not" in front of the whole sentence.

Solution. Write it as "$\forall$ students s $\exists$ exam e : s passed e". Flipping both quantifiers and negating the inside gives "$\exists$ student s $\forall$ exams e : s did not pass e". In words: **some student passed no exam**. The tempting answer "no student passed any exam" is the negation of "some student passed some exam", a different statement.

Remark. Each form of statement suggests how to prove or refute it. To prove $\forall x \in S : P(x)$, take an arbitrary $x \in S$ and show $P(x)$. To prove $\exists x \in S : P(x)$, produce a witness. To refute a $\forall$, produce one counterexample. To refute an $\exists$, show every candidate fails. The next section develops these moves into proof techniques.

0.1.6 Exercises

0.1.6.1 A. Check your understanding

Exercise. A1

(a) Decide whether each statement is true: (i) "$3 < 2$ or 3 is odd"; (ii) "if $3 < 2$, then $3 = 5$"; (iii) "if 3 is odd, then $3 < 2$".

(b) Write the converse and the contrapositive of "if an integer n is divisible by 4, then n is even". Which of the three statements are true for every integer n?

(c) Negate $\forall x \in \mathbb{R}\ \exists n \in \mathbb{Z} : n > x$, with no $\neg$ in front. Which of the statement and its negation is true?

(d) True or false: $\neg(P \wedge Q) \equiv \neg P \wedge \neg Q$ for all statements P, Q. Give a reason.

Solution.

(a) (i) True: the second part is true, and a disjunction needs only one true part. (ii) True: the hypothesis $3 < 2$ is false, so the implication is vacuously true. (iii) False: the hypothesis is true and the conclusion is false, which is the one case where an implication fails.

(b) The converse is "if n is even, then n is divisible by 4". The contrapositive is "if n is odd, then n is not divisible by 4". The original is true: $n = 4k$ gives $n = 2(2k)$. The contrapositive is true by Theorem 0.1.1. The converse is false: $n = 2$ is even and not divisible by 4.

(c) By Definition 0.1.4, the negation is $\exists x \in \mathbb{R}\ \forall n \in \mathbb{Z} : n \leq x$. The original is true: given x, some integer exceeds x (for instance the integer part of $|x|$ plus 1). So the negation is false.

(d) False. Take P true and Q false. Then $P \wedge Q$ is false, so $\neg(P \wedge Q)$ is true, while $\neg P \wedge \neg Q$ is false because $\neg P$ is false. The correct law is $\neg(P \wedge Q) \equiv \neg P \vee \neg Q$, by Theorem 0.1.2.

0.1.6.2 B. Practice

Exercise. B1: Negating definitions

Write the negation of each definition as a quantified statement with no $\neg$ in front, and then say it in words.

(a) A function f from $\mathbb{R}$ to $\mathbb{R}$ is **bounded** if $\exists M \in \mathbb{R}\ \forall x \in \mathbb{R} : |f(x)| \leq M$.

(b) A function f from A to B is injective if $\forall a_1, a_2 \in A : (a_1 \neq a_2 \Rightarrow f(a_1) \neq f(a_2))$. Compare your answer with Example (What "not injective" means).

(c) "Every pair of reals is a combination of $(1, 0)$ and $(1, 1)$": $\forall p, q \in \mathbb{R}\ \exists a, b \in \mathbb{R} : (p, q) =$

$a(1,0)+b(1,1)$, where pairs are added and scaled entry by entry.

Solution.

(a) By Definition 0.1.4, f is not bounded if and only if $\forall M \in \mathbb{R}\ \exists x \in \mathbb{R} : |f(x)| > M$. In words: whatever bound M is proposed, some value of f exceeds it in absolute value.

(b) Flip both quantifiers and negate the implication using law (c) of Theorem 0.1.2: $\exists a_1, a_2 \in A : a_1 \neq a_2 \wedge f(a_1) = f(a_2)$. In words: two different inputs share an output. This is the same condition as in Example (What "not injective" means). That is expected: the given implication is the contrapositive of $f(a_1) = f(a_2) \Rightarrow a_1 = a_2$, so by Theorem 0.1.1 the two definitions agree, and so do their negations.

(c) Flip all the quantifiers and negate the equation: $\exists p, q \in \mathbb{R}\ \forall a, b \in \mathbb{R} : (p,q) \neq a(1,0)+b(1,1)$. In words: some pair is not a combination of $(1,0)$ and $(1,1)$, whatever coefficients are used. (The original is in fact true, since $(p-q)(1,0)+q(1,1)=(p,q)$, so this negation is false.)

Exercise. B2: Integers versus reals

For each statement, decide whether it is true when x and y range over $\mathbb{Z}$, and whether it is true when they range over $\mathbb{R}$. Justify your answer.

(a) $\forall x\ \exists y : x < y < x+1$.

(b) $\forall x\ (x \neq 0 \Rightarrow \exists y : xy = 1)$.

(c) $\exists x\ \forall y : xy = y$.

(d) $\forall x\ \exists y : y^2 = x$.

Solution.

(a) Over $\mathbb{Z}$: false. Take $x = 0$. No integer y satisfies $0 < y < 1$. Over $\mathbb{R}$: true. Given x, take $y = x + \frac{1}{2}$; then $x < y < x+1$.

(b) Over $\mathbb{Z}$: false. Take $x = 2$, which is nonzero. If $2y = 1$ for an integer y, then 1 would be even, which is false. Over $\mathbb{R}$: true. Given $x \neq 0$, take $y = 1/x$, which exists because $x \neq 0$, and then $xy = 1$.

(c) True over both. Take $x = 1$, which lies in $\mathbb{Z}$ and in $\mathbb{R}$. Then $xy = y$ for every y. Here one x works for all y, so the order $\exists x\,\forall y$ causes no difficulty.

(d) False over both. Take $x = -1$. For every y, integer or real, $y^2 \geq 0 > -1$ by Example (Quantified statements), so $y^2 \neq -1$.

0.1.6.3 C. Going deeper

Exercise. C1: Exporting a hypothesis

Prove that $P \Rightarrow (Q \Rightarrow R)$ is logically equivalent to $(P \wedge Q) \Rightarrow R$ for all statements P, Q, R. Explain in one sentence why this lets a theorem with two hypotheses be stated either way.

Solution. Let P, Q, R be statements. By Definition 0.1.2, an implication is false exactly when its hypothesis is true and its conclusion is false.

The statement $(P \wedge Q) \Rightarrow R$ is false exactly when $P \wedge Q$ is true and R is false, that is, when P is true, Q is true, and R is false.

The statement $P \Rightarrow (Q \Rightarrow R)$ is false exactly when P is true and $Q \Rightarrow R$ is false. By the same definition, $Q \Rightarrow R$ is false exactly when Q is true and R is false. So $P \Rightarrow (Q \Rightarrow R)$ is false exactly when P is true, Q is true, and R is false.

Both statements are false in exactly the same row of the eight-row truth table, and true in the other seven. Hence they are logically equivalent.

Consequently "if P, then (if Q, then R)" and "if P and Q, then R" say the same thing, so a theorem may list its hypotheses one after another or all together.

Exercise. C2: When can a converse fail too?

(a) Determine whether there exist statements P and Q such that both $P \Rightarrow Q$ and its converse $Q \Rightarrow P$ are false. Justify your answer.

(b) Give predicates $P(n)$ and $Q(n)$ about integers such that both $\forall n \in \mathbb{Z} : (P(n) \Rightarrow Q(n))$ and $\forall n \in \mathbb{Z} : (Q(n) \Rightarrow P(n))$ are false.

(c) Explain why (a) and (b) do not contradict each other.

Solution.

(a) No such statements exist. Suppose $P \Rightarrow Q$ is false. By Definition 0.1.2, P is true and Q is false. Then $Q \Rightarrow P$ has a false hypothesis, so it is true. Hence an implication and its converse can never both be false.

(b) Let $P(n)$ be "n is even" and $Q(n)$ be "n is divisible by 3". The first universal statement is false: $n = 2$ is even and not divisible by 3. The second is false: $n = 3$ is divisible by 3 and not even.

(c) In (a), P and Q are single statements, each with one fixed truth value. In (b), each universal statement is refuted by its own counterexample, and the two counterexamples are **different** integers. At $n = 2$ the converse $Q(2) \Rightarrow P(2)$ is vacuously true, and at $n = 3$ the original $P(3) \Rightarrow Q(3)$ is vacuously true, in line with (a). A "for every" statement can fail at one place while its converse fails somewhere else.

0.2 How Proofs Are Built

The previous section taught us to read statements: what an implication claims, what the quantifiers demand, and how to negate. Now we write proofs of them. A proof is a chain of statements, each one justified by a definition, a hypothesis, or an earlier result, that ends in the statement we want. This section collects the handful of shapes that almost every proof in this book takes, shows each on a small fact about integers, and ends with the conventions that the book's proofs follow.

Throughout, a theorem will usually come with two blocks. The **Idea** says how the argument

was found, often by working backwards from the goal. The **Proof** then presents the argument forwards, one justified step at a time. Read the Idea first, then check the Proof line by line.

0.2.1 Direct proof

Most statements have the form "for every x, if $P(x)$ then $Q(x)$". The direct way to prove this follows the shape of the statement:

1. **Let** x be arbitrary. By Definition 0.1.3, a universal statement is a claim about each element, so we must not pick a convenient one.
2. **Suppose** $P(x)$. By Definition 0.1.2, the implication can only fail when $P(x)$ is true, so that is the only case we need to handle.
3. **Unwind** $P(x)$ and $Q(x)$ into their definitions. Then push forward from what we have towards what we want, often working backwards from the goal on scrap paper first.

Before we can prove anything about integers, we need definitions to unwind.

Definition 0.2.1. Even and odd

An integer n is **even** if there **exists** an integer k with $n = 2k$. An integer n is **odd** if there **exists** an integer k with $n = 2k + 1$.

So $0 = 2 \cdot 0$ is even, $-6 = 2 \cdot (-3)$ is even, and $7 = 2 \cdot 3 + 1$ is odd. Each definition is an existential statement: to show a number is even, we must produce the integer k. We use the school fact that **every integer is even or odd, and not both**; an exercise at the end of this section proves a more general fact about remainders.

Proposition 0.2.1. Sum of even integers

For all integers m and n, if m and n are even, then $m + n$ is even.

Idea. Givens: $m = 2a$ and $n = 2b$ for some integers a, b. Goal: $m + n = 2k$ for some integer k. So we must *find* k. Adding the givens gives $m + n = 2a + 2b = 2(a + b)$, and $k = a + b$ is the witness.

Proof. Let m and n be even integers. By Definition 0.2.1, there exist integers a and b with $m = 2a$ and $n = 2b$. Then

$$m + n = 2a + 2b = 2(a + b).$$

Since $a + b$ is an integer, Definition 0.2.1 shows that $m + n$ is even. This proves the claim. ∎

The pattern is worth naming. The hypothesis "m is even" is an **existence** statement, so using it hands us an integer a to work with. The goal "$m + n$ is even" is also an existence statement, so proving it means **producing** an integer, here $a + b$. The two letters a and b must be different: m and n are arbitrary, so their halves have nothing to do with each other.

Warning. **A proof of a "for every" statement cannot start with a specific example.** Writing "4+6 = 10, which is even" checks one case and proves nothing about all even integers. Equally, writing $m = 2k$ and $n = 2k$ with the **same** k silently assumes $m = n$, so it only proves that $2m$ is even.

0.2.2 Proof by contrapositive

Sometimes the direct route stalls. Consider the statement "if n^2 is even, then n is even". Going directly, we would suppose $n^2 = 2k$ and try to say something about n. But $n = \pm\sqrt{2k}$ is not visibly of the form $2 \cdot (\text{integer})$, and we are stuck.

By Theorem 0.1.1, "if P then Q" is equivalent to "if not Q then not P". Here the contrapositive is "if n is not even, then n^2 is not even", that is, "if n is odd, then n^2 is odd". Now the hypothesis hands us a formula for n, and squaring a formula is easy.

Proposition 0.2.2. Even squares

For every integer n, if n^2 is even, then n is even.

Idea. The hypothesis n^2 even says little about n, but "n odd" gives $n = 2k+1$ explicitly. So prove the contrapositive: from $n = 2k + 1$, expand n^2 and look for the form $2(\ldots) + 1$.

Proof. Let n be an integer. By Theorem 0.1.1, it suffices to prove that if n is not even, then n^2 is not even. Suppose n is not even. Since every integer is even or odd, n is odd, so by Definition 0.2.1 there is an integer k with $n = 2k + 1$. Then

$$n^2 = 4k^2 + 4k + 1 = 2(2k^2 + 2k) + 1.$$

Since $2k^2 + 2k$ is an integer, n^2 is odd by Definition 0.2.1. Since no integer is both even and odd, n^2 is not even. This proves the contrapositive, and hence the statement.

■

A contrapositive proof is a direct proof of a different, equivalent implication. Use it when the negated conclusion gives more to work with than the hypothesis does.

0.2.3 Proof by contradiction

To prove a statement S by **contradiction**, suppose that S is false and deduce something impossible: a statement together with its negation. Why does this work? We have shown that the implication $\neg S \Rightarrow (R \wedge \neg R)$ is true, for some statement R. Its conclusion $R \wedge \neg R$ is false in every row, so by Definition 0.1.2 the implication can only be true if its hypothesis $\neg S$ is false. Hence S is true.

Contradiction is the natural tool for statements that say something **cannot** happen, or that there is **no** object of some kind. Such statements give nothing to start from directly, while their negation hands us an object to study.

Statements about fractions and primes need the language of divisibility.

Definition 0.2.2. Divisibility and primes

Let a and b be integers. We say a **divides** b, written $a \mid b$, if there **exists** an integer k with $b = ak$. An integer p is **prime** if $p \geq 2$ and the only **positive** divisors of p are 1 and p.

For example, $3 \mid 12$ since $12 = 3 \cdot 4$, and $3 \nmid 7$. The first primes are $2, 3, 5, 7, 11$. The integer 1 is **not** prime, because the definition requires $p \geq 2$; this convention is what makes factorizations into primes unique. In this language, n is even exactly when $2 \mid n$.

A real number is **rational** if it equals a/b for some integers a and b with $b \neq 0$, and **irrational** otherwise. We use the school fact that every rational number can be written **in lowest terms**: as a/b with $b \geq 1$, where no integer $d \geq 2$ divides both a and b. (We justify this fact at the end of this section.)

Theorem 0.2.3. Irrationality of $\sqrt{2}$

There is no rational number whose square is 2. That is, $\sqrt{2}$ is irrational.

Idea. "There is no rational x with $x^2 = 2$" gives nothing to hold, so suppose there is one, and write it in lowest terms a/b. Clearing denominators gives $a^2 = 2b^2$, an integer equation. Then Proposition 0.2.2 applies twice: first a is even, then (after substituting $a = 2c$) b is even. That contradicts lowest terms.

Proof. Suppose, for a contradiction, that some rational number x satisfies $x^2 = 2$. Write $x = a/b$ in lowest terms, so $b \geq 1$ and no integer $d \geq 2$ divides both a and b. Then $a^2/b^2 = 2$, and multiplying by b^2, which is nonzero, gives

$$a^2 = 2b^2.$$

Therefore a^2 is even by Definition 0.2.1, and hence a is even by Proposition 0.2.2. Write $a = 2c$ with c an integer. Substituting gives $4c^2 = 2b^2$, so $b^2 = 2c^2$ is even. By Proposition 0.2.2 again, b is even. Thus 2 divides both a and b, which contradicts the choice of a/b in lowest terms. Hence no rational number has square 2.

■

The contradiction came from the lowest-terms condition, which is the hypothesis we chose to impose. That is typical: in a contradiction proof, look for the assumption you have not yet used, because the contradiction usually comes from there.

Remark. If a contradiction proof of "$P \Rightarrow Q$" assumes P and $\neg Q$ and then only ever derives $\neg P$, it is really a contrapositive proof in disguise. Write it as a contrapositive; it is shorter and clearer.

A second classic contradiction proof, that there are infinitely many primes, needs one more tool

and appears at the end of this section.

0.2.4 Proof by cases

When a hypothesis splits naturally into possibilities, handle each one separately. If $P_1 \vee P_2$ is true, and both $P_1 \Rightarrow Q$ and $P_2 \Rightarrow Q$ are true, then Q is true: whichever of P_1, P_2 holds, its implication delivers Q. The cases must **cover every possibility**. They are allowed to overlap.

Proposition 0.2.4. $n^2 + n$ **is even**

For every integer n, the integer $n^2 + n$ is even.

Idea. Nothing about n is given, but every integer is even or odd, and each case gives a formula for n. Factoring first, $n^2 + n = n(n+1)$, makes each case a one-line computation.

Proof. Let n be an integer. Then $n^2 + n = n(n+1)$. Since every integer is even or odd, there are two cases.

Case 1: n is even. By Definition 0.2.1, $n = 2k$ for some integer k. Then $n(n+1) = 2\big(k(n+1)\big)$, and $k(n+1)$ is an integer.

Case 2: n is odd. By Definition 0.2.1, $n = 2k+1$ for some integer k. Then $n+1 = 2(k+1)$, so $n(n+1) = 2\big(n(k+1)\big)$, and $n(k+1)$ is an integer.

In both cases $n^2 + n$ is even by Definition 0.2.1. This shows the claim for every integer n.

■

0.2.5 Proving "if and only if"

By the definition of the biconditional in the previous section, $P \Leftrightarrow Q$ means $(P \Rightarrow Q) \wedge (Q \Rightarrow P)$. So a proof of an "iff" has **two directions**, labeled ($\Rightarrow$) and ($\Leftarrow$), and each is proved by any method we like. Often one direction is already known.

Proposition 0.2.5. Parity of a square

For every integer n, n is even if and only if n^2 is even.

Proof. Let n be an integer.

($\Rightarrow$) Suppose n is even, so $n = 2k$ for some integer k by Definition 0.2.1. Then $n^2 = 4k^2 = 2(2k^2)$, and $2k^2$ is an integer, so n^2 is even.

($\Leftarrow$) This is Proposition 0.2.2.

Both directions hold, so n is even if and only if n^2 is even.

■

Warning. **Reversing the steps is not automatic.** A common slip is to "prove the converse" by rewriting the same argument backwards without checking that every step reverses. For instance, from $x = 2$ we get $x^2 = 4$; but from $x^2 = 4$ we only get $x = 2$ **or** $x = -2$. Each direction needs its own justification.

0.2.6 Existence and uniqueness

To prove "there exists x with $P(x)$", the most direct method is to **produce a witness** and check it. For example, "there exists an integer n with $n^2 = n + 2$" is proved by $n = 2$: indeed $2^2 = 4 = 2 + 2$.

To prove "there is **at most one** x with $P(x)$", the standard move is: *suppose x and x' both satisfy P, and show $x = x'$.* Notice that we do not assume $x \neq x'$; we simply take two objects with the property and prove they are the same. "There exists a **unique** x" means both existence and at most one, so it takes both arguments.

Here is an example that we will meet again when we study fields and groups. A **binary operation** $*$ on a collection S is a rule that assigns to each ordered pair a, b of elements of S an element $a * b$ of S. Addition and multiplication of integers are binary operations on $\mathbb{Z}$. An **identity element** for $*$ is an element e of S such that $e * a = a$ and $a * e = a$ for **every** a in S.

Proposition 0.2.6. Uniqueness of the identity

Let $*$ be a binary operation on S. Then $*$ has at most one identity element.

Idea. Suppose e and e' are both identities. The only thing we know about them is how they behave in a product, so look at the one product that involves both, $e * e'$, and evaluate it in two ways.

Proof. Suppose e and e' are both identity elements for $*$. Since e' is an identity, $e * e' = e$, taking $a = e$ in its defining property. Since e is an identity, $e * e' = e'$, taking $a = e'$ in its defining property. Therefore $e = e * e' = e'$. This shows that any two identity elements are equal. ∎

So 0 is **the** identity for addition on $\mathbb{Z}$, and 1 is **the** identity for multiplication. The result says nothing about existence. Subtraction on $\mathbb{Z}$ has no identity at all: $a - e = a$ for every a forces $e = 0$, but then $e - a = -a$, which is not a when $a = 1$. Uniqueness proofs of exactly this shape will recur for zero vectors, negatives, inverse matrices, and inverse functions.

0.2.7 Counterexamples

By Definition 0.1.4, the negation of "$\forall x : P(x)$" is "$\exists x : \neg P(x)$". So to **refute** a universal statement, one witness suffices: a single x where $P(x)$ fails. Such an x is a **counterexample**. Conversely, no number of examples can prove a universal statement over an infinite range.

Example. Refuting "every prime is odd"

Decide whether every prime number is odd. Justify your answer.

Solution. False. The statement is $\forall p$ prime : p is odd, so by Definition 0.1.4 one prime that is not odd refutes it. Take $p = 2$. It is prime, since $2 \geq 2$ and a positive divisor of 2 is at most 2, so it is 1 or 2. It is even, since $2 = 2 \cdot 1$, and hence not odd. So 2 is a counterexample.

Warning. **Many examples prove nothing.** The number $n^2 + n + 41$ is prime for every integer n from 0 to 39. Yet the statement "$n^2 + n + 41$ is prime for every $n \geq 0$" is false: at $n = 40$ we get $40^2 + 40 + 41 = 1681 = 41^2$, which is not prime.

0.2.8 Mathematical induction

Some statements are really infinitely many statements, one for each natural number: "for every $n \geq 1$, $1 + 2 + \cdots + n = n(n + 1)/2$". Checking cases one at a time never finishes. Induction replaces the infinite list of checks by two finite tasks, like knocking over a row of dominoes: push the first one, and make sure each domino knocks over the next.

We take this as a basic property of the integers, not something to prove.

Theorem 0.2.7. Principle of Mathematical Induction

Let n_0 be an integer, and for each integer $n \geq n_0$ let $P(n)$ be a statement. Suppose

(i) (**base case**) $P(n_0)$ is true, and

(ii) (**inductive step**) for **every** integer $n \geq n_0$, if $P(n)$ is true then $P(n + 1)$ is true.

Then $P(n)$ is true for every integer $n \geq n_0$.

In the inductive step, the assumption "$P(n)$ is true" is called the **induction hypothesis**. It is not circular: we do not assume the whole conclusion, only that the statement holds at one arbitrary n, and we prove that it then holds at $n + 1$.

Proposition 0.2.8. Sum of the first n integers

For every integer $n \geq 1$,

$$1 + 2 + \cdots + n = \frac{n(n + 1)}{2}.$$

Idea. Passing from n to $n + 1$ adds exactly one term, $n + 1$, to the left side. So the step reduces to the algebra identity $\frac{n(n+1)}{2} + (n + 1) = \frac{(n+1)(n+2)}{2}$, which holds after taking out the factor $n + 1$.

Proof. We prove this by induction on n, using Theorem 0.2.7 with $n_0 = 1$. Let $P(n)$ be the statement $1 + 2 + \cdots + n = n(n+1)/2$.

Base case. For $n = 1$ the left side is 1 and the right side is $1 \cdot 2/2 = 1$. So $P(1)$ is true.

Inductive step. Let $n \geq 1$ and suppose $P(n)$ is true. Then

$$\begin{aligned} 1 + 2 + \cdots + n + (n+1) &= \frac{n(n+1)}{2} + (n+1) \\ &= (n+1)\Big(\frac{n}{2} + 1\Big) \\ &= \frac{(n+1)(n+2)}{2}, \end{aligned}$$

where the first equality uses the induction hypothesis $P(n)$. This is $P(n+1)$.

By Theorem 0.2.7, $P(n)$ is true for every integer $n \geq 1$, as claimed.

■

The base case does not have to be 1, and the step does not have to be an equation. In this book the **natural numbers** are $\mathbb{N} = \{0, 1, 2, \dots\}$: they **include** 0. (Some books start at 1; the next section fixes our convention formally.)

Proposition 0.2.9. $2^n > n$

For every natural number n, we have $2^n > n$.

Proof. We prove this by induction on n, with $n_0 = 0$. For $n = 0$, $2^0 = 1 > 0$. Now let $n \geq 0$ and suppose $2^n > n$. Then

$$2^{n+1} = 2^n + 2^n > n + 2^n \geq n + 1,$$

where the strict inequality uses the induction hypothesis and the last step uses $2^n \geq 1$. Hence $2^{n+1} > n + 1$. By Theorem 0.2.7, $2^n > n$ for every $n \in \mathbb{N}$.

■

The inductive step must work for **every** $n \geq n_0$, including the smallest ones. The following fake proof fails exactly there.

Warning. **The "all horses are the same color" fallacy.** *Claim:* in every group of $n \geq 1$ horses, all horses have the same color. *"Proof":* for $n = 1$ there is nothing to show. Suppose the claim holds for groups of n horses, and take a group of $n+1$ horses $h_1, \dots, h_{n+1}$. The group $h_1, \dots, h_n$ is single-colored by the induction hypothesis, and so is $h_2, \dots, h_{n+1}$. The two groups share a horse, so all $n+1$ horses have its color. The claim is plainly false, so this "proof" has a gap. Before opening the check below, try to find the exact value of n where the step breaks.

Quick check. Where does the inductive step in the horse argument fail?

Solution. It fails when passing from $n = 1$ to $n + 1 = 2$. The two groups are then $\{h_1\}$ and $\{h_2\}$, which share **no** horse, so nothing links the color of h_1 to that of h_2. The sentence "the two groups share a horse" is true only when $n \geq 2$. Since the step $P(1) \Rightarrow P(2)$ is not proved, the chain of dominoes breaks at the very first push, and the base case alone gives nothing.

0.2.9 Strong induction and well-ordering

In some arguments, knowing $P(n)$ alone is not enough to reach $P(n+1)$: we need the statement at some **earlier** value, and we do not know in advance which one. Strong induction allows the hypothesis at all earlier values at once.

Theorem 0.2.10. Strong induction

Let n_0 be an integer, and for each integer $n \geq n_0$ let $P(n)$ be a statement. Suppose that for **every** integer $n \geq n_0$,

$$\text{if } P(k) \text{ is true for all integers } k \text{ with } n_0 \leq k < n, \text{ then } P(n) \text{ is true.}$$

Then $P(n)$ is true for every integer $n \geq n_0$.

Idea. Ordinary induction applies to the stronger statement $Q(n)$: "$P(k)$ holds for all k from n_0 to n". The hypothesis is exactly what pushes $Q(n)$ to $Q(n+1)$. At $n = n_0$ the condition "for all k with $n_0 \leq k < n_0$" is vacuous, so the hypothesis already contains the base case.

Proof. For each integer $n \geq n_0$, let $Q(n)$ be the statement "$P(k)$ is true for all integers k with $n_0 \leq k \leq n$". We prove $Q(n)$ for all $n \geq n_0$ by Theorem 0.2.7.

Base case. There is no integer k with $n_0 \leq k < n_0$, so the condition "$P(k)$ is true for all such k" is vacuously true. By the hypothesis with $n = n_0$, $P(n_0)$ is true. Hence $Q(n_0)$ is true.

Inductive step. Let $n \geq n_0$ and suppose $Q(n)$ is true, so $P(k)$ holds for all k with $n_0 \leq k < n+1$. By the hypothesis applied to $n+1$, $P(n+1)$ is true. Together with $Q(n)$, this gives $Q(n+1)$.

By Theorem 0.2.7, $Q(n)$ holds for all $n \geq n_0$. In particular $P(n)$ holds for all $n \geq n_0$, since $Q(n)$ includes $k = n$. This proves the theorem. ∎

The same property of the natural numbers can be phrased without any statements $P(n)$ at all: there is no endless strictly decreasing sequence of natural numbers, so every search for a smallest example ends.

Theorem 0.2.11. Well-Ordering Principle

Every nonempty collection of natural numbers has a least element: an element m of the collection with $m \leq s$ for every s in the collection.

We take this as a basic property too. It is equivalent to the Principle of Mathematical Induction, in the sense that each can be proved from the other, but we will not need that. Well-ordering is the natural language for "take a smallest counterexample" arguments, which are contradiction proofs in disguise.

Here is a result that needs strong induction, because the smaller number we reach is not $n - 1$.

Proposition 0.2.12. Existence of a prime factor

Every integer $n \geq 2$ has a prime divisor.

Idea. If n is prime, n itself works. If not, $n = ab$ with $1 < a < n$, and a prime dividing a also divides n. The factor a can be any size below n, so we need the statement for all smaller values, which is strong induction.

Proof. We use Theorem 0.2.10 with $n_0 = 2$. Let $n \geq 2$, and suppose every integer k with $2 \leq k < n$ has a prime divisor.

Case 1: n is prime. Then $n \mid n$, so n is a prime divisor of n.

Case 2: n is not prime. Since $n \geq 2$, Definition 0.2.2 gives a positive divisor a of n with $a \neq 1$ and $a \neq n$. Write $n = ab$ with b an integer. Since a and n are positive, $b \geq 1$, and $b \neq 1$ because $a \neq n$; so $b \geq 2$ and $a = n/b < n$. Thus $2 \leq a < n$, and by the hypothesis a has a prime divisor p, say $a = pc$. Then $n = p(cb)$, so $p \mid n$.

In both cases n has a prime divisor. By Theorem 0.2.10, this holds for every $n \geq 2$.

■

Now the promised contradiction proof.

Theorem 0.2.13. Infinitely many primes

There are infinitely many prime numbers.

Idea. Suppose the primes form a finite list. Build a number that no prime on the list can divide: multiply them all and add 1. Dividing by any p_i leaves remainder 1. But Proposition 0.2.12 says this number has some prime divisor, and that prime must be on the list.

Proof. Suppose, for a contradiction, that there are only finitely many primes, and list them

all as $p_1, p_2, \ldots, p_k$. Let

$$N = p_1 p_2 \cdots p_k + 1.$$

The list is nonempty, since 2 is prime, and each $p_i \geq 2$; hence $N \geq 3$. By Proposition 0.2.12, N has a prime divisor p. Since the list contains every prime, $p = p_i$ for some i. Then $p_i \mid N$ and $p_i \mid p_1 p_2 \cdots p_k$, so there are integers s, t with $N = p_i s$ and $p_1 \cdots p_k = p_i t$. Subtracting gives $1 = p_i(s - t)$. Since $s - t$ is an integer and $p_i \geq 2$, this is impossible: $p_i(s - t)$ is 0 when $s = t$, and has absolute value at least 2 otherwise. This contradiction shows that there are infinitely many primes.

■

Remark. Well-ordering also justifies the "lowest terms" fact used in Theorem 0.2.3. Let x be rational. There is at least one way of writing $x = a/b$ with $b \geq 1$: if $x = a/b$ with $b < 0$, then also $x = (-a)/(-b)$ with $-b \geq 1$. Among all such ways, the collection of possible denominators b is a nonempty collection of natural numbers, so by Theorem 0.2.11 it has a least element. Take a representation a/b with that least b. If some $d \geq 2$ divided both, say $a = da'$ and $b = db'$, then $x = a'/b'$ with $1 \leq b' < b$, contradicting minimality. So a/b is in lowest terms.

0.2.10 Writing proofs in this book

The proofs in this book follow a few conventions. Knowing them makes proofs faster to read, and copying them makes your own proofs easier to check.

- **The first sentence does mathematics.** Proofs open with "Let ...", "Suppose ...", "Since ...", or "By ...". The motivation has already been given in the Idea block.
- **Every step names its reason.** Each line cites a hypothesis, a definition, or an earlier result by name, as in "by Proposition 0.2.2". When a display hides a reason, a clause such as "where the first equality uses the induction hypothesis" points to it.
- **Structure is marked.** Two directions of an iff are labeled ($\Rightarrow$) and ($\Leftarrow$), cases are labeled *Case 1* and *Case 2*, and inductions name their base case and inductive step.
- **No hidden gaps.** Words like "obviously" and "clearly" never stand in for an argument. If a step is short, it is written out; if it is routine and has been done before, the proof says so.
- **The ending restates the result**: "This shows ...", "This proves ...", "as claimed".

How to read a proof here. First read the statement and identify the givens and the goal. Then read the Idea, which says why the argument has the shape it has. Then read the Proof with a pencil. At each line, ask which earlier line or result justifies it; if you cannot answer, that is the line to reread. Finally, ask which hypothesis each step used. If a hypothesis was never used, either it was not needed or you have missed a step.

0.2.11 Exercises

0.2.11.1 A. Check your understanding

Exercise. A1

For each statement, name the proof technique you would use first (direct, contrapositive,

contradiction, cases, induction, or counterexample), and give a one-sentence reason. Then carry out the proof or refutation.

(a) The sum of two odd integers is even.

(b) For every integer n, if $3n+2$ is odd, then n is odd.

(c) For every integer $n \geq 1$, $3 \mid 4^n - 1$.

(d) For every integer $n \geq 1$, $n^2 \geq 2n$.

(e) There is no largest integer.

Solution.

(a) *Direct.* The hypothesis gives formulas to add. Let $m = 2a+1$ and $n = 2b+1$ with a, b integers, by Definition 0.2.1. Then $m+n = 2(a+b+1)$, which is even.

(b) *Contrapositive.* The conclusion's negation "n even" gives a formula, while "$3n+2$ odd" does not. By Theorem 0.1.1 it suffices to show: if n is even then $3n+2$ is even. If $n = 2k$, then $3n+2 = 2(3k+1)$, which is even, hence not odd.

(c) *Induction.* It is a statement for every $n \geq 1$, and $4^{n+1} - 1$ is built from $4^n - 1$. For $n = 1$, $4 - 1 = 3 = 3 \cdot 1$. If $4^n - 1 = 3k$, then $4^{n+1} - 1 = 4(4^n - 1) + 3 = 3(4k+1)$. By Theorem 0.2.7 the claim holds for all $n \geq 1$.

(d) *Counterexample.* The statement is false, so a single witness refutes it. At $n = 1$, $n^2 = 1 < 2 = 2n$.

(e) *Contradiction.* It says an object does not exist, so assume it does. Suppose m is a largest integer. Then $m+1$ is an integer with $m+1 > m$, contradicting the choice of m. Hence there is no largest integer.

0.2.11.2 B. Practice

Exercise. B1: Contrapositive

Prove that for every integer n, if n^2 is odd, then n is odd.

Solution. Let n be an integer. By Theorem 0.1.1, it suffices to prove that if n is not odd, then n^2 is not odd. Suppose n is not odd. Since every integer is even or odd, n is even, so $n = 2k$ for some integer k by Definition 0.2.1. Then $n^2 = 4k^2 = 2(2k^2)$, so n^2 is even. Since no integer is both even and odd, n^2 is not odd. This proves the statement.

Exercise. B2: Induction

Prove that for every integer $n \geq 1$, the sum of the first n odd numbers is n^2:

$$1 + 3 + 5 + \cdots + (2n-1) = n^2.$$

Solution. We prove this by induction on n, using Theorem 0.2.7 with $n_0 = 1$. Let $P(n)$ be the displayed statement.

Base case. For $n = 1$ the left side is 1 and the right side is $1^2 = 1$, so $P(1)$ is true.

Inductive step. Let $n \geq 1$ and suppose $P(n)$ is true. The $(n+1)$-st odd number is $2(n+1)-1 = 2n+1$. Then

$$1 + 3 + \cdots + (2n-1) + (2n+1) = n^2 + 2n + 1 = (n+1)^2,$$

where the first equality uses the induction hypothesis. This is $P(n+1)$.

By Theorem 0.2.7, $P(n)$ holds for every $n \geq 1$.

Exercise. B3: $\sqrt{3}$ is irrational

You may use the school fact that every integer has exactly one of the forms $3k$, $3k+1$, $3k+2$ with k an integer.

(a) Prove that for every integer n, if $3 \mid n^2$ then $3 \mid n$.

(b) Hence prove that there is no rational number whose square is 3.

Solution.

(a) Let n be an integer. By Theorem 0.1.1, it suffices to show that if $3 \nmid n$ then $3 \nmid n^2$. Suppose $3 \nmid n$. Then $n = 3k+1$ or $n = 3k+2$ for some integer k.
Case 1: $n = 3k+1$. Then $n^2 = 9k^2 + 6k + 1 = 3(3k^2 + 2k) + 1$.
Case 2: $n = 3k+2$. Then $n^2 = 9k^2 + 12k + 4 = 3(3k^2 + 4k + 1) + 1$.
In both cases n^2 has the form $3j+1$ with j an integer. Since an integer has only one of the three forms, n^2 is not of the form $3j$, so $3 \nmid n^2$. This proves (a).

(b) Suppose, for a contradiction, that a rational number x satisfies $x^2 = 3$. Write $x = a/b$ in lowest terms, so $b \geq 1$ and no integer $d \geq 2$ divides both a and b. Multiplying $a^2/b^2 = 3$ by $b^2 \neq 0$ gives $a^2 = 3b^2$. Therefore $3 \mid a^2$, and by (a), $3 \mid a$. Write $a = 3c$. Then $9c^2 = 3b^2$, so $b^2 = 3c^2$ and $3 \mid b^2$. By (a) again, $3 \mid b$. Thus 3 divides both a and b, contradicting lowest terms. Hence no rational number has square 3.

0.2.11.3 C. Going deeper

Exercise. C1: Stamps of 4 and 5

(a) Show that each of $12, 13, 14, 15$ can be written as $4a + 5b$ with $a, b \in \mathbb{N}$.

(b) Prove by strong induction that every integer $n \geq 12$ can be written as $4a + 5b$ with $a, b \in \mathbb{N}$.

(c) Show that 11 cannot be written in this form, so the bound 12 in (b) cannot be lowered to 11.

Hint: for $n \geq 16$, look at $n - 4$.

Solution.

(a) $12 = 4 \cdot 3 + 5 \cdot 0$, $13 = 4 \cdot 2 + 5 \cdot 1$, $14 = 4 \cdot 1 + 5 \cdot 2$, and $15 = 4 \cdot 0 + 5 \cdot 3$.

(b) We use Theorem 0.2.10 with $n_0 = 12$. Let $n \geq 12$, and suppose every integer k with $12 \leq k < n$ can be written as $4a + 5b$ with $a, b \in \mathbb{N}$.
Case 1: $12 \leq n \leq 15$. Then n has the required form by (a).
Case 2: $n \geq 16$. Then $k = n-4$ satisfies $12 \leq k < n$, so by the hypothesis $n-4 = 4a+5b$ with $a, b \in \mathbb{N}$. Hence $n = 4(a+1) + 5b$, and $a + 1 \in \mathbb{N}$.
In both cases n has the required form. By Theorem 0.2.10, every integer $n \geq 12$ does.

(c) Suppose $11 = 4a + 5b$ with $a, b \in \mathbb{N}$. Since $5b \leq 11$, $b \in \{0, 1, 2\}$. If $b = 0$ then $4a = 11$; if $b = 1$ then $4a = 6$; if $b = 2$ then $4a = 1$. None of $11, 6, 1$ is a multiple of 4, so no such a exists. Hence 11 is not of the form $4a + 5b$.

Exercise. C2: Division with remainder

Let a and b be integers with $b \geq 1$. Let R be the collection of all **nonnegative** integers of the form $a - qb$ with q an integer.

(a) Show that R is nonempty.

(b) Let r be the least element of R, which exists by Theorem 0.2.11, and write $r = a - qb$. Prove that $0 \leq r < b$.

(c) Prove that the pair (q, r) is unique: if $a = qb + r = q'b + r'$ with $0 \leq r < b$ and $0 \leq r' < b$, then $q = q'$ and $r = r'$.

Hint for (a): try $q = -|a|$.

Solution.

(a) Take $q = -|a|$. Then $a - qb = a + |a|b$. Since $b \geq 1$ and $|a| \geq 0$, we have $|a|b \geq |a|$, so $a + |a|b \geq a + |a| \geq 0$, where the last step uses $|a| \geq -a$. Hence $a + |a|b$ belongs to R, and R is nonempty.

(b) By (a) and Theorem 0.2.11, R has a least element $r = a - qb$ with $q \in \mathbb{Z}$. Since r belongs to R, $r \geq 0$. Suppose, for a contradiction, that $r \geq b$. Then $r - b = a - (q+1)b \geq 0$, and $q + 1 \in \mathbb{Z}$, so $r - b$ belongs to R. But $b \geq 1$ gives $r - b < r$, contradicting the choice of r as the least element of R. Hence $0 \leq r < b$.

(c) Suppose $qb + r = q'b + r'$ with $0 \leq r, r' < b$. Then $r' - r = (q - q')b$, so $b \mid r' - r$. From the bounds, $-b < r' - r < b$. Suppose $q \neq q'$. Then $|q - q'| \geq 1$, so $|r' - r| = |q - q'|\, b \geq b$, contradicting $-b < r' - r < b$. Hence $q = q'$, and then $r' - r = 0$, so $r = r'$. This proves uniqueness.

Together, (a)–(c) show that for integers a and $b \geq 1$ there are unique integers q, r with $a = qb + r$ and $0 \leq r < b$. With $b = 2$, this is the fact that every integer is even or odd, and not both.

0.3 Sets

Every object in linear algebra is a set with some extra structure. A vector space is a set on which we can add and scale, a subspace is a subset of it, and a linear map sends elements of one set to elements of another. Before the structure arrives, we need the plain language of sets: how to describe them, when two of them are equal, and how to combine them. The most important

move in this section is proving that two sets are equal by proving two inclusions, a move we will use for subspaces, spans, kernels, and images throughout the book.

0.3.1 Sets, elements, and set-builder notation

A **set** is a collection of objects, called its **elements**. We write $x \in A$ for "x is an element of A" (also "x belongs to A", "x lies in A"), and $x \notin A$ for its negation. A set is determined by its elements and nothing else: it has no order and no repetition. So $\{1, 2\}$, $\{2, 1\}$, and $\{1, 1, 2\}$ all describe the same set, whose elements are 1 and 2. We make this precise below.

Small sets can be listed between braces. Infinite sets we meet constantly have fixed names.

Definition 0.3.1. Common number sets

We write

- $\mathbb{N} = \{0, 1, 2, 3, \dots\}$ for the **natural numbers**; in this book $\mathbb{N}$ **includes** 0;
- $\mathbb{Z} = \{\dots, -2, -1, 0, 1, 2, \dots\}$ for the **integers**;
- $\mathbb{Q}$ for the **rational numbers**, the numbers a/b with $a, b \in \mathbb{Z}$ and $b \neq 0$;
- $\mathbb{R}$ for the **real numbers**;
- $\mathbb{C}$ for the **complex numbers**, studied later in this chapter.

Books disagree about whether $0 \in \mathbb{N}$. With our convention, the Well-Ordering Principle (Theorem 0.2.11) applies to collections that contain 0, and "$n \in \mathbb{N}$" means $n \geq 0$. When we need positive integers only, we say "$n \geq 1$". For real numbers $a \leq b$ we also use **intervals**: $[a, b]$ is the set of real x with $a \leq x \leq b$, (a, b) the set with $a < x < b$, and $[a, b)$, $(a, b]$ are defined in the same way.

Listing elements fails for most sets we care about, such as "the real numbers whose square is less than 4". Instead we describe a set by a property, and select from a set we already have.

Definition 0.3.2. Set-builder notation

Let S be a set and let $P(x)$ be a predicate for $x \in S$. The set

$$\{x \in S : P(x)\}$$

is the set whose elements are **exactly** those $x \in S$ for which $P(x)$ is true.

In words: an object y belongs to $\{x \in S : P(x)\}$ if and only if $y \in S$ **and** $P(y)$ is true. Both parts matter. The ambient set S says where we are looking, and the predicate says what we keep. The colon is read "such that"; some books use a vertical bar instead. When it is plain which sets the elements come from, as in the unions and intersections below, we sometimes leave the ambient set out and write $\{x : P(x)\}$.

A second form describes the elements by a formula: $\{2k : k \in \mathbb{Z}\}$ is the set of all numbers of the form $2k$ with k an integer. An object y belongs to it exactly when $y = 2k$ for **some** $k \in \mathbb{Z}$. This form will describe sets such as "all combinations of given vectors" in Chapter 1.

Example. Reading set-builder notation

Describe each set more simply.

(a) $\{x \in \mathbb{R} : x^2 < 4\}$.

(b) $\{x \in \mathbb{Z} : x^2 < 4\}$.

(c) $\{n \in \mathbb{Z} : n \text{ is even}\}$ and $\{2k : k \in \mathbb{Z}\}$.

(d) $\{(x, y) : x, y \in \mathbb{R},\ x + y = 1\}$, a set of ordered pairs (made precise under Cartesian products below).

Solution.

(a) A real number satisfies $x^2 < 4$ exactly when $-2 < x < 2$, so the set is the interval $(-2, 2)$.

(b) The same predicate, selected from $\mathbb{Z}$ instead, keeps only integers: the set is $\{-1, 0, 1\}$. Changing the ambient set changes the set.

(c) Both are the set $\{\ldots, -4, -2, 0, 2, 4, \ldots\}$ of even integers. The first selects the integers with a property; the second produces each element by the formula $2k$. By the definition of "even" (Definition 0.2.1), an integer is even exactly when it equals $2k$ for some $k \in \mathbb{Z}$.

(d) The pairs of reals whose coordinates add to 1: the line in the plane through $(1, 0)$ and $(0, 1)$.

What if no element of S satisfies the predicate? We still want a set, and it has no elements.

Definition 0.3.3. Empty set

The **empty set** $\varnothing$ is the set with **no** elements: $x \notin \varnothing$ for every object x.

For example, $\{x \in \mathbb{R} : x^2 = -1\} = \varnothing$, since $x^2 \geq 0$ for every real x (Example (Quantified statements)). Also $\{n \in \mathbb{Z} : 0 < n < 1\} = \varnothing$. We will see shortly that there is only **one** empty set, so these two descriptions give the same set.

Statements about all elements of $\varnothing$ are vacuously true: "every element of $\varnothing$ is a prime number" has no element that could fail, just as in the degenerate case discussed after Example (Quantified statements).

Warning. **$\varnothing$, $\{\varnothing\}$, and $\{0\}$ are three different sets.** The set $\varnothing$ has no elements. The set $\{\varnothing\}$ has exactly one element, namely the empty set. The set $\{0\}$ has exactly one element, namely the number 0. A box containing an empty box is not an empty box. In Chapter 1 this matters: $\{0\}$ will be a subspace, and $\varnothing$ never is.

0.3.2 Subsets

Most sets in this book live inside a bigger set: the even integers inside $\mathbb{Z}$, a line inside the plane, and later a subspace inside a vector space. The first thing to check about a candidate subspace will be that it sits inside the space at all. We need a word for "sits inside".

A is a subset of B when everything in A is also in B.

Definition 0.3.4. Subset

Let A and B be sets. We say A is a **subset** of B, and write $A \subseteq B$, if

$$\forall x : (x \in A \Rightarrow x \in B).$$

If $A \subseteq B$ and $A \neq B$, then A is a **proper** subset of B, written $A \subsetneq B$.

In words: $A \subseteq B$ says that **every** element of A is an element of B. It says nothing about elements of B; B may contain many more. The definition is an implication inside a "for every", so the proof template of the previous section applies: to show $A \subseteq B$, **let** $x \in A$ and show $x \in B$.

By Definition 0.1.4 and Theorem 0.1.2, the negation reads

$$A \not\subseteq B \iff \exists x : x \in A \text{ and } x \notin B.$$

So to show A is **not** a subset of B, exhibit one element of A outside B.

Example. Subsets

Decide whether each inclusion holds.

(a) $\mathbb{N} \subseteq \mathbb{Z} \subseteq \mathbb{Q} \subseteq \mathbb{R}$.
(b) $\{1, 2\} \subseteq \{1, 2, 3\}$ and $\{1, 4\} \subseteq \{1, 2, 3\}$.
(c) $A \subseteq A$ and $\varnothing \subseteq A$, for any set A.
(d) $\{x \in \mathbb{R} : x^2 = 1\} \subseteq \mathbb{Z}$.

Solution.

(a) All hold. Every natural number is an integer; every integer n equals $n/1$, so it is rational; every rational number is real.
(b) The first holds: each of $1, 2$ is an element of $\{1, 2, 3\}$. The second fails, although it differs from the first in only one element: 1 passes, but $4 \in \{1, 4\}$ and $4 \notin \{1, 2, 3\}$, so 4 is a witness for $\{1, 4\} \not\subseteq \{1, 2, 3\}$.
(c) Both hold. For $A \subseteq A$: let $x \in A$; then $x \in A$. For $\varnothing \subseteq A$: the implication $x \in \varnothing \Rightarrow x \in A$ has a false hypothesis for every x by Definition 0.3.3, so it is vacuously true (Definition 0.1.2). The empty set is a subset of every set, including itself.
(d) Holds. Let $x \in \mathbb{R}$ with $x^2 = 1$. Then $(x - 1)(x + 1) = 0$, so $x = 1$ or $x = -1$, and both are integers.

Warning. **$\in$ and $\subseteq$ are different relations.** $x \in A$ says the object x is one of the elements of A. $A \subseteq B$ compares two sets, element by element. So $1 \in \{1, 2\}$ is true while $1 \subseteq \{1, 2\}$ makes no sense (1 is not a set here); and $\{1\} \subseteq \{1, 2\}$ is true while $\{1\} \in \{1, 2\}$ is false.

Quick check. Let $X = \{\{1, 2\}, 3\}$. Is $\{1, 2\} \subseteq X$? Is $\{1, 2\} \in X$?

Solution. The set X has exactly two elements: the set $\{1,2\}$ and the number 3. So $\{1,2\} \in X$ is **true**. But $\{1,2\} \subseteq X$ is **false**: $1 \in \{1,2\}$, and 1 is neither the set $\{1,2\}$ nor the number 3, so $1 \notin X$.

0.3.3 Equality of sets and double inclusion

We said a set is determined by its elements. Here is the precise version.

Definition 0.3.5. Set equality

Two sets A and B are **equal**, written $A = B$, if they have exactly the same elements:

$$\forall x : (x \in A \Leftrightarrow x \in B).$$

So $\{1,2\} = \{2,1\} = \{1,1,2\}$: each object belongs to one of these sets exactly when it is 1 or 2. Order and repetition in a list are features of how we **write** a set, not of the set.

Proving a biconditional for every x at once is awkward. It is much easier to prove two implications separately, and that is exactly what the next result allows.

Theorem 0.3.1. Double inclusion

Let A and B be sets. Then $A = B$ if and only if $A \subseteq B$ and $B \subseteq A$.

Proof. By Definition 0.3.5, $A = B$ means that for every x, both $x \in A \Rightarrow x \in B$ and $x \in B \Rightarrow x \in A$ hold, since a biconditional is the conjunction of an implication and its converse. A statement of the form "for every x, $P(x)$ and $Q(x)$" holds exactly when "for every x, $P(x)$" and "for every x, $Q(x)$" both hold, by Definition 0.1.3. Hence $A = B$ holds exactly when $\forall x\,(x \in A \Rightarrow x \in B)$ and $\forall x\,(x \in B \Rightarrow x \in A)$ both hold. By Definition 0.3.4, these are $A \subseteq B$ and $B \subseteq A$. This proves the theorem. ∎

This is the move **"prove set equality by double inclusion"**. The proof is written in two labeled halves, ($\subseteq$) and ($\supseteq$), each starting "Let $x \in \ldots$". Here is a first consequence.

Remark. There is only one empty set. If E and E' are sets with no elements, then $E \subseteq E'$ and $E' \subseteq E$ hold vacuously, as in Example (Subsets) (c), so $E = E'$ by Theorem 0.3.1. This is why we may say **the** empty set.

Example. A solution set by double inclusion

Prove that $\{x \in \mathbb{R} : x^2 - 3x + 2 = 0\} = \{1,2\}$.

Solution. Write $S = \{x \in \mathbb{R} : x^2 - 3x + 2 = 0\}$. By Theorem 0.3.1 it suffices to prove $S \subseteq \{1, 2\}$ and $\{1, 2\} \subseteq S$.

($\subseteq$) Let $x \in S$. Then x is real and $x^2 - 3x + 2 = 0$. Factoring, $(x - 1)(x - 2) = 0$. A product of two real numbers is zero only if one factor is zero, so $x = 1$ or $x = 2$. Hence $x \in \{1, 2\}$.

($\supseteq$) Let $x \in \{1, 2\}$, so $x = 1$ or $x = 2$. Both are real. If $x = 1$, then $1 - 3 + 2 = 0$; if $x = 2$, then $4 - 6 + 2 = 0$. In either case $x \in S$.

This shows $S = \{1, 2\}$.

Solving an equation "at school" usually produces only the ($\subseteq$) half: "if x is a solution, then x is 1 or 2". The ($\supseteq$) half, checking that the candidates really are solutions, is the part that gets forgotten. It matters whenever a step in the solving is not reversible, such as squaring both sides.

0.3.4 Union, intersection, and difference

Given two sets, there are three basic ways to form a new one: keep what is in either, keep what is in both, or keep what is in the first but not the second. Each is a set-builder description whose predicate uses a connective from the section on logic.

Definition 0.3.6. Union

The **union** of sets A and B is

$$A \cup B = \{x : x \in A \text{ or } x \in B\}.$$

Definition 0.3.7. Intersection

The **intersection** of sets A and B is

$$A \cap B = \{x : x \in A \text{ and } x \in B\}.$$

We call A and B **disjoint** if $A \cap B = \varnothing$.

Definition 0.3.8. Set difference and complement

The **difference** of sets A and B is

$$A \setminus B = \{x \in A : x \notin B\}.$$

When all sets under discussion are subsets of one fixed set U, the **complement** of $B \subseteq U$ is $B^c = U \setminus B$.

In words: the union keeps an element that lies in **at least one** of the sets (the "or" is inclusive, as always); the intersection keeps an element that lies in **both**; the difference $A \setminus B$ keeps the elements of A that are **not** in B. The complement depends on the choice of U: the complement of $\mathbb{N}$ is the set of negative integers inside $\mathbb{Z}$, but a much bigger set inside $\mathbb{R}$.

Example. Operations on finite sets

Let $A = \{1, 2, 3, 4\}$ and $B = \{3, 4, 5, 6\}$. Compute $A \cup B$, $A \cap B$, $A \setminus B$, and $B \setminus A$.

Solution. Checking each element against the definitions: $A \cup B = \{1, 2, 3, 4, 5, 6\}$; $A \cap B = \{3, 4\}$; $A \setminus B = \{1, 2\}$; and $B \setminus A = \{5, 6\}$. Since $A \setminus B \neq B \setminus A$, the difference is not symmetric in its two arguments, unlike union and intersection.

Example. Operations on intervals

Let $A = [0, 2]$ and $B = [1, 3]$. Compute $A \cup B$, $A \cap B$, and $A \setminus B$.

Solution. A real x lies in $A \cup B$ when $0 \leq x \leq 2$ or $1 \leq x \leq 3$, which happens exactly when $0 \leq x \leq 3$; so $A \cup B = [0, 3]$. It lies in $A \cap B$ when both hold, that is, when $1 \leq x \leq 2$; so $A \cap B = [1, 2]$. It lies in $A \setminus B$ when $0 \leq x \leq 2$ and not $1 \leq x \leq 3$. Given $x \leq 2$, "not $1 \leq x \leq 3$" means $x < 1$; so $A \setminus B = [0, 1)$. The endpoint 1 is excluded because $1 \in B$.

Unions, intersections, and differences interact through laws that mirror those for "or", "and", and "not". The first pair is De Morgan's laws, the set version of Theorem 0.1.2.

Theorem 0.3.2. De Morgan's laws for sets

Let A, B, and C be sets. Then

(a) $A \setminus (B \cup C) = (A \setminus B) \cap (A \setminus C)$;
(b) $A \setminus (B \cap C) = (A \setminus B) \cup (A \setminus C)$.

In particular, if $B, C \subseteq U$, then $(B \cup C)^c = B^c \cap C^c$ and $(B \cap C)^c = B^c \cup C^c$.

Idea. Unwind membership into connectives. "$x \notin B \cup C$" is "not ($x \in B$ or $x \in C$)", and De Morgan for statements turns that into "$x \notin B$ and $x \notin C$". So each law is its logical counterpart, applied one element at a time. We prove (a) by double inclusion; (b) is an exercise.

Proof. We prove (a) using Theorem 0.3.1.

($\subseteq$) Let $x \in A \setminus (B \cup C)$. By Definition 0.3.8, $x \in A$ and $x \notin B \cup C$. By Definition 0.3.6, the second says that "$x \in B$ or $x \in C$" is false, so by Theorem 0.1.2, $x \notin B$ and $x \notin C$. Since $x \in A$ and $x \notin B$, we have $x \in A \setminus B$; since $x \in A$ and $x \notin C$, we have $x \in A \setminus C$. Hence $x \in (A \setminus B) \cap (A \setminus C)$ by Definition 0.3.7.

($\supseteq$) Let $x \in (A \setminus B) \cap (A \setminus C)$. By Definition 0.3.7 and Definition 0.3.8, $x \in A$, $x \notin B$, and $x \notin C$. By Theorem 0.1.2, "$x \in B$ or $x \in C$" is false, so $x \notin B \cup C$. Together with $x \in A$, this gives $x \in A \setminus (B \cup C)$.

This proves (a). For the complement form, take $A = U$ and use Definition 0.3.8: $(B \cup C)^c =$

$U \setminus (B \cup C) = (U \setminus B) \cap (U \setminus C) = B^c \cap C^c$, and similarly for (b).

■

The second pair of laws says each of $\cap$ and $\cup$ distributes over the other, like multiplication over addition, except that here **both** directions hold.

Theorem 0.3.3. Distributive laws for sets

Let A, B, and C be sets. Then

(a) $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$;
(b) $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$.

Idea. Double inclusion again, and in each half the "or" in a union forces a split into cases. For (b) ($\supseteq$), an element of both $A \cup B$ and $A \cup C$ is either in A, and then done, or not in A, and then it must be in B and in C. We prove (b); (a) is an exercise.

Proof. We prove (b) using Theorem 0.3.1.

($\subseteq$) Let $x \in A \cup (B \cap C)$. By Definition 0.3.6, $x \in A$ or $x \in B \cap C$.

Case 1: $x \in A$. Then $x \in A \cup B$ and $x \in A \cup C$ by Definition 0.3.6.

Case 2: $x \in B \cap C$. Then $x \in B$ and $x \in C$ by Definition 0.3.7, so again $x \in A \cup B$ and $x \in A \cup C$.

In both cases $x \in (A \cup B) \cap (A \cup C)$ by Definition 0.3.7.

($\supseteq$) Let $x \in (A \cup B) \cap (A \cup C)$, so $x \in A \cup B$ and $x \in A \cup C$.

Case 1: $x \in A$. Then $x \in A \cup (B \cap C)$.

Case 2: $x \notin A$. Since $x \in A \cup B$ and $x \notin A$, we get $x \in B$. Since $x \in A \cup C$ and $x \notin A$, we get $x \in C$. Hence $x \in B \cap C$, so $x \in A \cup (B \cap C)$.

In both cases $x \in A \cup (B \cap C)$. This proves (b).

■

Remark. In Chapter 1 we will see that the intersection of two subspaces is always a subspace, while their union usually is not: the two axes in the plane are subspaces, but $(1,0) + (0,1) = (1,1)$ lies on neither. The union is replaced there by a different construction, the sum of subspaces.

0.3.5 Indexed families of sets

Two sets are not always enough. In Chapter 1 we will need "the intersection of **all** subspaces that contain a given set", and there are usually infinitely many of them. So we need unions and

intersections of any number of sets at once.

An indexed family is a list of sets with one set for each label; its union keeps what lies in some member, its intersection what lies in every member.

Definition 0.3.9. Unions and intersections of indexed families

Let I be a set, and for each $i \in I$ let A_i be a set. The collection $(A_i)_{i\in I}$ is an **indexed family** of sets, with **index set** I. Its **union** is

$$\bigcup_{i\in I} A_i = \{x : \exists i \in I : x \in A_i\},$$

and, if I is **nonempty**, its **intersection** is

$$\bigcap_{i\in I} A_i = \{x : \forall i \in I : x \in A_i\}.$$

In words: x lies in the union when it lies in **at least one** A_i, and in the intersection when it lies in **every** A_i. When $I = \{1, 2\}$, these are $A_1 \cup A_2$ and $A_1 \cap A_2$. When $I = \{1, 2, 3, \dots\}$, we also write $\bigcup_{n=1}^{\infty} A_n$ and $\bigcap_{n=1}^{\infty} A_n$.

Why I must be nonempty for the intersection. If $I = \varnothing$, the condition "$\forall i \in I : x \in A_i$" is vacuously true for **every** object x, so the "intersection" would contain everything, which is not a set we want. (When all sets live inside a fixed U, one sometimes defines the empty intersection to be U.) The empty union causes no such trouble: "$\exists i \in \varnothing$" is always false, so $\bigcup_{i\in\varnothing} A_i = \varnothing$.

Example. Indexed unions and intersections of intervals

For each integer $n \geq 1$, let $A_n = [1/n, 1]$ and $B_n = (0, 1/n)$. Show that

$$\bigcup_{n=1}^{\infty} A_n = (0, 1] \qquad \text{and} \qquad \bigcap_{n=1}^{\infty} B_n = \varnothing.$$

You may use the fact that for every real $x > 0$ there is an integer $n \geq 1$ with $1/n < x$.

Solution. The union. ($\subseteq$) Let $x \in \bigcup_n A_n$. By Definition 0.3.9, $x \in A_n$ for some $n \geq 1$, so $0 < 1/n \leq x \leq 1$. Hence $x \in (0, 1]$. ($\supseteq$) Let $x \in (0, 1]$. Since $x > 0$, there is an integer $n \geq 1$ with $1/n < x$. Then $1/n \leq x \leq 1$, so $x \in A_n$, and hence x lies in the union. By Theorem 0.3.1, the union is $(0, 1]$.

The intersection. Suppose, for a contradiction, that some x lies in $\bigcap_n B_n$. Then $x \in B_1 = (0, 1)$, so $x > 0$. Choose an integer $n \geq 1$ with $1/n < x$. Then $x \notin (0, 1/n) = B_n$, contradicting $x \in B_n$ for every n. Hence the intersection has no elements, so it equals $\varnothing$.

Each B_n is nonempty (it contains $1/(2n)$), and each is contained in the one before. Yet their intersection is empty. Infinite intersections can shrink to nothing even when no single step does.

De Morgan's laws hold for indexed families with the same proofs: for a nonempty index set I, $A \setminus \bigcup_i B_i = \bigcap_i (A \setminus B_i)$ and $A \setminus \bigcap_i B_i = \bigcup_i (A \setminus B_i)$. The "or" of the finite version becomes "there exists", and the "and" becomes "for every", exactly as in Definition 0.1.4.

0.3.6 Cartesian products and n-tuples

A set forgets order: $\{1,2\} = \{2,1\}$. But a point in the plane is not like that. The point $(1,2)$ is different from $(2,1)$. We need objects that remember order.

An **ordered pair** (a,b) has a first entry a and a second entry b, and two ordered pairs are equal exactly when their entries agree in order:

$$(a,b) = (a',b') \iff a = a' \text{ and } b = b'.$$

Definition 0.3.10. Cartesian product

Let A and B be sets. The **Cartesian product** of A and B is the set of ordered pairs

$$A \times B = \{(a,b) : a \in A,\ b \in B\}.$$

More generally, for sets $A_1, \dots, A_n$, the product $A_1 \times \cdots \times A_n$ is the set of all ordered lists $(a_1, \dots, a_n)$ with $a_i \in A_i$ for each i.

For example, $\{1,2\} \times \{x,y\} = \{(1,x),(1,y),(2,x),(2,y)\}$. The order of the factors matters: $(1,x) \in \{1,2\} \times \{x,y\}$, but $(1,x) \notin \{x,y\} \times \{1,2\}$, whose pairs have a letter first. A degenerate case: $A \times \varnothing = \varnothing$ for every set A, since no pair can have a second entry from $\varnothing$.

Definition 0.3.11. n-tuples

Let A be a set and $n \geq 1$ an integer. An **n-tuple** of elements of A is an ordered list $(a_1, \dots, a_n)$ with each $a_i \in A$. The set of all of them is

$$A^n = \underbrace{A \times A \times \cdots \times A}_{n \text{ factors}}.$$

Two n-tuples are equal exactly when $a_i = b_i$ for **every** $i = 1, \dots, n$.

Example. Sets of tuples

Describe some elements of $\mathbb{R}^2$, $\mathbb{Z}^3$, and $\{0,1\}^3$, and count the elements of $\{0,1\}^3$.

Solution. The set $\mathbb{R}^2$ is the plane: it contains $(0,0)$, $(1,-2)$, and $(\sqrt{2},\pi)$. The set $\mathbb{Z}^3$ contains $(1,0,-4)$ but not $(1,\frac{1}{2},0)$, whose second entry is not an integer. The set $\{0,1\}^3$ consists of the lists of three bits:

$$\{0,1\}^3 = \{(0,0,0),(0,0,1),(0,1,0),(0,1,1),(1,0,0),(1,0,1),(1,1,0),(1,1,1)\}.$$

There are two choices for each of the three entries, giving $2 \cdot 2 \cdot 2 = 8$ elements. In the same way, if A has k elements then A^n has k^n elements.

In Chapter 1, $\mathbb{R}^n$ and more generally F^n become the first examples of vector spaces, with n-tuples added and scaled entry by entry.

0.3.7 Power sets

Sets can themselves be elements of sets, as the warning about $\{\varnothing\}$ showed. One such set collects every subset of a given set.

Definition 0.3.12. Power set

The **power set** of a set S is the set of all subsets of S:

$$\mathcal{P}(S) = \{X : X \subseteq S\}.$$

So $X \in \mathcal{P}(S)$ means exactly $X \subseteq S$. For example,

$$\mathcal{P}(\{1,2\}) = \{\varnothing, \{1\}, \{2\}, \{1,2\}\},$$

which has 4 elements. The degenerate case is instructive: $\mathcal{P}(\varnothing) = \{\varnothing\}$ has **one** element, because the empty set has exactly one subset, itself. For a set with n elements the power set has 2^n elements; the last exercise proves this by induction.

0.3.8 Exercises

0.3.8.1 A. Check your understanding

Exercise. A1

Decide whether each statement is true or false. Give a one-line reason.

(a) $0 \in \mathbb{N}$.
(b) $\varnothing \in \varnothing$.
(c) $\varnothing \subseteq \varnothing$.
(d) $\{0\} = \varnothing$.
(e) $\{1,2\} = \{2,1,1\}$ and $(1,2) = (2,1)$.
(f) $\{1\} \in \mathcal{P}(\{1,2\})$ and $\{1\} \subseteq \mathcal{P}(\{1,2\})$.

Solution.

(a) True. By Definition 0.3.1, $\mathbb{N}$ includes 0 in this book.
(b) False. By Definition 0.3.3, nothing is an element of $\varnothing$, not even $\varnothing$.
(c) True. The empty set is a subset of every set, vacuously, as in Example (Subsets).
(d) False. $0 \in \{0\}$ but $0 \notin \varnothing$, so the two sets do not have the same elements (Definition 0.3.5).
(e) The first is true: both sets have exactly the elements 1 and 2. The second is false: ordered pairs are equal only when their first entries agree, and $1 \neq 2$.
(f) The first is true: $\{1\} \subseteq \{1,2\}$, so $\{1\} \in \mathcal{P}(\{1,2\})$ by Definition 0.3.12. The second is false: $1 \in \{1\}$, but 1 is a number, not a subset of $\{1,2\}$, so $1 \notin \mathcal{P}(\{1,2\})$.

0.3.8.2 B. Practice

Exercise. B1: The second De Morgan law

Let A, B, C be sets. Prove that $A \setminus (B \cap C) = (A \setminus B) \cup (A \setminus C)$, which is part (b) of Theorem 0.3.2.

Solution. We use Theorem 0.3.1.

($\subseteq$) Let $x \in A \setminus (B \cap C)$. By Definition 0.3.8, $x \in A$ and $x \notin B \cap C$. By Definition 0.3.7, "$x \in B$ and $x \in C$" is false, so by Theorem 0.1.2, $x \notin B$ or $x \notin C$. If $x \notin B$, then $x \in A \setminus B$; if $x \notin C$, then $x \in A \setminus C$. In either case $x \in (A \setminus B) \cup (A \setminus C)$ by Definition 0.3.6.

($\supseteq$) Let $x \in (A \setminus B) \cup (A \setminus C)$. By Definition 0.3.6, $x \in A \setminus B$ or $x \in A \setminus C$. In the first case $x \in A$ and $x \notin B$; in the second, $x \in A$ and $x \notin C$. In both cases $x \in A$, and "$x \notin B$ or $x \notin C$" is true, so by Theorem 0.1.2 "$x \in B$ and $x \in C$" is false, that is, $x \notin B \cap C$. Hence $x \in A \setminus (B \cap C)$.

This proves the equality.

Exercise. B2: A distributive law

Let A, B, C be sets. Prove that $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$, which is part (a) of Theorem 0.3.3.

Solution. We use Theorem 0.3.1.

($\subseteq$) Let $x \in A \cap (B \cup C)$. Then $x \in A$ and $x \in B \cup C$, so $x \in B$ or $x \in C$. If $x \in B$, then $x \in A \cap B$; if $x \in C$, then $x \in A \cap C$. In either case $x \in (A \cap B) \cup (A \cap C)$.

($\supseteq$) Let $x \in (A \cap B) \cup (A \cap C)$. If $x \in A \cap B$, then $x \in A$ and $x \in B \subseteq B \cup C$. If $x \in A \cap C$, then $x \in A$ and $x \in C \subseteq B \cup C$. In either case $x \in A$ and $x \in B \cup C$, so $x \in A \cap (B \cup C)$.

This proves the equality.

Exercise. B3: An indexed intersection

Prove that

$$\bigcap_{n=1}^{\infty} \left(-\frac{1}{n}, \frac{1}{n}\right) = \{0\}.$$

You may use the fact that for every real $x > 0$ there is an integer $n \geq 1$ with $1/n < x$.

Solution. Write $B_n = (-1/n, 1/n)$ and $B = \bigcap_{n \geq 1} B_n$. We use Theorem 0.3.1.

($\supseteq$) For every $n \geq 1$, $-1/n < 0 < 1/n$, so $0 \in B_n$. By Definition 0.3.9, $0 \in B$.

($\subseteq$) Let $x \in B$. Suppose, for a contradiction, that $x \neq 0$. Then $|x| > 0$, so there is an integer $n \geq 1$ with $1/n < |x|$. Then $x \geq 1/n$ or $x \leq -1/n$, so $x \notin B_n$. This contradicts $x \in B$,

which requires $x \in B_n$ for **every** n. Hence $x = 0$, that is, $x \in \{0\}$.

This shows $B = \{0\}$.

0.3.8.3 C. Going deeper

Exercise. C1: When union equals intersection

Let A and B be sets. Prove that $A \cup B = A \cap B$ if and only if $A = B$.

Solution. ($\Leftarrow$) Suppose $A = B$. Then $A \cup B = A \cup A$ and $A \cap B = A \cap A$. By Definition 0.3.6 and Definition 0.3.7, "$x \in A$ or $x \in A$" and "$x \in A$ and $x \in A$" are both equivalent to $x \in A$, so $A \cup A = A = A \cap A$. Hence $A \cup B = A \cap B$.

($\Rightarrow$) Suppose $A \cup B = A \cap B$. We show $A = B$ using Theorem 0.3.1.

($\subseteq$) Let $x \in A$. Then $x \in A \cup B$ by Definition 0.3.6. Since $A \cup B = A \cap B$, we get $x \in A \cap B$, and in particular $x \in B$.

($\supseteq$) Swapping the roles of A and B in the previous paragraph, which is allowed since $A \cup B = B \cup A$ and $A \cap B = B \cap A$, gives $B \subseteq A$.

Hence $A = B$. This proves the equivalence.

Exercise. C2: Counting subsets

For a finite set S, write $|S|$ for its number of elements. The aim is to prove that $|\mathcal{P}(S)| = 2^{|S|}$ for every finite set S.

(a) Check the claim when $S = \varnothing$.

(b) Let S have $n + 1$ elements, fix $s \in S$, and let $S' = S \setminus \{s\}$. Show that every subset X of S belongs to exactly one of the two collections: the subsets of S with $s \notin X$, and the subsets of S with $s \in X$.

(c) Show that the subsets of S with $s \notin X$ are exactly the subsets of S', and that $X \mapsto X \setminus \{s\}$ and $T \mapsto T \cup \{s\}$ pair off the subsets of S containing s with the subsets of S', each undoing the other.

(d) Hence prove the claim by induction on $n = |S|$.

Hint: in (c), use double inclusion to check that $(X \setminus \{s\}) \cup \{s\} = X$ *when* $s \in X$.

Solution.

(a) The only subset of $\varnothing$ is $\varnothing$ itself, since any $X \subseteq \varnothing$ has no elements and so equals $\varnothing$ by Theorem 0.3.1. So $\mathcal{P}(\varnothing) = \{\varnothing\}$ has $1 = 2^0$ element.

(b) Let $X \subseteq S$. The statement $s \in X$ is either true or false, and not both. So X lies in exactly one of the two collections.

(c) *First collection.* If $X \subseteq S$ and $s \notin X$, then every $x \in X$ lies in S and is not s, so $x \in S'$; hence $X \subseteq S'$. Conversely, if $T \subseteq S'$, then $T \subseteq S$ and $s \notin T$, since no element of S' equals s. So the first collection is exactly $\mathcal{P}(S')$.

Second collection. If $X \subseteq S$ with $s \in X$, then $X \setminus \{s\} \subseteq S'$. If $T \subseteq S'$, then $T \cup \{s\} \subseteq S$ contains s. Now we check that the two operations undo each other. For X containing s: an element lies in $(X \setminus \{s\}) \cup \{s\}$ exactly when it is in X and not s, or it is s; since $s \in X$, this is exactly when it is in X. So $(X \setminus \{s\}) \cup \{s\} = X$. For $T \subseteq S'$: an element lies in $(T \cup \{s\}) \setminus \{s\}$ exactly when it is in T or equal to s, and not equal to s; since $s \notin T$, this is exactly when it is in T. So $(T \cup \{s\}) \setminus \{s\} = T$. Hence the two operations pair the subsets of S containing s one-to-one with the subsets of S', and the two collections have the same number of elements.

(d) Let $P(n)$ be the statement "every set with n elements has exactly 2^n subsets". We use Theorem 0.2.7 with $n_0 = 0$. The base case $P(0)$ is (a), since the only set with no elements is $\varnothing$. Let $n \geq 0$, suppose $P(n)$, and let S have $n+1$ elements. Fix $s \in S$, which exists since $n+1 \geq 1$, and let $S' = S \setminus \{s\}$, a set with n elements. By (b), the number of subsets of S is the size of the first collection plus the size of the second. By (c) and the induction hypothesis, each of these sizes equals $|\mathcal{P}(S')| = 2^n$. Hence S has $2^n + 2^n = 2^{n+1}$ subsets, which is $P(n+1)$. By Theorem 0.2.7, $P(n)$ holds for every $n \in \mathbb{N}$, as claimed.

0.4 Functions

Sets are the objects of this chapter. The next question is how to move between them, and the tool is a function. Almost everything in linear algebra is a question about a function: a linear map is a function, a matrix will turn out to encode one, and "does $\mathbf{Ax} = \mathbf{b}$ have a solution, and is it unique?" is a question about whether a function is surjective and injective. This section fixes the language of functions, composition, and inverses, and proves the handful of facts about them that the rest of the book uses without comment.

0.4.1 What a function is

At school a function is usually a formula, such as $x^2 + 1$. A formula alone is not enough. The same formula x^2 behaves differently depending on which numbers we feed it and where we allow the answers to land, and whether an equation $x^2 = c$ can be solved depends on exactly that information. So a function carries three pieces of data, not one.

A function is a machine with a fixed input set and a fixed target set that turns each input into exactly one output.

Definition 0.4.1. Function

Let X and Y be sets. A **function** (or **map**) $f\colon X \to Y$ is a rule that assigns to **every** element $x \in X$ **exactly one** element $f(x) \in Y$. The set X is the **domain** of f, the set Y is its **codomain**, and $f(x)$ is the **value** of f at x. We also write $x \mapsto f(x)$.

Two functions $f\colon X \to Y$ and $g\colon X' \to Y'$ are **equal** if $X = X'$, $Y = Y'$, and $f(x) = g(x)$ for every $x \in X$.

In words: the first clause says every input gets an output, so nothing in the domain is left out. The second says the output is unique, so the rule never hesitates between two values. The codomain is where outputs are *allowed* to land; nothing says every element of Y must be hit.

If you want a definition that avoids the vague word "rule", use the Cartesian product (Defini-

tion 0.3.10): a function $X \to Y$ is a subset $G \subseteq X \times Y$ such that for every $x \in X$ there is exactly one $y \in Y$ with $(x, y) \in G$. The set G is the **graph** of f, and $f(x)$ is that unique y. Nothing in this book depends on which version you prefer.

Example. First examples

Check each of the following against the two clauses of Definition 0.4.1.

(a) $f\colon \mathbb{R} \to \mathbb{R}$, $f(x) = 3x - 5$. Every real x gives one real number $3x - 5$.

(b) $s\colon \mathbb{R}^2 \to \mathbb{R}$, $s(x, y) = x + y$. The domain is a Cartesian product; each pair gives one sum.

(c) $\ell\colon \{\text{finite strings of the letters } a, b\} \to \mathbb{N}$, sending a string to its length. The empty string has length 0, which is why it matters that $\mathbb{N}$ contains 0.

(d) For any set X, the rule $x \mapsto x$ is a function $X \to X$. It looks too simple to matter, but it will play the role that the number 1 plays for multiplication.

(e) For any set Y, there is exactly one function $\varnothing \to Y$, the **empty function**. Both clauses are about elements of the domain, and there are none, so they hold vacuously. It is a degenerate case that makes statements like "for every set X" true without exceptions.

The failures are just as instructive. The rule "$x \mapsto$ the real number y with $y^2 = x$" is **not** a function $\mathbb{R} \to \mathbb{R}$. At $x = -1$ there is no such y, so the first clause fails. At $x = 4$ there are two, $y = 2$ and $y = -2$, so the second clause fails. Repairing both problems gives a genuine function $[0, \infty) \to \mathbb{R}$, $x \mapsto \sqrt{x}$, where $\sqrt{x}$ means the nonnegative root.

The equality clause says the codomain is part of the function. The functions $\mathbb{R} \to \mathbb{R}$, $x \mapsto x^2$ and $\mathbb{R} \to [0, \infty)$, $x \mapsto x^2$ have the same values but are **different** functions. This is not pedantry: we will see in a moment that one of them is surjective and the other is not.

Quick check. Let $f, g\colon \mathbb{R} \to \mathbb{R}$ be $f(x) = |x|$ and $g(x) = \sqrt{x^2}$. Are f and g equal?

Solution. Yes. They have the same domain and the same codomain, and for every real x we have $\sqrt{x^2} = |x|$, because $|x|$ is the nonnegative number whose square is x^2. Different formulas can describe the same function; what counts is the values.

0.4.2 Images and preimages

Often we want to push a whole set through a function, or ask which inputs land in a given region. Both operations get a name.

Definition 0.4.2. Image and preimage

Let $f\colon X \to Y$ be a function, $A \subseteq X$, and $B \subseteq Y$.

- The **image** of A under f is $f(A) \coloneqq \{f(a) : a \in A\} \subseteq Y$. The set $f(X)$ is called the **image of** f.
- The **preimage** of B under f is $f^{-1}(B) \coloneqq \{x \in X : f(x) \in B\} \subseteq X$.

In words: $f(A)$ collects the outputs of the inputs in A. The preimage $f^{-1}(B)$ collects every input whose output lies in B. So $y \in f(A)$ means "$y = f(a)$ for **some** $a \in A$", while $x \in f^{-1}(B)$ means "$f(x) \in B$", with no existential quantifier at all.

Example. Images and preimages of the squaring map

Let $q\colon \mathbb{R} \to \mathbb{R}$, $q(x) = x^2$. Compute $q([-1,2])$, $q^{-1}([1,4])$, and $q^{-1}([-5,-1])$.

Solution. As x runs over $[-1,2]$, the square x^2 takes every value in $[0,4]$: values in $[0,1]$ come from $[0,1]$, values in $[1,4]$ from $[1,2]$, and no square is negative or exceeds $2^2 = 4$ on this interval. Hence $q([-1,2]) = [0,4]$.

Next, $x^2 \in [1,4]$ means $1 \le |x| \le 2$. Hence $q^{-1}([1,4]) = [-2,-1] \cup [1,2]$.

Finally, no real square lies in $[-5,-1]$, so $q^{-1}([-5,-1]) = \varnothing$. A preimage may be empty; that is not an error.

Warning. The symbol $f^{-1}(B)$ is defined for **every** function, invertible or not; it does not mean that an inverse function exists. For the same reason, pushing forward and pulling back need not undo each other. With $q(x) = x^2$ on $\mathbb{R}$: for $B = [-1,4]$ we get $q^{-1}(B) = [-2,2]$ and $q(q^{-1}(B)) = [0,4] \neq B$; for $A = [0,1]$ we get $q(A) = [0,1]$ and $q^{-1}(q(A)) = [-1,1] \neq A$.

What does survive is one inclusion in each direction.

Proposition 0.4.1. Image and preimage inclusions

Let $f\colon X \to Y$, $A \subseteq X$, and $B \subseteq Y$. Then

$$f(f^{-1}(B)) \subseteq B \qquad \text{and} \qquad A \subseteq f^{-1}(f(A)).$$

Proof. Let $y \in f(f^{-1}(B))$. By Definition 0.4.2, $y = f(x)$ for some $x \in f^{-1}(B)$, and $x \in f^{-1}(B)$ means $f(x) \in B$. Hence $y \in B$.

Let $a \in A$. Then $f(a) \in f(A)$ by definition of the image, so $a \in f^{-1}(f(A))$ by definition of the preimage. This proves both inclusions. ∎

The warning shows both inclusions can be strict. In Exercise (C3: When images and preimages undo each other) you will find exactly which property of f turns each one into an equality.

0.4.3 Injective, surjective, bijective

Given $f\colon X \to Y$ and a target $y \in Y$, the most basic question is how many solutions the equation $f(x) = y$ has. There are two separate issues: whether there is **at least one** solution, and whether there is **at most one**. In linear algebra these become "does $\mathbf{Ax} = \mathbf{b}$ have a solution?" and "is the solution unique?", and they are answered by different tools. So the two

halves deserve separate names.

Injective means no two inputs collide; surjective means every target is hit; bijective means both, so every equation $f(x) = y$ has exactly one solution.

Definition 0.4.3. Injective, surjective, bijective

Let $f\colon X \to Y$ be a function.

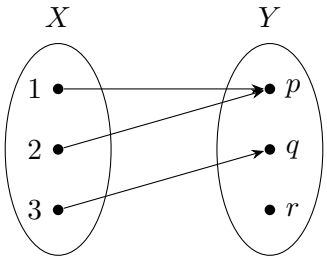
- f is **injective** (one-to-one) if **for all** $x_1, x_2 \in X$, $f(x_1) = f(x_2)$ implies $x_1 = x_2$.
- f is **surjective** (onto) if **for every** $y \in Y$ **there exists** $x \in X$ with $f(x) = y$. Equivalently, $f(X) = Y$.
- f is **bijective** if it is both injective and surjective.

In words: injectivity says that if two inputs give the same output, they were the same input. Equivalently, by the contrapositive (Theorem 0.1.1), distinct inputs give distinct outputs. Surjectivity says the image of f fills the whole codomain. The quantifiers matter: in surjectivity the x is allowed to depend on y.

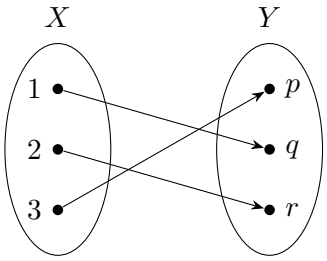
Negating these, as in Example (What "not injective" means) and Example (What "not surjective" means), tells us what a counterexample looks like. The map f is **not injective** when there exist $x_1 \neq x_2$ with $f(x_1) = f(x_2)$: one collision is enough. It is **not surjective** when there exists $y \in Y$ that equals $f(x)$ for no $x \in X$: one missed target is enough.

The standard moves. To prove f injective, write "Suppose $f(x_1) = f(x_2)$" and work towards $x_1 = x_2$. To prove f surjective, write "Let $y \in Y$" and then produce an x, usually found backwards by solving $f(x) = y$ on scrap paper, and check forwards that $f(x) = y$.

The picture below shows maps between three-element sets, drawn as arrows. A collision is two arrows landing on one point; a missed target is a point with no arrow.



collision at p, nothing reaches r: neither injective nor surjective



exactly one arrow into each point: bijective

Example. Deciding injectivity and surjectivity

Determine which of the following functions are injective, surjective, or bijective.

(a) $f\colon \mathbb{R} \to \mathbb{R}$, $f(x) = 3x - 5$.

(b) $g\colon \mathbb{Z} \to \mathbb{Z}$, $g(n) = 2n$.

(c) $s\colon \mathbb{R}^2 \to \mathbb{R}$, $s(x, y) = x + y$.

(d) $\mathrm{id}_X\colon X \to X$, $x \mapsto x$, for any set X.

Solution.

(a) Bijective. Suppose $f(x_1) = f(x_2)$. Then $3x_1 - 5 = 3x_2 - 5$, so $3x_1 = 3x_2$, and dividing by $3 \neq 0$ gives $x_1 = x_2$. Hence f is injective. Let $y \in \mathbb{R}$. Solving $3x - 5 = y$ suggests $x = (y + 5)/3$, and indeed $f((y + 5)/3) = (y + 5) - 5 = y$. Hence f is surjective.

(b) Injective, not surjective. Suppose $2n_1 = 2n_2$; then $n_1 = n_2$. But 1 is not hit: $2n = 1$ has no integer solution, since $2n$ is even. Note that the same formula $x \mapsto 2x$ on $\mathbb{R} \to \mathbb{R}$ *is* surjective; the failure comes from the domain and codomain, not the formula.

(c) Surjective, not injective. Let $t \in \mathbb{R}$; then $s(t, 0) = t$, so s is surjective. But $s(1, 0) = 1 = s(0, 1)$ while $(1, 0) \neq (0, 1)$, a collision, so s is not injective.

(d) Bijective. If $\mathrm{id}_X(x_1) = \mathrm{id}_X(x_2)$ then $x_1 = x_2$ directly, and every $x \in X$ is hit by itself. This degenerate case is the model of a bijection and reappears as the identity function below.

The cleanest way to see that domain and codomain are part of the question is to keep the formula $x \mapsto x^2$ fixed and change only the sets. Write $[0, \infty)$ for the nonnegative reals.

Function	Injective?	Surjective?
$\mathbb{R} \to \mathbb{R}$	no: $(-1)^2 = 1^2$	no: -1 is not a square
$\mathbb{R} \to [0, \infty)$	no: $(-1)^2 = 1^2$	yes: $y = (\sqrt{y})^2$
$[0, \infty) \to \mathbb{R}$	yes	no: -1 is not a square
$[0, \infty) \to [0, \infty)$	yes	yes

For the injective rows: if $x_1, x_2 \geq 0$ and $x_1^2 = x_2^2$, then $(x_1 - x_2)(x_1 + x_2) = 0$. Either $x_1 = x_2$, or $x_1 + x_2 = 0$, which for nonnegative numbers forces $x_1 = x_2 = 0$. In both cases $x_1 = x_2$.

Read the table as a set of minimal changes. Shrinking the domain from $\mathbb{R}$ to $[0, \infty)$ removes the collision between -1 and 1, so the injectivity clause now holds. Shrinking the codomain removes the missed targets, so the surjectivity clause now holds. The formula never changed.

Warning. Do not confuse injectivity with the statement "$x_1 = x_2 \Rightarrow f(x_1) = f(x_2)$". That implication is true for **every** function; it is the second clause of Definition 0.4.1. Injectivity is its converse, "$f(x_1) = f(x_2) \Rightarrow x_1 = x_2$", and $x \mapsto x^2$ on $\mathbb{R}$ satisfies the first but not the second.

Why these definitions. Surjectivity depends on the codomain, and injectivity on the domain, so neither is a property of a formula alone; this is why Definition 0.4.1 insists on both sets. The names are descriptive: "in-jective" throws the domain *into* the codomain without collisions, and "sur-jective" (French *sur*, onto) lands *onto* all of it. In Chapter 2 we will meet a fast test for linear maps: a linear map is injective exactly when the only input sent to 0 is 0. That test is special to linear maps and fails for general functions, which is one reason the general definitions come first.

0.4.4 Composition

If the outputs of one function can be fed into another, we can run them in sequence. This is the operation that matrix multiplication will describe in Chapter 4.

Definition 0.4.4. Composition

Let $f\colon X \to Y$ and $g\colon Y \to Z$ be functions. The **composition** $g \circ f\colon X \to Z$ is defined by

$$(g \circ f)(x) \coloneqq g\big(f(x)\big) \qquad \text{for every } x \in X.$$

Read $g \circ f$ as "g after f": the function written on the **right** acts **first**. The definition needs the codomain of f to equal the domain of g, so that $g(f(x))$ makes sense.

Example. Order matters

(a) Let $f, g\colon \mathbb{R} \to \mathbb{R}$ be $f(x) = x + 3$ and $g(x) = 2x$. Compute $g \circ f$ and $f \circ g$.

(b) Let $q\colon \mathbb{R} \to [0, \infty)$, $q(x) = x^2$, and $r\colon [0, \infty) \to \mathbb{R}$, $r(y) = \sqrt{y}$. Compute $r \circ q$ and $q \circ r$.

Solution.

(a) $(g \circ f)(x) = g(x + 3) = 2x + 6$, while $(f \circ g)(x) = f(2x) = 2x + 3$. At $x = 0$ these are $6 \neq 3$, so $g \circ f \neq f \circ g$. Composition is not commutative, even when both orders make sense.

(b) $r \circ q\colon \mathbb{R} \to \mathbb{R}$ is $x \mapsto \sqrt{x^2} = |x|$. In the other order, $q \circ r\colon [0, \infty) \to [0, \infty)$ is $y \mapsto (\sqrt{y})^2 = y$. So $q \circ r$ does nothing, while $r \circ q$ forgets the sign of x. The two orders do not even have the same domain. We will return to this pair when we discuss one-sided inverses.

Although the order matters, the grouping does not. This is what lets us write $h \circ g \circ f$ without brackets.

Theorem 0.4.2. Associativity of composition

Let $f\colon X \to Y$, $g\colon Y \to Z$, and $h\colon Z \to W$ be functions. Then $(h \circ g) \circ f = h \circ (g \circ f)$.

Proof. Both sides are functions $X \to W$. Let $x \in X$. By Definition 0.4.4 applied twice on each side,

$$\big((h \circ g) \circ f\big)(x) = (h \circ g)\big(f(x)\big) = h\big(g(f(x))\big) = h\big((g \circ f)(x)\big) = \big(h \circ (g \circ f)\big)(x).$$

Since the values agree at every x, the two functions are equal by Definition 0.4.1.

■

Next, the properties of the previous subsection pass through composition.

Theorem 0.4.3. Composition preserves injectivity and surjectivity

Let $f\colon X \to Y$ and $g\colon Y \to Z$ be functions.

(a) If f and g are injective, then $g \circ f$ is injective.

(b) If f and g are surjective, then $g \circ f$ is surjective.

(c) If f and g are bijective, then $g \circ f$ is bijective.

Idea. For (a), start from a collision of $g \circ f$ and peel off one function at a time, outermost first: injectivity of g removes g, then injectivity of f removes f. For (b), run backwards: reach z from some y using g, then reach that y from some x using f.

Proof.

(a) Suppose $(g \circ f)(x_1) = (g \circ f)(x_2)$, that is, $g(f(x_1)) = g(f(x_2))$. Since g is injective, $f(x_1) = f(x_2)$. Since f is injective, $x_1 = x_2$. Hence $g \circ f$ is injective.

(b) Let $z \in Z$. Since g is surjective, there is $y \in Y$ with $g(y) = z$. Since f is surjective, there is $x \in X$ with $f(x) = y$. Then $(g \circ f)(x) = g(y) = z$. Hence $g \circ f$ is surjective.

(c) If f and g are bijective, then $g \circ f$ is injective by (a) and surjective by (b), so it is bijective.

■

The converse of (a) is false, but half of it survives, and that half is surprisingly useful. Suppose $g \circ f$ is injective. A collision of f would be a collision of $g \circ f$, so f must be injective; but g is only used on the image $f(X)$, so g may still collide outside it. The surjective case is dual.

Theorem 0.4.4. Partial converse for compositions

Let $f\colon X \to Y$ and $g\colon Y \to Z$ be functions.

(a) If $g \circ f$ is injective, then f is injective.

(b) If $g \circ f$ is surjective, then g is surjective.

Proof.

(a) Suppose $f(x_1) = f(x_2)$. Applying g to both sides gives $(g \circ f)(x_1) = (g \circ f)(x_2)$. Since $g \circ f$ is injective, $x_1 = x_2$. Hence f is injective.

(b) Let $z \in Z$. Since $g \circ f$ is surjective, there is $x \in X$ with $g(f(x)) = z$. Put $y = f(x) \in Y$; then $g(y) = z$. Hence g is surjective.

■

Remark. The other halves really fail. Take $X = Z = \{1\}$, $Y = \{1, 2\}$, $f(1) = 1$, and $g(1) = g(2) = 1$. Then $g \circ f = \mathrm{id}_{\{1\}}$ is bijective, yet g is not injective ($g(1) = g(2)$) and f is not surjective (nothing maps to 2).

0.4.5 Identity and one-sided inverses

The function that does nothing deserves a name, because "undoing f" means "composing with f gives the function that does nothing".

Definition 0.4.5. Identity function

For a set X, the **identity function** on X is $\mathrm{id}_X \colon X \to X$, $\mathrm{id}_X(x) = x$.

For every $f \colon X \to Y$ we have $f \circ \mathrm{id}_X = f$ and $\mathrm{id}_Y \circ f = f$, since both sides send x to $f(x)$. Notice the subscripts: on the right of f the identity lives on the domain, on the left on the codomain.

Now look again at Example (Order matters) (b). There $q \circ r = \mathrm{id}_{[0,\infty)}$, but $r \circ q = |\cdot| \neq \mathrm{id}_{\mathbb{R}}$. So r undoes q from one side only. This one-sided behavior is common enough to need a word.

Definition 0.4.6. Left and right inverses

Let $f \colon X \to Y$. A function $g \colon Y \to X$ is a **left inverse** of f if $g \circ f = \mathrm{id}_X$, and a **right inverse** of f if $f \circ g = \mathrm{id}_Y$.

"Left" and "right" refer to where g is written in the composition. A left inverse recovers the input from the output: $g(f(x)) = x$. A right inverse picks, for each target y, an input that reaches it: $f(g(y)) = y$. In Example (Order matters) (b), r is a right inverse of q, and q is a left inverse of r. But r is **not** a left inverse of q, because $r(q(-1)) = 1 \neq -1$.

The descriptions "recovers the input" and "reaches every target" should remind you of injective and surjective. They are exactly that.

Theorem 0.4.5. Left inverses and injectivity

Let $f \colon X \to Y$ with X **nonempty**. Then f has a left inverse if and only if f is injective.

Idea. ($\Rightarrow$) is Theorem 0.4.4 with $g \circ f = \mathrm{id}_X$, which is injective. For ($\Leftarrow$) we must define $g(y)$ for every $y \in Y$. If y is in the image, injectivity gives a **unique** x with $f(x) = y$, and $g(y)$ must be that x. If y is not in the image, the condition $g \circ f = \mathrm{id}_X$ says nothing about $g(y)$, so we send y anywhere in X. That is where nonemptiness is used.

Proof. ($\Rightarrow$) Suppose $g \circ f = \mathrm{id}_X$. The identity id_X is injective, so f is injective by Theorem 0.4.4 (a).

($\Leftarrow$) Suppose f is injective. Since $X \neq \varnothing$, we may fix some $x_0 \in X$. Define $g \colon Y \to X$ as follows. If $y \in f(X)$, there is some $x \in X$ with $f(x) = y$, and this x is unique because f is injective; set $g(y) = x$. If $y \notin f(X)$, set $g(y) = x_0$. Every y receives exactly one value, so g is a function. For $x \in X$, the element $y = f(x)$ lies in $f(X)$, and the unique input sent to it is x, so $g(f(x)) = x$. Hence $g \circ f = \mathrm{id}_X$, as claimed.

■

The hypothesis $X \neq \varnothing$ is not decoration. The empty function $\varnothing \to \{1\}$ is injective (there are no two inputs to collide), but there is no function $\{1\} \to \varnothing$ at all, since 1 would have nowhere to go.

Theorem 0.4.6. Right inverses and surjectivity

Let $f\colon X \to Y$. Then f has a right inverse if and only if f is surjective.

Proof. ($\Rightarrow$) Suppose $f \circ g = \mathrm{id}_Y$. The identity id_Y is surjective, so f is surjective by Theorem 0.4.4 (b).

($\Leftarrow$) Suppose f is surjective. For each $y \in Y$ the set $f^{-1}(\{y\})$ is nonempty, so we may choose one element of it and call it $g(y)$. This defines $g\colon Y \to X$ with $f(g(y)) = y$ for every $y \in Y$, that is, $f \circ g = \mathrm{id}_Y$.

■

When Y is infinite, the step "choose one element from each of infinitely many nonempty sets at once" is an axiom of set theory, the **axiom of choice**; we use it freely and never mention it again in this chapter.

Unlike an inverse in the usual sense, one-sided inverses are rarely unique. For $q\colon \mathbb{R} \to [0, \infty)$, $x \mapsto x^2$, both $y \mapsto \sqrt{y}$ and $y \mapsto -\sqrt{y}$ are right inverses: each picks one of the two square roots.

Quick check. Give a function $f\colon \mathbb{N} \to \mathbb{N}$ that has a left inverse but no right inverse.

Solution. Take $f(n) = n + 1$. It is injective, since $n_1 + 1 = n_2 + 1$ gives $n_1 = n_2$, so by Theorem 0.4.5 it has a left inverse; explicitly, $g(m) = m - 1$ for $m \geq 1$ and $g(0) = 0$ satisfies $g(f(n)) = n$. It is not surjective, since $n + 1 = 0$ has no solution in $\mathbb{N}$. So by Theorem 0.4.6 it has no right inverse.

0.4.6 Inverses

When a function has an inverse on both sides, the two one-sided inverses turn out to coincide, and we get the notion of inverse we know from e^x and $\ln x$.

Definition 0.4.7. Inverse function

Let $f\colon X \to Y$. A function $g\colon Y \to X$ is an **inverse** of f if

$$g \circ f = \mathrm{id}_X \qquad \text{and} \qquad f \circ g = \mathrm{id}_Y .$$

If f has an inverse, f is **invertible**.

Both conditions are needed: Example (Order matters) (b) shows a pair with one but not the other. The next theorem says which functions are invertible and that "the" inverse is justified.

Theorem 0.4.7. Bijective if and only if invertible

A function $f\colon X \to Y$ is invertible if and only if it is bijective. Moreover, an inverse of f, when it exists, is **unique**; we denote it $f^{-1}\colon Y \to X$.

Idea. For ($\Leftarrow$) one could combine the two one-sided theorems, but that route needs $X \neq \varnothing$ and a choice. A bijection makes both unnecessary: every y has **exactly one** preimage, so $f^{-1}(y)$ is forced. For uniqueness, the standard move is to assume two inverses g and h, and sandwich: $g = g \circ (f \circ h) = (g \circ f) \circ h = h$.

Proof. ($\Rightarrow$) Suppose g is an inverse of f. Since $g \circ f = \mathrm{id}_X$ is injective, f is injective by Theorem 0.4.4 (a). Since $f \circ g = \mathrm{id}_Y$ is surjective, f is surjective by Theorem 0.4.4 (b). Hence f is bijective.

($\Leftarrow$) Suppose f is bijective, and let $y \in Y$. Since f is surjective, some $x \in X$ has $f(x) = y$, and since f is injective, this x is unique. Define $g(y)$ to be this x. Then $f(g(y)) = y$ for every $y \in Y$, so $f \circ g = \mathrm{id}_Y$. For $x \in X$, the unique input sent to $f(x)$ is x itself, so $g(f(x)) = x$, and $g \circ f = \mathrm{id}_X$. Hence g is an inverse of f.

(Uniqueness) Suppose g and h are both inverses of f. Then

$$g = g \circ \mathrm{id}_Y = g \circ (f \circ h) = (g \circ f) \circ h = \mathrm{id}_X \circ h = h,$$

where the third equality is Theorem 0.4.2. This proves the theorem.

■

The uniqueness computation used only $g \circ f = \mathrm{id}_X$ and $f \circ h = \mathrm{id}_Y$. So it proves more: **if f has a left inverse g and a right inverse h, then $g = h$**, and this common function is the inverse. The same sandwich will prove that a matrix has at most one inverse.

Warning. The notation f^{-1} now has two meanings. The preimage $f^{-1}(B)$ of a **set** $B \subseteq Y$ exists for every f. The inverse function f^{-1}, applied to an **element** $y \in Y$, exists only for bijective f. Writing $q^{-1}(4)$ for $q(x) = x^2$ on $\mathbb{R}$ is meaningless, while $q^{-1}(\{4\}) = \{-2, 2\}$ is fine. When f is bijective the two agree: $f^{-1}(\{y\}) = \{f^{-1}(y)\}$.

Example. Finding inverses

For each function, decide whether it is invertible, and if so find its inverse.

(a) $f\colon \mathbb{R} \to \mathbb{R}$, $f(x) = 5 - 2x$.
(b) $t\colon \mathbb{R}^2 \to \mathbb{R}^2$, $t(x, y) = (y, x)$.
(c) $g\colon \mathbb{R} \to \mathbb{R}$, $g(x) = x^2$.
(d) $h\colon [0, \infty) \to [0, \infty)$, $h(x) = x^2$.

Solution.

(a) Solving $5 - 2x = y$ gives $x = (5 - y)/2$. Put $k(y) = (5 - y)/2$. Check: $k(f(x)) = (5 - (5 - 2x))/2 = x$ and $f(k(y)) = 5 - (5 - y) = y$. So k is an inverse of f, and $f^{-1}(y) = (5 - y)/2$.

(b) $t(t(x, y)) = t(y, x) = (x, y)$, so $t \circ t = \mathrm{id}_{\mathbb{R}^2}$. Hence t is its own inverse: $t^{-1} = t$.

(c) Not invertible. $g(-1) = g(1)$, so g is not injective, hence not bijective, and Theorem 0.4.7 says it has no inverse.

(d) Invertible. By the table in the previous subsection h is bijective, and the inverse is $h^{-1}(y) = \sqrt{y}$: indeed $\sqrt{x^2} = x$ for $x \geq 0$ and $(\sqrt{y})^2 = y$ for $y \geq 0$.

To undo two steps, undo the last one first. When you dress, the shirt goes on before the jacket; to undress, the jacket comes off first.

Theorem 0.4.8. Inverse of a composition

Let $f\colon X \to Y$ and $g\colon Y \to Z$ be bijective. Then $g \circ f$ is bijective and

$$(g \circ f)^{-1} = f^{-1} \circ g^{-1}.$$

Proof. Using Theorem 0.4.2 to regroup,

$$(f^{-1} \circ g^{-1}) \circ (g \circ f) = f^{-1} \circ (g^{-1} \circ g) \circ f = f^{-1} \circ \mathrm{id}_Y \circ f = f^{-1} \circ f = \mathrm{id}_X,$$

and in the same way $(g \circ f) \circ (f^{-1} \circ g^{-1}) = g \circ (f \circ f^{-1}) \circ g^{-1} = g \circ g^{-1} = \mathrm{id}_Z$. Hence $f^{-1} \circ g^{-1}$ is an inverse of $g \circ f$. By Theorem 0.4.7, $g \circ f$ is bijective and its unique inverse is $f^{-1} \circ g^{-1}$, as claimed. ∎

The same reversal will appear for matrices as $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$, and for the same reason.

0.4.7 Maps between finite sets

For functions between two finite sets **of the same size**, one of the two conditions comes for free. This is the pigeonhole principle, and it is the ancestor of a move used constantly from Chapter 2 on: instead of checking both injectivity and surjectivity, count, and check only one.

We write $|X|$ for the number of elements of a finite set X. We use two counting facts about lists without proof: a list of n entries with a repeated entry contains at most $n - 1$ distinct elements, and a subset of a finite set Y with $|Y|$ elements is all of Y.

Theorem 0.4.9. Injective if and only if surjective for finite sets of equal size

Let X and Y be finite sets with $|X| = |Y|$, and let $f\colon X \to Y$. Then f is injective if and only if f is surjective. In that case f is bijective.

Idea. List the elements of X as $x_1, \dots, x_n$; the image $f(X)$ is then listed by $f(x_1), \dots, f(x_n)$. Injective means this list has no repeats, so it has n distinct entries and must fill Y. Surjective means the list fills Y, which has n elements, so there is no room for a repeat.

Proof. Let $n = |X| = |Y|$, and write $X = \{x_1, \dots, x_n\}$ with the x_i distinct. By Definition 0.4.2, $f(X)$ consists of the entries of the list $f(x_1), \dots, f(x_n)$.

($\Rightarrow$) Suppose f is injective. Since the x_i are distinct, the entries $f(x_1), \dots, f(x_n)$ are distinct. Hence $f(X)$ is a subset of Y with $n = |Y|$ elements, so $f(X) = Y$, and f is surjective.

($\Leftarrow$) We prove the contrapositive (Theorem 0.1.1). Suppose f is not injective. Then $f(x_i) = f(x_j)$ for some $i \neq j$, so the list $f(x_1), \dots, f(x_n)$ has a repeated entry and $f(X)$ has at most $n - 1$ elements. Hence $f(X) \neq Y$, and f is not surjective.

So if f is injective or surjective, the two directions show that it is both, that is, bijective.

■

For a finite set X, this applies to every $f\colon X \to X$: such a map is injective if and only if it is surjective. In Chapter 2 the same shape of statement returns for linear maps $V \to V$ on a finite-dimensional space, with dimension playing the role of the number of elements.

Warning. The theorem needs finiteness. The map $g\colon \mathbb{N} \to \mathbb{N}$, $g(n) = 2n$, is injective but misses every odd number, so it is not surjective, even though domain and codomain are the same set. Infinite sets have room for an injection into a proper part of themselves.

0.4.8 Exercises

0.4.8.1 A. Check your understanding

Exercise. A1

(a) State what it means for $f\colon X \to Y$ to be injective, and what it means to be surjective.

(b) True or false: the functions $\mathbb{R} \to \mathbb{R}$, $x \mapsto x^3$ and $\mathbb{R} \to \mathbb{R}$, $x \mapsto x \cdot x \cdot x$ are equal. Give a reason.

(c) True or false: if $g \circ f$ is injective, then g is injective. Give a reason.

(d) Write down the first line of a proof that a function f is injective, and the first line of a proof that it is surjective.

(e) True or false: the preimage $f^{-1}(B)$ is defined only when f is bijective.

(f) True or false: every injective function $\{1, 2, 3\} \to \{1, 2, 3\}$ is bijective. Name the result you use.

Solution.

(a) f is injective if for all $x_1, x_2 \in X$, $f(x_1) = f(x_2)$ implies $x_1 = x_2$. It is surjective if for every $y \in Y$ there exists $x \in X$ with $f(x) = y$.

(b) True. They have the same domain and codomain, and $x^3 = x \cdot x \cdot x$ for every real x. Equality of functions is about values, not formulas.

(c) False. In the remark after Theorem 0.4.4, $g \circ f = \mathrm{id}_{\{1\}}$ is injective but $g(1) = g(2)$. Only f is forced to be injective.

(d) Injective: "Suppose $x_1, x_2 \in X$ satisfy $f(x_1) = f(x_2)$." Surjective: "Let $y \in Y$."

(e) False. $f^{-1}(B) = \{x \in X : f(x) \in B\}$ makes sense for every function; see Definition 0.4.2 and the warning after it.

(f) True, by Theorem 0.4.9: domain and codomain are finite with the same number of elements, so an injective map is surjective, hence bijective.

0.4.8.2 B. Practice

Exercise. B1: Injective or surjective?

Determine whether each function is injective, and whether it is surjective. Justify your answers.

(a) $f\colon \mathbb{Z} \to \mathbb{Z}$, $f(n) = 3n + 1$.

(b) $g\colon \mathbb{R}^2 \to \mathbb{R}^2$, $g(x, y) = (x + y, x - y)$.

(c) $h\colon \mathbb{N} \to \mathbb{N}$, $h(n) = \lfloor n/2 \rfloor$, the largest integer not exceeding $n/2$.

(d) $p\colon \mathbb{R} \to \mathbb{R}^2$, $p(t) = (t, t^2)$.

Solution.

(a) Injective, not surjective. Suppose $3n_1 + 1 = 3n_2 + 1$. Then $3n_1 = 3n_2$, so $n_1 = n_2$. But 0 is not hit: $3n + 1 = 0$ would give $n = -1/3 \notin \mathbb{Z}$.

(b) Bijective. Suppose $g(x_1, y_1) = g(x_2, y_2)$. Then $x_1 + y_1 = x_2 + y_2$ and $x_1 - y_1 = x_2 - y_2$. Adding the two equations gives $2x_1 = 2x_2$, and subtracting gives $2y_1 = 2y_2$, so $(x_1, y_1) = (x_2, y_2)$. Hence g is injective. Let $(u, v) \in \mathbb{R}^2$. Then

$$g\Big(\frac{u+v}{2}, \frac{u-v}{2}\Big) = \Big(\frac{u+v}{2} + \frac{u-v}{2}, \frac{u+v}{2} - \frac{u-v}{2}\Big) = (u, v).$$

Hence g is surjective.

(c) Surjective, not injective. Let $m \in \mathbb{N}$. Then $h(2m) = \lfloor m \rfloor = m$, so h is surjective. But $h(0) = 0 = h(1)$, a collision, so h is not injective.

(d) Injective, not surjective. Suppose $(t_1, t_1^2) = (t_2, t_2^2)$. Comparing first coordinates gives $t_1 = t_2$. But $(0, 1)$ is not hit: $p(t) = (0, 1)$ forces $t = 0$ and then $t^2 = 0 \neq 1$.

Exercise. B2: Inverses

(a) Show that $f\colon \mathbb{R} \to \mathbb{R}$, $f(x) = 7 - 4x$, is invertible and find f^{-1}.

(b) Let $D = \mathbb{R} \setminus \{1\}$ and define $k\colon D \to D$ by $k(x) = \dfrac{x+1}{x-1}$. Show that k really takes values in D, and that $k \circ k = \mathrm{id}_D$. Hence deduce that k is bijective and find k^{-1}.

(c) With g as in B1(b), find g^{-1}. Hence find the inverse of $g \circ g$ without computing $g \circ g$ first.

Solution.

(a) Put $e(y) = (7-y)/4$. Then $e(f(x)) = (7-(7-4x))/4 = x$ and $f(e(y)) = 7-(7-y) = y$. So e is an inverse of f, and by Theorem 0.4.7 f is invertible with $f^{-1}(y) = (7-y)/4$.

(b) For $x \in D$ the denominator $x-1$ is nonzero, so $k(x)$ is a real number. If $k(x) = 1$, then $x+1 = x-1$, that is, $1 = -1$, which is false. Hence $k(x) \in D$. Next, for $x \in D$,

$$k(k(x)) = \frac{\frac{x+1}{x-1} + 1}{\frac{x+1}{x-1} - 1} = \frac{(x+1)+(x-1)}{(x+1)-(x-1)} = \frac{2x}{2} = x,$$

where the second equality multiplies numerator and denominator by $x-1 \neq 0$. So $k \circ k = \mathrm{id}_D$. This says k is an inverse of itself, so k is invertible, hence bijective by Theorem 0.4.7, and $k^{-1} = k$.

(c) The solution of B1(b) shows that $e(u,v) = \left(\frac{u+v}{2}, \frac{u-v}{2}\right)$ satisfies $g \circ e = \mathrm{id}_{\mathbb{R}^2}$. Also $e(g(x,y)) = e(x+y, x-y) = \left(\frac{2x}{2}, \frac{2y}{2}\right) = (x,y)$, so $e \circ g = \mathrm{id}_{\mathbb{R}^2}$. Hence $g^{-1} = e$. By Theorem 0.4.8,

$$(g \circ g)^{-1} = g^{-1} \circ g^{-1}, \qquad g^{-1}(g^{-1}(u,v)) = g^{-1}\Big(\frac{u+v}{2}, \frac{u-v}{2}\Big) = \Big(\frac{u}{2}, \frac{v}{2}\Big).$$

(As a check, $g(g(x,y)) = g(x+y, x-y) = (2x, 2y)$, which $(u,v) \mapsto (u/2, v/2)$ indeed undoes.)

Exercise. B3: Images of unions and intersections

Let $f\colon X \to Y$ and $A_1, A_2 \subseteq X$.

(a) Prove that $f(A_1 \cup A_2) = f(A_1) \cup f(A_2)$.

(b) Prove that $f(A_1 \cap A_2) \subseteq f(A_1) \cap f(A_2)$.

(c) Using $q(x) = x^2$ on $\mathbb{R}$, give an example where the inclusion in (b) is strict.

Solution.

(a) We use double inclusion (Theorem 0.3.1).
($\subseteq$) Let $y \in f(A_1 \cup A_2)$. Then $y = f(a)$ for some $a \in A_1 \cup A_2$. If $a \in A_1$, then $y \in f(A_1)$; if $a \in A_2$, then $y \in f(A_2)$. In either case $y \in f(A_1) \cup f(A_2)$.
($\supseteq$) Let $y \in f(A_1) \cup f(A_2)$. If $y \in f(A_1)$, then $y = f(a)$ for some $a \in A_1 \subseteq A_1 \cup A_2$, so $y \in f(A_1 \cup A_2)$. The case $y \in f(A_2)$ is the same with the roles of A_1 and A_2 swapped. This proves the equality.

(b) Let $y \in f(A_1 \cap A_2)$. Then $y = f(a)$ for some $a \in A_1 \cap A_2$. Since $a \in A_1$, $y \in f(A_1)$, and since $a \in A_2$, $y \in f(A_2)$. Hence $y \in f(A_1) \cap f(A_2)$.

(c) Take $A_1 = \{-1\}$ and $A_2 = \{1\}$. Then $A_1 \cap A_2 = \varnothing$, so $q(A_1 \cap A_2) = \varnothing$, while $q(A_1) \cap q(A_2) = \{1\} \cap \{1\} = \{1\}$. The inclusion is strict. The collision $q(-1) = q(1)$ is what goes wrong; for an injective f, equality holds in (b).

0.4.8.3 C. Going deeper

Exercise. C1: Left inverses are not unique

(a) Let $f\colon X \to Y$ be injective but **not** surjective, with X having at least two elements. Prove that f has at least two different left inverses.

(b) Find two different left inverses of $f\colon \mathbb{N} \to \mathbb{N}$, $f(n) = n + 1$.

(c) Explain why a bijective function has exactly one left inverse.

Hint for (a): look at the proof of Theorem 0.4.5 and ask where it made a free choice.

Solution.

(a) Since f is not surjective, there is $y_0 \in Y \setminus f(X)$. Since X has at least two elements, there are $x_0 \neq x_1$ in X. As in the proof of Theorem 0.4.5, for $y \in f(X)$ let $\bar{y}$ be the unique $x \in X$ with $f(x) = y$. Define $g_0, g_1\colon Y \to X$ by $g_i(y) = \bar{y}$ for $y \in f(X)$, and $g_i(y) = x_i$ for $y \notin f(X)$. The proof of Theorem 0.4.5 shows that each g_i is a left inverse of f, since $g_i(f(x)) = x$ uses only the values on $f(X)$. But $g_0(y_0) = x_0 \neq x_1 = g_1(y_0)$, so $g_0 \neq g_1$.

(b) The only element missed by f is 0. Define $g_0(m) = m - 1$ for $m \geq 1$ and $g_0(0) = 0$; define $g_1(m) = m - 1$ for $m \geq 1$ and $g_1(0) = 5$. Both satisfy $g_i(n + 1) = n$ for every $n \in \mathbb{N}$, and $g_0(0) \neq g_1(0)$.

(c) Let f be bijective, and let g be any left inverse. By Theorem 0.4.7, f has an inverse f^{-1}, which in particular is a right inverse. By the observation following that theorem, a left inverse and a right inverse of the same function are equal, so $g = f^{-1}$. Hence every left inverse equals f^{-1}. Conversely, f^{-1} is itself a left inverse, since $f^{-1} \circ f = \mathrm{id}_X$. So f has exactly one left inverse.

Exercise. C2: Self-maps of finite and infinite sets

(a) Let X be a finite set and $f\colon X \to X$ injective. Prove that f is bijective.

(b) Show that (a) fails for $X = \mathbb{N}$: give an injective $f\colon \mathbb{N} \to \mathbb{N}$ that is not bijective.

(c) Give a surjective $h\colon \mathbb{N} \to \mathbb{N}$ that is not bijective.

(d) Let X be finite and $f, g\colon X \to X$ with $g \circ f = \mathrm{id}_X$. Prove that $f \circ g = \mathrm{id}_X$ as well.

Hint for (d): first show that f is bijective.

Solution.

(a) The domain and codomain are the same finite set, so $|X| = |X|$. By Theorem 0.4.9, the injective map f is bijective.

(b) $f(n) = n + 1$ is injective, since $n_1 + 1 = n_2 + 1$ gives $n_1 = n_2$, but 0 is not of the form $n + 1$ with $n \in \mathbb{N}$. Hence f is not surjective, so not bijective.

(c) Take $h(n) = \lfloor n/2 \rfloor$. By B1(c) it is surjective but not injective.

(d) Since $g \circ f = \mathrm{id}_X$ is injective, f is injective by Theorem 0.4.4 (a). By (a), f is bijective, so it has an inverse f^{-1} by Theorem 0.4.7. Then

$$g = g \circ \mathrm{id}_X = g \circ (f \circ f^{-1}) = (g \circ f) \circ f^{-1} = \mathrm{id}_X \circ f^{-1} = f^{-1},$$

using Theorem 0.4.2. Hence $f \circ g = f \circ f^{-1} = \mathrm{id}_X$, as claimed. (For $X = \mathbb{N}$ this fails: with $f(n) = n + 1$ and g_0 from C1(b), $g_0 \circ f = \mathrm{id}_{\mathbb{N}}$ but $f(g_0(0)) = 1 \neq 0$. The matrix version of (d), "$\mathbf{AB} = \mathbf{I}$ implies $\mathbf{BA} = \mathbf{I}$ for square matrices", is proved in Chapter 3.)

Exercise. C3: When images and preimages undo each other

Let $f\colon X \to Y$.

(a) Prove that $f(f^{-1}(B)) = B$ for **every** $B \subseteq Y$ if and only if f is surjective.

(b) Prove that $f^{-1}(f(A)) = A$ for **every** $A \subseteq X$ if and only if f is injective.

Hint: for each direction you need only one well-chosen set; try a one-element set.

Solution.

(a) ($\Rightarrow$) Suppose the equality holds for every B, and let $y \in Y$. Taking $B = \{y\}$ gives $f(f^{-1}(\{y\})) = \{y\}$, so $y \in f(f^{-1}(\{y\}))$. Hence $y = f(x)$ for some x, and f is surjective. ($\Leftarrow$) Suppose f is surjective, and let $B \subseteq Y$. The inclusion $f(f^{-1}(B)) \subseteq B$ is Proposition 0.4.1. For the reverse, let $y \in B$. Since f is surjective, $y = f(x)$ for some $x \in X$. Then $f(x) \in B$, so $x \in f^{-1}(B)$, and hence $y = f(x) \in f(f^{-1}(B))$. By Theorem 0.3.1, $f(f^{-1}(B)) = B$.

(b) ($\Rightarrow$) Suppose the equality holds for every A, and suppose $f(x_1) = f(x_2)$. Taking $A = \{x_1\}$ gives $f^{-1}(\{f(x_1)\}) = \{x_1\}$. Since $f(x_2) = f(x_1)$, we have $x_2 \in f^{-1}(\{f(x_1)\}) = \{x_1\}$, so $x_2 = x_1$. Hence f is injective.

($\Leftarrow$) Suppose f is injective, and let $A \subseteq X$. The inclusion $A \subseteq f^{-1}(f(A))$ is Proposition 0.4.1. For the reverse, let $x \in f^{-1}(f(A))$. Then $f(x) \in f(A)$, so $f(x) = f(a)$ for some $a \in A$. Since f is injective, $x = a \in A$. By Theorem 0.3.1, $f^{-1}(f(A)) = A$.

0.5 Relations, Equivalence Classes, and Quotients

Mathematics constantly declares different-looking things to be "the same": $\frac{1}{2}$ and $\frac{2}{4}$ are the same number, 14:00 and 2 p.m. are the same time, and in Chapter 4 two vectors will count as the same when their difference lies in a fixed subspace. This section makes "the same, for present purposes" precise. The notion is an equivalence relation. Gluing equivalent things together produces a new set, the quotient, and the one real danger of working with quotients is defining a function on them that secretly depends on a choice. We will learn to check for that.

0.5.1 Relations

Before we can say what a good notion of sameness is, we need a word for any way of comparing two elements of a set, good or bad. Examples are $\leq$ on numbers, "divides" on integers, and "has the same birthday as" on people. What such a comparison really amounts to is the list of pairs that are related, and a list of pairs is a subset of the Cartesian product (Definition 0.3.10).

Definition 0.5.1. Relation

A **relation** on a set X is a subset $R \subseteq X \times X$. When $(x, y) \in R$ we say x **is related to** y and write $x \, R \, y$, or $x \sim y$ when the relation is fixed.

Example. Relations

(a) On $\mathbb{Z}$: $a \leq b$. The pairs $(1, 2)$ and $(2, 2)$ belong to the relation, $(2, 1)$ does not.

(b) On $\mathbb{Z}$: $a \mid b$ ("a divides b"), meaning $b = ka$ for some $k \in \mathbb{Z}$. So $3 \mid 12$ and $3 \mid 0$, but $3 \nmid 5$.

(c) On $\mathbb{R}$: $x \sim y$ if $x^2 + y^2 = 1$. As a subset of $\mathbb{R} \times \mathbb{R} = \mathbb{R}^2$ this relation is the unit circle, so a relation can literally be a picture.

(d) On any set X: the **empty relation** $R = \varnothing$, in which nothing is related to anything, and the **full relation** $R = X \times X$, in which everything is related to everything. These two extremes are useful test cases for every property below.

Every subset of $X \times X$ is a relation, so the definition has no nonexamples. The interesting question is which properties a given relation has.

0.5.2 Equivalence relations

What must any reasonable notion of "x is the same as y" satisfy? Think of "has the same birthday as". Everyone has the same birthday as themselves. If you share a birthday with me, I share one with you. If x shares a birthday with y, and y with z, then x shares one with z. Those three features are exactly what we ask for.

An equivalence relation is a way of declaring elements "the same" that behaves like equality: everything is the same as itself, sameness goes both ways, and it can be chained.

Definition 0.5.2. Equivalence relation

A relation $\sim$ on a set X is an **equivalence relation** if it satisfies the following three conditions.

(E1) **Reflexive:** $x \sim x$ **for every** $x \in X$.

(E2) **Symmetric:** for all $x, y \in X$, if $x \sim y$ then $y \sim x$.

(E3) **Transitive:** for all $x, y, z \in X$, if $x \sim y$ and $y \sim z$ then $x \sim z$.

In words: (E1) asks every single element to be related to itself, with no exceptions. (E2) says the relation cannot be one-way. (E3) says a chain of two links can be shortened to one link, and so, by applying it repeatedly, can any finite chain. When $x \sim y$ we say x and y are **equivalent**.

The standard move for checking an equivalence relation is to write three short paragraphs, one per condition, each starting from its hypothesis ("Let $x \in X$.", "Suppose $x \sim y$.", "Suppose $x \sim y$ and $y \sim z$.").

Example. Equivalence relations

Show that each of the following is an equivalence relation.

(a) Equality on any set X: $x \sim y$ if $x = y$.

(b) **Congruence modulo** n on $\mathbb{Z}$, for a fixed integer $n \geq 1$: $a \equiv b \pmod{n}$ if $n \mid a - b$.

(c) "Same length" on the set S of finite strings of the letters a, b: $u \sim v$ if u and v have the same number of letters.

(d) "Same value": for a function $f\colon X \to Y$, put $x \sim_f x'$ if $f(x) = f(x')$.

(e) Fractions: on $P = \mathbb{Z} \times (\mathbb{Z} \setminus \{0\})$, put $(a, b) \sim (c, d)$ if $ad = bc$.

Solution.

(a) Every x equals itself; $x = y$ gives $y = x$; $x = y$ and $y = z$ give $x = z$. Equality is the degenerate case: it is the equivalence relation that identifies nothing new, and the model the other examples imitate.

(b) *Reflexive.* Let $a \in \mathbb{Z}$. Then $a - a = 0 = 0 \cdot n$, so $n \mid a - a$. *Symmetric.* Suppose $a - b = kn$ with $k \in \mathbb{Z}$. Then $b - a = (-k)n$, so $n \mid b - a$. *Transitive.* Suppose $a - b = kn$ and $b - c = \ell n$ with $k, \ell \in \mathbb{Z}$. Then $a - c = (a - b) + (b - c) = (k + \ell)n$, so $n \mid a - c$. For $n = 2$ this is "same parity": $a \equiv b \pmod{2}$ says a and b are both even or both odd.

(c) This is the special case of (d) in which $f\colon S \to \mathbb{N}$ sends a string to its length, so it follows from (d), whose proof does not use (c).

(d) *Reflexive:* $f(x) = f(x)$. *Symmetric:* $f(x) = f(x')$ gives $f(x') = f(x)$. *Transitive:* $f(x) = f(x')$ and $f(x') = f(x'')$ give $f(x) = f(x'')$. Each condition is inherited from the same property of equality in Y. Many equivalence relations are of this form: "same distance from the origin" on $\mathbb{R}^2$ is $\sim_f$ for $f(x, y) = x^2 + y^2$.

(e) *Reflexive:* $ab = ba$, so $(a, b) \sim (a, b)$. *Symmetric:* $ad = bc$ gives $cb = da$, which says $(c, d) \sim (a, b)$. *Transitive:* suppose $ad = bc$ and $cf = de$. Multiplying the first equation by f and the second by b gives $adf = bcf = bde$. Hence $d(af - be) = 0$. Since $d \neq 0$, we get $af = be$, that is, $(a, b) \sim (e, f)$. Transitivity is exactly where second entries must be nonzero: with $d = 0$ allowed, $(1, 1) \sim (0, 0) \sim (2, 1)$ but $(1, 1) \not\sim (2, 1)$.

Minimal changes to these examples show that each condition can fail on its own.

Example. Relations that are not equivalence relations

Decide which of (E1)–(E3) each relation satisfies.

(a) $\leq$ on $\mathbb{Z}$.

(b) On $\mathbb{R}$: $x \sim y$ if $|x - y| \leq 1$.

(c) On $\mathbb{R}$: $x \sim y$ if $xy > 0$.

Solution.

(a) Change "=" to "$\leq$". Reflexivity survives ($a \leq a$) and so does transitivity ($a \leq b \leq c$ gives $a \leq c$). **Symmetry fails:** $1 \leq 2$ but $2 \not\leq 1$.

(b) Change "$|x - y| = 0$" (which is equality) to "$|x - y| \leq 1$". Reflexivity survives, since $|x - x| = 0$, and so does symmetry, since $|y - x| = |x - y|$. **Transitivity fails:** $0 \sim 1$ and $1 \sim 2$, but $|0 - 2| = 2 > 1$. "Close to" cannot be chained; many small steps make

a big one.

(c) Symmetry holds, since $yx = xy$. Transitivity holds: if $xy > 0$ and $yz > 0$, then $xy^2z > 0$, and since $y^2 > 0$ this gives $xz > 0$. **Reflexivity fails, at exactly one point:** $0 \cdot 0 = 0$, so $0 \not\sim 0$. On $\mathbb{R} \setminus \{0\}$ the same rule is an equivalence relation ("same sign").

Warning. A tempting argument claims that (E1) follows from (E2) and (E3): "if $x \sim y$, then $y \sim x$ by symmetry, so $x \sim x$ by transitivity." The gap is the word "if". The argument needs some y with $x \sim y$, and there may be none. In Example (Relations that are not equivalence relations) (c), the element 0 is related to nothing, which is exactly why it escapes. Reflexivity must always be checked on its own.

Quick check. On $\mathbb{Z}$, define $a \sim b$ if $ab \geq 0$. Is $\sim$ an equivalence relation?

Solution. No. It is reflexive ($a^2 \geq 0$) and symmetric, but not transitive: $1 \sim 0$ and $0 \sim -1$ since both products are 0, yet $1 \cdot (-1) = -1 < 0$, so $1 \not\sim -1$. Compare Example (Relations that are not equivalence relations) (c): moving from $>$ to $\geq$ repairs reflexivity at 0 but lets 0 link the positives to the negatives.

Why these three conditions. They are precisely what the next subsection needs. Reflexivity will guarantee that every element lies in its own class, so no class is empty and nothing is lost. Symmetry and transitivity together will guarantee that two classes are either identical or disjoint. Drop any one of them and the picture of X cut cleanly into pieces falls apart, as the nonexamples above show.

Remark. Two equivalence relations on matrices are central later: *row equivalence*, which records that one matrix can be reached from another by row operations, and *similarity*, which records that two square matrices describe the same linear map in different coordinates. Both are introduced once matrices are available, later in this chapter, and studied in Chapters 3 and 4.

0.5.3 Equivalence classes and partitions

Once we have decided which elements count as the same, it is natural to collect everything that is the same as a given element into one set.

Definition 0.5.3. Equivalence class

Let $\sim$ be an equivalence relation on X, and let $x \in X$. The **equivalence class** of x is

$$[x] \coloneqq \{y \in X : y \sim x\}.$$

Any element of $[x]$ is called a **representative** of the class.

The word "representative" is chosen carefully. A class is one set, but it usually has many names: $[x] = [y]$ is possible with $x \neq y$.

Example. Equivalence classes

(a) For equality on X, $[x] = \{x\}$. Every class has one element.

(b) For congruence modulo 3 on $\mathbb{Z}$:

$$[0] = \{\dots, -6, -3, 0, 3, 6, \dots\},$$
$$[1] = \{\dots, -5, -2, 1, 4, 7, \dots\},$$
$$[2] = \{\dots, -4, -1, 2, 5, 8, \dots\}.$$

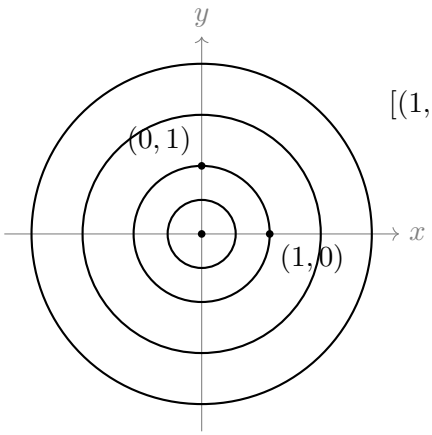
Here $[0] = [3] = [-3]$ and $[2] = [-1]$: one class, many names.

(c) For "same length" on strings of a, b, the class of ab is $\{aa, ab, ba, bb\}$, and the class of the empty string contains only the empty string.

(d) For the full relation $X \times X$ on a nonempty set X, there is a single class, $[x] = X$.

(e) For "same distance from the origin" on $\mathbb{R}^2$, the class of (x, y) is the circle of radius $\sqrt{x^2 + y^2}$ centered at the origin. The class of the origin is the one-point set $\{(0, 0)\}$.

In example (e), the classes look like this. Each circle is one class; the origin on its own is another.



$[(1, 0)] = [(0, 1)]$ is the unit circle

The picture suggests two facts: every point lies on exactly one circle, and two different circles never meet. Both hold for every equivalence relation, and the proof is where the three conditions do their work.

Lemma 0.5.1. Classes are equal or disjoint

Let $\sim$ be an equivalence relation on X, and let $a, b \in X$.

(a) $[a] = [b]$ if and only if $a \sim b$.

(b) Either $[a] = [b]$ or $[a] \cap [b] = \varnothing$.

Proof.

(a) ($\Rightarrow$) Suppose $[a] = [b]$. By reflexivity $a \sim a$, so $a \in [a] = [b]$, which means $a \sim b$.
($\Leftarrow$) Suppose $a \sim b$. Let $x \in [a]$, so $x \sim a$. Since $a \sim b$, transitivity gives $x \sim b$, so $x \in [b]$. Hence $[a] \subseteq [b]$. By symmetry $b \sim a$, and swapping the roles of a and b in the

argument gives $[b] \subseteq [a]$. By Theorem 0.3.1, $[a] = [b]$.

(b) Suppose $[a] \cap [b] \neq \varnothing$, and let $c \in [a] \cap [b]$. Then $c \sim a$ and $c \sim b$. By symmetry $a \sim c$, and by transitivity $a \sim b$. By (a), $[a] = [b]$.

■

Part (a) is the working form of the lemma: to show two classes are equal, show one representative is equivalent to the other. The picture of nonoverlapping pieces that cover everything has its own name.

Definition 0.5.4. Partition

A **partition** of a set X is a collection $\mathcal{P}$ of subsets of X, called **blocks**, such that:

(P1) every block is **nonempty**;

(P2) **distinct** blocks are disjoint: if $B, B' \in \mathcal{P}$ and $B \neq B'$, then $B \cap B' = \varnothing$;

(P3) the blocks cover X: every $x \in X$ lies in some block.

By (P2) and (P3), every element of X lies in **exactly one** block. For example, $\{\{1,3\},\{2\}\}$ is a partition of $\{1,2,3\}$, while $\{\{1,2\},\{2,3\}\}$ is not, because the distinct blocks share 2 and (P2) fails.

Quick check. Is $\{\{1,2,3\},\varnothing\}$ a partition of $\{1,2,3\}$? Is $\{\{1\},\{2\}\}$?

Solution. Neither. The first contains the empty block, so (P1) fails, even though (P2) and (P3) hold. The second does not cover 3, so (P3) fails.

The main theorem says that equivalence relations and partitions are two descriptions of the same thing: "which elements are the same" and "how X is cut into pieces".

Theorem 0.5.2. Equivalence relations and partitions

Let X be a set.

(a) If $\sim$ is an equivalence relation on X, then the set of its equivalence classes is a partition of X.

(b) If $\mathcal{P}$ is a partition of X, then the relation "$x \sim_{\mathcal{P}} y$ if x and y lie in the same block" is an equivalence relation on X, and its equivalence classes are exactly the blocks of $\mathcal{P}$.

(c) The two constructions undo each other: starting from $\sim$, the relation defined in (b) by the partition in (a) is $\sim$ again.

Idea. Part (a) is the lemma plus reflexivity: (P1) and (P3) come from $x \in [x]$, and (P2) is Lemma 0.5.1 (b). Part (b) runs the other way, and each block condition pays for one equivalence condition: covering gives reflexivity, and disjointness gives transitivity, because a middle element y can lie in only one block. Part (c) is Lemma 0.5.1 (a) read as a sentence about blocks.

Proof.

(a) Let $\mathcal{P} = \{[x] : x \in X\}$. For every $x \in X$, reflexivity gives $x \in [x]$. Hence every class is nonempty, which is (P1), and every element lies in some class, which is (P3). If $[a] \neq [b]$, then $[a] \cap [b] = \varnothing$ by Lemma 0.5.1 (b), which is (P2). Hence $\mathcal{P}$ is a partition of X.

(b) *Reflexive.* Let $x \in X$. By (P3), x lies in some block, so $x \sim_{\mathcal{P}} x$. *Symmetric.* "x and y lie in a common block" does not depend on the order of x and y. *Transitive.* Suppose $x, y \in B$ and $y, z \in B'$ for blocks B, B'. Then $y \in B \cap B'$, so $B = B'$ by (P2). Hence $x, z \in B$, and $x \sim_{\mathcal{P}} z$.
Now let $B \in \mathcal{P}$ and $x \in B$; we show $[x] = B$, where the class is taken for $\sim_{\mathcal{P}}$. ($\subseteq$) If $y \in [x]$, then y and x lie in a common block B'; since $x \in B \cap B'$, (P2) gives $B' = B$, so $y \in B$. ($\supseteq$) If $y \in B$, then y and x both lie in B, so $y \sim_{\mathcal{P}} x$ and $y \in [x]$. Hence $[x] = B$ by Theorem 0.3.1. Every class $[x]$ is therefore the block containing x, which exists by (P3); and every block B is a class, namely $[x]$ for any $x \in B$, which exists by (P1). So the classes are exactly the blocks.

(c) Let $\sim$ be an equivalence relation, and let $\approx$ be the relation built by (b) from the classes of $\sim$. If $x \sim y$, then $x, y \in [y]$, so $x \approx y$. Conversely, if $x \approx y$, then $x, y \in [z]$ for some z; thus $x \sim z$ and $y \sim z$, and symmetry and transitivity give $x \sim y$. Hence $\approx$ and $\sim$ are the same relation. This proves the theorem.

■

So whenever you meet an equivalence relation, you may think of it as a way of cutting the set into nonoverlapping pieces, and vice versa. Exercise C2 uses this to count equivalence relations by counting partitions instead.

0.5.4 Quotient sets

Now we take the decisive step: treat each piece as a single new element. The time of day is the classic case. Hours counted from some starting moment are integers, but a clock only records the class of the hour modulo 12.

Definition 0.5.5. Quotient set

Let $\sim$ be an equivalence relation on X. The **quotient set** of X by $\sim$ is the set of all equivalence classes,

$$X/\!\sim \;:= \{[x] : x \in X\}.$$

The function $\pi\colon X \to X/\!\sim$, $\pi(x) = [x]$, is the **quotient map**.

The elements of $X/\!\sim$ are sets, but once the quotient is formed we stop looking inside them and treat each class as a single point. The quotient map is surjective, since every class is $[x] = \pi(x)$ for some x. By Lemma 0.5.1 (a), $\pi(x) = \pi(y)$ exactly when $x \sim y$: the quotient map makes equivalent elements literally equal.

Example. Quotient sets

(a) For an integer $n \geq 1$, the quotient of $\mathbb{Z}$ by congruence modulo n is written $\mathbb{Z}/n\mathbb{Z}$. Show that it has exactly n elements, namely $[0], [1], \ldots, [n-1]$.

(b) Describe the quotient of the set S of strings in a, b by "same length", and of $\mathbb{R}^2$ by "same distance from the origin".

Solution.

(a) Let $a \in \mathbb{Z}$. By division with remainder (Exercise (C2: Division with remainder)), $a = qn + r$ with $q, r \in \mathbb{Z}$ and $0 \leq r \leq n-1$. Then $a - r = qn$, so $a \equiv r \pmod{n}$, and $[a] = [r]$ by Lemma 0.5.1 (a). Hence every class is one of $[0], \ldots, [n-1]$. These are distinct: if $0 \leq r < s \leq n-1$, then $0 < s - r < n$, so $s - r$ is not a multiple of n, hence $r \not\equiv s$ and $[r] \neq [s]$ by Lemma 0.5.1 (a). So $\mathbb{Z}/n\mathbb{Z} = \{[0], [1], \ldots, [n-1]\}$ has exactly n elements.

(b) Two strings are identified exactly when they have the same length, so there is one class for each length $k \in \mathbb{N}$: the set of all 2^k strings of length k. Similarly, $\mathbb{R}^2$ modulo "same distance" has one class for each radius $r \in [0, \infty)$: the circle of radius r, or the origin when $r = 0$. In both cases the quotient is "the same as" the set of values $\mathbb{N}$ or $[0, \infty)$, and the next subsection makes that precise.

0.5.5 Functions on a quotient

We want to compute with classes. On a clock, "three hours after 11 o'clock" is 2 o'clock, which is really addition on $\mathbb{Z}/12\mathbb{Z}$. The natural way to define such an operation is through representatives: to add two classes, pick an element of each, add them, and take the class of the result. The danger is that a class has many representatives, and different choices might give different answers. When they do, the "definition" does not define a function at all.

Here is that failure in its simplest form.

Example. The numerator is not well defined

Let $P = \mathbb{Z} \times (\mathbb{Z} \setminus \{0\})$ with $(a, b) \sim (c, d)$ if $ad = bc$, as in Example (Equivalence relations) (e). Its classes are the rational numbers: the class of (a, b) is the fraction a/b, with all its other names. Is "$[(a, b)] \mapsto a$" a function $P/\!\sim \to \mathbb{Z}$?

Solution. No. Since $1 \cdot 4 = 2 \cdot 2$, we have $(1, 2) \sim (2, 4)$, so $[(1, 2)] = [(2, 4)]$ by Lemma 0.5.1 (a). The rule assigns to this one class the value 1 through one name and 2 through the other. So it does not assign **exactly one** value to each element of the domain, and Definition 0.4.1 fails. "The numerator of $\frac{1}{2}$" is a statement about how a fraction is written, not about the number.

By contrast, "$[(a, b)] \mapsto 1$ if $a = 0$, and 0 otherwise" is fine. If $(a, b) \sim (c, d)$ and $a = 0$, then $bc = ad = 0$, so $c = 0$ because $b \neq 0$. By symmetry, $c = 0$ forces $a = 0$. So equivalent pairs always get the same value.

The contrast in that example is the whole story, and the following theorem turns it into a test. A function f on X passes to the quotient exactly when it cannot tell equivalent elements apart.

Theorem 0.5.3. Functions on a quotient

Let $\sim$ be an equivalence relation on X, let $\pi\colon X \to X/\!\sim$ be the quotient map, and let $f\colon X \to Y$ be a function.

(a) If $f(a) = f(b)$ **whenever** $a \sim b$, then there is **exactly one** function $\bar{f}\colon X/\!\sim \to Y$ with $\bar{f}([x]) = f(x)$ for every $x \in X$, that is, $\bar{f} \circ \pi = f$.

(b) Conversely, if some function $\bar{f}\colon X/\!\sim \to Y$ satisfies $\bar{f}\circ\pi = f$, then $f(a) = f(b)$ whenever $a \sim b$.

Idea. The formula $\bar{f}([x]) = f(x)$ leaves no freedom: if $\bar{f}$ exists, this is its value. The only question is whether the formula makes sense, since the class $C = [x]$ may also be called $[x']$. So look at **all** the values f takes on C at once, the image $f(C)$. A function needs exactly one value; $f(C)$ is nonempty because C is, and has at most one element because all elements of C are equivalent.

Proof.

(a) Let $C \in X/\!\sim$. By Definition 0.5.5, $C = [x]$ for some $x \in X$, so C is nonempty, and hence the image $f(C)$ is nonempty. If $x', x'' \in C$, then $[x'] = C = [x'']$ by Lemma 0.5.1 (a) applied to $x' \sim x$ and $x'' \sim x$; hence $x' \sim x''$ by the same lemma, and $f(x') = f(x'')$ by hypothesis. Therefore $f(C)$ has exactly one element, and we define $\bar{f}(C)$ to be that element. This assigns exactly one value to each class, so $\bar{f}$ is a function. For $x \in X$, we have $x \in [x]$ by reflexivity, so $\bar{f}([x]) \in f([x]) = \{f(x)\}$, that is, $\bar{f}([x]) = f(x)$.
For uniqueness, suppose $g\colon X/\!\sim \to Y$ also satisfies $g([x]) = f(x)$ for every x. Every class is $[x]$ for some x, and there $g([x]) = f(x) = \bar{f}([x])$. Hence $g = \bar{f}$.

(b) Suppose $\bar{f} \circ \pi = f$ and $a \sim b$. By Lemma 0.5.1 (a), $[a] = [b]$. Hence $f(a) = \bar{f}([a]) = \bar{f}([b]) = f(b)$, as claimed.

■

In practice we rarely name f and $\bar{f}$ separately. We write down a rule "$[x] \mapsto$ (expression in x)" and check that equivalent representatives give the same value. That check is what people mean when they say a function is **well defined**. For example, "same value" in Example (Equivalence relations) (d) satisfies the hypothesis by its very definition, so every $f\colon X \to Y$ induces a function $X/\!\sim_f \to Y$. This $\bar{f}$ is injective: $\bar{f}([x]) = \bar{f}([x'])$ means $f(x) = f(x')$, which is $x \sim_f x'$, so $[x] = [x']$. This makes precise the remark that $\mathbb{R}^2$ modulo "same distance" is "the same as" $[0, \infty)$: the induced map $[(x, y)] \mapsto \sqrt{x^2 + y^2}$ is a bijection onto $[0, \infty)$.

The most important application in this chapter is arithmetic modulo n. The operation takes two classes, so we check both representatives at once.

Example. Addition modulo n is well defined

Fix an integer $n \geq 1$. Show that the rule $[a]+[b] \coloneqq [a+b]$ defines a function $\mathbb{Z}/n\mathbb{Z} \times \mathbb{Z}/n\mathbb{Z} \to \mathbb{Z}/n\mathbb{Z}$. Then compute $[3] + [2]$ in $\mathbb{Z}/4\mathbb{Z}$.

Solution. We must show that the value $[a+b]$ depends only on the classes $[a]$ and $[b]$, not on the representatives a and b. Suppose $[a]=[a']$ and $[b]=[b']$. By Lemma 0.5.1 (a), $a \equiv a'$ and $b \equiv b' \pmod{n}$, so $a-a'=kn$ and $b-b'=\ell n$ for some $k, \ell \in \mathbb{Z}$. Then

$$(a+b)-(a'+b')=(a-a')+(b-b')=(k+\ell)n,$$

so $a+b \equiv a'+b' \pmod{n}$, and $[a+b]=[a'+b']$ by the same lemma. Hence every choice of representatives gives the same class, and the rule defines a function.

In $\mathbb{Z}/4\mathbb{Z}$, $[3]+[2]=[5]=[1]$, since $5-1=4$. Choosing other names gives the same answer, as it must: $[7]+[-2]=[5]=[1]$.

This is the same principle as Theorem 0.5.3, applied to a function of two variables. The same check (Exercise (B3: Multiplication modulo n)) shows multiplication $[a][b] \coloneqq [ab]$ is well defined. With these operations, $\mathbb{Z}/p\mathbb{Z}$ for a prime p will become our first finite field later in this chapter, and the identical pattern, now with "$v \sim w$ if $v-w \in U$", will build quotient vector spaces in Chapter 4.

Warning. A rule written in terms of a representative is not automatically a function on classes. Define it, then check it. For example, "$[a] \mapsto a$" from $\mathbb{Z}/5\mathbb{Z}$ to $\mathbb{Z}$ is not a function, because $[0]=[5]$ would be sent to both 0 and 5. The check is not a formality: forgetting it is the most common error in any argument that uses quotients.

Quick check. Which of these rules defines a function? (i) $\mathbb{Z}/6\mathbb{Z} \to \mathbb{Z}/3\mathbb{Z}$, $[a]_6 \mapsto [a]_3$. (ii) $\mathbb{Z}/6\mathbb{Z} \to \mathbb{Z}/4\mathbb{Z}$, $[a]_6 \mapsto [a]_4$. Here the subscript records the modulus.

Solution.

(i) Yes. If $[a]_6=[a']_6$, then $a-a'=6k=3(2k)$, so $[a]_3=[a']_3$.

(ii) No. $[0]_6=[6]_6$, but $[0]_4 \neq [6]_4=[2]_4$, since $6-0=6$ is not a multiple of 4. The first works because $3 \mid 6$; the second fails because $4 \nmid 6$.

0.5.6 Exercises

0.5.6.1 A. Check your understanding

Exercise. A1

(a) State the three conditions for a relation $\sim$ on X to be an equivalence relation.

(b) True or false: a symmetric and transitive relation is automatically reflexive. Give a reason.

(c) True or false: if two equivalence classes $[a]$ and $[b]$ have an element in common, then $a \sim b$.

(d) You want to define a function on $X/\sim$ by a formula in a representative x. What must

you check?

(e) How many elements does $\mathbb{Z}/5\mathbb{Z}$ have? Is $[12] = [-3]$ in $\mathbb{Z}/5\mathbb{Z}$?

Solution.

(a) Reflexive: $x \sim x$ for every $x \in X$. Symmetric: $x \sim y$ implies $y \sim x$. Transitive: $x \sim y$ and $y \sim z$ imply $x \sim z$.

(b) False. On $\mathbb{R}$, $x \sim y$ if $xy > 0$ is symmetric and transitive, but $0 \not\sim 0$; see Example (Relations that are not equivalence relations) (c) and the warning after it.

(c) True. By Lemma 0.5.1 (b) the classes are equal, and by part (a) of the same lemma $a \sim b$.

(d) That equivalent representatives give the same value: if $x \sim x'$, the formula gives the same result for x and x'. By Theorem 0.5.3 this is exactly what is needed.

(e) Five, by Example (Quotient sets) (a). Yes: $12-(-3) = 15 = 3\cdot 5$, so $12 \equiv -3 \pmod 5$, and both classes equal $[2]$.

0.5.6.2 B. Practice

Exercise. B1: Which are equivalence relations?

Determine which of the following are equivalence relations. For those that are, describe the equivalence classes. For those that are not, name a condition that fails with a specific witness.

(a) On $\mathbb{Z}$: $a \sim b$ if $a + b$ is even.

(b) On $\mathbb{Z}$: $a \sim b$ if $a - b$ is odd.

(c) On $\mathbb{R}$: $x \sim y$ if $x - y \in \mathbb{Z}$.

(d) On $\{1, 2, 3, \dots\}$: $m \sim n$ if m and n have a common divisor greater than 1.

(e) On the set of all subsets of $\{1, 2, 3\}$: $A \sim B$ if A and B have the same number of elements.

Solution.

(a) Equivalence relation. *Reflexive:* $a + a = 2a$ is even. *Symmetric:* $b + a = a + b$. *Transitive:* if $a + b$ and $b + c$ are even, then $a + c = (a + b) + (b + c) - 2b$ is a sum of even integers, hence even (Proposition 0.2.1). There are two classes: the even integers and the odd integers. (In fact $a + b$ is even exactly when $a - b = (a + b) - 2b$ is even, so this is congruence modulo 2.)

(b) Not an equivalence relation. Reflexivity fails: $0 - 0 = 0$ is even, so $0 \not\sim 0$. (Transitivity also fails: $0 \sim 1$ and $1 \sim 2$, but $0 - 2$ is even.)

(c) Equivalence relation. *Reflexive:* $x - x = 0 \in \mathbb{Z}$. *Symmetric:* if $x - y = k \in \mathbb{Z}$, then $y - x = -k \in \mathbb{Z}$. *Transitive:* if $x - y = k$ and $y - z = \ell$ with $k, \ell \in \mathbb{Z}$, then $x - z = k + \ell \in \mathbb{Z}$. The class of x is $\{x + k : k \in \mathbb{Z}\}$, the points at integer distance from x in either direction. Each class contains exactly one number in $[0, 1)$, namely $x - \lfloor x \rfloor$.

(d) Not an equivalence relation. Reflexivity fails at 1: the only positive divisor of 1 is 1. Transitivity also fails: $2 \sim 6$ (common divisor 2) and $6 \sim 3$ (common divisor 3), but 2 and 3 have no common divisor greater than 1.

(e) Equivalence relation, since it is "same value" (Example (Equivalence relations) (d)) for the function $A \mapsto |A|$. There are four classes: $\{\varnothing\}$; $\{\{1\},\{2\},\{3\}\}$; $\{\{1,2\},\{1,3\},\{2,3\}\}$; $\{\{1,2,3\}\}$.

Exercise. B2: Parallel lines as classes

On $\mathbb{R}^2$, define $(x,y) \sim (x',y')$ if $x - y = x' - y'$.

(a) Prove that $\sim$ is an equivalence relation.

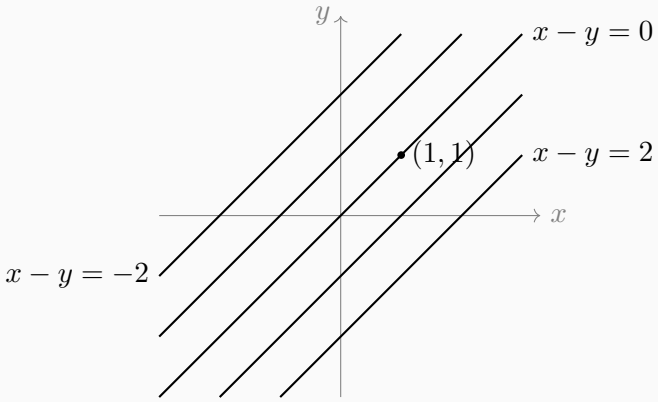
(b) Describe the class of (x_0, y_0) geometrically, and draw several classes.

(c) Show that $[(x,y)] \mapsto x - y$ is a well-defined function $\mathbb{R}^2/\!\sim \to \mathbb{R}$, and that it is a bijection.

Solution.

(a) This is "same value" for $f\colon \mathbb{R}^2 \to \mathbb{R}$, $f(x,y) = x - y$, so it is an equivalence relation by Example (Equivalence relations) (d). Directly: $x - y = x - y$; equality of the numbers $x - y$ and $x' - y'$ is symmetric and transitive.

(b) Put $c = x_0 - y_0$. Then $[(x_0,y_0)] = \{(x,y) \in \mathbb{R}^2 : x - y = c\} = \{(x,y) : y = x - c\}$, the line of slope 1 through (x_0, y_0), which meets the y-axis at $(0, -c)$. The classes are all the lines parallel to $y = x$: each point of the plane lies on exactly one of them, as Theorem 0.5.2 predicts.



(c) Let $f(x,y) = x - y$. If $(x,y) \sim (x',y')$, then $f(x,y) = f(x',y')$ by the definition of $\sim$. By Theorem 0.5.3 (a), there is a function $\bar{f}\colon \mathbb{R}^2/\!\sim \to \mathbb{R}$ with $\bar{f}([(x,y)]) = x - y$.

Injective. Suppose $\bar{f}([(x,y)]) = \bar{f}([(x',y')])$. Then $x - y = x' - y'$, so $(x,y) \sim (x',y')$, and $[(x,y)] = [(x',y')]$ by Lemma 0.5.1 (a).

Surjective. Let $c \in \mathbb{R}$. Then $\bar{f}([(c,0)]) = c - 0 = c$.

Hence $\bar{f}$ is a bijection: the set of parallel lines is labeled, one to one, by the real number c.

Exercise. B3: Multiplication modulo n

Fix an integer $n \geq 1$.

(a) Prove that $[a][b] \coloneqq [ab]$ is a well-defined operation on $\mathbb{Z}/n\mathbb{Z}$.

(b) Write out the multiplication table of $\mathbb{Z}/4\mathbb{Z}$, using the representatives $0, 1, 2, 3$.

(c) Hence find a class $[a] \neq [0]$ in $\mathbb{Z}/4\mathbb{Z}$ with $[a][a] = [0]$.

Solution.

(a) Suppose $[a] = [a']$ and $[b] = [b']$. By Lemma 0.5.1 (a), $a - a' = kn$ and $b - b' = \ell n$ for some $k, \ell \in \mathbb{Z}$. Then

$$ab - a'b' = ab - a'b + a'b - a'b' = (a - a')b + a'(b - b') = (kb + a'\ell)\, n,$$

so $ab \equiv a'b' \pmod{n}$ and $[ab] = [a'b']$ by the same lemma. Hence the value does not depend on the representatives, and the operation is well defined.

(b) Reducing each product modulo 4:

$\cdot$	$[0]$	$[1]$	$[2]$	$[3]$
$[0]$	$[0]$	$[0]$	$[0]$	$[0]$
$[1]$	$[0]$	$[1]$	$[2]$	$[3]$
$[2]$	$[0]$	$[2]$	$[0]$	$[2]$
$[3]$	$[0]$	$[3]$	$[2]$	$[1]$

For instance $[2][3] = [6] = [2]$ and $[3][3] = [9] = [1]$.

(c) From the table, $[2][2] = [4] = [0]$, while $[2] \neq [0]$ since $4 \nmid 2$. So a product of nonzero classes can be zero in $\mathbb{Z}/4\mathbb{Z}$, something that never happens with ordinary numbers.

0.5.6.3 C. Going deeper

Exercise. C1: Squares and powers modulo n

(a) Let $n \geq 1$. Prove that $[x] \mapsto [x^2]$ is a well-defined function $\mathbb{Z}/n\mathbb{Z} \to \mathbb{Z}/n\mathbb{Z}$.

(b) On $\mathbb{N}$, let $\equiv_3$ be congruence modulo 3. Show that the rule $[x] \mapsto [2^x]_3$ does **not** define a function from $\mathbb{N}/\equiv_3$ to $\mathbb{Z}/3\mathbb{Z}$.

(c) On $\mathbb{N}$, let $\equiv_2$ be congruence modulo 2. Prove that $[x] \mapsto [2^x]_3$ **does** define a function from $\mathbb{N}/\equiv_2$ to $\mathbb{Z}/3\mathbb{Z}$.

Hint for (c): compute 2^2 modulo 3.

Solution.

(a) Suppose $[x] = [x']$. By Exercise (B3: Multiplication modulo n) (a) with $a = b = x$ and $a' = b' = x'$, we get $[x][x] = [x'][x']$, that is, $[x^2] = [x'^2]$. Hence the value depends only on the class, and the function is well defined.

(b) In $\mathbb{N}$, $0 \equiv 3 \pmod{3}$, so $[0] = [3]$. The rule sends this class to $[2^0]_3 = [1]_3$ through the representative 0, and to $[2^3]_3 = [8]_3 = [2]_3$ through the representative 3. Since $8 - 1 = 7$ is not a multiple of 3, $[1]_3 \neq [2]_3$. So one class receives two different values, and the rule is not a function.

(c) By Theorem 0.5.3 (a), it suffices to show that $x \equiv x' \pmod{2}$ implies $2^x \equiv 2^{x'} \pmod{3}$. By symmetry of both relations we may assume $x \leq x'$, so $x' = x + 2k$ with $k \in \mathbb{N}$.

First, $4^k \equiv 1 \pmod 3$ for every $k \in \mathbb{N}$, by induction on k (Theorem 0.2.7): $4^0 = 1$, and if $4^k - 1 = 3m$, then $4^{k+1} - 1 = 4(4^k - 1) + 3 = 3(4m + 1)$. Hence

$$2^{x'} - 2^x = 2^x(4^k - 1)$$

is a multiple of 3, so $[2^{x'}]_3 = [2^x]_3$, as claimed. The difference between (b) and (c) is that the powers of 2 modulo 3 repeat with period 2, and 3 is not a multiple of 2.

Exercise. C2: Counting equivalence relations

(a) List all partitions of $\{1, 2, 3\}$.

(b) Hence deduce that there are exactly 5 equivalence relations on $\{1, 2, 3\}$.

(c) Write the equivalence relation corresponding to the partition $\{\{1, 3\}, \{2\}\}$ as a subset of $\{1, 2, 3\} \times \{1, 2, 3\}$.

Solution.

(a) Sort by the number of blocks. One block: $\{\{1, 2, 3\}\}$. Two blocks: one block is a single element and the other the remaining pair, giving $\{\{1\}, \{2, 3\}\}$, $\{\{2\}, \{1, 3\}\}$, $\{\{3\}, \{1, 2\}\}$. Three blocks: $\{\{1\}, \{2\}, \{3\}\}$. Four or more nonempty disjoint blocks are impossible in a three-element set. That is 5 partitions.

(b) By Theorem 0.5.2 (a), every equivalence relation gives a partition, its set of classes. By Theorem 0.5.2 (c), the relation can be recovered from that partition, so distinct equivalence relations give distinct partitions. By Theorem 0.5.2 (b), every partition arises this way. Hence equivalence relations on $\{1, 2, 3\}$ correspond one to one with partitions, and there are exactly 5 of them.

(c) Two elements are related exactly when they lie in a common block:

$$\{(1, 1), (2, 2), (3, 3), (1, 3), (3, 1)\}.$$

Exercise. C3: Intersections and unions of equivalence relations

Let R_1 and R_2 be equivalence relations on a set X, viewed as subsets of $X \times X$.

(a) Prove that $R_1 \cap R_2$ is an equivalence relation on X.

(b) Show by an example that $R_1 \cup R_2$ need not be an equivalence relation. Which condition fails?

(c) For congruence modulo 2 and modulo 3 on $\mathbb{Z}$, identify $R_1 \cap R_2$ as a familiar relation.

Hint for (b): try congruence modulo 2 and modulo 3 on $\mathbb{Z}$.

Solution.

(a) *Reflexive.* Let $x \in X$. Then $(x, x) \in R_1$ and $(x, x) \in R_2$, since both are reflexive, so $(x, x) \in R_1 \cap R_2$. *Symmetric.* Suppose $(x, y) \in R_1 \cap R_2$. Since R_1 and R_2 are symmetric, $(y, x) \in R_1$ and $(y, x) \in R_2$, so $(y, x) \in R_1 \cap R_2$. *Transitive.* Suppose $(x, y), (y, z) \in R_1 \cap R_2$. Both pairs lie in R_1, which is transitive, so $(x, z) \in R_1$; likewise

$(x, z) \in R_2$. Hence $(x, z) \in R_1 \cap R_2$.

(b) Let R_1 be congruence modulo 2 and R_2 congruence modulo 3 on $\mathbb{Z}$. The union is still reflexive and symmetric, since each of R_1, R_2 is. But transitivity fails: $(0, 2) \in R_1$ since $2 \mid 2$, and $(2, 5) \in R_2$ since $3 \mid 3$, yet $(0, 5)$ lies in neither, because 5 is not a multiple of 2 or of 3.

(c) $(a, b) \in R_1 \cap R_2$ means $2 \mid a - b$ and $3 \mid a - b$. This holds exactly when $6 \mid a - b$. Indeed, if $a - b = 2k$ and $a - b = 3m$ with $k, m \in \mathbb{Z}$, then $a - b = 3(a - b) - 2(a - b) = 3 \cdot 2k - 2 \cdot 3m = 6(k - m)$. Conversely, if $a - b = 6j$, then $a - b = 2(3j) = 3(2j)$. So $R_1 \cap R_2$ is congruence modulo 6. The same phenomenon, where intersections behave well and unions do not, will reappear for subspaces in Chapter 1.

0.6 Complex Numbers

The equation $x^2 + 1 = 0$ has no real solution, because the square of a real number is never negative. For most of school mathematics this gap is harmless. In linear algebra it is not: later in the book we meet real matrices, such as a rotation of the plane, whose most useful data cannot be found among the real numbers. This section builds the complex numbers, the number system that closes the gap, and collects the facts about them that the rest of the book uses.

0.6.1 The complex numbers as pairs of reals

The quick way to introduce complex numbers is to say "invent a symbol i with $i^2 = -1$ and compute as usual". That raises a fair worry. How do we know that such a symbol does not lead to a contradiction, the way "invent a number z with $0 \cdot z = 1$" does? The answer is to build the new numbers out of objects we already trust: pairs of real numbers.

Here is the recurring expression that suggests the definition. If we pretend $i^2 = -1$ and expand, we get

$$(a + bi)(c + di) = ac + adi + bci + bd\,i^2 = (ac - bd) + (ad + bc)i.$$

So a "number $a+bi$" is really the pair (a, b), and the product should be the pair $(ac-bd, ad+bc)$. We take this formula as the definition, so nothing is pretended.

A complex number is a pair of real numbers, multiplied as if $i^2 = -1$.

Definition 0.6.1. Complex Numbers

The set of **complex numbers** is $\mathbb{C} \coloneqq \mathbb{R}^2 = \{(a, b) : a, b \in \mathbb{R}\}$, with addition and multiplication defined **for every** $(a, b), (c, d) \in \mathbb{C}$ by

$$(a, b) + (c, d) \coloneqq (a + c,\ b + d), \qquad (a, b)(c, d) \coloneqq (ac - bd,\ ad + bc).$$

We write a for $(a, 0)$ and i for $(0, 1)$, so that $(a, b) = a + bi$. For $z = a + bi$ with $a, b \in \mathbb{R}$, the **real part** of z is $\operatorname{Re} z \coloneqq a$ and the **imaginary part** of z is $\operatorname{Im} z \coloneqq b$.

In words: addition is coordinate by coordinate, as for points of the plane. Multiplication is the formula found by expanding. The last sentence is a naming convention: the real number a is identified with the point $(a, 0)$ on the horizontal axis, and i is the point one unit up. Note that the imaginary part b is a **real** number; it is not bi.

Two things need checking before this convention is safe.

- **The notation $a + bi$ is honest.** By the definitions, $a + bi = (a, 0) + (b, 0)(0, 1) = (a, 0) + (0, b) = (a, b)$. Two complex numbers are equal exactly when they are equal as pairs, so $a+bi = c+di$ with $a, b, c, d \in \mathbb{R}$ means $a = c$ **and** $b = d$. This is how we "compare real and imaginary parts".
- **Identifying a with $(a, 0)$ does not change real arithmetic.** We have $(a, 0) + (c, 0) = (a + c, 0)$ and $(a, 0)(c, 0) = (ac - 0, 0 + 0) = (ac, 0)$. So adding or multiplying two real numbers inside $\mathbb{C}$ gives the same answer as in $\mathbb{R}$.

With these in place, the equation that motivated everything is now a computation: $i^2 = (0, 1)(0, 1) = (0 \cdot 0 - 1 \cdot 1,\ 0 \cdot 1 + 1 \cdot 0) = (-1, 0) = -1$.

Example. First Computations in $\mathbb{C}$

Compute $(2 + 3i)(1 - i)$, the powers $i^0, i^1, \ldots, i^7$, and the product $(a + bi)(a - bi)$ for real a, b.

Solution. By the multiplication rule with $a = 2$, $b = 3$, $c = 1$, $d = -1$,

$$(2 + 3i)(1 - i) = \big(2 \cdot 1 - 3 \cdot (-1)\big) + \big(2 \cdot (-1) + 3 \cdot 1\big)i = 5 + i.$$

For the powers, $i^0 = 1$, $i^1 = i$, $i^2 = -1$, $i^3 = i^2 \cdot i = -i$, and $i^4 = (-i)i = -i^2 = 1$. From then on the pattern repeats with period 4: $i^5 = i$, $i^6 = -1$, $i^7 = -i$.

Finally, $(a + bi)(a - bi) = \big(a^2 - b(-b)\big) + \big(a(-b) + ba\big)i = a^2 + b^2$. This product is a real number, and it is **nonnegative**. We will meet it again as $|z|^2$.

The degenerate case deserves a look too. A real number $a = a + 0i$ is a complex number whose imaginary part is zero, and the number $0 = 0 + 0i$ is the only complex number with both parts zero. Every computation with real numbers is also a computation in $\mathbb{C}$, by the second check above.

A nonexample by minimal change. Change one sign in the multiplication rule: define $(a, b) * (c, d) := (ac + bd,\ ad + bc)$, which is what expanding gives if we pretend $j^2 = +1$ instead. Addition is unchanged, and this product is still commutative and has $(1, 0)$ as a unit. But

$$(1, 1) * (1, -1) = (1 - 1,\ -1 + 1) = (0, 0),$$

so two nonzero elements multiply to zero. Division also breaks down: $(1, 1) * (c, d) = (c+d,\ c+d)$ can never equal the unit $(1, 0)$, because that would need $c + d = 1$ and $c + d = 0$ at once. So the nonzero element $(1, 1)$ has no inverse. The single sign in $ac - bd$ is what makes the complex numbers a number system in which we can divide, as we are about to see.

Quick check. Using the definition, compute i^{10} and $(1 + i)^2$. Is $\mathrm{Im}(3 - 4i)$ equal to $-4i$?

Solution. Since the powers of i repeat with period 4 and $10 = 8 + 2$, we get $i^{10} = i^2 = -1$. Next, $(1 + i)^2 = (1 - 1) + (1 + 1)i = 2i$. No: the imaginary part is the real number -4, not $-4i$.

The usual rules of arithmetic hold. Before computing freely, we should confirm that re-

arranging, regrouping, and expanding brackets are legal in $\mathbb{C}$. The definition does not promise this; it has to be checked.

Proposition 0.6.1. Arithmetic Laws in $\mathbb{C}$

For all $z, w, u \in \mathbb{C}$:

(a) $z + w = w + z$ and $zw = wz$;
(b) $(z + w) + u = z + (w + u)$ and $(zw)u = z(wu)$;
(c) $z(w + u) = zw + zu$;
(d) $z + 0 = z$, $z \cdot 1 = z$, and $z + (-z) = 0$, where $-(a + bi) \coloneqq (-a) + (-b)i$.

Idea. Every law reduces to the corresponding law for real numbers once both sides are written as pairs. Addition is coordinate by coordinate, so its laws are immediate. For multiplication the only real work is associativity: expand both sides and see that they produce the same four terms in each coordinate.

Proof. Let $z = a + bi$, $w = c + di$, and $u = e + fi$ with $a, b, c, d, e, f \in \mathbb{R}$. Since addition acts coordinate by coordinate, the laws for addition in (a), (b), and (d) follow from the same laws in $\mathbb{R}$, applied to each coordinate.

(a) By definition $zw = (ac - bd) + (ad + bc)i$ and $wz = (ca - db) + (cb + da)i$. These are equal because real multiplication and addition are commutative.

(b) Expanding with the definition twice gives

$$(zw)u = \big((ac - bd)e - (ad + bc)f\big) + \big((ac - bd)f + (ad + bc)e\big)i$$

and

$$z(wu) = \big(a(ce - df) - b(cf + de)\big) + \big(a(cf + de) + b(ce - df)\big)i.$$

Multiplying out, both real parts equal $ace - adf - bcf - bde$ and both imaginary parts equal $acf + ade + bce - bdf$. Hence $(zw)u = z(wu)$.

(c) We have $w + u = (c + e) + (d + f)i$, so

$$\begin{aligned} z(w + u) &= \big(a(c + e) - b(d + f)\big) + \big(a(d + f) + b(c + e)\big)i \\ &= \big((ac - bd) + (ae - bf)\big) + \big((ad + bc) + (af + be)\big)i, \end{aligned}$$

which is $zw + zu$ by the definitions of the product and the sum.

(d) With $1 = 1 + 0i$, we get $z \cdot 1 = (a \cdot 1 - b \cdot 0) + (a \cdot 0 + b \cdot 1)i = a + bi = z$. Also $z + (-z) = (a - a) + (b - b)i = 0$. This proves the proposition.

■

From now on we compute in $\mathbb{C}$ exactly as in $\mathbb{R}$: expand brackets, collect terms, and replace i^2 by -1. The proposition is what makes that habit legitimate.

0.6.2 Conjugate and modulus

Division is still missing. To divide by $3+4i$ we want some number to multiply by that lands in the real numbers, where division is familiar. The computation $(a+bi)(a-bi) = a^2+b^2$ in Example (First Computations in $\mathbb{C}$) shows exactly which number does the job. Both ingredients get names.

Conjugating reflects a point in the real axis; the modulus is its distance from the origin.

Definition 0.6.2. Conjugate and Modulus

Let $z = a+bi \in \mathbb{C}$ with $a, b \in \mathbb{R}$. The **complex conjugate** of z is $\overline{z} := a-bi$. The **modulus** (or absolute value) of z is the real number $|z| := \sqrt{a^2+b^2} \geq 0$, the nonnegative square root.

In words: the conjugate keeps the real part and flips the sign of the imaginary part. The modulus is the length of the segment from 0 to the point (a, b), by Pythagoras. For a real number $a = a+0i$, we get $\overline{a} = a$ and $|a| = \sqrt{a^2}$, the usual absolute value, so the notation does not clash with its real meaning. For example, $\overline{3+4i} = 3-4i$, $|3+4i| = 5$, $\overline{i} = -i$, and $|i| = 1$.

The following rules are used constantly, in this section and much later for complex inner products.

Theorem 0.6.2. Properties of Conjugation and Modulus

For all $z, w \in \mathbb{C}$:

(a) $\overline{z+w} = \overline{z} + \overline{w}$ and $\overline{\overline{z}} = z$;
(b) $\overline{zw} = \overline{z}\,\overline{w}$;
(c) $z\overline{z} = |z|^2$ and $z + \overline{z} = 2\operatorname{Re} z$;
(d) $|zw| = |z|\,|w|$ and $|\overline{z}| = |z|$.

Idea. Parts (a)–(c) are direct computations with real and imaginary parts. For (d), expanding $|zw|$ directly is legitimate but messy. It is cleaner to square: by (c) a squared modulus is a product with a conjugate, and (b) lets the conjugate split across the product.

Proof. Let $z = a+bi$ and $w = c+di$ with $a, b, c, d \in \mathbb{R}$.

(a) We have $\overline{z+w} = (a+c)-(b+d)i = (a-bi)+(c-di) = \overline{z}+\overline{w}$, and $\overline{\overline{z}} = \overline{a-bi} = a+bi = z$.

(b) Since $zw = (ac-bd)+(ad+bc)i$, we get $\overline{zw} = (ac-bd)-(ad+bc)i$. On the other hand,

$$\begin{aligned}\overline{z}\,\overline{w} &= (a-bi)(c-di)\\ &= (ac-(-b)(-d)) + (a(-d)+(-b)c)i\\ &= (ac-bd)-(ad+bc)i.\end{aligned}$$

(c) By the computation in Example (First Computations in $\mathbb{C}$), $z\overline{z} = a^2+b^2 = |z|^2$. Also $z+\overline{z} = 2a = 2\operatorname{Re} z$.

(d) By (c), then (b), then Proposition 0.6.1 to reorder the four factors,

$$|zw|^2 = (zw)\overline{zw} = zw\overline{z}\,\overline{w} = (z\overline{z})(w\overline{w}) = |z|^2|w|^2.$$

Both $|zw|$ and $|z||w|$ are nonnegative real numbers with the same square, so they are equal. Finally $|\overline{z}| = \sqrt{a^2 + (-b)^2} = |z|$. This proves the theorem.

■

The first payoff is division. If $z \neq 0$, then $|z|^2 = a^2 + b^2$ is a **strictly** positive real number, so we may divide by it.

Proposition 0.6.3. Inverse of a Complex Number

Let $z \in \mathbb{C}$ with $z \neq 0$. Then $w := \dfrac{1}{|z|^2}\,\overline{z}$ satisfies $zw = 1$. We write z^{-1} or $\dfrac{1}{z}$ for this w.

Proof. Write $z = a + bi$. Since $z \neq 0$, at least one of a, b is nonzero, so $|z|^2 = a^2 + b^2 > 0$. By Proposition 0.6.1 and Theorem 0.6.2 (c),

$$zw = \frac{1}{|z|^2}\,(z\overline{z}) = \frac{1}{|z|^2}\,|z|^2 = 1,$$

as claimed.

■

In practice, "multiply top and bottom by the conjugate". For instance

$$\frac{1}{3+4i} = \frac{3-4i}{(3+4i)(3-4i)} = \frac{3-4i}{25},$$
$$\frac{5+i}{1-i} = \frac{(5+i)(1+i)}{(1-i)(1+i)} = \frac{4+6i}{2} = 2+3i,$$

which agrees with $(2+3i)(1-i) = 5+i$ from Example (First Computations in $\mathbb{C}$). Only one inverse exists; this is a special case of a general fact about fields proved in the next section.

The triangle inequality. The modulus measures distance: $|z-w|$ is the distance between the points z and w of the plane. The side of a triangle is never longer than the other two sides together, and this is the algebraic form of that fact.

Theorem 0.6.4. Triangle Inequality

For all $z, w \in \mathbb{C}$,

$$|z+w| \leq |z| + |w|.$$

Idea. Moduli are square roots, which are awkward to add. Both sides are nonnegative, so it suffices to compare their squares. Expand $|z+w|^2$ using $|u|^2 = u\overline{u}$: the cross terms combine into $2\,\mathrm{Re}(z\overline{w})$. A real part is never bigger than the modulus, and that single estimate finishes the proof.

Proof. First, for any $u = x + yi$ with $x, y \in \mathbb{R}$, we have $\operatorname{Re} u = x \le |x| = \sqrt{x^2} \le \sqrt{x^2 + y^2} = |u|$, since the square root is increasing.

Now let $z, w \in \mathbb{C}$. By Theorem 0.6.2 (c), (a) and Proposition 0.6.1,

$$|z + w|^2 = (z + w)(\overline{z} + \overline{w}) = z\overline{z} + z\overline{w} + w\overline{z} + w\overline{w} = |z|^2 + (z\overline{w} + \overline{z\overline{w}}) + |w|^2,$$

where we used $w\overline{z} = \overline{z\overline{w}}$, which follows from Theorem 0.6.2 (b) and (a). By Theorem 0.6.2 (c) the bracket equals $2\operatorname{Re}(z\overline{w})$. Therefore, using $\operatorname{Re} u \le |u|$ and then Theorem 0.6.2 (d),

$$|z + w|^2 \le |z|^2 + 2|z\overline{w}| + |w|^2 = |z|^2 + 2|z||w| + |w|^2 = \left(|z| + |w|\right)^2.$$

Both $|z + w|$ and $|z| + |w|$ are nonnegative, so taking square roots preserves the inequality. This proves the triangle inequality.

■

The route "square, expand, bound the cross term" is worth remembering: it is exactly the route we will take for norms in inner product spaces. The inequality is often strict. For $z = 1$ and $w = -1$, the left side is 0 and the right side is 2.

0.6.3 Two things that do not carry over from the reals

Complex numbers obey the rules of arithmetic, but not every habit from the real line survives.

The first loss is order. On $\mathbb{R}$ we use two facts about positivity: for every x, exactly one of $x > 0$, $x = 0$, $-x > 0$ holds, and sums and products of positive numbers are positive. Suppose some relation $>$ on $\mathbb{C}$ had both properties. Since $i \neq 0$, either $i > 0$ or $-i > 0$, and in both cases the product rule gives $i^2 = (-i)^2 = -1 > 0$. The same argument applied to $1 \neq 0$ gives $1 = 1^2 = (-1)^2 > 0$. Then both 1 and -1 are positive, so their sum 0 is positive, which contradicts "exactly one".

Warning. There is no ordering on $\mathbb{C}$ compatible with addition and multiplication, so an inequality such as $z < w$ between complex numbers means nothing. Inequalities in this book compare **real** numbers: moduli $|z|$, real parts, or quantities known to be real. Writing $i > 0$ leads, as shown above, to $0 > 0$.

The second loss is the square-root rule $\sqrt{x}\sqrt{y} = \sqrt{xy}$, which on $\mathbb{R}$ is only claimed for $x, y \ge 0$.

Warning. Every nonzero complex number has **two** square roots, and no choice of "the" square root makes $\sqrt{z}\sqrt{w} = \sqrt{zw}$ true for all z, w. With $\sqrt{-1} = i$: $\sqrt{-1} \cdot \sqrt{-1} = i^2 = -1$, but $\sqrt{(-1)(-1)} = \sqrt{1} = 1$. Avoid the symbol $\sqrt{z}$ for nonreal z; say "a square root of z" and solve $w^2 = z$ instead.

Solving $w^2 = z$ is a finite computation, done by comparing real and imaginary parts.

Example. Square Roots of $5-12i$

Find all $w \in \mathbb{C}$ with $w^2 = 5-12i$.

Solution. Write $w = x+yi$ with $x, y \in \mathbb{R}$. Then $w^2 = (x^2-y^2)+2xyi$, so comparing real and imaginary parts, $w^2 = 5-12i$ holds exactly when

$$x^2 - y^2 = 5 \quad \text{and} \quad xy = -6.$$

Taking moduli in $w^2 = 5-12i$ and using Theorem 0.6.2 (d) gives $|w|^2 = |5-12i| = \sqrt{25+144} = 13$, that is $x^2+y^2 = 13$. Adding and subtracting this and the first equation gives $x^2 = 9$ and $y^2 = 4$, so $x = \pm 3$ and $y = \pm 2$. The condition $xy = -6 < 0$ says x and y have opposite signs. Hence $w = 3-2i$ or $w = -3+2i$. Both work: $(3-2i)^2 = 9-12i+4i^2 = 5-12i$, and $(-3+2i)^2 = (3-2i)^2$.

0.6.4 Polar form: multiplication rotates and scales

Addition in $\mathbb{C}$ is the familiar addition of arrows in the plane. Multiplication has a geometric meaning too, and it only becomes visible when we describe a point by its distance from 0 and its angle, instead of by its coordinates.

Let $z \neq 0$ and put $r = |z| > 0$. The number z/r has modulus 1 by Theorem 0.6.2 (d), so it lies on the unit circle. From trigonometry, every point (x, y) with $x^2+y^2 = 1$ is $(\cos\theta, \sin\theta)$ for some real θ. We abbreviate

$$e^{i\theta} := \cos\theta + i\sin\theta \qquad (\theta \in \mathbb{R}),$$

and then every nonzero complex number can be written as

$$z = re^{i\theta}, \qquad r = |z| > 0, \quad \theta \in \mathbb{R}.$$

This is the **polar form** of z, and θ is **an** argument of z. The word "an" matters: θ and $\theta + 2\pi k$ give the same point for every integer k. Conversely, if $re^{i\theta} = se^{i\varphi}$ with $r, s > 0$, then taking moduli gives $r = s$ (as $|e^{i\theta}| = \sqrt{\cos^2\theta + \sin^2\theta} = 1$). Comparing real and imaginary parts then gives $\cos\theta = \cos\varphi$ and $\sin\theta = \sin\varphi$, which from trigonometry means $\theta - \varphi$ is an integer multiple of 2π.

For now $e^{i\theta}$ is only a name for $\cos\theta + i\sin\theta$. The next result is the reason the name is a good one: these numbers multiply the way exponentials do.

Theorem 0.6.5. Multiplication in Polar Form

For all $\theta, \varphi \in \mathbb{R}$,

$$e^{i\theta}e^{i\varphi} = e^{i(\theta+\varphi)}.$$

Consequently, if $z = re^{i\theta}$ and $w = se^{i\varphi}$ with $r, s > 0$, then $zw = rs\,e^{i(\theta+\varphi)}$.

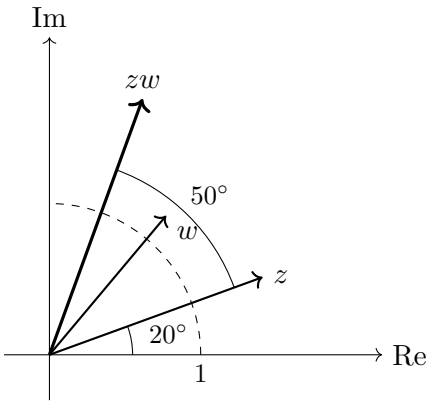
Proof. By the multiplication rule,

$$\begin{aligned} e^{i\theta}e^{i\varphi} &= (\cos\theta\cos\varphi - \sin\theta\sin\varphi) + (\cos\theta\sin\varphi + \sin\theta\cos\varphi)i \\ &= \cos(\theta+\varphi) + i\sin(\theta+\varphi), \end{aligned}$$

where the last equality uses the addition formulas for cosine and sine. This is $e^{i(\theta+\varphi)}$. For the second statement, $zw = (rs)(e^{i\theta}e^{i\varphi})$ by Proposition 0.6.1, and the first statement finishes the proof.

■

In words: **to multiply two complex numbers, multiply their moduli and add their arguments**. Multiplying by a fixed $w = se^{i\varphi}$ is therefore a rotation of the plane about 0 by the angle φ, combined with a stretch by the factor s. The picture shows z, w, and zw for $|z| = 1.5$, $|w| = 1.2$, and arguments $20°$ and $50°$.



The arrow for zw has length $1.5 \times 1.2 = 1.8$ and makes the angle $20° + 50° = 70°$ with the positive real axis: it is the arrow for z, turned by w's angle and stretched by w's length. The small arc marks the argument of z, the large arc marks the extra turn by the argument of w, and the dashed arc is the unit circle, where the numbers $e^{i\theta}$ live.

Two special cases are worth isolating. Multiplying by $i = e^{i\pi/2}$ is a quarter turn counterclockwise, which explains $i^2 = -1$: two quarter turns send 1 to -1. And $e^{i\theta}e^{-i\theta} = e^0 = 1$, so $(e^{i\theta})^{-1} = e^{-i\theta} = \overline{e^{i\theta}}$.

Repeating Theorem 0.6.5 gives powers.

Theorem 0.6.6. De Moivre's Formula

For every $\theta \in \mathbb{R}$ and every $n \in \mathbb{N}$,

$$(e^{i\theta})^n = e^{in\theta}, \quad \text{that is,} \quad (\cos\theta + i\sin\theta)^n = \cos n\theta + i\sin n\theta.$$

Consequently $(re^{i\theta})^n = r^n e^{in\theta}$ for every real $r > 0$.

Proof. We use induction on n (Theorem 0.2.7). For $n = 0$, both sides equal 1, since $e^{i\cdot 0} = \cos 0 + i\sin 0 = 1$. Suppose $(e^{i\theta})^n = e^{in\theta}$ for some $n \in \mathbb{N}$. Then $(e^{i\theta})^{n+1} = (e^{i\theta})^n e^{i\theta} = e^{in\theta}e^{i\theta} = e^{i(n+1)\theta}$ by the induction hypothesis and Theorem 0.6.5. The consequence follows by the same induction, since $(re^{i\theta})^{n+1} = (r^n e^{in\theta})(re^{i\theta}) = r^{n+1}e^{i(n+1)\theta}$ by Theorem 0.6.5.

■

Powers that are painful to expand become one line in polar form.

Example. A Power via Polar Form

Compute $(\sqrt{3}+i)^6$.

Solution. The modulus is $|\sqrt{3}+i| = \sqrt{3+1} = 2$, and $\sqrt{3}+i = 2(\frac{\sqrt{3}}{2} + \frac{1}{2}i) = 2(\cos\frac{\pi}{6} + i\sin\frac{\pi}{6}) = 2e^{i\pi/6}$. By Theorem 0.6.6,

$$(\sqrt{3}+i)^6 = 2^6 e^{i\cdot 6\pi/6} = 64(\cos\pi + i\sin\pi) = -64.$$

Expanding $(\sqrt{3}+i)^6$ with the binomial theorem gives the same answer after seven terms and some cancellation.

Quick check. Use polar form to compute $(1+i)^8$.

Solution. We have $|1+i| = \sqrt{2}$ and $1+i = \sqrt{2}(\frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2}}i) = \sqrt{2}\,e^{i\pi/4}$. By Theorem 0.6.6, $(1+i)^8 = (\sqrt{2})^8 e^{i\cdot 8\pi/4} = 16e^{2\pi i} = 16$. As a cross-check, $(1+i)^2 = 2i$, so $(1+i)^8 = (2i)^4 = 16i^4 = 16$.

0.6.5 Roots of unity

Polar form also turns the equation $z^n = c$ into a question about angles. The most important case is $c = 1$.

Example. Roots of Unity

Let $n \geq 1$ be an integer. Find all $z \in \mathbb{C}$ with $z^n = 1$, and draw them for $n = 6$.

Solution. Since $0^n = 0 \neq 1$, every solution is nonzero, so we may write $z = re^{i\theta}$ with $r > 0$. By Theorem 0.6.6, $z^n = r^n e^{in\theta}$, and $1 = 1 \cdot e^{i0}$. By uniqueness of the polar form up to multiples of 2π, $z^n = 1$ holds exactly when $r^n = 1$ and $n\theta = 2\pi k$ for some integer k. Since $r > 0$ is real, $r^n = 1$ forces $r = 1$. Hence the solutions are

$$e^{2\pi ik/n}, \qquad k \in \mathbb{Z}.$$

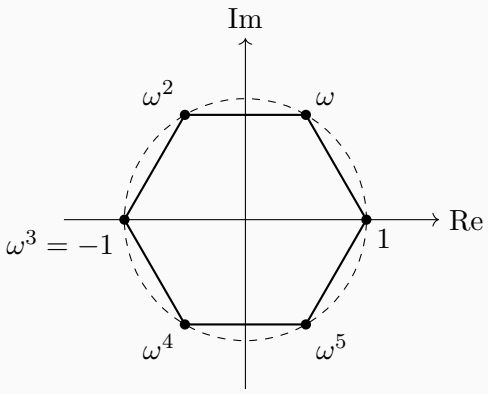
Put $\omega := e^{2\pi i/n}$. By Theorem 0.6.6, $e^{2\pi ik/n} = \omega^k$ for $k \geq 0$. Two integers k, l give the same number exactly when $2\pi(k-l)/n$ is a multiple of 2π, that is, when n divides $k-l$. So there are exactly n distinct solutions,

$$1,\ \omega,\ \omega^2,\ \ldots,\ \omega^{n-1},$$

the **n-th roots of unity**. They sit on the unit circle at equally spaced angles. For $n \geq 3$ they are the corners of a regular polygon with n corners, one of them at 1; for $n = 2$ they are ± 1.

For $n = 6$, the angles are multiples of $60°$, and the roots are

$$1, \quad \tfrac{1}{2} + \tfrac{\sqrt{3}}{2}i, \quad -\tfrac{1}{2} + \tfrac{\sqrt{3}}{2}i, \quad -1, \quad -\tfrac{1}{2} - \tfrac{\sqrt{3}}{2}i, \quad \tfrac{1}{2} - \tfrac{\sqrt{3}}{2}i.$$



The six roots sum to 0: since $\omega^3 = e^{i\pi} = -1$, we have $\omega^{k+3} = -\omega^k$, so the roots cancel in the opposite pairs $\{1, \omega^3\}$, $\{\omega, \omega^4\}$, $\{\omega^2, \omega^5\}$. In the picture, the hexagon is symmetric about its center 0. The sum of the n-th roots of unity is 0 for every $n \geq 2$; see Exercise (C2: Roots of Unity Sum to Zero).

0.6.6 The Fundamental Theorem of Algebra

We built $\mathbb{C}$ so that one polynomial equation, $z^2 + 1 = 0$, has a solution. It turns out that nothing more ever needs to be added: every polynomial equation already has a solution in $\mathbb{C}$. Here a **complex polynomial of degree** n means an expression $a_n z^n + \cdots + a_1 z + a_0$ with $a_0, \ldots, a_n \in \mathbb{C}$ and $a_n \neq 0$; polynomials are studied properly in a later section of this chapter.

Theorem 0.6.7. Fundamental Theorem of Algebra, Statement

Let $n \geq 1$ and let $a_0, a_1, \ldots, a_n \in \mathbb{C}$ with $a_n \neq 0$. Then there exists $z \in \mathbb{C}$ with

$$a_n z^n + a_{n-1} z^{n-1} + \cdots + a_1 z + a_0 = 0.$$

We do not prove this theorem here. Its proof needs tools we have not built, and it is proved in Chapter 6. Nothing in Chapters 0–5 depends on it.

Remark. Combined with division of polynomials (also in Chapter 6), the theorem gives the form in which it is usually applied: every complex polynomial of degree $n \geq 1$ **splits into linear factors**, $a_n z^n + \cdots + a_0 = a_n(z - z_1)(z - z_2)\cdots(z - z_n)$ for some $z_1, \ldots, z_n \in \mathbb{C}$, not necessarily distinct. For linear algebra this is the fact that every complex square matrix has an eigenvalue, the starting point of the theory of canonical forms.

One companion fact needs no heavy machinery. If the coefficients $a_0, \ldots, a_n$ are **real** and $p(z) = a_n z^n + \cdots + a_0$ has $p(z_0) = 0$, then applying Theorem 0.6.2 (a) and (b) term by term, and $\overline{a_k} = a_k$, gives $p(\overline{z_0}) = \overline{p(z_0)} = \overline{0} = 0$. So the nonreal roots of a real polynomial come in conjugate pairs. For example, $z^2 - 2z + 5$ has the roots $1 + 2i$ and $1 - 2i$: indeed $(1 + 2i)^2 - 2(1 + 2i) + 5 = (-3 + 4i) - (2 + 4i) + 5 = 0$.

0.6.7 Exercises

0.6.7.1 A. Check your understanding

Exercise. A1

(a) Define the complex conjugate $\overline{z}$ and the modulus $|z|$ of $z = a + bi$, where $a, b \in \mathbb{R}$.
(b) State the triangle inequality for complex numbers.
(c) True or false: $|z + w| = |z| + |w|$ for all $z, w \in \mathbb{C}$. Justify your answer.
(d) True or false: $z\overline{z}$ is a nonnegative real number for every $z \in \mathbb{C}$. Justify your answer.
(e) True or false: if $z, w \in \mathbb{C}$ and $z^2 = w^2$, then $z = w$. Justify your answer.
(f) State the Fundamental Theorem of Algebra.

Solution.

(a) See Definition 0.6.2: $\overline{z} = a - bi$ and $|z| = \sqrt{a^2 + b^2} \geq 0$.
(b) See Theorem 0.6.4: $|z + w| \leq |z| + |w|$ for all $z, w \in \mathbb{C}$.
(c) False. For $z = 1$ and $w = -1$, $|z + w| = 0$ but $|z| + |w| = 2$.
(d) True. By Theorem 0.6.2 (c), $z\overline{z} = |z|^2 = a^2 + b^2$, which is real and nonnegative.
(e) False. For $z = 1$ and $w = -1$, $z^2 = w^2 = 1$ but $z \neq w$. (What is true is that $z^2 = w^2$ implies $z = w$ or $z = -w$.)
(f) See Theorem 0.6.7: every complex polynomial $a_n z^n + \cdots + a_0$ with $n \geq 1$ and $a_n \neq 0$ has a root in $\mathbb{C}$.

0.6.7.2 B. Practice

Exercise. B1: Arithmetic

Write each of the following in the form $a + bi$ with $a, b \in \mathbb{R}$, or as a real number where indicated.

(a) $(3 - 2i)(4 + i)$.
(b) $\dfrac{1 + 2i}{2 - i}$.
(c) $|(3 + 4i)(5 - 12i)|$, without expanding the product.
(d) $\operatorname{Re} \dfrac{1}{1 + i}$ and $\operatorname{Im} \dfrac{1}{1 + i}$.

Solution.

(a) Expanding and using $i^2 = -1$: $(3 - 2i)(4 + i) = 12 + 3i - 8i - 2i^2 = 14 - 5i$.
(b) Multiplying top and bottom by $\overline{2 - i} = 2 + i$ (Proposition 0.6.3),

$$\frac{1 + 2i}{2 - i} = \frac{(1 + 2i)(2 + i)}{(2 - i)(2 + i)} = \frac{2 + i + 4i + 2i^2}{4 + 1} = \frac{5i}{5} = i.$$

(c) By Theorem 0.6.2 (d), $|(3+4i)(5-12i)| = |3+4i|\,|5-12i| = \sqrt{25} \cdot \sqrt{169} = 5 \cdot 13 = 65$.

(d) By Proposition 0.6.3, $\dfrac{1}{1+i} = \dfrac{1-i}{|1+i|^2} = \dfrac{1-i}{2} = \dfrac{1}{2} - \dfrac{1}{2}i$. Hence the real part is $\frac{1}{2}$ and the imaginary part is $-\frac{1}{2}$.

Exercise. B2: A Square Root

(a) Find all $w \in \mathbb{C}$ with $w^2 = -3 + 4i$.

(b) Hence solve $z^2 + 2z + 4 - 4i = 0$.

Hint: for (b), complete the square.

Solution.

(a) Write $w = x + yi$ with $x, y \in \mathbb{R}$. Comparing real and imaginary parts in $(x^2 - y^2) + 2xyi = -3 + 4i$ gives $x^2 - y^2 = -3$ and $xy = 2$. Taking moduli, by Theorem 0.6.2 (d), $x^2 + y^2 = |w|^2 = |-3 + 4i| = 5$. Adding and subtracting, $x^2 = 1$ and $y^2 = 4$. Since $xy = 2 > 0$, x and y have the same sign. Therefore $w = 1 + 2i$ or $w = -1 - 2i$. Check: $(1 + 2i)^2 = 1 + 4i - 4 = -3 + 4i$, and $(-1 - 2i)^2 = (1 + 2i)^2$.

(b) Completing the square, $z^2 + 2z + 4 - 4i = (z+1)^2 - 1 + 4 - 4i = (z+1)^2 - (-3+4i)$. So the equation holds exactly when $(z + 1)^2 = -3 + 4i$. By (a), this means $z + 1 = 1 + 2i$ or $z + 1 = -1 - 2i$. Hence the solutions are $z = 2i$ and $z = -2 - 2i$.

Exercise. B3

Let $z \in \mathbb{C}$ with $z \neq 0$. Prove that $\dfrac{1}{z} = \overline{z}$ if and only if $|z| = 1$.

Solution. By Proposition 0.6.3, $\dfrac{1}{z} = \dfrac{1}{|z|^2}\overline{z}$.

($\Leftarrow$) Suppose $|z| = 1$. Then $\dfrac{1}{z} = \dfrac{1}{1}\overline{z} = \overline{z}$.

($\Rightarrow$) Suppose $\dfrac{1}{z} = \overline{z}$. Then $\dfrac{1}{|z|^2}\overline{z} = \overline{z}$. Since $z \neq 0$, also $\overline{z} \neq 0$, so multiplying both sides by $(\overline{z})^{-1}$ gives $\dfrac{1}{|z|^2} = 1$, that is, $|z|^2 = 1$. As $|z| \geq 0$, this forces $|z| = 1$. This proves the equivalence.

Exercise. B4: Polar Form

(a) Compute $(1 - \sqrt{3}\,i)^6$.

(b) Find all $z \in \mathbb{C}$ with $z^3 = -8i$, in the form $a + bi$.

Solution.

(a) We have $|1-\sqrt{3}i| = \sqrt{1+3} = 2$ and $1-\sqrt{3}i = 2\big(\cos(-\frac{\pi}{3}) + i\sin(-\frac{\pi}{3})\big) = 2e^{-i\pi/3}$. By Theorem 0.6.6, $(1-\sqrt{3}i)^6 = 64e^{-2\pi i} = 64(\cos 2\pi - i\sin 2\pi) = 64$.

(b) Since $0^3 \neq -8i$, write $z = re^{i\theta}$ with $r > 0$. Also $-8i = 8e^{-i\pi/2}$. By Theorem 0.6.6, $z^3 = r^3 e^{3i\theta}$. By uniqueness of polar form up to multiples of 2π, $z^3 = -8i$ holds exactly when $r^3 = 8$ and $3\theta = -\frac{\pi}{2} + 2\pi k$ for some integer k. Therefore $r = 2$ and $\theta = -\frac{\pi}{6} + \frac{2\pi k}{3}$. The values $k = 0, 1, 2$ give $\theta = -\frac{\pi}{6}, \frac{\pi}{2}, \frac{7\pi}{6}$, and every other k differs from one of these by a multiple of 3, which changes θ by a multiple of 2π. Hence the solutions are

$$2e^{-i\pi/6} = \sqrt{3} - i, \qquad 2e^{i\pi/2} = 2i, \qquad 2e^{7i\pi/6} = -\sqrt{3} - i.$$

Check: $(2i)^3 = 8i^3 = -8i$.

0.6.7.3 C. Going deeper

Exercise. C1: Reverse Triangle Inequality

Prove that $\big||z| - |w|\big| \leq |z - w|$ for all $z, w \in \mathbb{C}$.

Hint: write $z = (z - w) + w$.

Solution. Let $z, w \in \mathbb{C}$. Since $z = (z-w) + w$, the triangle inequality (Theorem 0.6.4) gives $|z| \leq |z-w| + |w|$, so $|z| - |w| \leq |z-w|$. Swapping the roles of z and w gives $|w| - |z| \leq |w-z|$. Moreover $|w - z| = |(-1)(z - w)| = |-1|\,|z - w| = |z - w|$ by Theorem 0.6.2 (d). The real number $\big||z| - |w|\big|$ equals one of $|z| - |w|$ and $|w| - |z|$, and both are at most $|z - w|$. This proves the inequality.

Exercise. C2: Roots of Unity Sum to Zero

Let $n \geq 2$ and $\omega = e^{2\pi i/n}$.

(a) Prove that $1 + \omega + \omega^2 + \cdots + \omega^{n-1} = 0$.

(b) Deduce that $\displaystyle\sum_{k=0}^{n-1} \cos\frac{2\pi k}{n} = 0$ and $\displaystyle\sum_{k=0}^{n-1} \sin\frac{2\pi k}{n} = 0$.

Hint: compare S with ωS, where S is the sum in (a).

Solution.

(a) Let $S = 1 + \omega + \cdots + \omega^{n-1}$. By Theorem 0.6.6, $\omega^n = e^{2\pi i} = 1$. Therefore

$$\omega S = \omega + \omega^2 + \cdots + \omega^{n-1} + \omega^n = \omega + \omega^2 + \cdots + \omega^{n-1} + 1 = S,$$

so $(\omega - 1)S = 0$. Since $n \geq 2$, we have $0 < 2\pi/n \leq \pi$, so $\operatorname{Re}\omega = \cos(2\pi/n) < 1$ and $\omega \neq 1$. Hence $\omega - 1 \neq 0$, and multiplying $(\omega - 1)S = 0$ by $(\omega - 1)^{-1}$ (Proposition 0.6.3) gives $S = 0$.

(b) By Theorem 0.6.6, $\omega^k = \cos\frac{2\pi k}{n} + i\sin\frac{2\pi k}{n}$ for each k. Since addition in $\mathbb{C}$ is coordinate by coordinate, the real part of S is $\sum_{k=0}^{n-1}\cos\frac{2\pi k}{n}$ and the imaginary part is $\sum_{k=0}^{n-1}\sin\frac{2\pi k}{n}$. By (a), $S = 0$, so both sums are 0.

Exercise. C3: Symmetries of $\mathbb{C}$ Fixing $\mathbb{R}$

Let $\sigma\colon \mathbb{C} \to \mathbb{C}$ be a function with $\sigma(z+w) = \sigma(z) + \sigma(w)$ and $\sigma(zw) = \sigma(z)\sigma(w)$ for all $z, w \in \mathbb{C}$, and $\sigma(a) = a$ for every $a \in \mathbb{R}$.

(a) Prove that $\sigma(i)^2 = -1$.

(b) Prove that the only complex numbers u with $u^2 = -1$ are i and $-i$.

(c) Deduce that either $\sigma(z) = z$ for every $z \in \mathbb{C}$, or $\sigma(z) = \overline{z}$ for every $z \in \mathbb{C}$.

(d) Show that $z \mapsto \overline{z}$ does satisfy all three conditions on σ.

Hint: for (b), write $u = x + yi$ and compare real and imaginary parts.

Solution.

(a) Since σ respects products and fixes the real number -1, we get $\sigma(i)^2 = \sigma(i \cdot i) = \sigma(-1) = -1$.

(b) Let $u = x + yi$ with $x, y \in \mathbb{R}$ and $u^2 = -1$. Comparing real and imaginary parts in $(x^2 - y^2) + 2xyi = -1$ gives $x^2 - y^2 = -1$ and $xy = 0$. If $y = 0$, then $x^2 = -1$, which is impossible for real x. Hence $x = 0$, so $y^2 = 1$ and $y = \pm 1$. Therefore $u = i$ or $u = -i$, and both satisfy $u^2 = -1$.

(c) By (a) and (b), $\sigma(i) = i$ or $\sigma(i) = -i$. Let $z = a + bi$ with $a, b \in \mathbb{R}$. Since σ respects sums and products and fixes a and b,

$$\sigma(z) = \sigma(a) + \sigma(b)\sigma(i) = a + b\,\sigma(i).$$

If $\sigma(i) = i$, this gives $\sigma(z) = z$ for every z. If $\sigma(i) = -i$, it gives $\sigma(z) = a - bi = \overline{z}$ for every z.

(d) By Theorem 0.6.2 (a) and (b), conjugation respects sums and products. For $a \in \mathbb{R}$, $\overline{a} = \overline{a + 0i} = a - 0i = a$. So conjugation satisfies all three conditions.

0.7 Fields

Linear algebra is about equations like $2x + 3y = 7$, and so far we have met several number systems in which such equations make sense: $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$, and the arithmetic of remainders $\mathbb{Z}/n\mathbb{Z}$. Which of them can serve as the numbers of linear algebra? This section answers by isolating the rules that solving linear equations actually uses. A number system obeying those rules is called a field, and fields are the scalars of every later chapter.

0.7.1 What solving linear equations needs

Here is a small system, solved the way you learned at school:

$$\begin{aligned} 2x + 3y &= 7, \\ x - y &= 1. \end{aligned}$$

Subtracting twice the second equation from the first gives $5y = 5$. Dividing by 5 gives $y = 1$, and then $x = 1 + y = 2$. Check: $2 \cdot 2 + 3 \cdot 1 = 7$ and $2 - 1 = 1$.

Look at what we used. We added and subtracted equations, multiplied an equation by a number, and divided by a **nonzero** number. Silently we also rearranged terms, regrouped them, and

expanded brackets. Every step of Gaussian elimination is of this kind, and every one of them works in $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$.

Now try the integers. The equation $2x = 1$ involves only integers, yet it has no integer solution, because the step "divide by 2" leaves $\mathbb{Z}$. Elimination stalls as soon as a pivot has no inverse. So what separates $\mathbb{Z}$ from $\mathbb{Q}$ is division, and division is exactly what we must demand.

A field is a number system where you can add, subtract, multiply, and divide by anything nonzero, with the usual rules.

0.7.2 The definition

Definition 0.7.1. Field

A **field** is a set F together with two functions, **addition** $+\colon F \times F \to F$, $(a, b) \mapsto a + b$, and **multiplication** $\cdot\colon F \times F \to F$, $(a, b) \mapsto a \cdot b = ab$, satisfying the following axioms.

(F1) $a + b = b + a$ **for all** $a, b \in F$.

(F2) $(a + b) + c = a + (b + c)$ **for all** $a, b, c \in F$.

(F3) There exists an element $0 \in F$ such that $a + 0 = a$ **for every** $a \in F$.

(F4) **For every** $a \in F$ there exists an element $-a \in F$ such that $a + (-a) = 0$.

(F5) $(ab)c = a(bc)$ **for all** $a, b, c \in F$.

(F6) $ab = ba$ **for all** $a, b \in F$.

(F7) There exists an element $1 \in F$ with $1 \neq 0$ such that $a \cdot 1 = a$ **for every** $a \in F$.

(F8) **For every nonzero** $a \in F$ there exists an element $a^{-1} \in F$ such that $aa^{-1} = 1$.

(F9) $a(b + c) = ab + ac$ **for all** $a, b, c \in F$.

Read the axioms in groups.

- **(F1)–(F4)** are about addition alone. It is commutative and associative, it has a neutral element 0, and every element has a negative. Subtraction is then defined by $a - b := a + (-b)$.
- **(F5)–(F8)** say the same things about multiplication, with one change: only **nonzero** elements are required to have inverses. Division is defined by $a/b := ab^{-1}$ for $b \neq 0$. We never divide by 0, and the axioms never ask us to.
- **(F7)** also insists that $1 \neq 0$: the two neutral elements are different.
- **(F9)** is the one rule linking the two operations. It is what lets us expand brackets.

Note the order of quantifiers in (F3): **one** element 0 works **for every** a. This is stronger than "for every a there is some element z_a with $a + z_a = a$". In (F4) the order is reversed, as it must be: the negative $-a$ depends on a.

We often say "F is a field" when the operations are clear, and write $(F, +, \cdot)$ when they need to be named. Since addition and multiplication are functions into F, sums and products of elements of F stay in F; when we build a field inside a bigger number system, this "closure" is the first thing to check.

0.7.3 First consequences of the axioms

The notation in the definition hides a question. We wrote "0", "$-a$", and "a^{-1}" as if each were a single well-defined element. The axioms only say that such elements **exist**. If a field could contain two different elements behaving like 0, the symbol 0 would be ambiguous, and so would $a - b$ and a/b. The next theorem removes this worry and proves the familiar rules of algebra from the nine axioms.

Theorem 0.7.1. Basic Properties of Fields

Let F be a field and let $a, b, c \in F$.

(a) The element 0 in (F3) is unique, and the element 1 in (F7) is unique.
(b) (Cancellation) If $a + b = a + c$, then $b = c$. If $ab = ac$ and $a \neq 0$, then $b = c$.
(c) The negative $-a$ in (F4) is unique, and for $a \neq 0$ the inverse a^{-1} in (F8) is unique.
(d) $0 \cdot a = 0$.
(e) $(-1)a = -a$.
(f) If $ab = 0$, then $a = 0$ or $b = 0$.

Idea. Each part is a short chain of axioms, and the chain is found backwards from the goal. For uniqueness we use the standard move: suppose two elements both have the property, and show they are equal. For (d) there is no axiom about $0 \cdot a$ at all, so we must create a situation where one applies. The trick is to rewrite 0 in a fancy way as $0 + 0$, expand with (F9), and cancel. For (f), the hypothesis $a \neq 0$ is there to be spent on (F8).

We prove the uniqueness of 0, additive cancellation, the uniqueness of negatives, (d), and (f). The remaining statements are Exercise (B1: The Remaining Basic Properties).

Proof.

(a) Suppose 0 and $0'$ both satisfy (F3). Applying (F3) for $0'$ to the element 0, and (F3) for 0 to the element $0'$, and using (F1),

$$0 = 0 + 0' = 0' + 0 = 0'.$$

So the zero element is unique.

(b) Suppose $a + b = a + c$. Adding $-a$ on the left of both sides and using (F2),

$$((-a) + a) + b = ((-a) + a) + c.$$

By (F1) and (F4), $(-a) + a = a + (-a) = 0$. Hence $0 + b = 0 + c$, and by (F1) and (F3) this says $b = c$.

(c) Suppose b and c both satisfy (F4) for a, that is, $a + b = 0$ and $a + c = 0$. Then $a + b = a + c$, so $b = c$ by additive cancellation in (b).

(d) By (F3), $0 + 0 = 0$. Therefore, by (F9),

$$a \cdot 0 = a(0 + 0) = a \cdot 0 + a \cdot 0.$$

By (F3), also $a\cdot 0 = a\cdot 0+0$. Comparing, $a\cdot 0+0 = a\cdot 0+a\cdot 0$, and additive cancellation (b) gives $0 = a\cdot 0$. Finally $0\cdot a = a\cdot 0$ by (F6), so $0\cdot a = 0$.

(e) Suppose $ab = 0$. If $a = 0$ there is nothing to prove, so suppose $a \neq 0$. By (F8), a^{-1} exists. Then, by (F7), (F8), (F6), and (F5),

$$b = b\cdot 1 = b(aa^{-1}) = (ba)a^{-1} = (ab)a^{-1} = 0\cdot a^{-1} = 0,$$

where the last equality is (d). This shows that $a = 0$ or $b = 0$.

■

Two small consequences are used without comment from now on: $0 + a = a$ and $1\cdot a = a$, by (F1), (F3), (F6), and (F7). More importantly, every identity of school algebra that is proved from these rules, such as $(a+b)^2 = a^2 + 2ab + b^2$ or $(-a)(-b) = ab$, now holds in every field. Part (f) is the property we will use most: a product of nonzero elements is nonzero.

0.7.4 Examples

We check the standard examples first, then an exotic one, then a degenerate case.

Example. The Fields $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$

Explain why $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$, with their usual addition and multiplication, are fields.

Solution. For $\mathbb{R}$, the axioms (F1)–(F9) are the rules of arithmetic for real numbers that we take as familiar, and $1 \neq 0$.

For $\mathbb{Q}$, the sum, product, and negative of fractions are fractions, and so is the inverse $\frac{q}{p}$ of a nonzero fraction $\frac{p}{q}$. So the operations of $\mathbb{R}$ restrict to functions $\mathbb{Q}\times\mathbb{Q}\to\mathbb{Q}$, and $0, 1, -a, a^{-1}$ lie in $\mathbb{Q}$ whenever a does. The identities in (F1), (F2), (F5), (F6), and (F9) hold for all real numbers, so in particular for rational ones. Hence $\mathbb{Q}$ is a field.

For $\mathbb{C}$, the axioms (F1)–(F6) and (F9), together with the existence of $1 = 1 + 0i$, are Proposition 0.6.1. We have $1 = (1,0) \neq (0,0) = 0$. Finally (F8) is Proposition 0.6.3: a nonzero z has the inverse $\overline{z}/|z|^2$. Hence $\mathbb{C}$ is a field.

The next example is a field strictly between $\mathbb{Q}$ and $\mathbb{R}$. It shows that the inverse axiom (F8) can take real work to check.

Example. $\mathbb{Q}(\sqrt{2})$ Is a Field

Let $\mathbb{Q}(\sqrt{2}) := \{a + b\sqrt{2} : a, b \in \mathbb{Q}\} \subseteq \mathbb{R}$, with the addition and multiplication of $\mathbb{R}$. Show that $\mathbb{Q}(\sqrt{2})$ is a field, and find the inverse of $1 + 2\sqrt{2}$.

Solution. Let $x = a + b\sqrt{2}$ and $y = c + d\sqrt{2}$ with $a, b, c, d \in \mathbb{Q}$.

Closure. We have $x + y = (a + c) + (b + d)\sqrt{2}$ and $xy = (ac + 2bd) + (ad + bc)\sqrt{2}$, using $\sqrt{2}\cdot\sqrt{2} = 2$. The coefficients are rational, so $x + y, xy \in \mathbb{Q}(\sqrt{2})$.

Axioms that are identities. (F1), (F2), (F5), (F6), and (F9) hold for all real numbers, hence for the elements of $\mathbb{Q}(\sqrt{2})$.

Neutral elements and negatives. $0 = 0 + 0\sqrt{2}$ and $1 = 1 + 0\sqrt{2}$ lie in $\mathbb{Q}(\sqrt{2})$, and $1 \neq 0$. The negative $-x = (-a) + (-b)\sqrt{2}$ lies in $\mathbb{Q}(\sqrt{2})$.

Inverses. Suppose $x \neq 0$. We first claim $a - b\sqrt{2} \neq 0$. If $b \neq 0$ and $a - b\sqrt{2} = 0$, then $\sqrt{2} = a/b$ would be rational, contradicting Theorem 0.2.3. If $b = 0$, then $a = x \neq 0$, so $a - b\sqrt{2} = a \neq 0$. Therefore $a - b\sqrt{2} \neq 0$, and by Theorem 0.7.1 (f) in $\mathbb{R}$,

$$N := (a + b\sqrt{2})(a - b\sqrt{2}) = a^2 - 2b^2 \neq 0.$$

Since N is a nonzero rational number, the element

$$x' := \frac{a}{N} + \frac{-b}{N}\sqrt{2} \in \mathbb{Q}(\sqrt{2})$$

satisfies $xx' = \frac{1}{N}(a + b\sqrt{2})(a - b\sqrt{2}) = 1$. So (F8) holds, and $\mathbb{Q}(\sqrt{2})$ is a field.

For $x = 1 + 2\sqrt{2}$ we get $N = 1 - 8 = -7$, so $x^{-1} = -\frac{1}{7} + \frac{2}{7}\sqrt{2} = \frac{-1+2\sqrt{2}}{7}$. Check: $(1 + 2\sqrt{2})(-1 + 2\sqrt{2}) = -1 + 2\sqrt{2} - 2\sqrt{2} + 8 = 7$.

The method is "rationalize the denominator", the same move as multiplying by the conjugate in $\mathbb{C}$. The irrationality of $\sqrt{2}$ is exactly what guarantees the new denominator is not zero.

Quick check. Is $\{a + b\sqrt{3} : a, b \in \mathbb{Q}\}$, with the operations of $\mathbb{R}$, a field?

Solution. Yes, by the same argument as Example ($\mathbb{Q}(\sqrt{2})$ Is a Field). Products stay in the set because $\sqrt{3} \cdot \sqrt{3} = 3$ is rational. For inverses, the number $a^2 - 3b^2$ is nonzero whenever $a + b\sqrt{3} \neq 0$, because $\sqrt{3}$ is irrational (Exercise (B3: $\sqrt{3}$ is irrational)). The inverse is $\frac{a-b\sqrt{3}}{a^2-3b^2}$.

The degenerate case. Let $F = \{0\}$ with $0 + 0 = 0$ and $0 \cdot 0 = 0$. Every axiom except one holds: (F1), (F2), (F5), (F6), and (F9) are identities between elements that all equal 0, (F3) and (F4) hold with $-0 = 0$, and (F8) holds vacuously because there is no nonzero element. The only failure is (F7): the one element that could serve as 1 is 0, and the axiom demands $1 \neq 0$. So $\{0\}$ is not a field, and every field has at least the two elements 0 and 1. This one-element system is the reason the clause $1 \neq 0$ is there.

0.7.5 Finite fields

Every field so far is infinite. The arithmetic of remainders gives finite ones, and they are genuinely useful: computer science and coding theory do linear algebra over them all the time.

Recall from Definition 0.5.2 and Definition 0.5.5 that for an integer $n \geq 1$, congruence modulo n is an equivalence relation on $\mathbb{Z}$. We write $[a]$ for the class of a and $\mathbb{Z}/n\mathbb{Z}$ for the quotient set. By division with remainder (Exercise (C2: Division with remainder)), $\mathbb{Z}/n\mathbb{Z} = \{[0], [1], \ldots, [n-1]\}$, and these n classes are distinct. Also $[a] = [b]$ exactly when n divides $a - b$.

Example. Arithmetic on $\mathbb{Z}/n\mathbb{Z}$ and the Field $\mathbb{F}_2$

Let $n \geq 1$. Define $[a] + [b] := [a + b]$ and $[a][b] := [ab]$. Show that these operations are well defined, that they satisfy every field axiom except possibly (F8) and the clause $1 \neq 0$, and that $1 \neq 0$ holds exactly when $n \geq 2$. Write out the tables for $n = 2$.

Solution. Well-definedness. The formulas use representatives, so we must check that the answer does not depend on which representatives we pick. For addition this is Example (Addition modulo n is well defined). For multiplication, suppose $[a] = [a']$ and $[b] = [b']$, so n divides $a - a'$ and $b - b'$. Then

$$ab - a'b' = a(b - b') + (a - a')b'$$

is divisible by n, so $[ab] = [a'b']$.

Axioms. Every identity in (F1), (F2), (F5), (F6), (F9) transfers from $\mathbb{Z}$, because both sides can be computed on representatives. For instance, for (F9),

$$[a]([b] + [c]) = [a][b + c] = [a(b + c)] = [ab + ac] = [ab] + [ac] = [a][b] + [a][c].$$

The same one-line computation works for the other four. The class $[0]$ satisfies (F3) since $[a] + [0] = [a]$, the class $[-a]$ is a negative of $[a]$ as in (F4), and $[1]$ satisfies $[a][1] = [a]$.

The clause $1 \neq 0$. We have $[1] = [0]$ exactly when n divides 1, that is, when $n = 1$. So $[1] \neq [0]$ exactly when $n \geq 2$.

Tables for $n = 2$. Writing $0, 1$ for $[0], [1]$, the set $\mathbb{F}_2 := \mathbb{Z}/2\mathbb{Z} = \{0, 1\}$ has the tables

$+$	0	1
0	0	1
1	1	0

$\cdot$	0	1
0	0	0
1	0	1

The only surprise is $1 + 1 = 0$, since 2 is even. Here (F8) holds as well, since the only nonzero element is 1 and $1 \cdot 1 = 1$. Hence $\mathbb{F}_2$ is a field.

So for $n \geq 2$, whether $\mathbb{Z}/n\mathbb{Z}$ is a field comes down to one question: does every nonzero class have an inverse? To find an inverse of $[a]$ we need an integer x with $[ax] = [1]$, that is, $ax + ny = 1$ for some integer y. The following lemma produces such x and y.

Lemma 0.7.2. Bézout's Lemma for Integers

Let $a, n \in \mathbb{Z}$ with $n \geq 1$, and suppose that the only positive integer dividing both a and n is 1. Then there exist $x, y \in \mathbb{Z}$ with

$$ax + ny = 1.$$

Idea. Consider all positive integers of the form $ax + ny$. There is a smallest one, d, by well-ordering. The plan: ① show d divides a and n; ② conclude $d = 1$ from the hypothesis. For ①, divide a by d: the remainder is again of the form $ax + ny$ and is smaller than d, so minimality forces it to be 0.

Proof. Let $S = \{ax + ny : x, y \in \mathbb{Z}\} \cap \{1, 2, 3, \dots\}$. Since $n = a \cdot 0 + n \cdot 1 \geq 1$, we have $n \in S$, so S is a nonempty subset of $\mathbb{N}$. By the well-ordering principle (Theorem 0.2.11), S has a smallest element $d = ax_0 + ny_0$.

Next we divide a by d, using well-ordering a second time. Let R be the collection of nonnegative integers of the form $a - qd$ with $q \in \mathbb{Z}$. Taking $q = -|a|$ gives $a + |a|d$, which is nonnegative because $d \geq 1$ forces $|a|d \geq |a| \geq -a$; so R is nonempty, and by Theorem 0.2.11 it has a least element $r = a - qd \geq 0$. If $r \geq d$, then $r - d = a - (q + 1)d$ would lie in R and be smaller than r, contradicting minimality. Hence $0 \leq r < d$, and $a = qd + r$; this is the division with remainder of Exercise (C2: Division with remainder), repeated here in the one case the proof needs. Now

$$r = a - q(ax_0 + ny_0) = a(1 - qx_0) + n(-qy_0)$$

has the form $ax + ny$. If $r \geq 1$, then $r \in S$ and $r < d$, contradicting the minimality of d. Hence $r = 0$, and d divides a. The same argument with n in place of a shows that d divides n.

Therefore d is a positive integer dividing both a and n. By hypothesis $d = 1$, so $ax_0 + ny_0 = 1$, as claimed.

■

Now we can decide exactly when $\mathbb{Z}/n\mathbb{Z}$ is a field.

Theorem 0.7.3. $\mathbb{Z}/n\mathbb{Z}$ Is a Field Exactly for Prime n

Let $n \geq 2$ be an integer. Then $\mathbb{Z}/n\mathbb{Z}$ is a field if and only if n is prime.

Idea. By Example (Arithmetic on $\mathbb{Z}/n\mathbb{Z}$ and the Field $\mathbb{F}_2$), only (F8) is in question. If n is prime, a nonzero class $[a]$ has a not divisible by n, so a and n share no factor, and Bézout hands us the inverse. If $n = rs$ factors, then $[r][s] = [n] = [0]$ with both factors nonzero, which no field allows.

Proof. By Example (Arithmetic on $\mathbb{Z}/n\mathbb{Z}$ and the Field $\mathbb{F}_2$), $\mathbb{Z}/n\mathbb{Z}$ satisfies all field axioms except possibly (F8), including $[1] \neq [0]$ since $n \geq 2$.

($\Leftarrow$) Suppose $n = p$ is prime, and let $[a] \neq [0]$. Then p does not divide a. A positive integer dividing p is 1 or p, and p does not divide a, so the only positive common divisor of a and p is 1. By Lemma 0.7.2 there are $x, y \in \mathbb{Z}$ with $ax + py = 1$. Since $ax - 1 = -py$ is divisible by p,

$$[a][x] = [ax] = [1].$$

So $[x]$ is an inverse of $[a]$, and (F8) holds. Hence $\mathbb{Z}/p\mathbb{Z}$ is a field.

($\Rightarrow$) We prove the contrapositive. Suppose $n \geq 2$ is not prime. Then $n = rs$ for some integers r, s with $1 < r < n$ and $1 < s < n$. Since $0 < r < n$, n does not divide r, so $[r] \neq [0]$; likewise $[s] \neq [0]$. But $[r][s] = [rs] = [n] = [0]$. By Theorem 0.7.1 (f), this cannot happen in a field. Hence $\mathbb{Z}/n\mathbb{Z}$ is not a field. This proves the theorem.

■

For a prime p we write $\mathbb{F}_p := \mathbb{Z}/p\mathbb{Z}$, and usually drop the brackets, writing its elements as $0, 1, \ldots, p-1$ and computing "mod p". So $\mathbb{F}_2, \mathbb{F}_3, \mathbb{F}_5, \mathbb{F}_7, \ldots$ are fields, while $\mathbb{Z}/4\mathbb{Z}$, $\mathbb{Z}/6\mathbb{Z}$, and $\mathbb{Z}/9\mathbb{Z}$ are not.

Quick check. Find the inverse of 3 in $\mathbb{F}_7$.

Solution. We need x with $3x$ leaving remainder 1 on division by 7. Trying $x = 1, 2, \ldots$: $3 \cdot 5 = 15 = 2 \cdot 7 + 1$. So $3^{-1} = 5$ in $\mathbb{F}_7$. (By Theorem 0.7.1 (c) this is the only inverse.)

Because $\mathbb{F}_p$ is a field, the elimination from the start of the section runs unchanged in it. Only the arithmetic is new.

Example. Elimination over $\mathbb{F}_5$

Solve the system $2x + y = 3$, $3x + 2y = 4$ over $\mathbb{F}_5$.

Solution. In $\mathbb{F}_5$ the inverses are $1^{-1} = 1$, $2^{-1} = 3$, $3^{-1} = 2$, $4^{-1} = 4$, since $2 \cdot 3 = 6 = 5 + 1$ and $4 \cdot 4 = 16 = 15 + 1$. Multiply the first equation by $2^{-1} = 3$: since $3 \cdot 2 = 1$, $3 \cdot 1 = 3$, and $3 \cdot 3 = 9 = 4$, it becomes $x + 3y = 4$. Subtract 3 times this from the second equation: the x-terms cancel, and we get

$$(2 - 9)y = 4 - 12, \quad \text{that is,} \quad 3y = 2,$$

since $-7 = -10 + 3$ and $-8 = -10 + 2$. Multiply by $3^{-1} = 2$: $y = 4$. Then $x = 4 - 3y = 4 - 12 = -8 = 2$. Check: $2 \cdot 2 + 4 = 8 = 3$ and $3 \cdot 2 + 2 \cdot 4 = 14 = 4$ in $\mathbb{F}_5$. Hence $(x, y) = (2, 4)$ is the unique solution.

0.7.6 Nonexamples

Each of these fails some part of the slogan, and it pays to name the exact clause that fails.

- $\mathbb{Z}$**.** Everything but (F8) holds: $\mathbb{Z}$ is closed under $+$ and $\cdot$, contains 0, 1, and negatives, and inherits the identities from $\mathbb{R}$. But $2 \neq 0$ has no inverse in $\mathbb{Z}$, since $2m$ is even and never equals 1. Changing $\mathbb{Z}$ to $\mathbb{Q}$ is the minimal change that repairs it.
- $\mathbb{Z}/6\mathbb{Z}$**.** This is $\mathbb{F}_5$ with the modulus changed from 5 to 6. All axioms but (F8) still hold (Example (Arithmetic on $\mathbb{Z}/n\mathbb{Z}$ and the Field $\mathbb{F}_2$)), but $[2][3] = [0]$ with $[2], [3] \neq [0]$. By Theorem 0.7.1 (f), $[2]$ cannot have an inverse; indeed $[2][x]$ is always one of $[0], [2], [4]$.
- 2×2 **real matrices**, with the matrix operations you know from school (they are treated properly later in this chapter). The identity matrix plays the role of 1. Two axioms fail. (F6) fails:

$$\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \qquad \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}.$$

(F8) fails too: $\mathbf{E} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$ is not the zero matrix, but every product $\mathbf{EX}$ has second row zero, so $\mathbf{EX}$ is never the identity.

Warning. "$[a] \neq [0]$" does not mean $[a]$ can be divided by in $\mathbb{Z}/n\mathbb{Z}$. In $\mathbb{Z}/6\mathbb{Z}$ the equation $[2]x = [1]$ has no solution, and $[2]x = [4]$ has two, $x = [2]$ and $x = [5]$. Before dividing modulo n, check that n is prime, or at least that a and n share no factor.

0.7.7 Why these axioms

Each clause is there because something we rely on breaks without it.

- **Without (F8)**, elimination stalls, as $2x = 1$ over $\mathbb{Z}$ showed. Worse, the theorems of the next chapters fail: over $\mathbb{Z}$, the analogues of vector spaces (called modules) need not have bases, and "a set of n independent vectors in an n-dimensional space spans it" becomes false. This book always works over a field.
- **Without (F6)**, scalars would not commute with each other. The matrices above show such systems exist, and there are even systems with division but without commutativity. Linear algebra constantly swaps scalars, as in $a(bv) = b(av)$ for a vector v (Chapter 1), so we keep commutativity.
- **Without** $1 \neq 0$, the one-element system $\{0\}$ would count, and every statement of the form "a field has a nonzero element" would fail. Keeping it costs nothing.
- **The name** "field" is historical and carries no extra meaning.

Remark. Many identities we are used to are consequences of the axioms and hold in **every** field, including $\mathbb{F}_2$. But properties that involve **order** or **size**, such as "squares are nonnegative" or "$1+1+\cdots+1 \neq 0$", are not among the axioms. The first already fails in $\mathbb{C}$, where $i^2 = -1$, as the warning on ordering in the previous section explains. The second fails in $\mathbb{F}_p$, which is the subject of the last part of this section.

0.7.8 The characteristic of a field

In $\mathbb{Q}$ we can add 1 to itself as often as we like without reaching 0. In $\mathbb{F}_2$ two copies already give 0, and in $\mathbb{F}_p$ it takes p copies. This number is the most basic invariant of a field, and it decides whether familiar arguments such as "$2x = 0$ implies $x = 0$" are valid.

First some notation. For $a \in F$ and $m \in \mathbb{N}$, define the **integer multiple** $m \cdot a$ by $0 \cdot a := 0$ and $(m+1) \cdot a := m \cdot a + a$. So $m \cdot a = a + a + \cdots + a$ with m terms. This agrees with the field product when $m = 0$, by Theorem 0.7.1 (d).

Lemma 0.7.4. Integer Multiples

Let F be a field, $a \in F$, and $m, k \in \mathbb{N}$. Then

$$(m \cdot 1) + (k \cdot 1) = (m + k) \cdot 1,$$
$$(m \cdot 1)(k \cdot 1) = (mk) \cdot 1,$$
$$(m \cdot 1)\, a = m \cdot a.$$

Proof. All three follow by induction on one variable (Theorem 0.2.7). For the first, induct on k: the case $k = 0$ is $m \cdot 1 + 0 = m \cdot 1$ by (F3), and if it holds for k, then by (F2) and the definition,

$$m \cdot 1 + (k+1) \cdot 1 = (m \cdot 1 + k \cdot 1) + 1 = (m+k) \cdot 1 + 1 = (m+k+1) \cdot 1.$$

For the third, induct on m: the case $m = 0$ is $0 \cdot a = 0$ by Theorem 0.7.1 (d), and if it holds for m, then by (F6), (F9), and (F7),

$$((m+1) \cdot 1)\, a = (m \cdot 1 + 1)\, a = (m \cdot 1)\, a + 1 \cdot a = m \cdot a + a = (m+1) \cdot a.$$

The second is the third with $a = k \cdot 1$, combined with $m \cdot (k \cdot 1) = (mk) \cdot 1$, which follows from the first by induction on m.

■

Definition 0.7.2. Characteristic

Let F be a field. If there is a **positive** integer n with $n \cdot 1 = 0$, the **characteristic** of F is the **smallest** such n. If there is no such n, the characteristic of F is 0.

In words: count how many copies of 1 must be added to reach 0. If it never happens, the convention is to say "characteristic 0", not "characteristic ∞": in that case the only integer $m \geq 0$ with $m \cdot 1 = 0$ is $m = 0$.

- $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$ have characteristic 0, because $n \cdot 1 = n \neq 0$ for every positive integer n.
- $\mathbb{F}_p$ has characteristic p: $n \cdot [1] = [n]$, which is $[0]$ exactly when p divides n, and the smallest positive such n is p.
- $\mathbb{F}_2$ is the smallest case: characteristic 2, because $1 + 1 = 0$.

Can a field have characteristic 6? In $\mathbb{Z}/6\mathbb{Z}$, six copies of $[1]$ give $[0]$, but $\mathbb{Z}/6\mathbb{Z}$ is not a field. This is no accident.

Theorem 0.7.5. The Characteristic Is 0 or Prime

The characteristic of a field is either 0 or a prime number.

Idea. Suppose the characteristic n factors as rs with smaller factors. Then the elements $r \cdot 1$ and $s \cdot 1$ are nonzero, by minimality of n, but their product is $n \cdot 1 = 0$. A field has no such pairs. This is the same argument as in Theorem 0.7.3, run inside an arbitrary field.

Proof. Let F be a field of characteristic $n \neq 0$. Since the integer multiple $1 \cdot 1 = 0 + 1 = 1$ is nonzero by (F7), we have $n \neq 1$, so $n \geq 2$. Suppose, for a contradiction, that n is not prime. Then $n = rs$ with integers $1 < r < n$ and $1 < s < n$. By Lemma 0.7.4,

$$(r \cdot 1)(s \cdot 1) = (rs) \cdot 1 = n \cdot 1 = 0.$$

By Theorem 0.7.1 (f), $r \cdot 1 = 0$ or $s \cdot 1 = 0$. Either way we have a positive integer smaller than n whose multiple of 1 is 0, contradicting the minimality of n. Hence n is prime.

■

So the characteristic is 0, 2, 3, 5, 7, and so on. Together with Exercise (C2: Finite Fields Have Prime Characteristic), this says a finite field always has prime characteristic.

Warning. In a field of characteristic 2, such as $\mathbb{F}_2$, we have $1+1=0$, so $-1=1$ and $x=-x$ for **every** x. The familiar argument "$x=-x$, so $2x=0$, so $x=0$" divides by $2=1+1$, which is 0 there. For example, in $\mathbb{F}_2$ the element 1 satisfies $1=-1$ but is not 0. This is why later results, notably on symmetric bilinear forms in Chapter 14, carry the hypothesis "characteristic $\neq 2$".

With fields in hand, the stage for linear algebra is set: in Chapter 1, a vector space is defined **over a field** F, and every result proved there holds at once for $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$, $\mathbb{Q}(\sqrt{2})$, and $\mathbb{F}_p$.

0.7.9 Exercises

0.7.9.1 A. Check your understanding

Exercise. A1

(a) Define what it means for a set F with two operations to be a field.

(b) True or false: in every field, $a^2=0$ implies $a=0$. Justify your answer.

(c) True or false: $\mathbb{Z}/8\mathbb{Z}$ is a field. Justify your answer.

(d) True or false: in every field, $1+1\neq 0$. Justify your answer.

(e) What are the characteristics of $\mathbb{C}$ and of $\mathbb{F}_7$?

(f) The set $\mathbb{N}$ with the usual addition and multiplication is not a field. Name two axioms that fail, each with a witness.

Solution.

(a) See Definition 0.7.1: two operations satisfying (F1)–(F9), including $1\neq 0$.

(b) True. If $a\cdot a=0$, then $a=0$ or $a=0$ by Theorem 0.7.1 (f).

(c) False. Since $8=2\cdot 4$ is not prime, $\mathbb{Z}/8\mathbb{Z}$ is not a field by Theorem 0.7.3; concretely $[2][4]=[0]$ with $[2],[4]\neq[0]$.

(d) False. In $\mathbb{F}_2$, $1+1=0$.

(e) $\mathbb{C}$ has characteristic 0, and $\mathbb{F}_7$ has characteristic 7 (see the examples after Definition 0.7.2).

(f) (F4) fails: $1\in\mathbb{N}$, but no $b\in\mathbb{N}$ has $1+b=0$. (F8) fails: $2\neq 0$, but no $b\in\mathbb{N}$ has $2b=1$.

0.7.9.2 B. Practice

Exercise. B1: The Remaining Basic Properties

Let F be a field and $a,b,c\in F$. Using only the axioms and the parts of Theorem 0.7.1

proved in the text, prove:

(a) the element 1 in (F7) is unique;

(b) if $ab = ac$ and $a \neq 0$, then $b = c$;

(c) for $a \neq 0$, the inverse a^{-1} is unique;

(d) $(-1)a = -a$.

Hint: for (d), show that $a + (-1)a = 0$.

Solution.

(a) Suppose 1 and $1'$ both satisfy (F7). Then, by (F7) for $1'$, (F6), and (F7) for 1, $1 = 1 \cdot 1' = 1' \cdot 1 = 1'$.

(b) Suppose $ab = ac$ and $a \neq 0$. By (F8), a^{-1} exists, and multiplying both sides by it gives $(ab)a^{-1} = (ac)a^{-1}$. By (F6) and (F5), $(ab)a^{-1} = (ba)a^{-1} = b(aa^{-1}) = b \cdot 1 = b$, using (F8) and (F7) at the end. In the same way $(ac)a^{-1} = c$. Hence $b = c$.

(c) Suppose b and c both satisfy (F8) for $a \neq 0$: $ab = 1 = ac$. By (b), $b = c$.

(d) By (F7) and (F6), $a = 1 \cdot a$. Therefore, using (F6) and (F9),

$$a + (-1)a = 1 \cdot a + (-1)a = a \cdot 1 + a(-1) = a\big(1 + (-1)\big) = a \cdot 0 = 0,$$

where the last two equalities use (F4) and Theorem 0.7.1 (d). So $(-1)a$ satisfies (F4) for a. By uniqueness of negatives (Theorem 0.7.1 (c)), $(-1)a = -a$.

Exercise. B2: Solving Equations in $\mathbb{F}_7$

(a) Solve $3x + 5 = 2$ in $\mathbb{F}_7$.

(b) Solve the system $x + 2y = 3$, $3x + y = 1$ over $\mathbb{F}_7$.

Solution.

(a) Adding -5 to both sides gives $3x = 2 - 5 = -3 = 4$ in $\mathbb{F}_7$. Since $3 \cdot 5 = 15 = 1$ in $\mathbb{F}_7$, we have $3^{-1} = 5$. Multiplying by 5, $x = 20 = 6$. Check: $3 \cdot 6 + 5 = 23 = 2$ in $\mathbb{F}_7$. Hence $x = 6$.

(b) Subtract 3 times the first equation from the second: $(1-6)y = 1-9$, that is, $-5y = -8$, or $2y = 6$ in $\mathbb{F}_7$. Since $2 \cdot 4 = 8 = 1$, $2^{-1} = 4$, so $y = 24 = 3$. Then $x = 3 - 2y = 3 - 6 = -3 = 4$. Check: $4 + 2 \cdot 3 = 10 = 3$ and $3 \cdot 4 + 3 = 15 = 1$ in $\mathbb{F}_7$. Hence $(x, y) = (4, 3)$.

Exercise. B3: Which Are Fields?

Determine which of the following are fields, with the stated operations. Justify your answer.

(a) $\mathbb{Z}/9\mathbb{Z}$, with the operations of Example (Arithmetic on $\mathbb{Z}/n\mathbb{Z}$ and the Field $\mathbb{F}_2$).

(b) $\mathbb{Q}(i) := \{a + bi : a, b \in \mathbb{Q}\}$, with the operations of $\mathbb{C}$.

(c) $\{a+bi : a,b \in \mathbb{Z}\}$, with the operations of $\mathbb{C}$.

(d) $\mathbb{F}_5$.

Solution.

(a) Not a field. Since $9 = 3 \cdot 3$ is not prime, Theorem 0.7.3 applies; concretely $[3][3] = [0]$ with $[3] \neq [0]$, which is impossible in a field by Theorem 0.7.1 (f).

(b) A field. Let $x = a+bi$ and $y = c+di$ with $a,b,c,d \in \mathbb{Q}$. Then $x+y = (a+c)+(b+d)i$, $xy = (ac-bd)+(ad+bc)i$, and $-x = -a-bi$ have rational parts, and $0, 1 \in \mathbb{Q}(i)$ with $1 \neq 0$. The identities (F1), (F2), (F5), (F6), (F9) hold because they hold in $\mathbb{C}$ (Example (The Fields $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$)). If $x \neq 0$, then a^2+b^2 is a nonzero rational number, and by Proposition 0.6.3 $x^{-1} = \frac{a}{a^2+b^2} - \frac{b}{a^2+b^2}i$, which lies in $\mathbb{Q}(i)$. So (F8) holds.

(c) Not a field: (F8) fails for 2. Suppose $u = a+bi$ with $a,b \in \mathbb{Z}$ satisfied $2u = 1$. This is also an equation in the field $\mathbb{C}$, where $\frac{1}{2}$ is an inverse of 2; by uniqueness of inverses in $\mathbb{C}$ (Theorem 0.7.1 (c)), $u = \frac{1}{2}$. But $\frac{1}{2}$ has real part $\frac{1}{2} \notin \mathbb{Z}$, a contradiction.

(d) A field, by Theorem 0.7.3, since 5 is prime.

0.7.9.3 C. Going deeper

Exercise. C1: A Field with Four Elements

Let $F = \{a+b\alpha : a,b \in \mathbb{F}_2\}$, a set of four formal symbols $0, 1, \alpha, \beta$, where $\beta := 1+\alpha$. Define

$$(a+b\alpha)+(c+d\alpha) := (a+c)+(b+d)\alpha,$$
$$(a+b\alpha)(c+d\alpha) := (ac+bd)+(ad+bc+bd)\alpha,$$

with all coefficient arithmetic in $\mathbb{F}_2$. (The product is what expanding gives if we replace α^2 by $\alpha+1$.)

(a) Write out the addition and multiplication tables of F.

(b) Prove that F is a field.

(c) Explain why there is no bijection $\varphi\colon F \to \mathbb{Z}/4\mathbb{Z}$ with $\varphi(x+y) = \varphi(x)+\varphi(y)$ and $\varphi(xy) = \varphi(x)\varphi(y)$ for all $x, y \in F$. So this four-element field is not "$\mathbb{Z}/4\mathbb{Z}$ in disguise".

Hint: for (c), look for a nonzero element of $\mathbb{Z}/4\mathbb{Z}$ whose square is zero.

Solution.

(a) Using the formulas, with $1+1=0$ in $\mathbb{F}_2$:

$+$	0	1	α	β
0	0	1	α	β
1	1	0	β	α
α	α	β	0	1
β	β	α	1	0

$\cdot$	0	1	α	β
0	0	0	0	0
1	0	1	α	β
α	0	α	β	1
β	0	β	1	α

For instance, $\alpha \cdot \alpha = (0+1\alpha)(0+1\alpha) = (0+1)+(0+0+1)\alpha = 1+\alpha = \beta$, and

$\alpha\beta = (0+1\alpha)(1+1\alpha) = (0+1)+(0+1+1)\alpha = 1$.

(b) Let $x = a+b\alpha$, $y = c+d\alpha$, $z = e+f\alpha$ with $a,\dots,f \in \mathbb{F}_2$. Since $\mathbb{F}_2$ is a field (Example (Arithmetic on $\mathbb{Z}/n\mathbb{Z}$ and the Field $\mathbb{F}_2$)), we may compute with the coefficients using the rules of algebra.
Addition. It acts coordinate by coordinate, so (F1) and (F2) follow from the same laws in $\mathbb{F}_2$, the element $0 = 0+0\alpha$ satisfies (F3), and $-x = x$ satisfies (F4) because $a+a = 0$ and $b+b = 0$ in $\mathbb{F}_2$.
(F6). The formula for xy is unchanged when (a,b) and (c,d) are swapped: $ac+bd = ca+db$ and $ad+bc+bd = cb+da+db$.
(F7). With $1 = 1+0\alpha$, $x\cdot 1 = (a+0)+(0+b+0)\alpha = x$, and $1 \neq 0$ since the coefficients differ.
(F9). Since $y+z = (c+e)+(d+f)\alpha$,

$$x(y+z) = (a(c+e)+b(d+f)) + (a(d+f)+b(c+e)+b(d+f))\alpha,$$

and expanding the coefficients shows this is $((ac+bd)+(ae+bf)) + ((ad+bc+bd)+(af+be+bf))\alpha = xy+xz$.
(F5). Write $xy = g+h\alpha$ with $g = ac+bd$, $h = ad+bc+bd$. Then $(xy)z = (ge+hf)+(gf+he+hf)\alpha$. Expanding,

$$ge+hf = ace+adf+bcf+bde+bdf,$$
$$gf+he+hf = acf+ade+adf+bce+bcf+bde,$$

where in the second we used $bdf+bdf = 0$ in $\mathbb{F}_2$. Similarly $yz = k+l\alpha$ with $k = ce+df$, $l = cf+de+df$, and $x(yz) = (ak+bl)+(al+bk+bl)\alpha$. Expanding gives $ak+bl = ace+adf+bcf+bde+bdf$ and $al+bk+bl = acf+ade+adf+bce+bcf+bde$, again using $bdf+bdf = 0$. Hence $(xy)z = x(yz)$.
(F8). From the table, $1\cdot 1 = 1$, $\alpha\beta = 1$, and $\beta\alpha = 1$, so every nonzero element has an inverse.
This proves that F is a field.

(c) Suppose such a bijection φ existed. First, $\varphi(0) = \varphi(0+0) = \varphi(0)+\varphi(0)$, so additive cancellation in $\mathbb{Z}/4\mathbb{Z}$ (adding $-\varphi(0)$, which only uses (F1)–(F4), valid there by Example (Arithmetic on $\mathbb{Z}/n\mathbb{Z}$ and the Field $\mathbb{F}_2$)) gives $\varphi(0) = [0]$. Since φ is surjective, there is $x \in F$ with $\varphi(x) = [2]$. As $[2] \neq [0] = \varphi(0)$, we have $x \neq 0$. But

$$\varphi(x\cdot x) = \varphi(x)\varphi(x) = [2][2] = [4] = [0] = \varphi(0),$$

so $x\cdot x = 0$ because φ is injective. This contradicts Theorem 0.7.1 (f) in the field F (from (b)). Hence no such φ exists.

Exercise. C2: Finite Fields Have Prime Characteristic

(a) Let F be a finite field. Prove that the characteristic of F is a prime p.

(b) Let F be a field of characteristic $p > 0$. Prove that $p\cdot a = 0$ for every $a \in F$.

Hint: for (a), the elements $1\cdot 1, 2\cdot 1, 3\cdot 1, \dots$ cannot all be different.

Solution.

(a) The elements $m\cdot 1$ for $m = 1,2,3,\dots$ all lie in the finite set F, so they cannot all be

distinct. Hence there are positive integers $k < m$ with $m \cdot 1 = k \cdot 1$. By Lemma 0.7.4, $m \cdot 1 = k \cdot 1 + (m-k) \cdot 1$, so $k \cdot 1 + (m-k) \cdot 1 = k \cdot 1 + 0$, and additive cancellation (Theorem 0.7.1 (b)) gives $(m-k) \cdot 1 = 0$ with $m - k \geq 1$. Therefore the characteristic of F is not 0. By Theorem 0.7.5, it is a prime p.

(b) Let $a \in F$. By Lemma 0.7.4 and the definition of characteristic, $p \cdot a = (p \cdot 1)\, a = 0 \cdot a = 0$, where the last step is Theorem 0.7.1 (d).

Exercise. C3: The Frobenius Identity

Let p be a prime.

(a) Let $a, b \in \mathbb{Z}$ with p dividing ab. Prove that p divides a or p divides b.

(b) Deduce that p divides the binomial coefficient $\binom{p}{k}$ for $0 < k < p$.

(c) Let F be a field of characteristic p. Prove that $(x+y)^p = x^p + y^p$ for all $x, y \in F$.

You may use the binomial theorem $(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} \cdot (x^k y^{n-k})$ in any field, with integer multiples as in Lemma 0.7.4; its inductive proof uses only the field axioms.

Hint: for (a), use Lemma 0.7.2. For (b), use $p! = \binom{p}{k}\, k!\, (p-k)!$.

Solution.

(a) Suppose p does not divide a. A positive integer dividing p is 1 or p, so the only positive common divisor of a and p is 1. By Lemma 0.7.2 there are $x, y \in \mathbb{Z}$ with $ax + py = 1$. Multiplying by b, $b = (ab)x + p(by)$. Both terms on the right are divisible by p, the first because p divides ab. Hence p divides b.

(b) First, by (a) and induction on the number of factors, if p divides a product of integers then it divides one of the factors. The integers $1, 2, \ldots, p-1$ are not divisible by p, so p divides neither $k!$ nor $(p-k)!$ for $0 < k < p$, and hence, by (a) once more, not their product. Now p divides $p! = \binom{p}{k} \cdot (k!\,(p-k)!)$. By (a), p divides $\binom{p}{k}$.

(c) Let $x, y \in F$. By the binomial theorem,

$$(x+y)^p = x^p + y^p + \sum_{k=1}^{p-1} \binom{p}{k} \cdot (x^k y^{p-k}),$$

since $\binom{p}{0} = \binom{p}{p} = 1$. Fix $0 < k < p$. By (b), $\binom{p}{k} = pm$ for some $m \in \mathbb{N}$. By Lemma 0.7.4,

$$\binom{p}{k} \cdot (x^k y^{p-k}) = ((pm) \cdot 1)(x^k y^{p-k}) = (p \cdot 1)(m \cdot 1)(x^k y^{p-k}) = 0,$$

because $p \cdot 1 = 0$ in characteristic p and 0 times anything is 0 (Theorem 0.7.1 (d)). So every middle term vanishes, and $(x+y)^p = x^p + y^p$, as claimed.

0.8 Polynomials

We now have fields: number systems in which we can add, subtract, multiply, and divide by anything nonzero. The next objects built on top of a field are polynomials. They will be

everywhere in this book: as examples of vectors in Chapter 1, as the characteristic and minimal polynomials of a matrix later on, and as the reason some fields behave differently from others. This section says precisely what a polynomial is, introduces its degree, and explains why a polynomial is **not** the same thing as the function it defines.

0.8.1 What a polynomial is

At school, a polynomial is "an expression like $3x^2 - x + 2$". That is a good picture, but it leaves one question open: what is x? One tempting answer is that x is a number to be plugged in, so a polynomial is really a function. This answer breaks down as soon as we leave the familiar number systems.

Work over the field $\mathbb{F}_2 = \{0, 1\}$ of Example (Arithmetic on $\mathbb{Z}/n\mathbb{Z}$ and the Field $\mathbb{F}_2$), where $1 + 1 = 0$. The expressions $x^2 + x$ and 0 look different. But if we plug in the two elements of $\mathbb{F}_2$, we get $0^2 + 0 = 0$ and $1^2 + 1 = 1 + 1 = 0$. As functions on $\mathbb{F}_2$, the two are the same. If a polynomial *were* its function, then $x^2 + x$ would be the zero polynomial, and it would have no sensible degree. We want $x^2 + x$ to be a genuine polynomial of degree 2, so we must define polynomials by their coefficients alone.

A polynomial is its list of coefficients, and nothing else.

Definition 0.8.1. Polynomial

Let F be a field. A **polynomial over** F in the indeterminate x is a sequence $(a_0, a_1, a_2, \dots)$ of elements of F, indexed by $\mathbb{N}$, such that **only finitely many** a_k are nonzero. That is, there is some $N \in \mathbb{N}$ with $a_k = 0$ **for every** $k > N$.

The element a_k is the **coefficient of** x^k. We write the polynomial as

$$p = a_0 + a_1 x + a_2 x^2 + \cdots + a_N x^N = \sum_{k=0}^{N} a_k x^k.$$

Two polynomials $p = \sum a_k x^k$ and $q = \sum b_k x^k$ are **equal** if and only if $a_k = b_k$ **for every** $k \in \mathbb{N}$.

In words: a polynomial is an infinite list of coefficients that is eventually all zero. The expression $\sum_{k=0}^{N} a_k x^k$ is only a way of displaying the list: x^k marks the position k, and we stop writing once the remaining coefficients are all zero. Two polynomials are equal exactly when they agree position by position. A sequence indexed by $\mathbb{N}$ is a function $\mathbb{N} \to F$, so this is just equality of functions in the sense of Definition 0.4.1.

There is one thing to check about the notation. The number N in the display is not unique: if $a_k = 0$ for every $k > N$, the same is true for every $k > N + 1$. Writing $\sum_{k=0}^{N} a_k x^k$ or $\sum_{k=0}^{N+1} a_k x^k$ only adds a term $0 \cdot x^{N+1}$, which we agree to drop. So the display names the same sequence whichever admissible N we use. We also drop terms with zero coefficient in the middle, write x for $1 \cdot x^1$, and write a_0 for $a_0 x^0$.

Example. Polynomials over various fields

Read off the coefficient sequence of each polynomial, and check it against Definition 0.8.1.

$$3x^2 - x + 2 \in \mathbb{R}[x], \qquad x^5 + \pi x^3 - \sqrt{2} \in \mathbb{R}[x],$$
$$(1+i)x - i \in \mathbb{C}[x], \qquad x^2 + x \in \mathbb{F}_2[x],$$
$$7, \qquad 0.$$

(The symbol $F[x]$ for the set of all polynomials over F is defined formally in Definition 0.8.2 below.)

Solution. The first is the sequence $(2, -1, 3, 0, 0, \dots)$ of real numbers. It has finitely many nonzero terms, so it is a polynomial over $\mathbb{R}$.

The second is $(-\sqrt{2}, 0, 0, \pi, 0, 1, 0, \dots)$. Its coefficients are real but irrational; the definition only asks that they lie in the field.

The third is $(-i, 1+i, 0, \dots)$, a polynomial over $\mathbb{C}$.

The fourth is $(0, 1, 1, 0, \dots)$ with entries in $\mathbb{F}_2$. It is **not** the zero sequence, so it is not the zero polynomial, even though it vanishes at both points of $\mathbb{F}_2$.

The constant 7 is the sequence $(7, 0, 0, \dots)$. It is a polynomial over $\mathbb{Q}$, $\mathbb{R}$, or $\mathbb{C}$. Over $\mathbb{F}_2$ it would be $(1, 0, 0, \dots)$, because $7 = 1$ there.

The last, $0 = (0, 0, 0, \dots)$, is the **zero polynomial**. It is the degenerate case: every coefficient is zero, and the finiteness condition holds with $N = 0$. It matters because it is the one polynomial whose degree needs a special convention, as we will see.

Now the nonexample, by a minimal change. Take the polynomial $1 + x + x^2$, whose sequence is $(1, 1, 1, 0, 0, \dots)$, and change every later 0 to 1. The result $(1, 1, 1, 1, \dots)$, informally "$1 + x + x^2 + x^3 + \cdots$", is still a sequence of elements of $\mathbb{R}$. What fails is the clause **only finitely many** coefficients are nonzero, so this is **not** a polynomial. (It is a *formal power series*, which we will not need.) Expressions such as $x^{-1} + 1$ and $\sqrt{x}$ are not polynomials either: they are not of the form $\sum a_k x^k$ with $k \in \mathbb{N}$ at all.

Why insist on finitely many nonzero coefficients? Because we want to plug elements of the field into polynomials, which means forming the sum $\sum a_k c^k$, and in a field only finite sums make sense. Finiteness is also what gives every nonzero polynomial a highest power, its degree.

0.8.2 Adding and multiplying polynomials

The coefficient picture tells us how to add and multiply. Adding $3x^2 - x + 2$ and $x + 5$ at school means collecting like powers, so we add coefficients position by position. Multiplying means expanding every product $a_i x^i \cdot b_j x^j = a_i b_j x^{i+j}$ and collecting like powers, so the coefficient of x^k collects all pairs (i, j) with $i + j = k$.

Definition 0.8.2. The polynomial ring $F[x]$

Let F be a field. The set of **all** polynomials over F is denoted $F[x]$. For $p = \sum a_k x^k$ and

$q = \sum b_k x^k$ in $F[x]$, define the **sum** $p+q$ and the **product** pq by their coefficients:

$$p+q = \sum_k (a_k + b_k)x^k, \qquad pq = \sum_k \Big(\sum_{i=0}^{k} a_i b_{k-i}\Big)x^k.$$

We must check that $p+q$ and pq are again polynomials, that is, that their coefficient sequences are eventually zero. Suppose $a_i = 0$ for every $i > m$ and $b_j = 0$ for every $j > n$. If $k > \max(m,n)$, then $a_k + b_k = 0 + 0 = 0$. If $k > m+n$, then in each product $a_i b_{k-i}$ either $i > m$ or $k - i > n$, because otherwise $k = i + (k-i) \le m+n$. So every term is 0, and the coefficient of x^k in pq is 0. Hence both are polynomials, and addition and multiplication are well defined on $F[x]$.

With this multiplication, $x \cdot x$ has coefficient sequence $(0,0,1,0,\dots)$, and by induction on k the k-th power of the polynomial x is the sequence with a 1 in position k and 0 elsewhere. So the notation x^k in Definition 0.8.1 agrees with the actual power, and $\sum a_k x^k$ is an honest sum of products.

Example. A product, coefficient by coefficient

In $\mathbb{R}[x]$, compute $(3x^2 - x + 2)(x+1)$ from the definition.

Solution. Here $(a_0, a_1, a_2) = (2, -1, 3)$ and $(b_0, b_1) = (1, 1)$, with all other coefficients 0. The coefficient of x^k in the product is $\sum_{i=0}^{k} a_i b_{k-i}$:

$$\begin{aligned} k=0:&\ a_0b_0 = 2,\\ k=1:&\ a_0b_1 + a_1b_0 = 2 - 1 = 1,\\ k=2:&\ a_0b_2 + a_1b_1 + a_2b_0 = 0 - 1 + 3 = 2,\\ k=3:&\ a_0b_3 + a_1b_2 + a_2b_1 + a_3b_0 = 0 + 0 + 3 + 0 = 3, \end{aligned}$$

and every coefficient with $k \ge 4$ is 0 by the argument above (here $m+n = 3$). Hence $(3x^2 - x + 2)(x+1) = 3x^3 + 2x^2 + x + 2$, which is what expanding by hand gives.

The operations obey the rules we expect. We record them once, so that later sections can use them freely.

Theorem 0.8.1. Arithmetic in $F[x]$

Let F be a field and $p, q, r \in F[x]$. Then:

1. $p + q = q + p$ and $(p+q) + r = p + (q + r)$;
2. $p + 0 = p$, and $p + (-p) = 0$, where $-p = \sum(-a_k)x^k$;
3. $pq = qp$ and $(pq)r = p(qr)$;
4. $1 \cdot p = p$, where 1 is the constant polynomial $(1,0,0,\dots)$;
5. $p(q+r) = pq + pr$.

Idea. Two polynomials are equal when their coefficients agree, so each identity reduces to an identity between coefficients, which the field axioms supply. For products it helps to write the coefficient of x^k in pq symmetrically, as a sum over all pairs (i,j) of natural numbers with $i + j = k$.

Proof. Let $p = \sum a_k x^k$, $q = \sum b_k x^k$, and $r = \sum c_k x^k$. By Definition 0.8.1, it suffices to compare the coefficient of x^k on both sides, for each $k \in \mathbb{N}$.

Parts 1, 2, and 4 hold because in each position they are the corresponding field axioms of Definition 0.7.1: for example, the coefficient of x^k in $p+q$ is $a_k + b_k = b_k + a_k$, the coefficient in $q+p$. For 4, the coefficient of x^k in $1 \cdot p$ is $\sum_{i+j=k} e_i a_j$ with $e_0 = 1$ and $e_i = 0$ for $i \geq 1$, which equals a_k.

For 3, the coefficient of x^k in pq is $\sum_{i+j=k} a_i b_j$. Since multiplication in F is commutative, this equals $\sum_{i+j=k} b_j a_i$, the coefficient in qp. Next, the coefficient of x^k in $(pq)r$ is

$$\sum_{s+l=k} \Big(\sum_{i+j=s} a_i b_j \Big) c_l = \sum_{i+j+l=k} a_i b_j c_l,$$

where the equality uses distributivity and associativity in F. The same computation for $p(qr)$ gives $\sum_{i+s=k} a_i (\sum_{j+l=s} b_j c_l) = \sum_{i+j+l=k} a_i b_j c_l$. Therefore $(pq)r = p(qr)$.

For 5, the coefficient of x^k in $p(q+r)$ is $\sum_{i+j=k} a_i (b_j + c_j) = \sum_{i+j=k} a_i b_j + \sum_{i+j=k} a_i c_j$ by distributivity in F, which is the coefficient in $pq + pr$. This proves the theorem.

■

A set with an addition and a multiplication satisfying these rules is called a **commutative ring**; this is where the name "polynomial ring" comes from. We will not need the word again in this chapter. What matters is that we may now calculate with polynomials exactly as at school.

Quick check. Over $\mathbb{R}$, are $(x+1)^2$ and x^2+2x+1 equal as polynomials? Over $\mathbb{F}_2$, are $(x+1)^2$ and x^2+1 equal as polynomials?

Solution. Both answers are yes, but for a reason that depends on the field. Expanding gives $(x+1)^2 = x^2 + 2x + 1$, with coefficient sequence $(1, 2, 1, 0, \dots)$. Over $\mathbb{R}$ this is exactly the sequence of $x^2 + 2x + 1$. Over $\mathbb{F}_2$ we have $2 = 1 + 1 = 0$, so the sequence is $(1, 0, 1, 0, \dots)$, which is the sequence of $x^2 + 1$.

0.8.3 Degree

The most useful single number attached to a polynomial is the highest power of x that actually occurs. Every nonzero polynomial has one: it has at least one nonzero coefficient and only finitely many, so there is a largest position holding a nonzero coefficient. The zero polynomial has no such position, and we must decide what to do with it.

Definition 0.8.3. Degree

Let $p = \sum a_k x^k \in F[x]$.

- If $p \neq 0$, the **degree** of p, written $\deg p$, is the **largest** $n \in \mathbb{N}$ with $a_n \neq 0$. The coefficient a_n is the **leading coefficient** of p, and p is **monic** if its leading coefficient is 1.

- The zero polynomial has degree $\deg 0 = -\infty$.

We use the conventions $-\infty < n$ and $-\infty + n = n + (-\infty) = -\infty$ for every $n \in \mathbb{N}$, and $-\infty + (-\infty) = -\infty$.

Examples, simplest first. A nonzero constant c has degree 0; its leading coefficient is c. The polynomial $x + 5 \in \mathbb{R}[x]$ is monic of degree 1. The polynomial $3x^3 + 2x^2 + x + 2$ from Example (A product, coefficient by coefficient) has degree 3 and leading coefficient 3. Over $\mathbb{F}_2$, $x^2 + x$ has degree 2, even though it vanishes at every point of $\mathbb{F}_2$.

Why $-\infty$, and not 0 or -1? Look at the formula we expect: the degree of a product is the sum of the degrees. With $\deg 0 = 0$ it fails at once, since $0 \cdot x = 0$ would give $0 = \deg 0 = \deg 0 + \deg x = 1$. With $\deg 0 = -1$ it fails too: $-1 = \deg(0 \cdot x) \neq \deg 0 + \deg x = 0$. The formula needs $\deg 0 + n = \deg 0$ for every n, and $-\infty$ is exactly the symbol with that property. The convention keeps a formula true, and that is its only job.

It turns out that the degree of a sum is controlled by the larger degree, and the degree of a product is the sum of the degrees:

Theorem 0.8.2. Degree of a sum

Let F be a field and $p, q \in F[x]$. Then

$$\deg(p + q) \leq \max(\deg p, \deg q),$$

with **equality** whenever $\deg p \neq \deg q$.

Proof. Let $p = \sum a_k x^k$, $q = \sum b_k x^k$, and $M = \max(\deg p, \deg q)$. If $M = -\infty$, then $p = q = 0$, so $p + q = 0$ and both sides are $-\infty$. Suppose $M \in \mathbb{N}$. For every $k > M$ we have $k > \deg p$ and $k > \deg q$, so $a_k = b_k = 0$ by Definition 0.8.3. Hence the coefficient $a_k + b_k$ of x^k in $p + q$ is 0 for every $k > M$, and therefore $\deg(p + q) \leq M$.

Now suppose $\deg p \neq \deg q$; by symmetry we may assume $\deg p > \deg q$, so $M = \deg p \in \mathbb{N}$. Then $a_M \neq 0$ and $b_M = 0$, because $M > \deg q$. The coefficient of x^M in $p + q$ is $a_M + 0 = a_M \neq 0$, so $\deg(p + q) \geq M$. Together with the first part, $\deg(p + q) = M$. This proves the theorem.

■

The inequality can be strict: over any field, $\deg(x + (-x)) = \deg 0 = -\infty < 1$. The leading terms cancel. That is the only way the degree of a sum can drop.

Theorem 0.8.3. Degree of a product

Let F be a field and $p, q \in F[x]$. Then

$$\deg(pq) = \deg p + \deg q.$$

If p and q are nonzero, the leading coefficient of pq is the product of their leading coefficients.

Idea. Expanding $(a_m x^m + \text{lower terms})(b_n x^n + \text{lower terms})$, the only way to reach x^{m+n} is $a_m x^m \cdot b_n x^n$, and nothing reaches beyond. So the top coefficient is $a_m b_n$. The one real question is whether $a_m b_n$ could be 0. In a field it cannot, and this is exactly where the field axioms enter.

Proof. If $p = 0$ or $q = 0$, then $pq = 0$ by the definition of the product, so both sides equal $-\infty$ by the conventions in Definition 0.8.3. Suppose now $p = \sum a_k x^k$ and $q = \sum b_k x^k$ are nonzero, with $m = \deg p$ and $n = \deg q$. Then $a_m \neq 0$, $b_n \neq 0$, $a_i = 0$ for $i > m$, and $b_j = 0$ for $j > n$.

We showed after Definition 0.8.2 that the coefficient of x^k in pq is 0 for every $k > m + n$. For $k = m + n$, the coefficient is $\sum_{i+j=m+n} a_i b_j$. A term with $i > m$ has $a_i = 0$, and a term with $i < m$ has $j = m + n - i > n$, so $b_j = 0$. Hence only $i = m$, $j = n$ survives, and the coefficient is $a_m b_n$.

Since $a_m \neq 0$ and $b_n \neq 0$, Theorem 0.7.1 gives $a_m b_n \neq 0$, because a product of nonzero elements of a field is nonzero. Therefore $\deg(pq) = m + n$ and the leading coefficient of pq is $a_m b_n$, as claimed.

■

Warning. The field axioms are doing real work here. The same definitions make sense with coefficients in $\mathbb{Z}/6\mathbb{Z}$, which is **not** a field. There $(2x+1)(3x^2+x) = 6x^3+5x^2+x = 5x^2+x$, because $2 \cdot 3 = 6 = 0$. The factors have degrees 1 and 2, but the product has degree 2, not 3.

The first consequence is that polynomials over a field behave like integers: a product can only be zero if a factor is, and nonzero factors can be canceled.

Corollary 0.8.4. No zero divisors in $F[x]$

Let F be a field and $p, q, r \in F[x]$.

1. If $pq = 0$, then $p = 0$ or $q = 0$.
2. If $pq = pr$ and $p \neq 0$, then $q = r$.

Proof. For 1, we prove the contrapositive. Suppose $p \neq 0$ and $q \neq 0$. Then $\deg p, \deg q \in \mathbb{N}$, so by Theorem 0.8.3, $\deg(pq) = \deg p + \deg q \in \mathbb{N}$. In particular $\deg(pq) \neq -\infty$, so $pq \neq 0$.

For 2, suppose $pq = pr$ and $p \neq 0$. By Theorem 0.8.1, $p(q - r) = pq - pr = 0$. By part 1 and $p \neq 0$, we get $q - r = 0$, that is, $q = r$. This proves the corollary.

■

Quick check. Let $p, q \in \mathbb{R}[x]$ with $\deg p = 3$ and $\deg q = 2$. What are $\deg(pq)$, $\deg(p + q)$, and $\deg(p^2)$? Can you say what $\deg(p - x^3)$ is?

Solution. By Theorem 0.8.3, $\deg(pq) = 5$ and $\deg(p^2) = 6$. Since the degrees differ, Theorem 0.8.2 gives $\deg(p+q) = 3$. The degree of $p - x^3$ is not determined: if the leading coefficient of p is 1, the x^3 terms cancel and the degree is at most 2 (it could even be $-\infty$ if $p = x^3$); otherwise it is 3.

0.8.4 Polynomials of bounded degree

In linear algebra we constantly need "all polynomials up to a given size". For instance, the quadratics $ax^2 + bx + c$ are described by three numbers, much like points of $\mathbb{R}^3$. So we give such collections names.

Definition 0.8.4. Polynomials of degree at most n

Let F be a field and $n \in \mathbb{N}$. The set of polynomials over F of degree **at most** n is

$$F[x]_{\leq n} = \{p \in F[x] : \deg p \leq n\} = \{a_0 + a_1 x + \cdots + a_n x^n : a_0, \ldots, a_n \in F\}.$$

The two descriptions agree: $\deg p \leq n$ says exactly that every coefficient beyond position n is zero, and the coefficients $a_0, \ldots, a_n$ are then free, any of them allowed to be zero. In particular the zero polynomial lies in $F[x]_{\leq n}$, since $-\infty \leq n$.

Example. Inside $\mathbb{R}[x]_{\leq 2}$

Which of $3x^2 - 2x + 1$, $5x + 7$, -4, 0, and $x^3 - x^3 + x$ lie in $\mathbb{R}[x]_{\leq 2}$? Which lie in $\mathbb{R}[x]_{\leq 0}$?

Solution. All five lie in $\mathbb{R}[x]_{\leq 2}$. Their degrees are 2, 1, 0, $-\infty$, and 1 (the last is the polynomial x, since $x^3 - x^3 = 0$ as polynomials). Only -4 and 0 lie in $\mathbb{R}[x]_{\leq 0}$, which is the set of constant polynomials, 0 included.

Sometimes we need the degree to be exactly n, for instance when we speak of "a quadratic" and mean that the x^2 term is really there.

Definition 0.8.5. Polynomials of degree exactly n

Let F be a field and $n \in \mathbb{N}$. The set of polynomials over F of degree **exactly** n is

$$\begin{aligned} F[x]_{=n} &= \{p \in F[x] : \deg p = n\} \\ &= \{a_0 + a_1 x + \cdots + a_n x^n : a_0, \ldots, a_n \in F,\ a_n \neq 0\}. \end{aligned}$$

Example. Comparing the two notations

For each of $x^2 + 1$, $3x - 2$, and 0, decide whether it lies in $\mathbb{R}[x]_{\leq 2}$ and whether it lies in $\mathbb{R}[x]_{=2}$.

Solution. The polynomial x^2+1 has degree 2, so it lies in both. The polynomial $3x-2$ has degree $1 \le 2$, so it lies in $\mathbb{R}[x]_{\le 2}$ but **not** in $\mathbb{R}[x]_{=2}$. The zero polynomial has degree $-\infty \le 2$, so it lies in $\mathbb{R}[x]_{\le 2}$, but $-\infty \neq 2$, so it is **not** in $\mathbb{R}[x]_{=2}$. In general $F[x]_{\le n} = F[x]_{=n} \cup F[x]_{\le n-1}$ for $n \ge 1$, and the union is disjoint.

The two sets differ in one small clause, $a_n \neq 0$, and that clause changes their behavior completely.

Warning. $F[x]_{\le n}$ is closed under addition, but $F[x]_{=n}$ is **not**. For example x^2+x and $-x^2+1$ both lie in $\mathbb{R}[x]_{=2}$, but their sum $x+1$ has degree 1. The leading terms cancel. Also $F[x]_{=n}$ does not contain 0. In Chapter 1 this is why $F[x]_{\le n}$ is a subspace and $F[x]_{=n}$ is not.

The positive half is Theorem 0.8.2 at work: if $\deg p \le n$ and $\deg q \le n$, then $\deg(p+q) \le \max(\deg p, \deg q) \le n$. Multiplying by a constant $c \in F$ also cannot raise the degree, by Theorem 0.8.3.

0.8.5 Evaluation and roots

A polynomial is not a function, but every polynomial *defines* a function: plug an element of the field in for x.

Definition 0.8.6. Evaluation and roots

Let F be a field, $p = \sum_{k=0}^{N} a_k x^k \in F[x]$, and $c \in F$. The **value** of p at c is the element

$$p(c) = \sum_{k=0}^{N} a_k c^k = a_0 + a_1 c + \cdots + a_N c^N \in F,$$

where $c^0 = 1$. An element $c \in F$ is a **root** of p if $p(c) = 0$. The function $F \to F$, $c \mapsto p(c)$, is the **polynomial function** of p.

The value does not depend on the choice of N: enlarging N adds terms $0 \cdot c^k = 0$ (by Theorem 0.7.1), which do not change the sum. So $p(c)$ is well defined.

For example, $p = x^2 - 3x + 2 \in \mathbb{R}[x]$ has $p(0) = 2$ and $p(1) = 1 - 3 + 2 = 0$, so 1 is a root. Over $\mathbb{F}_3 = \{0, 1, 2\}$, the polynomial $x^3 + 2$ has $1^3 + 2 = 3 = 0$, so 1 is a root; the other two values are $0 + 2 = 2$ and $8 + 2 = 10 = 1$, so 1 is its only root in $\mathbb{F}_3$.

Evaluation turns the operations of $F[x]$ into the operations of F:

Theorem 0.8.5. Evaluation respects sums and products

Let F be a field, $p, q \in F[x]$, and $c \in F$. Then

$$(p+q)(c) = p(c) + q(c), \qquad (pq)(c) = p(c)\, q(c),$$

and the constant polynomial a has value a at c.

Proof. Let $p = \sum_{i=0}^{m} a_i x^i$ and $q = \sum_{j=0}^{n} b_j x^j$, where we choose m and n large enough that all nonzero coefficients are included. The statement about constants is immediate from Definition 0.8.6. For sums, taking $N = \max(m, n)$ and padding with zero coefficients, $(p+q)(c) = \sum_{k=0}^{N}(a_k + b_k)c^k = \sum_{k=0}^{N} a_k c^k + \sum_{k=0}^{N} b_k c^k$ by distributivity in F.

For products, all nonzero coefficients of pq occur at positions $k \le m+n$, so

$$\begin{aligned}(pq)(c) &= \sum_{k=0}^{m+n}\Big(\sum_{i+j=k} a_i b_j\Big)c^k \\ &= \sum_{i=0}^{m}\sum_{j=0}^{n} a_i b_j c^i c^j = \sum_{i=0}^{m}\sum_{j=0}^{n}(a_i c^i)(b_j c^j) \\ &= \Big(\sum_{i=0}^{m} a_i c^i\Big)\Big(\sum_{j=0}^{n} b_j c^j\Big),\end{aligned}$$

where the first equality is Definition 0.8.2, the second uses distributivity and $c^{i+j} = c^i c^j$ and regroups the pairs (i, j), the third uses **commutativity** of multiplication in F to move c^i past b_j, and the last is distributivity again. This shows $(pq)(c) = p(c)q(c)$.

■

Note which step used commutativity. In the next section we will plug a square matrix $\mathbf{A}$ into a polynomial. Products of matrices do not commute in general, but the proof only moves **scalars** past powers of $\mathbf{A}$, and multiplies powers of the same $\mathbf{A}$; both are still allowed for matrices, so the same identity will hold there.

As a first use, a root of a factor is a root of the product: if $q(c) = 0$, then $(pq)(c) = p(c) \cdot 0 = 0$. Conversely, if $(pq)(c) = 0$ then $p(c)q(c) = 0$, so $p(c) = 0$ or $q(c) = 0$ by Theorem 0.7.1. This is how one solves $x^2 - 3x + 2 = (x-1)(x-2) = 0$ at school.

Now we can return to the question from the start of the section.

Example. Different polynomials, same function

Over $\mathbb{F}_2$, compare the polynomials $x^2 + x$ and 0, and the polynomials x^2 and x, both as polynomials and as functions $\mathbb{F}_2 \to \mathbb{F}_2$.

Solution. As polynomials, x^2+x has coefficient sequence $(0, 1, 1, 0, \dots)$ and 0 has $(0, 0, 0, \dots)$. They differ in position 1, so $x^2 + x \neq 0$ by Definition 0.8.1. Its degree is 2.

As functions, $\mathbb{F}_2 = \{0, 1\}$ has two elements, and

$$0^2 + 0 = 0, \qquad 1^2 + 1 = 1 + 1 = 0.$$

So the polynomial function of x^2+x is the zero function, the same as that of 0. Both elements of $\mathbb{F}_2$ are roots of a polynomial of degree 2 that is not the zero polynomial.

Similarly $x^2 \neq x$ as polynomials, but $0^2 = 0$ and $1^2 = 1$, so they define the same function. Indeed, by Theorem 0.8.5, two polynomials define the same function exactly when their difference is a polynomial whose function is zero, and $x^2 - x = x^2 + x$ in $\mathbb{F}_2[x]$.

Warning. **A polynomial is not a function.** Two different polynomials can define the same polynomial function, as $x^2 + x$ and 0 do over $\mathbb{F}_2$. So "$p(c) = 0$ for every $c \in F$" does **not** imply $p = 0$ in general, and we may never read off coefficients or a degree from the values alone. This happens over every finite field: if $F = \{c_1, \dots, c_q\}$, then $(x-c_1)(x-c_2)\cdots(x-c_q)$ has degree q by Theorem 0.8.3, yet vanishes at every c_i by Theorem 0.8.5. Over $\mathbb{F}_p$, Exercise (C2: A nonzero polynomial vanishing everywhere) below gives the shorter example $x^p - x$.

Over an infinite field such as $\mathbb{Q}$, $\mathbb{R}$, or $\mathbb{C}$, the problem disappears: a polynomial whose function is zero is the zero polynomial, so polynomials and polynomial functions can be identified. We prove this in Chapter 3 (in the section on applications of linear systems), once we can count roots. Until then, whenever we say two polynomials are equal we mean their coefficients agree.

0.8.6 Looking ahead

Two tools will be built on this section later.

Remark. **Division with remainder.** Over a field, if $f, g \in F[x]$ and $g \neq 0$, there are unique $q, r \in F[x]$ with $f = qg + r$ and $\deg r < \deg g$. This is long division of polynomials, and the convention $\deg 0 = -\infty$ makes "$\deg r < \deg g$" cover the case $r = 0$. We state it here only for orientation; it is proved in Chapter 6 and is not used before then.

Remark. **Polynomials in a matrix.** In the next section, once matrices and their products are available, we substitute a square matrix $\mathbf{A}$ for x, replacing the constant term a_0 by $a_0\mathbf{I}$. This is how the characteristic and minimal polynomials of later chapters act on matrices.

0.8.7 Exercises

0.8.7.1 A. Check your understanding

Exercise. A1

(a) Define the degree of a polynomial over a field F, including the zero polynomial. State the formula for $\deg(pq)$.

(b) Is $1 + x + x^2 + x^3 + \cdots$, with every coefficient equal to 1, a polynomial over $\mathbb{R}$? Justify your answer.

(c) Determine whether the following statement is true: "for all $p, q \in F[x]$, $\deg(p + q) = \max(\deg p, \deg q)$." Justify your answer.

(d) Determine whether the following statement is true: "if $p \in \mathbb{F}_2[x]$ and $p(c) = 0$ for every $c \in \mathbb{F}_2$, then $p = 0$." Justify your answer.

(e) Explain why $F[x]_{=2}$ is **not** closed under addition, while $F[x]_{\leq 2}$ is.

Solution.

(a) For $p = \sum a_k x^k \neq 0$, $\deg p$ is the largest n with $a_n \neq 0$; and $\deg 0 = -\infty$. For all $p, q \in F[x]$, $\deg(pq) = \deg p + \deg q$ (Theorem 0.8.3).

(b) No. Its coefficient sequence $(1, 1, 1, \dots)$ has infinitely many nonzero terms, so it fails the clause "only finitely many a_k are nonzero" of Definition 0.8.1.

(c) False. Take $p = x$ and $q = -x$. Then $p+q = 0$ has degree $-\infty$, while $\max(\deg p, \deg q) = 1$. Only the inequality $\leq$ holds in general (Theorem 0.8.2).

(d) False. By Example (Different polynomials, same function), $p = x^2 + x$ satisfies $p(0) = p(1) = 0$, but $p \neq 0$.

(e) The sum of x^2 and $-x^2 + 1$, both of degree exactly 2, is 1, of degree 0. For $F[x]_{\leq 2}$: if $\deg p \leq 2$ and $\deg q \leq 2$, then $\deg(p + q) \leq \max(\deg p, \deg q) \leq 2$ by Theorem 0.8.2.

0.8.7.2 B. Practice

Exercise. B1: Sums, products, and degrees

(a) In $\mathbb{R}[x]$, let $p = 2x^3 - x + 4$ and $q = -2x^3 + x^2 + 3$. Compute $p+q$ and pq, and give their degrees. Which inequality or formula from this section does each degree illustrate?

(b) In $\mathbb{F}_3[x]$, compute $(x + 2)(x^2 + x + 1)$. Hence find all roots in $\mathbb{F}_3$ of the product.

Solution.

(a) Adding coefficients gives $p + q = x^2 - x + 7$, of degree 2. Here $\deg p = \deg q = 3$ and the x^3 terms cancel, so the inequality of Theorem 0.8.2 is strict. Expanding,

$$pq = -4x^6 + 2x^5 + 2x^4 - 3x^3 + 4x^2 - 3x + 12,$$

of degree $6 = 3 + 3$, as Theorem 0.8.3 predicts; its leading coefficient is $2 \cdot (-2) = -4$.

(b) Expanding, $(x + 2)(x^2 + x + 1) = x^3 + 3x^2 + 3x + 2$. In $\mathbb{F}_3$ we have $3 = 0$, so the product is $x^3 + 2$. By Theorem 0.8.5 and Theorem 0.7.1, c is a root of the product if and only if $c + 2 = 0$ or $c^2 + c + 1 = 0$. The first gives $c = 1$. For the second, the values at $0, 1, 2$ are 1, $3 = 0$, and $7 = 1$. Hence the only root in $\mathbb{F}_3$ is $c = 1$, which occurs in both factors.

Exercise. B2: Degree of a composition

For $p = \sum_{k=0}^{m} a_k x^k$ and q in $F[x]$, define the **composition** $p \circ q = \sum_{k=0}^{m} a_k q^k \in F[x]$, where $q^0 = 1$.

(a) Prove that if p and q are nonconstant, then $\deg(p \circ q) = \deg p \cdot \deg q$.

(b) Give an example showing the formula fails when q is constant.

Hint: first find $\deg(q^k)$.

Solution.

(a) Let $m = \deg p \geq 1$ and $d = \deg q \geq 1$, so $a_m \neq 0$. By Theorem 0.8.3 and induction on k (Theorem 0.2.7), $\deg(q^k) = kd$ for every $k \in \mathbb{N}$: the case $k = 0$ is $\deg 1 = 0$, and $\deg(q^{k+1}) = \deg(q^k) + \deg q = kd + d$. Hence, again by Theorem 0.8.3, $\deg(a_m q^m) = 0 + md = md$, and for $k < m$, $\deg(a_k q^k) \leq kd < md$ (it is kd or $-\infty$). Let $s = \sum_{k=0}^{m-1} a_k q^k$. Applying Theorem 0.8.2 $m - 1$ times gives $\deg s \leq \max_{k<m} \deg(a_k q^k) < md$. Since

$\deg s \neq \deg(a_m q^m)$, the equality case of Theorem 0.8.2 gives $\deg(p \circ q) = \deg(s + a_m q^m) = md$. This proves $\deg(p \circ q) = \deg p \cdot \deg q$.

(b) Take $p = x - 1$ and $q = 1$. Then $p \circ q = 1 - 1 = 0$ has degree $-\infty$, while $\deg p \cdot \deg q = 1 \cdot 0 = 0$.

Exercise. B3: All small polynomials over $\mathbb{F}_2$

List all polynomials in $\mathbb{F}_2[x]_{\leq 2}$. For each, give its degree and its polynomial function (the values at 0 and 1). Hence determine which of them define the zero function, and explain why every function $\mathbb{F}_2 \to \mathbb{F}_2$ is the polynomial function of exactly two of them.

Solution. A polynomial $a_0 + a_1 x + a_2 x^2$ with $a_i \in \mathbb{F}_2$ has value a_0 at 0 and $a_0 + a_1 + a_2$ at 1. With $2^3 = 8$ choices of coefficients:

polynomial	degree	value at 0	value at 1
0	$-\infty$	0	0
1	0	1	1
x	1	0	1
$x + 1$	1	1	0
x^2	2	0	1
$x^2 + 1$	2	1	0
$x^2 + x$	2	0	0
$x^2 + x + 1$	2	1	1

The zero function requires $a_0 = 0$ and $a_0 + a_1 + a_2 = 0$, that is, $a_0 = 0$ and $a_1 = a_2$. Hence exactly 0 and $x^2 + x$ define the zero function.

A function $f\colon \mathbb{F}_2 \to \mathbb{F}_2$ is determined by the pair $(f(0), f(1))$. Given such a pair (u, v), the polynomial $a_0 + a_1 x + a_2 x^2$ defines f exactly when $a_0 = u$ and $a_1 + a_2 = v - u$. The first fixes a_0, and for each of the two choices of a_2 the second fixes $a_1 = v - u - a_2$. Therefore exactly two of the eight polynomials define f, and the table confirms this.

0.8.7.3 C. Going deeper

Exercise. C1: Roots of a quadratic

Let F be a field.

(a) Let $p = ax^2 + bx + c \in F[x]$ and $r \in F$ with $p(r) = 0$. Prove that $p = (x - r)\big(ax + (b + ar)\big)$.

(b) Deduce that a nonzero polynomial in $F[x]_{\leq 2}$ has at most 2 roots in F.

(c) The same definitions make sense with coefficients in $\mathbb{Z}/8\mathbb{Z}$. Show that $x^2 - 1$ has four roots there, and name the step of (b) that breaks.

Hint for (b): use (a) together with Theorem 0.8.5.

Solution.

(a) Since $p(r) = 0$, we have $ar^2 + br + c = 0$, so $c = -ar^2 - br$. Expanding with Theorem 0.8.1,

$$\begin{aligned}(x - r)(ax + (b + ar)) &= ax^2 + (b + ar)x - arx - r(b + ar)\\ &= ax^2 + bx - (br + ar^2) = ax^2 + bx + c = p.\end{aligned}$$

(b) Let $p \neq 0$ with $\deg p \leq 2$. *Case 1.* $\deg p = 0$. Then $p = c \neq 0$ is constant and $p(r) = c \neq 0$ for every r, so p has no roots. *Case 2.* $\deg p = 1$, say $p = bx + c$ with $b \neq 0$. Then $p(r) = 0$ if and only if $r = -b^{-1}c$, so p has exactly one root. *Case 3.* $\deg p = 2$, say $p = ax^2 + bx + c$ with $a \neq 0$. If p has no root we are done. Otherwise let r be a root. By (a), $p = (x - r)q$ with $q = ax + (b + ar)$, of degree 1 since $a \neq 0$. Let s be any root of p. By Theorem 0.8.5, $0 = p(s) = (s - r)q(s)$, so $s = r$ or $q(s) = 0$ by Theorem 0.7.1. By Case 2, q has exactly one root. Hence every root of p is r or the root of q, and p has at most 2 roots. This proves the claim.

(c) In $\mathbb{Z}/8\mathbb{Z}$, $1^2 = 1$, $3^2 = 9 = 1$, $5^2 = 25 = 1$, and $7^2 = 49 = 1$, so $1, 3, 5, 7$ are four roots of $x^2 - 1$. The argument in (b) breaks at "$(s - r)q(s) = 0$ implies $s - r = 0$ or $q(s) = 0$": for $r = 1$, $s = 3$ we get $(3 - 1)(3 + 1) = 2 \cdot 4 = 0$ with both factors nonzero. That step needs the field property that a product of nonzero elements is nonzero.

Exercise. C2: A nonzero polynomial vanishing everywhere

Let p be a prime and $F = \mathbb{F}_p$.

(a) Fix $a \in F \setminus \{0\}$. Prove that the map $F \setminus \{0\} \to F \setminus \{0\}$, $b \mapsto ab$, is a bijection.

(b) Deduce that $a^{p-1} = 1$ for every $a \in F \setminus \{0\}$.

(c) Hence show that the polynomial $x^p - x \in \mathbb{F}_p[x]$ is not the zero polynomial but that its polynomial function is the zero function.

Hint for (b): multiply together all the elements of $F \setminus \{0\}$ in two ways.

Solution.

(a) The map lands in $F \setminus \{0\}$, since a product of nonzero elements of a field is nonzero (Theorem 0.7.1). It is injective: if $ab = ab'$, multiplying by a^{-1} gives $b = b'$. The set $F \setminus \{0\}$ is finite with $p - 1$ elements, so an injective map from it to itself is surjective by Theorem 0.4.9, hence bijective.

(b) Let P be the product of all $p - 1$ elements of $F \setminus \{0\}$; it is nonzero by Theorem 0.7.1. Since $b \mapsto ab$ is a bijection by (a), the elements ab, as b runs over $F \setminus \{0\}$, are the same elements in a different order. Multiplication in F is commutative and associative, so the product does not depend on the order, and

$$P = \prod_{b \neq 0} (ab) = a^{p-1} \prod_{b \neq 0} b = a^{p-1} P.$$

Multiplying by P^{-1} gives $a^{p-1} = 1$.

(c) The coefficient of x^p in $x^p - x$ is $1 \neq 0$, so the polynomial is nonzero, of degree p. For its function: $0^p - 0 = 0$, and for $a \neq 0$, (b) gives $a^p - a = a \cdot a^{p-1} - a = a - a = 0$.

Hence every element of $\mathbb{F}_p$ is a root, and the polynomial function of $x^p - x$ is the zero function. For $p = 2$ this is $x^2 - x = x^2 + x$, the polynomial of Example (Different polynomials, same function).

Exercise. C3: Which polynomials have inverses?

(a) Let F be a field and $p, q \in F[x]$ with $pq = 1$. Prove that p and q are nonzero constants.

(b) Show that (a) fails with coefficients in $\mathbb{Z}/4\mathbb{Z}$ by computing $(2x + 1)^2$ there.

Solution.

(a) Since $pq = 1 \neq 0$, neither p nor q is the zero polynomial, so $\deg p, \deg q \in \mathbb{N}$. By Theorem 0.8.3, $\deg p + \deg q = \deg 1 = 0$. Two natural numbers with sum 0 are both 0, so $\deg p = \deg q = 0$. Therefore p and q are nonzero constants.

(b) In $\mathbb{Z}/4\mathbb{Z}$, $(2x + 1)^2 = 4x^2 + 4x + 1 = 1$, because $4 = 0$. So $2x + 1$ has degree 1 and is its own inverse. The proof of (a) breaks at Theorem 0.8.3, which fails here because the leading coefficient $2 \cdot 2 = 4 = 0$.

0.9 Matrices

You have probably met matrices as tables of numbers that are multiplied by a strange rule. This section rebuilds them over an arbitrary field F and answers the questions a computational course leaves open. Why is the product defined the way it is? Which rules of school algebra survive, and which fail? What does it mean to divide by a matrix? Everything later in the book, from Gaussian elimination in Chapter 3 to the matrix of a linear map in Chapter 4, rests on the answers.

Throughout, F is a field (Definition 0.7.1). Nothing here uses more about F than the field axioms, so every statement holds over $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$, and $\mathbb{F}_p$ alike.

0.9.1 Matrices and their entries

Definition 0.9.1. Matrix

Let F be a field and $m, n \geq 1$ integers. An $m \times n$ **matrix over** F is a rectangular array

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}$$

with m rows and n columns, whose **entries** a_{ij} lie in F. Formally, it is a function $\{1, \ldots, m\} \times \{1, \ldots, n\} \to F$, $(i, j) \mapsto a_{ij}$. The set of all such matrices is $M_{m \times n}(F)$, and we write $M_n(F) = M_{n \times n}(F)$ for the **square** matrices.

The entry a_{ij} sits in row i and column j: **row first, column second**. We write $\mathbf{A} = (a_{ij})$,

and $(\mathbf{A})_{ij}$ for the (i,j)-entry when the matrix has no letter of its own. Two matrices are **equal** when they have the **same size** and the same entry in every position; this is equality of functions (Definition 0.4.1).

Two sizes deserve their own names. A matrix in $M_{n\times 1}(F)$ is a **column vector**, and we write

$$F^n = M_{n\times 1}(F), \qquad \mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}.$$

In running text we write $\mathbf{x} = (x_1, \ldots, x_n)$ and still mean a column. Columns are printed in bold. A matrix in $M_{1\times n}(F)$ is a **row vector**. For $j = 1, \ldots, n$, the column $\mathbf{e}_j \in F^n$ has 1 in position j and 0 elsewhere.

We will often cut a matrix into its columns. If $\mathbf{a}_j = (a_{1j}, \ldots, a_{mj}) \in F^m$ is the j-th column of $\mathbf{A}$, we write

$$\mathbf{A} = \begin{pmatrix} \mathbf{a}_1 & \mathbf{a}_2 & \cdots & \mathbf{a}_n \end{pmatrix}.$$

The i-th row of $\mathbf{A}$ is the row vector $\begin{pmatrix} a_{i1} & \cdots & a_{in} \end{pmatrix} \in M_{1\times n}(F)$.

Example. Matrices over different fields

Identify the size, the field, and a few entries of each matrix:

$$\mathbf{A} = \begin{pmatrix} 1 & -2 & 0 \\ 3 & 1 & 4 \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} i & 1+i \\ 0 & 2-i \end{pmatrix}, \quad \mathbf{C} = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix}, \quad \mathbf{D} = \begin{pmatrix} 5 \end{pmatrix}.$$

Here $\mathbf{C}$ is meant over $\mathbb{F}_2$. Also write out the 2×2 real matrix $\mathbf{E}$ with $e_{ij} = i + 2j$.

Solution. $\mathbf{A} \in M_{2\times 3}(\mathbb{R})$, with $a_{12} = -2$ and $a_{23} = 4$; its second column is $\mathbf{a}_2 = (-2, 1)$. $\mathbf{B} \in M_2(\mathbb{C})$, with $b_{12} = 1 + i$ and $b_{21} = 0$. $\mathbf{C} \in M_3(\mathbb{F}_2)$; every entry is 0 or 1. $\mathbf{D} \in M_1(\mathbb{R})$ is the degenerate case, a 1×1 matrix. It behaves exactly like the number 5, and 1×1 matrices are how numbers will appear inside matrix formulas. Finally $e_{11} = 3$, $e_{12} = 5$, $e_{21} = 4$, $e_{22} = 6$, so $\mathbf{E} = \begin{pmatrix} 3 & 5 \\ 4 & 6 \end{pmatrix}$. Swapping the roles of i and j would give a different matrix, which is why the order "row, column" matters.

0.9.2 Special matrices

A few shapes of matrix occur so often that they have names. The **main diagonal** of a square matrix $\mathbf{A} \in M_n(F)$ consists of the entries $a_{11}, a_{22}, \ldots, a_{nn}$.

Definition 0.9.2. Zero matrix

The **zero matrix** $\mathbf{O} = \mathbf{O}_{m\times n} \in M_{m\times n}(F)$ is the matrix all of whose entries are 0.

Definition 0.9.3. Identity matrix

The **identity matrix** $\mathbf{I}_n \in M_n(F)$ has entries

$$(\mathbf{I}_n)_{ij} = \delta_{ij} = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases}$$

That is, 1 on the main diagonal and 0 elsewhere. We write $\mathbf{I}$ when the size is clear. The symbol δ_{ij} is the **Kronecker delta**.

The columns of $\mathbf{I}_n$ are exactly $\mathbf{e}_1, \ldots, \mathbf{e}_n$.

Definition 0.9.4. Diagonal matrix

A square matrix $\mathbf{A} \in M_n(F)$ is **diagonal** if $a_{ij} = 0$ **whenever** $i \neq j$. We write $\operatorname{diag}(d_1, \ldots, d_n)$ for the diagonal matrix with diagonal entries $d_1, \ldots, d_n$.

Definition 0.9.5. Upper triangular matrix

A square matrix $\mathbf{A} \in M_n(F)$ is **upper triangular** if $a_{ij} = 0$ **whenever** $i > j$, that is, every entry **strictly below** the main diagonal is zero. It is **lower triangular** if $a_{ij} = 0$ whenever $i < j$.

Example. Zero and identity matrices

Write out $\mathbf{O}_{2\times 3}$, $\mathbf{I}_2$, and $\mathbf{I}_3$. Which of them are diagonal?

Solution.

$$\mathbf{O}_{2\times 3} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad \mathbf{I}_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \mathbf{I}_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

$\mathbf{I}_2 = \operatorname{diag}(1, 1)$ and $\mathbf{I}_3 = \operatorname{diag}(1, 1, 1)$ are diagonal. The matrix $\mathbf{O}_{2\times 3}$ is **not** diagonal, because the definition only applies to square matrices; the square zero matrix $\mathbf{O}_{n\times n} = \operatorname{diag}(0, \ldots, 0)$ is diagonal.

Example. Diagonal and triangular

Classify each matrix as diagonal, upper triangular, lower triangular, or none of these:

$$\mathbf{P} = \begin{pmatrix} 3 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad \mathbf{Q} = \begin{pmatrix} 1 & 2 & 3 \\ 0 & 4 & 5 \\ 0 & 0 & 6 \end{pmatrix}, \quad \mathbf{R} = \begin{pmatrix} 1 & 2 & 3 \\ 0 & 4 & 5 \\ 0 & 7 & 6 \end{pmatrix}.$$

Solution. $\mathbf{P} = \operatorname{diag}(3, -1, 0)$ is diagonal. A zero on the diagonal is allowed: the definition only constrains off-diagonal entries. $\mathbf{P}$ is also both upper and lower triangular, and in general a matrix is diagonal exactly when it is both. $\mathbf{Q}$ is upper triangular but not lower triangular,

since $q_{12} = 2 \neq 0$. $\mathbf{R}$ differs from $\mathbf{Q}$ in one entry, $r_{32} = 7$. All other entries below the diagonal are still zero, but the clause "$a_{ij} = 0$ whenever $i > j$" fails at $(i, j) = (3, 2)$, so $\mathbf{R}$ is not upper triangular. It is not lower triangular either.

0.9.3 Addition and scalar multiplication

Matrices of the same size are added, and multiplied by scalars, one entry at a time.

Definition 0.9.6. Matrix addition

Let $\mathbf{A} = (a_{ij})$ and $\mathbf{B} = (b_{ij})$ be in $M_{m\times n}(F)$, of the **same size**. Their **sum** is $\mathbf{A} + \mathbf{B} \in M_{m\times n}(F)$ with $(\mathbf{A} + \mathbf{B})_{ij} = a_{ij} + b_{ij}$.

Definition 0.9.7. Scalar multiplication

Let $\mathbf{A} = (a_{ij}) \in M_{m\times n}(F)$ and $c \in F$. The **scalar multiple** $c\mathbf{A} \in M_{m\times n}(F)$ has entries $(c\mathbf{A})_{ij} = c\, a_{ij}$.

Definition 0.9.8. Additive inverse

The **additive inverse** of $\mathbf{A} \in M_{m\times n}(F)$ is $-\mathbf{A} = (-1)\mathbf{A}$, with entries $-a_{ij}$. We write $\mathbf{A} - \mathbf{B} = \mathbf{A} + (-\mathbf{B})$.

Matrices of different sizes are never added: $\begin{pmatrix} 1 & 2 \end{pmatrix} + \begin{pmatrix} 1 \\ 2 \end{pmatrix}$ is not defined.

Example. Adding and scaling

Let $\mathbf{A} = \begin{pmatrix} 1 & -2 & 0 \\ 3 & 1 & 4 \end{pmatrix}$ and $\mathbf{B} = \begin{pmatrix} 2 & 2 & -1 \\ 0 & -5 & 1 \end{pmatrix}$ in $M_{2\times 3}(\mathbb{R})$. Compute $\mathbf{A} + \mathbf{B}$, $3\mathbf{A}$, and $\mathbf{A} - 2\mathbf{B}$.

Solution. Entry by entry,

$$\mathbf{A} + \mathbf{B} = \begin{pmatrix} 3 & 0 & -1 \\ 3 & -4 & 5 \end{pmatrix},$$

$$3\mathbf{A} = \begin{pmatrix} 3 & -6 & 0 \\ 9 & 3 & 12 \end{pmatrix},$$

$$\mathbf{A} - 2\mathbf{B} = \begin{pmatrix} 1-4 & -2-4 & 0+2 \\ 3-0 & 1+10 & 4-2 \end{pmatrix} = \begin{pmatrix} -3 & -6 & 2 \\ 3 & 11 & 2 \end{pmatrix}.$$

Because both operations work entry by entry, they inherit the rules of F:

Theorem 0.9.1. Properties of addition and scalar multiplication

Let $\mathbf{A}, \mathbf{B}, \mathbf{C} \in M_{m\times n}(F)$ and $c, d \in F$. Then:

1. $\mathbf{A}+\mathbf{B} = \mathbf{B}+\mathbf{A}$;
2. $(\mathbf{A}+\mathbf{B})+\mathbf{C} = \mathbf{A}+(\mathbf{B}+\mathbf{C})$;
3. $\mathbf{A}+\mathbf{O} = \mathbf{A}$;
4. $\mathbf{A}+(-\mathbf{A}) = \mathbf{O}$;
5. $c(\mathbf{A}+\mathbf{B}) = c\mathbf{A}+c\mathbf{B}$ and $(c+d)\mathbf{A} = c\mathbf{A}+d\mathbf{A}$;
6. $(cd)\mathbf{A} = c(d\mathbf{A})$ and $1\mathbf{A} = \mathbf{A}$.

Proof. Two matrices of the same size are equal when their entries agree, so it suffices to compare (i,j)-entries. For 1, $(\mathbf{A}+\mathbf{B})_{ij} = a_{ij}+b_{ij} = b_{ij}+a_{ij} = (\mathbf{B}+\mathbf{A})_{ij}$, where the middle equality is commutativity of addition in F. For 5, $(c(\mathbf{A}+\mathbf{B}))_{ij} = c(a_{ij}+b_{ij}) = ca_{ij}+cb_{ij} = (c\mathbf{A}+c\mathbf{B})_{ij}$ by distributivity in F. Each remaining identity reduces in the same way to the corresponding field axiom of Definition 0.7.1 in each entry. This proves the theorem. ∎

These eight rules will reappear in Chapter 1 as the axioms of a vector space: $M_{m\times n}(F)$ is one of our standard examples.

0.9.4 Matrix multiplication

Addition was entrywise. The obvious guess for multiplication is also entrywise, but that product is almost useless. The product that matters is the one that records **substitution**, and it looks very different.

Suppose new variables y_1, y_2 are given in terms of x_1, x_2, and z_1, z_2 in terms of y_1, y_2:

$$\begin{aligned} y_1 &= 2x_1 + x_2 & \qquad z_1 &= y_1 + 2y_2 \\ y_2 &= x_1 + 4x_2 & \qquad z_2 &= 3y_1 - y_2 \end{aligned}$$

Record each substitution by its table of coefficients: $\mathbf{B} = \begin{pmatrix} 2 & 1 \\ 1 & 4 \end{pmatrix}$ for $\mathbf{y}$ in terms of $\mathbf{x}$, and $\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 3 & -1 \end{pmatrix}$ for $\mathbf{z}$ in terms of $\mathbf{y}$. Substituting,

$$\begin{aligned} z_1 &= (2x_1 + x_2) + 2(x_1 + 4x_2) \\ &= (1\cdot 2 + 2\cdot 1)x_1 + (1\cdot 1 + 2\cdot 4)x_2 = 4x_1 + 9x_2, \\ z_2 &= 3(2x_1 + x_2) - (x_1 + 4x_2) \\ &= (3\cdot 2 + (-1)\cdot 1)x_1 + (3\cdot 1 + (-1)\cdot 4)x_2 = 5x_1 - x_2. \end{aligned}$$

So $\mathbf{z}$ in terms of $\mathbf{x}$ has coefficient table $\begin{pmatrix} 4 & 9 \\ 5 & -1 \end{pmatrix}$. Look at how each entry arose: the coefficient of x_k in z_i is $a_{i1}b_{1k} + a_{i2}b_{2k}$, row i of $\mathbf{A}$ against column k of $\mathbf{B}$. The entrywise product $\begin{pmatrix} 2 & 2 \\ 3 & -4 \end{pmatrix}$ has nothing to do with the substitution. This forces the definition.

To multiply matrices, run each row of the first against each column of the second: the product of two matrices is the table of the combined substitution.

Definition 0.9.9. Matrix multiplication

Let $\mathbf{A} = (a_{ij}) \in M_{m\times n}(F)$ and $\mathbf{B} = (b_{jk}) \in M_{n\times p}(F)$, so that **the number of columns of $\mathbf{A}$ equals the number of rows of $\mathbf{B}$**. Their **product** is the matrix $\mathbf{AB} \in M_{m\times p}(F)$ with entries

$$(\mathbf{AB})_{ik} = \sum_{j=1}^{n} a_{ij}b_{jk} = a_{i1}b_{1k} + a_{i2}b_{2k} + \cdots + a_{in}b_{nk}$$

for $1 \le i \le m$ and $1 \le k \le p$. If the sizes do not match, $\mathbf{AB}$ is **not defined**.

In words: to get the (i,k)-entry of $\mathbf{AB}$, walk along row i of $\mathbf{A}$ and down column k of $\mathbf{B}$ at the same time, multiply the entries you meet in pairs, and add. The size condition says that the row and the column have the same length n, so the pairing makes sense. The product has as many rows as $\mathbf{A}$ and as many columns as $\mathbf{B}$:

$$\underset{m\times n}{\mathbf{A}}\ \underset{n\times p}{\mathbf{B}} = \underset{m\times p}{\mathbf{AB}}.$$

In particular, for $\mathbf{A} \in M_{m\times n}(F)$ and a column $\mathbf{x} \in F^n$, the product $\mathbf{Ax} \in F^m$ is defined, with i-th entry $a_{i1}x_1 + \cdots + a_{in}x_n$. A system of m linear equations in n unknowns is therefore a single equation $\mathbf{Ax} = \mathbf{b}$.

Example. A product of square matrices

Let $\mathbf{A} = \begin{pmatrix} 2 & -1 \\ 0 & 3 \end{pmatrix}$ and $\mathbf{B} = \begin{pmatrix} 1 & 4 \\ -2 & 1 \end{pmatrix}$ in $M_2(\mathbb{R})$. Compute $\mathbf{AB}$ and $\mathbf{BA}$.

Solution. Row i of $\mathbf{A}$ against column k of $\mathbf{B}$:

$$\mathbf{AB} = \begin{pmatrix} 2\cdot 1 + (-1)(-2) & 2\cdot 4 + (-1)\cdot 1 \\ 0\cdot 1 + 3\cdot(-2) & 0\cdot 4 + 3\cdot 1 \end{pmatrix} = \begin{pmatrix} 4 & 7 \\ -6 & 3 \end{pmatrix},$$

$$\mathbf{BA} = \begin{pmatrix} 1\cdot 2 + 4\cdot 0 & 1\cdot(-1) + 4\cdot 3 \\ -2\cdot 2 + 1\cdot 0 & (-2)(-1) + 1\cdot 3 \end{pmatrix} = \begin{pmatrix} 2 & 11 \\ -4 & 5 \end{pmatrix}.$$

Both products are defined, and they are different.

Example. Products of rectangular matrices

Let $\mathbf{A} = \begin{pmatrix} 1 & 0 & 2 \\ -1 & 3 & 1 \end{pmatrix} \in M_{2\times 3}(\mathbb{R})$ and $\mathbf{B} = \begin{pmatrix} 3 & 1 \\ 2 & 1 \\ 1 & 0 \end{pmatrix} \in M_{3\times 2}(\mathbb{R})$. Determine the sizes of $\mathbf{AB}$ and $\mathbf{BA}$ and compute them. Then compute $\begin{pmatrix} 1 & 2 \end{pmatrix}\begin{pmatrix} 3 \\ 4 \end{pmatrix}$ and $\begin{pmatrix} 3 \\ 4 \end{pmatrix}\begin{pmatrix} 1 & 2 \end{pmatrix}$.

Solution. $\mathbf{AB}$ is $(2\times 3)(3\times 2)$, so it is 2×2; $\mathbf{BA}$ is $(3\times 2)(2\times 3)$, so it is 3×3. Computing

row against column,

$$\mathbf{AB} = \begin{pmatrix} 3+0+2 & 1+0+0 \\ -3+6+1 & -1+3+0 \end{pmatrix} = \begin{pmatrix} 5 & 1 \\ 4 & 2 \end{pmatrix}, \qquad \mathbf{BA} = \begin{pmatrix} 2 & 3 & 7 \\ 1 & 3 & 5 \\ 1 & 0 & 2 \end{pmatrix}.$$

For instance $(\mathbf{BA})_{13} = 3\cdot 2 + 1\cdot 1 = 7$. The last two products show the extreme cases: a row times a column is 1×1, namely $\begin{pmatrix} 1\cdot 3 + 2\cdot 4 \end{pmatrix} = \begin{pmatrix} 11 \end{pmatrix}$, while a column times a row is 2×2, namely $\begin{pmatrix} 3 & 6 \\ 4 & 8 \end{pmatrix}$.

Here is a nonexample by minimal change. Take the same six entries as $\mathbf{B}$, but arrange them in two rows: $\mathbf{C} = \begin{pmatrix} 3 & 2 & 1 \\ 1 & 1 & 0 \end{pmatrix} \in M_{2\times 3}(\mathbb{R})$. The entries are still real numbers and $\mathbf{A}$ is unchanged. But $\mathbf{A}$ has 3 columns and $\mathbf{C}$ has 2 rows, so forming $\mathbf{AC}$ would pair a row of length 3 with a column of length 2. The clause "**the number of columns of A equals the number of rows of B**" fails, and $\mathbf{AC}$ is **not defined**.

Why this definition, and not the entrywise one? Because it is the only rule under which "the table of a combined substitution is the product of the tables". In Chapter 4 the same fact becomes: the matrix of a composition of linear maps is the product of their matrices. Every property of matrix multiplication, good or bad, comes from this.

0.9.5 Three views of a product

The entry formula is how one computes a product by hand, but it hides the structure. Two other descriptions, by columns and by rows, are what we will actually think with. It turns out that multiplying by $\mathbf{A}$ on the left acts on each column of $\mathbf{B}$ separately, and multiplying by $\mathbf{B}$ on the right acts on each row of $\mathbf{A}$ separately:

Theorem 0.9.2. Three views of a product

Let $\mathbf{A} \in M_{m\times n}(F)$ and $\mathbf{B} \in M_{n\times p}(F)$, and write $\mathbf{B} = \begin{pmatrix} \mathbf{b}_1 & \cdots & \mathbf{b}_p \end{pmatrix}$ with columns $\mathbf{b}_k \in F^n$. Then:

1. **(Entries)** $(\mathbf{AB})_{ik} = \sum_{j=1}^{n} a_{ij} b_{jk}$ for all i, k;
2. **(Columns)** the k-th column of $\mathbf{AB}$ is $\mathbf{Ab}_k$, that is,
$$\mathbf{AB} = \begin{pmatrix} \mathbf{Ab}_1 & \mathbf{Ab}_2 & \cdots & \mathbf{Ab}_p \end{pmatrix};$$
3. **(Rows)** the i-th row of $\mathbf{AB}$ is (row i of $\mathbf{A}$)$\,\mathbf{B}$.

Proof. Part 1 is Definition 0.9.9. For part 2, fix k. The column $\mathbf{b}_k$ has entries $b_{1k}, \ldots, b_{nk}$, so by Definition 0.9.9 the i-th entry of $\mathbf{Ab}_k \in F^m$ is $\sum_{j=1}^{n} a_{ij} b_{jk}$. By part 1 this is $(\mathbf{AB})_{ik}$, the i-th entry of the k-th column of $\mathbf{AB}$. Hence the k-th column of $\mathbf{AB}$ is $\mathbf{Ab}_k$.

For part 3, fix i, and let $\mathbf{r} = \begin{pmatrix} a_{i1} & \cdots & a_{in} \end{pmatrix}$ be row i of $\mathbf{A}$. Then $\mathbf{rB}$ is a $1 \times p$ matrix whose k-th entry is $\sum_{j=1}^{n} a_{ij} b_{jk} = (\mathbf{AB})_{ik}$, again by Definition 0.9.9. This is the k-th entry of row i of $\mathbf{AB}$. This proves the theorem.

■

The column view reduces every matrix product to products $\mathbf{Ax}$ of a matrix with a single column. So it pays to understand $\mathbf{Ax}$ well, and here is the fact to understand:

Theorem 0.9.3. $\mathbf{Ax}$ is a combination of the columns of $\mathbf{A}$

Let $\mathbf{A} = \begin{pmatrix} \mathbf{a}_1 & \cdots & \mathbf{a}_n \end{pmatrix} \in M_{m\times n}(F)$ with columns $\mathbf{a}_j \in F^m$, and let $\mathbf{x} = (x_1, \ldots, x_n) \in F^n$. Then

$$\mathbf{Ax} = x_1\mathbf{a}_1 + x_2\mathbf{a}_2 + \cdots + x_n\mathbf{a}_n.$$

In particular $\mathbf{Ae}_j = \mathbf{a}_j$ for each j. Similarly, if $\mathbf{r}_1, \ldots, \mathbf{r}_m$ are the rows of $\mathbf{A}$ and $\mathbf{y} = \begin{pmatrix} y_1 & \cdots & y_m \end{pmatrix}$ is a row vector, then $\mathbf{yA} = y_1\mathbf{r}_1 + \cdots + y_m\mathbf{r}_m$.

Proof. Both sides are in F^m, so we compare i-th entries. By Definition 0.9.9, the i-th entry of $\mathbf{Ax}$ is $\sum_{j=1}^{n} a_{ij}x_j$. On the right, the i-th entry of $x_j\mathbf{a}_j$ is x_ja_{ij} by Definition 0.9.7, so by Definition 0.9.6 the i-th entry of $\sum_j x_j\mathbf{a}_j$ is $\sum_{j=1}^{n} x_ja_{ij}$. These agree because multiplication in F is commutative. Taking $\mathbf{x} = \mathbf{e}_j$, whose only nonzero entry is $x_j = 1$, gives $\mathbf{Ae}_j = \mathbf{a}_j$. The statement for rows follows in the same way, comparing the k-th entries $\sum_i y_ia_{ik}$. This proves the theorem.

■

Read the formula as a sentence: **the entries of $\mathbf{x}$ are the weights, and the columns of $\mathbf{A}$ are what gets weighted.** So $\mathbf{Ax} = \mathbf{b}$ has a solution exactly when $\mathbf{b}$ can be built from the columns of $\mathbf{A}$. This single observation drives much of Chapters 2 and 3. Combined with Theorem 0.9.2, it also says that **every column of $\mathbf{AB}$ is a combination of the columns of $\mathbf{A}$**, with weights taken from the corresponding column of $\mathbf{B}$, and **every row of $\mathbf{AB}$ is a combination of the rows of $\mathbf{B}$**.

Example. One product, three ways

Let

$$\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 3 & -1 \\ 0 & 4 \end{pmatrix} \in M_{3\times 2}(\mathbb{R}), \qquad \mathbf{B} = \begin{pmatrix} 2 & 1 \\ -1 & 1 \end{pmatrix} \in M_2(\mathbb{R}), \qquad \mathbf{x} = \begin{pmatrix} 2 \\ -1 \end{pmatrix}.$$

Compute $\mathbf{Ax}$ as a combination of the columns of $\mathbf{A}$. Then compute $\mathbf{AB}$ by columns, check one row by the row view, and check one entry by the entry formula.

Solution. The columns of $\mathbf{A}$ are $\mathbf{a}_1 = (1, 3, 0)$ and $\mathbf{a}_2 = (2, -1, 4)$. By Theorem 0.9.3,

$$\mathbf{Ax} = 2\mathbf{a}_1 - \mathbf{a}_2 = \begin{pmatrix} 2 \\ 6 \\ 0 \end{pmatrix} - \begin{pmatrix} 2 \\ -1 \\ 4 \end{pmatrix} = \begin{pmatrix} 0 \\ 7 \\ -4 \end{pmatrix}.$$

Columns. The first column of $\mathbf{B}$ is $\mathbf{b}_1 = (2, -1) = \mathbf{x}$, so the first column of $\mathbf{AB}$ is $\mathbf{Ab}_1 = (0, 7, -4)$, just computed. The second column is $\mathbf{b}_2 = (1, 1)$, so $\mathbf{Ab}_2 = \mathbf{a}_1 + \mathbf{a}_2 = (3, 2, 4)$.

Hence

$$\mathbf{AB} = \begin{pmatrix} 0 & 3 \\ 7 & 2 \\ -4 & 4 \end{pmatrix}.$$

Rows. Row 2 of $\mathbf{A}$ is $\begin{pmatrix} 3 & -1 \end{pmatrix}$, and by Theorem 0.9.3 (rows version) $\begin{pmatrix} 3 & -1 \end{pmatrix}\mathbf{B} = 3\begin{pmatrix} 2 & 1 \end{pmatrix} - \begin{pmatrix} -1 & 1 \end{pmatrix} = \begin{pmatrix} 7 & 2 \end{pmatrix}$, which is row 2 of $\mathbf{AB}$.

Entries. $(\mathbf{AB})_{32} = a_{31}b_{12} + a_{32}b_{22} = 0 \cdot 1 + 4 \cdot 1 = 4$, as in the answer.

The column view needed nothing beyond scaling and adding columns of $\mathbf{A}$. When the columns of $\mathbf{B}$ are simple, it is usually the fastest route.

Quick check. Without multiplying out, what is the second column of $\begin{pmatrix} 5 & -1 & 7 \\ 2 & 0 & 3 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & 0 \\ 4 & 1 \end{pmatrix}$?

Solution. The second column of the right factor is $\mathbf{e}_3 = (0, 0, 1)$. By Theorem 0.9.2 and Theorem 0.9.3, the second column of the product is $\mathbf{A}\mathbf{e}_3$, the third column of the left factor: $(7, 3)$.

0.9.6 The rules of multiplication

Matrix multiplication keeps the most important rules of arithmetic: associativity, both distributive laws, and an identity.

Theorem 0.9.4. Properties of matrix multiplication

Let $c \in F$, and let $\mathbf{A}, \mathbf{A}' \in M_{m\times n}(F)$, $\mathbf{B}, \mathbf{B}' \in M_{n\times p}(F)$, and $\mathbf{C} \in M_{p\times q}(F)$. Then:

1. **(Associativity)** $(\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC})$;
2. **(Distributivity)** $\mathbf{A}(\mathbf{B} + \mathbf{B}') = \mathbf{AB} + \mathbf{AB}'$ and $(\mathbf{A} + \mathbf{A}')\mathbf{B} = \mathbf{AB} + \mathbf{A}'\mathbf{B}$;
3. **(Identity)** $\mathbf{I}_m\mathbf{A} = \mathbf{A} = \mathbf{A}\mathbf{I}_n$;
4. **(Scalars)** $c(\mathbf{AB}) = (c\mathbf{A})\mathbf{B} = \mathbf{A}(c\mathbf{B})$;
5. **(Zero)** $\mathbf{O}_{\ell\times m}\mathbf{A} = \mathbf{O}_{\ell\times n}$ and $\mathbf{A}\,\mathbf{O}_{n\times r} = \mathbf{O}_{m\times r}$ for all $\ell, r \geq 1$.

Idea. Each identity is checked entry by entry. The only one with content is associativity. Both sides have (i, l)-entry equal to the double sum $\sum_j \sum_k a_{ij}b_{jk}c_{kl}$, added up in two different orders, so the proof is "expand, swap the order of the finite sums, and fold back up". In the substitution picture it is no surprise: substituting in two stages gives the same result whichever pair of stages we combine first.

Proof. For 1, both sides lie in $M_{m\times q}(F)$. Fix i and l. By Definition 0.9.9 applied twice,

$$\big((\mathbf{AB})\mathbf{C}\big)_{il} = \sum_{k=1}^{p} (\mathbf{AB})_{ik}\, c_{kl} = \sum_{k=1}^{p} \Big(\sum_{j=1}^{n} a_{ij} b_{jk}\Big) c_{kl} = \sum_{k=1}^{p}\sum_{j=1}^{n} a_{ij} b_{jk} c_{kl},$$

where the last step uses distributivity in F. The same computation gives

$$\big(\mathbf{A}(\mathbf{BC})\big)_{il} = \sum_{j=1}^{n} a_{ij} (\mathbf{BC})_{jl} = \sum_{j=1}^{n} a_{ij} \Big(\sum_{k=1}^{p} b_{jk} c_{kl}\Big) = \sum_{j=1}^{n}\sum_{k=1}^{p} a_{ij} b_{jk} c_{kl}.$$

The two double sums add the same np terms in different orders, so they are equal, by commutativity and associativity of addition in F. Hence $(\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC})$.

For 2, $\big(\mathbf{A}(\mathbf{B}+\mathbf{B}')\big)_{ik} = \sum_j a_{ij}(b_{jk}+b'_{jk}) = \sum_j a_{ij}b_{jk} + \sum_j a_{ij}b'_{jk} = (\mathbf{AB}+\mathbf{AB}')_{ik}$ by distributivity in F; the other law is the same with the roles of the factors swapped.

For 3, $(\mathbf{I}_m\mathbf{A})_{ik} = \sum_{j=1}^{m} \delta_{ij} a_{jk} = a_{ik}$, because by Definition 0.9.3 only the term $j = i$ survives. Likewise $(\mathbf{A}\mathbf{I}_n)_{ik} = \sum_j a_{ij}\delta_{jk} = a_{ik}$.

For 4, $(c(\mathbf{AB}))_{ik} = c\sum_j a_{ij}b_{jk} = \sum_j (ca_{ij})b_{jk} = ((c\mathbf{A})\mathbf{B})_{ik}$, and moving c next to b_{jk} instead, using commutativity in F, gives $(\mathbf{A}(c\mathbf{B}))_{ik}$. For 5, every term $0\cdot a_{jk}$ or $a_{ij}\cdot 0$ is 0 by Theorem 0.7.1. This proves the theorem.

■

Associativity means we may write $\mathbf{ABC}$ without brackets, and more generally a product of any number of matrices, as long as consecutive sizes match.

Just as important is what is **missing** from the list. There is no commutative law, and no cancellation law. Each of the following failures will cost you a wrong proof if you forget it.

Warning. **$\mathbf{AB} \neq \mathbf{BA}$ in general.** Take $\mathbf{A} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$ and $\mathbf{B} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}$. Then $\mathbf{AB} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$ but $\mathbf{BA} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$. Before swapping two factors in a proof, you must know that they commute. (And if $\mathbf{A}$ is 2×3 and $\mathbf{B}$ is 3×2, then $\mathbf{AB}$ and $\mathbf{BA}$ do not even have the same size.)

Warning. **A product can be zero with both factors nonzero.** In $M_2(\mathbb{R})$,

$$\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}.$$

So $M_n(F)$ is not a field for $n \geq 2$, and "$\mathbf{AB} = \mathbf{O}$ implies $\mathbf{A} = \mathbf{O}$ or $\mathbf{B} = \mathbf{O}$" is false.

Warning. **Cancellation fails: $\mathbf{AC} = \mathbf{BC}$ does not imply $\mathbf{A} = \mathbf{B}$.** With

$$\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} 1 & 2 \\ 3 & 9 \end{pmatrix}, \quad \mathbf{C} = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix},$$

we get $\mathbf{AC} = \mathbf{BC} = \begin{pmatrix} 1 & 1 \\ 3 & 3 \end{pmatrix}$ although $\mathbf{A} \neq \mathbf{B}$. By the column view, both columns of $\mathbf{C}$ are $\mathbf{e}_1$, so $\mathbf{AC}$ and $\mathbf{BC}$ only see the first columns of $\mathbf{A}$ and $\mathbf{B}$, which agree. Canceling $\mathbf{C}$ is legal only when $\mathbf{C}$ is invertible, as we will see below.

Quick check. Find a nonzero $\mathbf{A} \in M_2(\mathbb{R})$ with $\mathbf{A}^2 = \mathbf{AA} = \mathbf{O}$.

Solution. Take $\mathbf{A} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$. By the column view, $\mathbf{Ae}_1 = \mathbf{0}$ and $\mathbf{Ae}_2 = \mathbf{e}_1$, so the columns of $\mathbf{A}^2$ are $\mathbf{A0} = \mathbf{0}$ and $\mathbf{Ae}_1 = \mathbf{0}$. Hence $\mathbf{A}^2 = \mathbf{O}$ although $\mathbf{A} \neq \mathbf{O}$. This small matrix is the source of a remarkable number of counterexamples in this book; it will reappear often.

0.9.7 The transpose

Rows and columns play symmetric roles in a matrix, and the transpose exchanges them.

Definition 0.9.10. Transpose

Let $\mathbf{A} = (a_{ij}) \in M_{m\times n}(F)$. The **transpose** of $\mathbf{A}$ is the matrix $\mathbf{A}^\top \in M_{n\times m}(F)$ with entries

$$(\mathbf{A}^\top)_{ij} = a_{ji}.$$

In words: row i of $\mathbf{A}$ becomes column i of $\mathbf{A}^\top$, and the size flips from $m \times n$ to $n \times m$. A square matrix is reflected in its main diagonal.

Definition 0.9.11. Symmetric and skew-symmetric matrices

A square matrix $\mathbf{A} \in M_n(F)$ is **symmetric** if $\mathbf{A}^\top = \mathbf{A}$, that is, $a_{ij} = a_{ji}$ for all i, j. It is **skew-symmetric** if $\mathbf{A}^\top = -\mathbf{A}$, that is, $a_{ij} = -a_{ji}$ for all i, j.

Example. Transposes, symmetric and skew-symmetric matrices

Compute the transpose of $\mathbf{A} = \begin{pmatrix} 1 & -2 & 0 \\ 3 & 1 & 4 \end{pmatrix}$. Decide whether

$$\mathbf{S} = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 4 & 5 \\ 3 & 5 & 6 \end{pmatrix}, \qquad \mathbf{K} = \begin{pmatrix} 0 & 2 & -1 \\ -2 & 0 & 3 \\ 1 & -3 & 0 \end{pmatrix}, \qquad \mathbf{S}' = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 4 & 5 \\ 3 & 6 & 6 \end{pmatrix}$$

over $\mathbb{R}$ are symmetric or skew-symmetric.

Solution. The rows of $\mathbf{A}$ become columns: $\mathbf{A}^\top = \begin{pmatrix} 1 & 3 \\ -2 & 1 \\ 0 & 4 \end{pmatrix} \in M_{3\times 2}(\mathbb{R})$.

$\mathbf{S}$ is symmetric: $s_{12} = s_{21} = 2$, $s_{13} = s_{31} = 3$, $s_{23} = s_{32} = 5$. $\mathbf{K}$ is skew-symmetric: $k_{12} = 2 = -k_{21}$, $k_{13} = -1 = -k_{31}$, $k_{23} = 3 = -k_{32}$, and each diagonal entry satisfies $k_{ii} = -k_{ii}$, so it is 0. The matrix $\mathbf{S}'$ differs from $\mathbf{S}$ only in the entry $(3, 2)$. Every other symmetric pair still matches, but $s'_{23} = 5 \neq 6 = s'_{32}$, so $\mathbf{S}'$ is **not** symmetric. A 1×1 matrix is always symmetric, and the identity $\mathbf{I}_n$ and every diagonal matrix are symmetric.

Over $\mathbb{R}$, a skew-symmetric matrix has zero diagonal, since $a_{ii} = -a_{ii}$ forces $2a_{ii} = 0$. Over $\mathbb{F}_2$, where $-1 = 1$, "skew-symmetric" and "symmetric" mean the same thing, and $\mathbf{I}_2$ is skew-symmetric. The characteristic of the field (Definition 0.7.2) matters.

The transpose is compatible with the operations, except that it **reverses** products:

Theorem 0.9.5. Properties of the transpose

Let $c \in F$. For matrices of sizes for which each expression is defined:

1. $(\mathbf{A}^\top)^\top = \mathbf{A}$;
2. $(\mathbf{A} + \mathbf{B})^\top = \mathbf{A}^\top + \mathbf{B}^\top$;
3. $(c\mathbf{A})^\top = c\mathbf{A}^\top$;
4. $(\mathbf{AB})^\top = \mathbf{B}^\top\mathbf{A}^\top$ for $\mathbf{A} \in M_{m\times n}(F)$ and $\mathbf{B} \in M_{n\times p}(F)$.

Idea. The order reverses for a size reason before any computation: $(\mathbf{AB})^\top$ is $p \times m$, and $\mathbf{B}^\top\mathbf{A}^\top$ is $(p \times n)(n \times m) = p \times m$, while $\mathbf{A}^\top\mathbf{B}^\top$ is $(n \times m)(p \times n)$, usually not even defined. Then compare entries.

Proof. For 1, $((\mathbf{A}^\top)^\top)_{ij} = (\mathbf{A}^\top)_{ji} = a_{ij}$ by Definition 0.9.10. For 2 and 3, $((\mathbf{A} + \mathbf{B})^\top)_{ij} = a_{ji} + b_{ji} = (\mathbf{A}^\top + \mathbf{B}^\top)_{ij}$ and $((c\mathbf{A})^\top)_{ij} = ca_{ji} = (c\mathbf{A}^\top)_{ij}$.

For 4, both sides lie in $M_{p\times m}(F)$. Fix $1 \leq i \leq p$ and $1 \leq k \leq m$. By Definition 0.9.10 and Definition 0.9.9,

$$(\mathbf{AB})^\top)_{ik} = (\mathbf{AB})_{ki} = \sum_{j=1}^{n} a_{kj}b_{ji},$$

$$(\mathbf{B}^\top\mathbf{A}^\top)_{ik} = \sum_{j=1}^{n} (\mathbf{B}^\top)_{ij}(\mathbf{A}^\top)_{jk} = \sum_{j=1}^{n} b_{ji}a_{kj}.$$

The two sums agree term by term because multiplication in F is commutative. Hence $(\mathbf{AB})^\top = \mathbf{B}^\top\mathbf{A}^\top$. This proves the theorem.

■

For example, for every $\mathbf{A} \in M_{m\times n}(F)$, the square matrix $\mathbf{A}^\top\mathbf{A}$ is symmetric: by parts 4 and

1, $(\mathbf{A}^\top\mathbf{A})^\top = \mathbf{A}^\top(\mathbf{A}^\top)^\top = \mathbf{A}^\top\mathbf{A}$. Matrices of this form are central to least squares and the singular value decomposition much later on.

0.9.8 The trace

The trace is the simplest number attached to a square matrix, and it has one surprising property.

Definition 0.9.12. Trace

The **trace** of a square matrix $\mathbf{A} = (a_{ij}) \in M_n(F)$ is the sum of its diagonal entries:

$$\operatorname{tr}\mathbf{A} = \sum_{i=1}^{n} a_{ii} = a_{11} + a_{22} + \cdots + a_{nn} \in F.$$

Example. Computing traces

Compute $\operatorname{tr}\mathbf{A}$ for $\mathbf{A} = \begin{pmatrix} 2 & 7 & -1 \\ 0 & -3 & 5 \\ 4 & 1 & 6 \end{pmatrix} \in M_3(\mathbb{R})$, and $\operatorname{tr}\mathbf{I}_n$ over $\mathbb{R}$ and over $\mathbb{F}_2$.

Solution. $\operatorname{tr}\mathbf{A} = 2+(-3)+6 = 5$; the off-diagonal entries play no role. $\operatorname{tr}\mathbf{I}_n = 1+\cdots+1 = n\cdot 1$. Over $\mathbb{R}$ this is n. Over $\mathbb{F}_2$ it is 0 when n is even and 1 when n is odd. The trace is **not** defined for a nonsquare matrix.

Theorem 0.9.6. Properties of the trace

Let $\mathbf{A}, \mathbf{B} \in M_n(F)$ and $c \in F$. Then:

1. $\operatorname{tr}(\mathbf{A}+\mathbf{B}) = \operatorname{tr}\mathbf{A} + \operatorname{tr}\mathbf{B}$ and $\operatorname{tr}(c\mathbf{A}) = c\operatorname{tr}\mathbf{A}$;
2. $\operatorname{tr}(\mathbf{A}^\top) = \operatorname{tr}\mathbf{A}$;
3. $\operatorname{tr}(\mathbf{AB}) = \operatorname{tr}(\mathbf{BA})$. More generally, this holds for all $\mathbf{A} \in M_{m\times n}(F)$ and $\mathbf{B} \in M_{n\times m}(F)$, where $\mathbf{AB}$ is $m \times m$ and $\mathbf{BA}$ is $n \times n$.

Idea. Part 3 is surprising because $\mathbf{AB} \neq \mathbf{BA}$ in general. But the trace only sees the diagonal, and $\operatorname{tr}(\mathbf{AB}) = \sum_i \sum_j a_{ij}b_{ji}$ is a sum over all pairs (i, j) of the product of a_{ij} with the "mirror" entry b_{ji}. That description does not care which matrix is written first.

Proof. Parts 1 and 2 follow from the definitions: the diagonal entries of $\mathbf{A}+\mathbf{B}$, $c\mathbf{A}$, and $\mathbf{A}^\top$ are $a_{ii} + b_{ii}$, ca_{ii}, and a_{ii}.

For 3, let $\mathbf{A} \in M_{m\times n}(F)$ and $\mathbf{B} \in M_{n\times m}(F)$. By Definition 0.9.12 and Definition 0.9.9,

$$\operatorname{tr}(\mathbf{AB}) = \sum_{i=1}^{m}(\mathbf{AB})_{ii} = \sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}b_{ji} = \sum_{j=1}^{n}\sum_{i=1}^{m} b_{ji}a_{ij} = \sum_{j=1}^{n}(\mathbf{BA})_{jj} = \operatorname{tr}(\mathbf{BA}),$$

where the middle equality swaps the order of the finite double sum and uses commutativity

of multiplication in F. This proves the theorem.

■

A first payoff: for all $\mathbf{A}, \mathbf{B} \in M_n(F)$, the matrix $\mathbf{AB} - \mathbf{BA}$ has trace $\operatorname{tr}(\mathbf{AB}) - \operatorname{tr}(\mathbf{BA}) = 0$. So over $\mathbb{R}$ it can never equal $\mathbf{I}_n$, whose trace is n (Exercise (C1: $\mathbf{AB} - \mathbf{BA} = \mathbf{I}$ is impossible, or is it?)). The trace will also turn out to be one of the quantities that does not change when we change coordinates.

It is tempting to conclude that the trace of a product ignores the order of **all** the factors. It does not.

Warning. $\operatorname{tr}(\mathbf{ABC}) \neq \operatorname{tr}(\mathbf{BAC})$ **in general.** Take $\mathbf{A} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$, $\mathbf{B} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$, and $\mathbf{C} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}$. Then $\mathbf{ABC} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$ has trace 1, but $\mathbf{BA} = \mathbf{O}$, so $\mathbf{BAC} = \mathbf{O}$ has trace 0. What **is** true is that a **cyclic** rotation preserves the trace: $\operatorname{tr}(\mathbf{ABC}) = \operatorname{tr}((\mathbf{AB})\mathbf{C}) = \operatorname{tr}(\mathbf{C}(\mathbf{AB})) = \operatorname{tr}(\mathbf{CAB})$ by part 3.

0.9.9 Inverse matrices

In a field we solve $ax = b$ by multiplying by a^{-1}, and this works exactly when $a \neq 0$. We would like to solve a system $\mathbf{Ax} = \mathbf{b}$ the same way, by "dividing by $\mathbf{A}$". What should that mean, and for which $\mathbf{A}$ is it possible?

The naive answer, "every nonzero matrix", fails. Let $\mathbf{N} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \neq \mathbf{O}$. For **any** $\mathbf{B} \in M_2(F)$, the first column of $\mathbf{BN}$ is $\mathbf{B}$ times the first column of $\mathbf{N}$, that is $\mathbf{B0} = \mathbf{0}$, by Theorem 0.9.2. The first column of $\mathbf{I}_2$ is $\mathbf{e}_1 \neq \mathbf{0}$, so $\mathbf{BN} \neq \mathbf{I}_2$: no matrix undoes $\mathbf{N}$ from the left. Dividing needs a genuine two-sided inverse, and not every nonzero matrix has one.

An invertible matrix is a square matrix whose effect can be undone by multiplying, on either side.

Definition 0.9.13. Invertible matrix

A **square** matrix $\mathbf{A} \in M_n(F)$ is **invertible** if there exists $\mathbf{B} \in M_n(F)$ such that

$$\mathbf{AB} = \mathbf{I}_n \quad \textbf{and} \quad \mathbf{BA} = \mathbf{I}_n.$$

Such a $\mathbf{B}$ is called **an inverse** of $\mathbf{A}$. A square matrix that is not invertible is called **singular**.

In words: $\mathbf{A}$ is invertible when some matrix $\mathbf{B}$ of the same size cancels it **from the right** ($\mathbf{AB} = \mathbf{I}_n$) **and** from the left ($\mathbf{BA} = \mathbf{I}_n$). The definition asks for one $\mathbf{B}$ doing both jobs, and it only speaks about square matrices.

Is "an inverse" really "the inverse"? Yes: an inverse, if it exists, is unique. This is part 1 of the next theorem, and it lets us write $\mathbf{A}^{-1}$. The remaining parts are what inverses are for.

Theorem 0.9.7. Properties of inverses

Let $\mathbf{A}, \mathbf{B} \in M_n(F)$.

1. **(Uniqueness)** If $\mathbf{B}$ and $\mathbf{B}'$ are both inverses of $\mathbf{A}$, then $\mathbf{B} = \mathbf{B}'$. We write $\mathbf{A}^{-1}$ for **the** inverse of an invertible $\mathbf{A}$.
2. If $\mathbf{A}$ is invertible, then so is $\mathbf{A}^{-1}$, and $(\mathbf{A}^{-1})^{-1} = \mathbf{A}$.
3. If $\mathbf{A}$ and $\mathbf{B}$ are invertible, then so is $\mathbf{AB}$, and $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$.
4. If $\mathbf{A}$ is invertible, then so is $\mathbf{A}^\top$, and $(\mathbf{A}^\top)^{-1} = (\mathbf{A}^{-1})^\top$.
5. If $\mathbf{A}$ is invertible and $c \in F$ is **nonzero**, then $c\mathbf{A}$ is invertible, and $(c\mathbf{A})^{-1} = c^{-1}\mathbf{A}^{-1}$.
6. **(Cancellation)** If $\mathbf{A}$ is invertible, then $\mathbf{AX} = \mathbf{AY}$ implies $\mathbf{X} = \mathbf{Y}$ for all $\mathbf{X}, \mathbf{Y} \in M_{n\times p}(F)$, and $\mathbf{XA} = \mathbf{YA}$ implies $\mathbf{X} = \mathbf{Y}$ for all $\mathbf{X}, \mathbf{Y} \in M_{p\times n}(F)$.
7. **(Solving)** If $\mathbf{A}$ is invertible, then for every $\mathbf{b} \in F^n$ the equation $\mathbf{Ax} = \mathbf{b}$ has **exactly one** solution $\mathbf{x} \in F^n$, namely $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$.

Idea. Uniqueness is the standard trick for "prove unique by assuming two": sandwich $\mathbf{B}'$ between $\mathbf{B}$ and $\mathbf{A}$, and use associativity to collapse the product both ways. For the other parts the candidate inverse is written in the statement, so we only multiply it out on **both** sides, as the definition demands. The order in part 3 is the "socks and shoes" rule: to undo "first $\mathbf{B}$, then $\mathbf{A}$", undo $\mathbf{A}$ first.

Proof. For 1, suppose $\mathbf{AB} = \mathbf{BA} = \mathbf{I}_n$ and $\mathbf{AB}' = \mathbf{B}'\mathbf{A} = \mathbf{I}_n$. By Theorem 0.9.4 (identity and associativity),

$$\mathbf{B} = \mathbf{BI}_n = \mathbf{B}(\mathbf{AB}') = (\mathbf{BA})\mathbf{B}' = \mathbf{I}_n\mathbf{B}' = \mathbf{B}'.$$

For 2, the equations $\mathbf{AA}^{-1} = \mathbf{I}_n$ and $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}_n$ say precisely that $\mathbf{A}$ is an inverse of $\mathbf{A}^{-1}$. By part 1 it is the inverse.

For 3, by associativity,

$$(\mathbf{AB})(\mathbf{B}^{-1}\mathbf{A}^{-1}) = \mathbf{A}(\mathbf{BB}^{-1})\mathbf{A}^{-1} = \mathbf{AI}_n\mathbf{A}^{-1} = \mathbf{AA}^{-1} = \mathbf{I}_n,$$

and similarly $(\mathbf{B}^{-1}\mathbf{A}^{-1})(\mathbf{AB}) = \mathbf{B}^{-1}(\mathbf{A}^{-1}\mathbf{A})\mathbf{B} = \mathbf{B}^{-1}\mathbf{B} = \mathbf{I}_n$. Hence $\mathbf{B}^{-1}\mathbf{A}^{-1}$ is an inverse of $\mathbf{AB}$, and by part 1 it is $(\mathbf{AB})^{-1}$.

For 4, by Theorem 0.9.5, $\mathbf{A}^\top(\mathbf{A}^{-1})^\top = (\mathbf{A}^{-1}\mathbf{A})^\top = \mathbf{I}_n^\top = \mathbf{I}_n$ and $(\mathbf{A}^{-1})^\top\mathbf{A}^\top = (\mathbf{AA}^{-1})^\top = \mathbf{I}_n$, where $\mathbf{I}_n^\top = \mathbf{I}_n$ since $\mathbf{I}_n$ is symmetric. Hence $(\mathbf{A}^{-1})^\top$ is the inverse of $\mathbf{A}^\top$.

For 5, since $c \neq 0$, the inverse c^{-1} exists in F. By Theorem 0.9.4 (scalars), $(c\mathbf{A})(c^{-1}\mathbf{A}^{-1}) = (cc^{-1})(\mathbf{AA}^{-1}) = \mathbf{I}_n$, and similarly in the other order.

For 6, suppose $\mathbf{AX} = \mathbf{AY}$. Multiplying on the left by $\mathbf{A}^{-1}$ and using associativity, $\mathbf{X} = \mathbf{I}_n\mathbf{X} = (\mathbf{A}^{-1}\mathbf{A})\mathbf{X} = \mathbf{A}^{-1}(\mathbf{AX}) = \mathbf{A}^{-1}(\mathbf{AY}) = (\mathbf{A}^{-1}\mathbf{A})\mathbf{Y} = \mathbf{Y}$. The statement for $\mathbf{XA} = \mathbf{YA}$ follows by multiplying on the right by $\mathbf{A}^{-1}$.

For 7, $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$ is a solution, since $\mathbf{A}(\mathbf{A}^{-1}\mathbf{b}) = (\mathbf{AA}^{-1})\mathbf{b} = \mathbf{b}$. If $\mathbf{x}$ is any solution, then $\mathbf{Ax} = \mathbf{b} = \mathbf{A}(\mathbf{A}^{-1}\mathbf{b})$, and part 6 (with $p = 1$) gives $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$. This proves the theorem.

■

Part 6 is the repair of the failed cancellation law from the warnings above: **you may cancel an invertible matrix**, on the side where it stands, and nothing else.

Example. Some invertible matrices

Show that each of the following is invertible and find its inverse: $\mathbf{I}_n$; a diagonal matrix $\mathbf{D} = \operatorname{diag}(d_1, \ldots, d_n)$ with **all** $d_i \neq 0$; $\mathbf{A} = \begin{pmatrix} 2 & 1 \\ 5 & 3 \end{pmatrix} \in M_2(\mathbb{R})$; and $\mathbf{U} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \in M_2(\mathbb{F}_2)$.

Solution. $\mathbf{I}_n\mathbf{I}_n = \mathbf{I}_n$, so $\mathbf{I}_n$ is its own inverse. This is the degenerate case: $\mathbf{I}_n$ plays the role that 1 plays among numbers.

For $\mathbf{D}$, let $\mathbf{E} = \operatorname{diag}(d_1^{-1}, \ldots, d_n^{-1})$, which exists since every $d_i \neq 0$. Then $(\mathbf{DE})_{ik} = \sum_j d_{ij}e_{jk}$, and only $j = i = k$ can contribute, so $\mathbf{DE} = \operatorname{diag}(d_1d_1^{-1}, \ldots, d_nd_n^{-1}) = \mathbf{I}_n$; likewise $\mathbf{ED} = \mathbf{I}_n$. Hence $\mathbf{D}^{-1} = \mathbf{E}$. For $n = 1$ this says the 1×1 matrix (a) is invertible when $a \neq 0$, with inverse (a^{-1}), exactly as in the field.

For $\mathbf{A}$, let $\mathbf{B} = \begin{pmatrix} 3 & -1 \\ -5 & 2 \end{pmatrix}$. Then

$$\mathbf{AB} = \begin{pmatrix} 6-5 & -2+2 \\ 15-15 & -5+6 \end{pmatrix} = \mathbf{I}_2, \qquad \mathbf{BA} = \begin{pmatrix} 6-5 & 3-3 \\ -10+10 & -5+6 \end{pmatrix} = \mathbf{I}_2,$$

so $\mathbf{A}^{-1} = \mathbf{B}$. (Where $\mathbf{B}$ came from is explained by Theorem 0.9.8 below.)

For $\mathbf{U}$ over $\mathbb{F}_2$, $\mathbf{U}^2 = \begin{pmatrix} 1 & 1+1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, since $1 + 1 = 0$. So $\mathbf{U}$ is its own inverse. Over $\mathbb{R}$ the same array would have inverse $\begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix}$: the answer depends on the field.

Now a nonexample by minimal change. The matrix $\operatorname{diag}(1, 2)$ is invertible, with inverse $\operatorname{diag}(1, \frac{1}{2})$. Change the entry 2 to 0: $\mathbf{A} = \operatorname{diag}(1, 0)$ is still diagonal and still nonzero. But for any $\mathbf{B} \in M_2(F)$, the second column of $\mathbf{BA}$ is $\mathbf{B}$ times the second column of $\mathbf{A}$, which is $\mathbf{B0} = \mathbf{0}$, while the second column of $\mathbf{I}_2$ is $\mathbf{e}_2$. So $\mathbf{BA} \neq \mathbf{I}_2$ for every $\mathbf{B}$: the clause "**there exists B** with … $\mathbf{BA} = \mathbf{I}_n$" fails, and $\operatorname{diag}(1, 0)$ is singular. The same argument shows that any square matrix with a zero column is singular, and in particular so is every zero matrix $\mathbf{O}_{n\times n}$.

Why does the definition insist on a **square** matrix and on **both** equations? For nonsquare matrices one-sided inverses can exist without the other side: with $\mathbf{C} = \begin{pmatrix} 1 & 0 \end{pmatrix} \in M_{1\times 2}(\mathbb{R})$ and $\mathbf{D} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$, we get $\mathbf{CD} = \begin{pmatrix} 1 \end{pmatrix} = \mathbf{I}_1$ but $\mathbf{DC} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \neq \mathbf{I}_2$. Demanding both equations makes the inverse unique (part 1) and makes both cancellation laws (part 6) hold.

Warning. **For square matrices, $\mathbf{AB} = \mathbf{I}_n$ already implies $\mathbf{BA} = \mathbf{I}_n$. This is true, but it is not obvious, and we do not know it yet.** Nothing in this chapter proves it: the entries of $\mathbf{AB}$ and $\mathbf{BA}$ are different sums. It is proved in Chapter 3, using row reduction. Until then, to show that $\mathbf{B}$ is the inverse of $\mathbf{A}$, check **both** $\mathbf{AB} = \mathbf{I}_n$ and $\mathbf{BA} = \mathbf{I}_n$. (The nonsquare example above shows that some argument is needed.)

For 2×2 matrices, invertibility is decided by one number.

Theorem 0.9.8. Inverse of a 2×2 matrix

Let $\mathbf{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in M_2(F)$. Then $\mathbf{A}$ is invertible **if and only if** $ad - bc \neq 0$. In that case,

$$\mathbf{A}^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}.$$

Idea. Where does the formula come from? Try to kill the off-diagonal entries of $\mathbf{A}$ by multiplying with a matrix built from its own entries. Swapping a and d and negating b and c gives $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix} = (ad - bc)\mathbf{I}_2$. If $ad - bc \neq 0$, divide by it. If $ad - bc = 0$, the same computation shows that $\mathbf{A}$ sends a nonzero column to $\mathbf{0}$, and an invertible matrix cannot do that, since its inverse would have to send $\mathbf{0}$ back to a nonzero column.

Proof. Write $\Delta = ad - bc$ and $\tilde{\mathbf{A}} = \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$. By Definition 0.9.9 and commutativity of multiplication in F,

$$\mathbf{A}\tilde{\mathbf{A}} = \begin{pmatrix} ad - bc & -ab + ba \\ cd - dc & -cb + da \end{pmatrix} = \Delta \mathbf{I}_2,$$

$$\tilde{\mathbf{A}}\mathbf{A} = \begin{pmatrix} da - bc & db - bd \\ -ca + ac & -cb + ad \end{pmatrix} = \Delta \mathbf{I}_2.$$

($\Leftarrow$) Suppose $\Delta \neq 0$, so Δ^{-1} exists in F. By Theorem 0.9.4 (scalars), $\mathbf{A}(\Delta^{-1}\tilde{\mathbf{A}}) = \Delta^{-1}(\mathbf{A}\tilde{\mathbf{A}}) = \Delta^{-1}\Delta\mathbf{I}_2 = \mathbf{I}_2$, and likewise $(\Delta^{-1}\tilde{\mathbf{A}})\mathbf{A} = \mathbf{I}_2$. Hence $\mathbf{A}$ is invertible with $\mathbf{A}^{-1} = \Delta^{-1}\tilde{\mathbf{A}}$, which is the stated formula.

($\Rightarrow$) We prove the contrapositive. Suppose $\Delta = 0$. We first find $\mathbf{x} \in F^2$ with $\mathbf{x} \neq \mathbf{0}$ and $\mathbf{A}\mathbf{x} = \mathbf{0}$.

Case 1. $\mathbf{A} = \mathbf{O}$. Take $\mathbf{x} = \mathbf{e}_1$.

Case 2. $c \neq 0$ or $d \neq 0$. Take $\mathbf{x} = (d, -c) \neq \mathbf{0}$. Then $\mathbf{A}\mathbf{x} = (ad - bc,\ cd - dc) = (\Delta, 0) = \mathbf{0}$.

Case 3. $c = d = 0$ but $a \neq 0$ or $b \neq 0$. Take $\mathbf{x} = (b, -a) \neq \mathbf{0}$. Then $\mathbf{A}\mathbf{x} = (ab - ba,\ cb - da) = (0, 0) = \mathbf{0}$, since $c = d = 0$.

These cases cover every $\mathbf{A}$. Now suppose, for a contradiction, that $\mathbf{A}$ is invertible. Then by Theorem 0.9.4,

$$\mathbf{x} = \mathbf{I}_2\mathbf{x} = (\mathbf{A}^{-1}\mathbf{A})\mathbf{x} = \mathbf{A}^{-1}(\mathbf{A}\mathbf{x}) = \mathbf{A}^{-1}\mathbf{0} = \mathbf{0},$$

contradicting $\mathbf{x} \neq \mathbf{0}$. Hence $\mathbf{A}$ is not invertible. This proves the theorem.

■

The number $ad - bc$ will return in Chapter 7 as the determinant of $\mathbf{A}$. Here it arose from a single multiplication, and the proof used nothing beyond the field axioms, so the theorem holds

over every field. For instance, $\begin{pmatrix} 2 & 1 \\ 5 & 3 \end{pmatrix}$ has $ad - bc = 6 - 5 = 1$, which explains the inverse found in Example (Some invertible matrices); and $\begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix}$ has $ad - bc = 0$, so it is singular, with $\mathbf{A}(2, -1) = \mathbf{0}$.

0.9.10 Powers and polynomials of a matrix

A square matrix can be multiplied by itself, so we can form powers, and then plug the matrix into a polynomial (Definition 0.8.1). This is how polynomials act on matrices throughout the book.

Definition 0.9.14. Powers and polynomials of a matrix

Let $\mathbf{A} \in M_n(F)$. The **powers** of $\mathbf{A}$ are defined recursively by

$$\mathbf{A}^0 = \mathbf{I}_n, \qquad \mathbf{A}^{k+1} = \mathbf{A}^k \mathbf{A} \quad (k \in \mathbb{N}).$$

For a polynomial $p = a_0 + a_1 x + \cdots + a_N x^N \in F[x]$, the matrix $p(\mathbf{A})$ is

$$p(\mathbf{A}) = a_0 \mathbf{I}_n + a_1 \mathbf{A} + a_2 \mathbf{A}^2 + \cdots + a_N \mathbf{A}^N \in M_n(F).$$

Two points in this definition deserve attention. The **constant term** a_0 becomes $a_0 \mathbf{I}_n$, because a_0 alone is a number and cannot be added to matrices; this matches $x^0 = 1$ with $\mathbf{A}^0 = \mathbf{I}_n$. And $p(\mathbf{A})$ is well defined, because enlarging N only adds terms $0 \cdot \mathbf{A}^k = \mathbf{O}$, exactly as for evaluation at a number (Definition 0.8.6).

The next theorem says that plugging in a matrix respects sums and products, just as plugging in a number does (Theorem 0.8.5).

Theorem 0.9.9. Polynomials in one matrix

Let $\mathbf{A} \in M_n(F)$ and $p, q \in F[x]$. Then:

1. $\mathbf{A}^j \mathbf{A}^k = \mathbf{A}^{j+k}$ for all $j, k \in \mathbb{N}$;
2. $(p + q)(\mathbf{A}) = p(\mathbf{A}) + q(\mathbf{A})$ and $(pq)(\mathbf{A}) = p(\mathbf{A})\, q(\mathbf{A})$;
3. $p(\mathbf{A})\, q(\mathbf{A}) = q(\mathbf{A})\, p(\mathbf{A})$.

Idea. In the proof of Theorem 0.8.5, the only delicate step moved c^i past a coefficient b_j. Here the coefficients are scalars, and scalars can be moved freely through matrix products. The other step, $c^i c^j = c^{i+j}$, becomes part 1, which is associativity plus induction. Part 3 is then free, because $pq = qp$ in $F[x]$.

Proof. For 1, fix j and use induction on k (Theorem 0.2.7). For $k = 0$, $\mathbf{A}^j \mathbf{A}^0 = \mathbf{A}^j \mathbf{I}_n = \mathbf{A}^j$. If $\mathbf{A}^j \mathbf{A}^k = \mathbf{A}^{j+k}$, then by Definition 0.9.14 and associativity (Theorem 0.9.4),

$$\mathbf{A}^j \mathbf{A}^{k+1} = \mathbf{A}^j (\mathbf{A}^k \mathbf{A}) = (\mathbf{A}^j \mathbf{A}^k) \mathbf{A} = \mathbf{A}^{j+k} \mathbf{A} = \mathbf{A}^{j+k+1}.$$

For 2, let $p = \sum_{i=0}^{m} a_i x^i$ and $q = \sum_{j=0}^{l} b_j x^j$. The statement for sums follows from Theorem 0.9.1, since $(a_i + b_i)\mathbf{A}^i = a_i\mathbf{A}^i + b_i\mathbf{A}^i$. For products, by Definition 0.8.2 the coefficient of x^k in pq is $\sum_{i+j=k} a_i b_j$, so

$$(pq)(\mathbf{A}) = \sum_k \Big(\sum_{i+j=k} a_i b_j \Big) \mathbf{A}^k = \sum_{i=0}^{m} \sum_{j=0}^{l} (a_i b_j) \mathbf{A}^i \mathbf{A}^j$$
$$= \Big(\sum_{i=0}^{m} a_i \mathbf{A}^i \Big) \Big(\sum_{j=0}^{l} b_j \mathbf{A}^j \Big) = p(\mathbf{A})\, q(\mathbf{A}),$$

where the second equality uses part 1 and Theorem 0.9.1, and the third uses distributivity and the scalar rule $(a_i\mathbf{A}^i)(b_j\mathbf{A}^j) = (a_i b_j)\mathbf{A}^i\mathbf{A}^j$ from Theorem 0.9.4.

For 3, $pq = qp$ in $F[x]$ by Theorem 0.8.1. Hence by part 2, $p(\mathbf{A})q(\mathbf{A}) = (pq)(\mathbf{A}) = (qp)(\mathbf{A}) = q(\mathbf{A})p(\mathbf{A})$. This proves the theorem.

■

Example. Plugging a matrix into a polynomial

Let $\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 0 & 3 \end{pmatrix} \in M_2(\mathbb{R})$ and $p = x^2 - 4x + 3$. Compute $p(\mathbf{A})$ directly, and again using the factorization $p = (x-1)(x-3)$.

Solution. First $\mathbf{A}^2 = \begin{pmatrix} 1 & 2 \\ 0 & 3 \end{pmatrix}\begin{pmatrix} 1 & 2 \\ 0 & 3 \end{pmatrix} = \begin{pmatrix} 1 & 8 \\ 0 & 9 \end{pmatrix}$. Hence

$$p(\mathbf{A}) = \mathbf{A}^2 - 4\mathbf{A} + 3\mathbf{I}_2 = \begin{pmatrix} 1 & 8 \\ 0 & 9 \end{pmatrix} - \begin{pmatrix} 4 & 8 \\ 0 & 12 \end{pmatrix} + \begin{pmatrix} 3 & 0 \\ 0 & 3 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}.$$

The constant term contributed $3\mathbf{I}_2$, not "3". By Theorem 0.9.9, $p(\mathbf{A}) = (\mathbf{A} - \mathbf{I}_2)(\mathbf{A} - 3\mathbf{I}_2)$, and indeed

$$(\mathbf{A} - \mathbf{I}_2)(\mathbf{A} - 3\mathbf{I}_2) = \begin{pmatrix} 0 & 2 \\ 0 & 2 \end{pmatrix}\begin{pmatrix} -2 & 2 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}.$$

Note that neither factor is zero; this is the second warning of the multiplication part in action. That the diagonal entries $1, 3$ of $\mathbf{A}$ are the roots of p is no accident, as we will see with eigenvalues.

Theorem 0.9.9 is about polynomials in **one** matrix. Identities with **two** different matrices are another matter.

Warning. **School identities fail for matrices that do not commute.** Expanding with distributivity, $(\mathbf{A}+\mathbf{B})^2 = \mathbf{A}^2 + \mathbf{AB} + \mathbf{BA} + \mathbf{B}^2$, and this equals $\mathbf{A}^2 + 2\mathbf{AB} + \mathbf{B}^2$ only when $\mathbf{AB} = \mathbf{BA}$. With $\mathbf{A} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$ and $\mathbf{B} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}$: $(\mathbf{A}+\mathbf{B})^2 = \mathbf{I}_2$, but $\mathbf{A}^2 + 2\mathbf{AB} + \mathbf{B}^2 = \begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix}$. Before using a polynomial identity valid for numbers, check that all the matrices

involved commute; polynomials in a single $\mathbf{A}$ always do.

0.9.11 Two equivalence relations on matrices

We close with two relations on matrices that the rest of the book studies in depth. Both are equivalence relations (Definition 0.5.2), and the proofs are short exercises in using Theorem 0.9.7.

Example. Similarity

For $\mathbf{A}, \mathbf{B} \in M_n(F)$, say $\mathbf{A}$ is **similar** to $\mathbf{B}$, written $\mathbf{A} \sim \mathbf{B}$, if there is an **invertible** $\mathbf{P} \in M_n(F)$ with $\mathbf{B} = \mathbf{P}^{-1}\mathbf{A}\mathbf{P}$. Prove that similarity is an equivalence relation on $M_n(F)$. Then check that $\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 0 & 3 \end{pmatrix}$ is similar to $\operatorname{diag}(1,3)$ using $\mathbf{P} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$, and show that $\operatorname{diag}(1,2)$ is **not** similar to $\mathbf{I}_2$ over $\mathbb{R}$.

Solution. We check the three conditions of Definition 0.5.2.

Reflexive. Let $\mathbf{A} \in M_n(F)$. The identity $\mathbf{I}_n$ is invertible with $\mathbf{I}_n^{-1} = \mathbf{I}_n$ (Example (Some invertible matrices)), and $\mathbf{I}_n^{-1}\mathbf{A}\mathbf{I}_n = \mathbf{A}$. Hence $\mathbf{A} \sim \mathbf{A}$.

Symmetric. Suppose $\mathbf{A} \sim \mathbf{B}$, so $\mathbf{B} = \mathbf{P}^{-1}\mathbf{A}\mathbf{P}$ with $\mathbf{P}$ invertible. Multiplying on the left by $\mathbf{P}$ and on the right by $\mathbf{P}^{-1}$, and using associativity, $\mathbf{P}\mathbf{B}\mathbf{P}^{-1} = (\mathbf{P}\mathbf{P}^{-1})\mathbf{A}(\mathbf{P}\mathbf{P}^{-1}) = \mathbf{A}$. By Theorem 0.9.7 (part 2), $\mathbf{Q} = \mathbf{P}^{-1}$ is invertible with $\mathbf{Q}^{-1} = \mathbf{P}$, so $\mathbf{A} = \mathbf{Q}^{-1}\mathbf{B}\mathbf{Q}$. Hence $\mathbf{B} \sim \mathbf{A}$.

Transitive. Suppose $\mathbf{A} \sim \mathbf{B}$ and $\mathbf{B} \sim \mathbf{C}$, so $\mathbf{B} = \mathbf{P}^{-1}\mathbf{A}\mathbf{P}$ and $\mathbf{C} = \mathbf{Q}^{-1}\mathbf{B}\mathbf{Q}$ with $\mathbf{P}, \mathbf{Q}$ invertible. Then

$$\mathbf{C} = \mathbf{Q}^{-1}(\mathbf{P}^{-1}\mathbf{A}\mathbf{P})\mathbf{Q} = (\mathbf{Q}^{-1}\mathbf{P}^{-1})\mathbf{A}(\mathbf{P}\mathbf{Q}) = (\mathbf{P}\mathbf{Q})^{-1}\mathbf{A}(\mathbf{P}\mathbf{Q}),$$

where the last step is Theorem 0.9.7 (part 3), which also says $\mathbf{P}\mathbf{Q}$ is invertible. Hence $\mathbf{A} \sim \mathbf{C}$. This shows that similarity is an equivalence relation.

For the numerical part, $\mathbf{P}^{-1} = \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix}$ by Theorem 0.9.8 (here $ad - bc = 1$), and

$$\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & 2 \\ 0 & 3 \end{pmatrix}\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ 0 & 3 \end{pmatrix}\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix}.$$

Finally, if $\mathbf{P}^{-1}\mathbf{I}_2\mathbf{P} = \operatorname{diag}(1,2)$ for some invertible $\mathbf{P}$, then $\operatorname{diag}(1,2) = \mathbf{P}^{-1}\mathbf{P} = \mathbf{I}_2$, which is false. (Alternatively: similar matrices have equal traces by Theorem 0.9.6, and $\operatorname{tr}\operatorname{diag}(1,2) = 3 \neq 2 = \operatorname{tr}\mathbf{I}_2$.) Hence $\operatorname{diag}(1,2) \not\sim \mathbf{I}_2$.

In Chapter 4 we will see that similar matrices describe the same linear map in two different coordinate systems. That is why a quantity like the trace, which similar matrices share, is really a property of the map and not of the matrix. Deciding when two matrices are similar is one of the main problems of this book.

Example. Row equivalence

For $\mathbf{A}, \mathbf{B} \in M_{m\times n}(F)$, say $\mathbf{A}$ is **row equivalent** to $\mathbf{B}$ if there is an **invertible** $\mathbf{E} \in M_m(F)$ with $\mathbf{B} = \mathbf{EA}$. Prove that row equivalence is an equivalence relation on $M_{m\times n}(F)$. Then, with $\mathbf{E} = \begin{pmatrix} 1 & 0 \\ -2 & 1 \end{pmatrix}$, compute $\mathbf{EA}$ for $\mathbf{A} = \begin{pmatrix} 1 & 2 & 1 \\ 2 & 5 & 0 \end{pmatrix}$ and describe what $\mathbf{E}$ did to the rows of $\mathbf{A}$.

Solution. Reflexive. $\mathbf{A} = \mathbf{I}_m\mathbf{A}$ and $\mathbf{I}_m$ is invertible, so $\mathbf{A}$ is row equivalent to itself.

Symmetric. Suppose $\mathbf{B} = \mathbf{EA}$ with $\mathbf{E}$ invertible. Multiplying on the left by $\mathbf{E}^{-1}$, $\mathbf{E}^{-1}\mathbf{B} = (\mathbf{E}^{-1}\mathbf{E})\mathbf{A} = \mathbf{A}$, and $\mathbf{E}^{-1}$ is invertible by Theorem 0.9.7 (part 2). So $\mathbf{B}$ is row equivalent to $\mathbf{A}$.

Transitive. Suppose $\mathbf{B} = \mathbf{EA}$ and $\mathbf{C} = \mathbf{E}'\mathbf{B}$ with $\mathbf{E}, \mathbf{E}'$ invertible. Then $\mathbf{C} = \mathbf{E}'(\mathbf{EA}) = (\mathbf{E}'\mathbf{E})\mathbf{A}$ by associativity, and $\mathbf{E}'\mathbf{E}$ is invertible by Theorem 0.9.7 (part 3). So $\mathbf{A}$ is row equivalent to $\mathbf{C}$. This shows that row equivalence is an equivalence relation.

For the computation, $\mathbf{E}$ is invertible, with inverse $\begin{pmatrix} 1 & 0 \\ 2 & 1 \end{pmatrix}$ (check both products, or use Theorem 0.9.8 with $ad - bc = 1$). By the row view of Theorem 0.9.2 and Theorem 0.9.3, row 1 of $\mathbf{EA}$ is $1 \cdot$ (row 1 of $\mathbf{A}$) and row 2 is $-2 \cdot$ (row 1) $+ 1 \cdot$ (row 2):

$$\mathbf{EA} = \begin{pmatrix} 1 & 2 & 1 \\ 2-2 & 5-4 & 0-2 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 1 \\ 0 & 1 & -2 \end{pmatrix}.$$

So $\mathbf{E}$ subtracted twice row 1 from row 2, which is the first step of Gaussian elimination on $\mathbf{A}$.

In Chapter 3 we will see that every elementary row operation is multiplication on the left by an invertible matrix, and that row equivalence is exactly "reachable by a sequence of row operations". This is why Gaussian elimination never changes the solutions of a linear system.

0.9.12 Exercises

0.9.12.1 A. Check your understanding

Exercise. A1

(a) Let $\mathbf{A} \in M_{2\times 3}(F)$ and $\mathbf{B} \in M_{3\times 4}(F)$. Which of $\mathbf{AB}$, $\mathbf{BA}$, $\mathbf{A}^\top\mathbf{A}$, and $\mathbf{B}^\top\mathbf{A}^\top$ are defined? Give the size of each one that is.

(b) Define what it means for $\mathbf{A} \in M_n(F)$ to be invertible.

(c) Determine whether the following statement is true: "if $\mathbf{A}, \mathbf{B} \in M_n(F)$ and $\mathbf{AB} = \mathbf{O}$, then $\mathbf{A} = \mathbf{O}$ or $\mathbf{B} = \mathbf{O}$." Justify your answer.

(d) Determine whether the following statement is true: "$(\mathbf{A} + \mathbf{B})(\mathbf{A} - \mathbf{B}) = \mathbf{A}^2 - \mathbf{B}^2$ for all $\mathbf{A}, \mathbf{B} \in M_2(\mathbb{R})$." Justify your answer.

(e) Let $\mathbf{A} \in M_3(F)$ have columns $\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3$. Express $\mathbf{A}(2, 0, -1)$ and the second column of $\mathbf{A}^2$ in terms of the columns of $\mathbf{A}$ and $\mathbf{A}$ itself.

Solution.

(a) $\mathbf{AB}$ is $(2\times 3)(3\times 4)$, defined, of size 2×4. $\mathbf{BA}$ would pair rows of length 4 with columns of length 2, so it is not defined. $\mathbf{A}^\top\mathbf{A}$ is $(3\times 2)(2\times 3)$, defined, of size 3×3. $\mathbf{B}^\top\mathbf{A}^\top$ is $(4\times 3)(3\times 2)$, defined, of size 4×2; it equals $(\mathbf{AB})^\top$ by Theorem 0.9.5.

(b) $\mathbf{A}$ is invertible if there exists $\mathbf{B}\in M_n(F)$ with $\mathbf{AB}=\mathbf{I}_n$ **and** $\mathbf{BA}=\mathbf{I}_n$.

(c) False. $\begin{pmatrix}1&1\\1&1\end{pmatrix}\begin{pmatrix}1&-1\\-1&1\end{pmatrix}=\mathbf{O}$ with both factors nonzero.

(d) False. Expanding gives $(\mathbf{A}+\mathbf{B})(\mathbf{A}-\mathbf{B})=\mathbf{A}^2-\mathbf{AB}+\mathbf{BA}-\mathbf{B}^2$, which differs from $\mathbf{A}^2-\mathbf{B}^2$ unless $\mathbf{AB}=\mathbf{BA}$. For $\mathbf{A}=\begin{pmatrix}0&1\\0&0\end{pmatrix}$ and $\mathbf{B}=\begin{pmatrix}0&0\\1&0\end{pmatrix}$, we have $\mathbf{A}^2=\mathbf{B}^2=\mathbf{O}$, but $(\mathbf{A}+\mathbf{B})(\mathbf{A}-\mathbf{B})=\begin{pmatrix}0&1\\1&0\end{pmatrix}\begin{pmatrix}0&1\\-1&0\end{pmatrix}=\begin{pmatrix}-1&0\\0&1\end{pmatrix}\neq\mathbf{O}$.

(e) By Theorem 0.9.3, $\mathbf{A}(2,0,-1)=2\mathbf{a}_1-\mathbf{a}_3$. By Theorem 0.9.2, the second column of $\mathbf{A}^2=\mathbf{AA}$ is $\mathbf{Aa}_2$.

0.9.12.2 B. Practice

Exercise. B1: Products and sizes

Let

$$\mathbf{A}=\begin{pmatrix}1&-1&2\\0&3&1\end{pmatrix},\quad \mathbf{B}=\begin{pmatrix}2&1\\1&0\\0&4\end{pmatrix},\quad \mathbf{C}=\begin{pmatrix}1&2\\3&4\end{pmatrix}$$

over $\mathbb{R}$.

(a) Compute $\mathbf{AB}$ and $\mathbf{BA}$.

(b) Determine which of $\mathbf{AC}$ and $\mathbf{CA}$ is defined, and compute it.

(c) Compute $\mathbf{Ax}$ for $\mathbf{x}=(2,0,-1)$ as a combination of the columns of $\mathbf{A}$.

Solution.

(a) $\mathbf{AB}$ is 2×2 and $\mathbf{BA}$ is 3×3. Row against column,

$$\mathbf{AB}=\begin{pmatrix}2-1+0&1+0+8\\0+3+0&0+0+4\end{pmatrix}=\begin{pmatrix}1&9\\3&4\end{pmatrix},\qquad \mathbf{BA}=\begin{pmatrix}2&1&5\\1&-1&2\\0&12&4\end{pmatrix}.$$

For instance $(\mathbf{BA})_{12}=2\cdot(-1)+1\cdot 3=1$ and $(\mathbf{BA})_{32}=0\cdot(-1)+4\cdot 3=12$.

(b) $\mathbf{AC}$ would be $(2\times 3)(2\times 2)$, and $3\neq 2$, so it is not defined. $\mathbf{CA}$ is $(2\times 2)(2\times 3)$, of size 2×3:

$$\mathbf{CA}=\begin{pmatrix}1&-1+6&2+2\\3&-3+12&6+4\end{pmatrix}=\begin{pmatrix}1&5&4\\3&9&10\end{pmatrix}.$$

(c) By Theorem 0.9.3, $\mathbf{Ax}=2\mathbf{a}_1+0\mathbf{a}_2-\mathbf{a}_3=2(1,0)-(2,1)=(0,-1)$.

Exercise. B2: Traces of products

(a) With $\mathbf{A}$ and $\mathbf{B}$ as in Exercise (B1: Products and sizes), verify that $\operatorname{tr}(\mathbf{AB}) = \operatorname{tr}(\mathbf{BA})$, although $\mathbf{AB}$ and $\mathbf{BA}$ have different sizes.

(b) Let $\mathbf{A} = \operatorname{diag}(1, 2)$, $\mathbf{B} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$, and $\mathbf{C} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}$. Show that $\operatorname{tr}(\mathbf{ABC}) \neq \operatorname{tr}(\mathbf{BAC})$, and verify that $\operatorname{tr}(\mathbf{ABC}) = \operatorname{tr}(\mathbf{CAB})$.

Solution.

(a) From Exercise (B1: Products and sizes), $\operatorname{tr}(\mathbf{AB}) = 1+4 = 5$ and $\operatorname{tr}(\mathbf{BA}) = 2+(-1)+4 = 5$, as Theorem 0.9.6 predicts.

(b) By the column view, $\mathbf{BC}$ has columns $\mathbf{Be}_2 = \mathbf{e}_1$ and $\mathbf{B0} = \mathbf{0}$, so $\mathbf{BC} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$, and $\mathbf{ABC} = \mathbf{A}(\mathbf{BC}) = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$, with trace 1. Next, $\mathbf{BA} = \begin{pmatrix} 0 & 2 \\ 0 & 0 \end{pmatrix}$, so $\mathbf{BAC} = \begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix}$, with trace 2. Hence $\operatorname{tr}(\mathbf{ABC}) = 1 \neq 2 = \operatorname{tr}(\mathbf{BAC})$. Finally $\mathbf{AB} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$ and $\mathbf{CAB} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$, with trace $1 = \operatorname{tr}(\mathbf{ABC})$, as the cyclic rule predicts.

Exercise. B3: Inverses of 2×2 matrices

(a) Find the inverse of $\begin{pmatrix} 3 & 4 \\ 2 & 3 \end{pmatrix} \in M_2(\mathbb{R})$, and check both products.

(b) Show that $\mathbf{A} = \begin{pmatrix} 2 & 6 \\ 1 & 3 \end{pmatrix}$ is not invertible by finding a nonzero $\mathbf{x}$ with $\mathbf{Ax} = \mathbf{0}$.

(c) Find the inverse of $\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \in M_2(\mathbb{F}_5)$.

(d) For which $t \in \mathbb{R}$ is $\begin{pmatrix} t & 2 \\ 2 & t \end{pmatrix}$ invertible?

Solution.

(a) Here $ad-bc = 9-8 = 1 \neq 0$, so by Theorem 0.9.8 the inverse is $\mathbf{B} = \begin{pmatrix} 3 & -4 \\ -2 & 3 \end{pmatrix}$. Check: $\begin{pmatrix} 3 & 4 \\ 2 & 3 \end{pmatrix}\mathbf{B} = \begin{pmatrix} 9-8 & -12+12 \\ 6-6 & -8+9 \end{pmatrix} = \mathbf{I}_2$ and $\mathbf{B}\begin{pmatrix} 3 & 4 \\ 2 & 3 \end{pmatrix} = \begin{pmatrix} 9-8 & 12-12 \\ -6+6 & -8+9 \end{pmatrix} = \mathbf{I}_2$.

(b) Here $ad - bc = 6 - 6 = 0$. Following the proof of Theorem 0.9.8, take $\mathbf{x} = (d, -c) = (3, -1) \neq \mathbf{0}$. Then $\mathbf{Ax} = (6-6, 3-3) = \mathbf{0}$. If $\mathbf{A}$ were invertible, then $\mathbf{x} = \mathbf{A}^{-1}(\mathbf{Ax}) = \mathbf{0}$, a contradiction. Hence $\mathbf{A}$ is not invertible.

(c) In $\mathbb{F}_5$, $ad-bc = 4-6 = -2 = 3 \neq 0$, and $3^{-1} = 2$ since $3 \cdot 2 = 6 = 1$. By Theorem 0.9.8 the inverse is $2\begin{pmatrix} 4 & -2 \\ -3 & 1 \end{pmatrix} = \begin{pmatrix} 8 & -4 \\ -6 & 2 \end{pmatrix} = \begin{pmatrix} 3 & 1 \\ 4 & 2 \end{pmatrix}$. Check: $\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}\begin{pmatrix} 3 & 1 \\ 4 & 2 \end{pmatrix} =$

$\begin{pmatrix} 11 & 5 \\ 25 & 11 \end{pmatrix} = \mathbf{I}_2$ and $\begin{pmatrix} 3 & 1 \\ 4 & 2 \end{pmatrix}\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} = \begin{pmatrix} 6 & 10 \\ 10 & 16 \end{pmatrix} = \mathbf{I}_2$ in $\mathbb{F}_5$.

(d) By Theorem 0.9.8, it is invertible if and only if $t^2 - 4 \neq 0$, that is, $t \neq 2$ and $t \neq -2$.

Exercise. B4: Transposes of products

(a) Prove that $(\mathbf{A}_1\mathbf{A}_2\cdots\mathbf{A}_k)^\top = \mathbf{A}_k^\top \cdots \mathbf{A}_2^\top \mathbf{A}_1^\top$ for all $k \geq 1$ and matrices $\mathbf{A}_1, \ldots, \mathbf{A}_k$ whose sizes allow the product.

(b) Let $\mathbf{A} \in M_n(F)$. Prove that $\mathbf{A} + \mathbf{A}^\top$ is symmetric and $\mathbf{A} - \mathbf{A}^\top$ is skew-symmetric.

Solution.

(a) We use induction on k (Theorem 0.2.7). For $k = 1$ there is nothing to prove. Suppose the formula holds for products of k matrices, and let $\mathbf{A}_1, \ldots, \mathbf{A}_{k+1}$ be given. By associativity, $\mathbf{A}_1 \cdots \mathbf{A}_{k+1} = (\mathbf{A}_1 \cdots \mathbf{A}_k)\mathbf{A}_{k+1}$. By Theorem 0.9.5 (part 4) and then the induction hypothesis,

$$(\mathbf{A}_1 \cdots \mathbf{A}_{k+1})^\top = \mathbf{A}_{k+1}^\top (\mathbf{A}_1 \cdots \mathbf{A}_k)^\top = \mathbf{A}_{k+1}^\top \mathbf{A}_k^\top \cdots \mathbf{A}_1^\top.$$

This proves the formula for all $k \geq 1$.

(b) By Theorem 0.9.5 (parts 1–3), $(\mathbf{A} + \mathbf{A}^\top)^\top = \mathbf{A}^\top + (\mathbf{A}^\top)^\top = \mathbf{A}^\top + \mathbf{A} = \mathbf{A} + \mathbf{A}^\top$, so $\mathbf{A} + \mathbf{A}^\top$ is symmetric. Similarly $(\mathbf{A} - \mathbf{A}^\top)^\top = \mathbf{A}^\top - \mathbf{A} = -(\mathbf{A} - \mathbf{A}^\top)$, so $\mathbf{A} - \mathbf{A}^\top$ is skew-symmetric.

0.9.12.3 C. Going deeper

Exercise. C1: $\mathbf{AB} - \mathbf{BA} = \mathbf{I}$ is impossible, or is it?

(a) Prove that there are no $\mathbf{A}, \mathbf{B} \in M_n(\mathbb{R})$, for any $n \geq 1$, with $\mathbf{AB} - \mathbf{BA} = \mathbf{I}_n$.

(b) Show that the argument of (a) breaks down in $M_2(\mathbb{F}_2)$, and that $\mathbf{A} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}$, $\mathbf{B} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$ satisfy $\mathbf{AB} - \mathbf{BA} = \mathbf{I}_2$ over $\mathbb{F}_2$.

Hint: consider the trace of both sides.

Solution.

(a) Suppose, for a contradiction, that $\mathbf{AB} - \mathbf{BA} = \mathbf{I}_n$. Taking traces and using Theorem 0.9.6 (parts 1 and 3),

$$n = \operatorname{tr} \mathbf{I}_n = \operatorname{tr}(\mathbf{AB} - \mathbf{BA}) = \operatorname{tr}(\mathbf{AB}) - \operatorname{tr}(\mathbf{BA}) = 0.$$

This contradicts $n \geq 1$. Hence no such $\mathbf{A}, \mathbf{B}$ exist.

(b) Over $\mathbb{F}_2$ the same computation gives $\operatorname{tr} \mathbf{I}_2 = 1 + 1 = 0$, so taking traces yields $0 = 0$ and no contradiction. For the given matrices, the column view gives $\mathbf{AB} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$

and $\mathbf{BA} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$, so $\mathbf{AB} - \mathbf{BA} = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \mathbf{I}_2$, because $-1 = 1$ in $\mathbb{F}_2$.

Exercise. C2: Triangular matrices with nonzero diagonal

Let F be a field.

(a) Let $\mathbf{U} = \begin{pmatrix} a & b \\ 0 & d \end{pmatrix}$ with $a, d \neq 0$. Show that $\mathbf{U}$ is invertible and that $\mathbf{U}^{-1}$ is upper triangular.

(b) Let $\mathbf{U} = \begin{pmatrix} a & b & c \\ 0 & d & e \\ 0 & 0 & f \end{pmatrix}$ with $a, d, f \neq 0$. Show that

$$\mathbf{V} = \begin{pmatrix} a^{-1} & -b(ad)^{-1} & (be - cd)(adf)^{-1} \\ 0 & d^{-1} & -e(df)^{-1} \\ 0 & 0 & f^{-1} \end{pmatrix}$$

is the inverse of $\mathbf{U}$.

(c) Explain how the entries of $\mathbf{V}$ can be found, column by column, by solving $\mathbf{U}\mathbf{v}_k = \mathbf{e}_k$.

The general case, for all n, follows from the methods of Chapter 3.

Solution.

(a) Here $ad - b \cdot 0 = ad \neq 0$, since a product of nonzero elements of a field is nonzero (Theorem 0.7.1). By Theorem 0.9.8, $\mathbf{U}$ is invertible with $\mathbf{U}^{-1} = (ad)^{-1} \begin{pmatrix} d & -b \\ 0 & a \end{pmatrix} = \begin{pmatrix} a^{-1} & -b(ad)^{-1} \\ 0 & d^{-1} \end{pmatrix}$, which is upper triangular.

(b) The inverses a^{-1}, d^{-1}, f^{-1} exist, and $ad, df, adf \neq 0$ by Theorem 0.7.1, so $\mathbf{V}$ is defined. We check $\mathbf{UV} = \mathbf{I}_3$ entry by entry. The diagonal entries are $a \cdot a^{-1} = 1$, $d \cdot d^{-1} = 1$, $f \cdot f^{-1} = 1$. Below the diagonal both matrices are zero, so $(\mathbf{UV})_{21} = (\mathbf{UV})_{31} = (\mathbf{UV})_{32} = 0$. Above the diagonal,

$$\begin{aligned} (\mathbf{UV})_{12} &= a \cdot \big(-b(ad)^{-1}\big) + bd^{-1} = -bd^{-1} + bd^{-1} = 0, \\ (\mathbf{UV})_{23} &= d \cdot \big(-e(df)^{-1}\big) + ef^{-1} = -ef^{-1} + ef^{-1} = 0, \\ (\mathbf{UV})_{13} &= a(be - cd)(adf)^{-1} + b\big(-e(df)^{-1}\big) + cf^{-1} \\ &= (be - cd)(df)^{-1} - be(df)^{-1} + cd(df)^{-1} = 0. \end{aligned}$$

Hence $\mathbf{UV} = \mathbf{I}_3$. For $\mathbf{VU}$, again the diagonal entries are 1 and the entries below the diagonal are 0, and

$$\begin{aligned} (\mathbf{VU})_{12} &= a^{-1}b - b(ad)^{-1}d = a^{-1}b - a^{-1}b = 0, \\ (\mathbf{VU})_{23} &= d^{-1}e - e(df)^{-1}f = 0, \\ (\mathbf{VU})_{13} &= a^{-1}c - b(ad)^{-1}e + (be - cd)(adf)^{-1}f \\ &= a^{-1}c - be(ad)^{-1} + be(ad)^{-1} - cd(ad)^{-1} = 0, \end{aligned}$$

using $cd(ad)^{-1} = ca^{-1}$. Hence $\mathbf{VU} = \mathbf{I}_3$, and $\mathbf{V} = \mathbf{U}^{-1}$. Both products had to be checked, since we may not yet use that one implies the other.

(c) By Theorem 0.9.2, $\mathbf{UV} = \mathbf{I}_3$ says exactly that the columns $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ of $\mathbf{V}$ satisfy $\mathbf{U}\mathbf{v}_k = \mathbf{e}_k$. Because $\mathbf{U}$ is upper triangular, each system can be solved from the bottom row up. For example, $\mathbf{U}\mathbf{v}_3 = \mathbf{e}_3$ reads $fx_3 = 1$, $dx_2 + ex_3 = 0$, $ax_1 + bx_2 + cx_3 = 0$, giving $x_3 = f^{-1}$, then $x_2 = -e(df)^{-1}$, then $x_1 = (be - cd)(adf)^{-1}$, which is the third column of $\mathbf{V}$. (This only finds a candidate satisfying $\mathbf{UV} = \mathbf{I}_3$; the check $\mathbf{VU} = \mathbf{I}_3$ in (b) is still needed.)

Exercise. C3: Matrices that commute

Let $\mathbf{N} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \in M_2(\mathbb{R})$.

(a) Find all $\mathbf{X} \in M_2(\mathbb{R})$ with $\mathbf{XN} = \mathbf{NX}$.

(b) Deduce that the only matrices $\mathbf{X} \in M_2(\mathbb{R})$ with $\mathbf{XY} = \mathbf{YX}$ **for every** $\mathbf{Y} \in M_2(\mathbb{R})$ are the scalar multiples $c\mathbf{I}_2$, $c \in \mathbb{R}$.

Solution.

(a) Let $\mathbf{X} = \begin{pmatrix} p & q \\ r & s \end{pmatrix}$. By the column view, $\mathbf{XN} = \begin{pmatrix} \mathbf{X0} & \mathbf{Xe}_1 \end{pmatrix} = \begin{pmatrix} 0 & p \\ 0 & r \end{pmatrix}$, and by the row view, $\mathbf{NX}$ has rows (row 2 of $\mathbf{X}$) and $\mathbf{0}$: $\mathbf{NX} = \begin{pmatrix} r & s \\ 0 & 0 \end{pmatrix}$. Comparing entries, $\mathbf{XN} = \mathbf{NX}$ if and only if $r = 0$ and $p = s$, that is, $\mathbf{X} = \begin{pmatrix} p & q \\ 0 & p \end{pmatrix} = p\mathbf{I}_2 + q\mathbf{N}$ with $p, q \in \mathbb{R}$ arbitrary.

(b) Every $c\mathbf{I}_2$ commutes with every $\mathbf{Y}$, by Theorem 0.9.4: $(c\mathbf{I}_2)\mathbf{Y} = c\mathbf{Y} = \mathbf{Y}(c\mathbf{I}_2)$. Conversely, suppose $\mathbf{XY} = \mathbf{YX}$ for every $\mathbf{Y}$. Taking $\mathbf{Y} = \mathbf{N}$, part (a) gives $\mathbf{X} = \begin{pmatrix} p & q \\ 0 & p \end{pmatrix}$. Taking $\mathbf{Y} = \mathbf{N}^\top = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}$, we compute $\mathbf{XN}^\top = \begin{pmatrix} q & 0 \\ p & 0 \end{pmatrix}$ and $\mathbf{N}^\top\mathbf{X} = \begin{pmatrix} 0 & 0 \\ p & q \end{pmatrix}$. Equality forces $q = 0$. Hence $\mathbf{X} = p\mathbf{I}_2$, as claimed.

0.10 Groups and Permutations

Several structures in this chapter have quietly behaved alike. The invertible $n \times n$ matrices, the nonzero elements of a field, and the bijections of a set can each be combined, have an element that does nothing, and can be undone. This short section gives that common structure a name, proves its basic rules once, and introduces the permutations that the determinant in Chapter 7 is built from. We keep it brief: groups are a language for us here, not a subject.

0.10.1 Groups

Look at four situations side by side. In $\mathrm{GL}_n(F)$, the set of invertible matrices in $M_n(F)$, we multiply matrices, $\mathbf{I}_n$ does nothing, and each $\mathbf{A}$ is undone by $\mathbf{A}^{-1}$ (Theorem 0.9.7). In a field F,

we multiply nonzero elements, 1 does nothing, and each a is undone by a^{-1}. The four rotations of a square about its center (by $0°, 90°, 180°, 270°$) can be performed one after another, the rotation by $0°$ does nothing, and each rotation is undone by rotating back. The bijections $X \to X$ of a set are composed, id_X does nothing, and each is undone by its inverse function (Theorem 0.4.7). The objects differ completely, but the rules we use are the same three. So we give the pattern a name.

A group is a set with one way of combining elements, in which combining is associative, some element does nothing, and every element can be undone.

Definition 0.10.1. Group

A **group** is a set G together with a **binary operation** $G \times G \to G$, $(a, b) \mapsto ab$, such that:

(G1) **(Associativity)** $(ab)c = a(bc)$ **for all** $a, b, c \in G$;

(G2) **(Identity)** there is an element $e \in G$ with $ea = ae = a$ **for every** $a \in G$;

(G3) **(Inverses) for every** $a \in G$ there is $b \in G$ with $ab = ba = e$, where e is the element of (G2).

The group is **abelian** if moreover $ab = ba$ for all $a, b \in G$.

In words: the operation is a function on pairs, so the combination of two elements of G **lies in** G again; this "closure" is built into the phrase "binary operation $G \times G \to G$". (G1) lets us drop brackets. (G2) asks for an element that changes nothing, from **both** sides. (G3) asks that each element can be undone, again from both sides, back to that same e. When the operation is addition we write $a + b$, 0, and $-a$ instead of ab, e, and a^{-1}; this notation is only used for abelian groups.

Example. Groups

Check that each of the following is a group, and decide which are abelian: $(\mathbb{Z}, +)$; $(F, +)$ and $(F \setminus \{0\}, \cdot)$ for a field F; $\mathrm{GL}_n(F)$ under matrix multiplication; the set S_n of bijections $\{1, \dots, n\} \to \{1, \dots, n\}$ under composition; and a one-element set $\{e\}$ with $ee = e$.

Solution. $(\mathbb{Z}, +)$: the sum of integers is an integer, addition is associative, 0 is an identity, and $-a$ undoes a. It is abelian. $(F, +)$ is a group for the same reasons, by the field axioms (Definition 0.7.1); it is abelian.

$(F\setminus\{0\}, \cdot)$: the product of two nonzero elements is nonzero by Theorem 0.7.1, so the operation lands in $F \setminus \{0\}$. Associativity and the identity $1 \neq 0$ come from the field axioms, and each $a \neq 0$ has an inverse a^{-1}, which is nonzero since $aa^{-1} = 1 \neq 0$. It is abelian.

$\mathrm{GL}_n(F)$: by Theorem 0.9.7 (part 3), a product of invertible matrices is invertible, so the operation lands in $\mathrm{GL}_n(F)$. Associativity holds by Theorem 0.9.4, $\mathbf{I}_n$ is invertible and is an identity, and $\mathbf{A}^{-1} \in \mathrm{GL}_n(F)$ undoes $\mathbf{A}$ by part 2. For $n \geq 2$ it is **not** abelian: for $n = 2$, $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$, while the product in the other order is $\begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}$. (Over $\mathbb{F}_2$ read 2 as 0; the two products still differ.) For $n \geq 3$, put these blocks in the top-left corner and 1's further down the diagonal. $\mathrm{GL}_1(F)$ is just $F \setminus \{0\}$, and is abelian.

S_n: a composition of bijections is a bijection (Theorem 0.4.3), composition is associative

(Theorem 0.4.2), id is an identity, and each bijection has an inverse function (Theorem 0.4.7), which is again a bijection. We will see below that S_n is not abelian for $n \geq 3$.

$\{e\}$: all three axioms reduce to $ee = e$. This **trivial group** is the degenerate case; it reappears as the smallest subgroup of every group.

Now nonexamples, each a minimal change of an example. Replace addition on $\mathbb{Z}$ by multiplication. The operation still lands in $\mathbb{Z}$, it is associative, and 1 is an identity. But (G3) fails: there is no integer b with $2b = 1$. So $(\mathbb{Z}, \cdot)$ is **not** a group. Similarly, replace $\mathrm{GL}_n(F)$ by all of $M_n(F)$ under multiplication: (G1) and (G2) still hold with $\mathbf{I}_n$, but the zero matrix has no inverse, so (G3) fails. Keep $\mathbb{N}$ under $+$: the identity 0 is there, but 1 has no inverse in $\mathbb{N}$.

Warning. **The operation must land in the set.** $\mathrm{GL}_n(F)$ under matrix **addition** is not a group: $\mathbf{I}_n$ and $-\mathbf{I}_n$ are invertible, but $\mathbf{I}_n + (-\mathbf{I}_n) = \mathbf{O}$ is not, so addition is not even a binary operation on $\mathrm{GL}_n(F)$. Before checking (G1)–(G3), check that combining two elements of the set gives an element of the set.

The rules we proved separately for fields and for matrices follow from the axioms alone. It is worth proving them once:

Theorem 0.10.1. Basic properties of groups

Let G be a group and $a, b, c \in G$.

1. The identity element of G is unique.
2. The inverse of a is unique. We denote it by a^{-1}.
3. $(a^{-1})^{-1} = a$ and $(ab)^{-1} = b^{-1}a^{-1}$.
4. **(Cancellation)** If $ab = ac$, then $b = c$. If $ba = ca$, then $b = c$.

Proof. For 1, suppose e and e' both satisfy (G2). Then $e = ee'$, using that e' is an identity, and $ee' = e'$, using that e is one. Hence $e = e'$.

For 2, suppose b and b' both satisfy $ab = ba = e$ and $ab' = b'a = e$. By (G2) and (G1),

$$b = be = b(ab') = (ba)b' = eb' = b'.$$

For 3, the equations $aa^{-1} = a^{-1}a = e$ say that a is an inverse of a^{-1}, so $(a^{-1})^{-1} = a$ by part 2. Next, by (G1),

$$(ab)(b^{-1}a^{-1}) = a(bb^{-1})a^{-1} = aea^{-1} = aa^{-1} = e,$$

and similarly $(b^{-1}a^{-1})(ab) = b^{-1}(a^{-1}a)b = e$. By part 2, $(ab)^{-1} = b^{-1}a^{-1}$.

For 4, suppose $ab = ac$. Multiplying on the left by a^{-1} and using (G1), $b = (a^{-1}a)b = a^{-1}(ab) = a^{-1}(ac) = (a^{-1}a)c = c$. The second statement follows by multiplying on the right by a^{-1}. This proves the theorem.

■

These are the arguments of Theorem 0.9.7 with matrices erased. That is the point of the definition: a proof that uses only (G1)–(G3) holds in every group at once.

0.10.2 Subgroups

After a structure comes the substructure. The even integers sit inside $(\mathbb{Z}, +)$ and form a group on their own, with the same operation.

Definition 0.10.2. Subgroup

Let G be a group. A subset $H \subseteq G$ is a **subgroup** of G if $e \in H$, and $ab \in H$ and $a^{-1} \in H$ **for all** $a, b \in H$.

Then the operation of G restricts to a binary operation on H, and H is a group with it: (G1) holds in H because it holds in G, and (G2), (G3) hold because e and the inverses lie in H. In practice one checks a single condition:

Theorem 0.10.2. Subgroup test

Let G be a group and $H \subseteq G$. Then H is a subgroup of G if and only if $H \neq \emptyset$ and $ab^{-1} \in H$ for all $a, b \in H$.

Proof. ($\Rightarrow$) If H is a subgroup, then $e \in H$, so $H \neq \emptyset$. For $a, b \in H$ we have $b^{-1} \in H$, and then $ab^{-1} \in H$.

($\Leftarrow$) Suppose $H \neq \emptyset$ and $ab^{-1} \in H$ for all $a, b \in H$. Pick $h \in H$. Taking $a = b = h$ gives $e = hh^{-1} \in H$. For $b \in H$, taking $a = e$ gives $b^{-1} = eb^{-1} \in H$. For $a, b \in H$, we now know $b^{-1} \in H$, so $ab = a(b^{-1})^{-1} \in H$ by Theorem 0.10.1. Hence H is a subgroup. This proves the theorem.

■

For example, $2\mathbb{Z} = \{2k : k \in \mathbb{Z}\}$ is a subgroup of $(\mathbb{Z}, +)$. In additive notation the test reads "$a - b \in H$", and $2k - 2l = 2(k - l) \in 2\mathbb{Z}$, while $0 \in 2\mathbb{Z}$. By contrast $\mathbb{N} \subseteq \mathbb{Z}$ is nonempty and closed under $+$, but $0 - 1 = -1 \notin \mathbb{N}$: the test fails, because $\mathbb{N}$ lacks inverses. The set of odd integers fails more badly, since it does not contain 0. In $\mathrm{GL}_n(F)$, the invertible diagonal matrices form a subgroup: $\mathbf{I}_n$ is one, and by Example (Some invertible matrices) the inverse of $\mathrm{diag}(d_1, \ldots, d_n)$ is $\mathrm{diag}(d_1^{-1}, \ldots, d_n^{-1})$, so $\mathbf{D}\mathbf{E}^{-1}$ is again diagonal with nonzero diagonal entries. Every group G has the subgroups $\{e\}$ and G.

0.10.3 Homomorphisms

After objects come the maps between them that respect the structure.

Definition 0.10.3. Group homomorphism

Let G and H be groups. A function $\varphi\colon G \to H$ is a **homomorphism** if $\varphi(ab) = \varphi(a)\varphi(b)$ **for all** $a, b \in G$. A **bijective** homomorphism is an **isomorphism**. The **kernel** of φ is $\ker \varphi = \{g \in G : \varphi(g) = e_H\}$, where e_H is the identity of H.

The product ab on the left is taken in G, and $\varphi(a)\varphi(b)$ on the right in H. For example, $F \setminus \{0\} \to \mathrm{GL}_2(F)$, $a \mapsto \mathrm{diag}(a, 1)$, is a homomorphism, since $\mathrm{diag}(ab, 1) = \mathrm{diag}(a, 1)\,\mathrm{diag}(b, 1)$; its kernel is $\{1\}$. The map $\mathbb{Z} \to \mathbb{Z}/n\mathbb{Z}$, $a \mapsto [a]$, is a homomorphism of additive groups, because $[a + b] = [a] + [b]$ is exactly how addition on $\mathbb{Z}/n\mathbb{Z}$ was defined (Example (Addition modulo n is well defined)); its kernel is $n\mathbb{Z}$. On the other hand, $x \mapsto x + 1$ from $(\mathbb{R}, +)$ to itself is not a homomorphism: writing $\varphi(x) = x + 1$, we get $\varphi(0 + 0) = 1$ but $\varphi(0) + \varphi(0) = 2$.

Theorem 0.10.3. Basic properties of homomorphisms

Let $\varphi\colon G \to H$ be a homomorphism of groups. Then $\varphi(e_G) = e_H$, $\varphi(a^{-1}) = \varphi(a)^{-1}$ for every $a \in G$, and $\ker \varphi$ is a subgroup of G.

Proof. Since $\varphi(e_G)\varphi(e_G) = \varphi(e_G e_G) = \varphi(e_G) = \varphi(e_G)e_H$, cancellation (Theorem 0.10.1) gives $\varphi(e_G) = e_H$. Next, $\varphi(a)\varphi(a^{-1}) = \varphi(aa^{-1}) = \varphi(e_G) = e_H$, and similarly $\varphi(a^{-1})\varphi(a) = e_H$, so $\varphi(a^{-1}) = \varphi(a)^{-1}$ by uniqueness of inverses. Finally, $e_G \in \ker \varphi$, and for $a, b \in \ker \varphi$,

$$\varphi(ab^{-1}) = \varphi(a)\varphi(b)^{-1} = e_H e_H^{-1} = e_H,$$

so $ab^{-1} \in \ker \varphi$. By Theorem 0.10.2, $\ker \varphi$ is a subgroup. This proves the theorem. ∎

Two homomorphisms will matter later: the determinant $\mathrm{GL}_n(F) \to F \setminus \{0\}$ and the sign $S_n \to \{1, -1\}$, both in Chapter 7. Linear maps in Chapter 2 are, among other things, homomorphisms of additive groups, and their kernels will be one of our main tools.

0.10.4 Permutations

A **permutation** of $\{1, \dots, n\}$ is a bijection $\{1, \dots, n\} \to \{1, \dots, n\}$, and S_n is the group of all of them, the **symmetric group**. It has $n!$ elements: there are n choices for $\sigma(1)$, then $n - 1$ for $\sigma(2)$, and so on.

A permutation σ can be written in **two-line notation**, with $\sigma(i)$ below i. A more compact notation lists what happens along a loop. For **distinct** $i_1, \dots, i_k \in \{1, \dots, n\}$, the **cycle** $(i_1\ i_2\ \cdots\ i_k)$ is the permutation sending

$$i_1 \mapsto i_2 \mapsto \cdots \mapsto i_k \mapsto i_1$$

and fixing every other element. A cycle of length 2 is a **transposition**: $(i\ j)$ swaps i and j. The same cycle can be written starting at any of its entries: $(1\ 2\ 3) = (2\ 3\ 1) = (3\ 1\ 2)$, but $(1\ 3\ 2)$ is different.

The product $\sigma\tau$ in S_n means the composition $\sigma \circ \tau$: **first apply τ, then σ**. So products of cycles are read **from right to left**.

Example. Computing with cycles

In S_3, compute $(1\ 3)(1\ 2)$ and $(1\ 2)(1\ 3)$. In S_5, write $\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 3 & 5 & 1 & 2 & 4 \end{pmatrix}$ (two-line notation) as a product of cycles.

Solution. For $(1\ 3)(1\ 2)$, follow each element through $(1\ 2)$ first, then $(1\ 3)$: $1 \mapsto 2 \mapsto 2$, $2 \mapsto 1 \mapsto 3$, $3 \mapsto 3 \mapsto 1$. So $(1\ 3)(1\ 2) = (1\ 2\ 3)$. For $(1\ 2)(1\ 3)$: $1 \mapsto 3 \mapsto 3$, $3 \mapsto 1 \mapsto 2$, $2 \mapsto 2 \mapsto 1$. So $(1\ 2)(1\ 3) = (1\ 3\ 2)$. The two products differ, so S_3 is not abelian; the same two transpositions show that S_n is not abelian for every $n \geq 3$.

For σ, start at 1 and follow: $1 \mapsto 3 \mapsto 1$, a loop $(1\ 3)$. The smallest element not yet used is 2: $2 \mapsto 5 \mapsto 4 \mapsto 2$, a loop $(2\ 5\ 4)$. Every element is now used, so $\sigma = (1\ 3)(2\ 5\ 4)$. These two cycles move disjoint sets of elements, so they can be applied in either order.

Warning. **Mind the order.** $\sigma\tau$ applies τ first, exactly as for functions, and as for matrices: $(\mathbf{AB})\mathbf{x} = \mathbf{A}(\mathbf{Bx})$ applies $\mathbf{B}$ first. Reading a product of cycles from left to right gives the wrong answer, as the two different products $(1\ 3)(1\ 2) \neq (1\ 2)(1\ 3)$ above show.

Quick check. Compute $(1\ 2\ 3)(1\ 2)$ in S_3.

Solution. Apply $(1\ 2)$ first: $1 \mapsto 2 \mapsto 3$, $2 \mapsto 1 \mapsto 2$, $3 \mapsto 3 \mapsto 1$. So 1 and 3 are swapped and 2 is fixed: $(1\ 2\ 3)(1\ 2) = (1\ 3)$.

The example showed $(1\ 2\ 3) = (1\ 3)(1\ 2)$: a 3-cycle is a product of transpositions. Every permutation is, and the proof is a model induction on n.

Theorem 0.10.4. Transpositions generate S_n

Let $n \geq 2$. Every permutation in S_n is a product of transpositions.

Idea. Induction on n: what is the smaller object inside a permutation of $n+1$ elements? If σ fixes $n+1$, it is really a permutation of $\{1, \ldots, n\}$. If not, one transposition repairs that: swap $\sigma(n+1)$ back to $n+1$. The result fixes $n+1$, so the induction hypothesis applies, and we undo the swap.

Proof. We use induction on $n \geq 2$ (Theorem 0.2.7). For $n = 2$, $S_2 = \{\mathrm{id}, (1\ 2)\}$, and $\mathrm{id} = (1\ 2)(1\ 2)$.

Suppose every permutation in S_n is a product of transpositions, and let $\sigma \in S_{n+1}$.

Case 1. $\sigma(n+1) = n+1$. Since σ is injective, $\sigma(i) \neq n+1$ for $i \leq n$, so σ restricts to a function $\sigma' \colon \{1, \ldots, n\} \to \{1, \ldots, n\}$. It is injective because σ is, and hence bijective by Theorem 0.4.9. By the induction hypothesis, $\sigma' = \tau_1 \cdots \tau_r$ with transpositions $\tau_j \in S_n$. Regard each τ_j as a transposition in S_{n+1} fixing $n+1$. Then $\tau_1 \cdots \tau_r$ agrees with $\sigma' = \sigma$ on $\{1, \ldots, n\}$ and fixes $n+1$, as σ does. Hence $\sigma = \tau_1 \cdots \tau_r$.

Case 2. $\sigma(n+1) = k \neq n+1$. Let $\tau = (k\ \ n{+}1)$. Then $(\tau\sigma)(n+1) = \tau(k) = n+1$, so by Case 1, $\tau\sigma = \tau_1 \cdots \tau_r$ for some transpositions τ_j. Since $\tau\tau = \mathrm{id}$, multiplying on the left by τ

gives $\sigma = \tau\tau_1 \cdots \tau_r$, a product of transpositions.

This completes the induction, and proves the theorem.

■

The product is not unique: $(1\ 2\ 3) = (1\ 3)(1\ 2) = (1\ 2)(2\ 3)$, and one can always insert $(1\ 2)(1\ 2) = \mathrm{id}$. What turns out to be unique is whether the number of transpositions is even or odd. That fact defines the **sign** of a permutation, which we do not define here; it is the subject of Chapter 7, where it builds the determinant.

0.10.5 Exercises

0.10.5.1 A. Check your understanding

Exercise. A1

(a) State the three axioms of a group.

(b) Is $\mathbb{Z}$ with the operation $(a, b) \mapsto a - b$ a group? Justify your answer.

(c) Determine whether the following statement is true: "in every group, $(ab)^{-1} = a^{-1}b^{-1}$." Justify your answer.

(d) Compute $(1\ 3)(1\ 2\ 3)$ in S_3.

(e) Explain why $\{\mathbf{A} \in M_2(\mathbb{R}) : \mathbf{A} \text{ is not invertible}\}$ is **not** a subgroup of $\mathrm{GL}_2(\mathbb{R})$.

Solution.

(a) (G1) $(ab)c = a(bc)$ for all a, b, c; (G2) there is e with $ea = ae = a$ for all a; (G3) for every a there is b with $ab = ba = e$. The operation must be a binary operation $G \times G \to G$.

(b) No. Associativity fails: $(1 - 1) - 1 = -1$, but $1 - (1 - 1) = 1$.

(c) False. By Theorem 0.10.1, $(ab)^{-1} = b^{-1}a^{-1}$, and this can differ from $a^{-1}b^{-1}$. In S_3 take $a = (1\ 2)$ and $b = (1\ 3)$, which are their own inverses. By Example (Computing with cycles), $ab = (1\ 3\ 2)$, whose inverse is $(1\ 2\ 3)$, while $a^{-1}b^{-1} = ab = (1\ 3\ 2)$. The statement is true in abelian groups.

(d) Apply $(1\ 2\ 3)$ first: $1 \mapsto 2 \mapsto 2$, $2 \mapsto 3 \mapsto 1$, $3 \mapsto 1 \mapsto 3$. Hence $(1\ 3)(1\ 2\ 3) = (1\ 2)$.

(e) It is not even a subset of $\mathrm{GL}_2(\mathbb{R})$, since its elements are not invertible; for instance it does not contain the identity $\mathbf{I}_2$.

0.10.5.2 B. Practice

Exercise. B1: The table of S_3

List the six elements of S_3 in cycle notation and write out the table whose entry in row x and column y is xy. Hence find all elements $x \in S_3$ with $x^2 = \mathrm{id}$, and all x that commute with every element of S_3.

Solution. The elements are $\mathrm{id}, (1\ 2), (1\ 3), (2\ 3), (1\ 2\ 3), (1\ 3\ 2)$. Computing each product right to left, as in Example (Computing with cycles), the entry in row x, column y is xy:

xy	id	$(1\,2)$	$(1\,3)$	$(2\,3)$	$(1\,2\,3)$	$(1\,3\,2)$
id	id	$(1\,2)$	$(1\,3)$	$(2\,3)$	$(1\,2\,3)$	$(1\,3\,2)$
$(1\,2)$	$(1\,2)$	id	$(1\,3\,2)$	$(1\,2\,3)$	$(2\,3)$	$(1\,3)$
$(1\,3)$	$(1\,3)$	$(1\,2\,3)$	id	$(1\,3\,2)$	$(1\,2)$	$(2\,3)$
$(2\,3)$	$(2\,3)$	$(1\,3\,2)$	$(1\,2\,3)$	id	$(1\,3)$	$(1\,2)$
$(1\,2\,3)$	$(1\,2\,3)$	$(1\,3)$	$(2\,3)$	$(1\,2)$	$(1\,3\,2)$	id
$(1\,3\,2)$	$(1\,3\,2)$	$(2\,3)$	$(1\,2)$	$(1\,3)$	id	$(1\,2\,3)$

For example, the entry in row $(1\ 2)$, column $(2\ 3)$ is $(1\ 2)(2\ 3)$: $1 \mapsto 1 \mapsto 2$, $2 \mapsto 3 \mapsto 3$, $3 \mapsto 2 \mapsto 1$, which is $(1\ 2\ 3)$.

Reading the diagonal, $x^2 = \mathrm{id}$ exactly for $x = \mathrm{id}, (1\ 2), (1\ 3), (2\ 3)$. An element x commutes with every element exactly when row x equals column x. This holds for id. It fails for every other element: $(1\ 2)(1\ 3) = (1\ 3\ 2) \neq (1\ 2\ 3) = (1\ 3)(1\ 2)$ handles $(1\ 2)$ and $(1\ 3)$; $(2\ 3)(1\ 2) = (1\ 3\ 2) \neq (1\ 2\ 3) = (1\ 2)(2\ 3)$ handles $(2\ 3)$; and $(1\ 2\ 3)(1\ 2) = (1\ 3) \neq (2\ 3) = (1\ 2)(1\ 2\ 3)$ handles $(1\ 2\ 3)$ and, since $(1\ 3\ 2)(1\ 2) = (2\ 3) \neq (1\ 3) = (1\ 2)(1\ 3\ 2)$, also $(1\ 3\ 2)$. Hence only id commutes with everything.

Exercise. B2: Invertible upper triangular matrices

Let F be a field and $T = \left\{ \begin{pmatrix} a & b \\ 0 & d \end{pmatrix} : a, b, d \in F,\ a \neq 0,\ d \neq 0 \right\}$.

(a) Prove that T is a subgroup of $\mathrm{GL}_2(F)$.

(b) For which fields F is T abelian? Justify your answer.

Solution.

(a) Let $\mathbf{X} = \begin{pmatrix} a & b \\ 0 & d \end{pmatrix} \in T$. Since $a, d \neq 0$, $ad - b \cdot 0 = ad \neq 0$ by Theorem 0.7.1, so $\mathbf{X}$ is invertible by Theorem 0.9.8. Hence $T \subseteq \mathrm{GL}_2(F)$, and $T \neq \emptyset$ since $\mathbf{I}_2 \in T$. Now let $\mathbf{Y} = \begin{pmatrix} a' & b' \\ 0 & d' \end{pmatrix} \in T$. By Theorem 0.9.8, $\mathbf{Y}^{-1} = (a'd')^{-1} \begin{pmatrix} d' & -b' \\ 0 & a' \end{pmatrix} = \begin{pmatrix} a'^{-1} & -b'(a'd')^{-1} \\ 0 & d'^{-1} \end{pmatrix}$. Therefore

$$\mathbf{X}\mathbf{Y}^{-1} = \begin{pmatrix} aa'^{-1} & -ab'(a'd')^{-1} + bd'^{-1} \\ 0 & dd'^{-1} \end{pmatrix},$$

whose $(2,1)$-entry is 0 and whose diagonal entries aa'^{-1}, dd'^{-1} are nonzero by Theorem 0.7.1. So $\mathbf{X}\mathbf{Y}^{-1} \in T$, and T is a subgroup by Theorem 0.10.2.

(b) It depends on the field. *Case 1: F has an element c with $c \neq 0$ and $c \neq 1$* (for example $c = 2$ in $\mathbb{R}$ or $\mathbb{F}_3$). Let $\mathbf{X} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ and $\mathbf{Y} = \mathrm{diag}(1, c)$, both in T. Then

$\mathbf{XY} = \begin{pmatrix} 1 & c \\ 0 & c \end{pmatrix}$ and $\mathbf{YX} = \begin{pmatrix} 1 & 1 \\ 0 & c \end{pmatrix}$, which differ in the $(1,2)$-entry since $c \neq 1$. So T is **not** abelian. *Case 2:* $F = \{0,1\}$, that is, $F = \mathbb{F}_2$. Then $a = d = 1$ is forced, so $T = \left\{ \begin{pmatrix} 1 & b \\ 0 & 1 \end{pmatrix} : b \in \mathbb{F}_2 \right\}$, and $\begin{pmatrix} 1 & b \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & b' \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & b+b' \\ 0 & 1 \end{pmatrix}$ is symmetric in b, b'. So T **is** abelian over $\mathbb{F}_2$, and over every other field it is not.

Exercise. B3: The exponential is an isomorphism

Let $\mathbb{R}_{>0} = \{x \in \mathbb{R} : x > 0\}$. Show that $(\mathbb{R}_{>0}, \cdot)$ is a group, and that $\exp\colon (\mathbb{R}, +) \to (\mathbb{R}_{>0}, \cdot)$, $x \mapsto e^x$, is an isomorphism. What is its kernel? You may use the familiar facts $e^{x+y} = e^x e^y$, $e^x > 0$, and that $\ln\colon \mathbb{R}_{>0} \to \mathbb{R}$ satisfies $\ln(e^x) = x$ and $e^{\ln y} = y$.

Solution. A product of positive reals is positive, so multiplication is a binary operation on $\mathbb{R}_{>0}$. It is associative, $1 \in \mathbb{R}_{>0}$ is an identity, and for $y > 0$ the inverse $1/y$ is positive. Hence $(\mathbb{R}_{>0}, \cdot)$ is a group.

Since $e^x > 0$, exp maps $\mathbb{R}$ into $\mathbb{R}_{>0}$, and $e^{x+y} = e^x e^y$ says exactly that it is a homomorphism from $(\mathbb{R}, +)$ to $(\mathbb{R}_{>0}, \cdot)$. The function ln satisfies $\ln \circ \exp = \mathrm{id}_{\mathbb{R}}$ and $\exp \circ \ln = \mathrm{id}_{\mathbb{R}_{>0}}$, so exp is bijective by Theorem 0.4.7. Hence exp is an isomorphism. Its kernel is $\{x \in \mathbb{R} : e^x = 1\} = \{0\}$, since exp is injective and $e^0 = 1$.

0.10.5.3 C. Going deeper

Exercise. C1: Every element its own inverse

(a) Let G be a group in which $g^2 = gg = e$ for every $g \in G$. Prove that G is abelian.

(b) Show that $(M_2(\mathbb{F}_2), +)$ is an example of such a group, written additively.

Solution.

(a) Let $a, b \in G$. The hypothesis $gg = e$ says that every g is its own inverse, by uniqueness of inverses (Theorem 0.10.1). Applying this to ab, a, and b, and using Theorem 0.10.1,

$$ab = (ab)^{-1} = b^{-1}a^{-1} = ba.$$

Hence G is abelian.

(b) $(M_2(\mathbb{F}_2), +)$ is a group by Theorem 0.9.1: addition is associative, $\mathbf{O}$ is an identity, and $-\mathbf{A}$ is an inverse. In additive notation, "$g^2 = e$" reads $\mathbf{A} + \mathbf{A} = \mathbf{O}$, and indeed $(\mathbf{A}+\mathbf{A})_{ij} = a_{ij} + a_{ij} = (1+1)a_{ij} = 0$ in $\mathbb{F}_2$. (This group is abelian anyway, consistent with (a).)

Exercise. C2: The order of a product of disjoint cycles

The **order** of $\sigma \in S_n$ is the smallest integer $m \geq 1$ with $\sigma^m = \mathrm{id}$, where σ^m is the product of m copies of σ. Two cycles are **disjoint** if no element appears in both.

(a) Let $\sigma = (i_1 \ \cdots \ i_k)$ be a k-cycle. Prove that $\sigma^m = \mathrm{id}$ if and only if k divides m. Deduce that σ has order k.

(b) Prove that disjoint cycles σ and τ commute, and deduce $(\sigma\tau)^m = \sigma^m\tau^m$ for all $m \geq 1$.

(c) Let σ be a k-cycle and τ an l-cycle, disjoint. Prove that $\sigma\tau$ has order $\mathrm{lcm}(k, l)$.

(d) Find the order of $(1\ 2)(3\ 4\ 5) \in S_5$.

Hint for (c): σ^m moves only elements that appear in σ.

Solution.

(a) Index the entries of the cycle modulo k, so $\sigma(i_j) = i_{j+1}$ with $i_{k+1} = i_1$. By induction on m, $\sigma^m(i_j) = i_{j+m}$, indices taken modulo k, and σ^m fixes every element not in the cycle. Hence $\sigma^m = \mathrm{id}$ if and only if $i_{j+m} = i_j$ for all j. Since $i_1, \ldots, i_k$ are distinct, this holds if and only if $j + m \equiv j \pmod{k}$, that is, $k \mid m$. The smallest $m \geq 1$ with $k \mid m$ is k, so σ has order k.

(b) Let A and B be the disjoint sets of elements appearing in σ and τ. Note that σ maps A to A and fixes everything outside A, and τ does the same with B. Let $i \in \{1, \ldots, n\}$. If $i \in A$, then $i \notin B$ and $\sigma(i) \in A$, so $\sigma(i) \notin B$; hence $\sigma\tau(i) = \sigma(i) = \tau\sigma(i)$. The case $i \in B$ is the same with the roles swapped. If $i \notin A \cup B$, both sides give i. Hence $\sigma\tau = \tau\sigma$. Then $(\sigma\tau)^m = \sigma^m\tau^m$ follows by induction on m: $(\sigma\tau)^{m+1} = \sigma^m\tau^m\sigma\tau = \sigma^m\sigma\tau^m\tau = \sigma^{m+1}\tau^{m+1}$, moving σ past τ one factor at a time, which is allowed because they commute.

(c) By (b), $(\sigma\tau)^m = \sigma^m\tau^m$. We claim this is id if and only if $\sigma^m = \mathrm{id}$ and $\tau^m = \mathrm{id}$. One direction is clear. Conversely, suppose $\sigma^m\tau^m = \mathrm{id}$, and let $i \in A$. Then $i \notin B$, so $\tau^m(i) = i$ by the hint, and $\sigma^m(i) = \sigma^m\tau^m(i) = i$. Since σ^m also fixes every $i \notin A$, $\sigma^m = \mathrm{id}$, and then $\tau^m = \sigma^m\tau^m = \mathrm{id}$ as well. By (a), $(\sigma\tau)^m = \mathrm{id}$ if and only if $k \mid m$ and $l \mid m$. The smallest such $m \geq 1$ is $\mathrm{lcm}(k, l)$, which is therefore the order of $\sigma\tau$.

(d) $(1\ 2)$ and $(3\ 4\ 5)$ are disjoint cycles of lengths 2 and 3, so by (c) the order is $\mathrm{lcm}(2, 3) = 6$.

Chapter 1

Vector Spaces

Arrows in the plane can be added and stretched. So can solutions of a homogeneous system of equations, polynomials, matrices, and real-valued functions. In each case the same handful of rules governs the arithmetic, and none of the arguments we care about uses anything beyond those rules. This chapter names the structure, a **vector space**, so that every theorem proved once holds for all of these examples at the same time.

The second half of the chapter answers a question that sounds innocent: how big is a vector space? The spaces $\mathbb{R}^3$, $\mathbb{R}[x]_{\leq 2}$, and the symmetric 2×2 matrices are all infinite sets, yet each is "three-dimensional" in a precise sense. Making that sense precise, and proving that it does not depend on any choices, takes the Exchange Theorem, the most important proof in the chapter. Everything later in the book measures with the dimension defined here.

What you need. From Chapter 0: fields (Definition 0.7.1) and their basic arithmetic (Theorem 0.7.1), sets and double inclusion (Theorem 0.3.1), functions, and induction (Theorem 0.2.7). Polynomials (Definition 0.8.1) and matrices (Definition 0.9.1) supply many of the examples.

1.0.1 Roadmap

- **Vector spaces.** The eight axioms, a gallery of examples and nonexamples, and the first consequences of the axioms.
- **Subspaces.** A subset that is a vector space in its own right, recognized by three checks.
- **Linear combinations and span.** Everything that can be built from a list of vectors, and the smallest subspace containing them.
- **Linear independence.** When a list contains no redundant vector, and the Linear Dependence Lemma that finds the redundant one when it does.
- **Bases and coordinates.** Lists that span without waste; every vector gets unique coordinates, and every finite spanning list can be trimmed to a basis.
- **The Exchange Theorem and dimension.** Independent lists are never longer than spanning lists, so all bases have the same length; extending bases and counting instead of checking.
- **Sums and direct sums.** Combining subspaces, the dimension formula, and splitting a space into independent pieces.
- **Infinite-dimensional spaces.** Every vector space has a basis (via Zorn's lemma), and which finite-dimensional facts fail without finiteness.
- **Changing the field.** The same set viewed over a smaller or larger field: restriction of scalars, the tower law, and complexification.

1.0.2 Named moves

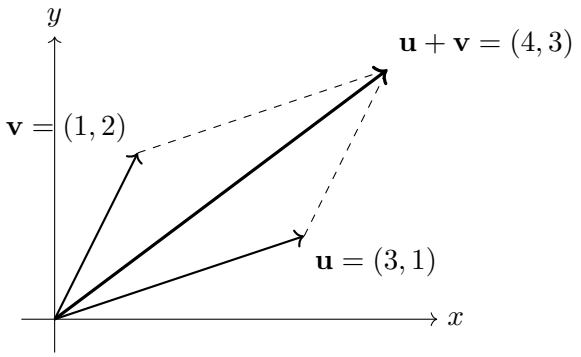
- **The three-check subspace test.** Zero vector, closed under addition, closed under scaling, each check ending with its own sentence.
- **"Let $a_1v_1 + \cdots + a_kv_k = 0$."** The first line of every independence proof; the goal is then to show every a_i is zero.
- **Exchange one vector at a time.** Swap a vector of an independent list into a spanning list without losing the span.
- **Count instead of check.** In a space of dimension n, an independent list of length n is automatically a basis, and so is a spanning list of length n.
- **Basis of the smallest space, then extend.** For any dimension formula, start from a basis of the smallest subspace in sight, extend it, and count.

1.1 Vector Spaces

Chapter 0 met several kinds of objects that can be added and scaled: pairs of real numbers, polynomials, matrices. Each time we checked the same handful of rules by hand. This section collects those rules into one definition, the vector space, and proves the first consequences once, so that they hold in every example at the same time.

1.1.1 One set of rules in three places

Start with the plane $\mathbb{R}^2$. A pair $\mathbf{u} = (3, 1)$ can be pictured as an arrow from the origin to the point $(3, 1)$. Two arrows are added tip to tail, which is the same as taking the diagonal of the parallelogram they span, and an arrow is scaled by stretching it.



The picture shows one rule at once: going along $\mathbf{u}$ and then $\mathbf{v}$ ends at the same corner as going along $\mathbf{v}$ and then $\mathbf{u}$, so $\mathbf{u} + \mathbf{v} = \mathbf{v} + \mathbf{u}$. Other rules are just as familiar. Scaling a sum is the same as adding the scaled arrows, $2(\mathbf{u} + \mathbf{v}) = 2\mathbf{u} + 2\mathbf{v}$. Adding the zero arrow changes nothing, and every arrow is undone by its reverse.

Now look at polynomials of degree at most 2. We add them coefficient by coefficient,

$$(1 + 2x - x^2) + (3 - x + 4x^2) = 4 + x + 3x^2,$$

and scale them the same way, $2(1 + 2x - x^2) = 2 + 4x - 2x^2$. There is no picture of arrows here, yet the rules are identical: order does not matter in a sum, the zero polynomial changes nothing, and $-1 - 2x + x^2$ undoes $1 + 2x - x^2$.

The same happens for 2×2 real matrices. Theorem 0.9.1 lists the rules for adding matrices and multiplying them by numbers, and they are again the same rules.

Arrows, polynomials, and matrices are very different objects, but many arguments about them only use these rules. If we prove a statement from the rules alone, it holds for arrows, polynomials, and matrices at once, and for every other system obeying the rules. So we list the rules once and give the systems that obey them a name.

A vector space is a set whose elements can be added and scaled by numbers from a field, with the usual rules of arithmetic.

1.1.2 The definition

Definition 1.1.1. Vector Space

Let F be a field. A **vector space over** F is a set V together with two functions

$$+ \colon V \times V \to V,\ (u, v) \mapsto u + v, \qquad \cdot \colon F \times V \to V,\ (a, v) \mapsto av,$$

called **addition** and **scalar multiplication**, such that the following axioms hold.

(VS1) $u + v = v + u$ **for all** $u, v \in V$.
(VS2) $(u + v) + w = u + (v + w)$ **for all** $u, v, w \in V$.
(VS3) There exists an element $0 \in V$ such that $v + 0 = v$ **for every** $v \in V$.
(VS4) **For every** $v \in V$ there exists an element $v' \in V$ such that $v + v' = 0$.
(VS5) $a(u + v) = au + av$ **for all** $a \in F$ and $u, v \in V$.
(VS6) $(a + b)v = av + bv$ **for all** $a, b \in F$ and $v \in V$.
(VS7) $a(bv) = (ab)v$ **for all** $a, b \in F$ and $v \in V$.
(VS8) $1v = v$ **for every** $v \in V$, where 1 is the identity element of F.

Read the definition clause by clause.

- **The two functions.** Addition takes two elements of V and returns an element **of** V. Scalar multiplication takes a number from F and an element of V and returns an element **of** V. So "the sum stays in the set" and "a multiple stays in the set" are not extra axioms: they are built into the word "function into V". When we test a candidate, this closure is the first thing to check.
- **(VS1)–(VS4)** concern addition alone. It is commutative and associative, there is a vector 0 that changes nothing, and every vector can be undone. These are the axioms (F1)–(F4) of Definition 0.7.1, now for vectors.
- **(VS5) and (VS6)** are the two distributive laws. In (VS6) the symbol $+$ appears twice with two meanings: on the left it adds scalars in F, on the right it adds vectors in V.
- **(VS7)** says scaling by b and then by a is scaling by the product ab, computed in F.
- **(VS8)** says the scalar 1 acts as it should. It looks too obvious to state, but we will see below that it cannot be dropped.

Note the order of quantifiers in (VS3): **one** vector 0 works for **every** v. In (VS4) the vector v' is allowed to depend on v.

Definition 1.1.2. Vectors and Scalars

Let V be a vector space over F. The elements of V are called **vectors**, and the elements of F are called **scalars**. The element 0 in (VS3) is a **zero vector** of V, and an element v' as

in (VS4) is an **additive inverse** of v.

A vector space over $\mathbb{R}$ is also called a **real vector space**, and one over $\mathbb{C}$ a **complex vector space**. When the operations are clear from context, we say "V is a vector space over F" instead of naming the triple $(V, +, \cdot)$. The field is part of the data, and we will see shortly that changing it changes the space.

We wrote "**a** zero vector" and "**an** additive inverse" on purpose. The axioms only say these exist. That there is only one of each is a theorem, proved later in this section.

Warning. **Two different zeros.** The scalar $0 \in F$ and the zero vector $0 \in V$ are different objects. In $M_2(\mathbb{R})$, the scalar is the number 0 and the zero vector is the zero matrix; in $\mathbb{R}^3$, the zero vector is $(0, 0, 0)$. Only the zero vector of F^n and the zero matrix are bold, as every coordinate vector and every matrix is: the zero vector of F^n is $\mathbf{0}$, and the zero matrix is a bold letter O, $\mathbf{O}$; in an abstract space V the zero vector 0 is printed like the scalar, so the place it stands in tells the two apart. An equation such as $0v = 0$ has a scalar on the left and a vector on the right, and mixing them up produces statements that make no sense, such as "$v = 0$" for a matrix v.

1.1.3 Examples

We start with the families we will use throughout the book. For each, the work is to say what the operations are and why the axioms hold. We check some axioms in full and point out the pattern for the rest.

Example. The Standard Vector Spaces

Let F be a field. Explain why each of the following is a vector space over F.

(a) F^n, with entrywise addition and scalar multiplication.

(b) $M_{m\times n}(F)$, with the matrix operations of Definition 0.9.6 and Definition 0.9.7.

(c) $F[x]$, with the addition of Definition 0.8.2 and $c\sum a_k x^k \coloneqq \sum (ca_k)x^k$.

(d) For a set X, the set F^X of **all** functions $X \to F$, with the **pointwise** operations

$$(f+g)(t) \coloneqq f(t) + g(t), \qquad (af)(t) \coloneqq af(t) \qquad \text{for all } t \in X.$$

(e) The set $F^{\mathbb{N}}$ of all sequences $(s_0, s_1, s_2, \dots)$ of elements of F, with termwise operations.

Solution.

(b) The sum and scalar multiple of $m \times n$ matrices are $m \times n$ matrices, so both operations are functions into $M_{m\times n}(F)$. The eight axioms are exactly the rules of Theorem 0.9.1, with the zero matrix as $\mathbf{O}$ and $-\mathbf{A}$ as an additive inverse of $\mathbf{A}$.

(c) Chapter 0 defines $F^n = M_{n\times 1}(F)$, the column vectors, and the entrywise operations are the matrix operations. So (a) is the case of (b) with n rows and 1 column. The zero vector is $(0, \dots, 0)$.

(d) The formulas define functions $f + g$ and af from X to F, so the operations land in F^X. Two functions $X \to F$ are equal when they take the same value at every $t \in X$ (Definition 0.4.1), so each axiom is checked one point at a time. For (VS3), let 0 be the

zero function, $0(t) = 0$ for all t. Then $(f+0)(t) = f(t)+0 = f(t)$ for every t, by (F3), so $f + 0 = f$. For (VS4), let $f'(t) \coloneqq -f(t)$. Then $(f + f')(t) = f(t) + (-f(t)) = 0 = 0(t)$ by (F4), so $f + f' = 0$. For (VS6), let $a, b \in F$. For every $t \in X$,

$$((a + b)f)(t) = (a + b)f(t) = af(t) + bf(t) = (af)(t) + (bf)(t) = (af + bf)(t),$$

where the second equality is (F9) together with (F6) in F. Hence $(a + b)f = af + bf$. The other five axioms reduce in the same way to a field axiom at each point: (VS1) to (F1), (VS2) to (F2), (VS5) to (F9), (VS7) to (F5), and (VS8) to (F7) with (F6).

(e) A sequence $(s_0, s_1, \dots)$ is a function $\mathbb{N} \to F$, $k \mapsto s_k$, and the termwise operations are the pointwise ones. So (e) is (d) with $X = \mathbb{N}$.

(f) Let $p = \sum a_k x^k \in F[x]$ and $c \in F$. Only finitely many a_k are nonzero, and $a_k = 0$ forces $ca_k = 0$, so cp is again a polynomial. The sum of polynomials is a polynomial by Definition 0.8.2. By Theorem 0.8.1 (parts 1 and 2), addition satisfies (VS1)–(VS4), with the zero polynomial as 0 and $-p$ as an additive inverse. Since polynomials are equal when their coefficients agree (Definition 0.8.1), (VS5)–(VS8) are checked coefficient by coefficient. For instance, the coefficient of x^k in $(c+d)p$ is $(c+d)a_k = ca_k + da_k$, which is the coefficient of x^k in $cp + dp$. The others reduce in the same way: (VS5) to (F9), (VS7) to (F5), and (VS8) to (F7) with (F6).

The pattern is worth naming. Whenever the operations are defined **entry by entry**, **point by point**, or **coefficient by coefficient**, each axiom for vectors is a field axiom applied in each slot. In fact (a), (b), and (e) are all instances of (d): a column in F^n is a function $\{1, \dots, n\} \to F$, and a matrix in $M_{m\times n}(F)$ is a function on the set of positions $\{1, \dots, m\} \times \{1, \dots, n\}$.

A **subset** of a known space often works just as well, provided the operations do not lead out of it.

The set $F[x]_{\le n}$ of Definition 0.8.4 is a vector space over F, with the operations of $F[x]$. Closure is the only real issue. If $\deg p \le n$ and $\deg q \le n$, then $\deg(p + q) \le \max(\deg p, \deg q) \le n$ by Theorem 0.8.2. If $a_k = 0$ for all $k > n$, then $ca_k = 0$ for all $k > n$, so $\deg(cp) \le n$. The zero polynomial has degree $-\infty \le n$, and $-p$ has the same coefficients as p up to sign, so it stays in $F[x]_{\le n}$. The six axioms (VS1), (VS2), (VS5)–(VS8) are equations that hold for **all** polynomials, so in particular for those of degree at most n.

The next example applies the same pattern to a set of real numbers, with a smaller field of scalars.

Example. $\mathbb{Q}(\sqrt{2})$ over $\mathbb{Q}$

Let $\mathbb{Q}(\sqrt{2}) = \{a + b\sqrt{2} : a, b \in \mathbb{Q}\}$, with the addition of $\mathbb{R}$ and with scalar multiplication $c(a + b\sqrt{2}) \coloneqq ca + cb\sqrt{2}$ for $c \in \mathbb{Q}$. Show that $\mathbb{Q}(\sqrt{2})$ is a vector space over $\mathbb{Q}$.

Solution. Let $u = a + b\sqrt{2}$ and $v = a' + b'\sqrt{2}$ with $a, b, a', b' \in \mathbb{Q}$, and let $c \in \mathbb{Q}$.

Closure. We have $u + v = (a + a') + (b + b')\sqrt{2}$ and $cu = ca + cb\sqrt{2}$. Sums and products of rationals are rational, so both lie in $\mathbb{Q}(\sqrt{2})$. Hence addition is a function $\mathbb{Q}(\sqrt{2}) \times \mathbb{Q}(\sqrt{2}) \to \mathbb{Q}(\sqrt{2})$, and scalar multiplication is a function $\mathbb{Q} \times \mathbb{Q}(\sqrt{2}) \to \mathbb{Q}(\sqrt{2})$.

Zero and additive inverses. The number $0 = 0 + 0\sqrt{2}$ lies in $\mathbb{Q}(\sqrt{2})$ and satisfies $u + 0 = u$, so (VS3) holds. The number $-u = (-a) + (-b)\sqrt{2}$ lies in $\mathbb{Q}(\sqrt{2})$ and satisfies $u + (-u) = 0$,

so (VS4) holds.

The identities. Scalar multiplication by $c \in \mathbb{Q}$ is the multiplication of real numbers, since $c(a + b\sqrt{2}) = ca + cb\sqrt{2}$ in $\mathbb{R}$. So each of (VS1), (VS2), and (VS5)–(VS8) is an identity between real numbers, and it holds by the field axioms of $\mathbb{R}$: (VS1) is (F1), (VS2) is (F2), (VS5) is (F9), (VS6) is (F9) with (F6), (VS7) is (F5), and (VS8) is (F7) with (F6). Hence $\mathbb{Q}(\sqrt{2})$ is a vector space over $\mathbb{Q}$.

We saw in Example ($\mathbb{Q}(\sqrt{2})$ Is a Field) that $\mathbb{Q}(\sqrt{2})$ is even a field. As a vector space over $\mathbb{Q}$, we simply ignore how to multiply two of its elements together and keep only multiplication by rationals.

The same set can carry more than one field of scalars. The next example makes this concrete.

Example. The Same Set over Different Fields

Explain why $\mathbb{C}$ is a vector space over $\mathbb{C}$ and also over $\mathbb{R}$, and why $\mathbb{R}$ is a vector space over $\mathbb{R}$ and also over $\mathbb{Q}$. In each case the addition is the usual one and the scalar multiplication is the usual product.

Solution. $\mathbb{C}$ over $\mathbb{C}$ is F^1 for $F = \mathbb{C}$, covered by Example (The Standard Vector Spaces) (a). For $\mathbb{C}$ over $\mathbb{R}$, a real number r is the complex number $r + 0i$, and rz is their complex product, which lies in $\mathbb{C}$. Each axiom is an identity among complex numbers in which some of the numbers happen to be real, so it holds by Proposition 0.6.1: (VS5) and (VS6) by distributivity together with commutativity, (VS7) by associativity of multiplication, and (VS8) since $1 \cdot z = z$. The argument for $\mathbb{R}$ over $\mathbb{Q}$ is the same, with the field axioms of $\mathbb{R}$ in place of the laws of $\mathbb{C}$.

As sets, $\mathbb{C}$ over $\mathbb{C}$ and $\mathbb{C}$ over $\mathbb{R}$ are the same. As vector spaces they differ, because the scalars available differ. In $\mathbb{C}$ over $\mathbb{C}$, the vector i is a scalar multiple of the vector 1. In $\mathbb{C}$ over $\mathbb{R}$ it is not, since $r \cdot 1 = i$ has no real solution r. Later in this chapter this difference becomes a difference in dimension. That is why we always say **over which field**.

The degenerate case. The set $\{0\}$ with one element, with $0 + 0 = 0$ and $a0 = 0$ for every $a \in F$, is a vector space over any field. Every axiom is an equation between elements that all equal 0, and 0 serves as the zero vector and as its own additive inverse. It is called the **zero space**. It matters because it turns up constantly as an answer: the only solution of a system, or the smallest vector space inside every other one. Unlike a field, which must contain $1 \neq 0$ (Definition 0.7.1), a vector space is allowed to have a single element, since no axiom asks for a vector different from 0.

A final example looks strange, but it satisfies every axiom.

Example. Positive Reals with Exotic Operations

Let $V = \mathbb{R}_{>0} = \{t \in \mathbb{R} : t > 0\}$. For $u, v \in V$ and $c \in \mathbb{R}$, define

$$u \boxplus v \coloneqq uv, \qquad c \boxdot v \coloneqq v^c.$$

Show that these operations are functions into V, find the zero vector and the additive inverse

of v, and check (VS6) and (VS8).

Solution. Closure. For $u, v > 0$ and $c \in \mathbb{R}$, the product uv is positive and the power v^c is positive. So $\boxplus\colon V \times V \to V$ and $\boxdot\colon \mathbb{R} \times V \to V$.

(VS3). We need $z \in V$ with $v \boxplus z = v$, that is, $vz = v$, for every $v > 0$. The number $z = 1$ works, since $v \cdot 1 = v$. So the zero vector of V is the real number 1.

(VS4). For $v \in V$, we need $v' \in V$ with $v \boxplus v' = 1$, that is, $vv' = 1$. The number $v' = 1/v$ is positive and works. So the additive inverse of v is $1/v$.

(VS6). For $a, b \in \mathbb{R}$ and $v \in V$, by the laws of exponents,

$$(a + b) \boxdot v = v^{a+b} = v^a v^b = (a \boxdot v) \boxplus (b \boxdot v).$$

(VS8). $1 \boxdot v = v^1 = v$.

The remaining axioms (VS1), (VS2), (VS5), and (VS7) are Exercise (B3: Finishing the Positive Reals). So V is a real vector space.

This example earns its place for one reason. The words "zero vector" and "addition" describe **roles** fixed by the axioms, not the familiar number 0 or the familiar $+$. Here the vector playing the role of 0 is the number 1, and "adding" means multiplying.

Quick check. Explain why each of the following, with the usual operations, is a vector space over $\mathbb{R}$.

(a) The set S of twice-differentiable functions $f\colon \mathbb{R} \to \mathbb{R}$ with $f'' + f = 0$.

(b) The set of symmetric matrices $\mathbf{A} \in M_2(\mathbb{R})$, that is, those with $\mathbf{A}^\top = \mathbf{A}$.

Solution. Both are subsets of known spaces, so we use the pattern after Example (The Standard Vector Spaces): check closure, the zero vector, and additive inverses; the other six axioms are identities inherited from the big space.

(a) $S \subseteq \mathbb{R}^{\mathbb{R}}$. If $f, g \in S$ and $c \in \mathbb{R}$, then by the sum and constant multiple rules of calculus, $f+g$ and cf are twice differentiable, with $(f+g)''+(f+g) = (f''+f)+(g''+g) = 0$ and $(cf)'' + cf = c(f'' + f) = 0$. So both lie in S. The zero function 0 satisfies $0'' + 0 = 0$, so $0 \in S$. If $f \in S$, its additive inverse in $\mathbb{R}^{\mathbb{R}}$ is $t \mapsto -f(t)$, and $-f(t) = (-1)f(t)$ by Theorem 0.7.1 (e); so it is the function $(-1)f$, which lies in S by closure under scaling. The identities hold in $\mathbb{R}^{\mathbb{R}}$ (Example (The Standard Vector Spaces) (d)), hence in S. For instance sin and cos lie in S.

(b) By Theorem 0.9.5, $(\mathbf{A} + \mathbf{B})^\top = \mathbf{A}^\top + \mathbf{B}^\top = \mathbf{A} + \mathbf{B}$ and $(c\mathbf{A})^\top = c\mathbf{A}^\top = c\mathbf{A}$ when $\mathbf{A}^\top = \mathbf{A}$ and $\mathbf{B}^\top = \mathbf{B}$. The zero matrix is symmetric, and the additive inverse $-\mathbf{A}$, which Chapter 0 defines as $(-1)\mathbf{A}$, is symmetric by closure under scaling. The identities hold in $M_2(\mathbb{R})$, hence for symmetric matrices.

1.1.4 Nonexamples

Each nonexample below is a small change to an example. We say what still works and name the exact clause that fails.

- **Change scalar multiplication on $\mathbb{R}^2$ to $c(x, y) \coloneqq (cx, 0)$**, keeping the usual addition. Addition is untouched, so (VS1)–(VS4) hold. (VS5) holds: $c((x, y) + (x', y')) = (c(x + x'), 0) = (cx, 0) + (cx', 0)$. In the same way (VS6) and (VS7) hold. But $1(0, 1) = (0, 0) \neq (0, 1)$, so **(VS8) fails**.
- **Change the line $x + 2y = 0$ to $x + 2y = 3$.** Let $L = \{(x, y) \in \mathbb{R}^2 : x + 2y = 3\}$ with the operations of $\mathbb{R}^2$. The points $(3, 0)$ and $(1, 1)$ lie on L, but their sum $(4, 1)$ has $4 + 2 \cdot 1 = 6 \neq 3$. So addition is **not a function** $L \times L \to L$, and L is not a vector space with these operations.
- **Change "degree at most 2" to "degree exactly 2".** In $\mathbb{R}[x]_{=2}$ (Definition 0.8.5), $x^2 + 2x$ and $-x^2 + 3$ have degree 2, but their sum $2x + 3$ has degree 1. Again addition leaves the set. There is no zero vector either: the zero polynomial has degree $-\infty$, not 2.
- **Change the entries from real to integer**, keeping $\mathbb{R}$ as the field. For $\mathbb{Z}^2 \subseteq \mathbb{R}^2$, addition is fine, but $\frac{1}{2}(1, 0) = (\frac{1}{2}, 0) \notin \mathbb{Z}^2$. So scalar multiplication is **not a function** $\mathbb{R} \times \mathbb{Z}^2 \to \mathbb{Z}^2$.

1.1.5 Why these axioms

Each axiom is there for a reason, and a few conventions deserve a comment.

- **(VS8) cannot be dropped.** The first nonexample satisfies all other axioms. Without (VS8), the "scalar multiplication" $cv = 0$ for all c and v would also be allowed, and scaling would carry no information at all.
- **(VS1) is in fact redundant.** Commutativity of addition follows from the other seven axioms, by expanding $(1 + 1)(u + v)$ in two ways (Exercise (C1: Commutativity Comes for Free)). It is kept because the proof is fiddly and nobody wants to repeat it.
- **Closure is not optional.** Two of the nonexamples fail only because an operation leads out of the set. That is why the definition asks for **functions into** V.
- **The scalars form a field.** We use division by nonzero scalars in the zero product law at the end of the next subsection. The warning after it shows what goes wrong without division.
- **The name.** "Vector" originally meant an arrow in physics. The word is kept for elements of any vector space, even when they are polynomials or functions.

1.1.6 First consequences of the axioms

The definition says a zero vector and additive inverses **exist**, not that they are unique. Before writing 0 and $-v$ as if each named one thing, we prove uniqueness. We also prove that familiar rules such as $0v = 0$ follow from the axioms. None of this is obvious from the axioms: in Example (Positive Reals with Exotic Operations), $0 \boxdot v = v^0 = 1$, which happens to be the zero vector there, and it takes a proof to see that this is no coincidence.

Throughout, we use one fact about functions without further comment: if two inputs are equal, the outputs are equal. So from $v = w$ we may conclude $u + v = u + w$ and $av = aw$.

The first tool lets us cancel a vector from both sides of an equation.

Theorem 1.1.1. Left Cancellation Law

Let V be a vector space over F and let $u, v, w \in V$. If $u + v = u + w$, then $v = w$.

Proof. Suppose $u + v = u + w$. By (VS4) there is $u' \in V$ with $u + u' = 0$. Adding u' on the left of both sides gives $u' + (u + v) = u' + (u + w)$, and by (VS2),

$$(u' + u) + v = (u' + u) + w.$$

By (VS1) and the choice of u', $u' + u = u + u' = 0$. Hence $0 + v = 0 + w$. By (VS1) this is $v + 0 = w + 0$, and by (VS3), $v = w$.

■

Theorem 1.1.2. Right Cancellation Law

Let V be a vector space over F and let $u, v, w \in V$. If $v + u = w + u$, then $v = w$.

Proof. By (VS1), the hypothesis says $u + v = u + w$. By Theorem 1.1.1, $v = w$.

■

Uniqueness uses the standard move from Chapter 0: suppose there are two, and show they are equal.

Theorem 1.1.3. Zero Vector Is Unique

Let V be a vector space over F. There is exactly one element $0 \in V$ satisfying (VS3).

Proof. Existence is (VS3). Suppose 0 and $0'$ both satisfy (VS3). Applying (VS3) for $0'$ to the vector 0, then (VS1), then (VS3) for 0 to the vector $0'$,

$$0 = 0 + 0' = 0' + 0 = 0'.$$

This shows the zero vector is unique.

■

Theorem 1.1.4. Additive Inverse Is Unique

Let V be a vector space over F and $v \in V$. There is exactly one $v' \in V$ with $v + v' = 0$.

Proof. Existence is (VS4). Suppose $v + v' = 0$ and $v + v'' = 0$. Then $v + v' = v + v''$, and Theorem 1.1.1 gives $v' = v''$.

■

Since the inverse is unique, it deserves a name.

Definition 1.1.3. Notation for the Additive Inverse

Let V be a vector space over F and $v \in V$. The unique additive inverse of v is denoted $-v$. For $u, v \in V$ we write $u - v := u + (-v)$.

For example, $v + (-v) = 0$ and $(-v) + v = 0$ by (VS1), so v is the additive inverse of $-v$. By Theorem 1.1.4, $-(-v) = v$.

Remark. Axioms (VS1)–(VS4) say that V with addition is an abelian group (Definition 0.10.1). So the four results above are also special cases of Theorem 0.10.1. We proved them directly to practice citing the axioms.

Now to scalars. No axiom mentions $0v$, so we must create a situation in which one applies. The move is to write the scalar 0 in a fancy way, as $0+0$, expand with a distributive law, and cancel. We will call it **expand zero and cancel**.

Theorem 1.1.5. Zero Scalar Gives the Zero Vector

Let V be a vector space over F. Then $0v = 0$ for every $v \in V$.

Proof. Let $v \in V$. Since $0 + 0 = 0$ in F by (F3), (VS6) gives

$$0v + 0v = (0 + 0)v = 0v = 0v + 0,$$

where the last equality is (VS3). By Theorem 1.1.1, $0v = 0$. ■

The same move, with the zero vector in place of the zero scalar, handles $a0$.

Theorem 1.1.6. Scalar Multiple of the Zero Vector

Let V be a vector space over F. Then $a0 = 0$ for every $a \in F$.

Proof. Let $a \in F$. Since $0 + 0 = 0$ by (VS3), (VS5) gives

$$a0 + a0 = a(0 + 0) = a0 = a0 + 0,$$

where the last equality is (VS3). By Theorem 1.1.1, $a0 = 0$. ■

To identify a vector as $-v$, it is enough to show it adds with v to 0, by Theorem 1.1.4. This gives the next two results.

Theorem 1.1.7. Negation as Scalar Multiplication

Let V be a vector space over F. Then $(-1)v = -v$ for every $v \in V$.

Proof. Let $v \in V$. By (VS8), (VS6), (F4), and Theorem 1.1.5,

$$v + (-1)v = 1v + (-1)v = (1 + (-1))v = 0v = 0.$$

So $(-1)v$ is an additive inverse of v, and by Theorem 1.1.4 it equals $-v$.

■

Theorem 1.1.8. Negative Scalars

Let V be a vector space over F. Then $(-a)v = -(av)$ for every $a \in F$ and $v \in V$.

Proof. Let $a \in F$ and $v \in V$. By (VS6), (F4), and Theorem 1.1.5,

$$av + (-a)v = (a + (-a))v = 0v = 0.$$

So $(-a)v$ is an additive inverse of av, and by Theorem 1.1.4 it equals $-(av)$.

■

The companion rules $a(-v) = -(av)$ and $a(u - v) = au - av$ are Exercise (B2: More Rules for Signs). From now on, we compute with $+$, $-$, and scalars as in school algebra, citing these results when a step is not a single axiom.

The last result is the one we will use most. In a field, a product of nonzero numbers is nonzero (Theorem 0.7.1 (f)). The same holds for a scalar times a vector.

Theorem 1.1.9. Zero Product Law

Let V be a vector space over F, let $a \in F$ and $v \in V$. Then $av = 0$ if and only if $a = 0$ or $v = 0$.

Idea. The direction ($\Leftarrow$) is the two zero results above, Theorem 1.1.5 and Theorem 1.1.6. For ($\Rightarrow$), the only way to get rid of a is to multiply by a^{-1}, and that is exactly where the hypothesis $a \neq 0$ and the field axiom (F8) are spent.

Proof. ($\Leftarrow$) If $a = 0$, then $av = 0$ by Theorem 1.1.5. If $v = 0$, then $av = 0$ by Theorem 1.1.6.

($\Rightarrow$) Suppose $av = 0$. If $a = 0$ there is nothing to prove, so suppose $a \neq 0$. By (F8), a has an inverse $a^{-1} \in F$. Then

$$v = 1v = (a^{-1}a)v = a^{-1}(av) = a^{-1}0 = 0,$$

by (VS8), (F8) with (F6), (VS7), the hypothesis, and Theorem 1.1.6. This shows $a = 0$ or $v = 0$.

■

Warning. **The zero product law needs division by scalars.** Suppose we allowed integer "scalars" and let $V = \mathbb{F}_2 = \{0, 1\}$ with its own addition, where $n \cdot v$ is v for odd n and 0 for even n (for $n \geq 0$ this is $v + \cdots + v$ with n terms). Every rule of the shape (VS1)–(VS8) still holds. But $2 \cdot 1 = 1 + 1 = 0$, although $2 \neq 0$ in $\mathbb{Z}$ and $1 \neq 0$ in V. The proof above breaks exactly where it multiplies by a^{-1}: the integer 2 has no inverse in $\mathbb{Z}$. This is one reason vector spaces are defined **over a field**.

As a first use, here is a fact about sizes. Let $v \neq 0$ be a vector in a space over $\mathbb{R}$. If $av = bv$, then $(a - b)v = av - bv = 0$ by (VS6) and Theorem 1.1.8, so $a = b$ by Theorem 1.1.9. So the multiples av, $a \in \mathbb{R}$, are all different, and a real vector space other than $\{0\}$ has infinitely many elements (Exercise (C2: How Big Is a Vector Space?) treats a general field). This is our first glimpse of the question at the heart of this chapter: all these spaces are infinite sets, so how should we measure their size? The next section starts on the answer by looking at the vector spaces that sit inside a given one.

1.1.7 Exercises

1.1.7.1 A. Check your understanding

Exercise. A1

(a) State the eight axioms of a vector space over a field F, and say which of them involve scalars.

(b) Identify the zero vector in each of: $M_{2\times 3}(\mathbb{R})$; $\mathbb{R}[x]_{\leq 2}$; $\mathbb{R}^{\mathbb{R}}$; the space $\mathbb{R}_{>0}$ of Example (Positive Reals with Exotic Operations).

(c) True or false: a vector space can have exactly two elements. Justify your answer.

(d) True or false: in every vector space, $v + v = 0$ implies $v = 0$. Justify your answer.

(e) True or false: if $av = bv$ and $v \neq 0$, then $a = b$. Justify your answer.

Solution.

(a) See Definition 1.1.1. The axioms (VS5)–(VS8) involve scalars; (VS1)–(VS4) involve only addition.

(b) The zero matrix of size 2×3; the zero polynomial; the zero function $t \mapsto 0$; the number 1 (Example (Positive Reals with Exotic Operations)).

(c) True. Let $F = \mathbb{F}_2$. By Example (The Standard Vector Spaces) (a), $\mathbb{F}_2^1 = \mathbb{F}_2 = \{0, 1\}$ is a vector space over $\mathbb{F}_2$, and it has exactly two elements.

(d) False. In the vector space $\mathbb{F}_2$ over $\mathbb{F}_2$ from (c), the vector $1 \neq 0$ satisfies $1 + 1 = 0$. (Over $\mathbb{R}$ the statement is true: $v + v = (1 + 1)v = 2v$ by (VS8) and (VS6), and $2 \neq 0$, so $v = 0$ by Theorem 1.1.9.)

(e) True. By (VS6) and Theorem 1.1.8, $(a - b)v = av + (-b)v = av - bv = 0$. Since $v \neq 0$,

Theorem 1.1.9 gives $a - b = 0$, so $a = b$.

1.1.7.2 B. Practice

Exercise. B1: Which Are Vector Spaces?

Determine which of the following are vector spaces. Justify your answer.

(a) $\mathbb{R}^2$ over $\mathbb{R}$, with the usual scalar multiplication but with addition $(x, y) + (x', y') \coloneqq (x + x', 0)$.

(b) The set of upper triangular matrices in $M_2(\mathbb{R})$, over $\mathbb{R}$, with the matrix operations.

(c) $\{p \in \mathbb{R}[x] : p(1) = 0\}$, over $\mathbb{R}$, with the operations of $\mathbb{R}[x]$.

(d) $\{p \in \mathbb{R}[x] : p(1) = 1\}$, over $\mathbb{R}$, with the operations of $\mathbb{R}[x]$.

(e) $\mathbb{R}^2$ over $\mathbb{C}$, with $c(x, y) \coloneqq (cx, cy)$ and the usual addition.

Solution.

(a) Not a vector space: (VS3) fails. Suppose $\mathbf{0} = (z_1, z_2)$ satisfied (VS3). Then $(0, 1) + (z_1, z_2) = (z_1, 0)$ would equal $(0, 1)$, which is impossible since the second entries differ.

(b) A vector space. Let $\mathbf{A} = \begin{pmatrix} a & b \\ 0 & d \end{pmatrix}$ and $\mathbf{B} = \begin{pmatrix} a' & b' \\ 0 & d' \end{pmatrix}$ be upper triangular and $c \in \mathbb{R}$. Then $\mathbf{A} + \mathbf{B} = \begin{pmatrix} a + a' & b + b' \\ 0 & d + d' \end{pmatrix}$ and $c\mathbf{A} = \begin{pmatrix} ca & cb \\ 0 & cd \end{pmatrix}$ are upper triangular, so the operations land in the set. The zero matrix is upper triangular, and so is $-\mathbf{A} = (-1)\mathbf{A}$. The identities (VS1), (VS2), (VS5)–(VS8) hold for all matrices in $M_2(\mathbb{R})$ by Example (The Standard Vector Spaces) (b), hence for upper triangular ones.

(c) A vector space. Let $U = \{p \in \mathbb{R}[x] : p(1) = 0\}$, $p, q \in U$, and $c \in \mathbb{R}$. By Theorem 0.8.5, $(p + q)(1) = p(1) + q(1) = 0$. Also cp is the product of the constant polynomial c with p, since by the product formula of Definition 0.8.2 that product has coefficients ca_k. So $(cp)(1) = c\,p(1) = 0$. Hence $p + q, cp \in U$. The zero polynomial has value 0 at 1, and $-p = (-1)p \in U$ by closure under scaling. The identities hold in $\mathbb{R}[x]$ by Example (The Standard Vector Spaces) (c), hence in U.

(d) Not a vector space: addition leaves the set. The constant polynomial 1 lies in the set, but $(1 + 1)(1) = 2 \neq 1$, so $1 + 1$ does not.

(e) Not a vector space: scalar multiplication leaves the set. We have $i \in \mathbb{C}$ and $(1, 0) \in \mathbb{R}^2$, but $i(1, 0) = (i, 0) \notin \mathbb{R}^2$. So the rule is not a function $\mathbb{C} \times \mathbb{R}^2 \to \mathbb{R}^2$.

Exercise. B2: More Rules for Signs

Let V be a vector space over F, let $a, b \in F$ and $u, v \in V$. Prove that:

(a) $a(-v) = -(av)$;

(b) $(-a)(-v) = av$;

(c) $a(u - v) = au - av$ and $(a - b)v = av - bv$.

Hint: to show a vector equals $-(av)$, use Theorem 1.1.4.

Solution.

(a) By (VS5), (VS4), and Theorem 1.1.6,

$$av + a(-v) = a(v + (-v)) = a0 = 0.$$

So $a(-v)$ is an additive inverse of av. By Theorem 1.1.4, $a(-v) = -(av)$.

(b) By (a) applied with the scalar $-a$, then Theorem 1.1.8, then the remark after Definition 1.1.3,

$$(-a)(-v) = -((-a)v) = -(-(av)) = av.$$

(c) By Definition 1.1.3, (VS5), and (a),

$$a(u - v) = a(u + (-v)) = au + a(-v) = au + (-(av)) = au - av.$$

Similarly, since $a - b = a + (-b)$ in F, (VS6) and Theorem 1.1.8 give $(a - b)v = av + (-b)v = av + (-(bv)) = av - bv$.

Exercise. B3: Finishing the Positive Reals

Let $V = \mathbb{R}_{>0}$ with the operations $\boxplus$ and $\boxdot$ of Example (Positive Reals with Exotic Operations).

(a) Verify (VS1), (VS2), (VS5), and (VS7). Hence V is a vector space over $\mathbb{R}$.

(b) Compute the vector $(2 \boxdot 3) \boxplus ((-1) \boxdot 6)$.

(c) Find the vector $-(2 \boxdot 3)$, and check your answer against Theorem 1.1.8.

Solution.

(a) Let $u, v, w \in V$ and $a, b \in \mathbb{R}$. By the laws of real multiplication and of exponents for positive bases:

- (VS1): $u \boxplus v = uv = vu = v \boxplus u$.
- (VS2): $(u \boxplus v) \boxplus w = (uv)w = u(vw) = u \boxplus (v \boxplus w)$.
- (VS5): $a \boxdot (u \boxplus v) = (uv)^a = u^a v^a = (a \boxdot u) \boxplus (a \boxdot v)$.
- (VS7): $a \boxdot (b \boxdot v) = (v^b)^a = v^{ab} = (ab) \boxdot v$.

Together with closure, (VS3), (VS4), (VS6), and (VS8) from Example (Positive Reals with Exotic Operations), all eight axioms hold. Hence V is a vector space over $\mathbb{R}$.

(b) $2 \boxdot 3 = 3^2 = 9$ and $(-1) \boxdot 6 = 6^{-1} = \frac{1}{6}$. So the vector is $9 \boxplus \frac{1}{6} = 9 \cdot \frac{1}{6} = \frac{3}{2}$.

(c) By Example (Positive Reals with Exotic Operations), the additive inverse of a vector v is $1/v$. Since $2 \boxdot 3 = 9$, we get $-(2 \boxdot 3) = \frac{1}{9}$. By Theorem 1.1.8, this should equal $(-2) \boxdot 3 = 3^{-2} = \frac{1}{9}$, and it does.

1.1.7.3 C. Going deeper

Exercise. C1: Commutativity Comes for Free

Let V be a set with operations $+ \colon V \times V \to V$ and $\cdot \colon F \times V \to V$ satisfying (VS2)–(VS8), but **not necessarily** (VS1). Fix 0 as in (VS3).

(a) Let $v \in V$ and v' as in (VS4). Put $w = v' + v$. Show that $w + w = w$, and deduce that $w = 0$.

(b) Show that $0 + v = v$ for every $v \in V$.

(c) Deduce that both cancellation laws hold: $u+v = u+w$ implies $v = w$, and $v+u = w+u$ implies $v = w$.

(d) By expanding $(1 + 1)(u + v)$ in two ways, prove that $u + v = v + u$ for all $u, v \in V$.

Hint: in (a), add an additive inverse of w on the right. The proof of Theorem 1.1.1 used (VS1), so it cannot be quoted in (c).

Solution. Throughout we use only (VS2)–(VS8), and we drop brackets in sums whenever (VS2) allows it.

(a) Since $v + v' = 0$ and $v' + 0 = v'$ by (VS3),

$$w + w = v' + (v + v') + v = v' + 0 + v = v' + v = w.$$

By (VS4) choose w' with $w + w' = 0$. Adding w' on the right of $w + w = w$ gives $w + (w + w') = w + w'$, that is, $w + 0 = 0$. By (VS3), $w = 0$. So $v' + v = 0$ as well.

(b) Let $v \in V$ and v' as in (VS4). By (a), (VS2), and (VS3),

$$0 + v = (v + v') + v = v + (v' + v) = v + 0 = v.$$

(c) Suppose $u + v = u + w$, and let u' be as in (VS4). Adding u' on the left and using (VS2), $(u' + u) + v = (u' + u) + w$. By (a), $u' + u = 0$, so $0 + v = 0 + w$, and (b) gives $v = w$. Now suppose $v + u = w + u$. Adding u' on the right and using (VS2), $v + (u + u') = w + (u + u')$, so $v + 0 = w + 0$, and (VS3) gives $v = w$.

(d) Let $u, v \in V$. On the one hand, by (VS6) and (VS8),

$$(1 + 1)(u + v) = 1(u + v) + 1(u + v) = u + v + u + v.$$

On the other hand, by (VS5), then (VS6) and (VS8),

$$(1 + 1)(u + v) = (1 + 1)u + (1 + 1)v = u + u + v + v.$$

Hence $u + (v + u) + v = u + (u + v) + v$, grouping by (VS2). Canceling u on the left and then v on the right, by (c), gives $v + u = u + v$. This proves (VS1).

Exercise. C2: How Big Is a Vector Space?

(a) Let F be an infinite field and $V \neq \{0\}$ a vector space over F. Prove that V is an infinite set.

(b) Let p be a prime and $n \geq 1$. How many elements does the vector space $\mathbb{F}_p^n$ over $\mathbb{F}_p$ have? Explain why this does not contradict (a).

Hint: for (a), fix $v \neq 0$ and consider the function $F \to V$, $a \mapsto av$.

Solution.

(a) Since $V \neq \{0\}$ and $0 \in V$, there is $v \in V$ with $v \neq 0$. Consider $\varphi\colon F \to V$, $\varphi(a) = av$. Suppose $\varphi(a) = \varphi(b)$, so $av = bv$. By Exercise (B2: More Rules for Signs) (c), $(a-b)v = av - bv = 0$. Since $v \neq 0$, Theorem 1.1.9 gives $a - b = 0$, so $a = b$. Therefore φ is injective. An injective function from an infinite set has an infinite image, so V, which contains the image of φ, is infinite.

(b) An element of $\mathbb{F}_p^n$ is a list $(x_1, \ldots, x_n)$ with each x_i chosen freely from the p elements of $\mathbb{F}_p$. So $\mathbb{F}_p^n$ has p^n elements. There is no contradiction, because (a) assumes the field is infinite and $\mathbb{F}_p$ is finite. The proof of (a) still shows that $a \mapsto av$ is injective, so a nonzero space over $\mathbb{F}_p$ has at least p elements.

Exercise. C3: Shrinking the Field

Let F be a field and let $K \subseteq F$ be a subset with $0, 1 \in K$ that is closed under addition, multiplication, negatives, and inverses of nonzero elements. (So K is itself a field with the operations of F; for example $\mathbb{Q} \subseteq \mathbb{R} \subseteq \mathbb{C}$.)

(a) Let V be a vector space over F. Prove that V, with the same addition and with scalar multiplication restricted to $K \times V$, is a vector space over K.

(b) Deduce that $\mathbb{C}^n$ is a vector space over $\mathbb{R}$ and over $\mathbb{Q}$.

(c) Explain why $\mathbb{R}$, with its usual addition and multiplication, is **not** a vector space over $\mathbb{C}$.

Solution.

(a) The addition is unchanged, so it is a function $V \times V \to V$ satisfying (VS1)–(VS4). Scalar multiplication $F \times V \to V$ restricts to a function $K \times V \to V$, since $K \subseteq F$. The axioms (VS5)–(VS7) are statements about **all** scalars in F, so they hold in particular for all scalars in K; the operations $a + b$ and ab appearing in them are the same in K as in F. For (VS8), the identity of K is the identity 1 of F, and $1v = v$ holds in V over F. Hence V is a vector space over K.

(b) $\mathbb{C}^n$ is a vector space over $\mathbb{C}$ by Example (The Standard Vector Spaces) (a). The subsets $\mathbb{R} \subseteq \mathbb{C}$ and $\mathbb{Q} \subseteq \mathbb{C}$ contain 0 and 1 and are closed under the four operations, so (a) applies to both.

(c) The only candidate for scalar multiplication is $(c, t) \mapsto ct$, the product in $\mathbb{C}$. But it does not land in $\mathbb{R}$: $i \in \mathbb{C}$ and $1 \in \mathbb{R}$, while $i \cdot 1 = i \notin \mathbb{R}$. So there is no function $\mathbb{C} \times \mathbb{R} \to \mathbb{R}$ given by this rule, and $\mathbb{R}$ is not a vector space over $\mathbb{C}$ with these operations.

1.2 Subspaces

After defining a structure, the next step is to look at the smaller copies of it that sit inside. In the previous section we showed three times, for $F[x]_{\leq n}$, $\mathbb{Q}(\sqrt{2})$, and the solutions of $f'' + f = 0$, that a subset of a known vector space is itself a vector space, and each time the same few checks did the work. This section turns that pattern into a theorem, the subspace test, and makes its three checks a routine we will use for the rest of the book.

1.2.1 Spaces inside spaces

Consider the set

$$P = \{(x, y, z) \in \mathbb{R}^3 : x + y - z = 0\},$$

a plane through the origin in $\mathbb{R}^3$. It contains $(1, 0, 1)$ and $(0, 1, 1)$, and also their sum $(1, 1, 2)$, since $1 + 1 - 2 = 0$. Stretching a vector of P keeps it in the plane. So P, with the addition and scalar multiplication of $\mathbb{R}^3$, looks like a vector space in its own right.

To prove it we could check all eight axioms again. That is wasteful, and the previous section already showed why. The axioms (VS1), (VS2), and (VS5)–(VS8) are equations that hold for **all** vectors of $\mathbb{R}^3$, so they hold in particular for the vectors of P. What can go wrong is different: a sum or a multiple might leave P, and the zero vector or an additive inverse might not lie in P. The question of this section is which of these checks are really needed.

A subspace is a subset that is a vector space in its own right, using the same addition and scalar multiplication.

Definition 1.2.1. Subspace

Let V be a vector space over F. A subset $U \subseteq V$ is a **subspace** of V if U is **itself a vector space over** F when it is given the addition and the scalar multiplication of V.

For that sentence to make sense, U must first be closed under those two operations:

$$u + w \in U \quad \text{and} \quad au \in U \qquad \text{for all } u, w \in U \text{ and all } a \in F.$$

In words: take the plus and the dot that V already has, use them only on vectors of U, and ask whether U is a vector space with them. Nothing new is defined; U borrows everything from V. The displayed condition is what makes the borrowing possible, because an operation on U must land **in** U: if $u, w \in U$ but $u + w \notin U$, then addition is not an operation on U at all, and the question of the axioms never arises. Once the condition holds, the eight axioms are asked of U exactly as they were asked of V in the previous section. In particular U must contain **a** zero vector of its own, and every vector of U must have an additive inverse **lying in** U. The field does not change: a subspace of a space over F is again a space over F.

Two words of the definition carry all the weight. **Itself a vector space** is the whole content. **The same operations** is what distinguishes a subspace from a subset that merely happens to be a vector space under some other rule: $\mathbb{R}_{>0}$ with the operations of Example (Positive Reals with Exotic Operations) is a vector space and a subset of $\mathbb{R}$, but it is not a subspace of $\mathbb{R}$, because those are not the operations of $\mathbb{R}$.

There is a question hidden in "a zero vector of its own". Could U have a zero vector different from the zero vector of V? The proof of the subspace test below shows that it cannot.

The two extreme cases come first.

Example. Trivial Subspaces

Let V be a vector space over F. Show that V and $\{0\}$ are subspaces of V, and that the empty set $\varnothing$ is not.

Solution. For V itself, the restricted operations are the operations of V, and V is a vector space by assumption.

For $\{0\}$, the only sum is $0+0=0$ by (VS3), and $a0=0$ for every $a \in F$ by Theorem 1.1.6. So both operations stay in $\{0\}$, and $\{0\}$ with them is the zero space, a vector space as noted in the previous section.

For $\varnothing$, axiom (VS3) demands that some element 0 exists in the set. The empty set has no elements, so (VS3) fails and $\varnothing$ is not a subspace.

The subspace $\{0\}$ is the **zero subspace**. A subspace other than V is called **proper**. The zero subspace is small, but it is the answer to many questions later: the solutions of a system with only the trivial solution, or the intersection of two lines through the origin that are not equal.

1.2.2 The subspace test

The following theorem says that three checks suffice. Two of them are the closure conditions of the definition, and the third replaces all of the axioms.

Theorem 1.2.1. Subspace Test

Let V be a vector space over F and let $U \subseteq V$. Then U is a subspace of V if and only if the following three conditions hold.

(1) $0 \in U$, where 0 is the zero vector of V.
(2) U is **closed under addition**: $u+w \in U$ for all $u, w \in U$.
(3) U is **closed under scalar multiplication**: $au \in U$ for all $a \in F$ and $u \in U$.

Idea. Sort the eight axioms into two kinds. Six of them, (VS1), (VS2), and (VS5)–(VS8), say "an equation holds **for all** vectors"; they are true in V, so they are automatically true for the fewer vectors in U. The other two, (VS3) and (VS4), say "**there exists** a vector", and a witness must be found **inside** U. Condition (1) supplies the zero. For inverses, the results of the previous section give $-u=(-1)u$, so closure under scaling supplies them for free. For ($\Rightarrow$), the only nonobvious point is that the zero vector of U is the zero vector of V; cancellation in V settles it.

Proof. ($\Rightarrow$) Suppose U is a subspace of V. Conditions (2) and (3) are the closure condition of Definition 1.2.1. By (VS3) in U, there is $z \in U$ with $u+z=u$ for every $u \in U$; in particular $z+z=z$. By (VS3) in V, also $z+0=z$. Hence $z+z=z+0$, and Theorem 1.1.1 in V gives $z=0$. Therefore $0 \in U$, which is (1).

($\Leftarrow$) Suppose (1), (2), and (3) hold. By (2) and (3), addition and scalar multiplication of V restrict to functions $U \times U \to U$ and $F \times U \to U$. We check the axioms for U with these operations.

- (VS1), (VS2), (VS5), (VS6), (VS7), and (VS8) state that certain equations hold for all vectors and scalars. They hold for all vectors of V, since V is a vector space, so in particular for all vectors of U.

- (VS3): by (1), $0 \in U$, and $u + 0 = u$ for every $u \in U$ by (VS3) in V.
- (VS4): let $u \in U$. By (3), $(-1)u \in U$. By Theorem 1.1.7, $(-1)u = -u$, so $u + (-1)u = 0$. Thus u has an additive inverse in U.

Hence U is a vector space over F with the operations of V, that is, a subspace of V.

■

From now on we never check eight axioms for a subset of a known vector space. We check three things, always in the same order and with the same wording. This is the first named move of the chapter.

The three-check move. To show that a subset U of a vector space V is a subspace:

- make sure U is a subset of V (what are its elements, and do they lie in V?);
- **(1) Zero vector.** Show $0 \in U$ by checking that 0 satisfies the condition defining U. End with "Therefore $0 \in U$."
- **(2) Closed under addition.** Let $u, w \in U$. Write down what membership in U means for each, then regroup $u + w$ so that those conditions appear. End with "This shows that U is closed under addition."
- **(3) Closed under scalar multiplication.** Let $a \in F$ and $u \in U$, and do the same for au. End with "This shows that U is closed under scalar multiplication."
- Conclude: "Hence U is a subspace of V by the subspace test (Theorem 1.2.1)."

To show that U is **not** a subspace, one failure suffices. Check (1) first, since it is the cheapest, and if it holds, look for a concrete pair of vectors or a concrete scalar that leads out of U.

Remark. Condition (1) may be replaced by "U is nonempty". Indeed, if U contains some u, then (3) gives $0u \in U$, and $0u = 0$ by Theorem 1.1.5. We still prefer (1): the zero vector is the easiest element to test, and when it is missing we have a refutation in one line.

Warning. **The empty set is never a subspace.** It is tempting to check only (2) and (3). But $\varnothing$ satisfies both, vacuously: there are no vectors in it to add or scale. It fails only (1). So whichever version of the test you use, do not skip the first condition, and remember that "$0 \in U$" and "U is nonempty" are interchangeable only in the presence of (3).

1.2.3 Examples: the three-check move

Every example below follows the move exactly. The first is the model.

Example. A Line through the Origin

Show that $U = \{(x, y) \in \mathbb{R}^2 : x + 2y = 0\}$ is a subspace of $\mathbb{R}^2$.

Solution. By definition, U is a subset of $\mathbb{R}^2$.

(1) Zero vector. The zero vector $(0, 0)$ satisfies $0 + 2 \cdot 0 = 0$. Therefore $\mathbf{0} \in U$.

(2) Closed under addition. Let $\mathbf{u} = (x_1, y_1)$ and $\mathbf{w} = (x_2, y_2)$ be in U, so that $x_1 + 2y_1 = 0$ and $x_2 + 2y_2 = 0$. Then $\mathbf{u} + \mathbf{w} = (x_1 + x_2, y_1 + y_2)$, and

$$(x_1 + x_2) + 2(y_1 + y_2) = (x_1 + 2y_1) + (x_2 + 2y_2) = 0 + 0 = 0.$$

Therefore $\mathbf{u} + \mathbf{w} \in U$. This shows that U is closed under addition.

(3) Closed under scalar multiplication. Let $a \in \mathbb{R}$ and $\mathbf{u} = (x, y) \in U$, so that $x + 2y = 0$. Then $a\mathbf{u} = (ax, ay)$, and

$$ax + 2(ay) = a(x + 2y) = a \cdot 0 = 0.$$

Therefore $a\mathbf{u} \in U$. This shows that U is closed under scalar multiplication.

Hence U is a subspace of $\mathbb{R}^2$ by the subspace test (Theorem 1.2.1).

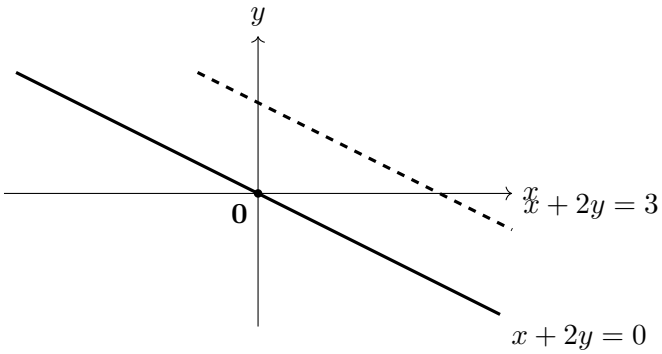
Change the right-hand side from 0 to 3, and everything breaks.

Example. A Line Missing the Origin

Show that $L = \{(x, y) \in \mathbb{R}^2 : x + 2y = 3\}$ is **not** a subspace of $\mathbb{R}^2$.

Solution. Condition (1) fails: the zero vector has $0 + 2 \cdot 0 = 0 \neq 3$, so $\mathbf{0} \notin L$. By the subspace test (Theorem 1.2.1), L is not a subspace of $\mathbb{R}^2$.

The other two conditions fail as well, as we saw in the previous section: $(3, 0) + (1, 1) = (4, 1) \notin L$, and $2(3, 0) = (6, 0) \notin L$. One failure is enough, though, and (1) is the cheapest to find.



The solid line U passes through the origin; the dashed line L is parallel to it and misses the origin. The line L is a translate of a subspace, a kind of set we will study in Chapter 4, but it is not a subspace.

The line U is the set of solutions of one homogeneous equation. The same three checks work for any number of homogeneous equations, written as one matrix equation.

Example. Solutions of a Homogeneous System

Let $\mathbf{A} \in M_{m \times n}(F)$. Show that $N = \{\mathbf{x} \in F^n : \mathbf{A}\mathbf{x} = \mathbf{0}\}$ is a subspace of F^n.

Solution. By definition, N is a subset of F^n. Recall that $F^n = M_{n\times 1}(F)$, so $\mathbf{Ax} \in F^m$ is a matrix product, and we may use Theorem 0.9.4.

(1) Zero vector. By the zero rule of Theorem 0.9.4, $\mathbf{A0} = \mathbf{0}$, where the first $\mathbf{0}$ is the zero column in F^n and the second is the zero column in F^m. Therefore $\mathbf{0} \in N$.

(2) Closed under addition. Let $\mathbf{x}, \mathbf{y} \in N$, so that $\mathbf{Ax} = \mathbf{0}$ and $\mathbf{Ay} = \mathbf{0}$. By distributivity (Theorem 0.9.4),

$$\mathbf{A}(\mathbf{x} + \mathbf{y}) = \mathbf{Ax} + \mathbf{Ay} = \mathbf{0} + \mathbf{0} = \mathbf{0}.$$

Therefore $\mathbf{x} + \mathbf{y} \in N$. This shows that N is closed under addition.

(3) Closed under scalar multiplication. Let $a \in F$ and $\mathbf{x} \in N$. By the scalar rule of Theorem 0.9.4 and Theorem 1.1.6 in F^m,

$$\mathbf{A}(a\mathbf{x}) = a(\mathbf{Ax}) = a\mathbf{0} = \mathbf{0}.$$

Therefore $a\mathbf{x} \in N$. This shows that N is closed under scalar multiplication.

Hence N is a subspace of F^n by the subspace test (Theorem 1.2.1).

With $\mathbf{A} = \begin{pmatrix} 1 & 2 \end{pmatrix}$ and $F = \mathbb{R}$, this is Example (A Line through the Origin) again, and with $\mathbf{A} = \begin{pmatrix} 1 & 1 & -1 \end{pmatrix}$ it is the plane P from the start of the section. The word **homogeneous** (right-hand side $\mathbf{0}$) is essential: Example (A Line Missing the Origin) is a system with right-hand side 3. This subspace will be called the null space of $\mathbf{A}$ in Chapter 3.

Next, polynomials. The pair "degree at most n" and "degree exactly n" is another minimal change. The previous section checked the first of them by hand; here it is as a three-check move.

Example. Polynomials of Bounded Degree

Let $n \in \mathbb{N}$. Show that $F[x]_{\leq n}$ is a subspace of $F[x]$.

Solution. By Definition 0.8.4, $F[x]_{\leq n}$ is a subset of $F[x]$.

(1) Zero vector. The zero polynomial has degree $-\infty \leq n$ (Definition 0.8.3). Therefore $0 \in F[x]_{\leq n}$.

(2) Closed under addition. Let $p, q \in F[x]_{\leq n}$, so that $\deg p \leq n$ and $\deg q \leq n$. By Theorem 0.8.2, $\deg(p+q) \leq \max(\deg p, \deg q) \leq n$. Therefore $p + q \in F[x]_{\leq n}$. This shows that $F[x]_{\leq n}$ is closed under addition.

(3) Closed under scalar multiplication. Let $c \in F$ and $p = a_0 + a_1x + \cdots + a_nx^n \in F[x]_{\leq n}$. Then $cp = ca_0 + ca_1x + \cdots + ca_nx^n$, whose coefficients of x^k for $k > n$ are all 0. So $\deg(cp) \leq n$, and therefore $cp \in F[x]_{\leq n}$. This shows that $F[x]_{\leq n}$ is closed under scalar multiplication.

Hence $F[x]_{\leq n}$ is a subspace of $F[x]$ by the subspace test (Theorem 1.2.1).

Example. Polynomials of Exact Degree

Let $n \in \mathbb{N}$. Show that $F[x]_{=n}$ is **not** a subspace of $F[x]$.

Solution. Condition (1) fails: the zero polynomial has degree $-\infty \neq n$, so $0 \notin F[x]_{=n}$. By the subspace test (Theorem 1.2.1), $F[x]_{=n}$ is not a subspace of $F[x]$.

Even if we added the zero polynomial by hand, condition (2) would still fail for $n \geq 1$: the polynomials $x^n + 1$ and $-x^n$ have degree exactly n, but their sum is the constant 1, of degree 0. The leading terms cancel, and nothing in "degree exactly n" prevents it.

The same move works in function spaces. Here the facts we need come from calculus.

Example. Differentiable Functions

Let D be the set of differentiable functions $\mathbb{R} \to \mathbb{R}$. Show that D is a subspace of $\mathbb{R}^{\mathbb{R}}$.

Solution. Every element of D is a function $\mathbb{R} \to \mathbb{R}$, so $D \subseteq \mathbb{R}^{\mathbb{R}}$.

(1) Zero vector. The zero vector of $\mathbb{R}^{\mathbb{R}}$ is the zero function (Example (The Standard Vector Spaces)), which is constant and hence differentiable. Therefore $0 \in D$.

(2) Closed under addition. Let $f, g \in D$. By the sum rule of calculus, $f + g$ is differentiable, with $(f + g)' = f' + g'$. Therefore $f + g \in D$. This shows that D is closed under addition.

(3) Closed under scalar multiplication. Let $a \in \mathbb{R}$ and $f \in D$. By the constant multiple rule, af is differentiable, with $(af)' = af'$. Therefore $af \in D$. This shows that D is closed under scalar multiplication.

Hence D is a subspace of $\mathbb{R}^{\mathbb{R}}$ by the subspace test (Theorem 1.2.1).

The same three checks show that the continuous functions $\mathbb{R} \to \mathbb{R}$ form a subspace of $\mathbb{R}^{\mathbb{R}}$, and that D is a subspace of it. A subspace of a subspace is a subspace, since the operations are the same throughout.

Finally, two families of matrices. Each uses one rule from Chapter 0 for all three checks.

Example. Symmetric Matrices

Show that the set $S = \{\mathbf{A} \in M_n(F) : \mathbf{A}^\top = \mathbf{A}\}$ of symmetric matrices (Definition 0.9.11) is a subspace of $M_n(F)$.

Solution. By definition, S is a subset of $M_n(F)$.

(1) Zero vector. Every entry of the zero matrix is 0, so $(\mathbf{O})_{ji} = 0 = (\mathbf{O})_{ij}$ for all i, j, and the zero matrix is symmetric. Therefore $\mathbf{O} \in S$.

(2) Closed under addition. Let $\mathbf{A}, \mathbf{B} \in S$, so that $\mathbf{A}^\top = \mathbf{A}$ and $\mathbf{B}^\top = \mathbf{B}$. By Theo-

rem 0.9.5,

$$(\mathbf{A}+\mathbf{B})^\top = \mathbf{A}^\top + \mathbf{B}^\top = \mathbf{A}+\mathbf{B}.$$

Therefore $\mathbf{A}+\mathbf{B} \in S$. This shows that S is closed under addition.

(3) Closed under scalar multiplication. Let $c \in F$ and $\mathbf{A} \in S$. By Theorem 0.9.5, $(c\mathbf{A})^\top = c\mathbf{A}^\top = c\mathbf{A}$. Therefore $c\mathbf{A} \in S$. This shows that S is closed under scalar multiplication.

Hence S is a subspace of $M_n(F)$ by the subspace test (Theorem 1.2.1).

Example. Trace-Free Matrices

Show that the set $T = \{\mathbf{A} \in M_n(F) : \operatorname{tr}\mathbf{A} = 0\}$ of trace-free matrices is a subspace of $M_n(F)$.

Solution. By definition, T is a subset of $M_n(F)$.

(1) Zero vector. The zero matrix has trace $0+\cdots+0=0$. Therefore $\mathbf{O} \in T$.

(2) Closed under addition. Let $\mathbf{A}, \mathbf{B} \in T$, so that $\operatorname{tr}\mathbf{A} = \operatorname{tr}\mathbf{B} = 0$. By Theorem 0.9.6, $\operatorname{tr}(\mathbf{A}+\mathbf{B}) = \operatorname{tr}\mathbf{A} + \operatorname{tr}\mathbf{B} = 0+0=0$. Therefore $\mathbf{A}+\mathbf{B} \in T$. This shows that T is closed under addition.

(3) Closed under scalar multiplication. Let $c \in F$ and $\mathbf{A} \in T$. By Theorem 0.9.6, $\operatorname{tr}(c\mathbf{A}) = c\operatorname{tr}\mathbf{A} = c\cdot 0 = 0$. Therefore $c\mathbf{A} \in T$. This shows that T is closed under scalar multiplication.

Hence T is a subspace of $M_n(F)$ by the subspace test (Theorem 1.2.1).

The pattern in the last two examples, and in Example (Solutions of a Homogeneous System), is the same: the condition defining the set is "some rule that respects sums and scalar multiples sends the element to zero". Chapter 2 gives such rules a name, linear maps, and this pattern becomes a single theorem about their kernels.

1.2.4 Nonexamples

Here are two more subsets of $\mathbb{R}^2$ that fail, each in a different way.

- **The unit circle** $\{(x,y) : x^2+y^2=1\}$ fails (1), since $0^2+0^2 \neq 1$. It fails (2) and (3) as well: $(1,0)+(0,1)=(1,1)$ and $2(1,0)=(2,0)$ are not on the circle.
- **The set** $Q = \{(x,y) : xy \geq 0\}$, the first and third quadrants together with both axes. Here (1) holds, since $0\cdot 0 = 0 \geq 0$. Condition (3) holds too: if $xy \geq 0$ and $a \in \mathbb{R}$, then $(ax)(ay) = a^2xy \geq 0$. But (2) fails: $(1,0)$ and $(0,-1)$ lie in Q, while their sum $(1,-1)$ has $1\cdot(-1) = -1 < 0$. So Q is not a subspace.

The set Q shows that (3) does not imply (2). The next check shows that (2) does not imply (3).

Quick check. Find a subset of $\mathbb{R}^2$ that contains $\mathbf{0}$ and is closed under addition, but is not a subspace. Which condition of the subspace test fails?

Solution. Take the closed first quadrant $U = \{(x, y) : x \geq 0 \text{ and } y \geq 0\}$. It contains $(0, 0)$, and if $x_1, y_1, x_2, y_2 \geq 0$, then $x_1 + x_2 \geq 0$ and $y_1 + y_2 \geq 0$, so it is closed under addition. Condition (3) fails: $(1, 0) \in U$ but $(-1)(1, 0) = (-1, 0) \notin U$. Another answer is $\mathbb{Z}^2$: it contains $\mathbf{0}$ and is closed under addition, but $\frac{1}{2}(1, 0) \notin \mathbb{Z}^2$.

1.2.5 Why this definition

The definition insists on **the same** operations. This is not a formality. The positive reals $\mathbb{R}_{>0}$ form a vector space over $\mathbb{R}$ (Example (Positive Reals with Exotic Operations)), and $\mathbb{R}_{>0}$ is a subset of the vector space $\mathbb{R}$. But $\mathbb{R}_{>0}$ is **not** a subspace of $\mathbb{R}$: with the operations of $\mathbb{R}$, it does not contain the zero vector 0, and $(-1) \cdot 1 = -1$ leaves it. Its vector space structure uses different operations, multiplication and powers, and those do not count.

The definition also keeps the field. For example, $\mathbb{R}^2$ is a subset of $\mathbb{C}^2$, and it is closed under addition and under multiplication by **real** scalars. But it is not a subspace of $\mathbb{C}^2$ over $\mathbb{C}$, because $i(1, 0) = (i, 0) \notin \mathbb{R}^2$. Condition (3) quantifies over **every** scalar of the field.

Finally, why "(1) $0 \in U$" rather than "U is nonempty"? The two are equivalent given (3), as the remark after the test explains. We choose the version that is quicker to check and quicker to refute.

1.2.6 Intersections of subspaces

Given several subspaces, can we build new ones? The first way is to take the vectors they have in common. Recall from Definition 0.3.9 that the intersection of a **nonempty** family of sets $(U_i)_{i \in I}$ consists of the elements lying in every U_i.

Theorem 1.2.2. Intersection of Subspaces

Let V be a vector space over F, and let $(U_i)_{i \in I}$ be a family of subspaces of V with I nonempty. Then $\bigcap_{i \in I} U_i$ is a subspace of V.

Proof. Let $U = \bigcap_{i \in I} U_i$. Each U_i is a subset of V, so $U \subseteq V$. We use the subspace test (Theorem 1.2.1) for U, and for each U_i its direction ($\Rightarrow$).

(1) For every $i \in I$, $0 \in U_i$, since U_i is a subspace. Therefore $0 \in U$.

(2) Let $u, w \in U$. Then $u, w \in U_i$ for every $i \in I$. Since each U_i is closed under addition, $u + w \in U_i$ for every $i \in I$. Therefore $u + w \in U$. This shows that U is closed under addition.

(3) Let $a \in F$ and $u \in U$. Then $u \in U_i$ for every $i \in I$, and since each U_i is closed under scalar multiplication, $au \in U_i$ for every $i \in I$. Therefore $au \in U$. This shows that U is closed under scalar multiplication.

Hence U is a subspace of V by the subspace test (Theorem 1.2.1).

■

The index set may be infinite, and this matters. In the next section we will build the smallest

subspace containing a given set of vectors, and one way to describe it is as the intersection of **all** subspaces containing that set.

For two subspaces, the theorem gives familiar pictures: two different planes through the origin in $\mathbb{R}^3$ meet in a line through the origin. It also explains Example (Solutions of a Homogeneous System) from a new angle. The solution set of $\mathbf{Ax} = \mathbf{0}$ is the intersection of the solution sets of its m individual equations, each of which is a subspace of F^n. An algebraic example: in $M_2(F)$, the matrices that are both symmetric and trace-free (Example (Symmetric Matrices), Example (Trace-Free Matrices)) are those of the form $\begin{pmatrix} a & b \\ b & -a \end{pmatrix}$, and they form a subspace.

1.2.7 Unions of subspaces

Intersections behave well. Unions do not, and the reason is visible in the simplest picture.

Example. The Union of the Two Axes

In $\mathbb{R}^2$, let $U = \{(x, 0) : x \in \mathbb{R}\}$ be the x-axis and $W = \{(0, y) : y \in \mathbb{R}\}$ the y-axis. Show that U and W are subspaces of $\mathbb{R}^2$, but $U \cup W$ is not.

Solution. The x-axis is the solution set of the homogeneous equation $y = 0$, that is, of $\mathbf{Ax} = \mathbf{0}$ with $\mathbf{A} = \begin{pmatrix} 0 & 1 \end{pmatrix}$. By Example (Solutions of a Homogeneous System), U is a subspace. In the same way W is the solution set of $x = 0$, so it is a subspace.

For the union, condition (2) fails. We have $(1, 0) \in U \subseteq U \cup W$ and $(0, 1) \in W \subseteq U \cup W$, but

$$(1, 0) + (0, 1) = (1, 1),$$

which is on neither axis, since both of its entries are nonzero. So $(1, 1) \notin U \cup W$, and $U \cup W$ is not a subspace of $\mathbb{R}^2$.

The union passes (1) and (3): it contains $\mathbf{0}$, and scaling a vector on an axis keeps it on that axis. Only addition breaks it, by mixing a vector from each axis. That observation is the whole proof of the general fact.

Proposition 1.2.3. Union of Two Subspaces

Let V be a vector space over F, and let U and W be subspaces of V. Then $U \cup W$ is a subspace of V if and only if $U \subseteq W$ or $W \subseteq U$.

Idea. The direction ($\Leftarrow$) is immediate, because the union is then just the bigger of the two. For ($\Rightarrow$), a statement of the form "one of two inclusions holds" invites contradiction. If neither inclusion holds, there is a vector u of U outside W and a vector w of W outside U, exactly like $(1, 0)$ and $(0, 1)$ for the axes. Look at $u + w$. It must lie in U or in W. If it lies in U, subtracting u keeps us in U and leaves w, which is not in U.

Proof. ($\Leftarrow$) If $U \subseteq W$, then $U \cup W = W$, which is a subspace. If $W \subseteq U$, then $U \cup W = U$, which is a subspace.

($\Rightarrow$) Suppose $U \cup W$ is a subspace of V, and suppose, for a contradiction, that $U \not\subseteq W$ and $W \not\subseteq U$. Then there exist $u \in U$ with $u \notin W$, and $w \in W$ with $w \notin U$. Both lie in $U \cup W$, so $u + w \in U \cup W$ by condition (2) of the subspace test (Theorem 1.2.1).

Case 1: $u + w \in U$. Since U is a subspace and $u \in U$, condition (3) gives $(-1)u \in U$, and $(-1)u = -u$ by Theorem 1.1.7. By condition (2) for U, $(u + w) + (-u) \in U$. By (VS1), (VS2), (VS4), and (VS3),

$$(u + w) + (-u) = w + (u + (-u)) = w + 0 = w.$$

So $w \in U$, contradicting the choice of w.

Case 2: $u + w \in W$. Swapping the roles of U and W, and of u and w, in Case 1 gives $u \in W$, contradicting the choice of u.

Both cases lead to a contradiction. Hence $U \subseteq W$ or $W \subseteq U$.

■

Warning. **Do not combine subspaces by taking their union.** Apart from the case where one subspace already contains the other, the union of two subspaces is never a subspace, by Proposition 1.2.3. The two axes in $\mathbb{R}^2$ are the standard failing case: their union misses $(1, 1)$, and in fact misses every vector with two nonzero entries.

The right way to combine two subspaces U and W is to take all sums $u + w$ with $u \in U$ and $w \in W$. For the two axes this gives all of $\mathbb{R}^2$. This construction, the sum of subspaces, is the subject of a later section of this chapter; Exercise (C2: The Sum of Two Subspaces) gives a preview.

We now have a supply of vector spaces for free: every subspace of a known space is one, and every theorem of the previous section applies to it. In the next section we ask a sharper question. Given a few vectors, what is the smallest subspace that contains them? Its answer, the span, is the first step toward measuring the size of a vector space.

1.2.8 Exercises

1.2.8.1 A. Check your understanding

Exercise. A1

(a) Define what it means for a subset U of a vector space V over F to be a subspace.

(b) State the subspace test.

(c) True or false: $\varnothing$ is a subspace of every vector space. Justify your answer.

(d) True or false: if U is a subspace of V and W is a subspace of U, then W is a subspace of V. Justify your answer.

(e) True or false: the union of two subspaces is never a subspace. Justify your answer.

(f) Which condition of the subspace test would you check first to show that $\{(x, y, z) \in \mathbb{R}^3 : x + y + z = 1\}$ is not a subspace of $\mathbb{R}^3$? Carry out the check.

Solution.

(a) See Definition 1.2.1: U is closed under the addition and scalar multiplication of V, and with these operations it is a vector space over F.

(b) See Theorem 1.2.1: $U \subseteq V$ is a subspace if and only if $0 \in U$, U is closed under addition, and U is closed under scalar multiplication.

(c) False. The empty set does not contain 0, so condition (1) fails (Example (Trivial Subspaces)).

(d) True. Since W is a subspace of U, it contains the zero vector of U, which is 0 (the proof of Theorem 1.2.1 shows the zero vector of a subspace is the zero vector of the big space). The operations of U are those of V, so closure of W under the operations of U is closure under the operations of V. By the subspace test (Theorem 1.2.1), W is a subspace of V.

(e) False. If $U \subseteq W$, then $U \cup W = W$ is a subspace (Proposition 1.2.3). For example, $\{0\} \cup V = V$.

(f) Condition (1). The zero vector has $0 + 0 + 0 = 0 \neq 1$, so it is not in the set, and the set is not a subspace.

1.2.8.2 B. Practice

Exercise. B1: Which Are Subspaces?

Determine which of the following are subspaces of the given vector space over $\mathbb{R}$. Justify your answer.

(a) $\{(x, y, z) \in \mathbb{R}^3 : x - 2y + z = 0\}$ in $\mathbb{R}^3$.

(b) $\{(x, y, z) \in \mathbb{R}^3 : x = y^2\}$ in $\mathbb{R}^3$.

(c) $\{\mathbf{A} \in M_2(\mathbb{R}) : \mathbf{A} \text{ is invertible}\}$ in $M_2(\mathbb{R})$.

(d) $\{\mathbf{A} \in M_2(\mathbb{R}) : a_{12} = 2a_{21}\}$ in $M_2(\mathbb{R})$.

(e) $\{p \in \mathbb{R}[x] : p(2) = 0 \text{ and } p'(2) = 0\}$ in $\mathbb{R}[x]$, where p' is the derivative of p.

(f) The even functions $\{f \in \mathbb{R}^{\mathbb{R}} : f(-t) = f(t) \text{ for all } t \in \mathbb{R}\}$ in $\mathbb{R}^{\mathbb{R}}$.

(g) $\{f \in \mathbb{R}^{\mathbb{R}} : f(0) = 1\}$ in $\mathbb{R}^{\mathbb{R}}$.

(h) The continuous functions $\mathbb{R} \to \mathbb{R}$, in $\mathbb{R}^{\mathbb{R}}$.

Solution.

(a) A subspace. It is $\{\mathbf{x} \in \mathbb{R}^3 : \mathbf{Ax} = \mathbf{0}\}$ with $\mathbf{A} = \begin{pmatrix} 1 & -2 & 1 \end{pmatrix}$, a subspace by Example (Solutions of a Homogeneous System).

(b) Not a subspace: (3) fails. The vector $(1, 1, 0)$ lies in the set, since $1 = 1^2$. But $2(1, 1, 0) = (2, 2, 0)$ does not, since $2 \neq 2^2$.

(c) Not a subspace: (1) fails. The zero matrix is not invertible, because $\mathbf{O} \cdot \mathbf{B} = \mathbf{O} \neq \mathbf{I}_2$ for every $\mathbf{B} \in M_2(\mathbb{R})$.

(d) A subspace. Call the set U.

(1) The zero matrix has $a_{12} = 0 = 2 \cdot 0 = 2a_{21}$. Therefore $\mathbf{O} \in U$.
(2) Let $\mathbf{A}, \mathbf{B} \in U$, so $a_{12} = 2a_{21}$ and $b_{12} = 2b_{21}$. The $(1,2)$-entry of $\mathbf{A} + \mathbf{B}$ is $a_{12} + b_{12} = 2a_{21} + 2b_{21} = 2(a_{21} + b_{21})$, which is twice its $(2,1)$-entry. Therefore $\mathbf{A} + \mathbf{B} \in U$. This shows that U is closed under addition.
(3) Let $c \in \mathbb{R}$ and $\mathbf{A} \in U$. The $(1,2)$-entry of $c\mathbf{A}$ is $ca_{12} = c(2a_{21}) = 2(ca_{21})$, twice its $(2,1)$-entry. Therefore $c\mathbf{A} \in U$. This shows that U is closed under scalar multiplication.
Hence U is a subspace of $M_2(\mathbb{R})$ by the subspace test (Theorem 1.2.1).

(e) A subspace. Call the set U. We use Theorem 0.8.5 for values, and the sum and constant multiple rules for derivatives.
(1) The zero polynomial has value 0 at 2, and its derivative is the zero polynomial, also with value 0 at 2. Therefore $0 \in U$.
(2) Let $p, q \in U$. Then $(p+q)(2) = p(2) + q(2) = 0$ and $(p+q)'(2) = p'(2) + q'(2) = 0$. Therefore $p + q \in U$. This shows that U is closed under addition.
(3) Let $c \in \mathbb{R}$ and $p \in U$. Then $(cp)(2) = c\,p(2) = 0$ and $(cp)'(2) = c\,p'(2) = 0$. Therefore $cp \in U$. This shows that U is closed under scalar multiplication.
Hence U is a subspace of $\mathbb{R}[x]$ by the subspace test (Theorem 1.2.1).

(f) A subspace. Call the set E.
(1) The zero function satisfies $0(-t) = 0 = 0(t)$ for all t. Therefore $0 \in E$.
(2) Let $f, g \in E$. For every $t \in \mathbb{R}$, $(f+g)(-t) = f(-t)+g(-t) = f(t)+g(t) = (f+g)(t)$. Therefore $f + g \in E$. This shows that E is closed under addition.
(3) Let $a \in \mathbb{R}$ and $f \in E$. For every $t \in \mathbb{R}$, $(af)(-t) = af(-t) = af(t) = (af)(t)$. Therefore $af \in E$. This shows that E is closed under scalar multiplication.
Hence E is a subspace of $\mathbb{R}^{\mathbb{R}}$ by the subspace test (Theorem 1.2.1).

(g) Not a subspace: (1) fails. The zero function has value $0 \neq 1$ at 0.

(h) A subspace. **(1)** The zero function is constant, hence continuous. **(2)** A sum of continuous functions is continuous. **(3)** A constant multiple of a continuous function is continuous. Hence the continuous functions form a subspace of $\mathbb{R}^{\mathbb{R}}$ by the subspace test (Theorem 1.2.1).

Exercise. B2: Matrices Commuting with a Given Matrix

Fix $\mathbf{B} \in M_n(F)$, and let $C(\mathbf{B}) = \{\mathbf{A} \in M_n(F) : \mathbf{AB} = \mathbf{BA}\}$.

(a) Prove that $C(\mathbf{B})$ is a subspace of $M_n(F)$.

(b) Let $n = 2$, $F = \mathbb{R}$, and $\mathbf{B} = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$. Find all $\mathbf{A} \in C(\mathbf{B})$. Hence describe $C(\mathbf{B})$ in words.

Solution.

(a) By definition, $C(\mathbf{B}) \subseteq M_n(F)$. We use Theorem 0.9.4 throughout.
(1) By the zero rule, $\mathbf{OB} = \mathbf{O} = \mathbf{BO}$. Therefore $\mathbf{O} \in C(\mathbf{B})$.
(2) Let $\mathbf{A}_1, \mathbf{A}_2 \in C(\mathbf{B})$, so $\mathbf{A}_1\mathbf{B} = \mathbf{BA}_1$ and $\mathbf{A}_2\mathbf{B} = \mathbf{BA}_2$. By distributivity,

$$(\mathbf{A}_1 + \mathbf{A}_2)\mathbf{B} = \mathbf{A}_1\mathbf{B} + \mathbf{A}_2\mathbf{B} = \mathbf{BA}_1 + \mathbf{BA}_2 = \mathbf{B}(\mathbf{A}_1 + \mathbf{A}_2).$$

Therefore $\mathbf{A}_1 + \mathbf{A}_2 \in C(\mathbf{B})$. This shows that $C(\mathbf{B})$ is closed under addition.
(3) Let $c \in F$ and $\mathbf{A} \in C(\mathbf{B})$. By the scalar rule, $(c\mathbf{A})\mathbf{B} = c(\mathbf{AB}) = c(\mathbf{BA}) = \mathbf{B}(c\mathbf{A})$. Therefore $c\mathbf{A} \in C(\mathbf{B})$. This shows that $C(\mathbf{B})$ is closed under scalar multiplication.

Hence $C(\mathbf{B})$ is a subspace of $M_n(F)$ by the subspace test (Theorem 1.2.1).

(b) Let $\mathbf{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$. Then

$$\mathbf{AB} = \begin{pmatrix} a & 2b \\ c & 2d \end{pmatrix}, \qquad \mathbf{BA} = \begin{pmatrix} a & b \\ 2c & 2d \end{pmatrix}.$$

So $\mathbf{AB} = \mathbf{BA}$ if and only if $2b = b$ and $c = 2c$, that is, $b = 0$ and $c = 0$. Hence $C(\mathbf{B})$ is the set of diagonal matrices $\begin{pmatrix} a & 0 \\ 0 & d \end{pmatrix}$ with $a, d \in \mathbb{R}$.

Exercise. B3: Intersection and Union of Two Planes

In $\mathbb{R}^3$, let $U = \{(x, y, z) : x + y + z = 0\}$ and $W = \{(x, y, z) : x - y = 0\}$.

(a) Explain why U and W are subspaces of $\mathbb{R}^3$, and find $U \cap W$.

(b) Find $\mathbf{u} \in U$ with $\mathbf{u} \notin W$ and $\mathbf{w} \in W$ with $\mathbf{w} \notin U$. Hence show directly that $U \cup W$ is not a subspace of $\mathbb{R}^3$.

Solution.

(a) U and W are the solution sets of $\mathbf{Ax} = \mathbf{0}$ for $\mathbf{A} = \begin{pmatrix} 1 & 1 & 1 \end{pmatrix}$ and $\mathbf{A} = \begin{pmatrix} 1 & -1 & 0 \end{pmatrix}$, so they are subspaces by Example (Solutions of a Homogeneous System). A vector (x, y, z) lies in $U \cap W$ exactly when $x + y + z = 0$ and $x = y$. Substituting $y = x$ gives $z = -2x$. Hence

$$U \cap W = \{(t, t, -2t) : t \in \mathbb{R}\},$$

a line through the origin, which is a subspace by Theorem 1.2.2.

(b) Take $\mathbf{u} = (1, -1, 0)$, which lies in U since $1 - 1 + 0 = 0$, but not in W since $1 - (-1) = 2 \neq 0$. Take $\mathbf{w} = (1, 1, 0)$, which lies in W, but not in U since $1 + 1 + 0 = 2 \neq 0$. Both lie in $U \cup W$. Their sum is $(2, 0, 0)$, which is not in U (as $2 + 0 + 0 \neq 0$) and not in W (as $2 - 0 \neq 0$). Hence $U \cup W$ is not closed under addition, and it is not a subspace of $\mathbb{R}^3$.

1.2.8.3 C. Going deeper

Exercise. C1: Covering a Space by Subspaces

(a) Let V be a vector space over any field F. Prove that V is not the union of two proper subspaces.

(b) Show that $\mathbb{F}_2^2$ is the union of three proper subspaces.

Hint: for (a), use Proposition 1.2.3.

Solution.

(a) Suppose, for a contradiction, that $V = U \cup W$ with U and W proper subspaces of V. Then $U \cup W = V$ is a subspace of V, so by Proposition 1.2.3, $U \subseteq W$ or $W \subseteq U$.

In the first case $V = U \cup W = W$, and in the second $V = U$. Either way one of the subspaces equals V, contradicting that both are proper. Hence V is not the union of two proper subspaces.

(b) The space $\mathbb{F}_2^2$ has the four elements $(0,0), (1,0), (0,1), (1,1)$. Let $\mathbf{v}$ be one of the three nonzero elements, and let $U_{\mathbf{v}} = \{\mathbf{0}, \mathbf{v}\}$. We check that $U_{\mathbf{v}}$ is a subspace with the three-check move.

(1) $\mathbf{0} \in U_{\mathbf{v}}$ by construction.

(2) The possible sums are $\mathbf{0}+\mathbf{0} = \mathbf{0}$, $\mathbf{0}+\mathbf{v} = \mathbf{v}+\mathbf{0} = \mathbf{v}$, and $\mathbf{v}+\mathbf{v} = (1+1)\mathbf{v} = 0\mathbf{v} = \mathbf{0}$, using (VS8), (VS6), $1+1=0$ in $\mathbb{F}_2$, and Theorem 1.1.5. All lie in $U_{\mathbf{v}}$, so $U_{\mathbf{v}}$ is closed under addition.

(3) The only scalars are 0 and 1. For $\mathbf{u} \in U_{\mathbf{v}}$, $0\mathbf{u} = \mathbf{0}$ by Theorem 1.1.5 and $1\mathbf{u} = \mathbf{u}$ by (VS8), both in $U_{\mathbf{v}}$. So $U_{\mathbf{v}}$ is closed under scalar multiplication.

Hence each $U_{\mathbf{v}}$ is a subspace. Each has two elements, so it is proper. Every element of $\mathbb{F}_2^2$ is $\mathbf{0}$ or one of the three nonzero vectors, so

$$\mathbb{F}_2^2 = U_{(1,0)} \cup U_{(0,1)} \cup U_{(1,1)}.$$

So three proper subspaces can cover a vector space, although (a) shows two never can.

Exercise. C2: The Sum of Two Subspaces

Let U and W be subspaces of a vector space V over F, and let

$$U + W = \{u + w : u \in U,\ w \in W\}.$$

(a) Prove that $U + W$ is a subspace of V containing both U and W.

(b) Prove that if X is a subspace of V with $U \cup W \subseteq X$, then $U + W \subseteq X$.

(c) For the two axes U and W of Example (The Union of the Two Axes), show that $U + W = \mathbb{R}^2$.

Solution.

(a) Every $u + w$ is a vector of V, so $U + W \subseteq V$.
(1) Since $0 \in U$ and $0 \in W$, and $0 + 0 = 0$ by (VS3), we get $0 \in U + W$.
(2) Let $u_1 + w_1$ and $u_2 + w_2$ be in $U + W$, with $u_1, u_2 \in U$ and $w_1, w_2 \in W$. By (VS1) and (VS2),

$$(u_1 + w_1) + (u_2 + w_2) = (u_1 + u_2) + (w_1 + w_2).$$

Since U and W are closed under addition, $u_1 + u_2 \in U$ and $w_1 + w_2 \in W$. Therefore the sum lies in $U + W$. This shows that $U + W$ is closed under addition.
(3) Let $a \in F$ and $u + w \in U + W$. By (VS5), $a(u + w) = au + aw$, with $au \in U$ and $aw \in W$ by closure. Therefore $a(u + w) \in U + W$. This shows that $U + W$ is closed under scalar multiplication.
Hence $U + W$ is a subspace of V. For $u \in U$, we have $u = u + 0$ with $0 \in W$, so $u \in U + W$; thus $U \subseteq U + W$. Swapping the roles of U and W gives $W \subseteq U + W$, using (VS1).

(b) Let $u + w \in U + W$ with $u \in U$ and $w \in W$. Then $u, w \in U \cup W \subseteq X$. Since X is a

subspace, it is closed under addition, so $u + w \in X$. Hence $U + W \subseteq X$.

(c) By (a), $U + W \subseteq \mathbb{R}^2$. Conversely, let $(x, y) \in \mathbb{R}^2$. Then $(x, y) = (x, 0) + (0, y)$ with $(x, 0) \in U$ and $(0, y) \in W$, so $(x, y) \in U + W$. Hence $U + W = \mathbb{R}^2$, whereas $U \cup W$ is only the two axes.

Exercise. C3: A One-Line Test, and a Test over $\mathbb{F}_2$

Let V be a vector space over F and $U \subseteq V$.

(a) Prove that U is a subspace of V if and only if $U \neq \varnothing$ and $au + w \in U$ for all $a \in F$ and $u, w \in U$.

(b) Suppose $F = \mathbb{F}_2$. Prove that U is a subspace of V if and only if $0 \in U$ and U is closed under addition.

(c) Show by an example that the statement in (b) is false over $\mathbb{R}$.

Solution.

(a) ($\Rightarrow$) Suppose U is a subspace. By the subspace test (Theorem 1.2.1), $0 \in U$, so $U \neq \varnothing$. For $a \in F$ and $u, w \in U$, condition (3) gives $au \in U$, and then condition (2) gives $au + w \in U$.
($\Leftarrow$) Suppose $U \neq \varnothing$ and $au + w \in U$ for all $a \in F$, $u, w \in U$. Pick $u_0 \in U$. Taking $a = -1$ and $u = w = u_0$, and using Theorem 1.1.7 and (VS4),

$$(-1)u_0 + u_0 = -u_0 + u_0 = 0 \in U,$$

where $-u_0 + u_0 = u_0 + (-u_0)$ by (VS1). This is condition (1). For $u, w \in U$, taking $a = 1$ and using (VS8) gives $u + w = 1u + w \in U$, which is (2). For $a \in F$ and $u \in U$, taking $w = 0 \in U$ and using (VS3) gives $au = au + 0 \in U$, which is (3). By the subspace test (Theorem 1.2.1), U is a subspace.

(b) ($\Rightarrow$) This is part of the subspace test (Theorem 1.2.1). ($\Leftarrow$) Suppose $0 \in U$ and U is closed under addition. The only scalars in $\mathbb{F}_2$ are 0 and 1. For $u \in U$, $0u = 0 \in U$ by Theorem 1.1.5, and $1u = u \in U$ by (VS8). So U is closed under scalar multiplication, and by the subspace test (Theorem 1.2.1) U is a subspace.

(c) Take $V = \mathbb{R}^2$ and $U = \mathbb{Z}^2$. Then $\mathbf{0} \in U$, and the sum of two integer vectors is an integer vector. But $\frac{1}{2}(1, 0) = (\frac{1}{2}, 0) \notin U$, so U is not a subspace. Over $\mathbb{R}$ there are many more scalars to test than 0 and 1.

1.3 Linear Combinations and Span

We now have vector spaces and, inside them, subspaces. So far every subspace has been handed to us by a condition, such as "$x + 2y = 0$" or "$\mathbf{A}^\top = \mathbf{A}$". This section goes the other way. We start from a few vectors, build everything that can be built from them, and show that the result is always a subspace: the smallest one containing the vectors we started with. Along the way we meet the first way of measuring a space, by asking whether finitely many vectors are enough to build all of it.

1.3.1 Linear combinations

Take two vectors in $\mathbb{R}^3$, say $\mathbf{v}_1 = (1, 0, 1)$ and $\mathbf{v}_2 = (0, 1, 1)$. Which vectors can we build from them, using only the two operations a vector space gives us? Scaling gives $a\mathbf{v}_1$ and $b\mathbf{v}_2$, and adding gives

$$a\mathbf{v}_1 + b\mathbf{v}_2 = (a, b, a + b).$$

Adding or scaling again produces nothing new: for instance $2(a\mathbf{v}_1 + b\mathbf{v}_2) + (c\mathbf{v}_1 + d\mathbf{v}_2) = (2a + c)\mathbf{v}_1 + (2b + d)\mathbf{v}_2$ has the same shape. So the vectors we can reach are exactly those of the form $(a, b, a + b)$: the plane $z = x + y$ through the origin. The expression $a\mathbf{v}_1 + b\mathbf{v}_2$ will appear in almost every argument from now on, so it deserves a name.

The vectors we combine come as a **list** $(v_1, \ldots, v_k)$: a finite sequence of vectors, in order, in which repeats are allowed. A list of length $k = 0$ is the **empty list** $()$. Why lists rather than sets becomes important in the next section; for now it costs nothing.

Definition 1.3.1. Linear Combination

Let V be a vector space over F, and let $(v_1, \ldots, v_k)$ be a list of vectors in V, where $k \geq 0$. A **linear combination** of $(v_1, \ldots, v_k)$ is a vector of the form

$$a_1 v_1 + a_2 v_2 + \cdots + a_k v_k, \qquad \text{where } a_1, \ldots, a_k \in F.$$

The scalars a_i are the **coefficients** of the combination. A linear combination of the empty list is, by convention, the **empty sum** 0.

In words: scale each vector of the list by a scalar from F, then add up the results. Two small points hide in the notation. First, a sum of k vectors needs no brackets, because addition in V is associative (VS2); so the expression names a single vector. Second, the coefficients must lie in the field F we are working over. Changing the field can change what counts as a combination, as Exercise (C2: The Span Depends on the Field) shows.

The empty sum convention is the additive twin of $0! = 1$ and $x^0 = 1$, where a product with no factors is the neutral element 1 for multiplication: a sum with no terms is the neutral element 0 for addition. It will make several statements true without exceptions.

Example. Linear Combinations in Four Families

Show that each vector is a linear combination of the given list.

(a) $(3, -2, 1) \in \mathbb{R}^3$, of the list $((1, 0, 1), (0, 1, 1))$.

(b) $x^2 - 1 \in \mathbb{R}[x]$, of the list $((x + 1)^2, x + 1)$.

(c) The function $\sin^2$ in $\mathbb{R}^{\mathbb{R}}$, of the list $(1, \cos 2x)$, where 1 is the constant function.

(d) The zero vector of any vector space V, of any list $(v_1, \ldots, v_k)$ in V.

Solution.

(a) We need $a(1, 0, 1) + b(0, 1, 1) = (a, b, a + b)$ to equal $(3, -2, 1)$. The first two entries force $a = 3$ and $b = -2$, and then the third entry is $a + b = 1$, as required. Hence $(3, -2, 1) = 3(1, 0, 1) - 2(0, 1, 1)$.

(b) Expanding, $(x + 1)^2 - 2(x + 1) = x^2 + 2x + 1 - 2x - 2 = x^2 - 1$. So the coefficients are

1 and -2.

(c) The double-angle formula $\cos 2x = 1 - 2\sin^2 x$ holds for every real x. Hence $\sin^2 x = \frac{1}{2} \cdot 1 + (-\frac{1}{2}) \cos 2x$ for every x, which is an equality of functions by Definition 0.4.1. The coefficients $\frac{1}{2}$ and $-\frac{1}{2}$ are real, as they must be.

(d) Take every coefficient to be 0. By Theorem 1.1.5, $0v_i = 0$ for each i, so $0v_1 + \cdots + 0v_k = 0$. For the empty list, 0 is the empty sum. This degenerate case matters: it says the zero vector can always be built, from anything, even from nothing.

Notice that the question in (a) was a small system of equations in the unknown coefficients. That is typical: deciding whether a vector is a combination of a list means deciding whether a system of linear equations has a solution. Here we solve such systems by hand; Chapter 3 gives a systematic method.

1.3.2 The span

The span of some vectors is everything you can build from them by scaling and adding.

We have two kinds of input to cover. Usually we start from a list, as above. But sometimes the natural input is a set, possibly infinite, such as all the powers $1, x, x^2, \ldots$ in $F[x]$. A linear combination only ever uses finitely many vectors, so for a set we allow every finite list drawn from it.

Definition 1.3.2. Span

Let V be a vector space over F.

- The **span** of a list $(v_1, \ldots, v_k)$ of vectors in V is the set of **all** its linear combinations:

$$\operatorname{span}(v_1, \ldots, v_k) = \{a_1 v_1 + \cdots + a_k v_k : a_1, \ldots, a_k \in F\}.$$

- The **span** of a subset $S \subseteq V$, possibly infinite, is the set of all vectors that are linear combinations of **some finite** list of vectors in S:

$$\operatorname{span}(S) = \{a_1 s_1 + \cdots + a_k s_k : k \geq 0,\ s_1, \ldots, s_k \in S,\ a_1, \ldots, a_k \in F\}.$$

In particular $\operatorname{span}() = \operatorname{span}(\varnothing) = \{0\}$. When the field matters we write span_F.

In words: for a list, the span collects every choice of coefficients. For a set, we may first pick any finite number of vectors from S (the same vector more than once is allowed), and then any coefficients. The value $k = 0$ is allowed, and it contributes the empty sum 0. So $0 \in \operatorname{span}(S)$ for **every** S, including $S = \varnothing$.

One thing to check. A list $(v_1, \ldots, v_k)$ now has two spans: the span of the list, and the span of the set $\{v_1, \ldots, v_k\}$ of its entries. They agree. A combination of the list is a combination of a finite list from the set, so the first is contained in the second. Conversely, take a combination $b_1 w_1 + \cdots + b_m w_m$ where each w_r is one of the v_i; for each r, choose one index $i(r)$ with $w_r = v_{i(r)}$. Group the terms with the same index together, using (VS1), (VS2), and $bv + b'v = (b + b')v$ (VS6). The result is $a_1 v_1 + \cdots + a_k v_k$, where a_i is the sum of the b_r with $i(r) = i$ (and $a_i = 0$ if there are none, by Theorem 1.1.5). So the second is contained in the first. In particular the span of a list does not change if we reorder the list or repeat an entry.

Why $\operatorname{span}(\varnothing) = \{0\}$? The definition forces it through the empty sum, but it is also the only

sensible choice. We will prove that the span of S is the smallest subspace containing S. The smallest subspace containing nothing in particular is $\{0\}$, since every subspace contains 0. Any other convention would make that theorem false for $S = \varnothing$, and the empty list will turn out to be a basis of the zero space.

Example. Computing Spans

Describe each span.

(a) $\text{span}((1,0,1),(0,1,1))$ in $\mathbb{R}^3$. Is $(1,2,3)$ in it? Is $(1,2,4)$?
(b) $\text{span}(v)$ for a single vector v in a vector space V.
(c) $\text{span}(1, x^2)$ in $\mathbb{R}[x]$.
(d) $\text{span}(\{1, x, x^2, x^3, \dots\})$ in $F[x]$.

Solution.

(a) As computed at the start of the section, the combinations are $a(1,0,1) + b(0,1,1) = (a, b, a+b)$. So the span is $\{(x,y,z) \in \mathbb{R}^3 : z = x + y\}$, a plane through the origin. For membership, a vector (x,y,z) is in the span exactly when the entries fit: the first two entries force $a = x$ and $b = y$, and then we need $z = x + y$. Since $3 = 1 + 2$, the vector $(1,2,3) = 1(1,0,1) + 2(0,1,1)$ is in the span. Since $4 \neq 1 + 2$, the vector $(1,2,4)$ is not.
(b) $\text{span}(v) = \{av : a \in F\}$. If $v \neq 0$ and $V = \mathbb{R}^2$ or $\mathbb{R}^3$, this is the line through the origin in the direction of v. If $v = 0$, then $a0 = 0$ for every a (Theorem 1.1.6), so $\text{span}(0) = \{0\}$. This degenerate case shows that a span can be as small as the zero space even when the list is not empty.
(c) The combinations are $a + bx^2$ with $a, b \in \mathbb{R}$: the polynomials of degree at most 2 with no x term. For example $x^2 + x$ is **not** in the span, because its coefficient of x is $1 \neq 0$ and polynomials are equal only when all coefficients agree (Definition 0.8.1).
(d) Every polynomial $a_0 + a_1 x + \cdots + a_N x^N$ is a combination of the finite list $(1, x, \dots, x^N)$ of vectors in the set. Conversely every combination of finitely many powers of x is a polynomial. So the span is all of $F[x]$. No single polynomial here is a combination of *infinitely* many powers: each one uses only the powers up to its degree.

Quick check. Is $\text{span}((1,2))$ a subset of $\text{span}((2,4))$ in $\mathbb{R}^2$? Are they equal?

Solution. Yes to both. Since $(2,4) = 2(1,2)$, every $a(2,4) = (2a)(1,2)$ lies in $\text{span}((1,2))$. Since $(1,2) = \frac{1}{2}(2,4)$, every $a(1,2) = \frac{a}{2}(2,4)$ lies in $\text{span}((2,4))$. Both spans are the line $y = 2x$. Different lists can have the same span.

1.3.3 The span is the smallest subspace

The plane in Example (Computing Spans) (a) was a subspace. That was no accident. Adding two combinations gives a combination, and so does scaling one. And any subspace that contains the vectors must contain all their combinations, because a subspace is closed under exactly the operations that build them. So the span is both a subspace and the smallest one available:

Theorem 1.3.1. Span Is the Smallest Subspace

Let V be a vector space over F and let $S \subseteq V$. Then:

(1) $\operatorname{span}(S)$ is a subspace of V;
(2) $S \subseteq \operatorname{span}(S)$;
(3) if U is a subspace of V with $S \subseteq U$, then $\operatorname{span}(S) \subseteq U$.

The same holds for the span of a list, which equals the span of the set of its entries.

Idea. Part (1) is the three-check subspace test. Part (2) is the combination $1s$. Part (3) is "a subspace is closed under the operations that build combinations"; since a combination has k terms for an arbitrary k, we add one term at a time, by induction on k.

Proof.

(1) We apply Theorem 1.2.1.
- *Zero.* The empty list is a finite list of vectors in S, and its only combination is 0. Hence $0 \in \operatorname{span}(S)$.
- *Addition.* Let $x, y \in \operatorname{span}(S)$. Then $x = a_1 s_1 + \cdots + a_k s_k$ and $y = b_1 t_1 + \cdots + b_m t_m$ for some $s_i, t_j \in S$ and scalars $a_i, b_j \in F$. Therefore

$$x + y = a_1 s_1 + \cdots + a_k s_k + b_1 t_1 + \cdots + b_m t_m,$$

which is a combination of the list $(s_1, \ldots, s_k, t_1, \ldots, t_m)$ of vectors in S. (If some t_j equals some s_i, the list has a repeat, which lists allow.) Hence $x + y \in \operatorname{span}(S)$. This shows $\operatorname{span}(S)$ is closed under addition.
- *Scaling.* Let $x = a_1 s_1 + \cdots + a_k s_k \in \operatorname{span}(S)$ and $c \in F$. By (VS5), applied $k - 1$ times, and by (VS7) for each term,

$$cx = (ca_1)s_1 + \cdots + (ca_k)s_k,$$

which is a combination of the same list. Hence $cx \in \operatorname{span}(S)$. This shows $\operatorname{span}(S)$ is closed under scalar multiplication.

Therefore $\operatorname{span}(S)$ is a subspace of V.

(2) Let $s \in S$. By (VS8), $s = 1s$, which is a combination of the list (s). Hence $s \in \operatorname{span}(S)$.

(3) Let U be a subspace with $S \subseteq U$. By Theorem 1.2.1, U contains 0 and is closed under addition and scalar multiplication. We prove by induction on $k \geq 0$ (Theorem 0.2.7) that every combination $a_1 s_1 + \cdots + a_k s_k$ with $s_i \in S$ lies in U. For $k = 0$ the combination is $0 \in U$. Suppose the claim holds for k, and consider $a_1 s_1 + \cdots + a_{k+1} s_{k+1}$. By the induction hypothesis $x = a_1 s_1 + \cdots + a_k s_k \in U$. Since $s_{k+1} \in S \subseteq U$ and U is closed under scaling, $a_{k+1} s_{k+1} \in U$. Since U is closed under addition, $x + a_{k+1} s_{k+1} \in U$. This completes the induction, and shows $\operatorname{span}(S) \subseteq U$.

■

In one phrase: $\operatorname{span}(S)$ is a subspace containing S, and it sits inside every other one. So "the smallest subspace containing S" makes sense, and it is $\operatorname{span}(S)$. This gives a second, top-down description of the span, which does not mention combinations at all.

Corollary 1.3.2. Span as an Intersection

Let V be a vector space and $S \subseteq V$. Let $\mathcal{U}$ be the set of all subspaces U of V with $S \subseteq U$. Then

$$\operatorname{span}(S) = \bigcap_{U \in \mathcal{U}} U.$$

Proof. The index set $\mathcal{U}$ is nonempty, because V is a subspace of itself containing S; so the intersection is defined (Definition 0.3.9).

($\subseteq$) By Theorem 1.3.1 (3), $\operatorname{span}(S) \subseteq U$ for every $U \in \mathcal{U}$. Hence $\operatorname{span}(S)$ is contained in the intersection.

($\supseteq$) By Theorem 1.3.1 (1) and (2), $\operatorname{span}(S)$ is itself a member of $\mathcal{U}$. An intersection is contained in each of its members, so the intersection is contained in $\operatorname{span}(S)$.

By Theorem 0.3.1, the two sets are equal.

■

By Theorem 1.2.2, an intersection of subspaces is a subspace, so the right-hand side is a subspace for a reason that never mentions combinations. The two descriptions complement each other. The bottom-up one (combinations) tells us what the elements look like. The top-down one (intersection) is convenient for proving that the span lies inside something.

Warning. The span of S is **not** S, and it is a subspace **even when** S **is not**. For example $S = \{(1,0),(0,1)\}$ is a two-element set, and not a subspace of $\mathbb{R}^2$ (it does not contain $(0,0)$). Its span is the whole plane $\mathbb{R}^2$, since $(x,y) = x(1,0)+y(0,1)$. By Theorem 1.3.1, $\operatorname{span}(S) = S$ holds exactly when S is already a subspace. Also keep the two uses of the word apart: "the span of S" is a noun, a subspace; "S spans V" is a verb, defined next.

1.3.4 Spanning lists and finite-dimensional spaces

The most important case is when the span is everything. Then the list is a complete toolkit for the space: every vector can be built from it.

Definition 1.3.3. Spanning List, Spanning Set

Let V be a vector space over F. A list $(v_1,\dots,v_k)$ in V **spans** V, or is a **spanning list** of V, if $\operatorname{span}(v_1,\dots,v_k) = V$. A subset $S \subseteq V$ **spans** V, or is a **spanning set** of V, if $\operatorname{span}(S) = V$.

Since every span is contained in V, to show that a list spans V we must show one thing: **every** $v \in V$ can be written as $a_1v_1 + \cdots + a_kv_k$ for **some** scalars. The scalars may depend on v, and they need not be unique.

Example. Spanning Lists

Show that each list spans the given space.

(a) $(\mathbf{e}_1, \ldots, \mathbf{e}_n)$ in F^n, where $\mathbf{e}_i$ has 1 in position i and 0 elsewhere.
(b) $(1, x, x^2)$ in $F[x]_{\leq 2}$.
(c) $((1,1),(1,-1))$ in $\mathbb{R}^2$.
(d) $(\mathbf{E}_{11}, \mathbf{E}_{12}, \mathbf{E}_{21}, \mathbf{E}_{22})$ in $M_2(F)$, where $\mathbf{E}_{ij}$ has 1 in entry (i,j) and 0 elsewhere.
(e) The empty list, in the zero space $\{0\}$.

Solution.

(a) Let $(x_1, \ldots, x_n) \in F^n$. Adding entry by entry, $x_1\mathbf{e}_1 + \cdots + x_n\mathbf{e}_n = (x_1, \ldots, x_n)$. So every vector is a combination, with its own entries as coefficients.
(b) Every $p \in F[x]_{\leq 2}$ has the form $a_0 + a_1x + a_2x^2$ (Definition 0.8.4), which is the combination $a_0 \cdot 1 + a_1 \cdot x + a_2 \cdot x^2$.
(c) Let $(p,q) \in \mathbb{R}^2$. We look for a, b with $a(1,1) + b(1,-1) = (p,q)$, that is,

$$a + b = p, \qquad a - b = q.$$

Adding the equations gives $2a = p+q$, and subtracting gives $2b = p-q$. So $a = \frac{p+q}{2}$ and $b = \frac{p-q}{2}$. Check: $\frac{p+q}{2} + \frac{p-q}{2} = p$ and $\frac{p+q}{2} - \frac{p-q}{2} = q$. Hence $(p,q) = \frac{p+q}{2}(1,1) + \frac{p-q}{2}(1,-1)$, and the list spans $\mathbb{R}^2$. For instance $(3,5) = 4(1,1) - (1,-1)$.
(d) For any matrix, $\begin{pmatrix} a & b \\ c & d \end{pmatrix} = a\mathbf{E}_{11} + b\mathbf{E}_{12} + c\mathbf{E}_{21} + d\mathbf{E}_{22}$, by adding entry by entry.
(e) span() $= \{0\}$ by Definition 1.3.2. So the empty list spans the zero space. (So does the list (0).)

The solution of (3) divided by 2, which is fine in $\mathbb{R}$. Over $\mathbb{F}_2$, where $2 = 0$, the same list fails: there $1 = -1$, so both vectors equal $(1,1)$, and their span is $\{(0,0),(1,1)\} \neq \mathbb{F}_2^2$. Whether a list spans depends on the field.

Now the nonexample, by a minimal change of (3). Replace the list by $((1,2),(2,4))$. It still has two vectors, and neither is zero. But every combination is

$$a(1,2) + b(2,4) = (a + 2b)(1,2),$$

a vector whose second entry is twice its first. So $(0,1)$ is **not** reached, and the list does **not** span $\mathbb{R}^2$. The clause that fails is "**every** $v \in V$ is a combination". The two vectors point along the same line, so their span is only the line $y = 2x$.

Some spaces are spanned by a short list; others need infinitely many vectors. That distinction is the first coarse measurement of size in this book.

Definition 1.3.4. Finite-Dimensional

A vector space V over F is **finite-dimensional** if **some** finite list of vectors in V spans V. Otherwise V is **infinite-dimensional**.

By Example (Spanning Lists), F^n, $F[x]_{\leq 2}$, $M_2(F)$, and $\{0\}$ are finite-dimensional. The same argument as in (2) and (4) shows that $F[x]_{\leq n}$ and $M_{m\times n}(F)$ are finite-dimensional too. The word "dimension" is used before we have defined a number called dimension; that number, and

the fact that it is well defined, are the subject of section 6 of this chapter.

For the whole polynomial space no finite list is enough. The reason is degree: finitely many polynomials have a largest degree, and combining them never climbs above it.

Theorem 1.3.3. $F[x]$ **Is Infinite-Dimensional**

For every field F, the vector space $F[x]$ is infinite-dimensional.

Proof. Let $(p_1, \dots, p_k)$ be any finite list in $F[x]$. We show it does not span $F[x]$. Each $\deg p_i$ is either a natural number or $-\infty$, and there are finitely many, so we can choose $N \in \mathbb{N}$ with $\deg p_i \le N$ for every i.

Let $a_1, \dots, a_k \in F$. The coefficients of $a_i p_i$ are a_i times those of p_i (Example (The Standard Vector Spaces) (c)). Every coefficient of x^m with $m > N$ in p_i is 0, by Definition 0.8.3, and $a_i \cdot 0 = 0$ by (F6) and Theorem 0.7.1 (d); hence $\deg(a_i p_i) \le N$. By Theorem 0.8.2, $\deg(p+q) \le \max(\deg p, \deg q)$, so a sum of two polynomials of degree at most N has degree at most N. Applying this $k-1$ times, $\deg(a_1 p_1 + \cdots + a_k p_k) \le N$. (For $k = 0$ the combination is 0, of degree $-\infty \le N$.)

Hence every element of $\operatorname{span}(p_1, \dots, p_k)$ has degree at most N. But $x^{N+1} \in F[x]$ has degree $N+1 > N$, so $x^{N+1} \notin \operatorname{span}(p_1, \dots, p_k)$. Since the list was arbitrary, no finite list spans $F[x]$. This shows $F[x]$ is infinite-dimensional. ∎

(In the proof we wrote polynomials in plain italic, p, as in Chapter 0; they are vectors of $F[x]$ all the same.) The proof is the move "find a quantity that combinations cannot increase". The space $\mathbb{R}^{\mathbb{R}}$ of all functions is infinite-dimensional too, and so are the sequence spaces; we return to such spaces in section 8 of this chapter.

Remark. For an arbitrary subset $S \subseteq V$, finite or infinite, "S spans V" means that **every** vector of V is a combination of **some finite** list of vectors in S. Different vectors may use different finite lists, as in Example (Computing Spans) (d). We never form infinite sums: a vector space only lets us add two vectors at a time, so only finite sums have a meaning. The same "every finite list" reading will be used for independence of an arbitrary set in the next section.

1.3.5 Working with spans

Three facts make spans easy to manipulate. Each follows from Theorem 1.3.1 without writing a single combination.

Proposition 1.3.4. **Basic Properties of Span**

Let V be a vector space and $S, T \subseteq V$.

(1) If $T \subseteq \operatorname{span}(S)$, then $\operatorname{span}(T) \subseteq \operatorname{span}(S)$.

(2) (Monotonicity) If $T \subseteq S$, then $\operatorname{span}(T) \subseteq \operatorname{span}(S)$.

(3) $\operatorname{span}(\operatorname{span}(S)) = \operatorname{span}(S)$.

Proof.

(1) By Theorem 1.3.1 (1), $\operatorname{span}(S)$ is a subspace, and by hypothesis it contains T. By Theorem 1.3.1 (3) applied to T, $\operatorname{span}(T) \subseteq \operatorname{span}(S)$.

(2) By Theorem 1.3.1 (2), $T \subseteq S \subseteq \operatorname{span}(S)$. Now apply (1).

(3) Taking $T = \operatorname{span}(S)$ in (1) gives $\operatorname{span}(\operatorname{span}(S)) \subseteq \operatorname{span}(S)$. The reverse inclusion is Theorem 1.3.1 (2) applied to the set $\operatorname{span}(S)$.

■

Part (1) is the workhorse. To show $\operatorname{span}(T) \subseteq \operatorname{span}(S)$, it is enough to check that each vector **of** T is a combination of vectors of S; we do not have to handle every combination of T.

The next fact is the one later sections lean on most. If a vector can already be built from S, then adding it to S builds nothing new.

Theorem 1.3.5. Absorbing a Vector into a Span

Let V be a vector space, $S \subseteq V$ and $v \in V$. If $v \in \operatorname{span}(S)$, then

$$\operatorname{span}(S \cup \{v\}) = \operatorname{span}(S).$$

In particular, for a list: if $v \in \operatorname{span}(v_1, \ldots, v_k)$, then $\operatorname{span}(v_1, \ldots, v_k, v) = \operatorname{span}(v_1, \ldots, v_k)$, and the same holds with v inserted at any position of the list.

Proof. ($\supseteq$) Since $S \subseteq S \cup \{v\}$, monotonicity (Proposition 1.3.4 (2)) gives $\operatorname{span}(S) \subseteq \operatorname{span}(S \cup \{v\})$.

($\subseteq$) By Theorem 1.3.1 (2), $S \subseteq \operatorname{span}(S)$, and $v \in \operatorname{span}(S)$ by hypothesis. Hence $S \cup \{v\} \subseteq \operatorname{span}(S)$, and Proposition 1.3.4 (1) gives $\operatorname{span}(S \cup \{v\}) \subseteq \operatorname{span}(S)$.

For the list version, the span of a list equals the span of the set of its entries, whatever the order. The entries of $(v_1, \ldots, v_k, v)$ form the set $\{v_1, \ldots, v_k\} \cup \{v\}$, so the first part applies.

■

Concretely, the proof substitutes. If $v = c_1 v_1 + \cdots + c_k v_k$, then any combination $a_1 v_1 + \cdots + a_k v_k + bv$ equals $(a_1 + bc_1)v_1 + \cdots + (a_k + bc_k)v_k$, which no longer uses v. Read backwards, the lemma says that a vector lying in the span of the others can be **deleted** from a list without shrinking its span. The next section turns this into the idea of a redundant vector, and section 6 of this chapter uses it at every step of the Exchange Theorem.

Example. Two Lists with the Same Span

Show that $\operatorname{span}(1 + x, 1 - x) = \operatorname{span}(1, x)$ in $\mathbb{R}[x]$.

Solution. ($\subseteq$) Both $1+x = 1\cdot 1 + 1\cdot x$ and $1-x = 1\cdot 1 + (-1)\cdot x$ lie in $\operatorname{span}(1,x)$. By Proposition 1.3.4 (1), $\operatorname{span}(1+x, 1-x) \subseteq \operatorname{span}(1,x)$.

($\supseteq$) Conversely,

$$1 = \tfrac{1}{2}(1+x) + \tfrac{1}{2}(1-x), \qquad x = \tfrac{1}{2}(1+x) - \tfrac{1}{2}(1-x),$$

so both 1 and x lie in $\operatorname{span}(1+x, 1-x)$. By Proposition 1.3.4 (1) again, $\operatorname{span}(1,x) \subseteq \operatorname{span}(1+x, 1-x)$.

Hence the spans are equal; both are $\mathbb{R}[x]_{\leq 1}$. Notice that we only checked four vectors, not all combinations. We also used $\frac{1}{2}$, so the argument needs $2 \neq 0$ in the field.

Quick check. Is $x^2 + 1 \in \operatorname{span}(x^2 + x, x + 1)$ in $\mathbb{R}[x]$?

Solution. No. Suppose $a(x^2+x) + b(x+1) = x^2 + 1$. The left side is $ax^2 + (a+b)x + b$. Comparing coefficients (Definition 0.8.1): $a = 1$ from x^2, $b = 1$ from the constant term, and $a + b = 0$ from x. But then $a + b = 2 \neq 0$, a contradiction. So no such a, b exist.

1.3.6 Exercises

1.3.6.1 A. Check your understanding

Exercise. A1

(a) Define the span of a subset S of a vector space V. What is $\operatorname{span}(\varnothing)$, and why?

(b) Determine whether the following statement is true: "for a subset $S \subseteq V$, $\operatorname{span}(S) = S$ if and only if S is a subspace of V." Justify your answer.

(c) Determine whether the following statement is true: "if $\operatorname{span}(S) = \operatorname{span}(T)$, then $S = T$." Justify your answer.

(d) Is the zero vector in the span of every list? Is it in the span of the empty list?

(e) State what it means for V to be finite-dimensional, and name the quantity used to show that $F[x]$ is not.

Solution.

(a) $\operatorname{span}(S)$ is the set of all linear combinations $a_1 s_1 + \cdots + a_k s_k$ with $k \geq 0$, $s_i \in S$, and $a_i \in F$. $\operatorname{span}(\varnothing) = \{0\}$: the only combination with no vectors is the empty sum 0. This is also the smallest subspace containing $\varnothing$, as Theorem 1.3.1 requires.

(b) True. If $\operatorname{span}(S) = S$, then S is a subspace by Theorem 1.3.1 (1). Conversely, if S is a subspace, then $\operatorname{span}(S) \subseteq S$ by Theorem 1.3.1 (3) with $U = S$, and $S \subseteq \operatorname{span}(S)$ by (2).

(c) False. In $\mathbb{R}^2$, take $S = \{(1,0), (0,1)\}$ and $T = \{(1,1), (1,-1)\}$. Both span $\mathbb{R}^2$ (Example (Spanning Lists)), but $S \neq T$.

(d) Yes to both: take all coefficients 0, or, for the empty list, the empty sum. See Example (Linear Combinations in Four Families) (d).

(e) V is finite-dimensional if some finite list of vectors in V spans V. For $F[x]$ the quantity is the degree: combinations of a finite list have degree at most the largest degree in the list (Theorem 1.3.3).

1.3.6.2 B. Practice

Exercise. B1: Membership in a Span

Determine whether each vector lies in the given span. If it does, write it as a linear combination. Justify your answer.

(a) $(2,-1,5)$ and $(3,4,1)$, in $\operatorname{span}((1,1,0),(0,1,1)) \subseteq \mathbb{R}^3$.

(b) $2x^2-3x+1$ and x^2+x+1, in $\operatorname{span}(x^2-x, x-1) \subseteq \mathbb{R}[x]$.

Solution.

(a) A combination is $a(1,1,0)+b(0,1,1)=(a,a+b,b)$. For $(2,-1,5)$, the first and third entries force $a=2$ and $b=5$; then the middle entry would be $7 \neq -1$. Hence $(2,-1,5)$ is **not** in the span. For $(3,4,1)$, we get $a=3$, $b=1$, and the middle entry is 4, as required. Hence $(3,4,1)=3(1,1,0)+(0,1,1)$ is in the span.

(b) A combination is $a(x^2-x)+b(x-1)=ax^2+(b-a)x-b$. Compare coefficients (Definition 0.8.1). For $2x^2-3x+1$: $a=2$, $-b=1$ so $b=-1$, and then $b-a=-3$, as required. Hence $2x^2-3x+1=2(x^2-x)-(x-1)$ is in the span. For x^2+x+1: $a=1$, $b=-1$, and then $b-a=-2 \neq 1$. Hence x^2+x+1 is **not** in the span. (Every element p of this span has $p(1)=a-a+b-b=0$, and indeed x^2+x+1 has value 3 at 1.)

Exercise. B2: Two Lists with the Same Span

Prove that $\operatorname{span}((1,1,0),(0,1,1)) = \operatorname{span}((1,2,1),(1,0,-1))$ in $\mathbb{R}^3$.

Solution. Write $\mathbf{u}=(1,1,0)$, $\mathbf{w}=(0,1,1)$.

($\subseteq$) We have $\mathbf{u}=\frac{1}{2}(1,2,1)+\frac{1}{2}(1,0,-1)$ and $\mathbf{w}=\frac{1}{2}(1,2,1)-\frac{1}{2}(1,0,-1)$: indeed $\frac{1}{2}(2,2,0)=(1,1,0)$ and $\frac{1}{2}(0,2,2)=(0,1,1)$. So $\mathbf{u},\mathbf{w} \in \operatorname{span}((1,2,1),(1,0,-1))$, and Proposition 1.3.4 (1) gives the inclusion.

($\supseteq$) We have $(1,2,1)=\mathbf{u}+\mathbf{w}$ and $(1,0,-1)=\mathbf{u}-\mathbf{w}$. So both lie in $\operatorname{span}(\mathbf{u},\mathbf{w})$, and Proposition 1.3.4 (1) gives the reverse inclusion.

Hence the two spans are equal.

Exercise. B3: Spans in $\mathbb{R}^3$, Geometrically

Describe each span as the origin, a line through the origin, a plane through the origin (give

an equation), or all of $\mathbb{R}^3$. Justify your answer.

(a) $\text{span}((1,2,3))$.

(b) $\text{span}((1,0,0),(0,1,0),(1,1,0))$.

(c) $\text{span}((1,-1,0),(0,1,-1))$.

(d) $\text{span}((0,0,0))$.

Solution.

(a) The span is $\{(t,2t,3t) : t \in \mathbb{R}\}$: the line through the origin and $(1,2,3)$.

(b) Since $(1,1,0) = (1,0,0)+(0,1,0)$, Theorem 1.3.5 gives $\text{span}((1,0,0),(0,1,0),(1,1,0)) = \text{span}((1,0,0),(0,1,0)) = \{(a,b,0) : a,b \in \mathbb{R}\}$: the plane $z = 0$.

(c) A combination is $a(1,-1,0) + b(0,1,-1) = (a, b-a, -b)$, whose entries sum to 0. Conversely, if $x+y+z = 0$, take $a = x$ and $b = -z$; then $b - a = -z - x = y$, so $(x,y,z) = x(1,-1,0) - z(0,1,-1)$. Hence the span is the plane $x+y+z = 0$.

(d) $\text{span}((0,0,0)) = \{\mathbf{0}\}$, by Example (Computing Spans) (b): the origin alone.

1.3.6.3 C. Going deeper

Exercise. C1: Spans of Intersections and Unions

Let S, T be subsets of a vector space V.

(a) Prove that $\text{span}(S \cap T) \subseteq \text{span}(S) \cap \text{span}(T)$.

(b) Give an example in $\mathbb{R}^2$ where the inclusion in (a) is strict.

(c) Prove that $\text{span}(S \cup T) = \{s + t : s \in \text{span}(S),\ t \in \text{span}(T)\}$.

Hint for (c): in a combination of vectors of $S \cup T$, sort the terms into two groups.

Solution.

(a) Since $S \cap T \subseteq S$, monotonicity (Proposition 1.3.4 (2)) gives $\text{span}(S \cap T) \subseteq \text{span}(S)$. By symmetry, $\text{span}(S \cap T) \subseteq \text{span}(T)$. Hence $\text{span}(S \cap T) \subseteq \text{span}(S) \cap \text{span}(T)$.

(b) Take $S = \{(1,0),(0,1)\}$ and $T = \{(1,1)\}$. Then $S \cap T = \varnothing$, so $\text{span}(S \cap T) = \{\mathbf{0}\}$. But $\text{span}(S) = \mathbb{R}^2$ (Example (Spanning Lists)), so $\text{span}(S) \cap \text{span}(T) = \text{span}(T)$, the line $y = x$, which contains $(1,1) \neq \mathbf{0}$. The inclusion is strict.

(c) Write $W = \{s + t : s \in \text{span}(S),\ t \in \text{span}(T)\}$.
($\subseteq$) Let $v \in \text{span}(S \cup T)$, say $v = a_1u_1 + \cdots + a_ku_k$ with each $u_i \in S \cup T$. Let s be the sum of the terms a_iu_i with $u_i \in S$, and t the sum of the remaining terms, whose vectors lie in T. Since addition is commutative and associative, $v = s + t$. Now $s \in \text{span}(S)$ and $t \in \text{span}(T)$; if a group is empty, its sum is 0, which lies in every span. Hence $v \in W$.
($\supseteq$) Let $s \in \text{span}(S)$ and $t \in \text{span}(T)$. By monotonicity, both lie in $\text{span}(S \cup T)$, which is closed under addition by Theorem 1.3.1 (1). Hence $s + t \in \text{span}(S \cup T)$.
By Theorem 0.3.1, $\text{span}(S \cup T) = W$. (So W is the set $\text{span}(S) + \text{span}(T)$ of Exercise (C2: The Sum of Two Subspaces), the **sum** of the two subspaces, studied in section 7 of this chapter.)

Exercise. C2: The Span Depends on the Field

The set $\mathbb{C}^2$ is a vector space over $\mathbb{C}$. With the same addition, and scalar multiplication restricted to real scalars, it is also a vector space over $\mathbb{R}$, by Exercise (C3: Shrinking the Field): every axiom holds for all complex scalars, so in particular for real ones.

(a) Show that $((1,0),(0,1))$ spans $\mathbb{C}^2$ over $\mathbb{C}$.

(b) Explain why $((1,0),(0,1))$ does **not** span $\mathbb{C}^2$ over $\mathbb{R}$.

(c) Find a finite list that spans $\mathbb{C}^2$ over $\mathbb{R}$. Hence $\mathbb{C}^2$ is finite-dimensional over $\mathbb{R}$ as well.

Solution.

(a) For $(z,w) \in \mathbb{C}^2$, $(z,w) = z(1,0) + w(0,1)$ with $z, w \in \mathbb{C}$.

(b) Over $\mathbb{R}$ the coefficients must be real, and $a(1,0)+b(0,1) = (a,b)$ with $a, b \in \mathbb{R}$ has real entries. So $(i,0)$ is not in $\operatorname{span}_{\mathbb{R}}((1,0),(0,1))$.

(c) Take $((1,0),(i,0),(0,1),(0,i))$. Write $z = a + bi$ and $w = c + di$ with $a, b, c, d \in \mathbb{R}$ (Definition 0.6.1). Then

$$(z,w) = a(1,0) + b(i,0) + c(0,1) + d(0,i),$$

a combination with real coefficients. So this list spans $\mathbb{C}^2$ over $\mathbb{R}$, and $\mathbb{C}^2$ is finite-dimensional over $\mathbb{R}$ by Definition 1.3.4.

1.4 Linear Independence

A spanning list tells us that every vector of a space can be built. It says nothing about waste. The list $((1,0),(0,1),(1,1))$ spans $\mathbb{R}^2$, but its third vector adds nothing that the first two did not already reach. To measure a space by the length of a list, we need lists with no such waste. This section makes "no vector is wasted" precise, proves the lemma that finds a wasted vector whenever there is one, and shows that lists without waste give every vector in their span a unique description.

1.4.1 Redundant vectors

The previous section gave a way to shorten a list: if a vector lies in the span of the others, we may delete it without changing the span (Theorem 1.3.5). Call such a vector **redundant**. We want a test for "this list has no redundant vector".

A first guess is that a vector is wasted when it repeats the direction of another one, so the test would be "no vector is a scalar multiple of another". This guess fails. In $\mathbb{R}^3$, take

$$\mathbf{u} = (1,0,1), \qquad \mathbf{v} = (0,1,1), \qquad \mathbf{w} = (1,1,2).$$

No one of them is a multiple of another: a multiple of $\mathbf{u}$ has second entry 0, a multiple of $\mathbf{v}$ has first entry 0, a multiple $(c,c,2c)$ of $\mathbf{w}$ has equal first and second entries, and in each case neither of the other two vectors has that shape. Yet $\mathbf{w} = \mathbf{u} + \mathbf{v}$. All three lie in the plane $z = x + y$, and by Theorem 1.3.5, $\operatorname{span}(\mathbf{u},\mathbf{v},\mathbf{w}) = \operatorname{span}(\mathbf{u},\mathbf{v})$. The vector $\mathbf{w}$ is redundant even though it is not parallel to anything.

A second guess is correct but clumsy: "no vector is a linear combination of the others". For a

list of length k this is k separate conditions, one for each vector. Here is a way to fold them into one. Move everything in $\mathbf{w} = \mathbf{u} + \mathbf{v}$ to one side:

$$1\mathbf{u} + 1\mathbf{v} + (-1)\mathbf{w} = \mathbf{0}.$$

This is a way to build $\mathbf{0}$ from the list with coefficients that are **not all zero**. Conversely, suppose $a\mathbf{u} + b\mathbf{v} + c\mathbf{w} = \mathbf{0}$ with, say, $c \neq 0$. Then $\mathbf{w} = -\frac{a}{c}\mathbf{u} - \frac{b}{c}\mathbf{v}$, so $\mathbf{w}$ is redundant. The same works whichever coefficient is nonzero. So a redundant vector exists exactly when $\mathbf{0}$ can be built in a nontrivial way. Building $\mathbf{0}$ trivially, with every coefficient 0, is always possible, so the condition we want is that this is the only way.

A list is independent when the only way to build the zero vector from it is the trivial way.

1.4.2 Definition and examples

Definition 1.4.1. Linear Independence

Let V be a vector space over F and let $(v_1, \ldots, v_k)$ be a list of vectors in V, where $k \geq 0$. The list is **linearly independent** (over F) if, **for all** $a_1, \ldots, a_k \in F$,

$$a_1v_1 + a_2v_2 + \cdots + a_kv_k = 0 \quad \Longrightarrow \quad a_1 = a_2 = \cdots = a_k = 0.$$

The list is **linearly dependent** if it is not linearly independent: that is, if there exist $a_1, \ldots, a_k \in F$, **not all zero**, with $a_1v_1 + \cdots + a_kv_k = 0$. Such an equation is a **linear relation** among the vectors. The empty list is linearly independent.

In words: the hypothesis is an arbitrary linear combination that happens to equal 0. The conclusion says that **every** coefficient in it is 0. Dependence is the negation, obtained by the rules of Definition 0.1.4: "for all a, (combination $= 0 \Rightarrow$ all $a_i = 0$)" becomes "there exist a with combination $= 0$ and **not** all $a_i = 0$". This is exactly the negation worked out in Example (Negating "only the zero combination"). The empty list is independent because the condition is vacuous: there are no coefficients, so "all of them are 0" holds automatically.

The definition also dictates how every independence proof starts. It is a "for all" statement about coefficients, so we take arbitrary coefficients and assume the hypothesis:

Let $a_1v_1 + \cdots + a_kv_k = 0$**.** Then ... hence all $a_i = 0$.

Between "Let" and "hence", the work is to turn one vector equation into scalar equations. How we do that depends on what 0 means in the space: in F^n we compare entries, in $F[x]$ we compare coefficients, in a function space we evaluate at points, and in $M_{m\times n}(F)$ we compare entries.

Example. Independent Lists in Four Families

Show that each list is linearly independent.

(a) $(\mathbf{e}_1, \ldots, \mathbf{e}_n)$ in F^n.
(b) $(1, x, x^2)$ in $F[x]$.
(c) $((1,1),(1,-1))$ in $\mathbb{R}^2$.
(d) $(\sin, \cos)$ in the space $\mathbb{R}^{\mathbb{R}}$ of all functions $\mathbb{R} \to \mathbb{R}$.

Solution.

(a) Let $a_1\mathbf{e}_1 + \cdots + a_n\mathbf{e}_n = \mathbf{0}$. The left side is the vector $(a_1, \ldots, a_n)$, and $\mathbf{0} = (0, \ldots, 0)$. Comparing entries, $a_1 = \cdots = a_n = 0$. Hence the list is independent.

(b) Let $a + bx + cx^2 = 0$, where 0 is the **zero polynomial**, the polynomial with every coefficient 0. Two polynomials are equal exactly when all their coefficients agree (Definition 0.8.1). Comparing the coefficients of $1, x, x^2$ gives $a = b = c = 0$. Hence the list is independent.

(c) Let $a(1,1) + b(1,-1) = (0,0)$. Comparing entries, $a + b = 0$ and $a - b = 0$. Adding the two equations gives $2a = 0$, so $a = 0$ since $2 \neq 0$ in $\mathbb{R}$; then $b = -a = 0$. Hence the list is independent.

(d) Let $a\sin + b\cos = 0$, where 0 is the **zero function**. Equality of functions means equality at every point (Definition 0.4.1), so $a\sin t + b\cos t = 0$ for **every** $t \in \mathbb{R}$. We may choose convenient points. At $t = 0$: $a \cdot 0 + b \cdot 1 = 0$, so $b = 0$. At $t = \pi/2$: $a \cdot 1 + b \cdot 0 = 0$, so $a = 0$. Hence the list is independent.

In (d) we did not need the equation at every point, only at two well-chosen ones. That is the general move for functions: **choose good points**, where most terms vanish. Note the logic. Evaluating at a point can only prove independence; if the equations at our chosen points had allowed nonzero solutions, we would have learned nothing, and would have to try other points or another method.

The degenerate cases are worth settling once, because they come up inside every later argument.

Example. Degenerate Lists

Let V be a vector space over F. Decide whether each list is independent.

(a) A list $(v_1, \ldots, v_k)$ with some $v_j = 0$.

(b) A list (v) of length one.

(c) A list $(v_1, \ldots, v_k)$ with a repeat: $v_i = v_j$ for some $i \neq j$.

Solution.

(a) Dependent. Take $a_j = 1$ and every other coefficient 0. By Theorem 1.1.5 and Theorem 1.1.6, every term is 0, so the combination is 0, and the coefficient $a_j = 1$ is not zero.

(b) Independent if and only if $v \neq 0$. If $v = 0$, use (a). If $v \neq 0$, let $av = 0$; by Theorem 1.1.9, $a = 0$ or $v = 0$, and the second is false, so $a = 0$.

(c) Dependent. Take $a_i = 1$, $a_j = -1$, and every other coefficient 0. The combination is $v_i + (-1)v_j = v_i - v_j = 0$ by Theorem 1.1.7, since $v_i = v_j$, and $a_i = 1 \neq 0$.

Now the nonexample, by a minimal change. Add one vector to the independent list $((1,0),(0,1))$ of Example (Independent Lists in Four Families) (a), to get $((1,0),(0,1),(1,1))$. Each vector is still nonzero, and any two of the three still form an independent list. But

$$1(1,0) + 1(0,1) + (-1)(1,1) = (0,0)$$

with coefficients not all zero, so the clause "$\Longrightarrow a_1 = \cdots = a_k = 0$" fails, and the list is

dependent. The polynomial version behaves the same way: $(1, x)$ is independent, but $(1, x, 1+x)$ is dependent, because $1 \cdot 1 + 1 \cdot x + (-1)(1+x) = 0$.

Quick check. Is the list $((1,2,0),(0,1,1),(1,0,-2))$ in $\mathbb{R}^3$ linearly independent?

Solution. No. Let $a(1,2,0)+b(0,1,1)+c(1,0,-2) = \mathbf{0}$. Comparing entries: $a+c=0$, $2a+b=0$, $b-2c=0$. The first two give $c=-a$ and $b=-2a$, and then the third reads $-2a+2a=0$, which holds for every a. Taking $a=1$ gives the relation $(1,2,0)-2(0,1,1)-(1,0,-2)=\mathbf{0}$, with coefficients not all zero.

Two more kinds of question come up constantly: lists that depend on a parameter, and lists of matrices. The method does not change.

Example. A Parameter and a List of Matrices

(a) For which $k \in \mathbb{R}$ is the list $((1,1,0),(1,0,1),(0,1,k))$ in $\mathbb{R}^3$ linearly independent?

(b) Is the list $\left(\begin{pmatrix}1&0\\0&1\end{pmatrix}, \begin{pmatrix}0&1\\1&0\end{pmatrix}, \begin{pmatrix}0&1\\0&0\end{pmatrix}\right)$ in $M_2(\mathbb{R})$ linearly independent?

Solution.

(a) Let $a(1,1,0)+b(1,0,1)+c(0,1,k)=\mathbf{0}$. Comparing entries gives $a+b=0$, $a+c=0$, and $b+kc=0$. The first two give $b=-a$ and $c=-a$; substituting into the third, $-a-ka=-(1+k)a=0$.
If $k \neq -1$, then $1+k \neq 0$, so $a=0$ by Theorem 0.7.1, and then $b=c=0$. The list is independent.
If $k=-1$, the equations hold for every a; taking $a=1$ gives $(1,1,0)-(1,0,1)-(0,1,-1)=\mathbf{0}$ with coefficients not all zero. The list is dependent.
So the list is independent exactly when $k \neq -1$. Notice the shape of the argument: the parameter ends up as a factor multiplying one unknown, and the only question is when that factor can be zero.

(b) Let $a\begin{pmatrix}1&0\\0&1\end{pmatrix}+b\begin{pmatrix}0&1\\1&0\end{pmatrix}+c\begin{pmatrix}0&1\\0&0\end{pmatrix}=\begin{pmatrix}0&0\\0&0\end{pmatrix}$. The left side is $\begin{pmatrix}a&b+c\\b&a\end{pmatrix}$. Two matrices are equal when their entries agree (Definition 0.9.1), so $a=0$, $b+c=0$, and $b=0$; hence $c=0$ too. The list is independent.
For matrices, "compare entries" turns one matrix equation into four scalar equations, one per entry.

1.4.3 Why lists, and why this definition

Remark. **Why lists, not sets.** In a set, repeats disappear: $\{v,v\}=\{v\}$. If independence were defined for sets, the "list" (v,v), which has the obvious relation $v-v=0$, would silently become the independent set $\{v\}$ when $v \neq 0$. The repeat can also be hidden: $x+1$ and $1+x$ look different on the page but are the same vector. A list keeps both entries

and so records the dependence. The second reason is order. Coordinates with respect to a basis (next section) are read off in order, so $(\mathbf{e}_1, \mathbf{e}_2)$ and $(\mathbf{e}_2, \mathbf{e}_1)$ must be different objects. The Linear Dependence Lemma below also refers to "the vectors before v_j", which needs an order. Reordering a list does not change whether it is independent, since a relation can be rearranged term by term.

For an arbitrary subset $S \subseteq V$, finite or infinite, we say S is **linearly independent** if **every** finite list of **distinct** vectors of S is linearly independent. The word "distinct" is essential: without it, the list (s, s) would make every nonempty set dependent. For a finite set $S = \{v_1, \ldots, v_k\}$ with the v_i distinct, this agrees with independence of the list $(v_1, \ldots, v_k)$: every list of distinct vectors of S is a reordered sub-list, and a sub-list of an independent list is independent (Exercise (B3: Sub-Lists)). Together with "S spans V" from the previous section, this is the language section 8 of this chapter uses for spaces with no finite spanning list.

Why "all $a_i = 0$", and why is the empty list independent? Suppose we only asked that **some** a_i be 0. The list $((1,0),(0,1),(0,2))$ has the redundant vector $(0,2) = 2(0,1)$. But in any relation $a(1,0) + b(0,1) + c(0,2) = \mathbf{0}$, comparing first entries forces $a = 0$, so every relation has a zero coefficient, and the list would pass the weaker test. Only "**all** $a_i = 0$" rules out every nontrivial relation. The empty list is independent because it wastes nothing, and this convention makes the empty list a basis of the zero space in the next section, just as $\text{span}() = \{0\}$ makes it a spanning list.

Warning. **"Not all zero" is not "all nonzero".** A dependence relation needs **one** nonzero coefficient, not every coefficient nonzero. The list $((1,0),(0,1),(0,2))$ is dependent, by $0(1,0) + 2(0,1) - (0,2) = \mathbf{0}$. But no relation has all three coefficients nonzero: comparing first entries in $a(1,0) + b(0,1) + c(0,2) = \mathbf{0}$ forces $a = 0$. Hunting for a relation with no zero coefficient would wrongly conclude that the list is independent.

Warning. **Pairwise independence is not independence.** Checking pairs is not enough. In the list $((1,0,1),(0,1,1),(1,1,2))$ from the start of the section, every two of the three vectors form an independent list, but the three together are dependent, since $(1,0,1) + (0,1,1) - (1,1,2) = \mathbf{0}$. For three or more vectors, independence is a condition on all of them at once.

1.4.4 The Linear Dependence Lemma

The definition says that a dependent list has a relation. It does not say which vector is redundant, and the discussion at the start of the section only found a redundant vector where the coefficient was nonzero. The next lemma makes a definite choice, and chooses it so that the redundant vector depends only on vectors **before** it in the list. This one-directional form is what later proofs need, because they build lists from left to right.

Theorem 1.4.1. Linear Dependence Lemma

Let V be a vector space over F, and let $(v_1, \ldots, v_k)$ be a linearly dependent list in V. Then:

(1) there is an index $j \in \{1, \ldots, k\}$ with $v_j \in \text{span}(v_1, \ldots, v_{j-1})$. For $j = 1$ this means $v_1 = 0$, so if $v_1 \neq 0$, then $j \geq 2$;

(2) for any such j, removing v_j does not change the span:

$$\operatorname{span}(v_1, \ldots, v_{j-1}, v_{j+1}, \ldots, v_k) = \operatorname{span}(v_1, \ldots, v_k).$$

Idea. Take a relation with coefficients not all zero. We want to solve for one vector in terms of the vectors before it, so we need a vector whose coefficient is nonzero and which comes **after** every other vector with a nonzero coefficient. That is the **largest** index j with $a_j \neq 0$. Everything after it has coefficient zero and drops out, and we divide by a_j. Part (2) is then Theorem 1.3.5 read backwards.

Proof.

(1) Since the list is dependent, there are $a_1, \ldots, a_k \in F$, not all zero, with $a_1v_1 + \cdots + a_kv_k = 0$. The set $\{i : a_i \neq 0\}$ is nonempty and finite, so it has a largest element; call it j. Then $a_j \neq 0$, and $a_i = 0$ for every $i > j$. By Theorem 1.1.5 the terms with $i > j$ are 0, so

$$a_1v_1 + \cdots + a_{j-1}v_{j-1} + a_jv_j = 0.$$

Adding the negative of $a_1v_1 + \cdots + a_{j-1}v_{j-1}$ to both sides gives $a_jv_j = -(a_1v_1 + \cdots + a_{j-1}v_{j-1})$. Since $a_j \neq 0$ and F is a field, a_j has an inverse $a_j^{-1} \in F$ (Definition 0.7.1). Multiplying both sides by a_j^{-1}, and using $a_j^{-1}(a_jv_j) = (a_j^{-1}a_j)v_j = 1v_j = v_j$ on the left, and $-x = (-1)x$ (Theorem 1.1.7) with the distributive laws on the right, we get

$$v_j = (-a_j^{-1}a_1)v_1 + \cdots + (-a_j^{-1}a_{j-1})v_{j-1}.$$

Hence $v_j \in \operatorname{span}(v_1, \ldots, v_{j-1})$. When $j = 1$ the right side is the empty sum, so $v_1 = 0$; therefore if $v_1 \neq 0$, then $j \geq 2$.

(2) Let j be an index with $v_j \in \operatorname{span}(v_1, \ldots, v_{j-1})$, and let $L = (v_1, \ldots, v_{j-1}, v_{j+1}, \ldots, v_k)$. Each of $v_1, \ldots, v_{j-1}$ is an entry of L, so by monotonicity (Proposition 1.3.4 (2)) $\operatorname{span}(v_1, \ldots, v_{j-1}) \subseteq \operatorname{span}(L)$, and hence $v_j \in \operatorname{span}(L)$. By Theorem 1.3.5, inserting v_j back into L at position j does not change the span. That is, $\operatorname{span}(v_1, \ldots, v_k) = \operatorname{span}(L)$, as claimed.

■

The index j need not be unique, and the lemma does **not** say that every vector depends on earlier ones. A small example shows both.

Example. The Lemma on a Concrete List

In $\mathbb{R}^2$, let $\mathbf{v}_1 = (1, 1)$, $\mathbf{v}_2 = (1, -1)$, $\mathbf{v}_3 = (2, 0)$, $\mathbf{v}_4 = (0, 1)$. Find two different linear relations, and the index j that the proof of Theorem 1.4.1 extracts from each. Which indices j satisfy $\mathbf{v}_j \in \operatorname{span}(\mathbf{v}_1, \ldots, \mathbf{v}_{j-1})$?

Solution. First relation: $\mathbf{v}_1 + \mathbf{v}_2 - \mathbf{v}_3 = (1 + 1 - 2, 1 - 1 - 0) = (0, 0)$. The largest index with a nonzero coefficient is $j = 3$, and solving gives $\mathbf{v}_3 = \mathbf{v}_1 + \mathbf{v}_2 \in \operatorname{span}(\mathbf{v}_1, \mathbf{v}_2)$.

Second relation: $\mathbf{v}_1 - \mathbf{v}_2 - 2\mathbf{v}_4 = (1 - 1 - 0, 1 + 1 - 2) = (0, 0)$. Here the largest index is

$j = 4$, and $\mathbf{v}_4 = \frac{1}{2}\mathbf{v}_1 - \frac{1}{2}\mathbf{v}_2 \in \text{span}(\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3)$.

So $j = 3$ and $j = 4$ both qualify, and by part (2) of the lemma either $\mathbf{v}_3$ or $\mathbf{v}_4$ can be removed without changing the span. On the other hand, $j = 2$ does not qualify: $\mathbf{v}_2 = (1, -1)$ is not a multiple of $\mathbf{v}_1 = (1, 1)$, because a multiple (c, c) has equal entries. And $j = 1$ does not occur since $\mathbf{v}_1 \neq \mathbf{0}$. Removing $\mathbf{v}_3$ and then (by the lemma applied to $(\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_4)$, which is still dependent by the second relation) $\mathbf{v}_4$ leaves $(\mathbf{v}_1, \mathbf{v}_2)$, whose span is $\mathbb{R}^2$ by Example (Spanning Lists).

The lemma only locates vectors that depend on **earlier** ones. Here $\mathbf{v}_1$ and $\mathbf{v}_2$ could also be removed without changing the span, since $\mathbf{v}_1 = \mathbf{v}_3 - \mathbf{v}_2$ and $\mathbf{v}_2 = \mathbf{v}_3 - \mathbf{v}_1$, but the lemma never points at them.

Read the other way round, part (1) is a test for independence that works left to right.

Remark. **Independence, one vector at a time.** A list $(v_1, \ldots, v_k)$ is linearly independent **if and only if** no v_j lies in $\text{span}(v_1, \ldots, v_{j-1})$. The "if" direction is the contrapositive of part (1). For "only if", suppose $v_j = c_1 v_1 + \cdots + c_{j-1} v_{j-1}$. Then $c_1 v_1 + \cdots + c_{j-1} v_{j-1} + (-1) v_j = 0$, with zero coefficients on $v_{j+1}, \ldots, v_k$, is a relation whose coefficient -1 is nonzero. So the list is dependent. In words: a list is independent exactly when each vector reaches something new beyond the vectors before it. For instance $(1, x, x^2)$ passes: $1 \neq 0$, x is not a constant, and x^2 has degree too high to be in $\text{span}(1, x)$.

Combining the two parts of the lemma gives a clean statement about spans alone: a list is dependent exactly when it contains a vector that can be thrown away for free.

Theorem 1.4.2. Span Preservation

Let V be a vector space and $(v_1, \ldots, v_k)$ a list in V with $k \geq 1$. Then the list is linearly dependent **if and only if** there is an index j such that

$$\text{span}(v_1, \ldots, v_{j-1}, v_{j+1}, \ldots, v_k) = \text{span}(v_1, \ldots, v_k).$$

Proof. ($\Rightarrow$) By Theorem 1.4.1 (1) there is an index j with $v_j \in \text{span}(v_1, \ldots, v_{j-1})$, and by part (2) removing that v_j does not change the span.

($\Leftarrow$) Let L be the list with v_j removed, and suppose $\text{span}(L) = \text{span}(v_1, \ldots, v_k)$. By Theorem 1.3.1 (2), $v_j \in \text{span}(v_1, \ldots, v_k) = \text{span}(L)$. So $v_j = \sum_{i \neq j} c_i v_i$ for some scalars c_i, and therefore $\sum_{i \neq j} c_i v_i + (-1) v_j = 0$. The coefficient of v_j is $-1 \neq 0$, so the list is dependent. ∎

So a spanning list that is dependent can always be shortened and still span. Repeating this until the list becomes independent is the idea behind the next section's proof that every finite spanning list contains a basis.

1.4.5 Unique representations

Independence was introduced to rule out waste. Its most useful consequence is about uniqueness. If a vector could be built from a list in two different ways, subtracting the two recipes would give a nontrivial way to build 0. So for an independent list, every vector in the span has exactly one recipe.

Theorem 1.4.3. Independence and Uniqueness of Coefficients

Let V be a vector space over F and $(v_1, \dots, v_k)$ a list in V. The following are equivalent:

(a) $(v_1, \dots, v_k)$ is linearly independent;

(b) every $v \in \operatorname{span}(v_1, \dots, v_k)$ has **exactly one** representation as a linear combination: if $a_1v_1 + \cdots + a_kv_k = b_1v_1 + \cdots + b_kv_k$, then $a_i = b_i$ for every i.

Proof. (a $\Rightarrow$ b) Suppose the list is independent, and let $a_1v_1 + \cdots + a_kv_k = b_1v_1 + \cdots + b_kv_k$. Subtracting the right side from both sides and using $av - bv = (a-b)v$ (Theorem 1.1.8 and Definition 1.1.1) gives

$$(a_1 - b_1)v_1 + \cdots + (a_k - b_k)v_k = 0.$$

By independence, $a_i - b_i = 0$ for every i, that is, $a_i = b_i$.

(b $\Rightarrow$ a) Suppose (b) holds, and let $a_1v_1 + \cdots + a_kv_k = 0$. By Theorem 1.1.5, also $0v_1 + \cdots + 0v_k = 0$. These are two representations of $0 \in \operatorname{span}(v_1, \dots, v_k)$, so by (b), $a_i = 0$ for every i. Hence the list is independent.

■

Every vector of a span has **at least** one representation, by definition of the span; independence is what makes it **at most** one. For a dependent list uniqueness fails: with the list $((1,0),(0,1),(1,1))$,

$$(1,1) = 1(1,0) + 1(0,1) + 0(1,1) = 0(1,0) + 0(0,1) + 1(1,1).$$

In the next section, a list that is independent **and** spans is called a basis. By this theorem every vector then has a unique list of coefficients, its coordinates.

1.4.6 Polynomials of distinct degrees

For polynomials there is a test that needs no equations at all. The list $(x^2 + x, x - 3, 5)$ is independent, and the reason is visible: the degrees are 2, 1, and 0, and combining lower-degree polynomials can never produce a higher degree.

Theorem 1.4.4. Polynomials of Distinct Degrees Are Independent

Let F be a field and let $p_1, \dots, p_k \in F[x]$ be **nonzero** polynomials whose degrees are **pairwise distinct**. Then the list $(p_1, \dots, p_k)$ is linearly independent.

Idea. Put the polynomials in order of increasing degree. If the list were dependent, the Linear Dependence Lemma would give a polynomial lying in the span of the ones before it. Those all have smaller degree, so their span consists of polynomials of smaller degree, and

cannot contain it.

Proof. A linear relation among the p_i can be rearranged term by term, so reordering the list does not affect independence. Hence we may assume $\deg p_1 < \deg p_2 < \cdots < \deg p_k$.

Suppose, for a contradiction, that the list is dependent. Since $p_1 \neq 0$, Theorem 1.4.1 gives an index $j \geq 2$ with $p_j \in \operatorname{span}(p_1, \ldots, p_{j-1})$. Let $n = \deg p_{j-1}$, a natural number since $p_{j-1} \neq 0$. For every $i \leq j-1$ we have $\deg p_i \leq n$, so $p_1, \ldots, p_{j-1}$ lie in $F[x]_{\leq n}$, which is a subspace of $F[x]$ by Example (Polynomials of Bounded Degree). By Theorem 1.3.1 (3), $\operatorname{span}(p_1, \ldots, p_{j-1}) \subseteq F[x]_{\leq n}$. Hence $\deg p_j \leq n = \deg p_{j-1}$, contradicting $\deg p_{j-1} < \deg p_j$. Therefore the list is linearly independent.

■

Both hypotheses matter. The zero polynomial must be excluded, because a list containing 0 is dependent (Example (Degenerate Lists)). And the converse fails: distinct degrees are sufficient, not necessary. The list $(x, x+1)$ has equal degrees but is independent, since $ax + b(x+1) = (a+b)x + b = 0$ forces $b = 0$ and then $a = 0$.

1.4.7 Independence depends on the field

The same set can be a vector space over two fields, and the answer to "independent?" can change with the field, because the field decides which coefficients are allowed.

Warning. **Always ask: independent over which field?** In $\mathbb{C}^2$, let $\mathbf{v}_1 = (1, i)$ and $\mathbf{v}_2 = (i, -1)$. Over $\mathbb{C}$ the list $(\mathbf{v}_1, \mathbf{v}_2)$ is **dependent**: $i\mathbf{v}_1 = (i, i^2) = (i, -1) = \mathbf{v}_2$, so $i\mathbf{v}_1 + (-1)\mathbf{v}_2 = \mathbf{0}$ with coefficient $i \neq 0$. Over $\mathbb{R}$ (the same addition, but only real scalars, as in Exercise (C2: The Span Depends on the Field)) the list is **independent**: if $a\mathbf{v}_1 + b\mathbf{v}_2 = (a + bi, ai - b) = (0, 0)$ with $a, b \in \mathbb{R}$, then the first entry $a + bi = 0$ forces $a = b = 0$, since a complex number is zero only when its real and imaginary parts are. The relation over $\mathbb{C}$ used the coefficient i, which is not available over $\mathbb{R}$.

A second example of the same phenomenon: $((1,1),(1,-1))$ is independent in $\mathbb{R}^2$ (Example (Independent Lists in Four Families) (c)), but in $\mathbb{F}_2^2$, where $-1 = 1$, it is the list $((1,1),(1,1))$ with a repeat, hence dependent. The proof over $\mathbb{R}$ divided by 2, which is 0 in $\mathbb{F}_2$.

1.4.8 Exercises

1.4.8.1 A. Check your understanding

Exercise. A1

(a) Define what it means for a list $(v_1, \ldots, v_k)$ in a vector space over F to be linearly independent, and what it means for it to be linearly dependent.

(b) Determine whether the following statement is true: "a list of vectors is linearly dependent if and only if one of its vectors is a scalar multiple of another." Justify your answer.

(c) Is the empty list linearly independent? Is the list (0)?

(d) Determine whether the following statement is true: "if (v_1, v_2, v_3) is linearly dependent, then $v_3 \in \operatorname{span}(v_1, v_2)$." Justify your answer.

(e) To prove that a list of functions $\mathbb{R} \to \mathbb{R}$ is independent, what does "$= 0$" mean, and what is a standard way to extract scalar equations?

Solution.

(a) It is independent if for all $a_1, \dots, a_k \in F$, $a_1v_1 + \cdots + a_kv_k = 0$ implies $a_1 = \cdots = a_k = 0$. It is dependent if there are $a_1, \dots, a_k \in F$, not all zero, with $a_1v_1 + \cdots + a_kv_k = 0$.

(b) False. The list $((1,0,1),(0,1,1),(1,1,2))$ in $\mathbb{R}^3$ is dependent, since the third vector is the sum of the first two, but no vector is a multiple of another. (The "if" direction is true: $v_i = cv_j$ with $i \neq j$ gives the relation $v_i - cv_j = 0$ with coefficient $1 \neq 0$.)

(c) The empty list is independent: there are no coefficients, so the condition holds vacuously. The list (0) is dependent: $1 \cdot 0 = 0$ with coefficient $1 \neq 0$ (Example (Degenerate Lists)).

(d) False. Take $\mathbf{v}_1 = (1,0)$, $\mathbf{v}_2 = (2,0)$, $\mathbf{v}_3 = (0,1)$ in $\mathbb{R}^2$. The list is dependent, since $2\mathbf{v}_1 - \mathbf{v}_2 = \mathbf{0}$. But every combination of $\mathbf{v}_1, \mathbf{v}_2$ has second entry 0, so $\mathbf{v}_3 \notin \operatorname{span}(\mathbf{v}_1, \mathbf{v}_2)$. Theorem 1.4.1 only promises **some** j; here $j = 2$.

(e) It means the zero function: the combination takes the value 0 at **every** point. Evaluating at well-chosen points turns it into scalar equations for the coefficients.

1.4.8.2 B. Practice

Exercise. B1: Independent or Not?

Determine which of the following lists are linearly independent. Justify your answer.

(a) $((1,0,2),(0,1,1),(1,1,3))$ in $\mathbb{R}^3$.

(b) $((1,1,0),(0,1,1),(1,0,1))$ in $\mathbb{R}^3$.

(c) (x^2+1, x^2-1, x) in $\mathbb{R}[x]$.

(d) $\left(\begin{pmatrix}1&0\\0&1\end{pmatrix}, \begin{pmatrix}0&1\\1&0\end{pmatrix}, \begin{pmatrix}1&1\\1&1\end{pmatrix}\right)$ in $M_2(\mathbb{R})$.

(e) (f, g) in $\mathbb{R}^{\mathbb{R}}$, where $f(t) = e^t$ and $g(t) = e^{2t}$.

Solution.

(a) Dependent. Since $(1,0,2)+(0,1,1) = (1,1,3)$, we have $(1,0,2)+(0,1,1)-(1,1,3) = \mathbf{0}$, with coefficients not all zero.

(b) Independent. Let $a(1,1,0) + b(0,1,1) + c(1,0,1) = \mathbf{0}$. Comparing entries: $a + c = 0$, $a + b = 0$, $b + c = 0$. The first two give $c = -a$ and $b = -a$; then the third gives $-2a = 0$, so $a = 0$, and hence $b = c = 0$.

(c) Independent. Let $a(x^2+1)+b(x^2-1)+cx = 0$. The left side is $(a+b)x^2+cx+(a-b)$. Comparing coefficients (Definition 0.8.1): $a + b = 0$, $c = 0$, $a - b = 0$. Adding the first and third gives $2a = 0$, so $a = 0$, and then $b = 0$.

(d) Dependent. The first two matrices add up to the third, so $1 \cdot \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + 1 \cdot \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} + (-1) \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$ is the zero matrix.

(e) Independent. Let $af + bg = 0$, the zero function. Then $ae^t + be^{2t} = 0$ for every $t \in \mathbb{R}$. At $t = 0$: $a + b = 0$. At $t = \ln 2$: $2a + 4b = 0$, so $a + 2b = 0$. Subtracting the first equation gives $b = 0$, and then $a = 0$.

Exercise. B2: A Parameter

For which $t \in \mathbb{R}$ is the list $((1, t), (t, 1))$ in $\mathbb{R}^2$ linearly independent? Justify your answer.

Solution. Let $a(1, t) + b(t, 1) = (0, 0)$, that is, $a + tb = 0$ and $ta + b = 0$. The first equation gives $a = -tb$. Substituting into the second, $-t^2b + b = (1 - t^2)b = 0$.

If $t \neq \pm 1$, then $1 - t^2 \neq 0$, so $b = 0$ by Theorem 0.7.1, and then $a = -tb = 0$. The list is independent.

If $t = 1$, the list is $((1, 1), (1, 1))$, which has a repeat and is dependent. If $t = -1$, the list is $((1, -1), (-1, 1))$, and $1 \cdot (1, -1) + 1 \cdot (-1, 1) = (0, 0)$, so it is dependent.

Hence the list is linearly independent exactly when $t \neq 1$ and $t \neq -1$.

Exercise. B3: Sub-Lists

A **sub-list** of $(v_1, \ldots, v_k)$ is a list $(v_{i_1}, \ldots, v_{i_m})$ with $1 \leq i_1 < i_2 < \cdots < i_m \leq k$.

(a) Prove that every sub-list of a linearly independent list is linearly independent.

(b) Deduce that if some sub-list of a list is linearly dependent, then the whole list is linearly dependent.

Solution.

(a) Let $(v_1, \ldots, v_k)$ be independent, and let $b_1 v_{i_1} + \cdots + b_m v_{i_m} = 0$. Define $a_i = b_r$ if $i = i_r$ for some r (the indices i_r are distinct, so this is well defined), and $a_i = 0$ otherwise. By Theorem 1.1.5 the extra terms are 0, so $a_1 v_1 + \cdots + a_k v_k = b_1 v_{i_1} + \cdots + b_m v_{i_m} = 0$. By independence, every $a_i = 0$; in particular every $b_r = a_{i_r} = 0$. Hence the sub-list is independent.

(b) This is the contrapositive of (a): if the whole list were independent, every sub-list would be independent by (a).

1.4.8.3 C. Going deeper

Exercise. C1: Sums of Pairs, and the Field $\mathbb{F}_2$

Let V be a vector space over F and (u, v, w) a linearly independent list in V.

(a) Suppose $F = \mathbb{R}$. Prove that $(u+v, v+w, w+u)$ is linearly independent.

(b) Suppose $F = \mathbb{F}_2$. Prove that $(u+v, v+w, w+u)$ is linearly dependent, and give an example of an independent list (u, v, w) over $\mathbb{F}_2$, so that the statement is not vacuous.

(c) Which property of $\mathbb{R}$ did (a) use that $\mathbb{F}_2$ lacks? Deduce that (a) holds over every field in which $1 + 1 \neq 0$.

Solution.

(a) Let $a(u+v) + b(v+w) + c(w+u) = 0$. Expanding and regrouping with the axioms of Definition 1.1.1 gives $(a+c)u + (a+b)v + (b+c)w = 0$. By independence of (u, v, w), $a + c = 0$, $a + b = 0$, and $b + c = 0$. The first two give $c = -a$ and $b = -a$, and then the third gives $-2a = 0$. Since $2 \neq 0$ in $\mathbb{R}$, $a = 0$, and hence $b = c = 0$. This proves independence.

(b) Over $\mathbb{F}_2$,

$$\begin{aligned} 1(u+v) + 1(v+w) + 1(w+u) &= (1+1)u + (1+1)v + (1+1)w \\ &= 0u + 0v + 0w = 0, \end{aligned}$$

using $1 + 1 = 0$ in $\mathbb{F}_2$ and Theorem 1.1.5. The coefficients are $1 \neq 0$, so the list is dependent. For an example, $(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$ in $\mathbb{F}_2^3$ is independent by the argument of Example (Independent Lists in Four Families) (a), which works over any field.

(c) Part (a) divided by $2 = 1 + 1$, that is, used $2 \neq 0$, to pass from $-2a = 0$ to $a = 0$. In $\mathbb{F}_2$, $2 = 0$. In any field with $1 + 1 \neq 0$, the element 2 has an inverse, and the proof of (a) goes through word for word.

Exercise. C2: Adding a Vector outside the Span

Let $(v_1, \dots, v_k)$ be a linearly independent list in V, and let $v \in V$ with $v \notin \operatorname{span}(v_1, \dots, v_k)$. Prove that $(v_1, \dots, v_k, v)$ is linearly independent. Give an example showing that the hypothesis $v \notin \operatorname{span}(v_1, \dots, v_k)$ cannot be dropped.

Solution. Let $a_1 v_1 + \cdots + a_k v_k + av = 0$. Suppose $a \neq 0$. Then, as in the proof of Theorem 1.4.1, $v = (-a^{-1}a_1)v_1 + \cdots + (-a^{-1}a_k)v_k \in \operatorname{span}(v_1, \dots, v_k)$, contradicting the hypothesis. Hence $a = 0$, and the relation becomes $a_1 v_1 + \cdots + a_k v_k = 0$. By independence of $(v_1, \dots, v_k)$, $a_1 = \cdots = a_k = 0$. So all coefficients are zero, and the longer list is independent.

For the example, take $(\mathbf{e}_1)$ in $\mathbb{R}^2$ and $\mathbf{v} = 2\mathbf{e}_1 \in \operatorname{span}(\mathbf{e}_1)$. The list $(\mathbf{e}_1)$ is independent, but $(\mathbf{e}_1, 2\mathbf{e}_1)$ is dependent, by $2\mathbf{e}_1 - 2\mathbf{e}_1 = \mathbf{0}$.

Exercise. C3: $1, \sqrt{2}, \sqrt{3}$ over $\mathbb{Q}$

Regard $\mathbb{R}$ as a vector space over $\mathbb{Q}$. Prove that the list $(1, \sqrt{2}, \sqrt{3})$ is linearly independent over $\mathbb{Q}$. Is it linearly independent over $\mathbb{R}$?

Hint: you may use without proof that $\sqrt{3}$ and $\sqrt{6}$ are irrational (the proof of Theorem 0.2.3 adapts). Isolate one square root and square.

Solution. Let $a+b\sqrt{2}+c\sqrt{3}=0$ with $a,b,c\in\mathbb{Q}$. Then $a+b\sqrt{2}=-c\sqrt{3}$, and squaring both sides gives $a^2+2b^2+2ab\sqrt{2}=3c^2$.

If $ab\neq 0$, then $\sqrt{2}=\dfrac{3c^2-a^2-2b^2}{2ab}\in\mathbb{Q}$, contradicting Theorem 0.2.3. Hence $ab=0$, so $a=0$ or $b=0$.

Case 1: $b=0$. Then $a+c\sqrt{3}=0$. If $c\neq 0$, then $\sqrt{3}=-a/c\in\mathbb{Q}$, a contradiction. So $c=0$, and then $a=0$.

Case 2: $a=0$. Then $b\sqrt{2}+c\sqrt{3}=0$. Multiplying by $\sqrt{2}$ gives $2b+c\sqrt{6}=0$. If $c\neq 0$, then $\sqrt{6}=-2b/c\in\mathbb{Q}$, a contradiction. So $c=0$, then $b\sqrt{2}=0$, and $b=0$ since $\sqrt{2}\neq 0$.

In both cases $a=b=c=0$, so the list is linearly independent over $\mathbb{Q}$.

Over $\mathbb{R}$ it is dependent: $\sqrt{2}\cdot 1+(-1)\cdot\sqrt{2}+0\cdot\sqrt{3}=0$, and the coefficient $\sqrt{2}$ is a nonzero real number. Over $\mathbb{R}$ any two nonzero real numbers form a dependent list.

1.5 Bases and Coordinates

Spanning lists and independent lists each solve half of a problem. A spanning list reaches every vector of the space, but it may reach some vectors in more than one way. An independent list never wastes a vector, but it may not reach everything. This section puts the two halves together. The result, a basis, turns an abstract vector space into columns of numbers, and we prove that every space spanned by a finite list has one.

1.5.1 What makes coordinates work

Start with the plane. Every vector of $\mathbb{R}^2$ can be written as

$$(x,y)=x\mathbf{e}_1+y\mathbf{e}_2,$$

and the pair (x,y) is its address. Two separate facts make this addressing system work. First, **every** vector has an address: the list $(\mathbf{e}_1,\mathbf{e}_2)$ spans $\mathbb{R}^2$. Second, every vector has **only one** address: if $x\mathbf{e}_1+y\mathbf{e}_2=x'\mathbf{e}_1+y'\mathbf{e}_2$, then $(x,y)=(x',y')$, so $x=x'$ and $y=y'$.

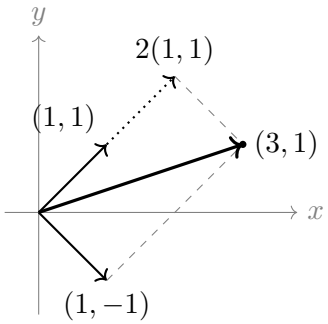
Nothing about this is special to $\mathbf{e}_1$ and $\mathbf{e}_2$. Take instead the list $((1,1),(1,-1))$. To find an address for (x,y) we solve

$$a(1,1)+b(1,-1)=(x,y),\qquad\text{that is,}\qquad a+b=x,\quad a-b=y.$$

Adding and subtracting the two equations gives $2a=x+y$ and $2b=x-y$. So there is a solution, and it is the only one:

$$a=\frac{x+y}{2},\qquad b=\frac{x-y}{2}.$$

For example $(3,1)=2(1,1)+1(1,-1)$, and no other pair of scalars works. The grid is slanted, but every point still has exactly one address.



Look at which property did which job. **Existence** of an address is spanning. **Uniqueness** of the address is independence: by Theorem 1.4.3, a list is independent exactly when every vector in its span has only one expression as a combination of it. A coordinate system is a list that does both jobs at once.

A basis is a list with enough vectors to build everything and no vector to spare.

1.5.2 Bases

Definition 1.5.1. Basis

Let V be a vector space over F. A **basis** of V is a **list** $(v_1, \ldots, v_n)$ of vectors in V such that

(B1) $(v_1, \ldots, v_n)$ is **linearly independent**, and
(B2) $(v_1, \ldots, v_n)$ **spans** V, that is, $\operatorname{span}(v_1, \ldots, v_n) = V$.

We usually name a basis with a calligraphic letter, $\mathcal{B} = (v_1, \ldots, v_n)$, and call n its **length**. The plural of basis is **bases**.

In words: (B1) says that the only combination of the v_i equal to 0 is the one with **all** coefficients zero, so no vector in the list is wasted. (B2) says that **every** vector of V is some combination of the v_i, so nothing is out of reach. The basis is a list, so its entries come in a definite order, and a basis is always a basis **of** a particular space: $(\mathbf{e}_1, \mathbf{e}_2)$ is a basis of $\mathbb{R}^2$, but not of $\mathbb{R}^3$.

We start with the bases that come with the running examples.

Example. Standard Bases

Show that each of the following lists is a basis. They are called the **standard bases**.

(a) $(\mathbf{e}_1, \ldots, \mathbf{e}_n)$ in F^n, where $\mathbf{e}_i$ has 1 in position i and 0 elsewhere.
(b) $(1, x, x^2, \ldots, x^n)$ in $F[x]_{\leq n}$.
(c) $(\mathbf{E}_{11}, \mathbf{E}_{12}, \ldots, \mathbf{E}_{1n}, \mathbf{E}_{21}, \ldots, \mathbf{E}_{mn})$ in $M_{m\times n}(F)$, where the **matrix unit** $\mathbf{E}_{ij}$ has 1 in entry (i, j) and 0 elsewhere, listed row by row.

Solution.

(a) (B2): every $(x_1, \ldots, x_n) \in F^n$ equals $x_1\mathbf{e}_1 + \cdots + x_n\mathbf{e}_n$. (B1): let $a_1\mathbf{e}_1 + \cdots + a_n\mathbf{e}_n = \mathbf{0}$. The left side is the column $(a_1, \ldots, a_n)$, so comparing entries gives $a_1 = \cdots = a_n = 0$.
(b) (B2): by Definition 0.8.4, every element of $F[x]_{\leq n}$ has the form $a_0 + a_1x + \cdots + a_nx^n$,

which is a combination of $1, x, \ldots, x^n$. (B1): let $a_0 \cdot 1 + a_1 x + \cdots + a_n x^n = 0$. Here 0 is the zero polynomial, all of whose coefficients are 0. Two polynomials are equal exactly when their coefficients agree (Definition 0.8.1), so $a_0 = a_1 = \cdots = a_n = 0$.

(c) (B2): a matrix $\mathbf{A} = (a_{ij})$ equals $\sum_{i,j} a_{ij}\mathbf{E}_{ij}$, since both sides have a_{ij} in entry (i, j). (B1): if $\sum_{i,j} c_{ij}\mathbf{E}_{ij}$ is the zero matrix, then its (i, j) entry c_{ij} is 0 for every i, j.

The next examples show that a space has other bases, that the field matters, and what happens in the smallest space.

Example. Three More Bases

Show that each list is a basis.

(a) $((1, 1), (1, -1))$ in $\mathbb{R}^2$.

(b) $(1, i)$ in $\mathbb{C}$, regarded as a vector space **over** $\mathbb{R}$.

(c) The empty list $()$ in the zero space $\{0\}$.

Solution.

(a) We showed above that for every $(x, y) \in \mathbb{R}^2$ the equation $a(1, 1) + b(1, -1) = (x, y)$ has a solution, so (B2) holds. Taking $(x, y) = (0, 0)$, the only solution is $a = \frac{0+0}{2} = 0$ and $b = 0$, so (B1) holds.

(b) The scalars are now real. (B2): every complex number is $a + bi = a \cdot 1 + b \cdot i$ with $a, b \in \mathbb{R}$ (Definition 0.6.1). (B1): let $a \cdot 1 + b \cdot i = 0$ with $a, b \in \mathbb{R}$. A complex number is 0 exactly when its real and imaginary parts are 0, so $a = b = 0$.

(c) The empty list is linearly independent: there are no coefficients, so the condition "all coefficients are zero" holds vacuously. Its span is $\{0\}$ by the convention for the empty span. So both (B1) and (B2) hold.

The degenerate case (c) matters more than it looks. Without it, the zero space would be the one space with no basis, and every theorem below would need an exception for it.

Nonexamples by minimal change. Add one vector to the standard basis of $\mathbb{R}^2$, making $((1, 0), (0, 1), (1, 1))$. Spanning (B2) still holds, since the first two vectors already span. But (B1) fails:

$$1 \cdot (1, 0) + 1 \cdot (0, 1) + (-1) \cdot (1, 1) = (0, 0),$$

with coefficients **not all zero**. The effect is exactly the loss of unique addresses: $(3, 1) = 3(1, 0) + 1(0, 1) + 0(1, 1) = 2(1, 0) + 0(0, 1) + 1(1, 1)$.

Now remove a vector instead. In $\mathbb{R}^3$, the list $((1, 0, 0), (0, 1, 0))$ is still independent, since $a(1, 0, 0) + b(0, 1, 0) = (a, b, 0) = \mathbf{0}$ forces $a = b = 0$. But (B2) fails: every combination $(a, b, 0)$ has third entry 0, so $(0, 0, 1)$ is out of reach.

Quick check. Is $((1, 2), (2, 1), (0, 0))$ a basis of $\mathbb{R}^2$? If not, which of (B1) and (B2) fails?

Solution. No. (B2) holds: $(1,2)$ and $(2,1)$ already span $\mathbb{R}^2$, because solving $a(1,2)+b(2,1) = (x,y)$ gives $a = \frac{2y-x}{3}$ and $b = \frac{2x-y}{3}$. But (B1) fails, because the list contains the zero vector: $0(1,2) + 0(2,1) + 1(0,0) = (0,0)$ with a nonzero coefficient.

Why this definition. Drop (B1), and addresses stop being unique, as the list $((1,0),(0,1),(1,1))$ showed. Drop (B2), and some vectors have no address at all. Keeping both is exactly what an addressing system needs. The name comes from the everyday word: the basis is what everything else is built on.

Remark. Why a list and not a set? Two reasons. Coordinates, defined below, are a column of numbers, and a column needs to know which scalar comes first. And a set would hide repeats: the list $((1,0),(1,0),(0,1))$ is dependent, since $1(1,0) - 1(1,0) = \mathbf{0}$, but as a set $\{(1,0),(1,0),(0,1)\} = \{(1,0),(0,1)\}$ it would look like the standard basis.

Warning. Say **a** basis, not **the** basis. A space usually has many bases; a nonzero space over an infinite field such as $\mathbb{R}$ has infinitely many. For instance $((1,0),(c,1))$ is a basis of $\mathbb{R}^2$ for **every** real number c, since $(x,y) = (x-cy)(1,0)+y(c,1)$, and this is the only solution. The standard basis is a convenient choice, not a canonical one.

1.5.3 Unique representation and coordinates

We found the definition by asking for addresses that exist and are unique. The first result confirms that this is exactly what a basis delivers, and nothing more.

Theorem 1.5.1. Unique Representation

Let V be a vector space over F, and let $\mathcal{B} = (v_1, \ldots, v_n)$ be a list in V. Then $\mathcal{B}$ is a basis of V if and only if every $v \in V$ can be written as

$$v = a_1 v_1 + \cdots + a_n v_n \qquad (a_1, \ldots, a_n \in F)$$

in **exactly one** way.

Proof. ($\Rightarrow$) Let $v \in V$. Since $\mathcal{B}$ spans V, there are scalars with $v = a_1 v_1 + \cdots + a_n v_n$. Suppose also $v = b_1 v_1 + \cdots + b_n v_n$. Subtracting the two expressions and regrouping with the vector space axioms,

$$(a_1 - b_1)v_1 + \cdots + (a_n - b_n)v_n = v - v = 0.$$

Since $\mathcal{B}$ is linearly independent, $a_i - b_i = 0$ for every i. Hence the two expressions are the same.

($\Leftarrow$) Every $v \in V$ is a combination of $\mathcal{B}$, so $\mathcal{B}$ spans V. For independence, let $a_1 v_1 + \cdots + a_n v_n = 0$. We also have $0 = 0v_1 + \cdots + 0v_n$. By uniqueness of the expression for the vector

0, all $a_i = 0$. Hence $\mathcal{B}$ is linearly independent, and $\mathcal{B}$ is a basis.

■

This is Theorem 1.4.3 in the special case where the span is the whole space. The theorem also gives a second way to check that a list is a basis: solve for the coefficients of a general vector, and see that there is always exactly one solution. We did exactly that for $((1,1),(1,-1))$.

Whenever something exists and is unique, it deserves a name.

Definition 1.5.2. Coordinates

Let $\mathcal{B} = (v_1, \dots, v_n)$ be a basis of a vector space V over F, and let $v \in V$. The unique scalars $a_1, \dots, a_n$ with $v = a_1 v_1 + \cdots + a_n v_n$ are the **coordinates** of v **with respect to** $\mathcal{B}$. The column

$$[v]_{\mathcal{B}} := \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix} \in F^n$$

is the **coordinate vector** of v with respect to $\mathcal{B}$.

In the standard basis of F^n, the coordinate vector of $(x_1, \dots, x_n)$ is the column $(x_1, \dots, x_n)$ itself. That is why we rarely notice coordinates in F^n. In other bases and other spaces they carry real information.

Example. Coordinates of a Polynomial

Show that $\mathcal{B} = (1, 1+x, 1+x+x^2)$ is a basis of $\mathbb{R}[x]_{\le 2}$, and find $[p]_{\mathcal{B}}$ for $p = 3 - x + 2x^2$.

Solution. We use Theorem 1.5.1. Let $q = q_0 + q_1 x + q_2 x^2 \in \mathbb{R}[x]_{\le 2}$ be arbitrary. For scalars a, b, c,

$$a \cdot 1 + b(1+x) + c(1+x+x^2) = (a+b+c) + (b+c)x + cx^2.$$

Two polynomials are equal exactly when their coefficients agree, so this equals q if and only if

$$c = q_2, \qquad b + c = q_1, \qquad a + b + c = q_0.$$

Solving from the top, $c = q_2$, then $b = q_1 - q_2$, then $a = q_0 - q_1$. So every q has exactly one expression, and $\mathcal{B}$ is a basis of $\mathbb{R}[x]_{\le 2}$.

For $p = 3 - x + 2x^2$ we have $q_0 = 3$, $q_1 = -1$, $q_2 = 2$. Hence $c = 2$, $b = -1 - 2 = -3$, and $a = 3 - (-1) = 4$:

$$[p]_{\mathcal{B}} = \begin{pmatrix} 4 \\ -3 \\ 2 \end{pmatrix}.$$

Check: $4 - 3(1+x) + 2(1+x+x^2) = (4 - 3 + 2) + (-3 + 2)x + 2x^2 = 3 - x + 2x^2$. In the standard basis $(1, x, x^2)$, the same polynomial has coordinate vector $(3, -1, 2)$.

Warning. Coordinates depend on the basis **and on its order**. Take $\mathbf{v} = (3, 1) \in \mathbb{R}^2$. With respect to the standard basis $\mathcal{E} = (\mathbf{e}_1, \mathbf{e}_2)$, $[\mathbf{v}]_{\mathcal{E}} = (3, 1)$. With respect to $\mathcal{B} = ((1, 1), (1, -1))$, $[\mathbf{v}]_{\mathcal{B}} = (2, 1)$. With respect to $\mathcal{B}' = ((1, -1), (1, 1))$, the same two vectors in the other order, $[\mathbf{v}]_{\mathcal{B}'} = (1, 2)$. A column of numbers means nothing until you say which ordered basis it refers to.

The point of coordinates is that they turn computations in V into computations with columns. That works because they respect the two operations.

Theorem 1.5.2. Coordinates Respect the Operations

Let $\mathcal{B} = (v_1, \ldots, v_n)$ be a basis of a vector space V over F. Then for all $u, v \in V$ and $c \in F$,

$$[u + v]_{\mathcal{B}} = [u]_{\mathcal{B}} + [v]_{\mathcal{B}}, \qquad [cv]_{\mathcal{B}} = c\,[v]_{\mathcal{B}}.$$

Moreover, every column $(a_1, \ldots, a_n) \in F^n$ is the coordinate vector of exactly one vector of V, namely $a_1 v_1 + \cdots + a_n v_n$.

Proof. Let $[u]_{\mathcal{B}} = (a_1, \ldots, a_n)$ and $[v]_{\mathcal{B}} = (b_1, \ldots, b_n)$. By the vector space axioms,

$$u + v = \sum_{i=1}^{n} a_i v_i + \sum_{i=1}^{n} b_i v_i = \sum_{i=1}^{n} (a_i + b_i) v_i, \qquad cv = c \sum_{i=1}^{n} b_i v_i = \sum_{i=1}^{n} (c b_i) v_i.$$

By Theorem 1.5.1 these are the only expressions of $u + v$ and cv in $\mathcal{B}$, so their coefficients are the coordinates: $[u + v]_{\mathcal{B}} = (a_i + b_i)_i$ and $[cv]_{\mathcal{B}} = (c b_i)_i$, as claimed.

For the last statement, let $(a_1, \ldots, a_n) \in F^n$ and $w = a_1 v_1 + \cdots + a_n v_n$. By Definition 1.5.2, $[w]_{\mathcal{B}} = (a_1, \ldots, a_n)$. If also $[w']_{\mathcal{B}} = (a_1, \ldots, a_n)$, then by definition $w' = a_1 v_1 + \cdots + a_n v_n = w$.

■

So a basis of length n sets up a bijection $v \mapsto [v]_{\mathcal{B}}$ between V and F^n that respects addition and scaling. Any question about sums and multiples in V can be answered in F^n instead. In Chapter 2 we will call such a correspondence an **isomorphism**.

1.5.4 Every finite spanning list contains a basis

Where do bases come from? A finite-dimensional space is, by Definition 1.3.4, a space spanned by some finite list. That list reaches everything but may contain waste, so the natural plan is to throw the waste away.

Care is needed about **what** to throw away. In $((1, 0), (0, 1), (1, 1))$ each vector is a combination of the other two. Any single one can go, but not two of them. So redundancy is not a property of one vector alone: it depends on which other vectors stay. The fix is to decide one vector at a time, reading from left to right, and to compare each vector only with the vectors already kept.

First, a small lemma that tells us when a kept list stays independent.

Lemma 1.5.3. Appending a Vector to an Independent List

Let $(u_1, \dots, u_k)$ be a linearly independent list in a vector space V, and let $v \in V$. Then $(u_1, \dots, u_k, v)$ is linearly independent **if and only if** $v \notin \operatorname{span}(u_1, \dots, u_k)$.

Proof. ($\Rightarrow$) Suppose $v \in \operatorname{span}(u_1, \dots, u_k)$, say $v = c_1u_1 + \cdots + c_ku_k$. Then $c_1u_1 + \cdots + c_ku_k + (-1)v = 0$, with the coefficient $-1 \neq 0$. So the longer list is dependent. This proves ($\Rightarrow$) by contraposition.

($\Leftarrow$) Suppose $v \notin \operatorname{span}(u_1, \dots, u_k)$, and let $a_1u_1 + \cdots + a_ku_k + bv = 0$. If $b \neq 0$, then b^{-1} exists and

$$v = -b^{-1}a_1u_1 - \cdots - b^{-1}a_ku_k \in \operatorname{span}(u_1, \dots, u_k),$$

a contradiction. Hence $b = 0$. Then $a_1u_1 + \cdots + a_ku_k = 0$, and the independence of $(u_1, \dots, u_k)$ gives $a_1 = \cdots = a_k = 0$. So all coefficients are zero, and the longer list is independent.

■

The direction ($\Leftarrow$) is Exercise (C2: Adding a Vector outside the Span). When $k = 0$, the lemma says that the one-vector list (v) is independent exactly when $v \notin \operatorname{span}() = \{0\}$, that is, when $v \neq 0$, in agreement with Example (Degenerate Lists).

Now the procedure. Given a list $(v_1, \dots, v_m)$ in V, **sifting** it means the following.

Algorithm. Start with the empty list $S_0 = ()$. For $r = 1, 2, \dots, m$ in turn:

- if $v_r \notin \operatorname{span}(S_{r-1})$, **keep** it: $S_r = (S_{r-1}, v_r)$;
- if $v_r \in \operatorname{span}(S_{r-1})$, **discard** it: $S_r = S_{r-1}$.

The final list S_m is the **sifted list** of $(v_1, \dots, v_m)$.

Here (S_{r-1}, v_r) means the list S_{r-1} with v_r appended at the end. Each S_r is a **sub-list** of $(v_1, \dots, v_m)$: it is obtained by deleting some entries and keeping the rest in their original order.

Theorem 1.5.4. Sifting Theorem

Let $(v_1, \dots, v_m)$ be a list in a vector space V, and let S_m be its sifted list. Then S_m is linearly independent and

$$\operatorname{span}(S_m) = \operatorname{span}(v_1, \dots, v_m).$$

In particular, every finite list that spans V contains a sub-list that is a basis of V.

Idea. Sifting protects two things at every step. A vector is kept only when it lies outside the span of what was kept before, so by Lemma 1.5.3 the kept list stays independent. A vector is discarded only when it is already in that span, so by Theorem 1.3.5 discarding it loses nothing. We turn "at every step" into an induction on r, the number of vectors read so far, with the invariant: S_r is independent and has the same span as $v_1, \dots, v_r$.

Proof. We prove by induction on $r \in \{0, 1, \dots, m\}$ the statement

$$P(r)\text{: the list } S_r \text{ is linearly independent and } \operatorname{span}(S_r) = \operatorname{span}(v_1, \dots, v_r).$$

Base case. S_0 is the empty list, which is linearly independent, and both sides of the span equation are the span of the empty list. So $P(0)$ holds.

Inductive step. Let $1 \le r \le m$ and suppose $P(r-1)$ holds. Put $W = \operatorname{span}(S_{r-1})$, so that $W = \operatorname{span}(v_1, \dots, v_{r-1})$ by $P(r-1)$.

Case 1: $v_r \in W$. Then $S_r = S_{r-1}$, which is independent. Since $v_r \in \operatorname{span}(v_1, \dots, v_{r-1})$, Theorem 1.3.5 gives $\operatorname{span}(v_1, \dots, v_r) = \operatorname{span}(v_1, \dots, v_{r-1}) = W = \operatorname{span}(S_r)$.

Case 2: $v_r \notin W$. Then $S_r = (S_{r-1}, v_r)$. Since S_{r-1} is independent and $v_r \notin \operatorname{span}(S_{r-1})$, Lemma 1.5.3 shows that S_r is independent. For the spans, recall from Theorem 1.3.1 that the span of a list is a subspace, and is contained in every subspace containing the entries of the list. Every entry of S_r is one of $v_1, \dots, v_r$, so $\operatorname{span}(S_r) \subseteq \operatorname{span}(v_1, \dots, v_r)$. Conversely, $\operatorname{span}(S_r)$ is a subspace containing the entries of S_{r-1}, so it contains W, and with it $v_1, \dots, v_{r-1}$; it also contains v_r. Hence $\operatorname{span}(v_1, \dots, v_r) \subseteq \operatorname{span}(S_r)$.

In both cases $P(r)$ holds. By induction $P(m)$ holds, which is the first statement.

Finally, suppose $(v_1, \dots, v_m)$ spans V. Then S_m is a sub-list of it that is independent and satisfies $\operatorname{span}(S_m) = V$. So S_m is a basis of V.

■

Alternatively, one can delete instead of keep. As long as the list is dependent, the Linear Dependence Lemma (Theorem 1.4.1) locates an entry lying in the span of the entries before it, and deleting that entry does not change the span. Each deletion shortens the list, so the process stops after at most m steps. Sifting performs the same idea in a single left-to-right pass.

The decision "is v_r in the span of the kept vectors?" is a small system of equations. For now we solve these systems by hand; Chapter 3 gives a systematic method.

Example. Sifting a List in $\mathbb{R}^3$

Sift the list $(\mathbf{v}_1, \dots, \mathbf{v}_6)$ in $\mathbb{R}^3$, where

$$\mathbf{v}_1 = (1,1,1), \quad \mathbf{v}_2 = (2,2,2), \quad \mathbf{v}_3 = (1,0,0),$$
$$\mathbf{v}_4 = (3,2,2), \quad \mathbf{v}_5 = (1,1,0), \quad \mathbf{v}_6 = (0,0,1).$$

Hence find a basis of $\mathbb{R}^3$ inside the list.

Solution. We go through the list once, comparing each vector with the vectors kept so far.

- $\mathbf{v}_1$: the kept list is empty, with span $\{\mathbf{0}\}$. Since $\mathbf{v}_1 \neq \mathbf{0}$, **keep** $\mathbf{v}_1$.
- $\mathbf{v}_2$: $\mathbf{v}_2 = 2\mathbf{v}_1 \in \operatorname{span}(\mathbf{v}_1)$. **Discard.**
- $\mathbf{v}_3$: if $(1,0,0) = a(1,1,1) = (a,a,a)$, the first entry gives $a = 1$ and the second gives $a = 0$, a contradiction. So $\mathbf{v}_3 \notin \operatorname{span}(\mathbf{v}_1)$. **Keep.**
- $\mathbf{v}_4$: $2\mathbf{v}_1 + \mathbf{v}_3 = (2,2,2) + (1,0,0) = (3,2,2) = \mathbf{v}_4$, so $\mathbf{v}_4 \in \operatorname{span}(\mathbf{v}_1, \mathbf{v}_3)$. **Discard.**
- $\mathbf{v}_5$: suppose $(1,1,0) = a(1,1,1) + b(1,0,0) = (a+b, a, a)$. The second entry gives $a = 1$

and the third gives $a = 0$, a contradiction. So $\mathbf{v}_5 \notin \operatorname{span}(\mathbf{v}_1, \mathbf{v}_3)$. **Keep.**
- $\mathbf{v}_6$: $\mathbf{v}_1 - \mathbf{v}_5 = (1,1,1) - (1,1,0) = (0,0,1) = \mathbf{v}_6$, so $\mathbf{v}_6 \in \operatorname{span}(\mathbf{v}_1, \mathbf{v}_3, \mathbf{v}_5)$. **Discard.**

The sifted list is $(\mathbf{v}_1, \mathbf{v}_3, \mathbf{v}_5) = ((1,1,1), (1,0,0), (1,1,0))$. By Theorem 1.5.4 it is independent and has the same span as the original list. That span is all of $\mathbb{R}^3$, since it contains $\mathbf{e}_1 = \mathbf{v}_3$, $\mathbf{e}_2 = \mathbf{v}_5 - \mathbf{v}_3$, and $\mathbf{e}_3 = \mathbf{v}_6$, and hence every combination of them. So $(\mathbf{v}_1, \mathbf{v}_3, \mathbf{v}_5)$ is a basis of $\mathbb{R}^3$.

Sifting has no memory of the future, so its output depends on the order of the input.

Example. The Order Matters

Sift $((1,0),(0,1),(1,1))$ and $((1,1),(1,0),(0,1))$ in $\mathbb{R}^2$.

Solution. For the first list: $(1,0) \neq \mathbf{0}$, so keep it. Every multiple of $(1,0)$ has second entry 0, so $(0,1) \notin \operatorname{span}((1,0))$; keep it. Then $(1,1) = (1,0) + (0,1)$; discard it. The result is $((1,0),(0,1))$.

For the second list: $(1,1) \neq \mathbf{0}$, so keep it. If $(1,0) = a(1,1) = (a,a)$, then $a = 1$ and $a = 0$, which is impossible; keep $(1,0)$. Then $(0,1) = (1,1) - (1,0)$; discard it. The result is $((1,1),(1,0))$.

Both results are bases of $\mathbb{R}^2$ (the lists span $\mathbb{R}^2$ because they contain $\mathbf{e}_1$ and $\mathbf{e}_2$), but they are different bases. A zero vector, by contrast, is discarded wherever it stands, because $\mathbf{0}$ lies in every span.

Quick check. Sift the list $(x, 2x, 1+x, 1)$ in $\mathbb{R}[x]_{\leq 1}$. Which vectors are kept?

Solution. Keep x, since $x \neq 0$. Discard $2x = 2 \cdot x$. Keep $1+x$: every multiple ax has constant term 0, but $1+x$ has constant term 1. Discard $1 = (1+x) - x$. The sifted list is $(x, 1+x)$.

The theorem answers the question we started with.

Corollary 1.5.5. Existence of Bases

Every finite-dimensional vector space has a basis.

Proof. Let V be finite-dimensional. By Definition 1.3.4, V is spanned by some finite list. By Theorem 1.5.4, that list contains a sub-list that is a basis of V.

■

For $V = \{0\}$ the proof still works: $\{0\}$ is spanned by the empty list, and sifting it returns the empty list, which is the basis found in Example (Three More Bases).

We now know that bases exist, and we can find one inside any finite spanning list. But a space has many bases. The next section proves the fact that makes them measurable: all bases of the same space have the same length.

1.5.5 Exercises

1.5.5.1 A. Check your understanding

Exercise. A1

(a) Define what it means for a list $(v_1, \ldots, v_n)$ to be a basis of a vector space V over F.

(b) True or false: every list that spans $\mathbb{R}^2$ is a basis of $\mathbb{R}^2$. Justify your answer.

(c) True or false: $((1,0),(0,1))$ and $((0,1),(1,0))$ are the same basis of $\mathbb{R}^2$. Justify your answer.

(d) What is a basis of the zero space $\{0\}$?

(e) When sifting a list, under what condition is the vector v_r discarded?

(f) True or false: if $\mathcal{B}$ is a basis of V and $[u]_\mathcal{B} = [v]_\mathcal{B}$, then $u = v$. Justify your answer.

Solution.

(a) See Definition 1.5.1: the list is linearly independent and spans V.

(b) False. $((1,0),(0,1),(1,1))$ spans $\mathbb{R}^2$, but it is dependent, since $(1,0)+(0,1)-(1,1) = \mathbf{0}$.

(c) False. A basis is a list, and the two lists have their entries in different orders. The difference is visible in coordinates: $(3,1)$ has coordinate vector $(3,1)$ in the first basis and $(1,3)$ in the second.

(d) The empty list $()$, by Example (Three More Bases) (c).

(e) When v_r lies in the span of the vectors kept before it, that is, $v_r \in \operatorname{span}(S_{r-1})$.

(f) True. If both coordinate vectors equal $(a_1, \ldots, a_n)$, then $u = a_1 v_1 + \cdots + a_n v_n = v$ by Definition 1.5.2, where $\mathcal{B} = (v_1, \ldots, v_n)$.

1.5.5.2 B. Practice

Exercise. B1: A Basis of $\mathbb{R}^3$ and Coordinates

Let $\mathcal{B} = ((1,1,0),(0,1,1),(1,0,1))$. Show that $\mathcal{B}$ is a basis of $\mathbb{R}^3$. Hence find $[\mathbf{v}]_\mathcal{B}$ for $\mathbf{v} = (2,3,5)$.

Solution. Let $(p,q,r) \in \mathbb{R}^3$. For scalars a, b, c,

$$a(1,1,0) + b(0,1,1) + c(1,0,1) = (a+c,\ a+b,\ b+c),$$

so this equals (p,q,r) if and only if $a+c = p$, $a+b = q$, and $b+c = r$. Adding the three equations gives $2(a+b+c) = p+q+r$, so $a+b+c = \frac{p+q+r}{2}$. Subtracting each equation from this one,

$$b = \frac{-p+q+r}{2}, \qquad c = \frac{p-q+r}{2}, \qquad a = \frac{p+q-r}{2}.$$

Conversely, these values satisfy the three equations: for instance $a + c = \frac{2p}{2} = p$, and the other two are checked the same way. So every vector of $\mathbb{R}^3$ has exactly one expression in $\mathcal{B}$, and $\mathcal{B}$ is a basis by Theorem 1.5.1.

For $(p, q, r) = (2, 3, 5)$: $a = \frac{2+3-5}{2} = 0$, $b = \frac{-2+3+5}{2} = 3$, $c = \frac{2-3+5}{2} = 2$. Check: $0(1, 1, 0) + 3(0, 1, 1) + 2(1, 0, 1) = (2, 3, 5)$. Hence $[\mathbf{v}]_{\mathcal{B}} = (0, 3, 2)$.

Exercise. B2: Sifting Polynomials

Sift the list $(1 + x,\ 2 + 2x,\ x^2,\ 1 + x + x^2,\ x,\ 1)$ in $\mathbb{R}[x]_{\leq 2}$, justifying each decision. Hence find a basis of $\mathbb{R}[x]_{\leq 2}$ contained in the list.

Solution. We read the list from left to right, comparing coefficients.

- $1 + x \neq 0$: **keep**.
- $2 + 2x = 2(1 + x)$: **discard**.
- x^2: every multiple $a(1 + x)$ has x^2-coefficient 0, while x^2 has x^2-coefficient 1. So $x^2 \notin \operatorname{span}(1 + x)$: **keep**.
- $1 + x + x^2 = (1 + x) + x^2$: **discard**.
- x: suppose $x = a(1 + x) + bx^2 = a + ax + bx^2$. The constant terms give $a = 0$ and the x-coefficients give $a = 1$, a contradiction. So $x \notin \operatorname{span}(1 + x, x^2)$: **keep**.
- $1 = (1 + x) - x$: **discard**.

The sifted list is $(1 + x, x^2, x)$. The original list contains 1, x, and x^2, so its span contains every combination of them, which is all of $\mathbb{R}[x]_{\leq 2}$ by Example (Standard Bases). Hence the list spans $\mathbb{R}[x]_{\leq 2}$, and by Theorem 1.5.4 the sifted list $(1 + x, x^2, x)$ is a basis of $\mathbb{R}[x]_{\leq 2}$.

Exercise. B3: Symmetric Matrices

Let $U = \{\mathbf{A} \in M_2(\mathbb{R}) : \mathbf{A}^\top = \mathbf{A}\}$. Show that $(\mathbf{E}_{11},\ \mathbf{E}_{12} + \mathbf{E}_{21},\ \mathbf{E}_{22})$ is a basis of U, and find the coordinate vector of $\begin{pmatrix} 2 & -1 \\ -1 & 5 \end{pmatrix}$ with respect to it.

Solution. Each of the three matrices is symmetric, so the list lies in U.

Spanning. Let $\mathbf{A} \in U$. Writing $\mathbf{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, the condition $\mathbf{A}^\top = \mathbf{A}$ says $c = b$. Hence

$$\mathbf{A} = \begin{pmatrix} a & b \\ b & d \end{pmatrix} = a\mathbf{E}_{11} + b(\mathbf{E}_{12} + \mathbf{E}_{21}) + d\mathbf{E}_{22}.$$

Independence. Let $a\mathbf{E}_{11} + b(\mathbf{E}_{12} + \mathbf{E}_{21}) + d\mathbf{E}_{22} = \mathbf{O}$. The left side is $\begin{pmatrix} a & b \\ b & d \end{pmatrix}$, so comparing entries with the zero matrix gives $a = b = d = 0$.

Hence the list is a basis of U. Reading off $a = 2$, $b = -1$, $d = 5$, the coordinate vector is $(2, -1, 5)$.

1.5.5.3 C. Going deeper

Exercise. C1: Polynomials Vanishing at 1

Let F be a field, $n \geq 1$, and $U = \{p \in F[x]_{\leq n} : p(1) = 0\}$.

(a) Prove that $(x - 1,\ x^2 - 1,\ \ldots,\ x^n - 1)$ is a basis of U.

(b) For $n = 3$, find the coordinate vector of $x^3 - 2x + 1$ with respect to this basis.

Hint: for $p = a_0 + a_1 x + \cdots + a_n x^n$, what does $p(1) = 0$ say about a_0?

Solution.

(a) For $p = a_0 + a_1 x + \cdots + a_n x^n$, evaluation gives $p(1) = a_0 + a_1 + \cdots + a_n$, since every power of 1 is 1. In particular each $x^k - 1$ lies in U, because $1 - 1 = 0$.
Spanning. Let $p \in U$. Then $a_0 = -(a_1 + \cdots + a_n)$, so

$$\begin{aligned} p &= a_1 x + \cdots + a_n x^n - (a_1 + \cdots + a_n) \\ &= a_1(x - 1) + a_2(x^2 - 1) + \cdots + a_n(x^n - 1). \end{aligned}$$

Independence. Let $c_1(x - 1) + \cdots + c_n(x^n - 1) = 0$. For $1 \leq k \leq n$, the coefficient of x^k on the left is c_k, since x^k appears in no other term. The zero polynomial has all coefficients 0, so $c_k = 0$ for every k.
Hence the list is a basis of U.

(b) For $p = x^3 - 2x + 1$ we have $p(1) = 0$, and $(a_1, a_2, a_3) = (-2, 0, 1)$. By the formula in (a), $p = -2(x - 1) + 0(x^2 - 1) + 1(x^3 - 1)$. Check: $-2x + 2 + x^3 - 1 = x^3 - 2x + 1$. The coordinate vector is $(-2, 0, 1)$.

Exercise. C2: The Same List over a Different Field

(a) Explain why $((1, 1), (1, -1))$ is **not** a basis of $\mathbb{F}_2^2$.

(b) Let F be any field. Prove that $((1, 1), (1, -1))$ is a basis of F^2 if and only if $1 + 1 \neq 0$ in F.

Solution.

(a) In $\mathbb{F}_2$ we have $-1 = 1$, so the list is $((1, 1), (1, 1))$. It is dependent: $1 \cdot (1, 1) + 1 \cdot (1, 1) = (1+1, 1+1) = (0, 0)$ with nonzero coefficients. (It also fails to span: every combination is (c, c), so $(1, 0)$ is out of reach.)

(b) Write $2 = 1 + 1 \in F$.

($\Leftarrow$) Suppose $2 \neq 0$, so 2^{-1} exists. Let $(x, y) \in F^2$. The equation $a(1, 1) + b(1, -1) = (x, y)$ says $a+b = x$ and $a-b = y$. If it holds, adding and subtracting gives $2a = x+y$ and $2b = x-y$, so $a = 2^{-1}(x + y)$ and $b = 2^{-1}(x - y)$. Conversely these values satisfy $a + b = 2^{-1} \cdot 2x = x$ and $a - b = 2^{-1} \cdot 2y = y$. So every vector has exactly one expression, and the list is a basis

by Theorem 1.5.1.

($\Rightarrow$) Suppose $2 = 0$. Then $1 = -1$, and the argument of (a) applies word for word: $1 \cdot (1,1) + 1 \cdot (1,-1) = (2,0) = (0,0)$, so the list is dependent and is not a basis. This proves ($\Rightarrow$) by contraposition.

Exercise. C3: Rescaling a Basis

Let $\mathcal{B} = (v_1, \ldots, v_n)$ be a basis of V over F, and let $c_1, \ldots, c_n \in F$ be **nonzero**.

(a) Prove that $\mathcal{B}' = (c_1 v_1, \ldots, c_n v_n)$ is a basis of V.

(b) Express $[v]_{\mathcal{B}'}$ in terms of $[v]_{\mathcal{B}}$.

(c) Explain why the conclusion of (a) can fail if some $c_i = 0$.

Solution.

(a) Let $v \in V$ and $[v]_{\mathcal{B}} = (a_1, \ldots, a_n)$. Since each $c_i \neq 0$, c_i^{-1} exists, and

$$v = \sum_{i=1}^{n} a_i v_i = \sum_{i=1}^{n} (a_i c_i^{-1})(c_i v_i).$$

So v has an expression in $\mathcal{B}'$. Suppose $v = \sum_i b_i (c_i v_i) = \sum_i (b_i c_i) v_i$. By uniqueness of the expression in $\mathcal{B}$ (Theorem 1.5.1), $b_i c_i = a_i$ for each i, so $b_i = a_i c_i^{-1}$. Hence the expression in $\mathcal{B}'$ is unique, and $\mathcal{B}'$ is a basis by Theorem 1.5.1.

(b) By (a), $[v]_{\mathcal{B}'} = (a_1 c_1^{-1}, \ldots, a_n c_n^{-1})$: divide the i-th coordinate by c_i.

(c) If $c_1 = 0$, then $\mathcal{B}'$ contains $0 v_1 = 0$, and $1 \cdot 0 + 0 \cdot c_2 v_2 + \cdots + 0 \cdot c_n v_n = 0$ is a combination with a nonzero coefficient. So $\mathcal{B}'$ is dependent and is not a basis.

1.6 The Exchange Theorem and Dimension

Every finite-dimensional space has a basis, and usually a great many of them. This section proves that they all have the same length, so that the length is a property of the space and not of the basis we happened to pick. That number is the dimension. The key step is the Steinitz Exchange Theorem, the most important proof of the chapter. With it we build the everyday tools of dimension counting: size bounds, "count instead of check", extension of independent lists to bases, and dimensions of subspaces.

1.6.1 How big is a vector space?

We would like to say that a plane is smaller than space, and that $M_2(\mathbb{R})$ and $\mathbb{R}[x]_{\leq 3}$ are, in some sense, the same size. The naive measure is the number of elements. It fails at once: $\mathbb{R}^2$, $M_2(\mathbb{R})$, and $\mathbb{R}[x]_{\leq 3}$ are all infinite sets, and counting elements cannot tell them apart.

A better measure counts **how many vectors it takes to build everything, without waste**, which is the length of a basis. The standard bases give

$$\mathbb{R}^2 : 2, \qquad M_2(\mathbb{R}) : 4 \ \ (\mathbf{E}_{11}, \mathbf{E}_{12}, \mathbf{E}_{21}, \mathbf{E}_{22}), \qquad \mathbb{R}[x]_{\leq 3} : 4 \ \ (1, x, x^2, x^3).$$

This matches intuition: a matrix $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ and a polynomial $a+bx+cx^2+dx^3$ are both described by four free numbers.

But there is a gap. A space has many bases. We counted the standard basis of $\mathbb{R}^2$; the basis $((1,1),(1,-1))$ also has length 2, but what guarantees that no basis of $\mathbb{R}^2$ has length 3? If some space had bases of two different lengths, "the length of a basis" would be meaningless. So before we can define anything, we need the following.

Goal. Any two bases of a finite-dimensional vector space have the same length.

A basis is independent **and** spanning, and the goal follows from a single inequality about the two properties separately: *an independent list is never longer than a spanning list.* Granting this, take two bases $\mathcal{B}$ and $\mathcal{C}$. Since $\mathcal{B}$ is independent and $\mathcal{C}$ spans, $\mathcal{B}$ is not longer than $\mathcal{C}$. Swapping the roles, $\mathcal{C}$ is not longer than $\mathcal{B}$. So they have the same length. Everything rests on the inequality.

1.6.2 The Steinitz Exchange Theorem

Why should an independent list be no longer than a spanning list? In $\mathbb{R}^2$ the spanning list $(\mathbf{e}_1, \mathbf{e}_2)$ has length 2, and three vectors in the plane always satisfy a linear relation, but "the plane is two-dimensional" is what we are trying to define, so we cannot use it. The idea is instead to feed the independent vectors into the spanning list one at a time, pushing out one spanning vector each time, while keeping the list spanning. If the independent vectors outnumbered the spanning ones, we would run out of vectors to push out.

Theorem 1.6.1. Steinitz Exchange Theorem

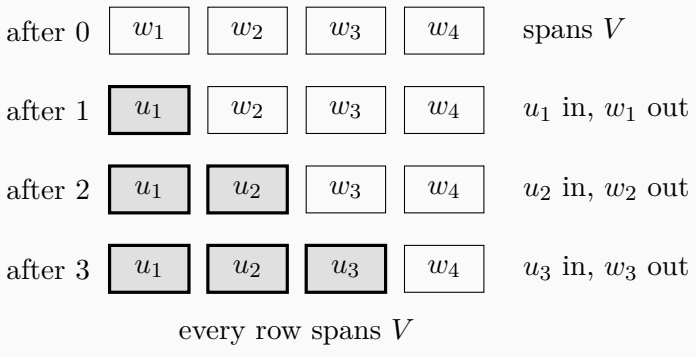
Let V be a vector space over F. Suppose that $(u_1, \ldots, u_m)$ is a **linearly independent** list in V and that $(w_1, \ldots, w_n)$ is a list that **spans** V. Then

(a) $m \leq n$, and

(b) after renumbering $w_1, \ldots, w_n$ suitably, the list $(u_1, \ldots, u_m, w_{m+1}, \ldots, w_n)$ spans V.

In words: an independent list is at most as long as a spanning list, and the independent vectors can **replace** m of the spanning vectors without losing the spanning property. "Renumbering" means listing the w's in a different order; which w's get replaced depends on the u's.

Idea. Picture a row of n slots holding $w_1, \ldots, w_n$; the row spans V. We push u_1 into the row and push one w out, then u_2, and so on, keeping the row spanning after every exchange. Here is the picture for $m = 3$, $n = 4$, with the w's numbered in the order they leave.



Which w may leave? Write the incoming u_{k+1} as a combination of the current row. Any w with a **nonzero** coefficient can be solved for, so it lies in the span of the other entries together with u_{k+1}, and removing it costs nothing. Such a w must exist: otherwise u_{k+1} would be a combination of $u_1, \dots, u_k$ alone, which independence forbids. In particular, while u's remain to be inserted, a w remains to be removed, and that is why $m \le n$.

Step roadmap. Induction on the number k of exchanges, with the invariant

> $P(k)$: $k \le n$, and after renumbering the w's, the list $(u_1, \dots, u_k, w_{k+1}, \dots, w_n)$ spans V.

Step 1 is the base case $k = 0$. **Step 2** writes u_{k+1} in the current list. **Step 3** is the Claim that some remaining w has a nonzero coefficient. **Step 4** exchanges that w for u_{k+1} and checks the span.

Proof. We prove $P(k)$ for $k = 0, 1, \dots, m$ by induction on k. The case $k = m$ is the theorem.

Step 1 (base case). For $k = 0$ the list in $P(0)$ is $(w_1, \dots, w_n)$, which spans V by hypothesis, and $0 \le n$. So $P(0)$ holds.

Step 2. Let $0 \le k < m$ and suppose $P(k)$ holds. Renumber the w's as $P(k)$ allows, so that

$$L_k = (u_1, \dots, u_k, w_{k+1}, \dots, w_n)$$

spans V. Since $k < m$, the vector u_{k+1} exists, and $u_{k+1} \in V = \operatorname{span}(L_k)$. Hence there are scalars $a_1, \dots, a_k, b_{k+1}, \dots, b_n \in F$ with

$$u_{k+1} = a_1 u_1 + \cdots + a_k u_k + b_{k+1} w_{k+1} + \cdots + b_n w_n. \qquad (*)$$

(If $k = n$, the list L_k contains no w's and the second group of terms is empty.)

Step 3.

Claim. We have $k + 1 \le n$, and $b_j \neq 0$ for some j with $k + 1 \le j \le n$.

Proof. Suppose not. Then either $k = n$, so that there are no b's at all, or $k < n$ and $b_{k+1} = \cdots = b_n = 0$. In both cases $(*)$ reads $u_{k+1} = a_1 u_1 + \cdots + a_k u_k$. Therefore

$$a_1 u_1 + \cdots + a_k u_k + (-1) u_{k+1} + 0 u_{k+2} + \cdots + 0 u_m = 0$$

is a linear relation among $u_1, \dots, u_m$ in which the coefficient of u_{k+1} is $-1 \neq 0$. This contradicts the linear independence of $(u_1, \dots, u_m)$. Hence there is at least one w in L_k, so $k + 1 \le n$, and some b_j with $j \ge k + 1$ is nonzero.

■

Step 4. Renumber $w_{k+1}, \dots, w_n$ among themselves so that $b_{k+1} \neq 0$. This only changes the order of entries of L_k, so L_k still spans V (the span of a list does not depend on the order of its entries), and $(*)$ still holds with the coefficients renumbered in the same way. Put

$$L_{k+1} = (u_1, \dots, u_{k+1}, w_{k+2}, \dots, w_n).$$

Since $b_{k+1} \neq 0$, the scalar b_{k+1}^{-1} exists, and solving $(*)$ for w_{k+1} gives

$$w_{k+1} = b_{k+1}^{-1}(u_{k+1} - a_1u_1 - \cdots - a_ku_k - b_{k+2}w_{k+2} - \cdots - b_nw_n).$$

The right side is a combination of the entries of L_{k+1}, so $w_{k+1} \in \operatorname{span}(L_{k+1})$. By Theorem 1.3.5, inserting w_{k+1} into L_{k+1} does not change the span:

$$\operatorname{span}(L_{k+1}) = \operatorname{span}(u_1, \ldots, u_{k+1}, w_{k+1}, w_{k+2}, \ldots, w_n).$$

The list on the right contains every entry of L_k, so its span is a subspace containing all entries of L_k, and hence contains $\operatorname{span}(L_k) = V$ by Theorem 1.3.1. Therefore $\operatorname{span}(L_{k+1}) = V$. Together with $k+1 \leq n$ from the Claim, this is $P(k+1)$.

By induction, $P(m)$ holds: $m \leq n$, and after renumbering the w's, $(u_1, \ldots, u_m, w_{m+1}, \ldots, w_n)$ spans V. This proves the theorem.

■

Part (a) is the inequality our goal needed. Part (b), the exchange itself, is a bonus that we will spend later in this section, to extend independent lists to bases.

Warning. The $\mathbf{w}$ that leaves must have a **nonzero** coefficient; an arbitrary $\mathbf{w}$ will not do. In $\mathbb{R}^2$, take the spanning list $(\mathbf{w}_1, \mathbf{w}_2) = ((1,0),(0,1))$ and the independent vector $\mathbf{u}_1 = (1,0) = 1\mathbf{w}_1 + 0\mathbf{w}_2$. Exchanging $\mathbf{u}_1$ for $\mathbf{w}_1$ gives $((1,0),(0,1))$, which spans. Exchanging $\mathbf{u}_1$ for $\mathbf{w}_2$, whose coefficient is 0, gives $((1,0),(1,0))$, which does not reach $(0,1)$.

The theorem already settles questions that would otherwise need a computation.

Quick check. Can the four polynomials $1+x$, $x+x^2$, x^2+1, $1+x+x^2$ be linearly independent in $\mathbb{R}[x]_{\leq 2}$?

Solution. No. The list $(1, x, x^2)$ spans $\mathbb{R}[x]_{\leq 2}$ and has length 3, so by Theorem 1.6.1 every independent list there has length at most $3 < 4$. Indeed $(1+x) + (x+x^2) + (x^2+1) - 2(1+x+x^2) = 0$.

1.6.3 Dimension

The goal of the section is now one line away.

Theorem 1.6.2. Basis Theorem

Let V be a finite-dimensional vector space. Then V has a basis, and any two bases of V have the same length.

Proof. By Corollary 1.5.5, V has a basis. Let $(v_1, \ldots, v_n)$ and $(w_1, \ldots, w_m)$ be bases of V. The first is linearly independent and the second spans V, so $n \leq m$ by Theorem 1.6.1.

Swapping the roles of the two bases gives $m \le n$. Hence $m = n$.

■

We started from a question: how big is a vector space? The Basis Theorem says that one number, the length of a basis, answers it without depending on any choice. So we give that number a name.

The dimension of a space is the number of vectors in any basis: how many independent directions the space has.

Definition 1.6.1. Dimension

Let V be a finite-dimensional vector space over F. The **dimension** of V over F, written $\dim_F V$ or simply $\dim V$, is the length of **any** basis of V.

In words: pick a basis of V, any basis at all, and count its vectors. The dimension is defined only for finite-dimensional spaces (Definition 1.3.4); a space spanned by no finite list is infinite-dimensional, and we do not assign it a number here. The subscript F records the field of scalars, which, as we will see, matters.

Well-definedness. Two things could go wrong with "the length of any basis": there might be no basis to count, or different bases might give different counts. The first is ruled out by Corollary 1.5.5 and the second by Theorem 1.6.2. That is exactly what the word **any** in the definition needs.

Example. Dimensions of the Running Examples

Find the dimension of each space.

(a) F^n, $M_{m\times n}(F)$, and $F[x]_{\le n}$.
(b) The zero space $\{0\}$.
(c) $\mathbb{C}$ over $\mathbb{C}$, and $\mathbb{C}$ over $\mathbb{R}$.
(d) $U = \{\mathbf{A} \in M_2(\mathbb{R}) : \mathbf{A}^\top = \mathbf{A}\}$, the symmetric 2×2 real matrices.
(e) $\mathbb{Q}(\sqrt{2}) = \{a + b\sqrt{2} : a, b \in \mathbb{Q}\}$ over $\mathbb{Q}$ (Example ($\mathbb{Q}(\sqrt{2})$ over $\mathbb{Q}$)).

Solution. In each case we exhibit one basis and count it.

(a) The standard bases of Example (Standard Bases) have lengths n, mn, and $n+1$. So $\dim F^n = n$, $\dim M_{m\times n}(F) = mn$, and $\dim F[x]_{\le n} = n+1$.
(b) The empty list is a basis of $\{0\}$ (Example (Three More Bases)), so $\dim\{0\} = 0$.
(c) Over $\mathbb{C}$, the list (1) is a basis of $\mathbb{C}$: every z equals $z \cdot 1$, and $z \cdot 1 = 0$ forces $z = 0$. So $\dim_{\mathbb{C}} \mathbb{C} = 1$. Over $\mathbb{R}$, the list $(1, i)$ is a basis (Example (Three More Bases)), so $\dim_{\mathbb{R}} \mathbb{C} = 2$.
(d) A symmetric matrix has the form $\begin{pmatrix} a & b \\ b & d \end{pmatrix} = a\mathbf{E}_{11} + b(\mathbf{E}_{12} + \mathbf{E}_{21}) + d\mathbf{E}_{22}$, and this combination is the zero matrix only when $a = b = d = 0$. So $(\mathbf{E}_{11}, \mathbf{E}_{12} + \mathbf{E}_{21}, \mathbf{E}_{22})$ is a basis of U, as in Exercise (B3: Symmetric Matrices), and $\dim U = 3$.
(e) Every element is $a \cdot 1 + b\sqrt{2}$ with $a, b \in \mathbb{Q}$, so $(1, \sqrt{2})$ spans. Let $a + b\sqrt{2} = 0$ with

$a, b \in \mathbb{Q}$. If $b \neq 0$, then $\sqrt{2} = -a/b \in \mathbb{Q}$, contradicting Theorem 0.2.3. So $b = 0$, and then $a = 0$. Hence $(1, \sqrt{2})$ is a basis, and $\dim_{\mathbb{Q}} \mathbb{Q}(\sqrt{2}) = 2$.

The zero space in (b) is the degenerate case: it has dimension 0 because its basis is empty, not because it has no elements. It has one element, 0.

Nonexample by minimal change. Remove the degree bound from $F[x]_{\leq n}$. The space $F[x]$ is not spanned by any finite list (Theorem 1.3.3), so it has no finite basis and no dimension in the sense of Definition 1.6.1. The hypothesis "finite-dimensional" in the Basis Theorem is exactly what fails.

Why this definition. Why a basis, and not some other list? Spanning lists can be made as long as we like by repeating vectors, and independent lists can be made as short as we like by deleting vectors, so neither length alone is an invariant. A basis sits exactly where the two meet: it is the longest possible independent list and the shortest possible spanning list, as we prove next. And why not count elements? Over $\mathbb{R}$ every nonzero space is infinite, so the count says nothing. Over a finite field the count does carry information, but it is determined by the dimension; see Exercise (C3: Counting Elements over a Finite Field).

Warning. The dimension of $F[x]_{\leq n}$ is $n+1$, **not** n. The basis $(1, x, \ldots, x^n)$ has one vector for each exponent $0, 1, \ldots, n$, and the constant 1 is easy to forget. For instance $\dim \mathbb{R}[x]_{\leq 2} = 3$.

Warning. Dimension depends on the field. The set $\mathbb{C}$ has dimension 1 over $\mathbb{C}$ but dimension 2 over $\mathbb{R}$, because over $\mathbb{R}$ the scalar i is no longer available and i must be counted as a separate basis vector. When more than one field is in sight, write $\dim_F$.

Quick check. What are $\dim M_{2\times 3}(F)$ and $\dim F[x]_{\leq 5}$? Which of them equals $\dim F^6$?

Solution. $\dim M_{2\times 3}(F) = 2 \cdot 3 = 6$ and $\dim F[x]_{\leq 5} = 5 + 1 = 6$. Both equal $\dim F^6 = 6$.

1.6.4 Size bounds

The first payoff turns the Exchange Theorem into statements about the number $\dim V$.

Theorem 1.6.3. Size Bounds

Let V be a vector space over F with $\dim V = n$.

(a) Every linearly independent list in V has length at most n. Hence every list in V of length **greater than** n is linearly dependent.

(b) Every list that spans V has length at least n. Hence no list of length **less than** n spans V.

Proof. Let $\mathcal{B}$ be a basis of V; it has length n.

(a) Let $\mathcal{L}$ be an independent list of length ℓ. Since $\mathcal{B}$ spans V, Theorem 1.6.1 gives $\ell \leq n$. The second sentence is the contrapositive.

(b) Let $\mathcal{L}$ be a list of length ℓ that spans V. Since $\mathcal{B}$ is independent, Theorem 1.6.1 gives $n \leq \ell$. The second sentence is the contrapositive.

■

So in a space of dimension n,

$$\text{length of an independent list} \;\leq\; n \;\leq\; \text{length of a spanning list.}$$

This settles many questions with no computation. The four vectors $(1,2,3)$, $(4,5,6)$, $(7,8,9)$, $(1,0,0)$ in $\mathbb{R}^3$ are dependent, because $4 > 3$. Three matrices never span $M_2(F)$, because $3 < 4$.

1.6.5 Count instead of check

To show a list is a basis we normally check two things, independence and spanning. When the length of the list equals the dimension, one check is enough. This is one of the most used facts in the book.

Theorem 1.6.4. Count Instead of Check

Let V be a vector space with $\dim V = n$, and let $\mathcal{L} = (v_1, \ldots, v_n)$ be a list in V of length exactly n.

(a) If $\mathcal{L}$ is linearly independent, then $\mathcal{L}$ is a basis of V.

(b) If $\mathcal{L}$ spans V, then $\mathcal{L}$ is a basis of V.

Idea. For (a), the exchange part of Theorem 1.6.1 does the work: we can push all n vectors of $\mathcal{L}$ into a basis of length n, and then no old vectors are left, so $\mathcal{L}$ alone spans. For (b), sift $\mathcal{L}$: the result is a basis, so it has length n, so nothing was discarded.

Proof.

(a) Let $(w_1, \ldots, w_n)$ be a basis of V; in particular it spans V. Since $\mathcal{L}$ is independent of length n, Theorem 1.6.1 (b) with $m = n$ says that, after renumbering the w's, the list $(v_1, \ldots, v_n, w_{n+1}, \ldots, w_n)$ spans V. This list contains no w's: it is $\mathcal{L}$. So $\mathcal{L}$ is independent and spans V, and it is a basis.

(b) By Theorem 1.5.4, $\mathcal{L}$ contains a sub-list that is a basis of V. By Theorem 1.6.2, that sub-list has length $\dim V = n$. The only sub-list of length n of a list of length n is the list itself. Hence $\mathcal{L}$ is a basis.

■

The theorem lets us choose the easier check, and independence is almost always easier: it is a single equation equal to 0, while spanning asks us to solve for an arbitrary vector.

Example. A Basis by Counting

Show that $(1,\ 1+x,\ (1+x)^2)$ is a basis of $\mathbb{R}[x]_{\leq 2}$.

Solution. By Example (Dimensions of the Running Examples), $\dim \mathbb{R}[x]_{\leq 2} = 3$, and the list has length 3. So by Theorem 1.6.4 (a) it suffices to check independence. Let

$$a \cdot 1 + b(1+x) + c(1+x)^2 = 0.$$

Expanding, the left side is $(a+b+c) + (b+2c)x + cx^2$. Comparing coefficients with the zero polynomial: $c = 0$; then $b + 2c = 0$ gives $b = 0$; then $a + b + c = 0$ gives $a = 0$. Hence the list is independent, with $3 = \dim \mathbb{R}[x]_{\leq 2}$ elements, and therefore it is a basis.

We never had to show that every polynomial of degree at most 2 is a combination of $1, 1+x, (1+x)^2$. The dimension count did it for us.

1.6.6 Extending to a basis

Sifting finds a basis **inside** a spanning list. The opposite direction starts from an independent list and grows it into a basis. This is the tool behind every dimension formula later in the book: start from a basis of the smallest space, then extend.

Theorem 1.6.5. Basis Extension Theorem

Let V be a finite-dimensional vector space with $\dim V = n$, and let $(u_1, \ldots, u_m)$ be a linearly independent list in V. Then $m \leq n$, and there are vectors $v_{m+1}, \ldots, v_n \in V$ such that

$$(u_1, \ldots, u_m, v_{m+1}, \ldots, v_n)$$

is a basis of V. Moreover, the v's can be chosen from any given basis of V.

Proof. Let $(w_1, \ldots, w_n)$ be a basis of V. It spans V, so by Theorem 1.6.1, $m \leq n$, and after renumbering the w's the list $(u_1, \ldots, u_m, w_{m+1}, \ldots, w_n)$ spans V. This list has length $m + (n-m) = n = \dim V$, so it is a basis by Theorem 1.6.4 (b). Take $v_j = w_j$ for $m < j \leq n$. ∎

Alternatively, sift the list $(u_1, \ldots, u_m, w_1, \ldots, w_n)$. It spans V, and sifting keeps every u_k: the sub-list $(u_1, \ldots, u_k)$ is independent (Exercise (B3: Sub-Lists)), so $u_k \notin \operatorname{span}(u_1, \ldots, u_{k-1})$ by Lemma 1.5.3. Sifting has the advantage of telling us **which** w's to add.

Example. Extending an Independent List

Extend the list $((1,1,0),(0,1,1))$ to a basis of $\mathbb{R}^3$.

Solution. The list is independent: $a(1,1,0) + b(0,1,1) = (a, a+b, b) = \mathbf{0}$ forces $a = b = 0$.

By Theorem 1.6.5, one vector from the standard basis will do, so we try $\mathbf{e}_1$. Let

$$a(1,1,0) + b(0,1,1) + c(1,0,0) = (a+c,\ a+b,\ b) = (0,0,0).$$

Then $b = 0$, so $a = 0$ from the second entry, so $c = 0$ from the first. Hence the extended list is independent, with $3 = \dim \mathbb{R}^3$ elements, and therefore a basis by Theorem 1.6.4 (a).

The extension is far from unique. The same computation with $\mathbf{e}_2$ gives $(a, a+b+c, b) = \mathbf{0}$, and with $\mathbf{e}_3$ gives $(a, a+b, b+c) = \mathbf{0}$; both force $a = b = c = 0$. So any one of $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$ completes this list to a basis.

1.6.7 Largest independent lists and smallest spanning lists

We claimed that a basis is where independence and spanning meet. Here is the precise statement. It does not need finite dimension, and it is the form of "basis" that generalizes to infinite-dimensional spaces later in this chapter.

Call an independent list $(v_1, \dots, v_n)$ in V **maximal independent** if for **every** $v \in V$ the longer list $(v_1, \dots, v_n, v)$ is dependent. Call a spanning list **minimal spanning** if deleting **any one** of its entries gives a list that does not span V.

Theorem 1.6.6. Maximal Independent and Minimal Spanning Lists

Let V be a vector space and $\mathcal{L} = (v_1, \dots, v_n)$ a list in V.

(a) $\mathcal{L}$ is a basis of V if and only if $\mathcal{L}$ is maximal independent.
(b) $\mathcal{L}$ is a basis of V if and only if $\mathcal{L}$ is minimal spanning.

Proof.

(a) ($\Rightarrow$) Let $\mathcal{L}$ be a basis and $v \in V$. Then $v \in \operatorname{span}(\mathcal{L}) = V$, so $(v_1, \dots, v_n, v)$ is dependent by Lemma 1.5.3. Hence $\mathcal{L}$ is maximal independent.
($\Leftarrow$) Let $\mathcal{L}$ be maximal independent and $v \in V$. The list $(v_1, \dots, v_n, v)$ is dependent while $\mathcal{L}$ is independent, so $v \in \operatorname{span}(\mathcal{L})$ by Lemma 1.5.3. Hence $\mathcal{L}$ spans V, and $\mathcal{L}$ is a basis.

(b) ($\Rightarrow$) Let $\mathcal{L}$ be a basis, and suppose that deleting v_j gives a list that still spans V. Then $v_j = \sum_{i \neq j} c_i v_i$ for some scalars, so $\sum_{i \neq j} c_i v_i + (-1) v_j = 0$, a linear relation with the coefficient $-1 \neq 0$. This contradicts independence. Hence $\mathcal{L}$ is minimal spanning.

($\Leftarrow$) Let $\mathcal{L}$ be minimal spanning. By Theorem 1.5.4, some sub-list S of $\mathcal{L}$ is a basis of V. Suppose $S \neq \mathcal{L}$. Then S omits some entry v_j, so every entry of S is an entry of the list $\mathcal{L}'$ obtained by deleting v_j. By Theorem 1.3.1, $\operatorname{span}(\mathcal{L}')$ contains $\operatorname{span}(S) = V$, so $\mathcal{L}'$ spans V, contradicting minimality. Hence $S = \mathcal{L}$, and $\mathcal{L}$ is a basis.

■

"Maximal" means "cannot be enlarged", which is different from "longest possible". In a finite-dimensional space the two agree: by Theorem 1.6.3 a longest independent list has length $\dim V$,

and by Theorem 1.6.4 every independent list of that length is a basis.

1.6.8 Dimensions of subspaces

A subspace should be no bigger than the space containing it. Proving this needs care, because of a tempting circular argument.

Theorem 1.6.7. Dimension of a Subspace

Let V be a finite-dimensional vector space and U a subspace of V. Then U is finite-dimensional, and $\dim U \leq \dim V$.

Idea. The tempting proof is: "take a basis of U; it is independent in V, so its length is at most $\dim V$". But we do not yet know that U **has** a basis. Corollary 1.5.5 needs a finite spanning list of U, and that is exactly what we are trying to prove. So we build from below instead. Independent lists inside U are independent in V, so their lengths are bounded by $\dim V$. Take one of the largest possible length. If it did not span U, we could append a vector of U outside its span and get a longer one.

Proof. Let $n = \dim V$. Independence of a list depends only on its vectors and the operations of V, which U shares, so every independent list in U is an independent list in V, and has length at most n by Theorem 1.6.3. The empty list is independent in U. Hence the set of lengths of independent lists in U is a nonempty set of integers in $\{0, 1, \ldots, n\}$, and it has a largest element $k \leq n$. Choose an independent list $(u_1, \ldots, u_k)$ in U of this length.

We claim $\operatorname{span}(u_1, \ldots, u_k) = U$. Since U is a subspace containing every u_i, Theorem 1.3.1 gives $\operatorname{span}(u_1, \ldots, u_k) \subseteq U$. Conversely, let $u \in U$, and suppose $u \notin \operatorname{span}(u_1, \ldots, u_k)$. By Lemma 1.5.3, $(u_1, \ldots, u_k, u)$ is an independent list in U of length $k + 1$, contradicting the maximality of k. Hence $u \in \operatorname{span}(u_1, \ldots, u_k)$.

Therefore $(u_1, \ldots, u_k)$ spans U, so U is finite-dimensional (Definition 1.3.4), and the list is a basis of U. Hence $\dim U = k \leq n = \dim V$.

■

When the dimensions are equal, there is no room left.

Theorem 1.6.8. Equal Dimension Forces Equality

Let V be a finite-dimensional vector space and U a subspace of V. If $\dim U = \dim V$, then $U = V$.

Proof. Let $n = \dim U = \dim V$; U is finite-dimensional by Theorem 1.6.7. Let $(u_1, \ldots, u_n)$ be a basis of U (Corollary 1.5.5). It is an independent list in V of length $n = \dim V$, so it is a basis of V by Theorem 1.6.4 (a). Hence $V = \operatorname{span}(u_1, \ldots, u_n) \subseteq U$, where the inclusion is Theorem 1.3.1 (3) applied to the subspace U. Since also $U \subseteq V$, we get $U = V$.

■

This is the dimension version of "count instead of check": to prove two subspaces are equal, prove **one** inclusion and compare dimensions. It is usually much less work than proving both inclusions.

Quick check. Let U be a subspace of $\mathbb{R}[x]_{\leq 2}$ that contains 1, $1+x$, and $(1+x)^2$. Must $U = \mathbb{R}[x]_{\leq 2}$?

Solution. Yes. By Example (A Basis by Counting), $(1, 1+x, (1+x)^2)$ is independent, and it lies in U, so $\dim U \geq 3$ by Theorem 1.6.3 applied in U (which is finite-dimensional by Theorem 1.6.7). Also $\dim U \leq \dim \mathbb{R}[x]_{\leq 2} = 3$. So $\dim U = 3$, and Theorem 1.6.8 gives $U = \mathbb{R}[x]_{\leq 2}$.

Warning. Both hypotheses of Theorem 1.6.8 matter. Containment is needed: the xy-plane and the xz-plane in $\mathbb{R}^3$ both have dimension 2, but neither contains the other, and they are different. Without finite dimension it fails too. In $F[x]$, the subspace U of polynomials with constant coefficient 0 has the basis-like list $x, x^2, x^3, \ldots$, which matches $1, x, x^2, \ldots$ one for one; yet $1 \notin U$. Later in this chapter we make this precise for infinite-dimensional spaces.

1.6.9 Exercises

1.6.9.1 A. Check your understanding

Exercise. A1

(a) State the Steinitz Exchange Theorem.

(b) True or false: $\dim \mathbb{R}[x]_{\leq 3} = 3$. Justify your answer.

(c) True or false: in a vector space of dimension 3, every spanning list of length 3 is a basis. Justify your answer.

(d) True or false: $M_2(\mathbb{R})$ contains a linearly independent list of length 5. Justify your answer.

(e) True or false: if U and W are subspaces of $\mathbb{R}^4$ with $\dim U = \dim W = 2$, then $U = W$. Justify your answer.

(f) What are $\dim_{\mathbb{C}} \mathbb{C}^2$ and $\dim_{\mathbb{R}} \mathbb{C}^2$?

Solution.

(a) See Theorem 1.6.1: if $(u_1, \ldots, u_m)$ is independent and $(w_1, \ldots, w_n)$ spans V, then $m \leq n$, and after renumbering the w's, $(u_1, \ldots, u_m, w_{m+1}, \ldots, w_n)$ spans V.

(b) False. The basis $(1, x, x^2, x^3)$ has length 4, so $\dim \mathbb{R}[x]_{\leq 3} = 4$.

(c) True, by Theorem 1.6.4 (b).

(d) False. $\dim M_2(\mathbb{R}) = 4$, and by Theorem 1.6.3 (a) every list of length $5 > 4$ is dependent.

(e) False. The subspaces $\operatorname{span}(\mathbf{e}_1, \mathbf{e}_2)$ and $\operatorname{span}(\mathbf{e}_3, \mathbf{e}_4)$ both have dimension 2, but $\mathbf{e}_1$ lies in the first and not in the second. Theorem 1.6.8 needs one subspace to be contained

in the other.

(f) $\dim_{\mathbb{C}} \mathbb{C}^2 = 2$, with basis $(\mathbf{e}_1, \mathbf{e}_2)$. Over $\mathbb{R}$, the list $(\mathbf{e}_1, i\mathbf{e}_1, \mathbf{e}_2, i\mathbf{e}_2)$ is a basis: every $(a+bi, c+di)$ with $a, b, c, d \in \mathbb{R}$ equals $a\mathbf{e}_1 + b(i\mathbf{e}_1) + c\mathbf{e}_2 + d(i\mathbf{e}_2)$, and this is $\mathbf{0}$ only if $a + bi = 0$ and $c + di = 0$, that is, $a = b = c = d = 0$. So $\dim_{\mathbb{R}} \mathbb{C}^2 = 4$.

1.6.9.2 B. Practice

Exercise. B1: Dimension and a Basis of a Subspace

Find a basis and the dimension of each subspace.

(a) $U_1 = \{(x_1, x_2, x_3, x_4) \in \mathbb{R}^4 : x_1 + x_2 = 0 \text{ and } x_3 - x_4 = 0\}$.

(b) $U_2 = \{\mathbf{A} \in M_3(\mathbb{R}) : \operatorname{tr} \mathbf{A} = 0\}$.

(c) $U_3 = \{p \in \mathbb{R}[x]_{\le 3} : p(1) = p(-1) = 0\}$.

Solution.

(a) A vector lies in U_1 exactly when $x_2 = -x_1$ and $x_3 = x_4$, that is, when it equals

$$(x_1, -x_1, x_4, x_4) = x_1(1, -1, 0, 0) + x_4(0, 0, 1, 1).$$

So $((1,-1,0,0),(0,0,1,1))$ spans U_1. It is independent: $a(1,-1,0,0) + b(0,0,1,1) = (a,-a,b,b) = \mathbf{0}$ forces $a = b = 0$. Hence it is a basis, and $\dim U_1 = 2$.

(b) Write $\mathbf{A} = (a_{ij})$. The condition $a_{11} + a_{22} + a_{33} = 0$ says $a_{33} = -a_{11} - a_{22}$, so

$$\mathbf{A} = \sum_{i \neq j} a_{ij}\mathbf{E}_{ij} + a_{11}(\mathbf{E}_{11} - \mathbf{E}_{33}) + a_{22}(\mathbf{E}_{22} - \mathbf{E}_{33}).$$

All eight matrices on the right have trace 0, so the list of the six $\mathbf{E}_{ij}$ with $i \neq j$, followed by $\mathbf{E}_{11} - \mathbf{E}_{33}$ and $\mathbf{E}_{22} - \mathbf{E}_{33}$, spans U_2. For independence, let $\sum_{i \neq j} c_{ij}\mathbf{E}_{ij} + d_1(\mathbf{E}_{11} - \mathbf{E}_{33}) + d_2(\mathbf{E}_{22} - \mathbf{E}_{33}) = \mathbf{O}$. The entry (i, j) with $i \neq j$ of the left side is c_{ij}, the entry $(1, 1)$ is d_1, and the entry $(2, 2)$ is d_2; all must be 0. Hence the list is a basis, and $\dim U_2 = 8$.

(c) Let $p = a_0 + a_1x + a_2x^2 + a_3x^3$. Then $p(1) = a_0 + a_1 + a_2 + a_3$ and $p(-1) = a_0 - a_1 + a_2 - a_3$. Adding and subtracting, $p \in U_3$ exactly when $a_0 + a_2 = 0$ and $a_1 + a_3 = 0$. Hence $p \in U_3$ exactly when

$$p = -a_2 - a_3x + a_2x^2 + a_3x^3 = a_2(x^2 - 1) + a_3(x^3 - x).$$

So (x^2-1, x^3-x) spans U_3. It is independent: $a(x^2-1)+b(x^3-x) = 0$ has x^3-coefficient b and x^2-coefficient a, so $a = b = 0$. Hence it is a basis, and $\dim U_3 = 2$.

Exercise. B2: Extending to a Basis of $\mathbb{R}^4$

Show that $((1,1,0,0),(0,1,1,0))$ is linearly independent, and extend it to a basis of $\mathbb{R}^4$ using vectors from the standard basis.

Solution. Let $\mathbf{u}_1 = (1,1,0,0)$ and $\mathbf{u}_2 = (0,1,1,0)$. If $a\mathbf{u}_1 + b\mathbf{u}_2 = (a, a+b, b, 0) = \mathbf{0}$, then $a = b = 0$, so the list is independent. By Theorem 1.6.5, two standard basis vectors complete it; we sift $(\mathbf{u}_1, \mathbf{u}_2, \mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_4)$ to find which.

- $\mathbf{u}_1 \neq \mathbf{0}$: keep. $\mathbf{u}_2 \notin \operatorname{span}(\mathbf{u}_1)$, since a multiple $(a, a, 0, 0)$ of $\mathbf{u}_1$ has third entry 0: keep.
- $\mathbf{e}_1$: if $(a, a+b, b, 0) = (1, 0, 0, 0)$, then $a = 1$ and $b = 0$, but then $a + b = 1 \neq 0$. So $\mathbf{e}_1 \notin \operatorname{span}(\mathbf{u}_1, \mathbf{u}_2)$: keep.
- $\mathbf{e}_2 = \mathbf{u}_1 - \mathbf{e}_1$: discard.
- $\mathbf{e}_3 = -\mathbf{u}_1 + \mathbf{u}_2 + \mathbf{e}_1$, since $-(1,1,0,0) + (0,1,1,0) + (1,0,0,0) = (0,0,1,0)$: discard.
- $\mathbf{e}_4$: every combination of $\mathbf{u}_1, \mathbf{u}_2, \mathbf{e}_1$ has fourth entry 0: keep.

The sifted list $(\mathbf{u}_1, \mathbf{u}_2, \mathbf{e}_1, \mathbf{e}_4)$ is independent by Theorem 1.5.4. It has $4 = \dim \mathbb{R}^4$ entries, so it is a basis of $\mathbb{R}^4$ by Theorem 1.6.4 (a).

Exercise. B3: Count Instead of Check

Show that $(1,\ x-1,\ (x-1)^2)$ is a basis of $\mathbb{R}[x]_{\leq 2}$. Hence find the coordinate vector of x^2 with respect to this basis.

Solution. Let $a + b(x-1) + c(x-1)^2 = 0$. Expanding, $(a - b + c) + (b - 2c)x + cx^2 = 0$. Comparing coefficients, $c = 0$, then $b = 2c = 0$, then $a = b - c = 0$. So the list is independent, with $3 = \dim \mathbb{R}[x]_{\leq 2}$ elements, and it is a basis by Theorem 1.6.4 (a).

Writing $x = (x-1) + 1$,

$$x^2 = \big((x-1)+1\big)^2 = 1 + 2(x-1) + (x-1)^2.$$

Hence the coordinate vector of x^2 is $(1, 2, 1)$.

1.6.9.3 C. Going deeper

Exercise. C1: Symmetric and Skew-Symmetric Matrices

Let $n \geq 1$. For a field F, let U_+ and U_- be the subspaces of symmetric and of skew-symmetric matrices in $M_n(F)$ (Definition 0.9.11): $U_+ = \{\mathbf{A} : \mathbf{A}^\top = \mathbf{A}\}$ and $U_- = \{\mathbf{A} : \mathbf{A}^\top = -\mathbf{A}\}$.

(a) For $F = \mathbb{R}$, prove that $\dim U_+ = \frac{n(n+1)}{2}$.

(b) For $F = \mathbb{R}$, prove that $\dim U_- = \frac{n(n-1)}{2}$.

(c) For $F = \mathbb{F}_2$, show that $U_- = U_+$, and find its dimension. Which step of (b) fails?

Solution.

(a) Consider the list $\mathcal{S}$ consisting of $\mathbf{E}_{ii}$ for $1 \leq i \leq n$ and $\mathbf{E}_{ij} + \mathbf{E}_{ji}$ for $1 \leq i < j \leq n$. Each is symmetric. If $\mathbf{A} = (a_{ij})$ is symmetric, then $a_{ji} = a_{ij}$, so

$$\mathbf{A} = \sum_i a_{ii}\mathbf{E}_{ii} + \sum_{i<j} a_{ij}(\mathbf{E}_{ij} + \mathbf{E}_{ji}),$$

because both sides have a_{ij} in entry (i,j) and $a_{ij} = a_{ji}$ in entry (j,i). So $\mathcal{S}$ spans U_+. If $\sum_i c_{ii}\mathbf{E}_{ii} + \sum_{i<j} c_{ij}(\mathbf{E}_{ij} + \mathbf{E}_{ji}) = \mathbf{O}$, then for $i \le j$ the entry (i,j) of the left side is c_{ij}, so every coefficient is 0. Hence $\mathcal{S}$ is a basis, of length $n + \frac{n(n-1)}{2} = \frac{n(n+1)}{2}$.

(b) If $\mathbf{A}^\top = -\mathbf{A}$, then comparing diagonal entries gives $a_{ii} = -a_{ii}$, so $2a_{ii} = 0$, and $a_{ii} = 0$ because $2 \neq 0$ in $\mathbb{R}$. Off the diagonal, $a_{ji} = -a_{ij}$. Hence

$$\mathbf{A} = \sum_{i<j} a_{ij}(\mathbf{E}_{ij} - \mathbf{E}_{ji}).$$

Each $\mathbf{E}_{ij} - \mathbf{E}_{ji}$ is skew-symmetric, so these matrices span U_-, and as in (a) the entry (i,j) with $i < j$ of a combination is its coefficient, so they are independent. There are $\frac{n(n-1)}{2}$ pairs $i < j$, so $\dim U_- = \frac{n(n-1)}{2}$.

(c) In $\mathbb{F}_2$, $-1 = 1$, so $-\mathbf{A} = \mathbf{A}$ for every matrix $\mathbf{A}$, and the conditions $\mathbf{A}^\top = -\mathbf{A}$ and $\mathbf{A}^\top = \mathbf{A}$ coincide. Hence $U_- = U_+$ over $\mathbb{F}_2$, which has dimension $\frac{n(n+1)}{2}$ by the argument of (a), which works over any field. The step of (b) that fails is "$2a_{ii} = 0$ implies $a_{ii} = 0$", which divides by $2 = 0$. For example, $\mathbf{I}_n \in U_-$ over $\mathbb{F}_2$, although its diagonal entries are not 0.

Exercise. C2: A complex space viewed over the reals

Let V be a vector space over $\mathbb{C}$ with $\dim_{\mathbb{C}} V = n$. Restricting the scalars to real numbers makes V a vector space over $\mathbb{R}$. Prove that $\dim_{\mathbb{R}} V = 2n$.

Hint: start from a basis $(v_1, \dots, v_n)$ over $\mathbb{C}$, and consider the vectors iv_k.

Solution. Let $(v_1, \dots, v_n)$ be a basis of V over $\mathbb{C}$. We show that $\mathcal{B} = (v_1, iv_1, \dots, v_n, iv_n)$ is a basis of V over $\mathbb{R}$.

Spanning. Let $v \in V$. Then $v = \sum_k z_k v_k$ with $z_k \in \mathbb{C}$. Write $z_k = a_k + b_k i$ with $a_k, b_k \in \mathbb{R}$. By the distributive and associative axioms of scalar multiplication, $z_k v_k = a_k v_k + (b_k i)v_k = a_k v_k + b_k(iv_k)$. Hence $v = \sum_k \left(a_k v_k + b_k(iv_k)\right)$, a real combination of $\mathcal{B}$.

Independence. Let $\sum_k \left(a_k v_k + b_k(iv_k)\right) = 0$ with $a_k, b_k \in \mathbb{R}$. Reading the same axioms backwards, $\sum_k (a_k + b_k i)v_k = 0$. Since $(v_1, \dots, v_n)$ is independent over $\mathbb{C}$, every $a_k + b_k i = 0$, so $a_k = b_k = 0$.

Hence $\mathcal{B}$ is a basis over $\mathbb{R}$ of length $2n$, and $\dim_{\mathbb{R}} V = 2n$.

Exercise. C3: Counting Elements over a Finite Field

Let p be a prime.

(a) Let V be a vector space over $\mathbb{F}_p$ with $\dim V = k$. Prove that V has exactly p^k elements.

(b) Deduce that no vector space over $\mathbb{F}_2$ has exactly 6 elements.

(c) Let K be a finite field that contains $\mathbb{F}_p$, with the same addition and multiplication. Explain why K is a finite-dimensional vector space over $\mathbb{F}_p$, and deduce that K has p^k

elements for some integer $k \geq 1$.

Hint: for (a), use coordinates.

Solution.

(a) Let $\mathcal{B}$ be a basis of V, of length k. By Theorem 1.5.2, $v \mapsto [v]_{\mathcal{B}}$ is a bijection from V to $\mathbb{F}_p^k$: every column is the coordinate vector of exactly one vector. So V has as many elements as $\mathbb{F}_p^k$. A column in $\mathbb{F}_p^k$ is a choice of k entries, each from the p elements of $\mathbb{F}_p$, so there are p^k columns. Hence V has p^k elements.

(b) A vector space over $\mathbb{F}_2$ with finitely many elements is spanned by the finite list of all its elements, so it is finite-dimensional, of some dimension k. By (a) it has 2^k elements. Since 6 is not a power of 2, no such space has 6 elements.

(c) Define scalar multiplication $\mathbb{F}_p \times K \to K$ by the multiplication of K. The vector space axioms are then instances of the field axioms of K: addition in K is commutative and associative with zero and negatives, multiplication is associative, $1 \cdot a = a$, and multiplication distributes over addition. So K is a vector space over $\mathbb{F}_p$. It is spanned by the finite list of all its elements, so it is finite-dimensional; let $k = \dim_{\mathbb{F}_p} K$. By (a), K has p^k elements. Since K contains $0 \neq 1$, it is not the zero space, so $k \geq 1$.

1.7 Sums and Direct Sums

We can now measure a single subspace: take a basis and count. This section asks how two or more subspaces combine. The intersection of two subspaces is again a subspace, but the union usually is not, so we need a better way to join them. The answer is the sum $U + W$. We then prove the dimension formula for sums and single out the best case, the direct sum, in which every vector splits into pieces in exactly one way.

1.7.1 The sum of subspaces

Recall from Theorem 1.2.2 that $U \cap W$ is a subspace whenever U and W are. It is the largest subspace inside both. The natural companion would be the smallest subspace containing both, and the first guess is the union $U \cup W$.

That guess fails. In $\mathbb{R}^2$, let U be the x-axis and W the y-axis. Then $(1, 0) \in U$ and $(0, 1) \in W$, but their sum $(1, 1)$ lies on neither axis, so $U \cup W$ is not closed under addition (Example (The Union of the Two Axes)). In fact Proposition 1.2.3 shows that $U \cup W$ is a subspace only in the dull case $U \subseteq W$ or $W \subseteq U$.

The failure tells us what is missing. Any subspace containing U and W must contain $u + w$ for every $u \in U$ and $w \in W$, because subspaces are closed under addition. So we put all these sums in from the start.

The sum of two subspaces is everything you can reach by adding one vector from each.

Definition 1.7.1. Sum of Subspaces

Let V be a vector space over F, and let U and W be subspaces of V. The **sum** of U and W

is the subset

$$U + W \coloneqq \{u + w : u \in U \text{ and } w \in W\}.$$

More generally, for subspaces $U_1, \ldots, U_k$ of V (with $k \geq 1$), their **sum** is

$$U_1 + \cdots + U_k \coloneqq \{u_1 + \cdots + u_k : u_i \in U_i \text{ for each } i = 1, \ldots, k\}.$$

In words: a vector v lies in $U + W$ when **there exist** some $u \in U$ and some $w \in W$ with $v = u + w$. The definition asks for one such pair. It does not say the pair is unique, and in general it is not. For k subspaces we pick **one** vector from **each** U_i and add them. We also write $\sum_{i=1}^{k} U_i$, and for $k = 1$ the sum is just U_1.

Here are some first examples.

- **The two axes.** In $\mathbb{R}^2$, with U the x-axis and W the y-axis, every (a, b) equals $(a, 0)+(0, b)$, so $U + W = \mathbb{R}^2$. The union is two lines; the sum is the whole plane.
- **Two planes in $\mathbb{R}^3$.** Let $P_{xy} = \{(x, y, 0)\}$ and $P_{yz} = \{(0, y, z)\}$. Every $(a, b, c) = (a, b, 0) + (0, 0, c)$ is a sum of a vector in P_{xy} and one in P_{yz}, so $P_{xy} + P_{yz} = \mathbb{R}^3$.
- **Polynomials.** In $F[x]_{\leq 2}$, let U be the constant polynomials and $W = \{bx+cx^2 : b, c \in F\}$. Then $a + bx + cx^2 = a + (bx + cx^2)$, so $U + W = F[x]_{\leq 2}$.
- **Matrices.** In $M_2(F)$, the upper triangular matrices plus the lower triangular matrices give everything:
$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} a & b \\ 0 & d \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ c & 0 \end{pmatrix}.$$
- **Degenerate cases.** For any subspace U, we have $U + \{0\} = U$ and $U + U = U$. In both, "$\supseteq$" holds because $u = u+0$, and "$\subseteq$" holds because U is closed under addition. So adding a subspace to itself produces nothing new: a sum is not a count.

Nonexample by minimal change. The definition makes sense for any two subsets $S, T \subseteq V$, but the result need not be a subspace. Take $S = \{(1, 0)\}$ and $T = \{(0, 1)\}$ in $\mathbb{R}^2$. Then $S + T = \{(1, 1)\}$, which does not contain $\mathbf{0}$. What changed is that S and T are not subspaces: the hypothesis "subspace" is exactly what makes the next theorem work.

Warning. The sum $U + W$ is **not** the union $U \cup W$. A vector of $U + W$ need not lie in U or in W: $(1, 1)$ lies in the sum of the two axes but on neither axis. Conversely, $U + W$ is usually far bigger than $U \cup W$.

The sum does what we built it to do.

Theorem 1.7.1. Sum of Subspaces Theorem

Let $U_1, \ldots, U_k$ be subspaces of a vector space V. Then $U_1 + \cdots + U_k$ is a subspace of V that contains each U_i. Moreover, if X is any subspace of V containing every U_i, then $U_1 + \cdots + U_k \subseteq X$. In other words, $U_1 + \cdots + U_k$ is the **smallest** subspace of V containing $U_1, \ldots, U_k$.

Proof. Write $S = U_1 + \cdots + U_k$. We apply the Subspace Test (Theorem 1.2.1).

(1) Each U_i contains 0, so $0 = 0 + \cdots + 0 \in S$.

(2) Let $u_1 + \cdots + u_k$ and $u'_1 + \cdots + u'_k$ be in S, with $u_i, u'_i \in U_i$. Regrouping with (VS1)

and (VS2),

$$(u_1 + \cdots + u_k) + (u_1' + \cdots + u_k') = (u_1 + u_1') + \cdots + (u_k + u_k'),$$

and $u_i + u_i' \in U_i$ because U_i is closed under addition. Hence S is closed under addition.

(3) Let $c \in F$. By the distributive axiom, $c(u_1 + \cdots + u_k) = cu_1 + \cdots + cu_k$, and $cu_i \in U_i$ because U_i is closed under scaling. Hence S is closed under scalar multiplication.

Therefore S is a subspace. It contains U_j: for $u \in U_j$, take $u_j = u$ and $u_i = 0 \in U_i$ for $i \neq j$. Finally, let X be a subspace containing every U_i. An element $u_1 + \cdots + u_k$ of S is a sum of vectors $u_i \in U_i \subseteq X$, so it lies in X since X is closed under addition. This shows $S \subseteq X$.

■

So the sum plays the role that the union could not. It is the subspace analogue of span, and the two notions fit together exactly. Recall from Theorem 1.3.1 that $\operatorname{span}(S)$ is the smallest subspace containing S.

Proposition 1.7.2. Sum of Spans

Let $L_1, \ldots, L_k$ be finite lists of vectors in V, and let L be the list obtained by writing $L_1, \ldots, L_k$ one after another. Then

$$\operatorname{span}(L_1) + \cdots + \operatorname{span}(L_k) = \operatorname{span}(L).$$

Proof. ($\subseteq$) The subspace $\operatorname{span}(L)$ contains every vector of each L_i, so by Theorem 1.3.1 it contains each $\operatorname{span}(L_i)$. By Theorem 1.7.1 it therefore contains $\operatorname{span}(L_1) + \cdots + \operatorname{span}(L_k)$.

($\supseteq$) The sum $\operatorname{span}(L_1) + \cdots + \operatorname{span}(L_k)$ is a subspace by Theorem 1.7.1, and it contains each $\operatorname{span}(L_i)$, hence every vector of L. By Theorem 1.3.1 it contains $\operatorname{span}(L)$.

■

In practice this means: **to find a spanning list of a sum, concatenate spanning lists of the summands.** That list may be dependent, and the dimension formula below tells us by how much.

Quick check. In $\mathbb{R}^3$, what is $\operatorname{span}((1,0,0)) + \operatorname{span}((0,1,0))$? Is it equal to $\operatorname{span}((1,0,0)) \cup \operatorname{span}((0,1,0))$?

Solution. By Proposition 1.7.2 the sum is $\operatorname{span}((1,0,0),(0,1,0)) = \{(a,b,0) : a,b \in \mathbb{R}\}$, the xy-plane. It is not the union: $(1,1,0)$ lies in the plane but on neither axis.

1.7.2 The dimension formula

If U and W are finite-dimensional, how big is $U + W$? Concatenating a basis of U and a basis of W gives a spanning list of $U + W$ of length $\dim U + \dim W$ (Proposition 1.7.2). That count is

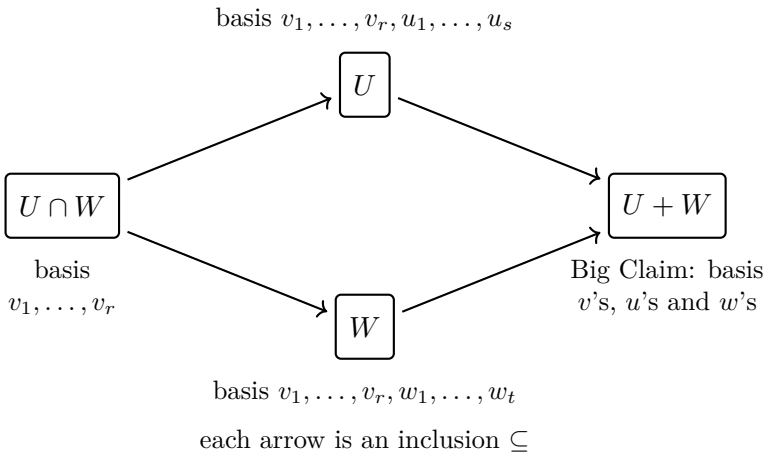
too large when U and W overlap. For the two planes P_{xy} and P_{yz} in $\mathbb{R}^3$, it would give $2+2=4$, but $P_{xy}+P_{yz}=\mathbb{R}^3$ has dimension 3. The overlap is the y-axis, of dimension 1, and it has been counted twice. The theorem says this is the whole story.

Theorem 1.7.3. Dimension Formula for Sums

Let U and W be finite-dimensional subspaces of a vector space V. Then $U+W$ is finite-dimensional, and

$$\dim(U+W)=\dim U+\dim W-\dim(U\cap W).$$

The picture behind the proof is the inclusion diagram. The smallest space, $U\cap W$, sits inside both U and W, and both sit inside $U+W$. We build bases from the bottom up.



Idea. This is a dimension formula, so we start from the **smallest** space and extend outward.

① Take a basis $(v_1,\dots,v_r)$ of $U\cap W$. ② Extend it to a basis of U by adding $u_1,\dots,u_s$, and separately to a basis of W by adding $w_1,\dots,w_t$. ③ **Big Claim:** the combined list of all $r+s+t$ vectors is a basis of $U+W$. Spanning is easy. For independence, move the w-part to the other side: the common vector lies in U and in W, hence in $U\cap W$, where the v's are a basis. ④ Count: $r+s+t=(r+s)+(r+t)-r$.

Proof. **Step 1: bases.** By Theorem 1.2.2, $U\cap W$ is a subspace of V; it lies inside U and uses the same operations, so it is also a subspace of U. Since U is finite-dimensional, so is $U\cap W$ (Theorem 1.6.7), and it has a basis $(v_1,\dots,v_r)$, where $r=\dim(U\cap W)$ (Corollary 1.5.5). This list is independent and lies in U, so by the Basis Extension Theorem (Theorem 1.6.5) there are vectors $u_1,\dots,u_s\in U$ such that $(v_1,\dots,v_r,u_1,\dots,u_s)$ is a basis of U. In the same way, since the list also lies in W, there are $w_1,\dots,w_t\in W$ such that $(v_1,\dots,v_r,w_1,\dots,w_t)$ is a basis of W. Thus $\dim U=r+s$ and $\dim W=r+t$.

Step 2: the Big Claim.

Claim. The list $\mathcal{B}=(v_1,\dots,v_r,u_1,\dots,u_s,w_1,\dots,w_t)$ is a basis of $U+W$.

Proof. Spanning. By Proposition 1.7.2,

$$U+W=\operatorname{span}(v_1,\dots,v_r,u_1,\dots,u_s)+\operatorname{span}(v_1,\dots,v_r,w_1,\dots,w_t)$$

is the span of the concatenated list. That list contains each v_i twice; removing the second copies does not change the span, since every removed vector already lies in the span of what remains (Theorem 1.3.5). Hence $\mathcal{B}$ spans $U + W$.

Independence. Let $a_i, b_j, c_k \in F$ satisfy

$$a_1v_1 + \cdots + a_rv_r + b_1u_1 + \cdots + b_su_s + c_1w_1 + \cdots + c_tw_t = 0.$$

Move the w-part across and name the common vector:

$$x \coloneqq a_1v_1 + \cdots + a_rv_r + b_1u_1 + \cdots + b_su_s = -c_1w_1 - \cdots - c_tw_t.$$

The middle expression lies in U, because all the v_i and u_j do. The right-hand expression lies in W, because all the w_k do. Hence $x \in U \cap W$. Since $(v_1, \ldots, v_r)$ spans $U \cap W$, there are $d_1, \ldots, d_r \in F$ with $x = d_1v_1 + \cdots + d_rv_r$. Comparing with the right-hand expression for x,

$$d_1v_1 + \cdots + d_rv_r + c_1w_1 + \cdots + c_tw_t = 0.$$

This is a combination of the basis $(v_1, \ldots, v_r, w_1, \ldots, w_t)$ of W, which is independent, so all $d_i = 0$ and all $c_k = 0$. Putting $c_1 = \cdots = c_t = 0$ into the original equation leaves

$$a_1v_1 + \cdots + a_rv_r + b_1u_1 + \cdots + b_su_s = 0,$$

a combination of the basis $(v_1, \ldots, v_r, u_1, \ldots, u_s)$ of U. By its independence, all $a_i = 0$ and all $b_j = 0$. Hence $\mathcal{B}$ is linearly independent (Definition 1.4.1).

■

Step 3: count. By the claim, $U + W$ has a finite basis of length $r + s + t$, so it is finite-dimensional with $\dim(U + W) = r + s + t$ (Definition 1.6.1). Therefore

$$\begin{aligned}\dim(U + W) &= r + s + t = (r + s) + (r + t) - r \\ &= \dim U + \dim W - \dim(U \cap W),\end{aligned}$$

as claimed.

■

The proof also works when some of r, s, t are 0: an empty list is independent and spans $\{0\}$, and every step reads correctly with empty sums. The move "basis of the smallest space, extend, Big Claim, count" is worth learning as a unit. It proves Rank–Nullity in Chapter 2 with the same four blocks.

The formula is often used to find one of the four numbers from the other three. Here are two uses, one by counting and one by computing.

Example. Two planes in $\mathbb{R}^3$

Let $U = \{(x, y, z) \in \mathbb{R}^3 : x + y + z = 0\}$ and $W = \{(x, y, z) \in \mathbb{R}^3 : x = z\}$. Show that $U + W = \mathbb{R}^3$, and find a basis of $\mathbb{R}^3$ of the shape used in the proof.

Solution. Solving $x + y + z = 0$ for y, every vector of U is $(x, -x - z, z) = x(1, -1, 0) + z(0, -1, 1)$. The two vectors are independent (look at the first and third entries), so $\dim U =$

2. Similarly every vector of W is $(x, y, x) = x(1, 0, 1) + y(0, 1, 0)$, and $\dim W = 2$.

A vector lies in $U \cap W$ when $x = z$ and $x + y + z = 0$, that is, $y = -2x$. So $U \cap W = \{(x, -2x, x)\} = \text{span}((1, -2, 1))$, which has dimension 1. By the dimension formula,

$$\dim(U + W) = 2 + 2 - 1 = 3 = \dim \mathbb{R}^3.$$

Since $U + W \subseteq \mathbb{R}^3$ and the dimensions agree, $U + W = \mathbb{R}^3$ by Theorem 1.6.8.

Following the proof: $(1, -2, 1)$ extends to a basis of U by $(1, -1, 0) \in U$. Indeed $(1, -1, 0)$ is not a multiple $(c, -2c, c)$ of $(1, -2, 1)$, since its first and third entries differ, so the list $((1, -2, 1), (1, -1, 0))$ is independent by Lemma 1.5.3; it has length $2 = \dim U$, so it is a basis of U (Theorem 1.6.4). It extends to a basis of W by $(0, 1, 0) \in W$, for the same reason: a multiple of $(1, -2, 1)$ with first entry 0 is $\mathbf{0}$. The Big Claim says $((1, -2, 1), (1, -1, 0), (0, 1, 0))$ is a basis of $U + W = \mathbb{R}^3$.

Example. Two planes in $\mathbb{R}^4$ from spanning lists

Let $U = \text{span}((1, 1, 0, 0), (0, 0, 1, 1))$ and $W = \text{span}((1, 0, 1, 0), (0, 1, 0, 1))$ in $\mathbb{R}^4$. Find $\dim(U \cap W)$ and $\dim(U + W)$, and hence a basis of $U + W$.

Solution. Each spanning list is independent: in $a(1, 1, 0, 0)+b(0, 0, 1, 1) = (a, a, b, b) = \mathbf{0}$, both a and b are 0, and similarly $c(1, 0, 1, 0) + d(0, 1, 0, 1) = (c, d, c, d)$. So $\dim U = \dim W = 2$.

A vector lies in $U \cap W$ exactly when it can be written in both forms, $(a, a, b, b) = (c, d, c, d)$. Comparing entries gives $a = c$, $a = d$, $b = c$, $b = d$, so $a = b = c = d$. Hence $U \cap W = \text{span}((1, 1, 1, 1))$ and $\dim(U \cap W) = 1$. By the dimension formula, $\dim(U+W) = 2+2-1 = 3$.

For a basis, follow the proof. The vector $(1, 1, 1, 1)$ together with $(1, 1, 0, 0)$ is a list in U, and it is independent by Lemma 1.5.3, because a multiple (c, c, c, c) of $(1, 1, 1, 1)$ never equals $(1, 1, 0, 0)$. It has length $2 = \dim U$, hence it is a basis of U (Theorem 1.6.4). Likewise $((1, 1, 1, 1), (1, 0, 1, 0))$ is a basis of W. By the Big Claim, $((1, 1, 1, 1), (1, 1, 0, 0), (1, 0, 1, 0))$ is a basis of $U + W$. In particular $U + W \neq \mathbb{R}^4$.

1.7.3 Direct sums

In a basis, every vector has **exactly one** expression as a combination (Theorem 1.5.1). Read that for F^n with the standard basis: $F^n = \text{span}(\mathbf{e}_1) + \cdots + \text{span}(\mathbf{e}_n)$, and each vector splits into pieces from these lines in exactly one way. We want the same property for subspaces that are not lines.

Sums do not always have it. In $\mathbb{R}^3 = P_{xy} + P_{yz}$,

$$(1, 1, 1) = (1, 1, 0) + (0, 0, 1) = (1, 0, 0) + (0, 1, 1),$$

two different splittings, because the planes share the y-axis and the y-component can be handed to either side. If we replace P_{yz} by the z-axis Z, then $(a, b, c) = (a, b, 0) + (0, 0, c)$ is the **only** splitting, since the Z-piece must have first two entries 0. We give this situation a name.

A sum is direct when every vector in it breaks into one piece from each subspace in exactly one way.

Definition 1.7.2. Direct Sum

Let $U_1, \dots, U_k$ be subspaces of a vector space V over F. The sum $U_1 + \cdots + U_k$ is **direct** if **every** vector $v \in U_1 + \cdots + U_k$ can be written in **exactly one** way as

$$v = u_1 + \cdots + u_k \quad \text{with } u_i \in U_i \text{ for each } i.$$

In that case we write the sum as $U_1 \oplus \cdots \oplus U_k$. We write $V = U_1 \oplus \cdots \oplus U_k$ to mean that $V = U_1 + \cdots + U_k$ **and** the sum is direct.

In words: existence of a splitting is automatic, since that is what membership in the sum means. The content of "direct" is **uniqueness**: if $u_1 + \cdots + u_k = u_1' + \cdots + u_k'$ with $u_i, u_i' \in U_i$, then $u_i = u_i'$ for **every** i. The symbol $\oplus$ does not build a new space. It is the same subspace $U_1 + \cdots + U_k$, with a property recorded in the notation. (Chapter 4 builds an "external" direct sum out of spaces that do not sit inside a common V.)

Examples.

- **Standard basis.** $F^n = \operatorname{span}(\mathbf{e}_1) \oplus \cdots \oplus \operatorname{span}(\mathbf{e}_n)$. More generally, if $(v_1, \dots, v_n)$ is a basis of V, then $V = \operatorname{span}(v_1) \oplus \cdots \oplus \operatorname{span}(v_n)$. A splitting $v = a_1 v_1 + \cdots + a_n v_n$ is a choice of coefficients, and Theorem 1.5.1 says there is exactly one.
- **Two splittings of the plane.** Let X be the x-axis, Y the y-axis, and $D = \{(t,t) : t \in \mathbb{R}\}$ the line $y = x$. Then $\mathbb{R}^2 = X \oplus Y$, with $(a,b) = (a,0) + (0,b)$ the only splitting. Also $\mathbb{R}^2 = X \oplus D$: if $(a,b) = (s,0) + (t,t)$, the second entry forces $t = b$, and then the first forces $s = a - b$. So $(a,b) = (a-b,0) + (b,b)$, and there is no other choice.
- **Polynomials.** $F[x]_{\le 2} = \operatorname{span}(1) \oplus \operatorname{span}(x, x^2)$. If $a + bx + cx^2 = a' + (b'x + c'x^2)$, comparing coefficients gives $a = a'$, $b = b'$, $c = c'$.
- **Degenerate.** $U \oplus \{0\} = U$ for every subspace U: the only possible $\{0\}$-piece is 0, so the U-piece of v must be v.

Nonexample by minimal change. Replace the z-axis in $\mathbb{R}^3 = P_{xy} \oplus Z$ by the plane P_{yz}. The sum is still all of $\mathbb{R}^3$, so existence still holds. What fails is uniqueness, as the two splittings of $(1,1,1)$ above show. So $\mathbb{R}^3 = P_{xy} + P_{yz}$, but the sum is **not** direct. Likewise upper plus lower triangular 2×2 matrices give $M_2(F)$, but a diagonal matrix can be put in either summand, so that sum is not direct.

Checking uniqueness for **every** vector looks like a lot of work. For two subspaces it reduces to one condition you can check in a line.

Theorem 1.7.4. Direct Sum Criteria for Two Subspaces

Let U and W be subspaces of a vector space V. The following are equivalent:

(a) the sum $U + W$ is direct;
(b) $U \cap W = \{0\}$.

If moreover U and W are finite-dimensional, they are also equivalent to

(c) $\dim(U + W) = \dim U + \dim W$.

Idea. Uniqueness questions go through differences. If $u + w = u' + w'$, then $u - u' = w' - w$ is a single vector lying in both U and W. So "no overlap" is exactly "no freedom in splitting". For

(c), the dimension formula says the overcount $\dim U + \dim W - \dim(U+W)$ is $\dim(U \cap W)$, which is 0 exactly when $U \cap W = \{0\}$.

Proof.

(a) $\Rightarrow$ (b). Suppose the sum is direct and let $v \in U \cap W$. Then $v = v + 0$ with $v \in U$, $0 \in W$, and $v = 0 + v$ with $0 \in U$, $v \in W$. These are two splittings of $v \in U + W$, so by uniqueness their U-pieces agree: $v = 0$. Hence $U \cap W = \{0\}$.

(b) $\Rightarrow$ (a). Suppose $U \cap W = \{0\}$, and let $u + w = u' + w'$ with $u, u' \in U$ and $w, w' \in W$. Then $u - u' = w' - w$. The left side lies in U and the right side in W, since both are subspaces. So this common vector lies in $U \cap W = \{0\}$, which gives $u = u'$ and $w = w'$. Hence every vector of $U + W$ has only one splitting.

(c) $\Leftrightarrow$ (c). Now let U and W be finite-dimensional. By the dimension formula (Theorem 1.7.3), (c) holds if and only if $\dim(U \cap W) = 0$. A space has dimension 0 exactly when the empty list is a basis of it, that is, when it equals the span of the empty list, which is $\{0\}$. So (c) holds if and only if $U \cap W = \{0\}$.

■

This is the everyday test: to show $V = U \oplus W$, show $V = U + W$ and $U \cap W = \{0\}$. In finite dimension there is a shortcut that avoids checking $V = U + W$ at all: if $U \cap W = \{0\}$ and $\dim U + \dim W = \dim V$, then $\dim(U + W) = \dim V$ by (c), so $U + W = V$ by Theorem 1.6.8. Count instead of check.

Example. A plane and a line in $\mathbb{R}^3$

Let $U = \{(x, y, z) \in \mathbb{R}^3 : x + y + z = 0\}$ and $L = \operatorname{span}((1, 1, 1))$. Show that $\mathbb{R}^3 = U \oplus L$, and split (a, b, c) into its two pieces.

Solution. From Example (Two planes in $\mathbb{R}^3$), $\dim U = 2$, and $\dim L = 1$ since $(1, 1, 1) \neq \mathbf{0}$. If $t(1, 1, 1) \in U$, then $t + t + t = 3t = 0$, so $t = 0$. Hence $U \cap L = \{\mathbf{0}\}$. By Theorem 1.7.4, $\dim(U + L) = 2 + 1 = 3$, so $U + L = \mathbb{R}^3$ by Theorem 1.6.8. Therefore $\mathbb{R}^3 = U \oplus L$.

To split (a, b, c), look for t with $(a, b, c) - t(1, 1, 1) \in U$, that is, $(a - t) + (b - t) + (c - t) = 0$. This gives $t = \frac{a+b+c}{3}$, and

$$(a, b, c) = \left(a - \tfrac{a+b+c}{3},\ b - \tfrac{a+b+c}{3},\ c - \tfrac{a+b+c}{3}\right) + \tfrac{a+b+c}{3}(1, 1, 1).$$

The L-piece is the average of the entries times $(1, 1, 1)$; the U-piece is what is left over.

The same splitting trick, "average with a reflected copy", works for functions and matrices.

Example. Even and odd functions

Let $V = \mathbb{R}^{\mathbb{R}}$ be the space of all functions $\mathbb{R} \to \mathbb{R}$, with $E = \{f : f(-x) = f(x) \text{ for all } x\}$ and $O = \{f : f(-x) = -f(x) \text{ for all } x\}$. Then $V = E \oplus O$.

Solution. Both are subspaces by the Subspace Test (Theorem 1.2.1): the zero function is even and odd, and the defining equations are preserved by sums and scalar multiples. For instance, if $f, g \in O$, then $(f+g)(-x) = -f(x) - g(x) = -(f+g)(x)$.

Sum. For any $f \in V$, put $f_E(x) = \frac{f(x)+f(-x)}{2}$ and $f_O(x) = \frac{f(x)-f(-x)}{2}$. Then $f = f_E + f_O$, and $f_E(-x) = \frac{f(-x)+f(x)}{2} = f_E(x)$, $f_O(-x) = \frac{f(-x)-f(x)}{2} = -f_O(x)$. So $V = E + O$.

Intersection. If $f \in E \cap O$, then $f(x) = f(-x) = -f(x)$ for every x, so $2f(x) = 0$ and $f(x) = 0$. Hence $E \cap O = \{0\}$, and $V = E \oplus O$ by Theorem 1.7.4.

For example, $e^x = \cosh x + \sinh x$ is the splitting of the exponential function.

Example. Symmetric and skew-symmetric matrices

Let F be a field of characteristic not 2 (for instance $\mathbb{R}$ or $\mathbb{C}$), and let $n \geq 1$. In $M_n(F)$, let $U_+ = \{\mathbf{A} : \mathbf{A}^\top = \mathbf{A}\}$ be the symmetric matrices and $U_- = \{\mathbf{A} : \mathbf{A}^\top = -\mathbf{A}\}$ the skew-symmetric ones (Definition 0.9.11). Then $M_n(F) = U_+ \oplus U_-$.

Solution. Both are subspaces: $\mathbf{O}^\top = \mathbf{O} = -\mathbf{O}$, and by Theorem 0.9.5, $(\mathbf{A}+\mathbf{B})^\top = \mathbf{A}^\top + \mathbf{B}^\top$ and $(c\mathbf{A})^\top = c\mathbf{A}^\top$, so the equations $\mathbf{A}^\top = \pm\mathbf{A}$ survive sums and scalar multiples.

Since the characteristic is not 2, we have $2 = 1 + 1 \neq 0$ in F, so $\frac{1}{2}$ exists (Definition 0.7.2). For $\mathbf{A} \in M_n(F)$,

$$\mathbf{A} = \tfrac{1}{2}(\mathbf{A} + \mathbf{A}^\top) + \tfrac{1}{2}(\mathbf{A} - \mathbf{A}^\top).$$

By Theorem 0.9.5, $\left(\frac{1}{2}(\mathbf{A} + \mathbf{A}^\top)\right)^\top = \frac{1}{2}(\mathbf{A}^\top + \mathbf{A})$ and $\left(\frac{1}{2}(\mathbf{A} - \mathbf{A}^\top)\right)^\top = \frac{1}{2}(\mathbf{A}^\top - \mathbf{A}) = -\frac{1}{2}(\mathbf{A} - \mathbf{A}^\top)$. So the first piece is in U_+, the second in U_-, and $M_n(F) = U_+ + U_-$.

If $\mathbf{A} \in U_+ \cap U_-$, then $\mathbf{A} = \mathbf{A}^\top = -\mathbf{A}$, so $2\mathbf{A} = \mathbf{O}$. Multiplying by $\frac{1}{2}$ gives $\mathbf{A} = \mathbf{O}$. Hence $U_+ \cap U_- = \{\mathbf{O}\}$, and $M_n(F) = U_+ \oplus U_-$ by Theorem 1.7.4. The pattern is the same as even plus odd: transpose plays the role of $x \mapsto -x$.

The hypothesis on the characteristic is used twice, both times to divide by 2. Exercise (C1: Symmetric Plus Skew over $\mathbb{F}_2$) shows what goes wrong over $\mathbb{F}_2$.

Quick check. Can two 2-dimensional subspaces of $\mathbb{R}^3$ form a direct sum?

Solution. No. If the sum $U + W$ were direct, Theorem 1.7.4 (c) would give $\dim(U + W) = 2 + 2 = 4$. But $U + W \subseteq \mathbb{R}^3$, so $\dim(U + W) \leq 3$ by Theorem 1.6.7. So two planes through $\mathbf{0}$ in $\mathbb{R}^3$ always share a nonzero vector.

1.7.4 Direct sums of several subspaces

For $k \geq 3$ subspaces, the tidy condition $U \cap W = \{0\}$ has no direct replacement by pairwise intersections; we will see why in a moment. What does carry over is the zero vector test, which

looks just like linear independence, and a condition that compares each subspace with the sum of **all the others**.

Theorem 1.7.5. Direct Sum Criteria for Several Subspaces

Let $U_1, \dots, U_k$ be subspaces of a vector space V, and write $S = U_1 + \cdots + U_k$. The following are equivalent:

(a) the sum S is direct;

(b) the only way to write $0 = u_1 + \cdots + u_k$ with $u_i \in U_i$ is $u_1 = \cdots = u_k = 0$;

(c) for **every** $j = 1, \dots, k$,
$$U_j \cap \sum_{i \neq j} U_i = \{0\}.$$

If moreover each U_i is finite-dimensional with a basis $\mathcal{B}_i$, and $\mathcal{B}$ is the list obtained by writing $\mathcal{B}_1, \dots, \mathcal{B}_k$ one after another, then (a)–(c) are also equivalent to

(d) $\mathcal{B}$ is a basis of S;

(e) $\dim S = \dim U_1 + \cdots + \dim U_k$.

Here $\sum_{i \neq j} U_i$ is the sum of the $k - 1$ subspaces other than U_j (and $\{0\}$ if $k = 1$).

Idea. Condition (b) is uniqueness for the single vector 0, and just as for independence, uniqueness for 0 gives uniqueness everywhere by taking differences. For (b) $\Leftrightarrow$ (c), a nontrivial way of writing 0 is the same thing as a nonzero vector u_j that equals minus the sum of the other pieces. For (d), a combination of $\mathcal{B}$ groups into one piece per U_i; independence of each $\mathcal{B}_i$ turns "piece is 0" into "coefficients are 0". Finally (d) $\Leftrightarrow$ (e) is count instead of check: $\mathcal{B}$ always spans S, so it is a basis exactly when its length is right.

Proof.

(a) $\Rightarrow$ (b). The vector $0 \in S$ has the splitting $0 = 0 + \cdots + 0$. If the sum is direct, it has no other, which is (b).

(b) $\Rightarrow$ (a). Assume (b), and let $u_1 + \cdots + u_k = u_1' + \cdots + u_k'$ with $u_i, u_i' \in U_i$. Subtracting and regrouping,
$$(u_1 - u_1') + \cdots + (u_k - u_k') = 0,$$
where $u_i - u_i' \in U_i$ because U_i is a subspace. By (b), $u_i - u_i' = 0$, that is, $u_i = u_i'$, for every i. Hence every vector of S has only one splitting.

(c) $\Rightarrow$ (c). Assume (b), fix j, and let $v \in U_j \cap \sum_{i \neq j} U_i$. Then $v = \sum_{i \neq j} u_i$ for some $u_i \in U_i$, and so
$$u_1 + \cdots + u_{j-1} + (-v) + u_{j+1} + \cdots + u_k = 0,$$
with $-v \in U_j$ in the j-th slot. By (b), every piece is 0; in particular $-v = 0$, so $v = 0$.

(d) $\Rightarrow$ (b). Assume (c), and let $u_1 + \cdots + u_k = 0$ with $u_i \in U_i$. Fix j. Then
$$u_j = -\sum_{i \neq j} u_i = \sum_{i \neq j} (-u_i).$$
The left side lies in U_j, and the right side lies in $\sum_{i \neq j} U_i$ since each $-u_i \in U_i$. By (c), $u_j = 0$. As j was arbitrary, all pieces are 0.

For the rest, assume each U_i is finite-dimensional with basis $\mathcal{B}_i = (b_{i1}, \dots, b_{im_i})$, where $m_i = \dim U_i$. By Proposition 1.7.2, $\mathcal{B}$ spans $S = \operatorname{span}(\mathcal{B}_1) + \cdots + \operatorname{span}(\mathcal{B}_k)$. In particular S is finite-dimensional.

(b) $\Rightarrow$ (d). Since $\mathcal{B}$ spans S, it remains to prove independence. Let $\sum_{i=1}^{k} \sum_{r=1}^{m_i} a_{ir} b_{ir} = 0$, and group the terms by subspace: put $u_i = \sum_{r=1}^{m_i} a_{ir} b_{ir} \in U_i$. Then $u_1 + \cdots + u_k = 0$, so each $u_i = 0$ by (b). Since $\mathcal{B}_i$ is independent, $a_{i1} = \cdots = a_{im_i} = 0$ for each i. Hence $\mathcal{B}$ is independent, and so a basis of S.

(c) $\Rightarrow$ (b). Assume (d), and let $u_1 + \cdots + u_k = 0$ with $u_i \in U_i$. Since $\mathcal{B}_i$ spans U_i, write $u_i = \sum_{r=1}^{m_i} a_{ir} b_{ir}$. Then $\sum_{i=1}^{k} \sum_{r=1}^{m_i} a_{ir} b_{ir} = 0$ is a combination of the independent list $\mathcal{B}$, so every $a_{ir} = 0$, and therefore every $u_i = 0$.

(d) $\Rightarrow$ (e). The list $\mathcal{B}$ has length $m_1 + \cdots + m_k$, so if it is a basis of S, then $\dim S = \dim U_1 + \cdots + \dim U_k$ by Definition 1.6.1.

(e) $\Rightarrow$ (d). The list $\mathcal{B}$ spans S and has length $m_1 + \cdots + m_k$, which equals $\dim S$ by (e). A spanning list whose length is the dimension is a basis (Theorem 1.6.4). This proves the theorem.

■

For $k = 2$, condition (c) reads $U_1 \cap U_2 = \{0\}$ and $U_2 \cap U_1 = \{0\}$, so we recover Theorem 1.7.4. The criteria (d) and (e) say that **bases of the pieces assemble into a basis of a direct sum, and dimensions add.**

For $k \geq 3$, the tempting shortcut is to check only the intersections of the subspaces two at a time. It does not work.

Warning. **Pairwise intersections $\{\mathbf{0}\}$ do not make a sum direct.** In $\mathbb{R}^2$, take the three lines X (the x-axis), Y (the y-axis), and D (the line $y = x$). Any two of them meet only in $\mathbf{0}$. But

$$(1, 0) + (0, 1) + (-1, -1) = (0, 0)$$

writes $\mathbf{0}$ with nonzero pieces from X, Y, D, so (b) fails and $X + Y + D$ is **not** direct. Condition (c) sees the problem: $D \cap (X + Y) = D \cap \mathbb{R}^2 = D \neq \{\mathbf{0}\}$.

The same example breaks another tempting generalization. For finite sets, $|A \cup B \cup C|$ is given by inclusion–exclusion, and the dimension formula looks like its two-set case. For three subspaces the analogous formula would be

$$\dim(U_1 + U_2 + U_3) \stackrel{?}{=} \sum_i \dim U_i - \sum_{i<j} \dim(U_i \cap U_j) + \dim(U_1 \cap U_2 \cap U_3).$$

For X, Y, D the left side is $\dim \mathbb{R}^2 = 2$, and the right side is $1+1+1-0-0-0+0 = 3$. So the formula is **false**. The reason is that $U + W$ is not a union: the dimension of $(U_1 + U_2) \cap U_3$ is not determined by the pairwise intersections. Exercise (C3: Three Subspaces) gives the correct replacement and a second failure.

1.7.5 Complements

Given a subspace U of V, can we always find a subspace W with $V = U \oplus W$? Such a W is called a **complement** of U in V. In $\mathbb{R}^3$, the plane U of Example (A plane and a line in $\mathbb{R}^3$) has the complement $\operatorname{span}((1, 1, 1))$. In finite dimension, complements always exist, and the proof is

again “basis of the smaller space, then extend”.

Theorem 1.7.6. Existence of Complements

Let V be a finite-dimensional vector space and U a subspace of V. Then there is a subspace W of V with $V = U \oplus W$. Every such W satisfies $\dim W = \dim V - \dim U$.

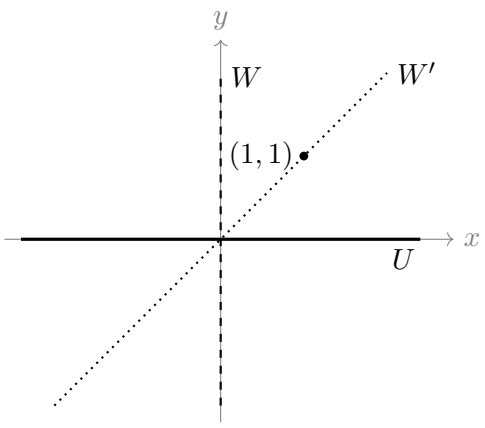
Proof. By Theorem 1.6.7, U is finite-dimensional; let $(u_1, \dots, u_m)$ be a basis of U (Corollary 1.5.5). This list is independent in V, so by the Basis Extension Theorem (Theorem 1.6.5) there are $w_1, \dots, w_r \in V$ such that $(u_1, \dots, u_m, w_1, \dots, w_r)$ is a basis of V. Put $W = \operatorname{span}(w_1, \dots, w_r)$, a subspace by Theorem 1.3.1.

By Proposition 1.7.2, $U + W = \operatorname{span}(u_1, \dots, u_m, w_1, \dots, w_r) = V$. The list $(w_1, \dots, w_r)$ is independent, since a relation $c_1 w_1 + \cdots + c_r w_r = 0$ is also a relation $0u_1 + \cdots + 0u_m + c_1 w_1 + \cdots + c_r w_r = 0$ among the basis vectors of V, forcing all $c_k = 0$. So $(w_1, \dots, w_r)$ is a basis of W. The concatenation of the bases of U and W is a basis of $U + W = V$, so the sum is direct by Theorem 1.7.5 ((d) $\Rightarrow$ (a)). Hence $V = U \oplus W$.

Finally, if W is any subspace with $V = U \oplus W$, then W is finite-dimensional (Theorem 1.6.7), and Theorem 1.7.4 (c) gives $\dim V = \dim U + \dim W$, as claimed.

■

The proof made a choice: the vectors $w_1, \dots, w_r$ that extend the basis. Different choices give different complements.



Warning. **A complement is not unique.** In $\mathbb{R}^2$, the x-axis U (solid) has the complement W, the y-axis (dashed), and also the complement W', the line $y = x$ (dotted), as shown in the examples after Definition 1.7.2. So “the complement of U” is meaningless, and $U \oplus W = U \oplus W'$ does **not** allow us to cancel U and conclude $W = W'$. What all complements share is their dimension. Also, a complement is not the set complement $V \setminus U$, which does not even contain 0.

Complements matter because a splitting $V = U \oplus W$ lets us study V one piece at a time. In Chapter 4 each such splitting gives a projection of V onto U along W, and the warning above becomes visible there: different complements give different projections. In Chapter 9 the eigenspaces of a diagonalizable operator split V as a direct sum, and that is what makes such

operators easy to understand.

1.7.6 Exercises

1.7.6.1 A. Check your understanding

Exercise. A1

(a) Define the sum $U_1 + \cdots + U_k$ of subspaces of V, and what it means for the sum to be direct.

(b) State the dimension formula for $\dim(U + W)$, with its hypotheses.

(c) True or false: if $V = U + W$ and $U \cap W = \{0\}$, then $V = U \oplus W$. Justify your answer.

(d) True or false: if U_1, U_2, U_3 are subspaces with $U_i \cap U_j = \{0\}$ for all $i \neq j$, then the sum $U_1 + U_2 + U_3$ is direct. Justify your answer.

(e) True or false: if U is a subspace of V with $U \neq V$, then $V \setminus U$ is a complement of U. Justify your answer.

(f) Name the method used to prove the dimension formula, in four steps.

Solution.

(a) See Definition 1.7.1: $U_1 + \cdots + U_k = \{u_1 + \cdots + u_k : u_i \in U_i\}$. By Definition 1.7.2, the sum is direct if every vector in it can be written as $u_1 + \cdots + u_k$ with $u_i \in U_i$ in exactly one way.

(b) If U and W are finite-dimensional subspaces of a vector space V, then $\dim(U + W) = \dim U + \dim W - \dim(U \cap W)$ (Theorem 1.7.3).

(c) True. By Theorem 1.7.4, $U \cap W = \{0\}$ means the sum $U + W$ is direct, and it equals V.

(d) False. The three lines X, Y, D in $\mathbb{R}^2$ (the axes and the line $y = x$) meet pairwise only in $\mathbf{0}$, but $(1, 0) + (0, 1) + (-1, -1) = \mathbf{0}$ is a nontrivial way to write $\mathbf{0}$, so the sum is not direct by Theorem 1.7.5.

(e) False. $V \setminus U$ does not contain 0, because $0 \in U$. So it is not a subspace, and cannot be a complement.

(f) Take a basis of the smallest space $U \cap W$; extend it to bases of U and of W (Basis Extension); prove the Big Claim that all the vectors together form a basis of $U + W$ (spanning, then independence by moving one block across and noting the common vector lies in $U \cap W$); count.

1.7.6.2 B. Practice

Exercise. B1: Two Planes in $\mathbb{R}^4$

Let $U = \operatorname{span}((1, 2, 0, 1), (0, 1, 1, 0))$ and $W = \operatorname{span}((1, 3, 1, 1), (0, 0, 1, 1))$ in $\mathbb{R}^4$.

(a) Show that $\dim U = \dim W = 2$.

(b) Find $U \cap W$ and its dimension.

(c) Hence find $\dim(U + W)$ and a basis of $U + W$.

Solution.

(a) If $a(1,2,0,1)+b(0,1,1,0) = (a, 2a+b, b, a) = \mathbf{0}$, then $a = 0$ and $b = 0$ from the first and third entries. So the spanning list of U is independent, hence a basis, and $\dim U = 2$. Similarly $c(1,3,1,1) + d(0,0,1,1) = (c, 3c, c+d, c+d) = \mathbf{0}$ forces $c = 0$, then $d = 0$, so $\dim W = 2$.

(b) A vector lies in $U \cap W$ exactly when $(a, 2a + b, b, a) = (c, 3c, c + d, c + d)$ for some $a, b, c, d \in \mathbb{R}$. The first entries give $a = c$. The second gives $2a + b = 3a$, so $b = a$. The third gives $a = a + d$, so $d = 0$; the fourth, $a = a + 0$, then holds. So the common vectors are $(a, 3a, a, a) = a(1, 3, 1, 1)$, and $U \cap W = \operatorname{span}((1, 3, 1, 1))$, of dimension 1.

(c) By the dimension formula (Theorem 1.7.3), $\dim(U + W) = 2 + 2 - 1 = 3$. Following its proof: $((1, 3, 1, 1), (0, 1, 1, 0))$ is a list in U (by (b), $(1, 3, 1, 1) \in U \cap W$). It is independent by Lemma 1.5.3, since a multiple $c(1, 3, 1, 1)$ with first entry 0 is $\mathbf{0} \neq (0, 1, 1, 0)$. It has length $2 = \dim U$, hence it is a basis of U by Theorem 1.6.4. Likewise $((1, 3, 1, 1), (0, 0, 1, 1))$ is a basis of W. By the Big Claim in the proof, $((1, 3, 1, 1), (0, 1, 1, 0), (0, 0, 1, 1))$ is a basis of $U + W$.

Exercise. B2: A Plane and a Line

Let $U = \{(x, y, z) \in \mathbb{R}^3 : x - 2y + z = 0\}$ and $L = \operatorname{span}((1, 0, 1))$.

(a) Prove that $\mathbb{R}^3 = U \oplus L$.

(b) Write $(2, 0, 4)$ as $\mathbf{u} + \mathbf{w}$ with $\mathbf{u} \in U$ and $\mathbf{w} \in L$.

Solution.

(a) Solving for x, every vector of U is $(2y - z, y, z) = y(2, 1, 0) + z(-1, 0, 1)$. These two vectors are independent (look at the second and third entries), so $\dim U = 2$. Since $(1, 0, 1) \neq \mathbf{0}$, $\dim L = 1$. If $t(1, 0, 1) \in U$, then $t - 0 + t = 2t = 0$, so $t = 0$; hence $U \cap L = \{\mathbf{0}\}$. By Theorem 1.7.4, $\dim(U + L) = 2 + 1 = 3$, so $U + L = \mathbb{R}^3$ by Theorem 1.6.8. Therefore $\mathbb{R}^3 = U \oplus L$.

(b) We need t with $(2, 0, 4) - t(1, 0, 1) = (2 - t, 0, 4 - t) \in U$, that is, $(2 - t) - 0 + (4 - t) = 0$. This gives $t = 3$. So

$$(2, 0, 4) = (-1, 0, 1) + 3(1, 0, 1),$$

and indeed $-1 - 0 + 1 = 0$, so $(-1, 0, 1) \in U$. By (a) this is the only such splitting.

Exercise. B3: Direct or Not?

Determine which of the following sums are direct. Justify your answer.

(a) $\operatorname{span}(1, x^2) + \operatorname{span}(x, x^3)$ in $\mathbb{R}[x]_{\leq 3}$.

(b) $T_{\text{up}} + T_{\text{low}}$ in $M_2(\mathbb{R})$, where T_{up} and T_{low} are the upper and lower triangular matrices.

(c) $\operatorname{span}((1, 0, 0)) + \operatorname{span}((0, 1, 0)) + \operatorname{span}((1, 1, 1))$ in $\mathbb{R}^3$.

(d) $\operatorname{span}((1, 1, 0)) + \operatorname{span}((0, 1, 1)) + \operatorname{span}((1, 0, -1))$ in $\mathbb{R}^3$.

Solution.

(a) Direct. If $p \in \operatorname{span}(1, x^2) \cap \operatorname{span}(x, x^3)$, then $p = a + bx^2 = cx + dx^3$. Comparing coefficients of $1, x, x^2, x^3$ gives $a = c = b = d = 0$, so $p = 0$. By Theorem 1.7.4 the sum is direct.

(b) Not direct. The identity matrix $\mathbf{I}_2$ is both upper and lower triangular, so $\mathbf{I}_2 \in T_{\text{up}} \cap T_{\text{low}} \neq \{\mathbf{O}\}$, and Theorem 1.7.4 applies.

(c) Direct. Each line has the basis given by its spanning vector, and we check that the concatenated list $((1,0,0),(0,1,0),(1,1,1))$ is independent. If $a(1,0,0) + b(0,1,0) + c(1,1,1) = (a+c, b+c, c) = \mathbf{0}$, then $c = 0$, hence $a = 0$ and $b = 0$. The list spans the sum by Proposition 1.7.2, so it is a basis of the sum, and the sum is direct by Theorem 1.7.5 ((d) $\Rightarrow$ (a)).

(d) Not direct, although any two of the three lines meet only in $\mathbf{0}$ (no spanning vector is a multiple of another). Indeed

$$(1,1,0) + (-1)(0,1,1) + (-1)(1,0,-1) = (0,0,0)$$

writes $\mathbf{0}$ as a sum of **nonzero** pieces from the three lines, so condition (b) of Theorem 1.7.5 fails.

1.7.6.3 C. Going deeper

Exercise. C1: Symmetric Plus Skew over $\mathbb{F}_2$

Let U_+ and U_- be the symmetric and skew-symmetric matrices in $M_2(\mathbb{F}_2)$.

(a) Show that $U_+ = U_-$.

(b) Deduce that $M_2(\mathbb{F}_2) \neq U_+ \oplus U_-$. Which step of Example (Symmetric and skew-symmetric matrices) fails, and why?

Solution.

(a) In $\mathbb{F}_2$ we have $1 + 1 = 0$, so $-a = a$ for every $a \in \mathbb{F}_2$, and hence $-\mathbf{A} = \mathbf{A}$ for every $\mathbf{A} \in M_2(\mathbb{F}_2)$. Therefore $\mathbf{A}^\top = -\mathbf{A}$ holds if and only if $\mathbf{A}^\top = \mathbf{A}$, which says $U_- = U_+$.

(b) By (a), $U_+ \cap U_- = U_+$, which contains $\mathbf{I}_2 \neq \mathbf{O}$. So the sum is not direct (Theorem 1.7.4). It is not even all of $M_2(\mathbb{F}_2)$: $U_+ + U_- = U_+ + U_+ = U_+$, and $\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$ is not symmetric. In Example (Symmetric and skew-symmetric matrices) both halves of the proof divide by 2: the splitting $\mathbf{A} = \frac{1}{2}(\mathbf{A} + \mathbf{A}^\top) + \frac{1}{2}(\mathbf{A} - \mathbf{A}^\top)$ and the step from $2\mathbf{A} = \mathbf{O}$ to $\mathbf{A} = \mathbf{O}$. In $\mathbb{F}_2$, $2 = 1 + 1 = 0$ has no inverse, so neither step is available. This is exactly the case of characteristic 2 excluded by the hypothesis.

Exercise. C2: Subspaces That Must Meet

(a) Let V be a vector space of dimension n, and let U, W be subspaces with $\dim U + \dim W > n$. Prove that $U \cap W \neq \{0\}$.

(b) Deduce that any two 3-dimensional subspaces of $\mathbb{R}[x]_{\leq 4}$ contain a common nonzero polynomial.

(c) Give an example showing that (a) fails if "$>$" is replaced by "$=$".

Solution.

(a) By Theorem 1.6.7, U, W, and $U+W$ are finite-dimensional, and $\dim(U+W) \leq \dim V = n$. By the dimension formula (Theorem 1.7.3),

$$\dim(U \cap W) = \dim U + \dim W - \dim(U+W) \geq \dim U + \dim W - n > 0.$$

A subspace of positive dimension has a basis with at least one vector, which is nonzero because a list containing 0 is dependent. Hence $U \cap W \neq \{0\}$.

(b) $\dim \mathbb{R}[x]_{\leq 4} = 5$ and $3+3 = 6 > 5$. By (a), the two subspaces meet in a nonzero polynomial.

(c) In $\mathbb{R}^2$ (with $n=2$), the x-axis and the y-axis have $1+1 = 2 = n$, but their intersection is $\{\mathbf{0}\}$.

Exercise. C3: Three Subspaces

(a) Let U_1, U_2, U_3 be subspaces of a vector space V. Show that $(U_1+U_2)+U_3 = U_1+U_2+U_3$.

(b) Now let U_1, U_2, U_3 be finite-dimensional. Prove that

$$\begin{aligned}\dim(U_1+U_2+U_3) = \dim U_1 + \dim U_2 + \dim U_3 \\ - \dim(U_1 \cap U_2) - \dim\big((U_1+U_2)\cap U_3\big).\end{aligned}$$

(c) In $\mathbb{R}^3$, let $U_1 = \{z=0\}$, $U_2 = \{y=0\}$, and $U_3 = \{y=z\}$. Compute $\dim(U_1+U_2+U_3)$, every $\dim(U_i \cap U_j)$, and $\dim(U_1 \cap U_2 \cap U_3)$. Show that the inclusion–exclusion formula fails, and compare with (b).

Hint: for (b), apply the dimension formula twice.

Solution.

(a) Both sides are subspaces by Theorem 1.7.1. The right side contains U_1 and U_2, hence U_1+U_2 (Theorem 1.7.1), and it contains U_3; so it contains $(U_1+U_2)+U_3$ by Theorem 1.7.1. Conversely, an element $u_1+u_2+u_3$ of the right side is the sum of $u_1+u_2 \in U_1+U_2$ and $u_3 \in U_3$, so it lies in the left side.

(b) By Theorem 1.7.3, U_1+U_2 is finite-dimensional with $\dim(U_1+U_2) = \dim U_1 + \dim U_2 - \dim(U_1 \cap U_2)$. Applying the formula again to U_1+U_2 and U_3, and using (a),

$$\dim(U_1+U_2+U_3) = \dim(U_1+U_2) + \dim U_3 - \dim\big((U_1+U_2)\cap U_3\big).$$

Substituting the first equation into the second gives the formula.

(c) Each U_i is a plane, so $\dim U_i = 2$: $U_1 = \operatorname{span}(\mathbf{e}_1, \mathbf{e}_2)$, $U_2 = \operatorname{span}(\mathbf{e}_1, \mathbf{e}_3)$, $U_3 = \operatorname{span}(\mathbf{e}_1, \mathbf{e}_2+\mathbf{e}_3)$, each with an evidently independent spanning list. All three contain the x-axis $\operatorname{span}(\mathbf{e}_1)$.

- $U_1 \cap U_2 = \{y=z=0\} = \operatorname{span}(\mathbf{e}_1)$.
- $U_1 \cap U_3 = \{z=0, y=z\} = \{y=z=0\} = \operatorname{span}(\mathbf{e}_1)$.
- $U_2 \cap U_3 = \{y=0, y=z\} = \operatorname{span}(\mathbf{e}_1)$.

- $U_1 \cap U_2 \cap U_3 = \operatorname{span}(\mathbf{e}_1)$.

So each of these has dimension 1. Since $U_1 + U_2$ contains $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$, it is $\mathbb{R}^3$, and so $U_1 + U_2 + U_3 = \mathbb{R}^3$ has dimension 3. The inclusion–exclusion formula would give

$$2 + 2 + 2 - 1 - 1 - 1 + 1 = 4 \neq 3,$$

so it fails. In (b), the last term is $(U_1 + U_2) \cap U_3 = \mathbb{R}^3 \cap U_3 = U_3$, of dimension 2, and indeed $2+2+2-1-2 = 3$. The inclusion–exclusion guess replaces this 2 by $\dim(U_1 \cap U_3) + \dim(U_2 \cap U_3) - \dim(U_1 \cap U_2 \cap U_3) = 1 + 1 - 1 = 1$, which is wrong because $(U_1 + U_2) \cap U_3$ is much bigger than $(U_1 \cap U_3) + (U_2 \cap U_3) = \operatorname{span}(\mathbf{e}_1)$.

1.8 Infinite-Dimensional Spaces

Every theorem about bases so far has assumed a finite spanning list. Yet several of our running examples have none: $F[x]$ is not finite-dimensional (Theorem 1.3.3), and neither are the sequences or the functions $\mathbb{R} \to \mathbb{R}$. Do such spaces have a basis at all? This section shows that they always do, using a principle from set theory called Zorn's Lemma, and then looks honestly at the price. The basis exists, but often nobody can write it down, and several finite-dimensional facts stop being true.

1.8.1 Bases of infinite sets of vectors

An infinite-dimensional space cannot have a finite basis, so a basis now has to be an infinite collection of vectors. We recall the conventions for arbitrary subsets, set up with the span (Definition 1.3.2) and with independence (Definition 1.4.1). A linear combination always involves **finitely many** vectors, because the axioms only let us add two vectors at a time.

Let $S \subseteq V$ be any subset, finite or infinite.

- $\operatorname{span}(S)$ is the set of all linear combinations $a_1 v_1 + \cdots + a_k v_k$ with $k \in \mathbb{N}$, $a_i \in F$, and $v_i \in S$.
- S is **linearly independent** if **every finite list of distinct vectors** of S is linearly independent in the sense of Definition 1.4.1.
- S is a **basis** of V if S is linearly independent and $\operatorname{span}(S) = V$.

For a finite set $S = \{v_1, \ldots, v_k\}$ of distinct vectors, these agree with the notions for the list $(v_1, \ldots, v_k)$, since a sub-list of an independent list is independent. The empty set is independent: its only finite list of distinct vectors is the empty list. In the infinite case we lose the ordering of a basis, and we will not need it.

Example. A Basis of $F[x]$

Show that $\{x^k : k \in \mathbb{N}\} = \{1, x, x^2, \ldots\}$ is a basis of $F[x]$.

Solution. Spanning. By Definition 0.8.1, every polynomial can be written as $a_0 + a_1 x + \cdots + a_N x^N$ for some $N \in \mathbb{N}$. This is a finite linear combination of $1, x, \ldots, x^N$.

Independence. Let $x^{k_1}, \ldots, x^{k_m}$ be distinct monomials and suppose $a_1 x^{k_1} + \cdots + a_m x^{k_m} = 0$. The left side is the polynomial whose coefficient of x^{k_j} is a_j, because the exponents k_j are distinct. The zero polynomial has every coefficient equal to 0, so by the definition of equality

of polynomials, $a_j = 0$ for every j.

Hence $\{x^k : k \in \mathbb{N}\}$ is a basis of $F[x]$.

The sequence space behaves differently. Let $F^{\mathbb{N}}$ be the space of all sequences $(a_0, a_1, a_2, \dots)$ with entries in F, with entrywise addition and scaling. For $i \in \mathbb{N}$ let e_i be the sequence with 1 in position i and 0 everywhere else.

Example. The Vectors e_i Do Not Span $F^{\mathbb{N}}$

Show that $\{e_i : i \in \mathbb{N}\}$ is linearly independent in $F^{\mathbb{N}}$, but is **not** a basis of $F^{\mathbb{N}}$.

Solution. Independence. Let $i_1, \dots, i_m$ be distinct and suppose $a_1 e_{i_1} + \cdots + a_m e_{i_m} = 0$. Position i_j of the left side is a_j, since only e_{i_j} has a nonzero entry there. Position i_j of the zero sequence is 0, so $a_j = 0$ for every j.

Not spanning. A linear combination $a_1 e_{i_1} + \cdots + a_m e_{i_m}$ has nonzero entries only in the m positions $i_1, \dots, i_m$. So every vector in $\operatorname{span}\{e_i : i \in \mathbb{N}\}$ has only finitely many nonzero entries. The sequence $(1, 1, 1, \dots)$ has infinitely many, so it is not in the span.

The same example shows that $F^{\mathbb{N}}$ is infinite-dimensional. For each n, the list $(e_0, \dots, e_n)$ is independent of length $n+1$. If a list of length m spanned $F^{\mathbb{N}}$, then $n + 1 \le m$ for every n by Theorem 1.6.1, which is impossible.

Warning. It is tempting to write $(1, 1, 1, \dots) = e_0 + e_1 + e_2 + \cdots$. In a vector space this sum **means nothing**: the axioms define sums of two vectors, hence of finitely many, and an infinite sum would need a notion of limit. So "the e_i span" is false, even though every sequence is "made of" its entries. Infinite sums return only in spaces with a notion of distance, much later.

Quick check. Let V be the space of formal power series $a_0 + a_1 x + a_2 x^2 + \cdots$ over F, that is, $F^{\mathbb{N}}$ with x^k standing for e_k. Is $\{1, x, x^2, \dots\}$ a basis of V?

Solution. No. It is independent, by the argument of Example (The Vectors e_i Do Not Span $F^{\mathbb{N}}$). But its span consists of the series with only finitely many nonzero coefficients, which are the polynomials. The series $1 + x + x^2 + \cdots$ is not among them, so the set does not span V.

So does $F^{\mathbb{N}}$ have a basis? In finite dimension we built bases by sifting (Theorem 1.5.4) or by extending one vector at a time (Theorem 1.6.5). Both constructions stop after finitely many steps. For $F^{\mathbb{N}}$ no step-by-step process can stop, and we need a different kind of existence argument.

1.8.2 Zorn's Lemma

Look again at what a basis is in finite dimension: an independent list that cannot be enlarged (Theorem 1.6.6). "Cannot be enlarged" is a statement about the family of all independent sets, ordered by inclusion. Zorn's Lemma is a general principle that produces objects which cannot be enlarged. To state it we need words for orderings.

A partial order is a way of comparing elements in which some pairs may be incomparable.

Definition 1.8.1. Partial Order, Chain, Upper Bound, Maximal Element

Let P be a set. A **partial order** on P is a relation $\leq$ on P (Definition 0.5.1) such that, **for all** $p, q, r \in P$:

(P1) $p \leq p$ (reflexive);

(P2) if $p \leq q$ and $q \leq p$, then $p = q$ (antisymmetric);

(P3) if $p \leq q$ and $q \leq r$, then $p \leq r$ (transitive).

Let $(P, \leq)$ be a partially ordered set.

- A **chain** in P is a subset $C \subseteq P$ such that **for all** $p, q \in C$, $p \leq q$ or $q \leq p$.
- An **upper bound** of a subset $C \subseteq P$ is an element $u \in P$ with $c \leq u$ **for every** $c \in C$.
- A **maximal element** of P is an element $m \in P$ such that **no** $p \in P$ satisfies $m \leq p$ and $p \neq m$.

In words: a chain is a subset in which any two elements can be compared. An upper bound of C sits above everything in C, but it only has to lie in P, **not** in C. A maximal element has nothing strictly above it. It is **not** required to lie above everything else.

Example. Inclusion as a Partial Order

Let X be a set and P the power set of X (Definition 0.3.12), ordered by $\subseteq$. Check that this is a partial order, and describe chains, upper bounds, and maximal elements when $X = \{1, 2\}$. Then do the same for the subsets of $\{1, 2\}$ with **at most one** element.

Solution. Every set satisfies $A \subseteq A$, which is (P1). If $A \subseteq B$ and $B \subseteq A$, then $A = B$ by Theorem 0.3.1, which is (P2). Inclusion is transitive, which is (P3).

For $X = \{1, 2\}$, the set $C = \{\varnothing, \{1\}, \{1, 2\}\}$ is a chain, while $\{1\}$ and $\{2\}$ are incomparable, so $\{\{1\}, \{2\}\}$ is not a chain. The set $\{1, 2\}$ is an upper bound of every subset of P, and it is the only maximal element.

Now let $Q = \{\varnothing, \{1\}, \{2\}\}$, still ordered by $\subseteq$. Both $\{1\}$ and $\{2\}$ are maximal: nothing in Q strictly contains either. Neither is above the other. The subset $\{\{1\}, \{2\}\}$ has no upper bound in Q.

Theorem 1.8.1. Zorn's Lemma

Let $(P, \leq)$ be a **nonempty** partially ordered set in which **every nonempty chain** has an upper bound in P. Then P has a maximal element.

We do not prove Zorn's Lemma, and no proof is possible from the other everyday rules of set theory. It is equivalent to the **axiom of choice**, the principle already used to build right inverses of surjections in Theorem 0.4.6: from any family of nonempty sets one may choose one element from each. Mathematicians accept it as an axiom, and we do too. In this section we point out each place where it is used.

Warning. "Maximal" is not "largest". The partially ordered set Q in Example (Inclusion as a Partial Order) has two maximal elements and no largest one. For bases this is the familiar fact that a space has many bases: in $\mathbb{R}^2$, both $\{\mathbf{e}_1, \mathbf{e}_2\}$ and $\{(1,1),(1,-1)\}$ are maximal independent sets, and neither contains the other.

1.8.3 Every vector space has a basis

We prove the more general extension statement first, since the existence of a basis is the case $S = \varnothing$.

Theorem 1.8.2. Basis Extension, General Form

Let V be a vector space over F, and let $S \subseteq V$ be a linearly independent subset. Then there is a basis B of V with $S \subseteq B$.

Idea. ① Let P be the family of independent sets containing S, ordered by $\subseteq$. ② A chain has the union as an upper bound, because an independence test uses only finitely many vectors, and finitely many members of a chain all fit inside one of them. ③ Zorn gives a maximal B. ④ If some v were outside $\operatorname{span}(B)$, then $B \cup \{v\}$ would be a bigger independent set, so B spans.

Proof. Let P be the set of all linearly independent subsets $T \subseteq V$ with $S \subseteq T$, partially ordered by $\subseteq$ (see Example (Inclusion as a Partial Order)). Since S is independent, $S \in P$, so P is nonempty.

Step 1: every nonempty chain has an upper bound. Let $C \subseteq P$ be a nonempty chain and let $U = \bigcup_{T \in C} T$. Every $T \in C$ contains S and C is nonempty, so $S \subseteq U$, and $T \subseteq U$ for every $T \in C$.

Claim. If $T_1, \dots, T_k \in C$ with $k \geq 1$, then one of them contains all the others.

Proof. We use induction on k. For $k = 1$ there is nothing to prove. Suppose the claim holds for k, and let $T_1, \dots, T_{k+1} \in C$. By the induction hypothesis some T_j contains $T_1, \dots, T_k$. Since C is a chain, $T_j \subseteq T_{k+1}$ or $T_{k+1} \subseteq T_j$. In the first case T_{k+1} contains all of $T_1, \dots, T_{k+1}$, and in the second case T_j does.

■

Now let $(u_1, \dots, u_k)$ be a list of distinct vectors of U. If $k = 0$, the empty list is independent. Otherwise, each u_i lies in some $T_i \in C$, and by the Claim one of $T_1, \dots, T_k$, say T_j, contains all the u_i. Since T_j is independent, the list $(u_1, \dots, u_k)$ is independent. Hence U is independent, so $U \in P$, and U is an upper bound of C.

Step 2: a maximal element. By Step 1 and Theorem 1.8.1, P has a maximal element B. Thus B is independent and $S \subseteq B$.

Step 3: B spans V. Suppose not, and let $v \in V$ with $v \notin \operatorname{span}(B)$. Then $v \notin B$, since $B \subseteq \operatorname{span}(B)$. We show $B \cup \{v\}$ is independent. Take a list of distinct vectors of $B \cup \{v\}$. If v is not in the list, the list lies in B and is independent. Otherwise, after reordering (which does not affect independence, as addition is commutative), the list is $(v, w_1, \ldots, w_m)$ with distinct $w_i \in B$. Let

$$av + b_1 w_1 + \cdots + b_m w_m = 0.$$

If $a \neq 0$, then $v = -a^{-1}(b_1 w_1 + \cdots + b_m w_m) \in \operatorname{span}(B)$, a contradiction. So $a = 0$, and then $b_1 = \cdots = b_m = 0$ because B is independent. Hence $B \cup \{v\} \in P$, it contains B, and it is not equal to B. This contradicts the maximality of B. Therefore $\operatorname{span}(B) = V$, and B is a basis of V containing S.

■

Notice where finiteness went. Steps 1 and 3 each look at only finitely many vectors at a time, which is all that linear algebra ever does. The one infinite step, "keep enlarging until you cannot", is handed to Zorn's Lemma.

Theorem 1.8.3. Every Vector Space Has a Basis

Every vector space V over a field F has a basis.

Proof. The empty set is linearly independent. By Theorem 1.8.2 with $S = \varnothing$, V has a basis. For $V = \{0\}$ that basis is $\varnothing$, since $\operatorname{span}(\varnothing) = \{0\}$.

■

So $F^{\mathbb{N}}$ has a basis, and so do $\mathbb{R}^{\mathbb{R}}$ and the continuous functions on $[0, 1]$. The theorem does not tell us what these bases are.

1.8.4 Existence without construction: $\mathbb{R}$ over $\mathbb{Q}$

The real numbers form a vector space over $\mathbb{Q}$: add reals as usual, and multiply a real by a rational. By Theorem 1.8.3 this space has a basis, often called a **Hamel basis**. Nobody has ever written one down, and nobody will. It is known that the existence of such a basis cannot be proved without some form of the axiom of choice. What we **can** prove is that this basis is infinite. The tool is counting.

A countable set is one whose elements can be listed as a sequence, possibly with repeats.

Definition 1.8.2. Countable Set

A set A is **countable** if $A = \varnothing$ or there is a **surjective** function $\mathbb{N} \to A$. Otherwise A is **uncountable**.

A surjection $\sigma\colon \mathbb{N} \to A$ lists A as $\sigma(0), \sigma(1), \sigma(2), \ldots$, and every element of A appears somewhere. Repeats are allowed, so every finite nonempty set is countable. The next lemma collects what we need.

Lemma 1.8.4. Countable Sets

(a) If A is countable and $f\colon A \to A'$ is surjective, then A' is countable.

(b) $\mathbb{N} \times \mathbb{N}$ is countable.

(c) If A and A' are countable, then $A \times A'$ is countable.

(d) $\mathbb{Q}$ is countable, and $\mathbb{Q}^n$ is countable for every $n \geq 1$.

Proof.

(a) If $A = \varnothing$, then $A' = f(A) = \varnothing$. Otherwise take a surjection $\sigma\colon \mathbb{N} \to A$. Then $f \circ \sigma\colon \mathbb{N} \to A'$ is surjective by Theorem 0.4.3 (a composite of surjections is surjective).

(b) For $n \in \mathbb{N}$, let 2^a be the largest power of 2 dividing $n+1$; it exists because $2^a \leq n+1$ bounds a. Then $(n+1)/2^a$ is odd, so it equals $2b+1$ for a unique $b \in \mathbb{N}$. Put $g(n) = (a, b)$. Given $(a, b) \in \mathbb{N} \times \mathbb{N}$, let $n = 2^a(2b+1) - 1$. The largest power of 2 dividing $2^a(2b+1)$ is 2^a, because $2^{a+1} \mid 2^a(2b+1)$ would force $2 \mid 2b+1$. Hence $g(n) = (a, b)$, and g is surjective.

(c) If A or A' is empty, so is $A \times A'$. Otherwise take surjections $\sigma\colon \mathbb{N} \to A$ and $\tau\colon \mathbb{N} \to A'$. The map $\mathbb{N} \times \mathbb{N} \to A \times A'$, $(m, n) \mapsto (\sigma(m), \tau(n))$, is surjective. By (b) and (a), $A \times A'$ is countable.

(d) The map $\mathbb{N} \to \mathbb{Z}$ sending an even m to $m/2$ and an odd m to $-(m+1)/2$ is surjective, so $\mathbb{Z}$ is countable. The map $\mathbb{Z} \times \mathbb{N} \to \mathbb{Q}$, $(p, q) \mapsto p/(q+1)$, is surjective, so $\mathbb{Q}$ is countable by (c) and (a). Finally $\mathbb{Q}^{n+1}$ is $\mathbb{Q}^n \times \mathbb{Q}$ after identifying $(q_1, \ldots, q_{n+1})$ with $((q_1, \ldots, q_n), q_{n+1})$, so induction on n with (c) shows that $\mathbb{Q}^n$ is countable for every $n \geq 1$.

■

The real numbers are not countable. This is a fact about $\mathbb{R}$, not about linear algebra, and its proof needs one property of $\mathbb{R}$ that we take from analysis: the **nested interval property**. If $I_0 \supseteq I_1 \supseteq I_2 \supseteq \cdots$ are closed bounded intervals $[a_n, b_n]$ with $a_n \leq b_n$, then some real number lies in every I_n.

Lemma 1.8.5. $\mathbb{R}$ Is Uncountable

Assuming the nested interval property, there is no surjection $\mathbb{N} \to \mathbb{R}$.

Proof. Let $r\colon \mathbb{N} \to \mathbb{R}$ be any function. We build closed intervals $I_0 \supseteq I_1 \supseteq \cdots$ of positive length with $r(n) \notin I_n$. Split $[0, 1]$ into the thirds $[0, \frac{1}{3}]$ and $[\frac{2}{3}, 1]$. They are disjoint, so $r(0)$ misses at least one of them; let I_0 be such a third. Given $I_n = [a, b]$ with $a < b$, the intervals $[a, a + \frac{b-a}{3}]$ and $[b - \frac{b-a}{3}, b]$ are disjoint and lie in I_n, so $r(n+1)$ misses one of them; let I_{n+1} be that one. By the nested interval property some x lies in every I_n. For each n, $x \in I_n$ and $r(n) \notin I_n$, so $x \neq r(n)$. Hence x is not a value of r, and r is not surjective.

■

Theorem 1.8.6. $\mathbb{R}$ Is Infinite-Dimensional over $\mathbb{Q}$

The vector space $\mathbb{R}$ over $\mathbb{Q}$ is not finite-dimensional.

Proof. Suppose a finite list spans $\mathbb{R}$ over $\mathbb{Q}$. Appending 1 keeps it spanning, so we may take a spanning list $(b_1, \dots, b_n)$ with $n \geq 1$. Then the map

$$\mathbb{Q}^n \to \mathbb{R}, \qquad (q_1, \dots, q_n) \mapsto q_1 b_1 + \cdots + q_n b_n$$

is surjective, since every real number is a rational linear combination of $b_1, \dots, b_n$. By Lemma 1.8.4 (d) and (a), $\mathbb{R}$ is countable. This contradicts Lemma 1.8.5. Hence $\mathbb{R}$ is not finite-dimensional over $\mathbb{Q}$.

■

The proof really shows more: **every** finite-dimensional vector space over $\mathbb{Q}$ is countable. It is honest about its inputs. Linear algebra supplied only "a spanning list gives a surjection from $\mathbb{Q}^n$"; the rest is counting and one property of the real line.

Quick check. Is $\mathbb{C}$ finite-dimensional as a vector space over $\mathbb{Q}$?

Solution. No. If $\mathbb{C}$ were countable, then $\mathbb{R} \subseteq \mathbb{C}$ would be countable too: pick any $y_0 \in \mathbb{R}$ and send $z \in \mathbb{C}$ to z if $z \in \mathbb{R}$ and to y_0 otherwise, a surjection $\mathbb{C} \to \mathbb{R}$, and apply Lemma 1.8.4 (a). That contradicts Lemma 1.8.5. So $\mathbb{C}$ is uncountable, and the proof of Theorem 1.8.6, with $\mathbb{C}$ in place of $\mathbb{R}$, shows it is not finite-dimensional over $\mathbb{Q}$.

1.8.5 What changes in infinite dimension

Several finite-dimensional theorems use a count, and counting is exactly what fails here.

Warning. **A proper subspace can be as big as the whole space.** In finite dimension, $U \subseteq V$ with $\dim U = \dim V$ forces $U = V$ (Theorem 1.6.8). Now let $U = \{p \in F[x] : \text{the constant coefficient of } p \text{ is } 0\}$. By the argument of Example (A Basis of $F[x]$), U has the basis $\{x, x^2, x^3, \dots\}$, and $x^k \leftrightarrow x^{k+1}$ matches it one-to-one with the basis $\{1, x, x^2, \dots\}$ of $F[x]$. Yet $1 \notin U$, so $U \neq F[x]$. The same happens in $F^{\mathbb{N}}$: the sequences with first entry 0 form a proper subspace, and deleting that first entry matches it one-to-one with all of $F^{\mathbb{N}}$, respecting addition and scaling. In Chapter 2 this becomes the shift map, which is injective but not surjective.

Remark. In an infinite-dimensional space, "$\dim V = \infty$" hides real differences of size: $F[x]$ has a countable basis, while every basis of $\mathbb{R}$ over $\mathbb{Q}$ is uncountable (Exercise (C2: Bases of $\mathbb{R}$ over $\mathbb{Q}$ Are Uncountable)), and set theory shows that all bases of one space have the same cardinality, a fact we will not develop.

Scope. From here on the book is mostly about finite-dimensional spaces, where bases are finite lists and counting works. Every theorem that needs finite dimension says so in its hypotheses. Infinite-dimensional spaces return with more structure, such as inner products and limits, where infinite sums can finally be given a meaning.

1.8.6 Exercises

1.8.6.1 A. Check your understanding

Exercise. A1

(a) Define what it means for an arbitrary subset S of a vector space V to be a basis of V.

(b) State Zorn's Lemma, including the meaning of "chain" and "upper bound".

(c) True or false: $\{e_i : i \in \mathbb{N}\}$ is a basis of $F^{\mathbb{N}}$. Justify your answer.

(d) True or false: every linearly independent subset of an infinite-dimensional space is infinite. Justify your answer.

(e) True or false: in a partially ordered set, a maximal element m satisfies $p \leq m$ for every p. Justify your answer.

Solution.

(a) S is a basis if every finite list of distinct vectors of S is linearly independent, and every vector of V is a linear combination of finitely many vectors of S.

(b) See Theorem 1.8.1 and Definition 1.8.1: a nonempty partially ordered set in which every nonempty chain (a subset whose elements are pairwise comparable) has an upper bound (an element of P above every element of the chain) has a maximal element.

(c) False. The set is independent, but $(1, 1, 1, \dots)$ is not a finite linear combination of the e_i (Example (The Vectors e_i Do Not Span $F^{\mathbb{N}}$)).

(d) False. The empty set and $\{x\} \subseteq F[x]$ are finite independent subsets of the infinite-dimensional space $F[x]$.

(e) False. In $Q = \{\varnothing, \{1\}, \{2\}\}$ ordered by $\subseteq$, the element $\{1\}$ is maximal, but $\{2\} \not\subseteq \{1\}$ (Example (Inclusion as a Partial Order)).

1.8.6.2 B. Practice

Exercise. B1: Finitely Supported Sequences

Let $F^{(\mathbb{N})} \subseteq F^{\mathbb{N}}$ be the set of sequences with only finitely many nonzero entries.

(a) Prove that $F^{(\mathbb{N})}$ is a subspace of $F^{\mathbb{N}}$.

(b) Prove that $\{e_i : i \in \mathbb{N}\}$ is a basis of $F^{(\mathbb{N})}$.

Solution.

(a) We use Theorem 1.2.1. (1) The zero sequence has no nonzero entries, so $0 \in F^{(\mathbb{N})}$. (2) Let $x, y \in F^{(\mathbb{N})}$, with nonzero entries only in finite sets of positions A and A'. If $i \notin A \cup A'$, then $x_i + y_i = 0 + 0 = 0$. So $x + y$ has nonzero entries only in the finite set

$A \cup A'$, and $x + y \in F^{(\mathbb{N})}$. (3) Let $c \in F$. If $i \notin A$, then $cx_i = c \cdot 0 = 0$, so $cx \in F^{(\mathbb{N})}$. Hence $F^{(\mathbb{N})}$ is a subspace.

(b) Each e_i lies in $F^{(\mathbb{N})}$, and the set is independent by Example (The Vectors e_i Do Not Span $F^{\mathbb{N}}$). For spanning, let $x \in F^{(\mathbb{N})}$ have nonzero entries only in positions $i_1, \dots, i_m$. Then x and $x_{i_1}e_{i_1} + \cdots + x_{i_m}e_{i_m}$ agree in positions $i_1, \dots, i_m$ and are both 0 elsewhere, so they are equal. Hence $\{e_i : i \in \mathbb{N}\}$ is a basis of $F^{(\mathbb{N})}$.

Exercise. B2: Exponentials Are Independent

For $k \in \mathbb{N}$ let $f_k \in \mathbb{R}^{\mathbb{R}}$ be $f_k(x) = e^{kx}$. Prove that $\{f_k : k \in \mathbb{N}\}$ is linearly independent. Hence deduce that $\mathbb{R}^{\mathbb{R}}$ is infinite-dimensional.

Hint: divide by the fastest-growing term.

Solution. Suppose not. Then there are distinct $k_1 < k_2 < \cdots < k_m$ and scalars $a_1, \dots, a_m$, not all zero, with $a_1e^{k_1x} + \cdots + a_me^{k_mx} = 0$ for every $x \in \mathbb{R}$. Discarding terms with zero coefficient, we may assume $a_m \neq 0$. Dividing by $e^{k_mx} > 0$ gives, for every $x \in \mathbb{R}$,

$$a_m = -\sum_{j=1}^{m-1} a_j e^{(k_j - k_m)x}.$$

Since $k_j - k_m < 0$ for $j < m$, each $e^{(k_j-k_m)x} \to 0$ as $x \to \infty$, so the right side tends to 0. The left side is the constant a_m, so $a_m = 0$, a contradiction. Hence the set is independent.

For each n, the list $(f_0, \dots, f_n)$ is independent of length $n+1$. If a list of length m spanned $\mathbb{R}^{\mathbb{R}}$, Theorem 1.6.1 would give $n + 1 \leq m$ for every n, which is impossible. Hence $\mathbb{R}^{\mathbb{R}}$ is infinite-dimensional.

Exercise. B3: Continuous Functions

Let $C[0,1]$ be the space of continuous functions $[0,1] \to \mathbb{R}$. Prove that $C[0,1]$ is infinite-dimensional.

Hint: use "tent" functions supported on disjoint intervals, and choose good points.

Solution. For $k \geq 1$, let $m_k = \frac{1}{2}\left(\frac{1}{k+1} + \frac{1}{k}\right)$ and $r_k = \frac{1}{2}\left(\frac{1}{k} - \frac{1}{k+1}\right) > 0$, the midpoint and half-length of $J_k = [\frac{1}{k+1}, \frac{1}{k}]$. Define

$$t_k(x) = \max(0,\ r_k - |x - m_k|).$$

This is continuous, as built from continuous functions by $|\cdot|$ and max. Also $t_k(x) > 0$ exactly when $|x - m_k| < r_k$, that is, when x lies in the open interval $(\frac{1}{k+1}, \frac{1}{k})$. These open intervals are pairwise disjoint, and m_j lies in the j-th one. Hence $t_k(m_j) = 0$ for $j \neq k$, while $t_k(m_k) = r_k$.

Let $a_1t_1 + \cdots + a_nt_n = 0$. Evaluating at m_j gives $a_jr_j = 0$, so $a_j = 0$ since $r_j \neq 0$. Hence $(t_1, \dots, t_n)$ is independent for every n. As in Exercise (B2: Exponentials Are Independent), Theorem 1.6.1 shows that no finite list spans $C[0,1]$.

1.8.6.3 C. Going deeper

Exercise. C1: Complements in Any Dimension

Let U be a subspace of a vector space V, of any dimension. Prove that there is a subspace W of V with $V = U \oplus W$, that is, every $v \in V$ can be written **uniquely** as $u + w$ with $u \in U$, $w \in W$.

Hint: Theorem 1.8.2.

Solution. By Theorem 1.8.3, U has a basis B_U. It is independent in V, so by Theorem 1.8.2 there is a basis B of V with $B_U \subseteq B$. Let $W = \operatorname{span}(B \setminus B_U)$, a subspace by Theorem 1.3.1.

Existence. Let $v \in V$. Since B spans, $v = \sum_i a_i u_i + \sum_j b_j w_j$, a finite sum with $u_i \in B_U$ and $w_j \in B \setminus B_U$. The first sum lies in U and the second in W.

$U \cap W = \{0\}$. Let $v \in U \cap W$. Since B_U spans U and $B \setminus B_U$ spans W, we have $v = \sum_i a_i u_i = \sum_j b_j w_j$ with distinct $u_i \in B_U$ and distinct $w_j \in B \setminus B_U$. Then $\sum_i a_i u_i - \sum_j b_j w_j = 0$ is a combination of a list of distinct vectors of B, because B_U and $B \setminus B_U$ are disjoint. Since B is independent, every a_i and b_j is 0, so $v = 0$.

Uniqueness. If $u + w = u' + w'$ with $u, u' \in U$ and $w, w' \in W$, then $u - u' = w' - w$ lies in $U \cap W = \{0\}$. Hence $u = u'$ and $w = w'$. This proves $V = U \oplus W$.

Exercise. C2: Bases of $\mathbb{R}$ over $\mathbb{Q}$ Are Uncountable

(a) Let $A_0, A_1, A_2, \ldots$ be countable sets. Prove that $\bigcup_{n \in \mathbb{N}} A_n$ is countable. (You may use the axiom of choice.)

(b) Let V be a vector space over $\mathbb{Q}$ with a countable basis. Prove that V is countable.

(c) Deduce that every basis of $\mathbb{R}$ over $\mathbb{Q}$ is uncountable.

Hint: for (b), write V as a union of spans of finite lists.

Solution.

(a) Let $A = \bigcup_n A_n$. If $A = \varnothing$ there is nothing to prove; otherwise fix $a^* \in A$. For each n with $A_n \neq \varnothing$, choose a surjection $\sigma_n \colon \mathbb{N} \to A_n$ (this is where choice is used). Define $h \colon \mathbb{N} \times \mathbb{N} \to A$ by $h(n, m) = \sigma_n(m)$ if $A_n \neq \varnothing$, and $h(n, m) = a^*$ otherwise. Every $a \in A$ lies in some nonempty A_n and equals $\sigma_n(m)$ for some m, so h is surjective. By Lemma 1.8.4 (b) and (a), A is countable.

(b) If the basis is empty, $V = \{0\}$ is countable. Otherwise list the basis as $\sigma(0), \sigma(1), \ldots$ for a surjection $\sigma \colon \mathbb{N} \to B$. Every $v \in V$ is a combination of finitely many basis vectors, all among $\sigma(0), \ldots, \sigma(n)$ for some n. Hence $V = \bigcup_n V_n$ with $V_n = \operatorname{span}(\sigma(0), \ldots, \sigma(n))$. The map $\mathbb{Q}^{n+1} \to V_n$, $(q_0, \ldots, q_n) \mapsto \sum_i q_i \sigma(i)$, is surjective, so each V_n is countable by Lemma 1.8.4 (d) and (a). By (a), V is countable.

(c) By Theorem 1.8.3, $\mathbb{R}$ has a basis over $\mathbb{Q}$. If some basis were countable, then $\mathbb{R}$ would be countable by (b), contradicting Lemma 1.8.5. Hence every basis is uncountable.

1.9 Changing the Field

Every vector space comes with its field, and we have already seen that the field matters: $\dim_{\mathbb{C}} \mathbb{C} = 1$ but $\dim_{\mathbb{R}} \mathbb{C} = 2$ (Example (Dimensions of the Running Examples)). This section explains the pattern behind that example. Shrinking the field multiplies the dimension by a fixed factor, the **tower law**. Going the other way, every real vector space can be enlarged to a complex one, its **complexification**, without changing the dimension. Both moves are used later, most of all when real matrices need complex eigenvalues.

1.9.1 Restricting the scalars

Start with the standard example. The space $\mathbb{C}^2$ over $\mathbb{C}$ has the basis $(\mathbf{e}_1, \mathbf{e}_2)$, so its dimension is 2. But nothing stops us from multiplying vectors of $\mathbb{C}^2$ only by **real** numbers. The addition is the same, and the axioms still hold, because they hold for all complex scalars. So $\mathbb{C}^2$ is also a vector space over $\mathbb{R}$. What is its dimension now?

Example. $\mathbb{C}^2$ as a Real Vector Space

Show that $(\mathbf{e}_1, i\mathbf{e}_1, \mathbf{e}_2, i\mathbf{e}_2)$ is a basis of $\mathbb{C}^2$ over $\mathbb{R}$, and write $(1+2i, 3-i)$ in it.

Solution. Spanning. Let $(z_1, z_2) \in \mathbb{C}^2$ and write $z_k = a_k + b_k i$ with $a_k, b_k \in \mathbb{R}$. Then

$$(z_1, z_2) = a_1\mathbf{e}_1 + b_1(i\mathbf{e}_1) + a_2\mathbf{e}_2 + b_2(i\mathbf{e}_2),$$

a combination with **real** coefficients. For example, $(1+2i, 3-i) = 1\mathbf{e}_1 + 2(i\mathbf{e}_1) + 3\mathbf{e}_2 + (-1)(i\mathbf{e}_2)$.

Independence. Let $a_1, b_1, a_2, b_2 \in \mathbb{R}$ with $a_1\mathbf{e}_1 + b_1(i\mathbf{e}_1) + a_2\mathbf{e}_2 + b_2(i\mathbf{e}_2) = \mathbf{0}$. The left side is $(a_1+b_1 i,\ a_2+b_2 i)$. Each entry is 0, and comparing real and imaginary parts (Definition 0.6.1) gives $a_1 = b_1 = a_2 = b_2 = 0$.

Hence $\dim_{\mathbb{R}} \mathbb{C}^2 = 4 = 2\dim_{\mathbb{C}} \mathbb{C}^2$.

The general setting has a big field L, a smaller field K inside it, and a vector space over L.

Definition 1.9.1. Subfield, Restriction of Scalars, Degree

Let L be a field. A **subfield** of L is a subset $K \subseteq L$ that contains 0 and 1 and is closed under addition, multiplication, negatives, and inverses of **nonzero** elements.

Let K be a subfield of L, and let V be a vector space over L. Keeping the addition of V and using scalar multiplication only for scalars in K makes V a vector space **over** K. We say V is obtained by **restriction of scalars**.

In particular L itself, a vector space over L, is a vector space over K. If it is finite-dimensional, its dimension is the **degree** $[L : K] := \dim_K L$.

In words: a subfield is a field sitting inside L with the same operations. Restriction of scalars forgets how to multiply by elements of L outside K.

There are two things to check. First, K with the operations of L is a field. Closure makes the operations land in K, the elements $0, 1$, negatives, and inverses are in K by assumption, and the remaining axioms are identities that hold for all elements of L, so in particular for those of K. Second, V over K is a vector space. Scalar multiplication $K \times V \to V$ lands in V, and each axiom of Definition 1.1.1 holds **for all** scalars in L, so in particular for all scalars in K.

We always write $\dim_K V$ or $\dim_L V$ when both fields are in play.

Example. Degrees of Some Fields

Find $[\mathbb{C} : \mathbb{R}]$, $[\mathbb{Q}(\sqrt{2}) : \mathbb{Q}]$, and $[K : K]$ for any field K. Is $[\mathbb{R} : \mathbb{Q}]$ defined?

Solution. $[\mathbb{C} : \mathbb{R}] = 2$. Every $z \in \mathbb{C}$ is $a \cdot 1 + b \cdot i$ with $a, b \in \mathbb{R}$, and $a + bi = 0$ forces $a = b = 0$ by Definition 0.6.1. So $(1, i)$ is a basis of $\mathbb{C}$ over $\mathbb{R}$.

$[\mathbb{Q}(\sqrt{2}) : \mathbb{Q}] = 2$. The set $\mathbb{Q}(\sqrt{2}) = \{a + b\sqrt{2} : a, b \in \mathbb{Q}\}$ is a subfield of $\mathbb{R}$ by Example ($\mathbb{Q}(\sqrt{2})$ Is a Field). The list $(1, \sqrt{2})$ spans it over $\mathbb{Q}$ by definition. For independence, let $a + b\sqrt{2} = 0$ with $a, b \in \mathbb{Q}$. If $b \neq 0$, then $\sqrt{2} = -a/b \in \mathbb{Q}$, contradicting Theorem 0.2.3. So $b = 0$, and then $a = 0$.

$[K : K] = 1$. The list (1) is a basis of K over K: every a equals $a \cdot 1$, and $a \cdot 1 = 0$ forces $a = 0$. This degenerate case says that restricting to the same field changes nothing.

$[\mathbb{R} : \mathbb{Q}]$ is not defined, because $\mathbb{R}$ is not finite-dimensional over $\mathbb{Q}$ (Theorem 1.8.6).

Restriction only goes **down**. The set $\mathbb{R}$ with the operations of $\mathbb{C}$ is **not** a vector space over $\mathbb{C}$: everything about addition works, but scalar multiplication leaves the set, since $i \cdot 1 = i \notin \mathbb{R}$. Enlarging the field needs a construction, which is the second half of this section.

1.9.2 The tower law

In Example ($\mathbb{C}^2$ as a Real Vector Space) we took a vector, wrote it in coordinates over $\mathbb{C}$, and then split each complex coordinate into real coordinates. The same two-stage process works for any fields.

Theorem 1.9.1. Tower Law

Let K be a subfield of L, and let V be a vector space over L. Suppose $(a_1, \ldots, a_m)$ is a basis of L over K and $(v_1, \ldots, v_n)$ is a basis of V over L. Then the list of the mn vectors

$$(a_1 v_1, \ldots, a_m v_1,\ a_1 v_2, \ldots, a_m v_2,\ \ldots,\ a_1 v_n, \ldots, a_m v_n)$$

is a basis of V over K. In particular,

$$\dim_K V = [L : K] \cdot \dim_L V.$$

Idea. Coordinates in two stages. A vector is an L-combination of the v_j; each L-coefficient is a K-combination of the a_i. Substituting gives a double sum $\sum_j \sum_i b_{ij}\, a_i v_j$ with $b_{ij} \in K$.

For independence, run the same two stages backwards: regroup the double sum by j, use independence over L, then independence over K.

Proof. Throughout we use two axioms of Definition 1.1.1 for scalars in L: $(c+c')x = cx+c'x$ and $c(c'x) = (cc')x$.

Spanning. Let $x \in V$. Since $(v_1, \dots, v_n)$ spans V over L, there are $c_1, \dots, c_n \in L$ with $x = \sum_{j=1}^{n} c_j v_j$. Since $(a_1, \dots, a_m)$ spans L over K, for each j there are $b_{1j}, \dots, b_{mj} \in K$ with $c_j = \sum_{i=1}^{m} b_{ij} a_i$. Therefore

$$x = \sum_{j=1}^{n} \Big(\sum_{i=1}^{m} b_{ij} a_i \Big) v_j = \sum_{j=1}^{n} \sum_{i=1}^{m} (b_{ij} a_i) v_j = \sum_{j=1}^{n} \sum_{i=1}^{m} b_{ij} (a_i v_j),$$

where the second equality uses $(c + c')x = cx + c'x$ and the third uses $c(c'x) = (cc')x$. This is a K-linear combination of the vectors $a_i v_j$.

Independence. Let $b_{ij} \in K$ with $\sum_{j=1}^{n} \sum_{i=1}^{m} b_{ij}(a_i v_j) = 0$. Reading the display above from right to left,

$$\sum_{j=1}^{n} \Big(\sum_{i=1}^{m} b_{ij} a_i \Big) v_j = 0.$$

Each coefficient $\sum_i b_{ij} a_i$ lies in L, and $(v_1, \dots, v_n)$ is independent over L. Hence $\sum_{i=1}^{m} b_{ij} a_i = 0$ for every j. For each fixed j, this is a K-combination of $(a_1, \dots, a_m)$ equal to 0, and that list is independent over K. Hence $b_{ij} = 0$ for all i and j.

So the mn vectors $a_i v_j$ form a basis of V over K, and $\dim_K V = mn = [L : K] \dim_L V$.

■

The name comes from the picture of fields stacked in a tower, $K \subseteq L$, with V on top: dimensions multiply as you go down. The most important case is $\mathbb{R} \subseteq \mathbb{C}$.

Theorem 1.9.2. Complex Spaces as Real Spaces

Let V be a vector space over $\mathbb{C}$ with basis $(v_1, \dots, v_n)$. Then $(v_1, iv_1, \dots, v_n, iv_n)$ is a basis of V over $\mathbb{R}$. In particular $\dim_{\mathbb{R}} V = 2 \dim_{\mathbb{C}} V$.

Proof. By Example (Degrees of Some Fields), $(1, i)$ is a basis of $\mathbb{C}$ over $\mathbb{R}$. By Theorem 1.9.1 with $a_1 = 1$ and $a_2 = i$, the list $(1v_1, iv_1, \dots, 1v_n, iv_n)$ is a basis of V over $\mathbb{R}$, and $1v_j = v_j$ by Definition 1.1.1. The dimension is $2n$.

■

This is the result of Exercise (C2: A complex space viewed over the reals), now as a case of the tower law. For example, $\dim_{\mathbb{R}} \mathbb{C}^n = 2n$, and $\dim_{\mathbb{R}} \mathbb{C}[x]_{\le n} = 2(n+1)$, with basis $(1, i, x, ix, \dots, x^n, ix^n)$.

Warning. A subset of $\mathbb{C}^n$ can be a subspace over $\mathbb{R}$ without being a subspace over $\mathbb{C}$. The real vectors $\mathbb{R}^n \subseteq \mathbb{C}^n$ contain $\mathbf{0}$ and are closed under addition and under multiplication by **real** scalars, so they form an $\mathbb{R}$-subspace of dimension n. But $\mathbf{e}_1 \in \mathbb{R}^n$ and $i\mathbf{e}_1 \notin \mathbb{R}^n$, so $\mathbb{R}^n$ is **not** a $\mathbb{C}$-subspace. Always ask: a subspace over which field?

Quick check. What is $\dim_{\mathbb{R}} M_{2\times 3}(\mathbb{C})$? What is $\dim_{\mathbb{Q}} \mathbb{Q}(\sqrt{2})^2$?

Solution. $M_{2\times 3}(\mathbb{C})$ has dimension 6 over $\mathbb{C}$ (Example (Dimensions of the Running Examples)), so $\dim_{\mathbb{R}} = 2 \cdot 6 = 12$ by Theorem 1.9.2. The space $\mathbb{Q}(\sqrt{2})^2$ has dimension 2 over the field $\mathbb{Q}(\sqrt{2})$, and $[\mathbb{Q}(\sqrt{2}) : \mathbb{Q}] = 2$ by Example (Degrees of Some Fields), so $\dim_{\mathbb{Q}} = 2 \cdot 2 = 4$ by Theorem 1.9.1.

1.9.3 Complexification

Now go up. A real matrix such as the rotation $\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ moves every nonzero vector of $\mathbb{R}^2$ off its own line. In $\mathbb{C}^2$, however, it sends $(1, -i)$ to $(i, 1) = i(1, -i)$. To use such complex vectors for an arbitrary real space V, not just $\mathbb{R}^n$, we need a complex space that contains V the way $\mathbb{C}^n$ contains $\mathbb{R}^n$.

The model is $\mathbb{C}^n$ itself. Every $\mathbf{z} \in \mathbb{C}^n$ splits as $\mathbf{z} = \mathbf{x} + i\mathbf{y}$ with $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, its real and imaginary parts, taken entry by entry. Multiplying by $a + bi$ with $a, b \in \mathbb{R}$ gives

$$(a + bi)(\mathbf{x} + i\mathbf{y}) = (a\mathbf{x} - b\mathbf{y}) + i(a\mathbf{y} + b\mathbf{x}).$$

The right side only uses real scalars acting on real vectors. So we can copy it for any real space, exactly as $\mathbb{C}$ was built from pairs of real numbers in Definition 0.6.1.

The complexification of V is the space of formal expressions $u + iv$ with $u, v \in V$, multiplied by complex numbers the way the rules of algebra dictate.

Definition 1.9.2. Complexification

Let V be a vector space over $\mathbb{R}$. The **complexification** of V is the set $V_{\mathbb{C}} := V \times V$ with addition and scalar multiplication by $\mathbb{C}$ defined, **for all** $(u, v), (u', v') \in V_{\mathbb{C}}$ and $a, b \in \mathbb{R}$, by

$$(u, v) + (u', v') := (u + u',\ v + v'), \qquad (a + bi)(u, v) := (au - bv,\ av + bu).$$

We write $u + iv$ for (u, v), and u for $(u, 0)$.

In words: addition is entry by entry. Multiplication by $a + bi$ is the formula forced by $i^2 = -1$. The notation is honest: $i(v, 0) = (0v - 10,\ 00 + 1v) = (0, v)$, by Theorem 1.1.5 and Theorem 1.1.6, so $(u, 0) + i(v, 0) = (u, v)$. Also a **real** scalar acts entrywise: $a(u, v) = (au, av)$. So $V_{\mathbb{C}}$, restricted to real scalars, is just two copies of V.

Why $V_{\mathbb{C}}$ is a vector space over $\mathbb{C}$. The operations land in $V \times V$. Addition is entrywise, so (commutativity, associativity, zero, negatives) follow from the same axioms in each entry of V, with zero $(0, 0)$ and negative $(-u, -v)$. We check the scalar axioms that involve the

multiplication of $\mathbb{C}$; below, $au - bv$ means $au + (-b)v$, which is the same by Theorem 1.1.8. Let $\alpha = a + bi$ and $\beta = c + di$ with $a, b, c, d \in \mathbb{R}$, and $z = (u, v)$.

- $1z = z$: we get $(1u - 0v,\ 1v + 0u) = (u, v)$, since $0v = 0$ by Theorem 1.1.5.
- $\alpha(\beta z) = (\alpha\beta)z$: since $\beta z = (cu - dv,\ cv + du)$, the axioms of V give

$$\begin{aligned}\alpha(\beta z) &= (a(cu - dv) - b(cv + du),\ a(cv + du) + b(cu - dv)) \\ &= ((ac - bd)u - (ad + bc)v,\ (ac - bd)v + (ad + bc)u).\end{aligned}$$

  Since $\alpha\beta = (ac - bd) + (ad + bc)i$, this is $(\alpha\beta)z$.
- $(\alpha + \beta)z = \alpha z + \beta z$: with $\alpha + \beta = (a + c) + (b + d)i$,

$$\begin{aligned}(\alpha + \beta)z &= ((a + c)u - (b + d)v,\ (a + c)v + (b + d)u) \\ &= (au - bv,\ av + bu) + (cu - dv,\ cv + du) \\ &= \alpha z + \beta z.\end{aligned}$$

The remaining axiom, $\alpha(z + w) = \alpha z + \alpha w$, is Exercise (B3: Working in a Complexification).

Here is the payoff: complexifying keeps the dimension, and even keeps the basis.

Theorem 1.9.3. A Real Basis Is a Complex Basis of the Complexification

Let V be a vector space over $\mathbb{R}$ with basis $(v_1, \ldots, v_n)$. Then $(v_1, \ldots, v_n)$, that is, $((v_1, 0), \ldots, (v_n, 0))$, is a basis of $V_{\mathbb{C}}$ over $\mathbb{C}$. In particular $\dim_{\mathbb{C}} V_{\mathbb{C}} = \dim_{\mathbb{R}} V$.

Idea. A complex coefficient $a_j + b_j i$ on v_j puts $a_j v_j$ into the real part and $b_j v_j$ into the imaginary part. So one complex combination of the v_j is the same as two real combinations, one for each part, and both questions (spanning, independence) split into two real questions about $(v_1, \ldots, v_n)$.

Proof. For $a, b \in \mathbb{R}$ and $x \in V$, Definition 1.9.2 and Theorem 1.1.6 give

$$(a + bi)(x, 0) = (ax - b0,\ a0 + bx) = (ax,\ bx).$$

Call this identity $(*)$.

Spanning. Let $(u, v) \in V_{\mathbb{C}}$. Since $(v_1, \ldots, v_n)$ spans V, there are $a_j, b_j \in \mathbb{R}$ with $u = \sum_j a_j v_j$ and $v = \sum_j b_j v_j$. By $(*)$ and entrywise addition,

$$\sum_{j=1}^{n} (a_j + b_j i)(v_j, 0) = \sum_{j=1}^{n} (a_j v_j,\ b_j v_j) = \Big(\sum_{j=1}^{n} a_j v_j,\ \sum_{j=1}^{n} b_j v_j\Big) = (u, v).$$

Independence. Let $\alpha_1, \ldots, \alpha_n \in \mathbb{C}$ with $\sum_j \alpha_j (v_j, 0) = (0, 0)$, and write $\alpha_j = a_j + b_j i$ with $a_j, b_j \in \mathbb{R}$. By the same computation, $(\sum_j a_j v_j,\ \sum_j b_j v_j) = (0, 0)$. Since $(v_1, \ldots, v_n)$ is independent over $\mathbb{R}$, all $a_j = 0$ and all $b_j = 0$. Hence every $\alpha_j = 0$.

Therefore $(v_1, \ldots, v_n)$ is a basis of $V_{\mathbb{C}}$, and $\dim_{\mathbb{C}} V_{\mathbb{C}} = n$.

■

As a consistency check, restricting $V_{\mathbb{C}}$ back to $\mathbb{R}$ gives $\dim_{\mathbb{R}} V_{\mathbb{C}} = 2n$ by Theorem 1.9.2, which

matches the fact that $V_{\mathbb{C}}$ over $\mathbb{R}$ is two copies of V. The same proof works word for word for an infinite basis set, since every combination involves finitely many vectors.

Example. The Complexification of $\mathbb{R}^n$ Is $\mathbb{C}^n$

Show that the map $(\mathbf{x}, \mathbf{y}) \mapsto \mathbf{x} + i\mathbf{y}$, from $(\mathbb{R}^n)_{\mathbb{C}}$ to $\mathbb{C}^n$, is a bijection that respects addition and multiplication by complex scalars.

Solution. Here $\mathbf{x}+i\mathbf{y}$ is computed in $\mathbb{C}^n$: its k-th entry is $x_k + y_k i$. Every $\mathbf{z} \in \mathbb{C}^n$ has unique real and imaginary parts entry by entry (Definition 0.6.1), so exactly one pair $(\mathbf{x}, \mathbf{y})$ of real vectors maps to $\mathbf{z}$. Hence the map is a bijection.

Addition is entrywise on both sides, so it is respected. For scaling, let $a, b \in \mathbb{R}$. In $(\mathbb{R}^n)_{\mathbb{C}}$, $(a+bi)(\mathbf{x}, \mathbf{y}) = (a\mathbf{x} - b\mathbf{y},\ a\mathbf{y} + b\mathbf{x})$, which maps to the vector with entries $(ax_k - by_k) + (ay_k + bx_k)i$. In $\mathbb{C}^n$, the k-th entry of $(a+bi)(\mathbf{x}+i\mathbf{y})$ is

$$(a+bi)(x_k + y_k i) = (ax_k - by_k) + (ay_k + bx_k)i.$$

The two agree, so the map respects scalar multiplication. It sends the basis $(\mathbf{e}_1, \ldots, \mathbf{e}_n)$ of $(\mathbb{R}^n)_{\mathbb{C}}$ from Theorem 1.9.3 to the standard basis of $\mathbb{C}^n$.

So $(\mathbb{R}^n)_{\mathbb{C}}$ is $\mathbb{C}^n$ in disguise; in Chapter 2 such a bijection will be called an isomorphism. In the same way $(M_n(\mathbb{R}))_{\mathbb{C}}$ is $M_n(\mathbb{C})$, and $(\mathbb{R}[x]_{\le n})_{\mathbb{C}}$ is $\mathbb{C}[x]_{\le n}$. The construction earns its keep in Chapter 9, where real matrices without real eigenvalues, like the rotation above, get complex ones, and in Chapter 12, in the spectral theory of real operators.

1.9.4 Exercises

1.9.4.1 A. Check your understanding

Exercise. A1

(a) Define the degree $[L : K]$ of a subfield $K \subseteq L$.

(b) State the Tower Law.

(c) What is $\dim_{\mathbb{R}} \mathbb{C}^3$? Give a basis.

(d) True or false: every $\mathbb{C}$-subspace of $\mathbb{C}^n$ is an $\mathbb{R}$-subspace. Justify your answer.

(e) What is $\dim_{\mathbb{C}} (M_2(\mathbb{R}))_{\mathbb{C}}$?

Solution.

(a) $[L : K] = \dim_K L$, where L is regarded as a vector space over K by restriction of scalars (Definition 1.9.1), when this dimension is finite.

(b) See Theorem 1.9.1: if (a_i) is a basis of L over K and (v_j) is a basis of V over L, then $(a_i v_j)$ is a basis of V over K, so $\dim_K V = [L : K] \dim_L V$.

(c) 6, with basis $(\mathbf{e}_1, i\mathbf{e}_1, \mathbf{e}_2, i\mathbf{e}_2, \mathbf{e}_3, i\mathbf{e}_3)$, by Theorem 1.9.2.

(d) True. A $\mathbb{C}$-subspace contains $\mathbf{0}$, is closed under addition, and is closed under multiplication by every complex number, in particular by every real number.

(e) 4: $\dim_{\mathbb{R}} M_2(\mathbb{R}) = 4$ (Example (Dimensions of the Running Examples)), and Theorem 1.9.3 keeps the dimension.

1.9.4.2 B. Practice

Exercise. B1: A Tower over $\mathbb{Q}$

Find $\dim_{\mathbb{Q}} \mathbb{Q}(\sqrt{2})^3$ and a basis of $\mathbb{Q}(\sqrt{2})^3$ over $\mathbb{Q}$.

Solution. By Example ($\mathbb{Q}(\sqrt{2})$ Is a Field), $\mathbb{Q}(\sqrt{2})$ is a field, so $\mathbb{Q}(\sqrt{2})^3$ is a vector space over it with basis $(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$ and dimension 3 (Example (Dimensions of the Running Examples)). By Example (Degrees of Some Fields), $(1, \sqrt{2})$ is a basis of $\mathbb{Q}(\sqrt{2})$ over the subfield $\mathbb{Q}$. By Theorem 1.9.1,

$$(\mathbf{e}_1, \sqrt{2}\mathbf{e}_1, \mathbf{e}_2, \sqrt{2}\mathbf{e}_2, \mathbf{e}_3, \sqrt{2}\mathbf{e}_3)$$

is a basis of $\mathbb{Q}(\sqrt{2})^3$ over $\mathbb{Q}$. Hence $\dim_{\mathbb{Q}} \mathbb{Q}(\sqrt{2})^3 = 2 \cdot 3 = 6$.

Exercise. B2: Complex Matrices over $\mathbb{R}$

Let $\mathbf{E}_{11}, \mathbf{E}_{12}, \mathbf{E}_{21}, \mathbf{E}_{22}$ be the matrix units of $M_2(\mathbb{C})$. Find a basis of $M_2(\mathbb{C})$ over $\mathbb{R}$, and write $\mathbf{A} = \begin{pmatrix} 1+i & 2 \\ -i & 3i \end{pmatrix}$ in it.

Solution. The matrix units form a basis of $M_2(\mathbb{C})$ over $\mathbb{C}$ (Example (Standard Bases)). By Theorem 1.9.2,

$$(\mathbf{E}_{11}, i\mathbf{E}_{11}, \mathbf{E}_{12}, i\mathbf{E}_{12}, \mathbf{E}_{21}, i\mathbf{E}_{21}, \mathbf{E}_{22}, i\mathbf{E}_{22})$$

is a basis of $M_2(\mathbb{C})$ over $\mathbb{R}$, so $\dim_{\mathbb{R}} M_2(\mathbb{C}) = 8$. Splitting each entry of $\mathbf{A}$ into real and imaginary parts,

$$\mathbf{A} = 1\mathbf{E}_{11} + 1(i\mathbf{E}_{11}) + 2\mathbf{E}_{12} + 0(i\mathbf{E}_{12}) + 0\mathbf{E}_{21} + (-1)(i\mathbf{E}_{21}) + 0\mathbf{E}_{22} + 3(i\mathbf{E}_{22}).$$

Exercise. B3: Working in a Complexification

Let V be a vector space over $\mathbb{R}$.

(a) Verify that $\alpha(z+w) = \alpha z + \alpha w$ for all $\alpha \in \mathbb{C}$ and $z, w \in V_{\mathbb{C}}$.

(b) In $(\mathbb{R}[x]_{\le 1})_{\mathbb{C}}$, compute $(2-i)((1+x)+ix)$ from Definition 1.9.2, and compare with the product $(2-i)(1+x+ix)$ computed in $\mathbb{C}[x]$.

Solution.

(a) Let $\alpha = a + bi$ with $a, b \in \mathbb{R}$, $z = (u, v)$, and $w = (u', v')$. Then $z + w = (u+u', v+v')$,

and by the distributive law in V and Theorem 1.1.8,

$$\begin{aligned}\alpha(z+w) &= (a(u+u') - b(v+v'),\ a(v+v') + b(u+u')) \\ &= ((au-bv) + (au'-bv'),\ (av+bu) + (av'+bu')),\end{aligned}$$

where we also rearranged the sums using commutativity and associativity of addition in V. The right side is $(au-bv,\ av+bu) + (au'-bv',\ av'+bu') = \alpha z + \alpha w$.

(b) Here $a=2$, $b=-1$, $u=1+x$, and $v=x$. Then $au-bv = 2+2x+x = 2+3x$ and $av+bu = 2x-(1+x) = -1+x$. So the product is $(2+3x)+i(-1+x)$. In $\mathbb{C}[x]$, $(2-i)(1+x+ix) = 2+2x+2ix-i-ix-i^2x = (2+3x)+i(-1+x)$, the same answer.

1.9.4.3 C. Going deeper

Exercise. C1: When Is a Real Subspace Complex?

Let $W \subseteq \mathbb{C}^n$ be a subspace over $\mathbb{R}$.

(a) Prove that W is a subspace over $\mathbb{C}$ if and only if $i\mathbf{w} \in W$ for every $\mathbf{w} \in W$.

(b) Deduce that if W is a subspace over $\mathbb{C}$, then $\dim_{\mathbb{R}} W$ is even.

(c) Show that $W_0 = \{(z, \overline{z}) : z \in \mathbb{C}\} \subseteq \mathbb{C}^2$ is a subspace over $\mathbb{R}$ of even real dimension that is **not** a subspace over $\mathbb{C}$. So the converse of (b) fails.

Solution.

(a) ($\Rightarrow$) If W is a $\mathbb{C}$-subspace, it is closed under multiplication by i.
($\Leftarrow$) Suppose $i\mathbf{w} \in W$ for every $\mathbf{w} \in W$. We use Theorem 1.2.1 over $\mathbb{C}$. Since W is an $\mathbb{R}$-subspace, $\mathbf{0} \in W$ and W is closed under addition. Let $\mathbf{w} \in W$ and $a+bi \in \mathbb{C}$ with $a, b \in \mathbb{R}$. Then $a\mathbf{w} \in W$ and $b(i\mathbf{w}) \in W$, because W is closed under real scalars, and $(a+bi)\mathbf{w} = a\mathbf{w} + b(i\mathbf{w})$ by the axioms of $\mathbb{C}^n$. This lies in W by closure under addition. Hence W is a $\mathbb{C}$-subspace.

(b) W is a subspace of the finite-dimensional $\mathbb{C}$-space $\mathbb{C}^n$, so by Theorem 1.6.7 it is finite-dimensional over $\mathbb{C}$, and by Corollary 1.5.5 it has a basis over $\mathbb{C}$, of some length k. Restricting the scalars of W to $\mathbb{R}$ gives the same real space as before, so $\dim_{\mathbb{R}} W = 2k$ by Theorem 1.9.2.

(c) Write $z = a+bi$ with $a, b \in \mathbb{R}$. Then $(z, \overline{z}) = (a+bi, a-bi) = a(1,1) + b(i,-i)$, so W_0 is the $\mathbb{R}$-span of $((1,1),(i,-i))$ and is an $\mathbb{R}$-subspace by Theorem 1.3.1. The list is independent over $\mathbb{R}$: $a(1,1)+b(i,-i) = \mathbf{0}$ gives $a+bi = 0$, so $a = b = 0$. Hence $\dim_{\mathbb{R}} W_0 = 2$. But $(1,1) \in W_0$ while $i(1,1) = (i,i) \notin W_0$, because $\overline{i} = -i \neq i$. By (a), W_0 is not a $\mathbb{C}$-subspace.

Exercise. C2: A Tower of Degree Four

Let $V = \{x + y\sqrt{3} : x, y \in \mathbb{Q}(\sqrt{2})\} \subseteq \mathbb{R}$.

(a) Prove that $\sqrt{3} \notin \mathbb{Q}(\sqrt{2})$.

(b) Regard $\mathbb{R}$ as a vector space over its subfield $\mathbb{Q}(\sqrt{2})$. Prove that V is a subspace with basis $(1, \sqrt{3})$.

(c) Deduce that $\dim_{\mathbb{Q}} V = 4$, with basis $(1, \sqrt{2}, \sqrt{3}, \sqrt{6})$.

Hint for (a): square both sides, and argue with remainders on division by 3.

Solution.

(a) First, there are no integers p, q with $q \neq 0$, not both divisible by 3, and $cp^2 = 3q^2$ for $c \in \{1, 2\}$. Indeed, if $3 \nmid p$, then $p = 3t \pm 1$ and $cp^2 = c(9t^2 \pm 6t + 1)$ leaves remainder $c \in \{1, 2\}$ on division by 3, while $3q^2$ leaves remainder 0. So $3 \mid p$, say $p = 3s$. Then $9cs^2 = 3q^2$, so $q^2 = 3cs^2$, and the same remainder argument (a square not divisible by 3 leaves remainder 1) gives $3 \mid q$, a contradiction.
Now suppose $\sqrt{3} = a + b\sqrt{2}$ with $a, b \in \mathbb{Q}$. Squaring, $3 = a^2 + 2b^2 + 2ab\sqrt{2}$. If $ab \neq 0$, then $\sqrt{2} = (3 - a^2 - 2b^2)/(2ab) \in \mathbb{Q}$, contradicting Theorem 0.2.3. If $b = 0$, then $a^2 = 3$; writing $a = p/q$ in lowest terms gives $p^2 = 3q^2$ with p, q not both divisible by 3, which is impossible. If $a = 0$, then $2b^2 = 3$, and $b = p/q$ in lowest terms gives $2p^2 = 3q^2$, also impossible. Hence $\sqrt{3} \notin \mathbb{Q}(\sqrt{2})$.

(b) $\mathbb{Q}(\sqrt{2})$ is a subfield of $\mathbb{R}$ by Example ($\mathbb{Q}(\sqrt{2})$ Is a Field), so $\mathbb{R}$ is a vector space over it by restriction of scalars. We use Theorem 1.2.1. (1) $0 = 0 + 0\sqrt{3} \in V$. (2) $(x + y\sqrt{3}) + (x' + y'\sqrt{3}) = (x + x') + (y + y')\sqrt{3} \in V$, since $\mathbb{Q}(\sqrt{2})$ is closed under addition. (3) For $c \in \mathbb{Q}(\sqrt{2})$, $c(x + y\sqrt{3}) = cx + (cy)\sqrt{3} \in V$, since $\mathbb{Q}(\sqrt{2})$ is closed under multiplication. So V is a subspace.
The list $(1, \sqrt{3})$ spans V by the definition of V. For independence, let $x + y\sqrt{3} = 0$ with $x, y \in \mathbb{Q}(\sqrt{2})$. If $y \neq 0$, then $y^{-1} \in \mathbb{Q}(\sqrt{2})$ and $\sqrt{3} = -xy^{-1} \in \mathbb{Q}(\sqrt{2})$, contradicting (a). So $y = 0$, and then $x = 0$. Hence $(1, \sqrt{3})$ is a basis and $\dim_{\mathbb{Q}(\sqrt{2})} V = 2$.

(c) By Example (Degrees of Some Fields), $(1, \sqrt{2})$ is a basis of $\mathbb{Q}(\sqrt{2})$ over $\mathbb{Q}$. By Theorem 1.9.1 applied to $\mathbb{Q} \subseteq \mathbb{Q}(\sqrt{2})$ and the $\mathbb{Q}(\sqrt{2})$-space V, the list $(1 \cdot 1, \sqrt{2} \cdot 1, 1 \cdot \sqrt{3}, \sqrt{2} \cdot \sqrt{3}) = (1, \sqrt{2}, \sqrt{3}, \sqrt{6})$ is a basis of V over $\mathbb{Q}$. Hence $\dim_{\mathbb{Q}} V = 2 \cdot 2 = 4$.

Chapter 2

Linear Maps

Chapter 1 studied vector spaces one at a time. This chapter studies the functions between them that respect the structure: **linear maps**, the functions that send sums to sums and scalar multiples to scalar multiples. Differentiation of polynomials, the trace of a matrix, a rotation of the plane, a shift of a sequence, and multiplication by a matrix are all linear maps, and one theory covers them all.

The theory is built without coordinates, and that is the point of putting it here. Two subspaces measure how far a linear map is from being injective and from being surjective: its kernel and its image. The Rank–Nullity Theorem ties their dimensions to the dimension of the domain, and turns questions that look as though they need a computation into questions of counting — often it is enough to check one of injectivity and surjectivity and let the count supply the other. By the end we can say when two spaces are "the same": finite-dimensional spaces are classified by a single number, their dimension. No basis is chosen anywhere in this chapter; Chapter 4 chooses one and turns all of this into matrices.

What you need. From Chapter 0, functions and what injective, surjective, and bijective mean, composition and inverses (Definition 0.4.3, Definition 0.4.4, Theorem 0.4.3), and the counting fact this chapter has a linear version of: between finite sets of equal size, injective and surjective agree (Theorem 0.4.9). Also field arithmetic (Theorem 0.7.1), polynomials with their degree and evaluation (Definition 0.8.1, Theorem 0.8.3, Theorem 0.8.5), which supply half the examples, and, for §04 alone, polynomials of a matrix (Theorem 0.9.9). From Chapter 1, essentially all of it: vector spaces and the subspace test (Definition 1.1.1, Theorem 1.2.1), span (Definition 1.3.2, Theorem 1.3.1), independence (Definition 1.4.1), bases and coordinates (Definition 1.5.1, Example (Standard Bases), Theorem 1.5.2, Corollary 1.5.5), dimension with basis extension and the two counting theorems that replace a check (Definition 1.6.1, Theorem 1.6.5, Theorem 1.6.4, Theorem 1.6.3, Theorem 1.6.8), and sums and direct sums with the dimension formula (Theorem 1.7.4, Theorem 1.7.3), whose proof pattern is the one Rank–Nullity reuses.

Nothing in this chapter needs Chapter 3. Linear systems, row reduction, and the rank of a matrix come after it, and every proof here is written so that they can.

2.0.1 Roadmap

- **Linear maps.** The definition, a gallery of examples and nonexamples drawn from geometry, matrices, calculus, and algebra, and the fact that makes linear maps manageable: a linear map is determined by, and may be freely prescribed on, a basis.
- **Kernel and image.** The two subspaces attached to every linear map, how to compute an image from a spanning list, and the kernel test that makes injectivity cheap to check.

- **The Rank–Nullity Theorem.** $\operatorname{nullity} T + \operatorname{rank} T = \dim V$, proved by taking a basis of the kernel and extending it, and the corollaries that let a count replace a check.
- **The algebra of linear maps.** Adding, scaling, and composing maps; $\mathcal{L}(V, W)$ as a vector space and $\mathcal{L}(V)$ as an algebra; polynomials in an operator, and the relations they satisfy.
- **Invertible maps and isomorphisms.** When two spaces are "the same", why dimension alone decides it in finite dimension, and the Invertible Operator Theorem, which collects nine ways to recognize an invertible operator.

2.0.2 Named moves

- **Injective via the kernel.** To show T is injective, show that $Tv = 0$ forces $v = 0$. One quantifier instead of two.
- **Count instead of check.** When $\dim V = \dim W$ is finite, injective, surjective, and bijective are the same property; check whichever is easier, usually the kernel.
- **Basis of the kernel, then extend.** Take a basis of the smaller space, extend it to the whole, claim the added vectors do the remaining work, and count. This proves Rank–Nullity, and it proved the dimension formula for sums before it.
- **One equation is enough.** Between spaces of equal finite dimension, $ST = \mathrm{id}$ alone makes both maps invertible — the counting move again, in disguise.
- **Transport along an isomorphism.** An isomorphism carries bases to bases and preserves every statement written in terms of sums and scalar multiples, so a question about V may be asked in any space isomorphic to it.

2.1 Linear Maps

Chapter 1 studied vector spaces one at a time. This chapter studies the functions **between** vector spaces that respect addition and scaling. This section defines them, collects examples from geometry, matrices, calculus, and algebra, and proves the fact that makes them manageable: a linear map is completely determined by what it does to a basis.

2.1.1 Functions that behave like matrices

Objects come first, then the maps between them that respect their structure. For sets the maps are functions; for groups they are homomorphisms (Definition 0.10.3). For vector spaces the structure is addition and scalar multiplication, so we want functions that respect those two operations.

The standard example is a matrix. For $\mathbf{A} \in M_{m\times n}(F)$, the rules of matrix multiplication (Theorem 0.9.4) give

$$\mathbf{A}(\mathbf{x}+\mathbf{y}) = \mathbf{A}\mathbf{x} + \mathbf{A}\mathbf{y}, \qquad \mathbf{A}(c\mathbf{x}) = c(\mathbf{A}\mathbf{x}) \qquad \text{for all } \mathbf{x}, \mathbf{y} \in F^n,\ c \in F.$$

Now look at three functions that are not given by matrices at all.

- **Differentiation** on polynomials: $(p+q)' = p' + q'$ and $(cp)' = cp'$.
- **Trace** on square matrices: $\operatorname{tr}(\mathbf{A}+\mathbf{B}) = \operatorname{tr}\mathbf{A} + \operatorname{tr}\mathbf{B}$ and $\operatorname{tr}(c\mathbf{A}) = c\operatorname{tr}\mathbf{A}$ (Theorem 0.9.6).
- **Rotation** of the plane about the origin: rotating a parallelogram gives a parallelogram, so the rotation of a sum is the sum of the rotations, and rotating a stretched arrow gives the stretched rotated arrow.

The inputs are polynomials, matrices, and arrows, not columns. Yet all three obey exactly the two rules that matrices obey. Whatever we can prove from those two rules alone will hold for all of them at once. So we give the two rules a name.

A linear map is a function that preserves linear combinations.

2.1.2 The definition

Definition 2.1.1. Linear Map

Let V and W be vector spaces **over the same field** F. A function $T\colon V \to W$ is a **linear map** (or **linear transformation**) **over** F if

(LT1) $T(u+v) = T(u) + T(v)$ **for all** $u, v \in V$, and
(LT2) $T(cv) = cT(v)$ **for all** $c \in F$ and **all** $v \in V$.

A linear map $T\colon V \to V$ from a space to itself is also called a **linear operator** on V. We often write Tv for $T(v)$.

In words: (LT1), **additivity**, says that adding first and then applying T gives the same result as applying T first and then adding. The sum on the left is taken in V; the sum on the right is taken in W. (LT2), **homogeneity**, says the same for scaling: the scalar c can be pulled out of T. Both conditions must hold for **every** choice of vectors and scalars, so one counterexample is enough to show that a function is not linear.

The phrase "over the same field" matters. Scalars from F act on both sides of (LT2), so both spaces must accept them. We will see below that one function can be linear over $\mathbb{R}$ and not linear over $\mathbb{C}$. A general linear map is always named T, S, or R in this book, never f; a specific map may get a letter that recalls what it does, such as D for differentiation.

2.1.3 Examples

We check the examples against (LT1) and (LT2), grouped by where they come from.

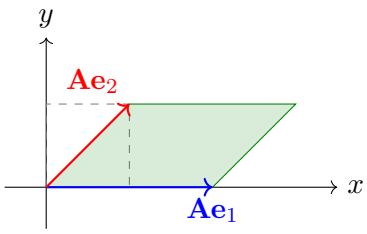
Matrices.

Example. Multiplication by a Matrix

Let $\mathbf{A} \in M_{m\times n}(F)$. Show that $T_{\mathbf{A}}\colon F^n \to F^m$, $T_{\mathbf{A}}(\mathbf{x}) = \mathbf{A}\mathbf{x}$, is a linear map.

Solution. Let $\mathbf{x}, \mathbf{y} \in F^n$ and $c \in F$. By the distributive law in Theorem 0.9.4, $T_{\mathbf{A}}(\mathbf{x}+\mathbf{y}) = \mathbf{A}(\mathbf{x}+\mathbf{y}) = \mathbf{A}\mathbf{x} + \mathbf{A}\mathbf{y} = T_{\mathbf{A}}(\mathbf{x}) + T_{\mathbf{A}}(\mathbf{y})$, which is (LT1). By the scalar rule in the same theorem, $T_{\mathbf{A}}(c\mathbf{x}) = \mathbf{A}(c\mathbf{x}) = c(\mathbf{A}\mathbf{x}) = cT_{\mathbf{A}}(\mathbf{x})$, which is (LT2). Hence $T_{\mathbf{A}}$ is linear.

We will use the name $T_{\mathbf{A}}$ for this map throughout the chapter. For $F = \mathbb{R}$ and $n = m = 2$, the map $T_{\mathbf{A}}$ sends the unit square to the parallelogram spanned by $\mathbf{A}\mathbf{e}_1$ and $\mathbf{A}\mathbf{e}_2$, the columns of $\mathbf{A}$ (Theorem 0.9.3).



The unit square (dashed) and its image under $\mathbf{A} = \begin{pmatrix} 2 & 1 \\ 0 & 1 \end{pmatrix}$.

Example. Maps Given by Formulas in Coordinates

(a) Show that $T\colon F^3 \to F^2$, $T(a_1, a_2, a_3) = (2a_1 + a_3,\ a_2 - a_1)$, is linear.

(b) Let $\mathcal{B} = (v_1, \ldots, v_n)$ be a basis of a vector space V over F. Show that the **coordinate map** $V \to F^n$, $v \mapsto [v]_{\mathcal{B}}$, is linear.

Solution.

(a) The formula is a matrix product:

$$\begin{pmatrix} 2 & 0 & 1 \\ -1 & 1 & 0 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} = \begin{pmatrix} 2a_1 + a_3 \\ -a_1 + a_2 \end{pmatrix}.$$

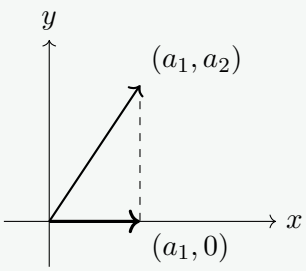
So $T = T_{\mathbf{A}}$ for this 2×3 matrix $\mathbf{A}$, and T is linear by Example (Multiplication by a Matrix). The same argument shows that every map $F^n \to F^m$ whose output entries are combinations of the input entries, with constant coefficients and no constant terms, is linear.

(b) By Theorem 1.5.2, $[u+v]_{\mathcal{B}} = [u]_{\mathcal{B}} + [v]_{\mathcal{B}}$ and $[cv]_{\mathcal{B}} = c[v]_{\mathcal{B}}$. These are (LT1) and (LT2).

Geometry.

Example. Projection onto an Axis

Show that $T\colon \mathbb{R}^2 \to \mathbb{R}^2$, $T(a_1, a_2) = (a_1, 0)$, is linear.



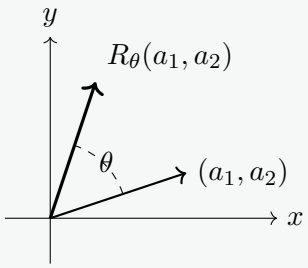
Solution. Let $\mathbf{u} = (a_1, a_2)$, $\mathbf{v} = (b_1, b_2)$, and $c \in \mathbb{R}$. Then

$$T(\mathbf{u}+\mathbf{v}) = (a_1+b_1, 0) = (a_1, 0) + (b_1, 0) = T(\mathbf{u}) + T(\mathbf{v}),$$
$$T(c\mathbf{u}) = (ca_1, 0) = c(a_1, 0) = cT(\mathbf{u}).$$

So (LT1) and (LT2) hold. The map drops each point vertically onto the x-axis, like a shadow cast by light from straight above.

Example. Rotation of the Plane

Let $\theta \in \mathbb{R}$, and let $R_\theta \colon \mathbb{R}^2 \to \mathbb{R}^2$ rotate each vector counterclockwise about the origin by the angle θ. Find a formula for R_θ, and deduce that it is linear.



Solution. Write $(a_1, a_2) = (r\cos\varphi, r\sin\varphi)$ in polar coordinates. Rotating adds θ to the angle and keeps the length r, so by the addition formulas for sine and cosine,

$$\begin{aligned} R_\theta(a_1, a_2) &= (r\cos(\varphi+\theta),\ r\sin(\varphi+\theta)) \\ &= (a_1\cos\theta - a_2\sin\theta,\ a_1\sin\theta + a_2\cos\theta). \end{aligned}$$

This is $T_{\mathbf{A}}$ for $\mathbf{A} = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$, so R_θ is linear by Example (Multiplication by a Matrix).

Calculus. For a polynomial $p = a_0 + a_1 x + \cdots + a_n x^n \in F[x]$, define its **(formal) derivative** as $p' = a_1 + 2a_2 x + \cdots + n a_n x^{n-1}$, where $k a_k$ means a_k added to itself k times. Over $\mathbb{R}$ this is the derivative of calculus.

Example. Differentiation

Show that $D \colon F[x] \to F[x]$, $D(p) = p'$, is linear. Show also that D maps $F[x]_{\le n}$ into $F[x]_{\le n-1}$ for $n \ge 1$.

Solution. Let $p = \sum_k a_k x^k$ and $q = \sum_k b_k x^k$, and $c \in F$. The coefficient of x^{k-1} in $D(p+q)$ is $k(a_k + b_k) = ka_k + kb_k$, which is the coefficient of x^{k-1} in $D(p) + D(q)$. The coefficient of x^{k-1} in $D(cp)$ is $k(ca_k) = c(ka_k)$, which is the coefficient in $cD(p)$. Both steps use only the field axioms. Two polynomials with the same coefficients are equal (Definition 0.8.1), so (LT1) and (LT2) hold. If $\deg p \le n$, the highest power in p' is at most x^{n-1}, so D restricts to a linear map $F[x]_{\le n} \to F[x]_{\le n-1}$.

Example. Integration

Show that $J\colon \mathbb{R}[x] \to \mathbb{R}[x]$, $J(p)(x) = \int_0^x p(t)\,\mathrm{d}t$, is linear.

Solution. On coefficients, J sends $a_0 + a_1x + \cdots + a_nx^n$ to $a_0x + \frac{a_1}{2}x^2 + \cdots + \frac{a_n}{n+1}x^{n+1}$, so it does map polynomials to polynomials. By linearity of the integral from calculus,

$$J(p+q)(x) = \int_0^x (p(t)+q(t))\,\mathrm{d}t = J(p)(x) + J(q)(x),$$
$$J(cp)(x) = \int_0^x cp(t)\,\mathrm{d}t = cJ(p)(x)$$

for every $x \in \mathbb{R}$. Hence J is linear. (The coefficient formula divides by $k+1$; over a field such as $\mathbb{F}_2$, where $1+1=0$, it breaks down.)

Algebra and sequences.

Example. Multiplication by x

Show that $M\colon F[x] \to F[x]$, $M(p) = xp$, is linear.

Solution. By the distributive laws in Theorem 0.8.1, $x(p+q) = xp + xq$, and $x(cp) = c(xp)$ because the product of polynomials is associative and commutative and the constant c is a polynomial. These are (LT1) and (LT2).

Example. Trace and Transpose

Show that the trace $\operatorname{tr}\colon M_n(F) \to F$ and the transpose $M_{m\times n}(F) \to M_{n\times m}(F)$, $\mathbf{A} \mapsto \mathbf{A}^\top$, are linear.

Solution. For the trace, Theorem 0.9.6 (1) says $\operatorname{tr}(\mathbf{A}+\mathbf{B}) = \operatorname{tr}\mathbf{A} + \operatorname{tr}\mathbf{B}$ and $\operatorname{tr}(c\mathbf{A}) = c\operatorname{tr}\mathbf{A}$, which are (LT1) and (LT2); the target is F, regarded as a vector space over itself. For the transpose, the (i,j) entry of $(\mathbf{A}+\mathbf{B})^\top$ is $a_{ji}+b_{ji}$, the (i,j) entry of $\mathbf{A}^\top + \mathbf{B}^\top$; and the (i,j) entry of $(c\mathbf{A})^\top$ is ca_{ji}, the (i,j) entry of $c\mathbf{A}^\top$.

Example. The Shift Maps on Sequences

On the space $F^{\mathbb{N}}$ of sequences (Example (The Standard Vector Spaces)), define the **right shift** and the **left shift**

$$R(s_0, s_1, s_2, \dots) = (0, s_0, s_1, \dots), \qquad L(s_0, s_1, s_2, \dots) = (s_1, s_2, s_3, \dots).$$

Show that both are linear.

Solution. The operations on $F^{\mathbb{N}}$ are termwise. The entry in position $k \geq 1$ of $R(s+t)$ is $s_{k-1}+t_{k-1}$, which is the entry of $R(s)+R(t)$; in position 0 both have $0 = 0+0$. Similarly, $R(cs)$ has entries $0 = c \cdot 0$ and cs_{k-1}, the entries of $cR(s)$. For L, the entry in position k of $L(s+t)$ is $s_{k+1}+t_{k+1}$, and of $L(cs)$ is cs_{k+1}, as required.

The right shift is the map promised in the warning of Chapter 1, §8, on what changes in infinite dimension: it matches $F^{\mathbb{N}}$ one-to-one with the sequences whose first entry is 0, respecting addition and scaling. In the next section we measure exactly how R and L fail to be bijective.

Example. Composition with a Fixed Function

Let X and Y be sets and $g\colon X \to Y$ a function. Show that $C_g\colon F^Y \to F^X$, $C_g(h) = h \circ g$, is linear. Show also that $\mathbb{R}[x] \to \mathbb{R}[x]$, $p(x) \mapsto p(x+1)$, is linear.

Solution. The operations on function spaces are pointwise (Example (The Standard Vector Spaces)). For $t \in X$,

$$C_g(h+k)(t) = (h+k)(g(t)) = h(g(t)) + k(g(t)) = \big(C_g(h) + C_g(k)\big)(t),$$

and $C_g(ch)(t) = c\,h(g(t)) = \big(cC_g(h)\big)(t)$. So (LT1) and (LT2) hold. The output is again a function, not a number; linearity is about the function h varying, while g stays fixed.

For the second claim, write $\sigma\colon \mathbb{R}[x] \to \mathbb{R}[x]$ for $p(x) \mapsto p(x+1)$; in full, if $p = \sum_k a_k x^k$, then $\sigma(p) = \sum_k a_k (x+1)^k$, a finite sum in $\mathbb{R}[x]$. Linearity is now read off the coefficients. The coefficients of $p+q$ are the $a_k + b_k$, so $\sigma(p+q) = \sum_k (a_k+b_k)(x+1)^k = \sigma(p) + \sigma(q)$ by Theorem 0.8.1, and $\sigma(cp) = \sum_k ca_k(x+1)^k = c\,\sigma(p)$ in the same way. So σ is linear. (It translates the graph of p one unit to the left.) One can also read σ as an instance of C_g with $g(t) = t+1$, once a polynomial and the function it defines are identified; that identification needs an argument, which Chapter 3 supplies.

The degenerate maps. Three maps are always available: the identity on any space, the inclusion of any subspace, and the zero map between any two spaces. They look too simple to matter. They matter because every general statement about linear maps must be true for them, so they are the first test of any conjecture.

Example. Identity Map

The identity $\mathrm{id}_V\colon V \to V$, $v \mapsto v$, is linear: $\mathrm{id}_V(u+v) = u+v = \mathrm{id}_V(u) + \mathrm{id}_V(v)$ and $\mathrm{id}_V(cv) = cv = c\,\mathrm{id}_V(v)$.

Example. Inclusion Map

If U is a subspace of V, the inclusion $\iota\colon U \to V$, $\iota(u) = u$, is linear, because the operations of U are those of V. Its formula is the identity's, but its codomain is bigger. For instance $\mathbb{R}[x]_{\leq 2} \to \mathbb{R}[x]$ is an inclusion.

Example. Zero Map

For any V and W over F, the zero map $0\colon V \to W$, $v \mapsto 0_W$, is linear: $0_W = 0_W + 0_W$ and $0_W = c0_W$ (Theorem 1.1.6).

2.1.4 Three consequences of linearity

The two axioms already force several other rules. The first is the quickest test for nonlinearity.

Theorem 2.1.1. Linear Maps Send Zero to Zero

Let $T\colon V \to W$ be linear. Then $T(0_V) = 0_W$.

Proof. By (LT1), $T(0_V) = T(0_V + 0_V) = T(0_V) + T(0_V)$. Adding $-T(0_V)$ to both sides gives $0_W = T(0_V)$. ■

The proof used only (LT1). Homogeneity with $c = 0$ would also give it: $T(0_V) = T(0v) = 0T(v) = 0_W$.

Theorem 2.1.2. Linear Maps Preserve Negatives

Let $T\colon V \to W$ be linear. Then $T(-v) = -T(v)$ and $T(u-v) = T(u) - T(v)$ for all $u, v \in V$.

Proof. By Theorem 1.1.7, $-v = (-1)v$. So (LT2) gives $T(-v) = (-1)T(v) = -T(v)$, using Theorem 1.1.7 again in W. Then (LT1) gives $T(u - v) = T(u + (-v)) = T(u) + T(-v) = T(u) - T(v)$. ■

Checking two conditions separately is often longer than necessary. One combined condition suffices.

Theorem 2.1.3. One-Line Test for Linearity

Let V and W be vector spaces over F, and let $T\colon V \to W$ be a function. Then T is linear **if and only if**

$$T(cu + v) = cT(u) + T(v) \qquad \text{for all } c \in F \text{ and all } u, v \in V.$$

Proof. ($\Rightarrow$) Suppose T is linear. By (LT1) and then (LT2), $T(cu + v) = T(cu) + T(v) = cT(u) + T(v)$.

($\Leftarrow$) Suppose the condition holds. Taking $c = 1$ gives $T(u + v) = T(u) + T(v)$, which is (LT1). Taking $u = v = 0$ and $c = 1$ gives $T(0) = T(0) + T(0)$, so $T(0) = 0$. Now taking $v = 0$ gives $T(cu) = cT(u) + T(0) = cT(u)$, which is (LT2).

■

Finally, the slogan. Two applications of the axioms handle a combination of two vectors; induction handles any finite combination.

Theorem 2.1.4. Linear Maps Preserve Linear Combinations

Let $T\colon V \to W$ be linear, let $v_1, \dots, v_k \in V$ and $a_1, \dots, a_k \in F$, where $k \geq 0$. Then

$$T(a_1 v_1 + \cdots + a_k v_k) = a_1 T(v_1) + \cdots + a_k T(v_k).$$

Proof. We use induction on k. For $k = 0$ both sides are empty sums, so the claim is $T(0) = 0$, which is Theorem 2.1.1. Let $k \geq 1$ and suppose the claim holds for $k-1$ vectors. By (LT1), then the induction hypothesis and (LT2),

$$T\Big(\sum_{i=1}^{k-1} a_i v_i + a_k v_k\Big) = T\Big(\sum_{i=1}^{k-1} a_i v_i\Big) + T(a_k v_k) = \sum_{i=1}^{k-1} a_i T(v_i) + a_k T(v_k).$$

This completes the induction.

■

In words: to apply T to a combination, apply it to each vector and keep the same coefficients. This is the property the slogan promised, and it is the property every later theorem about linear maps uses.

2.1.5 Nonexamples

A function fails to be linear as soon as **one** instance of (LT1) or (LT2) fails. The cheapest instance to try is 0, by Theorem 2.1.1.

Example. A Constant Function

Is $T\colon \mathbb{R} \to \mathbb{R}$, $T(x) = 5$, linear?

Solution. No. A linear map sends 0 to 0 by Theorem 2.1.1, but $T(0) = 5 \neq 0$.

Example. Translation

Let $\mathbf{b} \in \mathbb{R}^2$ be **nonzero**. Is the translation $T\colon \mathbb{R}^2 \to \mathbb{R}^2$, $T(\mathbf{v}) = \mathbf{v} + \mathbf{b}$, linear?

Solution. No: $T(\mathbf{0}) = \mathbf{b} \neq \mathbf{0}$, which is impossible for a linear map by Theorem 2.1.1. This is a minimal change of the identity map (Example (Identity Map)): the formula differs only by the constant $\mathbf{b}$, and that constant is exactly what breaks linearity. For the record, (LT1) fails

too: $T(\mathbf{u}+\mathbf{v}) = \mathbf{u}+\mathbf{v}+\mathbf{b}$, while $T(\mathbf{u})+T(\mathbf{v}) = \mathbf{u}+\mathbf{v}+2\mathbf{b}$. Maps of the form $\mathbf{v} \mapsto \mathbf{A}\mathbf{v}+\mathbf{b}$ are called **affine**; they are linear exactly when $\mathbf{b} = \mathbf{0}$.

Passing the zero test does not make a function linear.

Example. Squaring

Is $T\colon \mathbb{R}^2 \to \mathbb{R}^2$, $T(a_1, a_2) = (a_1^2, 0)$, linear?

Solution. No. Here $T(\mathbf{0}) = \mathbf{0}$, so the quick test says nothing. But (LT2) fails for $\mathbf{v} = (1,0)$ and $c = 2$: $T(2\mathbf{v}) = T(2,0) = (4,0)$, while $2T(\mathbf{v}) = 2(1,0) = (2,0)$. Compare with the projection $(a_1, a_2) \mapsto (a_1, 0)$ of Example (Projection onto an Axis): squaring the first entry is the one change, and it destroys homogeneity.

Example. Length

Is $T\colon \mathbb{R}^2 \to \mathbb{R}$, $T(x,y) = \sqrt{x^2+y^2}$, linear?

Solution. No. Take $\mathbf{v} = (1,0)$ and $c = -1$. Then $T(-\mathbf{v}) = 1$, but $-T(\mathbf{v}) = -1$, so (LT2) fails. Length does satisfy $T(c\mathbf{v}) = |c|T(\mathbf{v})$, which agrees with $cT(\mathbf{v})$ only for $c \geq 0$. Additivity fails too: $T((1,0)+(0,1)) = \sqrt{2}$, but $T(1,0)+T(0,1) = 2$.

The last nonexample shows that linearity depends on the field, not only on the function.

Example. Complex Conjugation: the Field Matters

Let $T\colon \mathbb{C} \to \mathbb{C}$, $T(z) = \overline{z}$. Is T linear when $\mathbb{C}$ is regarded as a vector space **over** $\mathbb{C}$? Over $\mathbb{R}$?

Solution. (LT1) holds in both cases, since $\overline{z+w} = \overline{z}+\overline{w}$ (Theorem 0.6.2).

Over $\mathbb{C}$: not linear. The scalars are all complex numbers. Take $c = i$ and $z = 1$: then $T(i \cdot 1) = \overline{i} = -i$, but $iT(1) = i \cdot 1 = i$. So (LT2) fails.

Over $\mathbb{R}$: linear. Now (LT2) is only required for real c. For $c \in \mathbb{R}$, $T(cz) = \overline{c}\,\overline{z} = c\,\overline{z} = cT(z)$, by Theorem 0.6.2 and $\overline{c} = c$.

Warning. **"Linear" here is not the "linear function" of school.** A function $f(x) = mx+b$ has a straight-line graph, but it is a linear map $\mathbb{R} \to \mathbb{R}$ only when $b = 0$: $f(0) = b$, and Theorem 2.1.1 demands $f(0) = 0$. In fact every linear map $T\colon \mathbb{R} \to \mathbb{R}$ has the form $T(x) = mx$: by (LT2), $T(x) = T(x \cdot 1) = xT(1)$, so $m = T(1)$. And $T(0) = 0$ is **necessary but not sufficient** for linearity, as Example (Squaring) shows.

Quick check. Which of these maps $\mathbb{R}^2 \to \mathbb{R}^2$ are linear? (i) $(x, y) \mapsto (y, x)$; (ii) $(x, y) \mapsto (x + y, 1)$; (iii) $(x, y) \mapsto (xy, 0)$.

Solution.

(i) Linear: it is $T_{\mathbf{A}}$ with $\mathbf{A} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$. (ii) Not linear: $(0, 0) \mapsto (0, 1) \neq \mathbf{0}$. (iii) Not linear, although $\mathbf{0} \mapsto \mathbf{0}$: $(2, 2) \mapsto (4, 0)$, but $2 \cdot T(1, 1) = 2(1, 0) = (2, 0)$, so (LT2) fails.

Why this definition. Could one condition be dropped? Complex conjugation on $\mathbb{C}$ over $\mathbb{C}$ satisfies (LT1) but not (LT2), so additivity alone is not enough. (Over $\mathbb{Q}$, surprisingly, additivity does imply homogeneity; see Exercise (C1: Over $\mathbb{Q}$, Additivity Is Enough).) Conversely, $T\colon \mathbb{R}^2 \to \mathbb{R}$, $T(x, y) = \sqrt[3]{x^3 + y^3}$ (the real cube root), satisfies (LT2), since $\sqrt[3]{c^3x^3 + c^3y^3} = c\sqrt[3]{x^3 + y^3}$ for every real c. But $T(1, 0) + T(0, 1) = 2$, while $T(1, 1) = \sqrt[3]{2}$, so (LT1) fails. Neither condition implies the other, and we need both to get Theorem 2.1.4. The name comes from the case $\mathbb{R} \to \mathbb{R}$ just described: the graph of a linear map is a line through the origin.

2.1.6 A linear map is determined by a basis

A function on an infinite set normally needs infinitely many values to describe it. A linear map needs far fewer. If $\mathcal{B} = (v_1, \ldots, v_n)$ is a basis of V, every $v \in V$ is a combination $a_1v_1 + \cdots + a_nv_n$, and Theorem 2.1.4 forces

$$T(v) = a_1T(v_1) + \cdots + a_nT(v_n).$$

So the n vectors $T(v_1), \ldots, T(v_n)$ already fix every value of T. The converse question is whether **any** choice of these n vectors comes from a linear map. It does, and that is the theorem.

Theorem 2.1.5. Linear Maps Are Determined by a Basis

Let V and W be vector spaces over F, let $\mathcal{B} = (v_1, \ldots, v_n)$ be a basis of V, and let $w_1, \ldots, w_n \in W$ be **any** vectors. Then there **exists exactly one** linear map $T\colon V \to W$ such that

$$T(v_i) = w_i \qquad \text{for } i = 1, \ldots, n.$$

It is given by $T(a_1v_1 + \cdots + a_nv_n) = a_1w_1 + \cdots + a_nw_n$ for all $a_1, \ldots, a_n \in F$.

Idea. **Uniqueness** is the computation above: any linear T with these values must be given by the formula. **Existence** asks us to *use* the formula as a definition. The only danger is that a vector might have two expressions $\sum a_iv_i$, giving two candidate values; the basis rules that out, because coordinates are unique. So: ① define $T(v)$ from the coordinates of v; ② check (LT1), (LT2), using that coordinates respect the operations; ③ check $T(v_i) = w_i$; ④ prove uniqueness by Theorem 2.1.4.

Proof. Existence. Let $v \in V$. By Theorem 1.5.1, there are **unique** scalars $a_1, \ldots, a_n \in F$ with $v = a_1v_1 + \cdots + a_nv_n$; they are the entries of $[v]_{\mathcal{B}}$ (Definition 1.5.2). Define

$$T(v) := a_1w_1 + \cdots + a_nw_n.$$

Since the scalars are unique, this assigns exactly one vector of W to each $v \in V$, so $T\colon V \to W$ is a well-defined function.

Let $u, v \in V$ and $c \in F$, with $[u]_{\mathcal{B}} = (b_1, \dots, b_n)$ and $[v]_{\mathcal{B}} = (a_1, \dots, a_n)$. By Theorem 1.5.2, $[cu + v]_{\mathcal{B}} = c[u]_{\mathcal{B}} + [v]_{\mathcal{B}} = (cb_1 + a_1, \dots, cb_n + a_n)$. Hence, by the definition of T and the vector space axioms in W,

$$T(cu + v) = \sum_{i=1}^{n} (cb_i + a_i) w_i = c \sum_{i=1}^{n} b_i w_i + \sum_{i=1}^{n} a_i w_i = cT(u) + T(v).$$

By Theorem 2.1.3, T is linear. Moreover, $v_i = 0v_1 + \cdots + 1v_i + \cdots + 0v_n$, so $[v_i]_{\mathcal{B}} = \mathbf{e}_i$ and $T(v_i) = w_i$.

Uniqueness. Let $S\colon V \to W$ be any linear map with $S(v_i) = w_i$ for all i. For $v = a_1 v_1 + \cdots + a_n v_n \in V$, Theorem 2.1.4 gives

$$S(v) = a_1 S(v_1) + \cdots + a_n S(v_n) = a_1 w_1 + \cdots + a_n w_n = T(v).$$

Since $S(v) = T(v)$ for every $v \in V$, the functions are equal: $S = T$. This proves the theorem.

■

The theorem has two halves, and they are used differently. Uniqueness says that **to check two linear maps are equal, check them on a basis**. Existence says that **to build a linear map, choose the images of a basis freely**. The same holds for the infinite bases of Chapter 1, §8, because every vector is still a unique **finite** combination of basis vectors; we record this as Theorem 2.1.6 below. In Chapter 4 this theorem turns into the matrix of a linear map: once bases are fixed, the n vectors $T(v_i)$, written in coordinates, are the columns.

Theorem 2.1.6. Linear Maps on an Arbitrary Basis

Let V and W be vector spaces over F, let B be a basis of V (possibly infinite, in the sense of Theorem 1.8.2), and for each $b \in B$ let $w_b \in W$ be any vector. Then there is exactly one linear map $T\colon V \to W$ with $Tb = w_b$ for every $b \in B$.

Proof. Let $v \in V$. Since B spans V and is independent, v is a finite combination $v = \sum_{b \in B} a_b b$ in which only finitely many a_b are nonzero, and these coefficients are unique: two such expressions differ by a finite combination of distinct elements of B equal to 0, whose coefficients all vanish by independence. Define $Tv = \sum_{b \in B} a_b w_b$, a finite sum. The coefficients of $v + u$ and of cv are $a_b + a'_b$ and ca_b by uniqueness, so T is linear, and $Tb = w_b$ because the coefficients of b are 1 at b and 0 elsewhere. If S is another linear map with $Sb = w_b$ for all b, then $Sv = \sum a_b Sb = Tv$ by Theorem 2.1.4. This proves the theorem.

■

Example. Finding a Formula from Values on a Basis

Show that there is exactly one linear map $T\colon \mathbb{R}^2 \to \mathbb{R}^3$ with

$$T(1, 1) = (1, 2, 0), \qquad T(1, -1) = (3, 0, 2),$$

and find $T(x, y)$.

Solution. First, $((1,1),(1,-1))$ is a basis of $\mathbb{R}^2$. It is independent: if $a(1,1)+b(1,-1)=\mathbf{0}$, then $a+b=0$ and $a-b=0$, so $a=b=0$. It has $2=\dim\mathbb{R}^2$ entries, so it is a basis (Theorem 1.6.4). By Theorem 2.1.5, exactly one such T exists.

To find it, write (x,y) in the basis. Solving $a(1,1)+b(1,-1)=(x,y)$, that is $a+b=x$ and $a-b=y$, gives $a=\frac{x+y}{2}$ and $b=\frac{x-y}{2}$. By the formula in the theorem,

$$T(x,y)=\frac{x+y}{2}(1,2,0)+\frac{x-y}{2}(3,0,2)=(2x-y,\ x+y,\ x-y).$$

Check: $T(1,1)=(1,2,0)$ and $T(1,-1)=(3,0,2)$.

Example. Too Many Conditions

Is there a linear map $T\colon \mathbb{R}^2\to\mathbb{R}^3$ with

$$T(1,0)=(2,3,4),\qquad T(1,1)=(3,4,5),\qquad T(2,3)=(4,5,6)?$$

Solution. No. As in Example (Finding a Formula from Values on a Basis), $((1,0),(1,1))$ is a basis of $\mathbb{R}^2$: if $a(1,0)+b(1,1)=\mathbf{0}$ then $b=0$ and $a+b=0$, so $a=b=0$, and two independent vectors in a 2-dimensional space form a basis (Theorem 1.6.4). By Theorem 2.1.5, the first two conditions already determine T, and the third condition must be checked against them, not imposed. Since $(2,3)=-1\cdot(1,0)+3\cdot(1,1)$, any linear T with the first two values satisfies

$$T(2,3)=-T(1,0)+3T(1,1)=-(2,3,4)+3(3,4,5)=(7,9,11)$$

by Theorem 2.1.4. As $(7,9,11)\neq(4,5,6)$, no such T exists. Had the third value been $(7,9,11)$, the answer would be yes, with exactly one such map.

Example. Recognizing Differentiation

Show that the differentiation map $D\colon \mathbb{R}[x]_{\leq 3}\to\mathbb{R}[x]_{\leq 2}$ is the **only** linear map $T\colon \mathbb{R}[x]_{\leq 3}\to\mathbb{R}[x]_{\leq 2}$ with

$$T(1)=0,\qquad T(x)=1,\qquad T(x^2)=2x,\qquad T(x^3)=3x^2.$$

Solution. The list $(1,x,x^2,x^3)$ is a basis of $\mathbb{R}[x]_{\leq 3}$ (Example (Standard Bases)). The map D is linear (Example (Differentiation)), and $D(1)=0$, $D(x)=1$, $D(x^2)=2x$, $D(x^3)=3x^2$, so D has the four prescribed values. By the uniqueness half of Theorem 2.1.5, every linear T with these values equals D. Four values, the derivatives of the monomials, pin down the derivative of every cubic.

Warning. **The values must be prescribed on a basis, or at least on an independent list.** On a dependent list, the values may contradict each other. There is **no** linear $T\colon \mathbb{R}^2\to$

$\mathbb{R}^2$ with $T(1,0) = (1,0)$ and $T(2,0) = (0,1)$: since $(2,0) = 2(1,0)$, (LT2) forces $T(2,0) = 2T(1,0) = (2,0) \neq (0,1)$. On the other hand, a list that does not span leaves T undetermined: many linear maps satisfy $T(1,0) = (1,0)$ alone.

Quick check. Let $T\colon \mathbb{R}^2 \to \mathbb{R}$ be linear with $T(1,2) = 3$ and $T(2,5) = 1$. Find $T(x,y)$.

Solution. Solving $a(1,2)+b(2,5) = (x,y)$, that is $a+2b = x$ and $2a+5b = y$, gives $b = y-2x$ and $a = 5x-2y$. Every (x,y) has exactly one such expression, so $((1,2),(2,5))$ is a basis of $\mathbb{R}^2$. Hence $T(x,y) = 3(5x-2y)+1\cdot(y-2x) = 13x-5y$. Check: $13-10 = 3$ and $26-25 = 1$.

We now have a supply of linear maps and a way to build new ones from a basis. The next section attaches to every linear map two subspaces that measure how far it is from being injective and from being surjective.

2.1.7 Exercises

2.1.7.1 A. Check your understanding

Exercise. A1

(a) Define a linear map $T\colon V \to W$ over F.

(b) State the theorem that says a linear map is determined by a basis, including both existence and uniqueness.

(c) True or false: if $T\colon V \to W$ satisfies $T(0) = 0$, then T is linear. Justify your answer.

(d) True or false: there is a linear map $T\colon \mathbb{R}^2 \to \mathbb{R}^2$ with $T(1,1) = (1,0)$ and $T(2,2) = (3,0)$. Justify your answer.

(e) True or false: complex conjugation $\mathbb{C} \to \mathbb{C}$ is linear over $\mathbb{R}$. Justify your answer.

Solution.

(a) See Definition 2.1.1: V, W are vector spaces over the same field F, and $T(u+v) = T(u)+T(v)$ and $T(cv) = cT(v)$ for all $u, v \in V$ and $c \in F$.

(b) See Theorem 2.1.5: if $(v_1, \ldots, v_n)$ is a basis of V and $w_1, \ldots, w_n \in W$ are arbitrary, there is one and only one linear $T\colon V \to W$ with $T(v_i) = w_i$ for all i.

(c) False. $T\colon \mathbb{R}^2 \to \mathbb{R}^2$, $T(a_1, a_2) = (a_1^2, 0)$, sends $\mathbf{0}$ to $\mathbf{0}$ but is not linear (Example (Squaring)).

(d) False. By (LT2), $T(2,2) = 2T(1,1) = (2,0) \neq (3,0)$.

(e) True. See Example (Complex Conjugation: the Field Matters): for real c, $\overline{cz} = c\overline{z}$, and $\overline{z+w} = \overline{z}+\overline{w}$.

2.1.7.2 B. Practice

Exercise. B1: Which Maps Are Linear?

Determine which of the following functions are linear maps over $\mathbb{R}$. Justify your answer.

(a) $T\colon \mathbb{R}^3 \to \mathbb{R}^2$, $T(x,y,z) = (x - 2z,\ 3y)$.

(b) $T\colon \mathbb{R}^2 \to \mathbb{R}^2$, $T(x,y) = (x+1,\ y)$.

(c) $T\colon M_2(\mathbb{R}) \to \mathbb{R}$, $T\begin{pmatrix} a & b \\ c & d \end{pmatrix} = ad - bc$.

(d) $T\colon \mathbb{R}[x]_{\le 2} \to \mathbb{R}[x]_{\le 3}$, $T(p) = x^2 p'$.

(e) $T\colon \mathbb{R}^2 \to \mathbb{R}$, $T(x,y) = xy$.

(f) $T\colon \mathbb{R}^{\mathbb{N}} \to \mathbb{R}^{\mathbb{N}}$, $T(s_0, s_1, s_2, \dots) = (s_1 - s_0,\ s_2 - s_1,\ s_3 - s_2,\ \dots)$.

Solution.

(a) Linear. $T = T_{\mathbf{A}}$ with $\mathbf{A} = \begin{pmatrix} 1 & 0 & -2 \\ 0 & 3 & 0 \end{pmatrix}$, which is linear by Example (Multiplication by a Matrix).

(b) Not linear. $T(0,0) = (1,0) \neq \mathbf{0}$, contradicting Theorem 2.1.1.

(c) Not linear. Take $\mathbf{I}_2$ and $c = 2$: $T(2\mathbf{I}_2) = 2 \cdot 2 - 0 = 4$, but $2T(\mathbf{I}_2) = 2 \cdot 1 = 2$. So (LT2) fails. (Additivity fails too: $T(\mathbf{E}_{11}) + T(\mathbf{E}_{22}) = 0 + 0 = 0$, but $T(\mathbf{E}_{11} + \mathbf{E}_{22}) = T(\mathbf{I}_2) = 1$.)

(d) Linear. For $p \in \mathbb{R}[x]_{\le 2}$, p' has degree at most 1, so $x^2 p' \in \mathbb{R}[x]_{\le 3}$. For p, q and $c \in \mathbb{R}$, by Example (Differentiation) and the distributive law, $T(cp + q) = x^2(cp' + q') = c\,x^2 p' + x^2 q' = cT(p) + T(q)$. By Theorem 2.1.3, T is linear.

(e) Not linear, although $T(0,0) = 0$. $T(2(1,1)) = T(2,2) = 4$, but $2T(1,1) = 2$.

(f) Linear. The entry in position k of $T(cs + t)$ is $(cs_{k+1} + t_{k+1}) - (cs_k + t_k) = c(s_{k+1} - s_k) + (t_{k+1} - t_k)$, which is the entry in position k of $cT(s) + T(t)$. By Theorem 2.1.3, T is linear.

Exercise. B2: A Formula from Values on a Basis

(a) Show that there is exactly one linear map $T\colon \mathbb{R}^2 \to \mathbb{R}^2$ with $T(2,1) = (1,4)$ and $T(1,1) = (0,3)$. Hence find $T(x,y)$ and a matrix $\mathbf{A}$ with $T = T_{\mathbf{A}}$.

(b) Let $S\colon \mathbb{R}[x]_{\le 2} \to \mathbb{R}$ be linear with $S(1) = 1$, $S(1+x) = 3$, and $S(1+x+x^2) = 6$. Find $S(a + bx + cx^2)$.

Solution.

(a) The list $((2,1),(1,1))$ is independent: if $a(2,1) + b(1,1) = \mathbf{0}$, then $2a + b = 0$ and $a + b = 0$, so $a = 0$ and then $b = 0$. It has $2 = \dim \mathbb{R}^2$ entries, so it is a basis (Theorem 1.6.4). By Theorem 2.1.5, exactly one such T exists. Solving $a(2,1) + b(1,1) = (x,y)$, that is $2a + b = x$ and $a + b = y$, gives $a = x - y$ and $b = 2y - x$. Hence

$$T(x,y) = (x-y)(1,4) + (2y-x)(0,3) = (x - y,\ x + 2y).$$

Check: $T(2,1) = (1,4)$ and $T(1,1) = (0,3)$. Hence $T = T_{\mathbf{A}}$ with $\mathbf{A} = \begin{pmatrix} 1 & -1 \\ 1 & 2 \end{pmatrix}$.

(b) The polynomials $1, 1+x, 1+x+x^2$ are nonzero of distinct degrees, so they are independent (Theorem 1.4.4); there are $3 = \dim \mathbb{R}[x]_{\leq 2}$ of them, so they form a basis (Theorem 1.6.4). By (LT1) and Theorem 2.1.2, $S(x) = S(1+x) - S(1) = 2$ and $S(x^2) = S(1+x+x^2) - S(1+x) = 3$. By Theorem 2.1.4,

$$S(a+bx+cx^2) = aS(1) + bS(x) + cS(x^2) = a + 2b + 3c.$$

Exercise. B3: Evaluation and Integration

(a) Prove that $\varphi\colon \mathbb{R}[x] \to \mathbb{R}$, $\varphi(p) = p(3)$, is linear.
(b) Prove that $\psi\colon \mathbb{R}[x] \to \mathbb{R}$, $\psi(p) = \int_0^1 p(t)\,\mathrm{d}t$, is linear.
(c) Hence find $\varphi(a+bx+cx^2)$ and $\psi(a+bx+cx^2)$.

Solution.

(a) Let $p, q \in \mathbb{R}[x]$ and $c \in \mathbb{R}$. By Theorem 0.8.5, $(cp+q)(3) = c\,p(3) + q(3)$, where cp is the product of the constant polynomial c with p, whose value at 3 is $c\,p(3)$. So $\varphi(cp+q) = c\varphi(p) + \varphi(q)$, and φ is linear by Theorem 2.1.3.
(b) By linearity of the integral, $\int_0^1 (cp(t)+q(t))\,\mathrm{d}t = c\int_0^1 p(t)\,\mathrm{d}t + \int_0^1 q(t)\,\mathrm{d}t$. So $\psi(cp+q) = c\psi(p) + \psi(q)$, and ψ is linear by Theorem 2.1.3.
(c) On the basis: $\varphi(1) = 1$, $\varphi(x) = 3$, $\varphi(x^2) = 9$, and $\psi(1) = 1$, $\psi(x) = \frac{1}{2}$, $\psi(x^2) = \frac{1}{3}$. By Theorem 2.1.4,

$$\varphi(a+bx+cx^2) = a + 3b + 9c, \qquad \psi(a+bx+cx^2) = a + \tfrac{1}{2}b + \tfrac{1}{3}c.$$

2.1.7.3 C. Going deeper

Exercise. C1: Over $\mathbb{Q}$, Additivity Is Enough

Let V and W be vector spaces over $\mathbb{Q}$, and let $T\colon V \to W$ be **additive**: $T(u+v) = T(u)+T(v)$ for all $u, v \in V$.

(a) Prove that $T(nv) = nT(v)$ for every $n \in \mathbb{N}$ and $v \in V$.
(b) Prove that $T(nv) = nT(v)$ for every $n \in \mathbb{Z}$.
(c) Prove that $T(qv) = qT(v)$ for every $q \in \mathbb{Q}$. Deduce that T is linear.

Hint: for (c), apply (b) to the vector qv.

Solution.

(a) We use induction on n. For $n = 0$: the proof of Theorem 2.1.1 used only additivity, so $T(0) = 0$, that is, $T(0v) = 0T(v)$. If $T(nv) = nT(v)$, then by additivity $T((n+1)v) = T(nv+v) = T(nv) + T(v) = nT(v) + T(v) = (n+1)T(v)$.
(b) Let $n \in \mathbb{N}$. By additivity and $T(0) = 0$, $0 = T(nv + (-n)v) = T(nv) + T((-n)v)$, so $T((-n)v) = -T(nv) = -nT(v) = (-n)T(v)$, using (a). Every integer is n or $-n$ for some $n \in \mathbb{N}$.

(c) Let $q = m/n$ with $m \in \mathbb{Z}$ and $n \geq 1$. By (b) applied to the vector qv, and then to v,

$$nT(qv) = T(nqv) = T(mv) = mT(v).$$

Multiplying by $\frac{1}{n} \in \mathbb{Q}$ gives $T(qv) = \frac{m}{n}T(v) = qT(v)$. So (LT2) holds for every scalar $q \in \mathbb{Q}$, and with additivity, which is (LT1), T is linear.

Exercise. C2: Over $\mathbb{C}$, Additivity Is Not Enough

(a) Give an additive map $\mathbb{C} \to \mathbb{C}$ that is **not** linear over $\mathbb{C}$, and explain which step of Exercise (C1: Over $\mathbb{Q}$, Additivity Is Enough) fails for it.

(b) Let V and W be vector spaces over $\mathbb{C}$. Regarding them also as spaces over $\mathbb{R}$, let $T\colon V \to W$ be linear **over** $\mathbb{R}$, and suppose $T(iv) = iT(v)$ for all $v \in V$. Prove that T is linear over $\mathbb{C}$.

Solution.

(a) Complex conjugation $T(z) = \overline{z}$ is additive and not $\mathbb{C}$-linear (Example (Complex Conjugation: the Field Matters)). The argument of Exercise (C1: Over $\mathbb{Q}$, Additivity Is Enough) still shows $T(qz) = qT(z)$ for rational q, and that is true for conjugation. What fails is that the rational numbers are not all the scalars: nothing in that argument reaches a scalar like i, which is not built from 1 by adding, subtracting, multiplying, and dividing.

(b) (LT1) holds because T is linear over $\mathbb{R}$. For (LT2), let $c = a + bi \in \mathbb{C}$ with $a, b \in \mathbb{R}$, and $v \in V$. By the vector space axioms, $cv = av + b(iv)$. Using (LT1), then $\mathbb{R}$-homogeneity, then the hypothesis,

$$\begin{aligned} T(cv) &= T(av) + T(b(iv)) = aT(v) + bT(iv) \\ &= aT(v) + b\,iT(v) = (a + bi)T(v) = cT(v). \end{aligned}$$

Hence T is linear over $\mathbb{C}$.

2.2 Kernel and Image

For any function, Chapter 0 asked two questions: is it injective, and is it surjective? For a function between sets, answering them means comparing outputs pair by pair and hunting for preimages point by point. For a linear map both questions reduce to a single subspace each. This section defines the two subspaces, the kernel and the image, shows how to compute them, and proves the test that makes injectivity cheap: a linear map is injective exactly when its kernel is zero.

2.2.1 Two questions, two subspaces

Let $T\colon V \to W$ be linear. Start with injectivity. Suppose two inputs collide, $T(u) = T(v)$. By Theorem 2.1.2 this is the same as

$$T(u - v) = T(u) - T(v) = 0.$$

So a collision between u and v is the same thing as their difference being sent to 0. To detect every collision at once, it is enough to know **which vectors are sent to** 0. For a general function this trick is unavailable, because there is no subtraction; linearity turns a question about all pairs into a question about one set.

Surjectivity asks whether every $w \in W$ is an output, so the set to know is **the set of all outputs**. For the matrix map $T_{\mathbf{A}}(\mathbf{x}) = \mathbf{Ax}$ both sets are concrete: the outputs are the vectors $\mathbf{b}$ for which $\mathbf{Ax} = \mathbf{b}$ can be solved, and the inputs sent to $\mathbf{0}$ are the solutions of $\mathbf{Ax} = \mathbf{0}$. Chapter 3 computes both. The definitions below give the two sets a name for every linear map. They are defined for what they measure.

The kernel is everything T sends to zero; the image is everything T reaches.

Definition 2.2.1. Kernel

Let $T\colon V \to W$ be a linear map over F. The **kernel** of T is the subset of the **domain**

$$\ker T \coloneqq \{v \in V : T(v) = 0_W\} \subseteq V.$$

Definition 2.2.2. Image

Let $T\colon V \to W$ be a linear map over F. The **image** of T is the subset of the **codomain**

$$\operatorname{im} T \coloneqq \{T(v) : v \in V\} = \{w \in W : w = T(v) \text{ for some } v \in V\} \subseteq W.$$

In words: a vector v is in $\ker T$ when v lives in V **and** T sends it to the zero vector of W. A vector w is in $\operatorname{im} T$ when w lives in W **and** there **exists** at least one input with output w; the input need not be unique. In the language of Definition 0.4.2, $\ker T = T^{-1}(\{0\})$ is the preimage of the zero vector, and $\operatorname{im} T = T(V)$ is the image of the function T. The kernel is also called the **null space** of T, and the image its **range**. For group homomorphisms, the kernel was defined the same way (Definition 0.10.3), with 0 playing the role of the identity element.

There is nothing to check for well-definedness: both sets are described by conditions that either hold or fail. What does need checking is that they are **subspaces**, not just subsets.

Theorem 2.2.1. The Kernel Is a Subspace

Let $T\colon V \to W$ be linear. Then $\ker T$ is a subspace of V.

Proof. We use the Subspace Test (Theorem 1.2.1).

(1) By Theorem 2.1.1, $T(0_V) = 0_W$, so $0_V \in \ker T$.

(2) Let $u, v \in \ker T$, so $T(u) = T(v) = 0_W$. By (LT1), $T(u+v) = T(u)+T(v) = 0_W+0_W = 0_W$. Therefore $u + v \in \ker T$.

(3) Let $v \in \ker T$ and $c \in F$. By (LT2), $T(cv) = cT(v) = c0_W = 0_W$, using Theorem 1.1.6. Therefore $cv \in \ker T$.

■

Theorem 2.2.2. The Image Is a Subspace

Let $T\colon V \to W$ be linear. Then:

(a) $\operatorname{im} T$ is a subspace of W;

(b) T is surjective if and only if $\operatorname{im} T = W$.

Proof.

(a) We use the Subspace Test (Theorem 1.2.1). (1) $0_W = T(0_V)$ by Theorem 2.1.1, so $0_W \in \operatorname{im} T$. (2) Let $w_1, w_2 \in \operatorname{im} T$, say $w_1 = T(v_1)$ and $w_2 = T(v_2)$ with $v_1, v_2 \in V$. By (LT1), $w_1 + w_2 = T(v_1 + v_2)$, and $v_1 + v_2 \in V$, so $w_1 + w_2 \in \operatorname{im} T$. (3) Let $w = T(v) \in \operatorname{im} T$ and $c \in F$. By (LT2), $cw = T(cv)$ with $cv \in V$, so $cw \in \operatorname{im} T$.

(b) The inclusion $\operatorname{im} T \subseteq W$ always holds. By Definition 0.4.3, T is surjective exactly when every $w \in W$ equals $T(v)$ for some v, that is, when $W \subseteq \operatorname{im} T$. So T is surjective if and only if $\operatorname{im} T = W$.

■

This pays off a pattern from Chapter 1, §2. The null space of a matrix (Example (Solutions of a Homogeneous System)), the symmetric matrices (Example (Symmetric Matrices)) and the trace-free matrices (Example (Trace-Free Matrices)) were each shown to be subspaces by a separate check, and that section promised a single theorem. Each is the kernel of a linear map ($T_{\mathbf{A}}$, $\mathbf{A} \mapsto \mathbf{A} - \mathbf{A}^{\top}$, and the trace), so Theorem 2.2.1 covers them all at once.

First examples. Start with the maps that have the most extreme kernels and images.

Example. The Zero Map and the Identity

Find the kernel and image of the zero map $0\colon V \to W$ and of the identity $\mathrm{id}_V\colon V \to V$.

Solution. The zero map sends every vector to 0_W, so $\ker 0 = V$ and $\operatorname{im} 0 = \{0_W\}$. The identity sends v to v, which is 0 only when $v = 0$, so $\ker \mathrm{id}_V = \{0\}$. Every v is its own image, so $\operatorname{im} \mathrm{id}_V = V$.

These two are the ends of the scale. The kernel is always between $\{0\}$ and V, and the image between $\{0\}$ and W. The zero map kills everything and reaches nothing but 0; the identity kills nothing and reaches everything.

Example. Projection onto an Axis

Find the kernel and image of the projection $T\colon \mathbb{R}^2 \to \mathbb{R}^2$, $T(a_1, a_2) = (a_1, 0)$, of Example (Projection onto an Axis).

Solution. $T(a_1, a_2) = (0, 0)$ exactly when $a_1 = 0$. So $\ker T = \{(0, a_2) : a_2 \in \mathbb{R}\}$, the y-axis. Every output has second entry 0, and every $(a_1, 0)$ is an output, namely $T(a_1, 0)$. So $\operatorname{im} T$ is the x-axis. The kernel is the direction the light shines along; the image is the screen the

shadow falls on.

A nonexample by minimal change. Replace $\mathbf{0}$ by another vector: for the projection, $\{\mathbf{v} : T(\mathbf{v}) = (1,0)\} = \{(1,a_2) : a_2 \in \mathbb{R}\}$. This is the vertical line through $(1,0)$. It does not contain $\mathbf{0}$, and it is not closed under addition, since $(1,0)+(1,0) = (2,0)$ lies off the line. So it is **not** a subspace. It is the kernel shifted by $(1,0)$. The same thing happens for systems in Chapter 3: the solution set of $\mathbf{Ax} = \mathbf{b}$ is one particular solution plus the set of solutions of $\mathbf{Ax} = \mathbf{0}$, which is $\ker T_{\mathbf{A}}$.

Why the preimage of 0? Because 0 is the one output that every linear map is guaranteed to hit, so its preimage is never empty, and because all other preimages are translates of it. If $T(p) = w$, then $T(v) = w$ exactly when $T(v-p) = 0$, that is, when $v \in p + \ker T$. The kernel describes the shape of every nonempty preimage at once.

Warning. **Keep track of where each subspace lives.** The kernel sits inside the **domain** V; the image sits inside the **codomain** W. For $T_{\mathbf{A}} : F^n \to F^m$, the kernel consists of columns with n entries and the image of columns with m entries, so "$\ker T = \operatorname{im} T$" does not even make sense unless $V = W$. And the image is **not** the codomain in general: the differentiation map $D : \mathbb{R}[x]_{\le 3} \to \mathbb{R}[x]_{\le 3}$ has codomain $\mathbb{R}[x]_{\le 3}$, but x^3 is not the derivative of any cubic, so $\operatorname{im} D \neq \mathbb{R}[x]_{\le 3}$.

2.2.2 Injectivity via the kernel

The computation that opened the section now becomes a theorem. It is the reason to prefer the kernel over the definition of injectivity whenever a map is linear.

Theorem 2.2.3. Kernel Test for Injectivity

Let $T : V \to W$ be linear. Then T is injective **if and only if** $\ker T = \{0\}$.

Proof. ($\Rightarrow$) Suppose T is injective, and let $v \in \ker T$. Then $T(v) = 0_W = T(0_V)$ by Theorem 2.1.1. Injectivity gives $v = 0_V$. Hence $\ker T \subseteq \{0\}$, and the reverse inclusion holds because $\ker T$ is a subspace (Theorem 2.2.1).

($\Leftarrow$) Suppose $\ker T = \{0\}$, and let $T(u) = T(v)$. By Theorem 2.1.2, $T(u-v) = T(u) - T(v) = 0$, so $u - v \in \ker T = \{0\}$. Hence $u = v$, and T is injective.

■

Compare the work. The definition asks: for **all pairs** u, v, does $T(u) = T(v)$ force $u = v$? The theorem asks only: for **each** v, does $T(v) = 0$ force $v = 0$? The second question is usually a homogeneous linear system or a coefficient comparison. For $T_{\mathbf{A}}$, the theorem says that $\mathbf{x} \mapsto \mathbf{Ax}$ is injective exactly when $\mathbf{Ax} = \mathbf{0}$ has only the trivial solution; Chapter 3 turns that into a test that elimination can run.

2.2.3 Computing the image

The image is defined by an existence condition, which is hard to check directly. When V has a spanning list, applying T to it gives a spanning list of $\operatorname{im} T$.

Theorem 2.2.4. The Image Is Spanned by the Images of a Spanning List

Let $T\colon V \to W$ be linear, and suppose $(v_1, \dots, v_k)$ spans V. Then

$$\operatorname{im} T = \operatorname{span}(T(v_1), \dots, T(v_k)).$$

In particular, if V is finite-dimensional, then so is $\operatorname{im} T$, even when W is not.

Proof. ($\supseteq$) Each $T(v_i)$ lies in $\operatorname{im} T$, which is a subspace by Theorem 2.2.2. By Theorem 1.3.1 (3), the span of these vectors is contained in $\operatorname{im} T$.

($\subseteq$) Let $w \in \operatorname{im} T$, say $w = T(v)$ with $v \in V$. Since $(v_1, \dots, v_k)$ spans V, $v = a_1 v_1 + \cdots + a_k v_k$ for some $a_i \in F$. By Theorem 2.1.4, $w = a_1 T(v_1) + \cdots + a_k T(v_k) \in \operatorname{span}(T(v_1), \dots, T(v_k))$.

If V is finite-dimensional, some finite list spans V (Definition 1.3.4), and its images form a finite list spanning $\operatorname{im} T$.

■

So the recipe for the image is: apply T to a basis (or any spanning list) of V, then sift the results to a basis if one is wanted (Theorem 1.5.4). The images of a basis need not be independent; the projection sends the basis $(\mathbf{e}_1, \mathbf{e}_2)$ to $(\mathbf{e}_1, \mathbf{0})$.

2.2.4 Examples

We now compute kernels and images across the families: polynomials, matrices, and sequences. In each case we then read off injectivity from the kernel and surjectivity from the image.

Example. Differentiation

Find the kernel and image of $D\colon \mathbb{R}[x]_{\le 3} \to \mathbb{R}[x]_{\le 2}$, $D(p) = p'$.

Solution. Kernel. Let $p = a_0 + a_1 x + a_2 x^2 + a_3 x^3$. Then $D(p) = a_1 + 2a_2 x + 3a_3 x^2$, which is the zero polynomial exactly when $a_1 = 2a_2 = 3a_3 = 0$. Since $2 \neq 0$ and $3 \neq 0$ in $\mathbb{R}$, this means $a_1 = a_2 = a_3 = 0$. So $\ker D = \{a_0 : a_0 \in \mathbb{R}\} = \operatorname{span}(1)$, the constants.

Image. By Theorem 2.2.4 applied to the basis $(1, x, x^2, x^3)$,

$$\begin{aligned} \operatorname{im} D &= \operatorname{span}(D(1), D(x), D(x^2), D(x^3)) = \operatorname{span}(0, 1, 2x, 3x^2) \\ &= \operatorname{span}(1, x, x^2) = \mathbb{R}[x]_{\le 2}. \end{aligned}$$

So D is surjective but not injective: p and $p + 5$ have the same derivative. Note $\dim \ker D + \dim \operatorname{im} D = 1 + 3 = 4 = \dim \mathbb{R}[x]_{\le 3}$.

Example. Trace

Find the kernel and image of $\operatorname{tr}\colon M_2(\mathbb{R}) \to \mathbb{R}$.

Solution. Kernel. A matrix $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ has trace 0 exactly when $d = -a$, that is, when it equals

$$\begin{pmatrix} a & b \\ c & -a \end{pmatrix} = a \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} + b\mathbf{E}_{12} + c\mathbf{E}_{21}.$$

So $\mathbf{E}_{11} - \mathbf{E}_{22}, \mathbf{E}_{12}, \mathbf{E}_{21}$ span $\ker \operatorname{tr}$. They are independent: the combination above is the zero matrix only if $a = b = c = 0$. Hence they form a basis, and $\dim \ker \operatorname{tr} = 3$.

Image. $\operatorname{im} \operatorname{tr}$ is a subspace of $\mathbb{R}$, and it contains $\operatorname{tr} \mathbf{E}_{11} = 1$. The lazy preimage $r\mathbf{E}_{11}$ of any $r \in \mathbb{R}$ shows $\operatorname{im} \operatorname{tr} = \mathbb{R}$. So the trace is surjective but not injective. Again $3 + 1 = 4 = \dim M_2(\mathbb{R})$.

The next two examples live in infinite-dimensional spaces, and the pattern "one property but not the other" goes the opposite way.

Example. Integration

Find the kernel and image of $J\colon \mathbb{R}[x] \to \mathbb{R}[x]$, $J(p)(x) = \int_0^x p(t)\,\mathrm{d}t$, of Example (Integration).

Solution. Kernel. For $p = a_0 + a_1 x + \cdots + a_n x^n$, $J(p) = a_0 x + \frac{a_1}{2} x^2 + \cdots + \frac{a_n}{n+1} x^{n+1}$. This is the zero polynomial only if every $\frac{a_k}{k+1} = 0$, that is, every $a_k = 0$. So $\ker J = \{0\}$, and J is injective by Theorem 2.2.3.

Image. Let $U = \{q \in \mathbb{R}[x] : q(0) = 0\}$, the polynomials with constant coefficient 0. Every $J(p)$ has constant coefficient 0, so $\operatorname{im} J \subseteq U$. Conversely, if $q = b_1 x + b_2 x^2 + \cdots + b_m x^m \in U$, then $p = b_1 + 2b_2 x + \cdots + m b_m x^{m-1}$ satisfies $J(p) = q$. Hence $\operatorname{im} J = U$. Since $1 \notin U$, J is **not** surjective.

Example. The Shifts: Injective but Not Surjective, and Conversely

Find the kernels and images of the right and left shifts $R, L\colon F^{\mathbb{N}} \to F^{\mathbb{N}}$ of Example (The Shift Maps on Sequences).

Solution. Right shift. $R(s_0, s_1, \dots) = (0, s_0, s_1, \dots)$ is the zero sequence only if every $s_k = 0$. So $\ker R = \{0\}$, and R is injective. Every output has first entry 0; conversely, $(0, t_1, t_2, \dots) = R(t_1, t_2, \dots)$. So $\operatorname{im} R = \{t \in F^{\mathbb{N}} : t_0 = 0\} \neq F^{\mathbb{N}}$, and R is not surjective: $e_0 = (1, 0, 0, \dots)$ is never an output.

Left shift. $L(s_0, s_1, s_2, \dots) = (s_1, s_2, \dots)$ is zero exactly when $s_1 = s_2 = \cdots = 0$, with s_0 arbitrary. So $\ker L = \operatorname{span}(e_0) \neq \{0\}$, and L is not injective. Every sequence is an output, since $L(R(t)) = t$. So $\operatorname{im} L = F^{\mathbb{N}}$, and L is surjective.

This pays off the pointer from Chapter 1, §8: R matches $F^{\mathbb{N}}$ one-to-one with its proper subspace $\operatorname{im} R$. For a map $V \to V$ on a **finite** set, injective and surjective are equivalent (Theorem 0.4.9). The shifts, and the pair J and D on $\mathbb{R}[x]$ (where $D(J(p)) = p$ makes D surjective, while $D(1) = 0$ makes it not injective), show that for linear maps on an infinite-dimensional space this equivalence fails. The next section proves that it holds again when V is finite-dimensional.

Quick check. Find $\ker T$ and $\operatorname{im} T$ for the map $T\colon \mathbb{R}^2 \to \mathbb{R}^3$, $T(x, y) = (2x - y,\ x + y,\ x - y)$, of Example (Finding a Formula from Values on a Basis). Is T injective? Surjective?

Solution. If $T(x, y) = \mathbf{0}$, then $x + y = 0$ and $x - y = 0$, so $x = y = 0$. Hence $\ker T = \{\mathbf{0}\}$, and T is injective. By Theorem 2.2.4, $\operatorname{im} T = \operatorname{span}(T(\mathbf{e}_1), T(\mathbf{e}_2)) = \operatorname{span}\big((2, 1, 1), (-1, 1, -1)\big)$, a plane in $\mathbb{R}^3$. A span of two vectors has dimension at most $2 < 3$, so $\operatorname{im} T \neq \mathbb{R}^3$ and T is not surjective.

2.2.5 What injective and surjective maps do to lists

Injectivity and surjectivity are statements about vectors. They translate into statements about lists: injective maps keep independent lists independent, and surjective maps keep spanning lists spanning. On a basis, the translation goes both ways.

Theorem 2.2.5. Injective Maps Preserve Independence, Surjective Maps Preserve Spanning

Let $T\colon V \to W$ be linear.

(a) If T is injective and $(v_1, \dots, v_k)$ is linearly independent in V, then $(T(v_1), \dots, T(v_k))$ is linearly independent in W.

(b) If T is surjective and $(v_1, \dots, v_k)$ spans V, then $(T(v_1), \dots, T(v_k))$ spans W.

(c) If $(v_1, \dots, v_n)$ is a basis of V, then T is injective if and only if $(T(v_1), \dots, T(v_n))$ is linearly independent, and T is surjective if and only if $(T(v_1), \dots, T(v_n))$ spans W.

Proof.

(a) Let $a_1 T(v_1) + \cdots + a_k T(v_k) = 0$. By Theorem 2.1.4, $T(a_1 v_1 + \cdots + a_k v_k) = 0$, so $a_1 v_1 + \cdots + a_k v_k \in \ker T$. Since T is injective, $\ker T = \{0\}$ (Theorem 2.2.3), so $a_1 v_1 + \cdots + a_k v_k = 0$. The independence of $(v_1, \dots, v_k)$ gives $a_1 = \cdots = a_k = 0$.

(b) By Theorem 2.2.4, $\operatorname{span}(T(v_1), \dots, T(v_k)) = \operatorname{im} T$, and $\operatorname{im} T = W$ because T is surjective (Theorem 2.2.2).

(c) *Injectivity.* ($\Rightarrow$) is (a). ($\Leftarrow$) Suppose the images are independent, and let $v \in \ker T$. Write $v = a_1 v_1 + \cdots + a_n v_n$. By Theorem 2.1.4, $0 = T(v) = a_1 T(v_1) + \cdots + a_n T(v_n)$, so all $a_i = 0$ and $v = 0$. Hence $\ker T = \{0\}$, and T is injective by Theorem 2.2.3.

Surjectivity. By Theorem 2.2.4, $\operatorname{im} T = \operatorname{span}(T(v_1), \dots, T(v_n))$. By Theorem 2.2.2, T is

surjective exactly when this span is W.

■

Warning. **Without injectivity, independence can be lost.** The projection $T(a_1, a_2) = (a_1, 0)$ sends the independent list $((1,0),(1,1))$ to $((1,0),(1,0))$, which is dependent. The kernel is to blame: $(1,1) - (1,0) = (0,1)$ lies in $\ker T$. Likewise, a map that is not surjective sends **no** list to a spanning list of W, since all the images lie in $\operatorname{im} T \neq W$: the inclusion $\mathbb{R}[x]_{\leq 2} \to \mathbb{R}[x]_{\leq 3}$ sends the basis $(1, x, x^2)$ to a list that misses x^3.

Part (c) turns a question about a map into a question about n vectors in W, which for vectors in F^m is a question Chapter 3 settles by elimination. In the next section, counting dimensions will let us check only **one** of injectivity and surjectivity when $\dim V = \dim W$.

2.2.6 Exercises

2.2.6.1 A. Check your understanding

Exercise. A1

(a) Define the kernel and the image of a linear map $T\colon V \to W$. Of which spaces are they subspaces?

(b) State the kernel test for injectivity.

(c) True or false: if $T(u) = T(v)$, then $u - v \in \ker T$. Justify your answer.

(d) True or false: for $D\colon \mathbb{R}[x]_{\leq 3} \to \mathbb{R}[x]_{\leq 3}$, $D(p) = p'$, we have $\operatorname{im} D = \mathbb{R}[x]_{\leq 3}$. Justify your answer.

(e) True or false: for a linear $T\colon V \to W$ and **any** $w \in W$, the set $\{v \in V : T(v) = w\}$ is a subspace of V. Justify your answer.

Solution.

(a) See Definition 2.2.1 and Definition 2.2.2: $\ker T = \{v \in V : T(v) = 0\}$, a subspace of V (Theorem 2.2.1); $\operatorname{im} T = \{T(v) : v \in V\}$, a subspace of W (Theorem 2.2.2).

(b) See Theorem 2.2.3: a linear map is injective if and only if its kernel is $\{0\}$.

(c) True. By Theorem 2.1.2, $T(u - v) = T(u) - T(v) = 0$.

(d) False. Every derivative of a polynomial of degree at most 3 has degree at most 2, so $x^3 \notin \operatorname{im} D$.

(e) False. If $w \neq 0$, the set does not contain 0, since $T(0) = 0 \neq w$. For example, for the projection of Example (Projection onto an Axis) and $w = (1,0)$, the set is the line $\{(1, a_2)\}$. It is a subspace exactly when $w = 0$, in which case it is $\ker T$.

2.2.6.2 B. Practice

Exercise. B1: Kernel and Image from the Definitions

For each map, find the kernel and the image, give a basis of each, and decide whether the map is injective and whether it is surjective.

(a) $D^2\colon \mathbb{R}[x]_{\le 3} \to \mathbb{R}[x]_{\le 3}$, where D is differentiation.
(b) $T\colon M_2(\mathbb{R}) \to M_2(\mathbb{R})$, $T(\mathbf{A}) = \mathbf{A} - \mathbf{A}^\top$.

Solution.

(a) Write $p = a_0 + a_1x + a_2x^2 + a_3x^3$. Then $D^2p = 2a_2 + 6a_3x$. *Kernel:* $D^2p = 0$ forces $a_2 = a_3 = 0$, so $\ker D^2 = \mathbb{R}[x]_{\le 1}$, with basis $(1, x)$. *Image:* the formula shows every D^2p lies in $\mathbb{R}[x]_{\le 1}$, and conversely $b_0 + b_1x = D^2(\frac{b_0}{2}x^2 + \frac{b_1}{6}x^3)$, so $\operatorname{im} D^2 = \mathbb{R}[x]_{\le 1}$, again with basis $(1, x)$. Since $\ker D^2 \neq \{0\}$, D^2 is not injective (Theorem 2.2.3); since $\operatorname{im} D^2 \neq \mathbb{R}[x]_{\le 3}$ (it omits x^2), it is not surjective.

(b) T is linear because the transpose is: $(\mathbf{A} + \mathbf{B})^\top = \mathbf{A}^\top + \mathbf{B}^\top$ and $(c\mathbf{A})^\top = c\mathbf{A}^\top$ (Theorem 0.9.5). *Kernel:* $\mathbf{A} - \mathbf{A}^\top = \mathbf{O}$ says $\mathbf{A} = \mathbf{A}^\top$, so $\ker T$ is the space of symmetric matrices, with basis $(\mathbf{E}_{11}, \mathbf{E}_{22}, \mathbf{E}_{12} + \mathbf{E}_{21})$. *Image:* each $\mathbf{A} - \mathbf{A}^\top$ is skew-symmetric, since $(\mathbf{A} - \mathbf{A}^\top)^\top = \mathbf{A}^\top - \mathbf{A} = -(\mathbf{A} - \mathbf{A}^\top)$; and every skew-symmetric $\mathbf{B} \in M_2(\mathbb{R})$ is a multiple of $\mathbf{E}_{12} - \mathbf{E}_{21}$, with $\mathbf{B} = T(\frac{1}{2}\mathbf{B})$. So $\operatorname{im} T = \operatorname{span}(\mathbf{E}_{12} - \mathbf{E}_{21})$, with basis $(\mathbf{E}_{12} - \mathbf{E}_{21})$. T is neither injective (the kernel is nonzero) nor surjective (the image is not all of $M_2(\mathbb{R})$).

Exercise. B2: Kernels and Images Beyond F^n

For each map, find a basis of the kernel and a basis of the image, and decide whether the map is injective and whether it is surjective.

(a) $T\colon \mathbb{R}[x]_{\le 2} \to \mathbb{R}[x]_{\le 2}$, $T(p) = p' + p$.
(b) $S\colon M_2(\mathbb{R}) \to M_2(\mathbb{R})$, $S(\mathbf{A}) = \mathbf{A} - \mathbf{A}^\top$.

Solution. Both maps are linear. For $c \in \mathbb{R}$, by linearity of differentiation (Example (Differentiation)), $T(cp + q) = (cp' + q') + (cp + q) = cT(p) + T(q)$. By linearity of the transpose (Example (Trace and Transpose)), $S(c\mathbf{A} + \mathbf{B}) = c\mathbf{A} + \mathbf{B} - c\mathbf{A}^\top - \mathbf{B}^\top = cS(\mathbf{A}) + S(\mathbf{B})$. Both are linear by Theorem 2.1.3.

(a) *Kernel.* Let $p = a + bx + cx^2$. Then $T(p) = (b + 2cx) + (a + bx + cx^2) = (a + b) + (b + 2c)x + cx^2$. This is zero exactly when $c = 0$, $b + 2c = 0$, and $a + b = 0$, that is, $c = b = a = 0$. So $\ker T = \{0\}$, with the empty list as basis, and T is injective.
Image. We show T is surjective. Given $q \in \mathbb{R}[x]_{\le 2}$, let $p = q - q' + q''$. Then $p \in \mathbb{R}[x]_{\le 2}$, and since $q''' = 0$,
$$T(p) = (q' - q'' + q''') + (q - q' + q'') = q.$$
So $\operatorname{im} T = \mathbb{R}[x]_{\le 2}$, with basis $(1, x, x^2)$, and T is surjective.

(b) *Kernel.* $S(\mathbf{A}) = \mathbf{O}$ exactly when $\mathbf{A}^\top = \mathbf{A}$. A matrix $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ is symmetric exactly when $b = c$, that is, when it equals $a\mathbf{E}_{11} + d\mathbf{E}_{22} + b(\mathbf{E}_{12} + \mathbf{E}_{21})$. These three matrices are independent (compare entries), so $(\mathbf{E}_{11}, \mathbf{E}_{22}, \mathbf{E}_{12} + \mathbf{E}_{21})$ is a basis of $\ker S$. S is not injective.

Image. By Theorem 2.2.4 with the basis $(\mathbf{E}_{11}, \mathbf{E}_{12}, \mathbf{E}_{21}, \mathbf{E}_{22})$, $\operatorname{im} S$ is spanned by $S(\mathbf{E}_{11}) = \mathbf{O}$,

$S(\mathbf{E}_{12}) = \mathbf{E}_{12} - \mathbf{E}_{21}$, $S(\mathbf{E}_{21}) = \mathbf{E}_{21} - \mathbf{E}_{12}$, and $S(\mathbf{E}_{22}) = \mathbf{O}$. So $\operatorname{im} S = \operatorname{span}(\mathbf{E}_{12} - \mathbf{E}_{21})$, with basis $(\mathbf{E}_{12} - \mathbf{E}_{21})$, the skew-symmetric matrices. Since $\mathbf{I}_2 \notin \operatorname{im} S$, S is not surjective.

2.2.6.3 C. Going deeper

In these exercises, for linear maps $T\colon U \to V$ and $S\colon V \to W$ we write ST for the composite $S \circ T\colon U \to W$, which applies T first. It is linear: $ST(cu + u') = S(cT(u) + T(u')) = cST(u) + ST(u')$, by Theorem 2.1.3 for T and then for S.

Exercise. C1: Kernels and Images of a Composite

Let $T\colon U \to V$ and $S\colon V \to W$ be linear.

(a) Prove that $\ker T \subseteq \ker(ST)$ and $\operatorname{im}(ST) \subseteq \operatorname{im} S$.

(b) Give an example, with $U = V = W = \mathbb{R}^2$, where $\ker T \neq \ker(ST)$.

(c) Deduce that if ST is injective, then T is injective, and if ST is surjective, then S is surjective.

Solution.

(a) Let $u \in \ker T$. Then $ST(u) = S(0) = 0$ by Theorem 2.1.1, so $u \in \ker(ST)$. Let $w \in \operatorname{im}(ST)$, say $w = S(T(u))$. Then $w = S(v)$ with $v = T(u) \in V$, so $w \in \operatorname{im} S$.

(b) Let $T = \mathrm{id}_{\mathbb{R}^2}$ and let S be the projection $S(a_1, a_2) = (a_1, 0)$. Then $\ker T = \{\mathbf{0}\}$, while $\ker(ST) = \ker S$ is the y-axis (Example (Projection onto an Axis)).

(c) If ST is injective, then $\ker(ST) = \{0\}$ (Theorem 2.2.3), so $\ker T = \{0\}$ by (a), and T is injective. If ST is surjective, then $W = \operatorname{im}(ST) \subseteq \operatorname{im} S \subseteq W$ by (a), so $\operatorname{im} S = W$ and S is surjective (Theorem 2.2.2).

Exercise. C2: When Is a Composite Injective?

Let $T\colon U \to V$ and $S\colon V \to W$ be linear, and suppose T is injective. Prove that ST is injective if and only if $\ker S \cap \operatorname{im} T = \{0\}$, that is, if and only if S is injective on $\operatorname{im} T$. Give an example where S is not injective but ST is.

Solution. ($\Rightarrow$) Suppose ST is injective, and let $v \in \ker S \cap \operatorname{im} T$. Then $v = T(u)$ for some $u \in U$, and $ST(u) = S(v) = 0$. So $u \in \ker(ST) = \{0\}$ (Theorem 2.2.3), and $v = T(0) = 0$.

($\Leftarrow$) Suppose $\ker S \cap \operatorname{im} T = \{0\}$, and let $u \in \ker(ST)$. Then $T(u) \in \operatorname{im} T$ and $S(T(u)) = 0$, so $T(u) \in \ker S \cap \operatorname{im} T = \{0\}$. Since T is injective, $u = 0$. Hence $\ker(ST) = \{0\}$, and ST is injective.

Example: let $T\colon \mathbb{R} \to \mathbb{R}^2$, $T(a) = (a, 0)$, which is injective with $\operatorname{im} T$ the x-axis, and let $S\colon \mathbb{R}^2 \to \mathbb{R}$, $S(a_1, a_2) = a_1$. Then $\ker S$ is the y-axis, so S is not injective, but $\ker S \cap \operatorname{im} T = \{\mathbf{0}\}$, and indeed $ST(a) = a$ is injective.

Exercise. C3: Differentiation over $\mathbb{F}_2$

Let $D\colon \mathbb{F}_2[x]_{\le 3} \to \mathbb{F}_2[x]_{\le 3}$ be the formal derivative (Example (Differentiation)). Find bases of $\ker D$ and $\operatorname{im} D$. Which step of Example (Differentiation) fails over $\mathbb{F}_2$?

Solution. For $p = a_0 + a_1x + a_2x^2 + a_3x^3$, $D(p) = a_1 + 2a_2x + 3a_3x^2$. In $\mathbb{F}_2$, $2 = 0$ and $3 = 1$, so $D(p) = a_1 + a_3x^2$.

Kernel. $D(p) = 0$ exactly when $a_1 = a_3 = 0$, so $\ker D = \{a_0 + a_2x^2\}$, with basis $(1, x^2)$. The polynomials 1 and x^2 are independent by Theorem 1.4.4.

Image. By Theorem 2.2.4, $\operatorname{im} D = \operatorname{span}(D(1), D(x), D(x^2), D(x^3)) = \operatorname{span}(0, 1, 0, x^2)$, with basis $(1, x^2)$.

The failing step is "$2a_2 = 0$ forces $a_2 = 0$", which divides by 2; over $\mathbb{F}_2$, $2 = 0$, and $D(x^2) = 2x = 0$ although x^2 is not constant. So over $\mathbb{F}_2$ the kernel of differentiation is bigger than the constants, and here $\ker D = \operatorname{im} D$.

2.3 The Rank–Nullity Theorem

The examples of the last section kept producing the same coincidence. Differentiation $\mathbb{R}[x]_{\le 3} \to \mathbb{R}[x]_{\le 2}$ has a kernel of dimension 1 and an image of dimension 3, and $1 + 3 = 4 = \dim \mathbb{R}[x]_{\le 3}$. The trace on $M_2(\mathbb{R})$ gives $3 + 1 = 4$ again. This section proves that the same equation holds for every linear map on a finite-dimensional space, with no matrices at all, and then turns it into the most useful counting tool of the chapter. Chapter 3 meets the equation again for matrices, counted a different way.

2.3.1 Rank and nullity

The two numbers in the equation are dimensions of the two subspaces of Definition 2.2.1 and Definition 2.2.2, so they are defined whenever those subspaces are finite-dimensional.

Definition 2.3.1. Nullity

Let $T\colon V \to W$ be a linear map with $\ker T$ finite-dimensional. The **nullity** of T is $\operatorname{nullity} T := \dim \ker T$.

Definition 2.3.2. Rank

Let $T\colon V \to W$ be a linear map with $\operatorname{im} T$ finite-dimensional. The **rank** of T is $\operatorname{rank} T := \dim \operatorname{im} T$.

If V is finite-dimensional, both numbers are defined: $\ker T$ is a subspace of V, hence finite-dimensional (Theorem 1.6.7), and $\operatorname{im} T$ is finite-dimensional by Theorem 2.2.4. The nullity counts the independent directions that T destroys; the rank counts the independent directions that survive in the output.

Example. Rank and Nullity in Examples

Find the nullity and rank of (a) differentiation $D\colon \mathbb{R}[x]_{\le 3} \to \mathbb{R}[x]_{\le 2}$; (b) the zero map and

the identity on a space of dimension n.

Solution.

(a) By Example (Differentiation), $\ker D = \operatorname{span}(1)$ and $\operatorname{im} D = \mathbb{R}[x]_{\leq 2}$. So $\operatorname{nullity} D = 1$ and $\operatorname{rank} D = 3$.

(b) By Example (The Zero Map and the Identity), the zero map $V \to V$ has nullity n and rank 0, and the identity has nullity 0 and rank n.

A nonexample. The integration map $J\colon \mathbb{R}[x] \to \mathbb{R}[x]$ of Example (Integration) has nullity 0, but its image contains the independent list $(x, x^2, \ldots, x^k)$ for every k, so the image is not finite-dimensional (Theorem 1.6.3) and $\operatorname{rank} J$ is not defined. What separates J from the examples above is that its domain is infinite-dimensional.

2.3.2 The theorem

We want to prove that for every linear $T\colon V \to W$ with V finite-dimensional,

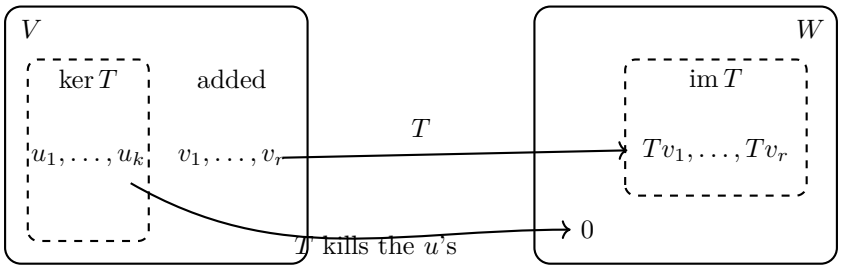
$$\dim \ker T + \dim \operatorname{im} T = \dim V.$$

This is a dimension formula, so we look for the proof pattern that proved the last one, the Dimension Formula for Sums (Theorem 1.7.3): take a basis of the **smallest** space, extend it, make a Big Claim about the vectors that were added, and count. Which space is the smallest? The kernel lives in V and the image lives in W, different spaces that cannot be compared. The only inclusion available is $\ker T \subseteq V$. So we start with a basis of $\ker T$ and extend it to a basis of V. The image has to be reached by applying T, and the kernel vectors contribute nothing when we do. So the natural candidate for a basis of $\operatorname{im} T$ is the images of the **added** vectors.

Theorem 2.3.1. Rank–Nullity Theorem

Let V be a **finite-dimensional** vector space, let W be any vector space over the same field, and let $T\colon V \to W$ be linear. Then $\ker T$ and $\operatorname{im} T$ are finite-dimensional, and

$$\operatorname{nullity} T + \operatorname{rank} T = \dim V.$$



Idea. The picture is the whole proof. The u's all go to 0; the v's carry everything else.

① Take a basis $(u_1, \ldots, u_k)$ of $\ker T$, the smallest space in sight. ② Extend it to a basis $(u_1, \ldots, u_k, v_1, \ldots, v_r)$ of V (Basis Extension). ③ **Big Claim:** $(Tv_1, \ldots, Tv_r)$ is a basis of $\operatorname{im} T$. *Spanning:* apply T to a general vector of V; the u-terms die because they are in the kernel. *Independence:* the only way to use linearity is backwards, turning $\sum b_j T v_j = 0$ into $T(\sum b_j v_j) = 0$. Then $\sum b_j v_j$ is in the kernel, where the u's are a basis, and the big basis of

V finishes it. ④ Count: $\dim V = k + r = \text{nullity}\, T + \text{rank}\, T$.

This is the move of Theorem 1.7.3: basis of the smallest space, extend, Big Claim, count. The new ingredient is that the Big Claim is about the **images** of the added vectors.

Proof. **Step 1: bases.** By Theorem 2.2.1, $\ker T$ is a subspace of V. Since V is finite-dimensional, so is $\ker T$ (Theorem 1.6.7), and it has a basis $(u_1, \dots, u_k)$ with $k = \text{nullity}\, T$ (Corollary 1.5.5). This list is linearly independent in V, so by the Basis Extension Theorem (Theorem 1.6.5) there are $v_1, \dots, v_r \in V$ such that $(u_1, \dots, u_k, v_1, \dots, v_r)$ is a basis of V. Thus $\dim V = k + r$.

Step 2: the Big Claim.

Claim. The list $(Tv_1, \dots, Tv_r)$ is a basis of $\text{im}\, T$.

Proof. Spanning. By Theorem 2.2.4 applied to the basis of Step 1,

$$\text{im}\, T = \text{span}(Tu_1, \dots, Tu_k, Tv_1, \dots, Tv_r).$$

Each $Tu_i = 0$, since $u_i \in \ker T$, and deleting a zero vector from a list does not change its span, because 0 lies in every span (Theorem 1.3.5). Hence $\text{im}\, T = \text{span}(Tv_1, \dots, Tv_r)$.

Independence. Let $b_1, \dots, b_r \in F$ satisfy $b_1 Tv_1 + \cdots + b_r Tv_r = 0$. By Theorem 2.1.4, $T(b_1 v_1 + \cdots + b_r v_r) = 0$, so $b_1 v_1 + \cdots + b_r v_r \in \ker T$. Since $(u_1, \dots, u_k)$ spans $\ker T$, there are $a_1, \dots, a_k \in F$ with

$$b_1 v_1 + \cdots + b_r v_r = a_1 u_1 + \cdots + a_k u_k.$$

Rearranging, $(-a_1)u_1 + \cdots + (-a_k)u_k + b_1 v_1 + \cdots + b_r v_r = 0$. This is a combination of the basis $(u_1, \dots, u_k, v_1, \dots, v_r)$ of V, which is linearly independent, so all its coefficients are 0. In particular $b_1 = \cdots = b_r = 0$. Hence $(Tv_1, \dots, Tv_r)$ is linearly independent.

■

Step 3: count. By the claim, $\text{im}\, T$ has a basis of length r, so it is finite-dimensional with $\text{rank}\, T = r$ (Definition 1.6.1). Therefore

$$\text{nullity}\, T + \text{rank}\, T = k + r = \dim V,$$

as claimed.

■

The proof reads correctly when $k = 0$ (injective T: the v's are a basis of V) and when $r = 0$ ($\ker T = V$: the claim says $\text{im}\, T$ has the empty basis). Notice what was **not** used: nothing about W beyond its being a vector space. The theorem counts only in the domain.

Applied to the matrix map $T_{\mathbf{A}} \colon F^n \to F^m$, the theorem says $\text{nullity}\, T_{\mathbf{A}} + \text{rank}\, T_{\mathbf{A}} = n$. In particular, **the solutions of a homogeneous system $\mathbf{Ax} = \mathbf{0}$ in n unknowns form a subspace of dimension** $n - \text{rank}\, T_{\mathbf{A}}$, whatever method is used to find the rank. Chapter 3 proves this equation a second way, by counting the free and pivot columns of a reduced matrix; here we counted basis vectors, and no row operation appeared.

In the online version you can test the count on any matrix map.

```
from sympy import Matrix

A = Matrix([
    [1, 2, 0, 3],
    [2, 4, 1, 8],
    [-1, -2, 1, -1],
])
rank = A.rank()                  # dim im T_A
nullity = len(A.nullspace())     # dim ker T_A: length of a basis of the kernel
print("rank =", rank, " nullity =", nullity, " dim of domain =", A.cols)
rank + nullity == A.cols
```

2.3.3 Count instead of check

Rank–Nullity links the two questions of the last section. Injectivity is about the kernel, surjectivity about the image, and the theorem says the two dimensions add up to a fixed number. So one controls the other.

Corollary 2.3.2. Consequences of Rank–Nullity

Let V and W be finite-dimensional vector spaces over F, and let $T\colon V \to W$ be linear.

(a) T is injective if and only if $\operatorname{rank} T = \dim V$.

(b) T is surjective if and only if $\operatorname{rank} T = \dim W$.

(c) If $\dim V > \dim W$, then T is **not** injective.

(d) If $\dim V < \dim W$, then T is **not** surjective.

(e) If $\dim V = \dim W$, then T is injective $\iff$ T is surjective $\iff$ T is bijective.

Proof.

(a) By Theorem 2.2.3, T is injective if and only if $\ker T = \{0\}$. A finite-dimensional space is $\{0\}$ exactly when its dimension is 0, because the empty list is a basis of $\{0\}$ and a basis of length 0 spans only $\{0\}$. So T is injective if and only if $\operatorname{nullity} T = 0$, which by Theorem 2.3.1 means $\operatorname{rank} T = \dim V$.

(b) If T is surjective, then $\operatorname{im} T = W$ (Theorem 2.2.2), so $\operatorname{rank} T = \dim W$. Conversely, if $\operatorname{rank} T = \dim W$, then $\operatorname{im} T$ is a subspace of W of the same dimension, so $\operatorname{im} T = W$ by Theorem 1.6.8, and T is surjective.

(c) Since $\operatorname{im} T$ is a subspace of W, $\operatorname{rank} T \leq \dim W < \dim V$ by Theorem 1.6.7. By (a), T is not injective.

(d) By Theorem 2.3.1, $\operatorname{rank} T = \dim V - \operatorname{nullity} T \leq \dim V < \dim W$. By (b), T is not surjective.

(e) If $\dim V = \dim W$, then (a) and (b) have the same right-hand side, so T is injective if and only if it is surjective. Hence either property gives both, which is bijectivity; and bijective maps are injective by definition.

■

Part (e) is the linear version of the pigeonhole principle for finite sets (Theorem 0.4.9), with

dimension in place of the number of elements. It is a **count instead of check**: to show that a map between spaces of the same finite dimension is bijective, check **only** the kernel, usually the easier half, and let the count supply surjectivity. For $T_{\mathbf{A}}$ with $\mathbf{A}$ square, it says that if $\mathbf{Ax} = \mathbf{0}$ has only the trivial solution then $\mathbf{Ax} = \mathbf{b}$ is solvable for every $\mathbf{b}$; Chapter 3 puts that equivalence at the center of its account of square systems. Parts (c) and (d) are the linear version of "no injection from a bigger finite set into a smaller one, and no surjection the other way".

Warning. **Both finiteness and equal dimension are needed.** Rank–Nullity requires V to be finite-dimensional; W may be anything, as for the inclusion $\mathbb{R}[x]_{\leq 2} \to \mathbb{R}[x]$, with nullity 0 and rank 3. Part (e) fails when the spaces are infinite-dimensional, even for an operator $V \to V$: the right shift R on $F^{\mathbb{N}}$ is injective but not surjective, and the left shift L is surjective but not injective (Example (The Shifts: Injective but Not Surjective, and Conversely)). And the hypothesis $\dim V = \dim W$ cannot be dropped: $T(x, y) = (2x - y,\ x + y,\ x - y)$ from $\mathbb{R}^2$ to $\mathbb{R}^3$ is injective but not surjective.

Quick check. Answer each question with a count.

(a) Can a linear map $\mathbb{R}^5 \to \mathbb{R}^2$ be injective?

(b) Is there a linear map $T\colon \mathbb{R}^6 \to \mathbb{R}^4$ with $\operatorname{nullity} T = 1$?

(c) A linear operator T on $\mathbb{R}[x]_{\leq 3}$ satisfies $\ker T = \{0\}$. Must T be surjective?

Solution.

(a) No, by Corollary 2.3.2 (c): $5 > 2$. Concretely, the rank is at most 2, so the nullity is at least 3.

(b) No. By Theorem 2.3.1 the rank would be $6 - 1 = 5$, but $\operatorname{im} T \subseteq \mathbb{R}^4$ has dimension at most 4.

(c) Yes. The domain and codomain are both $\mathbb{R}[x]_{\leq 3}$, of dimension 4, and T is injective, so it is surjective by Corollary 2.3.2 (e).

2.3.4 Existence with conditions on derivatives

The same argument works when some conditions are on derivatives, where no Vandermonde matrix is at hand.

Example. Prescribing Values and Slopes

Show that for all $a, b, c, d \in \mathbb{R}$ there is **exactly one** $p \in \mathbb{R}[x]_{\leq 3}$ with

$$p(0) = a, \qquad p(1) = b, \qquad p'(0) = c, \qquad p'(1) = d.$$

Hence find the cubic that rises from 0 at $x = 0$ to 1 at $x = 1$ with slope 0 at both ends.

Solution. Define $T\colon \mathbb{R}[x]_{\leq 3} \to \mathbb{R}^4$ by $T(p) = (p(0),\ p(1),\ p'(0),\ p'(1))$. Each entry is linear in p: evaluation respects sums and scalar multiples (Theorem 0.8.5), and so does differentiation

(Example (Differentiation)). Hence T is linear by Theorem 2.1.3.

Kernel. Let $p = c_0 + c_1x + c_2x^2 + c_3x^3$ with $T(p) = \mathbf{0}$. Then $p(0) = c_0 = 0$ and $p'(0) = c_1 = 0$. So $p = c_2x^2 + c_3x^3$, and the remaining conditions read

$$p(1) = c_2 + c_3 = 0, \qquad p'(1) = 2c_2 + 3c_3 = 0.$$

Subtracting twice the first from the second gives $c_3 = 0$, and then $c_2 = 0$. Hence $\ker T = \{0\}$.

Count. $\dim \mathbb{R}[x]_{\le 3} = 4 = \dim \mathbb{R}^4$, so T is bijective by Corollary 2.3.2 (e). Surjectivity is existence for every (a, b, c, d), and injectivity is uniqueness.

For the data $(0, 1, 0, 0)$, try $p = \alpha x^2 + \beta x^3$, which already has $p(0) = p'(0) = 0$. Then $p(1) = \alpha + \beta = 1$ and $p'(1) = 2\alpha + 3\beta = 0$ give $\beta = -2$ and $\alpha = 3$. So $p = 3x^2 - 2x^3$; by uniqueness it is the only such cubic.

2.3.5 Dimensions of subspaces cut out by conditions

Many subspaces are defined as "all vectors satisfying some linear conditions", which is to say as the kernel of a linear map. If the map is **surjective**, Rank–Nullity gives the dimension of the kernel at once, and surjectivity is often easy to see.

Example. Dimension of a Kernel from a Surjective Map

Find the dimension of each subspace.

(a) The trace-free matrices $U = \{\mathbf{A} \in M_n(F) : \operatorname{tr} \mathbf{A} = 0\}$ (Example (Trace-Free Matrices)).

(b) $U' = \{p \in \mathbb{R}[x]_{\le n} : \int_0^1 p(t)\,dt = 0\}$.

Solution.

(a) $U = \ker \operatorname{tr}$, where $\operatorname{tr}\colon M_n(F) \to F$ is linear (Example (Trace and Transpose)). Its image is a subspace of F (Theorem 2.2.2) containing $\operatorname{tr} \mathbf{E}_{11} = 1$, so it contains $c \cdot 1 = c$ for every $c \in F$; it is all of F, and $\operatorname{rank} \operatorname{tr} = 1$. Since $\dim M_n(F) = n^2$ (Example (Standard Bases)), Theorem 2.3.1 gives $\dim U = n^2 - 1$.

(b) $U' = \ker \psi$, where $\psi\colon \mathbb{R}[x]_{\le n} \to \mathbb{R}$, $\psi(p) = \int_0^1 p(t)\,dt$, is linear by Exercise (B3: Evaluation and Integration) (b). Since $\psi(1) = 1$, the image is a subspace of $\mathbb{R}$ containing 1, hence all of $\mathbb{R}$, so $\operatorname{rank} \psi = 1$. By Theorem 2.3.1, $\dim U' = (n + 1) - 1 = n$.

Without the theorem, (a) needs an explicit basis of $n^2 - 1$ matrices together with a proof of independence and spanning. With it, one matrix of trace 1 is enough.

2.3.6 Exercises

2.3.6.1 A. Check your understanding

Exercise. A1

(a) Define the rank and the nullity of a linear map.

(b) State the Rank–Nullity Theorem, with its hypotheses.

(c) True or false: there is an injective linear map $\mathbb{R}^4 \to \mathbb{R}^3$. Justify your answer.

(d) True or false: if V is finite-dimensional and $T\colon V \to V$ is linear with $\operatorname{im} T = V$, then T is injective. Justify your answer.

(e) True or false: every injective linear map $F^{\mathbb{N}} \to F^{\mathbb{N}}$ is surjective. Justify your answer.

(f) Name the four steps of the proof of Rank–Nullity.

Solution.

(a) See Definition 2.3.1 and Definition 2.3.2: $\operatorname{nullity} T = \dim \ker T$ and $\operatorname{rank} T = \dim \operatorname{im} T$, when these are finite-dimensional.

(b) See Theorem 2.3.1: if V is finite-dimensional and $T\colon V \to W$ is linear, then $\operatorname{nullity} T + \operatorname{rank} T = \dim V$.

(c) False. $4 > 3$, so no linear map $\mathbb{R}^4 \to \mathbb{R}^3$ is injective (Corollary 2.3.2 (c)).

(d) True. T is surjective, and its domain and codomain are the same finite-dimensional space, so T is injective by Corollary 2.3.2 (e).

(e) False. The right shift R is injective but not surjective (Example (The Shifts: Injective but Not Surjective, and Conversely)). The count fails because $F^{\mathbb{N}}$ is not finite-dimensional.

(f) Take a basis of $\ker T$; extend it to a basis of V; prove the Big Claim that the images of the added vectors form a basis of $\operatorname{im} T$ (spanning, then independence); count.

2.3.6.2 B. Practice

Exercise. B1: Verifying Rank–Nullity

For each map, find a basis of the kernel and a basis of the image, and verify that $\operatorname{nullity} T + \operatorname{rank} T = \dim V$.

(a) $T\colon \mathbb{R}[x]_{\leq 3} \to \mathbb{R}^2$, $T(p) = (p(0), p(1))$.

(b) $T\colon M_2(\mathbb{R}) \to M_2(\mathbb{R})$, $T(\mathbf{A}) = \mathbf{A} + \mathbf{A}^\top$.

Solution.

(a) T is linear by Theorem 0.8.5. *Kernel:* for $p = a_0 + a_1x + a_2x^2 + a_3x^3$, $T(p) = \mathbf{0}$ means $a_0 = 0$ and $a_1 + a_2 + a_3 = 0$, that is, $p = a_2(x^2 - x) + a_3(x^3 - x)$. The list $(x^2 - x, x^3 - x)$ spans $\ker T$, and it is independent by Theorem 1.4.4, so it is a basis. *Image:* $T(1) = (1, 1)$ and $T(x) = (0, 1)$ are independent, so they span a 2-dimensional subspace of $\mathbb{R}^2$, which is $\mathbb{R}^2$ (Theorem 1.6.8). So $\operatorname{im} T = \mathbb{R}^2$ with basis $((1, 1), (0, 1))$. Hence $2 + 2 = 4 = \dim \mathbb{R}[x]_{\leq 3}$.

(b) T is linear because the transpose is (Example (Trace and Transpose)). *Kernel:* $\mathbf{A} + \mathbf{A}^\top = \mathbf{O}$ means $a_{11} = a_{22} = 0$ and $a_{21} = -a_{12}$, so $\ker T = \operatorname{span}(\mathbf{E}_{12} - \mathbf{E}_{21})$, with basis $(\mathbf{E}_{12} - \mathbf{E}_{21})$. *Image:* each $\mathbf{A} + \mathbf{A}^\top$ is symmetric, since $(\mathbf{A} + \mathbf{A}^\top)^\top = \mathbf{A}^\top + \mathbf{A}$, and each

symmetric $\mathbf{B}$ equals $T(\frac{1}{2}\mathbf{B})$. So $\operatorname{im} T$ is the space of symmetric matrices, with basis $(\mathbf{E}_{11}, \mathbf{E}_{22}, \mathbf{E}_{12} + \mathbf{E}_{21})$ (as in Exercise (B2: Kernels and Images Beyond F^n)). Hence $1 + 3 = 4 = \dim M_2(\mathbb{R})$.

Exercise. B2: Bijective by the Kernel Alone

Let $n \geq 0$ and $T\colon \mathbb{R}[x]_{\leq n} \to \mathbb{R}[x]_{\leq n}$, $T(p) = p + xp'$.

(a) Show that T is a linear map into $\mathbb{R}[x]_{\leq n}$ and that $\ker T = \{0\}$. Deduce that T is bijective.

(b) Hence show that there is exactly one $p \in \mathbb{R}[x]_{\leq 2}$ with $p + xp' = 1 + x + x^2$, and find it.

Solution.

(a) For $p = \sum_{k=0}^{n} a_k x^k$, $xp' = \sum_{k=1}^{n} k a_k x^k$, so

$$T(p) = \sum_{k=0}^{n} (1 + k) a_k x^k \in \mathbb{R}[x]_{\leq n}.$$

Linearity: $T(cp+q) = (cp+q)+x(cp'+q') = c(p+xp')+(q+xq') = cT(p)+T(q)$, using Example (Differentiation). If $T(p) = 0$, then $(1 + k)a_k = 0$ for each k; since $1 + k \neq 0$ in $\mathbb{R}$, every $a_k = 0$. So $\ker T = \{0\}$, and T is injective (Theorem 2.2.3). Domain and codomain have the same finite dimension $n+1$, so T is bijective by Corollary 2.3.2 (e).

(b) With $n = 2$, surjectivity gives a solution and injectivity makes it unique. By the formula in (a), $T(a_0 + a_1 x + a_2 x^2) = a_0 + 2a_1 x + 3a_2 x^2$, so we need $a_0 = 1$, $2a_1 = 1$, $3a_2 = 1$. Hence $p = 1 + \frac{1}{2}x + \frac{1}{3}x^2$. Check: $xp' = \frac{1}{2}x + \frac{2}{3}x^2$, and $p + xp' = 1 + x + x^2$.

Exercise. B3: A Dimension via a Surjective Map

Let $U = \{p \in \mathbb{R}[x]_{\leq 4} : p(1) = p(2) = 0\}$. Use a surjective linear map to show that $\dim U = 3$. Hence find a basis of U.

Solution. Define $T\colon \mathbb{R}[x]_{\leq 4} \to \mathbb{R}^2$, $T(p) = (p(1), p(2))$. It is linear by Theorem 0.8.5, and $U = \ker T$. Its image contains $T(1) = (1, 1)$ and $T(x) = (1, 2)$, which are independent, so $\operatorname{im} T$ is a 2-dimensional subspace of $\mathbb{R}^2$ and equals $\mathbb{R}^2$ (Theorem 1.6.8). Thus $\operatorname{rank} T = 2$, and by Theorem 2.3.1, $\dim U = 5 - 2 = 3$.

The polynomials $(x - 1)(x - 2)$, $x(x - 1)(x - 2)$, and $x^2(x - 1)(x - 2)$ vanish at 1 and 2 and have degrees $2, 3, 4 \leq 4$, so they lie in U. They are nonzero of distinct degrees, hence independent (Theorem 1.4.4). There are $3 = \dim U$ of them, so they form a basis of U (Theorem 1.6.4).

2.3.6.3 C. Going deeper

Exercise. C1: A Lower Bound for the Rank of a Composite

Let U and V be finite-dimensional, let W be a vector space, all over F, and let $T\colon U \to V$ and $S\colon V \to W$ be linear. Write ST for the composite $S \circ T$, which is linear (see the note before Exercise (C1: Kernels and Images of a Composite)). Let $S'\colon \operatorname{im} T \to W$ be the restriction of S to $\operatorname{im} T$, that is, $S'(v) = S(v)$ for $v \in \operatorname{im} T$.

(a) Show that S' is linear, that $\operatorname{im} S' = \operatorname{im}(ST)$, and that $\ker S' = \ker S \cap \operatorname{im} T$.

(b) Deduce that $\operatorname{rank}(ST) = \operatorname{rank} T - \dim(\ker S \cap \operatorname{im} T)$.

(c) Hence prove **Sylvester's inequality**: $\operatorname{rank}(ST) \geq \operatorname{rank} S + \operatorname{rank} T - \dim V$.

Hint: for (c), compare $\ker S \cap \operatorname{im} T$ *with* $\ker S$.

Solution.

(a) S' satisfies (LT1) and (LT2) because S does and $\operatorname{im} T$ is a subspace (Theorem 2.2.2) with the operations of V. A vector w lies in $\operatorname{im} S'$ exactly when $w = S(v)$ with $v = T(u)$ for some $u \in U$, that is, when $w = ST(u)$; so $\operatorname{im} S' = \operatorname{im}(ST)$. A vector $v \in \operatorname{im} T$ lies in $\ker S'$ exactly when $S(v) = 0$; so $\ker S' = \ker S \cap \operatorname{im} T$.

(b) $\operatorname{im} T$ is finite-dimensional with $\dim \operatorname{im} T = \operatorname{rank} T$, since U is finite-dimensional (Theorem 2.2.4). By Theorem 2.3.1 applied to $S'\colon \operatorname{im} T \to W$ and by (a),

$$\operatorname{rank} T = \dim \ker S' + \dim \operatorname{im} S' = \dim(\ker S \cap \operatorname{im} T) + \operatorname{rank}(ST).$$

(c) $\ker S \cap \operatorname{im} T$ is a subspace of $\ker S$, so $\dim(\ker S \cap \operatorname{im} T) \leq \operatorname{nullity} S = \dim V - \operatorname{rank} S$, by Theorem 1.6.7 and Theorem 2.3.1 for S. Substituting into (b),

$$\operatorname{rank}(ST) \geq \operatorname{rank} T - (\dim V - \operatorname{rank} S) = \operatorname{rank} S + \operatorname{rank} T - \dim V.$$

Exercise. C2: When the Kernel Equals the Image

Let V be finite-dimensional and $T\colon V \to V$ linear with $\ker T = \operatorname{im} T$.

(a) Prove that $\dim V$ is even.

(b) Give an example of such a T on $\mathbb{R}^2$, and one on $\mathbb{R}^4$.

(c) Is there such a T on $\mathbb{R}^3$? Justify your answer.

Solution.

(a) By Theorem 2.3.1, $\dim V = \operatorname{nullity} T + \operatorname{rank} T = \dim \ker T + \dim \operatorname{im} T = 2 \dim \operatorname{im} T$, which is even.

(b) On $\mathbb{R}^2$, let $T(x, y) = (y, 0)$. Then $T(x, y) = \mathbf{0}$ exactly when $y = 0$, so $\ker T$ is the x-axis; and the outputs $(y, 0)$ fill the x-axis, so $\operatorname{im} T$ is the x-axis too. On $\mathbb{R}^4$, let $T(x_1, x_2, x_3, x_4) = (x_2, 0, x_4, 0)$. Its kernel is $\{x_2 = x_4 = 0\} = \operatorname{span}(\mathbf{e}_1, \mathbf{e}_3)$, and its image is $\{(a, 0, b, 0)\} = \operatorname{span}(\mathbf{e}_1, \mathbf{e}_3)$. Both maps are matrix maps, hence linear.

(c) No. By (a), $\dim V$ would be even, but $\dim \mathbb{R}^3 = 3$.

2.4 The Algebra of Linear Maps

So far we have studied one linear map at a time: its kernel, its image, and the count that ties them together. But linear maps rarely come alone. The map $p \mapsto p + p'$ is the identity plus differentiation, and rotating then projecting is one map followed by another. This section shows that linear maps can be added, scaled, and composed, that these operations obey almost all the rules of arithmetic, and that an operator can be plugged into a polynomial just as a square matrix can.

2.4.1 Adding and scaling linear maps

Look at three maps on $\mathbb{R}[x]_{\leq 2}$ that have already appeared in this chapter or will soon:

$$p \mapsto p + p', \qquad p \mapsto 3p', \qquad p \mapsto p' - 2p.$$

Each is built from two simpler maps, the identity id and differentiation D, by the recipe "apply both to p, then combine the outputs with the operations of the target space". The first is "$\mathrm{id} + D$", the second "$3D$", the third "$D - 2\,\mathrm{id}$". Instead of checking linearity for each such combination separately, we would like to check it once for the recipe. The recipe itself deserves a name.

Linear maps into the same space can be added and scaled output by output, and the results are again linear maps.

Definition 2.4.1. The space of linear maps

Let V and W be vector spaces over **the same field** F. We write $\mathcal{L}(V, W)$ for the set of **all** linear maps $V \to W$, and $\mathcal{L}(V) \coloneqq \mathcal{L}(V, V)$ for the set of linear operators on V (Definition 2.1.1). For $S, T \in \mathcal{L}(V, W)$ and $c \in F$, the **sum** $S + T$ and the **scalar multiple** cT are the functions $V \to W$ defined by

$$(S + T)(v) \coloneqq Sv + Tv, \qquad (cT)(v) \coloneqq c\,(Tv) \qquad \text{for every } v \in V.$$

As in Definition 2.1.1, we write Tv for $T(v)$ when no brackets are needed. In words: to evaluate $S + T$ at v, evaluate both maps at v and add the results **in** W; to evaluate cT, evaluate T and scale the result in W. The operations of V play no role in the definition. That is why S and T must have the **same codomain**: the sum $Sv + Tv$ has to be computed in one space. Both maps must also be defined on the same V.

We have used these operations before without naming them: the operations on F^X in Example (The Standard Vector Spaces) (d) are defined by the same pointwise recipe, with values added in F instead of W.

It turns out that the recipe never leaves the world of linear maps, and that $\mathcal{L}(V, W)$ is itself a vector space. This is the first sign that maps between vector spaces are as good as vectors.

Theorem 2.4.1. $\mathcal{L}(V, W)$ Is a Vector Space

Let V and W be vector spaces over F. If $S, T \in \mathcal{L}(V, W)$ and $c \in F$, then $S + T$ and cT are linear. With the operations of Definition 2.4.1, $\mathcal{L}(V, W)$ is a vector space over F. Its zero vector is the zero map $v \mapsto 0_W$, and the negative of T is $(-1)T \colon v \mapsto -Tv$.

Idea. There are two layers. The set W^V of **all** functions $V \to W$, with the pointwise operations, is a vector space, for the same reason F^X is: each axiom is checked one input at a time, where it becomes an axiom of W. Then $\mathcal{L}(V,W)$ sits inside W^V, so the Subspace Test finishes the job, and its three conditions are exactly "the zero map is linear", "a sum of linear maps is linear", and "a multiple of a linear map is linear".

Proof. Let W^V be the set of all functions $V \to W$, with the operations of Definition 2.4.1 applied to arbitrary functions. The verification in Example (The Standard Vector Spaces) (d) goes through word for word with W in place of F: two functions $V \to W$ are equal when they agree at every $v \in V$ (Definition 0.4.1), and at each v every axiom for W^V becomes the same axiom in W, with zero function $v \mapsto 0_W$ and negative $v \mapsto -f(v)$. Hence W^V is a vector space over F, and $\mathcal{L}(V,W) \subseteq W^V$. We apply Theorem 1.2.1.

(1) The zero map is linear: $0_W = 0_W + 0_W$ and $0_W = c\,0_W$ by Theorem 1.1.6. So $0 \in \mathcal{L}(V,W)$.

(2) Let $S, T \in \mathcal{L}(V,W)$, $u, v \in V$, and $a \in F$. By the definition of $S+T$, the linearity of S and T, and the axioms of W,

$$(S+T)(u+v) = Su + Sv + Tu + Tv = (S+T)u + (S+T)v,$$

$$(S+T)(av) = aSv + aTv = a\,(Sv + Tv) = a\,(S+T)v.$$

So $S+T \in \mathcal{L}(V,W)$.

(3) Let $c \in F$ and $T \in \mathcal{L}(V,W)$. Then $(cT)(u+v) = c\,(Tu+Tv) = (cT)u + (cT)v$, and $(cT)(av) = c\,(aTv) = (ca)Tv = (ac)Tv = a\,(cT)v$, using the axioms of W and commutativity of multiplication in F. So $cT \in \mathcal{L}(V,W)$.

By Theorem 1.2.1, $\mathcal{L}(V,W)$ is a subspace of W^V, hence a vector space over F with the same zero and negatives. The negative of T in W^V is $v \mapsto -Tv = (-1)Tv$, by Theorem 1.1.7. This proves the theorem.

■

The step that used commutativity of F is worth a second look: cT is linear because c can be moved past the scalar a. This is one place where "F is a field" is used silently.

Example. Sums of Matrix Maps and Other First Examples

For $\mathbf{A} \in M_{m\times n}(F)$, write $T_{\mathbf{A}}\colon F^n \to F^m$, $\mathbf{x} \mapsto \mathbf{A}\mathbf{x}$, for the matrix map of Example (Multiplication by a Matrix).

(a) Show that $T_{\mathbf{A}} + T_{\mathbf{B}} = T_{\mathbf{A}+\mathbf{B}}$ and $cT_{\mathbf{A}} = T_{c\mathbf{A}}$ for $\mathbf{A}, \mathbf{B} \in M_{m\times n}(F)$ and $c \in F$.

(b) Describe $\mathcal{L}(V,W)$ when $V = \{0\}$, and when $W = \{0\}$.

(c) On $\mathbb{R}[x]_{\le 2}$, write $T(p) = p' - 2p$ in terms of id and D, and compute $T(1+x^2)$.

Solution.

(a) Let $\mathbf{x} \in F^n$. By Definition 2.4.1 and distributivity (Theorem 0.9.4),

$$(T_{\mathbf{A}} + T_{\mathbf{B}})(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{x} = (\mathbf{A} + \mathbf{B})\mathbf{x} = T_{\mathbf{A}+\mathbf{B}}(\mathbf{x}),$$
$$(cT_{\mathbf{A}})(\mathbf{x}) = c(\mathbf{A}\mathbf{x}) = (c\mathbf{A})\mathbf{x} = T_{c\mathbf{A}}(\mathbf{x}).$$

The maps agree at every $\mathbf{x}$, so they are equal. So adding and scaling matrix maps is adding and scaling matrices.

(b) If $V = \{0\}$, a linear map must send 0 to 0_W, so the zero map is the **only** element: $\mathcal{L}(\{0\}, W) = \{0\}$. If $W = \{0\}$, every function $V \to \{0\}$ is the zero map, and it is linear, so again $\mathcal{L}(V, \{0\}) = \{0\}$. These degenerate cases matter because formulas like "$\dim \mathcal{L}(V, W) = \dim V \cdot \dim W$", which we prove in Chapter 4, must give 0 here, and they do.

(c) $T = D - 2\,\mathrm{id}$, which is linear by Theorem 2.4.1. With $p = 1 + x^2$, $p' = 2x$, so $T(p) = 2x - 2 - 2x^2$.

Here is a nonexample by minimal change. Keep D on $\mathbb{R}[x]_{\leq 2}$, but replace id by the evaluation map $E(p) = p(0) \in \mathbb{R}$. Both maps are linear and both are defined on $\mathbb{R}[x]_{\leq 2}$. But D lands in $\mathbb{R}[x]_{\leq 2}$ and E lands in $\mathbb{R}$, so $D + E$ would require adding a polynomial to a number. The clause that fails is "$S, T \in \mathcal{L}(V, W)$" with the **same** W: the sum is not defined.

2.4.2 Composition

Addition makes $\mathcal{L}(V, W)$ a vector space, but the operation that really matters for maps is doing one after another. For functions, Chapter 0 wrote this as $S \circ T$ (Definition 0.4.4). For linear maps we drop the circle.

Remark. **Notation.** For $T \in \mathcal{L}(U, V)$ and $S \in \mathcal{L}(V, W)$, we write $ST \coloneqq S \circ T \colon U \to W$, so $(ST)u = S(Tu)$. **The map written on the right acts first.** For an operator $T \in \mathcal{L}(V)$, $T^2 = TT$ means "apply T twice", never "square the output".

Is ST linear, and how does composition interact with the sum? It turns out that composition behaves like a multiplication: associative, with identities, and distributive over addition on both sides.

Theorem 2.4.2. Composition of Linear Maps

Let U, V, W, X be vector spaces over F.

(a) **(Linearity)** If $T \in \mathcal{L}(U, V)$ and $S \in \mathcal{L}(V, W)$, then $ST \in \mathcal{L}(U, W)$.

(b) **(Associativity)** If $T \in \mathcal{L}(U, V)$, $S \in \mathcal{L}(V, W)$, and $R \in \mathcal{L}(W, X)$, then $R(ST) = (RS)T$.

(c) **(Identities)** If $T \in \mathcal{L}(U, V)$, then $T\,\mathrm{id}_U = T = \mathrm{id}_V\,T$.

(d) **(Distributivity)** If $T, T_1, T_2 \in \mathcal{L}(U, V)$, $S, S_1, S_2 \in \mathcal{L}(V, W)$, and $c \in F$, then

$$(S_1 + S_2)T = S_1T + S_2T,$$
$$S(T_1 + T_2) = ST_1 + ST_2,$$
$$(cS)T = c(ST) = S(cT).$$

Idea. Parts (b) and (c) hold for all functions, so there is nothing new to do. Part (a) is a two-line check. The interesting point is in (d): the three identities are not equally cheap. In $(S_1 + S_2)T$ the map T acts first and the sum is taken afterwards, so the identity is just the definition of the sum. In $S(T_1 + T_2)$ the sum is taken **before** S acts, and pulling it out of S needs S to be linear. So we should watch where the linearity of the **left** factor is spent.

Proof.

(a) Let $u, u' \in U$ and $a \in F$. Since T and then S are linear,

$$ST(u + u') = S(Tu + Tu') = STu + STu', \qquad ST(au) = S(aTu) = a\,STu.$$

(b) and (c) hold for arbitrary functions, by Theorem 0.4.2 and the observation after Definition 0.4.5.

(c) Each side is a function $U \to W$, so we compare values at $u \in U$. By Definition 2.4.1,

$$(S_1 + S_2)Tu = S_1(Tu) + S_2(Tu) = (S_1T + S_2T)u,$$

with no linearity used. Next, since S is linear,

$$S(T_1 + T_2)u = S(T_1u + T_2u) = ST_1u + ST_2u = (ST_1 + ST_2)u.$$

Finally $(cS)(Tu) = c\,(STu)$ by definition, and $S(cTu) = c\,(STu)$ because S is linear. This proves the theorem.

■

In short, $\mathcal{L}(V)$ carries a vector space structure and a multiplication (composition) that is associative, has the unit id_V, distributes over addition, and lets scalars move freely through products. A vector space with such a multiplication is called an **algebra** over F; $M_n(F)$ is the other algebra we know (Theorem 0.9.4). The two will turn out to be the same algebra in disguise when we write operators as matrices in Chapter 4.

Example. Rotate, Then Project

Let $R\colon \mathbb{R}^2 \to \mathbb{R}^2$ be the rotation by $90°$ counterclockwise, $R(x, y) = (-y, x)$, and let $P\colon \mathbb{R}^2 \to \mathbb{R}^2$, $P(x, y) = (x, 0)$, be the projection onto the x-axis. Compute PR and RP.

Solution. In PR the rotation acts first:

$$PR(x, y) = P(-y, x) = (-y, 0), \qquad RP(x, y) = R(x, 0) = (0, x).$$

At $(1, 0)$: $PR(1, 0) = (0, 0)$, while $RP(1, 0) = (0, 1)$. So $PR \neq RP$. Geometrically, PR rotates $\mathbf{e}_1$ onto the y-axis and then squashes it to $\mathbf{0}$; RP leaves $\mathbf{e}_1$ alone under the projection and then rotates it to $\mathbf{e}_2$. In matrix form, $P = T_{\mathbf{A}}$ and $R = T_{\mathbf{B}}$ with $\mathbf{A} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$ and $\mathbf{B} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$, and indeed $\mathbf{AB} = \begin{pmatrix} 0 & -1 \\ 0 & 0 \end{pmatrix} \neq \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} = \mathbf{BA}$.

Warning. **Composition is not commutative, and the order is read right to left.** In Example (Rotate, Then Project), $PR \neq RP$. It can be worse: for $T \in \mathcal{L}(\mathbb{R}^2, \mathbb{R}^3)$ and $S \in \mathcal{L}(\mathbb{R}^3, \mathbb{R}^2)$, both $ST \in \mathcal{L}(\mathbb{R}^2)$ and $TS \in \mathcal{L}(\mathbb{R}^3)$ are defined, but they live on different spaces, so they cannot even be compared. And for $T \in \mathcal{L}(\mathbb{R}^2, \mathbb{R}^3)$, $S \in \mathcal{L}(\mathbb{R}^3, \mathbb{R}^4)$, ST is defined but TS is not. The thing you do first goes on the **right**.

Quick check. Let $T\colon \mathbb{R}^3 \to \mathbb{R}^2$ and $S\colon \mathbb{R}^2 \to \mathbb{R}^4$ be linear. Which of ST and TS is defined, and between which spaces does it go?

Solution. $ST = S \circ T$ first applies $T\colon \mathbb{R}^3 \to \mathbb{R}^2$, then $S\colon \mathbb{R}^2 \to \mathbb{R}^4$, so $ST \in \mathcal{L}(\mathbb{R}^3, \mathbb{R}^4)$. TS would apply S first, landing in $\mathbb{R}^4$, where T is not defined; so TS is not defined.

Because of the missing commutativity, school identities about products must be rechecked. By distributivity (Theorem 2.4.2 (d)),

$$(S+T)^2 = (S+T)(S+T) = S^2 + ST + TS + T^2,$$

which equals $S^2 + 2ST + T^2$ only if $ST = TS$. With P and R as above, $PR + RP\colon (x, y) \mapsto (-y, x)$, while $2PR\colon (x, y) \mapsto (-2y, 0)$; so $(P+R)^2 \neq P^2 + 2PR + R^2$. This is the operator version of the matrix warning in Chapter 0, §9.

2.4.3 Operators, powers, and polynomials

Composition of operators on one space V never leaves $\mathcal{L}(V)$, so we may compose an operator with itself as often as we like. Repeated composition leads to powers, and combinations of powers to polynomials, exactly as for square matrices (Definition 0.9.14).

Definition 2.4.2. Powers and Polynomials of an Operator

Let V be a vector space over F and $T \in \mathcal{L}(V)$. The **powers** of T are defined recursively by

$$T^0 := \mathrm{id}_V, \qquad T^{k+1} := T^k T \quad (k \in \mathbb{N}).$$

For $p = a_0 + a_1 x + \cdots + a_N x^N \in F[x]$, the operator $p(T) \in \mathcal{L}(V)$ is

$$p(T) := a_0\,\mathrm{id}_V + a_1 T + a_2 T^2 + \cdots + a_N T^N.$$

In words: T^k applies T exactly k times, and T^0 does nothing. In $p(T)$, the **constant term becomes** $a_0\,\mathrm{id}_V$, because an operator cannot be added to a scalar. Each T^k is linear by Theorem 2.4.2 (a) and induction, and $p(T)$ is linear by Theorem 2.4.1. As for matrices, $p(T)$ does not depend on how many zero coefficients we write, since $0 \cdot T^k$ is the zero map.

Plugging in an operator respects sums and products of polynomials, just as plugging in a matrix does.

Theorem 2.4.3. Polynomials in One Operator

Let $T \in \mathcal{L}(V)$ and $p, q \in F[x]$. Then:

(a) $T^jT^k = T^{j+k}$ for all $j, k \in \mathbb{N}$;
(b) $(p+q)(T) = p(T) + q(T)$ and $(pq)(T) = p(T)\,q(T)$;
(c) $p(T)\,q(T) = q(T)\,p(T)$.

Idea. The proof of Theorem 0.9.9 used four matrix rules: associativity, identity, distributivity, and "scalars move through products". Theorem 2.4.2 supplies exactly these four rules for operators. So the old proof transfers line by line. Part (c) is the surprise: T need not commute with other operators, but it commutes with every polynomial in itself.

Proof.

(a) Fix j and use induction on k (Theorem 0.2.7). For $k = 0$, $T^jT^0 = T^j\,\mathrm{id}_V = T^j$ by Theorem 2.4.2 (c). If $T^jT^k = T^{j+k}$, then by Definition 2.4.2 and associativity (Theorem 2.4.2 (b)),

$$T^jT^{k+1} = T^j(T^kT) = (T^jT^k)T = T^{j+k}T = T^{j+k+1}.$$

(b) Let $p = \sum_{i=0}^{m} a_i x^i$ and $q = \sum_{j=0}^{l} b_j x^j$. The statement for sums holds because $(a_i + b_i)T^i = a_iT^i + b_iT^i$ in the vector space $\mathcal{L}(V)$ (Theorem 2.4.1). For products, the coefficient of x^k in pq is $\sum_{i+j=k} a_ib_j$ (Definition 0.8.2), so

$$\begin{aligned}(pq)(T) &= \sum_k \Big(\sum_{i+j=k} a_ib_j \Big) T^k = \sum_{i=0}^{m}\sum_{j=0}^{l} (a_ib_j)\,T^iT^j \\ &= \Big(\sum_{i=0}^{m} a_iT^i\Big)\Big(\sum_{j=0}^{l} b_jT^j\Big) = p(T)\,q(T),\end{aligned}$$

where the second equality uses (a), and the third uses both distributive laws and $(a_iT^i)(b_jT^j) = (a_ib_j)\,T^iT^j$, all from Theorem 2.4.2 (d).

(c) Since $pq = qp$ in $F[x]$ (Theorem 0.8.1), part (b) gives $p(T)q(T) = (pq)(T) = (qp)(T) = q(T)p(T)$. This proves the theorem.

■

In practice, (b) says that **a factorization of p is a factorization of $p(T)$**: for instance $T^2 - 3T + 2\,\mathrm{id}_V = (T - \mathrm{id}_V)(T - 2\,\mathrm{id}_V) = (T - 2\,\mathrm{id}_V)(T - \mathrm{id}_V)$.

Example. Differentiation Is Nilpotent

Let $n \geq 0$ and let $D \in \mathcal{L}(F[x]_{\leq n})$ be (formal) differentiation (Example (Differentiation)), $D(a_0 + a_1x + \cdots + a_nx^n) = a_1 + 2a_2x + \cdots + na_nx^{n-1}$, where ka_k means a_k added to itself k times. Show that $D^{n+1} = 0$. Show also that $D^n \neq 0$ when $F = \mathbb{R}$.

Solution. For $0 \leq k \leq n$, $D(x^k) = kx^{k-1}$ if $k \geq 1$ and $D(1) = 0$. So D sends each x^k to a scalar multiple of x^{k-1}, or to 0 when $k = 0$. Applying this $k+1$ times, $D^{k+1}(x^k)$ is a scalar multiple of $D(1) = 0$, so $D^{k+1}(x^k) = 0$. Since $k \leq n$, also $D^{n+1}(x^k) = D^{n-k}(D^{k+1}(x^k)) = D^{n-k}(0) = 0$, by Theorem 2.4.3 (a).

Now let $p = a_0 + a_1 x + \cdots + a_n x^n$. Since D^{n+1} is linear, Theorem 2.1.4 gives

$$D^{n+1}(p) = a_0 D^{n+1}(1) + a_1 D^{n+1}(x) + \cdots + a_n D^{n+1}(x^n) = 0.$$

Hence $D^{n+1} = 0$.

Over $\mathbb{R}$, $D^n(x^n) = n(n-1)\cdots 1 = n! \neq 0$, so $D^n \neq 0$. Over $\mathbb{F}_2$ with $n = 2$, however, $D(x^2) = 2x = 0$, and $D^2 = 0$ already; the field matters for the exact power, but never for $D^{n+1} = 0$.

An operator T with $T^k = 0$ for some $k \geq 1$ is called **nilpotent**. The example shows that a nonzero operator can have a power equal to zero, just as the matrix $\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$ can. Differentiation is the model nilpotent operator, and it returns when we study canonical forms.

Example. Operators That Satisfy a Polynomial

With P and R as in Example (Rotate, Then Project), find nonzero polynomials p and q with $p(P) = 0$ and $q(R) = 0$. Also, for an arbitrary $p \in F[x]$, compute $p(\mathrm{id}_V)$ and $p(Z)$, where $Z \in \mathcal{L}(V)$ is the zero operator.

Solution. For the projection, $P^2(x, y) = P(x, 0) = (x, 0) = P(x, y)$, so $P^2 = P$ and $p = x^2 - x$ satisfies $p(P) = P^2 - P = 0$. By Theorem 2.4.3 (b), this says $P(P - \mathrm{id}) = 0$: whatever $P - \mathrm{id}$ produces is killed by P, and indeed $(P - \mathrm{id})(x, y) = (0, -y)$ lies on the y-axis.

For the rotation, $R^2(x, y) = R(-y, x) = (-x, -y)$, so $R^2 = -\mathrm{id}$, and $q = x^2 + 1$ satisfies $q(R) = R^2 + \mathrm{id} = 0$. Rotating twice by $90°$ is rotating by $180°$. Since q has no real roots, q has no real linear factor, so the trick "factor q, then factor $q(R)$" gives nothing over $\mathbb{R}$.

For $p = a_0 + a_1 x + \cdots + a_N x^N$: every power of id_V is id_V, so $p(\mathrm{id}_V) = (a_0 + a_1 + \cdots + a_N)\,\mathrm{id}_V = p(1)\,\mathrm{id}_V$. Every power Z^k with $k \geq 1$ is the zero map, but $Z^0 = \mathrm{id}_V$, so $p(Z) = a_0\,\mathrm{id}_V$, the scalar $p(0)$ times id_V. The second computation is the reason for the convention $T^0 = \mathrm{id}_V$: with it, "the constant term" of $p(T)$ is what survives when T is the zero operator.

Warning. **The constant term is $a_0\,\mathrm{id}_V$, not a_0.** Writing "$P^2 - P + 3$" for an operator is a type error: $P^2 - P$ is a map and 3 is a number. For $p = x^2 - x + 3$, $p(P) = P^2 - P + 3\,\mathrm{id}$, which is $3\,\mathrm{id}$ since $P^2 = P$. Forgetting the id leads to "$p(P) = 3$", which is not an operator at all.

Quick check. With P the projection above, compute $p(P)$ for $p = x^3 - 2x + 1$.

Solution. Since $P^2 = P$, also $P^3 = P^2 P = PP = P$. So $p(P) = P^3 - 2P + \mathrm{id} = P - 2P + \mathrm{id} = \mathrm{id} - P$, which is the map $(x, y) \mapsto (0, y)$: the projection onto the y-axis.

Polynomial relations such as $P^2 = P$, $R^2 = -\mathrm{id}$, and $D^{n+1} = 0$ carry real information about an

operator. In the next section we use one to write down an inverse without solving any equations, and later in the book the polynomials that kill an operator lead to the minimal polynomial and to canonical forms.

2.4.4 Exercises

2.4.4.1 A. Check your understanding

Exercise. A1

(a) For $S, T \in \mathcal{L}(V, W)$ and $c \in F$, define $S + T$ and cT. Why must S and T have the same codomain?

(b) True or false: for all $S, T \in \mathcal{L}(V)$, $ST = TS$. Justify your answer.

(c) Let $T \in \mathcal{L}(\mathbb{R}^5, \mathbb{R}^2)$ and $S \in \mathcal{L}(\mathbb{R}^2, \mathbb{R}^5)$. Which of ST, TS are defined, and in which spaces do they lie?

(d) For $T \in \mathcal{L}(V)$ and $p = 4$ (a constant polynomial), what is $p(T)$?

(e) True or false: for all $S, T \in \mathcal{L}(V)$, $(S+T)^2 = S^2 + 2ST + T^2$. Justify your answer.

Solution.

(a) $(S+T)(v) = Sv + Tv$ and $(cT)(v) = cTv$ for every $v \in V$ (Definition 2.4.1). The sum $Sv + Tv$ must be computed in a single vector space, so both values must lie in the same W.

(b) False. In Example (Rotate, Then Project), $PR(1,0) = (0,0)$ but $RP(1,0) = (0,1)$.

(c) Both are defined. ST applies T first: $ST \in \mathcal{L}(\mathbb{R}^5)$. TS applies S first: $TS \in \mathcal{L}(\mathbb{R}^2)$.

(d) $p(T) = 4\,\mathrm{id}_V$, by Definition 2.4.2.

(e) False. By Theorem 2.4.2 (d), $(S+T)^2 = S^2 + ST + TS + T^2$, which differs from $S^2 + 2ST + T^2$ by $TS - ST$. With $S = P$, $T = R$ of Example (Rotate, Then Project), $RP - PR$: $(x, y) \mapsto (y, x)$ is not the zero map, so the identity fails.

2.4.4.2 B. Practice

Exercise. B1: Computing compositions

Let $S, T \in \mathcal{L}(\mathbb{R}^2)$ be given by $S(x, y) = (x + y,\ y)$ and $T(x, y) = (y,\ x)$.

(a) Compute ST, TS, and $2S - T$.

(b) Compute S^2 and S^k for every $k \in \mathbb{N}$.

(c) Hence show that $(S - \mathrm{id})^2 = 0$.

Solution.

(a) Applying the right-hand map first,

$$ST(x, y) = S(y, x) = (x + y,\ x), \qquad TS(x, y) = T(x + y,\ y) = (y,\ x + y).$$

So $ST \neq TS$; for instance $ST(1,0) = (1,1)$ and $TS(1,0) = (0,1)$. By Definition 2.4.1,

$(2S - T)(x, y) = (2x + 2y,\ 2y) - (y,\ x) = (2x + y,\ 2y - x)$.

(b) $S^2(x, y) = S(x + y,\ y) = (x + 2y,\ y)$. We claim $S^k(x, y) = (x + ky,\ y)$ for all $k \in \mathbb{N}$, by induction on k. For $k = 0$, $S^0 = \mathrm{id}$ sends (x, y) to $(x + 0y,\ y)$. If the claim holds for k, then

$$S^{k+1}(x, y) = S^k(S(x, y)) = S^k(x + y,\ y) = (x + y + ky,\ y) = (x + (k + 1)y,\ y),$$

where the first equality is Definition 2.4.2, $S^{k+1} = S^k S$.

(c) $(S - \mathrm{id})(x, y) = (y,\ 0)$, so $(S - \mathrm{id})^2(x, y) = (S - \mathrm{id})(y, 0) = (0, 0)$. Hence $(S - \mathrm{id})^2 = 0$, that is, $S - \mathrm{id}$ is nilpotent. Expanding with Theorem 2.4.3 (b), $S^2 - 2S + \mathrm{id} = 0$, which matches (b): $(x + 2y, y) - 2(x + y, y) + (x, y) = (0, 0)$.

Exercise. B2: Operators that preserve a subspace

Let U be a subspace of V, and let

$$\mathcal{A}_U \coloneqq \{T \in \mathcal{L}(V) : Tu \in U \text{ for every } u \in U\}.$$

Prove that $\mathcal{A}_U$ is a subspace of $\mathcal{L}(V)$, that $\mathrm{id}_V \in \mathcal{A}_U$, and that $ST \in \mathcal{A}_U$ whenever $S, T \in \mathcal{A}_U$.

Solution. We use Theorem 1.2.1 in the vector space $\mathcal{L}(V)$ (Theorem 2.4.1).

(1) The zero map sends every $u \in U$ to 0, and $0 \in U$ because U is a subspace. So $0 \in \mathcal{A}_U$.

(2) Let $S, T \in \mathcal{A}_U$ and $u \in U$. Then $Su \in U$ and $Tu \in U$, so $(S + T)u = Su + Tu \in U$, since U is closed under addition. Therefore $S + T \in \mathcal{A}_U$.

(3) Let $c \in F$ and $T \in \mathcal{A}_U$. For $u \in U$, $(cT)u = cTu \in U$, since U is closed under scalar multiplication. Therefore $cT \in \mathcal{A}_U$.

Hence $\mathcal{A}_U$ is a subspace of $\mathcal{L}(V)$. Next, $\mathrm{id}_V u = u \in U$ for every $u \in U$, so $\mathrm{id}_V \in \mathcal{A}_U$. Finally let $S, T \in \mathcal{A}_U$ and $u \in U$. Then $Tu \in U$, and applying the hypothesis on S to the vector $Tu \in U$ gives $STu = S(Tu) \in U$. This shows $ST \in \mathcal{A}_U$.

Exercise. B3: A polynomial that kills D

Let $D \in \mathcal{L}(\mathbb{R}[x]_{\le 2})$ be differentiation.

(a) Find a nonzero polynomial p with $p(D) = 0$.

(b) Show that no nonzero polynomial q of degree at most 2 satisfies $q(D) = 0$.

(c) Hence, or otherwise, compute $(\mathrm{id} + D)^3$ as a combination of id, D, D^2, and evaluate $(\mathrm{id} + D)^3(x^2)$.

Solution.

(a) By Example (Differentiation Is Nilpotent) with $n = 2$, $D^3 = 0$. So $p = x^3$ works.

(b) Let $q = a + bx + cx^2$ with $q(D) = a\,\mathrm{id} + bD + cD^2 = 0$. Applying this operator to x^2,

with $D(x^2) = 2x$ and $D^2(x^2) = 2$,

$$0 = ax^2 + 2bx + 2c.$$

A polynomial is zero exactly when all its coefficients are, so $a = 0$, $2b = 0$, and $2c = 0$. Hence $a = b = c = 0$ and $q = 0$.

(c) The operator $\mathrm{id} + D$ is $q(D)$ for $q = 1 + x$. Applying the product rule of Theorem 2.4.3 (b) twice, $(\mathrm{id} + D)^3 = q(D)q(D)q(D) = (q^3)(D)$, and $q^3 = 1 + 3x + 3x^2 + x^3$. So

$$(\mathrm{id} + D)^3 = \mathrm{id} + 3D + 3D^2 + D^3 = \mathrm{id} + 3D + 3D^2,$$

using $D^3 = 0$ from (a). Therefore $(\mathrm{id} + D)^3(x^2) = x^2 + 3 \cdot 2x + 3 \cdot 2 = x^2 + 6x + 6$.

2.4.4.3 C. Going deeper

Exercise. C1: The Heisenberg relation

Let F be a field, let $D \in \mathcal{L}(F[x])$ be formal differentiation, $D(\sum_k a_k x^k) = \sum_{k \geq 1} k a_k x^{k-1}$ (Example (Differentiation)), and let $S \in \mathcal{L}(F[x])$ be multiplication by x, $S(p) = xp$ (the map M of Example (Multiplication by x)).

(a) Prove that $DS - SD = \mathrm{id}$.

(b) Let $T \in \mathcal{L}(F^n, F^m)$. Prove that $T = T_{\mathbf{A}}$, where $\mathbf{A} \in M_{m \times n}(F)$ is the matrix with columns $T\mathbf{e}_1, \ldots, T\mathbf{e}_n$.

(c) Now suppose that F has **characteristic** 0 (Definition 0.7.2), and let $n \geq 1$. Deduce that there are no operators $D', S' \in \mathcal{L}(F^n)$ with $D'S' - S'D' = \mathrm{id}$.

(d) Show that the assumption on the characteristic in (c) cannot be dropped.

Hint: for (c), use (b) and consider traces.

Solution.

(a) Both sides are linear (Theorem 2.4.1, Theorem 2.4.2). We first compare them on x^k, $k \in \mathbb{N}$. We have $DS(x^k) = D(x^{k+1}) = (k+1)x^k$. If $k \geq 1$, $SD(x^k) = x \cdot kx^{k-1} = kx^k$; if $k = 0$, $SD(1) = x \cdot 0 = 0 = 0 \cdot x^0$. In both cases

$$(DS - SD)(x^k) = (k+1)x^k - kx^k = x^k = \mathrm{id}(x^k).$$

Every $p \in F[x]$ is a finite combination $p = \sum_k a_k x^k$, so by Theorem 2.1.4, $(DS - SD)(p) = \sum_k a_k (DS - SD)(x^k) = \sum_k a_k x^k = p$. Hence $DS - SD = \mathrm{id}$.

(b) Let $\mathbf{x} = (x_1, \ldots, x_n) = x_1\mathbf{e}_1 + \cdots + x_n\mathbf{e}_n \in F^n$. By Theorem 2.1.4 and Theorem 0.9.3,

$$T\mathbf{x} = x_1 T\mathbf{e}_1 + \cdots + x_n T\mathbf{e}_n = \mathbf{A}\mathbf{x} = T_{\mathbf{A}}(\mathbf{x}).$$

Since $\mathbf{x}$ was arbitrary, $T = T_{\mathbf{A}}$.

(c) Suppose, for a contradiction, that $D', S' \in \mathcal{L}(F^n)$ satisfy $D'S' - S'D' = \mathrm{id}$. By (b), $D' = T_{\mathbf{A}}$ and $S' = T_{\mathbf{B}}$ for some $\mathbf{A}, \mathbf{B} \in M_n(F)$, and $\mathrm{id} = T_{\mathbf{I}_n}$. By associativity of matrix multiplication, $T_{\mathbf{A}}T_{\mathbf{B}}(\mathbf{x}) = \mathbf{A}(\mathbf{B}\mathbf{x}) = (\mathbf{A}\mathbf{B})\mathbf{x}$, so $T_{\mathbf{A}}T_{\mathbf{B}} = T_{\mathbf{A}\mathbf{B}}$, and similarly $T_{\mathbf{B}}T_{\mathbf{A}} = T_{\mathbf{B}\mathbf{A}}$. With Example (Sums of Matrix Maps and Other First Examples) (a),

$$T_{\mathbf{A}\mathbf{B}-\mathbf{B}\mathbf{A}} = T_{\mathbf{A}}T_{\mathbf{B}} - T_{\mathbf{B}}T_{\mathbf{A}} = \mathrm{id} = T_{\mathbf{I}_n}.$$

Evaluating at $\mathbf{e}_j$ gives equal j-th columns (Theorem 0.9.3), so $\mathbf{AB}-\mathbf{BA}=\mathbf{I}_n$. Taking traces and using Theorem 0.9.6,

$$n\cdot 1=\operatorname{tr}\mathbf{I}_n=\operatorname{tr}(\mathbf{AB})-\operatorname{tr}(\mathbf{BA})=0.$$

Since $n\geq 1$ and F has characteristic 0, $n\cdot 1\neq 0$ by Definition 0.7.2. This is a contradiction, so no such D',S' exist. In words: over $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$, or any field of characteristic 0, the relation $DS-SD=\mathrm{id}$ of (a) has no analogue on F^n, $n\geq 1$. Since every n-dimensional space is isomorphic to F^n, as we show in the next section, the same holds on every nonzero finite-dimensional space. (On the zero space $\mathrm{id}=0$, so the relation holds trivially there; that is why $n\geq 1$.)

(d) Over $\mathbb{F}_2$ (characteristic 2), take $\mathbf{A}=\begin{pmatrix}0&0\\1&0\end{pmatrix}$ and $\mathbf{B}=\begin{pmatrix}0&1\\0&0\end{pmatrix}$ as in Exercise (C1: **AB − BA = I is impossible, or is it?**). There $\mathbf{AB}-\mathbf{BA}=\begin{pmatrix}-1&0\\0&1\end{pmatrix}=\mathbf{I}_2$, because $-1=1$ in $\mathbb{F}_2$. By the computation in (c), $D'=T_{\mathbf{A}}$ and $S'=T_{\mathbf{B}}$ satisfy $D'S'-S'D'=T_{\mathbf{AB}-\mathbf{BA}}=\mathrm{id}$ on $\mathbb{F}_2^2$. So the conclusion of (c) fails without the characteristic assumption; the proof breaks exactly at $n\cdot 1\neq 0$, since $2\cdot 1=0$ in $\mathbb{F}_2$.

Exercise. C2: Commuting nilpotent operators

Let $S,T\in\mathcal{L}(V)$.

(a) Suppose $ST=TS$. Prove that $(S+T)^n=\sum_{k=0}^{n}\binom{n}{k}S^kT^{n-k}$ for every $n\in\mathbb{N}$, where $\binom{n}{k}S^kT^{n-k}$ means S^kT^{n-k} added to itself $\binom{n}{k}$ times.

(b) Deduce that if $ST=TS$, $S^a=0$, and $T^b=0$ for some $a,b\geq 1$, then $(S+T)^{a+b-1}=0$.

(c) Show that (b) fails without $ST=TS$: give nilpotent $S,T\in\mathcal{L}(\mathbb{R}^2)$ such that $S+T$ is not nilpotent.

Hint: for (c), try maps that send one standard basis vector to the other and the other to $\mathbf{0}$.

Solution.

(a) First, $T^jS=ST^j$ for all $j\in\mathbb{N}$, by induction on j: it holds for $j=0$, and if $T^jS=ST^j$, then $T^{j+1}S=T^j(TS)=T^j(ST)=(T^jS)T=(ST^j)T=ST^{j+1}$, using associativity (Theorem 2.4.2 (b)) and Theorem 2.4.3 (a) in the form $T^{j+1}=T^jT$.
Now induct on n. For $n=0$ both sides are id_V. Suppose the formula holds for n. By Theorem 2.4.2 (d),

$$(S+T)^{n+1}=(S+T)^n(S+T)=\sum_{k=0}^{n}\binom{n}{k}S^kT^{n-k}S+\sum_{k=0}^{n}\binom{n}{k}S^kT^{n-k+1}.$$

By the first paragraph, $S^kT^{n-k}S=S^k(T^{n-k}S)=S^kST^{n-k}=S^{k+1}T^{n-k}$. Shifting the index in the first sum, the coefficient of S^kT^{n+1-k} becomes $\binom{n}{k-1}+\binom{n}{k}=\binom{n+1}{k}$ (Pascal's rule, with $\binom{n}{-1}=\binom{n}{n+1}=0$). This is the formula for $n+1$.

(b) By (a) with $n=a+b-1$, $(S+T)^{a+b-1}$ is a sum of multiples of $S^kT^{a+b-1-k}$, $0\leq k\leq a+b-1$. If $k\geq a$, then $S^k=S^{k-a}S^a=0$ by Theorem 2.4.3 (a), so the term is 0 by Theorem 2.4.2 (d). If $k\leq a-1$, then $a+b-1-k\geq b$, so $T^{a+b-1-k}=0$ likewise. Every term vanishes, so $(S+T)^{a+b-1}=0$.

(c) Let $S(x,y) = (y,0)$ and $T(x,y) = (0,x)$. Then $S^2(x,y) = S(y,0) = (0,0)$ and $T^2(x,y) = T(0,x) = (0,0)$, so both are nilpotent. But $(S+T)(x,y) = (y,x)$, so $(S+T)^2(x,y) = (x,y)$, that is, $(S+T)^2 = \mathrm{id}$. Then $(S+T)^{2k} = \mathrm{id} \neq 0$ and $(S+T)^{2k+1} = S+T \neq 0$ for every k, so $S+T$ is not nilpotent. Here $ST(x,y) = (x,0) \neq (0,y) = TS(x,y)$, so the hypothesis of (b) indeed fails.

2.5 Invertible Maps and Isomorphisms

In Chapter 1 we noticed that a basis of length n turns every vector of V into a column in F^n, and that sums and multiples survive the translation. Then $\mathbb{R}[x]_{\leq 2}$ and $\mathbb{R}^3$ behave "the same", although they are different sets. This section makes "the same" precise. We define invertible linear maps and isomorphisms, prove that finite-dimensional spaces are classified by one number, their dimension, and collect everything we know about invertibility of an operator into one theorem.

2.5.1 When are two spaces the same?

The naive notion of sameness is equality of sets, and it fails at once. The polynomial $3 - x + 2x^2$ is not a column vector, so $\mathbb{R}[x]_{\leq 2} \neq \mathbb{R}^3$. Yet every computation in one space has a twin in the other. Match $a_0 + a_1 x + a_2 x^2$ with (a_0, a_1, a_2):

$$(1+2x) + (x - x^2) = 1 + 3x - x^2 \quad \longleftrightarrow \quad (1,2,0) + (0,1,-1) = (1,3,-1).$$

A reader who only sees one side cannot tell which space the computation happened in. Three features make this work. The matching is **one-to-one** and **surjective**, so nothing is lost or left over. It **respects sums and multiples** in the direction polynomial $\to$ column. And it respects them in the **reverse** direction too, so a conclusion drawn among columns can be carried back. In the language of this chapter: a linear map with a linear map going back.

An isomorphism is a perfect dictionary between two vector spaces: every word has exactly one translation, and translating respects sums and scalar multiples in both directions.

Definition 2.5.1. Invertible Linear Map

Let V and W be vector spaces over F, and let $T \in \mathcal{L}(V,W)$. Then T is **invertible** if there **exists a linear** map $S \in \mathcal{L}(W,V)$ such that

$$ST = \mathrm{id}_V \qquad \textbf{and} \qquad TS = \mathrm{id}_W .$$

In words: S undoes T on V (first T, then S, and we are back where we started), T undoes S on W, and the undoing map is itself linear. The definition asks for **one** map S doing **both** jobs.

Well-definedness. If S exists, it is an inverse **function** of T in the sense of Definition 0.4.7, and inverse functions are unique by Theorem 0.4.7. So we may speak of **the** inverse and write $T^{-1} := S$. The same theorem says an invertible T is a bijection.

Is the word "linear" in the definition doing any work? One could imagine a linear bijection whose inverse function is not linear, which would then fail the definition. It turns out this never happens:

Theorem 2.5.1. The Inverse of a Linear Bijection Is Linear

Let $T \in \mathcal{L}(V, W)$ be bijective. Then its inverse function $T^{-1}\colon W \to V$ is linear. Consequently, a linear map T is invertible if and only if it is bijective.

Idea. To show $T^{-1}(w_1 + w_2) = T^{-1}w_1 + T^{-1}w_2$, we cannot compute T^{-1} directly. But we can apply T, which we know is linear, to the right-hand side, and see that it lands on $w_1 + w_2$. Since T is injective, only one vector does that, namely $T^{-1}(w_1 + w_2)$.

Proof. Let $w_1, w_2 \in W$ and $c \in F$, and put $v_i \coloneqq T^{-1}w_i$, so that $Tv_i = w_i$. Since T is linear,

$$T(v_1 + v_2) = w_1 + w_2, \qquad T(cv_1) = cw_1.$$

Applying T^{-1} to both equations gives $T^{-1}(w_1 + w_2) = v_1 + v_2 = T^{-1}w_1 + T^{-1}w_2$ and $T^{-1}(cw_1) = cv_1 = cT^{-1}w_1$. Hence T^{-1} is linear.

For the last statement: if T is invertible, it is bijective by Theorem 0.4.7. Conversely, if T is bijective, its inverse function $S = T^{-1}$ satisfies $ST = \mathrm{id}_V$ and $TS = \mathrm{id}_W$ by Definition 0.4.7, and S is linear by the first part. So T is invertible. This proves the theorem.

■

So in practice we check invertibility by checking **bijectivity**, and never need to verify linearity of the inverse. Invertible linear maps are important enough to have a second name.

Definition 2.5.2. Isomorphism

Let V and W be vector spaces over **the same field** F. An **isomorphism** from V to W is an invertible linear map $T \in \mathcal{L}(V, W)$; equivalently, by Theorem 2.5.1, a **bijective** linear map.

Definition 2.5.3. Isomorphic Spaces

Vector spaces V and W over F are **isomorphic**, written $V \cong W$, if there **exists an** isomorphism $V \to W$.

In words: $V \cong W$ says that **some** perfect dictionary exists; it does not name one. The name comes from Greek *isos* (equal) and *morphe* (form): the two spaces have the same form, though not the same elements. An isomorphism $V \to V$ is also called an **invertible operator**.

Example. The Standard Isomorphisms

Show that each map is an isomorphism, and give its inverse.

(a) $\Phi\colon \mathbb{R}[x]_{\le n} \to \mathbb{R}^{n+1}$, $a_0 + a_1x + \cdots + a_nx^n \mapsto (a_0, a_1, \ldots, a_n)$.

(b) $\Psi\colon M_{m\times n}(F) \to F^{mn}$, $(a_{ij}) \mapsto (a_{11}, \ldots, a_{1n}, a_{21}, \ldots, a_{mn})$, listing the entries row by row.

(c) $\Theta\colon \mathbb{C} \to \mathbb{R}^2$, $a+bi \mapsto (a,b)$, where $\mathbb{C}$ is a vector space **over** $\mathbb{R}$.

(d) $T_{\mathbf{A}}\colon F^n \to F^n$, $\mathbf{x} \mapsto \mathbf{A}\mathbf{x}$, for an invertible $\mathbf{A} \in M_n(F)$; and id_V for any V.

Solution.

(a) *Linear.* Polynomials are added and scaled coefficient by coefficient, and so are columns, so $\Phi(p+q) = \Phi(p) + \Phi(q)$ and $\Phi(cp) = c\,\Phi(p)$. *Bijective.* The map $(a_0, \dots, a_n) \mapsto a_0 + a_1 x + \cdots + a_n x^n$ is a two-sided inverse function, because two polynomials are equal exactly when all their coefficients agree (Definition 0.8.1). By Theorem 2.5.1, Φ is an isomorphism, and Φ^{-1} is that map. Hence $\mathbb{R}[x]_{\le n} \cong \mathbb{R}^{n+1}$.

(b) Matrices are added and scaled entrywise, and so are vectors of F^{mn}, so Ψ is linear. The map that writes the first n entries of a vector into row 1, the next n into row 2, and so on, undoes Ψ on both sides. Hence $M_{m\times n}(F) \cong F^{mn}$.

(c) For $a, b, a', b', c \in \mathbb{R}$: $(a+bi) + (a'+b'i) = (a+a') + (b+b')i$ and $c(a+bi) = ca + cbi$, so Θ is linear over $\mathbb{R}$. Every complex number has unique real and imaginary parts (Definition 0.6.1), so $(a,b) \mapsto a+bi$ is a two-sided inverse. Hence $\mathbb{C} \cong \mathbb{R}^2$ **as real vector spaces**. In the same way Example (The Complexification of $\mathbb{R}^n$ Is $\mathbb{C}^n$) showed that $(\mathbf{x}, \mathbf{y}) \mapsto \mathbf{x} + i\mathbf{y}$ is a bijection $(\mathbb{R}^n)_{\mathbb{C}} \to \mathbb{C}^n$ respecting complex scalars; in our new language, $(\mathbb{R}^n)_{\mathbb{C}} \cong \mathbb{C}^n$ over $\mathbb{C}$.

(d) $T_{\mathbf{A}}$ is linear (Example (Multiplication by a Matrix)). Since $\mathbf{A}(\mathbf{A}^{-1}\mathbf{x}) = (\mathbf{A}\mathbf{A}^{-1})\mathbf{x} = \mathbf{x}$ and $\mathbf{A}^{-1}(\mathbf{A}\mathbf{x}) = \mathbf{x}$ by Theorem 0.9.4, the linear map $T_{\mathbf{A}^{-1}}$ is a two-sided inverse: $T_{\mathbf{A}}^{-1} = T_{\mathbf{A}^{-1}}$. The identity id_V is its own inverse. This degenerate example matters: it shows $V \cong V$, and even the zero space is isomorphic to itself, via the zero map $\{0\} \to \{0\}$, which is $\mathrm{id}_{\{0\}}$.

Here is a nonexample by minimal change. Take (a) with $n = 3$, but keep only three coefficients: $\Phi'\colon \mathbb{R}[x]_{\le 3} \to \mathbb{R}^3$, $a_0 + a_1 x + a_2 x^2 + a_3 x^3 \mapsto (a_0, a_1, a_2)$. It is still linear, and still surjective, since $(a_0, a_1, a_2) = \Phi'(a_0 + a_1 x + a_2 x^2)$. But $\Phi'(x^3) = \mathbf{0} = \Phi'(0)$, so Φ' is **not injective**. No S can satisfy $S\Phi' = \mathrm{id}$, because $S\Phi'(x^3) = S(\mathbf{0}) = 0 \neq x^3$. The clause $ST = \mathrm{id}_V$ is the one that fails.

Why this definition. Could we drop one of the two equations? The next example says no.

Example. One Equation Is Not Enough

On $\mathbb{R}[x]$, let D be differentiation and $J(p) := \int_0^x p(t)\,dt$, so that $J(x^k) = \frac{x^{k+1}}{k+1}$ (Example (Differentiation), Example (Integration)). On the sequence space $F^{\mathbb{N}}$, let $L(a_0, a_1, a_2, \dots) := (a_1, a_2, a_3, \dots)$ and $R(a_0, a_1, a_2, \dots) := (0, a_0, a_1, \dots)$ be the left and right shifts of Example (The Shift Maps on Sequences). Show that $DJ = \mathrm{id}$ and $LR = \mathrm{id}$, but that none of D, J, L, R is invertible.

Solution. All four maps are linear, by Example (Differentiation), Example (Integration), and Example (The Shift Maps on Sequences). For each $k \in \mathbb{N}$, $DJ(x^k) = D(\frac{x^{k+1}}{k+1}) = x^k$, so DJ and id agree on every x^k. Every polynomial is a finite combination of the x^k, so $DJ = \mathrm{id}$ by Theorem 2.1.4. Directly, $LR(a_0, a_1, \dots) = L(0, a_0, a_1, \dots) = (a_0, a_1, \dots)$, so $LR = \mathrm{id}$.

But $JD(1) = J(0) = 0 \neq 1$, and $RL(1, 0, 0, \dots) = R(0, 0, \dots) = 0$. In fact $D(1) = 0 = D(0)$ and $L(1, 0, 0, \dots) = 0 = L(0)$, so D and L are not injective; and $1 \notin \mathrm{im}\, J$ (every $J(p)$ has

constant coefficient 0) and $(1,0,0,\dots) \notin \operatorname{im} R$, so J and R are not surjective. None of the four is bijective, so none is invertible (Theorem 2.5.1).

So a one-sided inverse can exist without invertibility. The spaces here are infinite-dimensional, and that is not an accident: later in this section we prove that in finite dimension one equation does imply the other.

Warning. **Isomorphic is not equal, and an isomorphism is a choice.** $\mathbb{R}[x]_{\le 2} \cong \mathbb{R}^3$ does not make a polynomial a column. Worse, there are many isomorphisms, and none is singled out. The coefficient map of Example (The Standard Isomorphisms) sends x to $(0,1,0)$. The coordinate map for the basis $(1, 1+x, 1+x+x^2)$ of Example (Coordinates of a Polynomial) is also an isomorphism (we prove this below), and it sends $x = -1\cdot 1 + 1\cdot(1+x)$ to $(-1,1,0)$. A general n-dimensional space has no "standard" basis, so there is no natural isomorphism $F^n \to V$; every one depends on a choice. Chapter 5 returns to this point with dual spaces.

The first payoff of the definition is that an isomorphism carries all linear structure across, in both directions. We record the two facts we need.

Theorem 2.5.2. Isomorphisms Carry Bases to Bases

Let $T\colon V \to W$ be an isomorphism, and let $(v_1,\dots,v_n)$ be a basis of V. Then $(Tv_1,\dots,Tv_n)$ is a basis of W.

Proof. Since T is injective, the list $(Tv_1,\dots,Tv_n)$ is linearly independent by Theorem 2.2.5. By Theorem 2.2.4, it spans $\operatorname{im} T$, and $\operatorname{im} T = W$ because T is surjective. Hence it is a basis of W. ∎

Theorem 2.5.3. Isomorphism Is an Equivalence Relation

For vector spaces over F: $V \cong V$; if $V \cong W$ then $W \cong V$; and if $U \cong V$ and $V \cong W$ then $U \cong W$.

Proof. id_V is an isomorphism (Example (The Standard Isomorphisms) (d)). If $T\colon V \to W$ is an isomorphism, then T^{-1} is linear (Theorem 2.5.1) and has inverse T, so it is an isomorphism $W \to V$. If $T\colon U \to V$ and $S\colon V \to W$ are isomorphisms, then ST is linear by Theorem 2.4.2 and bijective by Theorem 0.4.3, hence an isomorphism by Theorem 2.5.1, with $(ST)^{-1} = T^{-1}S^{-1}$ by Theorem 0.4.8. This proves the theorem. ∎

So $\cong$ sorts vector spaces over F into classes (Definition 0.5.2). The next question is which spaces land in the same class.

2.5.2 Classification by dimension

By Theorem 2.5.2, isomorphic spaces have bases of the same length. Is that the **only** obstruction? Two spaces of dimension 6, say $M_{2\times 3}(\mathbb{R})$ and $\mathbb{R}[x]_{\leq 5}$, look nothing alike. The following theorem says that this does not matter: for finite-dimensional spaces, dimension is a **complete invariant**.

Theorem 2.5.4. Classification of Finite-Dimensional Spaces

Let V and W be finite-dimensional vector spaces over the same field F. Then

$$V \cong W \quad \Longleftrightarrow \quad \dim V = \dim W.$$

Idea. ($\Rightarrow$) is the payoff we just proved: an isomorphism moves a basis to a basis of the same length. For ($\Leftarrow$) we must **build** an isomorphism, and the only material we have is a basis on each side, $(v_1, \ldots, v_n)$ and $(w_1, \ldots, w_n)$. The obvious dictionary sends v_i to w_i. Linear maps can be prescribed freely on a basis, so this defines T, and the reverse prescription defines S. To see $ST = \mathrm{id}_V$ we do not compute: both sides are linear and agree on the basis, so they are equal. The hypothesis "same dimension" is exactly what makes the two bases the same length, so that "$v_i \mapsto w_i$" makes sense.

Proof. ($\Rightarrow$) Suppose $T\colon V \to W$ is an isomorphism, and let $(v_1, \ldots, v_n)$ be a basis of V, with $n = \dim V$ (Corollary 1.5.5). By Theorem 2.5.2, $(Tv_1, \ldots, Tv_n)$ is a basis of W, so $\dim W = n = \dim V$.

($\Leftarrow$) Suppose $\dim V = \dim W = n$, and choose bases $(v_1, \ldots, v_n)$ of V and $(w_1, \ldots, w_n)$ of W (Corollary 1.5.5). By Theorem 2.1.5 there are linear maps $T \in \mathcal{L}(V, W)$ with $Tv_i = w_i$ and $S \in \mathcal{L}(W, V)$ with $Sw_i = v_i$, for $i = 1, \ldots, n$. Then $ST \in \mathcal{L}(V)$ is linear (Theorem 2.4.2) and $STv_i = Sw_i = v_i = \mathrm{id}_V\, v_i$ for every i. Two linear maps that agree on a basis are equal, by the uniqueness part of Theorem 2.1.5, so $ST = \mathrm{id}_V$. Swapping the roles of V and W gives $TS = \mathrm{id}_W$. Hence T is invertible, and $V \cong W$. This proves the theorem. ∎

The theorem needs both spaces to be over the **same** field; this is part of Definition 2.5.3. $\mathbb{C}$ over $\mathbb{C}$ has dimension 1 and $\mathbb{R}$ over $\mathbb{R}$ has dimension 1, but the question "is $\mathbb{C} \cong \mathbb{R}$?" makes no sense, since a linear map must respect one common set of scalars.

Taking $W = F^n$ with its standard basis, the proof builds a specific isomorphism $V \to F^n$: the one sending v_i to $\mathbf{e}_i$. It is an old friend.

Corollary 2.5.5. The Coordinate Isomorphism

Let $\mathcal{B} = (v_1, \ldots, v_n)$ be a basis of a vector space V over F. Then the coordinate map

$$V \to F^n, \qquad v \mapsto [v]_{\mathcal{B}}$$

is an isomorphism, with inverse $(a_1, \ldots, a_n) \mapsto a_1 v_1 + \cdots + a_n v_n$. In particular, every n-dimensional vector space over F is isomorphic to F^n.

Proof. By Theorem 1.5.2, the coordinate map is linear, and every column $(a_1, \ldots, a_n) \in F^n$ is the coordinate vector of exactly one vector of V, namely $a_1v_1 + \cdots + a_nv_n$. So the coordinate map is bijective, with the stated inverse function, and it is an isomorphism by Theorem 2.5.1. ■

This is the promise of Chapter 1 kept: the correspondence $v \mapsto [v]_{\mathcal{B}}$ from Theorem 1.5.2 is an isomorphism. Its practical meaning is the slogan behind the rest of this chapter: **any computation with sums and multiples in an n-dimensional space can be done with columns in F^n**, at the price of choosing a basis.

Example. Sorting Spaces by Dimension

Which of the following real vector spaces are isomorphic to each other?

(i) $\mathbb{R}[x]_{\leq 5}$;
(ii) $M_{2\times 3}(\mathbb{R})$;
(iii) the upper triangular matrices in $M_3(\mathbb{R})$;
(iv) $\mathbb{C}^3$, regarded as a vector space over $\mathbb{R}$;
(v) $U = \{p \in \mathbb{R}[x]_{\leq 6} : p(1) = 0\}$;
(vi) $\mathbb{R}^5$.

Solution. By Theorem 2.5.4 it suffices to compute dimensions over $\mathbb{R}$.

(i) The basis $(1, x, \ldots, x^5)$ gives dimension 6. (ii) $\dim M_{2\times 3}(\mathbb{R}) = 2 \cdot 3 = 6$ (Example (Dimensions of the Running Examples)). (iii) An upper triangular matrix is $\sum_{i\leq j} a_{ij}\mathbf{E}_{ij}$, and this expression is unique because the coefficient of $\mathbf{E}_{ij}$ is the (i, j)-entry. By Theorem 1.5.1 the six matrix units $\mathbf{E}_{ij}$ with $1 \leq i \leq j \leq 3$ form a basis, so the dimension is 6. (iv) $\dim_{\mathbb{C}} \mathbb{C}^3 = 3$, so $\dim_{\mathbb{R}} \mathbb{C}^3 = 6$ by Theorem 1.9.2.

(ii) The evaluation map $E\colon \mathbb{R}[x]_{\leq 6} \to \mathbb{R}$, $p \mapsto p(1)$, is linear, and surjective because $E(c) = c$ for each constant c. Its kernel is U. By Theorem 2.3.1, $\dim U = \dim \mathbb{R}[x]_{\leq 6} - \operatorname{rank} E = 7 - 1 = 6$.

(iii) $\dim \mathbb{R}^5 = 5$.

Hence the spaces (i)–(v) are all isomorphic to each other (and to $\mathbb{R}^6$), while $\mathbb{R}^5$ is isomorphic to none of them. We never wrote down an isomorphism between, say, (iii) and (v); the theorem guarantees one.

Quick check. Is $M_2(\mathbb{R})$ isomorphic to $\mathbb{R}[x]_{\leq 3}$? Is it isomorphic to $\mathbb{R}[x]_{\leq 4}$?

Solution. $\dim M_2(\mathbb{R}) = 4 = \dim \mathbb{R}[x]_{\leq 3}$, so $M_2(\mathbb{R}) \cong \mathbb{R}[x]_{\leq 3}$ by Theorem 2.5.4. But $\dim \mathbb{R}[x]_{\leq 4} = 5 \neq 4$ (not 4: count $1, x, x^2, x^3, x^4$), so $M_2(\mathbb{R}) \ncong \mathbb{R}[x]_{\leq 4}$.

2.5.3 One-sided inverses

To know that T is invertible with inverse S, the definition asks for two equations, $ST = \mathrm{id}$ and $TS = \mathrm{id}$. Often only one is easy to check. When is one enough? Not always: Example (One Equation Is Not Enough) shows that it is not enough for operators on infinite-dimensional spaces. And it fails between spaces of different dimension: for $T\colon \mathbb{R}^2 \to \mathbb{R}^3$, $(x, y) \mapsto (x, y, 0)$, and $S\colon \mathbb{R}^3 \to \mathbb{R}^2$, $(x, y, z) \mapsto (x, y)$, we have $ST = \mathrm{id}_{\mathbb{R}^2}$ but $TS(0, 0, 1) = \mathbf{0}$, so $TS \neq \mathrm{id}_{\mathbb{R}^3}$. What survives is exactly the case that looks like "square":

Theorem 2.5.6. One-Sided Inverses of Linear Maps

Let V and W be finite-dimensional vector spaces over F with $\dim V = \dim W$. Let $T \in \mathcal{L}(V, W)$ and $S \in \mathcal{L}(W, V)$. If $ST = \mathrm{id}_V$, then $TS = \mathrm{id}_W$. In that case T and S are both isomorphisms, $S = T^{-1}$ and $T = S^{-1}$.

Idea. Without any dimension hypothesis, $ST = \mathrm{id}_V$ already gives half of what we need: T must be injective, since S recovers v from Tv. The missing half, surjectivity, is exactly what the Rank–Nullity Theorem (Theorem 2.3.1) supplies for free when $\dim V = \dim W$ is finite. Once T is known to be invertible, we cancel it: $S = (ST)T^{-1} = T^{-1}$. So the one hypothesis the infinite-dimensional counterexamples violate, finite equal dimension, is spent on a single line: injective $\Rightarrow$ surjective.

Proof. Suppose $ST = \mathrm{id}_V$, and let $v \in \ker T$. Then $v = \mathrm{id}_V\, v = S(Tv) = S0 = 0$, since S is linear. So $\ker T = \{0\}$, and T is injective by Theorem 2.2.3. Since V and W are finite-dimensional with $\dim V = \dim W$, Corollary 2.3.2 shows that T is also surjective. Hence T is bijective, and by Theorem 2.5.1 it has a linear inverse $T^{-1} \in \mathcal{L}(W, V)$. By Theorem 2.4.2 (b) and (c),

$$S = S\,\mathrm{id}_W = S(TT^{-1}) = (ST)T^{-1} = \mathrm{id}_V\, T^{-1} = T^{-1}.$$

Therefore $TS = TT^{-1} = \mathrm{id}_W$. Finally $S = T^{-1}$ is an isomorphism with inverse T (Theorem 2.5.3), so $T = S^{-1}$. This proves the theorem. ∎

Remark. The whole weight of the theorem falls on one step, "injective, hence bijective", and that step is Rank–Nullity and nothing else. Chapter 3 proves the matrix counterpart — if $\mathbf{AB} = \mathbf{I}_n$ for square $\mathbf{A}$ and $\mathbf{B}$, then $\mathbf{BA} = \mathbf{I}_n$ — by row reduction instead, reaching the same place by a different road. Once maps are written as matrices in Chapter 4, the two become one theorem seen from two sides.

In practice this means: **to verify that S is the inverse of T between spaces of the same finite dimension, one composition suffices.** By symmetry, $TS = \mathrm{id}_W$ also implies $ST = \mathrm{id}_V$: apply the theorem with the roles of T and S (and of V and W) swapped.

2.5.4 The Invertible Operator Theorem

For an operator on a finite-dimensional space, invertibility can now be detected in many different ways, and it is worth having them in one place. The theorem below is the first of several such lists in this book, and the others all grow from it: Chapter 3 gives the matrix version, and later chapters add tests by determinant, by eigenvalue, and by singular value.

Theorem 2.5.7. Invertible Operator Theorem

Let V be a **finite-dimensional** vector space over F, and let $T \in \mathcal{L}(V)$. The following are equivalent.

(a) T is invertible.
(b) T is injective.
(c) $\ker T = \{0\}$, that is, $\operatorname{nullity} T = 0$.
(d) T is surjective.
(e) $\operatorname{rank} T = \dim V$.
(f) There is $S \in \mathcal{L}(V)$ with $ST = \mathrm{id}_V$.
(g) There is $S \in \mathcal{L}(V)$ with $TS = \mathrm{id}_V$.
(h) T maps **some** basis of V to a basis of V.
(i) T maps **every** basis of V to a basis of V.

Idea. Nine conditions, but almost every arrow is a citation. The plan:

① the core, (a) ⇔ (b) ⇔ (c) ⇔ (d) ⇔ (e): kernel test for injectivity, Rank–Nullity for "injective ⇔ surjective", equal dimension for "image is everything";

② the one-sided inverses (f), (g): trivially implied by (a), and they imply (a) by Theorem 2.5.6 with $W = V$;

③ the bases, (a) ⇒ (i) ⇒ (h) ⇒ (d): isomorphisms carry bases to bases, some basis exists, and an image containing a basis is everything.

Proof. **(a) ⇒ (b).** An invertible map is bijective (Theorem 2.5.1).

(b) ⇔ (c). This is Theorem 2.2.3, and $\ker T = \{0\}$ is the same as $\dim \ker T = 0$.

(b) ⇔ (d). Since T maps the finite-dimensional space V to a space of the same dimension, this is Corollary 2.3.2.

(d) ⇔ (e). If $\operatorname{im} T = V$, then $\operatorname{rank} T = \dim V$. Conversely, if $\operatorname{rank} T = \dim V$, then $\operatorname{im} T$ is a subspace of V of dimension $\dim V$, so $\operatorname{im} T = V$ by Theorem 1.6.8.

(b) and (d) ⇒ (a). If T is injective, it is also surjective by (b) ⇒ (d), hence bijective, hence invertible by Theorem 2.5.1. Together with the arrows above, (a)–(e) are equivalent.

(a) ⇒ (f) and (a) ⇒ (g). Take $S = T^{-1}$.

(f) ⇒ (a) and (g) ⇒ (a). Apply Theorem 2.5.6 with $W = V$: if $ST = \mathrm{id}_V$, then T is invertible; if $TS = \mathrm{id}_V$, the same theorem with the roles of S and T swapped shows that S

is invertible with $S^{-1} = T$, so T is invertible.

(a) $\Rightarrow$ (i). This is Theorem 2.5.2.

(i) $\Rightarrow$ (h). V has a basis by Corollary 1.5.5, and by (i) its image is a basis.

(h) $\Rightarrow$ (d). Let $(v_1, \dots, v_n)$ be a basis whose image $(Tv_1, \dots, Tv_n)$ is a basis of V. Each Tv_i lies in the subspace $\operatorname{im} T$, so $V = \operatorname{span}(Tv_1, \dots, Tv_n) \subseteq \operatorname{im} T$ by Theorem 1.3.1. Hence $\operatorname{im} T = V$.

All nine conditions are equivalent. This proves the theorem.

■

As with the matrix theorem, the list is a menu. To prove that an operator **is** invertible, pick the cheapest item, usually (c): take v with $Tv = 0$ and show $v = 0$. To prove that it is **not**, exhibit one nonzero vector in the kernel.

Example. Invertibility Without Solving Equations

(a) Let $T \in \mathcal{L}(\mathbb{R}[x]_{\leq 2})$, $T(p) = p + xp'$. Show that T is invertible and find T^{-1}.

(b) Let V be **any** vector space and $T \in \mathcal{L}(V)$ with $T^2 = T + \mathrm{id}_V$. Show that T is invertible, with $T^{-1} = T - \mathrm{id}_V$.

Solution.

(a) Let $p = a + bx + cx^2$. Then $xp' = bx + 2cx^2$, so

$$T(a + bx + cx^2) = a + 2bx + 3cx^2.$$

If $T(p) = 0$, then $a = 2b = 3c = 0$, so $a = b = c = 0$ and $p = 0$. Hence $\ker T = \{0\}$, and T is invertible by Theorem 2.5.7 ((c) $\Rightarrow$ (a)). We did not have to check surjectivity: the count did it for us. For the inverse, let $S(a + bx + cx^2) := a + \frac{b}{2}x + \frac{c}{3}x^2$, which is linear. Then $TS(a + bx + cx^2) = a + 2 \cdot \frac{b}{2}x + 3 \cdot \frac{c}{3}x^2 = a + bx + cx^2$, so $TS = \mathrm{id}$, and by Theorem 2.5.6, $S = T^{-1}$.

(b) The operators T and $T - \mathrm{id}_V$ are polynomials in T, so they commute (Theorem 2.4.3 (c)). By Theorem 2.4.2 (d) and the hypothesis,

$$T(T - \mathrm{id}_V) = T^2 - T = \mathrm{id}_V = (T - \mathrm{id}_V)T.$$

So $T - \mathrm{id}_V$ is a two-sided inverse, and T is invertible by Definition 2.5.1. Here V may be infinite-dimensional, since we verified **both** equations.

Warning. **The Invertible Operator Theorem needs finite dimension, and it is about operators.** On $\mathbb{R}[x]$, differentiation D is surjective (every polynomial has an antiderivative $J(p)$) but not injective, and $S(p) = xp$ is injective but not surjective ($1 \notin \operatorname{im} S$). And for a map $T\colon V \to W$ between **different** spaces, "injective" does not give "surjective" unless $\dim V = \dim W$: the inclusion $(x, y) \mapsto (x, y, 0)$ of $\mathbb{R}^2$ into $\mathbb{R}^3$ is injective and not surjective.

Quick check. Is $T \in \mathcal{L}(M_2(\mathbb{R}))$, $T(\mathbf{A}) = \mathbf{A} + \mathbf{A}^\top$, invertible? Use the cheapest item of Theorem 2.5.7.

Solution. No. Let $\mathbf{A} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \neq \mathbf{O}$. Then $\mathbf{A}^\top = -\mathbf{A}$, so $T(\mathbf{A}) = \mathbf{O}$, and $\ker T \neq \{\mathbf{O}\}$. Item (c) fails, so T is not invertible.

Isomorphisms let us move a whole space into F^n. Chapter 4 moves **maps** as well: once bases are chosen, every linear map between finite-dimensional spaces becomes a matrix, and composition becomes matrix multiplication.

2.5.5 Exercises

2.5.5.1 A. Check your understanding

Exercise. A1

(a) Define what it means for $T \in \mathcal{L}(V, W)$ to be an isomorphism, and what $V \cong W$ means.

(b) True or false: $\mathbb{R}[x]_{\leq 3} \cong M_2(\mathbb{R})$. Justify your answer.

(c) True or false: if $\dim V = \dim W = 5$ and $T \in \mathcal{L}(V, W)$ is injective, then T is an isomorphism. Justify your answer.

(d) True or false: for every vector space V and all $S, T \in \mathcal{L}(V)$, $ST = \mathrm{id}_V$ implies $TS = \mathrm{id}_V$. Justify your answer.

(e) True or false: if $V \cong W$, then every linear map $V \to W$ is an isomorphism. Justify your answer.

(f) Which theorem would you use to show that a given operator on a finite-dimensional space is invertible by computing only its kernel?

Solution.

(a) T is an isomorphism if it is invertible, that is, there is $S \in \mathcal{L}(W, V)$ with $ST = \mathrm{id}_V$ and $TS = \mathrm{id}_W$; equivalently, T is a bijective linear map (Theorem 2.5.1). $V \cong W$ means that some isomorphism $V \to W$ exists.

(b) True. Both have dimension 4, so they are isomorphic by Theorem 2.5.4.

(c) True. By Corollary 2.3.2 an injective linear map between spaces of the same finite dimension is surjective, so T is bijective, hence an isomorphism.

(d) False. On $F^{\mathbb{N}}$, the shifts of Example (One Equation Is Not Enough) satisfy $LR = \mathrm{id}$ but $RL \neq \mathrm{id}$. The statement is true when V is finite-dimensional (Theorem 2.5.6).

(e) False. $\mathbb{R}^2 \cong \mathbb{R}^2$, but the zero map $\mathbb{R}^2 \to \mathbb{R}^2$ is not injective.

(f) The Invertible Operator Theorem, Theorem 2.5.7, (c) $\Rightarrow$ (a).

2.5.5.2 B. Practice

Exercise. B1: An explicit isomorphism

Let $U = \{\mathbf{A} \in M_2(\mathbb{R}) : \mathbf{A}^\top = \mathbf{A}\}$ be the space of symmetric 2×2 real matrices. Find an explicit isomorphism $U \to \mathbb{R}[x]_{\le 2}$, and verify that it is one.

Solution. Every $\mathbf{A} \in U$ has the form $\begin{pmatrix} a & b \\ b & c \end{pmatrix}$ with $a, b, c \in \mathbb{R}$. Define

$$T\begin{pmatrix} a & b \\ b & c \end{pmatrix} := a + bx + cx^2, \qquad S(a + bx + cx^2) := \begin{pmatrix} a & b \\ b & c \end{pmatrix}.$$

Linearity of T. Adding two symmetric matrices adds the entries a, b, c, and adding polynomials adds coefficients, so $T(\mathbf{A} + \mathbf{A}') = T(\mathbf{A}) + T(\mathbf{A}')$; similarly $T(\lambda\mathbf{A}) = \lambda T(\mathbf{A})$.

Inverse. S takes values in U, and $ST(\mathbf{A}) = \mathbf{A}$ for every $\mathbf{A} \in U$, since $\mathbf{A}$ is determined by its entries a, b, c. Also $TS(p) = p$ for every p, since p is determined by its coefficients. So T has a two-sided inverse and is bijective, hence an isomorphism by Theorem 2.5.1. (Alternatively: both spaces have dimension 3 by Example (Dimensions of the Running Examples), so by Theorem 2.5.6 the single equation $ST = \mathrm{id}$ would already suffice.)

Exercise. B2: Inverting $\mathrm{id} + D$

Let $n \ge 0$, and let $T \in \mathcal{L}(\mathbb{R}[x]_{\le n})$, $T(p) = p + p'$.

(a) Show that T is an isomorphism by computing its kernel.

(b) Show that $T^{-1} = \sum_{k=0}^{n} (-1)^k D^k$, where D is differentiation.

(c) For $n = 2$, find the polynomial p with $p + p' = x^2$.

Hint: for (b), use the polynomial identity $(1 + x)\sum_{k=0}^{n}(-x)^k = 1 + (-1)^n x^{n+1}$.

Solution.

(a) Let $p \in \ker T$, so $p' = -p$. If $p \neq 0$, let $d = \deg p \ge 0$. Differentiation lowers the degree of a nonzero polynomial by at least one (it kills constants), so $\deg p' < d = \deg(-p)$, contradicting $p' = -p$. Hence $p = 0$ and $\ker T = \{0\}$. By Theorem 2.5.7 ((c) $\Rightarrow$ (a)), T is an isomorphism.

(b) $T = \mathrm{id} + D = q(D)$ with $q = 1 + x$, and $S := \sum_{k=0}^{n}(-1)^k D^k = r(D)$ with $r = \sum_{k=0}^{n}(-x)^k$. The identity in the hint is the telescoping sum $\sum_{k=0}^{n}(-x)^k - \sum_{k=0}^{n}(-x)^{k+1} = 1 - (-x)^{n+1}$. By Theorem 2.4.3 (b),

$$TS = (qr)(D) = \mathrm{id} + (-1)^n D^{n+1} = \mathrm{id},$$

since $D^{n+1} = 0$ on $\mathbb{R}[x]_{\le n}$ (Example (Differentiation Is Nilpotent)). By Theorem 2.5.6, $S = T^{-1}$.

(c) $p = T^{-1}(x^2) = x^2 - D(x^2) + D^2(x^2) = x^2 - 2x + 2$. Check: $p + p' = (x^2 - 2x + 2) + (2x - 2) = x^2$.

Exercise. B3: Deciding isomorphism

Determine which of the following pairs of real vector spaces are isomorphic. Justify your answer.

(a) $\mathbb{R}^4$ and $M_2(\mathbb{R})$.
(b) $\mathbb{R}[x]_{\leq 3}$ and $\{\mathbf{A} \in M_3(\mathbb{R}) : \mathbf{A}^\top = -\mathbf{A}\}$.
(c) $\{p \in \mathbb{R}[x]_{\leq 4} : p(2) = 0\}$ and $\mathbb{R}^4$.
(d) $\mathbb{C}^2$ (over $\mathbb{R}$) and $\mathbb{R}[x]_{\leq 3}$.
(e) The plane $\{(x, y, z) \in \mathbb{R}^3 : x + y + z = 0\}$ and $\mathbb{C}$ (over $\mathbb{R}$).

Solution. By Theorem 2.5.4, in each case we compare dimensions over $\mathbb{R}$.

(a) Isomorphic: $\dim \mathbb{R}^4 = 4 = \dim M_2(\mathbb{R})$.
(b) Not isomorphic. $\dim \mathbb{R}[x]_{\leq 3} = 4$. If $\mathbf{A}^\top = -\mathbf{A}$, then $a_{ii} = -a_{ii}$, so $2a_{ii} = 0$ and $a_{ii} = 0$, and $a_{ji} = -a_{ij}$. Hence $\mathbf{A} = a_{12}(\mathbf{E}_{12} - \mathbf{E}_{21}) + a_{13}(\mathbf{E}_{13} - \mathbf{E}_{31}) + a_{23}(\mathbf{E}_{23} - \mathbf{E}_{32})$, and the coefficients are entries of $\mathbf{A}$, so this expression is unique. By Theorem 1.5.1 these three matrices form a basis, and the dimension is $3 \neq 4$.
(c) Isomorphic. The evaluation $p \mapsto p(2)$, $\mathbb{R}[x]_{\leq 4} \to \mathbb{R}$, is linear and surjective (constants), with kernel the given space, so by Theorem 2.3.1 its dimension is $5 - 1 = 4$.
(d) Isomorphic: $\dim_{\mathbb{R}} \mathbb{C}^2 = 2 \dim_{\mathbb{C}} \mathbb{C}^2 = 4$ by Theorem 1.9.2, and $\dim \mathbb{R}[x]_{\leq 3} = 4$.
(e) Isomorphic. The plane is the kernel of the surjective linear map $(x, y, z) \mapsto x + y + z$ onto $\mathbb{R}$, so it has dimension $3 - 1 = 2$ by Theorem 2.3.1; and $\dim_{\mathbb{R}} \mathbb{C} = 2$ (Example (Dimensions of the Running Examples)).

2.5.5.3 C. Going deeper

Exercise. C1: Invertible products

(a) Let V be finite-dimensional and $S, T \in \mathcal{L}(V)$. Prove that if ST is invertible, then S and T are both invertible. Deduce that TS is then invertible too.
(b) Show that (a) fails for $V = F^{\mathbb{N}}$: give $S, T \in \mathcal{L}(F^{\mathbb{N}})$ with ST invertible but neither S nor T invertible, and TS not invertible.

Hint: for (a), regroup $(ST)(ST)^{-1}$ and $(ST)^{-1}(ST)$.

Solution.

(a) The inverse $(ST)^{-1}$ lies in $\mathcal{L}(V)$ (Theorem 2.5.1), and $(ST)(ST)^{-1} = \mathrm{id}_V = (ST)^{-1}(ST)$. By associativity (Theorem 2.4.2 (b)), $S(T(ST)^{-1}) = \mathrm{id}_V$ and $((ST)^{-1}S)T = \mathrm{id}_V$. By Theorem 2.5.7, (g) $\Rightarrow$ (a) applied to S with right inverse $T(ST)^{-1}$ shows that S is invertible, and (f) $\Rightarrow$ (a) applied to T with left inverse $(ST)^{-1}S$ shows that T is invertible. Then TS is a composition of isomorphisms, hence an isomorphism by Theorem 2.5.3.
(b) Let $S = L$ and $T = R$ be the left and right shifts of Example (One Equation Is Not Enough). Then $ST = LR = \mathrm{id}$, which is invertible. But L is not injective and R is not surjective, so neither is invertible, and $TS = RL$ sends $(1, 0, 0, \dots)$ to 0, so it is not

injective and not invertible. The proof of (a) breaks at the appeal to Theorem 2.5.7, which needs finite dimension.

Exercise. C2: A space isomorphic to a proper subspace

(a) Let V be finite-dimensional and U a subspace of V with $U \cong V$. Prove that $U = V$.

(b) Let $U = \{p \in F[x] : p(0) = 0\}$, the polynomials with constant coefficient 0. Show that $T\colon F[x] \to U$, $T(p) = xp$, is an isomorphism. Hence (a) fails without finite dimension.

Solution.

(a) U is finite-dimensional by Theorem 1.6.7. Since $U \cong V$, Theorem 2.5.4 gives $\dim U = \dim V$, and then $U = V$ by Theorem 1.6.8.

(b) For $p = \sum_k a_k x^k$, $xp = \sum_k a_k x^{k+1}$ has constant coefficient 0, so T maps into U, and T is linear because multiplication in $F[x]$ distributes and commutes with scalars (Theorem 0.8.1). *Injective:* if $xp = 0$, then $p = 0$ by Corollary 0.8.4, since $x \neq 0$. *Surjective:* if $q = \sum_{k\geq 1} b_k x^k \in U$, then $q = x \sum_{k\geq 1} b_k x^{k-1} = T(\sum_{k\geq 1} b_k x^{k-1})$. So T is a bijective linear map, an isomorphism by Theorem 2.5.1, and $U \cong F[x]$. But $1 \notin U$, so $U \neq F[x]$. This is the phenomenon noted in Chapter 1, §8, now as an isomorphism.

Chapter 3

Linear Systems and Matrices

Chapter 1 kept asking questions that end in a system of equations. Is this vector in the span of those? Is this list independent? Extend this list to a basis. Each time we solved a small system by hand and promised a systematic method. This chapter keeps that promise. Every question about finitely many vectors in F^n turns into a linear system, and linear systems have a complete algorithm: Gaussian elimination.

We do not treat the algorithm as a recipe. We prove that row operations never change the answer, that the fully reduced form of a matrix does not depend on the choices made along the way, and that the numbers it produces, pivots and free variables, measure exactly the dimensions of Chapter 1. Along the way the invertible matrices get their first full characterization, and we prove a fact Chapter 0 had to postpone: for square matrices, $\mathbf{AB} = \mathbf{I}$ already forces $\mathbf{BA} = \mathbf{I}$.

What you need. From Chapter 0, matrix multiplication read column by column (Theorem 0.9.3) and invertible matrices (Definition 0.9.13). From Chapter 1, span (Definition 1.3.2), linear independence (Definition 1.4.1), bases (Definition 1.5.1), and dimension (Definition 1.6.1).

3.0.1 Roadmap

- **Three views of a linear system.** Intersecting lines and planes, combinations of columns, and a single matrix equation; the shape of the solution set as a particular solution plus the homogeneous solutions.
- **Gaussian elimination.** Row operations, echelon and reduced echelon forms, and how to read the solutions off the reduced form.
- **Uniqueness of the reduced form.** Different sequences of row operations always reach the same reduced matrix, so pivot columns are well defined.
- **Elementary matrices and inverses.** Row operations as left multiplication, the invertible matrix theorem, computing inverses by row reduction, and the permutation matrices that later chapters use to relabel coordinates.
- **Rank.** Row space, column space, and null space; row rank equals column rank; the Rouché–Capelli theorem; and a toolkit for the questions of Chapter 1.
- **LU factorization.** Storing elimination as a product of triangular matrices, with and without row swaps.
- **Applications.** Polynomial interpolation, balancing chemical equations, networks, and a puzzle over $\mathbb{F}_2$.

3.0.2 Named moves

- **Row operations do not change the solution set.** So simplify first, then read the answer.
- **Put it in reduced form, then read off.** Consistency, free variables, bases of the column, row, and null spaces all come from the reduced row echelon form.
- **A question about vectors is a question about a matrix.** Put the vectors in as columns; independence, spanning, and membership become statements about pivots.
- **Pivots count dimensions.** The number of pivots is the rank, and the free variables count the dimension of the null space.

3.1 Three Views of a Linear System

Chapter 1 kept running into the same kind of question. Is this vector in the span of those? Is this list independent? Every time, the question turned into a handful of equations in unknown coefficients, and we solved them by hand. This chapter gives those equations a name, looks at them from three sides, and turns solving them into an algorithm with a proof behind it. This first section sets up the language and answers a structural question before any computing: what can the set of all solutions look like?

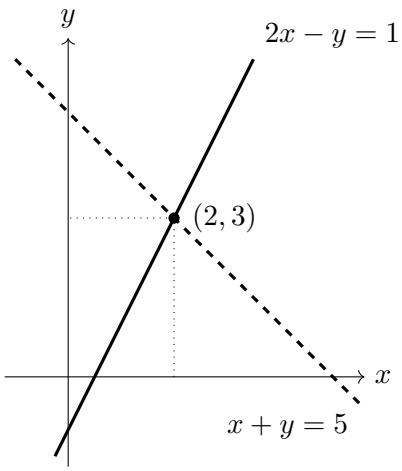
3.1.1 One system, three pictures

Here is a system of two equations in two unknowns, over $\mathbb{R}$:

$$\begin{aligned} 2x - y &= 1, \\ x + y &= 5. \end{aligned}$$

Adding the equations gives $3x = 6$, so $x = 2$, and then $y = 5 - x = 3$. We can read this one small system in three different ways, and each will be useful.

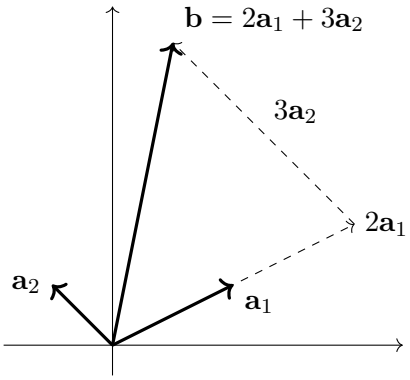
The row picture. Each equation describes a line in the plane: $y = 2x - 1$ and $y = 5 - x$. A solution is a point on **both** lines, so solving the system means intersecting them. The lines meet at $(2, 3)$.



The column picture. Stack the coefficients of x into one column and those of y into another. The system says

$$x \begin{pmatrix} 2 \\ 1 \end{pmatrix} + y \begin{pmatrix} -1 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 5 \end{pmatrix}.$$

Now the unknowns are **weights**. We are asked how much of the vector $\mathbf{a}_1 = (2, 1)$ and how much of $\mathbf{a}_2 = (-1, 1)$ to combine to reach $\mathbf{b} = (1, 5)$. The answer $x = 2$, $y = 3$ says: walk twice along $\mathbf{a}_1$, then three times along $\mathbf{a}_2$.



The matrix picture. Collect the coefficients into a matrix and the unknowns into a column. By the definition of matrix multiplication (Definition 0.9.9), the system is the single equation

$$\begin{pmatrix} 2 & -1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 1 \\ 5 \end{pmatrix}, \qquad \text{that is,} \qquad \mathbf{Ax} = \mathbf{b}.$$

The three pictures say the same thing, and they are tied together by one fact from Chapter 0: $\mathbf{Ax} = x_1\mathbf{a}_1 + \cdots + x_n\mathbf{a}_n$, a combination of the columns of $\mathbf{A}$ weighted by the entries of $\mathbf{x}$ (Theorem 0.9.3). The row picture is the one we draw. The column picture connects to Chapter 1, since it asks whether $\mathbf{b}$ lies in a span. The matrix picture is the one we compute and prove things with.

Remark. With more unknowns the pictures can no longer be drawn, but they do not change. An equation in three unknowns whose coefficients are not all zero describes a plane in $\mathbb{R}^3$, and a system asks where several planes meet. In the column picture, three unknowns mean three columns, and we ask whether a combination of them hits $\mathbf{b}$.

3.1.2 Linear systems

The example had two equations and two unknowns over $\mathbb{R}$. The general notion allows any numbers of each, and any field. Allowing any field is not generality for its own sake. Systems over $\mathbb{F}_2$ describe parity checks and switching puzzles, and systems over $\mathbb{Q}$ are what a computer can solve exactly.

A linear system is a finite list of equations, each saying that a fixed combination of the unknowns equals a fixed scalar.

Definition 3.1.1. Linear system

Let F be a field and let $m, n \geq 1$. A **system of m linear equations in n unknowns**

$x_1, \dots, x_n$ **over** F is a list of equations

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1, \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2, \\ &\vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= b_m, \end{aligned}$$

where all a_{ij} and b_i lie in F.

- The **coefficient matrix** is $\mathbf{A} = (a_{ij}) \in M_{m\times n}(F)$, the **right-hand side** is $\mathbf{b} = (b_1, \dots, b_m) \in F^m$, and the system is written $\mathbf{Ax} = \mathbf{b}$.
- The **augmented matrix** is the $m \times (n+1)$ matrix $[\,\mathbf{A} \mid \mathbf{b}\,]$ obtained by appending $\mathbf{b}$ to $\mathbf{A}$ as a last column.
- A **solution** is a vector $\mathbf{s} = (s_1, \dots, s_n) \in F^n$ such that **every** equation holds when $x_j = s_j$ for all j; equivalently, $\mathbf{As} = \mathbf{b}$. The **solution set** is the set of all solutions, a subset of F^n.
- The system is **consistent** if it has **at least one** solution, and **inconsistent** if its solution set is empty.

In words: the coefficient matrix records the left-hand sides, row i for equation i and column j for unknown x_j. The augmented matrix records the whole system, and it is the object we will manipulate in the next section. A solution is a single vector of F^n, and it must satisfy all m equations at once, not just some of them. "Equivalently, $\mathbf{As} = \mathbf{b}$" holds because the i-th entry of $\mathbf{As}$ is $a_{i1}s_1 + \cdots + a_{in}s_n$ by Definition 0.9.9, and two columns are equal exactly when all their entries agree.

Two small words deserve attention. A solution lives in F^n, with entries in **the field** F: the equation $2x = 1$ has the solution $\frac{1}{2}$ over $\mathbb{Q}$, but over $\mathbb{F}_2$ it reads $0 = 1$ and has none. And "consistent" means **at least one** solution, not exactly one.

Example. None, one, or a whole line

Find the solution set of each system over $\mathbb{R}$, and describe it in the row picture.

(a) $x + 2y = 1$, $2x + 4y = 5$.

(b) $2x - y = 1$, $x + y = 5$.

(c) $x + 2y = 1$, $2x + 4y = 2$.

Solution.

(a) Suppose (x, y) is a solution. Doubling the first equation gives $2x + 4y = 2$, while the second says $2x + 4y = 5$. Then $2 = 5$, which is false. So there is no solution, and the system is **inconsistent**. In the row picture, the lines $y = \frac{1}{2} - \frac{1}{2}x$ and $y = \frac{5}{4} - \frac{1}{2}x$ have the same slope and different intercepts: they are parallel and never meet.

(b) As computed at the start of the section, any solution has $x = 2$ and $y = 3$; conversely $2 \cdot 2 - 3 = 1$ and $2 + 3 = 5$. The solution set is $\{(2, 3)\}$, the single point where two nonparallel lines cross.

(c) The second equation is twice the first, so a pair (x, y) satisfies both exactly when it satisfies $x + 2y = 1$, that is, when $x = 1 - 2y$. Writing t for y, the solution set is $\{(1 - 2t, t) : t \in \mathbb{R}\}$. Both equations describe the **same** line, and every point on it is a

solution.

The three answers are the three things two lines can do: miss, cross, or coincide. Over $\mathbb{R}$, and over any infinite field, no other sizes are possible: once we know the shape of the solution set (Theorem 3.1.3 below), this is a short argument, which you will give in Exercise (C1: How many solutions?). For instance, no system over $\mathbb{R}$ has exactly two solutions. Over a finite field the picture is different.

Example. Exactly two solutions over $\mathbb{F}_2$

Find all solutions of the system $x + y + z = 1$, $y + z = 0$ over $\mathbb{F}_2$. Then compare with the same system over $\mathbb{R}$.

Solution. In $\mathbb{F}_2$ we have $-1 = 1$, so the second equation says $y = -z = z$. Substituting into the first gives $x + 2z = 1$, and $2z = 0$ in $\mathbb{F}_2$, so $x = 1$. Conversely, every $(1, z, z)$ satisfies both equations: $1+z+z = 1+2z = 1$ and $z+z = 0$. Since z can only be 0 or 1, the solution set is

$$\{(1,0,0),\ (1,1,1)\},$$

with exactly two elements.

Over $\mathbb{R}$ the second equation gives $y = -z$, and then the first gives $x = 1$. The solution set is $\{(1, -t, t) : t \in \mathbb{R}\}$, a line in $\mathbb{R}^3$ with infinitely many points. The equations are the same, but the field decides how many values the free unknown can take.

Now a nonexample, by a minimal change. The equation $x+y^2 = 1$ looks like one of the equations above, and it still has a left side built from the unknowns and a constant on the right. But the unknown y appears squared, so the left side is not of the form $a_1x_1 + a_2x_2$ with **fixed scalars** a_1, a_2. The clause that fails is "a fixed combination of the unknowns". Its solution set in $\mathbb{R}^2$ is a parabola, not a line, and nothing in this chapter applies to it.

Warning. **"Consistent" does not mean "has exactly one solution".** The system $x + 2y = 1$, $2x + 4y = 2$ of Example (None, one, or a whole line) (c) is consistent, and it has infinitely many solutions. Consistency only rules out the empty solution set. Whether the solution is unique is a separate question, answered by Corollary 3.1.4 below.

Why record the augmented matrix at all, when $\mathbf{Ax} = \mathbf{b}$ is shorter? Because solving a system never touches the names of the unknowns. Adding two equations adds their coefficients and their right-hand sides, and nothing else. The augmented matrix keeps exactly this data and throws away the symbols $x_1, \ldots, x_n$ and the plus signs. In the next section every manipulation is performed on it.

One kind of system is always consistent, because it has an obvious solution.

Definition 3.1.2. Homogeneous system

A linear system $\mathbf{Ax} = \mathbf{b}$ over F is **homogeneous** if $\mathbf{b} = \mathbf{0}$. The vector $\mathbf{0} \in F^n$ is a solution of every homogeneous system $\mathbf{Ax} = \mathbf{0}$, called the **trivial solution**. For any system $\mathbf{Ax} = \mathbf{b}$,

the system $\mathbf{Ax} = \mathbf{0}$, with the same coefficient matrix, is its **associated homogeneous system**.

The trivial solution exists by the zero rule of Theorem 0.9.4: $\mathbf{A0} = \mathbf{0}$. So for a homogeneous system the interesting question is never "is there a solution?" but "is there a **nontrivial** one?". For example, $x + 2y = 0$, $2x + 4y = 0$ has the nontrivial solution $(-2, 1)$, while $2x - y = 0$, $x + y = 0$ has only the trivial one: adding the equations gives $3x = 0$, so $x = 0$, and then $y = -x = 0$.

Quick check. Over $\mathbb{R}$, write the system $x_1 - x_2 + 3x_3 = 4$, $2x_2 + x_3 = 0$ in the matrix picture and the column picture, and give its augmented matrix. Is $(4, 0, 0)$ a solution? Is $(1, 0, 1)$?

Solution. With $\mathbf{A} = \begin{pmatrix} 1 & -1 & 3 \\ 0 & 2 & 1 \end{pmatrix}$ and $\mathbf{b} = (4, 0)$, the system is $\mathbf{Ax} = \mathbf{b}$. In the column picture it reads $x_1(1, 0) + x_2(-1, 2) + x_3(3, 1) = (4, 0)$. The augmented matrix is

$$[\,\mathbf{A} \mid \mathbf{b}\,] = \left[\begin{array}{ccc|c} 1 & -1 & 3 & 4 \\ 0 & 2 & 1 & 0 \end{array}\right].$$

For $(4, 0, 0)$: $4 - 0 + 0 = 4$ and $0 + 0 = 0$, so it is a solution. For $(1, 0, 1)$: the first equation gives $1 - 0 + 3 = 4$, but the second gives $0 + 1 = 1 \neq 0$. One failing equation is enough, so it is not a solution.

3.1.3 Consistency is a question about span

The column picture turns "does this system have a solution?" into a question from Chapter 1.

Theorem 3.1.1. Consistency and the span of the columns

Let $\mathbf{A} \in M_{m\times n}(F)$ have columns $\mathbf{a}_1, \dots, \mathbf{a}_n \in F^m$, and let $\mathbf{b} \in F^m$. Then the system $\mathbf{Ax} = \mathbf{b}$ is consistent **if and only if** $\mathbf{b} \in \operatorname{span}(\mathbf{a}_1, \dots, \mathbf{a}_n)$. More precisely, $\mathbf{s} = (s_1, \dots, s_n)$ is a solution exactly when $\mathbf{b} = s_1\mathbf{a}_1 + \cdots + s_n\mathbf{a}_n$.

Proof. By Theorem 0.9.3, $\mathbf{As} = s_1\mathbf{a}_1 + \cdots + s_n\mathbf{a}_n$ for every $\mathbf{s} \in F^n$. Hence $\mathbf{s}$ is a solution, that is $\mathbf{As} = \mathbf{b}$, exactly when $\mathbf{b} = s_1\mathbf{a}_1 + \cdots + s_n\mathbf{a}_n$. This proves the second statement.

($\Rightarrow$) If the system is consistent, it has a solution $\mathbf{s}$, and then $\mathbf{b} = s_1\mathbf{a}_1 + \cdots + s_n\mathbf{a}_n$ is a linear combination of the columns. By Definition 1.3.2, $\mathbf{b} \in \operatorname{span}(\mathbf{a}_1, \dots, \mathbf{a}_n)$.

($\Leftarrow$) If $\mathbf{b} \in \operatorname{span}(\mathbf{a}_1, \dots, \mathbf{a}_n)$, then $\mathbf{b} = s_1\mathbf{a}_1 + \cdots + s_n\mathbf{a}_n$ for some $s_1, \dots, s_n \in F$, and $\mathbf{s} = (s_1, \dots, s_n)$ is a solution by the first paragraph. Hence the system is consistent.

■

Read backwards, the theorem is a recipe for Chapter 1 questions: to decide whether $\mathbf{b}$ lies in the span of some vectors of F^m, put the vectors as the columns of a matrix and ask whether the system is consistent.

Example. Span membership as a system

In $\mathbb{R}^3$, let $\mathbf{v}_1 = (1,2,3)$ and $\mathbf{v}_2 = (0,1,2)$. Decide whether $(1,4,7)$ and $(1,4,8)$ lie in $\operatorname{span}(\mathbf{v}_1, \mathbf{v}_2)$.

Solution. Let $\mathbf{A}$ be the 3×2 matrix with columns $\mathbf{v}_1, \mathbf{v}_2$. By Theorem 3.1.1, $(1,4,7) \in \operatorname{span}(\mathbf{v}_1, \mathbf{v}_2)$ if and only if the system

$$a = 1, \qquad 2a + b = 4, \qquad 3a + 2b = 7$$

in the unknowns a, b is consistent. The first equation gives $a = 1$, and then the second gives $b = 2$. The third equation must also hold, and it does: $3 + 4 = 7$. Hence $(1,4,7) = \mathbf{v}_1 + 2\mathbf{v}_2$ lies in the span.

For $(1,4,8)$, the first two equations are the same, so any solution has $a = 1$ and $b = 2$. The third equation now demands $3 + 4 = 8$, which is false. The system is inconsistent, so $(1,4,8) \notin \operatorname{span}(\mathbf{v}_1, \mathbf{v}_2)$.

Geometrically, $\operatorname{span}(\mathbf{v}_1, \mathbf{v}_2)$ is a plane through $\mathbf{0}$ in $\mathbb{R}^3$. The first right-hand side lies on that plane and the second does not.

3.1.4 The shape of the solution set

We now describe the solution set of an arbitrary system, starting with the homogeneous case. Chapter 1 already did the work there: in Example (Solutions of a Homogeneous System) we checked, with the three conditions of the subspace test, that the solutions of $\mathbf{Ax} = \mathbf{0}$ form a subspace.

Theorem 3.1.2. Homogeneous solutions form a subspace

Let $\mathbf{A} \in M_{m\times n}(F)$. The solution set $N = \{\mathbf{x} \in F^n : \mathbf{Ax} = \mathbf{0}\}$ of the homogeneous system $\mathbf{Ax} = \mathbf{0}$ is a subspace of F^n.

Proof. This is Example (Solutions of a Homogeneous System): $\mathbf{A0} = \mathbf{0}$, and for $\mathbf{x}, \mathbf{y} \in N$ and $c \in F$, the rules of Theorem 0.9.4 give $\mathbf{A}(\mathbf{x} + \mathbf{y}) = \mathbf{Ax} + \mathbf{Ay} = \mathbf{0}$ and $\mathbf{A}(c\mathbf{x}) = c(\mathbf{Ax}) = \mathbf{0}$. By the subspace test (Theorem 1.2.1), N is a subspace of F^n. ∎

The solutions of $\mathbf{Ax} = \mathbf{b}$ with $\mathbf{b} \neq \mathbf{0}$ behave differently.

Warning. **The solution set of $\mathbf{Ax} = \mathbf{b}$ with $\mathbf{b} \neq \mathbf{0}$ is never a subspace.** A subspace must contain $\mathbf{0}$, but $\mathbf{A0} = \mathbf{0} \neq \mathbf{b}$, so $\mathbf{0}$ is not a solution. For example, the solutions of $x - y = 1$ in $\mathbb{R}^2$ form a line that misses the origin. The sum of the solutions $(1,0)$ and $(2,1)$ is $(3,1)$, and $3 - 1 = 2 \neq 1$: adding two solutions does not give a solution.

What does survive is this: the **difference** of two solutions of $\mathbf{Ax} = \mathbf{b}$ solves the homogeneous

system, since $\mathbf{A}(\mathbf{s}-\mathbf{s}') = \mathbf{b}-\mathbf{b} = \mathbf{0}$. So once one solution is known, every other one differs from it by a homogeneous solution. That is the whole structure.

Theorem 3.1.3. Structure of the solution set

Let $\mathbf{A} \in M_{m\times n}(F)$ and $\mathbf{b} \in F^m$, let N be the solution set of $\mathbf{Ax} = \mathbf{0}$, and suppose $\mathbf{p} \in F^n$ is **one** solution of $\mathbf{Ax} = \mathbf{b}$. Then the solution set of $\mathbf{Ax} = \mathbf{b}$ is

$$\mathbf{p} + N = \{\mathbf{p} + \mathbf{h} : \mathbf{h} \in N\}.$$

Idea. The theorem is "particular plus homogeneous". To prove equality of two sets, prove both inclusions. Going from a solution $\mathbf{s}$ to $\mathbf{p} + N$ uses the difference $\mathbf{s} - \mathbf{p}$, as observed above; going back uses the sum $\mathbf{A}(\mathbf{p} + \mathbf{h}) = \mathbf{Ap} + \mathbf{Ah}$.

Proof. Let S be the solution set of $\mathbf{Ax} = \mathbf{b}$. We use Theorem 0.3.1.

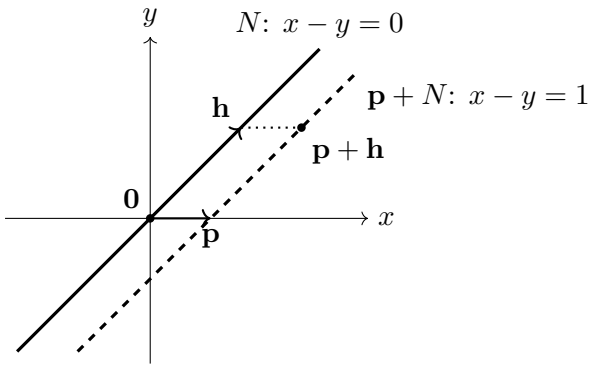
($\subseteq$) Let $\mathbf{s} \in S$, and put $\mathbf{h} = \mathbf{s} - \mathbf{p}$. By distributivity (Theorem 0.9.4), $\mathbf{Ah} = \mathbf{As} - \mathbf{Ap} = \mathbf{b} - \mathbf{b} = \mathbf{0}$, so $\mathbf{h} \in N$. Hence $\mathbf{s} = \mathbf{p} + \mathbf{h} \in \mathbf{p} + N$.

($\supseteq$) Let $\mathbf{h} \in N$. Then $\mathbf{A}(\mathbf{p} + \mathbf{h}) = \mathbf{Ap} + \mathbf{Ah} = \mathbf{b} + \mathbf{0} = \mathbf{b}$, again by distributivity, so $\mathbf{p} + \mathbf{h} \in S$.

Therefore $S = \mathbf{p} + N$, as claimed.

■

The theorem splits solving $\mathbf{Ax} = \mathbf{b}$ into two independent jobs: find **one** solution, and find **all** solutions of the homogeneous system. The set $\mathbf{p} + N$ is a translate of the subspace N by the vector $\mathbf{p}$, the kind of set that the picture in Example (A Line Missing the Origin) showed. For $x - y = 1$ in $\mathbb{R}^2$, the homogeneous solutions form the line $N = \operatorname{span}((1,1))$ through $\mathbf{0}$, one solution is $\mathbf{p} = (1,0)$, and the solution set is the parallel line through $\mathbf{p}$.



A system solved by substitution fits the same pattern once we collect the free parameter.

Example. Particular plus homogeneous

Solve $x_1 + 2x_2 - x_3 = 4$, $x_2 + x_3 = 1$ over $\mathbb{R}$, and write the solution set in the form $\mathbf{p} + \operatorname{span}(\mathbf{h})$.

Solution. The second equation gives $x_2 = 1 - x_3$. Substituting into the first, $x_1 = 4 - 2(1 - x_3) + x_3 = 2 + 3x_3$. The value of x_3 is not constrained, so write $x_3 = t$. Conversely, for every $t \in \mathbb{R}$ the vector $(2 + 3t, 1 - t, t)$ satisfies both equations, since $(2 + 3t) + 2(1 - t) - t = 4$ and $(1 - t) + t = 1$. Hence the solution set is

$$\{(2 + 3t, 1 - t, t) : t \in \mathbb{R}\} = \{(2, 1, 0) + t(3, -1, 1) : t \in \mathbb{R}\} = \mathbf{p} + \operatorname{span}(\mathbf{h}),$$

with $\mathbf{p} = (2, 1, 0)$ and $\mathbf{h} = (3, -1, 1)$. As Theorem 3.1.3 predicts, $\mathbf{p}$ is a solution (take $t = 0$), and $\mathbf{h}$ solves the homogeneous system: $3 - 2 - 1 = 0$ and $-1 + 1 = 0$. In the row picture, two planes in $\mathbb{R}^3$ meet in a line, and this line does not pass through $\mathbf{0}$.

3.1.5 When is the solution unique?

A consistent system has a unique solution exactly when there is nothing to add to the particular solution. The theorem just proved turns this into a statement about the homogeneous system, and the column picture turns that into independence.

Corollary 3.1.4. Uniqueness of solutions

Let $\mathbf{A} \in M_{m\times n}(F)$ have columns $\mathbf{a}_1, \dots, \mathbf{a}_n$, and let $\mathbf{b} \in F^m$ be such that $\mathbf{Ax} = \mathbf{b}$ is **consistent**. The following are equivalent:

(a) $\mathbf{Ax} = \mathbf{b}$ has exactly one solution;

(b) $\mathbf{Ax} = \mathbf{0}$ has only the trivial solution;

(c) the list $(\mathbf{a}_1, \dots, \mathbf{a}_n)$ is linearly independent.

Proof. Let N be the solution set of $\mathbf{Ax} = \mathbf{0}$. Since the system is consistent, it has a solution $\mathbf{p}$, and by Theorem 3.1.3 its solution set is $\mathbf{p} + N$.

(a $\Leftrightarrow$ b) The function $N \to \mathbf{p} + N$, $\mathbf{h} \mapsto \mathbf{p} + \mathbf{h}$, is surjective by definition of $\mathbf{p} + N$, and injective because $\mathbf{p} + \mathbf{h} = \mathbf{p} + \mathbf{h}'$ implies $\mathbf{h} = \mathbf{h}'$ by adding $-\mathbf{p}$ to both sides. So it is a bijection, and $\mathbf{p} + N$ has exactly one element if and only if N does. Since $\mathbf{0} \in N$ always, N has exactly one element if and only if $N = \{\mathbf{0}\}$, which is (b).

(b $\Leftrightarrow$ c) By Theorem 0.9.3, $\mathbf{Ax} = x_1\mathbf{a}_1 + \cdots + x_n\mathbf{a}_n$ for every $\mathbf{x} \in F^n$. So (b) says: for all $x_1, \dots, x_n \in F$, $x_1\mathbf{a}_1 + \cdots + x_n\mathbf{a}_n = \mathbf{0}$ implies $x_1 = \cdots = x_n = 0$. This is word for word Definition 1.4.1 for the list $(\mathbf{a}_1, \dots, \mathbf{a}_n)$.

■

Two remarks on the hypotheses. The equivalence of (b) and (c) does not need consistency; only (a) does. And consistency cannot be dropped from (a $\Leftrightarrow$ b): the system $x = 0$, $x = 1$ in one unknown has only the trivial solution in its homogeneous version ($x = 0$, $x = 0$), but it has no solution at all, so it certainly does not have exactly one.

So the three pictures answer the two basic questions about $\mathbf{Ax} = \mathbf{b}$ in the language of Chapter 1: a solution **exists** when $\mathbf{b}$ is in the span of the columns, and it is **unique** when the columns are independent. What we still lack is a reliable way to decide either question for a given matrix. That is the job of Gaussian elimination, in the next section.

3.1.6 Exercises

3.1.6.1 A. Check your understanding

Exercise. A1

(a) Define what it means for a linear system $\mathbf{Ax} = \mathbf{b}$ over F to be consistent, and what its augmented matrix is. What is the size of the augmented matrix of a system of m equations in n unknowns?

(b) Determine whether the following statement is true: "the solution set of $\mathbf{Ax} = \mathbf{b}$ is a subspace of F^n." Justify your answer.

(c) Determine whether the following statement is true: "if $\mathbf{Ax} = \mathbf{b}$ has two different solutions, then $\mathbf{Ax} = \mathbf{0}$ has a nontrivial solution." Justify your answer.

(d) Let $\mathbf{A} \in M_{3\times 5}(\mathbb{R})$. In which spaces do the unknown vector $\mathbf{x}$ and the right-hand side $\mathbf{b}$ of $\mathbf{Ax} = \mathbf{b}$ live?

(e) Name the theorem that turns "is $\mathbf{b}$ in the span of $\mathbf{a}_1, \ldots, \mathbf{a}_n$?" into a question about a linear system, and state which system.

Solution.

(a) It is consistent if it has at least one solution $\mathbf{s} \in F^n$, that is, some $\mathbf{s}$ with $\mathbf{As} = \mathbf{b}$. The augmented matrix is $[\,\mathbf{A} \mid \mathbf{b}\,]$, the matrix $\mathbf{A}$ with $\mathbf{b}$ appended as an extra last column. It has size $m \times (n+1)$.

(b) False in general. If $\mathbf{b} \neq \mathbf{0}$, then $\mathbf{A0} = \mathbf{0} \neq \mathbf{b}$, so $\mathbf{0}$ is not a solution and the solution set is not a subspace. (If $\mathbf{b} = \mathbf{0}$, it is a subspace, by Theorem 3.1.2.)

(c) True. If $\mathbf{s} \neq \mathbf{s}'$ are solutions, then $\mathbf{A}(\mathbf{s}-\mathbf{s}') = \mathbf{b}-\mathbf{b} = \mathbf{0}$ by distributivity, and $\mathbf{s}-\mathbf{s}' \neq \mathbf{0}$.

(d) $\mathbf{x} \in \mathbb{R}^5$, one entry per column of $\mathbf{A}$, and $\mathbf{b} \in \mathbb{R}^3$, one entry per row.

(e) Theorem 3.1.1: $\mathbf{b} \in \operatorname{span}(\mathbf{a}_1, \ldots, \mathbf{a}_n)$ if and only if $\mathbf{Ax} = \mathbf{b}$ is consistent, where $\mathbf{A}$ is the matrix with columns $\mathbf{a}_1, \ldots, \mathbf{a}_n$.

3.1.6.2 B. Practice

Exercise. B1: Three views

(a) Write the system $x_1 - x_2 + 2x_3 = 3$, $2x_1 + x_3 = 1$ over $\mathbb{R}$ in the matrix picture and in the column picture, and give its augmented matrix.

(b) Write out, as a list of equations, the system with augmented matrix $\left[\begin{array}{cc|c} 1 & 4 & 0 \\ 0 & 1 & 2 \\ 3 & -1 & 5 \end{array}\right]$ over $\mathbb{R}$. How many equations and unknowns does it have? Hence decide whether it is consistent.

Solution.

(a) The matrix picture is

$$\begin{pmatrix} 1 & -1 & 2 \\ 2 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 3 \\ 1 \end{pmatrix}.$$

The column picture is $x_1(1,2) + x_2(-1,0) + x_3(2,1) = (3,1)$. The augmented matrix is $\left[\begin{array}{ccc|c} 1 & -1 & 2 & 3 \\ 2 & 0 & 1 & 1 \end{array}\right]$.

(b) The system is $x_1 + 4x_2 = 0$, $x_2 = 2$, $3x_1 - x_2 = 5$: three equations in two unknowns. The second equation gives $x_2 = 2$, and then the first gives $x_1 = -8$. The third equation would require $-24 - 2 = 5$, which is false. Hence no vector satisfies all three equations, and the system is inconsistent.

Exercise. B2: In the span or not?

Determine whether the vector $\mathbf{b}$ lies in the span of the given vectors. Justify your answer, and when it does, write $\mathbf{b}$ as a linear combination of them.

(a) $\mathbf{b} = (2,1,4)$ and $(1,0,1), (0,1,2)$ in $\mathbb{R}^3$.

(b) $\mathbf{b} = (1,1,1)$ and $(1,2,0), (2,4,1)$ in $\mathbb{R}^3$.

(c) $\mathbf{b} = (1,1,0)$ and $(1,0,1), (0,1,1)$ in $\mathbb{F}_2^3$. Is the answer the same in $\mathbb{R}^3$?

Solution. By Theorem 3.1.1, in each case we decide whether the system $a\mathbf{v}_1 + b\mathbf{v}_2 = \mathbf{b}$ in the unknowns a, b is consistent.

(a) The system is $a = 2$, $b = 1$, $a + 2b = 4$. The first two equations force $a = 2$, $b = 1$, and then $a + 2b = 4$ holds. Hence $(2,1,4) = 2(1,0,1) + (0,1,2)$ is in the span.

(b) The system is $a + 2b = 1$, $2a + 4b = 1$, $b = 1$. Doubling the first equation gives $2a + 4b = 2$, contradicting the second, since $2 \neq 1$. Hence the system is inconsistent and $(1,1,1)$ is not in the span.

(c) The system is $a = 1$, $b = 1$, $a + b = 0$. The first two force $a = b = 1$, and then $a + b = 1 + 1 = 0$ in $\mathbb{F}_2$. Hence $(1,1,0) = (1,0,1) + (0,1,1)$ is in the span over $\mathbb{F}_2$. In $\mathbb{R}^3$ the same first two equations force $a = b = 1$, but $1 + 1 = 2 \neq 0$, so the system is inconsistent and $\mathbf{b}$ is **not** in the span. The answer depends on the field.

Exercise. B3: Solve and describe

Solve each system over $\mathbb{R}$ by substitution, and write its solution set in the form $\mathbf{p} + \text{span}(\dots)$. Check that your $\mathbf{p}$ is a solution and that your spanning vectors solve the associated homogeneous system.

(a) $x_1 + x_2 + x_3 = 2$, $x_2 - x_3 = 1$.

(b) $x_1 - 2x_2 + x_3 - x_4 = 1$, $x_3 + 2x_4 = 3$.

Solution.

(a) The second equation gives $x_2 = 1 + x_3$, and then the first gives $x_1 = 2 - (1 + x_3) - x_3 =$

$1-2x_3$. With $x_3 = t$ free, every solution is $(1-2t, 1+t, t)$, and conversely each such vector satisfies both equations. So the solution set is $(1,1,0)+\text{span}((-2,1,1))$. Check: $1+1+0=2$ and $1-0=1$; and $-2+1+1=0$, $1-1=0$.

(b) The second equation gives $x_3 = 3-2x_4$. The first then gives $x_1 = 1+2x_2-(3-2x_4)+x_4 = -2+2x_2+3x_4$. The unknowns $x_2 = s$ and $x_4 = t$ are free, and every solution is

$$(-2+2s+3t,\ s,\ 3-2t,\ t) = (-2,0,3,0) + s(2,1,0,0) + t(3,0,-2,1),$$

and conversely each such vector is a solution. So the solution set is $(-2,0,3,0) + \text{span}((2,1,0,0),(3,0,-2,1))$. Check: $-2-0+3-0=1$ and $3+0=3$; for $(2,1,0,0)$: $2-2+0-0=0$ and $0+0=0$; for $(3,0,-2,1)$: $3-0-2-1=0$ and $-2+2=0$.

3.1.6.3 C. Going deeper

Exercise. C1: How many solutions?

Let $\mathbf{A} \in M_{m\times n}(F)$ and $\mathbf{b} \in F^m$, and let N be the solution set of $\mathbf{Ax} = \mathbf{0}$.

(a) Suppose F is **infinite**. Prove that $\mathbf{Ax} = \mathbf{b}$ has either no solution, exactly one solution, or infinitely many solutions.

(b) Suppose F is **finite**. Prove that $\mathbf{Ax} = \mathbf{b}$ has either no solution or exactly $|N|$ solutions.

(c) Give a system over $\mathbb{F}_3$ with exactly three solutions.

Hint: for (a), if N contains a nonzero vector $\mathbf{h}$, look at its multiples.

Solution.

(a) Suppose the system is consistent, with a solution $\mathbf{p}$. By Theorem 3.1.3 its solution set is $\mathbf{p}+N$, and as in the proof of Corollary 3.1.4 the map $\mathbf{h} \mapsto \mathbf{p}+\mathbf{h}$ is a bijection $N \to \mathbf{p}+N$. If $N = \{\mathbf{0}\}$, there is exactly one solution. Otherwise, let $\mathbf{h} \in N$ with $\mathbf{h} \neq \mathbf{0}$. Since N is a subspace (Theorem 3.1.2), $c\mathbf{h} \in N$ for every $c \in F$. If $c\mathbf{h} = c'\mathbf{h}$, then $(c-c')\mathbf{h} = \mathbf{0}$, and since $\mathbf{h} \neq \mathbf{0}$, Theorem 1.1.9 gives $c = c'$. So $c \mapsto c\mathbf{h}$ is injective, and N contains infinitely many vectors because F is infinite. Hence $\mathbf{p}+N$ is infinite.

(b) If the system is consistent with solution $\mathbf{p}$, the bijection $N \to \mathbf{p}+N$ from (a) shows that the solution set has exactly $|N|$ elements. Otherwise it has none.

(c) The single equation $x+y=1$ over $\mathbb{F}_3$ (a system with $m=1$, $n=2$). For each of the three values of $y \in \mathbb{F}_3$ there is exactly one $x = 1-y$, so the solutions are $(1,0),(0,1),(2,2)$. Consistently with (b), the homogeneous equation $x+y=0$ has the three solutions $(0,0),(2,1),(1,2)$.

Exercise. C2: The right-hand sides that work

Let $\mathbf{A} \in M_{m\times n}(F)$ have columns $\mathbf{a}_1, \dots, \mathbf{a}_n$.

(a) Prove that the set $C = \{\mathbf{b} \in F^m : \mathbf{Ax} = \mathbf{b} \text{ is consistent}\}$ is a subspace of F^m.

(b) Deduce that if $n < m$, then there is some $\mathbf{b} \in F^m$ for which $\mathbf{Ax} = \mathbf{b}$ is inconsistent.

(c) For $\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 2 & 4 \\ 0 & 1 \end{pmatrix}$ over $\mathbb{R}$, find one such $\mathbf{b}$.

Solution.

(a) By Theorem 3.1.1, $\mathbf{b} \in C$ if and only if $\mathbf{b} \in \operatorname{span}(\mathbf{a}_1, \ldots, \mathbf{a}_n)$. So $C = \operatorname{span}(\mathbf{a}_1, \ldots, \mathbf{a}_n)$, which is a subspace of F^m by Theorem 1.3.1.

(b) By (a), C is spanned by the n vectors $\mathbf{a}_1, \ldots, \mathbf{a}_n$. Since $\dim F^m = m$ (Example (Dimensions of the Running Examples)), Theorem 1.6.3 says no list of fewer than m vectors spans F^m. As $n < m$, $C \neq F^m$, so some $\mathbf{b} \in F^m$ is not in C, and for that $\mathbf{b}$ the system is inconsistent.

(c) Take $\mathbf{b} = (1, 0, 0)$. A solution would satisfy $x + 2y = 1$ and $2x + 4y = 0$. Doubling the first gives $2x + 4y = 2 \neq 0$, a contradiction. So $\mathbf{Ax} = (1, 0, 0)$ is inconsistent.

3.2 Gaussian Elimination

Every reader of this book has solved a system of equations by eliminating unknowns: subtract a multiple of one equation from another, and a variable disappears. Done by hand on small systems, this feels like common sense. Done on a system of 50 equations, or over $\mathbb{F}_2$, or in a proof about all systems at once, it needs to be an algorithm, and we need to know that it never changes the answer. This section makes elimination precise, proves that it always reaches a standard shape, and shows how to read every solution from that shape.

3.2.1 Three moves that do not change the answer

Here is a system over $\mathbb{R}$, solved the way one would solve it on paper:

$$\begin{aligned} x + 2y + z &= 4, \\ 2x + 4y + 3z &= 9, \\ x + 3y + z &= 5. \end{aligned}$$

Subtract twice the first equation from the second: $z = 1$. Subtract the first from the third: $y = 1$. Then the first equation gives $x = 4 - 2 - 1 = 1$.

Only three kinds of move were used, or could reasonably be used: **reorder** the equations, **multiply** an equation by a nonzero number, and **add a multiple** of one equation to another. None of them looks at the names x, y, z. Each acts on the coefficients and the right-hand side of whole equations, that is, on the rows of the augmented matrix. The first step above, for instance, replaces row 2, $(2, 4, 3 \mid 9)$, by row 2 minus twice row 1, $(0, 0, 1 \mid 1)$.

Not every move that looks harmless is safe. Multiplying an equation by 0 turns it into $0 = 0$, and a constraint silently disappears. So we list the allowed moves carefully.

An elementary row operation is a reversible move on the rows of a matrix: swap two rows, rescale one row by a nonzero scalar, or add a multiple of one row to a different row.

Definition 3.2.1. Elementary row operations

Let F be a field and let $\mathbf{M} \in M_{m\times k}(F)$ have rows $\mathbf{r}_1, \dots, \mathbf{r}_m$. An **elementary row operation** on $\mathbf{M}$ is one of the following.

(1) **Swap** two rows: for $i \neq j$, exchange $\mathbf{r}_i$ and $\mathbf{r}_j$. Notation: $R_i \leftrightarrow R_j$.

(2) **Scale** a row: for an index i and a scalar $c \in F$ with $c \neq 0$, replace $\mathbf{r}_i$ by $c\mathbf{r}_i$. Notation: $R_i \to cR_i$.

(3) **Replace** a row by adding a multiple of **another** row: for indices $i \neq j$ and any $c \in F$, replace $\mathbf{r}_i$ by $\mathbf{r}_i + c\mathbf{r}_j$. Notation: $R_i \to R_i + cR_j$.

In each case the rows not mentioned are left unchanged.

In words: (1) reorders, (2) rescales one row by a **nonzero** scalar, and (3) changes **one** row, row i, using a **different** row j, which itself stays as it was. We write $R_i \to R_i - cR_j$ for $R_i \to R_i + (-c)R_j$. The notation R_i refers to the current row i at the moment the operation is performed.

Take $\mathbf{M} = \begin{pmatrix} 1 & 2 & 0 \\ 3 & 1 & 1 \end{pmatrix}$ over $\mathbb{R}$. Then $R_1 \leftrightarrow R_2$ gives $\begin{pmatrix} 3 & 1 & 1 \\ 1 & 2 & 0 \end{pmatrix}$, $R_1 \to -2R_1$ gives $\begin{pmatrix} -2 & -4 & 0 \\ 3 & 1 & 1 \end{pmatrix}$, and $R_2 \to R_2 - 3R_1$ gives $\begin{pmatrix} 1 & 2 & 0 \\ 0 & -5 & 1 \end{pmatrix}$. The degenerate case $m = 1$ is worth a glance: a one-row matrix admits no swap and no replacement, since both need two different rows, so its only elementary row operations are the scalings.

The minimal change that breaks the definition is to drop "$c \neq 0$". The move $R_2 \to 0R_2$ still replaces one row by a multiple of itself, but it sends $\begin{pmatrix} 1 & 2 & 0 \\ 3 & 1 & 1 \end{pmatrix}$ to $\begin{pmatrix} 1 & 2 & 0 \\ 0 & 0 & 0 \end{pmatrix}$, and no move of any kind can recover the lost row $(3, 1, 1)$ from that matrix. The failing clause is $c \neq 0$, and the damage is exactly irreversibility.

Warning. **Perform operations one at a time.** It is tempting to write "$R_1 \to R_1 - R_2$ and $R_2 \to R_2 - R_1$" as one step, using the old rows in both. That is not a sequence of elementary row operations, and it can destroy information. On the system $x = 1$, $x = 1$ it produces $0 = 0$, $0 = 0$, whose solution set is all of F, not $\{1\}$. Done in order, the second operation uses the **new** row 1, $(0 \mid 0)$, and leaves row 2 unchanged, so no information is lost.

Why these three and no others? They are exactly the moves of hand elimination, and each can be undone by a move of the same kind, which is what the theorem below needs. The condition $i \neq j$ in (3) is not fussiness: "add c times row i to row i" would be scaling by $1 + c$, which is illegal when $c = -1$ (Exercise C1).

An elementary row operation acts on the rows, but it treats each column separately: the new entry in position (i, l) is computed only from entries of column l. We record this, because it will be used again and again.

Lemma 3.2.1. Row operations act column by column

Let $\mathbf{M} = [\,\mathbf{M}_1 \mid \mathbf{M}_2\,] \in M_{m\times k}(F)$ be split into its first k_1 columns $\mathbf{M}_1$ and its remaining columns $\mathbf{M}_2$. If an elementary row operation turns $\mathbf{M}$ into $\mathbf{M}' = [\,\mathbf{M}_1' \mid \mathbf{M}_2'\,]$ (split in the

same place), then the same operation turns $\mathbf{M}_1$ into $\mathbf{M}_1'$ and $\mathbf{M}_2$ into $\mathbf{M}_2'$. In particular, an elementary row operation turns a zero column into a zero column.

Proof. Let m_{hl} be the entries of $\mathbf{M}$ and m'_{hl} those of $\mathbf{M}'$. For a swap $R_i \leftrightarrow R_j$: $m'_{il} = m_{jl}$, $m'_{jl} = m_{il}$, and $m'_{hl} = m_{hl}$ for $h \neq i, j$. For a scaling $R_i \to cR_i$: $m'_{il} = c\,m_{il}$, and other rows are unchanged. For a replacement $R_i \to R_i + cR_j$: $m'_{il} = m_{il} + c\,m_{jl}$, and other rows are unchanged. In each case, every entry of column l of $\mathbf{M}'$ is computed from column l of $\mathbf{M}$ alone, by the same rule for every l. Hence applying the operation to $\mathbf{M}_1$ or to $\mathbf{M}_2$ separately gives $\mathbf{M}_1'$ and $\mathbf{M}_2'$. If column l of $\mathbf{M}$ is zero, each formula gives 0, since $c \cdot 0 = 0$ and $0 + c \cdot 0 = 0$ (Theorem 0.7.1).

■

For an augmented matrix, the lemma says: operating on $[\,\mathbf{A} \mid \mathbf{b}\,]$ is the same as operating on $\mathbf{A}$ and on $\mathbf{b}$ with the same moves. Now we name what sequences of moves produce.

Definition 3.2.2. Row equivalence

Let $\mathbf{A}, \mathbf{B} \in M_{m\times k}(F)$. We say $\mathbf{A}$ is **row equivalent** to $\mathbf{B}$ if $\mathbf{B}$ can be obtained from $\mathbf{A}$ by a **finite** sequence of elementary row operations. The empty sequence is allowed, so every matrix is row equivalent to itself.

In Example (Row equivalence) of Chapter 0 we met a definition that looks different: $\mathbf{B} = \mathbf{EA}$ for an **invertible** matrix $\mathbf{E}$. In section 4 of this chapter we prove that the two definitions agree (Theorem 3.4.8), by writing each elementary row operation as multiplication by an invertible matrix. Until then, "row equivalent" means the definition above.

3.2.2 Row operations preserve solutions

Here is the promise from the start of the section, now as a theorem.

Theorem 3.2.2. Row operations preserve solutions

Let F be a field.

(a) **(Reversibility)** If an elementary row operation turns $\mathbf{M}$ into $\mathbf{M}'$, then an elementary row operation **of the same type** turns $\mathbf{M}'$ back into $\mathbf{M}$.

(b) Row equivalence is an equivalence relation on $M_{m\times k}(F)$.

(c) If $[\,\mathbf{A} \mid \mathbf{b}\,]$ and $[\,\mathbf{A}' \mid \mathbf{b}'\,]$ in $M_{m\times(n+1)}(F)$ are row equivalent, then the systems $\mathbf{Ax} = \mathbf{b}$ and $\mathbf{A}'\mathbf{x} = \mathbf{b}'$ have the **same** solution set.

Idea. Reversibility is checked type by type, and the check for scaling is the only place where $c \neq 0$ is used. For (c), every equation of the new system is a combination of equations of the old one, so an old solution satisfies the new system. That gives one inclusion. Instead of repeating the argument backwards, reversibility gives the other inclusion for free.

Proof.

(a) A swap $R_i \leftrightarrow R_j$ is undone by the same swap. A scaling $R_i \to cR_i$ is undone by $R_i \to c^{-1}R_i$, which is a legal scaling because $c \neq 0$ has an inverse $c^{-1} \neq 0$ in F (Definition 0.7.1): the row becomes $c^{-1}(c\mathbf{r}_i) = \mathbf{r}_i$. A replacement $R_i \to R_i + cR_j$ is undone by $R_i \to R_i - cR_j$. Here we use $i \neq j$: row j was not changed by the first operation, so the second one gives $(\mathbf{r}_i + c\mathbf{r}_j) - c\mathbf{r}_j = \mathbf{r}_i$.

(b) Reflexivity holds via the empty sequence. For symmetry, suppose operations $O_1, \ldots, O_s$, in this order, turn $\mathbf{A}$ into $\mathbf{B}$. By (a), each O_t has an undoing operation O'_t. Then $O'_s, O'_{s-1}, \ldots, O'_1$, in this order, turn $\mathbf{B}$ back into $\mathbf{A}$. For transitivity, if a sequence turns $\mathbf{A}$ into $\mathbf{B}$ and another turns $\mathbf{B}$ into $\mathbf{C}$, performing the first and then the second turns $\mathbf{A}$ into $\mathbf{C}$, and this is still a finite sequence.

(c) First suppose a **single** operation turns $[\,\mathbf{A} \mid \mathbf{b}\,]$ into $[\,\mathbf{A}' \mid \mathbf{b}'\,]$. Let S and S' be the two solution sets. Write $\mathbf{r}_h$ for row h of $\mathbf{A}$, so that equation h says $\mathbf{r}_h\mathbf{x} = b_h$, where $\mathbf{r}_h\mathbf{x}$ is a 1×1 product. Let $\mathbf{s} \in S$, so $\mathbf{r}_h\mathbf{s} = b_h$ for every h. Each equation of the new system holds at $\mathbf{s}$:

- after $R_i \leftrightarrow R_j$, the new equations are the old ones in another order;
- after $R_i \to cR_i$, the new equation i is $(c\mathbf{r}_i)\mathbf{x} = cb_i$, and $(c\mathbf{r}_i)\mathbf{s} = c(\mathbf{r}_i\mathbf{s}) = cb_i$;
- after $R_i \to R_i + cR_j$, the new equation i is $(\mathbf{r}_i + c\mathbf{r}_j)\mathbf{x} = b_i + cb_j$, and $(\mathbf{r}_i + c\mathbf{r}_j)\mathbf{s} = \mathbf{r}_i\mathbf{s} + c(\mathbf{r}_j\mathbf{s}) = b_i + cb_j$, by Theorem 0.9.4.

The other equations are unchanged. Hence $\mathbf{s} \in S'$, and $S \subseteq S'$. By (a), a single operation turns $[\,\mathbf{A}' \mid \mathbf{b}'\,]$ back into $[\,\mathbf{A} \mid \mathbf{b}\,]$, so the same argument gives $S' \subseteq S$. Therefore $S = S'$.

For a sequence of s operations, the intermediate augmented matrices all have the same solution set, by the single-operation case applied s times (formally, by induction on s). This proves (c).

■

This theorem is the license for everything that follows: we may transform a system as much as we like by row operations, and the solutions of the final system are the solutions of the original one. What it does **not** license is just as important.

Warning. **Scaling by 0 is not a row operation, and it changes the solution set.** Over $\mathbb{R}$, the system $x + y = 2$, $x - y = 0$ has exactly one solution, $(1, 1)$. Applying the illegal move $R_2 \to 0R_2$ gives $x + y = 2$, $0 = 0$, whose solutions $(2 - t, t)$ form a whole line. The inclusion $S \subseteq S'$ in the proof above still holds; what fails is reversibility, and with it $S' \subseteq S$.

Warning. **Column operations are not allowed.** Operations on the columns of $[\,\mathbf{A} \mid \mathbf{b}\,]$ mix up the unknowns. The system $x+y = 3$, $x-y = 1$ has the unique solution $(2, 1)$. Adding column 1 to column 2 of its augmented matrix gives $\left[\begin{array}{cc|c} 1 & 2 & 3 \\ 1 & 0 & 1 \end{array}\right]$, the system $x+2y = 3$, $x = 1$, whose only solution is $(1, 1)$. Even swapping two columns of $\mathbf{A}$ swaps the corresponding entries of every solution.

3.2.3 Echelon forms

Row operations can change a system a great deal. We now choose a target: a shape of matrix from which the solutions can be read off with no further work. Look at the finished computation from the start of the section. After eliminating, the augmented matrix could be brought to

$$\left[\begin{array}{ccc|c} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{array}\right],$$

which says $x = 1$, $y = 1$, $z = 1$ outright. Not every system can be brought to this shape, since not every system has exactly one solution. The general target is a staircase.

For a nonzero row of a matrix, its **leading entry** is its first nonzero entry, reading from the left.

A matrix is in row echelon form when its nonzero rows come first and their leading entries step strictly to the right as we go down.

Definition 3.2.3. Row echelon form

A matrix $\mathbf{R} \in M_{m\times k}(F)$ is in **row echelon form** if:

(E1) there is an integer r with $0 \le r \le m$ such that rows $1, \dots, r$ are nonzero and rows $r+1, \dots, m$ are zero; and

(E2) if the leading entry of row i lies in column j_i, for $i = 1, \dots, r$, then $j_1 < j_2 < \cdots < j_r$, **strictly**.

The leading entries of rows $1, \dots, r$ are the **pivots** of $\mathbf{R}$, and the columns $j_1, \dots, j_r$ are its **pivot columns**.

In words: (E1) says zero rows sit at the bottom, and (E2) says each pivot lies strictly to the right of the pivot above it. Two consequences are used constantly. Each row contains at most one pivot, so there are at most m pivots. And **every entry below a pivot is zero**: if $h > i$, then row h is zero or has its leading entry in column $j_h > j_i$, so its entry in column j_i is 0.

Definition 3.2.4. Reduced row echelon form

A matrix $\mathbf{R} \in M_{m\times k}(F)$ is in **reduced row echelon form** (**RREF**) if it is in row echelon form, with pivot columns $j_1 < \cdots < j_r$, and moreover:

(E3) every pivot equals 1: $r_{ij_i} = 1$ for $i = 1, \dots, r$; and

(E4) every pivot is the **only** nonzero entry in its column: $r_{hj_i} = 0$ for all $h \neq i$.

In words: (E3) normalizes the pivots, and (E4) clears the entries **above** each pivot as well as below. Together they say that the pivot column j_i is the standard basis vector $\mathbf{e}_i \in F^m$.

Example. Echelon or not?

Decide which of the following real matrices are in row echelon form, and which are in reduced

row echelon form. Give the pivot columns where they exist.

$$\mathbf{P} = \begin{pmatrix} 2 & 1 & 3 \\ 0 & 0 & -1 \\ 0 & 0 & 0 \end{pmatrix}, \quad \mathbf{Q} = \begin{pmatrix} 1 & 5 & 0 & 2 \\ 0 & 0 & 1 & 3 \end{pmatrix}, \quad \mathbf{T} = \begin{pmatrix} 1 & 0 & 2 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}, \quad \mathbf{Z} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}.$$

Solution. **P** is in row echelon form: rows 1 and 2 are nonzero, row 3 is zero, and the leading entries lie in columns $1 < 3$. The pivot columns are 1 and 3. It is **not** reduced: the pivot 2 is not 1, failing (E3), and column 3 has the nonzero entry 3 above the pivot -1, failing (E4).

Q is in reduced row echelon form, with pivot columns 1 and 3. Both pivots are 1, and columns 1 and 3 are $\mathbf{e}_1$ and $\mathbf{e}_2$. The entries $5, 2, 3$ in the nonpivot columns 2 and 4 are allowed to be anything.

T is **not** in row echelon form. Its leading entries lie in columns $1, 3, 2$, and $3 < 2$ fails, so (E2) fails. Swapping rows 2 and 3 would repair it.

Z is in reduced row echelon form, with $r = 0$: there are no pivots, and (E2)–(E4) hold vacuously. This degenerate case matters, because the zero matrix must have an RREF too, and it is itself.

For **Q**, a minimal change gives a nonexample: replacing the 0 in position $(1, 3)$ by 4 keeps (E1)–(E3), but column 3 now has the nonzero entry 4 above its pivot, so (E4) fails.

Why insist on the reduced form, when row echelon form already allows back substitution? Because a system in reduced form needs **no** substitution: each pivot variable appears in exactly one equation. More importantly, the reduced form turns out to be **unique**: a matrix is row equivalent to exactly one matrix in RREF, however the operations are chosen. That is proved in the next section, and it is what makes later definitions, such as the rank, independent of choices. Row echelon form is not unique, as the warning after the worked example shows.

Quick check. Is $\begin{pmatrix} 1 & 2 & 0 & -1 \\ 0 & 0 & 1 & 3 \\ 0 & 0 & 0 & 0 \end{pmatrix}$ in reduced row echelon form? If it is the coefficient matrix of a homogeneous system in x_1, x_2, x_3, x_4, which unknowns correspond to pivot columns?

Solution. Yes. Row 3 is the only zero row and it is last; the leading entries lie in columns $1 < 3$; both are 1; and columns 1 and 3 are $\mathbf{e}_1$ and $\mathbf{e}_2$. The pivot columns are 1 and 3, so the unknowns x_1 and x_3 correspond to pivots, and x_2, x_4 do not.

3.2.4 Gauss–Jordan elimination

We now show that every matrix can be brought to reduced row echelon form. The procedure works through the columns from left to right, keeping a count r of the pivots found so far.

Algorithm. **Gauss–Jordan elimination.** Input: $\mathbf{M} \in M_{m\times k}(F)$. Set $r = 0$. For $l = 1, 2, \ldots, k$ in turn:

1. Look for a row h with $h > r$ whose entry in column l is nonzero. If there is none, column l gets no pivot; go on to the next l.
2. Otherwise choose such an h. If $h \neq r+1$, perform $R_{r+1} \leftrightarrow R_h$.
3. Let $a \neq 0$ be the entry now in position $(r+1, l)$. Perform $R_{r+1} \to a^{-1}R_{r+1}$, so that this entry becomes 1.
4. For every row $g \neq r+1$ whose entry c in column l is nonzero, perform $R_g \to R_g - cR_{r+1}$.
5. Increase r by 1.

Output: the final matrix, which is in reduced row echelon form, with r pivots.

Step 2 involves a choice: any nonzero entry will do. By hand, we choose one that avoids fractions. If we skip the scaling in step 3 and clear only the entries **below** the pivot a, by $R_g \to R_g - ca^{-1}R_{r+1}$ for the rows $g > r+1$, the output is a row echelon form instead; this cheaper variant is **Gaussian elimination**, and it is followed by back substitution.

Theorem 3.2.3. Existence of the reduced row echelon form

Let F be a field. Every matrix $\mathbf{M} \in M_{m\times k}(F)$ is row equivalent to a matrix in reduced row echelon form.

Idea. We prove that the algorithm works, by induction on the number l of columns processed. The invariant is: the first l columns of the current matrix form a matrix in RREF. When a new column is processed, all the operations involve only rows below the pivots found so far, or add multiples of the new pivot row, which is zero in the first l columns. So the part already built is never disturbed. After all k columns, the whole matrix is in RREF.

Proof. For $0 \le l \le k$, let $P(l)$ be the statement: there is a matrix $\mathbf{M}_l$ row equivalent to $\mathbf{M}$ and an integer $r \ge 0$ such that the $m \times l$ matrix Ł formed by the first l columns of $\mathbf{M}_l$ is in reduced row echelon form with exactly r nonzero rows. We prove $P(l)$ for all l by induction (Theorem 0.2.7). For $l = 0$ take $\mathbf{M}_0 = \mathbf{M}$ and $r = 0$; there are no columns, so there is nothing to check.

Suppose $P(l)$ holds for some $l < k$, with $\mathbf{M}_l$, Ł, and r as described, and let $j_1 < \cdots < j_r \le l$ be the pivot columns of Ł. By (E1) for Ł, rows $r+1, \ldots, m$ of $\mathbf{M}_l$ are zero in columns $1, \ldots, l$. Let Ł′ denote the first $l+1$ columns of the matrix we are about to construct.

Case 1: every entry of column $l+1$ of $\mathbf{M}_l$ in rows $h > r$ is zero (this includes the case $r = m$). Put $\mathbf{M}_{l+1} = \mathbf{M}_l$. Rows $r+1, \ldots, m$ of Ł′ are zero, and rows $1, \ldots, r$ are nonzero with the same leading entries as in Ł, which lie in columns $\le l$. So (E1)–(E3) hold for Ł′ with the same r and pivot columns, and so does (E4), since the pivot columns of Ł′ are those of Ł and are unchanged. Hence Ł′ is in RREF and $P(l+1)$ holds.

Case 2: some row $h > r$ has a nonzero entry in column $l+1$. Then $r+1 \le h \le m$. Perform steps 2, 3, and 4 of the algorithm with this h and with column $l+1$; call the result $\mathbf{M}_{l+1}$. It is row equivalent to $\mathbf{M}_l$, hence to $\mathbf{M}$, by Theorem 3.2.2 (b).

These operations do not change columns $1, \dots, l$. Indeed, the swap in step 2 and the scaling in step 3 only move or rescale rows h and $r+1$, which are both zero in those columns. Each replacement $R_g \to R_g - cR_{r+1}$ in step 4 adds a multiple of row $r+1$, which is still zero in columns $1, \dots, l$. So the first l columns of $\mathbf{M}_{l+1}$ form the matrix Ł. In column $l+1$, step 3 makes the entry in row $r+1$ equal to 1, and each replacement in step 4 makes the entry of row g equal to $c - c \cdot 1 = 0$ without changing any other row. So column $l+1$ of $\mathbf{M}_{l+1}$ is $\mathbf{e}_{r+1}$.

We check that Ł′ is in RREF with $r+1$ nonzero rows. Rows $1, \dots, r$ of Ł′ have the same leading entries as in Ł, in columns $j_1 < \cdots < j_r \le l$. Row $r+1$ is zero in columns $1, \dots, l$ and has a 1 in column $l+1$, so its leading entry is in column $j_{r+1} := l+1 > j_r$. Rows $r+2, \dots, m$ are zero in columns $1, \dots, l$ and in column $l+1$, so they are zero. This gives (E1) and (E2). The pivots are 1, by (E3) for Ł and the construction, so (E3) holds. For (E4), the columns $j_1, \dots, j_r$ are unchanged from Ł, and column $l+1$ is $\mathbf{e}_{r+1}$. Hence $P(l+1)$ holds.

By induction $P(k)$ holds: $\mathbf{M}_k$ is row equivalent to $\mathbf{M}$, and its first k columns, which are all of $\mathbf{M}_k$, form a matrix in reduced row echelon form. This proves the theorem.

■

The proof is also the correctness proof of the algorithm: the matrix $\mathbf{M}_l$ is the state of the computation after column l. Note where the field axioms were used. Step 3 divides by $a \neq 0$, which needs (F8), and nothing else in the argument depends on which field we work over.

Example. Gauss–Jordan on a 3×4 system

Solve the following system over $\mathbb{R}$ by Gauss–Jordan elimination, naming each operation:

$$\begin{aligned} 2x_2 - 2x_3 + x_4 &= 3, \\ x_1 + x_2 + x_3 &= 3, \\ 2x_1 + x_2 + 3x_3 + x_4 &= 3. \end{aligned}$$

Solution. We work on the augmented matrix. **Column 1.** Row 1 has a 0 there, so we bring up row 2, whose entry is already 1. Then we clear the rest of column 1.

$$\left[\begin{array}{cccc|c} 0 & 2 & -2 & 1 & 3 \\ 1 & 1 & 1 & 0 & 3 \\ 2 & 1 & 3 & 1 & 3 \end{array}\right] \xrightarrow{R_1 \leftrightarrow R_2} \left[\begin{array}{cccc|c} 1 & 1 & 1 & 0 & 3 \\ 0 & 2 & -2 & 1 & 3 \\ 2 & 1 & 3 & 1 & 3 \end{array}\right]$$

$$\xrightarrow{R_3 \to R_3 - 2R_1} \left[\begin{array}{cccc|c} 1 & 1 & 1 & 0 & 3 \\ 0 & 2 & -2 & 1 & 3 \\ 0 & -1 & 1 & 1 & -3 \end{array}\right]$$

Column 2. The candidates below row 1 are 2 and -1. Choosing -1 avoids fractions, so we swap it up and scale.

$$\xrightarrow{R_2 \leftrightarrow R_3} \left[\begin{array}{cccc|c} 1 & 1 & 1 & 0 & 3 \\ 0 & -1 & 1 & 1 & -3 \\ 0 & 2 & -2 & 1 & 3 \end{array}\right] \xrightarrow{R_2 \to -R_2} \left[\begin{array}{cccc|c} 1 & 1 & 1 & 0 & 3 \\ 0 & 1 & -1 & -1 & 3 \\ 0 & 2 & -2 & 1 & 3 \end{array}\right]$$

Now clear column 2 above and below the pivot.

$$\xrightarrow{R_1\to R_1-R_2}\left[\begin{array}{cccc|c}1&0&2&1&0\\0&1&-1&-1&3\\0&2&-2&1&3\end{array}\right]\xrightarrow{R_3\to R_3-2R_2}\left[\begin{array}{cccc|c}1&0&2&1&0\\0&1&-1&-1&3\\0&0&0&3&-3\end{array}\right]$$

Column 3. The only entry below row 2 is 0, so column 3 gets no pivot. **Column 4.** The entry 3 in row 3 becomes the pivot.

$$\xrightarrow{R_3\to\frac{1}{3}R_3}\left[\begin{array}{cccc|c}1&0&2&1&0\\0&1&-1&-1&3\\0&0&0&1&-1\end{array}\right]$$

$$\xrightarrow{R_1\to R_1-R_3}\left[\begin{array}{cccc|c}1&0&2&0&1\\0&1&-1&-1&3\\0&0&0&1&-1\end{array}\right]$$

$$\xrightarrow{R_2\to R_2+R_3}\left[\begin{array}{cccc|c}1&0&2&0&1\\0&1&-1&0&2\\0&0&0&1&-1\end{array}\right]$$

Column 5. No rows remain below row 3, so we stop. The final matrix is in RREF, with pivot columns 1, 2, and 4. It is the system

$$x_1+2x_3=1,\qquad x_2-x_3=2,\qquad x_4=-1,$$

which by Theorem 3.2.2 has the same solutions as the original. The unknown x_3 is unconstrained; writing $x_3=t$, the solutions are

$$(1-2t,\ 2+t,\ t,\ -1)=(1,2,0,-1)+t(-2,1,1,0),\qquad t\in\mathbb{R}.$$

Check: $(1,2,0,-1)$ gives $4+0-1=3$, $1+2+0=3$, and $2+2+0-1=3$; and $(-2,1,1,0)$ gives $2-2+0=0$, $-2+1+1=0$, and $-4+1+3+0=0$. This is the "particular plus homogeneous" form of Theorem 3.1.3.

The same computation can be carried out in exact arithmetic. In the online version the cell below can be edited and run, which is a good way to check a hand computation.

```
from sympy import Matrix

M = Matrix([
    [0, 2, -2, 1, 3],
    [1, 1,  1, 0, 3],
    [2, 1,  3, 1, 3],
])
R, pivots = M.rref()   # the RREF and the (0-based) pivot columns
print(R)
print("pivot columns:", [j + 1 for j in pivots])
```

Warning. **Echelon forms are not unique; only the reduced form is.** Different choices in step 2 give different intermediate matrices. For $\begin{pmatrix}1&2\\3&4\end{pmatrix}$ over $\mathbb{R}$, the operation $R_2\to$

$R_2 - 3R_1$ gives the row echelon form $\begin{pmatrix} 1 & 2 \\ 0 & -2 \end{pmatrix}$. Starting instead with $R_1 \leftrightarrow R_2$ and then $R_2 \to R_2 - \frac{1}{3}R_1$ gives the row echelon form $\begin{pmatrix} 3 & 4 \\ 0 & \frac{2}{3} \end{pmatrix}$. Both continue to the same RREF, $\mathbf{I}_2$. That this always happens is the theorem of the next section; until it is proved, "the" RREF of a matrix is not yet justified.

3.2.5 Reading off the solutions

Once $[\,\mathbf{A} \mid \mathbf{b}\,]$ is in reduced row echelon form, the example shows what to do: a pivot in the last column signals a contradiction, and otherwise the unknowns without pivots can be chosen freely. We name the two kinds of unknowns.

Suppose $[\,\mathbf{A} \mid \mathbf{b}\,]$, with $\mathbf{A} \in M_{m\times n}(F)$, is row equivalent to a matrix $[\,\mathbf{R} \mid \mathbf{c}\,]$ in reduced row echelon form, where $\mathbf{R}$ is $m \times n$. For $l \le n$, the unknown x_l is a **pivot variable** if l is a pivot column of $[\,\mathbf{R} \mid \mathbf{c}\,]$, and a **free variable** otherwise. (By the uniqueness theorem of the next section, this does not depend on which sequence of operations was used.)

Theorem 3.2.4. Reading solutions from the RREF

Let $\mathbf{A} \in M_{m\times n}(F)$ and $\mathbf{b} \in F^m$, and suppose $[\,\mathbf{A} \mid \mathbf{b}\,]$ is row equivalent to a matrix $[\,\mathbf{R} \mid \mathbf{c}\,]$ in reduced row echelon form. Let r_{il} be the entries of $\mathbf{R}$ and $\mathbf{c} = (c_1, \dots, c_m)$.

(a) $\mathbf{Ax} = \mathbf{b}$ is inconsistent **if and only if** the last column, column $n+1$, is a pivot column of $[\,\mathbf{R} \mid \mathbf{c}\,]$, that is, some row of $[\,\mathbf{R} \mid \mathbf{c}\,]$ is $(0 \ \cdots \ 0 \mid 1)$.

(b) Suppose there is no pivot in the last column. Let $j_1 < \cdots < j_r$ be the pivot columns. Then for **every** choice of scalars $t_l \in F$, one for each free variable x_l, there is **exactly one** solution with $x_l = t_l$ for all free l, namely the one with

$$x_{j_i} = c_i - \sum_{l \text{ free}} r_{il}\, t_l \qquad (i = 1, \dots, r).$$

In particular, setting every free variable to 0 gives the solution $\mathbf{p}$ with $p_{j_i} = c_i$ and all other entries 0.

(c) A consistent system $\mathbf{Ax} = \mathbf{b}$ has exactly one solution **if and only if** it has no free variables, that is, each of the columns $1, \dots, n$ is a pivot column.

Proof. By Theorem 3.2.2 (c), $\mathbf{Ax} = \mathbf{b}$ has the same solution set as $\mathbf{Rx} = \mathbf{c}$, so we work with the latter.

(a, $\Leftarrow$) If row i of $[\,\mathbf{R} \mid \mathbf{c}\,]$ has its pivot in column $n+1$, then its first n entries are 0 and $c_i = 1$ by (E3). Equation i reads $0 = 1$, which is false in any field (Definition 0.7.1). So there is no solution.

(b) Suppose no pivot lies in column $n+1$. Then all pivots of $[\,\mathbf{R} \mid \mathbf{c}\,]$ lie in columns $j_1, \dots, j_r \le n$. By (E1), rows $r+1, \dots, m$ of $[\,\mathbf{R} \mid \mathbf{c}\,]$ are zero, so the corresponding equations read $0 = 0$ and hold for every $\mathbf{x}$. For $i \le r$, row i of $\mathbf{R}$ has $r_{ij_i} = 1$ by (E3) and $r_{ij_h} = 0$ for $h \ne i$ by (E4). So equation i reads

$$x_{j_i} + \sum_{l \text{ free}} r_{il}\, x_l = c_i.$$

Hence $\mathbf{x}$ is a solution if and only if $x_{j_i} = c_i - \sum_{l \text{ free}} r_{il} x_l$ for $i = 1, \dots, r$. Each unknown is either a pivot variable or a free variable, not both. So once the free variables are given the values t_l, these equations determine every pivot variable uniquely, and the resulting vector is a solution. This proves (b).

(a, $\Rightarrow$) This is the contrapositive of what (b) just showed: if there is no pivot in the last column, then choosing all $t_l = 0$ produces a solution, so the system is consistent.

(c) By (a), a consistent system has no pivot in the last column, so (b) applies. If there are no free variables, (b) gives exactly one solution. If x_l is free, the choices $t_l = 0$ and $t_l = 1$, with all other free variables 0, give two solutions whose l-th entries differ, since $0 \neq 1$. This proves (c).

■

In Example (Gauss–Jordan on a 3×4 system) the pivot columns were 1, 2, 4, the free variable was x_3, and (b) gave $x_1 = 1 - 2t$, $x_2 = 2 + t$, $x_4 = -1$, with $\mathbf{p} = (1, 2, 0, -1)$. The same recipe works over any field, and over a finite field the free variables can take only finitely many values.

Example. Elimination over $\mathbb{F}_2$

Solve the following system over $\mathbb{F}_2$:

$$x_1 + x_2 + x_4 = 1, \qquad x_2 + x_3 + x_4 = 0, \qquad x_1 + x_3 + x_4 = 1.$$

How many solutions are there? What changes if the same equations are read over $\mathbb{R}$?

Solution. In $\mathbb{F}_2$, $-1 = 1$, so $R_i \to R_i - R_j$ is the same operation as $R_i \to R_i + R_j$, and every nonzero pivot is already 1. **Column 1** has its pivot in row 1; clear row 3. **Column 2** has its pivot in row 2; clear rows 3 and 1.

$$\left[\begin{array}{cccc|c} 1 & 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 & 1 \end{array}\right] \xrightarrow{R_3 \to R_3 + R_1} \left[\begin{array}{cccc|c} 1 & 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 \end{array}\right] \xrightarrow{R_3 \to R_3 + R_2} \left[\begin{array}{cccc|c} 1 & 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{array}\right]$$

$$\xrightarrow{R_1 \to R_1 + R_2} \left[\begin{array}{cccc|c} 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{array}\right]$$

In the row $R_3 + R_1$, the entries $1 + 1$ became 0, and in $R_3 + R_2$ the entry in column 3 is $1 + 1 = 0$. **Column 3** has only a 0 below row 2, so no pivot. **Column 4** has its pivot in row 3; row 1 already has a 0 there, so only row 2 needs clearing.

$$\xrightarrow{R_2 \to R_2 + R_3} \left[\begin{array}{cccc|c} 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{array}\right]$$

This is in RREF, with pivot columns 1, 2, 4 and no pivot in the last column, so the system is consistent by Theorem 3.2.4 (a). The free variable is $x_3 = t$, and part (b) gives $x_1 = 1 - t = 1 + t$, $x_2 = -t = t$, $x_4 = 0$. Since $t \in \mathbb{F}_2 = \{0, 1\}$, there are exactly two solutions:

$$(1, 0, 0, 0) \quad \text{and} \quad (0, 1, 1, 0).$$

Check the second: $0+1+0=1$, $1+1+0=0$, $0+1+0=1$ in $\mathbb{F}_2$.

Over $\mathbb{R}$ the first step is $R_3 \to R_3 - R_1$, giving row 3 $= (0,-1,1,0 \mid 0)$, and adding row 2 then gives $(0,0,2,1 \mid 0)$. The entry 2 is nonzero in $\mathbb{R}$, so column 3 now gets a pivot. Completing the reduction gives pivot columns 1, 2, 3, free variable x_4, and infinitely many real solutions. The same equations lead to different pivot patterns over different fields, because $1+1=0$ in one and not in the other.

When a coefficient depends on a parameter, whether a column gets a pivot can depend on the parameter too. Then the reduction splits into cases, one for each answer to "is this entry zero?".

Example. A system with a parameter

For which $k \in \mathbb{R}$ does the system $x + ky = 1$, $kx + y = 1$ have no solution, exactly one solution, or infinitely many? Find the solutions when they exist.

Solution. The entry in position $(1,1)$ is 1 for every k, so it is a pivot with no case distinction:

$$\left[\begin{array}{cc|c} 1 & k & 1 \\ k & 1 & 1 \end{array}\right] \xrightarrow{R_2 \to R_2 - kR_1} \left[\begin{array}{cc|c} 1 & k & 1 \\ 0 & 1-k^2 & 1-k \end{array}\right].$$

(For $k=0$ this operation changes nothing, which is allowed.) Column 2 gets a pivot in row 2 exactly when $1-k^2 = (1-k)(1+k)$ is nonzero, and only then may we scale by its inverse. So we split into cases.

Case $k \neq \pm 1$. Then $1-k^2 \neq 0$, and $R_2 \to (1-k^2)^{-1}R_2$ turns row 2 into $(0,1 \mid \frac{1}{1+k})$, since $\frac{1-k}{(1-k)(1+k)} = \frac{1}{1+k}$. Then $R_1 \to R_1 - kR_2$ turns row 1 into $(1,0 \mid 1 - \frac{k}{1+k}) = (1,0 \mid \frac{1}{1+k})$. The RREF is $[\,\mathbf{I}_2 \mid \mathbf{c}\,]$ with $\mathbf{c} = (\frac{1}{1+k}, \frac{1}{1+k})$, and by Theorem 3.2.4 (c) this is the only solution. Check: $\frac{1}{1+k} + \frac{k}{1+k} = 1$ for both equations.

Case $k = 1$. Row 2 is $(0,0 \mid 0)$, and $\left[\begin{array}{cc|c} 1 & 1 & 1 \\ 0 & 0 & 0 \end{array}\right]$ is in RREF with pivot column 1. The free variable is $y = t$, and Theorem 3.2.4 (b) gives the solutions $(1-t, t)$, $t \in \mathbb{R}$: infinitely many.

Case $k = -1$. Row 2 is $(0,0 \mid 2)$, and $R_2 \to \frac{1}{2}R_2$ makes it $(0,0 \mid 1)$. The last column is a pivot column, so the system is inconsistent by Theorem 3.2.4 (a). Directly: $x - y = 1$ and $-x + y = 1$ add up to $0 = 2$.

Hence there is no solution for $k = -1$, there are infinitely many for $k = 1$, and there is exactly one otherwise. The trap is to scale by $(1-k^2)^{-1}$ before asking whether $1-k^2$ is zero: that step silently assumes $k \neq \pm 1$, and loses exactly the two interesting cases.

3.2.6 More unknowns than equations

Counting pivots gives a first general theorem. A matrix with m rows has at most m pivots, one per nonzero row. If there are more unknowns than that, some unknown must be free.

Corollary 3.2.5. More unknowns than equations

Let $\mathbf{A} \in M_{m\times n}(F)$ with $n > m$.

(a) The homogeneous system $\mathbf{Ax} = \mathbf{0}$ has a **nontrivial** solution.

(b) Consequently, every list of n vectors in F^m with $n > m$ is linearly dependent.

Proof.

(a) By Theorem 3.2.3, $[\,\mathbf{A} \mid \mathbf{0}\,]$ is row equivalent to a matrix $[\,\mathbf{R} \mid \mathbf{c}\,]$ in RREF. By Lemma 3.2.1, each operation keeps the last column zero, so $\mathbf{c} = \mathbf{0}$, and the last column is not a pivot column. Each pivot is the leading entry of a different row, so there are at most m pivots, hence at most $m < n$ pivot columns among the first n columns. So some x_l is a free variable. By Theorem 3.2.4 (b), with $t_l = 1$ and every other free variable 0, there is a solution whose l-th entry is $1 \neq 0$. This solution is nontrivial.

(b) Let $\mathbf{v}_1, \ldots, \mathbf{v}_n \in F^m$ and let $\mathbf{A} \in M_{m\times n}(F)$ have these columns. By (a) there is $\mathbf{x} \neq \mathbf{0}$ with $\mathbf{Ax} = \mathbf{0}$, and by Theorem 0.9.3 this says $x_1\mathbf{v}_1 + \cdots + x_n\mathbf{v}_n = \mathbf{0}$ with $x_1, \ldots, x_n$ not all zero. So the list is linearly dependent.

■

Part (b) is not new: since $\dim F^m = m$, it is also a case of Theorem 1.6.3, which rests on the Steinitz Exchange Theorem (Theorem 1.6.1). The proof here is independent of that one. It uses nothing from Chapter 1 beyond the definition of independence, only elimination and a count of pivots. Two proofs by different routes give some confidence that the dimension theory of Chapter 1 and the computations of this chapter will fit together, and the rest of the chapter makes that fit exact.

3.2.7 Exercises

3.2.7.1 A. Check your understanding

Exercise. A1

(a) List the three types of elementary row operation, including the conditions on the scalar and on the indices.

(b) Determine whether the following statement is true: "every matrix in row echelon form is in reduced row echelon form." Justify your answer.

(c) Determine whether the following statement is true: "if the RREF of $[\,\mathbf{A} \mid \mathbf{b}\,]$ has a zero row, then $\mathbf{Ax} = \mathbf{b}$ is inconsistent." Justify your answer.

(d) Determine whether the following statement is true: "a homogeneous system of 3 equations in 5 unknowns over $\mathbb{R}$ has infinitely many solutions." Justify your answer.

(e) Explain why swapping two columns of $\mathbf{A}$ in $[\,\mathbf{A} \mid \mathbf{b}\,]$ is **not** a safe step when solving $\mathbf{Ax} = \mathbf{b}$.

Solution.

(a) $R_i \leftrightarrow R_j$ with $i \neq j$; $R_i \to cR_i$ with $c \neq 0$; $R_i \to R_i + cR_j$ with $i \neq j$ and any $c \in F$.

(b) False. $\begin{pmatrix} 2 & 1 \\ 0 & 1 \end{pmatrix}$ is in row echelon form, but its first pivot is $2 \neq 1$, so (E3) fails; its

second column also has a nonzero entry above the pivot, so (E4) fails.

(c) False. A zero row is the equation $0 = 0$, which every vector satisfies. Inconsistency means a row $(0 \ \cdots \ 0 \mid 1)$ (Theorem 3.2.4 (a)). For example $\left[\begin{array}{c|c} 1 & 1 \\ 0 & 0 \end{array}\right]$ is the consistent system $x = 1$, $0 = 0$.

(d) True. By Corollary 3.2.5 ($5 > 3$) there is a nontrivial solution $\mathbf{h}$, and since $\mathbb{R}$ is infinite, the multiples $c\mathbf{h}$ ($c \in \mathbb{R}$) are infinitely many distinct solutions, as in Exercise (C1: How many solutions?).

(e) It swaps the roles of two unknowns: $\mathbf{s}$ solves the new system exactly when the vector obtained from $\mathbf{s}$ by swapping those two entries solves the old one. So the solution set changes, unless the swap happens to fix it. For instance $x = 1$, $y = 2$ becomes $y = 1$, $x = 2$.

3.2.7.2 B. Practice

Exercise. B1: Reduce and solve

Reduce the augmented matrix of each system over $\mathbb{R}$ to reduced row echelon form, naming each operation. Hence decide whether the system is consistent, and if so, describe all solutions.

(a) $x_1 + 2x_2 - x_3 = 1$, $2x_1 + 5x_2 + x_3 = 4$, $x_1 + 3x_2 + 2x_3 = 4$.

(b) $x_1 + x_2 + 2x_3 = 3$, $2x_1 + 3x_2 + 5x_3 = 7$, $x_1 + 2x_2 + 4x_3 = 5$.

(c) $x_1 - x_2 + 2x_3 + x_4 = 2$, $2x_1 - 2x_2 + 5x_3 + 4x_4 = 5$, $-x_1 + x_2 - x_3 + x_4 = -1$.

Solution.

(a)

$$\left[\begin{array}{ccc|c} 1 & 2 & -1 & 1 \\ 2 & 5 & 1 & 4 \\ 1 & 3 & 2 & 4 \end{array}\right] \xrightarrow{R_2 \to R_2 - 2R_1} \left[\begin{array}{ccc|c} 1 & 2 & -1 & 1 \\ 0 & 1 & 3 & 2 \\ 1 & 3 & 2 & 4 \end{array}\right] \xrightarrow{R_3 \to R_3 - R_1} \left[\begin{array}{ccc|c} 1 & 2 & -1 & 1 \\ 0 & 1 & 3 & 2 \\ 0 & 1 & 3 & 3 \end{array}\right]$$

$$\xrightarrow{R_1 \to R_1 - 2R_2} \left[\begin{array}{ccc|c} 1 & 0 & -7 & -3 \\ 0 & 1 & 3 & 2 \\ 0 & 1 & 3 & 3 \end{array}\right] \xrightarrow{R_3 \to R_3 - R_2} \left[\begin{array}{ccc|c} 1 & 0 & -7 & -3 \\ 0 & 1 & 3 & 2 \\ 0 & 0 & 0 & 1 \end{array}\right]$$

Column 3 gets no pivot. Column 4 has its pivot in row 3:

$$\xrightarrow{R_1 \to R_1 + 3R_3} \left[\begin{array}{ccc|c} 1 & 0 & -7 & 0 \\ 0 & 1 & 3 & 2 \\ 0 & 0 & 0 & 1 \end{array}\right] \xrightarrow{R_2 \to R_2 - 2R_3} \left[\begin{array}{ccc|c} 1 & 0 & -7 & 0 \\ 0 & 1 & 3 & 0 \\ 0 & 0 & 0 & 1 \end{array}\right].$$

The last column is a pivot column, so by Theorem 3.2.4 (a) the system is inconsistent. (Row 3 already said $0 = 1$ before the last two operations.)

(b)

$$\left[\begin{array}{ccc|c} 1 & 1 & 2 & 3 \\ 2 & 3 & 5 & 7 \\ 1 & 2 & 4 & 5 \end{array}\right] \xrightarrow{R_2 \to R_2 - 2R_1} \left[\begin{array}{ccc|c} 1 & 1 & 2 & 3 \\ 0 & 1 & 1 & 1 \\ 1 & 2 & 4 & 5 \end{array}\right] \xrightarrow{R_3 \to R_3 - R_1} \left[\begin{array}{ccc|c} 1 & 1 & 2 & 3 \\ 0 & 1 & 1 & 1 \\ 0 & 1 & 2 & 2 \end{array}\right]$$

$$\xrightarrow{R_1 \to R_1 - R_2} \left[\begin{array}{ccc|c} 1 & 0 & 1 & 2 \\ 0 & 1 & 1 & 1 \\ 0 & 1 & 2 & 2 \end{array}\right]$$

$$\xrightarrow{R_3 \to R_3 - R_2} \left[\begin{array}{ccc|c} 1 & 0 & 1 & 2 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{array}\right]$$

$$\xrightarrow{R_1 \to R_1 - R_3} \left[\begin{array}{ccc|c} 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{array}\right]$$

$$\xrightarrow{R_2 \to R_2 - R_3} \left[\begin{array}{ccc|c} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{array}\right].$$

Columns 1, 2, 3 are pivot columns and the last column is not, so by Theorem 3.2.4 (c) there is exactly one solution, $(1, 0, 1)$. Check: $1+0+2 = 3$, $2+0+5 = 7$, $1+0+4 = 5$.

(c)

$$\left[\begin{array}{cccc|c} 1 & -1 & 2 & 1 & 2 \\ 2 & -2 & 5 & 4 & 5 \\ -1 & 1 & -1 & 1 & -1 \end{array}\right] \xrightarrow{R_2 \to R_2 - 2R_1} \left[\begin{array}{cccc|c} 1 & -1 & 2 & 1 & 2 \\ 0 & 0 & 1 & 2 & 1 \\ -1 & 1 & -1 & 1 & -1 \end{array}\right]$$

$$\xrightarrow{R_3 \to R_3 + R_1} \left[\begin{array}{cccc|c} 1 & -1 & 2 & 1 & 2 \\ 0 & 0 & 1 & 2 & 1 \\ 0 & 0 & 1 & 2 & 1 \end{array}\right]$$

Column 2 has only zeros below row 1, so no pivot. Column 3 has its pivot in row 2:

$$\xrightarrow{R_1 \to R_1 - 2R_2} \left[\begin{array}{cccc|c} 1 & -1 & 0 & -3 & 0 \\ 0 & 0 & 1 & 2 & 1 \\ 0 & 0 & 1 & 2 & 1 \end{array}\right] \xrightarrow{R_3 \to R_3 - R_2} \left[\begin{array}{cccc|c} 1 & -1 & 0 & -3 & 0 \\ 0 & 0 & 1 & 2 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right].$$

Columns 4 and 5 have only a 0 in row 3, so this is the RREF. The pivot columns are 1 and 3; the free variables are $x_2 = s$ and $x_4 = t$. By Theorem 3.2.4 (b), $x_1 = s + 3t$ and $x_3 = 1 - 2t$. Hence the solution set is

$$\{(0, 0, 1, 0) + s(1, 1, 0, 0) + t(3, 0, -2, 1) : s, t \in \mathbb{R}\}.$$

Exercise. B2: A parameter

For which $k \in \mathbb{R}$ is the system

$$x + y - z = 2, \qquad x + 2y + z = 3, \qquad x + y + (k^2 - 5)z = k$$

inconsistent, which give exactly one solution, and which give infinitely many? Justify your answer, and find the solutions when they exist.

Solution.

$$\left[\begin{array}{ccc|c} 1 & 1 & -1 & 2 \\ 1 & 2 & 1 & 3 \\ 1 & 1 & k^2-5 & k \end{array}\right] \xrightarrow{R_2 \to R_2 - R_1} \left[\begin{array}{ccc|c} 1 & 1 & -1 & 2 \\ 0 & 1 & 2 & 1 \\ 1 & 1 & k^2-5 & k \end{array}\right]$$

$$\xrightarrow{R_3 \to R_3 - R_1} \left[\begin{array}{ccc|c} 1 & 1 & -1 & 2 \\ 0 & 1 & 2 & 1 \\ 0 & 0 & k^2-4 & k-2 \end{array}\right]$$

$$\xrightarrow{R_1 \to R_1 - R_2} \left[\begin{array}{ccc|c} 1 & 0 & -3 & 1 \\ 0 & 1 & 2 & 1 \\ 0 & 0 & k^2-4 & k-2 \end{array}\right].$$

Whether column 3 gets a pivot depends on whether $k^2 - 4 = (k-2)(k+2)$ is zero.

Case $k \neq \pm 2$. Then $k^2 - 4 \neq 0$, and $R_3 \to (k^2-4)^{-1} R_3$ turns row 3 into $(0, 0, 1 \mid \frac{1}{k+2})$, since $\frac{k-2}{(k-2)(k+2)} = \frac{1}{k+2}$. Then $R_1 \to R_1 + 3R_3$ and $R_2 \to R_2 - 2R_3$ give the RREF $[\,\mathbf{I}_3 \mid \mathbf{c}\,]$ with

$$\mathbf{c} = \left(\frac{k+5}{k+2}, \frac{k}{k+2}, \frac{1}{k+2}\right),$$

using $1 + \frac{3}{k+2} = \frac{k+5}{k+2}$ and $1 - \frac{2}{k+2} = \frac{k}{k+2}$. By Theorem 3.2.4 (c) there is exactly one solution, $\mathbf{c}$.

Case $k = 2$. Row 3 is zero, and the matrix is in RREF with pivot columns 1, 2 and free variable $z = t$. The solutions are $(1 + 3t, 1 - 2t, t)$, $t \in \mathbb{R}$: infinitely many.

Case $k = -2$. Row 3 is $(0, 0, 0 \mid -4)$, and $R_3 \to -\frac{1}{4} R_3$ makes it $(0, 0, 0 \mid 1)$. The last column is a pivot column, so the system is inconsistent.

Hence: inconsistent for $k = -2$, infinitely many solutions for $k = 2$, and exactly one solution otherwise.

Exercise. B3: Over $\mathbb{F}_3$

Solve each system over $\mathbb{F}_3$ by reducing to RREF, naming each operation, and list all solutions.

(a) $x + 2y + z = 1$, $2x + y + z = 0$, $x + z = 2$.

(b) $x + y + 2z = 1$, $2x + z = 2$.

Solution. In $\mathbb{F}_3$ the inverses are $1^{-1} = 1$ and $2^{-1} = 2$, since $2 \cdot 2 = 4 = 1$, and $-1 = 2$, $-2 = 1$.

(a)

$$\left[\begin{array}{ccc|c} 1 & 2 & 1 & 1 \\ 2 & 1 & 1 & 0 \\ 1 & 0 & 1 & 2 \end{array}\right] \xrightarrow{R_2 \to R_2 - 2R_1} \left[\begin{array}{ccc|c} 1 & 2 & 1 & 1 \\ 0 & 0 & 2 & 1 \\ 1 & 0 & 1 & 2 \end{array}\right] \xrightarrow{R_3 \to R_3 - R_1} \left[\begin{array}{ccc|c} 1 & 2 & 1 & 1 \\ 0 & 0 & 2 & 1 \\ 0 & 1 & 0 & 1 \end{array}\right]$$

Here $1 - 4 = -3 = 0$, $1 - 2 = -1 = 2$, $0 - 2 = 1$, and $0 - 2 = 1$, $2 - 1 = 1$. Column 2

has entries $0, 1$ below row 1, so we swap:

$$\xrightarrow{R_2 \leftrightarrow R_3} \left[\begin{array}{ccc|c} 1 & 2 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 2 & 1 \end{array}\right]$$

$$\xrightarrow{R_1 \to R_1 - 2R_2} \left[\begin{array}{ccc|c} 1 & 0 & 1 & 2 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 2 & 1 \end{array}\right]$$

$$\xrightarrow{R_3 \to 2R_3} \left[\begin{array}{ccc|c} 1 & 0 & 1 & 2 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 2 \end{array}\right]$$

$$\xrightarrow{R_1 \to R_1 - R_3} \left[\begin{array}{ccc|c} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 2 \end{array}\right].$$

There are no free variables, so the unique solution is $(0, 1, 2)$. Check: $0+2+2 = 4 = 1$, $0 + 1 + 2 = 3 = 0$, $0 + 2 = 2$.

(b)
$$\left[\begin{array}{ccc|c} 1 & 1 & 2 & 1 \\ 2 & 0 & 1 & 2 \end{array}\right] \xrightarrow{R_2 \to R_2 - 2R_1} \left[\begin{array}{ccc|c} 1 & 1 & 2 & 1 \\ 0 & 1 & 0 & 0 \end{array}\right] \xrightarrow{R_1 \to R_1 - R_2} \left[\begin{array}{ccc|c} 1 & 0 & 2 & 1 \\ 0 & 1 & 0 & 0 \end{array}\right],$$

using $0 - 2 = 1$, $1 - 4 = 0$, $2 - 2 = 0$. The free variable is $z = t$, and $x = 1 - 2t = 1 + t$, $y = 0$. As t runs through $0, 1, 2$, the solutions are $(1, 0, 0)$, $(2, 0, 1)$, $(0, 0, 2)$: exactly three.

3.2.7.3 C. Going deeper

Exercise. C1: Why $i \neq j$?

Consider the move "$R_i \to R_i + cR_i$", which the definition does **not** allow.

(a) Show that for $c = -1$ this move is not reversible: give a matrix on which its effect cannot be undone by any sequence of elementary row operations and moves of this kind, and a system whose solution set it changes.

(b) Show that for $c \neq -1$ the move is an elementary row operation of another type. Explain why the condition $i \neq j$ in type (3) therefore loses nothing.

Solution.

(a) The move replaces $\mathbf{r}_i$ by $\mathbf{r}_i + (-1)\mathbf{r}_i = \mathbf{0}$. Apply it to the 1×2 augmented matrix $[\,1 \mid 1\,]$ of the equation $x = 1$ over any field F: the result is $[\,0 \mid 0\,]$, the equation $0 = 0$. The solution set changes from $\{1\}$ to all of F. No sequence of moves can return to $[\,1 \mid 1\,]$: every elementary row operation, and every move "$R_i \to R_i + cR_i$", sends a zero matrix to a zero matrix, by the formulas in the proof of Lemma 3.2.1 and $c \cdot 0 = 0$.

(b) The move replaces $\mathbf{r}_i$ by $(1 + c)\mathbf{r}_i$. If $c \neq -1$, then $1 + c \neq 0$, and this is the scaling $R_i \to (1 + c)R_i$, an operation of type (2). So allowing $i = j$ in type (3) would add only moves that are already available (when $c \neq -1$) and one move that destroys information ($c = -1$). Requiring $i \neq j$ excludes exactly the bad case.

Exercise. C2: The cost of elimination

Let $\mathbf{A} \in M_n(F)$ and $\mathbf{b} \in F^n$, and suppose that Gauss–Jordan elimination on $[\,\mathbf{A} \mid \mathbf{b}\,]$ finds, for each $l = 1, \dots, n$, a nonzero entry in position (l, l) and so never swaps. Count as one **operation** each multiplication or division of two scalars, under these conventions: scaling row l costs one division for each entry in columns $l+1, \dots, n+1$ (the pivot is simply set to 1), and each replacement $R_g \to R_g - cR_l$ costs one multiplication for each entry in columns $l+1, \dots, n+1$ (the entry in column l is simply set to 0).

(a) Show that processing column l costs at most $n(n+1-l)$ operations.

(b) Hence show that the whole elimination costs at most $\frac{1}{2}(n^3 + n^2)$ operations.

(c) Roughly how many times more work is a system of $2n$ equations than one of n equations, for large n?

Solution.

(a) Scaling row l costs $n+1-l$ divisions. At most $n-1$ other rows need a replacement, each costing $n+1-l$ multiplications. The total is at most $(n+1-l) + (n-1)(n+1-l) = n(n+1-l)$.

(b) Summing (a) over $l = 1, \dots, n$, and substituting $q = n+1-l$, which runs through $1, \dots, n$,

$$\sum_{l=1}^{n} n(n+1-l) = n \sum_{q=1}^{n} q = n \cdot \frac{n(n+1)}{2} = \frac{n^3 + n^2}{2},$$

using the formula for $1 + \cdots + n$ (Proposition 0.2.8).

(c) The bound is dominated by $\frac{1}{2}n^3$ for large n, since $\frac{1}{2}n^2$ is small compared with it. Replacing n by $2n$ multiplies n^3 by 8, so doubling the size of the system costs roughly eight times as much.

Exercise. C3: Solvable for every right-hand side

Let $\mathbf{A} \in M_{m \times n}(F)$, and suppose that $\mathbf{Ax} = \mathbf{b}$ is consistent for **every** $\mathbf{b} \in F^m$. Prove that $m \le n$.

Hint: suppose $m > n$, reduce $\mathbf{A}$ to its RREF $\mathbf{R}$, and run the operations backwards on $[\,\mathbf{R} \mid \mathbf{e}_m\,]$.

Solution. Suppose, for a contradiction, that $m > n$. By Theorem 3.2.3, some sequence of operations $O_1, \dots, O_s$ turns $\mathbf{A}$ into a matrix $\mathbf{R}$ in RREF. By (E2), $\mathbf{R}$ has at most n pivots, one in each pivot column, so at most $n < m$ nonzero rows, and by (E1) its last row is zero. Therefore the last row of $[\,\mathbf{R} \mid \mathbf{e}_m\,]$ is $(0 \ \cdots \ 0 \mid 1)$, so the system $\mathbf{Rx} = \mathbf{e}_m$ has no solution.

By Theorem 3.2.2 (a), each O_t is undone by an operation O_t', and $O_s', \dots, O_1'$ turn $\mathbf{R}$ back into $\mathbf{A}$. Apply them to $[\,\mathbf{R} \mid \mathbf{e}_m\,]$. By Lemma 3.2.1 they act on the two blocks separately, so the result is $[\,\mathbf{A} \mid \mathbf{b}\,]$ for some $\mathbf{b} \in F^m$. By Theorem 3.2.2 (c), $\mathbf{Ax} = \mathbf{b}$ has the same solutions as $\mathbf{Rx} = \mathbf{e}_m$, namely none. This contradicts the hypothesis, so $m \le n$.

3.3 Uniqueness of the Reduced Form

Gauss–Jordan elimination involves choices: which nonzero entry to use as a pivot, and in what order to perform the operations. Different choices give different intermediate matrices, and different row echelon forms. The previous section claimed that they always end at the **same** reduced row echelon form. This matters more than it seems. Pivot columns, free variables, and later the rank of a matrix are all read off from the reduced form; if two people could reduce the same matrix to two different answers, none of these would be well defined. This section proves the claim.

3.3.1 What row operations remember

Row operations change the columns of a matrix. What do they keep? Take

$$\mathbf{A} = \begin{pmatrix} 1 & 2 & 1 & 3 \\ 2 & 4 & 0 & 2 \\ 3 & 6 & 2 & 7 \end{pmatrix} \quad \text{and its RREF} \quad \mathbf{R} = \begin{pmatrix} 1 & 2 & 0 & 1 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

over $\mathbb{R}$ (the computation is carried out in an example later in this section). The columns of $\mathbf{R}$ look nothing like those of $\mathbf{A}$. But look at how the columns depend on each other. In $\mathbf{R}$, column 2 is twice column 1, and column 4 is column 1 plus twice column 3. In $\mathbf{A}$, with columns $\mathbf{a}_1, \dots, \mathbf{a}_4$,

$$\mathbf{a}_2 = (2,4,6) = 2\mathbf{a}_1, \qquad \mathbf{a}_4 = (3,2,7) = (1,2,3) + 2(1,0,2) = \mathbf{a}_1 + 2\mathbf{a}_3.$$

The same relations, with the same coefficients. This is no accident. A relation $x_1\mathbf{a}_1 + \cdots + x_n\mathbf{a}_n = \mathbf{0}$ among the columns says $\mathbf{Ax} = \mathbf{0}$ (Theorem 0.9.3), and row operations do not change the solutions of a homogeneous system.

Lemma 3.3.1. Row equivalent matrices have the same homogeneous solutions

Let $\mathbf{A}, \mathbf{B} \in M_{m\times n}(F)$ be row equivalent. Then for every $\mathbf{x} \in F^n$, $\mathbf{Ax} = \mathbf{0}$ if and only if $\mathbf{Bx} = \mathbf{0}$.

Proof. Let a sequence of elementary row operations turn $\mathbf{A}$ into $\mathbf{B}$. By Lemma 3.2.1, the same sequence turns $[\,\mathbf{A} \mid \mathbf{0}\,]$ into $[\,\mathbf{B} \mid \mathbf{0}'\,]$, where $\mathbf{0}'$ is obtained from the zero column by row operations and so is again the zero column. Hence $[\,\mathbf{A} \mid \mathbf{0}\,]$ and $[\,\mathbf{B} \mid \mathbf{0}\,]$ are row equivalent, and by Theorem 3.2.2 (c) the systems $\mathbf{Ax} = \mathbf{0}$ and $\mathbf{Bx} = \mathbf{0}$ have the same solution set. ∎

Rewriting $\mathbf{Ax}$ as a combination of columns turns the lemma into the statement we observed.

Lemma 3.3.2. Row operations preserve column relations

Let $\mathbf{A}, \mathbf{B} \in M_{m\times n}(F)$ have columns $\mathbf{a}_1, \dots, \mathbf{a}_n$ and $\mathbf{b}_1, \dots, \mathbf{b}_n$, and suppose that $\mathbf{Ax} = \mathbf{0} \Leftrightarrow \mathbf{Bx} = \mathbf{0}$ for every $\mathbf{x} \in F^n$. This holds, for instance, if $\mathbf{A}$ and $\mathbf{B}$ are row equivalent (Lemma 3.3.1). Then:

(a) for all $x_1, \dots, x_n \in F$: $x_1\mathbf{a}_1 + \cdots + x_n\mathbf{a}_n = \mathbf{0}$ if and only if $x_1\mathbf{b}_1 + \cdots + x_n\mathbf{b}_n = \mathbf{0}$;

(b) for every index l and all scalars $d_1, \dots, d_{l-1}$: $\mathbf{a}_l = d_1\mathbf{a}_1 + \cdots + d_{l-1}\mathbf{a}_{l-1}$ if and only if

$\mathbf{b}_l = d_1\mathbf{b}_1 + \cdots + d_{l-1}\mathbf{b}_{l-1}$;

(c) for every index l: $\mathbf{a}_l \in \operatorname{span}(\mathbf{a}_1, \ldots, \mathbf{a}_{l-1})$ if and only if $\mathbf{b}_l \in \operatorname{span}(\mathbf{b}_1, \ldots, \mathbf{b}_{l-1})$.

Proof.

(a) By Theorem 0.9.3, $x_1\mathbf{a}_1 + \cdots + x_n\mathbf{a}_n = \mathbf{Ax}$ and $x_1\mathbf{b}_1 + \cdots + x_n\mathbf{b}_n = \mathbf{Bx}$, where $\mathbf{x} = (x_1, \ldots, x_n)$. The hypothesis says $\mathbf{Ax} = \mathbf{0} \Leftrightarrow \mathbf{Bx} = \mathbf{0}$.

(b) The equation $\mathbf{a}_l = d_1\mathbf{a}_1 + \cdots + d_{l-1}\mathbf{a}_{l-1}$ holds if and only if $d_1\mathbf{a}_1 + \cdots + d_{l-1}\mathbf{a}_{l-1} + (-1)\mathbf{a}_l + 0\mathbf{a}_{l+1} + \cdots + 0\mathbf{a}_n = \mathbf{0}$. By (a), with these coefficients, this holds if and only if the same combination of the $\mathbf{b}_j$ is $\mathbf{0}$, that is, $\mathbf{b}_l = d_1\mathbf{b}_1 + \cdots + d_{l-1}\mathbf{b}_{l-1}$.

(c) By Definition 1.3.2, $\mathbf{a}_l \in \operatorname{span}(\mathbf{a}_1, \ldots, \mathbf{a}_{l-1})$ means that the left side of (b) holds for **some** scalars $d_1, \ldots, d_{l-1}$. By (b) this happens if and only if the right side holds for some scalars, which means $\mathbf{b}_l \in \operatorname{span}(\mathbf{b}_1, \ldots, \mathbf{b}_{l-1})$.

■

Warning. **Row operations preserve column relations, not columns or their span.** For $\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 1 & 2 \end{pmatrix}$ over $\mathbb{R}$, the operation $R_2 \to R_2 - R_1$ gives $\begin{pmatrix} 1 & 2 \\ 0 & 0 \end{pmatrix}$. The relation "column 2 = 2· column 1" holds in both. But every column of $\mathbf{A}$ lies on the line $\operatorname{span}((1,1))$, and every column of the new matrix lies on the different line $\operatorname{span}((1,0))$. So the span of the columns changed. To answer a question about the columns of $\mathbf{A}$ themselves, find the relations from the reduced form, and then use them on the columns of $\mathbf{A}$.

3.3.2 The columns of a reduced matrix

In a matrix in RREF, the column relations can be read off at a glance: pivot columns are standard basis vectors, and a nonpivot column **displays its own coefficients**. We record this precisely, together with a description of the pivot columns that does not mention pivots at all.

Lemma 3.3.3. Columns of a matrix in RREF

Let $\mathbf{R} \in M_{m\times n}(F)$ be in reduced row echelon form, with entries r_{il}, columns $\mathbf{u}_1, \ldots, \mathbf{u}_n$, and pivot columns $j_1 < \cdots < j_r$.

(a) For each $i = 1, \ldots, r$, the pivot column $\mathbf{u}_{j_i}$ is $\mathbf{e}_i$.

(b) If l is **not** a pivot column, then

$$\mathbf{u}_l = \sum_{i\,:\,j_i<l} r_{il}\,\mathbf{e}_i = \sum_{i\,:\,j_i<l} r_{il}\,\mathbf{u}_{j_i}.$$

(c) Column l is a pivot column **if and only if** $\mathbf{u}_l \notin \operatorname{span}(\mathbf{u}_1, \ldots, \mathbf{u}_{l-1})$.

Proof.

(a) By (E3) the entry of $\mathbf{u}_{j_i}$ in row i is 1, and by (E4) its other entries are 0.

(b) Let l be a nonpivot column, and look at the entry r_{il} in each row i. If $i > r$, row i is

zero by (E1), so $r_{il} = 0$. If $i \le r$ and $j_i > l$, then r_{il} lies strictly before the leading entry of row i, so $r_{il} = 0$. The case $j_i = l$ does not occur, since l is not a pivot column. So the only possibly nonzero entries of $\mathbf{u}_l$ are the r_{il} with $j_i < l$, which gives the first expression; the second follows from (a).

(c) If l is not a pivot column, (b) writes $\mathbf{u}_l$ as a combination of the columns $\mathbf{u}_{j_i}$ with $j_i < l$, so $\mathbf{u}_l \in \operatorname{span}(\mathbf{u}_1, \dots, \mathbf{u}_{l-1})$. Conversely, let $l = j_i$ be a pivot column. For each $l' < j_i$, the entry $r_{il'}$ lies before the leading entry of row i, so it is 0. Hence every vector of $\operatorname{span}(\mathbf{u}_1, \dots, \mathbf{u}_{l-1})$, being a combination of vectors with i-th entry 0, has i-th entry 0. But $\mathbf{u}_l = \mathbf{e}_i$ has i-th entry $1 \neq 0$, so $\mathbf{u}_l \notin \operatorname{span}(\mathbf{u}_1, \dots, \mathbf{u}_{l-1})$.

■

Part (c) says: reading the columns from left to right, a column gets a pivot exactly when it is **new**, not a combination of the columns before it. That condition is about column relations only, and column relations survive row operations. This is the whole idea of the uniqueness proof.

3.3.3 The uniqueness theorem

We can now prove that the reduced form depends only on the homogeneous solutions, and hence not on the operations used to reach it.

Theorem 3.3.4. Uniqueness of the reduced row echelon form

Let $\mathbf{R}, \mathbf{R}' \in M_{m\times n}(F)$ be in reduced row echelon form, and suppose that $\mathbf{Rx} = \mathbf{0} \Leftrightarrow \mathbf{R}'\mathbf{x} = \mathbf{0}$ for every $\mathbf{x} \in F^n$. Then $\mathbf{R} = \mathbf{R}'$.

Consequently, every matrix $\mathbf{A} \in M_{m\times n}(F)$ is row equivalent to **exactly one** matrix in reduced row echelon form, called **the RREF of A**.

Idea. A column dependence $\sum_j x_j \mathbf{u}_j = \mathbf{0}$ is exactly a null vector $\mathbf{x}$, so $\mathbf{R}$ and $\mathbf{R}'$ have the same column relations. The pivot columns are exactly the columns that are not combinations of earlier columns (Lemma 3.3.3 (c)), so $\mathbf{R}$ and $\mathbf{R}'$ have the same pivot columns. A pivot column is a standard basis vector in both. A nonpivot column of an RREF matrix is determined by how it is built from the pivot columns before it (Lemma 3.3.3 (b)), and that recipe is a column relation, so it is the same in both. Hence every column agrees.

Proof. Let $\mathbf{u}_1, \dots, \mathbf{u}_n$ and $\mathbf{u}'_1, \dots, \mathbf{u}'_n$ be the columns of $\mathbf{R}$ and $\mathbf{R}'$, and r_{il} the entries of $\mathbf{R}$. By hypothesis, Lemma 3.3.2 applies to $\mathbf{R}$ and $\mathbf{R}'$.

Same pivot columns. For each l, by Lemma 3.3.2 (c), $\mathbf{u}_l \in \operatorname{span}(\mathbf{u}_1, \dots, \mathbf{u}_{l-1})$ if and only if $\mathbf{u}'_l \in \operatorname{span}(\mathbf{u}'_1, \dots, \mathbf{u}'_{l-1})$. By Lemma 3.3.3 (c) applied to $\mathbf{R}$ and to $\mathbf{R}'$, column l is a pivot column of $\mathbf{R}$ if and only if it is a pivot column of $\mathbf{R}'$. So both matrices have the same pivot columns $j_1 < \cdots < j_r$.

Pivot columns agree. By Lemma 3.3.3 (a), $\mathbf{u}_{j_i} = \mathbf{e}_i = \mathbf{u}'_{j_i}$ for $i = 1, \dots, r$.

Same nonpivot columns. Let l be a nonpivot column. By Lemma 3.3.3 (b) for $\mathbf{R}$,

$$\mathbf{u}_l = \sum_{i\,:\,j_i<l} r_{il}\,\mathbf{u}_{j_i}.$$

This is an equation of the form in Lemma 3.3.2 (b), with coefficient r_{il} on column j_i and 0 on the other earlier columns. By that lemma, the same equation holds for $\mathbf{R}'$:

$$\mathbf{u}'_l = \sum_{i\,:\,j_i<l} r_{il}\,\mathbf{u}'_{j_i} = \sum_{i\,:\,j_i<l} r_{il}\,\mathbf{e}_i = \mathbf{u}_l,$$

where the middle equality uses Lemma 3.3.3 (a) for $\mathbf{R}'$, and the last uses Lemma 3.3.3 (b) for $\mathbf{R}$.

Hence $\mathbf{u}_l = \mathbf{u}'_l$ for every l, and $\mathbf{R} = \mathbf{R}'$.

For the consequence, let $\mathbf{A} \in M_{m\times n}(F)$. By Theorem 3.2.3, $\mathbf{A}$ is row equivalent to at least one matrix in RREF. Suppose $\mathbf{A}$ is row equivalent to $\mathbf{R}$ and to $\mathbf{R}'$, both in RREF. Row equivalence is an equivalence relation (Theorem 3.2.2 (b)), so $\mathbf{R}$ is row equivalent to $\mathbf{R}'$. By Lemma 3.3.1, $\mathbf{Rx} = \mathbf{0} \Leftrightarrow \mathbf{R}'\mathbf{x} = \mathbf{0}$ for every $\mathbf{x}$, and by the first part $\mathbf{R} = \mathbf{R}'$. This proves the theorem.

■

Notice what the proof did **not** do: it never compared the two sequences of operations. A direct comparison of two runs of the algorithm would have to track every choice, and quickly becomes unmanageable. The proof instead found a description of the RREF, "pivots where a new column appears, and coefficients of the old columns elsewhere", that refers only to the null space, which the operations cannot change.

The theorem also gives a practical test: two matrices of the same size are row equivalent exactly when their RREFs are equal (you will write out the two-line proof in Exercise (B2: Row equivalence via the RREF)). For example, over $\mathbb{R}$, $\begin{pmatrix}1&2\\2&4\end{pmatrix}$ and $\begin{pmatrix}2&4\\0&0\end{pmatrix}$ both reduce to $\begin{pmatrix}1&2\\0&0\end{pmatrix}$, by $R_2 \to R_2 - 2R_1$ and by $R_1 \to \frac{1}{2}R_1$ respectively, so they are row equivalent. But $\begin{pmatrix}1&3\\2&6\end{pmatrix}$ reduces to $\begin{pmatrix}1&3\\0&0\end{pmatrix} \neq \begin{pmatrix}1&2\\0&0\end{pmatrix}$, so it is row equivalent to neither, even though all three matrices have the same pivot columns.

Example. Two echelon forms, one reduced form

Let $\mathbf{A} = \begin{pmatrix}1&2&1&3\\2&4&0&2\\3&6&2&7\end{pmatrix}$ over $\mathbb{R}$.

(a) Reduce $\mathbf{A}$ to a row echelon form in two ways: first using row 1 as the first pivot row, then starting with $R_1 \leftrightarrow R_2$.

(b) Continue both to reduced row echelon form.

(c) Read the column relations of $\mathbf{A}$ off the RREF.

Solution.

(a) First route:

$$\mathbf{A} \xrightarrow{R_2 \to R_2 - 2R_1} \begin{pmatrix} 1 & 2 & 1 & 3 \\ 0 & 0 & -2 & -4 \\ 3 & 6 & 2 & 7 \end{pmatrix} \xrightarrow{R_3 \to R_3 - 3R_1} \begin{pmatrix} 1 & 2 & 1 & 3 \\ 0 & 0 & -2 & -4 \\ 0 & 0 & -1 & -2 \end{pmatrix}$$
$$\xrightarrow{R_3 \to R_3 - \frac{1}{2}R_2} \begin{pmatrix} 1 & 2 & 1 & 3 \\ 0 & 0 & -2 & -4 \\ 0 & 0 & 0 & 0 \end{pmatrix} = \mathbf{E}_1.$$

Second route:

$$\mathbf{A} \xrightarrow{R_1 \leftrightarrow R_2} \begin{pmatrix} 2 & 4 & 0 & 2 \\ 1 & 2 & 1 & 3 \\ 3 & 6 & 2 & 7 \end{pmatrix} \xrightarrow{R_2 \to R_2 - \frac{1}{2}R_1} \begin{pmatrix} 2 & 4 & 0 & 2 \\ 0 & 0 & 1 & 2 \\ 3 & 6 & 2 & 7 \end{pmatrix}$$
$$\xrightarrow{R_3 \to R_3 - \frac{3}{2}R_1} \begin{pmatrix} 2 & 4 & 0 & 2 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 2 & 4 \end{pmatrix}$$
$$\xrightarrow{R_3 \to R_3 - 2R_2} \begin{pmatrix} 2 & 4 & 0 & 2 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 0 \end{pmatrix} = \mathbf{E}_2.$$

Both $\mathbf{E}_1$ and $\mathbf{E}_2$ are in row echelon form, and $\mathbf{E}_1 \neq \mathbf{E}_2$. Both have pivot columns 1 and 3.

(b) From $\mathbf{E}_1$:

$$\mathbf{E}_1 \xrightarrow{R_2 \to -\frac{1}{2}R_2} \begin{pmatrix} 1 & 2 & 1 & 3 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 0 \end{pmatrix} \xrightarrow{R_1 \to R_1 - R_2} \begin{pmatrix} 1 & 2 & 0 & 1 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

From $\mathbf{E}_2$:

$$\mathbf{E}_2 \xrightarrow{R_1 \to \frac{1}{2}R_1} \begin{pmatrix} 1 & 2 & 0 & 1 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

The two routes end at the same matrix $\mathbf{R}$, as Theorem 3.3.4 guarantees.

(c) By Lemma 3.3.3 (b), the nonpivot columns of $\mathbf{R}$ are $\mathbf{u}_2 = 2\mathbf{u}_1$ and $\mathbf{u}_4 = 1 \cdot \mathbf{u}_1 + 2\mathbf{u}_3$, reading the coefficients from the entries of each column. By Lemma 3.3.2 (b), the same relations hold for the columns of $\mathbf{A}$: $\mathbf{a}_2 = 2\mathbf{a}_1$ and $\mathbf{a}_4 = \mathbf{a}_1 + 2\mathbf{a}_3$. Check: $2(1,2,3) = (2,4,6)$ and $(1,2,3) + 2(1,0,2) = (3,2,7)$.

Both echelon forms in the example had the same pivot columns, although they were different matrices. That is also true in general, and it is the practical version of the theorem.

3.3.4 Pivot columns are well defined

Corollary 3.3.5. Pivot columns are well defined

Let $\mathbf{A} \in M_{m\times n}(F)$ have columns $\mathbf{a}_1, \ldots, \mathbf{a}_n$. Define the **pivot columns of A** to be the pivot columns of the RREF of $\mathbf{A}$.

(a) Column l is a pivot column of $\mathbf{A}$ **if and only if** $\mathbf{a}_l \notin \operatorname{span}(\mathbf{a}_1, \ldots, \mathbf{a}_{l-1})$.

(b) If $\mathbf{E}$ is **any** matrix in row echelon form that is row equivalent to $\mathbf{A}$, then the pivot columns of $\mathbf{E}$ are the pivot columns of $\mathbf{A}$. In particular, every row echelon form of $\mathbf{A}$ has the same number of nonzero rows.

Idea. Part (a) combines Lemma 3.3.3 (c) with the fact that column relations pass from the RREF back to $\mathbf{A}$. For (b), finish the reduction of $\mathbf{E}$: scale each pivot to 1 and clear the entries above the pivots. These operations never move a leading entry, so they produce a matrix in RREF with the same pivot columns as $\mathbf{E}$, and by uniqueness it is the RREF of $\mathbf{A}$.

Proof. The RREF of $\mathbf{A}$ is well defined by Theorem 3.3.4; let $\mathbf{u}_1, \ldots, \mathbf{u}_n$ be its columns.

(a) $\mathbf{A}$ and its RREF are row equivalent, so by Lemma 3.3.2 (c), $\mathbf{a}_l \in \operatorname{span}(\mathbf{a}_1, \ldots, \mathbf{a}_{l-1})$ if and only if $\mathbf{u}_l \in \operatorname{span}(\mathbf{u}_1, \ldots, \mathbf{u}_{l-1})$. By Lemma 3.3.3 (c), the latter fails exactly when l is a pivot column.

(b) Let $\mathbf{E}$ have entries e_{hl} and pivot columns $j_1 < \cdots < j_s$. First perform $R_i \to e_{ij_i}^{-1} R_i$ for $i = 1, \ldots, s$, which is legal since each pivot e_{ij_i} is nonzero. Scaling a row by a nonzero scalar does not change which of its entries are zero (Theorem 0.7.1 (d), (f)), so the result is still in row echelon form with the same pivot columns, and now every pivot is 1. Next, for $i = 1, 2, \ldots, s$ in turn, and for each $h < i$, perform $R_h \to R_h - cR_i$, where c is the current entry of row h in column j_i. Row i is zero in all columns before j_i, so this changes row h only in columns $\geq j_i > j_h$. Hence the leading entry of row h is still a 1 in column j_h, no zero row is touched, and the entries below every pivot stay 0. The operation makes the entry in position (h, j_i) equal to 0, and it does not spoil the entries $(h, j_{i'})$ with $i' < i$ cleared earlier, because row i has a 0 in column $j_{i'} < j_i$. At the end, every pivot column is a standard basis vector, so the matrix is in RREF, with pivot columns $j_1, \ldots, j_s$.

This matrix is row equivalent to $\mathbf{E}$, hence to $\mathbf{A}$ (Theorem 3.2.2 (b)). By Theorem 3.3.4 it is the RREF of $\mathbf{A}$. Therefore the pivot columns of $\mathbf{E}$ are those of $\mathbf{A}$. Since the nonzero rows of $\mathbf{E}$ correspond one-to-one to its pivots, their number s is the number of pivot columns of $\mathbf{A}$, whichever $\mathbf{E}$ we chose.

■

In practice, (b) means we may stop at any row echelon form to find the pivot columns, and (a) means we can sometimes find them with no elimination at all. Part (a) also says that the pivot columns of $\mathbf{A}$ are exactly the columns kept when the list $(\mathbf{a}_1, \ldots, \mathbf{a}_n)$ is sifted as in Theorem 1.5.4: a column is kept exactly when it is not in the span of the columns before it. The number of pivot columns is the first candidate for a measure of "how much a matrix contains"; later in this chapter it becomes the rank.

Quick check. Without computing any row operations, find the pivot columns of $\mathbf{A} = \begin{pmatrix} 1 & 2 & 0 & 1 \\ 2 & 4 & 1 & 3 \end{pmatrix}$ over $\mathbb{R}$.

Solution. Use Corollary 3.3.5 (a) column by column. $\mathbf{a}_1 = (1,2) \neq \mathbf{0}$, so it is not in $\text{span}() = \{\mathbf{0}\}$: pivot. $\mathbf{a}_2 = (2,4) = 2\mathbf{a}_1$: not a pivot. $\mathbf{a}_3 = (0,1)$ is not a multiple of $(1,2)$, and $\text{span}(\mathbf{a}_1, \mathbf{a}_2) = \text{span}((1,2))$: pivot. $\mathbf{a}_4 = (1,3) = \mathbf{a}_1 + \mathbf{a}_3$: not a pivot. The pivot columns are 1 and 3.

Warning. **Uniqueness is a property of the reduced form only.** A matrix has many row echelon forms (Example (Two echelon forms, one reduced form) gave two), and it is only their pivot positions that agree. Also, two different matrices in RREF with the **same** pivot columns need not be row equivalent: $\begin{pmatrix} 1 & 2 \end{pmatrix}$ and $\begin{pmatrix} 1 & 3 \end{pmatrix}$ are both in RREF with pivot column 1, but they are different RREF matrices, so by Theorem 3.3.4 neither is row equivalent to the other.

3.3.5 Exercises

3.3.5.1 A. Check your understanding

Exercise. A1

(a) State the uniqueness theorem for the reduced row echelon form.

(b) Determine whether the following statement is true: "if $\mathbf{A}$ and $\mathbf{B}$ are row equivalent, then the columns of $\mathbf{A}$ and the columns of $\mathbf{B}$ span the same subspace." Justify your answer.

(c) Determine whether the following statement is true: "if $\mathbf{A}$ and $\mathbf{B}$ are row equivalent and $\mathbf{a}_3 = \mathbf{a}_1 - \mathbf{a}_2$ for the columns of $\mathbf{A}$, then $\mathbf{b}_3 = \mathbf{b}_1 - \mathbf{b}_2$ for the columns of $\mathbf{B}$." Justify your answer.

(d) Determine whether the following statement is true: "any two row echelon forms of the same matrix have the same number of zero rows." Justify your answer.

(e) In the RREF of a matrix, what do the entries of a nonpivot column tell you about the original matrix?

Solution.

(a) If $\mathbf{R}, \mathbf{R}' \in M_{m\times n}(F)$ are in RREF and $\mathbf{Rx} = \mathbf{0} \Leftrightarrow \mathbf{R}'\mathbf{x} = \mathbf{0}$ for all $\mathbf{x} \in F^n$, then $\mathbf{R} = \mathbf{R}'$. Hence every matrix is row equivalent to exactly one matrix in RREF.

(b) False. $\begin{pmatrix} 1 & 2 \\ 1 & 2 \end{pmatrix}$ and $\begin{pmatrix} 1 & 2 \\ 0 & 0 \end{pmatrix}$ are row equivalent, but their columns span $\text{span}((1,1))$ and $\text{span}((1,0))$, which are different lines in $\mathbb{R}^2$.

(c) True, by Lemma 3.3.2 (b) with $l = 3$, $d_1 = 1$, $d_2 = -1$.

(d) True. By Corollary 3.3.5 (b), every row echelon form of an $m \times n$ matrix $\mathbf{A}$ has exactly as many nonzero rows as $\mathbf{A}$ has pivot columns, so the number of zero rows is m minus that number.

(e) If column l is not a pivot column, its entries r_{il} (for the pivot rows i with $j_i < l$) are the coefficients expressing column l of the original matrix as a combination of its pivot columns: $\mathbf{a}_l = \sum_{j_i<l} r_{il}\, \mathbf{a}_{j_i}$.

3.3.5.2 B. Practice

Exercise. B1: Column relations

Let $\mathbf{A} = \begin{pmatrix} 1 & -1 & 0 & 2 & 1 \\ 2 & -2 & 1 & 7 & 0 \\ 1 & -1 & 1 & 5 & 1 \end{pmatrix}$ over $\mathbb{R}$. Find the RREF of $\mathbf{A}$, naming each operation. Hence find the pivot columns of $\mathbf{A}$, and express each nonpivot column of $\mathbf{A}$ as a linear combination of the pivot columns before it. Check your answer on the columns of $\mathbf{A}$.

Solution.

$$\mathbf{A} \xrightarrow{R_2 \to R_2 - 2R_1} \begin{pmatrix} 1 & -1 & 0 & 2 & 1 \\ 0 & 0 & 1 & 3 & -2 \\ 1 & -1 & 1 & 5 & 1 \end{pmatrix} \xrightarrow{R_3 \to R_3 - R_1} \begin{pmatrix} 1 & -1 & 0 & 2 & 1 \\ 0 & 0 & 1 & 3 & -2 \\ 0 & 0 & 1 & 3 & 0 \end{pmatrix}$$

$$\xrightarrow{R_3 \to R_3 - R_2} \begin{pmatrix} 1 & -1 & 0 & 2 & 1 \\ 0 & 0 & 1 & 3 & -2 \\ 0 & 0 & 0 & 0 & 2 \end{pmatrix}$$

$$\xrightarrow{R_3 \to \frac{1}{2}R_3} \begin{pmatrix} 1 & -1 & 0 & 2 & 1 \\ 0 & 0 & 1 & 3 & -2 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

$$\xrightarrow{R_1 \to R_1 - R_3} \begin{pmatrix} 1 & -1 & 0 & 2 & 0 \\ 0 & 0 & 1 & 3 & -2 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

$$\xrightarrow{R_2 \to R_2 + 2R_3} \begin{pmatrix} 1 & -1 & 0 & 2 & 0 \\ 0 & 0 & 1 & 3 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

The pivot columns are 1, 3 and 5. By Lemma 3.3.3 (b), in the RREF column 2 is -1 times column 1, and column 4 is 2 times column 1 plus 3 times column 3. By Lemma 3.3.2 (b), the columns of $\mathbf{A}$ satisfy

$$\mathbf{a}_2 = -\mathbf{a}_1, \qquad \mathbf{a}_4 = 2\mathbf{a}_1 + 3\mathbf{a}_3.$$

Check: $-(1,2,1) = (-1,-2,-1) = \mathbf{a}_2$, and $2(1,2,1) + 3(0,1,1) = (2,7,5) = \mathbf{a}_4$.

Exercise. B2: Row equivalence via the RREF

Prove that two matrices $\mathbf{A}, \mathbf{B} \in M_{m\times n}(F)$ are row equivalent if and only if they have the same RREF.

Solution. ($\Rightarrow$) Suppose $\mathbf{A}$ and $\mathbf{B}$ are row equivalent, and let $\mathbf{R}$ be the RREF of $\mathbf{B}$. Then $\mathbf{A}$ is row equivalent to $\mathbf{B}$, and $\mathbf{B}$ to $\mathbf{R}$, so $\mathbf{A}$ is row equivalent to $\mathbf{R}$ by transitivity (Theorem 3.2.2 (b)). By the uniqueness in Theorem 3.3.4, $\mathbf{R}$ is the RREF of $\mathbf{A}$.

($\Leftarrow$) Suppose both have the RREF $\mathbf{R}$. Then $\mathbf{A}$ is row equivalent to $\mathbf{R}$, and by symmetry $\mathbf{R}$ is row equivalent to $\mathbf{B}$ (Theorem 3.2.2 (b)). By transitivity $\mathbf{A}$ is row equivalent to $\mathbf{B}$.

Exercise. B3: Row equivalent or not?

Determine whether the matrices in each pair are row equivalent over $\mathbb{R}$. Justify your answer.

(a) $\mathbf{P} = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 5 & 7 \end{pmatrix}$ and $\mathbf{Q} = \begin{pmatrix} 1 & 1 & 2 \\ 3 & 4 & 7 \end{pmatrix}$.

(b) $\mathbf{P} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \\ 0 & 1 & 2 \end{pmatrix}$ and $\mathbf{Q} = \begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & 2 \\ 1 & 1 & 2 \end{pmatrix}$.

Solution. By Exercise (B2: Row equivalence via the RREF), two matrices of the same size are row equivalent exactly when their RREFs are equal. So we compute both RREFs.

(a) Row reduce both:

$$\mathbf{P} \xrightarrow{R_2 \to R_2 - 2R_1} \begin{pmatrix} 1 & 2 & 3 \\ 0 & 1 & 1 \end{pmatrix} \xrightarrow{R_1 \to R_1 - 2R_2} \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix},$$

$$\mathbf{Q} \xrightarrow{R_2 \to R_2 - 3R_1} \begin{pmatrix} 1 & 1 & 2 \\ 0 & 1 & 1 \end{pmatrix} \xrightarrow{R_1 \to R_1 - R_2} \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}.$$

The RREFs are equal, so $\mathbf{P}$ and $\mathbf{Q}$ are row equivalent.

(b) $\mathbf{P} \xrightarrow{R_2 \to R_2 - R_1} \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \\ 0 & 1 & 2 \end{pmatrix} \xrightarrow{R_1 \to R_1 - R_2} \begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & 2 \\ 0 & 1 & 2 \end{pmatrix} \xrightarrow{R_3 \to R_3 - R_2} \begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & 2 \\ 0 & 0 & 0 \end{pmatrix}.$

$$\mathbf{Q} \xrightarrow{R_3 \to R_3 - R_1} \begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & 2 \\ 0 & 1 & 3 \end{pmatrix} \xrightarrow{R_3 \to R_3 - R_2} \begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{pmatrix}$$

$$\xrightarrow{R_1 \to R_1 + R_3} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{pmatrix} \xrightarrow{R_2 \to R_2 - 2R_3} \mathbf{I}_3.$$

The RREFs differ (one has a zero row, the other is $\mathbf{I}_3$), so $\mathbf{P}$ and $\mathbf{Q}$ are **not** row equivalent. Consistently with Corollary 3.3.5, $\mathbf{P}$ has pivot columns 1, 2 while $\mathbf{Q}$ has 1, 2, 3.

3.3.5.3 C. Going deeper

Exercise. C1: Counting reduced forms

(a) List, by pivot columns, all 2×3 matrices over $\mathbb{F}_2$ in reduced row echelon form, and show that there are exactly 15 of them.

(b) Deduce that $M_{2\times 3}(\mathbb{F}_2)$, which has 64 elements, splits into exactly 15 classes of row equivalent matrices.

(c) Show that over a field with q elements there are exactly $2q^2 + 2q + 3$ such matrices.

Solution. We count over a field F with q elements, sorting by the list of pivot columns. In a matrix in RREF, a pivot column is forced (it is $\mathbf{e}_i$), an entry before a leading entry or in a zero row is forced to be 0, and every other entry, one lying in a nonpivot column to the right of its row's pivot, is free.

(a) With $q = 2$, writing $*$ for a free entry:

- no pivots: $\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$, 1 matrix;
- pivot column 1 only: $\begin{pmatrix} 1 & * & * \\ 0 & 0 & 0 \end{pmatrix}$, 4 matrices;
- pivot column 2 only: $\begin{pmatrix} 0 & 1 & * \\ 0 & 0 & 0 \end{pmatrix}$, 2 matrices;
- pivot column 3 only: $\begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$, 1 matrix;
- pivot columns 1, 2: $\begin{pmatrix} 1 & 0 & * \\ 0 & 1 & * \end{pmatrix}$, 4 matrices;
- pivot columns 1, 3: $\begin{pmatrix} 1 & * & 0 \\ 0 & 0 & 1 \end{pmatrix}$, 2 matrices;
- pivot columns 2, 3: $\begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$, 1 matrix.

A 2×3 matrix has at most two pivots, so these are all the cases. The total is $1 + 4 + 2 + 1 + 4 + 2 + 1 = 15$.

(b) Row equivalence is an equivalence relation (Theorem 3.2.2 (b)), so $M_{2\times 3}(\mathbb{F}_2)$ is partitioned into equivalence classes (Theorem 0.5.2). By Theorem 3.3.4, each class contains exactly one matrix in RREF: at least one by Theorem 3.2.3, and at most one since two RREF matrices in the same class would be row equivalent, hence equal. So the classes correspond one-to-one to the 15 matrices of (a).

(c) The same list, with each free entry taking q values, gives $1 + q^2 + q + 1 + q^2 + q + 1 = 2q^2 + 2q + 3$. For $q = 2$ this is $8 + 4 + 3 = 15$, as in (a).

Exercise. C2: Deleting a column

Let $\mathbf{A} \in M_{m\times n}(F)$ with $n \geq 2$, and let $\mathbf{R}$ be its RREF.

(a) Let $\mathbf{A}'$ and $\mathbf{R}'$ be obtained from $\mathbf{A}$ and $\mathbf{R}$ by deleting the **last** column. Prove that $\mathbf{R}'$ is the RREF of $\mathbf{A}'$.

(b) Show by an example that the corresponding statement for deleting the **first** column is false.

Solution.

(a) Let a sequence of row operations turn $\mathbf{A}$ into $\mathbf{R}$. By Lemma 3.2.1, the same sequence turns $\mathbf{A}'$ into $\mathbf{R}'$, so $\mathbf{A}'$ is row equivalent to $\mathbf{R}'$. By Theorem 3.3.4 it suffices to show that $\mathbf{R}'$ is in RREF. Let $j_1 < \cdots < j_r$ be the pivot columns of $\mathbf{R}$. If $j_r < n$, then deleting column n does not change any leading entry or any pivot column, and the zero rows of $\mathbf{R}$ stay zero, so (E1)–(E4) still hold. If $j_r = n$, then row r of $\mathbf{R}$ is zero except for its pivot in column n, so in $\mathbf{R}'$ rows $r, \ldots, m$ are zero and rows $1, \ldots, r-1$

keep their leading entries in columns $j_1 < \cdots < j_{r-1}$ and their pivot columns. Again (E1)–(E4) hold, with $r-1$ pivots. Hence $\mathbf{R}'$ is in RREF, and it is the RREF of $\mathbf{A}'$.

(b) Let $\mathbf{A} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ over $\mathbb{R}$. Then $\mathbf{A} \xrightarrow{R_1 \to R_1 - R_2} \mathbf{I}_2$, so its RREF is $\mathbf{I}_2$, and deleting the first column of $\mathbf{I}_2$ gives $\begin{pmatrix} 0 \\ 1 \end{pmatrix}$. But deleting the first column of $\mathbf{A}$ gives $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$, whose RREF is $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$, via $R_2 \to R_2 - R_1$. These differ. The pivot of a later column depends on the columns before it, so removing an early column can move pivots.

3.4 Elementary Matrices and Inverses

Gaussian elimination is a list of instructions: swap these rows, scale that one, add a multiple of one row to another. So far it lives outside algebra, as something we do to a matrix. This section turns every row operation into a matrix, so that "apply an operation" becomes "multiply on the left". With that one observation the algorithm becomes a tool for proofs. We use it to prove the main theorem about invertible matrices, to show that a one-sided inverse of a square matrix is automatically two-sided, which Chapter 0 stated after Definition 0.9.13 but could not prove, and to compute inverses.

Throughout, F is a field, and we use the three elementary row operations of Definition 3.2.1 with their notation: the swap $R_i \leftrightarrow R_j$ ($i \neq j$), the scaling $R_i \to cR_i$ (c **nonzero**), and the replacement $R_i \to R_i + cR_j$ ($i \neq j$, any $c \in F$).

3.4.1 Row operations are matrix multiplications

We have already seen one row operation disguised as a product. In Example (Row equivalence) of Chapter 0, multiplying $\mathbf{A} = \begin{pmatrix} 1 & 2 & 1 \\ 2 & 5 & 0 \end{pmatrix}$ on the left by $\mathbf{E} = \begin{pmatrix} 1 & 0 \\ -2 & 1 \end{pmatrix}$ replaced row 2 by row 2 minus twice row 1, that is, it performed $R_2 \to R_2 - 2R_1$. Look at $\mathbf{E}$ itself: it is what that same operation does to the identity matrix $\mathbf{I}_2$. This is no accident, and it deserves a name.

An elementary matrix is the identity matrix after one row operation.

Definition 3.4.1. Elementary matrix

Let $m \geq 1$. A matrix $\mathbf{E} \in M_m(F)$ is an **elementary matrix** if it is obtained from $\mathbf{I}_m$ by performing **exactly one** elementary row operation. Writing $\mathbf{E}_{ij}$ for the matrix unit (1 in entry (i, j), 0 elsewhere; see Example (Standard Bases)), the three kinds are:

1. $\mathbf{P}_{ij}$, obtained by $R_i \leftrightarrow R_j$ with $i \neq j$: the identity with rows i and j swapped;
2. $\mathbf{D}_i(c) = \operatorname{diag}(1, \ldots, 1, c, 1, \ldots, 1)$, with c in position i, obtained by $R_i \to cR_i$ with $c \neq 0$;
3. $\mathbf{I}_m + c\mathbf{E}_{ij}$, obtained by $R_i \to R_i + cR_j$ with $i \neq j$ and $c \in F$: the identity with an extra c in entry (i, j).

In words: start from $\mathbf{I}_m$, do **one** operation, and record the result. The size m is the number of **rows** of the matrices we plan to operate on. Type 3 puts c in row i, column j: the row that changes is i, and the row that gets copied is j.

Examples. In $M_3(F)$,

$$\mathbf{P}_{13} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \qquad \mathbf{D}_2(5) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \qquad \mathbf{I}_3 - 3\mathbf{E}_{31} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -3 & 0 & 1 \end{pmatrix}$$

come from $R_1 \leftrightarrow R_3$, $R_2 \to 5R_2$ (over a field where $5 \neq 0$), and $R_3 \to R_3 - 3R_1$. The degenerate case is worth a look: $\mathbf{D}_1(1) = \mathbf{I}_m$ is elementary, since "multiply row 1 by 1" is a legal operation, and so is $\mathbf{I}_m + 0\mathbf{E}_{12}$. The identity is an elementary matrix. For $m = 1$ the elementary matrices are exactly the 1×1 matrices (c) with $c \neq 0$, since there are no two distinct rows to swap or combine.

Nonexample by minimal change. Change $c = 5$ in $\mathbf{D}_2(5)$ to $c = 0$. The matrix $\operatorname{diag}(1, 0, 1)$ is still diagonal and still differs from $\mathbf{I}_3$ in one entry, but "multiply row 2 by 0" is not an elementary row operation: the clause $c \neq 0$ fails. Likewise $\begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$ is a rearrangement of the rows of $\mathbf{I}_3$, but every one of its rows has moved, and a single swap moves only two rows. So it is not elementary. It is, however, the **product** $\mathbf{P}_{12}\mathbf{P}_{23}$ of two elementary matrices, as Exercise A1 at the end of this section checks.

The whole point of the definition is the next theorem. To perform a row operation on **any** matrix, multiply on the left by what that operation does to the identity.

Theorem 3.4.1. Row operations are left multiplications

Let ρ be an elementary row operation on matrices with m rows, and let $\mathbf{E} = \rho(\mathbf{I}_m)$ be the corresponding elementary matrix. Then for **every** $n \geq 1$ and every $\mathbf{A} \in M_{m\times n}(F)$,

$$\rho(\mathbf{A}) = \mathbf{E}\mathbf{A}.$$

Idea. Each operation builds every new row as a combination of the old rows, with coefficients that do not depend on the matrix: for $R_i \to R_i + cR_j$, new row i is $1 \cdot (\text{row } i) + c \cdot (\text{row } j)$, and every other row is copied. Feed in $\mathbf{I}_m$, whose rows are the standard row vectors, and the coefficients appear as the entries of $\mathbf{E}$. Then the row view of a product says $\mathbf{EA}$ forms exactly those combinations of the rows of $\mathbf{A}$.

Proof. Let $\mathbf{r}_1, \dots, \mathbf{r}_m$ be the rows of $\mathbf{A}$, and write $\mathbf{e}_l^\top$ for the l-th row of $\mathbf{I}_m$. By Definition 3.2.1, for each k there are scalars $c_{k1}, \dots, c_{km}$, depending on ρ but **not** on $\mathbf{A}$, such that row k of $\rho(\mathbf{A})$ is $c_{k1}\mathbf{r}_1 + \cdots + c_{km}\mathbf{r}_m$. Explicitly, every row that the operation does not touch has $c_{kk} = 1$ and all other $c_{kl} = 0$, and the touched rows are:

- for $R_i \leftrightarrow R_j$: row i is $\mathbf{r}_j$ and row j is $\mathbf{r}_i$;
- for $R_i \to cR_i$: row i is $c\mathbf{r}_i$;
- for $R_i \to R_i + cR_j$: row i is $\mathbf{r}_i + c\mathbf{r}_j$.

Applying ρ to $\mathbf{I}_m$, row k of $\mathbf{E}$ is $c_{k1}\mathbf{e}_1^\top + \cdots + c_{km}\mathbf{e}_m^\top = \begin{pmatrix} c_{k1} & \cdots & c_{km} \end{pmatrix}$. So the (k, l)-entry of $\mathbf{E}$ is c_{kl}. By Theorem 0.9.2 (rows), row k of $\mathbf{EA}$ is (row k of $\mathbf{E}$) $\mathbf{A}$, and by Theorem 0.9.3 (rows version) this equals $c_{k1}\mathbf{r}_1 + \cdots + c_{km}\mathbf{r}_m$, which is row k of $\rho(\mathbf{A})$. Since every row agrees, $\rho(\mathbf{A}) = \mathbf{EA}$. This proves the theorem.

■

So a sequence of operations $\rho_1, \rho_2, \dots, \rho_k$, performed in that order, turns $\mathbf{A}$ into

$$\rho_k(\cdots \rho_2(\rho_1(\mathbf{A}))\cdots) = \mathbf{E}_k \cdots \mathbf{E}_2\mathbf{E}_1\mathbf{A}, \qquad \mathbf{E}_i = \rho_i(\mathbf{I}_m),$$

by applying the theorem k times and using associativity. The whole of Gaussian elimination is a single matrix $\mathbf{E}_k \cdots \mathbf{E}_1$ standing on the left.

Warning. The **first** operation stands **rightmost**. Doing $R_1 \leftrightarrow R_2$ and then $R_1 \to R_1 + R_2$ on $\mathbf{I}_2$ gives $(\mathbf{I}_2 + \mathbf{E}_{12})\mathbf{P}_{12} = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}$, while the product in the other order is $\mathbf{P}_{12}(\mathbf{I}_2 + \mathbf{E}_{12}) = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}$. Also, multiplying by $\mathbf{E}$ on the **right** performs a **column** operation: $\mathbf{A}(\mathbf{I}_2 + \mathbf{E}_{12})$ adds column 1 to column 2, and column operations change solution sets.

Example. Three operations, three products

Let $\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 0 & 1 \\ 3 & 4 \end{pmatrix} \in M_{3\times 2}(\mathbb{R})$. Compute $\mathbf{P}_{13}\mathbf{A}$, $\mathbf{D}_2(5)\mathbf{A}$ and $(\mathbf{I}_3 - 3\mathbf{E}_{31})\mathbf{A}$, and check that each is the result of the corresponding row operation.

Solution. By the row view, row k of $\mathbf{EA}$ is the combination of the rows of $\mathbf{A}$ whose weights are row k of $\mathbf{E}$.

- Rows of $\mathbf{P}_{13}$ are $\mathbf{e}_3^\top, \mathbf{e}_2^\top, \mathbf{e}_1^\top$, so $\mathbf{P}_{13}\mathbf{A} = \begin{pmatrix} 3 & 4 \\ 0 & 1 \\ 1 & 2 \end{pmatrix}$: rows 1 and 3 are swapped.
- Row 2 of $\mathbf{D}_2(5)$ is $5\mathbf{e}_2^\top$, so $\mathbf{D}_2(5)\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 0 & 5 \\ 3 & 4 \end{pmatrix}$: row 2 is multiplied by 5.
- Row 3 of $\mathbf{I}_3 - 3\mathbf{E}_{31}$ is $\begin{pmatrix} -3 & 0 & 1 \end{pmatrix}$, so row 3 of the product is $-3(1,2) + (3,4) = (0,-2)$, and $(\mathbf{I}_3 - 3\mathbf{E}_{31})\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 0 & 1 \\ 0 & -2 \end{pmatrix}$: this is $R_3 \to R_3 - 3R_1$.

In each case the product agrees with the operation, as Theorem 3.4.1 predicts.

Quick check. Which 3×3 elementary matrix performs $R_2 \to R_2 + 4R_3$? Which operation undoes it, and what is its matrix?

Solution. Apply the operation to $\mathbf{I}_3$: row 2 becomes $\mathbf{e}_2^\top + 4\mathbf{e}_3^\top$, so the matrix is $\mathbf{I}_3 + 4\mathbf{E}_{23}$,

with 4 in entry $(2,3)$. The operation $R_2 \to R_2 - 4R_3$ undoes it, since row 3 is unchanged by both; its matrix is $\mathbf{I}_3 - 4\mathbf{E}_{23}$.

Every row operation can be undone by another row operation, and the matching statement for matrices is:

Theorem 3.4.2. Elementary matrices are invertible

Every elementary matrix is invertible, and its inverse is an elementary matrix of the same kind:

$$\mathbf{P}_{ij}^{-1} = \mathbf{P}_{ij}, \qquad \mathbf{D}_i(c)^{-1} = \mathbf{D}_i(c^{-1}), \qquad (\mathbf{I}_m + c\mathbf{E}_{ij})^{-1} = \mathbf{I}_m - c\mathbf{E}_{ij} \quad (i \neq j).$$

Proof. Let $\mathbf{E} = \rho(\mathbf{I}_m)$, and let ρ' be the operation on the same line of the display: $\rho' = \rho$ for a swap, $R_i \to c^{-1}R_i$ for $R_i \to cR_i$ (legal, since $c \neq 0$ gives $c^{-1} \neq 0$), and $R_i \to R_i - cR_j$ for $R_i \to R_i + cR_j$. In each case $\rho'(\rho(\mathbf{A})) = \mathbf{A}$ and $\rho(\rho'(\mathbf{A})) = \mathbf{A}$ for every $\mathbf{A}$, as in Theorem 3.2.2 (a): swapping twice restores the order; scaling by c and c^{-1} multiplies row i by $cc^{-1} = 1$; and adding $c\mathbf{r}_j$ and then $-c\mathbf{r}_j$ to row i restores row i, because $i \neq j$ means row j is not changed in between. Let $\mathbf{E}' = \rho'(\mathbf{I}_m)$. By Theorem 3.4.1,

$$\mathbf{E}'\mathbf{E} = \rho'(\mathbf{E}) = \rho'(\rho(\mathbf{I}_m)) = \mathbf{I}_m \quad \text{and} \quad \mathbf{E}\mathbf{E}' = \rho(\mathbf{E}') = \rho(\rho'(\mathbf{I}_m)) = \mathbf{I}_m.$$

Hence $\mathbf{E}$ is invertible with $\mathbf{E}^{-1} = \mathbf{E}'$, which is the matrix in the display. This proves the theorem.

■

The hypothesis $i \neq j$ did real work: "add c times row i to row i" is scaling by $1 + c$, which destroys row i when $c = -1$ and cannot be undone. That is why Definition 3.2.1 excludes it.

3.4.2 The invertible matrix theorem

Chapter 0 defined an invertible matrix by the existence of a two-sided inverse, and gave only one practical test, $ad - bc \neq 0$, for 2×2 matrices (Theorem 0.9.8). For larger matrices, searching for $\mathbf{B}$ with $\mathbf{AB} = \mathbf{BA} = \mathbf{I}_n$ directly is hopeless. What we want is a list of conditions, each of which we can check by a method we already have, that are all equivalent to invertibility. Then we may test whichever is cheapest.

We need one small fact about square matrices in reduced form.

Lemma 3.4.3. Square reduced matrices

Let $\mathbf{R} \in M_n(F)$ be in reduced row echelon form. Then either $\mathbf{R} = \mathbf{I}_n$, or the last row of $\mathbf{R}$ is zero.

Proof. Suppose the last row of $\mathbf{R}$ is not zero. Since zero rows of a matrix in echelon form come after all nonzero rows (Definition 3.2.3), no row of $\mathbf{R}$ is zero, so each of the n rows has a pivot. Let the pivot of row k be in column c_k. Pivots move strictly to the right as we go down, so $1 \leq c_1 < c_2 < \cdots < c_n \leq n$. A strictly increasing list of n integers in

$\{1, \ldots, n\}$ must be $1, 2, \ldots, n$, so $c_k = k$. By Definition 3.2.4, each pivot equals 1 and is the only nonzero entry of its column, so column k of $\mathbf{R}$ is $\mathbf{e}_k$ for every k. Hence $\mathbf{R} = \mathbf{I}_n$.

■

Here is the theorem. It collects, for a square matrix, facts about solving systems, about row reduction, and about the columns as vectors in the sense of Chapter 1.

Theorem 3.4.4. Invertible Matrix Theorem

Let $\mathbf{A} \in M_n(F)$, with columns $\mathbf{a}_1, \ldots, \mathbf{a}_n \in F^n$. The following are equivalent.

(a) $\mathbf{A}$ is invertible.
(b) The homogeneous system $\mathbf{Ax} = \mathbf{0}$ has only the trivial solution $\mathbf{x} = \mathbf{0}$.
(c) The reduced row echelon form of $\mathbf{A}$ is $\mathbf{I}_n$.
(d) $\mathbf{A}$ is a product of elementary matrices.
(e) For **every** $\mathbf{b} \in F^n$, the system $\mathbf{Ax} = \mathbf{b}$ has a solution.
(f) The columns $(\mathbf{a}_1, \ldots, \mathbf{a}_n)$ are linearly independent.
(g) The columns $(\mathbf{a}_1, \ldots, \mathbf{a}_n)$ span F^n.
(h) The columns $(\mathbf{a}_1, \ldots, \mathbf{a}_n)$ form a basis of F^n.

Idea. Eight conditions would need 56 implications if proved pairwise; a cycle through them needs far fewer. The plan:

① the core cycle (a) ⇒ (b) ⇒ (c) ⇒ (d) ⇒ (a), where each arrow uses one earlier tool (cancel by $\mathbf{A}^{-1}$; row reduction; elementary matrices; products of invertible matrices);

② (b) ⇔ (f), which is the column view $\mathbf{Ax} = x_1\mathbf{a}_1 + \cdots + x_n\mathbf{a}_n$;

③ (f) ⇔ (h) ⇔ (g), which is "count instead of check" in the n-dimensional space F^n;

④ (g) ⇔ (e), which is "consistent means $\mathbf{b}$ is in the column span".

The only arrow that needs a new argument is (b) ⇒ (c); the others quote a result we already have. An RREF that is not $\mathbf{I}_n$ has a zero row, so it really has fewer equations than unknowns, and such a homogeneous system has a nontrivial solution.

Proof. **(a) ⇒ (b).** Suppose $\mathbf{A}$ is invertible and $\mathbf{Ax} = \mathbf{0}$. Multiplying on the left by $\mathbf{A}^{-1}$ and using Theorem 0.9.4, $\mathbf{x} = (\mathbf{A}^{-1}\mathbf{A})\mathbf{x} = \mathbf{A}^{-1}(\mathbf{Ax}) = \mathbf{A}^{-1}\mathbf{0} = \mathbf{0}$.

(b) ⇒ (c). Suppose (b) holds. By Theorem 3.2.3, $\mathbf{A}$ is row equivalent to a matrix $\mathbf{R}$ in reduced row echelon form, and $\mathbf{R}$ is the RREF of $\mathbf{A}$ by Theorem 3.3.4. By Lemma 3.3.1, $\mathbf{Rx} = \mathbf{0}$ has the same solutions as $\mathbf{Ax} = \mathbf{0}$, so it too has only the trivial solution. Suppose, for a contradiction, that $\mathbf{R} \neq \mathbf{I}_n$. By Lemma 3.4.3, the last row of $\mathbf{R}$ is zero. If $n = 1$, then $\mathbf{R} = \mathbf{O}_1$ and $\mathbf{x} = (1)$ is a nontrivial solution. If $n \geq 2$, the last equation of $\mathbf{Rx} = \mathbf{0}$ reads $0 = 0$, so $\mathbf{Rx} = \mathbf{0}$ has the same solutions as the homogeneous system $\mathbf{R}'\mathbf{x} = \mathbf{0}$ formed by the first $n - 1$ rows. That system has $n - 1$ equations in $n > n - 1$ unknowns, so by Corollary 3.2.5 it has a nontrivial solution. Either way we contradict (b). Hence $\mathbf{R} = \mathbf{I}_n$.

(c) ⇒ (d). Suppose the RREF of $\mathbf{A}$ is $\mathbf{I}_n$. By Definition 3.2.2 there are row operations $\rho_1, \dots, \rho_k$ taking $\mathbf{A}$ to $\mathbf{I}_n$. If $k = 0$, then $\mathbf{A} = \mathbf{I}_n = \mathbf{D}_1(1)$ is elementary. Otherwise, with $\mathbf{E}_i = \rho_i(\mathbf{I}_n)$, Theorem 3.4.1 gives $\mathbf{E}_k \cdots \mathbf{E}_1 \mathbf{A} = \mathbf{I}_n$. Each $\mathbf{E}_i$ is invertible by Theorem 3.4.2, so multiplying on the left by $\mathbf{E}_k^{-1}$, then $\mathbf{E}_{k-1}^{-1}$, and so on, gives

$$\mathbf{A} = \mathbf{E}_1^{-1} \mathbf{E}_2^{-1} \cdots \mathbf{E}_k^{-1},$$

and each $\mathbf{E}_i^{-1}$ is elementary by Theorem 3.4.2.

(d) ⇒ (a). Suppose $\mathbf{A} = \mathbf{E}_1 \cdots \mathbf{E}_k$ with each $\mathbf{E}_i$ elementary and $k \geq 1$. Each $\mathbf{E}_i$ is invertible by Theorem 3.4.2, and a product of two invertible matrices is invertible by Theorem 0.9.7 (part 3). By induction on k, $\mathbf{A}$ is invertible.

So far (a), (b), (c), (d) are equivalent.

(b) ⇔ (f). By Theorem 0.9.3, $\mathbf{A}\mathbf{x} = x_1 \mathbf{a}_1 + \cdots + x_n \mathbf{a}_n$. So "$\mathbf{A}\mathbf{x} = \mathbf{0}$ only for $\mathbf{x} = \mathbf{0}$" says exactly that $x_1 \mathbf{a}_1 + \cdots + x_n \mathbf{a}_n = \mathbf{0}$ forces $x_1 = \cdots = x_n = 0$, which is Definition 1.4.1 for the columns.

(f) ⇔ (h) ⇔ (g). The columns form a list of exactly n vectors in F^n, and $\dim F^n = n$ (Example (Dimensions of the Running Examples)). By Theorem 1.6.4, such a list is a basis if it is linearly independent, and also if it spans. Conversely a basis is independent and spanning by Definition 1.5.1. Hence (f) ⇒ (h) ⇒ (g) ⇒ (h) ⇒ (f).

(g) ⇔ (e). By Theorem 3.1.1, $\mathbf{A}\mathbf{x} = \mathbf{b}$ has a solution if and only if $\mathbf{b} \in \operatorname{span}(\mathbf{a}_1, \dots, \mathbf{a}_n)$. So (e) says every $\mathbf{b} \in F^n$ lies in this span, that is, $F^n \subseteq \operatorname{span}(\mathbf{a}_1, \dots, \mathbf{a}_n)$. The reverse inclusion always holds, so (e) is equivalent to (g).

Combining the four parts, all eight conditions are equivalent. This proves the theorem.

■

In practice the theorem is a menu. To show a matrix **is** invertible, pick the most convenient item, very often (b): take a solution of $\mathbf{A}\mathbf{x} = \mathbf{0}$ and show it is $\mathbf{0}$. To show a matrix is **not** invertible, exhibit one nonzero solution of $\mathbf{A}\mathbf{x} = \mathbf{0}$, or one linear relation among the columns. The list grows in both directions. Chapter 2 has already given the operator form of it, in terms of kernels and images (Theorem 2.5.7), and later chapters add determinants in Chapter 7, eigenvalues in Chapter 9, and positive definiteness and singular values in Chapter 13.

Quick check. Is $\mathbf{A} = \begin{pmatrix} 1 & 2 & 3 \\ 0 & 1 & 1 \\ 1 & 3 & 4 \end{pmatrix} \in M_3(\mathbb{R})$ invertible? Use whichever item of Theorem 3.4.4 is quickest.

Solution. No. The third column is the sum of the first two: $(1, 0, 1) + (2, 1, 3) = (3, 1, 4)$. So $\mathbf{a}_1 + \mathbf{a}_2 - \mathbf{a}_3 = \mathbf{0}$ is a linear relation with coefficients not all zero, item (f) fails, and hence (a) fails. Equivalently, $\mathbf{x} = (1, 1, -1)$ is a nontrivial solution of $\mathbf{A}\mathbf{x} = \mathbf{0}$.

Warning. The theorem is about **square** matrices. For $\mathbf{A} = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix} \in M_{2\times 3}(\mathbb{R})$, the columns span $\mathbb{R}^2$ (the first two are $\mathbf{e}_1, \mathbf{e}_2$), so (g) holds, but $\mathbf{x} = (1, 1, -1)$ solves $\mathbf{Ax} = \mathbf{0}$, so (b) fails. The proof of (f) $\Leftrightarrow$ (g) used that the number of columns equals $\dim F^n$, and that is exactly what breaks.

3.4.3 One-sided inverses are two-sided

Chapter 0 left a gap on purpose. The definition of invertibility (Definition 0.9.13) asks for **both** $\mathbf{AB} = \mathbf{I}_n$ and $\mathbf{BA} = \mathbf{I}_n$. Chapter 0 then stated that, for square matrices, the first equation already implies the second, but that nothing there could prove it: the entries of $\mathbf{AB}$ and $\mathbf{BA}$ are different sums. Now we can prove it, and the proof uses the menu above instead of entries.

Theorem 3.4.5. One-sided inverses of square matrices

Let $\mathbf{A}, \mathbf{B} \in M_n(F)$. If $\mathbf{AB} = \mathbf{I}_n$, then $\mathbf{BA} = \mathbf{I}_n$. Hence $\mathbf{A}$ and $\mathbf{B}$ are both invertible, $\mathbf{B} = \mathbf{A}^{-1}$ and $\mathbf{A} = \mathbf{B}^{-1}$.

Idea. We cannot touch $\mathbf{BA}$ directly, so we aim at invertibility of one factor instead. Which one? From $\mathbf{AB} = \mathbf{I}_n$, a vector killed by $\mathbf{B}$ is killed by $\mathbf{AB} = \mathbf{I}_n$, so it is $\mathbf{0}$. That is item (b) for $\mathbf{B}$. Once $\mathbf{B}^{-1}$ exists, we may cancel it: $\mathbf{A} = (\mathbf{AB})\mathbf{B}^{-1} = \mathbf{B}^{-1}$.

Proof. Suppose $\mathbf{AB} = \mathbf{I}_n$, and let $\mathbf{x} \in F^n$ with $\mathbf{Bx} = \mathbf{0}$. Then, by Theorem 0.9.4,

$$\mathbf{x} = \mathbf{I}_n\mathbf{x} = (\mathbf{AB})\mathbf{x} = \mathbf{A}(\mathbf{Bx}) = \mathbf{A0} = \mathbf{0}.$$

So $\mathbf{Bx} = \mathbf{0}$ has only the trivial solution, and $\mathbf{B}$ is invertible by Theorem 3.4.4 ((b) $\Rightarrow$ (a)). Multiplying $\mathbf{AB} = \mathbf{I}_n$ on the right by $\mathbf{B}^{-1}$ and using associativity,

$$\mathbf{A} = \mathbf{A}(\mathbf{BB}^{-1}) = (\mathbf{AB})\mathbf{B}^{-1} = \mathbf{I}_n\mathbf{B}^{-1} = \mathbf{B}^{-1}.$$

Therefore $\mathbf{BA} = \mathbf{BB}^{-1} = \mathbf{I}_n$. Since $\mathbf{A} = \mathbf{B}^{-1}$ is invertible with inverse $\mathbf{B}$ by Theorem 0.9.7 (part 2), we get $\mathbf{A}^{-1} = \mathbf{B}$. This proves the theorem.

■

By symmetry, $\mathbf{BA} = \mathbf{I}_n$ also implies $\mathbf{AB} = \mathbf{I}_n$: apply the theorem with the roles of $\mathbf{A}$ and $\mathbf{B}$ swapped. So from now on, **to verify that $\mathbf{B}$ is the inverse of a square matrix $\mathbf{A}$, one product suffices**. We also get a useful consequence of the same kind: if $\mathbf{A}, \mathbf{B} \in M_n(F)$ and $\mathbf{AB}$ is invertible, then $\mathbf{A}$ and $\mathbf{B}$ are invertible (Exercise B3 at the end of this section).

Warning. For **nonsquare** matrices a one-sided inverse need not be two-sided. Let

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}, \qquad \mathbf{B} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix}.$$

Then $\mathbf{AB} = \mathbf{I}_2$, but $\mathbf{BA} = \operatorname{diag}(1,1,0) \neq \mathbf{I}_3$. In fact no $\mathbf{C} \in M_{3\times 2}(F)$ has $\mathbf{CA} = \mathbf{I}_3$: $\mathbf{Ae}_3 = \mathbf{0}$, so $\mathbf{CAe}_3 = \mathbf{0} \neq \mathbf{e}_3 = \mathbf{I}_3\mathbf{e}_3$. The proof above breaks because $\mathbf{Bx} = \mathbf{0} \Rightarrow \mathbf{x} = \mathbf{0}$ no longer gives invertibility: $\mathbf{B}$ is not square, so Theorem 3.4.4 does not apply.

3.4.4 Computing the inverse

Theorem 3.4.4 (c) says $\mathbf{A}$ is invertible exactly when it row reduces to $\mathbf{I}_n$. The proof of (c) $\Rightarrow$ (d) even produced a matrix: if $\mathbf{E}_k \cdots \mathbf{E}_1\mathbf{A} = \mathbf{I}_n$, then $\mathbf{E}_k \cdots \mathbf{E}_1$ is a left inverse, hence **the** inverse by Theorem 3.4.5. We do not want to multiply out the $\mathbf{E}_i$ by hand. The trick is to let the operations act on $\mathbf{I}_n$ at the same time as on $\mathbf{A}$: since $\mathbf{E}_k \cdots \mathbf{E}_1\mathbf{I}_n = \mathbf{E}_k \cdots \mathbf{E}_1$, the operations that turn $\mathbf{A}$ into $\mathbf{I}_n$ turn $\mathbf{I}_n$ into $\mathbf{A}^{-1}$. We write $[\mathbf{A} \mid \mathbf{I}_n]$ for the $n \times 2n$ matrix whose first n columns are those of $\mathbf{A}$ and whose last n columns are those of $\mathbf{I}_n$.

Algorithm. **Input:** $\mathbf{A} \in M_n(F)$.

1. Form the $n \times 2n$ matrix $[\mathbf{A} \mid \mathbf{I}_n]$.
2. Row reduce it to reduced row echelon form $[\mathbf{R} \mid \mathbf{C}]$, with $\mathbf{R}, \mathbf{C} \in M_n(F)$.
3. If $\mathbf{R} = \mathbf{I}_n$, output $\mathbf{A}^{-1} = \mathbf{C}$. Otherwise, output "$\mathbf{A}$ is not invertible".

In practice, stop as soon as a zero row appears in the left half: $\mathbf{A}$ is then not invertible (see the remark after the theorem).

Theorem 3.4.6. Inverse by row reduction

Let $\mathbf{A} \in M_n(F)$, and let $[\mathbf{R} \mid \mathbf{C}]$ be the reduced row echelon form of $[\mathbf{A} \mid \mathbf{I}_n]$, with $\mathbf{R}, \mathbf{C} \in M_n(F)$. Then $\mathbf{R}$ is the reduced row echelon form of $\mathbf{A}$. Consequently, $\mathbf{A}$ is invertible if and only if $\mathbf{R} = \mathbf{I}_n$, and in that case $\mathbf{A}^{-1} = \mathbf{C}$.

Idea. Two things need checking. First, the operations act on the two halves independently, so the left half is transformed exactly as $\mathbf{A}$ alone would be. Second, the left half of a reduced matrix is itself reduced, so it really is the RREF of $\mathbf{A}$. The first is the column view of a product; the second is a small check on the staircase shape. After that, $\mathbf{R} = \mathbf{EA}$ and $\mathbf{C} = \mathbf{EI}_n = \mathbf{E}$, and Theorem 3.4.5 finishes.

Proof. Let $\rho_1, \ldots, \rho_k$ be row operations taking $[\mathbf{A} \mid \mathbf{I}_n]$ to $[\mathbf{R} \mid \mathbf{C}]$, and put $\mathbf{E} = \mathbf{E}_k \cdots \mathbf{E}_1$ with $\mathbf{E}_i = \rho_i(\mathbf{I}_n)$ (and $\mathbf{E} = \mathbf{I}_n$ if $k = 0$). By Theorem 3.4.1, $[\mathbf{R} \mid \mathbf{C}] = \mathbf{E}\,[\mathbf{A} \mid \mathbf{I}_n]$. By Theorem 0.9.2 (columns), each column of $\mathbf{E}\,[\mathbf{A} \mid \mathbf{I}_n]$ is $\mathbf{E}$ times the corresponding column, so

$$\mathbf{R} = \mathbf{EA} \qquad \text{and} \qquad \mathbf{C} = \mathbf{EI}_n = \mathbf{E}.$$

Claim. $\mathbf{R}$ is in reduced row echelon form.

Proof. If row k of $\mathbf{R}$ is nonzero, its first nonzero entry is also the first nonzero entry of row k of $[\mathbf{R} \mid \mathbf{C}]$, so the pivots of the nonzero rows of $\mathbf{R}$ are pivots of $[\mathbf{R} \mid \mathbf{C}]$, in the same positions. Suppose row k of $\mathbf{R}$ is zero but some later row $l > k$ of $\mathbf{R}$ is nonzero. Row k of $[\mathbf{R} \mid \mathbf{C}]$ cannot be zero, since zero rows of $[\mathbf{R} \mid \mathbf{C}]$ come after its nonzero rows and row l is nonzero. So its pivot lies in a column $> n$, while the pivot of row l lies in a column $\leq n$, to the left:

this contradicts the staircase condition of Definition 3.2.3. Hence the zero rows of $\mathbf{R}$ come last, the pivots of $\mathbf{R}$ move strictly right, and each pivot of $\mathbf{R}$ is 1 and the only nonzero entry in its column, because this holds in $[\mathbf{R} \mid \mathbf{C}]$. So $\mathbf{R}$ satisfies Definition 3.2.4.

■

Now $\mathbf{R} = \mathbf{E}_k \cdots \mathbf{E}_1 \mathbf{A}$ is obtained from $\mathbf{A}$ by the operations $\rho_1, \ldots, \rho_k$ (Theorem 3.4.1), so $\mathbf{A}$ is row equivalent to $\mathbf{R}$, and $\mathbf{R}$ is in reduced form. By Theorem 3.3.4, $\mathbf{R}$ is the RREF of $\mathbf{A}$. By Theorem 3.4.4 ((a) $\Leftrightarrow$ (c)), $\mathbf{A}$ is invertible if and only if $\mathbf{R} = \mathbf{I}_n$. If $\mathbf{R} = \mathbf{I}_n$, then $\mathbf{EA} = \mathbf{I}_n$, and Theorem 3.4.5 gives $\mathbf{A}^{-1} = \mathbf{E} = \mathbf{C}$. This proves the theorem.

■

Remark. The early stop is justified the same way. If, after some operations, $[\mathbf{A} \mid \mathbf{I}_n]$ has become $[\mathbf{A}' \mid \mathbf{E}']$ with a zero row in $\mathbf{A}'$, then, exactly as in the proof, $\mathbf{E}'$ is the product of the elementary matrices of the operations used so far and $\mathbf{A}' = \mathbf{E}'\mathbf{A}$. So $\mathbf{E}'$ is invertible by Theorem 3.4.2 and Theorem 0.9.7 (part 3). A matrix with a zero row k is not invertible, since row k of $\mathbf{A}'\mathbf{X}$ is zero for every $\mathbf{X}$, so $\mathbf{A}'\mathbf{X} \neq \mathbf{I}_n$. If $\mathbf{A}$ were invertible, $\mathbf{A}' = \mathbf{E}'\mathbf{A}$ would be invertible by Theorem 0.9.7 (part 3). Hence $\mathbf{A}$ is not invertible.

Example. A 3×3 inverse

Find the inverse of $\mathbf{A} = \begin{pmatrix} 1 & 2 & 1 \\ 2 & 5 & 3 \\ 1 & 3 & 3 \end{pmatrix} \in M_3(\mathbb{R})$, or show that it does not exist.

Solution. Row reduce $[\mathbf{A} \mid \mathbf{I}_3]$, naming each operation.

$$\left[\begin{array}{ccc|ccc} 1 & 2 & 1 & 1 & 0 & 0 \\ 2 & 5 & 3 & 0 & 1 & 0 \\ 1 & 3 & 3 & 0 & 0 & 1 \end{array}\right] \xrightarrow[R_3 \to R_3 - R_1]{R_2 \to R_2 - 2R_1} \left[\begin{array}{ccc|ccc} 1 & 2 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & -2 & 1 & 0 \\ 0 & 1 & 2 & -1 & 0 & 1 \end{array}\right]$$

$$\xrightarrow{R_3 \to R_3 - R_2} \left[\begin{array}{ccc|ccc} 1 & 2 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & -2 & 1 & 0 \\ 0 & 0 & 1 & 1 & -1 & 1 \end{array}\right]$$

The left half now has three pivots, so no zero row will appear. Clear above the pivots, from the bottom up:

$$\xrightarrow[R_1 \to R_1 - R_3]{R_2 \to R_2 - R_3} \left[\begin{array}{ccc|ccc} 1 & 2 & 0 & 0 & 1 & -1 \\ 0 & 1 & 0 & -3 & 2 & -1 \\ 0 & 0 & 1 & 1 & -1 & 1 \end{array}\right] \xrightarrow{R_1 \to R_1 - 2R_2} \left[\begin{array}{ccc|ccc} 1 & 0 & 0 & 6 & -3 & 1 \\ 0 & 1 & 0 & -3 & 2 & -1 \\ 0 & 0 & 1 & 1 & -1 & 1 \end{array}\right].$$

The left half is $\mathbf{I}_3$, so by Theorem 3.4.6 $\mathbf{A}$ is invertible and

$$\mathbf{A}^{-1} = \begin{pmatrix} 6 & -3 & 1 \\ -3 & 2 & -1 \\ 1 & -1 & 1 \end{pmatrix}.$$

Check one product: row 1 of $\mathbf{A}$ times the columns of $\mathbf{A}^{-1}$ gives $6-6+1=1$, $-3+4-1=0$, $1-2+1=0$; rows 2 and 3 give $(0,1,0)$ and $(0,0,1)$ in the same way. So $\mathbf{A}\mathbf{A}^{-1}=\mathbf{I}_3$, and by Theorem 3.4.5 this single product is enough.

Example. When the algorithm stops

Decide whether $\mathbf{A}=\begin{pmatrix}1&2&3\\2&5&7\\1&3&4\end{pmatrix}\in M_3(\mathbb{R})$ is invertible.

Solution. Row reduce $[\mathbf{A}\mid\mathbf{I}_3]$:

$$\left[\begin{array}{ccc|ccc}1&2&3&1&0&0\\2&5&7&0&1&0\\1&3&4&0&0&1\end{array}\right]\xrightarrow[R_3\to R_3-R_1]{R_2\to R_2-2R_1}\left[\begin{array}{ccc|ccc}1&2&3&1&0&0\\0&1&1&-2&1&0\\0&1&1&-1&0&1\end{array}\right]$$

$$\xrightarrow{R_3\to R_3-R_2}\left[\begin{array}{ccc|ccc}1&2&3&1&0&0\\0&1&1&-2&1&0\\0&0&0&1&-1&1\end{array}\right].$$

A zero row has appeared in the left half, so by the remark after Theorem 3.4.6, $\mathbf{A}$ is not invertible. The right half of that row is not wasted: it records the combination of rows that produced zero. Row 3 of the left half is $1\cdot(\text{row }1)-1\cdot(\text{row }2)+1\cdot(\text{row }3)$ of $\mathbf{A}$, and indeed $(1,2,3)-(2,5,7)+(1,3,4)=(0,0,0)$.

Finally, the method writes an invertible matrix as a product of elementary matrices, as in Theorem 3.4.4 (d).

Example. Factoring into elementary matrices

Write $\mathbf{A}=\begin{pmatrix}2&1\\5&3\end{pmatrix}\in M_2(\mathbb{R})$ as a product of elementary matrices, and read off $\mathbf{A}^{-1}$.

Solution. Reduce $\mathbf{A}$ to $\mathbf{I}_2$, keeping integers:

$$\begin{pmatrix}2&1\\5&3\end{pmatrix}\xrightarrow{R_2\to R_2-2R_1}\begin{pmatrix}2&1\\1&1\end{pmatrix}\xrightarrow{R_1\to R_1-R_2}\begin{pmatrix}1&0\\1&1\end{pmatrix}\xrightarrow{R_2\to R_2-R_1}\begin{pmatrix}1&0\\0&1\end{pmatrix}.$$

The elementary matrices of these operations are $\mathbf{E}_1=\mathbf{I}_2-2\mathbf{E}_{21}$, $\mathbf{E}_2=\mathbf{I}_2-\mathbf{E}_{12}$, $\mathbf{E}_3=\mathbf{I}_2-\mathbf{E}_{21}$, and by Theorem 3.4.1, $\mathbf{E}_3\mathbf{E}_2\mathbf{E}_1\mathbf{A}=\mathbf{I}_2$. Multiplying on the left by $\mathbf{E}_3^{-1}$, then $\mathbf{E}_2^{-1}$, then $\mathbf{E}_1^{-1}$, and using Theorem 3.4.2,

$$\mathbf{A}=\mathbf{E}_1^{-1}\mathbf{E}_2^{-1}\mathbf{E}_3^{-1}=\begin{pmatrix}1&0\\2&1\end{pmatrix}\begin{pmatrix}1&1\\0&1\end{pmatrix}\begin{pmatrix}1&0\\1&1\end{pmatrix}.$$

Multiplying out confirms it: the last two factors give $\begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$, and then the first factor adds twice row 1 to row 2, giving $\begin{pmatrix} 2 & 1 \\ 5 & 3 \end{pmatrix}$. Also $\mathbf{E}_3\mathbf{E}_2\mathbf{E}_1$ is a left inverse of $\mathbf{A}$, hence $\mathbf{A}^{-1}$ by Theorem 3.4.5:

$$\mathbf{A}^{-1} = \mathbf{E}_3\mathbf{E}_2\mathbf{E}_1 = \begin{pmatrix} 3 & -1 \\ -5 & 2 \end{pmatrix},$$

which agrees with Theorem 0.9.8, since $ad - bc = 1$. The factorization is not unique: a different sequence of operations gives different factors.

3.4.5 Permutation matrices

A row swap is one elementary matrix. Doing several of them in a row is not, but the product is still a matrix of a single recognizable kind, and it is worth naming here because it turns up far beyond this chapter: it is how a relabeling of coordinates becomes a matrix.

Definition 3.4.2. Permutation matrix

Let $\sigma \in S_n$ be a permutation of $\{1, \dots, n\}$. The **permutation matrix** of σ is

$$\mathbf{P}_\sigma = \begin{pmatrix} \mathbf{e}_{\sigma(1)} & \mathbf{e}_{\sigma(2)} & \cdots & \mathbf{e}_{\sigma(n)} \end{pmatrix} \in M_n(F),$$

the matrix whose j-th column is $\mathbf{e}_{\sigma(j)}$. A matrix is a **permutation matrix** if it equals $\mathbf{P}_\sigma$ for some $\sigma \in S_n$.

Equivalently, a permutation matrix has exactly one 1 in each row and in each column, and zeros elsewhere. The identity is $\mathbf{P}_{\mathrm{id}}$, and $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ is $\mathbf{P}_\tau$ for the transposition $\tau = (1\ 2)$, which is also the elementary matrix $\mathbf{P}_{12}$. By contrast $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ is **not** a permutation matrix: each column is still nonzero, but its first row contains two 1's, and its second column $(1, 1)$ is not a standard basis vector.

Lemma 3.4.7. Permutation matrices

Let $\sigma, \tau \in S_n$.

(a) $\mathbf{P}_\sigma\mathbf{P}_\tau = \mathbf{P}_{\sigma\circ\tau}$. In particular a product of permutation matrices is a permutation matrix.

(b) $\mathbf{P}_\sigma$ is invertible and $\mathbf{P}_\sigma^{-1} = \mathbf{P}_\sigma^{\top}$.

(c) If $\tau = (r\ s)$ is a transposition, then $\mathbf{P}_\tau$ is the elementary matrix $\mathbf{P}_{rs}$ that swaps rows r and s, and $\mathbf{P}_\tau^2 = \mathbf{I}$.

Proof.

(a) By Theorem 0.9.3, $\mathbf{P}_\tau\mathbf{e}_j = \mathbf{e}_{\tau(j)}$ and so $\mathbf{P}_\sigma\mathbf{P}_\tau\mathbf{e}_j = \mathbf{P}_\sigma\mathbf{e}_{\tau(j)} = \mathbf{e}_{\sigma(\tau(j))}$. So the j-th column of $\mathbf{P}_\sigma\mathbf{P}_\tau$ is $\mathbf{e}_{(\sigma\circ\tau)(j)}$, for every j, which is the claim.

(b) By Theorem 0.9.2, $(\mathbf{P}_\sigma^{\top}\mathbf{P}_\sigma)_{ij} = \mathbf{e}_{\sigma(i)}^{\top}\mathbf{e}_{\sigma(j)}$, which is 1 if $\sigma(i) = \sigma(j)$ and 0 otherwise.

Since σ is injective, $\sigma(i) = \sigma(j)$ exactly when $i = j$. Hence $\mathbf{P}_\sigma^\top \mathbf{P}_\sigma = \mathbf{I}$, and by Theorem 3.4.5, $\mathbf{P}_\sigma$ is invertible with inverse $\mathbf{P}_\sigma^\top$.

(c) Swapping rows r and s of $\mathbf{I}$ produces a matrix whose column r is $\mathbf{e}_s$, whose column s is $\mathbf{e}_r$, and whose other columns j are $\mathbf{e}_j$. These are the columns $\mathbf{e}_{\tau(j)}$, so the swap matrix is $\mathbf{P}_\tau$. Since $\tau \circ \tau = \mathrm{id}$, part (a) gives $\mathbf{P}_\tau^2 = \mathbf{P}_{\mathrm{id}} = \mathbf{I}$. This proves the lemma.

■

By (c) and Theorem 3.4.1, $\mathbf{P}_\tau \mathbf{A}$ is $\mathbf{A}$ with rows r and s swapped, and by (a) every permutation matrix is a product of such swaps applied in turn, since for $n \geq 2$ every permutation is a product of transpositions (Theorem 0.10.4). So $\mathbf{P}_\sigma \mathbf{A}$ is always $\mathbf{A}$ with its rows rearranged.

3.4.6 Row equivalence, revisited

Two descriptions of "row equivalent" are now on the table. Example (Row equivalence) of Chapter 0 said $\mathbf{B} = \mathbf{EA}$ for some invertible $\mathbf{E}$. Definition 3.2.2 says $\mathbf{B}$ is reachable from $\mathbf{A}$ by finitely many row operations. Chapter 0 promised that they agree, and now it is one line each way.

Theorem 3.4.8. Row equivalence is left multiplication by an invertible matrix

Let $\mathbf{A}, \mathbf{B} \in M_{m\times n}(F)$. Then $\mathbf{A}$ is row equivalent to $\mathbf{B}$ (that is, $\mathbf{B}$ can be reached from $\mathbf{A}$ by row operations) if and only if $\mathbf{B} = \mathbf{EA}$ for some invertible $\mathbf{E} \in M_m(F)$.

Proof. ($\Rightarrow$) Suppose row operations $\rho_1, \dots, \rho_k$ take $\mathbf{A}$ to $\mathbf{B}$. By Theorem 3.4.1, $\mathbf{B} = \mathbf{E}_k \cdots \mathbf{E}_1 \mathbf{A}$ with $\mathbf{E}_i = \rho_i(\mathbf{I}_m)$, and $\mathbf{E} = \mathbf{E}_k \cdots \mathbf{E}_1$ is invertible by Theorem 3.4.2 and Theorem 0.9.7 (part 3). (If $k = 0$, take $\mathbf{E} = \mathbf{I}_m$.)

($\Leftarrow$) Suppose $\mathbf{B} = \mathbf{EA}$ with $\mathbf{E}$ invertible. By Theorem 3.4.4 ((a) $\Rightarrow$ (d)), $\mathbf{E} = \mathbf{E}_k \cdots \mathbf{E}_1$ for some elementary matrices $\mathbf{E}_i$. Let ρ_i be the operation with $\rho_i(\mathbf{I}_m) = \mathbf{E}_i$. By Theorem 3.4.1, performing ρ_1, then ρ_2, …, then ρ_k on $\mathbf{A}$ produces $\mathbf{E}_k \cdots \mathbf{E}_1 \mathbf{A} = \mathbf{B}$. Hence $\mathbf{A}$ is row equivalent to $\mathbf{B}$.

■

So the equivalence relation of Example (Row equivalence) is the one Gaussian elimination moves along, and "two matrices have the same RREF" is the same as "each is an invertible matrix times the other". In the next section this form, $\mathbf{B} = \mathbf{EA}$, is what makes row spaces and ranks easy to compare.

3.4.7 Exercises

3.4.7.1 A. Check your understanding

Exercise. A1

(a) Define an elementary matrix, and write down the 3×3 elementary matrix that performs $R_3 \to R_3 + 7R_1$.

(b) State four of the conditions in Theorem 3.4.4.

(c) True or false: the product of two elementary matrices is always an elementary matrix. Justify your answer.

(d) True or false: if $\mathbf{A} \in M_{2\times 3}(F)$, $\mathbf{B} \in M_{3\times 2}(F)$ and $\mathbf{AB} = \mathbf{I}_2$, then $\mathbf{BA} = \mathbf{I}_3$. Justify your answer.

(e) True or false: if $\mathbf{A} \in M_n(F)$ and $\mathbf{Ax} = \mathbf{e}_j$ has a solution for each $j = 1, \ldots, n$, then $\mathbf{A}$ is invertible. Justify your answer.

(f) Describe the method for computing the inverse of a square matrix, and how it detects a matrix that is not invertible.

Solution.

(a) An elementary matrix in $M_m(F)$ is one obtained from $\mathbf{I}_m$ by exactly one elementary row operation (Definition 3.4.1). Applying $R_3 \to R_3 + 7R_1$ to $\mathbf{I}_3$ gives $\mathbf{I}_3 + 7\mathbf{E}_{31} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 7 & 0 & 1 \end{pmatrix}$.

(b) For $\mathbf{A} \in M_n(F)$, each of the following is equivalent to $\mathbf{A}$ being invertible: $\mathbf{Ax} = \mathbf{0}$ has only the trivial solution; the RREF of $\mathbf{A}$ is $\mathbf{I}_n$; $\mathbf{A}$ is a product of elementary matrices; the columns of $\mathbf{A}$ form a basis of F^n.

(c) False. A single row operation changes at most two rows of $\mathbf{I}_3$ (a swap changes two, a scaling or an addition changes one), so every elementary 3×3 matrix has at least one row equal to the corresponding row of $\mathbf{I}_3$. But $\mathbf{P}_{12}\mathbf{P}_{23} = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$ (by the row view, its rows are rows 2, 1, 3 of $\mathbf{P}_{23}$) differs from $\mathbf{I}_3$ in every row. Hence it is a product of two elementary matrices that is not elementary.

(d) False. With $\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$ and $\mathbf{B} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix}$, $\mathbf{AB} = \mathbf{I}_2$ but $\mathbf{BA} = \operatorname{diag}(1, 1, 0) \neq \mathbf{I}_3$. Theorem 3.4.5 needs square matrices.

(e) True. Let $\mathbf{x}_j$ solve $\mathbf{Ax}_j = \mathbf{e}_j$, and let $\mathbf{B}$ have columns $\mathbf{x}_1, \ldots, \mathbf{x}_n$. By Theorem 0.9.2, $\mathbf{AB}$ has columns $\mathbf{e}_1, \ldots, \mathbf{e}_n$, so $\mathbf{AB} = \mathbf{I}_n$, and $\mathbf{A}$ is invertible by Theorem 3.4.5.

(f) Row reduce $[\mathbf{A} \mid \mathbf{I}_n]$ to reduced form $[\mathbf{R} \mid \mathbf{C}]$. If $\mathbf{R} = \mathbf{I}_n$, then $\mathbf{A}^{-1} = \mathbf{C}$; if $\mathbf{R} \neq \mathbf{I}_n$, in particular if a zero row appears in the left half at any stage, $\mathbf{A}$ is not invertible (Theorem 3.4.6).

3.4.7.2 B. Practice

Exercise. B1: Inverses by row reduction

For each matrix, find the inverse or show that it is not invertible.

(a) $\mathbf{A} = \begin{pmatrix} 1 & 0 & 2 \\ 2 & 1 & 3 \\ 4 & 1 & 8 \end{pmatrix} \in M_3(\mathbb{R})$.

(b) $\mathbf{B} = \begin{pmatrix} 1 & 2 & -1 \\ 0 & 1 & 1 \\ 2 & 5 & -1 \end{pmatrix} \in M_3(\mathbb{R})$.

(c) $\mathbf{C} = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix} \in M_3(\mathbb{F}_2)$.

Solution.

(a) Row reduce $[\mathbf{A} \mid \mathbf{I}_3]$:

$$\left[\begin{array}{ccc|ccc} 1 & 0 & 2 & 1 & 0 & 0 \\ 2 & 1 & 3 & 0 & 1 & 0 \\ 4 & 1 & 8 & 0 & 0 & 1 \end{array}\right] \xrightarrow[R_3 \to R_3 - 4R_1]{R_2 \to R_2 - 2R_1} \left[\begin{array}{ccc|ccc} 1 & 0 & 2 & 1 & 0 & 0 \\ 0 & 1 & -1 & -2 & 1 & 0 \\ 0 & 1 & 0 & -4 & 0 & 1 \end{array}\right]$$

$$\xrightarrow{R_3 \to R_3 - R_2} \left[\begin{array}{ccc|ccc} 1 & 0 & 2 & 1 & 0 & 0 \\ 0 & 1 & -1 & -2 & 1 & 0 \\ 0 & 0 & 1 & -2 & -1 & 1 \end{array}\right]$$

$$\xrightarrow[R_1 \to R_1 - 2R_3]{R_2 \to R_2 + R_3} \left[\begin{array}{ccc|ccc} 1 & 0 & 0 & 5 & 2 & -2 \\ 0 & 1 & 0 & -4 & 0 & 1 \\ 0 & 0 & 1 & -2 & -1 & 1 \end{array}\right].$$

The left half is $\mathbf{I}_3$, so by Theorem 3.4.6, $\mathbf{A}^{-1} = \begin{pmatrix} 5 & 2 & -2 \\ -4 & 0 & 1 \\ -2 & -1 & 1 \end{pmatrix}$. Check: row 1 of $\mathbf{A}$ against the columns of $\mathbf{A}^{-1}$ gives $5 - 4 = 1$, $2 - 2 = 0$, $-2 + 2 = 0$, and rows 2 and 3 give $(0, 1, 0)$ and $(0, 0, 1)$, so $\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}_3$, which suffices by Theorem 3.4.5.

(b) Row reduce $[\mathbf{B} \mid \mathbf{I}_3]$:

$$\left[\begin{array}{ccc|ccc} 1 & 2 & -1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 \\ 2 & 5 & -1 & 0 & 0 & 1 \end{array}\right] \xrightarrow{R_3 \to R_3 - 2R_1} \left[\begin{array}{ccc|ccc} 1 & 2 & -1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & -2 & 0 & 1 \end{array}\right]$$

$$\xrightarrow{R_3 \to R_3 - R_2} \left[\begin{array}{ccc|ccc} 1 & 2 & -1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & -2 & -1 & 1 \end{array}\right].$$

A zero row appears in the left half, so $\mathbf{B}$ is not invertible (remark after Theorem 3.4.6). The right half of that row says $-2(\text{row } 1) - (\text{row } 2) + (\text{row } 3) = \mathbf{0}$, that is, row 3 of $\mathbf{B}$ is $2(\text{row } 1) + (\text{row } 2)$.

(c) Over $\mathbb{F}_2$ we have $-1 = 1$, so subtracting a row is adding it. Row reduce $[\mathbf{C} \mid \mathbf{I}_3]$:

$$\left[\begin{array}{ccc|ccc} 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 & 1 \end{array}\right] \xrightarrow{R_3 \to R_3 + R_1} \left[\begin{array}{ccc|ccc} 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 \end{array}\right]$$

$$\xrightarrow{R_2 \to R_2 + R_3} \left[\begin{array}{ccc|ccc} 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 1 \end{array}\right]$$

$$\xrightarrow{R_1 \to R_1 + R_2} \left[\begin{array}{ccc|ccc} 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 1 \end{array}\right].$$

Hence $\mathbf{C}^{-1} = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 0 & 1 \end{pmatrix}$. Check in $\mathbb{F}_2$: row 1 of $\mathbf{C}$ is $(1, 1, 0)$, and against the columns of $\mathbf{C}^{-1}$ it gives $0 + 1 = 1$, $1 + 1 = 0$, $1 + 1 = 0$; rows 2 and 3 give $(0, 1, 0)$ and $(0, 0, 1)$ similarly.

Exercise. B2: A product of elementary matrices

Let $\mathbf{A} = \begin{pmatrix} 1 & 3 \\ 2 & 5 \end{pmatrix} \in M_2(\mathbb{R})$.

(a) Write $\mathbf{A}$ as a product of elementary matrices.

(b) Hence write $\mathbf{A}^{-1}$ as a product of elementary matrices, and compute it.

Solution.

(a) Reduce $\mathbf{A}$ to $\mathbf{I}_2$:

$$\begin{pmatrix} 1 & 3 \\ 2 & 5 \end{pmatrix} \xrightarrow{R_2 \to R_2 - 2R_1} \begin{pmatrix} 1 & 3 \\ 0 & -1 \end{pmatrix} \xrightarrow{R_2 \to -R_2} \begin{pmatrix} 1 & 3 \\ 0 & 1 \end{pmatrix} \xrightarrow{R_1 \to R_1 - 3R_2} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

With $\mathbf{E}_1 = \mathbf{I}_2 - 2\mathbf{E}_{21}$, $\mathbf{E}_2 = \mathbf{D}_2(-1)$, $\mathbf{E}_3 = \mathbf{I}_2 - 3\mathbf{E}_{12}$, Theorem 3.4.1 gives $\mathbf{E}_3\mathbf{E}_2\mathbf{E}_1\mathbf{A} = \mathbf{I}_2$. Multiplying on the left by $\mathbf{E}_3^{-1}, \mathbf{E}_2^{-1}, \mathbf{E}_1^{-1}$ in turn and using Theorem 3.4.2,

$$\mathbf{A} = \mathbf{E}_1^{-1}\mathbf{E}_2^{-1}\mathbf{E}_3^{-1} = \begin{pmatrix} 1 & 0 \\ 2 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} 1 & 3 \\ 0 & 1 \end{pmatrix}.$$

Check: the last two factors give $\begin{pmatrix} 1 & 3 \\ 0 & -1 \end{pmatrix}$, and adding twice row 1 to row 2 gives $\begin{pmatrix} 1 & 3 \\ 2 & 5 \end{pmatrix}$.

(b) Since $\mathbf{E}_3\mathbf{E}_2\mathbf{E}_1\mathbf{A} = \mathbf{I}_2$, Theorem 3.4.5 gives

$$\mathbf{A}^{-1} = \mathbf{E}_3\mathbf{E}_2\mathbf{E}_1 = \begin{pmatrix} 1 & -3 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -2 & 1 \end{pmatrix} = \begin{pmatrix} -5 & 3 \\ 2 & -1 \end{pmatrix}.$$

(The last two factors give $\begin{pmatrix} 1 & 0 \\ 2 & -1 \end{pmatrix}$, and $R_1 \to R_1 - 3R_2$ then gives the answer.) This agrees with Theorem 0.9.8, since $ad - bc = -1$.

Exercise. B3: Invertible products

Let $\mathbf{A}, \mathbf{B} \in M_n(F)$, and suppose $\mathbf{AB}$ is invertible. Prove that $\mathbf{A}$ and $\mathbf{B}$ are both invertible.

Solution. Let $\mathbf{C} = (\mathbf{AB})^{-1}$. Then $\mathbf{A}(\mathbf{BC}) = (\mathbf{AB})\mathbf{C} = \mathbf{I}_n$ by associativity, so $\mathbf{A}$ has the right inverse $\mathbf{BC}$, and by Theorem 3.4.5 $\mathbf{A}$ is invertible. Similarly $(\mathbf{CA})\mathbf{B} = \mathbf{C}(\mathbf{AB}) = \mathbf{I}_n$, so $\mathbf{B}$ has the left inverse $\mathbf{CA}$. By Theorem 3.4.5 with the roles of the two matrices swapped ($\mathbf{CA}$ times $\mathbf{B}$ is $\mathbf{I}_n$), $\mathbf{B}$ is invertible. This proves the claim.

3.4.7.3 C. Going deeper

Exercise. C1: Triangular matrices

Let $\mathbf{U} = (u_{ij}) \in M_n(F)$ be upper triangular (Definition 0.9.5).

(a) Prove that $\mathbf{U}$ is invertible if and only if $u_{ii} \neq 0$ for every i.

(b) Prove that if $\mathbf{U}$ is invertible, then $\mathbf{U}^{-1}$ is upper triangular.

Hint: for (a), use item (b) or item (f) of Theorem 3.4.4; for (b), consider the systems $\mathbf{Ux} = \mathbf{e}_k$.

Solution.

(a) ($\Leftarrow$) Suppose every $u_{ii} \neq 0$, and let $\mathbf{Ux} = \mathbf{0}$. Row i of this system reads $u_{ii}x_i + u_{i,i+1}x_{i+1} + \cdots + u_{in}x_n = 0$, since $u_{ij} = 0$ for $j < i$. We show $x_n = x_{n-1} = \cdots = x_1 = 0$ by downward induction. Row n is $u_{nn}x_n = 0$, and $u_{nn} \neq 0$, so $x_n = 0$. If $x_{i+1} = \cdots = x_n = 0$, row i becomes $u_{ii}x_i = 0$, so $x_i = 0$ since $u_{ii} \neq 0$. Hence $\mathbf{x} = \mathbf{0}$, and $\mathbf{U}$ is invertible by Theorem 3.4.4 ((b) $\Rightarrow$ (a)).
($\Rightarrow$) Suppose $u_{kk} = 0$ for some k. Every column $j \leq k$ of $\mathbf{U}$ has zero entries in rows $k, k+1, \ldots, n$: below the diagonal by triangularity, and in row k of column k by assumption. So the k columns $\mathbf{u}_1, \ldots, \mathbf{u}_k$ all lie in $\operatorname{span}(\mathbf{e}_1, \ldots, \mathbf{e}_{k-1})$, a subspace of dimension $k-1$ (its spanning list is a sub-list of the standard basis, hence independent). By Theorem 1.6.3 (a), a list of $k > k-1$ vectors there is linearly dependent. A nontrivial relation among $\mathbf{u}_1, \ldots, \mathbf{u}_k$, extended by zero coefficients on $\mathbf{u}_{k+1}, \ldots, \mathbf{u}_n$, is a nontrivial relation among all the columns. So the columns of $\mathbf{U}$ are dependent, and $\mathbf{U}$ is not invertible by Theorem 3.4.4 ((a) $\Rightarrow$ (f)). (For $k = 1$, the span is $\{\mathbf{0}\}$ and $\mathbf{u}_1 = \mathbf{0}$, so the same conclusion holds.)

(b) Suppose $\mathbf{U}$ is invertible, so every $u_{ii} \neq 0$ by (a). Fix k, and let $\mathbf{x}$ be the k-th column of $\mathbf{U}^{-1}$. By Theorem 0.9.2, $\mathbf{Ux}$ is the k-th column of $\mathbf{UU}^{-1} = \mathbf{I}_n$, so $\mathbf{Ux} = \mathbf{e}_k$. For $i > k$, row i of this system reads $u_{ii}x_i + \cdots + u_{in}x_n = 0$. By the same downward induction as in (a), running only over $i = n, n-1, \ldots, k+1$, we get $x_n = \cdots = x_{k+1} = 0$. So the k-th column of $\mathbf{U}^{-1}$ has zeros below position k. As k was arbitrary, $\mathbf{U}^{-1}$ is upper triangular.

Exercise. C2: All right inverses

Let $\mathbf{A} = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix} \in M_{2\times 3}(\mathbb{R})$.

(a) Find all $\mathbf{B} \in M_{3\times 2}(\mathbb{R})$ with $\mathbf{AB} = \mathbf{I}_2$.

(b) Show that there is no $\mathbf{C} \in M_{3\times 2}(\mathbb{R})$ with $\mathbf{CA} = \mathbf{I}_3$.

Solution.

(a) By Theorem 0.9.2, $\mathbf{AB} = \mathbf{I}_2$ means that the columns $\mathbf{b}_1, \mathbf{b}_2$ of $\mathbf{B}$ satisfy $\mathbf{Ab}_1 = \mathbf{e}_1$ and $\mathbf{Ab}_2 = \mathbf{e}_2$. The matrix $\mathbf{A}$ is already in reduced form with pivots in columns 1 and 2 and x_3 free, so $\mathbf{Ax} = \mathbf{0}$ means $x_1 = -x_3$, $x_2 = -x_3$, and its solutions are the multiples of $(-1,-1,1)$. One solution of $\mathbf{Ax} = \mathbf{e}_1$ is $(1,0,0)$, and one solution of $\mathbf{Ax} = \mathbf{e}_2$ is $(0,1,0)$. By Theorem 3.1.3,

$$\mathbf{b}_1 = (1,0,0) + s(-1,-1,1), \qquad \mathbf{b}_2 = (0,1,0) + t(-1,-1,1) \qquad (s,t \in \mathbb{R}).$$

Hence the right inverses of $\mathbf{A}$ are exactly

$$\mathbf{B} = \begin{pmatrix} 1-s & -t \\ -s & 1-t \\ s & t \end{pmatrix}, \qquad s,t \in \mathbb{R},$$

infinitely many of them.

(b) Suppose $\mathbf{CA} = \mathbf{I}_3$, and let $\mathbf{x} = (-1,-1,1)$, so $\mathbf{Ax} = \mathbf{0}$. Then $\mathbf{x} = \mathbf{I}_3\mathbf{x} = \mathbf{C}(\mathbf{Ax}) = \mathbf{C0} = \mathbf{0}$, a contradiction since $\mathbf{x} \neq \mathbf{0}$. Hence $\mathbf{A}$ has no left inverse.

Exercise. C3: Nilpotent perturbations of the identity

Let $\mathbf{A} \in M_n(F)$ and suppose $\mathbf{A}^k = \mathbf{O}$ for some integer $k \geq 1$.

(a) Prove that $\mathbf{I}_n + \mathbf{A}$ is invertible, by using item (b) of Theorem 3.4.4.

(b) Verify that $(\mathbf{I}_n + \mathbf{A})(\mathbf{I}_n - \mathbf{A} + \mathbf{A}^2 - \cdots + (-1)^{k-1}\mathbf{A}^{k-1}) = \mathbf{I}_n$, and deduce a formula for $(\mathbf{I}_n + \mathbf{A})^{-1}$.

Solution.

(a) Let $\mathbf{x} \in F^n$ with $(\mathbf{I}_n + \mathbf{A})\mathbf{x} = \mathbf{0}$. Then $\mathbf{Ax} = -\mathbf{x}$. Multiplying on the left by $\mathbf{A}$ and using this again, $\mathbf{A}^2\mathbf{x} = \mathbf{A}(-\mathbf{x}) = -\mathbf{Ax} = (-1)^2\mathbf{x}$, and inductively $\mathbf{A}^j\mathbf{x} = (-1)^j\mathbf{x}$ for every $j \geq 1$. Taking $j = k$, $\mathbf{0} = \mathbf{A}^k\mathbf{x} = (-1)^k\mathbf{x}$, so $\mathbf{x} = \mathbf{0}$ (multiply by $(-1)^k$). By Theorem 3.4.4 ((b) $\Rightarrow$ (a)), $\mathbf{I}_n + \mathbf{A}$ is invertible.

(b) Expanding with Theorem 0.9.4,

$$(\mathbf{I}_n + \mathbf{A})\sum_{j=0}^{k-1}(-1)^j\mathbf{A}^j = \sum_{j=0}^{k-1}(-1)^j\mathbf{A}^j + \sum_{j=0}^{k-1}(-1)^j\mathbf{A}^{j+1} = \mathbf{I}_n + (-1)^{k-1}\mathbf{A}^k = \mathbf{I}_n,$$

since the terms $(-1)^j\mathbf{A}^j$ for $1 \leq j \leq k-1$ in the first sum cancel against the terms

$(-1)^{j-1}\mathbf{A}^j$ of the second, and $\mathbf{A}^k = \mathbf{O}$. By Theorem 3.4.5, $(\mathbf{I}_n + \mathbf{A})^{-1} = \mathbf{I}_n - \mathbf{A} + \mathbf{A}^2 - \cdots + (-1)^{k-1}\mathbf{A}^{k-1}$.

3.5 Row Space, Column Space, Null Space, and Rank

We can now reduce any matrix and read off the solutions of any linear system. This section asks a coarser question: how much information does a matrix carry? Three subspaces come attached to every matrix $\mathbf{A}$: the vectors $\mathbf{Ax}$ it can produce, the combinations of its rows, and the vectors it sends to zero. Their dimensions are governed by a single number, the rank, which is the number of pivots. With it we settle when a system is solvable and how many solutions it has, and we turn the questions of Chapter 1 about vectors in F^m (independent? spanning? a basis of a sum or an intersection?) into a routine computation, which is the "systematic method" that Chapter 1 promised.

Throughout, F is a field and $\mathbf{A} \in M_{m\times n}(F)$ has columns $\mathbf{a}_1, \ldots, \mathbf{a}_n \in F^m$.

3.5.1 Three subspaces of a matrix

Here is the matrix we will follow through the section:

$$\mathbf{A} = \begin{pmatrix} 1 & 2 & 0 & 0 & -1 \\ 2 & 4 & 1 & 0 & 0 \\ 1 & 2 & 1 & 1 & 2 \\ 0 & 0 & 1 & 1 & 3 \end{pmatrix} \in M_{4\times 5}(\mathbb{R}).$$

Three questions about it have come up in this chapter, and each answer is a set. For which $\mathbf{b} \in \mathbb{R}^4$ can we solve $\mathbf{Ax} = \mathbf{b}$? By Theorem 3.1.1, exactly for $\mathbf{b}$ in the span of the columns. Which $\mathbf{x} \in \mathbb{R}^5$ solve $\mathbf{Ax} = \mathbf{0}$? By Theorem 3.1.2, a subspace of $\mathbb{R}^5$. And which new equations can we derive from the four by adding multiples of them? Their left-hand sides are exactly the combinations of the rows. These three sets recur so often that they get names.

The column space is everything $\mathbf{A}$ *can produce, the null space is everything* $\mathbf{A}$ *kills, and the row space is every left-hand side we can build from the equations of* $\mathbf{Ax} = \mathbf{b}$.

Definition 3.5.1. Column space

The **column space** of $\mathbf{A} \in M_{m\times n}(F)$ is the span of its columns,

$$\operatorname{col}(\mathbf{A}) \coloneqq \operatorname{span}(\mathbf{a}_1, \ldots, \mathbf{a}_n) = \{\mathbf{Ax} : \mathbf{x} \in F^n\} \subseteq F^m.$$

Definition 3.5.2. Row space

The **row space** of $\mathbf{A} \in M_{m\times n}(F)$ is the span of its rows,

$$\operatorname{row}(\mathbf{A}) \coloneqq \operatorname{span}(\text{row } 1, \ldots, \text{row } m) \subseteq M_{1\times n}(F).$$

Definition 3.5.3. Null space

The **null space** of $\mathbf{A} \in M_{m\times n}(F)$ is the solution set of the homogeneous system,

$$\operatorname{nul}(\mathbf{A}) := \{\mathbf{x} \in F^n : \mathbf{A}\mathbf{x} = \mathbf{0}\} \subseteq F^n.$$

In words: $\operatorname{col}(\mathbf{A})$ lives among the **outputs**, in F^m (one entry per row); $\operatorname{nul}(\mathbf{A})$ lives among the **inputs**, in F^n (one entry per column); and $\operatorname{row}(\mathbf{A})$ consists of row vectors of length n. The two descriptions of $\operatorname{col}(\mathbf{A})$ agree by Theorem 0.9.3: $\mathbf{A}\mathbf{x} = x_1\mathbf{a}_1 + \cdots + x_n\mathbf{a}_n$, and as $\mathbf{x}$ runs over F^n these are all the combinations of the columns.

Well-definedness. All three are subspaces. $\operatorname{col}(\mathbf{A})$ and $\operatorname{row}(\mathbf{A})$ are spans, hence subspaces by Theorem 1.3.1, of F^m and of $M_{1\times n}(F)$ respectively. $\operatorname{nul}(\mathbf{A})$ is a subspace of F^n by Theorem 3.1.2. The row vectors $M_{1\times n}(F)$ form an n-dimensional space, with the standard basis $\mathbf{e}_1^\top, \ldots, \mathbf{e}_n^\top$ (Example (Dimensions of the Running Examples)). Transposing turns row vectors into columns and respects sums and scalar multiples (Theorem 0.9.5), so it sends a combination of rows to the same combination of their transposes, and only the zero row transposes to $\mathbf{0}$. The transposes of the rows of $\mathbf{A}$ are the columns of $\mathbf{A}^\top$. Hence transposing sends $\operatorname{row}(\mathbf{A})$ onto $\operatorname{col}(\mathbf{A}^\top)$, a list of rows is independent exactly when the list of their transposes is, and transposing a basis of $\operatorname{row}(\mathbf{A})$ gives a basis of $\operatorname{col}(\mathbf{A}^\top)$. Therefore

$$\dim \operatorname{row}(\mathbf{A}) = \dim \operatorname{col}(\mathbf{A}^\top). \tag{$*$}$$

Examples.

- **The identity.** $\operatorname{col}(\mathbf{I}_n) = F^n$, since the columns $\mathbf{e}_1, \ldots, \mathbf{e}_n$ span; $\operatorname{row}(\mathbf{I}_n) = M_{1\times n}(F)$; and $\operatorname{nul}(\mathbf{I}_n) = \{\mathbf{0}\}$, since $\mathbf{I}_n\mathbf{x} = \mathbf{x}$.
- **The zero matrix** is the degenerate case. For $\mathbf{O} = \mathbf{O}_{m\times n}$, $\operatorname{col}(\mathbf{O}) = \operatorname{span}(\mathbf{0}, \ldots, \mathbf{0}) = \{\mathbf{0}\}$, $\operatorname{row}(\mathbf{O}) = \{\mathbf{0}\}$, and $\operatorname{nul}(\mathbf{O}) = F^n$, since every $\mathbf{x}$ is killed. It carries no information, and its column space is as small as possible while its null space is as large as possible. We will see that this trade-off is exact.
- **A single row.** For $\mathbf{A} = \begin{pmatrix} 1 & 1 & -1 \end{pmatrix} \in M_{1\times 3}(\mathbb{R})$, the columns are the scalars $1, 1, -1 \in \mathbb{R}^1$, so $\operatorname{col}(\mathbf{A}) = \mathbb{R}^1$, $\operatorname{row}(\mathbf{A}) = \operatorname{span}(\begin{pmatrix} 1 & 1 & -1 \end{pmatrix})$ is a line of row vectors, and $\operatorname{nul}(\mathbf{A})$ is the plane $x + y - z = 0$ in $\mathbb{R}^3$.

Nonexample by minimal change. Replace $\mathbf{0}$ by a nonzero right-hand side: the solution set $\{\mathbf{x} \in F^n : \mathbf{A}\mathbf{x} = \mathbf{b}\}$ with $\mathbf{b} \neq \mathbf{0}$. It is described by the same equations, and it is still of the form $\mathbf{p} + \operatorname{nul}(\mathbf{A})$ when nonempty (Theorem 3.1.3). But it does not contain $\mathbf{0}$, since $\mathbf{A}\mathbf{0} = \mathbf{0} \neq \mathbf{b}$, so it fails the first condition of Theorem 1.2.1.

Warning. The three spaces live in **different places**. For $\mathbf{A} \in M_{4\times 5}(\mathbb{R})$ above, $\operatorname{col}(\mathbf{A}) \subseteq \mathbb{R}^4$ but $\operatorname{nul}(\mathbf{A}) \subseteq \mathbb{R}^5$, so a statement such as "$\operatorname{nul}(\mathbf{A}) \subseteq \operatorname{col}(\mathbf{A})$" does not even make sense. Keep track by size: $\mathbf{A}\mathbf{x}$ needs $\mathbf{x}$ with n entries and returns a vector with m entries.

3.5.2 Rank and nullity

The column space measures how much $\mathbf{A}$ can reach, and the null space how much it collapses. We measure each by its dimension. Both are subspaces of finite-dimensional spaces, so both have a dimension by Theorem 1.6.7.

The rank counts the independent directions that $\mathbf{A}$ *can produce.*

Definition 3.5.4. Rank of a matrix

The **rank** of $\mathbf{A} \in M_{m\times n}(F)$ is $\operatorname{rank}\mathbf{A} := \dim\operatorname{col}(\mathbf{A})$.

Definition 3.5.5. Nullity of a matrix

The **nullity** of $\mathbf{A} \in M_{m\times n}(F)$ is $\operatorname{nullity}\mathbf{A} := \dim\operatorname{nul}(\mathbf{A})$.

From the examples above: $\operatorname{rank}\mathbf{I}_n = n$ and $\operatorname{nullity}\mathbf{I}_n = 0$; $\operatorname{rank}\mathbf{O}_{m\times n} = 0$ and $\operatorname{nullity}\mathbf{O}_{m\times n} = n$; and $\operatorname{rank}\begin{pmatrix}1 & 1 & -1\end{pmatrix} = 1$, with nullity 2. In each case rank plus nullity equals the number of columns. Two simple bounds hold at once: $\operatorname{rank}\mathbf{A} \le m$, because $\operatorname{col}(\mathbf{A}) \subseteq F^m$, and $\operatorname{rank}\mathbf{A} \le n$, because $\operatorname{col}(\mathbf{A})$ is spanned by n vectors (Theorem 1.6.3). Why define rank through columns and not rows? We could have, and one of the main theorems of this section says the answer would be the same.

3.5.3 What row operations preserve

To compute these spaces we will row reduce. So we first ask what a row operation does to each of them. Recall from Theorem 3.4.8 that $\mathbf{A}$ is row equivalent to $\mathbf{B}$ exactly when $\mathbf{B} = \mathbf{E}\mathbf{A}$ with $\mathbf{E} \in M_m(F)$ invertible; since row equivalence is symmetric (Theorem 3.2.2 (b)), we may simply say that $\mathbf{A}$ and $\mathbf{B}$ are row equivalent.

Theorem 3.5.1. Row operations and the three spaces

Let $\mathbf{A}, \mathbf{B} \in M_{m\times n}(F)$ be row equivalent. Then:

(a) $\operatorname{row}(\mathbf{B}) = \operatorname{row}(\mathbf{A})$;

(b) $\operatorname{nul}(\mathbf{B}) = \operatorname{nul}(\mathbf{A})$. Equivalently, writing $\mathbf{b}_1, \dots, \mathbf{b}_n$ for the columns of $\mathbf{B}$: for all $c_1, \dots, c_n \in F$, the columns of $\mathbf{A}$ satisfy $c_1\mathbf{a}_1 + \cdots + c_n\mathbf{a}_n = \mathbf{0}$ if and only if the columns of $\mathbf{B}$ satisfy $c_1\mathbf{b}_1 + \cdots + c_n\mathbf{b}_n = \mathbf{0}$;

(c) but $\operatorname{col}(\mathbf{B})$ and $\operatorname{col}(\mathbf{A})$ **need not** be equal.

Proof.

(a) Write $\mathbf{B} = \mathbf{E}\mathbf{A}$ with $\mathbf{E}$ invertible, by Theorem 3.4.8. By Theorem 0.9.2, row i of $\mathbf{E}\mathbf{A}$ is (row i of $\mathbf{E}$)$\mathbf{A}$, which by Theorem 0.9.3 (rows version) is a combination of the rows of $\mathbf{A}$. So every row of $\mathbf{B}$ lies in $\operatorname{row}(\mathbf{A})$, and $\operatorname{row}(\mathbf{B}) \subseteq \operatorname{row}(\mathbf{A})$ by Proposition 1.3.4 (1). Since $\mathbf{A} = \mathbf{E}^{-1}\mathbf{B}$ with $\mathbf{E}^{-1}$ invertible, the same argument gives $\operatorname{row}(\mathbf{A}) \subseteq \operatorname{row}(\mathbf{B})$.

(b) The equality $\operatorname{nul}(\mathbf{B}) = \operatorname{nul}(\mathbf{A})$ is Lemma 3.3.1. By Theorem 0.9.3, $c_1\mathbf{a}_1 + \cdots + c_n\mathbf{a}_n = \mathbf{A}\mathbf{c}$ and $c_1\mathbf{b}_1 + \cdots + c_n\mathbf{b}_n = \mathbf{B}\mathbf{c}$ with $\mathbf{c} = (c_1, \dots, c_n)$, so the second statement is the first one written out (this is Lemma 3.3.2).

(c) Let $\mathbf{A} = \begin{pmatrix}1 & 1\\ 1 & 1\end{pmatrix} \in M_2(\mathbb{R})$ and $\mathbf{B} = \begin{pmatrix}1 & 1\\ 0 & 0\end{pmatrix}$, obtained by $R_2 \to R_2 - R_1$. Then $\operatorname{col}(\mathbf{A}) = \operatorname{span}((1,1))$ and $\operatorname{col}(\mathbf{B}) = \operatorname{span}((1,0))$. These lines differ, since $(1,0)$ is not a multiple of $(1,1)$.

■

So row reduction keeps the row space and the **relations among the columns**, but it moves the columns themselves.

Warning. **Do not read** col(**A**) **off the reduced form.** In (c) above, the RREF of **A** is **B**, whose column space is the x-axis, while col(**A**) is the line $y = x$. What survives row reduction is the *pattern* of dependence: in both matrices column 2 equals column 1. So we use the RREF to decide **which** columns to take, and then take those columns **from A**.

3.5.4 Bases from the reduced form

Fix the notation for the rest of the section. Let $\mathbf{R} = (r_{kj})$ be the reduced row echelon form of $\mathbf{A}$ (Theorem 3.2.3, Theorem 3.3.4), with columns $\mathbf{r}_1, \dots, \mathbf{r}_n \in F^m$. Let $\mathbf{R}$ have r nonzero rows, and let the pivot of row k lie in column j_k, so that $j_1 < j_2 < \cdots < j_r$. These are the **pivot columns** of $\mathbf{A}$, well defined by Corollary 3.3.5. The remaining columns are the **free columns**. We use three facts from Definition 3.2.4:

- rows $r+1, \dots, m$ of $\mathbf{R}$ are zero;
- column j_k of $\mathbf{R}$ is $\mathbf{e}_k$ (the pivot is 1 and the only nonzero entry of its column);
- in row k, every entry to the left of the pivot is 0.

From the last two, if column j of $\mathbf{R}$ has exactly t pivot columns to its left, then **every entry of $\mathbf{r}_j$ in rows $t+1, \dots, m$ is zero, except the pivot entry of row $t+1$ when $j = j_{t+1}$.** Indeed, for $k \geq t+1$, row k is either zero or has its pivot in column $j_k \geq j$, and entries to the left of a pivot vanish. In the running example,

$$\mathbf{R} = \begin{pmatrix} 1 & 2 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 & 2 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

with $r = 3$, pivot columns $1, 3, 4$ and free columns $2, 5$ (the reduction is carried out in Example (Bases for the running example) below).

The column space comes first. In $\mathbf{R}$ the answer is visible: the pivot columns are $\mathbf{e}_1, \dots, \mathbf{e}_r$, and every column has zeros below row r. Row reduction does not preserve columns, but by Theorem 3.5.1 (b) it preserves every linear relation among them, and "is a basis" and "equals this combination" are statements about relations.

Theorem 3.5.2. Pivot columns form a basis of the column space

Let $\mathbf{A} \in M_{m\times n}(F)$ have pivot columns $j_1 < \cdots < j_r$, and let $\mathbf{R} = (r_{kj})$ be its RREF. Then:

(a) the pivot columns **of A**, $(\mathbf{a}_{j_1}, \dots, \mathbf{a}_{j_r})$, form a basis of col(**A**);

(b) if column j is free and t pivot columns lie to its left, then $\mathbf{a}_j = r_{1j}\mathbf{a}_{j_1} + \cdots + r_{tj}\mathbf{a}_{j_t}$;

(c) the pivot columns of $\mathbf{A}$ are exactly the vectors kept when $(\mathbf{a}_1, \dots, \mathbf{a}_n)$ is sifted as in Theorem 1.5.4.

Idea. Both claims are statements about linear relations among columns, so by Theorem 3.5.1 (b) it suffices to prove them for $\mathbf{R}$, where they can be seen by inspection. The pivot columns of $\mathbf{R}$ are $\mathbf{e}_1, \dots, \mathbf{e}_r$, which are independent. A free column of $\mathbf{R}$ is $\sum_k r_{kj}\mathbf{e}_k$, a combination of the pivot columns to its left with its own entries as coefficients. So we transfer two relations back to $\mathbf{A}$: "$\sum c_k \mathbf{a}_{j_k} = \mathbf{0}$ forces $c_1 = \cdots = c_r = 0$" and "$\mathbf{a}_j - \sum_k r_{kj}\mathbf{a}_{j_k} = \mathbf{0}$". Part (c) is a rereading of Corollary 3.3.5.

Proof. Throughout, we use Theorem 3.5.1 (b) in the form: for scalars $c_1, \dots, c_n$,

$$c_1\mathbf{a}_1 + \cdots + c_n\mathbf{a}_n = \mathbf{0} \iff c_1\mathbf{r}_1 + \cdots + c_n\mathbf{r}_n = \mathbf{0}. \tag{$\dagger$}$$

(b) Let column j be free, with t pivot columns $j_1, \dots, j_t$ to its left. By the fact in bold before the theorem, the entries of $\mathbf{r}_j$ in rows $t+1, \dots, m$ are zero, so $\mathbf{r}_j = r_{1j}\mathbf{e}_1 + \cdots + r_{tj}\mathbf{e}_t = r_{1j}\mathbf{r}_{j_1} + \cdots + r_{tj}\mathbf{r}_{j_t}$, since $\mathbf{r}_{j_k} = \mathbf{e}_k$. This is a linear relation among the columns of $\mathbf{R}$ with coefficient 1 on $\mathbf{r}_j$, $-r_{kj}$ on $\mathbf{r}_{j_k}$, and 0 elsewhere. By ($\dagger$), the columns of $\mathbf{A}$ satisfy the same relation, which is (b).

(c) *Spanning.* Each pivot column $\mathbf{a}_{j_k}$ lies in $\operatorname{span}(\mathbf{a}_{j_1}, \dots, \mathbf{a}_{j_r})$, and each free column does by (b). Hence $\operatorname{col}(\mathbf{A}) = \operatorname{span}(\mathbf{a}_1, \dots, \mathbf{a}_n) \subseteq \operatorname{span}(\mathbf{a}_{j_1}, \dots, \mathbf{a}_{j_r})$ by Proposition 1.3.4 (1). The reverse inclusion holds by Proposition 1.3.4 (2), since the pivot columns are among the columns.

Independence. Let $c_1\mathbf{a}_{j_1} + \cdots + c_r\mathbf{a}_{j_r} = \mathbf{0}$. This is a relation among all the columns of $\mathbf{A}$ with coefficient 0 on the free columns. By ($\dagger$), $c_1\mathbf{r}_{j_1} + \cdots + c_r\mathbf{r}_{j_r} = c_1\mathbf{e}_1 + \cdots + c_r\mathbf{e}_r = \mathbf{0}$. The k-th entry of the left side is c_k, so $c_1 = \cdots = c_r = 0$. Hence the pivot columns of $\mathbf{A}$ are linearly independent, and with spanning they form a basis of $\operatorname{col}(\mathbf{A})$ by Definition 1.5.1.

(c) By Corollary 3.3.5 (a), column j is a pivot column if and only if $\mathbf{a}_j \notin \operatorname{span}(\mathbf{a}_1, \dots, \mathbf{a}_{j-1})$. Sifting keeps $\mathbf{a}_j$ exactly when $\mathbf{a}_j$ is not in the span of the vectors kept before it, and by Theorem 1.5.4 (applied to $(\mathbf{a}_1, \dots, \mathbf{a}_{j-1})$) that span is $\operatorname{span}(\mathbf{a}_1, \dots, \mathbf{a}_{j-1})$. Hence the kept vectors are exactly the pivot columns. This proves the theorem.

■

This is sifting (Theorem 1.5.4) done by elimination. In Chapter 1, each step "is $\mathbf{v}_j$ in the span of the vectors kept so far?" was a separate small system. One row reduction answers all of them at once, and part (b) even gives the coefficients.

The row space and the null space are read off $\mathbf{R}$ directly.

Theorem 3.5.3. Basis of the row space

The nonzero rows of the RREF $\mathbf{R}$ of $\mathbf{A}$ form a basis of $\operatorname{row}(\mathbf{A})$.

Proof. By Theorem 3.5.1 (a), $\operatorname{row}(\mathbf{A}) = \operatorname{row}(\mathbf{R})$. The zero rows of $\mathbf{R}$ contribute nothing to a combination, so the r nonzero rows span $\operatorname{row}(\mathbf{R})$. For independence, let $c_1(\text{row } 1) + \cdots + c_r(\text{row } r) = \mathbf{0}$. Look at the entry in column j_k: since column j_k of $\mathbf{R}$ is $\mathbf{e}_k$, only row k has a nonzero entry there, equal to 1. So that entry of the combination is c_k, and $c_k = 0$ for each k. Hence the nonzero rows are a basis of $\operatorname{row}(\mathbf{A})$.

■

For the null space, each free variable gets one basis vector: set that free variable to 1, the other free variables to 0, and solve for the pivot variables.

Theorem 3.5.4. Basis of the null space

Let $\mathbf{A} \in M_{m\times n}(F)$ have RREF $\mathbf{R} = (r_{kj})$ with pivot columns $j_1 < \cdots < j_r$. For each free column f, define $\mathbf{s}_f \in F^n$ by

$$\begin{aligned}(\mathbf{s}_f)_f &= 1,\\ (\mathbf{s}_f)_{f'} &= 0 \text{ for every other free } f',\\ (\mathbf{s}_f)_{j_k} &= -r_{kf} \text{ for } k = 1,\dots,r.\end{aligned}$$

Then the vectors $\mathbf{s}_f$, one for each free column f, form a basis of $\operatorname{nul}(\mathbf{A})$. In particular $\operatorname{nullity}\mathbf{A} = n - r$, the number of free columns.

Proof. By Theorem 3.5.1 (b), $\operatorname{nul}(\mathbf{A}) = \operatorname{nul}(\mathbf{R})$. For $k \le r$, row k of $\mathbf{R}$ has entry 1 in column j_k, entry 0 in every other pivot column (since column $j_{k'}$ is $\mathbf{e}_{k'}$), and entries r_{kf} in the free columns. Rows $k > r$ are zero. Hence $\mathbf{Rx} = \mathbf{0}$ is equivalent to the r equations

$$x_{j_k} = -\sum_{f \text{ free}} r_{kf}\, x_f \qquad (k = 1,\dots,r). \tag{‡}$$

Spanning. Let $\mathbf{x} \in \operatorname{nul}(\mathbf{A})$, and put $\mathbf{y} = \sum_{f \text{ free}} x_f \mathbf{s}_f$. For a free column f, the f-th entry of $\mathbf{y}$ is x_f, since $\mathbf{s}_f$ has a 1 there and the other $\mathbf{s}_{f'}$ have 0. For a pivot column, the j_k-th entry of $\mathbf{y}$ is $\sum_f x_f(-r_{kf})$, which equals x_{j_k} by (‡). So $\mathbf{x} = \mathbf{y}$ lies in the span of the $\mathbf{s}_f$. Each $\mathbf{s}_f$ itself satisfies (‡), so it lies in $\operatorname{nul}(\mathbf{A})$.

Independence. If $\sum_f c_f \mathbf{s}_f = \mathbf{0}$, then the f-th entry of the left side is c_f, so every $c_f = 0$.

Hence the $\mathbf{s}_f$ form a basis of $\operatorname{nul}(\mathbf{A})$, and there are $n - r$ of them.

■

Example. Bases for the running example

Find bases of $\operatorname{col}(\mathbf{A})$, $\operatorname{row}(\mathbf{A})$, and $\operatorname{nul}(\mathbf{A})$, and the rank and nullity, for

$$\mathbf{A} = \begin{pmatrix} 1 & 2 & 0 & 0 & -1 \\ 2 & 4 & 1 & 0 & 0 \\ 1 & 2 & 1 & 1 & 2 \\ 0 & 0 & 1 & 1 & 3 \end{pmatrix} \in M_{4\times 5}(\mathbb{R}).$$

Solution. Row reduce:

$$\mathbf{A} \xrightarrow[R_3 \to R_3 - R_1]{R_2 \to R_2 - 2R_1} \begin{pmatrix} 1 & 2 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 & 2 \\ 0 & 0 & 1 & 1 & 3 \\ 0 & 0 & 1 & 1 & 3 \end{pmatrix} \xrightarrow[R_4 \to R_4 - R_2]{R_3 \to R_3 - R_2} \begin{pmatrix} 1 & 2 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 & 2 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix}$$

$$\xrightarrow{R_4 \to R_4 - R_3} \mathbf{R} = \begin{pmatrix} 1 & 2 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 & 2 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

The pivot columns are $1, 3, 4$, and the free columns are $2, 5$.

Column space. By Theorem 3.5.2 (a), the pivot columns **of A**,

$$\mathbf{a}_1 = (1, 2, 1, 0), \qquad \mathbf{a}_3 = (0, 1, 1, 1), \qquad \mathbf{a}_4 = (0, 0, 1, 1),$$

form a basis of $\operatorname{col}(\mathbf{A})$, so $\operatorname{rank} \mathbf{A} = 3$. By part (b), the free columns are $\mathbf{a}_2 = 2\mathbf{a}_1$ (from column 2 of $\mathbf{R}$) and $\mathbf{a}_5 = -\mathbf{a}_1 + 2\mathbf{a}_3 + \mathbf{a}_4$ (from column 5). Check: $-(1, 2, 1, 0) + 2(0, 1, 1, 1) + (0, 0, 1, 1) = (-1, 0, 2, 3) = \mathbf{a}_5$. Note that $\operatorname{col}(\mathbf{A}) \neq \operatorname{col}(\mathbf{R})$: every column of $\mathbf{R}$ has last entry 0, but $\mathbf{a}_3$ does not.

Row space. By Theorem 3.5.3, a basis is

$$\begin{pmatrix} 1 & 2 & 0 & 0 & -1 \end{pmatrix}, \quad \begin{pmatrix} 0 & 0 & 1 & 0 & 2 \end{pmatrix}, \quad \begin{pmatrix} 0 & 0 & 0 & 1 & 1 \end{pmatrix},$$

so $\dim \operatorname{row}(\mathbf{A}) = 3$.

Null space. The equations (‡) read $x_1 = -2x_2 + x_5$, $x_3 = -2x_5$, $x_4 = -x_5$. By Theorem 3.5.4, a basis is

$$\mathbf{s}_2 = (-2, 1, 0, 0, 0), \qquad \mathbf{s}_5 = (1, 0, -2, -1, 1),$$

so $\operatorname{nullity} \mathbf{A} = 2$. Check: $\mathbf{A}\mathbf{s}_2 = -2\mathbf{a}_1 + \mathbf{a}_2 = \mathbf{0}$ and $\mathbf{A}\mathbf{s}_5 = \mathbf{a}_1 - 2\mathbf{a}_3 - \mathbf{a}_4 + \mathbf{a}_5 = \mathbf{0}$, which are the two relations found above.

So $\operatorname{rank} \mathbf{A} = \dim \operatorname{row}(\mathbf{A}) = 3$ and $\operatorname{rank} \mathbf{A} + \operatorname{nullity} \mathbf{A} = 5$, the number of columns.

Quick check. Without computing anything, what is the largest possible rank of a 3×5 matrix, and the smallest possible nullity?

Solution. $\operatorname{rank} \mathbf{A} = \dim \operatorname{col}(\mathbf{A}) \leq 3$, since $\operatorname{col}(\mathbf{A}) \subseteq F^3$. By Theorem 3.5.4, the nullity is the number of free columns, $5 - r$, where $r \leq 3$ is the number of pivots (one per nonzero row). So the nullity is at least 2. For example $\begin{pmatrix} \mathbf{I}_3 & 0_{3\times 2} \end{pmatrix}$ attains rank 3 and nullity 2.

3.5.5 Row rank equals column rank

The example came out with $\dim \operatorname{row}(\mathbf{A}) = \dim \operatorname{col}(\mathbf{A})$. This is not luck, and it is surprising: the row space and the column space live in different spaces, and the rows and columns of a 4×5 matrix have nothing obvious in common. Counting pivots explains it.

Theorem 3.5.5. Row rank equals column rank

Let $\mathbf{A} \in M_{m\times n}(F)$, and let r be the number of pivots of its RREF. Then

$$\dim \operatorname{row}(\mathbf{A}) = \dim \operatorname{col}(\mathbf{A}) = r.$$

Consequently $\operatorname{rank} \mathbf{A} = \operatorname{rank} \mathbf{A}^\top$.

Proof. By Theorem 3.5.2 (a), $\operatorname{col}(\mathbf{A})$ has a basis consisting of the r pivot columns, so $\dim \operatorname{col}(\mathbf{A}) = r$. By Theorem 3.5.3, $\operatorname{row}(\mathbf{A})$ has a basis consisting of the nonzero rows of the RREF. Each nonzero row contains exactly one pivot and each pivot lies in a nonzero row, so there are r of them, and $\dim \operatorname{row}(\mathbf{A}) = r$. Finally, by $(*)$ at the start of this section, $\operatorname{rank} \mathbf{A}^\top = \dim \operatorname{col}(\mathbf{A}^\top) = \dim \operatorname{row}(\mathbf{A}) = r = \operatorname{rank} \mathbf{A}$. ■

The theorem makes the rank independent of **which** echelon form we compute. Gaussian elimination without the final clearing-up phase produces an echelon form that is not reduced, and different sequences of operations produce different echelon forms. They all give the same count.

Corollary 3.5.6. Rank from any echelon form

Let $\mathbf{A} \in M_{m\times n}(F)$, and let $\mathbf{G}$ be a matrix in row echelon form that is row equivalent to $\mathbf{A}$. Then the number of nonzero rows of $\mathbf{G}$ is $\operatorname{rank} \mathbf{A}$, and the columns of $\mathbf{G}$ containing its pivots are exactly the pivot columns of $\mathbf{A}$.

Proof. By Corollary 3.3.5 (b), the pivot columns of $\mathbf{G}$ are the pivot columns of $\mathbf{A}$. Each nonzero row of $\mathbf{G}$ has exactly one pivot, in a different column (Definition 3.2.3), so the number of nonzero rows of $\mathbf{G}$ equals the number of pivot columns of $\mathbf{A}$, which is $\operatorname{rank} \mathbf{A}$ by Theorem 3.5.5. ■

Warning. Rank is a property of $\mathbf{A}$, not of a computation, but it must be counted from a row echelon form. The number of nonzero rows of $\mathbf{A}$ itself is not the rank: $\begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix}$ has two nonzero rows and rank 1. Count nonzero rows only after reaching an echelon form. And the rank depends on the field when the entries are read in different fields: $\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$ has rank 2 over $\mathbb{R}$, but over $\mathbb{F}_2$, where $-1 = 1$, both rows are $\begin{pmatrix} 1 & 1 \end{pmatrix}$ and the rank is 1.

3.5.6 Rank–nullity and the Rouché–Capelli theorem

The trade-off between the zero matrix and the identity was exact, and now we can say why.

Theorem 3.5.7. Rank–Nullity for Matrices

For every $\mathbf{A} \in M_{m\times n}(F)$,

$$\operatorname{rank}\mathbf{A} + \operatorname{nullity}\mathbf{A} = n,$$

the number of **columns** of $\mathbf{A}$.

Proof. Let r be the number of pivot columns of $\mathbf{A}$. By Theorem 3.5.5, $\operatorname{rank}\mathbf{A} = r$. By Theorem 3.5.4, $\operatorname{nullity}\mathbf{A} = n - r$. Adding gives n. ∎

Each column of $\mathbf{A}$ is either a pivot column, contributing a direction to $\operatorname{col}(\mathbf{A})$, or a free column, contributing a direction to $\operatorname{nul}(\mathbf{A})$. Nothing is lost and nothing is counted twice. This is the same equation as the Rank–Nullity Theorem of Chapter 2, which proved it for linear maps by counting basis vectors, with no matrices at all.

Now the question this chapter started with. When does $\mathbf{Ax} = \mathbf{b}$ have a solution, and how many solutions are there? The first answer was $\mathbf{b} \in \operatorname{col}(\mathbf{A})$ (Theorem 3.1.1); here is a version that is a pure count.

Theorem 3.5.8. Rouché–Capelli Theorem

Let $\mathbf{A} \in M_{m\times n}(F)$, $\mathbf{b} \in F^m$, and let $[\mathbf{A} \mid \mathbf{b}] \in M_{m\times(n+1)}(F)$ be the augmented matrix.

(a) The system $\mathbf{Ax} = \mathbf{b}$ is consistent if and only if $\operatorname{rank}\mathbf{A} = \operatorname{rank}[\mathbf{A} \mid \mathbf{b}]$.

(b) Suppose it is consistent, let $\mathbf{p}$ be one solution, let $r = \operatorname{rank}\mathbf{A}$, and let $\mathbf{s}_1, \dots, \mathbf{s}_{n-r}$ be a basis of $\operatorname{nul}(\mathbf{A})$. Then every solution can be written as

$$\mathbf{x} = \mathbf{p} + t_1\mathbf{s}_1 + \cdots + t_{n-r}\mathbf{s}_{n-r}$$

for **exactly one** choice of $(t_1, \dots, t_{n-r}) \in F^{n-r}$. In words: the solutions form an $(n-r)$-parameter family.

(c) If moreover F is a finite field with q elements (for instance $\mathbb{F}_p$, with $q = p$), a consistent system has exactly $q^{\,n-\operatorname{rank}\mathbf{A}}$ solutions.

Idea. For (a), adding the column $\mathbf{b}$ can only enlarge the column space, and the dimension goes up exactly when $\mathbf{b}$ is new, that is, when $\mathbf{b} \notin \operatorname{col}(\mathbf{A})$. For (b), the structure theorem says solutions are $\mathbf{p}$ plus null vectors, and null vectors have unique coordinates in a basis. For (c), unique coordinates mean the solutions are in bijection with F^{n-r}, which has q^{n-r} elements.

Proof.

(a) By Definition 3.5.1, $\operatorname{col}([\mathbf{A} \mid \mathbf{b}]) = \operatorname{span}(\mathbf{a}_1, \ldots, \mathbf{a}_n, \mathbf{b})$, which contains $\operatorname{col}(\mathbf{A})$ by Proposition 1.3.4 (2).
($\Rightarrow$) Suppose the system is consistent. By Theorem 3.1.1, $\mathbf{b} \in \operatorname{span}(\mathbf{a}_1, \ldots, \mathbf{a}_n)$, so by Theorem 1.3.5 $\operatorname{col}([\mathbf{A} \mid \mathbf{b}]) = \operatorname{col}(\mathbf{A})$, and the ranks are equal.
($\Leftarrow$) Suppose $\operatorname{rank} \mathbf{A} = \operatorname{rank} [\mathbf{A} \mid \mathbf{b}]$. Then $\operatorname{col}(\mathbf{A})$ is a subspace of $\operatorname{col}([\mathbf{A} \mid \mathbf{b}])$ of the same dimension, so $\operatorname{col}(\mathbf{A}) = \operatorname{col}([\mathbf{A} \mid \mathbf{b}])$ by Theorem 1.6.8. In particular $\mathbf{b} \in \operatorname{col}(\mathbf{A})$, and the system is consistent by Theorem 3.1.1.

(b) By Theorem 3.5.7, $\operatorname{nullity} \mathbf{A} = n - r$, so a basis $\mathbf{s}_1, \ldots, \mathbf{s}_{n-r}$ of $\operatorname{nul}(\mathbf{A})$ exists (Theorem 3.5.4). By Theorem 3.1.3, $\mathbf{x}$ is a solution if and only if $\mathbf{x} - \mathbf{p} \in \operatorname{nul}(\mathbf{A})$. By Theorem 1.5.1, applied in the vector space $\operatorname{nul}(\mathbf{A})$ with this basis, each $\mathbf{x} - \mathbf{p} \in \operatorname{nul}(\mathbf{A})$ equals $t_1\mathbf{s}_1 + \cdots + t_{n-r}\mathbf{s}_{n-r}$ for exactly one $(t_1, \ldots, t_{n-r})$. Conversely every such expression is in $\operatorname{nul}(\mathbf{A})$, so $\mathbf{p}$ plus it is a solution. This proves (b).

(c) By (b), the rule $(t_1, \ldots, t_{n-r}) \mapsto \mathbf{p} + t_1\mathbf{s}_1 + \cdots + t_{n-r}\mathbf{s}_{n-r}$ is a bijection from F^{n-r} onto the solution set: every solution is hit, and by exactly one tuple. A tuple in F^{n-r} is a choice of $n - r$ entries, each one of q elements, so F^{n-r} has q^{n-r} elements. (When $r = n$ there is exactly one tuple, the empty one, and the basis of $\operatorname{nul}(\mathbf{A})$ is empty, so $\mathbf{p}$ is the only solution and $q^0 = 1$.) Hence the solution set has q^{n-r} elements.

■

Over an infinite field such as $\mathbb{Q}$, $\mathbb{R}$ or $\mathbb{C}$, part (b) gives the familiar trichotomy: no solution, exactly one solution (when $\operatorname{rank} \mathbf{A} = n$), or infinitely many (when $\operatorname{rank} \mathbf{A} < n$, since distinct values of t_1 give distinct solutions). Over a finite field "infinitely many" is replaced by an exact power of q. For example, over $\mathbb{F}_2$ a system can never have exactly 3 or 6 solutions.

Example. Counting solutions over $\mathbb{F}_3$

Over $\mathbb{F}_3$, consider

$$\mathbf{A} = \begin{pmatrix} 1 & 1 & 0 & 2 \\ 1 & 2 & 1 & 0 \\ 2 & 0 & 1 & 2 \end{pmatrix}, \qquad \mathbf{b} = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \qquad \mathbf{b}' = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}.$$

Decide whether $\mathbf{A}\mathbf{x} = \mathbf{b}$ and $\mathbf{A}\mathbf{x} = \mathbf{b}'$ are consistent, and count their solutions.

Solution. We reduce both augmented matrices at once, carrying two right-hand columns; arithmetic is modulo 3, so $-1 = 2$ and $-2 = 1$.

$$\left[\begin{array}{cccc|c|c} 1 & 1 & 0 & 2 & 1 & 1 \\ 1 & 2 & 1 & 0 & 0 & 0 \\ 2 & 0 & 1 & 2 & 1 & 0 \end{array}\right] \xrightarrow[R_3 \to R_3 - 2R_1]{R_2 \to R_2 - R_1} \left[\begin{array}{cccc|c|c} 1 & 1 & 0 & 2 & 1 & 1 \\ 0 & 1 & 1 & 1 & 2 & 2 \\ 0 & 1 & 1 & 1 & 2 & 1 \end{array}\right]$$

$$\xrightarrow[R_1 \to R_1 - R_2]{R_3 \to R_3 - R_2} \left[\begin{array}{cccc|c|c} 1 & 0 & 2 & 1 & 2 & 2 \\ 0 & 1 & 1 & 1 & 2 & 2 \\ 0 & 0 & 0 & 0 & 0 & 2 \end{array}\right].$$

Row operations act on each column separately, so deleting either right-hand column gives a reduction of the corresponding augmented matrix. For instance, in row 2, $(1, 2, 1, 0) -$

$(1,1,0,2) = (0,1,1,-2) = (0,1,1,1)$, and in row 3, $(2,0,1,2) - 2(1,1,0,2) = (0,-2,1,-2) = (0,1,1,1)$.

The left block is an echelon form of $\mathbf{A}$ with two nonzero rows, so $\operatorname{rank}\mathbf{A} = 2$ by Corollary 3.5.6. With $\mathbf{b}$, the augmented matrix $[\mathbf{A} \mid \mathbf{b}]$ has the same two nonzero rows (the last is zero), so $\operatorname{rank}[\mathbf{A} \mid \mathbf{b}] = 2 = \operatorname{rank}\mathbf{A}$, and by Theorem 3.5.8 (a) the system is consistent. By part (c) it has exactly $3^{4-2} = 9$ solutions. Explicitly, the reduced system is $x_1 = 2 - 2x_3 - x_4 = 2 + x_3 + 2x_4$ and $x_2 = 2 - x_3 - x_4 = 2 + 2x_3 + 2x_4$, with $x_3, x_4 \in \mathbb{F}_3$ free: nine choices.

With $\mathbf{b}'$, the last row $(0,0,0,0 \mid 2)$ is a nonzero row with its pivot in the last column, so $[\mathbf{A} \mid \mathbf{b}']$ has three nonzero rows in echelon form and $\operatorname{rank}[\mathbf{A} \mid \mathbf{b}'] = 3 \neq 2$. By Theorem 3.5.8 (a), $\mathbf{Ax} = \mathbf{b}'$ has no solution. Indeed, row 3 of $\mathbf{A}$ is the sum of rows 1 and 2, so any solution would need $b'_3 = b'_1 + b'_2 = 1$, but $b'_3 = 0$.

3.5.7 Rank of a product

Multiplying matrices cannot create information: every column of $\mathbf{AB}$ is a combination of the columns of $\mathbf{A}$, and every row is a combination of the rows of $\mathbf{B}$. The rank records this.

Theorem 3.5.9. Rank of a product

Let $\mathbf{A} \in M_{m\times n}(F)$ and $\mathbf{B} \in M_{n\times p}(F)$. Then

$$\operatorname{rank}(\mathbf{AB}) \leq \min(\operatorname{rank}\mathbf{A}, \operatorname{rank}\mathbf{B}).$$

Moreover, if $\mathbf{P} \in M_m(F)$ and $\mathbf{Q} \in M_n(F)$ are invertible, then $\operatorname{rank}(\mathbf{PA}) = \operatorname{rank}\mathbf{A} = \operatorname{rank}(\mathbf{AQ})$.

Proof. Let $\mathbf{b}_1, \ldots, \mathbf{b}_p \in F^n$ be the columns of $\mathbf{B}$. By Theorem 0.9.2, the columns of $\mathbf{AB}$ are $\mathbf{Ab}_1, \ldots, \mathbf{Ab}_p$, which lie in $\operatorname{col}(\mathbf{A})$ by Definition 3.5.1. By Proposition 1.3.4 (1), $\operatorname{col}(\mathbf{AB}) \subseteq \operatorname{col}(\mathbf{A})$, so $\operatorname{rank}(\mathbf{AB}) \leq \operatorname{rank}\mathbf{A}$ by Theorem 1.6.7. Similarly, row i of $\mathbf{AB}$ is (row i of $\mathbf{A}$)$\mathbf{B}$, a combination of the rows of $\mathbf{B}$ by Theorem 0.9.3 (rows version), so $\operatorname{row}(\mathbf{AB}) \subseteq \operatorname{row}(\mathbf{B})$. By Theorem 3.5.5, $\operatorname{rank}(\mathbf{AB}) = \dim\operatorname{row}(\mathbf{AB}) \leq \dim\operatorname{row}(\mathbf{B}) = \operatorname{rank}\mathbf{B}$.

For the second statement, apply the inequality twice: $\operatorname{rank}\mathbf{A} = \operatorname{rank}(\mathbf{P}^{-1}(\mathbf{PA})) \leq \operatorname{rank}(\mathbf{PA}) \leq \operatorname{rank}\mathbf{A}$, so equality holds throughout. The same sandwich with $\mathbf{A} = (\mathbf{AQ})\mathbf{Q}^{-1}$ gives $\operatorname{rank}(\mathbf{AQ}) = \operatorname{rank}\mathbf{A}$.

■

The equality case recovers something we already knew in a new form: row operations (left multiplication by an invertible $\mathbf{P}$) do not change the rank. The inequality is often strict. For $\mathbf{A} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$, both factors of $\mathbf{A} \cdot \mathbf{A}$ have rank 1, but $\mathbf{A}^2 = \mathbf{O}$ has rank 0.

3.5.8 A toolkit for questions from Chapter 1

In Chapter 1 we asked, again and again, whether some vectors in F^m are independent, whether they span, how to extend them to a basis, and what the sum and intersection of two subspaces

are. Each was answered by solving a small system by hand. Everything now reduces to one move: **put the vectors as columns of a matrix, row reduce, and read off pivots**.

Theorem 3.5.10. Independence and spanning by rank

Let $\mathbf{v}_1, \dots, \mathbf{v}_k \in F^m$, and let $\mathbf{A} = \begin{pmatrix} \mathbf{v}_1 & \cdots & \mathbf{v}_k \end{pmatrix} \in M_{m\times k}(F)$. Then:

(a) $(\mathbf{v}_1, \dots, \mathbf{v}_k)$ is linearly independent if and only if $\operatorname{rank}\mathbf{A} = k$, that is, every column is a pivot column;

(b) $(\mathbf{v}_1, \dots, \mathbf{v}_k)$ spans F^m if and only if $\operatorname{rank}\mathbf{A} = m$, that is, every row of the RREF has a pivot;

(c) $(\mathbf{v}_1, \dots, \mathbf{v}_k)$ is a basis of F^m if and only if $k = m = \operatorname{rank}\mathbf{A}$.

Proof.

(a) By Theorem 0.9.3, $c_1\mathbf{v}_1 + \cdots + c_k\mathbf{v}_k = \mathbf{A}\mathbf{c}$ with $\mathbf{c} = (c_1, \dots, c_k)$, so the list is independent if and only if $\operatorname{nul}(\mathbf{A}) = \{\mathbf{0}\}$, that is, $\operatorname{nullity}\mathbf{A} = 0$. By Theorem 3.5.7 this means $\operatorname{rank}\mathbf{A} = k$.

(b) The list spans F^m if and only if $\operatorname{col}(\mathbf{A}) = F^m$. Since $\operatorname{col}(\mathbf{A}) \subseteq F^m$ and $\dim F^m = m$, this holds if and only if $\operatorname{rank}\mathbf{A} = m$, by Theorem 1.6.8 for one direction and because equal spaces have equal dimension for the other. The number of pivots is $\operatorname{rank}\mathbf{A}$ by Theorem 3.5.5, and the RREF has m rows, each with at most one pivot.

(c) Combine (a) and (b) with Definition 1.5.1.

■

Example. Testing independence

Decide whether $\mathbf{v}_1 = (1, 2, 0, 1)$, $\mathbf{v}_2 = (2, 4, 1, 1)$, $\mathbf{v}_3 = (0, 0, 1, -1)$ are linearly independent in $\mathbb{R}^4$. If not, find a linear relation.

Solution. Row reduce the matrix with these columns:

$$\begin{pmatrix} 1 & 2 & 0 \\ 2 & 4 & 0 \\ 0 & 1 & 1 \\ 1 & 1 & -1 \end{pmatrix} \xrightarrow[R_4 \to R_4 - R_1]{R_2 \to R_2 - 2R_1} \begin{pmatrix} 1 & 2 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & -1 & -1 \end{pmatrix} \xrightarrow{R_2 \leftrightarrow R_3} \begin{pmatrix} 1 & 2 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & -1 & -1 \end{pmatrix} \xrightarrow[R_1 \to R_1 - 2R_2]{R_4 \to R_4 + R_2} \begin{pmatrix} 1 & 0 & -2 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Only columns 1 and 2 are pivot columns, so the rank is $2 < 3$, and the vectors are **dependent**

by Theorem 3.5.10 (a). Column 3 of the RREF gives the relation, by Theorem 3.5.2 (b): $\mathbf{v}_3 = -2\mathbf{v}_1 + \mathbf{v}_2$. Check: $-2(1,2,0,1)+(2,4,1,1) = (0,0,1,-1)$. Moreover $(\mathbf{v}_1, \mathbf{v}_2)$ is a basis of $\operatorname{span}(\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3)$ by Theorem 3.5.2 (a).

To **extend** an independent list to a basis of F^m, append the standard basis and keep the pivot columns. The appended vectors guarantee that the columns span F^m. The original vectors all survive, because by Theorem 3.5.2 (c) a column is discarded only if it lies in the span of the columns before it, and an independent list has no such vector (Lemma 1.5.3).

Example. Extending to a basis by elimination

Extend the independent list $(\mathbf{v}_1, \mathbf{v}_2) = ((1,2,0,1),(2,4,1,1))$ from Example (Testing independence) to a basis of $\mathbb{R}^4$.

Solution. Form $\mathbf{M} = \begin{pmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \mathbf{e}_1 & \mathbf{e}_2 & \mathbf{e}_3 & \mathbf{e}_4 \end{pmatrix}$ and row reduce:

$$\begin{pmatrix} 1 & 2 & 1 & 0 & 0 & 0 \\ 2 & 4 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \end{pmatrix} \xrightarrow[R_4 \to R_4 - R_1]{R_2 \to R_2 - 2R_1} \begin{pmatrix} 1 & 2 & 1 & 0 & 0 & 0 \\ 0 & 0 & -2 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & -1 & -1 & 0 & 0 & 1 \end{pmatrix}$$

$$\xrightarrow[\text{then } R_4 \to R_4 + R_2]{R_2 \leftrightarrow R_3} \begin{pmatrix} 1 & 2 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & -2 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 & 1 & 1 \end{pmatrix}$$

$$\xrightarrow[\text{then } R_3 \to -R_3]{R_3 \leftrightarrow R_4} \begin{pmatrix} 1 & 2 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & -1 & -1 \\ 0 & 0 & -2 & 1 & 0 & 0 \end{pmatrix} \xrightarrow{R_4 \to R_4 + 2R_3} \begin{pmatrix} 1 & 2 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & -1 & -1 \\ 0 & 0 & 0 & 1 & -2 & -2 \end{pmatrix}.$$

This is an echelon form with pivots in columns $1, 2, 3, 4$. By Corollary 3.5.6 these are the pivot columns of $\mathbf{M}$, so by Theorem 3.5.2 (a) the corresponding columns of $\mathbf{M}$ form a basis of $\operatorname{col}(\mathbf{M}) = \mathbb{R}^4$:

$$(\mathbf{v}_1, \mathbf{v}_2, \mathbf{e}_1, \mathbf{e}_2).$$

(Clearing above the pivots gives the RREF, whose column 5 says $\mathbf{e}_3 = -\mathbf{v}_1 + \mathbf{v}_2 - \mathbf{e}_1 - 2\mathbf{e}_2$; check: $-(1,2,0,1) + (2,4,1,1) - (1,0,0,0) - 2(0,1,0,0) = (0,0,1,0)$.)

For sums and intersections, one reduction does both jobs.

Proposition 3.5.11. Sum and intersection by elimination

Let $U, W \subseteq F^m$ be subspaces with bases $(\mathbf{u}_1, \ldots, \mathbf{u}_k)$ and $(\mathbf{w}_1, \ldots, \mathbf{w}_l)$, and let $\mathbf{M} = \begin{pmatrix} \mathbf{u}_1 & \cdots & \mathbf{u}_k & \mathbf{w}_1 & \cdots & \mathbf{w}_l \end{pmatrix} \in M_{m \times (k+l)}(F)$. Then:

(a) the pivot columns of $\mathbf{M}$ form a basis of $U + W$;

(b) for $\mathbf{x} = (x_1, \ldots, x_{k+l}) \in F^{k+l}$ write $\mathbf{u}_\mathbf{x} := x_1\mathbf{u}_1 + \cdots + x_k\mathbf{u}_k$, the combination of the

$\mathbf{u}$'s with the first k entries of $\mathbf{x}$ as coefficients. If $(\mathbf{n}_1, \dots, \mathbf{n}_d)$ is a basis of $\operatorname{nul}(\mathbf{M})$, then $(\mathbf{u}_{\mathbf{n}_1}, \dots, \mathbf{u}_{\mathbf{n}_d})$ is a basis of $U \cap W$.

In particular $\dim(U + W) = \operatorname{rank} \mathbf{M}$ and $\dim(U \cap W) = \operatorname{nullity} \mathbf{M}$.

Proof.

(a) By Proposition 1.7.2, $U + W = \operatorname{span}(\mathbf{u}_1, \dots, \mathbf{u}_k, \mathbf{w}_1, \dots, \mathbf{w}_l) = \operatorname{col}(\mathbf{M})$, and Theorem 3.5.2 (a) applies.

(b) By Theorem 0.9.3, $\mathbf{x} \in \operatorname{nul}(\mathbf{M})$ means $x_1\mathbf{u}_1 + \cdots + x_k\mathbf{u}_k = -(x_{k+1}\mathbf{w}_1 + \cdots + x_{k+l}\mathbf{w}_l)$, so $\mathbf{u}_{\mathbf{x}}$ lies in U and in W. Also $\mathbf{u}_{\mathbf{x}+\mathbf{y}} = \mathbf{u}_{\mathbf{x}} + \mathbf{u}_{\mathbf{y}}$ and $\mathbf{u}_{c\mathbf{x}} = c\,\mathbf{u}_{\mathbf{x}}$, by the vector space axioms in F^m.

Spanning. Let $\mathbf{v} \in U \cap W$. Since the $\mathbf{u}$'s span U and the $\mathbf{w}$'s span W, $\mathbf{v} = a_1\mathbf{u}_1 + \cdots + a_k\mathbf{u}_k = b_1\mathbf{w}_1 + \cdots + b_l\mathbf{w}_l$. Then $\mathbf{x} = (a_1, \dots, a_k, -b_1, \dots, -b_l) \in \operatorname{nul}(\mathbf{M})$ and $\mathbf{u}_{\mathbf{x}} = \mathbf{v}$. Writing $\mathbf{x} = \sum_i t_i\mathbf{n}_i$ gives $\mathbf{v} = \sum_i t_i\mathbf{u}_{\mathbf{n}_i}$.

Independence. Let $\sum_i t_i\mathbf{u}_{\mathbf{n}_i} = \mathbf{0}$, and put $\mathbf{x} = \sum_i t_i\mathbf{n}_i \in \operatorname{nul}(\mathbf{M})$. Then $\mathbf{u}_{\mathbf{x}} = x_1\mathbf{u}_1 + \cdots + x_k\mathbf{u}_k = \mathbf{0}$, so $x_1 = \cdots = x_k = 0$ because the $\mathbf{u}$'s are independent. Since $\mathbf{M}\mathbf{x} = \mathbf{0}$, also $x_{k+1}\mathbf{w}_1 + \cdots + x_{k+l}\mathbf{w}_l = \mathbf{0}$, so the remaining entries vanish because the $\mathbf{w}$'s are independent. Hence $\mathbf{x} = \mathbf{0}$, and $t_1 = \cdots = t_d = 0$ because the $\mathbf{n}_i$ are independent.

The dimension statements follow by counting the two bases, using Theorem 3.5.5 for $\operatorname{rank} \mathbf{M}$.

■

Adding the two counts and using Theorem 3.5.7 gives $\dim(U + W) + \dim(U \cap W) = \operatorname{rank} \mathbf{M} + \operatorname{nullity} \mathbf{M} = k + l = \dim U + \dim W$. This is the Dimension Formula for Sums (Theorem 1.7.3) again, now with a second, computational proof for subspaces of F^m. (Some texts use the matrix with columns $\mathbf{u}_1, \dots, \mathbf{u}_k, -\mathbf{w}_1, \dots, -\mathbf{w}_l$ instead of $\mathbf{M}$, so that a null vector gives equal combinations $\sum a_i\mathbf{u}_i = \sum b_j\mathbf{w}_j$ with no sign change. The two null spaces differ only by the signs of the last l entries.)

Example. A basis of a sum

Let $U = \operatorname{span}(\mathbf{u}_1, \mathbf{u}_2)$ and $W = \operatorname{span}(\mathbf{w}_1, \mathbf{w}_2)$ in $\mathbb{R}^4$, where

$$\mathbf{u}_1 = (1, 0, 1, 2), \quad \mathbf{u}_2 = (0, 1, 1, 1), \quad \mathbf{w}_1 = (1, 1, 0, 1), \quad \mathbf{w}_2 = (1, 3, 0, 1).$$

Find a basis and the dimension of $U + W$.

Solution. Each pair is independent, since neither vector is a multiple of the other (compare the first two entries), so $\dim U = \dim W = 2$. Row reduce $\mathbf{M} = \begin{pmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \mathbf{w}_1 & \mathbf{w}_2 \end{pmatrix}$:

$$\begin{pmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 3 \\ 1 & 1 & 0 & 0 \\ 2 & 1 & 1 & 1 \end{pmatrix} \xrightarrow[R_4 \to R_4 - 2R_1]{R_3 \to R_3 - R_1} \begin{pmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 3 \\ 0 & 1 & -1 & -1 \\ 0 & 1 & -1 & -1 \end{pmatrix} \xrightarrow[R_4 \to R_4 - R_2]{R_3 \to R_3 - R_2} \begin{pmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 3 \\ 0 & 0 & -2 & -4 \\ 0 & 0 & -2 & -4 \end{pmatrix}$$

$$\xrightarrow[\text{then } R_3 \to -\frac{1}{2}R_3]{R_4 \to R_4 - R_3} \begin{pmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 3 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 0 \end{pmatrix} \xrightarrow[R_2 \to R_2 - R_3]{R_1 \to R_1 - R_3} \mathbf{R} = \begin{pmatrix} 1 & 0 & 0 & -1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

The pivot columns are $1, 2, 3$. By Proposition 3.5.11 (a), $(\mathbf{u}_1, \mathbf{u}_2, \mathbf{w}_1)$ is a basis of $U + W$, and $\dim(U + W) = 3$. Column 4 of $\mathbf{R}$ says $\mathbf{w}_2 = -\mathbf{u}_1 + \mathbf{u}_2 + 2\mathbf{w}_1$; check: $-(1,0,1,2) + (0,1,1,1) + 2(1,1,0,1) = (1,3,0,1)$.

Example. A basis of an intersection

For U and W as in Example (A basis of a sum), find a basis of $U \cap W$, and check the Dimension Formula for Sums.

Solution. We reuse the RREF $\mathbf{R}$ of $\mathbf{M}$. Its only free column is column 4, so by Theorem 3.5.4 $\operatorname{nul}(\mathbf{M})$ has the basis $\mathbf{n} = (1, -1, -2, 1)$ (set $x_4 = 1$; then $x_1 = 1$, $x_2 = -1$, $x_3 = -2$). The relation $\mathbf{Mn} = \mathbf{0}$ reads $\mathbf{u}_1 - \mathbf{u}_2 - 2\mathbf{w}_1 + \mathbf{w}_2 = \mathbf{0}$, that is,

$$\mathbf{u}_1 - \mathbf{u}_2 = 2\mathbf{w}_1 - \mathbf{w}_2.$$

By Proposition 3.5.11 (b), $\mathbf{u_n} = \mathbf{u}_1 - \mathbf{u}_2 = (1, -1, 0, 1)$ is a basis of $U \cap W$. Check that it lies in W: $2(1,1,0,1) - (1,3,0,1) = (1,-1,0,1)$. So $\dim(U \cap W) = 1$, and

$$\dim U + \dim W - \dim(U \cap W) = 2 + 2 - 1 = 3 = \dim(U + W),$$

in agreement with Theorem 1.7.3.

3.5.9 Exercises

3.5.9.1 A. Check your understanding

Exercise. A1

(a) Define $\operatorname{col}(\mathbf{A})$, $\operatorname{row}(\mathbf{A})$, $\operatorname{nul}(\mathbf{A})$, and $\operatorname{rank}\mathbf{A}$ for $\mathbf{A} \in M_{m\times n}(F)$, and say which space each subspace lives in.

(b) State the Rank–Nullity Theorem for matrices and the Rouché–Capelli Theorem.

(c) True or false: if $\mathbf{R}$ is the RREF of $\mathbf{A}$, then $\operatorname{col}(\mathbf{R}) = \operatorname{col}(\mathbf{A})$. Justify your answer.

(d) True or false: there is a 4×6 real matrix with nullity 1. Justify your answer.

(e) True or false: for every $\mathbf{A} \in M_{m\times n}(F)$, $\operatorname{rank}\mathbf{A} = \operatorname{rank}\mathbf{A}^\top$. Justify your answer.

(f) Describe how to find a basis of $U \cap W$ for subspaces $U, W \subseteq F^m$ given by bases.

Solution.

(a) $\operatorname{col}(\mathbf{A})$ is the span of the columns, a subspace of F^m; $\operatorname{row}(\mathbf{A})$ is the span of the rows, a subspace of $M_{1\times n}(F)$; $\operatorname{nul}(\mathbf{A}) = \{\mathbf{x} \in F^n : \mathbf{Ax} = \mathbf{0}\}$, a subspace of F^n; and $\operatorname{rank}\mathbf{A} = \dim\operatorname{col}(\mathbf{A})$ (Definition 3.5.1, Definition 3.5.2, Definition 3.5.3, Definition 3.5.4).

(b) For $\mathbf{A} \in M_{m\times n}(F)$, $\operatorname{rank}\mathbf{A} + \operatorname{nullity}\mathbf{A} = n$ (Theorem 3.5.7). For $\mathbf{b} \in F^m$, $\mathbf{Ax} = \mathbf{b}$ is consistent if and only if $\operatorname{rank}\mathbf{A} = \operatorname{rank}[\mathbf{A} \mid \mathbf{b}]$; then the solutions are $\mathbf{p} + \operatorname{nul}(\mathbf{A})$, an $(n - \operatorname{rank}\mathbf{A})$-parameter family, with exactly $q^{n-\operatorname{rank}\mathbf{A}}$ solutions over a field with q elements (Theorem 3.5.8).

(c) False. For $\mathbf{A} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$, $\mathbf{R} = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}$, and $(1,1) \in \operatorname{col}(\mathbf{A})$ but $(1,1) \notin \operatorname{col}(\mathbf{R}) = \operatorname{span}((1,0))$. Row operations preserve the row space and the column relations, not the column space (Theorem 3.5.1).

(d) False. Its rank is at most 4, since the column space lies in $\mathbb{R}^4$, so by Theorem 3.5.7 its nullity is $6 - \operatorname{rank}\mathbf{A} \geq 2$.

(e) True. This is Theorem 3.5.5.

(f) Put the basis vectors of U and then of W as the columns of $\mathbf{M}$, find a basis of $\operatorname{nul}(\mathbf{M})$ from the RREF, and for each basis vector $\mathbf{n}$ take the combination of the $\mathbf{u}$'s with the first k entries of $\mathbf{n}$ as coefficients (Proposition 3.5.11).

3.5.9.2 B. Practice

Exercise. B1: The three subspaces of a 3×5 matrix

Let

$$\mathbf{A} = \begin{pmatrix} 1 & -1 & 2 & 0 & 3 \\ 2 & -2 & 5 & 1 & 4 \\ -1 & 1 & -1 & 1 & -5 \end{pmatrix} \in M_{3\times 5}(\mathbb{R}).$$

Find bases of $\operatorname{col}(\mathbf{A})$, $\operatorname{row}(\mathbf{A})$ and $\operatorname{nul}(\mathbf{A})$. Hence verify $\operatorname{rank}\mathbf{A} + \operatorname{nullity}\mathbf{A} = 5$.

Solution. Row reduce:

$$\mathbf{A} \xrightarrow[R_3\to R_3+R_1]{R_2\to R_2-2R_1} \begin{pmatrix} 1 & -1 & 2 & 0 & 3 \\ 0 & 0 & 1 & 1 & -2 \\ 0 & 0 & 1 & 1 & -2 \end{pmatrix} \xrightarrow[R_1\to R_1-2R_2]{R_3\to R_3-R_2} \mathbf{R} = \begin{pmatrix} 1 & -1 & 0 & -2 & 7 \\ 0 & 0 & 1 & 1 & -2 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

The pivot columns are 1 and 3; the free columns are $2, 4, 5$.

By Theorem 3.5.2, the pivot columns of $\mathbf{A}$, $(1,2,-1)$ and $(2,5,-1)$, form a basis of $\operatorname{col}(\mathbf{A})$. By Theorem 3.5.3, $\begin{pmatrix} 1 & -1 & 0 & -2 & 7 \end{pmatrix}$ and $\begin{pmatrix} 0 & 0 & 1 & 1 & -2 \end{pmatrix}$ form a basis of $\operatorname{row}(\mathbf{A})$. The reduced equations are $x_1 = x_2 + 2x_4 - 7x_5$ and $x_3 = -x_4 + 2x_5$, so by Theorem 3.5.4 a basis of $\operatorname{nul}(\mathbf{A})$ is

$$\mathbf{s}_2 = (1,1,0,0,0), \qquad \mathbf{s}_4 = (2,0,-1,1,0), \qquad \mathbf{s}_5 = (-7,0,2,0,1).$$

For example $\mathbf{A}\mathbf{s}_4 = 2\mathbf{a}_1 - \mathbf{a}_3 + \mathbf{a}_4 = (2,4,-2) - (2,5,-1) + (0,1,1) = \mathbf{0}$. Hence $\operatorname{rank}\mathbf{A} = 2$ and $\operatorname{nullity}\mathbf{A} = 3$, and $2 + 3 = 5$, as Theorem 3.5.7 requires.

Exercise. B2: Independence and extension in $\mathbb{R}^4$

Let $\mathbf{v}_1 = (1,0,2,1)$, $\mathbf{v}_2 = (2,1,3,1)$, $\mathbf{v}_3 = (0,1,-1,-1)$ in $\mathbb{R}^4$.

(a) Determine whether $(\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3)$ is linearly independent. Justify your answer.

(b) Find a basis of $\operatorname{span}(\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3)$, and extend it to a basis of $\mathbb{R}^4$ using standard basis vectors.

Solution.

(a) Row reduce the matrix with columns $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$:

$$\begin{pmatrix} 1 & 2 & 0 \\ 0 & 1 & 1 \\ 2 & 3 & -1 \\ 1 & 1 & -1 \end{pmatrix} \xrightarrow[R_4 \to R_4 - R_1]{R_3 \to R_3 - 2R_1} \begin{pmatrix} 1 & 2 & 0 \\ 0 & 1 & 1 \\ 0 & -1 & -1 \\ 0 & -1 & -1 \end{pmatrix} \xrightarrow[R_4 \to R_4 + R_2,\ R_1 \to R_1 - 2R_2]{R_3 \to R_3 + R_2} \begin{pmatrix} 1 & 0 & -2 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

The rank is $2 < 3$, so the list is dependent by Theorem 3.5.10 (a). Column 3 gives $\mathbf{v}_3 = -2\mathbf{v}_1 + \mathbf{v}_2$ (Theorem 3.5.2 (b)); check: $(-2, 0, -4, -2) + (2, 1, 3, 1) = (0, 1, -1, -1)$.

(b) By Theorem 3.5.2 (a), $(\mathbf{v}_1, \mathbf{v}_2)$ is a basis of $\operatorname{span}(\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3)$. To extend it, row reduce $\begin{pmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \mathbf{e}_1 & \mathbf{e}_2 & \mathbf{e}_3 & \mathbf{e}_4 \end{pmatrix}$:

$$\begin{pmatrix} 1 & 2 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 2 & 3 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \end{pmatrix} \xrightarrow[R_4 \to R_4 - R_1]{R_3 \to R_3 - 2R_1} \begin{pmatrix} 1 & 2 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & -1 & -2 & 0 & 1 & 0 \\ 0 & -1 & -1 & 0 & 0 & 1 \end{pmatrix}$$

$$\xrightarrow[R_4 \to R_4 + R_2]{R_3 \to R_3 + R_2} \begin{pmatrix} 1 & 2 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & -2 & 1 & 1 & 0 \\ 0 & 0 & -1 & 1 & 0 & 1 \end{pmatrix}$$

$$\xrightarrow[\text{then } R_3 \to -R_3]{R_3 \leftrightarrow R_4} \begin{pmatrix} 1 & 2 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & -1 \\ 0 & 0 & -2 & 1 & 1 & 0 \end{pmatrix} \xrightarrow{R_4 \to R_4 + 2R_3} \begin{pmatrix} 1 & 2 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & -1 \\ 0 & 0 & 0 & -1 & 1 & -2 \end{pmatrix}.$$

This echelon form has pivots in columns $1, 2, 3, 4$, which are the pivot columns by Corollary 3.5.6. By Theorem 3.5.2 (a) the corresponding columns form a basis of the column space, which is $\mathbb{R}^4$ since it contains $\mathbf{e}_1, \ldots, \mathbf{e}_4$. Hence $(\mathbf{v}_1, \mathbf{v}_2, \mathbf{e}_1, \mathbf{e}_2)$ is a basis of $\mathbb{R}^4$ extending $(\mathbf{v}_1, \mathbf{v}_2)$.

Exercise. B3: Sum and intersection in $\mathbb{R}^5$

Let $U = \operatorname{span}(\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3)$ and $W = \operatorname{span}(\mathbf{w}_1, \mathbf{w}_2)$ in $\mathbb{R}^5$, where

$$\mathbf{u}_1 = (1, 1, 0, 0, 1), \quad \mathbf{u}_2 = (0, 1, 1, 0, 0), \quad \mathbf{u}_3 = (1, 0, 0, 1, 0),$$
$$\mathbf{w}_1 = (1, 2, 2, 1, 2), \quad \mathbf{w}_2 = (0, 0, 1, 1, 1).$$

Find $\dim U$, $\dim W$, a basis of $U + W$, and a basis of $U \cap W$. Hence verify Theorem 1.7.3.

Solution. Row reduce $\mathbf{M} = \begin{pmatrix}\mathbf{u}_1 & \mathbf{u}_2 & \mathbf{u}_3 & \mathbf{w}_1 & \mathbf{w}_2\end{pmatrix}$:

$$\begin{pmatrix}1&0&1&1&0\\1&1&0&2&0\\0&1&0&2&1\\0&0&1&1&1\\1&0&0&2&1\end{pmatrix} \xrightarrow[R_5\to R_5-R_1]{R_2\to R_2-R_1} \begin{pmatrix}1&0&1&1&0\\0&1&-1&1&0\\0&1&0&2&1\\0&0&1&1&1\\0&0&-1&1&1\end{pmatrix}$$

$$\xrightarrow{R_3\to R_3-R_2} \begin{pmatrix}1&0&1&1&0\\0&1&-1&1&0\\0&0&1&1&1\\0&0&1&1&1\\0&0&-1&1&1\end{pmatrix}$$

$$\xrightarrow[R_5\to R_5+R_3]{R_4\to R_4-R_3} \begin{pmatrix}1&0&1&1&0\\0&1&-1&1&0\\0&0&1&1&1\\0&0&0&0&0\\0&0&0&2&2\end{pmatrix} \xrightarrow[\text{then } R_4\to\frac{1}{2}R_4]{R_4\leftrightarrow R_5} \begin{pmatrix}1&0&1&1&0\\0&1&-1&1&0\\0&0&1&1&1\\0&0&0&1&1\\0&0&0&0&0\end{pmatrix}.$$

Clearing above the pivots ($R_3 \to R_3-R_4$, $R_2 \to R_2-R_4$, $R_1 \to R_1-R_4$, then $R_2 \to R_2+R_3$, $R_1 \to R_1 - R_3$) gives

$$\mathbf{R} = \begin{pmatrix}1&0&0&0&-1\\0&1&0&0&-1\\0&0&1&0&0\\0&0&0&1&1\\0&0&0&0&0\end{pmatrix}.$$

The pivot columns are $1, 2, 3, 4$, so $(\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3, \mathbf{w}_1)$ is independent by Theorem 3.5.2 (a), and so are its sub-lists $(\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3)$ and $(\mathbf{w}_1)$; also $\mathbf{w}_2$ is not a multiple of $\mathbf{w}_1$ (compare first entries). Hence $\dim U = 3$ and $\dim W = 2$, with the given spanning lists as bases. By Proposition 3.5.11 (a), $(\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3, \mathbf{w}_1)$ is a basis of $U + W$, so $\dim(U + W) = 4$.

The only free column is column 5. Setting $x_5 = 1$ gives $\mathbf{n} = (1, 1, 0, -1, 1)$, a basis of $\mathrm{nul}(\mathbf{M})$ by Theorem 3.5.4. So $\mathbf{u}_1 + \mathbf{u}_2 - \mathbf{w}_1 + \mathbf{w}_2 = \mathbf{0}$, and by Proposition 3.5.11 (b), $\mathbf{u}_1 + \mathbf{u}_2 = (1, 2, 1, 0, 1)$ is a basis of $U \cap W$. Check: $\mathbf{w}_1 - \mathbf{w}_2 = (1, 2, 1, 0, 1)$. Hence $\dim(U \cap W) = 1$, and

$$\dim U + \dim W - \dim(U \cap W) = 3 + 2 - 1 = 4 = \dim(U + W).$$

3.5.9.3 C. Going deeper

Exercise. C1: Rank of a sum

Let $\mathbf{A}, \mathbf{B} \in M_{m\times n}(F)$. Prove that $\mathrm{rank}(\mathbf{A}+\mathbf{B}) \le \mathrm{rank}\,\mathbf{A} + \mathrm{rank}\,\mathbf{B}$. Give an example where the inequality is strict.

Solution. The j-th column of $\mathbf{A}+\mathbf{B}$ is $\mathbf{a}_j+\mathbf{b}_j$, which lies in $\mathrm{col}(\mathbf{A})+\mathrm{col}(\mathbf{B})$ (Definition 1.7.1). This sum is a subspace (Theorem 1.7.1), so by Theorem 1.3.1, $\mathrm{col}(\mathbf{A}+\mathbf{B}) \subseteq \mathrm{col}(\mathbf{A})+\mathrm{col}(\mathbf{B})$.

By Theorem 1.6.7 and Theorem 1.7.3,

$$\begin{aligned}\operatorname{rank}(\mathbf{A}+\mathbf{B}) &\le \dim(\operatorname{col}(\mathbf{A})+\operatorname{col}(\mathbf{B}))\\ &= \operatorname{rank}\mathbf{A}+\operatorname{rank}\mathbf{B}-\dim(\operatorname{col}(\mathbf{A})\cap\operatorname{col}(\mathbf{B}))\\ &\le \operatorname{rank}\mathbf{A}+\operatorname{rank}\mathbf{B}.\end{aligned}$$

For strictness, take $\mathbf{A}=\mathbf{I}_2$ and $\mathbf{B}=-\mathbf{I}_2$ over $\mathbb{R}$: $\operatorname{rank}(\mathbf{A}+\mathbf{B})=\operatorname{rank}\mathbf{O}=0<2+2$.

Exercise. C2: Matrices of rank one

Let $\mathbf{A}\in M_{m\times n}(F)$. Prove that $\operatorname{rank}\mathbf{A}=1$ if and only if $\mathbf{A}=\mathbf{x}\mathbf{y}^{\top}$ for some **nonzero** $\mathbf{x}\in F^m$ and $\mathbf{y}\in F^n$.

Solution. ($\Leftarrow$) Suppose $\mathbf{A}=\mathbf{x}\mathbf{y}^{\top}$ with $\mathbf{x},\mathbf{y}\neq\mathbf{0}$, and write $\mathbf{y}=(y_1,\dots,y_n)$. By Theorem 0.9.2, the j-th column of $\mathbf{x}\mathbf{y}^{\top}$ is $\mathbf{x}$ times the j-th column of $\mathbf{y}^{\top}$, the 1×1 matrix (y_j); that is, $y_j\mathbf{x}$. So every column lies in $\operatorname{span}(\mathbf{x})$, and $\operatorname{col}(\mathbf{A})\subseteq\operatorname{span}(\mathbf{x})$. Since $\mathbf{y}\neq\mathbf{0}$, some $y_j\neq 0$, and then $y_j\mathbf{x}\neq\mathbf{0}$ because $\mathbf{x}\neq\mathbf{0}$ (Theorem 1.1.9). The list $(\mathbf{x})$ is independent, since $a\mathbf{x}=\mathbf{0}$ forces $a=0$ (Theorem 1.1.9), so $\operatorname{span}(\mathbf{x})$ has dimension 1, and $\operatorname{rank}\mathbf{A}\le 1$ by Theorem 1.6.7. On the other hand $\operatorname{rank}\mathbf{A}\neq 0$: a space of dimension 0 has the empty list as a basis, so it equals $\operatorname{span}()=\{\mathbf{0}\}$, but $\operatorname{col}(\mathbf{A})$ contains the nonzero vector $y_j\mathbf{x}$. Hence $\operatorname{rank}\mathbf{A}=1$.

($\Rightarrow$) Suppose $\operatorname{rank}\mathbf{A}=1$, and let $(\mathbf{x})$ be a basis of $\operatorname{col}(\mathbf{A})$; then $\mathbf{x}\neq\mathbf{0}$. Each column $\mathbf{a}_j\in\operatorname{col}(\mathbf{A})$ equals $y_j\mathbf{x}$ for some $y_j\in F$. Put $\mathbf{y}=(y_1,\dots,y_n)$. By the computation in ($\Leftarrow$), the j-th column of $\mathbf{x}\mathbf{y}^{\top}$ is $y_j\mathbf{x}=\mathbf{a}_j$, so $\mathbf{A}=\mathbf{x}\mathbf{y}^{\top}$. If $\mathbf{y}=\mathbf{0}$, then $\mathbf{A}=\mathbf{O}$ would have rank 0, so $\mathbf{y}\neq\mathbf{0}$. This proves the claim.

Exercise. C3: The matrix $i+j$

For $n\ge 1$, let $\mathbf{A}_n\in M_n(\mathbb{R})$ have entries $a_{ij}=i+j$. Prove that $\operatorname{rank}\mathbf{A}_1=1$ and $\operatorname{rank}\mathbf{A}_n=2$ for every $n\ge 2$. Hence find $\operatorname{nullity}\mathbf{A}_n$.

Hint: write each row as a combination of two fixed row vectors.

Solution. For $n=1$, $\mathbf{A}_1=(2)\neq\mathbf{O}$, so $\operatorname{rank}\mathbf{A}_1=1$. Let $n\ge 2$, and put $\mathbf{1}=\begin{pmatrix}1 & 1 & \cdots & 1\end{pmatrix}$ and $\mathbf{d}=\begin{pmatrix}1 & 2 & \cdots & n\end{pmatrix}$ in $M_{1\times n}(\mathbb{R})$. Row i of $\mathbf{A}_n$ is $\begin{pmatrix}i+1 & \cdots & i+n\end{pmatrix}=i\mathbf{1}+\mathbf{d}$. So $\operatorname{row}(\mathbf{A}_n)\subseteq\operatorname{span}(\mathbf{1},\mathbf{d})$, and $\dim\operatorname{row}(\mathbf{A}_n)\le 2$ by Theorem 1.6.3.

Rows 1 and 2 are $\begin{pmatrix}2 & 3 & \cdots\end{pmatrix}$ and $\begin{pmatrix}3 & 4 & \cdots\end{pmatrix}$. If $c_1(\text{row }1)+c_2(\text{row }2)=\mathbf{0}$, the first two entries give $2c_1+3c_2=0$ and $3c_1+4c_2=0$. Subtracting the first from the second gives $c_1+c_2=0$, so $c_1=-c_2$, and then $-2c_2+3c_2=c_2=0$, so $c_1=0$. Hence $\operatorname{row}(\mathbf{A}_n)$ contains two independent vectors, and $\dim\operatorname{row}(\mathbf{A}_n)\ge 2$ by Theorem 1.6.3.

So $\dim\operatorname{row}(\mathbf{A}_n)=2$, and $\operatorname{rank}\mathbf{A}_n=2$ by Theorem 3.5.5. By Theorem 3.5.7, $\operatorname{nullity}\mathbf{A}_n=n-2$ for $n\ge 2$, and $\operatorname{nullity}\mathbf{A}_1=0$.

Exercise. C4: Counting right-hand sides over a finite field

Let F be a finite field with q elements, and let $\mathbf{A} \in M_{m\times n}(F)$ have rank r.

(a) Prove that exactly q^r vectors $\mathbf{b} \in F^m$ make $\mathbf{Ax} = \mathbf{b}$ consistent.

(b) By counting the pairs $(\mathbf{x}, \mathbf{b}) \in F^n \times F^m$ with $\mathbf{Ax} = \mathbf{b}$ in two ways, and using Theorem 3.1.3 but **not** Theorem 3.5.7, prove that $\operatorname{nul}(\mathbf{A})$ has exactly q^{n-r} elements. Deduce $\operatorname{nullity} \mathbf{A} = n - r$: a counting proof of Rank–Nullity over a finite field.

(c) Over $\mathbb{F}_2$, let $\mathbf{A} = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{pmatrix}$. How many $\mathbf{b} \in \mathbb{F}_2^3$ give a consistent system, and how many solutions does each such system have?

Solution.

(a) By Theorem 3.1.1, $\mathbf{Ax} = \mathbf{b}$ is consistent exactly when $\mathbf{b} \in \operatorname{col}(\mathbf{A})$. Let $(\mathbf{c}_1, \dots, \mathbf{c}_r)$ be a basis of $\operatorname{col}(\mathbf{A})$. By Theorem 1.5.1, $(t_1, \dots, t_r) \mapsto t_1\mathbf{c}_1 + \cdots + t_r\mathbf{c}_r$ is a bijection $F^r \to \operatorname{col}(\mathbf{A})$: every vector of $\operatorname{col}(\mathbf{A})$ is hit by exactly one tuple. Since F^r has q^r elements, so does $\operatorname{col}(\mathbf{A})$.

(b) Let $N = |\operatorname{nul}(\mathbf{A})|$. Each $\mathbf{x} \in F^n$ determines exactly one $\mathbf{b} = \mathbf{Ax}$, so there are q^n pairs. Now group the pairs by $\mathbf{b}$. If $\mathbf{b} \notin \operatorname{col}(\mathbf{A})$, there are none, by Theorem 3.1.1. If $\mathbf{b} \in \operatorname{col}(\mathbf{A})$, pick one solution $\mathbf{p}$; by Theorem 3.1.3 the solutions are $\mathbf{p} + \operatorname{nul}(\mathbf{A})$, and $\mathbf{h} \mapsto \mathbf{p} + \mathbf{h}$ is a bijection from $\operatorname{nul}(\mathbf{A})$ onto this set (it is surjective by definition, and $\mathbf{p} + \mathbf{h} = \mathbf{p} + \mathbf{h}'$ gives $\mathbf{h} = \mathbf{h}'$). So each of the q^r vectors $\mathbf{b}$ of (a) accounts for exactly N pairs, and there are $q^r N$ pairs in total. Hence $q^r N = q^n$, and $N = q^{n-r}$.
Let $d = \operatorname{nullity} \mathbf{A}$. The argument of (a), applied to a basis of $\operatorname{nul}(\mathbf{A})$ of length d, gives $N = q^d$. So $q^d = q^{n-r}$. Since $q \geq 2$, distinct exponents give distinct powers of q, so $d = n - r$. This is Theorem 3.5.7, proved here by counting alone.

(c) Over $\mathbb{F}_2$, $R_3 \to R_3 + R_1$ gives row 3 equal to $(0, 1, 1)$, and then $R_3 \to R_3 + R_2$ makes it zero. So an echelon form is $\begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \end{pmatrix}$, and $\operatorname{rank} \mathbf{A} = 2$ by Corollary 3.5.6. By (a), exactly $2^2 = 4$ of the 8 vectors $\mathbf{b}$ give a consistent system, and by Theorem 3.5.8 (c) each has $2^{3-2} = 2$ solutions. (The consistent $\mathbf{b}$ are those with $b_1 + b_2 + b_3 = 0$. The rows of $\mathbf{A}$ add up to zero, so every consistent $\mathbf{b}$ satisfies this equation; exactly 4 vectors of $\mathbb{F}_2^3$ satisfy it, and 4 are consistent, so the two sets coincide.)

3.6 LU and PLU Factorizations

Gaussian elimination solves one system $\mathbf{Ax} = \mathbf{b}$. In practice the matrix often stays fixed while the right-hand side changes: the same network under different loads, the same model fed different data. The row operations depend only on $\mathbf{A}$, so repeating them for every $\mathbf{b}$ wastes almost all of the work. This section stores the elimination as a factorization $\mathbf{A} = \mathbf{LU}$, a lower triangular matrix that remembers the operations times an upper triangular matrix that is their result. We prove when it exists and when it is unique, and we repair its failures with row swaps.

Throughout, F is a field and all matrices are square, in $M_n(F)$.

3.6.1 Triangular systems are cheap

Some systems need no elimination at all. Take

$$\mathbf{U} = \begin{pmatrix} 2 & 1 & -1 \\ 0 & 3 & 1 \\ 0 & 0 & -2 \end{pmatrix}, \qquad \mathbf{U}\mathbf{x} = \begin{pmatrix} -1 \\ -1 \\ -4 \end{pmatrix}.$$

The last equation mentions only x_3: $-2x_3 = -4$, so $x_3 = 2$. The second mentions x_2 and the known x_3: $3x_2 + 2 = -1$, so $x_2 = -1$. The first then gives $2x_1 - 1 - 2 = -1$, so $x_1 = 1$. This is **back substitution**: in an upper triangular system with nonzero diagonal entries, each equation introduces exactly one new unknown, read from the bottom up. A lower triangular system is solved the same way from the top down, by **forward substitution**.

Now suppose a matrix $\mathbf{A}$ factors as $\mathbf{A} = Ł\mathbf{U}$ with Ł lower triangular and $\mathbf{U}$ upper triangular. Then $\mathbf{A}\mathbf{x} = \mathbf{b}$ reads $Ł(\mathbf{U}\mathbf{x}) = \mathbf{b}$. Naming the inner vector $\mathbf{y} = \mathbf{U}\mathbf{x}$ splits the problem in two:

$$\text{first solve } Ł\mathbf{y} = \mathbf{b} \text{ for } \mathbf{y}, \qquad \text{then solve } \mathbf{U}\mathbf{x} = \mathbf{y} \text{ for } \mathbf{x}.$$

Both are triangular, so both are cheap. The expensive part, finding Ł and $\mathbf{U}$, is done once for all right-hand sides.

An LU factorization is Gaussian elimination written down once, so that it never has to be done again.

3.6.2 LU factorizations

Definition 3.6.1. LU factorization

A matrix $Ł \in M_n(F)$ is **unit lower triangular** if it is lower triangular and **every** diagonal entry equals 1. Let $\mathbf{A} \in M_n(F)$. An **LU factorization** of $\mathbf{A}$ is a pair $(Ł, \mathbf{U})$ of matrices in $M_n(F)$ such that

$$\mathbf{A} = Ł\mathbf{U}, \qquad Ł \text{ is } \textbf{unit} \text{ lower triangular}, \qquad \mathbf{U} \text{ is upper triangular}.$$

In words: Ł has 1's on the diagonal and zeros above it, and the entries below the diagonal are unrestricted. $\mathbf{U}$ has zeros strictly below the diagonal, and its diagonal entries are unrestricted; some of them may be 0. The definition says **an** factorization: whether there is only one is a theorem, and it needs a hypothesis.

Example. An LU factorization from elimination

Let

$$\mathbf{A} = \begin{pmatrix} 2 & 1 & -1 \\ 4 & 5 & -1 \\ -2 & 8 & 2 \end{pmatrix} \in M_3(\mathbb{Q}).$$

Reduce $\mathbf{A}$ to upper triangular form using only operations "subtract a multiple of a higher row from a lower row", and record each multiplier. Compare the result with the matrix $\mathbf{U}$ of the previous subsection.

Solution. The first pivot is 2. To clear column 1 below it we subtract $\ell_{21} = 4/2 = 2$ times row 1 from row 2, and $\ell_{31} = -2/2 = -1$ times row 1 from row 3:

$$\begin{pmatrix} 2 & 1 & -1 \\ 4 & 5 & -1 \\ -2 & 8 & 2 \end{pmatrix} \xrightarrow[R_3 \to R_3 - (-1)R_1]{R_2 \to R_2 - 2R_1} \begin{pmatrix} 2 & 1 & -1 \\ 0 & 3 & 1 \\ 0 & 9 & 1 \end{pmatrix}.$$

The second pivot is 3. We subtract $\ell_{32} = 9/3 = 3$ times row 2 from row 3:

$$\begin{pmatrix} 2 & 1 & -1 \\ 0 & 3 & 1 \\ 0 & 9 & 1 \end{pmatrix} \xrightarrow{R_3 \to R_3 - 3R_2} \begin{pmatrix} 2 & 1 & -1 \\ 0 & 3 & 1 \\ 0 & 0 & -2 \end{pmatrix} = \mathbf{U}.$$

Put the multipliers below the diagonal of a unit lower triangular matrix, each in the position of the entry it cleared:

$$Ł = \begin{pmatrix} 1 & 0 & 0 \\ \ell_{21} & 1 & 0 \\ \ell_{31} & \ell_{32} & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ -1 & 3 & 1 \end{pmatrix}.$$

Multiplying out, row by row,

$$Ł\mathbf{U} = \begin{pmatrix} 2 & 1 & -1 \\ 2 \cdot 2 & 2 \cdot 1 + 3 & 2 \cdot (-1) + 1 \\ -2 & -1 + 9 & 1 + 3 - 2 \end{pmatrix} = \begin{pmatrix} 2 & 1 & -1 \\ 4 & 5 & -1 \\ -2 & 8 & 2 \end{pmatrix} = \mathbf{A}.$$

So $(Ł, \mathbf{U})$ is an LU factorization of $\mathbf{A}$, and $\mathbf{U}$ is exactly the triangular matrix of the back substitution above.

The example suggests the general statement: elimination produces $\mathbf{U}$, and the multipliers, **with their signs as subtracted**, fill in Ł. We prove this below. First, the smallest cases.

- **Upper triangular A.** $\mathbf{A} = \mathbf{I} \cdot \mathbf{A}$, and $\mathbf{I}$ is unit lower triangular. So every upper triangular matrix, including $\mathbf{O}$ and $\mathbf{I}$, has the LU factorization $(\mathbf{I}, \mathbf{A})$.
- **1×1.** Every $(a) \in M_1(F)$ is upper triangular, so $(a) = (1)(a)$. Nothing can go wrong here, and that is exactly why the 2×2 failure below is instructive.

Nonexample by minimal change. The matrix $\begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}$ has the LU factorization

$$\begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 0 & -1 \end{pmatrix}.$$

Change the top-left entry from 1 to 0, and everything collapses.

Example. An invertible matrix with no LU factorization

Show that $\mathbf{A} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \in M_2(F)$ has **no** LU factorization, over any field F, although $\mathbf{A}$ is invertible.

Solution. $\mathbf{A}$ is invertible, since $\mathbf{AA} = \mathbf{I}_2$. Suppose, for a contradiction, that $\mathbf{A} = Ł\mathbf{U}$ with

$$Ł = \begin{pmatrix} 1 & 0 \\ \ell & 1 \end{pmatrix}, \qquad \mathbf{U} = \begin{pmatrix} u & v \\ 0 & w \end{pmatrix}.$$

Then

$$Ł\mathbf{U} = \begin{pmatrix} u & v \\ \ell u & \ell v + w \end{pmatrix}.$$

Comparing $(1,1)$-entries with $\mathbf{A}$ gives $u = 0$. Comparing $(2,1)$-entries gives $\ell u = 1$. But $\ell u = \ell \cdot 0 = 0$ by Theorem 0.7.1, and $0 \neq 1$ in a field. This contradiction shows that no LU factorization exists.

What failed is visible: the $(1,1)$-entry of any product $Ł\mathbf{U}$ is u_{11} alone, so a zero there forces the whole first column of $Ł\mathbf{U}$ to vanish. In elimination terms, the first pivot is 0 while the entry below it is not, and the only cure is to **swap** rows, which "subtract a multiple of a higher row from a lower row" cannot do. We return to this with permutation matrices.

Why "unit". Without the 1's on the diagonal of Ł, uniqueness would fail even for invertible matrices. If $\mathbf{A} = Ł\mathbf{U}$ and $\mathbf{D}$ is any invertible diagonal matrix, then also $\mathbf{A} = (Ł\mathbf{D})(\mathbf{D}^{-1}\mathbf{U})$, where $Ł\mathbf{D}$ is still lower triangular and $\mathbf{D}^{-1}\mathbf{U}$ still upper triangular. When Ł is invertible and $\mathbf{D} \neq \mathbf{I}$, this is a different pair, and over any field with more than two elements such a $\mathbf{D}$ exists. For instance $\begin{pmatrix} 2 \end{pmatrix} = \begin{pmatrix} 1 \end{pmatrix}\begin{pmatrix} 2 \end{pmatrix} = \begin{pmatrix} 2 \end{pmatrix}\begin{pmatrix} 1 \end{pmatrix} = \begin{pmatrix} 4 \end{pmatrix}\begin{pmatrix} 1/2 \end{pmatrix}$ over $\mathbb{Q}$. Fixing the diagonal of Ł removes this freedom and matches what elimination does: it never rescales a row.

Warning. **Invertible does not mean "has an LU factorization".** The matrix of Example (An invertible matrix with no LU factorization) is invertible and has none. Conversely, a singular matrix can have one: $\mathbf{O} = \mathbf{I} \cdot \mathbf{O}$. Existence of an LU factorization is a property of the **order of the rows**, not of invertibility.

3.6.3 Unit lower triangular matrices

Elimination builds Ł out of elementary matrices, and the uniqueness proof below divides one factorization by another. Both need the same facts: products and inverses of unit lower triangular matrices stay unit lower triangular. We also need to know exactly when a triangular matrix is invertible.

The key observation is a statement about where the nonzero entries of a product can sit.

Lemma 3.6.1. Unit lower triangular matrices are closed

Let $Ł, Ł' \in M_n(F)$ be unit lower triangular. Then:

(a) $ŁŁ'$ is unit lower triangular;

(b) Ł is invertible, and $Ł^{-1}$ is unit lower triangular.

Idea. Write $Ł = \mathbf{I} + \mathbf{N}$, where $\mathbf{N}$ has nonzero entries only strictly below the diagonal. Multiplying two such matrices pushes the nonzero entries **further** below the diagonal, so $\mathbf{N}^n = \mathbf{O}$. Then the geometric series $1/(1+t) = 1 - t + t^2 - \cdots$ stops after n terms and gives

the inverse exactly.

Proof. Say that $\mathbf{X} \in M_n(F)$ is **p-low** if $x_{ij} = 0$ whenever $i < j + p$. So 0-low means lower triangular, and 1-low means zero on and above the diagonal.

Claim. If $\mathbf{X}$ is p-low and $\mathbf{Y}$ is q-low, then $\mathbf{XY}$ is $(p+q)$-low. A matrix that is n-low is the zero matrix.

Proof. By Definition 0.9.9, $(\mathbf{XY})_{ij} = \sum_{l=1}^{n} x_{il} y_{lj}$. A term can be nonzero only if $i \geq l + p$ and $l \geq j + q$, which together give $i \geq j + p + q$. So if $i < j + p + q$, every term is 0 and $(\mathbf{XY})_{ij} = 0$. For the second statement, every pair of indices satisfies $i \leq n < 1 + n \leq j + n$, so every entry of an n-low matrix is 0.

■

Write $Ł = \mathbf{I} + \mathbf{N}$ and $Ł' = \mathbf{I} + \mathbf{N}'$. Since $Ł$ and $Ł'$ are unit lower triangular, $\mathbf{N}$ and $\mathbf{N}'$ are 1-low.

For (a), $ŁŁ' = \mathbf{I} + \mathbf{N} + \mathbf{N}' + \mathbf{NN}'$ by Theorem 0.9.4. Here $\mathbf{N} + \mathbf{N}'$ is 1-low, since a sum of entries that are both 0 is 0, and $\mathbf{NN}'$ is 2-low, hence 1-low, by the Claim. So $ŁŁ' - \mathbf{I}$ is 1-low, which says that $ŁŁ'$ is unit lower triangular.

For (b), let $\mathbf{K} = \mathbf{I} - \mathbf{N} + \mathbf{N}^2 - \cdots + (-1)^{n-1}\mathbf{N}^{n-1}$. Multiplying out, the sum telescopes:

$$(\mathbf{I} + \mathbf{N})\mathbf{K} = \mathbf{I} + (-1)^{n-1}\mathbf{N}^n = \mathbf{I},$$

where $\mathbf{N}^n = \mathbf{O}$ because $\mathbf{N}^n$ is n-low by the Claim applied $n-1$ times. The same computation gives $\mathbf{K}(\mathbf{I} + \mathbf{N}) = \mathbf{I}$, since $\mathbf{N}$ commutes with its own powers. Hence $Ł$ is invertible with $Ł^{-1} = \mathbf{K}$. Finally $\mathbf{K} - \mathbf{I} = \sum_{p=1}^{n-1}(-1)^p \mathbf{N}^p$ is a sum of 1-low matrices, by the Claim, so $\mathbf{K}$ is unit lower triangular. This proves the lemma.

■

Transposing everything, the same statements hold for **unit upper triangular** matrices (upper triangular with 1's on the diagonal).

The next lemma decides invertibility of a triangular matrix from its diagonal, without any computation beyond back substitution.

Lemma 3.6.2. Invertibility of a triangular matrix

An upper triangular matrix $\mathbf{U} \in M_n(F)$ is invertible if and only if **every** diagonal entry $u_{11}, \ldots, u_{nn}$ is nonzero.

Proof. ($\Leftarrow$) Suppose every $u_{kk} \neq 0$, and let $\mathbf{Ux} = \mathbf{0}$. We show $x_k = 0$ for $k = n, n-1, \ldots, 1$ in turn. Suppose $x_{k+1} = \cdots = x_n = 0$ (nothing is assumed when $k = n$). Row k of $\mathbf{Ux} = \mathbf{0}$ reads $u_{kk}x_k + \sum_{j>k} u_{kj}x_j = 0$, because $u_{kj} = 0$ for $j < k$. The sum vanishes, so $u_{kk}x_k = 0$, and $x_k = 0$ since $u_{kk} \neq 0$. Hence $\mathbf{x} = \mathbf{0}$ is the only solution of $\mathbf{Ux} = \mathbf{0}$, and $\mathbf{U}$ is invertible by Theorem 3.4.4.

($\Rightarrow$) We prove the contrapositive. Suppose $u_{kk} = 0$ for some k. For each $j \leq k$, the entries of

column j in rows $i \geq k$ are zero: if $i > j$ this is upper triangularity, and the only remaining case $i = j = k$ is $u_{kk} = 0$. So the first k columns of $\mathbf{U}$ lie in $W = \operatorname{span}(\mathbf{e}_1, \ldots, \mathbf{e}_{k-1})$. The list $(\mathbf{e}_1, \ldots, \mathbf{e}_{k-1})$ is independent and spans W, so $\dim W = k - 1$ (for $k = 1$, $W = \{\mathbf{0}\}$ has dimension 0). By Theorem 1.6.3 (a), these k columns are linearly dependent, so the columns of $\mathbf{U}$ are dependent, and $\mathbf{U}$ is not invertible by Theorem 3.4.4. This proves the lemma.

■

3.6.4 Elimination without row swaps

To prove that elimination always produces an LU factorization, we first need to say exactly which elimination we mean. It is Gaussian elimination restricted to one kind of operation: subtracting a multiple of a row from a row **below** it.

Algorithm. **Elimination without row swaps.** Input: $\mathbf{A} \in M_n(F)$. Put $\mathbf{A}^{(0)} = \mathbf{A}$. For $k = 1, 2, \ldots, n-1$, write $\mathbf{A}^{(k-1)} = (a_{ij})$ and do the following.

- If $a_{kk} \neq 0$, set $\ell_{ik} = a_{ik}/a_{kk}$ for $i = k+1, \ldots, n$.
- If $a_{kk} = 0$ and $a_{ik} = 0$ for **every** $i > k$, set $\ell_{ik} = 0$ for $i = k+1, \ldots, n$.
- If $a_{kk} = 0$ but $a_{ik} \neq 0$ for **some** $i > k$, **stop**: a row swap is needed.

If we did not stop, obtain $\mathbf{A}^{(k)}$ from $\mathbf{A}^{(k-1)}$ by replacing row i with (row i) $-\ell_{ik}\cdot$ (row k), for each $i = k+1, \ldots, n$.

If the loop finishes without stopping, we say **elimination without row swaps succeeds** on $\mathbf{A}$. The numbers ℓ_{ik} $(i > k)$ are its **multipliers**, and $\mathbf{U} = \mathbf{A}^{(n-1)}$ is its **result**.

In both nonstopping cases the new (i,k)-entry is $a_{ik} - \ell_{ik}a_{kk} = 0$: in the first case by the choice of ℓ_{ik}, in the second because both terms are already 0. The multipliers are well defined: row k is not changed during step k, so it does not matter in which order the rows $i > k$ are updated.

The proof will write each step as a matrix. By Definition 3.4.1, the elementary matrix of the operation $R_i \to R_i + cR_k$ $(i \neq k)$ is $\mathbf{I} + c\mathbf{E}_{ik}$, where $\mathbf{E}_{ik}$ is the matrix unit with 1 in entry (i,k). We will write the matrix unit as a product of a column and a row,

$$\mathbf{E}_{ik} = \mathbf{e}_i \mathbf{e}_k^{\top},$$

which holds because $\mathbf{e}_i\mathbf{e}_k^{\top}$ has (p,q)-entry $(\mathbf{e}_i)_p(\mathbf{e}_k)_q$, equal to 1 if $p = i$ and $q = k$, and 0 otherwise. We will use one more rule constantly: for a vector $\mathbf{v} \in F^n$, the 1×1 matrix $\mathbf{e}_k^{\top}\mathbf{v}$ is the k-th entry v_k, which we treat as a scalar. In particular $\mathbf{e}_k^{\top}\mathbf{e}_j = 0$ for $j \neq k$.

For $1 \leq k \leq n$ and a vector $\mathbf{v} \in F^n$ whose first k entries are zero, put

$$\mathbf{G}_k(\mathbf{v}) = \mathbf{I} + \mathbf{v}\,\mathbf{e}_k^{\top}.$$

Its only nonidentity entries are v_i in position (i,k), for $i > k$: it is the identity with the column vector $\mathbf{v}$ written into column k below the diagonal. So $\mathbf{G}_k(\mathbf{v})$ is unit lower triangular.

Lemma 3.6.3. Multiplier matrices

Let $1 \leq k \leq n$, and let $\mathbf{v}, \mathbf{v}_1, \ldots, \mathbf{v}_m \in F^n$ be vectors such that the first k entries of $\mathbf{v}$ are zero and, for each j, the first j entries of $\mathbf{v}_j$ are zero $(m \leq n)$. Then:

(a) $\mathbf{G}_k(\mathbf{v})$ is the product, in any order, of the elementary matrices $\mathbf{I} + v_i\mathbf{E}_{ik}$ for $i = k+1, \dots, n$;
(b) $\mathbf{G}_k(\mathbf{v})\mathbf{G}_k(-\mathbf{v}) = \mathbf{G}_k(-\mathbf{v})\mathbf{G}_k(\mathbf{v}) = \mathbf{I}$;
(c) $\mathbf{G}_1(\mathbf{v}_1)\mathbf{G}_2(\mathbf{v}_2)\cdots\mathbf{G}_m(\mathbf{v}_m) = \mathbf{I} + \mathbf{v}_1\mathbf{e}_1^\top + \mathbf{v}_2\mathbf{e}_2^\top + \cdots + \mathbf{v}_m\mathbf{e}_m^\top$.

Proof. For (a), let $i, j > k$. By Theorem 0.9.4,

$$(\mathbf{I} + a\,\mathbf{e}_i\mathbf{e}_k^\top)(\mathbf{I} + b\,\mathbf{e}_j\mathbf{e}_k^\top) = \mathbf{I} + a\,\mathbf{e}_i\mathbf{e}_k^\top + b\,\mathbf{e}_j\mathbf{e}_k^\top + ab\,\mathbf{e}_i(\mathbf{e}_k^\top\mathbf{e}_j)\mathbf{e}_k^\top,$$

and the last term is zero because $\mathbf{e}_k^\top\mathbf{e}_j = 0$ for $j \neq k$. Repeating this, one factor at a time, the product of the $\mathbf{I} + v_i\mathbf{E}_{ik} = \mathbf{I} + v_i\mathbf{e}_i\mathbf{e}_k^\top$ in any order is $\mathbf{I} + \sum_{i>k} v_i\mathbf{e}_i\mathbf{e}_k^\top = \mathbf{I} + \mathbf{v}\,\mathbf{e}_k^\top$, since $\mathbf{v} = \sum_{i>k} v_i\mathbf{e}_i$.

For (b), $(\mathbf{I} + \mathbf{v}\,\mathbf{e}_k^\top)(\mathbf{I} - \mathbf{v}\,\mathbf{e}_k^\top) = \mathbf{I} - \mathbf{v}(\mathbf{e}_k^\top\mathbf{v})\mathbf{e}_k^\top = \mathbf{I}$, because $\mathbf{e}_k^\top\mathbf{v} = v_k = 0$. Swapping the signs gives the other order.

For (c), we use induction on m; the case $m = 1$ is the definition. Suppose the formula holds for $m-1$. Then

$$\Big(\mathbf{I} + \sum_{j<m} \mathbf{v}_j\mathbf{e}_j^\top\Big)(\mathbf{I} + \mathbf{v}_m\mathbf{e}_m^\top) = \mathbf{I} + \sum_{j<m} \mathbf{v}_j\mathbf{e}_j^\top + \mathbf{v}_m\mathbf{e}_m^\top + \sum_{j<m} \mathbf{v}_j(\mathbf{e}_j^\top\mathbf{v}_m)\mathbf{e}_m^\top.$$

For $j < m$, $\mathbf{e}_j^\top\mathbf{v}_m$ is entry j of $\mathbf{v}_m$, which is zero because the first m entries of $\mathbf{v}_m$ are zero. So the last sum vanishes, and the formula holds for m. This proves the lemma.

■

Part (c) is the heart of the matter: when the factors come in the order $\mathbf{G}_1, \mathbf{G}_2, \dots$, they **do not interfere**, and each vector $\mathbf{v}_j$ lands untouched in column j.

Theorem 3.6.4. LU from elimination

Let $\mathbf{A} \in M_n(F)$, and suppose elimination without row swaps succeeds on $\mathbf{A}$, with multipliers ℓ_{ik} and result $\mathbf{U}$. Let Ł be the unit lower triangular matrix whose (i,k)-entry is ℓ_{ik} for every $i > k$. Then $\mathbf{U}$ is upper triangular and

$$\mathbf{A} = Ł\mathbf{U}.$$

In particular $\mathbf{A}$ has an LU factorization.

Idea. Each step of elimination is left multiplication by a product of elementary matrices, which collapses to one matrix $\mathbf{G}_k(-\boldsymbol{\ell}_k)$. So $\mathbf{U} = \mathbf{G}_{n-1}(-\boldsymbol{\ell}_{n-1})\cdots\mathbf{G}_1(-\boldsymbol{\ell}_1)\mathbf{A}$. To get $\mathbf{A}$ back, undo the steps in reverse order. Their inverses $\mathbf{G}_k(\boldsymbol{\ell}_k)$ then appear in the order $\mathbf{G}_1, \mathbf{G}_2, \dots$, which is exactly the order in which, by Lemma 3.6.3 (c), the multipliers drop into place.

Proof. Let $\mathbf{A}^{(0)}, \dots, \mathbf{A}^{(n-1)} = \mathbf{U}$ be the matrices produced by the algorithm. For each k, let $\boldsymbol{\ell}_k = \sum_{i>k} \ell_{ik}\mathbf{e}_i \in F^n$, a vector whose first k entries are zero.

Step 1: U is upper triangular. We show by induction on k that in $\mathbf{A}^{(k)}$ every entry below

the diagonal in columns $1, \dots, k$ is zero. For $k = 0$ there is nothing to prove. Suppose it holds for $\mathbf{A}^{(k-1)}$. Step k changes only rows $i > k$, replacing row i by row i minus a multiple of row k. In columns $j < k$, the entry of row k lies below the diagonal ($k > j$), so it is 0 by the induction hypothesis, and the entries of row i in these columns stay 0. In column k, the new entries below the diagonal are 0, as noted after the algorithm. This completes the induction, and for $k = n - 1$ it says that $\mathbf{U}$ is upper triangular.

Step 2: each step is a matrix. Step k performs the operations "add $-\ell_{ik}$ times row k to row i" for $i = k + 1, \dots, n$. By Theorem 3.4.1, applying them one after another multiplies $\mathbf{A}^{(k-1)}$ on the left by the corresponding elementary matrices $\mathbf{I} - \ell_{ik}\mathbf{E}_{ik}$. By Lemma 3.6.3 (a), their product is $\mathbf{G}_k(-\boldsymbol{\ell}_k)$. Hence $\mathbf{A}^{(k)} = \mathbf{G}_k(-\boldsymbol{\ell}_k)\mathbf{A}^{(k-1)}$, and therefore

$$\mathbf{U} = \mathbf{G}_{n-1}(-\boldsymbol{\ell}_{n-1}) \cdots \mathbf{G}_2(-\boldsymbol{\ell}_2)\, \mathbf{G}_1(-\boldsymbol{\ell}_1)\, \mathbf{A}.$$

Step 3: undo the steps. Multiply both sides on the left by $\mathbf{G}_{n-1}(\boldsymbol{\ell}_{n-1})$, then by $\mathbf{G}_{n-2}(\boldsymbol{\ell}_{n-2})$, and so on down to $\mathbf{G}_1(\boldsymbol{\ell}_1)$. By Lemma 3.6.3 (b), each one cancels the factor next to it, and we get

$$\mathbf{G}_1(\boldsymbol{\ell}_1)\, \mathbf{G}_2(\boldsymbol{\ell}_2) \cdots \mathbf{G}_{n-1}(\boldsymbol{\ell}_{n-1})\, \mathbf{U} = \mathbf{A}.$$

By Lemma 3.6.3 (c), the product on the left of $\mathbf{U}$ is $\mathbf{I} + \sum_{k=1}^{n-1} \boldsymbol{\ell}_k \mathbf{e}_k^{\top}$. Its (i,k)-entry for $i > k$ is ℓ_{ik}; its diagonal entries are 1; its entries above the diagonal are 0. That is, it is Ł. Hence $\mathbf{A} = \text{Ł}\mathbf{U}$ with Ł unit lower triangular and $\mathbf{U}$ upper triangular, as claimed.

■

So the multipliers are not something to be remembered separately: they **are** the lower triangular factor. When a computer performs elimination it stores each ℓ_{ik} in the slot (i,k) that it has just cleared, and at the end the one array holds Ł below the diagonal and $\mathbf{U}$ on and above it.

Warning. **The matrix that performs elimination, Ł^{-1}, is not Ł with its multipliers negated.** In Example (An LU factorization from elimination) the operations multiply $\mathbf{A}$ on the left by

$$\mathbf{G}_2(-\boldsymbol{\ell}_2)\, \mathbf{G}_1(-\boldsymbol{\ell}_1) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -3 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 7 & -3 & 1 \end{pmatrix},$$

and the entry $7 = 1 + (-3)(-2)$ mixes two multipliers. This matrix is Ł^{-1}, by Step 3 of the proof. It is Ł, whose factors come in the order $\mathbf{G}_1, \mathbf{G}_2$, that has the multipliers cleanly in place.

Quick check. Find the LU factorization of $\mathbf{A} = \begin{pmatrix} 2 & 4 \\ 6 & 5 \end{pmatrix}$ over $\mathbb{Q}$ by elimination, and check it by multiplying.

Solution. The pivot is 2 and the multiplier is $\ell_{21} = 6/2 = 3$. Subtracting 3 times row 1 from row 2 gives $\mathbf{U} = \begin{pmatrix} 2 & 4 \\ 0 & 5-12 \end{pmatrix} = \begin{pmatrix} 2 & 4 \\ 0 & -7 \end{pmatrix}$, and $Ł = \begin{pmatrix} 1 & 0 \\ 3 & 1 \end{pmatrix}$. Check: $Ł\mathbf{U} = \begin{pmatrix} 2 & 4 \\ 6 & 12-7 \end{pmatrix} = \mathbf{A}$.

3.6.5 Solving with an LU factorization

Once $\mathbf{A} = Ł\mathbf{U}$ is known, each system $\mathbf{Ax} = \mathbf{b}$ costs two substitutions.

Example. Two right-hand sides, one factorization

Let $\mathbf{A} = \begin{pmatrix} 2 & 1 & -1 \\ 4 & 5 & -1 \\ -2 & 8 & 2 \end{pmatrix}$ with the factorization $\mathbf{A} = Ł\mathbf{U}$ of Example (An LU factorization from elimination). Solve $\mathbf{Ax} = \mathbf{b}$ for

$$\mathbf{b} = \begin{pmatrix} -1 \\ -3 \\ -6 \end{pmatrix} \quad \text{and} \quad \mathbf{b}' = \begin{pmatrix} 3 \\ 7 \\ -2 \end{pmatrix}.$$

Solution. First right-hand side. Forward substitution in $Ł\mathbf{y} = \mathbf{b}$:

$$\begin{aligned} y_1 &= -1, \\ 2y_1 + y_2 &= -3 \Rightarrow y_2 = -1, \\ -y_1 + 3y_2 + y_3 &= -6 \Rightarrow y_3 = -6 - 1 + 3 = -4. \end{aligned}$$

Back substitution in $\mathbf{Ux} = \mathbf{y}$:

$$\begin{aligned} -2x_3 &= -4 \Rightarrow x_3 = 2, \\ 3x_2 + x_3 &= -1 \Rightarrow x_2 = -1, \\ 2x_1 + x_2 - x_3 &= -1 \Rightarrow x_1 = 1. \end{aligned}$$

So $\mathbf{x} = (1, -1, 2)$. Check: $\mathbf{Ax} = (2-1-2,\ 4-5-2,\ -2-8+4) = (-1, -3, -6) = \mathbf{b}$.

Second right-hand side. $Ł\mathbf{y} = \mathbf{b}'$: $y_1 = 3$, $y_2 = 7 - 2\cdot 3 = 1$, $y_3 = -2 + y_1 - 3y_2 = -2+3-3 = -2$. Then $\mathbf{Ux} = \mathbf{y}$: $-2x_3 = -2$, so $x_3 = 1$; $3x_2 + 1 = 1$, so $x_2 = 0$; $2x_1 + 0 - 1 = 3$, so $x_1 = 2$. So $\mathbf{x}' = (2, 0, 1)$, and indeed $\mathbf{Ax}' = (4+0-1,\ 8+0-1,\ -4+0+2) = (3, 7, -2)$.

Since the diagonal of $\mathbf{U}$ is $2, 3, -2$, all nonzero, $\mathbf{U}$ is invertible by Lemma 3.6.2, Ł is invertible by Lemma 3.6.1, and so $\mathbf{A} = Ł\mathbf{U}$ is invertible. Each solution found is therefore the only one.

How much is saved? Count the arithmetic operations $(+, -, \times, \div)$ for an $n \times n$ matrix. At step k of elimination there are $n-k$ rows below the pivot. Each needs one division for its multiplier, then $n-k$ multiplications and $n-k$ subtractions in the columns to the right of column k. With $j = n-k$, the total is

$$\sum_{j=1}^{n-1} j(2j+1) = \frac{n(n-1)(4n+1)}{6} \approx \frac{2}{3}n^3.$$

Forward substitution with a unit lower triangular Ł uses $n(n-1)$ operations, and back substitution n^2, so a pair of triangular solves costs $2n^2 - n$. For $n = 1000$ that is about 6.7×10^8 operations for the factorization, done once, against about 2×10^6 for each new right-hand side.

3.6.6 Uniqueness

The factorization of Example (An LU factorization from elimination) came from one particular procedure. Could another procedure give a different pair $(Ł, \mathbf{U})$? For invertible matrices, no.

Theorem 3.6.5. Uniqueness of the LU factorization

Let $\mathbf{A} \in M_n(F)$ be **invertible**. If $\mathbf{A} = Ł_1\mathbf{U}_1$ and $\mathbf{A} = Ł_2\mathbf{U}_2$ are LU factorizations, then $Ł_1 = Ł_2$ and $\mathbf{U}_1 = \mathbf{U}_2$.

Idea. Take the difference in multiplicative form: from $Ł_1\mathbf{U}_1 = Ł_2\mathbf{U}_2$ we get $(Ł_2^{-1}Ł_1)\mathbf{U}_1 = \mathbf{U}_2$. The matrix $\mathbf{M} = Ł_2^{-1}Ł_1$ is unit lower triangular, and multiplying $\mathbf{U}_1$ by it must leave an upper triangular matrix. Clearing below the diagonal column by column, the nonzero pivots of $\mathbf{U}_1$ force $\mathbf{M}$ to have nothing below its diagonal. Invertibility of $\mathbf{A}$ is what provides those nonzero pivots.

Proof. By Lemma 3.6.1 (b), $Ł_1$ and $Ł_2$ are invertible. Hence $\mathbf{U}_1 = Ł_1^{-1}\mathbf{A}$ is a product of invertible matrices, so it is invertible by Theorem 0.9.7, and by Lemma 3.6.2 its diagonal entries $u_{11}, \dots, u_{nn}$ are all nonzero.

Let $\mathbf{M} = Ł_2^{-1}Ł_1$. By Lemma 3.6.1, $\mathbf{M} = (m_{ij})$ is unit lower triangular. Multiplying $Ł_1\mathbf{U}_1 = Ł_2\mathbf{U}_2$ on the left by $Ł_2^{-1}$ gives $\mathbf{M}\mathbf{U}_1 = \mathbf{U}_2$, so $\mathbf{M}\mathbf{U}_1$ is upper triangular. We show $m_{ij} = 0$ for all $i > j$, by induction on the column j.

Fix j, and suppose $m_{ik} = 0$ whenever $k < j$ and $i > k$. Let $i > j$. Since $\mathbf{M}\mathbf{U}_1$ is upper triangular and $i > j$,

$$0 = (\mathbf{M}\mathbf{U}_1)_{ij} = \sum_{k=1}^{n} m_{ik}u_{kj} = \sum_{k \le j} m_{ik}u_{kj} = m_{ij}u_{jj},$$

where the third equality uses $u_{kj} = 0$ for $k > j$, and the last uses $m_{ik} = 0$ for $k < j$ (here $i > j > k$). Since $u_{jj} \neq 0$, we get $m_{ij} = 0$. This completes the induction, so $\mathbf{M}$ is lower triangular with 1's on the diagonal and nothing below it: $\mathbf{M} = \mathbf{I}$.

Hence $Ł_2^{-1}Ł_1 = \mathbf{I}$, and multiplying by $Ł_2$ gives $Ł_1 = Ł_2$. Then $\mathbf{U}_1 = Ł_1^{-1}\mathbf{A} = Ł_2^{-1}\mathbf{A} = \mathbf{U}_2$. This proves the theorem. ∎

Two consequences are worth saying. The LU factorization of an invertible matrix does not depend on how it was found, so it is legitimate to speak of **the** factors Ł and $\mathbf{U}$. And if elimination without row swaps succeeds on an invertible $\mathbf{A}$, then every pivot is nonzero: row k is not changed after step k, so the pivot of step k is u_{kk}, which is nonzero by the first paragraph of the proof. So the second case of the algorithm (a zero pivot over a zero column) never occurs.

Remark. The hypothesis "invertible" cannot be dropped. For every $c \in F$,

$$\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ c & 1 \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix},$$

since the first column of the right factor is zero. Taking $c = 0$ and $c = 1$ gives two different LU factorizations. The proof breaks exactly at "$u_{jj} \neq 0$", here for $j = 1$.

3.6.7 Row swaps and the PLU factorization

Example (An invertible matrix with no LU factorization) showed that some matrices, even invertible ones, have no LU factorization, because elimination meets a zero pivot with a nonzero entry below it. The cure in Gaussian elimination is to swap rows. We now show that the swaps can all be collected in front, as a single matrix that rearranges the rows of **A before** elimination starts. After that rearrangement, no swaps are needed.

Matrices that rearrange rows were named in Definition 3.4.2: $\mathbf{P}_\sigma$ has j-th column $\mathbf{e}_{\sigma(j)}$, a product of permutation matrices is one, $\mathbf{P}_\sigma^{-1} = \mathbf{P}_\sigma^{\top}$, and a transposition gives the row-swap elementary matrix (Lemma 3.4.7). One more fact about them is needed here and nowhere else, because it is about the Gauss transforms of this section: a swap that leaves the first k coordinates alone can be pushed through $\mathbf{G}_k$.

Lemma 3.6.6. Pushing a Swap Past a Gauss Transform

Let $1 \leq k < r < s \leq n$, let $\tau = (r\ s)$, and let $\mathbf{v} \in F^n$ have its first k entries zero. Then $\mathbf{P}_\tau \mathbf{v}$ has its first k entries zero, and

$$\mathbf{P}_\tau\, \mathbf{G}_k(\mathbf{v})\, \mathbf{P}_\tau = \mathbf{G}_k(\mathbf{P}_\tau \mathbf{v}).$$

Proof. Write $\mathbf{P} = \mathbf{P}_\tau$. Its effect on a vector is to exchange entries r and s, which are both beyond position k, so $\mathbf{P}\mathbf{v}$ still has its first k entries zero. Since $k \neq r, s$, we have $\mathbf{P}\mathbf{e}_k = \mathbf{e}_{\tau(k)} = \mathbf{e}_k$. By Lemma 3.4.7 (b) and (c), $\mathbf{P}^{\top} = \mathbf{P}^{-1} = \mathbf{P}$, so $\mathbf{e}_k^{\top}\mathbf{P} = \mathbf{e}_k^{\top}\mathbf{P}^{\top} = (\mathbf{P}\mathbf{e}_k)^{\top} = \mathbf{e}_k^{\top}$ by Theorem 0.9.5. Therefore

$$\mathbf{P}(\mathbf{I} + \mathbf{v}\,\mathbf{e}_k^{\top})\mathbf{P} = \mathbf{P}^2 + (\mathbf{P}\mathbf{v})(\mathbf{e}_k^{\top}\mathbf{P}) = \mathbf{I} + (\mathbf{P}\mathbf{v})\,\mathbf{e}_k^{\top} = \mathbf{G}_k(\mathbf{P}\mathbf{v}),$$

using $\mathbf{P}^2 = \mathbf{I}$ from Lemma 3.4.7 (c). This proves the lemma. ∎

Definition 3.6.2. PLU factorization

Let $\mathbf{A} \in M_n(F)$. A **PLU factorization** of $\mathbf{A}$ is a triple $(\mathbf{P}, Ł, \mathbf{U})$ of matrices in $M_n(F)$ such that

$$\mathbf{PA} = Ł\mathbf{U}, \qquad \mathbf{P} \text{ is a permutation matrix},$$
$$Ł \text{ is unit lower triangular}, \qquad \mathbf{U} \text{ is upper triangular}.$$

In words: after rearranging the rows of $\mathbf{A}$ by $\mathbf{P}$, the matrix has an LU factorization. By

Lemma 3.4.7 (b) this can also be written $\mathbf{A} = \mathbf{P}^{\top}\mathbf{L}\mathbf{U}$, which is why some books write the factorization as $\mathbf{A} = \mathbf{P}\mathbf{L}\mathbf{U}$ with their $\mathbf{P}$ our $\mathbf{P}^{\top}$. For example, $\mathbf{A} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ has the PLU factorization $(\mathbf{A}, \mathbf{I}, \mathbf{I})$, since $\mathbf{A}\mathbf{A} = \mathbf{I} = \mathbf{I} \cdot \mathbf{I}$. An LU factorization is the special case $\mathbf{P} = \mathbf{I}$.

To produce $\mathbf{P}$, run elimination and allow a swap whenever a zero pivot has a nonzero entry below it.

Algorithm. **Elimination with row swaps.** Input: $\mathbf{A} \in M_n(F)$. Put $\mathbf{A}^{(0)} = \mathbf{A}$. For $k = 1, 2, \dots, n-1$, with $\mathbf{A}^{(k-1)} = (a_{ij})$:

- If $a_{ik} = 0$ for **every** $i \geq k$, set $\mathbf{P}_k = \mathbf{I}$ and all $\ell_{ik} = 0$, and let $\mathbf{A}^{(k)} = \mathbf{A}^{(k-1)}$.
- Otherwise choose some $r_k \geq k$ with $a_{r_k k} \neq 0$. Let $\mathbf{P}_k$ be the matrix swapping rows k and r_k (so $\mathbf{P}_k = \mathbf{I}$ if $r_k = k$), and let $\mathbf{B} = \mathbf{P}_k \mathbf{A}^{(k-1)}$. Set $\ell_{ik} = b_{ik}/b_{kk}$ for $i > k$, and obtain $\mathbf{A}^{(k)}$ from $\mathbf{B}$ by replacing row i with (row i) $-\ell_{ik}\cdot$ (row k) for each $i > k$.

This version never stops. Here is the result it delivers.

Theorem 3.6.7. Existence of a PLU factorization

Every $\mathbf{A} \in M_n(F)$ has a PLU factorization.

Idea. **Step roadmap.** ① As before, each step is a matrix: $\mathbf{A}^{(k)} = \mathbf{G}_k(-\boldsymbol{\ell}_k)\mathbf{P}_k\mathbf{A}^{(k-1)}$, and the final matrix $\mathbf{U}$ is upper triangular. ② The swaps are tangled with the multiplier matrices. A swap of two rows below row k can be pushed to the right past $\mathbf{G}_k$ at the price of swapping the entries of the multiplier vector: this is Lemma 3.4.7 (d). ③ Once all swaps sit next to $\mathbf{A}$, the rest is the LU proof word for word. In practice, ② says: when you swap two rows, also swap the multipliers already stored in those rows.

Proof. Run elimination with row swaps on $\mathbf{A}$, producing $\mathbf{A}^{(0)}, \dots, \mathbf{A}^{(n-1)} = \mathbf{U}$, swap matrices $\mathbf{P}_1, \dots, \mathbf{P}_{n-1}$, and multipliers ℓ_{ik}. Let $\boldsymbol{\ell}_k = \sum_{i>k} \ell_{ik}\mathbf{e}_i$, whose first k entries are zero.

Step 1. Exactly as in Step 1 of the proof of Theorem 3.6.4, every entry of $\mathbf{A}^{(k)}$ below the diagonal in columns $1, \dots, k$ is zero. The only new point is the swap in step k: it exchanges rows k and $r_k \geq k$, whose entries in columns $j < k$ are below the diagonal and hence zero, so the swap does not disturb these columns. After the swap the pivot b_{kk} is nonzero, and the new entries below it are $b_{ik} - \ell_{ik}b_{kk} = 0$. Hence $\mathbf{U}$ is upper triangular. By Lemma 3.4.7 (c), Theorem 3.4.1 and Lemma 3.6.3 (a), as in Step 2 of that proof,

$$\mathbf{A}^{(k)} = \mathbf{G}_k(-\boldsymbol{\ell}_k)\,\mathbf{P}_k\,\mathbf{A}^{(k-1)} \qquad (k = 1, \dots, n-1).$$

This also holds in the first case of the algorithm, where $\mathbf{P}_k = \mathbf{I}$ and $\mathbf{G}_k(\mathbf{0}) = \mathbf{I}$.

Step 2. For $1 \leq k \leq m \leq n-1$ put $\boldsymbol{\ell}_k^{(m)} = \mathbf{P}_m\mathbf{P}_{m-1}\cdots\mathbf{P}_{k+1}\boldsymbol{\ell}_k$, with $\boldsymbol{\ell}_m^{(m)} = \boldsymbol{\ell}_m$. We claim that the first k entries of $\boldsymbol{\ell}_k^{(m)}$ are zero and that

$$\mathbf{G}_m(-\boldsymbol{\ell}_m)\mathbf{P}_m\cdots\mathbf{G}_1(-\boldsymbol{\ell}_1)\mathbf{P}_1 = \mathbf{G}_m(-\boldsymbol{\ell}_m^{(m)})\cdots\mathbf{G}_1(-\boldsymbol{\ell}_1^{(m)})\,\mathbf{P}_m\cdots\mathbf{P}_1.$$

For $m = 1$ both sides are $\mathbf{G}_1(-\boldsymbol{\ell}_1)\mathbf{P}_1$. Suppose the claim holds for m, and multiply it on the left by $\mathbf{G}_{m+1}(-\boldsymbol{\ell}_{m+1})\mathbf{P}_{m+1}$. The matrix $\mathbf{P}_{m+1}$ is $\mathbf{I}$ or swaps rows $m+1$ and $r_{m+1} > m+1$, both beyond every $k \le m$. So by Lemma 3.4.7 (c) and Lemma 3.6.6, for each $k \le m$ and each vector $\mathbf{v}$ with its first k entries zero,

$$\mathbf{P}_{m+1}\,\mathbf{G}_k(\mathbf{v}) = \mathbf{P}_{m+1}\,\mathbf{G}_k(\mathbf{v})\,\mathbf{P}_{m+1}\mathbf{P}_{m+1} = \mathbf{G}_k(\mathbf{P}_{m+1}\mathbf{v})\,\mathbf{P}_{m+1},$$

and $\mathbf{P}_{m+1}\mathbf{v}$ again has its first k entries zero. Applying this for $k = m, m-1, \dots, 1$ moves $\mathbf{P}_{m+1}$ to the right past every factor $\mathbf{G}_k$, replacing $\boldsymbol{\ell}_k^{(m)}$ by $\mathbf{P}_{m+1}\boldsymbol{\ell}_k^{(m)} = \boldsymbol{\ell}_k^{(m+1)}$ (note $\mathbf{P}_{m+1}(-\mathbf{v}) = -\mathbf{P}_{m+1}\mathbf{v}$). This is the claim for $m+1$.

Step 3. By Step 1, $\mathbf{U} = \mathbf{G}_{n-1}(-\boldsymbol{\ell}_{n-1})\mathbf{P}_{n-1}\cdots\mathbf{G}_1(-\boldsymbol{\ell}_1)\mathbf{P}_1\mathbf{A}$. Write $\boldsymbol{\ell}'_k = \boldsymbol{\ell}_k^{(n-1)}$ and $\mathbf{P} = \mathbf{P}_{n-1}\cdots\mathbf{P}_1$. By Step 2,

$$\mathbf{U} = \mathbf{G}_{n-1}(-\boldsymbol{\ell}'_{n-1})\cdots\mathbf{G}_1(-\boldsymbol{\ell}'_1)\,\mathbf{PA}.$$

$\mathbf{P}$ is a permutation matrix by Lemma 3.4.7 (a) and (c). Multiplying on the left by $\mathbf{G}_{n-1}(\boldsymbol{\ell}'_{n-1})$, then $\mathbf{G}_{n-2}(\boldsymbol{\ell}'_{n-2})$, and so on, and using Lemma 3.6.3 (b) and (c) as in Step 3 of the proof of Theorem 3.6.4,

$$\mathbf{PA} = Ł\mathbf{U}, \qquad Ł = \mathbf{I} + \sum_{k=1}^{n-1} \boldsymbol{\ell}'_k \mathbf{e}_k^{\top},$$

where Ł is unit lower triangular because the first k entries of each $\boldsymbol{\ell}'_k$ are zero. So $(\mathbf{P}, Ł, \mathbf{U})$ is a PLU factorization of $\mathbf{A}$. This proves the theorem.

■

Example. A swap in the middle

Find a PLU factorization of

$$\mathbf{A} = \begin{pmatrix} 1 & 2 & 1 \\ 2 & 4 & 3 \\ 3 & 7 & 2 \end{pmatrix} \in M_3(\mathbb{Q}).$$

Solution. Step 1. The pivot 1 is nonzero, so no swap. The multipliers are $\ell_{21} = 2$ and $\ell_{31} = 3$:

$$\mathbf{A} \xrightarrow[R_3 \to R_3 - 3R_1]{R_2 \to R_2 - 2R_1} \begin{pmatrix} 1 & 2 & 1 \\ 0 & 0 & 1 \\ 0 & 1 & -1 \end{pmatrix}, \qquad \boldsymbol{\ell}_1 = \begin{pmatrix} 0 \\ 2 \\ 3 \end{pmatrix}.$$

Step 2. The pivot position $(2,2)$ holds 0, but the entry below it is 1. Elimination without row swaps would stop here. We swap rows 2 and 3, so $\mathbf{P}_2 = \mathbf{P}_{23}$:

$$\xrightarrow{R_2 \leftrightarrow R_3} \begin{pmatrix} 1 & 2 & 1 \\ 0 & 1 & -1 \\ 0 & 0 & 1 \end{pmatrix} = \mathbf{U}.$$

The entry below the new pivot is already 0, so $\ell_{32} = 0$, and $\mathbf{U}$ is upper triangular.

By Step 2 of the proof, the stored multipliers of step 1 are swapped too: $\boldsymbol{\ell}'_1 = \mathbf{P}_2\boldsymbol{\ell}_1 = (0, 3, 2)$.

Hence

$$\mathbf{P} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}, \qquad Ł = \begin{pmatrix} 1 & 0 & 0 \\ 3 & 1 & 0 \\ 2 & 0 & 1 \end{pmatrix}.$$

Check: $\mathbf{PA}$ is $\mathbf{A}$ with rows 2 and 3 swapped, and

$$Ł\mathbf{U} = \begin{pmatrix} 1 & 2 & 1 \\ 3 & 6+1 & 3-1 \\ 2 & 4 & 2+1 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 1 \\ 3 & 7 & 2 \\ 2 & 4 & 3 \end{pmatrix} = \mathbf{PA}.$$

Notice where the multipliers ended up: 3 in row 2 and 2 in row 3, traveling with the rows they belong to. To solve $\mathbf{Ax} = \mathbf{b}$, multiply by $\mathbf{P}$ and solve $Ł\mathbf{Ux} = \mathbf{Pb}$ by two substitutions; since $\mathbf{P}$ is invertible, this changes no solutions. For $\mathbf{b} = (2, 3, 8)$, $\mathbf{Pb} = (2, 8, 3)$. Forward substitution in $Ł\mathbf{y} = \mathbf{Pb}$ gives $y_1 = 2$, $y_2 = 8 - 3 \cdot 2 = 2$, $y_3 = 3 - 2 \cdot 2 = -1$. Back substitution in $\mathbf{Ux} = \mathbf{y}$ gives $x_3 = -1$, $x_2 = 2 + x_3 = 1$, $x_1 = 2 - 2x_2 - x_3 = 1$. Check: $\mathbf{A}(1, 1, -1) = (1 + 2 - 1,\ 2 + 4 - 3,\ 3 + 7 - 2) = (2, 3, 8)$.

Remark. Over $\mathbb{R}$, computers do not choose just any nonzero pivot. They choose the entry of **largest absolute value** in the column, which is called **partial pivoting**. The reason is rounding. For $\mathbf{A} = \begin{pmatrix} \varepsilon & 1 \\ 1 & 1 \end{pmatrix}$ with $\varepsilon = 10^{-20}$, elimination without swaps gives the multiplier 10^{20} and $u_{22} = 1 - 10^{20}$, which in standard double-precision arithmetic is stored as -10^{20}. Multiplying back, the computed $Ł\mathbf{U}$ has $(2, 2)$-entry $10^{20} - 10^{20} = 0$ instead of 1. Swapping the rows first gives the harmless multiplier ε. Chapter 24 makes this precise.

3.6.8 Symmetric matrices: the LDL$^\top$ factorization

When $\mathbf{A}$ is symmetric, the two factors of $\mathbf{A} = Ł\mathbf{U}$ are secretly related: up to a diagonal matrix, $\mathbf{U}$ is the transpose of Ł. In Example (An LU factorization from elimination) the factors look unrelated, but that matrix is not symmetric. Try a symmetric one:

$$\begin{pmatrix} 1 & 3 & -2 \\ 3 & 7 & -8 \\ -2 & -8 & 5 \end{pmatrix} \xrightarrow[R_3 \to R_3 + 2R_1]{R_2 \to R_2 - 3R_1} \begin{pmatrix} 1 & 3 & -2 \\ 0 & -2 & -2 \\ 0 & -2 & 1 \end{pmatrix} \xrightarrow{R_3 \to R_3 - R_2} \begin{pmatrix} 1 & 3 & -2 \\ 0 & -2 & -2 \\ 0 & 0 & 3 \end{pmatrix} = \mathbf{U},$$

with multipliers $\ell_{21} = 3$, $\ell_{31} = -2$, $\ell_{32} = 1$. Divide each row of $\mathbf{U}$ by its diagonal entry: the rows become $(1, 3, -2)$, $(0, 1, 1)$, $(0, 0, 1)$. These are the columns of $Ł = \begin{pmatrix} 1 & 0 & 0 \\ 3 & 1 & 0 \\ -2 & 1 & 1 \end{pmatrix}$, read as rows. So $\mathbf{U} = \mathbf{D}Ł^\top$ with $\mathbf{D} = \operatorname{diag}(1, -2, 3)$, and $\mathbf{A} = Ł\mathbf{D}Ł^\top$.

Theorem 3.6.8. LDL$^\top$ factorization

Let $\mathbf{A} \in M_n(F)$ be **symmetric** and **invertible**, and suppose $\mathbf{A} = Ł\mathbf{U}$ is an LU factorization. Then the diagonal entries $u_{11}, \ldots, u_{nn}$ of $\mathbf{U}$ are nonzero, and with $\mathbf{D} = \operatorname{diag}(u_{11}, \ldots, u_{nn})$,

$$\mathbf{U} = \mathbf{D}Ł^\top \qquad \text{and} \qquad \mathbf{A} = Ł\mathbf{D}Ł^\top.$$

Idea. Pull the diagonal out of $\mathbf{U}$, writing $\mathbf{U} = \mathbf{D}\mathbf{V}$ with $\mathbf{V}$ unit upper triangular, so $\mathbf{A} = Ł\mathbf{D}\mathbf{V}$. Transposing, $\mathbf{A} = \mathbf{A}^\top = \mathbf{V}^\top(\mathbf{D}Ł^\top)$, which is a second LU factorization of the same invertible matrix. Uniqueness forces the two to agree, so $Ł = \mathbf{V}^\top$.

Proof. As in the proof of Theorem 3.6.5, $\mathbf{U} = Ł^{-1}\mathbf{A}$ is invertible, so its diagonal entries are nonzero by Lemma 3.6.2. Hence $\mathbf{D}$ is invertible, with $\mathbf{D}^{-1} = \operatorname{diag}(u_{11}^{-1}, \dots, u_{nn}^{-1})$. Let $\mathbf{V} = \mathbf{D}^{-1}\mathbf{U}$. Row i of $\mathbf{V}$ is u_{ii}^{-1} times row i of $\mathbf{U}$, so $\mathbf{V}$ is upper triangular with $v_{ii} = u_{ii}^{-1}u_{ii} = 1$, and $\mathbf{A} = Ł\mathbf{D}\mathbf{V}$.

Since $\mathbf{A}$ is symmetric and $\mathbf{D}^\top = \mathbf{D}$, Theorem 0.9.5 gives

$$\mathbf{A} = \mathbf{A}^\top = (Ł\mathbf{D}\mathbf{V})^\top = \mathbf{V}^\top\mathbf{D}Ł^\top = \mathbf{V}^\top(\mathbf{D}Ł^\top).$$

Here $\mathbf{V}^\top$ is unit lower triangular, because $\mathbf{V}$ is unit upper triangular. And $\mathbf{D}Ł^\top$ is upper triangular: its (i,j)-entry is $u_{ii}\ell_{ji}$, and $\ell_{ji} = 0$ when $j < i$, since Ł is lower triangular. So $(\mathbf{V}^\top, \mathbf{D}Ł^\top)$ is an LU factorization of the invertible matrix $\mathbf{A}$. By Theorem 3.6.5, $\mathbf{V}^\top = Ł$ and $\mathbf{D}Ł^\top = \mathbf{U}$. Therefore $\mathbf{A} = Ł\mathbf{U} = Ł\mathbf{D}Ł^\top$, as claimed.

■

The factorization stores about half as much: Ł and the diagonal $\mathbf{D}$ determine everything. Over $\mathbb{R}$, when every entry of $\mathbf{D}$ is positive, one can split $\mathbf{D} = \mathbf{D}^{1/2}\mathbf{D}^{1/2}$ and write $\mathbf{A} = (Ł\mathbf{D}^{1/2})(Ł\mathbf{D}^{1/2})^\top$; this is the **Cholesky factorization**, which Chapter 13 connects with positive definite matrices.

Remark. Symmetry alone does not give an LU factorization: $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ is symmetric and invertible, and has none by Example (An invertible matrix with no LU factorization). The theorem assumes the factorization exists and then improves it.

3.6.9 Exercises

3.6.9.1 A. Check your understanding

Exercise. A1

(a) Define an LU factorization of a matrix $\mathbf{A} \in M_n(F)$, and a PLU factorization.

(b) True or false: every invertible matrix has an LU factorization. Justify your answer.

(c) Suppose $\mathbf{A} = Ł\mathbf{U}$. In which order are the two triangular systems solved to find a solution of $\mathbf{A}\mathbf{x} = \mathbf{b}$, and what is each one?

(d) True or false: the product of two unit lower triangular $n \times n$ matrices is unit lower triangular. Justify your answer.

(e) True or false: the zero matrix in $M_2(F)$ has more than one LU factorization. Justify your answer.

(f) What is the inverse of a permutation matrix $\mathbf{P}$?

Solution.

(a) See Definition 3.6.1: $\mathbf{A} = \text{Ł}\mathbf{U}$ with Ł unit lower triangular and $\mathbf{U}$ upper triangular. See Definition 3.6.2: $\mathbf{PA} = \text{Ł}\mathbf{U}$ with $\mathbf{P}$ a permutation matrix, Ł unit lower triangular, and $\mathbf{U}$ upper triangular.

(b) False. $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ is invertible but has no LU factorization, by Example (An invertible matrix with no LU factorization).

(c) First $\text{Ł}\mathbf{y} = \mathbf{b}$, by forward substitution; then $\mathbf{U}\mathbf{x} = \mathbf{y}$, by back substitution. Then $\mathbf{A}\mathbf{x} = \text{Ł}(\mathbf{U}\mathbf{x}) = \text{Ł}\mathbf{y} = \mathbf{b}$.

(d) True, by Lemma 3.6.1 (a).

(e) True. For every $c \in F$, $\begin{pmatrix} 1 & 0 \\ c & 1 \end{pmatrix} \mathbf{O} = \mathbf{O}$, so $\left(\begin{pmatrix} 1 & 0 \\ c & 1 \end{pmatrix}, \mathbf{O} \right)$ is an LU factorization; $c = 0$ and $c = 1$ give two different ones. This does not contradict Theorem 3.6.5, because $\mathbf{O}$ is not invertible.

(f) $\mathbf{P}^{-1} = \mathbf{P}^{\top}$, by Lemma 3.4.7 (b).

3.6.9.2 B. Practice

Exercise. B1: An LU factorization

Find the LU factorization of

$$\mathbf{A} = \begin{pmatrix} 1 & 2 & 0 \\ -3 & -7 & 4 \\ 2 & 6 & -3 \end{pmatrix} \in M_3(\mathbb{Q})$$

by elimination without row swaps. Hence show that $\mathbf{A}$ is invertible.

Solution. Step 1. The pivot is 1, with multipliers $\ell_{21} = -3$ and $\ell_{31} = 2$:

$$\mathbf{A} \xrightarrow[R_3 \to R_3 - 2R_1]{R_2 \to R_2 - (-3)R_1} \begin{pmatrix} 1 & 2 & 0 \\ 0 & -1 & 4 \\ 0 & 2 & -3 \end{pmatrix}.$$

Step 2. The pivot is -1, with multiplier $\ell_{32} = 2/(-1) = -2$:

$$\xrightarrow{R_3 \to R_3 - (-2)R_2} \begin{pmatrix} 1 & 2 & 0 \\ 0 & -1 & 4 \\ 0 & 0 & 5 \end{pmatrix} = \mathbf{U}.$$

By Theorem 3.6.4,

$$\text{Ł} = \begin{pmatrix} 1 & 0 & 0 \\ -3 & 1 & 0 \\ 2 & -2 & 1 \end{pmatrix}, \qquad \mathbf{A} = \text{Ł}\mathbf{U}.$$

Check: $\text{Ł}\mathbf{U} = \begin{pmatrix} 1 & 2 & 0 \\ -3 & -6-1 & 4 \\ 2 & 4+2 & -8+5 \end{pmatrix} = \mathbf{A}$.

The diagonal entries $1, -1, 5$ of $\mathbf{U}$ are nonzero, so $\mathbf{U}$ is invertible by Lemma 3.6.2, and Ł is invertible by Lemma 3.6.1. Hence $\mathbf{A} = \text{Ł}\mathbf{U}$ is invertible by Theorem 0.9.7.

Exercise. B2: Two systems, one factorization

With $\mathbf{A} = \text{Ł}\mathbf{U}$ as in Exercise (B1: An LU factorization), solve $\mathbf{Ax} = \mathbf{b}$ for $\mathbf{b} = (3, -6, 5)$ and for $\mathbf{b} = (0, 1, -2)$, using only forward and back substitution.

Solution. *For* $\mathbf{b} = (3, -6, 5)$. $\text{Ł}\mathbf{y} = \mathbf{b}$: $y_1 = 3$; $-3y_1 + y_2 = -6$, so $y_2 = 3$; $2y_1 - 2y_2 + y_3 = 5$, so $y_3 = 5 - 6 + 6 = 5$. Then $\mathbf{Ux} = \mathbf{y}$: $5x_3 = 5$, so $x_3 = 1$; $-x_2 + 4x_3 = 3$, so $x_2 = 1$; $x_1 + 2x_2 = 3$, so $x_1 = 1$. Hence $\mathbf{x} = (1, 1, 1)$. Check: $\mathbf{Ax} = (1 + 2,\ -3 - 7 + 4,\ 2 + 6 - 3) = (3, -6, 5)$.

For $\mathbf{b} = (0, 1, -2)$. $\text{Ł}\mathbf{y} = \mathbf{b}$: $y_1 = 0$; $y_2 = 1 + 3y_1 = 1$; $y_3 = -2 - 2y_1 + 2y_2 = 0$. Then $\mathbf{Ux} = \mathbf{y}$: $5x_3 = 0$, so $x_3 = 0$; $-x_2 + 0 = 1$, so $x_2 = -1$; $x_1 + 2(-1) = 0$, so $x_1 = 2$. Hence $\mathbf{x} = (2, -1, 0)$. Check: $\mathbf{Ax} = (2 - 2,\ -6 + 7,\ 4 - 6) = (0, 1, -2)$.

Since $\mathbf{A}$ is invertible (Exercise (B1: An LU factorization)), each solution is the only one, by Theorem 0.9.7.

Exercise. B3: A PLU factorization

Let $\mathbf{A} = \begin{pmatrix} 0 & 1 & 2 \\ 1 & 1 & 1 \\ 2 & 3 & 1 \end{pmatrix} \in M_3(\mathbb{Q})$.

(a) Explain why $\mathbf{A}$ has no LU factorization.

(b) Find a PLU factorization of $\mathbf{A}$.

(c) Hence solve $\mathbf{Ax} = (0, 2, 7)$.

Solution.

(a) If $\mathbf{A} = \text{Ł}\mathbf{U}$, then the $(1,1)$-entry gives $u_{11} = a_{11} = 0$, and the $(2,1)$-entry gives $a_{21} = \ell_{21}u_{11} = 0$. But $a_{21} = 1 \neq 0$, a contradiction, exactly as in Example (An invertible matrix with no LU factorization).

(b) The first pivot position holds 0 and the entry below is 1, so swap rows 1 and 2, $\mathbf{P}_1 = \mathbf{P}_{12}$:

$$\mathbf{A} \xrightarrow{R_1 \leftrightarrow R_2} \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \\ 2 & 3 & 1 \end{pmatrix} \xrightarrow[R_3 \to R_3 - 2R_1]{R_2 \to R_2 - 0R_1} \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \\ 0 & 1 & -1 \end{pmatrix}$$

$$\xrightarrow{R_3 \to R_3 - 1R_2} \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \\ 0 & 0 & -3 \end{pmatrix} = \mathbf{U}.$$

The multipliers are $\ell_{21} = 0$, $\ell_{31} = 2$, $\ell_{32} = 1$, and no swap happens after step 1, so they need no rearranging. Hence

$$\mathbf{P} = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \qquad Ł = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 2 & 1 & 1 \end{pmatrix}.$$

Check: $\mathbf{PA} = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \\ 2 & 3 & 1 \end{pmatrix}$, and $Ł\mathbf{U} = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \\ 2 & 2+1 & 2+2-3 \end{pmatrix} = \mathbf{PA}$.

(c) $\mathbf{Ax} = \mathbf{b}$ holds if and only if $\mathbf{PAx} = \mathbf{Pb}$, since $\mathbf{P}$ is invertible (Lemma 3.4.7 (b)). Swapping the first two entries of $\mathbf{b} = (0, 2, 7)$, we solve $Ł\mathbf{Ux} = \mathbf{Pb} = (2, 0, 7)$. $Ł\mathbf{y} = (2, 0, 7)$: $y_1 = 2$, $y_2 = 0$, $y_3 = 7 - 2 \cdot 2 - 0 = 3$. $\mathbf{Ux} = \mathbf{y}$: $-3x_3 = 3$, so $x_3 = -1$; $x_2 + 2x_3 = 0$, so $x_2 = 2$; $x_1 + x_2 + x_3 = 2$, so $x_1 = 1$. Hence $\mathbf{x} = (1, 2, -1)$. Check: $\mathbf{Ax} = (0 + 2 - 2,\ 1 + 2 - 1,\ 2 + 6 - 1) = (0, 2, 7)$.

3.6.9.3 C. Going deeper

Exercise. C1: Leading principal submatrices

For $\mathbf{A} \in M_n(F)$ and $1 \le k \le n$, the **leading principal submatrix** $\mathbf{A}_k \in M_k(F)$ is the top-left $k \times k$ corner of $\mathbf{A}$, with entries a_{ij} for $i, j \le k$.

(a) Let Ł be lower triangular and $\mathbf{U}$ upper triangular in $M_n(F)$. Show that $(Ł\mathbf{U})_k = Ł_k\mathbf{U}_k$ for every k.

(b) Deduce that if $\mathbf{A}$ has an LU factorization $\mathbf{A} = Ł\mathbf{U}$ with $\mathbf{U}$ invertible, then every leading principal submatrix $\mathbf{A}_1, \dots, \mathbf{A}_n$ is invertible.

(c) Prove the converse: if $\mathbf{A}_1, \dots, \mathbf{A}_n$ are all invertible, then $\mathbf{A}$ has an LU factorization with $\mathbf{U}$ invertible.

Hint for (c): induction on n. Write $\mathbf{A}$ with $\mathbf{A}_{n-1}$ in its top-left corner, a column $\mathbf{b}$, a row $\mathbf{c}^\top$, and a corner entry d, and look for Ł and $\mathbf{U}$ of the same shape.

Solution.

(a) Let $i, j \le k$. By Definition 0.9.9, $(Ł\mathbf{U})_{ij} = \sum_{l=1}^{n} \ell_{il} u_{lj}$. If $l > k$, then $l > i$, so $\ell_{il} = 0$ because Ł is lower triangular. Hence $(Ł\mathbf{U})_{ij} = \sum_{l=1}^{k} \ell_{il} u_{lj} = (Ł_k\mathbf{U}_k)_{ij}$, as claimed.

(b) Suppose $\mathbf{A} = Ł\mathbf{U}$ with $\mathbf{U}$ invertible. By (a), $\mathbf{A}_k = Ł_k\mathbf{U}_k$. The matrix $Ł_k$ is unit lower triangular, so it is invertible by Lemma 3.6.1. By Lemma 3.6.2, $u_{11}, \dots, u_{nn}$ are nonzero; $\mathbf{U}_k$ is upper triangular with diagonal $u_{11}, \dots, u_{kk}$, so $\mathbf{U}_k$ is invertible by the same lemma. Hence $\mathbf{A}_k$ is invertible by Theorem 0.9.7.

(c) We prove by induction on n that every $\mathbf{A} \in M_n(F)$ whose leading principal submatrices are all invertible has an LU factorization with $\mathbf{U}$ invertible. For $n = 1$, $\mathbf{A} = (a_{11})$ with $a_{11} \neq 0$, and $\mathbf{A} = (1)(a_{11})$ works.
Let $n \ge 2$ and suppose the statement holds for $n - 1$. Write

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{n-1} & \mathbf{b} \\ \mathbf{c}^\top & d \end{pmatrix}, \qquad \mathbf{b}, \mathbf{c} \in F^{n-1},\ d \in F.$$

The leading principal submatrices of $\mathbf{A}_{n-1}$ are $\mathbf{A}_1, \dots, \mathbf{A}_{n-1}$, all invertible, so by the

induction hypothesis $\mathbf{A}_{n-1} = Ł'\mathbf{U}'$ with $Ł'$ unit lower triangular and $\mathbf{U}'$ upper triangular and invertible. Put $\mathbf{y} = Ł'^{-1}\mathbf{b}$, $\mathbf{x}^\top = \mathbf{c}^\top \mathbf{U}'^{-1}$, and $z = d - \mathbf{x}^\top \mathbf{y}$, and let

$$Ł = \begin{pmatrix} Ł' & \mathbf{0} \\ \mathbf{x}^\top & 1 \end{pmatrix}, \qquad \mathbf{U} = \begin{pmatrix} \mathbf{U}' & \mathbf{y} \\ \mathbf{0}^\top & z \end{pmatrix}.$$

Then $Ł$ is unit lower triangular and $\mathbf{U}$ is upper triangular. Computing $Ł\mathbf{U}$ entry by entry, splitting each sum $\sum_{l=1}^{n}$ into $l \le n-1$ and $l = n$: the top-left block is $Ł'\mathbf{U}' + \mathbf{0}\,\mathbf{0}^\top = \mathbf{A}_{n-1}$; the top-right column is $Ł'\mathbf{y} + \mathbf{0}z = \mathbf{b}$; the bottom-left row is $\mathbf{x}^\top \mathbf{U}' + 1 \cdot \mathbf{0}^\top = \mathbf{c}^\top$; the corner is $\mathbf{x}^\top \mathbf{y} + z = d$. Hence $\mathbf{A} = Ł\mathbf{U}$.
Finally $\mathbf{A} = \mathbf{A}_n$ is invertible by hypothesis, and $Ł$ is invertible by Lemma 3.6.1, so $\mathbf{U} = Ł^{-1}\mathbf{A}$ is invertible by Theorem 0.9.7. This completes the induction.

In the language of elimination: z is the n-th pivot, and (c) shows that invertible leading corners are exactly what keeps every pivot nonzero.

Exercise. C2: An $\mathbf{LDL}^\top$ factorization

Let $\mathbf{A} = \begin{pmatrix} 4 & -4 & 8 \\ -4 & 5 & -11 \\ 8 & -11 & 27 \end{pmatrix} \in M_3(\mathbb{R})$.

(a) Find the LU factorization of $\mathbf{A}$, show that $\mathbf{A}$ is invertible, and write $\mathbf{A} = Ł\mathbf{D}Ł^\top$.

(b) Deduce that $\mathbf{x}^\top \mathbf{A}\mathbf{x} > 0$ for every nonzero $\mathbf{x} \in \mathbb{R}^3$.

Solution.

(a) Step 1: pivot 4, multipliers $\ell_{21} = -1$, $\ell_{31} = 2$. Row 2 becomes $(-4, 5, -11) + (4, -4, 8) = (0, 1, -3)$, row 3 becomes $(8, -11, 27) - 2(4, -4, 8) = (0, -3, 11)$. Step 2: pivot 1, multiplier $\ell_{32} = -3$; row 3 becomes $(0, -3, 11) + 3(0, 1, -3) = (0, 0, 2)$. Hence

$$Ł = \begin{pmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 2 & -3 & 1 \end{pmatrix}, \qquad \mathbf{U} = \begin{pmatrix} 4 & -4 & 8 \\ 0 & 1 & -3 \\ 0 & 0 & 2 \end{pmatrix},$$

and $\mathbf{A} = Ł\mathbf{U}$ by Theorem 3.6.4. The diagonal $4, 1, 2$ of $\mathbf{U}$ is nonzero, so $\mathbf{U}$ is invertible (Lemma 3.6.2), $Ł$ is invertible (Lemma 3.6.1), and $\mathbf{A} = Ł\mathbf{U}$ is invertible. $\mathbf{A}$ is symmetric, so by Theorem 3.6.8, with $\mathbf{D} = \operatorname{diag}(4, 1, 2)$,

$$\mathbf{A} = Ł\mathbf{D}Ł^\top, \qquad \mathbf{D}Ł^\top = \begin{pmatrix} 4 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{pmatrix} \begin{pmatrix} 1 & -1 & 2 \\ 0 & 1 & -3 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 4 & -4 & 8 \\ 0 & 1 & -3 \\ 0 & 0 & 2 \end{pmatrix} = \mathbf{U},$$

which confirms the theorem directly.

(b) Let $\mathbf{x} \neq \mathbf{0}$ and put $\mathbf{y} = Ł^\top \mathbf{x}$. By Lemma 3.6.1, $Ł$ is invertible, so $Ł^\top$ is invertible by Theorem 0.9.7 (4), so $\mathbf{y} \neq \mathbf{0}$ (otherwise $\mathbf{x} = (Ł^\top)^{-1}\mathbf{y} = \mathbf{0}$). By Theorem 0.9.5, $\mathbf{y}^\top = \mathbf{x}^\top Ł$, so

$$\mathbf{x}^\top \mathbf{A}\mathbf{x} = \mathbf{x}^\top Ł\mathbf{D}Ł^\top \mathbf{x} = \mathbf{y}^\top \mathbf{D}\mathbf{y} = 4y_1^2 + y_2^2 + 2y_3^2,$$

reading the 1×1 matrix as a real number. Each term is ≥ 0, and at least one $y_i \neq 0$, which makes its term strictly positive. Hence $\mathbf{x}^\top \mathbf{A}\mathbf{x} > 0$.

Exercise. C3: Counting matrices without an LU factorization

(a) Let $\mathbf{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in M_2(F)$. Prove that $\mathbf{A}$ has an LU factorization if and only if $a \neq 0$ or $c = 0$.

(b) Let q be a prime. How many matrices in $M_2(\mathbb{F}_q)$ have **no** LU factorization? List them for $q = 2$.

Solution.

(a) ($\Rightarrow$) Suppose $\mathbf{A} = \text{Ł}\mathbf{U}$ with $\text{Ł} = \begin{pmatrix} 1 & 0 \\ \ell & 1 \end{pmatrix}$ and $\mathbf{U} = \begin{pmatrix} u & v \\ 0 & w \end{pmatrix}$. As in Example (An invertible matrix with no LU factorization), $\text{Ł}\mathbf{U} = \begin{pmatrix} u & v \\ \ell u & \ell v + w \end{pmatrix}$, so $a = u$ and $c = \ell u = \ell a$. If $a = 0$, then $c = 0$. Hence $a \neq 0$ or $c = 0$.
($\Leftarrow$) If $a \neq 0$, elimination without row swaps succeeds with multiplier $\ell = c/a$, and indeed

$$\begin{pmatrix} 1 & 0 \\ c/a & 1 \end{pmatrix} \begin{pmatrix} a & b \\ 0 & d - cb/a \end{pmatrix} = \begin{pmatrix} a & b \\ c & cb/a + d - cb/a \end{pmatrix} = \mathbf{A}.$$

If $c = 0$, then $\mathbf{A}$ is upper triangular and $\mathbf{A} = \mathbf{I}\mathbf{A}$.

(b) By (a), $\mathbf{A}$ has no LU factorization exactly when $a = 0$ and $c \neq 0$. There are $q - 1$ choices of c, and q choices each of b and d, so there are $(q-1)q^2$ such matrices. For $q = 2$ this gives 4 of the 16 matrices: $a = 0$, $c = 1$, and $b, d \in \{0, 1\}$, that is,

$$\begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}, \quad \begin{pmatrix} 0 & 0 \\ 1 & 1 \end{pmatrix}, \quad \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}.$$

3.7 Applications of Linear Systems

We now have a complete engine: elimination decides whether $\mathbf{A}\mathbf{x} = \mathbf{b}$ is consistent, rank–nullity counts the free parameters, and the null space describes the freedom. This section runs the engine on four problems from outside linear algebra. In each one the modeling step is where the thinking happens: which quantities are unknown, why the constraints are linear, and over which field. Each model is also honest about what it leaves out.

3.7.1 Polynomial interpolation

Suppose we know the values of an unknown polynomial at a few points, and want the polynomial. Through two points with different x-values there is exactly one line. Through three there is exactly one parabola, possibly degenerate. The general question: given points $(x_0, y_0), \ldots, (x_n, y_n)$ in F^2, is there a polynomial $p \in F[x]_{\leq n}$ with

$$p(x_0) = y_0, \quad p(x_1) = y_1, \quad \ldots, \quad p(x_n) = y_n,$$

and is it unique?

The unknowns are the coefficients. Writing $p = c_0 + c_1x + \cdots + c_nx^n$, the condition $p(x_i) = y_i$ reads $c_0 + c_1x_i + \cdots + c_nx_i^n = y_i$, by Definition 0.8.6. The x_i and y_i are known numbers, so this is a **linear** equation in $c_0, \ldots, c_n$, even though p itself is not linear. The $n+1$ conditions form a square system

$$\mathbf{Vc} = \mathbf{y}, \qquad \mathbf{V} = \begin{pmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^n \\ 1 & x_1 & x_1^2 & \cdots & x_1^n \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^n \end{pmatrix}, \quad \mathbf{c} = \begin{pmatrix} c_0 \\ \vdots \\ c_n \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} y_0 \\ \vdots \\ y_n \end{pmatrix}.$$

By Theorem 3.4.4, this has exactly one solution for **every** $\mathbf{y}$ as soon as $\mathbf{Vc} = \mathbf{0}$ has only the trivial solution. And $\mathbf{Vc} = \mathbf{0}$ says that the polynomial with coefficients $\mathbf{c}$ has the $n+1$ roots $x_0, \ldots, x_n$. So everything hinges on a fact about polynomials: a nonzero polynomial of degree at most n cannot have $n+1$ distinct roots. Chapter 0 proved this only for degree at most 2, in Exercise (C1: Roots of a quadratic), so we prove it now in general. It takes two lemmas.

The first is long division by $x - c$, carried out one leading term at a time.

Lemma 3.7.1. Factor Theorem for a Linear Divisor

Let $p \in F[x]$ and $c \in F$. Then there is $q \in F[x]$ such that

$$p = (x - c)\, q + p(c).$$

In particular, if c is a root of p, then $p = (x - c)\, q$.

Idea. Divide as at school. If p has leading term a_dx^d, subtract $a_dx^{d-1}(x - c)$, which kills the leading term and leaves a polynomial of smaller degree. Repeat until only a constant remains. To turn "repeat" into a proof, use induction on the degree. The constant left over is found by evaluating at c, where the factor $x - c$ vanishes.

Proof. We first show that $p = (x - c)q + r$ for some $q \in F[x]$ and some constant $r \in F$, by strong induction on $\deg p$ (Theorem 0.2.10), treating $p = 0$ together with the constants. If $p = 0$ or $\deg p = 0$, then p is a constant, and $q = 0$, $r = p$ work.

Let $d \geq 1$, and suppose the statement holds for every polynomial of degree less than d, and for 0. Let $\deg p = d$ with leading coefficient a_d, and put

$$p_1 = p - a_dx^{d-1}(x - c) = p - a_dx^d + c\, a_dx^{d-1}.$$

The coefficient of x^d in p_1 is $a_d - a_d = 0$, and all higher coefficients are 0, so $p_1 = 0$ or $\deg p_1 < d$. By the induction hypothesis, $p_1 = (x - c)q_1 + r$ with $r \in F$. Hence

$$p = (x - c)(a_dx^{d-1} + q_1) + r,$$

which is the statement for p, with $q = a_dx^{d-1} + q_1$.

Now evaluate at c. By Theorem 0.8.5, $p(c) = (c - c)\, q(c) + r = 0 \cdot q(c) + r = r$. So $r = p(c)$, and if $p(c) = 0$, then $p = (x - c)q$. This proves the lemma.

■

Lemma 3.7.2. Root bound

Let $n \in \mathbb{N}$, and let $p \in F[x]$ be **nonzero** with $\deg p \leq n$. Then p has **at most** n distinct roots in F.

Proof. We use induction on n. If $n = 0$, then $p = a$ is a nonzero constant, and $p(c) = a \neq 0$ for every $c \in F$, so p has no roots.

Let $n \geq 1$, suppose the statement holds for $n-1$, and let $p \neq 0$ with $\deg p \leq n$. If p has no root, we are done. Otherwise let c be a root. By Lemma 3.7.1, $p = (x-c)q$ for some $q \in F[x]$. Then $q \neq 0$, since otherwise $p = 0$, and by Theorem 0.8.3, $\deg p = 1 + \deg q$, so $\deg q \leq n-1$. By the induction hypothesis, q has at most $n-1$ distinct roots.

Let $s \in F$ be any root of p. By Theorem 0.8.5, $0 = p(s) = (s-c)\,q(s)$, so $s - c = 0$ or $q(s) = 0$ by Theorem 0.7.1. Hence every root of p is c or a root of q, and p has at most $1 + (n-1) = n$ distinct roots. This completes the induction.

■

The field axioms did real work in the last step: a product of two nonzero elements is nonzero. Over $\mathbb{Z}/8\mathbb{Z}$, which is not a field, $x^2 - 1$ has the four roots $1, 3, 5, 7$, as Exercise (C1: Roots of a quadratic) showed. And the lemma is sharp: over $\mathbb{F}_2$, the polynomial $x^2 + x$ of Example (Different polynomials, same function) has degree 2 and the two roots 0 and 1.

The lemma also settles a question left open in Chapter 0. If F is **infinite** and the polynomial function of p is zero, then every element of F is a root of p. If p were nonzero, this would be more than $\deg p$ roots, so $p = 0$. Over an infinite field, a polynomial is therefore determined by its values: if $p(c) = q(c)$ for every $c \in F$, then $p - q$ has zero polynomial function by Theorem 0.8.5, so $p - q = 0$.

Theorem 3.7.3. Polynomial interpolation

Let $x_0, x_1, \ldots, x_n \in F$ be **distinct**, and let $y_0, y_1, \ldots, y_n \in F$ be arbitrary. Then there is **exactly one** polynomial $p \in F[x]_{\leq n}$ with $p(x_i) = y_i$ for $i = 0, 1, \ldots, n$.

Proof. A polynomial $p \in F[x]_{\leq n}$ is the same thing as its coefficient vector $\mathbf{c} = (c_0, \ldots, c_n) \in F^{n+1}$, since two polynomials are equal exactly when their coefficients agree (Definition 0.8.1). As computed above, $p(x_i) = y_i$ for all i if and only if $\mathbf{V}\mathbf{c} = \mathbf{y}$, where $\mathbf{V} \in M_{n+1}(F)$ has (i,j)-entry x_i^j (rows and columns numbered from 0 to n).

Let $\mathbf{V}\mathbf{c} = \mathbf{0}$, and let $q = c_0 + c_1 x + \cdots + c_n x^n$. Then $q(x_i) = 0$ for every i, so q has the $n+1$ distinct roots $x_0, \ldots, x_n$, while $\deg q \leq n$. By Lemma 3.7.2, q cannot be nonzero, so $q = 0$ and $\mathbf{c} = \mathbf{0}$. Hence $\mathbf{V}\mathbf{c} = \mathbf{0}$ has only the trivial solution, and $\mathbf{V}$ is invertible by Theorem 3.4.4. By Theorem 0.9.7 (7), $\mathbf{V}\mathbf{c} = \mathbf{y}$ has exactly one solution $\mathbf{c} = \mathbf{V}^{-1}\mathbf{y}$. The polynomial with these coefficients is the unique p required. This proves the theorem.

■

The proof is a model of how this chapter turns questions into systems. Existence for every right-hand side and uniqueness are both properties of the square matrix $\mathbf{V}$, and for a square matrix they come together. So we only had to prove uniqueness for the zero data, and that was a statement about roots. Chapter 6 gives an explicit formula for p, the Lagrange interpolation formula. Here is the computation by elimination.

Example. A cubic through four points

Find the polynomial $p \in \mathbb{Q}[x]_{\le 3}$ whose graph passes through $(-1, 2)$, $(0, 1)$, $(1, 2)$, and $(2, 11)$.

Solution. The x-values $-1, 0, 1, 2$ are distinct, so by Theorem 3.7.3 there is exactly one such $p = c_0 + c_1x + c_2x^2 + c_3x^3$. The conditions $p(x_i) = y_i$ give the augmented matrix below. We first swap the rows for $x = -1$ and $x = 0$, so that the simplest row leads:

$$\left(\begin{array}{cccc|c} 1 & -1 & 1 & -1 & 2 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 & 2 \\ 1 & 2 & 4 & 8 & 11 \end{array}\right) \xrightarrow{R_1 \leftrightarrow R_2} \left(\begin{array}{cccc|c} 1 & 0 & 0 & 0 & 1 \\ 1 & -1 & 1 & -1 & 2 \\ 1 & 1 & 1 & 1 & 2 \\ 1 & 2 & 4 & 8 & 11 \end{array}\right)$$

$$\xrightarrow[R_4 \to R_4 - R_1]{R_2 \to R_2 - R_1,\ R_3 \to R_3 - R_1} \left(\begin{array}{cccc|c} 1 & 0 & 0 & 0 & 1 \\ 0 & -1 & 1 & -1 & 1 \\ 0 & 1 & 1 & 1 & 1 \\ 0 & 2 & 4 & 8 & 10 \end{array}\right).$$

Next clear column 2 below the pivot -1:

$$\xrightarrow[R_4 \to R_4 + 2R_2]{R_3 \to R_3 + R_2} \left(\begin{array}{cccc|c} 1 & 0 & 0 & 0 & 1 \\ 0 & -1 & 1 & -1 & 1 \\ 0 & 0 & 2 & 0 & 2 \\ 0 & 0 & 6 & 6 & 12 \end{array}\right) \xrightarrow{R_4 \to R_4 - 3R_3} \left(\begin{array}{cccc|c} 1 & 0 & 0 & 0 & 1 \\ 0 & -1 & 1 & -1 & 1 \\ 0 & 0 & 2 & 0 & 2 \\ 0 & 0 & 0 & 6 & 6 \end{array}\right).$$

Back substitution: $6c_3 = 6$, so $c_3 = 1$; $2c_2 = 2$, so $c_2 = 1$; $-c_1 + c_2 - c_3 = 1$, so $c_1 = -1$; and $c_0 = 1$. Hence

$$p = 1 - x + x^2 + x^3.$$

Check: $p(-1) = 1+1+1-1 = 2$, $p(0) = 1$, $p(1) = 1-1+1+1 = 2$, $p(2) = 1-2+4+8 = 11$.

Warning. **The x-values must be distinct as elements of F.** If $x_0 = x_1$, the first two equations of $\mathbf{Vc} = \mathbf{y}$ have the same left-hand side, so the system is inconsistent for any data with $y_0 \neq y_1$. This depends on the field: over $\mathbb{F}_3$ the numbers -1 and 2 are the **same** element, so the four points of Example (A cubic through four points) do not have distinct x-values there. Note also that the theorem promises degree **at most** n: the three points $(0, 0), (1, 1), (2, 2)$ are interpolated by $p = x$, of degree 1.

Quick check. Find the polynomial $p \in \mathbb{F}_3[x]_{\le 2}$ with $p(0) = 1$, $p(1) = 0$ and $p(2) = 0$.

Solution. The x-values $0, 1, 2$ are distinct in $\mathbb{F}_3$, so there is exactly one such p. Since 1 and 2 are roots, try $p = a(x-1)(x-2)$, which has degree at most 2 and vanishes at 1 and 2. Then $p(0) = a \cdot (-1)(-2) = 2a$, and $2a = 1$ gives $a = 2$, since $2 \cdot 2 = 4 = 1$ in $\mathbb{F}_3$. Expanding, $p = 2(x^2 - 3x + 2) = 2(x^2 + 2) = 2x^2 + 1$. Check: $p(0) = 1$, $p(1) = 3 = 0$, $p(2) = 9 = 0$. By uniqueness, this is the answer.

What the model leaves out. Interpolation fits the data **exactly**. When the y_i are measurements with errors, an exact fit of high degree follows the noise and can swing wildly between the data points. The honest tool there is a low-degree polynomial that fits approximately, found by least squares in Chapter 11.

3.7.2 Interpolation without a matrix

Theorem 3.7.3 came out of a square system and the Invertible Matrix Theorem. Rank–Nullity, from Chapter 2, reaches the same conclusion with no matrix written down, and the two arguments are worth seeing side by side.

Example. Interpolation as a Count

Let $x_0, \dots, x_n \in F$ be distinct, and define $E\colon F[x]_{\le n} \to F^{n+1}$ by $E(p) = (p(x_0), \dots, p(x_n))$. Show that E is bijective, and recover Theorem 3.7.3 from that alone.

Solution. Linear. For $p, q \in F[x]_{\le n}$ and $c \in F$, Theorem 0.8.5 gives $(cp+q)(x_i) = c\,p(x_i) + q(x_i)$ in each entry, so $E(cp+q) = cE(p) + E(q)$, and E is linear by Theorem 2.1.3.

Check the kernel. If $E(p) = \mathbf{0}$, then p has the $n+1$ distinct roots $x_0, \dots, x_n$ while $\deg p \le n$. A nonzero polynomial of degree at most n has at most n distinct roots (Lemma 3.7.2), so $p = 0$. Hence $\ker E = \{0\}$, and E is injective (Theorem 2.2.3).

Count. $\dim F[x]_{\le n} = n+1$, since $(1, x, \dots, x^n)$ is a basis (Example (Standard Bases)), and $\dim F^{n+1} = n+1$. By Corollary 2.3.2 (e), E is bijective.

Surjectivity of E says that for every $(y_0, \dots, y_n) \in F^{n+1}$ some $p \in F[x]_{\le n}$ has $p(x_i) = y_i$ for all i; injectivity says there is only one. This is Theorem 3.7.3.

Compare them. Both prove uniqueness for the zero data from the root bound (Lemma 3.7.2). The first then invoked the Invertible Matrix Theorem for the Vandermonde matrix; here the count $n+1 = n+1$ does the same job. Neither proof of existence constructs a polynomial. That is the price and the power of counting: it tells us a solution exists without saying what it is.

3.7.3 Balancing chemical equations

When ammonia burns in oxygen over a catalyst, it produces nitric oxide and water:

$$a\,\mathrm{NH_3} + b\,\mathrm{O_2} \longrightarrow c\,\mathrm{NO} + d\,\mathrm{H_2O}.$$

The coefficients a, b, c, d count molecules. The model rests on one physical law: **atoms are neither created nor destroyed**, so each element appears equally often on both sides. For

each element this gives one linear equation:

$$\begin{aligned} \text{N:} &\quad a = c, \\ \text{H:} &\quad 3a = 2d, \\ \text{O:} &\quad 2b = c + d. \end{aligned}$$

Moving everything to the left gives a homogeneous system, so the balanced coefficient vectors form the null space of its coefficient matrix. We want positive integers, but we compute over $\mathbb{Q}$. Elimination uses only the four field operations on integer data, so it never leaves $\mathbb{Q}$, and a rational solution can be scaled to an integer one at the end.

Example. Burning ammonia

Find all solutions of the system above over $\mathbb{Q}$, and hence the balanced equation with the smallest positive integer coefficients.

Solution. In the order (a, b, c, d), with rows N, H, O, the coefficient matrix is reduced as follows:

$$\begin{pmatrix} 1 & 0 & -1 & 0 \\ 3 & 0 & 0 & -2 \\ 0 & 2 & -1 & -1 \end{pmatrix} \xrightarrow{R_2 \to R_2 - 3R_1} \begin{pmatrix} 1 & 0 & -1 & 0 \\ 0 & 0 & 3 & -2 \\ 0 & 2 & -1 & -1 \end{pmatrix} \xrightarrow{R_2 \leftrightarrow R_3} \begin{pmatrix} 1 & 0 & -1 & 0 \\ 0 & 2 & -1 & -1 \\ 0 & 0 & 3 & -2 \end{pmatrix}.$$

Scale rows 2 and 3 by $\frac{1}{2}$ and $\frac{1}{3}$, then clear column 3 above its pivot:

$$\xrightarrow[R_3 \to \frac{1}{3}R_3]{R_2 \to \frac{1}{2}R_2} \begin{pmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & -\frac{1}{2} & -\frac{1}{2} \\ 0 & 0 & 1 & -\frac{2}{3} \end{pmatrix} \xrightarrow[R_2 \to R_2 + \frac{1}{2}R_3]{R_1 \to R_1 + R_3} \begin{pmatrix} 1 & 0 & 0 & -\frac{2}{3} \\ 0 & 1 & 0 & -\frac{5}{6} \\ 0 & 0 & 1 & -\frac{2}{3} \end{pmatrix}.$$

This is the RREF. The pivot columns are 1, 2, 3, and d is the only free variable. By Theorem 3.2.4, with $d = t$, the solutions are

$$(a, b, c, d) = t\left(\tfrac{2}{3}, \tfrac{5}{6}, \tfrac{2}{3}, 1\right), \qquad t \in \mathbb{Q}.$$

So the null space is a line: rank 3 and nullity $4 - 3 = 1$, as Theorem 3.5.7 predicts. The coefficients are integers exactly when $\frac{2}{3}t$, $\frac{5}{6}t$ and t are integers, which forces t to be a multiple of 6. The smallest positive choice $t = 6$ gives $(4, 5, 4, 6)$:

$$4\,\mathrm{NH_3} + 5\,\mathrm{O_2} \longrightarrow 4\,\mathrm{NO} + 6\,\mathrm{H_2O}.$$

Check: N: $4 = 4$. H: $12 = 12$. O: $10 = 4 + 6$.

Nullity 1 is what makes "the balanced equation" meaningful: every balancing is a multiple of one. Positivity was luck of the data here. Nothing in the algebra guarantees that the null space contains a vector with all entries positive, and when it does not, the proposed reaction cannot be balanced at all.

Warning. **A null space of dimension 2 or more does not give one balanced equation.** For $a\,\mathrm{H_2} + b\,\mathrm{O_2} \to c\,\mathrm{H_2O} + d\,\mathrm{H_2O_2}$, the equations $2a = 2c + 2d$ (hydrogen) and $2b = c + 2d$ (oxygen) have rank 2 and nullity 2. Both $(2, 1, 2, 0)$ and $(1, 1, 0, 1)$ balance, and so does every positive combination of them. The "reaction" is really two reactions, $2\mathrm{H_2} + \mathrm{O_2} \to 2\mathrm{H_2O}$ and

$H_2 + O_2 \to H_2O_2$, and the proportions depend on chemistry, not on algebra.

What the model leaves out. Conservation of atoms is necessary, not sufficient. The equations cannot tell whether a reaction actually happens, and reactions involving ions also conserve electric charge, which adds one more linear equation.

3.7.4 Flows in a network

A small town center has four intersections and five one-way streets. Cars enter and leave the center at some intersections. We measure the external traffic and want to know the traffic on each street.

The model has one assumption: at every intersection, as many cars leave per minute as arrive, since cars do not pile up or vanish. In electrical circuits the same rule for currents is **Kirchhoff's current law**. To write it down, number the intersections (**vertices**) $1, \dots, m$ and the streets (**edges**) $1, \dots, k$. Each edge e runs from a vertex $\text{tail}(e)$ to a **different** vertex $\text{head}(e)$. The unknown f_e is the flow along edge e. The data s_v is the net external supply at v: cars entering the network there per minute, with a negative sign for cars leaving. Conservation at v says

$$(\text{flow out of } v) - (\text{flow into } v) = s_v.$$

Definition 3.7.1. Incidence matrix

The **incidence matrix** of a network with vertices $1, \dots, m$ and edges $1, \dots, k$, where $k \geq 1$, is $\mathbf{N} \in M_{m\times k}(\mathbb{R})$ with

$$\mathbf{N}_{ve} = \begin{cases} 1 & \text{if } v = \text{tail}(e), \\ -1 & \text{if } v = \text{head}(e), \\ 0 & \text{otherwise.} \end{cases}$$

Row v of $\mathbf{Nf}$ is $\sum_e \mathbf{N}_{ve} f_e$, the flow out of v minus the flow into v. So conservation at every vertex is the single matrix equation $\mathbf{Nf} = \mathbf{s}$. Column e of $\mathbf{N}$ has exactly one 1 and one -1, so its entries sum to 0. Adding up all the conservation equations therefore gives $0 = s_1 + \cdots + s_m$: **total inflow equals total outflow**. This is necessary for a solution. The next result shows that, for a network in one piece, it is also sufficient.

A network is **connected** if for any two vertices v, w there is a sequence of vertices $v = w_0, w_1, \dots, w_t = w$ in which each consecutive pair is joined by an edge, in either direction.

Proposition 3.7.4. Rank of an incidence matrix

Let $\mathbf{N} \in M_{m\times k}(\mathbb{R})$ be the incidence matrix of a **connected** network with m vertices. Then $\operatorname{rank} \mathbf{N} = m - 1$.

Idea. Rank is easier to see through the transpose. A vector $\mathbf{y} \in \mathbb{R}^m$ assigns a number to each vertex, and $\mathbf{N}^\top \mathbf{y}$ records, for each edge, the difference of the numbers at its two ends. So $\mathbf{N}^\top \mathbf{y} = \mathbf{0}$ says that the numbers agree across every edge. Along a chain of edges they then agree everywhere, so the null space of $\mathbf{N}^\top$ is the line of constant vectors. Rank–nullity and "row rank equals column rank" finish.

Proof. Let $\mathbf{y} \in \mathbb{R}^m$. For each edge e, entry e of $\mathbf{N}^\top \mathbf{y}$ is $\sum_v \mathbf{N}_{ve} y_v = y_{\text{tail}(e)} - y_{\text{head}(e)}$, by Definition 0.9.10 and Definition 3.7.1. Hence $\mathbf{N}^\top \mathbf{y} = \mathbf{0}$ if and only if $y_u = y_w$ whenever u and w are joined by an edge.

Suppose $\mathbf{N}^\top \mathbf{y} = \mathbf{0}$. Given any vertex w, connectedness gives a sequence $1 = w_0, w_1, \ldots, w_t = w$ with consecutive vertices joined by edges, so $y_1 = y_{w_1} = \cdots = y_w$. Hence $\mathbf{y} = y_1 \mathbf{1}$, where $\mathbf{1} = (1, \ldots, 1) \in \mathbb{R}^m$. Conversely $\mathbf{N}^\top \mathbf{1} = \mathbf{0}$, since all differences vanish. So $\operatorname{nul}(\mathbf{N}^\top) = \operatorname{span}(\mathbf{1})$, and since $\mathbf{1} \neq \mathbf{0}$, its dimension is 1.

The matrix $\mathbf{N}^\top$ has m columns, so Theorem 3.5.7 gives $\operatorname{rank} \mathbf{N}^\top = m - 1$. By Theorem 3.5.5, $\operatorname{rank} \mathbf{N} = \operatorname{rank} \mathbf{N}^\top = m - 1$. This proves the proposition.

■

Corollary 3.7.5. Solvable flow problems

Let $\mathbf{N} \in M_{m \times k}(\mathbb{R})$ be the incidence matrix of a connected network, and let $\mathbf{s} \in \mathbb{R}^m$. Then $\mathbf{Nf} = \mathbf{s}$ has a solution if and only if $s_1 + \cdots + s_m = 0$. In that case the solutions form $\mathbf{f}_0 + \operatorname{nul}(\mathbf{N})$ for any one solution $\mathbf{f}_0$, and $\dim \operatorname{nul}(\mathbf{N}) = k - m + 1$.

Proof. Let $W = \{\mathbf{z} \in \mathbb{R}^m : z_1 + \cdots + z_m = 0\}$. It is the solution set of the homogeneous system with the $1 \times m$ coefficient matrix $\begin{pmatrix} 1 & \cdots & 1 \end{pmatrix}$, so it is a subspace by Theorem 3.1.2. That matrix has rank 1, since its column space is spanned by the nonzero column (1) and so equals $\mathbb{R}^1$. By Theorem 3.5.7, $\dim W = m - 1$.

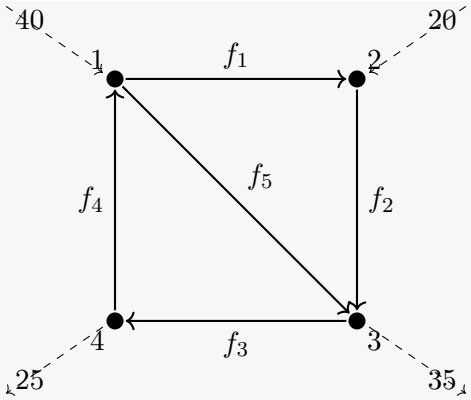
Each column of $\mathbf{N}$ has entries summing to 0, so it lies in W, and since W is a subspace, $\operatorname{col}(\mathbf{N}) \subseteq W$ by Theorem 1.3.1. Also $\dim \operatorname{col}(\mathbf{N}) = \operatorname{rank} \mathbf{N} = m - 1 = \dim W$, by Definition 3.5.4 and Proposition 3.7.4. Hence $\operatorname{col}(\mathbf{N}) = W$ by Theorem 1.6.8. By Theorem 3.1.1, $\mathbf{Nf} = \mathbf{s}$ is consistent if and only if $\mathbf{s} \in \operatorname{col}(\mathbf{N}) = W$, which is the first claim. The description of all solutions is Theorem 3.1.3, and $\dim \operatorname{nul}(\mathbf{N}) = k - (m - 1)$ by Theorem 3.5.7, since $\mathbf{N}$ has k columns.

■

The $k - m + 1$ free parameters have a clear meaning: they are flows around **loops**. Adding the same amount of traffic all the way around a closed circuit of streets changes no intersection's balance.

Example. Traffic in a town center

Four intersections are joined by one-way streets $1 \to 2$, $2 \to 3$, $3 \to 4$, $4 \to 1$, and $1 \to 3$, carrying unknown flows $f_1, \ldots, f_5$ (cars per minute) in that order. From outside, 40 cars per minute enter at intersection 1 and 20 at intersection 2, while 35 leave at intersection 3 and 25 at intersection 4.



Find all flows that satisfy conservation. If street $4 \to 1$ is closed for repairs and flows cannot be negative, what is the largest flow the diagonal street $1 \to 3$ may have to carry?

Solution. The network is connected, and $\mathbf{s} = (40, 20, -35, -25)$ sums to 0, so by Corollary 3.7.5 a solution exists and the solutions have $5 - 4 + 1 = 2$ free parameters. We row reduce $[\mathbf{N} \mid \mathbf{s}]$, with rows for vertices $1, 2, 3, 4$ and columns for $f_1, \dots, f_5$:

$$\left(\begin{array}{ccccc|c} 1 & 0 & 0 & -1 & 1 & 40 \\ -1 & 1 & 0 & 0 & 0 & 20 \\ 0 & -1 & 1 & 0 & -1 & -35 \\ 0 & 0 & -1 & 1 & 0 & -25 \end{array}\right) \xrightarrow{R_2 \to R_2 + R_1} \left(\begin{array}{ccccc|c} 1 & 0 & 0 & -1 & 1 & 40 \\ 0 & 1 & 0 & -1 & 1 & 60 \\ 0 & -1 & 1 & 0 & -1 & -35 \\ 0 & 0 & -1 & 1 & 0 & -25 \end{array}\right)$$

$$\xrightarrow{R_3 \to R_3 + R_2} \left(\begin{array}{ccccc|c} 1 & 0 & 0 & -1 & 1 & 40 \\ 0 & 1 & 0 & -1 & 1 & 60 \\ 0 & 0 & 1 & -1 & 0 & 25 \\ 0 & 0 & -1 & 1 & 0 & -25 \end{array}\right) \xrightarrow{R_4 \to R_4 + R_3} \left(\begin{array}{ccccc|c} 1 & 0 & 0 & -1 & 1 & 40 \\ 0 & 1 & 0 & -1 & 1 & 60 \\ 0 & 0 & 1 & -1 & 0 & 25 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right).$$

This is the RREF, with pivots in columns 1, 2, 3, so $\operatorname{rank} \mathbf{N} = 3 = m - 1$, as Proposition 3.7.4 predicts. The free variables are $f_4 = \alpha$ and $f_5 = \beta$, and by Theorem 3.2.4,

$$\mathbf{f} = \begin{pmatrix} 40 \\ 60 \\ 25 \\ 0 \\ 0 \end{pmatrix} + \alpha \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 0 \end{pmatrix} + \beta \begin{pmatrix} -1 \\ -1 \\ 0 \\ 0 \\ 1 \end{pmatrix}, \qquad \alpha, \beta \in \mathbb{R}.$$

The two null space vectors are the loops: α sends extra cars around $1 \to 2 \to 3 \to 4 \to 1$, and β moves cars from the route $1 \to 2 \to 3$ onto the diagonal $1 \to 3$.

With street $4 \to 1$ closed, $\alpha = f_4 = 0$ and $\mathbf{f} = (40 - \beta,\ 60 - \beta,\ 25,\ 0,\ \beta)$. Nonnegativity requires $\beta \geq 0$, $40 - \beta \geq 0$ and $60 - \beta \geq 0$, that is, $0 \leq \beta \leq 40$. So the diagonal may carry up to 40 cars per minute, which happens when street $1 \to 2$ is empty. Check, for $\mathbf{f} = (0, 20, 25, 0, 40)$: at vertex 1, out $0 + 40$ minus in 0 is 40; at 2, out 20 minus in 0 is 20; at 3, out 25 minus in $20 + 40$ is -35; at 4, out 0 minus in 25 is -25.

What the model leaves out. Linear algebra finds every flow that conserves cars, including physically meaningless ones with negative entries, and it cannot say which of the many valid flows drivers actually choose. Nonnegativity and street capacities turn the problem into one

about inequalities, which is the subject of linear programming. The matrix $\mathbf{N}\mathbf{N}^\top$ of this model, the graph Laplacian, returns in Chapter 24.

3.7.5 Lights out over $\mathbb{F}_2$

A puzzle has a 3×3 grid of lights, each on or off. Pressing a light **toggles** it and its horizontal and vertical neighbors (a corner has two neighbors, an edge light three, the center four). Starting from some pattern, the goal is to switch every light off. Can every starting pattern be solved?

The model rests on two observations about the puzzle, and together they say that the right field is $\mathbb{F}_2 = \{0, 1\}$.

1. Encode "on" as 1 and "off" as 0. Toggling a light is adding 1 in $\mathbb{F}_2$, because $0 + 1 = 1$ and $1 + 1 = 0$. The effect of several presses on a light is the sum of their effects, and since addition is commutative, **the order of presses does not matter**.
2. Pressing a button twice toggles each affected light twice, which is $1 + 1 = 0$: nothing. So a strategy is just the **set** of buttons pressed once, a vector $\mathbf{x} \in \mathbb{F}_2^9$.

Number the cells (i, j), row i, column j, and list them row by row. Pressing button (i, j) adds to the pattern a fixed vector $\mathbf{a}_{ij} \in \mathbb{F}_2^9$, with 1's at (i, j) and its neighbors. Let $\mathbf{A} \in M_9(\mathbb{F}_2)$ be the matrix with these columns. Starting from pattern $\mathbf{b}$ and pressing the buttons in $\mathbf{x}$, the final pattern is $\mathbf{b} + \mathbf{A}\mathbf{x}$ by Theorem 0.9.3, and we want it to be $\mathbf{0}$. Since $-\mathbf{b} = \mathbf{b}$ in $\mathbb{F}_2^9$, the puzzle is the linear system

$$\mathbf{A}\mathbf{x} = \mathbf{b} \quad \text{over } \mathbb{F}_2, \qquad \mathbf{A} = \begin{pmatrix} 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 \end{pmatrix}.$$

By Theorem 3.1.1, pattern $\mathbf{b}$ is solvable exactly when $\mathbf{b} \in \operatorname{col}(\mathbf{A})$. So "is every pattern solvable?" asks whether $\operatorname{rank} \mathbf{A} = 9$.

Example. Every pattern of the 3×3 puzzle is solvable

Show that $\operatorname{rank} \mathbf{A} = 9$. Deduce that every one of the $2^9 = 512$ starting patterns can be switched off, by exactly one set of presses.

Solution. We show that $\mathbf{A}\mathbf{x} = \mathbf{0}$ has only the trivial solution. Write the presses as a grid x_{ij}, and suppose $\mathbf{A}\mathbf{x} = \mathbf{0}$: the presses change no light. The equation for light (i, j) says that the presses at (i, j) and its neighbors add up to 0. We solve these equations row by row, remembering that $-u = u$ in $\mathbb{F}_2$. Name the top row $x_{11} = p$, $x_{12} = q$, $x_{13} = r$.

Lights of row 1 determine the presses of row 2. Light $(1, 1)$ gives $x_{11} + x_{12} + x_{21} = 0$, so $x_{21} = p + q$. Likewise lights $(1, 2)$ and $(1, 3)$ give

$$x_{22} = x_{11} + x_{12} + x_{13} = p + q + r, \qquad x_{23} = x_{12} + x_{13} = q + r.$$

Lights of row 2 determine the presses of row 3. Light $(2,1)$ gives $x_{31} = x_{11} + x_{21} + x_{22} = p + (p+q) + (p+q+r) = p + r$, using $3p = p$ and $2q = 0$. Light $(2,2)$ gives $x_{32} = x_{12} + x_{21} + x_{22} + x_{23} = q + (p+q) + (p+q+r) + (q+r) = 0$. Light $(2,3)$ gives $x_{33} = x_{13} + x_{22} + x_{23} = r + (p+q+r) + (q+r) = p + r$.

Lights of row 3 give three conditions on p, q, r.

$$\begin{aligned}
(3,1): &\quad x_{21} + x_{31} + x_{32} = (p+q) + (p+r) + 0 = q + r = 0,\\
(3,2): &\quad x_{22} + x_{31} + x_{32} + x_{33} = (p+q+r) + (p+r) + 0 + (p+r)\\
&\qquad\qquad = p + q + r = 0,\\
(3,3): &\quad x_{23} + x_{32} + x_{33} = (q+r) + 0 + (p+r) = p + q = 0.
\end{aligned}$$

The first and third give $r = q$ and $p = q$, and then the second gives $3q = q = 0$. So $p = q = r = 0$, and the formulas above make every $x_{ij} = 0$.

Hence the null space of $\mathbf{A}$ is $\{\mathbf{0}\}$, and rank $\mathbf{A} = 9 - 0 = 9$ by Theorem 3.5.7. So $\operatorname{col}(\mathbf{A})$ is a subspace of $\mathbb{F}_2^9$ of dimension 9, and $\operatorname{col}(\mathbf{A}) = \mathbb{F}_2^9$ by Theorem 1.6.8. Every pattern $\mathbf{b}$ lies in $\operatorname{col}(\mathbf{A})$, so every pattern is solvable. The solution is unique by Corollary 3.1.4, since $\mathbf{Ax} = \mathbf{0}$ has only the trivial solution. In particular exactly one of the 512 press sets solves each of the 512 patterns.

The same chase finds an actual solution: the constants from $\mathbf{b}$ simply ride along.

Example. Switching off the center

Only the center light is on. Which buttons should be pressed?

Solution. Now $\mathbf{b}$ has a 1 at $(2,2)$ only, so the equation for light $(2,2)$ has right-hand side 1 and all others 0. Rows 1 and 2 of lights give, as before, $x_{21} = p + q$, $x_{22} = p + q + r$, $x_{23} = q + r$, $x_{31} = p + r$, $x_{33} = p + r$, and now

$$x_{32} = 1 + x_{12} + x_{21} + x_{22} + x_{23} = 1.$$

The lights of row 3 give $(p+q) + (p+r) + 1 = 0$, $(p+q+r) + (p+r) + 1 + (p+r) = 0$ and $(q+r) + 1 + (p+r) = 0$, that is,

$$q + r = 1, \qquad p + q + r = 1, \qquad p + q = 1.$$

Adding the first two gives $p = 0$; then $q = 1$ and $r = 0$. Substituting, $x_{21} = x_{22} = x_{23} = 1$, $x_{31} = x_{33} = 0$, and $x_{32} = 1$. So we press the center and its four neighbors, a plus sign:

	×	
×	×	×
	×	

Check, counting how often each light is toggled. The center is toggled by all five presses: odd, so it goes from on to off. A corner, say $(1,1)$, is toggled by presses at $(1,2)$ and $(2,1)$: even, so it stays off. An edge light, say $(1,2)$, is toggled by presses at $(1,2)$ and $(2,2)$: even, so it stays off. By the symmetry of the plus sign the other corners and edges behave the same way.

Quick check. Starting from all lights **off**, which set of presses produces the pattern with only the center light **on**?

Solution. The same plus sign. Starting from $\mathbf{0}$, presses $\mathbf{x}$ produce the pattern $\mathbf{Ax}$, so we need $\mathbf{Ax} = \mathbf{b}$ with $\mathbf{b}$ the center pattern, which is exactly the system just solved. Over $\mathbb{F}_2$, turning a pattern on and turning it off are the same problem.

Remark. Larger boards behave differently. A computation over $\mathbb{F}_2$, by the same elimination, gives $\operatorname{rank} \mathbf{A} = 12$ for the 4×4 board and $\operatorname{rank} \mathbf{A} = 23$ for the 5×5 board. On the 5×5 board, $\operatorname{col}(\mathbf{A})$ has dimension 23, so it contains 2^{23} patterns (coordinates in a basis match its elements with $\mathbb{F}_2^{23}$), and only one pattern in four can be switched off.

What the model leaves out. Nothing, which is rare. The puzzle is linear over $\mathbb{F}_2$ by its rules, so the model is exact. It even predicts features that are not obvious when playing, such as the irrelevance of the order of presses.

3.7.6 Exercises

3.7.6.1 A. Check your understanding

Exercise. A1

(a) State the polynomial interpolation theorem, including its hypotheses.

(b) True or false: there is a polynomial $p \in \mathbb{R}[x]_{\leq 1}$ with $p(1) = 2$ and $p(1) = 3$. Which hypothesis of the theorem fails?

(c) In balancing a chemical equation, what does it mean if the null space of the coefficient matrix has dimension 1? Dimension 2?

(d) Why is lights out modeled over $\mathbb{F}_2$ and not over $\mathbb{R}$?

(e) A connected network has 6 vertices and 9 edges. What is the rank of its incidence matrix, and how many free parameters does the solution set of a consistent flow problem have?

Solution.

(a) See Theorem 3.7.3: for **distinct** $x_0, \ldots, x_n \in F$ and any $y_0, \ldots, y_n \in F$, there is exactly one $p \in F[x]_{\leq n}$ with $p(x_i) = y_i$ for all i.

(b) False: $p(1)$ is one number, and $2 \neq 3$. The theorem needs distinct x-values, and here $x_0 = x_1 = 1$.

(c) Dimension 1: all balancings are multiples of one vector, so there is one balanced equation up to scaling (provided a positive multiple exists). Dimension 2 or more: there are independent balancings, and the equation is a mixture of several reactions whose proportions algebra cannot decide.

(d) A light has two states, and toggling twice restores it, so toggling is adding 1 with $1 + 1 = 0$, which is arithmetic in $\mathbb{F}_2$. Over $\mathbb{R}$, pressing a button twice would count as

2, not 0.

(e) By Proposition 3.7.4 the rank is $6-1=5$. By Corollary 3.7.5 a consistent problem has $9-6+1=4$ free parameters.

3.7.6.2 B. Practice

Exercise. B1: Interpolating four points

Find the polynomial $p \in \mathbb{Q}[x]_{\le 3}$ with $p(0)=-1$, $p(1)=1$, $p(2)=3$, and $p(3)=11$. Justify that there is only one.

Solution. The x-values $0,1,2,3$ are distinct, so there is exactly one such $p = c_0+c_1x+c_2x^2+c_3x^3$, by Theorem 3.7.3. Its coefficients solve

$$\left(\begin{array}{cccc|c} 1&0&0&0&-1\\ 1&1&1&1&1\\ 1&2&4&8&3\\ 1&3&9&27&11 \end{array}\right) \xrightarrow[R_4\to R_4-R_1]{R_2\to R_2-R_1,\ R_3\to R_3-R_1} \left(\begin{array}{cccc|c} 1&0&0&0&-1\\ 0&1&1&1&2\\ 0&2&4&8&4\\ 0&3&9&27&12 \end{array}\right)$$

$$\xrightarrow[R_4\to R_4-3R_2]{R_3\to R_3-2R_2} \left(\begin{array}{cccc|c} 1&0&0&0&-1\\ 0&1&1&1&2\\ 0&0&2&6&0\\ 0&0&6&24&6 \end{array}\right)$$

and then R_4-3R_3 turns the last row into $(0,0,0,6 \mid 6)$. Back substitution: $c_3=1$; $2c_2+6=0$, so $c_2=-3$; $c_1-3+1=2$, so $c_1=4$; $c_0=-1$. Hence

$$p = -1+4x-3x^2+x^3.$$

Check: $p(0)=-1$, $p(1)=-1+4-3+1=1$, $p(2)=-1+8-12+8=3$, $p(3)=-1+12-27+27=11$.

Exercise. B2: Balancing a redox reaction

Potassium permanganate reacts with hydrochloric acid:

$$a\,\mathrm{KMnO_4} + b\,\mathrm{HCl} \longrightarrow c\,\mathrm{KCl} + d\,\mathrm{MnCl_2} + e\,\mathrm{H_2O} + f\,\mathrm{Cl_2}.$$

Write the conservation equations for K, Mn, O, H, and Cl, find all rational solutions, and hence the balanced equation with the smallest positive integer coefficients.

Solution. Counting atoms of each element on both sides:

$$\text{K: } a=c, \qquad \text{Mn: } a=d, \qquad \text{O: } 4a=e, \qquad \text{H: } b=2e, \qquad \text{Cl: } b=c+2d+2f.$$

In the order (a,b,c,d,e,f), the coefficient matrix has rows $(1,0,-1,0,0,0)$, $(1,0,0,-1,0,0)$, $(4,0,0,0,-1,0)$, $(0,1,0,0,-2,0)$, and $(0,1,-1,-2,0,-2)$. The equations are nearly solved already, so we eliminate by substitution, which is row reduction written in words. The K, Mn, and O equations give $c=a$, $d=a$, $e=4a$. The H equation gives $b=2e=8a$. The

Cl equation gives $2f = b - c - 2d = 8a - a - 2a = 5a$, so $a = \frac{2}{5}f$. Taking $f = t$ as the free variable, the RREF of the coefficient matrix is

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 & -\frac{2}{5} \\ 0 & 1 & 0 & 0 & 0 & -\frac{16}{5} \\ 0 & 0 & 1 & 0 & 0 & -\frac{2}{5} \\ 0 & 0 & 0 & 1 & 0 & -\frac{2}{5} \\ 0 & 0 & 0 & 0 & 1 & -\frac{8}{5} \end{pmatrix},$$

and the solutions are $(a, b, c, d, e, f) = t\left(\frac{2}{5}, \frac{16}{5}, \frac{2}{5}, \frac{2}{5}, \frac{8}{5}, 1\right)$, $t \in \mathbb{Q}$. The rank is 5 and the nullity $6 - 5 = 1$, so the balancing is unique up to scaling. All entries are integers exactly when t is a multiple of 5, and $t = 5$ gives

$$2\,\mathrm{KMnO_4} + 16\,\mathrm{HCl} \longrightarrow 2\,\mathrm{KCl} + 2\,\mathrm{MnCl_2} + 8\,\mathrm{H_2O} + 5\,\mathrm{Cl_2}.$$

Check: K $2 = 2$; Mn $2 = 2$; O $8 = 8$; H $16 = 16$; Cl $16 = 2 + 4 + 10$.

Exercise. B3: A ring road

Four intersections lie on a ring of one-way streets $1 \to 2$, $2 \to 3$, $3 \to 4$, $4 \to 1$, with flows f_1, f_2, f_3, f_4 in that order. From outside, 30 cars per minute enter at intersection 1 and 10 at intersection 3, while 15 leave at intersection 2 and 25 at intersection 4.

(a) Write down the incidence matrix $\mathbf{N}$ and the supply vector $\mathbf{s}$, and explain without computation why the system $\mathbf{Nf} = \mathbf{s}$ is consistent.

(b) Find all solutions.

(c) If all flows must be nonnegative, what is the smallest possible flow on each street?

Solution.

(a) With rows for the vertices and columns for the edges,

$$\mathbf{N} = \begin{pmatrix} 1 & 0 & 0 & -1 \\ -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & -1 & 1 \end{pmatrix}, \qquad \mathbf{s} = \begin{pmatrix} 30 \\ -15 \\ 10 \\ -25 \end{pmatrix}.$$

The ring is connected, and $30 - 15 + 10 - 25 = 0$, so the system is consistent by Corollary 3.7.5.

(b) Row reducing $[\mathbf{N} \mid \mathbf{s}]$: $R_2 + R_1$ gives $(0, 1, 0, -1 \mid 15)$; then $R_3 + R_2$ gives $(0, 0, 1, -1 \mid 25)$; then $R_4 + R_3$ gives the zero row. The pivots are in columns 1, 2, 3, and $f_4 = t$ is free, so

$$\mathbf{f} = (30 + t,\ 15 + t,\ 25 + t,\ t), \qquad t \in \mathbb{R}.$$

There is $4 - 4 + 1 = 1$ parameter, the flow around the ring. Check at $t = 0$: vertex 1, $30 - 0 = 30$; vertex 2, $15 - 30 = -15$; vertex 3, $25 - 15 = 10$; vertex 4, $0 - 25 = -25$.

(c) All flows are nonnegative exactly when $t \geq 0$, since then $30 + t, 15 + t, 25 + t \geq 0$ too. Each flow increases with t, so all are smallest at $t = 0$: $f = (30, 15, 25, 0)$. No car needs to use street $4 \to 1$.

3.7.6.3 C. Going deeper

Exercise. C1: The 2 × 2 puzzle

Consider lights out on a 2×2 grid, where pressing a light toggles it and its two horizontal and vertical neighbors. Number the cells $1 = (1,1)$, $2 = (1,2)$, $3 = (2,1)$, $4 = (2,2)$.

(a) Write down the matrix $\mathbf{A} \in M_4(\mathbb{F}_2)$ of the puzzle.

(b) Show that $\mathbf{A}^2 = \mathbf{I}$.

(c) Deduce that every pattern is solvable by exactly one set of presses, and find the presses that switch off the pattern with only light 1 on.

Solution.

(a) Button j toggles every light except the one diagonally opposite. So

$$\mathbf{A} = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 \end{pmatrix}.$$

(b) Let $\mathbf{J}$ be the 4×4 all-ones matrix and $\mathbf{Q}$ the permutation matrix exchanging $1 \leftrightarrow 4$ and $2 \leftrightarrow 3$, which has 1's on the anti-diagonal. Then $\mathbf{A} = \mathbf{J} - \mathbf{Q} = \mathbf{J} + \mathbf{Q}$ over $\mathbb{F}_2$. Every entry of $\mathbf{J}^2$ is $1+1+1+1 = 0$, so $\mathbf{J}^2 = \mathbf{O}$. Each row and each column of $\mathbf{Q}$ has a single 1, so $\mathbf{JQ} = \mathbf{J}$ and $\mathbf{QJ} = \mathbf{J}$. And $\mathbf{Q} = \mathbf{P}_\sigma$ for $\sigma = (1\ 4)(2\ 3)$, with $\sigma \circ \sigma = \mathrm{id}$, so $\mathbf{Q}^2 = \mathbf{P}_{\mathrm{id}} = \mathbf{I}$ by Lemma 3.4.7 (a). By Theorem 0.9.4,

$$\mathbf{A}^2 = \mathbf{J}^2 + \mathbf{JQ} + \mathbf{QJ} + \mathbf{Q}^2 = \mathbf{O} + \mathbf{J} + \mathbf{J} + \mathbf{I} = \mathbf{I},$$

because $\mathbf{J} + \mathbf{J} = 2\mathbf{J} = \mathbf{O}$ over $\mathbb{F}_2$.

(c) By (b), $\mathbf{A}$ is invertible with $\mathbf{A}^{-1} = \mathbf{A}$. By Theorem 0.9.7 (7), $\mathbf{Ax} = \mathbf{b}$ has exactly one solution $\mathbf{x} = \mathbf{Ab}$ for every $\mathbf{b} \in \mathbb{F}_2^4$. For $\mathbf{b} = \mathbf{e}_1$, $\mathbf{x} = \mathbf{Ae}_1 = (1,1,1,0)$: press lights $1, 2, 3$. Check: light 1 is toggled by presses 1, 2, 3, three times, so it goes off; light 2 by presses 1 and 2; light 3 by presses 1 and 3; light 4 by presses 2 and 3; each of these is toggled twice and stays off.

Exercise. C2: Interpolating values and slopes

For $p = c_0 + c_1 x + c_2 x^2 + c_3 x^3 \in F[x]_{\le 3}$, define its (formal) derivative $p' = c_1 + 2c_2 x + 3c_3 x^2$.

(a) Prove that for **every** field F and all $a, b, c, d \in F$ there is exactly one $p \in F[x]_{\le 3}$ with

$$p(0) = a, \qquad p'(0) = b, \qquad p(1) = c, \qquad p'(1) = d.$$

(b) Find this p over $\mathbb{Q}$ for $(a, b, c, d) = (0, 1, 1, 0)$.

Hint: in (a), check that elimination never divides by an element that could be zero in some field.

Solution.

(a) The four conditions are linear in $c_0, \dots, c_3$: $p(0) = c_0$, $p'(0) = c_1$, $p(1) = c_0+c_1+c_2+c_3$, $p'(1) = c_1 + 2c_2 + 3c_3$. Row reducing the augmented matrix, using only additions of multiples of rows,

$$\left(\begin{array}{cccc|c} 1 & 0 & 0 & 0 & a \\ 0 & 1 & 0 & 0 & b \\ 1 & 1 & 1 & 1 & c \\ 0 & 1 & 2 & 3 & d \end{array}\right) \xrightarrow[R_4 \to R_4 - R_2]{R_3 \to R_3 - R_1 - R_2} \left(\begin{array}{cccc|c} 1 & 0 & 0 & 0 & a \\ 0 & 1 & 0 & 0 & b \\ 0 & 0 & 1 & 1 & c-a-b \\ 0 & 0 & 2 & 3 & d-b \end{array}\right)$$

$$\xrightarrow{R_4 \to R_4 - 2R_3} \left(\begin{array}{cccc|c} 1 & 0 & 0 & 0 & a \\ 0 & 1 & 0 & 0 & b \\ 0 & 0 & 1 & 1 & c-a-b \\ 0 & 0 & 0 & 1 & 2a+b-2c+d \end{array}\right).$$

Here $3 - 2 = 1$ holds in every field. The coefficient matrix is now upper triangular with every diagonal entry equal to 1, so it is invertible by Lemma 3.6.2, and back substitution gives the unique solution without any division:

$$\begin{aligned} c_3 &= 2a + b - 2c + d, \\ c_2 &= (c - a - b) - c_3 = -3a - 2b + 3c - d, \\ c_1 &= b, \qquad c_0 = a. \end{aligned}$$

Row operations preserve the solution set (Theorem 3.2.2), so the original system also has exactly this one solution, for every field F.

(b) With $(a, b, c, d) = (0, 1, 1, 0)$: $c_3 = 0 + 1 - 2 + 0 = -1$, $c_2 = 0 - 2 + 3 - 0 = 1$, $c_1 = 1$, $c_0 = 0$. So $p = x + x^2 - x^3$. Check: $p(0) = 0$; $p' = 1 + 2x - 3x^2$, so $p'(0) = 1$; $p(1) = 1 + 1 - 1 = 1$; $p'(1) = 1 + 2 - 3 = 0$.

Exercise. C3: Networks in several pieces

Let $\mathbf{N} \in M_{m\times k}(\mathbb{R})$ be the incidence matrix of a network with m vertices. Say that vertices v and w are **linked** if they are joined by a sequence of vertices in which consecutive ones share an edge, in either direction (every vertex is linked to itself by the sequence of length zero).

(a) Show that "linked" is an equivalence relation. Its equivalence classes are called the **components** of the network; let c be their number.

(b) Prove that $\operatorname{rank} \mathbf{N} = m - c$.

Solution.

(a) *Reflexive:* by the sequence of length zero. *Symmetric:* reversing a sequence from v to w gives one from w to v, and consecutive vertices still share an edge. *Transitive:* following a sequence from u to v by one from v to w gives a sequence from u to w. So "linked" is an equivalence relation (Definition 0.5.2).

(b) Let $C_1, \dots, C_c$ be the components, and let $\mathbf{1}_{C_r} \in \mathbb{R}^m$ have entry 1 at the vertices of C_r and 0 elsewhere. As in the proof of Proposition 3.7.4, $\mathbf{N}^\top \mathbf{y} = \mathbf{0}$ if and only if

$y_u = y_w$ for every edge joining u and w. Following sequences of edges, this holds if and only if $\mathbf{y}$ is constant on each component, that is, $\mathbf{y} = \sum_r \lambda_r \mathbf{1}_{C_r}$ with λ_r the common value on C_r. (Conversely each $\mathbf{1}_{C_r}$ is in the null space, since both ends of an edge lie in the same component.) So $(\mathbf{1}_{C_1}, \ldots, \mathbf{1}_{C_c})$ spans $\operatorname{nul}(\mathbf{N}^\top)$. It is independent: if $\sum_r \lambda_r \mathbf{1}_{C_r} = \mathbf{0}$, then reading the entry at any vertex of C_r, where only the r-th vector is nonzero because the components are disjoint (Theorem 0.5.2), gives $\lambda_r = 0$. Hence $\dim \operatorname{nul}(\mathbf{N}^\top) = c$. Since $\mathbf{N}^\top$ has m columns, Theorem 3.5.7 gives $\operatorname{rank} \mathbf{N}^\top = m - c$, and Theorem 3.5.5 gives $\operatorname{rank} \mathbf{N} = m - c$, as claimed.

Chapter 4

Matrices of Maps

Chapter 2 built the theory of linear maps without ever choosing a basis, and Chapter 3 built the machinery of elimination for matrices. This chapter joins them. Fix a basis of the domain and one of the target, write each $T(v_j)$ in coordinates, and its coordinates become the columns of a matrix. Every linear map between finite-dimensional spaces becomes a matrix, composition becomes matrix multiplication — which is why the product is defined the way it is — and every question about a map becomes a question about an array of scalars that Chapter 3 knows how to reduce.

What we gain is computation; what we give up is uniqueness, because the matrix depends on the bases. So the chapter's real subject is the dependence itself. Changing basis conjugates the matrix, $\mathbf{A} \mapsto \mathbf{P}^{-1}\mathbf{A}\mathbf{P}$, and two matrices represent the same operator exactly when they are **similar** — the question that the rest of the book spends five chapters answering. When both bases may be chosen freely the answer is immediate and disappointing in its simplicity: every matrix is equivalent to $\begin{pmatrix} \mathbf{I}_r & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{pmatrix}$, and rank is the only invariant. The chapter closes with two constructions that the coordinate machinery makes routine: products and quotients of spaces with the isomorphism theorems, and projections and the trace, a number attached to an operator that no choice of basis can change.

What you need. From Chapter 2, all of it, and most of all the freedom to prescribe a linear map on a basis (Theorem 2.1.5), which is what makes the correspondence between maps and matrices a bijection: kernel and image with the injectivity test (Definition 2.2.1, Definition 2.2.2, Theorem 2.2.1, Theorem 2.2.2, Theorem 2.2.3, Theorem 2.2.4), rank, nullity, and Rank–Nullity with its counting corollaries (Definition 2.3.2, Definition 2.3.1, Theorem 2.3.1, Corollary 2.3.2), the algebra $\mathcal{L}(V)$ and polynomials of an operator (Definition 2.4.1, Theorem 2.4.1, Theorem 2.4.2, Definition 2.4.2), and isomorphisms — above all the coordinate isomorphism $v \mapsto [v]_{\mathcal{B}}$, which every proof here runs through (Corollary 2.5.5, Theorem 2.5.1, Theorem 2.5.4, Theorem 2.5.6, Theorem 2.5.7).

From Chapter 3, the matrix side of the same story: row reduction and the reduced form it produces (Definition 3.2.4, Theorem 3.2.3, Theorem 3.3.4), elementary matrices and the Invertible Matrix Theorem (Theorem 3.4.1, Theorem 3.4.4, Theorem 3.4.5), and the subspaces and numbers attached to a matrix, which §01 shows are the map's own (Definition 3.5.3, Definition 3.5.1, Definition 3.5.4, Theorem 3.5.4, Theorem 3.5.2, Theorem 3.5.5, Theorem 3.5.7, Theorem 3.5.9).

From Chapter 0, matrix arithmetic, since that is what a matrix of a map is made of: the product read three ways, its algebra, the transpose and similarity of matrices (Definition 0.9.9, Theorem 0.9.2, Theorem 0.9.4, Definition 0.9.10, Theorem 0.9.5, Example (Similarity)), the trace with

its cyclic property (Theorem 0.9.6), and — for §04 only — equivalence classes and the check that a function on a quotient is well defined (Definition 0.5.2, Theorem 0.5.3). From Chapter 1, bases, coordinates, and dimension (Definition 1.5.1, Definition 1.6.1, Theorem 1.5.2, Theorem 1.6.4, Theorem 1.6.5), and direct sums with complements (Theorem 1.7.4, Theorem 1.7.6), which §05 reads through operators.

4.0.1 Roadmap

- **The matrix of a linear map.** Coordinates turn maps into matrices, composition into matrix multiplication, and $\mathcal{L}(V, W)$ into $M_{m\times n}(F)$; rank and invertibility of a map are read off any matrix for it.
- **Change of basis and similarity.** How the matrix changes when the bases do, the change-of-coordinates matrix, and similarity as the relation "same operator, different basis", with the invariants it preserves.
- **Rank, equivalence, and factorizations.** With both bases free, every matrix reduces to $\begin{pmatrix} \mathbf{I}_r & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{pmatrix}$; rank factorizations, and the rank inequalities that follow from them.
- **Products and quotient spaces.** Building new spaces from old — the external product, and the quotient V/U whose vectors are cosets — with the isomorphism theorems that relate them.
- **Projections and the trace of an operator.** Direct sums seen through the idempotent operators they produce, and a trace that belongs to the operator rather than to any of its matrices.

4.0.2 Named moves

- **Abstract space → write down a matrix.** Choose a basis that makes the map simple, compute with the matrix, and translate back through the coordinate isomorphism.
- **The change-of-basis square.** Do not recompute a matrix in a new basis: go around the square instead of across it, $[T]_{\mathcal{B}'}^{\mathcal{B}'} = \mathbf{P}^{-1}[T]_{\mathcal{B}}^{\mathcal{B}}\mathbf{P}$.
- **Free both bases.** When nothing forces the bases to match, choose them to suit the map; that is what turns an arbitrary matrix into a rank normal form.
- **Factor through the image.** Writing a rank r map as a surjection onto an r-dimensional space followed by an injection proves the rank inequalities in one move.
- **Check well-definedness on cosets.** Every operation on a quotient space, and every map out of one, begins by showing the answer does not depend on the representative chosen.
- **Recognize a direct sum by an idempotent.** $P^2 = P$ splits V as $\operatorname{im} P \oplus \ker P$, and conversely; a projection carries the complement, not just the image.

4.1 The Matrix of a Linear Map

An isomorphism $V \cong F^n$ turns the vectors of an abstract space into columns. This section does the same for **maps**. Once a basis is fixed on each side, every linear map between finite-dimensional spaces becomes a matrix, applying the map becomes multiplying by that matrix, and composing maps becomes multiplying matrices. This last fact finally explains the strange multiplication rule of Chapter 0, and it gives us the working method of the rest of the book: to study an abstract map, write down its matrix, compute there, and translate back.

4.1.1 Writing a map as a matrix

We have seen one family of linear maps that is completely described by a table of numbers: the matrix maps $T_{\mathbf{A}}\colon F^n \to F^m$, $\mathbf{x} \mapsto \mathbf{A}\mathbf{x}$. In fact **every** linear map $T\colon F^n \to F^m$ is one of them. By Theorem 2.1.4 and Theorem 0.9.3, for $\mathbf{x} = x_1\mathbf{e}_1 + \cdots + x_n\mathbf{e}_n$,

$$T\mathbf{x} = x_1 T\mathbf{e}_1 + \cdots + x_n T\mathbf{e}_n = \mathbf{A}\mathbf{x}, \qquad \text{where } \mathbf{A} = \begin{pmatrix} T\mathbf{e}_1 & \cdots & T\mathbf{e}_n \end{pmatrix}.$$

The j-th column of $\mathbf{A}$ is the image of the j-th basis vector. We want the same description for maps between arbitrary finite-dimensional spaces, such as differentiation on polynomials.

Take $D\colon \mathbb{R}[x]_{\le 3} \to \mathbb{R}[x]_{\le 2}$. Then

$$D(a_0 + a_1 x + a_2 x^2 + a_3 x^3) = a_1 + 2a_2 x + 3a_3 x^2.$$

A polynomial is not a column, but its coordinates are. In the bases $(1, x, x^2, x^3)$ and $(1, x, x^2)$, the input has coordinates (a_0, a_1, a_2, a_3) and the output has coordinates $(a_1, 2a_2, 3a_3)$, and

$$\begin{pmatrix} a_1 \\ 2a_2 \\ 3a_3 \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 3 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{pmatrix}.$$

Look at the columns of this matrix: $(0,0,0)$, $(1,0,0)$, $(0,2,0)$, $(0,0,3)$. They are the coordinates of $D(1) = 0$, $D(x) = 1$, $D(x^2) = 2x$, $D(x^3) = 3x^2$. The standard example generalizes: record the images of the input basis, in coordinates of the output basis, as columns.

The matrix of a linear map lists, column by column, where the input basis vectors go, written in coordinates of the output basis.

Definition 4.1.1. Matrix of a Linear Map

Let V and W be finite-dimensional vector spaces over F, with **ordered** bases $\mathcal{B} = (v_1, \ldots, v_n)$ of V and $\mathcal{C} = (w_1, \ldots, w_m)$ of W, and let $T \in \mathcal{L}(V, W)$. The **matrix of T with respect to $\mathcal{B}$ and $\mathcal{C}$** is the $m \times n$ matrix

$$[T]_{\mathcal{B}}^{\mathcal{C}} := \begin{pmatrix} [Tv_1]_{\mathcal{C}} & [Tv_2]_{\mathcal{C}} & \cdots & [Tv_n]_{\mathcal{C}} \end{pmatrix} \in M_{m\times n}(F),$$

whose j-th column is the coordinate vector of Tv_j with respect to $\mathcal{C}$. Equivalently, its entries a_{ij} are the **unique** scalars with

$$Tv_j = a_{1j}w_1 + a_{2j}w_2 + \cdots + a_{mj}w_m \qquad (j = 1, \ldots, n).$$

When $V = W$ and $\mathcal{B} = \mathcal{C}$, we call $[T]_{\mathcal{B}}^{\mathcal{B}}$ the **matrix of the operator T with respect to $\mathcal{B}$**; the basis is written in both slots even though it is the same one on each side.

In words, clause by clause:

- **The input basis $\mathcal{B}$ is the subscript, the output basis $\mathcal{C}$ is the superscript.** The subscript sits next to where the input goes, as in $[T]_{\mathcal{B}}^{\mathcal{C}}[v]_{\mathcal{B}}$ below.
- **Columns correspond to input basis vectors**, so there are $n = \dim V$ of them; **rows correspond to output basis vectors**, so there are $m = \dim W$. The size is "dim(codomain) × dim(domain)", the same order as $M_{m\times n}$ for $T_{\mathbf{A}}\colon F^n \to F^m$.
- Column j answers one question: **where does v_j go?**

Well-definedness. Each Tv_j lies in W, and $\mathcal{C}$ is a basis, so $[Tv_j]_{\mathcal{C}}$ exists and is unique by Theorem 1.5.1. So the matrix is determined by T, $\mathcal{B}$, and $\mathcal{C}$. It is **not** determined by T alone: change either basis, or even the **order** of the vectors in one of them, and the matrix changes. That is why the bases are ordered.

Example. Differentiation on Cubic Polynomials

Let $\mathcal{E}_3 = (1, x, x^2, x^3)$ and $\mathcal{E}_2 = (1, x, x^2)$. Find $[D]_{\mathcal{E}_3}^{\mathcal{E}_2}$ for $D\colon \mathbb{R}[x]_{\le 3} \to \mathbb{R}[x]_{\le 2}$, and $[D]_{\mathcal{E}_3}^{\mathcal{E}_3}$ for D regarded as an operator on $\mathbb{R}[x]_{\le 3}$.

Solution. The images of the input basis are $D(1) = 0$, $D(x) = 1$, $D(x^2) = 2x$, $D(x^3) = 3x^2$. In $\mathcal{E}_2$ their coordinates are $(0,0,0)$, $(1,0,0)$, $(0,2,0)$, $(0,0,3)$; in $\mathcal{E}_3$ the same vectors have a fourth coordinate 0. So

$$[D]_{\mathcal{E}_3}^{\mathcal{E}_2} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 3 \end{pmatrix}, \qquad [D]_{\mathcal{E}_3}^{\mathcal{E}_3} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 3 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

The same rule $p \mapsto p'$ gives a 3×4 or a 4×4 matrix, depending on the codomain we declare. The superdiagonal $1, 2, 3$ records "$x^k \mapsto kx^{k-1}$".

Example. Matrix Maps and the Identity

(a) Let $\mathbf{A} \in M_{m\times n}(F)$, and let $\mathcal{E}_n$, $\mathcal{E}_m$ be the standard bases of F^n, F^m. Show that $[T_{\mathbf{A}}]_{\mathcal{E}_n}^{\mathcal{E}_m} = \mathbf{A}$.

(b) Let $\mathcal{B}$ be any basis of an n-dimensional space V. Find $[\mathrm{id}_V]_{\mathcal{B}}^{\mathcal{B}}$ and the matrix of the zero map $V \to W$ for any bases.

Solution.

(a) The j-th column of $[T_{\mathbf{A}}]_{\mathcal{E}_n}^{\mathcal{E}_m}$ is $[\mathbf{A}\mathbf{e}_j]_{\mathcal{E}_m} = \mathbf{A}\mathbf{e}_j$, since coordinates in the standard basis of F^m are the entries themselves; and $\mathbf{A}\mathbf{e}_j$ is the j-th column of $\mathbf{A}$ (Theorem 0.9.3). So the matrix of a matrix map in the standard bases is the matrix we started with.

(b) $\mathrm{id}_V\, v_j = v_j = 0v_1 + \cdots + 1v_j + \cdots + 0v_n$, so the j-th column of $[\mathrm{id}_V]_{\mathcal{B}}^{\mathcal{B}}$ is $\mathbf{e}_j$, and $[\mathrm{id}_V]_{\mathcal{B}}^{\mathcal{B}} = \mathbf{I}_n$. The zero map sends every v_j to 0, whose coordinates are all 0, so its matrix is the zero matrix of size $\dim W \times \dim V$, whatever the bases. These degenerate cases fix the expectations "identity map, identity matrix" and "zero map, zero matrix". The first needs the **same** basis on both sides, as the warning below shows.

Example. The Transpose Map

Let $T\colon M_2(F) \to M_2(F)$, $T(\mathbf{X}) = \mathbf{X}^{\top}$, and let $\mathcal{E} = (\mathbf{E}_{11}, \mathbf{E}_{12}, \mathbf{E}_{21}, \mathbf{E}_{22})$. Find $[T]_{\mathcal{E}}^{\mathcal{E}}$.

Solution. T is linear by Theorem 0.9.5. Its values on the basis are $\mathbf{E}_{11}^{\top} = \mathbf{E}_{11}$, $\mathbf{E}_{12}^{\top} = \mathbf{E}_{21}$, $\mathbf{E}_{21}^{\top} = \mathbf{E}_{12}$, $\mathbf{E}_{22}^{\top} = \mathbf{E}_{22}$. In $\mathcal{E}$, these have coordinates $\mathbf{e}_1, \mathbf{e}_3, \mathbf{e}_2, \mathbf{e}_4$. So

$$[T]_{\mathcal{E}}^{\mathcal{E}} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

A map on a space of 2×2 matrices is described by a 4×4 matrix, because $\dim M_2(F) = 4$. Do not confuse the matrix $\mathbf{X}$ that T acts on with the matrix $[T]_{\mathcal{E}}^{\mathcal{E}}$ that describes T.

Example. A Reflection in Two Bases

Let $T\colon \mathbb{R}^2 \to \mathbb{R}^2$ be the reflection across the line $y = x$, $T(x, y) = (y, x)$. Find $[T]_{\mathcal{E}}^{\mathcal{E}}$ for the standard basis $\mathcal{E} = (\mathbf{e}_1, \mathbf{e}_2)$, and $[T]_{\mathcal{B}}^{\mathcal{B}}$ for $\mathcal{B} = (\mathbf{u}_1, \mathbf{u}_2)$ with $\mathbf{u}_1 = (1, 1)$ and $\mathbf{u}_2 = (1, -1)$.

Solution. $T\mathbf{e}_1 = (0, 1) = \mathbf{e}_2$ and $T\mathbf{e}_2 = \mathbf{e}_1$, so $[T]_{\mathcal{E}}^{\mathcal{E}} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$.

The vectors $\mathbf{u}_1, \mathbf{u}_2$ are independent (neither is a multiple of the other), hence a basis of $\mathbb{R}^2$ by Theorem 1.6.4. The first lies **on** the mirror line and the second is **perpendicular** to it: $T\mathbf{u}_1 = (1, 1) = 1\mathbf{u}_1 + 0\mathbf{u}_2$ and $T\mathbf{u}_2 = (-1, 1) = 0\mathbf{u}_1 + (-1)\mathbf{u}_2$. So

$$[T]_{\mathcal{B}}^{\mathcal{B}} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

In the basis adapted to the geometry, the reflection is visibly "keep one direction, flip the other". The same map has a diagonal matrix in one basis and a nondiagonal one in another. How the two matrices are related is the subject of the next section.

The matrix also works in reverse: from $[T]_{\mathcal{B}}^{\mathcal{B}}$ alone we can recover T. Take $\mathbf{v} = (3, 1)$. Its coordinates in $\mathcal{B}$ solve $a\mathbf{u}_1 + b\mathbf{u}_2 = (3, 1)$, that is, $a + b = 3$ and $a - b = 1$, so $[\mathbf{v}]_{\mathcal{B}} = (2, 1)$, that is, $\mathbf{v} = 2\mathbf{u}_1 + \mathbf{u}_2$. By linearity, $T\mathbf{v} = 2T\mathbf{u}_1 + T\mathbf{u}_2$. The columns of $[T]_{\mathcal{B}}^{\mathcal{B}}$ say $T\mathbf{u}_1 = 1\mathbf{u}_1 + 0\mathbf{u}_2$ and $T\mathbf{u}_2 = 0\mathbf{u}_1 + (-1)\mathbf{u}_2$, so $T\mathbf{v} = 2\mathbf{u}_1 - \mathbf{u}_2 = (1, 3)$, which is indeed (y, x) at $(3, 1)$. In coordinates, we turned $(2, 1)$ into $(2, -1) = [T]_{\mathcal{B}}^{\mathcal{B}} \begin{pmatrix} 2 \\ 1 \end{pmatrix}$; the next subsection proves that this "coordinates in, multiply, combine out" recipe always works.

Here is a nonexample by minimal change. In Example (Differentiation on Cubic Polynomials), write the coordinate vectors $[D(x^j)]_{\mathcal{E}_2}$ as **rows** instead of columns. The numbers are the same, but the result is the 4×3 matrix $\begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{pmatrix}$, and it cannot even be multiplied by the coordinate vector (a_0, a_1, a_2, a_3) of an input polynomial. The clause that fails is "the j-th **column** is $[Tv_j]_{\mathcal{C}}$".

Why this definition. Why columns, and why coordinates in $\mathcal{C}$? Because these are the choices under which "apply T" becomes "multiply by the matrix", as the D computation above sug-

gested; we prove this next. With rows, we would get the transpose, and the natural formula $[Tv]_{\mathcal{C}} = \mathbf{A}[v]_{\mathcal{B}}$ would fail. The convention that the input basis is the subscript is chosen to make that formula easy to read: the $\mathcal{B}$ of the matrix sits next to the $\mathcal{B}$ of the vector.

Warning. **Three ways to get a matrix of a map wrong.** (1) **Columns, not rows.** The j-th **column** is $[Tv_j]_{\mathcal{C}}$, as in the nonexample above. (2) **Order matters.** In Example (A Reflection in Two Bases), reordering the standard basis as $(\mathbf{e}_2, \mathbf{e}_1)$ leaves the matrix $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ unchanged, but reordering $\mathcal{B}$ as $(\mathbf{u}_2, \mathbf{u}_1)$ gives $\begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} \neq [T]_{\mathcal{B}}^{\mathcal{B}}$. (3) **Input subscript, output superscript.** $[T]_{\mathcal{B}}^{\mathcal{C}}$ and $[T]_{\mathcal{C}}^{\mathcal{B}}$ are different objects, and for $T\colon V \to W$ with $V \neq W$ only one of them makes sense. Also $[\mathrm{id}_V]_{\mathcal{B}}^{\mathcal{C}} \neq \mathbf{I}_n$ when $\mathcal{B} \neq \mathcal{C}$: for $\mathcal{B} = (\mathbf{e}_1, \mathbf{e}_2)$ and $\mathcal{C} = (\mathbf{e}_2, \mathbf{e}_1)$ in $\mathbb{R}^2$, $[\mathrm{id}]_{\mathcal{B}}^{\mathcal{C}} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$.

Quick check. What is the size of $[T]_{\mathcal{B}}^{\mathcal{C}}$ for a linear map $T\colon M_{2\times 3}(\mathbb{R}) \to \mathbb{R}[x]_{\leq 2}$? And what is the first column of $[T]_{\mathcal{E}}^{\mathcal{E}'}$ for $T\colon \mathbb{R}[x]_{\leq 2} \to \mathbb{R}^2$, $T(p) = (p(0), p(1))$, with $\mathcal{E} = (1, x, x^2)$ and $\mathcal{E}'$ the standard basis of $\mathbb{R}^2$?

Solution. $\dim \mathbb{R}[x]_{\leq 2} = 3$ and $\dim M_{2\times 3}(\mathbb{R}) = 6$, so the matrix is 3×6: rows for the codomain, columns for the domain. For the second map, the first column is the coordinate vector of the image of the first basis vector 1: $T(1) = (1, 1)$, so the column is $(1, 1)$.

4.1.2 The fundamental square

Here is the promise behind the definition: once coordinates are taken on both sides, applying T **is** multiplying by its matrix.

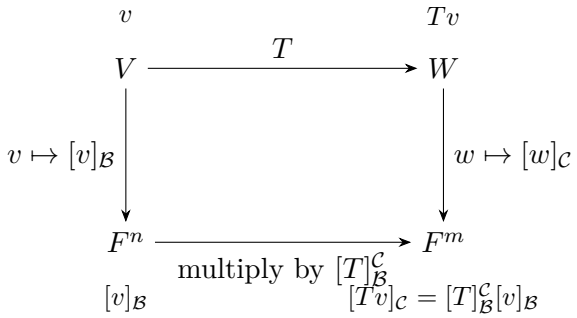
Theorem 4.1.1. Coordinates of an Image

Let V and W be finite-dimensional vector spaces over F with ordered bases $\mathcal{B} = (v_1, \dots, v_n)$ and $\mathcal{C} = (w_1, \dots, w_m)$, and let $T \in \mathcal{L}(V, W)$. Then for every $v \in V$,

$$[Tv]_{\mathcal{C}} = [T]_{\mathcal{B}}^{\mathcal{C}}\,[v]_{\mathcal{B}}.$$

Moreover, $[T]_{\mathcal{B}}^{\mathcal{C}}$ is the **only** matrix $\mathbf{A} \in M_{m\times n}(F)$ with $[Tv]_{\mathcal{C}} = \mathbf{A}[v]_{\mathcal{B}}$ for all $v \in V$.

The statement says that the following square **commutes**: starting from v in the top-left corner, the two routes to the bottom-right corner give the same column.



The top row is the abstract world, where T acts on vectors of V. The bottom row is the matrix world, where a matrix acts on columns. The vertical arrows are the coordinate isomorphisms of Corollary 2.5.5, so nothing is lost going down or coming back up. **Right then down** (apply T, then take coordinates) equals **down then right** (take coordinates, then multiply by the matrix).

Idea. Both sides are linear in v, so it is enough to check the formula on the basis vectors v_j, and there it is the definition: $[v_j]_{\mathcal{B}} = \mathbf{e}_j$, and multiplying a matrix by $\mathbf{e}_j$ picks out its j-th column, which was defined to be $[Tv_j]_{\mathcal{C}}$. Written out for a general v, this becomes a one-line chain. Uniqueness uses the same observation backwards: any $\mathbf{A}$ that works must have $[Tv_j]_{\mathcal{C}}$ as its j-th column.

Proof. Let $v \in V$ and $[v]_{\mathcal{B}} = (x_1, \dots, x_n)$, so that $v = x_1 v_1 + \cdots + x_n v_n$. Since T is linear, Theorem 2.1.4 gives $Tv = x_1 Tv_1 + \cdots + x_n Tv_n$. Since taking coordinates respects sums and scalar multiples (Theorem 1.5.2),

$$[Tv]_{\mathcal{C}} = x_1[Tv_1]_{\mathcal{C}} + \cdots + x_n[Tv_n]_{\mathcal{C}} = [T]_{\mathcal{B}}^{\mathcal{C}} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = [T]_{\mathcal{B}}^{\mathcal{C}}[v]_{\mathcal{B}},$$

where the middle equality is Theorem 0.9.3 applied to the columns $[Tv_j]_{\mathcal{C}}$ of $[T]_{\mathcal{B}}^{\mathcal{C}}$.

For uniqueness, suppose $\mathbf{A} \in M_{m\times n}(F)$ satisfies $[Tv]_{\mathcal{C}} = \mathbf{A}[v]_{\mathcal{B}}$ for all v. Taking $v = v_j$, whose coordinate vector is $\mathbf{e}_j$, the j-th column of $\mathbf{A}$ is $\mathbf{A}\mathbf{e}_j = [Tv_j]_{\mathcal{C}}$ (Theorem 0.9.3), which is the j-th column of $[T]_{\mathcal{B}}^{\mathcal{C}}$. Hence $\mathbf{A} = [T]_{\mathcal{B}}^{\mathcal{C}}$. This proves the theorem. ∎

This is the dictionary that lets us "write down a matrix first". A question about T (its kernel, its image, a preimage) becomes a question about a matrix, which Chapter 3 taught us to answer by row reduction; the coordinate isomorphisms carry the answer back.

Example. Kernel and Image Through a Matrix

Let $T\colon \mathbb{R}[x]_{\le 2} \to \mathbb{R}^2$, $T(p) = (p(1), p(-1))$. Find $\ker T$ and decide whether T is surjective, by working with the matrix of T in the bases $\mathcal{E} = (1, x, x^2)$ and the standard basis $\mathcal{E}'$ of $\mathbb{R}^2$.

Solution. Write down the matrix. $T(1) = (1,1)$, $T(x) = (1,-1)$, $T(x^2) = (1,1)$, so

$$\mathbf{A} = [T]_{\mathcal{E}}^{\mathcal{E}'} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & -1 & 1 \end{pmatrix}.$$

Solve in the matrix world. Subtracting the first row from the second gives $\begin{pmatrix} 1 & 1 & 1 \\ 0 & -2 & 0 \end{pmatrix}$, so $\mathbf{Ax} = \mathbf{0}$ means $x_2 = 0$ and $x_1 = -x_3$: $\operatorname{nul}(\mathbf{A}) = \operatorname{span}((-1,0,1))$. Both rows are pivot rows, so $\operatorname{rank}\mathbf{A} = 2$.

Translate back. By Theorem 4.1.1, $T(p) = \mathbf{0}$ exactly when $\mathbf{A}[p]_{\mathcal{E}} = \mathbf{0}$ (the only vector with zero coordinates is $\mathbf{0}$). The coordinate vector $(-1,0,1)$ belongs to $p = -1 + x^2$. So

$$\ker T = \operatorname{span}(x^2 - 1),$$

as a check confirms: $x^2 - 1$ vanishes at 1 and -1. By Theorem 2.3.1, $\operatorname{rank} T = 3 - 1 = 2 = \dim \mathbb{R}^2$, so $\operatorname{im} T = \mathbb{R}^2$ by Theorem 1.6.8, and T is surjective: any two values $p(1)$, $p(-1)$ can be prescribed.

4.1.3 Composition is matrix multiplication

In Chapter 0, §9 we asked why matrices are multiplied row-against-column. The answer given there was **substitution**. If $\mathbf{y} = \mathbf{Bx}$ and $\mathbf{z} = \mathbf{Ay}$, with

$$\mathbf{B} = \begin{pmatrix} 2 & 1 \\ 1 & 4 \end{pmatrix}, \qquad \mathbf{A} = \begin{pmatrix} 1 & 2 \\ 3 & -1 \end{pmatrix},$$

then substituting gives $\mathbf{z} = \begin{pmatrix} 4 & 9 \\ 5 & -1 \end{pmatrix}\mathbf{x}$, and the table of the combined substitution turned out to be $\mathbf{AB}$. At the time we promised that this fact would become a statement about linear maps. We can now say what it is. Each substitution is a linear map, $T_{\mathbf{B}}\colon \mathbf{x} \mapsto \mathbf{y}$ and $T_{\mathbf{A}}\colon \mathbf{y} \mapsto \mathbf{z}$, and substituting one into the other is **composing** them: $\mathbf{z} = T_{\mathbf{A}}(T_{\mathbf{B}}\mathbf{x}) = (T_{\mathbf{A}}T_{\mathbf{B}})\mathbf{x}$. So the Chapter 0 computation says $T_{\mathbf{A}}T_{\mathbf{B}} = T_{\mathbf{AB}}$. Here is the general statement, for arbitrary spaces and bases.

Theorem 4.1.2. Matrix of a Composition

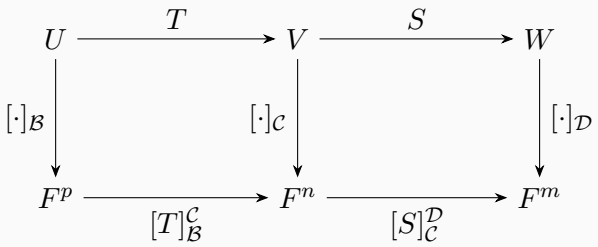
Let U, V, W be finite-dimensional vector spaces over F with ordered bases $\mathcal{B}$, $\mathcal{C}$, $\mathcal{D}$ respectively. Let $T \in \mathcal{L}(U,V)$ and $S \in \mathcal{L}(V,W)$. Then

$$[ST]_{\mathcal{B}}^{\mathcal{D}} = [S]_{\mathcal{C}}^{\mathcal{D}}\,[T]_{\mathcal{B}}^{\mathcal{C}}.$$

Notice how the notation checks itself: the middle basis $\mathcal{C}$ appears as the output of T and the input of S, and it "cancels", leaving $\mathcal{B}$ below and $\mathcal{D}$ above. The order of the factors is the order of the maps: the map applied first is on the right, in both ST and the product.

Idea. Glue two commutative squares side by side. The left square is T with its matrix, the right square is S with its matrix, and they share the middle column $V \to F^{\dim V}$. Going along the top from U to W is ST; going along the bottom is multiplication by $[T]_{\mathcal{B}}^{\mathcal{C}}$ followed by multiplication by $[S]_{\mathcal{C}}^{\mathcal{D}}$, which is multiplication by their product. The big rectangle commutes because both small squares do, and the uniqueness part of Theorem 4.1.1 identifies the

product as the matrix of ST.



Proof. Let $p = \dim U$ and $m = \dim W$, and let $u \in U$. Applying Theorem 4.1.1 first to S and the vector Tu, then to T and u,

$$[STu]_{\mathcal{D}} = [S(Tu)]_{\mathcal{D}} = [S]_{\mathcal{C}}^{\mathcal{D}}[Tu]_{\mathcal{C}} = [S]_{\mathcal{C}}^{\mathcal{D}}([T]_{\mathcal{B}}^{\mathcal{C}}[u]_{\mathcal{B}}) = ([S]_{\mathcal{C}}^{\mathcal{D}}[T]_{\mathcal{B}}^{\mathcal{C}})[u]_{\mathcal{B}},$$

where the last equality is associativity of matrix multiplication (Theorem 0.9.4). The map ST is linear (Theorem 2.4.2), and the $m \times p$ matrix $[S]_{\mathcal{C}}^{\mathcal{D}}[T]_{\mathcal{B}}^{\mathcal{C}}$ satisfies $[STu]_{\mathcal{D}} = ([S]_{\mathcal{C}}^{\mathcal{D}}[T]_{\mathcal{B}}^{\mathcal{C}})[u]_{\mathcal{B}}$ for every $u \in U$. By the uniqueness part of Theorem 4.1.1, it equals $[ST]_{\mathcal{B}}^{\mathcal{D}}$. This proves the theorem.

■

This is the payoff of the hook in Chapter 0, §9 (Definition 0.9.9). The row-times-column rule was not an arbitrary convention: it is the **only** rule under which the matrix of a composition is the product of the matrices, because composing linear maps is substituting one coordinate formula into another. Everything Chapter 0 proved about matrix multiplication now has a meaning for maps, and conversely:

- **associativity** of matrix multiplication is associativity of composition (Theorem 0.4.2);
- $\mathbf{AB} \neq \mathbf{BA}$ in general is $ST \neq TS$ (Example (Rotate, Then Project));
- the size condition "columns of $\mathbf{A}$ = rows of $\mathbf{B}$" is "the codomain of the first map is the domain of the second";
- in particular $T_{\mathbf{A}}T_{\mathbf{B}} = T_{\mathbf{AB}}$: apply the theorem with standard bases and Example (Matrix Maps and the Identity) (a).

Example. Multiply by x, Then Differentiate

Let $S\colon \mathbb{R}[x]_{\leq 2} \to \mathbb{R}[x]_{\leq 3}$, $S(p) = xp$, and $D\colon \mathbb{R}[x]_{\leq 3} \to \mathbb{R}[x]_{\leq 2}$, $D(p) = p'$, with the monomial bases $\mathcal{E}_2 = (1, x, x^2)$ and $\mathcal{E}_3 = (1, x, x^2, x^3)$. Compute $[DS]_{\mathcal{E}_2}^{\mathcal{E}_2}$ and $[SD]_{\mathcal{E}_3}^{\mathcal{E}_3}$, both directly and as products.

Solution. Matrices of the factors. $S(1) = x$, $S(x) = x^2$, $S(x^2) = x^3$, and $[D]_{\mathcal{E}_3}^{\mathcal{E}_2}$ was found in Example (Differentiation on Cubic Polynomials):

$$[S]_{\mathcal{E}_2}^{\mathcal{E}_3} = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \qquad [D]_{\mathcal{E}_3}^{\mathcal{E}_2} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 3 \end{pmatrix}.$$

Products. By Theorem 4.1.2,

$$[DS]_{\mathcal{E}_2}^{\mathcal{E}_2} = [D]_{\mathcal{E}_3}^{\mathcal{E}_2}[S]_{\mathcal{E}_2}^{\mathcal{E}_3} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{pmatrix}, \qquad [SD]_{\mathcal{E}_3}^{\mathcal{E}_3} = [S]_{\mathcal{E}_2}^{\mathcal{E}_3}[D]_{\mathcal{E}_3}^{\mathcal{E}_2} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 3 \end{pmatrix}.$$

Directly. $DS(x^k) = (x^{k+1})' = (k+1)x^k$ for $k = 0, 1, 2$, giving the diagonal $1, 2, 3$. And $SD(x^k) = x \cdot kx^{k-1} = kx^k$ for $k \geq 1$, with $SD(1) = 0$, giving the diagonal $0, 1, 2, 3$. Both computations agree. The two composites are not even the same size: DS is an operator on $\mathbb{R}[x]_{\leq 2}$ and SD an operator on $\mathbb{R}[x]_{\leq 3}$.

For operators, the theorem applies with $\mathcal{B} = \mathcal{C} = \mathcal{D}$, and gives powers and polynomials.

Corollary 4.1.3. Matrices of Powers and Polynomials

Let $\mathcal{B} = (v_1, \ldots, v_n)$ be a basis of a finite-dimensional space V, let $T, S \in \mathcal{L}(V)$, $c \in F$, and $p \in F[x]$. Then $[ST]_{\mathcal{B}}^{\mathcal{B}} = [S]_{\mathcal{B}}^{\mathcal{B}}[T]_{\mathcal{B}}^{\mathcal{B}}$, $[S+T]_{\mathcal{B}}^{\mathcal{B}} = [S]_{\mathcal{B}}^{\mathcal{B}} + [T]_{\mathcal{B}}^{\mathcal{B}}$, $[cT]_{\mathcal{B}}^{\mathcal{B}} = c[T]_{\mathcal{B}}^{\mathcal{B}}$, $[T^k]_{\mathcal{B}}^{\mathcal{B}} = ([T]_{\mathcal{B}}^{\mathcal{B}})^k$ for $k \in \mathbb{N}$, and

$$[p(T)]_{\mathcal{B}}^{\mathcal{B}} = p([T]_{\mathcal{B}}^{\mathcal{B}}).$$

Proof. The first formula is Theorem 4.1.2 with $\mathcal{B} = \mathcal{C} = \mathcal{D}$. For sums and scalar multiples, the j-th column of $[S+T]_{\mathcal{B}}^{\mathcal{B}}$ is $[Sv_j + Tv_j]_{\mathcal{B}} = [Sv_j]_{\mathcal{B}} + [Tv_j]_{\mathcal{B}}$ by Theorem 1.5.2, and similarly $[cTv_j]_{\mathcal{B}} = c[Tv_j]_{\mathcal{B}}$. For powers we induct on k: $[T^0]_{\mathcal{B}}^{\mathcal{B}} = [\mathrm{id}_V]_{\mathcal{B}}^{\mathcal{B}} = \mathbf{I}_n$ by Example (Matrix Maps and the Identity) (b), and if $[T^k]_{\mathcal{B}}^{\mathcal{B}} = ([T]_{\mathcal{B}}^{\mathcal{B}})^k$, then $[T^{k+1}]_{\mathcal{B}}^{\mathcal{B}} = [T^k T]_{\mathcal{B}}^{\mathcal{B}} = ([T]_{\mathcal{B}}^{\mathcal{B}})^k [T]_{\mathcal{B}}^{\mathcal{B}} = ([T]_{\mathcal{B}}^{\mathcal{B}})^{k+1}$, by the first formula and Definition 0.9.14. Finally, for $p = \sum_k a_k x^k$, combining the three rules gives $[p(T)]_{\mathcal{B}}^{\mathcal{B}} = \sum_k a_k [T^k]_{\mathcal{B}}^{\mathcal{B}} = \sum_k a_k ([T]_{\mathcal{B}}^{\mathcal{B}})^k = p([T]_{\mathcal{B}}^{\mathcal{B}})$, using Definition 2.4.2 and Definition 0.9.14.

■

For example, with $[D]_{\mathcal{E}_3}^{\mathcal{E}_3}$ from Example (Differentiation on Cubic Polynomials), $([D]_{\mathcal{E}_3}^{\mathcal{E}_3})^4 = [D^4]_{\mathcal{E}_3}^{\mathcal{E}_3} = \mathbf{O}$, recovering Example (Differentiation Is Nilpotent) in matrix form.

4.1.4 Linear maps are matrices

The sums in Corollary 4.1.3 work between different spaces too, and together with the construction of maps from values on a basis they show that, after choosing bases, there is no difference at all between linear maps and matrices.

Theorem 4.1.4. $\mathcal{L}(V, W) \cong M_{m \times n}(F)$

Let V and W be finite-dimensional vector spaces over F with ordered bases $\mathcal{B} = (v_1, \ldots, v_n)$ and $\mathcal{C} = (w_1, \ldots, w_m)$. Then the map

$$\Phi \colon \mathcal{L}(V, W) \to M_{m \times n}(F), \qquad T \mapsto [T]_{\mathcal{B}}^{\mathcal{C}},$$

is an isomorphism. In particular, $\dim \mathcal{L}(V, W) = mn = (\dim V)(\dim W)$.

Idea. Three things to check. *Linear:* columns of $[S+T]_{\mathcal{B}}^{\mathcal{C}}$ are coordinates of $Sv_j + Tv_j$, and coordinates respect sums. *Injective:* a map whose matrix is zero kills every basis vector, so it is the zero map. *Surjective:* a matrix prescribes the value of each v_j, and a linear map with prescribed values on a basis exists. The last two are the uniqueness and existence halves of the same theorem about bases, which is why a linear map is "nothing but" its values on a basis.

Proof. Linear. Let $S, T \in \mathcal{L}(V, W)$ and $c \in F$. By Definition 2.4.1 and Theorem 1.5.2, the j-th column of $[S+T]_{\mathcal{B}}^{\mathcal{C}}$ is $[Sv_j+Tv_j]_{\mathcal{C}} = [Sv_j]_{\mathcal{C}}+[Tv_j]_{\mathcal{C}}$, and that of $[cT]_{\mathcal{B}}^{\mathcal{C}}$ is $[cTv_j]_{\mathcal{C}} = c[Tv_j]_{\mathcal{C}}$. Hence $\Phi(S+T) = \Phi(S) + \Phi(T)$ and $\Phi(cT) = c\,\Phi(T)$.

Injective. Suppose $\Phi(T) = \mathbf{O}$. Then $[Tv_j]_{\mathcal{C}} = \mathbf{0}$, so $Tv_j = 0$ for every j. The zero map also sends every v_j to 0, so $T = 0$ by the uniqueness part of Theorem 2.1.5. Hence $\ker \Phi = \{0\}$, and Φ is injective by Theorem 2.2.3.

Surjective. Let $\mathbf{A} = (a_{ij}) \in M_{m\times n}(F)$. By the existence part of Theorem 2.1.5 there is $T \in \mathcal{L}(V, W)$ with $Tv_j = a_{1j}w_1 + \cdots + a_{mj}w_m$ for each j. By Definition 4.1.1, $[T]_{\mathcal{B}}^{\mathcal{C}} = \mathbf{A}$.

So Φ is a bijective linear map, an isomorphism by Theorem 2.5.1. For the dimension, we cannot quote Theorem 2.5.4 directly, since we do not yet know that $\mathcal{L}(V, W)$ is finite-dimensional. Instead, $M_{m\times n}(F)$ has a basis of mn matrix units (Example (Dimensions of the Running Examples)), and the inverse isomorphism Φ^{-1} carries it to a basis of $\mathcal{L}(V, W)$ of length mn (Theorem 2.5.2, applied to Φ^{-1}, which is an isomorphism by Theorem 2.5.3). Hence $\dim \mathcal{L}(V, W) = mn$. This proves the theorem.

■

With Theorem 4.1.2, the isomorphism Φ also turns composition into multiplication. So, for $V = W$ and $\mathcal{B} = \mathcal{C}$, the algebra $\mathcal{L}(V)$ of the previous sections and the algebra $M_n(F)$ of Chapter 0 are the same algebra, written in two languages. The dictionary depends on the basis, as every isomorphism to F^n does.

4.1.5 Reading rank and invertibility from a matrix

Chapter 3 developed tools for the rank and the invertibility of matrices. The square of Theorem 4.1.1 transfers all of them to linear maps.

Theorem 4.1.5. Rank and Invertibility from the Matrix

Let V and W be finite-dimensional vector spaces over F with ordered bases $\mathcal{B}$ and $\mathcal{C}$, let $T \in \mathcal{L}(V, W)$, and let $\mathbf{A} = [T]_{\mathcal{B}}^{\mathcal{C}}$.

(a) The coordinate map $v \mapsto [v]_{\mathcal{B}}$ restricts to an isomorphism $\ker T \to \mathrm{nul}(\mathbf{A})$, and the coordinate map $w \mapsto [w]_{\mathcal{C}}$ restricts to an isomorphism $\operatorname{im} T \to \mathrm{col}(\mathbf{A})$. Hence

$$\operatorname{rank} T = \operatorname{rank} \mathbf{A} \qquad \text{and} \qquad \operatorname{nullity} T = \operatorname{nullity} \mathbf{A}.$$

(b) Suppose $\dim V = \dim W$. Then T is invertible if and only if $\mathbf{A}$ is invertible, and in

that case $[T^{-1}]_{\mathcal{C}}^{\mathcal{B}} = \mathbf{A}^{-1}$.

Idea. For (a), the commutative square says a vector is killed by T exactly when its coordinate column is killed by $\mathbf{A}$, and the outputs of T, in coordinates, are exactly the outputs of $\mathbf{A}$. Isomorphic subspaces have the same dimension. For (b), $T^{-1}T = \mathrm{id}_V$ translates, by the composition theorem, into a matrix equation with $\mathbf{I}_n$ on the right; conversely an inverse matrix is the matrix of **some** map, by the isomorphism $\mathcal{L}(W, V) \cong M_n(F)$, and that map is the inverse of T.

Proof. Let $n = \dim V$ and $m = \dim W$, and write $C_{\mathcal{B}} \colon V \to F^n$ and $C_{\mathcal{C}} \colon W \to F^m$ for the coordinate isomorphisms (Corollary 2.5.5).

(a) *Kernel.* For $v \in V$: $v \in \ker T$ if and only if $Tv = 0$ if and only if $[Tv]_{\mathcal{C}} = \mathbf{0}$ (since $C_{\mathcal{C}}$ is injective) if and only if $\mathbf{A}[v]_{\mathcal{B}} = \mathbf{0}$ (Theorem 4.1.1) if and only if $[v]_{\mathcal{B}} \in \mathrm{nul}(\mathbf{A})$. So $C_{\mathcal{B}}$ maps $\ker T$ into $\mathrm{nul}(\mathbf{A})$, and every $\mathbf{x} \in \mathrm{nul}(\mathbf{A})$ is $[v]_{\mathcal{B}}$ for some $v \in V$ (surjectivity of $C_{\mathcal{B}}$), which then lies in $\ker T$. Thus the restriction of $C_{\mathcal{B}}$ is a linear map $\ker T \to \mathrm{nul}(\mathbf{A})$, injective because $C_{\mathcal{B}}$ is, and surjective. By Theorem 2.5.1 it is an isomorphism.
Image. By Theorem 4.1.1, $C_{\mathcal{C}}(Tv) = \mathbf{A}\, C_{\mathcal{B}}(v)$ for every $v \in V$. As v runs over V, $C_{\mathcal{B}}(v)$ runs over all of F^n, so $\{C_{\mathcal{C}}(w) : w \in \mathrm{im}\, T\} = \{\mathbf{A}\mathbf{x} : \mathbf{x} \in F^n\} = \mathrm{col}(\mathbf{A})$ (Definition 3.5.1). So the restriction of $C_{\mathcal{C}}$ is a linear map $\mathrm{im}\, T \to \mathrm{col}(\mathbf{A})$, surjective by this computation and injective because $C_{\mathcal{C}}$ is; hence an isomorphism.
The subspaces involved are finite-dimensional (Theorem 1.6.7), so Theorem 2.5.4 gives $\dim \ker T = \dim \mathrm{nul}(\mathbf{A})$ and $\dim \mathrm{im}\, T = \dim \mathrm{col}(\mathbf{A})$, which are the stated equalities by Definition 2.3.1, Definition 2.3.2, Definition 3.5.5, and Definition 3.5.4.

(b) ($\Rightarrow$) Suppose T is invertible. By Theorem 4.1.2 and Example (Matrix Maps and the Identity) (b),

$$[T^{-1}]_{\mathcal{C}}^{\mathcal{B}}\, \mathbf{A} = [T^{-1}T]_{\mathcal{B}}^{\mathcal{B}} = [\mathrm{id}_V]_{\mathcal{B}}^{\mathcal{B}} = \mathbf{I}_n, \qquad \mathbf{A}\, [T^{-1}]_{\mathcal{C}}^{\mathcal{B}} = [TT^{-1}]_{\mathcal{C}}^{\mathcal{C}} = [\mathrm{id}_W]_{\mathcal{C}}^{\mathcal{C}} = \mathbf{I}_n.$$

So $\mathbf{A}$ is invertible with $\mathbf{A}^{-1} = [T^{-1}]_{\mathcal{C}}^{\mathcal{B}}$ (Definition 0.9.13).

($\Leftarrow$) Suppose $\mathbf{A}$ is invertible. By Theorem 4.1.4 (for $\mathcal{L}(W, V)$ with bases $\mathcal{C}$ and $\mathcal{B}$), there is $S \in \mathcal{L}(W, V)$ with $[S]_{\mathcal{C}}^{\mathcal{B}} = \mathbf{A}^{-1}$. Then by Theorem 4.1.2, $[ST]_{\mathcal{B}}^{\mathcal{B}} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}_n = [\mathrm{id}_V]_{\mathcal{B}}^{\mathcal{B}}$. Since $T' \mapsto [T']_{\mathcal{B}}^{\mathcal{B}}$ is injective on $\mathcal{L}(V)$ (Theorem 4.1.4), $ST = \mathrm{id}_V$. By Theorem 2.5.6, T is invertible with $T^{-1} = S$. This proves the theorem.

■

So we may add one more item to the Invertible Operator Theorem (Theorem 2.5.7): for $T \in \mathcal{L}(V)$ and **any** basis $\mathcal{B}$ of V, T is invertible if and only if $[T]_{\mathcal{B}}^{\mathcal{B}}$ is invertible. Every item of the Invertible Matrix Theorem (Theorem 3.4.4) now becomes a test for operators.

The first thing the square buys is a way to compute a kernel and an image for a map given by a matrix: the two subspaces are the null space and the column space, and elimination finds them.

Example. A Matrix Map

Let $\mathbf{A} = \begin{pmatrix} 1 & 2 & 0 & 3 \\ 2 & 4 & 1 & 8 \\ -1 & -2 & 1 & -1 \end{pmatrix} \in M_{3\times 4}(\mathbb{R})$. Find bases of $\ker T_{\mathbf{A}}$ and $\operatorname{im} T_{\mathbf{A}}$ for $T_{\mathbf{A}}\colon \mathbb{R}^4 \to \mathbb{R}^3$, and decide whether $T_{\mathbf{A}}$ is injective or surjective.

Solution. By definition, $\ker T_{\mathbf{A}} = \{\mathbf{x} : \mathbf{A}\mathbf{x} = \mathbf{0}\} = \operatorname{nul}(\mathbf{A})$ (Definition 3.5.3) and $\operatorname{im} T_{\mathbf{A}} = \{\mathbf{A}\mathbf{x} : \mathbf{x} \in \mathbb{R}^4\} = \operatorname{col}(\mathbf{A})$ (Definition 3.5.1). So Chapter 3's methods apply directly. Row reduce:

$$\mathbf{A} \xrightarrow[R_3+R_1]{R_2-2R_1} \begin{pmatrix} 1 & 2 & 0 & 3 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 1 & 2 \end{pmatrix} \xrightarrow{R_3-R_2} \begin{pmatrix} 1 & 2 & 0 & 3 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

This is the RREF, with pivot columns 1 and 3 and free columns 2 and 4. By Theorem 3.5.4, a basis of $\ker T_{\mathbf{A}}$ is

$$((-2,1,0,0),\ (-3,0,-2,1)).$$

By Theorem 3.5.2, the pivot columns of $\mathbf{A}$ form a basis of $\operatorname{im} T_{\mathbf{A}}$: $((1,2,-1),\ (0,1,1))$.

Since $\ker T_{\mathbf{A}} \neq \{\mathbf{0}\}$, $T_{\mathbf{A}}$ is not injective (Theorem 2.2.3); for instance $T_{\mathbf{A}}(-2,1,0,0) = \mathbf{0} = T_{\mathbf{A}}(\mathbf{0})$. Since $\dim \operatorname{im} T_{\mathbf{A}} = 2 < 3 = \dim \mathbb{R}^3$, $\operatorname{im} T_{\mathbf{A}} \neq \mathbb{R}^3$, so $T_{\mathbf{A}}$ is not surjective (Theorem 2.2.2).

Example. Rank of $\mathbf{X} \mapsto \mathbf{X} + \mathbf{X}^\top$

Let $T\colon M_2(\mathbb{R}) \to M_2(\mathbb{R})$, $T(\mathbf{X}) = \mathbf{X} + \mathbf{X}^\top$, and $\mathcal{E} = (\mathbf{E}_{11}, \mathbf{E}_{12}, \mathbf{E}_{21}, \mathbf{E}_{22})$. Find $[T]_{\mathcal{E}}^{\mathcal{E}}$, use it to find $\operatorname{rank} T$, a basis of $\ker T$, and a basis of $\operatorname{im} T$, and decide whether T is invertible. Then do the same for the transpose map of Example (The Transpose Map).

Solution. Matrix. $T(\mathbf{E}_{11}) = 2\mathbf{E}_{11}$, $T(\mathbf{E}_{12}) = \mathbf{E}_{12} + \mathbf{E}_{21} = T(\mathbf{E}_{21})$, $T(\mathbf{E}_{22}) = 2\mathbf{E}_{22}$. So

$$\mathbf{A} = [T]_{\mathcal{E}}^{\mathcal{E}} = \begin{pmatrix} 2 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 2 \end{pmatrix}.$$

Matrix world. Subtracting row 2 from row 3 turns row 3 into a zero row; moving it to the bottom gives an echelon form with three nonzero rows and pivots in columns 1, 2, 4. So $\operatorname{rank} \mathbf{A} = 3$ and $\operatorname{nullity} \mathbf{A} = 1$. The free column is column 3, and $\mathbf{A}\mathbf{x} = \mathbf{0}$ gives $x_1 = x_4 = 0$, $x_2 = -x_3$, so $\operatorname{nul}(\mathbf{A}) = \operatorname{span}((0,-1,1,0))$. By Theorem 3.5.2, the pivot columns $(2,0,0,0)$, $(0,1,1,0)$, $(0,0,0,2)$ form a basis of $\operatorname{col}(\mathbf{A})$.

Translate back. By Theorem 4.1.5, $\operatorname{rank} T = 3$. The column $(0,-1,1,0)$ is the coordinate vector of $\mathbf{E}_{21} - \mathbf{E}_{12}$, so $\ker T = \operatorname{span}\left(\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}\right)$, the skew-symmetric matrices. The basis of $\operatorname{col}(\mathbf{A})$ translates to $2\mathbf{E}_{11}$, $\mathbf{E}_{12} + \mathbf{E}_{21}$, $2\mathbf{E}_{22}$, a basis of $\operatorname{im} T$, which is therefore the space of symmetric matrices. Since $\operatorname{rank} \mathbf{A} = 3 < 4$, $\mathbf{A}$ is not invertible (Theorem 3.4.4), and neither is T.

Transpose. The matrix $\mathbf{P} = [\mathbf{X} \mapsto \mathbf{X}^\top]_{\mathcal{E}}^{\mathcal{E}}$ from Example (The Transpose Map) satisfies $\mathbf{P}^2 = \mathbf{I}_4$ (swapping two coordinates twice does nothing). So $\mathbf{P}$ is invertible with $\mathbf{P}^{-1} = \mathbf{P}$,

and by Theorem 4.1.5 (b) the transpose map is invertible and is its own inverse, which is $(\mathbf{X}^\top)^\top = \mathbf{X}$ again.

We now have both halves of the working method: a basis turns spaces into F^n and maps into matrices, and every computation can be done with Chapter 3's tools. What we have not yet controlled is the choice. Example (A Reflection in Two Bases) shows that one map can have a complicated matrix in one basis and a diagonal one in another; the next section finds the formula relating the two.

4.1.6 Exercises

4.1.6.1 A. Check your understanding

Exercise. A1

(a) Define $[T]_{\mathcal{B}}^{\mathcal{C}}$ for $T \in \mathcal{L}(V, W)$ and ordered bases $\mathcal{B}$ of V, $\mathcal{C}$ of W.

(b) State the formula relating $[Tv]_{\mathcal{C}}$ and $[v]_{\mathcal{B}}$, and the formula for the matrix of a composition.

(c) True or false: the j-th **row** of $[T]_{\mathcal{B}}^{\mathcal{C}}$ is $[Tv_j]_{\mathcal{C}}$. Justify your answer.

(d) True or false: $[\mathrm{id}_V]_{\mathcal{B}}^{\mathcal{C}} = \mathbf{I}_n$ for any two bases $\mathcal{B}$, $\mathcal{C}$ of an n-dimensional space V. Justify your answer.

(e) What is $\dim \mathcal{L}(\mathbb{R}^3, \mathbb{R}[x]_{\le 2})$?

Solution.

(a) If $\mathcal{B} = (v_1, \dots, v_n)$, then $[T]_{\mathcal{B}}^{\mathcal{C}}$ is the $\dim W \times n$ matrix whose j-th column is $[Tv_j]_{\mathcal{C}}$ (Definition 4.1.1).

(b) $[Tv]_{\mathcal{C}} = [T]_{\mathcal{B}}^{\mathcal{C}}[v]_{\mathcal{B}}$ (Theorem 4.1.1), and $[ST]_{\mathcal{B}}^{\mathcal{D}} = [S]_{\mathcal{C}}^{\mathcal{D}}[T]_{\mathcal{B}}^{\mathcal{C}}$ (Theorem 4.1.2).

(c) False. It is the j-th **column**. For $D\colon \mathbb{R}[x]_{\le 3} \to \mathbb{R}[x]_{\le 2}$ in Example (Differentiation on Cubic Polynomials), $[D(x)]_{\mathcal{E}_2} = (1, 0, 0)$ is the second column, while the second row is $(0, 0, 2, 0)$.

(d) False. In $\mathbb{R}^2$ with $\mathcal{B} = (\mathbf{e}_1, \mathbf{e}_2)$ and $\mathcal{C} = (\mathbf{e}_2, \mathbf{e}_1)$, $[\mathbf{e}_1]_{\mathcal{C}} = (0, 1)$, so the first column of $[\mathrm{id}]_{\mathcal{B}}^{\mathcal{C}}$ is $(0, 1) \neq \mathbf{e}_1$. It is true when $\mathcal{B} = \mathcal{C}$ (Example (Matrix Maps and the Identity) (b)).

(e) $3 \cdot 3 = 9$, by Theorem 4.1.4.

4.1.6.2 B. Practice

Exercise. B1: Matrices in given bases

Find the matrix of each operator in the given basis. (Both maps are linear; you may take this as given.)

(a) $T\colon \mathbb{R}[x]_{\le 2} \to \mathbb{R}[x]_{\le 2}$, $T(p) = p(x + 1)$, in $\mathcal{E} = (1, x, x^2)$.

(b) $T\colon M_2(\mathbb{R}) \to M_2(\mathbb{R})$, $T(\mathbf{X}) = \mathbf{BX}$ with $\mathbf{B} = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$, in $\mathcal{E} = (\mathbf{E}_{11}, \mathbf{E}_{12}, \mathbf{E}_{21}, \mathbf{E}_{22})$.

Solution.

(a) $T(1) = 1$, $T(x) = x+1 = 1+x$, $T(x^2) = (x+1)^2 = 1+2x+x^2$. Their coordinates are $(1,0,0)$, $(1,1,0)$, $(1,2,1)$, so

$$[T]_{\mathcal{E}}^{\mathcal{E}} = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{pmatrix}.$$

(b) Multiplying $\mathbf{B}$ by a matrix unit $\mathbf{E}_{kj}$ on the right keeps column k of $\mathbf{B}$ and moves it to column j:

$$\mathbf{B}\mathbf{E}_{11} = \begin{pmatrix} 1 & 0 \\ 3 & 0 \end{pmatrix}, \quad \mathbf{B}\mathbf{E}_{12} = \begin{pmatrix} 0 & 1 \\ 0 & 3 \end{pmatrix}, \quad \mathbf{B}\mathbf{E}_{21} = \begin{pmatrix} 2 & 0 \\ 4 & 0 \end{pmatrix}, \quad \mathbf{B}\mathbf{E}_{22} = \begin{pmatrix} 0 & 2 \\ 0 & 4 \end{pmatrix}.$$

Reading entries in the order $(1,1),(1,2),(2,1),(2,2)$ gives the columns $(1,0,3,0)$, $(0,1,0,3)$, $(2,0,4,0)$, $(0,2,0,4)$, so

$$[T]_{\mathcal{E}}^{\mathcal{E}} = \begin{pmatrix} 1 & 0 & 2 & 0 \\ 0 & 1 & 0 & 2 \\ 3 & 0 & 4 & 0 \\ 0 & 3 & 0 & 4 \end{pmatrix}.$$

Exercise. B2: Recovering a map from its matrix

Let $\mathcal{B} = (1,\, 1+x,\, x^2)$, a basis of $\mathbb{R}[x]_{\le 2}$, and let $\mathcal{E}'$ be the standard basis of $\mathbb{R}^2$. A linear map $T\colon \mathbb{R}[x]_{\le 2} \to \mathbb{R}^2$ has

$$[T]_{\mathcal{B}}^{\mathcal{E}'} = \begin{pmatrix} 1 & 0 & 2 \\ 0 & 1 & -1 \end{pmatrix}.$$

Find $T(a+bx+cx^2)$. Hence find a basis of $\ker T$.

Solution. First find the coordinates of $p = a+bx+cx^2$ in $\mathcal{B}$: $p = (a-b)\cdot 1 + b(1+x) + cx^2$, so $[p]_{\mathcal{B}} = (a-b,\, b,\, c)$, unique by Theorem 1.5.1. By Theorem 4.1.1, and since coordinates in $\mathcal{E}'$ are the entries,

$$T(p) = \begin{pmatrix} 1 & 0 & 2 \\ 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} a-b \\ b \\ c \end{pmatrix} = (a-b+2c,\; b-c).$$

Hence $T(p) = \mathbf{0}$ if and only if $b = c$ and $a = b-2c = -c$, that is, $p = c(-1+x+x^2)$. So (x^2+x-1) is a basis of $\ker T$. Check: $[x^2+x-1]_{\mathcal{B}} = (-2,1,1)$, and $\begin{pmatrix} 1 & 0 & 2 \\ 0 & 1 & -1 \end{pmatrix}(-2,1,1) = (0,0)$.

Exercise. B3: A composition computed two ways

Let $T\colon \mathbb{R}[x]_{\le 1} \to \mathbb{R}[x]_{\le 2}$, $T(p) = xp$, and $S\colon \mathbb{R}[x]_{\le 2} \to \mathbb{R}^2$, $S(p) = (p(0), p(1))$. Use the

monomial bases and the standard basis of $\mathbb{R}^2$. Compute the matrix of ST directly, and again as a product of the matrices of S and T.

Solution. Let $\mathcal{E}_1 = (1, x)$, $\mathcal{E}_2 = (1, x, x^2)$, and $\mathcal{E}'$ the standard basis of $\mathbb{R}^2$.

Factors. $T(1) = x$, $T(x) = x^2$, so $[T]_{\mathcal{E}_1}^{\mathcal{E}_2} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}$. $S(1) = (1, 1)$, $S(x) = (0, 1)$, $S(x^2) = (0, 1)$, so $[S]_{\mathcal{E}_2}^{\mathcal{E}'} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \end{pmatrix}$.

Product. By Theorem 4.1.2,

$$[ST]_{\mathcal{E}_1}^{\mathcal{E}'} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 1 & 1 \end{pmatrix}.$$

Directly. $ST(1) = S(x) = (0, 1)$ and $ST(x) = S(x^2) = (0, 1)$, so the columns are $(0, 1)$ and $(0, 1)$, in agreement. In words, $ST(a + bx) = (0,\ a + b)$.

4.1.6.3 C. Going deeper

Exercise. C1: Upper triangular matrices and invariant subspaces

Let V be finite-dimensional and $T \in \mathcal{L}(V)$. A subspace U of V is **T-invariant** if $Tu \in U$ for every $u \in U$. Let $\mathcal{B} = (v_1, \dots, v_n)$ be a basis of V, and for $1 \le k \le n$ let $U_k = \operatorname{span}(v_1, \dots, v_k)$. Prove that $[T]_{\mathcal{B}}^{\mathcal{B}}$ is upper triangular if and only if every U_k is T-invariant.

Solution. Let $[T]_{\mathcal{B}}^{\mathcal{B}} = (a_{ij})$, so $Tv_j = \sum_{i=1}^{n} a_{ij} v_i$ with unique coefficients (Definition 4.1.1). By Definition 0.9.5, $[T]_{\mathcal{B}}^{\mathcal{B}}$ is upper triangular if and only if $a_{ij} = 0$ whenever $i > j$, that is, if and only if

$$Tv_j \in U_j \quad \text{for every } j = 1, \dots, n. \tag{$*$}$$

Indeed, if $a_{ij} = 0$ for $i > j$, then $Tv_j = \sum_{i \le j} a_{ij} v_i \in U_j$; conversely, if $Tv_j = \sum_{i \le j} b_i v_i$, then by uniqueness of coordinates $a_{ij} = 0$ for $i > j$.

($\Rightarrow$) Assume $(*)$, and fix k. For $j \le k$, $Tv_j \in U_j \subseteq U_k$. Let $u = \sum_{j \le k} c_j v_j \in U_k$. Then $Tu = \sum_{j \le k} c_j Tv_j$ by Theorem 2.1.4, a combination of vectors of the subspace U_k, so $Tu \in U_k$ by Theorem 1.3.1. Hence U_k is T-invariant.

($\Leftarrow$) Assume every U_k is T-invariant. For each j, $v_j \in U_j$, so $Tv_j \in U_j$ by invariance of U_j. This is $(*)$, so $[T]_{\mathcal{B}}^{\mathcal{B}}$ is upper triangular.

Exercise. C2: Integrating with a matrix

Let $f_1(x) = e^x$, $f_2(x) = xe^x$, $f_3(x) = x^2 e^x$, and let $V = \operatorname{span}(f_1, f_2, f_3)$ inside the real vector space of all functions $\mathbb{R} \to \mathbb{R}$ (Example (The Standard Vector Spaces) (d)).

(a) Prove that $\mathcal{B} = (f_1, f_2, f_3)$ is a basis of V.

(b) Show that differentiation maps V into V, and find $[D]_{\mathcal{B}}^{\mathcal{B}}$ for $D \in \mathcal{L}(V)$, $D(f) = f'$.

(c) Hence show that D is invertible on V, and use $([D]_{\mathcal{B}}^{\mathcal{B}})^{-1}$ to find a function $g \in V$ with $g' = x^2 e^x$.

Hint: for (a), divide a vanishing combination by e^x and count roots.

Solution.

(a) $\mathcal{B}$ spans V by definition. Suppose $af_1 + bf_2 + cf_3 = 0$, the zero function. Then $(a + bx + cx^2)e^x = 0$ for every $x \in \mathbb{R}$, and $e^x \neq 0$, so the polynomial $a + bx + cx^2$ vanishes at every real number. It has infinitely many roots, so it is the zero polynomial by Lemma 3.7.2, and $a = b = c = 0$. Hence $\mathcal{B}$ is independent, and a basis of V.

(b) By the product rule,

$$f_1' = e^x = f_1, \qquad f_2' = e^x + xe^x = f_1 + f_2, \qquad f_3' = 2xe^x + x^2e^x = 2f_2 + f_3.$$

Each lies in V, and D is linear, so D maps every combination of f_1, f_2, f_3 into V (Theorem 2.1.4). Reading off coordinates as columns,

$$[D]_{\mathcal{B}}^{\mathcal{B}} = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{pmatrix}.$$

(c) The matrix is upper triangular with nonzero diagonal entries, and one checks directly that

$$\begin{pmatrix} 1 & -1 & 2 \\ 0 & 1 & -2 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{pmatrix} = \mathbf{I}_3,$$

so $[D]_{\mathcal{B}}^{\mathcal{B}}$ is invertible with this inverse (Theorem 3.4.5). By Theorem 4.1.5 (b), D is invertible on V and $[D^{-1}]_{\mathcal{B}}^{\mathcal{B}} = ([D]_{\mathcal{B}}^{\mathcal{B}})^{-1}$. We want $g = D^{-1}(f_3)$. Since $[f_3]_{\mathcal{B}} = (0, 0, 1)$, Theorem 4.1.1 gives

$$[g]_{\mathcal{B}} = \begin{pmatrix} 1 & -1 & 2 \\ 0 & 1 & -2 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 2 \\ -2 \\ 1 \end{pmatrix},$$

so $g = 2f_1 - 2f_2 + f_3 = (x^2 - 2x + 2)e^x$. Check: $g' = (2x - 2)e^x + (x^2 - 2x + 2)e^x = x^2e^x$. No integration by parts was needed.

4.2 Change of Basis and Similarity

The matrix $[T]_{\mathcal{B}}^{\mathcal{C}}$ of a linear map depends on two choices of basis, and different choices can give very different matrices for the same map. This section answers two questions left open by that fact. How are two matrices of the same map related? And which properties of a matrix belong to the map itself, so that no choice of basis can change them? The answers are the change-of-basis formula, drawn as a commutative square, and the relation of similarity.

4.2.1 Change-of-coordinates matrices

Recall the definition of the matrix of a map (Definition 4.1.1). If $\mathcal{B} = (v_1, \dots, v_n)$ is a basis of V and $\mathcal{C}$ is a basis of W, then $[T]_{\mathcal{B}}^{\mathcal{C}}$ is the matrix whose j-th column is $[Tv_j]_{\mathcal{C}}$, and it turns coordinates into coordinates: $[Tv]_{\mathcal{C}} = [T]_{\mathcal{B}}^{\mathcal{C}}[v]_{\mathcal{B}}$ (Theorem 4.1.1).

Recall the reflection from the previous section (Example (A Reflection in Two Bases)), which motivates this one. Let $R\colon \mathbb{R}^2 \to \mathbb{R}^2$ be the reflection across the line $y = x$, so $R(x, y) = (y, x)$. In the standard basis $\mathcal{E} = (\mathbf{e}_1, \mathbf{e}_2)$, we have $R\mathbf{e}_1 = \mathbf{e}_2$ and $R\mathbf{e}_2 = \mathbf{e}_1$, so

$$[R]_{\mathcal{E}}^{\mathcal{E}} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

Now use the basis $\mathcal{B} = ((1, 1), (1, -1))$, one vector on the mirror line and one perpendicular to it. Then $R(1, 1) = (1, 1)$ and $R(1, -1) = (-1, 1) = -(1, -1)$, so

$$[R]_{\mathcal{B}}^{\mathcal{B}} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

The second matrix shows at a glance what R does: it keeps one direction and flips the other. The two matrices describe the same map, so there must be a rule converting one into the other. The first step is to convert the **coordinates** of a single vector.

Take $\mathbf{v} = (3, 1)$. Its standard coordinates are $[\mathbf{v}]_{\mathcal{E}} = (3, 1)$. In the basis $\mathcal{B}$, we solve $a(1, 1) + b(1, -1) = (3, 1)$, which gives $a = 2$, $b = 1$, so $[\mathbf{v}]_{\mathcal{B}} = (2, 1)$. The same vector has two different coordinate columns, and we want a matrix that translates between them. Since the identity map id_V sends every vector to itself, the matrix of id_V, with one basis on the input side and the other on the output side, should do exactly this job.

A change-of-coordinates matrix is the matrix of "do nothing", read with one basis on the way in and another on the way out.

Definition 4.2.1. Change-of-coordinates matrix

Let V be a finite-dimensional vector space over F, and let $\mathcal{B} = (v_1, \dots, v_n)$ and $\mathcal{C}$ be **ordered** bases of V. The **change-of-coordinates matrix from $\mathcal{B}$ to $\mathcal{C}$** is the $n \times n$ matrix

$$[\mathrm{id}]_{\mathcal{B}}^{\mathcal{C}} \coloneqq [\mathrm{id}_V]_{\mathcal{B}}^{\mathcal{C}},$$

the matrix of the identity map $\mathrm{id}_V\colon V \to V$ with respect to the input basis $\mathcal{B}$ and the output basis $\mathcal{C}$. Its j-th column is $[v_j]_{\mathcal{C}}$.

In words: to build $[\mathrm{id}]_{\mathcal{B}}^{\mathcal{C}}$, take the vectors of the **old** basis $\mathcal{B}$ one at a time, write each in the **new** basis $\mathcal{C}$, and put the coordinate columns side by side. The column description is Definition 4.1.1 with $T = \mathrm{id}_V$, since $\mathrm{id}_V(v_j) = v_j$. Both bases must be bases of the **same** space V, and both are ordered: reordering either one permutes the columns or the rows.

Examples.

- **Into the standard basis of F^n.** If $\mathcal{B} = (\mathbf{b}_1, \dots, \mathbf{b}_n)$ is a basis of F^n and $\mathcal{E}$ is the standard basis, then $[\mathbf{b}_j]_{\mathcal{E}} = \mathbf{b}_j$. So $[\mathrm{id}]_{\mathcal{B}}^{\mathcal{E}}$ is the matrix whose columns are the vectors of $\mathcal{B}$ themselves. For $\mathcal{B} = ((1, 1), (1, -1))$ in $\mathbb{R}^2$, $[\mathrm{id}]_{\mathcal{B}}^{\mathcal{E}} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$, and indeed $\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}\begin{pmatrix} 2 \\ 1 \end{pmatrix} = \begin{pmatrix} 3 \\ 1 \end{pmatrix}$ turns $[\mathbf{v}]_{\mathcal{B}}$ into $[\mathbf{v}]_{\mathcal{E}}$ for $\mathbf{v} = (3, 1)$.

- **Out of the standard basis.** In the other direction, $[\mathrm{id}]_{\mathcal{E}}^{\mathcal{B}}$ has columns $[\mathbf{e}_1]_{\mathcal{B}}$ and $[\mathbf{e}_2]_{\mathcal{B}}$. Solving $a(1,1)+b(1,-1)=(1,0)$ gives $a=b=\frac{1}{2}$, and solving for $(0,1)$ gives $a=\frac{1}{2}$, $b=-\frac{1}{2}$. So $[\mathrm{id}]_{\mathcal{E}}^{\mathcal{B}}=\frac{1}{2}\begin{pmatrix}1 & 1\\ 1 & -1\end{pmatrix}$. This is the inverse of the previous matrix, which the next theorem explains.
- **Polynomials.** In $F[x]_{\leq 2}$, let $\mathcal{B}=(1,x,x^2)$ and $\mathcal{C}=(1,x-1,(x-1)^2)$. Expanding $(x-1)^2=1-2x+x^2$, the vectors of $\mathcal{C}$ have $\mathcal{B}$-coordinates $(1,0,0)$, $(-1,1,0)$, and $(1,-2,1)$, so
$$[\mathrm{id}]_{\mathcal{C}}^{\mathcal{B}}=\begin{pmatrix}1 & -1 & 1\\ 0 & 1 & -2\\ 0 & 0 & 1\end{pmatrix}.$$
- **Degenerate case.** If $\mathcal{C}=\mathcal{B}$, then $[v_j]_{\mathcal{B}}=\mathbf{e}_j$, so $[\mathrm{id}]_{\mathcal{B}}^{\mathcal{B}}=\mathbf{I}_n$. Changing nothing is recorded by the identity matrix, as it should be.

Nonexample by minimal change. Replace the basis $((1,1),(1,-1))$ by the list $\mathcal{L}=((1,1),(2,2))$. We can still write the matrix with these vectors as columns, but $\mathcal{L}$ is not a basis, so "coordinates with respect to $\mathcal{L}$" are not defined: $(2,2)=2(1,1)+0(2,2)=0(1,1)+1(2,2)$ has two coordinate columns, and $(1,0)$ has none. The clause that fails is "$\mathcal{B}$ is a basis", which Definition 1.5.2 needs for coordinates to exist and be unique.

Warning. **Watch the direction.** The columns of $[\mathrm{id}]_{\mathcal{B}}^{\mathcal{C}}$ are the **$\mathcal{B}$-vectors written in $\mathcal{C}$**, and the matrix turns $\mathcal{B}$-coordinates into $\mathcal{C}$-coordinates. Using it the other way gives nonsense. With $\mathcal{B}=((1,1),(1,-1))$, the matrix $\begin{pmatrix}1 & 1\\ 1 & -1\end{pmatrix}=[\mathrm{id}]_{\mathcal{B}}^{\mathcal{E}}$ applied to $[\mathbf{v}]_{\mathcal{E}}=(3,1)$ gives $(4,2)$, which is **not** $[\mathbf{v}]_{\mathcal{B}}=(2,1)$. A reliable test: the first column must be the first **input** basis vector written in the **output** basis. Here $(1,1)=1\cdot\mathbf{e}_1+1\cdot\mathbf{e}_2$, so the column $(1,1)$ belongs to $[\mathrm{id}]_{\mathcal{B}}^{\mathcal{E}}$, which accepts $\mathcal{B}$-coordinates.

The examples suggest three facts: the matrix converts coordinates, converting twice multiplies the matrices, and converting back inverts the matrix. All three follow from results about matrices of maps, applied to the identity.

Theorem 4.2.1. Change of Coordinates

Let V be a vector space of dimension n, and let $\mathcal{B}$, $\mathcal{C}$, $\mathcal{D}$ be bases of V.

(a) For every $v\in V$, $[v]_{\mathcal{C}}=[\mathrm{id}]_{\mathcal{B}}^{\mathcal{C}}[v]_{\mathcal{B}}$. Moreover, $[\mathrm{id}]_{\mathcal{B}}^{\mathcal{C}}$ is the **only** matrix $M\in M_n(F)$ with $[v]_{\mathcal{C}}=M[v]_{\mathcal{B}}$ for every $v\in V$.

(b) $[\mathrm{id}]_{\mathcal{B}}^{\mathcal{D}}=[\mathrm{id}]_{\mathcal{C}}^{\mathcal{D}}[\mathrm{id}]_{\mathcal{B}}^{\mathcal{C}}$.

(c) $[\mathrm{id}]_{\mathcal{B}}^{\mathcal{C}}$ is invertible, and $\left([\mathrm{id}]_{\mathcal{B}}^{\mathcal{C}}\right)^{-1}=[\mathrm{id}]_{\mathcal{C}}^{\mathcal{B}}$.

Proof.

(a) Apply Theorem 4.1.1, including its uniqueness clause, to id_V with input basis $\mathcal{B}$ and output basis $\mathcal{C}$: since $\mathrm{id}_V(v)=v$, it says $[v]_{\mathcal{C}}=[\mathrm{id}]_{\mathcal{B}}^{\mathcal{C}}[v]_{\mathcal{B}}$ for every v, and that no other matrix has this property.

(b) Since $\mathrm{id}_V=\mathrm{id}_V\circ\mathrm{id}_V$, Theorem 4.1.2, with $\mathcal{B}$ on the input, $\mathcal{C}$ in the middle, and $\mathcal{D}$ on the output, gives $[\mathrm{id}]_{\mathcal{B}}^{\mathcal{D}}=[\mathrm{id}]_{\mathcal{C}}^{\mathcal{D}}[\mathrm{id}]_{\mathcal{B}}^{\mathcal{C}}$.

(c) Taking $\mathcal{D} = \mathcal{B}$ in (b) gives $[\mathrm{id}]_{\mathcal{C}}^{\mathcal{B}}[\mathrm{id}]_{\mathcal{B}}^{\mathcal{C}} = [\mathrm{id}]_{\mathcal{B}}^{\mathcal{B}} = \mathbf{I}_n$, where the last equality holds because, writing $\mathcal{B} = (v_1, \dots, v_n)$, the j-th column of $[\mathrm{id}]_{\mathcal{B}}^{\mathcal{B}}$ is $[v_j]_{\mathcal{B}} = \mathbf{e}_j$. Swapping the roles of $\mathcal{B}$ and $\mathcal{C}$ gives $[\mathrm{id}]_{\mathcal{B}}^{\mathcal{C}}[\mathrm{id}]_{\mathcal{C}}^{\mathcal{B}} = \mathbf{I}_n$. By Definition 0.9.13, $[\mathrm{id}]_{\mathcal{B}}^{\mathcal{C}}$ is invertible with inverse $[\mathrm{id}]_{\mathcal{C}}^{\mathcal{B}}$.

■

So every change-of-coordinates matrix is invertible. The converse is also true, and it will matter when we compare matrices: every invertible matrix is a change-of-coordinates matrix, for a suitable new basis.

Proposition 4.2.2. Invertible matrices are changes of basis

Let $\mathcal{B} = (v_1, \dots, v_n)$ be a basis of V, and let $\mathbf{P} = (p_{ij}) \in M_n(F)$ be invertible. Put

$$v_j' := p_{1j}v_1 + p_{2j}v_2 + \cdots + p_{nj}v_n \qquad (j = 1, \dots, n).$$

Then $\mathcal{B}' = (v_1', \dots, v_n')$ is a basis of V, and $[\mathrm{id}]_{\mathcal{B}'}^{\mathcal{B}} = \mathbf{P}$.

Proof. By Definition 1.5.2, $[v_j']_{\mathcal{B}}$ is the j-th column $\mathbf{p}_j$ of $\mathbf{P}$. Let $c_1, \dots, c_n \in F$ satisfy $c_1v_1' + \cdots + c_nv_n' = 0$. Taking $\mathcal{B}$-coordinates and using Theorem 1.5.2,

$$c_1\mathbf{p}_1 + \cdots + c_n\mathbf{p}_n = [0]_{\mathcal{B}} = \mathbf{0},$$

that is, $\mathbf{Pc} = \mathbf{0}$ with $\mathbf{c} = (c_1, \dots, c_n)$, by Theorem 0.9.3. Since $\mathbf{P}$ is invertible, Theorem 3.4.4 ((a) $\Rightarrow$ (b)) gives $\mathbf{c} = \mathbf{0}$. So $\mathcal{B}'$ is linearly independent. It has length $n = \dim V$, hence it is a basis by Theorem 1.6.4 (a). Finally, the j-th column of $[\mathrm{id}]_{\mathcal{B}'}^{\mathcal{B}}$ is $[v_j']_{\mathcal{B}} = \mathbf{p}_j$, so $[\mathrm{id}]_{\mathcal{B}'}^{\mathcal{B}} = \mathbf{P}$.

■

Example. Coordinates around the point 1

In $\mathbb{R}[x]_{\le 2}$, let $\mathcal{B} = (1, x, x^2)$ and $\mathcal{C} = (1, x-1, (x-1)^2)$. Find $[\mathrm{id}]_{\mathcal{B}}^{\mathcal{C}}$, and use it to write $p = 2 + 3x + x^2$ in powers of $x - 1$.

Solution. From the examples above, $[\mathrm{id}]_{\mathcal{C}}^{\mathcal{B}} = \begin{pmatrix} 1 & -1 & 1 \\ 0 & 1 & -2 \\ 0 & 0 & 1 \end{pmatrix}$. By Theorem 4.2.1 (c), $[\mathrm{id}]_{\mathcal{B}}^{\mathcal{C}}$ is its inverse. We can also read it off directly: $1 = 1$, $x = 1 + (x-1)$, and $x^2 = (1 + (x-1))^2 = 1 + 2(x-1) + (x-1)^2$. So

$$[\mathrm{id}]_{\mathcal{B}}^{\mathcal{C}} = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{pmatrix},$$

and multiplying the two matrices gives $\mathbf{I}_3$, as the theorem predicts. By Theorem 4.2.1 (a),

$$[p]_{\mathcal{C}} = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 2 \\ 3 \\ 1 \end{pmatrix} = \begin{pmatrix} 6 \\ 5 \\ 1 \end{pmatrix}.$$

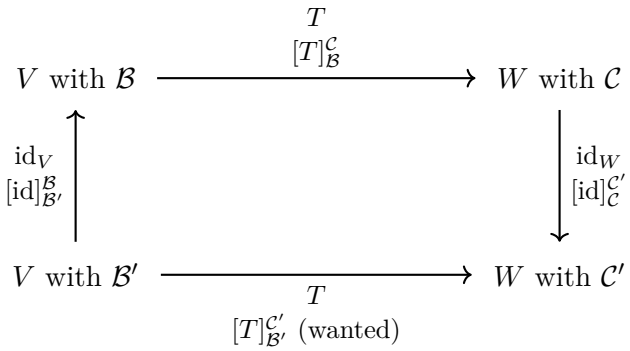
Hence $2 + 3x + x^2 = 6 + 5(x-1) + (x-1)^2$. Check at $x = 1$: both sides equal 6. The $\mathcal{C}$-coordinates are $p(1)$, $p'(1)$, and $p''(1)/2$, the Taylor coefficients at 1.

4.2.2 Change of basis for linear maps

Now return to a linear map $T\colon V \to W$ between finite-dimensional spaces. Suppose we know its matrix $[T]_{\mathcal{B}}^{\mathcal{C}}$ and want the matrix $[T]_{\mathcal{B}'}^{\mathcal{C}'}$ for new bases $\mathcal{B}'$ of V and $\mathcal{C}'$ of W. The trick is to write T in a slightly more elaborate way that does nothing:

$$T = \mathrm{id}_W \circ T \circ \mathrm{id}_V .$$

Each factor can now carry its own pair of bases, and the matrix of the composite is the product of the matrices. The picture keeps track of which basis sits where.



Going along the bottom edge applies T. Going up, across the top, and down applies $\mathrm{id}_W \circ T \circ \mathrm{id}_V$, which is the same map. So the two routes must have the same matrix. Reading the long route, the step done **first** (up the left side) is written on the **right**.

Theorem 4.2.3. Change of Basis for Linear Maps

Let V and W be finite-dimensional vector spaces over F, let $\mathcal{B}, \mathcal{B}'$ be bases of V and $\mathcal{C}, \mathcal{C}'$ bases of W, and let $T\colon V \to W$ be linear. Then

$$[T]_{\mathcal{B}'}^{\mathcal{C}'} = [\mathrm{id}]_{\mathcal{C}}^{\mathcal{C}'}\,[T]_{\mathcal{B}}^{\mathcal{C}}\,[\mathrm{id}]_{\mathcal{B}'}^{\mathcal{B}}.$$

In particular, for an operator $T \in \mathcal{L}(V)$ and bases $\mathcal{B}, \mathcal{B}'$ of V, with $\mathbf{P} = [\mathrm{id}]_{\mathcal{B}'}^{\mathcal{B}}$,

$$[T]_{\mathcal{B}'}^{\mathcal{B}'} = \mathbf{P}^{-1}\,[T]_{\mathcal{B}}^{\mathcal{B}}\,\mathbf{P}.$$

Idea. Nothing is computed: the formula is the change-of-basis square read along its long route. ① Write $T = \mathrm{id}_W \circ T \circ \mathrm{id}_V$. ② Give the middle factor the bases whose matrix we know. ③ Apply the composition rule twice. Do not memorize the formula; redraw the square and read it off, right to left.

Proof. Since $T = \mathrm{id}_W \circ (T \circ \mathrm{id}_V)$, Theorem 4.1.2 with bases $\mathcal{B}'$, $\mathcal{C}$, $\mathcal{C}'$ gives $[T]_{\mathcal{B}'}^{\mathcal{C}'} = [\mathrm{id}]_{\mathcal{C}}^{\mathcal{C}'}\,[T \circ \mathrm{id}_V]_{\mathcal{B}'}^{\mathcal{C}}$. Applying Theorem 4.1.2 again, with bases $\mathcal{B}'$, $\mathcal{B}$, $\mathcal{C}$, gives $[T \circ \mathrm{id}_V]_{\mathcal{B}'}^{\mathcal{C}} = [T]_{\mathcal{B}}^{\mathcal{C}}\,[\mathrm{id}]_{\mathcal{B}'}^{\mathcal{B}}$. Combining the two equations proves the first formula.

For an operator, take $W = V$, $\mathcal{C} = \mathcal{B}$, and $\mathcal{C}' = \mathcal{B}'$. The first formula becomes $[T]_{\mathcal{B}'}^{\mathcal{B}'} =$

$[\mathrm{id}]^{\mathcal{B}'}_{\mathcal{B}}\,[T]^{\mathcal{B}}_{\mathcal{B}}\,\mathbf{P}$, and $[\mathrm{id}]^{\mathcal{B}'}_{\mathcal{B}} = \mathbf{P}^{-1}$ by Theorem 4.2.1 (c). This proves the theorem.

■

We will call this move **the change-of-basis square**: whenever a question involves the same map in two bases, draw the square, label the arrows with bases, and read the formula off the long route. It also explains the word "coordinates": a matrix is a map seen through a choice of coordinates, and the side arrows re-express those coordinates.

Example. The reflection in a mirror-adapted basis

Let $R(x, y) = (y, x)$ on $\mathbb{R}^2$, and $\mathcal{B} = ((1, 1), (1, -1))$. Compute $[R]^{\mathcal{B}}_{\mathcal{B}}$ from $[R]^{\mathcal{E}}_{\mathcal{E}}$ using Theorem 4.2.3, and compare with the direct computation at the start of the section.

Solution. Here $\mathbf{P} = [\mathrm{id}]^{\mathcal{E}}_{\mathcal{B}} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$, with inverse $\mathbf{P}^{-1} = [\mathrm{id}]^{\mathcal{B}}_{\mathcal{E}} = \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$, found in the examples after Definition 4.2.1. Then

$$\begin{aligned} \mathbf{P}^{-1}[R]^{\mathcal{E}}_{\mathcal{E}}\mathbf{P} &= \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \\ &= \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} = \frac{1}{2}\begin{pmatrix} 2 & 0 \\ 0 & -2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \end{aligned}$$

which agrees with the direct computation. The direct route was shorter here, because we chose $\mathcal{B}$ so that R acts simply on it. The formula earns its keep when the new matrix is not visible by inspection, or when we want to go back from the simple matrix to the standard one: $[R]^{\mathcal{E}}_{\mathcal{E}} = \mathbf{P}\,[R]^{\mathcal{B}}_{\mathcal{B}}\,\mathbf{P}^{-1}$.

The next example changes basis in a polynomial space, where "standard" does not mean "best".

Example. Differentiation with a rescaled basis

Let $D\colon \mathbb{R}[x]_{\le 2} \to \mathbb{R}[x]_{\le 2}$ be differentiation, $\mathcal{B} = (1, x, x^2)$, and $\mathcal{B}' = (1, x, \frac{1}{2}x^2)$. Find $[D]^{\mathcal{B}'}_{\mathcal{B}'}$.

Solution. In $\mathcal{B}$, $D(1) = 0$, $D(x) = 1$, and $D(x^2) = 2x$, so $\mathbf{A} = [D]^{\mathcal{B}}_{\mathcal{B}} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \end{pmatrix}$. The vectors of $\mathcal{B}'$ have $\mathcal{B}$-coordinates $\mathbf{e}_1$, $\mathbf{e}_2$, $\frac{1}{2}\mathbf{e}_3$, so $\mathbf{P} = [\mathrm{id}]^{\mathcal{B}}_{\mathcal{B}'} = \mathrm{diag}(1, 1, \frac{1}{2})$ and $\mathbf{P}^{-1} = \mathrm{diag}(1, 1, 2)$. Multiplying a matrix on the left by a diagonal matrix scales its rows, and on the right scales its columns, so

$$\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \begin{pmatrix} 0 & 1\cdot 1\cdot 1 & 0 \\ 0 & 0 & 1\cdot 2\cdot \frac{1}{2} \\ 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}.$$

Directly: $D(1) = 0$, $D(x) = 1$, $D(\frac{1}{2}x^2) = x$, so D shifts each vector of $\mathcal{B}'$ to the previous one, and the matrix has 1's just above the diagonal. The factor 2 has disappeared into the basis.

Quick check. Let $T\colon V \to W$, and suppose we change only the basis of W, from $\mathcal{C}$ to $\mathcal{C}'$, keeping $\mathcal{B}$. Which of the two side matrices in Theorem 4.2.3 becomes the identity, and what is the formula?

Solution. With $\mathcal{B}' = \mathcal{B}$, the left side of the square is $[\mathrm{id}]_{\mathcal{B}}^{\mathcal{B}} = \mathbf{I}_n$. The formula becomes $[T]_{\mathcal{B}}^{\mathcal{C}'} = [\mathrm{id}]_{\mathcal{C}}^{\mathcal{C}'}\,[T]_{\mathcal{B}}^{\mathcal{C}}$: only a multiplication on the left. Changing the output basis acts on the rows; changing the input basis acts on the columns.

4.2.3 Similar matrices

For an operator, the two bases on the two sides of the square are the same, and the formula takes the special shape $\mathbf{P}^{-1}\mathbf{A}\mathbf{P}$. Recall from Example (Similarity) that we already gave this shape a name in Chapter 0 and checked that it defines an equivalence relation. We now record it as a definition, because it is exactly the relation "same operator, different basis".

Two square matrices are similar when one is obtained from the other by a change of basis applied on both sides at once.

Definition 4.2.2. Similar matrices

Let $\mathbf{A}, \mathbf{B} \in M_n(F)$. We say $\mathbf{A}$ is **similar** to $\mathbf{B}$, written $\mathbf{A} \sim \mathbf{B}$, if there **exists** an **invertible** matrix $\mathbf{P} \in M_n(F)$ such that

$$\mathbf{B} = \mathbf{P}^{-1}\mathbf{A}\mathbf{P}.$$

In words: the **same** $\mathbf{P}$ appears on both sides, once inverted. By Example (Similarity), $\sim$ is an equivalence relation, so we may say "$\mathbf{A}$ and $\mathbf{B}$ are similar" without worrying about the order. The matrix $\mathbf{P}$ is not unique: for instance $\mathbf{A} = \mathbf{P}^{-1}\mathbf{A}\mathbf{P}$ holds for every invertible $\mathbf{P}$ that commutes with $\mathbf{A}$, such as any nonzero multiple of $\mathbf{I}_n$. Writing $\mathbf{B} = \mathbf{Q}\mathbf{A}\mathbf{Q}^{-1}$ defines the same relation, with $\mathbf{Q} = \mathbf{P}^{-1}$.

Examples.

- **The reflection.** $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \sim \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$, by Example (The reflection in a mirror-adapted basis).
- **A triangular matrix.** $\begin{pmatrix} 1 & 2 \\ 0 & 3 \end{pmatrix} \sim \mathrm{diag}(1, 3)$, as computed in Example (Similarity) with $\mathbf{P} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$.
- **Reordering a basis.** $\mathrm{diag}(1, 2) \sim \mathrm{diag}(2, 1)$: with $\mathbf{P} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = \mathbf{P}^{-1}$, we get $\mathbf{P}^{-1}\,\mathrm{diag}(1, 2)\mathbf{P} = \mathrm{diag}(2, 1)$. This is the operator with matrix $\mathrm{diag}(1, 2)$ in (v_1, v_2), read in the reordered basis (v_2, v_1).
- **Degenerate case: scalar matrices.** For $c \in F$, $\mathbf{P}^{-1}(c\mathbf{I}_n)\mathbf{P} = c\mathbf{P}^{-1}\mathbf{P} = c\mathbf{I}_n$ for every

invertible $\mathbf{P}$. So $c\mathbf{I}_n$ is similar only to itself. In particular $\mathbf{I}_n$ and the zero matrix are alone in their classes: the identity operator looks like $\mathbf{I}_n$ in **every** basis.

Nonexample by minimal change. Allow two different invertible matrices, $\mathbf{B} = \mathbf{QAP}$. With $\mathbf{A} = \mathbf{I}_2$, $\mathbf{Q} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$, and $\mathbf{P} = \mathbf{I}_2$, we get $\mathbf{B} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \neq \mathbf{I}_2$. So $\mathbf{B} = \mathbf{QAP}$ holds, but $\mathbf{B}$ is not similar to $\mathbf{I}_2$, which is similar only to itself. The clause that failed is "the **same** $\mathbf{P}$ on both sides, once inverted". The relaxed relation is also useful; it is the subject of the next section.

The definition was designed to capture one idea, and the next theorem says it captures exactly that idea.

Theorem 4.2.4. Similar Matrices Represent the Same Operator

Let $\mathbf{A}, \mathbf{B} \in M_n(F)$.

(a) If V is a vector space of dimension n, $T \in \mathcal{L}(V)$, and $\mathcal{B}$,$\mathcal{C}$ are bases of V, then $[T]_{\mathcal{B}}^{\mathcal{B}} \sim [T]_{\mathcal{C}}^{\mathcal{C}}$.

(b) Conversely, suppose $\mathbf{A} \sim \mathbf{B}$. Let V be any vector space of dimension n with a basis $\mathcal{B}$, and let $T \in \mathcal{L}(V)$ be an operator with $[T]_{\mathcal{B}}^{\mathcal{B}} = \mathbf{A}$. Then there is a basis $\mathcal{C}$ of V with $[T]_{\mathcal{C}}^{\mathcal{C}} = \mathbf{B}$.

(c) Such an operator T in (b) always exists. In particular, $\mathbf{A} \sim \mathbf{B}$ if and only if $\mathbf{A}$ and $\mathbf{B}$ are the matrices of the operator $\mathbf{x} \mapsto \mathbf{Ax}$ on F^n in the standard basis and in some basis of F^n, respectively.

Proof.

(a) By Theorem 4.2.3, $[T]_{\mathcal{C}}^{\mathcal{C}} = \mathbf{P}^{-1}[T]_{\mathcal{B}}^{\mathcal{B}}\mathbf{P}$ with $\mathbf{P} = [\mathrm{id}]_{\mathcal{C}}^{\mathcal{B}}$, which is invertible by Theorem 4.2.1 (c).

(b) Suppose $\mathbf{B} = \mathbf{P}^{-1}\mathbf{AP}$ with $\mathbf{P}$ invertible. By Proposition 4.2.2, there is a basis $\mathcal{C}$ of V with $[\mathrm{id}]_{\mathcal{C}}^{\mathcal{B}} = \mathbf{P}$. By Theorem 4.2.3, $[T]_{\mathcal{C}}^{\mathcal{C}} = \mathbf{P}^{-1}[T]_{\mathcal{B}}^{\mathcal{B}}\mathbf{P} = \mathbf{P}^{-1}\mathbf{AP} = \mathbf{B}$.

(c) Write $\mathcal{B} = (v_1, \ldots, v_n)$. By Theorem 2.1.5, there is a linear map $T\colon V \to V$ with $Tv_j = a_{1j}v_1 + \cdots + a_{nj}v_n$ for each j. The j-th column of $[T]_{\mathcal{B}}^{\mathcal{B}}$ is $[Tv_j]_{\mathcal{B}}$, which is the j-th column of $\mathbf{A}$; so $[T]_{\mathcal{B}}^{\mathcal{B}} = \mathbf{A}$. For $V = F^n$ with the standard basis $\mathcal{E}$, the operator $\mathbf{x} \mapsto \mathbf{Ax}$ works, since $\mathbf{Ae}_j$ is the j-th column of $\mathbf{A}$ (Theorem 0.9.3). The last statement now follows from (a) and (b).

■

This is the promise of Chapter 0 kept: similar matrices describe the same operator in two coordinate systems. So a question of the form "are $\mathbf{A}$ and $\mathbf{B}$ similar?" is really the question "is there a basis in which the operator $\mathbf{x} \mapsto \mathbf{Ax}$ has matrix $\mathbf{B}$?". Deciding it is one of the central problems of the book, and it will take until Chapter 10 to answer it in general.

4.2.4 Similarity invariants

If a quantity computed from a matrix does not change under $\mathbf{A} \mapsto \mathbf{P}^{-1}\mathbf{AP}$, then it is really a property of the operator, and we call it a **similarity invariant**. Invariants are the cheap way to prove that two matrices are **not** similar: find one invariant on which they differ. The trace is the first example, and it is the reason the trace will make sense for operators in this chapter's

last section.

Theorem 4.2.5. Trace Is a Similarity Invariant

If $\mathbf{A}, \mathbf{B} \in M_n(F)$ are similar, then $\operatorname{tr}\mathbf{A} = \operatorname{tr}\mathbf{B}$.

Proof. Let $\mathbf{B} = \mathbf{P}^{-1}\mathbf{A}\mathbf{P}$ with $\mathbf{P}$ invertible. By Theorem 0.9.6 (3), applied to the matrices $\mathbf{P}^{-1}$ and $\mathbf{AP}$,

$$\operatorname{tr}\mathbf{B} = \operatorname{tr}\left(\mathbf{P}^{-1}(\mathbf{AP})\right) = \operatorname{tr}\left((\mathbf{AP})\mathbf{P}^{-1}\right) = \operatorname{tr}\left(\mathbf{A}(\mathbf{P}\mathbf{P}^{-1})\right) = \operatorname{tr}\mathbf{A},$$

where the third equality is associativity of matrix multiplication. This proves the theorem. ∎

The proof never used that $\mathbf{B}$ is "similar" in any geometric sense, only that the trace ignores the order of two factors. Several other quantities survive a change of basis, for simpler reasons.

Proposition 4.2.6. More similarity invariants

Let $\mathbf{A}, \mathbf{B} \in M_n(F)$ with $\mathbf{B} = \mathbf{P}^{-1}\mathbf{A}\mathbf{P}$ for an invertible $\mathbf{P}$.

(a) $\operatorname{rank}\mathbf{A} = \operatorname{rank}\mathbf{B}$.

(b) $\mathbf{A}$ is invertible if and only if $\mathbf{B}$ is invertible.

(c) For every polynomial $p \in F[x]$, $p(\mathbf{B}) = \mathbf{P}^{-1}p(\mathbf{A})\mathbf{P}$. In particular, $p(\mathbf{A}) = \mathbf{O}$ if and only if $p(\mathbf{B}) = \mathbf{O}$; for example, $\mathbf{A}^2 = \mathbf{O}$ if and only if $\mathbf{B}^2 = \mathbf{O}$.

Proof.

(a) By Theorem 3.5.9, multiplying by invertible matrices on either side does not change the rank, so $\operatorname{rank}(\mathbf{P}^{-1}\mathbf{AP}) = \operatorname{rank}(\mathbf{AP}) = \operatorname{rank}\mathbf{A}$.

(b) If $\mathbf{A}$ is invertible, then $\mathbf{B} = \mathbf{P}^{-1}\mathbf{AP}$ is a product of invertible matrices, hence invertible by Theorem 0.9.7 (part 3). Since $\mathbf{A} = \mathbf{PBP}^{-1} = (\mathbf{P}^{-1})^{-1}\mathbf{BP}^{-1}$, the same argument with the roles swapped gives the converse.

(c) First, $\mathbf{B}^k = \mathbf{P}^{-1}\mathbf{A}^k\mathbf{P}$ for every $k \geq 0$, by induction on k: for $k = 0$ both sides are $\mathbf{I}_n$, and if it holds for k, then $\mathbf{B}^{k+1} = \mathbf{B}^k\mathbf{B} = (\mathbf{P}^{-1}\mathbf{A}^k\mathbf{P})(\mathbf{P}^{-1}\mathbf{AP}) = \mathbf{P}^{-1}\mathbf{A}^k(\mathbf{P}\mathbf{P}^{-1})\mathbf{AP} = \mathbf{P}^{-1}\mathbf{A}^{k+1}\mathbf{P}$. Now let $p = c_0 + c_1x + \cdots + c_dx^d$. By Definition 0.9.14 and the distributive and scalar laws of Theorem 0.9.4,

$$p(\mathbf{B}) = \sum_{k=0}^{d} c_k\mathbf{B}^k = \sum_{k=0}^{d} c_k\mathbf{P}^{-1}\mathbf{A}^k\mathbf{P} = \mathbf{P}^{-1}\Big(\sum_{k=0}^{d} c_k\mathbf{A}^k\Big)\mathbf{P} = \mathbf{P}^{-1}p(\mathbf{A})\mathbf{P}.$$

If $p(\mathbf{A}) = \mathbf{O}$, then $p(\mathbf{B}) = \mathbf{P}^{-1}\mathbf{OP} = \mathbf{O}$. Conversely, $\mathbf{A} = \mathbf{PBP}^{-1}$ is of the same form with $\mathbf{P}^{-1}$ in place of $\mathbf{P}$, so $p(\mathbf{B}) = \mathbf{O}$ implies $p(\mathbf{A}) = \mathbf{O}$. ∎

For example, $\begin{pmatrix} 1 & 2 \\ 0 & 3 \end{pmatrix}$ and $\begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$ are **not** similar, because their traces are 4 and 3. And $\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$

is not similar to the zero matrix, because their ranks are 1 and 0 (or because the zero matrix is similar only to itself).

It is tempting to turn this around and decide similarity by comparing a list of invariants. That fails.

Warning. **Shared invariants do not imply similarity.** Let $\mathbf{J} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$. Then $\mathbf{I}_2$ and $\mathbf{J}$ have the same trace 2, the same rank 2, both are invertible, and both satisfy $(\mathbf{X} - \mathbf{I}_2)^2 = \mathbf{O}$. But they are **not** similar: for every invertible $\mathbf{P}$, $\mathbf{P}^{-1}\mathbf{I}_2\mathbf{P} = \mathbf{I}_2 \neq \mathbf{J}$. A sharper invariant does separate them: combining (a) and (c) of Proposition 4.2.6, similar matrices $\mathbf{A}$ and $\mathbf{B}$ have $\operatorname{rank} p(\mathbf{A}) = \operatorname{rank} p(\mathbf{B})$ for every polynomial p, and $\operatorname{rank}(\mathbf{I}_2 - \mathbf{I}_2) = 0$ while $\operatorname{rank}(\mathbf{J} - \mathbf{I}_2) = 1$. Which invariants suffice in general is the question Chapter 10 answers.

Quick check. Are $\operatorname{diag}(2, 5)$ and $\operatorname{diag}(5, 2)$ similar? Are $\begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$ and $\begin{pmatrix} 2 & 1 \\ 0 & 2 \end{pmatrix}$ similar?

Solution. The first pair is similar: with $\mathbf{P} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$, its own inverse, $\mathbf{P}^{-1} \operatorname{diag}(2, 5)\mathbf{P} = \operatorname{diag}(5, 2)$. The second pair is not: $2\mathbf{I}_2$ is a scalar matrix, so it is similar only to itself, although both matrices have trace 4 and rank 2.

4.2.5 A basis adapted to the operator

The reflection became diagonal once we chose one basis vector the map keeps and one it flips. The same idea works whenever a map is simple on two complementary pieces of the space. For a map of rank 1 on $\mathbb{R}^2$, the natural pieces are the image and the kernel, provided they are two different lines. (They need not be: for $\mathbf{x} \mapsto \mathbf{N}\mathbf{x}$ with $\mathbf{N} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$, both are $\operatorname{span}(\mathbf{e}_1)$, and Exercise (C3: Every nonzero 2×2 matrix with square zero) shows what to do instead.)

Example. Making a matrix diagonal with a well-chosen basis

Let $T\colon \mathbb{R}^2 \to \mathbb{R}^2$, $T(x, y) = (2x - y,\ 2x - y)$. Find a basis $\mathcal{B}$ of $\mathbb{R}^2$ such that $[T]_{\mathcal{B}}^{\mathcal{B}}$ is diagonal, and check the answer with Theorem 4.2.3.

Solution. In the standard basis $\mathbf{A} = [T]_{\mathcal{E}}^{\mathcal{E}} = \begin{pmatrix} 2 & -1 \\ 2 & -1 \end{pmatrix}$.

Both entries of $T(x, y)$ are equal, so $\operatorname{im} T \subseteq \operatorname{span}((1, 1))$, and $T(1, 1) = (1, 1)$ shows $\operatorname{im} T = \operatorname{span}((1, 1))$. The kernel is $\{(x, y) : y = 2x\} = \operatorname{span}((1, 2))$. So T fixes $(1, 1)$ and kills $(1, 2)$. These two vectors are not multiples of each other, so $\mathcal{B} = ((1, 1), (1, 2))$ is independent, and

it is a basis of $\mathbb{R}^2$ by Theorem 1.6.4. Since $T(1,1) = 1\cdot(1,1) + 0\cdot(1,2)$ and $T(1,2) = \mathbf{0}$,

$$[T]_{\mathcal{B}}^{\mathcal{B}} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}.$$

Check: $\mathbf{P} = [\mathrm{id}]_{\mathcal{B}}^{\mathcal{E}} = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}$ has $ad - bc = 1$, so $\mathbf{P}^{-1} = \begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}$ by Theorem 0.9.8. Then

$$\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}\begin{pmatrix} 2 & -1 \\ 2 & -1 \end{pmatrix}\begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix} = \begin{pmatrix} 2 & -1 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}.$$

As Theorem 4.2.5 and Proposition 4.2.6 require, $\operatorname{tr}\mathbf{A} = 1$ and $\operatorname{rank}\mathbf{A} = 1$ match the diagonal matrix. Its form also explains a property of $\mathbf{A}$ that is not obvious from its entries: $\operatorname{diag}(1,0)^2 = \operatorname{diag}(1,0)$, so $\mathbf{A}^2 = \mathbf{A}$ by Proposition 4.2.6 (c) applied to $p = x^2 - x$.

The basis in this example consisted of vectors that T sends to multiples of themselves. Chapter 9 turns this into a method: such vectors are called eigenvectors, and an operator has a diagonal matrix exactly when there is a basis of them. Operators with $T^2 = T$, like this one, are the projections of the last section of this chapter.

4.2.6 Exercises

4.2.6.1 A. Check your understanding

Exercise. A1

(a) Define the change-of-coordinates matrix $[\mathrm{id}]_{\mathcal{B}}^{\mathcal{C}}$, and say what its columns are.

(b) State the change-of-basis formula for an operator $T \in \mathcal{L}(V)$ and bases $\mathcal{B}, \mathcal{B}'$, saying exactly which matrix $\mathbf{P}$ is.

(c) True or false: $[\mathrm{id}]_{\mathcal{B}}^{\mathcal{C}} = [\mathrm{id}]_{\mathcal{C}}^{\mathcal{B}}$ for all bases $\mathcal{B}, \mathcal{C}$ of V. Justify your answer.

(d) True or false: if $\mathbf{A} \sim \mathbf{B}$, then $\operatorname{tr}\mathbf{A} = \operatorname{tr}\mathbf{B}$ and $\operatorname{rank}\mathbf{A} = \operatorname{rank}\mathbf{B}$. Justify your answer.

(e) True or false: if $\operatorname{tr}\mathbf{A} = \operatorname{tr}\mathbf{B}$ and $\operatorname{rank}\mathbf{A} = \operatorname{rank}\mathbf{B}$, then $\mathbf{A} \sim \mathbf{B}$. Justify your answer.

(f) Describe the method for recovering the change-of-basis formula without memorizing it.

Solution.

(a) For bases $\mathcal{B} = (v_1, \ldots, v_n)$ and $\mathcal{C}$ of V, $[\mathrm{id}]_{\mathcal{B}}^{\mathcal{C}}$ is the matrix of id_V with input basis $\mathcal{B}$ and output basis $\mathcal{C}$ (Definition 4.2.1). Its j-th column is $[v_j]_{\mathcal{C}}$.

(b) $[T]_{\mathcal{B}'}^{\mathcal{B}'} = \mathbf{P}^{-1}[T]_{\mathcal{B}}^{\mathcal{B}}\mathbf{P}$ with $\mathbf{P} = [\mathrm{id}]_{\mathcal{B}'}^{\mathcal{B}}$, whose columns are the $\mathcal{B}$-coordinates of the vectors of $\mathcal{B}'$ (Theorem 4.2.3).

(c) False. By Theorem 4.2.1 (c) the two matrices are inverse to each other, and an invertible matrix usually differs from its inverse: for $\mathcal{B} = ((1,1),(1,-1))$ and the standard basis $\mathcal{E}$ of $\mathbb{R}^2$, $[\mathrm{id}]_{\mathcal{B}}^{\mathcal{E}} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$ but $[\mathrm{id}]_{\mathcal{E}}^{\mathcal{B}} = \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$.

(d) True, by Theorem 4.2.5 and Proposition 4.2.6 (a).

(e) False. $\mathbf{I}_2$ and $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ both have trace 2 and rank 2, but $\mathbf{I}_2$ is similar only to itself.

(f) Draw the change-of-basis square: T along the top and bottom with the old and new bases, identity maps up the left side and down the right side. Since $T = \mathrm{id}_W \circ T \circ \mathrm{id}_V$, the matrix of the bottom arrow is the product of the matrices along the long route, with the first step on the right.

4.2.6.2 B. Practice

Exercise. B1: Change-of-coordinates matrices

(a) In $\mathbb{R}^2$, let $\mathcal{B} = ((1,2),(3,5))$ and $\mathcal{C} = ((1,1),(1,2))$. Find $[\mathrm{id}]_{\mathcal{B}}^{\mathcal{C}}$. If $[\mathbf{v}]_{\mathcal{B}} = (1,-1)$, find $[\mathbf{v}]_{\mathcal{C}}$, and check your answer by computing $\mathbf{v}$ both ways.

(b) In $\mathbb{R}[x]_{\le 2}$, let $\mathcal{B} = (1, x, x^2)$ and $\mathcal{C} = (1, 1+x, 1+x+x^2)$. Find $[\mathrm{id}]_{\mathcal{C}}^{\mathcal{B}}$ and $[\mathrm{id}]_{\mathcal{B}}^{\mathcal{C}}$. Hence find the $\mathcal{C}$-coordinates of $3 - x + 2x^2$.

Solution.

(a) The columns are the $\mathcal{C}$-coordinates of $(1,2)$ and $(3,5)$. Solving $a(1,1)+b(1,2)=(1,2)$ gives $a+b=1$, $a+2b=2$, so $b=1$, $a=0$. Solving $a(1,1)+b(1,2)=(3,5)$ gives $b=2$, $a=1$. Hence

$$[\mathrm{id}]_{\mathcal{B}}^{\mathcal{C}} = \begin{pmatrix} 0 & 1 \\ 1 & 2 \end{pmatrix}, \qquad [\mathbf{v}]_{\mathcal{C}} = \begin{pmatrix} 0 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} 1 \\ -1 \end{pmatrix} = \begin{pmatrix} -1 \\ -1 \end{pmatrix}$$

by Theorem 4.2.1 (a). Check: from $\mathcal{B}$, $\mathbf{v} = (1,2) - (3,5) = (-2,-3)$; from $\mathcal{C}$, $\mathbf{v} = -(1,1) - (1,2) = (-2,-3)$.

(b) The vectors of $\mathcal{C}$ have $\mathcal{B}$-coordinates $(1,0,0)$, $(1,1,0)$, $(1,1,1)$, so $[\mathrm{id}]_{\mathcal{C}}^{\mathcal{B}} = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}$.

For the other direction, write the vectors of $\mathcal{B}$ in $\mathcal{C}$: $1 = 1$, $x = (1+x) - 1$, $x^2 = (1+x+x^2) - (1+x)$. So

$$[\mathrm{id}]_{\mathcal{B}}^{\mathcal{C}} = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \\ 0 & 0 & 1 \end{pmatrix},$$

and multiplying the two matrices gives $\mathbf{I}_3$, as Theorem 4.2.1 (c) requires. Hence

$$[3 - x + 2x^2]_{\mathcal{C}} = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 3 \\ -1 \\ 2 \end{pmatrix} = \begin{pmatrix} 4 \\ -3 \\ 2 \end{pmatrix}.$$

Check: $4 - 3(1+x) + 2(1+x+x^2) = 3 - x + 2x^2$.

Exercise. B2: A matrix in a new basis, two ways

Let $T\colon \mathbb{R}^2 \to \mathbb{R}^2$, $T(x,y) = (4x - 2y,\ x + y)$, and $\mathcal{B} = ((1,1),(2,1))$.

(a) Compute $[T]_{\mathcal{B}}^{\mathcal{B}}$ directly from Definition 4.1.1.

(b) Compute it again with Theorem 4.2.3, starting from the standard matrix.

Solution.

(a) $T(1,1) = (2,2) = 2(1,1) + 0(2,1)$ and $T(2,1) = (6,3) = 0(1,1) + 3(2,1)$. Hence $[T]_{\mathcal{B}}^{\mathcal{B}} = \operatorname{diag}(2,3)$.

(b) $\mathbf{A} = [T]_{\mathcal{E}}^{\mathcal{E}} = \begin{pmatrix} 4 & -2 \\ 1 & 1 \end{pmatrix}$ and $\mathbf{P} = [\mathrm{id}]_{\mathcal{B}}^{\mathcal{E}} = \begin{pmatrix} 1 & 2 \\ 1 & 1 \end{pmatrix}$. Here $ad - bc = -1$, so by Theorem 0.9.8 $\mathbf{P}^{-1} = \begin{pmatrix} -1 & 2 \\ 1 & -1 \end{pmatrix}$. Then

$$\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \begin{pmatrix} -1 & 2 \\ 1 & -1 \end{pmatrix}\begin{pmatrix} 4 & -2 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} 1 & 2 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} -2 & 4 \\ 3 & -3 \end{pmatrix}\begin{pmatrix} 1 & 2 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix},$$

in agreement with (a).

Exercise. B3: Similar or not?

Determine which of the following pairs of real matrices are similar. Justify your answer, either with an invariant or with an explicit invertible $\mathbf{P}$.

(a) $\begin{pmatrix} 1 & 2 \\ 0 & 3 \end{pmatrix}$ and $\begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$.

(b) $\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$ and $\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$.

(c) $\begin{pmatrix} 1 & 1 \\ 0 & 2 \end{pmatrix}$ and $\begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$.

(d) $\begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$ and $\begin{pmatrix} 2 & 1 \\ 0 & 2 \end{pmatrix}$.

(e) $\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$ and $\begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix}$.

Solution.

(a) Not similar: the traces are 4 and 3, contradicting Theorem 4.2.5.

(b) Not similar: the ranks are 1 and 0, contradicting Proposition 4.2.6 (a).

(c) Similar. Let $\mathbf{A}$ be the first matrix. We look for a basis in which $\mathbf{x} \mapsto \mathbf{A}\mathbf{x}$ is diagonal: $\mathbf{A}\mathbf{e}_1 = \mathbf{e}_1$, and $\mathbf{A}(1,1) = (2,2) = 2(1,1)$. With $\mathbf{P} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$, $\mathbf{P}^{-1} = \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix}$ and

$$\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & 1 \\ 0 & 2 \end{pmatrix}\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ 0 & 2 \end{pmatrix}\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}.$$

(d) Not similar: $2\mathbf{I}_2$ is a scalar matrix, so $\mathbf{P}^{-1}(2\mathbf{I}_2)\mathbf{P} = 2\mathbf{I}_2$ for every invertible $\mathbf{P}$, and the second matrix is not $2\mathbf{I}_2$. The traces and ranks agree, so invariants alone would not have decided this.

(e) Similar. Let $\mathbf{A}$ be the first matrix. $\mathbf{A}(1,1) = (2,2)$ and $\mathbf{A}(1,-1) = (0,0)$. With

$$\mathbf{P} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \text{ and } \mathbf{P}^{-1} = \tfrac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix},$$

$$\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}\begin{pmatrix} 2 & 0 \\ 2 & 0 \end{pmatrix} = \frac{1}{2}\begin{pmatrix} 4 & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix},$$

where $\mathbf{A}\mathbf{P} = \begin{pmatrix} 2 & 0 \\ 2 & 0 \end{pmatrix}$ has columns $\mathbf{A}(1,1)$ and $\mathbf{A}(1,-1)$.

4.2.6.3 C. Going deeper

Exercise. C1: Matrices similar only to themselves

Let $n \geq 1$ and $\mathbf{A} \in M_n(F)$.

(a) Prove that $\mathbf{A}$ is similar only to itself if and only if $\mathbf{A} = c\mathbf{I}_n$ for some $c \in F$.

(b) Show that $\mathbf{N} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$ and $\mathbf{N}^\top = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}$ are similar, and explain the answer in terms of the operator $\mathbf{x} \mapsto \mathbf{N}\mathbf{x}$ and a reordered basis.

Hint: for (a), use the invertible matrices $\mathbf{I}_n + \mathbf{E}_{ij}$ with $i \neq j$.

Solution.

(a) ($\Leftarrow$) If $\mathbf{A} = c\mathbf{I}_n$, then $\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = c\mathbf{P}^{-1}\mathbf{P} = c\mathbf{I}_n = \mathbf{A}$ for every invertible $\mathbf{P}$, so the only matrix similar to $\mathbf{A}$ is $\mathbf{A}$.
($\Rightarrow$) Suppose $\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \mathbf{A}$ for every invertible $\mathbf{P}$, that is, $\mathbf{A}\mathbf{P} = \mathbf{P}\mathbf{A}$. If $n = 1$, $\mathbf{A} = (a_{11}) = a_{11}\mathbf{I}_1$. Let $n \geq 2$, and fix $i \neq j$. The matrix $\mathbf{I}_n + \mathbf{E}_{ij}$ is invertible, with inverse $\mathbf{I}_n - \mathbf{E}_{ij}$, since $\mathbf{E}_{ij}\mathbf{E}_{ij} = \mathbf{O}$ for $i \neq j$. So $\mathbf{A}(\mathbf{I}_n + \mathbf{E}_{ij}) = (\mathbf{I}_n + \mathbf{E}_{ij})\mathbf{A}$, which gives $\mathbf{A}\mathbf{E}_{ij} = \mathbf{E}_{ij}\mathbf{A}$. Compare the entries in position (k,l): by Definition 0.9.9,

$$(\mathbf{A}\mathbf{E}_{ij})_{kl} = \begin{cases} a_{ki} & l = j \\ 0 & l \neq j \end{cases}, \qquad (\mathbf{E}_{ij}\mathbf{A})_{kl} = \begin{cases} a_{jl} & k = i \\ 0 & k \neq i. \end{cases}$$

Taking $l = j$ and $k \neq i$ gives $a_{ki} = 0$. Taking $k = i$ and $l = j$ gives $a_{ii} = a_{jj}$. Since $i \neq j$ were arbitrary (and every index i has some $j \neq i$ because $n \geq 2$), all off-diagonal entries of $\mathbf{A}$ vanish and all diagonal entries are equal to $c = a_{11}$. Hence $\mathbf{A} = c\mathbf{I}_n$.

(b) Let $\mathbf{P} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$, so $\mathbf{P}^{-1} = \mathbf{P}$. Then $\mathbf{N}\mathbf{P} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$ and $\mathbf{P}^{-1}\mathbf{N}\mathbf{P} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} = \mathbf{N}^\top$.
In terms of operators: $T\mathbf{x} = \mathbf{N}\mathbf{x}$ has $T\mathbf{e}_1 = \mathbf{0}$ and $T\mathbf{e}_2 = \mathbf{e}_1$. In the reordered basis $\mathcal{B} = (\mathbf{e}_2, \mathbf{e}_1)$, the first basis vector goes to the second and the second goes to $\mathbf{0}$, so $[T]_{\mathcal{B}}^{\mathcal{B}} = \mathbf{N}^\top$. The same operator has both matrices, as Theorem 4.2.4 predicts.

Exercise. C2: AB and BA

Let $\mathbf{A}, \mathbf{B} \in M_n(F)$.

(a) Prove that if $\mathbf{A}$ is invertible, then $\mathbf{A}\mathbf{B} \sim \mathbf{B}\mathbf{A}$.

(b) Show that (a) fails without the invertibility hypothesis, even though $\operatorname{tr}(\mathbf{AB}) = \operatorname{tr}(\mathbf{BA})$ always holds.

Solution.

(a) Since $\mathbf{A}$ is invertible, $\mathbf{A}^{-1}(\mathbf{AB})\mathbf{A} = (\mathbf{A}^{-1}\mathbf{A})(\mathbf{BA}) = \mathbf{BA}$ by associativity. With $\mathbf{P} = \mathbf{A}$ this is $\mathbf{BA} = \mathbf{P}^{-1}(\mathbf{AB})\mathbf{P}$, so $\mathbf{AB} \sim \mathbf{BA}$.

(b) Take $\mathbf{A} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$ and $\mathbf{B} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$. Then $\mathbf{AB} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$, of rank 1, and $\mathbf{BA} = \mathbf{O}$, of rank 0. By Proposition 4.2.6 (a), $\mathbf{AB} \not\sim \mathbf{BA}$. Both traces are 0, in agreement with Theorem 0.9.6; the trace simply cannot detect the difference.

Exercise. C3: Every nonzero 2×2 matrix with square zero

Let $\mathbf{N} \in M_2(F)$ with $\mathbf{N} \neq \mathbf{O}$ and $\mathbf{N}^2 = \mathbf{O}$, and let $T\colon F^2 \to F^2$, $T\mathbf{x} = \mathbf{N}\mathbf{x}$.

(a) Explain why there is $\mathbf{v} \in F^2$ with $\mathbf{N}\mathbf{v} \neq \mathbf{0}$, and show that $\mathcal{B} = (\mathbf{N}\mathbf{v}, \mathbf{v})$ is a basis of F^2.

(b) Compute $[T]_{\mathcal{B}}^{\mathcal{B}}$, and deduce that $\mathbf{N} \sim \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$.

(c) Hence show that $\begin{pmatrix} 2 & -4 \\ 1 & -2 \end{pmatrix} \sim \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$ in $M_2(\mathbb{R})$, with an explicit $\mathbf{P}$.

Solution.

(a) Since $\mathbf{N} \neq \mathbf{O}$, some column $\mathbf{N}\mathbf{e}_j$ is nonzero (Theorem 0.9.3); take $\mathbf{v} = \mathbf{e}_j$. Let $a\mathbf{N}\mathbf{v} + b\mathbf{v} = \mathbf{0}$. Multiplying by $\mathbf{N}$ and using $\mathbf{N}^2 = \mathbf{O}$ gives $b\mathbf{N}\mathbf{v} = \mathbf{0}$, so $b = 0$ because $\mathbf{N}\mathbf{v} \neq \mathbf{0}$ (Theorem 1.1.9). Then $a\mathbf{N}\mathbf{v} = \mathbf{0}$ forces $a = 0$ for the same reason. So $\mathcal{B}$ is linearly independent of length $2 = \dim F^2$, hence a basis by Theorem 1.6.4.

(b) $T(\mathbf{N}\mathbf{v}) = \mathbf{N}^2\mathbf{v} = \mathbf{0}$ and $T\mathbf{v} = \mathbf{N}\mathbf{v} = 1 \cdot \mathbf{N}\mathbf{v} + 0 \cdot \mathbf{v}$. The columns of $[T]_{\mathcal{B}}^{\mathcal{B}}$ are therefore $(0,0)$ and $(1,0)$, so $[T]_{\mathcal{B}}^{\mathcal{B}} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$. Also $[T]_{\mathcal{E}}^{\mathcal{E}} = \mathbf{N}$. By Theorem 4.2.4 (a), $\mathbf{N} \sim \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$.

(c) Let $\mathbf{N} = \begin{pmatrix} 2 & -4 \\ 1 & -2 \end{pmatrix}$. Then $\mathbf{N}^2 = \begin{pmatrix} 4-4 & -8+8 \\ 2-2 & -4+4 \end{pmatrix} = \mathbf{O}$ and $\mathbf{N} \neq \mathbf{O}$. Take $\mathbf{v} = \mathbf{e}_1$, so $\mathbf{N}\mathbf{v} = (2,1)$, and $\mathcal{B} = ((2,1),(1,0))$. By Theorem 4.2.3, $\mathbf{P} = [\mathrm{id}]_{\mathcal{B}}^{\mathcal{E}} = \begin{pmatrix} 2 & 1 \\ 1 & 0 \end{pmatrix}$ satisfies $\mathbf{P}^{-1}\mathbf{N}\mathbf{P} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$. Check: $\mathbf{P}^{-1} = \begin{pmatrix} 0 & 1 \\ 1 & -2 \end{pmatrix}$ by Theorem 0.9.8 (here $ad - bc = -1$), $\mathbf{N}\mathbf{P} = \begin{pmatrix} 0 & 2 \\ 0 & 1 \end{pmatrix}$, and $\mathbf{P}^{-1}\mathbf{N}\mathbf{P} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$.

4.3 Rank, Equivalence, and Factorizations

Similarity compares two matrices of one operator, where the same basis is used for input and output, and the last section showed that deciding it is hard. This section asks what happens when the two bases may be chosen independently, as they may for any map $T\colon V \to W$. The answer is strikingly simple: the rank is then the only thing that matters. On the way we meet the column–row factorization, which writes every matrix as a product through a space of dimension equal to its rank, and we collect the basic inequalities for the rank of a composition.

4.3.1 The column–row factorization

In Chapter 3 we computed the rank of a matrix by row reduction, and the reduction gave more than a number. It produced a basis of the column space, the pivot columns of $\mathbf{A}$ (Theorem 3.5.2), and it expressed every other column in terms of them, with coefficients read off the reduced row echelon form. Recording both pieces of information as matrices gives a factorization.

Theorem 4.3.1. Column–Row Factorization

Let $\mathbf{A} \in M_{m\times n}(F)$ have rank $r \geq 1$, with pivot columns $j_1 < \cdots < j_r$. Let $\mathbf{C} \in M_{m\times r}(F)$ be the matrix whose columns are the pivot columns $\mathbf{a}_{j_1}, \ldots, \mathbf{a}_{j_r}$ of $\mathbf{A}$, and let $\mathbf{R} \in M_{r\times n}(F)$ consist of the first r rows of the reduced row echelon form of $\mathbf{A}$, that is, its nonzero rows. Then

$$\mathbf{A} = \mathbf{CR}.$$

Moreover, the columns of $\mathbf{C}$ are linearly independent and the rows of $\mathbf{R}$ are linearly independent.

Idea. Column l of $\mathbf{CR}$ is $\mathbf{C}$ times column l of $\mathbf{R}$, that is, a combination of the pivot columns of $\mathbf{A}$ with the entries of column l of the RREF as coefficients. For a pivot column the coefficients are a standard basis vector, and for a free column they are exactly the coefficients that Theorem 3.5.2 (b) provides. So the factorization is that theorem written as a matrix identity.

Proof. Let $\mathbf{E}$ be the RREF of $\mathbf{A}$, with entries e_{kl}. By Theorem 3.5.5, $\mathbf{E}$ has exactly r pivots, so its nonzero rows are rows $1, \ldots, r$. Fix a column index l. By Theorem 0.9.3, column l of $\mathbf{CR}$ is

$$\mathbf{C}\begin{pmatrix} e_{1l} \\ \vdots \\ e_{rl} \end{pmatrix} = e_{1l}\mathbf{a}_{j_1} + \cdots + e_{rl}\mathbf{a}_{j_r}.$$

Case 1: $l = j_i$ is a pivot column. By Lemma 3.3.3 (a), column j_i of $\mathbf{E}$ is $\mathbf{e}_i$, so the combination is $\mathbf{a}_{j_i} = \mathbf{a}_l$.

Case 2: l is a free column. Let t be the number of pivot columns to the left of l. For $k > t$, the leading entry of row k of $\mathbf{E}$ lies in column $j_k > l$, so $e_{kl} = 0$. The combination is therefore $e_{1l}\mathbf{a}_{j_1} + \cdots + e_{tl}\mathbf{a}_{j_t}$, which equals $\mathbf{a}_l$ by Theorem 3.5.2 (b).

In both cases column l of $\mathbf{CR}$ is column l of $\mathbf{A}$; hence $\mathbf{A} = \mathbf{CR}$. The columns of $\mathbf{C}$ form a basis of $\operatorname{col}(\mathbf{A})$ by Theorem 3.5.2 (a), so they are independent. The rows of $\mathbf{R}$ are the

nonzero rows of $\mathbf{E}$, which form a basis of $\mathrm{row}(\mathbf{A})$ by Theorem 3.5.3, so they are independent. This proves the theorem.

■

In words: every matrix of rank r is a "tall" matrix with independent columns times a "wide" matrix with independent rows, and the middle size is exactly r. The factorization cannot be squeezed further.

Proposition 4.3.2. Rank as the size of the thinnest factorization

Let $\mathbf{A} \in M_{m\times n}(F)$ be nonzero. Then $\operatorname{rank}\mathbf{A}$ is the **smallest** integer $s \geq 1$ such that $\mathbf{A} = \mathbf{BD}$ for some $\mathbf{B} \in M_{m\times s}(F)$ and $\mathbf{D} \in M_{s\times n}(F)$.

Proof. Let $r = \operatorname{rank}\mathbf{A}$; then $r \geq 1$ since $\mathbf{A} \neq \mathbf{O}$. By Theorem 4.3.1, a factorization with $s = r$ exists. Conversely, suppose $\mathbf{A} = \mathbf{BD}$ with $\mathbf{B} \in M_{m\times s}(F)$. By Theorem 3.5.9, $r = \operatorname{rank}(\mathbf{BD}) \leq \operatorname{rank}\mathbf{B}$, and $\operatorname{rank}\mathbf{B} \leq s$ because $\mathrm{col}(\mathbf{B})$ is spanned by s columns (Theorem 1.6.3). Hence $r \leq s$.

■

This gives a second view of "row rank equals column rank". Let $\mathbf{A} \neq \mathbf{O}$, and let $\mathbf{B}$ be the matrix whose $s = \dim\mathrm{col}(\mathbf{A})$ columns form a basis of $\mathrm{col}(\mathbf{A})$. Each column $\mathbf{a}_j$ of $\mathbf{A}$ is $\mathbf{Bd}_j$ for its coordinate column $\mathbf{d}_j$ (Theorem 0.9.3), so $\mathbf{A} = \mathbf{BD}$ with $\mathbf{D} = \begin{pmatrix}\mathbf{d}_1 & \cdots & \mathbf{d}_n\end{pmatrix}$. Then every row of $\mathbf{A}$ is a row of $\mathbf{B}$ times $\mathbf{D}$, a combination of the s rows of $\mathbf{D}$ (Theorem 0.9.2 (3) and the row form of Theorem 0.9.3), so $\dim\mathrm{row}(\mathbf{A}) \leq s = \dim\mathrm{col}(\mathbf{A})$. Applying the same argument to $\mathbf{A}^{\top}$ gives the reverse inequality. In the language of maps, the factorization says that $\mathbf{x} \mapsto \mathbf{Ax}$ can be done in two steps, $F^n \to F^r \to F^m$: first a surjection onto a space of dimension r, then an injection.

Example. A column–row factorization

Find the column–row factorization of $\mathbf{A} = \begin{pmatrix}1 & 2 & 0 & 3\\ 2 & 4 & 1 & 4\\ 3 & 6 & 1 & 7\end{pmatrix} \in M_{3\times 4}(\mathbb{R})$.

Solution. Row reduce. Subtracting 2 times row 1 from row 2 and 3 times row 1 from row 3 gives rows $(1,2,0,3)$, $(0,0,1,-2)$, $(0,0,1,-2)$. Subtracting row 2 from row 3 gives the RREF

$$\begin{pmatrix}1 & 2 & 0 & 3\\ 0 & 0 & 1 & -2\\ 0 & 0 & 0 & 0\end{pmatrix}.$$

The pivot columns are columns 1 and 3, so $r = 2$, and by Theorem 4.3.1

$$\mathbf{A} = \mathbf{CR} = \begin{pmatrix}1 & 0\\ 2 & 1\\ 3 & 1\end{pmatrix}\begin{pmatrix}1 & 2 & 0 & 3\\ 0 & 0 & 1 & -2\end{pmatrix}.$$

Check by rows: row 2 of $\mathbf{CR}$ is $2(1,2,0,3)+(0,0,1,-2)=(2,4,1,4)$, and row 3 is $3(1,2,0,3)+(0,0,1,-2)=(3,6,1,7)$. The last column of $\mathbf{R}$ says $\mathbf{a}_4 = 3\mathbf{a}_1 - 2\mathbf{a}_3$, and indeed $3(1,2,3)-2(0,1,1)=(3,4,7)$.

4.3.2 Equivalent matrices

Go back to the change-of-basis square for a map $T: V \to W$ (Theorem 4.2.3):

$$[T]_{\mathcal{B}'}^{\mathcal{C}'} = [\mathrm{id}]_{\mathcal{C}}^{\mathcal{C}'}\,[T]_{\mathcal{B}}^{\mathcal{C}}\,[\mathrm{id}]_{\mathcal{B}'}^{\mathcal{B}}.$$

When $V \neq W$, nothing ties the basis of V to the basis of W. So the two outer matrices are **unrelated** invertible matrices, of sizes $m \times m$ and $n \times n$. Any two matrices of the same map are related in this way, and this recurring shape deserves a name.

Two matrices are equivalent when they describe the same linear map, with the input basis and the output basis each changed freely.

Definition 4.3.1. Equivalent matrices

Let $\mathbf{A}, \mathbf{B} \in M_{m\times n}(F)$. We say $\mathbf{A}$ is **equivalent** to $\mathbf{B}$ if there **exist** invertible matrices $\mathbf{Q} \in M_m(F)$ and $\mathbf{P} \in M_n(F)$ such that

$$\mathbf{B} = \mathbf{QAP}.$$

In words: we may multiply on the left by **any** invertible matrix and on the right by **any** invertible matrix, and the two need not be related. Both sizes are forced: $\mathbf{Q}$ acts on the m rows and $\mathbf{P}$ on the n columns. This is an equivalence relation on $M_{m\times n}(F)$: $\mathbf{A} = \mathbf{I}_m\mathbf{A}\mathbf{I}_n$; if $\mathbf{B} = \mathbf{QAP}$ then $\mathbf{A} = \mathbf{Q}^{-1}\mathbf{B}\mathbf{P}^{-1}$; and if also $\mathbf{C} = \mathbf{Q}'\mathbf{B}\mathbf{P}'$, then $\mathbf{C} = (\mathbf{Q}'\mathbf{Q})\mathbf{A}(\mathbf{P}\mathbf{P}')$, where products of invertible matrices are invertible by Theorem 0.9.7 (part 3).

Examples.

- **Row equivalence.** If $\mathbf{B}$ is obtained from $\mathbf{A}$ by row operations, then $\mathbf{B} = \mathbf{EA}$ with $\mathbf{E}$ invertible (Theorem 3.4.8), so $\mathbf{A}$ and $\mathbf{B}$ are equivalent with $\mathbf{Q} = \mathbf{E}$ and $\mathbf{P} = \mathbf{I}_n$. Column operations correspond to multiplying on the right, and they are allowed too.
- **Similarity.** If $\mathbf{B} = \mathbf{P}^{-1}\mathbf{AP}$, then $\mathbf{A}$ and $\mathbf{B}$ are equivalent, with $\mathbf{Q} = \mathbf{P}^{-1}$.
- **The identity and a swap.** $\mathbf{I}_2$ and $S = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ are equivalent, since $S = S\mathbf{I}_2\mathbf{I}_2$. They are not similar, because $\mathbf{I}_2$ is similar only to itself.
- **Degenerate case.** The zero matrix $\mathbf{O}_{m\times n}$ is equivalent only to itself, since $\mathbf{QOP} = \mathbf{O}$. It is the matrix of the zero map in every pair of bases.

Nonexample by minimal change. Keep $\mathbf{B} = \mathbf{QAP}$ but drop the requirement that $\mathbf{Q}$ be invertible. Then every matrix could be sent to the zero matrix with $\mathbf{Q} = \mathbf{O}$, and the relation would not even be symmetric: $\mathbf{O} = \mathbf{O}\cdot\mathbf{I}_2\cdot\mathbf{I}_2$, but $\mathbf{I}_2 \neq \mathbf{QOP}$ for any $\mathbf{Q}, \mathbf{P}$. The clause "invertible" is what makes $\mathbf{QAP}$ a change of basis rather than an arbitrary composition.

Warning. **Equivalent is much weaker than similar.** The matrices $\mathbf{I}_2$ and $J = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ are equivalent, since $J = J\mathbf{I}_2\mathbf{I}_2$ and J is invertible, but they are not similar (see the warning

in the previous section). For square matrices, "equivalent" allows a different change of basis on each side, and that freedom destroys almost all the structure an operator has.

How much structure survives? The rank does, since multiplying by invertible matrices does not change it (Theorem 3.5.9). The theorem below says that nothing else survives. The target is the simplest matrix of each rank.

For $0 \le r \le \min(m, n)$, let $\mathbf{N}_r \in M_{m\times n}(F)$ be the block matrix

$$\mathbf{N}_r \coloneqq \begin{pmatrix} \mathbf{I}_r & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{pmatrix},$$

with 1 in positions $(1,1), \dots, (r,r)$ and 0 everywhere else. Its first r columns are $\mathbf{e}_1, \dots, \mathbf{e}_r \in F^m$, and its remaining columns are zero.

Theorem 4.3.3. Rank Normal Form

(a) Let V and W be finite-dimensional with $\dim V = n$ and $\dim W = m$, and let $T\colon V \to W$ be linear of rank r. Then there are bases $\mathcal{B}$ of V and $\mathcal{C}$ of W such that $[T]_{\mathcal{B}}^{\mathcal{C}} = \mathbf{N}_r$.

(b) Every $\mathbf{A} \in M_{m\times n}(F)$ is equivalent to $\mathbf{N}_r$, where $r = \operatorname{rank}\mathbf{A}$.

(c) Two matrices $\mathbf{A}, \mathbf{B} \in M_{m\times n}(F)$ are equivalent if and only if $\operatorname{rank}\mathbf{A} = \operatorname{rank}\mathbf{B}$.

Idea. Work backwards from the goal. $[T]_{\mathcal{B}}^{\mathcal{C}} = \mathbf{N}_r$ says: T sends the first r vectors of $\mathcal{B}$ to the first r vectors of $\mathcal{C}$, and sends the last $n-r$ vectors of $\mathcal{B}$ to 0. So ① the last $n-r$ vectors of $\mathcal{B}$ should be a basis of $\ker T$, and by Rank–Nullity that is exactly the right number; ② the first r vectors of $\mathcal{B}$ extend it to a basis of V; ③ their images are forced to be the first r vectors of $\mathcal{C}$, and we must show they are independent, which follows by counting; ④ extend them to a basis of W. These are the bases of the proof of Rank–Nullity, one on each side.

Proof.

(a) By the Rank–Nullity Theorem (Theorem 2.3.1), $\dim\ker T = n - r$. Let $(u_{r+1}, \dots, u_n)$ be a basis of $\ker T$, a subspace of V by Theorem 2.2.1. By the Basis Extension Theorem (Theorem 1.6.5) there are $u_1, \dots, u_r \in V$ such that $\mathcal{B} = (u_1, \dots, u_r, u_{r+1}, \dots, u_n)$ is a basis of V; the new vectors may be listed first, since reordering a basis gives a basis. Put $w_i = Tu_i$ for $i = 1, \dots, r$.
By Theorem 2.2.4, $\operatorname{im} T = \operatorname{span}(Tu_1, \dots, Tu_n)$. Since $Tu_j = 0$ for $j > r$, removing these vectors does not change the span, so $(w_1, \dots, w_r)$ spans $\operatorname{im} T$. This list has length $r = \dim\operatorname{im} T$, so it is a basis of $\operatorname{im} T$ by Theorem 1.6.4 (b); in particular it is linearly independent in W. By the Basis Extension Theorem there are $w_{r+1}, \dots, w_m \in W$ such that $\mathcal{C} = (w_1, \dots, w_m)$ is a basis of W.
Now read off the columns of $[T]_{\mathcal{B}}^{\mathcal{C}}$ (Definition 4.1.1). For $i \le r$, $Tu_i = w_i$, whose $\mathcal{C}$-coordinate vector is $\mathbf{e}_i$. For $j > r$, $Tu_j = 0$, whose coordinate vector is $\mathbf{0}$. These are the columns of $\mathbf{N}_r$.

(b) Let $T\colon F^n \to F^m$, $T\mathbf{x} = \mathbf{A}\mathbf{x}$, and let $\mathcal{E}_n$, $\mathcal{E}_m$ be the standard bases. The j-th column of $[T]_{\mathcal{E}_n}^{\mathcal{E}_m}$ is $[\mathbf{A}\mathbf{e}_j]_{\mathcal{E}_m} = \mathbf{A}\mathbf{e}_j$, the j-th column of $\mathbf{A}$ (Theorem 0.9.3), so $[T]_{\mathcal{E}_n}^{\mathcal{E}_m} = \mathbf{A}$. By Theorem 4.1.5, $\operatorname{rank} T = \operatorname{rank}\mathbf{A} = r$. Take $\mathcal{B}$, $\mathcal{C}$ as in (a). By Theorem 4.2.3,

$$\mathbf{N}_r = [T]_{\mathcal{B}}^{\mathcal{C}} = [\mathrm{id}]_{\mathcal{E}_m}^{\mathcal{C}}\, \mathbf{A}\, [\mathrm{id}]_{\mathcal{B}}^{\mathcal{E}_n},$$

and both outer matrices are invertible by Theorem 4.2.1 (c). Hence $\mathbf{A}$ is equivalent to $\mathbf{N}_r$.

(c) ($\Rightarrow$) If $\mathbf{B} = \mathbf{QAP}$ with $\mathbf{Q}, \mathbf{P}$ invertible, then $\operatorname{rank}\mathbf{B} = \operatorname{rank}\mathbf{A}$ by Theorem 3.5.9. ($\Leftarrow$) If $\operatorname{rank}\mathbf{A} = \operatorname{rank}\mathbf{B} = r$, then by (b) both $\mathbf{A}$ and $\mathbf{B}$ are equivalent to $\mathbf{N}_r$. Since equivalence is symmetric and transitive, $\mathbf{A}$ is equivalent to $\mathbf{B}$. This proves the theorem.

■

So the equivalence classes of $M_{m\times n}(F)$ are exactly the sets of matrices of rank r, for $r = 0, 1, \dots, \min(m, n)$: there are $\min(m, n) + 1$ of them, whatever the field. Every linear map between finite-dimensional spaces looks like "keep r coordinates and delete the rest" once the bases on both sides are chosen well. **The right pair of bases makes any single map trivial.**

Remark. Contrast this with operators. For $T \in \mathcal{L}(V)$ we must use **one** basis for input and output, and the relation becomes similarity. There the rank is far from the whole story: $\mathbf{I}_2$ and $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ have the same rank but are not similar. Classifying operators up to a change of one basis is a much harder problem, answered in Chapter 10 by the Jordan and rational canonical forms.

The proof is also an algorithm. To find $\mathbf{Q}$ and $\mathbf{P}$ for a concrete matrix, compute a basis of $\operatorname{nul}(\mathbf{A})$, extend it, and take images.

Example. Bringing a 2×3 matrix to normal form

Let $\mathbf{A} = \begin{pmatrix} 1 & 2 & 1 \\ 2 & 4 & 3 \end{pmatrix} \in M_{2\times 3}(\mathbb{R})$. Find invertible $\mathbf{Q} \in M_2(\mathbb{R})$ and $\mathbf{P} \in M_3(\mathbb{R})$ with $\mathbf{QAP} = \mathbf{N}_r$.

Solution. Subtracting 2 times row 1 from row 2 gives $\begin{pmatrix} 1 & 2 & 1 \\ 0 & 0 & 1 \end{pmatrix}$, so $r = \operatorname{rank}\mathbf{A} = 2$, and the solutions of $\mathbf{Ax} = \mathbf{0}$ satisfy $x_3 = 0$, $x_1 = -2x_2$. Hence $\operatorname{nul}(\mathbf{A}) = \operatorname{span}((-2, 1, 0))$, as Rank–Nullity predicts $(3 - 2 = 1)$.

Following the proof, let $\mathbf{u}_3 = (-2, 1, 0)$, and extend with $\mathbf{u}_1 = \mathbf{e}_1$, $\mathbf{u}_2 = \mathbf{e}_3$. The list $(\mathbf{e}_1, \mathbf{e}_3, \mathbf{u}_3)$ is independent: in $a\mathbf{e}_1 + b\mathbf{e}_3 + c(-2, 1, 0) = (a - 2c, c, b) = \mathbf{0}$, the second and third entries give $c = b = 0$, and then $a = 0$. So it is a basis $\mathcal{B}$ of $\mathbb{R}^3$. The images are $\mathbf{w}_1 = \mathbf{Ae}_1 = (1, 2)$ and $\mathbf{w}_2 = \mathbf{Ae}_3 = (1, 3)$, which already form a basis $\mathcal{C}$ of $\mathbb{R}^2$ since $r = m = 2$. Therefore

$$\mathbf{P} = [\mathrm{id}]_{\mathcal{B}}^{\mathcal{E}_3} = \begin{pmatrix} 1 & 0 & -2 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}, \qquad \mathbf{Q} = [\mathrm{id}]_{\mathcal{E}_2}^{\mathcal{C}} = \begin{pmatrix} 1 & 1 \\ 2 & 3 \end{pmatrix}^{-1} = \begin{pmatrix} 3 & -1 \\ -2 & 1 \end{pmatrix},$$

using Theorem 4.2.1 (c) and Theorem 0.9.8. Check: $\mathbf{AP} = \begin{pmatrix} 1 & 1 & 0 \\ 2 & 3 & 0 \end{pmatrix}$ (the columns are $\mathbf{w}_1$,

$\mathbf{w}_2$, $\mathbf{0}$), and

$$\mathbf{QAP} = \begin{pmatrix} 3 & -1 \\ -2 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & 0 \\ 2 & 3 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} = \mathbf{N}_2.$$

The pair $(\mathbf{Q}, \mathbf{P})$ is far from unique; any choice of extension gives another.

Quick check. How many equivalence classes does $M_{3\times 2}(F)$ have, and which is the class of $\begin{pmatrix} 1 & 2 \\ 2 & 4 \\ 0 & 0 \end{pmatrix}$?

Solution. Ranks in $M_{3\times 2}(F)$ range over $0, 1, 2$, so there are $\min(3, 2) + 1 = 3$ classes by Theorem 4.3.3 (c). The given matrix is nonzero and its second column is twice the first, so its rank is 1, and it lies in the class of $\mathbf{N}_1$: all matrices of rank 1.

4.3.3 Rank inequalities for compositions

The rank of a composition cannot exceed the rank of either factor, and it cannot fall too far either. Both facts come from the same observation: ST is S applied to the subspace $\operatorname{im} T$. For a subspace $X \subseteq W$ and a linear map $S\colon W \to U$, the **restriction** $S|_X\colon X \to U$, $x \mapsto Sx$, is linear, since its values are those of S. Its kernel is $\ker S \cap X$.

Theorem 4.3.4. Rank of a Composition

Let $T\colon V \to W$ and $S\colon W \to U$ be linear, with V and W finite-dimensional. Then

$$\operatorname{rank}(ST) \le \min(\operatorname{rank} S, \operatorname{rank} T).$$

Proof. Consider the restriction $S' = S|_{\operatorname{im} T}\colon \operatorname{im} T \to U$. For $v \in V$, $(ST)(v) = S(Tv) = S'(Tv)$, so $\operatorname{im}(ST) \subseteq \operatorname{im} S'$; conversely every $S'w$ with $w = Tv \in \operatorname{im} T$ equals $(ST)(v)$. Hence $\operatorname{im}(ST) = \operatorname{im} S'$. Since $\operatorname{im} T$ is finite-dimensional (a subspace of W, Theorem 1.6.7), Rank–Nullity (Theorem 2.3.1) applied to S' gives

$$\operatorname{rank}(ST) = \dim \operatorname{im} S' = \dim \operatorname{im} T - \dim \ker S' \le \operatorname{rank} T.$$

Also $\operatorname{im}(ST) = \operatorname{im} S' \subseteq \operatorname{im} S$, and $\operatorname{im} S$ is finite-dimensional by Rank–Nullity for S, so $\operatorname{rank}(ST) \le \operatorname{rank} S$ by Theorem 1.6.7. This proves the theorem. ∎

The lower bound uses the same restriction, but now we control its kernel instead of ignoring it.

Theorem 4.3.5. Sylvester's Rank Inequality

Let $T\colon V \to W$ and $S\colon W \to U$ be linear, with V and W finite-dimensional. Then

$$\operatorname{rank}(ST) \geq \operatorname{rank} S + \operatorname{rank} T - \dim W.$$

In matrix form: for $\mathbf{A} \in M_{m\times n}(F)$ and $\mathbf{B} \in M_{n\times p}(F)$, $\operatorname{rank}(\mathbf{AB}) \geq \operatorname{rank}\mathbf{A} + \operatorname{rank}\mathbf{B} - n$.

Idea. Of the $\operatorname{rank} T$ dimensions that reach W, the map S can destroy only those lying in $\ker S$, and $\ker S$ has dimension $\dim W - \operatorname{rank} S$. Rank–Nullity for the restriction $S|_{\operatorname{im} T}$ turns this sentence into an inequality.

Proof. Let $S' = S|_{\operatorname{im} T}$, so that $\operatorname{im} S' = \operatorname{im}(ST)$ as in the proof of Theorem 4.3.4, and $\ker S' = \ker S \cap \operatorname{im} T \subseteq \ker S$. By Rank–Nullity for S' and then for S,

$$\begin{aligned}\operatorname{rank} T &= \operatorname{rank}(ST) + \dim(\ker S \cap \operatorname{im} T)\\ &\leq \operatorname{rank}(ST) + \dim\ker S = \operatorname{rank}(ST) + \dim W - \operatorname{rank} S,\end{aligned}$$

where the inequality is Theorem 1.6.7. Rearranging gives the claim. For matrices, apply it to $S\mathbf{y} = \mathbf{Ay}$ on $W = F^n$ and $T\mathbf{x} = \mathbf{Bx}$ on $V = F^p$: the matrix of ST in standard bases is $\mathbf{AB}$ by Theorem 4.1.2, and ranks of maps and matrices agree by Theorem 4.1.5. ∎

If you worked Exercise (C1: A Lower Bound for the Rank of a Composite), you have seen this proof already; we record the result here as a theorem so that later arguments can cite it. A typical use: if $\mathbf{AB} = \mathbf{O}$ with $\mathbf{A} \in M_{m\times n}(F)$ and $\mathbf{B} \in M_{n\times p}(F)$, then $0 \geq \operatorname{rank}\mathbf{A} + \operatorname{rank}\mathbf{B} - n$, that is,

$$\operatorname{rank}\mathbf{A} + \operatorname{rank}\mathbf{B} \leq n.$$

For instance, there are no $\mathbf{A}, \mathbf{B} \in M_3(\mathbb{R})$ of rank 2 with $\mathbf{AB} = \mathbf{O}$.

Warning. **The rank of a product depends on the order of the factors.** For $\mathbf{A} = \begin{pmatrix}1 & 0\\ 0 & 0\end{pmatrix}$ and $\mathbf{B} = \begin{pmatrix}0 & 1\\ 0 & 0\end{pmatrix}$, $\mathbf{AB} = \mathbf{B}$ has rank 1 but $\mathbf{BA} = \mathbf{O}$ has rank 0. Theorem 3.5.9 and Sylvester's inequality bound $\operatorname{rank}(\mathbf{AB})$ from both sides; they do not determine it, and they say nothing that would make $\operatorname{rank}(\mathbf{AB}) = \operatorname{rank}(\mathbf{BA})$.

4.3.4 Exercises

4.3.4.1 A. Check your understanding

Exercise. A1

(a) Define what it means for $\mathbf{A}, \mathbf{B} \in M_{m\times n}(F)$ to be equivalent.

(b) State the Rank Normal Form theorem for a matrix $\mathbf{A} \in M_{m\times n}(F)$.

(c) True or false: equivalent square matrices are similar. Justify your answer.

(d) True or false: $\operatorname{rank}(\mathbf{AB}) = \operatorname{rank}(\mathbf{BA})$ for all $\mathbf{A}, \mathbf{B} \in M_n(F)$. Justify your answer.

(e) How many equivalence classes does $M_{2\times 5}(\mathbb{F}_7)$ have?

(f) In the proof of the normal form for $T\colon V \to W$, which vectors form the end of the basis of V, and which vectors form the start of the basis of W?

Solution.

(a) There are invertible $\mathbf{Q} \in M_m(F)$ and $\mathbf{P} \in M_n(F)$ with $\mathbf{B} = \mathbf{QAP}$ (Definition 4.3.1).

(b) If $r = \operatorname{rank} \mathbf{A}$, there are invertible $\mathbf{Q}, \mathbf{P}$ with $\mathbf{QAP} = \begin{pmatrix} \mathbf{I}_r & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{pmatrix}$; and two $m \times n$ matrices are equivalent if and only if they have the same rank (Theorem 4.3.3).

(c) False. $\mathbf{I}_2$ and $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ both have rank 2, so they are equivalent, but $\mathbf{I}_2$ is similar only to itself.

(d) False. With $\mathbf{A} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$ and $\mathbf{B} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$, $\operatorname{rank}(\mathbf{AB}) = 1$ and $\operatorname{rank}(\mathbf{BA}) = 0$.

(e) The possible ranks are $0, 1, 2$, so there are 3 classes. The field plays no role.

(f) The last $n - r$ vectors of the basis of V form a basis of $\ker T$. The first r vectors of the basis of W are the images of the first r basis vectors of V, and they form a basis of $\operatorname{im} T$.

4.3.4.2 B. Practice

Exercise. B1: A column–row factorization

Find the column–row factorization of

$$\mathbf{A} = \begin{pmatrix} 1 & -1 & 2 & 0 \\ 2 & -2 & 5 & 1 \\ 1 & -1 & 3 & 1 \end{pmatrix} \in M_{3\times 4}(\mathbb{R}).$$

Hence write the second and fourth columns of $\mathbf{A}$ as combinations of its pivot columns.

Solution. Subtract 2 times row 1 from row 2 and row 1 from row 3: the rows become $(1, -1, 2, 0)$, $(0, 0, 1, 1)$, $(0, 0, 1, 1)$. Subtract row 2 from row 3, and then 2 times row 2 from row 1:

$$\mathrm{RREF} = \begin{pmatrix} 1 & -1 & 0 & -2 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

The pivot columns are 1 and 3, so $\operatorname{rank} \mathbf{A} = 2$, and by Theorem 4.3.1

$$\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 2 & 5 \\ 1 & 3 \end{pmatrix} \begin{pmatrix} 1 & -1 & 0 & -2 \\ 0 & 0 & 1 & 1 \end{pmatrix}.$$

Check by rows: $2(1, -1, 0, -2) + 5(0, 0, 1, 1) = (2, -2, 5, 1)$ and $(1, -1, 0, -2) + 3(0, 0, 1, 1) = (1, -1, 3, 1)$; row 1 is $(1, -1, 0, -2) + 2(0, 0, 1, 1)$. Reading columns 2 and 4 of $\mathbf{R}$: $\mathbf{a}_2 = -\mathbf{a}_1$ and $\mathbf{a}_4 = -2\mathbf{a}_1 + \mathbf{a}_3$. Indeed $-2(1, 2, 1) + (2, 5, 3) = (0, 1, 1) = \mathbf{a}_4$.

Exercise. B2: Normal form of a rank-one matrix

Let $\mathbf{A} = \begin{pmatrix} 1 & -1 & 2 \\ -2 & 2 & -4 \end{pmatrix} \in M_{2\times 3}(\mathbb{R})$. Find invertible $\mathbf{Q} \in M_2(\mathbb{R})$ and $\mathbf{P} \in M_3(\mathbb{R})$ such that $\mathbf{QAP} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$, following the proof of Theorem 4.3.3.

Solution. Row 2 is -2 times row 1, and $\mathbf{A} \neq \mathbf{O}$, so $\operatorname{rank}\mathbf{A} = 1$. The null space is $\{x_1 - x_2 + 2x_3 = 0\}$; setting the free variables x_2, x_3 to $(1,0)$ and $(0,1)$ gives the basis $\mathbf{u}_2 = (1,1,0)$, $\mathbf{u}_3 = (-2,0,1)$. Extend with $\mathbf{u}_1 = \mathbf{e}_1$: in $a\mathbf{e}_1 + b\mathbf{u}_2 + c\mathbf{u}_3 = (a+b-2c, b, c) = \mathbf{0}$, the last two entries give $b = c = 0$ and then $a = 0$, so $\mathcal{B} = (\mathbf{e}_1, \mathbf{u}_2, \mathbf{u}_3)$ is a basis of $\mathbb{R}^3$. Its first image is $\mathbf{w}_1 = \mathbf{A}\mathbf{e}_1 = (1,-2)$, which we extend to the basis $\mathcal{C} = ((1,-2), \mathbf{e}_2)$ of $\mathbb{R}^2$ (independent, since $a(1,-2) + b(0,1) = (a, -2a+b) = \mathbf{0}$ forces $a = b = 0$). Then

$$\mathbf{P} = \begin{pmatrix} 1 & 1 & -2 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \qquad \mathbf{Q} = \begin{pmatrix} 1 & 0 \\ -2 & 1 \end{pmatrix}^{-1} = \begin{pmatrix} 1 & 0 \\ 2 & 1 \end{pmatrix}.$$

Check: $\mathbf{AP}$ has columns $\mathbf{A}\mathbf{e}_1 = (1,-2)$, $\mathbf{A}\mathbf{u}_2 = \mathbf{0}$, $\mathbf{A}\mathbf{u}_3 = \mathbf{0}$, and $\mathbf{Q}(1,-2) = (1,0)$. So $\mathbf{QAP} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$. Both $\mathbf{Q}$ and $\mathbf{P}$ are invertible by Theorem 4.2.1 (c).

Exercise. B3: Rank-one pieces

Let $\mathbf{A} \in M_{m\times n}(F)$.

(a) Use Theorem 4.3.1 to show that if $\operatorname{rank}\mathbf{A} = 1$, then $\mathbf{A} = \mathbf{x}\mathbf{y}^{\top}$ for some nonzero $\mathbf{x} \in F^m$ and $\mathbf{y} \in F^n$.

(b) Show that if $\operatorname{rank}\mathbf{A} = r \geq 1$, then $\mathbf{A}$ is a sum of r matrices of rank 1, and that it is not a sum of fewer than r matrices of rank 1.

(c) Write $\begin{pmatrix} 1 & 2 & 0 & 3 \\ 2 & 4 & 1 & 4 \\ 3 & 6 & 1 & 7 \end{pmatrix}$ as a sum of two rank-one matrices.

Solution.

(a) With $r = 1$, $\mathbf{C} \in M_{m\times 1}(F)$ is a single column $\mathbf{x}$, and $\mathbf{R} \in M_{1\times n}(F)$ is a single row, which we write as $\mathbf{y}^{\top}$ with $\mathbf{y} \in F^n$. So $\mathbf{A} = \mathbf{x}\mathbf{y}^{\top}$. Both are nonzero: the columns of $\mathbf{C}$ and the rows of $\mathbf{R}$ are linearly independent by Theorem 4.3.1, and a list containing a zero vector is dependent. (This recovers one direction of Exercise (C2: Matrices of rank one).)

(b) Write $\mathbf{A} = \mathbf{CR}$ with columns $\mathbf{c}_1, \dots, \mathbf{c}_r$ of $\mathbf{C}$ and rows $\mathbf{r}_1, \dots, \mathbf{r}_r$ of $\mathbf{R}$, and let c_{ik}, r_{kl} be the entries of $\mathbf{C}$, $\mathbf{R}$. The $m \times n$ matrix $\mathbf{c}_k\mathbf{r}_k$ has (i,l)-entry $c_{ik}r_{kl}$, so by Theorem 0.9.2

(1),

$$(\mathbf{CR})_{il} = \sum_{k=1}^{r} c_{ik} r_{kl} = \sum_{k=1}^{r} (\mathbf{c}_k \mathbf{r}_k)_{il}, \qquad \text{that is,} \qquad \mathbf{A} = \mathbf{c}_1\mathbf{r}_1 + \cdots + \mathbf{c}_r\mathbf{r}_r.$$

Each $\mathbf{c}_k\mathbf{r}_k$ has rank 1 by the ($\Leftarrow$) direction of Exercise (C2: Matrices of rank one), since $\mathbf{c}_k \neq \mathbf{0}$ (a column of $\mathbf{C}$, whose columns are independent) and $\mathbf{r}_k \neq \mathbf{0}$ (a row of $\mathbf{R}$, whose rows are independent). If $\mathbf{A} = \mathbf{A}_1 + \cdots + \mathbf{A}_s$ with each $\operatorname{rank}\mathbf{A}_i = 1$, then by Exercise (C1: Rank of a sum) applied repeatedly, $r = \operatorname{rank}\mathbf{A} \leq \operatorname{rank}\mathbf{A}_1 + \cdots + \operatorname{rank}\mathbf{A}_s = s$.

(c) By Example (A column–row factorization), $\mathbf{A} = \mathbf{CR}$ with $\mathbf{C} = \begin{pmatrix} 1 & 0 \\ 2 & 1 \\ 3 & 1 \end{pmatrix}$ and $\mathbf{R} = \begin{pmatrix} 1 & 2 & 0 & 3 \\ 0 & 0 & 1 & -2 \end{pmatrix}$, so

$$\mathbf{A} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} \begin{pmatrix} 1 & 2 & 0 & 3 \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} \begin{pmatrix} 0 & 0 & 1 & -2 \end{pmatrix}$$
$$= \begin{pmatrix} 1 & 2 & 0 & 3 \\ 2 & 4 & 0 & 6 \\ 3 & 6 & 0 & 9 \end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -2 \\ 0 & 0 & 1 & -2 \end{pmatrix}.$$

4.3.4.3 C. Going deeper

Exercise. C1: Frobenius's rank inequality

Let $\mathbf{A} \in M_{m\times n}(F)$, $\mathbf{B} \in M_{n\times p}(F)$, and $\mathbf{C} \in M_{p\times q}(F)$. Write $N = \operatorname{nul}(\mathbf{A}) \subseteq F^n$, and let $S\colon F^n \to F^m$ be $S\mathbf{y} = \mathbf{Ay}$.

(a) Show that $\operatorname{col}(\mathbf{AB}) = \{\mathbf{Ay} : \mathbf{y} \in \operatorname{col}(\mathbf{B})\}$, and deduce from Rank–Nullity for $S|_{\operatorname{col}(\mathbf{B})}$ that $\operatorname{rank}(\mathbf{AB}) = \operatorname{rank}\mathbf{B} - \dim(N \cap \operatorname{col}(\mathbf{B}))$.

(b) Show similarly that $\operatorname{rank}(\mathbf{ABC}) = \operatorname{rank}(\mathbf{BC}) - \dim(N \cap \operatorname{col}(\mathbf{BC}))$.

(c) Prove that $\operatorname{col}(\mathbf{BC}) \subseteq \operatorname{col}(\mathbf{B})$, and deduce **Frobenius's inequality**

$$\operatorname{rank}(\mathbf{ABC}) + \operatorname{rank}\mathbf{B} \geq \operatorname{rank}(\mathbf{AB}) + \operatorname{rank}(\mathbf{BC}).$$

(d) Deduce Sylvester's inequality for matrices from (c).

Hint: for (d), look for a special choice of the middle factor.

Solution.

(a) By Theorem 0.9.2, every vector $\mathbf{ABx}$ equals $\mathbf{Ay}$ with $\mathbf{y} = \mathbf{Bx} \in \operatorname{col}(\mathbf{B})$; conversely every $\mathbf{y} \in \operatorname{col}(\mathbf{B})$ is $\mathbf{Bx}$ for some $\mathbf{x}$, and then $\mathbf{Ay} = (\mathbf{AB})\mathbf{x} \in \operatorname{col}(\mathbf{AB})$. So $\operatorname{col}(\mathbf{AB}) = \operatorname{im}(S|_{\operatorname{col}(\mathbf{B})})$. The kernel of $S|_{\operatorname{col}(\mathbf{B})}$ is $N \cap \operatorname{col}(\mathbf{B})$, and $\operatorname{col}(\mathbf{B})$ is finite-dimensional, so by Theorem 2.3.1,

$$\operatorname{rank}\mathbf{B} = \dim\operatorname{col}(\mathbf{B}) = \operatorname{rank}(\mathbf{AB}) + \dim(N \cap \operatorname{col}(\mathbf{B})).$$

(b) The same argument with $\mathbf{BC}$ in place of $\mathbf{B}$ gives $\operatorname{col}(\mathbf{ABC}) = \operatorname{im}(S|_{\operatorname{col}(\mathbf{BC})})$, with kernel $N \cap \operatorname{col}(\mathbf{BC})$, so $\operatorname{rank}(\mathbf{BC}) = \operatorname{rank}(\mathbf{ABC}) + \dim(N \cap \operatorname{col}(\mathbf{BC}))$.

(c) Every vector $\mathbf{BCx} = \mathbf{B}(\mathbf{Cx})$ lies in $\operatorname{col}(\mathbf{B})$, so $\operatorname{col}(\mathbf{BC}) \subseteq \operatorname{col}(\mathbf{B})$. Hence $N \cap \operatorname{col}(\mathbf{BC}) \subseteq N \cap \operatorname{col}(\mathbf{B})$, and by Theorem 1.6.7, $\dim(N \cap \operatorname{col}(\mathbf{BC})) \leq \dim(N \cap \operatorname{col}(\mathbf{B}))$. By (b) and then (a),

$$\begin{aligned} \operatorname{rank}(\mathbf{ABC}) &= \operatorname{rank}(\mathbf{BC}) - \dim(N \cap \operatorname{col}(\mathbf{BC})) \\ &\geq \operatorname{rank}(\mathbf{BC}) - \dim(N \cap \operatorname{col}(\mathbf{B})) \\ &= \operatorname{rank}(\mathbf{BC}) - \operatorname{rank}\mathbf{B} + \operatorname{rank}(\mathbf{AB}). \end{aligned}$$

Rearranging gives Frobenius's inequality.

(d) Let $\mathbf{A} \in M_{m\times n}(F)$ and $\mathbf{C} \in M_{n\times q}(F)$, and apply (c) with $p = n$ and $\mathbf{B} = \mathbf{I}_n$. Then $\operatorname{rank}(\mathbf{AC}) + n \geq \operatorname{rank}\mathbf{A} + \operatorname{rank}\mathbf{C}$, since $\operatorname{rank}\mathbf{I}_n = n$. This is Theorem 4.3.5 for matrices.

Exercise. C2: Counting over $\mathbb{F}_2$

(a) Show that $M_2(\mathbb{F}_2)$ contains exactly 1 matrix of rank 0, 6 of rank 2, and 9 of rank 1.

(b) Deduce the sizes of the equivalence classes of $M_2(\mathbb{F}_2)$.

(c) Compare with similarity: how many matrices of $M_2(\mathbb{F}_2)$ are similar to $\mathbf{I}_2$? Is the class of $\mathbf{I}_2$ under equivalence a union of several similarity classes?

Solution.

(a) The only matrix of rank 0 is $\mathbf{O}$. A matrix has rank 2 exactly when its two columns span $\mathbb{F}_2^2$, that is (Theorem 1.6.4), when they form a basis of $\mathbb{F}_2^2$. The first column can be any of the 3 nonzero vectors of $\mathbb{F}_2^2$. The second must lie outside the span of the first, which has 2 elements, so there are $4 - 2 = 2$ choices. That gives $3 \cdot 2 = 6$ matrices of rank 2. Since $|M_2(\mathbb{F}_2)| = 2^4 = 16$, the remaining $16 - 1 - 6 = 9$ matrices have rank 1.

(b) By Theorem 4.3.3 (c), the classes are the sets of matrices of a fixed rank, of sizes 1, 9, and 6.

(c) $\mathbf{I}_2$ is similar only to itself, so its similarity class has 1 element, while its equivalence class, the invertible matrices, has 6. Since similar matrices are equivalent, each equivalence class is a union of similarity classes, and here the class of the 6 invertible matrices contains the similarity class $\{\mathbf{I}_2\}$ together with others. For example $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ is invertible but not similar to $\mathbf{I}_2$, so there are at least two similarity classes inside.

4.4 Products and Quotient Spaces

So far every new vector space in this book has been found **inside** an old one, as a subspace. This section builds new spaces in two other ways. The first glues spaces together side by side, making a product; it is the "external" direct sum promised in Chapter 1. The second, the quotient space V/U, is subtler and more important: it collapses a subspace U to a single point and keeps only what V knows beyond U. Quotients let us state precisely what a linear map remembers and what it forgets, in the three Isomorphism Theorems.

4.4.1 Products of vector spaces

In Chapter 1 we wrote $V = U \oplus W$ for subspaces U and W of one space V (Definition 1.7.2). The two pieces already lived in a common space, and "direct" was a property of how they sat there. But often we want to combine two spaces that have nothing to do with each other, such as $\mathbb{R}^2$ and $\mathbb{R}[x]_{\leq 1}$, into one space in which each keeps its identity. There is no common space to add inside, so we build one.

The model is F^2. A vector of F^2 is a pair (a, b) of scalars, and we add and scale pairs entry by entry. Nothing in that recipe needs the entries to be scalars: the entries could be vectors, taken from any vector spaces over F.

The product of vector spaces is the space of lists with one entry from each space, added and scaled entry by entry.

Definition 4.4.1. Product of vector spaces

Let $V_1, \ldots, V_k$ be vector spaces over the **same** field F. Their **product** (or **external direct sum**) is the Cartesian product (Definition 0.3.10)

$$V_1 \times \cdots \times V_k = \{(v_1, \ldots, v_k) : v_i \in V_i \text{ for each } i\},$$

with addition and scalar multiplication defined **entrywise**: for $a \in F$,

$$(v_1, \ldots, v_k) + (w_1, \ldots, w_k) \coloneqq (v_1 + w_1, \ldots, v_k + w_k),$$
$$a(v_1, \ldots, v_k) \coloneqq (av_1, \ldots, av_k).$$

In words: the i-th entry of a sum is the sum of the i-th entries, computed **in** V_i, and similarly for scalar multiples. Two elements are equal exactly when all their entries are equal. Every operation happens inside one V_i at a time, which is why all the V_i must share the field F: the same scalar a has to act on every entry.

It is a vector space. Both operations land in $V_1 \times \cdots \times V_k$, since $v_i + w_i \in V_i$ and $av_i \in V_i$. Each axiom of Definition 1.1.1 is an equation between lists, so it holds when it holds in each entry, and in each entry it is the same axiom in V_i. For example, (VS6) reads

$$\begin{aligned}(a+b)(v_1, \ldots, v_k) &= ((a+b)v_1, \ldots, (a+b)v_k) \\ &= (av_1 + bv_1, \ldots, av_k + bv_k) \\ &= a(v_1, \ldots, v_k) + b(v_1, \ldots, v_k),\end{aligned}$$

using (VS6) in each V_i. The zero vector is $(0, \ldots, 0)$, with the zero of V_i in the i-th entry, and the additive inverse of $(v_1, \ldots, v_k)$ is $(-v_1, \ldots, -v_k)$. The other axioms are checked in exactly the same way.

Examples.

- **The standard example.** $F \times \cdots \times F$ (n factors) is F^n with its usual operations.
- **Unrelated spaces.** An element of $\mathbb{R}^2 \times \mathbb{R}[x]_{\leq 1}$ is a pair $((a, b), c + dx)$, and $((1, 0), x) + ((2, 5), 3 - x) = ((3, 5), 3)$.
- **Augmented data.** A linear system $\mathbf{Ax} = \mathbf{b}$ with m equations in n unknowns is a single element $(\mathbf{A}, \mathbf{b})$ of $M_{m\times n}(F) \times F^m$.
- **Degenerate cases.** For $k = 1$ the product is V_1 itself, with entries written in parentheses. And $V \times \{0\}$ consists of the pairs $(v, 0)$; the map $v \mapsto (v, 0)$ is a linear bijection $V \to V \times \{0\}$, so a zero factor adds nothing.

Nonexample by minimal change. Take $V_1 = \mathbb{C}$ as a vector space over $\mathbb{C}$ and $V_2 = \mathbb{R}$ over $\mathbb{R}$. The set $\mathbb{C} \times \mathbb{R}$ is fine, but the entrywise rule $i \cdot (z, t) = (iz, it)$ fails: $it \notin \mathbb{R}$ for $t \neq 0$. The clause that fails is "over the **same** field". Restricting the scalars of $\mathbb{C}$ to $\mathbb{R}$ (Definition 1.9.1) repairs it, and $\mathbb{C} \times \mathbb{R}$ over $\mathbb{R}$ is a vector space of dimension 3.

Bases of the factors combine into a basis of the product. For $v \in V_i$, write $\iota_i(v)$ for the list with v in the i-th entry and 0 in every other entry. The map $\iota_i \colon V_i \to V_1 \times \cdots \times V_k$ is linear, because the operations are entrywise.

Theorem 4.4.1. Dimension of a Product

Let $V_1, \ldots, V_k$ be finite-dimensional vector spaces over F, and let $(b_{i1}, \ldots, b_{in_i})$ be a basis of V_i for each i. Then the list of all vectors $\iota_i(b_{ij})$, for $i = 1, \ldots, k$ and $j = 1, \ldots, n_i$, is a basis of $V_1 \times \cdots \times V_k$. In particular,

$$\dim(V_1 \times \cdots \times V_k) = \dim V_1 + \cdots + \dim V_k.$$

Proof. Spanning. Let $(v_1, \ldots, v_k)$ be in the product. Since the operations are entrywise, $(v_1, \ldots, v_k) = \iota_1(v_1) + \cdots + \iota_k(v_k)$. Each v_i is a combination $\sum_j c_{ij} b_{ij}$, and ι_i is linear, so $\iota_i(v_i) = \sum_j c_{ij}\, \iota_i(b_{ij})$ (Theorem 2.1.4). Hence $(v_1, \ldots, v_k)$ is a combination of the vectors $\iota_i(b_{ij})$.

Independence. Let $\sum_{i=1}^{k} \sum_{j=1}^{n_i} c_{ij}\, \iota_i(b_{ij}) = 0$. The i-th entry of the left side is $\sum_j c_{ij} b_{ij}$, because every $\iota_{i'}(b_{i'j})$ with $i' \neq i$ has 0 there. So $\sum_j c_{ij} b_{ij} = 0$ in V_i for each i, and the independence of the basis of V_i gives $c_{ij} = 0$ for all j.

The list has $n_1 + \cdots + n_k$ vectors, which gives the dimension (Definition 1.6.1).

■

For instance, $\dim(\mathbb{R}^2 \times \mathbb{R}[x]_{\leq 1}) = 2+2 = 4$, so this product is isomorphic to $\mathbb{R}^4$ by Theorem 2.5.4.

How does the external product relate to the internal direct sum of Chapter 1? If $U_1, \ldots, U_k$ are subspaces of V, there is a natural linear map from the product to V: add up the entries. The sum is direct exactly when this map loses no information.

Theorem 4.4.2. Internal and External Direct Sums

(a) Let $U_1, \ldots, U_k$ be subspaces of a vector space V, and let

$$S \colon U_1 \times \cdots \times U_k \to V, \qquad S(u_1, \ldots, u_k) = u_1 + \cdots + u_k.$$

Then S is linear with $\operatorname{im} S = U_1 + \cdots + U_k$. It is injective if and only if the sum $U_1 + \cdots + U_k$ is direct. In that case S is an isomorphism from $U_1 \times \cdots \times U_k$ onto $U_1 \oplus \cdots \oplus U_k$.

(b) Conversely, let $V_1, \ldots, V_k$ be vector spaces over F. Then the images $\iota_i(V_i)$ are subspaces of $V_1 \times \cdots \times V_k$, each isomorphic to V_i via ι_i, and

$$V_1 \times \cdots \times V_k = \iota_1(V_1) \oplus \cdots \oplus \iota_k(V_k).$$

Proof.

(a) For lists (u_i), (u_i'), and $a \in F$, regrouping the finite sums gives $S((u_i)+(u_i')) = \sum_i (u_i+u_i') = S((u_i))+S((u_i'))$ and $S(a(u_i)) = \sum_i au_i = aS((u_i))$, so S is linear. Its image is $\{u_1+\cdots+u_k : u_i \in U_i\}$, which is $U_1+\cdots+U_k$ by Definition 1.7.1. Its kernel consists of the lists with $u_1+\cdots+u_k = 0$. By Theorem 2.2.3, S is injective if and only if the only such list is $(0,\dots,0)$, which is condition (b) of Theorem 1.7.5, equivalent to the sum being direct. In that case S, viewed as a map onto its image, is a linear bijection, hence an isomorphism (Definition 2.5.2).

(b) Each ι_i is linear and injective (its kernel is $\{0\}$, since $\iota_i(v)$ has v as its i-th entry), so $\iota_i(V_i)$ is a subspace (Theorem 2.2.2) and ι_i is a bijection onto it. Every list satisfies $(v_1,\dots,v_k) = \iota_1(v_1)+\cdots+\iota_k(v_k)$, so the product is the sum of the subspaces $\iota_i(V_i)$. If $\iota_1(v_1)+\cdots+\iota_k(v_k) = 0$, then $(v_1,\dots,v_k) = (0,\dots,0)$, so every piece $\iota_i(v_i)$ is 0. By Theorem 1.7.5 ((b) $\Rightarrow$ (a)) the sum is direct.

■

So internal and external direct sums are the same thing seen from two sides. An internal direct sum is a product that happens to live inside V; a product is an internal direct sum of copies of its factors. Many books write $\oplus$ for both. This book keeps $\times$ for the external construction, so that $U \oplus W$ always refers to subspaces of one space.

Quick check. Use Theorem 4.4.2 (a) and Rank–Nullity to explain why, for **finite-dimensional** subspaces U, W of V, the sum $U+W$ is direct if and only if $\dim(U+W) = \dim U + \dim W$.

Solution. The addition map $S\colon U \times W \to V$ has image $U+W$, and $\dim(U\times W) = \dim U + \dim W$ by Theorem 4.4.1. By Theorem 2.3.1, $\dim U + \dim W = \dim(U+W) + \dim\ker S$. So $\dim(U+W) = \dim U + \dim W$ exactly when $\ker S = \{0\}$, that is, when S is injective, which by the theorem means the sum is direct. This recovers Theorem 1.7.4 (a) $\Leftrightarrow$ (c).

4.4.2 Cosets

The second construction starts from a picture you have already seen. In Chapter 1 we compared the line $U = \{x+2y=0\}$ through the origin with the parallel line $L = \{x+2y=3\}$, which is not a subspace. In Chapter 3 we found that the solutions of a consistent system $\mathbf{Ax}=\mathbf{b}$ form a set $\mathbf{p}+\mathrm{nul}(\mathbf{A})$, a subspace shifted by one particular solution (Theorem 3.1.3). Both are translates of a subspace. They deserve a name.

Definition 4.4.2. Coset

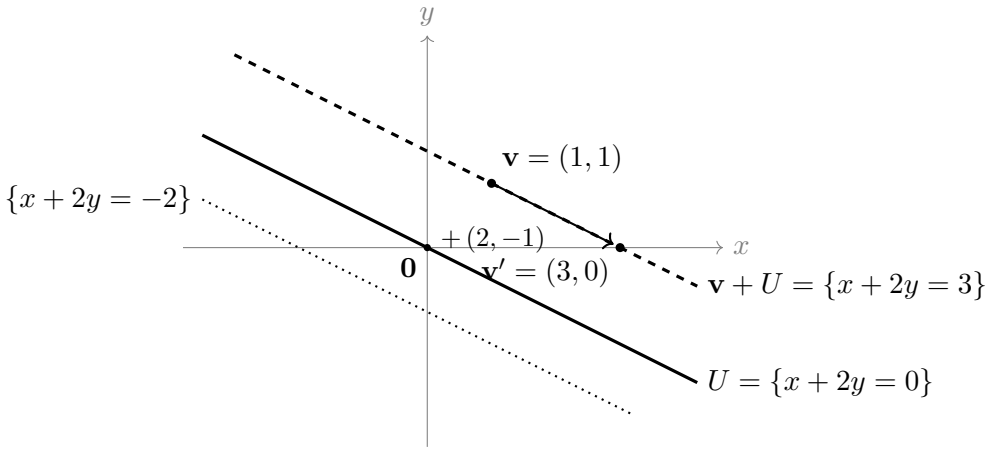
Let U be a subspace of a vector space V, and let $v \in V$. The **coset** of U through v is the subset

$$v+U \coloneqq \{v+u : u \in U\} \subseteq V.$$

Any $w \in v+U$ is called a **representative** of the coset.

In $\mathbb{R}^2$, with $U = \{x+2y=0\} = \mathrm{span}((2,-1))$, the coset through $(3,0)$ is the line $\{x+2y=3\}$:

a point $(3,0)+t(2,-1)=(3+2t,-t)$ satisfies $x+2y=3$, and every solution of $x+2y=3$ is of this form with $t=-y$. In general $(a,b)+U=\{x+2y=a+2b\}$, so the cosets of U are exactly the lines parallel to U.



The picture shows three cosets of U: U itself (solid), the line through $\mathbf{v}=(1,1)$ (dashed), and another parallel line (dotted). The points $\mathbf{v}=(1,1)$ and $\mathbf{v}'=(3,0)$ lie on the same dashed line, because $\mathbf{v}'-\mathbf{v}=(2,-1)\in U$; so they are two representatives of **one** coset. That observation is the whole theory of cosets.

Lemma 4.4.3. When two cosets are equal

Let U be a subspace of V. For $v,w\in V$, write $v\sim w$ if $v-w\in U$.

(a) $\sim$ is an equivalence relation on V, and the equivalence class of v is the coset $v+U$.

(b) $v+U=w+U$ if and only if $v-w\in U$.

(c) Two cosets of U are either equal or disjoint, and every vector of V lies in exactly one coset.

Proof.

(a) We check Definition 0.5.2. (E1) $v-v=0\in U$. (E2) If $v-w\in U$, then $w-v=(-1)(v-w)\in U$ by Theorem 1.1.7 and closure under scalar multiplication. (E3) If $v-w\in U$ and $w-x\in U$, then $v-x=(v-w)+(w-x)\in U$ by closure under addition. For the class, $w\sim v$ means $w-v=u$ for some $u\in U$, that is, $w=v+u\in v+U$; so the class $[v]$ of Definition 0.5.3 is $v+U$.

(b) By (a) and Lemma 0.5.1 (a), $v+U=w+U$ if and only if $v\sim w$, that is, $v-w\in U$.

(c) This is Lemma 0.5.1 (b), together with $v\in v+U$ (take $u=0$).

■

More examples.

- **Vertical lines.** In $\mathbb{R}^3$, let $U=\operatorname{span}(\mathbf{e}_3)$, the z-axis. The coset $(a,b,c)+U$ is the vertical line through $(a,b,0)$, and $(a,b,c)+U=(a',b',c')+U$ exactly when $a=a'$ and $b=b'$. A coset is determined by the shadow of any of its points on the xy-plane.
- **Polynomials with a prescribed value.** In $F[x]_{\le n}$, let $U=\{p:p(0)=0\}$, a subspace (the kernel of evaluation at 0). Then $p+U=q+U$ exactly when $(p-q)(0)=0$, that is,

$p(0) = q(0)$. So $p + U$ is the set of all polynomials with the same constant term as p.
- **Degenerate cases.** If $U = \{0\}$, each coset $v + U = \{v\}$ is a single vector. If $U = V$, there is only one coset, $v + V = V$, because $v - w \in V$ always.

Nonexample by minimal change. Try to use the line $L = \{x + 2y = 3\}$, which is not a subspace, in place of U. The rule "$\mathbf{v} \sim \mathbf{w}$ if $\mathbf{v} - \mathbf{w} \in L$" is not even reflexive: $\mathbf{v} - \mathbf{v} = \mathbf{0} \notin L$. What fails is the hypothesis that U is a subspace, which (E1), (E2), and (E3) each used.

Warning. **A coset has many names, and it is usually not a subspace.** In the example above, $(1,1) + U$ and $(3,0) + U$ are the **same** set, although $(1,1) \neq (3,0)$. Never conclude $v = w$ from $v + U = w + U$; conclude only $v - w \in U$. And $v + U$ is a subspace only when it contains 0, that is, when $v \in U$ and so $v + U = U$. The dashed line above is not a subspace.

4.4.3 The quotient space

Now we want to treat each coset as a single point. The motivation is a kind of deliberate forgetting. Suppose we only care about vectors of $\mathbb{R}^2$ "up to" adding something from U, for instance because we only ever measure the quantity $x + 2y$. Then $(1,1)$ and $(3,0)$ carry the same information, and the space of all the information we keep is the set of lines parallel to U. We would like this set to be a vector space, so that we can add and scale lines.

The recipe is the one used for $\mathbb{Z}/n\mathbb{Z}$ in Chapter 0 (Example (Addition modulo n is well defined)): define the operations through representatives, then check that the choice of representatives does not matter.

The quotient space V/U is V with everything in U declared to be zero.

Definition 4.4.3. Quotient space

Let U be a subspace of a vector space V over F. The **quotient space** V/U (read "V modulo U") is the set of all cosets of U,

$$V/U \coloneqq \{v + U : v \in V\},$$

with the operations

$$(v + U) + (w + U) \coloneqq (v + w) + U,$$
$$a(v + U) \coloneqq av + U \qquad (v, w \in V,\ a \in F).$$

The **quotient map** is $\pi\colon V \to V/U$, $\pi(v) = v + U$.

In words: an element of V/U is a whole coset, a subset of V. By Lemma 4.4.3 (a), V/U is the quotient set of V by the relation $v \sim w \Leftrightarrow v - w \in U$ (Definition 0.5.5), and π is its quotient map. To add two cosets, pick a representative of each, add them in V, and take the coset of the result. To scale a coset, scale a representative. The operations are written with representatives, so before anything else we must check that they are **well defined**.

The danger is real. By Lemma 4.4.3, $(1,1) + U$ and $(3,0) + U$ are one element of $\mathbb{R}^2/U$. If the sum $((1,1) + U) + (\mathbf{w} + U)$ depended on which of the two names we used, the "definition" would assign two values to one pair of cosets, and it would not define a function at all (Definition 0.4.1). The theorem below rules this out, and then shows that the operations make V/U a vector space.

Theorem 4.4.4. The Quotient Space Is a Vector Space

Let U be a subspace of a vector space V over F.

(a) The addition and scalar multiplication in Definition 4.4.3 are well defined: the right-hand sides depend only on the cosets $v+U$ and $w+U$, not on the representatives v and w.

(b) With these operations V/U is a vector space over F. Its zero vector is $0+U=U$, and the additive inverse of $v+U$ is $(-v)+U$.

(c) The quotient map $\pi\colon V\to V/U$ is linear and surjective, and $\ker\pi=U$.

Idea. Chapter 0's Theorem 0.5.3 turns a function on V that is constant on classes into a function on V/U. It is stated for functions of one variable, so we use it one variable at a time. ① For scalar multiplication by a fixed a: if $v-v'\in U$, then $av-av'=a(v-v')\in U$. ② For addition, first freeze the second summand w and vary the first; then show that the resulting function on cosets does not change when w is replaced by an equivalent w'. Each check is one line and uses exactly one closure property of U. Once the operations exist, every axiom for cosets is the same axiom in V, passed through π.

Proof. Throughout, $v\sim v'$ means $v-v'\in U$, and by Lemma 4.4.3 (b), $v\sim v'$ exactly when $v+U=v'+U$.

(a) *Scalar multiplication.* Fix $a\in F$, and let $g_a\colon V\to V/U$, $g_a(v)=av+U$. If $v\sim v'$, then $av-av'=a(v-v')\in U$ by (VS5) and closure of U under scalar multiplication, so $g_a(v)=g_a(v')$. By Theorem 0.5.3 (a), there is exactly one function $\bar{g}_a\colon V/U\to V/U$ with $\bar{g}_a(v+U)=av+U$ for every $v\in V$. We define $a(v+U)\coloneqq\bar{g}_a(v+U)$.
Addition, first variable. Fix $w\in V$, and let $f_w\colon V\to V/U$, $f_w(v)=(v+w)+U$. If $v\sim v'$, then $(v+w)-(v'+w)=v-v'\in U$, so $f_w(v)=f_w(v')$. By Theorem 0.5.3 (a), there is exactly one function $\bar{f}_w\colon V/U\to V/U$ with $\bar{f}_w(v+U)=(v+w)+U$ for every v.
Addition, second variable. Let $\mathcal{F}$ be the set of all functions $V/U\to V/U$, and let $h\colon V\to\mathcal{F}$, $h(w)=\bar{f}_w$. Suppose $w\sim w'$. For every $v\in V$, $(v+w)-(v+w')=w-w'\in U$, so

$$\bar{f}_w(v+U)=(v+w)+U=(v+w')+U=\bar{f}_{w'}(v+U).$$

Since every element of V/U has the form $v+U$, the functions $\bar{f}_w$ and $\bar{f}_{w'}$ are equal, that is, $h(w)=h(w')$. By Theorem 0.5.3 (a) again, there is exactly one function $\bar{h}\colon V/U\to\mathcal{F}$ with $\bar{h}(w+U)=\bar{f}_w$. We define $X+Y\coloneqq\bar{h}(Y)(X)$ for $X,Y\in V/U$. Then

$$(v+U)+(w+U)=\bar{h}(w+U)(v+U)=\bar{f}_w(v+U)=(v+w)+U,$$

which is the formula of Definition 4.4.3, and its value depends only on the two cosets.

(b) Every element of V/U is $\pi(v)=v+U$ for some v, and by (a) we have $\pi(v)+\pi(w)=\pi(v+w)$ and $a\pi(v)=\pi(av)$. Let $X=\pi(x)$, $Y=\pi(y)$, $Z=\pi(z)$ be in V/U, and $a,b\in F$. Each axiom follows by moving π outside, using the axiom in V, and moving π back:

- (VS1) $X+Y=\pi(x+y)=\pi(y+x)=Y+X$.

- (VS2) $(X+Y)+Z = \pi((x+y)+z) = \pi(x+(y+z)) = X+(Y+Z)$.
- (VS3) $X+\pi(0) = \pi(x+0) = \pi(x) = X$, so $\pi(0) = 0+U = U$ is a zero vector.
- (VS4) $X+\pi(-x) = \pi(x+(-x)) = \pi(0)$, so $\pi(-x) = (-x)+U$ is an additive inverse of X.
- (VS5) $a(X+Y) = \pi(a(x+y)) = \pi(ax+ay) = aX+aY$.
- (VS6) $(a+b)X = \pi((a+b)x) = \pi(ax+bx) = aX+bX$.
- (VS7) $a(bX) = \pi(a(bx)) = \pi((ab)x) = (ab)X$.
- (VS8) $1X = \pi(1x) = \pi(x) = X$.

Hence V/U is a vector space over F.

(c) The identities $\pi(v+w) = \pi(v)+\pi(w)$ and $\pi(av) = a\pi(v)$ from (b) say that π is linear. It is surjective by the definition of V/U. Finally, $v \in \ker\pi$ means $v+U = 0+U$, which by Lemma 4.4.3 (b) means $v = v-0 \in U$. So $\ker\pi = U$.

■

Part (c) has a pleasant reading: **every subspace is a kernel**. Given $U \subseteq V$, the quotient map is a linear map whose kernel is exactly U.

Examples.

- **Parallel lines.** In $\mathbb{R}^2/U$ with $U = \{x+2y=0\}$, write ℓ_c for the line $\{x+2y=c\}$. Since $(a,b) \in \ell_{a+2b}$, the definition gives $\ell_c + \ell_d = \ell_{c+d}$ and $t\,\ell_c = \ell_{tc}$: to add two parallel lines, add their "heights" c and d. The zero vector is $\ell_0 = U$.
- **Constant terms.** In $F[x]_{\le n}/U$ with $U = \{p : p(0) = 0\}$, the coset $p+U$ is determined by $p(0)$, and $(p+U)+(q+U) = (p+q)+U$ is determined by $p(0)+q(0)$. All that survives of a polynomial is its constant term.
- **Degenerate cases.** $V/\{0\}$ has the cosets $\{v\}$, added exactly as in V: nothing is forgotten. V/V has the single element V, so it is the zero space: everything is forgotten.

Nonexample by minimal change. Replace the line U by the union of the two axes, $S = \operatorname{span}(\mathbf{e}_1) \cup \operatorname{span}(\mathbf{e}_2) \subseteq \mathbb{R}^2$. It still contains $\mathbf{0}$ and is still closed under scalar multiplication, but it is not closed under addition. Then "$\mathbf{v}-\mathbf{w} \in S$" is not transitive: $(1,1)-(0,1) = (1,0) \in S$ and $(0,1)-(0,0) = (0,1) \in S$, but $(1,1)-(0,0) = (1,1) \notin S$. Concretely, the "cosets" $(1,1)+S$ and $(0,0)+S = S$ both contain $(0,1) = (1,1)+(-1,0)$, yet they are different sets, since $\mathbf{0} \in S$ but $(1,1)+\mathbf{s} = \mathbf{0}$ would need $\mathbf{s} = (-1,-1) \notin S$. The sets $\mathbf{v}+S$ do not split $\mathbb{R}^2$ into disjoint pieces, so there is no set "$\mathbb{R}^2/S$" of the kind the definition needs. The clause that fails is "U is a subspace", specifically closure under addition.

Why this definition. Each part of the hypothesis that U is a subspace has a job: $0 \in U$ makes $\sim$ reflexive; closure under scalar multiplication makes $\sim$ symmetric and scaling of cosets well defined; closure under addition makes $\sim$ transitive and addition of cosets well defined. The nonexample shows what goes wrong when one of them is missing. The notation V/U and the phrase "modulo U" come from $\mathbb{Z}/n\mathbb{Z}$, where integers are identified when their difference lies in $n\mathbb{Z}$.

Warning. V/U **is not a subspace of** V**.** Its elements are subsets of V, not vectors of V, and its addition is a new operation. For example, $\mathbb{R}^3/\operatorname{span}(\mathbf{e}_3)$ is not the xy-plane: its elements are vertical lines. The xy-plane is a **complement** of the z-axis, one of many (Chapter 1), whereas V/U involves no choice at all. The two are isomorphic, as we show below, but they are different objects.

In finite dimension the size of a quotient is exactly what one would guess from the picture: we lose $\dim U$ dimensions.

Theorem 4.4.5. Dimension of a Quotient

Let V be finite-dimensional and U a subspace of V. Then V/U is finite-dimensional and

$$\dim(V/U) = \dim V - \dim U.$$

More precisely, if $(u_1, \dots, u_m, w_1, \dots, w_s)$ is a basis of V such that $(u_1, \dots, u_m)$ is a basis of U, then $(w_1 + U, \dots, w_s + U)$ is a basis of V/U.

Proof. By Theorem 4.4.4 (c), $\pi\colon V \to V/U$ is linear, surjective, and has kernel U. By the Rank–Nullity Theorem (Theorem 2.3.1), $\operatorname{im}\pi = V/U$ is finite-dimensional and $\dim V = \dim U + \dim(V/U)$, which is the formula.

For the basis, let $v + U \in V/U$ and write $v = \sum_i a_i u_i + \sum_j b_j w_j$. Since π is linear and $\pi(u_i) = U$ is the zero vector, Theorem 2.1.4 gives $v + U = \sum_j b_j(w_j + U)$. So the list $(w_1 + U, \dots, w_s + U)$ spans V/U. Its length is $s = \dim V - \dim U = \dim(V/U)$, so it is a basis by Theorem 1.6.4 (b).

■

Such a basis always exists: take a basis of U and extend it (Theorem 1.6.5). So to find a basis of a quotient, **extend a basis of U to one of V and keep the cosets of the new vectors.**

Example. A basis of a quotient of $\mathbb{R}^3$

Let $U = \operatorname{span}((1,1,1)) \subseteq \mathbb{R}^3$. Find a basis of $\mathbb{R}^3/U$, and write $(2,5,-1) + U$ in it.

Solution. Extend $(1,1,1)$ by $\mathbf{e}_1$ and $\mathbf{e}_2$. The list $((1,1,1), \mathbf{e}_1, \mathbf{e}_2)$ is independent: in $a(1,1,1) + b\mathbf{e}_1 + c\mathbf{e}_2 = (a+b, a+c, a) = \mathbf{0}$, the third entry gives $a = 0$, and then $b = c = 0$. It has length 3, so it is a basis of $\mathbb{R}^3$ (Theorem 1.6.4). By Theorem 4.4.5, $(\mathbf{e}_1 + U, \mathbf{e}_2 + U)$ is a basis of $\mathbb{R}^3/U$, which has dimension $3 - 1 = 2$.

To write $(2,5,-1) + U$, solve $(2,5,-1) = a(1,1,1) + b\mathbf{e}_1 + c\mathbf{e}_2$: the third entry gives $a = -1$, then $b = 3$ and $c = 6$. Hence

$$(2,5,-1) + U = 3(\mathbf{e}_1 + U) + 6(\mathbf{e}_2 + U),$$

and indeed $(2,5,-1) - (3,6,0) = (-1,-1,-1) \in U$. The coordinates $(3,6)$ are $(x - z, y - z)$ evaluated at $(2,5,-1)$: they are exactly the quantities that do not change when a multiple of $(1,1,1)$ is added.

The warning above said that a quotient is not a complement. The next result says how they are related: any complement is a concrete model of the quotient.

Proposition 4.4.6. A complement is a model of the quotient

Let U be a subspace of V, and let W be a subspace with $V = U \oplus W$. Then the restriction of the quotient map, $W \to V/U$, $w \mapsto w + U$, is an isomorphism.

Proof. The restriction of the linear map π to W is linear. *Injective:* if $w+U = U$ with $w \in W$, then $w \in U$ by Lemma 4.4.3 (b), so $w \in U \cap W = \{0\}$ (Theorem 1.7.4); by Theorem 2.2.3 the restriction is injective. *Surjective:* given $v+U$, write $v = u+w$ with $u \in U$, $w \in W$; then $v - w = u \in U$, so $v + U = w + U$. A linear bijection is an isomorphism (Definition 2.5.2). ∎

So every complement of U is isomorphic to V/U, which gives a second proof that all complements have the same dimension. The quotient is the choice-free object that all the complements imitate.

Quick check. In $\mathbb{R}^2/U$ with $U = \operatorname{span}((1,1))$, is $(2,3) + U = (0,1) + U$? Is $(1,0) + U = (0,1) + U$? What is $\dim(\mathbb{R}^2/U)$?

Solution. By Lemma 4.4.3 (b): $(2,3) - (0,1) = (2,2) \in U$, so the first two cosets are equal. $(1,0) - (0,1) = (1,-1) \notin U$, so the second pair are different cosets. By Theorem 4.4.5, $\dim(\mathbb{R}^2/U) = 2 - 1 = 1$.

4.4.4 The First Isomorphism Theorem

A linear map $T\colon V \to W$ forgets exactly the information in its kernel: $Tv = Tv'$ if and only if $T(v - v') = 0$, that is, $v - v' \in \ker T$. In the language of this section, T takes the same value at two vectors exactly when they lie in the same coset of $\ker T$. So T is really a function of the coset, and on cosets it forgets nothing.

Theorem 4.4.7. First Isomorphism Theorem

Let $T\colon V \to W$ be a linear map, and let $K = \ker T$. Then the rule

$$\bar{T}\colon V/K \to W, \qquad \bar{T}(v + K) = Tv$$

defines a linear map, which is injective with $\operatorname{im}\bar{T} = \operatorname{im} T$, and satisfies $\bar{T} \circ \pi = T$. Hence

$$V/\ker T \cong \operatorname{im} T.$$

Proof. Well defined. K is a subspace by Theorem 2.2.1, so V/K is a vector space. If $v + K = v' + K$, then $v - v' \in K$ by Lemma 4.4.3 (b), so $Tv - Tv' = T(v - v') = 0$. Hence T is constant on each coset of K, and by Theorem 0.5.3 (a) there is exactly one function $\bar{T}\colon V/K \to W$ with $\bar{T}(v + K) = Tv$, that is, $\bar{T} \circ \pi = T$.

Linear. For $v, w \in V$ and $a \in F$, by Definition 4.4.3 and the linearity of T,

$$\begin{aligned}\bar{T}((v + K) + (w + K)) &= \bar{T}((v + w) + K) = T(v + w)\\ &= Tv + Tw = \bar{T}(v + K) + \bar{T}(w + K),\end{aligned}$$

and similarly $\bar{T}(a(v+K)) = \bar{T}(av+K) = T(av) = a\bar{T}(v+K)$.

Injective. If $\bar{T}(v+K) = 0$, then $Tv = 0$, so $v \in K$ and $v + K = K$, the zero of V/K. By Theorem 2.2.3, $\bar{T}$ is injective.

Image. The values of $\bar{T}$ are the vectors Tv, $v \in V$, so $\operatorname{im}\bar{T} = \operatorname{im} T$. Regarded as a map $V/K \to \operatorname{im} T$, $\bar{T}$ is a linear bijection, hence an isomorphism (Definition 2.5.2).

■

So every linear map factors as "collapse the kernel, then embed": $T = \bar{T} \circ \pi$, with π surjective and $\bar{T}$ injective. When V is finite-dimensional, taking dimensions with Theorem 4.4.5 gives $\dim V - \dim\ker T = \dim\operatorname{im} T$, which is Rank–Nullity again. The theorem is stronger than the count, because it names the isomorphism, and it holds in infinite dimension too.

The practical use runs the other way: to identify a quotient V/U, find a surjective linear map out of V whose kernel is exactly U.

Example. Identifying quotients

Use the First Isomorphism Theorem to show:

(a) $\mathbb{R}^3/L \cong \mathbb{R}^2$, where $L = \operatorname{span}((1,1,1))$;

(b) $\mathbb{R}[x]_{\le n}/U \cong \mathbb{R}$, where $U = \{p : p(0) = 0\}$.

Solution.

(a) Let $T\colon \mathbb{R}^3 \to \mathbb{R}^2$, $T(x,y,z) = (x - z,\ y - z)$, which is linear because each entry is a linear expression. It is surjective, since $T(a,b,0) = (a,b)$. Its kernel is $\{x = z,\ y = z\} = \{(t,t,t)\} = L$. By Theorem 4.4.7, $(x,y,z) + L \mapsto (x - z, y - z)$ is an isomorphism $\mathbb{R}^3/L \to \mathbb{R}^2$. This explains the coordinates found in Example (A basis of a quotient of $\mathbb{R}^3$). The check that it is well defined is visible: adding (t,t,t) to (x,y,z) does not change $x - z$ or $y - z$.

(b) Let $\mathrm{ev}_0\colon \mathbb{R}[x]_{\le n} \to \mathbb{R}$, $p \mapsto p(0)$. It is linear by Theorem 0.8.5, surjective because the constant polynomial c maps to c, and its kernel is U by definition. By Theorem 4.4.7, $p + U \mapsto p(0)$ is an isomorphism $\mathbb{R}[x]_{\le n}/U \to \mathbb{R}$. In particular $\dim(\mathbb{R}[x]_{\le n}/U) = 1$ by Theorem 2.5.4, so $\dim U = (n+1) - 1 = n$ by Theorem 4.4.5.

4.4.5 The Second and Third Isomorphism Theorems

Two more identifications come from the same method: write down a surjective map, compute its kernel, and apply the First Isomorphism Theorem. The first compares a sum of subspaces with an intersection.

Theorem 4.4.8. Second Isomorphism Theorem

Let U and W be subspaces of a vector space V. Then $U \cap W$ is a subspace of U, W is a

subspace of $U + W$, and

$$U/(U \cap W) \to (U + W)/W, \qquad u + (U \cap W) \mapsto u + W$$

is a well-defined isomorphism. Hence $(U + W)/W \cong U/(U \cap W)$.

Idea. Picture two planes U and W in $\mathbb{R}^3$ meeting in a line. Modulo W, a vector of $U + W$ is represented by its U-part alone, since the W-part is forgotten. The only vectors of U that become zero are those already in W. So the map "$u \mapsto u + W$" from U is surjective, with kernel $U \cap W$.

Proof. $U \cap W$ is a subspace (Theorem 1.2.2) contained in U, and W is a subspace contained in $U + W$ (Theorem 1.7.1), so both quotients exist. Let $\pi_W : U + W \to (U + W)/W$ be the quotient map, and let $S : U \to (U + W)/W$, $Su = u + W$, be its restriction to U; S is linear because π_W is (Theorem 4.4.4).

S is surjective. An element of $(U + W)/W$ is $(u + w) + W$ with $u \in U$, $w \in W$. Since $(u + w) - u = w \in W$, Lemma 4.4.3 (b) gives $(u + w) + W = u + W = Su$.

$\ker S = U \cap W$. For $u \in U$, $Su = W$ means $u \in W$ (Lemma 4.4.3 (b)), that is, $u \in U \cap W$.

By Theorem 4.4.7 applied to S, the rule $u + (U \cap W) \mapsto Su = u + W$ is a well-defined injective linear map with image $\operatorname{im} S = (U + W)/W$, hence an isomorphism. (Directly: if $u - u' \in U \cap W$, then $u - u' \in W$, so $u + W = u' + W$, and the value does not depend on the representative.)

■

If U and W are finite-dimensional, then so is $U + W$ (it is spanned by the union of bases, Proposition 1.7.2), and Theorem 4.4.5 turns the theorem into $\dim(U + W) - \dim W = \dim U - \dim(U \cap W)$. That is the dimension formula for sums (Theorem 1.7.3), now as a consequence of an isomorphism rather than of a basis computation. An exercise below asks for the details.

The third theorem says that quotients can be taken in stages.

Theorem 4.4.9. Third Isomorphism Theorem

Let $U \subseteq W \subseteq V$ be subspaces. Then $W/U = \{w + U : w \in W\}$ is a subspace of V/U, and

$$(V/U)/(W/U) \to V/W, \qquad (v + U) + W/U \mapsto v + W$$

is a well-defined isomorphism. Hence $(V/U)/(W/U) \cong V/W$.

Proof. Let $S : V/U \to V/W$, $S(v + U) = v + W$.

Well defined. If $v + U = v' + U$, then $v - v' \in U \subseteq W$, so $v + W = v' + W$ by Lemma 4.4.3 (b). By Theorem 0.5.3 (a), applied to $V \to V/W$, $v \mapsto v + W$, the rule defines a function on V/U.

Linear and surjective.

$$\begin{aligned} S((v+U)+(w+U)) &= S((v+w)+U) = (v+w)+W \\ &= S(v+U)+S(w+U), \end{aligned}$$

and similarly for scalars, by Definition 4.4.3 in both quotients. Every $v+W$ equals $S(v+U)$.

Kernel. $S(v+U) = W$ means $v \in W$. If $v \in W$, then $v+U \in W/U$. Conversely, if $v+U = w+U$ with $w \in W$, then $v-w \in U \subseteq W$, so $v \in W$. Hence $\ker S = W/U$. In particular W/U is a subspace of V/U, by Theorem 2.2.1.

By Theorem 4.4.7 applied to S, $(V/U)/(W/U) \cong \operatorname{im} S = V/W$, via $(v+U)+W/U \mapsto S(v+U) = v+W$.

■

The notation W/U is used with two meanings, and they agree: the quotient space of W by U, and the set of cosets $w+U$ with $w \in W$, sitting inside V/U. The coset $w+U$ is the same subset of V in both readings, and the operations agree because both are computed with representatives. For example, with $V = \mathbb{R}^3$, U the z-axis, and W the xz-plane, $V/W \cong \mathbb{R}$ through the y-coordinate, and the theorem says that first forgetting z and then forgetting x leaves the same thing as forgetting both at once.

4.4.6 The universal property of the quotient

The proofs of the three theorems repeated one move: a linear map that vanishes on U passes to V/U. We record that move once, in its general form. It is the standard way to **define** linear maps out of a quotient.

Theorem 4.4.10. Universal Property of the Quotient

Let U be a subspace of V, $\pi\colon V \to V/U$ the quotient map, and $T\colon V \to W$ a linear map.

(a) If $U \subseteq \ker T$, then there is **exactly one** linear map $\bar{T}\colon V/U \to W$ with $\bar{T} \circ \pi = T$, namely $\bar{T}(v+U) = Tv$. Its image is $\operatorname{im} T$, and it is injective if and only if $U = \ker T$.

(b) Conversely, if some function $\bar{T}\colon V/U \to W$ satisfies $\bar{T} \circ \pi = T$, then $U \subseteq \ker T$.

Proof.

(a) If $v-v' \in U \subseteq \ker T$, then $Tv = Tv'$. By Theorem 0.5.3 (a) there is exactly one function $\bar{T}$ with $\bar{T}(v+U) = Tv$, and it is linear by the computation in the proof of Theorem 4.4.7, which used only this formula. Any linear $\bar{T}$ with $\bar{T} \circ \pi = T$ satisfies this formula, so it is this one. The image is $\{Tv\} = \operatorname{im} T$. The kernel of $\bar{T}$ is $\{v+U : v \in \ker T\}$; it is $\{U\}$, the zero subspace, exactly when every $v \in \ker T$ lies in U, that is, when $\ker T \subseteq U$, which together with $U \subseteq \ker T$ means $U = \ker T$. Apply Theorem 2.2.3.

(b) Let $u \in U$. Then $\pi(u) = U = \pi(0)$, so $Tu = \bar{T}(\pi(u)) = \bar{T}(\pi(0)) = T(0) = 0$ by Theorem 2.1.1. Hence $U \subseteq \ker T$.

■

In words: **linear maps out of V/U are the same as linear maps out of V that kill U.**

The First Isomorphism Theorem is the case $U = \ker T$. For example, the map $T\colon \mathbb{R}^3 \to \mathbb{R}$, $T(x,y,z) = x$, kills the z-axis U, so it induces $\bar{T}\colon \mathbb{R}^3/U \to \mathbb{R}$, $(x,y,z)+U \mapsto x$. Since $\ker T$ is the yz-plane, strictly larger than U, this $\bar{T}$ is not injective: it also kills the coset $\mathbf{e}_2 + U$.

Quotients will return in Chapter 9, where an operator T with $T(U) \subseteq U$ induces an operator on V/U, and this is the engine of induction arguments on the dimension.

4.4.7 Exercises

4.4.7.1 A. Check your understanding

Exercise. A1

(a) Define the quotient space V/U and its two operations.

(b) State the condition, in terms of $v - w$, for $v + U = w + U$.

(c) True or false: V/U is a subspace of V. Justify your answer.

(d) State the First Isomorphism Theorem.

(e) If U is a 2-dimensional subspace of $\mathbb{R}^5$, what is $\dim(\mathbb{R}^5/U)$? What is $\dim(\mathbb{R}^5 \times U)$?

(f) What must be checked before the formula $(v+U)+(w+U) = (v+w)+U$ can be used as a definition, and which property of U makes the check work?

Solution.

(a) For a subspace U of V, $V/U = \{v+U : v \in V\}$, with $(v+U)+(w+U) = (v+w)+U$ and $a(v+U) = av+U$ (Definition 4.4.3).

(b) $v+U = w+U$ if and only if $v - w \in U$ (Lemma 4.4.3).

(c) False. The elements of V/U are cosets, which are subsets of V, not vectors of V.

(d) If $T\colon V \to W$ is linear, then $v + \ker T \mapsto Tv$ is a well-defined isomorphism $V/\ker T \to \operatorname{im} T$ (Theorem 4.4.7).

(e) $5 - 2 = 3$ by Theorem 4.4.5, and $5 + 2 = 7$ by Theorem 4.4.1.

(f) That the right-hand side does not depend on the representatives: if $v - v' \in U$ and $w - w' \in U$, then $(v+w)+U = (v'+w')+U$. This works because $(v+w)-(v'+w') = (v-v')+(w-w')$ and U is closed under addition.

4.4.7.2 B. Practice

Exercise. B1: A basis of a quotient of $\mathbb{R}^4$

Let $U = \operatorname{span}((1,0,1,0),(0,1,0,1)) \subseteq \mathbb{R}^4$.

(a) Find $\dim(\mathbb{R}^4/U)$ and a basis of $\mathbb{R}^4/U$.

(b) Write $(3,1,4,1)+U$ in your basis.

(c) Decide whether $(1,2,3,4)+U = (3,4,1,2)+U$.

Solution.

(a) The two spanning vectors are independent (look at the first two entries), so $\dim U = 2$

and $\dim(\mathbb{R}^4/U) = 4 - 2 = 2$ by Theorem 4.4.5. Extend by $\mathbf{e}_1, \mathbf{e}_2$. If $a(1,0,1,0) + b(0,1,0,1) + c\mathbf{e}_1 + d\mathbf{e}_2 = (a+c, b+d, a, b) = \mathbf{0}$, then $a = b = 0$ from the last two entries and $c = d = 0$ from the first two. So the four vectors are independent, hence a basis of $\mathbb{R}^4$ (Theorem 1.6.4), and by Theorem 4.4.5 $(\mathbf{e}_1 + U, \mathbf{e}_2 + U)$ is a basis of $\mathbb{R}^4/U$.

(b) Solve $(3,1,4,1) = (a+c, b+d, a, b)$: $a = 4$, $b = 1$, $c = -1$, $d = 0$. Since the U-part becomes zero in the quotient, $(3,1,4,1) + U = -1(\mathbf{e}_1 + U) + 0(\mathbf{e}_2 + U)$. Check: $(3,1,4,1) + \mathbf{e}_1 = (4,1,4,1) = 4(1,0,1,0) + (0,1,0,1) \in U$.

(c) The difference is $(1,2,3,4) - (3,4,1,2) = (-2,-2,2,2)$. A vector (a,b,a,b) of U has equal first and third entries, but $-2 \neq 2$. So the difference is not in U, and the cosets are different by Lemma 4.4.3.

Exercise. B2: A map that factors through a quotient

Let $T\colon \mathbb{R}[x]_{\leq 2} \to \mathbb{R}^2$, $T(p) = (p(1), p'(1))$, and $U = \operatorname{span}((x-1)^2)$.

(a) Show that $U \subseteq \ker T$, so that T induces a linear map $\bar{T}\colon \mathbb{R}[x]_{\leq 2}/U \to \mathbb{R}^2$ with $\bar{T}(p + U) = T(p)$.

(b) Prove that $\bar{T}$ is an isomorphism.

(c) Compute $\bar{T}(x^2 + U)$, and find the coset $\bar{T}^{-1}(0,1)$.

Solution.

(a) T is linear, since evaluation and differentiation are. For $q = (x-1)^2$, $q(1) = 0$ and $q'(1) = 2(1-1) = 0$, so $T(q) = \mathbf{0}$, and hence $T(cq) = \mathbf{0}$ for all c. So $U \subseteq \ker T$, and Theorem 4.4.10 (a) gives the linear map $\bar{T}$.

(b) $T(1) = (1,0)$ and $T(x-1) = (0,1)$, so T is surjective and $\operatorname{rank} T = 2$. By Theorem 2.3.1, $\dim \ker T = 3 - 2 = 1 = \dim U$, so $\ker T = U$ by Theorem 1.6.8. By Theorem 4.4.10 (a), $\bar{T}$ is injective with image $\operatorname{im} T = \mathbb{R}^2$. Hence $\bar{T}$ is a linear bijection, an isomorphism.

(c) $\bar{T}(x^2 + U) = T(x^2) = (1,2)$. Since $T(x-1) = (0,1)$, $\bar{T}^{-1}(0,1) = (x-1) + U$, the set of all $x - 1 + c(x-1)^2$.

Exercise. B3: Matrices modulo scalars

Let $n \geq 1$ and $S = \{c\mathbf{I}_n : c \in F\} \subseteq M_n(F)$.

(a) Show that S is a subspace and find $\dim(M_n(F)/S)$.

(b) For $n = 2$, find a basis of $M_2(F)/S$.

(c) Show that $\mathbf{A} + S = \mathbf{B} + S$ if and only if $\mathbf{A} - \mathbf{B}$ is a scalar matrix, and decide whether $\begin{pmatrix} 3 & 1 \\ 0 & 5 \end{pmatrix} + S = \begin{pmatrix} 1 & 1 \\ 0 & 3 \end{pmatrix} + S$.

Solution.

(a) $S = \operatorname{span}(\mathbf{I}_n)$ is a subspace by Theorem 1.3.1, of dimension 1 since $\mathbf{I}_n \neq \mathbf{O}$. By Theorem 4.4.5, $\dim(M_n(F)/S) = n^2 - 1$.

(b) Extend $(\mathbf{I}_2)$ by $\mathbf{E}_{11}, \mathbf{E}_{12}, \mathbf{E}_{21}$. If $a\mathbf{I}_2 + b\mathbf{E}_{11} + c\mathbf{E}_{12} + d\mathbf{E}_{21} = \begin{pmatrix} a+b & c \\ d & a \end{pmatrix} = \mathbf{O}$, then $a = c = d = 0$ and then $b = 0$. So the four matrices are independent, hence a basis of $M_2(F)$ (Theorem 1.6.4), and $(\mathbf{E}_{11} + S, \mathbf{E}_{12} + S, \mathbf{E}_{21} + S)$ is a basis of $M_2(F)/S$ by Theorem 4.4.5.

(c) This is Lemma 4.4.3 (b) for the subspace S. Here the difference is $\begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} = 2\mathbf{I}_2 \in S$, so the two cosets are equal.

4.4.7.3 C. Going deeper

Exercise. C1: The dimension formula from isomorphisms

Let U and W be finite-dimensional subspaces of a vector space V.

(a) Using Theorem 4.4.8 and Theorem 4.4.5, prove that $\dim(U + W) = \dim U + \dim W - \dim(U \cap W)$.

(b) Give a second proof: apply Rank–Nullity to the addition map $S\colon U \times W \to V$, $S(u, w) = u + w$, after showing that $u \mapsto (u, -u)$ is an isomorphism from $U \cap W$ onto $\ker S$.

Solution.

(a) $U + W$ is finite-dimensional, being spanned by a basis of U followed by a basis of W (Proposition 1.7.2), and $U \cap W \subseteq U$ is finite-dimensional (Theorem 1.6.7). By Theorem 4.4.8, $(U + W)/W \cong U/(U \cap W)$, and isomorphic finite-dimensional spaces have equal dimension (Theorem 2.5.4). Applying Theorem 4.4.5 to both sides,

$$\dim(U + W) - \dim W = \dim U - \dim(U \cap W),$$

which rearranges to the formula.

(b) The map S is linear with image $U + W$ by Theorem 4.4.2 (a). Let $J\colon U \cap W \to U \times W$, $Ju = (u, -u)$; it is well defined because $u \in U$ and $-u \in W$, and linear because the operations are entrywise. It is injective, since $(u, -u) = (0, 0)$ forces $u = 0$. Its image lies in $\ker S$, since $u + (-u) = 0$. Conversely, if $(u, w) \in \ker S$, then $w = -u$; so $u = -w \in W$, hence $u \in U \cap W$ and $(u, w) = Ju$. Thus J is an isomorphism onto $\ker S$, and $\dim \ker S = \dim(U \cap W)$. By Theorem 4.4.1 and Theorem 2.3.1,

$$\dim U + \dim W = \dim(U \times W) = \dim(U + W) + \dim(U \cap W).$$

Exercise. C2: The operator induced on a quotient

Let V be finite-dimensional, $T \in \mathcal{L}(V)$, and U a subspace with $Tu \in U$ for every $u \in U$.

(a) Prove that $\bar{T}(v + U) \coloneqq Tv + U$ is a well-defined linear operator on V/U.

(b) Let $(u_1, \dots, u_m)$ be a basis of U, extended to a basis $\mathcal{B} = (u_1, \dots, u_m, w_1, \dots, w_s)$ of V. Show that

$$[T]^{\mathcal{B}}_{\mathcal{B}} = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{O} & D \end{pmatrix},$$

where $\mathbf{A} \in M_m(F)$ is the matrix of the restriction $T|_U \colon U \to U$ in $(u_1, \dots, u_m)$, $D \in M_s(F)$ is the matrix of $\bar{T}$ in $(w_1 + U, \dots, w_s + U)$, and $\mathbf{O}$ is the $s \times m$ zero matrix.

(c) Check (b) when T is differentiation on $\mathbb{R}[x]_{\le 2}$, with $U = \mathbb{R}[x]_{\le 1}$ and $\mathcal{B} = (1, x, x^2)$.

Hint: for (a), apply Theorem 4.4.10 to $\pi \circ T$.

Solution.

(a) The map $\pi \circ T \colon V \to V/U$, $v \mapsto Tv + U$, is linear as a composition of linear maps (Theorem 2.4.2). For $u \in U$, $Tu \in U$, so $(\pi \circ T)(u) = U$, the zero of V/U; thus $U \subseteq \ker(\pi \circ T)$. By Theorem 4.4.10 (a) there is exactly one linear map $\bar{T} \colon V/U \to V/U$ with $\bar{T}(v + U) = Tv + U$.

(b) For $j \le m$, $Tu_j \in U$, so it is a combination of $u_1, \dots, u_m$ only, and its $\mathcal{B}$-coordinates are 0 in the last s places. These first m columns, restricted to their first m entries, are the columns of the matrix $\mathbf{A}$ of $T|_U$ (which maps U to U by hypothesis and is linear). For the remaining columns, write $Tw_k = \sum_i b_{ik} u_i + \sum_l d_{lk} w_l$. Passing to V/U, where $u_i + U = U$ is zero, gives $\bar{T}(w_k + U) = \sum_l d_{lk}(w_l + U)$. So the $s \times s$ matrix (d_{lk}) is the matrix D of $\bar{T}$ in the basis $(w_1 + U, \dots, w_s + U)$ of Theorem 4.4.5, and the bottom-right block of $[T]_{\mathcal{B}}^{\mathcal{B}}$ is D.

(c) $T(1) = 0$, $T(x) = 1$, $T(x^2) = 2x$, so $[T]_{\mathcal{B}}^{\mathcal{B}} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \end{pmatrix}$. Differentiation maps $U = \mathbb{R}[x]_{\le 1}$ into itself. The block $\mathbf{A} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$ is the matrix of T on U in $(1, x)$, the lower-left block is $\begin{pmatrix} 0 & 0 \end{pmatrix}$, and the block D is the 1×1 matrix $\mathbf{O}_1$: indeed $\bar{T}(x^2 + U) = 2x + U = U$, the zero coset, since $2x \in U$.

Exercise. C3: Counting cosets over a finite field

Let $F = \mathbb{F}_q$ be a finite field with q elements, V a vector space over F of dimension n, and U a subspace of dimension m.

(a) Show that $|V| = q^n$.

(b) Show that every coset $v + U$ has exactly $|U|$ elements, and deduce $|V/U| = q^{n-m}$ by counting.

(c) Check that (b) agrees with Theorem 4.4.5.

Solution.

(a) Choose a basis $(v_1, \dots, v_n)$. By Theorem 1.5.1, $(a_1, \dots, a_n) \mapsto a_1 v_1 + \cdots + a_n v_n$ is a bijection $F^n \to V$, and F^n has q^n elements.

(b) The map $U \to v + U$, $u \mapsto v + u$, is surjective by Definition 4.4.2 and injective because $v + u = v + u'$ implies $u = u'$. So $|v + U| = |U| = q^m$ by (a) applied to U. By Lemma 4.4.3 (c), the cosets partition V into $|V/U|$ disjoint sets of size q^m, so $q^n = |V/U| \cdot q^m$, and $|V/U| = q^{n-m}$.

(c) By Theorem 4.4.5, V/U is a vector space over $\mathbb{F}_q$ of dimension $n - m$, so by (a) it has q^{n-m} elements, in agreement with (b). For example, $\mathbb{F}_2^3$ modulo a line has $8/2 = 4$

cosets.

4.5 Projections and the Trace of an Operator

Chapter 1 ended its discussion of complements with a promise: a splitting $V = U \oplus W$ should give a projection of V onto U along W, and different complements should give different projections. This section keeps that promise. Projections are the operators that are simplest in a well-chosen basis, and computing their matrices there leads to a second topic: the trace of an operator, a number attached to $T \in \mathcal{L}(V)$ that no choice of basis can change. We end with involutions, which split a space into two pieces in the same way.

4.5.1 Projections

Suppose $V = U \oplus W$. By Definition 1.7.2, every $v \in V$ can be written as $v = u + w$ with $u \in U$ and $w \in W$ in **exactly one** way. So the rule "$v \mapsto u$", keep the U-piece and throw away the W-piece, is a well-defined function $V \to V$. In $\mathbb{R}^3 = \{\text{the } xy\text{-plane}\} \oplus \{\text{the } z\text{-axis}\}$, it sends (a, b, c) to $(a, b, 0)$: the shadow of a point on the floor when light falls straight down.

What property does such a map have that singles it out? Apply it twice. The output u already lies in U, and its splitting is $u = u + 0$, so applying the rule again returns u. Doing it twice is the same as doing it once. We take this as the definition, because it can be checked without knowing U and W in advance.

A projection is an operator that leaves alone everything it has already produced.

Definition 4.5.1. Projection

Let V be a vector space over F. An operator $P \in \mathcal{L}(V)$ is a **projection** if

$$P^2 = P,$$

that is, $P(Pv) = Pv$ **for every** $v \in V$.

In words: P^2 is the composite PP (Definition 2.4.2), so the condition says that P fixes each of its own outputs. Equivalently, $Px = x$ **for every** $x \in \operatorname{im} P$: if $x = Pv$, then $Px = P^2(v) = Pv = x$, and conversely this for $x = Pv$ is $P^2(v) = Pv$. No inner product and no notion of angle is involved. The projections of this section are often called **oblique** projections, to distinguish them from the orthogonal projections met later, in the chapters on inner product spaces.

Examples.

- **Onto an axis.** $P(x, y) = (x, 0)$ on F^2. Then $P^2(x, y) = P(x, 0) = (x, 0)$, so $P^2 = P$.
- **A rank-one projection.** $T(x, y) = (2x - y,\ 2x - y)$ on $\mathbb{R}^2$ from Example (Making a matrix diagonal with a well-chosen basis). We saw there that its matrix $\mathbf{A}$ satisfies $\mathbf{A}^2 = \mathbf{A}$; directly, $T(2x - y, 2x - y) = (4x - 2y - 2x + y, \dots) = (2x - y, 2x - y)$.
- **Constant term.** On $F[x]_{\le n}$, $P(p) = p(0)$, the constant polynomial with value $p(0)$. The constant polynomial c has value c at 0, so $P(P(p)) = P(p)$.
- **Degenerate cases.** The identity id_V and the zero operator are projections, since $\mathrm{id}_V^2 = \mathrm{id}_V$ and $0^2 = 0$. They correspond to keeping everything and keeping nothing.

Nonexample by minimal change. Change the first example to $T(x, y) = (2x, 0)$ on $\mathbb{R}^2$. Its image is still the x-axis and its kernel is still the y-axis. But $T^2(x, y) = (4x, 0) \neq T(x, y)$ when $x \neq 0$: T does not fix the vectors of its own image, since $T(1, 0) = (2, 0)$. So having "the right image and kernel" is not enough; the clause $P^2 = P$ is what forces P to be the identity on its image. Likewise the reflection $R(x, y) = (y, x)$ is not a projection: $R^2 = \mathrm{id} \neq R$.

Here is the promised correspondence. It says that projections and direct sum decompositions are two descriptions of the same thing.

Theorem 4.5.1. Projections and Direct Sums

Let V be a vector space.

(a) If $P \in \mathcal{L}(V)$ is a projection, then

$$V = \operatorname{im} P \oplus \ker P,$$

P is the identity on $\operatorname{im} P$, and P is zero on $\ker P$.

(b) Conversely, suppose $V = U \oplus W$. Then there is **exactly one** operator $P \in \mathcal{L}(V)$ with $Pu = u$ for all $u \in U$ and $Pw = 0$ for all $w \in W$, namely $P(u + w) = u$. It is a projection, with $\operatorname{im} P = U$ and $\ker P = W$. It is called the **projection onto U along W**.

Idea. For (a), every vector splits as "what P keeps" plus "what P removes": $v = Pv + (v - Pv)$. The first piece is in the image by definition, and $P^2 = P$ is exactly what puts the second piece in the kernel. For (b), uniqueness of the splitting makes the rule a function, and uniqueness again makes it linear.

Proof.

(a) Let $v \in V$. Then $v = Pv + (v - Pv)$, where $Pv \in \operatorname{im} P$ and $P(v - Pv) = Pv - P^2(v) = 0$ because $P^2 = P$. Hence $V = \operatorname{im} P + \ker P$. If $x \in \operatorname{im} P \cap \ker P$, then $x = Px$, as shown after Definition 4.5.1, and $Px = 0$; so $x = 0$. By Theorem 1.7.4, $V = \operatorname{im} P \oplus \ker P$. We have also seen that P fixes $\operatorname{im} P$, and P is zero on $\ker P$ by definition.

(b) *Existence.* By Definition 1.7.2, each $v \in V$ has exactly one splitting $v = u + w$, so $Pv := u$ defines a function $V \to V$. It is linear: if $v = u + w$ and $v' = u' + w'$, then $v + v' = (u + u') + (w + w')$ and $av = au + aw$ are splittings, because U and W are subspaces, so by uniqueness $P(v + v') = u + u' = Pv + Pv'$ and $P(av) = au = aPv$. For $u \in U$ the splitting is $u = u + 0$, so $Pu = u$; for $w \in W$ it is $w = 0 + w$, so $Pw = 0$.

Properties. Every value Pv lies in U, and every $u \in U$ is Pu; so $\operatorname{im} P = U$. Then $P(Pv) = Pv$ because P fixes U; so $P^2 = P$. Finally $P(u + w) = u$ is 0 exactly when $u = 0$, that is, when $v = w \in W$; so $\ker P = W$.

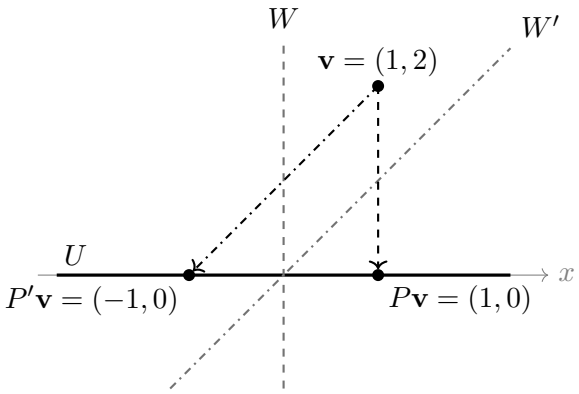
Uniqueness. If $P' \in \mathcal{L}(V)$ is identity on U and zero on W, then $P'(u + w) = P'u + P'w = u$ by linearity, so $P' = P$.

■

So a projection **is** a direct sum decomposition $V = U \oplus W$, recorded as an operator: U is its

image and W its kernel. Given P, the decomposition is $\operatorname{im} P \oplus \ker P$; given the decomposition, P is the projection onto U along W. The two constructions undo each other, by the "exactly one" in (b).

Now the promise from Chapter 1. There, the x-axis U in $\mathbb{R}^2$ had many complements, for instance the y-axis W and the line W' spanned by $(1,1)$. Each gives a projection onto the same line U.



To project $\mathbf{v}$ onto U along a complement, slide $\mathbf{v}$ parallel to that complement until it hits U. Along the y-axis, $(1,2) = (1,0) + (0,2)$ gives $P\mathbf{v} = (1,0)$ (dashed arrow). Along W', $(1,2) = (-1,0) + (2,2)$ gives $P'\mathbf{v} = (-1,0)$ (dash-dotted arrow). In general $P(x,y) = (x,0)$ and $P'(x,y) = (x-y,0)$, with standard matrices $\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$ and $\begin{pmatrix} 1 & -1 \\ 0 & 0 \end{pmatrix}$.

Warning. **"Projection onto U" is incomplete: the direction "along W" matters.** The two operators above are both projections, both have image U, and they are different: $P(1,2) = (1,0)$ but $P'(1,2) = (-1,0)$. Since U has infinitely many complements in $\mathbb{R}^2$, it is the image of infinitely many projections. A projection is determined by its image **and** its kernel together.

In finite dimension, the direct sum $V = \operatorname{im} P \oplus \ker P$ hands us a basis in which P is as simple as possible.

Corollary 4.5.2. The matrix of a projection

Let V be finite-dimensional and $P \in \mathcal{L}(V)$ a projection with $\operatorname{rank} P = r$. Let $(u_1, \dots, u_r)$ be a basis of $\operatorname{im} P$ and $(w_1, \dots, w_s)$ a basis of $\ker P$. Then $\mathcal{B} = (u_1, \dots, u_r, w_1, \dots, w_s)$ is a basis of V, and

$$[P]_{\mathcal{B}}^{\mathcal{B}} = \begin{pmatrix} \mathbf{I}_r & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{pmatrix} = \operatorname{diag}(1, \dots, 1, 0, \dots, 0),$$

with r ones.

Proof. By Theorem 4.5.1 (a), $V = \operatorname{im} P \oplus \ker P$, so the concatenated list $\mathcal{B}$ is a basis of V by Theorem 1.7.5 ((a) $\Rightarrow$ (d)). Since $Pu_i = u_i$ and $Pw_j = 0$, the columns of $[P]_{\mathcal{B}}^{\mathcal{B}}$ are $\mathbf{e}_1, \dots, \mathbf{e}_r$ followed by s zero columns (Definition 4.1.1).

■

Conversely, an operator whose matrix in some basis is $\mathrm{diag}(1,\dots,1,0,\dots,0)$ is a projection, since that matrix squares to itself. With the change-of-basis square, this gives a way to write down any projection in standard coordinates.

Example. Projecting onto a plane along a line

In $\mathbb{R}^3$, let $U = \{(x,y,z) : x+y+z = 0\}$ and $L = \mathrm{span}((1,1,1))$, so that $\mathbb{R}^3 = U \oplus L$ by Example (A plane and a line in $\mathbb{R}^3$). Find the standard matrix of the projection P onto U along L, and check that it squares to itself.

Solution. Direct route. By Example (A plane and a line in $\mathbb{R}^3$), (a,b,c) splits as $(a-m, b-m, c-m) + m(1,1,1)$ with $m = \frac{a+b+c}{3}$, and the first piece lies in U. So

$$P(a,b,c) = \left(\tfrac{2a-b-c}{3},\ \tfrac{-a+2b-c}{3},\ \tfrac{-a-b+2c}{3}\right), \qquad [P]^{\mathcal{E}}_{\mathcal{E}} = \frac{1}{3}\begin{pmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{pmatrix}.$$

Adapted basis. $(1,-1,0)$ and $(0,-1,1)$ lie in U and are independent (look at the first and third entries), so they form a basis of the 2-dimensional plane U. With $\mathcal{B} = ((1,-1,0),(0,-1,1),(1,1,1))$, Corollary 4.5.2 gives $[P]^{\mathcal{B}}_{\mathcal{B}} = \mathrm{diag}(1,1,0)$, and by Theorem 4.2.3 $[P]^{\mathcal{E}}_{\mathcal{E}} = M\,\mathrm{diag}(1,1,0)\,M^{-1}$ with $M = [\mathrm{id}]^{\mathcal{E}}_{\mathcal{B}}$, whose columns are the vectors of $\mathcal{B}$. Multiplying out gives the same matrix as the direct route.

Check. Let $\mathbf{A}$ be the matrix found. The first column of $\mathbf{A}^2$ is $\mathbf{A}$ applied to $\frac{1}{3}(2,-1,-1)$, which is $\frac{1}{9}(4+1+1,\ -2-2+1,\ -2+1-2) = \frac{1}{3}(2,-1,-1)$, the first column of $\mathbf{A}$; the other columns work the same way by symmetry of the entries. Also $\mathbf{A}(1,1,1) = \mathbf{0}$ and $\mathbf{A}(1,-1,0) = (1,-1,0)$, as a projection onto U along L must satisfy. The division by 3 needs $3 \neq 0$ in the field, which holds over $\mathbb{R}$.

Quick check. Is $P = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}$ (acting on $\mathbb{R}^2$) a projection? If so, onto which line, and along which line?

Solution. $P^2 = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix} = P$, so yes. Its image is spanned by its columns, $\mathrm{span}((1,0))$, the x-axis. Its kernel is $\{x+y=0\} = \mathrm{span}((1,-1))$. By Theorem 4.5.1, it is the projection onto the x-axis along the line $y = -x$.

4.5.2 The trace of an operator

For a projection of rank r, the diagonal matrix of Corollary 4.5.2 has trace $1+\cdots+1 = r\cdot 1$. But traces are defined for matrices, and the same projection has many other matrices, such as the 3×3 matrix of Example (Projecting onto a plane along a line), whose trace is $\frac{2+2+2}{3} = 2$. The agreement is no accident. In Theorem 4.2.5 we proved that similar matrices have the same trace, and by Theorem 4.2.4 all matrices of one operator are similar. So the trace belongs to

the operator.

The trace of an operator is the trace of its matrix in any basis; the basis does not matter.

Definition 4.5.2. Trace of an operator

Let V be a finite-dimensional vector space over F with $\dim V \geq 1$, and let $T \in \mathcal{L}(V)$. The **trace** of T is

$$\operatorname{tr} T := \operatorname{tr}[T]_{\mathcal{B}}^{\mathcal{B}},$$

where $\mathcal{B}$ is **any** basis of V. (For $V = \{0\}$ we set $\operatorname{tr} T = 0$.)

In words: to find $\operatorname{tr} T$, choose a basis, write down the matrix of T with that **same** basis on input and output, and add the diagonal entries.

Well-definedness. The definition names a basis but claims the answer does not depend on it. Let $\mathcal{B}$ and $\mathcal{C}$ be bases of V. By Theorem 4.2.3, $[T]_{\mathcal{C}}^{\mathcal{C}} = P^{-1}[T]_{\mathcal{B}}^{\mathcal{B}}P$ with $P = [\mathrm{id}]_{\mathcal{C}}^{\mathcal{B}}$ invertible, so the two matrices are similar, and by Theorem 4.2.5 they have the same trace. Hence $\operatorname{tr} T$ depends only on T.

Examples.

- **Identity.** In every basis, $[\mathrm{id}_V]_{\mathcal{B}}^{\mathcal{B}} = \mathbf{I}_n$, so $\operatorname{tr} \mathrm{id}_V = n \cdot 1$, where $n = \dim V$. Over $\mathbb{R}$ this is n; over $\mathbb{F}_2$ it is 0 when n is even.
- **Differentiation.** On $F[x]_{\leq n}$ with the basis $(1, x, \ldots, x^n)$, $D(x^k) = kx^{k-1}$ has no x^k-component, so every diagonal entry of the matrix is 0 and $\operatorname{tr} D = 0$.
- **A shift of the variable.** On $\mathbb{R}[x]_{\leq 2}$, let $T(p) = p(x+1)$. In $(1, x, x^2)$, $T(1) = 1$, $T(x) = 1 + x$, $T(x^2) = 1 + 2x + x^2$, so the matrix is $\begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{pmatrix}$ and $\operatorname{tr} T = 3$.
- **The reflection.** $R(x, y) = (y, x)$ has trace 0, whether computed from $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ or from $\operatorname{diag}(1, -1)$ (Example (The reflection in a mirror-adapted basis)).

Warning. **The trace needs the same basis on both sides.** For an operator $T \in \mathcal{L}(V)$, the matrix $[T]_{\mathcal{B}}^{\mathcal{C}}$ with $\mathcal{B} \neq \mathcal{C}$ is square, but its trace depends on the bases. For id on $\mathbb{R}^2$, with $\mathcal{E}$ the standard basis and $\mathcal{C} = ((2,0),(0,1))$, we get $[\mathrm{id}]_{\mathcal{E}}^{\mathcal{C}} = \operatorname{diag}(\frac{1}{2}, 1)$, of trace $\frac{3}{2}$, while $\operatorname{tr} \mathrm{id} = 2$. Different bases on input and output give equivalent matrices, not similar ones, and equivalence does not preserve the trace. For the same reason there is no trace of a map $V \to W$ between different spaces.

The rules for the trace of a matrix carry over at once.

Theorem 4.5.3. Properties of the Trace of an Operator

Let V be finite-dimensional with $n = \dim V$, let $S, T \in \mathcal{L}(V)$ and $c \in F$. Then

$$\begin{aligned} \operatorname{tr}(S+T) &= \operatorname{tr} S + \operatorname{tr} T, \\ \operatorname{tr}(cT) &= c \operatorname{tr} T, \\ \operatorname{tr}(ST) &= \operatorname{tr}(TS), \\ \operatorname{tr} \mathrm{id}_V &= n \cdot 1. \end{aligned}$$

Proof. For $n = 0$ all traces are 0. Otherwise fix a basis $\mathcal{B}$ of V. The map $T \mapsto [T]_{\mathcal{B}}^{\mathcal{B}}$ is linear by Theorem 4.1.4, so $[S+T]_{\mathcal{B}}^{\mathcal{B}} = [S]_{\mathcal{B}}^{\mathcal{B}} + [T]_{\mathcal{B}}^{\mathcal{B}}$ and $[cT]_{\mathcal{B}}^{\mathcal{B}} = c[T]_{\mathcal{B}}^{\mathcal{B}}$; and $[ST]_{\mathcal{B}}^{\mathcal{B}} = [S]_{\mathcal{B}}^{\mathcal{B}}[T]_{\mathcal{B}}^{\mathcal{B}}$ by Theorem 4.1.2, and likewise for TS. The first three identities now follow from Theorem 0.9.6, since $\operatorname{tr} T$ may be computed in $\mathcal{B}$. The last was computed in the examples. ∎

A first consequence: on a finite-dimensional space V of dimension $n \geq 1$ over a field in which $n \cdot 1 \neq 0$, such as $\mathbb{R}$, there are no operators with $ST - TS = \mathrm{id}_V$, since the left side has trace 0 and the right side has trace $n \cdot 1$.

Now the trace of a projection. Over $\mathbb{R}$ it is simply the rank, but the precise statement is an equation in the field, and reading it as an equation of integers needs a hypothesis.

Theorem 4.5.4. Trace of a Projection

Let V be a finite-dimensional vector space over F and $P \in \mathcal{L}(V)$ a projection. Then

$$\operatorname{tr} P = (\operatorname{rank} P) \cdot 1 \quad \text{in } F.$$

If F has characteristic 0, or characteristic $p > \dim V$, then $\operatorname{rank} P$ is determined by $\operatorname{tr} P$: it is the unique integer r with $0 \leq r \leq \dim V$ and $r \cdot 1 = \operatorname{tr} P$. In particular $\operatorname{rank} P = \operatorname{tr} P$ for projections on real or complex spaces.

Proof. Let $r = \operatorname{rank} P$. If $V = \{0\}$, both sides are 0. Otherwise, by Corollary 4.5.2 there is a basis $\mathcal{B}$ with $[P]_{\mathcal{B}}^{\mathcal{B}} = \operatorname{diag}(1, \ldots, 1, 0, \ldots, 0)$ with r ones, so $\operatorname{tr} P = r \cdot 1$ by Definition 4.5.2.

For the second statement, let $n = \dim V$. Since $0 \leq r \leq n$ and $r \cdot 1 = \operatorname{tr} P$, it remains to show uniqueness. Suppose $k \cdot 1 = l \cdot 1$ for integers $0 \leq l \leq k \leq n$. By Lemma 0.7.4, $(k-l) \cdot 1 + l \cdot 1 = k \cdot 1 = l \cdot 1$, and adding $-(l \cdot 1)$ to both sides gives $(k-l) \cdot 1 = 0$. If $k - l > 0$, then by Definition 0.7.2 the characteristic of F is nonzero and at most $k - l \leq n$, which contradicts the hypothesis. Hence $k = l$. In $\mathbb{R}$ and $\mathbb{C}$, $r \cdot 1$ is the number r. ∎

The hypothesis cannot be dropped. Over $\mathbb{F}_2$, the identity of $\mathbb{F}_2^2$ is a projection of rank 2, and its trace is $1 + 1 = 0$, the same as the trace of the zero projection, of rank 0. The equation $\operatorname{tr} P = (\operatorname{rank} P) \cdot 1$ still holds there; what fails is recovering the integer $\operatorname{rank} P$ from the field element $\operatorname{tr} P$.

4.5.3 Involutions and reflections

A projection satisfies $P^2 = P$. The other simple polynomial relation is $T^2 = \mathrm{id}_V$: doing T twice changes nothing. Such an operator is called an **involution**. Examples are the reflection $R(x, y) = (y, x)$, the transpose $\mathbf{A} \mapsto \mathbf{A}^\top$ on $M_n(F)$, and $f(x) \mapsto f(-x)$ on functions $\mathbb{R} \to \mathbb{R}$. In Chapter 1 the last two produced direct sums: symmetric plus skew-symmetric matrices (Example (Symmetric and skew-symmetric matrices)), and even plus odd functions (Example (Even and odd functions)). The next theorem explains both at once. For an involution T, write

$$E_+ := \{v \in V : Tv = v\}, \qquad E_- := \{v \in V : Tv = -v\}.$$

These are subspaces, since $E_+ = \ker(T - \mathrm{id}_V)$ and $E_- = \ker(T + \mathrm{id}_V)$ (Theorem 2.2.1).

Theorem 4.5.5. Involutions Split the Space

Let F be a field of characteristic not 2, V a vector space over F, and $T \in \mathcal{L}(V)$ with $T^2 = \mathrm{id}_V$. Then

$$V = E_+ \oplus E_-,$$

and the projection onto E_+ along E_- is $\frac{1}{2}(\mathrm{id}_V + T)$.

Idea. Copy the even/odd split: $v = \frac{1}{2}(v + Tv) + \frac{1}{2}(v - Tv)$. Applying T swaps v and Tv, so the first piece is fixed and the second changes sign. Both halves of the argument divide by 2, which is where the characteristic enters.

Proof. Since the characteristic is not 2, $2 = 1 + 1 \neq 0$ in F (Definition 0.7.2), so $\frac{1}{2}$ exists. Let $v \in V$, and put $v_+ = \frac{1}{2}(v + Tv)$ and $v_- = \frac{1}{2}(v - Tv)$. Then $v = v_+ + v_-$, and by linearity and $T^2 = \mathrm{id}_V$,

$$Tv_+ = \tfrac{1}{2}(Tv + v) = v_+, \qquad Tv_- = \tfrac{1}{2}(Tv - v) = -v_-.$$

So $v_+ \in E_+$, $v_- \in E_-$, and $V = E_+ + E_-$. If $v \in E_+ \cap E_-$, then $v = Tv = -v$, so $2v = 0$ and $v = \frac{1}{2}(2v) = 0$. By Theorem 1.7.4, $V = E_+ \oplus E_-$.

The operator $\frac{1}{2}(\mathrm{id}_V + T)$ is linear by Theorem 2.4.1, sends v to v_+, fixes E_+ (where $Tv = v$), and kills E_- (where $Tv = -v$). By the uniqueness in Theorem 4.5.1 (b), it is the projection onto E_+ along E_-.

■

For the transpose, E_+ is the symmetric and E_- the skew-symmetric matrices; for $f(x) \mapsto f(-x)$, they are the even and odd functions; for the reflection $R(x, y) = (y, x)$, they are the mirror line $y = x$ and the perpendicular line $y = -x$, and the adapted basis $((1, 1), (1, -1))$ is the one of Example (The reflection in a mirror-adapted basis). In finite dimension, a basis of E_+ followed by a basis of E_- gives $[T]_{\mathcal{B}}^{\mathcal{B}} = \mathrm{diag}(1, \ldots, 1, -1, \ldots, -1)$, and so $\operatorname{tr} T = (\dim E_+) \cdot 1 - (\dim E_-) \cdot 1$. In Chapter 9 the nonzero vectors of E_+ and E_- will be called eigenvectors with eigenvalues 1 and -1.

Warning. **In characteristic 2 an involution need not split the space.** Over $\mathbb{F}_2$, let

$T\mathbf{x} = \mathbf{J}\mathbf{x}$ with $\mathbf{J} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$. Then $\mathbf{J}^2 = \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix} = \mathbf{I}_2$, since $2 = 0$. But $-1 = 1$, so $E_- = E_+ = \ker(\mathbf{J} - \mathbf{I}_2) = \operatorname{span}(\mathbf{e}_1)$, and $E_+ + E_- = \operatorname{span}(\mathbf{e}_1) \neq \mathbb{F}_2^2$.

4.5.4 Exercises

4.5.4.1 A. Check your understanding

Exercise. A1

(a) Define a projection $P \in \mathcal{L}(V)$, and state the correspondence between projections and direct sum decompositions.

(b) True or false: a projection is determined by its image. Justify your answer.

(c) True or false: an operator T on $\mathbb{R}^2$ with $\operatorname{im} T$ the x-axis and $\ker T$ the y-axis is a projection. Justify your answer.

(d) Explain why $\operatorname{tr} T$ does not depend on the basis used to compute it.

(e) True or false: for every field F and every projection P on a finite-dimensional space over F, the integer $\operatorname{rank} P$ can be recovered from $\operatorname{tr} P \in F$. Justify your answer.

(f) Where does the hypothesis "characteristic not 2" enter the proof of Theorem 4.5.5?

Solution.

(a) P is a projection if $P^2 = P$ (Definition 4.5.1). By Theorem 4.5.1, a projection gives $V = \operatorname{im} P \oplus \ker P$, and each decomposition $V = U \oplus W$ gives exactly one projection with image U and kernel W.

(b) False. On $\mathbb{R}^2$, $P(x, y) = (x, 0)$ and $P'(x, y) = (x - y, 0)$ are different projections with the same image, the x-axis. A projection is determined by its image and its kernel together.

(c) False. $T(x, y) = (2x, 0)$ has this image and kernel, but $T^2 \neq T$.

(d) Matrices of T in two bases are similar by Theorem 4.2.3, and similar matrices have equal traces by Theorem 4.2.5.

(e) False. Over $\mathbb{F}_2$, id on $\mathbb{F}_2^2$ has rank 2 and the zero operator has rank 0, but both have trace 0. It is true when the characteristic is 0 or larger than $\dim V$ (Theorem 4.5.4).

(f) Twice: to define $v_\pm = \frac{1}{2}(v \pm Tv)$, and to pass from $2v = 0$ to $v = 0$. Both need $2 \neq 0$ in F.

4.5.4.2 B. Practice

Exercise. B1: A projection in $\mathbb{R}^3$

Let $U = \{(x, y, z) \in \mathbb{R}^3 : x + y - z = 0\}$ and $L = \operatorname{span}((1, 1, 1))$.

(a) Show that $\mathbb{R}^3 = U \oplus L$.

(b) Find the standard matrix $\mathbf{A}$ of the projection P onto U along L.

(c) Verify that $\mathbf{A}^2 = \mathbf{A}$, and that $\operatorname{tr} \mathbf{A} = \operatorname{rank} \mathbf{A}$.

Solution.

(a) U is the kernel of $(x, y, z) \mapsto x + y - z$, which is nonzero; its image is then a nonzero subspace of $\mathbb{R}$, hence all of $\mathbb{R}$ (Theorem 1.6.8), so $\dim U = 3 - 1 = 2$ by Theorem 2.3.1. If $t(1, 1, 1) \in U$, then $t+t-t = t = 0$, so $U \cap L = \{\mathbf{0}\}$. By Theorem 1.7.4, $\dim(U+L) = 2 + 1 = 3$, so $U + L = \mathbb{R}^3$ by Theorem 1.6.8, and $\mathbb{R}^3 = U \oplus L$.

(b) Split $(a, b, c) = \mathbf{u} + t(1, 1, 1)$ with $\mathbf{u} = (a - t, b - t, c - t) \in U$: the condition $(a - t) + (b - t) - (c - t) = 0$ gives $t = a + b - c$. So

$$P(a, b, c) = (-b + c,\ -a + c,\ -a - b + 2c), \qquad \mathbf{A} = \begin{pmatrix} 0 & -1 & 1 \\ -1 & 0 & 1 \\ -1 & -1 & 2 \end{pmatrix}.$$

Check: $\mathbf{A}(1, 1, 1) = \mathbf{0}$, and for the vectors $(1, 0, 1), (0, 1, 1) \in U$, $\mathbf{A}(1, 0, 1) = (1, 0, 1)$ and $\mathbf{A}(0, 1, 1) = (0, 1, 1)$.

(c) Row by row,

$$\mathbf{A}^2 = \begin{pmatrix} 0+1-1 & 0+0-1 & 0-1+2 \\ 0+0-1 & 1+0-1 & -1+0+2 \\ 0+1-2 & 1+0-2 & -1-1+4 \end{pmatrix} = \begin{pmatrix} 0 & -1 & 1 \\ -1 & 0 & 1 \\ -1 & -1 & 2 \end{pmatrix} = \mathbf{A}.$$

$\operatorname{tr} \mathbf{A} = 0 + 0 + 2 = 2$, and $\operatorname{rank} \mathbf{A} = \operatorname{rank} P = \dim U = 2$ by Theorem 4.5.1 (b), in agreement with Theorem 4.5.4.

Exercise. B2: The complementary projection

Let $P \in \mathcal{L}(V)$ be a projection.

(a) Prove that $\mathrm{id}_V - P$ is a projection.

(b) Prove that $\operatorname{im}(\mathrm{id}_V - P) = \ker P$ and $\ker(\mathrm{id}_V - P) = \operatorname{im} P$.

(c) Deduce that if P is the projection onto U along W, then $\mathrm{id}_V - P$ is the projection onto W along U.

Solution.

(a) By Theorem 2.4.2 (composition distributes over sums of maps), $(\mathrm{id}_V - P)^2 = \mathrm{id}_V - 2P + P^2 = \mathrm{id}_V - 2P + P = \mathrm{id}_V - P$.

(b) ($\subseteq$) $P((\mathrm{id}_V - P)(v)) = Pv - P^2(v) = 0$, so $\operatorname{im}(\mathrm{id}_V - P) \subseteq \ker P$. ($\supseteq$) If $Pv = 0$, then $v = (\mathrm{id}_V - P)(v) \in \operatorname{im}(\mathrm{id}_V - P)$. For the kernel: $(\mathrm{id}_V - P)(v) = 0$ means $v = Pv$. Such v lies in $\operatorname{im} P$, and conversely every $x \in \operatorname{im} P$ satisfies $Px = x$, as shown after Definition 4.5.1. So $\ker(\mathrm{id}_V - P) = \operatorname{im} P$.

(c) By Theorem 4.5.1 (b), $\operatorname{im} P = U$ and $\ker P = W$. By (a) and (b), $\mathrm{id}_V - P$ is a projection with image W and kernel U. By the uniqueness in Theorem 4.5.1 (b), applied to $V = W \oplus U$, it is the projection onto W along U.

Exercise. B3: Traces of operators on matrices

(a) Let $\tau\colon M_n(F) \to M_n(F)$, $\tau(\mathbf{A}) = \mathbf{A}^\top$. Show that $\operatorname{tr}\tau = n \cdot 1$.

(b) Fix $\mathbf{B} \in M_2(F)$, and let $L_{\mathbf{B}}\colon M_2(F) \to M_2(F)$, $L_{\mathbf{B}}(\mathbf{X}) = \mathbf{BX}$. Show that $\operatorname{tr} L_{\mathbf{B}} = 2\operatorname{tr}\mathbf{B}$.

Hint: use the basis of matrix units $\mathbf{E}_{ij}$.

Solution.

(a) Order the basis $(\mathbf{E}_{ij})$ of $M_n(F)$ in any way. Since $\tau(\mathbf{E}_{ij}) = \mathbf{E}_{ji}$, the coordinate column of $\tau(\mathbf{E}_{ij})$ has a 1 in the position of $\mathbf{E}_{ji}$ and 0 elsewhere. The diagonal entry of the matrix in the column of $\mathbf{E}_{ij}$ is therefore 1 if $\mathbf{E}_{ji} = \mathbf{E}_{ij}$, that is $i = j$, and 0 otherwise. There are n basis vectors $\mathbf{E}_{ii}$, so $\operatorname{tr}\tau = n \cdot 1$ by Definition 4.5.2. (Consistently with Theorem 4.5.5 over $\mathbb{R}$: symmetric matrices have dimension $\frac{n(n+1)}{2}$, skew-symmetric ones $\frac{n(n-1)}{2}$, and the difference is n.)

(b) Use the basis $(\mathbf{E}_{11}, \mathbf{E}_{12}, \mathbf{E}_{21}, \mathbf{E}_{22})$ and write $\mathbf{B} = (b_{ij})$. The product $\mathbf{BE}_{ij}$ has the i-th column of $\mathbf{B}$ in column j and zeros elsewhere, so $\mathbf{BE}_{ij} = b_{1i}\mathbf{E}_{1j} + b_{2i}\mathbf{E}_{2j}$. The diagonal entry in the column of $\mathbf{E}_{ij}$ is the coefficient of $\mathbf{E}_{ij}$ itself, which is b_{ii}. Adding over the four basis vectors, $\operatorname{tr} L_{\mathbf{B}} = b_{11} + b_{11} + b_{22} + b_{22} = 2\operatorname{tr}\mathbf{B}$.

4.5.4.3 C. Going deeper

Exercise. C1: When is a sum of projections a projection?

Let F have characteristic not 2, and let $P, Q \in \mathcal{L}(V)$ be projections.

(a) Prove that $P + Q$ is a projection if and only if $PQ = QP = 0$.

(b) Show that (a) fails over $\mathbb{F}_2$.

(c) Suppose V is finite-dimensional over $\mathbb{R}$ and $P+Q$ is a projection. Deduce that $\operatorname{rank}(P+Q) = \operatorname{rank} P + \operatorname{rank} Q$.

Hint: for ($\Rightarrow$) in (a), use the relation $PQ + QP = 0$ together with $P^2 = P$.

Solution.

(a) By Theorem 2.4.2, $(P+Q)^2 = P^2 + PQ + QP + Q^2 = P + Q + PQ + QP$.
($\Leftarrow$) If $PQ = QP = 0$, this is $P + Q$.
($\Rightarrow$) If $(P+Q)^2 = P + Q$, then $PQ + QP = 0$. Multiplying on the left by P and using $P^2 = P$ gives $PQ + PQP = 0$; multiplying on the right by P gives $PQP + QP = 0$. Subtracting, $PQ - QP = 0$, so $PQ = QP$. Then $2PQ = PQ + QP = 0$, and since $2 \neq 0$ in F, $PQ = 0$. Hence also $QP = 0$.

(b) Over $\mathbb{F}_2$, let $P = Q = \mathrm{id}_V$ with $V \neq \{0\}$. Then $P + Q = 2\,\mathrm{id}_V = 0$, which is a projection, but $PQ = \mathrm{id}_V \neq 0$. The step from $2PQ = 0$ to $PQ = 0$ is the one that fails.

(c) By Theorem 4.5.4 over $\mathbb{R}$, and Theorem 4.5.3,

$$\operatorname{rank}(P+Q) = \operatorname{tr}(P+Q) = \operatorname{tr} P + \operatorname{tr} Q = \operatorname{rank} P + \operatorname{rank} Q.$$

Exercise. C2: The trace is the only trace-like functional

Let $n \geq 1$, and let $\varphi\colon M_n(F) \to F$ be linear with $\varphi(\mathbf{AB}) = \varphi(\mathbf{BA})$ for all $\mathbf{A}, \mathbf{B} \in M_n(F)$.

(a) Using $\mathbf{E}_{ij}\mathbf{E}_{jk} = \mathbf{E}_{ik}$ and $\mathbf{E}_{jk}\mathbf{E}_{ij} = \mathbf{O}$ for $k \neq i$, show that $\varphi(\mathbf{E}_{ik}) = 0$ whenever $i \neq k$.

(b) Show that $\varphi(\mathbf{E}_{ii}) = \varphi(\mathbf{E}_{jj})$ for all i, j.

(c) Deduce that $\varphi = c\ \mathrm{tr}$ for some $c \in F$, and conversely that every $c\ \mathrm{tr}$ has the property.

Solution. Recall that $\mathbf{E}_{ij}\mathbf{E}_{kl}$ is $\mathbf{E}_{il}$ if $j = k$ and $\mathbf{O}$ otherwise, by Definition 0.9.9.

(a) Let $i \neq k$. Then $\mathbf{E}_{ik} = \mathbf{E}_{ii}\mathbf{E}_{ik}$ and $\mathbf{E}_{ik}\mathbf{E}_{ii} = \mathbf{O}$, because $k \neq i$. By the hypothesis and linearity, $\varphi(\mathbf{E}_{ik}) = \varphi(\mathbf{E}_{ii}\mathbf{E}_{ik}) = \varphi(\mathbf{E}_{ik}\mathbf{E}_{ii}) = \varphi(\mathbf{O}) = 0$.

(b) $\mathbf{E}_{ii} = \mathbf{E}_{ij}\mathbf{E}_{ji}$ and $\mathbf{E}_{jj} = \mathbf{E}_{ji}\mathbf{E}_{ij}$. Hence $\varphi(\mathbf{E}_{ii}) = \varphi(\mathbf{E}_{ij}\mathbf{E}_{ji}) = \varphi(\mathbf{E}_{ji}\mathbf{E}_{ij}) = \varphi(\mathbf{E}_{jj})$.

(c) Let $c = \varphi(\mathbf{E}_{11})$. For $\mathbf{A} = (a_{ij}) = \sum_{i,j} a_{ij}\mathbf{E}_{ij}$, linearity with (a) and (b) gives

$$\varphi(\mathbf{A}) = \sum_{i,j} a_{ij}\varphi(\mathbf{E}_{ij}) = \sum_{i} a_{ii}\varphi(\mathbf{E}_{ii}) = c\sum_{i} a_{ii} = c\ \mathrm{tr}\,\mathbf{A}.$$

Conversely, $c\ \mathrm{tr}$ is linear and $c\ \mathrm{tr}(\mathbf{AB}) = c\ \mathrm{tr}(\mathbf{BA})$ by Theorem 0.9.6. So the linear functionals with $\varphi(\mathbf{AB}) = \varphi(\mathbf{BA})$ are exactly the multiples of the trace.

Chapter 5

Duality

A linear map $V \to F$ turns each vector into a single number. The i-th coordinate in a basis does this, and so do evaluating a polynomial at a point, integrating a function, and taking the trace of a matrix. These **linear functionals** are the measurements we can make on a vector space, and together they form a vector space of their own, the dual space V^*.

The dual space mirrors V with every inclusion reversed. A subspace corresponds to the functionals that vanish on it. An injective map corresponds to a surjective one. The kernel of a map corresponds to the image of its dual map, whose matrix is the transpose. Passing to the dual turns a hard question into an easy one surprisingly often. It also explains facts we already know from a new angle: why a subspace of F^n is always the solution set of a homogeneous system, and why row rank equals column rank.

What you need. From Chapter 0, the transpose and its arithmetic (Definition 0.9.10, Theorem 0.9.5) and matrix multiplication read column by column (Theorem 0.9.3), together with polynomials (Definition 0.8.1), which supply the infinite-dimensional example of §01. From Chapter 1, bases, dimension, and complements (Definition 1.5.1, Definition 1.6.1, Theorem 1.7.6), and bases of infinite-dimensional spaces (Theorem 1.8.2) for the remarks on infinite dimension. From Chapter 2, linear maps and the space $\mathcal{L}(V, W)$ (Definition 2.1.1, Theorem 2.4.1) and the Rank–Nullity Theorem (Theorem 2.3.1). From Chapter 3, the row space and null space of a matrix (Definition 3.5.2, Definition 3.5.3) and inverses by row reduction (Theorem 3.4.6). From Chapter 4, the matrix of a map and change of basis (Definition 4.1.1, Theorem 4.2.1), and quotient spaces with the First Isomorphism Theorem (Definition 4.4.3, Theorem 4.4.7).

5.0.1 Roadmap

- **Linear functionals and the dual space.** Measurements as vectors; dual bases; separating vectors by functionals.
- **Annihilators.** The functionals that vanish on a subspace, and the formula $\dim U + \dim U^0 = \dim V$.
- **The double dual.** A natural isomorphism $V \to V^{**}$ that needs no basis, and what "natural" means.
- **The dual map and the transpose.** Pulling functionals back along a map; its matrix is the transpose; kernels and images trade places.
- **Duality for subspaces and quotients.** The dual of a subspace is a quotient of the dual, and the dual of a quotient is an annihilator; hyperplanes and codimension.
- **Hyperplanes and linear systems revisited.** Every subspace is cut out by equations; converting between parametric and implicit descriptions.

5.0.2 Named moves

- **Test against functionals.** To show a vector is zero, show that every functional (or every vector of a dual basis) sends it to zero.
- **Pass to the dual and count.** Replace a subspace by its annihilator and use $\dim U + \dim U^0 = \dim V$.
- **Natural or chosen?** Ask whether an isomorphism needs a basis. $V \cong V^*$ does; $V \cong V^{**}$ does not.
- **Dualize a statement.** Injective becomes surjective, kernel becomes image, a matrix becomes its transpose, and inclusions reverse.

5.1 Linear Functionals and the Dual Space

Chapter 2 studied linear maps $V \to W$ into arbitrary targets. The smallest nonzero target is the field F itself, and maps into it are everywhere: a coordinate, the value of a polynomial at a point, an integral, the trace. This section names these maps and shows that they form a vector space V^*, of the same dimension as V when V is finite-dimensional. It also proves the fact that makes them useful: every nonzero vector is detected by at least one of them.

5.1.1 Measuring a vector by a number

After studying vector spaces and the maps between them, we look at the simplest possible target. Here are four maps we have already met, from four different families.

- **A coordinate.** Fix a basis $\mathcal{B} = (v_1, \dots, v_n)$ of V. The map $V \to F$ sending v to the first entry of $[v]_{\mathcal{B}}$ respects sums and multiples, by Theorem 1.5.2.
- **Evaluation.** The map $F[x] \to F$, $p \mapsto p(3)$, satisfies $(p+q)(3) = p(3) + q(3)$ and $(cp)(3) = c\,p(3)$, by Theorem 0.8.5 (a constant c is a polynomial with value c).
- **Integration.** The map $\mathbb{R}[x] \to \mathbb{R}$, $p \mapsto \int_0^1 p(t)\,dt$, is linear by the linearity of the integral from calculus.
- **Trace.** The map $\operatorname{tr}\colon M_n(F) \to F$ is linear (Example (Trace and Transpose)).

The inputs are columns, polynomials, and matrices. The outputs are always single scalars. Each map is a way of **measuring** a vector by one number, and each measurement respects sums and scalar multiples. The same expression keeps recurring, "a linear map whose target is F", and we will see that these maps form a space with a structure of its own. So we give them a name.

A linear functional is a linear measurement: it assigns a scalar to each vector, and the measurement of a linear combination is the same combination of the measurements.

Definition 5.1.1. Linear Functional

Let V be a vector space over F, and regard F as a vector space over itself. A **linear functional** on V is a linear map $\varphi\colon V \to F$. That is, **for all** $u, v \in V$ and **all** $c \in F$,

$$\varphi(u+v) = \varphi(u) + \varphi(v), \qquad \varphi(cv) = c\,\varphi(v).$$

Definition 5.1.2. Dual Space

Let V be a vector space over F. The **dual space** of V is the set of **all** linear functionals on

V,

$$V^* \coloneqq \mathcal{L}(V, F),$$

with the pointwise operations $(\varphi+\psi)(v) \coloneqq \varphi(v)+\psi(v)$ and $(c\varphi)(v) \coloneqq c\,\varphi(v)$ for $\varphi, \psi \in V^*$, $c \in F$, and $v \in V$.

In words: a linear functional is a linear map (Definition 2.1.1) whose codomain is the one-dimensional space F. Its values lie in the **same** field F over which V is a vector space. The first condition says that measuring a sum gives the sum of the measurements, and the second says that scaling a vector scales its measurement. The dual space collects **every** such measurement, and it adds them the natural way: the measurement $\varphi+\psi$ reports the sum of the two readings. We write functionals with Greek letters φ, ψ, which keeps them apart from vectors of V.

Well-definedness. Two things need checking: that $\varphi+\psi$ and $c\varphi$ are again linear, and that the vector space axioms hold. Both are the case $W = F$ of Theorem 2.4.1. So V^* is a vector space over F. Its zero vector is the **zero functional** $v \mapsto 0$, and the negative of φ is $v \mapsto -\varphi(v)$. Note that V^* is a set of **functions on** V, not a subset of V.

Example. Four Families of Functionals

Show that each map is a linear functional.

(a) $\varepsilon_c \colon F[x] \to F$, $p \mapsto p(c)$, for a fixed $c \in F$ (**evaluation at** c).

(b) $\iota \colon \mathbb{R}[x]_{\le 2} \to \mathbb{R}$, $p \mapsto \int_0^1 p(t)\,\mathrm{d}t$.

(c) $\operatorname{tr} \colon M_n(F) \to F$.

(d) The zero functional on any V. Also find the dual space of the zero space $\{0\}$.

Solution.

(a) By Theorem 0.8.5, $(p+q)(c) = p(c)+q(c)$, and, regarding ap as the product of the constant polynomial a with p, $(ap)(c) = a(c)\,p(c) = a\,p(c)$, because the constant polynomial a has value a at c. These are the two conditions of Definition 5.1.1.

(b) By linearity of the integral, $\int_0^1 (p+q) = \int_0^1 p + \int_0^1 q$ and $\int_0^1 ap = a\int_0^1 p$. On coefficients, $\iota(a_0 + a_1x + a_2x^2) = a_0 + \frac{a_1}{2} + \frac{a_2}{3}$.

(c) This is Theorem 0.9.6 (1): $\operatorname{tr}(\mathbf{A}+\mathbf{B}) = \operatorname{tr}\mathbf{A} + \operatorname{tr}\mathbf{B}$ and $\operatorname{tr}(c\mathbf{A}) = c\operatorname{tr}\mathbf{A}$.

(d) $0 = 0+0$ and $0 = c\cdot 0$ in F, so the zero functional is linear. On $\{0\}$, a linear functional must send 0 to 0 (Theorem 2.1.1), and there is only one vector. So the zero functional is the only element, and $\{0\}^* = \{0\}$. This degenerate case matters: it is the case $n = 0$ of the count $\dim V^* = \dim V$ proved below.

The family F^n can be described completely. A row vector times a column is a scalar, and every functional on F^n arises this way.

Theorem 5.1.1. Functionals on F^n Are Row Vectors

For $\mathbf{a} = (a_1, \dots, a_n) \in F^n$, define $\varphi_{\mathbf{a}} \colon F^n \to F$ by

$$\varphi_{\mathbf{a}}(\mathbf{x}) \coloneqq \mathbf{a}^{\top}\mathbf{x} = a_1x_1 + \cdots + a_nx_n.$$

Then $\varphi_{\mathbf{a}} \in (F^n)^*$. Every $\varphi \in (F^n)^*$ equals $\varphi_{\mathbf{a}}$ for **exactly one a**, namely $(\varphi(\mathbf{e}_1), \dots, \varphi(\mathbf{e}_n))$. The map $F^n \to (F^n)^*$, $\mathbf{a} \mapsto \varphi_{\mathbf{a}}$, is an isomorphism.

Proof. The map $\varphi_{\mathbf{a}}$ is $T_{\mathbf{a}^\top}$ for the $1 \times n$ matrix $\mathbf{a}^\top$, with a 1×1 matrix read as a scalar, so it is linear by Example (Multiplication by a Matrix). Let $\varphi \in (F^n)^*$ and put $a_i := \varphi(\mathbf{e}_i)$. Then $\varphi_{\mathbf{a}}(\mathbf{e}_i) = a_i = \varphi(\mathbf{e}_i)$ for each i. Two linear maps that agree on the basis $(\mathbf{e}_1, \dots, \mathbf{e}_n)$ are equal by the uniqueness part of Theorem 2.1.5, so $\varphi = \varphi_{\mathbf{a}}$. If $\varphi_{\mathbf{a}} = \varphi_{\mathbf{b}}$, then $a_i = \varphi_{\mathbf{a}}(\mathbf{e}_i) = \varphi_{\mathbf{b}}(\mathbf{e}_i) = b_i$ for every i, so $\mathbf{a} = \mathbf{b}$.

Hence $\mathbf{a} \mapsto \varphi_{\mathbf{a}}$ is a bijection $F^n \to (F^n)^*$. It is linear, since $\varphi_{\mathbf{a}+c\mathbf{b}}(\mathbf{x}) = \sum_i (a_i + cb_i)x_i = \varphi_{\mathbf{a}}(\mathbf{x}) + c\,\varphi_{\mathbf{b}}(\mathbf{x})$ for every $\mathbf{x}$. By Theorem 2.5.1 it is an isomorphism.

■

So in F^n a functional **is** a row vector acting by matrix multiplication, and its $1 \times n$ matrix in the standard bases is $(\varphi(\mathbf{e}_1) \ \cdots \ \varphi(\mathbf{e}_n))$. This is how we compute with functionals: rows act on columns.

Nonexamples by minimal change. Each of the following changes one of the examples slightly. In each case something still works, and we name a condition that fails.

- On $\mathbb{R}^2$, the length $(x, y) \mapsto \sqrt{x^2 + y^2}$ (Example (Length)). It sends $\mathbf{0}$ to 0 and respects **positive** scalars. But with $c = -1$ and $\mathbf{v} = (1, 0)$ it gives 1, not -1, so the scaling condition fails.
- On $\mathbb{R}[x]$, squaring the evaluation at 0: $p \mapsto p(0)^2$. The outputs are still scalars. But for $p = q = 1$ we get $(1 + 1)^2 = 4 \neq 1 + 1$, so additivity fails.
- On $M_2(F)$, replace the trace $a+d$ of $\mathbf{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ by $ad-bc$. Then $\mathbf{E}_{11} \mapsto 0$ and $\mathbf{E}_{22} \mapsto 0$, but $\mathbf{E}_{11} + \mathbf{E}_{22} = \mathbf{I}_2 \mapsto 1$, so additivity fails. The expression $ad - bc$ is linear in each **row** separately, which is a different property.

Why this definition. Why single out the target F? Because nothing is lost. A linear map $T\colon V \to F^m$ is the same thing as m functionals: if $\varphi_i(v)$ denotes the i-th entry of Tv, then each φ_i is linear and $Tv = (\varphi_1(v), \dots, \varphi_m(v))$; conversely any m functionals stacked this way give a linear map. So functionals are the "rows" of linear maps, and understanding them one at a time is a way of understanding all maps into F^m. The name comes from analysis, where the vectors were themselves functions and a map such as $f \mapsto \int_0^1 f$ was called a "functional", a function of a function. The word "dual" will explain itself in this chapter: every statement about V has a mirror statement about V^*.

Warning. **The values must lie in the field of scalars of V.** Regard $\mathbb{C}$ as a vector space over $\mathbb{C}$. The real part $z \mapsto \operatorname{Re} z$ is additive, and it respects **real** scalars, but it is **not** a linear functional on $\mathbb{C}$ over $\mathbb{C}$: with $c = i$ and $z = 1$, $\operatorname{Re}(i \cdot 1) = 0 \neq i \cdot \operatorname{Re}(1) = i$. It is a functional on $\mathbb{C}$ regarded as a vector space over $\mathbb{R}$. The dual space depends on the field.

The first payoff is a count that holds for **every** nonzero functional: it hits every scalar, and in finite dimension it kills exactly one dimension.

Proposition 5.1.2. A Non-Zero Functional Is Surjective

Let V be a vector space over F, and let $\varphi \in V^*$ be **nonzero**. Then φ is surjective. If V is finite-dimensional, then $\dim \ker \varphi = \dim V - 1$.

Proof. Since $\varphi \neq 0$, there is $u \in V$ with $c := \varphi(u) \neq 0$. For every $a \in F$, linearity gives $\varphi(ac^{-1}u) = ac^{-1}c = a$, so φ is surjective. Hence $\operatorname{im}\varphi = F$ and $\operatorname{rank}\varphi = 1$. If V is finite-dimensional, the Rank–Nullity Theorem (Theorem 2.3.1) gives $\dim \ker \varphi = \dim V - 1$. ■

For example, $\ker \operatorname{tr}$ in $M_n(F)$ has dimension $n^2 - 1$, and the polynomials in $F[x]_{\leq n}$ with $p(c) = 0$ form a subspace of dimension n. Every nonzero functional cuts out a subspace one dimension smaller than the whole space. Later in this chapter these kernels, the hyperplanes, become the building blocks of every subspace.

Quick check. Which of the following are linear functionals? (a) $p \mapsto p(0) + 1$ on $\mathbb{R}[x]$. (b) $p \mapsto p(1) - 2p'(0)$ on $\mathbb{R}[x]$. (c) $\mathbf{A} \mapsto a_{12}$ on $M_2(F)$.

Solution.

(a) No. It sends the zero polynomial to 1, but a linear map sends 0 to 0 (Theorem 2.1.1). (b) Yes. Evaluation at 1, differentiation, and evaluation at 0 are linear, and V^* is closed under linear combinations and composition with linear maps (Theorem 2.4.1, Theorem 2.4.2). (c) Yes. It is the coordinate of $\mathbf{A}$ at $\mathbf{E}_{12}$ in the standard basis of Example (Standard Bases), and coordinates are linear (Theorem 1.5.2).

5.1.2 The dual basis

Return to the coordinate functionals. Let $\mathcal{B} = (v_1, \dots, v_n)$ be a basis of V, and let $\varphi_i(v)$ be the i-th entry of $[v]_{\mathcal{B}}$. The basis vector v_j has coordinate vector $\mathbf{e}_j$, so

$$\varphi_i(v_j) = \delta_{ij} = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j, \end{cases}$$

with the Kronecker delta of Definition 0.9.3. Each φ_i reads off one coordinate and ignores the others. These n functionals are the natural candidates for a basis of V^*, and the theorem says they are one.

Theorem 5.1.3. The Dual Basis

Let V be a finite-dimensional vector space over F, and let $\mathcal{B} = (v_1, \dots, v_n)$ be a basis of V.

(a) There are **unique** $\varphi_1, \dots, \varphi_n \in V^*$ with $\varphi_i(v_j) = \delta_{ij}$ for all i, j. For every $v \in V$, $\varphi_i(v)$ is the i-th entry of $[v]_{\mathcal{B}}$.

(b) The list $(\varphi_1, \dots, \varphi_n)$ is a basis of V^*.

(c) For every $v \in V$ and every $\varphi \in V^*$,

$$v = \sum_{i=1}^{n} \varphi_i(v)\, v_i, \qquad \varphi = \sum_{i=1}^{n} \varphi(v_i)\, \varphi_i.$$

Idea. ① A functional is a linear map into F, and a linear map may be prescribed freely on a basis, so existence and uniqueness in (a) are one citation. ② Independence: apply a vanishing combination $\sum c_i\varphi_i$ to v_j; the delta kills every term but c_j. ③ Spanning: we cannot count, because $\dim V^*$ is exactly what we do not know yet. So we guess the coefficients. If $\varphi = \sum c_i\varphi_i$, then applying both sides to v_j forces $c_j = \varphi(v_j)$. With this guess, the two sides are linear maps that agree on a basis, hence equal.

Proof.

(a) For each i, apply Theorem 2.1.5 with $W = F$ and the prescribed values $\delta_{i1}, \dots, \delta_{in} \in F$. It gives exactly one linear map $\varphi_i \colon V \to F$ with $\varphi_i(v_j) = \delta_{ij}$ for all j. If $v = a_1v_1 + \cdots + a_nv_n$, then by Theorem 2.1.4

$$\varphi_i(v) = \sum_{j=1}^{n} a_j\varphi_i(v_j) = \sum_{j=1}^{n} a_j\delta_{ij} = a_i,$$

the i-th entry of $[v]_{\mathcal{B}}$ (Definition 1.5.2).

(b) *Independence.* Let $c_1\varphi_1 + \cdots + c_n\varphi_n = 0$ in V^*. Applying both sides to v_j and using the pointwise operations of Definition 5.1.2,

$$0 = \sum_{i=1}^{n} c_i\varphi_i(v_j) = \sum_{i=1}^{n} c_i\delta_{ij} = c_j$$

for each j. Hence $(\varphi_1, \dots, \varphi_n)$ is linearly independent.
Spanning. Let $\varphi \in V^*$, and put $\psi := \sum_{i=1}^{n} \varphi(v_i)\varphi_i \in V^*$. For each j, $\psi(v_j) = \sum_i \varphi(v_i)\delta_{ij} = \varphi(v_j)$. So ψ and φ are linear maps $V \to F$ that agree on the basis $\mathcal{B}$, and $\psi = \varphi$ by the uniqueness part of Theorem 2.1.5. Hence $\varphi \in \operatorname{span}(\varphi_1, \dots, \varphi_n)$, and $(\varphi_1, \dots, \varphi_n)$ is a basis of V^*.

(c) The second formula was proved in the spanning step. The first is the definition of coordinates, since $\varphi_i(v)$ is the i-th coordinate of v by (a). This proves the theorem.

■

The two formulas in (c) are mirror images. A vector is recovered from its measurements by the dual basis; a functional is recovered from its values on the basis.

Definition 5.1.3. Dual Basis

Let $\mathcal{B} = (v_1, \dots, v_n)$ be a basis of a finite-dimensional vector space V over F. The basis $(\varphi_1, \dots, \varphi_n)$ of V^* given by Theorem 5.1.3, characterized by $\varphi_i(v_j) = \delta_{ij}$, is **the dual basis** of $\mathcal{B}$. We write it $\mathcal{B}^*$.

Counting the dual basis gives the dimension of the dual space at once.

Corollary 5.1.4. Dimension of the Dual Space

Let V be a finite-dimensional vector space over F. Then V^* is finite-dimensional and $\dim V^* = \dim V$. In particular $V^* \cong V$.

Proof. Let $n = \dim V$ and choose a basis of V (Corollary 1.5.5). By Theorem 5.1.3 (b), its dual basis is a basis of V^* of length n. Hence $\dim V^* = n$, and $V^* \cong V$ by Theorem 2.5.4. ∎

For instance $\dim M_{m\times n}(F)^* = mn$ and $\dim F[x]^*_{\le n} = n+1$.

Remark. Many books write v_i^* for the dual basis functional φ_i. We avoid that notation for two reasons. It suggests that φ_i is determined by v_i alone, which is false (see the warning below). And in this book the star on a map is reserved for adjoints, which come much later. We write φ_i, and speak of "the dual basis of $\mathcal{B}$".

How do we find a dual basis in practice? In F^n, functionals are rows, and the conditions $\varphi_i(v_j) = \delta_{ij}$ turn into a single matrix equation.

Example. The Dual Basis Is Read Off the Inverse Matrix

(a) Find the dual basis of $\mathcal{B} = ((1,1),(1,-1))$ in $\mathbb{R}^2$.

(b) More generally, let $\mathcal{B} = (\mathbf{v}_1, \dots, \mathbf{v}_n)$ be a basis of F^n, and let $\mathbf{P} = \begin{pmatrix} \mathbf{v}_1 & \cdots & \mathbf{v}_n \end{pmatrix}$. Show that $\mathbf{P}$ is invertible and that the dual basis of $\mathcal{B}$ is $\varphi_i(\mathbf{x}) = \mathbf{r}_i\mathbf{x}$, where $\mathbf{r}_i$ is the i-th **row** of $\mathbf{P}^{-1}$.

Solution.

(a) By Theorem 5.1.1, $\varphi_1(x,y) = px+qy$ for some $p,q \in \mathbb{R}$, that is, $\varphi_1(\mathbf{x}) = \begin{pmatrix} p & q \end{pmatrix}\mathbf{x}$. The conditions $\varphi_1(1,1) = 1$ and $\varphi_1(1,-1) = 0$ say $\begin{pmatrix} p & q \end{pmatrix}\mathbf{P} = \begin{pmatrix} 1 & 0 \end{pmatrix}$ with $\mathbf{P} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$. Likewise the row $\mathbf{r}_2$ of φ_2 satisfies $\mathbf{r}_2\mathbf{P} = \begin{pmatrix} 0 & 1 \end{pmatrix}$. Stacking the two rows into a matrix $\mathbf{R}$, the four conditions become $\mathbf{RP} = \mathbf{I}_2$, so $\mathbf{R} = \mathbf{P}^{-1}$. By Theorem 0.9.8, with $ad - bc = -2$,

$$\mathbf{P}^{-1} = \frac{1}{-2}\begin{pmatrix} -1 & -1 \\ -1 & 1 \end{pmatrix} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{pmatrix}.$$

Hence

$$\varphi_1(x,y) = \frac{x+y}{2}, \qquad \varphi_2(x,y) = \frac{x-y}{2}.$$

Check: $\varphi_1(1,1) = 1$, $\varphi_1(1,-1) = 0$, $\varphi_2(1,1) = 0$, $\varphi_2(1,-1) = 1$. And the first formula of Theorem 5.1.3 (c) for $\mathbf{v} = (3,1)$ reads $(3,1) = 2(1,1) + 1(1,-1)$, which is true. The second formula expresses a functional in the dual basis through its values on $\mathcal{B}$: for $\psi(x,y) = x$, the values are $\psi(1,1) = 1$ and $\psi(1,-1) = 1$, so $\psi = \varphi_1 + \varphi_2$, and indeed $\frac{x+y}{2} + \frac{x-y}{2} = x$.

(b) The columns of $\mathbf{P}$ form a basis of F^n, so $\mathbf{P}$ is invertible by Theorem 3.4.4. Let $\mathbf{R} \in M_n(F)$ have rows $\mathbf{r}_1, \dots, \mathbf{r}_n$, and let $\psi_i(\mathbf{x}) := \mathbf{r}_i\mathbf{x}$. By the entry formula for a product (Theorem 0.9.2), $\psi_i(\mathbf{v}_j) = \mathbf{r}_i\mathbf{v}_j = (\mathbf{RP})_{ij}$. Hence $\psi_i(\mathbf{v}_j) = \delta_{ij}$ for all i, j if and only if $\mathbf{RP} = \mathbf{I}_n$, if and only if $\mathbf{R} = \mathbf{P}^{-1}$ (multiply by $\mathbf{P}^{-1}$ on the right). Now take $\mathbf{R} = \mathbf{P}^{-1}$: the functionals $\psi_i(\mathbf{x}) = \mathbf{r}_i\mathbf{x}$ are linear (Theorem 5.1.1) and satisfy $\psi_i(\mathbf{v}_j) = \delta_{ij}$, so by the uniqueness in Theorem 5.1.3 (a) they are the dual basis.

This agrees with Chapter 2: $\mathbf{P} = [\mathrm{id}]_{\mathcal{B}}^{\mathcal{E}}$ and $\mathbf{P}^{-1} = [\mathrm{id}]_{\mathcal{E}}^{\mathcal{B}}$ (Theorem 4.2.1), so $\mathbf{P}^{-1}\mathbf{x} = [\mathbf{x}]_{\mathcal{B}}$, and row i of $\mathbf{P}^{-1}$ computes the i-th coordinate, as Theorem 5.1.3 (a) predicts. The columns of $\mathbf{P}$ are the basis; the rows of $\mathbf{P}^{-1}$ are the dual basis.

Warning. **A dual basis functional depends on the whole basis, not only on its own vector.** In $\mathbb{R}^2$, the basis $((1,0),(0,1))$ has first dual functional $\varphi_1(x,y) = x$. The basis $((1,0),(1,1))$ has the same first vector, but $\mathbf{P}^{-1} = \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix}$, so its first dual functional is $\varphi_1(x,y) = x - y$. Changing $\mathbf{v}_2$ changed φ_1, because φ_1 must vanish on $\mathbf{v}_2$.

The same method works away from F^n, once we recognize the coordinates.

Example. Dual Bases in $\mathbb{R}[x]_{\le 2}$

(a) Show that the dual basis of $(1, x, x^2)$ in $\mathbb{R}[x]_{\le 2}$ is $(\varphi_0, \varphi_1, \varphi_2)$ with $\varphi_0(p) = p(0)$, $\varphi_1(p) = p'(0)$, and $\varphi_2(p) = \frac{1}{2}p''(0)$. (We index from 0 to match the exponents.)

(b) Let $\ell_{-1} = \frac{1}{2}x(x-1)$, $\ell_0 = 1 - x^2$, and $\ell_1 = \frac{1}{2}x(x+1)$. Show that $(\ell_{-1}, \ell_0, \ell_1)$ is a basis of $\mathbb{R}[x]_{\le 2}$ whose dual basis is $(\varepsilon_{-1}, \varepsilon_0, \varepsilon_1)$, the evaluations at $-1, 0, 1$. Deduce that $p = p(-1)\ell_{-1} + p(0)\ell_0 + p(1)\ell_1$ for every $p \in \mathbb{R}[x]_{\le 2}$.

Solution.

(a) Each φ_k is linear, as a composition of differentiation (Example (Differentiation)) with evaluation at 0, scaled. For $p = a_0 + a_1x + a_2x^2$, we have $p(0) = a_0$, $p'(0) = a_1$, and $\frac{1}{2}p''(0) = \frac{1}{2} \cdot 2a_2 = a_2$. So $\varphi_k(x^j) = \delta_{kj}$ for $j, k \in \{0,1,2\}$, and by the uniqueness in Theorem 5.1.3 (a) these are the dual basis. Over a general field F, the dual basis of $(1, x, x^2)$ is still "the coefficient of x^k", but the formula $\frac{1}{2}p''(0)$ needs $2 \neq 0$ in F.

(b) The values of the ℓ's at the three points are

$$\begin{aligned} &\ell_{-1}(-1) = 1,\ \ell_{-1}(0) = 0,\ \ell_{-1}(1) = 0; \\ &\ell_0(-1) = 0,\ \ell_0(0) = 1,\ \ell_0(1) = 0; \\ &\ell_1(-1) = 0,\ \ell_1(0) = 0,\ \ell_1(1) = 1. \end{aligned}$$

That is, $\varepsilon_i(\ell_j) = \delta_{ij}$ for $i, j \in \{-1, 0, 1\}$. *Basis.* Let $c_{-1}\ell_{-1} + c_0\ell_0 + c_1\ell_1 = 0$. Evaluating at $j \in \{-1, 0, 1\}$ gives $c_j = 0$. So the list is independent, of length $3 = \dim \mathbb{R}[x]_{\le 2}$, hence a basis by Theorem 1.6.4. *Dual basis.* The evaluations are linear (Example (Four Families of Functionals)) and satisfy $\varepsilon_i(\ell_j) = \delta_{ij}$, so they are the dual basis by the uniqueness in Theorem 5.1.3 (a). The first formula of Theorem 5.1.3 (c) now reads $p = \sum_i \varepsilon_i(p)\,\ell_i = p(-1)\ell_{-1} + p(0)\ell_0 + p(1)\ell_1$.

The last formula rebuilds a polynomial from three of its values. It is a preview of the

Lagrange interpolation formula of Chapter 6, and it explains where the ℓ's come from: they are the basis dual to the evaluations.

Quick check. Let (v_1, v_2, v_3) be a basis of V with dual basis $(\varphi_1, \varphi_2, \varphi_3)$. Compute $\varphi_2(3v_1 - v_2 + 5v_3)$ and $(2\varphi_1 + \varphi_3)(v_3)$.

Solution. $\varphi_2(3v_1 - v_2 + 5v_3) = 3\cdot 0 - 1\cdot 1 + 5\cdot 0 = -1$, the second coordinate. $(2\varphi_1 + \varphi_3)(v_3) = 2\cdot 0 + 1 = 1$.

Warning. $V \cong V^*$ **is true in finite dimension, but no isomorphism is singled out.** Corollary 5.1.4 gives an isomorphism by sending $v_i \mapsto \varphi_i$, and that depends on the basis. In $\mathbb{R}^2$ with the standard basis, it sends $(1,0)$ to the functional $(x,y) \mapsto x$. With the basis $((1,0),(1,1))$, it sends the **same** vector $(1,0)$ to $(x,y) \mapsto x - y$, by the warning above. A vector does not come with "its" functional. Later in this chapter we make precise the sense in which V^* needs a choice to be identified with V, while the double dual V^{**} does not.

5.1.3 Testing against functionals

In F^n a vector is zero exactly when all its entries are zero, and the entries are the values of the coordinate functionals. The same holds in any finite-dimensional space, and it gives a flexible way to prove that a vector is zero.

Theorem 5.1.5. Functionals Separate Vectors

Let V be a finite-dimensional vector space over F.

(a) If $v \in V$ and $v \neq 0$, then there is $\varphi \in V^*$ with $\varphi(v) = 1$.

(b) If $v \in V$ satisfies $\varphi(v) = 0$ for **every** $\varphi \in V^*$, then $v = 0$. It suffices that $\varphi_i(v) = 0$ for the functionals φ_i of **one** dual basis.

(c) For $u, w \in V$: $u = w$ if and only if $\varphi(u) = \varphi(w)$ for every $\varphi \in V^*$.

Idea. We need a functional that does not vanish at v, and the only functionals we can build at will are dual basis functionals. The first one takes the value 1 on the first basis vector. So make v the first vector of a basis: this is possible exactly because $v \neq 0$.

Proof.

(a) Since $v \neq 0$, the list (v) is linearly independent: if $av = 0$, then $a = 0$ by Theorem 1.1.9. By the Basis Extension Theorem (Theorem 1.6.5), there is a basis $(v, v_2, \ldots, v_n)$ of V. Let $(\varphi_1, \ldots, \varphi_n)$ be its dual basis (Theorem 5.1.3). Then $\varphi_1(v) = 1$.

(b) The first statement is the contrapositive of (a). For the second, if $(\varphi_1, \ldots, \varphi_n)$ is the dual basis of a basis $(v_1, \ldots, v_n)$ and every $\varphi_i(v) = 0$, then $v = \sum_i \varphi_i(v) v_i = 0$ by

Theorem 5.1.3 (c).

(c) If $u = w$, the values agree. Conversely, if $\varphi(u) = \varphi(w)$ for every φ, then $\varphi(u - w) = \varphi(u) - \varphi(w) = 0$ for every φ, so $u - w = 0$ by (b).

■

This is the signature move of the chapter: **to show a vector is zero, test it against every functional, and a dual basis suffices.** It turns one equation in V into scalar equations, which are often easier to reach.

Remark. Parts (a), (b) (first sentence) and (c) hold for **every** vector space V; only (a) needs a new construction. By Theorem 1.8.2, $\{v\}$ extends to a basis B of V. Every $w \in V$ is a finite combination of distinct vectors of B, with coefficients that are unique because B is independent. Let $\varphi(w)$ be the coefficient of v in this expansion (and 0 if v does not occur). Adding or scaling expansions gives expansions, so by uniqueness coefficients add and scale, and φ is linear with $\varphi(v) = 1$. The price is Zorn's Lemma (Theorem 1.8.1): for $\mathbb{R}$ over $\mathbb{Q}$, no one has an explicit formula for such a functional.

Example. A Matrix Killed by Every Trace

Let $\mathbf{A} \in M_n(F)$ satisfy $\operatorname{tr}(\mathbf{AX}) = 0$ for every $\mathbf{X} \in M_n(F)$. Show that $\mathbf{A} = \mathbf{O}$. Deduce that every linear functional on $M_n(F)$ is $\mathbf{X} \mapsto \operatorname{tr}(\mathbf{AX})$ for exactly one $\mathbf{A} \in M_n(F)$.

Solution. The matrix units $\mathbf{E}_{ij}$ form a basis of $M_n(F)$ (Example (Standard Bases)), and $\mathbf{A} = \sum_{i,j} a_{ij}\mathbf{E}_{ij}$. By Theorem 5.1.3 (a), its dual basis consists of the coordinate functionals $\varepsilon_{ij}(\mathbf{A}) = a_{ij}$. We compute $\varepsilon_{ij}(\mathbf{A})$ as a trace. By Definition 0.9.9, $(\mathbf{AE}_{ji})_{kk} = \sum_l a_{kl}(\mathbf{E}_{ji})_{lk}$, and $(\mathbf{E}_{ji})_{lk} = 1$ only for $l = j$ and $k = i$. So the only nonzero diagonal entry of $\mathbf{AE}_{ji}$ is $(\mathbf{AE}_{ji})_{ii} = a_{ij}$, and

$$\operatorname{tr}(\mathbf{AE}_{ji}) = a_{ij} = \varepsilon_{ij}(\mathbf{A}).$$

Taking $\mathbf{X} = \mathbf{E}_{ji}$ in the hypothesis gives $\varepsilon_{ij}(\mathbf{A}) = 0$ for all i, j. By Theorem 5.1.5 (b), $\mathbf{A} = \mathbf{O}$.

For the deduction, let $\Theta\colon M_n(F) \to M_n(F)^*$, $\Theta(\mathbf{A})(\mathbf{X}) := \operatorname{tr}(\mathbf{AX})$. Each $\Theta(\mathbf{A})$ is linear in $\mathbf{X}$, and Θ is linear in $\mathbf{A}$, both by Theorem 0.9.6 (1) and the distributive laws of Theorem 0.9.4. We just showed $\ker\Theta = \{\mathbf{O}\}$, so Θ is injective (Theorem 2.2.3). Since $\dim M_n(F)^* = n^2 = \dim M_n(F)$ (Corollary 5.1.4), Θ is bijective by Corollary 2.3.2 (e). This is the matrix analogue of Theorem 5.1.1.

5.1.4 When V is infinite-dimensional

The main results of the last two subsections assumed finite dimension, or used a finite basis. That hypothesis does real work. Consider $F[x]$, with its basis $\{1, x, x^2, \dots\}$ (Example (A Basis of $F[x]$)). For $k \in \mathbb{N}$ let

$$\varphi_k(p) := \text{the coefficient of } x^k \text{ in } p.$$

These are the coordinate functionals, and $\varphi_k(x^j) = \delta_{kj}$, exactly as in Theorem 5.1.3. But they are not a basis of the dual space.

Proposition 5.1.6. The Coordinate Functionals Do Not Span $F[x]^*$

The functionals φ_k ($k \in \mathbb{N}$) lie in $F[x]^*$ and form a linearly independent set. But they do **not** span $F[x]^*$: the evaluation $\varepsilon_1 : p \mapsto p(1)$ is **not** a linear combination of finitely many φ_k.

Proof. Polynomials are added and scaled coefficient by coefficient (Definition 0.8.1), so each φ_k is linear. *Independence.* Let $k_1, \dots, k_m$ be distinct and $c_1\varphi_{k_1} + \cdots + c_m\varphi_{k_m} = 0$. Applying both sides to x^{k_j} gives $c_j = 0$, since $\varphi_{k_i}(x^{k_j}) = \delta_{k_i k_j}$.

Not spanning. The evaluation ε_1 is linear by Example (Four Families of Functionals) (a), and $\varepsilon_1(x^N) = 1^N = 1$ for every $N \in \mathbb{N}$. Suppose, for a contradiction, that $\varepsilon_1 = c_1\varphi_{k_1} + \cdots + c_m\varphi_{k_m}$ for some $m \in \mathbb{N}$, distinct k_j, and $c_j \in F$. Choose $N \in \mathbb{N}$ larger than every k_j. Then $\varphi_{k_j}(x^N) = 0$ for every j, so the right-hand side sends x^N to 0, while $\varepsilon_1(x^N) = 1$. This contradiction shows $\varepsilon_1 \notin \operatorname{span}\{\varphi_k : k \in \mathbb{N}\}$.

■

Where did the proof of Theorem 5.1.3 break? In the spanning step. The candidate $\sum_i \varphi(v_i)\varphi_i$ for $\varphi = \varepsilon_1$ would be $\varphi_0 + \varphi_1 + \varphi_2 + \cdots$, an infinite sum, and infinite sums mean nothing in a vector space (the warning in Chapter 1, §8). A functional on $F[x]$ may take nonzero values on infinitely many basis vectors, while a finite combination of the φ_k cannot.

Remark. More is true, and we state it without proof. For every infinite-dimensional V, a basis of V^* is **strictly larger**, as an infinite set, than a basis of V; in particular $V^* \ncong V$. So "$\dim V^* = \dim V$" is a genuinely finite-dimensional fact, and this is why finite dimension appears in the hypotheses of most theorems in this chapter. A first step towards seeing how large $F[x]^*$ is appears in Exercise C4 below.

5.1.5 Exercises

5.1.5.1 A. Check your understanding

Exercise. A1

(a) Define a linear functional on a vector space V over F, and the dual space V^*.

(b) True or false: $p \mapsto p(1)^2$ is a linear functional on $\mathbb{R}[x]$. Justify your answer.

(c) True or false: if (v_1, v_2) and (v_1, w_2) are bases of V, then the first functionals of their dual bases are equal. Justify your answer.

(d) What is $\dim M_{2\times 3}(F)^*$?

(e) State the formula that expresses $\varphi \in V^*$ in the dual basis of $(v_1, \dots, v_n)$.

(f) Name the move you would use to prove that a vector v of a finite-dimensional space is 0 without computing v itself.

Solution.

(a) See Definition 5.1.1 and Definition 5.1.2: a linear functional is a linear map $V \to F$,

and $V^* = \mathcal{L}(V, F)$ with pointwise addition and scaling.

(b) False. For $p = 1$ and $c = 2$, the polynomial $cp = 2$ is sent to $2^2 = 4$, while c times the image of p is $2 \cdot 1^2 = 2$. So the scaling condition fails.

(c) False. By the warning after Example (The Dual Basis Is Read Off the Inverse Matrix), in $\mathbb{R}^2$ the bases $((1,0),(0,1))$ and $((1,0),(1,1))$ have first dual functionals x and $x - y$.

(d) 6, by Corollary 5.1.4, since $\dim M_{2\times 3}(F) = 6$.

(e) $\varphi = \sum_{i=1}^{n} \varphi(v_i)\varphi_i$, by Theorem 5.1.3 (c).

(f) Test against functionals: show $\varphi(v) = 0$ for every $\varphi \in V^*$, or only for the functionals of one dual basis, and apply Theorem 5.1.5 (b).

5.1.5.2 B. Practice

Exercise. B1: Finding dual bases

(a) Find the dual basis of $((1,0,1),(0,1,1),(1,1,1))$ in $\mathbb{R}^3$.

(b) Show that the dual basis of $(1, x-1, (x-1)^2)$ in $\mathbb{R}[x]_{\le 2}$ is $p \mapsto p(1)$, $p \mapsto p'(1)$, $p \mapsto \frac{1}{2}p''(1)$. Hence write $p = 2 + 3x - x^2$ in this basis.

Solution.

(a) Let $\mathbf{P}$ have the three vectors as columns. Row reduction of $[\mathbf{P} \mid \mathbf{I}_3]$ (Theorem 3.4.6) gives

$$\mathbf{P} = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}, \qquad \mathbf{P}^{-1} = \begin{pmatrix} 0 & -1 & 1 \\ -1 & 0 & 1 \\ 1 & 1 & -1 \end{pmatrix},$$

and one checks $\mathbf{P}^{-1}\mathbf{P} = \mathbf{I}_3$. By Example (The Dual Basis Is Read Off the Inverse Matrix) (b), the dual basis is

$$\varphi_1(x,y,z) = -y + z, \qquad \varphi_2(x,y,z) = -x + z, \qquad \varphi_3(x,y,z) = x + y - z.$$

For instance $\varphi_3(1,0,1) = 0$, $\varphi_3(0,1,1) = 0$, and $\varphi_3(1,1,1) = 1$.

(b) The three maps are linear, as combinations of differentiation and evaluation at 1. Call them ψ_1, ψ_2, ψ_3. On the basis:

$$\psi_1(1) = 1,\ \psi_1(x-1) = 0,\ \psi_1((x-1)^2) = 0;$$
$$\psi_2(1) = 0,\ \psi_2(x-1) = 1,\ \psi_2((x-1)^2) = 2(1-1) = 0;$$
$$\psi_3(1) = 0,\ \psi_3(x-1) = 0,\ \psi_3((x-1)^2) = \tfrac{1}{2}\cdot 2 = 1.$$

So ψ_i takes the value δ_{ij} on the j-th basis vector, and by the uniqueness in Theorem 5.1.3 (a) they are the dual basis. (The list $(1, x-1, (x-1)^2)$ is a basis: it has distinct degrees $0, 1, 2$, so it is independent by Theorem 1.4.4, and it has length $3 = \dim \mathbb{R}[x]_{\le 2}$, so Theorem 1.6.4 applies.) For $p = 2+3x-x^2$: $p(1) = 4$, $p'(1) = 3-2 = 1$, $\frac{1}{2}p''(1) = -1$. By Theorem 5.1.3 (c),

$$p = 4 + (x-1) - (x-1)^2.$$

Check: $4 + x - 1 - (x^2 - 2x + 1) = 2 + 3x - x^2$.

Exercise. B2: Which maps are functionals?

Determine which of the following are linear functionals. Justify your answer.

(a) $\mathbb{R}^3 \to \mathbb{R}$, $(x, y, z) \mapsto 2x - z$.
(b) $\mathbb{R}^3 \to \mathbb{R}$, $(x, y, z) \mapsto xy$.
(c) $M_2(\mathbb{R}) \to \mathbb{R}$, $\mathbf{A} \mapsto a_{12} - 3a_{21}$.
(d) $\mathbb{R}[x] \to \mathbb{R}$, $p \mapsto \int_0^1 t\,p(t)\,dt$.
(e) $\mathbb{R}[x]_{\le 2} \to \mathbb{R}$, $p \mapsto p(0) + p'(0) + 1$.
(f) $\mathbb{C}^2 \to \mathbb{C}$, $(z, w) \mapsto z + \overline{w}$, where $\mathbb{C}^2$ is a vector space over $\mathbb{C}$.

Solution.

(a) Yes: it is $\varphi_{\mathbf{a}}$ with $\mathbf{a} = (2, 0, -1)$, linear by Theorem 5.1.1.
(b) No. $(1, 1, 0) \mapsto 1$, but $2(1, 1, 0) = (2, 2, 0) \mapsto 4 \neq 2$, so the scaling condition fails.
(c) Yes: it is $\varepsilon_{12} - 3\varepsilon_{21}$, a combination of coordinate functionals for the standard basis, and V^* is a vector space (Definition 5.1.2).
(d) Yes. By linearity of the integral, $\int_0^1 t(p+q)(t)\,dt = \int_0^1 tp(t)\,dt + \int_0^1 tq(t)\,dt$ and $\int_0^1 t(cp)(t)\,dt = c\int_0^1 tp(t)\,dt$.
(e) No. The zero polynomial is sent to $1 \neq 0$, contradicting Theorem 2.1.1.
(f) No. With $c = i$ and $(z, w) = (0, 1)$: $(0, i) \mapsto \overline{i} = -i$, while $i \cdot (0 + \overline{1}) = i$. The map is additive and respects real scalars, but it is not linear over $\mathbb{C}$.

Exercise. B3: Coordinates of a functional

(a) Let (φ_1, φ_2) be the dual basis of $((1, 1), (1, -1))$ from Example (The Dual Basis Is Read Off the Inverse Matrix). Write $\psi(x, y) = 3x + 5y$ as a combination of φ_1 and φ_2.
(b) Let $(\varphi_0, \varphi_1, \varphi_2)$ be the dual basis of $(1, x, x^2)$ from Example (Dual Bases in $\mathbb{R}[x]_{\le 2}$). Write $\iota(p) = \int_0^1 p(t)\,dt$ as a combination of them.

Solution.

(a) By Theorem 5.1.3 (c), $\psi = \psi(1,1)\varphi_1 + \psi(1,-1)\varphi_2 = 8\varphi_1 - 2\varphi_2$. Check: $8\cdot\frac{x+y}{2} - 2\cdot\frac{x-y}{2} = 4x + 4y - x + y = 3x + 5y$.
(b) By Theorem 5.1.3 (c), $\iota = \iota(1)\varphi_0 + \iota(x)\varphi_1 + \iota(x^2)\varphi_2 = \varphi_0 + \frac{1}{2}\varphi_1 + \frac{1}{3}\varphi_2$. Check: for $p = a_0 + a_1 x + a_2 x^2$, the right-hand side is $a_0 + \frac{a_1}{2} + \frac{a_2}{3} = \int_0^1 p$.

5.1.5.3 C. Going deeper

Exercise. C1: Functionals with the same kernel

Let V be **any** vector space over F, and let $\varphi, \psi \in V^*$ with $\ker\varphi = \ker\psi$. Prove that $\psi = c\varphi$ for some $c \in F$. Is the converse true?

Hint: if $\varphi \neq 0$, fix u with $\varphi(u) = 1$ and consider $v - \varphi(v)u$.

Solution. If $\varphi = 0$, then $\ker\psi = \ker\varphi = V$, so $\psi = 0 = 1 \cdot \varphi$. Suppose $\varphi \neq 0$. By Proposition 5.1.2 there is $u \in V$ with $\varphi(u) = 1$. Let $v \in V$. Then

$$\varphi(v - \varphi(v)u) = \varphi(v) - \varphi(v)\varphi(u) = 0,$$

so $v - \varphi(v)u \in \ker\varphi = \ker\psi$. Therefore $0 = \psi(v) - \varphi(v)\psi(u)$, that is, $\psi(v) = \psi(u)\varphi(v)$. Since v was arbitrary, $\psi = c\varphi$ with $c = \psi(u)$.

The converse holds only for $c \neq 0$: if $\psi = c\varphi$ with $c \neq 0$, then $\psi(v) = 0 \iff \varphi(v) = 0$. For $c = 0$ and $\varphi \neq 0$, $\ker\psi = V \neq \ker\varphi$.

Exercise. C2: Functionals vanishing where others vanish

Let V be a vector space over F, and let $\varphi_1, \ldots, \varphi_k, \psi \in V^*$.

(a) Prove that if $\ker\varphi_1 \cap \cdots \cap \ker\varphi_k \subseteq \ker\psi$, then $\psi \in \operatorname{span}(\varphi_1, \ldots, \varphi_k)$.

(b) Deduce that if V is finite-dimensional, then $(\varphi_1, \ldots, \varphi_k)$ spans V^* if and only if $\ker\varphi_1 \cap \cdots \cap \ker\varphi_k = \{0\}$.

Hint: for (a), consider the linear map $T\colon V \to F^k$, $v \mapsto (\varphi_1(v), \ldots, \varphi_k(v))$.

Solution.

(a) Let T be as in the hint; it is linear because each φ_i is. Its kernel is $\ker\varphi_1 \cap \cdots \cap \ker\varphi_k$. Define $S\colon \operatorname{im} T \to F$ by $S(Tv) := \psi(v)$. This is well defined: if $Tv = Tv'$, then $v - v' \in \ker T \subseteq \ker\psi$, so $\psi(v) = \psi(v')$. It is linear, since $S(cTv + Tv') = S(T(cv + v')) = \psi(cv + v') = cS(Tv) + S(Tv')$.
Now extend S to F^k. The subspace $\operatorname{im} T \subseteq F^k$ has a basis $(\mathbf{y}_1, \ldots, \mathbf{y}_r)$ (Theorem 1.6.7, Corollary 1.5.5), which extends to a basis $(\mathbf{y}_1, \ldots, \mathbf{y}_k)$ of F^k by Theorem 1.6.5. By Theorem 2.1.5 there is $\tilde{S} \in (F^k)^*$ with $\tilde{S}(\mathbf{y}_j) = S(\mathbf{y}_j)$ for $j \leq r$ and $\tilde{S}(\mathbf{y}_j) = 0$ for $j > r$. Then $\tilde{S}$ and S agree on a basis of $\operatorname{im} T$, so $\tilde{S} = S$ on $\operatorname{im} T$. By Theorem 5.1.1, $\tilde{S}(z_1, \ldots, z_k) = a_1 z_1 + \cdots + a_k z_k$ for some $a_i \in F$. Hence for every $v \in V$,

$$\psi(v) = S(Tv) = \tilde{S}(\varphi_1(v), \ldots, \varphi_k(v)) = a_1\varphi_1(v) + \cdots + a_k\varphi_k(v),$$

so $\psi = a_1\varphi_1 + \cdots + a_k\varphi_k$. Note that V was arbitrary; only F^k needed to be finite-dimensional.

(b) ($\Leftarrow$) If the intersection is $\{0\}$, it lies in $\ker\psi$ for every $\psi \in V^*$, so every ψ is in the span by (a). ($\Rightarrow$) Suppose the φ_i span V^*, and let v lie in every $\ker\varphi_i$. Every $\psi \in V^*$ is a combination $\sum c_i\varphi_i$, so $\psi(v) = \sum c_i\varphi_i(v) = 0$. By Theorem 5.1.5 (b), $v = 0$.

Exercise. C3: Evaluations as a dual basis

Let $n \in \mathbb{N}$.

(a) Let $x_0, \ldots, x_n \in F$ be **distinct**. Prove that the evaluations $(\varepsilon_{x_0}, \ldots, \varepsilon_{x_n})$ form a basis of $F[x]^*_{\leq n}$, and that they are the dual basis of some basis of $F[x]_{\leq n}$.

(b) Over $\mathbb{F}_2$ with $n = 2$, show that the evaluations at the elements of $\mathbb{F}_2$ do **not** span $\mathbb{F}_2[x]^*_{\leq 2}$.

Hint: for (a), use Theorem 3.7.3 to build the basis first.

Solution.

(a) For each j, Theorem 3.7.3 gives $\ell_j \in F[x]_{\le n}$ with $\ell_j(x_i) = \delta_{ij}$ for $i = 0, \dots, n$, because the x_i are distinct. If $\sum_j c_j \ell_j = 0$, evaluating at x_i gives $c_i = 0$. So $(\ell_0, \dots, \ell_n)$ is independent of length $n+1 = \dim F[x]_{\le n}$, hence a basis by Theorem 1.6.4. The evaluations are linear (Example (Four Families of Functionals)) and $\varepsilon_{x_i}(\ell_j) = \delta_{ij}$, so by the uniqueness in Theorem 5.1.3 (a) they are the dual basis of $(\ell_0, \dots, \ell_n)$, and by Theorem 5.1.3 (b) a basis of $F[x]^*_{\le n}$.

(b) $\mathbb{F}_2 = \{0, 1\}$ has only two elements, so only ε_0 and ε_1 are available, and $\dim \mathbb{F}_2[x]^*_{\le 2} = 3$ by Corollary 5.1.4. Two functionals cannot span a 3-dimensional space (Theorem 1.6.3). Concretely, $p = x^2 + x$ is nonzero but $p(0) = 0$ and $p(1) = 1 + 1 = 0$; the coefficient functional of x^2 takes the value 1 on p, so it is not a combination of $\varepsilon_0, \varepsilon_1$. The hypothesis "distinct points" in (a) needs $n+1$ distinct elements of F, which a small field may not have.

Exercise. C4: How large is $F[x]^*$?

Let $F^{\mathbb{N}}$ be the space of sequences with entries in F.

(a) Prove that $\Phi\colon F[x]^* \to F^{\mathbb{N}}$, $\varphi \mapsto (\varphi(1), \varphi(x), \varphi(x^2), \dots)$, is an isomorphism.

(b) Identify $\Phi(\varphi_k)$ for the coordinate functionals φ_k of Proposition 5.1.6, and $\Phi(\varepsilon_1)$. Explain how Proposition 5.1.6 matches Example (The Vectors e_i Do Not Span $F^{\mathbb{N}}$).

Hint: for surjectivity, remember that a polynomial has only finitely many nonzero coefficients.

Solution.

(a) *Linear.* The k-th entry of $\Phi(c\varphi + \psi)$ is $(c\varphi + \psi)(x^k) = c\varphi(x^k) + \psi(x^k)$, by the pointwise operations of Definition 5.1.2. *Injective.* If $\Phi(\varphi) = 0$, then φ vanishes on the basis $\{x^k\}$ of Example (A Basis of $F[x]$). Every polynomial is a finite combination of the x^k, so $\varphi = 0$ by Theorem 2.1.4; hence Φ is injective by Theorem 2.2.3. *Surjective.* Given $(a_0, a_1, \dots) \in F^{\mathbb{N}}$, define $\varphi(b_0 + b_1 x + \cdots + b_N x^N) := a_0 b_0 + \cdots + a_N b_N$. This is a finite sum, it is well defined because coefficients are unique, and it is linear because coefficients add and scale; it is the linear map with $\varphi(x^k) = a_k$ given by Theorem 2.1.6. Then $\Phi(\varphi) = (a_0, a_1, \dots)$. By Theorem 2.5.1, Φ is an isomorphism.

(b) $\Phi(\varphi_k) = (\delta_{k0}, \delta_{k1}, \dots) = e_k$, and $\Phi(\varepsilon_1) = (1, 1, 1, \dots)$. An isomorphism carries spans to spans, so $\varepsilon_1 \in \operatorname{span}\{\varphi_k\}$ would give $(1, 1, 1, \dots) \in \operatorname{span}\{e_k\}$. Example (The Vectors e_i Do Not Span $F^{\mathbb{N}}$) showed this is false, because finite combinations of the e_k have only finitely many nonzero entries. So $F[x]$, with its basis $\{x^k\}$ indexed by $\mathbb{N}$, has a dual as large as the whole sequence space $F^{\mathbb{N}}$.

5.2 Annihilators

Section 1 built the dual space V^*, the space of linear measurements on V. Now we connect subspaces of V with subspaces of V^*. A plane in $\mathbb{R}^3$ can be described by a spanning list or by

an equation, and an equation is a functional. This section collects all the equations that a set of vectors satisfies into one subspace of V^*, the annihilator, counts its dimension, and shows that it turns sums into intersections and intersections into sums.

5.2.1 The equations a subspace satisfies

Every object comes with its subobjects. The dual space V^* has subspaces of its own, and the question is which of them are attached to subspaces of V. The standard example shows the way.

Let $U = \operatorname{span}((1,0,1),(0,1,1))$, a plane in $\mathbb{R}^3$. The functional $\varphi(x,y,z) = x+y-z$ vanishes on both spanning vectors, so by linearity it vanishes on all of U: the plane satisfies the equation $x+y-z=0$. So do $2x+2y-2z=0$ and the trivial equation $0=0$. Which functionals vanish on U? By Theorem 5.1.1 a functional is $\psi(x,y,z) = ax+by+cz$. It vanishes on the two spanning vectors if and only if $a+c=0$ and $b+c=0$, that is, $\psi = -c(x+y-z) = -c\,\varphi$. So the equations of the plane form the **line** $\operatorname{span}(\varphi)$ in $(\mathbb{R}^3)^*$.

For the z-axis $\operatorname{span}((0,0,1))$, the same computation gives $c=0$: the equations are $ax+by=0$, a **plane** in $(\mathbb{R}^3)^*$. A bigger subspace satisfies fewer equations, and in both cases the dimensions add up to 3. The set of equations deserves a name, and the pattern deserves a theorem.

The annihilator of a set of vectors is the space of all linear equations that every one of those vectors satisfies.

Definition 5.2.1. Annihilator

Let V be a vector space over F, and let $S \subseteq V$ be **any subset**. The **annihilator** of S is

$$S^0 := \{\varphi \in V^* : \varphi(s) = 0 \text{ for all } s \in S\}.$$

In words: S^0 is a subset of the **dual space** V^*, not of V. A functional belongs to it when it vanishes at **every** vector of S; a single nonzero value disqualifies it. The set S needs no structure: it may be a subspace, a finite list of vectors, or even empty. The superscript 0 recalls the value that the functionals take on S.

Well-definedness. The definition makes sense for every subset. What needs checking is that the result is a subspace, which is what makes its dimension meaningful.

Proposition 5.2.1. The Annihilator Is a Subspace

Let V be a vector space over F, and let $S \subseteq V$. Then S^0 is a subspace of V^*.

Proof. We use the subspace test (Theorem 1.2.1) in V^*. The zero functional sends every $s \in S$ to 0, so it lies in S^0. Let $\varphi, \psi \in S^0$ and $c \in F$. For every $s \in S$, the pointwise operations of Definition 5.1.2 give

$$(\varphi+\psi)(s) = \varphi(s)+\psi(s) = 0+0 = 0, \qquad (c\varphi)(s) = c\,\varphi(s) = c\cdot 0 = 0.$$

Hence $\varphi+\psi \in S^0$ and $c\varphi \in S^0$, and S^0 is a subspace of V^*.

■

Example. First Annihilators

(a) For any vector space V, find $\{0\}^0$, $\varnothing^0$, and V^0.

(b) In $M_2(F)$, find the annihilator of the single matrix $\mathbf{E}_{11}$.

(c) In $\mathbb{R}[x]_{\leq 2}$, find the annihilator of $S = \{1, x\}$.

Solution.

(a) Every functional sends 0 to 0 (Theorem 2.1.1), so $\{0\}^0 = V^*$. The condition "for all $s \in \varnothing$" holds vacuously, so $\varnothing^0 = V^*$. A functional in V^0 takes the value 0 at every vector, so it is the zero functional, and $V^0 = \{0\}$. These degenerate cases fix the direction of everything that follows: the smallest subspace of V has the largest annihilator, and the largest has the smallest.

(b) Let $\varepsilon_{ij}(\mathbf{A}) = a_{ij}$ be the dual basis of the standard basis $(\mathbf{E}_{11}, \mathbf{E}_{12}, \mathbf{E}_{21}, \mathbf{E}_{22})$. By Theorem 5.1.3 (c), every $\psi \in M_2(F)^*$ is $\psi = \sum_{i,j} \psi(\mathbf{E}_{ij})\, \varepsilon_{ij}$. Such a ψ lies in $\{\mathbf{E}_{11}\}^0$ exactly when $\psi(\mathbf{E}_{11}) = 0$. Hence

$$\{\mathbf{E}_{11}\}^0 = \operatorname{span}(\varepsilon_{12}, \varepsilon_{21}, \varepsilon_{22}),$$

the functionals $\mathbf{A} \mapsto c_{12}a_{12} + c_{21}a_{21} + c_{22}a_{22}$. It has dimension 3, and $3 + 1 = 4 = \dim M_2(F)$.

(c) Let $(\varphi_0, \varphi_1, \varphi_2)$ be the dual basis of $(1, x, x^2)$, so φ_k reads off the coefficient of x^k. A functional $\psi = \psi(1)\varphi_0 + \psi(x)\varphi_1 + \psi(x^2)\varphi_2$ lies in S^0 exactly when $\psi(1) = \psi(x) = 0$, that is, when $\psi = \psi(x^2)\varphi_2$. So $S^0 = \operatorname{span}(\varphi_2)$: the only equations satisfied by both 1 and x are the multiples of "the coefficient of x^2 is 0".

Nonexample by minimal change. Shift the plane of the opening example off the origin: $A = \{(x, y, z) \in \mathbb{R}^3 : x + y - z = 1\}$. The set is still cut out by the same left-hand side, but $\varphi(x, y, z) = x + y - z$ takes the value 1, not 0, on A, so the defining clause $\varphi(s) = 0$ fails and $\varphi \notin A^0$. In fact $A^0 = \{0\}$: the set A contains $(1, 0, 0)$, $(0, 1, 0)$, and $(0, 0, -1)$, and a functional $ax + by + cz$ vanishing at these three vectors has $a = b = c = 0$. Annihilators record only **homogeneous** equations.

Why this definition. We allow any subset S, not only subspaces, because subspaces are usually handed to us by spanning lists, and we want to test a functional on the list alone. The theorem at the end of this section shows that this loses nothing: $S^0 = (\operatorname{span} S)^0$. The name is literal. To annihilate is to reduce to nothing, and the functionals in S^0 reduce every vector of S to 0.

Warning. **U^0 lives in V^*, and it is not a complement of U.** It is tempting to turn the functional $ax+by$ into the vector (a, b) and to picture U^0 as a subspace of F^2 "perpendicular" to U. Over $\mathbb{F}_2$ this picture fails completely. Let $U = \operatorname{span}((1, 1)) \subseteq \mathbb{F}_2^2$. A functional $ax + by$ vanishes on $(1, 1)$ exactly when $a + b = 0$, that is, $a = b$, so $U^0 = \{0,\ x + y\}$. The functional $x+y$ corresponds to the vector $(1, 1)$, which lies **in** U. So the "picture" of U^0 is U itself, not a complement. Perpendicularity needs an inner product, which is the orthogonal complement of Chapter 11; the annihilator needs nothing.

The payoff of the definition is the pattern we saw in the opening example. The equations of a

subspace take up exactly the dimensions that the subspace leaves free.

5.2.2 Counting equations

Theorem 5.2.2. Dimension of the Annihilator

Let V be a finite-dimensional vector space over F, and let U be a subspace of V. Then

$$\dim U + \dim U^0 = \dim V.$$

More precisely, if $(u_1, \dots, u_m, v_{m+1}, \dots, v_n)$ is a basis of V extending a basis $(u_1, \dots, u_m)$ of U, with dual basis $(\varphi_1, \dots, \varphi_n)$, then $(\varphi_{m+1}, \dots, \varphi_n)$ is a basis of U^0.

Idea. This is a dimension formula, so we start from the smallest subspace, U, take a basis and extend it to V. The dual basis now splits in two. The last $n - m$ functionals vanish on every u_i, by the delta, so they vanish on U. The first m do not. To see that nothing else is in U^0, expand a functional $\psi \in U^0$ in the dual basis: its coefficients are its values on the basis, and the first m values are $\psi(u_i) = 0$.

Proof. Since U is a subspace of the finite-dimensional space V, it has a basis $(u_1, \dots, u_m)$ (Theorem 1.6.7, Corollary 1.5.5), and by the Basis Extension Theorem (Theorem 1.6.5) it extends to a basis $\mathcal{B} = (u_1, \dots, u_m, v_{m+1}, \dots, v_n)$ of V, where $n = \dim V$. Let $(\varphi_1, \dots, \varphi_n)$ be the dual basis of $\mathcal{B}$ (Theorem 5.1.3).

The φ_j with $j > m$ lie in U^0. Let $j > m$ and $u \in U$, say $u = a_1 u_1 + \cdots + a_m u_m$. By Theorem 2.1.4, $\varphi_j(u) = \sum_{i=1}^{m} a_i \varphi_j(u_i) = 0$, since $\varphi_j(u_i) = \delta_{ji} = 0$ for $i \le m < j$.

They span U^0. Let $\psi \in U^0$. By Theorem 5.1.3 (c),

$$\psi = \sum_{i=1}^{m} \psi(u_i)\, \varphi_i + \sum_{j=m+1}^{n} \psi(v_j)\, \varphi_j = \sum_{j=m+1}^{n} \psi(v_j)\, \varphi_j,$$

because $\psi(u_i) = 0$ for $u_i \in U$. Hence $U^0 \subseteq \operatorname{span}(\varphi_{m+1}, \dots, \varphi_n)$, and the reverse inclusion holds because U^0 is a subspace (Proposition 5.2.1, Theorem 1.3.1).

They are independent, being part of the basis $(\varphi_1, \dots, \varphi_n)$ of V^*. Therefore $(\varphi_{m+1}, \dots, \varphi_n)$ is a basis of U^0, and $\dim U^0 = n - m = \dim V - \dim U$. This proves the theorem. ∎

Alternatively, restrict functionals to U. The map $R\colon V^* \to U^*$, $\varphi \mapsto \varphi|_U$, is linear, and its kernel is U^0 by definition. It is surjective: given $\xi \in U^*$, Theorem 2.1.5 gives $\varphi \in V^*$ with $\varphi(u_i) = \xi(u_i)$ and $\varphi(v_j) = 0$, and then $\varphi|_U = \xi$ because both agree on a basis of U. By the Rank–Nullity Theorem (Theorem 2.3.1) and Corollary 5.1.4, $\dim V = \dim V^* = \dim U^0 + \dim U^* = \dim U^0 + \dim U$. This route shows **why** the formula holds: a functional on V is a functional on U plus a choice of values off U, and U^0 measures that choice.

The theorem is the second signature move of the chapter: **pass to the dual and count.** One immediate consequence: if U is a **proper** subspace of a finite-dimensional V, then $\dim U^0 > 0$,

so some **nonzero** functional vanishes on U. Equivalently, a list spans a finite-dimensional V if and only if the only functional that kills every vector of the list is 0: apply the previous sentence to the span U of the list, and note that a functional kills every vector of the list exactly when it kills U, by Theorem 2.1.4.

In F^n the whole computation becomes elimination, because functionals are rows.

Example. An Annihilator as a Null Space

Find a basis of U^0 for $U = \operatorname{span}((1,2,0),(0,1,1)) \subseteq \mathbb{R}^3$, and check the dimension formula.

Solution. By Theorem 5.1.1, every functional is $\varphi_{\mathbf{a}}(\mathbf{x}) = \mathbf{a}^\top \mathbf{x}$ for a unique $\mathbf{a} = (a_1, a_2, a_3)$. By linearity, $\varphi_{\mathbf{a}}$ vanishes on U if and only if it vanishes on the two spanning vectors (a combination of vectors killed by $\varphi_{\mathbf{a}}$ is killed, by Theorem 2.1.4). These two conditions are

$$a_1 + 2a_2 = 0, \qquad a_2 + a_3 = 0,$$

$$\text{that is,} \qquad \mathbf{A}\mathbf{a} = \mathbf{0} \quad \text{with} \quad \mathbf{A} = \begin{pmatrix} 1 & 2 & 0 \\ 0 & 1 & 1 \end{pmatrix},$$

the matrix whose **rows** are the spanning vectors. So $\varphi_{\mathbf{a}} \in U^0$ if and only if $\mathbf{a} \in \operatorname{nul}(\mathbf{A})$. Subtracting twice row 2 from row 1 gives the RREF

$$\begin{pmatrix} 1 & 0 & -2 \\ 0 & 1 & 1 \end{pmatrix},$$

with a_3 free. By Theorem 3.5.4, $\operatorname{nul}(\mathbf{A}) = \operatorname{span}((2,-1,1))$. Hence $U^0 = \operatorname{span}(\psi)$ with

$$\psi(x,y,z) = 2x - y + z.$$

Check: $\psi(1,2,0) = 2 - 2 + 0 = 0$ and $\psi(0,1,1) = -1 + 1 = 0$. The two spanning vectors are independent (neither is a multiple of the other), so $\dim U = 2$, and $\dim U + \dim U^0 = 2 + 1 = 3$, as Theorem 5.2.2 predicts. As a by-product, U is the plane $2x - y + z = 0$: $U \subseteq \ker \psi$, both have dimension 2 (Proposition 5.1.2), and Theorem 1.6.8 gives equality.

The same recipe works in general. For $U = \operatorname{span}(\mathbf{u}_1, \dots, \mathbf{u}_m) \subseteq F^n$, let $\mathbf{A}$ have rows $\mathbf{u}_1^\top, \dots, \mathbf{u}_m^\top$. Then $U^0 = \{\varphi_{\mathbf{a}} : \mathbf{a} \in \operatorname{nul}(\mathbf{A})\}$, and the isomorphism of Theorem 5.1.1 carries a basis of $\operatorname{nul}(\mathbf{A})$ to a basis of U^0.

Away from F^n, counting often replaces the computation entirely.

Example. The Equations of a Root Condition

Let $U = \{p \in \mathbb{R}[x]_{\le 2} : p(1) = 0\}$. Find U^0.

Solution. Let $\varepsilon_1(p) = p(1)$, a linear functional (Example (Four Families of Functionals)). By definition $U = \ker \varepsilon_1$, so $\varepsilon_1 \in U^0$. Since $\varepsilon_1(1) = 1$, $\varepsilon_1 \neq 0$, and Proposition 5.1.2 gives $\dim U = 3 - 1 = 2$. By Theorem 5.2.2, $\dim U^0 = 3 - 2 = 1$. Now $\operatorname{span}(\varepsilon_1)$ is a subspace of U^0 (Proposition 5.2.1) of dimension 1, so $U^0 = \operatorname{span}(\varepsilon_1)$ by Theorem 1.6.8.

We never found a basis of U and never solved an equation. The conclusion is concrete: a functional on $\mathbb{R}[x]_{\leq 2}$ vanishes on every polynomial with a root at 1 **only if** it is a multiple of $p \mapsto p(1)$.

Quick check. Let $U = \operatorname{span}((1,1,1)) \subseteq \mathbb{R}^3$. What is $\dim U^0$? Name a basis of U^0 without solving a system.

Solution. $\dim U^0 = 3 - 1 = 2$ by Theorem 5.2.2. The functionals $x - y$ and $y - z$ vanish at $(1,1,1)$, hence on U, and they are independent: in $a(x-y) + b(y-z) = 0$, the coefficient of x gives $a = 0$, and then $b = 0$. Two independent vectors in a 2-dimensional space form a basis (Theorem 1.6.4).

5.2.3 Sums, intersections, and spans

Taking annihilators reverses inclusions: more vectors satisfy fewer common equations. It also exchanges sums with intersections. In $\mathbb{R}^3$, let U be the xy-plane and W the xz-plane. Then $U^0 = \operatorname{span}(z)$ and $W^0 = \operatorname{span}(y)$, writing x, y, z for the coordinate functionals. The intersection $U \cap W$ is the x-axis, with annihilator $\operatorname{span}(y,z) = U^0 + W^0$. The sum $U + W = \mathbb{R}^3$ has annihilator $\{0\} = U^0 \cap W^0$. The theorem says this always happens.

Theorem 5.2.3. Properties of Annihilators

Let V be a vector space over F.

(a) If $S \subseteq T \subseteq V$, then $T^0 \subseteq S^0$.

(b) For every subset $S \subseteq V$, $S^0 = (\operatorname{span} S)^0$.

(c) For subspaces U, W of V, $(U + W)^0 = U^0 \cap W^0$.

(d) If V is **finite-dimensional**, then for subspaces U, W of V, $(U \cap W)^0 = U^0 + W^0$.

Idea. Parts (a)–(c) are unwinding definitions: fix a functional in one side and test it on the vectors of the other. In (d), the inclusion $U^0 + W^0 \subseteq (U \cap W)^0$ is also direct. The reverse inclusion is the hard half. Given φ vanishing on $U \cap W$, we would have to **split** it as $\psi + \chi$ with ψ killing U and χ killing W, and nothing in sight tells us how. So we count instead: one inclusion plus equal dimensions gives equality. The plan is ① compute $\dim(U^0 + W^0)$ by the dimension formula for sums inside V^*; ② replace $U^0 \cap W^0$ by $(U+W)^0$ using (c); ③ turn every annihilator dimension into a dimension in V with Theorem 5.2.2; ④ recognize the dimension formula in V. Finite dimension is spent in step ③.

Proof.

(a) Let $\varphi \in T^0$. Every $s \in S$ lies in T, so $\varphi(s) = 0$. Hence $\varphi \in S^0$.

(b) Since $S \subseteq \operatorname{span} S$ (Theorem 1.3.1), part (a) gives $(\operatorname{span} S)^0 \subseteq S^0$. Conversely, let $\varphi \in S^0$ and $v \in \operatorname{span} S$, say $v = a_1 s_1 + \cdots + a_k s_k$ with $s_i \in S$ and $k \geq 0$. By

Theorem 2.1.4, $\varphi(v) = \sum_i a_i\varphi(s_i) = 0$. Hence $\varphi \in (\operatorname{span} S)^0$.

(c) Since $U \subseteq U + W$ and $W \subseteq U + W$ (Theorem 1.7.1), part (a) gives $(U + W)^0 \subseteq U^0$ and $(U + W)^0 \subseteq W^0$, so $(U + W)^0 \subseteq U^0 \cap W^0$. Conversely, let $\varphi \in U^0 \cap W^0$, and let $u + w \in U + W$ with $u \in U$, $w \in W$. Then $\varphi(u + w) = \varphi(u) + \varphi(w) = 0 + 0 = 0$. Hence $\varphi \in (U + W)^0$.

(d) ($\supseteq$) Let $\varphi = \psi + \chi$ with $\psi \in U^0$ and $\chi \in W^0$, and let $v \in U \cap W$. Since $v \in U$ and $v \in W$, $\varphi(v) = \psi(v) + \chi(v) = 0 + 0 = 0$. Hence $U^0 + W^0 \subseteq (U \cap W)^0$.

($\subseteq$) Let $n = \dim V$. The subspaces U, W, $U \cap W$ (Theorem 1.2.2), and $U + W$ (Theorem 1.7.1) of V are finite-dimensional by Theorem 1.6.7. The space V^* has dimension n by Corollary 5.1.4, so its subspaces U^0, W^0 are finite-dimensional too. By the Dimension Formula for Sums (Theorem 1.7.3) in V^*, then (c),

$$\begin{aligned}\dim(U^0 + W^0) &= \dim U^0 + \dim W^0 - \dim(U^0 \cap W^0) \\ &= \dim U^0 + \dim W^0 - \dim(U + W)^0.\end{aligned}$$

By Theorem 5.2.2 applied to U, W, and $U + W$, the right-hand side is

$$\begin{aligned}&(n - \dim U) + (n - \dim W) - (n - \dim(U + W)) \\ &\quad = n - (\dim U + \dim W - \dim(U + W)) \\ &\quad = n - \dim(U \cap W),\end{aligned}$$

where the last equality is Theorem 1.7.3 in V. By Theorem 5.2.2 once more, $n - \dim(U \cap W) = \dim(U \cap W)^0$. So $U^0 + W^0$ is a subspace of $(U \cap W)^0$ (by ($\supseteq$) and Theorem 1.7.1) with the same finite dimension, and Theorem 1.6.8 gives $U^0 + W^0 = (U \cap W)^0$. This proves the theorem.

■

Part (b) is what made the computations of this section legitimate: to find the annihilator of a subspace, test functionals on a spanning list only. Parts (c) and (d) form a dictionary. A subspace given as a **sum** has equations obtained by **intersecting** equation spaces; a subspace given as an **intersection** has equations obtained by **adding** them. Later in this chapter this dictionary becomes a practical method for computing sums and intersections of subspaces.

Remark. Parts (a)–(c) and the inclusion ($\supseteq$) in (d) hold in every vector space; only the counting step used finite dimension. The equality (d) is in fact true for **every** V, but the proof must then construct the splitting $\varphi = \psi + \chi$ by hand; Exercise C1 below does this. In finite dimension, (a) also becomes strict: if $U \subsetneq W$, then $\dim U < \dim W$ (Theorem 1.6.8), so $\dim W^0 < \dim U^0$ and $W^0 \subsetneq U^0$.

5.2.4 Exercises

5.2.4.1 A. Check your understanding

Exercise. A1

(a) Define the annihilator S^0 of a subset S of a vector space V.

(b) True or false: U^0 is a subspace of V. Justify your answer.

(c) Let U be a 2-dimensional subspace of $\mathbb{R}^5$. What is $\dim U^0$?

(d) For a finite-dimensional V, what are $\{0\}^0$ and V^0?

(e) True or false: if $U \subsetneq W$ are subspaces of a finite-dimensional V, then $W^0 \subsetneq U^0$. Justify your answer.

(f) To compute U^0 for $U = \operatorname{span}(\mathbf{u}_1, \dots, \mathbf{u}_m) \subseteq F^n$, which matrix do you row reduce, and why?

Solution.

(a) $S^0 = \{\varphi \in V^* : \varphi(s) = 0 \text{ for all } s \in S\}$ (Definition 5.2.1).

(b) False. U^0 is a subspace of the dual space V^* (Proposition 5.2.1); its elements are functionals, not vectors of V.

(c) $\dim U^0 = 5 - 2 = 3$, by Theorem 5.2.2.

(d) $\{0\}^0 = V^*$ and $V^0 = \{0\}$, by Example (First Annihilators) (a).

(e) True. By Theorem 5.2.3 (a), $W^0 \subseteq U^0$. Since $U \subsetneq W$, $\dim U < \dim W$ by Theorem 1.6.8, so $\dim W^0 = \dim V - \dim W < \dim V - \dim U = \dim U^0$. Hence the inclusion is strict.

(f) The matrix $\mathbf{A}$ whose **rows** are $\mathbf{u}_1^\top, \dots, \mathbf{u}_m^\top$. By Theorem 5.1.1 a functional is $\mathbf{x} \mapsto \mathbf{a}^\top \mathbf{x}$, and it kills every $\mathbf{u}_i$ exactly when $\mathbf{A}\mathbf{a} = \mathbf{0}$; so U^0 corresponds to $\operatorname{nul}(\mathbf{A})$ (Example (An Annihilator as a Null Space)).

5.2.4.2 B. Practice

Exercise. B1: Computing annihilators

(a) Let $U = \operatorname{span}((1,2,1,0),(2,4,3,1),(1,2,2,1)) \subseteq \mathbb{R}^4$. Find a basis of U^0, and verify that $\dim U + \dim U^0 = 4$.

(b) Let $U = \{p \in \mathbb{R}[x]_{\le 3} : p(1) = p(-1) = 0\}$. Show that $U^0 = \operatorname{span}(\varepsilon_1, \varepsilon_{-1})$, where $\varepsilon_c(p) = p(c)$.

Solution.

(a) Let $\mathbf{A}$ have the three spanning vectors as rows. As in Example (An Annihilator as a Null Space), $\varphi_{\mathbf{a}} \in U^0$ if and only if $\mathbf{A}\mathbf{a} = \mathbf{0}$. Subtracting 2 times row 1 from row 2 and row 1 from row 3 gives rows $(1,2,1,0)$, $(0,0,1,1)$, $(0,0,1,1)$; subtracting row 2 from row 3 and then row 2 from row 1 gives the RREF

$$\begin{pmatrix} 1 & 2 & 0 & -1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

The free variables are a_2 and a_4, and Theorem 3.5.4 gives

$$\operatorname{nul}(\mathbf{A}) = \operatorname{span}((-2,1,0,0),(1,0,-1,1)).$$

Hence U^0 has the basis

$$\psi_1(x_1,x_2,x_3,x_4) = -2x_1 + x_2, \qquad \psi_2(x_1,x_2,x_3,x_4) = x_1 - x_3 + x_4.$$

Check on the first two spanning vectors: $\psi_1(1,2,1,0) = 0$, $\psi_2(1,2,1,0) = 0$, $\psi_1(2,4,3,1) = 0$, $\psi_2(2,4,3,1) = 2-3+1 = 0$; the third vector is their difference, $(2,4,3,1)-(1,2,1,0)$. The RREF has two pivots, so $\dim U = \dim \operatorname{row}(\mathbf{A}) = 2$ (Theorem 3.5.5), and $2+2=4$.

(b) By definition $\varepsilon_1, \varepsilon_{-1} \in U^0$. *Independent:* if $a\varepsilon_1 + b\varepsilon_{-1} = 0$, applying it to $x+1$ gives $2a = 0$ and applying it to $x-1$ gives $-2b = 0$. *Dimension of U:* the map $T\colon \mathbb{R}[x]_{\leq 3} \to \mathbb{R}^2$, $p \mapsto (p(1), p(-1))$, is linear with kernel U, and its image contains $T(x+1) = (2,0)$ and $T(x-1) = (0,-2)$, which span $\mathbb{R}^2$. So $\operatorname{rank} T = 2$ and $\dim U = 4-2 = 2$ by Theorem 2.3.1. By Theorem 5.2.2, $\dim U^0 = 2$. Hence the 2-dimensional subspace $\operatorname{span}(\varepsilon_1, \varepsilon_{-1})$ of U^0 is all of U^0 (Theorem 1.6.8).

Exercise. B2: Equations of the symmetric matrices

Let $U = \{\mathbf{A} \in M_2(\mathbb{R}) : \mathbf{A}^\top = \mathbf{A}\}$. Find U^0, and verify the dimension formula.

Solution. Every symmetric matrix is $a\mathbf{E}_{11} + b(\mathbf{E}_{12}+\mathbf{E}_{21}) + c\mathbf{E}_{22}$, so $U = \operatorname{span}(\mathbf{E}_{11}, \mathbf{E}_{12}+\mathbf{E}_{21}, \mathbf{E}_{22})$. Let (ε_{ij}) be the dual basis of the standard basis, so that every $\psi \in M_2(\mathbb{R})^*$ is $\psi = \sum_{i,j} c_{ij}\varepsilon_{ij}$ with $c_{ij} = \psi(\mathbf{E}_{ij})$ (Theorem 5.1.3 (c)). By Theorem 5.2.3 (b), $\psi \in U^0$ if and only if ψ kills the three spanning matrices:

$$c_{11} = 0, \qquad c_{12} + c_{21} = 0, \qquad c_{22} = 0.$$

So $\psi = c_{12}(\varepsilon_{12} - \varepsilon_{21})$, and $U^0 = \operatorname{span}(\varepsilon_{12} - \varepsilon_{21})$: the only equation of the symmetric matrices, up to scaling, is $a_{12} - a_{21} = 0$. The three spanning matrices are independent (they involve disjoint sets of entries), so $\dim U = 3$, and $\dim U + \dim U^0 = 3+1 = 4 = \dim M_2(\mathbb{R})$.

Exercise. B3: The equations of a hyperplane

Let V be a finite-dimensional vector space over F, and let $\varphi \in V^*$ be nonzero. Prove that $(\ker \varphi)^0 = \operatorname{span}(\varphi)$.

Solution. By definition of the kernel, φ vanishes on $\ker \varphi$, so $\varphi \in (\ker \varphi)^0$, and $\operatorname{span}(\varphi) \subseteq (\ker \varphi)^0$ by Proposition 5.2.1. Since $\varphi \neq 0$, $\dim \operatorname{span}(\varphi) = 1$, and Proposition 5.1.2 gives $\dim \ker \varphi = \dim V - 1$. By Theorem 5.2.2, $\dim(\ker \varphi)^0 = \dim V - (\dim V - 1) = 1$. A 1-dimensional subspace of a 1-dimensional space is the whole space (Theorem 1.6.8), so $(\ker \varphi)^0 = \operatorname{span}(\varphi)$.

5.2.4.3 C. Going deeper

Exercise. C1: The intersection formula without counting

Let U, W be subspaces of a vector space V over F, and let $\varphi \in (U \cap W)^0$.

(a) Suppose V is finite-dimensional. Construct $\psi \in U^0$ and $\chi \in W^0$ with $\varphi = \psi + \chi$, without using any dimension formula.

(b) Explain how to modify the construction when V is infinite-dimensional. Deduce that $(U \cap W)^0 = U^0 + W^0$ holds in every vector space.

Hint: extend a basis of $U \cap W$ to a basis of U and, separately, to a basis of W.

Solution.

(a) Let $(x_1, \dots, x_k)$ be a basis of $U \cap W$ (Theorem 1.2.2, Theorem 1.6.7). Extend it to a basis $(x_1, \dots, x_k, u_1, \dots, u_p)$ of U and to a basis $(x_1, \dots, x_k, w_1, \dots, w_q)$ of W (Theorem 1.6.5).
The combined list $(x_1, \dots, x_k, u_1, \dots, u_p, w_1, \dots, w_q)$ is independent. Let $\sum a_i x_i + \sum b_i u_i + \sum c_j w_j = 0$. Then $\sum c_j w_j = -\sum a_i x_i - \sum b_i u_i$ lies in W and in U, hence in $U \cap W$, so $\sum c_j w_j = \sum d_i x_i$ for some $d_i \in F$. Since (x, w) is a basis of W, all c_j and d_i are 0. Then $\sum a_i x_i + \sum b_i u_i = 0$, and since (x, u) is a basis of U, all a_i and b_i are 0.
Extend the combined list to a basis of V by vectors $y_1, \dots, y_r$ (Theorem 1.6.5). By Theorem 2.1.5 there is $\psi \in V^*$ with

$$\psi(x_i) = 0, \qquad \psi(u_i) = 0, \qquad \psi(w_j) = \varphi(w_j), \qquad \psi(y_l) = 0.$$

Then ψ kills a basis of U, so $\psi \in U^0$ (Theorem 5.2.3 (b)). Put $\chi := \varphi - \psi$. Then $\chi(x_i) = \varphi(x_i) - 0 = 0$, because $x_i \in U \cap W$ and $\varphi \in (U \cap W)^0$; and $\chi(w_j) = \varphi(w_j) - \varphi(w_j) = 0$. So χ kills a basis of W, and $\chi \in W^0$. Hence $\varphi = \psi + \chi \in U^0 + W^0$.

(b) Replace the finite lists by sets. By Theorem 1.8.3, $U \cap W$ has a basis X, and by Theorem 1.8.2 it extends to a basis $X \cup P$ of U and to a basis $X \cup Q$ of W, with $P \cap X = Q \cap X = \varnothing$. First, $P \cap Q = \varnothing$: a vector of $P \cap Q$ lies in $U \cap W = \operatorname{span} X$, which contradicts the independence of $X \cup P$. So every vector of $X \cup P \cup Q$ lies in exactly one of X, P, Q. A linear relation among finitely many distinct vectors of $X \cup P \cup Q$ splits into an X-part, a P-part, and a Q-part, and the independence argument of (a) applies word for word; so $X \cup P \cup Q$ is independent. By Theorem 1.8.2 it extends to a basis B of V. Every $v \in V$ is a finite combination $v = \sum_b c_b b$ of distinct vectors of B, with unique coefficients because B is independent. Define

$$\psi(v) := \sum_{b \in Q} c_b \, \varphi(b),$$

a finite sum over the vectors of Q that occur. Adding or scaling expansions gives expansions, so by uniqueness the coefficients add and scale, and ψ is linear. It vanishes on $X \cup P$ and agrees with φ on Q, so the rest of (a) is unchanged: $\psi \in U^0$ and $\chi = \varphi - \psi$ kills the basis $X \cup Q$ of W. Together with the inclusion ($\supseteq$) of Theorem 5.2.3 (d), which used no finiteness, this gives $(U \cap W)^0 = U^0 + W^0$ in every vector space.

Exercise. C2: Equations of a double root

Let $U = \{p \in \mathbb{R}[x]_{\le 3} : p(1) = 0 \text{ and } p'(1) = 0\}$, and let $\varepsilon_1(p) = p(1)$ and $\eta_1(p) = p'(1)$. Prove that $U^0 = \operatorname{span}(\varepsilon_1, \eta_1)$. Deduce that $\psi(p) = p(0) - p(2) + 2p'(1)$ does **not** vanish on U.

Solution. Both ε_1 and η_1 are linear (evaluation, and differentiation followed by evaluation), and they vanish on U by definition. *Independent:* if $a\varepsilon_1 + b\eta_1 = 0$, applying it to 1 gives

$a = 0$, and applying it to $x-1$ gives $b = 0$. *Dimension of U:* the linear map $T\colon \mathbb{R}[x]_{\leq 3} \to \mathbb{R}^2$, $p \mapsto (p(1), p'(1))$, has kernel U, and $T(1) = (1,0)$, $T(x-1) = (0,1)$, so T is surjective. By Theorem 2.3.1, $\dim U = 4 - 2 = 2$, and by Theorem 5.2.2, $\dim U^0 = 2$. The 2-dimensional subspace $\operatorname{span}(\varepsilon_1, \eta_1)$ of U^0 is therefore all of U^0 (Theorem 1.6.8).

For the deduction, suppose $\psi \in U^0$. Then $\psi = a\varepsilon_1 + b\eta_1$. Applying both sides to 1 gives $1 - 1 + 0 = a$, so $a = 0$; applying them to $x - 1$ gives $-1 - 1 + 2 = b$, so $b = 0$. Then $\psi = 0$. But $\psi(x^2) = 0 - 4 + 4 = 0$ and $\psi(x^3) = 0 - 8 + 6 = -2 \neq 0$, a contradiction. Hence $\psi \notin U^0$. Check with an explicit witness: $p = x(x-1)^2 \in U$ has $\psi(p) = 0 - 2 + 0 = -2 \neq 0$.

Exercise. C3: Dual of a direct sum decomposition

(a) Let V be finite-dimensional with $V = U \oplus W$. Prove that $V^* = U^0 \oplus W^0$.

(b) In $\mathbb{R}^2$, let $U = \operatorname{span}((1,0))$ and $W = \operatorname{span}((1,1))$. Write $\psi(x,y) = x$ as $\psi_U + \psi_W$ with $\psi_U \in U^0$ and $\psi_W \in W^0$.

Solution.

(a) Since the sum is direct, $U \cap W = \{0\}$ (Theorem 1.7.4), and $U + W = V$. By Theorem 5.2.3 (c) and Example (First Annihilators) (a), $U^0 \cap W^0 = (U+W)^0 = V^0 = \{0\}$. By Theorem 5.2.3 (d), $U^0 + W^0 = (U \cap W)^0 = \{0\}^0 = V^*$. So $V^* = U^0 + W^0$, and the sum is direct by Theorem 1.7.4.

(b) A functional $ax + by$ vanishes on $(1,0)$ if and only if $a = 0$, so $U^0 = \operatorname{span}(y)$. It vanishes on $(1,1)$ if and only if $b = -a$, so $W^0 = \operatorname{span}(x - y)$. Solving $x = s\,y + t(x-y)$ gives $t = 1$ and $s - t = 0$, so $s = 1$:

$$\psi = \underbrace{y}_{\in U^0} + \underbrace{(x-y)}_{\in W^0}.$$

By (a) this decomposition is unique.

5.3 The Double Dual

In a finite-dimensional space, the dual basis gives an isomorphism $V \cong V^*$, sending v_i to φ_i. But that isomorphism depends on the basis: change the basis and the same vector goes to a different functional. Nothing stops us from taking duals twice. The dual V^* is a vector space, so it has a dual of its own, V^{**}, and two chosen bases give $V \cong V^* \cong V^{**}$. This section shows that the composite isomorphism $V \to V^{**}$ needs **no** choice at all: it is built from evaluation alone. We use it to make one meaning of the word "natural" precise, and to prove that a subspace can be recovered from its annihilator. At the end we see that in infinite dimension the construction still makes sense, but it is no longer an isomorphism.

5.3.1 Measuring the measurements

We have studied a space V and its space of measurements $V^* = \mathcal{L}(V, F)$ (Definition 5.1.2). The next object is the space of measurements **of** V^*.

Look at the expression $\varphi(v)$, with $\varphi \in V^*$ and $v \in V$. It has two inputs. If we fix φ and let

v vary, we get the functional φ itself. If we fix v and let φ vary, we get a rule that takes a functional and returns a number. For example, take $\mathbf{v} = (3, -1) \in \mathbb{R}^2$. A functional on $\mathbb{R}^2$ is $\varphi(x, y) = ax + by$ for unique a, b (Theorem 5.1.1), and the rule returns $\varphi(3, -1) = 3a - b$. So the fixed vector $(3, -1)$ measures each measurement. This rule appears every time we test a vector against functionals, so it deserves a name.

A vector measures a functional by letting the functional act on it.

Definition 5.3.1. Double Dual and Evaluation Map

Let V be a vector space over F. The **double dual** of V is $V^{**} := (V^*)^*$, the space of linear functionals on V^*.

For $v \in V$, **evaluation at** v is the function

$$\mathrm{ev}_v \colon V^* \to F, \qquad \mathrm{ev}_v(\varphi) := \varphi(v).$$

The **evaluation map** of V is

$$\mathrm{ev}_V \colon V \to V^{**}, \qquad v \mapsto \mathrm{ev}_v\,.$$

When only one space is in play we write ev for ev_V.

In words, clause by clause:

- V^{**} consists of **linear** maps $V^* \to F$. Its elements take a whole functional as input.
- ev_v is **defined on** V^*, not on V. The vector v is fixed once and for all; the variable is φ.
- The evaluation map sends each vector v to the function ev_v. A vector in the subscript, ev_v, gives a single element of V^{**}; a space in the subscript, ev_V, gives the whole map.

Well-definedness. The definition claims that $\mathrm{ev}_v \in V^{**}$, so we must check that ev_v is linear. The operations on $V^* = \mathcal{L}(V, F)$ are pointwise (Definition 2.4.1). So for $\varphi, \psi \in V^*$ and $a, b \in F$,

$$\mathrm{ev}_v(a\varphi + b\psi) = (a\varphi + b\psi)(v) = a\varphi(v) + b\psi(v) = a\,\mathrm{ev}_v(\varphi) + b\,\mathrm{ev}_v(\psi).$$

Notice what the definition did **not** use: no basis, no coordinates, no choice of any kind. Only v and the meaning of "functional" appear.

Example. Evaluation in Three Spaces

(a) Let $V = F^n$. Describe $\mathrm{ev}_{\mathbf{x}}$ for $\mathbf{x} \in F^n$ in terms of row vectors.

(b) Let $V = \mathbb{R}[x]_{\le 2}$, and let $(\varphi_0, \varphi_1, \varphi_2)$ be the dual basis of $(1, x, x^2)$, so that $\varphi_k(p)$ is the coefficient of x^k in p (Example (Dual Bases in $\mathbb{R}[x]_{\le 2}$)). Compute ev_{x^2} on $\varphi_0, \varphi_1, \varphi_2$ and on the functional $\varepsilon_3(p) = p(3)$. Which element of V^{**} is ev_{x^2}?

(c) Describe ev_0, and the evaluation map of $V = \{0\}$.

Solution.

(a) By Theorem 5.1.1, every $\varphi \in (F^n)^*$ has the form $\varphi(\mathbf{y}) = \mathbf{a}^\top \mathbf{y}$ for a unique $\mathbf{a} \in F^n$. So $\mathrm{ev}_{\mathbf{x}}(\varphi) = \mathbf{a}^\top \mathbf{x}$: the vector $\mathbf{x}$ measures the row $\mathbf{a}^\top$ by standing to its right in the product. The same product $\mathbf{a}^\top \mathbf{x}$ is read as "$\mathbf{a}^\top$ measures $\mathbf{x}$" or "$\mathbf{x}$ measures $\mathbf{a}^\top$", depending on which factor we hold fixed.

(b) $\mathrm{ev}_{x^2}(\varphi_k) = \varphi_k(x^2)$, the coefficient of x^k in x^2. So the values on $\varphi_0, \varphi_1, \varphi_2$ are $0, 0, 1$.

Also $\mathrm{ev}_{x^2}(\varepsilon_3) = \varepsilon_3(x^2) = 9$. A linear functional on V^* is determined by its values on the basis $(\varphi_0, \varphi_1, \varphi_2)$ of V^* (Theorem 2.1.5). The values $0, 0, 1$ say that ev_{x^2} is the third vector of the dual basis of $(\varphi_0, \varphi_1, \varphi_2)$ (Theorem 5.1.3). In the same way, ev_1 and ev_x are the first and second vectors of that dual basis. We will see that this is no accident.

(c) $\mathrm{ev}_0(\varphi) = \varphi(0) = 0$ for every φ, because a linear map sends 0 to 0 (Theorem 2.1.1). So ev_0 is the zero functional on V^*. If $V = \{0\}$, the only functional on V is the zero functional, so $V^* = \{0\}$ and $V^{**} = \{0\}$, and ev_V is the only map $\{0\} \to \{0\}$. This degenerate case is a first sign that ev loses nothing: the only vector it sends to zero is 0.

Here is a nonexample by minimal change. On $V = \mathbb{R}^2$, replace $\mathrm{ev}_{\mathbf{v}}$ by $q_{\mathbf{v}}(\varphi) \coloneqq \varphi(\mathbf{v})^2$. It is still a function $V^* \to \mathbb{R}$, and it is still built without any choice. But take $\mathbf{v} = (1, 0)$ and $\varphi(x, y) = x$. Then $q_{\mathbf{v}}(2\varphi) = 2^2 = 4$, while $2q_{\mathbf{v}}(\varphi) = 2$. The clause that fails is **linearity** on V^*, so $q_{\mathbf{v}} \notin V^{**}$.

Why this definition. Given only a vector v and a functional φ, the one number we can form without choosing anything is $\varphi(v)$; anything else, such as $\varphi(v)^2$, is built from it and usually breaks linearity. The pointwise operations of Definition 2.4.1 are exactly what make $\varphi \mapsto \varphi(v)$ linear. The name "double dual" is literal: it is the dual of the dual.

Warning. **Two different "evaluations".** On $V = \mathbb{R}[x]_{\leq 2}$, the functional $\varepsilon_3 \colon p \mapsto p(3)$ lies in V^*: it evaluates **polynomials at a number**. The element $\mathrm{ev}_p \colon \varphi \mapsto \varphi(p)$ lies in V^{**}: it evaluates **functionals at a polynomial**. They meet in one number, $\mathrm{ev}_p(\varepsilon_3) = p(3) = \varepsilon_3(p)$, read from opposite sides. Also, ev_v is not the vector v: it is a function on V^*.

Quick check. Let $\mathbf{v} = (1, 2) \in \mathbb{R}^2$. Compute $\mathrm{ev}_{\mathbf{v}}(\varphi)$ for $\varphi(x, y) = 3x - y$. Is there a **nonzero** $\psi \in (\mathbb{R}^2)^*$ with $\mathrm{ev}_{\mathbf{v}}(\psi) = 0$?

Solution. $\mathrm{ev}_{\mathbf{v}}(\varphi) = \varphi(1, 2) = 3 - 2 = 1$. Yes: $\psi(x, y) = 2x - y$ is nonzero and $\psi(1, 2) = 0$. A single $\mathrm{ev}_{\mathbf{v}}$ kills many functionals; what the next theorem says is that for $\mathbf{v} \neq \mathbf{0}$ it does not kill **all** of them.

5.3.2 The evaluation map never loses information

The first result says that ev is a linear map, and that it is injective in every vector space, of any dimension.

Theorem 5.3.1. The Evaluation Map Is Linear and Injective

Let V be any vector space over F. Then the evaluation map $\mathrm{ev}_V \colon V \to V^{**}$ is linear and injective.

Idea. Linearity is an identity between two elements of V^{**}, that is, between two functions on V^*; compare them at an arbitrary φ. For injectivity, use the kernel test. If $\mathrm{ev}_v = 0$, then **every** functional vanishes at v. This is the signature move of the chapter: to show a vector

is zero, test it against every functional. In finite dimension Theorem 5.1.5 does the job. In general we repeat its proof with the Basis Extension Theorem for arbitrary spaces.

Proof. *Linear.* Let $u, v \in V$ and $a, b \in F$. For every $\varphi \in V^*$, since φ is linear and the operations on V^{**} are pointwise (Definition 2.4.1),

$$\mathrm{ev}_{au+bv}(\varphi) = \varphi(au + bv) = a\varphi(u) + b\varphi(v) = (a\,\mathrm{ev}_u + b\,\mathrm{ev}_v)(\varphi).$$

Two functions on V^* that agree at every φ are equal, so $\mathrm{ev}_{au+bv} = a\,\mathrm{ev}_u + b\,\mathrm{ev}_v$.

Injective. By Theorem 2.2.3 it suffices to show that $\mathrm{ev}_v = 0$ implies $v = 0$. We prove the contrapositive. Let $v \neq 0$. If V is finite-dimensional, Theorem 5.1.5 (a) gives $\varphi \in V^*$ with $\varphi(v) = 1$. In general, $\{v\}$ is linearly independent, so by Theorem 1.8.2 there is a basis B of V with $v \in B$. By Theorem 2.1.6, which allows infinite bases, there is a linear map $\varphi\colon V \to F$ with $\varphi(v) = 1$ and $\varphi(b) = 0$ for every $b \in B \setminus \{v\}$. In either case $\varphi \in V^*$ and $\mathrm{ev}_v(\varphi) = \varphi(v) = 1 \neq 0$, so $\mathrm{ev}_v \neq 0$. This proves that ev_V is injective.

■

In infinite dimension the proof repeats the remark after Theorem 5.1.5, and Zorn's Lemma enters through Theorem 1.8.2. In finite dimension no choice principle is needed.

So every vector v is completely remembered by the list of all its measurements $\varphi(v)$. In finite dimension a count turns "injective" into "bijective".

Theorem 5.3.2. The Double Dual Isomorphism

Let V be a finite-dimensional vector space over F. Then the evaluation map $\mathrm{ev}_V\colon V \to V^{**}$ is an isomorphism. Moreover, if $\mathcal{B} = (v_1, \ldots, v_n)$ is a basis of V with dual basis $\mathcal{B}^* = (\varphi_1, \ldots, \varphi_n)$, then $(\mathrm{ev}_{v_1}, \ldots, \mathrm{ev}_{v_n})$ is the dual basis of $\mathcal{B}^*$.

Idea. Count instead of check. The dual basis has as many vectors as the basis, so $\dim V^* = \dim V$, and the same holds one level up. An injective linear map between spaces of the same finite dimension is bijective. For the second statement, "dual basis" is the condition $\theta_i(\varphi_j) = \delta_{ij}$, and for $\theta_i = \mathrm{ev}_{v_i}$ that condition is the definition of φ_j read backwards.

Proof. Let $n = \dim V$. By Corollary 5.1.4, V^* is finite-dimensional with $\dim V^* = n$, and applying it again to V^* gives $\dim V^{**} = n$. By Theorem 5.3.1, ev_V is linear and injective. Since $\dim V = \dim V^{**}$ is finite, Corollary 2.3.2 (e) shows that ev_V is bijective, hence an isomorphism (Definition 2.5.2).

For the second statement, $\mathrm{ev}_{v_i}(\varphi_j) = \varphi_j(v_i) = \delta_{ji} = \delta_{ij}$ for all i, j, by the definition of the dual basis. By the uniqueness in Theorem 5.1.3 (a), applied to the basis $\mathcal{B}^*$ of V^*, the dual basis of $\mathcal{B}^*$ (Definition 5.1.3) is the only list $(\theta_1, \ldots, \theta_n)$ in V^{**} with $\theta_i(\varphi_j) = \delta_{ij}$. Hence it is $(\mathrm{ev}_{v_1}, \ldots, \mathrm{ev}_{v_n})$, as claimed.

■

The theorem says: **every** linear functional on V^* is evaluation at exactly one vector. To find the vector, read off its coordinates from the values on a dual basis. For instance, on $V = \mathbb{R}[x]_{\leq 2}$ with the coefficient functionals $\varphi_0, \varphi_1, \varphi_2$ of Example (Evaluation in Three Spaces), let $\theta \in V^{**}$ be the functional with $\theta(\varphi_0) = 2$, $\theta(\varphi_1) = 0$, $\theta(\varphi_2) = -1$. By the second statement, $\theta = 2\,\mathrm{ev}_1 - \mathrm{ev}_{x^2} = \mathrm{ev}_{2-x^2}$, since both sides agree on the basis $(\varphi_0, \varphi_1, \varphi_2)$ and ev is linear. So, for example, $\theta(\varepsilon_2) = \varepsilon_2(2 - x^2) = -2$, where $\varepsilon_2(p) = p(2)$. Part (b) of Example (Evaluation in Three Spaces) was the case $\theta = \mathrm{ev}_{x^2}$.

5.3.3 What "natural" means

Both $V \cong V^*$ and $V \cong V^{**}$ hold in finite dimension, and both follow from counting dimensions. Yet they are very different isomorphisms. The difference is whether a choice is needed.

The isomorphism $V \to V^*$ depends on the basis. For a basis $\mathcal{B} = (v_1, \ldots, v_n)$ of V, let $\Theta_{\mathcal{B}} \colon V \to V^*$ be the linear map with $\Theta_{\mathcal{B}}(v_i) = \varphi_i$ (Theorem 2.1.5). It sends a basis to a basis, so it is an isomorphism, by the argument in the proof of Theorem 2.5.4. Now take $V = \mathbb{R}$, a space of dimension 1. With the basis (1), the dual basis is $\varphi(t) = t$, and $\Theta_{(1)}(1)$ is $t \mapsto t$. With the basis (2), the dual basis is $\varphi(t) = t/2$, since that is the functional sending 2 to 1. So $\Theta_{(2)}(2) = (t \mapsto t/2)$, and by linearity $\Theta_{(2)}(1) = \frac{1}{2}\Theta_{(2)}(2)$ is $t \mapsto t/4$. The same vector 1 goes to two different functionals. Every basis is as good as any other, so none of these isomorphisms is preferred.

The isomorphism $V \to V^{**}$ does not. Build it the same way, choosing bases twice. First $\Theta_{\mathcal{B}} \colon V \to V^*$ sends $v_i \mapsto \varphi_i$. Then, using the basis $\mathcal{B}^*$ of V^*, the map $\Theta_{\mathcal{B}^*} \colon V^* \to V^{**}$ sends each φ_i to the i-th vector of the dual basis of $\mathcal{B}^*$. By Theorem 5.3.2, that vector is ev_{v_i}. So the composite $\Theta_{\mathcal{B}^*}\Theta_{\mathcal{B}}$ sends $v_i \mapsto \mathrm{ev}_{v_i}$ for every i. The linear map ev_V does the same, so by the uniqueness part of Theorem 2.1.5,

$$\Theta_{\mathcal{B}^*}\,\Theta_{\mathcal{B}} = \mathrm{ev}_V \qquad \text{for every basis } \mathcal{B} \text{ of } V.$$

Each factor depends on $\mathcal{B}$, but the composite does not: the two choices cancel. This is the first meaning of **natural**: an isomorphism is natural if it can be defined without choosing anything, and ev_V is defined by the formula $\mathrm{ev}_v(\varphi) = \varphi(v)$ alone.

There is a second, sharper meaning. A choice-free construction should be compatible with every linear map: if $T \colon V \to W$, then first applying T and then identifying W with W^{**} should agree with first identifying V with V^{**} and then applying "the map that T induces on double duals". To state this we need to know how a linear map acts on functionals. That is the **dual map**, the subject of the next section, where we prove that ev is compatible with every linear map, and see that no family of isomorphisms $V \to V^*$ can be.

Quick check. In $V = \mathbb{R}$ with the basis $\mathcal{B} = (2)$, compute $\Theta_{\mathcal{B}^*}\Theta_{\mathcal{B}}(1)$ as an element of V^{**}, and evaluate it on $\psi(t) = 5t$. Compare with $\mathrm{ev}_1(\psi)$.

Solution. The dual basis of $\mathcal{B}$ is $\varphi_1(t) = t/2$, so $\Theta_{\mathcal{B}}(1) = \frac{1}{2}\Theta_{\mathcal{B}}(2) = \frac{1}{2}\varphi_1$. The dual basis of $\mathcal{B}^* = (\varphi_1)$ is the $\theta \in V^{**}$ with $\theta(\varphi_1) = 1$, and $\mathrm{ev}_2(\varphi_1) = \varphi_1(2) = 1$, so $\theta = \mathrm{ev}_2$ and $\Theta_{\mathcal{B}^*}(\varphi_1) = \mathrm{ev}_2$. Hence $\Theta_{\mathcal{B}^*}\Theta_{\mathcal{B}}(1) = \frac{1}{2}\,\mathrm{ev}_2 = \mathrm{ev}_1$, using linearity of ev. On ψ it gives $\mathrm{ev}_1(\psi) = \psi(1) = 5$. The factor $\frac{1}{2}$ from the first step and the factor 2 hidden in ev_2 cancel, exactly as the general computation predicts.

Remark. Because ev_V is natural, it is common to **identify** a finite-dimensional V with V^{**}, and to regard a vector v as the functional $\varphi \mapsto \varphi(v)$ on V^*. We will do this only with an explicit "under the identification $v \leftrightarrow \mathrm{ev}_v$", as in the next corollary.

5.3.4 Consequences in finite dimension

Throughout this subsection, V is finite-dimensional. The isomorphism ev_V lets us run arguments about V^* "one level up" and bring the answer back to V.

The dual basis construction goes from a basis of V to a basis of V^*. Does every basis of V^* arise this way? A list of n independent measurements should be the coordinate functions of some basis, and the double dual shows it is.

Corollary 5.3.3. Every Basis of the Dual Is a Dual Basis

Let V be a finite-dimensional vector space over F, and let $(\psi_1, \dots, \psi_n)$ be a basis of V^*. Then there is exactly one basis $(v_1, \dots, v_n)$ of V whose dual basis is $(\psi_1, \dots, \psi_n)$, that is, with

$$\psi_i(v_j) = \delta_{ij} \qquad \text{for all } i, j.$$

Idea. Go up one level, where the problem is already solved. The basis $(\psi_1, \dots, \psi_n)$ of V^* has a dual basis $(\theta_1, \dots, \theta_n)$ in V^{**}. Every θ_j is evaluation at some vector, because ev is surjective. Those vectors are the answer, since $\theta_j(\psi_i) = \delta_{ij}$ says $\psi_i(v_j) = \delta_{ij}$. Uniqueness is a "test against functionals" argument.

Proof. Existence. Here V^* is finite-dimensional (Corollary 5.1.4). By Theorem 5.1.3 applied to V^*, there is a basis $(\theta_1, \dots, \theta_n)$ of V^{**} with $\theta_j(\psi_i) = \delta_{ij}$. By Theorem 5.3.2, ev_V is an isomorphism, so there are unique $v_j \in V$ with $\mathrm{ev}_{v_j} = \theta_j$. Its inverse is an isomorphism (Theorem 2.5.3), so $(v_1, \dots, v_n)$ is a basis of V by Theorem 2.5.2. For all i, j,

$$\psi_i(v_j) = \mathrm{ev}_{v_j}(\psi_i) = \theta_j(\psi_i) = \delta_{ij}.$$

By the uniqueness in Theorem 5.1.3 (a), the dual basis of $(v_1, \dots, v_n)$ is $(\psi_1, \dots, \psi_n)$.

Uniqueness. Suppose $(u_1, \dots, u_n)$ is also a basis with $\psi_i(u_j) = \delta_{ij}$. Fix j. Then $\psi_i(v_j - u_j) = \delta_{ij} - \delta_{ij} = 0$ for every i. Every $\varphi \in V^*$ is a combination of the ψ_i, so $\varphi(v_j - u_j) = 0$ for every $\varphi \in V^*$. By Theorem 5.1.5 (b), $v_j - u_j = 0$. Hence $u_j = v_j$ for every j.

■

In coordinates this is a matrix inversion, the mirror image of Example (The Dual Basis Is Read Off the Inverse Matrix). There, a basis was given and its dual basis appeared as the rows of an inverse matrix. Here the functionals are given, and the basis appears as the columns of an inverse matrix.

Example. A Basis With Prescribed Dual Basis

On $V = \mathbb{R}[x]_{\le 2}$, consider the functionals

$$\psi_1(p) = p(0), \qquad \psi_2(p) = p(1), \qquad \psi_3(p) = p'(1).$$

Show that (ψ_1, ψ_2, ψ_3) is a basis of V^*, and find the basis (p_1, p_2, p_3) of V whose dual basis it is.

Solution. Write down a matrix. For $p = a + bx + cx^2$, with $[p]_{\mathcal{E}} = (a, b, c)$ in $\mathcal{E} = (1, x, x^2)$,

$$\psi_1(p) = a, \qquad \psi_2(p) = a + b + c, \qquad \psi_3(p) = b + 2c.$$

So $\psi_i(p) = \mathbf{r}_i [p]_{\mathcal{E}}$, where $\mathbf{r}_i$ is row i of

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 2 \end{pmatrix}.$$

Translate the condition. Let $\mathbf{P}$ be the 3×3 matrix whose j-th column is $[p_j]_{\mathcal{E}}$. By Theorem 0.9.2, the (i, j) entry of $\mathbf{AP}$ is $\mathbf{r}_i [p_j]_{\mathcal{E}} = \psi_i(p_j)$. So the condition $\psi_i(p_j) = \delta_{ij}$ says exactly $\mathbf{AP} = \mathbf{I}_3$.

Solve. Row reduction (Theorem 3.4.6) gives

$$\mathbf{A}^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ -2 & 2 & -1 \\ 1 & -1 & 1 \end{pmatrix},$$

and one checks $\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}_3$, so $\mathbf{A}$ is invertible by Theorem 3.4.5. Multiplying $\mathbf{AP} = \mathbf{I}_3$ on the left by $\mathbf{A}^{-1}$ forces $\mathbf{P} = \mathbf{A}^{-1}$. Since the columns of the invertible matrix $\mathbf{P}$ are independent, (p_1, p_2, p_3) is a basis of V (via the coordinate isomorphism, Corollary 2.5.5). By the uniqueness in Theorem 5.1.3 (a), its dual basis is (ψ_1, ψ_2, ψ_3), which in particular is a basis of V^*. Reading the columns,

$$p_1 = 1 - 2x + x^2 = (1 - x)^2, \qquad p_2 = 2x - x^2, \qquad p_3 = x^2 - x.$$

Check. $p_1(0) = 1$, $p_1(1) = 0$, $p_1'(1) = 0$; $p_2(0) = 0$, $p_2(1) = 1$, $p_2'(1) = 2 - 2 = 0$; $p_3(0) = 0$, $p_3(1) = 0$, $p_3'(1) = 1$.

So every $p \in \mathbb{R}[x]_{\le 2}$ is $p = p(0)\, p_1 + p(1)\, p_2 + p'(1)\, p_3$: the coordinates of p in this basis are the three measurements.

The next consequence is the one we will use most. A subspace U determines the set U^0 of functionals vanishing on it. Going back, the functionals in U^0 vanish together on some set of vectors, which contains U. Is it exactly U?

Corollary 5.3.4. The Annihilator of the Annihilator

Let V be a finite-dimensional vector space over F, and let U be a subspace of V. Write $U^{00} := (U^0)^0 \subseteq V^{**}$. Then

$$U^{00} = \mathrm{ev}_V(U) = \{\mathrm{ev}_u : u \in U\}.$$

So, under the identification $v \leftrightarrow \mathrm{ev}_v$, we have $U^{00} = U$. Equivalently,

$$U = \{v \in V : \varphi(v) = 0 \text{ for every } \varphi \in U^0\}.$$

Idea. One inclusion is a tautology: if $u \in U$ and φ kills U, then $\mathrm{ev}_u(\varphi) = \varphi(u) = 0$. For the other, count. The dimension formula applied twice, once in V and once in V^*, gives $\dim U^{00} = \dim U$, and ev is injective, so $\dim \mathrm{ev}_V(U) = \dim U$ too. One inclusion plus equal dimension gives equality.

Proof. Let $n = \dim V$, so $\dim V^* = n$ by Corollary 5.1.4.

Inclusion. Let $u \in U$. For every $\varphi \in U^0$, $\mathrm{ev}_u(\varphi) = \varphi(u) = 0$ by Definition 5.2.1. Hence $\mathrm{ev}_u \in (U^0)^0$, and $\mathrm{ev}_V(U) \subseteq U^{00}$.

Dimensions. By Theorem 5.2.2 in V, $\dim U^0 = n - \dim U$. By Theorem 5.2.2 in the n-dimensional space V^*, applied to its subspace U^0,

$$\dim U^{00} = n - \dim U^0 = n - (n - \dim U) = \dim U.$$

The restriction of ev_V to U is linear and injective (Theorem 5.3.1), so it is an isomorphism onto its image $\mathrm{ev}_V(U)$, and $\dim \mathrm{ev}_V(U) = \dim U$ (Theorem 2.5.4). So $\mathrm{ev}_V(U)$ is a subspace of U^{00} of the same finite dimension, and Theorem 1.6.8 gives $\mathrm{ev}_V(U) = U^{00}$.

The equivalent form. Let $v \in V$. Then $\varphi(v) = 0$ for every $\varphi \in U^0$ if and only if $\mathrm{ev}_v \in U^{00} = \mathrm{ev}_V(U)$, if and only if $\mathrm{ev}_v = \mathrm{ev}_u$ for some $u \in U$. Since ev_V is injective, this happens if and only if $v \in U$.

■

The last form says: **a subspace is exactly the common zero set of the functionals that vanish on it.** No vector outside U sneaks through all the equations. In section 6 of this chapter this becomes the statement that every subspace of F^n is the solution set of a homogeneous linear system. An immediate consequence is that the annihilator loses no information.

Corollary 5.3.5. Subspaces Are Determined by Their Annihilators

Let V be a finite-dimensional vector space, and let U and W be subspaces of V. Then $U = W$ if and only if $U^0 = W^0$.

Proof. ($\Rightarrow$) is immediate. ($\Leftarrow$) Suppose $U^0 = W^0$. By Corollary 5.3.4, $U = \{v : \varphi(v) = 0 \text{ for all } \varphi \in U^0\} = \{v : \varphi(v) = 0 \text{ for all } \varphi \in W^0\} = W$.

■

Quick check. Is Corollary 5.3.5 still true if U and W are allowed to be arbitrary **subsets** of V?

Solution. No. In $\mathbb{R}^2$, take $U = \{(1,0)\}$ and $W = \{(2,0)\}$. A functional $\varphi(x,y) = ax + by$ vanishes at $(1,0)$ exactly when $a = 0$, and at $(2,0)$ exactly when $2a = 0$, that is, again when $a = 0$. So $U^0 = W^0$ but $U \neq W$. An annihilator only sees the span, $U^0 = (\operatorname{span} U)^0$ (Theorem 5.2.3 (b)).

5.3.5 Infinite dimension: the evaluation map is not surjective

Nothing in Definition 5.3.1 or Theorem 5.3.1 needed finite dimension. The count in Theorem 5.3.2 did.

Warning. $V \cong V^{**}$ **through** ev **is a finite-dimensional fact.** For $V = F[x]$, the evaluation map is injective but **not** surjective: there is a linear functional on $F[x]^*$ that is not evaluation at any polynomial. The next example constructs one.

Example. A Functional on $F[x]^*$ That Is Not an Evaluation

Let $\varphi_k \in F[x]^*$ $(k \in \mathbb{N})$ be the coordinate functionals, $\varphi_k(p) =$ the coefficient of x^k in p, and let $\varepsilon_1(p) = p(1)$, as in Proposition 5.1.6. Show that there is $\theta \in F[x]^{**}$ with $\theta \neq \mathrm{ev}_p$ for every $p \in F[x]$.

Solution. Step 1: $S = \{\varphi_k : k \in \mathbb{N}\} \cup \{\varepsilon_1\}$ is linearly independent. By Proposition 5.1.6, the φ_k form an independent set and $\varepsilon_1 \notin \operatorname{span}\{\varphi_k : k \in \mathbb{N}\}$. Let L be a finite list of distinct elements of S. If ε_1 is not in L, then L is independent. Otherwise, list ε_1 last; the others are independent and ε_1 is not in their span, so L is independent by Lemma 1.5.3.

Step 2: build θ. By Theorem 1.8.2 there is a basis B of $F[x]^*$ containing S. By Theorem 2.1.6 for the (infinite) basis B, there is a linear $\theta\colon F[x]^* \to F$ with $\theta(\varepsilon_1) = 1$ and $\theta(\beta) = 0$ for every $\beta \in B \setminus \{\varepsilon_1\}$. In particular $\theta(\varphi_k) = 0$ for all k.

Step 3: θ is not an evaluation. Suppose $\theta = \mathrm{ev}_p$ for some $p \in F[x]$. Then for every k, the coefficient of x^k in p is $\varphi_k(p) = \mathrm{ev}_p(\varphi_k) = \theta(\varphi_k) = 0$. So $p = 0$, and $\theta = \mathrm{ev}_0 = 0$. This contradicts $\theta(\varepsilon_1) = 1$. Hence $\theta \notin \mathrm{ev}(F[x])$, and ev is not surjective.

The functional θ exists, but nobody can write it down: it comes from Zorn's Lemma through Theorem 1.8.2. The deeper reason for the failure is size. $F[x]^*$ is so much bigger than $F[x]$ (Exercise (C4: How large is $F[x]^*$?) identifies it with the whole sequence space $F^{\mathbb{N}}$) that $F[x]^{**}$ is bigger still, and in fact no isomorphism $F[x] \cong F[x]^{**}$ exists at all, a statement about the sizes of bases that we do not prove. In finite dimension, sizes cannot grow, and the evaluation map is the perfect dictionary of Theorem 5.3.2.

5.3.6 Exercises

5.3.6.1 A. Check your understanding

Exercise. A1

(a) Define ev_v for $v \in V$, and the evaluation map ev_V. In which space does ev_v live?

(b) State a hypothesis on V under which ev_V is an isomorphism, and a space for which it is not.

(c) True or false: if V is finite-dimensional and $v \neq 0$, then $\mathrm{ev}_v(\varphi) \neq 0$ for **every** nonzero $\varphi \in V^*$. Justify your answer.

(d) True or false: if U and W are subspaces of $\mathbb{R}^5$ with $U^0 = W^0$, then $U = W$. Justify your answer.

(e) Name the method used to prove $U^{00} = \mathrm{ev}_V(U)$ in Corollary 5.3.4.

Solution.

(a) $\mathrm{ev}_v \colon V^* \to F$, $\mathrm{ev}_v(\varphi) = \varphi(v)$, and $\mathrm{ev}_V \colon V \to V^{**}$, $v \mapsto \mathrm{ev}_v$ (Definition 5.3.1). ev_v lives in V^{**}.

(b) If V is finite-dimensional, ev_V is an isomorphism (Theorem 5.3.2). For $V = F[x]$ it is not surjective (Example (A Functional on $F[x]^*$ That Is Not an Evaluation)).

(c) False. It says only that **some** φ has $\varphi(\mathbf{v}) \neq 0$ (Theorem 5.3.1). In $\mathbb{R}^2$, $\mathbf{v} = (1, 2)$ and $\varphi(x, y) = 2x - y \neq 0$ give $\mathrm{ev}_{\mathbf{v}}(\varphi) = 0$.

(d) True, by Corollary 5.3.5, since $\mathbb{R}^5$ is finite-dimensional.

(e) One inclusion plus equal dimension: $\mathrm{ev}_V(U) \subseteq U^{00}$ directly, then $\dim U^{00} = \dim U = \dim \mathrm{ev}_V(U)$ from the dimension formula for annihilators (twice) and injectivity of ev, and Theorem 1.6.8.

5.3.6.2 B. Practice

Exercise. B1: A basis with a given dual basis

(a) Let $\psi_1(x, y) = 2x + y$ and $\psi_2(x, y) = x + y$ on $\mathbb{R}^2$. Find the basis $(\mathbf{v}_1, \mathbf{v}_2)$ of $\mathbb{R}^2$ whose dual basis is (ψ_1, ψ_2).

(b) Let $\psi_1(p) = p(0)$ and $\psi_2(p) = \int_0^1 p(x)\,dx$ on $\mathbb{R}[x]_{\leq 1}$. Find the basis (p_1, p_2) of $\mathbb{R}[x]_{\leq 1}$ whose dual basis is (ψ_1, ψ_2).

Solution.

(a) As in Example (A Basis With Prescribed Dual Basis), $\psi_i(\mathbf{v}) = \mathbf{r}_i \mathbf{v}$ with $\mathbf{r}_i$ the rows of $\mathbf{A} = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$, and the condition $\psi_i(\mathbf{v}_j) = \delta_{ij}$ says $\mathbf{AP} = \mathbf{I}_2$ for $\mathbf{P} = \begin{pmatrix} \mathbf{v}_1 & \mathbf{v}_2 \end{pmatrix}$. Here $2 \cdot 1 - 1 \cdot 1 = 1 \neq 0$, so by Theorem 0.9.8

$$\mathbf{A}^{-1} = \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix},$$

$\mathbf{P} = \mathbf{A}^{-1}$, so $\mathbf{v}_1 = (1, -1)$ and $\mathbf{v}_2 = (-1, 2)$. These are independent (the columns of an invertible matrix), hence a basis. Check: $\psi_1(\mathbf{v}_1) = 1$, $\psi_2(\mathbf{v}_1) = 0$, $\psi_1(\mathbf{v}_2) = 0$, $\psi_2(\mathbf{v}_2) = 1$. By the uniqueness part of Theorem 5.1.3, the dual basis of $(\mathbf{v}_1, \mathbf{v}_2)$ is (ψ_1, ψ_2).

(b) For $p = a + bx$, $\psi_1(p) = a$ and $\psi_2(p) = a + \frac{1}{2}b$. So $\mathbf{A} = \begin{pmatrix} 1 & 0 \\ 1 & 1/2 \end{pmatrix}$ and

$$\mathbf{P} = \mathbf{A}^{-1} = \begin{pmatrix} 1 & 0 \\ -2 & 2 \end{pmatrix},$$

whose columns are the coordinates of $p_1 = 1 - 2x$ and $p_2 = 2x$ in $(1, x)$. Check: $p_1(0) = 1$, $\int_0^1 (1 - 2x)\,dx = 1 - 1 = 0$, $p_2(0) = 0$, $\int_0^1 2x\,dx = 1$. As in (a), (p_1, p_2) is a basis with dual basis (ψ_1, ψ_2).

Exercise. B2: $U^{00} = U$ for a line

Let $U = \operatorname{span}((1, 2, -1)) \subseteq \mathbb{R}^3$.

(a) Find a basis of U^0.

(b) Find all $\mathbf{v} \in \mathbb{R}^3$ with $\varphi(\mathbf{v}) = 0$ for every $\varphi \in U^0$, and check that the answer is U.

(c) Hence verify that $U^{00} = \operatorname{ev}(U)$, and confirm the dimension count.

Solution.

(a) A functional $\varphi(x, y, z) = ax + by + cz$ (Theorem 5.1.1) lies in U^0 if and only if it vanishes at $(1, 2, -1)$ (Theorem 5.2.3 (b)), that is, $a + 2b - c = 0$. The solutions are $(a, b, c) = s(-2, 1, 0) + t(1, 0, 1)$, so $\varphi_1(x, y, z) = 2x - y$ and $\varphi_2(x, y, z) = x + z$ span U^0 (we replaced $(-2, 1, 0)$ by its negative). They are independent, since neither is a multiple of the other, so (φ_1, φ_2) is a basis of U^0. This matches $\dim U^0 = 3 - 1 = 2$ (Theorem 5.2.2).

(b) Every $\varphi \in U^0$ is a combination of φ_1, φ_2, so $\varphi(\mathbf{v}) = 0$ for all $\varphi \in U^0$ if and only if $\varphi_1(\mathbf{v}) = \varphi_2(\mathbf{v}) = 0$, that is, $2x - y = 0$ and $x + z = 0$. With $x = t$: $y = 2t$, $z = -t$, so $\mathbf{v} = t(1, 2, -1)$. The set is exactly U.

(c) Let $\theta \in U^{00} \subseteq (\mathbb{R}^3)^{**}$. By Theorem 5.3.2, $\theta = \operatorname{ev}_{\mathbf{v}}$ for some $\mathbf{v}$, and $\theta \in (U^0)^0$ means $\varphi(\mathbf{v}) = 0$ for all $\varphi \in U^0$. By (b), $\mathbf{v} \in U$, so $U^{00} \subseteq \operatorname{ev}(U)$; the reverse inclusion is immediate from Definition 5.2.1. Dimensions: $\dim U^{00} = 3 - \dim U^0 = 1 = \dim U$.

Exercise. B3: Recognizing an evaluation

Let $V = \mathbb{R}[x]_{\le 2}$, with the coefficient functionals $\varphi_0, \varphi_1, \varphi_2$ of Example (Evaluation in Three Spaces). Let $\theta \in V^{**}$ be the linear functional with $\theta(\varphi_0) = 1$, $\theta(\varphi_1) = -3$, $\theta(\varphi_2) = 2$. Find $p \in V$ with $\theta = \operatorname{ev}_p$. Hence compute $\theta(\varepsilon_1)$ and $\theta(\psi)$, where $\varepsilon_1(q) = q(1)$ and $\psi(q) = q'(0)$.

Solution. By Theorem 5.3.2, $(\operatorname{ev}_1, \operatorname{ev}_x, \operatorname{ev}_{x^2})$ is the dual basis of $(\varphi_0, \varphi_1, \varphi_2)$. Hence $\operatorname{ev}_1 - 3\operatorname{ev}_x + 2\operatorname{ev}_{x^2}$ takes the values $1, -3, 2$ on $\varphi_0, \varphi_1, \varphi_2$. So does θ, and a linear functional is determined by its values on a basis (Theorem 2.1.5). Since ev is linear (Theorem 5.3.1),

$$\theta = \operatorname{ev}_1 - 3\operatorname{ev}_x + 2\operatorname{ev}_{x^2} = \operatorname{ev}_p, \qquad p = 1 - 3x + 2x^2.$$

Hence $\theta(\varepsilon_1) = \varepsilon_1(p) = 1 - 3 + 2 = 0$ and $\theta(\psi) = \psi(p) = p'(0) = -3$.

5.3.6.3 C. Going deeper

Exercise. C1: Subspaces of V and of V^*

Let V be a finite-dimensional vector space over F. Prove that $U \mapsto U^0$ is a bijection from the set of subspaces of V to the set of subspaces of V^*, and that it reverses inclusions in both directions: $U \subseteq W$ if and only if $W^0 \subseteq U^0$.

Hint: for surjectivity, given a subspace $X \subseteq V^*$, consider $U = \{v \in V : \varphi(v) = 0$ for all $\varphi \in X\}$.

Solution. Let $n = \dim V = \dim V^*$ (Corollary 5.1.4). By Proposition 5.2.1, U^0 is a subspace of V^*, so the map is well defined.

Injective. This is Corollary 5.3.5.

Inclusions. If $U \subseteq W$, then $W^0 \subseteq U^0$ by Theorem 5.2.3 (a). Conversely, suppose $W^0 \subseteq U^0$, and let $u \in U$. Every $\varphi \in W^0$ lies in U^0, so $\varphi(u) = 0$. By Corollary 5.3.4 (the equivalent form, for W), $u \in W$. Hence $U \subseteq W$.

Surjective. Let X be a subspace of V^*, and let $U = \{v \in V : \varphi(v) = 0$ for all $\varphi \in X\}$. For $v \in V$, $v \in U$ if and only if $\mathrm{ev}_v(\varphi) = 0$ for all $\varphi \in X$ if and only if $\mathrm{ev}_v \in X^0$, where $X^0 \subseteq V^{**}$. So $U = \mathrm{ev}_V^{-1}(X^0)$, and since ev_V is an isomorphism (Theorem 5.3.2), U is a subspace and ev_V restricts to an isomorphism $U \to X^0$. By Theorem 5.2.2 in V^*, $\dim U = \dim X^0 = n - \dim X$. By definition of U, every $\varphi \in X$ vanishes on U, so $X \subseteq U^0$. By Theorem 5.2.2 in V, $\dim U^0 = n - \dim U = \dim X$. By Theorem 1.6.8, $X = U^0$. This proves the claim.

Exercise. C2: Spans cut out by equations

Let V be finite-dimensional and $S \subseteq V$ any subset.

(a) Prove that $\{v \in V : \varphi(v) = 0$ for every $\varphi \in S^0\} = \mathrm{span}(S)$.

(b) Deduce: a vector v lies in $\mathrm{span}(S)$ if and only if every functional that vanishes on S vanishes at v.

(c) Show by an example that (a) can fail if "$\mathrm{span}(S)$" is replaced by "S".

Solution.

(a) $\mathrm{span}(S)$ is a subspace (Theorem 1.3.1), and $S^0 = (\mathrm{span}\, S)^0$ by Theorem 5.2.3 (b). Applying Corollary 5.3.4 to $U = \mathrm{span}(S)$,

$$\{v : \varphi(v) = 0 \ \forall \varphi \in S^0\} = \{v : \varphi(v) = 0 \ \forall \varphi \in (\mathrm{span}\, S)^0\} = \mathrm{span}(S).$$

(b) This is (a) read for a single vector v.

(c) Take $S = \{(1, 0)\} \subseteq \mathbb{R}^2$. Then S^0 consists of the functionals $(x, y) \mapsto by$, and their common zero set is the whole x-axis, which contains $(2, 0) \notin S$.

Exercise. C3: Where finite dimension is needed

Let $V = F[x]$, let φ_k and ε_1 be as in Example (A Functional on $F[x]^*$ That Is Not an Evaluation), and let $X = \operatorname{span}\{\varphi_k : k \in \mathbb{N}\} \subseteq V^*$.

(a) Show that $\{p \in F[x] : \varphi(p) = 0 \text{ for all } \varphi \in X\} = \{0\}$.

(b) Deduce that $X \neq U^0$ for **every** subspace U of $F[x]$. So the map $U \mapsto U^0$ of Exercise (C1: Subspaces of V and of V^*) is not surjective for $F[x]$.

(c) Which step of the solution of Exercise (C1: Subspaces of V and of V^*) fails here?

Solution.

(a) If $\varphi(p) = 0$ for all $\varphi \in X$, then in particular $\varphi_k(p) = 0$ for every k, so every coefficient of p is 0 and $p = 0$. Conversely $\varphi(0) = 0$ for every φ.

(b) Suppose $X = U^0$. Every $\varphi \in X = U^0$ vanishes on U, so $U \subseteq \{p : \varphi(p) = 0 \ \forall \varphi \in X\} = \{0\}$ by (a). Hence $U = \{0\}$ and $U^0 = V^*$ (every functional vanishes at 0). But $\varepsilon_1 \in V^*$ and $\varepsilon_1 \notin X$, by Proposition 5.1.6. So $X \neq V^* = U^0$, a contradiction.

(c) The dimension count. In Exercise (C1: Subspaces of V and of V^*) we used that ev_V is surjective and that $\dim U + \dim U^0 = \dim V$ to force $X = U^0$ from $X \subseteq U^0$. Here $U = \{0\}$ and $X \subsetneq U^0 = V^*$: the inclusion holds, but there is no finite dimension to compare, and a proper subspace of an infinite-dimensional space can be "as large as" the whole space.

5.4 The Dual Map and the Transpose

So far each space has had its own dual, but the duals of different spaces have not talked to each other. Linear algebra is about maps, so the question is: what does a linear map $T\colon V \to W$ do to functionals? This section shows that T moves vectors forward but moves measurements **backward**, giving the dual map $T'\colon W^* \to V^*$. In dual bases its matrix is the transpose of the matrix of T, which finally explains what the transpose of Chapter 0 means. The kernel and image of T' are annihilators of the image and kernel of T, and counting dimensions gives a third proof that row rank equals column rank. At the end we return to the double dual and prove that the evaluation map is compatible with every linear map.

5.4.1 Pulling functionals back

After spaces come the maps between them. We have a linear map $T\colon V \to W$ and the two spaces of measurements V^* and W^*. We want to know how T relates them.

Start with the standard example, a matrix map $T_{\mathbf{A}}\colon F^n \to F^m$, $\mathbf{x} \mapsto \mathbf{A}\mathbf{x}$, with $\mathbf{A} \in M_{m\times n}(F)$. A functional on F^m is $\psi(\mathbf{y}) = \mathbf{b}^\top\mathbf{y}$ for a unique $\mathbf{b} \in F^m$ (Theorem 5.1.1). Given ψ, there is one obvious way to measure a vector $\mathbf{x} \in F^n$: move it to F^m with $T_{\mathbf{A}}$, then measure. The result is

$$\psi(T_{\mathbf{A}}\mathbf{x}) = \mathbf{b}^\top\mathbf{A}\mathbf{x} = (\mathbf{A}^\top\mathbf{b})^\top\mathbf{x},$$

using $(\mathbf{A}^\top\mathbf{b})^\top = \mathbf{b}^\top\mathbf{A}$ (Theorem 0.9.5). So "move, then measure" is again a functional, now on F^n, and it is given by the vector $\mathbf{A}^\top\mathbf{b}$. The transpose appears by itself. The recipe "compose with T" makes sense for any linear map, and it deserves a name.

The dual map pulls a measurement of W back to a measurement of V: first move the vector with T, then measure it in W.

Definition 5.4.1. Dual Map

Let V and W be vector spaces over F, and let $T \in \mathcal{L}(V, W)$. The **dual map** of T is the map

$$T' \colon W^* \to V^*, \qquad T'(\psi) \coloneqq \psi T = \psi \circ T.$$

That is, $(T'\psi)(v) = \psi(Tv)$ for every $\psi \in W^*$ and every $v \in V$.

In words, clause by clause:

- The input of T' is a functional on the **codomain** W, and the output is a functional on the **domain** V. The arrow is reversed.
- The formula $(T'\psi)(v) = \psi(Tv)$ has two steps: to measure v with $T'\psi$, send v to W with T, then apply ψ there.
- No basis and no finite dimension are involved.

Well-definedness. Two things must be checked. First, $T'\psi$ must lie in V^*: it is the composite of the linear maps $T \colon V \to W$ and $\psi \colon W \to F$, so it is linear by Theorem 2.4.2 (a). Second, T' should be linear, so that $T' \in \mathcal{L}(W^*, V^*)$. By the distributive laws in Theorem 2.4.2 (d), for $\psi_1, \psi_2 \in W^*$ and $c \in F$,

$$\begin{aligned} T'(\psi_1 + \psi_2) &= (\psi_1 + \psi_2)T = \psi_1 T + \psi_2 T = T'\psi_1 + T'\psi_2, \\ T'(c\psi) &= (c\psi)T = c(\psi T) = cT'\psi. \end{aligned}$$

Example. Dual Maps in Four Settings

(a) For $\mathbf{A} \in M_{m\times n}(F)$, describe $T'_{\mathbf{A}}$ in terms of the vectors $\mathbf{b}$ and $\mathbf{A}^\top \mathbf{b}$ that represent functionals.

(b) Let $D \colon \mathbb{R}[x]_{\le 3} \to \mathbb{R}[x]_{\le 2}$, $D(p) = p'$, and let $\varepsilon_2 \in (\mathbb{R}[x]_{\le 2})^*$ be $\varepsilon_2(q) = q(2)$. Compute $D'(\varepsilon_2)$.

(c) Let $S \colon F[x] \to F[x]$, $S(p) = xp$, and let $\varphi_k \in F[x]^*$ send a polynomial to its coefficient of x^k. Compute $S'(\varphi_k)$ for every $k \in \mathbb{N}$.

(d) Find $(\mathrm{id}_V)'$, the dual map of the zero map $V \to W$, and the dual map of the inclusion $\iota \colon U \to V$, $\iota(u) = u$, of a subspace $U \subseteq V$.

Solution.

(a) If $\psi(\mathbf{y}) = \mathbf{b}^\top \mathbf{y}$, the computation before the definition shows $(T'_{\mathbf{A}}\psi)(\mathbf{x}) = (\mathbf{A}^\top \mathbf{b})^\top \mathbf{x}$. So, identifying a functional with its vector, $T'_{\mathbf{A}}$ is $\mathbf{b} \mapsto \mathbf{A}^\top \mathbf{b}$. In row language: $T'_{\mathbf{A}}$ sends the row $\mathbf{b}^\top$ to the row $\mathbf{b}^\top \mathbf{A}$.

(b) For $p \in \mathbb{R}[x]_{\le 3}$, $(D'\varepsilon_2)(p) = \varepsilon_2(Dp) = p'(2)$. So D' turns "the value at 2" into "the slope at 2".

(c) $(S'\varphi_k)(p) = \varphi_k(xp)$. Multiplying by x shifts every coefficient up one place, so the coefficient of x^k in xp is the coefficient of x^{k-1} in p when $k \ge 1$, and 0 when $k = 0$. Hence $S'\varphi_k = \varphi_{k-1}$ for $k \ge 1$, and $S'\varphi_0 = 0$. Here $F[x]$ is infinite-dimensional, and the definition works all the same.

(d) $(\mathrm{id}_V)'\varphi = \varphi \circ \mathrm{id}_V = \varphi$, so $(\mathrm{id}_V)' = \mathrm{id}_{V^*}$. For the zero map $0\colon V \to W$, $(0'\psi)(v) = \psi(0) = 0$ (Theorem 2.1.1), so $0'$ is the zero map $W^* \to V^*$. For the inclusion, $(\iota'\varphi)(u) = \varphi(u)$ for $u \in U$, so $\iota'\varphi = \varphi|_U$ is the **restriction** of φ to U. The restriction map $V^* \to U^*$ is thus a dual map; we study it further in the next section.

Here is a nonexample by minimal change. Try to push functionals **forward** instead: for $T\colon V \to W$ and $\varphi \in V^*$, look for a $\psi \in W^*$ with $\psi T = \varphi$, and call it "$T_*\varphi$". Take $T\colon \mathbb{R} \to \mathbb{R}^2$, $T(t) = (t, 0)$, and $\varphi(t) = t$. Then $\psi(x, y) = x + cy$ satisfies $\psi(T(t)) = t$ for **every** $c \in \mathbb{R}$, so there is no single answer. And for the zero map $\mathbb{R} \to \mathbb{R}^2$, no ψ at all satisfies $\psi \circ 0 = \varphi$. What fails is well-definedness: a function must assign **exactly one** value. Composition with T goes only one way, on the right, and that forces the direction $W^* \to V^*$.

Why this definition. From $T\colon V \to W$ and $\psi\colon W \to F$, the only composite we can form is ψT; the nonexample shows that the other direction is not a function. Linearity of T' comes for free from the distributive laws. The name "dual map" says that T' is the map T induces between the duals. It is also called the **transpose** of T, for a reason that the matrix theorem below makes exact. We write T' rather than T^*, because T^* is reserved for the adjoint of Chapter 11, which is related to T' but needs an inner product.

Warning. **The dual map goes backwards.** For $T\colon V \to W$, the dual map is $T'\colon W^* \to V^*$, **not** $V^* \to W^*$. As a result, composition reverses order: if $T\colon U \to V$ and $S\colon V \to W$, then $(ST)' = T'S'$, just as $(\mathbf{AB})^\top = \mathbf{B}^\top\mathbf{A}^\top$. The other order $S'T'$ does not even make sense: T' lands in U^*, where S' is not defined.

Quick check. Let $T\colon \mathbb{R}^3 \to \mathbb{R}^2$, $T(x, y, z) = (x - z, 2y)$, and $\psi(u, v) = u + v$. Compute $T'\psi$. If T' is written as a matrix in some bases, what is its size?

Solution. $(T'\psi)(x, y, z) = \psi(x - z, 2y) = x + 2y - z$. The map T' goes from $(\mathbb{R}^2)^*$, of dimension 2, to $(\mathbb{R}^3)^*$, of dimension 3 (Corollary 5.1.4), so its matrix is 3×2, the size of the transpose of the 2×3 matrix of T.

The first payoff is that duals behave well under the operations on maps, with one reversal.

Theorem 5.4.1. Properties of the Dual Map

Let U, V, W be vector spaces over F.

(a) For $S, T \in \mathcal{L}(V, W)$ and $a, b \in F$, $(aS + bT)' = aS' + bT'$.
(b) For $T \in \mathcal{L}(U, V)$ and $S \in \mathcal{L}(V, W)$, $(ST)' = T'S'$.
(c) $(\mathrm{id}_V)' = \mathrm{id}_{V^*}$.
(d) If $T \in \mathcal{L}(V, W)$ is invertible, then T' is invertible and $(T')^{-1} = (T^{-1})'$.

Idea. Each identity is an equality of maps into a dual space, so we evaluate both sides at a functional, and then evaluate the resulting functionals at a vector. After that, each line is the

definition of T' read twice. Part (d) needs no computation: apply (b) and (c) to $T^{-1}T = \mathrm{id}_V$ and $TT^{-1} = \mathrm{id}_W$.

Proof.

(a) Let $\psi \in W^*$ and $v \in V$. Using the linearity of ψ and the pointwise operations (Definition 2.4.1),

$$((aS + bT)'\psi)(v) = \psi(aSv + bTv) = a\psi(Sv) + b\psi(Tv) = (aS'\psi + bT'\psi)(v).$$

Hence $(aS + bT)'\psi = (aS' + bT')\psi$ for every ψ, and $(aS + bT)' = aS' + bT'$.

(b) Let $\psi \in W^*$. By Definition 5.4.1 and associativity of composition (Theorem 2.4.2 (b)),

$$(ST)'\psi = \psi(ST) = (\psi S)T = T'(\psi S) = T'(S'\psi) = (T'S')\psi.$$

Hence $(ST)' = T'S'$.

(c) This is Example (Dual Maps in Four Settings) (d).

(d) Suppose T is invertible with inverse $T^{-1} \in \mathcal{L}(W, V)$. By (b) and (c),

$$T'(T^{-1})' = (T^{-1}T)' = (\mathrm{id}_V)' = \mathrm{id}_{V^*}, \qquad (T^{-1})'T' = (TT^{-1})' = (\mathrm{id}_W)' = \mathrm{id}_{W^*}.$$

The map $(T^{-1})'$ is linear, so by Definition 2.5.1 T' is invertible with inverse $(T^{-1})'$. This proves the theorem.

■

Part (a) says that $T \mapsto T'$ is itself a linear map $\mathcal{L}(V, W) \to \mathcal{L}(W^*, V^*)$. Part (b) is the reversal of the warning. Parts (c) and (d) say that dualizing turns isomorphisms into isomorphisms, so isomorphic spaces have isomorphic duals.

5.4.2 The matrix of the dual map

Now we fix bases and compute. Let $\mathcal{B} = (v_1, \ldots, v_n)$ be a basis of V with dual basis $\mathcal{B}^* = (\varphi_1, \ldots, \varphi_n)$ (Definition 5.1.3), and $\mathcal{C} = (w_1, \ldots, w_m)$ a basis of W with dual basis $\mathcal{C}^* = (\psi_1, \ldots, \psi_m)$. The map T has the $m \times n$ matrix $[T]_{\mathcal{B}}^{\mathcal{C}}$ (Definition 4.1.1). The dual map $T' \colon W^* \to V^*$ has input basis $\mathcal{C}^*$ and output basis $\mathcal{B}^*$, so its matrix is $[T']_{\mathcal{C}^*}^{\mathcal{B}^*}$, of size $n \times m$. The size already suggests the answer, and Example (Dual Maps in Four Settings) (a) suggests it for matrix maps.

Theorem 5.4.2. The Matrix of the Dual Map Is the Transpose

Let V and W be finite-dimensional vector spaces over F, with ordered bases $\mathcal{B}$ of V and $\mathcal{C}$ of W, and let $\mathcal{B}^*$ and $\mathcal{C}^*$ be their dual bases. For every $T \in \mathcal{L}(V, W)$,

$$[T']_{\mathcal{C}^*}^{\mathcal{B}^*} = \left([T]_{\mathcal{B}}^{\mathcal{C}}\right)^{\top}.$$

Idea. Both matrices are tables of the same numbers $\psi_i(Tv_j)$. In $[T]_{\mathcal{B}}^{\mathcal{C}}$, the (i, j) entry is the w_i-coordinate of Tv_j, and a dual basis vector **reads off** a coordinate: that coordinate is

$\psi_i(Tv_j)$. In $[T']_{\mathcal{C}^*}^{\mathcal{B}^*}$, column i holds the $\mathcal{B}^*$-coordinates of $T'\psi_i$, and the φ_j-coordinate of a functional is its value at v_j: that is $(T'\psi_i)(v_j) = \psi_i(Tv_j)$. The same number sits at (i,j) in one matrix and at (j,i) in the other.

Proof. Write $\mathcal{B} = (v_1, \dots, v_n)$, $\mathcal{B}^* = (\varphi_1, \dots, \varphi_n)$, $\mathcal{C} = (w_1, \dots, w_m)$, $\mathcal{C}^* = (\psi_1, \dots, \psi_m)$, and $\mathbf{A} = (a_{kj}) = [T]_{\mathcal{B}}^{\mathcal{C}}$. By Definition 4.1.1, a_{ij} is the i-th entry of $[Tv_j]_{\mathcal{C}}$, and by Theorem 5.1.3 (a) that entry is read off by ψ_i:

$$\psi_i(Tv_j) = a_{ij}. \tag{$*$}$$

Next, by Theorem 5.1.3 (c), every $\varphi \in V^*$ satisfies $\varphi = \sum_{j=1}^{n} \varphi(v_j)\varphi_j$. Since $\mathcal{B}^*$ is a basis, coordinates in it are unique (Theorem 1.5.1), so $[\varphi]_{\mathcal{B}^*} = (\varphi(v_1), \dots, \varphi(v_n))$.

By Definition 4.1.1, column i of $[T']_{\mathcal{C}^*}^{\mathcal{B}^*}$ is $[T'\psi_i]_{\mathcal{B}^*}$. By the previous paragraph and $(*)$, its j-th entry is

$$(T'\psi_i)(v_j) = \psi_i(Tv_j) = a_{ij}.$$

So the (j,i) entry of $[T']_{\mathcal{C}^*}^{\mathcal{B}^*}$ is a_{ij}, which is the (j,i) entry of $\mathbf{A}^\top$ (Definition 0.9.10). Both matrices are $n \times m$, so $[T']_{\mathcal{C}^*}^{\mathcal{B}^*} = \mathbf{A}^\top$. This proves the theorem. ∎

This is the meaning of the transpose. In Chapter 0, Definition 0.9.10 was a rule for rearranging entries. Now we see it as the matrix shadow of pulling functionals back: **transposing a matrix is dualizing the map**, provided we use dual bases on both sides. Every property of the transpose becomes a property of dual maps, and conversely; for instance, Theorem 5.4.1 (b) and this theorem together give $(\mathbf{AB})^\top = \mathbf{B}^\top\mathbf{A}^\top$ a second time (Exercise B3 below).

Example. A 2×3 Example

Let $T\colon \mathbb{R}[x]_{\le 2} \to \mathbb{R}^2$, $T(p) = (p(1), p'(1))$. Use the basis $\mathcal{B} = (1, x, x^2)$, whose dual basis $\mathcal{B}^* = (\varphi_0, \varphi_1, \varphi_2)$ consists of $\varphi_0(p) = p(0)$, $\varphi_1(p) = p'(0)$, $\varphi_2(p) = p''(0)/2$ (Example (Dual Bases in $\mathbb{R}[x]_{\le 2}$)), and the standard basis $\mathcal{C} = (\mathbf{e}_1, \mathbf{e}_2)$ of $\mathbb{R}^2$, with dual basis $\mathcal{C}^* = (\eta_1, \eta_2)$, $\eta_1(u,v) = u$, $\eta_2(u,v) = v$.

(a) Find $\mathbf{A} = [T]_{\mathcal{B}}^{\mathcal{C}}$.

(b) Compute $T'\eta_1$ and $T'\eta_2$ directly as combinations of $\varphi_0, \varphi_1, \varphi_2$, write down $[T']_{\mathcal{C}^*}^{\mathcal{B}^*}$, and compare with $\mathbf{A}^\top$.

(c) Use the matrix to compute $T'\psi$ for $\psi = -2\eta_1 + \eta_2$, and check the answer directly.

Solution.

(a) $T(1) = (1,0)$, $T(x) = (1,1)$, $T(x^2) = (1,2)$. These are the columns:

$$\mathbf{A} = [T]_{\mathcal{B}}^{\mathcal{C}} = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \end{pmatrix}.$$

(b) Let $p = a + bx + cx^2$, so $\varphi_0(p) = a$, $\varphi_1(p) = b$, $\varphi_2(p) = c$. Then

$$(T'\eta_1)(p) = \eta_1(Tp) = p(1) = a + b + c,$$
$$(T'\eta_2)(p) = \eta_2(Tp) = p'(1) = b + 2c.$$

So $T'\eta_1 = \varphi_0 + \varphi_1 + \varphi_2$ and $T'\eta_2 = \varphi_1 + 2\varphi_2$. Their coordinate vectors in $\mathcal{B}^*$ are the columns of

$$[T']_{\mathcal{C}^*}^{\mathcal{B}^*} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{pmatrix} = \mathbf{A}^\top,$$

as Theorem 5.4.2 predicts. Read in words, the first column is Taylor's formula $p(1) = p(0) + p'(0) + \frac{1}{2}p''(0)$ for quadratics, and the second is $p'(1) = p'(0) + p''(0)$.

(c) $[\psi]_{\mathcal{C}^*} = (-2, 1)$. By Theorem 4.1.1,

$$[T'\psi]_{\mathcal{B}^*} = \mathbf{A}^\top \begin{pmatrix} -2 \\ 1 \end{pmatrix} = \begin{pmatrix} -2 \\ -1 \\ 0 \end{pmatrix},$$

so $T'\psi = -2\varphi_0 - \varphi_1$. Directly: $(T'\psi)(p) = -2p(1) + p'(1) = -2(a + b + c) + (b + 2c) = -2a - b$, which is $(-2\varphi_0 - \varphi_1)(p)$.

Warning. **Dual bases on both sides.** The transpose rule needs the **dual** bases $\mathcal{C}^*$ and $\mathcal{B}^*$. With other bases of W^* and V^* the matrix of T' is usually not $\mathbf{A}^\top$. In Example (A 2×3 Example), keep $\mathcal{B}^*$ but use the basis $(\eta_1 + \eta_2, \eta_2)$ of $(\mathbb{R}^2)^*$: then $T'(\eta_1 + \eta_2) = \varphi_0 + 2\varphi_1 + 3\varphi_2$, so the first column is $(1, 2, 3)$, not the first column $(1, 1, 1)$ of $\mathbf{A}^\top$.

5.4.3 Kernel and image of the dual map

The kernel and image of T' are subspaces of W^* and V^*. The natural candidates for them are annihilators of the subspaces that T defines, and the dictionary swaps kernel and image.

Theorem 5.4.3. Kernel and Image of the Dual Map

Let V and W be vector spaces over F, and let $T \in \mathcal{L}(V, W)$.

(a) $\ker T' = (\operatorname{im} T)^0$.

(b) If V and W are finite-dimensional, then $\operatorname{im} T' = (\ker T)^0$.

Idea.

(a) is a chain of definitions: $\psi T = 0$ says that ψ vanishes on every Tv, that is, on $\operatorname{im} T$. For (b), one inclusion is just as easy: ψT kills every vector that T kills. The reverse inclusion asks us to **extend** a functional given on V to one of the form ψT, which is a construction. Instead we count. Rank–Nullity for T', together with (a) and the dimension formula for annihilators, gives $\dim \operatorname{im} T' = \operatorname{rank} T$; the dimension formula

in V and Rank–Nullity for T give $\dim(\ker T)^0 = \operatorname{rank} T$. Finite dimension is spent exactly on these counts.

Proof.

(a) Let $\psi \in W^*$. Then $\psi \in \ker T'$ if and only if $\psi T = 0$ if and only if $\psi(Tv) = 0$ for every $v \in V$ if and only if $\psi(w) = 0$ for every $w \in \operatorname{im} T$ (Definition 2.2.2) if and only if $\psi \in (\operatorname{im} T)^0$ (Definition 5.2.1).

(b) Let $n = \dim V$ and $m = \dim W$, so $\dim V^* = n$ and $\dim W^* = m$ (Corollary 5.1.4).

Inclusion. Let $\varphi \in \operatorname{im} T'$, say $\varphi = \psi T$ with $\psi \in W^*$. For $u \in \ker T$, $\varphi(u) = \psi(Tu) = \psi(0) = 0$. Hence $\varphi \in (\ker T)^0$, and $\operatorname{im} T' \subseteq (\ker T)^0$.

Dimensions. By the Rank–Nullity Theorem (Theorem 2.3.1) for $T' \colon W^* \to V^*$, then (a), then Theorem 5.2.2 for the subspace $\operatorname{im} T$ of W,

$$\begin{aligned} \dim \operatorname{im} T' &= m - \dim \ker T' = m - \dim(\operatorname{im} T)^0 \\ &= m - (m - \dim \operatorname{im} T) = \operatorname{rank} T. \end{aligned}$$

By Theorem 5.2.2 for the subspace $\ker T$ of V, and Rank–Nullity for T,

$$\dim(\ker T)^0 = n - \dim \ker T = \operatorname{rank} T.$$

So $\operatorname{im} T'$ is a subspace of the finite-dimensional space $(\ker T)^0$ with the same dimension, and Theorem 1.6.8 gives $\operatorname{im} T' = (\ker T)^0$. This proves the theorem.

■

For a matrix map, this is a statement about rows. By Example (Dual Maps in Four Settings) (a), $T'_{\mathbf{A}}$ sends the row $\mathbf{b}^\top$ to $\mathbf{b}^\top\mathbf{A}$, so $\operatorname{im} T'_{\mathbf{A}}$ is the set of combinations of the rows of $\mathbf{A}$: it is the row space, read as functionals. Part (b) then says that **the functionals vanishing on** $\operatorname{nul}(\mathbf{A})$ **are exactly the combinations of the rows of** $\mathbf{A}$. Part (a) says that the functionals vanishing on $\operatorname{col}(\mathbf{A})$ are the rows $\mathbf{b}^\top$ with $\mathbf{b}^\top\mathbf{A} = \mathbf{0}^\top$, that is, $\mathbf{A}^\top\mathbf{b} = \mathbf{0}$.

Example. Kernel and Image of a Dual Map

Let $T = T_{\mathbf{A}} \colon \mathbb{R}^3 \to \mathbb{R}^3$ with

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 2 \\ 0 & 1 & -1 \\ 1 & 1 & 1 \end{pmatrix}.$$

Find bases of $\ker T'$ and $\operatorname{im} T'$, written as explicit functionals, and check Theorem 5.4.3.

Solution. The matrix world. The third row of $\mathbf{A}$ is the sum of the first two, and subtracting them leaves the RREF $\begin{pmatrix} 1 & 0 & 2 \\ 0 & 1 & -1 \\ 0 & 0 & 0 \end{pmatrix}$. So $\operatorname{rank} \mathbf{A} = 2$, and $\mathbf{Ax} = \mathbf{0}$ gives $x = -2z$, $y = z$: $\ker T = \operatorname{span}((-2, 1, 1))$. The pivot columns give $\operatorname{im} T = \operatorname{span}((1, 0, 1), (0, 1, 1))$ (Theorem 3.5.2).

Kernel of T'. By Theorem 5.4.2 with standard bases and Example (Matrix Maps and the Identity) (a), a functional $\psi(u,v,w) = b_1u + b_2v + b_3w$ lies in $\ker T'$ if and only if $\mathbf{A}^\top\mathbf{b} = \mathbf{0}$:

$$b_1 + b_3 = 0, \qquad b_2 + b_3 = 0, \qquad 2b_1 - b_2 + b_3 = 0.$$

The first two give $\mathbf{b} = t(1,1,-1)$, and these satisfy the third. So $\ker T' = \operatorname{span}(\psi_0)$ with $\psi_0(u,v,w) = u + v - w$. *Check:* $\psi_0(1,0,1) = 0$ and $\psi_0(0,1,1) = 0$, so ψ_0 vanishes on $\operatorname{im} T$; and $\dim(\operatorname{im} T)^0 = 3 - 2 = 1$ (Theorem 5.2.2). So $\ker T' = (\operatorname{im} T)^0$.

Image of T'. By Example (Dual Maps in Four Settings) (a), $\operatorname{im} T'$ is spanned by the rows of $\mathbf{A}$ read as functionals, and the first two rows already span (the third is their sum):

$$\varphi_1(x,y,z) = x + 2z, \qquad \varphi_2(x,y,z) = y - z.$$

They are independent (neither is a multiple of the other), so (φ_1, φ_2) is a basis of $\operatorname{im} T'$. *Check:* $\varphi_1(-2,1,1) = 0$ and $\varphi_2(-2,1,1) = 0$, so both vanish on $\ker T$, and $\dim(\ker T)^0 = 3 - 1 = 2$. So $\operatorname{im} T' = (\ker T)^0$.

Counting dimensions in the proof of Theorem 5.4.3 already gave $\dim \operatorname{im} T' = \operatorname{rank} T$. We record it, together with its most famous consequence.

Corollary 5.4.4. Rank of the Dual Map

(a) Let V and W be finite-dimensional vector spaces over F, and $T \in \mathcal{L}(V,W)$. Then $\operatorname{rank} T' = \operatorname{rank} T$.

(b) For every $\mathbf{A} \in M_{m\times n}(F)$, $\operatorname{rank} \mathbf{A}^\top = \operatorname{rank} \mathbf{A}$. Equivalently, $\dim \operatorname{row}(\mathbf{A}) = \dim \operatorname{col}(\mathbf{A})$: **row rank equals column rank**.

Proof.

(a) This is the first display in the proof of Theorem 5.4.3 (b).

(b) Let $\mathcal{E}_n$, $\mathcal{E}_m$ be the standard bases. By Example (Matrix Maps and the Identity) (a), $[T_{\mathbf{A}}]_{\mathcal{E}_n}^{\mathcal{E}_m} = \mathbf{A}$, so by Theorem 5.4.2, $[T'_{\mathbf{A}}]_{\mathcal{E}_m^*}^{\mathcal{E}_n^*} = \mathbf{A}^\top$. By Theorem 4.1.5 (a), applied to $T_{\mathbf{A}}$ and to $T'_{\mathbf{A}}$, and by (a),

$$\operatorname{rank} \mathbf{A} = \operatorname{rank} T_{\mathbf{A}} = \operatorname{rank} T'_{\mathbf{A}} = \operatorname{rank} \mathbf{A}^\top.$$

Finally, the transpose $M_{1\times n}(F) \to F^n$ is linear (Theorem 0.9.5) and bijective (it is its own inverse), and it sends row i of $\mathbf{A}$ to column i of $\mathbf{A}^\top$. By Theorem 2.1.4 it maps $\operatorname{row}(\mathbf{A})$ onto $\operatorname{col}(\mathbf{A}^\top)$, so these spaces are isomorphic and $\dim \operatorname{row}(\mathbf{A}) = \dim \operatorname{col}(\mathbf{A}^\top) = \operatorname{rank} \mathbf{A}^\top = \operatorname{rank} \mathbf{A} = \dim \operatorname{col}(\mathbf{A})$ (Theorem 2.5.4, Definition 3.5.4).

■

This is the third proof of this fact in the book, and each one explains it differently.

- **Chapter 3** (Theorem 3.5.5) counted pivots. Row reduction preserves the row space and the relations among the columns, and in the RREF both dimensions visibly equal the number of pivots. The proof is an algorithm.
- **Chapter 4, §03** factored $\mathbf{A} = \mathbf{BD}$ through a space of dimension $\dim \operatorname{col}(\mathbf{A})$, after

Proposition 4.3.2, and saw that the rows of $\mathbf{A}$ are combinations of the few rows of $\mathbf{D}$. The proof is a factorization.

- **Here** there are no row operations and no factorizations. The equality is the dimension formula $\dim U + \dim U^0 = \dim V$, applied once to $\operatorname{im} T$ in W and once to $\ker T$ in V, glued together by Rank–Nullity. The rows of $\mathbf{A}$ are the functionals that $T'_{\mathbf{A}}$ produces, and the columns span the image of $T_{\mathbf{A}}$; the two counts agree because pulling back measurements loses exactly as much as pushing forward vectors.

(The dimension formula for annihilators in section 2 was proved by extending a basis of U, without using ranks of matrices, so this proof does not go in a circle.)

The dictionary also swaps injective and surjective.

Corollary 5.4.5. Injective and Surjective Swap Under Duality

Let V and W be finite-dimensional vector spaces over F, and $T \in \mathcal{L}(V, W)$. Then:

(a) T is surjective if and only if T' is injective;

(b) T is injective if and only if T' is surjective.

Proof. Let $n = \dim V$ and $m = \dim W$. A finite-dimensional space is $\{0\}$ exactly when its dimension is 0.

(a) By Theorem 2.2.3 and Theorem 5.4.3 (a), T' is injective if and only if $(\operatorname{im} T)^0 = \{0\}$. By Theorem 5.2.2, $\dim(\operatorname{im} T)^0 = m - \operatorname{rank} T$, so this holds if and only if $\operatorname{rank} T = m$, that is, if and only if T is surjective (Corollary 2.3.2 (b)).

(b) T' is surjective if and only if $\operatorname{im} T' = V^*$, if and only if $\dim \operatorname{im} T' = n$ (by Theorem 1.6.8 for one direction, since $\dim V^* = n$). By Corollary 5.4.4 (a), $\dim \operatorname{im} T' = \operatorname{rank} T$, so this holds if and only if $\operatorname{rank} T = n$, that is, if and only if T is injective (Corollary 2.3.2 (a)).

■

For $T_{\mathbf{A}}$, part (a) reads: $\mathbf{Ax} = \mathbf{b}$ is solvable for every $\mathbf{b}$ exactly when $\mathbf{A}^\top \mathbf{y} = \mathbf{0}$ has only the trivial solution. So a question about the columns of $\mathbf{A}$ turns into a question about the columns of $\mathbf{A}^\top$, the rows of $\mathbf{A}$. Exercise C1 below sharpens this into a test for a single $\mathbf{b}$.

5.4.4 Naturality of the double dual

In the previous section we called $\mathrm{ev}_V : V \to V^{**}$ natural because it needs no choice, and promised a sharper sense: compatibility with every linear map. With dual maps available, we can state it. A linear map $T: V \to W$ has a dual $T': W^* \to V^*$, and that has a dual in turn,

$$T'' := (T')' : V^{**} \to W^{**},$$

which points **forward** again, because we reversed twice.

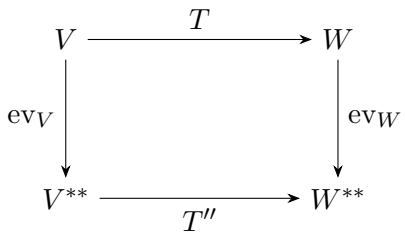
Theorem 5.4.6. The Evaluation Map Is Natural

Let V and W be vector spaces over F, and let $T \in \mathcal{L}(V, W)$. Then

$$T''\, \mathrm{ev}_V = \mathrm{ev}_W\, T.$$

That is, $T''(\mathrm{ev}_v) = \mathrm{ev}_{Tv}$ for every $v \in V$. If V and W are finite-dimensional, then $T'' = \mathrm{ev}_W \, T \, \mathrm{ev}_V^{-1}$.

The statement says that the square below commutes, for **every** linear map T: going right then down (apply T, then evaluate) equals going down then right (evaluate, then apply T'').



Proof. Let $v \in V$. Both $T''(\mathrm{ev}_v)$ and ev_{Tv} are functionals on W^*, so let $\psi \in W^*$. By Definition 5.4.1 applied to T', then to T,

$$(T''(\mathrm{ev}_v))(\psi) = \mathrm{ev}_v(T'\psi) = (T'\psi)(v) = \psi(Tv) = \mathrm{ev}_{Tv}(\psi).$$

Hence $T''(\mathrm{ev}_v) = \mathrm{ev}_{Tv}$ for every v, which is $T'' \, \mathrm{ev}_V = \mathrm{ev}_W \, T$. If V is finite-dimensional, ev_V is an isomorphism (Theorem 5.3.2); composing on the right with ev_V^{-1} gives $T'' = \mathrm{ev}_W \, T \, \mathrm{ev}_V^{-1}$.

■

So, once V is identified with V^{**} and W with W^{**} through evaluation, the double dual map T'' **is** T. The matrices agree: by Theorem 5.3.2, ev carries $\mathcal{B}$ and $\mathcal{C}$ to the dual bases $\mathcal{B}^{**}$ and $\mathcal{C}^{**}$ of $\mathcal{B}^*$ and $\mathcal{C}^*$, and two applications of Theorem 5.4.2 give $[T'']_{\mathcal{B}^{**}}^{\mathcal{C}^{**}} = (([T]_{\mathcal{B}}^{\mathcal{C}})^{\top})^{\top} = [T]_{\mathcal{B}}^{\mathcal{C}}$ (Theorem 0.9.5). The rule $(\mathbf{A}^{\top})^{\top} = \mathbf{A}$ is the matrix form of naturality.

Remark. No such statement holds for $V \cong V^*$. Suppose that for every finite-dimensional real space V we had an isomorphism $\theta_V : V \to V^*$, compatible with isomorphisms in the only way the directions allow: $\theta_V = T' \, \theta_W \, T$ for every isomorphism $T : V \to W$. Take $V = W = \mathbb{R}$ and $T = 2\,\mathrm{id}$. Then $T' = 2\,\mathrm{id}$ by Theorem 5.4.1 (a) and (c), so $\theta_{\mathbb{R}} = 2 \cdot 2 \cdot \theta_{\mathbb{R}} = 4\theta_{\mathbb{R}}$, which forces $\theta_{\mathbb{R}} = 0$, not an isomorphism. The dual space is isomorphic to V, but not naturally.

5.4.5 Exercises

5.4.5.1 A. Check your understanding

Exercise. A1

(a) Define the dual map T' of $T \in \mathcal{L}(V, W)$, including its domain and codomain.

(b) State the relation between $[T']_{\mathcal{C}^*}^{\mathcal{B}^*}$ and $[T]_{\mathcal{B}}^{\mathcal{C}}$, with its hypotheses.

(c) True or false: $(ST)' = S'T'$ whenever ST is defined. Justify your answer.

(d) True or false: $\ker T' = (\ker T)^0$ for every linear map between finite-dimensional spaces. Justify your answer.

(e) True or false: for $T \in \mathcal{L}(\mathbb{R}^4, \mathbb{R}^3)$, the dual map T' is never surjective. Justify your answer.

(f) Name the two tools combined to prove $\operatorname{rank} T' = \operatorname{rank} T$.

Solution.

(a) For $T\colon V \to W$, $T'\colon W^* \to V^*$, $T'(\psi) = \psi T$ (Definition 5.4.1).

(b) If V, W are finite-dimensional with bases $\mathcal{B}, \mathcal{C}$ and dual bases $\mathcal{B}^*, \mathcal{C}^*$, then $[T']_{\mathcal{C}^*}^{\mathcal{B}^*} = ([T]_{\mathcal{B}}^{\mathcal{C}})^\top$ (Theorem 5.4.2).

(c) False. The correct rule is $(ST)' = T'S'$ (Theorem 5.4.1 (b)); for $T\colon U \to V$, $S\colon V \to W$ with $U \neq W$, the composite $S'T'$ is not even defined.

(d) False. $\ker T' = (\operatorname{im} T)^0$. For the zero map $T\colon \mathbb{R} \to \mathbb{R}$, $\ker T = \mathbb{R}$, so $(\ker T)^0 = \{0\}$, while $T' = 0$ has kernel $\mathbb{R}^* \neq \{0\}$.

(e) True. T' maps $(\mathbb{R}^3)^*$, of dimension 3, to $(\mathbb{R}^4)^*$, of dimension 4, so it cannot be surjective (Corollary 2.3.2 (d)). This matches Corollary 5.4.5: T cannot be injective.

(f) The Rank–Nullity Theorem (Theorem 2.3.1) and the dimension formula for annihilators, $\dim U + \dim U^0 = \dim V$ (Theorem 5.2.2), together with $\ker T' = (\operatorname{im} T)^0$.

5.4.5.2 B. Practice

Exercise. B1: Differentiation in dual bases

Let $D\colon \mathbb{R}[x]_{\le 2} \to \mathbb{R}[x]_{\le 2}$, $D(p) = p'$, and let $\mathcal{B} = (1, x, x^2)$ with dual basis $\mathcal{B}^* = (\varphi_0, \varphi_1, \varphi_2)$, $\varphi_k(p) =$ the coefficient of x^k in p. Compute $D'\varphi_0$, $D'\varphi_1$, $D'\varphi_2$ directly, write down $[D']_{\mathcal{B}^*}^{\mathcal{B}^*}$, and check that it is $([D]_{\mathcal{B}}^{\mathcal{B}})^\top$. Hence describe $\ker D'$ and $\operatorname{im} D'$, and verify Theorem 5.4.3.

Solution. Let $p = a + bx + cx^2$, so $Dp = b + 2cx$. Then $(D'\varphi_0)(p) = \varphi_0(Dp) = b$, $(D'\varphi_1)(p) = 2c$, $(D'\varphi_2)(p) = 0$. Hence

$$D'\varphi_0 = \varphi_1, \qquad D'\varphi_1 = 2\varphi_2, \qquad D'\varphi_2 = 0.$$

The coordinate vectors $(0,1,0)$, $(0,0,2)$, $(0,0,0)$ are the columns of $[D']_{\mathcal{B}^*}^{\mathcal{B}^*}$. On the other side, $D(1) = 0$, $D(x) = 1$, $D(x^2) = 2x$, so

$$[D]_{\mathcal{B}}^{\mathcal{B}} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \end{pmatrix}, \qquad [D']_{\mathcal{B}^*}^{\mathcal{B}^*} = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 2 & 0 \end{pmatrix} = ([D]_{\mathcal{B}}^{\mathcal{B}})^\top,$$

as Theorem 5.4.2 predicts.

From the formulas, $D'(c_0\varphi_0 + c_1\varphi_1 + c_2\varphi_2) = c_0\varphi_1 + 2c_1\varphi_2$, which is 0 if and only if $c_0 = c_1 = 0$. So $\ker D' = \operatorname{span}(\varphi_2)$ and $\operatorname{im} D' = \operatorname{span}(\varphi_1, \varphi_2)$. On the other hand $\operatorname{im} D = \mathbb{R}[x]_{\le 1} = \operatorname{span}(1, x)$, and φ_2 is exactly the functional (up to scalars) vanishing on it, since $c_0\varphi_0 + c_1\varphi_1 + c_2\varphi_2$ kills 1 and x if and only if $c_0 = c_1 = 0$. So $\ker D' = (\operatorname{im} D)^0$. Also $\ker D = \operatorname{span}(1)$, and a functional kills 1 if and only if $c_0 = 0$, so $(\ker D)^0 = \operatorname{span}(\varphi_1, \varphi_2) = \operatorname{im} D'$.

Exercise. B2: Kernel and image of a dual map

Let $\mathbf{A} = \begin{pmatrix} 1 & 2 & -1 \\ -2 & -4 & 2 \end{pmatrix}$ and $T = T_{\mathbf{A}} \colon \mathbb{R}^3 \to \mathbb{R}^2$. Find bases of $\ker T'$ and $\operatorname{im} T'$ as explicit functionals, and verify $\dim \ker T' + \dim \operatorname{im} T' = 2$ and $\operatorname{rank} T' = \operatorname{rank} T$.

Solution. The second row is -2 times the first, so $\operatorname{rank} \mathbf{A} = 1$, $\operatorname{im} T = \operatorname{col}(\mathbf{A}) = \operatorname{span}((1,-2))$, and $\ker T = \{(x,y,z) : x + 2y - z = 0\}$, of dimension 2.

Kernel of T'. As in Example (Kernel and Image of a Dual Map), $\psi(u,v) = b_1 u + b_2 v$ lies in $\ker T'$ if and only if $\mathbf{A}^{\top}\mathbf{b} = \mathbf{0}$, that is, $b_1 - 2b_2 = 0$, $2b_1 - 4b_2 = 0$, $-b_1 + 2b_2 = 0$. So $\mathbf{b} = t(2,1)$, and $\ker T' = \operatorname{span}(\psi_0)$ with $\psi_0(u,v) = 2u + v$. Check: $\psi_0(1,-2) = 0$, so $\psi_0 \in (\operatorname{im} T)^0$, which has dimension $2 - 1 = 1$.

Image of T'. By Example (Dual Maps in Four Settings) (a), $\operatorname{im} T'$ is spanned by the rows of $\mathbf{A}$ read as functionals, $x + 2y - z$ and $-2x - 4y + 2z$. The second is -2 times the first, so $\operatorname{im} T' = \operatorname{span}(\varphi_0)$ with $\varphi_0(x,y,z) = x + 2y - z$. Check: φ_0 vanishes on $\ker T$ by the description of $\ker T$, and $\dim(\ker T)^0 = 3 - 2 = 1$.

Hence $\dim \ker T' + \dim \operatorname{im} T' = 1 + 1 = 2 = \dim(\mathbb{R}^2)^*$, and $\operatorname{rank} T' = 1 = \operatorname{rank} T$.

Exercise. B3: The transpose of a product, again

Let $\mathbf{A} \in M_{m\times n}(F)$ and $\mathbf{B} \in M_{n\times p}(F)$. Using $T_{\mathbf{A}} T_{\mathbf{B}} = T_{\mathbf{AB}}$ (Chapter 4, §01), Theorem 5.4.1 (b) and Theorem 5.4.2, give a new proof that $(\mathbf{AB})^{\top} = \mathbf{B}^{\top}\mathbf{A}^{\top}$.

Solution. Let $\mathcal{E}_p, \mathcal{E}_n, \mathcal{E}_m$ be the standard bases, with dual bases $\mathcal{E}_p^*, \mathcal{E}_n^*, \mathcal{E}_m^*$. By Example (Matrix Maps and the Identity) (a), $[T_{\mathbf{A}}]_{\mathcal{E}_n}^{\mathcal{E}_m} = \mathbf{A}$, $[T_{\mathbf{B}}]_{\mathcal{E}_p}^{\mathcal{E}_n} = \mathbf{B}$, and $[T_{\mathbf{AB}}]_{\mathcal{E}_p}^{\mathcal{E}_m} = \mathbf{AB}$. Since $T_{\mathbf{AB}} = T_{\mathbf{A}} T_{\mathbf{B}}$, Theorem 5.4.1 (b) gives $T'_{\mathbf{AB}} = T'_{\mathbf{B}} T'_{\mathbf{A}}$. Taking matrices, with Theorem 4.1.2 in the middle and Theorem 5.4.2 on the outside,

$$(\mathbf{AB})^{\top} = [T'_{\mathbf{AB}}]_{\mathcal{E}_m^*}^{\mathcal{E}_p^*} = [T'_{\mathbf{B}} T'_{\mathbf{A}}]_{\mathcal{E}_m^*}^{\mathcal{E}_p^*} = [T'_{\mathbf{B}}]_{\mathcal{E}_n^*}^{\mathcal{E}_p^*} [T'_{\mathbf{A}}]_{\mathcal{E}_m^*}^{\mathcal{E}_n^*} = \mathbf{B}^{\top}\mathbf{A}^{\top}.$$

The reversal of order is now explained: it is the reversal of arrows under duality.

5.4.5.3 C. Going deeper

Exercise. C1: The Fredholm alternative

(a) Let V, W be finite-dimensional, $T \in \mathcal{L}(V,W)$ and $w \in W$. Prove that $Tv = w$ has a solution $v \in V$ if and only if $\psi(w) = 0$ for every $\psi \in \ker T'$.

(b) Deduce: for $\mathbf{A} \in M_{m\times n}(F)$ and $\mathbf{b} \in F^m$, the system $\mathbf{Ax} = \mathbf{b}$ is consistent if and only if $\mathbf{y}^{\top}\mathbf{b} = 0$ for every $\mathbf{y} \in F^m$ with $\mathbf{A}^{\top}\mathbf{y} = \mathbf{0}$.

(c) Let $\mathbf{A} = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{pmatrix}$. Find a single linear condition on $\mathbf{b} = (b_1, b_2, b_3) \in \mathbb{R}^3$ that is equivalent to consistency of $\mathbf{Ax} = \mathbf{b}$.

Hint: for (a), combine Theorem 5.4.3 with the description of a subspace by its annihilator in the previous section.

Solution.

(a) $Tv = w$ has a solution if and only if $w \in \operatorname{im} T$. By Theorem 5.4.3 (a), $\ker T' = (\operatorname{im} T)^0$. By Corollary 5.3.4 applied to the subspace $\operatorname{im} T$ of the finite-dimensional space W,

$$\begin{aligned}\operatorname{im} T &= \{w \in W : \psi(w) = 0 \text{ for every } \psi \in (\operatorname{im} T)^0\} \\ &= \{w \in W : \psi(w) = 0 \text{ for every } \psi \in \ker T'\}.\end{aligned}$$

This is the claim.

(b) Apply (a) to $T_{\mathbf{A}} \colon F^n \to F^m$. By Theorem 5.1.1, every $\psi \in (F^m)^*$ is $\psi(\mathbf{u}) = \mathbf{y}^{\top}\mathbf{u}$ for a unique $\mathbf{y} \in F^m$. By Example (Dual Maps in Four Settings) (a), $T'_{\mathbf{A}}\psi$ is represented by $\mathbf{A}^{\top}\mathbf{y}$, and the zero functional by $\mathbf{0}$; by uniqueness of the representing vector, $\psi \in \ker T'_{\mathbf{A}}$ if and only if $\mathbf{A}^{\top}\mathbf{y} = \mathbf{0}$. For such ψ, $\psi(\mathbf{b}) = \mathbf{y}^{\top}\mathbf{b}$. So (a) says: $\mathbf{Ax} = \mathbf{b}$ is consistent if and only if $\mathbf{y}^{\top}\mathbf{b} = 0$ for every $\mathbf{y}$ with $\mathbf{A}^{\top}\mathbf{y} = \mathbf{0}$.

(c) $\mathbf{A}^{\top}\mathbf{y} = \mathbf{0}$ reads $y_1 + y_2 + y_3 = 0$ and $y_1 + 2y_2 + 3y_3 = 0$. Subtracting, $y_2 + 2y_3 = 0$, so $\mathbf{y} = t(1, -2, 1)$. By (b), $\mathbf{Ax} = \mathbf{b}$ is consistent if and only if $\mathbf{y}^{\top}\mathbf{b} = 0$ for all these $\mathbf{y}$, that is, if and only if

$$b_1 - 2b_2 + b_3 = 0.$$

In words, the system $x_1 + kx_2 = b_k$ $(k = 1, 2, 3)$ asks for a line through the points $(1, b_1), (2, b_2), (3, b_3)$, and it exists exactly when the second difference $b_3 - 2b_2 + b_1$ vanishes. No inner product was used: $\mathbf{y}^{\top}\mathbf{b}$ is just a functional applied to $\mathbf{b}$.

Exercise. C2: Invariant subspaces and annihilators

Let V be finite-dimensional, $T \in \mathcal{L}(V)$, and U a subspace of V. Recall that U is **T-invariant** if $Tu \in U$ for every $u \in U$. Prove that U is T-invariant if and only if U^0 is T'-invariant.

Solution. ($\Rightarrow$) Suppose U is T-invariant, and let $\varphi \in U^0$. For every $u \in U$, $(T'\varphi)(u) = \varphi(Tu) = 0$, because $Tu \in U$. Hence $T'\varphi \in U^0$, and U^0 is T'-invariant.

($\Leftarrow$) Suppose U^0 is T'-invariant, and let $u \in U$. For every $\varphi \in U^0$, $T'\varphi \in U^0$, so $\varphi(Tu) = (T'\varphi)(u) = 0$. Thus Tu is killed by every functional in U^0. By Corollary 5.3.4, $Tu \in U$. Hence U is T-invariant.

Exercise. C3: Duality without finite dimension

Let V and W be arbitrary vector spaces over F, and $T \in \mathcal{L}(V, W)$.

(a) Prove that if T is surjective, then T' is injective.

(b) Prove that if T is injective, then T' is surjective.

(c) Check (b) on $S \colon F[x] \to F[x]$, $S(p) = xp$, from Example (Dual Maps in Four Settings) (c): given $\varphi \in F[x]^*$, find ψ with $S'\psi = \varphi$. Is S' injective?

Hint: for (b), use Theorem 1.8.2 in W.

Solution.

(a) By Theorem 5.4.3 (a), which needs no finite dimension, $\ker T' = (\operatorname{im} T)^0 = W^0$. A functional vanishing on all of W is the zero functional, so $\ker T' = \{0\}$ and T' is injective (Theorem 2.2.3).

(b) Let $\varphi \in V^*$. By Theorem 1.8.3, V has a basis B. Since T is injective, T is injective on B, and $T(B)$ is linearly independent: a finite combination $\sum c_i T b_i = T(\sum c_i b_i) = 0$ with distinct b_i forces $\sum c_i b_i = 0$ (Theorem 2.2.3), hence all $c_i = 0$. By Theorem 1.8.2 there is a basis $C \supseteq T(B)$ of W. By Theorem 2.1.6, there is $\psi \in W^*$ with $\psi(Tb) = \varphi(b)$ for $b \in B$, and $\psi(c) = 0$ for $c \in C \setminus T(B)$; this is well defined because each element of $T(B)$ is Tb for exactly one b. Then ψT and φ are linear and agree on the basis B, so $T'\psi = \psi T = \varphi$ by the uniqueness part of Theorem 2.1.6. Hence T' is surjective.

(c) Here $B = \{1, x, x^2, \dots\}$ and $S(B) = \{x, x^2, \dots\}$, extended to a basis by adding 1. The recipe gives $\psi(x^{k+1}) = \varphi(x^k)$ for $k \geq 0$ and $\psi(1) = 0$. Then $(S'\psi)(x^k) = \psi(x^{k+1}) = \varphi(x^k)$, so $S'\psi = \varphi$. The dual map S' is not injective: $S'\varphi_0 = 0$ by Example (Dual Maps in Four Settings) (c). This matches the contrapositive of (a): S is not surjective, since $1 \notin \operatorname{im} S$, and indeed φ_0 is a nonzero functional vanishing on $\operatorname{im} S$.

5.5 Duality for Subspaces and Quotients

A subspace U of V gives two new vector spaces: U itself, and the quotient V/U of Chapter 4. Each has a dual space, and so far U^* and $(V/U)^*$ have nothing to do with V^*. This section connects them. The answer is a swap: the dual of the quotient is a **subspace** of V^*, namely the annihilator U^0, and the dual of the subspace is a **quotient** of V^*, namely V^*/U^0. Both identifications come from the First Isomorphism Theorem, and neither needs a basis. At the end we use them to describe hyperplanes, the subspaces cut out by a single equation, in any dimension.

5.5.1 Restricting functionals to a subspace

Let U be a subspace of V. A functional $\varphi \in V^*$ measures every vector of V, so in particular it measures the vectors of U. Forgetting the rest gives the **restriction** $\varphi|_U \colon U \to F$, $u \mapsto \varphi(u)$, which is a functional on U. This defines the restriction map

$$\rho \colon V^* \to U^*, \qquad \rho(\varphi) = \varphi|_U.$$

It is not a new kind of map. Let $\iota \colon U \to V$, $\iota(u) = u$, be the inclusion. Its dual map (Definition 5.4.1) sends φ to $\varphi \circ \iota$, and $(\varphi \circ \iota)(u) = \varphi(u)$. So $\rho = \iota'$, and in particular ρ is linear.

Two questions decide what ρ does. Which functionals restrict to zero? By Definition 5.2.1, exactly those in U^0. And does every functional on U come from one on V? That is an extension problem, and it has a positive answer in every dimension.

Theorem 5.5.1. Dual of a Subspace

Let V be a vector space over F, of any dimension, and let U be a subspace of V. The

restriction map $\rho\colon V^* \to U^*$, $\rho(\varphi) = \varphi|_U$, is linear and surjective, and $\ker \rho = U^0$. Hence

$$V^*/U^0 \to U^*, \qquad \varphi + U^0 \mapsto \varphi|_U$$

is a well-defined isomorphism, and $U^* \cong V^*/U^0$.

Idea. The kernel is the definition of U^0 read aloud. The real content is surjectivity: given ψ on U, we must extend it to all of V. A basis tells us how. Extend a basis of U to a basis of V; a linear map may be prescribed freely on a basis, so declare the extension to agree with ψ on the basis of U and to be 0 on the new basis vectors. Then the First Isomorphism Theorem turns "surjective with kernel U^0" into the isomorphism.

Proof. We saw above that $\rho = \iota'$ is linear. For $\varphi \in V^*$, $\rho(\varphi) = 0$ means $\varphi(u) = 0$ for every $u \in U$, which is $\varphi \in U^0$ by Definition 5.2.1. So $\ker \rho = U^0$.

Surjective. Let $\psi \in U^*$. By Theorem 1.8.3, U has a basis B_U. It is a linearly independent subset of V, so by Theorem 1.8.2 there is a basis B of V with $B_U \subseteq B$. By Theorem 2.1.6 there is a linear map $\varphi\colon V \to F$ with

$$\varphi(b) = \psi(b) \quad \text{for } b \in B_U, \qquad \varphi(b) = 0 \quad \text{for } b \in B \setminus B_U.$$

Then $\varphi|_U$ and ψ are linear maps $U \to F$ that agree on the basis B_U of U, so $\varphi|_U = \psi$ by the uniqueness part of Theorem 2.1.6. Thus $\rho(\varphi) = \psi$, and ρ is surjective. (If V is finite-dimensional, the finite versions Corollary 1.5.5, Theorem 1.6.5 and Theorem 2.1.5 suffice, and no choice principle is needed.)

By the First Isomorphism Theorem (Theorem 4.4.7), $\varphi + \ker \rho \mapsto \rho(\varphi)$ is a well-defined isomorphism $V^*/\ker \rho \to \operatorname{im} \rho$. Since $\ker \rho = U^0$ and $\operatorname{im} \rho = U^*$, this is the claimed isomorphism.

■

In words: **a functional on U is a functional on V, up to adding something that vanishes on U.** Two extensions of the same ψ differ by an element of U^0, and every element of U^0 can be added.

When V is finite-dimensional, taking dimensions recovers the count of this chapter. By Corollary 5.1.4, $\dim U^* = \dim U$ and $\dim V^* = \dim V$, and by Theorem 4.4.5, $\dim(V^*/U^0) = \dim V - \dim U^0$. So the isomorphism says $\dim U = \dim V - \dim U^0$, which is Theorem 5.2.2 again.

5.5.2 Functionals on a quotient

Now the other space. A functional ψ on V/U measures cosets. Composing with the quotient map $\pi\colon V \to V/U$, $\pi(v) = v + U$, turns it into a measurement of vectors, $v \mapsto \psi(v + U)$. This functional on V cannot tell apart two vectors in the same coset, so in particular it is zero on U. The theorem says that nothing is lost and nothing else appears: the functionals on V/U are **exactly** the functionals on V that vanish on U.

Theorem 5.5.2. Dual of a Quotient

Let V be a vector space over F, of any dimension, let U be a subspace, and let $\pi\colon V \to V/U$ be the quotient map. Then the dual map

$$\pi'\colon (V/U)^* \to V^*, \qquad \pi'(\psi) = \psi \circ \pi,$$

is injective, and its image is U^0. Hence π' is an isomorphism from $(V/U)^*$ onto U^0, and $(V/U)^* \cong U^0$.

Idea. Injectivity uses only that π is surjective: if $\psi \circ \pi$ is zero, then ψ is zero on every coset. For the image, one inclusion is "π kills U". The other inclusion asks us to build a functional on V/U from a functional on V that vanishes on U, and that is exactly what the universal property of the quotient does.

Proof. The map π' is linear by Definition 5.4.1.

Injective. Suppose $\pi'(\psi) = \psi \circ \pi = 0$. Every element of V/U is $\pi(v)$ for some $v \in V$, because π is surjective (Theorem 4.4.4 (c)). Then $\psi(\pi(v)) = 0$ for every v, so $\psi = 0$. By Theorem 2.2.3, π' is injective.

$\operatorname{im}\pi' \subseteq U^0$. For $u \in U$, $\pi(u) = U$ is the zero vector of V/U, so $(\psi \circ \pi)(u) = \psi(0) = 0$. Hence $\psi \circ \pi \in U^0$.

$U^0 \subseteq \operatorname{im}\pi'$. Let $\varphi \in U^0$. Then $U \subseteq \ker\varphi$, so by Theorem 4.4.10 (a) there is a linear map $\psi\colon V/U \to F$ with $\psi \circ \pi = \varphi$. This ψ lies in $(V/U)^*$, and $\pi'(\psi) = \varphi$.

So π' is an injective linear map with image U^0; regarded as a map onto U^0 it is a linear bijection, hence an isomorphism.

■

Concretely, the inverse isomorphism $U^0 \to (V/U)^*$ is $\varphi \mapsto \bar{\varphi}$, where $\bar{\varphi}(v + U) = \varphi(v)$. The formula is well defined precisely because φ vanishes on U: if $v - v' \in U$, then $\varphi(v) - \varphi(v') = \varphi(v - v') = 0$.

The two theorems fit together as a mirror:

	subspace $U \subseteq V$	quotient V/U
its dual is	a **quotient** of V^*: $U^* \cong V^*/U^0$	a **subspace** of V^*: $(V/U)^* \cong U^0$
the map	ι': restrict	π': compose with π
the map is	surjective	injective

The last row is no accident: ι is injective and π is surjective, and dualizing swaps the two properties.

Alternatively, in finite dimension both theorems follow from Theorem 5.4.3, which says $\ker T' = (\operatorname{im} T)^0$ and $\operatorname{im} T' = (\ker T)^0$. For $T = \pi$: $\operatorname{im}\pi = V/U$, whose annihilator is $\{0\}$, and

$\ker \pi = U$, so π' is injective with image U^0. For $T = \iota$: $\operatorname{im} \iota = U$, so $\ker \iota' = U^0$, and $\ker \iota = \{0\}$, whose annihilator is all of U^*, so ι' is surjective. The proofs above avoid dimension counts, which is why they work in every dimension.

Quick check. Let V be a vector space of dimension 6 and U a subspace of dimension 2. Find $\dim U^0$, $\dim(V/U)^*$, $\dim U^*$, and $\dim(V^*/U^0)$, and check that the two theorems are consistent with these numbers.

Solution. By Theorem 5.2.2, $\dim U^0 = 6 - 2 = 4$. By Theorem 4.4.5 and Corollary 5.1.4, $\dim(V/U)^* = \dim(V/U) = 4$, matching $(V/U)^* \cong U^0$. Next, $\dim U^* = \dim U = 2$, and $\dim(V^*/U^0) = 6 - 4 = 2$, matching $U^* \cong V^*/U^0$.

5.5.3 A line in $\mathbb{R}^3$, in coordinates

The theorems name their isomorphisms without choosing anything. To see them act, we now choose bases and compute. Recall that a functional on $\mathbb{R}^3$ is determined by its values on $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$ (Theorem 5.1.1). So every $\varphi \in (\mathbb{R}^3)^*$ has the form $\varphi(x, y, z) = ax + by + cz$, with $(a, b, c) = (\varphi(\mathbf{e}_1), \varphi(\mathbf{e}_2), \varphi(\mathbf{e}_3))$. We write such a functional simply as $ax + by + cz$.

Example. Both isomorphisms for a line

Let $\mathbf{u} = (1, 2, -1)$ and $U = \operatorname{span}(\mathbf{u}) \subseteq \mathbb{R}^3$.

(a) Find a basis of U^0.

(b) Find a basis of $\mathbb{R}^3/U$, its dual basis (ψ_1, ψ_2) of $(\mathbb{R}^3/U)^*$, and the images $\pi'(\psi_1), \pi'(\psi_2)$.

(c) Describe the restriction map $\rho\colon (\mathbb{R}^3)^* \to U^*$ and the isomorphism $(\mathbb{R}^3)^*/U^0 \to U^*$.

Solution.

(a) By Theorem 5.2.3 (b), $U^0 = \{\mathbf{u}\}^0$, so $ax + by + cz \in U^0$ exactly when $a + 2b - c = 0$, that is, $c = a + 2b$. Taking $(a, b) = (1, 0)$ and $(0, 1)$ gives

$$\varphi_1 = x + z, \qquad \varphi_2 = y + 2z.$$

They are independent (look at the coefficients of x and y), and $\dim U^0 = 3 - 1 = 2$ by Theorem 5.2.2, so (φ_1, φ_2) is a basis of U^0 (Theorem 1.6.4).

(b) The list $(\mathbf{u}, \mathbf{e}_1, \mathbf{e}_2)$ is independent: in $a\mathbf{u} + b\mathbf{e}_1 + c\mathbf{e}_2 = (a + b, 2a + c, -a) = \mathbf{0}$, the third entry gives $a = 0$, and then $b = c = 0$. It has length 3, so it is a basis of $\mathbb{R}^3$, and by Theorem 4.4.5 $(\mathbf{e}_1 + U, \mathbf{e}_2 + U)$ is a basis of $\mathbb{R}^3/U$. Its dual basis consists of the ψ_i with $\psi_i(\mathbf{e}_j + U) = \delta_{ij}$ (Theorem 5.1.3).
The functional $\pi'(\psi_1) = \psi_1 \circ \pi$ on $\mathbb{R}^3$ takes the values $\psi_1(\mathbf{e}_1 + U) = 1$ at $\mathbf{e}_1$, $\psi_1(\mathbf{e}_2 + U) = 0$ at $\mathbf{e}_2$, and $\psi_1(U) = 0$ at $\mathbf{u}$. Writing it as $ax + by + cz$, this says $a = 1$, $b = 0$, and $a + 2b - c = 0$, so $c = 1$. Similarly $\pi'(\psi_2)$ has $a = 0$, $b = 1$, $c = 2$. Hence

$$\pi'(\psi_1) = x + z = \varphi_1, \qquad \pi'(\psi_2) = y + 2z = \varphi_2.$$

So π' carries the dual basis of $(\mathbb{R}^3/U)^*$ onto the basis of U^0 found in (a), as Theo-

rem 5.5.2 predicts. Read backwards: $\psi_1((x,y,z)+U) = x+z$ and $\psi_2((x,y,z)+U) = y+2z$. These are well defined, since adding $t\mathbf{u} = (t, 2t, -t)$ changes $x+z$ by $t-t=0$ and $y+2z$ by $2t-2t=0$.

(c) The space U^* is one-dimensional, with basis the functional ξ determined by $\xi(\mathbf{u}) = 1$ (the dual basis of $(\mathbf{u})$). A functional $\varphi = ax+by+cz$ restricts to $\varphi|_U = \varphi(\mathbf{u})\,\xi = (a+2b-c)\,\xi$, since both sides take the value $a+2b-c$ at $\mathbf{u}$. So ρ is "evaluate at $\mathbf{u}$". Its kernel is $\{a+2b-c=0\} = U^0$, and it is surjective, since $\rho(x) = \xi$. The isomorphism of Theorem 5.5.1 is

$$(ax+by+cz)+U^0 \;\mapsto\; (a+2b-c)\,\xi.$$

For instance $x+U^0 = -z+U^0$, because $x-(-z) = \varphi_1 \in U^0$; both restrict to ξ.

5.5.4 Why "natural" matters here

In the example, bases entered only when we wanted numbers. The maps ρ and π' were defined before any basis was chosen: restrict a functional, or compose it with π. That is what makes the two isomorphisms **natural**, in the sense of the discussion of the double dual. They are part of the structure of V and U, not of a coordinate system.

There is a tempting shortcut that is **not** natural. Since U^* is a quotient of V^*, one might try to find a copy of U^* **inside** V^*, as a subspace. This can be done, but only by making a choice. Pick a complement W of U, so $V = U \oplus W$. A functional that vanishes on W is determined by its values on U, so the restriction ρ maps W^0 isomorphically onto U^* (Exercise (C2: Complements and a choice of copy of U^*)). But different complements give different subspaces W^0.

Warning. U^* **is not a subspace of** V^***.** Its elements are functionals on U, which cannot be evaluated at vectors outside U. To place U^* inside V^* one must extend each functional, and extensions are not unique. In the example above, the complement $W_1 = \operatorname{span}(\mathbf{e}_1, \mathbf{e}_2)$ of U gives $W_1^0 = \operatorname{span}(z)$, while $W_2 = \operatorname{span}(\mathbf{e}_1, \mathbf{e}_3)$ gives $W_2^0 = \operatorname{span}(y)$. These are different lines in $(\mathbb{R}^3)^*$, and both restrict isomorphically onto U^*: $z|_U = -\xi$ and $y|_U = 2\xi$. The choice-free statement is $U^* \cong V^*/U^0$, a quotient.

The same caution applies on the other side, in reverse. $(V/U)^*$ **is** naturally a subspace of V^*, namely U^0, but V/U itself is not a subspace of V. Duality turns the quotient, which has no natural home inside V, into a subspace, which does.

5.5.5 Codimension and hyperplanes

The quotient measures how far a subspace is from filling the whole space. In $\mathbb{R}^3$ a plane through the origin misses "one direction" and a line misses "two directions". For finite-dimensional V the number of missing directions is $\dim V - \dim U$. But this formula is useless in $F[x]$, where $\dim V$ and $\dim U$ may both be infinite, although $U = \{p : p(0) = 0\}$ misses only one direction: every polynomial is a constant plus an element of U. The quotient gives the right number in every case.

The codimension of U counts the directions in V that U is missing.

Definition 5.5.1. Codimension

Let U be a subspace of a vector space V over F. If V/U is finite-dimensional, the **codimension** of U in V is

$$\operatorname{codim} U := \dim(V/U).$$

If V/U is infinite-dimensional, we say that U has **infinite codimension**.

In words: we do not compare the sizes of U and V. We collapse U to zero and measure what is left. The codimension depends on the ambient space, so when that is unclear we write "codimension of U **in** V".

Definition 5.5.2. Hyperplane

A **hyperplane** in V is a subspace H of V of codimension **exactly** 1, that is, with $\dim(V/H) = 1$.

Examples.

- **Finite dimension.** If V is finite-dimensional, then $\operatorname{codim} U = \dim V - \dim U$ by Theorem 4.4.5. The hyperplanes of F^n are its subspaces of dimension $n-1$: the lines through $\mathbf{0}$ in F^2 and the planes through $\mathbf{0}$ in F^3.
- **Trace.** In $M_n(F)$ with $n \geq 1$, the trace-zero matrices are the kernel of $\operatorname{tr}\colon M_n(F) \to F$. This map is linear and surjective, since $\operatorname{tr}(c\mathbf{E}_{11}) = c$, so $M_n(F)/\ker\operatorname{tr} \cong F$ by Theorem 4.4.7, and the trace-zero matrices form a hyperplane.
- **Polynomials.** In $F[x]$, let $U = \{p : p(0) = 0\}$. Evaluation at 0 is a linear map $F[x] \to F$ with kernel U, and it is surjective because the constant c maps to c. By Theorem 4.4.7, $F[x]/U \cong F$, so U is a hyperplane in the infinite-dimensional space $F[x]$. In the same way $\{p : p(0) = p(1) = 0\}$ has codimension 2 (Exercise (B3: Dimension computations)).
- **Degenerate cases.** $\operatorname{codim} V = 0$, since V/V is the zero space, and $\operatorname{codim}\{0\} = \dim V$ when V is finite-dimensional. In particular V itself is **not** a hyperplane, and the only hyperplane of a one-dimensional space is $\{0\}$. The zero space $\{0\}$ has no hyperplanes at all.

Nonexample by minimal change. In $\mathbb{R}^3$, the plane $\{x+y+z=0\}$ is a hyperplane. Replace it by the line $\{x+y+z=0,\ x-y=0\} = \operatorname{span}((1,1,-2))$. It is still a subspace, but its codimension is $3-1=2$, so the clause "codimension **exactly** 1" fails. Replace it instead by $\{x+y+z=1\}$: now it is not a subspace at all, so it has no codimension. It is a coset of a hyperplane, an object we return to in the next section.

Why this definition. Codimension is the dual notion of dimension, and the duality theorems make this precise. If U has finite codimension, then by Theorem 5.5.2, Corollary 5.1.4, and Theorem 2.5.4,

$$\operatorname{codim} U = \dim(V/U) = \dim(V/U)^* = \dim U^0,$$

and this holds even when V is infinite-dimensional. So **dimension counts vectors, and codimension counts independent equations**. For finite-dimensional V this is Theorem 5.2.2 again: $\dim U^0 = \dim V - \dim U$.

Warning. **In infinite dimension a hyperplane need not be "smaller" than the whole space.** In $F[x]$, the hyperplane $U = \{p : p(0) = 0\} = \{xq : q \in F[x]\}$ is isomorphic to $F[x]$ itself, via $q \mapsto xq$: this map is linear, injective (if $xq = 0$ then $q = 0$), and onto U. So

you cannot detect codimension by comparing U with V up to isomorphism, and the formula $\operatorname{codim} U = \dim V - \dim U$ must never be used outside finite dimension. Always compute $\dim(V/U)$, or count equations.

Hyperplanes are exactly the subspaces cut out by **one** nontrivial linear equation. In F^n this is familiar: $a_1x_1 + \cdots + a_nx_n = 0$ with the a_i not all zero defines a subspace of dimension $n-1$. The quotient makes the statement true in every dimension.

Theorem 5.5.3. Hyperplanes Are Kernels of Functionals

Let V be a vector space over F, of any dimension, and let H be a subspace of V. Then H is a hyperplane if and only if $H = \ker\varphi$ for some **nonzero** $\varphi \in V^*$.

Idea. A nonzero functional maps onto the one-dimensional space F, so the First Isomorphism Theorem makes $V/\ker\varphi$ one-dimensional. Conversely, if V/H is one-dimensional, it "is" F after choosing an isomorphism, and composing that isomorphism with π gives a functional whose kernel is H.

Proof. ($\Leftarrow$) Suppose $\varphi \in V^*$ is nonzero. Then $\operatorname{im}\varphi$ is a subspace of F (Theorem 2.2.2) containing a nonzero scalar, so $\dim\operatorname{im}\varphi = 1 = \dim F$, and $\operatorname{im}\varphi = F$ by Theorem 1.6.8. By Theorem 4.4.7, $V/\ker\varphi \cong F$, so $\dim(V/\ker\varphi) = 1$, and $\ker\varphi$ is a hyperplane.

($\Rightarrow$) Suppose $\dim(V/H) = 1$. By Theorem 2.5.4 there is an isomorphism $\theta\colon V/H \to F$. Let $\varphi = \theta \circ \pi$, where $\pi\colon V \to V/H$ is the quotient map. It is linear as a composition of linear maps (Theorem 2.4.2), so $\varphi \in V^*$. For $v \in V$, $\varphi(v) = 0$ means $\theta(v+H) = 0$, which holds exactly when $v + H$ is the zero coset, since θ is injective; and $v + H = H$ means $v \in H$ by Lemma 4.4.3 (b). So $\ker\varphi = H$. Finally $\varphi \neq 0$: since $V/H \neq \{0\}$, we have $H \neq V$, so some $v \notin H = \ker\varphi$.

■

The functional is not unique, but almost: two functionals with the same kernel are scalar multiples of each other (Exercise (C1: Functionals with the same kernel); Exercise (B2: Functionals with the same kernel) gives a second proof). So hyperplanes of V correspond to nonzero functionals **up to a nonzero scalar**, just as the plane $x + y + z = 0$ is also $2x + 2y + 2z = 0$.

A hyperplane together with any one vector outside it fills the space.

Remark. Let $H = \ker\varphi$ be a hyperplane and $v_0 \notin H$. Then $V = H \oplus \operatorname{span}(v_0)$. Indeed $\varphi(v_0) \neq 0$, and every $v \in V$ splits as

$$v = \left(v - \tfrac{\varphi(v)}{\varphi(v_0)}\, v_0\right) + \tfrac{\varphi(v)}{\varphi(v_0)}\, v_0,$$

where the first bracket lies in $\ker\varphi$ because φ of it is $\varphi(v) - \varphi(v) = 0$. And $H \cap \operatorname{span}(v_0) = \{0\}$, since $\varphi(cv_0) = c\varphi(v_0) = 0$ forces $c = 0$.

Quick check. Decide which of the following are hyperplanes, and for each hyperplane give a nonzero functional with that kernel.

(a) The symmetric matrices in $M_2(F)$.

(b) The symmetric matrices in $M_3(F)$.

(c) $\{p \in F[x] : p(1) = 0\}$ in $F[x]$.

Solution.

(a) Yes. $\mathbf{A}$ is symmetric exactly when $a_{12} - a_{21} = 0$, so the symmetric matrices are the kernel of $\mathbf{A} \mapsto a_{12} - a_{21}$, which is linear and nonzero (it takes the value 1 at $\mathbf{E}_{12}$). By Theorem 5.5.3 this is a hyperplane.

(b) No. The symmetric 3×3 matrices form a subspace of dimension 6 (choose the entries on and above the diagonal freely), so the codimension is $9 - 6 = 3 \neq 1$.

(c) Yes. It is the kernel of $p \mapsto p(1)$, which is linear (Theorem 0.8.5) and nonzero (it sends the constant 1 to 1).

5.5.6 Exercises

5.5.6.1 A. Check your understanding

Exercise. A1

(a) State the theorem identifying $(V/U)^*$ with a subspace of V^*, including the map.

(b) Define the codimension of a subspace, and define a hyperplane.

(c) True or false: U^* is a subspace of V^*. Justify your answer.

(d) Let U be a subspace of dimension 2 in F^5. Find $\dim(F^5/U)^*$ and $\operatorname{codim} U$.

(e) Is $\{(x_1, x_2, x_3, x_4) \in \mathbb{R}^4 : x_1 = x_2 = 0\}$ a hyperplane in $\mathbb{R}^4$? Justify your answer.

(f) True or false: every nonzero linear functional $\varphi\colon V \to F$ is surjective. Justify your answer.

Solution.

(a) For a subspace U of V with quotient map π, the dual map $\pi'\colon (V/U)^* \to V^*$, $\psi \mapsto \psi \circ \pi$, is injective with image U^0; so $(V/U)^* \cong U^0$ (Theorem 5.5.2).

(b) If V/U is finite-dimensional, $\operatorname{codim} U = \dim(V/U)$; otherwise U has infinite codimension. A hyperplane is a subspace of codimension exactly 1.

(c) False. Elements of U^* are defined only on U, so they are not elements of V^*. What is true is $U^* \cong V^*/U^0$, a quotient (Theorem 5.5.1).

(d) $\dim(F^5/U)^* = \dim(F^5/U) = 5 - 2 = 3$, by Corollary 5.1.4 and Theorem 4.4.5. So $\operatorname{codim} U = 3$.

(e) No. The subspace is $\operatorname{span}(\mathbf{e}_3, \mathbf{e}_4)$, of dimension 2, so its codimension is $4 - 2 = 2$.

(f) True, in any dimension. This is Proposition 5.1.2: $\operatorname{im}\varphi$ is a subspace of the one-dimensional space F containing a nonzero element, so it is F.

5.5.6.2 B. Practice

Exercise. B1: The dual of a quotient of $\mathbb{R}^4$

Let $U = \operatorname{span}((1,-1,0,2),(0,1,1,-1)) \subseteq \mathbb{R}^4$, and write functionals on $\mathbb{R}^4$ as $a_1x_1 + a_2x_2 + a_3x_3 + a_4x_4$.

(a) Find a basis (φ_1, φ_2) of U^0.

(b) Let $\psi_1, \psi_2 \in (\mathbb{R}^4/U)^*$ be the functionals with $\pi'(\psi_i) = \varphi_i$. Compute ψ_1 and ψ_2 at the coset $(2,0,1,3) + U$.

(c) Find a basis of $\mathbb{R}^4/U$ whose dual basis is (ψ_1, ψ_2).

Solution.

(a) A functional lies in U^0 exactly when it vanishes at both spanning vectors (Theorem 5.2.3 (b)), that is,

$$a_1 - a_2 + 2a_4 = 0, \qquad a_2 + a_3 - a_4 = 0.$$

The second gives $a_3 = a_4 - a_2$, and the first gives $a_1 = a_2 - 2a_4$. Taking $(a_2, a_4) = (1,0)$ and $(-1,-1)$ gives $(1,1,-1,0)$ and $(1,-1,0,-1)$, that is,

$$\varphi_1 = x_1 + x_2 - x_3, \qquad \varphi_2 = x_1 - x_2 - x_4.$$

Check: $\varphi_1(1,-1,0,2) = 0$, $\varphi_1(0,1,1,-1) = 0$, $\varphi_2(1,-1,0,2) = 1+1-2 = 0$, $\varphi_2(0,1,1,-1) = -1+1 = 0$. They are independent (compare the coefficients of x_3 and x_4). The spanning vectors of U are independent (compare first entries, then second), so $\dim U^0 = 4 - 2 = 2$ by Theorem 5.2.2, and (φ_1, φ_2) is a basis by Theorem 1.6.4.

(b) By Theorem 5.5.2, $\psi_i(\mathbf{v} + U) = \varphi_i(\mathbf{v})$. So $\psi_1((2,0,1,3) + U) = 2 + 0 - 1 = 1$ and $\psi_2((2,0,1,3) + U) = 2 - 0 - 3 = -1$.

(c) We need cosets c_1, c_2 with $\psi_i(c_j) = \delta_{ij}$. The vector $-\mathbf{e}_3$ has $\varphi_1(-\mathbf{e}_3) = 1$ and $\varphi_2(-\mathbf{e}_3) = 0$, and $-\mathbf{e}_4$ has $\varphi_1(-\mathbf{e}_4) = 0$, $\varphi_2(-\mathbf{e}_4) = 1$. Put $c_1 = -\mathbf{e}_3 + U$ and $c_2 = -\mathbf{e}_4 + U$. If $ac_1 + bc_2 = 0$, applying ψ_1 and ψ_2 gives $a = 0$ and $b = 0$; so (c_1, c_2) is independent, and since $\dim(\mathbb{R}^4/U) = 2$ it is a basis (Theorem 1.6.4). By construction $\psi_i(c_j) = \delta_{ij}$, so (ψ_1, ψ_2) is its dual basis (Theorem 5.1.3).

Exercise. B2: Functionals with the same kernel

Let V be a vector space over F, of any dimension, and let $\varphi, \psi \in V^*$ be nonzero. Prove that $\ker\varphi = \ker\psi$ if and only if $\psi = c\varphi$ for some nonzero $c \in F$. (Exercise (C1: Functionals with the same kernel) proved ($\Rightarrow$) by a different route; here use the hyperplane decomposition.)

Hint: the remark after Theorem 5.5.3.

Solution. ($\Leftarrow$) If $\psi = c\varphi$ with $c \neq 0$, then $\psi(v) = 0$ if and only if $c\varphi(v) = 0$, if and only if $\varphi(v) = 0$, since c is invertible. So $\ker\psi = \ker\varphi$.

($\Rightarrow$) Suppose $\ker\varphi = \ker\psi = H$. Since $\varphi \neq 0$, choose v_0 with $\varphi(v_0) \neq 0$; then $v_0 \notin H$, so also $\psi(v_0) \neq 0$. Put $c = \psi(v_0)/\varphi(v_0)$, which is nonzero, and $\chi = \psi - c\varphi \in V^*$. Then χ vanishes on H, because both ψ and φ do, and $\chi(v_0) = \psi(v_0) - c\varphi(v_0) = 0$. By Theorem 5.5.3 H is

a hyperplane, and by the remark after it $V = H \oplus \operatorname{span}(v_0)$. A linear map vanishing on H and on v_0 vanishes on every $h + av_0$, hence on V. So $\chi = 0$, that is, $\psi = c\varphi$.

Exercise. B3: Dimension computations

(a) Let $\dim V = 7$ and $\dim U = 3$. Find $\dim U^*$, $\dim(V^*/U^0)$, $\dim(V/U)^*$, and $\operatorname{codim} U$.

(b) In $V = F[x]_{\leq 3}$, let $U = \{p : p(0) = p(1) = 0\}$. Using the map $T\colon V \to F^2$, $T(p) = (p(0), p(1))$, find $\operatorname{codim} U$ and $\dim U$, and write U as an intersection of two hyperplanes.

(c) Hence find a basis of U^0 consisting of evaluation functionals.

Solution.

(a) $\dim U^* = 3$ by Corollary 5.1.4. $\dim U^0 = 7-3 = 4$ by Theorem 5.2.2, so $\dim(V^*/U^0) = 7 - 4 = 3$ (Theorem 4.4.5), consistent with Theorem 5.5.1. $\dim(V/U)^* = \dim(V/U) = 4$, consistent with Theorem 5.5.2, and $\operatorname{codim} U = 4$.

(b) T is linear because evaluation is (Theorem 0.8.5), and $\ker T = U$. It is surjective: $T(1) = (1, 1)$ and $T(x) = (0, 1)$ span F^2. By Theorem 4.4.7, $V/U \cong F^2$, so $\operatorname{codim} U = 2$, and $\dim U = 4 - 2 = 2$ by Theorem 4.4.5. Writing $\varepsilon_0(p) = p(0)$ and $\varepsilon_1(p) = p(1)$, we have $U = \ker \varepsilon_0 \cap \ker \varepsilon_1$, and both kernels are hyperplanes by Theorem 5.5.3, since both functionals send 1 to 1.

(c) Both ε_0 and ε_1 lie in U^0. They are independent: if $a\varepsilon_0 + b\varepsilon_1 = 0$, evaluating at x gives $b = 0$, and then evaluating at 1 gives $a = 0$. Since $\dim U^0 = 4 - 2 = 2$, $(\varepsilon_0, \varepsilon_1)$ is a basis of U^0 (Theorem 1.6.4).

5.5.6.3 C. Going deeper

Exercise. C1: Subspaces of finite codimension are intersections of hyperplanes

Let V be a vector space over F, of any dimension, and let U be a subspace of finite codimension $k \geq 1$.

(a) Prove that there are nonzero $\varphi_1, \dots, \varphi_k \in U^0$ with $U = \ker \varphi_1 \cap \cdots \cap \ker \varphi_k$. Deduce that U is an intersection of k hyperplanes.

(b) Prove that U is not an intersection of fewer than k hyperplanes.

Hint: for (a), take a dual basis of $(V/U)^*$; for (b), consider the map $V \to F^m$, $v \mapsto (\varphi_1(v), \dots, \varphi_m(v))$.

Solution.

(a) Let $\pi\colon V \to V/U$ be the quotient map. Since $\dim(V/U) = k$, choose a basis $(c_1, \dots, c_k)$ of V/U and let $(\psi_1, \dots, \psi_k)$ be its dual basis (Theorem 5.1.3). Put $\varphi_i = \psi_i \circ \pi$. By Theorem 5.5.2, $\varphi_i \in U^0$, and $\varphi_i \neq 0$ because π' is injective and $\psi_i \neq 0$.
Since each φ_i lies in U^0, $U \subseteq \ker \varphi_i$ for each i. Conversely, let $v \in \ker \varphi_1 \cap \cdots \cap \ker \varphi_k$, and write $v + U = \sum_j a_j c_j$. Applying ψ_i gives $a_i = \psi_i(v + U) = \varphi_i(v) = 0$ for every i. So $v + U$ is the zero coset, and $v \in U$ by Lemma 4.4.3 (b). This proves $U = \bigcap_i \ker \varphi_i$. Each $\ker \varphi_i$ is a hyperplane by Theorem 5.5.3.

(b) Suppose $U = H_1 \cap \cdots \cap H_m$ with hyperplanes H_i. By Theorem 5.5.3, $H_i = \ker \varphi_i$ for nonzero $\varphi_i \in V^*$. The map $S\colon V \to F^m$, $S(v) = (\varphi_1(v), \ldots, \varphi_m(v))$, is linear, since each entry is, and $\ker S = \bigcap_i \ker \varphi_i = U$. By Theorem 4.4.7, $V/U \cong \operatorname{im} S$, a subspace of F^m. Hence $k = \dim(V/U) = \dim \operatorname{im} S \leq m$ by Theorem 1.6.7. So at least k hyperplanes are needed.

Exercise. C2: Complements and a choice of copy of U^*

Let V be finite-dimensional, and let U, W be subspaces with $V = U \oplus W$.

(a) Prove that $V^* = U^0 \oplus W^0$ without using Theorem 5.2.3 (d). (This is Exercise (C3: Dual of a direct sum decomposition) (a), which used it.)

(b) Deduce that the restriction map $\rho\colon V^* \to U^*$ restricts to an isomorphism $W^0 \to U^*$.

(c) In $\mathbb{R}^2$, let $U = \operatorname{span}(\mathbf{e}_1)$, $W_1 = \operatorname{span}(\mathbf{e}_2)$, and $W_2 = \operatorname{span}((1,1))$. Find W_1^0 and W_2^0, and find the element of each that restricts to the functional $\xi \in U^*$ with $\xi(\mathbf{e}_1) = 1$. Conclude that the copy of U^* inside V^* depends on the complement.

Solution.

(a) Let $n = \dim V$. If $\varphi \in U^0 \cap W^0$, then φ vanishes on U and on W, hence on every $u + w$, that is, on $U + W = V$; so $\varphi = 0$. By Theorem 5.2.2, $\dim U^0 + \dim W^0 = (n - \dim U) + (n - \dim W) = 2n - n = n$, using $\dim U + \dim W = n$ for a direct sum (Theorem 1.7.4). By Theorem 1.7.3, $\dim(U^0 + W^0) = n - 0 = n = \dim V^*$, so $U^0 + W^0 = V^*$ (Theorem 1.6.8), and the sum is direct because the intersection is $\{0\}$.

(b) The restriction of ρ to W^0 is linear. Its kernel is $W^0 \cap \ker \rho = W^0 \cap U^0 = \{0\}$ by Theorem 5.5.1 and (a), so it is injective. Its domain has dimension $n - \dim W = \dim U = \dim U^*$. An injective linear map between spaces of the same finite dimension is an isomorphism (Corollary 2.3.2 (e)).

(c) Write functionals on $\mathbb{R}^2$ as $ax + by$. W_1^0 consists of those with $b = 0$, so $W_1^0 = \operatorname{span}(x)$. W_2^0 consists of those with $a + b = 0$, so $W_2^0 = \operatorname{span}(x - y)$. Restricting to U: $x|_U$ and $(x - y)|_U$ both take the value 1 at $\mathbf{e}_1$, so both equal ξ. The elements x and $x - y$ are different, so the subspaces $W_1^0 \neq W_2^0$ of V^* are two different copies of U^*, one for each complement.

5.6 Hyperplanes and Linear Systems Revisited

Chapter 3 met subspaces of F^n in two disguises. A **parametric** description lists spanning vectors, as in $\operatorname{col}(\mathbf{A})$. An **implicit** description lists equations, as in $\operatorname{nul}(\mathbf{A})$. We moved from equations to spanning vectors by elimination, but we never asked whether the reverse is always possible, or how many equations it takes. Duality answers both questions at once: the equations of a subspace U are the elements of U^0. This section turns that sentence into an algorithm, proves once more that row rank and nullity add up to n, and ends the chapter with a dictionary between a space and its dual.

5.6.1 Row vectors are functionals

Recall from Theorem 5.1.1 how functionals on F^n look. For $\mathbf{a} = (a_1, \dots, a_n) \in F^n$, put

$$\varphi_{\mathbf{a}} \colon F^n \to F, \qquad \varphi_{\mathbf{a}}(\mathbf{x}) = \mathbf{a}^{\top}\mathbf{x} = a_1x_1 + \cdots + a_nx_n.$$

By that theorem, every $\varphi \in (F^n)^*$ is $\varphi_{\mathbf{a}}$ for exactly one $\mathbf{a}$, namely $a_j = \varphi(\mathbf{e}_j)$, and $\mathbf{a} \mapsto \varphi_{\mathbf{a}}$ is an isomorphism $F^n \to (F^n)^*$.

The functional $\varphi_{\mathbf{a}}$ is "multiply by the **row** vector $\mathbf{a}^{\top}$". So the identification is really between functionals and row vectors, and that is what makes a system of equations a dual object. Write the rows of $\mathbf{A} \in M_{m\times n}(F)$ as $\mathbf{a}_1^{\top}, \dots, \mathbf{a}_m^{\top}$ with $\mathbf{a}_i \in F^n$. The i-th entry of $\mathbf{Ax}$ is $\mathbf{a}_i^{\top}\mathbf{x} = \varphi_{\mathbf{a}_i}(\mathbf{x})$, so the homogeneous system $\mathbf{Ax} = \mathbf{0}$ says exactly that $\varphi_{\mathbf{a}_1}(\mathbf{x}) = \cdots = \varphi_{\mathbf{a}_m}(\mathbf{x}) = 0$. Hence

$$\operatorname{nul}(\mathbf{A}) = \ker \varphi_{\mathbf{a}_1} \cap \cdots \cap \ker \varphi_{\mathbf{a}_m}.$$

Each nonzero row contributes a hyperplane (Theorem 5.5.3), and the null space is their intersection. **An equation is a functional, and a system is a list of functionals.**

The identification uses the standard basis of F^n, and nothing else. Under it, the coordinate functionals $\mathbf{x} \mapsto x_i$, which form the dual basis of $(\mathbf{e}_1, \dots, \mathbf{e}_n)$, correspond to $\mathbf{e}_1, \dots, \mathbf{e}_n$, that is, to the rows $\mathbf{e}_1^{\top}, \dots, \mathbf{e}_n^{\top}$.

5.6.2 Every subspace is a solution set

To describe U by equations, we need equations that hold on U, that is, functionals in U^0, and enough of them that they hold **only** on U. The first requirement is the definition of U^0. The second follows from the separation of points, applied in the quotient.

Lemma 5.6.1. A subspace is cut out by its annihilator

Let V be a finite-dimensional vector space and U a subspace. For $v \in V$,

$$v \in U \quad \Longleftrightarrow \quad \varphi(v) = 0 \text{ for every } \varphi \in U^0.$$

If $(\varphi_1, \dots, \varphi_k)$ is a basis of U^0, then $U = \ker \varphi_1 \cap \cdots \cap \ker \varphi_k$.

Proof. ($\Rightarrow$) This is Definition 5.2.1.

($\Leftarrow$) Suppose $v \notin U$. Then $v + U$ is a nonzero vector of V/U by Lemma 4.4.3 (b), and V/U is finite-dimensional (Theorem 4.4.5). By Theorem 5.1.5 there is $\psi \in (V/U)^*$ with $\psi(v + U) \neq 0$. By Theorem 5.5.2, $\varphi = \psi \circ \pi$ lies in U^0, and $\varphi(v) = \psi(v + U) \neq 0$. This proves the contrapositive.

For the last statement, $U \subseteq \ker \varphi_i$ for each i since $\varphi_i \in U^0$. Conversely, if $\varphi_i(v) = 0$ for all i, then every $\varphi = \sum_i c_i\varphi_i \in U^0$ has $\varphi(v) = 0$, so $v \in U$ by ($\Leftarrow$).

■

The lemma is Corollary 5.3.4 in disguise. That corollary says $U^{00} = U$ under $V \cong V^{**}$; the lemma unpacks it into a statement about vectors. In F^n, with rows in place of functionals, it becomes the main theorem of this section.

Theorem 5.6.2. Every Subspace Is a Solution Set

Let U be a subspace of F^n with $\dim U = k$.

(a) There is a matrix $\mathbf{A} \in M_{(n-k)\times n}(F)$ with $U = \operatorname{nul}(\mathbf{A})$. One may take as rows of $\mathbf{A}$ the vectors $\mathbf{a}_1^\top, \dots, \mathbf{a}_{n-k}^\top$ for any $\mathbf{a}_i \in F^n$ such that $(\varphi_{\mathbf{a}_1}, \dots, \varphi_{\mathbf{a}_{n-k}})$ is a basis of U^0.

(b) If $\mathbf{B} \in M_{m\times n}(F)$ and $\operatorname{nul}(\mathbf{B}) = U$, then $m \geq n - k$.

So U is the solution set of a homogeneous system of exactly $n - k = \operatorname{codim} U$ equations, and of no system with fewer. (When $k = n$, that is $U = F^n$, this is the empty system, with no equations.)

Proof.

(a) By Theorem 5.2.2, $\dim U^0 = n - k$. Choose a basis of U^0; by the identification above it is $(\varphi_{\mathbf{a}_1}, \dots, \varphi_{\mathbf{a}_{n-k}})$ for some $\mathbf{a}_i \in F^n$. Let $\mathbf{A}$ have rows $\mathbf{a}_1^\top, \dots, \mathbf{a}_{n-k}^\top$. Then

$$\operatorname{nul}(\mathbf{A}) = \ker \varphi_{\mathbf{a}_1} \cap \cdots \cap \ker \varphi_{\mathbf{a}_{n-k}} = U,$$

where the second equality is Lemma 5.6.1.

(b) By Theorem 3.5.7, $\operatorname{rank} \mathbf{B} = n - \operatorname{nullity} \mathbf{B} = n - k$. By Theorem 3.5.5, $\operatorname{rank} \mathbf{B} = \dim \operatorname{row}(\mathbf{B})$. The row space is spanned by the m rows of $\mathbf{B}$, so $m \geq \dim \operatorname{row}(\mathbf{B})$ by Theorem 1.6.3 (b). Hence $m \geq n - k$.

■

The count is the one promised by codimension: **dimension counts free parameters, codimension counts independent equations**, and in F^n the two add up to n.

Example. Equations for a plane in $\mathbb{R}^4$

Let $U = \operatorname{span}(\mathbf{u}_1, \mathbf{u}_2) \subseteq \mathbb{R}^4$, where $\mathbf{u}_1 = (1, 1, 0, 1)$ and $\mathbf{u}_2 = (0, 1, 1, 1)$. Find a system of linear equations whose solution set is U.

Solution. The two vectors are independent (compare first entries, then third), so $\dim U = 2$, and we expect $4 - 2 = 2$ equations. For $\mathbf{a} = (a_1, a_2, a_3, a_4)$, the functional $\varphi_{\mathbf{a}}$ lies in U^0 exactly when it vanishes at both spanning vectors (Theorem 5.2.3 (b)):

$$a_1 + a_2 + a_4 = 0, \qquad a_2 + a_3 + a_4 = 0.$$

This is the system $\mathbf{Ma} = \mathbf{0}$ for the matrix $\mathbf{M}$ whose **rows** are $\mathbf{u}_1^\top$ and $\mathbf{u}_2^\top$. Subtracting the second row from the first gives the RREF

$$\mathbf{M} = \begin{pmatrix} 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 \end{pmatrix} \longrightarrow \begin{pmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 1 & 1 \end{pmatrix},$$

with free variables a_3, a_4. By Theorem 3.5.4, a basis of the solutions is $(1, -1, 1, 0)$ (from $a_3 = 1, a_4 = 0$) and $(0, -1, 0, 1)$ (from $a_3 = 0, a_4 = 1$). The corresponding functionals form

a basis of U^0, and by Theorem 5.6.2,

$$U = \{\mathbf{x} \in \mathbb{R}^4 : x_1 - x_2 + x_3 = 0, \quad x_4 - x_2 = 0\}.$$

Check. Both equations hold at $\mathbf{u}_1$: $1-1+0=0$ and $1-1=0$; and at $\mathbf{u}_2$: $0-1+1=0$ and $1-1=0$. The coefficient matrix $\begin{pmatrix} 1 & -1 & 1 & 0 \\ 0 & -1 & 0 & 1 \end{pmatrix}$ has rank 2, so its null space has dimension $4-2=2$, and it contains the independent vectors $\mathbf{u}_1, \mathbf{u}_2$; hence it equals U (Theorem 1.6.8). The implicit description says: U consists of the vectors with $x_4 = x_2$ and $x_1 + x_3 = x_2$.

5.6.3 From spans to equations and back

The example is an instance of a general procedure. Both directions of the conversion are "compute a null space", applied to transposed data.

Algorithm. **Parametric to implicit.** Input: vectors $\mathbf{v}_1, \dots, \mathbf{v}_p \in F^n$ spanning U.

1. Form $\mathbf{M} \in M_{p\times n}(F)$ whose i-th **row** is $\mathbf{v}_i^\top$.
2. Row reduce $\mathbf{M}$ and find a basis $\mathbf{n}_1, \dots, \mathbf{n}_d$ of $\operatorname{nul}(\mathbf{M})$ as in Theorem 3.5.4.
3. Output the matrix $\mathbf{A} \in M_{d\times n}(F)$ whose j-th row is $\mathbf{n}_j^\top$, that is, the equations $\mathbf{n}_j^\top \mathbf{x} = 0$ for $j = 1, \dots, d$.

Implicit to parametric. Input: $\mathbf{A} \in M_{m\times n}(F)$. Output: a basis of $\operatorname{nul}(\mathbf{A})$, found by row reduction as in Theorem 3.5.4.

The second half was proved correct in Chapter 3. The first half is new.

Theorem 5.6.3. Correctness of the Conversion

Let $\mathbf{v}_1, \dots, \mathbf{v}_p \in F^n$, $U = \operatorname{span}(\mathbf{v}_1, \dots, \mathbf{v}_p)$, and let $\mathbf{M}$ and $\mathbf{A}$ be as in the algorithm. Then $\operatorname{nul}(\mathbf{A}) = U$, and $\mathbf{A}$ has exactly $d = n - \dim U$ rows, which are linearly independent.

Idea. The only thing to see is what step 2 computes. The i-th entry of $\mathbf{Mn}$ is $\mathbf{v}_i^\top \mathbf{n} = \varphi_{\mathbf{n}}(\mathbf{v}_i)$. So $\mathbf{n} \in \operatorname{nul}(\mathbf{M})$ says that the functional $\varphi_{\mathbf{n}}$ vanishes at every $\mathbf{v}_i$, that is, $\varphi_{\mathbf{n}} \in U^0$. Step 2 therefore computes a basis of U^0, and the theorem of the previous subsection does the rest.

Proof. For $\mathbf{n} \in F^n$, the i-th entry of $\mathbf{Mn}$ is $\mathbf{v}_i^\top \mathbf{n} = \mathbf{n}^\top \mathbf{v}_i = \varphi_{\mathbf{n}}(\mathbf{v}_i)$, where the middle equality holds because both sides equal $\sum_j (\mathbf{v}_i)_j n_j$. Hence

$$\mathbf{n} \in \operatorname{nul}(\mathbf{M}) \iff \varphi_{\mathbf{n}}(\mathbf{v}_i) = 0 \text{ for all } i \iff \varphi_{\mathbf{n}} \in \{\mathbf{v}_1, \dots, \mathbf{v}_p\}^0 = U^0,$$

where the last equality is Theorem 5.2.3 (b). So the isomorphism $\mathbf{n} \mapsto \varphi_{\mathbf{n}}$ maps $\operatorname{nul}(\mathbf{M})$ onto U^0. An isomorphism carries a basis to a basis (Theorem 2.5.2), so $(\varphi_{\mathbf{n}_1}, \dots, \varphi_{\mathbf{n}_d})$ is a basis of U^0. In particular $d = \dim U^0 = n - \dim U$ by Theorem 5.2.2, and the $\mathbf{n}_j$, hence the rows $\mathbf{n}_j^\top$, are independent. By Theorem 5.6.2 (a), the matrix $\mathbf{A}$ with rows $\mathbf{n}_1^\top, \dots, \mathbf{n}_d^\top$ satisfies

$\operatorname{nul}(\mathbf{A}) = U$.

■

The algorithm does not need the $\mathbf{v}_i$ to be independent. Redundant spanning vectors only produce zero rows in the RREF of $\mathbf{M}$, and the count $d = n - \operatorname{rank}\mathbf{M}$ automatically uses $\dim U = \operatorname{rank}\mathbf{M}$.

Running both halves in turn is a useful self-check. Starting from $\mathbf{A}$ in Example (Equations for a plane in $\mathbb{R}^4$) and computing $\operatorname{nul}(\mathbf{A})$ must return a basis of U; starting from that basis and running the first half must return equations with the same solution set, though not necessarily the same equations.

Warning. **Equations for a subspace are far from unique, and more equations do not mean a smaller solution set.** In Example (Equations for a plane in $\mathbb{R}^4$), the system $x_1 - x_2 + x_3 = 0$, $x_4 - x_2 = 0$, $x_1 + x_3 - x_4 = 0$ has the **same** solution set U: the third equation is the sum of the first and the negative of the second, so it adds a functional already in U^0. Three equations in $\mathbb{R}^4$ cut out a two-dimensional subspace here, not a one-dimensional one. Each new equation lowers the dimension by one **only if** its functional is not a combination of the previous ones. A theorem later in this section makes this precise.

5.6.4 Affine solution sets

An inhomogeneous equation $\mathbf{a}^\top\mathbf{x} = b$ with $\mathbf{a} \neq \mathbf{0}$ defines a set that is not a subspace when $b \neq 0$. It is a translate of the hyperplane $H = \ker\varphi_{\mathbf{a}}$: if $\mathbf{p}$ is one solution, then by Theorem 3.1.3 the solutions are $\mathbf{p} + H$. We call a coset $\mathbf{p} + H$ of a hyperplane H (Definition 4.4.2) an **affine hyperplane**. In $\mathbb{R}^3$, the plane $x + y + z = 1$ is an affine hyperplane, parallel to the hyperplane $x + y + z = 0$.

A system $\mathbf{Ax} = \mathbf{b}$ with rows $\mathbf{a}_i^\top$ and right-hand sides b_i asks for the vectors lying on all the sets $\{\mathbf{x} : \mathbf{a}_i^\top\mathbf{x} = b_i\}$ at once. Each such set is an affine hyperplane if $\mathbf{a}_i \neq \mathbf{0}$; if $\mathbf{a}_i = \mathbf{0}$, it is all of F^n (when $b_i = 0$) or empty (when $b_i \neq 0$). So **the solution set of a linear system is an intersection of affine hyperplanes**, apart from these trivial rows. The duality theorems give the converse.

Proposition 5.6.4. Cosets as intersections of affine hyperplanes

Let U be a subspace of F^n of dimension $k < n$, and let $\mathbf{p} \in F^n$. Let $\mathbf{A} \in M_{(n-k)\times n}(F)$ be as in Theorem 5.6.2 (a), with rows $\mathbf{a}_1^\top, \dots, \mathbf{a}_{n-k}^\top$. Then

$$\mathbf{p} + U = \{\mathbf{x} \in F^n : \mathbf{Ax} = \mathbf{Ap}\} = \bigcap_{i=1}^{n-k} \{\mathbf{x} : \mathbf{a}_i^\top\mathbf{x} = \mathbf{a}_i^\top\mathbf{p}\},$$

an intersection of $n - k$ affine hyperplanes.

Proof. By Theorem 5.6.2, $\operatorname{nul}(\mathbf{A}) = U$. The vector $\mathbf{p}$ is one solution of $\mathbf{Ax} = \mathbf{Ap}$, so by Theorem 3.1.3 its solution set is $\mathbf{p} + \operatorname{nul}(\mathbf{A}) = \mathbf{p} + U$. The i-th equation reads $\mathbf{a}_i^\top\mathbf{x} = \mathbf{a}_i^\top\mathbf{p}$. Each $\varphi_{\mathbf{a}_i}$ is nonzero, because it belongs to a basis of U^0; so $\ker\varphi_{\mathbf{a}_i}$ is a hyperplane (Theorem 5.5.3), and $\{\mathbf{x} : \mathbf{a}_i^\top\mathbf{x} = \mathbf{a}_i^\top\mathbf{p}\} = \mathbf{p} + \ker\varphi_{\mathbf{a}_i}$ is an affine hyperplane, again by Theorem 3.1.3 applied to this one equation.

■

Unlike hyperplanes, which all contain $\mathbf{0}$, affine hyperplanes can be disjoint: $x_1 + x_2 = 0$ and $x_1 + x_2 = 1$ are parallel lines in $\mathbb{R}^2$. That is the geometric picture of an inconsistent system.

5.6.5 The row space annihilates the null space

In Chapter 3, Theorem 3.5.7 said $\operatorname{rank}\mathbf{A} + \operatorname{nullity}\mathbf{A} = n$, and its proof counted pivot and free columns in the RREF. The duality of this chapter explains the same equation without elimination: the rows of $\mathbf{A}$ are exactly the equations satisfied by the null space.

Theorem 5.6.5. The Row Space Is the Annihilator of the Null Space

Let $\mathbf{A} \in M_{m\times n}(F)$. Identifying a row vector $\mathbf{a}^\top \in M_{1\times n}(F)$ with the functional $\varphi_{\mathbf{a}} \in (F^n)^*$,

$$\{\varphi_{\mathbf{a}} : \mathbf{a}^\top \in \operatorname{row}(\mathbf{A})\} = \operatorname{nul}(\mathbf{A})^0.$$

Consequently $\dim \operatorname{row}(\mathbf{A}) + \dim \operatorname{nul}(\mathbf{A}) = n$.

Idea. Let $T = T_{\mathbf{A}} \colon F^n \to F^m$. Its kernel is $\operatorname{nul}(\mathbf{A})$, and Theorem 5.4.3 says $\operatorname{im} T' = (\ker T)^0$. So it is enough to see that $\operatorname{im} T'$ is the row space. Pull back a functional $\varphi_{\mathbf{y}}$ on F^m: the result is $\mathbf{x} \mapsto \mathbf{y}^\top\mathbf{A}\mathbf{x}$, which is multiplication by the row $\mathbf{y}^\top\mathbf{A}$, the combination of the rows of $\mathbf{A}$ with coefficients $y_1, \dots, y_m$. The transpose of Chapter 0 appears exactly here: $\mathbf{y}^\top\mathbf{A} = (\mathbf{A}^\top\mathbf{y})^\top$.

Proof. Let $T = T_{\mathbf{A}} \colon F^n \to F^m$, $T(\mathbf{x}) = \mathbf{A}\mathbf{x}$, so that $\ker T = \operatorname{nul}(\mathbf{A})$. Every functional on F^m is $\varphi_{\mathbf{y}}$ for a unique $\mathbf{y} \in F^m$, as recalled at the start of this section. By Definition 5.4.1, associativity of matrix multiplication, and Theorem 0.9.5,

$$\begin{aligned} T'(\varphi_{\mathbf{y}})(\mathbf{x}) &= \varphi_{\mathbf{y}}(\mathbf{A}\mathbf{x}) = \mathbf{y}^\top(\mathbf{A}\mathbf{x}) = (\mathbf{y}^\top\mathbf{A})\mathbf{x} \\ &= (\mathbf{A}^\top\mathbf{y})^\top\mathbf{x} = \varphi_{\mathbf{A}^\top\mathbf{y}}(\mathbf{x}) \qquad (\mathbf{x} \in F^n), \end{aligned}$$

so $T'(\varphi_{\mathbf{y}}) = \varphi_{\mathbf{A}^\top\mathbf{y}}$. The columns of $\mathbf{A}^\top$ are $\mathbf{a}_1, \dots, \mathbf{a}_m$, where $\mathbf{a}_i^\top$ are the rows of $\mathbf{A}$, so $\mathbf{A}^\top\mathbf{y} = y_1\mathbf{a}_1 + \cdots + y_m\mathbf{a}_m$ by Theorem 0.9.3. As $\mathbf{y}$ runs over F^m, these are all the combinations of $\mathbf{a}_1, \dots, \mathbf{a}_m$, and since transposing respects combinations (Theorem 0.9.5), they are exactly the vectors $\mathbf{a}$ with $\mathbf{a}^\top \in \operatorname{row}(\mathbf{A})$ (Definition 3.5.2). Hence

$$\operatorname{im} T' = \{\varphi_{\mathbf{a}} : \mathbf{a}^\top \in \operatorname{row}(\mathbf{A})\}.$$

By Theorem 5.4.3, $\operatorname{im} T' = (\ker T)^0 = \operatorname{nul}(\mathbf{A})^0$, which is the first claim.

For the count, $\mathbf{a}^\top \mapsto \varphi_{\mathbf{a}}$ is an isomorphism $M_{1\times n}(F) \to (F^n)^*$ (transposition followed by $\mathbf{a} \mapsto \varphi_{\mathbf{a}}$), so $\dim \operatorname{row}(\mathbf{A}) = \dim \operatorname{nul}(\mathbf{A})^0 = n - \dim \operatorname{nul}(\mathbf{A})$ by Theorem 5.2.2.

■

The count is Rank–Nullity for matrices once more, now as a statement about annihilators: $\operatorname{rank}\mathbf{A} = \dim \operatorname{row}(\mathbf{A})$ counts independent equations and $\operatorname{nullity}\mathbf{A}$ counts free parameters. It also says that the given equations generate **all** the equations satisfied by $\operatorname{nul}(\mathbf{A})$: any linear

equation that holds on the solution set of $\mathbf{Ax} = \mathbf{0}$ is a combination of the rows of $\mathbf{A}$.

Corollary 5.6.6. Same solutions, same row space

Let $\mathbf{A} \in M_{m\times n}(F)$ and $\mathbf{B} \in M_{p\times n}(F)$. Then $\operatorname{nul}(\mathbf{A}) = \operatorname{nul}(\mathbf{B})$ if and only if $\operatorname{row}(\mathbf{A}) = \operatorname{row}(\mathbf{B})$.

Proof. ($\Leftarrow$) If $\operatorname{row}(\mathbf{A}) = \operatorname{row}(\mathbf{B})$, then $\operatorname{nul}(\mathbf{A})^0 = \operatorname{nul}(\mathbf{B})^0$ by Theorem 5.6.5, so $\operatorname{nul}(\mathbf{A}) = \operatorname{nul}(\mathbf{B})$ by Corollary 5.3.5.

($\Rightarrow$) If $\operatorname{nul}(\mathbf{A}) = \operatorname{nul}(\mathbf{B})$, then $\{\varphi_{\mathbf{a}} : \mathbf{a}^\top \in \operatorname{row}(\mathbf{A})\} = \operatorname{nul}(\mathbf{A})^0 = \{\varphi_{\mathbf{a}} : \mathbf{a}^\top \in \operatorname{row}(\mathbf{B})\}$ by Theorem 5.6.5. Since $\mathbf{a}^\top \mapsto \varphi_{\mathbf{a}}$ is injective, $\operatorname{row}(\mathbf{A}) = \operatorname{row}(\mathbf{B})$.

■

Chapter 3 showed that row operations preserve the null space (Theorem 3.5.1). The corollary gives a converse in terms of row spaces: two systems in the same unknowns have the same solutions **exactly** when each equation of one is a combination of the equations of the other.

5.6.6 Sums and intersections by duality

Parametric and implicit descriptions are good at different things. A **sum** is easy parametrically: $U + W$ is spanned by a spanning list of U followed by one of W (Proposition 1.7.2). An **intersection** is easy implicitly: if $U = \operatorname{nul}(\mathbf{A})$ and $W = \operatorname{nul}(\mathbf{B})$, then $\mathbf{x} \in U \cap W$ exactly when $\mathbf{x}$ satisfies both systems, so

$$U \cap W = \operatorname{nul}\begin{pmatrix}\mathbf{A}\\ \mathbf{B}\end{pmatrix},$$

the null space of the matrix with the rows of $\mathbf{A}$ above the rows of $\mathbf{B}$. Duality explains why each is easy where it is. The rows of the stacked matrix span $\operatorname{row}(\mathbf{A}) + \operatorname{row}(\mathbf{B})$, which corresponds to $U^0 + W^0 = (U \cap W)^0$ by Theorem 5.6.5 and Theorem 5.2.3 (d); and $(U + W)^0 = U^0 \cap W^0$ says that the equations of a sum are the equations common to both pieces.

Example. An intersection computed from equations

In $\mathbb{R}^4$, let $U = \operatorname{span}((1,0,1,2),(0,1,1,1))$ and $W = \operatorname{span}((1,1,0,1),(1,3,0,1))$, as in Example (A basis of an intersection). Find $U \cap W$ by first finding equations for U and for W.

Solution. Equations for U. A row (a_1, a_2, a_3, a_4) annihilates both spanning vectors when $a_1 + a_3 + 2a_4 = 0$ and $a_2 + a_3 + a_4 = 0$. This system is already in RREF, with free a_3, a_4; the basic solutions are $(-1,-1,1,0)$ and $(-2,-1,0,1)$. Changing signs, which does not change the kernels,

$$U = \{x_1 + x_2 - x_3 = 0, \quad 2x_1 + x_2 - x_4 = 0\}.$$

Equations for W. Here $a_1 + a_2 + a_4 = 0$ and $a_1 + 3a_2 + a_4 = 0$. Subtracting gives $a_2 = 0$, and then $a_1 = -a_4$, with a_3 free. The basic solutions are $(0,0,1,0)$ and $(-1,0,0,1)$, so

$$W = \{x_3 = 0, \quad x_4 - x_1 = 0\}.$$

Intersection. Stack the four equations. From W: $x_3 = 0$ and $x_4 = x_1$. Then $x_1 + x_2 - x_3 = 0$ gives $x_2 = -x_1$, and the last equation $2x_1 + x_2 - x_4 = 2x_1 - x_1 - x_1 = 0$ holds automatically. So

$$U \cap W = \{(t, -t, 0, t) : t \in \mathbb{R}\} = \operatorname{span}((1, -1, 0, 1)),$$

in agreement with Chapter 3.

The redundancy of the fourth equation is not a coincidence. The four rows span $U^0 + W^0 = (U \cap W)^0$, which has dimension $4 - 1 = 3$, so exactly one of the four equations must be a consequence of the others.

Contrast with Chapter 3. There we put the four spanning vectors as **columns** of one matrix and read the intersection off a null vector, a relation $\mathbf{u}_1 - \mathbf{u}_2 = 2\mathbf{w}_1 - \mathbf{w}_2$. Here we computed two small null spaces (the equations) and then one more (the intersection). The equation method pays off when the subspaces are given by equations in the first place, or when several intersections with the same U are needed: the equations of U are computed once.

5.6.7 Independent functionals

The warning above said that k equations cut out codimension k only when they are independent. We now make that precise, in any dimension. Given $\varphi_1, \dots, \varphi_k \in V^*$, bundle them into one linear map

$$\Phi \colon V \to F^k, \qquad \Phi(v) = (\varphi_1(v), \dots, \varphi_k(v)),$$

which is linear because each entry is. Its kernel is $\ker \varphi_1 \cap \cdots \cap \ker \varphi_k$, the common solution set of the k equations.

Theorem 5.6.7. Independent Functionals

Let V be a vector space over F, of any dimension, and let $\varphi_1, \dots, \varphi_k \in V^*$. The following are equivalent:

(a) $(\varphi_1, \dots, \varphi_k)$ is linearly independent in V^*;
(b) $\Phi \colon V \to F^k$ is surjective;
(c) $\ker \varphi_1 \cap \cdots \cap \ker \varphi_k$ has codimension exactly k.

Idea. A dependence $\sum c_i \varphi_i = 0$ says that $\sum c_i y_i = 0$ for every $\mathbf{y} = \Phi(v)$ in the image: the image lies in the kernel of a nonzero functional on F^k. Conversely, if the image is a proper subspace of F^k, some nonzero functional on F^k kills it, and its coefficients give a dependence. The First Isomorphism Theorem turns "surjective" into "codimension k".

Proof.

(a) $\Rightarrow$ (b). Suppose Φ is not surjective. Then $\operatorname{im}\Phi$ is a proper subspace of F^k, so $\dim \operatorname{im}\Phi < k$ and, by Theorem 5.2.2, $(\operatorname{im}\Phi)^0 \neq \{0\}$. Take a nonzero functional in it; as at the start of this section it is $\varphi_{\mathbf{c}}$, $\mathbf{y} \mapsto c_1 y_1 + \cdots + c_k y_k$, for a nonzero $\mathbf{c} \in F^k$. For every $v \in V$, $\Phi(v) \in \operatorname{im}\Phi$, so

$$(c_1\varphi_1 + \cdots + c_k\varphi_k)(v) = \varphi_{\mathbf{c}}(\Phi(v)) = 0.$$

Thus $c_1\varphi_1 + \cdots + c_k\varphi_k = 0$ with $\mathbf{c} \neq \mathbf{0}$, and the list is dependent. This proves the contrapositive.

(b) $\Rightarrow$ (a). Suppose Φ is surjective and $c_1\varphi_1 + \cdots + c_k\varphi_k = 0$. For each j, choose v_j with $\Phi(v_j) = \mathbf{e}_j$, that is, $\varphi_i(v_j) = \delta_{ij}$. Evaluating the relation at v_j gives $c_j = 0$. So the list is independent.

(c) $\Leftrightarrow$ (c). By Theorem 4.4.7, $V/\ker\Phi \cong \operatorname{im}\Phi$, and $\ker\Phi = \bigcap_i \ker\varphi_i$. So the codimension of $\bigcap_i \ker\varphi_i$ is $\dim\operatorname{im}\Phi$, which equals k exactly when $\operatorname{im}\Phi = F^k$ (Theorem 1.6.8).

■

For $V = F^n$ and $\varphi_i = \varphi_{\mathbf{a}_i}$, the map Φ is $T_{\mathbf{A}}$, where $\mathbf{A}$ has rows $\mathbf{a}_i^\top$. The theorem then says: the rows of $\mathbf{A}$ are independent if and only if $\mathbf{Ax} = \mathbf{b}$ is solvable for **every** $\mathbf{b} \in F^k$. Independent equations never contradict each other.

Quick check. In $\mathbb{R}^3$, consider $\varphi_1 = x + y$, $\varphi_2 = y + z$, $\varphi_3 = x - z$. What is the codimension of $\ker\varphi_1 \cap \ker\varphi_2 \cap \ker\varphi_3$? Is the system $x + y = 1$, $y + z = 1$, $x - z = 1$ solvable?

Solution. $\varphi_3 = \varphi_1 - \varphi_2$, so the list is dependent, while (φ_1, φ_2) is independent (look at the coefficients of x and z). The three kernels intersect in $\ker\varphi_1 \cap \ker\varphi_2$, of codimension 2 by Theorem 5.6.7, so the intersection is a line, not $\{\mathbf{0}\}$. The system is not solvable: any solution would give $1 = x - z = (x + y) - (y + z) = 1 - 1 = 0$. This is (b) failing: $(1, 1, 1) \notin \operatorname{im}\Phi$.

5.6.8 Summary and transfer

The chapter has built a dictionary. On the left is a finite-dimensional space V of dimension n, or a linear map $T\colon V \to W$ between finite-dimensional spaces; on the right is its mirror image in the dual. Every entry has been proved in this chapter, and every arrow reverses inclusions or directions.

In V (vectors)	In V^* (measurements)
a subspace U, $\dim U = k$	its annihilator U^0, $\dim U^0 = n - k$
$U \subseteq W$	$W^0 \subseteq U^0$
$U + W$	$U^0 \cap W^0$
$U \cap W$	$U^0 + W^0$
U itself, recovered as U^{00}	U^0 determines U
the quotient V/U	the subspace $U^0 \cong (V/U)^*$
the subspace U	the quotient $V^*/U^0 \cong U^*$
a hyperplane H	a nonzero functional, up to scale
a spanning list (parametric form)	a list of equations (implicit form)
dimension (free parameters)	codimension (independent equations)
a basis $(v_1, \ldots, v_n)$	its dual basis $(\varphi_1, \ldots, \varphi_n)$
$T\colon V \to W$	$T'\colon W^* \to V^*$, backwards
T injective	T' surjective
T surjective	T' injective
$\ker T$	$\operatorname{im} T' = (\ker T)^0$

In V (vectors)	In V^* (measurements)
$\operatorname{im} T$	$\ker T' = (\operatorname{im} T)^0$
$\operatorname{rank} T$	$\operatorname{rank} T' = \operatorname{rank} T$
the matrix $\mathbf{A} = [T]_{\mathcal{B}}^{\mathcal{C}}$	the transpose $\mathbf{A}^\top = [T']_{\mathcal{C}^*}^{\mathcal{B}^*}$
$\operatorname{nul}(\mathbf{A})$	$\operatorname{row}(\mathbf{A})$, as $\operatorname{nul}(\mathbf{A})^0$
V	V^{**}, naturally isomorphic via ev_V

Three moves make the dictionary usable, and they transfer to any situation with a pairing between two spaces.

- **Test against functionals.** To show a vector is zero, or lies in a subspace, check that enough functionals vanish on it (Theorem 5.1.5, Lemma 5.6.1).
- **Pass to the dual and count.** A question about a subspace can be translated into one about its annihilator, where $\dim U + \dim U^0 = n$ often settles it. The third proof of row rank = column rank (Corollary 5.4.4) and the count in Theorem 5.6.5 both work this way.
- **Prefer the natural map.** Isomorphisms such as $V \cong V^{**}$, $(V/U)^* \cong U^0$, and $U^* \cong V^*/U^0$ are defined without a basis, so they can be used in any coordinates and survive changes of basis. Isomorphisms such as $V \cong V^*$ depend on a choice, and a proof that uses one must say which.

In Chapter 11, an inner product will give a **chosen** isomorphism $V \cong V^*$, and the annihilator U^0 will reappear inside V itself as the orthogonal complement. The dictionary above is the part of that story that needs no inner product.

5.6.9 Exercises

5.6.9.1 A. Check your understanding

Exercise. A1

(a) Explain how a row vector $\mathbf{a}^\top$ (with $\mathbf{a} \in F^n$) defines a linear functional on F^n, and why every functional arises this way.

(b) What is the least number of homogeneous linear equations whose solution set is a given 3-dimensional subspace of F^7?

(c) True or false: every system of three homogeneous equations in $\mathbb{R}^5$ has a solution set of dimension 2. Justify your answer.

(d) State the relation between $\operatorname{row}(\mathbf{A})$ and $\operatorname{nul}(\mathbf{A})$ in terms of annihilators.

(e) True or false: if $\mathbf{A}$ and $\mathbf{B}$ have n columns and $\operatorname{nul}(\mathbf{A}) = \operatorname{nul}(\mathbf{B})$, then $\mathbf{A}$ and $\mathbf{B}$ have the same number of rows. Justify your answer.

(f) Name the dual counterparts of: a sum of subspaces; an injective linear map; the matrix of a linear map.

Solution.

(a) $\varphi_{\mathbf{a}}(\mathbf{x}) = \mathbf{a}^\top\mathbf{x}$ is linear by the rules of matrix multiplication (Theorem 0.9.4). A functional φ equals $\varphi_{\mathbf{a}}$ with $a_j = \varphi(\mathbf{e}_j)$, because both agree on the standard basis (Theorem 2.1.5); this is Theorem 5.1.1.

(b) $7 - 3 = 4$, by Theorem 5.6.2.

(c) False. The equations may be dependent: $x_1 = 0$ written three times has a solution set of dimension 4. The dimension is $5 - \operatorname{rank} \mathbf{A}$, which is 2 only when the three rows are independent.

(d) Identifying rows with functionals, $\operatorname{row}(\mathbf{A}) = \operatorname{nul}(\mathbf{A})^0$ (Theorem 5.6.5).

(e) False. By Corollary 5.6.6 they have the same row space, but not necessarily the same number of rows: $\mathbf{A} = \begin{pmatrix} 1 & 0 \end{pmatrix}$ and $\mathbf{B} = \begin{pmatrix} 1 & 0 \\ 2 & 0 \end{pmatrix}$ both have null space $\operatorname{span}(\mathbf{e}_2)$.

(f) An intersection of annihilators, $(U + W)^0 = U^0 \cap W^0$; a surjective dual map; the transpose matrix.

5.6.9.2 B. Practice

Exercise. B1: Equations for a span

Let $U = \operatorname{span}((1, 2, 0, -1), (2, 1, 1, 0), (3, 3, 1, -1)) \subseteq \mathbb{R}^4$.

(a) Find $\dim U$.

(b) Find a system of homogeneous linear equations with solution set U, using as few equations as possible.

(c) Hence write the affine plane $(1, 0, 0, 0) + U$ as the solution set of a linear system.

Solution.

(a) The third vector is the sum of the first two, so $U = \operatorname{span}((1, 2, 0, -1), (2, 1, 1, 0))$ by Theorem 1.3.5. These two vectors are independent, since neither is a multiple of the other, so they form a basis of U and $\dim U = 2$.

(b) By Theorem 5.6.3, we need a basis of $\operatorname{nul}(\mathbf{M})$, where $\mathbf{M}$ has the three vectors as rows. Row reduction gives

$$\mathbf{M} = \begin{pmatrix} 1 & 2 & 0 & -1 \\ 2 & 1 & 1 & 0 \\ 3 & 3 & 1 & -1 \end{pmatrix} \longrightarrow \begin{pmatrix} 1 & 0 & \frac{2}{3} & \frac{1}{3} \\ 0 & 1 & -\frac{1}{3} & -\frac{2}{3} \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

The free variables are the third and fourth. Setting them to $(3, 0)$ and $(0, 3)$ to clear denominators gives the null vectors $(-2, 1, 3, 0)$ and $(-1, 2, 0, 3)$. So

$$U = \{\mathbf{x} \in \mathbb{R}^4 : -2x_1 + x_2 + 3x_3 = 0, \quad -x_1 + 2x_2 + 3x_4 = 0\}.$$

Check on the first two spanning vectors, which suffice since they span U: $-2+2+0 = 0$, $-1 + 4 - 3 = 0$; $-4 + 1 + 3 = 0$, $-2 + 2 + 0 = 0$. Two equations is the minimum, by Theorem 5.6.2 (b).

(c) By Proposition 5.6.4, replace each right-hand side by the value of the left-hand side at $(1, 0, 0, 0)$:

$$(1, 0, 0, 0) + U = \{\mathbf{x} : -2x_1 + x_2 + 3x_3 = -2, \quad -x_1 + 2x_2 + 3x_4 = -1\}.$$

Exercise. B2: A spanning list for a system

Let $U = \{\mathbf{x} \in \mathbb{R}^4 : x_1 + x_2 - x_3 + 2x_4 = 0, \quad 2x_1 + 2x_2 + x_3 + x_4 = 0\}$.

(a) Find a basis of U.

(b) Run the parametric-to-implicit algorithm on your basis, and check that the resulting equations have the same solution set as the given ones, by comparing row spaces.

Solution.

(a) Row reduce the coefficient matrix: subtract twice row 1 from row 2 to get $(0, 0, 3, -3)$, divide by 3, then add row 2 to row 1:

$$\begin{pmatrix} 1 & 1 & -1 & 2 \\ 2 & 2 & 1 & 1 \end{pmatrix} \longrightarrow \begin{pmatrix} 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & -1 \end{pmatrix}.$$

The pivots are in columns 1 and 3, and x_2, x_4 are free. By Theorem 3.5.4, a basis is $\mathbf{n}_1 = (-1, 1, 0, 0)$ and $\mathbf{n}_2 = (-1, 0, 1, 1)$.

(b) Put $\mathbf{n}_1^\top, \mathbf{n}_2^\top$ as the rows of $\mathbf{M}$ and solve $\mathbf{Ma} = \mathbf{0}$: $-a_1 + a_2 = 0$ and $-a_1 + a_3 + a_4 = 0$. So $a_2 = a_1$ and $a_3 = a_1 - a_4$, with a_1, a_4 free; the basic solutions are $(1, 1, 1, 0)$ and $(0, 0, -1, 1)$. The new system is $x_1 + x_2 + x_3 = 0$, $-x_3 + x_4 = 0$. Its rows lie in the row space of the original matrix, which equals the row space of the RREF (Theorem 3.5.1), since $(1, 1, 1, 0) = (1, 1, 0, 1) + (0, 0, 1, -1)$ and $(0, 0, -1, 1) = -(0, 0, 1, -1)$. Both row spaces are two-dimensional, so they are equal (Theorem 1.6.8), and the two systems have the same solution set by Corollary 5.6.6.

Exercise. B3: Two planes in $\mathbb{R}^4$

Let $U = \operatorname{span}((1, 0, 0, 1), (0, 1, 1, 0))$ and $W = \operatorname{span}((1, 1, 0, 0), (0, 0, 1, 1))$ in $\mathbb{R}^4$.

(a) Find two equations for U and two for W.

(b) Use them to find $U \cap W$.

(c) Find a single equation for $U + W$, using $(U + W)^0 = U^0 \cap W^0$.

Solution.

(a) Each pair is independent, so each plane has dimension 2 and needs two equations. The functionals $x_1 - x_4$ and $x_2 - x_3$ vanish at $(1, 0, 0, 1)$ and $(0, 1, 1, 0)$, and they are independent; since $\dim U^0 = 2$ (Theorem 5.2.2), they form a basis of U^0. So $U = \{x_1 = x_4, \ x_2 = x_3\}$ by Theorem 5.6.2. In the same way $x_1 - x_2$ and $x_3 - x_4$ form a basis of W^0, and $W = \{x_1 = x_2, \ x_3 = x_4\}$.

(b) $U \cap W$ is the solution set of all four equations: $x_1 = x_4$, $x_2 = x_3$, $x_1 = x_2$, $x_3 = x_4$. They force $x_1 = x_2 = x_3 = x_4$, so $U \cap W = \operatorname{span}((1, 1, 1, 1))$.

(c) A functional in $U^0 \cap W^0$ has the form $a(x_1 - x_4) + b(x_2 - x_3) = c(x_1 - x_2) + d(x_3 - x_4)$. Comparing coefficients of x_1, x_2, x_3, x_4: $a = c$, $b = -c$, $-b = d$, $-a = -d$. So $a = c = d$ and $b = -a$, and $U^0 \cap W^0 = \operatorname{span}(x_1 - x_2 + x_3 - x_4)$. By Theorem 5.2.3 (c) and Lemma 5.6.1,

$$U + W = \{\mathbf{x} : x_1 - x_2 + x_3 - x_4 = 0\},$$

a hyperplane. Consistently, $\dim(U+W) = 2+2-1 = 3$ by Theorem 1.7.3.

5.6.9.3 C. Going deeper

Exercise. C1: Avoiding finitely many subspaces

Let F be an **infinite** field, V a vector space over F with $1 \le \dim V = n < \infty$, and $U_1, \dots, U_k$ proper subspaces of V.

(a) Prove that for each i there is a nonzero $\varphi_i \in V^*$ with $U_i \subseteq \ker \varphi_i$.

(b) Let $(b_1, \dots, b_n)$ be a basis of V, and for $t \in F$ put $v(t) = b_1 + tb_2 + t^2 b_3 + \cdots + t^{n-1} b_n$. Show that $t \mapsto \varphi_i(v(t))$ is given by a nonzero polynomial of degree at most $n-1$.

(c) Deduce that some vector of V lies in none of $U_1, \dots, U_k$. In other words, V is not a union of finitely many proper subspaces.

(d) Show that the hypothesis that F is infinite cannot be dropped: over $\mathbb{F}_q$, the space $\mathbb{F}_q^2$ is the union of $q+1$ proper subspaces.

Hint: for (c), use Lemma 3.7.2.

Solution.

(a) Since $U_i \neq V$, pick $v \notin U_i$. By Lemma 5.6.1 there is $\varphi_i \in U_i^0$ with $\varphi_i(v) \neq 0$. Then $\varphi_i \neq 0$ and $U_i \subseteq \ker \varphi_i$.

(b) By linearity, $\varphi_i(v(t)) = \sum_{j=1}^{n} \varphi_i(b_j)\, t^{j-1}$, which is the value at t of the polynomial $p_i(x) = \sum_{j=1}^{n} \varphi_i(b_j)\, x^{j-1}$ of degree at most $n-1$. Its coefficients are $\varphi_i(b_1), \dots, \varphi_i(b_n)$, which are not all zero, since otherwise φ_i would vanish on a basis and be zero (Theorem 2.1.5). So $p_i \neq 0$.

(c) By Lemma 3.7.2, each p_i has at most $n-1$ roots in F, so together they have at most $k(n-1)$ roots. Since F is infinite, there is $t_0 \in F$ that is a root of none of them. Then $\varphi_i(v(t_0)) = p_i(t_0) \neq 0$, so $v(t_0) \notin \ker \varphi_i \supseteq U_i$, for every i.

(d) $\mathbb{F}_q^2$ has $q^2 - 1$ nonzero vectors, and each line through $\mathbf{0}$ contains $q-1$ of them (the nonzero multiples of one vector). Two distinct lines meet only in $\mathbf{0}$, so there are $(q^2-1)/(q-1) = q+1$ lines, and every nonzero vector lies on one of them. Hence $\mathbb{F}_q^2$ is the union of its $q+1$ lines, each a proper subspace. For $q = 2$ these are $\operatorname{span}((1,0))$, $\operatorname{span}((0,1))$, and $\operatorname{span}((1,1))$, as in Exercise (C1: Covering a Space by Subspaces). The proof of (c) breaks exactly where it used infinitely many t: here $k(n-1) = q+1$ exceeds $|F| = q$.

Exercise. C2: The least number of equations for a system

Let $\mathbf{A} \in M_{m\times n}(F)$ and $\mathbf{b} \in F^m$, and suppose the solution set S of $\mathbf{Ax} = \mathbf{b}$ is nonempty.

(a) Prove that if $\mathbf{B} \in M_{p\times n}(F)$, $\mathbf{c} \in F^p$ and $\mathbf{Bx} = \mathbf{c}$ has the same solution set S, then $\operatorname{nul}(\mathbf{B}) = \operatorname{nul}(\mathbf{A})$.

(b) Deduce that every system with solution set S has at least $\operatorname{rank} \mathbf{A}$ equations, and that some system with solution set S has exactly $\operatorname{rank} \mathbf{A}$ equations.

(c) Show by an example that (b) fails without the hypothesis $S \neq \varnothing$.

Solution.

(a) Fix $\mathbf{p} \in S$. By Theorem 3.1.3, $S = \mathbf{p}+\text{nul}(\mathbf{A})$ and also $S = \mathbf{p}+\text{nul}(\mathbf{B})$. For $\mathbf{h} \in \text{nul}(\mathbf{A})$, $\mathbf{p} + \mathbf{h} \in S = \mathbf{p} + \text{nul}(\mathbf{B})$, so $\mathbf{h} \in \text{nul}(\mathbf{B})$. Swapping the roles of $\mathbf{A}$ and $\mathbf{B}$ gives the reverse inclusion, so $\text{nul}(\mathbf{A}) = \text{nul}(\mathbf{B})$.

(b) Let $k = \dim \text{nul}(\mathbf{A}) = n - \text{rank}\,\mathbf{A}$ (Theorem 3.5.7). If $\mathbf{Bx} = \mathbf{c}$ with p equations has solution set S, then $\text{nul}(\mathbf{B}) = \text{nul}(\mathbf{A})$ by (a), so $p \geq n - k = \text{rank}\,\mathbf{A}$ by Theorem 5.6.2 (b). Conversely, take $\mathbf{A}_0 \in M_{(n-k)\times n}(F)$ with $\text{nul}(\mathbf{A}_0) = \text{nul}(\mathbf{A})$, which exists by Theorem 5.6.2 (a) (if $k = n$, take the empty system, whose solution set is $F^n = S$). By Proposition 5.6.4, or directly by Theorem 3.1.3, the system $\mathbf{A}_0\mathbf{x} = \mathbf{A}_0\mathbf{p}$ has solution set $\mathbf{p} + \text{nul}(\mathbf{A}) = S$, and it has $n - k = \text{rank}\,\mathbf{A}$ equations.

(c) Over any field, $0x_1 = 1$ (in F^1) has empty solution set and coefficient matrix of rank 0, yet no system with 0 equations has empty solution set, since the empty system is solved by every vector. So at least one equation is needed, which is more than $\text{rank}\,\mathbf{A} = 0$.

Chapter 6

Polynomials

Integers and polynomials share one arithmetic. Both can be divided with a remainder, and from that single tool everything else follows: greatest common divisors, Bézout identities, unique factorization into primes or irreducibles, and the Chinese Remainder Theorem. This chapter develops that arithmetic for $F[x]$ in full, because from Chapter 9 on the structure of an operator is read off from polynomials: which polynomials kill it, how they factor, and which of their factors are coprime.

Along the way we keep two promises made in Chapter 0: we prove division with remainder, stated there only for orientation, and the Fundamental Theorem of Algebra, that every nonconstant complex polynomial has a root. We also prove that polynomials factor uniquely into irreducibles, just as integers factor uniquely into primes. The chapter ends by plugging an operator into a polynomial, where the arithmetic of $F[x]$ turns into a splitting of the space into pieces.

What you need. From Chapter 0, induction and well-ordering (Theorem 0.2.10, Theorem 0.2.11), divisibility of integers and Bézout's lemma (Definition 0.2.2, Lemma 0.7.2), the fields $\mathbb{F}_p$ (Theorem 0.7.3), polynomials, degree and evaluation (Definition 0.8.1, Definition 0.8.3, Theorem 0.8.3, Theorem 0.8.5), and complex numbers (Theorem 0.6.6). From Chapter 1, the standard bases of the polynomial spaces and their dimensions (Example (Standard Bases), Example (Dimensions of the Running Examples)), counting instead of checking (Theorem 1.6.4, Theorem 1.6.3), and direct sums (Definition 1.7.2, Theorem 1.7.4), which is what the kernel splitting lemma of §07 produces. From Chapter 2, polynomials of an operator (Definition 2.4.2). From Chapter 3, the factor theorem and root bound for linear factors (Lemma 3.7.1, Lemma 3.7.2) and interpolation (Theorem 3.7.3). From Chapter 4, quotient spaces and the First Isomorphism Theorem (Definition 4.4.3, Theorem 4.4.7). From Chapter 5, dual bases (Theorem 5.1.3).

6.0.1 Roadmap

- **Division with remainder.** Long division always works over a field, with a unique quotient and remainder; the remainder and factor theorems in general.
- **Ideals and greatest common divisors.** Every ideal of $F[x]$ is generated by one polynomial; gcds, Bézout identities, and the Euclidean algorithm.
- **Irreducible polynomials and unique factorization.** The primes of $F[x]$, and why factorization into them is unique.
- **Roots, multiplicity, and derivatives.** Repeated roots detected by the formal derivative, and Taylor expansion of a polynomial.
- **The Fundamental Theorem of Algebra.** A complete proof, and what it says about real polynomials.
- **Interpolation and the Chinese Remainder Theorem.** The Lagrange formula, solving

simultaneous congruences, and the quotient-space picture behind both.
- **Polynomials of an operator.** Every operator on a finite-dimensional space satisfies a polynomial, and coprime factors split the space.

6.0.2 Named moves

- **Divide and look at the remainder.** Most statements about divisibility reduce to one division.
- **The smallest-degree element generates.** In an ideal, a nonzero element of least degree divides every other element.
- **Evaluate at a root.** Plugging a root of the divisor into $f = qg + r$ kills qg and reads off information about the remainder.
- **Compare degrees.** Uniqueness of quotients, factorizations, and interpolants usually comes from a degree count.
- **Bézout gives projections.** When $\gcd(p, q) = 1$, write $1 = ap + bq$; evaluated at an operator, the two terms split the space.

6.1 Division with Remainder

Integers can be divided with a remainder: $17 = 3 \cdot 5 + 2$, and the remainder 2 is smaller than the divisor 5. Polynomials over a field can be divided in the same way, with "smaller" measured by degree. Chapter 0 stated this fact without proof, and Chapter 3 proved the special case of dividing by $x - c$. This section proves division with remainder in general, over any field, and draws the first consequences: the language of divisibility, the Remainder and Factor Theorems for every divisor $x - c$, a sharper root bound, and the fact that over an infinite field a polynomial is determined by its values. Throughout, F is a field unless we say otherwise, and it matters: the warning in the second subsection shows division failing when the coefficients only form $\mathbb{Z}$.

6.1.1 Long division by hand

At school one divides $x^4 + 2x^3 - x + 3$ by $x^2 + x - 1$ with a layout that imitates long division of numbers. Let us do it in words, so that we can see what each step does.

Write $f = x^4 + 2x^3 - x + 3$ and $g = x^2 + x - 1$, in $\mathbb{Q}[x]$.

1. The leading term of f is x^4, and the leading term of g is x^2. The quotient $x^4/x^2 = x^2$ is the first term of the answer. Subtract $x^2 g = x^4 + x^3 - x^2$:
$$f - x^2 g = x^3 + x^2 - x + 3.$$
2. The leading term is now x^3, and $x^3/x^2 = x$. Subtract $xg = x^3 + x^2 - x$:
$$f - x^2 g - xg = 3.$$
3. What is left, 3, has degree 0, smaller than $\deg g = 2$. No multiple $c\,x^k g$ with $k \geq 0$ can remove it without creating terms of degree at least 2. We stop.

So $f = (x^2 + x)\, g + 3$: the **quotient** is $x^2 + x$ and the **remainder** is 3. Expanding $(x^2 + x)(x^2 + x - 1) = x^4 + 2x^3 - x$ confirms it.

Two questions are hidden in this routine. **Why does it always stop?** Each step removes the current leading term, so the degree of what is left drops by at least one. Degrees are natural numbers, and a strictly decreasing sequence of natural numbers cannot go on forever. **Why**

is the answer unique? Someone else might hope to find a different quotient and a different remainder of degree less than 2. We will see that they cannot, by comparing degrees.

There is also a third, quieter question. In step 1 we divided the leading coefficient of the current polynomial by the leading coefficient of g. Here that coefficient was 1. In general it is some $b \neq 0$, and we need b^{-1}. Over a field this always exists. That is exactly where the field enters.

6.1.2 The Division Theorem

The routine above works for every pair of polynomials with nonzero divisor. Here is the precise statement.

Theorem 6.1.1. Polynomial Division with Remainder

Let F be a field, and let $f, g \in F[x]$ with $g \neq 0$. Then there exist **unique** polynomials $q, r \in F[x]$ such that

$$f = qg + r \qquad \text{and} \qquad \deg r < \deg g.$$

The polynomial q is the **quotient** and r is the **remainder** of f on division by g. Because $\deg 0 = -\infty$, the condition $\deg r < \deg g$ allows $r = 0$, which is the case "g goes into f exactly".

Idea. Existence is the hand routine turned into an induction on $\deg f$. If $\deg f < \deg g$, there is nothing to do: $q = 0$, $r = f$. Otherwise subtract $(a/b)\,x^{n-m}g$, where ax^n and bx^m are the leading terms of f and g. This kills the leading term of f, so the induction hypothesis applies to what is left. Uniqueness is "take the difference": if $qg+r = q'g+r'$, then $(q-q')g = r'-r$. The right side has degree less than $\deg g$, while a nonzero multiple of g has degree at least $\deg g$.

Proof. Let $m = \deg g$, which lies in $\mathbb{N}$ since $g \neq 0$, and let $b \neq 0$ be the leading coefficient of g.

Existence. If $f = 0$, then $q = 0$ and $r = 0$ work. For nonzero f we prove, by strong induction on $n \in \mathbb{N}$ (Theorem 0.2.10 with $n_0 = 0$), the statement

> $P(n)$: for every $f \in F[x]$ with $\deg f = n$, there are $q, r \in F[x]$ with $f = qg + r$ and $\deg r < m$.

Let $n \in \mathbb{N}$, assume $P(k)$ for all $k < n$, and let $\deg f = n$ with leading coefficient a.

Case 1: $n < m$. Take $q = 0$ and $r = f$. Then $f = 0 \cdot g + f$ and $\deg r = n < m$.

Case 2: $n \geq m$. Since F is a field and $b \neq 0$, the inverse b^{-1} exists. Put

$$f_1 = f - ab^{-1}x^{n-m}g.$$

By Theorem 0.8.3, the polynomial $ab^{-1}x^{n-m}g$ has degree $(n-m)+m = n$ and leading coefficient $ab^{-1}\cdot b = a$. So the coefficient of x^n in f_1 is $a-a = 0$, and all coefficients of higher powers are 0 as well. Hence $f_1 = 0$ or $\deg f_1 < n$. If $f_1 = 0$, then $f = ab^{-1}x^{n-m}g + 0$, and we take $q = ab^{-1}x^{n-m}$, $r = 0$. Otherwise $k = \deg f_1 < n$, and the induction hypothesis $P(k)$ gives q_1, r_1 with $f_1 = q_1 g + r_1$ and $\deg r_1 < m$. Then

$$f = f_1 + ab^{-1}x^{n-m}g = (ab^{-1}x^{n-m} + q_1)g + r_1,$$

so $q = ab^{-1}x^{n-m} + q_1$ and $r = r_1$ work. This proves $P(n)$, and by strong induction $P(n)$ holds for every $n \in \mathbb{N}$.

Uniqueness. Suppose $f = qg + r = q'g + r'$ with $\deg r < m$ and $\deg r' < m$. Then $(q - q')g = r' - r$. By Theorem 0.8.2, $\deg(r' - r) \leq \max(\deg r', \deg r) < m$, since $-r$ has the same degree as r. Suppose $q \neq q'$. Then $\deg(q - q') \geq 0$, and Theorem 0.8.3 gives $\deg((q-q')g) = \deg(q-q') + m \geq m$, a contradiction. Hence $q = q'$. Then $r' - r = 0 \cdot g = 0$, so $r = r'$. This proves the theorem.

■

The field axioms did essential work in two places: the inverse b^{-1} in Case 2, and Theorem 0.8.3 in the uniqueness step, which needs a product of nonzero coefficients to be nonzero. The proof is also an algorithm. Case 2, repeated, is the long division of the previous subsection, and the induction guarantees that it stops after at most $\deg f - \deg g + 1$ subtractions.

The hypothesis "over a field" is not decoration. **The leading coefficient of g must be invertible.**

Warning. **Division with remainder fails over $\mathbb{Z}$.** Write $\mathbb{Z}[x]$ for the polynomials in $\mathbb{Q}[x]$ whose coefficients are all integers. Try to divide $f = x^2$ by $g = 2x$ inside $\mathbb{Z}[x]$: we want $q, r \in \mathbb{Z}[x]$ with $x^2 = 2x \cdot q + r$ and $\deg r < 1$, so r is a constant. Comparing coefficients of x^2 gives $2q_1 = 1$, where q_1 is the coefficient of x in q. No integer q_1 satisfies this, so no such q, r exist. Over $\mathbb{Q}$ there is no problem: $x^2 = 2x \cdot \frac{1}{2}x + 0$. The obstacle is precisely that 2 has no inverse in $\mathbb{Z}$.

Division by a divisor whose leading coefficient **is** invertible still works over $\mathbb{Z}$, for example by a monic polynomial; see Exercise C1.

Quick check. Why must the inequality $\deg r < \deg g$ be strict? Find two different pairs (q, r) with $x^2 = q \cdot (x^2 + 1) + r$ and $\deg r \leq 2$.

Solution. Both $x^2 = 1 \cdot (x^2 + 1) + (-1)$ and $x^2 = 0 \cdot (x^2 + 1) + x^2$ work, with remainders of degree 0 and 2. With $\deg r \leq \deg g$ allowed, uniqueness fails, because we may move one copy of g between qg and r. The strict inequality is exactly what rules this out: in the uniqueness proof, $\deg(r' - r) < \deg g$ is the step that forces $q = q'$.

Now a division in which the leading coefficient of the divisor is not 1, so fractions appear.

Example. Long Division with a Non-Monic Divisor

In $\mathbb{Q}[x]$, divide $f = 2x^4 - 3x^3 + x^2 + 4x - 1$ by $g = 2x^2 - x + 3$ with remainder.

Solution. We follow Case 2 of the proof, where $b = 2$ and $b^{-1} = \frac{1}{2}$.

Step 1. The leading term of f is $2x^4$, and $2x^4/2x^2 = x^2$. Subtracting $x^2 g = 2x^4 - x^3 + 3x^2$

leaves

$$f - x^2 g = -2x^3 - 2x^2 + 4x - 1.$$

Step 2. The leading term is $-2x^3$, and $-2x^3/2x^2 = -x$. Subtracting $-xg = -2x^3 + x^2 - 3x$ leaves

$$-3x^2 + 7x - 1.$$

Step 3. The leading term is $-3x^2$, and $-3x^2/2x^2 = -\frac{3}{2}$. Subtracting $-\frac{3}{2}g = -3x^2 + \frac{3}{2}x - \frac{9}{2}$ leaves

$$(7 - \tfrac{3}{2})x + (-1 + \tfrac{9}{2}) = \tfrac{11}{2}x + \tfrac{7}{2}.$$

This has degree $1 < 2 = \deg g$, so we stop. Hence

$$2x^4 - 3x^3 + x^2 + 4x - 1 = (x^2 - x - \tfrac{3}{2})(2x^2 - x + 3) + \tfrac{11}{2}x + \tfrac{7}{2},$$

with quotient $q = x^2 - x - \frac{3}{2}$ and remainder $r = \frac{11}{2}x + \frac{7}{2}$. By the uniqueness in Theorem 6.1.1, any other correct method must produce the same q and r.

Notice that the quotient has a noninteger coefficient although f and g have integer coefficients. This is the warning above in action: over $\mathbb{Z}$ this division would be impossible.

6.1.3 Divisibility

With integers, "3 divides 12" means that 12 is an integer multiple of 3 (Definition 0.2.2), and it is the case "remainder zero" of integer division. Polynomials get the same word. The field is part of the meaning, so we put it in the definition.

g divides f when f is a polynomial multiple of g, with the multiplier taken over the same field.

Definition 6.1.1. Divisibility, Associates

Let F be a field and $f, g \in F[x]$.

- We say g **divides** f **over** F, and write $g \mid f$, if there **exists** $h \in F[x]$ with $f = gh$. Then g is a **divisor** (or **factor**) of f, and f is a **multiple** of g. We write $g \nmid f$ if g does not divide f.
- We say f and g are **associates** if $f = cg$ for some **nonzero** constant $c \in F$.

In words: $g \mid f$ asks for a multiplier h **in** $F[x]$, so the same pair can behave differently over different fields. Associates differ only by a nonzero scalar factor. Since c^{-1} exists, "associates" is symmetric: $f = cg$ gives $g = c^{-1} f$.

Examples, simplest first.

- **Degenerate cases.** Every g divides 0, since $0 = g \cdot 0$. On the other hand $0 \mid f$ only for $f = 0$, since $0 \cdot h = 0$. Every nonzero constant c divides every f, since $f = c \cdot (c^{-1} f)$. These cases matter: they say that nonzero constants are invisible to divisibility, just as ± 1 are for integers.
- In $\mathbb{Q}[x]$, $x - 1 \mid x^2 - 1$, since $x^2 - 1 = (x - 1)(x + 1)$. Also $2x - 2 \mid x^2 - 1$, since $x^2 - 1 = (2x - 2) \cdot \frac{1}{2}(x + 1)$. The polynomials $x - 1$ and $2x - 2$ are associates.
- In $\mathbb{F}_2[x]$, $x + 1 \mid x^2 + 1$, since $(x + 1)^2 = x^2 + 2x + 1 = x^2 + 1$ there. In $\mathbb{Q}[x]$, however, $x + 1 \nmid x^2 + 1$, as the next paragraph shows.
- In $\mathbb{C}[x]$, $x - i \mid x^2 + 1 = (x - i)(x + i)$.

For a nonexample by minimal change, keep x^2+1 and $x+1$ but move from $\mathbb{F}_2$ to $\mathbb{Q}$. Both are still polynomials, and division with remainder still applies. What fails is the clause "there **exists** $h \in F[x]$". If $x^2+1=(x+1)h$ with $h \in \mathbb{Q}[x]$, then evaluating at -1 (Theorem 0.8.5) gives $2 = 0 \cdot h(-1) = 0$, which is false in $\mathbb{Q}$. Over $\mathbb{F}_2$ the same computation reads $0=0$, and there is no contradiction.

Warning. Divisibility depends on where the multiplier may live. In $\mathbb{Q}[x]$, $2x+2 \mid x+1$, because $x+1=(2x+2)\cdot\frac{1}{2}$. If we insisted on integer coefficients, $2x+2$ would **not** divide $x+1$: any integer multiple $(2x+2)h$ has all coefficients even. Over a field, a nonzero scalar factor never affects divisibility, so it is pointless to ask whether $x-1$ or $3x-3$ is "the" factor of x^2-1.

Why this definition? Because it makes "divides" the case of remainder zero, which is the link to division we need. The basic rules follow quickly.

Proposition 6.1.2. Rules for Divisibility

Let F be a field and $f,g,h \in F[x]$.

(a) If $f \mid g$ and $g \mid h$, then $f \mid h$.

(b) If $h \mid f$ and $h \mid g$, then $h \mid af+bg$ **for all** $a,b \in F[x]$.

(c) If $g \mid f$ and $f \neq 0$, then $\deg g \leq \deg f$.

(d) $f \mid g$ and $g \mid f$ if and only if f and g are associates or both are 0. In particular, if $f \mid g$, $g \mid f$, and f,g are both monic, then $f=g$.

(e) If $g \neq 0$, then $g \mid f$ if and only if the remainder of f on division by g is 0.

Proof.

(a) Write $g=fu$ and $h=gv$ with $u,v \in F[x]$. Then $h=f(uv)$ by Theorem 0.8.1, so $f \mid h$.

(b) Write $f=hu$ and $g=hv$. Then $af+bg=h(au+bv)$ by Theorem 0.8.1, so $h \mid af+bg$.

(c) Write $f=gu$. Since $f \neq 0$, both g and u are nonzero, so $\deg u \geq 0$. By Theorem 0.8.3, $\deg f = \deg g + \deg u \geq \deg g$.

(d) ($\Leftarrow$) If $f=cg$ with $c \neq 0$, then $g \mid f$ and $g=c^{-1}f$ gives $f \mid g$. If $f=g=0$, both divide each other. ($\Rightarrow$) Suppose $g=fu$ and $f=gv$. If $f=0$, then $g=0\cdot u=0$, so both are 0. If $f \neq 0$, then $f=f(uv)$, so $f(1-uv)=0$, and Corollary 0.8.4 gives $uv=1$. By Theorem 0.8.3, $\deg u+\deg v=\deg 1=0$ with both degrees in $\mathbb{N}$, so u and v are nonzero constants. Hence $f=vg$ with $v \neq 0$, and f,g are associates. If both are monic, comparing leading coefficients in $f=vg$ gives $1=v\cdot 1$, so $f=g$.

(e) If the remainder is 0, then $f=qg$, so $g \mid f$. Conversely, if $f=hg$, then $f=hg+0$ with $\deg 0=-\infty<\deg g$. This is a division with remainder, so by the uniqueness in Theorem 6.1.1, the remainder of f on division by g is 0. This proves the proposition.

■

Part (d) is the polynomial form of a move we will use again and again: **to show two monic polynomials are equal, show that each divides the other.** It also gives each nonzero polynomial a canonical representative. If $f \neq 0$ has leading coefficient a, then $a^{-1}f$ is monic

and an associate of f, and it is the **only** monic associate: two monic associates divide each other, so they are equal by (d). For example, the monic associate of $3x^2 - 6$ in $\mathbb{Q}[x]$ is $x^2 - 2$. The greatest common divisor in the next section will be defined as a monic polynomial for exactly this reason.

6.1.4 The Remainder and Factor Theorems

Division by a polynomial of degree 1 leaves a remainder of degree less than 1, that is, a constant. Which constant? Chapter 3 already answered this. The Factor Theorem for a Linear Divisor (Lemma 3.7.1) showed, by an induction that was long division in disguise, that $p = (x-c)q+p(c)$ for some q. We now read that lemma through Theorem 6.1.1. What is new is that q and the constant are **unique**, so $p(c)$ is **the** remainder in the sense of division.

Theorem 6.1.3. Remainder Theorem and Factor Theorem

Let F be a field, $f \in F[x]$, and $c \in F$.

(a) The remainder of f on division by $x - c$ is the constant $f(c)$.

(b) c is a root of f if and only if $x - c \mid f$.

Proof.

(a) By Lemma 3.7.1, $f = (x - c)q + f(c)$ for some $q \in F[x]$. The constant $f(c)$ has degree 0 or $-\infty$, which is less than $\deg(x - c) = 1$. So this is a division with remainder, and by the uniqueness in Theorem 6.1.1, $f(c)$ is the remainder.

(b) By Proposition 6.1.2 (e), $x - c \mid f$ if and only if the remainder is 0, which by (a) means $f(c) = 0$. This proves the theorem.

■

In other words, the Factor Theorem of Chapter 3 is really a statement about division. Division alone would also give it: dividing f by $x - c$ leaves a constant remainder r, and evaluating $f = (x-c)q+r$ at c gives $r = f(c)$. Since Chapter 3 already proved the lemma, we cite it rather than prove it again, and we will use the name "Remainder Theorem" for both parts.

Example. Remainders Without Dividing

(a) Find the remainder of $f = x^5 - 3x^3 + 2x + 7$ on division by $x - 2$ in $\mathbb{Q}[x]$. (b) Show that $x + 1 \nmid x^2 + 1$ in $\mathbb{Q}[x]$, but $x + 1 \mid x^2 + 1$ in $\mathbb{F}_2[x]$. (c) Find the remainder of x^5 on division by $x^2 - x$ in $\mathbb{Q}[x]$, again without long division.

Solution.

(a) By Theorem 6.1.3 (a), the remainder is $f(2) = 32 - 24 + 4 + 7 = 19$. No long division is needed.

(b) Here $x + 1 = x - (-1)$. In $\mathbb{Q}$, $(-1)^2 + 1 = 2 \neq 0$, so -1 is not a root and $x + 1 \nmid x^2 + 1$ by part (b) of the theorem; the remainder is 2. In $\mathbb{F}_2$, $-1 = 1$ and $1^2 + 1 = 0$, so $x + 1 \mid x^2 + 1$. Indeed $x^2 + 1 = (x + 1)^2$ in $\mathbb{F}_2[x]$. The same polynomial symbols, over two fields, give different answers.

(c) The divisor has degree 2, so by Theorem 6.1.1, $x^5 = q\,(x^2 - x) + r$ with $\deg r \leq 1$, say $r = ax+b$. The divisor $x^2-x = x(x-1)$ vanishes at 0 and at 1, so evaluating both sides there (Theorem 0.8.5) kills the term $q\,(x^2-x)$: at 0, $0 = b$; at 1, $1 = a+b$. Hence $a = 1$, $b = 0$, and the remainder is x. As a check, $x^5-x = x(x^4-1) = (x^2-x)(x^3+x^2+x+1)$. The move is: **evaluate at the roots of the divisor**, one equation per root.

The Factor Theorem peels off one root at a time. Doing it for several distinct roots at once strengthens the Root bound of Chapter 3 (Lemma 3.7.2), which said that a nonzero polynomial of degree at most n has at most n distinct roots. The stronger version says **why** there are so few: every root contributes its own linear factor.

Corollary 6.1.4. Distinct Roots Give a Product of Linear Factors

Let F be a field, $f \in F[x]$, and let $c_1, \ldots, c_k \in F$ be **distinct** roots of f, with $k \geq 1$. Then

$$(x - c_1)(x - c_2) \cdots (x - c_k) \mid f.$$

Consequently, if $f \neq 0$, then $k \leq \deg f$: a nonzero polynomial of degree n has at most n distinct roots in F.

Idea. Peel off $x-c_1$ by the Factor Theorem. What remains, h, must still vanish at c_2, because $f(c_2) = (c_2 - c_1)h(c_2)$ and $c_2 - c_1 \neq 0$. So peel off $x - c_2$ from h, and so on. Distinctness is what makes the factors $c_k - c_i$ nonzero.

Proof. We use induction on k. For $k = 1$, this is Theorem 6.1.3 (b).

Let $k \geq 2$, and assume the statement for $k - 1$ distinct roots. By the induction hypothesis, $f = (x - c_1) \cdots (x - c_{k-1})\, h$ for some $h \in F[x]$. Evaluating at c_k with Theorem 0.8.5,

$$0 = f(c_k) = (c_k - c_1) \cdots (c_k - c_{k-1})\, h(c_k).$$

Since the c_i are distinct, each $c_k - c_i$ with $i < k$ is nonzero, and a product of nonzero elements of a field is nonzero (Theorem 0.7.1). Hence $h(c_k) = 0$, and Theorem 6.1.3 (b) gives $h = (x - c_k)h_1$ with $h_1 \in F[x]$. Therefore $f = (x - c_1) \cdots (x - c_k)\, h_1$, which completes the induction.

For the count, suppose $f \neq 0$ and write $f = (x - c_1) \cdots (x - c_k)\, h_1$ as above. Then $h_1 \neq 0$, and by Theorem 0.8.3, $\deg f = k + \deg h_1 \geq k$. This proves the corollary.

■

The count is Lemma 3.7.2 again, now as a consequence of a divisibility statement. The divisibility statement is the more useful one. For instance, it tells us at once that a polynomial vanishing at 1, 2, and 3 is a multiple of $(x - 1)(x - 2)(x - 3)$.

6.1.5 Polynomials and polynomial functions

Chapter 0 warned that a polynomial is not the function it defines: over $\mathbb{F}_2$, the nonzero polynomial $x^2 + x$ has the zero function (Example (Different polynomials, same function)). It also

promised that over an infinite field the distinction disappears. Chapter 3 kept that promise in a paragraph right after Lemma 3.7.2. We record the fact here as a theorem, in a slightly sharper form that counts how many values are needed.

Theorem 6.1.5. Values Determine a Polynomial

Let F be a field and $p, q \in F[x]$.

(a) Let $n \in \mathbb{N}$ with $\deg p \le n$ and $\deg q \le n$. If $p(c) = q(c)$ for at least $n+1$ **distinct** elements $c \in F$, then $p = q$.

(b) If F is **infinite** and $p(c) = q(c)$ for **every** $c \in F$, then $p = q$.

Proof.

(a) Let $h = p - q$. By Theorem 0.8.2, $\deg h \le n$. By Theorem 0.8.5, $h(c) = p(c) - q(c) = 0$ for each of the given $n+1$ distinct elements, so h has at least $n+1$ distinct roots. By Lemma 3.7.2, a nonzero polynomial of degree at most n has at most n distinct roots, so h cannot be nonzero. Hence $h = 0$ and $p = q$.

(b) Let $n = \max(\deg p, \deg q, 0) \in \mathbb{N}$. Since F is infinite, it contains $n+1$ distinct elements, and p and q agree at each of them. Part (a) gives $p = q$. This proves the theorem.

■

So over $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$, we may compare polynomials by comparing their functions, and we may read off coefficients from values. Part (a) holds over **every** field, including finite ones: a quadratic over $\mathbb{F}_5$ is determined by its values at any three distinct points.

Warning. Part (b) is false over every finite field. Over $\mathbb{F}_p$, the polynomial $x^p - x$ is nonzero but vanishes at every element (Exercise (C2: A nonzero polynomial vanishing everywhere)). So x^p and x define the same function $\mathbb{F}_p \to \mathbb{F}_p$ although they are different polynomials. There is no contradiction with (a): the degrees are at most p, and (a) would need agreement at $p+1$ distinct points, but $\mathbb{F}_p$ has only p elements.

Quick check. Over $\mathbb{F}_3$, do the polynomials $x^3 + 2x$ and 0 define the same function? Does this contradict Theorem 6.1.5?

Solution. Yes. The values of $x^3 + 2x$ at $0, 1, 2$ are 0, $1 + 2 = 3 = 0$, and $8 + 4 = 12 = 0$. So the function is zero, although the polynomial is not. There is no contradiction: $\mathbb{F}_3$ is finite, so part (b) does not apply, and part (a) with $n = 3$ would need 4 distinct points. In fact $x^3 + 2x = x^3 - x$, the polynomial of the warning with $p = 3$.

Division with remainder is the one tool this chapter keeps using. In the next section it shows that every set of polynomials closed under the right operations consists of the multiples of a single polynomial, and that gives greatest common divisors.

6.1.6 Exercises

6.1.6.1 A. Check your understanding

Exercise. A1

(a) State the theorem on division with remainder in $F[x]$, including every hypothesis.

(b) Determine whether the following statement is true: "for every field F and all $f, g \in F[x]$, if $f \mid g$ and $g \mid f$, then $f = g$." Justify your answer. Does your answer change if $F = \mathbb{F}_2$?

(c) Let $c \in F$ be nonzero. What are the quotient and the remainder of f on division by the constant polynomial c?

(d) Name the method for finding the remainder of f on division by $x - c$ without dividing.

(e) Determine whether the following statement is true: "if F is a field and $p \in F[x]$ satisfies $p(c) = 0$ for every $c \in F$, then $p = 0$." Justify your answer.

Solution.

(a) Let F be a field and $f, g \in F[x]$ with $g \neq 0$. Then there are unique $q, r \in F[x]$ with $f = qg + r$ and $\deg r < \deg g$ (Theorem 6.1.1).

(b) False in general. In $\mathbb{Q}[x]$, take $f = x$ and $g = 2x$. Then $g = 2f$ and $f = \frac{1}{2}g$, so each divides the other, but $f \neq g$. By Proposition 6.1.2 (d), mutual divisibility only gives associates. Over $\mathbb{F}_2$, the statement is **true**: the only nonzero constant in $\mathbb{F}_2$ is 1, so associates are equal, and if $f \mid g$ and $g \mid f$, then either both are 0 or $f = 1 \cdot g$.

(c) The remainder must have degree less than $\deg c = 0$, so $r = 0$. Then $f = qc$, so $q = c^{-1}f$. In particular every polynomial is divisible by every nonzero constant.

(d) The Remainder Theorem (Theorem 6.1.3): the remainder is the constant $f(c)$.

(e) False. Over $\mathbb{F}_2$, $p = x^2 + x$ is nonzero but $p(0) = p(1) = 0$ (Example (Different polynomials, same function)). The statement is true when F is infinite, by Theorem 6.1.5 (b).

6.1.6.2 B. Practice

Exercise. B1: Dividing Over Two Fields

(a) In $\mathbb{Q}[x]$, divide $f = x^4 - 3x^3 + 2x - 5$ by $g = 3x^2 - 1$ with remainder.

(b) In $\mathbb{F}_5[x]$, divide $f = x^4 + 2x^3 + 3x + 1$ by $g = 2x^2 + 3$ with remainder. Hence decide whether $g \mid f$ in $\mathbb{F}_5[x]$.

Hint for (b): first find the inverse of 2 in $\mathbb{F}_5$.

Solution.

(a) The leading coefficient of g is 3, with inverse $\frac{1}{3}$. First, $x^4/3x^2 = \frac{1}{3}x^2$, and $f - \frac{1}{3}x^2 g = -3x^3 + \frac{1}{3}x^2 + 2x - 5$. Next, $-3x^3/3x^2 = -x$, and subtracting $-xg = -3x^3 + x$ leaves $\frac{1}{3}x^2 + x - 5$. Next, $\frac{1}{3}x^2/3x^2 = \frac{1}{9}$, and subtracting $\frac{1}{9}g = \frac{1}{3}x^2 - \frac{1}{9}$ leaves $x - \frac{44}{9}$, of degree

$1 < 2$. Hence

$$x^4 - 3x^3 + 2x - 5 = (\tfrac{1}{3}x^2 - x + \tfrac{1}{9})(3x^2 - 1) + x - \tfrac{44}{9},$$

with quotient $\frac{1}{3}x^2 - x + \frac{1}{9}$ and remainder $x - \frac{44}{9}$, which are the only ones by Theorem 6.1.1.

(b) In $\mathbb{F}_5$, $2 \cdot 3 = 6 = 1$, so $2^{-1} = 3$. First, the leading term x^4 divided by $2x^2$ is $3x^2$, and $3x^2 g = 6x^4 + 9x^2 = x^4 + 4x^2$. Subtracting leaves $2x^3 - 4x^2 + 3x + 1 = 2x^3 + x^2 + 3x + 1$. Next, $2x^3/2x^2 = x$, and subtracting $xg = 2x^3 + 3x$ leaves $x^2 + 1$. Next, $x^2/2x^2 = 3$, and subtracting $3g = 6x^2 + 9 = x^2 + 4$ leaves $1 - 4 = -3 = 2$. Hence

$$x^4 + 2x^3 + 3x + 1 = (3x^2 + x + 3)(2x^2 + 3) + 2 \quad \text{in } \mathbb{F}_5[x].$$

The remainder 2 is nonzero, so $g \nmid f$ in $\mathbb{F}_5[x]$, by Proposition 6.1.2 (e).

Exercise. B2: A Factor of $x^n - c^n$

Let F be a field, $c \in F$, and $n \geq 1$. Prove that $x - c \mid x^n - c^n$ in $F[x]$, and verify that the quotient is $x^{n-1} + c\,x^{n-2} + \cdots + c^{n-2}x + c^{n-1}$.

Solution. The value of $f = x^n - c^n$ at c is $c^n - c^n = 0$, so c is a root of f, and Theorem 6.1.3 (b) gives $x - c \mid f$. For the quotient, let $s = \sum_{j=0}^{n-1} c^j x^{n-1-j}$. Then

$$(x - c)s = \sum_{j=0}^{n-1} c^j x^{n-j} - \sum_{j=0}^{n-1} c^{j+1} x^{n-1-j} = \sum_{j=0}^{n-1} c^j x^{n-j} - \sum_{j=1}^{n} c^j x^{n-j} = x^n - c^n,$$

since the two sums share the terms with $1 \leq j \leq n-1$, leaving $c^0 x^n$ from the first and $c^n x^0$ from the second. This shows $x^n - c^n = (x - c)s$.

Exercise. B3: A Remainder Without Long Division

(a) In $\mathbb{Q}[x]$, find the remainder of x^{100} on division by $x^2 - 1$ without long division.

(b) Show that the answer to (a) is the remainder of x^{100} on division by $x^2 - 1$ over **every** field F, including $\mathbb{F}_2$. Explain why the method of (a) alone does not settle the case $F = \mathbb{F}_2$.

Hint for (a): part (c) of Example (Remainders Without Dividing). Hint for (b): imitate the identity of B2.

Solution.

(a) By Theorem 6.1.1, $x^{100} = q(x^2 - 1) + r$ with $\deg r \leq 1$, say $r = ax + b$. Evaluating at 1 and -1, where $x^2 - 1$ vanishes (Theorem 0.8.5), gives $1 = a + b$ and $1 = (-1)^{100} = -a + b$. Adding and subtracting, $2b = 2$ and $2a = 0$, so $b = 1$ and $a = 0$. The remainder is 1.

(b) Over any field, let $s = \sum_{j=0}^{49} x^{2j}$. Then

$$(x^2-1)s = \sum_{j=0}^{49} x^{2j+2} - \sum_{j=0}^{49} x^{2j} = x^{100} - 1,$$

because the terms $x^2, x^4, \ldots, x^{98}$ occur in both sums. Hence $x^{100} = s\cdot(x^2-1)+1$, and $\deg 1 = 0 < 2$, so by the uniqueness in Theorem 6.1.1 the remainder is 1 over every field. The method of (a) divides by 2, and over $\mathbb{F}_2$ we have $2 = 0$; also $1 = -1$ there, so evaluating at ± 1 gives only the single equation $a + b = 1$, which does not determine a and b.

6.1.6.3 C. Going deeper

Exercise. C1: Division by a Monic Polynomial Over $\mathbb{Z}$

As in the warning after Theorem 6.1.1, $\mathbb{Z}[x]$ denotes the set of polynomials in $\mathbb{Q}[x]$ all of whose coefficients are integers. It is closed under addition and multiplication.

(a) Let $f, g \in \mathbb{Z}[x]$ with g **monic**. Prove that there exist $q, r \in \mathbb{Z}[x]$ with $f = qg + r$ and $\deg r < \deg g$.

(b) Prove that q and r in (a) are unique, even among $q, r \in \mathbb{Q}[x]$.

(c) Does (a) remain true if the leading coefficient of g is -1 instead of 1? Justify your answer.

Hint for (a): find the one step in the proof of Theorem 6.1.1 that leaves $\mathbb{Z}[x]$.

Solution.

(a) We repeat the existence proof of Theorem 6.1.1, checking that every polynomial stays in $\mathbb{Z}[x]$. Let $m = \deg g$. The only step that could leave $\mathbb{Z}[x]$ is forming $ab^{-1}x^{n-m}g$, where b is the leading coefficient of g. Here $b = 1$, so $ab^{-1} = a \in \mathbb{Z}$ and $f_1 = f - ax^{n-m}g \in \mathbb{Z}[x]$. The strong induction on $\deg f$ (Theorem 0.2.10) then runs word for word inside $\mathbb{Z}[x]$: in Case 1, $q = 0$ and $r = f$ are in $\mathbb{Z}[x]$; in Case 2, $q = ax^{n-m} + q_1$ and $r = r_1$ are in $\mathbb{Z}[x]$ because q_1, r_1 are, by the induction hypothesis applied to $f_1 \in \mathbb{Z}[x]$. This proves existence in $\mathbb{Z}[x]$.

(b) Suppose $f = qg + r = q'g + r'$ with $q, q', r, r' \in \mathbb{Q}[x]$ and $\deg r, \deg r' < \deg g$. Since $\mathbb{Q}$ is a field and $g \neq 0$, the uniqueness part of Theorem 6.1.1 in $\mathbb{Q}[x]$ gives $q = q'$ and $r = r'$. In particular the pair found in (a) is the unique one.

(c) Yes. If the leading coefficient is $b = -1$, then $b^{-1} = -1 \in \mathbb{Z}$, so $ab^{-1} = -a \in \mathbb{Z}$, and the argument of (a) goes through unchanged. What matters is that the leading coefficient has an inverse **in** $\mathbb{Z}$, which is true exactly for ± 1.

Exercise. C2: Remainders on Division by a Quadratic

Let F be a field, $f \in F[x]$, and $a, b \in F$.

(a) Suppose $a \neq b$. Prove that the remainder of f on division by $(x-a)(x-b)$ is

$$r = f(a)\,\frac{x-b}{a-b} + f(b)\,\frac{x-a}{b-a},$$

where $\frac{x-b}{a-b}$ means $(a-b)^{-1}(x-b)$.

(b) Now let $a = b$. Write $f = (x-a)h + f(a)$ by Theorem 6.1.3. Prove that the remainder of f on division by $(x-a)^2$ is $h(a)(x-a) + f(a)$.

(c) Hence find the remainder of x^{100} on division by $x^2 - 3x + 2$ in $\mathbb{Q}[x]$.

Solution.

(a) Since $a \neq b$, $a-b$ is invertible, so r is a well-defined polynomial of degree at most 1. By Theorem 6.1.1, $f = q(x-a)(x-b) + s$ with $\deg s \leq 1$. Evaluating at a and at b (Theorem 0.8.5) gives $s(a) = f(a)$ and $s(b) = f(b)$. Also $r(a) = f(a)\cdot 1 + f(b)\cdot 0 = f(a)$ and $r(b) = f(b)$. So s and r have degree at most 1 and agree at the 2 distinct points a, b. By Theorem 6.1.5 (a) with $n = 1$, $s = r$. This proves the formula.

(b) By Theorem 6.1.3 (a) applied to h, $h = (x-a)k + h(a)$ for some $k \in F[x]$. Substituting,

$$f = (x-a)\big((x-a)k + h(a)\big) + f(a) = k\,(x-a)^2 + \big(h(a)(x-a) + f(a)\big).$$

The last bracket has degree at most $1 < 2 = \deg(x-a)^2$, so by the uniqueness in Theorem 6.1.1 it is the remainder.

(c) Here $x^2 - 3x + 2 = (x-1)(x-2)$, with $a = 1 \neq b = 2$, $f(1) = 1$, and $f(2) = 2^{100}$. By (a),

$$r = 1 \cdot \frac{x-2}{1-2} + 2^{100} \cdot \frac{x-1}{2-1} = (2^{100} - 1)x + (2 - 2^{100}).$$

As a check, $r(1) = 1$ and $r(2) = 2^{101} - 2 + 2 - 2^{100} = 2^{100}$.

Exercise. C3: Equal Functions Over $\mathbb{F}_p$

Let p be a prime and $f, g \in \mathbb{F}_p[x]$.

(a) Prove that $f(c) = g(c)$ for every $c \in \mathbb{F}_p$ if and only if $x^p - x \mid f - g$ in $\mathbb{F}_p[x]$.

(b) Deduce that f and its remainder on division by $x^p - x$ define the same function $\mathbb{F}_p \to \mathbb{F}_p$. Use this to find a polynomial of degree less than 5 that defines the same function on $\mathbb{F}_5$ as $x^7 + x$.

Hint for (a): Exercise (C2: A nonzero polynomial vanishing everywhere) and Corollary 6.1.4.

Solution.

(a) ($\Leftarrow$) Suppose $f - g = (x^p - x)h$. By Exercise (C2: A nonzero polynomial vanishing everywhere), $c^p - c = 0$ for every $c \in \mathbb{F}_p$. By Theorem 0.8.5, $f(c) - g(c) = (c^p - c)h(c) = 0$ for every c. ($\Rightarrow$) Suppose $f(c) = g(c)$ for every $c \in \mathbb{F}_p$. Then the p distinct elements $0, 1, \ldots, p-1$ of $\mathbb{F}_p$ are roots of $f - g$. Let $P = \prod_{c \in \mathbb{F}_p}(x - c)$. By Corollary 6.1.4, $P \mid f - g$. It remains to show $P = x^p - x$. By Exercise (C2: A nonzero polynomial vanishing everywhere), every $c \in \mathbb{F}_p$ is a root of $x^p - x$, so Corollary 6.1.4 gives $x^p - x = Pu$ for some u. Both P and $x^p - x$ are monic of degree p (Theorem 0.8.3), so $\deg u = 0$, and

comparing leading coefficients gives $u = 1$. Hence $P = x^p - x$, and $x^p - x \mid f - g$.

(b) Write $f = q(x^p - x) + r$ by Theorem 6.1.1. Then $x^p - x \mid f - r$, so f and r define the same function by (a). For $f = x^7 + x$ over $\mathbb{F}_5$: $x^7 = x^2 \cdot x^5 = x^2(x^5 - x) + x^3$, so $x^7 + x = x^2(x^5 - x) + (x^3 + x)$ with $\deg(x^3 + x) = 3 < 5$. Hence $x^3 + x$ defines the same function as $x^7 + x$ on $\mathbb{F}_5$.

6.2 Ideals and Greatest Common Divisors

Division with remainder is now available over every field. This section uses it to answer a question about two polynomials f and g: what do all the combinations $af + bg$, with **polynomial** coefficients a and b, look like? The answer is that they are exactly the multiples of one polynomial, and that polynomial is the greatest common divisor of f and g. Along the way we meet ideals, the subsets of $F[x]$ adapted to multiplication, and we prove that each of them is generated by a single polynomial. The Euclidean algorithm then computes the greatest common divisor and the coefficients a, b by repeated division. Throughout, F is a field.

6.2.1 All the combinations of two polynomials

Take $f = x^2 - 1$ and $g = x^2 + x - 2$ in $\mathbb{Q}[x]$, and look at the set

$$\{af + bg : a, b \in \mathbb{Q}[x]\}$$

of all their combinations with polynomial coefficients. It contains f (take $a = 1, b = 0$), g, and $g - f = x - 1$.

Since $f = (x - 1)(x + 1)$ and $g = (x - 1)(x + 2)$, every combination is

$$af + bg = (x - 1)\big(a(x + 1) + b(x + 2)\big),$$

a multiple of $x - 1$. Conversely, every multiple $h(x - 1)$ is a combination, namely $(-h)f + hg$. So the set of combinations is **exactly** the set of multiples of $x - 1$. Its nonzero members of smallest degree are the nonzero scalar multiples of $x - 1$. By Proposition 6.1.2 (c), no nonzero constant is a multiple of $x - 1$, so there is nothing of degree 0.

Two features of this set will carry the whole section. It is closed under addition, and it absorbs multiplication by **any** polynomial: if $af + bg$ is in the set, so is $h(af + bg) = (ha)f + (hb)g$. Sets with these two features appear everywhere, so they get a name.

6.2.2 Ideals

In Chapter 1 we studied $F[x]$ as a vector space and its subspaces: subsets closed under addition and under multiplication by **scalars**. But $F[x]$ also has a multiplication of its own. The subsets that respect it are closed under multiplication by **polynomials**. Three recurring examples share this feature:

- the multiples of a fixed polynomial p;
- the combinations $af + bg$ of two fixed polynomials;
- the polynomials that vanish at a fixed point c: if $f(c) = 0$, then $(hf)(c) = h(c) \cdot 0 = 0$ for every h.

An ideal is a set of polynomials that is closed under addition and swallows multiplication by anything.

Definition 6.2.1. Ideal of $F[x]$

Let F be a field. A subset $I \subseteq F[x]$ is an **ideal** of $F[x]$ if

(I1) I is **nonempty**;
(I2) $f + g \in I$ **for all** $f, g \in I$;
(I3) $hf \in I$ **for all** $f \in I$ and **every** $h \in F[x]$.

For $f_1, \dots, f_k \in F[x]$, write

$$\langle f_1, \dots, f_k \rangle := \{a_1 f_1 + \cdots + a_k f_k : a_1, \dots, a_k \in F[x]\}.$$

In particular $\langle p \rangle = \{hp : h \in F[x]\}$ is the set of all multiples of p.

In words: (I1) rules out the empty set. (I2) says sums stay inside. (I3) is the new clause: a member of I multiplied by **any** polynomial, not just by a constant, is still a member. Taking $h = 0$ in (I3) shows $0 \in I$, and taking $h = -1$ shows $-f \in I$. Taking $h = c$ constant shows I is closed under scalar multiplication, so **every ideal is a subspace** of $F[x]$ by Theorem 1.2.1.

The angle brackets here have nothing to do with the inner products of later chapters: $\langle f, g \rangle$ is a set of polynomials, not a number.

Well-definedness. We have named the set $\langle f_1, \dots, f_k \rangle$, and must check that it is an ideal. It is in fact the smallest ideal containing $f_1, \dots, f_k$.

Proposition 6.2.1. The Ideal Generated by Finitely Many Polynomials

Let F be a field and $f_1, \dots, f_k \in F[x]$, with $k \geq 1$. Then $\langle f_1, \dots, f_k \rangle$ is an ideal of $F[x]$ containing $f_1, \dots, f_k$. Moreover, every ideal J of $F[x]$ with $f_1, \dots, f_k \in J$ satisfies $\langle f_1, \dots, f_k \rangle \subseteq J$.

Proof. Let $I = \langle f_1, \dots, f_k \rangle$. Taking $a_i = 1$ and $a_j = 0$ for $j \neq i$ shows $f_i \in I$; in particular (I1) holds. For (I2), $\sum a_i f_i + \sum b_i f_i = \sum (a_i + b_i) f_i \in I$. For (I3), $h \sum a_i f_i = \sum (h a_i) f_i \in I$, using Theorem 0.8.1 in both cases. So I is an ideal.

Now let J be an ideal containing every f_i. By (I3) for J, each $a_i f_i \in J$, and by (I2) applied $k - 1$ times, $a_1 f_1 + \cdots + a_k f_k \in J$. Hence $I \subseteq J$. This proves the proposition. ∎

The ideal $\langle f_1, \dots, f_k \rangle$ is the **ideal generated by** $f_1, \dots, f_k$. An ideal of the form $\langle p \rangle$, generated by one polynomial, is called **principal**.

Example. First Ideals

Check each of the following against Definition 6.2.1.

(a) $\{0\}$ and $F[x]$; and any ideal containing a nonzero constant.
(b) $I_c = \{f \in F[x] : f(c) = 0\}$ for a fixed $c \in F$.

(c) For $\mathbf{A} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \in M_2(F)$, the set $\{p \in F[x] : p(\mathbf{A}) = \mathbf{O}\}$.

Solution.

(a) $\{0\} = \langle 0 \rangle$ and $F[x] = \langle 1 \rangle$ are ideals by Proposition 6.2.1. They are the degenerate cases, and the second one is the one to recognize: if an ideal I contains a nonzero constant c, then $1 = c^{-1}c \in I$ by (I3), and then $h = h \cdot 1 \in I$ for every h, so $I = F[x]$. An ideal that contains a nonzero constant is everything.

(b) $0 \in I_c$. If $f(c) = g(c) = 0$, then $(f+g)(c) = 0$ and $(hf)(c) = h(c)f(c) = 0$ for every $h \in F[x]$, by Theorem 0.8.5. So I_c is an ideal. By Theorem 6.1.3 (b), $f(c) = 0$ exactly when $x - c \mid f$, so $I_c = \langle x - c \rangle$.

(c) Call the set I. The zero polynomial is in I. If $p(\mathbf{A}) = q(\mathbf{A}) = \mathbf{O}$, then by Theorem 0.9.9, $(p+q)(\mathbf{A}) = p(\mathbf{A}) + q(\mathbf{A}) = \mathbf{O}$ and $(hp)(\mathbf{A}) = h(\mathbf{A})\,p(\mathbf{A}) = h(\mathbf{A}) \cdot \mathbf{O} = \mathbf{O}$ for every h. So I is an ideal. It contains x^2, since $\mathbf{A}^2 = \mathbf{O}$. It contains no nonzero polynomial of degree at most 1: $(a + bx)(\mathbf{A}) = \begin{pmatrix} a & b \\ 0 & a \end{pmatrix} = \mathbf{O}$ forces $a = b = 0$. The same argument works for any linear operator T in place of $\mathbf{A}$, with Theorem 2.4.3. In Chapter 9 the ideal of all polynomials killing an operator is where the minimal polynomial comes from.

Now the nonexample, by a minimal change. Take $F[x]_{\leq n}$ for some $n \in \mathbb{N}$. It is nonempty, closed under addition, and closed under multiplication by **constants**: it is a subspace of $F[x]$ (Example (Polynomials of Bounded Degree)). But it fails (I3): $x^n \in F[x]_{\leq n}$ and $x \in F[x]$, while $x \cdot x^n = x^{n+1} \notin F[x]_{\leq n}$. The exact clause that fails is "for **every** $h \in F[x]$".

Why insist on every h, and not just constants? Because the main theorem below divides a member f of I by another member d and needs the remainder $f - qd$ to lie in I again. The quotient q is an arbitrary polynomial, so we need $qd \in I$. With only closure under scalars, the argument breaks, and indeed $F[x]_{\leq n}$ is not the set of multiples of any single polynomial. The name "ideal" is historical and says nothing about the definition.

Warning. Do not confuse $\langle f, g \rangle$ with $\operatorname{span}(f, g)$. The span uses **scalar** coefficients, the ideal uses **polynomial** coefficients. For example, $\operatorname{span}(x) = \{cx : c \in F\}$ is a line in $F[x]$, while $\langle x \rangle$ contains $x^2, x^3, \ldots$ and is infinite-dimensional. Every ideal is a subspace, but most subspaces, such as $F[x]_{\leq n}$ or $\operatorname{span}(x)$, are not ideals.

Quick check. Is $J = \{f \in \mathbb{R}[x] : f(1) = 0 \text{ or } f(2) = 0\}$ an ideal of $\mathbb{R}[x]$?

Solution. No. J is nonempty and satisfies (I3), since $(hf)(c) = h(c)f(c)$. But it fails (I2): $x - 1 \in J$ and $x - 2 \in J$, while their sum $2x - 3$ takes the values -1 at 1 and 1 at 2, so $2x - 3 \notin J$. The set with "and" in place of "or" is an ideal.

6.2.3 Every ideal is generated by one polynomial

The payoff of the definition is that ideals of $F[x]$ are very simple: each one is the set of multiples of a single polynomial. The proof is the argument of the opening example, made general. We have met it before for integers, in Lemma 0.7.2, where the smallest positive combination $ax+ny$ turned out to divide a and n.

Theorem 6.2.2. Every Ideal of $F[x]$ Is Principal

Let F be a field and I an ideal of $F[x]$. If $I \neq \{0\}$, there is **exactly one monic** polynomial d with $I = \langle d \rangle$. Moreover, d is the unique monic polynomial of least degree in I.

Idea. Pick a nonzero $d \in I$ of smallest degree, made monic. Every multiple of d is in I by (I3). For the other inclusion, divide any $f \in I$ by d: the remainder $r = f - qd$ lies in I and has degree smaller than $\deg d$. Minimality leaves only one option, $r = 0$. Uniqueness is mutual divisibility of monic polynomials.

Proof. Suppose $I \neq \{0\}$. The degrees of the nonzero elements of I form a nonempty set of natural numbers, so by Theorem 0.2.11 it has a least element m. Choose $g \in I$ with $\deg g = m$, and let c be its leading coefficient. Put $d = c^{-1}g$. Then $d \in I$ by (I3), and d is monic of degree m.

($\supseteq$) For every $h \in F[x]$, $hd \in I$ by (I3). Hence $\langle d \rangle \subseteq I$.

($\subseteq$) Let $f \in I$. By Theorem 6.1.1, $f = qd + r$ with $\deg r < m$. By (I3), $-qd \in I$, and by (I2), $r = f + (-qd) \in I$. If $r \neq 0$, then r is a nonzero element of I of degree less than m, contradicting the choice of m. Hence $r = 0$, so $f = qd \in \langle d \rangle$.

Therefore $I = \langle d \rangle$. For uniqueness, let d' be monic with $I = \langle d' \rangle$. Then $d \in \langle d' \rangle$ and $d' \in \langle d \rangle$, so $d' \mid d$ and $d \mid d'$. Both are monic, so $d = d'$ by Proposition 6.1.2 (d).

Finally, let $e \in I$ be monic of least degree m. Then $d \mid e$, so $e = ud$ with $\deg u = 0$ by Theorem 0.8.3, and comparing leading coefficients gives $u = 1$. Hence $e = d$. This proves the theorem.

■

For $I = \{0\}$ the only generator is 0, since $\langle p \rangle$ contains p. So every ideal of $F[x]$ is principal, with a canonical generator: 0 for the zero ideal, and the monic polynomial of least degree otherwise. The named move is: **the smallest-degree element of an ideal generates it.** In Example (First Ideals) (c), the ideal contains x^2 and nothing nonzero of smaller degree, so it equals $\langle x^2 \rangle$.

The theorem also turns divisibility into containment. For $p, q \in F[x]$,

$$\langle p \rangle \subseteq \langle q \rangle \quad \Longleftrightarrow \quad q \mid p,$$

because $\langle p \rangle \subseteq \langle q \rangle$ holds exactly when $p \in \langle q \rangle$, by Proposition 6.2.1. The bigger the ideal, the smaller the generator.

Quick check. Find the monic generator of $\langle x^2 - 1,\ x^3 - 1\rangle \subseteq \mathbb{Q}[x]$ by finding an element of small degree in it.

Solution. The ideal contains $(x^3 - 1) - x(x^2 - 1) = x - 1$. Both generators are multiples of $x - 1$, since $x^2 - 1 = (x - 1)(x + 1)$ and $x^3 - 1 = (x - 1)(x^2 + x + 1)$, so every combination is a multiple of $x - 1$ by Proposition 6.1.2 (b). A nonzero constant is not a multiple of $x - 1$, so $x - 1$ has least degree in the ideal. By Theorem 6.2.2, the ideal is $\langle x - 1\rangle$.

6.2.4 The greatest common divisor

Apply Theorem 6.2.2 to the ideal $\langle f, g\rangle$ of all combinations. It has a canonical generator, and in the opening example that generator, $x - 1$, was a common divisor of f and g. A naive attempt to define "the greatest common divisor" as "a common divisor of largest degree" runs into a problem at once: $x - 1$, $2x - 2$, and $-\frac{1}{3}x + \frac{1}{3}$ are all common divisors of $x^2 - 1$ and $x^2 + x - 2$ of the largest possible degree. Degree cannot tell associates apart, and it also does not say how the common divisors relate to each other. The generator of the ideal fixes both problems.

The gcd of f and g is the monic polynomial whose multiples are exactly the combinations of f and g.

Definition 6.2.2. Greatest Common Divisor

Let F be a field and $f, g \in F[x]$. If f and g are **not both zero**, the **greatest common divisor** $\gcd(f, g)$ is the **unique monic** polynomial $d \in F[x]$ with

$$\langle f, g\rangle = \langle d\rangle.$$

We set $\gcd(0, 0) = 0$.

In words: form the ideal of all combinations $af + bg$ with $a, b \in F[x]$; it is generated by one polynomial; take the monic one. The definition refers to $F[x]$ twice, in the coefficients a, b and in the generator d, so at first sight the gcd depends on the field. We show below that it does not change when the field is enlarged.

Well-definedness. By Proposition 6.2.1, $\langle f, g\rangle$ is an ideal. If f and g are not both zero, it contains a nonzero element, so by Theorem 6.2.2 there is exactly one monic d with $\langle f, g\rangle = \langle d\rangle$. The convention $\gcd(0, 0) = 0$ keeps the equation $\langle f, g\rangle = \langle \gcd(f, g)\rangle$ true in the remaining case. The definition is symmetric: $\gcd(f, g) = \gcd(g, f)$.

Example. First Greatest Common Divisors

(a) Let $f \neq 0$ have leading coefficient a. Find $\gcd(f, 0)$ and $\gcd(f, c)$ for a nonzero constant c.

(b) Find $\gcd(x^2 - 1, x^2 + x - 2)$ in $\mathbb{Q}[x]$.

(c) Find $\gcd(x^2 + 1, x + 1)$ in $\mathbb{Q}[x]$ and in $\mathbb{F}_2[x]$.

Solution.

(a) $\langle f, 0\rangle = \{af : a \in F[x]\} = \langle f\rangle = \langle a^{-1}f\rangle$, and $a^{-1}f$ is monic, so $\gcd(f, 0) = a^{-1}f$, the monic associate of f. This degenerate case shows why the gcd must be made monic: f itself generates the ideal too, but is not the canonical choice. Next, $\langle f, c\rangle$ contains the nonzero constant c, so it is $F[x] = \langle 1\rangle$ by Example (First Ideals) (a). Hence $\gcd(f, c) = 1$.

(b) The opening example showed $\langle x^2 - 1, x^2 + x - 2\rangle = \langle x - 1\rangle$, and $x - 1$ is monic. So the gcd is $x - 1$.

(c) In $\mathbb{Q}[x]$: dividing, $x^2+1 = (x-1)(x+1)+2$. So $2 = (x^2+1)-(x-1)(x+1)$ lies in the ideal, which is therefore $\mathbb{Q}[x]$ by Example (First Ideals) (a). Hence gcd = 1. In $\mathbb{F}_2[x]$: $x^2+1 = (x+1)^2$, so $x^2+1 \in \langle x+1\rangle$, and $\langle x^2+1, x+1\rangle = \langle x+1\rangle$ by Proposition 6.2.1. Hence gcd $= x + 1$. The symbols are the same, but $\mathbb{F}_2$ is not a field containing $\mathbb{Q}$, so the two answers need not agree.

For a nonexample by minimal change, keep (b) but replace $x-1$ by the associate $3x-3$. It is still a common divisor of largest degree and still generates the ideal, but it is **not** $\gcd(x^2-1, x^2+x-2)$: the clause "monic" fails. Or keep "monic" but lower the degree: 1 is a monic common divisor, and it does not generate the ideal, since $1 \notin \langle x - 1\rangle$.

Why define the gcd through the ideal, rather than as "the largest common divisor"? Because the ideal definition delivers three things at once: existence and uniqueness (from Theorem 6.2.2), the fact that the gcd is a combination of f and g, and the fact that it is "greatest" in the strongest sense, namely that every common divisor **divides** it. Here is the last point.

Theorem 6.2.3. Characterization of the gcd

Let F be a field, $f, g \in F[x]$, and $d = \gcd(f, g)$.

(a) $d \mid f$ and $d \mid g$.

(b) If $h \in F[x]$ satisfies $h \mid f$ and $h \mid g$, then $h \mid d$.

(c) If f, g are not both zero, then d is the **only** monic polynomial with properties (a) and (b).

Proof.

(a) $f = 1 \cdot f + 0 \cdot g \in \langle f, g\rangle = \langle d\rangle$, so $d \mid f$. Similarly $d \mid g$.

(b) Since $d \in \langle d\rangle = \langle f, g\rangle$, there are $a, b \in F[x]$ with $d = af + bg$. If $h \mid f$ and $h \mid g$, then $h \mid af + bg = d$ by Proposition 6.1.2 (b).

(c) Let e be monic with properties (a) and (b). Since d is a common divisor of f and g, property (b) for e gives $d \mid e$. Since e is a common divisor, property (b) for d gives $e \mid d$. Both are monic (e by assumption, and d by Definition 6.2.2, since f, g are not both zero), so $d = e$ by Proposition 6.1.2 (d). This proves the theorem.

■

So $\gcd(f, g)$ is a common divisor that every common divisor divides. If f, g are not both zero, then $d \neq 0$, and by Proposition 6.1.2 (c) every common divisor has degree at most $\deg d$;

"greatest" is true for degrees as well. The proof of (b) used one fact that deserves its own statement, because it is the most useful property of the gcd.

Corollary 6.2.4. Bézout's Identity for Polynomials

Let F be a field and $f, g \in F[x]$. Then there exist $a, b \in F[x]$ with

$$\gcd(f, g) = af + bg.$$

Proof. By Definition 6.2.2, $\gcd(f, g) \in \langle \gcd(f, g) \rangle = \langle f, g \rangle$, and the elements of $\langle f, g \rangle$ are exactly the combinations $af + bg$. ∎

The coefficients are not unique: if $d = af + bg$, then also $d = (a + g)f + (b - f)g$. Bézout's identity is the polynomial version of Lemma 0.7.2, and it will be the tool behind the Chinese Remainder Theorem and the splitting of kernels later in this chapter.

6.2.5 The Euclidean algorithm

The definition tells us what the gcd is, but not how to find it. Factoring f and g is usually hard. Division is easy, and one observation turns division into an algorithm: dividing does not change the ideal.

Lemma 6.2.5. Dividing Does Not Change the Ideal

Let F be a field and $f, g, q, r \in F[x]$ with $f = qg + r$. Then $\langle f, g \rangle = \langle g, r \rangle$. Consequently $\gcd(f, g) = \gcd(g, r)$.

Proof. ($\supseteq$) The ideal $\langle f, g \rangle$ contains g, and $r = f + (-q)g$ by (I2) and (I3). By Proposition 6.2.1, $\langle g, r \rangle \subseteq \langle f, g \rangle$.

($\subseteq$) The ideal $\langle g, r \rangle$ contains g and $f = qg + r$. By Proposition 6.2.1, $\langle f, g \rangle \subseteq \langle g, r \rangle$.

Hence the two ideals are equal. The gcd is the canonical generator of the ideal (Definition 6.2.2), and f, g are both zero exactly when g, r are, so $\gcd(f, g) = \gcd(g, r)$. ∎

Now divide repeatedly. Each division replaces the pair (f, g) by (g, r), a pair with the same gcd and a smaller second entry.

Theorem 6.2.6. Euclidean Algorithm

Let F be a field and $f, g \in F[x]$ with $g \neq 0$. Put $r_{-1} = f$ and $r_0 = g$. As long as $r_k \neq 0$, divide r_{k-1} by r_k (Theorem 6.1.1):

$$r_{k-1} = q_{k+1} r_k + r_{k+1}, \qquad \deg r_{k+1} < \deg r_k.$$

Then:

(a) the process stops: there is an index n with $0 \le n \le \deg g$ such that $r_n \neq 0$ and $r_{n+1} = 0$;
(b) if c is the leading coefficient of this last nonzero remainder r_n, then $\gcd(f, g) = c^{-1} r_n$.

Idea. Two things to see. The degrees $\deg r_0 > \deg r_1 > \deg r_2 > \cdots$ strictly decrease, and natural numbers cannot decrease forever, so some remainder is 0. By Lemma 6.2.5, the ideal $\langle r_{k-1}, r_k \rangle$ never changes along the way; at the end it is $\langle r_n, 0 \rangle = \langle r_n \rangle$, and making r_n monic gives the gcd.

Proof.

(a) As long as $r_1, \ldots, r_j$ are all nonzero, their degrees are natural numbers with $\deg g = \deg r_0 > \deg r_1 > \cdots > \deg r_j \ge 0$. So $r_1, \ldots, r_j$ are j distinct values in $\{0, 1, \ldots, \deg g - 1\}$, and $j \le \deg g$. Hence the remainders cannot all be nonzero for $j = \deg g + 1$: there is a first index $n + 1$ with $r_{n+1} = 0$, and $n \le \deg g$. Also $n \ge 0$ since $r_0 = g \neq 0$.
(b) We show by induction on k that $\langle r_{k-1}, r_k \rangle = \langle f, g \rangle$ for $k = 0, 1, \ldots, n + 1$. For $k = 0$ this is the definition of r_{-1}, r_0. If it holds for some $k \le n$, then $r_{k-1} = q_{k+1} r_k + r_{k+1}$, and Lemma 6.2.5 gives $\langle r_k, r_{k+1} \rangle = \langle r_{k-1}, r_k \rangle = \langle f, g \rangle$.

For $k = n + 1$ this reads $\langle f, g \rangle = \langle r_n, 0 \rangle = \langle r_n \rangle$. Since $c \neq 0$, $\langle r_n \rangle = \langle c^{-1} r_n \rangle$, and $c^{-1} r_n$ is monic. Because $g \neq 0$, Definition 6.2.2 gives $\gcd(f, g) = c^{-1} r_n$. This proves the theorem. ∎

If $g = 0$, no division is needed: $\gcd(f, 0)$ is the monic associate of f (or 0), as in Example (First Greatest Common Divisors) (a).

The algorithm also produces Bézout coefficients. Each remainder is a combination of the previous two, $r_{k+1} = r_{k-1} - q_{k+1} r_k$, so by induction each remainder is a combination of f and g. Keeping track of the coefficients gives the following procedure.

Algorithm. **Extended Euclidean algorithm.** Input: $f, g \in F[x]$ with $g \neq 0$.

1. Set $(r_{-1}, a_{-1}, b_{-1}) = (f, 1, 0)$ and $(r_0, a_0, b_0) = (g, 0, 1)$.
2. For $k = 0, 1, 2, \ldots$, as long as $r_k \neq 0$: divide $r_{k-1} = q_{k+1} r_k + r_{k+1}$, and set
$$a_{k+1} = a_{k-1} - q_{k+1} a_k, \qquad b_{k+1} = b_{k-1} - q_{k+1} b_k.$$
3. When $r_{n+1} = 0$, let c be the leading coefficient of r_n. Output $\gcd(f, g) = c^{-1} r_n$ and the coefficients $a = c^{-1} a_n$, $b = c^{-1} b_n$.

The invariant $r_k = a_k f + b_k g$ holds for $k = -1, 0$ by the choice in step 1. If it holds for $k - 1$ and k, then
$$r_{k+1} = r_{k-1} - q_{k+1} r_k = (a_{k-1} - q_{k+1} a_k) f + (b_{k-1} - q_{k+1} b_k) g = a_{k+1} f + b_{k+1} g.$$
So at the end $\gcd(f, g) = c^{-1} r_n = af + bg$. By hand, it is often quicker to write each division as "remainder = ..." and substitute backwards, which is the same computation.

Example. The Euclidean Algorithm with Back-Substitution

In $\mathbb{Q}[x]$, let $f = x^4 - x^3 - 3x^2 + x + 2$ and $g = x^3 - 2x - 1$. Find $\gcd(f, g)$ and polynomials a, b with $\gcd(f, g) = af + bg$.

Solution. Divisions. First divide f by g. The leading terms give x, and $f - xg = -x^3 - x^2 + 2x + 2$; then -1, and adding g leaves $-x^2 + 1$. So

$$f = (x - 1)\, g + r_1, \qquad r_1 = 1 - x^2.$$

Next divide g by r_1. The leading terms give $x^3/(-x^2) = -x$, and $g - (-x)r_1 = x^3 - 2x - 1 + x - x^3 = -x - 1$. So

$$g = (-x)\, r_1 + r_2, \qquad r_2 = -x - 1.$$

Next divide r_1 by r_2. Since $(x - 1)(-x - 1) = -x^2 + 1$,

$$r_1 = (x - 1)\, r_2 + 0.$$

The last nonzero remainder is $r_2 = -x - 1$, with leading coefficient -1. By Theorem 6.2.6, $\gcd(f, g) = x + 1$.

Back-substitution. From the first two divisions,

$$r_1 = f - (x - 1)g, \qquad r_2 = g + x\, r_1.$$

Substituting the first into the second,

$$r_2 = g + x(f - (x - 1)g) = xf + (1 + x - x^2)\, g.$$

Multiplying by $c^{-1} = -1$,

$$x + 1 = (-x)\, f + (x^2 - x - 1)\, g.$$

In the notation of the algorithm, $(a_1, b_1) = (1, 1 - x)$ and $(a_2, b_2) = (x, 1 + x - x^2)$.

Checks. Expanding, $-x(x^4 - x^3 - 3x^2 + x + 2) = -x^5 + x^4 + 3x^3 - x^2 - 2x$ and $(x^2 - x - 1)(x^3 - 2x - 1) = x^5 - x^4 - 3x^3 + x^2 + 3x + 1$, which sum to $x + 1$. As an independent check that $x + 1$ divides both, $f(-1) = 1 + 1 - 3 - 1 + 2 = 0$ and $g(-1) = -1 + 2 - 1 = 0$, so Theorem 6.1.3 applies.

6.2.6 Coprime polynomials

The most important value a gcd can take is 1. Then f and g share no nonconstant factor, and Bézout's identity writes 1 itself as a combination.

Definition 6.2.3. Coprime Polynomials

Let F be a field. Polynomials $f, g \in F[x]$ are **coprime** (or **relatively prime**) **over** F if $\gcd(f, g) = 1$.

Proposition 6.2.7. Coprime Means 1 Is a Combination

Let F be a field and $f, g \in F[x]$. Then f and g are coprime if and only if there exist $a, b \in F[x]$ with $af + bg = 1$.

Proof. ($\Rightarrow$) This is Corollary 6.2.4 with $\gcd(f, g) = 1$.

($\Leftarrow$) If $af + bg = 1$, then $1 \in \langle f, g \rangle$, so $\langle f, g \rangle = F[x] = \langle 1 \rangle$ by Example (First Ideals) (a). In particular f, g are not both zero, and the monic generator is 1. Hence $\gcd(f, g) = 1$. ■

Examples: for **distinct** $c, c' \in F$, the polynomials $x - c$ and $x - c'$ are coprime, since

$$(c' - c)^{-1}(x - c) - (c' - c)^{-1}(x - c') = (c' - c)^{-1}(c' - c) = 1.$$

A nonzero constant is coprime to every polynomial (Example (First Greatest Common Divisors) (a)). The polynomial 0 is coprime only to the nonzero constants. Nonexample: $x^2 - 1$ and $x + 1$ are not coprime, since $x + 1$ divides both, so $x + 1 \mid \gcd$ by Theorem 6.2.3 (b), and the gcd has degree at least 1.

Coprimality is what lets a divisor pass through a product.

Theorem 6.2.8. A Coprime Factor Can Be Canceled

Let F be a field and $p, q, r \in F[x]$. If $p \mid qr$ and $\gcd(p, q) = 1$, then $p \mid r$.

Proof. By Proposition 6.2.7, there are $a, b \in F[x]$ with $ap + bq = 1$. Multiplying by r, $r = (ar)p + b(qr)$. Here $p \mid p$ and $p \mid qr$ by hypothesis, so $p \mid r$ by Proposition 6.1.2 (b). This proves the theorem. ■

Warning. The coprimality hypothesis cannot be dropped. Take $p = x^2$ and $q = r = x$. Then $p \mid qr = x^2$, but $p \nmid r$, since $\deg x^2 > \deg x$ (Proposition 6.1.2 (c)). Here $\gcd(p, q) = x \neq 1$: part of p is absorbed by q and part by r.

In the next section, Theorem 6.2.8 becomes Euclid's lemma for irreducible polynomials, which is the heart of unique factorization.

Least common multiples, briefly. For $f, g \in F[x]$, the intersection $\langle f \rangle \cap \langle g \rangle$ is again an ideal: it contains 0, so (I1) holds, and (I2) and (I3) hold in each of $\langle f \rangle$ and $\langle g \rangle$, hence in the intersection. Its members are the common multiples of f and g.

Definition 6.2.4. Least Common Multiple

Let F be a field and $f, g \in F[x]$. The **least common multiple** $\operatorname{lcm}(f, g)$ is the monic generator of the ideal $\langle f \rangle \cap \langle g \rangle$ if this ideal is nonzero, and 0 otherwise.

Exactly as for the gcd, $\operatorname{lcm}(f, g)$ is a common multiple of f and g that divides every common multiple, by Theorem 6.2.2. If $f = 0$ or $g = 0$, the only common multiple is 0; otherwise fg is a nonzero common multiple. For monic f and g, $\gcd(f, g) \cdot \operatorname{lcm}(f, g) = fg$; this is proved in Exercise C3. For example, in $\mathbb{Q}[x]$, $\gcd(x^2 - 1, x^2 + x - 2) = x - 1$, so $\operatorname{lcm}(x^2 - 1, x^2 + x - 2) = (x - 1)(x + 1)(x + 2)$.

6.2.7 The gcd does not depend on the field

A polynomial with rational coefficients is also a polynomial with real and with complex coefficients. Its gcd with another rational polynomial can be computed in $\mathbb{Q}[x]$, $\mathbb{R}[x]$, or $\mathbb{C}[x]$, and the definition does not obviously give the same answer, because the coefficients a, b in $\langle f, g \rangle$ range over a larger set. The Euclidean algorithm shows the answers agree. Recall that a subfield K of a field L is a subset closed under the field operations (Definition 1.9.1); then $K[x] \subseteq L[x]$, with the same addition and multiplication.

Proposition 6.2.9. The gcd Is the Same Over a Larger Field

Let K be a subfield of a field L, and let $f, g \in K[x]$. Then the greatest common divisor of f and g computed in $K[x]$ equals the one computed in $L[x]$. In particular, f and g are coprime over K if and only if they are coprime over L.

Proof. If $g = 0$, both gcds are the monic associate of f (or 0), by Example (First Greatest Common Divisors) (a), and this does not depend on the field. Let $g \neq 0$, and run the Euclidean algorithm in $K[x]$: $r_{k-1} = q_{k+1} r_k + r_{k+1}$ with $\deg r_{k+1} < \deg r_k$, all in $K[x]$. Each of these equations also holds in $L[x]$, with the same degree condition. By the uniqueness in Theorem 6.1.1 over L, they are exactly the divisions that the algorithm performs in $L[x]$. So both runs produce the same last nonzero remainder r_n, and Theorem 6.2.6 gives the same gcd $c^{-1} r_n$ in both. This proves the proposition. ∎

Warning. The gcd does not see the field, but **factorization does**. Over $\mathbb{Q}$, the polynomial $x^2 - 2$ has no root, by Theorem 0.2.3, so it has no factor $x - c$ with $c \in \mathbb{Q}$ (Theorem 6.1.3). Over $\mathbb{R}$ it factors as $(x - \sqrt{2})(x + \sqrt{2})$. Yet $\gcd(x^2 - 2, x^2 - 2x)$, say, is the same over $\mathbb{Q}$ and over $\mathbb{R}$. So "coprime" is a robust notion, while "cannot be factored" depends on the field; the next section makes the second notion precise.

A useful consequence: if $f, g \in \mathbb{Q}[x]$ are coprime over $\mathbb{Q}$, then they have **no common root in** $\mathbb{C}$. Indeed, $af + bg = 1$ with $a, b \in \mathbb{Q}[x] \subseteq \mathbb{C}[x]$, and evaluating at a common root z would give $0 = 1$ by Theorem 0.8.5. So common complex roots can be detected by a computation that never leaves $\mathbb{Q}$.

6.2.8 Exercises

6.2.8.1 A. Check your understanding

Exercise. A1

(a) Define an ideal of $F[x]$.

(b) State the theorem describing all ideals of $F[x]$.

(c) Determine whether the following statement is true: "for all $f, g \in F[x]$, $\text{span}(f, g) = \langle f, g \rangle$." Justify your answer.

(d) Is $\gcd(2x - 2, 4x^2 - 4) = 2x - 2$ in $\mathbb{Q}[x]$? Justify your answer.

(e) Determine whether the following statement is true: "if $af + bg = x - 1$ for some $a, b \in \mathbb{Q}[x]$, then $\gcd(f, g) = x - 1$." Justify your answer.

(f) Name an algorithm that computes $\gcd(f, g)$ using only division with remainder (Theorem 6.1.1).

Solution.

(a) A subset $I \subseteq F[x]$ is an ideal if it is nonempty, $f + g \in I$ for all $f, g \in I$, and $hf \in I$ for all $f \in I$ and every $h \in F[x]$ (Definition 6.2.1).

(b) If $I \neq \{0\}$ is an ideal of $F[x]$, there is exactly one monic d with $I = \langle d \rangle$, namely the monic polynomial of least degree in I (Theorem 6.2.2).

(c) False. Take $f = g = x$. Then $\text{span}(x, x) = \{cx : c \in F\}$ does not contain x^2, while $x^2 = x \cdot x \in \langle x, x \rangle$. The span allows only scalar coefficients.

(d) No. The gcd is monic by definition. Since $4x^2 - 4 = (2x - 2)(2x + 2)$, we have $\langle 2x - 2, 4x^2 - 4 \rangle = \langle 2x - 2 \rangle = \langle x - 1 \rangle$, so the gcd is $x - 1$.

(e) False. Take $f = x$ and $g = x - 1$, with $a = 0$, $b = 1$. Then $af + bg = x - 1$, but $\gcd(x, x - 1) = 1$, because $x - (x - 1) = 1$ (Proposition 6.2.7). What is true is that $\gcd(f, g)$ divides every combination $af + bg$, by Theorem 6.2.3 (a) and Proposition 6.1.2 (b).

(f) The Euclidean algorithm (Theorem 6.2.6).

6.2.8.2 B. Practice

Exercise. B1: gcd and Bézout Coefficients of Two Cubics

In $\mathbb{Q}[x]$, let $f = x^3 + x^2 + x + 1$ and $g = x^3 + x + 2$. Use the Euclidean algorithm to find $d = \gcd(f, g)$ and polynomials $a, b \in \mathbb{Q}[x]$ with $d = af + bg$. Hence find all complex numbers that are roots of both f and g.

Solution. Divide f by g: $f = 1 \cdot g + r_1$ with $r_1 = f - g = x^2 - 1$. Divide g by r_1: $x \cdot r_1 = x^3 - x$, and $g - x r_1 = 2x + 2$, so $g = x\, r_1 + r_2$ with $r_2 = 2x + 2$. Divide r_1 by r_2: $x^2 - 1 = \frac{1}{2}(x - 1)(2x + 2) + 0$. The last nonzero remainder is $r_2 = 2x + 2$, with leading coefficient 2, so $d = x + 1$ by Theorem 6.2.6.

Back-substituting, $r_2 = g - xr_1 = g - x(f-g) = -xf + (1+x)g$. Dividing by 2,

$$x + 1 = \left(-\frac{x}{2}\right) f + \frac{x+1}{2}\, g.$$

Check: $-\frac{x}{2}(x^3+x^2+x+1) + \frac{x+1}{2}(x^3+x+2)$ equals

$$\tfrac{1}{2}(-x^4 - x^3 - x^2 - x + x^4 + x^3 + x^2 + 3x + 2) = x + 1.$$

Hence: let $z \in \mathbb{C}$. If $f(z) = g(z) = 0$, then evaluating the identity above in $\mathbb{C}$ (Theorem 0.8.5) gives $z + 1 = 0$. Conversely, $x + 1$ divides f and g (Theorem 6.2.3 (a)), so -1 is a root of both. Therefore -1 is the only common complex root.

Exercise. B2: Ideal or Not?

Determine which of the following subsets of $\mathbb{R}[x]$ are ideals. For each ideal, find its monic generator. Justify your answers.

(a) $S_1 = \{f \in \mathbb{R}[x] : f(1) = f(-1)\}$.

(b) $S_2 = \{f \in \mathbb{R}[x] : f(0) = 0 \text{ and } f(1) = 0\}$.

(c) $S_3 = \{f \in \mathbb{R}[x] : \deg f \geq 3\} \cup \{0\}$.

(d) $S_4 = \{f \in \mathbb{R}[x] : f(i) = 0\}$, where $f(i)$ means the value at $i \in \mathbb{C}$ of f regarded as a complex polynomial.

Solution.

(a) Not an ideal. S_1 is a subspace, and $1 \in S_1$. But $x \cdot 1 = x$ has $x(1) = 1 \neq -1 = x(-1)$, so $x \notin S_1$, and (I3) fails.

(b) An ideal. The argument of Example (First Ideals) (b), applied at both points, gives (I1)–(I3). If $f \in S_2$, then 0 and 1 are distinct roots of f, so $x(x-1) \mid f$ by Corollary 6.1.4. Conversely every multiple of $x^2 - x$ vanishes at 0 and 1. Hence $S_2 = \langle x^2 - x \rangle$, with monic generator $x^2 - x$.

(c) Not an ideal. $x^3 \in S_3$ and $-x^3 + x \in S_3$, but their sum x has degree 1 and is nonzero, so (I2) fails.

(d) An ideal: evaluation at i respects sums and products (Theorem 0.8.5 over $\mathbb{C}$), so the argument of Example (First Ideals) (b) applies. It contains $x^2 + 1$. A nonzero element of degree at most 1 would be $a + bx$ with $a, b \in \mathbb{R}$ not both zero and $a + bi = 0$, which is impossible since the real and imaginary parts give $a = b = 0$. So the least degree of a nonzero element is 2. If d is the monic generator (Theorem 6.2.2), then $\deg d = 2$ and $d \mid x^2 + 1$, so $x^2 + 1 = ud$ with u constant, and comparing leading coefficients gives $u = 1$. Hence $S_4 = \langle x^2 + 1 \rangle$.

Exercise. B3: The gcd of $x^m - 1$ and $x^n - 1$

Let F be a field. For integers $m, n \geq 0$, not both zero, let $\gcd(m, n)$ denote the largest positive integer dividing both.

(a) For $m \geq n \geq 1$, verify that $x^m - 1 = x^{m-n}(x^n - 1) + (x^{m-n} - 1)$.

(b) Prove that $\gcd(x^m - 1, x^n - 1) = x^{\gcd(m,n)} - 1$ in $F[x]$ for all $m, n \geq 0$ not both zero, where $x^0 - 1 = 0$.

(c) Hence compute $\gcd(x^{12} - 1, x^{18} - 1)$.

Hint for (b): Theorem 6.2.6, run on the exponents.

Solution.

(a) Expanding, $x^{m-n}(x^n - 1) + x^{m-n} - 1 = x^m - x^{m-n} + x^{m-n} - 1 = x^m - 1$.

(b) We use strong induction on $s = m + n \geq 1$ (Theorem 0.2.10). If $n = 0$, then $m \geq 1$, $\gcd(m, 0) = m$, and $\gcd(x^m - 1, 0)$ is the monic associate of $x^m - 1$, which is $x^m - 1$ itself (Example (First Greatest Common Divisors) (a)). The case $m = 0$ is the same, since both gcds are symmetric. Now let $m, n \geq 1$; by symmetry we may assume $m \geq n$. By (a) and Lemma 6.2.5,

$$\gcd(x^m - 1, x^n - 1) = \gcd(x^n - 1, x^{m-n} - 1).$$

The pair $(n, m - n)$ is not both zero and has sum $m < m + n$, so by the induction hypothesis the right side is $x^{\gcd(n,m-n)} - 1$. Finally, $\gcd(n, m-n) = \gcd(m, n)$: if $k \mid m$ and $k \mid n$, then $k \mid m - n$; and if $k \mid n$ and $k \mid m - n$, then $k \mid (m - n) + n = m$. So the two pairs have the same positive common divisors, and the same largest one. This completes the induction.

(c) By (b), $\gcd(x^{12} - 1, x^{18} - 1) = x^{\gcd(12,18)} - 1 = x^6 - 1$.

6.2.8.3 C. Going deeper

Exercise. C1: Polynomials Vanishing on a Set

Let F be a field and $S \subseteq F$ nonempty. Put $I(S) = \{f \in F[x] : f(s) = 0 \text{ for all } s \in S\}$.

(a) Prove that $I(S)$ is an ideal of $F[x]$.

(b) Suppose $S = \{s_1, \ldots, s_k\}$ is finite, with the s_i distinct. Prove that $I(S) = \langle (x - s_1) \cdots (x - s_k) \rangle$.

(c) Suppose S is infinite. Prove that $I(S) = \{0\}$.

(d) Deduce from (b) and Exercise (C3: Equal Functions Over $\mathbb{F}_p$) that $I(\mathbb{F}_p) = \langle x^p - x \rangle$ for every prime p.

Solution.

(a) The zero polynomial vanishes everywhere, so $I(S) \neq \varnothing$. If $f, g \in I(S)$ and $h \in F[x]$, then for every $s \in S$, Theorem 0.8.5 gives $(f+g)(s) = 0+0 = 0$ and $(hf)(s) = h(s) \cdot 0 = 0$. So (I2) and (I3) hold.

(b) Let $P = (x - s_1) \cdots (x - s_k)$. Every multiple hP vanishes at each s_i, since $P(s_i) = 0$; so $\langle P \rangle \subseteq I(S)$. Conversely, if $f \in I(S)$, then $s_1, \ldots, s_k$ are distinct roots of f, and Corollary 6.1.4 gives $P \mid f$. Hence $I(S) = \langle P \rangle$, and since P is monic, it is the monic generator.

(c) Let $f \in I(S)$, and suppose $f \neq 0$. Since S is infinite, it contains $\deg f + 1$ distinct

elements, all roots of f. This contradicts Corollary 6.1.4. Hence $f = 0$.

(d) By (b), $I(\mathbb{F}_p) = \langle P \rangle$ with $P = \prod_{c \in \mathbb{F}_p}(x - c)$. The solution of Exercise (C3: Equal Functions Over $\mathbb{F}_p$) (a) showed $P = x^p - x$. Hence $I(\mathbb{F}_p) = \langle x^p - x \rangle$.

Exercise. C2: A Non-Principal Ideal in Two Variables

A **polynomial in two variables** over F is a finite sum $P = \sum_{i,j} a_{ij} x^i y^j$ with $a_{ij} \in F$, only finitely many nonzero; two are equal when all coefficients agree. They are added coefficient by coefficient and multiplied by expanding and collecting, and the set $F[x, y]$ of all of them satisfies the laws of Theorem 0.8.1, which you may use. Ideals of $F[x, y]$ and $\langle P_1, \ldots, P_k \rangle$ are defined exactly as in Definition 6.2.1, and Proposition 6.2.1 holds with the same proof. For $P \neq 0$, let $\deg_y P$ be the largest j with $a_{ij} \neq 0$ for some i, and define $\deg_x P$ similarly.

(a) Prove that $\deg_y(PQ) = \deg_y P + \deg_y Q$ for nonzero $P, Q \in F[x, y]$. By symmetry the same holds for $\deg_x$.

(b) Show that $\langle x, y \rangle$ consists of the polynomials whose constant coefficient a_{00} is 0.

(c) Prove that $\langle x, y \rangle \neq \langle D \rangle$ for every $D \in F[x, y]$.

(d) Explain precisely which step of the proof of Theorem 6.2.2 has no analogue in $F[x, y]$.

Hint for (a): write $P = \sum_j p_j(x) y^j$ with $p_j \in F[x]$, and use Corollary 0.8.4.

Solution.

(a) Group terms by the power of y: $P = \sum_{j=0}^{s} p_j y^j$ and $Q = \sum_{j=0}^{t} q_j y^j$ with $p_j, q_j \in F[x]$, $s = \deg_y P$, $t = \deg_y Q$, so $p_s \neq 0$ and $q_t \neq 0$. Expanding, the terms of PQ with y-power greater than $s+t$ are all zero, and the terms with y^{s+t} come only from $p_s y^s \cdot q_t y^t$, giving $(p_s q_t) y^{s+t}$. By Corollary 0.8.4, $p_s q_t \neq 0$ in $F[x]$, so some coefficient of $x^i y^{s+t}$ in PQ is nonzero. Hence $\deg_y(PQ) = s + t$.

(b) Every $Ax + By$ has constant coefficient 0, since each term of Ax and By involves x or y to a positive power. Conversely, if $a_{00} = 0$, then every term $a_{ij} x^i y^j$ with $i \geq 1$ is a multiple of x, and every term with $i = 0$, $j \geq 1$ is a multiple of y. Grouping them, $P = Ax + By$ for some $A, B \in F[x, y]$.

(c) Suppose $\langle x, y \rangle = \langle D \rangle$. Since $x \in \langle D \rangle$, $x = HD$ for some H, so $D \neq 0$ and $H \neq 0$. By (a), $0 = \deg_y x = \deg_y H + \deg_y D$, so $\deg_y D = 0$: D involves no y. Similarly $y = KD$ gives $\deg_x D = 0$. So D is a nonzero constant. But $D \in \langle x, y \rangle$, so its constant coefficient is 0 by (b), that is, $D = 0$. This contradiction shows that $\langle x, y \rangle$ is not principal.

(d) The proof of Theorem 6.2.2 picks a nonzero $d \in I$ of least degree, writes each $f \in I$ as $f = qd + r$ with r smaller than d by Theorem 6.1.1, notes $r = f - qd \in I$, and concludes $r = 0$ from minimality. Measure size in $F[x, y]$ by the **total degree**, the largest $i + j$ with $a_{ij} \neq 0$. Choosing d still works: by (b), every nonzero element of $\langle x, y \rangle$ has total degree at least 1, and those of total degree 1 are the $ax + by$ with $(a, b) \neq (0, 0)$. Let d be any of them. Then $\langle d \rangle \subseteq \langle x, y \rangle$ by (I3), and the inclusion is strict by (c), so some $f \in \langle x, y \rangle$ is not a multiple of d. For this f and **every** $Q \in F[x, y]$, the element $R = f - Qd$ lies in $\langle x, y \rangle$ and is nonzero, so its total degree is at least 1, the total degree of d. So no remainder smaller than d exists: the step with no analogue is **division with remainder**, the appeal to Theorem 6.1.1. Concretely, for $d = x$ and $f = y$, every term of Qx contains x, so the term y survives in $y - Qx$. The underlying reason is that,

viewed as a polynomial in y with coefficients in $F[x]$, the leading coefficient x of d is not invertible in $F[x]$, which is exactly the obstacle of the warning after Theorem 6.1.1.

Exercise. C3: gcd Times lcm

Let F be a field and $f, g \in F[x]$ **monic**. Put $d = \gcd(f, g)$ and write $f = df_1$, $g = dg_1$ with $f_1, g_1 \in F[x]$.

(a) Prove that $\gcd(f_1, g_1) = 1$.

(b) Prove that $m = df_1g_1$ is a monic common multiple of f and g that divides every common multiple of f and g.

(c) Deduce that $\operatorname{lcm}(f, g) = df_1g_1$, and hence $\gcd(f, g) \cdot \operatorname{lcm}(f, g) = fg$.

(d) Show by an example in $\mathbb{Q}[x]$ that the formula in (c) fails if f and g are not monic.

Hint for (b): use Theorem 6.2.8.

Solution.

(a) The quotients exist by Theorem 6.2.3 (a), and $d \neq 0$ since $f \neq 0$. By Corollary 6.2.4, $d = af + bg = d(af_1 + bg_1)$. Canceling $d \neq 0$ (Corollary 0.8.4 (2)) gives $af_1 + bg_1 = 1$, so $\gcd(f_1, g_1) = 1$ by Proposition 6.2.7.

(b) Comparing leading coefficients in $f = df_1$ with f, d monic shows f_1 is monic; likewise g_1. So m is monic, and $m = fg_1 = gf_1$ is a common multiple. Let M be any common multiple, say $M = fu$. Since $g \mid M$, write $M = gv$. Then $df_1u = dg_1v$, and canceling d gives $f_1u = g_1v$, so $g_1 \mid f_1u$. By (a) and Theorem 6.2.8, $g_1 \mid u$, say $u = g_1w$. Hence $M = fg_1w = mw$, and $m \mid M$.

(c) By (b), $m \in \langle f \rangle \cap \langle g \rangle$, and every element of this ideal is a multiple of m. So $\langle f \rangle \cap \langle g \rangle = \langle m \rangle$, and m is monic, hence $\operatorname{lcm}(f, g) = m$ by Definition 6.2.4 and Theorem 6.2.2. Then $\gcd(f, g) \cdot \operatorname{lcm}(f, g) = d \cdot df_1g_1 = (df_1)(dg_1) = fg$.

(d) Take $f = g = 2x$. Then $\gcd(f, g) = x$ and $\operatorname{lcm}(f, g) = x$, both monic, so their product is x^2, while $fg = 4x^2$. In general the two sides differ by the product of the leading coefficients.

6.3 Irreducible Polynomials and Unique Factorization

Every integer $n \geq 2$ is a product of primes, and Chapter 0 remarked that the convention "1 is not prime" is what makes such factorizations unique. Polynomials over a field behave the same way, with irreducible polynomials in the role of primes and nonzero constants in the role of ± 1. This section defines irreducible polynomials, proves Euclid's lemma from the gcd machinery of the previous section, and proves that every nonconstant polynomial factors into irreducibles in exactly one way. The catch is that "irreducible" depends on the field: the same polynomial may factor over $\mathbb{R}$ but not over $\mathbb{Q}$, or over $\mathbb{F}_5$ but not over $\mathbb{F}_3$. So we also collect tools for deciding irreducibility over specific fields: roots in low degree, the Rational Root Theorem, and Eisenstein's criterion. Throughout, F is a field.

6.3.1 Irreducible polynomials

For integers, the building blocks are the primes: integers $p \geq 2$ whose only positive divisors are 1 and p (Definition 0.2.2). To transplant this to $F[x]$, recall from the first section that every nonzero constant divides every polynomial, and that multiplying by a nonzero constant changes nothing for divisibility. So the trivial divisors of p are the nonzero constants and the associates cp. A building block should have no others.

An irreducible polynomial is a nonconstant polynomial that cannot be split into two factors of smaller degree.

Definition 6.3.1. Irreducible Polynomial

Let F be a field. A polynomial $p \in F[x]$ is **irreducible over** F if

(1) p is **nonconstant**, that is, $\deg p \geq 1$; and
(2) whenever $p = gh$ with $g, h \in F[x]$, **one of** g, h is a constant.

A nonconstant polynomial that is not irreducible is **reducible over** F.

In words: condition (1) excludes 0 and the nonzero constants. Condition (2) says every factorization of p is trivial: one factor is a constant, which is nonzero because $p \neq 0$, and the other is then an associate of p. The phrase **over** F is part of the notion, since g and h must lie in $F[x]$.

Since degrees add (Theorem 0.8.3), a factorization $p = gh$ with $p \neq 0$ has $\deg g + \deg h = \deg p$. So g is constant exactly when $\deg h = \deg p$. This gives a working form of the definition: **a nonconstant p is reducible if and only if $p = gh$ with $1 \leq \deg g < \deg p$ and $1 \leq \deg h < \deg p$.**

Example. First Irreducible and Reducible Polynomials

(a) Show that every polynomial of degree 1 is irreducible over every field.
(b) Show that the constants 0 and 2 are not irreducible over $\mathbb{Q}$.
(c) Show that $x^2 + x$ is reducible over every field, and that x^3 is reducible over every field.

Solution.

(a) If $\deg p = 1$ and $p = gh$, then $\deg g + \deg h = 1$ with both degrees in $\mathbb{N}$, so one of them is 0: one factor is a constant. Condition (1) holds since $\deg p = 1$. Hence p is irreducible, whatever the field. For example x, $2x + 4$, and $x - \sqrt{2}$ (over $\mathbb{R}$) are irreducible.
(b) Both are constants, so condition (1) fails. They are the degenerate cases, and they matter: 2 divides everything in $\mathbb{Q}[x]$, like 1 among the integers, and it is not a building block.
(c) $x^2 + x = x(x + 1)$ with both factors of degree 1, and $x^3 = x \cdot x^2$ with factors of degrees 1 and 2. In each case both factors are nonconstant, so condition (2) fails.

Why insist on "nonconstant"? For the same reason 1 is not a prime. If nonzero constants were allowed as building blocks, factorizations could never be unique: $x = 2 \cdot \frac{1}{2}x = 2 \cdot 3 \cdot \frac{1}{6}x = \cdots$ over $\mathbb{Q}$, with any number of constant factors. The factorization theorem below pulls out one

constant in front and uses only nonconstant, monic irreducibles.

For polynomials of degree 2 and 3, irreducibility is decided by roots alone.

Theorem 6.3.1. Irreducibility in Degree 2 and 3

Let F be a field and $p \in F[x]$.

(a) p has a factor of degree 1 in $F[x]$ if and only if p has a root in F. In particular, if $\deg p \geq 2$ and p has a root in F, then p is reducible over F.

(b) If $\deg p \in \{2, 3\}$, then p is irreducible over F if and only if p has **no** root in F.

Proof.

(a) ($\Leftarrow$) If $c \in F$ is a root, then $p = (x - c)q$ by Theorem 6.1.3 (b). ($\Rightarrow$) If $p = (ax + b)q$ with $a \neq 0$, then $c = -a^{-1}b$ satisfies $ac + b = 0$, so $p(c) = 0 \cdot q(c) = 0$ by Theorem 0.8.5. For the last sentence, if $\deg p \geq 2$ and $p = (x - c)q$, then $\deg q = \deg p - 1 \geq 1$, so both factors are nonconstant of degree less than $\deg p$, and p is reducible.

(b) We prove both directions by contraposition. ($\Rightarrow$) If p has a root, it is reducible by (a), since $\deg p \geq 2$. ($\Leftarrow$) Suppose p is reducible. By the working form of the definition, $p = gh$ with $\deg g, \deg h \geq 1$ and $\deg g + \deg h = \deg p \in \{2, 3\}$. Two integers that are at least 1 and sum to 2 or 3 cannot both be at least 2, so one of g, h has degree 1. By (a), p has a root in F. This proves the theorem.

■

Now the standard examples, each checked with the theorem.

- $x^2 + 1$ **over** $\mathbb{R}$ **and over** $\mathbb{C}$**.** For real c, $c^2 + 1 \geq 1 > 0$, so there is no real root, and $x^2 + 1$ is irreducible over $\mathbb{R}$ (and over $\mathbb{Q}$). Over $\mathbb{C}$, $x^2 + 1 = (x - i)(x + i)$ is reducible.
- $x^2 - 2$ **over** $\mathbb{Q}$ **and over** $\mathbb{R}$**.** A rational root would be a rational number with square 2, which does not exist (Theorem 0.2.3). So $x^2 - 2$ is irreducible over $\mathbb{Q}$. Over $\mathbb{R}$, $x^2 - 2 = (x - \sqrt{2})(x + \sqrt{2})$.
- $x^2 + x + 1$ **over** $\mathbb{F}_2$**.** The values at 0 and 1 are 1 and $1 + 1 + 1 = 1$. No root, so it is irreducible over $\mathbb{F}_2$. It is the only irreducible polynomial of degree 2 over $\mathbb{F}_2$ (Exercise B3).

The nonexample by minimal change is $x^2 + x$ over $\mathbb{F}_2$: change the constant term of $x^2 + x + 1$ from 1 to 0, and 0 becomes a root, so $x^2 + x = x(x + 1)$ is reducible. What fails is condition (2).

Warning. "No roots" implies "irreducible" **only in degrees 2 and 3**. The polynomial $(x^2 + 1)^2 = x^4 + 2x^2 + 1$ has no real root, since its values are at least 1, but it is reducible over $\mathbb{R}$: it is a product of two factors of degree 2. In degree 4 and higher, a polynomial can split into nonlinear factors without having any root.

Quick check. Over which of the fields $\mathbb{Q}$, $\mathbb{Q}(\sqrt{2})$ (Example ($\mathbb{Q}(\sqrt{2})$ Is a Field)), $\mathbb{R}$ is $x^2 - 2$ irreducible? Is $x^2 + 1$ irreducible over $\mathbb{F}_5$?

Solution. Over $\mathbb{Q}$ only. The field $\mathbb{Q}(\sqrt{2})$ contains $\sqrt{2}$, a root of x^2-2, so $x^2-2=(x-\sqrt{2})(x+\sqrt{2})$ is reducible there, and likewise over $\mathbb{R}$; over $\mathbb{Q}$ it is irreducible, as shown above. Over $\mathbb{F}_5$, $2^2+1=5=0$, so 2 is a root and x^2+1 is reducible by Theorem 6.3.1: indeed $x^2+1=(x-2)(x+2)$ in $\mathbb{F}_5[x]$.

6.3.2 Euclid's lemma

What makes primes useful is not their definition but a property the definition implies: if a prime divides a product, it divides one of the factors. For polynomials this follows quickly from the previous section. The first step is that an irreducible polynomial has only two possible gcds with anything.

Lemma 6.3.2. The gcd with an Irreducible Polynomial

Let F be a field, $p\in F[x]$ irreducible over F, and $f\in F[x]$. Then either $p\mid f$ or $\gcd(p,f)=1$.

Proof. Let $d=\gcd(p,f)$. Since $p\neq 0$, d is monic, and $d\mid p$ by Theorem 6.2.3 (a), say $p=du$. Since p is irreducible, d or u is a constant. If d is a constant, then $d=1$, since d is monic. If u is a constant, it is nonzero, and $p=ud$ divides every multiple of d; since $d\mid f$ by Theorem 6.2.3 (a), Proposition 6.1.2 (a) gives $p\mid f$. This proves the lemma. ∎

Theorem 6.3.3. Euclid's Lemma for Polynomials

Let F be a field and $p\in F[x]$ irreducible over F. If $q,r\in F[x]$ and $p\mid qr$, then $p\mid q$ or $p\mid r$.

Proof. Suppose $p\nmid q$. By Lemma 6.3.2, $\gcd(p,q)=1$. Since $p\mid qr$, Theorem 6.2.8 gives $p\mid r$. This proves the theorem. ∎

The hypothesis "irreducible" is essential. The polynomial x^2 divides $x\cdot x$ but divides neither factor; the lemma does not apply because $x^2=x\cdot x$ is reducible. Over $\mathbb{F}_2$, $x^2+1=(x+1)^2$ divides $(x+1)(x+1)$ but not $x+1$, for the same reason.

For unique factorization we need the lemma for products of any length, and we need to know what "divides" means between two monic irreducibles.

Corollary 6.3.4. Euclid's Lemma for Several Factors

Let F be a field and $p\in F[x]$ irreducible over F.

(a) If $q_1,\dots,q_k\in F[x]$ with $k\geq 1$ and $p\mid q_1q_2\cdots q_k$, then $p\mid q_i$ for some i.

(b) If q is irreducible over F, $p\mid q$, and p,q are both monic, then $p=q$.

Proof.

(a) We use induction on k. For $k = 1$ there is nothing to prove. Let $k \geq 2$ and assume the statement for $k-1$ factors. Since $p \mid (q_1 \cdots q_{k-1})\, q_k$, Theorem 6.3.3 gives $p \mid q_1 \cdots q_{k-1}$ or $p \mid q_k$. In the first case the induction hypothesis gives $p \mid q_i$ for some $i \leq k-1$. Either way the statement holds for k.

(b) Write $q = pu$. Since q is irreducible, p or u is constant. But p is not constant, being irreducible, so u is a nonzero constant. Comparing leading coefficients of the monic polynomials q and p gives $u = 1$, so $p = q$.

■

6.3.3 Unique factorization

Here is the theorem that Chapter 0's remark about primes was pointing towards, now for polynomials. The constant in front is forced to be the leading coefficient; the rest is a product of monic irreducibles, unique up to order.

Theorem 6.3.5. Unique Factorization in $F[x]$

Let F be a field and $f \in F[x]$ nonconstant.

(a) **Existence.** There are a nonzero constant $c \in F$, an integer $k \geq 1$, and monic polynomials $p_1, \ldots, p_k \in F[x]$, each irreducible over F, with

$$f = c\, p_1 p_2 \cdots p_k.$$

(b) **Uniqueness.** If also $f = c'\, q_1 q_2 \cdots q_l$ with $c' \in F$ nonzero, $l \geq 1$, and $q_1, \ldots, q_l$ monic and irreducible over F, then $c = c'$, $k = l$, and after renumbering the q_j we have $p_i = q_i$ for every i.

In both factorizations, the constant is the leading coefficient of f.

Idea. **Existence:** if f is irreducible, stop; otherwise split it into two factors of smaller degree and factor each of them, which is strong induction on the degree. **Uniqueness:** compare leading coefficients to get $c = c'$. Then p_k divides $q_1 \cdots q_l$, so by Euclid's lemma it equals some q_j. Cancel it from both sides; the two shorter factorizations are equal by induction on the number of factors k. The one thing to watch is that after canceling, the right-hand side must still have at least one factor.

Proof. First a remark used in both parts. A product of monic polynomials is monic, by Theorem 0.8.3 and induction on the number of factors. So if $f = c\, p_1 \cdots p_k$ with the p_i monic, then the leading coefficient of f is $c \cdot 1 = c$. This proves the final sentence of the theorem.

(a) We prove by strong induction on $n \geq 1$ (Theorem 0.2.10 with $n_0 = 1$) that every f of degree n has a factorization as in (a). Let $\deg f = n$, assume the statement for all degrees $1 \leq m < n$, and let c be the leading coefficient of f.
Case 1: f is irreducible over F. Then $f = c\,(c^{-1} f)$. The polynomial $c^{-1} f$ is monic,

and it is irreducible: it is nonconstant, and if $c^{-1}f = gh$, then $f = (cg)h$, so cg or h is constant, hence g or h is constant. So $k = 1$, $p_1 = c^{-1}f$ works.
Case 2: f is reducible over F. By the working form of the definition, $f = gh$ with $1 \le \deg g < n$ and $1 \le \deg h < n$. By the induction hypothesis, $g = a\,p_1 \cdots p_s$ and $h = b\,p_{s+1} \cdots p_{s+t}$ with a, b nonzero constants, $s, t \ge 1$, and all p_i monic irreducible. Then $f = (ab)\,p_1 \cdots p_{s+t}$, with $ab \neq 0$ by Theorem 0.7.1. This is a factorization as in (a), with $k = s + t$.

(b) By the first remark, c and c' are both the leading coefficient of f, so $c = c'$. Multiplying by c^{-1}, we get $p_1 \cdots p_k = q_1 \cdots q_l$. So it suffices to prove, by induction on $k \ge 1$, the statement

$U(k)$: whenever $p_1 \cdots p_k = q_1 \cdots q_l$ with $l \ge 1$ and all p_i, q_j monic and irreducible over F, we have $k = l$, and after renumbering the q_j, $p_i = q_i$ for every i.

We first note: if a monic irreducible q equals a product $r_1 \cdots r_m$ of $m \ge 2$ monic irreducibles, then $q = r_1 \cdot (r_2 \cdots r_m)$ with both factors nonconstant, since each r_i has degree at least 1; this contradicts the irreducibility of q. So **a monic irreducible is never a product of two or more monic irreducibles.** $(*)$

Base case $k = 1$. Then $p_1 = q_1 \cdots q_l$. By $(*)$, $l = 1$, and so $p_1 = q_1$.

Inductive step. Let $k \ge 2$ and assume $U(k-1)$. Suppose $p_1 \cdots p_k = q_1 \cdots q_l$. If $l = 1$, then $q_1 = p_1 \cdots p_k$ with $k \ge 2$, which is impossible by $(*)$. So $l \ge 2$. Now $p_k \mid q_1 \cdots q_l$, so by Corollary 6.3.4 (a), $p_k \mid q_j$ for some j, and by Corollary 6.3.4 (b), $p_k = q_j$. Renumber the q's so that $j = l$. Then

$$(p_1 \cdots p_{k-1})\,p_k = (q_1 \cdots q_{l-1})\,p_k,$$

and since $p_k \neq 0$, Corollary 0.8.4 (2) lets us cancel it: $p_1 \cdots p_{k-1} = q_1 \cdots q_{l-1}$, with $k-1 \ge 1$ and $l-1 \ge 1$ factors. By $U(k-1)$, $k-1 = l-1$, and after renumbering $q_1, \ldots, q_{l-1}$, $p_i = q_i$ for $i \le k-1$. Together with $p_k = q_k$, this proves $U(k)$.

By induction, $U(k)$ holds for every $k \ge 1$, which proves (b).

■

Grouping equal factors, every nonconstant $f \in F[x]$ can be written uniquely as

$$f = c\,p_1^{e_1} p_2^{e_2} \cdots p_s^{e_s}$$

with c the leading coefficient, $p_1, \ldots, p_s$ **distinct** monic irreducibles over F, and exponents $e_i \ge 1$. The monic irreducible factors of the characteristic and minimal polynomials of an operator will organize the structure theory of Chapters 9 and 10.

Warning. Uniqueness needs the normalization "monic". Without it, constants can move between factors: over $\mathbb{Q}$, $2x^2 - 2 = (2x-2)(x+1) = (x-1)(2x+2) = (\frac{1}{2}x - \frac{1}{2})(4x+4)$. All of these are "factorizations into irreducibles", and they differ only by associates. The theorem removes this freedom by collecting all constants into c. Uniqueness also refers to a **fixed** field: over $\mathbb{Q}$ the factorization of $x^2 - 2$ is $x^2 - 2$ itself, and over $\mathbb{R}$ it is $(x - \sqrt{2})(x + \sqrt{2})$.

The next example shows the field dependence in one table.

Example. Factoring $x^4 - 1$ Over Six Fields

Factor $x^4 - 1$ into a constant times monic irreducibles over $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$, $\mathbb{F}_2$, $\mathbb{F}_3$, and, for contrast, $\mathbb{F}_5$.

Solution. Over every field, $x^4 - 1 = (x^2 - 1)(x^2 + 1) = (x - 1)(x + 1)(x^2 + 1)$, by expanding. The factors $x - 1$ and $x + 1$ are irreducible by Example (First Irreducible and Reducible Polynomials) (a). Everything depends on $x^2 + 1$, which by Theorem 6.3.1 is irreducible exactly when it has no root in the field. The constant in front is always the leading coefficient 1.

- $\mathbb{Q}$ **and** $\mathbb{R}$**.** For real c, $c^2 + 1 \geq 1$, so there is no root, and $x^2 + 1$ is irreducible.
- $\mathbb{C}$**.** $x^2 + 1 = (x - i)(x + i)$.
- $\mathbb{F}_2$**.** Here $-1 = 1$, so $x - 1 = x + 1$, and $x^2 + 1 = (x + 1)^2$ because $2x = 0$. All four factors coincide.
- $\mathbb{F}_3$**.** The values of $x^2 + 1$ at $0, 1, 2$ are $1, 2, 5 = 2$. No root, so $x^2 + 1$ is irreducible. Also $x - 1 = x + 2$.
- $\mathbb{F}_5$**.** $2^2 + 1 = 5 = 0$ and $3^2 + 1 = 10 = 0$, so $x^2 + 1 = (x - 2)(x - 3) = (x + 3)(x + 2)$.

field	factorization of $x^4 - 1$ into monic irreducibles
$\mathbb{Q}$	$(x - 1)(x + 1)(x^2 + 1)$
$\mathbb{R}$	$(x - 1)(x + 1)(x^2 + 1)$
$\mathbb{C}$	$(x - 1)(x + 1)(x - i)(x + i)$
$\mathbb{F}_2$	$(x + 1)^4$
$\mathbb{F}_3$	$(x + 1)(x + 2)(x^2 + 1)$
$\mathbb{F}_5$	$(x + 1)(x + 2)(x + 3)(x + 4)$

By Theorem 6.3.5, each row is **the** factorization over that field, up to order. Notice that the numbers of irreducible factors are $3, 3, 4, 4, 3, 4$, and that over $\mathbb{F}_2$ a repeated factor appears although $x^4 - 1$ has no repeated factor over the other fields.

6.3.4 Tools over $\mathbb{Q}$: rational roots and Eisenstein's criterion

Over $\mathbb{Q}$, Theorem 6.3.1 reduces irreducibility of a quadratic or cubic to the question of rational roots. There are infinitely many rational numbers to try, but for integer coefficients only finitely many candidates can work. We write $\mathbb{Z}[x]$ for the polynomials in $\mathbb{Q}[x]$ whose coefficients are all integers, as in the first section. Every rational number can be written as u/v with $u \in \mathbb{Z}$, $v \geq 1$, and the only positive integer dividing both u and v equal to 1: take $v \geq 1$ as small as possible, since a common divisor $e > 1$ would give the representation $(u/e)/(v/e)$ with a smaller denominator. We say u/v is **in lowest terms**.

Theorem 6.3.6. Rational Root Theorem

Let $f = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_0 \in \mathbb{Z}[x]$ with $n \geq 1$ and $a_n \neq 0$. If $u/v \in \mathbb{Q}$ is in lowest terms and $f(u/v) = 0$, then $u \mid a_0$ and $v \mid a_n$.

Idea. Clear denominators: $v^n f(u/v) = a_n u^n + a_{n-1}u^{n-1}v + \cdots + a_0 v^n = 0$. Every term except the last contains u, so u divides $a_0 v^n$; every term except the first contains v, so v divides $a_n u^n$. Since u and v share no factor, the powers of v can be stripped off one at a time, using Bézout for integers.

Proof. Multiplying $f(u/v) = 0$ by v^n gives

$$a_n u^n + a_{n-1}u^{n-1}v + \cdots + a_1 u v^{n-1} + a_0 v^n = 0.$$

Hence $a_0 v^n = -u(a_n u^{n-1} + a_{n-1}u^{n-2}v + \cdots + a_1 v^{n-1})$, so $u \mid a_0 v^n$; and $a_n u^n = -v(a_{n-1}u^{n-1} + \cdots + a_0 v^{n-1})$, so $v \mid a_n u^n$.

Since $v \geq 1$ and the only positive common divisor of u and v is 1, Lemma 0.7.2 gives integers s, t with $us + vt = 1$. If $b \in \mathbb{Z}$ and $k \geq 1$ with $u \mid bv^k$, then

$$bv^{k-1} = bv^{k-1}(us + vt) = u\,(bv^{k-1}s) + t\,(bv^k)$$

is a sum of two multiples of u, so $u \mid bv^{k-1}$. Applying this n times to $b = a_0$ gives $u \mid a_0$. Similarly, if $v \mid cu^k$ with $k \geq 1$, then $cu^{k-1} = s\,(cu^k) + v\,(cu^{k-1}t)$, so $v \mid cu^{k-1}$; applying this n times to $c = a_n$ gives $v \mid a_n$. This proves the theorem. ∎

A polynomial in $\mathbb{Q}[x]$ with fractional coefficients can first be multiplied by a common denominator; this does not change its roots or its irreducibility.

Example. A Cubic That Is Irreducible Over $\mathbb{Q}$

Decide whether $f = 2x^3 - 3x^2 + 4x - 1$ is irreducible over $\mathbb{Q}$.

Solution. By Theorem 6.3.6, a rational root u/v in lowest terms has $u \mid -1$ and $v \mid 2$ with $v \geq 1$. So $u = \pm 1$ and $v \in \{1, 2\}$, and the candidates are $1, -1, \frac{1}{2}, -\frac{1}{2}$. Evaluating,

$$\begin{aligned} f(1) &= 2, \qquad f(-1) = -10, \\ f(\tfrac{1}{2}) &= \tfrac{1}{4} - \tfrac{3}{4} + 2 - 1 = \tfrac{1}{2}, \\ f(-\tfrac{1}{2}) &= -\tfrac{1}{4} - \tfrac{3}{4} - 2 - 1 = -4. \end{aligned}$$

None is 0, so f has no rational root. Since $\deg f = 3$, Theorem 6.3.1 shows that f is irreducible over $\mathbb{Q}$.

Beyond degree 3, roots are not enough, as the warning after Theorem 6.3.1 showed. One classical criterion handles many polynomials of any degree at once. It looks at divisibility of the coefficients by a prime.

Theorem 6.3.7. Eisenstein's Criterion

Let $f = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_0 \in \mathbb{Z}[x]$ with $n \geq 1$, and let p be a prime such that

(1) $p \nmid a_n$;
(2) $p \mid a_i$ for every $i = 0, 1, \ldots, n-1$;
(3) $p^2 \nmid a_0$.

Then f is irreducible over $\mathbb{Q}$.

The proof needs one more idea, **Gauss's lemma**: a polynomial with integer coefficients that factors over $\mathbb{Q}$ already factors over $\mathbb{Z}$, with factors of the same degrees. With it, one reduces the coefficients modulo p and compares. Exercise C2 of this section proves Gauss's lemma and then Eisenstein's criterion, step by step; no result of this chapter depends on the criterion. Here we only use it in examples.

Example. Irreducible Polynomials of Every Degree Over $\mathbb{Q}$

(a) Show that $x^n - 2$ is irreducible over $\mathbb{Q}$ for every $n \geq 1$. (b) Show that $x^4 + 10x + 5$ is irreducible over $\mathbb{Q}$.

Solution.

(a) Take $p = 2$. The leading coefficient 1 is not divisible by 2; the coefficients $a_1, \ldots, a_{n-1}$ are 0 and $a_0 = -2$, all divisible by 2; and $4 \nmid -2$. By Theorem 6.3.7, $x^n - 2$ is irreducible over $\mathbb{Q}$. So over $\mathbb{Q}$ there are irreducible polynomials of every degree. (Over $\mathbb{R}$, by contrast, $x^n - 2$ has the root $\sqrt[n]{2}$ and is reducible for $n \geq 2$.)
(b) Take $p = 5$: $5 \nmid 1$; 5 divides $0, 0, 10, 5$; and $25 \nmid 5$. So $x^4 + 10x + 5$ is irreducible over $\mathbb{Q}$. Roots alone could not have shown this, since a quartic may factor into two quadratics.

Eisenstein's criterion is sufficient, not necessary. The polynomial $x^2 + 1$ is irreducible over $\mathbb{Q}$, but no prime divides its constant term 1, so condition (2) fails for every p.

6.3.5 Exercises

6.3.5.1 A. Check your understanding

Exercise. A1

(a) Define what it means for $p \in F[x]$ to be irreducible over F.
(b) State the Unique Factorization Theorem in $F[x]$.
(c) Determine whether the following statement is true: "a polynomial in $\mathbb{R}[x]$ of degree at least 1 with no real root is irreducible over $\mathbb{R}$." Justify your answer.
(d) Is the constant 2 irreducible over $\mathbb{Q}$? Is $2x + 4$?
(e) Is $x^2 + 1$ irreducible over $\mathbb{F}_5$? Justify your answer.
(f) Name a tool that reduces the irreducibility of a cubic in $\mathbb{Z}[x]$ over $\mathbb{Q}$ to finitely many evaluations.

Solution.

(a) p is irreducible over F if $\deg p \geq 1$ and, whenever $p = gh$ with $g, h \in F[x]$, one of g, h is a constant (Definition 6.3.1).

(b) Every nonconstant $f \in F[x]$ can be written as $f = c\,p_1 \cdots p_k$ with $c \in F$ nonzero and $p_1, \ldots, p_k$ monic and irreducible over F; the constant c is the leading coefficient of f, and the p_i are unique up to order (Theorem 6.3.5).

(c) False. $(x^2+1)^2$ has no real root, but it is the product of two nonconstant factors, so it is reducible over $\mathbb{R}$. The statement is true in degrees 2 and 3 only (Theorem 6.3.1).

(d) No: 2 is constant, and irreducible polynomials are nonconstant by definition. Yes: $2x + 4$ has degree 1, so it is irreducible (Example (First Irreducible and Reducible Polynomials) (a)).

(e) No. In $\mathbb{F}_5$, $2^2 + 1 = 0$, so $x^2 + 1 = (x-2)(x+2)$ is reducible by Theorem 6.3.1.

(f) The Rational Root Theorem (Theorem 6.3.6), combined with Theorem 6.3.1.

6.3.5.2 B. Practice

Exercise. B1: Irreducible or Not?

Determine whether each polynomial is irreducible over each of the given fields. Justify your answers.

(a) $x^2 + 2$ over $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{F}_2$, and $\mathbb{F}_3$.

(b) $x^3 + x + 1$ over $\mathbb{F}_2$ and over $\mathbb{F}_3$.

(c) $x^3 - 3x - 1$ over $\mathbb{Q}$.

(d) $x^4 + 4$ over $\mathbb{Q}$.

(e) $x^5 - 6x + 3$ over $\mathbb{Q}$.

Solution.

(a) Over $\mathbb{Q}$ and $\mathbb{R}$: for real c, $c^2 + 2 \geq 2 > 0$, so there is no root, and $x^2 + 2$ is irreducible by Theorem 6.3.1. Over $\mathbb{F}_2$: $2 = 0$, so $x^2 + 2 = x \cdot x$ is reducible. Over $\mathbb{F}_3$: $2 = -1$, so $x^2 + 2 = x^2 - 1 = (x-1)(x+1)$ is reducible.

(b) Over $\mathbb{F}_2$: the values at 0 and 1 are 1 and $3 = 1$, so there is no root, and the cubic is irreducible by Theorem 6.3.1. Over $\mathbb{F}_3$: the value at 1 is $3 = 0$, so it is reducible; indeed $x^3+x+1 = (x-1)(x^2+x+2)$ in $\mathbb{F}_3[x]$, as expanding shows: $x^3+x^2+2x-x^2-x-2 = x^3 + x - 2 = x^3 + x + 1$.

(c) By Theorem 6.3.6, a rational root u/v in lowest terms has $u \mid -1$ and $v \mid 1$, so it is 1 or -1. The values are $1 - 3 - 1 = -3$ and $-1 + 3 - 1 = 1$. No rational root, so the cubic is irreducible over $\mathbb{Q}$ by Theorem 6.3.1.

(d) Reducible, although it has no rational root (its values are at least 4): expanding, $(x^2 + 2x + 2)(x^2 - 2x + 2) = x^4 - 4x^2 + 4 + 4x^2 = x^4 + 4$. This is the warning after Theorem 6.3.1 in action.

(e) Irreducible by Theorem 6.3.7 with $p = 3$: $3 \nmid 1$; 3 divides the coefficients $0, 0, 0, -6, 3$ of $x^4, x^3, x^2, x, 1$; and $9 \nmid 3$.

Exercise. B2: Factoring $x^6 - 1$ Over $\mathbb{Q}$

Factor $x^6 - 1$ into monic irreducible polynomials over $\mathbb{Q}$, and justify that each factor is irreducible. Hence find its factorization into monic irreducibles over $\mathbb{F}_2$ as well. Is it obtained by reducing the coefficients of each rational factor modulo 2? Explain why, in general, this has to be checked rather than assumed.

Solution. Using $x^6 - 1 = (x^3 - 1)(x^3 + 1)$ and Exercise (B2: A Factor of $x^n - c^n$) (with $c = 1$ and $c = -1$),

$$x^6 - 1 = (x - 1)(x^2 + x + 1)(x + 1)(x^2 - x + 1),$$

which one checks by expanding: $(x-1)(x^2+x+1) = x^3 - 1$ and $(x+1)(x^2-x+1) = x^3+1$. The linear factors are irreducible. For the quadratics, $c^2 \pm c + 1 = (c \pm \frac{1}{2})^2 + \frac{3}{4} > 0$ for real c, so they have no rational root and are irreducible by Theorem 6.3.1. By Theorem 6.3.5 this is the factorization over $\mathbb{Q}$.

Over $\mathbb{F}_2$, reducing coefficients gives $(x+1)(x^2+x+1)(x+1)(x^2+x+1)$, since $-1 = 1$. Here $x+1$ is irreducible and x^2+x+1 is irreducible over $\mathbb{F}_2$ (it has no root), so the factorization over $\mathbb{F}_2$ is $(x+1)^2(x^2+x+1)^2$. So yes, in this case the factorization is obtained by reduction, but only because the reduced factors happen to stay irreducible, and that has to be checked: two distinct rational factors, $x - 1$ and $x + 1$, became equal, and in general a factor that is irreducible over $\mathbb{Q}$ may become reducible after reduction, as $x^2 + 1$ does over $\mathbb{F}_2$ in Example (Factoring $x^4 - 1$ Over Six Fields).

Exercise. B3: Small Irreducibles Over $\mathbb{F}_2$

List all irreducible polynomials over $\mathbb{F}_2$ of degree at most 3. Justify that your list is complete.

Hint: a polynomial over $\mathbb{F}_2$ has the root 0 exactly when its constant term is 0, and the root 1 exactly when its number of nonzero terms is even.

Solution. Every nonzero polynomial over $\mathbb{F}_2$ is monic, since its leading coefficient is 1. For $f = \sum a_k x^k$, $f(0) = a_0$ and $f(1) = \sum a_k$, which is 0 exactly when the number of nonzero coefficients is even.

Degree 1: x and $x + 1$, both irreducible (Example (First Irreducible and Reducible Polynomials) (a)).

Degrees 2 and 3: by Theorem 6.3.1, f is irreducible exactly when $f(0) = 1$ and $f(1) = 1$, that is, the constant term is 1 and the number of nonzero terms is odd. In degree 2, the polynomials with constant term 1 are x^2+1 (two terms) and x^2+x+1 (three terms), so only $x^2 + x + 1$ is irreducible. In degree 3, they are $x^3 + 1$ (two terms), $x^3 + x + 1$ and $x^3 + x^2 + 1$ (three terms), and $x^3 + x^2 + x + 1$ (four terms). So the irreducible ones are $x^3 + x + 1$ and $x^3 + x^2 + 1$.

Hence the complete list is x, $x + 1$, $x^2 + x + 1$, $x^3 + x + 1$, $x^3 + x^2 + 1$.

6.3.5.3 C. Going deeper

Exercise. C1: Infinitely Many Irreducibles

Let F be any field.

(a) Prove that there are infinitely many monic irreducible polynomials over F.

(b) Deduce that for every prime p and every $N \in \mathbb{N}$, there is an irreducible polynomial over $\mathbb{F}_p$ of degree greater than N.

(c) Explain why (a) is immediate when F is infinite, and use Exercise (B3: Small Irreducibles Over $\mathbb{F}_2$) to name an irreducible polynomial of degree 3 over $\mathbb{F}_2$.

Hint for (a): imitate the proof that there are infinitely many primes (Theorem 0.2.13).

Solution.

(a) Suppose, for a contradiction, that $p_1, \dots, p_k$ are all the monic irreducible polynomials over F. The list is nonempty, since x is monic and irreducible. Let $f = p_1 p_2 \cdots p_k + 1$. Each p_i has degree at least 1, so $\deg(p_1 \cdots p_k) \geq 1$ by Theorem 0.8.3, and f is nonconstant by Theorem 0.8.2. By Theorem 6.3.5 (a), f has a monic irreducible factor, which must be some p_j. Then $p_j \mid f$ and $p_j \mid p_1 \cdots p_k$, so $p_j \mid f - p_1 \cdots p_k = 1$ by Proposition 6.1.2 (b). By Proposition 6.1.2 (c), $\deg p_j \leq \deg 1 = 0$, contradicting $\deg p_j \geq 1$. Hence there are infinitely many monic irreducibles.

(b) A monic polynomial of degree n over $\mathbb{F}_p$ is determined by its n lower coefficients, so there are exactly p^n of them, and there are $1 + p + \cdots + p^N$ monic polynomials of degree at most N, a finite number. By (a), infinitely many monic irreducibles exist, so some of them have degree greater than N.

(c) If F is infinite, the polynomials $x - c$ for $c \in F$ are infinitely many distinct monic polynomials of degree 1, all irreducible by Example (First Irreducible and Reducible Polynomials) (a). Over a finite field there are only finitely many polynomials of degree 1, so the content of (a) is that irreducibles of higher and higher degree keep appearing. Over $\mathbb{F}_2$, for example, $x^3 + x + 1$ is irreducible of degree 3 by Exercise (B3: Small Irreducibles Over $\mathbb{F}_2$).

Exercise. C2: Gauss's Lemma and Eisenstein's Criterion

Let $\mathbb{Z}[x]$ be as in Theorem 6.3.6, and let p be a prime. For $g = \sum b_i x^i \in \mathbb{Z}[x]$, define its **reduction modulo** p as $\bar{g} = \sum [b_i] x^i \in \mathbb{F}_p[x]$, where $[b]$ is the class of b in $\mathbb{F}_p = \mathbb{Z}/p\mathbb{Z}$ (Example (Arithmetic on $\mathbb{Z}/n\mathbb{Z}$ and the Field $\mathbb{F}_2$)).

(a) Prove that $\overline{g+h} = \bar{g} + \bar{h}$ and $\overline{gh} = \bar{g}\,\bar{h}$ for all $g, h \in \mathbb{Z}[x]$, and that $\deg \bar{g} \leq \deg g$.

(b) Let $g, h \in \mathbb{Z}[x]$, and suppose p divides every coefficient of gh. Prove that p divides every coefficient of g or every coefficient of h.

(c) **(Gauss's lemma.)** Let $f \in \mathbb{Z}[x]$, and suppose $f = gh$ with $g, h \in \mathbb{Q}[x]$. Prove that there is a nonzero $\lambda \in \mathbb{Q}$ with $\lambda g \in \mathbb{Z}[x]$ and $\lambda^{-1} h \in \mathbb{Z}[x]$.

(d) Let F be a field, $c \in F$ nonzero, $n \in \mathbb{N}$, and $u, v \in F[x]$ with $uv = cx^n$. Prove that $u = \beta x^s$ and $v = \gamma x^t$ for some nonzero $\beta, \gamma \in F$ and $s, t \in \mathbb{N}$.

(e) Prove Theorem 6.3.7.

Hint for (c): among all ways of writing $kf = (\alpha g)(\beta h)$ with $\alpha, \beta \in \mathbb{Q}$, $\alpha\beta = k$ a positive

integer, and $\alpha g, \beta h \in \mathbb{Z}[x]$, take k as small as possible. Hint for (e): suppose $f = gh$ with g, h nonconstant, use (c), and reduce modulo p.

Solution.

(a) By Example (Arithmetic on $\mathbb{Z}/n\mathbb{Z}$ and the Field $\mathbb{F}_2$), $[a]+[b] = [a+b]$ and $[a][b] = [ab]$. Let $g = \sum b_i x^i$ and $h = \sum c_i x^i$. The coefficient of x^k in $g+h$ is $b_k + c_k$, whose class is $[b_k] + [c_k]$, the coefficient of x^k in $\bar{g} + \bar{h}$. The coefficient of x^k in gh is $\sum_{i+j=k} b_i c_j$, whose class is $\sum_{i+j=k} [b_i][c_j]$, the coefficient of x^k in $\bar{g}\,\bar{h}$ (Definition 0.8.2). Finally, if $b_i = 0$ then $[b_i] = 0$, so $\bar{g}$ has no nonzero coefficient beyond position $\deg g$.

(b) The hypothesis says $\overline{gh} = 0$. By (a), $\bar{g}\,\bar{h} = 0$ in $\mathbb{F}_p[x]$. Since p is prime, $\mathbb{F}_p$ is a field (Theorem 0.7.3), so Corollary 0.8.4 gives $\bar{g} = 0$ or $\bar{h} = 0$. That is, p divides every coefficient of g or every coefficient of h.

(c) Choose positive integers m, n with $mg \in \mathbb{Z}[x]$ and $nh \in \mathbb{Z}[x]$, for instance the products of the denominators of the coefficients. Let K be the set of positive integers k for which there are $\alpha, \beta \in \mathbb{Q}$ with $\alpha\beta = k$, $\alpha g \in \mathbb{Z}[x]$, and $\beta h \in \mathbb{Z}[x]$. Then $mn \in K$, so by Theorem 0.2.11 K has a least element k, with corresponding α, β. Suppose $k > 1$. By Proposition 0.2.12, some prime q divides k. Now $(\alpha g)(\beta h) = \alpha\beta\, gh = kf$, and every coefficient of kf is divisible by q. By (b) with the prime q, q divides every coefficient of αg or every coefficient of βh; say the first (the other case is symmetric). Then $(\alpha/q)g \in \mathbb{Z}[x]$, $\beta h \in \mathbb{Z}[x]$, and $(\alpha/q)\beta = k/q$ is a positive integer smaller than k, which lies in K. This contradicts the minimality of k. Hence $k = 1$, so $\beta = \alpha^{-1}$, and $\lambda = \alpha$ works.

(d) Since $uv \neq 0$, both u, v are nonzero. Write $u = \sum u_i x^i$, $v = \sum v_j x^j$, let i_0 and j_0 be the smallest indices with $u_{i_0} \neq 0$ and $v_{j_0} \neq 0$, and let $s = \deg u$, $t = \deg v$. The coefficient of $x^{i_0+j_0}$ in uv is $\sum_{i+j=i_0+j_0} u_i v_j$. A term with $i < i_0$ has $u_i = 0$, and a term with $i > i_0$ has $j < j_0$, so $v_j = 0$. So this coefficient is $u_{i_0} v_{j_0} \neq 0$. The only nonzero coefficient of cx^n is at x^n, so $i_0 + j_0 = n$. By Theorem 0.8.3, also $s + t = n$. Since $i_0 \leq s$ and $j_0 \leq t$, this forces $i_0 = s$ and $j_0 = t$. So u has exactly one nonzero coefficient, $u = u_s x^s$, and likewise $v = v_t x^t$.

(e) Suppose f is reducible over $\mathbb{Q}$: $f = gh$ with $g, h \in \mathbb{Q}[x]$, $\deg g = s \geq 1$, $\deg h = t \geq 1$, and $s+t = n$. By (c), replacing g, h by $\lambda g, \lambda^{-1} h$ (which have the same degrees), we may assume $g, h \in \mathbb{Z}[x]$. Write g_0, h_0 for their constant coefficients, so $a_0 = g_0 h_0$. Reduce modulo p. By conditions (1) and (2), $\bar{f} = [a_n]x^n$ with $[a_n] \neq 0$. By (a), $\bar{g}\,\bar{h} = [a_n]x^n$, so by (d), $\bar{g} = \beta x^{s'}$ and $\bar{h} = \gamma x^{t'}$ with $\beta, \gamma \neq 0$. By Theorem 0.8.3, $s' + t' = n = s + t$, while $s' \leq s$ and $t' \leq t$ by (a). Hence $s' = s \geq 1$ and $t' = t \geq 1$. So the constant coefficients of $\bar{g}$ and $\bar{h}$ are 0, that is, $p \mid g_0$ and $p \mid h_0$. Then $p^2 \mid g_0 h_0 = a_0$, contradicting condition (3). Therefore f is irreducible over $\mathbb{Q}$; it is nonconstant since $n \geq 1$.

6.4 Roots, Multiplicity, and Derivatives

Chapter 3 bounded the number of roots of a polynomial, and the first section of this chapter restated the bound as Corollary 6.1.4: a nonzero polynomial of degree n has at most n roots. The bound counts each root once, and so it misses something. The polynomial $(x-1)^2$ has degree 2 but only one root, and the root 1 is somehow "used twice". This section makes "used twice" precise with the multiplicity of a root, sharpens the root bound, and then gives a test for repeated roots that needs no factoring: the formal derivative. The derivative behaves differently in characteristic p, and we say exactly where. The section ends with Taylor expansion of a

polynomial, which turns out to be a dual basis from Chapter 5.

6.4.1 Multiplicity of a root

By the Factor Theorem (Lemma 3.7.1, restated in general as Theorem 6.1.3), c is a root of f exactly when $x - c$ divides f. A repeated root should mean that a higher power of $x - c$ divides f. So we measure how many factors $x - c$ can be pulled out.

The multiplicity of c counts how many times $x - c$ divides f.

Definition 6.4.1. Multiplicity of a Root

Let F be a field, $f \in F[x]$ **nonzero**, and $c \in F$. The **multiplicity of c as a root of f**, written $\text{mult}_c(f)$, is the **largest** $m \in \mathbb{N}$ such that $(x - c)^m \mid f$.

The root c is **simple** if $\text{mult}_c(f) = 1$, and **multiple** (or **repeated**) if $\text{mult}_c(f) \geq 2$.

In words: keep dividing f by $x - c$ as long as the division leaves no remainder, and count the successful divisions. Divisibility is meant in the sense of Definition 6.1.1, so $(x - c)^m \mid f$ says $f = (x - c)^m g$ for some $g \in F[x]$.

There is something to check: a largest such m must exist. The value $m = 0$ always works, since $(x - c)^0 = 1$ divides every polynomial. And if $f = (x - c)^m g$, then $g \neq 0$ because $f \neq 0$, so Theorem 0.8.3 gives $\deg f = m + \deg g \geq m$. The admissible m form a nonempty set of natural numbers bounded by $\deg f$, so it has a largest element. Moreover $\text{mult}_c(f) \geq 1$ exactly when $x - c \mid f$, that is, exactly when c is a root. So multiplicity 0 means "not a root".

The hypothesis $f \neq 0$ is not decoration. Every power $(x-c)^m$ divides the zero polynomial, since $0 = (x - c)^m \cdot 0$. There is no largest m, and the zero polynomial has no multiplicities.

A definition by "largest m" is awkward to use directly, because to confirm m we must rule out all larger powers. The next lemma replaces it with a single test at c.

Lemma 6.4.1. Multiplicity via the Cofactor

Let $f \in F[x]$ be nonzero, $c \in F$, and $m \in \mathbb{N}$. Then $\text{mult}_c(f) = m$ if and only if

$$f = (x - c)^m g \quad \text{for some } g \in F[x] \text{ with } g(c) \neq 0.$$

Proof. ($\Rightarrow$) Suppose $\text{mult}_c(f) = m$. Then $f = (x - c)^m g$ for some $g \in F[x]$. If $g(c) = 0$, then $g = (x - c)h$ by Lemma 3.7.1, so $f = (x - c)^{m+1} h$, contradicting the maximality of m. Hence $g(c) \neq 0$.

($\Leftarrow$) Suppose $f = (x-c)^m g$ with $g(c) \neq 0$. Then $(x-c)^m \mid f$, so $\text{mult}_c(f) \geq m$. Suppose, for a contradiction, that $(x-c)^{m+1} \mid f$, say $f = (x-c)^{m+1} h$. Then $(x-c)^m g = (x-c)^m (x-c)h$. Since $(x-c)^m \neq 0$, cancellation (Corollary 0.8.4) gives $g = (x-c)h$, and evaluating at c gives $g(c) = 0$ by Theorem 0.8.5, a contradiction. Every higher power $(x - c)^k$ with $k > m + 1$ would also give $(x - c)^{m+1} \mid f$. Hence $\text{mult}_c(f) = m$.

■

The first payoff is that multiplicities add when polynomials multiply, just as exponents do.

Theorem 6.4.2. Multiplicity of a Product

Let $f, g \in F[x]$ be nonzero and $c \in F$. Then $\operatorname{mult}_c(fg) = \operatorname{mult}_c(f) + \operatorname{mult}_c(g)$.

Proof. Let $m = \operatorname{mult}_c(f)$ and $n = \operatorname{mult}_c(g)$. By Lemma 6.4.1, $f = (x-c)^m u$ and $g = (x-c)^n v$ with $u(c) \neq 0$ and $v(c) \neq 0$. Then $fg = (x-c)^{m+n} uv$, and $(uv)(c) = u(c)v(c) \neq 0$ by Theorem 0.8.5 and Theorem 0.7.1. By Lemma 6.4.1 again, $\operatorname{mult}_c(fg) = m + n$. ∎

Example. Finding a Multiplicity by Repeated Division

Let $f = x^4 - 2x^3 + 2x - 1 \in \mathbb{Q}[x]$. Find the multiplicity of 1 and of -1 as roots of f.

Solution. We have $f(1) = 1 - 2 + 2 - 1 = 0$, so we divide by $x - 1$. Long division gives

$$f = (x-1)(x^3 - x^2 - x + 1).$$

The cofactor $g_1 = x^3 - x^2 - x + 1$ has $g_1(1) = 0$, so we divide again: $g_1 = (x-1)(x^2 - 1)$. The new cofactor $g_2 = x^2 - 1$ has $g_2(1) = 0$, and $g_2 = (x-1)(x+1)$. Now $g_3 = x + 1$ has $g_3(1) = 2 \neq 0$. Hence

$$f = (x-1)^3(x+1), \qquad g_3(1) \neq 0,$$

and Lemma 6.4.1 gives $\operatorname{mult}_1(f) = 3$. For -1, write $f = (x+1) \cdot (x-1)^3$, whose cofactor $(x-1)^3$ takes the value $(-2)^3 = -8 \neq 0$ at -1. So $\operatorname{mult}_{-1}(f) = 1$, a simple root. The multiplicities add up to $3 + 1 = 4 = \deg f$.

Warning. **Multiplicity is not visible from the set of roots.** The polynomials $(x-1)^3(x+1)$ and $(x-1)(x+1)^3$ have the same roots $\{1, -1\}$, the same degree, and the same leading coefficient, but they are different polynomials. "The roots of f" as a set loses information; the roots **with their multiplicities** do not, as the next theorem shows for polynomials that split.

Quick check. Over $\mathbb{R}$, what is the multiplicity of 0 as a root of $x^3(x^2+1)$? Of 2 as a root of $(x^2-4)^2$?

Solution. For $x^3(x^2+1)$, the cofactor $x^2 + 1$ is $1 \neq 0$ at 0, so the multiplicity is 3 by Lemma 6.4.1. For $(x^2-4)^2 = (x-2)^2(x+2)^2$, the cofactor $(x+2)^2$ is $16 \neq 0$ at 2, so the multiplicity is 2.

6.4.2 Counting roots with multiplicity

In Example (Finding a Multiplicity by Repeated Division) the multiplicities added up to exactly the degree. That cannot always happen: $x^2+1 \in \mathbb{R}[x]$ has degree 2 and no real roots at all. The difference is whether f breaks into linear factors, so we give that a name.

Definition 6.4.2. Splitting into Linear Factors

Let F be a field. A nonzero $f \in F[x]$ of degree n **splits over** F if there are $a \in F$ and $a_1, \dots, a_n \in F$, not necessarily distinct, with

$$f = a(x-a_1)(x-a_2)\cdots(x-a_n).$$

A nonzero constant splits (take $n=0$ and the empty product 1). The polynomial $x^2-1 = (x-1)(x+1)$ splits over $\mathbb{Q}$. The polynomial x^2-2 does not split over $\mathbb{Q}$, since a linear factor $x-a_1$ with $a_1 \in \mathbb{Q}$ would give a rational root, but it splits over $\mathbb{R}$ as $(x-\sqrt{2})(x+\sqrt{2})$. The phrase **over** F carries weight: the a_i must lie in F.

It turns out that the multiplicities always fit inside the degree, and they fill it exactly when f splits:

Theorem 6.4.3. Roots Counted with Multiplicity

Let F be a field and $f \in F[x]$ nonzero. Let $c_1, \dots, c_k$ be the distinct roots of f in F, with multiplicities $m_1, \dots, m_k$. Then

$$f = (x-c_1)^{m_1}\cdots(x-c_k)^{m_k}\, g$$

for some $g \in F[x]$ with **no** root in F. Consequently

$$m_1 + \cdots + m_k \le \deg f,$$

with equality if and only if f splits over F.

Idea. Peel off the roots one at a time. After removing $(x-c_1)^{m_1}$, the other roots c_j are still roots of the cofactor, because $c_j - c_1 \neq 0$, and by Theorem 6.4.2 they keep their multiplicities. Whatever is left at the end has no roots, and its degree is exactly the gap between $\deg f$ and $\sum m_i$. The gap is zero exactly when the leftover is a constant.

Proof. The set of roots is finite by Corollary 6.1.4, so the list $c_1, \dots, c_k$ makes sense. We prove the following claim by induction on $j \in \{0, 1, \dots, k\}$: $f = (x-c_1)^{m_1}\cdots(x-c_j)^{m_j}\, g_j$ for some nonzero $g_j \in F[x]$ with

$$\operatorname{mult}_{c_i}(g_j) = 0 \text{ for } i \le j, \qquad \operatorname{mult}_{c_i}(g_j) = m_i \text{ for } i > j.$$

For $j=0$ take $g_0 = f$. Suppose the claim holds for some $j < k$. By Lemma 6.4.1 applied to g_j, we have $g_j = (x-c_{j+1})^{m_{j+1}} g_{j+1}$ with $g_{j+1}(c_{j+1}) \neq 0$, which gives the product formula for $j+1$ and $\operatorname{mult}_{c_{j+1}}(g_{j+1}) = 0$. For $i \neq j+1$, the polynomial $(x-c_{j+1})^{m_{j+1}}$ takes the value $(c_i - c_{j+1})^{m_{j+1}} \neq 0$ at c_i, since the roots are distinct. So its multiplicity at c_i is 0 by

Lemma 6.4.1 (with cofactor the polynomial itself), and Theorem 6.4.2 gives

$$\operatorname{mult}_{c_i}(g_j) = 0 + \operatorname{mult}_{c_i}(g_{j+1}).$$

This is 0 for $i \le j$ and m_i for $i > j+1$, which completes the induction.

Put $g = g_k$. By the claim, $\operatorname{mult}_{c_i}(g) = 0$, that is, $g(c_i) \neq 0$, for every i. If $d \in F$ were a root of g, then $f(d) = (d-c_1)^{m_1} \cdots (d-c_k)^{m_k} g(d) = 0$, so d would be one of the c_i, which is impossible. Hence g has no root in F.

Since $f \neq 0$, also $g \neq 0$, and Theorem 0.8.3 gives $\deg f = m_1 + \cdots + m_k + \deg g \geq m_1 + \cdots + m_k$.

($\Rightarrow$) If equality holds, then $\deg g = 0$, so $g = a$ is a nonzero constant and $f = a(x-c_1)^{m_1} \cdots (x-c_k)^{m_k}$ splits over F.

($\Leftarrow$) Suppose $f = a(x-a_1)\cdots(x-a_n)$ with $n = \deg f$. Group equal factors: $f = a(x-d_1)^{n_1} \cdots (x-d_l)^{n_l}$ with $d_1, \ldots, d_l$ distinct and $n_1 + \cdots + n_l = n$. By the computation above, the cofactor $a\prod_{s \neq r}(x-d_s)^{n_s}$ is nonzero at d_r, so $\operatorname{mult}_{d_r}(f) = n_r$ by Lemma 6.4.1. Every root d of f satisfies $a\prod_s (d-d_s)^{n_s} = 0$, so d is one of the d_r. Hence the roots of f are exactly $d_1, \ldots, d_l$, and their multiplicities add up to $n = \deg f$. This proves the theorem.

■

The theorem sharpens Corollary 6.1.4, because the number k of distinct roots is at most $m_1 + \cdots + m_k$. It also shows that a split polynomial is determined by its leading coefficient and its roots with multiplicities. Whether a given polynomial splits depends on the field: the fifth-degree polynomial $(x-1)^3(x^2+1)$ has root multiplicities adding up to $3 < 5$ over $\mathbb{R}$, but it splits over $\mathbb{C}$ as $(x-1)^3(x-i)(x+i)$. In the next section we prove that over $\mathbb{C}$ **every** polynomial splits.

6.4.3 The formal derivative

How do we recognize a repeated root without factoring? Over $\mathbb{R}$ calculus has an answer. At a simple root the graph of f crosses the axis with nonzero slope, while at a double root, as for $(x-1)^2$ at 1, it touches the axis and the slope is 0. We would like to use this over every field, including $\mathbb{F}_p$, where there are no limits and no slopes. So we keep the formula of the derivative and throw away the limit.

The formal derivative is the power rule, applied coefficient by coefficient.

Definition 6.4.3. Formal Derivative

Let F be a field and $f = a_0 + a_1 x + \cdots + a_n x^n \in F[x]$. The **formal derivative** of f is

$$f' = a_1 + 2a_2 x + 3a_3 x^2 + \cdots + n a_n x^{n-1} = \sum_{k=1}^{n} k a_k x^{k-1},$$

where ka_k is the integer multiple of a_k (the sum of k copies of a_k). Higher derivatives are defined by $f^{(0)} = f$ and $f^{(k+1)} = (f^{(k)})'$; we also write $f'' = f^{(2)}$.

In words: each monomial $a_k x^k$ becomes $k a_k x^{k-1}$, and constants disappear. This is the map D of Example (Differentiation), where we already saw that $D\colon F[x] \to F[x]$ is linear. Over $\mathbb{R}$ it agrees with the derivative of calculus on polynomial functions, since calculus gives the same

power rule. By Lemma 0.7.4, $ka_k = (k \cdot 1)a_k$, so the integer k acts through the element $k \cdot 1$ of F. **That element can be** 0, and this is the source of every difference between characteristic 0 and characteristic p.

Some examples. Over $\mathbb{Q}$, $(x^4 - 2x^3 + 2x - 1)' = 4x^3 - 6x^2 + 2$. Over any field, the derivative of a constant is 0 and $(x - c)' = 1$. Over $\mathbb{F}_3$, $(x^3 + x)' = 3x^2 + 1 = 1$, because $3 \cdot 1 = 0$ there. Over $\mathbb{F}_2$, $(x^2)' = 2x = 0$: a nonconstant polynomial with derivative zero.

In characteristic 0 the last example cannot happen. If $\deg f = n \geq 1$, the coefficient of x^{n-1} in f' is $(n \cdot 1)a_n$, a product of two nonzero elements, so $\deg f' = n - 1$ and in particular $f' \neq 0$. In characteristic p the same coefficient vanishes whenever $p \mid n$, because then $n \cdot 1 = (n/p) \cdot (p \cdot 1) = 0$ by Lemma 0.7.4.

The derivative obeys the rules we know from calculus. The proofs cannot use limits, so they go through monomials.

Theorem 6.4.4. Rules for the Formal Derivative

Let F be a field, $f, g \in F[x]$, $a, b, c \in F$, and $m \geq 1$ an integer. Then:

(a) $(af + bg)' = af' + bg'$, and $a' = 0$ for the constant polynomial a;
(b) $(fg)' = f'g + fg'$ (product rule);
(c) $(g^m)' = m\, g^{m-1} g'$; in particular $\big((x - c)^m\big)' = m(x - c)^{m-1}$.

Idea. Both sides of the product rule are linear in f for fixed g, and linear in g for fixed f. So it is enough to check the rule when f and g are monomials, where it is the identity $i+j = i+j$ for exponents. Checking on monomials, and extending by linearity, is the standard way to prove any identity about D.

Proof.

(a) Linearity is Example (Differentiation). A constant has no coefficients in positions $k \geq 1$, so its derivative is 0.

(b) First let $f = x^i$ and $g = x^j$ with $i, j \in \mathbb{N}$. If $i, j \geq 1$, then

$$(x^{i+j})' = (i + j)x^{i+j-1} = ix^{i-1}x^j + x^i\, jx^{j-1} = f'g + fg',$$

using $(i + j) \cdot 1 = i \cdot 1 + j \cdot 1$ from Lemma 0.7.4. If $i = 0$, then $f = 1$, $f' = 0$, and both sides equal jx^{j-1} (or 0 when $j = 0$); the case $j = 0$ is symmetric.
Now let $f = \sum_i a_i x^i$ and $g = \sum_j b_j x^j$. By Theorem 0.8.1, $fg = \sum_{i,j} a_i b_j\, x^i x^j$, a finite sum. Applying (a) and the monomial case,

$$\begin{aligned}(fg)' &= \sum_{i,j} a_i b_j \big((x^i)' x^j + x^i (x^j)'\big) \\ &= \Big(\sum_i a_i (x^i)'\Big)\Big(\sum_j b_j x^j\Big) + \Big(\sum_i a_i x^i\Big)\Big(\sum_j b_j (x^j)'\Big) \\ &= f'g + fg',\end{aligned}$$

where the middle equality regroups with the distributive laws, and the last uses (a) again.

(c) We use induction on m. For $m = 1$ both sides are g'. If $(g^m)' = mg^{m-1}g'$, then by (b)

$$(g^{m+1})' = (g^m g)' = mg^{m-1}g'\,g + g^m g' = (m+1)g^m g'.$$

For $g = x - c$ we have $g' = 1$ by (a). This proves the theorem.

■

6.4.4 Repeated roots and the derivative

Now we can turn the calculus picture into a theorem valid over every field.

Theorem 6.4.5. Repeated Root Test

Let F be a field, $f \in F[x]$ nonzero, and $c \in F$.

(a) c is a multiple root of f if and only if $f(c) = 0$ and $f'(c) = 0$.

(b) Suppose $\mathrm{mult}_c(f) = m \geq 1$ and $m \cdot 1 \neq 0$ in F (this holds automatically if F has characteristic 0). Then $f' \neq 0$ and $\mathrm{mult}_c(f') = m - 1$.

Idea. Write $f = (x - c)g$ and differentiate: $f' = g + (x - c)g'$, so $f'(c) = g(c)$. The derivative at c reads off exactly the cofactor's value, which is what Lemma 6.4.1 needs. For (b), differentiate $(x - c)^m g$ instead; the factor m appears, and it must not be 0 in F.

Proof.

(a) ($\Rightarrow$) Suppose $f = (x-c)^2 g$. By Theorem 6.4.4 (b) and (c), $f' = 2(x-c)g + (x-c)^2 g'$, and evaluating at c gives $f(c) = 0$ and $f'(c) = 0$.
($\Leftarrow$) Suppose $f(c) = 0$ and $f'(c) = 0$. By Lemma 3.7.1, $f = (x-c)g$ for some $g \in F[x]$. By Theorem 6.4.4, $f' = g + (x-c)g'$, so $0 = f'(c) = g(c)$. By Lemma 3.7.1 again, $g = (x-c)h$, so $f = (x-c)^2 h$ and $\mathrm{mult}_c(f) \geq 2$.

(b) By Lemma 6.4.1, $f = (x-c)^m g$ with $g(c) \neq 0$. By Theorem 6.4.4,

$$f' = m(x-c)^{m-1}g + (x-c)^m g' = (x-c)^{m-1}\,h, \qquad h := mg + (x-c)g'.$$

Then $h(c) = (m \cdot 1)\,g(c)$, a product of two nonzero elements of F, so $h(c) \neq 0$. In particular $h \neq 0$, so $f' \neq 0$ by Corollary 0.8.4, and $\mathrm{mult}_c(f') = m - 1$ by Lemma 6.4.1.

■

Part (a) holds over **every** field. Part (b) needs its hypothesis. Over $\mathbb{F}_2$, let $f = x^2(x+1) = x^3 + x^2$. Then $\mathrm{mult}_0(f) = 2$, and $f' = 3x^2 + 2x = x^2$, so $\mathrm{mult}_0(f') = 2$, not 1. Here $m \cdot 1 = 2 \cdot 1 = 0$, and the factor m killed the term that should have lowered the multiplicity.

Part (a) finds a repeated root at a known point c. To decide whether **any** repeated root exists, without knowing the roots, we use the gcd of f and f', which the Euclidean algorithm (Theorem 6.2.6) computes by division alone. Repeated roots are a special case of repeated irreducible factors: c is a multiple root exactly when the irreducible polynomial $x - c$ appears squared.

Corollary 6.4.6. Repeated Factors and $\gcd(f, f')$

Let F be a field and $f \in F[x]$ nonconstant.

(a) (**Any field.**) If $p \in F[x]$ is irreducible and $p^2 \mid f$, then $p \mid \gcd(f, f')$. Hence, if $\gcd(f, f') = 1$, then no irreducible polynomial divides f twice; in particular f has no multiple root in F.

(b) (**Characteristic 0.**) If F has characteristic 0 and $\gcd(f, f') \neq 1$, then $p^2 \mid f$ for some irreducible $p \in F[x]$.

Idea. For (a), differentiate p^2h: every term still contains p. For (b), run it backwards. An irreducible p dividing both $f = pg$ and $f' = p'g + pg'$ must divide $p'g$. If p does not divide p', then Euclid's lemma pushes p into g, and $p^2 \mid f$. The only question is whether $p \mid p'$ can happen, and that is exactly where the characteristic enters.

Proof.

(a) Write $f = p^2h$. By Theorem 6.4.4, $f' = 2pp'h + p^2h' = p(2p'h + ph')$. So p is a common divisor of f and f', and $p \mid \gcd(f, f')$ by Theorem 6.2.3. If $\gcd(f, f') = 1$, this would give $p \mid 1$, which is impossible because p is nonconstant (Definition 6.3.1) and a nonzero multiple of p has degree at least $\deg p \geq 1$ by Theorem 0.8.3. If c were a multiple root, then $(x - c)^2 \mid f$. But $x - c$ is irreducible (Definition 6.3.1), because a product of two nonconstant polynomials has degree at least 2. This is again impossible.

(b) Let $d = \gcd(f, f') \neq 1$. Since $d \mid f$ and $f \neq 0$, $d \neq 0$; as d is monic and $d \neq 1$, $\deg d \geq 1$. By Theorem 6.3.5, d has a monic irreducible factor p. Then $p \mid f$ and $p \mid f'$ (Theorem 6.2.3). Write $f = pg$. By Theorem 6.4.4, $f' = p'g + pg'$, so p divides $f' - pg' = p'g$.

Since F has characteristic 0 and $\deg p \geq 1$, we showed after Definition 6.4.3 that $p' \neq 0$ and $\deg p' = \deg p - 1$. A nonzero multiple of p has degree at least $\deg p$, so $p \nmid p'$. By Euclid's lemma (Theorem 6.3.3), $p \mid g$. Hence $p^2 \mid pg = f$. This proves the corollary.

■

Warning. **In characteristic p, a nonconstant polynomial can have derivative 0.** Over $\mathbb{F}_p$, $(x^p)' = (p \cdot 1)x^{p-1} = 0$. The proof of (b) used that an irreducible factor q of f has $q' \neq 0$ of smaller degree, so that $q \nmid q'$; in characteristic p that step is no longer automatic, because q' may be 0. So the correct statements are: $\gcd(f, f') = 1$ implies no repeated irreducible factor, over **any** field; the converse is proved here **only in characteristic** 0. Do not use "repeated factor if and only if $\gcd(f, f') \neq 1$" over a field of characteristic p without checking that no irreducible factor of f has derivative 0.

Remark. The converse really can fail in characteristic p, though not over the fields $\mathbb{F}_p$. Over the field of rational functions in a variable t with coefficients in $\mathbb{F}_p$, which this book does not construct, the polynomial $x^p - t$ is irreducible, so it has no repeated factor, yet its derivative is 0.

Example. Detecting a Repeated Root with a gcd

Decide whether $f = x^4 - 2x^3 + 2x - 1 \in \mathbb{Q}[x]$ has a repeated irreducible factor, using only $\gcd(f, f')$. Does f have a multiple root in $\mathbb{C}$?

Solution. Here $f' = 4x^3 - 6x^2 + 2$. Run the Euclidean algorithm (Theorem 6.2.6). Dividing f by f',

$$f = \left(\tfrac{1}{4}x - \tfrac{1}{8}\right)f' + \left(-\tfrac{3}{4}x^2 + \tfrac{3}{2}x - \tfrac{3}{4}\right),$$

and the remainder is $-\frac{3}{4}(x^2 - 2x + 1)$. Dividing f' by $x^2 - 2x + 1$,

$$f' = (4x + 2)(x^2 - 2x + 1) + 0.$$

(We divided by the monic multiple of the remainder; scaling a divisor by a nonzero constant changes only the quotient, so the next remainder is still 0.) The last nonzero remainder, made monic, is the gcd: $\gcd(f, f') = x^2 - 2x + 1 = (x - 1)^2 \neq 1$. By Corollary 6.4.6 (b), since $\mathbb{Q}$ has characteristic 0, f has a repeated irreducible factor. Indeed $x - 1$ divides the gcd, so 1 is a common root of f and f', and by Theorem 6.4.5 (a) it is a multiple root. This matches Example (Finding a Multiplicity by Repeated Division), where $\mathrm{mult}_1(f) = 3$, and Theorem 6.4.5 (b) gives $\mathrm{mult}_1(f') = 2$, consistent with the factor $(x - 1)^2$ of the gcd.

So yes, f has a multiple root in $\mathbb{C}$, namely $1 \in \mathbb{Q}$. The gcd test is useful over $\mathbb{C}$ even for polynomials with rational coefficients: the Euclidean algorithm uses only divisions in $\mathbb{Q}[x]$, and division with remainder is unique (Theorem 6.1.1), so running it in $\mathbb{C}[x]$ produces the same remainders and the same gcd. Had the gcd been 1, part (a) over $\mathbb{C}$ would have ruled out every multiple complex root, without our finding a single root.

Quick check. Over $\mathbb{R}$, $f = x^3$ satisfies $f(0) = f'(0) = 0$. Does Theorem 6.4.5 say that 0 has multiplicity 2?

Solution. No. Part (a) says only that the multiplicity is **at least** 2. The actual multiplicity is 3. Part (b) applies, since $\mathbb{R}$ has characteristic 0: $\mathrm{mult}_0(f') = \mathrm{mult}_0(3x^2) = 2 = 3 - 1$. To pin the multiplicity down with derivatives, find the first k with $f^{(k)}(0) \neq 0$: here $f''(0) = 0$ and $f'''(0) = 6 \neq 0$, so $k = 3$.

6.4.5 Taylor expansion of a polynomial

The answer to the check suggests that, in characteristic 0, the derivatives of f at c know everything about f near c. Calculus makes this precise with Taylor's formula. For polynomials the formula is exact, and it has a clean linear-algebra meaning: the numbers $f^{(k)}(c)/k!$ are the coordinates of f in the basis of powers of $x - c$, so the maps $f \mapsto f^{(k)}(c)/k!$ form a dual basis in the sense of Definition 5.1.3. In Example (Dual Bases in $\mathbb{R}[x]_{\leq 2}$) we saw this for $c = 0$ and $n = 2$.

We need one derivative computation. For integers $0 \leq j \leq k$, repeated use of Theorem 6.4.4 (c)

and (a) gives

$$\left((x-c)^k\right)^{(j)} = k(k-1)\cdots(k-j+1)\,(x-c)^{k-j},$$

by induction on j, and $\left((x-c)^k\right)^{(j)} = 0$ for $j > k$, since the k-th derivative is the constant $k!$. Evaluating at c, every term with $j < k$ vanishes because of the factor $(x-c)^{k-j}$. Hence

$$\left((x-c)^k\right)^{(j)}(c) = \begin{cases} k! \cdot 1 & \text{if } j = k, \\ 0 & \text{if } j \neq k. \end{cases}$$

Theorem 6.4.7. Taylor Expansion of a Polynomial

Let F be a field, $n \in \mathbb{N}$, and $c \in F$.

(a) The list $\mathcal{B}_c = (1, x-c, (x-c)^2, \ldots, (x-c)^n)$ is a basis of $F[x]_{\leq n}$.

(b) Suppose F has characteristic 0, or characteristic $p > n$. For $k = 0, \ldots, n$ define $\varphi_k \colon F[x]_{\leq n} \to F$ by $\varphi_k(f) = \dfrac{f^{(k)}(c)}{k!}$. Then $(\varphi_0, \ldots, \varphi_n)$ is the dual basis of $\mathcal{B}_c$, and every $f \in F[x]_{\leq n}$ satisfies

$$f = \sum_{k=0}^{n} \frac{f^{(k)}(c)}{k!}\,(x-c)^k.$$

Idea. ① The powers of $x-c$ have degrees $0, 1, \ldots, n$, so no combination of them can cancel down to zero; there are $n+1$ of them, so counting finishes (a). ② The computation above says exactly $\varphi_j\left((x-c)^k\right) = \delta_{jk}$, once we may divide by $k!$. ③ A list of functionals with the δ property is the dual basis, and the dual basis expansion $v = \sum \varphi_k(v) v_k$ of Theorem 5.1.3 is Taylor's formula.

Proof.

(a) Let $b_0 + b_1(x-c) + \cdots + b_n(x-c)^n = 0$, and suppose not all b_k are zero. Let k be the largest index with $b_k \neq 0$. By Theorem 0.8.3, $b_k(x-c)^k$ has degree k, and each earlier term has degree less than k, so by Theorem 0.8.2 the sum has degree $k \geq 0$. This contradicts the sum being 0, of degree $-\infty$. Hence $\mathcal{B}_c$ is linearly independent. It has $n+1$ elements and $\dim F[x]_{\leq n} = n+1$ by Example (Standard Bases), so $\mathcal{B}_c$ is a basis by Theorem 1.6.4 (a).

(b) First, $k! \cdot 1 \neq 0$ in F for $0 \leq k \leq n$. By Lemma 0.7.4, $k! \cdot 1 = (1 \cdot 1)(2 \cdot 1)\cdots(k \cdot 1)$. In characteristic 0 each factor is nonzero by Definition 0.7.2; in characteristic $p > n$, each factor $j \cdot 1$ with $1 \leq j \leq k < p$ is nonzero by the minimality of p. A product of nonzero elements is nonzero (Theorem 0.7.1). So φ_k is defined.

Each φ_k is linear: $f \mapsto f^{(k)}$ is linear as a composite of copies of D (Example (Differentiation)), evaluation at c is linear by Theorem 0.8.5, and so is multiplication by $(k!)^{-1}$. The derivatives of a polynomial of degree at most n have degree at most n, so $\varphi_k \in F[x]^*_{\leq n}$. By the computation before the theorem,

$$\varphi_j\left((x-c)^k\right) = \frac{1}{j!}\left((x-c)^k\right)^{(j)}(c) = \delta_{jk} \qquad (0 \leq j, k \leq n).$$

By the uniqueness in Theorem 5.1.3 (a), $(\varphi_0, \ldots, \varphi_n)$ is the dual basis of $\mathcal{B}_c$. The first formula of Theorem 5.1.3 (c), $f = \sum_k \varphi_k(f)(x-c)^k$, is the Taylor expansion. This proves

the theorem.

■

Example. Re-centering a Cubic

Expand $f = x^3 - 2x + 5 \in \mathbb{Q}[x]$ in powers of $x - 1$.

Solution. The derivatives are $f' = 3x^2 - 2$, $f'' = 6x$, $f''' = 6$. At $c = 1$: $f(1) = 4$, $f'(1) = 1$, $f''(1) = 6$, $f'''(1) = 6$. By Theorem 6.4.7 (b), since $\mathbb{Q}$ has characteristic 0,

$$\begin{aligned} f &= 4 + 1 \cdot (x-1) + \frac{6}{2}(x-1)^2 + \frac{6}{6}(x-1)^3 \\ &= 4 + (x-1) + 3(x-1)^2 + (x-1)^3. \end{aligned}$$

Check by expanding: $(x-1)^3 + 3(x-1)^2 + (x-1) + 4$ equals

$$(x^3 - 3x^2 + 3x - 1) + (3x^2 - 6x + 3) + (x - 1) + 4 = x^3 - 2x + 5.$$

The expansion also reads off root information at 1: the constant term $f(1) = 4 \neq 0$, so 1 is not a root. In general, $\mathrm{mult}_c(f)$ is the index of the first nonzero coordinate of f in the basis $\mathcal{B}_c$ (over any field, by part (a); in characteristic 0 these coordinates are the Taylor coefficients), because $f = (x-c)^m(b_m + b_{m+1}(x-c) + \cdots)$ and the bracket is b_m at c.

Warning. **Taylor's formula needs $k! \neq 0$.** Over $\mathbb{F}_2$, take $f = x^2$ and $c = 0$. Then $f' = 2x = 0$, so **every** derivative of f, at every point, is 0, even though $f \neq 0$. No formula in the values $f^{(k)}(0)$ can recover the coefficient 1 of x^2. Part (a) of Theorem 6.4.7 still holds over $\mathbb{F}_2$: x^2 has coordinates $(0, 0, 1)$ in $(1, x, x^2)$. What fails is that these coordinates are derivatives divided by $k!$.

The two halves of this section meet in Chapter 9, where a matrix is diagonalizable exactly when its minimal polynomial splits and has no repeated roots; the gcd test above is how such a condition is checked in practice. Before that, the next section proves that over $\mathbb{C}$ the sum of the multiplicities always equals the degree.

6.4.6 Exercises

6.4.6.1 A. Check your understanding

Exercise. A1

(a) Define the multiplicity of c as a root of a nonzero $f \in F[x]$. Why is $f \neq 0$ required?

(b) True or false: if $f(c) = f'(c) = 0$, then c has multiplicity exactly 2. Justify your answer.

(c) True or false: over every field, $f' = 0$ implies that f is constant. Justify your answer.

(d) True or false: if $f \in \mathbb{R}[x]$ has degree 4 and the multiplicities of its real roots add up to

4, then f splits over $\mathbb{R}$. Justify your answer.

(e) True or false: if $f \in \mathbb{Q}[x]$ is nonconstant and $\gcd(f, f') \neq 1$, then f has a repeated irreducible factor. Justify your answer.

(f) Name the dual basis (Theorem 5.1.3) of $(1, x-c, \ldots, (x-c)^n)$ in $\mathbb{R}[x]_{\leq n}$.

Solution.

(a) It is the largest $m \in \mathbb{N}$ with $(x-c)^m \mid f$ (Definition 6.4.1). For $f = 0$, every power divides f, so no largest m exists.

(b) False. For $f = x^3$ and $c = 0$, $f(0) = f'(0) = 0$, but $\mathrm{mult}_0(f) = 3$. Theorem 6.4.5 (a) only gives multiplicity at least 2.

(c) False. Over $\mathbb{F}_2$, $(x^2)' = 2x = 0$. (In characteristic 0 it is true: $\deg f' = \deg f - 1$ whenever $\deg f \geq 1$.)

(d) True. By Theorem 6.4.3, the sum of multiplicities equals the degree if and only if f splits over $\mathbb{R}$.

(e) True. $\mathbb{Q}$ has characteristic 0, so this is Corollary 6.4.6 (b).

(f) $\varphi_k(f) = f^{(k)}(c)/k!$ for $k = 0, \ldots, n$, by Theorem 6.4.7 (b); $\mathbb{R}$ has characteristic 0.

6.4.6.2 B. Practice

Exercise. B1: Multiplicities over different fields

(a) Let $f = x^5 - 3x^4 + 4x^3 - 4x^2 + 3x - 1$. Find $\mathrm{mult}_1(f)$ in $\mathbb{Q}[x]$. Hence decide whether f splits over $\mathbb{R}$, and whether it splits over $\mathbb{C}$.

(b) In $\mathbb{F}_2[x]$, find the roots of $g = x^4 + x^2$ and their multiplicities. Does g split over $\mathbb{F}_2$?

Solution.

(a) Here $f(1) = 1 - 3 + 4 - 4 + 3 - 1 = 0$. Dividing by $x - 1$ repeatedly,

$$f = (x-1)(x^4 - 2x^3 + 2x^2 - 2x + 1),$$
$$x^4 - 2x^3 + 2x^2 - 2x + 1 = (x-1)(x^3 - x^2 + x - 1),$$

and $x^3 - x^2 + x - 1 = (x-1)(x^2+1)$. Since $x^2 + 1$ takes the value $2 \neq 0$ at 1, Lemma 6.4.1 gives $f = (x-1)^3(x^2+1)$ and $\mathrm{mult}_1(f) = 3$. Over $\mathbb{R}$, $x^2 + 1$ has no root ($c^2 + 1 \geq 1$), so the only real root of f is 1, and $3 < 5 = \deg f$. By Theorem 6.4.3, f does not split over $\mathbb{R}$. Over $\mathbb{C}$, $x^2 + 1 = (x-i)(x+i)$, so $f = (x-1)^3(x-i)(x+i)$ splits.

(b) In $\mathbb{F}_2[x]$, $x^4 + x^2 = x^2(x^2+1)$, and $x^2 + 1 = (x+1)^2$ because $(x+1)^2 = x^2 + 2x + 1 = x^2 + 1$. Hence $g = x^2(x+1)^2 = (x-0)^2(x-1)^2$, using $-1 = 1$. The cofactor $(x+1)^2$ is 1 at 0, and x^2 is 1 at 1, so $\mathrm{mult}_0(g) = \mathrm{mult}_1(g) = 2$ by Lemma 6.4.1. These are the only elements of $\mathbb{F}_2$, the multiplicities add to $4 = \deg g$, and g splits over $\mathbb{F}_2$.

Exercise. B2: A gcd test

Decide whether $f = x^4 - 2x^2 + 1 \in \mathbb{Q}[x]$ has a repeated root, by computing $\gcd(f, f')$ with the Euclidean algorithm. Hence factor f and give all multiplicities.

Solution. We have $f' = 4x^3 - 4x$. Dividing, $f = \frac{1}{4}x \cdot f' + (-x^2 + 1)$, since $\frac{1}{4}x(4x^3 - 4x) = x^4 - x^2$. Next, $f' = (-4x)(-x^2 + 1) + 0$. So the last nonzero remainder is $-x^2 + 1$, and by Theorem 6.2.6 $\gcd(f, f') = x^2 - 1 \neq 1$. Its roots 1 and -1 are common roots of f and f', since $x^2 - 1 = (x - 1)(x + 1)$ divides both. By Theorem 6.4.5 (a), 1 and -1 are multiple roots of f.

To factor, note that $(x^2 - 1)^2 = x^4 - 2x^2 + 1 = f$, so $f = (x - 1)^2(x + 1)^2$. By Lemma 6.4.1, $\mathrm{mult}_1(f) = 2$ (cofactor $(x + 1)^2$, value 4 at 1) and $\mathrm{mult}_{-1}(f) = 2$.

Exercise. B3: The multiplicity of 1

Let $n \geq 2$ and $f = x^n - nx + n - 1 \in \mathbb{Q}[x]$. Prove that 1 is a root of f of multiplicity exactly 2.

Solution. We have $f(1) = 1 - n + n - 1 = 0$, $f' = nx^{n-1} - n$, so $f'(1) = 0$, and $f'' = n(n - 1)x^{n-2}$, so $f''(1) = n(n - 1) \neq 0$ since $n \geq 2$. In particular $f \neq 0$. By Theorem 6.4.5 (a), $m := \mathrm{mult}_1(f) \geq 2$. Since $\mathbb{Q}$ has characteristic 0, part (b) gives $\mathrm{mult}_1(f') = m - 1 \geq 1$. Apply (a) to $f' \neq 0$: since $f''(1) \neq 0$, 1 is not a multiple root of f', so $m - 1 = 1$. Hence $m = 2$.

6.4.6.3 C. Going deeper

Exercise. C1: Real roots of the derivative

Let $f \in \mathbb{R}[x]$ have degree $n \geq 2$ and split over $\mathbb{R}$. Prove that f' splits over $\mathbb{R}$. You may use **Rolle's theorem** from calculus: if $a < b$ are real and $f(a) = f(b) = 0$, then $f'(\xi) = 0$ for some ξ with $a < \xi < b$.

Hint: count the roots of f' with multiplicity, and compare with $\deg f'$.

Solution. Let $c_1 < c_2 < \cdots < c_k$ be the distinct real roots of f, with multiplicities $m_1, \ldots, m_k$. Since f splits over $\mathbb{R}$, Theorem 6.4.3 gives $m_1 + \cdots + m_k = n$.

By Rolle's theorem, for each $i = 1, \ldots, k - 1$ there is ξ_i with $c_i < \xi_i < c_{i+1}$ and $f'(\xi_i) = 0$. These $k - 1$ numbers are distinct, since they lie in disjoint open intervals, and none of them is a c_j. Moreover, each c_i with $m_i \geq 2$ is a root of f' of multiplicity $m_i - 1$, by Theorem 6.4.5 (b) (characteristic 0); for $m_i = 1$ the count $m_i - 1 = 0$ is also correct.

Since $\deg f = n \geq 1$ and $\mathbb{R}$ has characteristic 0, $\deg f' = n - 1$, so $f' \neq 0$. Adding the multiplicities of the distinct roots $\xi_1, \ldots, \xi_{k-1}, c_1, \ldots, c_k$ of f' gives at least

$$(k - 1) + \sum_{i=1}^{k} (m_i - 1) = (k - 1) + n - k = n - 1 = \deg f'.$$

By Theorem 6.4.3 the sum of the multiplicities of all roots of f' is at most $\deg f'$, so it equals

$\deg f'$, and f' splits over $\mathbb{R}$.

Exercise. C2: $x^p - x$ over $\mathbb{F}_p$

Let p be a prime and $f = x^p - x \in \mathbb{F}_p[x]$.

(a) Compute f' and deduce that f has no repeated irreducible factor.

(b) Using Exercise (C2: A nonzero polynomial vanishing everywhere), deduce that $x^p - x = \prod_{a \in \mathbb{F}_p}(x - a)$.

(c) Explain why the argument of (a) gives nothing for $g = x^p - 1$. Then prove that $g = (x-1)^p$ in $\mathbb{F}_p[x]$.

Hint for (c): for $0 < k < p$, use the integer identity $k\binom{p}{k} = p\binom{p-1}{k-1}$.

Solution.

(a) By Definition 6.4.3, $f' = px^{p-1} - 1 = -1$, since $p \cdot 1 = 0$ in $\mathbb{F}_p$. Every common divisor of f and -1 divides 1, so $\gcd(f, f') = 1$, and Corollary 6.4.6 (a), which holds over every field, shows f has no repeated irreducible factor.

(b) By Exercise (C2: A nonzero polynomial vanishing everywhere) (c), every $a \in \mathbb{F}_p$ is a root of f, so each of the p elements has multiplicity at least 1. The multiplicities add up to at least $p = \deg f$, and at most $\deg f$ by Theorem 6.4.3. So they add up to exactly p, and each is 1. By Theorem 6.4.3, $f = \prod_{b \in \mathbb{F}_p}(x - b)\, g$ with $\deg g = p - p = 0$, so g is a constant; comparing leading coefficients (Theorem 0.8.3), $g = 1$.

(c) Here $g' = px^{p-1} = 0$, so $\gcd(g, g') = g \neq 1$ (the monic generator of the ideal generated by g and 0). Part (a) of Corollary 6.4.6 needs gcd = 1, so it says nothing, and part (b) is only available in characteristic 0. Still, $g(1) = 0$ and $g'(1) = 0$, so 1 is a multiple root by Theorem 6.4.5 (a).
For the factorization, expand with the binomial theorem (valid in the commutative ring $\mathbb{F}_p[x]$, Theorem 0.8.1): $(x-1)^p = \sum_{k=0}^{p} \binom{p}{k}(-1)^{p-k}x^k$, with the integers $\binom{p}{k}$ acting as integer multiples. For $0 < k < p$, the identity $k\binom{p}{k} = p\binom{p-1}{k-1}$ and Lemma 0.7.4 give $(k \cdot 1)(\binom{p}{k} \cdot 1) = (p \cdot 1)(\binom{p-1}{k-1} \cdot 1) = 0$ in $\mathbb{F}_p$. Since $0 < k < p$, $k \cdot 1 \neq 0$ by Definition 0.7.2, so $\binom{p}{k} \cdot 1 = 0$ (Theorem 0.7.1). Hence $(x-1)^p = x^p + (-1)^p$. For odd p this is $x^p - 1$; for $p = 2$ it is $x^2 + 1 = x^2 - 1$. So $g = (x-1)^p$, and $\mathrm{mult}_1(g) = p$ by Lemma 6.4.1.

Exercise. C3: Removing repeated factors

Let F be a field and $f = a\, p_1^{e_1} \cdots p_r^{e_r} \in F[x]$, where $a \in F \setminus \{0\}$, the p_i are **distinct** monic irreducible polynomials, and $e_i \geq 1$.

(a) Suppose F has characteristic 0. Prove that $\gcd(f, f') = p_1^{e_1 - 1} \cdots p_r^{e_r - 1}$, and deduce that $f / \gcd(f, f') = a\, p_1 \cdots p_r$.

(b) Over $\mathbb{F}_2$, let $f = x^2(x+1)$. Compute $f / \gcd(f, f')$ and show that the conclusion of (a) fails.

Hint for (a): write $D = p_1^{e_1 - 1} \cdots p_r^{e_r - 1}$. Show $D \mid \gcd(f, f')$; for the reverse, show $p_i^{e_i} \nmid f'$ using Euclid's lemma.

Solution.

(a) Let $D = \prod_i p_i^{e_i - 1}$ and $d = \gcd(f, f')$. By Theorem 6.4.4 (b) and (c), applied repeatedly,

$$f' = a \sum_{i=1}^{r} e_i \, p_i^{e_i - 1} p_i' \prod_{j \neq i} p_j^{e_j}.$$

Every term is divisible by D, and so is f. Hence $D \mid d$ by Theorem 6.2.3. Conversely, $d \mid f$, so $f = dh$. Factoring d and h into monic irreducibles and comparing with the factorization of f by the uniqueness in Theorem 6.3.5, $d = \prod_i p_i^{k_i}$ with $0 \leq k_i \leq e_i$ (as d is monic). Suppose $k_i = e_i$ for some i. Then $p_i^{e_i} \mid f'$. In the formula for f' every term with index $j \neq i$ contains $p_i^{e_i}$, so $p_i^{e_i}$ divides the remaining term $ae_i p_i^{e_i - 1} p_i' \prod_{j \neq i} p_j^{e_j}$. Canceling $p_i^{e_i - 1}$ (Corollary 0.8.4), $p_i \mid (ae_i)\, p_i' \prod_{j \neq i} p_j^{e_j}$. Here $ae_i = a(e_i \cdot 1)$ is a nonzero constant (characteristic 0), $p_i \nmid p_i'$ because $p_i' \neq 0$ has smaller degree, and $p_i \nmid p_j$ for $j \neq i$ because both are monic irreducible and distinct. Applying Theorem 6.3.3 repeatedly gives a contradiction. Hence $k_i \leq e_i - 1$ for all i, so $d \mid D$. Both are monic and divide each other, so they have the same degree and $d = D$. Finally $f = D \cdot a\, p_1 \cdots p_r$, so $f/d = a\, p_1 \cdots p_r$.

(b) In $\mathbb{F}_2[x]$, $f = x^3 + x^2$ and $f' = 3x^2 + 2x = x^2$. Since $x^2 \mid f$, $\gcd(f, f') = x^2$, and $f / \gcd(f, f') = x + 1$. The formula of (a) would predict $\gcd(f, f') = x$ and $f / \gcd(f, f') = x(x + 1)$. The irreducible factor x has been lost. The step that fails is "$ae_i \neq 0$": here $e_1 = 2$ and $2 \cdot 1 = 0$ in $\mathbb{F}_2$.

6.5 The Fundamental Theorem of Algebra

Chapter 0 stated the Fundamental Theorem of Algebra (Theorem 0.6.7) and promised a proof. Everything algebraic that the proof needs is now in place: division, the Factor Theorem, multiplicities, and the polar form of complex numbers with De Moivre's formula. What algebra alone cannot supply is a property of the real numbers that goes beyond the field axioms, and some such property must enter, because $\mathbb{Q}$ is a field too and there $x^2 - 2$ has no root. We use completeness, in the form of the extreme value theorem. This section isolates the analytic input in two clearly labeled facts, proves everything else in full, and then draws the consequences for real polynomials: conjugate pairs, irreducible factors of degree at most 2, and real roots of odd-degree polynomials.

6.5.1 The goal and the plan

Here is the theorem we want, in the language of this chapter. A polynomial $p \in \mathbb{C}[x]$ of degree $n \geq 1$ is exactly the expression $a_n x^n + \cdots + a_0$ with $a_n \neq 0$ of Theorem 0.6.7.

Theorem 6.5.1. Fundamental Theorem of Algebra

Every nonconstant polynomial $p \in \mathbb{C}[x]$ has a root in $\mathbb{C}$.

We prove it after three lemmas. The proof is a minimum argument, and it runs in three steps.

① **Growth.** Far from the origin, $|p(z)|$ is large, because the leading term $a_n z^n$ outweighs all the others. We will prove this with an explicit radius.

② **A minimum exists.** On a large closed disc, the continuous function $z \mapsto |p(z)|$ attains a smallest value. By ①, nothing outside the disc does better, so this is the smallest value of $|p|$ on all of $\mathbb{C}$.

③ **A nonzero value is never the minimum.** If $p(z_0) \neq 0$, we can move from z_0 a little in a well-chosen direction and make $|p|$ strictly smaller. This is d'Alembert's lemma, and it is pure algebra of $\mathbb{C}$.

Putting ② and ③ together, the minimum value must be 0, and the point where it is attained is a root.

Steps ① and ③ are proved from scratch. Step ② is where analysis enters, and we say exactly how.

Remark. **What we assume from analysis.** The proof uses exactly two facts from real analysis, and nothing else: no limits, no derivatives, no intermediate value theorem. We identify $\mathbb{C}$ with $\mathbb{R}^2$ as in Definition 0.6.1, so that $|z-w|$ is the usual distance in the plane.

(A1) **Continuity of polynomial functions.** For every $p \in \mathbb{C}[x]$, the function $\mathbb{C} \to \mathbb{R}$, $z \mapsto |p(z)|$, is continuous.

(A2) **Extreme value theorem on a closed disc.** Let $\rho \geq 0$ and $D = \{z \in \mathbb{C} : |z| \leq \rho\}$. If $g\colon D \to \mathbb{R}$ is continuous, then there is $w \in D$ with $g(w) \leq g(z)$ for **every** $z \in D$.

Fact (A2) is where the completeness of $\mathbb{R}$ is used. Everything else below is the arithmetic of $\mathbb{C}$ from Chapter 0, including polar form and Theorem 0.6.6.

6.5.2 Step 1: growth

We use three small facts about the modulus repeatedly. For $z, w, z_1, \dots, z_m \in \mathbb{C}$ and $j \in \mathbb{N}$:

$$|z^j| = |z|^j, \qquad |z_1 + \cdots + z_m| \leq |z_1| + \cdots + |z_m|, \qquad |z+w| \geq |z| - |w|.$$

The first follows from Theorem 0.6.2 (d) by induction on j, and the second from Theorem 0.6.4 by induction on m. For the third, apply the triangle inequality to $z = (z+w) + (-w)$ and use $|-w| = |w|$.

The leading term wins as soon as $|z|$ exceeds a radius computed from the coefficients.

Lemma 6.5.2. Growth of a Polynomial

Let $p = a_n x^n + \cdots + a_1 x + a_0 \in \mathbb{C}[x]$ with $n \geq 1$ and $a_n \neq 0$. Put

$$A = |a_0| + |a_1| + \cdots + |a_{n-1}|, \qquad R = \max\Big\{1, \frac{2A}{|a_n|}\Big\}.$$

Then for every $z \in \mathbb{C}$ with $|z| \geq R$,

$$|p(z)| \geq \tfrac{1}{2}|a_n|\,|z|^n \geq \tfrac{1}{2}|a_n|\,|z|.$$

In particular, for every real $M \geq 0$, we have $|p(z)| > M$ whenever $|z| \geq R$ and $|z| > 2M/|a_n|$.

Idea. Compare the leading term with the rest. When $|z| \geq 1$, every lower power $|z|^j$ is at most $|z|^{n-1}$, so the rest is at most $A|z|^{n-1}$, while the leading term is $|a_n||z|^n$. The rest is at most half the leading term once $A \leq \frac{1}{2}|a_n||z|$, which is how R was found.

Proof. Let $|z| \geq R$. Since $|z| \geq 1$, we have $|z|^j \leq |z|^{n-1}$ for $0 \leq j \leq n-1$. By the triangle inequality (Theorem 0.6.4),

$$\Big|\sum_{j=0}^{n-1} a_j z^j\Big| \leq \sum_{j=0}^{n-1} |a_j||z|^j \leq A|z|^{n-1}.$$

Since $|z| \geq 2A/|a_n|$, we have $A \leq \frac{1}{2}|a_n||z|$. By the reverse triangle inequality,

$$|p(z)| \geq |a_n||z|^n - A|z|^{n-1} \geq |a_n||z|^n - \tfrac{1}{2}|a_n||z|^n = \tfrac{1}{2}|a_n||z|^n.$$

Since $|z| \geq 1$ and $n \geq 1$, $|z|^n \geq |z|$, which gives the second inequality. Finally, if also $|z| > 2M/|a_n|$, then $|p(z)| \geq \frac{1}{2}|a_n||z| > M$. This proves the lemma.

■

For example, for $p = x^3 - 5x + 1$ we get $A = 6$ and $R = 12$, so $|p(z)| \geq \frac{1}{2}|z|^3$ whenever $|z| \geq 12$. The radius is far from optimal, and it does not need to be. We only need **some** radius beyond which $|p|$ is large.

6.5.3 Step 2: the minimum is attained

Lemma 6.5.3. The Minimum of $|p|$ Is Attained

Let $p \in \mathbb{C}[x]$. Then there is $z_0 \in \mathbb{C}$ with $|p(z_0)| \leq |p(z)|$ for **every** $z \in \mathbb{C}$.

Idea. Use (A2) on a disc so big that outside it, $|p|$ exceeds $|p(0)|$. The disc contains 0, so its minimum is at most $|p(0)|$, and the points outside cannot compete. Lemma 6.5.2 tells us how big "so big" is.

Proof. If p is constant, every z_0 works. Suppose $p = a_n x^n + \cdots + a_0$ with $n \geq 1$ and $a_n \neq 0$, and let R be the radius of Lemma 6.5.2. Put $M = |p(0)|$ and

$$\rho = R + \frac{2M}{|a_n|}, \qquad D = \{z \in \mathbb{C} : |z| \leq \rho\}.$$

By (A1), $z \mapsto |p(z)|$ is continuous on D, so by (A2) there is $z_0 \in D$ with $|p(z_0)| \leq |p(z)|$ for every $z \in D$.

Now let $z \notin D$, so $|z| > \rho$. Then $|z| > R$, and $|z| > 2M/|a_n|$ because $R \geq 1 > 0$. By Lemma 6.5.2, $|p(z)| > M = |p(0)|$. Since $0 \in D$, $|p(0)| \geq |p(z_0)|$. Hence $|p(z)| > |p(z_0)|$ for

$z \notin D$, and $|p(z)| \geq |p(z_0)|$ for $z \in D$. This proves the lemma.

■

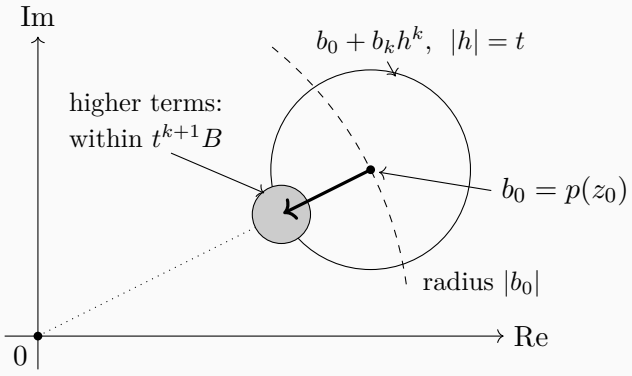
6.5.4 Step 3: d'Alembert's lemma

It remains to show that a nonzero value of $|p|$ is never the smallest one. Near z_0, write $p(z_0+h) = b_0 + b_k h^k +$ (terms of higher degree in h), where $b_0 = p(z_0)$ and $b_k h^k$ is the first nonconstant term that is actually present. For small h, the term $b_k h^k$ dominates the later ones, so we choose the direction of h to make $b_k h^k$ point from b_0 straight toward 0.

Lemma 6.5.4. d'Alembert's Lemma

Let $p \in \mathbb{C}[x]$ be nonconstant, and let $z_0 \in \mathbb{C}$ with $p(z_0) \neq 0$. Then there is $z \in \mathbb{C}$ with $|p(z)| < |p(z_0)|$.

Idea. Picture the values $p(z_0 + h)$ in the plane. As $h = tu$ runs over a small circle $|h| = t$, the term $b_0 + b_k h^k$ runs k times around the circle of radius $t^k|b_k|$ centered at $b_0 = p(z_0)$. One point of that circle lies on the segment from b_0 to 0, at distance $|b_0| - t^k|b_k|$ from 0. Reaching it needs a k-th root of a unit complex number, which De Moivre supplies. The higher terms move the value by at most a multiple of t^{k+1}, a blur much smaller than the gain $t^k|b_k|$ when t is small.



The dashed arc is part of the circle of radius $|b_0|$ about 0, the old value of $|p|$. The solid circle shows where $b_0 + b_k h^k$ goes for $|h| = t$. The arrow is the chosen direction, and the shaded disc is the uncertainty caused by the higher terms. The whole shaded disc lies strictly inside the dashed circle, which is the inequality $|p(z_0 + h)| < |p(z_0)|$.

Proof. Let $n = \deg p \geq 1$.

Step 1: re-center at z_0. By Theorem 6.4.7 (a), $(1, x - z_0, \ldots, (x - z_0)^n)$ is a basis of $\mathbb{C}[x]_{\leq n}$, so $p = \sum_{j=0}^{n} b_j (x - z_0)^j$ for some $b_0, \ldots, b_n \in \mathbb{C}$. Evaluating at z_0 gives $b_0 = p(z_0) \neq 0$. The coefficient of x^n on the right is b_n, since $(x - z_0)^j$ has degree j and $(x - z_0)^n$ is monic; so b_n is the leading coefficient of p, and $b_n \neq 0$. By Theorem 0.8.5, for every $h \in \mathbb{C}$,

$$p(z_0 + h) = \sum_{j=0}^{n} b_j h^j.$$

Since $b_n \neq 0$ and $n \geq 1$, there is a **smallest** $k \geq 1$ with $b_k \neq 0$. Then $b_1 = \cdots = b_{k-1} = 0$, and

$$p(z_0 + h) = b_0 + b_k h^k + \sum_{j=k+1}^{n} b_j h^j. \tag{$*$}$$

Step 2: choose the direction. Since $b_0 \neq 0$ and $b_k \neq 0$, the number $w = -b_0/b_k$ is nonzero. Write it in polar form, $w = |w|e^{i\alpha}$ with $\alpha \in \mathbb{R}$, and put $u = e^{i\alpha/k}$. Then $|u| = 1$, and $u^k = e^{i\alpha}$ by Theorem 0.6.6. Using $b_k w = -b_0$ and $|w| = |b_0|/|b_k|$ (Theorem 0.6.2 (d)),

$$b_k u^k = \frac{b_k w}{|w|} = -\frac{|b_k|}{|b_0|}\, b_0.$$

Hence, for every real $t > 0$,

$$b_0 + b_k (tu)^k = b_0\Big(1 - t^k \frac{|b_k|}{|b_0|}\Big).$$

Step 3: choose the size. Put $B = |b_{k+1}| + \cdots + |b_n|$ (so $B = 0$ if $k = n$), and

$$t = \min\Big\{1,\ \frac{|b_0|}{|b_k|},\ \frac{|b_k|}{2B}\Big\},$$

where the last entry is omitted if $B = 0$. Then $t > 0$, and t has three properties.

(i) $0 < t^k |b_k|/|b_0| \leq 1$. Indeed, if $|b_0|/|b_k| \geq 1$, then $t^k \leq 1 \leq |b_0|/|b_k|$; otherwise $t \leq |b_0|/|b_k| < 1$, and $t^k \leq t \leq |b_0|/|b_k|$.

(ii) $tB \leq \frac{1}{2}|b_k|$. This is trivial if $B = 0$, and follows from $t \leq |b_k|/(2B)$ otherwise.

(iii) $t^j \leq t^{k+1}$ for every $j \geq k+1$, since $0 < t \leq 1$.

Step 4: estimate. Let $z = z_0 + tu$. By Step 2 and (i), the number $1 - t^k|b_k|/|b_0|$ is real and lies in $[0, 1)$, so

$$|b_0 + b_k(tu)^k| = |b_0|\Big(1 - t^k \frac{|b_k|}{|b_0|}\Big) = |b_0| - t^k |b_k|.$$

By the triangle inequality (Theorem 0.6.4), $|u| = 1$ and (iii),

$$\Big|\sum_{j=k+1}^{n} b_j (tu)^j\Big| \leq \sum_{j=k+1}^{n} |b_j| t^j \leq t^{k+1} B.$$

Combining these with $(*)$, the triangle inequality, and (ii),

$$|p(z)| \leq |b_0| - t^k|b_k| + t^{k+1}B = |b_0| - t^k(|b_k| - tB) \leq |b_0| - \tfrac{1}{2}t^k|b_k| < |b_0|,$$

where the last step uses $t > 0$ and $b_k \neq 0$. Since $|b_0| = |p(z_0)|$, this proves the lemma.

■

Every hypothesis did a job. Nonconstant gave the index k. The condition $p(z_0) \neq 0$ let us divide by $|b_0|$ and fix a direction toward 0. And the field $\mathbb{C}$ supplied the k-th root u, the one step that has no analogue over $\mathbb{R}$.

Example. One Step Downhill

Let $p = x^3 + x^2 + 1 \in \mathbb{C}[x]$ and $z_0 = 0$. Follow the proof of Lemma 6.5.4 to find z with $|p(z)| < |p(0)|$.

Solution. Here $p(0) = 1$ and the expansion at $z_0 = 0$ is p itself: $b_0 = 1$, $b_1 = 0$, $b_2 = 1$, $b_3 = 1$. The first nonzero b_j with $j \geq 1$ is b_2, so $k = 2$ and $B = |b_3| = 1$.

Direction: $w = -b_0/b_2 = -1 = e^{i\pi}$, so $\alpha = \pi$ and $u = e^{i\pi/2} = i$. Indeed $b_2 u^2 = -1$ points from $b_0 = 1$ toward 0.

Size: $t = \min\{1,\ 1/1,\ 1/(2 \cdot 1)\} = \frac{1}{2}$. So $z = \frac{i}{2}$, and

$$p(\tfrac{i}{2}) = -\tfrac{i}{8} - \tfrac{1}{4} + 1 = \tfrac{3}{4} - \tfrac{i}{8}, \qquad |p(\tfrac{i}{2})| = \sqrt{\tfrac{9}{16} + \tfrac{1}{64}} = \frac{\sqrt{37}}{8} \approx 0.76 < 1.$$

The proof guarantees $|p(z)| \leq 1 - \frac{1}{2} \cdot \frac{1}{4} \cdot 1 = \frac{7}{8}$, and $\sqrt{37}/8 \leq \frac{7}{8}$ since $37 \leq 49$. The direction matters: along the real axis, $p(\frac{1}{2}) = \frac{1}{8} + \frac{1}{4} + 1 = \frac{11}{8} > 1$, because there $b_2 h^2 = \frac{1}{4}$ points away from 0.

6.5.5 The proof of the theorem

The three steps now combine into a proof of Theorem 6.5.1.

Proof. Let $p \in \mathbb{C}[x]$ be nonconstant. By Lemma 6.5.3 there is $z_0 \in \mathbb{C}$ with $|p(z_0)| \leq |p(z)|$ for every $z \in \mathbb{C}$. Suppose, for a contradiction, that $p(z_0) \neq 0$. By Lemma 6.5.4 there is z with $|p(z)| < |p(z_0)|$, contradicting the choice of z_0. Hence $p(z_0) = 0$, and z_0 is a root of p. ∎

This proves Chapter 0's Theorem 0.6.7: a polynomial $a_n z^n + \cdots + a_0$ with $n \geq 1$ and $a_n \neq 0$ is a nonconstant element of $\mathbb{C}[x]$, and we have just shown that it has a root.

Quick check. The polynomial $p = x^2 + 1 \in \mathbb{R}[x]$ has no real root, yet the function $c \mapsto |p(c)|$ on $\mathbb{R}$ has a minimum, at $c = 0$. Which step of the argument breaks if we try to run it over $\mathbb{R}$?

Solution. Growth and the attained minimum both hold on $\mathbb{R}$: the minimum value is $|p(0)| = 1$. The step that fails is d'Alembert's lemma. At $z_0 = 0$ we have $b_0 = 1$, $k = 2$, $b_2 = 1$, so we need a direction u with $u^2 = w = -1$. There is no such real u, and indeed $p(h) = 1 + h^2 \geq 1$ for every real h. Over $\mathbb{C}$, $u = i$ works and $p(ti) = 1 - t^2 < 1$ for $0 < t \leq 1$.

Warning. **The theorem is an existence theorem.** It says a root exists; it does not say how to find one, and the proof does not produce one. There is no formula for the roots of a general polynomial of degree 5 or more built from the coefficients by $+, -, \times, \div$ and

k-th roots (a theorem from Galois theory, which we do not prove). Numerical root-finding algorithms approximate roots; that is a different subject, and the Fundamental Theorem of Algebra is not one of those algorithms.

6.5.6 Complex polynomials split

Combining the theorem with the root count of the previous section gives the form in which it is used most.

Corollary 6.5.5. Complex Polynomials Split

(a) Every nonzero $p \in \mathbb{C}[x]$ splits over $\mathbb{C}$: if $\deg p = n$ and p has leading coefficient a, then $p = a(x-z_1)(x-z_2)\cdots(x-z_n)$ for some $z_1, \ldots, z_n \in \mathbb{C}$. In particular the multiplicities of the roots of p add up to n.

(b) The irreducible polynomials in $\mathbb{C}[x]$ are exactly the polynomials of degree 1.

Proof.

(a) By Theorem 6.4.3, $p = (x-c_1)^{m_1}\cdots(x-c_k)^{m_k}\,g$, where $c_1, \ldots, c_k$ are the distinct roots of p in $\mathbb{C}$ and $g \in \mathbb{C}[x]$ has no root in $\mathbb{C}$. By Theorem 6.5.1, g is constant, and $g \neq 0$ because $p \neq 0$. So p splits over $\mathbb{C}$, and the multiplicities add up to n by the same theorem. Comparing leading coefficients (Theorem 0.8.3), the constant is a.

(b) A polynomial of degree 1 is irreducible, since a factorization into two nonconstant factors would have degree at least 2. Conversely, let $p \in \mathbb{C}[x]$ be irreducible. It is nonconstant (Definition 6.3.1), so by (a) it has a factor $x - z_1$, say $p = (x-z_1)q$. Since p is irreducible and $x - z_1$ is not constant, q is a nonzero constant, and $\deg p = 1$.

■

So in $\mathbb{C}[x]$, unique factorization (Theorem 6.3.5) is completely explicit: the monic irreducible factors are the $x-z$, and the exponent of $x-z$ is the multiplicity of z. In Chapter 9 this is what gives every operator on a nonzero finite-dimensional complex vector space an eigenvalue. (That fact also has a proof using only linear algebra, and one can run the logic in the other direction; we will not need it.)

Example. Factoring x^4+1 over $\mathbb{C}$ and over $\mathbb{R}$

Factor $f = x^4 + 1$ into irreducible polynomials in $\mathbb{C}[x]$ and in $\mathbb{R}[x]$.

Solution. Roots. We solve $z^4 = -1 = e^{i\pi}$ as in Example (Roots of Unity). Write $z = re^{i\theta}$ with $r > 0$; by Theorem 0.6.6, $z^4 = r^4 e^{4i\theta}$. Comparing polar forms, $r^4 = 1$, so $r = 1$, and $4\theta = \pi + 2\pi j$ for an integer j. The values $j = 0, 1, 2, 3$ give four distinct roots

$$z_0 = e^{i\pi/4} = \tfrac{\sqrt{2}}{2} + \tfrac{\sqrt{2}}{2}i, \quad z_1 = e^{3i\pi/4} = -\tfrac{\sqrt{2}}{2} + \tfrac{\sqrt{2}}{2}i, \quad z_2 = \overline{z_1}, \quad z_3 = \overline{z_0}.$$

Over $\mathbb{C}$. Four distinct roots, each of multiplicity at least 1, and $\deg f = 4$, so by Theo-

rem 6.4.3 each multiplicity is 1 and

$$x^4 + 1 = (x - z_0)(x - z_1)(x - z_2)(x - z_3),$$

the leading coefficient being 1. The factors are irreducible by Corollary 6.5.5 (b).

Over $\mathbb{R}$. Pair each root with its conjugate. For any $z \in \mathbb{C}$,

$$(x - z)(x - \overline{z}) = x^2 - (z + \overline{z})x + z\overline{z} = x^2 - 2\,\mathrm{Re}(z)\,x + |z|^2 \in \mathbb{R}[x],$$

by Theorem 0.6.2 (c). With $\mathrm{Re}\, z_0 = \frac{\sqrt{2}}{2}$, $\mathrm{Re}\, z_1 = -\frac{\sqrt{2}}{2}$, and $|z_0| = |z_1| = 1$,

$$x^4 + 1 = (x^2 - \sqrt{2}\,x + 1)(x^2 + \sqrt{2}\,x + 1).$$

Check: the product is $(x^2+1)^2 - 2x^2 = x^4 + 1$. Each quadratic has nonreal roots only, hence no real root, so it is irreducible over $\mathbb{R}$ by Theorem 6.3.1. Note that $x^4 + 1$ has no real root but is **reducible** over $\mathbb{R}$.

6.5.7 Real polynomials

A real polynomial is also a complex polynomial, so the theorem applies to it; the roots it provides may be nonreal. They are not scattered arbitrarily. Chapter 0 observed that nonreal roots of a real polynomial come in conjugate pairs; we now show that the pairs also share their multiplicity. For $g = \sum_j c_j x^j \in \mathbb{C}[x]$ write $\overline{g} = \sum_j \overline{c_j}\, x^j$, the polynomial with conjugated coefficients.

Theorem 6.5.6. Conjugate Roots Have Equal Multiplicity

Let $f \in \mathbb{R}[x]$ be nonzero, regarded as an element of $\mathbb{C}[x]$, and let $z \in \mathbb{C}$. Then $f(\overline{z}) = \overline{f(z)}$. In particular z is a root of f if and only if $\overline{z}$ is, and

$$\mathrm{mult}_{\overline{z}}(f) = \mathrm{mult}_z(f).$$

Idea. Conjugation respects sums and products of complex numbers, so applying it to all coefficients respects sums and products of polynomials. A real polynomial is unchanged by it. So conjugating a factorization $f = (x - z)^m g$ gives a factorization $f = (x - \overline{z})^m \overline{g}$, and the cofactor test transfers.

Proof. Let $g = \sum_j c_j x^j$ and $h = \sum_j d_j x^j$ be in $\mathbb{C}[x]$, and $z \in \mathbb{C}$. By Theorem 0.6.2 (a) and (b), the coefficient of x^k in $\overline{gh}$ is $\overline{\sum_{i+j=k} c_i d_j} = \sum_{i+j=k} \overline{c_i}\,\overline{d_j}$, which is the coefficient of x^k in $\overline{g}\,\overline{h}$. So $\overline{gh} = \overline{g}\,\overline{h}$. By the same parts, $\overline{g}(\overline{z}) = \sum_j \overline{c_j}\,\overline{z}^j = \overline{\sum_j c_j z^j} = \overline{g(z)}$.

Since f has real coefficients, $\overline{f} = f$. Hence $f(\overline{z}) = \overline{f}(\overline{z}) = \overline{f(z)}$, and $f(z) = 0$ if and only if $f(\overline{z}) = 0$.

Let $m = \mathrm{mult}_z(f)$. By Lemma 6.4.1, $f = (x - z)^m g$ with $g \in \mathbb{C}[x]$ and $g(z) \neq 0$. Conjugating, and using $\overline{x - z} = x - \overline{z}$,

$$f = \overline{f} = (x - \overline{z})^m\, \overline{g}, \qquad \overline{g}(\overline{z}) = \overline{g(z)} \neq 0.$$

By Lemma 6.4.1 again, $\operatorname{mult}_{\overline{z}}(f) = m$. This proves the theorem.

■

For instance, a real polynomial with $2+i$ as a root of multiplicity 3 also has $2-i$ as a root of multiplicity 3, so by Theorem 6.4.3 its degree is at least 6. Now the irreducible real polynomials can be listed.

Corollary 6.5.7. Irreducible Real Polynomials

The irreducible polynomials in $\mathbb{R}[x]$ are exactly the polynomials of degree 1, and the polynomials of degree 2 with no real root. In particular, every irreducible polynomial in $\mathbb{R}[x]$ has degree 1 or 2.

Idea. An irreducible real p has a complex root z by the theorem. If z is real, $x-z$ is a real factor and p is linear. If not, $\overline{z}$ is another root, and $(x-z)(x-\overline{z})$ is a **real** quadratic dividing p. The only subtle point is that it divides p in $\mathbb{R}[x]$ and not merely in $\mathbb{C}[x]$; uniqueness of division settles that.

Proof. ($\Leftarrow$) A polynomial of degree 1 is irreducible, since a product of two nonconstant polynomials has degree at least 2 (Theorem 0.8.3). A polynomial of degree 2 with no real root is irreducible by Theorem 6.3.1.

($\Rightarrow$) Let $p \in \mathbb{R}[x]$ be irreducible. It is nonconstant, so by Theorem 6.5.1 it has a root $z \in \mathbb{C}$.

Case 1: $z \in \mathbb{R}$. By Lemma 3.7.1 in $\mathbb{R}[x]$, $p = (x-z)q$ with $q \in \mathbb{R}[x]$. Since p is irreducible and $x-z$ is not constant, q is a nonzero constant, so $\deg p = 1$.

Case 2: $z \notin \mathbb{R}$. Then $\overline{z} \neq z$, and $\overline{z}$ is also a root by Theorem 6.5.6. Applying Corollary 6.1.4 in $\mathbb{C}[x]$ to the two distinct roots $z, \overline{z}$ gives $p = (x-z)(x-\overline{z})\,h$ for some $h \in \mathbb{C}[x]$. Put $q = (x-z)(x-\overline{z}) = x^2 - 2\operatorname{Re}(z)x + |z|^2$, which lies in $\mathbb{R}[x]$. Divide in $\mathbb{R}[x]$ (Theorem 6.1.1): $p = qh_1 + r$ with $h_1, r \in \mathbb{R}[x]$ and $\deg r < 2$. This is also a division with remainder in $\mathbb{C}[x]$, and so is $p = qh + 0$. By the uniqueness in Theorem 6.1.1 over $\mathbb{C}$, $r = 0$ and $h = h_1 \in \mathbb{R}[x]$. So $p = qh_1$ in $\mathbb{R}[x]$. Since p is irreducible and q is not constant, h_1 is a nonzero constant, and $\deg p = 2$. Finally, p has no real root, because a real root c would give a factor $x-c$ and $p = (x-c)q'$ with $\deg q' = 1$, contradicting irreducibility. This proves the corollary.

■

Remark. Completing the square decides which quadratics occur: $x^2 + \beta x + \gamma = \left(x + \frac{\beta}{2}\right)^2 - \frac{\beta^2 - 4\gamma}{4}$ has a real root if and only if $\beta^2 - 4\gamma \geq 0$. So the monic irreducible real quadratics are those with $\beta^2 < 4\gamma$.

Unique factorization in $\mathbb{R}[x]$ now takes a concrete form.

Corollary 6.5.8. Factorization over $\mathbb{R}$

Every nonconstant $f \in \mathbb{R}[x]$ can be written as

$$f = a\,(x - c_1)\cdots(x - c_r)\;q_1 \cdots q_s,$$

where $a \in \mathbb{R}$ is the leading coefficient of f, $c_1, \ldots, c_r \in \mathbb{R}$ (not necessarily distinct), and $q_1, \ldots, q_s \in \mathbb{R}[x]$ are monic quadratics with no real root. Then $r + 2s = \deg f$, and r is the sum of the multiplicities of the real roots of f. The factorization is unique up to the order of the factors.

Proof. By Theorem 6.3.5, $f = a\,p_1 \cdots p_k$ with $p_i \in \mathbb{R}[x]$ monic irreducible, unique up to order. By Corollary 6.5.7 each p_i is either $x - c$ with $c \in \mathbb{R}$ or a monic quadratic with no real root; list the first kind as $x - c_1, \ldots, x - c_r$ and the second as $q_1, \ldots, q_s$. Then $\deg f = r + 2s$ by Theorem 0.8.3. For a real number c, each q_j has multiplicity 0 at c, and $x - c_i$ has multiplicity 1 or 0 according as $c_i = c$ or not; by Theorem 6.4.2, $\mathrm{mult}_c(f)$ is the number of indices i with $c_i = c$. Adding over the real roots of f gives r. ∎

The last consequence is one that calculus proves with the intermediate value theorem. Here it falls out of counting degrees.

Corollary 6.5.9. Odd Degree Forces a Real Root

Every $f \in \mathbb{R}[x]$ of odd degree has a root in $\mathbb{R}$.

Proof. By Corollary 6.5.8, $\deg f = r + 2s$. Since $\deg f$ is odd, r is odd, so $r \geq 1$, and c_1 is a real root of f. ∎

Warning. **No real root does not mean irreducible over $\mathbb{R}$.** By Corollary 6.5.7, an irreducible real polynomial has degree at most 2; so $x^4 + 1$, which has no real root, is still reducible over $\mathbb{R}$ (Example (Factoring $x^4 + 1$ over $\mathbb{C}$ and over $\mathbb{R}$)). The test "reducible if and only if it has a root" is only valid in degrees 2 and 3 (Theorem 6.3.1).

6.5.8 Exercises

6.5.8.1 A. Check your understanding

Exercise. A1

(a) State the Fundamental Theorem of Algebra, and list the two facts from analysis used in its proof in this section.

(b) True or false: every polynomial in $\mathbb{R}[x]$ of degree 4 has a real root. Justify your answer.

(c) True or false: if $f \in \mathbb{R}[x]$ has $2 + i$ as a root of multiplicity 3, then $\deg f \geq 6$. Justify your answer.

(d) True or false: every polynomial in $\mathbb{R}[x]$ of degree 5 has a real root. Justify your answer.

(e) True or false: every irreducible polynomial in $\mathbb{F}_2[x]$ has degree at most 2. Justify your answer.

(f) In the proof of d'Alembert's lemma (Lemma 6.5.4), which property of $\mathbb{C}$ is used that $\mathbb{R}$ lacks?

Solution.

(a) Every nonconstant $p \in \mathbb{C}[x]$ has a root in $\mathbb{C}$ (Theorem 6.5.1). The analytic facts are (A1) continuity of $z \mapsto |p(z)|$ on $\mathbb{C}$, and (A2) the extreme value theorem for continuous real functions on a closed disc.

(b) False. $x^4 + 1 \geq 1$ at every real number, so it has no real root.

(c) True. By Theorem 6.5.6, $2 - i$ is also a root of multiplicity 3, and $2 - i \neq 2 + i$. By Theorem 6.4.3, $\deg f \geq 3 + 3 = 6$.

(d) True, by Corollary 6.5.9.

(e) False. $x^3 + x + 1$ takes the value 1 at both 0 and 1, so it has no root in $\mathbb{F}_2$, and a cubic without a root is irreducible by Theorem 6.3.1. Corollary 6.5.7 is special to $\mathbb{R}$.

(f) Every nonzero complex number has a k-th root; the proof takes a k-th root of a complex number of modulus 1 by Theorem 0.6.6. Over $\mathbb{R}$, -1 has no square root.

6.5.8.2 B. Practice

Exercise. B1: $x^4 + 4$

Find all complex roots of $x^4 + 4$, and factor it into irreducible polynomials in $\mathbb{C}[x]$ and in $\mathbb{R}[x]$.

Solution. Write $z = re^{i\theta}$ with $r > 0$. By Theorem 0.6.6, $z^4 = r^4 e^{4i\theta}$, and $-4 = 4e^{i\pi}$. Comparing polar forms, $r^4 = 4$, so $r = \sqrt{2}$ (the only positive real number with fourth power 4, since $t \mapsto t^4$ is strictly increasing on positive reals), and $4\theta = \pi + 2\pi j$, $j \in \mathbb{Z}$. The values $j = 0, 1, 2, 3$ give the four distinct roots

$$\sqrt{2}e^{i\pi/4} = 1 + i, \quad \sqrt{2}e^{3i\pi/4} = -1 + i, \quad \sqrt{2}e^{5i\pi/4} = -1 - i, \quad \sqrt{2}e^{7i\pi/4} = 1 - i.$$

By Theorem 6.4.3, four distinct roots of a degree-4 monic polynomial each have multiplicity 1, and

$$x^4 + 4 = (x - 1 - i)(x - 1 + i)(x + 1 - i)(x + 1 + i) \quad \text{in } \mathbb{C}[x].$$

Pairing conjugates with $(x - z)(x - \overline{z}) = x^2 - 2\,\mathrm{Re}(z)x + |z|^2$, and $|1 \pm i|^2 = |-1 \pm i|^2 = 2$,

$$x^4 + 4 = (x^2 - 2x + 2)(x^2 + 2x + 2) \quad \text{in } \mathbb{R}[x].$$

Check: $(x^2+2)^2-(2x)^2 = x^4+4x^2+4-4x^2 = x^4+4$. Both quadratics have $\beta^2-4\gamma = 4-8 < 0$, so they have no real roots and are irreducible by Corollary 6.5.7.

Exercise. B2: $x^6 + 1$

Find the factorization of x^6+1 into irreducible polynomials in $\mathbb{R}[x]$.

Solution. The roots solve $z^6 = -1 = e^{i\pi}$. As in B1, $z = e^{i\theta}$ with $6\theta = \pi + 2\pi j$, and $j = 0, \dots, 5$ give six distinct roots at the angles $30°, 90°, 150°, 210°, 270°, 330°$:

$$\tfrac{\sqrt{3}}{2} \pm \tfrac{1}{2}i, \qquad \pm i, \qquad -\tfrac{\sqrt{3}}{2} \pm \tfrac{1}{2}i.$$

Each has multiplicity 1 by Theorem 6.4.3, and none is real. They form three conjugate pairs, each of modulus 1, with real parts $\frac{\sqrt{3}}{2}$, 0, $-\frac{\sqrt{3}}{2}$. Pairing them as in Example (Factoring x^4+1 over $\mathbb{C}$ and over $\mathbb{R}$),

$$x^6 + 1 = (x^2+1)(x^2 - \sqrt{3}\,x + 1)(x^2 + \sqrt{3}\,x + 1).$$

Check: $(x^2 - \sqrt{3}x + 1)(x^2 + \sqrt{3}x + 1) = (x^2+1)^2 - 3x^2 = x^4 - x^2 + 1$, and $(x^2+1)(x^4 - x^2 + 1) = x^6 + 1$. Each quadratic factor has only nonreal roots, so it is irreducible over $\mathbb{R}$ by Corollary 6.5.7. By the uniqueness in Corollary 6.5.8, this is the factorization.

Exercise. B3: Real cubics

Let $f \in \mathbb{R}[x]$ have degree 3, and suppose the multiplicities of the real roots of f add up to 1.

(a) Prove that $f = a(x-c)(x-z)(x-\overline{z})$ for some $a, c \in \mathbb{R}$ and some $z \in \mathbb{C} \setminus \mathbb{R}$, and that z and $\overline{z}$ are simple roots.

(b) Show that the hypothesis cannot be weakened to "f has exactly one real root" by considering $(x-1)^3$.

Solution.

(a) By Corollary 6.5.8, $f = a(x-c_1)\cdots(x-c_r)\,q_1 \cdots q_s$ with $r + 2s = 3$ and r equal to the sum of the real multiplicities, which is 1. So $r = 1$, $s = 1$, and $f = a(x-c)q$ with $c \in \mathbb{R}$ and q a monic real quadratic without real roots. By Theorem 6.5.1, q has a root $z \in \mathbb{C}$, which is not real. By Theorem 6.5.6, $\overline{z} \neq z$ is also a root of q. The two multiplicities in q are each at least 1 and add up to at most 2 (Theorem 6.4.3), so they equal 1 and $q = (x-z)(x-\overline{z})$, the leading coefficient being 1. Finally, by Theorem 6.4.2, $\mathrm{mult}_z(f) = \mathrm{mult}_z(a(x-c)) + \mathrm{mult}_z(q) = 0 + 1 = 1$, since $a(z-c) \neq 0$; similarly for $\overline{z}$.

(b) $(x-1)^3$ has exactly one real root, 1, but no nonreal roots at all: it already splits over $\mathbb{R}$. Here the real multiplicities add up to 3, not 1.

Exercise. B4: Running the proof on an example

Let $p = x^4 - x^2 + 2 \in \mathbb{C}[x]$.

(a) Compute the radius R of Lemma 6.5.2 for p.

(b) Follow the proof of Lemma 6.5.4 at $z_0 = 0$: find k, B, u, and t, and verify that $|p(tu)| < |p(0)|$.

(c) Compute $p(i/2)$, and explain why the direction i does not work here, although it did

in Example (One Step Downhill).

Solution.

(a) Here $a_4 = 1$ and $A = |2| + |0| + |-1| + |0| = 3$, so $R = \max\{1, 6\} = 6$. Hence $|p(z)| \geq \frac{1}{2}|z|^4$ for $|z| \geq 6$.

(b) At $z_0 = 0$ the coefficients are $b_0 = 2$, $b_1 = 0$, $b_2 = -1$, $b_3 = 0$, $b_4 = 1$. So $k = 2$ and $B = |b_3| + |b_4| = 1$. Then $w = -b_0/b_2 = 2 = 2e^{i\cdot 0}$, so $\alpha = 0$ and $u = 1$. Next $t = \min\{1,\ 2/1,\ 1/2\} = \frac{1}{2}$. So $z = \frac{1}{2}$, and $p(\frac{1}{2}) = \frac{1}{16} - \frac{1}{4} + 2 = \frac{29}{16} < 2 = |p(0)|$. The proof guarantees at most $2 - \frac{1}{2} \cdot \frac{1}{4} \cdot 1 = \frac{15}{8} = \frac{30}{16}$, consistent with $\frac{29}{16}$.

(c) $p(i/2) = \frac{1}{16} + \frac{1}{4} + 2 = \frac{37}{16} > 2$. The right direction depends on the sign of b_k relative to b_0: we need b_2u^2 to point from b_0 toward 0, that is, b_2u^2 a negative multiple of $b_0 = 2$. With $b_2 = -1$ this means $u^2 > 0$, so $u = \pm 1$; for $u = i$, $b_2u^2 = 1$ points away from 0. In Example (One Step Downhill), $b_2 = +1$ and $u = i$ was the right choice.

6.5.8.3 C. Going deeper

Exercise. C1: Fields where the theorem fails

Call a field F **algebraically closed** if every nonconstant polynomial in $F[x]$ has a root in F.

(a) Let p be a prime. Show that $x^p - x + 1$ has no root in $\mathbb{F}_p$. More generally, show that no finite field is algebraically closed.

(b) Give explicit witnesses that $\mathbb{R}$ and $\mathbb{Q}$ are not algebraically closed.

(c) For your witness over $\mathbb{R}$, explain which step of the proof of the Fundamental Theorem of Algebra (Theorem 6.5.1) breaks.

(d) For $f = x^2 - 2$ over $\mathbb{Q}$, prove that $r \mapsto |f(r)|$ has **no** minimum on $\mathbb{Q}$. Which step of the proof breaks now?

Hint for (d): if r_0 were a minimum point with $r_0^2 > \frac{2}{5}$, compare r_0 with $r_1 = \dfrac{r_0^2 + 2}{2r_0}$.

Solution.

(a) By Exercise (C2: A nonzero polynomial vanishing everywhere) (c), $a^p - a = 0$ for every $a \in \mathbb{F}_p$, so $a^p - a + 1 = 1 \neq 0$. For a finite field $F = \{c_1, \ldots, c_q\}$, let $g = (x - c_1) \cdots (x - c_q) + 1$. It has degree $q \geq 2$ (Theorem 0.8.3), and $g(c_i) = 0 + 1 = 1 \neq 0$ for every i by Theorem 0.8.5. So g is a nonconstant polynomial without a root in F.

(b) Over $\mathbb{R}$, $x^2 + 1$ has no root, since $c^2 + 1 \geq 1$. Over $\mathbb{Q}$, $x^2 - 2$ has no root, since $\sqrt{2}$ is irrational (Theorem 0.2.3) and the only real roots are $\pm\sqrt{2}$.

(c) For $x^2 + 1$ over $\mathbb{R}$, growth holds, and $c \mapsto |c^2 + 1|$ attains its minimum 1 at $c = 0$. The step that breaks is d'Alembert's lemma (Lemma 6.5.4): at $z_0 = 0$ we have $b_0 = 1$, $k = 2$, $b_2 = 1$, and we need u with $u^2 = -1$, which does not exist in $\mathbb{R}$.

(d) Suppose $r_0 \in \mathbb{Q}$ satisfies $|r_0^2 - 2| \leq |r^2 - 2|$ for every $r \in \mathbb{Q}$. Put $m = |r_0^2 - 2|$; then $m > 0$ by (b).
Case 1: $r_0^2 \leq \frac{2}{5}$. Then $m = 2 - r_0^2 \geq \frac{8}{5} > 1 = |1^2 - 2|$, contradicting minimality at $r = 1$.

Case 2: $r_0^2 > \frac{2}{5}$. Then $r_0 \neq 0$, and $r_1 = (r_0^2+2)/(2r_0) \in \mathbb{Q}$. Expanding,

$$r_1^2 - 2 = \frac{(r_0^2+2)^2 - 8r_0^2}{4r_0^2} = \frac{(r_0^2-2)^2}{4r_0^2}, \qquad \text{so} \qquad |r_1^2-2| = m \cdot \frac{m}{4r_0^2}.$$

We claim $m < 4r_0^2$. If $r_0^2 \geq 2$, then $m = r_0^2 - 2 < 4r_0^2$. If $\frac{2}{5} < r_0^2 < 2$, then $m = 2 - r_0^2 < \frac{8}{5} < 4r_0^2$. Hence $|r_1^2 - 2| < m$, contradicting minimality.
So there is no minimum. The step of the proof that breaks is Lemma 6.5.3, which rests on the extreme value theorem (A2). That fact is about the completeness of $\mathbb{R}$, and $\mathbb{Q}$ is not complete: the values $|r^2-2|$ get arbitrarily close to 0 without reaching it.

Exercise. C2: Sums of two squares

Let $f \in \mathbb{R}[x]$ be monic with **no** real root.

(a) Prove that $\deg f$ is even and that $f(c) > 0$ for every $c \in \mathbb{R}$.

(b) Prove that $f = g^2 + h^2$ for some $g, h \in \mathbb{R}[x]$.

Hint for (b): each irreducible factor of f is $(x-z)(x-\overline{z})$. Collect one root from each pair into a complex polynomial G, and split G into real and imaginary parts.

Solution.

(a) By Corollary 6.5.8, $f = (x-c_1)\cdots(x-c_r)\,q_1\cdots q_s$, with leading coefficient 1 since f is monic. A factor $x - c_1$ would give the real root c_1, so $r = 0$ and $\deg f = 2s$ is even. (If $s = 0$, then $f = 1$ and the claims hold.) Each $q_j = x^2 + \beta x + \gamma$ has no real root, so by the remark after Corollary 6.5.7, $\beta^2 < 4\gamma$. For real c,

$$q_j(c) = \Big(c + \frac{\beta}{2}\Big)^2 + \frac{4\gamma - \beta^2}{4} > 0.$$

By Theorem 0.8.5, $f(c) = q_1(c)\cdots q_s(c)$ is a product of positive reals, so $f(c) > 0$.

(b) For each j, q_j is nonconstant, so by Theorem 6.5.1 it has a root $z_j \in \mathbb{C}$, not real. As in Exercise (B3: Real cubics) (a), $q_j = (x - z_j)(x - \overline{z_j})$. Let $G = (x-z_1)\cdots(x-z_s) \in \mathbb{C}[x]$. By the product rule for conjugate polynomials proved in Theorem 6.5.6, $\overline{G} = (x-\overline{z_1})\cdots(x-\overline{z_s})$, so $f = G\,\overline{G}$ (commutativity, Theorem 0.8.1).
Write each coefficient of G as $\alpha_k + i\beta_k$ with $\alpha_k, \beta_k \in \mathbb{R}$, and put $g = \sum_k \alpha_k x^k$ and $h = \sum_k \beta_k x^k$ in $\mathbb{R}[x]$. Then $G = g + ih$ and $\overline{G} = g - ih$, so

$$f = (g+ih)(g-ih) = g^2 - i^2h^2 + (ihg - igh) = g^2 + h^2.$$

For example, for x^4+1 take $z_1 = e^{i\pi/4}$ and $z_2 = e^{3i\pi/4}$ from Example (Factoring x^4+1 over $\mathbb{C}$ and over $\mathbb{R}$). Then $z_1 + z_2 = i\sqrt{2}$ and $z_1 z_2 = e^{i\pi} = -1$, so $G = x^2 - i\sqrt{2}\,x - 1$, $g = x^2 - 1$, $h = -\sqrt{2}\,x$, and $x^4 + 1 = (x^2-1)^2 + 2x^2$.

6.6 Interpolation and the Chinese Remainder Theorem

Chapter 3 proved that through $n+1$ points with distinct x-values there passes exactly one polynomial of degree at most n (Theorem 3.7.3). The proof went through an invertible matrix, and it never wrote the polynomial down. This section gives the formula. Then it reads interpolation in the arithmetic of this chapter: "$p(c) = y$" says that p leaves remainder y on division by $x - c$. Prescribing remainders modulo several pairwise coprime polynomials at once is the Chinese Remainder Theorem. We prove it twice, once with Bézout's identity, which gives a formula, and once by counting dimensions in the quotient spaces $F[x]/\langle m \rangle$.

6.6.1 The Lagrange basis

Suppose we want the polynomial of degree at most n taking prescribed values $y_0, \dots, y_n$ at distinct points $c_0, \dots, c_n$. The problem is linear in the data: if p fits the values (y_i) and q fits (y_i'), then $p + q$ fits $(y_i + y_i')$, by Theorem 0.8.5. So it is enough to solve the simplest data sets, a single 1 and zeros everywhere else, and add up.

The polynomial that vanishes at every c_j with $j \neq i$ is easy to write: $\prod_{j \neq i}(x - c_j)$. At c_i it takes the value $\prod_{j \neq i}(c_i - c_j)$, which is a product of nonzero scalars, so we can divide by it. In Example (Dual Bases in $\mathbb{R}[x]_{\leq 2}$) we met these polynomials for the three points $-1, 0, 1$. Here is the general case.

The i-th Lagrange polynomial is the polynomial of degree n that is 1 at the i-th node and 0 at every other node.

Definition 6.6.1. Lagrange Basis

Let $n \in \mathbb{N}$, and let $c_0, c_1, \dots, c_n \in F$ be **distinct**. For $i = 0, \dots, n$, the i-th **Lagrange polynomial** for these nodes is

$$\ell_i(x) := \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{x - c_j}{c_i - c_j} \in F[x].$$

The list $(\ell_0, \dots, \ell_n)$ is the **Lagrange basis** for the nodes $c_0, \dots, c_n$.

In words: the numerator makes ℓ_i vanish at every node except c_i, and the denominator rescales so that the value at c_i is exactly 1. The points c_i are called **nodes**.

Well-definedness. Each denominator $c_i - c_j$ with $j \neq i$ is nonzero because the nodes are **distinct**, so the quotients exist in F. The product has n linear factors, each with leading coefficient $(c_i - c_j)^{-1} \neq 0$, so $\deg \ell_i = n$ by Theorem 0.8.3. The name "basis" is justified by the theorem below.

Examples.

- **One node.** For $n = 0$ the product is empty, so $\ell_0 = 1$. The degenerate case is consistent: a polynomial of degree at most 0 with value y_0 at c_0 is the constant $y_0 = y_0 \ell_0$.
- **Two nodes.** For $c_0 = 0$, $c_1 = 1$ we get $\ell_0 = \frac{x-1}{0-1} = 1 - x$ and $\ell_1 = \frac{x-0}{1-0} = x$. The line through $(0, y_0)$ and $(1, y_1)$ is $y_0(1 - x) + y_1 x$.
- **Three nodes.** For $-1, 0, 1$, $\ell_{-1} = \frac{x(x-1)}{(-1)(-2)} = \frac{1}{2}x(x-1)$, $\ell_0 = \frac{(x+1)(x-1)}{(1)(-1)} = 1 - x^2$, and

$\ell_1 = \frac{1}{2}x(x+1)$, the polynomials of Example (Dual Bases in $\mathbb{R}[x]_{\leq 2}$). (Here we index the nodes by their values.)

Nonexample by minimal change. Take the nodes $0, 1, 1$. The numerator of ℓ_1 is still defined, $(x-0)(x-1)$, but the denominator contains $c_1 - c_2 = 0$, which has no inverse. No rescaling helps: a polynomial that vanishes at $c_2 = 1$ cannot take the value 1 at $c_1 = 1$. The clause that fails is "distinct".

Everything about the Lagrange polynomials follows from one table of values, the Kronecker delta $\ell_i(c_j) = \delta_{ij}$.

Theorem 6.6.1. Lagrange Interpolation

Let $n \in \mathbb{N}$, let $c_0, \ldots, c_n \in F$ be distinct, and let $(\ell_0, \ldots, \ell_n)$ be the Lagrange basis for these nodes.

(a) $\ell_i(c_j) = \delta_{ij}$ for all $i, j \in \{0, \ldots, n\}$.

(b) $(\ell_0, \ldots, \ell_n)$ is a basis of $F[x]_{\leq n}$, and every $p \in F[x]_{\leq n}$ satisfies

$$p = \sum_{i=0}^{n} p(c_i)\, \ell_i.$$

(c) For all $y_0, \ldots, y_n \in F$, the unique $p \in F[x]_{\leq n}$ with $p(c_i) = y_i$ for every i is

$$p = \sum_{i=0}^{n} y_i\, \ell_i.$$

Idea. ① The delta table is a direct evaluation. ② Evaluating a vanishing combination $\sum a_i \ell_i$ at c_j kills every term but a_j, so the list is independent, and it has the right length, so we count instead of checking spanning. ③ The expansion of p and the interpolation formula are the same computation: both sides of a candidate identity are polynomials of degree at most n with the same values at $n+1$ nodes.

Proof.

(a) By Theorem 0.8.5, $\ell_i(c_j) = \prod_{k \neq i} \frac{c_j - c_k}{c_i - c_k}$. If $j = i$, every factor is 1, so $\ell_i(c_i) = 1$. If $j \neq i$, the factor with $k = j$ is 0, so $\ell_i(c_j) = 0$.

(b) Each ℓ_i has degree n, so it lies in $F[x]_{\leq n}$. Let $a_0\ell_0 + \cdots + a_n\ell_n = 0$. Evaluating at c_j and using (a) gives $0 = \sum_i a_i \delta_{ij} = a_j$ for each j. Hence the list is linearly independent. It has length $n+1 = \dim F[x]_{\leq n}$ (Example (Standard Bases)), so it is a basis by Theorem 1.6.4 (a).
Now let $p \in F[x]_{\leq n}$, and put $q = p - \sum_i p(c_i)\ell_i \in F[x]_{\leq n}$. By (a), $q(c_j) = p(c_j) - p(c_j) = 0$ for every j. So q has the $n+1$ distinct roots $c_0, \ldots, c_n$ and degree at most n. By Corollary 6.1.4, $q = 0$, which is the expansion of p.

(c) By (a), $(\sum_i y_i \ell_i)(c_j) = \sum_i y_i \delta_{ij} = y_j$, so $\sum_i y_i \ell_i$ has the required values, and it lies in $F[x]_{\leq n}$. By Theorem 3.7.3, it is the only such polynomial. This proves the theorem.

■

So Chapter 3's existence statement now comes with a formula, and the uniqueness half of (c) is the expansion in (b) read backwards: a polynomial of degree at most n is determined by its values at the nodes, because those values are its coordinates in the Lagrange basis.

Example. The Chapter 3 Cubic by Lagrange's Formula

Use Theorem 6.6.1 to find the $p \in \mathbb{Q}[x]_{\le 3}$ with $p(-1) = 2$, $p(0) = 1$, $p(1) = 2$, and $p(2) = 11$, and compare with Example (A cubic through four points).

Solution. The nodes $-1, 0, 1, 2$ are distinct. Write $\ell_{-1}, \ell_0, \ell_1, \ell_2$ for the Lagrange polynomials, indexed by the node. The denominators are $\prod_{j \neq i}(c_i - c_j)$:

$$\ell_{-1} = \frac{x(x-1)(x-2)}{(-1)(-2)(-3)} = -\tfrac{1}{6}x(x-1)(x-2),$$
$$\ell_0 = \frac{(x+1)(x-1)(x-2)}{(1)(-1)(-2)} = \tfrac{1}{2}(x+1)(x-1)(x-2),$$
$$\ell_1 = \frac{(x+1)x(x-2)}{(2)(1)(-1)} = -\tfrac{1}{2}(x+1)x(x-2),$$
$$\ell_2 = \frac{(x+1)x(x-1)}{(3)(2)(1)} = \tfrac{1}{6}(x+1)x(x-1).$$

Expanding,

$$\ell_{-1} = -\tfrac{1}{6}x^3 + \tfrac{1}{2}x^2 - \tfrac{1}{3}x,$$
$$\ell_0 = \tfrac{1}{2}x^3 - x^2 - \tfrac{1}{2}x + 1,$$
$$\ell_1 = -\tfrac{1}{2}x^3 + \tfrac{1}{2}x^2 + x,$$
$$\ell_2 = \tfrac{1}{6}x^3 - \tfrac{1}{6}x.$$

By Theorem 6.6.1 (c), $p = 2\ell_{-1} + \ell_0 + 2\ell_1 + 11\ell_2$. Collecting coefficients:

$$x^3\colon \ -\tfrac{1}{3} + \tfrac{1}{2} - 1 + \tfrac{11}{6} = 1,$$
$$x^2\colon \ 1 - 1 + 1 + 0 = 1,$$
$$x\colon \ -\tfrac{2}{3} - \tfrac{1}{2} + 2 - \tfrac{11}{6} = -1,$$
$$1\colon \ 1.$$

Hence $p = x^3 + x^2 - x + 1$, the polynomial found by elimination in Example (A cubic through four points), as uniqueness demands. The elimination solved a 4×4 system for these particular values. The Lagrange polynomials were computed once, and would serve for any other values at the same nodes.

Warning. **The interpolant has degree at most n, not exactly n.** Through the three points $(0,0), (1,1), (2,2)$ the formula gives $0 \cdot \ell_0 + 1 \cdot \ell_1 + 2\ell_2$, where $\ell_1 = -x(x-2)$ and $\ell_2 = \tfrac{1}{2}x(x-1)$; this is $-x^2 + 2x + x^2 - x = x$. The degree-2 terms cancel, because the points lie on a line. Each ℓ_i has degree exactly n, but a combination of them need not.

Quick check. Let $n \geq 1$ and let $(\ell_0, \dots, \ell_n)$ be the Lagrange basis for distinct nodes $c_0, \dots, c_n$. What is $\sum_{i=0}^{n} c_i \ell_i$?

Solution. It is the polynomial x. Indeed $x \in F[x]_{\leq n}$ since $n \geq 1$, and its values at the nodes are $c_0, \dots, c_n$, so Theorem 6.6.1 (b) gives $x = \sum_i c_i \ell_i$.

6.6.2 Evaluations are the dual basis

The delta table $\ell_i(c_j) = \delta_{ij}$ can be read the other way round. For $c \in F$, evaluation at c is the linear functional $\varepsilon_c \colon p \mapsto p(c)$ (Example (Four Families of Functionals)), which we restrict to $F[x]_{\leq n}$. Then the table says $\varepsilon_{c_j}(\ell_i) = \delta_{ji}$: the evaluations at the nodes behave on the Lagrange basis exactly as a dual basis must (Theorem 5.1.3). Exercise (C3: Evaluations as a dual basis) proved from Theorem 3.7.3 that the evaluations form the dual basis of **some** basis. We can now name that basis.

Corollary 6.6.2. Evaluations Are Dual to the Lagrange Basis

Let $c_0, \dots, c_n \in F$ be distinct, with Lagrange basis $(\ell_0, \dots, \ell_n)$. Then $(\varepsilon_{c_0}, \dots, \varepsilon_{c_n})$ is the dual basis of $(\ell_0, \dots, \ell_n)$ in $F[x]^*_{\leq n}$. In particular, every linear functional φ on $F[x]_{\leq n}$ is a combination of evaluations at the nodes:

$$\varphi = \sum_{i=0}^{n} \varphi(\ell_i)\, \varepsilon_{c_i}.$$

Proof. The evaluations are linear (Example (Four Families of Functionals)), and $\varepsilon_{c_i}(\ell_j) = \ell_j(c_i) = \delta_{ij}$ by Theorem 6.6.1 (a). Since $(\ell_0, \dots, \ell_n)$ is a basis (Theorem 6.6.1 (b)), the uniqueness in Theorem 5.1.3 (a) shows that the evaluations are its dual basis. The formula for φ is the second formula of Theorem 5.1.3 (c). ∎

The first formula of Theorem 5.1.3 (c), $p = \sum_i \varepsilon_{c_i}(p)\, \ell_i$, is the expansion in Theorem 6.6.1 (b). So Lagrange's formula is the statement "a vector is recovered from its dual-basis coordinates", for the space $F[x]_{\leq n}$ and the measurements "value at c_i".

The formula for φ has a practical reading: **any** linear measurement of a polynomial of degree at most n can be computed from $n+1$ of its values, with weights $\varphi(\ell_i)$ that do not depend on the polynomial.

Example. Integrals and Slopes from Values

(a) Find $w_0, w_1, w_2 \in \mathbb{R}$ such that $\int_0^1 p(t)\,\mathrm{d}t = w_0\, p(0) + w_1\, p(\tfrac{1}{2}) + w_2\, p(1)$ for every $p \in \mathbb{R}[x]_{\leq 2}$.

(b) Express $p \mapsto p'(0)$ on $\mathbb{R}[x]_{\leq 2}$ as a combination of $\varepsilon_{-1}, \varepsilon_0, \varepsilon_1$.

Solution.

(a) Integration over $[0,1]$ is a linear functional on $\mathbb{R}[x]_{\le 2}$ (Example (Four Families of Functionals)). For the nodes $0, \frac{1}{2}, 1$,

$$\ell_0 = \frac{(x-\frac{1}{2})(x-1)}{(-\frac{1}{2})(-1)} = 2x^2 - 3x + 1,$$
$$\ell_{1/2} = \frac{x(x-1)}{(\frac{1}{2})(-\frac{1}{2})} = -4x^2 + 4x,$$
$$\ell_1 = \frac{x(x-\frac{1}{2})}{(1)(\frac{1}{2})} = 2x^2 - x.$$

Their integrals are $\frac{2}{3} - \frac{3}{2} + 1 = \frac{1}{6}$, $-\frac{4}{3} + 2 = \frac{2}{3}$, and $\frac{2}{3} - \frac{1}{2} = \frac{1}{6}$. By Corollary 6.6.2,

$$\int_0^1 p(t)\,\mathrm{d}t = \tfrac{1}{6}\,p(0) + \tfrac{2}{3}\,p(\tfrac{1}{2}) + \tfrac{1}{6}\,p(1) \qquad \text{for all } p \in \mathbb{R}[x]_{\le 2}.$$

Check with $p = x^2$: the right side is $\frac{2}{3}\cdot\frac{1}{4} + \frac{1}{6} = \frac{1}{3} = \int_0^1 t^2\,\mathrm{d}t$. This is Simpson's rule, which here is exact rather than approximate because p has degree at most 2.

(b) With $\ell_{-1} = \frac{1}{2}x(x-1)$, $\ell_0 = 1 - x^2$, and $\ell_1 = \frac{1}{2}x(x+1)$, the derivatives at 0 are $-\frac{1}{2}$, 0, and $\frac{1}{2}$. Hence $p'(0) = \frac{1}{2}(p(1) - p(-1))$ for every $p \in \mathbb{R}[x]_{\le 2}$, the symmetric difference quotient, exact on quadratics.

6.6.3 Congruences and the Chinese Remainder Theorem

Now we change the point of view. By Theorem 6.1.3, the remainder of p on division by $x - c$ is $p(c)$. So the interpolation conditions $p(c_i) = y_i$ say that p leaves the remainder y_i on division by $x - c_i$, for each i. Nothing in this reformulation needs the divisors to be linear. What happens if we prescribe the remainders modulo arbitrary polynomials $m_1, \dots, m_k$? The integers faced the same question long ago, which explains the name of the answer. We first need the language of congruences, copied from $\mathbb{Z}$.

Definition 6.6.2. Congruence Modulo a Polynomial

Let $m \in F[x]$. Two polynomials $f, g \in F[x]$ are **congruent modulo** m, written $f \equiv g \pmod{m}$, if $m \mid f - g$.

For example, $x^2 \equiv -1 \pmod{x^2+1}$, and $f \equiv f(c) \pmod{x-c}$ for every f, by Theorem 6.1.3. Congruence modulo m is an equivalence relation, and it respects sums and products: if $f \equiv f'$ and $g \equiv g' \pmod{m}$, then $f+g-(f'+g') = (f-f')+(g-g')$ and $fg - f'g' = f(g-g')+(f-f')g'$ are divisible by m. By Theorem 6.1.1, if $m \neq 0$, every f is congruent modulo m to **exactly one** polynomial of degree less than $\deg m$, its remainder: $f = qm + r$ gives $f \equiv r$, and if $f \equiv r \equiv r'$ with both degrees below $\deg m$, then $m \mid r - r'$ with $\deg(r - r') < \deg m$, which forces $r - r' = 0$.

Interpolation needed distinct nodes. The general condition is that the moduli share no factor, and the proof needs a way to pass from "coprime in pairs" to "coprime to the product".

Lemma 6.6.3. Coprime to Each, Coprime to the Product

(a) Let $p, q_1, \dots, q_s \in F[x]$. If p is coprime to each q_j, then there are $a, b \in F[x]$ with $ap + b\,q_1 \cdots q_s = 1$, and p is coprime to $q_1 \cdots q_s$.

(b) Let $m_1, \dots, m_k \in F[x]$ be nonzero and **pairwise coprime**, that is, $\gcd(m_i, m_j) = 1$ whenever $i \neq j$. If $m_i \mid h$ for every i, then $m_1 \cdots m_k \mid h$.

Proof.

(a) By Corollary 6.2.4, for each j there are a_j, b_j with $a_j p + b_j q_j = 1$. Multiplying these s identities,

$$1 = \prod_{j=1}^{s} (a_j p + b_j q_j) = ap + (b_1 \cdots b_s)\, q_1 \cdots q_s$$

for some $a \in F[x]$, because every term of the expanded product except $(b_1 q_1) \cdots (b_s q_s)$ contains a factor $a_j p$. Let $d = \gcd(p, q_1 \cdots q_s)$. By Theorem 6.2.3, d divides p and $q_1 \cdots q_s$, hence divides $ap + b\,q_1 \cdots q_s = 1$, where $b = b_1 \cdots b_s$. So d is a nonzero constant by Theorem 0.8.3, and being monic, $d = 1$. Hence p and $q_1 \cdots q_s$ are coprime (Definition 6.2.3).

(b) We use induction on k. For $k = 1$ there is nothing to prove. Let $k \geq 2$. The moduli $m_1, \dots, m_{k-1}$ are pairwise coprime and each divides h, so by the induction hypothesis $M := m_1 \cdots m_{k-1}$ divides h, say $h = Mg$. Since $m_k \mid h = Mg$, and m_k is coprime to M by (a), Theorem 6.2.8 gives $m_k \mid g$, say $g = m_k g'$. Then $h = m_1 \cdots m_k\, g'$, as claimed.

■

Theorem 6.6.4. Chinese Remainder Theorem for Polynomials

Let $m_1, \dots, m_k \in F[x]$ be nonzero and pairwise coprime, and put $m = m_1 \cdots m_k$. For all $r_1, \dots, r_k \in F[x]$:

(a) there is $f \in F[x]$ with $f \equiv r_i \pmod{m_i}$ for every i;

(b) if f and g both satisfy (a), then $f \equiv g \pmod{m}$;

(c) there is **exactly one** such f with $\deg f < \deg m$.

Idea. Copy Lagrange. There, ℓ_i was 1 at c_i and 0 at the other nodes; in the new language, $\ell_i \equiv 1 \pmod{x - c_i}$ and $\ell_i \equiv 0 \pmod{x - c_j}$. So we look for e_i with $e_i \equiv 1 \pmod{m_i}$ and $e_i \equiv 0 \pmod{m_j}$ for $j \neq i$, and then $f = \sum_i r_i e_i$ works. To be 0 modulo every other m_j, e_i should be a multiple of $M_i = \prod_{j \neq i} m_j$; to be 1 modulo m_i, it should be $1 - a\,m_i$. Bézout's identity $a\,m_i + b\,M_i = 1$ provides exactly such an element: $e_i = b\,M_i$.

Proof.

(a) Fix i, and let $M_i = \prod_{j \neq i} m_j$. Since m_i is coprime to each m_j with $j \neq i$, Lemma 6.6.3 (a) gives $a_i, b_i \in F[x]$ with $a_i m_i + b_i M_i = 1$. Put $e_i = b_i M_i$. Then $e_i - 1 = -a_i m_i$, so $e_i \equiv 1 \pmod{m_i}$, and $m_j \mid M_i \mid e_i$ for $j \neq i$, so $e_i \equiv 0 \pmod{m_j}$. Let $f =$

$r_1e_1 + \cdots + r_ke_k$. Since congruence modulo m_j respects sums and products,

$$f \equiv r_1 \cdot 0 + \cdots + r_j \cdot 1 + \cdots + r_k \cdot 0 = r_j \pmod{m_j}$$

for each j.

(b) If f and g satisfy (a), then $f - g \equiv r_i - r_i = 0 \pmod{m_i}$, that is, $m_i \mid f - g$, for every i. By Lemma 6.6.3 (b), $m \mid f - g$.

(c) Since each $m_i \neq 0$, $m \neq 0$ (Theorem 0.8.3). Let f be as in (a), and let f_0 be its remainder on division by m (Theorem 6.1.1). Then $m_i \mid m \mid f - f_0$, so $f_0 \equiv f \equiv r_i \pmod{m_i}$ for each i, and $\deg f_0 < \deg m$. If g is another solution with $\deg g < \deg m$, then $m \mid f_0 - g$ by (b), and $\deg(f_0 - g) < \deg m$, so $f_0 = g$. This proves the theorem. ■

Lagrange is the case $m_i = x - c_i$. For distinct nodes, $(x - c_j) - (x - c_i) = c_i - c_j$ is a nonzero constant, so $\frac{1}{c_i - c_j}(x - c_j) - \frac{1}{c_i - c_j}(x - c_i) = 1$, and any common divisor of $x - c_i$ and $x - c_j$ divides 1: the moduli are pairwise coprime. The congruence $f \equiv y_i \pmod{x - c_i}$ means $f(c_i) = y_i$ by Theorem 6.1.3, and $\deg m = n + 1$. So part (c) is Theorem 3.7.3. The element e_i has degree less than $n + 1$ after reduction modulo m, is 1 at c_i and 0 at the other nodes, so after reduction it is ℓ_i.

Example. A Linear and a Quadratic Modulus

Find the polynomial $f \in \mathbb{Q}[x]$ of degree less than 3 with

$$f \equiv 1 \pmod{x - 2}, \qquad f \equiv 2x - 1 \pmod{x^2 + 1}.$$

Solution. Bézout. Divide $x^2 + 1$ by $x - 2$: $x^2 + 1 = (x + 2)(x - 2) + 5$. Hence

$$\tfrac{1}{5}(x^2 + 1) - \tfrac{1}{5}(x + 2)(x - 2) = 1,$$

and the moduli are coprime, since any common divisor divides 1.

The elements e_i. Put $e_1 = \frac{1}{5}(x^2 + 1)$ and $e_2 = -\frac{1}{5}(x + 2)(x - 2) = 1 - e_1$. Then e_1 is a multiple of $x^2 + 1$ and $e_1 = 1 + \frac{1}{5}(x + 2)(x - 2) \equiv 1 \pmod{x - 2}$; symmetrically, $e_2 \equiv 0 \pmod{x - 2}$ and $e_2 \equiv 1 \pmod{x^2 + 1}$.

Combine and reduce. By the proof of Theorem 6.6.4, $f \equiv 1 \cdot e_1 + (2x - 1)e_2 \pmod{(x - 2)(x^2 + 1)}$. Expanding,

$$e_1 + (2x - 1)e_2 = \tfrac{1}{5}(x^2 + 1) - \tfrac{1}{5}(2x - 1)(x^2 - 4) = \tfrac{1}{5}(-2x^3 + 2x^2 + 8x - 3).$$

Dividing by $m = (x - 2)(x^2 + 1) = x^3 - 2x^2 + x - 2$ gives quotient $-\frac{2}{5}$ and remainder

$$f = \tfrac{1}{5}(-2x^2 + 10x - 7).$$

Check. $f(2) = \frac{1}{5}(-8 + 20 - 7) = 1$, so $f \equiv 1 \pmod{x - 2}$. Modulo $x^2 + 1$, $x^2 \equiv -1$, so $f \equiv \frac{1}{5}(2 + 10x - 7) = 2x - 1$.

Warning. **The moduli must be coprime in pairs; coprime as a whole family is not enough.** Without coprimality a system can be unsolvable: with $m_1 = x$ and $m_2 = x^2$, the system $f \equiv 0 \pmod{x}$, $f \equiv 1 \pmod{x^2}$ has no solution: the first condition says $f(0) = 0$, the second says $f = 1 + x^2 q$, so $f(0) = 1$. For three moduli, x, $x-1$, and $x(x-1)$ have no common factor, yet the first and third share x, and $f \equiv 0 \pmod{x}$, $f \equiv 1 \pmod{x(x-1)}$ is again impossible.

Example. Values and Slopes by CRT

(a) Let $c, a, b \in F$ and $f \in F[x]$. Show that $f \equiv a + b(x-c) \pmod{(x-c)^2}$ if and only if $f(c) = a$ and $f'(c) = b$.

(b) Find the $p \in \mathbb{Q}[x]_{\leq 3}$ with $p(0) = 2$, $p'(0) = -1$, $p(1) = 1$, and $p'(1) = 3$.

Solution.

(a) Divide by $(x-c)^2$: $f = q(x-c)^2 + r$ with $\deg r \leq 1$ (Theorem 6.1.1). Writing $r = r_0 + r_1(x-c)$, which is possible because $(1, x-c)$ is a basis of $F[x]_{\leq 1}$ by Theorem 1.4.4 and Theorem 1.6.4, the uniqueness of the remainder says $f \equiv a + b(x-c)$ exactly when $r_0 = a$ and $r_1 = b$. Now $f(c) = r_0$. By Theorem 6.4.4, $f' = q'(x-c)^2 + 2q(x-c) + r_1$, so $f'(c) = r_1$. Hence the congruence holds if and only if $f(c) = a$ and $f'(c) = b$.

(b) By (a), the conditions are $p \equiv 2 - x \pmod{x^2}$ and $p \equiv 1 + 3(x-1) \pmod{(x-1)^2}$. The moduli are coprime, since

$$(2x+1)(x-1)^2 - (2x-3)x^2 = (2x^3 - 3x^2 + 1) - (2x^3 - 3x^2) = 1.$$

So $e_0 = (2x+1)(x-1)^2$ is $\equiv 0 \pmod{(x-1)^2}$ and $\equiv 1 \pmod{x^2}$, while $e_1 = -(2x-3)x^2 = 3x^2 - 2x^3$ is $\equiv 0 \pmod{x^2}$ and $\equiv 1 \pmod{(x-1)^2}$. The polynomial e_1 is the smooth step $3x^2 - 2x^3$ of Example (Prescribing Values and Slopes): value 0 and slope 0 at 0, value 1 and slope 0 at 1. By Theorem 6.6.4, p is the remainder of

$$(2-x)e_0 + (3x-2)e_1 = -8x^4 + 20x^3 - 12x^2 - x + 2$$

on division by $x^2(x-1)^2 = x^4 - 2x^3 + x^2$. Adding $8(x^4 - 2x^3 + x^2)$ gives

$$p = 4x^3 - 4x^2 - x + 2.$$

Check. $p(0) = 2$; $p' = 12x^2 - 8x - 1$, so $p'(0) = -1$; $p(1) = 4 - 4 - 1 + 2 = 1$; $p'(1) = 12 - 8 - 1 = 3$.

The same recipe handles any finite set of nodes c_i with multiplicities k_i: the moduli $(x-c_i)^{k_i}$ are pairwise coprime for distinct nodes, by Lemma 6.6.3 (a) applied twice: first $x - c_i$ is coprime to $(x-c_j)^{k_j}$, and then $(x-c_j)^{k_j}$ is coprime to $(x-c_i)^{k_i}$. So for any prescribed remainders modulo $(x-c_i)^{k_i}$, there is exactly one polynomial of degree less than $k_1 + \cdots + k_s$ with those remainders. In characteristic 0, the Taylor expansion of Theorem 6.4.7 translates "the remainder modulo $(x-c)^k$" into the values $f(c), f'(c), \ldots, f^{(k-1)}(c)$, and this is **Hermite interpolation**.

6.6.4 The quotient space $F[x]/\langle m\rangle$

Congruence modulo m is equality of cosets. The ideal $\langle m\rangle$ is a subspace of $F[x]$: it is nonempty and closed under addition and under multiplication by constant polynomials (Definition 6.2.1). So the quotient space $F[x]/\langle m\rangle$ of Definition 4.4.3 exists, and by Lemma 4.4.3 (b),

$$f+\langle m\rangle = g+\langle m\rangle \iff f-g\in\langle m\rangle \iff f\equiv g \pmod m.$$

An element of $F[x]/\langle m\rangle$ is a polynomial "known only modulo m". We cannot quote Theorem 4.4.5 for its dimension, because $F[x]$ is infinite-dimensional (Theorem 1.3.3). The remainder does the job instead.

Theorem 6.6.5. Dimension of $F[x]/\langle m\rangle$

Let $m\in F[x]$ be nonzero of degree d. Then $F[x]/\langle m\rangle$ is finite-dimensional with

$$\dim F[x]/\langle m\rangle = \deg m = d.$$

If $d\geq 1$, the cosets $(1+\langle m\rangle, x+\langle m\rangle, \dots, x^{d-1}+\langle m\rangle)$ form a basis.

Proof. If $d=0$, then m is a nonzero constant, and every $f=(m^{-1}f)m$ lies in $\langle m\rangle$. So $\langle m\rangle = F[x]$ and the quotient is the zero space, of dimension 0.

Let $d\geq 1$, and define $\rho\colon F[x]\to F[x]_{\leq d-1}$ by letting $\rho(f)$ be the remainder of f on division by m (Theorem 6.1.1). *Linear.* If $f=q_1m+\rho(f)$ and $g=q_2m+\rho(g)$, then for $c\in F$,

$$cf+g=(cq_1+q_2)m+(c\rho(f)+\rho(g)),$$

and $\deg(c\rho(f)+\rho(g))<d$ by Theorem 0.8.2. By the uniqueness in Theorem 6.1.1, $\rho(cf+g)=c\rho(f)+\rho(g)$. *Surjective.* If $\deg r<d$, then $r=0\cdot m+r$, so $\rho(r)=r$. *Kernel.* $\rho(f)=0$ exactly when $f=qm$ for some q, that is, $f\in\langle m\rangle$.

By the First Isomorphism Theorem (Theorem 4.4.7), $\bar\rho\colon F[x]/\langle m\rangle\to F[x]_{\leq d-1}$, $f+\langle m\rangle\mapsto\rho(f)$, is a well-defined injective linear map with image $\operatorname{im}\rho = F[x]_{\leq d-1}$, so it is an isomorphism. Since $\bar\rho(x^j+\langle m\rangle)=x^j$ for $j<d$, and $(1,x,\dots,x^{d-1})$ is a basis of $F[x]_{\leq d-1}$ (Example (Standard Bases)), the inverse isomorphism $\bar\rho^{-1}$ carries it to the basis $(x^j+\langle m\rangle)_{j<d}$ (Theorem 2.5.2). Hence $\dim F[x]/\langle m\rangle = d$. ∎

For example, $\mathbb{R}[x]/\langle x^2+1\rangle$ has basis $(1+\langle x^2+1\rangle, x+\langle x^2+1\rangle)$, so each element is $a+bx+\langle x^2+1\rangle$ for unique $a,b\in\mathbb{R}$.

Remark. Cosets can also be **multiplied**: $(f+\langle m\rangle)(g+\langle m\rangle) := fg+\langle m\rangle$ is well defined, because congruence modulo m respects products. In $\mathbb{R}[x]/\langle x^2+1\rangle$, $x\cdot x = x^2\equiv -1$, so the coset of x squares to minus the coset of 1, and $(a+bx)(a'+b'x)\equiv(aa'-bb')+(ab'+a'b)x$: this is multiplication in $\mathbb{C}$, with the coset of x playing the role of i.

The Chinese Remainder Theorem now becomes a statement about linear maps. Reducing a polynomial modulo each m_i records its k remainders; the theorem says that this record, taken

modulo m, is a perfect dictionary. Recall from Definition 4.4.1 that $V_1 \times \cdots \times V_k$ is the space of lists with entrywise operations.

Theorem 6.6.6. CRT as an Isomorphism

Let $m_1, \dots, m_k \in F[x]$ be nonzero and pairwise coprime, and put $m = m_1 \cdots m_k$. Then

$$\begin{aligned} \Phi\colon F[x]/\langle m\rangle &\to F[x]/\langle m_1\rangle \times \cdots \times F[x]/\langle m_k\rangle, \\ f + \langle m\rangle &\mapsto (f + \langle m_1\rangle, \dots, f + \langle m_k\rangle), \end{aligned}$$

is a well-defined isomorphism of vector spaces.

Idea. The map that does not care about well-definedness is $\Psi\colon f \mapsto (f + \langle m_i\rangle)_i$ on $F[x]$ itself. Its kernel is the set of polynomials divisible by every m_i, which is $\langle m\rangle$ by Lemma 6.6.3. So the First Isomorphism Theorem makes Φ well defined and injective in one stroke. For surjectivity we do **not** construct a preimage. We count: both sides have dimension $\deg m$, because degrees add under products.

Proof. Let $\Psi\colon F[x] \to F[x]/\langle m_1\rangle \times \cdots \times F[x]/\langle m_k\rangle$, $\Psi(f) = (f + \langle m_1\rangle, \dots, f + \langle m_k\rangle)$. Each entry is a quotient map, which is linear by Theorem 4.4.4 (c), and the operations in the product are entrywise, so Ψ is linear. By Lemma 4.4.3 (b), $\Psi(f) = 0$ if and only if $m_i \mid f$ for every i. If $m \mid f$, then $m_i \mid f$ for all i, since $m_i \mid m$; conversely, if every $m_i \mid f$, then $m \mid f$ by Lemma 6.6.3 (b). Hence $\ker \Psi = \langle m\rangle$.

By the First Isomorphism Theorem (Theorem 4.4.7), the rule $f + \langle m\rangle \mapsto \Psi(f)$ is a well-defined injective linear map; this is Φ. By Theorem 6.6.5, Theorem 0.8.3, and Theorem 4.4.1,

$$\begin{aligned} \dim F[x]/\langle m\rangle = \deg m &= \sum_{i=1}^{k} \deg m_i = \sum_{i=1}^{k} \dim F[x]/\langle m_i\rangle \\ &= \dim\Big(F[x]/\langle m_1\rangle \times \cdots \times F[x]/\langle m_k\rangle\Big). \end{aligned}$$

Since Φ is injective between spaces of the same finite dimension, it is bijective by Corollary 2.3.2 (e), hence an isomorphism (Definition 2.5.2).

■

Surjectivity of Φ is the existence part of Theorem 6.6.4, and injectivity is the uniqueness part. So the count has given a second proof of the Chinese Remainder Theorem, one that never uses Bézout's identity for existence. It is the same trade as in Example (Interpolation as a Count): counting proves that a solution exists, and the Bézout construction says what it is. By the remark above, Φ also respects products of cosets, since reduction modulo each m_i does. In Chapter 9 this splitting of $F[x]/\langle m\rangle$ into pieces, one for each coprime factor of m, reappears as a splitting of a space on which an operator acts. The next section begins that story.

Quick check. What is $\dim \mathbb{Q}[x]/\langle x^3 - x\rangle$? Into which three spaces does Theorem 6.6.6 split it, and what does Φ do to the coset of x^2?

Solution. By Theorem 6.6.5, the dimension is 3. Since $x^3 - x = x(x-1)(x+1)$ with distinct linear factors, which are pairwise coprime, Φ maps it isomorphically onto $\mathbb{Q}[x]/\langle x\rangle \times \mathbb{Q}[x]/\langle x-1\rangle \times \mathbb{Q}[x]/\langle x+1\rangle$, a product of three one-dimensional spaces. Modulo $x - c$, a polynomial is congruent to its value at c, so $\Phi(x^2 + \langle x^3 - x\rangle)$ is the list of cosets of the constants $0, 1, 1$: the values of x^2 at $0, 1, -1$.

6.6.5 Exercises

6.6.5.1 A. Check your understanding

Exercise. A1

(a) Define the Lagrange basis for distinct nodes $c_0, \dots, c_n$, and state the value of ℓ_i at c_j.

(b) True or false: for four distinct nodes, the polynomial of degree at most 3 through the four points always has degree exactly 3. Justify your answer.

(c) State the Chinese Remainder Theorem for polynomials, with all hypotheses.

(d) True or false: if $\gcd(m_1, m_2, m_3) = 1$, then every system $f \equiv r_i \pmod{m_i}$, $i = 1, 2, 3$, has a solution. Justify your answer.

(e) What is $\dim_{\mathbb{R}} \mathbb{R}[x]/\langle x^4 + 1\rangle$? Give a basis.

(f) Which functional on $F[x]_{\le n}$ is the i-th vector of the dual basis (Theorem 5.1.3) of the Lagrange basis?

Solution.

(a) For distinct $c_0, \dots, c_n \in F$, $\ell_i = \prod_{j \neq i} \frac{x - c_j}{c_i - c_j}$ (Definition 6.6.1), and $\ell_i(c_j) = \delta_{ij}$ (Theorem 6.6.1 (a)).

(b) False. The points $(0,0), (1,0), (2,0), (3,0)$ are fitted by the zero polynomial, of degree $-\infty$; by uniqueness (Theorem 3.7.3), that is the interpolant.

(c) If $m_1, \dots, m_k \in F[x]$ are nonzero and pairwise coprime and $m = m_1 \cdots m_k$, then for all $r_1, \dots, r_k \in F[x]$ there is f with $f \equiv r_i \pmod{m_i}$ for each i, unique modulo m, and exactly one such f has $\deg f < \deg m$ (Theorem 6.6.4).

(d) False. x, $x-1$, $x(x-1)$ have gcd 1 as a family, but $f \equiv 0 \pmod{x}$, $f \equiv 0 \pmod{x-1}$, $f \equiv 1 \pmod{x(x-1)}$ has no solution: the last condition gives $f(0) = 1$, contradicting the first.

(e) 4, with basis $(1, x, x^2, x^3)$ of cosets modulo $\langle x^4 + 1\rangle$, by Theorem 6.6.5.

(f) Evaluation at the i-th node, $\varepsilon_{c_i} : p \mapsto p(c_i)$ (Corollary 6.6.2).

6.6.5.2 B. Practice

Exercise. B1: A Lagrange interpolant

Let $c_0, c_1, c_2, c_3 = 0, 1, 2, 3$. Write down the Lagrange basis for these nodes in factored form, and use it to find the $p \in \mathbb{Q}[x]_{\le 3}$ with $p(0) = 1$, $p(1) = 0$, $p(2) = 1$, and $p(3) = 10$. Hence compute $p(4)$.

Solution. The denominators $\prod_{j\neq i}(c_i - c_j)$ are $(-1)(-2)(-3) = -6$, $(1)(-1)(-2) = 2$, $(2)(1)(-1) = -2$, and $(3)(2)(1) = 6$. So

$$\begin{aligned}\ell_0 &= -\tfrac{1}{6}(x-1)(x-2)(x-3),\\ \ell_1 &= \tfrac{1}{2}x(x-2)(x-3),\\ \ell_2 &= -\tfrac{1}{2}x(x-1)(x-3),\\ \ell_3 &= \tfrac{1}{6}x(x-1)(x-2).\end{aligned}$$

By Theorem 6.6.1 (c), $p = \ell_0 + 0\cdot\ell_1 + \ell_2 + 10\ell_3$. Expanding each needed term,

$$\begin{aligned}\ell_0 &= -\tfrac{1}{6}(x^3 - 6x^2 + 11x - 6),\\ \ell_2 &= -\tfrac{1}{2}(x^3 - 4x^2 + 3x),\\ 10\ell_3 &= \tfrac{5}{3}(x^3 - 3x^2 + 2x).\end{aligned}$$

The coefficient of x^3 is $-\frac{1}{6} - \frac{1}{2} + \frac{5}{3} = 1$; of x^2, $1 + 2 - 5 = -2$; of x, $-\frac{11}{6} - \frac{3}{2} + \frac{10}{3} = 0$; the constant is 1. Hence

$$p = x^3 - 2x^2 + 1.$$

Check: $p(0) = 1$, $p(1) = 0$, $p(2) = 8 - 8 + 1 = 1$, $p(3) = 27 - 18 + 1 = 10$. Hence $p(4) = 64 - 32 + 1 = 33$.

Exercise. B2: A CRT system

Find the polynomial $f \in \mathbb{Q}[x]$ of degree less than 3 with $f \equiv 2 \pmod{x-1}$ and $f \equiv x \pmod{x^2+1}$. Hence describe all $g \in \mathbb{Q}[x]$ satisfying both congruences.

Solution. Bézout. $x^2 + 1 = (x+1)(x-1) + 2$, so $\frac{1}{2}(x^2+1) - \frac{1}{2}(x+1)(x-1) = 1$, and the moduli are coprime.

The elements e_i. $e_1 = \frac{1}{2}(x^2+1)$ is a multiple of x^2+1 and $e_1 = 1 + \frac{1}{2}(x+1)(x-1) \equiv 1 \pmod{x-1}$. And $e_2 = 1 - e_1 = -\frac{1}{2}(x+1)(x-1)$ is a multiple of $x-1$ with $e_2 \equiv 1 \pmod{x^2+1}$.

Combine and reduce. By the proof of Theorem 6.6.4, $f \equiv 2e_1 + xe_2 = (x^2+1) - \frac{1}{2}x(x^2-1) = -\frac{1}{2}x^3 + x^2 + \frac{1}{2}x + 1$. Modulo $m = (x-1)(x^2+1) = x^3 - x^2 + x - 1$, add $\frac{1}{2}m$:

$$f = -\tfrac{1}{2}x^3 + x^2 + \tfrac{1}{2}x + 1 + \tfrac{1}{2}(x^3 - x^2 + x - 1) = \tfrac{1}{2}x^2 + x + \tfrac{1}{2} = \tfrac{1}{2}(x+1)^2.$$

Check. $f(1) = 2$. Modulo x^2+1, $f \equiv \frac{1}{2}(-1) + x + \frac{1}{2} = x$.

By Theorem 6.6.4 (b), and since every $g \equiv f \pmod{m}$ satisfies both congruences (because $x-1$ and x^2+1 divide m), the solutions are exactly $g = \frac{1}{2}(x+1)^2 + h\,(x-1)(x^2+1)$ with $h \in \mathbb{Q}[x]$.

Exercise. B3: The Lagrange basis sums to one

Let $c_0, \dots, c_n \in F$ be distinct with Lagrange basis $(\ell_0, \dots, \ell_n)$.

(a) Prove that $\ell_0 + \ell_1 + \cdots + \ell_n = 1$.

(b) Prove more generally that $\sum_{i=0}^{n} c_i^k \, \ell_i = x^k$ for $k = 0, 1, \ldots, n$ (with $c^0 = 1$).

(c) For $n = 2$ and nodes $0, 1, 2$, verify (a) by expanding.

Solution.

(a) This is (b) with $k = 0$.

(b) Let $0 \le k \le n$. Then $x^k \in F[x]_{\le n}$, and its value at c_i is c_i^k (Definition 0.8.6). By Theorem 6.6.1 (b), $x^k = \sum_i c_i^k \ell_i$.

(c) $\ell_0 = \frac{(x-1)(x-2)}{2} = \frac{1}{2}x^2 - \frac{3}{2}x + 1$, $\ell_1 = \frac{x(x-2)}{-1} = -x^2 + 2x$, $\ell_2 = \frac{x(x-1)}{2} = \frac{1}{2}x^2 - \frac{1}{2}x$. Adding: x^2 has coefficient $\frac{1}{2} - 1 + \frac{1}{2} = 0$, x has $-\frac{3}{2} + 2 - \frac{1}{2} = 0$, and the constant is 1. So the sum is 1.

6.6.5.3 C. Going deeper

Exercise. C1: Interpolation over a finite field

Let p be a prime and $F = \mathbb{F}_p$.

(a) Explain why Theorem 6.6.1 cannot be applied with more than p nodes.

(b) Let $E\colon \mathbb{F}_p[x]_{\le p} \to \mathbb{F}_p^p$, $E(f) = (f(0), f(1), \ldots, f(p-1))$. Prove that E is linear and surjective but not injective, and find a nonzero polynomial in $\ker E$.

(c) Deduce that every function $\mathbb{F}_p \to \mathbb{F}_p$ is the polynomial function of exactly one polynomial of degree at most $p-1$, but of more than one polynomial of degree at most p.

Solution.

(a) The theorem needs distinct nodes, and $\mathbb{F}_p$ has only p elements, so there are no $p+1$ distinct nodes.

(b) *Linear.* Each entry $f \mapsto f(c)$ is linear by Theorem 0.8.5. *Surjective.* The p elements $0, 1, \ldots, p-1$ of $\mathbb{F}_p$ are distinct nodes, so for every $(y_0, \ldots, y_{p-1})$, Theorem 6.6.1 (c) with $n = p-1$ gives $f = \sum_i y_i \ell_i \in \mathbb{F}_p[x]_{\le p-1} \subseteq \mathbb{F}_p[x]_{\le p}$ with $E(f) = (y_0, \ldots, y_{p-1})$. *Not injective.* $\dim \mathbb{F}_p[x]_{\le p} = p + 1 > p = \dim \mathbb{F}_p^p$, so E is not injective by Corollary 2.3.2 (c). Concretely, $g = \prod_{c \in \mathbb{F}_p}(x - c)$ is monic of degree p, hence nonzero, and $g(c) = 0$ for every $c \in \mathbb{F}_p$, because one factor vanishes (Theorem 0.8.5).

(c) A function $\mathbb{F}_p \to \mathbb{F}_p$ is the same as its list of values $(y_0, \ldots, y_{p-1})$. By (b), the restriction of E to $\mathbb{F}_p[x]_{\le p-1}$ is surjective, and it is injective because a polynomial of degree at most $p-1$ vanishing at p distinct points is 0 (Corollary 6.1.4). So exactly one polynomial of degree at most $p-1$ has the given values. In degree at most p, if f has the given values, so does $f + g \ne f$, with g from (b). Unlike over $\mathbb{R}$, a polynomial over $\mathbb{F}_p$ is not determined by its function.

Exercise. C2: Newton's form of the interpolant

Let $c_0, \ldots, c_n \in F$ be distinct. Put $N_0 = 1$ and $N_k = (x - c_0)(x - c_1) \cdots (x - c_{k-1})$ for

$1 \le k \le n+1$.

(a) Prove that $(N_0, \dots, N_n)$ is a basis of $F[x]_{\le n}$.

(b) Let $p = \sum_{k=0}^{n} a_k N_k$. Show that $p(c_i) = \sum_{k=0}^{i} a_k N_k(c_i)$ and $N_i(c_i) \neq 0$. Explain why the coefficients $a_0, a_1, \dots$ of the interpolant can be found one at a time, and show that $a_0 = y_0$ and $a_1 = \frac{y_1 - y_0}{c_1 - c_0}$.

(c) Let $p_n \in F[x]_{\le n}$ interpolate $y_0, \dots, y_n$ at $c_0, \dots, c_n$, and let c_{n+1} be a further node, distinct from the others, with a value y_{n+1}. Prove that the interpolant $p_{n+1} \in F[x]_{\le n+1}$ is

$$p_{n+1} = p_n + \frac{y_{n+1} - p_n(c_{n+1})}{N_{n+1}(c_{n+1})} N_{n+1}.$$

(d) Find the Newton form of the interpolant of Exercise (B1: A Lagrange interpolant), and then the interpolant after adding the point $(4, 0)$. Why is this more economical than recomputing with Lagrange's formula?

Hint: for (c), check the values of the right-hand side at all $n+2$ nodes.

Solution.

(a) N_k is a product of k monic linear factors, so it is monic of degree k (Theorem 0.8.3). The polynomials $N_0, \dots, N_n$ are nonzero with pairwise distinct degrees, hence linearly independent by Theorem 1.4.4. The list has length $n+1 = \dim F[x]_{\le n}$, so it is a basis by Theorem 1.6.4 (a).

(b) For $k > i$, N_k contains the factor $x - c_i$, so $N_k(c_i) = 0$; this gives the formula for $p(c_i)$. And $N_i(c_i) = \prod_{j<i}(c_i - c_j)$ is a product of nonzero scalars, since the nodes are distinct, so it is nonzero. The conditions $p(c_i) = y_i$ therefore form a lower triangular system: the i-th equation involves only $a_0, \dots, a_i$, with coefficient $N_i(c_i) \neq 0$ on a_i. Solving in the order $i = 0, 1, \dots, n$ determines each a_i from the earlier ones. For $i = 0$: $p(c_0) = a_0$, so $a_0 = y_0$. For $i = 1$: $y_1 = a_0 + a_1(c_1 - c_0)$, so $a_1 = \frac{y_1 - y_0}{c_1 - c_0}$, the slope of the chord, the first "divided difference".

(c) Let q be the right-hand side; the division is legal because $N_{n+1}(c_{n+1}) \neq 0$ by (b). Then $\deg q \le n+1$. For $i \le n$, $N_{n+1}(c_i) = 0$, so $q(c_i) = p_n(c_i) = y_i$. At c_{n+1}, $q(c_{n+1}) = p_n(c_{n+1}) + y_{n+1} - p_n(c_{n+1}) = y_{n+1}$. By Theorem 3.7.3, $q = p_{n+1}$.

(d) With nodes $0, 1, 2, 3$: $N_1 = x$, $N_2 = x(x-1)$, $N_3 = x(x-1)(x-2)$. Forward substitution: $a_0 = p(0) = 1$; $0 = p(1) = 1 + a_1$, so $a_1 = -1$; $1 = p(2) = 1 - 2 + 2a_2$, so $a_2 = 1$; $10 = p(3) = 1 - 3 + 6 + 6a_3$, so $a_3 = 1$. Thus

$$p_3 = 1 - x + x(x-1) + x(x-1)(x-2) = x^3 - 2x^2 + 1,$$

agreeing with Exercise (B1: A Lagrange interpolant). Adding $(4, 0)$: $p_3(4) = 33$ and $N_4(4) = 4 \cdot 3 \cdot 2 \cdot 1 = 24$, so by (c)

$$\begin{aligned} p_4 &= x^3 - 2x^2 + 1 - \tfrac{33}{24}\, x(x-1)(x-2)(x-3) \\ &= x^3 - 2x^2 + 1 - \tfrac{11}{8}\, x(x-1)(x-2)(x-3). \end{aligned}$$

One checks $p_4(4) = 33 - \frac{11}{8} \cdot 24 = 0$. Adding a node changes **every** Lagrange polynomial, since each ℓ_i acquires a new factor; in Newton's form the old coefficients $a_0, \dots, a_n$ survive and only one new coefficient is computed.

6.7 Polynomials of an Operator

Chapter 2 taught us to substitute an operator into a polynomial: for $T \in \mathcal{L}(V)$ and $p \in F[x]$, the operator $p(T)$ replaces each x^k by T^k (Definition 2.4.2). That section used the idea for a few examples. With the arithmetic of this chapter in hand, the idea becomes a method. Divisibility, ideals, and Bézout's identity in $F[x]$ turn into statements about T: every operator on a finite-dimensional space satisfies a polynomial equation, the polynomials it satisfies form an ideal, and a factorization of such a polynomial into coprime pieces splits the space. The last statement, the kernel splitting lemma, is the engine that Chapter 9 will run to take operators apart.

6.7.1 Evaluating at an operator

Fix $T \in \mathcal{L}(V)$, where V is any vector space over F. Theorem 2.4.3 already says that substituting T respects sums and products of polynomials, and that any two polynomials in T commute. The new point of view is to treat "substitute T" as a single map from $F[x]$ to $\mathcal{L}(V)$, and to ask what it preserves. Both sides are vector spaces, and both have a multiplication. The map respects all of the structure.

Theorem 6.7.1. Evaluation at an Operator

Let V be a vector space over F and $T \in \mathcal{L}(V)$. Then the map

$$F[x] \to \mathcal{L}(V), \qquad p \mapsto p(T),$$

has the following properties.

(a) It is **linear**: $(cp+q)(T) = c\,p(T) + q(T)$ for all $p, q \in F[x]$ and $c \in F$.

(b) It is **multiplicative** and sends 1 to id_V: $(pq)(T) = p(T)\,q(T)$ and $1(T) = \mathrm{id}_V$.

(c) Its image $\{p(T) : p \in F[x]\}$ is a subspace of $\mathcal{L}(V)$, closed under composition, in which any two operators commute.

(d) If $S \in \mathcal{L}(V)$ commutes with T, then S commutes with $p(T)$ for every $p \in F[x]$.

Proof.

(a) Let $p = \sum_k a_k x^k$ and $q = \sum_k b_k x^k$. The coefficient of x^k in $cp+q$ is $ca_k + b_k$, so by Definition 2.4.2 and the vector space laws in $\mathcal{L}(V)$ (Theorem 2.4.1),

$$(cp+q)(T) = \sum_k (ca_k + b_k)T^k = c\sum_k a_k T^k + \sum_k b_k T^k = c\,p(T) + q(T).$$

(b) The product rule is Theorem 2.4.3 (b). The constant polynomial 1 gives $1 \cdot T^0 = \mathrm{id}_V$.

(c) The image of a linear map is a subspace (Theorem 2.2.2). For polynomials p, q, the composite $p(T)q(T)$ equals $(pq)(T)$ by (b), so it lies in the image, and $p(T)q(T) = q(T)p(T)$ by Theorem 2.4.3 (c).

(d) Suppose $ST = TS$. We show $ST^k = T^k S$ by induction on k. For $k = 0$, both sides are S. If $ST^k = T^k S$, then by associativity (Theorem 2.4.2)

$$ST^{k+1} = (ST^k)T = T^k(ST) = T^k(TS) = T^{k+1}S.$$

For $p = \sum_k a_k x^k$, distributivity (Theorem 2.4.2) gives $S\,p(T) = \sum_k a_k\,ST^k = \sum_k a_k\,T^k S = p(T)\,S$.

■

Parts (a) and (b) together say that $p \mapsto p(T)$ is a **homomorphism of algebras**: it translates every identity built from sums, scalar multiples, and products of polynomials into the same identity for operators. (This map should not be confused with the evaluation map $V \to V^{**}$ of Chapter 5; here we evaluate polynomials, not functionals.) Part (d) is what we will use most. It says that anything commuting with T commutes with the whole family of polynomials in T.

For matrices nothing new is needed. If V is finite-dimensional with basis $\mathcal{B}$, then $[p(T)]_{\mathcal{B}}^{\mathcal{B}} = p([T]_{\mathcal{B}}^{\mathcal{B}})$ by Corollary 4.1.3. So every statement in this section holds verbatim for $\mathbf{A} \in M_n(F)$, applied to the operator $T_{\mathbf{A}}$.

Example. Two Ways to Evaluate

Let $\mathbf{A} = \begin{pmatrix} 3 & 1 \\ -1 & 1 \end{pmatrix} \in M_2(\mathbb{Q})$. Compute $p(\mathbf{A})$ for $p = x^2 - 4x + 4$ and $q = x^2 - 3x$, once by expanding and once by factoring.

Solution. Expanding. $\mathbf{A}^2 = \begin{pmatrix} 9-1 & 3+1 \\ -3-1 & -1+1 \end{pmatrix} = \begin{pmatrix} 8 & 4 \\ -4 & 0 \end{pmatrix}$. Hence

$$p(\mathbf{A}) = \begin{pmatrix} 8-12+4 & 4-4 \\ -4+4 & 0-4+4 \end{pmatrix} = \mathbf{O},$$

$$q(\mathbf{A}) = \begin{pmatrix} 8-9 & 4-3 \\ -4+3 & 0-3 \end{pmatrix} = \begin{pmatrix} -1 & 1 \\ -1 & -3 \end{pmatrix}.$$

Factoring. $p = (x-2)^2$, so $p(\mathbf{A}) = (\mathbf{A}-2\mathbf{I})^2$ by Theorem 6.7.1 (b). Here $\mathbf{A}-2\mathbf{I} = \begin{pmatrix} 1 & 1 \\ -1 & -1 \end{pmatrix}$, whose square is $\begin{pmatrix} 1-1 & 1-1 \\ -1+1 & -1+1 \end{pmatrix} = \mathbf{O}$. Likewise $q = x(x-3)$, so

$$q(\mathbf{A}) = \mathbf{A}(\mathbf{A} - 3\mathbf{I}) = \begin{pmatrix} 3 & 1 \\ -1 & 1 \end{pmatrix}\begin{pmatrix} 0 & 1 \\ -1 & -2 \end{pmatrix} = \begin{pmatrix} -1 & 1 \\ -1 & -3 \end{pmatrix},$$

in agreement with the expansion. The factored form showed at a glance why $p(\mathbf{A}) = \mathbf{O}$: the matrix $\mathbf{A} - 2\mathbf{I}$ squares to zero.

Warning. **A product of polynomials in T can vanish without either factor vanishing.** In $F[x]$, $pq = 0$ forces $p = 0$ or $q = 0$ (Corollary 0.8.4). Evaluation at T does not carry this over. For the projection $P(x, y) = (x, 0)$ on F^2, the polynomials $p = x$ and $q = x-1$ give $p(P)\,q(P) = P^2 - P = 0$, while $p(P) = P \neq 0$ and $q(P) = P - \mathrm{id} \neq 0$. The map $p \mapsto p(T)$ respects products, but it is not injective, and it is exactly its kernel that makes this possible.

6.7.2 Annihilating polynomials

In Chapter 2 we saw operators satisfying polynomial equations: a projection satisfies $P^2 - P = 0$, the rotation of $\mathbb{R}^2$ by a right angle satisfies $R^2 + \mathrm{id} = 0$ (Example (Operators That Satisfy a Polynomial)), and differentiation on $F[x]_{\le n}$ satisfies $D^{n+1} = 0$ (Example (Differentiation Is Nilpotent)). Each such equation is a nonzero polynomial in the kernel of $p \mapsto p(T)$. Does every operator have one? We need a word for these polynomials.

Definition 6.7.1. Annihilating Polynomial

Let $T \in \mathcal{L}(V)$. A polynomial $p \in F[x]$ **annihilates** T if $p(T) = 0$, the zero operator on V.

For example, $x^2 - x$ annihilates every projection, $x^2 + 1$ annihilates the rotation R, and the zero polynomial annihilates every operator. On the zero space $V = \{0\}$, even the constant 1 annihilates T, because $1(T) = \mathrm{id}_V$ is the zero map there. On $V \neq \{0\}$, a nonzero constant c never annihilates, since $c\,\mathrm{id}_V \neq 0$. (The word "annihilate" is also used for the annihilator U^0 of Chapter 5; the two uses share the idea "sends to zero" and nothing more.)

In finite dimension a counting argument settles existence. The space $\mathcal{L}(V)$ is finite-dimensional, so the powers of T cannot all be independent.

Theorem 6.7.2. Every Operator Satisfies a Polynomial

Let V be a finite-dimensional vector space over F with $\dim V = n$, and let $T \in \mathcal{L}(V)$. Then there is a **nonzero** $p \in F[x]$ with $\deg p \le n^2$ and $p(T) = 0$. The same holds for every $\mathbf{A} \in M_n(F)$.

Proof. If $n = 0$, the constant 1 works, as noted above. Let $n \ge 1$. By Theorem 4.1.4, $\dim \mathcal{L}(V) = n \cdot n = n^2$. The list $(\mathrm{id}_V, T, T^2, \dots, T^{n^2})$ in $\mathcal{L}(V)$ has length $n^2 + 1 > n^2$, so it is linearly dependent by Theorem 1.6.3 (a). Hence there are $a_0, \dots, a_{n^2} \in F$, **not all zero**, with $a_0\,\mathrm{id}_V + a_1 T + \cdots + a_{n^2} T^{n^2} = 0$. The polynomial $p = a_0 + a_1 x + \cdots + a_{n^2} x^{n^2}$ is nonzero, has degree at most n^2, and satisfies $p(T) = 0$ by Definition 2.4.2. For $\mathbf{A} \in M_n(F)$, apply the same argument in $M_n(F)$, which has dimension n^2 (Example (Dimensions of the Running Examples)).

■

The bound n^2 is crude. In Chapter 9, the Cayley–Hamilton Theorem will produce an annihilating polynomial of degree exactly n. The finite dimension is essential: on $F[x]$, let S be multiplication by x, $S(f) = xf$. Then $S^k(1) = x^k$, so $p(S)(1) = p$ for every $p \in F[x]$, and $p(S) \neq 0$ whenever $p \neq 0$.

The proof is also an algorithm: search for the first power of $\mathbf{A}$ that is a combination of the earlier ones. This is a question about linear dependence, which elimination answers (Theorem 3.5.10).

Example. Finding an Annihilating Polynomial

Find a nonzero polynomial of least possible degree that annihilates $\mathbf{B} = \begin{pmatrix} 2 & 1 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \in M_3(\mathbb{Q})$.

Solution. *Degree* 1. $\mathbf{B}$ is not a scalar multiple of $\mathbf{I}_3$ (its $(1,2)$-entry is nonzero), so $(\mathbf{I}_3, \mathbf{B})$ is independent, and no nonzero polynomial of degree at most 1 annihilates $\mathbf{B}$.

Degree 2. Compute

$$\mathbf{B}^2 = \begin{pmatrix} 4 & 3 & 3 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

We look for a, b, c with $a\mathbf{I}_3 + b\mathbf{B} + c\mathbf{B}^2 = \mathbf{O}$. Reading entries: position $(1,2)$ gives $b + 3c = 0$; position $(1,1)$ gives $a + 2b + 4c = 0$; position $(2,2)$ gives $a + b + c = 0$. The positions $(1,3)$ and $(3,3)$ repeat the $(1,2)$ and $(2,2)$ equations, and all other entries are $0 = 0$. With $c = 1$: $b = -3$ and $a = 2$, and indeed $2 - 6 + 4 = 0$. So

$$\mathbf{B}^2 - 3\mathbf{B} + 2\mathbf{I}_3 = \mathbf{O},$$

and $p = x^2 - 3x + 2 = (x-1)(x-2)$ annihilates $\mathbf{B}$. Check with the factored form: $\mathbf{B} - \mathbf{I}_3 = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$ and $\mathbf{B} - 2\mathbf{I}_3 = \begin{pmatrix} 0 & 1 & 1 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{pmatrix}$, whose product is $\begin{pmatrix} 0 & 1-1 & 1-1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} = \mathbf{O}$.

The bound of Theorem 6.7.2 allowed degree 9; the search stopped at degree 2.

Annihilating polynomials are far from unique: if p annihilates T, so does every multiple qp, since $(qp)(T) = q(T)p(T) = q(T) \cdot 0 = 0$. In Example (Finding an Annihilating Polynomial), $x^3 - 3x^2 + 2x$ annihilates $\mathbf{B}$ too. This closure under multiplication is the defining property of an ideal, and ideals of $F[x]$ are completely understood (Theorem 6.2.2).

Theorem 6.7.3. The Annihilating Polynomials Form an Ideal

Let V be a vector space over F and $T \in \mathcal{L}(V)$, and let

$$I_T := \{p \in F[x] : p(T) = 0\}.$$

(a) I_T is an ideal of $F[x]$.

(b) Either $I_T = \{0\}$, or there is a **unique monic** $m \in F[x]$ with $I_T = \langle m \rangle$. In the second case, for every $p \in F[x]$,

$$p(T) = 0 \iff m \mid p,$$

and m has the least degree among the nonzero polynomials annihilating T.

(c) If V is finite-dimensional, then $I_T \neq \{0\}$, and $\deg m \leq (\dim V)^2$.

Proof.

(a) By Theorem 6.7.1 (a), I_T is the kernel of a linear map, so it is a subspace (Theorem 2.2.1); in particular it contains 0 and is closed under addition. If $p \in I_T$ and

$q \in F[x]$, then $(qp)(T) = q(T)\,p(T) = 0$ by Theorem 6.7.1 (b), so $qp \in I_T$. By Definition 6.2.1, I_T is an ideal.

(b) By Theorem 6.2.2, either $I_T = \{0\}$, or $I_T = \langle m \rangle$ for a unique monic m, which has least degree among the nonzero elements of I_T. In the second case, $p(T) = 0$ means $p \in \langle m \rangle$, that is, $p = qm$ for some q, that is, $m \mid p$.

(c) By Theorem 6.7.2, I_T contains a nonzero polynomial of degree at most $(\dim V)^2$. So $I_T \neq \{0\}$, and $\deg m \leq (\dim V)^2$ because m has the least degree.

■

This monic generator is one of the most important invariants of an operator. **Chapter 9 names it the minimal polynomial of** T and relates its roots to eigenvalues; here it is simply "the monic generator of I_T". Part (b) turns every polynomial identity satisfied by T into a divisibility statement about one fixed polynomial.

Examples.

- **Projections.** Let P be a projection with $P \neq 0$ and $P \neq \mathrm{id}_V$. Then $x^2 - x = x(x-1) \in I_P$, so the monic generator divides $x(x-1)$. By unique factorization (Theorem 6.3.5), its monic divisors are 1, x, $x-1$, and $x(x-1)$. The first three do not annihilate P: $1(P) = \mathrm{id}_V \neq 0$, $x(P) = P \neq 0$, and $(x-1)(P) = P - \mathrm{id}_V \neq 0$. So the generator is $x^2 - x$.
- **Degenerate cases.** For the identity on $V \neq \{0\}$, the generator is $x - 1$; for the zero operator on $V \neq \{0\}$, it is x; on $V = \{0\}$, it is 1.
- **The matrix B.** In Example (Finding an Annihilating Polynomial), no nonzero polynomial of degree at most 1 annihilates $\mathbf{B}$, and $x^2 - 3x + 2$ does, so the generator is $x^2 - 3x + 2$.
- **Infinite dimension.** For multiplication by x on $F[x]$, we saw that $I_S = \{0\}$.

Quick check. Let $R(x, y) = (-y, x)$ be the rotation of $\mathbb{R}^2$ by a right angle. What is the monic generator of $I_R \subseteq \mathbb{R}[x]$?

Solution. $R^2 = -\mathrm{id}$, so $x^2 + 1 \in I_R$, and the generator divides $x^2 + 1$. Over $\mathbb{R}$, $x^2 + 1$ has no root, so its only monic divisors are 1 and $x^2 + 1$ (a monic divisor of degree 1 would be $x - c$ with c a root). Since $1(R) = \mathrm{id} \neq 0$, the generator is $x^2 + 1$.

6.7.3 Kernels of polynomials in T

The subspaces we can build from T and a polynomial p are $\ker p(T)$ and $\operatorname{im} p(T)$. For $p = x - \lambda$, the kernel $\ker(T - \lambda\,\mathrm{id}_V)$ consists of the vectors that T merely scales by λ; for a projection, $\ker P$ and $\ker(P - \mathrm{id}_V)$ are the two pieces of the direct sum. These subspaces have a property that general subspaces lack: T does not move vectors out of them.

Theorem 6.7.4. T **Preserves Kernels and Images of Polynomials in** T

(a) Let $R, S \in \mathcal{L}(V)$ with $SR = RS$. Then $Sv \in \ker R$ for every $v \in \ker R$, and $Sw \in \operatorname{im} R$ for every $w \in \operatorname{im} R$.

(b) In particular, for $T \in \mathcal{L}(V)$ and $p, q \in F[x]$, the operator $q(T)$ maps $\ker p(T)$ into $\ker p(T)$ and $\operatorname{im} p(T)$ into $\operatorname{im} p(T)$. Taking $q = x$: T maps $\ker p(T)$ and $\operatorname{im} p(T)$ into

themselves.

Proof.

(a) Let $v \in \ker R$. Then $R(Sv) = S(Rv) = S0 = 0$, so $Sv \in \ker R$. Let $w = Ru \in \operatorname{im} R$. Then $Sw = S(Ru) = R(Su) \in \operatorname{im} R$.

(b) By Theorem 6.7.1 (c), $q(T)$ commutes with $p(T)$, so (a) applies with $R = p(T)$ and $S = q(T)$. For $q = x$, $q(T) = T$.

■

In Chapter 9, a subspace that an operator maps into itself will be called an invariant subspace, and the notion will be developed properly there. For now the theorem is what matters: whenever $U = \ker p(T)$, the rule $u \mapsto Tu$ is an operator on U, and so is $u \mapsto q(T)u$.

The inclusion can fail for subspaces not built from T. For the rotation R of the check above, R moves the x-axis to the y-axis, so the x-axis is not of the form $\ker p(R)$ or $\operatorname{im} p(R)$ for any p.

6.7.4 The kernel splitting lemma

Chapter 2 contains two theorems with the same shape. A projection satisfies $P(P - \mathrm{id}_V) = 0$, and Theorem 4.5.1 splits V as $\operatorname{im} P \oplus \ker P$. An involution in characteristic not 2 satisfies $(T - \mathrm{id}_V)(T + \mathrm{id}_V) = 0$, and Theorem 4.5.5 splits V as $\ker(T - \mathrm{id}_V) \oplus \ker(T + \mathrm{id}_V)$. In both, a polynomial annihilating T factors into two pieces with no common factor, and V is the direct sum of the kernels of the pieces. In both, the projections turned out to be polynomials in T: P itself, and $\frac{1}{2}(\mathrm{id}_V + T)$. The general statement needs only the arithmetic of $F[x]$.

Theorem 6.7.5. Kernel Splitting Lemma

Let V be a vector space over F, let $T \in \mathcal{L}(V)$, and let $p, q \in F[x]$ be **coprime**. Choose $a, b \in F[x]$ with $ap + bq = 1$ (Corollary 6.2.4), and put

$$W = \ker(pq)(T), \qquad E_1 = b(T)\,q(T), \qquad E_2 = a(T)\,p(T).$$

Then:

(a) $W = \ker p(T) \oplus \ker q(T)$;

(b) $E_1 + E_2 = \mathrm{id}_V$, and for every $w \in W$, the splitting of w in (a) is

$$w = E_1 w + E_2 w, \qquad E_1 w \in \ker p(T), \quad E_2 w \in \ker q(T);$$

so E_1 maps W into W and acts on W as the projection onto $\ker p(T)$ along $\ker q(T)$, and E_2 acts on W as the projection onto $\ker q(T)$ along $\ker p(T)$;

(c) in particular, if $(pq)(T) = 0$, then

$$V = \ker p(T) \oplus \ker q(T),$$

and the projections of V onto the two summands along each other are the polynomials $(bq)(T)$ and $(ap)(T)$ in T.

Idea. Substitute T into Bézout's identity: $\mathrm{id}_V = a(T)p(T) + b(T)q(T)$. Applied to a vector w, this writes $w = E_1 w + E_2 w$, and it does so for **every** vector of V. Now look at the pieces. ① $p(T)$ kills $E_1 w = b(T)q(T)w$, because $p(T)b(T)q(T) = b(T)\,(pq)(T)$ and $(pq)(T)w = 0$ for $w \in W$: the factor q in E_1 combines with p to the whole product. Symmetrically $q(T)$ kills $E_2 w$. ② If w is killed by both $p(T)$ and $q(T)$, then both pieces are 0, so $w = 0$. ③ Direct sum by the intersection criterion; the projections are read off from uniqueness. Coprimality enters only through the identity $ap + bq = 1$, and without it step ② has nothing to stand on.

Proof. Since $ap + bq = 1$, Theorem 6.7.1 gives

$$\mathrm{id}_V = a(T)p(T) + b(T)q(T) = E_2 + E_1.$$

Throughout, polynomials in T commute with each other (Theorem 6.7.1 (c)).

The kernels lie in W. If $p(T)v = 0$, then $(pq)(T)v = q(T)\,p(T)v = 0$; so $\ker p(T) \subseteq W$, and likewise $\ker q(T) \subseteq W$.

The pieces. Let $w \in W$. Then $w = \mathrm{id}_V\, w = E_1 w + E_2 w$. Moreover

$$p(T)(E_1 w) = p(T)\,b(T)\,q(T)w = b(T)\,(pq)(T)w = b(T)0 = 0,$$

so $E_1 w \in \ker p(T)$, and similarly $q(T)(E_2 w) = a(T)\,(pq)(T)w = 0$, so $E_2 w \in \ker q(T)$. Hence $W = \ker p(T) + \ker q(T)$.

The intersection. Let $v \in \ker p(T) \cap \ker q(T)$. Then $E_1 v = b(T)\,q(T)v = 0$ and $E_2 v = a(T)\,p(T)v = 0$, so $v = E_1 v + E_2 v = 0$. By Theorem 1.7.4, the sum is direct, which proves (a).

The projections. The splitting of $w \in W$ in (a) is unique (Definition 1.7.2), and we have exhibited it as $w = E_1 w + E_2 w$; this proves the formula in (b). In particular $E_1 w \in \ker p(T) \subseteq W$, so E_1 maps W into W. As an operator on W, E_1 is linear, fixes every $u \in \ker p(T)$ (whose splitting is $u = u + 0$), and kills every vector of $\ker q(T)$ (whose splitting is $0 + v$). By the uniqueness in Theorem 4.5.1 (b), applied to $W = \ker p(T) \oplus \ker q(T)$, it is the projection onto $\ker p(T)$ along $\ker q(T)$. Swapping the roles of p and q gives the statement for E_2.

(c) If $(pq)(T) = 0$, then $W = V$, and (a) and (b) become the statements in (c), with $E_1 = (bq)(T)$ and $E_2 = (ap)(T)$ by Theorem 6.7.1 (b). This proves the theorem.

■

Bézout's coefficients a, b are not unique, but the projections on W are, by Theorem 4.5.1 (b). So different choices of a, b give polynomials bq that may differ, while $(bq)(T)$ always does the same thing on W. When $(pq)(T) = 0$, the lemma turns a factorization of an annihilating polynomial into a direct sum decomposition of the whole space, with explicit projections, and each summand is carried into itself by T (Theorem 6.7.4).

Example. Recovering Projections and Direct Sums

Let $P \in \mathcal{L}(V)$ be a projection. Apply Theorem 6.7.5 with $p = x$ and $q = x - 1$, and compare with Theorem 4.5.1 (a).

Solution. Coprime. $1 \cdot x + (-1)(x-1) = 1$, so we may take $a = 1$, $b = -1$, and any common divisor of x and $x-1$ divides 1. *Annihilating.* $(pq)(P) = P^2 - P = 0$ because P is a projection.

By Theorem 6.7.5 (c),

$$V = \ker P \oplus \ker(P - \mathrm{id}_V),$$

with projection $E_1 = b(P)q(P) = -(P - \mathrm{id}_V) = \mathrm{id}_V - P$ onto $\ker P$ along $\ker(P - \mathrm{id}_V)$, and $E_2 = a(P)p(P) = P$ onto $\ker(P - \mathrm{id}_V)$ along $\ker P$.

To compare, note that $\ker(P - \mathrm{id}_V) = \operatorname{im} P$. ($\subseteq$) If $Pv = v$, then $v \in \operatorname{im} P$. ($\supseteq$) If $v = Pu$, then $Pv = P^2u = Pu = v$. So the lemma gives $V = \operatorname{im} P \oplus \ker P$, with P the projection onto $\operatorname{im} P$ along $\ker P$: this is Theorem 4.5.1 (a). It also reproves Exercise (B2: The complementary projection), that $\mathrm{id}_V - P$ is the complementary projection.

Example. Recovering the Involution Decomposition

Let F have characteristic not 2, and let $T \in \mathcal{L}(V)$ with $T^2 = \mathrm{id}_V$. Apply Theorem 6.7.5 with $p = x-1$ and $q = x+1$, and compare with Theorem 4.5.5. What goes wrong in characteristic 2?

Solution. Coprime. Since $2 \neq 0$ in F (Definition 0.7.2),

$$-\tfrac{1}{2}(x-1) + \tfrac{1}{2}(x+1) = 1,$$

so $a = -\frac{1}{2}$ and $b = \frac{1}{2}$ work, and a common divisor of p and q divides 1. *Annihilating.* $(pq)(T) = T^2 - \mathrm{id}_V = 0$.

By Theorem 6.7.5 (c),

$$V = \ker(T - \mathrm{id}_V) \oplus \ker(T + \mathrm{id}_V) = E_+ \oplus E_-,$$

and the projection onto E_+ along E_- is $b(T)q(T) = \frac{1}{2}(T + \mathrm{id}_V)$. These are exactly the conclusions of Theorem 4.5.5. The other projection is $a(T)p(T) = \frac{1}{2}(\mathrm{id}_V - T)$.

In characteristic 2, $x + 1 = x - 1$, so $p = q$ and $\gcd(p, q) = x - 1 \neq 1$. The lemma does not apply, and the warning after Theorem 4.5.5 shows that the conclusion really fails there.

The lemma applies just as well when neither factor is linear, and then it finds subspaces that are not visible from single vectors.

Example. A Splitting with a Repeated Factor

Let $\mathbf{A} = \begin{pmatrix} 0 & 1 & -1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} \in M_3(\mathbb{Q})$. Verify that $\mathbf{A}^2(\mathbf{A} - \mathbf{I}_3) = \mathbf{O}$, and use Theorem 6.7.5 with $p = x^2$, $q = x - 1$ to split $\mathbb{Q}^3$, giving both projections as polynomials in $\mathbf{A}$.

Solution. The relation. $\mathbf{A}^2 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}$, and $\mathbf{A}^3 = \mathbf{A} \cdot \mathbf{A}^2 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} = \mathbf{A}^2$. So $\mathbf{A}^2(\mathbf{A} - \mathbf{I}_3) = \mathbf{A}^3 - \mathbf{A}^2 = \mathbf{O}$.

Bézout. $x^2 = (x+1)(x-1) + 1$, so $1 \cdot x^2 + (-(x+1))(x-1) = 1$: take $a = 1$, $b = -(x+1)$. Then

$$\begin{aligned} \mathbf{E}_1 &= b(\mathbf{A})q(\mathbf{A}) = -(\mathbf{A} + \mathbf{I}_3)(\mathbf{A} - \mathbf{I}_3) = \mathbf{I}_3 - \mathbf{A}^2 \\ &= \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & -1 \\ 0 & 0 & 0 \end{pmatrix}, \\ \mathbf{E}_2 &= a(\mathbf{A})p(\mathbf{A}) = \mathbf{A}^2. \end{aligned}$$

The summands. $\ker \mathbf{A}^2 = \{(x, y, z) : z = 0\} = \operatorname{span}(\mathbf{e}_1, \mathbf{e}_2)$, and $\ker(\mathbf{A} - \mathbf{I}_3)$ is cut out by $-x + y - z = 0$, $-y + z = 0$, $0 = 0$, so it is $\operatorname{span}((0, 1, 1))$. By Theorem 6.7.5 (c),

$$\mathbb{Q}^3 = \operatorname{span}(\mathbf{e}_1, \mathbf{e}_2) \oplus \operatorname{span}((0, 1, 1)),$$

and for $\mathbf{v} = (x, y, z)$, the splitting is $\mathbf{E}_1\mathbf{v} + \mathbf{E}_2\mathbf{v} = (x, y - z, 0) + (0, z, z)$. One checks directly that $(x, y - z, 0)$ has third entry 0 and $(0, z, z)$ is a multiple of $(0, 1, 1)$.

Note that $\ker \mathbf{A} = \operatorname{span}(\mathbf{e}_1)$ is strictly smaller than $\ker \mathbf{A}^2$. The factor $p = x^2$ could not be replaced by x: the polynomial $x(x-1)$ does not annihilate $\mathbf{A}$, since $\mathbf{A}^2 - \mathbf{A} \neq \mathbf{O}$.

Warning. **Without coprimality the kernels need not fill $\ker(pq)(T)$, or even form a direct sum.** Let $T = T_{\mathbf{N}}$ with $\mathbf{N} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$ on F^2, and $p = q = x$. Then $(pq)(T) = \mathbf{N}^2 = \mathbf{O}$, so $\ker(pq)(T) = F^2$. But $\ker p(T) = \ker q(T) = \ker \mathbf{N} = \operatorname{span}(\mathbf{e}_1)$. The two kernels are equal, so their sum is not direct, and $\operatorname{span}(\mathbf{e}_1) + \operatorname{span}(\mathbf{e}_1) = \operatorname{span}(\mathbf{e}_1) \neq F^2$. Here $\gcd(x, x) = x$, and no identity $ax + bx = 1$ exists, since its left side vanishes at 0.

Quick check. Let $T \in \mathcal{L}(V)$ satisfy $T^2 = 3T$, over a field with $3 \neq 0$. Split V with the lemma, and write the projection onto $\ker T$ as a polynomial in T.

Solution. $x(x-3)$ annihilates T. Since $\frac{1}{3}x - \frac{1}{3}(x-3) = 1$, $p = x$ and $q = x - 3$ are coprime with $a = \frac{1}{3}$, $b = -\frac{1}{3}$. By Theorem 6.7.5 (c), $V = \ker T \oplus \ker(T - 3\,\mathrm{id}_V)$, and the projection onto $\ker T$ along $\ker(T - 3\,\mathrm{id}_V)$ is $b(T)q(T) = -\frac{1}{3}(T - 3\,\mathrm{id}_V) = \mathrm{id}_V - \frac{1}{3}T$.

In Chapter 9 the lemma is used with an annihilating polynomial factored into powers of distinct irreducibles, $m = p_1^{e_1} \cdots p_k^{e_k}$. Distinct monic irreducibles are coprime, and so are their powers, so the lemma with k factors (Exercise (C2: Splitting with k coprime factors)) splits V into the pieces $\ker p_i(T)^{e_i}$. That is the primary decomposition, and the canonical forms of Chapter 10 are built on it.

6.7.5 Exercises

6.7.5.1 A. Check your understanding

Exercise. A1

(a) State what it means that $p \mapsto p(T)$ is linear and multiplicative.

(b) True or false: if $p(T)\,q(T) = 0$, then $p(T) = 0$ or $q(T) = 0$. Justify your answer.

(c) Why does every operator on a finite-dimensional space have a nonzero annihilating polynomial? Is the same true on $F[x]$?

(d) True or false: if p annihilates T and $p \mid q$, then q annihilates T. Justify your answer.

(e) State the kernel splitting lemma for $(pq)(T) = 0$, and say where coprimality is used in its proof.

(f) True or false: for every subspace U of V and every $T \in \mathcal{L}(V)$, T maps U into U. Justify your answer.

Solution.

(a) $(cp + q)(T) = c\,p(T) + q(T)$ and $(pq)(T) = p(T)\,q(T)$ for all $p, q \in F[x]$ and $c \in F$ (Theorem 6.7.1).

(b) False. For the projection $P(x, y) = (x, 0)$, $P(P - \mathrm{id}) = 0$ while $P \neq 0$ and $P - \mathrm{id} \neq 0$.

(c) $\mathcal{L}(V)$ has dimension n^2, so the $n^2 + 1$ operators $\mathrm{id}_V, T, \dots, T^{n^2}$ are linearly dependent, and a dependence is a nonzero annihilating polynomial (Theorem 6.7.2). On $F[x]$, multiplication by x has no nonzero annihilating polynomial, since $p(S)(1) = p$.

(d) True. If $q = hp$, then $q(T) = h(T)\,p(T) = 0$. This is the ideal property (Theorem 6.7.3).

(e) If p, q are coprime and $(pq)(T) = 0$, then $V = \ker p(T) \oplus \ker q(T)$, and the projections are polynomials in T (Theorem 6.7.5). Coprimality gives $ap + bq = 1$, hence $\mathrm{id}_V = a(T)p(T) + b(T)q(T)$, which both produces the splitting of each vector and shows the intersection is $\{0\}$.

(f) False. The rotation $R(x, y) = (-y, x)$ maps the x-axis to the y-axis. The statement is true for $U = \ker p(T)$ and $U = \operatorname{im} p(T)$ (Theorem 6.7.4).

6.7.5.2 B. Practice

Exercise. B1: A polynomial of a 2×2 matrix

Let $\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 3 & -1 \end{pmatrix} \in M_2(\mathbb{Q})$.

(a) Compute $p(\mathbf{A})$ for $p = x^2 - 7$ and for $q = x^2 + x - 5$.

(b) Hence show that $\mathbf{A}$ is invertible and find $\mathbf{A}^{-1}$ as a polynomial in $\mathbf{A}$.

Solution.

(a) $\mathbf{A}^2 = \begin{pmatrix} 1+6 & 2-2 \\ 3-3 & 6+1 \end{pmatrix} = 7\mathbf{I}_2$. Hence $p(\mathbf{A}) = \mathbf{A}^2 - 7\mathbf{I}_2 = \mathbf{O}$. For q, divide: $q =$

$(x^2-7)+(x+2)$, so by Theorem 6.7.1 $q(\mathbf{A}) = p(\mathbf{A}) + \mathbf{A} + 2\mathbf{I}_2 = \begin{pmatrix} 3 & 2 \\ 3 & 1 \end{pmatrix}$. (Directly: $7\mathbf{I}_2 + \mathbf{A} - 5\mathbf{I}_2 = \mathbf{A} + 2\mathbf{I}_2$.)

(b) From $\mathbf{A}^2 = 7\mathbf{I}_2$, $\mathbf{A}(\frac{1}{7}\mathbf{A}) = (\frac{1}{7}\mathbf{A})\mathbf{A} = \mathbf{I}_2$. So $\mathbf{A}$ is invertible with $\mathbf{A}^{-1} = \frac{1}{7}\mathbf{A} = \begin{pmatrix} \frac{1}{7} & \frac{2}{7} \\ \frac{3}{7} & -\frac{1}{7} \end{pmatrix}$.

Exercise. B2: An annihilating polynomial by elimination

Let $\mathbf{C} = \begin{pmatrix} 0 & 0 & 2 \\ 1 & 0 & -1 \\ 0 & 1 & 2 \end{pmatrix} \in M_3(\mathbb{Q})$.

(a) Compute $\mathbf{C}^2$ and $\mathbf{C}^3$, and show that $(\mathbf{I}_3, \mathbf{C}, \mathbf{C}^2)$ is linearly independent.

(b) Find a linear dependence among $\mathbf{I}_3, \mathbf{C}, \mathbf{C}^2, \mathbf{C}^3$, and hence a monic polynomial p of degree 3 with $p(\mathbf{C}) = \mathbf{O}$.

(c) Hence show that $\mathbf{C}$ is invertible and write $\mathbf{C}^{-1}$ as a polynomial in $\mathbf{C}$.

Solution.

(a) Multiplying out,

$$\mathbf{C}^2 = \begin{pmatrix} 0 & 2 & 4 \\ 0 & -1 & 0 \\ 1 & 2 & 3 \end{pmatrix}, \qquad \mathbf{C}^3 = \mathbf{C}\cdot\mathbf{C}^2 = \begin{pmatrix} 2 & 4 & 6 \\ -1 & 0 & 1 \\ 2 & 3 & 6 \end{pmatrix}.$$

Let $a\mathbf{I}_3 + b\mathbf{C} + c\mathbf{C}^2 = \mathbf{O}$. The $(2,1)$-entry gives $b = 0$; the $(3,1)$-entry gives $c = 0$; then the $(1,1)$-entry gives $a = 0$. So the list is independent.

(b) Let $a\mathbf{I}_3 + b\mathbf{C} + c\mathbf{C}^2 + d\mathbf{C}^3 = \mathbf{O}$. Reading entries: $(2,1)$: $b - d = 0$; $(3,1)$: $c + 2d = 0$; $(1,1)$: $a + 2d = 0$. With $d = 1$: $b = 1$, $c = -2$, $a = -2$. Check all entries of $-2\mathbf{I}_3 + \mathbf{C} - 2\mathbf{C}^2 + \mathbf{C}^3$: row 1 is $(-2+0-0+2,\ 0-4+4,\ 2-8+6) = (0,0,0)$; row 2 is $(1-0-1,\ -2+0+2+0,\ -1-0+1) = (0,0,0)$; row 3 is $(0-2+2,\ 1-4+3,\ -2+2-6+6) = (0,0,0)$. Hence

$$p = x^3 - 2x^2 + x - 2 = (x-2)(x^2+1)$$

satisfies $p(\mathbf{C}) = \mathbf{O}$. By (a) no nonzero polynomial of degree at most 2 annihilates $\mathbf{C}$, so p is the monic generator of I_C (Theorem 6.7.3).

(c) $\mathbf{C}^3 - 2\mathbf{C}^2 + \mathbf{C} = 2\mathbf{I}_3$, that is, $\mathbf{C}(\mathbf{C}^2 - 2\mathbf{C} + \mathbf{I}_3) = (\mathbf{C}^2 - 2\mathbf{C} + \mathbf{I}_3)\mathbf{C} = 2\mathbf{I}_3$, using Theorem 6.7.1 (c) for the second equality. Since $2 \neq 0$ in $\mathbb{Q}$, $\mathbf{C}$ is invertible with

$$\mathbf{C}^{-1} = \tfrac{1}{2}(\mathbf{C}^2 - 2\mathbf{C} + \mathbf{I}_3) = \begin{pmatrix} \frac{1}{2} & 1 & 0 \\ -1 & 0 & 1 \\ \frac{1}{2} & 0 & 0 \end{pmatrix}.$$

Exercise. B3: Operators with $T^3 = T$

Let F have characteristic not 2, and let $T \in \mathcal{L}(V)$ with $T^3 = T$.

(a) Use Theorem 6.7.5 with $p = x$ and $q = x^2 - 1$ to show $V = \ker T \oplus \ker(T^2 - \mathrm{id}_V)$, with projection $\mathrm{id}_V - T^2$ onto $\ker T$.

(b) Show that $V = \ker T \oplus \ker(T - \mathrm{id}_V) \oplus \ker(T + \mathrm{id}_V)$.

(c) Verify that the projections onto the three summands in (b) are $\mathrm{id}_V - T^2$, $\frac{1}{2}(T^2 + T)$, and $\frac{1}{2}(T^2 - T)$, in the sense that these three operators sum to id_V and each maps V into the corresponding kernel.

Solution.

(a) $x \cdot x + (-1)(x^2 - 1) = 1$: take $a = x$, $b = -1$. Then p and q are coprime, and $(pq)(T) = T^3 - T = 0$. By Theorem 6.7.5 (c), $V = \ker T \oplus \ker(T^2 - \mathrm{id}_V)$, and the projection onto $\ker T$ is $b(T)q(T) = -(T^2 - \mathrm{id}_V) = \mathrm{id}_V - T^2$.

(b) Let $W = \ker(T^2 - \mathrm{id}_V)$. Apply Theorem 6.7.5 (a) with $p_1 = x - 1$, $q_1 = x + 1$, which are coprime as in Example (Recovering the Involution Decomposition): $W = \ker(p_1 q_1)(T) = \ker(T - \mathrm{id}_V) \oplus \ker(T + \mathrm{id}_V)$. Write $K_0 = \ker T$, $K_+ = \ker(T - \mathrm{id}_V)$, $K_- = \ker(T + \mathrm{id}_V)$. Then $V = K_0 + K_+ + K_-$ by (a). If $u_0 + u_+ + u_- = 0$ with $u_0 \in K_0$, $u_\pm \in K_\pm$, then $u_+ + u_- \in W$, so directness in (a) gives $u_0 = 0$ and $u_+ + u_- = 0$, and directness of $W = K_+ \oplus K_-$ gives $u_+ = u_- = 0$. By Theorem 1.7.5 ((b) $\Rightarrow$ (a)), the sum is direct.

(c) The sum is $\mathrm{id}_V - T^2 + \frac{1}{2}T^2 + \frac{1}{2}T + \frac{1}{2}T^2 - \frac{1}{2}T = \mathrm{id}_V$. Using $T^3 = T$: $T(\mathrm{id}_V - T^2) = T - T^3 = 0$; $(T - \mathrm{id}_V)\frac{1}{2}(T^2 + T) = \frac{1}{2}(T^3 - T) = 0$; $(T + \mathrm{id}_V)\frac{1}{2}(T^2 - T) = \frac{1}{2}(T^3 - T) = 0$. So the three operators map V into K_0, K_+, K_- respectively. For $v \in V$, v is the sum of its three images, and by the uniqueness of the splitting in (b) these are the components of v. Hence each operator is the projection onto its summand along the sum of the other two.

6.7.5.3 C. Going deeper

Exercise. C1: Inverses are polynomials

Let V be a vector space over F and $T \in \mathcal{L}(V)$.

(a) Suppose $p(T) = 0$ for some $p \in F[x]$ with $p(0) \neq 0$. Prove that T is invertible and that $T^{-1} = g(T)$ for some $g \in F[x]$.

(b) Deduce that if V is finite-dimensional and T is invertible, then T^{-1} is a polynomial in T.

(c) Show that the monic generator m of I_T, for V finite-dimensional and nonzero, satisfies $m(0) \neq 0$ if and only if T is invertible.

Hint: for (b), write a nonzero annihilating polynomial as $x^k p_1$ with $p_1(0) \neq 0$.

Solution.

(a) Write $p = c + xg_0$, where $c = p(0) \neq 0$ and $g_0 \in F[x]$. Then $0 = p(T) = c\,\mathrm{id}_V + T\,g_0(T)$, so $T\,g_0(T) = -c\,\mathrm{id}_V$. Put $g = -c^{-1}g_0$. Then $T\,g(T) = \mathrm{id}_V$, and $g(T)\,T = T\,g(T) = \mathrm{id}_V$ by Theorem 6.7.1 (c). Hence T is invertible with $T^{-1} = g(T)$.

(b) If $V = \{0\}$, then $T^{-1} = \mathrm{id}_V = 1(T)$. Otherwise, by Theorem 6.7.2 there is a nonzero p

with $p(T) = 0$. Let k be the least index with a nonzero coefficient in p; then $p = x^k p_1$ with $p_1(0) \neq 0$. Now $0 = p(T) = T^k\, p_1(T)$. Since T is invertible, so is T^k, with inverse $(T^{-1})^k$, and composing on the left with it gives $p_1(T) = 0$. By (a), T^{-1} is a polynomial in T.

(c) ($\Rightarrow$) If $m(0) \neq 0$, apply (a) to m. ($\Leftarrow$) Suppose T is invertible and $m(0) = 0$. Then $m = x\, m_1$ with m_1 monic of degree $\deg m - 1$, and $0 = m(T) = T\, m_1(T)$. Composing with T^{-1} gives $m_1(T) = 0$, so $m \mid m_1$ by Theorem 6.7.3 (b). Since $V \neq \{0\}$, m is not constant, so $m_1 \neq 0$ and $\deg m \leq \deg m_1 = \deg m - 1$ by Theorem 0.8.3, a contradiction. Hence $m(0) \neq 0$.

Exercise. C2: Splitting with k coprime factors

Let $T \in \mathcal{L}(V)$, and let $p_1, \dots, p_k \in F[x]$ be nonzero and pairwise coprime, with $k \geq 2$. Put $p = p_1 \cdots p_k$ and $W = \ker p(T)$.

(a) Use Theorem 6.6.4 to find $e_1, \dots, e_k \in F[x]$ with $e_i \equiv 1 \pmod{p_i}$ and $e_i \equiv 0 \pmod{p_j}$ for $j \neq i$. Show that $p \mid e_1 + \cdots + e_k - 1$ and that $p \mid p_i e_i$ for each i.

(b) Prove that for every $w \in W$, $w = e_1(T)w + \cdots + e_k(T)w$ with $e_i(T)w \in \ker p_i(T)$.

(c) Prove that $W = \ker p_1(T) \oplus \cdots \oplus \ker p_k(T)$, and that $e_i(T)$ is the identity on $\ker p_i(T)$ and zero on $\ker p_j(T)$ for $j \neq i$.

(d) Deduce: if $p(T) = 0$, then $V = \ker p_1(T) \oplus \cdots \oplus \ker p_k(T)$, and the projection onto the i-th summand along the others is the polynomial $e_i(T)$.

Hint: for the second divisibility in (a), use Lemma 6.6.3 (b).

Solution.

(a) By Theorem 6.6.4 (a), applied to the pairwise coprime moduli $p_1, \dots, p_k$ with $r_i = 1$ and $r_j = 0$ for $j \neq i$, such an e_i exists. The sum $s = e_1 + \cdots + e_k$ satisfies $s \equiv 1 \pmod{p_j}$ for each j, since exactly one summand is $\equiv 1$ and the others are $\equiv 0$. The constant 1 satisfies the same congruences, so $s \equiv 1 \pmod{p}$ by Theorem 6.6.4 (b). For the second claim, $p_j \mid e_i$ for every $j \neq i$, and these p_j are pairwise coprime, so $\prod_{j \neq i} p_j \mid e_i$ by Lemma 6.6.3 (b). Multiplying by p_i, $p \mid p_i e_i$.

(b) Write $e_1 + \cdots + e_k = 1 + hp$ and $p_i e_i = h_i p$. By Theorem 6.7.1, for $w \in W$,

$$e_1(T)w + \cdots + e_k(T)w = w + h(T)\, p(T)w = w,$$
$$p_i(T)(e_i(T)w) = h_i(T)\, p(T)w = 0.$$

(c) Each $\ker p_i(T) \subseteq W$, since $p(T) = \left(\prod_{j \neq i} p_j\right)(T)\, p_i(T)$. With (b), $W = \ker p_1(T) + \cdots + \ker p_k(T)$. Let $u_j \in \ker p_j(T)$. For $j \neq i$, $p_j \mid e_i$, say $e_i = g\, p_j$, so $e_i(T)u_j = g(T)\, p_j(T)u_j = 0$. And $e_i = 1 + g' p_i$, so $e_i(T)u_i = u_i + g'(T)\, p_i(T)u_i = u_i$. This proves the second statement. Now suppose $u_1 + \cdots + u_k = 0$ with $u_j \in \ker p_j(T)$. Applying $e_i(T)$ gives $u_i = 0$, for each i. By Theorem 1.7.5 ((b) $\Rightarrow$ (a)), the sum is direct.

(d) If $p(T) = 0$, then $W = V$. By (c), V is the direct sum, and $e_i(T) \in \mathcal{L}(V)$ is the identity on the i-th summand and zero on the others. For $v = u_1 + \cdots + u_k$, $e_i(T)v = u_i$, which is the projection onto $\ker p_i(T)$ along the sum of the other summands. For $k = 2$ this is Theorem 6.7.5 (c).

Exercise. C3: The same rotation over $\mathbb{R}$ and over $\mathbb{C}$

Let $\mathbf{A} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$, so that $\mathbf{A}^2 + \mathbf{I}_2 = \mathbf{O}$.

(a) Over $\mathbb{C}$, apply Theorem 6.7.5 to $T_{\mathbf{A}}$ on $\mathbb{C}^2$ with $p = x - i$ and $q = x + i$. Find both kernels and write the projection onto $\ker(\mathbf{A} - i\mathbf{I}_2)$ as a polynomial in $\mathbf{A}$.

(b) Over $\mathbb{R}$, show that the only subspaces of $\mathbb{R}^2$ that $T_{\mathbf{A}}$ maps into themselves are $\{\mathbf{0}\}$ and $\mathbb{R}^2$. Explain why this matches the fact that x^2+1 admits no factorization into coprime nonconstant factors in $\mathbb{R}[x]$.

Solution.

(a) $(x + i) - (x - i) = 2i$, so $\frac{i}{2}(x - i) - \frac{i}{2}(x + i) = \frac{i}{2}(-2i) = 1$: take $a = \frac{i}{2}$ and $b = -\frac{i}{2}$. So p, q are coprime, and $(pq)(\mathbf{A}) = \mathbf{A}^2 + \mathbf{I}_2 = \mathbf{O}$. By Theorem 6.7.5 (c), $\mathbb{C}^2 = \ker(\mathbf{A} - i\mathbf{I}_2) \oplus \ker(\mathbf{A} + i\mathbf{I}_2)$. Solving, $(\mathbf{A} - i\mathbf{I}_2)(z, w) = (-iz - w,\ z - iw) = \mathbf{0}$ gives $w = -iz$, so $\ker(\mathbf{A} - i\mathbf{I}_2) = \operatorname{span}((1, -i))$; similarly $\ker(\mathbf{A} + i\mathbf{I}_2) = \operatorname{span}((1, i))$. The projection onto $\ker(\mathbf{A} - i\mathbf{I}_2)$ is

$$b(\mathbf{A})q(\mathbf{A}) = -\tfrac{i}{2}(\mathbf{A} + i\mathbf{I}_2) = \tfrac{1}{2}(\mathbf{I}_2 - i\mathbf{A}) = \begin{pmatrix} \frac{1}{2} & \frac{i}{2} \\ -\frac{i}{2} & \frac{1}{2} \end{pmatrix}.$$

Check: it sends $(1, -i)$ to $(\frac{1}{2} + \frac{1}{2},\ -\frac{i}{2} - \frac{i}{2}) = (1, -i)$ and $(1, i)$ to $(\frac{1}{2} - \frac{1}{2},\ -\frac{i}{2} + \frac{i}{2}) = \mathbf{0}$.

(b) A subspace of $\mathbb{R}^2$ other than $\{\mathbf{0}\}$ and $\mathbb{R}^2$ has dimension 1 (Theorem 1.6.7), so it is $\operatorname{span}(\mathbf{v})$ with $\mathbf{v} = (s, t) \neq \mathbf{0}$. If $\mathbf{Av} \in \operatorname{span}(\mathbf{v})$, then $(-t, s) = c(s, t)$ for some $c \in \mathbb{R}$, so $-t = cs$ and $s = ct$, giving $s = -c^2 s$ and $t = -c^2 t$. Since $1 + c^2 > 0$, $s = t = 0$, a contradiction. So only $\{\mathbf{0}\}$ and $\mathbb{R}^2$ are mapped into themselves.

Over $\mathbb{R}$, $x^2 + 1$ has no root, so its only monic divisors are 1 and $x^2 + 1$, and the only factorizations $x^2 + 1 = pq$ into monic factors have $p = 1$ or $q = 1$. The lemma then gives only $\ker p(T) \oplus \ker q(T) = \{\mathbf{0}\} \oplus \mathbb{R}^2$, and by (b) there is no finer splitting into subspaces preserved by $T_{\mathbf{A}}$ at all. Enlarging the field to $\mathbb{C}$ creates the coprime factors $x \mp i$, and with them the splitting of (a).

Chapter 7

Determinants

A square matrix is invertible or it is not, and so far deciding which has meant row reducing it. The determinant packages that decision into a single number. The same number measures how the linear map scales areas and volumes, and its sign records whether the map preserves or reverses orientation. It also produces the characteristic polynomial, the gateway to eigenvalues in Chapter 9.

This chapter does not start from a formula. It starts from what a signed volume must do. Scaling one edge scales the volume. Sliding one edge along another does not change it. The unit cube has volume 1. We prove that exactly one function has these properties, write it down, and then derive every property of the determinant from that uniqueness: multiplicativity, the transpose rule, cofactor expansion, the inverse formula, and Cramer's rule. Nothing is assumed, and no formula is used before it is proved.

What you need. From Chapter 0, permutations and transpositions (Theorem 0.10.4) and the characteristic of a field (Definition 0.7.2). From Chapter 1, linear independence and the Linear Dependence Lemma (Definition 1.4.1, Theorem 1.4.1). From Chapter 3, elementary matrices and the invertible matrix theorem (Definition 3.4.1, Theorem 3.4.4). From Chapter 2, the matrix of a linear map and similarity (Definition 4.1.1, Theorem 4.2.4). From Chapter 5, only the word **linear functional** (Definition 5.1.1), which §02 borrows to say what a multilinear form is in each argument separately; no result of that chapter is used in a proof here. From Chapter 6, the root bound for polynomials (Corollary 6.1.4).

7.0.1 Roadmap

- **Signed area and volume.** Three rules that any reasonable notion of signed volume obeys, and the formula $ad - bc$ forced by them.
- **Multilinear alternating forms.** The rules made precise for functions of n vectors, and their first consequences.
- **Permutations and sign.** Counting inversions, and why the parity of a permutation is well defined.
- **Existence and uniqueness of the determinant.** Exactly one alternating form takes the value 1 on the identity; the Leibniz formula.
- **Row operations and multiplicativity.** Computing determinants by elimination, $\det(\mathbf{AB}) = \det \mathbf{A} \det \mathbf{B}$, and the determinant of an operator.
- **Cofactor expansion, the adjugate, and Cramer's rule.** Expanding along a row or column, and a formula for the inverse.
- **Minors, rank, and Cauchy–Binet.** Rank read off from minors, and the determinant of a product of rectangular matrices.

- **Special determinants.** Vandermonde, tridiagonal, and block triangular matrices, and rank-one updates.
- **Orientation and volume.** The determinant as a volume scaling factor, and the two orientations of $\mathbb{R}^n$.
- **The characteristic polynomial.** $\det(x\mathbf{I} - \mathbf{A})$, its coefficients, and why similar matrices share it.

7.0.2 Named moves

- **Derive from the rules, not the formula.** Prove determinant facts from multilinearity, alternation and normalization, so no property is assumed before it is proved.
- **Uniqueness does the work.** To prove two expressions agree, show both are alternating multilinear in the columns and agree on the identity.
- **Build from elementary matrices.** Check a claim for elementary matrices, then extend it to products.
- **Look-alike matrices.** A matrix with two equal rows has determinant zero; expanding such a matrix proves the adjugate identity.
- **Induct on size.** Expand along a row or column to reduce an $n \times n$ determinant to smaller ones.

7.1 Signed Area and Volume

In Chapter 0 a single number, $ad - bc$, decided whether a 2×2 matrix is invertible (Theorem 0.9.8). For an $n \times n$ matrix, Chapter 3 answered the same question by row reduction, which is a procedure, not a number. Is there one number for every square matrix that does the job of $ad - bc$? This section does not define it yet. Instead it asks what such a number should **mean**, finds the answer in area and volume, and extracts a short list of rules. From the rules alone, $ad - bc$ comes out, with no other choice possible.

7.1.1 Area as a function of two vectors

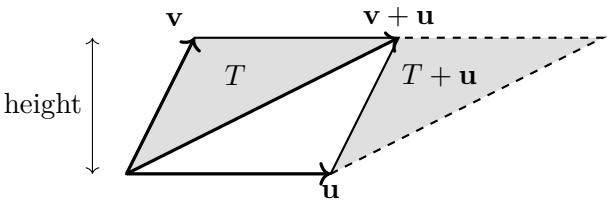
Take two vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^2$ and the parallelogram they span,

$$P(\mathbf{u}, \mathbf{v}) = \{s\mathbf{u} + t\mathbf{v} : 0 \le s \le 1,\ 0 \le t \le 1\}.$$

Its corners are $\mathbf{0}$, $\mathbf{u}$, $\mathbf{v}$, and $\mathbf{u} + \mathbf{v}$. We want to understand its area as a **function of the two vectors**, without drawing a coordinate grid and counting squares. Four features are visible without any formula.

Scaling a side scales the area. Replacing $\mathbf{v}$ by $2\mathbf{v}$ stacks two copies of $P(\mathbf{u}, \mathbf{v})$ on top of each other, so the area doubles. Replacing $\mathbf{v}$ by $3\mathbf{v}$ triples it, and replacing $\mathbf{v}$ by $\frac{1}{2}\mathbf{v}$ halves it.

Shearing does not change the area. Replace $\mathbf{v}$ by $\mathbf{v} + c\mathbf{u}$. The top edge slides along its own line, parallel to the base $\mathbf{u}$. The base stays the same, and so does the height, so the area stays the same. The picture shows $c = 1$: the triangle cut off on the left of $P(\mathbf{u}, \mathbf{v})$ reappears, moved by $\mathbf{u}$, on the right of $P(\mathbf{u}, \mathbf{v} + \mathbf{u})$.



The solid outline is $P(\mathbf{u}, \mathbf{v})$ and the dashed one completes $P(\mathbf{u}, \mathbf{v} + \mathbf{u})$. Both consist of the unshaded middle triangle plus a shaded triangle, and the two shaded triangles T and $T + \mathbf{u}$ are translates of each other.

Equal sides give area zero. $P(\mathbf{u}, \mathbf{u})$ is a segment, which has no area. More generally, if $\mathbf{v}$ is a multiple of $\mathbf{u}$, the parallelogram is flat.

The unit square has area one. $P(\mathbf{e}_1, \mathbf{e}_2)$ is the unit square. This fixes the unit of measurement.

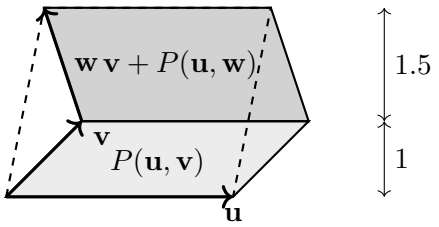
7.1.2 Why the area needs a sign

The first feature is not quite right as stated. Replacing $\mathbf{v}$ by $-\mathbf{v}$ moves the parallelogram to the other side of the line of $\mathbf{u}$ (the point reflection $\mathbf{x} \mapsto \mathbf{u} - \mathbf{x}$ carries $P(\mathbf{u}, \mathbf{v})$ onto $P(\mathbf{u}, -\mathbf{v})$) and does **not** change its area, so ordinary area satisfies $\text{area}(\mathbf{u}, c\mathbf{v}) = |c|\,\text{area}(\mathbf{u}, \mathbf{v})$, with an absolute value. That absolute value spoils a second, very useful property.

Fix the base $\mathbf{u}$ and let the second vector vary. The area is base times height, and heights along a fixed direction add up: if $\mathbf{v}$ reaches height 1 above the line of $\mathbf{u}$ and $\mathbf{w}$ reaches height 1.5, then $\mathbf{v} + \mathbf{w}$ reaches height 2.5. So we expect

$$\text{area}(\mathbf{u}, \mathbf{v} + \mathbf{w}) = \text{area}(\mathbf{u}, \mathbf{v}) + \text{area}(\mathbf{u}, \mathbf{w}).$$

The picture shows it: $P(\mathbf{u}, \mathbf{w})$, moved up by $\mathbf{v}$, sits exactly on top of $P(\mathbf{u}, \mathbf{v})$, and the two together have the same base and total height as $P(\mathbf{u}, \mathbf{v} + \mathbf{w})$ (dashed).



Now take $\mathbf{w} = -\mathbf{v}$. Then $\mathbf{v} + \mathbf{w} = \mathbf{0}$ and the left side is $\text{area}(\mathbf{u}, \mathbf{0}) = 0$, while the right side is $2\,\text{area}(\mathbf{u}, \mathbf{v})$, which is positive. Additivity fails, because $-\mathbf{v}$ points **below** the line of $\mathbf{u}$ and its height should count as negative. The repair is to give area a sign.

Signed area is ordinary area, counted positive when the second vector lies on the counterclockwise side of the first, and negative when it lies on the clockwise side.

So $P(\mathbf{e}_1, \mathbf{e}_2)$ has signed area $+1$, because $\mathbf{e}_2$ is a quarter turn counterclockwise from $\mathbf{e}_1$, and $P(\mathbf{e}_2, \mathbf{e}_1)$ has signed area -1. With the sign in place, scaling holds with c instead of $|c|$, and heights add with their signs, so additivity holds for all $\mathbf{v}, \mathbf{w}$.

Write $D(\mathbf{u}, \mathbf{v})$ for the signed area. The pictures suggest that $D\colon \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$ obeys the following rules, **for all** $\mathbf{u}, \mathbf{v}, \mathbf{w} \in \mathbb{R}^2$ and **all** $c \in \mathbb{R}$:

(R1) **(Scaling)** $D(c\mathbf{u}, \mathbf{v}) = cD(\mathbf{u}, \mathbf{v})$ and $D(\mathbf{u}, c\mathbf{v}) = cD(\mathbf{u}, \mathbf{v})$.

(R2) **(Shear)** $D(\mathbf{u}+c\mathbf{v},\mathbf{v}) = D(\mathbf{u},\mathbf{v})$ and $D(\mathbf{u},\mathbf{v}+c\mathbf{u}) = D(\mathbf{u},\mathbf{v})$.

(R3) **(Equal sides)** $D(\mathbf{u},\mathbf{u}) = 0$.

(R4) **(Normalization)** $D(\mathbf{e}_1,\mathbf{e}_2) = 1$.

(R5) **(Additivity)** $D(\mathbf{u}+\mathbf{w},\mathbf{v}) = D(\mathbf{u},\mathbf{v}) + D(\mathbf{w},\mathbf{v})$ and $D(\mathbf{u},\mathbf{v}+\mathbf{w}) = D(\mathbf{u},\mathbf{v}) + D(\mathbf{u},\mathbf{w})$.

These are rules read off from pictures, not yet a definition. The question is what they force.

7.1.3 The rules force $ad - bc$

A 2×2 matrix $\mathbf{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ has columns $\mathbf{u} = (a,c)$ and $\mathbf{v} = (b,d)$, and $P(\mathbf{u},\mathbf{v})$ is the image of the unit square under $\mathbf{x} \mapsto \mathbf{A}\mathbf{x}$. We now show that the first four rules leave no freedom at all. The method is the column version of elimination: use shears to clear entries until the parallelogram becomes a rectangle.

Example. $ad - bc$ from the Rules

Let $D\colon \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$ be **any** function satisfying (R1)–(R4). Show that $D((a,c),(b,d)) = ad - bc$ for all $a,b,c,d \in \mathbb{R}$.

Solution. Write $\mathbf{u} = (a,c)$ and $\mathbf{v} = (b,d)$. Two consequences of (R1) come first. Taking the scalar 0 in (R1) gives $D(\mathbf{u},\mathbf{0}) = D(\mathbf{u},0\mathbf{v}) = 0\cdot D(\mathbf{u},\mathbf{v}) = 0$, and likewise $D(\mathbf{0},\mathbf{v}) = 0$. Using (R1) twice and then (R4),

$$\begin{aligned} D((p,0),(0,q)) &= D(p\mathbf{e}_1, q\mathbf{e}_2) = pq\, D(\mathbf{e}_1,\mathbf{e}_2) \\ &= pq \qquad \text{for all } p,q \in \mathbb{R}. \end{aligned} \tag{$*$}$$

Case 1: $a \neq 0$. By the second shear rule with scalar $-b/a$,

$$D(\mathbf{u},\mathbf{v}) = D\big(\mathbf{u}, \mathbf{v} - \tfrac{b}{a}\mathbf{u}\big) = D\big(\mathbf{u},(0,d')\big), \qquad \text{where } d' = d - \tfrac{bc}{a}.$$

If $d' = 0$, this is $D(\mathbf{u},\mathbf{0}) = 0 = ad'$. If $d' \neq 0$, the first shear rule with scalar $-c/d'$ clears the second entry of $\mathbf{u}$:

$$D\big(\mathbf{u},(0,d')\big) = D\big(\mathbf{u} - \tfrac{c}{d'}(0,d'),(0,d')\big) = D\big((a,0),(0,d')\big) = ad'$$

by $(*)$. In both sub-cases $D(\mathbf{u},\mathbf{v}) = ad' = ad - bc$.

Case 2: $a = 0$ *and* $b \neq 0$. By the first shear rule with scalar 1, $D(\mathbf{u},\mathbf{v}) = D(\mathbf{u}+\mathbf{v},\mathbf{v})$, and $\mathbf{u}+\mathbf{v} = (b,c+d)$ has first entry $b \neq 0$. Case 1, applied to the columns $(b,c+d)$ and (b,d), gives

$$D(\mathbf{u},\mathbf{v}) = bd - b(c+d) = -bc = ad - bc,$$

since $a = 0$.

Case 3: $a = b = 0$. If $c = 0$, then $\mathbf{u} = \mathbf{0}$ and $D(\mathbf{u},\mathbf{v}) = 0$. If $c \neq 0$, the second shear rule with scalar $-d/c$ gives $D(\mathbf{u},\mathbf{v}) = D\big(\mathbf{u},\mathbf{v} - \tfrac{d}{c}\mathbf{u}\big) = D(\mathbf{u},\mathbf{0}) = 0$. In both sub-cases $D(\mathbf{u},\mathbf{v}) = 0 = ad - bc$.

The three cases cover all a,b,c,d, so $D((a,c),(b,d)) = ad - bc$, as claimed.

So **at most one** function obeys (R1)–(R4). Conversely, $(a, c), (b, d) \mapsto ad-bc$ obeys all five rules; for instance, for the second shear rule, replacing (b, d) by $(b+ta, d+tc)$ gives $a(d+tc)-(b+ta)c = ad - bc$. The other checks are just as short. Hence **exactly one** function satisfies (R1)–(R4), and it satisfies (R5) as well.

Two things in this computation are worth noticing. First, (R3) was never used. It follows from the others: $D(\mathbf{u}, \mathbf{u}) = D(\mathbf{u}, \mathbf{u}-\mathbf{u}) = D(\mathbf{u}, \mathbf{0}) = 0$, by a shear and then scaling by 0. Second, the argument used only addition, subtraction, multiplication, and division by nonzero numbers. So over **any** field F, exactly one function $F^2 \times F^2 \to F$ satisfies (R1)–(R4), namely $ad - bc$. The picture needs $\mathbb{R}$; the algebra does not.

There is also a shorter route, if we allow additivity. Expanding $\mathbf{u} = a\mathbf{e}_1 + c\mathbf{e}_2$ and $\mathbf{v} = b\mathbf{e}_1 + d\mathbf{e}_2$ with (R5) and (R1) gives four terms:

$$D(\mathbf{u}, \mathbf{v}) = ab\, D(\mathbf{e}_1, \mathbf{e}_1) + ad\, D(\mathbf{e}_1, \mathbf{e}_2) + cb\, D(\mathbf{e}_2, \mathbf{e}_1) + cd\, D(\mathbf{e}_2, \mathbf{e}_2).$$

The first and last vanish by (R3), $D(\mathbf{e}_1, \mathbf{e}_2) = 1$ by (R4), and $D(\mathbf{e}_2, \mathbf{e}_1) = -1$ (see the Quick check below). This expansion is the route that generalizes to n vectors later in the chapter, while the elimination route of Example ($ad - bc$ from the Rules) becomes the practical way to compute.

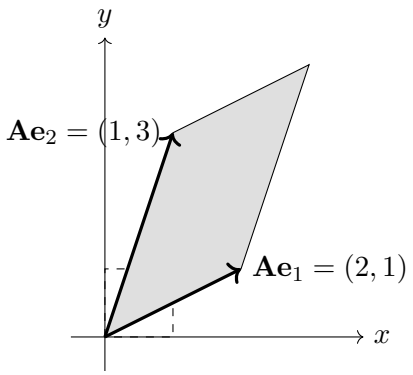
Quick check. Using only (R1) and (R2), show that $D(\mathbf{u}, \mathbf{v}) = -D(\mathbf{v}, \mathbf{u})$ for all $\mathbf{u}, \mathbf{v}$. Deduce $D(\mathbf{e}_2, \mathbf{e}_1)$.

Solution. Apply shears and one scaling:

$$\begin{aligned} D(\mathbf{u}, \mathbf{v}) &\overset{(R2)}{=} D(\mathbf{u}+\mathbf{v}, \mathbf{v}) \\ &\overset{(R2)}{=} D\big(\mathbf{u}+\mathbf{v}, \mathbf{v}-(\mathbf{u}+\mathbf{v})\big) = D(\mathbf{u}+\mathbf{v}, -\mathbf{u}) \\ &\overset{(R2)}{=} D\big((\mathbf{u}+\mathbf{v})+(-\mathbf{u}), -\mathbf{u}\big) = D(\mathbf{v}, -\mathbf{u}) \\ &\overset{(R1)}{=} -D(\mathbf{v}, \mathbf{u}). \end{aligned}$$

With $\mathbf{u} = \mathbf{e}_1$, $\mathbf{v} = \mathbf{e}_2$, and (R4), $D(\mathbf{e}_2, \mathbf{e}_1) = -D(\mathbf{e}_1, \mathbf{e}_2) = -1$. This agrees with $ad - bc = 0 \cdot 0 - 1 \cdot 1$ for the columns $(0, 1), (1, 0)$.

Now compare with Chapter 0. By Theorem 0.9.8, $\mathbf{A}$ is invertible **if and only if** $ad - bc \neq 0$. In the language of this section: $\mathbf{A}$ is invertible exactly when the parallelogram spanned by its columns is **not flat**, that is, when the columns do not lie on a common line through the origin. The number that decides invertibility and the signed area of the image of the unit square are the same number. Drag the images of $\mathbf{e}_1$ and $\mathbf{e}_2$ below and watch the signed area pass through 0 exactly when the two columns line up.



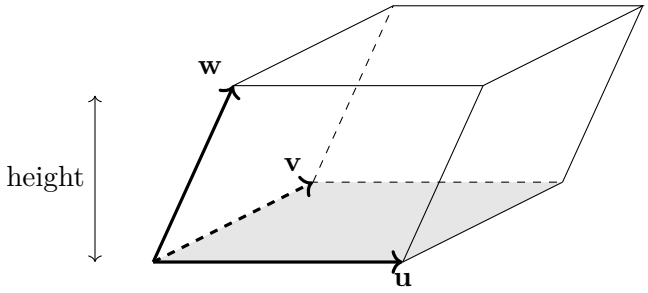
The unit square (dashed) and its image under $\mathbf{A} = \begin{pmatrix} 2 & 1 \\ 1 & 3 \end{pmatrix}$, a parallelogram of signed area $2 \cdot 3 - 1 \cdot 1 = 5$. The second column lies counterclockwise from the first, so the sign is positive.

7.1.4 Volume in three dimensions

The same story plays out one dimension up. Three vectors $\mathbf{u}, \mathbf{v}, \mathbf{w} \in \mathbb{R}^3$ span a **parallelepiped**

$$P(\mathbf{u}, \mathbf{v}, \mathbf{w}) = \{r\mathbf{u} + s\mathbf{v} + t\mathbf{w} : 0 \le r, s, t \le 1\},$$

a slanted box whose faces are parallelograms.



Its volume is the area of the base $P(\mathbf{u}, \mathbf{v})$ (shaded) times the height of $\mathbf{w}$ above the plane of the base. Each feature from the plane has a counterpart:

- scaling one edge scales the volume;
- replacing $\mathbf{w}$ by $\mathbf{w} + a\mathbf{u} + b\mathbf{v}$ slides the top face within its own plane, so the base and height, hence the volume, are unchanged, and similarly for shears of $\mathbf{u}$ or $\mathbf{v}$;
- if two edges are equal, the box is flat and the volume is 0;
- the unit cube $P(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$ has volume 1;
- with a sign attached, heights add, so the signed volume is additive in each edge.

The sign in three dimensions records whether $(\mathbf{u}, \mathbf{v}, \mathbf{w})$ is arranged like $(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$, the right-hand rule. We make this precise as orientation later in the chapter. For now the point is that the rules are **the same rules**, now for a function $D(\mathbf{u}, \mathbf{v}, \mathbf{w})$ of three vectors. The Quick check above also carries over word for word: shearing and scaling two of the arguments while the third stays fixed shows that swapping two arguments changes the sign.

7.1.5 What we will prove

Here is the goal for the next three sections. Let F be a field and $n \ge 1$.

> There is **exactly one** function D of n vectors in F^n that is linear in each argument separately, is 0 whenever two arguments are equal, and takes the value 1 at

$(\mathbf{e}_1, \ldots, \mathbf{e}_n)$.

Its value on the columns of a matrix $\mathbf{A}$ will be the determinant $\det \mathbf{A}$. The plan follows the shorter route for $n = 2$. Expanding every argument in the standard basis gives n^n terms. The rule for equal arguments kills every term with a repeated basis vector, which leaves one term for each rearrangement $(\mathbf{e}_{\sigma(1)}, \ldots, \mathbf{e}_{\sigma(n)})$ of the standard basis. Each surviving value is ± 1, reached from $D(\mathbf{e}_1, \ldots, \mathbf{e}_n) = 1$ by swaps, with one sign change per swap. That creates two tasks:

1. Name and study functions of several vectors with these rules. This is the next section, on multilinear alternating forms.
2. Make sure the sign of a rearrangement is well defined, even though it can be reached by different sequences of swaps. This is the section on permutations and sign.

The existence and uniqueness theorem itself comes after both.

Warning. **Sarrus's rule is a 3×3 coincidence.** For a 3×3 matrix, the "diagonals" shortcut (add the three products along the down-right diagonals of the array with its first two columns copied to the right, subtract the three along the up-right diagonals) gives the correct six terms, because there are exactly $3! = 6$ rearrangements. The same picture for 4×4 gives $4 + 4 = 8$ products, while the expansion above has $4! = 24$ terms. It fails already for the matrix with columns $\mathbf{e}_2, \mathbf{e}_1, \mathbf{e}_3, \mathbf{e}_4$: every one of the eight wrapped diagonals passes through a zero entry, so the shortcut returns 0, but the swap rule gives $D(\mathbf{e}_2, \mathbf{e}_1, \mathbf{e}_3, \mathbf{e}_4) = -D(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_4) = -1$. Never use Sarrus's rule beyond 3×3, and never use it as a proof tool.

7.1.6 Exercises

7.1.6.1 A. Check your understanding

Exercise. A1

(a) State the four rules (R1)–(R4) for signed area.

(b) True or false: ordinary (unsigned) area satisfies $\text{area}(\mathbf{u}, \mathbf{v}+\mathbf{w}) = \text{area}(\mathbf{u}, \mathbf{v})+\text{area}(\mathbf{u}, \mathbf{w})$ for all $\mathbf{u}, \mathbf{v}, \mathbf{w} \in \mathbb{R}^2$. Justify your answer.

(c) Find the signed area of $P((1,3),(2,4))$. Is $\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$ invertible?

(d) Explain why any D satisfying (R1) and (R2) has $D(\mathbf{u}, 5\mathbf{u}) = 0$.

(e) How many terms does the full expansion for 4×4 matrices have, and how many products does the 4×4 analogue of Sarrus's rule use?

Solution.

(a) For all $\mathbf{u}, \mathbf{v} \in \mathbb{R}^2$ and $c \in \mathbb{R}$: $D(c\mathbf{u}, \mathbf{v}) = cD(\mathbf{u}, \mathbf{v}) = D(\mathbf{u}, c\mathbf{v})$; $D(\mathbf{u}+c\mathbf{v}, \mathbf{v}) = D(\mathbf{u}, \mathbf{v}) = D(\mathbf{u}, \mathbf{v} + c\mathbf{u})$; $D(\mathbf{u}, \mathbf{u}) = 0$; $D(\mathbf{e}_1, \mathbf{e}_2) = 1$.

(b) False. Take $\mathbf{u} = \mathbf{e}_1$, $\mathbf{v} = \mathbf{e}_2$, $\mathbf{w} = -\mathbf{e}_2$. The left side is the area of $P(\mathbf{e}_1, \mathbf{0})$, which is 0; the right side is $1 + 1 = 2$.

(c) With $a = 1$, $b = 2$, $c = 3$, $d = 4$, the signed area is $ad - bc = 4 - 6 = -2$. Since $-2 \neq 0$, the matrix is invertible by Theorem 0.9.8.

(d) By the second shear rule with scalar -5, $D(\mathbf{u}, 5\mathbf{u}) = D(\mathbf{u}, 5\mathbf{u} - 5\mathbf{u}) = D(\mathbf{u}, \mathbf{0})$, and $D(\mathbf{u}, \mathbf{0}) = D(\mathbf{u}, 0 \cdot \mathbf{u}) = 0 \cdot D(\mathbf{u}, \mathbf{u}) = 0$ by (R1).

(e) The expansion has $4! = 24$ terms, one for each rearrangement of $\mathbf{e}_1, \dots, \mathbf{e}_4$. The Sarrus-type diagram has only 8 products, so it must miss 16 of them.

7.1.6.2 B. Practice

Exercise. B1: Signed Areas from the Rules

Using only (R1)–(R4), in the manner of Example ($ad - bc$ from the Rules), compute the following. Then check each answer against $ad - bc$.

(a) $D((2,1),(4,5))$.

(b) $D((1,2),(3,6))$.

(c) $D((0,3),(2,0))$.

Solution.

(a) By the second shear rule with scalar -2, $D((2,1),(4,5)) = D((2,1),(4,5) - 2(2,1)) = D((2,1),(0,3))$. By the first shear rule with scalar $-\frac{1}{3}$, this equals $D((2,1) - \frac{1}{3}(0,3),(0,3)) = D((2,0),(0,3))$. By (R1) twice and (R4), $D(2\mathbf{e}_1, 3\mathbf{e}_2) = 6D(\mathbf{e}_1, \mathbf{e}_2) = 6$. Check: $2 \cdot 5 - 4 \cdot 1 = 6$.

(b) By the second shear rule with scalar -3, $D((1,2),(3,6)) = D((1,2),(0,0))$, and $D(\mathbf{u}, \mathbf{0}) = D(\mathbf{u}, 0 \cdot \mathbf{u}) = 0$ by (R1). So the value is 0. Check: $1 \cdot 6 - 3 \cdot 2 = 0$. The columns lie on one line, and the parallelogram is flat.

(c) By (R1) twice, $D(3\mathbf{e}_2, 2\mathbf{e}_1) = 6D(\mathbf{e}_2, \mathbf{e}_1)$. By the Quick check (which uses only (R1) and (R2)), $D(\mathbf{e}_2, \mathbf{e}_1) = -D(\mathbf{e}_1, \mathbf{e}_2) = -1$ by (R4). So the value is -6. Check: with columns $(0,3)$ and $(2,0)$, $a = 0$, $b = 2$, $c = 3$, $d = 0$, and $ad - bc = -6$. The sign is negative because $(2,0)$ lies clockwise from $(0,3)$.

Exercise. B2: Equal Sides Force the Swap Rule

Let $D\colon \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$ satisfy (R1), (R3) and (R5), but not necessarily (R2).

(a) Prove that $D(\mathbf{u}, \mathbf{v}) = -D(\mathbf{v}, \mathbf{u})$ for all $\mathbf{u}, \mathbf{v} \in \mathbb{R}^2$.

(b) Suppose moreover that (R4) holds. Deduce that $D((a,c),(b,d)) = ad - bc$.

Hint: expand $D(\mathbf{u} + \mathbf{v}, \mathbf{u} + \mathbf{v})$.

Solution.

(a) Let $\mathbf{u}, \mathbf{v} \in \mathbb{R}^2$. By (R3), $D(\mathbf{u} + \mathbf{v}, \mathbf{u} + \mathbf{v}) = 0$. Expanding with (R5) in the first argument and then in the second,

$$0 = D(\mathbf{u}, \mathbf{u}) + D(\mathbf{u}, \mathbf{v}) + D(\mathbf{v}, \mathbf{u}) + D(\mathbf{v}, \mathbf{v}) = D(\mathbf{u}, \mathbf{v}) + D(\mathbf{v}, \mathbf{u}),$$

where the second equality uses (R3) twice. Hence $D(\mathbf{u}, \mathbf{v}) = -D(\mathbf{v}, \mathbf{u})$.

(b) Write $(a,c) = a\mathbf{e}_1 + c\mathbf{e}_2$ and $(b,d) = b\mathbf{e}_1 + d\mathbf{e}_2$. By (R5) and (R1) in each argument,

$$D((a,c),(b,d)) = ab\,D(\mathbf{e}_1,\mathbf{e}_1) + ad\,D(\mathbf{e}_1,\mathbf{e}_2) + cb\,D(\mathbf{e}_2,\mathbf{e}_1) + cd\,D(\mathbf{e}_2,\mathbf{e}_2).$$

By (R3) the first and last terms are 0. By (R4), $D(\mathbf{e}_1,\mathbf{e}_2) = 1$, and by (a), $D(\mathbf{e}_2,\mathbf{e}_1) = -1$. Hence $D((a,c),(b,d)) = ad - bc$.

7.1.6.3 C. Going deeper

Exercise. C1: Area of a Triangle

Let $\mathbf{p},\mathbf{q},\mathbf{r} \in \mathbb{R}^2$.

(a) Explain why the triangle with vertices $\mathbf{p},\mathbf{q},\mathbf{r}$ has area $\frac{1}{2}|D(\mathbf{q}-\mathbf{p},\mathbf{r}-\mathbf{p})|$, where $D(\mathbf{u},\mathbf{v})$ is the signed area.

(b) Find the area of the triangle with vertices $(1,2)$, $(4,3)$, $(2,6)$.

(c) Hence decide whether the points $(0,1)$, $(2,4)$, $(6,10)$ lie on a line. Justify your answer.

Solution.

(a) Translating by $-\mathbf{p}$ does not change areas, so the triangle has the same area as the triangle with vertices $\mathbf{0}$, $\mathbf{u} = \mathbf{q}-\mathbf{p}$, $\mathbf{v} = \mathbf{r}-\mathbf{p}$. That triangle is half of the parallelogram $P(\mathbf{u},\mathbf{v})$: the diagonal from $\mathbf{u}$ to $\mathbf{v}$ cuts $P(\mathbf{u},\mathbf{v})$ into the triangle with vertices $\mathbf{0},\mathbf{u},\mathbf{v}$ and the triangle with vertices $\mathbf{u}+\mathbf{v},\mathbf{v},\mathbf{u}$, and the point reflection $\mathbf{x} \mapsto \mathbf{u}+\mathbf{v}-\mathbf{x}$ maps the first onto the second, so they have equal area. The area of $P(\mathbf{u},\mathbf{v})$ is $|D(\mathbf{u},\mathbf{v})|$, since the signed area differs from the area only by its sign. Hence the triangle has area $\frac{1}{2}|D(\mathbf{q}-\mathbf{p},\mathbf{r}-\mathbf{p})|$.

(b) Here $\mathbf{q}-\mathbf{p} = (3,1)$ and $\mathbf{r}-\mathbf{p} = (1,4)$, so $D = 3\cdot 4 - 1\cdot 1 = 11$ and the area is $\frac{11}{2}$.

(c) With $\mathbf{p} = (0,1)$, $\mathbf{q}-\mathbf{p} = (2,3)$ and $\mathbf{r}-\mathbf{p} = (6,9)$, so $D = 2\cdot 9 - 6\cdot 3 = 0$ and the "triangle" has area 0. Indeed $(6,9) = 3(2,3)$, so $\mathbf{r}-\mathbf{p}$ is a multiple of $\mathbf{q}-\mathbf{p}$, and all three points lie on the line through $(0,1)$ with direction $(2,3)$. Yes, they are collinear.

Exercise. C2: The 1×1 Case

For $n = 1$, the rules ask for a function $f\colon F \to F$ of one vector in $F^1 = F$.

(a) Explain why the rule "zero whenever two arguments are equal" imposes no condition when $n = 1$.

(b) Prove that the only function $f\colon F \to F$ that is linear with $f(1) = 1$ is $f(a) = a$. Hence the determinant of a 1×1 matrix (a) must be a.

Solution.

(a) The rule speaks about two **different** positions $i \neq j$ holding equal vectors. With one argument there is no pair of positions, so the condition is vacuously true (Example (A vacuously true implication)).

(b) Let f be linear with $f(1) = 1$, and let $a \in F$. Since $a = a\cdot 1$, homogeneity gives

$f(a) = f(a \cdot 1) = af(1) = a$. Conversely, $a \mapsto a$ is linear and sends 1 to 1. Hence $f(a) = a$ is the only such function.

7.2 Multilinear Alternating Forms

The previous section found that signed area obeys a few rules: it is linear in each side separately, it vanishes when two sides are equal, and swapping the sides changes its sign. This section gives these rules names for functions of k vectors in an arbitrary vector space, and derives everything that follows from them before any formula exists. The main result is that "zero on equal arguments" already implies the swap rule, the shear rule, and vanishing on dependent lists. The swap rule, in turn, gives back "zero on equal arguments" only when $1 + 1 \neq 0$ in the field.

Throughout, V is a vector space over a field F, $k \geq 1$ is an integer, and $V^k = V \times \cdots \times V$ is the set of k-tuples of vectors (Definition 0.3.11).

7.2.1 Multilinear forms

Signed area is not a linear function of the pair $(\mathbf{u}, \mathbf{v})$: doubling both sides multiplies the area by four, not by two. What it does have is linearity in $\mathbf{u}$ when $\mathbf{v}$ is held fixed, and linearity in $\mathbf{v}$ when $\mathbf{u}$ is held fixed. This "one slot at a time" linearity recurs throughout the book: in the dot product, later in inner products and bilinear forms, and in the determinant. So we give it a name.

A multilinear form is a function of several vectors that is linear in each vector while the others are held fixed.

Definition 7.2.1. Multilinear Form

A function $f\colon V^k \to F$ is **k-linear**, or a **multilinear form** on V, if **for every** position $i \in \{1, \dots, k\}$ and **every** choice of $v_1, \dots, v_{i-1}, v_{i+1}, \dots, v_k \in V$, the function

$$V \to F, \qquad x \mapsto f(v_1, \dots, v_{i-1}, x, v_{i+1}, \dots, v_k)$$

is linear. A 2-linear form is called **bilinear**.

In words: fix all arguments except the one in position i. What remains is a function of one vector, and it must be a linear functional (Definition 5.1.1). Written out, for all $x, y \in V$ and $c \in F$,

$$f(\dots, x + y, \dots) = f(\dots, x, \dots) + f(\dots, y, \dots),$$
$$f(\dots, cx, \dots) = c\, f(\dots, x, \dots),$$

where the dots stand for the same fixed vectors on both sides. The condition is imposed for **every** position i, one at a time.

Some examples:

- **Signed area.** On F^2, $f(\mathbf{u}, \mathbf{v}) = u_1v_2 - u_2v_1$ is bilinear. For fixed $\mathbf{v}$, the map $\mathbf{u} \mapsto v_2u_1 - v_1u_2$ is a linear functional on F^2, and for fixed $\mathbf{u}$, so is $\mathbf{v} \mapsto u_1v_2 - u_2v_1$.
- **The dot product.** On F^n, $(\mathbf{u}, \mathbf{v}) \mapsto u_1v_1 + \cdots + u_nv_n$ is bilinear, by the same one-slot check.
- **Products of functionals.** For $\varphi_1, \dots, \varphi_k \in V^*$, $f(v_1, \dots, v_k) = \varphi_1(v_1)\varphi_2(v_2)\cdots\varphi_k(v_k)$ is k-linear: in position i it is the linear functional φ_i multiplied by the fixed scalar $\prod_{j \neq i} \varphi_j(v_j)$.

For instance, on $F[x]$, $(p, q) \mapsto p(0)q(1)$ is bilinear, since evaluation at a point is linear (Theorem 0.8.5).

- **The degenerate case $k = 1$.** A 1-linear form is exactly a linear functional on V. So multilinear forms generalize the dual space of Chapter 5.

Warning. **Multilinear is not linear.** A bilinear form is **not** a linear function of the pair. For the signed area $f(\mathbf{u}, \mathbf{v}) = u_1v_2 - u_2v_1$ on $\mathbb{R}^2$, scaling both arguments by 2 gives $f(2\mathbf{e}_1, 2\mathbf{e}_2) = 4$, not $2f(\mathbf{e}_1, \mathbf{e}_2) = 2$. Adding in both slots at once is not allowed either: $f(\mathbf{e}_1 + \mathbf{0}, \mathbf{e}_1 + \mathbf{e}_2) = f(\mathbf{e}_1, \mathbf{e}_1 + \mathbf{e}_2) = 1$, while $f(\mathbf{e}_1, \mathbf{e}_1) + f(\mathbf{0}, \mathbf{e}_2) = 0$. Pull out scalars and split sums **one slot at a time**.

7.2.2 Alternating and skew-symmetric forms

The second rule of signed area was geometric: if two edges coincide, the parallelogram or box is flat. There are two natural ways to put this into algebra. One says the value is **zero** when two arguments are equal. The other says the value **changes sign** when two arguments are swapped. They look like two phrasings of one idea, but they are not the same, and we must choose.

An alternating form vanishes as soon as two of its arguments coincide.

Definition 7.2.2. Alternating Form

A k-linear form $f\colon V^k \to F$ is **alternating** if $f(v_1, \ldots, v_k) = 0$ **whenever** $v_i = v_j$ for **some** pair of positions $i \neq j$.

Definition 7.2.3. Skew-Symmetric Form

A k-linear form $f\colon V^k \to F$ is **skew-symmetric** if, **for all** $v_1, \ldots, v_k \in V$ and **all** positions $i < j$, swapping the arguments in positions i and j multiplies the value by -1:

$$f(v_1, \ldots, v_j, \ldots, v_i, \ldots, v_k) = -f(v_1, \ldots, v_i, \ldots, v_j, \ldots, v_k),$$

where on the left v_j sits in position i, v_i sits in position j, and all other positions are unchanged.

In words: "alternating" asks for the value 0 as soon as **any two** positions, adjacent or not, hold the same vector, whatever the remaining arguments are. It is a condition on the inputs where two entries coincide, and says nothing directly about other inputs. "Skew-symmetric" is a condition on **every** input: it compares the value at a list with the value at the list with two entries exchanged.

Example. Alternating Forms

Check that each of the following is an alternating form.

(a) On F^2, for a fixed $c \in F$: $f_c(\mathbf{u}, \mathbf{v}) = c(u_1v_2 - u_2v_1)$.

(b) On F^3, for fixed coordinate indices $1 \leq r < s \leq 3$: $g_{rs}(\mathbf{u}, \mathbf{v}) = u_rv_s - u_sv_r$.

(c) On $F[x]$: $h(p,q) = p(0)q(1) - p(1)q(0)$.

(d) On any V, with $k = 1$: any linear functional $\varphi \in V^*$.

Solution.

(a) Bilinearity is the one-slot check above, multiplied by the constant c. With $k = 2$ the only pair of positions is $(1,2)$, and $f_c(\mathbf{u},\mathbf{u}) = c(u_1u_2 - u_2u_1) = 0$ by commutativity in F. So f_c is alternating. For $c = 0$ it is the zero form, which is alternating for the silly reason that it is zero everywhere; for $c = 1$ it is the signed area of the previous section.

(b) In each slot g_{rs} is a linear functional, as in (a), and $g_{rs}(\mathbf{u},\mathbf{u}) = u_ru_s - u_su_r = 0$. Over $\mathbb{R}$ these three forms are, up to sign and order, the components of the cross product: $\mathbf{u} \times \mathbf{v} = (g_{23}(\mathbf{u},\mathbf{v}), -g_{13}(\mathbf{u},\mathbf{v}), g_{12}(\mathbf{u},\mathbf{v}))$. This example shows that the number of arguments need not equal the dimension: here $k = 2$ and $\dim F^3 = 3$.

(c) Evaluation at 0 and at 1 are linear in p (Theorem 0.8.5), so for fixed q the map $p \mapsto q(1)\,p(0) - q(0)\,p(1)$ is linear, and similarly in q. Also $h(p,p) = p(0)p(1) - p(1)p(0) = 0$. It is not the zero form: $h(1,x) = 1 \cdot 1 - 1 \cdot 0 = 1$.

(d) A linear functional is 1-linear. With a single position there is **no** pair $i \neq j$, so the alternating condition holds vacuously (Example (A vacuously true implication)). This degenerate case matters because it shows that "alternating" is purely a condition about **pairs** of positions; the interesting content starts at $k = 2$.

Now a nonexample, by a minimal change of (a). Over $\mathbb{R}$, replace the minus sign by a plus: $s(\mathbf{u},\mathbf{v}) = u_1v_2 + u_2v_1$. Bilinearity survives, by the same one-slot check. But $s(\mathbf{e}_1+\mathbf{e}_2, \mathbf{e}_1+\mathbf{e}_2) = 1+1 = 2 \neq 0$, so the alternating clause fails. The dot product fails in the same way, $\mathbf{e}_1 \cdot \mathbf{e}_1 = 1$. A different minimal change keeps the alternating clause and breaks the other one. The unsigned area $|u_1v_2 - u_2v_1|$ on $\mathbb{R}^2$ is still 0 whenever $\mathbf{u} = \mathbf{v}$, but it is not linear in $\mathbf{u}$: at $\mathbf{v} = \mathbf{e}_2$ it sends $\mathbf{e}_1 \mapsto 1$ and $-\mathbf{e}_1 \mapsto 1 \neq -1$. This is the failure that forced a sign onto area in the previous section.

Quick check. Is $f(\mathbf{u},\mathbf{v}) = u_1v_2$ on F^2 alternating? Is it skew-symmetric?

Solution. It is bilinear, but neither. With $\mathbf{u} = \mathbf{v} = \mathbf{e}_1 + \mathbf{e}_2$, $f(\mathbf{u},\mathbf{u}) = 1 \cdot 1 = 1 \neq 0$, so it is not alternating. And $f(\mathbf{e}_1,\mathbf{e}_2) = 1$ while $f(\mathbf{e}_2,\mathbf{e}_1) = 0$; skew-symmetry would need $0 = -1$, which fails in every field.

Why this definition. We chose "zero when two arguments are equal" as the basic notion, and in the form "**any** two positions". There are two reasons. First, it is what the picture says directly: two equal edges make a flat box. Second, as we prove next, it implies the swap rule over **every** field, while the swap rule implies it only when $1 + 1 \neq 0$. Allowing any pair of positions, not just adjacent ones, also spares us a separate argument moving equal entries next to each other.

7.2.3 Consequences of the alternating condition

Everything that elimination does to a determinant will come from the following theorem. Its four parts are the column operations of Chapter 3 seen through a form: swap two arguments, add a multiple of one to another, scale one, and the extreme case of a dependent list.

Theorem 7.2.1. Properties of Alternating Forms

Let $f\colon V^k \to F$ be an alternating k-linear form, let $v_1, \dots, v_k \in V$, and let $i \neq j$ be positions in $\{1, \dots, k\}$.

(a) **(Swap)** Exchanging the arguments in positions i and j multiplies $f(v_1, \dots, v_k)$ by -1. In particular, f is skew-symmetric.

(b) **(Shear)** For every $c \in F$, replacing v_i by $v_i + cv_j$ does not change $f(v_1, \dots, v_k)$.

(c) **(Scaling)** For every $c \in F$, replacing v_i by cv_i multiplies $f(v_1, \dots, v_k)$ by c.

(d) **(Dependence)** If the list $(v_1, \dots, v_k)$ is linearly dependent, then $f(v_1, \dots, v_k) = 0$. In particular, $f(v_1, \dots, v_k) = 0$ if some $v_i = 0$.

Idea. Parts (a)–(c) involve at most the two slots i and j, so freeze the other $k-2$ arguments. What is left is a bilinear form $g(x, y)$ with $g(x, x) = 0$, and the k-argument statements become two-argument statements about g. For the swap, the move is to feed g a sum in both slots, $g(x + y, x + y) = 0$, and expand: the two "square" terms die and the two cross terms must cancel. For (d), write one vector as a combination of the others and expand in its slot: every resulting term has a repeated argument.

Proof. Suppose first $i < j$; the case $i > j$ is the same with the roles of i and j exchanged. For $x, y \in V$, let

$$g(x, y) = f(v_1, \dots, v_{i-1}, x, v_{i+1}, \dots, v_{j-1}, y, v_{j+1}, \dots, v_k),$$

the value of f with x in position i, y in position j, and v_m in every other position m. Since f is k-linear, g is bilinear, and since f is alternating, $g(x, x) = 0$ for every $x \in V$. In this notation $f(v_1, \dots, v_k) = g(v_i, v_j)$.

(c) This is linearity of f in position i: $g(cv_i, v_j) = c\, g(v_i, v_j)$.

(d) By linearity in the first slot of g and $g(v_j, v_j) = 0$,

$$g(v_i + cv_j, v_j) = g(v_i, v_j) + c\, g(v_j, v_j) = g(v_i, v_j).$$

(e) By bilinearity of g,

$$\begin{aligned} 0 &= g(v_i + v_j, v_i + v_j) \\ &= g(v_i, v_i) + g(v_i, v_j) + g(v_j, v_i) + g(v_j, v_j) \\ &= g(v_i, v_j) + g(v_j, v_i), \end{aligned}$$

where the first and last equalities use $g(x, x) = 0$. Hence $g(v_j, v_i) = -g(v_i, v_j)$, which is the swap statement. Since $v_1, \dots, v_k$ and $i < j$ were arbitrary, f is skew-symmetric (Definition 7.2.3).

(f) Suppose $(v_1, \dots, v_k)$ is linearly dependent. By the Linear Dependence Lemma (Theorem 1.4.1), there is an index j with $v_j \in \operatorname{span}(v_1, \dots, v_{j-1})$. If $j = 1$, then $v_1 = 0 = 0 \cdot v_1$, and by (c) with $c = 0$, $f(v_1, \dots, v_k) = 0 \cdot f(v_1, \dots, v_k) = 0$. If $j \geq 2$, write $v_j = a_1 v_1 + \cdots + a_{j-1} v_{j-1}$ with $a_m \in F$. By linearity of f in position j,

$$f(v_1, \dots, v_k) = \sum_{m=1}^{j-1} a_m \, f(v_1, \dots, v_{j-1}, v_m, v_{j+1}, \dots, v_k).$$

In the m-th term, the vector v_m sits both in position m and in position $j \neq m$, so the term is 0 because f is alternating. Hence $f(v_1, \dots, v_k) = 0$. The last sentence of (d) holds because a list containing 0 is dependent. This proves the theorem.

■

For the determinant, (a)–(c) will be the rules for column operations, and (d) explains at once why a matrix with dependent columns must have determinant 0. Part (d) also bounds the number of arguments: on a space of dimension n, any $k > n$ vectors are dependent, so a nonzero alternating form needs $k \leq n$ (Exercise C1).

7.2.4 Alternating versus skew-symmetric

Part (a) says alternating forms are skew-symmetric. The converse depends on the field.

Theorem 7.2.2. Skew-Symmetric and Alternating Forms

Let $f\colon V^k \to F$ be a k-linear form.

(a) If f is alternating, then f is skew-symmetric.

(b) If f is skew-symmetric and the characteristic of F is **not** 2, then f is alternating.

Proof.

(a) is Theorem 7.2.1 (a).

(b) Suppose $v_i = v_j$ for some $i < j$. Exchanging the arguments in positions i and j does not change the list, so skew-symmetry gives $f(v_1, \dots, v_k) = -f(v_1, \dots, v_k)$, that is, $(1 + 1)\, f(v_1, \dots, v_k) = 0$. If $1 + 1 = 0$ in F, then $2 \cdot 1 = 0$ while $1 \cdot 1 = 1 \neq 0$, so the smallest positive n with $n \cdot 1 = 0$ would be 2, and the characteristic of F would be 2 (Definition 0.7.2). So $1 + 1 \neq 0$, and multiplying by $(1 + 1)^{-1}$ gives $f(v_1, \dots, v_k) = 0$. Hence f is alternating.

■

Warning. **Skew-symmetric does not imply alternating over $\mathbb{F}_2$.** In $\mathbb{F}_2$ we have $-1 = 1$, so "skew-symmetric" just means "symmetric". Take $f(\mathbf{u}, \mathbf{v}) = u_1 v_1$ on $\mathbb{F}_2^2$. It is bilinear, and $f(\mathbf{v}, \mathbf{u}) = v_1 u_1 = u_1 v_1 = -f(\mathbf{u}, \mathbf{v})$, so it is skew-symmetric. But $f(\mathbf{e}_1, \mathbf{e}_1) = 1 \neq 0$, so it is **not** alternating. Checking all 16 pairs $(\mathbf{u}, \mathbf{v})$ confirms both claims: $f(\mathbf{u}, \mathbf{v}) = f(\mathbf{v}, \mathbf{u})$ always, and $f(\mathbf{u}, \mathbf{u}) = u_1$ is 1 for $\mathbf{u} = (1, 0)$ and $\mathbf{u} = (1, 1)$. This is why the determinant is built on alternating forms, not skew-symmetric ones: the construction must work over $\mathbb{F}_2$ too.

7.2.5 The space of alternating forms

Alternating forms can be added and scaled, and the result is again alternating. This is what makes the uniqueness question later in the chapter a question about dimension.

Theorem 7.2.3. Alternating Forms Form a Vector Space

The k-linear forms on V form a subspace of the vector space F^{V^k} of all functions $V^k \to F$, with the pointwise operations of Example (The Standard Vector Spaces) (d). The alternating k-linear forms on V form a subspace of it. In particular, both are vector spaces over F.

Proof. We use the Subspace Test (Theorem 1.2.1) twice.

The zero function is k-linear, since in each position it is the zero functional, and it is alternating, since all its values are 0.

Let f and g be k-linear and $a \in F$. Fix a position i and vectors in the other positions, and write $f(\dots, x, \dots)$ for the value with x in position i. For $x, y \in V$ and $c \in F$, the pointwise definition and linearity of f and g in position i give

$$\begin{aligned}(f+g)(\dots, x+y, \dots) &= f(\dots, x, \dots) + f(\dots, y, \dots) \\ &\quad + g(\dots, x, \dots) + g(\dots, y, \dots) \\ &= (f+g)(\dots, x, \dots) + (f+g)(\dots, y, \dots)\end{aligned}$$

and

$$(f+g)(\dots, cx, \dots) = c\,f(\dots, x, \dots) + c\,g(\dots, x, \dots) = c\,(f+g)(\dots, x, \dots).$$

So $f+g$ is k-linear. In the same way $(af)(\dots, x+y, \dots) = af(\dots, x, \dots) + af(\dots, y, \dots)$ and $(af)(\dots, cx, \dots) = c\,(af)(\dots, x, \dots)$, using commutativity in F, so af is k-linear.

If moreover f and g are alternating and $v_i = v_j$ for some $i \neq j$, then $(f+g)(v_1, \dots, v_k) = 0 + 0 = 0$ and $(af)(v_1, \dots, v_k) = a \cdot 0 = 0$. So the alternating forms are closed under both operations. This proves the theorem. ∎

For $V = F^2$ and $k = 2$, Exercise B2 shows that every alternating bilinear form is one of the forms f_c of Example (Alternating Forms) (a), with $c = f(\mathbf{e}_1, \mathbf{e}_2)$. So that space is spanned by the single nonzero form f_1, and has dimension 1. Later in this chapter we prove the same for all n: the alternating n-linear forms on F^n form a space of dimension exactly 1, spanned by the determinant. The missing ingredient is how the value of an alternating form changes when its arguments are rearranged by more than one swap, which is the subject of the next section.

7.2.6 Exercises

7.2.6.1 A. Check your understanding

Exercise. A1

(a) Define an alternating k-linear form on a vector space V over F.

(b) True or false: a bilinear form f on $\mathbb{R}^2$ with $f(\mathbf{e}_1, \mathbf{e}_1) = f(\mathbf{e}_2, \mathbf{e}_2) = 0$ is alternating. Justify your answer.

(c) True or false: every skew-symmetric bilinear form on $\mathbb{R}^5$ is alternating. Justify your answer.

(d) Let f be an alternating 3-linear form on $\mathbb{R}^3$ and $\mathbf{u}, \mathbf{v} \in \mathbb{R}^3$. What is $f(\mathbf{u}, \mathbf{v}, 2\mathbf{u} - \mathbf{v})$?

(e) Name the move that proves an alternating form changes sign under a swap.

Solution.

(a) A function $f\colon V^k \to F$ that is linear in each argument when the others are fixed (Definition 7.2.1), and satisfies $f(v_1, \dots, v_k) = 0$ whenever $v_i = v_j$ for some $i \neq j$ (Definition 7.2.2).

(b) False. The form $s(\mathbf{u}, \mathbf{v}) = u_1v_2 + u_2v_1$ is bilinear with $s(\mathbf{e}_1, \mathbf{e}_1) = s(\mathbf{e}_2, \mathbf{e}_2) = 0$, but $s(\mathbf{e}_1 + \mathbf{e}_2, \mathbf{e}_1 + \mathbf{e}_2) = 2 \neq 0$. Checking the basis vectors is not enough, because the alternating condition is about **all** vectors.

(c) True. $\mathbb{R}$ has characteristic 0, so Theorem 7.2.2 (b) applies.

(d) 0. The list $(\mathbf{u}, \mathbf{v}, 2\mathbf{u} - \mathbf{v})$ is linearly dependent, since $2\mathbf{u} - \mathbf{v} - (2\mathbf{u} - \mathbf{v}) = \mathbf{0}$ with coefficients $2, -1, -1$ not all zero. By Theorem 7.2.1 (d) the value is 0.

(e) Put the sum of the two arguments into both positions, use $f(\dots, x+y, \dots, x+y, \dots) = 0$, and expand by linearity in both positions; the terms with equal arguments vanish and the two cross terms must cancel.

7.2.6.2 B. Practice

Exercise. B1: Multilinear or Alternating?

Determine which of the following functions are multilinear, and which of those are alternating. Justify your answer.

(a) $\mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$, $(\mathbf{u}, \mathbf{v}) \mapsto 3u_1v_2 - 3u_2v_1$.

(b) $\mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$, $(\mathbf{u}, \mathbf{v}) \mapsto u_1v_2 - u_2v_1 + 1$.

(c) $\mathbb{R}^3 \times \mathbb{R}^3 \times \mathbb{R}^3 \to \mathbb{R}$, $(\mathbf{u}, \mathbf{v}, \mathbf{w}) \mapsto u_1v_2w_3$.

(d) $M_2(\mathbb{R}) \times M_2(\mathbb{R}) \to \mathbb{R}$, $(\mathbf{A}, \mathbf{B}) \mapsto \operatorname{tr}(\mathbf{AB}) - \operatorname{tr}(\mathbf{BA})$.

(e) $\mathbb{R}[x] \times \mathbb{R}[x] \to \mathbb{R}$, $(p, q) \mapsto p(1)q'(1) - p'(1)q(1)$.

Solution.

(a) Multilinear and alternating: it is f_3 from Example (Alternating Forms) (a) with $c = 3$.

(b) Not multilinear. A linear functional sends $\mathbf{0}$ to 0, but at $\mathbf{u} = \mathbf{0}$ the map $\mathbf{v} \mapsto 0-0+1 = 1$ does not. So it is not an alternating form either; in any case its value at $(\mathbf{u}, \mathbf{u})$ is $1 \neq 0$.

(c) Multilinear: in each position it is a coordinate functional times a fixed scalar, as for products of functionals. Not alternating: with $\mathbf{u} = \mathbf{v} = \mathbf{e}_1 + \mathbf{e}_2$ and $\mathbf{w} = \mathbf{e}_3$, the value is $1 \cdot 1 \cdot 1 = 1 \neq 0$.

(d) By Theorem 0.9.6, $\operatorname{tr}(\mathbf{AB}) = \operatorname{tr}(\mathbf{BA})$, so this is the zero function. The zero function is multilinear and alternating (Theorem 7.2.3). This is the degenerate case.

(e) Multilinear: $p \mapsto p(1)$ and $p \mapsto p'(1)$ are linear, the second as a composition of differen-

tiation (Example (Differentiation)) with evaluation (Theorem 0.8.5). So for fixed q, $p \mapsto q'(1)\,p(1) - q(1)\,p'(1)$ is linear, and similarly in q. Alternating: $p(1)p'(1) - p'(1)p(1) = 0$.

Exercise. B2: Alternating Bilinear Forms on F^2

Let f be an alternating bilinear form on F^2. Prove that

$$f(\mathbf{u}, \mathbf{v}) = f(\mathbf{e}_1, \mathbf{e}_2)\,(u_1 v_2 - u_2 v_1) \qquad \text{for all } \mathbf{u}, \mathbf{v} \in F^2.$$

Hence show that the alternating bilinear forms on F^2 form a space of dimension 1.

Solution. Write $\mathbf{u} = u_1\mathbf{e}_1 + u_2\mathbf{e}_2$ and $\mathbf{v} = v_1\mathbf{e}_1 + v_2\mathbf{e}_2$. By linearity in the first and then in the second argument,

$$f(\mathbf{u}, \mathbf{v}) = u_1 v_1 f(\mathbf{e}_1, \mathbf{e}_1) + u_1 v_2 f(\mathbf{e}_1, \mathbf{e}_2) + u_2 v_1 f(\mathbf{e}_2, \mathbf{e}_1) + u_2 v_2 f(\mathbf{e}_2, \mathbf{e}_2).$$

Since f is alternating, $f(\mathbf{e}_1, \mathbf{e}_1) = f(\mathbf{e}_2, \mathbf{e}_2) = 0$, and by Theorem 7.2.1 (a), $f(\mathbf{e}_2, \mathbf{e}_1) = -f(\mathbf{e}_1, \mathbf{e}_2)$. Hence $f(\mathbf{u}, \mathbf{v}) = f(\mathbf{e}_1, \mathbf{e}_2)(u_1 v_2 - u_2 v_1)$.

By Theorem 7.2.3 these forms form a vector space, and we have just shown $f = f(\mathbf{e}_1, \mathbf{e}_2)\, f_1$, where $f_1(\mathbf{u}, \mathbf{v}) = u_1 v_2 - u_2 v_1$ is alternating by Example (Alternating Forms) (a). So (f_1) spans the space. It is nonzero, since $f_1(\mathbf{e}_1, \mathbf{e}_2) = 1$, so the list (f_1) is linearly independent and hence a basis. The dimension is 1.

Exercise. B3: The Triple Product

For $\mathbf{v}, \mathbf{w} \in \mathbb{R}^3$ let $\mathbf{v} \times \mathbf{w} = (v_2 w_3 - v_3 w_2,\ v_3 w_1 - v_1 w_3,\ v_1 w_2 - v_2 w_1)$, and define $f(\mathbf{u}, \mathbf{v}, \mathbf{w}) = u_1(\mathbf{v} \times \mathbf{w})_1 + u_2(\mathbf{v} \times \mathbf{w})_2 + u_3(\mathbf{v} \times \mathbf{w})_3$. Show that f is an alternating 3-linear form on $\mathbb{R}^3$ with $f(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3) = 1$.

Solution. Multilinear. Each component of $\mathbf{v} \times \mathbf{w}$ is $g_{rs}(\mathbf{v}, \mathbf{w})$ or $-g_{rs}(\mathbf{v}, \mathbf{w})$ for a form of Example (Alternating Forms) (b), hence bilinear in $(\mathbf{v}, \mathbf{w})$. For fixed $\mathbf{v}, \mathbf{w}$, f is a linear functional of $\mathbf{u}$ with coefficients $(\mathbf{v} \times \mathbf{w})_m$. For fixed $\mathbf{u}$ and $\mathbf{w}$, $f = \sum_m u_m (\mathbf{v} \times \mathbf{w})_m$ is a combination of linear functionals of $\mathbf{v}$, hence linear in $\mathbf{v}$; likewise in $\mathbf{w}$.

Alternating. There are three pairs of positions. If $\mathbf{v} = \mathbf{w}$, every component $v_r v_s - v_s v_r$ is 0, so $\mathbf{v} \times \mathbf{v} = \mathbf{0}$ and $f = 0$. If $\mathbf{u} = \mathbf{v}$,

$$f(\mathbf{u}, \mathbf{u}, \mathbf{w}) = u_1(u_2 w_3 - u_3 w_2) + u_2(u_3 w_1 - u_1 w_3) + u_3(u_1 w_2 - u_2 w_1) = 0,$$

since the six terms cancel in pairs: $u_1 u_2 w_3$ with $-u_2 u_1 w_3$, $-u_1 u_3 w_2$ with $u_3 u_1 w_2$, and $u_2 u_3 w_1$ with $-u_3 u_2 w_1$. If $\mathbf{u} = \mathbf{w}$, the same expansion $u_1(v_2 u_3 - v_3 u_2) + u_2(v_3 u_1 - v_1 u_3) + u_3(v_1 u_2 - v_2 u_1)$ cancels in pairs in the same way.

Value. $\mathbf{e}_2 \times \mathbf{e}_3 = (1 \cdot 1 - 0 \cdot 0,\ 0 \cdot 0 - 0 \cdot 1,\ 0 \cdot 0 - 1 \cdot 0) = (1, 0, 0)$, so $f(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3) = 1$.

7.2.6.3 C. Going deeper

Exercise. C1: Too Many Arguments

Let V be finite-dimensional with $\dim V = n$.

(a) Prove that every alternating k-linear form on V with $k > n$ is the zero form.

(b) Deduce that the only alternating 3-linear form on F^2 is zero, and give a nonzero alternating 2-linear form on F^2 to show that the bound in (a) cannot be lowered to $k \geq n$.

Solution.

(a) Let f be an alternating k-linear form with $k > n$, and let $v_1, \dots, v_k \in V$. By Theorem 1.6.3 (a), a list of length $k > \dim V$ is linearly dependent. By Theorem 7.2.1 (d), $f(v_1, \dots, v_k) = 0$. Since the vectors were arbitrary, f is the zero form.

(b) Here $\dim F^2 = 2 < 3$, so (a) applies. For $k = n = 2$, the form $f_1(\mathbf{u}, \mathbf{v}) = u_1 v_2 - u_2 v_1$ of Example (Alternating Forms) (a) is alternating with $f_1(\mathbf{e}_1, \mathbf{e}_2) = 1 \neq 0$.

Exercise. C2: Bilinear Forms over $\mathbb{F}_2$

(a) Let f be a bilinear form on F^2 and write $b_{rs} = f(\mathbf{e}_r, \mathbf{e}_s)$. Show that $f(\mathbf{u}, \mathbf{v}) = \sum_{r,s=1}^{2} b_{rs} u_r v_s$, and that conversely every choice of four scalars b_{rs} defines a bilinear form this way. Hence there are exactly 16 bilinear forms on $\mathbb{F}_2^2$.

(b) Prove that a bilinear form on $\mathbb{F}_2^2$ is skew-symmetric if and only if $b_{12} = b_{21}$. How many are there?

(c) Prove that a bilinear form on $\mathbb{F}_2^2$ is alternating if and only if $b_{11} = b_{22} = 0$ and $b_{12} = b_{21}$. How many are there?

(d) Over $\mathbb{R}$, how do the analogous two sets of bilinear forms on $\mathbb{R}^2$ compare? Justify your answer.

Solution.

(a) Expanding $\mathbf{u} = u_1\mathbf{e}_1 + u_2\mathbf{e}_2$ and $\mathbf{v} = v_1\mathbf{e}_1 + v_2\mathbf{e}_2$ by linearity in each argument gives $f(\mathbf{u}, \mathbf{v}) = \sum_{r,s} u_r v_s f(\mathbf{e}_r, \mathbf{e}_s) = \sum_{r,s} b_{rs} u_r v_s$. Conversely, for any scalars b_{rs}, the function $\sum_{r,s} b_{rs} u_r v_s$ is linear in $\mathbf{u}$ for fixed $\mathbf{v}$ (a combination of the coordinates u_r with coefficients $\sum_s b_{rs} v_s$) and similarly in $\mathbf{v}$, and it takes the value b_{rs} at $(\mathbf{e}_r, \mathbf{e}_s)$. So bilinear forms on F^2 correspond exactly to quadruples $(b_{11}, b_{12}, b_{21}, b_{22})$, and over $\mathbb{F}_2$ there are $2^4 = 16$.

(b) In $\mathbb{F}_2$, $-a = a$ for every a, so f is skew-symmetric exactly when $f(\mathbf{v}, \mathbf{u}) = f(\mathbf{u}, \mathbf{v})$ for all $\mathbf{u}, \mathbf{v}$. ($\Rightarrow$) Taking $\mathbf{u} = \mathbf{e}_1$, $\mathbf{v} = \mathbf{e}_2$ gives $b_{21} = b_{12}$. ($\Leftarrow$) If $b_{12} = b_{21}$, then $b_{rs} = b_{sr}$ for all r, s, and $f(\mathbf{v}, \mathbf{u}) = \sum_{r,s} b_{rs} v_r u_s = \sum_{r,s} b_{sr} u_s v_r = f(\mathbf{u}, \mathbf{v})$. The entries b_{11}, b_{22} and the common value $b_{12} = b_{21}$ are free, so there are $2^3 = 8$ skew-symmetric forms.

(c) ($\Rightarrow$) If f is alternating, then $b_{11} = f(\mathbf{e}_1, \mathbf{e}_1) = 0$ and $b_{22} = 0$, and $0 = f(\mathbf{e}_1 + \mathbf{e}_2, \mathbf{e}_1 + \mathbf{e}_2) = b_{11} + b_{12} + b_{21} + b_{22} = b_{12} + b_{21}$, so $b_{21} = -b_{12} = b_{12}$. ($\Leftarrow$) With $b_{11} = b_{22} = 0$ and $b_{12} = b_{21} = b$, $f(\mathbf{u}, \mathbf{u}) = b(u_1 u_2 + u_2 u_1) = (1 + 1)\, b\, u_1 u_2 = 0$. Only $b \in \{0, 1\}$ is free, so there are 2 alternating forms. Hence $8 - 2 = 6$ skew-symmetric forms over $\mathbb{F}_2$ are not alternating; $u_1 v_1$ (with $b_{11} = 1$, all others 0) is one of them, as in the warning. A brute-force check of all 16 forms against all pairs in $\mathbb{F}_2^2$ gives the same counts, 8 and 2.

(d) Over $\mathbb{R}$ the two sets are **equal**. Alternating forms are skew-symmetric by Theorem 7.2.2 (a), and since $\mathbb{R}$ has characteristic 0, skew-symmetric forms are alternating by part (b) of that theorem. Concretely, both conditions say $b_{11} = b_{22} = 0$ and $b_{21} = -b_{12}$.

7.3 Permutations and Sign

An alternating form changes sign when two of its arguments are swapped (Theorem 7.2.1). To compare $f(v_3, v_1, v_2)$ with $f(v_1, v_2, v_3)$, we reach one list from the other by swaps and collect a factor -1 per swap. But the same rearrangement can be reached by different sequences of swaps, and Chapter 0 showed that even their number is not unique. If one route used an even number of swaps and another an odd number, we would get $f(v_1, v_2, v_3) = -f(v_1, v_2, v_3)$ for every alternating form, which is absurd for signed volume. This section rules that out. It attaches to each permutation a sign defined without any choices, proves that the sign is multiplicative, and deduces that the parity of the number of transpositions is well defined, as Chapter 0 promised.

7.3.1 Permutations, recalled

Recall from Chapter 0 that S_n is the group of bijections $\{1, \dots, n\} \to \{1, \dots, n\}$, with $n!$ elements. A permutation can be written in two-line notation, with $\sigma(i)$ below i, or as a product of cycles. The product $\sigma\tau$ is the composition $\sigma \circ \tau$: **first** τ**, then** σ, so products of cycles are read from right to left (Example (Computing with cycles)). A transposition $(i\ j)$ swaps i and j and fixes everything else, and for $n \geq 2$ every permutation is a product of transpositions (Theorem 0.10.4). The product is far from unique, since

$$(1\ 2\ 3) = (1\ 3)(1\ 2) = (1\ 2)(2\ 3) = (1\ 3)(1\ 2)(1\ 2)(1\ 2).$$

Here the numbers of transpositions are $2, 2, 4$, all even. The goal is to show this is no accident.

Why can we not simply **define** the sign of σ as $(-1)^r$, where σ is a product of r transpositions? Because that definition depends on a choice of product, and we would have to prove that every choice gives the same answer. That is the statement we are trying to prove. We need a definition that involves no choice at all.

7.3.2 Inversions and the sign

To measure how much a permutation scrambles $1, 2, \dots, n$, look at the bottom row $\sigma(1), \dots, \sigma(n)$ of its two-line notation and count the pairs of entries that are in the wrong order. A sorted row has none, and each swap of two neighboring entries changes the count by exactly one. So the count is a natural candidate for tracking the parity of swaps.

The sign of a permutation records whether it puts an even or an odd number of pairs out of order.

Definition 7.3.1. Inversion

Let $\sigma \in S_n$. An **inversion** of σ is a pair (i, j) of positions with $1 \leq i < j \leq n$ **and** $\sigma(i) > \sigma(j)$. We write $\mathrm{inv}(\sigma)$ for the number of inversions of σ.

Definition 7.3.2. Sign of a Permutation

The **sign** of $\sigma \in S_n$ is

$$\operatorname{sgn}\sigma = (-1)^{\operatorname{inv}(\sigma)} \in \{1, -1\}.$$

The permutation σ is **even** if $\operatorname{sgn}\sigma = 1$ and **odd** if $\operatorname{sgn}\sigma = -1$.

In words: an inversion is a pair of positions whose order σ **reverses**. The condition $i < j$ makes each pair of positions count once, and $\sigma(i) > \sigma(j)$ says the values come out in the opposite order. Since $\operatorname{inv}(\sigma)$ is the size of a subset of the finite set of pairs $i < j$, it is a well-defined integer between 0 and $n(n-1)/2$, and there is nothing to choose. That is the whole reason for this definition.

The identity. Its bottom row is sorted, so $\operatorname{inv}(\mathrm{id}) = 0$ and id is even. In $S_1 = \{\mathrm{id}\}$ this is the only permutation, with sign 1. The degenerate case matters: the identity is the empty product of transpositions, and its sign $1 = (-1)^0$ is what a product of 0 swaps should give.

All of S_3. Here is each permutation with its bottom row and inversions.

σ	id	(1 2)	(2 3)	(1 3)	(1 2 3)	(1 3 2)
bottom row	1 2 3	2 1 3	1 3 2	3 2 1	2 3 1	3 1 2
inversions	none	$(1,2)$	$(2,3)$	$(1,2)$, $(1,3)$, $(2,3)$	$(1,3)$, $(2,3)$	$(1,2)$, $(1,3)$
sgn	1	-1	-1	-1	1	1

For (1 2 3), the bottom row is $\sigma(1), \sigma(2), \sigma(3) = 2, 3, 1$, and the out-of-order pairs of positions are $(1,3)$, since $2 > 1$, and $(2,3)$, since $3 > 1$.

Two things stand out. The three transpositions are odd, although (1 3) has three inversions, not one. And the two 3-cycles are even, consistent with (1 2 3) = (1 3)(1 2) being a product of two transpositions.

Why all pairs, not just neighbors. A tempting shortcut counts only **adjacent** positions $(i, i+1)$ with $\sigma(i) > \sigma(i+1)$. For (1 3), with bottom row $3, 2, 1$, this count is 2, which would make (1 3) "even". But (1 3) = (1 2)(2 3)(1 2), as one checks by following $1 \mapsto 2 \mapsto 3 \mapsto 3$, $2 \mapsto 1 \mapsto 1 \mapsto 2$, $3 \mapsto 3 \mapsto 2 \mapsto 1$, and each factor has adjacent count 1. A count that is to be multiplicative must give $(-1)^3 = -1$ here, so the adjacent count is the wrong one.

Quick check. Find sgn(1 4) in S_4 by counting inversions.

Solution. The bottom row is $4, 2, 3, 1$. The pairs of positions in the wrong order are $(1,2), (1,3), (1,4)$, since 4 is larger than everything after it, and $(2,4), (3,4)$, since $2 > 1$ and $3 > 1$. The pair $(2,3)$ is in order. So $\operatorname{inv}(1\ 4) = 5$ and $\operatorname{sgn}(1\ 4) = -1$.

7.3.3 The sign is multiplicative

The key property is that signs multiply when permutations are composed.

Theorem 7.3.1. The Sign Is Multiplicative

For all $\sigma, \tau \in S_n$,

$$\operatorname{sgn}(\sigma\tau) = \operatorname{sgn}\sigma \cdot \operatorname{sgn}\tau.$$

Idea. Count inversions one pair at a time, and record each pair as $+1$ (order kept) or -1 (order reversed). The sign is the product of these ± 1's. Applying τ and then σ moves a pair of positions $\{i, j\}$ to $\{\tau(i), \tau(j)\}$ and then to $\{\sigma\tau(i), \sigma\tau(j)\}$. The order is reversed overall exactly when it is reversed at one of the two stages but not both, which is multiplication of ± 1's. Finally, as $\{i, j\}$ runs over all pairs, so does $\{\tau(i), \tau(j)\}$, so the contributions of σ multiply to $\operatorname{sgn}\sigma$, just in a different order.

Proof. For a nonzero integer m let $s(m) = 1$ if $m > 0$ and $s(m) = -1$ if $m < 0$. Then $s(mm') = s(m)s(m')$ for nonzero integers m, m', by the sign rules for products of integers.

Let $\mathcal{P}$ be the set of 2-element subsets $\{i, j\}$ of $\{1, \dots, n\}$. For $\rho \in S_n$ and $\{i, j\} \in \mathcal{P}$, define

$$\varepsilon_\rho(\{i, j\}) = s\big((\rho(i) - \rho(j))(i - j)\big).$$

This is defined: $i \neq j$ and $\rho(i) \neq \rho(j)$ since ρ is injective, so the product is nonzero; and exchanging the names i and j negates both factors, so the product does not depend on the order in which the subset is written. If $i < j$, then $i - j < 0$, so $\varepsilon_\rho(\{i, j\}) = -1$ exactly when $\rho(i) > \rho(j)$, that is, when (i, j) is an inversion of ρ. Hence, by Definition 7.3.2,

$$\operatorname{sgn}\rho = \prod_{P \in \mathcal{P}} \varepsilon_\rho(P). \tag{1}$$

Now let $\{i, j\} \in \mathcal{P}$. Since $\tau(i) \neq \tau(j)$, we have $(\tau(i) - \tau(j))^2 > 0$. Multiplying inside s by this positive integer does not change the value, so

$$\begin{aligned} \varepsilon_{\sigma\tau}(\{i, j\}) &= s\Big(\big(\sigma(\tau(i)) - \sigma(\tau(j))\big)\big(\tau(i) - \tau(j)\big) \cdot \big(\tau(i) - \tau(j)\big)(i - j)\Big) \\ &= \varepsilon_\sigma(\{\tau(i), \tau(j)\})\ \varepsilon_\tau(\{i, j\}). \end{aligned}$$

Let $\hat{\tau} \colon \mathcal{P} \to \mathcal{P}$, $\{i, j\} \mapsto \{\tau(i), \tau(j)\}$; the image has two elements because τ is injective. The same construction with τ^{-1} gives an inverse function, so $\hat{\tau}$ is a bijection (Theorem 0.4.7). Taking the product over all $P \in \mathcal{P}$ and using (1) three times,

$$\operatorname{sgn}(\sigma\tau) = \prod_{P \in \mathcal{P}} \varepsilon_\sigma(\hat{\tau}(P)) \cdot \prod_{P \in \mathcal{P}} \varepsilon_\tau(P) = \prod_{Q \in \mathcal{P}} \varepsilon_\sigma(Q) \cdot \operatorname{sgn}\tau = \operatorname{sgn}\sigma \cdot \operatorname{sgn}\tau,$$

where the second equality substitutes $Q = \hat{\tau}(P)$, which runs over $\mathcal{P}$ exactly once because $\hat{\tau}$ is a bijection, and reorders the finite product of integers. This proves the theorem.

■

By induction on the number of factors, $\operatorname{sgn}(\sigma_1\sigma_2\cdots\sigma_r) = \operatorname{sgn}\sigma_1 \cdots \operatorname{sgn}\sigma_r$. In the language of Chapter 0, the set $\{1, -1\}$ is a group under multiplication (it contains 1, and it is closed under products and inverses, since $(-1)^{-1} = -1$), and the theorem says exactly that

$$\operatorname{sgn} \colon S_n \to \{1, -1\}$$

is a group homomorphism (Definition 0.10.3).

7.3.4 Transpositions and parity

Now we can compute the sign of the building blocks.

Corollary 7.3.2. Sign of a Transposition

Every transposition is odd: $\operatorname{sgn}(i\ j) = -1$ for all $i \neq j$ in $\{1, \dots, n\}$.

Proof. Since $(i\ j) = (j\ i)$, we may assume $i < j$. Let $\tau = (i\ j)$ and let $p < q$ be positions. We decide when $\tau(p) > \tau(q)$.

- If neither p nor q lies in $\{i, j\}$, then $\tau(p) = p < q = \tau(q)$: no inversion.
- If $(p, q) = (i, j)$, then $\tau(i) = j > i = \tau(j)$: one inversion.
- If $p = i$ and $q \neq j$, then $\tau(p) = j$ and $\tau(q) = q$, an inversion exactly when $i < q < j$. This gives $j - i - 1$ inversions.
- If $q = j$ and $p \neq i$, then $\tau(p) = p$ and $\tau(q) = i$, an inversion exactly when $i < p < j$. This gives $j - i - 1$ inversions.
- If $p = j$ (so $q > j$), then $\tau(p) = i < q = \tau(q)$; if $q = i$ (so $p < i$), then $\tau(p) = p < j = \tau(q)$: no inversion.

These cases cover every pair, so $\operatorname{inv}(\tau) = 1 + 2(j - i - 1)$, an odd number. Hence $\operatorname{sgn} \tau = -1$.

■

The Quick check above is the case $(i, j) = (1, 4)$: $1 + 2 \cdot 2 = 5$ inversions. Combining the corollary with multiplicativity gives the fact that Chapter 0 left open.

Corollary 7.3.3. Parity Is Well Defined

Let $n \geq 2$ and $\sigma \in S_n$. If $\sigma = \tau_1 \cdots \tau_r$ is a product of r transpositions, then $\operatorname{sgn} \sigma = (-1)^r$. Consequently, if also $\sigma = \tau'_1 \cdots \tau'_s$ with transpositions τ'_m, then r and s are both even or both odd.

Proof. By Theorem 7.3.1, extended to r factors, and Corollary 7.3.2, $\operatorname{sgn} \sigma = \operatorname{sgn} \tau_1 \cdots \operatorname{sgn} \tau_r = (-1)^r$. In the same way $\operatorname{sgn} \sigma = (-1)^s$. Hence $(-1)^r = (-1)^s$, so $r - s$ is even.

■

So a permutation is even exactly when it is a product of an even number of transpositions, and the "number of swaps" has a well-defined parity even though the number itself is not defined.

Proposition 7.3.4. Sign of a Cycle

A k-cycle has sign $(-1)^{k-1}$. In particular, cycles of even length are odd and cycles of odd length are even.

Proof. For $k = 1$ the cycle is id, with sign $1 = (-1)^0$. Let $k \geq 2$ and let $\gamma = (i_1\ i_2\ \cdots\ i_k)$. We claim

$$\gamma = (i_1\ i_k)(i_1\ i_{k-1}) \cdots (i_1\ i_3)(i_1\ i_2).$$

Apply the factors from right to left. The element i_1 is sent to i_2 by $(i_1\ i_2)$, and no later factor involves i_2. For $2 \leq m \leq k-1$, the element i_m is fixed by the factors before $(i_1\ i_m)$, sent to i_1 by it, then to i_{m+1} by $(i_1\ i_{m+1})$, and no later factor involves i_{m+1}. The element i_k is fixed until the last factor $(i_1\ i_k)$ sends it to i_1. Elements outside the cycle are fixed by every factor. This agrees with γ everywhere, proving the claim. The product has $k-1$ transpositions, so $\operatorname{sgn}\gamma = (-1)^{k-1}$ by Corollary 7.3.3.

■

Warning. **Only the parity of the number of transpositions is determined, and it is not the parity of the cycle length.** The identity in S_3 equals $(1\ 2)(1\ 2)$ and also $(1\ 2)(1\ 3)(1\ 2)(1\ 3)(1\ 2)(1\ 3)$, since $((1\ 2)(1\ 3))^3 = (1\ 3\ 2)^3 = \mathrm{id}$; the numbers 2 and 6 differ, only their parity agrees. And a 3-cycle, "of length 3", is **even**: $(1\ 2\ 3) = (1\ 3)(1\ 2)$. One more trap concerns order: $(1\ 3)(1\ 2) = (1\ 2\ 3)$ but $(1\ 2)(1\ 3) = (1\ 3\ 2)$. Multiplying in the wrong order gives the wrong permutation and the wrong bottom row, even though, by Theorem 7.3.1, it happens to give the right sign.

7.3.5 The alternating group

Since sgn is a homomorphism, its kernel is a subgroup of S_n (Theorem 0.10.3).

Definition 7.3.3. Alternating Group

The **alternating group** A_n is the kernel of $\operatorname{sgn}\colon S_n \to \{1, -1\}$, that is, the subgroup of **even** permutations in S_n.

For example, $A_3 = \{\mathrm{id}, (1\ 2\ 3), (1\ 3\ 2)\}$ by the table above, and $A_2 = \{\mathrm{id}\}$. For $n \geq 2$, exactly half of the permutations are even (Exercise C1). The odd permutations do **not** form a subgroup: they do not contain id. The name is no coincidence: as we see below and in the next section, rearranging the arguments of an alternating form by an even permutation leaves its value unchanged.

7.3.6 Back to alternating forms

Here is why this section exists. Let f be an alternating 3-linear form and compare $f(v_3, v_1, v_2)$ with $f(v_1, v_2, v_3)$. Swapping the first two arguments of (v_3, v_1, v_2) gives (v_1, v_3, v_2), and swapping the last two then gives (v_1, v_2, v_3). By Theorem 7.2.1 (a) each swap costs a factor -1, so $f(v_3, v_1, v_2) = (-1)^2 f(v_1, v_2, v_3) = f(v_1, v_2, v_3)$. The list (v_3, v_1, v_2) is $(v_{\sigma(1)}, v_{\sigma(2)}, v_{\sigma(3)})$ for $\sigma = (1\ 3\ 2)$, with bottom row $3, 1, 2$, and the factor 1 we found is $\operatorname{sgn}\sigma$.

A different sequence of swaps could have been used, possibly a longer one. By Corollary 7.3.3, every such sequence has an even number of swaps, so every route gives the same factor. In general, the next section proves that $f(v_{\sigma(1)}, \ldots, v_{\sigma(k)}) = \operatorname{sgn}\sigma \cdot f(v_1, \ldots, v_k)$ for every alternating k-linear form, and uses it to turn each of the $n!$ surviving terms of the expansion sketched in

the first section of this chapter into $\pm D(\mathbf{e}_1, \ldots, \mathbf{e}_n)$. Without the corollary, the factor computed swap by swap could depend on the route, and could not be identified with $\operatorname{sgn}\sigma$.

7.3.7 Computing signs in practice

There are now two independent ways to find a sign: count inversions, or factor into cycles and use Proposition 7.3.4 with multiplicativity. Doing both is a good check.

Example. A Sign Computed Two Ways

Find the sign of the permutation $\sigma \in S_6$ with two-line notation $\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 4 & 5 & 3 & 6 & 2 & 1 \end{pmatrix}$, first by counting inversions and then from its cycles.

Solution. Inversions. Go through the bottom row $4, 5, 3, 6, 2, 1$ from left to right, and for each entry count the **later** entries that are smaller:

position i	1	2	3	4	5	6
$\sigma(i)$	4	5	3	6	2	1
later and smaller	3, 2, 1	3, 2, 1	2, 1	2, 1	1	none
count	3	3	2	2	1	0

So $\operatorname{inv}(\sigma) = 3+3+2+2+1+0 = 11$, which is odd, and $\operatorname{sgn}\sigma = -1$.

Cycles. Following elements as in Example (Computing with cycles): $1 \mapsto 4 \mapsto 6 \mapsto 1$, then $2 \mapsto 5 \mapsto 2$, and $3 \mapsto 3$. So $\sigma = (1\ 4\ 6)(2\ 5)$. By Proposition 7.3.4, the 3-cycle has sign 1 and the transposition has sign -1, so by Theorem 7.3.1, $\operatorname{sgn}\sigma = 1 \cdot (-1) = -1$. The two methods agree.

The following cell checks the example, and also tests Theorem 7.3.1 and Proposition 7.3.4 on all of S_5 by brute force. Permutations are stored as tuples of values $(\sigma(1), \ldots, \sigma(n))$, shifted to start at 0.

```
from itertools import permutations, combinations

def inv(s):
    return sum(1 for i, j in combinations(range(len(s)), 2) if s[i] > s[j])

def sgn(s):
    return (-1) ** inv(s)

def compose(s, t):          # (s t)(i) = s(t(i)): first t, then s
    return tuple(s[t[i]] for i in range(len(t)))

def cycle(entries, n):      # the cycle (entries[0] entries[1] ...), 0-based
    s = list(range(n))
    for a, b in zip(entries, entries[1:] + entries[:1]):
        s[a] = b
    return tuple(s)
```

```
sigma = (4, 5, 3, 6, 2, 1)
sigma0 = tuple(x - 1 for x in sigma)
print("inversions:", inv(sigma0), " sign:", sgn(sigma0))
print("(1 4 6)(2 5) equals sigma:", compose(cycle([0, 3, 5], 6), cycle([1, 4], 6))
    == sigma0)

S5 = list(permutations(range(5)))
print("multiplicative on S_5:", all(sgn(compose(s, t)) == sgn(s) * sgn(t) for s in
    S5 for t in S5))
print("k-cycles have sign (-1)^(k-1):", all(sgn(cycle(list(range(k)), 5)) == (-1) **
     (k - 1) for k in range(1, 6)))
```

7.3.8 Exercises

7.3.8.1 A. Check your understanding

Exercise. A1

(a) Define an inversion of $\sigma \in S_n$ and the sign $\operatorname{sgn}\sigma$.

(b) True or false: every 3-cycle is odd. Justify your answer.

(c) True or false: some permutation is a product of 5 transpositions and also a product of 8 transpositions. Justify your answer.

(d) What is the kernel of $\operatorname{sgn}\colon S_n \to \{1,-1\}$? Is the set of odd permutations a subgroup?

(e) True or false: $\operatorname{sgn}(\sigma\tau) = \operatorname{sgn}(\tau\sigma)$ for all $\sigma, \tau \in S_n$, even when $\sigma\tau \neq \tau\sigma$. Justify your answer.

(f) Let f be an alternating 3-linear form. Express $f(v_2, v_3, v_1)$ in terms of $f(v_1, v_2, v_3)$.

Solution.

(a) An inversion of σ is a pair (i, j) with $1 \le i < j \le n$ and $\sigma(i) > \sigma(j)$ (Definition 7.3.1); $\operatorname{sgn}\sigma = (-1)^{\operatorname{inv}(\sigma)}$ (Definition 7.3.2).

(b) False. By Proposition 7.3.4 a 3-cycle has sign $(-1)^2 = 1$, so it is even; for instance $(1\ 2\ 3) = (1\ 3)(1\ 2)$.

(c) False. By Corollary 7.3.3 the numbers of transpositions in two such products have the same parity, and 5 and 8 do not.

(d) The kernel is the alternating group A_n of even permutations (Definition 7.3.3). The odd permutations do not form a subgroup: id is even, so it is not among them.

(e) True. By Theorem 7.3.1 both sides equal $\operatorname{sgn}\sigma \cdot \operatorname{sgn}\tau$, and multiplication of integers is commutative. For example $(1\ 3)(1\ 2) \neq (1\ 2)(1\ 3)$, but both are 3-cycles, with sign 1.

(f) Swapping the first two arguments of (v_2, v_3, v_1) gives (v_3, v_2, v_1), and swapping the first and third then gives (v_1, v_2, v_3). Each swap costs a factor -1 by Theorem 7.2.1 (a), so $f(v_2, v_3, v_1) = f(v_1, v_2, v_3)$. This matches the sign of the rearrangement σ with bottom row $2, 3, 1$, namely $\sigma = (1\ 2\ 3)$, which is even.

7.3.8.2 B. Practice

Exercise. B1: Computing Signs

Find the sign of each permutation, both by counting inversions and by using cycles.

(a) $\sigma \in S_4$ with two-line notation $\begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 4 & 1 & 3 \end{pmatrix}$.

(b) $(1\ 2)(3\ 4\ 5) \in S_5$.

(c) $(1\ 2\ 3)(2\ 3\ 4) \in S_4$.

Solution.

(a) *Inversions.* In the bottom row $2, 4, 1, 3$, the entry 2 is followed by the smaller entry 1, the entry 4 by 1 and 3, and the entry 1 by nothing smaller. So $\operatorname{inv}(\sigma) = 1 + 2 + 0 = 3$ and $\operatorname{sgn}\sigma = -1$. *Cycles.* $1 \mapsto 2 \mapsto 4 \mapsto 3 \mapsto 1$, so $\sigma = (1\ 2\ 4\ 3)$, a 4-cycle, with sign $(-1)^3 = -1$ by Proposition 7.3.4.

(b) *Cycles.* By Proposition 7.3.4 and Theorem 7.3.1, the sign is $(-1) \cdot (-1)^2 = -1$. *Inversions.* The permutation sends $1 \mapsto 2$, $2 \mapsto 1$, $3 \mapsto 4$, $4 \mapsto 5$, $5 \mapsto 3$, so the bottom row is $2, 1, 4, 5, 3$. The inversions are $(1, 2)$, $(3, 5)$ and $(4, 5)$, so there are 3, and the sign is -1.

(c) *Cycles.* Both factors are 3-cycles, so the sign is $1 \cdot 1 = 1$. *Inversions.* Apply $(2\ 3\ 4)$ first: $1 \mapsto 1 \mapsto 2$, $2 \mapsto 3 \mapsto 1$, $3 \mapsto 4 \mapsto 4$, $4 \mapsto 2 \mapsto 3$. The bottom row is $2, 1, 4, 3$, with inversions $(1, 2)$ and $(3, 4)$. So the sign is $(-1)^2 = 1$, and in fact the product is $(1\ 2)(3\ 4)$.

Exercise. B2: Sign of the Inverse

Prove that $\operatorname{sgn}(\sigma^{-1}) = \operatorname{sgn}\sigma$ for every $\sigma \in S_n$.

Solution. Let $\sigma \in S_n$. By Theorem 7.3.1, $\operatorname{sgn}\sigma \cdot \operatorname{sgn}(\sigma^{-1}) = \operatorname{sgn}(\sigma\sigma^{-1}) = \operatorname{sgn}(\mathrm{id}) = 1$, since id has no inversions. Since $\operatorname{sgn}\sigma \in \{1, -1\}$ and $(\pm 1)^{-1} = \pm 1$, multiplying both sides by $\operatorname{sgn}\sigma$ gives $\operatorname{sgn}(\sigma^{-1}) = \operatorname{sgn}\sigma$.

Exercise. B3: The Reversal Permutation

Let $\rho \in S_n$ be the reversal, $\rho(i) = n + 1 - i$. Show that $\operatorname{inv}(\rho) = n(n-1)/2$. Hence find the n for which ρ is even, and decide the sign of ρ for $n = 6$ and $n = 9$.

Solution. For $1 \le i < j \le n$, $\rho(i) = n + 1 - i > n + 1 - j = \rho(j)$. So **every** pair $i < j$ is an inversion, and $\operatorname{inv}(\rho)$ is the number of 2-element subsets of $\{1, \dots, n\}$, which is $n(n-1)/2$. Hence $\operatorname{sgn}\rho = (-1)^{n(n-1)/2}$.

Write $n = 4q + t$ with $t \in \{0, 1, 2, 3\}$. Then $n(n-1)/2$ is $2q(4q-1)$ for $t = 0$, $2q(4q+1)$ for $t = 1$, $(4q+1)(2q+1)$ for $t = 2$, and $(2q+1)(4q+3)$ for $t = 3$. The first two are even and the last two are odd. So ρ is even exactly when n leaves remainder 0 or 1 on division by 4. For $n = 6$, $\operatorname{inv}(\rho) = 15$ and ρ is odd; for $n = 9$, $\operatorname{inv}(\rho) = 36$ and ρ is even.

7.3.8.3 C. Going deeper

Exercise. C1: Half the Permutations Are Even

Let $n \geq 2$. Prove that exactly $n!/2$ permutations in S_n are even.

Hint: consider $\sigma \mapsto (1\ 2)\sigma$.

Solution. Let E and O be the sets of even and odd permutations; they are disjoint and $E \cup O = S_n$, so $|E| + |O| = n!$. Let $\tau = (1\ 2)$, which exists since $n \geq 2$, and define $L\colon S_n \to S_n$, $L(\sigma) = \tau\sigma$. By Theorem 7.3.1 and Corollary 7.3.2, $\operatorname{sgn} L(\sigma) = -\operatorname{sgn}\sigma$, so L maps E into O and O into E. Since $\tau\tau = \mathrm{id}$, $L(L(\sigma)) = \sigma$, so L restricted to E is a bijection $E \to O$ with inverse L restricted to O (Theorem 0.4.7). Hence $|E| = |O| = n!/2$.

Exercise. C2: Sign from the Number of Cycles

Suppose $\sigma \in S_n$ is written as a product $\sigma = \gamma_1\gamma_2\cdots\gamma_c$ of **disjoint** cycles (no element appears in two of them) whose lengths $k_1, \ldots, k_c$ add up to n, where fixed points are included as cycles of length 1. (Such a product always exists: follow each element around its loop, as in Example (Computing with cycles).)

(a) Prove that $\operatorname{sgn}\sigma = (-1)^{n-c}$.

(b) Use (a) to find the sign of $\sigma = (1\ 4\ 6)(2\ 5) \in S_6$, and compare with Example (A Sign Computed Two Ways).

(c) Deduce that an n-cycle in S_n and the identity in S_n have the same sign exactly when n is odd.

Solution.

(a) By Theorem 7.3.1 and Proposition 7.3.4,

$$\operatorname{sgn}\sigma = \prod_{m=1}^{c} \operatorname{sgn}\gamma_m = \prod_{m=1}^{c} (-1)^{k_m - 1} = (-1)^{(k_1+\cdots+k_c)-c} = (-1)^{n-c}.$$

(b) Including the fixed point 3 as the cycle (3), $\sigma = (1\ 4\ 6)(2\ 5)(3)$, with lengths $3+2+1 = 6$ and $c = 3$. So $\operatorname{sgn}\sigma = (-1)^{6-3} = -1$, as in Example (A Sign Computed Two Ways).

(c) An n-cycle has $c = 1$ and sign $(-1)^{n-1}$; the identity consists of n cycles of length 1, so $c = n$ and its sign is $(-1)^0 = 1$. These agree exactly when $n - 1$ is even, that is, when n is odd.

7.4 Existence and Uniqueness of the Determinant

In the first section of this chapter, signed area obeyed three rules: it is linear in each side, it vanishes when two sides are equal, and the unit square has area 1. In the second section we named functions obeying the first two rules, alternating multilinear forms, and in the third we

built the sign of a permutation. This section puts the pieces together. We prove that on F^n there is **exactly one** alternating n-linear form that takes the value 1 on the standard basis, and we find a formula for it. That form is the determinant.

7.4.1 The goal and the plan

Here is the statement we are after.

> There is exactly one function $D\colon F^n \times \cdots \times F^n \to F$ (with n factors) that is n-linear (Definition 7.2.1), alternating (Definition 7.2.2), and satisfies $D(\mathbf{e}_1, \dots, \mathbf{e}_n) = 1$.

It has two halves of a very different nature. **Uniqueness** says: if such a D exists, its values are forced. **Existence** says: some function really has all three properties. We prove uniqueness first, because its proof is how we *find* the formula. Suppose D exists, expand everything using the rules, and see what is left. The formula that comes out is then the only candidate, and existence reduces to checking that the candidate obeys the rules.

Throughout, a list of n columns $\mathbf{a}_1, \dots, \mathbf{a}_n \in F^n$ and the matrix $\mathbf{A} = \begin{pmatrix} \mathbf{a}_1 & \cdots & \mathbf{a}_n \end{pmatrix} \in M_n(F)$ carry the same information, and we pass between them freely. We write a_{ij} for the i-th entry of $\mathbf{a}_j$, so that

$$\mathbf{a}_j = a_{1j}\mathbf{e}_1 + a_{2j}\mathbf{e}_2 + \cdots + a_{nj}\mathbf{e}_n = \sum_{i=1}^{n} a_{ij}\mathbf{e}_i.$$

7.4.2 Rearranging the arguments

Expanding a form on basis vectors will produce values like $f(\mathbf{e}_3, \mathbf{e}_1, \mathbf{e}_2)$: the standard basis in a scrambled order. Alternating forms handle scrambling by a sign, and the sign is exactly the one built from inversions.

Lemma 7.4.1. Permuting the arguments of an alternating form

Let f be an alternating n-linear form on a vector space V over F, let $v_1, \dots, v_n \in V$, and let $\sigma \in S_n$. Then

$$f(v_{\sigma(1)}, v_{\sigma(2)}, \dots, v_{\sigma(n)}) = \operatorname{sgn}(\sigma)\, f(v_1, v_2, \dots, v_n).$$

Idea. A transposition swaps two arguments, and each swap costs a factor -1 by Theorem 7.2.1. Every permutation is a product of transpositions, so the value picks up $(-1)^k$ for a product of k of them. Multiplicativity of the sign says that $(-1)^k$ is $\operatorname{sgn}(\sigma)$, whichever product we chose.

Proof. If $n = 1$, then $S_1 = \{\mathrm{id}\}$ and $\operatorname{sgn}(\mathrm{id}) = 1$, since the identity has no inversions (Definition 7.3.2); there is nothing to prove. Let $n \geq 2$. For $\pi \in S_n$, write $f_\pi = f(v_{\pi(1)}, \dots, v_{\pi(n)})$.

First let $\pi \in S_n$ and let $\tau = (p\ q)$ be a transposition. The j-th argument of $f_{\pi\tau}$ is $v_{\pi(\tau(j))}$. For $j \neq p, q$ this is $v_{\pi(j)}$; for $j = p$ it is $v_{\pi(q)}$; for $j = q$ it is $v_{\pi(p)}$. So the arguments of $f_{\pi\tau}$ are those of f_π with positions p and q exchanged. By Theorem 7.2.1, swapping two arguments changes the sign, so

$$f_{\pi\tau} = -f_\pi. \tag{$*$}$$

Now let $\sigma \in S_n$. By Theorem 0.10.4, $\sigma = \tau_1\tau_2\cdots\tau_k$ for some transpositions $\tau_1,\dots,\tau_k$ with $k \geq 1$. Put $\pi_0 = \mathrm{id}$ and $\pi_m = \tau_1\cdots\tau_m$, so that $\pi_m = \pi_{m-1}\tau_m$. Applying $(*)$ k times,

$$f_\sigma = f_{\pi_k} = -f_{\pi_{k-1}} = \cdots = (-1)^k f_{\pi_0} = (-1)^k f(v_1,\dots,v_n).$$

On the other hand, by Theorem 7.3.1 and Corollary 7.3.2, $\operatorname{sgn}(\sigma) = \operatorname{sgn}(\tau_1)\cdots\operatorname{sgn}(\tau_k) = (-1)^k$. This proves the lemma.

■

The proof chose a factorization of σ into transpositions, and different choices can have different lengths. The answer does not depend on the choice, because only the parity of k enters, and Corollary 7.3.3 says the parity is determined by σ. In the proof this is absorbed into the equation $(-1)^k = \operatorname{sgn}(\sigma)$.

7.4.3 Uniqueness

We now expand an alternating form completely. It helps to see the case $n = 2$ first.

Theorem 7.4.2. Uniqueness of Alternating Forms

Let f be an alternating n-linear form on F^n, and let $\mathbf{a}_1,\dots,\mathbf{a}_n \in F^n$ have entries a_{ij} as above. Then

$$f(\mathbf{a}_1,\dots,\mathbf{a}_n) = f(\mathbf{e}_1,\dots,\mathbf{e}_n)\sum_{\sigma\in S_n}\operatorname{sgn}(\sigma)\,a_{\sigma(1)1}\,a_{\sigma(2)2}\cdots a_{\sigma(n)n}.$$

In particular, an alternating n-linear form on F^n is determined by the single value $f(\mathbf{e}_1,\dots,\mathbf{e}_n)$.

Idea. For $n = 2$, write $\mathbf{a}_1 = a_{11}\mathbf{e}_1 + a_{21}\mathbf{e}_2$ and $\mathbf{a}_2 = a_{12}\mathbf{e}_1 + a_{22}\mathbf{e}_2$. Linearity in the first argument, then in the second, gives four terms:

$$\begin{aligned} f(\mathbf{a}_1,\mathbf{a}_2) &= a_{11}\,f(\mathbf{e}_1,\mathbf{a}_2) + a_{21}\,f(\mathbf{e}_2,\mathbf{a}_2)\\ &= a_{11}a_{12}\,f(\mathbf{e}_1,\mathbf{e}_1) + a_{11}a_{22}\,f(\mathbf{e}_1,\mathbf{e}_2)\\ &\quad + a_{21}a_{12}\,f(\mathbf{e}_2,\mathbf{e}_1) + a_{21}a_{22}\,f(\mathbf{e}_2,\mathbf{e}_2). \end{aligned}$$

The first and last terms vanish because two arguments are equal, and $f(\mathbf{e}_2,\mathbf{e}_1) = -f(\mathbf{e}_1,\mathbf{e}_2)$. What is left is $(a_{11}a_{22} - a_{21}a_{12})\,f(\mathbf{e}_1,\mathbf{e}_2)$, the formula for $n = 2$.

In general the four terms become one term for each way of choosing, for every argument j, which basis vector $\mathbf{e}_{\varphi(j)}$ to take from $\mathbf{a}_j$. A choice is a function $\varphi\colon\{1,\dots,n\}\to\{1,\dots,n\}$, and there are n^n of them. The ones that pick some basis vector twice die. The survivors pick every basis vector once, so they are permutations, and Lemma 7.4.1 converts each into a sign.

Proof. Write $[k] = \{1,\dots,k\}$ for $k \geq 1$. We first expand the arguments one at a time.

Claim. For each $k \in \{0, 1, \dots, n\}$,

$$f(\mathbf{a}_1, \dots, \mathbf{a}_n) = \sum_{\varphi\colon [k]\to[n]} a_{\varphi(1)1} \cdots a_{\varphi(k)k}\, f(\mathbf{e}_{\varphi(1)}, \dots, \mathbf{e}_{\varphi(k)}, \mathbf{a}_{k+1}, \dots, \mathbf{a}_n),$$

where the sum runs over **all** functions $\varphi\colon [k] \to [n]$. For $k = 0$ we read the right side as the single term $f(\mathbf{a}_1, \dots, \mathbf{a}_n)$.

Proof. We use induction on k. For $k = 0$ both sides are $f(\mathbf{a}_1, \dots, \mathbf{a}_n)$. Suppose the claim holds for some $k < n$. In each term, the argument in position $k+1$ is $\mathbf{a}_{k+1} = \sum_{i=1}^{n} a_{i,k+1}\mathbf{e}_i$. Since f is linear in argument $k+1$,

$$f(\mathbf{e}_{\varphi(1)}, \dots, \mathbf{e}_{\varphi(k)}, \mathbf{a}_{k+1}, \dots, \mathbf{a}_n) = \sum_{i=1}^{n} a_{i,k+1}\, f(\mathbf{e}_{\varphi(1)}, \dots, \mathbf{e}_{\varphi(k)}, \mathbf{e}_i, \mathbf{a}_{k+2}, \dots, \mathbf{a}_n).$$

Substituting this into the induction hypothesis gives a sum over all pairs (φ, i) with $\varphi\colon [k] \to [n]$ and $i \in [n]$. Such a pair is the same thing as a function $\psi\colon [k+1] \to [n]$, namely $\psi(j) = \varphi(j)$ for $j \le k$ and $\psi(k+1) = i$; and each ψ arises from exactly one pair. The term for the pair (φ, i) is $a_{\psi(1)1} \cdots a_{\psi(k+1),k+1}\, f(\mathbf{e}_{\psi(1)}, \dots, \mathbf{e}_{\psi(k+1)}, \mathbf{a}_{k+2}, \dots, \mathbf{a}_n)$. This is the claim for $k+1$.

■

Taking $k = n$ in the claim,

$$f(\mathbf{a}_1, \dots, \mathbf{a}_n) = \sum_{\varphi\colon [n]\to[n]} a_{\varphi(1)1} \cdots a_{\varphi(n)n}\, f(\mathbf{e}_{\varphi(1)}, \dots, \mathbf{e}_{\varphi(n)}). \tag{$**$}$$

We sort the functions φ into two kinds.

Case 1: φ is not injective. Then $\varphi(p) = \varphi(q)$ for some $p \neq q$, so the arguments in positions p and q of $f(\mathbf{e}_{\varphi(1)}, \dots, \mathbf{e}_{\varphi(n)})$ are equal. Since f is alternating, this value is 0, and the whole term is 0.

Case 2: φ is injective. Since $[n]$ is finite, φ is bijective by Theorem 0.4.9, so $\varphi = \sigma$ for some $\sigma \in S_n$. By Lemma 7.4.1 with $\mathbf{v}_i = \mathbf{e}_i$,

$$f(\mathbf{e}_{\sigma(1)}, \dots, \mathbf{e}_{\sigma(n)}) = \operatorname{sgn}(\sigma)\, f(\mathbf{e}_1, \dots, \mathbf{e}_n).$$

Hence only the terms of Case 2 survive in $(**)$, there is exactly one for each $\sigma \in S_n$, and

$$f(\mathbf{a}_1, \dots, \mathbf{a}_n) = \sum_{\sigma\in S_n} a_{\sigma(1)1} \cdots a_{\sigma(n)n}\, \operatorname{sgn}(\sigma)\, f(\mathbf{e}_1, \dots, \mathbf{e}_n),$$

which is the stated formula after moving the scalar $f(\mathbf{e}_1, \dots, \mathbf{e}_n)$ to the front. This proves the theorem.

■

So the three rules leave no freedom at all. If a normalized alternating form exists, it **must** be

$$\mathbf{a}_1, \dots, \mathbf{a}_n \longmapsto \sum_{\sigma\in S_n} \operatorname{sgn}(\sigma)\, a_{\sigma(1)1} \cdots a_{\sigma(n)n}.$$

Out of n^n terms, the rules killed all but $n!$. For $n = 3$ that is 6 out of 27.

7.4.4 Existence

The candidate has been found. Now we check it against the rules, forgetting where it came from.

Theorem 7.4.3. Existence: the Leibniz Formula

Define $D\colon F^n \times \cdots \times F^n \to F$ by

$$D(\mathbf{a}_1, \dots, \mathbf{a}_n) = \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma)\, a_{\sigma(1)1}\, a_{\sigma(2)2} \cdots a_{\sigma(n)n},$$

where a_{ij} is the i-th entry of $\mathbf{a}_j$. Then D is an alternating n-linear form on F^n, and $D(\mathbf{e}_1, \dots, \mathbf{e}_n) = 1$.

Idea. Linearity is visible: each product takes exactly one entry from each column. For alternation, suppose columns p and q are equal. Pair each σ with $\sigma(p\ q)$. The two permutations have opposite signs. Their products use the same entries, except that the entries taken from columns p and q trade places, and those two columns are equal. So each pair contributes $x - x = 0$.

Why pair terms instead of arguing "swapping the equal columns changes D to $-D$, so $D = -D$ and $D = 0$"? Because $D = -D$ gives $2D = 0$, and over a field with $1 + 1 = 0$, such as $\mathbb{F}_2$, that says nothing. The pairing argument never divides by 2.

Proof. Multilinearity. Fix $k \in [n]$ and fix all columns except the k-th. For a fixed σ, the product $a_{\sigma(1)1} \cdots a_{\sigma(n)n}$ contains exactly one entry of $\mathbf{a}_k$, namely $a_{\sigma(k)k}$, the $\sigma(k)$-th coordinate of $\mathbf{a}_k$. So the term for σ equals $c_\sigma\, a_{\sigma(k)k}$, where $c_\sigma = \operatorname{sgn}(\sigma) \prod_{j \neq k} a_{\sigma(j)j}$ does not depend on $\mathbf{a}_k$. The map $\mathbf{a}_k \mapsto a_{\sigma(k)k}$ is linear, since coordinates of a sum are sums of coordinates and coordinates of $c\mathbf{a}_k$ are c times the coordinates. So each term is linear in $\mathbf{a}_k$, and hence so is their sum D.

Alternation. Suppose $\mathbf{a}_p = \mathbf{a}_q$ with $p < q$, that is, $a_{ip} = a_{iq}$ for all i. Let $\tau = (p\ q)$, let $A_n = \{\sigma \in S_n : \operatorname{sgn}(\sigma) = 1\}$ be the even permutations (Definition 7.3.3), so that $S_n \setminus A_n = \{\sigma \in S_n : \operatorname{sgn}(\sigma) = -1\}$ consists of the odd ones. Since $\operatorname{sgn}(\sigma\tau) = \operatorname{sgn}(\sigma)\operatorname{sgn}(\tau) = -\operatorname{sgn}(\sigma)$ by Theorem 7.3.1 and Corollary 7.3.2, the map $\sigma \mapsto \sigma\tau$ sends A_n to $S_n \setminus A_n$. It is a bijection $A_n \to S_n \setminus A_n$, because $\rho \mapsto \rho\tau$ sends $S_n \setminus A_n$ back to A_n and $\sigma\tau\tau = \sigma$. Every $\sigma \in S_n$ has sign 1 or -1, so

$$D(\mathbf{a}_1, \dots, \mathbf{a}_n) = \sum_{\sigma \in A_n} \Big(\prod_{j=1}^{n} a_{\sigma(j)j} - \prod_{j=1}^{n} a_{\sigma\tau(j)\,j} \Big).$$

Fix $\sigma \in A_n$ and compare the two products factor by factor. For $j \neq p, q$, $\sigma\tau(j) = \sigma(j)$, so the factors agree. For $j = p$, the second product has $a_{\sigma(q)p} = a_{\sigma(q)q}$, which is the factor of the first product at $j = q$. For $j = q$, the second product has $a_{\sigma(p)q} = a_{\sigma(p)p}$, the factor of the first product at $j = p$. So the two products consist of the same n factors in a different order, and they are equal because multiplication in F is commutative. Every bracket is 0, and $D(\mathbf{a}_1, \dots, \mathbf{a}_n) = 0$.

Normalization. For $\mathbf{a}_j = \mathbf{e}_j$, a_{ij} is 1 if $i = j$ and 0 otherwise. The product $a_{\sigma(1)1} \cdots a_{\sigma(n)n}$ is therefore 1 if $\sigma(j) = j$ for all j, and 0 otherwise. Only $\sigma = \mathrm{id}$ contributes, and $\operatorname{sgn}(\mathrm{id}) = 1$. Hence $D(\mathbf{e}_1, \dots, \mathbf{e}_n) = 1$. This proves the theorem.

■

Uniqueness and existence together answer the question of the first section: the rules of signed area single out exactly one function.

7.4.5 The determinant

We now have a function of square matrices that is forced on us by three reasonable rules, and it deserves a name. It is the central notion of this chapter.

The determinant is the only way to attach a number to n vectors in F^n that is linear in each vector, vanishes when two vectors coincide, and gives 1 on the standard basis.

Definition 7.4.1. Determinant

Let $n \geq 1$ and $\mathbf{A} = (a_{ij}) \in M_n(F)$. The **determinant** of $\mathbf{A}$ is

$$\det \mathbf{A} := \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma)\, a_{\sigma(1)1}\, a_{\sigma(2)2} \cdots a_{\sigma(n)n} \ \in F.$$

For columns $\mathbf{a}_1, \ldots, \mathbf{a}_n \in F^n$ we also write $\det(\mathbf{a}_1, \ldots, \mathbf{a}_n)$ for the determinant of the matrix $\begin{pmatrix} \mathbf{a}_1 & \cdots & \mathbf{a}_n \end{pmatrix}$.

In words: run over **all** $n!$ permutations σ. For each, pick from column j the entry in row $\sigma(j)$; this picks one entry in every column and, because σ is a bijection, one entry in every row. Multiply the n picked entries, attach the sign of σ, and add up. The determinant is defined **only for square matrices**, and its value lies in the field F of the entries.

Well-definedness. The formula is a finite sum, so it always gives a number. What the name "**the** determinant" really claims is that nothing else satisfies the three rules, and that is Theorem 7.4.2; Theorem 7.4.3 shows det does satisfy them. The sign $\operatorname{sgn}(\sigma)$ is fixed by counting inversions (Definition 7.3.2), and by Corollary 7.3.3 any factorization of σ into transpositions gives the same sign, so either method may be used to evaluate the formula.

Examples.

- $n = 1$. $S_1 = \{\mathrm{id}\}$, so $\det(a) = a$. A degenerate case, but it is the base of every induction on size.
- $n = 2$. $S_2 = \{\mathrm{id}, (1\ 2)\}$, with signs 1 and -1. The identity picks $a_{11}a_{22}$, and $(1\ 2)$ picks $a_{21}a_{12}$. So
$$\det \begin{pmatrix} a & b \\ c & d \end{pmatrix} = ad - cb,$$
the value derived from the rules in Example ($ad - bc$ from the Rules), and the number whose nonvanishing decides invertibility in Theorem 0.9.8.
- $n = 3$. The six permutations of S_3 and their signs (a 3-cycle has sign 1, a transposition -1) give
$$\begin{aligned} \det \mathbf{A} = {} & a_{11}a_{22}a_{33} - a_{21}a_{12}a_{33} - a_{31}a_{22}a_{13} - a_{11}a_{32}a_{23} \\ & + a_{21}a_{32}a_{13} + a_{31}a_{12}a_{23}, \end{aligned}$$
the terms coming from $\mathrm{id}, (1\ 2), (1\ 3), (2\ 3), (1\ 2\ 3), (1\ 3\ 2)$ in this order. For instance $\sigma = (1\ 2\ 3)$ has $\sigma(1) = 2$, $\sigma(2) = 3$, $\sigma(3) = 1$, and picks $a_{21}a_{32}a_{13}$.

- **A numerical 3×3.** For $\mathbf{A}=\begin{pmatrix}2&1&3\\0&4&1\\5&2&1\end{pmatrix}$, the six terms in the same order are 8, -0, -60, -4, 0 and 5, so $\det\mathbf{A}=8-60-4+5=-51$.
- **Identity and zero.** $\det\mathbf{I}_n=1$ by the normalization in Theorem 7.4.3, and $\det\mathbf{O}=0$, since every product contains a zero factor.

Nonexample by minimal change. Delete the sign. The **permanent**

$$\operatorname{per}\mathbf{A}=\sum_{\sigma}a_{\sigma(1)1}\cdots a_{\sigma(n)n}$$

is still n-linear in the columns, by the same argument as in Theorem 7.4.3, and still gives 1 at $\mathbf{I}_n$. But it is not alternating: $\operatorname{per}\begin{pmatrix}1&1\\1&1\end{pmatrix}=1+1=2$, which is nonzero over $\mathbb{R}$ although the two columns are equal. The sign is exactly what makes the pairing argument cancel.

Why this definition. Each rule has a job. Without normalization, the uniqueness theorem only pins the form down up to the scalar $f(\mathbf{e}_1,\dots,\mathbf{e}_n)$. Without alternation, the permanent and many other functions qualify. Multilinearity is what let us expand at all. The choice of **columns** rather than rows is a convention: we show below that the transpose has the same determinant, so rows would give the same function. We have avoided "Sarrus's rule" on purpose. It is a picture of the six-term formula for $n=3$ and nothing more, and for $n=4$ its diagonal pattern would produce 8 products while the Leibniz formula has 24.

Warning. **The determinant is not linear.** It is linear in **each column separately**, not in the matrix. For $n=2$, $\det\mathbf{I}_2=1$ but $\det(\mathbf{I}_2+\mathbf{I}_2)=\det(2\mathbf{I}_2)=4\neq 1+1$. In general each product in the Leibniz formula has n factors, and scaling $\mathbf{A}$ scales every factor, so

$$\det(c\mathbf{A})=c^n\det\mathbf{A},\qquad\text{in particular}\qquad\det(-\mathbf{A})=(-1)^n\det\mathbf{A}.$$

So $\det(-\mathbf{A})=\det\mathbf{A}$ for 2×2 matrices, not $-\det\mathbf{A}$.

Quick check. Which permutations give a nonzero term in the determinant of $\begin{pmatrix}0&0&1\\0&2&0\\3&0&0\end{pmatrix}$, and what is the determinant?

Solution. Column 1 has its only nonzero entry in row 3, column 2 in row 2, and column 3 in row 1. A nonzero term needs $a_{\sigma(j)j}\neq 0$ for every j, so $\sigma(1)=3$, $\sigma(2)=2$, $\sigma(3)=1$, that is, $\sigma=(1\ 3)$, of sign -1. The determinant is $-3\cdot 2\cdot 1=-6$.

The payoff for the forms of the second section is immediate. There we saw that the alternating n-linear forms on F^n form a vector space (Theorem 7.2.3) and promised its dimension.

Corollary 7.4.4. Alternating n-Forms on F^n Form a Line

The vector space of alternating n-linear forms on F^n has dimension 1, with basis (det).

Explicitly, every alternating n-linear form f on F^n satisfies

$$f(\mathbf{a}_1, \dots, \mathbf{a}_n) = f(\mathbf{e}_1, \dots, \mathbf{e}_n)\, \det(\mathbf{a}_1, \dots, \mathbf{a}_n).$$

Proof. By Theorem 7.4.3, det is an element of this space, and it is not the zero form since $\det(\mathbf{e}_1, \dots, \mathbf{e}_n) = 1$. So the list (det) is linearly independent. By Theorem 7.4.2 and Definition 7.4.1, every f in the space satisfies the displayed formula, so $f = f(\mathbf{e}_1, \dots, \mathbf{e}_n) \det$ lies in the span of det. Hence (det) is a basis, and the dimension is 1. ∎

This corollary is the most useful form of uniqueness. To prove that some function g of n columns is a multiple of the determinant, check that g is multilinear and alternating, and compute $g(\mathbf{e}_1, \dots, \mathbf{e}_n)$. The multiplicativity of the determinant in the next section is proved exactly this way.

7.4.6 Transposes and triangular matrices

The definition read the matrix column by column. Reading it row by row gives nothing new.

Theorem 7.4.5. Determinant of the Transpose

For every $\mathbf{A} \in M_n(F)$, $\det \mathbf{A}^\top = \det \mathbf{A}$. Consequently, $\det \mathbf{A}$ is also an alternating n-linear function of the **rows** of $\mathbf{A}$, with value 1 at $\mathbf{I}_n$.

Idea. A term of $\det \mathbf{A}^\top$ picks the entries $a_{j\sigma(j)}$, one in each row. The same n positions are also "one in each column": position $(j, \sigma(j))$ is $(\sigma^{-1}(i), i)$ with $i = \sigma(j)$. So the term is the term of $\det \mathbf{A}$ for σ^{-1}, and σ^{-1} has the same sign as σ.

Proof. The (i,j)-entry of $\mathbf{A}^\top$ is a_{ji} (Definition 0.9.10), so by Definition 7.4.1

$$\det \mathbf{A}^\top = \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma)\, a_{1\sigma(1)}\, a_{2\sigma(2)} \cdots a_{n\sigma(n)}.$$

Fix σ. As j runs through $[n]$, $i = \sigma(j)$ runs through $[n]$ exactly once, and $j = \sigma^{-1}(i)$. Since multiplication in F is commutative, we may reorder the factors by i:

$$\prod_{j=1}^{n} a_{j\sigma(j)} = \prod_{i=1}^{n} a_{\sigma^{-1}(i)\, i}.$$

Next, $\operatorname{sgn}(\sigma)\operatorname{sgn}(\sigma^{-1}) = \operatorname{sgn}(\sigma\sigma^{-1}) = \operatorname{sgn}(\mathrm{id}) = 1$ by Theorem 7.3.1, and both signs lie in $\{1, -1\}$, so $\operatorname{sgn}(\sigma^{-1}) = \operatorname{sgn}(\sigma)$. Therefore

$$\det \mathbf{A}^\top = \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma^{-1}) \prod_{i=1}^{n} a_{\sigma^{-1}(i)\, i} = \sum_{\rho \in S_n} \operatorname{sgn}(\rho) \prod_{i=1}^{n} a_{\rho(i)\, i} = \det \mathbf{A},$$

where the second equality substitutes $\rho = \sigma^{-1}$; this is allowed because $\sigma \mapsto \sigma^{-1}$ is a bijection of S_n onto itself, being its own inverse.

For the second statement, let $\mathbf{r}_1, \dots, \mathbf{r}_n$ be the rows of $\mathbf{A}$. The columns of $\mathbf{A}^\top$ are $\mathbf{r}_1^\top, \dots, \mathbf{r}_n^\top$, so $\det \mathbf{A} = \det \mathbf{A}^\top = \det(\mathbf{r}_1^\top, \dots, \mathbf{r}_n^\top)$. The map $\mathbf{r} \mapsto \mathbf{r}^\top$ is linear (Theorem 0.9.5), so by Theorem 7.4.3 the function $(\mathbf{r}_1, \dots, \mathbf{r}_n) \mapsto \det(\mathbf{r}_1^\top, \dots, \mathbf{r}_n^\top)$ is linear in each $\mathbf{r}_i$. If $\mathbf{r}_p = \mathbf{r}_q$ for some $p \neq q$, then $\mathbf{r}_p^\top = \mathbf{r}_q^\top$, and the value is 0. Finally $\mathbf{I}_n^\top = \mathbf{I}_n$. This proves the theorem.

■

So every statement about columns has a twin about rows. In particular Theorem 7.2.1 applies to rows: swapping two rows changes the sign of det, and adding a multiple of one row to another does not change it. The next section turns this into a method of computation.

Before that, one class of matrices whose determinant can be read off directly. Recall that $\mathbf{A}$ is upper triangular if $a_{ij} = 0$ whenever $i > j$ (Definition 0.9.5).

Theorem 7.4.6. Determinant of a Triangular Matrix

If $\mathbf{A} \in M_n(F)$ is upper triangular or lower triangular, then

$$\det \mathbf{A} = a_{11}\, a_{22} \cdots a_{nn}.$$

Idea. Which permutations can pick only entries on or above the diagonal? Picking $a_{\sigma(j)j}$ with $\sigma(j) \leq j$ for every j forces $\sigma = \mathrm{id}$: the numbers $\sigma(1), \dots, \sigma(n)$ are $1, \dots, n$ in some order, so they have the same sum, and none of them is allowed to exceed its partner.

Proof. Suppose $\mathbf{A}$ is upper triangular, and let $\sigma \in S_n$ with $\sigma \neq \mathrm{id}$. We claim $\sigma(j) > j$ for some j. Otherwise $\sigma(j) \leq j$ for all j. Since σ is a bijection of $[n]$, $\sum_{j=1}^n \sigma(j) = \sum_{j=1}^n j$, so $\sum_{j=1}^n (j - \sigma(j)) = 0$ is a sum of nonnegative integers equal to 0. Hence $\sigma(j) = j$ for every j, contradicting $\sigma \neq \mathrm{id}$. For such j, $a_{\sigma(j)j} = 0$ by upper triangularity, so the term of σ is 0. Only the identity remains, and $\det \mathbf{A} = \operatorname{sgn}(\mathrm{id})\, a_{11} \cdots a_{nn} = a_{11} \cdots a_{nn}$.

If $\mathbf{A}$ is lower triangular, then $\mathbf{A}^\top$ is upper triangular with the same diagonal, and $\det \mathbf{A} = \det \mathbf{A}^\top = a_{11} \cdots a_{nn}$ by Theorem 7.4.5. This proves the theorem.

■

For example, $\det \begin{pmatrix} 2 & 7 & -1 \\ 0 & 3 & 5 \\ 0 & 0 & -4 \end{pmatrix} = 2 \cdot 3 \cdot (-4) = -24$: six terms, five of which vanish. Diagonal matrices are a special case, $\det \operatorname{diag}(d_1, \dots, d_n) = d_1 \cdots d_n$. Together with Chapter 3's result that an upper triangular matrix is invertible exactly when its diagonal entries are nonzero (Lemma 3.6.2), this already shows that for upper triangular matrices, $\mathbf{A}$ is invertible if and only if $\det \mathbf{A} \neq 0$. The next section proves this for all square matrices.

7.4.7 Determinants over a commutative ring

The characteristic polynomial at the end of this chapter is the determinant of a matrix whose entries are **polynomials**, such as $\begin{pmatrix} x-1 & -2 \\ -3 & x-4 \end{pmatrix}$. Polynomials form no field, since x has no inverse in $F[x]$. So we record exactly what this section used.

Remark. **Commutative rings.** A **commutative ring** is a set R with an addition and a multiplication satisfying all the axioms of a field (Definition 0.7.1) except, possibly, the existence of multiplicative inverses (F8). The integers $\mathbb{Z}$ and the polynomials $F[x]$ are examples (Theorem 0.8.1). For $\mathbf{A} \in M_n(R)$, the Leibniz formula of Definition 7.4.1 still defines $\det \mathbf{A} \in R$, reading $\operatorname{sgn}(\sigma)\, r$ as r or $-r$.

Every proof in this section uses only addition, subtraction, and multiplication, together with commutativity, and never a multiplicative inverse. So the following hold verbatim over any commutative ring R, with "linear in each argument" meaning additive and compatible with multiplication by elements of R: Lemma 7.4.1, Theorem 7.4.2, Theorem 7.4.3, Theorem 7.4.5, and Theorem 7.4.6. The only outside input is the swap rule of Theorem 7.2.1, and it too is division-free: expanding $0 = f(\dots, u+v, \dots, u+v, \dots)$ by additivity leaves $f(\dots, u, \dots, v, \dots) + f(\dots, v, \dots, u, \dots) = 0$. The parts of Theorem 7.2.1 about dependent arguments are not used; they involve solving for one vector in terms of others, which needs division.

In the characteristic-polynomial section we will apply these results with $R = F[x]$. The next two sections flag, in the same way, which of their proofs survive the passage to a ring.

7.4.8 Exercises

7.4.8.1 A. Check your understanding

Exercise. A1

(a) State the three properties that characterize $\det\colon M_n(F) \to F$, and write down the Leibniz formula.

(b) How many terms does the Leibniz formula have for $n = 4$? How many of the corresponding permutations have sign 1?

(c) True or false: $\det(\mathbf{A}+\mathbf{B}) = \det \mathbf{A} + \det \mathbf{B}$ for all $\mathbf{A}, \mathbf{B} \in M_2(\mathbb{R})$. Justify your answer.

(d) True or false: $\det(-\mathbf{A}) = -\det \mathbf{A}$ for all $\mathbf{A} \in M_3(\mathbb{R})$. Justify your answer.

(e) In the proof of Theorem 7.4.2, why does a term with a noninjective φ vanish?

(f) Explain why the existence proof pairs σ with $\sigma(p\ q)$ instead of arguing "$D = -D$, so $D = 0$".

Solution.

(a) det is n-linear in the columns, alternating (zero whenever two columns are equal), and $\det \mathbf{I}_n = 1$. By Theorem 7.4.2 and Theorem 7.4.3 it is the only such function, and $\det \mathbf{A} = \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma)\, a_{\sigma(1)1} \cdots a_{\sigma(n)n}$.

(b) One term per permutation: $4! = 24$. In the proof of Theorem 7.4.3, $\sigma \mapsto \sigma(1\ 2)$ is

a bijection from the permutations of sign 1 to those of sign -1, so each set has 12 elements.

(c) False. For $\mathbf{A} = \mathbf{B} = \mathbf{I}_2$, $\det(\mathbf{A}+\mathbf{B}) = \det(2\mathbf{I}_2) = 4$ but $\det\mathbf{A} + \det\mathbf{B} = 2$.

(d) True. Each Leibniz product has 3 factors, each multiplied by -1, so $\det(-\mathbf{A}) = (-1)^3 \det\mathbf{A} = -\det\mathbf{A}$. (For $n = 2$ it is false.)

(e) If $\varphi(p) = \varphi(q)$ with $p \neq q$, the form is evaluated with the equal arguments $\mathbf{e}_{\varphi(p)}$ in positions p and q, and an alternating form vanishes there.

(f) $D = -D$ only gives $2D = 0$. Over a field in which $2 = 0$, such as $\mathbb{F}_2$, this does not imply $D = 0$. The pairing shows that the terms cancel in pairs, which works over every field.

7.4.8.2 B. Practice

Exercise. B1: Six-term determinants

Using the Leibniz formula, listing all six terms, compute the determinants of

$$\mathbf{A} = \begin{pmatrix} 3 & 1 & 2 \\ 1 & 0 & 4 \\ 2 & 5 & 1 \end{pmatrix}, \qquad \mathbf{B} = \begin{pmatrix} 1 & -1 & 2 \\ 0 & 3 & 1 \\ 4 & 2 & -2 \end{pmatrix} \in M_3(\mathbb{R}).$$

Then compute $\det\mathbf{A}^\top$ in the same way, and hence verify Theorem 7.4.5 for this $\mathbf{A}$.

Solution. We use the order $\mathrm{id}, (1\ 2), (1\ 3), (2\ 3), (1\ 2\ 3), (1\ 3\ 2)$ and the six-term formula from the examples after Definition 7.4.1:

$$a_{11}a_{22}a_{33},\ -a_{21}a_{12}a_{33},\ -a_{31}a_{22}a_{13},\ -a_{11}a_{32}a_{23},\ a_{21}a_{32}a_{13},\ a_{31}a_{12}a_{23}.$$

For $\mathbf{A}$: $3\cdot0\cdot1 = 0$; $-1\cdot1\cdot1 = -1$; $-2\cdot0\cdot2 = 0$; $-3\cdot5\cdot4 = -60$; $1\cdot5\cdot2 = 10$; $2\cdot1\cdot4 = 8$. Hence $\det\mathbf{A} = 0 - 1 + 0 - 60 + 10 + 8 = -43$.

For $\mathbf{B}$: $1\cdot3\cdot(-2) = -6$; $-0\cdot(-1)\cdot(-2) = 0$; $-4\cdot3\cdot2 = -24$; $-1\cdot2\cdot1 = -2$; $0\cdot2\cdot2 = 0$; $4\cdot(-1)\cdot1 = -4$. Hence $\det\mathbf{B} = -6 + 0 - 24 - 2 + 0 - 4 = -36$.

For $\mathbf{A}^\top = \begin{pmatrix} 3 & 1 & 2 \\ 1 & 0 & 5 \\ 2 & 4 & 1 \end{pmatrix}$: $3\cdot0\cdot1 = 0$; $-1\cdot1\cdot1 = -1$; $-2\cdot0\cdot2 = 0$; $-3\cdot4\cdot5 = -60$; $1\cdot4\cdot2 = 8$; $2\cdot1\cdot5 = 10$. The sum is $-43 = \det\mathbf{A}$, as Theorem 7.4.5 predicts; the six terms are the same numbers, with the two 3-cycles exchanged, because $(1\ 2\ 3)^{-1} = (1\ 3\ 2)$.

Exercise. B2: The determinant of $\mathbf{P}_\sigma$

Let $\sigma \in S_n$ and let $\mathbf{P}_\sigma = \begin{pmatrix} \mathbf{e}_{\sigma(1)} & \cdots & \mathbf{e}_{\sigma(n)} \end{pmatrix}$ be its permutation matrix (Definition 3.4.2).

(a) Prove that $\det\mathbf{P}_\sigma = \mathrm{sgn}(\sigma)$.

(b) Hence compute $\det \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$ and $\det \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}$.

Solution.

(a) By Theorem 7.4.3, det is an alternating n-linear form on F^n. By Lemma 7.4.1 with $\mathbf{v}_i = \mathbf{e}_i$,

$$\det \mathbf{P}_\sigma = \det(\mathbf{e}_{\sigma(1)}, \dots, \mathbf{e}_{\sigma(n)}) = \operatorname{sgn}(\sigma) \det(\mathbf{e}_1, \dots, \mathbf{e}_n) = \operatorname{sgn}(\sigma) \det \mathbf{I}_n = \operatorname{sgn}(\sigma).$$

(b) The first matrix has columns $\mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_1$, so it is $\mathbf{P}_\sigma$ with $\sigma(1) = 2$, $\sigma(2) = 3$, $\sigma(3) = 1$, that is, $\sigma = (1\ 2\ 3)$. A 3-cycle is $(1\ 3)(1\ 2)$ (Example (Computing with cycles)), a product of two transpositions, so its sign is 1 and the determinant is 1. The second matrix has columns $\mathbf{e}_2, \mathbf{e}_1, \mathbf{e}_4, \mathbf{e}_3$, so it is $\mathbf{P}_\sigma$ with $\sigma = (1\ 2)(3\ 4)$, of sign $(-1)^2 = 1$ by Theorem 7.3.1 and Corollary 7.3.2. Its determinant is 1.

Exercise. B3: A zero row

Let $\mathbf{A} \in M_n(F)$ have a zero row. Prove directly from Definition 7.4.1, without using Theorem 7.4.5, that $\det \mathbf{A} = 0$. Then prove the same for a zero column.

Solution. Suppose row i of $\mathbf{A}$ is zero, and let $\sigma \in S_n$. Since σ is surjective, there is j with $\sigma(j) = i$. The factor $a_{\sigma(j)j} = a_{ij}$ of the term for σ is 0, so the term is 0. This holds for every σ, so $\det \mathbf{A} = 0$.

Suppose column j is zero. For every σ, the term for σ contains the factor $a_{\sigma(j)j} = 0$, so again every term vanishes and $\det \mathbf{A} = 0$.

7.4.8.3 C. Going deeper

Exercise. C1: Too many zeros

Let $\mathbf{A} \in M_n(F)$.

(a) Prove that if more than $n^2 - n$ entries of $\mathbf{A}$ are zero, then $\det \mathbf{A} = 0$.

(b) Show that the bound is sharp: give, for each n, a matrix with exactly $n^2 - n$ zero entries and nonzero determinant.

Solution.

(a) If more than $n^2 - n$ entries are zero, then fewer than n entries are nonzero. For $\sigma \in S_n$, the term for σ is a product of the n entries in positions $(\sigma(1), 1), \dots, (\sigma(n), n)$. These positions are distinct, since their column indices are distinct. They cannot all hold nonzero entries, because there are fewer than n such entries. So at least one factor is 0, and the term is 0. Hence $\det \mathbf{A} = 0$.

(b) $\mathbf{I}_n$ has exactly $n^2 - n$ zero entries, and $\det \mathbf{I}_n = 1 \neq 0$.

Exercise. C2: Counting the cost

A computer performs about 10^9 multiplications per second.

(a) Show that evaluating the Leibniz formula term by term for an $n \times n$ matrix takes $n!\,(n-1)$ multiplications. Estimate the time this takes for $n = 20$.

(b) Gaussian elimination (Chapter 3) brings $\mathbf{A}$ to upper triangular form. Suppose no row swaps are needed. At step k (for $k = 1, \dots, n-1$), each of the $n-k$ rows below the pivot requires one division to find its multiplier and $n-k$ multiplications to update its entries in columns $k+1, \dots, n$. Show that the total number of multiplications and divisions is $\frac{(n-1)n(n+1)}{3}$, and evaluate it for $n = 20$.

In the next section we will see that the determinant can be read off from this triangular form with $n-1$ further multiplications.

Solution.

(a) There is one term for each $\sigma \in S_n$, that is, $n!$ terms, and each is a product of n entries, which takes $n-1$ multiplications. For $n = 20$, $20! \cdot 19 = 46\,225\,138\,155\,356\,160\,000 \approx 4.6 \times 10^{19}$. At 10^9 per second this takes about 4.6×10^{10} seconds, roughly 1 460 years.

(b) Step k costs $(n-k)(1+(n-k))$ operations. With $m = n-k$ running from 1 to $n-1$, the total is

$$\begin{aligned}\sum_{m=1}^{n-1} m(m+1) &= \sum_{m=1}^{n-1} m^2 + \sum_{m=1}^{n-1} m \\ &= \frac{(n-1)n(2n-1)}{6} + \frac{(n-1)n}{2} \\ &= \frac{(n-1)n\big((2n-1)+3\big)}{6} = \frac{(n-1)n(n+1)}{3}.\end{aligned}$$

For $n = 20$ this is $\frac{19 \cdot 20 \cdot 21}{3} = 2\,660$, a few microseconds. The Leibniz formula is a definition and a proof tool, not an algorithm.

Exercise. C3: The permanent, and a change of field

For $\mathbf{A} \in M_n(F)$ define $\operatorname{per} \mathbf{A} = \sum_{\sigma \in S_n} a_{\sigma(1)1} \cdots a_{\sigma(n)n}$.

(a) Prove that per is n-linear in the columns and $\operatorname{per} \mathbf{I}_n = 1$.

(b) Let $n \geq 2$. Prove that over $\mathbb{R}$, per is **not** alternating.

(c) Prove that over $\mathbb{F}_2$, $\operatorname{per} \mathbf{A} = \det \mathbf{A}$ for every $\mathbf{A} \in M_n(\mathbb{F}_2)$. Explain why this does not contradict (b) and Theorem 7.4.2.

Solution.

(a) The proof of multilinearity in Theorem 7.4.3 never used the signs: for fixed k, the term

for σ is a constant times $a_{\sigma(k)k}$, which is linear in the k-th column. The normalization argument there also applies: only $\sigma = \mathrm{id}$ contributes at $\mathbf{I}_n$, with product 1.

(b) Let $\mathbf{J} \in M_n(\mathbb{R})$ have every entry equal to 1. Its first two columns are equal, but every term of $\operatorname{per}\mathbf{J}$ is 1, so $\operatorname{per}\mathbf{J} = n! \neq 0$ in $\mathbb{R}$.

(c) In $\mathbb{F}_2$, $-1 = 1$, so $\operatorname{sgn}(\sigma)\cdot r = r$ for every r, and the two formulas agree term by term. There is no contradiction: (b) is a statement over $\mathbb{R}$, where $n! \neq 0$. Over $\mathbb{F}_2$, $\operatorname{per}\mathbf{J} = n!\cdot 1 = 0$ for $n \geq 2$, and per = det is alternating, as Theorem 7.4.3 guarantees. Uniqueness is about a fixed field, and over each field there is exactly one normalized alternating form.

7.5 Row Operations and Multiplicativity

The previous section defined the determinant by the Leibniz formula and proved it is the only normalized alternating form. The formula settles every theoretical question in principle, but with $n!$ terms it is useless for computing, and it hides the two facts that make determinants worth having. The first is that $\det\mathbf{A} \neq 0$ exactly when $\mathbf{A}$ is invertible. The second is that $\det(\mathbf{AB}) = \det\mathbf{A}\det\mathbf{B}$. This section proves both. Along the way we get a fast algorithm, and we extend the determinant from matrices to operators.

7.5.1 Row operations

Row reduction is the workhorse of Chapter 3, and the three rules of a determinant say precisely how it reacts to each elementary row operation. Everything follows from Theorem 7.4.5, which lets us treat det as a function of the rows.

Theorem 7.5.1. Row Operations and the Determinant

Let $\mathbf{A} \in M_n(F)$ and let $\mathbf{B}$ be obtained from $\mathbf{A}$ by one operation.

(a) **(Swap)** If $\mathbf{B}$ is obtained by $R_i \leftrightarrow R_j$ with $i \neq j$, then $\det\mathbf{B} = -\det\mathbf{A}$.

(b) **(Scale)** If $\mathbf{B}$ is obtained by $R_i \to cR_i$ with $c \in F$, then $\det\mathbf{B} = c\det\mathbf{A}$.

(c) **(Replace)** If $\mathbf{B}$ is obtained by $R_i \to R_i + cR_j$ with $i \neq j$ and $c \in F$, then $\det\mathbf{B} = \det\mathbf{A}$.

(d) The same three statements hold for the corresponding operations on **columns**.

(e) For the elementary matrices of Definition 3.4.1,

$$\det\mathbf{P}_{ij} = -1, \qquad \det\mathbf{D}_i(c) = c, \qquad \det(\mathbf{I}_n + c\mathbf{E}_{ij}) = 1 \quad (i \neq j),$$

and $\det(\mathbf{EA}) = \det\mathbf{E}\det\mathbf{A}$ for every elementary matrix $\mathbf{E} \in M_n(F)$.

Proof. By Theorem 7.4.5, det is an alternating n-linear function of the rows $\mathbf{r}_1, \dots, \mathbf{r}_n$ of $\mathbf{A}$. We write $\det(\mathbf{r}_1, \dots, \mathbf{r}_n)$ for it in this proof.

(b) This is linearity in row i.

(c) By linearity in row i,

$$\begin{aligned}\det\mathbf{B} &= \det(\dots, \mathbf{r}_i + c\mathbf{r}_j, \dots, \mathbf{r}_j, \dots)\\ &= \det(\dots, \mathbf{r}_i, \dots, \mathbf{r}_j, \dots) + c\det(\dots, \mathbf{r}_j, \dots, \mathbf{r}_j, \dots),\end{aligned}$$

where the displayed arguments sit in positions i and j. The last determinant has two equal rows, so it is 0. Hence $\det \mathbf{B} = \det \mathbf{A}$.

(d) This is the swap rule of Theorem 7.2.1, applied to the rows.

(e) By Theorem 7.4.3, det is an alternating n-linear function of the columns, and the arguments for (a), (b), (c) apply word for word.

(f) The matrices $\mathbf{P}_{ij}$, $\mathbf{D}_i(c)$, and $\mathbf{I}_n + c\mathbf{E}_{ij}$ are obtained from $\mathbf{I}_n$ by a swap, a scaling, and a replacement respectively. By (a), (b), (c), and $\det \mathbf{I}_n = 1$, their determinants are -1, c, and 1. Now let $\mathbf{E} = \rho(\mathbf{I}_n)$ be elementary. By Theorem 3.4.1, $\mathbf{EA} = \rho(\mathbf{A})$. By (a), (b), or (c), $\det \rho(\mathbf{A})$ is $\det \mathbf{A}$ times -1, c, or 1 respectively, and this factor is $\det \mathbf{E}$. Hence $\det(\mathbf{EA}) = \det \mathbf{E} \det \mathbf{A}$. This proves the theorem.

■

Read part (e) as a bookkeeping rule. A swap flips the sign, a scaling by c multiplies by c, and a replacement changes nothing. Replacement, the operation used most in elimination, is free.

Warning. **A combined move is not a replacement.** The move $R_2 \to 3R_2 - R_1$ looks harmless but multiplies the determinant by 3: it is the scaling $R_2 \to 3R_2$ followed by the replacement $R_2 \to R_2 - R_1$. For $\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 1 & 3 \end{pmatrix}$, with $\det \mathbf{A} = 1$, it produces $\begin{pmatrix} 1 & 2 \\ 2 & 7 \end{pmatrix}$, whose determinant is 3. Keep replacements in the form "row i plus a multiple of another row", with coefficient 1 on row i.

Since elimination brings any square matrix to an upper triangular one, and triangular determinants are products of diagonal entries (Theorem 7.4.6), we have an algorithm.

Example. A 4×4 Determinant by Elimination

Compute $\det \mathbf{A}$ for

$$\mathbf{A} = \begin{pmatrix} 0 & 2 & 4 & 2 \\ 1 & 1 & 0 & 3 \\ 2 & 3 & 1 & 5 \\ -1 & 2 & 3 & 1 \end{pmatrix} \in M_4(\mathbb{R}).$$

Solution. We record each operation and its effect, using Theorem 7.5.1.

Step 1. The $(1,1)$-entry is 0, so swap: $R_1 \leftrightarrow R_2$. This multiplies the determinant by -1:

$$\det \mathbf{A} = -\det \begin{pmatrix} 1 & 1 & 0 & 3 \\ 0 & 2 & 4 & 2 \\ 2 & 3 & 1 & 5 \\ -1 & 2 & 3 & 1 \end{pmatrix}.$$

Step 2. Clear column 1 with the replacements $R_3 \to R_3 - 2R_1$ and $R_4 \to R_4 + R_1$, which

change nothing. Row 2 has a common factor 2; pulling it out is linearity in row 2:

$$\det \mathbf{A} = -\det \begin{pmatrix} 1 & 1 & 0 & 3 \\ 0 & 2 & 4 & 2 \\ 0 & 1 & 1 & -1 \\ 0 & 3 & 3 & 4 \end{pmatrix} = -2 \det \begin{pmatrix} 1 & 1 & 0 & 3 \\ 0 & 1 & 2 & 1 \\ 0 & 1 & 1 & -1 \\ 0 & 3 & 3 & 4 \end{pmatrix}.$$

Step 3. Clear column 2: $R_3 \to R_3 - R_2$ and $R_4 \to R_4 - 3R_2$. Then clear column 3: $R_4 \to R_4 - 3R_3$. None of these changes the determinant:

$$\det \mathbf{A} = -2 \det \begin{pmatrix} 1 & 1 & 0 & 3 \\ 0 & 1 & 2 & 1 \\ 0 & 0 & -1 & -2 \\ 0 & 0 & -3 & 1 \end{pmatrix} = -2 \det \begin{pmatrix} 1 & 1 & 0 & 3 \\ 0 & 1 & 2 & 1 \\ 0 & 0 & -1 & -2 \\ 0 & 0 & 0 & 7 \end{pmatrix}.$$

Step 4. The last matrix is upper triangular, so by Theorem 7.4.6 its determinant is $1 \cdot 1 \cdot (-1) \cdot 7 = -7$. Hence $\det \mathbf{A} = -2 \cdot (-7) = 14$.

Elimination costs roughly $n^3/3$ arithmetic operations, and reading off the triangular determinant takes only $n - 1$ more multiplications, against $n!\,(n-1)$ multiplications for the Leibniz formula (Exercise (C2: Counting the cost)). For $n = 20$ that is a few thousand operations against about 4.6×10^{19}. Every practical determinant computation is some form of elimination.

7.5.2 Determinant and invertibility

The same bookkeeping answers the invertibility question. Row reduction multiplies the determinant by nonzero factors only, so it cannot turn a nonzero determinant into zero or back. And the end of the reduction is either $\mathbf{I}_n$ or a matrix with a zero row.

Theorem 7.5.2. Invertible if and only if Non-Zero Determinant

A matrix $\mathbf{A} \in M_n(F)$ is invertible if and only if $\det \mathbf{A} \neq 0$.

Proof. Let $\mathbf{R}$ be the reduced row echelon form of $\mathbf{A}$ (Theorem 3.2.3, Theorem 3.3.4). By Definition 3.2.2 and Theorem 3.4.1, there are elementary matrices $\mathbf{E}_1, \ldots, \mathbf{E}_k$ (possibly $k = 0$) with $\mathbf{R} = \mathbf{E}_k \cdots \mathbf{E}_1 \mathbf{A}$. Applying Theorem 7.5.1 (e) k times,

$$\det \mathbf{R} = \det \mathbf{E}_k \cdots \det \mathbf{E}_1 \det \mathbf{A}.$$

Each $\det \mathbf{E}_m$ is -1, 1, or some $c \neq 0$, since a scaling operation requires $c \neq 0$. A product of nonzero elements of a field is nonzero (Theorem 0.7.1), so $\det \mathbf{R} = 0$ if and only if $\det \mathbf{A} = 0$.

By Lemma 3.4.3, either $\mathbf{R} = \mathbf{I}_n$, and then $\det \mathbf{R} = 1 \neq 0$; or the last row of $\mathbf{R}$ is zero. In the second case, scaling that row by 0 leaves $\mathbf{R}$ unchanged, so $\det \mathbf{R} = 0 \cdot \det \mathbf{R} = 0$ by Theorem 7.5.1 (b). Hence $\det \mathbf{A} \neq 0$ if and only if $\mathbf{R} = \mathbf{I}_n$, which by Theorem 3.4.4 holds if and only if $\mathbf{A}$ is invertible.

■

This is the promised addition to the list of equivalent conditions. We collect it with the conditions from Chapter 2 about the map $T_{\mathbf{A}}\colon \mathbf{x} \mapsto \mathbf{A}\mathbf{x}$ (Theorem 2.5.7), and with two conditions about rows that the determinant makes easy.

Theorem 7.5.3. Invertible Matrix Theorem, with Determinants

Let $\mathbf{A} \in M_n(F)$ and $T_{\mathbf{A}}\colon F^n \to F^n$, $\mathbf{x} \mapsto \mathbf{A}\mathbf{x}$. The following are equivalent, and each is equivalent to each of the eight conditions of Theorem 3.4.4.

(a) $\mathbf{A}$ is invertible.
(b) $T_{\mathbf{A}}$ is injective.
(c) $T_{\mathbf{A}}$ is surjective.
(d) $\operatorname{rank}\mathbf{A} = n$.
(e) $\mathbf{A}^\top$ is invertible.
(f) The rows of $\mathbf{A}$ are linearly independent.
(g) $\det\mathbf{A} \neq 0$.

Proof. Item (a) is item (a) of Theorem 3.4.4, so it suffices to connect each new item with that theorem.

(a) $\Leftrightarrow$ (g). This is Theorem 7.5.2.

(a) $\Leftrightarrow$ (b). By Theorem 2.5.7 ((b) $\Leftrightarrow$ (c)) with $V = F^n$, $T_{\mathbf{A}}$ is injective if and only if $\ker T_{\mathbf{A}} = \{\mathbf{x} : \mathbf{A}\mathbf{x} = \mathbf{0}\}$ is $\{\mathbf{0}\}$. That is item (b) of Theorem 3.4.4.

(b) $\Leftrightarrow$ (c). This is Theorem 2.5.7 ((b) $\Leftrightarrow$ (d)), since F^n is finite-dimensional.

(c) $\Leftrightarrow$ (d). By Theorem 0.9.3, $\operatorname{im} T_{\mathbf{A}}$ is the set of combinations of the columns of $\mathbf{A}$, that is, $\operatorname{col}(\mathbf{A})$. So $T_{\mathbf{A}}$ is surjective if and only if $\operatorname{col}(\mathbf{A}) = F^n$. Since $\operatorname{col}(\mathbf{A})$ is a subspace of F^n, this holds if and only if $\operatorname{rank}\mathbf{A} = \dim\operatorname{col}(\mathbf{A}) = n$ (Definition 3.5.4, Theorem 1.6.8).

(a) $\Leftrightarrow$ (e). By Theorem 7.4.5, $\det\mathbf{A}^\top = \det\mathbf{A}$. Applying (a) $\Leftrightarrow$ (g) to $\mathbf{A}$ and to $\mathbf{A}^\top$, both "$\mathbf{A}$ invertible" and "$\mathbf{A}^\top$ invertible" are equivalent to $\det\mathbf{A} \neq 0$.

(e) $\Leftrightarrow$ (f). Let $\mathbf{r}_1, \dots, \mathbf{r}_n$ be the rows of $\mathbf{A}$. The columns of $\mathbf{A}^\top$ are $\mathbf{r}_1^\top, \dots, \mathbf{r}_n^\top$. By Theorem 0.9.5, $c_1\mathbf{r}_1 + \cdots + c_n\mathbf{r}_n = \mathbf{0}$ if and only if $c_1\mathbf{r}_1^\top + \cdots + c_n\mathbf{r}_n^\top = \mathbf{0}^\top$, so the rows of $\mathbf{A}$ are independent exactly when the columns of $\mathbf{A}^\top$ are. By Theorem 3.4.4 ((a) $\Leftrightarrow$ (f)) applied to $\mathbf{A}^\top$, this holds exactly when $\mathbf{A}^\top$ is invertible. This proves the theorem.

■

The determinant is now one more item on the menu, and for the right problems it is the best one. For a concrete matrix with a parameter, one polynomial in the parameter decides invertibility at once. For an abstract matrix, such as $\mathbf{A} + \mathbf{I}$ when $\mathbf{A}^k = \mathbf{O}$, the determinant is usually hopeless, and the kernel is the item to try first.

Quick check. For which $t \in \mathbb{R}$ is $\begin{pmatrix} 1 & t \\ t & 4 \end{pmatrix}$ invertible? Over $\mathbb{F}_3$, for which t?

Solution. The determinant is $4 - t^2 = (2 - t)(2 + t)$. Over $\mathbb{R}$ the matrix is invertible exactly for $t \neq \pm 2$. Over $\mathbb{F}_3$, $4 = 1$, so the determinant is $1 - t^2$, which is 1 for $t = 0$ and 0 for $t = 1, 2$; the matrix is invertible only for $t = 0$.

7.5.3 Multiplicativity

Now the second fact. A natural first attempt is to multiply out the Leibniz formula for $\mathbf{AB}$, whose entries are sums, and regroup. That is legitimate, but the bookkeeping is horrible. The uniqueness theorem does the work instead: we exhibit $\mathbf{B} \mapsto \det(\mathbf{AB})$ as an alternating form in the columns of $\mathbf{B}$, and then it has no choice.

Theorem 7.5.4. Multiplicativity of the Determinant

For all $\mathbf{A}, \mathbf{B} \in M_n(F)$,

$$\det(\mathbf{AB}) = \det \mathbf{A} \det \mathbf{B}.$$

Idea. Fix $\mathbf{A}$. The j-th column of $\mathbf{AB}$ is $\mathbf{Ab}_j$, so $g(\mathbf{b}_1, \dots, \mathbf{b}_n) = \det(\mathbf{Ab}_1, \dots, \mathbf{Ab}_n)$ is $\det(\mathbf{AB})$ viewed as a function of the columns of $\mathbf{B}$. It inherits linearity in each $\mathbf{b}_j$ from det and from matrix multiplication, and equal columns $\mathbf{b}_p = \mathbf{b}_q$ give equal columns $\mathbf{Ab}_p = \mathbf{Ab}_q$. So g is an alternating form, hence a multiple of det, and the multiple is $g(\mathbf{e}_1, \dots, \mathbf{e}_n) = \det \mathbf{A}$.

Proof. Fix $\mathbf{A} \in M_n(F)$, and define $g\colon F^n \times \cdots \times F^n \to F$ by

$$g(\mathbf{b}_1, \dots, \mathbf{b}_n) = \det(\mathbf{Ab}_1, \dots, \mathbf{Ab}_n).$$

By Theorem 0.9.2 (columns), $\mathbf{AB}$ has columns $\mathbf{Ab}_1, \dots, \mathbf{Ab}_n$ when $\mathbf{B}$ has columns $\mathbf{b}_1, \dots, \mathbf{b}_n$, so $g(\mathbf{b}_1, \dots, \mathbf{b}_n) = \det(\mathbf{AB})$.

g is n-linear. Fix k and all $\mathbf{b}_j$ with $j \neq k$. The map $\mathbf{b}_k \mapsto \mathbf{Ab}_k$ is linear by Theorem 0.9.4, and det is linear in its k-th argument by Theorem 7.4.3. A composite of linear maps is linear, so g is linear in $\mathbf{b}_k$.

g is alternating. If $\mathbf{b}_p = \mathbf{b}_q$ with $p \neq q$, then $\mathbf{Ab}_p = \mathbf{Ab}_q$, so $g(\mathbf{b}_1, \dots, \mathbf{b}_n) = 0$ since det is alternating.

By Corollary 7.4.4, $g(\mathbf{b}_1, \dots, \mathbf{b}_n) = g(\mathbf{e}_1, \dots, \mathbf{e}_n) \det(\mathbf{b}_1, \dots, \mathbf{b}_n)$. Finally $\mathbf{Ae}_j$ is the j-th column of $\mathbf{A}$ (Theorem 0.9.3), so $g(\mathbf{e}_1, \dots, \mathbf{e}_n) = \det \mathbf{A}$. Hence $\det(\mathbf{AB}) = \det \mathbf{A} \det \mathbf{B}$.

■

Remark. **Over a commutative ring.** This proof uses only Corollary 7.4.4 (in the explicit form of Theorem 7.4.2), linearity of matrix multiplication, and the Leibniz formula. None of these divides. So $\det(\mathbf{AB}) = \det \mathbf{A} \det \mathbf{B}$ holds for $\mathbf{A}, \mathbf{B} \in M_n(R)$ over any commutative ring R, such as $F[x]$.

Alternatively, the building blocks of Chapter 3 give a proof. If $\mathbf{A}$ is invertible, then $\mathbf{A} = \mathbf{E}_1 \cdots \mathbf{E}_k$

with elementary $\mathbf{E}_m$ (Theorem 3.4.4 (d)), and peeling off one factor at a time with Theorem 7.5.1 (e) gives $\det(\mathbf{AB}) = \det \mathbf{E}_1 \cdots \det \mathbf{E}_k \det \mathbf{B}$; the case $\mathbf{B} = \mathbf{I}_n$ of the same computation shows $\det \mathbf{A} = \det \mathbf{E}_1 \cdots \det \mathbf{E}_k$. If $\mathbf{A}$ is not invertible, neither is $\mathbf{AB}$: otherwise $(\mathbf{AB})\mathbf{C} = \mathbf{I}_n$ for some $\mathbf{C}$, so $\mathbf{A}(\mathbf{BC}) = \mathbf{I}_n$ and $\mathbf{A}$ would be invertible by Theorem 3.4.5. Then both sides are 0 by Theorem 7.5.2. This route leans on row reduction, which divides, so unlike the first proof it needs a field.

For reference, here are the two identities used most often, now both proved.

Theorem 7.5.5. Properties of Determinants

Let $\mathbf{A}, \mathbf{B} \in M_n(F)$. Then:

1. $\det(\mathbf{AB}) = \det \mathbf{A} \det \mathbf{B}$;
2. $\det \mathbf{A}^\top = \det \mathbf{A}$.

Proof. Part 1 is Theorem 7.5.4 and part 2 is Theorem 7.4.5. ∎

The first consequences are immediate.

Corollary 7.5.6. Determinant of the Inverse

If $\mathbf{A} \in M_n(F)$ is invertible, then $\det \mathbf{A} \neq 0$ and $\det(\mathbf{A}^{-1}) = (\det \mathbf{A})^{-1}$.

Proof. By Theorem 7.5.4, $\det \mathbf{A} \det(\mathbf{A}^{-1}) = \det(\mathbf{A}\mathbf{A}^{-1}) = \det \mathbf{I}_n = 1$. Hence $\det \mathbf{A} \neq 0$, and $\det(\mathbf{A}^{-1})$ is the inverse of $\det \mathbf{A}$ in F. ∎

Warning. **Multiplicativity does not make** det **additive, and it does not make products commute.** For $\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$ and $\mathbf{B} = \begin{pmatrix} 0 & 1 \\ 1 & 5 \end{pmatrix}$, $\mathbf{AB} = \begin{pmatrix} 2 & 11 \\ 4 & 23 \end{pmatrix} \neq \begin{pmatrix} 3 & 4 \\ 16 & 22 \end{pmatrix} = \mathbf{BA}$, yet $\det(\mathbf{AB}) = \det(\mathbf{BA}) = 2$, since both equal $\det \mathbf{A} \det \mathbf{B} = (-2)(-1)$. On the other hand, $\det(\mathbf{A} + \mathbf{B})$ is not determined by $\det \mathbf{A}$ and $\det \mathbf{B}$ at all: $\mathbf{I}_2$ and $\mathbf{I}_2$ have determinants $1, 1$ and $\det(\mathbf{I}_2 + \mathbf{I}_2) = 4$, while $\mathbf{I}_2$ and $\mathbf{C} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$ also have determinants $1, 1$ but $\det(\mathbf{I}_2 + \mathbf{C}) = \det \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix} = 2$.

Quick check. Let $\mathbf{A} = \begin{pmatrix} 2 & 1 \\ 7 & 4 \end{pmatrix}$. Without multiplying any matrices, find $\det(\mathbf{A}^5)$, $\det(\mathbf{A}^{-1})$, and $\det(\mathbf{A}^\top \mathbf{A})$.

Solution. $\det \mathbf{A} = 8-7 = 1$. By Theorem 7.5.4, $\det(\mathbf{A}^5) = (\det \mathbf{A})^5 = 1$. By Corollary 7.5.6, $\det(\mathbf{A}^{-1}) = 1^{-1} = 1$. By Theorem 7.5.5, $\det(\mathbf{A}^\top \mathbf{A}) = \det \mathbf{A}^\top \det \mathbf{A} = (\det \mathbf{A})^2 = 1$.

7.5.4 The determinant of an operator

In Chapter 2 the trace passed from matrices to operators because similar matrices have the same trace (Definition 4.5.2). Multiplicativity gives the same for determinants.

Corollary 7.5.7. Determinant Is a Similarity Invariant

If $\mathbf{A}, \mathbf{B} \in M_n(F)$ are similar, then $\det \mathbf{A} = \det \mathbf{B}$.

Proof. Let $\mathbf{B} = \mathbf{P}^{-1}\mathbf{A}\mathbf{P}$ with $\mathbf{P}$ invertible. By Theorem 7.5.4 and Corollary 7.5.6,

$$\det \mathbf{B} = \det(\mathbf{P}^{-1}) \det \mathbf{A} \det \mathbf{P} = (\det \mathbf{P})^{-1} \det \mathbf{P} \det \mathbf{A} = \det \mathbf{A},$$

where the middle step uses commutativity of multiplication in F. ∎

The converse fails: $\mathbf{I}_2$ and $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ both have determinant 1, but the only matrix similar to $\mathbf{I}_2$ is $\mathbf{P}^{-1}\mathbf{I}_2\mathbf{P} = \mathbf{I}_2$ itself.

By Theorem 4.2.4, all matrices of one operator in the various bases are similar to each other. So the corollary lets us define:

Definition 7.5.1. Determinant of an operator

Let V be a finite-dimensional vector space over F with $\dim V \geq 1$, and let $T \in \mathcal{L}(V)$. The **determinant** of T is

$$\det T := \det [T]_{\mathcal{B}}^{\mathcal{B}},$$

where $\mathcal{B}$ is **any** basis of V. For $V = \{0\}$ we set $\det T = 1$.

Well-definedness. If $\mathcal{B}$ and $\mathcal{C}$ are bases of V, then $[T]_{\mathcal{C}}^{\mathcal{C}} = \mathbf{P}^{-1}[T]_{\mathcal{B}}^{\mathcal{B}}\mathbf{P}$ with $\mathbf{P} = [\mathrm{id}]_{\mathcal{C}}^{\mathcal{B}}$ invertible, by Theorem 4.2.3 and Theorem 4.2.4 (a). By Corollary 7.5.7 the two determinants agree. The convention for $V = \{0\}$ is the empty product, and it keeps the next theorem true in that case.

Examples.

- **Scalars.** For $c \in F$ and $\dim V = n \geq 1$, $[c\,\mathrm{id}_V]_{\mathcal{B}}^{\mathcal{B}} = c\mathbf{I}_n$ in every basis, so $\det(c\,\mathrm{id}_V) = c^n$ by Theorem 7.4.6.
- **A reflection.** $R(x, y) = (y, x)$ on $\mathbb{R}^2$ has standard matrix $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$, of determinant -1. In the basis $((1, 1), (1, -1))$ its matrix is $\mathrm{diag}(1, -1)$ (Example (The reflection in a mirror-adapted basis)), again of determinant -1, as it must be.
- **Differentiation.** On $\mathbb{R}[x]_{\leq 3}$ with basis $(1, x, x^2, x^3)$, the matrix of $\mathbf{D}$ is upper triangular with zero diagonal (Example (Differentiation on Cubic Polynomials)), so $\det \mathbf{D} = 0$.

The determinant of a map $T\colon V \to W$ between **different** spaces is not defined, even when $\dim V = \dim W$. A matrix $[T]_{\mathcal{B}}^{\mathcal{C}}$ exists, but changing $\mathcal{C}$ alone multiplies it on the left by an arbitrary invertible matrix, which can change its determinant to any nonzero value. For $\mathrm{id}\colon \mathbb{R}^2 \to \mathbb{R}^2$ with input basis $\mathcal{E}$ and output basis $((2,0),(0,1))$, the matrix is $\mathrm{diag}(\frac{1}{2}, 1)$, of determinant $\frac{1}{2}$, not 1.

Theorem 7.5.8. Properties of the Determinant of an Operator

Let V be a finite-dimensional vector space over F and $S, T \in \mathcal{L}(V)$. Then

$$\det(ST) = \det S \det T, \qquad \det \mathrm{id}_V = 1,$$

and T is invertible if and only if $\det T \neq 0$.

Proof. If $V = \{0\}$, all determinants are 1 and the only operator is id_V, which is invertible. Otherwise fix a basis $\mathcal{B}$ of V. By Theorem 4.1.2, $[ST]_{\mathcal{B}}^{\mathcal{B}} = [S]_{\mathcal{B}}^{\mathcal{B}}[T]_{\mathcal{B}}^{\mathcal{B}}$, so $\det(ST) = \det S \det T$ by Theorem 7.5.4. Next $[\mathrm{id}_V]_{\mathcal{B}}^{\mathcal{B}} = \mathbf{I}_n$, of determinant 1. Finally, by Theorem 4.1.5 (b), T is invertible if and only if $[T]_{\mathcal{B}}^{\mathcal{B}}$ is, which by Theorem 7.5.2 holds if and only if $\det T = \det [T]_{\mathcal{B}}^{\mathcal{B}} \neq 0$.

■

So the determinant is a property of the operator, not of the coordinates. Differentiation on $\mathbb{R}[x]_{\leq 3}$ is not invertible because its determinant is 0, whichever basis we compute in. In the section on orientation and volume we will see what the number $\det T$ measures geometrically.

7.5.5 The determinant as a homomorphism

In Chapter 0 we met groups and homomorphisms, and promised that the determinant would be one. The general linear group $\mathrm{GL}_n(F)$ is the group of invertible matrices in $M_n(F)$ under multiplication, and $F \setminus \{0\}$ is a group under multiplication (Example (Groups)).

Theorem 7.5.9. The Determinant Is a Group Homomorphism

Let $n \geq 1$. The determinant restricts to a **surjective** group homomorphism

$$\det\colon \mathrm{GL}_n(F) \to F \setminus \{0\}.$$

Its kernel is the **special linear group** $\mathrm{SL}_n(F) \coloneqq \{\mathbf{A} \in M_n(F) : \det \mathbf{A} = 1\}$, which is therefore a subgroup of $\mathrm{GL}_n(F)$.

Proof. By Theorem 7.5.2, $\det \mathbf{A} \neq 0$ for $\mathbf{A} \in \mathrm{GL}_n(F)$, so det maps $\mathrm{GL}_n(F)$ into $F \setminus \{0\}$, and Theorem 7.5.4 says it is a homomorphism (Definition 0.10.3). For $c \neq 0$, the matrix $\mathrm{diag}(c, 1, \ldots, 1)$ has determinant c by Theorem 7.4.6, which is nonzero, so it lies in $\mathrm{GL}_n(F)$; hence det is surjective. A matrix with determinant 1 is invertible by Theorem 7.5.2, so the kernel $\{\mathbf{A} \in \mathrm{GL}_n(F) : \det \mathbf{A} = 1\}$ is $\mathrm{SL}_n(F)$. It is a subgroup by Theorem 0.10.3.

■

For example, $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ and the elementary matrices $\mathbf{I}_n + c\mathbf{E}_{ij}$ lie in $\mathrm{SL}_n(F)$, while $\mathbf{P}_{12}$, of determinant -1, does not. Over $\mathbb{R}$, the rotation matrices $\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$ have determinant $\cos^2\theta + \sin^2\theta = 1$ and lie in $\mathrm{SL}_2(\mathbb{R})$. The sign $\operatorname{sgn}\colon S_n \to \{1, -1\}$ of the third section is the other homomorphism promised in Chapter 0, and the two are linked: $\det \mathbf{P}_\sigma = \operatorname{sgn}(\sigma)$ (Exercise (B2: The determinant of $\mathbf{P}_\sigma$)), and $\mathbf{P}_\sigma \mathbf{P}_\tau = \mathbf{P}_{\sigma\tau}$ (Lemma 3.4.7).

7.5.6 Which proofs survive over a ring

Remark. **Commutative rings.** Over a commutative ring R (as in the remark at the end of the previous section), the following proofs of this section use no division and remain valid for matrices in $M_n(R)$: Theorem 7.5.1, with $c \in R$ arbitrary in (b) and (c); Theorem 7.5.4, by its first proof; and Theorem 7.5.5. The proofs of Corollary 7.5.6 and Corollary 7.5.7 also survive for matrices that have an inverse in $M_n(R)$ (then $\det \mathbf{A} \det(\mathbf{A}^{-1}) = 1$, so $\det \mathbf{A}$ has an inverse in R), for instance an invertible $\mathbf{P} \in M_n(F)$ regarded inside $M_n(F[x])$; this is the case needed for characteristic polynomials. What fails over a ring is the converse direction: a matrix whose determinant is nonzero need not be invertible, since det may be a nonzero element with no inverse, such as $2 \in \mathbb{Z}$ or $x \in F[x]$. By contrast, Theorem 7.5.2, Theorem 7.5.3, and the "Alternatively" proof of multiplicativity run through row reduction, which divides by pivots, and they are proved here only over a field.

7.5.7 Exercises

7.5.7.1 A. Check your understanding

Exercise. A1

(a) State how each of the three elementary row operations changes the determinant.

(b) True or false: the move $R_2 \to 3R_2 - R_1$ does not change the determinant. Justify your answer.

(c) True or false: if $\mathbf{A}, \mathbf{B} \in M_n(F)$ have $\det \mathbf{A} = \det \mathbf{B}$, then $\mathbf{A}$ and $\mathbf{B}$ are similar. Justify your answer.

(d) True or false: $\det(\mathbf{AB}) = \det(\mathbf{BA})$ for all $\mathbf{A}, \mathbf{B} \in M_n(F)$. Justify your answer.

(e) Explain why $\det T$ does not depend on the basis used to compute it.

(f) Is $\det\colon M_n(F) \to F$ a group homomorphism? Explain.

Solution.

(a) By Theorem 7.5.1: a swap multiplies it by -1, scaling a row by c multiplies it by c, and adding a multiple of one row to a different row leaves it unchanged.

(b) False. The move is $R_2 \to 3R_2$ followed by $R_2 \to R_2 - R_1$, so it multiplies the determinant by 3. For $\mathbf{I}_2$ it gives $\begin{pmatrix} 1 & 0 \\ -1 & 3 \end{pmatrix}$, of determinant $3 \neq 1$.

(c) False. $\mathbf{I}_2$ and $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ both have determinant 1, but $\mathbf{P}^{-1}\mathbf{I}_2\mathbf{P} = \mathbf{I}_2$ for every invertible $\mathbf{P}$, so $\mathbf{I}_2$ is similar only to itself.

(d) True. Both equal $\det \mathbf{A} \det \mathbf{B}$ by Theorem 7.5.4, since multiplication in F is commutative.

(e) Matrices of T in two bases are similar (Theorem 4.2.3), and similar matrices have equal determinants (Corollary 7.5.7).

(f) No: $M_n(F)$ is not a group under multiplication, since the zero matrix has no inverse. The restriction $\det\colon \mathrm{GL}_n(F) \to F \setminus \{0\}$ is a homomorphism (Theorem 7.5.9).

7.5.7.2 B. Practice

Exercise. B1: Determinants by elimination

Compute the determinants of the following real matrices by row reduction, recording the effect of each operation.

$$\mathbf{A} = \begin{pmatrix} 1 & 2 & 0 & 1 \\ 2 & 3 & 1 & 4 \\ 0 & 1 & 3 & 2 \\ -1 & 1 & 2 & 5 \end{pmatrix}, \qquad \mathbf{B} = \begin{pmatrix} 2 & -1 & 3 & 1 \\ 4 & -2 & 7 & 3 \\ -2 & 2 & -1 & 4 \\ 6 & -3 & 9 & 5 \end{pmatrix}.$$

Solution. For $\mathbf{A}$: the replacements $R_2 \to R_2 - 2R_1$ and $R_4 \to R_4 + R_1$ give rows $(1,2,0,1)$, $(0,-1,1,2)$, $(0,1,3,2)$, $(0,3,2,6)$. Then $R_3 \to R_3 + R_2$ and $R_4 \to R_4 + 3R_2$ give $(0,0,4,4)$ and $(0,0,5,12)$. Pulling the factor 4 out of row 3 leaves $(0,0,1,1)$, and $R_4 \to R_4 - 5R_3$ gives $(0,0,0,7)$. None of the replacements changes the determinant (Theorem 7.5.1), so

$$\det \mathbf{A} = 4 \det \begin{pmatrix} 1 & 2 & 0 & 1 \\ 0 & -1 & 1 & 2 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 7 \end{pmatrix} = 4 \cdot (1 \cdot (-1) \cdot 1 \cdot 7) = -28,$$

by Theorem 7.4.6.

For $\mathbf{B}$: $R_2 \to R_2 - 2R_1$, $R_3 \to R_3 + R_1$, $R_4 \to R_4 - 3R_1$ give rows $(2,-1,3,1)$, $(0,0,1,1)$, $(0,1,2,5)$, $(0,0,0,2)$, without changing the determinant. The $(2,2)$-entry is 0, so swap $R_2 \leftrightarrow R_3$, which multiplies the determinant by -1. The result is upper triangular with diagonal $2,1,1,2$, so

$$\det \mathbf{B} = -(2 \cdot 1 \cdot 1 \cdot 2) = -4.$$

Exercise. B2: A parameter

Let $t \in \mathbb{R}$ and $\mathbf{A}_t = \begin{pmatrix} t & 1 & 1 \\ 1 & t & 1 \\ 1 & 1 & t \end{pmatrix}$. Find $\det \mathbf{A}_t$ as a factored polynomial in t. Hence determine for which t the matrix $\mathbf{A}_t$ is invertible.

Solution. The replacements $R_1 \to R_1 + R_2$ and $R_1 \to R_1 + R_3$ do not change the determinant and make row 1 equal to $(t+2, t+2, t+2)$. By linearity in row 1 (this holds for every value of $t+2$, including 0),

$$\det \mathbf{A}_t = (t+2) \det \begin{pmatrix} 1 & 1 & 1 \\ 1 & t & 1 \\ 1 & 1 & t \end{pmatrix}$$
$$= (t+2) \det \begin{pmatrix} 1 & 1 & 1 \\ 0 & t-1 & 0 \\ 0 & 0 & t-1 \end{pmatrix} = (t+2)(t-1)^2,$$

where the second step uses $R_2 \to R_2 - R_1$, $R_3 \to R_3 - R_1$, and the last uses Theorem 7.4.6. By Theorem 7.5.2, $\mathbf{A}_t$ is invertible if and only if $t \neq 1$ and $t \neq -2$.

Exercise. B3: The shift operator

Let $T\colon \mathbb{R}[x]_{\leq 3} \to \mathbb{R}[x]_{\leq 3}$, $T(p) = p(x+1)$. Find $\det T$. Hence show that T is invertible.

Solution. In the basis $\mathcal{B} = (1, x, x^2, x^3)$, $T(1) = 1$, $T(x) = 1 + x$, $T(x^2) = 1 + 2x + x^2$, and $T(x^3) = 1 + 3x + 3x^2 + x^3$. So

$$[T]_{\mathcal{B}}^{\mathcal{B}} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 \\ 0 & 0 & 1 & 3 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

which is upper triangular with ones on the diagonal. By Definition 7.5.1 and Theorem 7.4.6, $\det T = 1$. Since $\det T \neq 0$, T is invertible by Theorem 7.5.8. (Its inverse is $p \mapsto p(x-1)$.)

7.5.7.3 C. Going deeper

Exercise. C1: Orthogonal-type matrices

(a) Let $\mathbf{A} \in M_n(\mathbb{R})$ satisfy $\mathbf{A}^\top \mathbf{A} = \mathbf{I}_n$. Prove that $\det \mathbf{A} = 1$ or $\det \mathbf{A} = -1$.

(b) Show that both values occur, for every $n \geq 1$.

(c) Over $\mathbb{F}_2$, what can $\det \mathbf{A}$ be if $\mathbf{A}^\top \mathbf{A} = \mathbf{I}_n$?

Solution.

(a) By Theorem 7.5.5, $1 = \det \mathbf{I}_n = \det(\mathbf{A}^\top \mathbf{A}) = \det \mathbf{A}^\top \det \mathbf{A} = (\det \mathbf{A})^2$. A real number with square 1 is 1 or -1.

(b) $\mathbf{I}_n^\top \mathbf{I}_n = \mathbf{I}_n$ and $\det \mathbf{I}_n = 1$. For $\mathbf{D} = \operatorname{diag}(-1, 1, \ldots, 1)$, $\mathbf{D}^\top \mathbf{D} = \mathbf{D}^2 = \mathbf{I}_n$ and $\det \mathbf{D} = -1$ by Theorem 7.4.6.

(c) The same computation gives $(\det \mathbf{A})^2 = 1$. In $\mathbb{F}_2$, $\det \mathbf{A} \in \{0, 1\}$ and $0^2 = 0$, so $\det \mathbf{A} = 1$. This agrees with (a), since $-1 = 1$ in $\mathbb{F}_2$.

Exercise. C2: Left multiplication on matrices

Let $\mathbf{A} \in M_n(F)$, and let $L_{\mathbf{A}} \colon M_n(F) \to M_n(F)$, $L_{\mathbf{A}}(\mathbf{X}) = \mathbf{AX}$. This is a linear operator on a space of dimension n^2. Prove that

$$\det L_{\mathbf{A}} = (\det \mathbf{A})^n.$$

(a) Order the matrix units column by column,

$$\mathcal{B} = (\mathbf{E}_{11}, \mathbf{E}_{21}, \dots, \mathbf{E}_{n1}, \mathbf{E}_{12}, \dots, \mathbf{E}_{n2}, \dots, \mathbf{E}_{1n}, \dots, \mathbf{E}_{nn}).$$

Show that $[L_{\mathbf{A}}]_{\mathcal{B}}^{\mathcal{B}}$ consists of n copies of $\mathbf{A}$ placed along the diagonal, with zeros elsewhere.

(b) Show that $L_{\mathbf{AB}} = L_{\mathbf{A}} L_{\mathbf{B}}$, and prove the formula when $\mathbf{A}$ is an elementary matrix.

(c) Prove the formula for invertible $\mathbf{A}$, and then for noninvertible $\mathbf{A}$.

Hint: for a swap matrix, identify $[L_{\mathbf{P}_{ij}}]_{\mathcal{B}}^{\mathcal{B}}$ as a permutation matrix and use Exercise (B2: The determinant of $\mathbf{P}_\sigma$).

Solution.

(a) By Theorem 0.9.2, the l-th column of $\mathbf{AE}_{kl}$ is $\mathbf{Ae}_k$, the k-th column of $\mathbf{A}$, and the other columns are zero. So $\mathbf{AE}_{kl} = \sum_i a_{ik}\mathbf{E}_{il}$. In $\mathcal{B}$, the coordinates of this vector are zero outside the l-th block of n positions, and inside that block they are the k-th column of $\mathbf{A}$. Hence the columns of $[L_{\mathbf{A}}]_{\mathcal{B}}^{\mathcal{B}}$ belonging to the l-th block are those of $\mathbf{A}$, placed in the rows of the l-th block: $[L_{\mathbf{A}}]_{\mathcal{B}}^{\mathcal{B}}$ has n copies of $\mathbf{A}$ along the diagonal.

(b) $L_{\mathbf{AB}}(\mathbf{X}) = \mathbf{ABX} = L_{\mathbf{A}}(L_{\mathbf{B}}(\mathbf{X}))$ by associativity. Now let $\mathbf{E}$ be elementary. If $\mathbf{E} = \mathbf{I}_n + c\mathbf{E}_{ij}$ with $i \neq j$, then by (a) $[L_{\mathbf{E}}]_{\mathcal{B}}^{\mathcal{B}}$ is triangular (upper if $i < j$, lower if $i > j$) with all diagonal entries 1, so $\det L_{\mathbf{E}} = 1 = (\det \mathbf{E})^n$ by Theorem 7.4.6 and Theorem 7.5.1 (e). If $\mathbf{E} = \mathbf{D}_i(c)$, then $[L_{\mathbf{E}}]_{\mathcal{B}}^{\mathcal{B}}$ is diagonal with c appearing n times and 1 elsewhere, so $\det L_{\mathbf{E}} = c^n = (\det \mathbf{E})^n$. If $\mathbf{E} = \mathbf{P}_{ij}$, then $[L_{\mathbf{E}}]_{\mathcal{B}}^{\mathcal{B}}$ has exactly one 1 in each row and column, so it is a permutation matrix $\mathbf{P}_\sigma$ (Definition 3.4.2), and σ exchanges the positions of $\mathbf{E}_{il}$ and $\mathbf{E}_{jl}$ for each l: σ is a product of n disjoint transpositions. By Exercise (B2: The determinant of $\mathbf{P}_\sigma$), Theorem 7.3.1, and Corollary 7.3.2, $\det L_{\mathbf{E}} = \operatorname{sgn}(\sigma) = (-1)^n = (\det \mathbf{E})^n$.

(c) If $\mathbf{A}$ is invertible, then $\mathbf{A} = \mathbf{E}_1 \cdots \mathbf{E}_k$ with elementary $\mathbf{E}_m$ (Theorem 3.4.4 (d)). By (b) and Theorem 7.5.8,

$$\begin{aligned} \det L_{\mathbf{A}} &= \det L_{\mathbf{E}_1} \cdots \det L_{\mathbf{E}_k} = (\det \mathbf{E}_1)^n \cdots (\det \mathbf{E}_k)^n \\ &= (\det \mathbf{E}_1 \cdots \det \mathbf{E}_k)^n = (\det \mathbf{A})^n, \end{aligned}$$

where the last step is Theorem 7.5.4. If $\mathbf{A}$ is not invertible, there is $\mathbf{v} \neq \mathbf{0}$ with $\mathbf{Av} = \mathbf{0}$ (Theorem 3.4.4 (b)). The matrix $\mathbf{X} = \begin{pmatrix} \mathbf{v} & \mathbf{0} & \cdots & \mathbf{0} \end{pmatrix}$ is nonzero and $\mathbf{AX} = \mathbf{O}$, so $L_{\mathbf{A}}$ is not injective, hence not invertible, and $\det L_{\mathbf{A}} = 0$ by Theorem 7.5.8. Also $\det \mathbf{A} = 0$ by Theorem 7.5.2, so $(\det \mathbf{A})^n = 0$ as $n \geq 1$.

Exercise. C3: Nonsquare factors

(a) Give $\mathbf{A} \in M_{2\times 1}(\mathbb{R})$ and $\mathbf{B} \in M_{1\times 2}(\mathbb{R})$ with $\det(\mathbf{BA}) \neq \det(\mathbf{AB})$.

(b) Let $m > n$, $\mathbf{A} \in M_{m\times n}(F)$, and $\mathbf{B} \in M_{n\times m}(F)$. Prove that $\det(\mathbf{AB}) = 0$.

(c) Explain why (a) does not contradict Theorem 7.5.4.

Solution.

(a) Let $\mathbf{A} = \begin{pmatrix}1\\0\end{pmatrix}$ and $\mathbf{B} = \begin{pmatrix}1 & 0\end{pmatrix}$. Then $\mathbf{BA} = (1)$, of determinant 1, and $\mathbf{AB} = \begin{pmatrix}1 & 0\\0 & 0\end{pmatrix}$, of determinant 0.

(b) By Theorem 3.5.9, $\operatorname{rank}(\mathbf{AB}) \le \operatorname{rank}\mathbf{A} \le n$, since $\operatorname{col}(\mathbf{A})$ is spanned by n columns. So $\operatorname{rank}(\mathbf{AB}) < m$, and the $m \times m$ matrix $\mathbf{AB}$ is not invertible by Theorem 7.5.3 (d). Hence $\det(\mathbf{AB}) = 0$ by Theorem 7.5.2.

(c) Theorem 7.5.4 concerns two **square** matrices of the same size, and then $\det(\mathbf{AB}) = \det(\mathbf{BA})$. Here $\mathbf{A}$ and $\mathbf{B}$ are not square, $\det\mathbf{A}$ and $\det\mathbf{B}$ are not defined, and $\mathbf{AB}$, $\mathbf{BA}$ have different sizes.

7.6 Cofactor Expansion, the Adjugate, and Cramer's Rule

We can now compute determinants by elimination and we know what they decide. This section adds a third view: a determinant of size n is a combination of determinants of size $n-1$. This recursive formula, the Laplace expansion, is rarely the fastest way to compute, but it has two important consequences. It gives an explicit formula for the inverse of a matrix, through the adjugate, and an explicit formula for the solution of a square linear system, Cramer's rule. Both rest on one identity built from the entries alone, with no division, and that identity holds for matrices of polynomials too, which is what later chapters need.

7.6.1 Minors and cofactors

Group the six terms of a 3×3 determinant (from the examples after Definition 7.4.1) by the entry they take from column 1:

$$\det\mathbf{A} = a_{11}\,(a_{22}a_{33} - a_{32}a_{23}) - a_{21}\,(a_{12}a_{33} - a_{32}a_{13}) + a_{31}\,(a_{12}a_{23} - a_{22}a_{13}).$$

Each bracket is a 2×2 determinant. The bracket beside a_{11} is the determinant of what remains of $\mathbf{A}$ after deleting row 1 and column 1; the bracket beside a_{21} is what remains after deleting row 2 and column 1; and similarly for a_{31}. The signs alternate. These smaller determinants recur constantly, so they get names.

A minor is the determinant of what is left after crossing out one row and one column; a cofactor is that minor with a checkerboard sign.

Definition 7.6.1. Minor and cofactor

Let $n \ge 2$, $\mathbf{A} \in M_n(F)$, and $i, j \in \{1, \dots, n\}$. The **(i,j)-minor** of $\mathbf{A}$ is

$$M_{ij} \coloneqq \det(\text{the } (n-1)\times(n-1) \text{ matrix obtained from } \mathbf{A} \text{ by deleting row } i \text{ and column } j).$$

The (i,j)**-cofactor** of $\mathbf{A}$ is $C_{ij} \coloneqq (-1)^{i+j} M_{ij}$.

The rows and columns that remain keep their original order. The sign $(-1)^{i+j}$ depends only on the position, and follows a checkerboard starting with $+$ in the top-left corner:

$$\begin{pmatrix} + & - & + & \cdots \\ - & + & - & \cdots \\ + & - & + & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

When the matrix matters we write $M_{ij}(\mathbf{A})$ and $C_{ij}(\mathbf{A})$.

Examples. For $\mathbf{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, the minors are 1×1 determinants: $M_{11} = d$, $M_{12} = c$, $M_{21} = b$, $M_{22} = a$, and the cofactors are $C_{11} = d$, $C_{12} = -c$, $C_{21} = -b$, $C_{22} = a$. For the matrix $\mathbf{A} = \begin{pmatrix} 2 & 1 & 3 \\ 0 & 4 & 1 \\ 5 & 2 & 1 \end{pmatrix}$ of the previous sections, the cofactors along column 1 are

$$C_{11} = +\det \begin{pmatrix} 4 & 1 \\ 2 & 1 \end{pmatrix} = 2,$$
$$C_{21} = -\det \begin{pmatrix} 1 & 3 \\ 2 & 1 \end{pmatrix} = 5,$$
$$C_{31} = +\det \begin{pmatrix} 1 & 3 \\ 4 & 1 \end{pmatrix} = -11,$$

and $a_{11}C_{11} + a_{21}C_{21} + a_{31}C_{31} = 4 + 0 - 55 = -51$, which is $\det \mathbf{A}$ as computed before. For the identity $\mathbf{I}_n$, $C_{ii} = \det \mathbf{I}_{n-1} = 1$, and $C_{ij} = 0$ for $i \neq j$, because deleting row i and a different column j leaves a matrix with a zero column.

Nonexample by minimal change. A minor and a cofactor differ only by a sign, and forgetting the sign is the commonest slip. For the 2×2 matrix above, $M_{12} = c$ but $C_{12} = -c$; using minors throughout would turn $a_{11}C_{11} + a_{21}C_{21} = ad - cb$ into $ad + cb$, which is the permanent, not the determinant.

7.6.2 The Laplace expansion

The 3×3 grouping above is a special case of a general identity, valid along **every** row and **every** column.

Theorem 7.6.1. Laplace Expansion

Let $n \geq 2$ and $\mathbf{A} \in M_n(F)$, with cofactors C_{ij}.

(a) **(Along column j)** For each $j \in \{1, \ldots, n\}$,

$$\det \mathbf{A} = a_{1j}C_{1j} + a_{2j}C_{2j} + \cdots + a_{nj}C_{nj} = \sum_{i=1}^{n} a_{ij}C_{ij}.$$

(b) **(Along row i)** For each $i \in \{1, \ldots, n\}$,

$$\det \mathbf{A} = a_{i1}C_{i1} + a_{i2}C_{i2} + \cdots + a_{in}C_{in} = \sum_{j=1}^{n} a_{ij}C_{ij}.$$

Idea. The plan for (a):

① Write column j as $\sum_i a_{ij}\mathbf{e}_i$ and use linearity in column j. This leaves n determinants, the i-th with column j replaced by $\mathbf{e}_i$.

② In such a determinant, move the lone 1 to the top-left corner. To keep the other rows and columns in their original order, use **adjacent** swaps: $j-1$ column swaps and $i-1$ row swaps, costing $(-1)^{(i-1)+(j-1)} = (-1)^{i+j}$.

③ A matrix whose first column is $\mathbf{e}_1$ has the same determinant as its lower-right $(n-1) \times (n-1)$ block: only permutations fixing 1 contribute to the Leibniz formula. After ②, that block is exactly the matrix defining M_{ij}.

Part (b) is part (a) for the transpose.

Proof.

(a) Fix j. For $i \in \{1, \ldots, n\}$, let $\mathbf{B}^{(i)}$ be the matrix obtained from $\mathbf{A}$ by replacing column j with $\mathbf{e}_i$.
Claim. $\det \mathbf{B}^{(i)} = C_{ij}$.
Proof. Write $\mathbf{B} = \mathbf{B}^{(i)}$, with columns $\mathbf{b}_1, \ldots, \mathbf{b}_n$, so $\mathbf{b}_j = \mathbf{e}_i$ and $\mathbf{b}_k$ is the k-th column of $\mathbf{A}$ for $k \neq j$.
Moving the column. Swap columns $j-1$ and j, then columns $j-2$ and $j-1$, and so on, ending with columns 1 and 2. After these $j-1$ adjacent swaps the columns are $\mathbf{b}_j, \mathbf{b}_1, \ldots, \mathbf{b}_{j-1}, \mathbf{b}_{j+1}, \ldots, \mathbf{b}_n$: the column $\mathbf{b}_j$ is first and the others are in their original order. Call this matrix $\mathbf{B}_1$. By Theorem 7.5.1 (d), each swap multiplies the determinant by -1, so $\det \mathbf{B}_1 = (-1)^{j-1} \det \mathbf{B}$.
Moving the row. In the same way, move row i of $\mathbf{B}_1$ to the top by $i-1$ adjacent row swaps, keeping the other rows in their original order. Call the result $\mathbf{B}_2$. By Theorem 7.5.1 (a), $\det \mathbf{B}_2 = (-1)^{i-1} \det \mathbf{B}_1$.
The shape of $\mathbf{B}_2$. The first column of $\mathbf{B}_1$ is $\mathbf{e}_i$, and moving row i to the top turns it into $\mathbf{e}_1$. For $p, q \geq 2$, the (p,q)-entry of $\mathbf{B}_2$ comes from the rows of $\mathbf{A}$ other than i and the columns of $\mathbf{A}$ other than j, both in their original order. So the lower-right $(n-1) \times (n-1)$ block of $\mathbf{B}_2$ is the matrix $\mathbf{A}'$ obtained from $\mathbf{A}$ by deleting row i and column j. Writing b'_{pq} for the entries of $\mathbf{B}_2$, we have $b'_{11} = 1$, $b'_{p1} = 0$ for $p \geq 2$, and $b'_{pq} = a'_{p-1,q-1}$ for $p, q \geq 2$.
The block determinant. In the Leibniz formula for $\det \mathbf{B}_2$, the factor $b'_{\sigma(1)1}$ is 0 unless $\sigma(1) = 1$. Let $G = \{\sigma \in S_n : \sigma(1) = 1\}$. For $\sigma \in G$, define $\rho \in S_{n-1}$ by $\rho(k) = \sigma(k+1) - 1$; this is a bijection of $\{1, \ldots, n-1\}$, because σ maps $\{2, \ldots, n\}$ bijectively onto itself. The map $\sigma \mapsto \rho$ is a bijection $G \to S_{n-1}$, with inverse $\rho \mapsto \sigma$, $\sigma(1) = 1$, $\sigma(k+1) = \rho(k)+1$. An inversion of σ (Definition 7.3.1) is a pair $p < q$ with $\sigma(p) > \sigma(q)$; it never has $p = 1$, since $\sigma(1) = 1$ is smallest, and for $p, q \geq 2$ it is exactly an inversion

$(p-1, q-1)$ of ρ. So $\operatorname{sgn}(\sigma) = \operatorname{sgn}(\rho)$ by Definition 7.3.2. Therefore

$$\det \mathbf{B}_2 = \sum_{\sigma \in G} \operatorname{sgn}(\sigma) \cdot 1 \cdot \prod_{q=2}^{n} b'_{\sigma(q)q} = \sum_{\rho \in S_{n-1}} \operatorname{sgn}(\rho) \prod_{k=1}^{n-1} a'_{\rho(k)k} = \det \mathbf{A}' = M_{ij}.$$

Combining, $\det \mathbf{B} = (-1)^{j-1}(-1)^{i-1} \det \mathbf{B}_2 = (-1)^{i+j} M_{ij} = C_{ij}$.

■

Now the j-th column of $\mathbf{A}$ is $\sum_{i=1}^{n} a_{ij}\mathbf{e}_i$, and the other columns of $\mathbf{A}$ and of each $\mathbf{B}^{(i)}$ agree. By linearity of det in column j (Theorem 7.4.3) and the claim,

$$\det \mathbf{A} = \sum_{i=1}^{n} a_{ij} \det \mathbf{B}^{(i)} = \sum_{i=1}^{n} a_{ij} C_{ij}.$$

(b) Fix i. The (k,i)-entry of $\mathbf{A}^\top$ is a_{ik}. Deleting row k and column i from $\mathbf{A}^\top$ gives the transpose of the matrix obtained from $\mathbf{A}$ by deleting row i and column k, so $M_{ki}(\mathbf{A}^\top) = M_{ik}(\mathbf{A})$ by Theorem 7.4.5, and hence $C_{ki}(\mathbf{A}^\top) = C_{ik}(\mathbf{A})$. Applying (a) to $\mathbf{A}^\top$ along column i, and Theorem 7.4.5 again,

$$\det \mathbf{A} = \det \mathbf{A}^\top = \sum_{k=1}^{n} a_{ik}\, C_{ki}(\mathbf{A}^\top) = \sum_{k=1}^{n} a_{ik}\, C_{ik}(\mathbf{A}).$$

This proves the theorem.

■

Quick check. In step ② of the Idea, why not move column j to the front with the single swap of columns 1 and j?

Solution. A single swap costs only one factor -1, but it puts the old first column in position j, so the lower-right block is no longer the matrix defining M_{ij}: its columns are out of order. For $j \geq 3$, putting that column back in place within the block takes $j-2$ further adjacent swaps, and the total sign is again $(-1)^{1+(j-2)} = (-1)^{j-1}$. Adjacent swaps keep the count honest from the start.

Remark. **Over a commutative ring.** The proof uses multilinearity and the swap rule for columns and rows, the Leibniz formula, and Theorem 7.4.5. All of these are division-free (see the remarks at the ends of the previous two sections). So the Laplace expansion holds for matrices with entries in any commutative ring R, for instance $F[x]$.

The expansion is most useful along a row or column with many zeros, since each zero entry kills its term without computing the minor.

Example. A 4×4 Expansion

Compute $\det \mathbf{A}$ for

$$\mathbf{A} = \begin{pmatrix} 3 & 1 & 0 & 2 \\ 1 & 0 & 2 & 0 \\ 2 & 4 & 1 & 3 \\ 0 & 1 & 5 & 2 \end{pmatrix} \in M_4(\mathbb{R})$$

by expanding along a suitable row.

Solution. Row 2 has two zeros. Its signs in the checkerboard are $-, +, -, +$. By Theorem 7.6.1 (b) along row 2,

$$\det \mathbf{A} = 1 \cdot C_{21} + 0 \cdot C_{22} + 2 \cdot C_{23} + 0 \cdot C_{24} = -M_{21} - 2M_{23}.$$

Deleting row 2 and column 1, and then row 2 and column 3,

$$M_{21} = \det \begin{pmatrix} 1 & 0 & 2 \\ 4 & 1 & 3 \\ 1 & 5 & 2 \end{pmatrix}, \qquad M_{23} = \det \begin{pmatrix} 3 & 1 & 2 \\ 2 & 4 & 3 \\ 0 & 1 & 2 \end{pmatrix}.$$

Expand M_{21} along its first row, which has a zero:

$$M_{21} = 1 \cdot \det \begin{pmatrix} 1 & 3 \\ 5 & 2 \end{pmatrix} - 0 + 2 \cdot \det \begin{pmatrix} 4 & 1 \\ 1 & 5 \end{pmatrix} = (2 - 15) + 2(20 - 1) = 25.$$

Expand M_{23} along its third row $(0, 1, 2)$, with signs $+, -, +$:

$$M_{23} = -1 \cdot \det \begin{pmatrix} 3 & 2 \\ 2 & 3 \end{pmatrix} + 2 \cdot \det \begin{pmatrix} 3 & 1 \\ 2 & 4 \end{pmatrix} = -(9 - 4) + 2(12 - 2) = 15.$$

Hence $\det \mathbf{A} = -25 - 2 \cdot 15 = -55$.

Warning. **The sign belongs to the position, not to the place in the row.** Along row 2 the signs start with $-$. Using the first-row pattern $+, -, +, -$ in the example above would give $1 \cdot 25 + 2 \cdot 15 = 55$, the wrong sign. Always read the sign as $(-1)^{i+j}$.

Expansion alone is slow: expanding fully down to 1×1 determinants reproduces all $n!$ Leibniz terms. In practice one first creates zeros with replacements, which do not change the determinant, and then expands along the line of zeros.

7.6.3 The adjugate

In Chapter 0 we inverted a 2×2 matrix by multiplying it by $\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$ (Theorem 0.9.8). Compare with the cofactors listed after Definition 7.6.1: $C_{11} = d$, $C_{12} = -c$, $C_{21} = -b$, $C_{22} = a$. The matrix of Chapter 0 is the matrix of cofactors, **transposed**. That recurring matrix deserves a name in every size.

The adjugate is the transposed matrix of cofactors.

Definition 7.6.2. Adjugate

Let $n \geq 2$ and $\mathbf{A} \in M_n(F)$. The **adjugate** of $\mathbf{A}$ is the matrix $\operatorname{adj} \mathbf{A} \in M_n(F)$ with entries

$$(\operatorname{adj} \mathbf{A})_{ij} \coloneqq C_{ji},$$

that is, the **transpose** of the matrix (C_{ij}) of cofactors. For $n = 1$ we set $\operatorname{adj} \mathbf{A} \coloneqq (1)$.

In words: the entry in row i, column j of $\operatorname{adj} \mathbf{A}$ is the cofactor from row j, column i of $\mathbf{A}$. The adjugate is defined for **every** square matrix, invertible or not.

Examples. For $n = 2$, $\operatorname{adj} \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$. For $\mathbf{I}_n$, the cofactors computed above give $\operatorname{adj} \mathbf{I}_n = \mathbf{I}_n$. For

$$\mathbf{B} = \begin{pmatrix} 1 & 2 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 3 \end{pmatrix},$$

the nine cofactors, row by row, are $C_{11} = 3$, $C_{12} = 1$, $C_{13} = -1$; $C_{21} = -6$, $C_{22} = 3$, $C_{23} = 2$; $C_{31} = 2$, $C_{32} = -1$, $C_{33} = 1$. Transposing,

$$\operatorname{adj} \mathbf{B} = \begin{pmatrix} 3 & -6 & 2 \\ 1 & 3 & -1 \\ -1 & 2 & 1 \end{pmatrix}, \qquad \mathbf{B} \operatorname{adj} \mathbf{B} = \begin{pmatrix} 5 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 5 \end{pmatrix},$$

and $\det \mathbf{B} = 1 \cdot 3 + 2 \cdot 1 + 0 \cdot (-1) = 5$, by expansion along row 1.

Nonexample by minimal change. Skip the transpose. The cofactor matrix $\begin{pmatrix} 3 & 1 & -1 \\ -6 & 3 & 2 \\ 2 & -1 & 1 \end{pmatrix}$ has the same entries in other places, and $\mathbf{B}$ times it is $\begin{pmatrix} -9 & 7 & 3 \\ -4 & 2 & 3 \\ 9 & -2 & 2 \end{pmatrix}$, which is not a multiple of $\mathbf{I}_3$. The transpose is what lines up row i of $\mathbf{B}$ with the cofactors **of row** i.

The product $\mathbf{B} \operatorname{adj} \mathbf{B} = (\det \mathbf{B})\mathbf{I}_3$ is no accident. Its diagonal entries are row expansions of $\det \mathbf{B}$. Its off-diagonal entries are expansions too, of a matrix that looks like $\mathbf{B}$ but has a repeated row.

Theorem 7.6.2. The Adjugate Identity

For every $\mathbf{A} \in M_n(F)$,

$$\mathbf{A} \operatorname{adj} \mathbf{A} = \operatorname{adj} \mathbf{A}\, \mathbf{A} = (\det \mathbf{A})\, \mathbf{I}_n.$$

Idea. The (i, k)-entry of $\mathbf{A} \operatorname{adj} \mathbf{A}$ is $\sum_j a_{ij} C_{kj}$: the entries of row i against the cofactors of row k. If $i = k$, this is the expansion of $\det \mathbf{A}$ along row i. If $i \neq k$, build a look-alike matrix $\tilde{\mathbf{A}}$: copy $\mathbf{A}$, but overwrite row k with row i. The cofactors along row k never see row k, because computing them deletes it, so $\tilde{\mathbf{A}}$ has the same cofactors along row k as $\mathbf{A}$. Expanding $\det \tilde{\mathbf{A}}$ along row k gives exactly $\sum_j a_{ij} C_{kj}$, and $\det \tilde{\mathbf{A}} = 0$ because two rows are equal.

Proof. For $n = 1$, $\mathbf{A} = (a)$, $\operatorname{adj}\mathbf{A} = (1)$, and both products are $(a) = (\det\mathbf{A})\mathbf{I}_1$. Let $n \geq 2$.

The product $\mathbf{A}\operatorname{adj}\mathbf{A}$. By Theorem 0.9.2 and Definition 7.6.2, its (i,k)-entry is

$$\sum_{j=1}^{n} a_{ij}\,(\operatorname{adj}\mathbf{A})_{jk} = \sum_{j=1}^{n} a_{ij}\,C_{kj}.$$

If $i = k$, this is $\det\mathbf{A}$ by Theorem 7.6.1 (b) along row i. Suppose $i \neq k$, and let $\tilde{\mathbf{A}}$ be the matrix obtained from $\mathbf{A}$ by replacing row k with a copy of row i. Rows i and k of $\tilde{\mathbf{A}}$ are equal, so $\det\tilde{\mathbf{A}} = 0$ by Theorem 7.4.5. Expanding along row k by Theorem 7.6.1 (b),

$$0 = \det\tilde{\mathbf{A}} = \sum_{j=1}^{n} \tilde{a}_{kj}\,C_{kj}(\tilde{\mathbf{A}}).$$

Here $\tilde{a}_{kj} = a_{ij}$. The minor $M_{kj}(\tilde{\mathbf{A}})$ is computed from $\tilde{\mathbf{A}}$ with row k deleted, and outside row k the matrices $\tilde{\mathbf{A}}$ and $\mathbf{A}$ agree; so $C_{kj}(\tilde{\mathbf{A}}) = C_{kj}(\mathbf{A})$. Hence $\sum_j a_{ij}C_{kj} = 0$. Therefore $\mathbf{A}\operatorname{adj}\mathbf{A} = (\det\mathbf{A})\mathbf{I}_n$.

The product $\operatorname{adj}\mathbf{A}\,\mathbf{A}$. Its (i,k)-entry is $\sum_j C_{ji}\,a_{jk}$. If $i = k$, this is $\det\mathbf{A}$ by Theorem 7.6.1 (a) along column i. If $i \neq k$, let $\hat{\mathbf{A}}$ be $\mathbf{A}$ with column i replaced by a copy of column k. It has two equal columns, so $\det\hat{\mathbf{A}} = 0$ by Theorem 7.4.3. Expanding along column i, $0 = \sum_j \hat{a}_{ji}\,C_{ji}(\hat{\mathbf{A}}) = \sum_j a_{jk}\,C_{ji}(\mathbf{A})$, since deleting column i removes the only column where $\hat{\mathbf{A}}$ and $\mathbf{A}$ differ. Therefore $\operatorname{adj}\mathbf{A}\,\mathbf{A} = (\det\mathbf{A})\mathbf{I}_n$. This proves the theorem.

■

Remark. **Over a commutative ring.** This proof uses only Theorem 7.6.1 and the alternating property, so the adjugate identity holds for $\mathbf{A} \in M_n(R)$ over any commutative ring R. For $R = F[x]$, the matrix $\operatorname{adj}(x\mathbf{I} - \mathbf{A})$ has polynomial entries and $(x\mathbf{I} - \mathbf{A})\operatorname{adj}(x\mathbf{I} - \mathbf{A}) = \det(x\mathbf{I} - \mathbf{A})\,\mathbf{I}$. This identity reappears in Chapter 9. Over R the identity also shows that $\mathbf{A}$ has an inverse in $M_n(R)$ as soon as $\det\mathbf{A}$ has an inverse in R, namely $(\det\mathbf{A})^{-1}\operatorname{adj}\mathbf{A}$.

Dividing by the determinant, when we may, gives the inverse.

Corollary 7.6.3. Inverse via the Adjugate

Let $\mathbf{A} \in M_n(F)$ with $\det\mathbf{A} \neq 0$. Then $\mathbf{A}$ is invertible and

$$\mathbf{A}^{-1} = \frac{1}{\det\mathbf{A}}\operatorname{adj}\mathbf{A}.$$

Proof. By Theorem 7.6.2, $\mathbf{A}\big((\det\mathbf{A})^{-1}\operatorname{adj}\mathbf{A}\big) = (\det\mathbf{A})^{-1}(\det\mathbf{A})\mathbf{I}_n = \mathbf{I}_n$, using that scalars commute with matrices (Theorem 0.9.4). By Theorem 3.4.5, $\mathbf{A}$ is invertible with inverse $(\det\mathbf{A})^{-1}\operatorname{adj}\mathbf{A}$.

■

For $n = 2$ this is Theorem 0.9.8 again. For the matrix $\mathbf{B}$ above, $\mathbf{B}^{-1} = \frac{1}{5}\begin{pmatrix} 3 & -6 & 2 \\ 1 & 3 & -1 \\ -1 & 2 & 1 \end{pmatrix}$.

Warning. **The adjugate exists for singular matrices too.** Nothing in Definition 7.6.2 needs $\det \mathbf{A} \neq 0$. For $S = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \\ 1 & 0 & 1 \end{pmatrix}$, with $\det S = 0$, one computes $\operatorname{adj} S = \begin{pmatrix} 4 & -2 & 0 \\ 4 & -2 & 0 \\ -4 & 2 & 0 \end{pmatrix} \neq \mathbf{O}$, and $S \operatorname{adj} S = \mathbf{O}$, as Theorem 7.6.2 predicts. So "$\operatorname{adj} \mathbf{A}$ exists" says nothing about invertibility, and the formula $\mathbf{A}^{-1} = \operatorname{adj} \mathbf{A} / \det \mathbf{A}$ must never be used before checking $\det \mathbf{A} \neq 0$.

Quick check. What are $\operatorname{adj}(c\mathbf{I}_n)$ for $c \in F$ and $n \geq 2$, and $\operatorname{adj} \mathbf{O}$ for the zero matrix in $M_n(F)$, $n \geq 2$? What is $\operatorname{adj} \mathbf{O}$ for $n = 1$?

Solution. Deleting row i and column i from $c\mathbf{I}_n$ leaves $c\mathbf{I}_{n-1}$, of determinant c^{n-1}; deleting row i and a different column leaves a zero column, of determinant 0. So $\operatorname{adj}(c\mathbf{I}_n) = c^{n-1}\mathbf{I}_n$; in particular $\operatorname{adj} \mathbf{O} = \mathbf{O}$ for $n \geq 2$, since $0^{n-1} = 0$. For $n = 1$, $\operatorname{adj} \mathbf{O}_1 = (1)$ by definition, and indeed $\mathbf{O}_1(1) = \mathbf{O}_1 = (\det \mathbf{O}_1)\mathbf{I}_1$.

7.6.4 Cramer's rule

For an invertible $\mathbf{A}$, the system $\mathbf{A}\mathbf{x} = \mathbf{b}$ has exactly one solution. The adjugate expresses it in closed form, and the closed form has a direct proof by linearity that shows where it comes from. For $k \in \{1, \ldots, n\}$, write $\mathbf{A}_k(\mathbf{b})$ for the matrix obtained from $\mathbf{A}$ by replacing column k with $\mathbf{b}$.

Theorem 7.6.4. Cramer's Rule

Let $\mathbf{A} \in M_n(F)$ be invertible and $\mathbf{b} \in F^n$. The unique solution $\mathbf{x}$ of $\mathbf{A}\mathbf{x} = \mathbf{b}$ is given by

$$x_k = \frac{\det \mathbf{A}_k(\mathbf{b})}{\det \mathbf{A}}, \qquad k = 1, \ldots, n.$$

Idea. If $\mathbf{A}\mathbf{x} = \mathbf{b}$, then $\mathbf{b} = x_1\mathbf{a}_1 + \cdots + x_n\mathbf{a}_n$. Put this combination into column k and expand by linearity. Every term except the k-th has the column $\mathbf{a}_j$ twice, so it vanishes, and what remains is $x_k \det \mathbf{A}$.

Proof. By Theorem 0.9.7 (7), $\mathbf{A}\mathbf{x} = \mathbf{b}$ has exactly one solution $\mathbf{x}$. By Theorem 0.9.3, $\mathbf{b} =$

$\sum_{j=1}^{n} x_j \mathbf{a}_j$, where $\mathbf{a}_j$ is the j-th column of $\mathbf{A}$. Since det is linear in column k (Theorem 7.4.3),

$$\det \mathbf{A}_k(\mathbf{b}) = \sum_{j=1}^{n} x_j \det \mathbf{A}_k(\mathbf{a}_j).$$

For $j \neq k$, the matrix $\mathbf{A}_k(\mathbf{a}_j)$ has the column $\mathbf{a}_j$ in positions j and k, so its determinant is 0. For $j = k$, $\mathbf{A}_k(\mathbf{a}_k) = \mathbf{A}$. Hence $\det \mathbf{A}_k(\mathbf{b}) = x_k \det \mathbf{A}$. Since $\mathbf{A}$ is invertible, $\det \mathbf{A} \neq 0$ by Theorem 7.5.2, and dividing gives the formula.

■

Alternatively, $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b} = (\det \mathbf{A})^{-1} \operatorname{adj} \mathbf{A}\, \mathbf{b}$ by Corollary 7.6.3, and the k-th entry of $\operatorname{adj} \mathbf{A}\, \mathbf{b}$ is $\sum_j C_{jk} b_j$, which is the expansion of $\det \mathbf{A}_k(\mathbf{b})$ along column k (Theorem 7.6.1 (a)); the cofactors along column k of $\mathbf{A}_k(\mathbf{b})$ and of $\mathbf{A}$ agree, because computing them deletes column k.

For example, the system $2x+y = 5$, $7x+4y = 18$ has coefficient matrix of determinant $8-7 = 1$, so

$$x = \det \begin{pmatrix} 5 & 1 \\ 18 & 4 \end{pmatrix} = 2, \qquad y = \det \begin{pmatrix} 2 & 5 \\ 7 & 18 \end{pmatrix} = 1,$$

and indeed $2 \cdot 2 + 1 = 5$ and $7 \cdot 2 + 4 = 18$.

Warning. **Cramer's rule is a theoretical tool, not an algorithm.** It needs $n + 1$ determinants of size n. Evaluated by the Leibniz formula this is about $(n + 1)!$ products; even with each determinant computed by elimination it costs about $n^4/3$ operations, while Gaussian elimination solves the system once in about $n^3/3$. For $n = 20$, term-by-term Leibniz evaluation at 10^9 multiplications per second would take tens of thousands of years. Use Cramer's rule to understand **how** the solution depends on $\mathbf{A}$ and $\mathbf{b}$, and elimination to find it.

7.6.5 Integer matrices

The adjugate formula shows that each entry of $\mathbf{A}^{-1}$ is a polynomial in the entries of $\mathbf{A}$ divided by the single polynomial $\det \mathbf{A}$. In particular, the entries of $\mathbf{A}^{-1}$ change continuously with those of $\mathbf{A}$ over $\mathbb{R}$, as long as $\det \mathbf{A} \neq 0$. Here is an arithmetic consequence. A matrix with integer entries lies in $M_n(\mathbb{Q})$, and we may ask when its inverse again has integer entries.

Corollary 7.6.5. Integer Matrices with Integer Inverses

Let $\mathbf{A} \in M_n(\mathbb{Q})$ have all entries in $\mathbb{Z}$.

(a) $\det \mathbf{A} \in \mathbb{Z}$, and $\operatorname{adj} \mathbf{A}$ has all entries in $\mathbb{Z}$.

(b) $\mathbf{A}$ is invertible and $\mathbf{A}^{-1}$ has all entries in $\mathbb{Z}$ if and only if $\det \mathbf{A} = 1$ or $\det \mathbf{A} = -1$.

Proof.

(a) The Leibniz formula (Definition 7.4.1) is a sum of products of entries with signs ± 1, so the determinant of a matrix with integer entries is an integer. Each cofactor is $\pm$ such a determinant, so $\operatorname{adj} \mathbf{A}$ has integer entries (for $n = 1$, $\operatorname{adj} \mathbf{A} = (1)$).

(b) ($\Leftarrow$) If $\det \mathbf{A} = \pm 1$, then $\mathbf{A}$ is invertible and $\mathbf{A}^{-1} = \pm \operatorname{adj} \mathbf{A}$ by Corollary 7.6.3, which has integer entries by (a). ($\Rightarrow$) Suppose $\mathbf{A}^{-1}$ has integer entries. By Theorem 7.5.4, $\det \mathbf{A} \det(\mathbf{A}^{-1}) = \det \mathbf{I}_n = 1$, and both factors are integers by (a). The only integers whose product is 1 are $1 \cdot 1$ and $(-1)(-1)$, so $\det \mathbf{A} = \pm 1$.

■

For example, $\begin{pmatrix} 2 & 1 \\ 5 & 3 \end{pmatrix}$ has determinant 1 and inverse $\begin{pmatrix} 3 & -1 \\ -5 & 2 \end{pmatrix}$, while $\begin{pmatrix} 2 & 1 \\ 0 & 1 \end{pmatrix}$ has determinant 2 and inverse $\begin{pmatrix} 1/2 & -1/2 \\ 0 & 1 \end{pmatrix}$. In the language of the remark after Theorem 7.6.2, part (b) says that a matrix over the ring $\mathbb{Z}$ is invertible over $\mathbb{Z}$ exactly when its determinant is invertible in $\mathbb{Z}$.

7.6.6 Exercises

7.6.6.1 A. Check your understanding

Exercise. A1

(a) Define the (i,j)-minor, the (i,j)-cofactor, and the adjugate of $\mathbf{A} \in M_n(F)$, $n \geq 2$.

(b) What are the signs $(-1)^{i+j}$ of the cofactors in positions $(2,3)$ and $(4,4)$?

(c) True or false: $\operatorname{adj} \mathbf{A}$ is defined only when $\mathbf{A}$ is invertible. Justify your answer.

(d) True or false: Cramer's rule (Theorem 7.6.4) gives the solutions of every square linear system $\mathbf{Ax} = \mathbf{b}$. Justify your answer.

(e) In the proof of Theorem 7.6.2, which matrix is expanded to show that $\sum_j a_{ij} C_{kj} = 0$ for $i \neq k$, and why is its determinant 0?

(f) True or false: if $\mathbf{A}$ has integer entries and $\det \mathbf{A} = 2$, then $\mathbf{A}^{-1}$ has integer entries. Justify your answer.

Solution.

(a) M_{ij} is the determinant of the $(n-1) \times (n-1)$ matrix obtained by deleting row i and column j; $C_{ij} = (-1)^{i+j} M_{ij}$; $\operatorname{adj} \mathbf{A}$ is the matrix with (i,j)-entry C_{ji} (Definition 7.6.1, Definition 7.6.2).

(b) $(-1)^5 = -1$ and $(-1)^8 = 1$.

(c) False. It is defined for every square matrix; for instance $\operatorname{adj} \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$, although the matrix is singular.

(d) False. It requires $\det \mathbf{A} \neq 0$. For $\mathbf{A} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$ and $\mathbf{b} = (1,1)$, the system has infinitely many solutions, and the formula would divide by 0.

(e) The matrix $\tilde{\mathbf{A}}$ obtained from $\mathbf{A}$ by replacing row k with row i, expanded along row k. It has two equal rows, i and k, so its determinant is 0.

(f) False. By Corollary 7.6.5, an integer inverse forces $\det \mathbf{A} = \pm 1$. For instance $\operatorname{diag}(2,1)$ has inverse $\operatorname{diag}(\tfrac{1}{2}, 1)$.

7.6.6.2 B. Practice

Exercise. B1: Adjugate and inverse

Let $\mathbf{A} = \begin{pmatrix} 1 & 2 & -1 \\ 0 & 1 & 3 \\ 2 & 1 & 1 \end{pmatrix} \in M_3(\mathbb{R})$. Compute all nine cofactors, $\operatorname{adj}\mathbf{A}$ and $\det\mathbf{A}$. Verify that $\mathbf{A}\operatorname{adj}\mathbf{A} = (\det\mathbf{A})\mathbf{I}_3$, and hence write down $\mathbf{A}^{-1}$.

Solution. The cofactors are

$$\begin{aligned} C_{11} &= +(1\cdot 1 - 3\cdot 1) = -2, & C_{12} &= -(0\cdot 1 - 3\cdot 2) = 6, \\ C_{13} &= +(0\cdot 1 - 1\cdot 2) = -2, & C_{21} &= -(2\cdot 1 - (-1)\cdot 1) = -3, \\ C_{22} &= +(1\cdot 1 - (-1)\cdot 2) = 3, & C_{23} &= -(1\cdot 1 - 2\cdot 2) = 3, \\ C_{31} &= +(2\cdot 3 - (-1)\cdot 1) = 7, & C_{32} &= -(1\cdot 3 - (-1)\cdot 0) = -3, \\ C_{33} &= +(1\cdot 1 - 2\cdot 0) = 1. \end{aligned}$$

Transposing the cofactor matrix (Definition 7.6.2),

$$\operatorname{adj}\mathbf{A} = \begin{pmatrix} -2 & -3 & 7 \\ 6 & 3 & -3 \\ -2 & 3 & 1 \end{pmatrix}.$$

Expanding along row 1 (Theorem 7.6.1), $\det\mathbf{A} = 1\cdot(-2) + 2\cdot 6 + (-1)(-2) = 12$. Row 1 of $\mathbf{A}$ against the columns of $\operatorname{adj}\mathbf{A}$ gives $-2+12+2 = 12$, $-3+6-3 = 0$, $7-6-1 = 0$; row 2 gives $0+6-6 = 0$, $0+3+9 = 12$, $0-3+3 = 0$; row 3 gives $-4+6-2 = 0$, $-6+3+3 = 0$, $14-3+1 = 12$. So $\mathbf{A}\operatorname{adj}\mathbf{A} = 12\mathbf{I}_3$. By Corollary 7.6.3,

$$\mathbf{A}^{-1} = \frac{1}{12}\begin{pmatrix} -2 & -3 & 7 \\ 6 & 3 & -3 \\ -2 & 3 & 1 \end{pmatrix}.$$

Exercise. B2: Cramer's rule

Use Cramer's rule to solve

$$x_1 + 2x_2 - x_3 = -3, \qquad x_2 + 3x_3 = 5, \qquad 2x_1 + x_2 + x_3 = 3$$

over $\mathbb{R}$. Check your answer by substitution.

Solution. The coefficient matrix is the matrix $\mathbf{A}$ of Exercise (B1: Adjugate and inverse),

with $\det \mathbf{A} = 12 \neq 0$, and $\mathbf{b} = (-3, 5, 3)$. Expanding each $\mathbf{A}_k(\mathbf{b})$ along its first row,

$$\begin{aligned}
\det \mathbf{A}_1(\mathbf{b}) &= \det \begin{pmatrix} -3 & 2 & -1 \\ 5 & 1 & 3 \\ 3 & 1 & 1 \end{pmatrix} \\
&= -3(1-3) - 2(5-9) - 1(5-3) = 6 + 8 - 2 = 12, \\
\det \mathbf{A}_2(\mathbf{b}) &= \det \begin{pmatrix} 1 & -3 & -1 \\ 0 & 5 & 3 \\ 2 & 3 & 1 \end{pmatrix} \\
&= 1(5-9) + 3(0-6) - 1(0-10) = -4 - 18 + 10 = -12, \\
\det \mathbf{A}_3(\mathbf{b}) &= \det \begin{pmatrix} 1 & 2 & -3 \\ 0 & 1 & 5 \\ 2 & 1 & 3 \end{pmatrix} \\
&= 1(3-5) - 2(0-10) - 3(0-2) = -2 + 20 + 6 = 24.
\end{aligned}$$

By Theorem 7.6.4, $x_1 = 12/12 = 1$, $x_2 = -12/12 = -1$, $x_3 = 24/12 = 2$. Check: $1 - 2 - 2 = -3$, $-1 + 6 = 5$, $2 - 1 + 2 = 3$.

Exercise. B3: Choosing the best line

Compute $\det \mathbf{N}$ for

$$\mathbf{N} = \begin{pmatrix} 2 & 0 & 0 & 3 \\ 1 & 4 & 0 & -1 \\ 0 & 5 & 0 & 2 \\ 3 & 1 & 2 & 0 \end{pmatrix} \in M_4(\mathbb{R}),$$

expanding along the row or column that needs the least work.

Solution. Column 3 has a single nonzero entry, $a_{43} = 2$, whose sign is $(-1)^{4+3} = -1$. By Theorem 7.6.1 (a) along column 3, $\det \mathbf{N} = -2M_{43}$, where

$$\begin{aligned}
M_{43} &= \det \begin{pmatrix} 2 & 0 & 3 \\ 1 & 4 & -1 \\ 0 & 5 & 2 \end{pmatrix} \\
&= 2(4 \cdot 2 - (-1) \cdot 5) - 0 + 3(1 \cdot 5 - 4 \cdot 0) = 26 + 15 = 41,
\end{aligned}$$

expanding along the first row. Hence $\det \mathbf{N} = -82$.

7.6.6.3 C. Going deeper

Exercise. C1: The determinant of the adjugate

Let $n \geq 2$ and $\mathbf{A} \in M_n(F)$. Prove that $\det(\operatorname{adj} \mathbf{A}) = (\det \mathbf{A})^{n-1}$.

Hint: treat $\det \mathbf{A} \neq 0$ *and* $\det \mathbf{A} = 0$ *separately; in the second case, ask whether* $\operatorname{adj} \mathbf{A}$ *can be invertible.*

Solution. Case 1: $\det \mathbf{A} \neq 0$. By Theorem 7.6.2 and Theorem 7.5.4, $\det \mathbf{A} \det(\operatorname{adj} \mathbf{A}) = \det\left((\det \mathbf{A})\mathbf{I}_n\right) = (\det \mathbf{A})^n$, where the last step is Theorem 7.4.6. Dividing by $\det \mathbf{A} \neq 0$ gives $\det(\operatorname{adj} \mathbf{A}) = (\det \mathbf{A})^{n-1}$.

Case 2: $\det \mathbf{A} = 0$. Then $\mathbf{A} \operatorname{adj} \mathbf{A} = \mathbf{O}$ by Theorem 7.6.2. Suppose, for a contradiction, that $\operatorname{adj} \mathbf{A}$ is invertible. Then $\mathbf{A} = (\mathbf{A} \operatorname{adj} \mathbf{A})(\operatorname{adj} \mathbf{A})^{-1} = \mathbf{O}$. But every minor of the zero matrix is the determinant of a zero matrix of size $n-1 \geq 1$, which is 0, so $\operatorname{adj} \mathbf{O} = \mathbf{O}$, which is not invertible. This contradiction shows that $\operatorname{adj} \mathbf{A}$ is not invertible, so $\det(\operatorname{adj} \mathbf{A}) = 0 = (\det \mathbf{A})^{n-1}$ by Theorem 7.5.2, since $n - 1 \geq 1$.

Exercise. C2: The rank of the adjugate

Let $n \geq 2$ and $\mathbf{A} \in M_n(F)$. Prove that

$$\operatorname{rank}(\operatorname{adj} \mathbf{A}) = \begin{cases} n & \text{if } \operatorname{rank} \mathbf{A} = n, \\ 1 & \text{if } \operatorname{rank} \mathbf{A} = n-1, \\ 0 & \text{if } \operatorname{rank} \mathbf{A} \leq n-2. \end{cases}$$

(a) Prove the case $\operatorname{rank} \mathbf{A} = n$.

(b) Suppose $\operatorname{rank} \mathbf{A} \leq n-2$. Prove that every minor M_{ij} is 0.

(c) Suppose $\operatorname{rank} \mathbf{A} = n-1$. Prove that every column of $\operatorname{adj} \mathbf{A}$ lies in $\operatorname{nul}(\mathbf{A})$, and deduce $\operatorname{rank}(\operatorname{adj} \mathbf{A}) \leq 1$.

(d) Suppose $\operatorname{rank} \mathbf{A} = n-1$. Prove that some minor M_{ij} is nonzero, and conclude.

Hint for (d): choose $n-1$ independent columns of $\mathbf{A}$, then $n-1$ independent rows of the resulting $n \times (n-1)$ matrix.

Solution.

(a) If $\operatorname{rank} \mathbf{A} = n$, then $\mathbf{A}$ is invertible by Theorem 7.5.3, so $\det \mathbf{A} \neq 0$ and $\operatorname{adj} \mathbf{A} = (\det \mathbf{A})\mathbf{A}^{-1}$ by Corollary 7.6.3. This is invertible, since $\mathbf{A}^{-1}$ is and $\det \mathbf{A} \neq 0$ (Theorem 0.9.7), so $\operatorname{rank}(\operatorname{adj} \mathbf{A}) = n$.

(b) Fix i, j, and let $\mathbf{A}'$ be $\mathbf{A}$ with row i and column j deleted. The $n-1$ columns of $\mathbf{A}$ other than column j lie in $\operatorname{col}(\mathbf{A})$, of dimension $\operatorname{rank} \mathbf{A} \leq n-2$, so they are linearly dependent by Theorem 1.6.3: $\sum_{k \neq j} c_k \mathbf{a}_k = \mathbf{0}$ with the c_k not all zero. Deleting the i-th entry of every vector preserves this relation, so the columns of $\mathbf{A}'$ are dependent. By Theorem 3.4.4, $\mathbf{A}'$ is not invertible, and $M_{ij} = \det \mathbf{A}' = 0$ by Theorem 7.5.2. Hence $\operatorname{adj} \mathbf{A} = \mathbf{O}$ and $\operatorname{rank}(\operatorname{adj} \mathbf{A}) = 0$.

(c) Since $\operatorname{rank} \mathbf{A} < n$, $\det \mathbf{A} = 0$ by Theorem 7.5.3, so $\mathbf{A} \operatorname{adj} \mathbf{A} = \mathbf{O}$ by Theorem 7.6.2. By Theorem 0.9.2, $\mathbf{A}$ kills each column of $\operatorname{adj} \mathbf{A}$, so $\operatorname{col}(\operatorname{adj} \mathbf{A}) \subseteq \operatorname{nul}(\mathbf{A})$. By Theorem 3.5.7, $\dim \operatorname{nul}(\mathbf{A}) = n - (n-1) = 1$. Hence $\operatorname{rank}(\operatorname{adj} \mathbf{A}) \leq 1$.

(d) By Theorem 3.5.2, the pivot columns of $\mathbf{A}$ form a basis of $\operatorname{col}(\mathbf{A})$; there are $n-1$ of them, so exactly one column index j is not a pivot column. Let $\mathbf{A}_0 \in M_{n \times (n-1)}(F)$ consist of the other $n-1$ columns. They are independent, so $\operatorname{rank} \mathbf{A}_0 = n-1$, and $\dim \operatorname{row}(\mathbf{A}_0) = n-1$ by Theorem 3.5.5. The n rows of $\mathbf{A}_0$ span $\operatorname{row}(\mathbf{A}_0)$, so by Theorem 1.5.4 some $n-1$ of them are independent; since $n-1$ independent vectors in a space of dimension $n-1$ form a basis, sifting leaves out exactly one row, say row i. Deleting row i from $\mathbf{A}_0$ gives an $(n-1) \times (n-1)$ matrix $\mathbf{A}'$ with independent rows,

so $\det \mathbf{A}' \neq 0$ by Theorem 7.5.3. But $\mathbf{A}'$ is $\mathbf{A}$ with row i and column j deleted, so $M_{ij} \neq 0$, hence $C_{ij} \neq 0$ and $\operatorname{adj} \mathbf{A} \neq \mathbf{O}$. With (c), $\operatorname{rank}(\operatorname{adj} \mathbf{A}) = 1$.

Exercise. C3: An adjugate over $F[x]$

Work in $M_2(\mathbb{R}[x])$, using the remarks after Theorem 7.6.1 and Theorem 7.6.2.

(a) For $\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$, compute $\det(x\mathbf{I} - \mathbf{A})$ and $\operatorname{adj}(x\mathbf{I} - \mathbf{A})$, and verify $(x\mathbf{I} - \mathbf{A}) \operatorname{adj}(x\mathbf{I} - \mathbf{A}) = \det(x\mathbf{I} - \mathbf{A})\, \mathbf{I}$ by multiplying out.

(b) For a general $\mathbf{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in M_2(\mathbb{R})$, show that $\operatorname{adj}(x\mathbf{I} - \mathbf{A}) = x\mathbf{I} - \operatorname{adj} \mathbf{A}$ and $\mathbf{A} + \operatorname{adj} \mathbf{A} = (\operatorname{tr} \mathbf{A})\mathbf{I}$.

(c) Deduce from (b), without using Theorem 7.6.2 over $\mathbb{R}[x]$, that $(x\mathbf{I} - \mathbf{A}) \operatorname{adj}(x\mathbf{I} - \mathbf{A}) = (x^2 - (\operatorname{tr} \mathbf{A})x + \det \mathbf{A})\mathbf{I}$, and check that $x^2 - (\operatorname{tr} \mathbf{A})x + \det \mathbf{A} = \det(x\mathbf{I} - \mathbf{A})$.

Solution.

(a) $x\mathbf{I} - \mathbf{A} = \begin{pmatrix} x-1 & -2 \\ -3 & x-4 \end{pmatrix}$, so $\det(x\mathbf{I} - \mathbf{A}) = (x-1)(x-4) - 6 = x^2 - 5x - 2$, and by Definition 7.6.2, $\operatorname{adj}(x\mathbf{I} - \mathbf{A}) = \begin{pmatrix} x-4 & 2 \\ 3 & x-1 \end{pmatrix}$. The product has entries $(x-1)(x-4) - 6 = x^2 - 5x - 2$, $2(x-1) - 2(x-1) = 0$, $-3(x-4) + 3(x-4) = 0$ and $-6 + (x-4)(x-1) = x^2 - 5x - 2$, so it is $(x^2 - 5x - 2)\mathbf{I}$.

(b) $x\mathbf{I} - \mathbf{A} = \begin{pmatrix} x-a & -b \\ -c & x-d \end{pmatrix}$, so $\operatorname{adj}(x\mathbf{I} - \mathbf{A}) = \begin{pmatrix} x-d & b \\ c & x-a \end{pmatrix} = x\mathbf{I} - \begin{pmatrix} d & -b \\ -c & a \end{pmatrix} = x\mathbf{I} - \operatorname{adj} \mathbf{A}$. Also $\mathbf{A} + \operatorname{adj} \mathbf{A} = \begin{pmatrix} a+d & 0 \\ 0 & a+d \end{pmatrix} = (\operatorname{tr} \mathbf{A})\mathbf{I}$.

(c) Since $x\mathbf{I}$ commutes with every matrix, expanding with (b) gives

$$\begin{aligned}(x\mathbf{I} - \mathbf{A})(x\mathbf{I} - \operatorname{adj} \mathbf{A}) &= x^2\mathbf{I} - x(\mathbf{A} + \operatorname{adj} \mathbf{A}) + \mathbf{A} \operatorname{adj} \mathbf{A} \\ &= x^2\mathbf{I} - (\operatorname{tr} \mathbf{A})x\mathbf{I} + (ad - bc)\mathbf{I},\end{aligned}$$

where $\mathbf{A} \operatorname{adj} \mathbf{A} = (ad - bc)\mathbf{I}$ is the computation of Theorem 0.9.8 over $\mathbb{R}$. Finally $\det(x\mathbf{I} - \mathbf{A}) = (x-a)(x-d) - bc = x^2 - (a+d)x + (ad - bc)$, which is $x^2 - (\operatorname{tr} \mathbf{A})x + \det \mathbf{A}$.

7.7 Minors, Rank, and Cauchy–Binet

The determinant of a square matrix answers a yes-or-no question: is the matrix invertible? A rectangular matrix has no determinant, and a singular square matrix has determinant 0, which says nothing about *how* singular it is. Yet every matrix is full of square pieces, and each piece has a determinant. This section shows that these small determinants measure the rank exactly. It then answers a second question left open by multiplicativity: if $\mathbf{A}$ is $m \times n$ and $\mathbf{B}$ is $n \times m$, the product $\mathbf{AB}$ is square, so what is $\det(\mathbf{AB})$? The answer, the Cauchy–Binet formula, is proved by the same multilinear expansion that gave uniqueness of the determinant in Theorem 7.4.2.

7.7.1 Submatrices and minors

In Definition 7.6.1 we deleted one row and one column of a square matrix and took the determinant of what was left. That is a special case of a more flexible operation: keep **any** k rows and **any** k columns. The recurring object is the square block sitting at the crossings of the chosen rows and columns, and it deserves a name.

A minor is the determinant of a square block cut out of a matrix by choosing some rows and the same number of columns.

For a positive integer m, write $[m] = \{1, 2, \dots, m\}$.

Definition 7.7.1. Submatrix and Minor

Let $\mathbf{A} = (a_{ij}) \in M_{m\times n}(F)$, and let $1 \le k \le \min(m, n)$. Let $I = \{i_1 < i_2 < \cdots < i_k\} \subseteq [m]$ and $J = \{j_1 < j_2 < \cdots < j_k\} \subseteq [n]$ be sets with **exactly** k elements each, listed in **increasing** order. The **submatrix** of $\mathbf{A}$ on rows I and columns J is

$$\mathbf{A}_{I,J} := \left(a_{i_p j_q}\right)_{p,q=1}^{k} \in M_k(F),$$

the $k \times k$ matrix whose (p, q)-entry is $a_{i_p j_q}$. Its determinant $\det \mathbf{A}_{I,J}$ is a $k \times k$ **minor** of $\mathbf{A}$.

In words: cross out every row not in I and every column not in J, and push the surviving entries together **without changing their order**. The chosen rows need not be adjacent, and neither need the columns. The sets I and J have the **same** size, so the block is square and has a determinant.

A 1×1 minor is a single entry, $\det \mathbf{A}_{\{i\},\{j\}} = a_{ij}$. For a square $\mathbf{A} \in M_n(F)$ with $n \ge 2$, the minor M_{ij} of Definition 7.6.1 is the $(n-1) \times (n-1)$ minor with $I = [n] \setminus \{i\}$ and $J = [n] \setminus \{j\}$. The only $n \times n$ minor of a square matrix is $\det \mathbf{A}$ itself. An $m \times n$ matrix has $\binom{m}{k}\binom{n}{k}$ minors of size k, one for each choice of I and J.

Example. Let

$$\mathbf{A} = \begin{pmatrix} 1 & 2 & 0 \\ 0 & 1 & 3 \end{pmatrix} \in M_{2\times 3}(\mathbb{Q}).$$

Its 2×2 minors all use $I = \{1, 2\}$, and there are three choices of J:

$$\det \mathbf{A}_{I,\{1,2\}} = \det \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix} = 1,$$

$$\det \mathbf{A}_{I,\{1,3\}} = \det \begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix} = 3,$$

$$\det \mathbf{A}_{I,\{2,3\}} = \det \begin{pmatrix} 2 & 0 \\ 1 & 3 \end{pmatrix} = 6.$$

Its six 1×1 minors are its entries. It has no 3×3 minors, since it has only two rows.

A degenerate example. Every minor of the zero matrix $\mathbf{O}_{m\times n}$ is 0: each block is a zero matrix, whose determinant is 0 by Theorem 7.4.6. This case matters because it is exactly the matrix of rank 0, and the theorem below has to treat it separately.

Nonexample by minimal change. Take the same rows and columns but list the columns of $\mathbf{A}$ above in the order $3, 1$: the block $\begin{pmatrix} 0 & 1 \\ 3 & 0 \end{pmatrix}$ has determinant -3, not 3. Every entry is still an

entry of $\mathbf{A}$ from rows $\{1,2\}$ and columns $\{1,3\}$; what fails is the clause **listed in increasing order**. Reordering columns changes the determinant by a sign (Theorem 7.2.1), so the order convention is what makes "the" minor on (I,J) a single well-defined number.

Why this definition. Insisting on increasing order costs nothing and removes all sign ambiguity. Insisting on $|I| = |J|$ is forced: only square blocks have determinants.

Warning. **Rank is not read off principal minors.** A **principal** minor uses the same index set for rows and columns ($I = J$). It is tempting to test rank using only these, but they can all vanish for a nonzero matrix. For $\mathbf{N} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$, the principal minors are 0, 0, and $\det \mathbf{N} = 0$, yet $\operatorname{rank} \mathbf{N} = 1$. The nonzero minor is the off-diagonal entry $\det \mathbf{N}_{\{1\},\{2\}} = 1$. The theorem below uses **all** minors.

Quick check. How many 2×2 minors does a 3×4 matrix have? How many 3×3 minors?

Solution. A 2×2 minor needs 2 of the 3 rows and 2 of the 4 columns: $\binom{3}{2}\binom{4}{2} = 3 \cdot 6 = 18$. A 3×3 minor uses all three rows and 3 of the 4 columns: $\binom{3}{3}\binom{4}{3} = 4$.

7.7.2 Rank via minors

A nonzero determinant of a square block means the block is invertible, so its columns are independent. Independent columns inside a block come from independent columns of the whole matrix. So a large nonzero minor forces large rank. The converse, that large rank produces a large nonzero minor, needs us to find the right rows as well as the right columns. Both directions together give a description of rank using only determinants:

Theorem 7.7.1. Rank via Minors

Let $\mathbf{A} \in M_{m\times n}(F)$ be **nonzero**.

(a) $\operatorname{rank} \mathbf{A}$ is the **largest** integer k such that $\mathbf{A}$ has a **nonzero** $k \times k$ minor.

(b) Equivalently, for each $1 \le k \le \min(m,n)$: $\operatorname{rank} \mathbf{A} \ge k$ if and only if some $k \times k$ minor of $\mathbf{A}$ is nonzero; and $\operatorname{rank} \mathbf{A} < k$ if and only if **every** $k \times k$ minor of $\mathbf{A}$ is zero.

Idea. Two inequalities. ① A nonzero $k \times k$ minor on rows I and columns J makes $\mathbf{A}_{I,J}$ invertible; a dependence among the columns of $\mathbf{A}$ indexed by J would restrict to a dependence among the columns of $\mathbf{A}_{I,J}$. So $\operatorname{rank} \mathbf{A} \ge k$. ② If $\operatorname{rank} \mathbf{A} = r$, choose r independent columns (pivot columns). The $m \times r$ matrix they form has rank r, so by row rank = column rank it has r independent rows. Those rows cut out an $r \times r$ block with independent rows, hence a nonzero determinant.

Proof. **Step 1: a nonzero $k \times k$ minor forces $\operatorname{rank} \mathbf{A} \ge k$.** Let $I = \{i_1 < \cdots < i_k\}$ and $J = \{j_1 < \cdots < j_k\}$ with $\det \mathbf{A}_{I,J} \neq 0$. Write $\mathbf{a}_1, \ldots, \mathbf{a}_n$ for the columns of $\mathbf{A}$. Suppose

$c_1\mathbf{a}_{j_1}+\cdots+c_k\mathbf{a}_{j_k}=\mathbf{0}$ with $c_1,\dots,c_k\in F$. Reading only the entries in rows $i_1,\dots,i_k$ of this vector equation gives $\sum_q c_q\,a_{i_p j_q}=0$ for $p=1,\dots,k$, that is, $\mathbf{A}_{I,J}\mathbf{c}=\mathbf{0}$ with $\mathbf{c}=(c_1,\dots,c_k)$. By Theorem 7.5.2, $\mathbf{A}_{I,J}$ is invertible, so $\mathbf{c}=\mathbf{0}$ by Theorem 3.4.4. Hence $(\mathbf{a}_{j_1},\dots,\mathbf{a}_{j_k})$ is a linearly independent list in $\operatorname{col}(\mathbf{A})$, and by Theorem 1.6.3 (a), $k\le\dim\operatorname{col}(\mathbf{A})=\operatorname{rank}\mathbf{A}$.

Step 2: A has a nonzero $r\times r$ minor, where $r=\operatorname{rank}\mathbf{A}$. Since $\mathbf{A}\neq\mathbf{O}$, some column is nonzero, so $r\ge 1$. By Theorem 3.5.2 (a), the pivot columns $\mathbf{a}_{j_1},\dots,\mathbf{a}_{j_r}$ $(j_1<\cdots<j_r)$ are linearly independent. Let $J=\{j_1,\dots,j_r\}$ and let $\mathbf{A}'=\mathbf{A}_{[m],J}\in M_{m\times r}(F)$ be the matrix with these columns. By Theorem 3.5.10 (a), $\operatorname{rank}\mathbf{A}'=r$, so by Theorem 3.5.5 the row space of $\mathbf{A}'$ has dimension r. The m rows of $\mathbf{A}'$ span its row space, so by Theorem 1.5.4 some r of them, in rows $i_1<\cdots<i_r$, form a basis of it; in particular they are linearly independent. These are exactly the rows of $\mathbf{B}=\mathbf{A}_{I,J}\in M_r(F)$, where $I=\{i_1,\dots,i_r\}$. Transposition is a linear bijection from $1\times r$ rows to columns in F^r, so the columns of $\mathbf{B}^{\top}$ are linearly independent. By Theorem 3.4.4, $\mathbf{B}^{\top}$ is invertible, so $\det\mathbf{B}^{\top}\neq 0$ by Theorem 7.5.2, and $\det\mathbf{B}=\det\mathbf{B}^{\top}\neq 0$ by Theorem 7.4.5.

Conclusion. By Step 2 there is a nonzero minor of size r, and by Step 1 there is none of size larger than r. This proves (a). For (b): if $\operatorname{rank}\mathbf{A}\ge k$, then Step 2 gives a nonzero $r\times r$ minor with $r\ge k$, and we still need one of size exactly k. Expanding that $r\times r$ determinant along any row (Theorem 7.6.1) writes it as a combination of $(r-1)\times(r-1)$ minors of the block, each of which is also an $(r-1)\times(r-1)$ minor of $\mathbf{A}$; since the determinant is nonzero, one of these minors is nonzero. Repeating $r-k$ times gives a nonzero $k\times k$ minor. Conversely, a nonzero $k\times k$ minor gives $\operatorname{rank}\mathbf{A}\ge k$ by Step 1. The second equivalence in (b) is the negation of the first.

■

The theorem turns "rank at most $k-1$" into a finite list of polynomial equations in the entries: all $k\times k$ minors vanish. That is why rank can only **drop** when the entries satisfy special equations, as the next example shows.

Example. Rank of a Family of Matrices

For $t\in\mathbb{Q}$, let

$$\mathbf{A}(t)=\begin{pmatrix}1&2&1\\2&t&3\\3&6&t\end{pmatrix}.$$

Find $\operatorname{rank}\mathbf{A}(t)$ for every t.

Solution. The only 3×3 minor is $\det\mathbf{A}(t)$. By cofactor expansion along the first row (Theorem 7.6.1),

$$\begin{aligned}\det\mathbf{A}(t)&=1\cdot(t^2-18)-2\cdot(2t-9)+1\cdot(12-3t)\\&=t^2-7t+12=(t-3)(t-4).\end{aligned}$$

So for $t\neq 3,4$, $\mathbf{A}(t)$ has a nonzero 3×3 minor and $\operatorname{rank}\mathbf{A}(t)=3$ by Theorem 7.7.1.

For $t\in\{3,4\}$ every 3×3 minor is zero, so $\operatorname{rank}\mathbf{A}(t)\le 2$. Look for a 2×2 minor that does

not involve t: rows $\{1,2\}$ and columns $\{1,3\}$ give

$$\det\begin{pmatrix}1 & 1\\ 2 & 3\end{pmatrix} = 1 \neq 0.$$

Hence $\operatorname{rank}\mathbf{A}(3) = \operatorname{rank}\mathbf{A}(4) = 2$. The same minor shows $\operatorname{rank}\mathbf{A}(t) \geq 2$ for every t, so rank 1 never occurs.

Remark. To **compute** the rank of one given matrix, elimination is far cheaper: a 10×10 matrix has $\binom{10}{5}^2 = 63504$ minors of size 5. Minors earn their keep in theory. They show that rank is unchanged by transposing (the minors of $\mathbf{A}^\top$ are the minors of $\mathbf{A}$, by Theorem 7.4.5) and by enlarging the field (Exercise (C2: Rank does not depend on the field)), and they describe the matrices of rank below k by polynomial equations.

7.7.3 The Cauchy–Binet formula

Multiplicativity, Theorem 7.5.4, handles $\det(\mathbf{AB})$ when both factors are square. Now let $\mathbf{A} \in M_{m\times n}(F)$ and $\mathbf{B} \in M_{n\times m}(F)$. The product $\mathbf{AB}$ is $m \times m$, so it has a determinant, but $\det\mathbf{A}$ and $\det\mathbf{B}$ do not exist. The natural candidates to replace them are the largest minors: the $m \times m$ minors of $\mathbf{A}$, which keep all rows and choose m columns S, and the $m \times m$ minors of $\mathbf{B}$, which keep all columns and choose the **same** m rows S.

Theorem 7.7.2. Cauchy–Binet Formula

Let $\mathbf{A} \in M_{m\times n}(F)$ and $\mathbf{B} \in M_{n\times m}(F)$. Then

$$\det(\mathbf{AB}) = \sum_{\substack{S\subseteq[n]\\ |S|=m}} \det\mathbf{A}_{[m],S}\,\det\mathbf{B}_{S,[m]},$$

the sum running over all m-element subsets S of $[n]$. In particular:

(a) if $m > n$, the sum is empty and $\det(\mathbf{AB}) = 0$;

(b) if $m = n$, the only subset is $S = [n]$ and the formula is $\det(\mathbf{AB}) = \det\mathbf{A}\det\mathbf{B}$.

Idea. Treat det as a function of the m columns of $\mathbf{AB}$. Column k of $\mathbf{AB}$ is $\mathbf{Ab}_k = \sum_j b_{jk}\mathbf{a}_j$, a combination of the columns of $\mathbf{A}$. Expand by multilinearity in all m columns at once, exactly as in the uniqueness proof: one term for each way of picking a column $\mathbf{a}_{\varphi(k)}$ of $\mathbf{A}$ in slot k. ① Terms that pick the same column twice vanish, because the determinant is alternating. ② The surviving picks use m **distinct** columns, forming a set S, in some order; putting them in increasing order costs the sign of the reordering (Lemma 7.4.1). ③ Collecting all orders of the same S produces a Leibniz sum, which is $\det\mathbf{B}_{S,[m]}$.

Proof. Write $\mathbf{a}_1,\dots,\mathbf{a}_n \in F^m$ for the columns of $\mathbf{A}$ and b_{jk} for the entries of $\mathbf{B}$. For $\mathbf{x}_1,\dots,\mathbf{x}_m \in F^m$, the function $D(\mathbf{x}_1,\dots,\mathbf{x}_m) = \det(\mathbf{x}_1,\dots,\mathbf{x}_m)$ is an alternating m-linear form by Theorem 7.4.3.

By Theorem 0.9.2, column k of $\mathbf{AB}$ is $\mathbf{A}\mathbf{b}_k = \sum_{j=1}^{n} b_{jk}\mathbf{a}_j$. Expanding D by linearity in each of its m arguments in turn,

$$\det(\mathbf{AB}) = D\Big(\sum_{j=1}^{n} b_{j1}\mathbf{a}_j, \dots, \sum_{j=1}^{n} b_{jm}\mathbf{a}_j\Big) = \sum_{\varphi} \Big(\prod_{k=1}^{m} b_{\varphi(k)k}\Big)\, D(\mathbf{a}_{\varphi(1)}, \dots, \mathbf{a}_{\varphi(m)}),$$

where φ runs over **all** functions $[m] \to [n]$. If φ is not injective, two arguments of D are equal and the term is 0 because D is alternating (Definition 7.2.2). So only injective φ contribute. In particular, if $m > n$ there is no injective φ, and $\det(\mathbf{AB}) = 0$, which proves (a).

An injective φ has an image $S = \{s_1 < \cdots < s_m\} \subseteq [n]$ with m elements, and then $\varphi(k) = s_{\pi(k)}$ for exactly one $\pi \in S_m$. Conversely each pair (S, π) gives an injective φ. By Lemma 7.4.1 with $\mathbf{v}_p = \mathbf{a}_{s_p}$,

$$D(\mathbf{a}_{\varphi(1)}, \dots, \mathbf{a}_{\varphi(m)}) = \operatorname{sgn}(\pi)\, D(\mathbf{a}_{s_1}, \dots, \mathbf{a}_{s_m}) = \operatorname{sgn}(\pi)\, \det \mathbf{A}_{[m],S},$$

since $\mathbf{A}_{[m],S}$ has columns $\mathbf{a}_{s_1}, \dots, \mathbf{a}_{s_m}$. Hence

$$\det(\mathbf{AB}) = \sum_{|S|=m} \det \mathbf{A}_{[m],S} \sum_{\pi \in S_m} \operatorname{sgn}(\pi) \prod_{k=1}^{m} b_{s_{\pi(k)}k}.$$

The matrix $\mathbf{B}_{S,[m]}$ has (p,k)-entry $b_{s_p k}$, so its $(\pi(k),k)$-entry is $b_{s_{\pi(k)}k}$, and the inner sum is $\det \mathbf{B}_{S,[m]}$ by Definition 7.4.1. This proves the formula. For $m = n$, the only S is $[n]$, with $\mathbf{A}_{[n],[n]} = \mathbf{A}$ and $\mathbf{B}_{[n],[n]} = \mathbf{B}$, which gives (b).

■

Part (b) is a second proof of multiplicativity, and part (a) matches what rank predicts: if $m > n$, then $\operatorname{rank}(\mathbf{AB}) \le \operatorname{rank} \mathbf{A} \le n < m$ by Theorem 3.5.9, so $\mathbf{AB}$ is not invertible. The formula also passes a sanity check at $m = 1$: a $1 \times n$ row times an $n \times 1$ column is the 1×1 matrix $\sum_j a_{1j} b_{j1}$, and the 1×1 minors are single entries.

Remark. The proof uses only sums, differences, and products of entries, together with Lemma 7.4.1, which is division-free by the remark in §4. So, like the results flagged in §4–§6, the Cauchy–Binet formula holds for matrices with entries in any commutative ring.

Example. Cauchy–Binet for a 2×3 times 3×2 Product

Let

$$\mathbf{A} = \begin{pmatrix} 1 & 2 & 0 \\ 0 & 1 & 3 \end{pmatrix}, \qquad \mathbf{B} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 2 & -1 \end{pmatrix}$$

over $\mathbb{Q}$. Compute $\det(\mathbf{AB})$ directly and by the Cauchy–Binet formula. Then compute $\det(\mathbf{BA})$.

Solution. Directly. $\mathbf{AB} = \begin{pmatrix} 1+2+0 & 0+2+0 \\ 0+1+6 & 0+1-3 \end{pmatrix} = \begin{pmatrix} 3 & 2 \\ 7 & -2 \end{pmatrix}$, so $\det(\mathbf{AB}) = -6 - 14 = -20$.

By Cauchy–Binet. Here $m = 2$, $n = 3$, and there are three subsets S. The minors of $\mathbf{A}$ were computed after Definition 7.7.1. The minors of $\mathbf{B}$ keep both columns and rows S:

S	$\det \mathbf{A}_{[2],S}$	$\det \mathbf{B}_{S,[2]}$	product
$\{1,2\}$	1	$\det \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} = 1$	1
$\{1,3\}$	3	$\det \begin{pmatrix} 1 & 0 \\ 2 & -1 \end{pmatrix} = -1$	-3
$\{2,3\}$	6	$\det \begin{pmatrix} 1 & 1 \\ 2 & -1 \end{pmatrix} = -3$	-18

The sum is $1 - 3 - 18 = -20$, in agreement.

The other product. $\mathbf{BA}$ is 3×3, and here the roles are $m = 3 > n = 2$, so $\det(\mathbf{BA}) = 0$ by Theorem 7.7.2 (a). Indeed $\mathbf{BA} = \begin{pmatrix} 1 & 2 & 0 \\ 1 & 3 & 3 \\ 2 & 3 & -3 \end{pmatrix}$, whose third row is 3 times the first minus the second, so $\det(\mathbf{BA}) = 0$.

Warning. **For nonsquare factors, $\det(\mathbf{AB})$ and $\det(\mathbf{BA})$ differ.** For square matrices $\det(\mathbf{AB}) = \det \mathbf{A} \det \mathbf{B} = \det(\mathbf{BA})$. In the example above, $\det(\mathbf{AB}) = -20$ while $\det(\mathbf{BA}) = 0$. An even smaller case: $\mathbf{A} = \begin{pmatrix} 1 & 1 \end{pmatrix}$ and $\mathbf{B} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ give $\mathbf{AB} = (2)$ and $\mathbf{BA} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$, with determinants 2 and 0. Also, do not write "$\det \mathbf{A} \det \mathbf{B}$" for nonsquare $\mathbf{A}$, $\mathbf{B}$: those symbols are undefined, and the sum over S is what replaces them.

Quick check. Let $\mathbf{u}, \mathbf{v} \in F^n$ with $n \geq 2$. Without computing, what is $\det(\mathbf{u}\mathbf{v}^{\top})$?

Solution. $\mathbf{u}\mathbf{v}^{\top}$ is the product of the $n \times 1$ matrix $\mathbf{u}$ and the $1 \times n$ matrix $\mathbf{v}^{\top}$. Here the product is $n \times n$ and the middle size is $1 < n$, so $\det(\mathbf{u}\mathbf{v}^{\top}) = 0$ by Theorem 7.7.2 (a).

7.7.4 Sums of squares over the reals

Over $\mathbb{R}$, take $\mathbf{B} = \mathbf{A}^{\top}$. The minors of $\mathbf{A}^{\top}$ are the minors of $\mathbf{A}$ with rows and columns exchanged, so Cauchy–Binet writes $\det(\mathbf{A}\mathbf{A}^{\top})$ as a sum of squares.

Corollary 7.7.3. Gram Determinants Are Nonnegative

Let $\mathbf{A} \in M_{m\times n}(\mathbb{R})$. Then

$$\det(\mathbf{A}\mathbf{A}^{\top}) = \sum_{\substack{S\subseteq[n]\\|S|=m}} \left(\det \mathbf{A}_{[m],S}\right)^2 \;\geq\; 0,$$

with equality if and only if $\operatorname{rank}\mathbf{A} < m$.

Proof. By Definition 0.9.10, the (p,k)-entry of $(\mathbf{A}^{\top})_{S,[m]}$ is the (k,s_p)-entry of $\mathbf{A}$, so $(\mathbf{A}^{\top})_{S,[m]} = (\mathbf{A}_{[m],S})^{\top}$. By Theorem 7.4.5, $\det(\mathbf{A}^{\top})_{S,[m]} = \det \mathbf{A}_{[m],S}$. Substituting into Theorem 7.7.2 with $\mathbf{B} = \mathbf{A}^{\top}$ gives the formula. A sum of squares of real numbers is ≥ 0, and it is 0 if and only if every term is 0. Every $m \times m$ minor of $\mathbf{A}$ has the form $\det \mathbf{A}_{[m],S}$, since an $m \times m$ minor must use all m rows. If $m > n$, the sum is empty and $\operatorname{rank}\mathbf{A} \leq n < m$. Otherwise, by Theorem 7.7.1 (b) (or directly if $\mathbf{A} = \mathbf{O}$), all these minors vanish if and only if $\operatorname{rank}\mathbf{A} < m$. ∎

The matrix $\mathbf{A}\mathbf{A}^{\top}$ collects the dot products of the rows of $\mathbf{A}$ with each other. Such matrices of pairwise inner products are **Gram matrices**; Chapters 11 and 13 study them, and there this corollary becomes the statement that a Gram matrix is positive semidefinite. For $m = 2$ the corollary is a classical identity about two vectors, Exercise (B3: The Lagrange identity). Over $\mathbb{C}$ or $\mathbb{F}_p$ the conclusion "≥ 0" has no meaning, and the equality case can fail: over $\mathbb{C}$, $\mathbf{A} = \begin{pmatrix} 1 & i \end{pmatrix}$ has rank 1 but $\mathbf{A}\mathbf{A}^{\top} = (1 + i^2) = \mathbf{O}_1$.

7.7.5 Exercises

7.7.5.1 A. Check your understanding

Exercise. A1

(a) Define the $k \times k$ minor of $\mathbf{A} \in M_{m\times n}(F)$ on rows I and columns J.

(b) True or false: if every principal minor of a square matrix is zero, the matrix is zero. Justify your answer.

(c) True or false: if $\mathbf{A} \in M_{4\times 5}(F)$ has a nonzero 3×3 minor, then $\operatorname{rank}\mathbf{A} \geq 3$. Justify your answer.

(d) True or false: if $\mathbf{A} \in M_3(F)$ has a zero 2×2 minor, then $\operatorname{rank}\mathbf{A} \leq 1$. Justify your answer.

(e) State the Cauchy–Binet formula, and say what it gives when $m > n$.

(f) Let $\mathbf{A} \in M_{2\times 3}(\mathbb{R})$. Why is $\det(\mathbf{A}^{\top}\mathbf{A}) = 0$?

Solution.

(a) For $I = \{i_1 < \cdots < i_k\} \subseteq [m]$ and $J = \{j_1 < \cdots < j_k\} \subseteq [n]$, it is $\det \mathbf{A}_{I,J}$, where $\mathbf{A}_{I,J} \in M_k(F)$ has (p,q)-entry $a_{i_p j_q}$ (Definition 7.7.1).

(b) False. $\mathbf{N} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$ has principal minors $0, 0, 0$ but $\mathbf{N} \neq \mathbf{O}$.

(c) True, by Step 1 of the proof of Theorem 7.7.1, or by part (b) of that theorem.

(d) False. The rank is decided by whether **some** minor is nonzero, not by one zero minor. $\mathbf{I}_3$ has the zero 2×2 minor on rows $\{1,2\}$ and columns $\{2,3\}$, namely $\det\begin{pmatrix}0&0\\1&0\end{pmatrix}=0$, but $\operatorname{rank}\mathbf{I}_3=3$.

(e) For $\mathbf{A}\in M_{m\times n}(F)$ and $\mathbf{B}\in M_{n\times m}(F)$, $\det(\mathbf{AB})=\sum_{|S|=m}\det\mathbf{A}_{[m],S}\det\mathbf{B}_{S,[m]}$, summed over m-subsets $S\subseteq[n]$. If $m>n$ there are no such subsets and $\det(\mathbf{AB})=0$.

(f) $\mathbf{A}^{\top}\mathbf{A}$ is 3×3, the product of a 3×2 and a 2×3 matrix. By Theorem 7.7.2 (a) with $m=3>n=2$, $\det(\mathbf{A}^{\top}\mathbf{A})=0$.

7.7.5.2 B. Practice

Exercise. B1: Rank with a parameter

For $s\in\mathbb{Q}$, let $\mathbf{A}(s)=\begin{pmatrix}1&s&1\\s&1&1\\1&1&s\end{pmatrix}$. Use minors to find $\operatorname{rank}\mathbf{A}(s)$ for every s.

Solution. By cofactor expansion along the first row (Theorem 7.6.1),

$$\begin{aligned}\det\mathbf{A}(s)&=1\cdot(s-1)-s\,(s^2-1)+1\cdot(s-1)\\&=-s^3+3s-2=-(s-1)^2(s+2).\end{aligned}$$

If $s\neq 1,-2$, this 3×3 minor is nonzero, so $\operatorname{rank}\mathbf{A}(s)=3$ by Theorem 7.7.1.

If $s=-2$, then $\det\mathbf{A}(-2)=0$, so $\operatorname{rank}\mathbf{A}(-2)\le 2$. The minor on rows $\{1,2\}$ and columns $\{1,2\}$ is $\det\begin{pmatrix}1&-2\\-2&1\end{pmatrix}=1-4=-3\neq 0$. Hence $\operatorname{rank}\mathbf{A}(-2)=2$.

If $s=1$, every entry is 1. Every 2×2 minor is $\det\begin{pmatrix}1&1\\1&1\end{pmatrix}=0$, and the 1×1 minors are nonzero. Hence $\operatorname{rank}\mathbf{A}(1)=1$.

Exercise. B2: Checking Cauchy–Binet

Let $\mathbf{A}=\begin{pmatrix}1&0&2\\-1&1&1\end{pmatrix}$ and $\mathbf{B}=\begin{pmatrix}2&1\\0&1\\1&-1\end{pmatrix}$ over $\mathbb{Q}$.

(a) Compute $\det(\mathbf{AB})$ directly.

(b) Compute $\det(\mathbf{AB})$ by the Cauchy–Binet formula.

(c) Hence, without multiplying, find $\det(\mathbf{BA})$.

Solution.

(a) $\mathbf{AB} = \begin{pmatrix} 2+0+2 & 1+0-2 \\ -2+0+1 & -1+1-1 \end{pmatrix} = \begin{pmatrix} 4 & -1 \\ -1 & -1 \end{pmatrix}$, so $\det(\mathbf{AB}) = -4-1 = -5$.

(b) The three subsets of $[3]$ of size 2 give

$$S=\{1,2\}:\ \det\begin{pmatrix}1&0\\-1&1\end{pmatrix}\det\begin{pmatrix}2&1\\0&1\end{pmatrix}=1\cdot 2=2,$$
$$S=\{1,3\}:\ \det\begin{pmatrix}1&2\\-1&1\end{pmatrix}\det\begin{pmatrix}2&1\\1&-1\end{pmatrix}=3\cdot(-3)=-9,$$
$$S=\{2,3\}:\ \det\begin{pmatrix}0&2\\1&1\end{pmatrix}\det\begin{pmatrix}0&1\\1&-1\end{pmatrix}=(-2)(-1)=2.$$

The sum is $2-9+2=-5$, agreeing with (a).

(c) $\mathbf{BA} \in M_3(\mathbb{Q})$ is a product through F^2, so $\det(\mathbf{BA}) = 0$ by Theorem 7.7.2 (a).

Exercise. B3: The Lagrange identity

Let $\mathbf{a} = (a_1, \dots, a_n)$ and $\mathbf{b} = (b_1, \dots, b_n)$ be vectors in $\mathbb{R}^n$, $n \geq 2$, and write $\mathbf{a} \cdot \mathbf{b} = \sum_i a_i b_i$.

(a) Let $\mathbf{M} \in M_{2\times n}(\mathbb{R})$ have rows $\mathbf{a}^\top$ and $\mathbf{b}^\top$. Compute $\mathbf{M}\mathbf{M}^\top$.

(b) Hence prove the **Lagrange identity**

$$(\mathbf{a}\cdot\mathbf{a})(\mathbf{b}\cdot\mathbf{b}) - (\mathbf{a}\cdot\mathbf{b})^2 = \sum_{1\leq i<j\leq n} (a_i b_j - a_j b_i)^2.$$

(c) Check it for $\mathbf{a} = (1,2,3)$ and $\mathbf{b} = (4,5,6)$, and deduce $(\mathbf{a}\cdot\mathbf{b})^2 \leq (\mathbf{a}\cdot\mathbf{a})(\mathbf{b}\cdot\mathbf{b})$ for all $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$.

Solution.

(a) By Theorem 0.9.2, the (p,q)-entry of $\mathbf{M}\mathbf{M}^\top$ is row p of $\mathbf{M}$ times column q of $\mathbf{M}^\top$, which is row q of $\mathbf{M}$ transposed. Hence

$$\mathbf{M}\mathbf{M}^\top = \begin{pmatrix} \mathbf{a}\cdot\mathbf{a} & \mathbf{a}\cdot\mathbf{b} \\ \mathbf{b}\cdot\mathbf{a} & \mathbf{b}\cdot\mathbf{b} \end{pmatrix}, \qquad \mathbf{b}\cdot\mathbf{a} = \mathbf{a}\cdot\mathbf{b}.$$

(b) The left side of the identity is $\det(\mathbf{M}\mathbf{M}^\top)$. By Corollary 7.7.3 with $m = 2$, it equals the sum over $S = \{i < j\}$ of $(\det \mathbf{M}_{[2],S})^2$, and $\mathbf{M}_{[2],\{i,j\}} = \begin{pmatrix} a_i & a_j \\ b_i & b_j \end{pmatrix}$ has determinant $a_i b_j - a_j b_i$. This proves the identity.

(c) Left side: $\mathbf{a}\cdot\mathbf{a} = 14$, $\mathbf{b}\cdot\mathbf{b} = 77$, $\mathbf{a}\cdot\mathbf{b} = 32$, so $14\cdot 77 - 32^2 = 1078 - 1024 = 54$. Right side: the pairs $(1,2), (1,3), (2,3)$ give $5-8=-3$, $6-12=-6$, $12-15=-3$, and $9+36+9=54$. In general the right side is a sum of squares of reals, hence ≥ 0, so $(\mathbf{a}\cdot\mathbf{b})^2 \leq (\mathbf{a}\cdot\mathbf{a})(\mathbf{b}\cdot\mathbf{b})$.

7.7.5.3 C. Going deeper

Exercise. C1: Counting spanning trees

Consider the network with vertices $1, 2, 3, 4$ and five edges, oriented as $e_1 = 1 \to 2$, $e_2 = 2 \to 3$, $e_3 = 3 \to 4$, $e_4 = 4 \to 1$, $e_5 = 1 \to 3$. A **spanning tree** is a set of 3 edges that connects all four vertices (ignoring orientation). Let $\mathbf{N} \in M_{4\times 5}(\mathbb{R})$ be its incidence matrix (Definition 3.7.1), let $Ł = \mathbf{N}\mathbf{N}^\top$, and let $\mathbf{N}_0$ and $Ł_0$ be $\mathbf{N}$ and $Ł$ with the row (and, for $Ł$, the column) of vertex 4 deleted.

(a) Compute $Ł$ and $\det Ł_0$.

(b) List the spanning trees and count them.

(c) Show that $Ł_0 = \mathbf{N}_0\mathbf{N}_0^\top$, and use Corollary 7.7.3 to write $\det Ł_0$ as a sum over 3-element sets of edges.

(d) Compute the minor of $\mathbf{N}_0$ for the tree $\{e_1, e_2, e_3\}$ and for the nontree $\{e_1, e_2, e_5\}$. Explain, for these two sets, why the answers are ± 1 and 0.

The **matrix-tree theorem** says that for every connected network, $\det Ł_0$ equals the number of spanning trees; parts (c) and (d) are the heart of its proof, which Chapter 24 completes.

Hint: for (d), a set of edges containing a cycle gives dependent columns.

Solution.

(a) Column e of $\mathbf{N}$ has 1 at the tail and -1 at the head:

$$\mathbf{N} = \begin{pmatrix} 1 & 0 & 0 & -1 & 1 \\ -1 & 1 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & -1 \\ 0 & 0 & -1 & 1 & 0 \end{pmatrix}, \qquad Ł = \mathbf{N}\mathbf{N}^\top = \begin{pmatrix} 3 & -1 & -1 & -1 \\ -1 & 2 & -1 & 0 \\ -1 & -1 & 3 & -1 \\ -1 & 0 & -1 & 2 \end{pmatrix}.$$

(Diagonal entries count the edges at a vertex; off-diagonal entries are -1 for each edge joining two vertices.) Expanding along the first row,

$$\begin{aligned} \det Ł_0 &= \det \begin{pmatrix} 3 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 3 \end{pmatrix} \\ &= 3(6-1) + 1\cdot(-3-1) - 1\cdot(1+2) = 15 - 4 - 3 = 8. \end{aligned}$$

(b) There are $\binom{5}{3} = 10$ sets of three edges. Three edges on four vertices connect everything exactly when they contain no cycle, and the only 3-edge cycles are the triangles $\{e_1, e_2, e_5\}$ (vertices $1, 2, 3$) and $\{e_3, e_4, e_5\}$ (vertices $1, 3, 4$). So there are $10 - 2 = 8$ spanning trees, matching $\det Ł_0$.

(c) Deleting row 4 of $\mathbf{N}$ deletes row 4 of $\mathbf{N}\mathbf{N}^\top$ (row view) and column 4 of $\mathbf{N}\mathbf{N}^\top$ (column view of $\mathbf{N}^\top$), by Theorem 0.9.2. So $Ł_0 = \mathbf{N}_0\mathbf{N}_0^\top$, with $\mathbf{N}_0 \in M_{3\times 5}(\mathbb{R})$, and Corollary 7.7.3 gives $\det Ł_0 = \sum_{|S|=3} (\det(\mathbf{N}_0)_{[3],S})^2$, one term for each set S of three edges.

(d) For $S = \{e_1, e_2, e_3\}$, $(\mathbf{N}_0)_{[3],S} = \begin{pmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & -1 & 1 \end{pmatrix}$, which is lower triangular with determinant 1 (Theorem 7.4.6). For $S = \{e_1, e_2, e_5\}$, the columns are $(1, -1, 0)$, $(0, 1, -1)$, $(1, 0, -1)$, and the third is the sum of the first two: going around the cycle $1 \to 2 \to 3$

equals going directly $1 \to 3$. So the columns are dependent and the minor is 0 by Theorem 7.2.1. Thus the nontree contributes 0^2 and the tree contributes 1^2.

Exercise. C2: Rank does not depend on the field

Let K be a subfield of a field L (Definition 1.9.1), and let $\mathbf{A} \in M_{m\times n}(K)$. We may also regard $\mathbf{A}$ as a matrix in $M_{m\times n}(L)$.

(a) Explain why each minor $\det \mathbf{A}_{I,J}$ is the same element whether computed in K or in L.

(b) Prove that the rank of $\mathbf{A}$ over K equals its rank over L.

(c) Deduce: if a system $\mathbf{Ax} = \mathbf{0}$ with rational coefficients has a nonzero complex solution, it has a nonzero rational solution.

Solution.

(a) By Definition 7.4.1, $\det \mathbf{A}_{I,J}$ is a finite sum of $\pm$ products of entries of $\mathbf{A}$. The entries lie in K, and K has the same addition and multiplication as L, so the sum is the same element of $K \subseteq L$.

(b) If $\mathbf{A} = \mathbf{O}$, both ranks are 0. Otherwise, by Theorem 7.7.1 (a) applied over K and over L, each rank is the largest k for which some $k \times k$ minor is nonzero. By (a) the list of minors, and which of them are nonzero, is the same in both cases. Hence the two ranks agree.

(c) Take $K = \mathbb{Q}$, $L = \mathbb{C}$. A nonzero complex solution means $\operatorname{nullity} \mathbf{A} \geq 1$ over $\mathbb{C}$, so $\operatorname{rank}_{\mathbb{C}} \mathbf{A} \leq n-1$ by Theorem 3.5.7. By (b), $\operatorname{rank}_{\mathbb{Q}} \mathbf{A} \leq n-1$, so $\operatorname{nullity} \mathbf{A} \geq 1$ over $\mathbb{Q}$, and a nonzero rational solution exists.

7.8 Special Determinants

A determinant of size n has $n!$ terms, so a general formula for a large matrix is hopeless. Some matrices, though, have so much structure that their determinants collapse to a short product or a recurrence. This section computes the most useful of them: the Vandermonde determinant, which governs polynomial interpolation; tridiagonal determinants, which satisfy a three-term recurrence; block triangular determinants, which split into smaller ones; and the determinant of a rank-one update of the identity. The proofs share one idea: find the structure first (a polynomial with known roots, a row with two nonzero entries, a pattern of zeros) and let the defining properties of the determinant do the rest.

7.8.1 The Vandermonde determinant

Chapter 3 proved that there is exactly one polynomial of degree at most n through $n+1$ points with distinct first coordinates (Theorem 3.7.3). Behind it sat a square matrix of powers of the nodes. Its determinant turns out to be a product of all the differences of the nodes, and that product explains at a glance why distinct nodes are exactly what is needed.

For $x_1, \dots, x_n \in F$, the **Vandermonde matrix** is

$$\mathbf{V}(x_1, \dots, x_n) := \begin{pmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^{n-1} \\ 1 & x_2 & x_2^2 & \cdots & x_2^{n-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^{n-1} \end{pmatrix} \in M_n(F),$$

whose (i, j)-entry is x_i^{j-1}. Row i lists the powers $1, x_i, \dots, x_i^{n-1}$ of the i-th node. For $n = 1$ it is the 1×1 matrix (1), and for $n = 2$ its determinant is $x_2 - x_1$.

Theorem 7.8.1. Vandermonde Determinant

Let $n \geq 1$ and $x_1, \dots, x_n \in F$. Then

$$\det \mathbf{V}(x_1, \dots, x_n) = \prod_{1 \leq i < j \leq n} (x_j - x_i),$$

where the empty product (for $n = 1$) is 1.

Idea. Induction on n. Replace the last node x_n by a variable t. Expanding along the last row shows the determinant is a polynomial $f(t)$ of degree at most $n - 1$, with leading coefficient the smaller Vandermonde determinant. Setting $t = x_i$ for $i < n$ makes two rows equal, so $x_1, \dots, x_{n-1}$ are roots of f. A polynomial of degree at most $n - 1$ with these $n - 1$ roots is its leading coefficient times $\prod_{i<n}(t - x_i)$. Now put $t = x_n$ back. The roots argument needs the x_i distinct; if they are not, both sides are 0 anyway.

Proof. We use induction on n (Theorem 0.2.7). For $n = 1$, $\det(1) = 1$, the empty product.

Let $n \geq 2$, assume the formula for $n-1$ nodes, and write $\mathbf{V} = \mathbf{V}(x_1, \dots, x_n)$. Let $C_{n1}, \dots, C_{nn}$ be the cofactors of $\mathbf{V}$ along row n (Definition 7.6.1). Each C_{nj} is computed after deleting row n, so it does not involve x_n. Define

$$f(t) := \sum_{j=1}^{n} C_{nj}\, t^{j-1} \in F[t].$$

For $c \in F$, let $\mathbf{V}_c$ be $\mathbf{V}$ with row n replaced by $(1, c, \dots, c^{n-1})$. Then $\mathbf{V}_c$ has the same cofactors along row n as $\mathbf{V}$, so by Theorem 7.6.1 along row n,

$$\det \mathbf{V}_c = \sum_{j=1}^{n} c^{j-1} C_{nj} = f(c). \qquad (*)$$

In particular $\det \mathbf{V} = f(x_n)$. The coefficient of t^{n-1} in f is $C_{nn} = (-1)^{2n} \det \mathbf{V}(x_1, \dots, x_{n-1})$, since deleting row n and column n of $\mathbf{V}$ leaves the Vandermonde matrix of the first $n - 1$ nodes.

Case 1: two of $x_1, \dots, x_{n-1}$ are equal, say $x_p = x_q$ with $p < q < n$. Then rows p and q of $\mathbf{V}$ are equal. By Theorem 7.4.5, $\det \mathbf{V} = \det \mathbf{V}^\top$, and $\mathbf{V}^\top$ has two equal columns, so $\det \mathbf{V} = 0$ because the determinant is alternating (Definition 7.2.2). The product also vanishes, since it contains the factor $x_q - x_p = 0$.

Case 2: $x_1, \dots, x_{n-1}$ are distinct. For $i < n$, the matrix $\mathbf{V}_{x_i}$ has rows i and n equal, so $f(x_i) = \det \mathbf{V}_{x_i} = 0$ by $(*)$ and the argument of Case 1. Hence $x_1, \dots, x_{n-1}$ are $n-1$ distinct roots of f, and by Corollary 6.1.4,

$$g(t) \coloneqq (t - x_1)(t - x_2) \cdots (t - x_{n-1}) \text{ divides } f.$$

Write $f = hg$ with $h \in F[t]$. If $h \neq 0$, then $\deg h = \deg f - (n-1) \leq 0$ by Theorem 0.8.3, since $\deg f \leq n-1$; so in every case h is a constant $h \in F$. As g is monic of degree $n-1$, the coefficient of t^{n-1} in hg is h, so $h = C_{nn} = \det \mathbf{V}(x_1, \dots, x_{n-1})$. Evaluating at x_n (Theorem 0.8.5) and using the induction hypothesis,

$$\begin{aligned} \det \mathbf{V} = f(x_n) &= \det \mathbf{V}(x_1, \dots, x_{n-1}) \prod_{i=1}^{n-1} (x_n - x_i) \\ &= \prod_{1 \leq i < j \leq n-1} (x_j - x_i) \prod_{i=1}^{n-1} (x_n - x_i). \end{aligned}$$

The two products together run over all pairs $i < j \leq n$, the second one collecting exactly the pairs with $j = n$. This proves the formula for n, and the theorem follows by induction.

■

Example. A 3×3 Vandermonde Determinant

Compute $\det \mathbf{V}(1, 2, 4)$ from the formula and by cofactor expansion.

Solution. The formula gives $(2-1)(4-1)(4-2) = 1 \cdot 3 \cdot 2 = 6$. Directly, expanding

$$\mathbf{V}(1, 2, 4) = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 4 & 16 \end{pmatrix}$$

along the first row gives $1 \cdot (32 - 16) - 1 \cdot (16 - 4) + 1 \cdot (4 - 2) = 16 - 12 + 2 = 6$.

The product vanishes exactly when a factor does, because a field has no zero divisors. So:

Corollary 7.8.2. Distinct Nodes Give an Invertible Vandermonde Matrix

$\mathbf{V}(x_1, \dots, x_n)$ is invertible if and only if $x_1, \dots, x_n$ are **distinct**.

Proof. In a field, a product is nonzero if and only if every factor is nonzero (Theorem 0.7.1 (f), applied repeatedly). So by Theorem 7.8.1, $\det \mathbf{V}(x_1, \dots, x_n) \neq 0$ if and only if $x_j \neq x_i$ for all $i < j$. By Theorem 7.5.2, this is equivalent to invertibility.

■

Now the promised two-line proof of Theorem 3.7.3. Given distinct $x_0, \dots, x_n$ and values $y_0, \dots, y_n$, a polynomial $p = c_0 + c_1 x + \cdots + c_n x^n$ satisfies $p(x_i) = y_i$ for all i if and only if

$\mathbf{V}(x_0,\dots,x_n)\,\mathbf{c} = \mathbf{y}$, since row i of this system reads $c_0 + c_1x_i + \cdots + c_nx_i^n = y_i$. By Corollary 7.8.2 the matrix is invertible, so the system has exactly one solution $\mathbf{c} = \mathbf{V}(x_0,\dots,x_n)^{-1}\mathbf{y}$ (Theorem 3.4.4), and the interpolating polynomial exists and is unique.

Remark. This is not a proof from nothing. The root bound of Chapter 6 (Corollary 6.1.4) did the real work inside Theorem 7.8.1, exactly as the root bound did in Chapter 3 (Lemma 3.7.2). What the determinant adds is an explicit number: it says how close to singular the interpolation problem is, since nodes that nearly coincide make the product small.

Warning. **Watch the shape and the order.** The formula is for rows $(1, x_i, \dots, x_i^{n-1})$ starting at the power 0, and for the factors $x_j - x_i$ with $i < j$, "later node minus earlier node". Dropping the column of ones changes the answer: $\det\begin{pmatrix} x_1 & x_1^2 \\ x_2 & x_2^2 \end{pmatrix} = x_1x_2(x_2 - x_1)$, not $x_2 - x_1$. Writing the factors as $x_i - x_j$ instead flips the sign of each of the $\binom{n}{2}$ factors, so it multiplies the answer by $(-1)^{\binom{n}{2}}$: already for $n = 2$ and $n = 3$ this is -1, a wrong sign.

Quick check. Without expanding, find $\det \mathbf{V}(1, 2, 3, 5)$.

Solution. The pairs $i < j$ give $(2-1)(3-1)(5-1)(3-2)(5-2)(5-3) = 1\cdot2\cdot4\cdot1\cdot3\cdot2 = 48$.

7.8.2 Tridiagonal determinants

Many matrices in applications have nonzero entries only on the diagonal and next to it: discretized differential equations, chains of springs, the recurrences of orthogonal polynomials. Such a matrix is **tridiagonal**. Its determinant is not a closed formula in general, but it obeys a recurrence in the size, because the last row has only two nonzero entries.

Let $\mathbf{T}_n \in M_n(F)$ have diagonal entries $a_1,\dots,a_n$, entries $b_1,\dots,b_{n-1}$ just above the diagonal ($(i, i+1)$-entry b_i), entries $c_1,\dots,c_{n-1}$ just below it ($(i+1, i)$-entry c_i), and zeros elsewhere:

$$\mathbf{T}_n = \begin{pmatrix} a_1 & b_1 & & & \\ c_1 & a_2 & b_2 & & \\ & c_2 & \ddots & \ddots & \\ & & \ddots & a_{n-1} & b_{n-1} \\ & & & c_{n-1} & a_n \end{pmatrix}.$$

For $1 \le k \le n$, the top-left $k \times k$ block of $\mathbf{T}_n$ is the tridiagonal matrix $\mathbf{T}_k$ built from $a_1,\dots,a_k$, $b_1,\dots,b_{k-1}$, $c_1,\dots,c_{k-1}$. Write $D_k = \det \mathbf{T}_k$, and set $D_0 = 1$.

Theorem 7.8.3. Tridiagonal Recurrence

With this notation, $D_1 = a_1$, and for $2 \le k \le n$,

$$D_k = a_k\, D_{k-1} - b_{k-1}c_{k-1}\, D_{k-2}.$$

Proof. $D_1 = \det(a_1) = a_1$. For $k = 2$, $D_2 = a_1a_2 - b_1c_1 = a_2D_1 - b_1c_1D_0$. Let $k \ge 3$. Row k of $\mathbf{T}_k$ has only two possibly nonzero entries, c_{k-1} in column $k-1$ and a_k in column k. By Theorem 7.6.1 along row k,

$$D_k = c_{k-1}(-1)^{2k-1}M_{k,k-1} + a_k(-1)^{2k}M_{kk} = a_kM_{kk} - c_{k-1}M_{k,k-1},$$

where $M_{kk} = \det \mathbf{T}_{k-1} = D_{k-1}$. The minor $M_{k,k-1}$ is the determinant of the $(k-1)\times(k-1)$ matrix obtained by deleting row k and column $k-1$. Its last column consists of rows $1,\dots,k-1$ of column k of $\mathbf{T}_k$, which is b_{k-1} in the last position and 0 elsewhere. Expanding along this last column (Theorem 7.6.1), with sign $(-1)^{(k-1)+(k-1)} = 1$, gives $M_{k,k-1} = b_{k-1}\det \mathbf{T}_{k-2} = b_{k-1}D_{k-2}$, because deleting the last row and column leaves rows and columns $1,\dots,k-2$ of $\mathbf{T}_k$. Substituting proves the recurrence. ∎

Example. Fibonacci Numbers as Determinants

Let $\mathbf{T}_n$ have 1 on the diagonal, 1 above it, and -1 below it. Find $D_n = \det \mathbf{T}_n$.

Solution. Here $a_k = 1$ and $b_{k-1}c_{k-1} = 1\cdot(-1) = -1$, so Theorem 7.8.3 gives $D_k = D_{k-1} + D_{k-2}$ for $k \ge 2$, with $D_0 = 1$ and $D_1 = 1$. Hence

$$D_1, D_2, D_3, \cdots = 1, 2, 3, 5, 8, 13, 21, 34, \dots,$$

the Fibonacci numbers: $D_n = F_{n+1}$, where $F_1 = F_2 = 1$ and $F_k = F_{k-1} + F_{k-2}$. For instance

$$D_3 = \det\begin{pmatrix} 1 & 1 & 0 \\ -1 & 1 & 1 \\ 0 & -1 & 1 \end{pmatrix} = 1\cdot(1+1) - 1\cdot(-1-0) + 0 = 3.$$

Only the products $b_{k-1}c_{k-1}$ enter the recurrence, never b_{k-1} and c_{k-1} separately. For example, for a real number t, putting $2t$ on the diagonal and 1 on both neighboring lines gives $D_k = 2t\, D_{k-1} - D_{k-2}$, so $D_1 = 2t$, $D_2 = 4t^2 - 1$, $D_3 = 8t^3 - 4t$. These are the Chebyshev polynomials of the second kind.

7.8.3 Block triangular matrices

A matrix can be cut into blocks by a horizontal and a vertical line. When the block in the lower left is zero, the determinant should see only the two square blocks on the diagonal, just as a triangular determinant sees only diagonal entries. This result is used constantly, for instance in Chapter 8 for Schur complements.

Theorem 7.8.4. Block Triangular Determinant

Let $\mathbf{A} \in M_k(F)$, $\mathbf{D} \in M_l(F)$ and $\mathbf{B} \in M_{k\times l}(F)$, with $k, l \geq 1$. Then

$$\det \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{O} & \mathbf{D} \end{pmatrix} = \det \mathbf{A} \cdot \det \mathbf{D},$$

where $\mathbf{O}$ is the $l \times k$ zero matrix. The same holds for $\begin{pmatrix} \mathbf{A} & \mathbf{O} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}$ with $\mathbf{C} \in M_{l\times k}(F)$.

Idea. In the Leibniz formula for the whole matrix, a term picks one entry from each column. The first k columns have zeros below row k, so a nonzero term must pick rows $1, \ldots, k$ for the first k columns, and is then forced to pick rows $k+1, \ldots, n$ for the last l columns. Such a permutation is a permutation of each block, its inversions are the inversions inside the blocks, and the Leibniz sum factors into two Leibniz sums.

Proof. Let $n = k + l$ and $\mathbf{M} = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{O} & \mathbf{D} \end{pmatrix} = (m_{ij})$. By Definition 7.4.1, $\det \mathbf{M} = \sum_{\sigma\in S_n} \operatorname{sgn}(\sigma) \prod_{j=1}^{n} m_{\sigma(j)j}$. Since $m_{ij} = 0$ for $i > k \geq j$, the term of σ is 0 unless $\sigma(j) \leq k$ for every $j \leq k$. For such σ, the injective map σ sends $\{1, \ldots, k\}$ into itself, hence onto itself (a finite set), and so it sends $\{k+1, \ldots, n\}$ onto itself. Therefore σ determines $\alpha \in S_k$ and $\beta \in S_l$ by

$$\alpha(j) = \sigma(j) \quad (j \leq k), \qquad \beta(q) = \sigma(k+q) - k \quad (q \leq l),$$

and every pair (α, β) arises from exactly one such σ.

We compare signs through inversions (Definition 7.3.1). A pair $i < j$ with $i, j \leq k$ is an inversion of σ exactly when it is one of α; a pair $k + p < k + q$ is an inversion of σ exactly when $p < q$ is one of β; and a pair $i \leq k < j$ is never an inversion, since $\sigma(i) \leq k < \sigma(j)$. By Definition 7.3.2, $\operatorname{sgn}(\sigma) = \operatorname{sgn}(\alpha)\operatorname{sgn}(\beta)$. Also $m_{\sigma(j)j} = a_{\alpha(j)j}$ for $j \leq k$ and $m_{\sigma(k+q),k+q} = d_{\beta(q)q}$. Hence

$$\det \mathbf{M} = \sum_{\alpha\in S_k} \sum_{\beta\in S_l} \operatorname{sgn}(\alpha)\operatorname{sgn}(\beta) \prod_{j=1}^{k} a_{\alpha(j)j} \prod_{q=1}^{l} d_{\beta(q)q} = \det \mathbf{A} \cdot \det \mathbf{D},$$

using Definition 7.4.1 twice. For the lower block triangular case, the transpose of $\begin{pmatrix} \mathbf{A} & \mathbf{O} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}$ is $\begin{pmatrix} \mathbf{A}^\top & \mathbf{C}^\top \\ \mathbf{O} & \mathbf{D}^\top \end{pmatrix}$, so by Theorem 7.4.5 and the first case its determinant is $\det \mathbf{A}^\top \det \mathbf{D}^\top = \det \mathbf{A} \det \mathbf{D}$.

■

The proof uses only sums and products of entries, so, like the results flagged in §4–§6, it holds for matrices over any commutative ring; §10 uses this over $F[x]$. By induction on the number of blocks, a matrix with square blocks $\mathbf{A}_1, \ldots, \mathbf{A}_r$ on the diagonal and zeros below them has determinant $\det \mathbf{A}_1 \cdots \det \mathbf{A}_r$; Theorem 7.4.6 is the case where every block is 1×1.

For example,

$$\det\begin{pmatrix} 2 & 1 & 5 & -3 \\ 1 & 1 & 0 & 7 \\ 0 & 0 & 3 & 1 \\ 0 & 0 & 4 & 2 \end{pmatrix} = \det\begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix} \det\begin{pmatrix} 3 & 1 \\ 4 & 2 \end{pmatrix} = 1 \cdot 2 = 2,$$

and the entries $5, -3, 0, 7$ of the upper right block play no role.

Warning. **There is no "block $ad - bc$" rule.** For a general block matrix, $\det\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix} \neq \det\mathbf{A}\det\mathbf{D} - \det\mathbf{B}\det\mathbf{C}$. Take $\mathbf{A} = \mathbf{D} = \mathbf{O}$ and $\mathbf{B} = \mathbf{C} = \mathbf{I}_2$: the 4×4 matrix $\begin{pmatrix} \mathbf{O} & \mathbf{I}_2 \\ \mathbf{I}_2 & \mathbf{O} \end{pmatrix}$ is the permutation matrix of $(1\ 3)(2\ 4)$, a product of two transpositions, so its determinant is $+1$; the rule would give $0 \cdot 0 - 1 \cdot 1 = -1$. The zero block in the corner is what the theorem needs. The correct formulas for general blocks use Schur complements, in Chapter 8.

Quick check. Is $\det\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{O} & \mathbf{D} \end{pmatrix} = \det\mathbf{A}\det\mathbf{D}$ meaningful when $\mathbf{A}$ is 2×3 and $\mathbf{D}$ is 3×2 (so the whole matrix is 5×5)?

Solution. No. The theorem requires $\mathbf{A}$ and $\mathbf{D}$ square; here $\det\mathbf{A}$ and $\det\mathbf{D}$ are undefined. In fact the whole determinant is 0 whatever the entries: the first 3 columns have their nonzero entries only in the first 2 rows, so they are linearly dependent in a 2-dimensional space, and a determinant with dependent columns is 0 (Theorem 7.2.1).

7.8.4 Rank-one updates

Changing a matrix by a rank-one matrix $\mathbf{u}\mathbf{v}^\top$ is the smallest possible change in the sense of rank, and it happens often: updating one observation in statistics, modifying one row of a system, or adding the same number to every entry. The determinant of the updated identity turns out to be a single dot product.

Theorem 7.8.5. Matrix Determinant Lemma

Let $\mathbf{u}, \mathbf{v} \in F^n$. Then

$$\det(\mathbf{I}_n + \mathbf{u}\mathbf{v}^\top) = 1 + \mathbf{v}^\top\mathbf{u}.$$

Idea. Border $\mathbf{I}_n + \mathbf{u}\mathbf{v}^\top$ by an extra column $\mathbf{u}$ and an extra row $(\mathbf{0}^\top, 1)$, giving a block triangular matrix $\mathbf{M}$ with the same determinant. Then clear the update by elimination with blocks. Rows: add $\mathbf{v}^\top$ times the top n rows to the new bottom row (multiply by Ł on the left). Columns: subtract the last column times $\mathbf{v}^\top$ from the first n columns (multiply by Ł′ on the right). The factors Ł, Ł′ have determinant 1, and the result is block triangular with $\mathbf{I}_n$ and the 1×1 block $1 + \mathbf{v}^\top\mathbf{u}$ on the diagonal.

Proof. Consider the $(n+1)\times(n+1)$ block matrices

$$Ł = \begin{pmatrix} \mathbf{I}_n & \mathbf{0} \\ \mathbf{v}^\top & 1 \end{pmatrix}, \qquad \mathbf{M} = \begin{pmatrix} \mathbf{I}_n + \mathbf{u}\mathbf{v}^\top & \mathbf{u} \\ \mathbf{0}^\top & 1 \end{pmatrix}, \qquad Ł' = \begin{pmatrix} \mathbf{I}_n & \mathbf{0} \\ -\mathbf{v}^\top & 1 \end{pmatrix}.$$

Splitting the sum $\sum_j$ in Theorem 0.9.2 (1) into the indices $j \le n$ and $j = n+1$ shows that matrices partitioned in this way multiply block by block, like 2×2 matrices whose entries are blocks. Hence

$$Ł\mathbf{M} = \begin{pmatrix} \mathbf{I}_n + \mathbf{u}\mathbf{v}^\top & \mathbf{u} \\ \mathbf{v}^\top + (\mathbf{v}^\top\mathbf{u})\mathbf{v}^\top & \mathbf{v}^\top\mathbf{u} + 1 \end{pmatrix},$$

where we used $\mathbf{v}^\top(\mathbf{u}\mathbf{v}^\top) = (\mathbf{v}^\top\mathbf{u})\mathbf{v}^\top$ by associativity, $\mathbf{v}^\top\mathbf{u}$ being a 1×1 matrix, that is, a scalar. Multiplying on the right by $Ł'$,

$$Ł\mathbf{M}Ł' = \begin{pmatrix} \mathbf{I}_n + \mathbf{u}\mathbf{v}^\top - \mathbf{u}\mathbf{v}^\top & \mathbf{u} \\ \mathbf{v}^\top + (\mathbf{v}^\top\mathbf{u})\mathbf{v}^\top - (1+\mathbf{v}^\top\mathbf{u})\mathbf{v}^\top & 1 + \mathbf{v}^\top\mathbf{u} \end{pmatrix} = \begin{pmatrix} \mathbf{I}_n & \mathbf{u} \\ \mathbf{0}^\top & 1 + \mathbf{v}^\top\mathbf{u} \end{pmatrix}.$$

By Theorem 7.8.4, $\det Ł = \det \mathbf{I}_n \cdot 1 = 1$, $\det Ł' = 1$, $\det \mathbf{M} = \det(\mathbf{I}_n + \mathbf{u}\mathbf{v}^\top)\cdot 1$, and $\det(Ł\mathbf{M}Ł') = \det \mathbf{I}_n\,(1 + \mathbf{v}^\top\mathbf{u}) = 1 + \mathbf{v}^\top\mathbf{u}$. By Theorem 7.5.4, $\det(Ł\mathbf{M}Ł') = \det Ł \det \mathbf{M} \det Ł' = \det(\mathbf{I}_n + \mathbf{u}\mathbf{v}^\top)$. This proves the lemma.

■

Example. a on the Diagonal, 1 Elsewhere

Let $a \in \mathbb{R}$ and $n \ge 1$, and let $\mathbf{A}_n \in M_n(\mathbb{R})$ have every diagonal entry a and every other entry 1. Find $\det \mathbf{A}_n$.

Solution. Let $\mathbf{1} = (1,\dots,1) \in \mathbb{R}^n$. Every entry of $\mathbf{1}\mathbf{1}^\top$ is 1, so $\mathbf{A}_n = (a-1)\mathbf{I}_n + \mathbf{1}\mathbf{1}^\top$.

Case $a \ne 1$. Then $\mathbf{A}_n = (a-1)(\mathbf{I}_n + \mathbf{u}\mathbf{1}^\top)$ with $\mathbf{u} = \frac{1}{a-1}\mathbf{1}$. Scaling an $n\times n$ matrix by c multiplies its determinant by c^n (multilinearity in each of the n columns), so by Theorem 7.8.5

$$\det \mathbf{A}_n = (a-1)^n\Big(1 + \frac{n}{a-1}\Big) = (a-1)^{n-1}(a-1+n),$$

since $\mathbf{1}^\top\mathbf{u} = \frac{n}{a-1}$.

Case $a = 1$. For $n \ge 2$, all columns are equal, so $\det \mathbf{A}_n = 0$, which the formula also gives. For $n = 1$, $\det \mathbf{A}_1 = 1$, and the formula gives $(a-1)^0 \cdot 1 = 1$.

So $\det \mathbf{A}_n = (a-1)^{n-1}(a+n-1)$ for every a. For $n = 3$ this is $(a-1)^2(a+2)$; with $a = 3$ it gives $4\cdot 5 = 20$.

Warning. **The lemma is about $\mathbf{I}_n + \mathbf{u}\mathbf{v}^\top$, not $\mathbf{u}\mathbf{v}^\top$ or $\mathbf{u}^\top\mathbf{v}$.** The rank-one matrix $\mathbf{u}\mathbf{v}^\top$ is $n\times n$ and, for $n \ge 2$, has determinant 0; the scalar $\mathbf{v}^\top\mathbf{u}$ is its trace. Do not confuse $\det(\mathbf{I}_n + \mathbf{u}\mathbf{v}^\top) = 1 + \mathbf{v}^\top\mathbf{u}$ with $\det(\mathbf{I}_n) + \det(\mathbf{u}\mathbf{v}^\top) = 1$ (for $n \ge 2$), which is false whenever $\mathbf{v}^\top\mathbf{u} \ne 0$: for $\mathbf{u} = \mathbf{v} = \mathbf{e}_1$, $\mathbf{I}_n + \mathbf{e}_1\mathbf{e}_1^\top = \operatorname{diag}(2,1,\dots,1)$ has determinant 2.

7.8.5 Exercises

7.8.5.1 A. Check your understanding

Exercise. A1

(a) State the Vandermonde determinant formula, including the order of the factors.

(b) True or false: $\mathbf{V}(x_1, \ldots, x_n)$ is invertible whenever $x_1, \ldots, x_n$ are nonzero. Justify your answer.

(c) Write down the recurrence for the determinant of a tridiagonal matrix, and say which row the proof expands along.

(d) True or false: $\det \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix} = \det \mathbf{A} \det \mathbf{D}$ whenever $\mathbf{A}$ and $\mathbf{D}$ are square and $\mathbf{B} = \mathbf{O}$. Justify your answer.

(e) True or false: $\det(\mathbf{I}_n + \mathbf{u}\mathbf{v}^\top) = \det(\mathbf{I}_n + \mathbf{v}\mathbf{u}^\top)$ for all $\mathbf{u}, \mathbf{v} \in F^n$. Justify your answer.

Solution.

(a) $\det \mathbf{V}(x_1, \ldots, x_n) = \prod_{1 \le i < j \le n} (x_j - x_i)$, where $\mathbf{V}$ has (i, j)-entry x_i^{j-1} (Theorem 7.8.1).

(b) False. $\mathbf{V}(1, 1) = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$ has nonzero nodes and determinant 0. The condition is that the nodes are **distinct** (Corollary 7.8.2); 0 is allowed as a node.

(c) $D_k = a_k D_{k-1} - b_{k-1} c_{k-1} D_{k-2}$, with $D_0 = 1$, $D_1 = a_1$ (Theorem 7.8.3). The proof expands along the last row, which has only two nonzero entries.

(d) True. With $\mathbf{B} = \mathbf{O}$ the matrix is lower block triangular, and Theorem 7.8.4 applies.

(e) True. Both equal $1 + \mathbf{v}^\top \mathbf{u} = 1 + \mathbf{u}^\top \mathbf{v}$ by Theorem 7.8.5, since $\mathbf{v}^\top \mathbf{u} = \sum_i v_i u_i = \mathbf{u}^\top \mathbf{v}$.

7.8.5.2 B. Practice

Exercise. B1: A 4×4 Vandermonde determinant

Compute $\det \begin{pmatrix} 1 & -1 & 1 & -1 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 2 & 4 & 8 \end{pmatrix}$. Hence decide whether there is exactly one polynomial $p \in \mathbb{R}[x]_{\le 3}$ with $p(-1) = 3$, $p(0) = 1$, $p(1) = 0$, $p(2) = 5$.

Solution. Row i is $(1, x_i, x_i^2, x_i^3)$ with nodes $-1, 0, 1, 2$, so this is $\mathbf{V}(-1, 0, 1, 2)$. By Theorem 7.8.1 its determinant is

$$(0 - (-1))(1 - (-1))(2 - (-1))(1 - 0)(2 - 0)(2 - 1) = 1 \cdot 2 \cdot 3 \cdot 1 \cdot 2 \cdot 1 = 12.$$

(Expand along row 2, which is $(1, 0, 0, 0)$: the determinant is $(-1)^{2+1} \det \begin{pmatrix} -1 & 1 & -1 \\ 1 & 1 & 1 \\ 2 & 4 & 8 \end{pmatrix}$, that is $-(-1(8 - 4) - 1(8 - 2) - 1(4 - 2)) = -(-4 - 6 - 2) = 12$.) Hence the coefficient

matrix of the interpolation system is invertible, and there is exactly one $p \in \mathbb{R}[x]_{\le 3}$ with the given values.

Exercise. B2: The second-difference matrix

Let $\mathbf{K}_n \in M_n(\mathbb{R})$ have 2 on the diagonal, -1 directly above and below it, and 0 elsewhere. Prove that $\det \mathbf{K}_n = n+1$. Hence show that $\mathbf{K}_n$ is invertible for every n.

Solution. Here $a_k = 2$ and $b_{k-1}c_{k-1} = (-1)(-1) = 1$. By Theorem 7.8.3, $D_k = 2D_{k-1} - D_{k-2}$ for $k \ge 2$, with $D_0 = 1$, $D_1 = 2$. We show $D_k = k+1$ by strong induction on k (Theorem 0.2.10). It holds for $k = 0, 1$. If it holds for $k-1$ and $k-2$, then $D_k = 2k - (k-1) = k+1$. Hence $\det \mathbf{K}_n = D_n = n+1 \neq 0$, and $\mathbf{K}_n$ is invertible by Theorem 7.5.2.

Exercise. B3: A rank-one update

Let $\mathbf{u} = (1, 2, 3)$ and $\mathbf{v} = (1, -1, 2)$ in $\mathbb{Q}^3$.

(a) Use Theorem 7.8.5 to find $\det(\mathbf{I}_3 + \mathbf{u}\mathbf{v}^{\top})$.

(b) Write out $\mathbf{I}_3 + \mathbf{u}\mathbf{v}^{\top}$ and check your answer by elimination.

(c) Find all $t \in \mathbb{Q}$ for which $\mathbf{I}_3 + t\,\mathbf{u}\mathbf{v}^{\top}$ is **not** invertible.

Solution.

(a) $\mathbf{v}^{\top}\mathbf{u} = 1 - 2 + 6 = 5$, so $\det(\mathbf{I}_3 + \mathbf{u}\mathbf{v}^{\top}) = 6$.

(b) $\mathbf{u}\mathbf{v}^{\top} = \begin{pmatrix} 1 & -1 & 2 \\ 2 & -2 & 4 \\ 3 & -3 & 6 \end{pmatrix}$, so $\mathbf{I}_3 + \mathbf{u}\mathbf{v}^{\top} = \begin{pmatrix} 2 & -1 & 2 \\ 2 & -1 & 4 \\ 3 & -3 & 7 \end{pmatrix}$. Subtract row 1 from row 2, which does not change the determinant (Theorem 7.5.1), to get row 2 equal to $(0, 0, 2)$. Expanding along that row,

$$\det = 2 \cdot (-1)^{2+3} \det \begin{pmatrix} 2 & -1 \\ 3 & -3 \end{pmatrix} = -2(-6+3) = 6.$$

(c) $\mathbf{I}_3 + t\,\mathbf{u}\mathbf{v}^{\top} = \mathbf{I}_3 + (t\mathbf{u})\mathbf{v}^{\top}$, whose determinant is $1 + t\,\mathbf{v}^{\top}\mathbf{u} = 1 + 5t$. It is zero exactly for $t = -\frac{1}{5}$, and by Theorem 7.5.2 that is the only t for which the matrix is not invertible.

7.8.5.3 C. Going deeper

Exercise. C1: The Cauchy determinant

Let $x_1, \dots, x_n, y_1, \dots, y_n \in F$ with $x_i + y_j \neq 0$ for all i, j, and let $\mathbf{Q}_n \in M_n(F)$ have (i,j)-entry $\dfrac{1}{x_i + y_j}$. The goal is

$$\det \mathbf{Q}_n = \frac{\prod_{1 \le i < j \le n} (x_j - x_i)(y_j - y_i)}{\prod_{i=1}^{n} \prod_{j=1}^{n} (x_i + y_j)}.$$

(a) For $i < n$, subtract row n from row i. Show that the new (i,j)-entry is $\dfrac{x_n - x_i}{(x_i + y_j)(x_n + y_j)}$, and deduce that $\det \mathbf{Q}_n = \dfrac{\prod_{i<n}(x_n - x_i)}{\prod_j (x_n + y_j)} \det \mathbf{Q}'$, where $\mathbf{Q}'$ agrees with $\mathbf{Q}_n$ in rows $1, \ldots, n-1$ and has last row $(1, \ldots, 1)$.

(b) In $\mathbf{Q}'$, for $j < n$, subtract column n from column j. Deduce that $\det \mathbf{Q}' = \dfrac{\prod_{j<n}(y_n - y_j)}{\prod_{i<n}(x_i + y_n)} \det \mathbf{Q}_{n-1}$, where $\mathbf{Q}_{n-1}$ is built from $x_1, \ldots, x_{n-1}$ and $y_1, \ldots, y_{n-1}$.

(c) Prove the formula by induction on n.

(d) The **Hilbert matrix** $\mathbf{H}_n$ has (i,j)-entry $\frac{1}{i+j-1}$. Show that it is a Cauchy matrix, and use (c) to compute $\det \mathbf{H}_3$.

Hint: in (a) and (b), factor scalars out of whole rows and columns using multilinearity.

Solution.

(a) For $i < n$, $\frac{1}{x_i+y_j} - \frac{1}{x_n+y_j} = \frac{(x_n+y_j)-(x_i+y_j)}{(x_i+y_j)(x_n+y_j)} = \frac{x_n-x_i}{(x_i+y_j)(x_n+y_j)}$. These row operations do not change the determinant (Theorem 7.5.1). In the new matrix, row $i < n$ has the factor $x_n - x_i$ in every entry, and column j has the factor $\frac{1}{x_n+y_j}$ in every entry (in row n, $\frac{1}{x_n+y_j} = \frac{1}{x_n+y_j} \cdot 1$). Pulling these scalars out of each row and column by multilinearity (Theorem 7.5.1, in its row and column forms) leaves the matrix $\mathbf{Q}'$ with entries $\frac{1}{x_i+y_j}$ in rows $i < n$ and 1 in row n. This gives the stated formula.

(b) For $j < n$, subtracting column n from column j leaves the determinant unchanged, makes the last row $(0, \ldots, 0, 1)$, and for $i < n$ gives $\frac{1}{x_i+y_j} - \frac{1}{x_i+y_n} = \frac{y_n-y_j}{(x_i+y_j)(x_i+y_n)}$. Pull $y_n - y_j$ out of column j ($j < n$) and $\frac{1}{x_i+y_n}$ out of row i ($i < n$); the last row $(0, \ldots, 0, 1)$ is unaffected by the column factors because its first $n-1$ entries are 0, and row factors do not touch row n. What remains has $\frac{1}{x_i+y_j}$ in positions $i, j < n$, entries $(x_i + y_n) \cdot \frac{1}{x_i+y_n} = 1$ in column n above the corner, and last row $(0, \ldots, 0, 1)$. Expanding along the last row (Theorem 7.6.1) gives $\det \mathbf{Q}_{n-1}$, with sign $(-1)^{2n} = 1$.

(c) For $n = 1$, $\det \mathbf{Q}_1 = \frac{1}{x_1+y_1}$, which is the formula with empty products upstairs. For $n \geq 2$, combining (a) and (b),

$$\det \mathbf{Q}_n = \frac{\prod_{i<n}(x_n - x_i)(y_n - y_i)}{\prod_{j \leq n}(x_n + y_j) \prod_{i<n}(x_i + y_n)} \det \mathbf{Q}_{n-1}.$$

The numerator supplies exactly the factors of $\prod_{i<j}(x_j - x_i)(y_j - y_i)$ with $j = n$, and the denominator supplies exactly the factors $x_i + y_j$ with $\max(i,j) = n$, each once. By the induction hypothesis, $\det \mathbf{Q}_{n-1}$ supplies all the remaining factors. This proves the formula.

(d) With $x_i = i$ and $y_j = j - 1$, $x_i + y_j = i + j - 1 \geq 1$, so $\mathbf{H}_n$ is a Cauchy matrix over $\mathbb{Q}$. For $n = 3$, the numerator is $\big((2-1)(3-1)(3-2)\big)^2 = 4$, since $y_j - y_i = j - i = x_j - x_i$. The denominator is the product of all $i + j - 1$: the rows give $1 \cdot 2 \cdot 3 = 6$, $2 \cdot 3 \cdot 4 = 24$, $3 \cdot 4 \cdot 5 = 60$, in total 8640. Hence $\det \mathbf{H}_3 = \frac{4}{8640} = \frac{1}{2160}$. (In general $\det \mathbf{H}_n = c_n^4 / c_{2n}$ with $c_n = \prod_{k=1}^{n-1} k!$; we state this without proof. These determinants shrink extremely fast, which is why Hilbert matrices are a standard test of numerical methods.)

Exercise. C2: A circulant determinant over $\mathbb{C}$

Let $a, b, c \in \mathbb{C}$, let $\omega = e^{2\pi i/3}$, and let

$$\mathbf{Z} = \begin{pmatrix} a & b & c \\ c & a & b \\ b & c & a \end{pmatrix}.$$

(a) For $k = 0, 1, 2$, let $\mathbf{w}_k = (1, \omega^k, \omega^{2k})$ and $\lambda_k = a + b\omega^k + c\omega^{2k}$. Show that $\mathbf{Z}\mathbf{w}_k = \lambda_k \mathbf{w}_k$.

(b) Let $\mathbf{W}$ be the matrix with columns $\mathbf{w}_0, \mathbf{w}_1, \mathbf{w}_2$. Show that $\mathbf{Z}\mathbf{W} = \mathbf{W}\,\mathrm{diag}(\lambda_0, \lambda_1, \lambda_2)$ and that $\det \mathbf{W} \neq 0$. Deduce $\det \mathbf{Z} = \lambda_0\lambda_1\lambda_2$.

(c) Using $\omega^3 = 1$ and $1 + \omega + \omega^2 = 0$, deduce $\det \mathbf{Z} = a^3 + b^3 + c^3 - 3abc$, and check this against the Leibniz formula.

Circulant matrices of every size are treated in Chapter 12.

Solution.

(a) By Example (Roots of Unity), $\omega^3 = 1$, so $\omega^{-k} = \omega^{2k}$ and $\omega^{-2k} = \omega^k$. Row 1 of $\mathbf{Z}\mathbf{w}_k$ is $a + b\omega^k + c\omega^{2k} = \lambda_k$. Row 2 is $c + a\omega^k + b\omega^{2k} = \omega^k(c\omega^{-k} + a + b\omega^k) = \omega^k(a + b\omega^k + c\omega^{2k}) = \omega^k\lambda_k$. Row 3 is $b + c\omega^k + a\omega^{2k} = \omega^{2k}(b\omega^{-2k} + c\omega^{-k} + a) = \omega^{2k}\lambda_k$. Hence $\mathbf{Z}\mathbf{w}_k = \lambda_k\mathbf{w}_k$.

(b) By Theorem 0.9.2, column k of $\mathbf{Z}\mathbf{W}$ is $\mathbf{Z}\mathbf{w}_k = \lambda_k\mathbf{w}_k$, column k of $\mathbf{W}\,\mathrm{diag}(\lambda_0, \lambda_1, \lambda_2)$. The (i, k)-entry of $\mathbf{W}$ is $(\omega^k)^{i-1}$, so $\mathbf{W} = \mathbf{V}(1, \omega, \omega^2)^\top$. The nodes $1, \omega, \omega^2$ are distinct (they are $e^{2\pi ik/3}$ for $k = 0, 1, 2$), so $\det \mathbf{W} = \det \mathbf{V}(1, \omega, \omega^2) \neq 0$ by Theorem 7.4.5 and Corollary 7.8.2. By Theorem 7.5.4 and Theorem 7.4.6, $\det \mathbf{Z} \det \mathbf{W} = \det \mathbf{W}\,\lambda_0\lambda_1\lambda_2$, and canceling $\det \mathbf{W} \neq 0$ gives $\det \mathbf{Z} = \lambda_0\lambda_1\lambda_2$.

(c) Multiply out $\lambda_1\lambda_2 = (a + b\omega + c\omega^2)(a + b\omega^2 + c\omega)$. The nine terms are a^2, $ab\omega^2 + ab\omega$, $ac\omega + ac\omega^2$, $b^2\omega^3$, $bc\omega^2$, $bc\omega^4 = bc\omega$, $c^2\omega^3$. Using $\omega^3 = 1$ and $\omega + \omega^2 = -1$,

$$\lambda_1\lambda_2 = a^2 + b^2 + c^2 - ab - ac - bc.$$

Then $\det \mathbf{Z} = (a + b + c)(a^2 + b^2 + c^2 - ab - ac - bc) = a^3 + b^3 + c^3 - 3abc$, expanding and canceling. The Leibniz formula for $\mathbf{Z}$ gives the three even terms $a \cdot a \cdot a$, $c \cdot c \cdot c$ (from $\sigma = (1\ 2\ 3)$), and $b \cdot b \cdot b$, and the three odd terms, each equal to abc with sign $-$: the same answer.

Exercise. C3: Rank-one updates of an invertible matrix

Let $\mathbf{A} \in M_n(F)$ be invertible and $\mathbf{u}, \mathbf{v} \in F^n$.

(a) Prove that $\det(\mathbf{A} + \mathbf{u}\mathbf{v}^\top) = \det \mathbf{A}\,(1 + \mathbf{v}^\top\mathbf{A}^{-1}\mathbf{u})$.

(b) Deduce that $\mathbf{A} + \mathbf{u}\mathbf{v}^\top$ is invertible if and only if $\mathbf{v}^\top\mathbf{A}^{-1}\mathbf{u} \neq -1$.

(c) Give an example over $\mathbb{Q}$ with $n = 2$ where $\mathbf{A}$ is invertible, $\mathbf{u}\mathbf{v}^\top \neq \mathbf{O}$, and $\mathbf{A} + \mathbf{u}\mathbf{v}^\top$ is **not** invertible.

Solution.

(a) Since $\mathbf{A}$ is invertible, $\mathbf{A} + \mathbf{u}\mathbf{v}^\top = \mathbf{A}(\mathbf{I}_n + \mathbf{A}^{-1}\mathbf{u}\mathbf{v}^\top) = \mathbf{A}(\mathbf{I}_n + (\mathbf{A}^{-1}\mathbf{u})\mathbf{v}^\top)$ by Theo-

rem 0.9.4. By Theorem 7.5.4 and Theorem 7.8.5 applied to the vectors $\mathbf{A}^{-1}\mathbf{u}$ and $\mathbf{v}$, $\det(\mathbf{A}+\mathbf{u}\mathbf{v}^\top) = \det\mathbf{A}\,(1+\mathbf{v}^\top\mathbf{A}^{-1}\mathbf{u})$.

(b) $\det\mathbf{A} \neq 0$ by Theorem 7.5.2, so $\det(\mathbf{A}+\mathbf{u}\mathbf{v}^\top) \neq 0$ exactly when $1+\mathbf{v}^\top\mathbf{A}^{-1}\mathbf{u} \neq 0$. Apply Theorem 7.5.2 again.

(c) Take $\mathbf{A}=\mathbf{I}_2$, $\mathbf{u}=(-1,0)$, $\mathbf{v}=(1,0)$. Then $\mathbf{v}^\top\mathbf{u}=-1$, and indeed $\mathbf{A}+\mathbf{u}\mathbf{v}^\top = \begin{pmatrix}0&0\\0&1\end{pmatrix}$ is not invertible.

7.9 Orientation and Volume

The chapter began in §1 with a picture: the area of a parallelogram, with a sign attached, obeys a few simple rules, and those rules pinned down $ad-bc$. We have since built the determinant in every size from the same rules. It is time to return to geometry and say precisely what the determinant measures. There are two halves to the answer. Its absolute value measures volume, and its sign measures orientation, the difference between a left hand and a right hand. Both halves need care. "Volume" in $\mathbb{R}^7$ has no meaning until someone defines it, and "orientation" has no meaning at all over $\mathbb{C}$.

Throughout this section the field is $\mathbb{R}$.

7.9.1 What should the volume of a parallelepiped be?

For vectors $\mathbf{a}_1,\dots,\mathbf{a}_n \in \mathbb{R}^n$, the **parallelepiped** they span is the set

$$P(\mathbf{a}_1,\dots,\mathbf{a}_n) := \{t_1\mathbf{a}_1+\cdots+t_n\mathbf{a}_n : 0 \le t_i \le 1 \text{ for every } i\}.$$

For $n=2$ it is the parallelogram with sides $\mathbf{a}_1,\mathbf{a}_2$ at the origin, and for $n=3$ a slanted box. For $n=1$ it is the segment from 0 to $\mathbf{a}_1$.

We would like to say "its volume is $|\det\begin{pmatrix}\mathbf{a}_1 & \cdots & \mathbf{a}_n\end{pmatrix}|$". But a theorem needs two sides, and one side is missing: we have no definition of the volume of a subset of $\mathbb{R}^n$. School geometry supplies one for $n=2$ and $n=3$ through "base times height", and the proper general notion, Lebesgue measure, belongs to a course in measure theory. So we proceed honestly in three steps. We **define** the volume of a parallelepiped by the determinant. We **prove** that this is the only function obeying the rules that any reasonable notion of volume must obey. And we **state**, without proof, that it agrees with Lebesgue measure.

The volume of a parallelepiped is the size of the determinant of its edge vectors.

Definition 7.9.1. Volume of a Parallelepiped

Let $n \ge 1$ and $\mathbf{a}_1,\dots,\mathbf{a}_n \in \mathbb{R}^n$. The ($n$-dimensional) **volume** of the parallelepiped spanned by $\mathbf{a}_1,\dots,\mathbf{a}_n$ is

$$\operatorname{vol}(\mathbf{a}_1,\dots,\mathbf{a}_n) := \left|\det\begin{pmatrix}\mathbf{a}_1 & \mathbf{a}_2 & \cdots & \mathbf{a}_n\end{pmatrix}\right| \ge 0.$$

In words: stack the n edge vectors as the columns of a square matrix, take its determinant, and forget the sign. There must be **exactly** n vectors in $\mathbb{R}^n$, so that the matrix is square.

Examples.

- The unit cube $P(\mathbf{e}_1, \dots, \mathbf{e}_n)$ has volume $|\det \mathbf{I}_n| = 1$.
- A box with edges $d_1\mathbf{e}_1, \dots, d_n\mathbf{e}_n$ has volume $|\det \operatorname{diag}(d_1, \dots, d_n)| = |d_1 \cdots d_n|$ by Theorem 7.4.6: the product of its side lengths, as it should be.
- The parallelepiped spanned by $(1, 2, 0)$, $(0, 1, 1)$, $(1, 0, 3)$ in $\mathbb{R}^3$ has volume

$$\left| \det \begin{pmatrix} 1 & 0 & 1 \\ 2 & 1 & 0 \\ 0 & 1 & 3 \end{pmatrix} \right| = |1 \cdot (3 - 0) - 0 + 1 \cdot (2 - 0)| = 5,$$

  expanding along the first row.
- **Degenerate case.** If $\mathbf{a}_1, \dots, \mathbf{a}_n$ are linearly dependent, the parallelepiped is flattened into a proper subspace, and its volume is 0 by Theorem 7.2.1. For $n = 2$, $P((1, 2), (2, 4))$ is the segment from $(0, 0)$ to $(3, 6)$, with area $|4 - 4| = 0$. This case matters: "volume zero" is the geometric face of "not invertible".

Nonexample by minimal change. Drop one vector: $(1, 2, 0)$ and $(0, 1, 1)$ span a genuine parallelogram in $\mathbb{R}^3$, but the matrix $\begin{pmatrix} 1 & 0 \\ 2 & 1 \\ 0 & 1 \end{pmatrix}$ is 3×2 and has no determinant. The clause that fails is **exactly n vectors in $\mathbb{R}^n$**. The 3-dimensional volume of that flat parallelogram would be 0, and its area needs a different formula (Exercise (C1: Area of a triangle in space)).

Why this definition. Reordering the vectors multiplies the determinant by a sign (Theorem 7.2.1), which the absolute value removes, so the volume does not depend on the order in which the edges are listed. The absolute value is also what makes volume nonnegative. The sign we threw away is not garbage: it is the orientation, the subject of the second half of this section.

Warning. **Volume is not linear.** Scaling **one** edge by 2 doubles the volume, but scaling the whole parallelepiped by 2 multiplies it by 2^n: $\operatorname{vol}(2\mathbf{a}_1, \dots, 2\mathbf{a}_n) = |\det(2\mathbf{A})| = 2^n \operatorname{vol}(\mathbf{a}_1, \dots, \mathbf{a}_n)$, where $\mathbf{A}$ has columns $\mathbf{a}_1, \dots, \mathbf{a}_n$. A cube of side 2 in $\mathbb{R}^3$ has volume 8, not 2; this failure persists even with the sign kept, since det is linear in each edge separately, not in all edges at once. The absolute value also destroys additivity in a single edge: $\operatorname{vol}(\mathbf{a}_1 + \mathbf{b}_1, \mathbf{a}_2, \dots)$ is in general not $\operatorname{vol}(\mathbf{a}_1, \mathbf{a}_2, \dots) + \operatorname{vol}(\mathbf{b}_1, \mathbf{a}_2, \dots)$: with $\mathbf{a}_1 = \mathbf{e}_1$, $\mathbf{b}_1 = -\mathbf{e}_1$ in $\mathbb{R}^2$, and $\mathbf{a}_2 = \mathbf{e}_2$, the left side is 0 and the right side is 2.

7.9.2 Volume is forced by its rules

Why should anyone accept Definition 7.9.1? Because of what any notion of volume of parallelepipeds must satisfy. The pictures in §1 showed three facts about area, and they have obvious n-dimensional versions: stretching one edge by a factor c multiplies the volume by $|c|$; sliding one edge parallel to another (a shear) does not change the volume; and the unit cube has volume 1. The next theorem says that these three rules alone determine the volume completely. So any definition that obeys them, whether from school geometry or from measure theory, must agree with ours.

Theorem 7.9.1. Volume Is Determined by Its Rules

Let $n \geq 1$, and let g be a function assigning a real number $g(\mathbf{a}_1, \dots, \mathbf{a}_n)$ to every list of n vectors in $\mathbb{R}^n$. Suppose that for every list, every $i \neq j$, and every $c \in \mathbb{R}$:

(V1) $g(\mathbf{a}_1, \dots, c\mathbf{a}_i, \dots, \mathbf{a}_n) = |c|\, g(\mathbf{a}_1, \dots, \mathbf{a}_i, \dots, \mathbf{a}_n)$ (only slot i changes);
(V2) $g(\mathbf{a}_1, \dots, \mathbf{a}_i + c\mathbf{a}_j, \dots, \mathbf{a}_n) = g(\mathbf{a}_1, \dots, \mathbf{a}_i, \dots, \mathbf{a}_n)$ (only slot i changes);
(V3) $g(\mathbf{e}_1, \dots, \mathbf{e}_n) = 1$.

Then $g(\mathbf{a}_1, \dots, \mathbf{a}_n) = \mathrm{vol}(\mathbf{a}_1, \dots, \mathbf{a}_n)$ for every list. Moreover, vol itself satisfies (V1)–(V3).

Idea. Think of g as a function of the matrix $\mathbf{M}$ with columns $\mathbf{a}_1, \dots, \mathbf{a}_n$. ① If the columns are dependent, shear one column to $\mathbf{0}$ and scale by 0: $g = 0$. ② Each elementary column operation changes g by the factor $|\det \mathbf{E}|$. Scaling and shearing are (V1) and (V2); a swap is not among the rules, but three shears and a sign change perform it. ③ An invertible $\mathbf{M}$ is a product of elementary matrices, so peel them off one at a time until only $\mathbf{I}$ is left.

Proof. For $\mathbf{M} \in M_n(\mathbb{R})$ with columns $\mathbf{a}_1, \dots, \mathbf{a}_n$, write $h(\mathbf{M}) = g(\mathbf{a}_1, \dots, \mathbf{a}_n)$.

Step 1: dependent columns give $h(\mathbf{M}) = 0$. Suppose the columns are linearly dependent. By Theorem 1.4.1, some $\mathbf{a}_i = c_1\mathbf{a}_1 + \cdots + c_{i-1}\mathbf{a}_{i-1}$ (for $i = 1$, $\mathbf{a}_1 = \mathbf{0}$). Applying (V2) to slot i with the scalars $-c_1, \dots, -c_{i-1}$, one at a time, does not change g, and it replaces $\mathbf{a}_i$ by $\mathbf{0}$ while the other slots stay fixed. Then (V1) with $c = 0$ gives $g(\dots, \mathbf{0}, \dots) = g(\dots, 0\cdot\mathbf{0}, \dots) = 0 \cdot g(\dots, \mathbf{0}, \dots) = 0$. Hence $h(\mathbf{M}) = 0$.

Step 2: $h(\mathbf{ME}) = |\det \mathbf{E}|\, h(\mathbf{M})$ for every elementary matrix $\mathbf{E}$. By Theorem 3.4.1 and Theorem 0.9.5, $\mathbf{ME} = (\mathbf{E}^\top \mathbf{M}^\top)^\top$, and $\mathbf{E}^\top$ is the elementary matrix $\mathbf{P}_{ij}$, $\mathbf{D}_i(c)$, or $\mathbf{I} + c\mathbf{E}_{ji}$ when $\mathbf{E}$ is $\mathbf{P}_{ij}$, $\mathbf{D}_i(c)$, or $\mathbf{I} + c\mathbf{E}_{ij}$. So right multiplication by $\mathbf{E}$ performs a column operation: swap columns i and j; multiply column i by $c \neq 0$; or add c times column i to column j. The determinants are $\det \mathbf{P}_{ij} = -1$, $\det \mathbf{D}_i(c) = c$, and $\det(\mathbf{I} + c\mathbf{E}_{ij}) = 1$ by Theorem 7.5.1.

- For $\mathbf{E} = \mathbf{I} + c\mathbf{E}_{ij}$, (V2) gives $h(\mathbf{ME}) = h(\mathbf{M}) = |1|\, h(\mathbf{M})$.
- For $\mathbf{E} = \mathbf{D}_i(c)$, (V1) gives $h(\mathbf{ME}) = |c|\, h(\mathbf{M})$.
- For $\mathbf{E} = \mathbf{P}_{ij}$, watch only slots i and j, holding $\mathbf{x}$ and $\mathbf{y}$. By (V2) three times, $(\mathbf{x}, \mathbf{y}) \to (\mathbf{x}+\mathbf{y}, \mathbf{y}) \to (\mathbf{x}+\mathbf{y}, -\mathbf{x}) \to (\mathbf{y}, -\mathbf{x})$ leaves g unchanged, where the second step subtracts slot i from slot j and the third adds slot j to slot i. By (V1) with $c = -1$ on slot j, g at $(\mathbf{y}, \mathbf{x})$ equals $|-1|$ times g at $(\mathbf{y}, -\mathbf{x})$. Hence $h(\mathbf{MP}_{ij}) = h(\mathbf{M}) = |-1|\, h(\mathbf{M})$.

Step 3: conclusion. If the columns of $\mathbf{M}$ are dependent, then $h(\mathbf{M}) = 0$ by Step 1 and $\det \mathbf{M} = 0$ by Theorem 7.2.1, so $h(\mathbf{M}) = |\det \mathbf{M}|$. Otherwise $\mathbf{M}$ is invertible by Theorem 3.4.4, hence $\mathbf{M} = \mathbf{E}_1\mathbf{E}_2 \cdots \mathbf{E}_k$ is a product of elementary matrices by the same theorem. Applying Step 2 k times, starting from $\mathbf{I}$, and then (V3),

$$h(\mathbf{M}) = h(\mathbf{I}\mathbf{E}_1 \cdots \mathbf{E}_k) = |\det \mathbf{E}_1| \cdots |\det \mathbf{E}_k|\, h(\mathbf{I}) = |\det(\mathbf{E}_1 \cdots \mathbf{E}_k)| = |\det \mathbf{M}|,$$

where the third equality uses Theorem 7.5.4 and $|xy| = |x||y|$. This proves $g = \mathrm{vol}$.

Finally, vol satisfies (V1) because the determinant is linear in each column (Theorem 7.4.3) and $|c\,d| = |c||d|$; (V2) by Theorem 7.2.1; and (V3) because $\det \mathbf{I}_n = 1$.

■

Agreement with area and volume from school. For $n = 2$, let $g(\mathbf{a}_1, \mathbf{a}_2)$ be the area of the parallelogram $P(\mathbf{a}_1, \mathbf{a}_2)$ in the sense of school geometry, base times height. The pictures of §1 are exactly (V1)–(V3) for this g: stretching a side by c stretches the base or the height by

$|c|$; a shear $\mathbf{a}_1 \mapsto \mathbf{a}_1 + c\mathbf{a}_2$ slides the opposite side along a line parallel to the base $\mathbf{a}_2$, keeping base and height; and the unit square has area 1. Hence, by Theorem 7.9.1, school area equals $|\det\begin{pmatrix}\mathbf{a}_1 & \mathbf{a}_2\end{pmatrix}|$. For $n = 3$ the same argument works with "area of base times height". What the theorem cannot supply is the geometric input itself: it takes the three rules as given.

Agreement with Lebesgue measure. In measure theory, every reasonable subset $S \subseteq \mathbb{R}^n$ has an n-dimensional volume, its Lebesgue measure. One can prove that the Lebesgue measure of $P(\mathbf{a}_1, \dots, \mathbf{a}_n)$ satisfies (V1)–(V3), and hence equals $\mathrm{vol}(\mathbf{a}_1, \dots, \mathbf{a}_n)$. We state this without proof; nothing in this book depends on it.

Quick check. Which of the rules (V1)–(V3) does the function $g(\mathbf{a}_1, \mathbf{a}_2) = |\det\begin{pmatrix}\mathbf{a}_1 & \mathbf{a}_2\end{pmatrix}|^2$ on $\mathbb{R}^2$ break?

Solution. It satisfies (V2), since shears do not change the determinant, and (V3), since $1^2 = 1$. It breaks (V1): $g(2\mathbf{e}_1, \mathbf{e}_2) = 4$, while $|2|\, g(\mathbf{e}_1, \mathbf{e}_2) = 2$. So the theorem does not apply, and indeed $g \neq \mathrm{vol}$.

7.9.3 Linear maps scale volume

Applying a linear map to a parallelepiped gives another parallelepiped: by linearity, $T(\sum t_i \mathbf{a}_i) = \sum t_i T\mathbf{a}_i$, so $T(P(\mathbf{a}_1, \dots, \mathbf{a}_n)) = P(T\mathbf{a}_1, \dots, T\mathbf{a}_n)$. How does the volume change? Multiplicativity answers this in one line, and the answer does not depend on the parallelepiped.

Theorem 7.9.2. Linear Maps Scale Volume by $|\det|$

Let $T \in \mathcal{L}(\mathbb{R}^n)$ and $\mathbf{a}_1, \dots, \mathbf{a}_n \in \mathbb{R}^n$. Then

$$\mathrm{vol}(T\mathbf{a}_1, \dots, T\mathbf{a}_n) = |\det T| \cdot \mathrm{vol}(\mathbf{a}_1, \dots, \mathbf{a}_n).$$

In particular, the image of the unit cube under T has volume $|\det T|$.

Proof. Let $\mathbf{B}$ be the matrix of T in the standard basis, so $T\mathbf{x} = \mathbf{B}\mathbf{x}$ for all $\mathbf{x}$ (Theorem 4.1.1) and $\det T = \det \mathbf{B}$ (Definition 7.5.1). Let $\mathbf{A} = \begin{pmatrix}\mathbf{a}_1 & \cdots & \mathbf{a}_n\end{pmatrix}$. By Theorem 0.9.2, $\mathbf{BA} = \begin{pmatrix}\mathbf{B}\mathbf{a}_1 & \cdots & \mathbf{B}\mathbf{a}_n\end{pmatrix}$. By Theorem 7.5.4,

$$\mathrm{vol}(T\mathbf{a}_1, \dots, T\mathbf{a}_n) = |\det(\mathbf{BA})| = |\det \mathbf{B}|\,|\det \mathbf{A}| = |\det T|\ \mathrm{vol}(\mathbf{a}_1, \dots, \mathbf{a}_n).$$

For the unit cube take $\mathbf{a}_i = \mathbf{e}_i$, with $\mathrm{vol}(\mathbf{e}_1, \dots, \mathbf{e}_n) = 1$.

■

This is the meaning of the determinant of a real operator: **the factor by which it scales every volume**. An operator with $\det T = 0$ squashes $\mathbb{R}^n$ into a lower-dimensional subspace, where n-dimensional volume is 0. In multivariable calculus the same fact, applied to the linear approximation of a smooth map, becomes the Jacobian factor $|\det \mathbf{J}|$ in the change-of-variables formula for integrals.

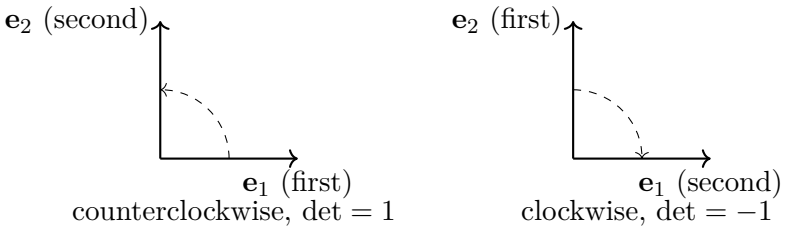
Example. Area of an Image

Let $T\colon \mathbb{R}^2 \to \mathbb{R}^2$ be $T(x, y) = (3x+y,\ x+2y)$. Find the area of the image of the unit square, and of the image of the parallelogram $P((1,1),(-1,2))$.

Solution. The standard matrix is $\mathbf{B} = \begin{pmatrix} 3 & 1 \\ 1 & 2 \end{pmatrix}$, with $\det T = 6 - 1 = 5$. By Theorem 7.9.2 the image of the unit square, the parallelogram spanned by $(3,1)$ and $(1,2)$, has area 5. The parallelogram $P((1,1),(-1,2))$ has area $|2+1| = 3$, so its image has area $5 \cdot 3 = 15$. Check directly: $T(1,1) = (4,3)$ and $T(-1,2) = (-1,3)$, and $|\det \begin{pmatrix} 4 & -1 \\ 3 & 3 \end{pmatrix}| = |12+3| = 15$.

7.9.4 Orientation

Now the sign. In $\mathbb{R}^2$, the ordered basis $(\mathbf{e}_1, \mathbf{e}_2)$ turns **counterclockwise** from its first vector to its second, while $(\mathbf{e}_2, \mathbf{e}_1)$ turns clockwise. No rotation carries one arrangement to the other: a rotation of the plane keeps counterclockwise turns counterclockwise. The determinant sees the difference: $\det \begin{pmatrix} \mathbf{e}_1 & \mathbf{e}_2 \end{pmatrix} = 1 > 0$ and $\det \begin{pmatrix} \mathbf{e}_2 & \mathbf{e}_1 \end{pmatrix} = -1 < 0$.



To compare two bases of an arbitrary real space, we compare them with each other rather than with a fixed standard basis, using the change-of-coordinates matrix of Chapter 4.

Two bases have the same orientation when the matrix converting one into the other has positive determinant.

Definition 7.9.2. Orientation

Let V be a vector space **over** $\mathbb{R}$ with $\dim V = n \geq 1$, and let $\mathcal{B}$ and $\mathcal{C}$ be **ordered** bases of V. We say $\mathcal{B}$ and $\mathcal{C}$ have the **same orientation** if

$$\det [\mathrm{id}]_{\mathcal{B}}^{\mathcal{C}} > 0,$$

and **opposite orientations** if $\det [\mathrm{id}]_{\mathcal{B}}^{\mathcal{C}} < 0$.

In words: write each vector of $\mathcal{B}$ in coordinates with respect to $\mathcal{C}$, put the coordinate columns side by side (Definition 4.2.1), and look at the **sign** of the determinant. Exactly one of the two cases occurs, because $[\mathrm{id}]_{\mathcal{B}}^{\mathcal{C}}$ is invertible (Theorem 4.2.1 (c)), so its determinant is nonzero (Theorem 7.5.2), and a nonzero real number is either positive or negative. The bases must be **ordered**: reordering a basis can change the answer.

For $V = \mathbb{R}^n$ there is a shortcut. If $\mathbf{B}$ and $\mathbf{C}$ are the matrices whose columns are the vectors of

$\mathcal{B}$ and $\mathcal{C}$, then $[\mathrm{id}]_{\mathcal{B}}^{\mathcal{E}} = \mathbf{B}$ and $[\mathrm{id}]_{\mathcal{C}}^{\mathcal{E}} = \mathbf{C}$ for the standard basis $\mathcal{E}$. By Theorem 4.2.1 (b) and (c), $[\mathrm{id}]_{\mathcal{B}}^{\mathcal{C}} = \mathbf{C}^{-1}\mathbf{B}$. Since $\det \mathbf{C} \det \mathbf{C}^{-1} = \det \mathbf{I} = 1$ (Theorem 7.5.4), $\det[\mathrm{id}]_{\mathcal{B}}^{\mathcal{C}} = \det \mathbf{B} / \det \mathbf{C}$. So $\mathcal{B}$ **and** $\mathcal{C}$ **have the same orientation if and only if** $\det \mathbf{B}$ **and** $\det \mathbf{C}$ **have the same sign.**

Examples.

- In $\mathbb{R}^2$, $(\mathbf{e}_1, \mathbf{e}_2)$ and $(\mathbf{e}_2, \mathbf{e}_1)$ have opposite orientations: $\det \mathbf{I}_2 = 1$ and $\det \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = -1$.
- In $\mathbb{R}^3$, $(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$ and $(\mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_1)$ have the **same** orientation. The second matrix is the permutation matrix of a 3-cycle, whose sign is $(-1)^{3-1} = 1$. A cyclic shift of three vectors is two swaps.
- In $V = \mathbb{R}[x]_{\leq 1}$, the bases $\mathcal{B} = (1, x)$ and $\mathcal{C} = (x, 1 + x)$ have $1 = -x + (1 + x)$ and $x = x + 0 \cdot (1 + x)$, so $[\mathrm{id}]_{\mathcal{B}}^{\mathcal{C}} = \begin{pmatrix} -1 & 1 \\ 1 & 0 \end{pmatrix}$, with determinant -1. They have opposite orientations. Orientation makes sense in any real space, with no picture at all.
- **Degenerate case,** $n = 1$**.** Bases of $\mathbb{R}$ are single nonzero numbers (a), and $[\mathrm{id}]_{(a)}^{(b)} = (a/b)$. So (a) and (b) have the same orientation exactly when a and b have the same sign. The two orientations of a line are its two directions.

Nonexample by minimal change. In $\mathbb{R}^2$, $(-\mathbf{e}_1, -\mathbf{e}_2)$ has the same orientation as $(\mathbf{e}_1, \mathbf{e}_2)$: $\det(-\mathbf{I}_2) = 1$, and indeed it is the rotation of $(\mathbf{e}_1, \mathbf{e}_2)$ by a half-turn. Change only **one** sign, to $(\mathbf{e}_1, -\mathbf{e}_2)$: the determinant becomes -1, and the orientation flips. What changed is the sign of the determinant, which is the whole content of the definition. In $\mathbb{R}^3$, by contrast, $-\mathbf{I}_3$ has determinant -1: negating all three vectors reverses orientation.

Why this definition. Using the change-of-coordinates matrix, rather than the matrices $\mathbf{B}$ and $\mathbf{C}$ above, makes the definition work in every real space, including spaces like $\mathbb{R}[x]_{\leq 1}$ that have no standard picture. Requiring $\det > 0$ rather than $\det = 1$ is what gives a coarse, two-valued answer: rescaling a basis vector by a positive number should not change its handedness.

The payoff is that "same orientation" really sorts all bases into two families, as the pictures suggest.

Theorem 7.9.3. There Are Exactly Two Orientations

Let V be a vector space over $\mathbb{R}$ with $\dim V = n \geq 1$.

(a) "Having the same orientation" is an equivalence relation on the set of ordered bases of V.

(b) It has **exactly two** equivalence classes. If $\mathcal{B} = (v_1, \ldots, v_n)$, then $\mathcal{B}^- = (-v_1, v_2, \ldots, v_n)$ lies in the class not containing $\mathcal{B}$.

Idea. The three axioms are three facts about change-of-coordinates matrices: $[\mathrm{id}]_{\mathcal{B}}^{\mathcal{B}} = \mathbf{I}$; the matrix back is the inverse; and matrices compose along a chain $\mathcal{B} \to \mathcal{C} \to \mathcal{D}$. Determinants turn these into $1 > 0$, $1/\text{positive} > 0$, and $\text{positive} \cdot \text{positive} > 0$. For "exactly two", flipping one vector gives a second class, and "negative times negative is positive" shows there is no third.

Proof. Write $d(\mathcal{B},\mathcal{C}) = \det[\mathrm{id}]_{\mathcal{B}}^{\mathcal{C}}$. By Theorem 4.2.1 (b) and Theorem 7.5.4, for bases $\mathcal{B},\mathcal{C},\mathcal{D}$,

$$d(\mathcal{B},\mathcal{D}) = d(\mathcal{C},\mathcal{D})\, d(\mathcal{B},\mathcal{C}). \tag{$*$}$$

(a) *Reflexive:* $[\mathrm{id}]_{\mathcal{B}}^{\mathcal{B}} = \mathbf{I}_n$, so $d(\mathcal{B},\mathcal{B}) = 1 > 0$. *Symmetric:* by $(*)$ with $\mathcal{D} = \mathcal{B}$, $d(\mathcal{C},\mathcal{B})\, d(\mathcal{B},\mathcal{C}) = d(\mathcal{B},\mathcal{B}) = 1$, so $d(\mathcal{C},\mathcal{B}) = 1/d(\mathcal{B},\mathcal{C})$, which is positive when $d(\mathcal{B},\mathcal{C})$ is. *Transitive:* if $d(\mathcal{B},\mathcal{C}) > 0$ and $d(\mathcal{C},\mathcal{D}) > 0$, then $d(\mathcal{B},\mathcal{D}) > 0$ by $(*)$, since a product of positive reals is positive. By Definition 0.5.2, this is an equivalence relation.

(b) *At least two classes.* $\mathcal{B}^-$ is a basis, and $[\mathrm{id}]_{\mathcal{B}^-}^{\mathcal{B}} = \mathrm{diag}(-1,1,\dots,1)$, because $-v_1 = (-1)v_1$ and $v_j = 1 \cdot v_j$. Its determinant is -1 by Theorem 7.4.6. So $d(\mathcal{B}^-,\mathcal{B}) < 0$, and $\mathcal{B}^-$ and $\mathcal{B}$ lie in different classes.

At most two classes. Let $\mathcal{C}$ and $\mathcal{D}$ be bases, neither of which has the same orientation as $\mathcal{B}$. Then $d(\mathcal{C},\mathcal{B}) < 0$ and $d(\mathcal{D},\mathcal{B}) < 0$, and by symmetry $d(\mathcal{B},\mathcal{D}) < 0$ as well. By $(*)$, $d(\mathcal{C},\mathcal{D}) = d(\mathcal{B},\mathcal{D})\, d(\mathcal{C},\mathcal{B})$ is a product of two negative reals, hence positive. So all bases outside the class of $\mathcal{B}$ lie in one class, and there are exactly two classes.

■

Choosing one of the two classes is called **orienting** V; its bases are then called **positively oriented**, and the others **negatively oriented**. On $\mathbb{R}^n$ the standard choice is the class of $(\mathbf{e}_1,\dots,\mathbf{e}_n)$. By the shortcut above, $(\mathbf{a}_1,\dots,\mathbf{a}_n)$ is positively oriented if and only if $\det\begin{pmatrix}\mathbf{a}_1 & \cdots & \mathbf{a}_n\end{pmatrix} > 0$. So the determinant of a real square matrix carries two independent pieces of information: its absolute value is a volume, and its sign is an orientation.

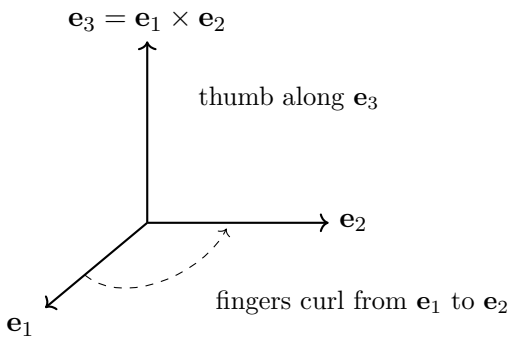
Warning. **Orientation needs $\mathbb{R}$.** The proof used two facts about the reals: every nonzero number is positive or negative, and products and inverses of positive numbers are positive. Over $\mathbb{C}$ there is no such splitting. Suppose we tried "$\mathcal{B}$ and $\mathcal{C}$ have the same orientation if $d(\mathcal{B},\mathcal{C})$ has positive real part". In $\mathbb{C}^1$, with $z = e^{i\pi/3}$, the bases $(\mathbf{e}_1)$, $(z\mathbf{e}_1)$, $(z^2\mathbf{e}_1)$ give $d((\mathbf{e}_1),(z\mathbf{e}_1)) = z^{-1}$ and $d((z\mathbf{e}_1),(z^2\mathbf{e}_1)) = z^{-1}$, both with real part $\frac{1}{2} > 0$, but $d((\mathbf{e}_1),(z^2\mathbf{e}_1)) = z^{-2}$ has real part $-\frac{1}{2}$: transitivity fails. Over $\mathbb{F}_2$, $-1 = 1$, so the second basis $\mathcal{B}^-$ of the proof is $\mathcal{B}$ itself. The theorem holds verbatim over any field with a compatible order, such as $\mathbb{Q}$, but not over $\mathbb{C}$ or a finite field.

Quick check. Do $((1,2),(3,4))$ and $((3,4),(1,2))$ have the same orientation? Does $((1,2),(3,4))$ have the standard orientation of $\mathbb{R}^2$?

Solution. $\det\begin{pmatrix}1 & 3\\ 2 & 4\end{pmatrix} = 4 - 6 = -2$ and $\det\begin{pmatrix}3 & 1\\ 4 & 2\end{pmatrix} = 6 - 4 = 2$. The signs differ, so the two bases have opposite orientations. The first has negative determinant, so it is negatively oriented: it is **not** in the class of $(\mathbf{e}_1,\mathbf{e}_2)$.

7.9.5 The right-hand rule

In $\mathbb{R}^3$, the standard orientation is the one described by the right hand: curl the fingers of the right hand from the first basis vector toward the second, and the thumb points along the third. For $(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$ with the usual picture of the axes, this is the familiar arrangement below.



The **cross product** of $\mathbf{a} = (a_1, a_2, a_3)$ and $\mathbf{b} = (b_1, b_2, b_3)$ in $\mathbb{R}^3$ is

$$\mathbf{a} \times \mathbf{b} \coloneqq (a_2b_3 - a_3b_2,\ a_3b_1 - a_1b_3,\ a_1b_2 - a_2b_1).$$

Physics defines it by the right-hand rule. Here is the linear-algebra reason the two descriptions match: $(\mathbf{a}, \mathbf{b}, \mathbf{a} \times \mathbf{b})$ is always positively oriented.

Expand $\det\begin{pmatrix}\mathbf{a} & \mathbf{b} & \mathbf{a} \times \mathbf{b}\end{pmatrix}$ along its third column (Theorem 7.6.1). The cofactors of that column are $C_{13} = a_2b_3 - a_3b_2$, $C_{23} = -(a_1b_3 - a_3b_1)$ and $C_{33} = a_1b_2 - a_2b_1$, which are exactly the three components of $\mathbf{a} \times \mathbf{b}$. Hence

$$\det\begin{pmatrix}\mathbf{a} & \mathbf{b} & \mathbf{a} \times \mathbf{b}\end{pmatrix} = (a_2b_3 - a_3b_2)^2 + (a_3b_1 - a_1b_3)^2 + (a_1b_2 - a_2b_1)^2 \ \geq\ 0.$$

If $\mathbf{a}$ and $\mathbf{b}$ are linearly independent, the 3×2 matrix $\begin{pmatrix}\mathbf{a} & \mathbf{b}\end{pmatrix}$ has rank 2, so one of its 2×2 minors, which are $\pm$ the components of $\mathbf{a} \times \mathbf{b}$, is nonzero (Theorem 7.7.1). Then the determinant is strictly positive, so $(\mathbf{a}, \mathbf{b}, \mathbf{a} \times \mathbf{b})$ is a basis (Theorem 7.5.2) with the standard orientation. For example, $\mathbf{e}_1 \times \mathbf{e}_2 = (0, 0, 1) = \mathbf{e}_3$, matching the picture.

7.9.6 Orientation-preserving maps

An invertible operator carries each basis to a basis. Does it keep the orientation? The answer is the same for every basis and depends only on the sign of $\det T$.

Proposition 7.9.4. Determinant Sign and Orientation

Let V be a vector space over $\mathbb{R}$ with $\dim V = n \geq 1$, let $T \in \mathcal{L}(V)$ be invertible, and let $\mathcal{B} = (v_1, \ldots, v_n)$ be any basis of V. Then $T\mathcal{B} = (Tv_1, \ldots, Tv_n)$ is a basis of V, and

$$[\mathrm{id}]_{T\mathcal{B}}^{\mathcal{B}} = [T]_{\mathcal{B}}^{\mathcal{B}}.$$

Hence $\mathcal{B}$ and $T\mathcal{B}$ have the same orientation if and only if $\det T > 0$, and this does not depend on $\mathcal{B}$.

Proof. Since T is invertible, it is an isomorphism, so $T\mathcal{B}$ is a basis by Theorem 2.5.2. By Definition 4.2.1, column j of $[\mathrm{id}]_{T\mathcal{B}}^{\mathcal{B}}$ is $[Tv_j]_{\mathcal{B}}$, which is column j of $[T]_{\mathcal{B}}^{\mathcal{B}}$ by Definition 4.1.1. Hence the two matrices are equal, and by Definition 7.5.1, $\det[\mathrm{id}]_{T\mathcal{B}}^{\mathcal{B}} = \det T$. By Defini-

tion 7.9.2 and the symmetry in Theorem 7.9.3 (a), $\mathcal{B}$ and $T\mathcal{B}$ have the same orientation exactly when $\det T > 0$.

■

An invertible operator with $\det T > 0$ is **orientation-preserving**, and one with $\det T < 0$ is **orientation-reversing**. Rotations of the plane preserve orientation and reflections reverse it (Exercise (C2: Rotations and reflections)). In $\mathbb{R}^3$, the map $\mathbf{x} \mapsto -\mathbf{x}$ reverses orientation, since $\det(-\mathbf{I}_3) = -1$: a right hand, pushed through the origin, becomes a left hand.

7.9.7 Exercises

7.9.7.1 A. Check your understanding

Exercise. A1

(a) State the definition of the volume of the parallelepiped spanned by $\mathbf{a}_1, \dots, \mathbf{a}_n \in \mathbb{R}^n$, and say what it is when the vectors are dependent.

(b) True or false: $\operatorname{vol}(\mathbf{a}_1 + \mathbf{b}, \mathbf{a}_2, \dots, \mathbf{a}_n) = \operatorname{vol}(\mathbf{a}_1, \dots, \mathbf{a}_n) + \operatorname{vol}(\mathbf{b}, \mathbf{a}_2, \dots, \mathbf{a}_n)$. Justify your answer.

(c) Which three rules characterize volume in Theorem 7.9.1?

(d) True or false: if $\det T = -2$, then T doubles every volume. Justify your answer.

(e) Define "same orientation" for two ordered bases of a real vector space.

(f) True or false: for every $n \geq 1$, the map $\mathbf{x} \mapsto -\mathbf{x}$ on $\mathbb{R}^n$ reverses orientation. Justify your answer.

Solution.

(a) $\operatorname{vol}(\mathbf{a}_1, \dots, \mathbf{a}_n) = \left|\det \begin{pmatrix} \mathbf{a}_1 & \cdots & \mathbf{a}_n \end{pmatrix}\right|$ (Definition 7.9.1). For dependent vectors it is 0, by Theorem 7.2.1.

(b) False. With $n = 2$, $\mathbf{a}_1 = \mathbf{e}_1$, $\mathbf{b} = -\mathbf{e}_1$, $\mathbf{a}_2 = \mathbf{e}_2$: the left side is $\operatorname{vol}(\mathbf{0}, \mathbf{e}_2) = 0$, the right side is $1 + 1 = 2$. The absolute value destroys additivity.

(c) (V1) scaling one vector by c scales the volume by $|c|$; (V2) adding a multiple of one vector to another does not change it; (V3) the unit cube has volume 1.

(d) True. By Theorem 7.9.2, $\operatorname{vol}(T\mathbf{a}_1, \dots, T\mathbf{a}_n) = |-2| \operatorname{vol}(\mathbf{a}_1, \dots, \mathbf{a}_n)$. The sign only affects orientation.

(e) $\mathcal{B}$ and $\mathcal{C}$ have the same orientation if $\det[\mathrm{id}]_{\mathcal{B}}^{\mathcal{C}} > 0$ (Definition 7.9.2).

(f) False. $\det(-\mathbf{I}_n) = (-1)^n$, so by Proposition 7.9.4 the map reverses orientation for odd n and preserves it for even n. In $\mathbb{R}^2$ it is the rotation by a half-turn.

7.9.7.2 B. Practice

Exercise. B1: Volume of a parallelepiped

Find the volume of the parallelepiped in $\mathbb{R}^3$ spanned by $(2, 1, 0)$, $(1, 3, 1)$, and $(0, 1, 1)$. Hence find the volume of its image under the map $T(x, y, z) = (x + y,\ y + z,\ 2z)$.

Solution. Expanding along the first row,

$$\det \begin{pmatrix} 2 & 1 & 0 \\ 1 & 3 & 1 \\ 0 & 1 & 1 \end{pmatrix} = 2(3-1) - 1 \cdot (1-0) + 0 = 3,$$

so the volume is 3. The standard matrix of T is $\begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 2 \end{pmatrix}$, upper triangular with determinant 2 (Theorem 7.4.6). By Theorem 7.9.2 the image has volume $2 \cdot 3 = 6$.

Exercise. B2: Comparing orientations

In $\mathbb{R}^3$, let

$$\mathcal{B} = ((1,1,0),(0,1,1),(1,0,1)),$$
$$\mathcal{C} = ((0,1,0),(1,0,0),(1,1,1)),$$
$$\mathcal{D} = ((1,0,1),(1,1,0),(0,1,1)).$$

Determine which pairs have the same orientation, and which are positively oriented. Justify your answer.

Solution. By the shortcut after Definition 7.9.2, compare signs of the determinants of the matrices with these columns. Expanding along the first row:

$$\det \begin{pmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix} = 1 \cdot (1-0) - 0 + 1 \cdot (1-0) = 2, \qquad \det \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} = -1,$$

the second by expanding along the last row: $1 \cdot \det \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = -1$. The list $\mathcal{D}$ is $\mathcal{B}$ shifted cyclically, $(\mathbf{b}_3, \mathbf{b}_1, \mathbf{b}_2)$; moving the last column to the front takes two column swaps, so by Theorem 7.2.1 its determinant is $(-1)^2 \cdot 2 = 2$. Hence $\mathcal{B}$ and $\mathcal{D}$ are positively oriented and have the same orientation; $\mathcal{C}$ is negatively oriented, opposite to both.

Exercise. B3: Areas of images

Let $T(x, y) = (2x + y,\ -x + 3y)$ on $\mathbb{R}^2$.

(a) Find the area of the image of the unit square under T.

(b) Find the area of the image of the parallelogram $P((1,0),(1,2))$, first by Theorem 7.9.2 and then directly.

(c) Is T orientation-preserving?

Solution.

(a) The standard matrix $\begin{pmatrix} 2 & 1 \\ -1 & 3 \end{pmatrix}$ has determinant $6 + 1 = 7$, so the image of the unit square has area 7.

(b) $\mathrm{vol}((1,0),(1,2)) = |\det \begin{pmatrix} 1 & 1 \\ 0 & 2 \end{pmatrix}| = 2$, so the image has area $7 \cdot 2 = 14$. Directly, $T(1,0) = (2,-1)$ and $T(1,2) = (4,5)$, and $|\det \begin{pmatrix} 2 & 4 \\ -1 & 5 \end{pmatrix}| = |10 + 4| = 14$.

(c) Yes: $\det T = 7 > 0$, so by Proposition 7.9.4 T preserves orientation.

7.9.7.3 C. Going deeper

Exercise. C1: Area of a triangle in space

For $\mathbf{u}, \mathbf{v} \in \mathbb{R}^3$, let $\mathbf{A} = \begin{pmatrix} \mathbf{u} & \mathbf{v} \end{pmatrix} \in M_{3\times 2}(\mathbb{R})$. We **define** the area of the parallelogram $P(\mathbf{u}, \mathbf{v}) \subseteq \mathbb{R}^3$ to be $\sqrt{\det(\mathbf{A}^\top \mathbf{A})}$; Chapter 13 justifies this with inner products, where $\det(\mathbf{A}^\top \mathbf{A})$ is recognized as a Gram determinant.

(a) Show that $\det(\mathbf{A}^\top \mathbf{A}) \geq 0$, so the square root exists, and that $\det(\mathbf{A}^\top \mathbf{A}) = (u_2 v_3 - u_3 v_2)^2 + (u_3 v_1 - u_1 v_3)^2 + (u_1 v_2 - u_2 v_1)^2$.

(b) Show that if $\mathbf{u}$ and $\mathbf{v}$ lie in the plane $z = 0$, this area equals the area $\mathrm{vol}((u_1, u_2), (v_1, v_2))$ of Definition 7.9.1 in $\mathbb{R}^2$.

(c) Find the area of the triangle with vertices $(1,0,0)$, $(0,2,0)$, $(0,0,3)$.

Hint: apply Corollary 7.7.3 to $\mathbf{A}^\top$.

Solution.

(a) $\mathbf{A}^\top \in M_{2\times 3}(\mathbb{R})$ and $(\mathbf{A}^\top)^\top = \mathbf{A}$, so by Corollary 7.7.3 applied to $\mathbf{A}^\top$, $\det(\mathbf{A}^\top \mathbf{A}) = \sum_{|S|=2} (\det(\mathbf{A}^\top)_{[2],S})^2 \geq 0$. The 2×2 minors of $\mathbf{A}^\top = \begin{pmatrix} u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \end{pmatrix}$ are $u_1 v_2 - u_2 v_1$, $u_1 v_3 - u_3 v_1$, $u_2 v_3 - u_3 v_2$ for $S = \{1,2\}, \{1,3\}, \{2,3\}$; squaring removes the signs, giving the stated sum.

(b) If $u_3 = v_3 = 0$, two of the three squares vanish, and $\det(\mathbf{A}^\top \mathbf{A}) = (u_1 v_2 - u_2 v_1)^2$. Hence $\sqrt{\det(\mathbf{A}^\top \mathbf{A})} = |u_1 v_2 - u_2 v_1| = |\det \begin{pmatrix} u_1 & v_1 \\ u_2 & v_2 \end{pmatrix}|$, which is $\mathrm{vol}((u_1, u_2), (v_1, v_2))$.

(c) The triangle is half of the parallelogram spanned by $\mathbf{u} = (0,2,0) - (1,0,0) = (-1,2,0)$ and $\mathbf{v} = (0,0,3) - (1,0,0) = (-1,0,3)$. By (a), $\det(\mathbf{A}^\top \mathbf{A}) = (6-0)^2 + (0+3)^2 + (0+2)^2 = 36 + 9 + 4 = 49$. As a check, $\mathbf{A}^\top \mathbf{A} = \begin{pmatrix} 5 & 1 \\ 1 & 10 \end{pmatrix}$ has determinant $50 - 1 = 49$. So the parallelogram has area 7, and the triangle has area $\frac{7}{2}$.

Exercise. C2: Rotations and reflections

For $\theta \in \mathbb{R}$, let $\mathbf{R}_\theta = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$ (rotation by θ) and $\mathbf{S}_\theta = \begin{pmatrix} \cos 2\theta & \sin 2\theta \\ \sin 2\theta & -\cos 2\theta \end{pmatrix}$ (reflection

in the line through 0 at angle θ).

(a) Show that every rotation preserves orientation and area, and every reflection reverses orientation and preserves area.

(b) Show that $\mathbf{S}_\theta \mathbf{S}_\varphi$ is orientation-preserving for all θ, φ. (It is in fact a rotation, but you need not show this.)

(c) Prove that no $\mathbf{M} \in M_2(\mathbb{R})$ satisfies $\mathbf{M}^2 = \mathbf{S}_0 = \operatorname{diag}(1, -1)$. Does the same argument rule out $\mathbf{M}^2 = -\mathbf{I}_2$?

Solution.

(a) $\det \mathbf{R}_\theta = \cos^2\theta + \sin^2\theta = 1 > 0$ and $\det \mathbf{S}_\theta = -\cos^2 2\theta - \sin^2 2\theta = -1 < 0$. By Proposition 7.9.4, rotations preserve and reflections reverse orientation. Both have $|\det| = 1$, so both preserve area by Theorem 7.9.2.

(b) By Theorem 7.5.4, $\det(\mathbf{S}_\theta \mathbf{S}_\varphi) = (-1)(-1) = 1 > 0$.

(c) If $\mathbf{M}^2 = \mathbf{S}_0$, then $(\det \mathbf{M})^2 = \det \mathbf{S}_0 = -1$ by Theorem 7.5.4, impossible for a real number $\det \mathbf{M}$. The argument says nothing about $-\mathbf{I}_2$, since $\det(-\mathbf{I}_2) = 1$; and indeed $\mathbf{R}_{\pi/2}^2 = \mathbf{R}_{\pi/2}\mathbf{R}_{\pi/2} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}^2 = -\mathbf{I}_2$. An orientation-reversing map can never be the square of a real map, since the square of anything preserves orientation.

7.10 The Characteristic Polynomial

For a square matrix $\mathbf{A}$ over F and a scalar $c \in F$, when is $c\mathbf{I} - \mathbf{A}$ invertible? The question matters: $c\mathbf{I} - \mathbf{A}$ fails to be invertible exactly when $\mathbf{A}\mathbf{v} = c\mathbf{v}$ for some nonzero vector $\mathbf{v}$, the situation Chapter 9 is built around. The determinant answers the question for each single c. This section lets c vary. Replacing the number c by a variable x turns $\det(c\mathbf{I} - \mathbf{A})$ into a polynomial, the characteristic polynomial, and the scalars we are looking for become its roots. The price is that $x\mathbf{I} - \mathbf{A}$ is a matrix of **polynomials**, not of field elements, so we must check which determinant facts survive there. The remarks at the ends of §4, §5, and §6 did exactly that bookkeeping, and here we cash them in.

7.10.1 From a number to a variable

Take $\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \in M_2(\mathbb{Q})$. For each $c \in \mathbb{Q}$,

$$\det(c\mathbf{I} - \mathbf{A}) = \det \begin{pmatrix} c-1 & -2 \\ -3 & c-4 \end{pmatrix} = (c-1)(c-4) - 6 = c^2 - 5c - 2.$$

The same expression $c^2 - 5c - 2$ appears whatever c is. It is the value at c of one polynomial, and that polynomial deserves a name. To define it directly, without first choosing c, we put the variable x into the matrix and take the determinant of

$$x\mathbf{I} - \mathbf{A} = \begin{pmatrix} x-1 & -2 \\ -3 & x-4 \end{pmatrix},$$

a matrix whose entries lie in $F[x]$.

Is that allowed? $F[x]$ is not a field, but it is a **commutative ring** in the sense of the remark in §4: all the field axioms hold except the existence of inverses (Theorem 0.8.1). The Leibniz formula of Definition 7.4.1 uses only sums, differences and products of entries, so it defines $\det \mathbf{M} \in F[x]$ for every $\mathbf{M} \in M_n(F[x])$.

The characteristic polynomial of $\mathbf{A}$ *is the determinant of* $x\mathbf{I} - \mathbf{A}$*, computed with polynomial entries.*

Definition 7.10.1. Characteristic Polynomial

Let $n \geq 1$ and $\mathbf{A} = (a_{ij}) \in M_n(F)$. The **characteristic polynomial** of $\mathbf{A}$ is

$$p_{\mathbf{A}}(x) := \det(x\mathbf{I}_n - \mathbf{A}) = \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma) \prod_{j=1}^{n} \left(x\,\delta_{\sigma(j)j} - a_{\sigma(j)j}\right) \in F[x],$$

where $\delta_{ij} = 1$ if $i = j$ and 0 otherwise, so that $x\delta_{ij} - a_{ij}$ is the (i,j)-entry of $x\mathbf{I}_n - \mathbf{A}$. The determinant is the Leibniz sum of Definition 7.4.1, computed in the commutative ring $F[x]$.

In words: subtract $\mathbf{A}$ from x times the identity, so that x appears only on the diagonal; then take the determinant exactly as for a matrix of numbers, multiplying and adding polynomials instead of scalars. The result is **one** polynomial, not a number.

Well-definedness. Each term of the Leibniz sum is a product of n polynomials, and the sum is finite, so $p_{\mathbf{A}}$ is a polynomial. It depends only on $\mathbf{A}$.

How to compute it. By the remarks in §4, §5, and §6, the following results were proved without dividing, and so hold for matrices over $F[x]$: Theorem 7.4.5 and Theorem 7.4.6 (§4), Theorem 7.5.4 by its first proof (§5), Theorem 7.6.1 (§6), and Theorem 7.8.4 (§8). So we may expand $\det(x\mathbf{I} - \mathbf{A})$ along any row or column, and read it off from triangular or block triangular shape. What we may **not** do is row reduce $x\mathbf{I} - \mathbf{A}$ by dividing by a polynomial entry, as the warning below explains.

Examples.

- 1×1. For $\mathbf{A} = (a)$, $p_{\mathbf{A}}(x) = x - a$.
- 2×2. For $\mathbf{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$,

$$\begin{aligned} p_{\mathbf{A}}(x) &= (x-a)(x-d) - bc = x^2 - (a+d)x + (ad - bc) \\ &= x^2 - (\operatorname{tr}\mathbf{A})\,x + \det\mathbf{A}. \end{aligned}$$

  For $\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$ this is $x^2 - 5x - 2$, as above.
- **Triangular.** If $\mathbf{A}$ is upper or lower triangular, so is $x\mathbf{I} - \mathbf{A}$, with diagonal entries $x - a_{ii}$. By Theorem 7.4.6 over $F[x]$, $p_{\mathbf{A}}(x) = (x - a_{11})(x - a_{22}) \cdots (x - a_{nn})$.
- **Degenerate cases.** The zero matrix has $p_{\mathbf{O}}(x) = x^n$, and the identity has $p_{\mathbf{I}_n}(x) = (x-1)^n$; both are triangular. These two polynomials are different, although both matrices are as simple as possible, which already shows that $p_{\mathbf{A}}$ records something about $\mathbf{A}$.

Example. A 3×3 Characteristic Polynomial

Find $p_{\mathbf{A}}$ for $\mathbf{A} = \begin{pmatrix} 2 & 0 & 1 \\ 1 & 3 & 1 \\ 1 & 0 & 2 \end{pmatrix} \in M_3(\mathbb{Q})$.

Solution. We have

$$x\mathbf{I} - \mathbf{A} = \begin{pmatrix} x-2 & 0 & -1 \\ -1 & x-3 & -1 \\ -1 & 0 & x-2 \end{pmatrix}.$$

Column 2 has a single nonzero entry, $x-3$ in position $(2,2)$. Expanding along it (Theorem 7.6.1 over $F[x]$),

$$\begin{aligned} p_{\mathbf{A}}(x) &= (x-3)\,(-1)^{2+2} \det \begin{pmatrix} x-2 & -1 \\ -1 & x-2 \end{pmatrix} \\ &= (x-3)((x-2)^2 - 1) = (x-1)(x-3)^2. \end{aligned}$$

Multiplied out, $p_{\mathbf{A}}(x) = x^3 - 7x^2 + 15x - 9$. Compare with $\operatorname{tr}\mathbf{A} = 2+3+2 = 7$ and $\det\mathbf{A} = 9$ (expand $\mathbf{A}$ along column 2: $3(4-1) = 9$).

A companion-type example. Let $\mathbf{A} = \begin{pmatrix} 0 & 0 & 6 \\ 1 & 0 & -11 \\ 0 & 1 & 6 \end{pmatrix}$, with ones just below the diagonal, a last column $(6, -11, 6)$, and zeros elsewhere. Expanding $x\mathbf{I} - \mathbf{A} = \begin{pmatrix} x & 0 & -6 \\ -1 & x & 11 \\ 0 & -1 & x-6 \end{pmatrix}$ along the first row (Theorem 7.6.1 over $F[x]$),

$$\begin{aligned} p_{\mathbf{A}}(x) &= x \det \begin{pmatrix} x & 11 \\ -1 & x-6 \end{pmatrix} + (-6) \det \begin{pmatrix} -1 & x \\ 0 & -1 \end{pmatrix} \\ &= x(x^2 - 6x + 11) - 6 = x^3 - 6x^2 + 11x - 6, \end{aligned}$$

which factors as $(x-1)(x-2)(x-3)$. The last column of $\mathbf{A}$ lists the coefficients $-6, 11, -6$ of $1, x, x^2$ in $p_{\mathbf{A}}$, with their signs changed. This is no accident: Exercise (C1: Companion matrices) shows that the same pattern works in every size.

Nonexample by minimal change. Swap the order of subtraction: $\det(\mathbf{A} - x\mathbf{I})$. Each of the n columns of $\mathbf{A} - x\mathbf{I}$ is -1 times the corresponding column of $x\mathbf{I} - \mathbf{A}$, so by multilinearity $\det(\mathbf{A} - x\mathbf{I}) = (-1)^n p_{\mathbf{A}}(x)$. For $\mathbf{A} = (a)$ this is $a - x$, whose leading coefficient is -1. What still works: the roots are the same. What fails: the result is **not monic** when n is odd.

Why this definition. Some books use $\det(\mathbf{A} - x\mathbf{I})$. We use $x\mathbf{I} - \mathbf{A}$ so that the characteristic polynomial is always **monic** (Theorem 7.10.2 below), which makes statements like "$p_{\mathbf{A}} = (x-1)(x-3)^2$" true without sign bookkeeping. The name comes from the fact that the roots of this polynomial are the "characteristic values" of $\mathbf{A}$, which Chapter 9 calls eigenvalues.

Warning. **Do not row reduce $x\mathbf{I} - \mathbf{A}$ by dividing by polynomials.** Over a field, a matrix with nonzero determinant is invertible (Theorem 7.5.2), but that proof runs through elimination, which divides by pivots, and it fails over $F[x]$. For $\mathbf{A} = \mathbf{O}_1 \in M_1(F)$, the matrix $x\mathbf{I} - \mathbf{A} = (x)$ has determinant $x \neq 0$, yet it has no inverse in $M_1(F[x])$: if $x\,q(x) = 1$ for a polynomial q, then $q \neq 0$, and $1 + \deg q = \deg 1 = 0$ by Theorem 0.8.3, which is impossible

since $\deg q \geq 0$. Likewise, "divide row 1 of $x\mathbf{I} - \mathbf{A}$ by $x - 1$" produces entries that are not polynomials and that make no sense at $x = 1$. Compute $p_{\mathbf{A}}$ by expansion, by triangular or block shape, or by the Leibniz formula.

Before looking at the coefficients, we record the fact that makes $p_{\mathbf{A}}$ useful: evaluating it at a scalar c gives the same result as putting c in place of x before taking the determinant.

Lemma 7.10.1. Evaluating the Characteristic Polynomial

Let $\mathbf{A} \in M_n(F)$ and $c \in F$. Then $p_{\mathbf{A}}(c) = \det(c\mathbf{I}_n - \mathbf{A})$.

Proof. By Theorem 0.8.5, evaluation at c turns sums of polynomials into sums and products into products, and sends a constant a to a. Applying it to the Leibniz sum in Definition 7.10.1 term by term,

$$p_{\mathbf{A}}(c) = \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma) \prod_{j=1}^{n} (c\,\delta_{\sigma(j)j} - a_{\sigma(j)j}),$$

and $c\,\delta_{ij} - a_{ij}$ is the (i, j)-entry of $c\mathbf{I}_n - \mathbf{A}$. By Definition 7.4.1, the right side is $\det(c\mathbf{I}_n - \mathbf{A})$. ∎

7.10.2 The shape of the characteristic polynomial

The 2×2 example suggests a pattern: the leading coefficient is 1, the next coefficient is minus the trace, and the constant term is the determinant up to sign. The same holds in every size.

Theorem 7.10.2. Coefficients of the Characteristic Polynomial

Let $n \geq 1$ and $\mathbf{A} \in M_n(F)$. Then $p_{\mathbf{A}}$ is **monic of degree** n, and

$$p_{\mathbf{A}}(x) = x^n - (\operatorname{tr} \mathbf{A})\, x^{n-1} + \cdots + (-1)^n \det \mathbf{A},$$

that is, the coefficient of x^{n-1} is $-\operatorname{tr} \mathbf{A}$ and the constant term is $(-1)^n \det \mathbf{A}$. (For $n = 1$ these two statements concern the same coefficient, and both say it is $-a_{11}$.)

Idea. Look at the Leibniz sum term by term and count powers of x. A factor $x\delta_{\sigma(j)j} - a_{\sigma(j)j}$ contains x only if $\sigma(j) = j$. A permutation other than the identity moves at least **two** indices, so its term has degree at most $n - 2$ and cannot affect the top two coefficients. The identity contributes $\prod_j (x - a_{jj})$, whose top two coefficients are 1 and $-\sum_j a_{jj}$. For the constant term, set $x = 0$.

Proof. Write $m_{ij} = x\delta_{ij} - a_{ij}$ for the entries of $x\mathbf{I} - \mathbf{A}$. By Definition 7.10.1, $p_{\mathbf{A}} = \sum_\sigma t_\sigma$ with $t_\sigma = \operatorname{sgn}(\sigma) \prod_j m_{\sigma(j)j}$.

Terms with $\sigma \neq$ id. If $\sigma(j) \neq j$ for some j, then $\sigma(\sigma(j)) \neq \sigma(j)$, because σ is injective. So σ moves at least two indices, and at most $n - 2$ indices are fixed. For a moved index j, $m_{\sigma(j)j} = -a_{\sigma(j)j}$ has degree at most 0; for a fixed index it has degree at most 1. By

Theorem 0.8.3, $\deg t_\sigma \le n-2$.

The term $\sigma = \mathrm{id}$. We show by induction on $k \ge 1$ that $\prod_{j=1}^{k}(x-a_{jj}) = x^k - s_k x^{k-1} + r_k$ with $s_k = a_{11}+\cdots+a_{kk}$ and $\deg r_k \le k-2$. For $k=1$, take $r_1 = 0$. If it holds for k, then with $a = a_{k+1,k+1}$,

$$(x^k - s_k x^{k-1} + r_k)(x-a) = x^{k+1} - (s_k + a)x^k + (s_k a x^{k-1} + r_k x - a r_k),$$

and the bracket has degree at most $k-1$ by Theorem 0.8.2 and Theorem 0.8.3. This completes the induction. Since $\mathrm{sgn}(\mathrm{id}) = 1$, $t_{\mathrm{id}} = x^n - (\operatorname{tr}\mathbf{A})\,x^{n-1} + r_n$ with $\deg r_n \le n-2$.

Adding, $p_{\mathbf{A}} = x^n - (\operatorname{tr}\mathbf{A})\,x^{n-1} + \left(r_n + \sum_{\sigma\neq\mathrm{id}} t_\sigma\right)$, and the bracket has degree at most $n-2$ by Theorem 0.8.2. Hence $p_{\mathbf{A}}$ is monic of degree n, with coefficient $-\operatorname{tr}\mathbf{A}$ at x^{n-1} (Definition 0.9.12).

Constant term. The constant term of a polynomial is its value at 0 (Definition 0.8.6). By Lemma 7.10.1 with $c = 0$, $p_{\mathbf{A}}(0) = \det(-\mathbf{A})$. The matrix $-\mathbf{A}$ is obtained from $\mathbf{A}$ by multiplying each of its n columns by -1, so $\det(-\mathbf{A}) = (-1)^n \det\mathbf{A}$ by Theorem 7.5.1 (d). This proves the theorem.

■

So the characteristic polynomial packages two invariants we already know, the trace and the determinant, and adds $n-2$ more coefficients in between. For $n = 2$ nothing is in between, which is why $p_{\mathbf{A}}(x) = x^2 - (\operatorname{tr}\mathbf{A})x + \det\mathbf{A}$ for every 2×2 matrix.

Quick check. Without expanding any determinant, find the coefficient of x^2 and the constant term of $p_{\mathbf{A}}$ for $\mathbf{A} = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}$, given that $\det\mathbf{A} = 2$.

Solution. By Theorem 7.10.2, the coefficient of x^2 is $-\operatorname{tr}\mathbf{A} = 0$, and the constant term is $(-1)^3 \det\mathbf{A} = -2$. (In full, $p_{\mathbf{A}}(x) = x^3 - 3x - 2 = (x-2)(x+1)^2$.)

7.10.3 Similar matrices, same polynomial

Similar matrices describe the same operator in different bases (Theorem 4.2.4), so everything intrinsic to the operator should agree for them. The trace and determinant do (Theorem 4.2.5, Corollary 7.5.7). The characteristic polynomial, which contains both, does too, and the proof is the proof for the determinant, run over $F[x]$.

Theorem 7.10.3. Similar Matrices Have the Same Characteristic Polynomial

If $\mathbf{A}, \mathbf{B} \in M_n(F)$ are similar, then $p_{\mathbf{A}} = p_{\mathbf{B}}$.

Proof. Let $\mathbf{B} = \mathbf{P}^{-1}\mathbf{A}\mathbf{P}$ with $\mathbf{P} \in M_n(F)$ invertible. Regard $\mathbf{P}$, $\mathbf{P}^{-1}$, $\mathbf{A}$, and $\mathbf{I}$ as matrices over $F[x]$ whose entries happen to be constants. The laws of matrix arithmetic in

Theorem 0.9.4 are proved from the ring axioms alone, so they hold in $M_n(F[x])$. Hence

$$\mathbf{P}^{-1}(x\mathbf{I}-\mathbf{A})\mathbf{P} = x\,\mathbf{P}^{-1}\mathbf{P} - \mathbf{P}^{-1}\mathbf{A}\mathbf{P} = x\mathbf{I}-\mathbf{B}.$$

By Theorem 7.5.4, valid over $F[x]$ by the remark in §5,

$$\begin{aligned} p_{\mathbf{B}} &= \det\left(\mathbf{P}^{-1}(x\mathbf{I}-\mathbf{A})\mathbf{P}\right) = \det(\mathbf{P}^{-1})\,\det(x\mathbf{I}-\mathbf{A})\,\det\mathbf{P} \\ &= \det(\mathbf{P}^{-1})\det(\mathbf{P})\,p_{\mathbf{A}}, \end{aligned}$$

using commutativity of $F[x]$ in the last step. Finally $\det(\mathbf{P}^{-1})\det\mathbf{P} = \det(\mathbf{P}^{-1}\mathbf{P}) = \det\mathbf{I} = 1$ by Theorem 7.5.4 over F. Hence $p_{\mathbf{B}} = p_{\mathbf{A}}$.

■

This is exactly the case singled out in the remark in §5: the proof of Corollary 7.5.7 survives over $F[x]$ because $\mathbf{P}$ is invertible already over F. The theorem lets us attach a characteristic polynomial to an operator.

Definition 7.10.2. Characteristic Polynomial of an Operator

Let V be a vector space over F with $\dim V = n \geq 1$, and let $T \in \mathcal{L}(V)$. The **characteristic polynomial** of T is

$$p_T(x) \coloneqq p_{[T]_{\mathcal{B}}^{\mathcal{B}}}(x) = \det\left(x\mathbf{I}_n - [T]_{\mathcal{B}}^{\mathcal{B}}\right),$$

where $\mathcal{B}$ is **any** basis of V.

This does not depend on $\mathcal{B}$: for bases $\mathcal{B}$ and $\mathcal{C}$, the matrices $[T]_{\mathcal{B}}^{\mathcal{B}}$ and $[T]_{\mathcal{C}}^{\mathcal{C}}$ are similar by Theorem 4.2.4 (a), so they have the same characteristic polynomial by Theorem 7.10.3. By Theorem 7.10.2, p_T is monic of degree $\dim V$, with coefficient $-\operatorname{tr} T$ at x^{n-1} and constant term $(-1)^n \det T$ (Definition 4.5.2, Definition 7.5.1).

Example. Characteristic Polynomials of Two Operators

On $V = \mathbb{R}[x]_{\leq 2}$, find the characteristic polynomials of differentiation $D(q) = q'$ and of the shift $S(q)(x) = q(x+1)$.

Solution. Use the basis $\mathcal{B} = (1, x, x^2)$. Since $D(1) = 0$, $D(x) = 1$, $D(x^2) = 2x$, and $S(1) = 1$, $S(x) = 1 + x$, $S(x^2) = 1 + 2x + x^2$,

$$[D]_{\mathcal{B}}^{\mathcal{B}} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \end{pmatrix}, \qquad [S]_{\mathcal{B}}^{\mathcal{B}} = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{pmatrix}.$$

Both are upper triangular, so $p_D(t) = t^3$ and $p_S(t) = (t-1)^3$. (We name the variable t here, since x is already the variable of the polynomials in V.) As a check, $\operatorname{tr} S = 3$ and $\det S = 1$, matching the coefficient -3 of t^2 and the constant term $(-1)^3 \cdot 1$.

Warning. **Equal characteristic polynomials do not make matrices similar.** The converse of Theorem 7.10.3 is false. Recall the pair from the warning in Chapter 4, §02:

$\mathbf{I}_2$ and $\mathbf{J} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ both have characteristic polynomial $(x-1)^2$, since both are triangular with diagonal $1, 1$. But $\mathbf{I}_2$ is similar only to itself, so $\mathbf{J} \not\sim \mathbf{I}_2$. In particular the characteristic polynomial does not determine the matrix, not even up to similarity. Deciding similarity needs finer invariants; Chapter 10 finds a complete set.

7.10.4 Roots and singular matrices

We can now answer the opening question. The determinant tests invertibility of $c\mathbf{I} - \mathbf{A}$ for one c at a time; the characteristic polynomial tests all c at once, and the answer is a finite set.

Theorem 7.10.4. Roots of the Characteristic Polynomial

Let $\mathbf{A} \in M_n(F)$ with $n \geq 1$, and let $c \in F$.

(a) The following are equivalent: (i) $p_{\mathbf{A}}(c) = 0$; (ii) $c\mathbf{I} - \mathbf{A}$ is **not** invertible; (iii) there is a **nonzero** $\mathbf{v} \in F^n$ with $\mathbf{A}\mathbf{v} = c\mathbf{v}$.

(b) There are **at most** n scalars $c \in F$ with these properties.

Proof.

(a) By Lemma 7.10.1, $p_{\mathbf{A}}(c) = \det(c\mathbf{I} - \mathbf{A})$, which is a determinant of a matrix over the **field** F. By Theorem 7.5.2, it is zero if and only if $c\mathbf{I} - \mathbf{A}$ is not invertible; this is (i) $\Leftrightarrow$ (ii). By Theorem 3.4.4, $c\mathbf{I} - \mathbf{A}$ is not invertible if and only if $(c\mathbf{I} - \mathbf{A})\mathbf{v} = \mathbf{0}$ for some $\mathbf{v} \neq \mathbf{0}$, and $(c\mathbf{I} - \mathbf{A})\mathbf{v} = c\mathbf{v} - \mathbf{A}\mathbf{v}$ by Theorem 0.9.4; this is (ii) $\Leftrightarrow$ (iii).

(b) By Theorem 7.10.2, $p_{\mathbf{A}}$ is monic of degree n, in particular nonzero. By Corollary 6.1.4, it has at most n distinct roots in F, and by (a) these are exactly the scalars in question.

■

Notice where the evaluation happens: **first** substitute c, getting a matrix over the field F, **then** use the invertibility test, which is valid over fields. That order is what the warning about $F[x]$ above requires.

The same holds for an operator T on V with $\dim V = n \geq 1$: $p_T(c) = 0$ if and only if $c\,\mathrm{id}_V - T$ is not invertible, by (a) applied to $[T]_{\mathcal{B}}^{\mathcal{B}}$ together with Theorem 4.1.5 (b), since $[c\,\mathrm{id}_V - T]_{\mathcal{B}}^{\mathcal{B}} = c\mathbf{I} - [T]_{\mathcal{B}}^{\mathcal{B}}$.

The scalars c of this theorem are the **eigenvalues** of $\mathbf{A}$, and the vectors $\mathbf{v}$ in (iii) its eigenvectors. They are the subject of Chapter 9, which starts from this theorem.

Examples.

- For $\mathbf{A}$ in Example (A 3×3 Characteristic Polynomial), $p_{\mathbf{A}} = (x-1)(x-3)^2$, so $c\mathbf{I} - \mathbf{A}$ is singular exactly for $c = 1$ and $c = 3$. For instance $3\mathbf{I} - \mathbf{A} = \begin{pmatrix} 1 & 0 & -1 \\ -1 & 0 & -1 \\ -1 & 0 & 1 \end{pmatrix}$ has a zero column, and $\mathbf{v} = \mathbf{e}_2$ satisfies $\mathbf{A}\mathbf{e}_2 = 3\mathbf{e}_2$.
- The rotation $\mathbf{R} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ has $p_{\mathbf{R}} = x^2 + 1$. Over $\mathbb{R}$ this has no root, so $c\mathbf{I} - \mathbf{R}$ is

invertible for **every** real c: no nonzero real vector is merely stretched by a quarter-turn. Over $\mathbb{C}$, the roots $\pm i$ appear, and for example $\mathbf{R}(1,-i) = (i,1) = i\,(1,-i)$. The field matters.

Quick check. For which $c \in \mathbb{Q}$ is $c\mathbf{I} - \mathbf{A}$ invertible, where $\mathbf{A} = \begin{pmatrix} 5 & -2 \\ 3 & 0 \end{pmatrix}$?

Solution. $p_{\mathbf{A}}(x) = x^2 - (\operatorname{tr}\mathbf{A})x + \det\mathbf{A} = x^2 - 5x + 6 = (x-2)(x-3)$. By Theorem 7.10.4, $c\mathbf{I} - \mathbf{A}$ is invertible exactly for $c \neq 2, 3$.

7.10.5 A tempting false proof

Here is a statement that is true, and important: **every square matrix satisfies its own characteristic polynomial**, $p_{\mathbf{A}}(\mathbf{A}) = \mathbf{O}$. This is the Cayley–Hamilton Theorem, proved in Chapter 9. Here $p_{\mathbf{A}}(\mathbf{A})$ means substituting the matrix $\mathbf{A}$ into the polynomial $p_{\mathbf{A}}$, replacing each x^k by $\mathbf{A}^k$ and the constant term a_0 by $a_0\mathbf{I}$ (Definition 0.9.14). For a 2×2 matrix it says $\mathbf{A}^2 - (\operatorname{tr}\mathbf{A})\mathbf{A} + (\det\mathbf{A})\mathbf{I} = \mathbf{O}$, which you can check by multiplying out.

A one-line "proof" suggests itself, and it is wrong.

Warning. **The fake proof of Cayley–Hamilton.** "Since $p_{\mathbf{A}}(x) = \det(x\mathbf{I} - \mathbf{A})$, put $x = \mathbf{A}$: $p_{\mathbf{A}}(\mathbf{A}) = \det(\mathbf{A}\mathbf{I} - \mathbf{A}) = \det(\mathbf{O}) = 0$." This is invalid for two reasons. First, the two sides are different kinds of object: $p_{\mathbf{A}}(\mathbf{A})$ is an $n \times n$ **matrix**, while $\det(\cdots)$ is a **scalar**, so "$= 0$" compares a matrix with a number. Second, Lemma 7.10.1 lets us substitute only a **scalar** $c \in F$ into the entries of $x\mathbf{I} - \mathbf{A}$. Substituting a matrix for x inside the entries would produce a matrix whose entries are matrices, and "$\mathbf{A}\mathbf{I} - \mathbf{A}$" silently treats the x in the entry $x - a_{11}$ as if the entry became $\mathbf{A} - a_{11}$, which is not what substitution into a polynomial means.

The same "argument" proves false statements, which shows it is not a proof. Replace det by tr: let $q(x) = \operatorname{tr}(x\mathbf{I} - \mathbf{A}) = nx - \operatorname{tr}\mathbf{A}$. The fake reasoning gives "$q(\mathbf{A}) = \operatorname{tr}(\mathbf{A}\mathbf{I} - \mathbf{A}) = \operatorname{tr}\mathbf{O} = 0$". But for $\mathbf{A} = \operatorname{diag}(1,0)$, $q(x) = 2x - 1$ and $q(\mathbf{A}) = 2\mathbf{A} - \mathbf{I} = \operatorname{diag}(1,-1) \neq \mathbf{O}$.

7.10.6 Exercises

7.10.6.1 A. Check your understanding

Exercise. A1

(a) Define the characteristic polynomial of $\mathbf{A} \in M_n(F)$, and say in which ring its determinant is computed.

(b) State the degree, leading coefficient, coefficient of x^{n-1}, and constant term of $p_{\mathbf{A}}$.

(c) True or false: $\det(\mathbf{A} - x\mathbf{I}) = p_{\mathbf{A}}(x)$ for every $\mathbf{A} \in M_3(F)$. Justify your answer.

(d) True or false: if $p_{\mathbf{A}} = p_{\mathbf{B}}$, then $\mathbf{A} \sim \mathbf{B}$. Justify your answer.

(e) What does $p_{\mathbf{A}}(c) = 0$ say about the matrix $c\mathbf{I} - \mathbf{A}$? About vectors?

(f) Explain in one sentence why "$p_{\mathbf{A}}(\mathbf{A}) = \det(\mathbf{A}\mathbf{I} - \mathbf{A}) = 0$" is not a proof.

Solution.

(a) $p_{\mathbf{A}}(x) = \det(x\mathbf{I}_n - \mathbf{A})$, the Leibniz sum computed in the commutative ring $F[x]$ (Definition 7.10.1).

(b) Degree n, leading coefficient 1, coefficient $-\operatorname{tr}\mathbf{A}$ at x^{n-1}, constant term $(-1)^n \det\mathbf{A}$ (Theorem 7.10.2).

(c) False. $\det(\mathbf{A} - x\mathbf{I}) = (-1)^3 p_{\mathbf{A}}(x) = -p_{\mathbf{A}}(x)$, since each of the three columns changes sign. For $\mathbf{A} = \mathbf{O}$, $\det(-x\mathbf{I}_3) = -x^3 \neq x^3 = p_{\mathbf{A}}(x)$.

(d) False. $\mathbf{I}_2$ and $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ both have $p = (x-1)^2$ but are not similar.

(e) $c\mathbf{I} - \mathbf{A}$ is not invertible; equivalently there is a nonzero $\mathbf{v}$ with $\mathbf{A}\mathbf{v} = c\mathbf{v}$ (Theorem 7.10.4).

(f) $p_{\mathbf{A}}(\mathbf{A})$ is a matrix obtained by substituting $\mathbf{A}$ into a polynomial, while $\det(\mathbf{A}\mathbf{I} - \mathbf{A})$ is a scalar, and substitution into the entries of $x\mathbf{I} - \mathbf{A}$ is only justified for scalars (Lemma 7.10.1).

7.10.6.2 B. Practice

Exercise. B1: Computing characteristic polynomials

Find the characteristic polynomial of each matrix over $\mathbb{Q}$, factor it as far as possible over $\mathbb{Q}$, and check the trace and determinant against Theorem 7.10.2.

$$\mathbf{A} = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{pmatrix}, \qquad \mathbf{B} = \begin{pmatrix} 2 & 1 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 1 & 3 \\ 0 & 0 & 1 & -1 \end{pmatrix}.$$

Hence list the $c \in \mathbb{Q}$ for which $c\mathbf{I} - \mathbf{A}$, respectively $c\mathbf{I} - \mathbf{B}$, is not invertible.

Solution. For $\mathbf{A}$. Expanding $x\mathbf{I} - \mathbf{A} = \begin{pmatrix} x-1 & -1 & 0 \\ 0 & x-1 & -1 \\ -1 & 0 & x-1 \end{pmatrix}$ along the first row (Theorem 7.6.1 over $F[x]$),

$$\begin{aligned} p_{\mathbf{A}}(x) &= (x-1)\big((x-1)^2 - 0\big) - (-1)(0-1) + 0 \\ &= (x-1)^3 - 1 = x^3 - 3x^2 + 3x - 2. \end{aligned}$$

Check: $-\operatorname{tr}\mathbf{A} = -3$; $\det\mathbf{A} = 1\cdot(1-0) - 1\cdot(0-1) + 0 = 2$, and $(-1)^3 \cdot 2 = -2$. Since $p_{\mathbf{A}}(2) = 1 - 1 = 0$, divide by $x-2$: $p_{\mathbf{A}} = (x-2)(x^2 - x + 1)$. The quadratic has discriminant $1 - 4 = -3 < 0$, so no rational (indeed no real) root. By Theorem 7.10.4, $c\mathbf{I} - \mathbf{A}$ is singular only for $c = 2$.

For $\mathbf{B}$. $x\mathbf{I}-\mathbf{B}$ is block upper triangular with diagonal blocks $x\mathbf{I}-\begin{pmatrix}2&1\\0&2\end{pmatrix}$ and $x\mathbf{I}-\begin{pmatrix}1&3\\1&-1\end{pmatrix}$. By Theorem 7.8.4 over $F[x]$ and the 2×2 formula,

$$p_{\mathbf{B}}(x)=(x-2)^2\left(x^2-0\cdot x+(-1-3)\right)=(x-2)^2(x^2-4)=(x-2)^3(x+2).$$

Multiplied out, $p_{\mathbf{B}}=x^4-4x^3+16x-16$. Check: $-\operatorname{tr}\mathbf{B}=-(2+2+1-1)=-4$; $\det\mathbf{B}=4\cdot(-4)=-16$ by Theorem 7.8.4, and $(-1)^4(-16)=-16$. So $c\mathbf{I}-\mathbf{B}$ is singular exactly for $c=2$ and $c=-2$.

Exercise. B2: Transpose

Prove that $p_{\mathbf{A}^\top}=p_{\mathbf{A}}$ for every $\mathbf{A}\in M_n(F)$. Hence show that $c\mathbf{I}-\mathbf{A}$ is invertible if and only if $c\mathbf{I}-\mathbf{A}^\top$ is.

Solution. By Theorem 0.9.5, $(x\mathbf{I}-\mathbf{A})^\top=x\mathbf{I}^\top-\mathbf{A}^\top=x\mathbf{I}-\mathbf{A}^\top$; these rules are proved entrywise from the ring axioms, so they hold over $F[x]$. By Theorem 7.4.5, which holds over $F[x]$ by the remark in §4,

$$p_{\mathbf{A}^\top}=\det(x\mathbf{I}-\mathbf{A}^\top)=\det\left((x\mathbf{I}-\mathbf{A})^\top\right)=\det(x\mathbf{I}-\mathbf{A})=p_{\mathbf{A}}.$$

By Theorem 7.10.4, $c\mathbf{I}-\mathbf{A}$ is not invertible exactly when $p_{\mathbf{A}}(c)=0$, which is exactly when $p_{\mathbf{A}^\top}(c)=0$, that is, when $c\mathbf{I}-\mathbf{A}^\top$ is not invertible.

Exercise. B3: AB and BA

Let $\mathbf{A},\mathbf{B}\in M_n(F)$ with $\mathbf{A}$ invertible.

(a) Show that $\mathbf{BA}=\mathbf{A}^{-1}(\mathbf{AB})\mathbf{A}$, and deduce $p_{\mathbf{AB}}=p_{\mathbf{BA}}$.

(b) Deduce that $\operatorname{tr}(\mathbf{AB})=\operatorname{tr}(\mathbf{BA})$ and $\det(\mathbf{AB})=\det(\mathbf{BA})$ in this case.

(c) Check (a) for $\mathbf{A}=\begin{pmatrix}1&1\\0&1\end{pmatrix}$, $\mathbf{B}=\begin{pmatrix}0&0\\1&0\end{pmatrix}$, and note that $\mathbf{AB}\neq\mathbf{BA}$.

(Chapter 9 removes the hypothesis that $\mathbf{A}$ is invertible.)

Solution.

(a) By associativity, $\mathbf{A}^{-1}(\mathbf{AB})\mathbf{A}=(\mathbf{A}^{-1}\mathbf{A})(\mathbf{BA})=\mathbf{BA}$. So $\mathbf{BA}\sim\mathbf{AB}$ (Definition 4.2.2, with $\mathbf{P}=\mathbf{A}$), and $p_{\mathbf{AB}}=p_{\mathbf{BA}}$ by Theorem 7.10.3.

(b) By Theorem 7.10.2, the coefficient of x^{n-1} is $-\operatorname{tr}(\mathbf{AB})$ in $p_{\mathbf{AB}}$ and $-\operatorname{tr}(\mathbf{BA})$ in $p_{\mathbf{BA}}$; the constant terms are $(-1)^n\det(\mathbf{AB})$ and $(-1)^n\det(\mathbf{BA})$. Equal polynomials have equal coefficients.

(c) $\mathbf{AB}=\begin{pmatrix}1&0\\1&0\end{pmatrix}$ and $\mathbf{BA}=\begin{pmatrix}0&0\\1&1\end{pmatrix}$, so $\mathbf{AB}\neq\mathbf{BA}$. Both have trace 1 and determinant 0, so both characteristic polynomials are $x^2-x=x(x-1)$.

7.10.6.3 C. Going deeper

Exercise. C1: Companion matrices

Let $n \geq 2$ and $p(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0 \in F[x]$. The **companion matrix** of p is the matrix $\mathbf{C}(p) \in M_n(F)$ with 1 in each position $(i+1, i)$ just below the diagonal, last column $(-a_0, -a_1, \ldots, -a_{n-1})$, and 0 elsewhere:

$$\mathbf{C}(p) = \begin{pmatrix} 0 & 0 & \cdots & 0 & -a_0 \\ 1 & 0 & \cdots & 0 & -a_1 \\ 0 & 1 & \cdots & 0 & -a_2 \\ \vdots & & \ddots & & \vdots \\ 0 & 0 & \cdots & 1 & -a_{n-1} \end{pmatrix}.$$

(The matrix $\mathbf{C}$ of Exercise (B2: An annihilating polynomial by elimination) is $\mathbf{C}(x^3 - 2x^2 + x - 2)$.)

(a) Let $\mathbf{M} = x\mathbf{I} - \mathbf{C}(p)$. For each i, show that deleting row i and column n of $\mathbf{M}$ leaves a block diagonal matrix whose blocks are an $(i-1) \times (i-1)$ lower triangular matrix with diagonal $x, \ldots, x$ and an $(n-i) \times (n-i)$ upper triangular matrix with diagonal $-1, \ldots, -1$ (either block may be absent when $i = 1$ or $i = n$). Deduce that the cofactor C_{in} of $\mathbf{M}$ equals x^{i-1}.

(b) Deduce, by expanding along the last column, that $p_{\mathbf{C}(p)} = p$.

(c) Verify (b) for $p = x^4 + x^3 - 2x + 3$ by writing $\mathbf{C}(p)$ and checking the coefficients predicted by Theorem 7.10.2.

(d) Deduce that every monic polynomial of degree $n \geq 1$ over F is the characteristic polynomial of some $n \times n$ matrix over F.

Solution.

(a) The entries of $\mathbf{M}$ are: x at (j, j) for $j < n$; -1 at $(j+1, j)$; a_{j-1} at (j, n) for $j < n$; $x + a_{n-1}$ at (n, n); 0 elsewhere. Delete row i and column n. In the remaining matrix, rows $1, \ldots, i-1$ have nonzero entries only in columns $j-1$ and j for row j, all at most $i-1$; and rows $i+1, \ldots, n$ have nonzero entries only in columns $j-1$ and $j \leq n-1$, all at least i. So the matrix is block diagonal with blocks on rows and columns $\{1, \ldots, i-1\}$ and on rows $\{i+1, \ldots, n\}$, columns $\{i, \ldots, n-1\}$. The first block has x on its diagonal and -1 just below it, so it is lower triangular. In the second block, row j of $\mathbf{M}$ (for $j > i$) has -1 in column $j-1$, which is the diagonal position of the block, and x in column j, one step to the right; so it is upper triangular with diagonal $-1, \ldots, -1$. By Theorem 7.8.4 and Theorem 7.4.6 over $F[x]$ (both division-free), the minor is $M_{in} = x^{i-1}(-1)^{n-i}$, where an absent block contributes the factor 1 (for $i = 1$ or $i = n$ the matrix is a single triangular block). Hence $C_{in} = (-1)^{i+n}(-1)^{n-i}x^{i-1} = (-1)^{2n}x^{i-1} = x^{i-1}$.

(b) By Theorem 7.6.1 along column n, valid over $F[x]$,

$$\begin{aligned} p_{\mathbf{C}(p)} = \det \mathbf{M} &= \sum_{i=1}^{n-1} a_{i-1}\, x^{i-1} + (x + a_{n-1})\, x^{n-1} \\ &= a_0 + a_1 x + \cdots + a_{n-1}x^{n-1} + x^n = p. \end{aligned}$$

(c) Here $a_3 = 1$, $a_2 = 0$, $a_1 = -2$, $a_0 = 3$, so

$$\mathbf{C}(p) = \begin{pmatrix} 0 & 0 & 0 & -3 \\ 1 & 0 & 0 & 2 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix}.$$

Its trace is -1, so Theorem 7.10.2 predicts the coefficient $1 = a_3$ at x^3, as in p. Its determinant, expanding along the first row, whose only nonzero entry is -3 in position $(1,4)$, is $(-3)(-1)^{1+4} \det \mathbf{I}_3 = 3$, since deleting row 1 and column 4 leaves $\mathbf{I}_3$. So the predicted constant term is $(-1)^4 \cdot 3 = 3 = a_0$.

(d) For $n \geq 2$, take $\mathbf{C}(p)$ and use (b). For $n = 1$, $p = x + a_0$ is $p_{(-a_0)}$.

Exercise. C2: Nilpotent 2×2 matrices

Let $\mathbf{A} \in M_2(F)$ satisfy $\mathbf{A}^k = \mathbf{O}$ for some $k \geq 1$.

(a) Show that the constant term of $p_{\mathbf{A}}$ is 0.

(b) By multiplying out, verify that $\mathbf{A}^2 - (\operatorname{tr} \mathbf{A})\mathbf{A} + (\det \mathbf{A})\mathbf{I} = \mathbf{O}$ for every $\mathbf{A} \in M_2(F)$.

(c) Deduce that $\operatorname{tr} \mathbf{A} = 0$, and hence $p_{\mathbf{A}} = x^2$ and $\mathbf{A}^2 = \mathbf{O}$.

(d) Give an example showing that $\mathbf{A}^2 = \mathbf{O}$ does not force $\mathbf{A} = \mathbf{O}$.

Solution.

(a) By Theorem 7.5.4, $(\det \mathbf{A})^k = \det(\mathbf{A}^k) = \det \mathbf{O} = 0$, so $\det \mathbf{A} = 0$ because a field has no zero divisors (Theorem 0.7.1). The constant term is $(-1)^2 \det \mathbf{A} = 0$ by Theorem 7.10.2.

(b) Let $\mathbf{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$. Then $\mathbf{A}^2 = \begin{pmatrix} a^2 + bc & ab + bd \\ ca + dc & bc + d^2 \end{pmatrix}$ and $(a+d)\mathbf{A} = \begin{pmatrix} a^2 + ad & ab + bd \\ ac + dc & ad + d^2 \end{pmatrix}$. The difference is $\begin{pmatrix} bc - ad & 0 \\ 0 & bc - ad \end{pmatrix} = -(\det \mathbf{A})\mathbf{I}$, as claimed.

(c) By (a) and (b), $\mathbf{A}^2 = (\operatorname{tr} \mathbf{A})\mathbf{A}$. By induction, $\mathbf{A}^m = (\operatorname{tr} \mathbf{A})^{m-1}\mathbf{A}$ for all $m \geq 1$: if it holds for m, then $\mathbf{A}^{m+1} = (\operatorname{tr} \mathbf{A})^{m-1}\mathbf{A}^2 = (\operatorname{tr} \mathbf{A})^m \mathbf{A}$. If $\mathbf{A} = \mathbf{O}$, then $\operatorname{tr} \mathbf{A} = 0$. If $\mathbf{A} \neq \mathbf{O}$, then $\mathbf{O} = \mathbf{A}^k = (\operatorname{tr} \mathbf{A})^{k-1}\mathbf{A}$ forces $(\operatorname{tr} \mathbf{A})^{k-1} = 0$, since a nonzero matrix times a scalar is zero only if the scalar is zero; here $k \geq 2$, because $k = 1$ would give $\mathbf{A} = \mathbf{O}$. Hence $\operatorname{tr} \mathbf{A} = 0$ by Theorem 0.7.1 (f), applied repeatedly. So $p_{\mathbf{A}} = x^2 - 0 \cdot x + 0 = x^2$, and $\mathbf{A}^2 = (\operatorname{tr} \mathbf{A})\mathbf{A} = \mathbf{O}$.

(d) $\mathbf{A} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$ has $\mathbf{A}^2 = \mathbf{O}$ and $\mathbf{A} \neq \mathbf{O}$.

Exercise. C3: Odd size over the reals

Let n be odd and $\mathbf{A} \in M_n(\mathbb{R})$.

(a) Prove that there is $c \in \mathbb{R}$ and a nonzero $\mathbf{v} \in \mathbb{R}^n$ with $\mathbf{A}\mathbf{v} = c\mathbf{v}$.

(b) Show by an example that (a) fails for $n = 2$.

(c) Show that (a) holds for every $\mathbf{A} \in M_n(\mathbb{C})$ and every $n \geq 1$.

Solution.

(a) By Theorem 7.10.2, $p_{\mathbf{A}} \in \mathbb{R}[x]$ has odd degree n. By Corollary 6.5.9, it has a real root c. By Theorem 7.10.4, there is a nonzero $\mathbf{v} \in \mathbb{R}^n$ with $\mathbf{Av} = c\mathbf{v}$.

(b) The rotation $\mathbf{R} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ has $p_{\mathbf{R}} = x^2 + 1$, with no real root; by Theorem 7.10.4 no real c and nonzero real $\mathbf{v}$ satisfy $\mathbf{Rv} = c\mathbf{v}$.

(c) $p_{\mathbf{A}} \in \mathbb{C}[x]$ has degree $n \geq 1$, so it has a root $c \in \mathbb{C}$ by Theorem 6.5.1. Apply Theorem 7.10.4 over $F = \mathbb{C}$.

Chapter 8

Block Matrices

A large matrix often has structure that its entries hide. Split it into blocks, smaller matrices arranged in a grid, and the structure shows. A direct sum of subspaces becomes a block-diagonal matrix. A subspace that an operator maps into itself becomes a block of zeros in the corner. A system of equations that splits into two groups becomes a two-by-two system whose "entries" are matrices.

This chapter makes block arithmetic rigorous and then puts it to work. Multiplying block by block, eliminating a whole block at once, and taking the Schur complement $\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B}$ turn statements about big matrices into statements about small ones: ranks, determinants, and inverses. The Kronecker product rewrites matrix equations such as $\mathbf{A}\mathbf{X} + \mathbf{X}\mathbf{B} = \mathbf{C}$ as ordinary linear systems. The chapter ends with a question that looks like it needs heavy machinery but does not: which matrices can be written as a commutator $\mathbf{X}\mathbf{Y} - \mathbf{Y}\mathbf{X}$?

What you need. From Chapter 0, matrix multiplication (Definition 0.9.9) and polynomials without zero divisors (Corollary 0.8.4). From Chapter 1, subspaces, spans, and bases with the dimension formula for a sum (Theorem 1.3.1, Theorem 1.6.5, Theorem 1.6.3, Theorem 1.6.7, Proposition 1.7.2, Theorem 1.7.3), the unique representation of a vector in a basis (Theorem 1.5.1, Theorem 1.4.3), and the direct sum criteria (Theorem 1.7.5), which is what §01 reads off a block decomposition; §05 also uses the standard bases and the trace-free matrices as examples (Example (Standard Bases), Example (Trace-Free Matrices), Theorem 1.1.9). From Chapter 3, rank (Definition 3.5.4). From Chapter 2, the matrix of a linear map and change of basis (Definition 4.1.1, Theorem 4.2.3), quotient spaces (Definition 4.4.3), the rank inequalities (Theorem 4.3.5), and the trace (Definition 4.5.2). From Chapter 7, determinants of block triangular matrices and the characteristic polynomial (Theorem 7.8.4, Theorem 7.10.2).

8.0.1 Roadmap

- **Partitions and block multiplication.** Multiplying matrices block by block, and what block-diagonal and block-triangular shapes say about a linear map.
- **Rank and block matrices.** Rank inequalities proved by clearing blocks, including Frobenius's inequality.
- **Schur complements.** The Schur complement, the determinant, rank, and inverse of a block matrix, the Woodbury formula, and determinants with commuting blocks over any field.
- **The Kronecker product.** A product of matrices of different sizes, and matrix equations turned into linear systems.
- **Commutators and trace-zero matrices.** Every trace-zero matrix over a field of characteristic zero is a commutator.

8.0.2 Named moves

- **Partition to match the structure.** Choose the blocks so that the zeros you know about sit in whole blocks.
- **Eliminate a block.** Multiply by an invertible block triangular matrix to clear a block, just as a row operation clears an entry.
- **Pass to the Schur complement.** Reduce a question about $\begin{pmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{C} & \mathbf{D}\end{pmatrix}$ with $\mathbf{A}$ invertible to a question about $\mathbf{A}$ and $\mathbf{D}-\mathbf{C}\mathbf{A}^{-1}\mathbf{B}$.
- **Vectorize.** Turn a linear equation in an unknown matrix $\mathbf{X}$ into a linear system in the entries of $\mathbf{X}$.
- **Adjoin an indeterminate.** To remove an invertibility hypothesis, replace $\mathbf{D}$ by $\mathbf{D}+x\mathbf{I}$, work over $F[x]$, and set $x=0$ at the end.

8.1 Partitions and Block Multiplication

A large matrix often has structure that its individual entries hide: a corner of zeros, a copy of the identity, a smaller matrix repeated. Chapter 7 already met one such shape, the block triangular matrix, and computed its determinant from two smaller determinants. This section makes the idea systematic. We cut matrices into rectangular pieces, prove that products can be computed piece by piece, and use the pieces to read off structure: direct sums become block diagonal matrices, and a subspace that an operator maps into itself becomes a zero block in the corner.

8.1.1 Cutting a matrix into blocks

Take two 4×4 matrices over $\mathbb{Q}$ whose top-right 2×2 corners are zero:

$$\mathbf{M}=\left(\begin{array}{cc|cc} 1 & 2 & 0 & 0\\ 0 & 1 & 0 & 0\\ \hline 1 & 0 & 2 & 1\\ 2 & 1 & 1 & 1\end{array}\right),\qquad \mathbf{N}=\left(\begin{array}{cc|cc} 1 & -1 & 0 & 0\\ 1 & 0 & 0 & 0\\ \hline 0 & 1 & 1 & 0\\ 1 & 1 & 3 & 1\end{array}\right).$$

The lines cut each matrix into four 2×2 pieces. Name them: $\mathbf{M}=\begin{pmatrix}\mathbf{A} & \mathbf{O}\\ \mathbf{C} & \mathbf{D}\end{pmatrix}$ and $\mathbf{N}=\begin{pmatrix}\mathbf{E} & \mathbf{O}\\ \mathbf{G} & \mathbf{H}\end{pmatrix}$, with $\mathbf{A}=\begin{pmatrix}1 & 2\\ 0 & 1\end{pmatrix}$, $\mathbf{C}=\begin{pmatrix}1 & 0\\ 2 & 1\end{pmatrix}$, and so on. If we pretend that the pieces are numbers and multiply as for 2×2 matrices, keeping the order of every product, we get

$$\mathbf{MN}\overset{?}{=}\begin{pmatrix}\mathbf{AE}+\mathbf{OG} & \mathbf{AO}+\mathbf{OH}\\ \mathbf{CE}+\mathbf{DG} & \mathbf{CO}+\mathbf{DH}\end{pmatrix}=\begin{pmatrix}\mathbf{AE} & \mathbf{O}\\ \mathbf{CE}+\mathbf{DG} & \mathbf{DH}\end{pmatrix}.$$

With $\mathbf{AE}=\begin{pmatrix}3 & -1\\ 1 & 0\end{pmatrix}$, $\mathbf{CE}+\mathbf{DG}=\begin{pmatrix}2 & 2\\ 4 & 0\end{pmatrix}$, and $\mathbf{DH}=\begin{pmatrix}5 & 1\\ 4 & 1\end{pmatrix}$, this predicts

$$\mathbf{MN}=\begin{pmatrix}3 & -1 & 0 & 0\\ 1 & 0 & 0 & 0\\ 2 & 2 & 5 & 1\\ 4 & 0 & 4 & 1\end{pmatrix},$$

and multiplying out the sixteen entries directly gives exactly this matrix. Two things happened. The zero corner of the product came for free, with no arithmetic. And the product of two 4×4 matrices became four products of 2×2 matrices, **AE**, **CE**, **DG**, and **DH**; the other four products in the block formula have a zero factor and cost nothing. To use this safely we need to say precisely what "cutting" means and prove that the pretend multiplication is always correct.

Draw horizontal lines between rows and vertical lines between columns; the pieces are the blocks.

Definition 8.1.1. Block Partition

Let $\mathbf{A} = (a_{ij}) \in M_{m\times n}(F)$. A **row partition** of $\mathbf{A}$ is a list of **positive** integers $m_1, \dots, m_a$ with $m_1 + \cdots + m_a = m$, and a **column partition** is a list of positive integers $n_1, \dots, n_b$ with $n_1 + \cdots + n_b = n$. Put $\mu_r = m_1 + \cdots + m_r$ and $\nu_s = n_1 + \cdots + n_s$, with $\mu_0 = \nu_0 = 0$. For $1 \le r \le a$ and $1 \le s \le b$, the (r, s) **block** of $\mathbf{A}$ is the matrix $\mathbf{A}_{rs} \in M_{m_r\times n_s}(F)$ with entries

$$(\mathbf{A}_{rs})_{ij} = a_{\mu_{r-1}+i,\ \nu_{s-1}+j} \qquad (1 \le i \le m_r,\ 1 \le j \le n_s).$$

We then write $\mathbf{A} = (\mathbf{A}_{rs})$, displayed as an $a \times b$ array of blocks, and call $\mathbf{A}$ a **block matrix**.

Let $\mathbf{B} \in M_{n\times p}(F)$. Partitions of $\mathbf{A}$ and $\mathbf{B}$ are **conformable** if the column partition of $\mathbf{A}$ and the row partition of $\mathbf{B}$ are **the same list** $n_1, \dots, n_b$, in the same order.

In words: the row partition says where the horizontal lines go. Block row r consists of the rows with indices in $I_r = \{\mu_{r-1}+1, \dots, \mu_r\}$, a run of m_r consecutive rows, and similarly block column s consists of the columns in $J_s = \{\nu_{s-1} + 1, \dots, \nu_s\}$. The block $\mathbf{A}_{rs}$ is what survives when we keep only the rows in I_r and the columns in J_s, renumbered from 1. Conformability says that the vertical lines through $\mathbf{A}$ sit at the same places as the horizontal lines through $\mathbf{B}$. That is exactly what will make each product "block of $\mathbf{A}$ times block of $\mathbf{B}$" defined.

Well-definedness. The runs $I_1, \dots, I_a$ are disjoint and cover $\{1, \dots, m\}$, and the same holds for $J_1, \dots, J_b$. So every entry of $\mathbf{A}$ lies in exactly one block. Conversely, if we are given matrices $\mathbf{A}_{rs} \in M_{m_r\times n_s}(F)$, there is exactly one $\mathbf{A} \in M_{m\times n}(F)$ with these blocks. Block sizes must fit: all blocks in block row r have m_r rows, and all blocks in block column s have n_s columns. We allow only positive parts, so no block is empty.

Examples.

- **Every entry a block.** With $m_r = 1$ and $n_s = 1$ for all r, s, the block $\mathbf{A}_{rs}$ is the 1×1 matrix (a_{rs}). A matrix is its own finest block matrix.
- **One block.** With $a = b = 1$, the single block is $\mathbf{A}$ itself. This degenerate partition matters because a statement about block matrices then says something about ordinary matrices, and the statement must not break.
- **Columns as blocks.** Row partition m, column partition $1, 1, \dots, 1$. The blocks are the n columns of $\mathbf{A}$: $\mathbf{A} = \begin{pmatrix}\mathbf{a}_1 & \cdots & \mathbf{a}_n\end{pmatrix}$, as in Chapter 0.
- **The hook.** For $\mathbf{M}$ and $\mathbf{N}$ above, both partitions are $2, 2$, so the column partition of $\mathbf{M}$ equals the row partition of $\mathbf{N}$: they are conformable.

Nonexample by minimal change. Keep $\mathbf{M}$ with column partition $2, 2$, but cut $\mathbf{N}$ with row partition $1, 3$. The product $\mathbf{MN}$ is still defined, and both partitions are legitimate. What fails is conformability: the first block column of $\mathbf{M}$ has 2 columns, while the first block row of $\mathbf{N}$ has 1 row, so the "product" $\mathbf{A} \cdot \mathbf{N}_{11}$ of a 2×2 matrix with a 1×2 matrix is undefined.

Quick check. Let $\mathbf{A} \in M_{3\times 5}(F)$ have row partition $1, 2$ and column partition $3, 2$, and let $\mathbf{B} \in M_{5\times 4}(F)$ have row partition $2, 3$. Are the partitions conformable? What are the sizes of the blocks $\mathbf{A}_{21}$ and $\mathbf{B}_{2t}$?

Solution. No. The column partition of $\mathbf{A}$ is $3, 2$ and the row partition of $\mathbf{B}$ is $2, 3$. Both lists sum to 5, but they differ in order, and conformability asks for the same list. The block $\mathbf{A}_{21}$ is 2×3 (block row 2 has 2 rows, block column 1 has 3 columns), and every $\mathbf{B}_{2t}$ has 3 rows.

8.1.2 Block multiplication

Here is the theorem that justifies the computation in the hook. The proof is nothing but the entry formula for a product, with the sum over the middle index cut into the runs $J_1, \dots, J_b$.

Theorem 8.1.1. Block Multiplication

Let $\mathbf{A} \in M_{m\times n}(F)$ and $\mathbf{B} \in M_{n\times p}(F)$ be conformably partitioned: $\mathbf{A}$ with row partition $m_1, \dots, m_a$ and column partition $n_1, \dots, n_b$, and $\mathbf{B}$ with row partition $n_1, \dots, n_b$ and column partition $p_1, \dots, p_c$. Partition $\mathbf{AB} \in M_{m\times p}(F)$ with row partition $m_1, \dots, m_a$ and column partition $p_1, \dots, p_c$. Then for all $1 \le r \le a$ and $1 \le t \le c$,

$$(\mathbf{AB})_{rt} = \sum_{s=1}^{b} \mathbf{A}_{rs}\mathbf{B}_{st} = \mathbf{A}_{r1}\mathbf{B}_{1t} + \mathbf{A}_{r2}\mathbf{B}_{2t} + \cdots + \mathbf{A}_{rb}\mathbf{B}_{bt}.$$

Each product $\mathbf{A}_{rs}\mathbf{B}_{st}$ is defined and lies in $M_{m_r\times p_t}(F)$.

Idea. Look at one entry of $\mathbf{AB}$, in row i^* and column k^*. It is a sum over the middle index $j = 1, \dots, n$. Cut that sum into the runs $J_1, \dots, J_b$ given by the common partition. The part over J_s pairs row i^* of $\mathbf{A}$, restricted to the columns of block column s, with column k^* of $\mathbf{B}$, restricted to the rows of block row s. That is one entry of $\mathbf{A}_{rs}\mathbf{B}_{st}$. The only real work is translating indices between the big matrix and the blocks.

Proof. Let μ_r, ν_s, ρ_t be the partial sums of the three partitions, as in Definition 8.1.1, and $J_s = \{\nu_{s-1} + 1, \dots, \nu_s\}$. The block $\mathbf{A}_{rs}$ is $m_r \times n_s$ and $\mathbf{B}_{st}$ is $n_s \times p_t$, so their product is defined and is $m_r \times p_t$, as is the block $(\mathbf{AB})_{rt}$.

Fix r, t, and let $1 \le i \le m_r$ and $1 \le k \le p_t$. Put $i^* = \mu_{r-1} + i$ and $k^* = \rho_{t-1} + k$. By Definition 8.1.1 and Definition 0.9.9,

$$\big((\mathbf{AB})_{rt}\big)_{ik} = (\mathbf{AB})_{i^*k^*} = \sum_{j=1}^{n} a_{i^*j}\, b_{jk^*}.$$

The sets $J_1, \dots, J_b$ are disjoint with union $\{1, \dots, n\}$, so the finite sum may be grouped according to them. In the group for J_s, substitute $j = \nu_{s-1} + l$; as l runs through $1, \dots, n_s$,

j runs through J_s exactly once. Hence

$$\begin{aligned}\big((\mathbf{AB})_{rt}\big)_{ik} &= \sum_{s=1}^{b} \sum_{j \in J_s} a_{i^* j}\, b_{j k^*} \\ &= \sum_{s=1}^{b} \sum_{l=1}^{n_s} a_{\mu_{r-1}+i,\ \nu_{s-1}+l}\ b_{\nu_{s-1}+l,\ \rho_{t-1}+k} \\ &= \sum_{s=1}^{b} \sum_{l=1}^{n_s} (\mathbf{A}_{rs})_{il}\, (\mathbf{B}_{st})_{lk},\end{aligned}$$

where the last step is the definition of the blocks $\mathbf{A}_{rs}$ and $\mathbf{B}_{st}$. By Definition 0.9.9 the inner sum is $(\mathbf{A}_{rs}\mathbf{B}_{st})_{ik}$, and by Definition 0.9.6 the outer sum is the (i,k)-entry of $\sum_s \mathbf{A}_{rs}\mathbf{B}_{st}$. Since i and k were arbitrary, $(\mathbf{AB})_{rt} = \sum_s \mathbf{A}_{rs}\mathbf{B}_{st}$. This proves the theorem.

■

So conformably partitioned matrices multiply like matrices whose entries are matrices. The commonest case is a 2×2 block matrix:

$$\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix} \begin{pmatrix} \mathbf{A}' & \mathbf{B}' \\ \mathbf{C}' & \mathbf{D}' \end{pmatrix} = \begin{pmatrix} \mathbf{AA}' + \mathbf{BC}' & \mathbf{AB}' + \mathbf{BD}' \\ \mathbf{CA}' + \mathbf{DC}' & \mathbf{CB}' + \mathbf{DD}' \end{pmatrix},$$

whenever the partitions are conformable. Sums and scalar multiples of matrices with the same partition are also computed blockwise, straight from Definition 0.9.6 and Definition 0.9.7, since each entry lies in exactly one block.

Over a commutative ring. The proof uses only sums and products of entries and the regrouping of a finite sum. So block multiplication holds verbatim for matrices with entries in any commutative ring R in the sense of Chapter 7 §4, for instance $R = F[x]$. Section 3 uses this.

Warning. **Keep the order inside every block product.** The top-left block of the product above is $\mathbf{AA}' + \mathbf{BC}'$, not $\mathbf{A}'\mathbf{A} + \mathbf{C}'\mathbf{B}$: blocks are matrices, and matrices do not commute. In the hook, $\mathbf{CE} + \mathbf{DG} = \begin{pmatrix} 2 & 2 \\ 4 & 0 \end{pmatrix}$, while $\mathbf{EC} + \mathbf{GD} = \begin{pmatrix} 0 & 0 \\ 4 & 2 \end{pmatrix}$. For the same reason, the formula $ad - bc$ for a 2×2 determinant does **not** survive with blocks: in general $\det \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix} \neq \det(\mathbf{AD} - \mathbf{BC})$. Take $\mathbf{A} = \mathbf{I}_2$, $\mathbf{B} = \mathbf{E}_{12}$, $\mathbf{C} = \mathbf{E}_{21}$, $\mathbf{D} = \mathbf{E}_{11}$ in $M_2(\mathbb{Q})$. Then $\mathbf{AD} - \mathbf{BC} = \mathbf{E}_{11} - \mathbf{E}_{12}\mathbf{E}_{21} = \mathbf{E}_{11} - \mathbf{E}_{11} = \mathbf{O}$, with determinant 0. But

$$\mathbf{M} = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix},$$

and subtracting row 4 from row 1, which does not change the determinant (Theorem 7.5.1), gives the permutation matrix that swaps $\mathbf{e}_1$ and $\mathbf{e}_4$, whose determinant is -1. So $\det \mathbf{M} = -1 \neq 0$. Section 3 shows that the formula does hold when $\mathbf{C}$ and $\mathbf{D}$ commute; here $\mathbf{CD} = \mathbf{E}_{21} \neq \mathbf{O} = \mathbf{DC}$.

Transposing a block matrix transposes each block and reflects the array of blocks.

Proposition 8.1.2. Transpose of a Block Matrix

Let $\mathbf{A} \in M_{m\times n}(F)$ have row partition $m_1, \dots, m_a$, column partition $n_1, \dots, n_b$, and blocks $\mathbf{A}_{rs}$. Give $\mathbf{A}^\top$ the row partition $n_1, \dots, n_b$ and the column partition $m_1, \dots, m_a$. Then $(\mathbf{A}^\top)_{sr} = (\mathbf{A}_{rs})^\top$. In particular $\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}^\top = \begin{pmatrix} \mathbf{A}^\top & \mathbf{C}^\top \\ \mathbf{B}^\top & \mathbf{D}^\top \end{pmatrix}$.

Proof. For $1 \le j \le n_s$ and $1 \le i \le m_r$, by Definition 8.1.1 and Definition 0.9.10, $\left((\mathbf{A}^\top)_{sr}\right)_{ji} = (\mathbf{A}^\top)_{\nu_{s-1}+j,\ \mu_{r-1}+i} = a_{\mu_{r-1}+i,\ \nu_{s-1}+j} = (\mathbf{A}_{rs})_{ij} = \left((\mathbf{A}_{rs})^\top\right)_{ji}$.

■

8.1.3 Rows, columns, and outer products

The finest and coarsest partitions turn Theorem 8.1.1 into facts we already know, and one mixed choice gives something new.

Partition $\mathbf{A} \in M_{m\times n}(F)$ trivially (one block) and $\mathbf{B} \in M_{n\times p}(F)$ into columns (row partition n, column partition $1, \dots, 1$). These are conformable, and the theorem says that the k-th block of $\mathbf{AB}$, its k-th column, is $\mathbf{A}\mathbf{b}_k$. That is the column view of Theorem 0.9.2. Partitioning $\mathbf{A}$ into rows and $\mathbf{B}$ trivially gives the row view in the same way.

Now cut $\mathbf{A}$ into columns and $\mathbf{B}$ into rows. Both middle partitions are $1, 1, \dots, 1$ (with n ones), so they are conformable, and the product has a single block.

Corollary 8.1.3. Outer Product Expansion

Let $\mathbf{A} \in M_{m\times n}(F)$ have columns $\mathbf{a}_1, \dots, \mathbf{a}_n \in F^m$, and let $\mathbf{B} \in M_{n\times p}(F)$ have rows $\mathbf{r}_1, \dots, \mathbf{r}_n \in M_{1\times p}(F)$. Then

$$\mathbf{AB} = \mathbf{a}_1\mathbf{r}_1 + \mathbf{a}_2\mathbf{r}_2 + \cdots + \mathbf{a}_n\mathbf{r}_n = \sum_{k=1}^{n} \mathbf{a}_k\mathbf{r}_k,$$

where each $\mathbf{a}_k\mathbf{r}_k \in M_{m\times p}(F)$ is column k of $\mathbf{A}$ times row k of $\mathbf{B}$.

Proof. Give $\mathbf{A}$ the row partition m and the column partition $1, \dots, 1$, so that its blocks are $\mathbf{A}_{1k} = \mathbf{a}_k$; give $\mathbf{B}$ the row partition $1, \dots, 1$ and the column partition p, so that $\mathbf{B}_{k1} = \mathbf{r}_k$. The partitions are conformable, and by Theorem 8.1.1 the single block of $\mathbf{AB}$ is $\sum_k \mathbf{A}_{1k}\mathbf{B}_{k1} = \sum_k \mathbf{a}_k\mathbf{r}_k$.

■

Each term $\mathbf{a}_k\mathbf{r}_k$ is a column times a row, an **outer product**, and it has rank at most 1: its columns are the multiples $r_{kj}\mathbf{a}_k$ of the single vector $\mathbf{a}_k$. So the corollary writes every product of an $m\times n$ and an $n\times p$ matrix as a sum of n matrices of rank at most one, in the spirit of Exercise (B3: Rank-one pieces), which writes a matrix of rank r as a sum of r rank-one matrices. For

example,

$$\begin{aligned}
&\begin{pmatrix} 1 & 2 \\ 3 & 4 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 2 \\ -1 & 1 & 1 \end{pmatrix} \\
&\quad = \begin{pmatrix} 1 \\ 3 \\ 0 \end{pmatrix} \begin{pmatrix} 1 & 0 & 2 \end{pmatrix} + \begin{pmatrix} 2 \\ 4 \\ 1 \end{pmatrix} \begin{pmatrix} -1 & 1 & 1 \end{pmatrix} \\
&\quad = \begin{pmatrix} 1 & 0 & 2 \\ 3 & 0 & 6 \\ 0 & 0 & 0 \end{pmatrix} + \begin{pmatrix} -2 & 2 & 2 \\ -4 & 4 & 4 \\ -1 & 1 & 1 \end{pmatrix} \\
&\quad = \begin{pmatrix} -1 & 2 & 4 \\ -1 & 4 & 10 \\ -1 & 1 & 1 \end{pmatrix},
\end{aligned}$$

which is the product computed entry by entry.

8.1.4 Block diagonal matrices

Diagonal matrices are the easiest matrices to compute with: they multiply, power, and invert entry by entry along the diagonal. Replace the diagonal entries by square blocks and all of this survives, one block at a time.

A block diagonal matrix is several square matrices placed corner to corner along the diagonal, with zeros everywhere else.

Definition 8.1.2. Block Diagonal Matrix

Let $k_1, \ldots, k_r$ be positive integers and $\mathbf{A}_i \in M_{k_i}(F)$ **square** for each i. Their **direct sum** $\mathbf{A}_1 \oplus \cdots \oplus \mathbf{A}_r$ is the matrix in $M_{k_1+\cdots+k_r}(F)$ which, with row partition and column partition both equal to $k_1, \ldots, k_r$, has (i,i) block $\mathbf{A}_i$ and (i,j) block $\mathbf{O}$ for $i \neq j$. A square matrix is **block diagonal with respect to** the partition $k_1, \ldots, k_r$ if it has this form. For two blocks,

$$\mathbf{A} \oplus \mathbf{B} = \begin{pmatrix} \mathbf{A} & \mathbf{O} \\ \mathbf{O} & \mathbf{B} \end{pmatrix}.$$

In words: the same partition is used for rows and columns, so the diagonal blocks are square; the blocks on the diagonal may be anything; every block off the diagonal is zero. Whether a matrix is block diagonal depends on the partition, which is part of the statement.

Examples.

- **Diagonal matrices.** $\operatorname{diag}(d_1, \ldots, d_n) = (d_1) \oplus \cdots \oplus (d_n)$, with all $k_i = 1$.
- **The identity.** $\mathbf{I}_k \oplus \mathbf{I}_l = \mathbf{I}_{k+l}$, and more generally $\mathbf{I}_k \oplus \mathbf{O}_l$ is the matrix of a projection (Corollary 4.5.2).
- **A genuine block.** $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \oplus (2) = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{pmatrix}$, with partition $2, 1$.
- **One block.** With $r = 1$, $\mathbf{A}_1$ itself. Every square matrix is block diagonal with respect to the trivial partition, which is why the partition must always be named.

Nonexample by minimal change. Change the $(1,3)$-entry of the third example from 0 to 1.

The partition $2,1$ still has square diagonal blocks $\begin{pmatrix}1&1\\0&1\end{pmatrix}$ and (2), and the $(2,1)$ block $\begin{pmatrix}0&0\end{pmatrix}$ is still zero. What fails is the clause that the $(1,2)$ block is zero: it is now $\begin{pmatrix}1\\0\end{pmatrix}$. The matrix is block upper triangular, which is treated below, but not block diagonal for this partition.

Why the symbol $\oplus$. The matrix $\mathbf{A}\oplus\mathbf{B}$ acts on $F^{k+l} = \operatorname{span}(\mathbf{e}_1,\dots,\mathbf{e}_k)\oplus\operatorname{span}(\mathbf{e}_{k+1},\dots,\mathbf{e}_{k+l})$ by sending each of the two summands into itself, acting as $\mathbf{A}$ on the first and as $\mathbf{B}$ on the second. The theorem on adapted bases below makes this precise for any operator and any direct sum, and it is the reason the same symbol is used.

Warning. **$\mathbf{A}\oplus\mathbf{B}$ is a matrix, not a subspace, and the order matters.** For subspaces, $U\oplus W$ and $W\oplus U$ are the same set. For matrices, $(1)\oplus(2) = \operatorname{diag}(1,2)$ and $(2)\oplus(1) = \operatorname{diag}(2,1)$ are different matrices. They are similar (swap the two basis vectors), but they are not equal.

Theorem 8.1.4. Arithmetic of Block Diagonal Matrices

Let $\mathbf{A},\mathbf{A}'\in M_k(F)$, $\mathbf{B},\mathbf{B}'\in M_l(F)$, and $c\in F$. Then:

(a) $(\mathbf{A}\oplus\mathbf{B})+(\mathbf{A}'\oplus\mathbf{B}') = (\mathbf{A}+\mathbf{A}')\oplus(\mathbf{B}+\mathbf{B}')$, $c(\mathbf{A}\oplus\mathbf{B}) = c\mathbf{A}\oplus c\mathbf{B}$, and $(\mathbf{A}\oplus\mathbf{B})(\mathbf{A}'\oplus\mathbf{B}') = \mathbf{A}\mathbf{A}'\oplus\mathbf{B}\mathbf{B}'$;

(b) $(\mathbf{A}\oplus\mathbf{B})^j = \mathbf{A}^j\oplus\mathbf{B}^j$ for every $j\in\mathbb{N}$, and $p(\mathbf{A}\oplus\mathbf{B}) = p(\mathbf{A})\oplus p(\mathbf{B})$ for every $p\in F[x]$;

(c) $\mathbf{A}\oplus\mathbf{B}$ is invertible if and only if $\mathbf{A}$ and $\mathbf{B}$ are both invertible, and then $(\mathbf{A}\oplus\mathbf{B})^{-1} = \mathbf{A}^{-1}\oplus\mathbf{B}^{-1}$;

(d) $\det(\mathbf{A}\oplus\mathbf{B}) = \det\mathbf{A}\det\mathbf{B}$ and $\operatorname{tr}(\mathbf{A}\oplus\mathbf{B}) = \operatorname{tr}\mathbf{A}+\operatorname{tr}\mathbf{B}$.

Proof.

(a) Sums and scalar multiples are computed blockwise, and by Theorem 8.1.1 the product has blocks $\mathbf{A}\mathbf{A}'+\mathbf{O}\cdot\mathbf{O} = \mathbf{A}\mathbf{A}'$, $\mathbf{A}\mathbf{O}+\mathbf{O}\mathbf{B}' = \mathbf{O}$, $\mathbf{O}\mathbf{A}'+\mathbf{B}\mathbf{O} = \mathbf{O}$, and $\mathbf{O}\cdot\mathbf{O}+\mathbf{B}\mathbf{B}' = \mathbf{B}\mathbf{B}'$.

(b) By induction on j. For $j=0$, $(\mathbf{A}\oplus\mathbf{B})^0 = \mathbf{I}_{k+l} = \mathbf{I}_k\oplus\mathbf{I}_l$ by Definition 0.9.14. If $(\mathbf{A}\oplus\mathbf{B})^j = \mathbf{A}^j\oplus\mathbf{B}^j$, then $(\mathbf{A}\oplus\mathbf{B})^{j+1} = (\mathbf{A}^j\oplus\mathbf{B}^j)(\mathbf{A}\oplus\mathbf{B}) = \mathbf{A}^{j+1}\oplus\mathbf{B}^{j+1}$ by (a). For $p = \sum_j c_j x^j$, (a) gives $p(\mathbf{A}\oplus\mathbf{B}) = \sum_j c_j(\mathbf{A}^j\oplus\mathbf{B}^j) = \left(\sum_j c_j\mathbf{A}^j\right)\oplus\left(\sum_j c_j\mathbf{B}^j\right) = p(\mathbf{A})\oplus p(\mathbf{B})$.

(c) The matrix $\mathbf{A}\oplus\mathbf{B}$ is block upper triangular with square diagonal blocks, so $\det(\mathbf{A}\oplus\mathbf{B}) = \det\mathbf{A}\det\mathbf{B}$ by Theorem 7.8.4. Its diagonal entries are those of $\mathbf{A}$ followed by those of $\mathbf{B}$, so the traces add (Definition 0.9.12).

(d) ($\Leftarrow$) By (a), $(\mathbf{A}\oplus\mathbf{B})(\mathbf{A}^{-1}\oplus\mathbf{B}^{-1}) = \mathbf{I}_k\oplus\mathbf{I}_l = \mathbf{I}_{k+l}$, so by Theorem 3.4.5 $\mathbf{A}\oplus\mathbf{B}$ is invertible with inverse $\mathbf{A}^{-1}\oplus\mathbf{B}^{-1}$. ($\Rightarrow$) If $\mathbf{A}\oplus\mathbf{B}$ is invertible, then $\det\mathbf{A}\det\mathbf{B}\neq 0$ by (d) and Theorem 7.5.2. A product in a field is nonzero only if both factors are (Theorem 0.7.1 (d)), so $\det\mathbf{A}\neq 0\neq\det\mathbf{B}$, and both are invertible by Theorem 7.5.2.

■

By induction on the number of blocks, the same rules hold for $\mathbf{A}_1\oplus\cdots\oplus\mathbf{A}_r$: everything is done

block by block. For instance $\left(\begin{pmatrix}1&1\\0&1\end{pmatrix}\oplus(2)\right)^5=\begin{pmatrix}1&5\\0&1\end{pmatrix}\oplus(32)$, since $\begin{pmatrix}1&1\\0&1\end{pmatrix}^j=\begin{pmatrix}1&j\\0&1\end{pmatrix}$ by induction.

8.1.5 Block triangular matrices

Allow one more nonzero block: $\mathbf{M}=\begin{pmatrix}\mathbf{A}&\mathbf{B}\\\mathbf{O}&\mathbf{D}\end{pmatrix}$ with $\mathbf{A}\in M_k(F)$, $\mathbf{D}\in M_l(F)$ square, and $\mathbf{B}\in M_{k\times l}(F)$. Such a matrix is **block upper triangular**; with the zero block above the diagonal instead, **block lower triangular**. Chapter 7 computed $\det\mathbf{M}=\det\mathbf{A}\det\mathbf{D}$ (Theorem 7.8.4). Products and inverses keep the shape.

Theorem 8.1.5. Block Triangular Products and Inverses

Let $k,l\geq 1$, and let $\mathbf{M}=\begin{pmatrix}\mathbf{A}&\mathbf{B}\\\mathbf{O}&\mathbf{D}\end{pmatrix}$ and $\mathbf{M}'=\begin{pmatrix}\mathbf{A}'&\mathbf{B}'\\\mathbf{O}&\mathbf{D}'\end{pmatrix}$, with $\mathbf{A},\mathbf{A}'\in M_k(F)$, $\mathbf{D},\mathbf{D}'\in M_l(F)$, and $\mathbf{B},\mathbf{B}'\in M_{k\times l}(F)$.

(a) $\mathbf{MM}'=\begin{pmatrix}\mathbf{AA}'&\mathbf{AB}'+\mathbf{BD}'\\\mathbf{O}&\mathbf{DD}'\end{pmatrix}$; in particular it is block upper triangular.

(b) $\mathbf{M}$ is invertible if and only if $\mathbf{A}$ and $\mathbf{D}$ are both invertible.

(c) In that case

$$\mathbf{M}^{-1}=\begin{pmatrix}\mathbf{A}^{-1}&-\mathbf{A}^{-1}\mathbf{BD}^{-1}\\\mathbf{O}&\mathbf{D}^{-1}\end{pmatrix}.$$

Idea. Where does the formula come from? Solve $\mathbf{M}\begin{pmatrix}\mathbf{x}\\\mathbf{y}\end{pmatrix}=\begin{pmatrix}\mathbf{f}\\\mathbf{g}\end{pmatrix}$ by back substitution with blocks. The second block row says $\mathbf{Dy}=\mathbf{g}$, so $\mathbf{y}=\mathbf{D}^{-1}\mathbf{g}$. The first says $\mathbf{Ax}+\mathbf{By}=\mathbf{f}$, so $\mathbf{x}=\mathbf{A}^{-1}\mathbf{f}-\mathbf{A}^{-1}\mathbf{BD}^{-1}\mathbf{g}$. Reading off the coefficients of $\mathbf{f}$ and $\mathbf{g}$ gives the four blocks of $\mathbf{M}^{-1}$. The proof only has to check the answer.

Proof.

(a) By Theorem 8.1.1, the blocks of $\mathbf{MM}'$ are $\mathbf{AA}'+\mathbf{BO}=\mathbf{AA}'$, $\mathbf{AB}'+\mathbf{BD}'$, $\mathbf{OA}'+\mathbf{DO}=\mathbf{O}$, and $\mathbf{OB}'+\mathbf{DD}'=\mathbf{DD}'$.

(b) By Theorem 7.8.4, $\det\mathbf{M}=\det\mathbf{A}\det\mathbf{D}$. By Theorem 7.5.2, $\mathbf{M}$ is invertible if and only if $\det\mathbf{A}\det\mathbf{D}\neq 0$, which in a field happens exactly when $\det\mathbf{A}\neq 0$ and $\det\mathbf{D}\neq 0$ (Theorem 0.7.1 (f) and (d)), that is, when $\mathbf{A}$ and $\mathbf{D}$ are both invertible.

(c) Let $\mathbf{X}$ be the displayed matrix. By (a) with $\mathbf{A}'=\mathbf{A}^{-1}$, $\mathbf{B}'=-\mathbf{A}^{-1}\mathbf{BD}^{-1}$, and $\mathbf{D}'=\mathbf{D}^{-1}$,

$$\mathbf{MX}=\begin{pmatrix}\mathbf{AA}^{-1}&-\mathbf{AA}^{-1}\mathbf{BD}^{-1}+\mathbf{BD}^{-1}\\\mathbf{O}&\mathbf{DD}^{-1}\end{pmatrix}=\begin{pmatrix}\mathbf{I}_k&\mathbf{O}\\\mathbf{O}&\mathbf{I}_l\end{pmatrix}=\mathbf{I}_{k+l}.$$

By Theorem 3.4.5, $\mathbf{X}=\mathbf{M}^{-1}$.

■

Transposing with Proposition 8.1.2 turns this into the lower triangular version: if $\mathbf{A}$ and $\mathbf{D}$ are invertible, then $\begin{pmatrix} \mathbf{A} & \mathbf{O} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{A}^{-1} & \mathbf{O} \\ -\mathbf{D}^{-1}\mathbf{C}\mathbf{A}^{-1} & \mathbf{D}^{-1} \end{pmatrix}$. A special case is used constantly: $\begin{pmatrix} \mathbf{I}_k & \mathbf{X} \\ \mathbf{O} & \mathbf{I}_l \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{I}_k & -\mathbf{X} \\ \mathbf{O} & \mathbf{I}_l \end{pmatrix}$ for every $\mathbf{X} \in M_{k\times l}(F)$. Adding $\mathbf{X}$ times the second block of coordinates to the first is undone by subtracting it.

Quick check. The matrix $\mathbf{M} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$, with row partition $2, 1$ and column partition $1, 2$, has a zero lower-left block and is invertible. Does this contradict Theorem 8.1.5?

Solution. No. The theorem needs the diagonal blocks square, which means equal row and column partitions. Here the top-left block $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$ is 2×1 and the bottom-right block $\begin{pmatrix} 1 & 0 \end{pmatrix}$ is 1×2, so "$\mathbf{A}$ is invertible" does not even make sense. (The lower-left block is the 1×1 entry in position $(3, 1)$, which is 0.)

8.1.6 Matrices of operators in adapted bases

Now the shapes pay off. For an operator $T \in \mathcal{L}(V)$ and a subspace $U \subseteq V$, we say that T **maps U into U** if $Tu \in U$ for every $u \in U$. Then the **restriction** $T|_U \colon U \to U$, $u \mapsto Tu$, is an operator on U, linear because its values are those of T. Such subspaces are the subject of Chapter 9; here we only ask what they do to the matrix of T.

When $V = U \oplus W$ and a basis of U is followed by a basis of W, the combined list is a basis of V (Theorem 1.7.5), which we call **adapted** to the decomposition. If T maps both pieces into themselves, the matrix splits.

Theorem 8.1.6. Direct Sums Give Block Diagonal Matrices

Let $V = U \oplus W$ be finite-dimensional with $U \neq \{0\} \neq W$, let $\mathcal{B}_U = (u_1, \dots, u_k)$ and $\mathcal{B}_W = (w_1, \dots, w_l)$ be bases of U and W, and let $\mathcal{B} = (u_1, \dots, u_k, w_1, \dots, w_l)$. For $T \in \mathcal{L}(V)$, the following are equivalent:

(a) T maps U into U and W into W;
(b) $[T]_{\mathcal{B}}^{\mathcal{B}}$ is block diagonal with respect to the partition k, l.

In that case $[T]_{\mathcal{B}}^{\mathcal{B}} = [T|_U]_{\mathcal{B}_U}^{\mathcal{B}_U} \oplus [T|_W]_{\mathcal{B}_W}^{\mathcal{B}_W}$.

Proof. By Theorem 1.7.5, $\mathcal{B}$ is a basis of V. By Definition 4.1.1, column $j \le k$ of $[T]_{\mathcal{B}}^{\mathcal{B}}$ holds the coordinates of Tu_j, and column $k + q$ those of Tw_q.

(a) $\Rightarrow$ (b). Since $Tu_j \in U = \operatorname{span}(\mathcal{B}_U)$, write $Tu_j = \sum_i a_{ij} u_i$. By uniqueness of coordinates (Theorem 1.5.1), the last l entries of column j are 0 and the first k are $a_{1j}, \dots, a_{kj}$, which is column j of $[T|_U]_{\mathcal{B}_U}^{\mathcal{B}_U}$. In the same way $Tw_q = \sum_p d_{pq} w_p$, so column $k + q$

is 0 in its first k entries and equals column q of $[T|_W]_{\mathcal{B}_W}^{\mathcal{B}_W}$ in its last l. This shows $[T]_{\mathcal{B}}^{\mathcal{B}} = [T|_U]_{\mathcal{B}_U}^{\mathcal{B}_U} \oplus [T|_W]_{\mathcal{B}_W}^{\mathcal{B}_W}$.

(b) $\Rightarrow$ (a). If the $(2,1)$ block is zero, then the coordinates of each Tu_j vanish in the last l places, so $Tu_j \in \operatorname{span}(\mathcal{B}_U) = U$. For $u = \sum_j c_j u_j \in U$, linearity gives $Tu = \sum_j c_j Tu_j \in U$, since U is a subspace. The $(1,2)$ block being zero gives $Tw_q \in W$ in the same way, and hence T maps W into W.

■

For example, let $T = \frac{d^2}{dx^2}$ on $\mathbb{R}[x]_{\leq 3} = U \oplus W$ with $U = \operatorname{span}(1, x^2)$ (even polynomials) and $W = \operatorname{span}(x, x^3)$ (odd polynomials). Since $T(1) = 0$, $T(x^2) = 2$, $T(x) = 0$, and $T(x^3) = 6x$, T maps each piece into itself, and in the basis $\mathcal{B} = (1, x^2, x, x^3)$

$$[T]_{\mathcal{B}}^{\mathcal{B}} = \begin{pmatrix} 0 & 2 \\ 0 & 0 \end{pmatrix} \oplus \begin{pmatrix} 0 & 6 \\ 0 & 0 \end{pmatrix}.$$

In the standard order $(1, x, x^2, x^3)$ the same operator has the matrix with 2 in position $(1,3)$ and 6 in position $(2,4)$, which is not block diagonal for the partition $2,2$. **Ordering the basis to match the decomposition is part of the method.**

If T maps only one subspace U into itself, we can still take a basis of U first, but the remaining basis vectors need not be sent anywhere special. What survives is a zero block in the lower left, and the lower-right block turns out to describe the operator that T induces on the quotient V/U. This was Exercise (C2: The operator induced on a quotient) in Chapter 2; we now record it as a theorem, with its converse.

Theorem 8.1.7. Subspaces Mapped into Themselves Give Block Triangular Matrices

Let V be finite-dimensional, $T \in \mathcal{L}(V)$, and U a subspace with $\{0\} \neq U \neq V$. Let $\mathcal{B}_U = (u_1, \ldots, u_k)$ be a basis of U, extended to a basis $\mathcal{B} = (u_1, \ldots, u_k, w_1, \ldots, w_l)$ of V.

(a) T maps U into U if and only if the lower-left $l \times k$ block of $[T]_{\mathcal{B}}^{\mathcal{B}}$ is zero.

(b) In that case the rule $\bar{T}(v + U) = Tv + U$ defines an operator $\bar{T}$ on V/U, and

$$[T]_{\mathcal{B}}^{\mathcal{B}} = \begin{pmatrix} [T|_U]_{\mathcal{B}_U}^{\mathcal{B}_U} & \mathbf{B} \\ \mathbf{O} & [\bar{T}]_{\bar{\mathcal{B}}}^{\bar{\mathcal{B}}} \end{pmatrix}$$

for some $\mathbf{B} \in M_{k \times l}(F)$, where $\bar{\mathcal{B}} = (w_1 + U, \ldots, w_l + U)$ is a basis of V/U.

Idea. Read the matrix one column at a time. The first k columns describe Tu_j, which lie in U exactly when their w-coordinates vanish. The last l columns describe Tw_q, which may have any u-part; passing to V/U erases that part and leaves precisely the w-coordinates, which are the lower-right block.

Proof.

(a) Column $j \leq k$ of $[T]_{\mathcal{B}}^{\mathcal{B}}$ is the coordinate vector of Tu_j (Definition 4.1.1), and by Theorem 1.5.1 its last l entries vanish if and only if $Tu_j \in \operatorname{span}(u_1, \ldots, u_k) = U$. If T maps U into U, this holds for every j. Conversely, if it holds for every j, then for

$u = \sum_j c_j u_j \in U$ we get $Tu = \sum_j c_j Tu_j \in U$, because U is a subspace.

(b) Let $\pi\colon V \to V/U$ be the quotient map, linear with kernel U (Theorem 4.4.4). The composite πT is linear (Theorem 2.4.2), and for $u \in U$, $\pi(Tu) = 0$ because $Tu \in U$. So $U \subseteq \ker(\pi T)$, and by Theorem 4.4.10 (a) there is exactly one linear $\bar{T}\colon V/U \to V/U$ with $\bar{T}(v + U) = Tv + U$. By Theorem 4.4.5, $\bar{\mathcal{B}}$ is a basis of V/U.

By (a), the lower-left block is zero, and as in the proof of Theorem 8.1.6, the upper-left block is $[T|_U]^{\mathcal{B}_U}_{\mathcal{B}_U}$. For $1 \le q \le l$, write $Tw_q = \sum_{i=1}^{k} b_{iq} u_i + \sum_{p=1}^{l} d_{pq} w_p$, so that column $k+q$ of $[T]^{\mathcal{B}}_{\mathcal{B}}$ is $(b_{1q}, \dots, b_{kq}, d_{1q}, \dots, d_{lq})$. Applying the linear map π, and using $\pi(u_i) = 0$,

$$\bar{T}(w_q + U) = \pi(Tw_q) = \sum_{p=1}^{l} d_{pq}\,(w_p + U).$$

By Definition 4.1.1, column q of $[\bar{T}]^{\bar{\mathcal{B}}}_{\bar{\mathcal{B}}}$ is $(d_{1q}, \dots, d_{lq})$, which is the lower-right part of column $k+q$ of $[T]^{\mathcal{B}}_{\mathcal{B}}$. This proves (b), with $\mathbf{B} = (b_{iq})$.

■

The upper-left block depends only on the basis $\mathcal{B}_U$, and the lower-right block only on the cosets $w_1 + U, \dots, w_l + U$. So replacing each w_q by $w_q + u'_q$ with $u'_q \in U$ can change the top-right block $\mathbf{B}$, but leaves both diagonal blocks unchanged. The theorem splits T into two smaller operators, and every quantity that sees only diagonal blocks splits with it. By Theorem 7.8.4, $\det T = \det(T|_U)\det \bar{T}$. Applied over $F[x]$ (Chapter 7 §8 notes that the block triangular determinant holds over any commutative ring), $x\mathbf{I} - [T]^{\mathcal{B}}_{\mathcal{B}} = \begin{pmatrix} x\mathbf{I}_k - [T|_U]^{\mathcal{B}_U}_{\mathcal{B}_U} & -\mathbf{B} \\ \mathbf{O} & x\mathbf{I}_l - [\bar{T}]^{\bar{\mathcal{B}}}_{\bar{\mathcal{B}}} \end{pmatrix}$ gives $p_T = p_{T|_U}\, p_{\bar{T}}$. Chapter 9 builds on exactly this factorization.

Example. A Plane Mapped into Itself

Let $T\colon \mathbb{R}^3 \to \mathbb{R}^3$, $T(x, y, z) = (y, z, x)$, and $U = \{(x, y, z) : x + y + z = 0\}$. Show that T maps U into U, find $[T]^{\mathcal{B}}_{\mathcal{B}}$ for the basis $\mathcal{B} = (\mathbf{u}_1, \mathbf{u}_2, \mathbf{e}_3)$ with $\mathbf{u}_1 = (1, -1, 0)$, $\mathbf{u}_2 = (0, 1, -1)$, and interpret the lower-right block.

Solution. If $x + y + z = 0$, then the entries of $T(x, y, z) = (y, z, x)$ also sum to 0, so T maps U into U. The vectors $\mathbf{u}_1, \mathbf{u}_2$ lie in U and are independent (look at the first entries, then the last), and $\dim U = 2$ since U is the kernel of the nonzero functional $(x, y, z) \mapsto x + y + z$ (Theorem 2.3.1). So $(\mathbf{u}_1, \mathbf{u}_2)$ is a basis of U (Theorem 1.6.4), and $\mathcal{B}$ is a basis of $\mathbb{R}^3$: in a relation $a\mathbf{u}_1 + b\mathbf{u}_2 + c\mathbf{e}_3 = \mathbf{0}$, the entries sum to $c = 0$, and then $a = b = 0$, so $\mathcal{B}$ is an independent list of 3 vectors (Theorem 1.6.4). Now

$$\begin{aligned} T\mathbf{u}_1 &= (-1, 0, 1) = -\mathbf{u}_1 - \mathbf{u}_2, \\ T\mathbf{u}_2 &= (1, -1, 0) = \mathbf{u}_1, \\ T\mathbf{e}_3 &= (0, 1, 0) = \mathbf{u}_2 + \mathbf{e}_3, \end{aligned}$$

so

$$[T]^{\mathcal{B}}_{\mathcal{B}} = \left(\begin{array}{cc|c} -1 & 1 & 0 \\ -1 & 0 & 1 \\ \hline 0 & 0 & 1 \end{array}\right).$$

The lower-left block is zero, as Theorem 8.1.7 predicts. The lower-right block (1) says

$\bar{T}(\mathbf{e}_3+U) = \mathbf{e}_3+U$, so $\bar{T}$ is the identity on the line $\mathbb{R}^3/U$. Indeed $T\mathbf{v}-\mathbf{v} = (y-x, z-y, x-z)$ has entries summing to 0, so $T\mathbf{v}+U = \mathbf{v}+U$ for every $\mathbf{v}$: T moves vectors, but never changes the sum of their coordinates, and that sum is all that $\mathbb{R}^3/U$ remembers.

8.1.7 Exercises

8.1.7.1 A. Check your understanding

Exercise. A1

(a) Define what it means for partitions of $\mathbf{A} \in M_{m\times n}(F)$ and $\mathbf{B} \in M_{n\times p}(F)$ to be conformable.

(b) State the block multiplication formula for $\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}\begin{pmatrix} \mathbf{A}' & \mathbf{B}' \\ \mathbf{C}' & \mathbf{D}' \end{pmatrix}$.

(c) True or false: the top-left block of that product is $\mathbf{A}'\mathbf{A} + \mathbf{C}'\mathbf{B}$. Justify your answer.

(d) True or false: if $\mathbf{A} \in M_k(F)$ and $\mathbf{B} \in M_l(F)$, then $\mathbf{A} \oplus \mathbf{B} = \mathbf{B} \oplus \mathbf{A}$. Justify your answer.

(e) True or false: if $\mathbf{M} = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{O} & \mathbf{D} \end{pmatrix}$ with $\mathbf{A}$, $\mathbf{D}$ square is invertible, then $\mathbf{M}^{-1}$ is block upper triangular. Justify your answer.

(f) In a basis whose first k vectors form a basis of U, what does a zero lower-left block of $[T]_{\mathcal{B}}^{\mathcal{B}}$ say about T?

Solution.

(a) The column partition of $\mathbf{A}$ and the row partition of $\mathbf{B}$ are the same list $n_1, \dots, n_b$, in the same order (Definition 8.1.1).

(b) It is $\begin{pmatrix} \mathbf{A}\mathbf{A}' + \mathbf{B}\mathbf{C}' & \mathbf{A}\mathbf{B}' + \mathbf{B}\mathbf{D}' \\ \mathbf{C}\mathbf{A}' + \mathbf{D}\mathbf{C}' & \mathbf{C}\mathbf{B}' + \mathbf{D}\mathbf{D}' \end{pmatrix}$, provided the partitions are conformable (Theorem 8.1.1).

(c) False in general. The block is $\mathbf{A}\mathbf{A}' + \mathbf{B}\mathbf{C}'$, with each block of the left factor on the left. For 4×4 matrices with partition $2, 2$, take $\mathbf{A} = \mathbf{E}_{12}$, $\mathbf{A}' = \mathbf{E}_{21}$ in $M_2(F)$, and all other blocks zero. Then the top-left block of the product is $\mathbf{A}\mathbf{A}' = \mathbf{E}_{11}$, while $\mathbf{A}'\mathbf{A} + \mathbf{C}'\mathbf{B} = \mathbf{E}_{22}$.

(d) False. For $k = l = 1$, $(1) \oplus (2) = \operatorname{diag}(1, 2) \neq \operatorname{diag}(2, 1) = (2) \oplus (1)$.

(e) True. By Theorem 8.1.5 (b), $\mathbf{A}$ and $\mathbf{D}$ are invertible, and (c) gives $\mathbf{M}^{-1}$ with lower-left block $\mathbf{O}$.

(f) That T maps U into U (Theorem 8.1.7 (a)).

8.1.7.2 B. Practice

Exercise. B1: Multiplying block matrices

Let

$$\mathbf{M} = \left(\begin{array}{cc|c} 1 & 2 & 1 \\ 0 & 1 & -1 \\ \hline 2 & 0 & 3 \end{array}\right), \qquad \mathbf{N} = \left(\begin{array}{cc|c} 1 & 0 & 0 \\ 1 & 1 & 2 \\ \hline 1 & -1 & 1 \end{array}\right) \in M_3(\mathbb{Q}),$$

both with row and column partition $2, 1$.

(a) Compute $\mathbf{MN}$ blockwise, and check one block against the entry formula.

(b) Explain why the partitions would **not** be conformable if $\mathbf{N}$ were given the row partition $1, 2$ instead.

Solution.

(a) Write $\mathbf{M} = \begin{pmatrix} \mathbf{A} & \mathbf{b} \\ \mathbf{c} & d \end{pmatrix}$ and $\mathbf{N} = \begin{pmatrix} \mathbf{E} & \mathbf{f} \\ \mathbf{g} & h \end{pmatrix}$ with $\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix}$, $\mathbf{b} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$, $\mathbf{c} = \begin{pmatrix} 2 & 0 \end{pmatrix}$, $d = (3)$, $\mathbf{E} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$, $\mathbf{f} = \begin{pmatrix} 0 \\ 2 \end{pmatrix}$, $\mathbf{g} = \begin{pmatrix} 1 & -1 \end{pmatrix}$, $h = (1)$. The partitions are conformable (both middle partitions are $2, 1$), so by Theorem 8.1.1

$$\mathbf{AE} + \mathbf{bg} = \begin{pmatrix} 3 & 2 \\ 1 & 1 \end{pmatrix} + \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} = \begin{pmatrix} 4 & 1 \\ 0 & 2 \end{pmatrix},$$

$$\mathbf{Af} + \mathbf{b}h = \begin{pmatrix} 4 \\ 2 \end{pmatrix} + \begin{pmatrix} 1 \\ -1 \end{pmatrix} = \begin{pmatrix} 5 \\ 1 \end{pmatrix},$$

$$\mathbf{cE} + d\mathbf{g} = \begin{pmatrix} 2 & 0 \end{pmatrix} + \begin{pmatrix} 3 & -3 \end{pmatrix} = \begin{pmatrix} 5 & -3 \end{pmatrix}, \qquad \mathbf{cf} + dh = \mathbf{O}_1 + (3) = (3).$$

Hence

$$\mathbf{MN} = \begin{pmatrix} 4 & 1 & 5 \\ 0 & 2 & 1 \\ 5 & -3 & 3 \end{pmatrix}.$$

Check the $(3, 3)$-entry directly: row 3 of $\mathbf{M}$ is $(2, 0, 3)$ and column 3 of $\mathbf{N}$ is $(0, 2, 1)$, giving $0 + 0 + 3 = 3$.

(b) The column partition of $\mathbf{M}$ is $2, 1$, and the row partition $1, 2$ is a different list. Concretely, the top-left block $\mathbf{A}$ of $\mathbf{M}$ is 2×2, while the top-left block of $\mathbf{N}$ would be 1×2, and a 2×2 matrix cannot multiply a 1×2 matrix.

Exercise. B2: Inverting a block triangular matrix

Use Theorem 8.1.5 to find the inverse of

$$\mathbf{M} = \begin{pmatrix} 2 & 1 & 1 & 0 \\ 1 & 1 & 2 & 1 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 1 \end{pmatrix} \in M_4(\mathbb{Q}),$$

and check one block of $\mathbf{MM}^{-1}$.

Solution. With the partition $2,2$, $\mathbf{M} = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{O} & \mathbf{D} \end{pmatrix}$ with $\mathbf{A} = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$, $\mathbf{B} = \begin{pmatrix} 1 & 0 \\ 2 & 1 \end{pmatrix}$, $\mathbf{D} = \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix}$. Both $\det \mathbf{A} = 1$ and $\det \mathbf{D} = 1$ are nonzero, so by Theorem 0.9.8

$$\mathbf{A}^{-1} = \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix}, \qquad \mathbf{D}^{-1} = \begin{pmatrix} 1 & -2 \\ 0 & 1 \end{pmatrix}.$$

Then $\mathbf{A}^{-1}\mathbf{B} = \begin{pmatrix} -1 & -1 \\ 3 & 2 \end{pmatrix}$ and $-\mathbf{A}^{-1}\mathbf{B}\mathbf{D}^{-1} = -\begin{pmatrix} -1 & 1 \\ 3 & -4 \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ -3 & 4 \end{pmatrix}$. By Theorem 8.1.5 (c),

$$\mathbf{M}^{-1} = \begin{pmatrix} 1 & -1 & 1 & -1 \\ -1 & 2 & -3 & 4 \\ 0 & 0 & 1 & -2 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Check the top-right block of $\mathbf{M}\mathbf{M}^{-1}$:

$$\begin{aligned} \mathbf{A}(-\mathbf{A}^{-1}\mathbf{B}\mathbf{D}^{-1}) + \mathbf{B}\mathbf{D}^{-1} &= \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} 1 & -1 \\ -3 & 4 \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ 2 & 1 \end{pmatrix}\begin{pmatrix} 1 & -2 \\ 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} -1 & 2 \\ -2 & 3 \end{pmatrix} + \begin{pmatrix} 1 & -2 \\ 2 & -3 \end{pmatrix} = \mathbf{O}, \end{aligned}$$

as it must be.

Exercise. B3: A projection in an adapted basis

Let $\mathbf{P} = \begin{pmatrix} 0 & 1 & -1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \in M_3(\mathbb{R})$, $U = \operatorname{span}((1,1,0),(0,1,1))$, and $W = \operatorname{span}(\mathbf{e}_1)$.

(a) Show that $\mathbf{P}^2 = \mathbf{P}$, that $\mathbf{P}\mathbf{u} = \mathbf{u}$ for $\mathbf{u} \in U$ and $\mathbf{P}\mathbf{e}_1 = \mathbf{0}$, and that $\mathbb{R}^3 = U \oplus W$.

(b) Let $\mathcal{B} = ((1,1,0),(0,1,1),\mathbf{e}_1)$. Use Theorem 8.1.6 to write down $[T_{\mathbf{P}}]_{\mathcal{B}}^{\mathcal{B}}$ without computing, and confirm it as $\mathbf{S}^{-1}\mathbf{P}\mathbf{S}$, where $\mathbf{S}$ has the vectors of $\mathcal{B}$ as columns.

(c) Hence find $\mathbf{P}^{100}$ and $\operatorname{tr}\mathbf{P}$.

Solution.

(a) Multiplying, $\mathbf{P}^2$ has rows $(0,1,-1)$, $(0,1,0)$, $(0,0,1)$, so $\mathbf{P}^2 = \mathbf{P}$. Also $\mathbf{P}(1,1,0) = (1,1,0)$, $\mathbf{P}(0,1,1) = (0,1,1)$, and $\mathbf{P}\mathbf{e}_1 = \mathbf{0}$ (the first column is zero), so $\mathbf{P}$ fixes every vector of U by linearity. The three vectors of $\mathcal{B}$ are the columns of $\mathbf{S} = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 0 \end{pmatrix}$, and expanding along the last column, $\det \mathbf{S} = 1 \cdot (1 - 0) = 1 \neq 0$. So $\mathbf{S}$ is invertible (Theorem 7.5.2), its columns form a basis of $\mathbb{R}^3$ (Theorem 3.4.4), and $\mathbb{R}^3 = U \oplus W$ by Theorem 1.7.5.

(b) $T_{\mathbf{P}}$ maps U into U (as the identity) and W into W (as zero). By Theorem 8.1.6, $[T_{\mathbf{P}}]_{\mathcal{B}}^{\mathcal{B}} =$

$\mathbf{I}_2 \oplus \mathbf{O}_1 = \operatorname{diag}(1,1,0)$. By Theorem 4.2.3, $[T_{\mathbf{P}}]_{\mathcal{B}}^{\mathcal{B}} = \mathbf{S}^{-1}\mathbf{P}\mathbf{S}$. Here $\mathbf{S}^{-1} = \begin{pmatrix} 0 & 1 & -1 \\ 0 & 0 & 1 \\ 1 & -1 & 1 \end{pmatrix}$ (check $\mathbf{S}\mathbf{S}^{-1} = \mathbf{I}_3$), and $\mathbf{P}\mathbf{S}$ has columns $\mathbf{P}(1,1,0) = (1,1,0)$, $\mathbf{P}(0,1,1) = (0,1,1)$, $\mathbf{P}\mathbf{e}_1 = \mathbf{0}$, so $\mathbf{S}^{-1}\mathbf{P}\mathbf{S} = \mathbf{S}^{-1}\begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 1 & 0 \end{pmatrix} = \operatorname{diag}(1,1,0)$.

(c) $\mathbf{P} = \mathbf{S}\operatorname{diag}(1,1,0)\mathbf{S}^{-1}$, so $\mathbf{P}^{100} = \mathbf{S}\operatorname{diag}(1,1,0)^{100}\mathbf{S}^{-1} = \mathbf{S}\operatorname{diag}(1,1,0)\mathbf{S}^{-1} = \mathbf{P}$ by Theorem 8.1.4 (b) (or directly from $\mathbf{P}^2 = \mathbf{P}$). By Theorem 4.2.5, $\operatorname{tr}\mathbf{P} = \operatorname{tr}\operatorname{diag}(1,1,0) = 2$, which matches $0+1+1$.

8.1.7.3 C. Going deeper

Exercise. C1: Square-zero matrices of half rank

Let $n \geq 1$.

(a) For $\mathbf{X} \in M_n(F)$, let $\mathbf{N}_{\mathbf{X}} = \begin{pmatrix} \mathbf{O} & \mathbf{X} \\ \mathbf{O} & \mathbf{O} \end{pmatrix} \in M_{2n}(F)$. Show that $\mathbf{N}_{\mathbf{X}}^2 = \mathbf{O}$ and that $\operatorname{rank}\mathbf{N}_{\mathbf{X}} = \operatorname{rank}\mathbf{X}$.

(b) Let V have dimension $2n$, and let $T \in \mathcal{L}(V)$ satisfy $T^2 = 0$ and $\operatorname{rank}T = n$. Show that $\operatorname{im}T = \ker T$. Choose $w_1, \ldots, w_n$ with $(Tw_1, \ldots, Tw_n)$ a basis of $\operatorname{im}T$, and put $u_i = Tw_i$. Prove that $\mathcal{B} = (u_1, \ldots, u_n, w_1, \ldots, w_n)$ is a basis of V and that $[T]_{\mathcal{B}}^{\mathcal{B}} = \mathbf{N}_{\mathbf{I}_n}$.

(c) Deduce that for every invertible $\mathbf{X} \in M_n(F)$, the matrix $\mathbf{N}_{\mathbf{X}}$ is similar to $\mathbf{N}_{\mathbf{I}_n}$.

Hint: for the independence in (b), apply T.

Solution.

(a) By Theorem 8.1.1, $\mathbf{N}_{\mathbf{X}}^2 = \begin{pmatrix} \mathbf{O}\cdot\mathbf{O} + \mathbf{X}\cdot\mathbf{O} & \mathbf{O}\cdot\mathbf{X} + \mathbf{X}\cdot\mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{pmatrix} = \mathbf{O}$. The first n columns of $\mathbf{N}_{\mathbf{X}}$ are zero, and column $n+j$ is $\begin{pmatrix} \mathbf{x}_j \\ \mathbf{0} \end{pmatrix}$, where $\mathbf{x}_j$ is column j of $\mathbf{X}$. A combination $\sum_j c_j \begin{pmatrix} \mathbf{x}_j \\ \mathbf{0} \end{pmatrix}$ is zero exactly when $\sum_j c_j \mathbf{x}_j = \mathbf{0}$, so a list of these columns is independent exactly when the corresponding columns of $\mathbf{X}$ are. Hence $\dim\operatorname{col}(\mathbf{N}_{\mathbf{X}}) = \dim\operatorname{col}(\mathbf{X})$, that is, $\operatorname{rank}\mathbf{N}_{\mathbf{X}} = \operatorname{rank}\mathbf{X}$.

(b) If $v = Tx \in \operatorname{im}T$, then $Tv = T^2x = 0$, so $\operatorname{im}T \subseteq \ker T$. By Theorem 2.3.1, $\dim\ker T = 2n - n = n = \dim\operatorname{im}T$, so $\operatorname{im}T = \ker T$ by Theorem 1.6.8. The vectors w_i exist because $(Tw_1, \ldots, Tw_n)$ can be any basis of $\operatorname{im}T$, each of whose vectors is T of something.
Let $\sum_i a_i u_i + \sum_i b_i w_i = 0$. Applying T, and using $Tu_i = T^2 w_i = 0$, gives $\sum_i b_i u_i = 0$. The u_i form a basis of $\operatorname{im}T$, so all $b_i = 0$, and then $\sum_i a_i u_i = 0$ gives all $a_i = 0$. So $\mathcal{B}$ is a linearly independent list of $2n = \dim V$ vectors, hence a basis (Theorem 1.6.4). Since $Tu_i = 0$ and $Tw_i = u_i$, column $i \leq n$ of $[T]_{\mathcal{B}}^{\mathcal{B}}$ is zero and column $n+i$ is $\mathbf{e}_i$. So $[T]_{\mathcal{B}}^{\mathcal{B}} = \begin{pmatrix} \mathbf{O} & \mathbf{I}_n \\ \mathbf{O} & \mathbf{O} \end{pmatrix} = \mathbf{N}_{\mathbf{I}_n}$.

(c) If $\mathbf{X}$ is invertible, then by (a) the operator $T = T_{\mathbf{N}_{\mathbf{X}}}$ on F^{2n} satisfies $T^2 = 0$ and $\operatorname{rank}T = \operatorname{rank}\mathbf{N}_{\mathbf{X}} = \operatorname{rank}\mathbf{X} = n$ (Theorem 4.1.5, and Theorem 3.4.4 for $\operatorname{rank}\mathbf{X} = n$).

By (b) there is a basis $\mathcal{B}$ with $[T]_{\mathcal{B}}^{\mathcal{B}} = \mathbf{N}_{\mathbf{I}_n}$. The matrix of T in the standard basis is $\mathbf{N}_{\mathbf{X}}$, so $\mathbf{N}_{\mathbf{X}}$ and $\mathbf{N}_{\mathbf{I}_n}$ are similar by Theorem 4.2.4.

Exercise. C2: Powers of block triangular matrices

Let $\mathbf{A} \in M_k(F)$, $\mathbf{D} \in M_l(F)$, and $\mathbf{B} \in M_{k\times l}(F)$, and let $\mathbf{M} = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{O} & \mathbf{D} \end{pmatrix}$.

(a) Prove that for every $j \geq 1$,

$$\mathbf{M}^j = \begin{pmatrix} \mathbf{A}^j & \mathbf{S}_j \\ \mathbf{O} & \mathbf{D}^j \end{pmatrix}, \qquad \mathbf{S}_j = \sum_{i=0}^{j-1} \mathbf{A}^i \mathbf{B} \mathbf{D}^{j-1-i}.$$

(b) Deduce that if $\mathbf{D} = \mathbf{O}$, then $\mathbf{M}^j = \begin{pmatrix} \mathbf{A}^j & \mathbf{A}^{j-1}\mathbf{B} \\ \mathbf{O} & \mathbf{O} \end{pmatrix}$ for $j \geq 1$, and that if moreover $\mathbf{A}^r = \mathbf{O}$, then $\mathbf{M}^{r+1} = \mathbf{O}$.

(c) For $\mathbf{A} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$, $\mathbf{B} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$, and $\mathbf{D} = \mathbf{O}_1$, find the least j with $\mathbf{M}^j = \mathbf{O}$.

Solution.

(a) By induction on j. For $j = 1$, $\mathbf{S}_1 = \mathbf{A}^0\mathbf{B}\mathbf{D}^0 = \mathbf{B}$. Suppose the formula holds for j. By Theorem 8.1.5 (a),

$$\mathbf{M}^{j+1} = \mathbf{M}^j\mathbf{M} = \begin{pmatrix} \mathbf{A}^j\mathbf{A} & \mathbf{A}^j\mathbf{B} + \mathbf{S}_j\mathbf{D} \\ \mathbf{O} & \mathbf{D}^j\mathbf{D} \end{pmatrix},$$

and $\mathbf{A}^j\mathbf{B} + \mathbf{S}_j\mathbf{D} = \mathbf{A}^j\mathbf{B}\mathbf{D}^0 + \sum_{i=0}^{j-1} \mathbf{A}^i\mathbf{B}\mathbf{D}^{j-i} = \sum_{i=0}^{j} \mathbf{A}^i\mathbf{B}\mathbf{D}^{j-i} = \mathbf{S}_{j+1}$. This completes the induction.

(b) If $\mathbf{D} = \mathbf{O}$, every term of $\mathbf{S}_j$ with $j - 1 - i \geq 1$ contains a factor $\mathbf{D} = \mathbf{O}$, and only $i = j - 1$ survives, with $\mathbf{D}^0 = \mathbf{I}_l$. So $\mathbf{S}_j = \mathbf{A}^{j-1}\mathbf{B}$ and $\mathbf{D}^j = \mathbf{O}$. If also $\mathbf{A}^r = \mathbf{O}$, then $\mathbf{M}^{r+1}$ has blocks $\mathbf{A}^{r+1} = \mathbf{O}$ and $\mathbf{A}^r\mathbf{B} = \mathbf{O}$, so $\mathbf{M}^{r+1} = \mathbf{O}$.

(c) Here $\mathbf{A}^2 = \mathbf{O}$, so $\mathbf{M}^3 = \mathbf{O}$ by (b). But $\mathbf{M}^2 = \begin{pmatrix} \mathbf{A}^2 & \mathbf{A}\mathbf{B} \\ \mathbf{O} & \mathbf{O} \end{pmatrix}$ with $\mathbf{A}\mathbf{B} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \neq \mathbf{O}$, so $\mathbf{M}^2 \neq \mathbf{O}$; and $\mathbf{M} \neq \mathbf{O}$. The least such j is 3. Explicitly, $\mathbf{M} = \begin{pmatrix} 0 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$ and $\mathbf{M}^2 = \mathbf{E}_{13}$.

8.2 Rank and Block Matrices

Rank is the most robust number attached to a matrix, and Chapters 3 and 4 found several inequalities for it: the rank of a product is at most the rank of each factor, and Sylvester's inequality bounds it from below. Those proofs worked with subspaces. This section proves rank inequalities by a different and very mechanical method. We place the matrices we care about as blocks of one bigger matrix, and then change that matrix by block row and column operations,

which never change the rank, until its rank can be read off. The method gives short proofs of the Frobenius inequality and of identities such as $\operatorname{rank}(\mathbf{I}_m - \mathbf{AB}) + n = \operatorname{rank}(\mathbf{I}_n - \mathbf{BA}) + m$, which are awkward to reach with subspaces alone.

8.2.1 Three basic inequalities

We start with what the shape of a block matrix alone says about its rank. Write $\begin{pmatrix}\mathbf{A} & \mathbf{B}\end{pmatrix}$ for the matrix obtained by placing $\mathbf{B}$ to the right of $\mathbf{A}$ (same number of rows), and $\begin{pmatrix}\mathbf{A}\\ \mathbf{C}\end{pmatrix}$ for $\mathbf{C}$ below $\mathbf{A}$ (same number of columns). In these results the blocks need **not** be square.

Theorem 8.2.1. Rank Inequalities for Block Matrices

Let $\mathbf{A} \in M_{m\times k}(F)$, $\mathbf{B} \in M_{m\times l}(F)$, $\mathbf{C} \in M_{p\times k}(F)$, and $\mathbf{D} \in M_{p\times l}(F)$.

(a) $\max(\operatorname{rank}\mathbf{A}, \operatorname{rank}\mathbf{B}) \le \operatorname{rank}\begin{pmatrix}\mathbf{A} & \mathbf{B}\end{pmatrix} \le \operatorname{rank}\mathbf{A} + \operatorname{rank}\mathbf{B}$, and $\max(\operatorname{rank}\mathbf{A}, \operatorname{rank}\mathbf{C}) \le \operatorname{rank}\begin{pmatrix}\mathbf{A}\\ \mathbf{C}\end{pmatrix} \le \operatorname{rank}\mathbf{A} + \operatorname{rank}\mathbf{C}$.

(b) $\operatorname{rank}\begin{pmatrix}\mathbf{A} & \mathbf{O}\\ \mathbf{O} & \mathbf{D}\end{pmatrix} = \operatorname{rank}\mathbf{A} + \operatorname{rank}\mathbf{D}$.

(c) $\operatorname{rank}\begin{pmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{O} & \mathbf{D}\end{pmatrix} \ge \operatorname{rank}\mathbf{A} + \operatorname{rank}\mathbf{D}$ and $\operatorname{rank}\begin{pmatrix}\mathbf{A} & \mathbf{O}\\ \mathbf{C} & \mathbf{D}\end{pmatrix} \ge \operatorname{rank}\mathbf{A} + \operatorname{rank}\mathbf{D}$.

Idea. Rank is the dimension of the column space, so look at columns. In (a) the columns of $\begin{pmatrix}\mathbf{A} & \mathbf{B}\end{pmatrix}$ are those of $\mathbf{A}$ together with those of $\mathbf{B}$, and dimensions of sums are at most sums of dimensions. In (b) the columns coming from $\mathbf{A}$ and from $\mathbf{D}$ live in two subspaces that meet only in $\mathbf{0}$. In (c) we cannot use the column space of $\mathbf{B}$, which is unknown, so instead we pick independent columns of $\mathbf{A}$ and of $\mathbf{D}$ and show that the columns of the big matrix standing above them are still independent: the zero block lets us peel off the $\mathbf{D}$-part first.

Proof. For $\mathbf{y} \in F^m$ and $\mathbf{z} \in F^p$, write $\begin{pmatrix}\mathbf{y}\\ \mathbf{z}\end{pmatrix} \in F^{m+p}$ for the stacked vector.

(a) The columns of $\begin{pmatrix}\mathbf{A} & \mathbf{B}\end{pmatrix}$ are the columns of $\mathbf{A}$ followed by the columns of $\mathbf{B}$, so its column space is $\operatorname{col}(\mathbf{A}) + \operatorname{col}(\mathbf{B})$ by Proposition 1.7.2. It contains $\operatorname{col}(\mathbf{A})$ and $\operatorname{col}(\mathbf{B})$, which gives the lower bound by Theorem 1.6.7, and by Theorem 1.7.3

$$\operatorname{rank}\begin{pmatrix}\mathbf{A} & \mathbf{B}\end{pmatrix} = \operatorname{rank}\mathbf{A} + \operatorname{rank}\mathbf{B} - \dim(\operatorname{col}\mathbf{A} \cap \operatorname{col}\mathbf{B}) \le \operatorname{rank}\mathbf{A} + \operatorname{rank}\mathbf{B}.$$

For the stacked matrix, Proposition 8.1.2 gives $\begin{pmatrix}\mathbf{A}\\ \mathbf{C}\end{pmatrix}^{\top} = \begin{pmatrix}\mathbf{A}^{\top} & \mathbf{C}^{\top}\end{pmatrix}$, and a matrix and its transpose have the same rank (Theorem 3.5.5). Apply the first part to $\mathbf{A}^{\top}$ and $\mathbf{C}^{\top}$.

(b) Let $\mathbf{M} = \begin{pmatrix}\mathbf{A} & \mathbf{O}\\ \mathbf{O} & \mathbf{D}\end{pmatrix}$. Its first k columns are $\begin{pmatrix}\mathbf{a}_j\\ \mathbf{0}\end{pmatrix}$ and its last l columns are $\begin{pmatrix}\mathbf{0}\\ \mathbf{d}_q\end{pmatrix}$, where

$\mathbf{a}_j$, $\mathbf{d}_q$ are the columns of $\mathbf{A}$, $\mathbf{D}$. So $\operatorname{col}(\mathbf{M}) = Y + Z$ with $Y = \{\begin{pmatrix}\mathbf{y}\\ \mathbf{0}\end{pmatrix} : \mathbf{y} \in \operatorname{col}\mathbf{A}\}$ and $Z = \{\begin{pmatrix}\mathbf{0}\\ \mathbf{z}\end{pmatrix} : \mathbf{z} \in \operatorname{col}\mathbf{D}\}$. A vector in $Y \cap Z$ has both parts zero, so $Y \cap Z = \{\mathbf{0}\}$. The map $\mathbf{y} \mapsto \begin{pmatrix}\mathbf{y}\\ \mathbf{0}\end{pmatrix}$ is linear and injective from $\operatorname{col}\mathbf{A}$ onto Y, so $\dim Y = \operatorname{rank}\mathbf{A}$ (Theorem 2.5.4); likewise $\dim Z = \operatorname{rank}\mathbf{D}$. By Theorem 1.7.3, $\operatorname{rank}\mathbf{M} = \operatorname{rank}\mathbf{A} + \operatorname{rank}\mathbf{D} - 0$.

(c) Let $\mathbf{M} = \begin{pmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{O} & \mathbf{D}\end{pmatrix}$, $r = \operatorname{rank}\mathbf{A}$, and $s = \operatorname{rank}\mathbf{D}$. By Theorem 3.5.2 there are columns $\mathbf{a}_{j_1}, \dots, \mathbf{a}_{j_r}$ of $\mathbf{A}$ that are linearly independent, and columns $\mathbf{d}_{q_1}, \dots, \mathbf{d}_{q_s}$ of $\mathbf{D}$ that are linearly independent. The corresponding columns of $\mathbf{M}$ are $\begin{pmatrix}\mathbf{a}_{j_i}\\ \mathbf{0}\end{pmatrix}$ and $\begin{pmatrix}\mathbf{b}_{q_i}\\ \mathbf{d}_{q_i}\end{pmatrix}$, where $\mathbf{b}_q$ is column q of $\mathbf{B}$. Let

$$\sum_{i=1}^{r} \alpha_i \begin{pmatrix}\mathbf{a}_{j_i}\\ \mathbf{0}\end{pmatrix} + \sum_{i=1}^{s} \beta_i \begin{pmatrix}\mathbf{b}_{q_i}\\ \mathbf{d}_{q_i}\end{pmatrix} = \mathbf{0}.$$

The last p entries give $\sum_i \beta_i \mathbf{d}_{q_i} = \mathbf{0}$, so all $\beta_i = 0$. The first m entries then give $\sum_i \alpha_i \mathbf{a}_{j_i} = \mathbf{0}$, so all $\alpha_i = 0$. Hence $\operatorname{col}(\mathbf{M})$ contains $r + s$ linearly independent vectors, and $\operatorname{rank}\mathbf{M} \geq r + s$ by Theorem 1.6.3. For the lower triangular shape, $\begin{pmatrix}\mathbf{A} & \mathbf{O}\\ \mathbf{C} & \mathbf{D}\end{pmatrix}^{\top} = \begin{pmatrix}\mathbf{A}^{\top} & \mathbf{C}^{\top}\\ \mathbf{O} & \mathbf{D}^{\top}\end{pmatrix}$ by Proposition 8.1.2; apply the first case and Theorem 3.5.5.

■

All three inequalities, the two in (a) and the one in (c), can be strict, and small witnesses show it. For $\mathbf{A} = \mathbf{B} = (1)$, $\operatorname{rank}\begin{pmatrix}1 & 1\end{pmatrix} = 1 < 2$: the two column spaces overlap. For $\mathbf{A} = \mathbf{e}_1$ and $\mathbf{B} = \mathbf{e}_2$ in F^2, $\operatorname{rank}\begin{pmatrix}\mathbf{e}_1 & \mathbf{e}_2\end{pmatrix} = \operatorname{rank}\mathbf{I}_2 = 2 > \max(1,1)$: each block adds something the other lacks. For $\mathbf{A} = \mathbf{D} = \mathbf{O}_1$ and $\mathbf{B} = (1)$, $\operatorname{rank}\begin{pmatrix}0 & 1\\ 0 & 0\end{pmatrix} = 1 > 0 + 0$: the off-diagonal block contributes rank of its own. The proof of (a) says exactly when the upper bound is attained: $\operatorname{rank}\begin{pmatrix}\mathbf{A} & \mathbf{B}\end{pmatrix} = \operatorname{rank}\mathbf{A} + \operatorname{rank}\mathbf{B}$ if and only if $\operatorname{col}\mathbf{A} \cap \operatorname{col}\mathbf{B} = \{\mathbf{0}\}$.

Warning. **The ranks of the four blocks do not determine the rank of the whole.** Over $\mathbb{Q}$, $\begin{pmatrix}1 & 1\\ 1 & 1\end{pmatrix}$ and $\begin{pmatrix}1 & 1\\ 1 & -1\end{pmatrix}$ have all four 1×1 blocks of rank 1, but ranks 1 and 2. Only special shapes, like the zero blocks in (b) and (c), turn block ranks into information about the whole matrix.

Quick check. Without computing, find $\operatorname{rank}\begin{pmatrix}1 & 2 & 0 & 0 & 0\\ 2 & 4 & 0 & 0 & 0\\ 0 & 0 & 1 & 0 & 3\\ 0 & 0 & 0 & 1 & 5\end{pmatrix}$.

Solution. With row partition $2,2$ and column partition $2,3$, the matrix is $\begin{pmatrix} \mathbf{A} & \mathbf{O} \\ \mathbf{O} & \mathbf{D} \end{pmatrix}$ with $\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix}$ of rank 1 (the second row is twice the first) and $\mathbf{D} = \begin{pmatrix} 1 & 0 & 3 \\ 0 & 1 & 5 \end{pmatrix}$ of rank 2 (the first two columns are $\mathbf{e}_1, \mathbf{e}_2$). By Theorem 8.2.1 (b), the rank is $1 + 2 = 3$.

8.2.2 Block elimination

Gaussian elimination changes a matrix by row operations, which are multiplications by invertible matrices on the left, and the rank never changes. The same is true when the "rows" are whole block rows and the multipliers are matrices. The matrices that do this are the block triangular matrices with identity blocks on the diagonal from Theorem 8.1.5, together with block swaps.

Fix a partition of $\mathbf{M} = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix} \in M_{(m+p)\times(k+l)}(F)$ with $\mathbf{A} \in M_{m\times k}(F)$. For matrices $\mathbf{X} \in M_{m\times p}(F)$ and $\mathbf{Y} \in M_{k\times l}(F)$, Theorem 8.1.1 gives

$$\begin{pmatrix} \mathbf{I}_m & \mathbf{X} \\ \mathbf{O} & \mathbf{I}_p \end{pmatrix} \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix} = \begin{pmatrix} \mathbf{A} + \mathbf{X}\mathbf{C} & \mathbf{B} + \mathbf{X}\mathbf{D} \\ \mathbf{C} & \mathbf{D} \end{pmatrix},$$
$$\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix} \begin{pmatrix} \mathbf{I}_k & \mathbf{Y} \\ \mathbf{O} & \mathbf{I}_l \end{pmatrix} = \begin{pmatrix} \mathbf{A} & \mathbf{A}\mathbf{Y} + \mathbf{B} \\ \mathbf{C} & \mathbf{C}\mathbf{Y} + \mathbf{D} \end{pmatrix}.$$

In words: multiplying on the **left** adds $\mathbf{X}$ times block row 2 to block row 1, with $\mathbf{X}$ on the left of the row; multiplying on the **right** adds block column 1 times $\mathbf{Y}$ to block column 2, with $\mathbf{Y}$ on the right of the column. The lower triangular multipliers $\begin{pmatrix} \mathbf{I}_m & \mathbf{O} \\ \mathbf{X}' & \mathbf{I}_p \end{pmatrix}$ and $\begin{pmatrix} \mathbf{I}_k & \mathbf{O} \\ \mathbf{Y}' & \mathbf{I}_l \end{pmatrix}$ act the same way in the other direction. Sides matter: the multiplier sits on the side of the product where the matrix operation happens.

Theorem 8.2.2. Block Elimination Preserves Rank

Let $\mathbf{M} = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}$ be as above.

(a) For all $\mathbf{X} \in M_{m\times p}(F)$, $\mathbf{X}' \in M_{p\times m}(F)$, $\mathbf{Y} \in M_{k\times l}(F)$, and $\mathbf{Y}' \in M_{l\times k}(F)$, the matrices

$$\begin{pmatrix} \mathbf{I}_m & \mathbf{X} \\ \mathbf{O} & \mathbf{I}_p \end{pmatrix}, \quad \begin{pmatrix} \mathbf{I}_m & \mathbf{O} \\ \mathbf{X}' & \mathbf{I}_p \end{pmatrix}, \quad \begin{pmatrix} \mathbf{I}_k & \mathbf{Y} \\ \mathbf{O} & \mathbf{I}_l \end{pmatrix}, \quad \begin{pmatrix} \mathbf{I}_k & \mathbf{O} \\ \mathbf{Y}' & \mathbf{I}_l \end{pmatrix}$$

are invertible, and so are the block swaps $\begin{pmatrix} \mathbf{O} & \mathbf{I}_p \\ \mathbf{I}_m & \mathbf{O} \end{pmatrix} \in M_{m+p}(F)$ and $\begin{pmatrix} \mathbf{O} & \mathbf{I}_k \\ \mathbf{I}_l & \mathbf{O} \end{pmatrix} \in M_{k+l}(F)$. Hence multiplying $\mathbf{M}$ on the left or right by any product of such matrices does not change its rank. The swaps give $\begin{pmatrix} \mathbf{O} & \mathbf{I}_p \\ \mathbf{I}_m & \mathbf{O} \end{pmatrix} \mathbf{M} = \begin{pmatrix} \mathbf{C} & \mathbf{D} \\ \mathbf{A} & \mathbf{B} \end{pmatrix}$ and $\mathbf{M} \begin{pmatrix} \mathbf{O} & \mathbf{I}_k \\ \mathbf{I}_l & \mathbf{O} \end{pmatrix} = \begin{pmatrix} \mathbf{B} & \mathbf{A} \\ \mathbf{D} & \mathbf{C} \end{pmatrix}$.

(b) If $\operatorname{col}(\mathbf{B}) \subseteq \operatorname{col}(\mathbf{A})$, or if $\operatorname{row}(\mathbf{B}) \subseteq \operatorname{row}(\mathbf{D})$, then $\operatorname{rank}\begin{pmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{O} & \mathbf{D}\end{pmatrix} = \operatorname{rank}\mathbf{A} + \operatorname{rank}\mathbf{D}$.

Proof.

(a) By Theorem 8.1.5 (c), with identity blocks on the diagonal, $\begin{pmatrix}\mathbf{I}_m & \mathbf{X}\\ \mathbf{O} & \mathbf{I}_p\end{pmatrix}$ has inverse $\begin{pmatrix}\mathbf{I}_m & -\mathbf{X}\\ \mathbf{O} & \mathbf{I}_p\end{pmatrix}$; the lower triangular case follows by transposing (Proposition 8.1.2, Theorem 0.9.7). In the swap $\mathbf{S} = \begin{pmatrix}\mathbf{O} & \mathbf{I}_p\\ \mathbf{I}_m & \mathbf{O}\end{pmatrix}$ the row partition is p, m and the column partition is m, p. Let $\mathbf{S}' = \begin{pmatrix}\mathbf{O} & \mathbf{I}_m\\ \mathbf{I}_p & \mathbf{O}\end{pmatrix}$, with row partition m, p and column partition p, m. These are conformable, and Theorem 8.1.1 gives $\mathbf{S}'\mathbf{S} = \begin{pmatrix}\mathbf{I}_m & \mathbf{O}\\ \mathbf{O} & \mathbf{I}_p\end{pmatrix} = \mathbf{I}_{m+p}$, so $\mathbf{S}$ is invertible by Theorem 3.4.5; the other swap is the same with k, l in place of m, p. The two displayed products are Theorem 8.1.1. Rank is unchanged by invertible factors on either side (Theorem 3.5.9).

(b) Suppose $\operatorname{col}(\mathbf{B}) \subseteq \operatorname{col}(\mathbf{A})$. Each column $\mathbf{b}_q$ of $\mathbf{B}$ is then $\mathbf{A}\mathbf{y}_q$ for some $\mathbf{y}_q \in F^k$ (Definition 3.5.1, Theorem 0.9.3). Let $\mathbf{Y} \in M_{k\times l}(F)$ have columns $\mathbf{y}_q$, so that $\mathbf{AY} = \mathbf{B}$ by Theorem 0.9.2. By the computation before the theorem,

$$\begin{pmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{O} & \mathbf{D}\end{pmatrix}\begin{pmatrix}\mathbf{I}_k & -\mathbf{Y}\\ \mathbf{O} & \mathbf{I}_l\end{pmatrix} = \begin{pmatrix}\mathbf{A} & -\mathbf{AY}+\mathbf{B}\\ \mathbf{O} & \mathbf{D}\end{pmatrix} = \begin{pmatrix}\mathbf{A} & \mathbf{O}\\ \mathbf{O} & \mathbf{D}\end{pmatrix},$$

and (a) together with Theorem 8.2.1 (b) gives the claim. If $\operatorname{row}(\mathbf{B}) \subseteq \operatorname{row}(\mathbf{D})$, the same argument applied to rows gives $\mathbf{B} = \mathbf{XD}$ for some $\mathbf{X} \in M_{m\times p}(F)$, and $\begin{pmatrix}\mathbf{I}_m & -\mathbf{X}\\ \mathbf{O} & \mathbf{I}_p\end{pmatrix}\begin{pmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{O} & \mathbf{D}\end{pmatrix} = \begin{pmatrix}\mathbf{A} & \mathbf{O}\\ \mathbf{O} & \mathbf{D}\end{pmatrix}$.

■

Part (b) is the typical use of (a): a block that is a multiple of a neighboring block can be cleared, exactly as an entry below a pivot is cleared in Gaussian elimination. The zero block of Theorem 8.2.1 (c) is what lets the cleared matrix fall apart into a direct sum.

8.2.3 Frobenius's inequality by blocks

Recall Sylvester's rank inequality (Theorem 4.3.5): for $\mathbf{A} \in M_{m\times n}(F)$ and $\mathbf{B} \in M_{n\times p}(F)$, $\operatorname{rank}(\mathbf{AB}) \geq \operatorname{rank}\mathbf{A} + \operatorname{rank}\mathbf{B} - n$. In Exercise (C1: Frobenius's rank inequality) we met a stronger statement, Frobenius's inequality, and proved it by comparing intersections of subspaces with a null space. Here is a proof in which nothing is chosen: we write down one block matrix whose rank we know, and eliminate.

Theorem 8.2.3. Frobenius's Rank Inequality

Let $\mathbf{A} \in M_{m\times n}(F)$, $\mathbf{B} \in M_{n\times p}(F)$, and $\mathbf{C} \in M_{p\times q}(F)$. Then

$$\operatorname{rank}(\mathbf{ABC}) + \operatorname{rank}\mathbf{B} \geq \operatorname{rank}(\mathbf{AB}) + \operatorname{rank}(\mathbf{BC}).$$

Idea. The left side is the rank of $\begin{pmatrix} \mathbf{ABC} & \mathbf{O} \\ \mathbf{O} & \mathbf{B} \end{pmatrix}$, by Theorem 8.2.1 (b). The right side is a lower bound for any block triangular matrix with diagonal blocks $\mathbf{AB}$ and $\mathbf{BC}$, by part (c). So we want to turn the first matrix into the second by block elimination. Found backwards: $\mathbf{A}$ times the second block row puts $\mathbf{AB}$ into the top-right corner; then the second block column times $\mathbf{C}$ cancels $\mathbf{ABC}$ and puts $\mathbf{BC}$ into the bottom-left corner. A swap of the two block columns finishes the job.

Proof. Let $\mathbf{P} = \begin{pmatrix} \mathbf{ABC} & \mathbf{O} \\ \mathbf{O} & \mathbf{B} \end{pmatrix}$, with row partition m, n and column partition q, p. By Theorem 8.2.1 (b), $\operatorname{rank}\mathbf{P} = \operatorname{rank}(\mathbf{ABC}) + \operatorname{rank}\mathbf{B}$. By Theorem 8.1.1,

$$\begin{pmatrix} \mathbf{I}_m & \mathbf{A} \\ \mathbf{O} & \mathbf{I}_n \end{pmatrix} \mathbf{P} = \begin{pmatrix} \mathbf{ABC} & \mathbf{AB} \\ \mathbf{O} & \mathbf{B} \end{pmatrix},$$

$$\begin{aligned} \begin{pmatrix} \mathbf{ABC} & \mathbf{AB} \\ \mathbf{O} & \mathbf{B} \end{pmatrix} \begin{pmatrix} \mathbf{I}_q & \mathbf{O} \\ -\mathbf{C} & \mathbf{I}_p \end{pmatrix} &= \begin{pmatrix} \mathbf{ABC} - \mathbf{ABC} & \mathbf{AB} \\ -\mathbf{BC} & \mathbf{B} \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{O} & \mathbf{AB} \\ -\mathbf{BC} & \mathbf{B} \end{pmatrix}. \end{aligned}$$

Multiplying on the right by the block swap $\begin{pmatrix} \mathbf{O} & \mathbf{I}_q \\ \mathbf{I}_p & \mathbf{O} \end{pmatrix}$ exchanges the two block columns and gives $\mathbf{Q} = \begin{pmatrix} \mathbf{AB} & \mathbf{O} \\ \mathbf{B} & -\mathbf{BC} \end{pmatrix}$. By Theorem 8.2.2 (a), $\operatorname{rank}\mathbf{Q} = \operatorname{rank}\mathbf{P}$. By Theorem 8.2.1 (c), $\operatorname{rank}\mathbf{Q} \geq \operatorname{rank}(\mathbf{AB}) + \operatorname{rank}(-\mathbf{BC}) = \operatorname{rank}(\mathbf{AB}) + \operatorname{rank}(\mathbf{BC})$, where $\operatorname{rank}(-\mathbf{BC}) = \operatorname{rank}(\mathbf{BC})$ because $-\mathbf{BC}$ and $\mathbf{BC}$ have the same column space. Hence $\operatorname{rank}(\mathbf{ABC}) + \operatorname{rank}\mathbf{B} \geq \operatorname{rank}(\mathbf{AB}) + \operatorname{rank}(\mathbf{BC})$.

■

Taking $\mathbf{B} = \mathbf{I}_n$ gives $\operatorname{rank}(\mathbf{AC}) + n \geq \operatorname{rank}\mathbf{A} + \operatorname{rank}\mathbf{C}$ for $\mathbf{A} \in M_{m\times n}(F)$ and $\mathbf{C} \in M_{n\times q}(F)$, which is Sylvester's inequality again. Taking $\mathbf{A} = \mathbf{I}_n$ or $\mathbf{C} = \mathbf{I}_p$ gives an equality and no information, as it should. The inequality says that multiplying by $\mathbf{A}$ on the left loses at least as much rank from $\mathbf{BC}$ as from $\mathbf{B}$: the rank lost is $\operatorname{rank}(\mathbf{BC}) - \operatorname{rank}(\mathbf{ABC}) \leq \operatorname{rank}\mathbf{B} - \operatorname{rank}(\mathbf{AB})$. A smaller column space has less to lose.

8.2.4 A worked example: $\mathbf{I} - \mathbf{AB}$ and $\mathbf{I} - \mathbf{BA}$

Products in the two orders, $\mathbf{AB}$ and $\mathbf{BA}$, are different matrices of possibly different sizes, and Chapter 4 showed that even their ranks can differ. Yet $\mathbf{I} - \mathbf{AB}$ and $\mathbf{I} - \mathbf{BA}$ are tightly linked. The block matrix that links them contains $\mathbf{I}_m$, $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{I}_n$, and each identity block can be used as a pivot to clear its neighbors.

Example. Rank of $\mathbf{I}-\mathbf{AB}$ versus $\mathbf{I}-\mathbf{BA}$

Let $\mathbf{A}\in M_{m\times n}(F)$ and $\mathbf{B}\in M_{n\times m}(F)$.

(a) Prove that $\operatorname{rank}(\mathbf{I}_m-\mathbf{AB})+n=\operatorname{rank}(\mathbf{I}_n-\mathbf{BA})+m$.

(b) Deduce that for $m=n$, $\operatorname{rank}(\mathbf{I}_n-\mathbf{AB})=\operatorname{rank}(\mathbf{I}_n-\mathbf{BA})$, and $\mathbf{I}_n-\mathbf{AB}$ is invertible if and only if $\mathbf{I}_n-\mathbf{BA}$ is.

(c) Check (a) for $\mathbf{A}=\begin{pmatrix}1&1&0\\0&1&1\end{pmatrix}$ and $\mathbf{B}=\begin{pmatrix}1&0\\0&0\\0&1\end{pmatrix}$ over $\mathbb{Q}$.

Solution.

(a) Let $\mathbf{M}=\begin{pmatrix}\mathbf{I}_m&\mathbf{A}\\\mathbf{B}&\mathbf{I}_n\end{pmatrix}\in M_{m+n}(F)$, with partition m,n for rows and columns. We eliminate in two different ways.

Using $\mathbf{I}_m$ as the pivot. Subtract $\mathbf{B}$ times block row 1 from block row 2, then subtract block column 1 times $\mathbf{A}$ from block column 2. By Theorem 8.1.1,

$$\begin{pmatrix}\mathbf{I}_m&\mathbf{O}\\-\mathbf{B}&\mathbf{I}_n\end{pmatrix}\mathbf{M}=\begin{pmatrix}\mathbf{I}_m&\mathbf{A}\\\mathbf{O}&\mathbf{I}_n-\mathbf{BA}\end{pmatrix},$$

$$\begin{pmatrix}\mathbf{I}_m&\mathbf{A}\\\mathbf{O}&\mathbf{I}_n-\mathbf{BA}\end{pmatrix}\begin{pmatrix}\mathbf{I}_m&-\mathbf{A}\\\mathbf{O}&\mathbf{I}_n\end{pmatrix}=\begin{pmatrix}\mathbf{I}_m&\mathbf{O}\\\mathbf{O}&\mathbf{I}_n-\mathbf{BA}\end{pmatrix}.$$

By Theorem 8.2.2 (a) and Theorem 8.2.1 (b), $\operatorname{rank}\mathbf{M}=m+\operatorname{rank}(\mathbf{I}_n-\mathbf{BA})$.

Using $\mathbf{I}_n$ as the pivot. Subtract $\mathbf{A}$ times block row 2 from block row 1, then subtract block column 2 times $\mathbf{B}$ from block column 1:

$$\begin{pmatrix}\mathbf{I}_m&-\mathbf{A}\\\mathbf{O}&\mathbf{I}_n\end{pmatrix}\mathbf{M}=\begin{pmatrix}\mathbf{I}_m-\mathbf{AB}&\mathbf{O}\\\mathbf{B}&\mathbf{I}_n\end{pmatrix},$$

$$\begin{pmatrix}\mathbf{I}_m-\mathbf{AB}&\mathbf{O}\\\mathbf{B}&\mathbf{I}_n\end{pmatrix}\begin{pmatrix}\mathbf{I}_m&\mathbf{O}\\-\mathbf{B}&\mathbf{I}_n\end{pmatrix}=\begin{pmatrix}\mathbf{I}_m-\mathbf{AB}&\mathbf{O}\\\mathbf{O}&\mathbf{I}_n\end{pmatrix}.$$

So $\operatorname{rank}\mathbf{M}=\operatorname{rank}(\mathbf{I}_m-\mathbf{AB})+n$ in the same way. Comparing the two expressions for $\operatorname{rank}\mathbf{M}$ proves (a).

(b) For $m=n$ the two added terms are equal and cancel. A square matrix is invertible exactly when its rank is n (Theorem 7.5.3 (d)), so $\mathbf{I}_n-\mathbf{AB}$ is invertible if and only if $\mathbf{I}_n-\mathbf{BA}$ is.

(c) Here $\mathbf{AB}=\begin{pmatrix}1&0\\0&1\end{pmatrix}=\mathbf{I}_2$, so $\operatorname{rank}(\mathbf{I}_2-\mathbf{AB})=0$, and

$$\mathbf{BA}=\begin{pmatrix}1&1&0\\0&0&0\\0&1&1\end{pmatrix},\qquad \mathbf{I}_3-\mathbf{BA}=\begin{pmatrix}0&-1&0\\0&1&0\\0&-1&0\end{pmatrix},$$

which has rank 1. Indeed $0+3=1+2$. So $\mathbf{AB}=\mathbf{I}_2$ while $\mathbf{BA}\neq\mathbf{I}_3$, and the identity accounts for the difference exactly: $\mathbf{I}_3-\mathbf{BA}$ has rank $3-2=1$.

The trick is worth naming. **To compare two expressions, find one block matrix from which both can be obtained by elimination, and compute its rank twice.** Section 3 does the same with determinants and inverses, and the pair $\mathbf{I}-\mathbf{AB}$, $\mathbf{I}-\mathbf{BA}$ returns there as Sylvester's determinant identity.

8.2.5 Exercises

8.2.5.1 A. Check your understanding

Exercise. A1

(a) State the three inequalities of Theorem 8.2.1.

(b) True or false: $\operatorname{rank}\begin{pmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{O} & \mathbf{D}\end{pmatrix} = \operatorname{rank}\mathbf{A} + \operatorname{rank}\mathbf{D}$ for all blocks of compatible sizes. Justify your answer.

(c) Which matrix do you multiply $\begin{pmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{C} & \mathbf{D}\end{pmatrix}$ by, and on which side, to replace $\mathbf{C}$ by $\mathbf{C}-\mathbf{XA}$?

(d) True or false: $\operatorname{rank}\begin{pmatrix}\mathbf{A} & \mathbf{B}\end{pmatrix} = \operatorname{rank}\mathbf{A} + \operatorname{rank}\mathbf{B}$ whenever $\mathbf{A}$ and $\mathbf{B}$ have the same number of rows. Justify your answer.

(e) State Frobenius's inequality, and explain how Sylvester's inequality follows from it.

Solution.

(a) $\max(\operatorname{rank}\mathbf{A}, \operatorname{rank}\mathbf{B}) \le \operatorname{rank}\begin{pmatrix}\mathbf{A} & \mathbf{B}\end{pmatrix} \le \operatorname{rank}\mathbf{A} + \operatorname{rank}\mathbf{B}$ (and the same for stacked blocks); $\operatorname{rank}\begin{pmatrix}\mathbf{A} & \mathbf{O}\\ \mathbf{O} & \mathbf{D}\end{pmatrix} = \operatorname{rank}\mathbf{A} + \operatorname{rank}\mathbf{D}$; and both block triangular shapes have rank at least $\operatorname{rank}\mathbf{A} + \operatorname{rank}\mathbf{D}$.

(b) False. $\begin{pmatrix}0 & 1\\ 0 & 0\end{pmatrix}$ with 1×1 blocks has rank 1, but $\operatorname{rank}(0)+\operatorname{rank}(0) = 0$. Only $\ge$ holds in general.

(c) Multiply on the left by $\begin{pmatrix}\mathbf{I} & \mathbf{O}\\ -\mathbf{X} & \mathbf{I}\end{pmatrix}$: block row 2 becomes $\begin{pmatrix}\mathbf{C}-\mathbf{XA} & \mathbf{D}-\mathbf{XB}\end{pmatrix}$ by Theorem 8.1.1. The rank does not change (Theorem 8.2.2).

(d) False. $\begin{pmatrix}1 & 1\end{pmatrix}$ has rank 1, not $1+1$. Equality holds exactly when $\operatorname{col}\mathbf{A} \cap \operatorname{col}\mathbf{B} = \{\mathbf{0}\}$.

(e) $\operatorname{rank}(\mathbf{ABC}) + \operatorname{rank}\mathbf{B} \ge \operatorname{rank}(\mathbf{AB}) + \operatorname{rank}(\mathbf{BC})$ for matrices of compatible sizes. With $\mathbf{B} = \mathbf{I}_n$ it reads $\operatorname{rank}(\mathbf{AC}) + n \ge \operatorname{rank}\mathbf{A} + \operatorname{rank}\mathbf{C}$, which is Sylvester's inequality.

8.2.5.2 B. Practice

Exercise. B1: Ranks of block matrices

Let $\mathbf{A} = \begin{pmatrix}1 & 2\\ 2 & 4\end{pmatrix}$ and $\mathbf{D} = \begin{pmatrix}1 & 1\\ 0 & 1\end{pmatrix}$ in $M_2(\mathbb{Q})$. Find the rank of each 4×4 matrix, citing the result you use. In which case is the lower bound of Theorem 8.2.1 (c) strict?

(a) $\mathbf{M}_1 = \begin{pmatrix} \mathbf{A} & \mathbf{O} \\ \mathbf{O} & \mathbf{D} \end{pmatrix}$.

(b) $\mathbf{M}_2 = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{O} & \mathbf{D} \end{pmatrix}$ with $\mathbf{B} = \begin{pmatrix} 1 & 3 \\ 2 & 6 \end{pmatrix}$.

(c) $\mathbf{M}_3 = \begin{pmatrix} \mathbf{A} & \mathbf{I}_2 \\ \mathbf{O} & \mathbf{O} \end{pmatrix}$.

Solution. $\operatorname{rank}\mathbf{A} = 1$, since the second row is twice the first and $\mathbf{A} \neq \mathbf{O}$; $\operatorname{rank}\mathbf{D} = 2$, since $\det\mathbf{D} = 1 \neq 0$.

(a) By Theorem 8.2.1 (b), $\operatorname{rank}\mathbf{M}_1 = 1 + 2 = 3$.

(b) The columns of $\mathbf{B}$ are $(1,2)$ and $(3,6) = 3(1,2)$, and $(1,2)$ is the first column of $\mathbf{A}$. So $\operatorname{col}\mathbf{B} \subseteq \operatorname{col}\mathbf{A}$, and by Theorem 8.2.2 (b), $\operatorname{rank}\mathbf{M}_2 = 1 + 2 = 3$. (Explicitly $\mathbf{B} = \mathbf{AY}$ with $\mathbf{Y} = \begin{pmatrix} 1 & 3 \\ 0 & 0 \end{pmatrix}$.)

(c) The lower bound is $\operatorname{rank}\mathbf{A} + \operatorname{rank}\mathbf{O} = 1$. But the last two columns of $\mathbf{M}_3$ are $\mathbf{e}_1, \mathbf{e}_2$, and every column lies in $\operatorname{span}(\mathbf{e}_1, \mathbf{e}_2)$, so $\operatorname{rank}\mathbf{M}_3 = 2$. This is the strict case: the block $\mathbf{I}_2$ is not a multiple of $\mathbf{A}$ or of the zero block.

Exercise. B2: Shifted products

Let $\mathbf{A} \in M_{m\times n}(F)$, $\mathbf{B} \in M_{n\times m}(F)$, and $c \in F$ with $c \neq 0$.

(a) By adapting Example (Rank of $\mathbf{I}-\mathbf{AB}$ versus $\mathbf{I}-\mathbf{BA}$), prove that $\operatorname{rank}(c\mathbf{I}_m - \mathbf{AB}) + n = \operatorname{rank}(c\mathbf{I}_n - \mathbf{BA}) + m$.

(b) Check (a) over $\mathbb{Q}$ for $c = 2$, $\mathbf{A} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \\ 1 & 0 \end{pmatrix}$, and $\mathbf{B} = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$.

(c) Show by an example with $m = 1$, $n = 2$ that (a) can fail for $c = 0$.

Solution.

(a) Since $c \neq 0$, put $\mathbf{A}' = c^{-1}\mathbf{A}$. Then $c\mathbf{I}_m - \mathbf{AB} = c(\mathbf{I}_m - \mathbf{A}'\mathbf{B})$ and $c\mathbf{I}_n - \mathbf{BA} = c(\mathbf{I}_n - \mathbf{BA}')$. Multiplying a matrix by the nonzero scalar c does not change its column space, hence not its rank. By Example (Rank of $\mathbf{I} - \mathbf{AB}$ versus $\mathbf{I} - \mathbf{BA}$) (a) applied to $\mathbf{A}'$ and $\mathbf{B}$,

$$\begin{aligned} \operatorname{rank}(c\mathbf{I}_m - \mathbf{AB}) + n &= \operatorname{rank}(\mathbf{I}_m - \mathbf{A}'\mathbf{B}) + n \\ &= \operatorname{rank}(\mathbf{I}_n - \mathbf{BA}') + m \\ &= \operatorname{rank}(c\mathbf{I}_n - \mathbf{BA}) + m. \end{aligned}$$

(b) Here $m = 3$, $n = 2$, and

$$2\mathbf{I}_3 - \mathbf{AB} = \begin{pmatrix} 1 & -1 & -1 \\ 0 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix}, \qquad 2\mathbf{I}_2 - \mathbf{BA} = \begin{pmatrix} 0 & -1 \\ 0 & 1 \end{pmatrix},$$

since $\mathbf{AB} = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}$ and $\mathbf{BA} = \begin{pmatrix} 2 & 1 \\ 0 & 1 \end{pmatrix}$. In the first matrix, row 1 equals $-(\text{row } 2) - (\text{row } 3)$, and rows 2 and 3 are not multiples of each other, so its rank is 2 (Theorem 3.5.5). The second has rank 1. Indeed $2 + 2 = 1 + 3$.

(c) For $c = 0$ the identity would say $\operatorname{rank}(\mathbf{AB}) + n = \operatorname{rank}(\mathbf{BA}) + m$. Take $\mathbf{A} = \begin{pmatrix} 1 & 0 \end{pmatrix}$ and $\mathbf{B} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$: $\mathbf{AB} = (1)$ has rank 1, and $\mathbf{BA} = \mathbf{E}_{11} \in M_2(F)$ has rank 1, but $1 + 2 \neq 1 + 1$.

Exercise. B3: Idempotents by blocks

Let $\mathbf{A} \in M_n(F)$.

(a) Starting from $\begin{pmatrix} \mathbf{A} & \mathbf{O} \\ \mathbf{O} & \mathbf{I}_n - \mathbf{A} \end{pmatrix}$, use block elimination to prove

$$\operatorname{rank} \mathbf{A} + \operatorname{rank}(\mathbf{I}_n - \mathbf{A}) = n + \operatorname{rank}(\mathbf{A} - \mathbf{A}^2).$$

(b) Deduce that $\mathbf{A}^2 = \mathbf{A}$ if and only if $\operatorname{rank} \mathbf{A} + \operatorname{rank}(\mathbf{I}_n - \mathbf{A}) = n$.

(c) Compare (b) with what Chapter 4 says about projections.

Hint: first add block row 1 to block row 2, then add block column 2 to block column 1.

Solution.

(a) Write $\mathbf{I} = \mathbf{I}_n$. By Theorem 8.1.1, each step being multiplication by one of the matrices of Theorem 8.2.2 (a):

$$\begin{pmatrix} \mathbf{A} & \mathbf{O} \\ \mathbf{O} & \mathbf{I} - \mathbf{A} \end{pmatrix} \xrightarrow{\text{row}_2 + \text{row}_1} \begin{pmatrix} \mathbf{A} & \mathbf{O} \\ \mathbf{A} & \mathbf{I} - \mathbf{A} \end{pmatrix} \xrightarrow{\text{col}_1 + \text{col}_2} \begin{pmatrix} \mathbf{A} & \mathbf{O} \\ \mathbf{I} & \mathbf{I} - \mathbf{A} \end{pmatrix} \xrightarrow{\text{row}_1 - \mathbf{A}\,\text{row}_2} \begin{pmatrix} \mathbf{O} & \mathbf{A}^2 - \mathbf{A} \\ \mathbf{I} & \mathbf{I} - \mathbf{A} \end{pmatrix} \xrightarrow{\text{col}_2 - \text{col}_1(\mathbf{I} - \mathbf{A})} \begin{pmatrix} \mathbf{O} & \mathbf{A}^2 - \mathbf{A} \\ \mathbf{I} & \mathbf{O} \end{pmatrix}.$$

The multipliers are, in order, $\begin{pmatrix} \mathbf{I} & \mathbf{O} \\ \mathbf{I} & \mathbf{I} \end{pmatrix}$ on the left, $\begin{pmatrix} \mathbf{I} & \mathbf{O} \\ \mathbf{I} & \mathbf{I} \end{pmatrix}$ on the right, $\begin{pmatrix} \mathbf{I} & -\mathbf{A} \\ \mathbf{O} & \mathbf{I} \end{pmatrix}$ on the left, and $\begin{pmatrix} \mathbf{I} & -(\mathbf{I} - \mathbf{A}) \\ \mathbf{O} & \mathbf{I} \end{pmatrix}$ on the right. In the third step the top row becomes $\begin{pmatrix} \mathbf{A} - \mathbf{A}\mathbf{I} & \mathbf{O} - \mathbf{A}(\mathbf{I} - \mathbf{A}) \end{pmatrix} = \begin{pmatrix} \mathbf{O} & \mathbf{A}^2 - \mathbf{A} \end{pmatrix}$, and in the fourth the bottom-right block becomes $(\mathbf{I} - \mathbf{A}) - \mathbf{I}(\mathbf{I} - \mathbf{A}) = \mathbf{O}$. Swapping the block columns (Theorem 8.2.2 (a)) gives $\begin{pmatrix} \mathbf{A}^2 - \mathbf{A} & \mathbf{O} \\ \mathbf{O} & \mathbf{I} \end{pmatrix}$, of rank $\operatorname{rank}(\mathbf{A}^2 - \mathbf{A}) + n$ by Theorem 8.2.1 (b). The first

matrix has rank $\operatorname{rank}\mathbf{A} + \operatorname{rank}(\mathbf{I}-\mathbf{A})$, and rank is unchanged throughout. Since $\mathbf{A}^2 - \mathbf{A} = -(\mathbf{A}-\mathbf{A}^2)$ has the same rank as $\mathbf{A}-\mathbf{A}^2$, (a) follows.

(b) By (a), $\operatorname{rank}\mathbf{A}+\operatorname{rank}(\mathbf{I}-\mathbf{A}) = n$ if and only if $\operatorname{rank}(\mathbf{A}-\mathbf{A}^2) = 0$, that is, $\mathbf{A}-\mathbf{A}^2 = \mathbf{O}$.

(c) If $\mathbf{A}^2 = \mathbf{A}$, then $T_{\mathbf{A}}$ is a projection (Definition 4.5.1), so $F^n = \operatorname{im} T_{\mathbf{A}} \oplus \ker T_{\mathbf{A}}$ by Theorem 4.5.1, with $\ker T_{\mathbf{A}} = \operatorname{col}(\mathbf{I}-\mathbf{A})$ (a vector $\mathbf{v}$ with $\mathbf{A}\mathbf{v} = \mathbf{0}$ equals $(\mathbf{I}-\mathbf{A})\mathbf{v}$, and conversely $\mathbf{A}(\mathbf{I}-\mathbf{A}) = \mathbf{A}-\mathbf{A}^2 = \mathbf{O}$). So $\operatorname{rank}\mathbf{A} + \operatorname{rank}(\mathbf{I}-\mathbf{A}) = \dim\operatorname{im} T_{\mathbf{A}} + \dim\ker T_{\mathbf{A}} = n$, which is the forward direction of (b). The block argument also gives the converse, and it never mentions subspaces.

8.2.5.3 C. Going deeper

Exercise. C1: The rank of $\mathbf{A} - \mathbf{ABA}$

Let $\mathbf{A} \in M_{m\times n}(F)$ and $\mathbf{B} \in M_{n\times m}(F)$.

(a) Verify, by block multiplication, the following chain, in which each arrow is multiplication by an invertible block triangular matrix:

$$\begin{pmatrix}\mathbf{A} & \mathbf{O}\\ \mathbf{O} & \mathbf{I}_n-\mathbf{BA}\end{pmatrix} \to \begin{pmatrix}\mathbf{A} & \mathbf{A}\\ \mathbf{O} & \mathbf{I}_n-\mathbf{BA}\end{pmatrix} \to \begin{pmatrix}\mathbf{A} & \mathbf{A}\\ \mathbf{BA} & \mathbf{I}_n\end{pmatrix}$$
$$\to \begin{pmatrix}\mathbf{A}-\mathbf{ABA} & \mathbf{A}\\ \mathbf{O} & \mathbf{I}_n\end{pmatrix} \to \begin{pmatrix}\mathbf{A}-\mathbf{ABA} & \mathbf{O}\\ \mathbf{O} & \mathbf{I}_n\end{pmatrix}.$$

Name each multiplier and the side it acts on.

(b) Deduce that $\operatorname{rank}(\mathbf{A}-\mathbf{ABA}) = \operatorname{rank}\mathbf{A} + \operatorname{rank}(\mathbf{I}_n-\mathbf{BA}) - n$.

(c) Suppose $\mathbf{ABA} = \mathbf{A}$. Show that $\mathbf{BA}$ is idempotent and that $\operatorname{rank}(\mathbf{BA}) = \operatorname{rank}\mathbf{A}$.

Hint for (c): use (b), then Exercise (B3: Idempotents by blocks).

Solution.

(a) The partitions are m, n for rows and n, n for columns. By Theorem 8.1.1:

1. Right multiplication by $\begin{pmatrix}\mathbf{I}_n & \mathbf{I}_n\\ \mathbf{O} & \mathbf{I}_n\end{pmatrix}$ adds block column 1 to block column 2:
$$\begin{pmatrix}\mathbf{A} & \mathbf{A}+\mathbf{O}\\ \mathbf{O} & \mathbf{O}+(\mathbf{I}_n-\mathbf{BA})\end{pmatrix}.$$
2. Left multiplication by $\begin{pmatrix}\mathbf{I}_m & \mathbf{O}\\ \mathbf{B} & \mathbf{I}_n\end{pmatrix}$ adds $\mathbf{B}$ times block row 1 to block row 2: the new row is $\begin{pmatrix}\mathbf{BA} & \mathbf{BA}+\mathbf{I}_n-\mathbf{BA}\end{pmatrix} = \begin{pmatrix}\mathbf{BA} & \mathbf{I}_n\end{pmatrix}$.
3. Right multiplication by $\begin{pmatrix}\mathbf{I}_n & \mathbf{O}\\ -\mathbf{BA} & \mathbf{I}_n\end{pmatrix}$ subtracts block column 2 times $\mathbf{BA}$ from block column 1: the new column is $\begin{pmatrix}\mathbf{A}-\mathbf{ABA}\\ \mathbf{BA}-\mathbf{I}_n\mathbf{BA}\end{pmatrix} = \begin{pmatrix}\mathbf{A}-\mathbf{ABA}\\ \mathbf{O}\end{pmatrix}$.
4. Left multiplication by $\begin{pmatrix}\mathbf{I}_m & -\mathbf{A}\\ \mathbf{O} & \mathbf{I}_n\end{pmatrix}$ subtracts $\mathbf{A}$ times block row 2 from block row 1:

the new row is $\begin{pmatrix}\mathbf{A}-\mathbf{ABA}-\mathbf{A}\cdot\mathbf{O} & \mathbf{A}-\mathbf{AI}_n\end{pmatrix} = \begin{pmatrix}\mathbf{A}-\mathbf{ABA} & \mathbf{O}\end{pmatrix}$.

All four multipliers are invertible by Theorem 8.2.2 (a).

(b) By Theorem 8.2.2 (a) the first and last matrices have the same rank, and by Theorem 8.2.1 (b) these ranks are $\operatorname{rank}\mathbf{A} + \operatorname{rank}(\mathbf{I}_n - \mathbf{BA})$ and $\operatorname{rank}(\mathbf{A}-\mathbf{ABA}) + n$. Rearranging gives the formula.

(c) $(\mathbf{BA})^2 = \mathbf{B}(\mathbf{ABA}) = \mathbf{BA}$, so $\mathbf{BA}$ is idempotent. By (b), $0 = \operatorname{rank}(\mathbf{A}-\mathbf{ABA}) = \operatorname{rank}\mathbf{A} + \operatorname{rank}(\mathbf{I}_n - \mathbf{BA}) - n$, so $\operatorname{rank}(\mathbf{I}_n - \mathbf{BA}) = n - \operatorname{rank}\mathbf{A}$. By Exercise (B3: Idempotents by blocks) (b) applied to the idempotent $\mathbf{BA}$, $\operatorname{rank}(\mathbf{BA}) = n - \operatorname{rank}(\mathbf{I}_n - \mathbf{BA}) = \operatorname{rank}\mathbf{A}$.

Exercise. C2: Involutions and the field

Let $\mathbf{A} \in M_n(F)$.

(a) Suppose the characteristic of F is not 2. Starting from $\begin{pmatrix}\mathbf{A}-\mathbf{I}_n & \mathbf{O}\\ \mathbf{O} & \mathbf{A}+\mathbf{I}_n\end{pmatrix}$, prove that

$$\operatorname{rank}(\mathbf{A}-\mathbf{I}_n) + \operatorname{rank}(\mathbf{A}+\mathbf{I}_n) = n + \operatorname{rank}(\mathbf{A}^2-\mathbf{I}_n).$$

(b) Deduce that if $\mathbf{A}^2 = \mathbf{I}_n$ (and $\operatorname{char} F \neq 2$), then $\operatorname{rank}(\mathbf{A}-\mathbf{I}_n) + \operatorname{rank}(\mathbf{A}+\mathbf{I}_n) = n$.

(c) Show that (b) fails over $\mathbb{F}_2$, and point to the step of (a) that breaks.

Hint for (a): add block column 2 to block column 1, and subtract block row 2 from block row 1; then $-2\mathbf{I}_n$ is an invertible pivot.

Solution.

(a) Write $\mathbf{I} = \mathbf{I}_n$ and $\mathbf{P} = \mathbf{A}+\mathbf{I}$. Since $\operatorname{char} F \neq 2$, $2 \neq 0$ in F, so $\frac{1}{2}$ exists and $-2\mathbf{I}$ is invertible. By Theorem 8.1.1, with each multiplier from Theorem 8.2.2 (a):

$$\begin{pmatrix}\mathbf{A}-\mathbf{I} & \mathbf{O}\\ \mathbf{O} & \mathbf{P}\end{pmatrix} \xrightarrow{\text{col}_1+\text{col}_2} \begin{pmatrix}\mathbf{A}-\mathbf{I} & \mathbf{O}\\ \mathbf{P} & \mathbf{P}\end{pmatrix} \xrightarrow{\text{row}_1-\text{row}_2} \begin{pmatrix}-2\mathbf{I} & -\mathbf{P}\\ \mathbf{P} & \mathbf{P}\end{pmatrix}$$
$$\xrightarrow{\text{col}_2-\text{col}_1\cdot\frac{1}{2}\mathbf{P}} \begin{pmatrix}-2\mathbf{I} & \mathbf{O}\\ \mathbf{P} & \mathbf{P}-\frac{1}{2}\mathbf{P}^2\end{pmatrix}$$
$$\xrightarrow{\text{row}_2+\frac{1}{2}\mathbf{P}\,\text{row}_1} \begin{pmatrix}-2\mathbf{I} & \mathbf{O}\\ \mathbf{O} & \mathbf{P}-\frac{1}{2}\mathbf{P}^2\end{pmatrix}.$$

In the second step $(\mathbf{A}-\mathbf{I})-\mathbf{P} = -2\mathbf{I}$; in the third, $-\mathbf{P}-(-2\mathbf{I})\frac{1}{2}\mathbf{P} = \mathbf{O}$ and $\mathbf{P}-\mathbf{P}\cdot\frac{1}{2}\mathbf{P} = \mathbf{P}-\frac{1}{2}\mathbf{P}^2$; in the fourth, $\mathbf{P}+\frac{1}{2}\mathbf{P}(-2\mathbf{I}) = \mathbf{O}$. Finally

$$\mathbf{P}-\tfrac{1}{2}\mathbf{P}^2 = \tfrac{1}{2}\mathbf{P}(2\mathbf{I}-\mathbf{P}) = \tfrac{1}{2}(\mathbf{A}+\mathbf{I})(\mathbf{I}-\mathbf{A}) = \tfrac{1}{2}(\mathbf{I}-\mathbf{A}^2),$$

where the product $(\mathbf{A}+\mathbf{I})(\mathbf{I}-\mathbf{A}) = \mathbf{I}-\mathbf{A}+\mathbf{A}-\mathbf{A}^2$ expands by distributivity. The first matrix has rank $\operatorname{rank}(\mathbf{A}-\mathbf{I})+\operatorname{rank}(\mathbf{A}+\mathbf{I})$ and the last has rank $n+\operatorname{rank}(\frac{1}{2}(\mathbf{I}-\mathbf{A}^2)) = n + \operatorname{rank}(\mathbf{A}^2-\mathbf{I})$ by Theorem 8.2.1 (b), nonzero scalar multiples having the same column space. Rank is unchanged throughout.

(b) If $\mathbf{A}^2 = \mathbf{I}_n$, then $\operatorname{rank}(\mathbf{A}^2-\mathbf{I}_n) = 0$.

(c) Over $\mathbb{F}_2$, take $\mathbf{A} = \mathbf{I}_n$. Then $\mathbf{A}^2 = \mathbf{I}_n$, but $\mathbf{A}-\mathbf{I}_n = \mathbf{O}$ and $\mathbf{A}+\mathbf{I}_n = 2\mathbf{I}_n = \mathbf{O}$, so the left side is $0 \neq n$. The step that breaks is the pivot: $-2\mathbf{I}_n = \mathbf{O}$ is not invertible, and $\frac{1}{2}$ does not exist. (Chapter 2's decomposition of an involution into ± 1 parts, Theorem 4.5.5, needs the same hypothesis.)

8.3 Schur Complements

Section 1 inverted block triangular matrices, and Section 2 cleared blocks without changing the rank. A general block matrix $\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}$ has no zero block to help, but when one diagonal block is invertible, a single step of block elimination creates one. What is left in the opposite corner is a new matrix, the Schur complement, and it carries everything that the invertible block does not: the determinant, the rank, and the inverse of the whole matrix can all be read off from the block and its complement. This section introduces it, derives these formulas, uses the two complements of one matrix to prove the Sherman–Morrison–Woodbury formula, and ends with the determinant of a block matrix whose lower blocks commute, over every field.

8.3.1 Eliminating a block

Suppose we must solve a linear system whose unknowns come in two groups, $\mathbf{x} \in F^k$ and $\mathbf{y} \in F^l$:

$$\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix} \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} = \begin{pmatrix} \mathbf{f} \\ \mathbf{g} \end{pmatrix}, \qquad \text{that is,} \qquad \begin{aligned} \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{y} &= \mathbf{f}, \\ \mathbf{C}\mathbf{x} + \mathbf{D}\mathbf{y} &= \mathbf{g}, \end{aligned}$$

with $\mathbf{A} \in M_k(F)$ invertible. Do what we would do for two equations in two numbers: solve the first equation for $\mathbf{x}$, and substitute. The first equation gives $\mathbf{x} = \mathbf{A}^{-1}(\mathbf{f} - \mathbf{B}\mathbf{y})$. Substituting into the second,

$$\mathbf{C}\mathbf{A}^{-1}\mathbf{f} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B}\mathbf{y} + \mathbf{D}\mathbf{y} = \mathbf{g}, \qquad \text{so} \qquad (\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})\,\mathbf{y} = \mathbf{g} - \mathbf{C}\mathbf{A}^{-1}\mathbf{f}.$$

The unknown $\mathbf{x}$ has disappeared, and $\mathbf{y}$ solves a smaller system whose matrix is $\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B}$. Once $\mathbf{y}$ is known, $\mathbf{x} = \mathbf{A}^{-1}(\mathbf{f} - \mathbf{B}\mathbf{y})$. So the big system is solvable for every right-hand side exactly when the small matrix $\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B}$ allows it, and this matrix will keep appearing.

For $k = l = 1$ we have met it before: $d - cb/a = (ad - bc)/a$ is the determinant divided by the pivot a, which is what Gaussian elimination leaves in the corner after clearing the entry c.

The Schur complement of an invertible block is what remains in the opposite corner after that block has been used to eliminate its row and column.

Definition 8.3.1. Schur Complement

Let $\mathbf{M} = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix} \in M_{(k+p)\times(k+l)}(F)$ be a block matrix with $\mathbf{A} \in M_k(F)$, $\mathbf{B} \in M_{k\times l}(F)$, $\mathbf{C} \in M_{p\times k}(F)$, and $\mathbf{D} \in M_{p\times l}(F)$, where $k, p, l \geq 1$.

(a) If $\mathbf{A}$ is **invertible**, the **Schur complement of $\mathbf{A}$ in $\mathbf{M}$** is

$$\mathbf{M}/\mathbf{A} \coloneqq \mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B} \in M_{p\times l}(F).$$

(b) Symmetrically, if $\mathbf{M} = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}$ with $\mathbf{D} \in M_l(F)$ **invertible**, $\mathbf{A} \in M_{m\times k}(F)$, $\mathbf{B} \in$

$M_{m\times l}(F)$, and $\mathbf{C} \in M_{l\times k}(F)$, the **Schur complement of D in M** is $\mathbf{M}/\mathbf{D} \coloneqq \mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C} \in M_{m\times k}(F)$.

In words, clause by clause. The block whose complement we take must be **square and invertible**, since its inverse appears in the formula; the other blocks may be rectangular. The complement has the size of the **opposite corner**: $\mathbf{M}/\mathbf{A}$ has the shape of $\mathbf{D}$, and $\mathbf{M}/\mathbf{D}$ has the shape of $\mathbf{A}$. In $\mathbf{C}\mathbf{A}^{-1}\mathbf{B}$ the factors come in the only order the sizes allow: $\mathbf{C}$ is $p \times k$, $\mathbf{A}^{-1}$ is $k \times k$, $\mathbf{B}$ is $k \times l$. A way to remember it: go from the corner $\mathbf{D}$ to $\mathbf{A}$ along its row ($\mathbf{C}$), invert, and come back along its column ($\mathbf{B}$).

Well-definedness. $\mathbf{M}/\mathbf{A}$ depends on $\mathbf{M}$ and on the chosen partition, which says which corner is $\mathbf{A}$. The inverse $\mathbf{A}^{-1}$ is unique (Theorem 0.9.7), so there is no other choice to make.

Examples.

- **Scalars.** For $\mathbf{M} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ with $a \neq 0$, $\mathbf{M}/a = d - ca^{-1}b = (ad - bc)/a$. In particular $a \cdot (\mathbf{M}/a) = \det \mathbf{M}$.
- **An identity corner.** If $\mathbf{A} = \mathbf{I}_k$, then $\mathbf{M}/\mathbf{I}_k = \mathbf{D} - \mathbf{C}\mathbf{B}$. No inversion is needed at all.
- **Block triangular.** If $\mathbf{C} = \mathbf{O}$, then $\mathbf{M}/\mathbf{A} = \mathbf{D}$: nothing needs to be eliminated.
- **A 3×3 example.** For $\mathbf{M} = \left(\begin{array}{cc|c} 2 & 1 & 1 \\ 1 & 1 & 0 \\ \hline 3 & 1 & 4 \end{array}\right)$ over $\mathbb{Q}$, $\mathbf{A} = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$ has $\mathbf{A}^{-1} = \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix}$, and $\mathbf{C}\mathbf{A}^{-1} = \begin{pmatrix} 3 & 1 \end{pmatrix}\mathbf{A}^{-1} = \begin{pmatrix} 2 & -1 \end{pmatrix}$, so $\mathbf{M}/\mathbf{A} = 4 - \begin{pmatrix} 2 & -1 \end{pmatrix}\begin{pmatrix} 1 \\ 0 \end{pmatrix} = (2)$. This 1×1 complement is the last pivot of Gaussian elimination on $\mathbf{M}$: subtracting $\frac{1}{2}$ row 1 from row 2 and $\frac{3}{2}$ row 1 from row 3 gives the rows $(0, \frac{1}{2}, -\frac{1}{2})$ and $(0, -\frac{1}{2}, \frac{5}{2})$, and adding the first of these to the second gives $(0, 0, 2)$.

Nonexample by minimal change. For $\mathbf{M} = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}$, $\mathbf{M}/(1) = 0 - 1 \cdot 1 \cdot 1 = (-1)$. Change the corner entry 1 to 0: $\mathbf{M}' = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$. The blocks $\mathbf{B}$, $\mathbf{C}$, $\mathbf{D}$, and the partition are unchanged, and $\mathbf{M}'$ is still invertible, but $\mathbf{M}'/\mathbf{O}_1$ is **not defined**, because the clause "$\mathbf{A}$ is invertible" fails. Here $\mathbf{M}'/\mathbf{D}$ is not defined either, since $\mathbf{D} = \mathbf{O}_1$.

Why this definition. The formula is read off from the elimination in the hook, so the question is only about the name and the hypotheses. "Complement" refers to position: $\mathbf{M}/\mathbf{A}$ sits in the corner complementary to $\mathbf{A}$. The quotient notation $\mathbf{M}/\mathbf{A}$ is suggested by the first example, $\det(\mathbf{M}/a) = \det \mathbf{M}/\det a$, and the determinant formula below shows that this holds for blocks of every size. Invertibility of $\mathbf{A}$ cannot be dropped, since $\mathbf{A}^{-1}$ is in the formula; the price is that some invertible matrices, like $\mathbf{M}'$, have no Schur complement for a given partition.

Warning. **M can be invertible while A is not.** For $\mathbf{M}' = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$, neither corner is invertible, so neither Schur complement exists for the partition $1, 1$. Before using a formula below, check that the block you invert is invertible. If it is not, try the other corner ($\mathbf{M}/\mathbf{D}$), a different partition, or first permute rows: $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}\mathbf{M}' = \mathbf{I}_2$. The same issue appears in Gaussian elimination as a zero pivot.

Quick check. Let $\mathbf{M} = \left(\begin{array}{c|cc} 2 & 4 & 2 \\ \hline 1 & 3 & 5 \\ 3 & 1 & 0 \end{array}\right)$ with partition $1, 2$. Compute $\mathbf{M}/(2)$.

Solution. Here $\mathbf{A} = (2)$, $\mathbf{B} = \begin{pmatrix} 4 & 2 \end{pmatrix}$, $\mathbf{C} = \begin{pmatrix} 1 \\ 3 \end{pmatrix}$, $\mathbf{D} = \begin{pmatrix} 3 & 5 \\ 1 & 0 \end{pmatrix}$. Then $\mathbf{CA}^{-1}\mathbf{B} = \frac{1}{2}\begin{pmatrix} 4 & 2 \\ 12 & 6 \end{pmatrix} = \begin{pmatrix} 2 & 1 \\ 6 & 3 \end{pmatrix}$, so $\mathbf{M}/(2) = \begin{pmatrix} 1 & 4 \\ -5 & -3 \end{pmatrix}$. These are exactly the entries left in rows 2, 3 after clearing column 1 with the pivot 2.

8.3.2 The block LDU factorization

The elimination of the hook, written with matrices, factors $\mathbf{M}$ into a lower block triangular matrix, a block diagonal matrix, and an upper block triangular matrix. Everything else in this section follows from this factorization.

Theorem 8.3.1. Block LDU Factorization

Let $\mathbf{M} = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}$ be partitioned as in Definition 8.3.1.

(a) If $\mathbf{A} \in M_k(F)$ is invertible, then

$$\mathbf{M} = \begin{pmatrix} \mathbf{I}_k & \mathbf{O} \\ \mathbf{CA}^{-1} & \mathbf{I}_p \end{pmatrix} \begin{pmatrix} \mathbf{A} & \mathbf{O} \\ \mathbf{O} & \mathbf{M}/\mathbf{A} \end{pmatrix} \begin{pmatrix} \mathbf{I}_k & \mathbf{A}^{-1}\mathbf{B} \\ \mathbf{O} & \mathbf{I}_l \end{pmatrix}.$$

(b) If $\mathbf{D} \in M_l(F)$ is invertible, then

$$\mathbf{M} = \begin{pmatrix} \mathbf{I}_m & \mathbf{BD}^{-1} \\ \mathbf{O} & \mathbf{I}_l \end{pmatrix} \begin{pmatrix} \mathbf{M}/\mathbf{D} & \mathbf{O} \\ \mathbf{O} & \mathbf{D} \end{pmatrix} \begin{pmatrix} \mathbf{I}_k & \mathbf{O} \\ \mathbf{D}^{-1}\mathbf{C} & \mathbf{I}_l \end{pmatrix}.$$

Idea. The left factor in (a) records the row operation "subtract $\mathbf{CA}^{-1}$ times block row 1 from block row 2", which turns $\mathbf{C}$ into $\mathbf{O}$ and $\mathbf{D}$ into $\mathbf{M}/\mathbf{A}$. The right factor records the column operation "subtract block column 1 times $\mathbf{A}^{-1}\mathbf{B}$ from block column 2", which turns $\mathbf{B}$ into $\mathbf{O}$. Undoing both operations gives the factorization. To prove it we need not retrace the elimination: multiplying out is enough.

Proof.

(a) Write $\mathbf{S} = \mathbf{M}/\mathbf{A} = \mathbf{D} - \mathbf{CA}^{-1}\mathbf{B}$. The sizes are conformable throughout. By Theorem 8.1.1,

$$\begin{pmatrix} \mathbf{A} & \mathbf{O} \\ \mathbf{O} & \mathbf{S} \end{pmatrix} \begin{pmatrix} \mathbf{I}_k & \mathbf{A}^{-1}\mathbf{B} \\ \mathbf{O} & \mathbf{I}_l \end{pmatrix} = \begin{pmatrix} \mathbf{A} & \mathbf{AA}^{-1}\mathbf{B} \\ \mathbf{O} & \mathbf{S} \end{pmatrix} = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{O} & \mathbf{S} \end{pmatrix},$$

and then

$$\begin{pmatrix} \mathbf{I}_k & \mathbf{O} \\ \mathbf{C}\mathbf{A}^{-1} & \mathbf{I}_p \end{pmatrix} \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{O} & \mathbf{S} \end{pmatrix} = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C}\mathbf{A}^{-1}\mathbf{A} & \mathbf{C}\mathbf{A}^{-1}\mathbf{B} + \mathbf{S} \end{pmatrix} = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix},$$

since $\mathbf{C}\mathbf{A}^{-1}\mathbf{A} = \mathbf{C}$ and $\mathbf{C}\mathbf{A}^{-1}\mathbf{B} + (\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B}) = \mathbf{D}$. By associativity (Theorem 0.9.4) the triple product equals $\mathbf{M}$.

(b) Write $\mathbf{S}' = \mathbf{M}/\mathbf{D} = \mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C}$. By Theorem 8.1.1,

$$\begin{pmatrix} \mathbf{S}' & \mathbf{O} \\ \mathbf{O} & \mathbf{D} \end{pmatrix} \begin{pmatrix} \mathbf{I}_k & \mathbf{O} \\ \mathbf{D}^{-1}\mathbf{C} & \mathbf{I}_l \end{pmatrix} = \begin{pmatrix} \mathbf{S}' & \mathbf{O} \\ \mathbf{C} & \mathbf{D} \end{pmatrix},$$

$$\begin{pmatrix} \mathbf{I}_m & \mathbf{B}\mathbf{D}^{-1} \\ \mathbf{O} & \mathbf{I}_l \end{pmatrix} \begin{pmatrix} \mathbf{S}' & \mathbf{O} \\ \mathbf{C} & \mathbf{D} \end{pmatrix} = \begin{pmatrix} \mathbf{S}' + \mathbf{B}\mathbf{D}^{-1}\mathbf{C} & \mathbf{B}\mathbf{D}^{-1}\mathbf{D} \\ \mathbf{C} & \mathbf{D} \end{pmatrix} = \mathbf{M}.$$

■

The outer factors are block triangular with identity blocks on the diagonal. They are invertible, with inverses obtained by changing the sign of the off-diagonal block (Theorem 8.1.5), and they have determinant 1. So, up to these harmless factors, $\mathbf{M}$ **is** the block diagonal matrix $\mathbf{A} \oplus (\mathbf{M}/\mathbf{A})$ when $\mathbf{A}$ and $\mathbf{D}$ are square. Determinant, rank, and invertibility cannot tell the difference.

8.3.3 Determinant and rank

The outer factors of Theorem 8.3.1 do not change the determinant or the rank, so both can be read off from the middle factor.

Theorem 8.3.2. Schur Determinant Formula

Let $\mathbf{M} = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix} \in M_{k+l}(F)$ with $\mathbf{A} \in M_k(F)$, $\mathbf{D} \in M_l(F)$, and $k, l \geq 1$.

(a) If $\mathbf{A}$ is invertible, then $\det \mathbf{M} = \det \mathbf{A} \cdot \det(\mathbf{M}/\mathbf{A})$.

(b) If $\mathbf{D}$ is invertible, then $\det \mathbf{M} = \det \mathbf{D} \cdot \det(\mathbf{M}/\mathbf{D})$.

Proof.

(a) By Theorem 8.3.1 (a), $\mathbf{M} = Ł\,(\mathbf{A} \oplus (\mathbf{M}/\mathbf{A}))\,\mathbf{U}$ with $Ł = \begin{pmatrix} \mathbf{I}_k & \mathbf{O} \\ \mathbf{C}\mathbf{A}^{-1} & \mathbf{I}_l \end{pmatrix}$ and $\mathbf{U} = \begin{pmatrix} \mathbf{I}_k & \mathbf{A}^{-1}\mathbf{B} \\ \mathbf{O} & \mathbf{I}_l \end{pmatrix}$. Here $p = l$ because $\mathbf{M}$ is square, so $\mathbf{M}/\mathbf{A} \in M_l(F)$. By Theorem 7.8.4, $\det Ł = \det \mathbf{I}_k \det \mathbf{I}_l = 1$, $\det \mathbf{U} = 1$, and $\det(\mathbf{A} \oplus (\mathbf{M}/\mathbf{A})) = \det \mathbf{A} \det(\mathbf{M}/\mathbf{A})$. By Theorem 7.5.4,

$$\det \mathbf{M} = \det Ł \cdot \det \mathbf{A} \det(\mathbf{M}/\mathbf{A}) \cdot \det \mathbf{U} = \det \mathbf{A} \det(\mathbf{M}/\mathbf{A}).$$

(b) The same argument with Theorem 8.3.1 (b), where $\mathbf{M}/\mathbf{D} \in M_k(F)$ because $\mathbf{M}$ is square.

■

This is the correct replacement for the false rule $\det \mathbf{M} = \det(\mathbf{AD} - \mathbf{BC})$ of Section 1: the determinant of $\mathbf{M}$ is that of $\mathbf{A}$ times that of the complement. For $k = l = 1$ it is $a(d - ca^{-1}b) = ad - bc$. Chapter 7 promised this, and it relies on Theorem 7.8.4 in exactly the form proved there, with square diagonal blocks.

Theorem 8.3.3. Schur Rank Formula

Let $\mathbf{M} = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}$ be partitioned as in Definition 8.3.1. If $\mathbf{A} \in M_k(F)$ is invertible, then

$$\operatorname{rank} \mathbf{M} = k + \operatorname{rank}(\mathbf{M}/\mathbf{A}).$$

Symmetrically, if $\mathbf{D} \in M_l(F)$ is invertible, then $\operatorname{rank} \mathbf{M} = l + \operatorname{rank}(\mathbf{M}/\mathbf{D})$.

Proof. By Theorem 8.3.1 (a), $\mathbf{M} = Ł \begin{pmatrix} \mathbf{A} & \mathbf{O} \\ \mathbf{O} & \mathbf{M}/\mathbf{A} \end{pmatrix} \mathbf{U}$, where Ł and $\mathbf{U}$ are invertible by Theorem 8.2.2 (a). By Theorem 3.5.9, $\operatorname{rank} \mathbf{M}$ equals the rank of the middle factor, which is $\operatorname{rank} \mathbf{A} + \operatorname{rank}(\mathbf{M}/\mathbf{A}) = k + \operatorname{rank}(\mathbf{M}/\mathbf{A})$ by Theorem 8.2.1 (b) and Theorem 3.4.4. The second statement follows in the same way from Theorem 8.3.1 (b). ∎

In words: an invertible block of size k contributes exactly k to the rank, and all the remaining rank sits in its complement. For example, in Example (Rank of $\mathbf{I} - \mathbf{AB}$ versus $\mathbf{I} - \mathbf{BA}$) the matrix $\begin{pmatrix} \mathbf{I}_m & \mathbf{A} \\ \mathbf{B} & \mathbf{I}_n \end{pmatrix}$ has the two complements $\mathbf{I}_n - \mathbf{BA}$ and $\mathbf{I}_m - \mathbf{AB}$, and the theorem gives both expressions for its rank at once.

8.3.4 Inverting a block matrix

Since $\mathbf{A} \oplus (\mathbf{M}/\mathbf{A})$ is invertible exactly when $\mathbf{M}/\mathbf{A}$ is, the factorization also inverts $\mathbf{M}$: invert the three factors and multiply them in the reverse order.

Theorem 8.3.4. Block Inverse Formula

Let $\mathbf{M} = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix} \in M_{k+l}(F)$ with $\mathbf{A} \in M_k(F)$ and $\mathbf{D} \in M_l(F)$.

(a) Suppose $\mathbf{A}$ is invertible, and let $\mathbf{S} = \mathbf{M}/\mathbf{A}$. Then $\mathbf{M}$ is invertible if and only if $\mathbf{S}$ is invertible, and in that case

$$\mathbf{M}^{-1} = \begin{pmatrix} \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{B}\mathbf{S}^{-1}\mathbf{C}\mathbf{A}^{-1} & -\mathbf{A}^{-1}\mathbf{B}\mathbf{S}^{-1} \\ -\mathbf{S}^{-1}\mathbf{C}\mathbf{A}^{-1} & \mathbf{S}^{-1} \end{pmatrix}.$$

In particular the lower-right $l \times l$ block of $\mathbf{M}^{-1}$ is $(\mathbf{M}/\mathbf{A})^{-1}$.

(b) Suppose $\mathbf{D}$ is invertible, and let $\mathbf{S}' = \mathbf{M}/\mathbf{D}$. Then $\mathbf{M}$ is invertible if and only if $\mathbf{S}'$ is

invertible, and in that case

$$\mathbf{M}^{-1} = \begin{pmatrix} \mathbf{S}'^{-1} & -\mathbf{S}'^{-1}\mathbf{B}\mathbf{D}^{-1} \\ -\mathbf{D}^{-1}\mathbf{C}\mathbf{S}'^{-1} & \mathbf{D}^{-1} + \mathbf{D}^{-1}\mathbf{C}\mathbf{S}'^{-1}\mathbf{B}\mathbf{D}^{-1} \end{pmatrix}.$$

In particular the upper-left $k \times k$ block of $\mathbf{M}^{-1}$ is $(\mathbf{M}/\mathbf{D})^{-1}$.

Proof.

(a) By Theorem 8.3.3, $\operatorname{rank}\mathbf{M} = k + \operatorname{rank}\mathbf{S}$. So $\operatorname{rank}\mathbf{M} = k + l$ if and only if $\operatorname{rank}\mathbf{S} = l$, and by Theorem 3.4.4 this says that $\mathbf{M}$ is invertible if and only if $\mathbf{S}$ is.
Suppose $\mathbf{S}$ is invertible. In Theorem 8.3.1 (a), $\mathbf{M} = Ł(\mathbf{A} \oplus \mathbf{S})\mathbf{U}$, where $Ł^{-1} = \begin{pmatrix} \mathbf{I}_k & \mathbf{O} \\ -\mathbf{C}\mathbf{A}^{-1} & \mathbf{I}_l \end{pmatrix}$ and $\mathbf{U}^{-1} = \begin{pmatrix} \mathbf{I}_k & -\mathbf{A}^{-1}\mathbf{B} \\ \mathbf{O} & \mathbf{I}_l \end{pmatrix}$ by Theorem 8.1.5, and $(\mathbf{A} \oplus \mathbf{S})^{-1} = \mathbf{A}^{-1} \oplus \mathbf{S}^{-1}$ by Theorem 8.1.4 (c). By Theorem 0.9.7 (3), $\mathbf{M}^{-1} = \mathbf{U}^{-1}(\mathbf{A}^{-1} \oplus \mathbf{S}^{-1})Ł^{-1}$. By Theorem 8.1.1,

$$\mathbf{U}^{-1}(\mathbf{A}^{-1} \oplus \mathbf{S}^{-1}) = \begin{pmatrix} \mathbf{A}^{-1} & -\mathbf{A}^{-1}\mathbf{B}\mathbf{S}^{-1} \\ \mathbf{O} & \mathbf{S}^{-1} \end{pmatrix},$$

$$\begin{pmatrix} \mathbf{A}^{-1} & -\mathbf{A}^{-1}\mathbf{B}\mathbf{S}^{-1} \\ \mathbf{O} & \mathbf{S}^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{I}_k & \mathbf{O} \\ -\mathbf{C}\mathbf{A}^{-1} & \mathbf{I}_l \end{pmatrix}$$
$$= \begin{pmatrix} \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{B}\mathbf{S}^{-1}\mathbf{C}\mathbf{A}^{-1} & -\mathbf{A}^{-1}\mathbf{B}\mathbf{S}^{-1} \\ -\mathbf{S}^{-1}\mathbf{C}\mathbf{A}^{-1} & \mathbf{S}^{-1} \end{pmatrix}.$$

(b) The equivalence follows from the second statement of Theorem 8.3.3 in the same way. By Theorem 8.3.1 (b), $\mathbf{M}^{-1} = \begin{pmatrix} \mathbf{I}_k & \mathbf{O} \\ -\mathbf{D}^{-1}\mathbf{C} & \mathbf{I}_l \end{pmatrix} (\mathbf{S}'^{-1} \oplus \mathbf{D}^{-1}) \begin{pmatrix} \mathbf{I}_k & -\mathbf{B}\mathbf{D}^{-1} \\ \mathbf{O} & \mathbf{I}_l \end{pmatrix}$, and

$$\begin{pmatrix} \mathbf{I}_k & \mathbf{O} \\ -\mathbf{D}^{-1}\mathbf{C} & \mathbf{I}_l \end{pmatrix} \begin{pmatrix} \mathbf{S}'^{-1} & \mathbf{O} \\ \mathbf{O} & \mathbf{D}^{-1} \end{pmatrix} = \begin{pmatrix} \mathbf{S}'^{-1} & \mathbf{O} \\ -\mathbf{D}^{-1}\mathbf{C}\mathbf{S}'^{-1} & \mathbf{D}^{-1} \end{pmatrix},$$

whose product with $\begin{pmatrix} \mathbf{I}_k & -\mathbf{B}\mathbf{D}^{-1} \\ \mathbf{O} & \mathbf{I}_l \end{pmatrix}$ is the displayed matrix by Theorem 8.1.1.

■

Nobody should memorize these blocks. Remember instead that $\mathbf{M}^{-1}$ is the product of three easy inverses in reverse order, and that the corner opposite $\mathbf{A}$ in $\mathbf{M}^{-1}$ is the inverse of the complement of $\mathbf{A}$.

Example. A 4×4 Inverse through a 2×2 Schur Complement

Invert

$$\mathbf{M} = \left(\begin{array}{cc|cc} 1 & 1 & 0 & -1 \\ 1 & 2 & -1 & -1 \\ \hline -1 & -1 & 1 & 1 \\ 0 & 1 & 0 & -1 \end{array}\right) \in M_4(\mathbb{Q})$$

using Theorem 8.3.4, and find $\det\mathbf{M}$.

Solution. The blocks are $\mathbf{A} = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}$, $\mathbf{B} = \begin{pmatrix} 0 & -1 \\ -1 & -1 \end{pmatrix}$, $\mathbf{C} = \begin{pmatrix} -1 & -1 \\ 0 & 1 \end{pmatrix}$, $\mathbf{D} = \begin{pmatrix} 1 & 1 \\ 0 & -1 \end{pmatrix}$.

Step 1: invert $\mathbf{A}$. $\det \mathbf{A} = 1$, so $\mathbf{A}^{-1} = \begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}$ by Theorem 0.9.8.

Step 2: the two "half" products.

$$\mathbf{C}\mathbf{A}^{-1} = \begin{pmatrix} -1 & 0 \\ -1 & 1 \end{pmatrix}, \qquad \mathbf{A}^{-1}\mathbf{B} = \begin{pmatrix} 1 & -1 \\ -1 & 0 \end{pmatrix}, \qquad \mathbf{C}\mathbf{A}^{-1}\mathbf{B} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}.$$

Step 3: the complement and its inverse. $\mathbf{S} = \mathbf{M}/\mathbf{A} = \mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B} = \begin{pmatrix} 1 & 0 \\ 1 & -1 \end{pmatrix}$. Its determinant is $-1 \neq 0$, so $\mathbf{M}$ is invertible by Theorem 8.3.4 (a), and $\mathbf{S}^{-1} = \begin{pmatrix} 1 & 0 \\ 1 & -1 \end{pmatrix} = \mathbf{S}$ by Theorem 0.9.8. By Theorem 8.3.2, $\det \mathbf{M} = \det \mathbf{A} \det \mathbf{S} = -1$.

Step 4: the blocks of $\mathbf{M}^{-1}$. Using Step 2,

$$-\mathbf{A}^{-1}\mathbf{B}\mathbf{S}^{-1} = -\begin{pmatrix} 1 & -1 \\ -1 & 0 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 1 & -1 \end{pmatrix} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix},$$

$$-\mathbf{S}^{-1}\mathbf{C}\mathbf{A}^{-1} = -\begin{pmatrix} 1 & 0 \\ 1 & -1 \end{pmatrix}\begin{pmatrix} -1 & 0 \\ -1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

and

$$\begin{aligned} \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{B}\mathbf{S}^{-1}\mathbf{C}\mathbf{A}^{-1} &= \mathbf{A}^{-1} + (\mathbf{A}^{-1}\mathbf{B}\mathbf{S}^{-1})(\mathbf{C}\mathbf{A}^{-1}) \\ &= \begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix} + \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}\begin{pmatrix} -1 & 0 \\ -1 & 1 \end{pmatrix} \\ &= \begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix} + \begin{pmatrix} -1 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \end{aligned}$$

Hence

$$\mathbf{M}^{-1} = \left(\begin{array}{cc|cc} 1 & 0 & 0 & -1 \\ 0 & 1 & 1 & 0 \\ \hline 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & -1 \end{array}\right).$$

Check one row: row 1 of $\mathbf{M}$ times column 1 of $\mathbf{M}^{-1}$ is $1 + 0 + 0 + 0 = 1$, and times column 3 is $0 + 1 + 0 - 1 = 0$. All the work was with 2×2 matrices, and only two of them had to be inverted.

Remark. When $\mathbf{M}$ is real symmetric, $\mathbf{M}/\mathbf{A}$ is again symmetric, and Chapter 13 uses Schur complements to test whether $\mathbf{M}$ is positive definite one block at a time. Chapter 24 reads Gaussian elimination itself as a sequence of Schur complements of 1×1 pivots, which is the point of view behind block algorithms in numerical linear algebra.

8.3.5 The Sherman–Morrison–Woodbury formula

Chapter 7 proved the Matrix Determinant Lemma, $\det(\mathbf{I}_n + \mathbf{u}\mathbf{v}^\top) = 1 + \mathbf{v}^\top\mathbf{u}$ (Theorem 7.8.5), by bordering $\mathbf{I}_n + \mathbf{u}\mathbf{v}^\top$ into a larger block matrix. There is a matching statement about inverses: if we already know $\mathbf{A}^{-1}$ and change $\mathbf{A}$ by a low-rank term $\mathbf{UCV}$, the new inverse should cost only a small inversion. The trick is to find one block matrix that has $\mathbf{A}$ in one corner and $\mathbf{A} + \mathbf{UCV}$ as a Schur complement, and then to compute its inverse twice.

Theorem 8.3.5. Sherman–Morrison–Woodbury Formula

Let $n, k \geq 1$, let $\mathbf{A} \in M_n(F)$ and $\mathbf{C} \in M_k(F)$ be invertible, and let $\mathbf{U} \in M_{n\times k}(F)$ and $\mathbf{V} \in M_{k\times n}(F)$. Put $\mathbf{W} = \mathbf{C}^{-1} + \mathbf{V}\mathbf{A}^{-1}\mathbf{U} \in M_k(F)$.

(a) $\mathbf{A} + \mathbf{UCV}$ is invertible if and only if $\mathbf{W}$ is invertible.

(b) In that case

$$\begin{aligned}(\mathbf{A} + \mathbf{UCV})^{-1} &= \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{U}\,\mathbf{W}^{-1}\,\mathbf{V}\mathbf{A}^{-1} \\ &= \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{U}(\mathbf{C}^{-1} + \mathbf{V}\mathbf{A}^{-1}\mathbf{U})^{-1}\mathbf{V}\mathbf{A}^{-1}.\end{aligned}$$

(c) $\det(\mathbf{A} + \mathbf{UCV}) = \det\mathbf{A} \cdot \det\mathbf{C} \cdot \det\mathbf{W}$.

Idea. Look for a block matrix whose two Schur complements are $\mathbf{A} + \mathbf{UCV}$ and something involving $\mathbf{W}$. The complement of the lower-right corner of $\mathbf{M} = \begin{pmatrix}\mathbf{A} & \mathbf{U}\\ \mathbf{V} & \mathbf{D}\end{pmatrix}$ is $\mathbf{A} - \mathbf{U}\mathbf{D}^{-1}\mathbf{V}$, which is $\mathbf{A} + \mathbf{UCV}$ when $\mathbf{D}^{-1} = -\mathbf{C}$, that is, $\mathbf{D} = -\mathbf{C}^{-1}$. The complement of the upper-left corner is then $-\mathbf{C}^{-1} - \mathbf{V}\mathbf{A}^{-1}\mathbf{U} = -\mathbf{W}$. Now Theorem 8.3.4 computes the upper-left block of $\mathbf{M}^{-1}$ in two ways: as $(\mathbf{M}/\mathbf{D})^{-1} = (\mathbf{A} + \mathbf{UCV})^{-1}$, and from the formula in terms of $\mathbf{A}$ and $\mathbf{M}/\mathbf{A}$. Setting them equal is the Woodbury formula. The determinant formula of the previous subsection, used twice, gives (c).

Proof. Let $\mathbf{D} = -\mathbf{C}^{-1}$, which is invertible with $\mathbf{D}^{-1} = -\mathbf{C}$ by Theorem 0.9.7 (2) and (5), and let $\mathbf{M} = \begin{pmatrix}\mathbf{A} & \mathbf{U}\\ \mathbf{V} & \mathbf{D}\end{pmatrix} \in M_{n+k}(F)$. By Definition 8.3.1,

$$\begin{aligned}\mathbf{M}/\mathbf{A} &= \mathbf{D} - \mathbf{V}\mathbf{A}^{-1}\mathbf{U} = -\mathbf{C}^{-1} - \mathbf{V}\mathbf{A}^{-1}\mathbf{U} = -\mathbf{W}, \\ \mathbf{M}/\mathbf{D} &= \mathbf{A} - \mathbf{U}\mathbf{D}^{-1}\mathbf{V} = \mathbf{A} + \mathbf{UCV}.\end{aligned}$$

(a) By Theorem 8.3.4 (a) and (b), $\mathbf{M}$ is invertible if and only if $-\mathbf{W}$ is, and if and only if $\mathbf{A} + \mathbf{UCV}$ is. Since $-\mathbf{W}$ is invertible exactly when $\mathbf{W}$ is (Theorem 0.9.7 (5)), this proves (a).

(b) Suppose $\mathbf{W}$ is invertible, so $\mathbf{M}$ is invertible, and $(-\mathbf{W})^{-1} = -\mathbf{W}^{-1}$. The upper-left $n \times n$ block of $\mathbf{M}^{-1}$ is $(\mathbf{M}/\mathbf{D})^{-1} = (\mathbf{A} + \mathbf{UCV})^{-1}$ by Theorem 8.3.4 (b). By Theorem 8.3.4 (a), applied to $\mathbf{M}$, whose blocks are $\mathbf{A}, \mathbf{U}, \mathbf{V}, \mathbf{D}$, and with $\mathbf{S} = \mathbf{M}/\mathbf{A} = -\mathbf{W}$, the same block is

$$\mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{U}(-\mathbf{W})^{-1}\mathbf{V}\mathbf{A}^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{U}\mathbf{W}^{-1}\mathbf{V}\mathbf{A}^{-1}.$$

The two expressions are the same block of the same matrix, so they are equal.

(c) By Theorem 8.3.2 (a) and (b), $\det \mathbf{A} \det(-\mathbf{W}) = \det \mathbf{M} = \det \mathbf{D} \det(\mathbf{A} + \mathbf{UCV})$. Multiplying a $k \times k$ matrix by -1 multiplies each of its k rows by -1, so by Theorem 7.5.1 $\det(-\mathbf{W}) = (-1)^k \det \mathbf{W}$ and $\det \mathbf{D} = (-1)^k \det(\mathbf{C}^{-1}) = (-1)^k (\det \mathbf{C})^{-1}$, using Corollary 7.5.6. Hence $\det \mathbf{A}\,(-1)^k \det \mathbf{W} = (-1)^k (\det \mathbf{C})^{-1} \det(\mathbf{A} + \mathbf{UCV})$. Multiplying both sides by $(-1)^k \det \mathbf{C}$ gives (c).

■

The formula is useful when k is much smaller than n and $\mathbf{A}^{-1}$ is already known or easy (diagonal, say): the only new inversion is of the $k \times k$ matrix $\mathbf{W}$. The most common case is a rank-one update, $k = 1$ and $\mathbf{C} = (1)$.

Corollary 8.3.6. Sherman–Morrison Formula

Let $\mathbf{A} \in M_n(F)$ be invertible and $\mathbf{u}, \mathbf{v} \in F^n$. Then $\mathbf{A} + \mathbf{uv}^\top$ is invertible if and only if $1 + \mathbf{v}^\top \mathbf{A}^{-1} \mathbf{u} \neq 0$, and in that case

$$(\mathbf{A} + \mathbf{uv}^\top)^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1} \mathbf{u}\, \mathbf{v}^\top \mathbf{A}^{-1}}{1 + \mathbf{v}^\top \mathbf{A}^{-1} \mathbf{u}}.$$

Moreover $\det(\mathbf{A} + \mathbf{uv}^\top) = \det \mathbf{A}\,(1 + \mathbf{v}^\top \mathbf{A}^{-1} \mathbf{u})$.

Proof. Apply Theorem 8.3.5 with $k = 1$, $\mathbf{C} = (1)$, $\mathbf{U} = \mathbf{u}$, and $\mathbf{V} = \mathbf{v}^\top$. Then $\mathbf{W} = (1 + \mathbf{v}^\top \mathbf{A}^{-1} \mathbf{u})$ is a 1×1 matrix, invertible exactly when its entry is nonzero, with inverse the reciprocal. Multiplying by a 1×1 matrix is multiplying by its entry, which gives the displayed formula; and $\det \mathbf{W}$ is the entry itself.

■

For example, over $\mathbb{Q}$ let $\mathbf{A} = \operatorname{diag}(1, -1)$, so $\mathbf{A}^{-1} = \mathbf{A}$, and $\mathbf{u} = (1, 1)$. For $\mathbf{v} = (2, 1)$, $\mathbf{A}^{-1}\mathbf{u} = (1, -1)$, $\mathbf{v}^\top \mathbf{A}^{-1} = \begin{pmatrix} 2 & -1 \end{pmatrix}$, and $1 + \mathbf{v}^\top \mathbf{A}^{-1} \mathbf{u} = 1 + (2 - 1) = 2 \neq 0$. So $\mathbf{A} + \mathbf{uv}^\top = \begin{pmatrix} 3 & 1 \\ 2 & 0 \end{pmatrix}$ has determinant $\det \mathbf{A} \cdot 2 = -2$, and

$$\begin{aligned} (\mathbf{A} + \mathbf{uv}^\top)^{-1} &= \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} - \frac{1}{2} \begin{pmatrix} 1 \\ -1 \end{pmatrix} \begin{pmatrix} 2 & -1 \end{pmatrix} \\ &= \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} - \begin{pmatrix} 1 & -\frac{1}{2} \\ -1 & \frac{1}{2} \end{pmatrix} = \begin{pmatrix} 0 & \frac{1}{2} \\ 1 & -\frac{3}{2} \end{pmatrix}, \end{aligned}$$

which agrees with Theorem 0.9.8. For $\mathbf{v} = (1, 2)$ instead, $1 + \mathbf{v}^\top \mathbf{A}^{-1} \mathbf{u} = 1 + (1 - 2) = 0$, and indeed $\mathbf{A} + \mathbf{uv}^\top = \begin{pmatrix} 2 & 2 \\ 1 & 1 \end{pmatrix}$ is not invertible.

With $\mathbf{A} = \mathbf{I}_n$, the determinant statement is the Matrix Determinant Lemma of Chapter 7, now obtained as a special case of a formula about general low-rank updates, and it matches Exercise (C3: Rank-one updates of an invertible matrix).

8.3.6 Commuting blocks

Section 1 warned that $\det \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix} \neq \det(\mathbf{AD} - \mathbf{BC})$ in general. With the Schur complement we can see what goes wrong and what extra hypothesis repairs it. Take all four blocks in $M_n(F)$ and suppose $\mathbf{D}$ is invertible. By Theorem 8.3.2 (b),

$$\begin{aligned} \det \mathbf{M} &= \det \mathbf{D} \, \det(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C}) \\ &= \det\left((\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})\mathbf{D}\right) = \det(\mathbf{AD} - \mathbf{B}\mathbf{D}^{-1}\mathbf{CD}), \end{aligned}$$

using Theorem 7.5.4. If $\mathbf{C}$ and $\mathbf{D}$ commute, then $\mathbf{D}^{-1}\mathbf{CD} = \mathbf{D}^{-1}\mathbf{DC} = \mathbf{C}$, and the right side is $\det(\mathbf{AD} - \mathbf{BC})$. So commuting is exactly what lets $\mathbf{D}^{-1}$ cancel. The real question is how to remove the hypothesis that $\mathbf{D}$ is invertible, which is not needed for the conclusion to make sense.

Over $\mathbb{R}$ one could perturb: $\mathbf{D} + t\mathbf{I}_n$ is invertible for all but finitely many t, both sides are polynomials in t, and polynomials that agree at infinitely many points are equal. Over a finite field such as $\mathbb{F}_2$ there may be no invertible perturbation at all. We use instead a perturbation by an **indeterminate**: replace $\mathbf{D}$ by $\mathbf{D} + x\mathbf{I}_n$, a matrix with entries in $F[x]$. Its determinant is a monic polynomial, hence not the zero polynomial, and that is all the cancellation needs. First we record that evaluating polynomial entries commutes with matrix operations.

For $\mathbf{N} = (n_{ij})$ with entries in $F[x]$ and $c \in F$, write $\mathbf{N}(c) = (n_{ij}(c))$ for the matrix over F obtained by evaluating every entry at c. A matrix over F is a matrix over $F[x]$ with constant entries, and then $\mathbf{N}(c) = \mathbf{N}$.

Lemma 8.3.7. Evaluating a Matrix of Polynomials

Let $c \in F$, and let $\mathbf{N}, \mathbf{N}'$ be matrices with entries in $F[x]$ of sizes for which the operations below are defined. Then $(\mathbf{N}+\mathbf{N}')(c) = \mathbf{N}(c)+\mathbf{N}'(c)$, $(\mathbf{N}\mathbf{N}')(c) = \mathbf{N}(c)\mathbf{N}'(c)$, and, for square $\mathbf{N}$, $(\det \mathbf{N})(c) = \det(\mathbf{N}(c))$.

Proof. By Theorem 0.8.5, evaluation at c sends sums of polynomials to sums, products to products, and constants to themselves. The (i, j)-entry of $\mathbf{N}+\mathbf{N}'$ is $n_{ij}+n'_{ij}$ and that of $\mathbf{N}\mathbf{N}'$ is $\sum_l n_{il}n'_{lj}$, so evaluating gives the corresponding entries of $\mathbf{N}(c) + \mathbf{N}'(c)$ and $\mathbf{N}(c)\mathbf{N}'(c)$. The determinant $\det \mathbf{N}$ is the Leibniz sum $\sum_\sigma \operatorname{sgn}(\sigma) \prod_j n_{\sigma(j)j}$ of Definition 7.4.1, computed in $F[x]$, and evaluating it term by term gives the Leibniz sum for $\mathbf{N}(c)$. This is the argument of Lemma 7.10.1. ∎

Theorem 8.3.8. Determinant with Commuting Blocks

Let $n \geq 1$ and $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D} \in M_n(F)$ over **any** field F. If $\mathbf{CD} = \mathbf{DC}$, then

$$\det \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix} = \det(\mathbf{AD} - \mathbf{BC}).$$

Idea. ① Work over the commutative ring $F[x]$, replacing $\mathbf{D}$ by $\mathbf{D}_x = \mathbf{D} + x\mathbf{I}_n$, which still commutes with $\mathbf{C}$. ② Multiply $\mathbf{M}_x = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D}_x \end{pmatrix}$ on the right by $\begin{pmatrix} \mathbf{D}_x & \mathbf{O} \\ -\mathbf{C} & \mathbf{I}_n \end{pmatrix}$: the lower-left block becomes $\mathbf{C}\mathbf{D}_x - \mathbf{D}_x\mathbf{C} = \mathbf{O}$ and the upper-left block becomes $\mathbf{A}\mathbf{D}_x - \mathbf{B}\mathbf{C}$. This is the Schur computation above with $\mathbf{D}^{-1}$ cleared from the denominators. ③ Take determinants: $\det \mathbf{M}_x \det \mathbf{D}_x = \det(\mathbf{A}\mathbf{D}_x - \mathbf{B}\mathbf{C}) \det \mathbf{D}_x$. ④ Cancel $\det \mathbf{D}_x$, which is legal because it is a nonzero polynomial and $F[x]$ has no zero divisors. ⑤ Set $x = 0$. Each step uses only determinant facts that Chapter 7 proved over commutative rings.

Proof. Regard $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}$ as matrices over $F[x]$ with constant entries, and let $\mathbf{D}_x = \mathbf{D} + x\mathbf{I}_n$. The proofs of Theorem 0.9.4 use only the ring axioms, so they hold for matrices over $F[x]$; in particular

$$\mathbf{C}\mathbf{D}_x = \mathbf{C}\mathbf{D} + x\mathbf{C} = \mathbf{D}\mathbf{C} + x\mathbf{C} = \mathbf{D}_x\mathbf{C},$$

using the hypothesis $\mathbf{C}\mathbf{D} = \mathbf{D}\mathbf{C}$. Let

$$\mathbf{M}_x = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D}_x \end{pmatrix}, \qquad \mathbf{R} = \begin{pmatrix} \mathbf{D}_x & \mathbf{O} \\ -\mathbf{C} & \mathbf{I}_n \end{pmatrix},$$

both in $M_{2n}(F[x])$. By Theorem 8.1.1, which holds over $F[x]$ by the remark following it,

$$\mathbf{M}_x\mathbf{R} = \begin{pmatrix} \mathbf{A}\mathbf{D}_x - \mathbf{B}\mathbf{C} & \mathbf{B} \\ \mathbf{C}\mathbf{D}_x - \mathbf{D}_x\mathbf{C} & \mathbf{D}_x \end{pmatrix} = \begin{pmatrix} \mathbf{A}\mathbf{D}_x - \mathbf{B}\mathbf{C} & \mathbf{B} \\ \mathbf{O} & \mathbf{D}_x \end{pmatrix}.$$

By the remarks in Chapter 7 §5 and §8, Theorem 7.5.4 and Theorem 7.8.4 hold over the commutative ring $F[x]$. Hence $\det(\mathbf{M}_x\mathbf{R}) = \det \mathbf{M}_x \det \mathbf{R}$, with $\det \mathbf{R} = \det \mathbf{D}_x \det \mathbf{I}_n = \det \mathbf{D}_x$, and $\det(\mathbf{M}_x\mathbf{R}) = \det(\mathbf{A}\mathbf{D}_x - \mathbf{B}\mathbf{C}) \det \mathbf{D}_x$. So, in $F[x]$,

$$\det \mathbf{M}_x \cdot \det \mathbf{D}_x = \det(\mathbf{A}\mathbf{D}_x - \mathbf{B}\mathbf{C}) \cdot \det \mathbf{D}_x. \tag{$*$}$$

Now $\mathbf{D}_x = x\mathbf{I}_n - (-\mathbf{D})$, so $\det \mathbf{D}_x = p_{-\mathbf{D}}(x)$ is the characteristic polynomial of $-\mathbf{D} \in M_n(F)$ (Definition 7.10.1). By Theorem 7.10.2 it is monic of degree n, so $\det \mathbf{D}_x \neq 0$ in $F[x]$. By Corollary 0.8.4 (2), we may cancel it from $(*)$:

$$\det \mathbf{M}_x = \det(\mathbf{A}\mathbf{D}_x - \mathbf{B}\mathbf{C}) \quad \text{in } F[x].$$

Finally evaluate both sides at 0. By Lemma 8.3.7, $(\det \mathbf{M}_x)(0) = \det(\mathbf{M}_x(0)) = \det \mathbf{M}$, since $\mathbf{D}_x(0) = \mathbf{D}$; and $(\mathbf{A}\mathbf{D}_x - \mathbf{B}\mathbf{C})(0) = \mathbf{A}\mathbf{D} - \mathbf{B}\mathbf{C}$, so $(\det(\mathbf{A}\mathbf{D}_x - \mathbf{B}\mathbf{C}))(0) = \det(\mathbf{A}\mathbf{D} - \mathbf{B}\mathbf{C})$. Hence $\det \mathbf{M} = \det(\mathbf{A}\mathbf{D} - \mathbf{B}\mathbf{C})$.

■

The cancellation in $F[x]$ is the heart of the argument, so it is worth being precise about why it is legal. Over F we cannot divide by $\det \mathbf{D}$ when it is 0. Over $F[x]$ we still cannot divide, since $\det \mathbf{D}_x$ usually has no inverse, but we do not need to: cancellation only needs "$pq = pr$ and $p \neq 0$ imply $q = r$", and this holds in $F[x]$ because a product of nonzero polynomials is nonzero (Corollary 0.8.4). The perturbation $x\mathbf{I}_n$ makes $\det \mathbf{D}_x$ nonzero **as a polynomial** even when every value $\det(\mathbf{D} + c\mathbf{I}_n)$ with $c \in F$ is zero. For example, over $\mathbb{F}_2$, let $\mathbf{D} = \mathbf{E}_{22} = \operatorname{diag}(0, 1)$. Then $\det(\mathbf{D} + 0\mathbf{I}_2) = 0$ and $\det(\mathbf{D} + \mathbf{I}_2) = \det \operatorname{diag}(1, 0) = 0$, because $1 + 1 = 0$ in $\mathbb{F}_2$; no scalar shift makes $\mathbf{D}$ invertible. But $\det(\mathbf{D} + x\mathbf{I}_2) = x(x + 1) = x^2 + x$ is not the zero polynomial.

For example, over $\mathbb{Q}$, let $\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix}$, $\mathbf{B} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$, $\mathbf{D} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$, and $\mathbf{C} = 2\mathbf{I}_2 + \mathbf{D} = \begin{pmatrix} 2 & 1 \\ 0 & 2 \end{pmatrix}$, which commutes with $\mathbf{D}$. Here $\mathbf{D}$ is not invertible, so the Schur route is closed. The theorem gives $\det \mathbf{M} = \det(\mathbf{AD} - \mathbf{BC}) = \det \begin{pmatrix} -2 & 0 \\ -2 & -3 \end{pmatrix} = 6$.

Warning. **Which blocks commute decides which product appears.** The hypothesis $\mathbf{CD} = \mathbf{DC}$ gives $\det(\mathbf{AD} - \mathbf{BC})$. If instead $\mathbf{A}$ and $\mathbf{C}$ commute, the correct formula is $\det(\mathbf{AD} - \mathbf{CB})$ (Exercise (C3: When $\mathbf{A}$ and $\mathbf{C}$ commute)), with $\mathbf{C}$ and $\mathbf{B}$ in the other order. The two can differ: for the blocks $\mathbf{A} = \mathbf{I}_2$, $\mathbf{B} = \mathbf{E}_{12}$, $\mathbf{C} = \mathbf{E}_{21}$, $\mathbf{D} = \mathbf{E}_{11}$ of the warning in Section 1, $\mathbf{A}$ commutes with $\mathbf{C}$ but $\mathbf{C}$ does not commute with $\mathbf{D}$; indeed $\det \mathbf{M} = -1 = \det(\mathbf{E}_{11} - \mathbf{E}_{22}) = \det(\mathbf{AD} - \mathbf{CB})$, while $\det(\mathbf{AD} - \mathbf{BC}) = 0$.

8.3.7 Exercises

8.3.7.1 A. Check your understanding

Exercise. A1

(a) Define the Schur complement $\mathbf{M}/\mathbf{A}$, including the hypothesis on $\mathbf{A}$ and the size of $\mathbf{M}/\mathbf{A}$.

(b) State the block LDU factorization of $\mathbf{M}$ in terms of $\mathbf{A}$ and $\mathbf{M}/\mathbf{A}$.

(c) True or false: if $\mathbf{M} = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}$ is invertible and $\mathbf{A}$ is square, then $\mathbf{A}$ is invertible. Justify your answer.

(d) If $\mathbf{A}$ is invertible and $\mathbf{M}$ is invertible, which block of $\mathbf{M}^{-1}$ is $(\mathbf{M}/\mathbf{A})^{-1}$?

(e) True or false: for all $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D} \in M_n(F)$ with $\mathbf{CD} = \mathbf{DC}$, $\det \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix} = \det(\mathbf{AD} - \mathbf{BC})$, even if $\mathbf{D}$ is not invertible and $F = \mathbb{F}_2$. Justify your answer.

Solution.

(a) For $\mathbf{M} = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}$ with $\mathbf{A} \in M_k(F)$ **invertible**, $\mathbf{B} \in M_{k\times l}(F)$, $\mathbf{C} \in M_{p\times k}(F)$, $\mathbf{D} \in M_{p\times l}(F)$: $\mathbf{M}/\mathbf{A} = \mathbf{D} - \mathbf{CA}^{-1}\mathbf{B} \in M_{p\times l}(F)$, the size of $\mathbf{D}$ (Definition 8.3.1).

(b) $\mathbf{M} = \begin{pmatrix} \mathbf{I}_k & \mathbf{O} \\ \mathbf{CA}^{-1} & \mathbf{I}_p \end{pmatrix} \begin{pmatrix} \mathbf{A} & \mathbf{O} \\ \mathbf{O} & \mathbf{M}/\mathbf{A} \end{pmatrix} \begin{pmatrix} \mathbf{I}_k & \mathbf{A}^{-1}\mathbf{B} \\ \mathbf{O} & \mathbf{I}_l \end{pmatrix}$ (Theorem 8.3.1).

(c) False. $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ is invertible, but $\mathbf{A} = \mathbf{O}_1$ is not.

(d) The lower-right block, in the position of $\mathbf{D}$ (Theorem 8.3.4 (a)).

(e) True. Theorem 8.3.8 holds over every field and needs no invertibility; its proof cancels the nonzero polynomial $\det(\mathbf{D} + x\mathbf{I}_n)$ in $F[x]$.

8.3.7.2 B. Practice

Exercise. B1: A Schur complement, a determinant, and a rank

Let

$$\mathbf{M} = \left(\begin{array}{cc|cc} 2 & 1 & 1 & 0 \\ 1 & 1 & 3 & 1 \\ \hline 1 & 2 & 9 & 5 \\ 0 & 1 & 6 & 4 \end{array}\right) \in M_4(\mathbb{Q}).$$

Compute $\mathbf{M}/\mathbf{A}$ for the upper-left block $\mathbf{A}$. Hence find $\det \mathbf{M}$ and $\operatorname{rank} \mathbf{M}$, and decide whether $\mathbf{M}$ is invertible.

Solution. $\mathbf{A} = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$ has $\det \mathbf{A} = 1$, so $\mathbf{A}^{-1} = \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix}$ (Theorem 0.9.8). With $\mathbf{B} = \begin{pmatrix} 1 & 0 \\ 3 & 1 \end{pmatrix}$, $\mathbf{C} = \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix}$, $\mathbf{D} = \begin{pmatrix} 9 & 5 \\ 6 & 4 \end{pmatrix}$:

$$\mathbf{C}\mathbf{A}^{-1} = \begin{pmatrix} -1 & 3 \\ -1 & 2 \end{pmatrix}, \qquad \mathbf{C}\mathbf{A}^{-1}\mathbf{B} = \begin{pmatrix} 8 & 3 \\ 5 & 2 \end{pmatrix}, \qquad \mathbf{M}/\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 1 & 2 \end{pmatrix}.$$

By Theorem 8.3.2, $\det \mathbf{M} = \det \mathbf{A} \det(\mathbf{M}/\mathbf{A}) = 1 \cdot 0 = 0$. By Theorem 8.3.3, $\operatorname{rank} \mathbf{M} = 2 + \operatorname{rank}(\mathbf{M}/\mathbf{A}) = 2 + 1 = 3$, since $\mathbf{M}/\mathbf{A}$ is nonzero with equal rows. So $\mathbf{M}$ is not invertible (Theorem 7.5.3 (d)).

Exercise. B2: A rank-one update

Let $\mathbf{A} = \operatorname{diag}(1, 2, 4) \in M_3(\mathbb{Q})$, $\mathbf{u} = (1, 1, 1)$, and $\mathbf{v} = (1, -1, 2)$. Use Corollary 8.3.6 to find $\det(\mathbf{A} + \mathbf{u}\mathbf{v}^{\top})$ and $(\mathbf{A} + \mathbf{u}\mathbf{v}^{\top})^{-1}$, and check one entry.

Solution. $\mathbf{A}^{-1} = \operatorname{diag}(1, \tfrac{1}{2}, \tfrac{1}{4})$, so $\mathbf{A}^{-1}\mathbf{u} = (1, \tfrac{1}{2}, \tfrac{1}{4})$ and $\mathbf{v}^{\top}\mathbf{A}^{-1} = \begin{pmatrix} 1 & -\tfrac{1}{2} & \tfrac{1}{2} \end{pmatrix}$. Then $\mathbf{v}^{\top}\mathbf{A}^{-1}\mathbf{u} = 1 - \tfrac{1}{2} + \tfrac{1}{2} = 1$, so $1 + \mathbf{v}^{\top}\mathbf{A}^{-1}\mathbf{u} = 2 \neq 0$. By Corollary 8.3.6, $\mathbf{A} + \mathbf{u}\mathbf{v}^{\top}$ is invertible, $\det(\mathbf{A} + \mathbf{u}\mathbf{v}^{\top}) = 8 \cdot 2 = 16$, and

$$\begin{aligned} (\mathbf{A} + \mathbf{u}\mathbf{v}^{\top})^{-1} &= \operatorname{diag}(1, \tfrac{1}{2}, \tfrac{1}{4}) - \frac{1}{2}\begin{pmatrix} 1 \\ \tfrac{1}{2} \\ \tfrac{1}{4} \end{pmatrix}\begin{pmatrix} 1 & -\tfrac{1}{2} & \tfrac{1}{2} \end{pmatrix} \\ &= \begin{pmatrix} 1 & 0 & 0 \\ 0 & \tfrac{1}{2} & 0 \\ 0 & 0 & \tfrac{1}{4} \end{pmatrix} - \begin{pmatrix} \tfrac{1}{2} & -\tfrac{1}{4} & \tfrac{1}{4} \\ \tfrac{1}{4} & -\tfrac{1}{8} & \tfrac{1}{8} \\ \tfrac{1}{8} & -\tfrac{1}{16} & \tfrac{1}{16} \end{pmatrix} \\ &= \begin{pmatrix} \tfrac{1}{2} & \tfrac{1}{4} & -\tfrac{1}{4} \\ -\tfrac{1}{4} & \tfrac{5}{8} & -\tfrac{1}{8} \\ -\tfrac{1}{8} & \tfrac{1}{16} & \tfrac{3}{16} \end{pmatrix}. \end{aligned}$$

Check: $\mathbf{A}+\mathbf{u}\mathbf{v}^\top = \begin{pmatrix} 2 & -1 & 2 \\ 1 & 1 & 2 \\ 1 & -1 & 6 \end{pmatrix}$, and its row 1 times column 1 of the answer is $1+\frac{1}{4}-\frac{1}{4}=1$, times column 2 is $\frac{1}{2}-\frac{5}{8}+\frac{1}{8}=0$.

Exercise. B3: Identity corners

Let $\mathbf{B} \in M_{k\times l}(F)$, $\mathbf{C} \in M_{l\times k}(F)$, and $\mathbf{M} = \begin{pmatrix} \mathbf{I}_k & \mathbf{B} \\ \mathbf{C} & \mathbf{I}_l \end{pmatrix}$.

(a) Show that $\mathbf{M}$ is invertible if and only if $\mathbf{S} = \mathbf{I}_l - \mathbf{CB}$ is, and write down $\mathbf{M}^{-1}$ in terms of $\mathbf{S}^{-1}$.

(b) Deduce the inverse of $\begin{pmatrix} \mathbf{I}_k & \mathbf{B} \\ \mathbf{O} & \mathbf{I}_l \end{pmatrix}$.

(c) For $\mathbf{B} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$ and $\mathbf{C} = \begin{pmatrix} 1 & 1 \end{pmatrix}$ over $\mathbb{Q}$, compute $\mathbf{M}^{-1}$.

Solution.

(a) The block $\mathbf{I}_k$ is invertible and $\mathbf{M}/\mathbf{I}_k = \mathbf{I}_l - \mathbf{C}\mathbf{I}_k^{-1}\mathbf{B} = \mathbf{I}_l - \mathbf{CB}$. By Theorem 8.3.4 (a), $\mathbf{M}$ is invertible if and only if $\mathbf{I}_l - \mathbf{CB}$ is, and then, with $\mathbf{A}^{-1} = \mathbf{I}_k$,

$$\mathbf{M}^{-1} = \begin{pmatrix} \mathbf{I}_k + \mathbf{B}\mathbf{S}^{-1}\mathbf{C} & -\mathbf{B}\mathbf{S}^{-1} \\ -\mathbf{S}^{-1}\mathbf{C} & \mathbf{S}^{-1} \end{pmatrix}.$$

(b) With $\mathbf{C} = \mathbf{O}$, $\mathbf{S} = \mathbf{I}_l - \mathbf{CB} = \mathbf{I}_l$ and $\mathbf{S}^{-1} = \mathbf{I}_l$, so the inverse is $\begin{pmatrix} \mathbf{I}_k & -\mathbf{B} \\ \mathbf{O} & \mathbf{I}_l \end{pmatrix}$, in agreement with Theorem 8.1.5.

(c) $\mathbf{CB} = (3)$, so $\mathbf{S} = \mathbf{I}_1 - \mathbf{CB} = (-2)$ and $\mathbf{S}^{-1} = (-\frac{1}{2})$. Then $\mathbf{B}\mathbf{S}^{-1}\mathbf{C} = -\frac{1}{2}\begin{pmatrix} 1 & 1 \\ 2 & 2 \end{pmatrix}$, $-\mathbf{B}\mathbf{S}^{-1} = \begin{pmatrix} \frac{1}{2} \\ 1 \end{pmatrix}$, and $-\mathbf{S}^{-1}\mathbf{C} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \end{pmatrix}$, so

$$\mathbf{M}^{-1} = \begin{pmatrix} \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} \\ -1 & 0 & 1 \\ \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} \end{pmatrix}, \qquad \mathbf{M} = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 2 \\ 1 & 1 & 1 \end{pmatrix}.$$

Check: row 3 of $\mathbf{M}$ times column 3 of $\mathbf{M}^{-1}$ is $\frac{1}{2}+1-\frac{1}{2}=1$, and times column 1 is $\frac{1}{2}-1+\frac{1}{2}=0$.

8.3.7.3 C. Going deeper

Exercise. C1: Sylvester's determinant identity

Let $\mathbf{A} \in M_{m\times n}(F)$ and $\mathbf{B} \in M_{n\times m}(F)$ with $m, n \geq 1$.

(a) Compute both Schur complements of the identity blocks in $\mathbf{M} = \begin{pmatrix} \mathbf{I}_m & -\mathbf{A} \\ \mathbf{B} & \mathbf{I}_n \end{pmatrix}$, and deduce

$$\det(\mathbf{I}_m + \mathbf{AB}) = \det(\mathbf{I}_n + \mathbf{BA}).$$

(b) Deduce that for every nonzero $c \in F$, $\det(c\mathbf{I}_m - \mathbf{AB}) = c^{m-n}\det(c\mathbf{I}_n - \mathbf{BA})$.

(c) Deduce Theorem 7.8.5 from (a).

Solution.

(a) Both identity blocks are invertible. By Definition 8.3.1,

$$\mathbf{M}/\mathbf{I}_m = \mathbf{I}_n - \mathbf{B}\mathbf{I}_m^{-1}(-\mathbf{A}) = \mathbf{I}_n + \mathbf{BA},$$
$$\mathbf{M}/\mathbf{I}_n = \mathbf{I}_m - (-\mathbf{A})\mathbf{I}_n^{-1}\mathbf{B} = \mathbf{I}_m + \mathbf{AB}.$$

By Theorem 8.3.2 (a) and (b), $\det\mathbf{M} = \det\mathbf{I}_m \det(\mathbf{I}_n+\mathbf{BA})$ and $\det\mathbf{M} = \det\mathbf{I}_n \det(\mathbf{I}_m+\mathbf{AB})$. Hence $\det(\mathbf{I}_m + \mathbf{AB}) = \det(\mathbf{I}_n + \mathbf{BA})$.

(b) Let $\mathbf{A}' = -c^{-1}\mathbf{A}$. Then $c\mathbf{I}_m - \mathbf{AB} = c(\mathbf{I}_m + \mathbf{A}'\mathbf{B})$ and $c\mathbf{I}_n - \mathbf{BA} = c(\mathbf{I}_n + \mathbf{BA}')$. Multiplying an $m \times m$ matrix by c multiplies each of its m rows by c, so it multiplies the determinant by c^m (Theorem 7.5.1). By (a) applied to $\mathbf{A}'$ and $\mathbf{B}$,

$$\begin{aligned}\det(c\mathbf{I}_m - \mathbf{AB}) &= c^m\det(\mathbf{I}_m + \mathbf{A}'\mathbf{B}) = c^m\det(\mathbf{I}_n + \mathbf{BA}') \\ &= c^m c^{-n}\det(c\mathbf{I}_n - \mathbf{BA}) = c^{m-n}\det(c\mathbf{I}_n - \mathbf{BA}).\end{aligned}$$

(c) Let $\mathbf{u}, \mathbf{v} \in F^n$, and apply (a) with the sizes m, n there replaced by $n, 1$: $\mathbf{A} = \mathbf{u} \in M_{n\times 1}(F)$ and $\mathbf{B} = \mathbf{v}^\top \in M_{1\times n}(F)$. Then $\det(\mathbf{I}_n + \mathbf{uv}^\top) = \det(\mathbf{I}_1 + \mathbf{v}^\top\mathbf{u}) = 1 + \mathbf{v}^\top\mathbf{u}$, because the determinant of a 1×1 matrix is its entry.

Exercise. C2: Pivots as Schur complements

Let $\mathbf{A} \in M_n(F)$, and for $1 \le k \le n$ let $\mathbf{A}_k$ be its leading principal submatrix (the top-left $k \times k$ corner), as in Exercise (C1: Leading principal submatrices).

(a) Suppose $\mathbf{A}_{k-1}$ is invertible, where $2 \le k \le n$, and write $\mathbf{A}_k = \begin{pmatrix} \mathbf{A}_{k-1} & \mathbf{b} \\ \mathbf{c}^\top & d \end{pmatrix}$. Show that $s_k := \mathbf{A}_k/\mathbf{A}_{k-1}$ is the scalar $d - \mathbf{c}^\top\mathbf{A}_{k-1}^{-1}\mathbf{b}$, and that $\det\mathbf{A}_k = \det\mathbf{A}_{k-1}\cdot s_k$.

(b) Suppose $\mathbf{A}_1, \dots, \mathbf{A}_n$ are all invertible, and put $s_1 = a_{11}$. Prove that $\det\mathbf{A}_k = s_1 s_2 \cdots s_k$ for every k, and that every s_k is nonzero.

(c) Keep the hypothesis of (b). In the proof of Exercise (C1: Leading principal submatrices) (c), the last pivot is $z = d - \mathbf{x}^\top\mathbf{y}$ with $\mathbf{y} = Ł'^{-1}\mathbf{b}$ and $\mathbf{x}^\top = \mathbf{c}^\top\mathbf{U}'^{-1}$. Show that $z = \mathbf{A}_n/\mathbf{A}_{n-1}$. Deduce that the diagonal entries of $\mathbf{U}$ in the LU factorization are $u_{kk} = \det\mathbf{A}_k/\det\mathbf{A}_{k-1}$, with $\det\mathbf{A}_0 := 1$.

(d) Check (c) for $\mathbf{A} = \begin{pmatrix} 2 & 1 & 1 \\ 4 & 3 & 3 \\ 8 & 7 & 9 \end{pmatrix}$ over $\mathbb{Q}$.

Solution.

(a) With the partition $k-1,1$, Definition 8.3.1 gives the 1×1 matrix $\mathbf{A}_k/\mathbf{A}_{k-1} = d - \mathbf{c}^\top \mathbf{A}_{k-1}^{-1}\mathbf{b}$, which we identify with its entry. By Theorem 8.3.2 (a), $\det \mathbf{A}_k = \det \mathbf{A}_{k-1}\det(s_k) = \det \mathbf{A}_{k-1}\cdot s_k$.

(b) By induction on k. For $k=1$, $\det \mathbf{A}_1 = a_{11} = s_1$. If $\det \mathbf{A}_{k-1} = s_1\cdots s_{k-1}$, then by (a), which applies since $\mathbf{A}_{k-1}$ is invertible, $\det \mathbf{A}_k = s_1\cdots s_{k-1}s_k$. Since $\mathbf{A}_k$ is invertible, $\det \mathbf{A}_k \neq 0$ (Theorem 7.5.2), so no factor s_k is zero (Theorem 0.7.1 (d)).

(c) $\mathbf{A}_{n-1} = Ł'\mathbf{U}'$ with both factors invertible, so $\mathbf{A}_{n-1}^{-1} = \mathbf{U}'^{-1}Ł'^{-1}$ (Theorem 0.9.7 (3)) and $\mathbf{x}^\top\mathbf{y} = \mathbf{c}^\top\mathbf{U}'^{-1}Ł'^{-1}\mathbf{b} = \mathbf{c}^\top\mathbf{A}_{n-1}^{-1}\mathbf{b}$. Hence $z = d - \mathbf{c}^\top\mathbf{A}_{n-1}^{-1}\mathbf{b} = \mathbf{A}_n/\mathbf{A}_{n-1} = s_n$ by (a). The same construction applied to $\mathbf{A}_k$ (whose LU factors are the top-left corners $Ł_k, \mathbf{U}_k$ of $Ł, \mathbf{U}$, by Exercise (C1: Leading principal submatrices) (a) and Theorem 3.6.5) shows $u_{kk} = s_k$ for every k. By (a) and (b), $s_k = \det \mathbf{A}_k/\det \mathbf{A}_{k-1}$, with $s_1 = \det \mathbf{A}_1/1$.

(d) $\det \mathbf{A}_1 = 2$, $\det \mathbf{A}_2 = 6-4 = 2$, and expanding, $\det \mathbf{A}_3 = 2(27-21) - 1(36-24) + 1(28-24) = 12-12+4 = 4$. The predicted pivots are 2, $2/2 = 1$, $4/2 = 2$. Elimination gives $Ł = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 4 & 3 & 1 \end{pmatrix}$ and $\mathbf{U} = \begin{pmatrix} 2 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 2 \end{pmatrix}$: subtract 2 and 4 times row 1 to get rows $(0,1,1)$ and $(0,3,5)$, then 3 times the new row 2 to get $(0,0,2)$. The diagonal of $\mathbf{U}$ is $2,1,2$, as predicted.

Exercise. C3: When A and C commute

Let $n \geq 1$ and $\mathbf{A},\mathbf{B},\mathbf{C},\mathbf{D} \in M_n(F)$ with $\mathbf{AC} = \mathbf{CA}$, and $\mathbf{M} = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}$.

(a) Assume $\mathbf{A}$ is invertible. Use Theorem 8.3.2 to prove $\det \mathbf{M} = \det(\mathbf{AD} - \mathbf{CB})$.

(b) Remove the assumption that $\mathbf{A}$ is invertible, over any field, by adapting the proof of Theorem 8.3.8 with $\mathbf{A}_x = \mathbf{A} + x\mathbf{I}_n$.

(c) For $\mathbf{A} = \mathbf{I}_2$, $\mathbf{B} = \mathbf{E}_{12}$, $\mathbf{C} = \mathbf{E}_{21}$, $\mathbf{D} = \mathbf{E}_{11}$ in $M_2(\mathbb{Q})$, check that $\mathbf{AC} = \mathbf{CA}$, and compare $\det \mathbf{M}$, $\det(\mathbf{AD} - \mathbf{CB})$, and $\det(\mathbf{AD} - \mathbf{BC})$.

Hint for (b): multiply $\mathbf{M}_x$ on the left by a block matrix that clears the lower-left block.

Solution.

(a) By Theorem 8.3.2 (a) and Theorem 7.5.4,

$$\begin{aligned} \det \mathbf{M} &= \det \mathbf{A}\det(\mathbf{D} - \mathbf{CA}^{-1}\mathbf{B}) \\ &= \det\left(\mathbf{A}(\mathbf{D} - \mathbf{CA}^{-1}\mathbf{B})\right) = \det(\mathbf{AD} - \mathbf{ACA}^{-1}\mathbf{B}). \end{aligned}$$

Since $\mathbf{AC} = \mathbf{CA}$, $\mathbf{ACA}^{-1} = \mathbf{CAA}^{-1} = \mathbf{C}$, so $\det \mathbf{M} = \det(\mathbf{AD} - \mathbf{CB})$.

(b) Let $\mathbf{A}_x = \mathbf{A} + x\mathbf{I}_n$ over $F[x]$. Then $\mathbf{A}_x\mathbf{C} = \mathbf{AC} + x\mathbf{C} = \mathbf{CA} + x\mathbf{C} = \mathbf{CA}_x$. Let

$\mathbf{M}_x = \begin{pmatrix} \mathbf{A}_x & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}$ and $Ł = \begin{pmatrix} \mathbf{I}_n & \mathbf{O} \\ -\mathbf{C} & \mathbf{A}_x \end{pmatrix}$. By Theorem 8.1.1 over $F[x]$,

$$Ł\mathbf{M}_x = \begin{pmatrix} \mathbf{A}_x & \mathbf{B} \\ -\mathbf{C}\mathbf{A}_x + \mathbf{A}_x\mathbf{C} & -\mathbf{C}\mathbf{B} + \mathbf{A}_x\mathbf{D} \end{pmatrix} = \begin{pmatrix} \mathbf{A}_x & \mathbf{B} \\ \mathbf{O} & \mathbf{A}_x\mathbf{D} - \mathbf{C}\mathbf{B} \end{pmatrix}.$$

By Theorem 7.5.4 and Theorem 7.8.4 over $F[x]$ (Chapter 7 §5 and §8), $\det \mathbf{A}_x \det \mathbf{M}_x = \det \mathbf{A}_x \det(\mathbf{A}_x\mathbf{D} - \mathbf{C}\mathbf{B})$, using $\det Ł = \det \mathbf{I}_n \det \mathbf{A}_x$. As in the proof of Theorem 8.3.8, $\det \mathbf{A}_x = p_{-\mathbf{A}}(x)$ is monic, hence nonzero, and Corollary 0.8.4 (2) gives $\det \mathbf{M}_x = \det(\mathbf{A}_x\mathbf{D} - \mathbf{C}\mathbf{B})$. Evaluating at $x = 0$ with Lemma 8.3.7 gives $\det \mathbf{M} = \det(\mathbf{A}\mathbf{D} - \mathbf{C}\mathbf{B})$.

(c) $\mathbf{A}\mathbf{C} = \mathbf{C} = \mathbf{C}\mathbf{A}$, since $\mathbf{A} = \mathbf{I}_2$. As computed in Section 1, $\det \mathbf{M} = -1$. Also $\mathbf{A}\mathbf{D} - \mathbf{C}\mathbf{B} = \mathbf{E}_{11} - \mathbf{E}_{21}\mathbf{E}_{12} = \mathbf{E}_{11} - \mathbf{E}_{22} = \operatorname{diag}(1, -1)$, with determinant -1, in agreement with (b). But $\mathbf{A}\mathbf{D} - \mathbf{B}\mathbf{C} = \mathbf{E}_{11} - \mathbf{E}_{12}\mathbf{E}_{21} = \mathbf{E}_{11} - \mathbf{E}_{11} = \mathbf{O}$, with determinant 0. So with $\mathbf{A}\mathbf{C} = \mathbf{C}\mathbf{A}$ the product must be taken as $\mathbf{C}\mathbf{B}$, not $\mathbf{B}\mathbf{C}$; here $\mathbf{C}\mathbf{D} = \mathbf{E}_{21} \neq \mathbf{O} = \mathbf{D}\mathbf{C}$, so Theorem 8.3.8 does not apply.

8.4 The Kronecker Product

So far the unknown in a matrix equation has been a vector: $\mathbf{A}\mathbf{x} = \mathbf{b}$. Many natural questions have a matrix as the unknown instead. Which matrices commute with a given $\mathbf{A}$? Which $\mathbf{X}$ solve $\mathbf{A}\mathbf{X} + \mathbf{X}\mathbf{B} = \mathbf{C}$? Such an equation is still linear in the entries of $\mathbf{X}$, so in principle it is an ordinary linear system. The trouble is bookkeeping: we need to know what the coefficient matrix of that system is. This section builds it from $\mathbf{A}$ and $\mathbf{B}$ with one operation, the Kronecker product, and turns matrix equations into systems we already know how to solve.

8.4.1 From a matrix equation to a linear system

Take $\mathbf{A}, \mathbf{B} \in M_2(F)$ and an unknown $\mathbf{X} = (x_{ij}) \in M_2(F)$, and look at the equation $\mathbf{A}\mathbf{X}\mathbf{B} = \mathbf{C}$. By Definition 0.9.9, applied twice,

$$(\mathbf{A}\mathbf{X}\mathbf{B})_{rs} = \sum_{i=1}^{2} \sum_{j=1}^{2} a_{ri}\, x_{ij}\, b_{js}.$$

So each entry of $\mathbf{A}\mathbf{X}\mathbf{B}$ is a linear combination of the four unknowns $x_{11}, x_{21}, x_{12}, x_{22}$, and the coefficient of x_{ij} in entry (r, s) is the product $b_{js}a_{ri}$: **an entry of B times an entry of A**. Writing the four equations as a 4×4 system, every coefficient is such a product. For instance, if we list the unknowns column by column as $(x_{11}, x_{21}, x_{12}, x_{22})$ and the equations in the same order, the first row of the coefficient matrix is

$$\begin{pmatrix} b_{11}a_{11} & b_{11}a_{12} & b_{21}a_{11} & b_{21}a_{12} \end{pmatrix}.$$

The left half of this row is b_{11} times row 1 of $\mathbf{A}$, and the right half is b_{21} times row 1 of $\mathbf{A}$.

That is a recurring pattern: a big matrix whose entries are all the products of the entries of two small matrices, arranged in blocks. It deserves a name.

8.4.2 The Kronecker product

The Kronecker product of $\mathbf{A}$ *and* $\mathbf{B}$ *is the block matrix obtained by replacing each entry* a_{ij} *of* $\mathbf{A}$ *by the block* $a_{ij}\mathbf{B}$.

Definition 8.4.1. Kronecker product

Let $\mathbf{A} = (a_{ij}) \in M_{m\times n}(F)$ and $\mathbf{B} = (b_{kl}) \in M_{p\times q}(F)$, of **any** sizes. The **Kronecker product** of $\mathbf{A}$ and $\mathbf{B}$ is the $mp \times nq$ block matrix

$$\mathbf{A}\otimes\mathbf{B} \coloneqq \begin{pmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} & \cdots & a_{1n}\mathbf{B} \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} & \cdots & a_{2n}\mathbf{B} \\ \vdots & \vdots & & \vdots \\ a_{m1}\mathbf{B} & a_{m2}\mathbf{B} & \cdots & a_{mn}\mathbf{B} \end{pmatrix} \in M_{mp\times nq}(F),$$

with m rows and n columns of blocks, each block of size $p \times q$ (Definition 8.1.1).

In words: the outer layout is the layout of $\mathbf{A}$, and the inner pattern of every block is $\mathbf{B}$. The sizes multiply, rows with rows and columns with columns. No condition on the sizes is needed, unlike for an ordinary product.

To work with single entries we need to know where each product $a_{ij}b_{kl}$ sits. Block (i, j) occupies rows $(i-1)p+1, \ldots, ip$ and columns $(j-1)q+1, \ldots, jq$. Inside it, entry (k, l) is $a_{ij}b_{kl}$. So

$$(\mathbf{A}\otimes\mathbf{B})_{(i-1)p+k,\ (j-1)q+l} = a_{ij}\,b_{kl} \qquad (1\le i\le m,\ 1\le j\le n,\ 1\le k\le p,\ 1\le l\le q). \tag{$*$}$$

Every entry is reached exactly once. Division with remainder writes each row index $r \in \{1, \ldots, mp\}$ uniquely as $r = (i-1)p + k$ with $1 \le i \le m$ and $1 \le k \le p$ (divide $r-1$ by p), and likewise for columns. So $(*)$ is a complete description of $\mathbf{A}\otimes\mathbf{B}$, and it is the formula we use in proofs. In words, the rows of $\mathbf{A}\otimes\mathbf{B}$ are labeled by pairs (row of $\mathbf{A}$, row of $\mathbf{B}$) in dictionary order, and the columns by pairs (column of $\mathbf{A}$, column of $\mathbf{B}$) in dictionary order.

Examples.

1. **$\mathbf{A}$ 2×2 example.** With $\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$ and $\mathbf{B} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$,

$$\mathbf{A}\otimes\mathbf{B} = \begin{pmatrix} 1\mathbf{B} & 2\mathbf{B} \\ 3\mathbf{B} & 4\mathbf{B} \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 & 2 \\ 1 & 0 & 2 & 0 \\ 0 & 3 & 0 & 4 \\ 3 & 0 & 4 & 0 \end{pmatrix}.$$

   Check one entry against $(*)$: row $3 = (2-1)\cdot 2 + 1$ and column $2 = (1-1)\cdot 2 + 2$ give $a_{21}b_{12} = 3 \cdot 1 = 3$, as displayed.
2. **Vectors.** For column vectors $\mathbf{u} \in F^m$ and $\mathbf{v} \in F^p$ (so $n = q = 1$), $\mathbf{u}\otimes\mathbf{v} \in F^{mp}$ lists all products u_iv_k in dictionary order. In particular, for standard basis vectors, $\mathbf{e}_i \otimes \mathbf{e}_k = \mathbf{e}_{(i-1)p+k} \in F^{mp}$: the only nonzero product is $1 \cdot 1$ in position $(i-1)p+k$. For a column $\mathbf{u} \in F^m$ and a row $\mathbf{v}^\top$, $\mathbf{u}\otimes\mathbf{v}^\top$ is the $m \times p$ matrix with entries u_iv_k, which is the outer product $\mathbf{u}\mathbf{v}^\top$ of Corollary 8.1.3.
3. **Identity blocks.** $\mathbf{I}_n \otimes \mathbf{B}$ has $\mathbf{B}$ in each diagonal block and zero blocks elsewhere, so it is the block diagonal matrix $\mathbf{B}\oplus\cdots\oplus\mathbf{B}$ with n copies of $\mathbf{B}$ (Definition 8.1.2). By contrast, $\mathbf{A}\otimes\mathbf{I}_m$ replaces each entry a_{ij} by the $m \times m$ block $a_{ij}\mathbf{I}_m$. Taking both, $\mathbf{I}_n \otimes \mathbf{I}_m = \mathbf{I}_{nm}$: by $(*)$ its entries are $\delta_{ij}\delta_{kl}$, which is 1 exactly when the row and column labels (i, k) and (j, l) agree.
4. **The degenerate case 1×1.** If $\mathbf{A} = (a)$ is 1×1, then $\mathbf{A}\otimes\mathbf{B} = a\mathbf{B}$; if $\mathbf{B} = (b)$, then $\mathbf{A}\otimes\mathbf{B} = b\mathbf{A}$. So a 1×1 factor behaves like a scalar, and $(1)\otimes\mathbf{B} = \mathbf{B}$. This case is worth

keeping in mind: it shows that the Kronecker product extends scalar multiplication, which is why it is bilinear below.

Nonexample by minimal change. Swap the two factors of Example 1:

$$\mathbf{B}\otimes\mathbf{A}=\begin{pmatrix}0\mathbf{A} & 1\mathbf{A}\\ 1\mathbf{A} & 0\mathbf{A}\end{pmatrix}=\begin{pmatrix}0&0&1&2\\0&0&3&4\\1&2&0&0\\3&4&0&0\end{pmatrix}.$$

It has the same size and the same multiset of entries, since both list all products $a_{ij}b_{kl}$. What fails is the position formula $(*)$: in $\mathbf{B}\otimes\mathbf{A}$ the **outer** layout comes from $\mathbf{B}$, so the products sit in different places. For instance, the $(1,2)$-entry is $a_{11}b_{12}=1$ in $\mathbf{A}\otimes\mathbf{B}$ but $b_{11}a_{12}=0$ in $\mathbf{B}\otimes\mathbf{A}$.

Why this definition. The dictionary order in $(*)$ is a convention, and the other order would simply give $\mathbf{B}\otimes\mathbf{A}$. We fix it because it makes the block picture true: the Kronecker product is literally "$\mathbf{A}$ with entries replaced by blocks", so block multiplication (Theorem 8.1.1) applies to it directly. The symbol $\otimes$ anticipates the tensor product of Chapter 15, of which this is the matrix version.

Warning. **The Kronecker product is not commutative, and it is not the ordinary product.** The two matrices in Example 1 and the nonexample differ, so $\mathbf{A}\otimes\mathbf{B}\neq\mathbf{B}\otimes\mathbf{A}$ in general, even for square matrices of the same size. Also do not confuse $\mathbf{A}\otimes\mathbf{B}$ with $\mathbf{AB}$: for $\mathbf{A},\mathbf{B}\in M_2(F)$, $\mathbf{AB}$ is 2×2 while $\mathbf{A}\otimes\mathbf{B}$ is 4×4, and $\mathbf{A}\otimes\mathbf{B}$ is defined for all sizes. We will see below that $\mathbf{A}\otimes\mathbf{B}$ and $\mathbf{B}\otimes\mathbf{A}$ are always **similar** when $\mathbf{A}$ and $\mathbf{B}$ are square, so they share rank, trace, and determinant, but they are not equal.

8.4.3 Algebraic properties

The Kronecker product interacts well with every operation we have. The most useful rule is the one connecting it to ordinary multiplication: multiply the two left factors together and the two right factors together.

Theorem 8.4.1. Properties of the Kronecker Product

Let $c\in F$. Whenever the sizes make each side defined:

(a) **(Bilinearity)** $(\mathbf{A}+\mathbf{A}')\otimes\mathbf{B}=\mathbf{A}\otimes\mathbf{B}+\mathbf{A}'\otimes\mathbf{B}$, $\mathbf{A}\otimes(\mathbf{B}+\mathbf{B}')=\mathbf{A}\otimes\mathbf{B}+\mathbf{A}\otimes\mathbf{B}'$, and $(c\mathbf{A})\otimes\mathbf{B}=c(\mathbf{A}\otimes\mathbf{B})=\mathbf{A}\otimes(c\mathbf{B})$.

(b) **(Associativity)** $(\mathbf{A}\otimes\mathbf{B})\otimes\mathbf{C}=\mathbf{A}\otimes(\mathbf{B}\otimes\mathbf{C})$ for all $\mathbf{A}\in M_{m\times n}(F)$, $\mathbf{B}\in M_{p\times q}(F)$, $\mathbf{C}\in M_{r\times s}(F)$.

(c) **(Mixed product rule)** For $\mathbf{A}\in M_{m\times n}(F)$, $\mathbf{C}\in M_{n\times t}(F)$, $\mathbf{B}\in M_{p\times q}(F)$, and $\mathbf{D}\in M_{q\times u}(F)$,
$$(\mathbf{A}\otimes\mathbf{B})(\mathbf{C}\otimes\mathbf{D})=(\mathbf{AC})\otimes(\mathbf{BD}).$$

(d) **(Transpose)** $(\mathbf{A}\otimes\mathbf{B})^{\top}=\mathbf{A}^{\top}\otimes\mathbf{B}^{\top}$.

(e) **(Inverse)** If $\mathbf{A}\in M_n(F)$ and $\mathbf{B}\in M_m(F)$ are invertible, then $\mathbf{A}\otimes\mathbf{B}$ is invertible and $(\mathbf{A}\otimes\mathbf{B})^{-1}=\mathbf{A}^{-1}\otimes\mathbf{B}^{-1}$.

Idea. Parts (a), (b), and (d) are entry checks with $(*)$. For (c), both factors on the left are block matrices, and the columns of blocks of $\mathbf{A}\otimes\mathbf{B}$ (n of them, each q wide) match the rows of blocks of $\mathbf{C}\otimes\mathbf{D}$ (n of them, each q tall). So we may multiply blockwise, and each block of the product is a sum $\sum_j (a_{ij}\mathbf{B})(c_{jk}\mathbf{D})$ in which the scalars come out. Part (e) is (c) with $\mathbf{C}=\mathbf{A}^{-1}$, $\mathbf{D}=\mathbf{B}^{-1}$.

Proof.

(a) By $(*)$, the entry of $(\mathbf{A}+\mathbf{A}')\otimes\mathbf{B}$ in position $((i-1)p+k,(j-1)q+l)$ is $(a_{ij}+a'_{ij})b_{kl} = a_{ij}b_{kl}+a'_{ij}b_{kl}$, which is the entry of $\mathbf{A}\otimes\mathbf{B}+\mathbf{A}'\otimes\mathbf{B}$ in the same position. The other identities are checked the same way, using $(ca_{ij})b_{kl} = c(a_{ij}b_{kl}) = a_{ij}(cb_{kl})$.

(b) Both sides are $mpr\times nqs$. Let $1\le i\le m$, $1\le k\le p$, $1\le t\le r$, and similarly j,l,u for columns. Applying $(*)$ twice, the entry of $(\mathbf{A}\otimes\mathbf{B})\otimes\mathbf{C}$ in row $((i-1)p+k-1)r+t$ and column $((j-1)q+l-1)s+u$ is $(\mathbf{A}\otimes\mathbf{B})_{(i-1)p+k,\,(j-1)q+l}\,c_{tu} = a_{ij}b_{kl}c_{tu}$. Likewise the entry of $\mathbf{A}\otimes(\mathbf{B}\otimes\mathbf{C})$ in row $(i-1)pr+(k-1)r+t$ and column $(j-1)qs+(l-1)s+u$ is $a_{ij}(\mathbf{B}\otimes\mathbf{C})_{(k-1)r+t,\,(l-1)s+u} = a_{ij}b_{kl}c_{tu}$. Since $((i-1)p+k-1)r+t = (i-1)pr+(k-1)r+t$, and similarly for columns, the two matrices have the same entry in every position; every position arises from exactly one choice of indices, by division with remainder. Hence they are equal.

(c) Partition $\mathbf{A}\otimes\mathbf{B}$ into m rows and n columns of blocks as in Definition 8.4.1, each block $p\times q$, and $\mathbf{C}\otimes\mathbf{D}$ into n rows and t columns of blocks, each $q\times u$. The column partition of the first matrix (n parts of size q) equals the row partition of the second, so the partitions are conformable, and by Theorem 8.1.1 the (i,k) block of the product is

$$\sum_{j=1}^{n}(a_{ij}\mathbf{B})(c_{jk}\mathbf{D}) = \sum_{j=1}^{n} a_{ij}c_{jk}\,\mathbf{BD} = (\mathbf{AC})_{ik}\,\mathbf{BD},$$

where the first equality uses Theorem 0.9.4 (4) and the second uses Definition 0.9.9. This is the (i,k) block of $(\mathbf{AC})\otimes(\mathbf{BD})$, a matrix partitioned into m rows and t columns of $p\times u$ blocks. So the two matrices agree block by block, hence entry by entry.

(d) $(\mathbf{A}\otimes\mathbf{B})^{\top}$ is $nq\times mp$. By Definition 0.9.10 and $(*)$, its entry in row $(j-1)q+l$ and column $(i-1)p+k$ is $(\mathbf{A}\otimes\mathbf{B})_{(i-1)p+k,\,(j-1)q+l} = a_{ij}b_{kl} = (\mathbf{A}^{\top})_{ji}(\mathbf{B}^{\top})_{lk}$, which by $(*)$ is the entry of $\mathbf{A}^{\top}\otimes\mathbf{B}^{\top}$ in the same position.

(e) By (c) and Example 3, $(\mathbf{A}\otimes\mathbf{B})(\mathbf{A}^{-1}\otimes\mathbf{B}^{-1}) = (\mathbf{AA}^{-1})\otimes(\mathbf{BB}^{-1}) = \mathbf{I}_n\otimes\mathbf{I}_m = \mathbf{I}_{nm}$, and similarly $(\mathbf{A}^{-1}\otimes\mathbf{B}^{-1})(\mathbf{A}\otimes\mathbf{B}) = \mathbf{I}_{nm}$. By Definition 0.9.13, $\mathbf{A}\otimes\mathbf{B}$ is invertible with inverse $\mathbf{A}^{-1}\otimes\mathbf{B}^{-1}$.

■

The mixed product rule is the engine of everything that follows. One immediate consequence splits a Kronecker product of square matrices into two simpler factors: for $\mathbf{A}\in M_n(F)$ and $\mathbf{B}\in M_m(F)$,

$$\mathbf{A}\otimes\mathbf{B} = (\mathbf{A}\mathbf{I}_n)\otimes(\mathbf{I}_m\mathbf{B}) = (\mathbf{A}\otimes\mathbf{I}_m)(\mathbf{I}_n\otimes\mathbf{B}), \qquad (**)$$

and in the same way $\mathbf{A}\otimes\mathbf{B} = (\mathbf{I}_n\otimes\mathbf{B})(\mathbf{A}\otimes\mathbf{I}_m)$. So $\mathbf{A}\otimes\mathbf{I}_m$ and $\mathbf{I}_n\otimes\mathbf{B}$ commute, even though $\mathbf{A}$ and $\mathbf{B}$ have nothing to do with each other.

For example, with $\mathbf{A} = \begin{pmatrix}1 & -1\\ 0 & 2\end{pmatrix}$, $\mathbf{B} = \begin{pmatrix}2 & 1\\ 1 & 1\end{pmatrix}$, $\mathbf{C} = \begin{pmatrix}1 & 0\\ 1 & 1\end{pmatrix}$, and $\mathbf{D} = \begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix}$, multiplying

two 4×4 matrices directly gives

$$(\mathbf{A}\otimes\mathbf{B})(\mathbf{C}\otimes\mathbf{D}) = \begin{pmatrix} 2 & 1 & -2 & -1 \\ 1 & 1 & -1 & -1 \\ 0 & 0 & 4 & 2 \\ 0 & 0 & 2 & 2 \end{pmatrix}\begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 & -1 & -2 \\ 0 & 0 & -1 & -1 \\ 2 & 4 & 2 & 4 \\ 2 & 2 & 2 & 2 \end{pmatrix},$$

while the rule needs only two 2×2 products, $\mathbf{AC} = \begin{pmatrix} 0 & -1 \\ 2 & 2 \end{pmatrix}$ and $\mathbf{BD} = \begin{pmatrix} 1 & 2 \\ 1 & 1 \end{pmatrix}$, whose Kronecker product is the same matrix.

8.4.4 Swapping the factors

The warning said that $\mathbf{A}\otimes\mathbf{B}$ and $\mathbf{B}\otimes\mathbf{A}$ differ. They differ only in the order in which rows and columns are labeled: $\mathbf{A}\otimes\mathbf{B}$ uses (index of $\mathbf{A}$, index of $\mathbf{B}$) in dictionary order, $\mathbf{B}\otimes\mathbf{A}$ the reverse order. Relabeling rows and columns by the same permutation is conjugation by a permutation matrix, and that is what the next result makes precise.

First, a fact about permutation matrices. Recall that for $\sigma\in S_N$, the permutation matrix $\mathbf{P}_\sigma$ has j-th column $\mathbf{e}_{\sigma(j)}$, and $\mathbf{P}_\sigma^{-1}=\mathbf{P}_\sigma^{\top}$ (Definition 3.4.2, Lemma 3.4.7). For every $\mathbf{M}\in M_N(F)$,

$$(\mathbf{P}_\sigma^{\top}\mathbf{M}\mathbf{P}_\sigma)_{rs} = \mathbf{M}_{\sigma(r)\,\sigma(s)} \qquad (1\le r,s\le N). \tag{$\dagger$}$$

Indeed, by Theorem 0.9.2 column s of $\mathbf{M}\mathbf{P}_\sigma$ is $\mathbf{M}\mathbf{e}_{\sigma(s)}$, which is column $\sigma(s)$ of $\mathbf{M}$. Row r of $\mathbf{P}_\sigma^{\top}$ is $\mathbf{e}_{\sigma(r)}^{\top}$, so by the same theorem row r of $\mathbf{P}_\sigma^{\top}Ł$ is row $\sigma(r)$ of Ł, for any $Ł\in M_N(F)$. Combining the two gives ($\dagger$).

Proposition 8.4.2. Swapping the Factors Is a Permutation Similarity

Let $n,m\ge 1$. There is a permutation matrix $\mathbf{P}\in M_{nm}(F)$, depending only on n and m, such that

$$\mathbf{B}\otimes\mathbf{A} = \mathbf{P}^{-1}(\mathbf{A}\otimes\mathbf{B})\mathbf{P} \qquad \text{for all } \mathbf{A}\in M_n(F),\ \mathbf{B}\in M_m(F).$$

In particular $\mathbf{A}\otimes\mathbf{B}\sim\mathbf{B}\otimes\mathbf{A}$, and $\mathbf{I}_m\otimes\mathbf{A}=\mathbf{P}^{-1}(\mathbf{A}\otimes\mathbf{I}_m)\mathbf{P}$.

Proof. Every $r\in\{1,\dots,nm\}$ can be written uniquely as $r=(k-1)n+i$ with $1\le k\le m$ and $1\le i\le n$. Define $\sigma(r)\coloneqq(i-1)m+k$. As (k,i) runs over all pairs, so does (i,k), and each value $(i-1)m+k$ is taken by exactly one pair; hence σ is a bijection of $\{1,\dots,nm\}$, that is, $\sigma\in S_{nm}$. Put $\mathbf{P}=\mathbf{P}_\sigma$.

Let $r=(k-1)n+i$ and $s=(l-1)n+j$. By $(*)$ applied to $\mathbf{B}\otimes\mathbf{A}$, $(\mathbf{B}\otimes\mathbf{A})_{rs}=b_{kl}a_{ij}$. By $(*)$ applied to $\mathbf{A}\otimes\mathbf{B}$,

$$(\mathbf{A}\otimes\mathbf{B})_{\sigma(r)\,\sigma(s)} = (\mathbf{A}\otimes\mathbf{B})_{(i-1)m+k,\ (j-1)m+l} = a_{ij}b_{kl}.$$

So by ($\dagger$), $\mathbf{P}_\sigma^{\top}(\mathbf{A}\otimes\mathbf{B})\mathbf{P}_\sigma$ and $\mathbf{B}\otimes\mathbf{A}$ agree in every entry, and $\mathbf{P}_\sigma^{\top}=\mathbf{P}_\sigma^{-1}$ by Lemma 3.4.7 (b). The last claim is the case $\mathbf{B}=\mathbf{I}_m$. ∎

For example, when $n=m=2$, σ fixes 1 and 4 and swaps 2 and 3, so $\mathbf{P}$ is the elementary matrix $\mathbf{P}_{23}$: conjugating by it swaps rows $2,3$ and columns $2,3$. Doing this to $\mathbf{A}\otimes\mathbf{B}$ in Example 1 does

give $\mathbf{B} \otimes \mathbf{A}$, as a direct check shows. For rectangular $\mathbf{A} \in M_{m\times n}(F)$ and $\mathbf{B} \in M_{p\times q}(F)$, the same argument gives $\mathbf{B} \otimes \mathbf{A} = \mathbf{P}_\sigma^\top (\mathbf{A} \otimes \mathbf{B})\mathbf{P}_\tau$ with one permutation for the rows and another for the columns.

8.4.5 Rank, trace, and determinant

Each of the three basic numbers attached to a matrix behaves multiplicatively. The formulas for trace and rank are what one would guess. The determinant is subtler, because the sizes enter as exponents.

Theorem 8.4.3. Rank, Trace, and Determinant of a Kronecker Product

(a) For $\mathbf{A} \in M_{m\times n}(F)$ and $\mathbf{B} \in M_{p\times q}(F)$, $\operatorname{rank}(\mathbf{A} \otimes \mathbf{B}) = \operatorname{rank}\mathbf{A} \cdot \operatorname{rank}\mathbf{B}$.

(b) For $\mathbf{A} \in M_n(F)$ and $\mathbf{B} \in M_m(F)$, $\operatorname{tr}(\mathbf{A} \otimes \mathbf{B}) = \operatorname{tr}\mathbf{A} \cdot \operatorname{tr}\mathbf{B}$.

(c) For $\mathbf{A} \in M_n(F)$ and $\mathbf{B} \in M_m(F)$, $\det(\mathbf{A} \otimes \mathbf{B}) = (\det \mathbf{A})^m (\det \mathbf{B})^n$.

Idea. For rank, bring both factors to rank normal form, $\mathbf{A} = \mathbf{Q}\mathbf{N}_r\mathbf{P}$ and $\mathbf{B} = \mathbf{Q}'\mathbf{N}_s\mathbf{P}'$. The mixed product rule moves the invertible factors outside, where they do not change rank, and $\mathbf{N}_r \otimes \mathbf{N}_s$ visibly has rank rs. Trace is a sum over diagonal entries, and the diagonal entries of $\mathbf{A} \otimes \mathbf{B}$ are the products $a_{ii}b_{kk}$. For the determinant, split $\mathbf{A} \otimes \mathbf{B} = (\mathbf{A} \otimes \mathbf{I}_m)(\mathbf{I}_n \otimes \mathbf{B})$ by $(**)$. The second factor is block diagonal with n copies of $\mathbf{B}$. The first is not, but by Proposition 8.4.2 it is similar to $\mathbf{I}_m \otimes \mathbf{A}$, which is block diagonal with m copies of $\mathbf{A}$.

Proof.

(a) Let $r = \operatorname{rank}\mathbf{A}$ and $s = \operatorname{rank}\mathbf{B}$. By Theorem 4.3.3 (b) and Definition 4.3.1, there are invertible $\mathbf{Q}_1, \mathbf{P}_1, \mathbf{Q}_2, \mathbf{P}_2$ with $\mathbf{N}_r = \mathbf{Q}_1\mathbf{A}\mathbf{P}_1$ and $\mathbf{N}_s = \mathbf{Q}_2\mathbf{B}\mathbf{P}_2$, where $\mathbf{N}_r \in M_{m\times n}(F)$ and $\mathbf{N}_s \in M_{p\times q}(F)$ have $\mathbf{I}_r$, $\mathbf{I}_s$ in the top left corner and zeros elsewhere. Hence $\mathbf{A} = \mathbf{Q}_1^{-1}\mathbf{N}_r\mathbf{P}_1^{-1}$ and $\mathbf{B} = \mathbf{Q}_2^{-1}\mathbf{N}_s\mathbf{P}_2^{-1}$, and by Theorem 8.4.1 (c), used twice,

$$\mathbf{A} \otimes \mathbf{B} = (\mathbf{Q}_1^{-1} \otimes \mathbf{Q}_2^{-1})(\mathbf{N}_r \otimes \mathbf{N}_s)(\mathbf{P}_1^{-1} \otimes \mathbf{P}_2^{-1}).$$

The outer factors are invertible by Theorem 8.4.1 (e), so $\operatorname{rank}(\mathbf{A}\otimes\mathbf{B}) = \operatorname{rank}(\mathbf{N}_r \otimes \mathbf{N}_s)$ by Theorem 3.5.9.
By $(*)$, the entry of $\mathbf{N}_r \otimes \mathbf{N}_s$ in position $((i-1)p+k, (j-1)q+l)$ is 1 if $i = j \le r$ and $k = l \le s$, and 0 otherwise. So for each of the rs pairs (i,k) with $i \le r$, $k \le s$, column $(i-1)q+k$ equals $\mathbf{e}_{(i-1)p+k}$, and every other column is zero. Distinct pairs give distinct standard basis vectors, so these rs columns are linearly independent and span the column space. Hence $\operatorname{rank}(\mathbf{N}_r \otimes \mathbf{N}_s) = rs$ by Definition 3.5.4, which proves (a). (If $r = 0$ or $s = 0$, the matrix is zero and both sides are 0.)

(b) By $(*)$ with $p = q = m$, the diagonal entries of $\mathbf{A} \otimes \mathbf{B}$ are those with $(i-1)m + k = (j-1)m + l$, that is $i = j$ and $k = l$ by uniqueness of division with remainder, and they equal $a_{ii}b_{kk}$. By Definition 0.9.12,

$$\operatorname{tr}(\mathbf{A} \otimes \mathbf{B}) = \sum_{i=1}^{n}\sum_{k=1}^{m} a_{ii}b_{kk} = \Big(\sum_{i=1}^{n} a_{ii}\Big)\Big(\sum_{k=1}^{m} b_{kk}\Big) = \operatorname{tr}\mathbf{A} \cdot \operatorname{tr}\mathbf{B}.$$

(c) First, a block diagonal matrix $\mathbf{M}_1 \oplus \cdots \oplus \mathbf{M}_t$ with square blocks $\mathbf{M}_1, \dots, \mathbf{M}_t$ has determinant $\det\mathbf{M}_1 \cdots \det\mathbf{M}_t$. For $t = 1$ there is nothing to prove, and for $t \ge 2$ the

matrix is $\mathbf{M}_1 \oplus (\mathbf{M}_2 \oplus \cdots \oplus \mathbf{M}_t)$, so Theorem 8.1.4 (d) (or Theorem 7.8.4) and induction on t give the claim.

By Example 3, $\mathbf{I}_n \otimes \mathbf{B} = \mathbf{B} \oplus \cdots \oplus \mathbf{B}$ with n blocks, so $\det(\mathbf{I}_n \otimes \mathbf{B}) = (\det \mathbf{B})^n$. Likewise $\mathbf{I}_m \otimes \mathbf{A} = \mathbf{A} \oplus \cdots \oplus \mathbf{A}$ with m blocks, so $\det(\mathbf{I}_m \otimes \mathbf{A}) = (\det \mathbf{A})^m$. By Proposition 8.4.2, $\mathbf{A} \otimes \mathbf{I}_m$ is similar to $\mathbf{I}_m \otimes \mathbf{A}$, so $\det(\mathbf{A} \otimes \mathbf{I}_m) = (\det \mathbf{A})^m$ by Corollary 7.5.7. Finally, by $(**)$ and Theorem 7.5.4,

$$\det(\mathbf{A} \otimes \mathbf{B}) = \det(\mathbf{A} \otimes \mathbf{I}_m)\det(\mathbf{I}_n \otimes \mathbf{B}) = (\det \mathbf{A})^m (\det \mathbf{B})^n.$$

■

The exponents are easy to get backwards, and a size count fixes them. In $\mathbf{A} \otimes \mathbf{B}$ with $\mathbf{A} \in M_n(F)$, $\mathbf{B} \in M_m(F)$, the matrix $\mathbf{A}$ is "used" m times (once for each index of $\mathbf{B}$), so $\det \mathbf{A}$ appears to the power m, the size of the **other** factor. For $\mathbf{A} = c\mathbf{I}_n$ and $\mathbf{B} = \mathbf{I}_m$, $\mathbf{A} \otimes \mathbf{B} = c\mathbf{I}_{nm}$ has determinant $c^{nm} = (c^n)^m = (\det \mathbf{A})^m$, which confirms the exponent.

For the matrices of the example after Theorem 8.4.1, $\det \mathbf{A} = 2$, $\det \mathbf{B} = 1$, $\operatorname{tr} \mathbf{A} = 3$, and $\operatorname{tr} \mathbf{B} = 3$, so $\det(\mathbf{A} \otimes \mathbf{B}) = 2^2 \cdot 1^2 = 4$, $\operatorname{tr}(\mathbf{A} \otimes \mathbf{B}) = 9$, and $\operatorname{rank}(\mathbf{A} \otimes \mathbf{B}) = 2 \cdot 2 = 4$. Directly, the diagonal of $\mathbf{A} \otimes \mathbf{B}$ is $2, 1, 4, 2$, with sum 9.

Quick check. Let $\mathbf{A} \in M_2(\mathbb{R})$ have $\operatorname{tr} \mathbf{A} = 4$, $\det \mathbf{A} = 3$, and let $\mathbf{B} \in M_3(\mathbb{R})$ have $\operatorname{tr} \mathbf{B} = 2$, $\det \mathbf{B} = -1$. Find $\operatorname{tr}(\mathbf{A} \otimes \mathbf{B})$, $\det(\mathbf{A} \otimes \mathbf{B})$, and $\det(\mathbf{B} \otimes \mathbf{A})$.

Solution. By Theorem 8.4.3, $\operatorname{tr}(\mathbf{A} \otimes \mathbf{B}) = 4 \cdot 2 = 8$ and $\det(\mathbf{A} \otimes \mathbf{B}) = (\det \mathbf{A})^3 (\det \mathbf{B})^2 = 27 \cdot 1 = 27$. With the roles swapped, $\det(\mathbf{B} \otimes \mathbf{A}) = (\det \mathbf{B})^2 (\det \mathbf{A})^3 = 27$, as it must be since $\mathbf{B} \otimes \mathbf{A} \sim \mathbf{A} \otimes \mathbf{B}$ (Proposition 8.4.2). The common wrong answer $(\det \mathbf{A})^2 (\det \mathbf{B})^3 = -9$ puts each matrix's own size in its exponent.

8.4.6 Vectorization

We return to the question of the opening: turning a matrix equation into a linear system. For that we need to flatten the unknown matrix into a column, in an order that matches $(*)$.

To vectorize a matrix is to stack its columns into one long column.

Definition 8.4.2. Vectorization

Let $\mathbf{X} = (x_{ij}) \in M_{m \times n}(F)$ have columns $\mathbf{x}_1, \ldots, \mathbf{x}_n \in F^m$. Its **vectorization** is the column vector

$$\operatorname{vec} \mathbf{X} := \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_n \end{pmatrix} \in F^{mn},$$

whose entry in position $(j-1)m + i$ is x_{ij}, for $1 \le i \le m$, $1 \le j \le n$.

For example, $\operatorname{vec}\begin{pmatrix}1 & 2\\ 3 & 4\end{pmatrix} = (1,3,2,4)$, read down the first column and then down the second. For a column $\mathbf{x} \in F^m$ (the case $n = 1$), $\operatorname{vec}\mathbf{x} = \mathbf{x}$. The map $\operatorname{vec}\colon M_{m\times n}(F) \to F^{mn}$ is linear, since each entry of $\operatorname{vec}\mathbf{X}$ is an entry of $\mathbf{X}$, and bijective, since it only rearranges entries; so it is an isomorphism. It sends the matrix unit $\mathbf{E}_{ij}$ to $\mathbf{e}_{(j-1)m+i} = \mathbf{e}_j \otimes \mathbf{e}_i$ (Example 2, with the index of the column first). More generally, $\operatorname{vec}(\mathbf{u}\mathbf{v}^\top) = \mathbf{v} \otimes \mathbf{u}$ for $\mathbf{u} \in F^m$ and $\mathbf{v} \in F^n$: column j of $\mathbf{u}\mathbf{v}^\top$ is $v_j\mathbf{u}$.

Two conventions are in use, and they differ. Some books stack the **rows** instead, which exchanges the roles of the two factors in the next theorem. We always stack columns.

Theorem 8.4.4. The vec Identity

Let $\mathbf{A} \in M_{k\times m}(F)$, $\mathbf{X} \in M_{m\times n}(F)$, and $\mathbf{B} \in M_{n\times l}(F)$. Then

$$\operatorname{vec}(\mathbf{AXB}) = (\mathbf{B}^\top \otimes \mathbf{A})\operatorname{vec}\mathbf{X}.$$

Idea. Both sides are columns in F^{kl}. We compare one entry, using the formula for a triple product on the left and $(*)$ on the right. The factor $\mathbf{A}$ acts on the row index of $\mathbf{X}$, which is the inner index in vec; the factor $\mathbf{B}$ acts on the column index, which is the outer index. That is why $\mathbf{B}$ comes first, and it enters transposed because it multiplies $\mathbf{X}$ from the right.

Proof. Both sides lie in F^{kl}: $\mathbf{AXB} \in M_{k\times l}(F)$, and $\mathbf{B}^\top \otimes \mathbf{A} \in M_{lk\times nm}(F)$ acts on $\operatorname{vec}\mathbf{X} \in F^{mn}$. Fix $1 \le r \le k$ and $1 \le s \le l$, and consider position $(s-1)k + r$. By Definition 8.4.2 and Definition 0.9.9 applied twice, the left side has entry

$$(\mathbf{AXB})_{rs} = \sum_{i=1}^{m}\sum_{j=1}^{n} a_{ri}\, x_{ij}\, b_{js}.$$

On the right, write the column index $c \in \{1,\dots,nm\}$ of $\mathbf{B}^\top \otimes \mathbf{A}$ uniquely as $c = (j-1)m + i$ with $1 \le j \le n$, $1 \le i \le m$. By $(*)$, $(\mathbf{B}^\top \otimes \mathbf{A})_{(s-1)k+r,\ (j-1)m+i} = (\mathbf{B}^\top)_{sj}\, a_{ri} = b_{js}\, a_{ri}$, and by Definition 8.4.2 the c-th entry of $\operatorname{vec}\mathbf{X}$ is x_{ij}. Hence the entry of the right side in position $(s-1)k + r$ is

$$\sum_{c=1}^{nm} (\mathbf{B}^\top \otimes \mathbf{A})_{(s-1)k+r,\ c}\,(\operatorname{vec}\mathbf{X})_c = \sum_{j=1}^{n}\sum_{i=1}^{m} b_{js}\, a_{ri}\, x_{ij},$$

which equals the left side. As (r,s) ranges over all pairs, $(s-1)k+r$ ranges over all positions, so the two vectors are equal.

■

Taking $\mathbf{B} = \mathbf{I}_n$ or $\mathbf{A} = \mathbf{I}_m$ gives the two one-sided forms, which are used most often:

$$\operatorname{vec}(\mathbf{AX}) = (\mathbf{I}_n \otimes \mathbf{A})\operatorname{vec}\mathbf{X}, \qquad \operatorname{vec}(\mathbf{XB}) = (\mathbf{B}^\top \otimes \mathbf{I}_m)\operatorname{vec}\mathbf{X}.$$

The first says that multiplying $\mathbf{X}$ by $\mathbf{A}$ on the left multiplies each column of $\mathbf{X}$ by $\mathbf{A}$: $\mathbf{I}_n \otimes \mathbf{A} = \mathbf{A} \oplus \cdots \oplus \mathbf{A}$ acts on the stacked columns one at a time.

Now every linear matrix equation built from products becomes a linear system. The most

important one is the **Sylvester equation**

$$\mathbf{AX} + \mathbf{XB} = \mathbf{C}, \qquad \mathbf{A} \in M_m(F),\ \mathbf{B} \in M_n(F),\ \mathbf{C} \in M_{m\times n}(F),$$

for an unknown $\mathbf{X} \in M_{m\times n}(F)$. Applying vec, which is linear, and the two one-sided forms,

$$(\mathbf{I}_n \otimes \mathbf{A} + \mathbf{B}^\top \otimes \mathbf{I}_m)\ \operatorname{vec}\mathbf{X} = \operatorname{vec}\mathbf{C}.$$

So the Sylvester equation has **exactly one** solution for every $\mathbf{C}$ if and only if the $mn \times mn$ matrix $\mathbf{I}_n \otimes \mathbf{A} + \mathbf{B}^\top \otimes \mathbf{I}_m$ is invertible (Theorem 3.4.4, transported through the isomorphism vec). The equation $\mathbf{AX} = \mathbf{XA}$ of the opening, which describes the matrices commuting with $\mathbf{A}$, is the case $\mathbf{B} = -\mathbf{A}$, $\mathbf{C} = \mathbf{O}$: its solutions form the null space of $\mathbf{I}_m \otimes \mathbf{A} - \mathbf{A}^\top \otimes \mathbf{I}_m$.

Example. A Sylvester Equation as a Linear System

Solve $\mathbf{AX} + \mathbf{XB} = \mathbf{C}$ over $\mathbb{Q}$, where

$$\mathbf{A} = \begin{pmatrix} 1 & 1 \\ 0 & 2 \end{pmatrix}, \qquad \mathbf{B} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}, \qquad \mathbf{C} = \begin{pmatrix} 1 & -1 \\ 1 & 3 \end{pmatrix}.$$

Solution. Here $m = n = 2$. Compute the two Kronecker products from Definition 8.4.1:

$$\mathbf{I}_2 \otimes \mathbf{A} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 2 \end{pmatrix}, \qquad \mathbf{B}^\top \otimes \mathbf{I}_2 = \begin{pmatrix} 1\mathbf{I}_2 & 1\mathbf{I}_2 \\ 0\mathbf{I}_2 & 1\mathbf{I}_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

With $\operatorname{vec}\mathbf{X} = (x_{11}, x_{21}, x_{12}, x_{22})$ and $\operatorname{vec}\mathbf{C} = (1, 1, -1, 3)$, the system $(\mathbf{I}_2 \otimes \mathbf{A} + \mathbf{B}^\top \otimes \mathbf{I}_2)\operatorname{vec}\mathbf{X} = \operatorname{vec}\mathbf{C}$ is

$$\begin{pmatrix} 2 & 1 & 1 & 0 \\ 0 & 3 & 0 & 1 \\ 0 & 0 & 2 & 1 \\ 0 & 0 & 0 & 3 \end{pmatrix} \begin{pmatrix} x_{11} \\ x_{21} \\ x_{12} \\ x_{22} \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ -1 \\ 3 \end{pmatrix}.$$

The coefficient matrix is upper triangular with nonzero diagonal entries $2, 3, 2, 3$, so its determinant is $36 \neq 0$ (Theorem 7.4.6) and the solution is unique. Back substitution gives $x_{22} = 1$; then $2x_{12} + 1 = -1$, so $x_{12} = -1$; then $3x_{21} + 1 = 1$, so $x_{21} = 0$; then $2x_{11} + 0 - 1 = 1$, so $x_{11} = 1$. Hence

$$\mathbf{X} = \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix}.$$

Check: $\mathbf{AX} = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$ and $\mathbf{XB} = \begin{pmatrix} 0 & -1 \\ 1 & 1 \end{pmatrix}$, whose sum is $\mathbf{C}$. As a check on the bookkeeping, row 1 of the system is the $(1,1)$ entry of $\mathbf{AX} + \mathbf{XB}$, namely $(x_{11} + x_{21}) + (x_{11} + x_{12})$.

In this example the coefficient matrix came out triangular, with diagonal entries $a_{ii} + b_{jj}$. That is no accident when $\mathbf{A}$ is upper triangular and $\mathbf{B}$ lower triangular, as Exercise (C2: Kronecker sums) below shows. In general, whether $\mathbf{I}_n \otimes \mathbf{A} + \mathbf{B}^\top \otimes \mathbf{I}_m$ is invertible depends on the eigenvalues of $\mathbf{A}$ and $\mathbf{B}$: over $\mathbb{C}$, the equation $\mathbf{AX} - \mathbf{XB} = \mathbf{C}$ is uniquely solvable for every $\mathbf{C}$ exactly when $\mathbf{A}$ and $\mathbf{B}$ have no common eigenvalue. That criterion needs eigenvalues and is proved in Chapter 12, and its quantitative form appears in Chapter 20.

Warning. **Keep the order and the transpose in $\operatorname{vec}(\mathbf{AXB}) = (\mathbf{B}^\top \otimes \mathbf{A})\operatorname{vec}\mathbf{X}$.** Both $\mathbf{A}\otimes\mathbf{B}^\top$ and $\mathbf{B}\otimes\mathbf{A}$ look plausible, and both are wrong. With $\mathbf{X} = \mathbf{E}_{11} \in M_2(F)$, $\mathbf{A} = \mathbf{I}_2$, and $\mathbf{B} = \mathbf{E}_{21}$, we get $\mathbf{AXB} = \mathbf{E}_{11}\mathbf{E}_{21} = \mathbf{O}$, and indeed $(\mathbf{B}^\top \otimes \mathbf{I}_2)\,\mathbf{e}_1 = (\mathbf{E}_{12}\otimes\mathbf{I}_2)\mathbf{e}_1 = \mathbf{0}$, because the first column of $\mathbf{E}_{12}\otimes\mathbf{I}_2$ is zero. But $(\mathbf{B}\otimes\mathbf{I}_2)\,\mathbf{e}_1 = (\mathbf{E}_{21}\otimes\mathbf{I}_2)\mathbf{e}_1 = \mathbf{e}_3 \neq \mathbf{0}$. So $\mathbf{B}\otimes\mathbf{A}$ fails. A similar test with $\mathbf{A} = \mathbf{E}_{12}$, $\mathbf{X} = \mathbf{E}_{21}$, $\mathbf{B} = \mathbf{I}_2$ refutes $\mathbf{A}\otimes\mathbf{B}^\top$. If a formula is in doubt, test it on matrix units.

The Kronecker product is the coordinate shadow of a basis-free construction: the tensor product of linear maps, in Chapter 15, whose matrix in suitable bases is exactly $\mathbf{A}\otimes\mathbf{B}$. The dictionary order in $(*)$ will reappear there as an ordering of a basis $v_i \otimes w_k$.

8.4.7 Exercises

8.4.7.1 A. Check your understanding

Exercise. A1

(a) Define $\mathbf{A}\otimes\mathbf{B}$ for $\mathbf{A}\in M_{m\times n}(F)$ and $\mathbf{B}\in M_{p\times q}(F)$, and state its size.

(b) True or false: for all $\mathbf{A},\mathbf{B}\in M_2(F)$, $\mathbf{A}\otimes\mathbf{B} = \mathbf{B}\otimes\mathbf{A}$. Justify your answer.

(c) State the mixed product rule, including the conditions on the sizes.

(d) True or false: $\det(\mathbf{A}\otimes\mathbf{B}) = \det\mathbf{A}\det\mathbf{B}$ for all $\mathbf{A},\mathbf{B}\in M_2(\mathbb{R})$. Justify your answer.

(e) State the vec identity, and name the matrix $\mathbf{M}$ with $\operatorname{vec}(\mathbf{XB}) = \mathbf{M}\operatorname{vec}\mathbf{X}$ for $\mathbf{X}\in M_{m\times n}(F)$, $\mathbf{B}\in M_n(F)$.

Solution.

(a) $\mathbf{A}\otimes\mathbf{B}$ is the $mp\times nq$ block matrix with (i,j) block $a_{ij}\mathbf{B}$ (Definition 8.4.1); equivalently, its entry in position $((i-1)p+k,(j-1)q+l)$ is $a_{ij}b_{kl}$.

(b) False. For $\mathbf{A} = \begin{pmatrix}1&2\\3&4\end{pmatrix}$ and $\mathbf{B} = \begin{pmatrix}0&1\\1&0\end{pmatrix}$, the $(1,2)$ entry of $\mathbf{A}\otimes\mathbf{B}$ is $a_{11}b_{12} = 1$ and that of $\mathbf{B}\otimes\mathbf{A}$ is $b_{11}a_{12} = 0$. The two are similar (Proposition 8.4.2), not equal.

(c) If $\mathbf{A}\in M_{m\times n}(F)$, $\mathbf{C}\in M_{n\times t}(F)$, $\mathbf{B}\in M_{p\times q}(F)$, and $\mathbf{D}\in M_{q\times u}(F)$, then $(\mathbf{A}\otimes\mathbf{B})(\mathbf{C}\otimes\mathbf{D}) = (\mathbf{AC})\otimes(\mathbf{BD})$ (Theorem 8.4.1 (c)). The conditions are exactly that $\mathbf{AC}$ and $\mathbf{BD}$ are defined.

(d) False. By Theorem 8.4.3, $\det(\mathbf{A}\otimes\mathbf{B}) = (\det\mathbf{A})^2(\det\mathbf{B})^2$. For $\mathbf{A} = \mathbf{B} = 2\mathbf{I}_2$, $\mathbf{A}\otimes\mathbf{B} = 4\mathbf{I}_4$ has determinant 256, while $\det\mathbf{A}\det\mathbf{B} = 16$.

(e) $\operatorname{vec}(\mathbf{AXB}) = (\mathbf{B}^\top\otimes\mathbf{A})\operatorname{vec}\mathbf{X}$ (Theorem 8.4.4). With $\mathbf{A} = \mathbf{I}_m$, $\mathbf{M} = \mathbf{B}^\top\otimes\mathbf{I}_m$.

8.4.7.2 B. Practice

Exercise. B1: Computing with Kronecker products

Let $\mathbf{A} = \begin{pmatrix}2&0\\1&-1\end{pmatrix}$ and $\mathbf{B} = \begin{pmatrix}1&1\\0&3\end{pmatrix}$ over $\mathbb{Q}$.

(a) Write out $\mathbf{A}\otimes\mathbf{B}$ and $\mathbf{B}\otimes\mathbf{A}$.

(b) Compute tr, det, and rank of $\mathbf{A}\otimes\mathbf{B}$ from Theorem 8.4.3, and check the trace directly.

(c) Compute $\mathbf{A}^2\otimes\mathbf{B}^2$ and use it to write down $(\mathbf{A}\otimes\mathbf{B})^2$ without multiplying 4×4 matrices. Justify the step.

(d) Hence find $(\mathbf{A}\otimes\mathbf{B})^{-1}$.

Solution.

(a) By Definition 8.4.1,

$$\mathbf{A}\otimes\mathbf{B}=\begin{pmatrix}2\mathbf{B} & 0\mathbf{B}\\ 1\mathbf{B} & -1\mathbf{B}\end{pmatrix}=\begin{pmatrix}2&2&0&0\\0&6&0&0\\1&1&-1&-1\\0&3&0&-3\end{pmatrix},$$

$$\mathbf{B}\otimes\mathbf{A}=\begin{pmatrix}1\mathbf{A} & 1\mathbf{A}\\ 0\mathbf{A} & 3\mathbf{A}\end{pmatrix}=\begin{pmatrix}2&0&2&0\\1&-1&1&-1\\0&0&6&0\\0&0&3&-3\end{pmatrix}.$$

(b) $\operatorname{tr}\mathbf{A}=1$, $\operatorname{tr}\mathbf{B}=4$, $\det\mathbf{A}=-2$, $\det\mathbf{B}=3$, and both have rank 2. By Theorem 8.4.3, $\operatorname{tr}(\mathbf{A}\otimes\mathbf{B})=4$, $\det(\mathbf{A}\otimes\mathbf{B})=(-2)^2\cdot3^2=36$, and $\operatorname{rank}(\mathbf{A}\otimes\mathbf{B})=4$. Directly, the diagonal of $\mathbf{A}\otimes\mathbf{B}$ is $2,6,-1,-3$, with sum 4.

(c) By the mixed product rule (Theorem 8.4.1 (c)) with $\mathbf{C}=\mathbf{A}$ and $\mathbf{D}=\mathbf{B}$, $(\mathbf{A}\otimes\mathbf{B})^2=\mathbf{A}^2\otimes\mathbf{B}^2$. Here $\mathbf{A}^2=\begin{pmatrix}4&0\\1&1\end{pmatrix}$ and $\mathbf{B}^2=\begin{pmatrix}1&4\\0&9\end{pmatrix}$, so

$$(\mathbf{A}\otimes\mathbf{B})^2=\begin{pmatrix}4\mathbf{B}^2 & \mathbf{O}\\ \mathbf{B}^2 & \mathbf{B}^2\end{pmatrix}=\begin{pmatrix}4&16&0&0\\0&36&0&0\\1&4&1&4\\0&9&0&9\end{pmatrix}.$$

(d) $\mathbf{A}^{-1}=\frac{1}{-2}\begin{pmatrix}-1&0\\-1&2\end{pmatrix}=\begin{pmatrix}\frac12&0\\\frac12&-1\end{pmatrix}$ and $\mathbf{B}^{-1}=\frac13\begin{pmatrix}3&-1\\0&1\end{pmatrix}$ by Theorem 0.9.8. By Theorem 8.4.1 (e),

$$(\mathbf{A}\otimes\mathbf{B})^{-1}=\mathbf{A}^{-1}\otimes\mathbf{B}^{-1}=\begin{pmatrix}\frac12\mathbf{B}^{-1} & \mathbf{O}\\ \frac12\mathbf{B}^{-1} & -\mathbf{B}^{-1}\end{pmatrix}=\begin{pmatrix}\frac12&-\frac16&0&0\\0&\frac16&0&0\\\frac12&-\frac16&-1&\frac13\\0&\frac16&0&-\frac13\end{pmatrix}.$$

Multiplying by $\mathbf{A}\otimes\mathbf{B}$ gives $\mathbf{I}_4$.

Exercise. B2: One-sided equations

Let $\mathbf{A}\in M_m(F)$ and $\mathbf{C}\in M_{m\times n}(F)$, and consider $\mathbf{AX}=\mathbf{C}$ for $\mathbf{X}\in M_{m\times n}(F)$.

(a) Write the equation in the form $\mathbf{M}\operatorname{vec}\mathbf{X}=\operatorname{vec}\mathbf{C}$, and describe $\mathbf{M}$ as a block matrix.

(b) Explain why the system splits into n separate systems, one for each column of $\mathbf{X}$.

(c) Hence show that $\mathbf{AX} = \mathbf{C}$ has exactly one solution for every $\mathbf{C}$ if and only if $\mathbf{A}$ is invertible.

(d) Solve $\mathbf{XA} = \mathbf{C}$ over $\mathbb{Q}$ by the same method, where $\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 1 & 3 \end{pmatrix}$ and $\mathbf{C} = \begin{pmatrix} 3 & 7 \\ 0 & 1 \end{pmatrix}$.

Solution.

(a) By Theorem 8.4.4 with $\mathbf{B} = \mathbf{I}_n$, $\operatorname{vec}(\mathbf{AX}) = (\mathbf{I}_n \otimes \mathbf{A}) \operatorname{vec} \mathbf{X}$, so $\mathbf{M} = \mathbf{I}_n \otimes \mathbf{A} = \mathbf{A} \oplus \cdots \oplus \mathbf{A}$, with n blocks.

(b) Write $\operatorname{vec} \mathbf{X}$ and $\operatorname{vec} \mathbf{C}$ as stacks of the columns $\mathbf{x}_j$ and $\mathbf{c}_j$. By Theorem 8.1.1, $(\mathbf{A} \oplus \cdots \oplus \mathbf{A}) \operatorname{vec} \mathbf{X}$ is the stack of the $\mathbf{A}\mathbf{x}_j$. So the system says $\mathbf{A}\mathbf{x}_j = \mathbf{c}_j$ for $j = 1, \ldots, n$, and the unknowns of different columns never meet.

(c) ($\Leftarrow$) If $\mathbf{A}$ is invertible, then $\mathbf{AX} = \mathbf{C}$ forces $\mathbf{X} = \mathbf{A}^{-1}\mathbf{C}$, and this $\mathbf{X}$ is a solution. ($\Rightarrow$) Suppose $\mathbf{AX} = \mathbf{C}$ is solvable for every $\mathbf{C}$. Given $\mathbf{c} \in F^m$, take $\mathbf{C}$ with every column equal to $\mathbf{c}$; by (b), a solution gives $\mathbf{x}$ with $\mathbf{A}\mathbf{x} = \mathbf{c}$. So $\mathbf{A}\mathbf{x} = \mathbf{c}$ is solvable for every $\mathbf{c}$, and $\mathbf{A}$ is invertible by Theorem 3.4.4 ((e) $\Rightarrow$ (a)).

(d) By Theorem 8.4.4 with the left factor $\mathbf{I}_2$, $\operatorname{vec}(\mathbf{XA}) = (\mathbf{A}^\top \otimes \mathbf{I}_2) \operatorname{vec} \mathbf{X}$. With $\mathbf{A}^\top = \begin{pmatrix} 1 & 1 \\ 2 & 3 \end{pmatrix}$, $\operatorname{vec} \mathbf{X} = (x_{11}, x_{21}, x_{12}, x_{22})$, and $\operatorname{vec} \mathbf{C} = (3, 0, 7, 1)$, the system is

$$\begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 2 & 0 & 3 & 0 \\ 0 & 2 & 0 & 3 \end{pmatrix} \begin{pmatrix} x_{11} \\ x_{21} \\ x_{12} \\ x_{22} \end{pmatrix} = \begin{pmatrix} 3 \\ 0 \\ 7 \\ 1 \end{pmatrix}.$$

This time the system splits by the **rows** of $\mathbf{X}$: equations $1, 3$ involve only x_{11}, x_{12}, and equations $2, 4$ only x_{21}, x_{22}. From $x_{11} + x_{12} = 3$ and $2x_{11} + 3x_{12} = 7$ we get $x_{12} = 1$, $x_{11} = 2$. From $x_{21} + x_{22} = 0$ and $2x_{21} + 3x_{22} = 1$ we get $x_{22} = 1$, $x_{21} = -1$. Hence

$$\mathbf{X} = \begin{pmatrix} 2 & 1 \\ -1 & 1 \end{pmatrix}.$$

Check: $\mathbf{XA} = \begin{pmatrix} 2+1 & 4+3 \\ -1+1 & -2+3 \end{pmatrix} = \begin{pmatrix} 3 & 7 \\ 0 & 1 \end{pmatrix} = \mathbf{C}$.

Exercise. B3: Which identities hold?

Determine which of the following statements hold for **all** $\mathbf{A}, \mathbf{B}, \mathbf{C} \in M_2(\mathbb{R})$. Justify your answer.

(a) $(\mathbf{A} \otimes \mathbf{B})(\mathbf{A} \otimes \mathbf{B}) = \mathbf{A}^2 \otimes \mathbf{B}^2$.

(b) $(\mathbf{A} + \mathbf{B}) \otimes \mathbf{C} = \mathbf{A} \otimes \mathbf{C} + \mathbf{B} \otimes \mathbf{C}$.

(c) $(\mathbf{A} \otimes \mathbf{B})^\top = \mathbf{B}^\top \otimes \mathbf{A}^\top$.

(d) $\mathbf{A} \otimes \mathbf{B} = \mathbf{O}$ implies $\mathbf{A} = \mathbf{O}$ or $\mathbf{B} = \mathbf{O}$.

(e) $\operatorname{tr}(\mathbf{A} \otimes \mathbf{B}) = \operatorname{tr}(\mathbf{B} \otimes \mathbf{A})$.

(f) $\mathbf{A} \otimes \mathbf{B} + \mathbf{B} \otimes \mathbf{A} = 2(\mathbf{A} \otimes \mathbf{B})$.

Solution.

(a) Holds, by the mixed product rule (Theorem 8.4.1 (c)).

(b) Holds, by bilinearity (Theorem 8.4.1 (a)).

(c) Fails. By Theorem 8.4.1 (d), $(\mathbf{A}\otimes\mathbf{B})^\top = \mathbf{A}^\top\otimes\mathbf{B}^\top$, and this differs from $\mathbf{B}^\top\otimes\mathbf{A}^\top$ in general. Take $\mathbf{A} = \begin{pmatrix}1 & 3\\ 2 & 4\end{pmatrix}$ and $\mathbf{B} = \begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix} = \mathbf{B}^\top$. By $(*)$, the $(1,2)$ entry of $\mathbf{A}^\top\otimes\mathbf{B}^\top$ is $(\mathbf{A}^\top)_{11}b_{12} = 1$, while that of $\mathbf{B}^\top\otimes\mathbf{A}^\top$ is $b_{11}(\mathbf{A}^\top)_{12} = 0$. Unlike $(\mathbf{AB})^\top = \mathbf{B}^\top\mathbf{A}^\top$, the order of the factors does not reverse.

(d) Holds. If $\mathbf{A}\neq\mathbf{O}$ and $\mathbf{B}\neq\mathbf{O}$, then $\operatorname{rank}\mathbf{A}\geq 1$ and $\operatorname{rank}\mathbf{B}\geq 1$, so $\operatorname{rank}(\mathbf{A}\otimes\mathbf{B})\geq 1$ by Theorem 8.4.3 (a), and $\mathbf{A}\otimes\mathbf{B}\neq\mathbf{O}$. (Directly: if $a_{ij}\neq 0$ and $b_{kl}\neq 0$, the entry $a_{ij}b_{kl}$ of $\mathbf{A}\otimes\mathbf{B}$ is nonzero.)

(e) Holds: both equal $\operatorname{tr}\mathbf{A}\operatorname{tr}\mathbf{B}$ by Theorem 8.4.3 (b).

(f) Fails, since it is equivalent to $\mathbf{B}\otimes\mathbf{A} = \mathbf{A}\otimes\mathbf{B}$, which is false for the matrices of Example 1 after Definition 8.4.1 (see the nonexample there).

8.4.7.3 C. Going deeper

Exercise. C1: When is a Kronecker product invertible?

(a) Let $\mathbf{A}\in M_n(F)$ and $\mathbf{B}\in M_m(F)$. Prove that $\mathbf{A}\otimes\mathbf{B}$ is invertible if and only if both $\mathbf{A}$ and $\mathbf{B}$ are invertible.

(b) Let $\mathbf{A}\in M_{m\times n}(F)$ and $\mathbf{B}\in M_{p\times q}(F)$ with $mp = nq$, so that $\mathbf{A}\otimes\mathbf{B}$ is square. Prove that if $m\neq n$, then $\mathbf{A}\otimes\mathbf{B}$ is **not** invertible.

(c) Give an example of a 2×1 matrix $\mathbf{A}$ and a 1×2 matrix $\mathbf{B}$ for which $\mathbf{A}\otimes\mathbf{B}$ is a nonzero 2×2 matrix, and compute its rank.

Solution.

(a) By Theorem 8.4.3 (c), $\det(\mathbf{A}\otimes\mathbf{B}) = (\det\mathbf{A})^m(\det\mathbf{B})^n$. In a field a product is nonzero if and only if each factor is, so this is nonzero if and only if $\det\mathbf{A}\neq 0$ and $\det\mathbf{B}\neq 0$ (here $n, m\geq 1$). By Theorem 7.5.2, applied to all three matrices, $\mathbf{A}\otimes\mathbf{B}$ is invertible if and only if $\mathbf{A}$ and $\mathbf{B}$ are.

(b) Suppose $m < n$. Since $mp = nq$ and $m < n$, we get $p > q$. By Theorem 8.4.3 (a) and the bound $\operatorname{rank}\mathbf{M}\leq$ (number of rows) and $\leq$ (number of columns),

$$\operatorname{rank}(\mathbf{A}\otimes\mathbf{B}) = \operatorname{rank}\mathbf{A}\operatorname{rank}\mathbf{B}\leq m\cdot q < n\cdot q = nq,$$

so the $nq\times nq$ matrix $\mathbf{A}\otimes\mathbf{B}$ is not invertible (Theorem 7.5.3 (d)). If $m > n$, then $p < q$, and $\operatorname{rank}\mathbf{A}\operatorname{rank}\mathbf{B}\leq n\cdot p < m\cdot p = mp = nq$ in the same way.

(c) Take $\mathbf{A} = \begin{pmatrix}1\\ 2\end{pmatrix}$ and $\mathbf{B} = \begin{pmatrix}3 & 1\end{pmatrix}$. Then $\mathbf{A}\otimes\mathbf{B} = \begin{pmatrix}1\mathbf{B}\\ 2\mathbf{B}\end{pmatrix} = \begin{pmatrix}3 & 1\\ 6 & 2\end{pmatrix}$, the outer product $\mathbf{AB}$, of rank $1\cdot 1 = 1$: its second row is twice its first. So it is square and nonzero but singular, as (b) predicts.

Exercise. C2: Kronecker sums

Let $\mathbf{A} \in M_n(F)$ and $\mathbf{B} \in M_m(F)$, and put $\mathbf{K} \coloneqq \mathbf{A} \otimes \mathbf{I}_m + \mathbf{I}_n \otimes \mathbf{B} \in M_{nm}(F)$.

(a) Prove that $(\mathbf{A} \otimes \mathbf{I}_m)(\mathbf{I}_n \otimes \mathbf{B}) = (\mathbf{I}_n \otimes \mathbf{B})(\mathbf{A} \otimes \mathbf{I}_m)$, and deduce that $\mathbf{K}^2 = \mathbf{A}^2 \otimes \mathbf{I}_m + 2(\mathbf{A} \otimes \mathbf{B}) + \mathbf{I}_n \otimes \mathbf{B}^2$.

(b) Show that for $\mathbf{X} \in M_{m\times n}(F)$, $\operatorname{vec}(\mathbf{B}\mathbf{X} + \mathbf{X}\mathbf{A}^{\top}) = \mathbf{K} \operatorname{vec} \mathbf{X}$.

(c) Suppose $\mathbf{A}$ and $\mathbf{B}$ are upper triangular. Prove that $\mathbf{K}$ is upper triangular with diagonal entries $a_{ii} + b_{kk}$, and deduce that $\mathbf{B}\mathbf{X} + \mathbf{X}\mathbf{A}^{\top} = \mathbf{C}$ has exactly one solution for every $\mathbf{C} \in M_{m\times n}(F)$ if and only if $a_{ii} + b_{kk} \neq 0$ for all i, k.

(d) For $\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 0 & -1 \end{pmatrix}$ and $\mathbf{B} = \begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix}$ over $\mathbb{Q}$, decide whether $\mathbf{B}\mathbf{X} + \mathbf{X}\mathbf{A}^{\top} = \mathbf{C}$ is solvable for every $\mathbf{C}$. If not, find a $\mathbf{C}$ with no solution.

Solution.

(a) By the mixed product rule, both products equal $(\mathbf{A}\mathbf{I}_n) \otimes (\mathbf{I}_m\mathbf{B}) = (\mathbf{I}_n\mathbf{A}) \otimes (\mathbf{B}\mathbf{I}_m) = \mathbf{A} \otimes \mathbf{B}$, as in $(**)$. Expanding $\mathbf{K}^2$ by distributivity (Theorem 0.9.4) gives $(\mathbf{A} \otimes \mathbf{I}_m)^2 + (\mathbf{A} \otimes \mathbf{I}_m)(\mathbf{I}_n \otimes \mathbf{B}) + (\mathbf{I}_n \otimes \mathbf{B})(\mathbf{A} \otimes \mathbf{I}_m) + (\mathbf{I}_n \otimes \mathbf{B})^2$. The middle terms are both $\mathbf{A} \otimes \mathbf{B}$, and the outer ones are $\mathbf{A}^2 \otimes \mathbf{I}_m$ and $\mathbf{I}_n \otimes \mathbf{B}^2$ by the mixed product rule.

(b) By Theorem 8.4.4, $\operatorname{vec}(\mathbf{B}\mathbf{X}\mathbf{I}_n) = (\mathbf{I}_n \otimes \mathbf{B}) \operatorname{vec} \mathbf{X}$ and $\operatorname{vec}(\mathbf{I}_m\mathbf{X}\mathbf{A}^{\top}) = ((\mathbf{A}^{\top})^{\top} \otimes \mathbf{I}_m) \operatorname{vec} \mathbf{X} = (\mathbf{A} \otimes \mathbf{I}_m) \operatorname{vec} \mathbf{X}$. Add, using linearity of vec.

(c) By $(*)$, the entry of $\mathbf{A} \otimes \mathbf{I}_m$ in position $((i-1)m + k, (j-1)m + l)$ is $a_{ij}\delta_{kl}$, which is zero unless $i \leq j$ and $k = l$; in that case $(i-1)m + k \leq (j-1)m + l$. The entry of $\mathbf{I}_n \otimes \mathbf{B}$ there is $\delta_{ij}b_{kl}$, zero unless $i = j$ and $k \leq l$, and again the row index is at most the column index. So both are upper triangular, hence so is $\mathbf{K}$. On the diagonal, $i = j$ and $k = l$, and the entry is $a_{ii} + b_{kk}$. By Theorem 7.4.6, $\det \mathbf{K} = \prod_{i,k}(a_{ii} + b_{kk})$. By (b), the equation is $\mathbf{K} \operatorname{vec} \mathbf{X} = \operatorname{vec} \mathbf{C}$, and since vec is an isomorphism, it has exactly one solution for every $\mathbf{C}$ if and only if $\mathbf{K}$ is invertible (Theorem 3.4.4), that is $\det \mathbf{K} \neq 0$ (Theorem 7.5.2), that is every $a_{ii} + b_{kk} \neq 0$.

(d) The diagonal entries $a_{ii} + b_{kk}$ are $1 + 1 = 2$, $1 + 3 = 4$, $-1 + 1 = 0$, $-1 + 3 = 2$, so by (c) the equation is not solvable for every $\mathbf{C}$. Explicitly,

$$\mathbf{K} = \begin{pmatrix} 1\mathbf{I}_2 & 2\mathbf{I}_2 \\ 0\mathbf{I}_2 & -1\mathbf{I}_2 \end{pmatrix} + \begin{pmatrix} \mathbf{B} & \mathbf{O} \\ \mathbf{O} & \mathbf{B} \end{pmatrix} = \begin{pmatrix} 2 & 0 & 2 & 0 \\ 0 & 4 & 0 & 2 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 \end{pmatrix},$$

whose third row is zero. So $\mathbf{K} \operatorname{vec} \mathbf{X}$ always has third entry 0, and any $\mathbf{C}$ with $(\operatorname{vec} \mathbf{C})_3 = c_{12} \neq 0$ has no solution, for example $\mathbf{C} = \mathbf{E}_{12}$. (Directly: the $(1,2)$ entry of $\mathbf{B}\mathbf{X} + \mathbf{X}\mathbf{A}^{\top}$ is $x_{12} + (x_{11}a_{21} + x_{12}a_{22}) = x_{12} - x_{12} = 0$.)

8.5 Commutators and Trace-Zero Matrices

The expression $\mathbf{A}\mathbf{B} - \mathbf{B}\mathbf{A}$ has come up several times already. In Chapter 0 its trace was 0, which showed that $\mathbf{A}\mathbf{B} - \mathbf{B}\mathbf{A} = \mathbf{I}_n$ is impossible over $\mathbb{R}$ (Exercise (C1: $\mathbf{AB} - \mathbf{BA} = \mathbf{I}$ is impossible, or is it?)). In Chapter 2 the same trace argument showed that the relation $DS - SD = \mathrm{id}$

between differentiation and multiplication by x on $F[x]$ has no finite-dimensional analogue in characteristic 0 (Exercise (C1: The Heisenberg relation)). Both arguments use one direction only: every matrix of the form $\mathbf{AB}-\mathbf{BA}$ has trace zero. This section asks whether the converse holds. Is every trace-zero matrix of the form $\mathbf{AB}-\mathbf{BA}$? The answer is yes over a field of characteristic 0, and the proof is a good example of the block techniques of this chapter: move the matrix into a convenient shape by a similarity, one row and column at a time, and then solve an equation that has become entrywise.

Which fields. Commutators, their basic rules, the first similarity lemma, and the final theorem on spans work over **every** field F. The lemma on zero diagonals and Shoda's theorem assume that F has **characteristic** 0. We point out exactly where that hypothesis is used, and what happens in characteristic p.

8.5.1 Commutators

Two matrices $\mathbf{A}$ and $\mathbf{B}$ commute when $\mathbf{AB}=\mathbf{BA}$, and most pairs do not. The difference $\mathbf{AB}-\mathbf{BA}$ measures the failure, and it keeps recurring: in the trace arguments above, in the relation between differentiation and multiplication by x, and in the description of trace-like functionals (Exercise (C2: The trace is the only trace-like functional)). So we give it a name.

The commutator of two matrices is the amount by which they fail to commute.

Definition 8.5.1. Commutator

Let $n \geq 1$ and $\mathbf{A}, \mathbf{B} \in M_n(F)$. The **commutator** of $\mathbf{A}$ and $\mathbf{B}$ is

$$[\mathbf{A}, \mathbf{B}] := \mathbf{AB} - \mathbf{BA} \in M_n(F).$$

A matrix $\mathbf{C} \in M_n(F)$ is **a commutator** if $\mathbf{C} = [\mathbf{X}, \mathbf{Y}]$ for **some** $\mathbf{X}, \mathbf{Y} \in M_n(F)$. For operators $S, T \in \mathcal{L}(V)$ on a vector space V, we write likewise $[S, T] := ST - TS \in \mathcal{L}(V)$.

In words: $[\mathbf{A}, \mathbf{B}]$ is "$\mathbf{A}$ then $\mathbf{B}$" minus "$\mathbf{B}$ then $\mathbf{A}$", in that order, and it is the zero matrix exactly when $\mathbf{A}$ and $\mathbf{B}$ commute. The second sentence of the definition is about a single matrix: being a commutator means that **there exist** two matrices whose commutator it is, and it says nothing about which ones.

Examples.

1. **Degenerate cases.** $[\mathbf{A}, \mathbf{A}] = \mathbf{O}$ and $[\mathbf{A}, c\mathbf{I}_n] = c\mathbf{A} - c\mathbf{A} = \mathbf{O}$ for every $c \in F$. So the zero matrix is a commutator, in many ways. In $M_1(F) = F$ every commutator is $ab - ba = 0$, so for $n = 1$ the only commutator is 0; this is the smallest case, and it already matches the trace condition, since the only 1×1 matrix of trace 0 is $\mathbf{O}_1$.
2. **Matrix units.** Recall that $\mathbf{E}_{ij}\mathbf{E}_{kl}$ is $\mathbf{E}_{il}$ if $j = k$ and $\mathbf{O}$ otherwise (Definition 0.9.9). So for $i \neq k$,

$$[\mathbf{E}_{ij}, \mathbf{E}_{jk}] = \mathbf{E}_{ik} - \mathbf{O} = \mathbf{E}_{ik}, \qquad \text{and} \qquad [\mathbf{E}_{ij}, \mathbf{E}_{ji}] = \mathbf{E}_{ii} - \mathbf{E}_{jj} \quad (i \neq j).$$

   For instance $[\mathbf{E}_{12}, \mathbf{E}_{21}] = \operatorname{diag}(1, -1)$ in $M_2(F)$, and $[\mathbf{E}_{11}, \mathbf{E}_{12}] = \mathbf{E}_{12}$. Both have trace 0.
3. **Operators on an infinite-dimensional space.** On $F[x]$, with D differentiation and S multiplication by x, Exercise (C1: The Heisenberg relation) (a) shows $[D, S] = \mathrm{id}$. So the identity **is** a commutator of operators on $F[x]$. This does not contradict anything, because there is no trace on $\mathcal{L}(F[x])$: the trace was defined only on finite-dimensional spaces (Definition 4.5.2).

Nonexample by minimal change. Change the minus sign to a plus: $\mathbf{AB}+\mathbf{BA}$. It is still linear in $\mathbf{A}$ and in $\mathbf{B}$, and it is still built from the two products $\mathbf{AB}$ and $\mathbf{BA}$. What is lost is the trace: $\operatorname{tr}(\mathbf{AB}+\mathbf{BA}) = 2\operatorname{tr}(\mathbf{AB})$, which has no reason to vanish. For example $\mathbf{E}_{12}\mathbf{E}_{21}+\mathbf{E}_{21}\mathbf{E}_{12} = \mathbf{E}_{11}+\mathbf{E}_{22} = \mathbf{I}_2$, with trace 2. It also no longer detects commuting: $\mathbf{AA}+\mathbf{AA} = 2\mathbf{A}^2$ is nonzero for $\mathbf{A} = \mathbf{I}_n$ over $\mathbb{Q}$, although $\mathbf{A}$ commutes with itself.

Why this definition. The order in $\mathbf{AB}-\mathbf{BA}$ is a convention; the other order gives $[\mathbf{B},\mathbf{A}] = -[\mathbf{A},\mathbf{B}]$, which is a commutator too. The notion is invariant under a change of basis: a similarity applied to $\mathbf{X}$ and $\mathbf{Y}$ applies to $[\mathbf{X},\mathbf{Y}]$. So "being a commutator" is a property of the operator $\mathbf{x} \mapsto \mathbf{Cx}$, not of the basis used to write $\mathbf{C}$, and we are free to choose a convenient basis. This is what the proof of the main theorem exploits.

The basic rules are collected here.

Proposition 8.5.1. Rules for Commutators

Let $\mathbf{A},\mathbf{A}',\mathbf{B} \in M_n(F)$, $c \in F$, and let $\mathbf{P} \in M_n(F)$ be invertible. Then:

(a) $[\mathbf{A}+\mathbf{A}',\mathbf{B}] = [\mathbf{A},\mathbf{B}]+[\mathbf{A}',\mathbf{B}]$, $[c\mathbf{A},\mathbf{B}] = c[\mathbf{A},\mathbf{B}]$, and $[\mathbf{B},\mathbf{A}] = -[\mathbf{A},\mathbf{B}]$; in particular the commutator is linear in each slot.

(b) $\mathbf{P}^{-1}[\mathbf{A},\mathbf{B}]\mathbf{P} = [\mathbf{P}^{-1}\mathbf{AP},\mathbf{P}^{-1}\mathbf{BP}]$.

(c) $\operatorname{tr}[\mathbf{A},\mathbf{B}] = 0$.

Proof.

(a) By Theorem 0.9.4, $(\mathbf{A}+\mathbf{A}')\mathbf{B}-\mathbf{B}(\mathbf{A}+\mathbf{A}') = (\mathbf{AB}-\mathbf{BA})+(\mathbf{A}'\mathbf{B}-\mathbf{BA}')$ and $(c\mathbf{A})\mathbf{B}-\mathbf{B}(c\mathbf{A}) = c(\mathbf{AB}-\mathbf{BA})$. The last identity is $\mathbf{BA}-\mathbf{AB} = -(\mathbf{AB}-\mathbf{BA})$. Linearity in the second slot follows from the first slot and antisymmetry.

(b) Inserting $\mathbf{PP}^{-1} = \mathbf{I}_n$, $\mathbf{P}^{-1}\mathbf{ABP} = (\mathbf{P}^{-1}\mathbf{AP})(\mathbf{P}^{-1}\mathbf{BP})$, and likewise for $\mathbf{BA}$. Subtract, using distributivity.

(c) By Theorem 0.9.6 (1) and (3), $\operatorname{tr}(\mathbf{AB}-\mathbf{BA}) = \operatorname{tr}(\mathbf{AB})-\operatorname{tr}(\mathbf{BA}) = 0$.

■

So every commutator lies in the subspace of trace-zero matrices of Example (Trace-Free Matrices). The question of this section is whether it fills that subspace.

8.5.2 The identity in characteristic p

Rule (c) is the whole content of the Heisenberg exercises: if $[\mathbf{X},\mathbf{Y}] = \mathbf{I}_n$, then $n \cdot 1 = \operatorname{tr}\mathbf{I}_n = 0$, which is impossible when F has characteristic 0 (Definition 0.7.2). The obstruction is the number $n \cdot 1$, not the identity matrix itself. In characteristic p, the $p \times p$ identity has trace $p \cdot 1 = 0$, and the obstruction disappears. In fact the pair D, S of Example 3 survives in a finite form.

Example. The Identity as a Commutator in Characteristic p

Let F have prime characteristic p (for instance $F = \mathbb{F}_p$). Let $\mathbf{D}_p \in M_p(F)$ have $(k,k+1)$-entry $k \cdot 1$ for $1 \le k \le p-1$, and let $\mathbf{S}_p \in M_p(F)$ have $(k+1,k)$-entry 1 for $1 \le k \le p-1$, all other entries 0. Show that $[\mathbf{D}_p,\mathbf{S}_p] = \mathbf{I}_p$.

Solution. The columns of these matrices describe differentiation and multiplication by x on the list $1, x, \dots, x^{p-1}$, where x^p is replaced by 0. By Theorem 0.9.2, $\mathbf{S}_p\mathbf{e}_k = \mathbf{e}_{k+1}$ for $k < p$ and $\mathbf{S}_p\mathbf{e}_p = \mathbf{0}$, while $\mathbf{D}_p\mathbf{e}_1 = \mathbf{0}$ and $\mathbf{D}_p\mathbf{e}_k = (k-1)\mathbf{e}_{k-1}$ for $k \geq 2$. Hence

$$\mathbf{D}_p\mathbf{S}_p\mathbf{e}_k = k\,\mathbf{e}_k \ (k < p), \quad \mathbf{D}_p\mathbf{S}_p\mathbf{e}_p = \mathbf{0}, \qquad \mathbf{S}_p\mathbf{D}_p\mathbf{e}_k = (k-1)\,\mathbf{e}_k \ (1 \leq k \leq p),$$

where for $k = 1$ we used $\mathbf{S}_p\mathbf{D}_p\mathbf{e}_1 = \mathbf{0} = 0\,\mathbf{e}_1$. Subtracting, $[\mathbf{D}_p, \mathbf{S}_p]\mathbf{e}_k = \mathbf{e}_k$ for $k < p$, and $[\mathbf{D}_p, \mathbf{S}_p]\mathbf{e}_p = -(p-1)\mathbf{e}_p = (1-p)\mathbf{e}_p = \mathbf{e}_p$, because $p \cdot 1 = 0$ in F (Definition 0.7.2, Lemma 0.7.4). So every column of $[\mathbf{D}_p, \mathbf{S}_p]$ is the corresponding column of $\mathbf{I}_p$. For $p = 2$ this is the pair of Exercise (C1: The Heisenberg relation) (d), transposed.

Warning. **"The identity is never a commutator" is a statement about characteristic 0.** The trace test says that a commutator has trace 0, and $\operatorname{tr}\mathbf{I}_n = n \cdot 1$. Over $\mathbb{R}$ this is $n \neq 0$; over $\mathbb{F}_p$ it is 0 whenever p divides n, and then the test is silent. The example above shows that $\mathbf{I}_p$ really is a commutator over $\mathbb{F}_p$. Whenever an argument ends with "$n \cdot 1 = 0$, contradiction", check which fields it covers.

8.5.3 Trace-zero matrices with zero diagonal

We want to write a given trace-zero $\mathbf{A}$ as $[\mathbf{X}, \mathbf{Y}]$. Look first at the easiest choice of $\mathbf{X}$, a diagonal matrix $\mathbf{X} = \operatorname{diag}(x_1, \dots, x_n)$. For any $\mathbf{Y} = (y_{ij})$, Definition 0.9.9 gives $(\mathbf{XY})_{ij} = x_i y_{ij}$ and $(\mathbf{YX})_{ij} = y_{ij}x_j$, so

$$[\mathbf{X}, \mathbf{Y}]_{ij} = (x_i - x_j)\,y_{ij}. \tag{$*$}$$

The diagonal entries of $[\mathbf{X}, \mathbf{Y}]$ are all 0, and each off-diagonal entry can be anything, provided $x_i \neq x_j$. So a diagonal $\mathbf{X}$ with distinct entries produces **exactly** the matrices with zero diagonal. A trace-zero matrix need not have zero diagonal, but by Proposition 8.5.1 (b) we may replace $\mathbf{A}$ by any matrix similar to it. The plan is therefore:

① find a similarity that makes the diagonal of $\mathbf{A}$ zero; ② solve $(*)$ entrywise.

Step ① is the heart of the matter. We do it one diagonal entry at a time, starting with the $(1,1)$-entry.

A matrix is **scalar** if it equals $\lambda\mathbf{I}_n$ for some $\lambda \in F$. A scalar matrix is similar only to itself, since $\mathbf{P}^{-1}(\lambda\mathbf{I}_n)\mathbf{P} = \lambda\mathbf{I}_n$, so its diagonal cannot be changed. Every other matrix can at least be given a zero in the corner. This first lemma holds over every field.

Lemma 8.5.2. A Non-Scalar Matrix Has a Similar Matrix with Zero Corner

Let F be any field, let $n \geq 2$, and let $\mathbf{A} \in M_n(F)$ be **not** scalar. Then there is an invertible $\mathbf{P} \in M_n(F)$ such that the first column of $\mathbf{P}^{-1}\mathbf{AP}$ is $\mathbf{e}_2$. In particular

$$\mathbf{P}^{-1}\mathbf{AP} = \begin{pmatrix} 0 & \mathbf{r}^{\top} \\ \mathbf{e}_1 & \mathbf{A}_1 \end{pmatrix}$$

for some $\mathbf{r} \in F^{n-1}$ and $\mathbf{A}_1 \in M_{n-1}(F)$ with $\operatorname{tr}\mathbf{A}_1 = \operatorname{tr}\mathbf{A}$, where $\mathbf{e}_1 \in F^{n-1}$.

Idea. Work backwards from the goal. The first column of $\mathbf{P}^{-1}\mathbf{A}\mathbf{P}$ is the coordinate vector of $\mathbf{A}\mathbf{p}_1$ in the basis formed by the columns $\mathbf{p}_1, \dots, \mathbf{p}_n$ of $\mathbf{P}$. For it to be $\mathbf{e}_2$, we need $\mathbf{A}\mathbf{p}_1 = \mathbf{p}_2$. So we want a vector $\mathbf{v}$ such that $\mathbf{v}$ and $\mathbf{A}\mathbf{v}$ are independent, and then we build a basis starting $\mathbf{v}, \mathbf{A}\mathbf{v}$. Such a $\mathbf{v}$ exists unless $\mathbf{A}$ sends every vector to a multiple of itself, and a short argument shows that this forces $\mathbf{A}$ to be scalar.

Proof. We first show that there is $\mathbf{v} \in F^n$ with $(\mathbf{v}, \mathbf{A}\mathbf{v})$ linearly independent. Suppose not. Then for every $\mathbf{v} \neq \mathbf{0}$ there are $a, b \in F$, not both zero, with $a\mathbf{v} + b\mathbf{A}\mathbf{v} = \mathbf{0}$. If $b = 0$, then $a \neq 0$ and $a\mathbf{v} = \mathbf{0}$, which forces $\mathbf{v} = \mathbf{0}$ (Theorem 1.1.9), a contradiction; so $b \neq 0$ and $\mathbf{A}\mathbf{v} = \lambda_{\mathbf{v}}\mathbf{v}$ with $\lambda_{\mathbf{v}} = -a/b$. Apply this to $\mathbf{e}_i$, $\mathbf{e}_j$, and $\mathbf{e}_i + \mathbf{e}_j$ for $i \neq j$:

$$\lambda_{\mathbf{e}_i+\mathbf{e}_j}\mathbf{e}_i + \lambda_{\mathbf{e}_i+\mathbf{e}_j}\mathbf{e}_j = \mathbf{A}(\mathbf{e}_i + \mathbf{e}_j) = \mathbf{A}\mathbf{e}_i + \mathbf{A}\mathbf{e}_j = \lambda_{\mathbf{e}_i}\mathbf{e}_i + \lambda_{\mathbf{e}_j}\mathbf{e}_j.$$

Since $\mathbf{e}_i, \mathbf{e}_j$ are linearly independent, comparing coefficients (Theorem 1.4.3) gives $\lambda_{\mathbf{e}_i} = \lambda_{\mathbf{e}_i+\mathbf{e}_j} = \lambda_{\mathbf{e}_j}$. Hence all $\lambda_{\mathbf{e}_i}$ equal a common λ, so $\mathbf{A}\mathbf{e}_i = \lambda\mathbf{e}_i$ for every i, and by Theorem 0.9.2 the columns of $\mathbf{A}$ are those of $\lambda\mathbf{I}_n$. So $\mathbf{A} = \lambda\mathbf{I}_n$ is scalar, contrary to the hypothesis.

Take such a $\mathbf{v}$. By Theorem 1.6.5, the independent list $(\mathbf{v}, \mathbf{A}\mathbf{v})$ extends to a basis $(\mathbf{v}, \mathbf{A}\mathbf{v}, \mathbf{w}_3, \dots, \mathbf{w}_n)$ of F^n. Let $\mathbf{P}$ be the matrix with these columns; it is invertible by Theorem 3.4.4. Then $\mathbf{P}\mathbf{e}_1 = \mathbf{v}$ and $\mathbf{P}\mathbf{e}_2 = \mathbf{A}\mathbf{v}$ (Theorem 0.9.3), so the first column of $\mathbf{P}^{-1}\mathbf{A}\mathbf{P}$ is

$$\mathbf{P}^{-1}\mathbf{A}\mathbf{P}\mathbf{e}_1 = \mathbf{P}^{-1}\mathbf{A}\mathbf{v} = \mathbf{P}^{-1}\mathbf{P}\mathbf{e}_2 = \mathbf{e}_2.$$

Partitioning $\mathbf{P}^{-1}\mathbf{A}\mathbf{P}$ after the first row and column gives the displayed form. Finally, $\operatorname{tr}(\mathbf{P}^{-1}\mathbf{A}\mathbf{P}) = \operatorname{tr}\mathbf{A}$ by Theorem 4.2.5, and $\operatorname{tr}(\mathbf{P}^{-1}\mathbf{A}\mathbf{P}) = 0 + \operatorname{tr}\mathbf{A}_1$ by Definition 0.9.12, so $\operatorname{tr}\mathbf{A}_1 = \operatorname{tr}\mathbf{A}$.

■

Quick check. Over $\mathbb{Q}$, let $\mathbf{A} = \begin{pmatrix} 2 & 1 \\ 0 & -2 \end{pmatrix}$. Using $\mathbf{v} = \mathbf{e}_2$ in the proof of Lemma 8.5.2, find $\mathbf{P}$ and $\mathbf{P}^{-1}\mathbf{A}\mathbf{P}$, and check that its diagonal is zero.

Solution. $\mathbf{A}\mathbf{e}_2 = (1, -2)$, which is not a multiple of $\mathbf{e}_2$, so $(\mathbf{e}_2, \mathbf{A}\mathbf{e}_2)$ is independent and already a basis of $\mathbb{Q}^2$. Then $\mathbf{P} = \begin{pmatrix} 0 & 1 \\ 1 & -2 \end{pmatrix}$, with $\det\mathbf{P} = -1$ and $\mathbf{P}^{-1} = \begin{pmatrix} 2 & 1 \\ 1 & 0 \end{pmatrix}$. We get $\mathbf{A}\mathbf{P} = \begin{pmatrix} 1 & 0 \\ -2 & 4 \end{pmatrix}$ and $\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \begin{pmatrix} 0 & 4 \\ 1 & 0 \end{pmatrix}$. The first column is $\mathbf{e}_2$, as the lemma promises; the $(2,2)$-entry is 0 as well, because the trace is 0 and trace is a similarity invariant.

The lemma controls one diagonal entry and hands us a smaller matrix $\mathbf{A}_1$ with the same trace. That is exactly the setting for induction on the size. The only thing that can stop the induction is a scalar $\mathbf{A}_1$, and here the trace saves us, provided the field has characteristic 0.

Lemma 8.5.3. Trace Zero Means Similar to Zero Diagonal

Let F be a field of **characteristic** 0, let $n \geq 1$, and let $\mathbf{A} \in M_n(F)$ with $\operatorname{tr} \mathbf{A} = 0$. Then $\mathbf{A}$ is similar to a matrix all of whose diagonal entries are 0.

Idea. Induction on n. If $\mathbf{A}$ is scalar, $\mathbf{A} = \lambda \mathbf{I}_n$ with $n\lambda = 0$, and characteristic 0 forces $\lambda = 0$: then $\mathbf{A} = \mathbf{O}$ already has zero diagonal. Otherwise, Lemma 8.5.2 puts a 0 in the corner and leaves a trace-zero block $\mathbf{A}_1$ of size $n-1$. The induction hypothesis conjugates $\mathbf{A}_1$ by some $\mathbf{Q}$ into zero diagonal. Conjugating the whole matrix by the block diagonal matrix $(1) \oplus \mathbf{Q}$ does this to the lower block without disturbing the corner.

Proof. We use induction on n (Theorem 0.2.7). For $n = 1$, $\mathbf{A} = (\operatorname{tr} \mathbf{A}) = \mathbf{O}_1$, which has zero diagonal.

Let $n \geq 2$, and assume the statement for trace-zero matrices of size $n-1$.

Case 1: $\mathbf{A}$ *is scalar,* say $\mathbf{A} = \lambda \mathbf{I}_n$. Then $0 = \operatorname{tr} \mathbf{A} = n \cdot \lambda = (n \cdot 1)\lambda$ by Definition 0.9.12 and Lemma 0.7.4. Since F has characteristic 0 and $n \geq 1$, $n \cdot 1 \neq 0$ (Definition 0.7.2), so $\lambda = 0$. Hence $\mathbf{A} = \mathbf{O}$, which has zero diagonal and is similar to itself.

Case 2: $\mathbf{A}$ *is not scalar.* By Lemma 8.5.2 there is an invertible $\mathbf{P}$ with

$$\mathbf{M} \coloneqq \mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \begin{pmatrix} 0 & \mathbf{r}^{\top} \\ \mathbf{e}_1 & \mathbf{A}_1 \end{pmatrix}, \qquad \operatorname{tr} \mathbf{A}_1 = \operatorname{tr} \mathbf{A} = 0.$$

By the induction hypothesis there is an invertible $\mathbf{Q} \in M_{n-1}(F)$ such that $\mathbf{Z}_1 \coloneqq \mathbf{Q}^{-1}\mathbf{A}_1\mathbf{Q}$ has zero diagonal. Let $\mathbf{R} = (1) \oplus \mathbf{Q} = \begin{pmatrix} 1 & \mathbf{0}^{\top} \\ \mathbf{0} & \mathbf{Q} \end{pmatrix}$. By Theorem 8.1.4, $\mathbf{R}$ is invertible with $\mathbf{R}^{-1} = (1) \oplus \mathbf{Q}^{-1}$. By Theorem 8.1.1,

$$\begin{aligned} \mathbf{R}^{-1}\mathbf{M}\mathbf{R} &= \begin{pmatrix} 1 & \mathbf{0}^{\top} \\ \mathbf{0} & \mathbf{Q}^{-1} \end{pmatrix} \begin{pmatrix} 0 & \mathbf{r}^{\top} \\ \mathbf{e}_1 & \mathbf{A}_1 \end{pmatrix} \begin{pmatrix} 1 & \mathbf{0}^{\top} \\ \mathbf{0} & \mathbf{Q} \end{pmatrix} \\ &= \begin{pmatrix} 0 & \mathbf{r}^{\top}\mathbf{Q} \\ \mathbf{Q}^{-1}\mathbf{e}_1 & \mathbf{Q}^{-1}\mathbf{A}_1\mathbf{Q} \end{pmatrix} = \begin{pmatrix} 0 & \mathbf{r}^{\top}\mathbf{Q} \\ \mathbf{Q}^{-1}\mathbf{e}_1 & \mathbf{Z}_1 \end{pmatrix}. \end{aligned}$$

Its $(1,1)$-entry is 0, and its other diagonal entries are those of $\mathbf{Z}_1$, which are 0. Since $\mathbf{R}^{-1}\mathbf{M}\mathbf{R} = (\mathbf{P}\mathbf{R})^{-1}\mathbf{A}(\mathbf{P}\mathbf{R})$ by Theorem 0.9.7 (3), and $\mathbf{P}\mathbf{R}$ is invertible, $\mathbf{A}$ is similar to a matrix with zero diagonal. This completes the induction.

■

Remark. The characteristic was used only in Case 1, and only through $k \cdot 1 \neq 0$ for the sizes $k = 2, \ldots, n$ that the induction passes through. So the lemma holds, with the same proof, whenever F has characteristic 0 or characteristic $p > n$. Some hypothesis on the characteristic is needed: for $n = p$, over $\mathbb{F}_p$, the matrix $\mathbf{I}_p$ has trace $p \cdot 1 = 0$, but it is scalar, so it is similar only to itself, and its diagonal entries are 1. The same happens for $\lambda \mathbf{I}_n$ with $\lambda \neq 0$ whenever p divides n.

8.5.4 Shoda's theorem

With a zero diagonal available, the diagonal choice of $\mathbf{X}$ in $(*)$ finishes the job. All we need are n distinct scalars for the diagonal of $\mathbf{X}$, and characteristic 0 provides them.

Theorem 8.5.4. Shoda's Theorem

Let F be a field of **characteristic** 0, let $n \geq 1$, and let $\mathbf{A} \in M_n(F)$. Then $\mathbf{A}$ is a commutator, that is $\mathbf{A} = [\mathbf{X}, \mathbf{Y}]$ for some $\mathbf{X}, \mathbf{Y} \in M_n(F)$, if and only if $\operatorname{tr} \mathbf{A} = 0$.

Idea. One direction is Proposition 8.5.1 (c). For the other: ① by Lemma 8.5.3, $\mathbf{A} = \mathbf{PZP}^{-1}$ with $\mathbf{Z}$ of zero diagonal; ② take $\mathbf{X}_0 = \operatorname{diag}(0, 1, \ldots, n-1)$, whose entries are distinct in characteristic 0; ③ solve $[\mathbf{X}_0, \mathbf{Y}_0] = \mathbf{Z}$ entrywise with $(*)$, dividing by $x_i - x_j \neq 0$; ④ conjugate back with Proposition 8.5.1 (b).

Proof. ($\Rightarrow$) If $\mathbf{A} = [\mathbf{X}, \mathbf{Y}]$, then $\operatorname{tr} \mathbf{A} = 0$ by Proposition 8.5.1 (c).

($\Leftarrow$) Suppose $\operatorname{tr} \mathbf{A} = 0$. By Lemma 8.5.3 there is an invertible $\mathbf{P}$ such that $\mathbf{Z} = (z_{ij}) := \mathbf{P}^{-1}\mathbf{AP}$ has $z_{ii} = 0$ for all i.

For $1 \leq i \leq n$ let $x_i := (i-1) \cdot 1 \in F$, and $\mathbf{X}_0 := \operatorname{diag}(x_1, \ldots, x_n)$. For $i < j$, Lemma 0.7.4 gives $x_j - x_i = (j-i) \cdot 1$, which is nonzero because $j - i \geq 1$ and F has characteristic 0 (Definition 0.7.2). So the x_i are distinct, and we may define $\mathbf{Y}_0 = (y_{ij}) \in M_n(F)$ by

$$y_{ij} := \frac{z_{ij}}{x_i - x_j} \quad (i \neq j), \qquad y_{ii} := 0.$$

By $(*)$, $[\mathbf{X}_0, \mathbf{Y}_0]_{ij} = (x_i - x_j) y_{ij} = z_{ij}$ for $i \neq j$, and $[\mathbf{X}_0, \mathbf{Y}_0]_{ii} = 0 = z_{ii}$. Hence $[\mathbf{X}_0, \mathbf{Y}_0] = \mathbf{Z}$.

Let $\mathbf{X} = \mathbf{PX}_0\mathbf{P}^{-1}$ and $\mathbf{Y} = \mathbf{PY}_0\mathbf{P}^{-1}$. Applying Proposition 8.5.1 (b) with the invertible matrix $\mathbf{P}^{-1}$,

$$[\mathbf{X}, \mathbf{Y}] = [\mathbf{PX}_0\mathbf{P}^{-1}, \mathbf{PY}_0\mathbf{P}^{-1}] = \mathbf{P}[\mathbf{X}_0, \mathbf{Y}_0]\mathbf{P}^{-1} = \mathbf{PZP}^{-1} = \mathbf{A}.$$

This proves that $\mathbf{A}$ is a commutator.

■

The proof used characteristic 0 twice: in the lemma, to rule out nonzero scalar blocks of trace zero, and here, to find n distinct diagonal entries. The second use needs only $(j-i) \cdot 1 \neq 0$ for $1 \leq j - i \leq n - 1$, and with the remark after Lemma 8.5.3, the same proof works whenever the characteristic of F is 0 or larger than n. Neither use is essential for the **theorem**. In fact the statement "a square matrix over a field is a commutator if and only if its trace is 0" is true over **every** field. The proof in positive characteristic needs a different normal form and more work, and we do not give it. Example (The Identity as a Commutator in Characteristic p) is a small instance: $\mathbf{I}_p$ over $\mathbb{F}_p$ is a commutator, although Lemma 8.5.3 fails for it.

The proof is constructive, and it is worth running once in full.

Example. Writing a Trace-Zero 3×3 Matrix as a Commutator

Over $\mathbb{Q}$, write $\mathbf{A} = \begin{pmatrix} 1 & 2 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & -2 \end{pmatrix}$ as a commutator $[\mathbf{X}, \mathbf{Y}]$.

Solution. The trace is $1 + 1 - 2 = 0$. We follow the proofs step by step.

First corner. $\mathbf{A}$ is not scalar. Take $\mathbf{v} = \mathbf{e}_1$; then $\mathbf{Av} = (1, 0, 1)$ is not a multiple of $\mathbf{e}_1$. Extend $(\mathbf{e}_1, \mathbf{Ae}_1)$ by $\mathbf{e}_2$, which is not in their span, to a basis, and let

$$\mathbf{P}_1 = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}, \qquad \mathbf{M} = \mathbf{P}_1^{-1}\mathbf{A}\mathbf{P}_1 = \begin{pmatrix} 0 & 2 & 2 \\ 1 & -1 & 0 \\ 0 & 1 & 1 \end{pmatrix}.$$

(To check $\mathbf{M}$ without inverting, verify $\mathbf{P}_1\mathbf{M} = \mathbf{A}\mathbf{P}_1$: both equal $\begin{pmatrix} 1 & 1 & 2 \\ 0 & 1 & 1 \\ 1 & -1 & 0 \end{pmatrix}$.) The first column is $\mathbf{e}_2$, and the lower block is $\mathbf{A}_1 = \begin{pmatrix} -1 & 0 \\ 1 & 1 \end{pmatrix}$, with trace 0.

Second corner. $\mathbf{A}_1$ is not scalar. With $\mathbf{w} = \mathbf{e}_1 \in \mathbb{Q}^2$, $\mathbf{A}_1\mathbf{w} = (-1, 1)$ is independent of $\mathbf{e}_1$, so

$$\mathbf{Q} = \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix}, \qquad \mathbf{Q}^{-1} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \qquad \mathbf{Z}_1 = \mathbf{Q}^{-1}\mathbf{A}_1\mathbf{Q} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

Assemble. With $\mathbf{r}^\top = (2, 2)$, the proof of Lemma 8.5.3 gives $\mathbf{r}^\top\mathbf{Q} = (2, 0)$ and $\mathbf{Q}^{-1}\mathbf{e}_1 = \mathbf{e}_1$, so for $\mathbf{P} = \mathbf{P}_1((1) \oplus \mathbf{Q})$,

$$\mathbf{P} = \begin{pmatrix} 1 & 1 & -1 \\ 0 & 0 & 1 \\ 0 & 1 & -1 \end{pmatrix}, \qquad \mathbf{Z} = \mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \begin{pmatrix} 0 & 2 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}, \qquad \mathbf{P}^{-1} = \begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & 1 \\ 0 & 1 & 0 \end{pmatrix}.$$

Solve entrywise. With $\mathbf{X}_0 = \operatorname{diag}(0, 1, 2)$, the rule $y_{ij} = z_{ij}/(x_i - x_j)$ gives $y_{12} = \frac{2}{-1} = -2$, $y_{21} = \frac{1}{1} = 1$, $y_{23} = \frac{1}{-1} = -1$, $y_{32} = \frac{1}{1} = 1$, and $y_{13} = y_{31} = 0$:

$$\mathbf{Y}_0 = \begin{pmatrix} 0 & -2 & 0 \\ 1 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix}.$$

Conjugate back. $\mathbf{X} = \mathbf{P}\mathbf{X}_0\mathbf{P}^{-1}$ and $\mathbf{Y} = \mathbf{P}\mathbf{Y}_0\mathbf{P}^{-1}$ are

$$\mathbf{X} = \begin{pmatrix} 0 & -1 & 1 \\ 0 & 2 & 0 \\ 0 & -1 & 1 \end{pmatrix}, \qquad \mathbf{Y} = \begin{pmatrix} 1 & -4 & -4 \\ 0 & 1 & 1 \\ 1 & -2 & -2 \end{pmatrix}.$$

Check: $\mathbf{XY} = \begin{pmatrix} 1 & -3 & -3 \\ 0 & 2 & 2 \\ 1 & -3 & -3 \end{pmatrix}$ and $\mathbf{YX} = \begin{pmatrix} 0 & -5 & -3 \\ 0 & 1 & 1 \\ 0 & -3 & -1 \end{pmatrix}$, and $\mathbf{XY} - \mathbf{YX} = \mathbf{A}$.

Nothing about these $\mathbf{X}$ and $\mathbf{Y}$ is unique. Adding to $\mathbf{Y}$ any matrix that commutes with $\mathbf{X}$, for instance a polynomial in $\mathbf{X}$, gives another solution, by Proposition 8.5.1 (a).

8.5.5 The span of the commutators

Shoda's theorem is about single commutators. A weaker question, with an easier answer that holds over **every** field, is which matrices are **sums** of commutators. Since the trace-zero matrices form a subspace and contain all commutators, the span of the commutators lies inside it. Matrix units show that it is all of it.

Theorem 8.5.5. Commutators Span the Trace-Zero Matrices

Let F be any field and $n \geq 1$. The span of all commutators $[\mathbf{X}, \mathbf{Y}]$ with $\mathbf{X}, \mathbf{Y} \in M_n(F)$ is the subspace $\{\mathbf{A} \in M_n(F) : \operatorname{tr} \mathbf{A} = 0\}$. More precisely, every trace-zero matrix is a linear combination of the commutators

$$\mathbf{E}_{ij} = [\mathbf{E}_{ii}, \mathbf{E}_{ij}] \ (i \neq j) \qquad \text{and} \qquad \mathbf{E}_{ii} - \mathbf{E}_{nn} = [\mathbf{E}_{in}, \mathbf{E}_{ni}] \ (1 \leq i \leq n-1).$$

Proof. Let $W = \{\mathbf{A} \in M_n(F) : \operatorname{tr} \mathbf{A} = 0\}$, a subspace by Example (Trace-Free Matrices). By Proposition 8.5.1 (c), every commutator lies in W, so their span is contained in W (Theorem 1.3.1, the span being the smallest subspace containing them).

For the reverse inclusion, the displayed identities hold by Example 2 after Definition 8.5.1: for $i \neq j$, $[\mathbf{E}_{ii}, \mathbf{E}_{ij}] = \mathbf{E}_{ij}$, and for $i < n$, $[\mathbf{E}_{in}, \mathbf{E}_{ni}] = \mathbf{E}_{ii} - \mathbf{E}_{nn}$. Now let $\mathbf{A} = (a_{ij}) \in W$. Writing $\mathbf{A}$ in the basis of matrix units (Example (Standard Bases)) and using $a_{nn} = -\sum_{i<n} a_{ii}$,

$$\mathbf{A} = \sum_{i \neq j} a_{ij} \mathbf{E}_{ij} + \sum_{i=1}^{n-1} a_{ii} \mathbf{E}_{ii} + a_{nn} \mathbf{E}_{nn} = \sum_{i \neq j} a_{ij} \mathbf{E}_{ij} + \sum_{i=1}^{n-1} a_{ii} \left(\mathbf{E}_{ii} - \mathbf{E}_{nn} \right).$$

So $\mathbf{A}$ is a linear combination of commutators, hence lies in their span. (For $n = 1$, $W = \{\mathbf{O}\}$ and both sides are $\{\mathbf{O}\}$.)

■

The theorem lists $n(n-1) + (n-1) = n^2 - 1$ commutators spanning the trace-zero matrices. Since $\operatorname{tr} \colon M_n(F) \to F$ is surjective ($\operatorname{tr}(c\mathbf{E}_{11}) = c$), the Rank–Nullity Theorem (Theorem 2.3.1) gives that its kernel has dimension $n^2 - 1$, so by Theorem 1.6.4 (b) these commutators form a basis of it. Over a field of characteristic 0, Shoda's theorem says much more: the sum of commutators produced here can always be collapsed into a single commutator. In general the set of commutators is not obviously closed under addition, and that is exactly what makes Shoda's theorem a theorem.

Warning. **The span of a set is not the set.** Theorem 8.5.5 shows that every trace-zero matrix is a **sum** of commutators, over any field, by a two-line argument. It does not show that every trace-zero matrix **is** a commutator. For comparison, in $M_2(F)$ the matrices of

rank at most 1 span everything (the four matrix units have rank 1), yet $\mathbf{I}_2$ has rank 2: a set can span a subspace without filling it. For commutators it happens that the set does fill the subspace, but the proof needs the similarity argument of Lemma 8.5.3, not just matrix units.

The span theorem gives a clean proof of a fact from Chapter 4: a linear functional φ on $M_n(F)$ with $\varphi(\mathbf{AB}) = \varphi(\mathbf{BA})$ is a multiple of the trace. Such a φ kills every commutator, hence their span, which is the kernel of the trace; Exercise (C1: Trace-like functionals, again) below completes the argument.

8.5.6 Exercises

8.5.6.1 A. Check your understanding

Exercise. A1

(a) Define the commutator $[\mathbf{A}, \mathbf{B}]$, and state what it means for a matrix to be a commutator.

(b) State Shoda's theorem, including the hypothesis on the field.

(c) True or false: $[\mathbf{A}, \mathbf{B}] = [\mathbf{B}, \mathbf{A}]$ for all $\mathbf{A}, \mathbf{B} \in M_n(F)$. Justify your answer.

(d) True or false: over every field, every trace-zero matrix is similar to a matrix with zero diagonal. Justify your answer.

(e) Name the two places where the proof of Shoda's theorem uses characteristic 0.

(f) Explain why Theorem 8.5.5 does not by itself prove Shoda's theorem.

Solution.

(a) $[\mathbf{A}, \mathbf{B}] = \mathbf{AB} - \mathbf{BA}$ (Definition 8.5.1). A matrix $\mathbf{C} \in M_n(F)$ is a commutator if $\mathbf{C} = [\mathbf{X}, \mathbf{Y}]$ for some $\mathbf{X}, \mathbf{Y} \in M_n(F)$.

(b) If F has characteristic 0 and $\mathbf{A} \in M_n(F)$, then $\mathbf{A} = [\mathbf{X}, \mathbf{Y}]$ for some $\mathbf{X}, \mathbf{Y} \in M_n(F)$ if and only if $\operatorname{tr} \mathbf{A} = 0$ (Theorem 8.5.4).

(c) False. $[\mathbf{B}, \mathbf{A}] = -[\mathbf{A}, \mathbf{B}]$ (Proposition 8.5.1 (a)), so equality holds only when $[\mathbf{A}, \mathbf{B}] = \mathbf{O}$. For example $[\mathbf{E}_{12}, \mathbf{E}_{21}] = \operatorname{diag}(1, -1)$ and $[\mathbf{E}_{21}, \mathbf{E}_{12}] = \operatorname{diag}(-1, 1)$ over $\mathbb{Q}$.

(d) False. Over $\mathbb{F}_2$, $\mathbf{I}_2$ has trace $1 + 1 = 0$, but it is scalar, so every matrix similar to it equals $\mathbf{P}^{-1}\mathbf{I}_2\mathbf{P} = \mathbf{I}_2$, whose diagonal entries are 1.

(e) In Lemma 8.5.3, to conclude from $(n \cdot 1)\lambda = 0$ that a trace-zero scalar matrix $\lambda \mathbf{I}_n$ is $\mathbf{O}$; and in Theorem 8.5.4, to find n distinct scalars $x_i = (i-1) \cdot 1$ so that we can divide by $x_i - x_j$.

(f) It says every trace-zero matrix is a **linear combination** of commutators. Shoda's theorem says it is a **single** commutator. A sum of commutators need not visibly be a commutator, since $[\mathbf{X}, \mathbf{Y}] + [\mathbf{X}', \mathbf{Y}']$ has no evident expression as $[\mathbf{X}'', \mathbf{Y}'']$.

8.5.6.2 B. Practice

Exercise. B1: A 2×2 commutator

Over $\mathbb{Q}$, let $\mathbf{A} = \begin{pmatrix} 3 & -2 \\ 1 & -3 \end{pmatrix}$.

(a) Using $\mathbf{v} = \mathbf{e}_1$ as in Lemma 8.5.2, find an invertible $\mathbf{P}$ such that $\mathbf{Z} = \mathbf{P}^{-1}\mathbf{A}\mathbf{P}$ has zero diagonal.

(b) With $\mathbf{X}_0 = \operatorname{diag}(0, 1)$, find $\mathbf{Y}_0$ with $[\mathbf{X}_0, \mathbf{Y}_0] = \mathbf{Z}$.

(c) Hence write $\mathbf{A} = [\mathbf{X}, \mathbf{Y}]$, and check your answer by multiplying out.

Solution.

(a) $\operatorname{tr}\mathbf{A} = 0$, and $\mathbf{A}\mathbf{e}_1 = (3, 1)$ is not a multiple of $\mathbf{e}_1$, so $(\mathbf{e}_1, \mathbf{A}\mathbf{e}_1)$ is a basis of $\mathbb{Q}^2$. Let $\mathbf{P} = \begin{pmatrix} 1 & 3 \\ 0 & 1 \end{pmatrix}$, so $\mathbf{P}^{-1} = \begin{pmatrix} 1 & -3 \\ 0 & 1 \end{pmatrix}$. Then $\mathbf{A}\mathbf{P} = \begin{pmatrix} 3 & 7 \\ 1 & 0 \end{pmatrix}$ and

$$\mathbf{Z} = \mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \begin{pmatrix} 0 & 7 \\ 1 & 0 \end{pmatrix}.$$

The first column is $\mathbf{e}_2$ by Lemma 8.5.2, and the $(2,2)$-entry is 0 because the trace is 0.

(b) By $(*)$, $[\mathbf{X}_0, \mathbf{Y}]_{12} = (0-1)y_{12}$ and $[\mathbf{X}_0, \mathbf{Y}]_{21} = (1-0)y_{21}$, while the diagonal of $[\mathbf{X}_0, \mathbf{Y}]$ is zero. So $y_{12} = -7$, $y_{21} = 1$, and we may take $y_{11} = y_{22} = 0$: $\mathbf{Y}_0 = \begin{pmatrix} 0 & -7 \\ 1 & 0 \end{pmatrix}$.

(c) By the proof of Theorem 8.5.4, $\mathbf{X} = \mathbf{P}\mathbf{X}_0\mathbf{P}^{-1}$ and $\mathbf{Y} = \mathbf{P}\mathbf{Y}_0\mathbf{P}^{-1}$:

$$\mathbf{X} = \begin{pmatrix} 0 & 3 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & -3 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 3 \\ 0 & 1 \end{pmatrix}, \qquad \mathbf{Y} = \begin{pmatrix} 3 & -7 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 1 & -3 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 3 & -16 \\ 1 & -3 \end{pmatrix},$$

using $\mathbf{P}\mathbf{X}_0 = \begin{pmatrix} 0 & 3 \\ 0 & 1 \end{pmatrix}$ and $\mathbf{P}\mathbf{Y}_0 = \begin{pmatrix} 3 & -7 \\ 1 & 0 \end{pmatrix}$. *Check:* $\mathbf{X}\mathbf{Y} = \begin{pmatrix} 3 & -9 \\ 1 & -3 \end{pmatrix}$, $\mathbf{Y}\mathbf{X} = \begin{pmatrix} 0 & -7 \\ 0 & 0 \end{pmatrix}$, and $\mathbf{X}\mathbf{Y} - \mathbf{Y}\mathbf{X} = \begin{pmatrix} 3 & -2 \\ 1 & -3 \end{pmatrix} = \mathbf{A}$.

Exercise. B2: A commutator over $\mathbb{F}_2$

Work over $\mathbb{F}_2$, and let $\mathbf{A} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$.

(a) Show that $\operatorname{tr}\mathbf{A} = 0$ and that $\mathbf{A}$ is not scalar.

(b) Explain why $\mathbf{v} = \mathbf{e}_1$ cannot be used in Lemma 8.5.2, and use $\mathbf{v} = \mathbf{e}_2$ instead to find $\mathbf{P}$ with $\mathbf{P}^{-1}\mathbf{A}\mathbf{P}$ of zero diagonal.

(c) The field $\mathbb{F}_2$ has exactly two elements. Use them as the diagonal of $\mathbf{X}_0$ to write $\mathbf{A}$ as a commutator, by hand.

(d) Explain why the same method cannot write $\mathbf{I}_2 \in M_2(\mathbb{F}_2)$ as a commutator, and recall a pair $\mathbf{X}, \mathbf{Y}$ that does.

Solution.

(a) $\operatorname{tr}\mathbf{A} = 1+1 = 0$ in $\mathbb{F}_2$. $\mathbf{A}$ has a nonzero off-diagonal entry, while every scalar matrix has zero off-diagonal entries.

(b) $\mathbf{Ae}_1 = \mathbf{e}_1$, so $(\mathbf{e}_1, \mathbf{Ae}_1)$ is dependent. With $\mathbf{v} = \mathbf{e}_2$, $\mathbf{Ae}_2 = (1,1)$, which is not a multiple of $\mathbf{e}_2$, so $(\mathbf{e}_2, \mathbf{Ae}_2)$ is a basis. Let $\mathbf{P} = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}$; since $\det\mathbf{P} = -1 = 1$, $\mathbf{P}^{-1} = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}$ (check: $\mathbf{PP}^{-1} = \begin{pmatrix} 1 & 0 \\ 2 & 1 \end{pmatrix} = \mathbf{I}_2$ in $\mathbb{F}_2$). Then $\mathbf{AP} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$ and $\mathbf{Z} = \mathbf{P}^{-1}\mathbf{AP} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$.

(c) Take $\mathbf{X}_0 = \operatorname{diag}(0,1)$. By $(*)$, $[\mathbf{X}_0, \mathbf{Y}]_{12} = (0-1)y_{12} = y_{12}$ (as $-1 = 1$) and $[\mathbf{X}_0, \mathbf{Y}]_{21} = y_{21}$. So $\mathbf{Y}_0 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ gives $[\mathbf{X}_0, \mathbf{Y}_0] = \mathbf{Z}$. Conjugating back, $\mathbf{PX}_0 = \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix}$ and $\mathbf{PY}_0 = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$, so

$$\mathbf{X} = \mathbf{PX}_0\mathbf{P}^{-1} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}, \qquad \mathbf{Y} = \mathbf{PY}_0\mathbf{P}^{-1} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}.$$

Check: $\mathbf{XY} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$, $\mathbf{YX} = \begin{pmatrix} 2 & 0 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}$, so $[\mathbf{X}, \mathbf{Y}] = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} = \mathbf{A}$. (Here $\mathbf{Y} = \mathbf{A}$: indeed $[\mathbf{X}, \mathbf{A}] = \mathbf{A}$.)

(d) The method writes $\mathbf{Z} = \mathbf{P}^{-1}\mathbf{I}_2\mathbf{P}$ as $[\mathbf{X}_0, \mathbf{Y}_0]$ with $\mathbf{X}_0$ diagonal, and by $(*)$ such a commutator has zero diagonal. But $\mathbf{I}_2$ is scalar, so $\mathbf{P}^{-1}\mathbf{I}_2\mathbf{P} = \mathbf{I}_2$ for every invertible $\mathbf{P}$, and its diagonal entries are 1. So the method cannot start. Nevertheless $\mathbf{I}_2$ is a commutator: with $\mathbf{X} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}$ and $\mathbf{Y} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$, $[\mathbf{X}, \mathbf{Y}] = \operatorname{diag}(-1,1) = \mathbf{I}_2$ in $\mathbb{F}_2$ (Exercise (C1: The Heisenberg relation) (d), or Example (The Identity as a Commutator in Characteristic p) with $p = 2$).

Exercise. B3: Commuting with a diagonal matrix

Let F be any field, let $d_1, \dots, d_n \in F$ be **distinct**, let $\mathbf{D} = \operatorname{diag}(d_1, \dots, d_n)$, and define $L\colon M_n(F) \to M_n(F)$ by $L(\mathbf{X}) = [\mathbf{D}, \mathbf{X}]$.

(a) Show that L is linear.

(b) Prove that $\ker L$ is the set of diagonal matrices and $\operatorname{im} L$ is the set of matrices with zero diagonal.

(c) Check that your answer to (b) agrees with the Rank–Nullity Theorem.

(d) Hence decide whether $\begin{pmatrix} 0 & 1 & 2 \\ 3 & 0 & 4 \\ 5 & 6 & 0 \end{pmatrix} \in M_3(\mathbb{Q})$ is a commutator with $\mathbf{X} = \operatorname{diag}(1,2,3)$, and if so find $\mathbf{Y}$.

Solution.

(a) By Proposition 8.5.1 (a), $[\mathbf{D}, \mathbf{X} + \mathbf{X}'] = [\mathbf{D}, \mathbf{X}] + [\mathbf{D}, \mathbf{X}']$ and $[\mathbf{D}, c\mathbf{X}] = c[\mathbf{D}, \mathbf{X}]$.

(b) By $(*)$, $L(\mathbf{X})_{ij} = (d_i - d_j)x_{ij}$. Since $d_i - d_j \neq 0$ for $i \neq j$, $L(\mathbf{X}) = \mathbf{O}$ if and only if $x_{ij} = 0$ for all $i \neq j$, that is, $\mathbf{X}$ is diagonal. Every $L(\mathbf{X})$ has zero diagonal. Conversely, if $\mathbf{Z}$ has zero diagonal, the matrix $\mathbf{Y}$ with $y_{ij} = z_{ij}/(d_i - d_j)$ for $i \neq j$ and $y_{ii} = 0$ satisfies $L(\mathbf{Y}) = \mathbf{Z}$. So $\operatorname{im} L$ is exactly the set of zero-diagonal matrices.

(c) The diagonal matrices have basis $\mathbf{E}_{11}, \dots, \mathbf{E}_{nn}$, so $\operatorname{nullity} L = n$. The zero-diagonal matrices have basis $\mathbf{E}_{ij}$ $(i \neq j)$, so $\operatorname{rank} L = n^2 - n$. Their sum is $n^2 = \dim M_n(F)$, as Theorem 2.3.1 requires.

(d) Yes, since the matrix has zero diagonal and $1, 2, 3$ are distinct. With $d = (1, 2, 3)$, $y_{ij} = z_{ij}/(d_i - d_j)$ gives

$$\mathbf{Y} = \begin{pmatrix} 0 & \frac{1}{-1} & \frac{2}{-2} \\ \frac{3}{1} & 0 & \frac{4}{-1} \\ \frac{5}{2} & \frac{6}{1} & 0 \end{pmatrix} = \begin{pmatrix} 0 & -1 & -1 \\ 3 & 0 & -4 \\ \frac{5}{2} & 6 & 0 \end{pmatrix},$$

and by (b), $[\operatorname{diag}(1, 2, 3), \mathbf{Y}]$ is the given matrix.

8.5.6.3 C. Going deeper

Exercise. C1: Trace-like functionals, again

Let F be any field, $n \geq 1$, and let $\varphi\colon M_n(F) \to F$ be linear with $\varphi(\mathbf{AB}) = \varphi(\mathbf{BA})$ for all $\mathbf{A}, \mathbf{B} \in M_n(F)$.

(a) Prove that φ vanishes on every trace-zero matrix.

(b) Deduce that $\varphi = c \operatorname{tr}$, where $c = \varphi(\mathbf{E}_{11})$.

(c) Suppose F has characteristic 0 and, in addition, $\varphi(\mathbf{I}_n) = n \cdot 1$. Deduce that $\varphi = \operatorname{tr}$.

(d) Show that the conclusion of (c) fails over $\mathbb{F}_p$ when $n = p$.

Hint: for (a), use Theorem 8.5.5.

Solution.

(a) For all $\mathbf{X}, \mathbf{Y}$, $\varphi([\mathbf{X}, \mathbf{Y}]) = \varphi(\mathbf{XY}) - \varphi(\mathbf{YX}) = 0$ by linearity and the hypothesis. So the kernel of φ, a subspace, contains every commutator, hence their span (Theorem 1.3.1), which is the set of trace-zero matrices by Theorem 8.5.5.

(b) Let $\mathbf{A} \in M_n(F)$. Since $\operatorname{tr} \mathbf{E}_{11} = 1$, the matrix $\mathbf{A} - (\operatorname{tr} \mathbf{A})\mathbf{E}_{11}$ has trace 0 (Theorem 0.9.6 (1)), so by (a) and linearity,

$$0 = \varphi(\mathbf{A} - (\operatorname{tr} \mathbf{A})\mathbf{E}_{11}) = \varphi(\mathbf{A}) - (\operatorname{tr} \mathbf{A})\,\varphi(\mathbf{E}_{11}).$$

Hence $\varphi(\mathbf{A}) = c \operatorname{tr} \mathbf{A}$ with $c = \varphi(\mathbf{E}_{11})$. This recovers Exercise (C2: The trace is the only trace-like functional) by a different route: the matrix-unit computations are now packaged in Theorem 8.5.5.

(c) By (b), $n \cdot 1 = \varphi(\mathbf{I}_n) = c \operatorname{tr} \mathbf{I}_n = c\,(n \cdot 1)$. Since F has characteristic 0, $n \cdot 1 \neq 0$, so $c = 1$ and $\varphi = \operatorname{tr}$.

(d) Over $\mathbb{F}_p$ with $n = p$, take $\varphi = 0$. It is linear, satisfies $\varphi(\mathbf{AB}) = 0 = \varphi(\mathbf{BA})$, and $\varphi(\mathbf{I}_p) = 0 = p \cdot 1$. But $\varphi \neq \operatorname{tr}$, since $\operatorname{tr} \mathbf{E}_{11} = 1$. The step that fails is canceling $n \cdot 1 = 0$ in (c).

Exercise. C2: When A commutes with $[\mathbf{A},\mathbf{B}]$

Let F be a field of characteristic 0, let $\mathbf{A},\mathbf{B}\in M_n(F)$, put $\mathbf{C}=[\mathbf{A},\mathbf{B}]$, and suppose that $\mathbf{AC}=\mathbf{CA}$. The goal is to prove that $\mathbf{C}$ is **nilpotent**, that is, $\mathbf{C}^s=\mathbf{O}$ for some $s\ge 1$.

(a) Prove that $\mathbf{C}^k=[\mathbf{A},\mathbf{BC}^{k-1}]$ for every $k\ge 1$. Deduce that $\operatorname{tr}\mathbf{C}^k=0$ for every $k\ge 1$.

(b) Show that there are an integer $s\ge 1$ and $g\in F[x]$ with $g(0)\neq 0$ such that $\mathbf{C}^s g(\mathbf{C})=\mathbf{O}$.

(c) Let $U=\operatorname{col}(\mathbf{C}^s)$ and $r=\dim U$, and suppose $r\ge 1$. Extend a basis of U to a basis of F^n, and let $\mathbf{P}$ be the matrix with these columns. Show that

$$\mathbf{P}^{-1}\mathbf{C}\mathbf{P}=\begin{pmatrix}\mathbf{D} & *\\ \mathbf{O} & \mathbf{N}\end{pmatrix},\qquad \mathbf{D}\in M_r(F),$$

with $\mathbf{N}^s=\mathbf{O}$ and $g(\mathbf{D})=\mathbf{O}$.

(d) Write $g=g_0+xh$ with $g_0=g(0)\neq 0$. Show that $\mathbf{I}_r$ is a linear combination of $\mathbf{D}^s,\mathbf{D}^{s+1},\dots,\mathbf{D}^d$ for some $d\ge s$, and use (a) to reach a contradiction.

(e) Conclude that $\mathbf{C}^s=\mathbf{O}$.

Hint: in (c), compare the columns of $\mathbf{P}^{-1}\mathbf{C}^k\mathbf{P}$ for $k=1$ and $k=s$ with the coordinates of vectors of U; in (d), take traces.

Solution.

(a) First, $\mathbf{C}^{k-1}\mathbf{A}=\mathbf{AC}^{k-1}$ for all $k\ge 1$, by induction on k from $\mathbf{CA}=\mathbf{AC}$. Hence

$$\begin{aligned}[\mathbf{A},\mathbf{BC}^{k-1}]&=\mathbf{ABC}^{k-1}-\mathbf{BC}^{k-1}\mathbf{A}=\mathbf{ABC}^{k-1}-\mathbf{BAC}^{k-1}\\&=(\mathbf{AB}-\mathbf{BA})\mathbf{C}^{k-1}=\mathbf{C}^k.\end{aligned}$$

By Proposition 8.5.1 (c), $\operatorname{tr}\mathbf{C}^k=0$.

(b) By Theorem 6.7.2 there is a nonzero $f\in F[x]$ with $f(\mathbf{C})=\mathbf{O}$. Let $t\ge 0$ be the largest integer with x^t dividing f, and write $f=x^t g$; then $x\nmid g$, so $g(0)\neq 0$ by Theorem 6.1.3. Put $s=t+1\ge 1$. By Theorem 0.9.9, $\mathbf{C}^s g(\mathbf{C})=\mathbf{C}\,f(\mathbf{C})=\mathbf{O}$.

(c) Let $\mathbf{P}$ have columns $\mathbf{p}_1,\dots,\mathbf{p}_n$, where $(\mathbf{p}_1,\dots,\mathbf{p}_r)$ is a basis of U; $\mathbf{P}$ is invertible by Theorem 3.4.4. For any $k\ge 1$, column j of $\mathbf{P}^{-1}\mathbf{C}^k\mathbf{P}$ is $\mathbf{P}^{-1}(\mathbf{C}^k\mathbf{p}_j)$, the vector $\mathbf{c}$ with $\mathbf{Pc}=\mathbf{C}^k\mathbf{p}_j$, that is, the coefficients of $\mathbf{C}^k\mathbf{p}_j$ in terms of $\mathbf{p}_1,\dots,\mathbf{p}_n$ (Theorem 0.9.3). If $\mathbf{y}\in U$, then $\mathbf{y}=\mathbf{C}^s\mathbf{z}$ and $\mathbf{Cy}=\mathbf{C}^s(\mathbf{Cz})\in U$, so $\mathbf{C}$ maps U into U; hence for $j\le r$ the vector $\mathbf{Cp}_j$ is a combination of $\mathbf{p}_1,\dots,\mathbf{p}_r$, and column j of $\mathbf{P}^{-1}\mathbf{CP}$ has zeros below row r. This gives the block form. By Theorem 8.1.1 and induction on k, $(\mathbf{P}^{-1}\mathbf{CP})^k=\begin{pmatrix}\mathbf{D}^k & *\\ \mathbf{O} & \mathbf{N}^k\end{pmatrix}$, and by Proposition 4.2.6 (c) this is $\mathbf{P}^{-1}\mathbf{C}^k\mathbf{P}$; more generally $\mathbf{P}^{-1}q(\mathbf{C})\mathbf{P}=\begin{pmatrix}q(\mathbf{D}) & *\\ \mathbf{O} & q(\mathbf{N})\end{pmatrix}$ for every $q\in F[x]$.
For $k=s$, every vector $\mathbf{C}^s\mathbf{p}_j$ lies in U, so every column of $\mathbf{P}^{-1}\mathbf{C}^s\mathbf{P}$ is zero below row r, and $\mathbf{N}^s=\mathbf{O}$. For $q=g$: for $j\le r$, $\mathbf{p}_j\in U$ is $\mathbf{C}^s\mathbf{z}$ for some $\mathbf{z}$, so $g(\mathbf{C})\mathbf{p}_j=g(\mathbf{C})\mathbf{C}^s\mathbf{z}=\mathbf{0}$ by (b) and Theorem 0.9.9. So the first r columns of $\mathbf{P}^{-1}g(\mathbf{C})\mathbf{P}$ are zero, and in particular $g(\mathbf{D})=\mathbf{O}$.

(d) From $g(\mathbf{D})=g_0\mathbf{I}_r+\mathbf{D}h(\mathbf{D})=\mathbf{O}$ we get $\mathbf{I}_r=\mathbf{D}\,h_1(\mathbf{D})$ with $h_1=-g_0^{-1}h$. Since

polynomials in $\mathbf{D}$ commute (Theorem 0.9.9),

$$\mathbf{I}_r = (\mathbf{D}\, h_1(\mathbf{D}))^s = \mathbf{D}^s\, h_1(\mathbf{D})^s = \sum_{k=s}^{d} c_k \mathbf{D}^k$$

for some $c_k \in F$, as every term of $x^s h_1(x)^s$ has degree at least s. For $k \geq s$, $\mathbf{N}^k = \mathbf{N}^s \mathbf{N}^{k-s} = \mathbf{O}$, so by (c), Theorem 4.2.5, and Definition 0.9.12, $\operatorname{tr} \mathbf{C}^k = \operatorname{tr}(\mathbf{P}^{-1}\mathbf{C}^k\mathbf{P}) = \operatorname{tr} \mathbf{D}^k + \operatorname{tr} \mathbf{N}^k = \operatorname{tr} \mathbf{D}^k$. Taking traces,

$$r \cdot 1 = \operatorname{tr} \mathbf{I}_r = \sum_{k=s}^{d} c_k \operatorname{tr} \mathbf{D}^k = \sum_{k=s}^{d} c_k \operatorname{tr} \mathbf{C}^k = 0$$

by (a). But $r \geq 1$ and F has characteristic 0, so $r \cdot 1 \neq 0$, a contradiction.

(e) By (d), $r = 0$, so $\operatorname{col}(\mathbf{C}^s) = \{\mathbf{0}\}$ and $\mathbf{C}^s = \mathbf{O}$. For example, with $\mathbf{A} = \mathbf{E}_{12} + \mathbf{E}_{23}$ and $\mathbf{B} = \operatorname{diag}(0, 1, 2)$ in $M_3(\mathbb{Q})$, $\mathbf{C} = [\mathbf{A}, \mathbf{B}] = \mathbf{E}_{12} + \mathbf{E}_{23} = \mathbf{A}$, which commutes with $\mathbf{A}$, and indeed $\mathbf{C}^3 = \mathbf{O}$. In Chapter 9 the statement "$\operatorname{tr} \mathbf{C}^k = 0$ for all $k \geq 1$ forces nilpotency in characteristic 0" becomes a statement about eigenvalues.

Chapter 9

Eigenvalues and Diagonalization

Chapter 4 showed that one operator has many matrices, one for each basis, and asked how simple the matrix can be made. This chapter answers the first half of that question. The simplest possible matrix is diagonal, and a matrix of an operator is diagonal exactly when every basis vector is only stretched, never turned. Such vectors are **eigenvectors**, and the stretching factors are **eigenvalues**.

Not every operator has enough eigenvectors, and the chapter explains precisely when it does. Two polynomials carry the information: the characteristic polynomial from Chapter 7 and the **minimal polynomial**, the smallest polynomial that kills the operator. The Cayley–Hamilton Theorem ties them together. When eigenvectors run short, operators over $\mathbb{C}$ can still be made triangular, and the space splits into pieces on which the operator is easier to understand. The chapter ends with what diagonalization is for: powers of matrices, linear recurrences such as the Fibonacci numbers, and the long-run behavior of Markov chains.

What you need. From Chapter 0, the arithmetic every computation here runs on: degree, evaluation and the absence of zero divisors in $F[x]$ (Definition 0.8.1, Theorem 0.8.2, Theorem 0.8.3, Corollary 0.8.4, Theorem 0.8.5), field arithmetic (Theorem 0.7.1) and conjugation (Theorem 0.6.2), the matrix product with its three views, the transpose and polynomials of a matrix (Definition 0.9.9, Theorem 0.9.4, Theorem 0.9.3, Theorem 0.9.2, Theorem 0.9.5, Definition 0.9.14), and that a composition of bijections is a bijection (Theorem 0.4.3). From Chapter 1, the dimension theory, used constantly: subspaces and intersections (Theorem 1.2.1, Theorem 1.2.2), span and independence with the Linear Dependence Lemma (Definition 1.3.2, Theorem 1.4.1, Theorem 1.4.3, Lemma 1.5.3, Theorem 1.1.9), bases and the counting arguments (Theorem 1.5.1, Theorem 1.6.5, Theorem 1.6.4, Theorem 1.6.8, Theorem 1.6.3, Theorem 1.6.7), and above all sums and direct sums (Theorem 1.7.1, Theorem 1.7.5), since every decomposition theorem in the chapter is the claim that some sum is direct. From Chapter 3, that a one-sided inverse of a square matrix is two-sided (Theorem 3.4.5), and rank with Rank–Nullity for matrices (Theorem 3.5.7, Theorem 3.5.5). From Chapter 2, the matrix of a linear map, change of basis and similarity (Definition 4.1.1, Theorem 4.2.3, Theorem 4.2.4), and quotient spaces (Definition 4.4.3). From Chapter 6, polynomials of an operator, the ideal of polynomials that kill it, and the kernel splitting lemma (Theorem 6.7.1, Theorem 6.7.3, Theorem 6.7.5), and the Fundamental Theorem of Algebra (Theorem 6.5.1). From Chapter 7, the characteristic polynomial and the adjugate (Definition 7.10.1, Theorem 7.6.2). From Chapter 8, block triangular matrices (Theorem 8.1.7).

9.0.1 Roadmap

- **Invariant subspaces.** Subspaces an operator maps into themselves, and the block triangular matrices they produce.
- **Eigenvalues and eigenvectors.** Directions that are only stretched; eigenvalues as roots of the characteristic polynomial; every complex operator has one.
- **Algebraic and geometric multiplicity.** Two ways to count a repeated eigenvalue, and why the second never exceeds the first.
- **Diagonalization.** Exactly when an operator has a basis of eigenvectors, and how to compute with one.
- **The minimal polynomial.** The monic polynomial of least degree that kills an operator, and why its roots are the eigenvalues.
- **The Cayley–Hamilton Theorem.** Every square matrix satisfies its own characteristic polynomial.
- **Triangularization.** An operator has a triangular matrix exactly when its characteristic polynomial splits.
- **Primary decomposition.** The factorization of the minimal polynomial splits the space, and decides diagonalizability.
- **Commuting operators.** Operators that commute share eigenvectors, and can be triangularized or diagonalized together.
- **Spectral mapping, AB and BA.** Eigenvalues of polynomials in an operator, and why **AB** and **BA** share their nonzero eigenvalues.
- **Powers, recurrences, and Markov chains.** Diagonalization at work.

9.0.2 Named moves

- **Ask when the matrix is diagonal.** Unwinding the columns of a diagonal matrix gives the eigenvector equation $Tv = \lambda v$.
- **Eigenvalue means a kernel.** λ is an eigenvalue exactly when $T - \lambda I$ is not injective.
- **Apply $T - \lambda I$ to a relation.** Kills one eigenvector and shortens a dependence, the heart of "distinct eigenvalues give independent eigenvectors".
- **Turn a relation into a divisibility.** If $p(T) = 0$, then the minimal polynomial divides p.
- **One eigenvector, extend, recurse.** Split off an eigenvector, pass to the quotient, and use induction on dimension.
- **Search inside an eigenspace.** A commuting operator maps each eigenspace into itself, so look for a common eigenvector there.

9.1 Invariant Subspaces

Chapter 4 raised a question it could not answer: given an operator T on a finite-dimensional space, how simple can its matrix $[T]_{\mathcal{B}}^{\mathcal{B}}$ be made by choosing the basis $\mathcal{B}$ well? Chapter 8 found the first piece of an answer. A zero block in the lower-left corner of $[T]_{\mathcal{B}}^{\mathcal{B}}$ appears exactly when the first few basis vectors span a subspace that T maps into itself. This section gives that property a name, collects the standard examples, and shows that it can fail badly: some operators map almost no subspace into itself. It ends by recording how such subspaces cut an operator, and its characteristic polynomial, into smaller pieces. The next section looks at the smallest nonzero case, a line, and finds eigenvectors there.

9.1.1 Subspaces an operator maps into itself

After a vector space we studied its subspaces, and after linear maps their kernels and images. For an operator $T \in \mathcal{L}(V)$, the domain and the codomain are the same space, so a new question makes sense: which subspaces of V are compatible with T?

Here is the result that makes the question worth asking. Let V be finite-dimensional, $T \in \mathcal{L}(V)$, and U a subspace with $\{0\} \neq U \neq V$. Take a basis $(u_1, \dots, u_k)$ of U and extend it to a basis $\mathcal{B} = (u_1, \dots, u_k, w_1, \dots, w_l)$ of V. By Theorem 8.1.7, the lower-left $l \times k$ block of $[T]_{\mathcal{B}}^{\mathcal{B}}$ is zero **if and only if** $Tu \in U$ for every $u \in U$. For example, the cyclic shift $T(x, y, z) = (y, z, x)$ on $\mathbb{R}^3$ keeps the plane $x + y + z = 0$ in place, and in Example (A Plane Mapped into Itself) this produced a matrix with a zero corner. The property "T maps U into U" is what we need to control, so it gets a name.

A subspace is invariant under T when T never moves a vector of the subspace out of it.

Definition 9.1.1. Invariant Subspace

Let V be a vector space over F and $T \in \mathcal{L}(V)$. A subspace U of V is **invariant under** T, or **T-invariant**, if

$$Tu \in U \qquad \text{for every } u \in U,$$

that is, if $T(U) \subseteq U$.

In words: we start from a **subspace** U, not an arbitrary subset. We apply T to **every** vector of U, and each image must land back **in** U. Nothing is required of vectors outside U, and nothing says that a vector of U stays where it is: Tu may be a different vector of U. The condition is an inclusion $T(U) \subseteq U$, not an equality.

Checking "for every $u \in U$" looks like infinitely many checks. Linearity reduces it to finitely many when U has a finite spanning list.

Lemma 9.1.1. Invariance Can Be Checked on a Spanning List

Let $T \in \mathcal{L}(V)$ and $U = \operatorname{span}(u_1, \dots, u_k)$. Then U is T-invariant if and only if $Tu_i \in U$ for each $i = 1, \dots, k$.

Proof. ($\Rightarrow$) Each u_i lies in U. ($\Leftarrow$) Let $u = c_1u_1 + \cdots + c_ku_k \in U$. By Theorem 2.1.4, $Tu = c_1Tu_1 + \cdots + c_kTu_k$, a linear combination of vectors of U, which lies in U because U is a subspace (Theorem 1.2.1). ∎

Examples.

- **The trivial subspaces.** For every $T \in \mathcal{L}(V)$, both $\{0\}$ and V are T-invariant: $T0 = 0$ (Theorem 2.1.1), and $Tv \in V$ always. These are the degenerate cases, and they matter as a warning about what the word does **not** tell us: invariance carries information only for subspaces strictly between $\{0\}$ and V. At the other extreme, for $T = c\,\mathrm{id}_V$ **every** subspace is invariant, since $cu \in U$ whenever $u \in U$.
- **Kernel and image.** $\ker T$ is invariant: if $u \in \ker T$, then $Tu = 0 \in \ker T$. So is $\operatorname{im} T$: if

$w \in \operatorname{im} T$, then Tw is T of something, so $Tw \in \operatorname{im} T$. More generally, for every polynomial $p \in F[x]$, the subspaces $\ker p(T)$ and $\operatorname{im} p(T)$ are T-invariant, by Theorem 6.7.4 (b). For instance $\ker(T - c\,\mathrm{id}_V)$ is invariant for every scalar c.

- **A line in the plane.** Let $\mathbf{A} = \begin{pmatrix} 2 & 1 \\ 0 & 3 \end{pmatrix}$ act on $\mathbb{R}^2$. The line $\operatorname{span}(\mathbf{e}_1)$ is invariant, by Lemma 9.1.1, because $\mathbf{A}\mathbf{e}_1 = (2,0) = 2\mathbf{e}_1 \in \operatorname{span}(\mathbf{e}_1)$. The line $\operatorname{span}((1,1))$ is invariant as well, since $\mathbf{A}(1,1) = (3,3) = 3\,(1,1)$.
- **Polynomials of bounded degree.** Let $D\colon F[x] \to F[x]$ be differentiation (Example (Differentiation)). For each $k \in \mathbb{N}$, the subspace $F[x]_{\le k}$ is D-invariant, because differentiating does not raise the degree. It is spanned by $1, x, \ldots, x^k$, and $D(x^j) = jx^{j-1} \in F[x]_{\le k}$.

Nonexample by minimal change. Keep $\mathbf{A} = \begin{pmatrix} 2 & 1 \\ 0 & 3 \end{pmatrix}$ and replace the line $\operatorname{span}(\mathbf{e}_1)$ by the line $\operatorname{span}(\mathbf{e}_2)$. It is still a subspace, and $\mathbf{A}\mathbf{e}_2 = (1,3)$ is still some vector of $\mathbb{R}^2$. What fails is the clause "$T\mathbf{u} \in U$ for every $\mathbf{u} \in U$" at $\mathbf{u} = \mathbf{e}_2$: the vector $(1,3)$ is not a multiple of $\mathbf{e}_2$. In the same way, $\operatorname{span}(x, x^2)$ inside $\mathbb{R}[x]_{\le 3}$ is not D-invariant, because $D(x) = 1 \notin \operatorname{span}(x, x^2)$.

The examples with lines all have the same shape, which is worth recording now, because the whole next section grows out of it.

Proposition 9.1.2. Invariant Lines

Let $T \in \mathcal{L}(V)$ and $v \in V$ with $v \neq 0$. Then $\operatorname{span}(v)$ is T-invariant if and only if $Tv = \lambda v$ for some $\lambda \in F$.

Proof. By Lemma 9.1.1, $\operatorname{span}(v)$ is invariant if and only if $Tv \in \operatorname{span}(v)$, and $\operatorname{span}(v) = \{\lambda v : \lambda \in F\}$ by Definition 1.3.2. ∎

So an invariant line is a line on which T acts by stretching with a single factor λ. Does every operator have one? Not over $\mathbb{R}$.

Example. A Rotation Has No Invariant Lines

Let $R\colon \mathbb{R}^2 \to \mathbb{R}^2$, $R(x,y) = (-y,x)$, be the rotation by a right angle. Find all R-invariant subspaces of $\mathbb{R}^2$.

Solution. A subspace U of $\mathbb{R}^2$ has $\dim U \in \{0,1,2\}$ by Theorem 1.6.7. If $\dim U = 0$, then $U = \{\mathbf{0}\}$; if $\dim U = 2$, then $U = \mathbb{R}^2$ by Theorem 1.6.8. Both are invariant.

Suppose $\dim U = 1$. A basis of U is a single vector $\mathbf{v} = (a,b) \neq \mathbf{0}$, and $U = \operatorname{span}(\mathbf{v})$. If U were invariant, then by Proposition 9.1.2 there would be $\lambda \in \mathbb{R}$ with $R\mathbf{v} = \lambda\mathbf{v}$, that is,

$$-b = \lambda a, \qquad a = \lambda b.$$

Substituting the first equation into the second gives $a = \lambda b = \lambda(-\lambda a) = -\lambda^2 a$, so $(1+\lambda^2)a = 0$. Since λ is real, $1 + \lambda^2 \ge 1 > 0$, so $a = 0$. Then $b = -\lambda a = 0$ by the first equation, and $\mathbf{v} = \mathbf{0}$, a contradiction.

Hence the only R-invariant subspaces of $\mathbb{R}^2$ are $\{\mathbf{0}\}$ and $\mathbb{R}^2$. Geometrically, a quarter-turn moves every line through the origin onto the perpendicular line.

The same computation over $\mathbb{C}$ does not end in a contradiction: $1+\lambda^2=0$ has the solutions $\lambda=\pm i$, and we will see in the next section that $\mathbb{C}^2$ does contain R-invariant lines. Whether invariant lines exist depends on the field.

Why this definition. Two variations suggest themselves, and both are worse.

- *Demand* $T(U)=U$. Then $\ker T$ would fail whenever T is not injective, since then $T(\ker T)=\{0\}\neq\ker T$. But the kernel is exactly the kind of subspace that produces a zero block, and Theorem 8.1.7 needs only the inclusion.
- *Demand* $Tu=u$ *for every* $u\in U$. That says T acts as the identity on U, a far stronger condition. The line $\operatorname{span}((1,1))$ above is invariant under $\mathbf{A}$, but $\mathbf{A}$ triples its vectors.

The name records that the subspace, as a whole, is unchanged in the sense that it is carried into itself; its individual vectors may move.

Warning. **Invariant does not mean fixed.** The vectors of a T-invariant subspace may move, as long as they stay inside it. For the swap $T(x,y)=(y,x)$ on $\mathbb{R}^2$, the whole plane is invariant, yet $T(1,0)=(0,1)\neq(1,0)$. Conversely, containing a fixed vector does not make a subspace invariant. For $T(x,y,z)=(x,z,y+z)$ on $\mathbb{R}^3$, the plane $z=0$ contains $\mathbf{e}_1$, and $T\mathbf{e}_1=\mathbf{e}_1$; but it also contains $\mathbf{e}_2$, and $T\mathbf{e}_2=\mathbf{e}_3$ leaves the plane.

Quick check. Let $T(x,y)=(x+y,y)$ on $\mathbb{R}^2$. Is the x-axis T-invariant? Is the y-axis?

Solution. By Lemma 9.1.1 it suffices to test one spanning vector. $T\mathbf{e}_1=(1,0)=\mathbf{e}_1$, which lies on the x-axis, so the x-axis is invariant. $T\mathbf{e}_2=(1,1)$, which is not a multiple of $\mathbf{e}_2$, so the y-axis is **not** invariant.

9.1.2 Restriction and the induced operator

An invariant subspace U lets us split T into two smaller operators. One lives on U: since T sends U into U, we may simply forget the rest of V. The other lives on the quotient V/U, which records what U cannot see. Chapter 8 met both in the proof of Theorem 8.1.7, and Chapter 2 constructed the second in Exercise (C2: The operator induced on a quotient). We now give them names.

Definition 9.1.2. Restriction and Induced Operator

Let $T\in\mathcal{L}(V)$ and let U be a T-invariant subspace of V.

(a) The **restriction** of T to U is the operator $T|_U\colon U\to U$, $u\mapsto Tu$.

(b) The **induced operator** (or quotient operator) on V/U is $\bar{T}\colon V/U\to V/U$, $v+U\mapsto Tv+U$.

Well-definedness. For (a), the rule makes sense as a map **into** U only because U is invariant; it is linear because T is, and U carries the operations of V. For (b), the rule is phrased in terms of a representative v of the coset, so we must check that it does not depend on that choice. Suppose $v+U = v'+U$. Then $v - v' \in U$ (Lemma 4.4.3), so $Tv - Tv' = T(v-v') \in U$ by invariance, and hence $Tv + U = Tv' + U$. Linearity follows from the coset operations (Theorem 4.4.4): for instance

$$\begin{aligned}\bar{T}\big((v+U)+(w+U)\big) &= \bar{T}(v+w+U) = Tv+Tw+U \\ &= \bar{T}(v+U)+\bar{T}(w+U),\end{aligned}$$

and scalars work the same way. Invariance was used exactly once in each part, which is where the hypothesis earns its place.

Examples.

- **Differentiation.** Let D act on $V = \mathbb{R}[x]_{\le 2}$ and $U = \mathbb{R}[x]_{\le 1}$, which is D-invariant. The restriction $D|_U$ is differentiation on $\mathbb{R}[x]_{\le 1}$. The quotient V/U has dimension 1 (Theorem 4.4.5), with basis $x^2 + U$, and $\bar{D}(x^2+U) = 2x + U = 0 + U$, because $2x \in U$. So $\bar{D} = 0$.
- **The cyclic shift.** For $T(x,y,z) = (y,z,x)$ and the plane U where $x+y+z=0$, Example (A Plane Mapped into Itself) showed that $\bar{T}$ is the identity of the line $\mathbb{R}^3/U$: the shift permutes coordinates and so never changes their sum.
- **Degenerate cases.** For $U = V$, the restriction is T itself and V/V is the zero space. For $U = \{0\}$, the restriction is the operator on the zero space, and $\bar{T}$ is T in disguise, since $v + \{0\}$ is just v.

With these names, the matrix theorem of Chapter 8 reads as follows. We add the consequence for determinants and characteristic polynomials, which is what the rest of this chapter uses.

Theorem 9.1.3. Invariant Subspaces and Block Triangular Matrices

Let V be finite-dimensional, $T \in \mathcal{L}(V)$, and U a subspace with $\{0\} \neq U \neq V$. Let $\mathcal{B}_U = (u_1, \dots, u_k)$ be a basis of U, extended to a basis $\mathcal{B} = (u_1, \dots, u_k, w_1, \dots, w_l)$ of V, and put $\bar{\mathcal{B}} = (w_1 + U, \dots, w_l + U)$.

(a) U is T-invariant if and only if $[T]_{\mathcal{B}}^{\mathcal{B}}$ has zero lower-left $l \times k$ block.

(b) In that case

$$[T]_{\mathcal{B}}^{\mathcal{B}} = \begin{pmatrix} [T|_U]_{\mathcal{B}_U}^{\mathcal{B}_U} & \mathbf{B} \\ \mathbf{O} & [\bar{T}]_{\bar{\mathcal{B}}}^{\bar{\mathcal{B}}} \end{pmatrix}$$

for some $\mathbf{B} \in M_{k\times l}(F)$, and

$$\det T = \det(T|_U)\,\det \bar{T}, \qquad p_T = p_{T|_U}\, p_{\bar{T}}.$$

Proof. Part (a), and the block form in (b), are Theorem 8.1.7 in the language of Definition 9.1.1 and Definition 9.1.2; $\bar{\mathcal{B}}$ is a basis of V/U by Theorem 4.4.5.

Write $\mathbf{A} = [T|_U]_{\mathcal{B}_U}^{\mathcal{B}_U} \in M_k(F)$ and $\mathbf{D} = [\bar{T}]_{\bar{\mathcal{B}}}^{\bar{\mathcal{B}}} \in M_l(F)$. By Definition 7.5.1 and Theorem 7.8.4, $\det T = \det [T]_{\mathcal{B}}^{\mathcal{B}} = \det \mathbf{A} \det \mathbf{D} = \det(T|_U) \det \bar{T}$. For the characteristic polynomials,

$$x\mathbf{I}_{k+l} - [T]_{\mathcal{B}}^{\mathcal{B}} = \begin{pmatrix} x\mathbf{I}_k - \mathbf{A} & -\mathbf{B} \\ \mathbf{O} & x\mathbf{I}_l - \mathbf{D} \end{pmatrix}$$

is block upper triangular with square diagonal blocks, as a matrix over $F[x]$. The block tri-

angular determinant holds over the commutative ring $F[x]$ (the remark after Theorem 7.8.4), so $\det(x\mathbf{I} - [T]_{\mathcal{B}}^{\mathcal{B}}) = \det(x\mathbf{I}_k - \mathbf{A})\det(x\mathbf{I}_l - \mathbf{D})$. By Definition 7.10.2, this says $p_T = p_{T|_U}\, p_{\bar{T}}$.

■

The theorem is the reason invariant subspaces matter. It splits the study of T into the study of two operators on smaller spaces, and every quantity that sees only the diagonal blocks splits with it. For differentiation on $\mathbb{R}[x]_{\le 2}$, with $U = \mathbb{R}[x]_{\le 1}$ and $\mathcal{B} = (1, x, x^2)$,

$$[D]_{\mathcal{B}}^{\mathcal{B}} = \left(\begin{array}{cc|c} 0 & 1 & 0 \\ 0 & 0 & 2 \\ \hline 0 & 0 & 0 \end{array}\right),$$

with upper-left block $[D|_U]_{\mathcal{B}_U}^{\mathcal{B}_U} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$, lower-right block $[\bar{D}]_{\bar{\mathcal{B}}}^{\bar{\mathcal{B}}} = \mathbf{O}_1$, and $p_D = x^3 = x^2 \cdot x$. Section 7 of this chapter will run this splitting repeatedly, by induction on the dimension.

9.1.3 Sums, intersections, and direct sums

Invariant subspaces can be combined. Sums and intersections of invariant subspaces are invariant, in contrast to many other properties, and the proof is one line each.

Proposition 9.1.4. Sums and Intersections of Invariant Subspaces

Let $T \in \mathcal{L}(V)$ and let $U_1, \dots, U_k$ be T-invariant subspaces of V. Then $U_1 + \cdots + U_k$ and $U_1 \cap \cdots \cap U_k$ are T-invariant.

Proof. Both are subspaces, by Theorem 1.7.1 and Theorem 1.2.2. Let $u = u_1 + \cdots + u_k$ with $u_i \in U_i$. Then $Tu = Tu_1 + \cdots + Tu_k$ with $Tu_i \in U_i$ by invariance, so $Tu \in U_1 + \cdots + U_k$. If $u \in U_1 \cap \cdots \cap U_k$, then $Tu \in U_i$ for each i, by invariance of each U_i, so $Tu \in U_1 \cap \cdots \cap U_k$.

■

When V is a **direct** sum of invariant subspaces, all the off-diagonal blocks vanish, not just one corner. Chapter 8 proved this for two summands (Theorem 8.1.6). The diagonalization theorem of Section 4 needs any number of summands.

Theorem 9.1.5. Invariant Direct Sums and Block Diagonal Matrices

Let V be finite-dimensional with $V = U_1 \oplus \cdots \oplus U_k$, where each $U_i \neq \{0\}$ has a basis $\mathcal{B}_i$ of length n_i. Let $\mathcal{B}$ be the list $\mathcal{B}_1, \dots, \mathcal{B}_k$ written one after another, and $T \in \mathcal{L}(V)$. Then $\mathcal{B}$ is a basis of V, and the following are equivalent:

(a) each U_i is T-invariant;

(b) $[T]_{\mathcal{B}}^{\mathcal{B}}$ is block diagonal with respect to the partition $n_1, \dots, n_k$.

In that case

$$[T]_{\mathcal{B}}^{\mathcal{B}} = [T|_{U_1}]_{\mathcal{B}_1}^{\mathcal{B}_1} \oplus \cdots \oplus [T|_{U_k}]_{\mathcal{B}_k}^{\mathcal{B}_k}, \qquad p_T = p_{T|_{U_1}} \cdots p_{T|_{U_k}}.$$

Idea. Read the matrix one block of columns at a time. The columns of block j are the coordinates of Tb for the vectors b of $\mathcal{B}_j$. Such a Tb lies in U_j exactly when all its coordinates outside block j vanish, because a vector of U_j is already a combination of $\mathcal{B}_j$, and coordinates are unique.

Proof. By Theorem 1.7.5 ((a) $\Rightarrow$ (d)), $\mathcal{B}$ is a basis of V. Fix j and a vector b of $\mathcal{B}_j$; the column of $[T]_{\mathcal{B}}^{\mathcal{B}}$ belonging to b is $[Tb]_{\mathcal{B}}$ (Definition 4.1.1). We claim: $Tb \in U_j$ if and only if the entries of this column outside the rows of block j are all 0. If $Tb \in U_j = \operatorname{span}(\mathcal{B}_j)$, then Tb is a combination of $\mathcal{B}_j$ alone, which is a representation in the basis $\mathcal{B}$ with zero coefficients outside block j; by Theorem 1.5.1, it is **the** coordinate vector. Conversely, if those entries vanish, then Tb is a combination of $\mathcal{B}_j$, so it lies in U_j.

(a) $\Rightarrow$ (b). If U_j is invariant, then $Tb \in U_j$ for each b in $\mathcal{B}_j$, so by the claim every block (i, j) with $i \neq j$ is zero. As j was arbitrary, $[T]_{\mathcal{B}}^{\mathcal{B}}$ is block diagonal. Moreover the entries of column b in the rows of block j are the coordinates of $Tb = T|_{U_j} b$ in $\mathcal{B}_j$, so the (j, j) block is $[T|_{U_j}]_{\mathcal{B}_j}^{\mathcal{B}_j}$.

(b) $\Rightarrow$ (a). If the blocks (i, j) with $i \neq j$ vanish, the claim gives $Tb \in U_j$ for every b in $\mathcal{B}_j$, and U_j is invariant by Lemma 9.1.1.

Finally, $x\mathbf{I} - [T]_{\mathcal{B}}^{\mathcal{B}}$ is block diagonal with diagonal blocks $x\mathbf{I}_{n_j} - [T|_{U_j}]_{\mathcal{B}_j}^{\mathcal{B}_j}$. Its determinant over $F[x]$ is the product of the determinants of these blocks (the remark after Theorem 7.8.4, applied by induction on the number of blocks), which is $p_T = \prod_j p_{T|_{U_j}}$ by Definition 7.10.2.

■

So an operator that preserves each piece of a direct sum decomposition is, in a suitable basis, a list of independent smaller operators. The best possible case is a decomposition into invariant **lines**: then every block is 1×1, and $[T]_{\mathcal{B}}^{\mathcal{B}}$ is diagonal. By Proposition 9.1.2, that happens exactly when V has a basis of vectors v with $Tv = \lambda v$. This observation is the starting point of the next section.

A tempting shortcut is to assume that once one invariant subspace is found, a complement of it can be chosen invariant too, and the matrix becomes block diagonal. This fails.

Warning. **The complement of an invariant subspace need not be invariant.** Let $\mathbf{N} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$ act on F^2. The line $U = \operatorname{span}(\mathbf{e}_1)$ is invariant, since $\mathbf{N}\mathbf{e}_1 = \mathbf{0}$. Every complement of U is a line $W = \operatorname{span}((a, 1))$ for some $a \in F$ (a line not equal to U contains a vector with second entry 1). But $\mathbf{N}(a, 1) = (1, 0)$, which is not a multiple of $(a, 1)$. So **no** complement of U is $\mathbf{N}$-invariant, and no basis $\mathcal{B}$ makes $[T_{\mathbf{N}}]_{\mathcal{B}}^{\mathcal{B}}$ block diagonal with two 1×1 blocks.

Quick check. Let U_1, U_2 be T-invariant. Must the union $U_1 \cup U_2$ be T-invariant? Must it even be a subspace?

Solution. The question has a trap: "invariant subspace" requires a subspace, and a union of two subspaces is usually not one (Proposition 1.2.3). For $T = \mathrm{id}_{\mathbb{R}^2}$, the two axes are invariant, but their union is not a subspace, so the word does not apply. The subspace that plays the role of the union is the sum $U_1 + U_2$, which is invariant by Proposition 9.1.4.

9.1.4 Exercises

9.1.4.1 A. Check your understanding

Exercise. A1

(a) Define what it means for a subspace U of V to be invariant under $T \in \mathcal{L}(V)$.

(b) True or false: for every $T \in \mathcal{L}(V)$, the subspaces $\ker T$ and $\operatorname{im} T$ are T-invariant. Justify your answer.

(c) True or false: if U is T-invariant, then $Tu = u$ for every $u \in U$. Justify your answer.

(d) True or false: if U is invariant under S and under T, then U is invariant under $S + T$ and under ST. Justify your answer.

(e) For $v \neq 0$, what condition on Tv makes $\operatorname{span}(v)$ invariant?

(f) In a basis whose first k vectors form a basis of a T-invariant subspace U, which block of $[T]_{\mathcal{B}}^{\mathcal{B}}$ is zero, and what are the two diagonal blocks?

Solution.

(a) U is T-invariant if $Tu \in U$ for every $u \in U$ (Definition 9.1.1).

(b) True. If $u \in \ker T$, then $Tu = 0 \in \ker T$. If $w \in \operatorname{im} T$, then $Tw \in \operatorname{im} T$ by definition of the image.

(c) False. For $T = 2\,\mathrm{id}_{\mathbb{R}^2}$, the subspace $\mathbb{R}^2$ is invariant, but $T\mathbf{e}_1 = 2\mathbf{e}_1 \neq \mathbf{e}_1$.

(d) True. For $u \in U$, both Su and Tu lie in U, so $(S+T)u = Su + Tu \in U$ since U is a subspace. Also $Tu \in U$, so $(ST)u = S(Tu) \in U$ by invariance under S.

(e) $Tv = \lambda v$ for some scalar λ (Proposition 9.1.2).

(f) The lower-left block is zero. The upper-left block is $[T|_U]_{\mathcal{B}_U}^{\mathcal{B}_U}$ and the lower-right block is the matrix of the induced operator $\bar{T}$ on V/U (Theorem 9.1.3).

9.1.4.2 B. Practice

Exercise. B1: All invariant subspaces of a diagonal matrix

Let $\mathbf{A} = \operatorname{diag}(1, 2)$ act on $\mathbb{R}^2$. Find all $\mathbf{A}$-invariant subspaces of $\mathbb{R}^2$.

Solution. As in Example (A Rotation Has No Invariant Lines), a subspace of $\mathbb{R}^2$ is $\{\mathbf{0}\}$, $\mathbb{R}^2$, or a line $\operatorname{span}(\mathbf{v})$ with $\mathbf{v} = (a, b) \neq \mathbf{0}$. The first two are invariant. By Proposition 9.1.2, the line is invariant if and only if $\mathbf{Av} = \lambda\mathbf{v}$ for some $\lambda \in \mathbb{R}$, that is,

$$a = \lambda a, \qquad 2b = \lambda b.$$

If $a \neq 0$, the first equation gives $\lambda = 1$, and then $2b = b$ gives $b = 0$; so $\mathbf{v}$ is a multiple of $\mathbf{e}_1$. If $a = 0$, then $b \neq 0$, the second equation gives $\lambda = 2$, and $\mathbf{v}$ is a multiple of $\mathbf{e}_2$. Conversely $\mathbf{A}\mathbf{e}_1 = \mathbf{e}_1$ and $\mathbf{A}\mathbf{e}_2 = 2\mathbf{e}_2$. Hence the invariant subspaces are exactly

$$\{\mathbf{0}\}, \qquad \operatorname{span}(\mathbf{e}_1), \qquad \operatorname{span}(\mathbf{e}_2), \qquad \mathbb{R}^2.$$

Compare with the identity $\mathbf{I}_2$, for which every line is invariant: changing one diagonal entry from 1 to 2 destroys all lines except the two axes.

Exercise. B2: Kernels of $T - c$ id

Let $T \in \mathcal{L}(V)$ and $c \in F$.

(a) Prove directly from the definition that $\ker(T - c\,\mathrm{id}_V)$ is T-invariant.

(b) Hence show that **every** subspace of $\ker(T - c\,\mathrm{id}_V)$ is T-invariant.

Solution.

(a) Let $K = \ker(T - c\,\mathrm{id}_V)$, a subspace by Theorem 2.2.1. For $u \in K$ we have $Tu - cu = 0$, that is, $Tu = cu$. Since K is a subspace, $cu \in K$. Hence $Tu \in K$, and K is T-invariant.

(b) Let W be a subspace of K and $w \in W$. Since $w \in K$, part (a) gives $Tw = cw$, and $cw \in W$ because W is a subspace. Hence W is T-invariant. (On K, T acts as c id, and for a scalar operator every subspace is invariant.)

Exercise. B3: Invariance under differentiation

Let D be differentiation on $V = \mathbb{R}[x]_{\leq 3}$. Determine which of the following subspaces are D-invariant. Justify your answer.

(a) $U_1 = \operatorname{span}(1, x)$.

(b) $U_2 = \operatorname{span}(x, x^2)$.

(c) $U_3 = \{p \in V : p(0) = 0\}$.

(d) $U_4 = \operatorname{span}(x^2 + 2x,\ 2x + 2,\ 2)$.

(e) $U_5 = \operatorname{span}(1, x^2)$, the even polynomials in V.

Solution. We use Lemma 9.1.1 throughout.

(a) Invariant: $D(1) = 0 \in U_1$ and $D(x) = 1 \in U_1$.

(b) Not invariant: $D(x) = 1$, and $1 \notin \operatorname{span}(x, x^2)$ since every element of U_2 has constant term 0.

(c) Not invariant: $x \in U_3$, but $D(x) = 1$ and 1 evaluated at 0 is $1 \neq 0$.

(d) Invariant: $D(x^2 + 2x) = 2x + 2$, $D(2x + 2) = 2$, and $D(2) = 0$, all in U_4. (In fact $U_4 = \mathbb{R}[x]_{\leq 2}$, since the three spanning polynomials have degrees $2, 1, 0$ and so are independent by Theorem 1.4.4.)

(e) Not invariant: $D(x^2) = 2x$, which is odd and nonzero, so it is not in $\operatorname{span}(1, x^2)$.

9.1.4.3 C. Going deeper

Exercise. C1: A chain of invariant subspaces

Let F be any field and

$$\mathbf{N} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} \in M_3(F),$$

acting on F^3, so $\mathbf{N}\mathbf{e}_1 = \mathbf{0}$, $\mathbf{N}\mathbf{e}_2 = \mathbf{e}_1$, $\mathbf{N}\mathbf{e}_3 = \mathbf{e}_2$. Put $W_0 = \{\mathbf{0}\}$, $W_1 = \operatorname{span}(\mathbf{e}_1)$, $W_2 = \operatorname{span}(\mathbf{e}_1, \mathbf{e}_2)$, $W_3 = F^3$.

(a) Show that each W_j is $\mathbf{N}$-invariant.

(b) Let U be an $\mathbf{N}$-invariant subspace, and let j be the largest index such that some vector of U has a nonzero j-th entry (put $j = 0$ if $U = \{\mathbf{0}\}$). Prove that $U = W_j$.

(c) Deduce that the $\mathbf{N}$-invariant subspaces are exactly $W_0 \subset W_1 \subset W_2 \subset W_3$, and that any two of them are comparable under inclusion.

Hint: for (b), apply $\mathbf{N}$ and $\mathbf{N}^2$ to a vector with nonzero j-th entry.

Solution.

(a) W_0 and W_3 are invariant for every operator. By Lemma 9.1.1, W_1 is invariant since $\mathbf{N}\mathbf{e}_1 = \mathbf{0}$, and W_2 is invariant since $\mathbf{N}\mathbf{e}_1 = \mathbf{0}$ and $\mathbf{N}\mathbf{e}_2 = \mathbf{e}_1$ lie in W_2.

(b) If $U = \{\mathbf{0}\}$, then $j = 0$ and $U = W_0$. Otherwise $j \geq 1$. By the choice of j, every vector of U has entries 0 beyond position j, so $U \subseteq W_j$. Pick $\mathbf{v} = (a, b, c) \in U$ with nonzero j-th entry. We have $\mathbf{N}\mathbf{v} = (b, c, 0)$ and $\mathbf{N}^2\mathbf{v} = (c, 0, 0)$, and all of these lie in U by invariance.
Case $j = 3$. Here $c \neq 0$. Then $\mathbf{e}_1 = c^{-1}\mathbf{N}^2\mathbf{v} \in U$, $\mathbf{e}_2 = c^{-1}(\mathbf{N}\mathbf{v} - b\mathbf{e}_1) \in U$, and $\mathbf{e}_3 = c^{-1}(\mathbf{v} - a\mathbf{e}_1 - b\mathbf{e}_2) \in U$. So $W_3 \subseteq U$.
Case $j = 2$. Here $c = 0$ and $b \neq 0$. Then $\mathbf{e}_1 = b^{-1}\mathbf{N}\mathbf{v} \in U$ and $\mathbf{e}_2 = b^{-1}(\mathbf{v} - a\mathbf{e}_1) \in U$, so $W_2 \subseteq U$.
Case $j = 1$. Here $b = c = 0$ and $a \neq 0$, so $\mathbf{e}_1 = a^{-1}\mathbf{v} \in U$ and $W_1 \subseteq U$.
In each case $W_j \subseteq U \subseteq W_j$, so $U = W_j$.

(c) By (a) each W_j is invariant, and by (b) every invariant subspace is one of them. Since $W_0 \subset W_1 \subset W_2 \subset W_3$, any two are comparable. In particular $\mathbf{N}$ has exactly one invariant line, W_1, and exactly one invariant plane, W_2. Contrast Exercise (B1: All invariant subspaces of a diagonal matrix), where $\operatorname{diag}(1, 2)$ has two invariant lines (in the sense of Proposition 9.1.2), neither containing the other.

Exercise. C2: Invariance under the inverse

(a) Let V be finite-dimensional, $T \in \mathcal{L}(V)$ invertible, and U a T-invariant subspace. Prove that $T(U) = U$, and deduce that U is T^{-1}-invariant.

(b) Show that (a) fails without finite dimension. On $V = F^{\mathbb{Z}}$, the space of all functions $\mathbb{Z} \to F$ written as two-sided sequences $(s_n)_{n \in \mathbb{Z}}$, let $(Ss)_n = s_{n-1}$. Show that S is invertible, that $U = \{s : s_n = 0 \text{ for all } n < 0\}$ is S-invariant, and that U is **not** S^{-1}-invariant.

Hint: for (a), consider the restriction $T|_U$ and its kernel.

Solution.

(a) The restriction $T|_U \in \mathcal{L}(U)$ exists by invariance. Its kernel is $\ker T \cap U = \{0\}$, since T is injective, so $T|_U$ is injective (Theorem 2.2.3). The space U is finite-dimensional (Theorem 1.6.7), so by Theorem 2.5.7, $T|_U$ is surjective onto U. That is, $T(U) = U$. Now let $u \in U$. By surjectivity of $T|_U$, there is $u' \in U$ with $Tu' = u$. Applying T^{-1} gives $T^{-1}u = u' \in U$. Hence U is T^{-1}-invariant.

(b) $V = F^{\mathbb{Z}}$ is a vector space with pointwise operations (Example (The Standard Vector Spaces)), and S is linear because it only re-indexes entries. Define $(S's)_n = s_{n+1}$. Then $(S'Ss)_n = (Ss)_{n+1} = s_n$ and $(SS's)_n = (S's)_{n-1} = s_n$, so S is invertible with $S^{-1} = S'$.

U is a subspace (the conditions $s_n = 0$ are preserved by sums and scalar multiples). If $s \in U$ and $n < 0$, then $n - 1 < 0$, so $(Ss)_n = s_{n-1} = 0$; thus $Ss \in U$, and U is S-invariant.

Let $\delta \in U$ be the sequence with $\delta_0 = 1$ and $\delta_n = 0$ for $n \neq 0$. Then $(S^{-1}\delta)_{-1} = \delta_0 = 1 \neq 0$, so $S^{-1}\delta \notin U$. Hence U is not S^{-1}-invariant. In (a), the step that breaks is "$T|_U$ injective, hence surjective", which needs finite dimension: here $S|_U$ is injective but its image misses δ.

9.2 Eigenvalues and Eigenvectors

The previous section ended with an observation: an operator has a diagonal matrix exactly when the space splits into invariant lines, and a line $\operatorname{span}(v)$ is invariant exactly when Tv is a multiple of v. This section turns that observation into the central notion of the chapter. We name the vectors that an operator only stretches, and the stretching factors. We find the factors as roots of the characteristic polynomial of Chapter 7, and we prove the two facts that everything later rests on: eigenvectors for different factors are independent, and over $\mathbb{C}$ every operator has at least one eigenvector.

9.2.1 When is a matrix diagonal?

Chapter 4 asked how simple $[T]_{\mathcal{B}}^{\mathcal{B}}$ can be made by choosing $\mathcal{B}$. The simplest square matrices are the diagonal ones, so start with the wish and unwind it.

Let V be finite-dimensional with basis $\mathcal{B} = (v_1, \dots, v_n)$, and $T \in \mathcal{L}(V)$. By Definition 4.1.1, column j of $[T]_{\mathcal{B}}^{\mathcal{B}}$ holds the coordinates of Tv_j. The matrix is $\operatorname{diag}(\lambda_1, \dots, \lambda_n)$ exactly when column j is $\lambda_j \mathbf{e}_j$ for each j, that is,

$$Tv_j = \lambda_j v_j \qquad (j = 1, \dots, n).$$

So the wish forces an equation: every basis vector must be sent to a multiple of itself. We met this in Chapter 2, where the reflection $R(x, y) = (y, x)$ became $\operatorname{diag}(1, -1)$ in the basis $((1, 1), (1, -1))$ (Example (The reflection in a mirror-adapted basis)), precisely because $R(1, 1) = (1, 1)$ and $R(1, -1) = -(1, -1)$.

Where should we look for vectors with $Tv = \lambda v$? Rewrite λv as $\lambda \operatorname{id}_V(v)$. Then the equation says $(T - \lambda \operatorname{id}_V)v = 0$: the vector lies in a **kernel**. Kernels are subspaces we know how to compute, so this rewriting is a plan, not just a restatement.

An eigenvector of T is a nonzero vector that T only stretches, and the eigenvalue is the stretching factor.

Definition 9.2.1. Eigenvalue and Eigenvector

Let V be a vector space over F and $T \in \mathcal{L}(V)$. A scalar $\lambda \in F$ is an **eigenvalue** of T if there **exists** a **nonzero** vector $v \in V$ with

$$Tv = \lambda v.$$

Every **nonzero** $v \in V$ with $Tv = \lambda v$ is called **an eigenvector** of T for the eigenvalue λ.

For a matrix $\mathbf{A} \in M_n(F)$, the eigenvalues and eigenvectors of $\mathbf{A}$ are those of $T_{\mathbf{A}} \colon F^n \to F^n$: a scalar $\lambda \in F$ and a nonzero $\mathbf{v} \in F^n$ with $\mathbf{Av} = \lambda \mathbf{v}$.

In words: an eigenvalue is a scalar **in the field** F over which V is a vector space. It qualifies if **some** nonzero vector is sent to λ times itself. The vector must be nonzero; the scalar may be anything, including 0. An eigenvector is a witness for its eigenvalue, and there are always many witnesses: if v is one, so is cv for every $c \neq 0$, since $T(cv) = cTv = \lambda(cv)$.

Well-definedness. An eigenvector determines its eigenvalue. If $Tv = \lambda v$ and $Tv = \mu v$ with $v \neq 0$, then $(\lambda - \mu)v = 0$, and Theorem 1.1.9 gives $\lambda = \mu$. So "the eigenvalue of v" makes sense, while "the eigenvector of λ" does not.

The eigenvectors for one eigenvalue, together with 0, are exactly the kernel found above. That set is a subspace, and it deserves its own name.

Definition 9.2.2. Eigenspace

Let $T \in \mathcal{L}(V)$ and $\lambda \in F$. The **eigenspace** of T for λ is

$$E_\lambda(T) \coloneqq \ker(T - \lambda \operatorname{id}_V) = \{v \in V : Tv = \lambda v\}.$$

For $\mathbf{A} \in M_n(F)$ we write $E_\lambda(\mathbf{A}) = E_\lambda(T_{\mathbf{A}}) = \operatorname{nul}(\mathbf{A} - \lambda \mathbf{I}_n)$.

$E_\lambda(T)$ is a subspace by Theorem 2.2.1, and it is T-invariant by Theorem 6.7.4 (b) with $p = x - \lambda$. It is defined for **every** $\lambda \in F$, and λ is an eigenvalue exactly when $E_\lambda(T) \neq \{0\}$; in that case the eigenvectors for λ are the nonzero vectors of $E_\lambda(T)$. The eigenspace is the home of all eigenvectors for λ at once, and it is one of the invariant subspaces of the previous section: on it, T acts as λ id.

Examples.

- **Diagonal matrices.** Let $\mathbf{A} = \operatorname{diag}(d_1, \ldots, d_n)$. Then $\mathbf{A}\mathbf{e}_j = d_j \mathbf{e}_j$ and $\mathbf{e}_j \neq \mathbf{0}$, so each d_j is an eigenvalue with eigenvector $\mathbf{e}_j$. The standard basis is a basis of eigenvectors, which is the observation we started from, read backwards.
- **Degenerate cases.** On $V \neq \{0\}$, the identity has the eigenvalue 1 with $E_1(\operatorname{id}_V) = V$: **every** nonzero vector is an eigenvector. The zero operator has the eigenvalue 0 with $E_0 = V$. On $V = \{0\}$, no operator has any eigenvalue, because there is no nonzero vector to be a witness. That last case matters: it is why the existence theorem later in this section must assume $V \neq \{0\}$.
- **Projections.** Let $P \in \mathcal{L}(V)$ be a projection, $P^2 = P$ (Definition 4.5.1). If $Pv = \lambda v$ with $v \neq 0$, then $\lambda v = Pv = P^2 v = P(\lambda v) = \lambda^2 v$, so $(\lambda^2 - \lambda)v = 0$ and $\lambda(\lambda - 1) = 0$ by

Theorem 1.1.9. So the only possible eigenvalues are 0 and 1. By definition $E_0(P) = \ker P$. Also $E_1(P) = \operatorname{im} P$: a vector with $Pv = v$ is in the image, and conversely P is the identity on $\operatorname{im} P$ by Theorem 4.5.1 (a). So a vector is fixed by P exactly when it is in the image, and killed exactly when it is in the kernel.

- **Differentiation of polynomials.** Let D act on $V = \mathbb{R}[x]_{\le n}$ with $n \ge 1$. Suppose $Dp = \lambda p$ with $p \neq 0$. If $\lambda \neq 0$, then $\deg(\lambda p) = \deg p$, while Dp is either 0 or of degree $\deg p - 1$; neither equals $\deg p$ for $p \neq 0$. So $\lambda = 0$, and $E_0(D) = \ker D$ is the space of constant polynomials. Differentiation on $\mathbb{R}[x]_{\le n}$ has exactly one eigenvalue, and its eigenvectors span only a line.

Nonexample by minimal change. For $\mathbf{A} = \begin{pmatrix} 2 & 1 \\ 0 & 3 \end{pmatrix}$, the vector $\mathbf{e}_1$ is an eigenvector: $\mathbf{A}\mathbf{e}_1 = 2\mathbf{e}_1$. Move to the neighboring basis vector $\mathbf{e}_2$. It is still nonzero, and $\mathbf{A}\mathbf{e}_2 = (1, 3)$ is still a vector of $\mathbb{R}^2$. What fails is the equation $\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$: no λ makes $(1, 3) = (0, \lambda)$. So $\mathbf{e}_2$ is not an eigenvector, and the line $\operatorname{span}(\mathbf{e}_2)$ is not invariant, as we saw in the previous section.

Why this definition. The clause "**nonzero**" is what gives the notion content. The zero vector satisfies $T0 = \lambda 0$ for **every** $\lambda \in F$, so if we allowed $v = 0$, every scalar would be an eigenvalue of every operator. On the other hand we do **not** exclude $\lambda = 0$: the equation $Tv = 0v$ says $v \in \ker T$, and a nonzero kernel vector is exactly the kind of information we want to record. Requiring $\lambda \in F$, rather than in some larger field, makes the notion depend on the field, and this is deliberate: a real matrix may have no real eigenvalues, and then no real basis diagonalizes it. The prefix *eigen* is German for "own": an eigenvector is a direction that belongs to T, which T maps into itself.

Warning. **The eigenvalue may be 0; the eigenvector may not.** For $\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix}$, the vector $(2, -1)$ satisfies $\mathbf{A}(2, -1) = (0, 0) = 0 \cdot (2, -1)$, so 0 is an eigenvalue. By contrast, $\mathbf{A0} = 7 \cdot \mathbf{0}$ does **not** make 7 an eigenvalue. And say "**an** eigenvector for 0", never "the": $(2, -1)$, $(-4, 2)$, and $(200, -100)$ are all eigenvectors for 0.

Quick check. Let $T(x, y, z) = (x, z, y)$ on $\mathbb{R}^3$. Which of $\mathbf{e}_1$, $(0, 1, 1)$, $(0, 1, -1)$, $\mathbf{e}_2$ are eigenvectors of T, and for which eigenvalues?

Solution. $T\mathbf{e}_1 = \mathbf{e}_1$, so $\mathbf{e}_1$ is an eigenvector for 1. $T(0, 1, 1) = (0, 1, 1)$, an eigenvector for 1 as well. $T(0, 1, -1) = (0, -1, 1) = -(0, 1, -1)$, an eigenvector for -1. $T\mathbf{e}_2 = \mathbf{e}_3$, which is not a multiple of $\mathbf{e}_2$, so $\mathbf{e}_2$ is **not** an eigenvector. The three eigenvectors form a basis $\mathcal{B}$ of $\mathbb{R}^3$, and $[T]_{\mathcal{B}}^{\mathcal{B}} = \operatorname{diag}(1, 1, -1)$.

9.2.2 Finding eigenvalues

The definition asks for a nonzero vector in $\ker(T - \lambda \operatorname{id}_V)$, a question about one λ at a time. In finite dimension, Chapters 2 and 7 turn it into a test that handles all λ at once. The last item is the one we compute with.

Theorem 9.2.1. Characterizations of Eigenvalues

Let V be a finite-dimensional vector space over F with $\dim V = n \geq 1$, let $T \in \mathcal{L}(V)$ and $\lambda \in F$. The following are equivalent:

(a) λ is an eigenvalue of T;
(b) $E_\lambda(T) \neq \{0\}$, that is, $T - \lambda\,\mathrm{id}_V$ is **not injective**;
(c) $T - \lambda\,\mathrm{id}_V$ is **not invertible**;
(d) $\det(\lambda\,\mathrm{id}_V - T) = 0$;
(e) $p_T(\lambda) = 0$.

In particular T has at most n eigenvalues. For $\mathbf{A} \in M_n(F)$: λ is an eigenvalue of $\mathbf{A}$ if and only if $p_{\mathbf{A}}(\lambda) = 0$, if and only if $\lambda\mathbf{I} - \mathbf{A}$ is not invertible.

Proof.

(a) $\Leftrightarrow$ (b). By Definition 9.2.1 and Definition 9.2.2, λ is an eigenvalue exactly when $E_\lambda(T) = \ker(T - \lambda\,\mathrm{id}_V)$ contains a nonzero vector. By Theorem 2.2.3, this happens exactly when $T - \lambda\,\mathrm{id}_V$ is not injective.

(b) $\Leftrightarrow$ (c). V is finite-dimensional, so by Theorem 2.5.7 ((a) $\Leftrightarrow$ (b)) an operator on V is invertible if and only if it is injective.

(c) $\Leftrightarrow$ (d). The operators $\lambda\,\mathrm{id}_V - T$ and $T - \lambda\,\mathrm{id}_V$ differ by the factor -1, so one is invertible exactly when the other is. By Theorem 7.5.8, $\lambda\,\mathrm{id}_V - T$ is invertible if and only if $\det(\lambda\,\mathrm{id}_V - T) \neq 0$.

(d) $\Leftrightarrow$ (e). Fix a basis $\mathcal{B}$ of V and put $\mathbf{A} = [T]_{\mathcal{B}}^{\mathcal{B}}$. By Corollary 4.1.3, $[\lambda\,\mathrm{id}_V - T]_{\mathcal{B}}^{\mathcal{B}} = \lambda\mathbf{I}_n - \mathbf{A}$, and by Theorem 4.1.5 (b), $\lambda\,\mathrm{id}_V - T$ is invertible if and only if $\lambda\mathbf{I}_n - \mathbf{A}$ is. By Theorem 7.10.4 ((a), (i) $\Leftrightarrow$ (ii)), $\lambda\mathbf{I}_n - \mathbf{A}$ is not invertible if and only if $p_{\mathbf{A}}(\lambda) = 0$, and $p_{\mathbf{A}} = p_T$ by Definition 7.10.2.

The bound on the number of eigenvalues is Theorem 7.10.4 (b) applied to $\mathbf{A}$, since by (e) the eigenvalues of T are the roots of $p_{\mathbf{A}}$ in F. The statement for matrices is the case $T = T_{\mathbf{A}}$, $\mathcal{B}$ standard, $[T_{\mathbf{A}}]_{\mathcal{B}}^{\mathcal{B}} = \mathbf{A}$.

■

For matrices, this is Theorem 7.10.4 read in the new language: Chapter 7 found the scalars c with $c\mathbf{I} - \mathbf{A}$ singular, and those are the eigenvalues. It gives a procedure with two steps.

1. Compute $p_{\mathbf{A}}(x) = \det(x\mathbf{I} - \mathbf{A})$ and find its roots **in** F. These are the eigenvalues.
2. For each eigenvalue λ, solve the homogeneous system $(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}$ by elimination. Its solution set is $E_\lambda(\mathbf{A})$, and a basis of it lists "all" eigenvectors for λ (Theorem 3.5.4).

Example. Eigenvalues and Eigenspaces of a 2×2 Matrix

Find the eigenvalues and eigenspaces of $\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 3 & 2 \end{pmatrix} \in M_2(\mathbb{R})$.

Solution. Step 1. For a 2×2 matrix, $p_{\mathbf{A}}(x) = x^2 - (\operatorname{tr}\mathbf{A})x + \det\mathbf{A}$ (the examples after Definition 7.10.1). Here $\operatorname{tr}\mathbf{A} = 3$ and $\det\mathbf{A} = 2 - 6 = -4$, so

$$p_{\mathbf{A}}(x) = x^2 - 3x - 4 = (x-4)(x+1).$$

By Theorem 9.2.1, the eigenvalues are 4 and -1.

Step 2. For $\lambda = 4$,

$$\mathbf{A} - 4\mathbf{I} = \begin{pmatrix} -3 & 2 \\ 3 & -2 \end{pmatrix}.$$

The second row is -1 times the first, so the system reduces to $-3x + 2y = 0$, with solutions $(x, y) = t(2, 3)$. Hence $E_4(\mathbf{A}) = \operatorname{span}((2, 3))$. For $\lambda = -1$,

$$\mathbf{A} + \mathbf{I} = \begin{pmatrix} 2 & 2 \\ 3 & 3 \end{pmatrix},$$

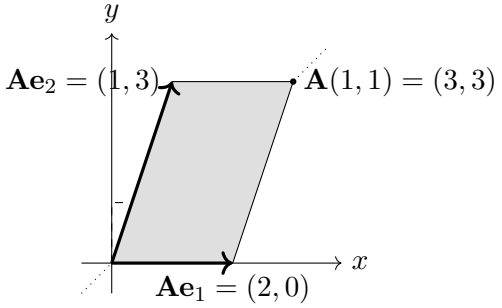
which reduces to $x + y = 0$, so $E_{-1}(\mathbf{A}) = \operatorname{span}((1, -1))$.

Check. $\mathbf{A}(2, 3) = (2 + 6, 6 + 6) = (8, 12) = 4\,(2, 3)$ and $\mathbf{A}(1, -1) = (1 - 2, 3 - 2) = (-1, 1) = -(1, -1)$. In the basis $((2, 3), (1, -1))$, the operator $T_{\mathbf{A}}$ has matrix $\operatorname{diag}(4, -1)$, by the computation at the start of this section.

In Step 2, the matrix $\mathbf{A} - \lambda\mathbf{I}$ **must** be singular; if elimination produces only the trivial solution, then λ was not a root of $p_{\mathbf{A}}$, and Step 1 contains an error. This is a built-in check.

Triangular matrices. If $\mathbf{A}$ is upper or lower triangular, then $p_{\mathbf{A}}(x) = (x - a_{11}) \cdots (x - a_{nn})$ (the examples after Definition 7.10.1), so the eigenvalues of a triangular matrix are its diagonal entries. For $\begin{pmatrix} 2 & 1 \\ 0 & 3 \end{pmatrix}$ they are 2 and 3, with eigenvectors $\mathbf{e}_1$ and $(1, 1)$, found in the previous section.

You can watch eigenvectors in the widget below. It shows the unit square and its image under $\mathbf{A} = \begin{pmatrix} 2 & 1 \\ 0 & 3 \end{pmatrix}$. The image $\mathbf{A}\mathbf{e}_1$ lies on the x-axis, which says that $\mathbf{e}_1$ is an eigenvector. The far corner of the image parallelogram is $\mathbf{A}(1, 1) = \mathbf{A}\mathbf{e}_1 + \mathbf{A}\mathbf{e}_2 = (3, 3)$, and it lies on the diagonal line through $(1, 1)$: so $(1, 1)$ is an eigenvector too, for the eigenvalue 3. Drag $\mathbf{A}\mathbf{e}_2$ to $(1, 2)$, which changes the matrix to $\begin{pmatrix} 2 & 1 \\ 0 & 2 \end{pmatrix}$. Now the far corner $(3, 2)$ leaves the diagonal, and only the x-axis is still mapped into itself. The next section explains what was lost.



The unit square (dashed) and its image under $\mathbf{A} = \begin{pmatrix} 2 & 1 \\ 0 & 3 \end{pmatrix}$. The image of $\mathbf{e}_1$ stays on the x-axis

(eigenvalue 2), and the image of $(1, 1)$ stays on the dotted diagonal (eigenvalue 3). The image of $\mathbf{e}_2$ leaves the y-axis, so $\mathbf{e}_2$ is not an eigenvector.

The set of all eigenvalues is used so often that it gets a symbol.

Definition 9.2.3. Spectrum

Let $T \in \mathcal{L}(V)$. The **spectrum** of T is the set $\operatorname{spec}(T) \subseteq F$ of all eigenvalues of T. For $\mathbf{A} \in M_n(F)$, $\operatorname{spec}(\mathbf{A}) \coloneqq \operatorname{spec}(T_{\mathbf{A}})$.

For instance $\operatorname{spec}\begin{pmatrix} 1 & 2 \\ 3 & 2 \end{pmatrix} = \{4, -1\}$, $\operatorname{spec}(\mathrm{id}_V) = \{1\}$ for $V \neq \{0\}$, and the spectrum of a projection is contained in $\{0, 1\}$. By Theorem 9.2.1, if $\dim V = n \geq 1$, then $\operatorname{spec}(T)$ is the set of roots of p_T in F, and it has at most n elements. It may be empty: the characteristic polynomial of the rotation $\mathbf{R} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ is $x^2 + 1$, which has no real root, so $\operatorname{spec}(\mathbf{R}) = \emptyset$ over $\mathbb{R}$. This matches Example (A Rotation Has No Invariant Lines): no invariant lines, no eigenvectors.

Infinite dimension. Without a characteristic polynomial, we go back to the definition, and anything can happen.

- **Differentiation of smooth functions.** Let V be the space of infinitely differentiable functions $\mathbb{R} \to \mathbb{R}$ (a subspace of $\mathbb{R}^{\mathbb{R}}$, by the argument of Example (Differentiable Functions) applied to all derivatives), and $Df = f'$. Here we take three facts from calculus as known: $e^{\lambda x}$ is smooth with derivative $\lambda e^{\lambda x}$; the product rule; and a function on $\mathbb{R}$ with derivative 0 everywhere is constant. For **every** $\lambda \in \mathbb{R}$, the function $f(x) = e^{\lambda x}$ is nonzero and satisfies $Df = \lambda f$. So $\operatorname{spec}(D) = \mathbb{R}$, an infinite set. In fact $E_\lambda(D) = \operatorname{span}(e^{\lambda x})$: if $f' = \lambda f$, then by the product rule $g(x) = e^{-\lambda x} f(x)$ has $g' = e^{-\lambda x}(f' - \lambda f) = 0$, so g is a constant c and $f = ce^{\lambda x}$.
- **The right shift.** Let $R(s_0, s_1, s_2, \dots) = (0, s_0, s_1, \dots)$ on $F^{\mathbb{N}}$ (Example (The Shift Maps on Sequences)). If $Rs = \lambda s$, compare entries: $0 = \lambda s_0$, and $s_k = \lambda s_{k+1}$ for all $k \geq 0$. If $\lambda = 0$, the second equation gives $s_k = 0$ for all k. If $\lambda \neq 0$, the first gives $s_0 = 0$, and then $s_{k+1} = \lambda^{-1} s_k$ gives $s_1 = 0$, $s_2 = 0$, and so on by induction. Either way $s = 0$. So $\operatorname{spec}(R) = \emptyset$, over **every** field.

Warning. **In infinite dimension, "not invertible" does not produce an eigenvalue.** The step (b) $\Leftrightarrow$ (c) of Theorem 9.2.1 used finite dimension. The right shift R is injective but not surjective (Example (The Shifts: Injective but Not Surjective, and Conversely)), so $R - 0 \cdot \mathrm{id}$ is not invertible, yet 0 is not an eigenvalue of R: nothing nonzero is sent to 0.

9.2.3 Distinct eigenvalues give independent eigenvectors

In Example (Eigenvalues and Eigenspaces of a 2×2 Matrix) the eigenvectors $(2, 3)$ and $(1, -1)$ for the eigenvalues 4 and -1 formed a basis. That is no accident. Eigenvectors for different eigenvalues can never be linearly dependent.

Theorem 9.2.2. Distinct Eigenvalues Give Independent Eigenvectors

Let $T \in \mathcal{L}(V)$, let $\lambda_1, \dots, \lambda_k \in F$ be **distinct** eigenvalues of T, and for each i let v_i be an eigenvector of T for λ_i. Then $(v_1, \dots, v_k)$ is linearly independent.

Idea. Try two vectors first. From $av_1 + bv_2 = 0$, we want to kill one term, and the operator $T - \lambda_2 \operatorname{id}_V$ does exactly that: it sends v_2 to 0 and v_1 to $(\lambda_1 - \lambda_2)v_1$. So $a(\lambda_1 - \lambda_2)v_1 = 0$, and since $\lambda_1 \neq \lambda_2$ and $v_1 \neq 0$, $a = 0$; then $b = 0$. In general, applying $T - \lambda_j \operatorname{id}_V$ to a dependence kills the j-th vector and rescales the others by nonzero factors, producing a **shorter** dependence. So a shortest dependence cannot exist. The Linear Dependence Lemma lets us pick that shortest one cleanly.

Proof. Suppose, for a contradiction, that $(v_1, \dots, v_k)$ is linearly dependent. By Theorem 1.4.1 (1), there is an index j with $v_j \in \operatorname{span}(v_1, \dots, v_{j-1})$; choose the **smallest** such j. Since $v_1 \neq 0$, $j \geq 2$. The list $(v_1, \dots, v_{j-1})$ is linearly independent: otherwise Theorem 1.4.1 (1), applied to it, would give a smaller index with the same property. Write

$$v_j = a_1 v_1 + \cdots + a_{j-1} v_{j-1} \qquad (a_i \in F).$$

Apply $T - \lambda_j \operatorname{id}_V$ to both sides. Since $(T - \lambda_j \operatorname{id}_V)v_i = (\lambda_i - \lambda_j)v_i$ for each i, linearity gives

$$0 = a_1(\lambda_1 - \lambda_j)v_1 + \cdots + a_{j-1}(\lambda_{j-1} - \lambda_j)v_{j-1}.$$

By independence of $(v_1, \dots, v_{j-1})$, every coefficient $a_i(\lambda_i - \lambda_j)$ is 0. The eigenvalues are distinct, so $\lambda_i - \lambda_j \neq 0$ for $i < j$, and hence $a_i = 0$ for all $i < j$. Then $v_j = 0$, which contradicts the fact that an eigenvector is nonzero. This proves that $(v_1, \dots, v_k)$ is linearly independent.

■

The theorem works in any vector space, finite-dimensional or not. For instance, for distinct real numbers $\lambda_1, \dots, \lambda_k$, the functions $e^{\lambda_1 x}, \dots, e^{\lambda_k x}$ are eigenvectors of differentiation for distinct eigenvalues, so they are linearly independent, with no computation at all. Two consequences follow at once, and the second is the first half of a promise made in Chapter 1, §7: eigenspaces sit inside V as a direct sum.

Corollary 9.2.3. Eigenspaces Form a Direct Sum

Let $T \in \mathcal{L}(V)$ and let $\lambda_1, \dots, \lambda_k \in F$ be distinct. Then the sum $E_{\lambda_1}(T) + \cdots + E_{\lambda_k}(T)$ is direct. If V is finite-dimensional, then

$$\dim E_{\lambda_1}(T) + \cdots + \dim E_{\lambda_k}(T) \leq \dim V,$$

and T has at most $\dim V$ distinct eigenvalues.

Proof. By Theorem 1.7.5 ((b) ⇒ (a)), it suffices to show: if $u_1 + \cdots + u_k = 0$ with $u_i \in E_{\lambda_i}(T)$, then every $u_i = 0$. Suppose not, and let S be the nonempty set of indices i with $u_i \neq 0$. For $i \in S$, u_i is an eigenvector of T for λ_i, and these eigenvalues are distinct. By Theorem 9.2.2, the list $(u_i)_{i \in S}$ is linearly independent. But $\sum_{i \in S} 1 \cdot u_i = \sum_{i=1}^{k} u_i = 0$ is a relation with all coefficients $1 \neq 0$, a contradiction. So the sum is direct.

If V is finite-dimensional, Theorem 1.7.5 ((a) $\Rightarrow$ (e)) gives $\dim(E_{\lambda_1}(T) \oplus \cdots \oplus E_{\lambda_k}(T)) = \sum_i \dim E_{\lambda_i}(T)$, and this is at most $\dim V$ by Theorem 1.6.7. If the λ_i are eigenvalues, each $\dim E_{\lambda_i}(T) \geq 1$, so $k \leq \dim V$.

■

The corollary gives a second, polynomial-free proof that an operator on an n-dimensional space has at most n eigenvalues. Section 4 shows that T has a diagonal matrix exactly when the inequality is an equality.

9.2.4 Existence over the complex numbers

We have seen operators without eigenvalues: the rotation of $\mathbb{R}^2$, and the right shift. The rotation fails because $\mathbb{R}$ is missing the roots of x^2+1; the shift fails because its space is infinite-dimensional. Remove both obstacles and eigenvalues always exist.

Theorem 9.2.4. Complex Operators Have Eigenvalues

Let V be a finite-dimensional vector space over $\mathbb{C}$ with $V \neq \{0\}$, and let $T \in \mathcal{L}(V)$. Then T has an eigenvalue. The same holds for every $\mathbf{A} \in M_n(\mathbb{C})$ with $n \geq 1$.

Proof. Let $n = \dim V \geq 1$. By Theorem 7.10.2 and Definition 7.10.2, $p_T \in \mathbb{C}[x]$ is monic of degree $n \geq 1$, so it is not constant. By the Fundamental Theorem of Algebra (Theorem 6.5.1), $p_T(\lambda) = 0$ for some $\lambda \in \mathbb{C}$. By Theorem 9.2.1 ((e) $\Rightarrow$ (a)), λ is an eigenvalue of T. For $\mathbf{A} \in M_n(\mathbb{C})$, apply this to $T_{\mathbf{A}}$.

■

Alternatively, determinants can be avoided altogether, and the proof then shows where the eigenvector comes from. By Theorem 6.7.2, there is a nonzero $p \in \mathbb{C}[x]$ with $p(T) = 0$. It is not a nonzero constant c, since $c\,\mathrm{id}_V \neq 0$ on $V \neq \{0\}$. By Corollary 6.5.5, $p = c(x - z_1)\cdots(x - z_m)$ with $c \neq 0$ and $m \geq 1$, so by Theorem 6.7.1 (a) and (b),

$$0 = p(T) = c\,(T - z_1\,\mathrm{id}_V)(T - z_2\,\mathrm{id}_V)\cdots(T - z_m\,\mathrm{id}_V).$$

Since $c \neq 0$, multiplying by c^{-1} shows that the composite $(T - z_1\,\mathrm{id}_V)\cdots(T - z_m\,\mathrm{id}_V)$ is the zero operator, which is not injective on $V \neq \{0\}$. A composition of injective maps is injective (Theorem 0.4.3 (a), applied $m-1$ times), so at least one factor $T - z_j\,\mathrm{id}_V$ is not injective, and z_j is an eigenvalue by Theorem 9.2.1 ((b) $\Rightarrow$ (a)).

Both routes use the Fundamental Theorem of Algebra, and both need all three hypotheses. Over $\mathbb{R}$ the rotation has no eigenvalue. In infinite dimension the right shift on $\mathbb{C}^{\mathbb{N}}$ has none, and indeed no nonzero polynomial annihilates it: for $p = a_0 + a_1 x + \cdots + a_N x^N \neq 0$, the sequence $p(R)(1,0,0,\dots) = (a_0, a_1, \dots, a_N, 0, \dots)$ is not zero. On $V = \{0\}$ there is nothing to be an eigenvector. This theorem pays off the promise made in Chapter 6, and it is the base step of every structure theorem over $\mathbb{C}$ in this book: find one eigenvector, then look at what is left.

Quick check. Does every operator on $\mathbb{C}^3$ have an invariant line? Does every operator on $\mathbb{R}^2$?

Solution. Every operator on $\mathbb{C}^3$ has one: by Theorem 9.2.4 it has an eigenvector $\mathbf{v}$, and $\operatorname{span}(\mathbf{v})$ is invariant by Proposition 9.1.2. Not every operator on $\mathbb{R}^2$: the rotation by a right angle has no invariant line (Example (A Rotation Has No Invariant Lines)).

9.2.5 Real matrices and complex eigenvalues

A real matrix may have no real eigenvalues, but it is also a complex matrix, and as such it has eigenvalues by Theorem 9.2.4. This is the situation Chapter 1, §9 prepared for. Let $\mathbf{A} \in M_n(\mathbb{R})$. The complexification of $\mathbb{R}^n$ is $\mathbb{C}^n$: the pair $(\mathbf{x}, \mathbf{y})$ is the vector $\mathbf{x} + i\mathbf{y}$ (Example (The Complexification of $\mathbb{R}^n$ Is $\mathbb{C}^n$)). The real operator $T_{\mathbf{A}}$ extends to the complexification by the only rule compatible with complex scalars,

$$\mathbf{x} + i\mathbf{y} \;\mapsto\; \mathbf{A}\mathbf{x} + i\mathbf{A}\mathbf{y} = \mathbf{A}(\mathbf{x} + i\mathbf{y}),$$

which is just $T_{\mathbf{A}}$ for $\mathbf{A}$ regarded as a matrix in $M_n(\mathbb{C})$. By Theorem 1.9.3, the real standard basis is also a complex basis of $\mathbb{C}^n$, and in it the extended operator still has matrix $\mathbf{A}$; in particular it has the same characteristic polynomial $p_{\mathbf{A}} \in \mathbb{R}[x] \subseteq \mathbb{C}[x]$. So the **complex eigenvalues** of a real matrix are the roots of $p_{\mathbf{A}}$ in $\mathbb{C}$, with eigenvectors in $\mathbb{C}^n$, and the real eigenvalues are among them. The complex ones cannot appear alone.

Theorem 9.2.5. Complex Eigenvalues of Real Matrices Come in Conjugate Pairs

Let $\mathbf{A} \in M_n(\mathbb{R})$, regarded as an element of $M_n(\mathbb{C})$. If $\lambda \in \mathbb{C}$ is an eigenvalue of $\mathbf{A}$ with eigenvector $\mathbf{v} \in \mathbb{C}^n$, then $\overline{\lambda}$ is an eigenvalue of $\mathbf{A}$ with eigenvector $\overline{\mathbf{v}}$, the vector of conjugated entries. Moreover:

(a) $\operatorname{mult}_{\lambda}(p_{\mathbf{A}}) = \operatorname{mult}_{\overline{\lambda}}(p_{\mathbf{A}})$;
(b) if $\lambda \notin \mathbb{R}$, then $(\mathbf{v}, \overline{\mathbf{v}})$ is linearly independent over $\mathbb{C}$.

Proof. Let $\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$ with $\mathbf{v} = (v_1, \dots, v_n) \neq \mathbf{0}$. For each k, the k-th entry of $\mathbf{A}\mathbf{v}$ is $\sum_j a_{kj} v_j$ (Definition 0.9.9). By Theorem 0.6.2 (a) and (b), and since $\overline{a_{kj}} = a_{kj}$ for the real entries of $\mathbf{A}$,

$$\overline{\textstyle\sum_j a_{kj} v_j} = \sum_j a_{kj}\overline{v_j}, \qquad \overline{\lambda v_k} = \overline{\lambda}\,\overline{v_k}.$$

Conjugating the equation $\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$ entry by entry therefore gives $\mathbf{A}\overline{\mathbf{v}} = \overline{\lambda}\,\overline{\mathbf{v}}$. Since $\overline{\overline{v_k}} = v_k$, $\overline{\mathbf{v}} = \mathbf{0}$ would force $\mathbf{v} = \mathbf{0}$; so $\overline{\mathbf{v}} \neq \mathbf{0}$, and $\overline{\mathbf{v}}$ is an eigenvector for $\overline{\lambda}$.

(a) The entries of $x\mathbf{I} - \mathbf{A}$ lie in $\mathbb{R}[x]$, so its determinant $p_{\mathbf{A}}$ lies in $\mathbb{R}[x]$ (Definition 7.10.1). By Theorem 6.5.6, $\operatorname{mult}_{\overline{\lambda}}(p_{\mathbf{A}}) = \operatorname{mult}_{\lambda}(p_{\mathbf{A}})$.

(b) If $\lambda \notin \mathbb{R}$, then $\overline{\lambda} \neq \lambda$, so $\mathbf{v}$ and $\overline{\mathbf{v}}$ are eigenvectors for distinct eigenvalues, and they are independent by Theorem 9.2.2.

■

For the rotation $\mathbf{R} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$, the vector $(1, -i)$ satisfies $\mathbf{R}(1, -i) = (i, 1) = i\,(1, -i)$, so i is an eigenvalue. Conjugating, $\mathbf{R}(1, i) = (-i, 1) = -i\,(1, i)$. Over $\mathbb{C}$, the two eigenvectors form a basis of $\mathbb{C}^2$, and the rotation has a diagonal matrix $\operatorname{diag}(i, -i)$ in it, although no real basis diagonalizes it. So the field matters for eigenvalues, and it will matter for diagonalization.

Example. A Real Matrix with Non-Real Eigenvalues

Find the eigenvalues of $\mathbf{A} = \begin{pmatrix} 3 & -2 \\ 4 & -1 \end{pmatrix}$ over $\mathbb{C}$, with an eigenvector for each.

Solution. Here $\operatorname{tr}\mathbf{A} = 2$ and $\det\mathbf{A} = -3 + 8 = 5$, so $p_{\mathbf{A}}(x) = x^2 - 2x + 5 = (x-1)^2 + 4$. Over $\mathbb{R}$ this has no root, so $\operatorname{spec}(\mathbf{A}) = \emptyset$ over $\mathbb{R}$. Over $\mathbb{C}$, $(x-1)^2 = -4$ gives the roots $1+2i$ and $1-2i$.

For $\lambda = 1 + 2i$,

$$\mathbf{A} - \lambda\mathbf{I} = \begin{pmatrix} 2-2i & -2 \\ 4 & -2-2i \end{pmatrix}.$$

The first row gives $(2-2i)x = 2y$, that is, $y = (1-i)x$; take $\mathbf{v} = (1, 1-i)$. Check the second row: $4 \cdot 1 + (-2-2i)(1-i) = 4 + (-2+2i-2i+2i^2) = 4 - 4 = 0$. So $\mathbf{v}$ is an eigenvector for $1+2i$. By Theorem 9.2.5, $\overline{\mathbf{v}} = (1, 1+i)$ is an eigenvector for $1-2i$, with no further computation.

9.2.6 The invertible matrix theorem grows

Chapter 2 collected the conditions equivalent to invertibility for an operator (Theorem 2.5.7), Chapter 3 gave the matrix form, the eight conditions of the Invertible Matrix Theorem (Theorem 3.4.4), and Chapter 7 added the determinant (Theorem 7.5.3). Eigenvalues give one more: a matrix is singular exactly when it kills a nonzero vector, which is to say, when it stretches some nonzero vector by the factor 0.

Theorem 9.2.6. Invertible Matrix Theorem, with Eigenvalues

Let $\mathbf{A} \in M_n(F)$ with $n \geq 1$. The following are equivalent, and each is equivalent to each of the conditions of Theorem 7.5.3 and Theorem 3.4.4.

(a) $\mathbf{A}$ is invertible.

(b) 0 is **not** an eigenvalue of $\mathbf{A}$, that is, $0 \notin \operatorname{spec}(\mathbf{A})$.

Likewise, an operator T on a finite-dimensional space $V \neq \{0\}$ is invertible if and only if $0 \notin \operatorname{spec}(T)$.

Proof. By Definition 9.2.2, $E_0(\mathbf{A}) = \ker(T_{\mathbf{A}} - 0\,\mathrm{id}) = \ker T_{\mathbf{A}}$. So $0 \notin \operatorname{spec}(\mathbf{A})$ if and only if $\ker T_{\mathbf{A}} = \{\mathbf{0}\}$, if and only if $T_{\mathbf{A}}$ is injective (Theorem 2.2.3), if and only if $\mathbf{A}$ is invertible (Theorem 7.5.3, (a) $\Leftrightarrow$ (b)). For an operator, $E_0(T) = \ker T$, and Theorem 2.5.7 ((a) $\Leftrightarrow$ (c)) finishes the same way. ■

For example, $\operatorname{spec}\begin{pmatrix} 1 & 2 \\ 3 & 2 \end{pmatrix} = \{4, -1\}$ does not contain 0, so that matrix is invertible, while $\begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix}$ has the eigenvalue 0 and is singular.

Eigenvalues are easy to compute for one matrix, and it is tempting to assume they behave well under the operations we use on matrices. Two such assumptions are false.

Warning. **Eigenvalues do not add or multiply.** Let $\mathbf{A} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$ and $\mathbf{B} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}$. Both are triangular with zero diagonal, so $\operatorname{spec}(\mathbf{A}) = \operatorname{spec}(\mathbf{B}) = \{0\}$. But $\mathbf{A} + \mathbf{B} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ has $p(x) = x^2 - 1$ and eigenvalues ± 1, which are not sums of eigenvalues of $\mathbf{A}$ and $\mathbf{B}$; and $\mathbf{AB} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$ has the eigenvalue 1, which is not a product of them.

Warning. **Row operations change eigenvalues.** Elimination preserves the solution set of $\mathbf{Ax} = \mathbf{0}$ (Theorem 3.2.2), not the solution sets of $\mathbf{Ax} = \lambda\mathbf{x}$. For $\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 3 & 2 \end{pmatrix}$, with eigenvalues 4 and -1, subtracting 3 times row 1 from row 2 gives the triangular matrix $\begin{pmatrix} 1 & 2 \\ 0 & -4 \end{pmatrix}$, whose eigenvalues are 1 and -4. So never row reduce $\mathbf{A}$ to find its eigenvalues; row reduce $\mathbf{A} - \lambda\mathbf{I}$ to find eigenvectors, once λ is known.

9.2.7 Exercises

9.2.7.1 A. Check your understanding

Exercise. A1

(a) Define eigenvalue, eigenvector, and eigenspace of an operator $T \in \mathcal{L}(V)$.

(b) True or false: 0 is an eigenvector of every operator. Justify your answer.

(c) True or false: if 0 is an eigenvalue of $\mathbf{A} \in M_n(F)$, then $\mathbf{A}$ is not invertible. Justify your answer.

(d) State four conditions equivalent to "λ is an eigenvalue of T" for an operator on a finite-dimensional space $V \neq \{0\}$.

(e) True or false: every operator on $\mathbb{R}^2$ has an eigenvalue. Justify your answer.

(f) True or false: the eigenvalues of $\mathbf{A}$ are the diagonal entries of a row echelon form of $\mathbf{A}$. Justify your answer.

Solution.

(a) $\lambda \in F$ is an eigenvalue of T if $Tv = \lambda v$ for some **nonzero** $v \in V$; each such nonzero v is an eigenvector for λ (Definition 9.2.1). The eigenspace is $E_\lambda(T) = \ker(T - \lambda \operatorname{id}_V)$ (Definition 9.2.2).

(b) False. Eigenvectors are nonzero by definition. (The equation $T0 = \lambda 0$ holds for every λ, which is why 0 is excluded.)

(c) True, by Theorem 9.2.6.

(d) By Theorem 9.2.1: $T - \lambda \operatorname{id}_V$ is not injective; $T - \lambda \operatorname{id}_V$ is not invertible; $\det(\lambda \operatorname{id}_V - T) =$

0; $p_T(\lambda) = 0$.

(e) False. The rotation $R(x, y) = (-y, x)$ has $p_R = x^2 + 1$, with no real root, so it has no real eigenvalue.

(f) False. Row operations change eigenvalues: $\begin{pmatrix} 1 & 2 \\ 3 & 2 \end{pmatrix}$ has eigenvalues $4, -1$, but the echelon form $\begin{pmatrix} 1 & 2 \\ 0 & -4 \end{pmatrix}$ has diagonal entries $1, -4$.

9.2.7.2 B. Practice

Exercise. B1: Computing eigenvalues and eigenspaces

Find the eigenvalues and a basis of each eigenspace of the following real matrices.

(a) $\mathbf{A} = \begin{pmatrix} 5 & 4 \\ -2 & -1 \end{pmatrix}$.

(b) $\mathbf{B} = \begin{pmatrix} 2 & 0 & 1 \\ -1 & 3 & 1 \\ 2 & -2 & 1 \end{pmatrix}$.

Hence write down a basis of $\mathbb{R}^2$ consisting of eigenvectors of $\mathbf{A}$, and a basis of $\mathbb{R}^3$ consisting of eigenvectors of $\mathbf{B}$.

Solution.

(a) $\operatorname{tr}\mathbf{A} = 4$ and $\det\mathbf{A} = -5 + 8 = 3$, so $p_{\mathbf{A}}(x) = x^2 - 4x + 3 = (x - 1)(x - 3)$, and $\operatorname{spec}(\mathbf{A}) = \{1, 3\}$ by Theorem 9.2.1. For $\lambda = 1$, $\mathbf{A} - \mathbf{I} = \begin{pmatrix} 4 & 4 \\ -2 & -2 \end{pmatrix}$ reduces to $x + y = 0$, so $E_1(\mathbf{A}) = \operatorname{span}((1, -1))$. For $\lambda = 3$, $\mathbf{A} - 3\mathbf{I} = \begin{pmatrix} 2 & 4 \\ -2 & -4 \end{pmatrix}$ reduces to $x + 2y = 0$, so $E_3(\mathbf{A}) = \operatorname{span}((2, -1))$. Check: $\mathbf{A}(1, -1) = (1, -1)$ and $\mathbf{A}(2, -1) = (6, -3) = 3(2, -1)$.

(b) Expanding $\det(x\mathbf{I} - \mathbf{B})$ along the first row,

$$\begin{aligned} p_{\mathbf{B}}(x) &= \det \begin{pmatrix} x - 2 & 0 & -1 \\ 1 & x - 3 & -1 \\ -2 & 2 & x - 1 \end{pmatrix} \\ &= (x - 2)\big[(x - 3)(x - 1) + 2\big] + (-1)\big[1 \cdot 2 - (x - 3)(-2)\big]. \end{aligned}$$

The first bracket is $x^2 - 4x + 5$ and the second is $2x - 4 = 2(x - 2)$, so

$$\begin{aligned} p_{\mathbf{B}}(x) &= (x - 2)(x^2 - 4x + 5) - 2(x - 2) \\ &= (x - 2)(x^2 - 4x + 3) = (x - 1)(x - 2)(x - 3). \end{aligned}$$

So $\operatorname{spec}(\mathbf{B}) = \{1, 2, 3\}$. Solving $(\mathbf{B} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}$:

- $\lambda = 1$: $\mathbf{B} - \mathbf{I} = \begin{pmatrix} 1 & 0 & 1 \\ -1 & 2 & 1 \\ 2 & -2 & 0 \end{pmatrix}$. The first row gives $z = -x$, the third gives $y = x$, and

the second is then $-x+2x-x=0$. So $E_1(\mathbf{B})=\operatorname{span}((1,1,-1))$.

- $\lambda=2$: $\mathbf{B}-2\mathbf{I}=\begin{pmatrix}0&0&1\\-1&1&1\\2&-2&-1\end{pmatrix}$. The first row gives $z=0$, and then the second gives $y=x$. So $E_2(\mathbf{B})=\operatorname{span}((1,1,0))$.
- $\lambda=3$: $\mathbf{B}-3\mathbf{I}=\begin{pmatrix}-1&0&1\\-1&0&1\\2&-2&-2\end{pmatrix}$. The first row gives $z=x$, and the third gives $2x-2y-2x=0$, so $y=0$. So $E_3(\mathbf{B})=\operatorname{span}((1,0,1))$.

Check: $\mathbf{B}(1,1,-1)=(1,1,-1)$, $\mathbf{B}(1,1,0)=(2,2,0)$, $\mathbf{B}(1,0,1)=(3,0,3)$.

Hence: the eigenvalues of $\mathbf{A}$ are distinct, so by Theorem 9.2.2 the eigenvectors $((1,-1),(2,-1))$ are independent, and two independent vectors in $\mathbb{R}^2$ form a basis (Theorem 1.6.4). In the same way $((1,1,-1),(1,1,0),(1,0,1))$ is a basis of $\mathbb{R}^3$ consisting of eigenvectors of $\mathbf{B}$.

Exercise. B2: The transpose map

Let F be a field of characteristic not 2, and let $\tau\colon M_2(F)\to M_2(F)$, $\tau(\mathbf{X})=\mathbf{X}^\top$.

(a) Show that every eigenvalue λ of τ satisfies $\lambda^2=1$.

(b) Find $\operatorname{spec}(\tau)$, and describe $E_1(\tau)$ and $E_{-1}(\tau)$ with a basis of each.

Solution.

(a) Let $\tau(\mathbf{X})=\lambda\mathbf{X}$ with $\mathbf{X}\neq\mathbf{O}$. Since $(\mathbf{X}^\top)^\top=\mathbf{X}$ (Theorem 0.9.5), $\mathbf{X}=\tau(\tau(\mathbf{X}))=\tau(\lambda\mathbf{X})=\lambda\tau(\mathbf{X})=\lambda^2\mathbf{X}$. So $(\lambda^2-1)\mathbf{X}=\mathbf{O}$ with $\mathbf{X}\neq\mathbf{O}$, and $\lambda^2=1$ by Theorem 1.1.9.

(b) From $\lambda^2-1=(\lambda-1)(\lambda+1)=0$, the only candidates are 1 and -1, which are different because the characteristic is not 2. $E_1(\tau)=\{\mathbf{X}:\mathbf{X}^\top=\mathbf{X}\}$ is the space of symmetric matrices. Writing $\mathbf{X}=\begin{pmatrix}a&b\\c&d\end{pmatrix}$, the condition is $b=c$, so $E_1(\tau)=\operatorname{span}(\mathbf{E}_{11},\mathbf{E}_{22},\mathbf{E}_{12}+\mathbf{E}_{21})$, and these three matrices are independent (compare entries). $E_{-1}(\tau)=\{\mathbf{X}:\mathbf{X}^\top=-\mathbf{X}\}$ is the space of skew-symmetric matrices. The condition $\mathbf{X}^\top=-\mathbf{X}$ says $a=-a$, $d=-d$, and $c=-b$; since $2\neq0$, $a=d=0$, so $E_{-1}(\tau)=\operatorname{span}(\mathbf{E}_{12}-\mathbf{E}_{21})$. Both eigenspaces are nonzero, so $\operatorname{spec}(\tau)=\{1,-1\}$. This agrees with Theorem 4.5.5: $\tau^2=\mathrm{id}$, and $M_2(F)=E_1(\tau)\oplus E_{-1}(\tau)$, with dimensions $3+1=4$.

Exercise. B3: Eigenvalues of powers and inverses

Let $T\in\mathcal{L}(V)$ and let v be an eigenvector of T for λ.

(a) Prove that v is an eigenvector of T^2 for λ^2, and more generally of $p(T)$ for $p(\lambda)$, for every $p\in F[x]$.

(b) Suppose T is invertible. Prove that $\lambda\neq0$ and that v is an eigenvector of T^{-1} for λ^{-1}.

(c) Show that the converse of the first statement in (a) fails: give $T\in\mathcal{L}(\mathbb{R}^2)$ such that -1 is an eigenvalue of T^2 but T has no eigenvalue.

Solution.

(a) $T^2v = T(Tv) = T(\lambda v) = \lambda Tv = \lambda^2 v$, and $v \neq 0$. By induction on k, $T^k v = \lambda^k v$: if $T^k v = \lambda^k v$, then $T^{k+1}v = T(\lambda^k v) = \lambda^{k+1}v$. For $p = \sum_k a_k x^k$, Definition 2.4.2 gives $p(T)v = \sum_k a_k T^k v = \sum_k a_k \lambda^k v = p(\lambda)v$.

(b) An invertible operator is injective, so $\ker T = \{0\}$ and $Tv \neq 0$; as $Tv = \lambda v$, $\lambda \neq 0$ by Theorem 1.1.9. Applying T^{-1} to $Tv = \lambda v$ gives $v = \lambda T^{-1}v$, so $T^{-1}v = \lambda^{-1}v$.

(c) Take the rotation $R(x, y) = (-y, x)$. It has no real eigenvalue (Example (A Rotation Has No Invariant Lines)). But $R^2(x, y) = R(-y, x) = (-x, -y)$, so $R^2 = -\mathrm{id}$, and every nonzero vector is an eigenvector of R^2 for -1. Part (a) would need a real λ with $\lambda^2 = -1$, and there is none.

9.2.7.3 C. Going deeper

Exercise. C1: Every vector an eigenvector

Let V be a vector space over F and $T \in \mathcal{L}(V)$. Suppose every nonzero vector of V is an eigenvector of T. Prove that $T = c\,\mathrm{id}_V$ for some $c \in F$.

Hint: compare the eigenvalues of v, w, and $v + w$.

Solution. If $V = \{0\}$, then $T = 0 = 1 \cdot \mathrm{id}_V$. Otherwise, for each $v \neq 0$ let λ_v be its eigenvalue, which is unique by the well-definedness remark after Definition 9.2.1. We show that $\lambda_v = \lambda_w$ for all nonzero v, w.

Case 1: (v, w) is linearly dependent. Then $w = av$ for some $a \neq 0$, and $Tw = aTv = \lambda_v(av) = \lambda_v w$. By uniqueness of the eigenvalue of w, $\lambda_w = \lambda_v$.

Case 2: (v, w) is linearly independent. Then $v + w \neq 0$, and

$$\lambda_{v+w}v + \lambda_{v+w}w = T(v + w) = Tv + Tw = \lambda_v v + \lambda_w w.$$

So $(\lambda_{v+w} - \lambda_v)v + (\lambda_{v+w} - \lambda_w)w = 0$, and by independence $\lambda_v = \lambda_{v+w} = \lambda_w$.

Fix $v_0 \neq 0$ and put $c = \lambda_{v_0}$. Then $Tv = cv$ for every $v \neq 0$, and also for $v = 0$. Hence $T = c\,\mathrm{id}_V$.

Exercise. C2: Nilpotent operators

An operator $T \in \mathcal{L}(V)$ is **nilpotent** if $T^k = 0$ for some integer $k \geq 1$.

(a) Let $V \neq \{0\}$ and T nilpotent. Prove that $\mathrm{spec}(T) = \{0\}$. (No finite dimension is needed.)

(b) Let $\mathbf{A} \in M_n(\mathbb{C})$, $n \geq 1$, be nilpotent. Deduce that $p_{\mathbf{A}}(x) = x^n$.

(c) Let $\mathbf{A} \in M_n(F)$, $n \geq 1$, be nilpotent. Prove that $\mathbf{I} + \mathbf{A}$ is invertible.

Solution.

(a) Let $Tv = \lambda v$ with $v \neq 0$. By Exercise (B3: Eigenvalues of powers and inverses) (a), $0 = T^k v = \lambda^k v$, so $\lambda^k = 0$ by Theorem 1.1.9, and hence $\lambda = 0$ (a product of nonzero elements of a field is nonzero, Theorem 0.7.1). So $\operatorname{spec}(T) \subseteq \{0\}$. Conversely, $T^k = 0$ is not injective on $V \neq \{0\}$; if T were injective, so would be the composite T^k (Theorem 0.4.3). So $\ker T \neq \{0\}$, and $0 \in \operatorname{spec}(T)$.

(b) By Corollary 6.5.5 and Theorem 7.10.2, $p_{\mathbf{A}} = (x - z_1)\cdots(x - z_n)$ for some $z_i \in \mathbb{C}$. Each z_i is a root of $p_{\mathbf{A}}$, hence an eigenvalue of $\mathbf{A}$ by Theorem 9.2.1, hence $z_i = 0$ by (a). So $p_{\mathbf{A}} = x^n$.

(c) By Theorem 9.2.6, it suffices to show $0 \notin \operatorname{spec}(\mathbf{I} + \mathbf{A})$. Suppose $(\mathbf{I} + \mathbf{A})\mathbf{v} = \mathbf{0}$ with $\mathbf{v} \neq \mathbf{0}$. Then $\mathbf{A}\mathbf{v} = -\mathbf{v}$, so $-1 \in \operatorname{spec}(\mathbf{A})$. By (a), $\operatorname{spec}(\mathbf{A}) = \{0\}$, but $-1 \neq 0$ in every field. This contradiction shows that $\mathbf{I} + \mathbf{A}$ is invertible.

Exercise. C3: Real matrices of odd size

(a) Prove that every $\mathbf{A} \in M_n(\mathbb{R})$ with n odd has a real eigenvalue.

(b) Deduce that every operator on a real vector space of dimension 3 has an invariant line.

(c) Show that (a) fails for $n = 4$: find $\mathbf{A} \in M_4(\mathbb{R})$ with no real eigenvalue.

Solution.

(a) By Theorem 7.10.2, $p_{\mathbf{A}} \in \mathbb{R}[x]$ is monic of degree n, which is odd. By Corollary 6.5.9, $p_{\mathbf{A}}(\lambda) = 0$ for some $\lambda \in \mathbb{R}$, and λ is an eigenvalue of $\mathbf{A}$ by Theorem 9.2.1.

(b) Let $\dim V = 3$ over $\mathbb{R}$ and $T \in \mathcal{L}(V)$. Choose a basis $\mathcal{B}$. By (a), the real matrix $[T]_{\mathcal{B}}^{\mathcal{B}}$ has a real eigenvalue λ, which is a root of $p_{[T]_{\mathcal{B}}^{\mathcal{B}}} = p_T$ (Theorem 9.2.1, Definition 7.10.2). So λ is an eigenvalue of T, again by Theorem 9.2.1, and T has an eigenvector v for it. By Proposition 9.1.2, $\operatorname{span}(v)$ is invariant.

(c) Let $\mathbf{R} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ and $\mathbf{A} = \mathbf{R} \oplus \mathbf{R} \in M_4(\mathbb{R})$. Then $x\mathbf{I}_4 - \mathbf{A} = (x\mathbf{I}_2 - \mathbf{R}) \oplus (x\mathbf{I}_2 - \mathbf{R})$ is block diagonal, so by the block triangular determinant over $\mathbb{R}[x]$ (the remark after Theorem 7.8.4), $p_{\mathbf{A}} = p_{\mathbf{R}}^2 = (x^2 + 1)^2$. For real λ, $(\lambda^2 + 1)^2 \geq 1 > 0$, so $p_{\mathbf{A}}$ has no real root, and $\mathbf{A}$ has no real eigenvalue. The argument of (a) needs odd degree.

9.3 Algebraic and Geometric Multiplicity

An eigenvalue can be counted in two ways. It is a root of the characteristic polynomial, and a root has a multiplicity. It also has an eigenspace, and a subspace has a dimension. The two numbers need not agree, and the gap between them is exactly what can stop an operator from having a diagonal matrix. This section defines both counts, proves that the dimension never exceeds the root multiplicity, and shows that the root multiplicities package the trace and the determinant. The next section turns the comparison into a test for diagonalizability.

9.3.1 Two ways to count an eigenvalue

Compare two matrices over any field F:

$$2\mathbf{I}_2 = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}, \qquad \mathbf{J} = \begin{pmatrix} 2 & 1 \\ 0 & 2 \end{pmatrix}.$$

Both are triangular with diagonal $2, 2$, so both have characteristic polynomial $(x-2)^2$ and spectrum $\{2\}$. By the only measure we have so far, the eigenvalue data agree. Yet the eigenvectors are very different. For $2\mathbf{I}_2$, every nonzero vector is an eigenvector and $E_2 = F^2$. For $\mathbf{J}$, the matrix $\mathbf{J} - 2\mathbf{I} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$ kills exactly the multiples of $\mathbf{e}_1$, so $E_2(\mathbf{J}) = \operatorname{span}(\mathbf{e}_1)$ is a line. The root 2 "counts twice" in both characteristic polynomials, but it supplies two independent eigenvectors for $2\mathbf{I}_2$ and only one for $\mathbf{J}$. Both numbers deserve names.

The algebraic multiplicity counts an eigenvalue as a root; the geometric multiplicity counts the independent eigenvectors it supplies.

Definition 9.3.1. Algebraic Multiplicity

Let V be a finite-dimensional vector space over F with $\dim V \geq 1$, let $T \in \mathcal{L}(V)$ and $\lambda \in F$. The **algebraic multiplicity** of λ for T is

$$a_T(\lambda) \coloneqq \operatorname{mult}_\lambda(p_T),$$

the multiplicity of λ as a root of the characteristic polynomial (Definition 6.4.1). For $\mathbf{A} \in M_n(F)$, $a_{\mathbf{A}}(\lambda) \coloneqq \operatorname{mult}_\lambda(p_{\mathbf{A}})$.

Definition 9.3.2. Geometric Multiplicity

Let V be a finite-dimensional vector space over F, $T \in \mathcal{L}(V)$, and $\lambda \in F$. The **geometric multiplicity** of λ for T is

$$g_T(\lambda) \coloneqq \dim E_\lambda(T) = \dim \ker(T - \lambda \operatorname{id}_V).$$

For $\mathbf{A} \in M_n(F)$, $g_{\mathbf{A}}(\lambda) \coloneqq \dim E_\lambda(\mathbf{A}) = n - \operatorname{rank}(\mathbf{A} - \lambda\mathbf{I}_n)$.

When T or $\mathbf{A}$ is clear from context we write simply $a(\lambda)$ and $g(\lambda)$.

In words: $a_T(\lambda)$ is the **largest** m such that $(x-\lambda)^m$ divides p_T. It is a statement about a polynomial and can be computed by factoring. $g_T(\lambda)$ is the dimension of the subspace of solutions of $Tv = \lambda v$, including 0. It is the largest number of **linearly independent** eigenvectors for λ, not the number of eigenvectors, which is infinite over an infinite field. For matrices the second formula comes from Theorem 3.5.7 applied to $\mathbf{A} - \lambda\mathbf{I}_n$, and it is the one we compute with: find the rank of $\mathbf{A} - \lambda\mathbf{I}$ by elimination and subtract from n.

Well-definedness. Both numbers are defined for every $\lambda \in F$, eigenvalue or not. By Theorem 9.2.1, λ is an eigenvalue if and only if $p_T(\lambda) = 0$, if and only if $a_T(\lambda) \geq 1$; and by Definition 9.2.1, if and only if $E_\lambda(T) \neq \{0\}$, that is, $g_T(\lambda) \geq 1$. So for a scalar that is not an eigenvalue both multiplicities are 0, and for an eigenvalue both are at least 1. Neither depends on a basis: p_T does not (Definition 7.10.2), and $E_\lambda(T)$ is defined without one. If $\mathcal{B}$ is a basis of V, then $a_T(\lambda) = a_{[T]_{\mathcal{B}}^{\mathcal{B}}}(\lambda)$ because $p_T = p_{[T]_{\mathcal{B}}^{\mathcal{B}}}$, and $g_T(\lambda) = g_{[T]_{\mathcal{B}}^{\mathcal{B}}}(\lambda)$ because the coordinate map is an isomorphism from $\ker(T - \lambda \operatorname{id}_V)$ onto $\operatorname{nul}([T]_{\mathcal{B}}^{\mathcal{B}} - \lambda\mathbf{I})$ (Theorem 4.1.5 (a) and Corollary 4.1.3).

Examples.

- **The hook.** For $2\mathbf{I}_2$: $a(2) = 2$ and $g(2) = 2 - \operatorname{rank}\mathbf{O} = 2$. For $\mathbf{J}$: $a(2) = 2$ and $g(2) = 2 - \operatorname{rank}\begin{pmatrix}0 & 1\\ 0 & 0\end{pmatrix} = 1$.
- **$\mathbf{A}$ 1×1 matrix.** For $\mathbf{A} = (c)$, $p_{\mathbf{A}} = x - c$, so $a(c) = 1$, and $\mathbf{A} - c\mathbf{I} = \mathbf{O}_1$ has rank 0, so $g(c) = 1$. This degenerate case already shows the pattern of the theorem below: the two counts can only differ when there is room, a repeated root.
- **Differentiation.** On $V = \mathbb{R}[x]_{\le n}$, the matrix of D in the basis $(1, x, \dots, x^n)$ is strictly upper triangular, so $p_D = x^{n+1}$ and $a_D(0) = n + 1$. But $E_0(D)$ is the line of constant polynomials (the examples after Definition 9.2.2), so $g_D(0) = 1$. The gap can be as large as n.
- **A matrix from Chapter 7.** For $\mathbf{A} = \begin{pmatrix}2 & 0 & 1\\ 1 & 3 & 1\\ 1 & 0 & 2\end{pmatrix}$, Example (A 3×3 Characteristic Polynomial) found $p_{\mathbf{A}} = (x-1)(x-3)^2$, so $a(1) = 1$ and $a(3) = 2$. The matrix $\mathbf{A} - 3\mathbf{I} = \begin{pmatrix}-1 & 0 & 1\\ 1 & 0 & 1\\ 1 & 0 & -1\end{pmatrix}$ has a zero column, and its first two rows are independent while the third is minus the first, so its rank is 2 and $g(3) = 3 - 2 = 1$. Similarly $\mathbf{A} - \mathbf{I} = \begin{pmatrix}1 & 0 & 1\\ 1 & 2 & 1\\ 1 & 0 & 1\end{pmatrix}$ has third row equal to the first and first two rows independent, so its rank is 2 and $g(1) = 1$.

Minimal change. Take the triangular matrices

$$\mathbf{B} = \begin{pmatrix}2 & 0 & 1\\ 0 & 2 & 0\\ 0 & 0 & 1\end{pmatrix}, \qquad \mathbf{B}' = \begin{pmatrix}2 & 1 & 1\\ 0 & 2 & 0\\ 0 & 0 & 1\end{pmatrix},$$

which differ in one entry. Both have $p = (x-2)^2(x-1)$, so $a(2) = 2$ for both: the algebraic multiplicity does not notice the change. But $\mathbf{B} - 2\mathbf{I}$ has nonzero rows $(0,0,1)$ and $(0,0,-1)$, rank 1, so $g_{\mathbf{B}}(2) = 2$; while $\mathbf{B}' - 2\mathbf{I}$ has rows $(0,1,1)$ and $(0,0,-1)$, rank 2, so $g_{\mathbf{B}'}(2) = 1$. One off-diagonal entry, invisible to p, removed an eigenvector.

Why these definitions. The algebraic multiplicity is the natural count for a polynomial: with it, the roots of p_T add up to $\deg p_T = \dim V$ whenever p_T splits (Theorem 6.4.3), and the trace and determinant come out right (Theorem 9.3.3 below). The geometric multiplicity is the natural count for eigenvectors: it is the number of basis vectors we can take from $E_\lambda(T)$ when we try to build a basis of eigenvectors. The names record where each number lives, in algebra (a polynomial) or in geometry (a subspace).

Warning. **The geometric multiplicity is not read off from p_T.** Two operators with the same characteristic polynomial can have different geometric multiplicities: $2\mathbf{I}_2$ and $\mathbf{J}$ above, or $\mathbf{B}$ and $\mathbf{B}'$. To find $g(\lambda)$ you must compute $\operatorname{rank}(\mathbf{A} - \lambda\mathbf{I})$. Also, $g(\lambda)$ is **not** "how many eigenvectors" λ has: over $\mathbb{R}$, $E_\lambda \neq \{0\}$ always contains infinitely many eigenvectors.

Quick check. Find $a(0)$ and $g(0)$ for $\mathbf{N} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \in M_3(F)$.

Solution. $\mathbf{N}$ is triangular with zero diagonal, so $p_{\mathbf{N}} = x^3$ and $a(0) = 3$. The rank of $\mathbf{N} - 0\mathbf{I} = \mathbf{N}$ is 1, so $g(0) = 3 - 1 = 2$, with $E_0(\mathbf{N}) = \operatorname{span}(\mathbf{e}_1, \mathbf{e}_3)$.

9.3.2 The geometric multiplicity never exceeds the algebraic

In every example so far, $g(\lambda) \le a(\lambda)$. That is always the case.

Theorem 9.3.1. Geometric Multiplicity Is at Most Algebraic Multiplicity

Let V be a finite-dimensional vector space over F with $\dim V = n \ge 1$, let $T \in \mathcal{L}(V)$, and let $\lambda \in F$ be an eigenvalue of T. Then

$$1 \le g_T(\lambda) \le a_T(\lambda).$$

Idea. Start from the smallest subspace in sight, $E_\lambda(T)$. Take a basis of it and extend it to a basis of V. In that basis, the first g columns of the matrix are $\lambda\mathbf{e}_1, \dots, \lambda\mathbf{e}_g$, so the matrix is block upper triangular with top-left block $\lambda\mathbf{I}_g$. The block triangular determinant then factors $(x - \lambda)^g$ out of p_T, and a root with a factor $(x - \lambda)^g$ has multiplicity at least g. The inequality can be strict because the other block may contribute further factors $x - \lambda$, as it does for $\mathbf{J}$.

Proof. Since λ is an eigenvalue, $E_\lambda(T) \neq \{0\}$, so $g = g_T(\lambda) \ge 1$. Let $(v_1, \dots, v_g)$ be a basis of $E_\lambda(T)$. By Theorem 1.6.5, it extends to a basis $\mathcal{B} = (v_1, \dots, v_g, w_1, \dots, w_{n-g})$ of V.

Case $g = n$. Then $Tv_i = \lambda v_i$ for every basis vector, so $[T]_{\mathcal{B}}^{\mathcal{B}} = \lambda\mathbf{I}_n$ by Definition 4.1.1, and $p_T = (x - \lambda)^n$ by the triangular case of Definition 7.10.1. Hence $a_T(\lambda) = n = g$.

Case $g < n$. For $i \le g$, $Tv_i = \lambda v_i$, so column i of $[T]_{\mathcal{B}}^{\mathcal{B}}$ is $\lambda\mathbf{e}_i$. Hence

$$[T]_{\mathcal{B}}^{\mathcal{B}} = \begin{pmatrix} \lambda\mathbf{I}_g & \mathbf{B} \\ \mathbf{O} & \mathbf{D} \end{pmatrix}, \qquad x\mathbf{I}_n - [T]_{\mathcal{B}}^{\mathcal{B}} = \begin{pmatrix} (x - \lambda)\mathbf{I}_g & -\mathbf{B} \\ \mathbf{O} & x\mathbf{I}_{n-g} - \mathbf{D} \end{pmatrix},$$

for some $\mathbf{B} \in M_{g \times (n-g)}(F)$ and $\mathbf{D} \in M_{n-g}(F)$. The second matrix is block upper triangular over $F[x]$ with square diagonal blocks. By Theorem 7.8.4, which holds over the commutative ring $F[x]$ (the remark after it), and Definition 7.10.2,

$$p_T = \det\left((x - \lambda)\mathbf{I}_g\right) \det(x\mathbf{I}_{n-g} - \mathbf{D}) = (x - \lambda)^g \, p_{\mathbf{D}},$$

where $\det((x - \lambda)\mathbf{I}_g) = (x - \lambda)^g$ because the matrix is diagonal. So $(x - \lambda)^g \mid p_T$. By

Definition 6.4.1, $a_T(\lambda)$, the **largest** m with $(x-\lambda)^m \mid p_T$, satisfies $a_T(\lambda) \geq g$. This proves the theorem.

■

The inequality is strict exactly when λ is also a root of $p_{\mathbf{D}}$, the characteristic polynomial of the "rest" of the operator. For $\mathbf{J}$, with $\mathcal{B} = (\mathbf{e}_1, \mathbf{e}_2)$, the block $\mathbf{D}$ is (2), and $p_{\mathbf{D}} = x - 2$ contributes the second factor.

A useful consequence: if $a(\lambda) = 1$, then $g(\lambda) = 1$. A **simple** root of p_T always has a one-dimensional eigenspace, and so the only eigenvalues whose eigenspaces need checking are the repeated ones.

Both multiplicities are properties of the operator, not of the basis, so they cannot distinguish similar matrices. That makes them tools for proving that two matrices are **not** similar.

Proposition 9.3.2. Similar Matrices Have the Same Multiplicities

Let $\mathbf{A}, \mathbf{B} \in M_n(F)$ with $\mathbf{A} \sim \mathbf{B}$, and let $\lambda \in F$. Then $a_{\mathbf{A}}(\lambda) = a_{\mathbf{B}}(\lambda)$ and $g_{\mathbf{A}}(\lambda) = g_{\mathbf{B}}(\lambda)$.

Proof. Let $\mathbf{B} = \mathbf{P}^{-1}\mathbf{A}\mathbf{P}$. By Theorem 7.10.3, $p_{\mathbf{A}} = p_{\mathbf{B}}$, so the algebraic multiplicities agree. Also $\mathbf{B} - \lambda\mathbf{I} = \mathbf{P}^{-1}\mathbf{A}\mathbf{P} - \lambda\mathbf{P}^{-1}\mathbf{P} = \mathbf{P}^{-1}(\mathbf{A} - \lambda\mathbf{I})\mathbf{P}$, so $\mathbf{B} - \lambda\mathbf{I} \sim \mathbf{A} - \lambda\mathbf{I}$, and these have the same rank by Proposition 4.2.6 (a). Hence $g_{\mathbf{B}}(\lambda) = n - \operatorname{rank}(\mathbf{B} - \lambda\mathbf{I}) = n - \operatorname{rank}(\mathbf{A} - \lambda\mathbf{I}) = g_{\mathbf{A}}(\lambda)$.

■

For example, $2\mathbf{I}_2$ and $\mathbf{J}$ have the same characteristic polynomial but $g(2) = 2$ versus $g(2) = 1$, so they are not similar. The same test shows $\mathbf{B} \not\sim \mathbf{B}'$ in the minimal change above.

Warning. **A repeated eigenvalue does not force a shortage of eigenvectors, and equal eigenvalues do not force similarity.** The identity $\mathbf{I}_2$ has the repeated eigenvalue 1, yet $g(1) = a(1) = 2$ and its matrix is as simple as possible. In the other direction, $\mathbf{I}_2$ and $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ have the same eigenvalues with the same algebraic multiplicities, and they are not similar. Even agreement of **both** multiplicities for every eigenvalue does not force similarity; Exercise (C2: Multiplicities do not determine similarity) gives two 4×4 matrices showing this.

9.3.3 Multiplicities, trace, and determinant

By Theorem 6.4.3, the algebraic multiplicities of the eigenvalues of T add up to at most $\deg p_T = \dim V$, with equality exactly when p_T splits over F. Over $\mathbb{C}$ this always happens (Corollary 6.5.5). When it does, p_T is determined by the eigenvalues and their algebraic multiplicities,

$$p_T = (x - \lambda_1)^{a(\lambda_1)} \cdots (x - \lambda_k)^{a(\lambda_k)},$$

and two of its coefficients are familiar: by Theorem 7.10.2, the coefficient of x^{n-1} is $-\operatorname{tr} T$ and the constant term is $(-1)^n \det T$. Comparing the two descriptions expresses trace and determinant through eigenvalues.

Theorem 9.3.3. Trace and Determinant from Eigenvalues

Let V be a finite-dimensional vector space over F with $\dim V = n \geq 1$, and let $T \in \mathcal{L}(V)$. Suppose p_T splits over F, and let $\lambda_1, \ldots, \lambda_k$ be the distinct eigenvalues of T. Then

$$a(\lambda_1) + \cdots + a(\lambda_k) = n, \qquad \operatorname{tr} T = \sum_{i=1}^{k} a(\lambda_i)\, \lambda_i, \qquad \det T = \prod_{i=1}^{k} \lambda_i^{\,a(\lambda_i)}.$$

In words: listing each eigenvalue as often as its algebraic multiplicity, the trace is the sum of the list and the determinant is its product. In particular this holds for every operator on a nonzero finite-dimensional complex space, and for every $\mathbf{A} \in M_n(\mathbb{C})$.

Idea. Rather than expanding a product of linear factors by hand, notice that the product is itself a characteristic polynomial: that of the diagonal matrix with the eigenvalues on its diagonal. Chapter 7 already knows the x^{n-1} coefficient and the constant term of any characteristic polynomial, so we read them off twice and compare.

Proof. Since p_T splits and is monic of degree n (Theorem 7.10.2), Definition 6.4.2 gives $p_T = (x - c_1)\cdots(x - c_n)$ for some $c_1, \ldots, c_n \in F$. For $c, \lambda \in F$, the polynomial $x - c$ has $\operatorname{mult}_\lambda(x - c) = 1$ if $c = \lambda$ and 0 otherwise, since λ is a root of $x - c$ only when $c = \lambda$ (Theorem 6.1.3). So by Theorem 6.4.2, $\operatorname{mult}_\lambda(p_T) = \sum_j \operatorname{mult}_\lambda(x - c_j)$ is the number of indices j with $c_j = \lambda$. In particular the values occurring among $c_1, \ldots, c_n$ are exactly the roots of p_T in F, which are $\lambda_1, \ldots, \lambda_k$ by Theorem 9.2.1, and λ_i occurs exactly $a(\lambda_i)$ times. Counting the n factors gives $\sum_i a(\lambda_i) = n$, and

$$c_1 + \cdots + c_n = \sum_i a(\lambda_i)\lambda_i, \qquad c_1 \cdots c_n = \prod_i \lambda_i^{\,a(\lambda_i)}.$$

Let $\mathbf{C} = \operatorname{diag}(c_1, \ldots, c_n)$. It is triangular, so $p_{\mathbf{C}} = (x - c_1)\cdots(x - c_n) = p_T$ (the examples after Definition 7.10.1). Apply Theorem 7.10.2 to both $\mathbf{C}$ and T: the coefficient of x^{n-1} in this one polynomial is $-\operatorname{tr}\mathbf{C} = -(c_1 + \cdots + c_n)$ and also $-\operatorname{tr} T$; the constant term is $(-1)^n \det \mathbf{C} = (-1)^n c_1 \cdots c_n$ (Theorem 7.4.6) and also $(-1)^n \det T$. Hence $\operatorname{tr} T = c_1 + \cdots + c_n$ and $\det T = c_1 \cdots c_n$, which are the displayed formulas. The complex case follows from Corollary 6.5.5.

■

Examples.

- For $\mathbf{A} = \begin{pmatrix} 2 & 0 & 1 \\ 1 & 3 & 1 \\ 1 & 0 & 2 \end{pmatrix}$, with $a(1) = 1$ and $a(3) = 2$: $\operatorname{tr}\mathbf{A} = 2 + 3 + 2 = 7 = 1 + 3 + 3$, and $\det \mathbf{A} = 9 = 1 \cdot 3 \cdot 3$.
- For the real matrix $\begin{pmatrix} 3 & -2 \\ 4 & -1 \end{pmatrix}$ of Example (A Real Matrix with Non-Real Eigenvalues), p does not split over $\mathbb{R}$, but it splits over $\mathbb{C}$ with roots $1 \pm 2i$. The theorem over $\mathbb{C}$ gives $(1 + 2i) + (1 - 2i) = 2 = \operatorname{tr}\mathbf{A}$ and $(1 + 2i)(1 - 2i) = 1 + 4 = 5 = \det \mathbf{A}$. The trace and determinant are real, while the eigenvalues are not.
- For a 2×2 matrix whose characteristic polynomial splits, the theorem says the eigenvalues

λ, μ (listed with multiplicity) satisfy $\lambda + \mu = \operatorname{tr} \mathbf{A}$ and $\lambda\mu = \det \mathbf{A}$. This is a quick way to guess eigenvalues: for $\begin{pmatrix} 1 & 2 \\ 3 & 2 \end{pmatrix}$, we want two numbers with sum 3 and product -4, namely 4 and -1.

Without splitting, the theorem says nothing. If p_T does not split, the eigenvalues in F, counted with algebraic multiplicity, do not account for all of p_T, and summing or multiplying them gives no information. For $\mathbf{A} = \begin{pmatrix} 3 & -2 \\ 4 & -1 \end{pmatrix}$ over $\mathbb{R}$ there are no eigenvalues: the multiplicities add up to $0 \neq 2$, the empty sum 0 is not $\operatorname{tr} \mathbf{A} = 2$, and the empty product 1 is not $\det \mathbf{A} = 5$. For the real rotation $\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$, with trace 0 and determinant 1, the empty sum and product happen to give the right values, but only by coincidence. For real matrices the remedy is to pass to $\mathbb{C}$, where $p_{\mathbf{A}}$ always splits, as in the second example above.

9.3.4 Exercises

9.3.4.1 A. Check your understanding

Exercise. A1

(a) Define the algebraic and the geometric multiplicity of $\lambda \in F$ for $\mathbf{A} \in M_n(F)$.

(b) State the inequality between them for an eigenvalue λ, and name the tool its proof uses.

(c) True or false: if $a(\lambda) = 1$, then $g(\lambda) = 1$. Justify your answer.

(d) True or false: for every $\mathbf{A} \in M_n(\mathbb{C})$, the geometric multiplicities of the eigenvalues add up to n. Justify your answer.

(e) True or false: if $\mathbf{A}, \mathbf{B} \in M_n(F)$ have the same characteristic polynomial, then $g_{\mathbf{A}}(\lambda) = g_{\mathbf{B}}(\lambda)$ for every λ. Justify your answer.

(f) For $\mathbf{A} \in M_3(\mathbb{C})$ with eigenvalues 2 (algebraic multiplicity 2) and -1, what are $\operatorname{tr} \mathbf{A}$ and $\det \mathbf{A}$?

Solution.

(a) $a_{\mathbf{A}}(\lambda) = \operatorname{mult}_\lambda(p_{\mathbf{A}})$, the largest m with $(x - \lambda)^m \mid p_{\mathbf{A}}$; $g_{\mathbf{A}}(\lambda) = \dim E_\lambda(\mathbf{A}) = n - \operatorname{rank}(\mathbf{A} - \lambda \mathbf{I})$ (Definition 9.3.1, Definition 9.3.2).

(b) $1 \leq g(\lambda) \leq a(\lambda)$ (Theorem 9.3.1). The proof extends a basis of E_λ to a basis of the space and uses the block triangular determinant (Theorem 7.8.4) over $F[x]$.

(c) True. An eigenvalue has $g(\lambda) \geq 1$, and $g(\lambda) \leq a(\lambda) = 1$.

(d) False. For $\mathbf{J} = \begin{pmatrix} 2 & 1 \\ 0 & 2 \end{pmatrix}$, the only eigenvalue is 2, with $g(2) = 1 \neq 2$. (The **algebraic** multiplicities do add up to n over $\mathbb{C}$.)

(e) False. $2\mathbf{I}_2$ and $\mathbf{J}$ both have $p = (x - 2)^2$, but $g(2) = 2$ and $g(2) = 1$.

(f) By Theorem 9.3.3, listing eigenvalues with algebraic multiplicity as $2, 2, -1$: $\operatorname{tr} \mathbf{A} = 3$ and $\det \mathbf{A} = -4$.

9.3.4.2 B. Practice

Exercise. B1: Computing multiplicities

For each real matrix, find every eigenvalue together with its algebraic and geometric multiplicity.

(a) $\mathbf{A}_1 = \begin{pmatrix} 4 & -1 \\ 1 & 2 \end{pmatrix}$.

(b) $\mathbf{A}_2 = \begin{pmatrix} 2 & 1 & 1 \\ 0 & 1 & -1 \\ -1 & -1 & 1 \end{pmatrix}$.

(c) $\mathbf{A}_3 = \begin{pmatrix} 2 & 1 & 1 \\ -1 & 0 & -1 \\ 1 & 1 & 2 \end{pmatrix}$, given that $p_{\mathbf{A}_3} = p_{\mathbf{A}_2}$.

(d) The matrix of the operator $S(q)(x) = q(x+1)$ on $\mathbb{R}[x]_{\le 2}$, in the basis $(1, x, x^2)$.

Solution.

(a) $\operatorname{tr}\mathbf{A}_1 = 6$, $\det\mathbf{A}_1 = 8+1 = 9$, so $p = x^2 - 6x + 9 = (x-3)^2$: the only eigenvalue is 3, with $a(3) = 2$. $\mathbf{A}_1 - 3\mathbf{I} = \begin{pmatrix} 1 & -1 \\ 1 & -1 \end{pmatrix}$ has rank 1, so $g(3) = 1$.

(b) Expand $\det(x\mathbf{I} - \mathbf{A}_2)$ along the first column:

$$\det\begin{pmatrix} x-2 & -1 & -1 \\ 0 & x-1 & 1 \\ 1 & 1 & x-1 \end{pmatrix} = (x-2)\big[(x-1)^2 - 1\big] + 1 \cdot \det\begin{pmatrix} -1 & -1 \\ x-1 & 1 \end{pmatrix}.$$

The first bracket is $x^2 - 2x = x(x-2)$ and the last determinant is $-1 + (x-1) = x-2$, so $p_{\mathbf{A}_2} = (x-2)(x^2-2x+1) = (x-1)^2(x-2)$. Thus $a(1) = 2$ and $a(2) = 1$. By Theorem 9.3.1, $g(2) = 1$. For $\lambda = 1$,

$$\mathbf{A}_2 - \mathbf{I} = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 0 & -1 \\ -1 & -1 & 0 \end{pmatrix}.$$

Row 3 is $-(\text{row } 1) - (\text{row } 2)$, and rows 1 and 2 are independent, so the rank is 2 and $g(1) = 1$.

(c) Here $a(1) = 2$ and $a(2) = 1$, as for $\mathbf{A}_2$, so $g(2) = 1$. Now

$$\mathbf{A}_3 - \mathbf{I} = \begin{pmatrix} 1 & 1 & 1 \\ -1 & -1 & -1 \\ 1 & 1 & 1 \end{pmatrix}$$

has all rows multiples of $(1,1,1)$, so its rank is 1 and $g(1) = 2$. By Proposition 9.3.2, $\mathbf{A}_2$ and $\mathbf{A}_3$ are not similar, although they have the same characteristic polynomial.

(d) $S(1) = 1$, $S(x) = 1+x$, $S(x^2) = 1+2x+x^2$, so the matrix is $\begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{pmatrix}$ (as in Example (Characteristic Polynomials of Two Operators)). It is triangular, so $p = (x-1)^3$

and $a(1) = 3$. The matrix minus $\mathbf{I}$ is $\begin{pmatrix} 0 & 1 & 1 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \end{pmatrix}$, of rank 2, so $g(1) = 1$: the only polynomials with $q(x+1) = q(x)$ in $\mathbb{R}[x]_{\le 2}$ are the constants.

Exercise. B2: A parameter that merges eigenvalues

For $t \in \mathbb{R}$, let $\mathbf{A}_t = \begin{pmatrix} t & 1 \\ -1 & 3 \end{pmatrix}$. Find all t for which $\mathbf{A}_t$ has a real eigenvalue λ with $g(\lambda) < a(\lambda)$.

Solution. $\operatorname{tr} \mathbf{A}_t = t+3$ and $\det \mathbf{A}_t = 3t+1$, so $p(x) = x^2 - (t+3)x + (3t+1)$. For a 2×2 matrix, $g(\lambda) < a(\lambda)$ is possible only if $a(\lambda) = 2$, because $a(\lambda) = 1$ forces $g(\lambda) = 1$ (Theorem 9.3.1). So p must have a repeated real root, which happens exactly when its discriminant vanishes:

$$(t+3)^2 - 4(3t+1) = t^2 - 6t + 5 = (t-1)(t-5) = 0.$$

If $t = 1$, then $p = x^2 - 4x + 4 = (x-2)^2$ and $\mathbf{A}_1 - 2\mathbf{I} = \begin{pmatrix} -1 & 1 \\ -1 & 1 \end{pmatrix}$, of rank 1, so $g(2) = 1 < 2 = a(2)$. If $t = 5$, then $p = (x-4)^2$ and $\mathbf{A}_5 - 4\mathbf{I} = \begin{pmatrix} 1 & 1 \\ -1 & -1 \end{pmatrix}$, of rank 1, so $g(4) = 1 < 2$. Hence the answer is $t \in \{1, 5\}$. (In both cases $g = a$ was impossible anyway: $g(\lambda) = 2$ would mean $\mathbf{A}_t - \lambda \mathbf{I} = \mathbf{O}$, but the off-diagonal entries of $\mathbf{A}_t$ are nonzero.)

Exercise. B3: Eigenvalues from trace and determinant

(a) Let $\mathbf{A} \in M_3(\mathbb{R})$ have the eigenvalue $1 + i$ (as a complex matrix) and $\det \mathbf{A} = 6$. Find all eigenvalues of $\mathbf{A}$ over $\mathbb{C}$, and $\operatorname{tr} \mathbf{A}$.

(b) Let $\mathbf{B} \in M_2(\mathbb{R})$ with $\operatorname{tr} \mathbf{B} = 6$ and $\det \mathbf{B} = 9$. Show that $\operatorname{spec}(\mathbf{B}) = \{3\}$ with $a(3) = 2$. Give one such $\mathbf{B}$ with $g(3) = 2$ and one with $g(3) = 1$.

Solution.

(a) By Theorem 9.2.5, $1 - i$ is also an eigenvalue. By Corollary 6.5.5, $p_{\mathbf{A}}$ splits over $\mathbb{C}$ into three linear factors, and two of them are $x - (1+i)$ and $x - (1-i)$; let μ be the third root. By Theorem 9.3.3, $6 = \det \mathbf{A} = (1+i)(1-i)\mu = 2\mu$, so $\mu = 3$. Hence the eigenvalues are $1 + i, 1 - i, 3$, and $\operatorname{tr} \mathbf{A} = (1+i) + (1-i) + 3 = 5$.

(b) By Theorem 7.10.2, $p_{\mathbf{B}} = x^2 - 6x + 9 = (x-3)^2$. So $\operatorname{spec}(\mathbf{B}) = \{3\}$ and $a(3) = 2$. For $\mathbf{B} = 3\mathbf{I}_2$, $g(3) = 2$. For $\mathbf{B} = \begin{pmatrix} 4 & -1 \\ 1 & 2 \end{pmatrix}$ of Exercise (B1: Computing multiplicities) (a), $\operatorname{tr} \mathbf{B} = 6$, $\det \mathbf{B} = 9$, and $g(3) = 1$.

9.3.4.3 C. Going deeper

Exercise. C1: Nilpotent matrices

Let $\mathbf{A} \in M_n(\mathbb{C})$, $n \geq 1$, be nilpotent ($\mathbf{A}^k = \mathbf{O}$ for some $k \geq 1$).

(a) Prove that $a(0) = n$ and $g(0) = n - \operatorname{rank} \mathbf{A}$.

(b) Deduce that $g(0) = a(0)$ if and only if $\mathbf{A} = \mathbf{O}$.

(c) For $\mathbf{N} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$, check that $\mathbf{N}^4 = \mathbf{O}$ and compute $g(0)$. How large can $a(0) - g(0)$ be for a nilpotent $n \times n$ matrix?

Solution.

(a) By Exercise (C2: Nilpotent operators) (b), $p_{\mathbf{A}} = x^n$, so $a(0) = \operatorname{mult}_0(x^n) = n$. By Definition 9.3.2, $g(0) = n - \operatorname{rank}(\mathbf{A} - 0\mathbf{I}) = n - \operatorname{rank} \mathbf{A}$.

(b) By (a), $g(0) = a(0)$ if and only if $n - \operatorname{rank} \mathbf{A} = n$, that is, $\operatorname{rank} \mathbf{A} = 0$, that is, $\mathbf{A} = \mathbf{O}$ (a matrix of rank 0 has no nonzero column).

(c) $\mathbf{N}\mathbf{e}_1 = \mathbf{0}$ and $\mathbf{N}\mathbf{e}_j = \mathbf{e}_{j-1}$ for $j \geq 2$. So $\mathbf{N}^j \mathbf{e}_i = \mathbf{0}$ once $j \geq i$, and $\mathbf{N}^4 \mathbf{e}_i = \mathbf{0}$ for $i \leq 4$; hence $\mathbf{N}^4 = \mathbf{O}$. The columns of $\mathbf{N}$ are $\mathbf{0}, \mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$, so $\operatorname{rank} \mathbf{N} = 3$ and $g(0) = 1$, while $a(0) = 4$. In general, a nilpotent $\mathbf{A}$ has $0 \in \operatorname{spec}(\mathbf{A})$ (Exercise (C2: Nilpotent operators) (a)), so $g(0) \geq 1$ and $a(0) - g(0) \leq n - 1$; the $n \times n$ analogue of $\mathbf{N}$, with ones just above the diagonal, has rank $n - 1$ and attains $a(0) - g(0) = n - 1$.

Exercise. C2: Multiplicities do not determine similarity

Let

$$\mathbf{M}_1 = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \qquad \mathbf{M}_2 = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \in M_4(\mathbb{R}).$$

(a) Show that $\mathbf{M}_1$ and $\mathbf{M}_2$ have the same eigenvalues with the same algebraic **and** geometric multiplicities.

(b) Compute $\operatorname{rank}(\mathbf{M}_1 - \mathbf{I})^2$ and $\operatorname{rank}(\mathbf{M}_2 - \mathbf{I})^2$. Deduce that $\mathbf{M}_1 \not\sim \mathbf{M}_2$.

(c) Construct a matrix $\mathbf{M} \in M_4(\mathbb{R})$ with $\operatorname{spec}(\mathbf{M}) = \{2, 5\}$, $a(2) = 3$, $g(2) = 1$, and $a(5) = g(5) = 1$.

Hint: for (b), if $\mathbf{B} = \mathbf{P}^{-1}\mathbf{A}\mathbf{P}$, *relate* $(\mathbf{B} - \mathbf{I})^2$ *to* $(\mathbf{A} - \mathbf{I})^2$.

Solution.

(a) Both matrices are upper triangular with diagonal $1, 1, 1, 1$, so both have $p = (x - 1)^4$, spectrum $\{1\}$, and $a(1) = 4$. The matrix $\mathbf{M}_1 - \mathbf{I}$ has columns $\mathbf{0}, \mathbf{e}_1, \mathbf{0}, \mathbf{e}_3$, of rank 2; $\mathbf{M}_2 - \mathbf{I}$ has columns $\mathbf{0}, \mathbf{e}_1, \mathbf{e}_2, \mathbf{0}$, also of rank 2. So $g(1) = 4 - 2 = 2$ for both.

(b) Let $\mathbf{N}_1 = \mathbf{M}_1 - \mathbf{I}$, so $\mathbf{N}_1\mathbf{e}_2 = \mathbf{e}_1$, $\mathbf{N}_1\mathbf{e}_4 = \mathbf{e}_3$, and $\mathbf{N}_1\mathbf{e}_1 = \mathbf{N}_1\mathbf{e}_3 = \mathbf{0}$. Then $\mathbf{N}_1^2$ kills every $\mathbf{e}_j$, so $\mathbf{N}_1^2 = \mathbf{O}$ and $\operatorname{rank} \mathbf{N}_1^2 = 0$. Let $\mathbf{N}_2 = \mathbf{M}_2 - \mathbf{I}$, so $\mathbf{N}_2\mathbf{e}_2 = \mathbf{e}_1$, $\mathbf{N}_2\mathbf{e}_3 = \mathbf{e}_2$,

and $\mathbf{N}_2\mathbf{e}_1 = \mathbf{N}_2\mathbf{e}_4 = \mathbf{0}$. Then $\mathbf{N}_2^2\mathbf{e}_3 = \mathbf{e}_1 \neq \mathbf{0}$ and $\mathbf{N}_2^2$ kills $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_4$, so $\mathbf{N}_2^2 = \mathbf{E}_{13}$ has rank 1.
Suppose $\mathbf{M}_2 = \mathbf{P}^{-1}\mathbf{M}_1\mathbf{P}$. With $p(x) = (x-1)^2$, Proposition 4.2.6 (c) gives $(\mathbf{M}_2-\mathbf{I})^2 = \mathbf{P}^{-1}(\mathbf{M}_1 - \mathbf{I})^2\mathbf{P}$, and then (a) of the same proposition gives equal ranks, $1 = 0$. This contradiction shows $\mathbf{M}_1 \not\sim \mathbf{M}_2$. So eigenvalues together with both multiplicities do not determine a matrix up to similarity; Chapter 10 finds invariants that do.

(c) Take

$$\mathbf{M} = \begin{pmatrix} 2 & 1 & 0 & 0 \\ 0 & 2 & 1 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 5 \end{pmatrix}.$$

It is triangular, so $p_{\mathbf{M}} = (x-2)^3(x-5)$, $a(2) = 3$, and $a(5) = 1$, hence $g(5) = 1$ by Theorem 9.3.1. The matrix $\mathbf{M} - 2\mathbf{I}$ has columns $\mathbf{0}, \mathbf{e}_1, \mathbf{e}_2, 3\mathbf{e}_4$, which has rank 3, so $g(2) = 4 - 3 = 1$.

9.4 Diagonalization

We can now answer the question of Chapter 4 in the best case. An operator has a diagonal matrix in some basis exactly when there are enough eigenvectors to form a basis, and the previous two sections give precise ways to count "enough": by dimensions of eigenspaces, or by comparing geometric with algebraic multiplicities. This section proves that these tests agree, shows how to find the diagonalizing basis for a concrete matrix, and uses it to compute powers. The payoff is a closed formula for the Fibonacci numbers. We also meet the operators for which the method fails, and see that the field is part of the answer.

9.4.1 Diagonalizable operators

Chapter 2 introduced similarity to compare the matrices of one operator in different bases (Theorem 4.2.4), and asked for the simplest representative. Section 2 unwound the wish "$[T]_{\mathcal{B}}^{\mathcal{B}}$ is diagonal" and found that it says exactly that every vector of $\mathcal{B}$ is an eigenvector. Diagonal matrices are the matrices we can compute with most easily: they add, multiply, power, and invert entry by entry. So the operators that admit such a basis deserve a name.

An operator is diagonalizable when some basis consists of vectors that the operator only stretches.

Definition 9.4.1. Diagonalizable

Let V be a finite-dimensional vector space over F. An operator $T \in \mathcal{L}(V)$ is **diagonalizable** if there **exists** a basis $\mathcal{B}$ of V such that $[T]_{\mathcal{B}}^{\mathcal{B}}$ is a diagonal matrix.

A matrix $\mathbf{A} \in M_n(F)$ is **diagonalizable over** F if it is similar to a diagonal matrix: there **exist** an **invertible** $\mathbf{P} \in M_n(F)$ and a diagonal $\mathbf{D} \in M_n(F)$ with $\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \mathbf{D}$, equivalently $\mathbf{A} = \mathbf{P}\mathbf{D}\mathbf{P}^{-1}$.

In words: for an operator, we are free to choose **any** basis, and we ask whether **one** of them makes the matrix diagonal; most bases will not. For a matrix, the change-of-basis matrix $\mathbf{P}$ and the diagonal entries of $\mathbf{D}$ must both have entries **in** F. The two notions match: $\mathbf{A}$ is diagonalizable over F if and only if $T_{\mathbf{A}} \in \mathcal{L}(F^n)$ is diagonalizable, by Theorem 4.2.4 (c), because the matrices of $T_{\mathbf{A}}$ in the bases of F^n are exactly the matrices similar to $\mathbf{A}$. Likewise, T is diagonalizable if

and only if $[T]_{\mathcal{C}}^{\mathcal{C}}$ is diagonalizable for one, equivalently every, basis $\mathcal{C}$ of V, by Theorem 4.2.4 (a) and (b).

Examples.

- **Diagonal matrices** are diagonalizable, with $\mathbf{P} = \mathbf{I}$. So is every scalar operator $c\,\mathrm{id}_V$, whose matrix is $c\mathbf{I}$ in every basis.
- **The degenerate cases.** Every 1×1 matrix is diagonal. On $V = \{0\}$, the empty basis gives the empty matrix, which counts as diagonal. These cases are harmless, but they remind us that diagonalizability is a question about repeated eigenvalues and missing eigenvectors, which need room to occur.
- **A 2×2 matrix.** In Example (Eigenvalues and Eigenspaces of a 2×2 Matrix), $\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 3 & 2 \end{pmatrix}$ has eigenvectors $(2, 3)$ for 4 and $(1, -1)$ for -1. They form a basis of $\mathbb{R}^2$, so $\mathbf{A}$ is diagonalizable, with $\mathbf{P} = \begin{pmatrix} 2 & 1 \\ 3 & -1 \end{pmatrix}$ and $\mathbf{D} = \mathrm{diag}(4, -1)$; we check this below.
- **Projections and involutions.** A projection P satisfies $V = \operatorname{im} P \oplus \ker P$ with P the identity on $\operatorname{im} P$ and zero on $\ker P$ (Theorem 4.5.1). A basis of $\operatorname{im} P$ followed by a basis of $\ker P$ is a basis $\mathcal{B}$ with $[P]_{\mathcal{B}}^{\mathcal{B}} = \mathrm{diag}(1, \ldots, 1, 0, \ldots, 0)$. In characteristic not 2, an involution splits $V = E_+ \oplus E_-$ (Theorem 4.5.5), and gives $\mathrm{diag}(1, \ldots, 1, -1, \ldots, -1)$ in the same way.

Nonexample by minimal change. Change the $(1, 2)$-entry of $\mathbf{I}_2$ from 0 to 1, giving $\mathbf{J} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ over any field F. It is still triangular with the single eigenvalue 1, and still invertible. What fails is the existence of a basis of eigenvectors: $E_1(\mathbf{J}) = \ker \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} = \mathrm{span}(\mathbf{e}_1)$ is a line, so every eigenvector of $\mathbf{J}$ is a multiple of $\mathbf{e}_1$, and no two eigenvectors are independent. By the unwinding in Section 2, no basis $\mathcal{B}$ makes $[T_{\mathbf{J}}]_{\mathcal{B}}^{\mathcal{B}}$ diagonal. (Directly: if $\mathbf{P}^{-1}\mathbf{J}\mathbf{P} = \mathbf{D}$ were diagonal, its diagonal entries would be eigenvalues of $\mathbf{J}$, so $\mathbf{D} = \mathbf{I}$, and then $\mathbf{J} = \mathbf{P}\mathbf{I}\mathbf{P}^{-1} = \mathbf{I}$, which is false.)

Why this definition. We ask for similarity to a diagonal matrix, not equality, because the basis is ours to choose; a matrix like $\begin{pmatrix} 1 & 2 \\ 3 & 2 \end{pmatrix}$ has nothing diagonal about it until the right basis is found. We insist that $\mathbf{P}$ and $\mathbf{D}$ have entries in F, because otherwise the word would not depend on the field and would hide a real phenomenon: the rotation $\mathbf{R} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ has no real eigenvector, so it is not diagonalizable over $\mathbb{R}$, while over $\mathbb{C}$ it becomes $\mathrm{diag}(i, -i)$ (Section 2). The name says what the definition promises: the operator can be **made** diagonal.

Warning. **Diagonalizability depends on the field, and even on its characteristic.** The rotation $\mathbf{R}$ is diagonalizable over $\mathbb{C}$ but not over $\mathbb{R}$. A subtler case: $\mathbf{S} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ is diagonalizable over $\mathbb{Q}$, with eigenvectors $(1, 1)$ for 1 and $(1, -1)$ for -1. Over $\mathbb{F}_2$ the same matrix has $p_{\mathbf{S}} = x^2 - 1 = (x - 1)^2$, since $-1 = 1$, and $\mathbf{S} - \mathbf{I} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$ has rank 1, so its only eigenvectors are the multiples of $(1, 1)$, and $\mathbf{S}$ is **not** diagonalizable over $\mathbb{F}_2$. Always say over which field.

9.4.2 The Diagonalization Theorem

How do we decide whether a basis of eigenvectors exists without searching for one? Eigenvectors for different eigenvalues are automatically independent (Theorem 9.2.2), so the only question is whether each eigenspace supplies enough of them. The theorem below lists five equivalent ways to say "enough".

Theorem 9.4.1. Diagonalization Theorem

Let V be a finite-dimensional vector space over F with $\dim V = n \geq 1$, let $T \in \mathcal{L}(V)$, and let $\lambda_1, \ldots, \lambda_k$ be the distinct eigenvalues of T (possibly $k = 0$). The following are equivalent:

(a) T is diagonalizable;

(b) V has a basis consisting of eigenvectors of T;

(c) $V = E_{\lambda_1}(T) \oplus \cdots \oplus E_{\lambda_k}(T)$;

(d) $\dim E_{\lambda_1}(T) + \cdots + \dim E_{\lambda_k}(T) = n$;

(e) p_T splits over F, and $g_T(\lambda_i) = a_T(\lambda_i)$ for every $i = 1, \ldots, k$.

In that case, if $\mathcal{B}$ is a basis of eigenvectors, then $[T]_{\mathcal{B}}^{\mathcal{B}}$ is diagonal, its diagonal entries are the eigenvalues of the corresponding basis vectors, and each λ_i appears on the diagonal exactly $a_T(\lambda_i) = g_T(\lambda_i)$ times.

(For $k = 0$ the sum in (c) is $\{0\}$ and the sum in (d) is 0; all five conditions are then false, since $n \geq 1$.)

Idea. **Plan:** prove (a) $\Rightarrow$ (b) $\Rightarrow$ (c) $\Rightarrow$ (d) $\Rightarrow$ (e) $\Rightarrow$ (b) $\Rightarrow$ (a).

① (a) $\Leftrightarrow$ (b) is the unwinding of columns from Section 2. ② From a basis of eigenvectors, the eigenspaces add up to V, and Corollary 9.2.3 makes the sum direct. ③ Dimensions add in a direct sum. ④ (d) $\Rightarrow$ (e) is a sandwich: $n = \sum g \leq \sum a \leq n$, so every inequality is an equality. ⑤ (e) $\Rightarrow$ (b) runs the sandwich backwards and glues bases of the eigenspaces together.

The hard work was done in Sections 2 and 3; this proof only assembles it.

Proof.

(a) $\Leftrightarrow$ (b). Let $\mathcal{B} = (v_1, \ldots, v_n)$ be a basis of V. By Definition 4.1.1, column j of $[T]_{\mathcal{B}}^{\mathcal{B}}$ is $[Tv_j]_{\mathcal{B}}$, and this column is $d_j \mathbf{e}_j$ if and only if $Tv_j = d_j v_j$. Hence $[T]_{\mathcal{B}}^{\mathcal{B}} = \operatorname{diag}(d_1, \ldots, d_n)$ if and only if $Tv_j = d_j v_j$ for every j. Since basis vectors are nonzero, this says that every v_j is an eigenvector of T, for the eigenvalue d_j. So a basis makes $[T]_{\mathcal{B}}^{\mathcal{B}}$ diagonal exactly when it consists of eigenvectors, which also proves the final statement about the diagonal entries, apart from the count.

(b) $\Rightarrow$ (c). Let $\mathcal{B}$ be a basis of eigenvectors. Each vector of $\mathcal{B}$ lies in $E_{\lambda_i}(T)$ for some i, so $\mathcal{B} \subseteq E_{\lambda_1}(T) + \cdots + E_{\lambda_k}(T)$. This sum is a subspace (Theorem 1.7.1) containing a spanning list of V, so it equals V. By Corollary 9.2.3 the sum is direct.

(c) $\Rightarrow$ (d). By Theorem 1.7.5 ((a) $\Rightarrow$ (e)), $n = \dim V = \sum_i \dim E_{\lambda_i}(T)$.

(d) $\Rightarrow$ (e). By Theorem 9.3.1, $g_T(\lambda_i) \leq a_T(\lambda_i)$ for each i. The eigenvalues are exactly the roots of p_T in F (Theorem 9.2.1), so by Theorem 6.4.3, $\sum_i a_T(\lambda_i) \leq \deg p_T = n$

(Theorem 7.10.2). With (d),

$$n = \sum_{i=1}^{k} g_T(\lambda_i) \le \sum_{i=1}^{k} a_T(\lambda_i) \le n.$$

Hence both inequalities are equalities. The second equality, $\sum_i a_T(\lambda_i) = \deg p_T$, means that p_T splits over F (Theorem 6.4.3). The first says $\sum_i \left(a_T(\lambda_i) - g_T(\lambda_i)\right) = 0$, a sum of nonnegative integers, so every term is 0 and $g_T(\lambda_i) = a_T(\lambda_i)$ for each i.

(e) $\Rightarrow$ (b). Since p_T splits, Theorem 6.4.3 gives $\sum_i a_T(\lambda_i) = n$, and so $\sum_i g_T(\lambda_i) = n$ by (e). For each i let $\mathcal{B}_i$ be a basis of $E_{\lambda_i}(T)$, of length $g_T(\lambda_i)$, and let $\mathcal{B}$ be the list $\mathcal{B}_1, \ldots, \mathcal{B}_k$. The sum $W = E_{\lambda_1}(T) + \cdots + E_{\lambda_k}(T)$ is direct (Corollary 9.2.3), so by Theorem 1.7.5 ((a) $\Rightarrow$ (d)), $\mathcal{B}$ is a basis of W, and $\dim W = n = \dim V$. By Theorem 1.6.8, $W = V$. So $\mathcal{B}$ is a basis of V, and its vectors are eigenvectors, being nonzero vectors of eigenspaces.

Finally, suppose $\mathcal{B}$ is a basis of eigenvectors, and let m_i be the number of its vectors with eigenvalue λ_i. By (a) $\Leftrightarrow$ (b), $[T]_{\mathcal{B}}^{\mathcal{B}}$ is diagonal with λ_i appearing m_i times, so $p_T = \prod_i (x-\lambda_i)^{m_i}$ (the triangular case after Definition 7.10.1). Since the λ_i are distinct, $(x-\lambda_i)^{m_i}$ divides p_T and the remaining factors do not vanish at λ_i, so $m_i = a_T(\lambda_i)$ by Lemma 6.4.1; and $a_T(\lambda_i) = g_T(\lambda_i)$ by (e). This proves the theorem.

■

This is the theorem promised in Chapter 1, §7: **an operator is diagonalizable exactly when its eigenspaces split V as a direct sum**, and then T acts on the i-th piece as the scalar λ_i. In the language of Theorem 9.1.5, the pieces are invariant, the blocks are $\lambda_i \mathbf{I}_{g(\lambda_i)}$, and the matrix is block diagonal with scalar blocks.

Condition (e) is the practical test. It separates two different ways to fail, and they have different remedies. If p_T does not split, the field is too small, and enlarging it (from $\mathbb{R}$ to $\mathbb{C}$) may help. If p_T splits but some $g(\lambda) < a(\lambda)$, there are genuinely too few eigenvectors, and no change of field helps; Chapter 10 explains what the best matrix is then.

The easiest case of (e) needs no rank computation at all.

Corollary 9.4.2. Distinct Eigenvalues Imply Diagonalizable

Let V be a finite-dimensional vector space over F with $\dim V = n \ge 1$, and let $T \in \mathcal{L}(V)$ have n distinct eigenvalues. Then T is diagonalizable. The same holds for $\mathbf{A} \in M_n(F)$ with n distinct eigenvalues in F.

Proof. Let $\lambda_1, \ldots, \lambda_n$ be the eigenvalues. Each λ_i is an eigenvalue, so $E_{\lambda_i}(T) \ne \{0\}$ (Definition 9.2.2) and $\dim E_{\lambda_i}(T) \ge 1$, so $\sum_i \dim E_{\lambda_i}(T) \ge n$, and by Corollary 9.2.3 the sum is at most n. Hence it equals n, and T is diagonalizable by Theorem 9.4.1 ((d) $\Rightarrow$ (a)).

■

So a matrix whose characteristic polynomial has n distinct roots in F is diagonalizable over F; for example $\begin{pmatrix} 2 & 1 \\ 0 & 3 \end{pmatrix}$ is, because its eigenvalues 2 and 3 are distinct. This is the generic situation,

but it is only a sufficient condition.

Warning. **The converse fails: diagonalizable matrices may have repeated eigenvalues.** $\mathbf{I}_n$ is diagonal, and its only eigenvalue is 1. Less trivially, the matrix of Example (Diagonalizing a 3×3 Matrix with a Repeated Eigenvalue) below has eigenvalues $1, 1, 3$ and is diagonalizable. A repeated eigenvalue is a reason to **check** $g(\lambda) = a(\lambda)$, not a verdict.

Quick check. Which of $\begin{pmatrix} 3 & 1 \\ 0 & 5 \end{pmatrix}$, $\begin{pmatrix} 3 & 1 \\ 0 & 3 \end{pmatrix}$, $\begin{pmatrix} 3 & 0 \\ 0 & 3 \end{pmatrix}$ are diagonalizable over $\mathbb{R}$?

Solution. The first has distinct eigenvalues $3, 5$, so it is diagonalizable by Corollary 9.4.2. The second has $a(3) = 2$ but rank $\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} = 1$, so $g(3) = 1 < a(3)$, and it is not diagonalizable by Theorem 9.4.1 (e). The third is already diagonal.

9.4.3 Computing P and D

For a matrix $\mathbf{A} \in M_n(F)$, a basis of eigenvectors turns into the matrices $\mathbf{P}$ and $\mathbf{D}$ of Definition 9.4.1 directly. Let $(\mathbf{v}_1, \dots, \mathbf{v}_n)$ be a basis of F^n with $\mathbf{A}\mathbf{v}_j = d_j \mathbf{v}_j$, let $\mathbf{P}$ be the matrix with **columns** $\mathbf{v}_1, \dots, \mathbf{v}_n$, and $\mathbf{D} = \operatorname{diag}(d_1, \dots, d_n)$. By the column view of matrix multiplication (Theorem 0.9.2), column j of $\mathbf{AP}$ is $\mathbf{A}\mathbf{v}_j = d_j \mathbf{v}_j$, and column j of $\mathbf{PD}$ is $\mathbf{P}(d_j \mathbf{e}_j) = d_j \mathbf{v}_j$. So

$$\mathbf{AP} = \mathbf{PD}.$$

The columns of $\mathbf{P}$ form a basis, so $\mathbf{P}$ is invertible (Theorem 3.4.4), and multiplying on the left by $\mathbf{P}^{-1}$ gives $\mathbf{P}^{-1}\mathbf{AP} = \mathbf{D}$. The **order** matters in one way only: the j-th column of $\mathbf{P}$ must be an eigenvector for the j-th diagonal entry of $\mathbf{D}$. Reordering the eigenvectors reorders the diagonal of $\mathbf{D}$ in the same way, and any order works.

For $\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 3 & 2 \end{pmatrix}$, with $\mathbf{P} = \begin{pmatrix} 2 & 1 \\ 3 & -1 \end{pmatrix}$ and $\mathbf{D} = \operatorname{diag}(4, -1)$: $\mathbf{AP} = \begin{pmatrix} 8 & -1 \\ 12 & 1 \end{pmatrix}$ and $\mathbf{PD} = \begin{pmatrix} 8 & -1 \\ 12 & 1 \end{pmatrix}$. Checking $\mathbf{AP} = \mathbf{PD}$ avoids computing $\mathbf{P}^{-1}$, and together with $\det \mathbf{P} = -5 \neq 0$ it proves $\mathbf{P}^{-1}\mathbf{AP} = \mathbf{D}$.

The full procedure, then: compute $p_{\mathbf{A}}$ and its roots; if it does not split over F, stop (not diagonalizable over F); for each eigenvalue compute a basis of $E_\lambda(\mathbf{A})$; if some $g(\lambda) < a(\lambda)$, stop; otherwise put the basis vectors as columns of $\mathbf{P}$.

Example. Diagonalizing a 3×3 Matrix with a Repeated Eigenvalue

Let

$$\mathbf{A} = \begin{pmatrix} 3 & 2 & -2 \\ -2 & -1 & 2 \\ -2 & -2 & 3 \end{pmatrix} \in M_3(\mathbb{R}).$$

Decide whether $\mathbf{A}$ is diagonalizable over $\mathbb{R}$. If it is, find an invertible $\mathbf{P}$ and a diagonal $\mathbf{D}$

with $\mathbf{P}^{-1}\mathbf{AP} = \mathbf{D}$.

Solution. Characteristic polynomial. Expanding along the first row (Theorem 7.6.1 over $\mathbb{R}[x]$),

$$\begin{aligned} p_{\mathbf{A}}(x) &= \det \begin{pmatrix} x-3 & -2 & 2 \\ 2 & x+1 & -2 \\ 2 & 2 & x-3 \end{pmatrix} \\ &= (x-3)\big[(x+1)(x-3)+4\big] + 2\big[2(x-3)+4\big] \\ &\quad + 2\big[4-2(x+1)\big]. \end{aligned}$$

The brackets are $x^2-2x+1 = (x-1)^2$, $2(x-1)$, and $-2(x-1)$, so the last two terms cancel and

$$p_{\mathbf{A}}(x) = (x-1)^2(x-3).$$

It splits over $\mathbb{R}$, with $a(1) = 2$ and $a(3) = 1$. As a check, $\operatorname{tr}\mathbf{A} = 5 = 1+1+3$ (Theorem 9.3.3).

Eigenspace for 1.

$$\mathbf{A} - \mathbf{I} = \begin{pmatrix} 2 & 2 & -2 \\ -2 & -2 & 2 \\ -2 & -2 & 2 \end{pmatrix}.$$

Every row is a multiple of $(1,1,-1)$, so the rank is 1 and $g(1) = 3-1 = 2 = a(1)$. The system reduces to $x+y-z = 0$; with the free variables y and z, a basis of $E_1(\mathbf{A})$ is $(-1,1,0)$, $(1,0,1)$ (Theorem 3.5.4).

Eigenspace for 3. Here $g(3) = 1 = a(3)$ by Theorem 9.3.1, so condition (e) of Theorem 9.4.1 holds and $\mathbf{A}$ is diagonalizable. To find the eigenvector,

$$\mathbf{A} - 3\mathbf{I} = \begin{pmatrix} 0 & 2 & -2 \\ -2 & -4 & 2 \\ -2 & -2 & 0 \end{pmatrix}.$$

Row 1 gives $y = z$, and row 3 gives $x = -y$; row 2 then reads $2y - 4y + 2y = 0$. So $E_3(\mathbf{A}) = \operatorname{span}((1,-1,-1))$.

The matrices. Put the eigenvectors as columns, in the order of the eigenvalues we want on the diagonal:

$$\mathbf{P} = \begin{pmatrix} -1 & 1 & 1 \\ 1 & 0 & -1 \\ 0 & 1 & -1 \end{pmatrix}, \qquad \mathbf{D} = \operatorname{diag}(1,1,3).$$

Check. $\mathbf{AP}$ has columns $\mathbf{A}(-1,1,0) = (-1,1,0)$, $\mathbf{A}(1,0,1) = (1,0,1)$, and $\mathbf{A}(1,-1,-1) = (3,-3,-3)$, which are the columns of $\mathbf{PD}$. Expanding along the first column, $\det\mathbf{P} = -1\cdot(0+1) - 1\cdot(-1-1) = 1 \neq 0$, so $\mathbf{P}$ is invertible, and in fact

$$\mathbf{P}^{-1} = \begin{pmatrix} 1 & 2 & -1 \\ 1 & 1 & 0 \\ 1 & 1 & -1 \end{pmatrix},$$

as multiplying out $\mathbf{PP}^{-1} = \mathbf{I}$ confirms. Hence $\mathbf{P}^{-1}\mathbf{AP} = \mathbf{D}$.

Had we listed the eigenvectors in the order $(1,-1,-1),(-1,1,0),(1,0,1)$, we would get $\mathbf{D} = \operatorname{diag}(3,1,1)$ instead. Both are correct.

9.4.4 Powers and polynomials

Why diagonalize? Powers of diagonal matrices are free: $\operatorname{diag}(d_1,\dots,d_n)^k = \operatorname{diag}(d_1^k,\dots,d_n^k)$. Similarity carries this over.

Theorem 9.4.3. Powers and Polynomials of a Diagonalizable Matrix

Let $\mathbf{A} = \mathbf{P}\mathbf{D}\mathbf{P}^{-1}$ with $\mathbf{P} \in M_n(F)$ invertible and $\mathbf{D} = \operatorname{diag}(d_1,\dots,d_n)$. Then for every $k \in \mathbb{N}$ and every $p \in F[x]$,

$$\mathbf{A}^k = \mathbf{P}\,\operatorname{diag}(d_1^k,\dots,d_n^k)\,\mathbf{P}^{-1}, \qquad p(\mathbf{A}) = \mathbf{P}\,\operatorname{diag}\big(p(d_1),\dots,p(d_n)\big)\,\mathbf{P}^{-1}.$$

If moreover every $d_i \neq 0$, then $\mathbf{A}$ is invertible and $\mathbf{A}^{-1} = \mathbf{P}\,\operatorname{diag}(d_1^{-1},\dots,d_n^{-1})\,\mathbf{P}^{-1}$.

Proof. Since $\mathbf{D} = \mathbf{P}^{-1}\mathbf{A}\mathbf{P}$, Proposition 4.2.6 (c) gives $p(\mathbf{D}) = \mathbf{P}^{-1}p(\mathbf{A})\mathbf{P}$, so $p(\mathbf{A}) = \mathbf{P}\,p(\mathbf{D})\,\mathbf{P}^{-1}$. A diagonal matrix is the direct sum $(d_1)\oplus\cdots\oplus(d_n)$ of 1×1 blocks, so by Theorem 8.1.4 (b), extended to n blocks by induction on the number of blocks, $p(\mathbf{D}) = (p(d_1))\oplus\cdots\oplus(p(d_n)) = \operatorname{diag}(p(d_1),\dots,p(d_n))$. Taking $p = x^k$ gives the formula for $\mathbf{A}^k$. If every $d_i \neq 0$, put $\mathbf{B} = \mathbf{P}\,\operatorname{diag}(d_1^{-1},\dots,d_n^{-1})\,\mathbf{P}^{-1}$; then $\mathbf{A}\mathbf{B} = \mathbf{P}\mathbf{D}\mathbf{P}^{-1}\mathbf{P}\,\operatorname{diag}(d_i^{-1})\,\mathbf{P}^{-1} = \mathbf{P}\,\operatorname{diag}(d_i d_i^{-1})\,\mathbf{P}^{-1} = \mathbf{P}\mathbf{P}^{-1} = \mathbf{I}$, and $\mathbf{B} = \mathbf{A}^{-1}$ by Theorem 3.4.5. ∎

In words: to apply a polynomial to a diagonalizable matrix, apply it to the eigenvalues and change basis back. The work is done once, when $\mathbf{P}$ is found; every power after that costs a single product of three matrices. For Example (Diagonalizing a 3×3 Matrix with a Repeated Eigenvalue),

$$\mathbf{A}^k = \mathbf{P}\,\operatorname{diag}(1,1,3^k)\,\mathbf{P}^{-1} = \begin{pmatrix} 3^k & 3^k-1 & 1-3^k \\ 1-3^k & 2-3^k & 3^k-1 \\ 1-3^k & 1-3^k & 3^k \end{pmatrix},$$

which is $\mathbf{I}$ for $k=0$ and $\mathbf{A}$ for $k=1$, as it must be.

The most famous application is to a recurrence. A linear recurrence of order two becomes a first-order recurrence for vectors, whose solution is a matrix power.

Example. A Closed Formula for the Fibonacci Numbers

The Fibonacci numbers are defined by $F_0 = 0$, $F_1 = 1$, and $F_{k+2} = F_{k+1} + F_k$ for $k \geq 0$. (Here F_k is a number, not a field; we work over $\mathbb{R}$.) Find a formula for F_k by diagonalizing a 2×2 matrix.

Solution. A vector recurrence. Put $\mathbf{x}_k = (F_{k+1}, F_k) \in \mathbb{R}^2$. The recurrence says

$$\mathbf{x}_{k+1} = \begin{pmatrix} F_{k+2} \\ F_{k+1} \end{pmatrix} = \begin{pmatrix} F_{k+1} + F_k \\ F_{k+1} \end{pmatrix} = \mathbf{A}\mathbf{x}_k, \qquad \mathbf{A} = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}.$$

By induction on k, $\mathbf{x}_k = \mathbf{A}^k \mathbf{x}_0$ with $\mathbf{x}_0 = (1, 0)$.

Eigenvalues. $\operatorname{tr} \mathbf{A} = 1$ and $\det \mathbf{A} = -1$, so $p_{\mathbf{A}}(x) = x^2 - x - 1$, with the two real roots

$$\varphi = \frac{1 + \sqrt{5}}{2}, \qquad \psi = \frac{1 - \sqrt{5}}{2}.$$

They are distinct, so $\mathbf{A}$ is diagonalizable over $\mathbb{R}$ (Corollary 9.4.2). Each root satisfies $\lambda^2 = \lambda + 1$, and we will use $\varphi + \psi = 1$, $\varphi\psi = -1$, and $\varphi - \psi = \sqrt{5}$.

Eigenvectors. For a root λ, the vector $(\lambda, 1)$ is an eigenvector: $\mathbf{A}(\lambda, 1) = (\lambda + 1, \lambda) = (\lambda^2, \lambda) = \lambda(\lambda, 1)$. So

$$\mathbf{P} = \begin{pmatrix} \varphi & \psi \\ 1 & 1 \end{pmatrix}, \qquad \mathbf{D} = \operatorname{diag}(\varphi, \psi), \qquad \mathbf{P}^{-1} = \frac{1}{\sqrt{5}} \begin{pmatrix} 1 & -\psi \\ -1 & \varphi \end{pmatrix},$$

where $\det \mathbf{P} = \varphi - \psi = \sqrt{5} \neq 0$ and $\mathbf{P}^{-1}$ comes from Theorem 0.9.8.

The power. By Theorem 9.4.3, $\mathbf{x}_k = \mathbf{P}\mathbf{D}^k\mathbf{P}^{-1}\mathbf{x}_0$. Working from the right:

$$\mathbf{P}^{-1}\mathbf{x}_0 = \frac{1}{\sqrt{5}} \begin{pmatrix} 1 \\ -1 \end{pmatrix},$$

$$\mathbf{D}^k\mathbf{P}^{-1}\mathbf{x}_0 = \frac{1}{\sqrt{5}} \begin{pmatrix} \varphi^k \\ -\psi^k \end{pmatrix},$$

$$\mathbf{P}\mathbf{D}^k\mathbf{P}^{-1}\mathbf{x}_0 = \frac{1}{\sqrt{5}} \begin{pmatrix} \varphi^{k+1} - \psi^{k+1} \\ \varphi^k - \psi^k \end{pmatrix}.$$

The second entry of $\mathbf{x}_k$ is F_k, so

$$F_k = \frac{\varphi^k - \psi^k}{\sqrt{5}} = \frac{1}{\sqrt{5}} \left[\left(\frac{1 + \sqrt{5}}{2} \right)^k - \left(\frac{1 - \sqrt{5}}{2} \right)^k \right].$$

Check. $k = 0$: 0. $k = 1$: $(\varphi - \psi)/\sqrt{5} = 1$. $k = 2$: $(\varphi^2 - \psi^2)/\sqrt{5} = (\varphi - \psi)(\varphi + \psi)/\sqrt{5} = 1$. And the right side satisfies the recurrence, because $\varphi^{k+2} = \varphi^{k+1} + \varphi^k$ (multiply $\varphi^2 = \varphi + 1$ by φ^k), and likewise for ψ.

Since $|\psi| < 1$, the term $\psi^k/\sqrt{5}$ has absolute value less than $1/2$, so F_k is the integer nearest to $\varphi^k/\sqrt{5}$; the Fibonacci numbers grow like powers of $\varphi \approx 1.618$, the larger eigenvalue.

The irrational numbers in the formula are forced. Over $\mathbb{Q}$, the polynomial $x^2 - x - 1$ has no root (by Theorem 6.3.6, the only candidates are ± 1, and neither works), so $\mathbf{A}$ has no eigenvalue in $\mathbb{Q}$ and is **not** diagonalizable over $\mathbb{Q}$, although all its entries and powers are integers. Chapter 7 met the same numbers as determinants of tridiagonal matrices (Example (Fibonacci Numbers as Determinants)); here the recurrence is solved rather than just recognized. Section 11 of this chapter treats general linear recurrences and Markov chains the same way.

9.4.5 When diagonalization fails

Two examples collect the ways Theorem 9.4.1 (e) can fail.

Too few eigenvectors. $\mathbf{J} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ is not diagonalizable over **any** field, because $p_{\mathbf{J}} = (x-1)^2$ splits but $g(1) = 1 < 2 = a(1)$. Enlarging the field changes nothing: the rank of $\mathbf{J} - \mathbf{I}$ is 1 over every field. Its powers are still easy, $\mathbf{J}^k = \begin{pmatrix} 1 & k \\ 0 & 1 \end{pmatrix}$ by induction, but they are not of the form $\mathbf{P}\,\mathrm{diag}(1,1)\,\mathbf{P}^{-1} = \mathbf{I}$; the entry k, growing linearly rather than like a power, is the footprint of the missing eigenvector. Chapter 10 shows that such blocks are the only additional ingredient.

Too small a field. The rotation $\mathbf{R} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ has $p_{\mathbf{R}} = x^2 + 1$, which does not split over $\mathbb{R}$, so $\mathbf{R}$ is not diagonalizable over $\mathbb{R}$. Over $\mathbb{C}$ it has the distinct eigenvalues $\pm i$ with eigenvectors $(1, -i)$ and $(1, i)$, so with $\mathbf{P} = \begin{pmatrix} 1 & 1 \\ -i & i \end{pmatrix}$,

$$\mathbf{P}^{-1}\mathbf{R}\mathbf{P} = \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix}.$$

Then $\mathbf{R}^k = \mathbf{P}\,\mathrm{diag}(i^k, (-i)^k)\,\mathbf{P}^{-1}$, and since $i^4 = 1$, $\mathbf{R}^4 = \mathbf{I}$: four quarter-turns make a full turn, as they should.

When a matrix depends on a parameter, the same test decides diagonalizability for every value at once.

Example. Diagonalizability with a Parameter

For which $t \in \mathbb{R}$ is $\mathbf{A}_t = \begin{pmatrix} 2 & t & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{pmatrix}$ diagonalizable over $\mathbb{R}$?

Solution. The matrix is upper triangular, so $p_{\mathbf{A}_t}(x) = (x-2)^2(x-3)$ for every t, by Theorem 7.4.6. This splits, with $a(2) = 2$ and $a(3) = 1$. Since $1 \le g(3) \le a(3) = 1$ by Theorem 9.3.1, only $\lambda = 2$ needs checking. Here

$$\mathbf{A}_t - 2\mathbf{I} = \begin{pmatrix} 0 & t & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

whose rank is 1 if $t = 0$ and 2 if $t \neq 0$. By rank–nullity for matrices (Theorem 3.5.7), $g(2) = 3 - \mathrm{rank}(\mathbf{A}_t - 2\mathbf{I})$ is 2 when $t = 0$ and 1 when $t \neq 0$. By Theorem 9.4.1 (e), $\mathbf{A}_t$ is diagonalizable exactly when $t = 0$.

Notice what did not change: the eigenvalues and their algebraic multiplicities are the same for every t. Only the geometric multiplicity, a rank, detects the difference.

9.4.6 Exercises

9.4.6.1 A. Check your understanding

Exercise. A1

(a) Define what it means for $\mathbf{A} \in M_n(F)$ to be diagonalizable over F.

(b) State the Diagonalization Theorem, with at least four equivalent conditions.

(c) True or false: a diagonalizable $n \times n$ matrix has n distinct eigenvalues. Justify your answer.

(d) True or false: every invertible matrix is diagonalizable. Justify your answer.

(e) True or false: every diagonalizable matrix is invertible. Justify your answer.

(f) If $\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \mathbf{D}$ is diagonal, what are the columns of $\mathbf{P}$, and how are they matched with the diagonal of $\mathbf{D}$?

Solution.

(a) There are an invertible $\mathbf{P} \in M_n(F)$ and a diagonal $\mathbf{D} \in M_n(F)$ with $\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \mathbf{D}$ (Definition 9.4.1).

(b) See Theorem 9.4.1: for $T \in \mathcal{L}(V)$, $\dim V = n \geq 1$, with distinct eigenvalues $\lambda_1, \ldots, \lambda_k$, the following are equivalent: T is diagonalizable; V has a basis of eigenvectors; $V = \bigoplus_i E_{\lambda_i}(T)$; $\sum_i \dim E_{\lambda_i}(T) = n$; p_T splits and $g(\lambda_i) = a(\lambda_i)$ for all i.

(c) False. $\mathbf{I}_n$ is diagonal, with the single eigenvalue 1.

(d) False. $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ has determinant 1 but is not diagonalizable.

(e) False. The zero matrix is diagonal and not invertible.

(f) The columns of $\mathbf{P}$ form a basis of eigenvectors of $\mathbf{A}$, and column j is an eigenvector for the j-th diagonal entry of $\mathbf{D}$, since $\mathbf{A}\mathbf{P} = \mathbf{P}\mathbf{D}$.

9.4.6.2 B. Practice

Exercise. B1: Diagonalize or show impossible

For each real matrix, either find an invertible $\mathbf{P}$ and a diagonal $\mathbf{D}$ with $\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \mathbf{D}$, or show that none exist.

(a) $\mathbf{A}_1 = \begin{pmatrix} 0 & 1 \\ -6 & 5 \end{pmatrix}$.

(b) $\mathbf{A}_2 = \begin{pmatrix} 1 & 1 & 1 \\ 2 & 0 & 1 \\ -3 & 3 & 2 \end{pmatrix}$.

(c) $\mathbf{A}_3 = \begin{pmatrix} -1 & 2 & -2 \\ -2 & 3 & -2 \\ 2 & -2 & 3 \end{pmatrix}$.

Solution.

(a) $\operatorname{tr}\mathbf{A}_1 = 5$, $\det\mathbf{A}_1 = 6$, so $p = x^2 - 5x + 6 = (x-2)(x-3)$. The eigenvalues are distinct, so $\mathbf{A}_1$ is diagonalizable (Corollary 9.4.2). $\mathbf{A}_1 - 2\mathbf{I} = \begin{pmatrix} -2 & 1 \\ -6 & 3 \end{pmatrix}$ gives $y = 2x$, eigenvector $(1,2)$; $\mathbf{A}_1 - 3\mathbf{I} = \begin{pmatrix} -3 & 1 \\ -6 & 2 \end{pmatrix}$ gives $y = 3x$, eigenvector $(1,3)$. So

$$\mathbf{P} = \begin{pmatrix} 1 & 1 \\ 2 & 3 \end{pmatrix}, \quad \mathbf{D} = \operatorname{diag}(2,3), \quad \mathbf{P}^{-1} = \begin{pmatrix} 3 & -1 \\ -2 & 1 \end{pmatrix},$$

and indeed $\mathbf{A}_1\mathbf{P} = \begin{pmatrix} 2 & 3 \\ 4 & 9 \end{pmatrix} = \mathbf{PD}$.

(b) Expanding along the first row,

$$\begin{aligned} p(x) &= \det \begin{pmatrix} x-1 & -1 & -1 \\ -2 & x & -1 \\ 3 & -3 & x-2 \end{pmatrix} \\ &= (x-1)\big[x(x-2) - 3\big] + \big[-2(x-2) + 3\big] - \big[6 - 3x\big]. \end{aligned}$$

The first bracket is $(x-3)(x+1)$, and the last two terms add up to $-2x+7-6+3x = x+1$. So $p(x) = (x+1)\big[(x-1)(x-3)+1\big] = (x+1)(x-2)^2$, and $a(2) = 2$. Now

$$\mathbf{A}_2 - 2\mathbf{I} = \begin{pmatrix} -1 & 1 & 1 \\ 2 & -2 & 1 \\ -3 & 3 & 0 \end{pmatrix}.$$

Row 3 equals row 1 minus row 2, and rows 1 and 2 are not proportional, so the rank is 2 and $g(2) = 1 < 2 = a(2)$. By Theorem 9.4.1 (e), $\mathbf{A}_2$ is **not** diagonalizable, over $\mathbb{R}$ or over $\mathbb{C}$.

(c) Expanding along the first row,

$$\begin{aligned} p(x) &= \det \begin{pmatrix} x+1 & -2 & 2 \\ 2 & x-3 & 2 \\ -2 & 2 & x-3 \end{pmatrix} \\ &= (x+1)\big[(x-3)^2 - 4\big] + 2\big[2(x-3) + 4\big] \\ &\quad + 2\big[4 + 2(x-3)\big]. \end{aligned}$$

The first bracket is $(x-5)(x-1)$ and the other two are $2(x-1)$ each, so $p(x) = (x-1)\big[(x+1)(x-5)+8\big] = (x-1)(x^2-4x+3) = (x-1)^2(x-3)$. For $\lambda = 1$, $\mathbf{A}_3 - \mathbf{I} = \begin{pmatrix} -2 & 2 & -2 \\ -2 & 2 & -2 \\ 2 & -2 & 2 \end{pmatrix}$ has rank 1, so $g(1) = 2 = a(1)$, and E_1 is the plane $x - y + z = 0$, with basis $(1,1,0), (-1,0,1)$. Also $g(3) = 1 = a(3)$, so $\mathbf{A}_3$ is diagonalizable. From $\mathbf{A}_3 - 3\mathbf{I} = \begin{pmatrix} -4 & 2 & -2 \\ -2 & 0 & -2 \\ 2 & -2 & 0 \end{pmatrix}$, row 3 gives $y = x$ and row 2 gives $z = -x$, so

$E_3 = \operatorname{span}((1,1,-1))$. Hence

$$\mathbf{P} = \begin{pmatrix} 1 & -1 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & -1 \end{pmatrix}, \qquad \mathbf{D} = \operatorname{diag}(1,1,3).$$

Check: $\mathbf{A}_3(1,1,0) = (1,1,0)$, $\mathbf{A}_3(-1,0,1) = (-1,0,1)$, $\mathbf{A}_3(1,1,-1) = (3,3,-3)$, so $\mathbf{A}_3\mathbf{P} = \mathbf{P}\mathbf{D}$; and $\det \mathbf{P} = 1(0-1) + 1(-1-0) + 1(1-0) = -1 \neq 0$.

Exercise. B2: A tenth power

Let $\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix} \in M_2(\mathbb{R})$.

(a) Diagonalize $\mathbf{A}$.

(b) Hence find a formula for $\mathbf{A}^k$, $k \in \mathbb{N}$, and compute $\mathbf{A}^{10}$.

Solution.

(a) $\operatorname{tr} \mathbf{A} = 2$, $\det \mathbf{A} = -3$, so $p = x^2 - 2x - 3 = (x-3)(x+1)$. $\mathbf{A} - 3\mathbf{I} = \begin{pmatrix} -2 & 2 \\ 2 & -2 \end{pmatrix}$ gives the eigenvector $(1,1)$, and $\mathbf{A} + \mathbf{I} = \begin{pmatrix} 2 & 2 \\ 2 & 2 \end{pmatrix}$ gives $(1,-1)$. So $\mathbf{P} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$, $\mathbf{D} = \operatorname{diag}(3,-1)$, and $\mathbf{P}^{-1} = \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$ by Theorem 0.9.8.

(b) By Theorem 9.4.3,

$$\mathbf{A}^k = \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}\begin{pmatrix} 3^k & 0 \\ 0 & (-1)^k \end{pmatrix}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} = \frac{1}{2}\begin{pmatrix} 3^k + (-1)^k & 3^k - (-1)^k \\ 3^k - (-1)^k & 3^k + (-1)^k \end{pmatrix}.$$

For $k = 1$ this gives $\frac{1}{2}\begin{pmatrix} 2 & 4 \\ 4 & 2 \end{pmatrix} = \mathbf{A}$. With $3^{10} = 59049$ and $(-1)^{10} = 1$,

$$\mathbf{A}^{10} = \begin{pmatrix} 29525 & 29524 \\ 29524 & 29525 \end{pmatrix}.$$

Exercise. B3: A matrix with a parameter

For $t \in \mathbb{R}$, let $\mathbf{A}_t = \begin{pmatrix} 1 & 1 & 0 \\ 0 & t & 0 \\ 0 & 0 & 2 \end{pmatrix}$. Determine all t for which $\mathbf{A}_t$ is diagonalizable over $\mathbb{R}$. Justify your answer.

Solution. $\mathbf{A}_t$ is upper triangular, so $p(x) = (x-1)(x-t)(x-2)$, which splits over $\mathbb{R}$.

Case $t \notin \{1,2\}$. The eigenvalues $1, t, 2$ are distinct, so $\mathbf{A}_t$ is diagonalizable (Corollary 9.4.2).

Case $t = 1$. Then $a(1) = 2$, and $\mathbf{A}_1 - \mathbf{I} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ has rank 2, so $g(1) = 1 < 2$. Not diagonalizable.

Case $t = 2$. Then $a(2) = 2$, and $\mathbf{A}_2 - 2\mathbf{I} = \begin{pmatrix} -1 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$ has rank 1, so $g(2) = 2 = a(2)$; also $g(1) = a(1) = 1$. Diagonalizable, with eigenvectors $\mathbf{e}_1$ for 1 and $(1, 1, 0), \mathbf{e}_3$ for 2.

Hence $\mathbf{A}_t$ is diagonalizable over $\mathbb{R}$ if and only if $t \neq 1$. In both exceptional cases a diagonal entry is repeated, but only for $t = 1$ is an eigenvector lost: the entry 1 in position $(1, 2)$ couples $\mathbf{e}_1$ and $\mathbf{e}_2$, and it matters only when their diagonal entries agree.

9.4.6.3 C. Going deeper

Exercise. C1: Matrices with $\mathbf{A}^k = \mathbf{I}$

Let $\mathbf{A} \in M_n(\mathbb{C})$, $n \geq 1$, and $k \geq 1$.

(a) Prove that if $\mathbf{A}^k = \mathbf{I}$, then every eigenvalue λ of $\mathbf{A}$ satisfies $\lambda^k = 1$.

(b) Prove that if $\mathbf{A}$ is diagonalizable and every eigenvalue satisfies $\lambda^k = 1$, then $\mathbf{A}^k = \mathbf{I}$.

(c) Show that the hypothesis "diagonalizable" in (b) cannot be dropped.

Solution.

(a) Let $\mathbf{Av} = \lambda\mathbf{v}$ with $\mathbf{v} \neq \mathbf{0}$. By Exercise (B3: Eigenvalues of powers and inverses) (a), $\mathbf{v} = \mathbf{Iv} = \mathbf{A}^k\mathbf{v} = \lambda^k\mathbf{v}$, so $(\lambda^k - 1)\mathbf{v} = \mathbf{0}$ and $\lambda^k = 1$ by Theorem 1.1.9.

(b) Write $\mathbf{A} = \mathbf{PDP}^{-1}$ with $\mathbf{D} = \operatorname{diag}(d_1, \ldots, d_n)$ (Definition 9.4.1). The d_i are eigenvalues of $\mathbf{A}$: $\mathbf{D}$ and $\mathbf{A}$ are similar, so $p_{\mathbf{D}} = p_{\mathbf{A}}$ (Theorem 7.10.3), and each d_i is a root of $p_{\mathbf{D}} = \prod_j (x - d_j)$. So $d_i^k = 1$ for all i, and by Theorem 9.4.3, $\mathbf{A}^k = \mathbf{P}\operatorname{diag}(d_1^k, \ldots, d_n^k)\,\mathbf{P}^{-1} = \mathbf{PIP}^{-1} = \mathbf{I}$.

(c) Let $\mathbf{J} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$. Its only eigenvalue is 1, and $1^k = 1$. But $\mathbf{J}^k = \begin{pmatrix} 1 & k \\ 0 & 1 \end{pmatrix}$ by induction ($\mathbf{J}^{k+1} = \mathbf{J}^k\mathbf{J}$), which is not $\mathbf{I}$ since $k \neq 0$ in $\mathbb{C}$. (Section 8 proves the converse of (b) in a strong form: over $\mathbb{C}$, $\mathbf{A}^k = \mathbf{I}$ forces $\mathbf{A}$ to be diagonalizable.)

Exercise. C2: Polynomials of diagonalizable matrices

(a) Prove that if $\mathbf{A} \in M_n(F)$ is diagonalizable over F, then so is $p(\mathbf{A})$ for every $p \in F[x]$, and that every eigenvalue of $p(\mathbf{A})$ has the form $p(\lambda)$ with $\lambda \in \operatorname{spec}(\mathbf{A})$.

(b) Show that the converse of the first statement fails: find a nondiagonalizable $\mathbf{A}$ and a polynomial p with $p(\mathbf{A})$ diagonalizable.

Solution.

(a) Let $\mathbf{A} = \mathbf{PDP}^{-1}$, $\mathbf{D} = \operatorname{diag}(d_1, \dots, d_n)$. By Theorem 9.4.3,

$$p(\mathbf{A}) = \mathbf{P} \operatorname{diag}(p(d_1), \dots, p(d_n)) \mathbf{P}^{-1},$$

so $p(\mathbf{A})$ is similar to a diagonal matrix, that is, diagonalizable. By Theorem 7.10.3, $p_{p(\mathbf{A})} = \prod_i (x - p(d_i))$, so by Theorem 9.2.1 every eigenvalue of $p(\mathbf{A})$ is some $p(d_i)$. As in Exercise (C1: Matrices with $\mathbf{A}^k = \mathbf{I}$) (b), each d_i is an eigenvalue of $\mathbf{A}$.

(b) Let $\mathbf{A} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$ and $p = x^2$. Then $\mathbf{A}$ is not diagonalizable: $p_{\mathbf{A}} = x^2$ gives $a(0) = 2$, while $\operatorname{rank} \mathbf{A} = 1$ gives $g(0) = 1$ (Theorem 9.4.1 (e)). But $p(\mathbf{A}) = \mathbf{A}^2 = \mathbf{O}$ is diagonal.

Exercise. C3: The transpose map on $M_n(F)$

Let $n \geq 1$ and $\tau \colon M_n(F) \to M_n(F)$, $\tau(\mathbf{X}) = \mathbf{X}^\top$.

(a) Suppose the characteristic of F is not 2. Prove that τ is diagonalizable, and find $\dim E_1(\tau)$ and $\dim E_{-1}(\tau)$.

(b) Under the same hypothesis, deduce p_τ, $\operatorname{tr} \tau$, and $\det \tau$.

(c) Now let $F = \mathbb{F}_2$ and $n \geq 2$. Prove that τ is **not** diagonalizable.

Hint: for the dimensions, use the matrices $\mathbf{E}_{ij} + \mathbf{E}_{ji}$ and $\mathbf{E}_{ij} - \mathbf{E}_{ji}$.

Solution.

(a) Since $(\mathbf{X}^\top)^\top = \mathbf{X}$ (Theorem 0.9.5), $\tau^2 = \mathrm{id}$. By Theorem 4.5.5, $M_n(F) = E_+ \oplus E_-$, where $E_+ = E_1(\tau)$ is the space of symmetric matrices and $E_- = E_{-1}(\tau)$ the space of skew-symmetric matrices.
A symmetric $\mathbf{X} = (x_{ij})$ is determined by its entries on and above the diagonal, and $\mathbf{X} = \sum_i x_{ii}\mathbf{E}_{ii} + \sum_{i<j} x_{ij}(\mathbf{E}_{ij} + \mathbf{E}_{ji})$. These $n + \binom{n}{2}$ matrices are symmetric and independent, since their nonzero entries sit in disjoint positions. So $\dim E_1(\tau) = n(n+1)/2$. A skew-symmetric $\mathbf{X}$ has $x_{ii} = -x_{ii}$, so $2x_{ii} = 0$ and $x_{ii} = 0$ (as $2 \neq 0$); and $\mathbf{X} = \sum_{i<j} x_{ij}(\mathbf{E}_{ij} - \mathbf{E}_{ji})$, with these $\binom{n}{2}$ matrices independent for the same reason. So $\dim E_{-1}(\tau) = n(n-1)/2$.
The eigenvalues of τ lie among ± 1 (as in Exercise (B2: The transpose map)), and $n(n+1)/2 + n(n-1)/2 = n^2 = \dim M_n(F)$. If $n \geq 2$, both eigenspaces are nonzero, and the dimensions of the eigenspaces add up to n^2; if $n = 1$, $\tau = \mathrm{id}$ and $E_1(\tau)$ is everything. Either way Theorem 9.4.1 ((d) $\Rightarrow$ (a)) shows that τ is diagonalizable.

(b) In a basis $\mathcal{B}$ of eigenvectors, $[\tau]_{\mathcal{B}}^{\mathcal{B}} = \operatorname{diag}(1, \dots, 1, -1, \dots, -1)$ with $n(n+1)/2$ ones and $n(n-1)/2$ entries -1 (Theorem 9.4.1). This matrix is triangular, so

$$\begin{aligned} p_\tau &= (x-1)^{n(n+1)/2}(x+1)^{n(n-1)/2}, \\ \operatorname{tr} \tau &= \frac{n(n+1)}{2} - \frac{n(n-1)}{2} = n, \\ \det \tau &= (-1)^{n(n-1)/2}. \end{aligned}$$

As a check on the trace: in the basis $\mathcal{E} = (\mathbf{E}_{ij})$, $\tau(\mathbf{E}_{ij}) = \mathbf{E}_{ji}$, so the diagonal entry of $[\tau]_{\mathcal{E}}^{\mathcal{E}}$ at $\mathbf{E}_{ij}$ is 1 if $i = j$ and 0 otherwise, and the trace is n.

(c) Over $\mathbb{F}_2$, $-1 = 1$. If $\tau(\mathbf{X}) = \lambda \mathbf{X}$ with $\mathbf{X} \neq \mathbf{O}$, then as in Exercise (B2: The transpose

map) (a), $\lambda^2 = 1$, so $(\lambda - 1)^2 = \lambda^2 - 2\lambda + 1 = \lambda^2 + 1 = 0$ and $\lambda = 1$. So $\operatorname{spec}(\tau) = \{1\}$ and $E_1(\tau)$ is the space of symmetric matrices. The spanning argument in (a) did not divide by 2, so $\dim E_1(\tau) = n(n+1)/2$. For $n \geq 2$, $n(n+1)/2 < n^2$, so condition (d) of Theorem 9.4.1 fails, and τ is not diagonalizable. For instance, $\mathbf{E}_{12}$ is not a sum of eigenvectors.

9.5 The Minimal Polynomial

Section 4 decided diagonalizability by counting eigenvectors, one eigenspace at a time. There is a second source of information about an operator, which Chapter 6 set up: the polynomials that kill it. On a finite-dimensional space they form an ideal of $F[x]$, and that ideal has a single monic generator. This section gives the generator its name, the minimal polynomial, and shows that its roots are exactly the eigenvalues. We learn two ways to compute it and check that similar matrices share it. The next section compares it with the characteristic polynomial, and Section 8 reads diagonalizability off its factorization.

9.5.1 The polynomial that divides all the others

Chapter 6 turned every polynomial identity satisfied by an operator into a statement about one ideal. For $T \in \mathcal{L}(V)$, the set

$$I_T = \{p \in F[x] : p(T) = 0\}$$

is an ideal of $F[x]$ (Theorem 6.7.3 (a)). If V is finite-dimensional, it contains a nonzero polynomial (Theorem 6.7.3 (c)). Then it consists of the multiples of a unique monic polynomial (Theorem 6.7.3 (b)).

Look at what this says for one operator. A projection P with $P \neq 0, \mathrm{id}_V$ satisfies $P^2 = P$, so $x^2 - x$ kills it. So do $x^3 - x^2$, $(x^2 - x)(x + 5)$, and $x^{10} - x^9$, and a short calculation shows each one is a multiple of $x^2 - x$. This is no accident. Every annihilating polynomial of P is a multiple of $x^2 - x$, which Chapter 6 found to be the monic generator of I_P. One polynomial controls the whole list, and whenever something in mathematics is unique we give it a name.

The minimal polynomial of T is the one monic polynomial that kills T and divides every other polynomial that kills T.

Definition 9.5.1. Minimal Polynomial

Let V be a **finite-dimensional** vector space over F and $T \in \mathcal{L}(V)$. The **minimal polynomial** of T is **the** unique **monic** polynomial $m_T \in F[x]$ with

$$I_T = \{p \in F[x] : p(T) = 0\} = \langle m_T \rangle.$$

Equivalently, m_T is the monic polynomial of **least degree** with $m_T(T) = 0$.

For $\mathbf{A} \in M_n(F)$, the minimal polynomial is $m_{\mathbf{A}} := m_{T_{\mathbf{A}}}$: the monic polynomial of least degree with $m_{\mathbf{A}}(\mathbf{A}) = \mathbf{O}$.

In words: we look at **all** polynomials p with $p(T) = 0$, where $p(T)$ replaces x^k by T^k and the constant term c by $c\,\mathrm{id}_V$ (Definition 2.4.2). They are exactly the multiples of one polynomial m_T. Among the generators of this ideal, which differ by nonzero constant factors, we pick **the monic** one. The second sentence of the definition gives the same polynomial described by size: it is the monic polynomial of **least degree** that kills T.

Well-definedness. Three things need checking, and Chapter 6 did all three. *Existence:* $I_T \neq \{0\}$ because V is finite-dimensional (Theorem 6.7.3 (c)). *Uniqueness:* a nonzero ideal of $F[x]$ has exactly one monic generator (Theorem 6.2.2). *The two descriptions agree:* the same theorem says that the monic generator is the unique monic polynomial of least degree in the ideal. For a matrix, $p(T_{\mathbf{A}}) = 0$ if and only if $p(\mathbf{A}) = \mathbf{O}$, because $[p(T_{\mathbf{A}})]_{\mathcal{E}}^{\mathcal{E}} = p(\mathbf{A})$ in the standard basis $\mathcal{E}$ (Corollary 4.1.3). So $I_{T_{\mathbf{A}}} = \{p : p(\mathbf{A}) = \mathbf{O}\}$, and the matrix version is the operator version for $T_{\mathbf{A}}$.

The name records the second description: m_T is the annihilating polynomial of minimal degree. The first description is the one we use in proofs, because it hands us a divisibility for free.

Theorem 9.5.1. Annihilating Polynomials Are Multiples of the Minimal Polynomial

Let V be finite-dimensional and $T \in \mathcal{L}(V)$. For every $p \in F[x]$,

$$p(T) = 0 \iff m_T \mid p.$$

In particular, a nonzero polynomial that annihilates T has degree at least $\deg m_T$. If $V \neq \{0\}$, then $\deg m_T \geq 1$.

Proof. By Definition 9.5.1, $p(T) = 0$ means $p \in \langle m_T \rangle$, that is, $p = q\, m_T$ for some $q \in F[x]$, that is, $m_T \mid p$. If $p \neq 0$ and $m_T \mid p$, then $\deg m_T \leq \deg p$ by Proposition 6.1.2 (c). Finally, the only monic polynomial of degree 0 is the constant 1, and $1(T) = \mathrm{id}_V \neq 0$ when $V \neq \{0\}$; so $m_T \neq 1$ and $\deg m_T \geq 1$.

■

This is the move we will use again and again: **a polynomial relation becomes a divisibility.** If we learn that $T^3 = T$, we do not need to know anything else about T to conclude that $m_T \mid x^3 - x = x(x-1)(x+1)$. That leaves only finitely many candidates for m_T. To list candidates like these we need to know the monic divisors of a factored polynomial, and unique factorization describes them exactly.

Lemma 9.5.2. Monic Divisors of a Factored Polynomial

Let $f = p_1^{e_1} \cdots p_k^{e_k} \in F[x]$, where $k \geq 1$, $p_1, \ldots, p_k$ are **distinct** monic irreducible polynomials, and $e_i \geq 1$. Then the monic divisors of f are exactly the polynomials

$$p_1^{f_1} \cdots p_k^{f_k} \qquad \text{with } 0 \leq f_i \leq e_i \text{ for each } i.$$

Idea. One direction is a product: the missing factors form the cofactor. For the other, factor a monic divisor d into irreducibles, multiply that factorization by one of the cofactor, and compare the result with the factorization of f. Uniqueness says the two lists of irreducibles agree, so the irreducibles in d are among the p_i, and none can occur more often in d than in f.

Proof. Each such product divides f, with cofactor $p_1^{e_1 - f_1} \cdots p_k^{e_k - f_k}$. Conversely, let d be monic with $f = dh$. If $d = 1$, take every $f_i = 0$. Otherwise d is nonconstant, and by

Theorem 6.3.5 (a) it is a product $q_1 \cdots q_s$ of monic irreducibles; the constant in front is the leading coefficient of d, which is 1. The cofactor h is nonzero (as $f \neq 0$), so it is either a nonzero constant c or, by the same theorem, $c\,r_1 \cdots r_t$ with r_j monic irreducible. Multiplying, $f = c\,q_1 \cdots q_s\,r_1 \cdots r_t$ (with no r's in the first case) is a factorization of f into a constant and monic irreducibles. By Theorem 6.3.5 (b), after renumbering it is the factorization $f = 1 \cdot p_1^{e_1} \cdots p_k^{e_k}$, in which p_i occurs exactly e_i times. So each q_j is one of the p_i, and p_i occurs at most e_i times among $q_1, \ldots, q_s$. Hence $d = p_1^{f_1} \cdots p_k^{f_k}$ with $0 \leq f_i \leq e_i$.

■

Examples. Each computation follows the same plan: find one annihilating polynomial, list its monic divisors with Lemma 9.5.2, and discard those that do not annihilate.

- **A diagonal matrix with a repeated entry.** Let $\mathbf{A} = \operatorname{diag}(1, 1, 2)$. Diagonal matrices multiply entry by entry, so

$$(\mathbf{A} - \mathbf{I})(\mathbf{A} - 2\mathbf{I}) = \operatorname{diag}(0, 0, 1)\ \operatorname{diag}(-1, -1, 0) = \mathbf{O}.$$

  By Theorem 9.5.1, $m_{\mathbf{A}}$ is a monic divisor of $(x-1)(x-2)$: one of 1, $x-1$, $x-2$, $(x-1)(x-2)$. None of the first three annihilates $\mathbf{A}$, since $\mathbf{I}$, $\mathbf{A} - \mathbf{I} = \operatorname{diag}(0, 0, 1)$, and $\mathbf{A} - 2\mathbf{I} = \operatorname{diag}(-1, -1, 0)$ are nonzero. So $m_{\mathbf{A}} = (x-1)(x-2)$. The repeated diagonal entry 1 appears only **once** in $m_{\mathbf{A}}$, while the characteristic polynomial is $p_{\mathbf{A}} = (x-1)^2(x-2)$.
- **A matrix with too few eigenvectors.** Let $\mathbf{J} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ over any field. Then $\mathbf{J} - \mathbf{I} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \neq \mathbf{O}$, and $(\mathbf{J} - \mathbf{I})^2 = \mathbf{O}$. So $m_{\mathbf{J}}$ is a monic divisor of $(x-1)^2$, namely 1, $x-1$, or $(x-1)^2$, and the first two do not annihilate $\mathbf{J}$. Hence $m_{\mathbf{J}} = (x-1)^2$, which equals $p_{\mathbf{J}}$.
- **Projections.** Chapter 6 found, by this plan, that a projection P with $P \neq 0$ and $P \neq \mathrm{id}_V$ has $m_P = x^2 - x$ (the examples after Theorem 6.7.3).
- **Differentiation.** Let D be differentiation on $\mathbb{R}[x]_{\leq n}$. Then $D^{n+1} = 0$ and $D^n \neq 0$ (Example (Differentiation Is Nilpotent)). By Lemma 9.5.2, the monic divisors of x^{n+1} are the powers x^j with $0 \leq j \leq n+1$. For $j \leq n$, $D^j \neq 0$, since $D^n = D^{n-j}D^j$ would otherwise be 0. Hence $m_D = x^{n+1}$, of degree $n+1 = \dim \mathbb{R}[x]_{\leq n}$.
- **Degenerate cases.** On $V \neq \{0\}$, the identity has $m = x - 1$ and the zero operator has $m = x$: each is killed by a polynomial of degree 1, which is the least possible degree. More generally $m_{c\,\mathrm{id}_V} = x - c$, and every 1×1 matrix (c) has $m = x - c$. On $V = \{0\}$, even the constant 1 annihilates, so $m_T = 1$. This last case is the reason for the hypothesis $V \neq \{0\}$ in Theorem 9.5.1.

Nonexample by minimal change. For $\mathbf{J}$ above, take $q = x - 1$ instead of $(x-1)^2$. It is still monic, it has the least degree a minimal polynomial can have on a nonzero space, and its only root is the eigenvalue 1 of $\mathbf{J}$. What fails is the clause $q(\mathbf{J}) = \mathbf{O}$: $q(\mathbf{J}) = \mathbf{J} - \mathbf{I} \neq \mathbf{O}$. In the other direction, $(x-1)^3$ is monic and does annihilate $\mathbf{J}$, but it fails the clause $I_{\mathbf{J}} = \langle q \rangle$: the annihilating polynomial $(x-1)^2$ is not a multiple of $(x-1)^3$.

Quick check. Find the minimal polynomial of $2\mathbf{I}_3$ and of $\mathbf{N} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} \in M_3(F)$.

Solution. $2\mathbf{I}_3 - 2\mathbf{I}_3 = \mathbf{O}$, so $m_{2\mathbf{I}_3}$ divides $x-2$; it is not 1, since $\mathbf{I}_3 \neq \mathbf{O}$. Hence $m_{2\mathbf{I}_3} = x-2$. For $\mathbf{N}$: $\mathbf{N}\mathbf{e}_1 = \mathbf{O}$, $\mathbf{N}\mathbf{e}_2 = \mathbf{e}_1$, $\mathbf{N}\mathbf{e}_3 = \mathbf{e}_2$, so $\mathbf{N}^2\mathbf{e}_3 = \mathbf{e}_1 \neq \mathbf{0}$ and $\mathbf{N}^3$ kills $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$. Thus $\mathbf{N}^3 = \mathbf{0} \neq \mathbf{N}^2$. By Lemma 9.5.2, $m_{\mathbf{N}} \in \{1, x, x^2, x^3\}$, and only x^3 annihilates $\mathbf{N}$ (the others give $\mathbf{I}_3, \mathbf{N}, \mathbf{N}^2 \neq \mathbf{O}$). So $m_{\mathbf{N}} = x^3$.

Why this definition. Each clause has a job.

- *Monic.* The polynomials m and $5m$ generate the same ideal, so "a generator" is determined only up to a nonzero constant. Insisting on leading coefficient 1 picks one, and makes statements like "$m_{\mathbf{A}} = (x-1)(x-2)$" meaningful.
- *Generator, not only least degree.* The two descriptions give the same polynomial. Stating the definition through the ideal puts Theorem 9.5.1 in front of us, and divisibility is the property we use.
- *Finite dimension.* Without it, I_T can be $\{0\}$, and then there is nothing to name. For multiplication by x on $F[x]$, $p(S)(1) = p$, so $p(S) \neq 0$ for every $p \neq 0$ (the remark after Theorem 6.7.2). The same happens for the right shift on $F^{\mathbb{N}}$.

Warning. **The minimal polynomial is not the characteristic polynomial.** For $\mathbf{I}_2$, $m = x-1$ while $p = (x-1)^2$; for $\operatorname{diag}(1,1,2)$ above, $m = (x-1)(x-2)$ while $p = (x-1)^2(x-2)$. The two can differ in their exponents, and m can have smaller degree than the size of the matrix. How they are related is the subject of the next section. Its main theorem is one that Chapter 7 warned against "proving" by substituting $x = \mathbf{A}$ into $\det(x\mathbf{I} - \mathbf{A})$; that shortcut is not a proof and must not be used here either.

9.5.2 The roots are the eigenvalues

The minimal polynomial is defined by an ideal, and eigenvalues by vectors. They are closely linked.

Theorem 9.5.3. Roots of the Minimal Polynomial

Let V be finite-dimensional, $T \in \mathcal{L}(V)$, and $\lambda \in F$. Then $m_T(\lambda) = 0$ if and only if λ is an eigenvalue of T. Consequently, if $\dim V \geq 1$, then m_T and p_T have the same roots in F.

Idea. ($\Leftarrow$) is a one-liner: apply $m_T(T) = 0$ to an eigenvector v. Since T scales v by λ, every polynomial in T scales v by the value of the polynomial at λ. For ($\Rightarrow$), a root gives a factor: $m_T = (x-\lambda)q$. The cofactor q has smaller degree, so by minimality it does **not** kill T, and some vector w survives it. Then $q(T)w$ is nonzero and is killed by $T - \lambda\,\mathrm{id}_V$. The eigenvector is manufactured by the part of m_T we removed.

Proof. ($\Leftarrow$) Let $v \neq 0$ with $Tv = \lambda v$. Then $T^j v = \lambda^j v$ for every $j \geq 0$, by induction: it holds for $j = 0$, and $T^{j+1}v = T^j(Tv) = T^j(\lambda v) = \lambda T^j v = \lambda^{j+1} v$, the first equality by Definition 2.4.2. Writing $p = \sum_j c_j x^j$, Definition 2.4.2 gives $p(T)v = \sum_j c_j T^j v = p(\lambda)v$ for every $p \in F[x]$. With $p = m_T$,

$$0 = m_T(T)v = m_T(\lambda)v.$$

Since $v \neq 0$, Theorem 1.1.9 gives $m_T(\lambda) = 0$.

($\Rightarrow$) Suppose $m_T(\lambda) = 0$. By Theorem 6.1.3 (b), $m_T = (x - \lambda)q$ for some $q \in F[x]$, and q is monic with $\deg q = \deg m_T - 1$ by Theorem 0.8.3. By Theorem 9.5.1, a nonzero polynomial of degree less than $\deg m_T$ does not annihilate T; so $q(T) \neq 0$, and there is $w \in V$ with $v \coloneqq q(T)w \neq 0$. By Theorem 6.7.1 (b),

$$(T - \lambda \operatorname{id}_V)v = (T - \lambda \operatorname{id}_V)\, q(T)w = m_T(T)w = 0.$$

So $Tv = \lambda v$ with $v \neq 0$, and λ is an eigenvalue of T.

If $\dim V \geq 1$, the eigenvalues of T are exactly the roots of p_T in F (Theorem 9.2.1, (a) $\Leftrightarrow$ (e)), so m_T and p_T have the same roots in F. This proves the theorem.

■

So the minimal polynomial knows the spectrum, just as the characteristic polynomial does, although the multiplicities of the roots may differ. The examples bear this out. $\operatorname{diag}(1, 1, 2)$ has spectrum $\{1, 2\}$, the roots of $(x - 1)(x - 2)$. $\mathbf{J}$ has spectrum $\{1\}$, the root of $(x - 1)^2$. The rotation R of $\mathbb{R}^2$ by a right angle has $m_R = x^2 + 1$ (the check after Theorem 6.7.3), with no real root, and indeed $\operatorname{spec}(R) = \emptyset$ over $\mathbb{R}$. Its minimal polynomial has no linear factor at all.

Two consequences are worth stating.

- **Invertibility.** Taking $\lambda = 0$: for $V \neq \{0\}$, T is invertible if and only if $m_T(0) \neq 0$. Indeed $m_T(0) = 0$ if and only if $0 \in \operatorname{spec}(T)$, if and only if $T = T - 0 \operatorname{id}_V$ is not invertible (Theorem 9.2.1, (a) $\Leftrightarrow$ (c) with $\lambda = 0$; for matrices this is Theorem 9.2.6). Chapter 6 proved the same statement without eigenvalues (Exercise (C1: Inverses are polynomials) (c)).
- **Over $\mathbb{C}$.** If $F = \mathbb{C}$ and $V \neq \{0\}$, then m_T splits (Corollary 6.5.5), and its roots are the distinct eigenvalues $\lambda_1, \ldots, \lambda_k$. So

  $$m_T = (x - \lambda_1)^{s_1} \cdots (x - \lambda_k)^{s_k} \qquad \text{with every } s_i \geq 1.$$

  Only the exponents remain to be found.

9.5.3 Computing the minimal polynomial

There are two practical methods. The first needs nothing but elimination. The second uses the eigenvalues, when we know them, to cut the search down to a few candidates.

Method 1: the first dependence among powers. The proof of Theorem 6.7.2 searched for a dependence among $\operatorname{id}_V, T, T^2, \ldots$. The **first** dependence that appears gives m_T exactly.

Proposition 9.5.4. The First Dependence Among the Powers

Let V be finite-dimensional and $T \in \mathcal{L}(V)$. Let $d \geq 0$ be the **least** integer such that the list $(\operatorname{id}_V, T, \ldots, T^d)$ in $\mathcal{L}(V)$ is linearly dependent. Then there are unique $c_0, \ldots, c_{d-1} \in F$ with

$$T^d = c_0 \operatorname{id}_V + c_1 T + \cdots + c_{d-1} T^{d-1},$$

and $m_T = x^d - c_{d-1}x^{d-1} - \cdots - c_1 x - c_0$. In particular $\deg m_T = d$. The same holds for $\mathbf{A} \in M_n(F)$, with powers of $\mathbf{A}$ in $M_n(F)$.

Proof. Such a d exists: with $n = \dim V$, the space $\mathcal{L}(V)$ has dimension n^2 (Theorem 4.1.4), so the list $(\mathrm{id}_V, T, \dots, T^{n^2})$, of length n^2+1, is linearly dependent (Theorem 1.6.3 (a)). By minimality of d, the list $(\mathrm{id}_V, \dots, T^{d-1})$ is independent (for $d = 0$ it is empty). By Lemma 1.5.3, since appending T^d makes it dependent, $T^d \in \mathrm{span}(\mathrm{id}_V, \dots, T^{d-1})$. The coefficients are unique by Theorem 1.4.3.

Let $q = x^d - c_{d-1}x^{d-1} - \cdots - c_0$. It is monic, and $q(T) = 0$ by the displayed relation. Let $p \neq 0$ with $\deg p < d$. Then $p(T) = \sum_{k<d} a_k T^k$ with the a_k not all zero, and this is not 0 because $(\mathrm{id}_V, \dots, T^{d-1})$ is independent. So no nonzero polynomial of degree less than d annihilates T, and q is the monic annihilating polynomial of least degree. By Definition 9.5.1, $q = m_T$.

■

Example. The Minimal Polynomial from the First Dependence

Find the minimal polynomial of

$$\mathbf{A} = \begin{pmatrix} 3 & 2 & -2 \\ -2 & -1 & 2 \\ -2 & -2 & 3 \end{pmatrix} \in M_3(\mathbb{R}),$$

the matrix of Example (Diagonalizing a 3×3 Matrix with a Repeated Eigenvalue), without using its eigenvalues.

Solution. *Degree* 1. $\mathbf{A}$ is not a scalar multiple of $\mathbf{I}_3$, since its $(1,2)$-entry is nonzero. So $(\mathbf{I}_3, \mathbf{A})$ is independent.

Degree 2. Multiplying out,

$$\mathbf{A}^2 = \begin{pmatrix} 9 & 8 & -8 \\ -8 & -7 & 8 \\ -8 & -8 & 9 \end{pmatrix}.$$

We look for a, b with $\mathbf{A}^2 = a\mathbf{A} + b\mathbf{I}_3$. The $(1,2)$-entries give $8 = 2a$, so $a = 4$. The $(1,1)$-entries give $9 = 3a + b$, so $b = -3$. Now check every entry of $4\mathbf{A} - 3\mathbf{I}_3$:

$$4\mathbf{A} - 3\mathbf{I}_3 = \begin{pmatrix} 12-3 & 8 & -8 \\ -8 & -4-3 & 8 \\ -8 & -8 & 12-3 \end{pmatrix} = \begin{pmatrix} 9 & 8 & -8 \\ -8 & -7 & 8 \\ -8 & -8 & 9 \end{pmatrix} = \mathbf{A}^2.$$

So $(\mathbf{I}_3, \mathbf{A}, \mathbf{A}^2)$ is dependent, and $d = 2$. By Proposition 9.5.4,

$$m_{\mathbf{A}} = x^2 - 4x + 3 = (x-1)(x-3).$$

This fits Theorem 9.5.3: Example (Diagonalizing a 3×3 Matrix with a Repeated Eigenvalue) found the eigenvalues 1 and 3, and $p_{\mathbf{A}} = (x-1)^2(x-3)$. Once again the repeated eigenvalue 1 appears only once in $m_{\mathbf{A}}$. Section 8 explains that this is exactly why $\mathbf{A}$ was diagonalizable.

Method 2: eigenvalues and trial exponents. Suppose p_T splits, $p_T = (x-\lambda_1)^{a_1} \cdots (x-\lambda_k)^{a_k}$ with distinct λ_i. By Theorem 9.5.3, every $x - \lambda_i$ divides m_T. If we find **any** annihilating

polynomial of the form $q = (x-\lambda_1)^{t_1}\cdots(x-\lambda_k)^{t_k}$, then $m_T \mid q$. By Lemma 9.5.2, $m_T = (x-\lambda_1)^{s_1}\cdots(x-\lambda_k)^{s_k}$ with $1 \le s_i \le t_i$, so only finitely many candidates remain, and we test them from the smallest degree up. The next section proves that $q = p_T$ always annihilates T, so the search always succeeds with $t_i \le a_i$. Here we verify that directly.

Example. The Minimal Polynomial from the Eigenvalues

Let $\mathbf{A} = \begin{pmatrix} 2 & 0 & 1 \\ 1 & 3 & 1 \\ 1 & 0 & 2 \end{pmatrix} \in M_3(\mathbb{R})$, with $p_{\mathbf{A}} = (x-1)(x-3)^2$ (Example (A 3×3 Characteristic Polynomial)). Find $m_{\mathbf{A}}$.

Solution. By Theorem 9.5.3, both $x-1$ and $x-3$ divide $m_{\mathbf{A}}$, so the candidate of least degree is $(x-1)(x-3)$. With

$$\mathbf{A}-\mathbf{I} = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 2 & 1 \\ 1 & 0 & 1 \end{pmatrix}, \qquad \mathbf{A}-3\mathbf{I} = \begin{pmatrix} -1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & -1 \end{pmatrix},$$

the product is

$$(\mathbf{A}-\mathbf{I})(\mathbf{A}-3\mathbf{I}) = \begin{pmatrix} 0 & 0 & 0 \\ 2 & 0 & 2 \\ 0 & 0 & 0 \end{pmatrix} \neq \mathbf{O}.$$

So $m_{\mathbf{A}} \neq (x-1)(x-3)$. Raise one exponent and try $(x-1)(x-3)^2$ (had this failed, the other candidate of degree 3, namely $(x-1)^2(x-3)$, would come next). Polynomials in $\mathbf{A}$ commute (Theorem 6.7.1 (c)), so we can multiply the matrix just found by $\mathbf{A}-3\mathbf{I}$ on the right:

$$(\mathbf{A}-\mathbf{I})(\mathbf{A}-3\mathbf{I})^2 = \begin{pmatrix} 0 & 0 & 0 \\ 2 & 0 & 2 \\ 0 & 0 & 0 \end{pmatrix}\begin{pmatrix} -1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & -1 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 \\ -2+2 & 0 & 2-2 \\ 0 & 0 & 0 \end{pmatrix} = \mathbf{O}.$$

So $m_{\mathbf{A}} \mid (x-1)(x-3)^2$, and by Lemma 9.5.2, $m_{\mathbf{A}} = (x-1)^s(x-3)^t$ with $1 \le s \le 1$ and $1 \le t \le 2$. The case $t = 1$ was excluded, so

$$m_{\mathbf{A}} = (x-1)(x-3)^2 = p_{\mathbf{A}}.$$

In Section 3 we found $g(3) = 1 < 2 = a(3)$ for this matrix. The exponent 2 of $x-3$ in $m_{\mathbf{A}}$ is the polynomial trace of the same shortage of eigenvectors.

The product in Method 2 need not be multiplied out completely: to rule out a candidate it is enough to find **one** nonzero entry, and to confirm one, the factored form usually makes the zero visible.

9.5.4 Similar matrices and block diagonal matrices

The minimal polynomial belongs to the operator, not to a basis. That is what makes it an invariant of similarity.

Theorem 9.5.5. Similar Matrices Have the Same Minimal Polynomial

Let V be finite-dimensional with a basis $\mathcal{B}$, and let $T \in \mathcal{L}(V)$. Then $m_T = m_{[T]_{\mathcal{B}}^{\mathcal{B}}}$. If $\mathbf{A}, \mathbf{B} \in M_n(F)$ are similar, then $m_{\mathbf{A}} = m_{\mathbf{B}}$.

Proof. By Corollary 4.1.3, $[p(T)]_{\mathcal{B}}^{\mathcal{B}} = p([T]_{\mathcal{B}}^{\mathcal{B}})$ for every $p \in F[x]$. An operator is zero exactly when its matrix is zero (Theorem 4.1.4), so $p(T) = 0$ if and only if $p([T]_{\mathcal{B}}^{\mathcal{B}}) = \mathbf{O}$. Thus T and $[T]_{\mathcal{B}}^{\mathcal{B}}$ have the same ideal of annihilating polynomials, and hence the same monic generator. If $\mathbf{B} = \mathbf{P}^{-1}\mathbf{A}\mathbf{P}$, then $p(\mathbf{B}) = \mathbf{P}^{-1}p(\mathbf{A})\mathbf{P}$ by Proposition 4.2.6 (c), so again $p(\mathbf{A}) = \mathbf{O}$ if and only if $p(\mathbf{B}) = \mathbf{O}$, and $m_{\mathbf{A}} = m_{\mathbf{B}}$. ■

The theorem gives a quick nonsimilarity test. For instance $\mathbf{I}_2$ and $\mathbf{J} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ have the same characteristic polynomial, but $m_{\mathbf{I}_2} = x - 1 \neq (x-1)^2 = m_{\mathbf{J}}$, so they are not similar.

Chapter 8 showed that a block diagonal matrix computes polynomials blockwise: $p(\mathbf{A} \oplus \mathbf{B}) = p(\mathbf{A}) \oplus p(\mathbf{B})$ (Theorem 8.1.4 (b)). So a polynomial kills $\mathbf{A} \oplus \mathbf{B}$ exactly when it kills both blocks. In the language of ideals, the answer is the least common multiple.

Proposition 9.5.6. Minimal Polynomial of a Block Diagonal Matrix

Let $\mathbf{A} \in M_k(F)$ and $\mathbf{B} \in M_l(F)$ with $k, l \geq 1$. Then $m_{\mathbf{A}\oplus\mathbf{B}} = \operatorname{lcm}(m_{\mathbf{A}}, m_{\mathbf{B}})$.

Proof. Let $p \in F[x]$. By Theorem 8.1.4 (b), $p(\mathbf{A}\oplus\mathbf{B}) = p(\mathbf{A})\oplus p(\mathbf{B})$, which is zero if and only if $p(\mathbf{A}) = \mathbf{O}$ and $p(\mathbf{B}) = \mathbf{O}$. By Theorem 9.5.1, this holds if and only if $p \in \langle m_{\mathbf{A}} \rangle \cap \langle m_{\mathbf{B}} \rangle$. So $I_{\mathbf{A}\oplus\mathbf{B}} = \langle m_{\mathbf{A}} \rangle \cap \langle m_{\mathbf{B}} \rangle$. This ideal contains $m_{\mathbf{A}\oplus\mathbf{B}} \neq 0$, and by Definition 6.2.4 its monic generator is $\operatorname{lcm}(m_{\mathbf{A}}, m_{\mathbf{B}})$. By Definition 9.5.1 its monic generator is also $m_{\mathbf{A}\oplus\mathbf{B}}$, and the monic generator is unique (Theorem 6.2.2). ■

For example, $\operatorname{diag}(1,1,2) = \mathbf{I}_2 \oplus (2)$ has $m = \operatorname{lcm}(x-1, x-2) = (x-1)(x-2)$, as computed above. And $\mathbf{J} \oplus (1)$ has $m = \operatorname{lcm}\left((x-1)^2, x-1\right) = (x-1)^2$: a block can contribute a factor that the other block already contains, and then the factor is not repeated. This is the difference from the characteristic polynomial, which multiplies: $p_{\mathbf{A}\oplus\mathbf{B}} = p_{\mathbf{A}}\, p_{\mathbf{B}}$.

Warning. **Equal minimal polynomials do not make matrices similar.** $\operatorname{diag}(1,1,2)$ and $\operatorname{diag}(1,2,2)$ both have minimal polynomial $(x-1)(x-2)$, by Proposition 9.5.6. But their characteristic polynomials $(x-1)^2(x-2)$ and $(x-1)(x-2)^2$ differ, so they are not

similar (Theorem 7.10.3).

Remark. Even the minimal and characteristic polynomials **together** do not decide similarity. With $\mathbf{J} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$, the matrices $\mathbf{J} \oplus \mathbf{J}$ and $\mathbf{J} \oplus \mathbf{I}_2$ in $M_4(F)$ both have characteristic polynomial $(x-1)^4$ and, by Proposition 9.5.6, minimal polynomial $(x-1)^2$, as $\operatorname{lcm}\left((x-1)^2, (x-1)^2\right)$ and $\operatorname{lcm}\left((x-1)^2, x-1\right)$ are both $(x-1)^2$. Yet $(\mathbf{J} \oplus \mathbf{J}) - \mathbf{I}_4$ has rank 2 and $(\mathbf{J} \oplus \mathbf{I}_2) - \mathbf{I}_4$ has rank 1; if $\mathbf{B} = \mathbf{P}^{-1}\mathbf{A}\mathbf{P}$, then $\mathbf{B} - \mathbf{I} = \mathbf{P}^{-1}(\mathbf{A} - \mathbf{I})\mathbf{P}$, so these ranks would have to agree (Proposition 4.2.6 (a)), and the two matrices are not similar. The invariant that does decide similarity is the subject of Chapter 10.

9.5.5 Exercises

9.5.5.1 A. Check your understanding

Exercise. A1

(a) Define the minimal polynomial of an operator T on a finite-dimensional space, and give its two equivalent descriptions.

(b) True or false: $m_{\mathbf{A}} = p_{\mathbf{A}}$ for every $\mathbf{A} \in M_n(F)$. Justify your answer.

(c) State the relation between the roots of m_T and the eigenvalues of T.

(d) Let $\mathbf{A} \neq \mathbf{O}$ satisfy $\mathbf{A}^3 = \mathbf{O}$. Which polynomials can $m_{\mathbf{A}}$ be?

(e) True or false: if $m_{\mathbf{A}} = m_{\mathbf{B}}$, then $\mathbf{A} \sim \mathbf{B}$. Justify your answer.

(f) Explain why multiplication by x on $F[x]$ has no minimal polynomial.

Solution.

(a) m_T is the unique monic polynomial with $\{p : p(T) = 0\} = \langle m_T \rangle$; equivalently, the monic polynomial of least degree with $m_T(T) = 0$ (Definition 9.5.1).

(b) False. For $\mathbf{I}_2$, $m = x - 1$ and $p = (x-1)^2$.

(c) They coincide: $m_T(\lambda) = 0$ if and only if λ is an eigenvalue of T (Theorem 9.5.3).

(d) By Theorem 9.5.1, $m_{\mathbf{A}} \mid x^3$, so $m_{\mathbf{A}} \in \{1, x, x^2, x^3\}$ by Lemma 9.5.2. It is not 1 (as $\mathbf{I} \neq \mathbf{O}$) and not x (as $\mathbf{A} \neq \mathbf{O}$). So $m_{\mathbf{A}} = x^2$ or $m_{\mathbf{A}} = x^3$, and both occur: $\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$ and the matrix $\mathbf{N}$ of the Quick check.

(e) False. $\operatorname{diag}(1, 1, 2)$ and $\operatorname{diag}(1, 2, 2)$ have the same minimal polynomial but different characteristic polynomials.

(f) For $S(f) = xf$, we have $p(S)(1) = p$, so $p(S) \neq 0$ for all $p \neq 0$. The ideal I_S is $\{0\}$ and has no monic generator. The space $F[x]$ is infinite-dimensional, so the definition does not apply.

9.5.5.2 B. Practice

Exercise. B1: Computing minimal polynomials

Find the minimal polynomial of each real matrix.

(a) $\mathbf{A}_1 = \begin{pmatrix} 4 & -1 \\ 1 & 2 \end{pmatrix}$.

(b) $\mathbf{A}_2 = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}$.

(c) $\mathbf{A}_3 = \begin{pmatrix} 2 & 1 & 1 \\ 0 & 1 & -1 \\ -1 & -1 & 1 \end{pmatrix}$, given that $p_{\mathbf{A}_3} = (x-1)^2(x-2)$.

(d) $\mathbf{A}_4 = \begin{pmatrix} 2 & 1 & 1 \\ -1 & 0 & -1 \\ 1 & 1 & 2 \end{pmatrix}$, given that $p_{\mathbf{A}_4} = (x-1)^2(x-2)$.

Hence decide whether $\mathbf{A}_3$ and $\mathbf{A}_4$ are similar.

Solution.

(a) $\mathbf{A}_1$ is not scalar, so $\deg m_{\mathbf{A}_1} \geq 2$. Here $\operatorname{tr}\mathbf{A}_1 = 6$ and $\det\mathbf{A}_1 = 9$, and $\mathbf{A}_1^2 = \begin{pmatrix} 15 & -6 \\ 6 & 3 \end{pmatrix} = 6\mathbf{A}_1 - 9\mathbf{I}_2$, as one checks entry by entry. By Proposition 9.5.4, $m_{\mathbf{A}_1} = x^2 - 6x + 9 = (x-3)^2$.

(b) $\mathbf{A}_2$ is not scalar. Every entry of $\mathbf{A}_2^2$ is $1+1+1 = 3$, so $\mathbf{A}_2^2 = 3\mathbf{A}_2$. By Proposition 9.5.4, $m_{\mathbf{A}_2} = x^2 - 3x = x(x-3)$.

(c) By Theorem 9.5.3, $(x-1)(x-2) \mid m_{\mathbf{A}_3}$. First candidate:

$$(\mathbf{A}_3 - \mathbf{I})(\mathbf{A}_3 - 2\mathbf{I}) = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 0 & -1 \\ -1 & -1 & 0 \end{pmatrix}\begin{pmatrix} 0 & 1 & 1 \\ 0 & -1 & -1 \\ -1 & -1 & -1 \end{pmatrix} = \begin{pmatrix} -1 & -1 & -1 \\ 1 & 1 & 1 \\ 0 & 0 & 0 \end{pmatrix} \neq \mathbf{O}.$$

Next candidate: multiplying this on the left by $\mathbf{A}_3 - \mathbf{I}$, the rows of the result are $(1,1,1)\mathbf{M}$, $(0,0,-1)\mathbf{M}$, and $(-1,-1,0)\mathbf{M}$, where $\mathbf{M}$ is the matrix just found. Since the rows of $\mathbf{M}$ sum to zero and its third row is zero, all three products are $\mathbf{0}$. So $(\mathbf{A}_3 - \mathbf{I})^2(\mathbf{A}_3 - 2\mathbf{I}) = \mathbf{O}$, and by Lemma 9.5.2, $m_{\mathbf{A}_3} = (x-1)^2(x-2)$.

(d) Again $(x-1)(x-2) \mid m_{\mathbf{A}_4}$, and

$$(\mathbf{A}_4 - \mathbf{I})(\mathbf{A}_4 - 2\mathbf{I}) = \begin{pmatrix} 1 & 1 & 1 \\ -1 & -1 & -1 \\ 1 & 1 & 1 \end{pmatrix}\begin{pmatrix} 0 & 1 & 1 \\ -1 & -2 & -1 \\ 1 & 1 & 0 \end{pmatrix} = \mathbf{O},$$

since each column of the second matrix has entries summing to 0 and each row of the first is a multiple of $(1,1,1)$. So $m_{\mathbf{A}_4} = (x-1)(x-2)$.

Hence $m_{\mathbf{A}_3} \neq m_{\mathbf{A}_4}$, and by Theorem 9.5.5 $\mathbf{A}_3$ and $\mathbf{A}_4$ are **not** similar, although their characteristic polynomials agree. (This matches Exercise (B1: Computing multiplicities), where the same two matrices appear as $\mathbf{A}_2$ and $\mathbf{A}_3$, with $g(1) = 1$ for the first and $g(1) = 2$ for the second.)

Exercise. B2: Operators with $T^2 = T$

Let $V \neq \{0\}$ be finite-dimensional and $T \in \mathcal{L}(V)$ with $T^2 = T$.

(a) Show that $m_T \in \{x,\ x-1,\ x(x-1)\}$.

(b) Show that $m_T = x$ exactly when $T = 0$, and $m_T = x - 1$ exactly when $T = \mathrm{id}_V$.

(c) What is m_T on $V = \{0\}$?

Solution.

(a) $x^2 - x = x(x-1)$ annihilates T, so $m_T \mid x(x-1)$ by Theorem 9.5.1. By Lemma 9.5.2, the monic divisors are $1, x, x-1, x(x-1)$. Since $V \neq \{0\}$, $\deg m_T \geq 1$, so $m_T \neq 1$.

(b) If $m_T = x$, then $T = m_T(T) = 0$. Conversely, if $T = 0$, then x annihilates T, so $m_T \mid x$, and $m_T = x$ by (a). Likewise $m_T = x - 1$ if and only if $T - \mathrm{id}_V = 0$. In all other cases $m_T = x(x-1)$.

(c) On $V = \{0\}$, $1(T) = \mathrm{id}_V = 0$, so $m_T = 1$.

Exercise. B3: Companion matrices

Let $n \geq 2$, $p = x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0 \in F[x]$, and let $\mathbf{C} = \mathbf{C}(p)$ be its companion matrix (Exercise (C1: Companion matrices)).

(a) Show that $\mathbf{C}\mathbf{e}_k = \mathbf{e}_{k+1}$ for $1 \leq k \leq n-1$, so $\mathbf{C}^k\mathbf{e}_1 = \mathbf{e}_{k+1}$ for $0 \leq k \leq n-1$.

(b) Deduce that no nonzero polynomial of degree less than n annihilates $\mathbf{C}$.

(c) Show that $p(\mathbf{C})\mathbf{e}_1 = \mathbf{0}$, and hence $p(\mathbf{C})\mathbf{e}_{k+1} = \mathbf{0}$ for all $k \leq n-1$.

(d) Deduce that $m_{\mathbf{C}} = p = p_{\mathbf{C}}$.

Solution.

(a) Column k of $\mathbf{C}$, for $k \leq n-1$, has a single 1 in row $k+1$, so $\mathbf{C}\mathbf{e}_k = \mathbf{e}_{k+1}$. By induction on k, $\mathbf{C}^k\mathbf{e}_1 = \mathbf{C}(\mathbf{C}^{k-1}\mathbf{e}_1) = \mathbf{C}\mathbf{e}_k = \mathbf{e}_{k+1}$.

(b) Let $q = \sum_{k=0}^{d} b_k x^k \neq 0$ with $d \leq n-1$. By (a), $q(\mathbf{C})\mathbf{e}_1 = \sum_k b_k \mathbf{C}^k\mathbf{e}_1 = \sum_k b_k \mathbf{e}_{k+1}$, which is not $\mathbf{0}$ because some $b_k \neq 0$. So $q(\mathbf{C}) \neq \mathbf{O}$.

(c) By (a), $\mathbf{C}^n\mathbf{e}_1 = \mathbf{C}\mathbf{e}_n$, the last column $(-a_0, \ldots, -a_{n-1}) = -\sum_{k=0}^{n-1} a_k \mathbf{e}_{k+1}$. Hence

$$p(\mathbf{C})\mathbf{e}_1 = \mathbf{C}^n\mathbf{e}_1 + \sum_{k=0}^{n-1} a_k \mathbf{C}^k\mathbf{e}_1 = -\sum_k a_k \mathbf{e}_{k+1} + \sum_k a_k \mathbf{e}_{k+1} = \mathbf{0}.$$

For $k \leq n-1$, polynomials in $\mathbf{C}$ commute (Theorem 6.7.1 (c)), so $p(\mathbf{C})\mathbf{e}_{k+1} = p(\mathbf{C})\mathbf{C}^k\mathbf{e}_1 = \mathbf{C}^k p(\mathbf{C})\mathbf{e}_1 = \mathbf{0}$.

(d) By (c), $p(\mathbf{C})$ kills the standard basis, so $p(\mathbf{C}) = \mathbf{O}$ and $m_{\mathbf{C}} \mid p$. By (b), $\deg m_{\mathbf{C}} \geq n = \deg p$. Two monic polynomials, one dividing the other, with $\deg m_{\mathbf{C}} \geq \deg p$, are equal (Proposition 6.1.2 (c), (d)). So $m_{\mathbf{C}} = p$, and $p = p_{\mathbf{C}}$ by Exercise (C1: Companion matrices) (b). Every monic polynomial of degree $n \geq 2$ is therefore the minimal polynomial of some $n \times n$ matrix.

9.5.5.3 C. Going deeper

Exercise. C1: Cyclic vectors

Let $\dim V = n \geq 1$ and $T \in \mathcal{L}(V)$. A vector v is a **cyclic vector** for T if $(v, Tv, \dots, T^{n-1}v)$ is a basis of V.

(a) Prove that if T has a cyclic vector, then $\deg m_T \geq n$.

(b) Suppose v is cyclic and $T^n v = -(a_0 v + a_1 Tv + \cdots + a_{n-1}T^{n-1}v)$. Show that the matrix of T in the basis $(v, \dots, T^{n-1}v)$ is the companion matrix of $p = x^n + a_{n-1}x^{n-1} + \cdots + a_0$ (for $n \geq 2$), and deduce $m_T = p_T = p$.

(c) Find a cyclic vector for $\mathbf{J} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$, and show that $\mathbf{I}_2$ has none.

(The converse of (a), that $\deg m_T = n$ forces a cyclic vector, is proved in Chapter 10.)

Solution.

(a) Let $q = \sum_{k=0}^{d} b_k x^k \neq 0$ with $d \leq n-1$. Then $q(T)v = \sum_k b_k T^k v \neq 0$, because the vectors $T^k v$, $k \leq n-1$, are independent and some $b_k \neq 0$. So $q(T) \neq 0$, and by Theorem 9.5.1, $\deg m_T \geq n$.

(b) Let $\mathcal{B} = (v, Tv, \dots, T^{n-1}v)$. For $k \leq n-2$, $T(T^k v) = T^{k+1}v$, the $(k+2)$-th basis vector, so column $k+1$ of $[T]_{\mathcal{B}}^{\mathcal{B}}$ is $\mathbf{e}_{k+2}$. The last column is $[T^n v]_{\mathcal{B}} = (-a_0, \dots, -a_{n-1})$. This is $\mathbf{C}(p)$. By Theorem 9.5.5 and Exercise (B3: Companion matrices) (d), $m_T = m_{\mathbf{C}(p)} = p$. By Definition 7.10.2 and Exercise (C1: Companion matrices) (b), $p_T = p_{\mathbf{C}(p)} = p$. (For $n = 1$, $Tv = -a_0 v$ and $T = -a_0 \operatorname{id}_V$, so $m_T = p_T = x + a_0$ directly.)

(c) $\mathbf{J}\mathbf{e}_2 = (1, 1)$, and $(\mathbf{e}_2, (1, 1))$ is independent, so $\mathbf{e}_2$ is cyclic. For $\mathbf{I}_2$, every $\mathbf{v}$ has $\mathbf{I}_2\mathbf{v} = \mathbf{v}$, so $(\mathbf{v}, \mathbf{I}_2\mathbf{v})$ is dependent. This matches (a): $\deg m_{\mathbf{I}_2} = 1 < 2$.

Exercise. C2: Minimal polynomials of the pieces

Let V be finite-dimensional, $T \in \mathcal{L}(V)$, and U a T-invariant subspace, with restriction $T|_U$ and induced operator $\bar{T}$ on V/U (Definition 9.1.2).

(a) Prove that $p(T|_U) = p(T)|_U$ and $p(\bar{T})(v + U) = p(T)v + U$ for every $p \in F[x]$.

(b) Deduce that $m_{T|_U} \mid m_T$ and $m_{\bar{T}} \mid m_T$.

(c) Prove that $m_T \mid m_{T|_U}\, m_{\bar{T}}$.

(d) For $\mathbf{J} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ and $U = \operatorname{span}(\mathbf{e}_1)$, compute all three minimal polynomials, and show

that m_T need not equal $\operatorname{lcm}(m_{T|_U}, m_{\bar{T}})$.

Hint: for (c), apply $m_{\bar{T}}(T)$ to v and ask where the result lies.

Solution.

(a) First, $(T|_U)^k = T^k|_U$ by induction on k: for $u \in U$, $(T|_U)^{k+1}u = T|_U((T|_U)^k u) = T(T^k u) = T^{k+1}u$. Taking linear combinations, $p(T|_U)u = p(T)u$. For the quotient, $\bar{T}^k(v+U) = T^k v + U$ by the same induction, using Definition 9.1.2 (b) at each step, and $p(\bar{T})(v+U) = \sum_k c_k(T^k v + U) = p(T)v + U$ by the coset operations (Theorem 4.4.4).

(b) By (a), $m_T(T|_U) = m_T(T)|_U = 0$ and $m_T(\bar{T})(v+U) = 0 + U$ for all v. By Theorem 9.5.1, applied to $T|_U \in \mathcal{L}(U)$ and to $\bar{T} \in \mathcal{L}(V/U)$, both finite-dimensional, $m_{T|_U} \mid m_T$ and $m_{\bar{T}} \mid m_T$.

(c) Let $v \in V$. By (a), $m_{\bar{T}}(T)v + U = m_{\bar{T}}(\bar{T})(v+U) = 0 + U$, so $u := m_{\bar{T}}(T)v \in U$. By (a) again, $m_{T|_U}(T)u = m_{T|_U}(T|_U)u = 0$. Hence $(m_{T|_U}\, m_{\bar{T}})(T)v = m_{T|_U}(T)\, m_{\bar{T}}(T)v = 0$ (Theorem 6.7.1 (b)). As v was arbitrary, $m_{T|_U} m_{\bar{T}}$ annihilates T, and $m_T \mid m_{T|_U} m_{\bar{T}}$ by Theorem 9.5.1.

(d) $\mathbf{J}\mathbf{e}_1 = \mathbf{e}_1$, so $T|_U = \mathrm{id}_U$ and $m_{T|_U} = x - 1$. The quotient F^2/U has basis $\mathbf{e}_2 + U$, and $\bar{T}(\mathbf{e}_2 + U) = (1,1) + U = \mathbf{e}_2 + U$, so $\bar{T} = \mathrm{id}$ and $m_{\bar{T}} = x - 1$. But $m_{\mathbf{J}} = (x-1)^2$, computed earlier. So $m_{\mathbf{J}} = m_{T|_U} m_{\bar{T}}$, while $\operatorname{lcm}(m_{T|_U}, m_{\bar{T}}) = x - 1 \neq m_{\mathbf{J}}$. Compare Proposition 9.5.6: when U has an invariant complement the lcm is right, but in general it is not.

9.6 The Cayley–Hamilton Theorem

An operator on an n-dimensional space now carries two monic polynomials. The characteristic polynomial p_T comes from determinants and has degree exactly n. The minimal polynomial m_T comes from the ideal of polynomials that kill T, and the previous section showed that it has the same roots as p_T. Chapter 7 stated, without proof, the fact that ties the two together: every square matrix satisfies its own characteristic polynomial. Chapter 7 also warned against a one-line fake proof of that statement. This section gives a real proof, built on the adjugate identity of Chapter 7 read over $F[x]$. It then draws the consequences: m_T divides p_T, inverses are polynomials, and high powers of a matrix reduce to low ones.

9.6.1 The statement, and a proof that is not one

Chapter 6 proved that every operator on an n-dimensional space is killed by some nonzero polynomial of degree at most n^2 (Theorem 6.7.2). The bound came from counting: $\mathcal{L}(V)$ has dimension n^2. Counting cannot do better, yet in every example so far a polynomial of degree at most n sufficed. The characteristic polynomial always has degree n, and it is the natural candidate.

Theorem 9.6.1. Cayley–Hamilton Theorem

Let $n \geq 1$ and $\mathbf{A} \in M_n(F)$. Then

$$p_{\mathbf{A}}(\mathbf{A}) = \mathbf{O}.$$

Likewise, if V is a finite-dimensional vector space over F with $\dim V \geq 1$ and $T \in \mathcal{L}(V)$, then $p_T(T) = 0$.

Here $p_{\mathbf{A}}(\mathbf{A})$ means what it always means (Definition 0.9.14): if $p_{\mathbf{A}} = x^n + c_{n-1}x^{n-1} + \cdots + c_1 x + c_0$, then

$$p_{\mathbf{A}}(\mathbf{A}) = \mathbf{A}^n + c_{n-1}\mathbf{A}^{n-1} + \cdots + c_1\mathbf{A} + c_0\mathbf{I}_n,$$

an $n \times n$ matrix, with the constant term becoming $c_0\mathbf{I}_n$.

For $n = 2$ the theorem can be checked by hand. With $\mathbf{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, $p_{\mathbf{A}} = x^2 - (\operatorname{tr}\mathbf{A})x + \det\mathbf{A}$, and Exercise (C2: Nilpotent 2×2 matrices) (b) multiplied out

$$\mathbf{A}^2 - (a+d)\mathbf{A} + (ad - bc)\mathbf{I}_2 = \mathbf{O}.$$

For $n = 3$, the same brute force already involves cubic expressions in nine entries, and for general n it is hopeless. We need an argument. There is a tempting one, and we name it before it can do any harm.

Warning. **The fake proof of Cayley–Hamilton, again.** "Since $p_{\mathbf{A}}(x) = \det(x\mathbf{I} - \mathbf{A})$, substitute $x = \mathbf{A}$ to get $p_{\mathbf{A}}(\mathbf{A}) = \det(\mathbf{A}\mathbf{I} - \mathbf{A}) = \det\mathbf{O} = 0$." Chapter 7 (the warning at the end of its §10) explained why this is not a proof. The left side $p_{\mathbf{A}}(\mathbf{A})$ is a **matrix**, while $\det(\cdots)$ is a **scalar**, so the equation compares objects of different kinds. And substituting $\mathbf{A}$ for x **inside the entries** of $x\mathbf{I} - \mathbf{A}$ is not what substituting into a polynomial means: the entry $x - a_{11}$ would have to become the matrix $\mathbf{A} - a_{11}\mathbf{I}$, and the determinant of a matrix whose entries are matrices is undefined. The same "argument" applied to $q(x) = \operatorname{tr}(x\mathbf{I} - \mathbf{A}) = nx - \operatorname{tr}\mathbf{A}$ would prove $q(\mathbf{A}) = \mathbf{O}$, which is false for $\mathbf{A} = \operatorname{diag}(1, 0)$: there $q(\mathbf{A}) = 2\mathbf{A} - \mathbf{I} = \operatorname{diag}(1, -1)$. Any correct proof must keep x a polynomial variable until the very end, and only then put a matrix in its place.

9.6.2 A proof through the adjugate

The adjugate identity of Chapter 7 says $\mathbf{M}\operatorname{adj}\mathbf{M} = (\det\mathbf{M})\,\mathbf{I}$ for every square matrix $\mathbf{M}$, and the remark after Theorem 7.6.2 notes that its proof works over any commutative ring. Over $F[x]$, with $\mathbf{M} = x\mathbf{I} - \mathbf{A}$, it reads

$$(x\mathbf{I} - \mathbf{A})\operatorname{adj}(x\mathbf{I} - \mathbf{A}) = p_{\mathbf{A}}(x)\,\mathbf{I}. \tag{1}$$

This is an identity between matrices of **polynomials**, and it contains $p_{\mathbf{A}}$ on the right. The proof extracts $p_{\mathbf{A}}(\mathbf{A})$ from it legally.

Idea. Try $n = 2$ first. By Exercise (C3: An adjugate over $F[x]$) (b) (done there over $\mathbb{R}$, but the computation works over any field), $\operatorname{adj}(x\mathbf{I} - \mathbf{A}) = x\mathbf{I} - \operatorname{adj}\mathbf{A}$, a polynomial in x with **matrix coefficients** $\mathbf{B}_1 = \mathbf{I}$ and $\mathbf{B}_0 = -\operatorname{adj}\mathbf{A}$. Expanding the left side of (1),

$$(x\mathbf{I} - \mathbf{A})(x\mathbf{I} - \operatorname{adj}\mathbf{A}) = x^2\,\mathbf{I} - x\,(\mathbf{A} + \operatorname{adj}\mathbf{A}) + \mathbf{A}\operatorname{adj}\mathbf{A}.$$

Comparing the coefficients of x^2, x, and 1 with those of $(x^2 - (\operatorname{tr}\mathbf{A})x + \det\mathbf{A})\mathbf{I}$ gives three matrix equations:

$$\mathbf{I} = \mathbf{I}, \qquad -(\mathbf{A} + \operatorname{adj}\mathbf{A}) = -(\operatorname{tr}\mathbf{A})\,\mathbf{I}, \qquad \mathbf{A}\operatorname{adj}\mathbf{A} = (\det\mathbf{A})\,\mathbf{I}.$$

Now multiply the first by $\mathbf{A}^2$, the second by $\mathbf{A}$, the third by $\mathbf{I}$, and add. The right sides add

up to $\mathbf{A}^2-(\operatorname{tr}\mathbf{A})\mathbf{A}+(\det\mathbf{A})\mathbf{I}=p_{\mathbf{A}}(\mathbf{A})$. The left sides add up to $\mathbf{A}^2-\mathbf{A}^2-\mathbf{A}\operatorname{adj}\mathbf{A}+\mathbf{A}\operatorname{adj}\mathbf{A}$, in which every term cancels against its neighbor. So $p_{\mathbf{A}}(\mathbf{A})=\mathbf{O}$. This is the whole proof; for general n there are $n+1$ equations and the cancellation is a telescoping sum. The variable x is never replaced by a matrix: we compare coefficients first, and only then multiply the resulting **matrix** equations by powers of $\mathbf{A}$.

Proof. Let $p_{\mathbf{A}}=c_nx^n+c_{n-1}x^{n-1}+\cdots+c_1x+c_0$ with $c_n=1$ (Theorem 7.10.2).

Step 1: the adjugate as a polynomial with matrix coefficients. If $n=1$, then $\operatorname{adj}(x\mathbf{I}-\mathbf{A})=(1)$ by Definition 7.6.2. If $n\geq 2$, each entry of $\operatorname{adj}(x\mathbf{I}-\mathbf{A})$ is a cofactor $\pm\det\mathbf{N}$, where $\mathbf{N}$ is an $(n-1)\times(n-1)$ submatrix of $x\mathbf{I}-\mathbf{A}$ (Definition 7.6.2, Definition 7.6.1). Every entry of $\mathbf{N}$ lies in $F[x]_{\leq 1}$. A product of $n-1$ such polynomials lies in $F[x]_{\leq n-1}$ (Theorem 0.8.3, the zero polynomial being harmless), and so does a sum of such products (Theorem 0.8.2); in particular the Leibniz sum $\det\mathbf{N}$ does (Definition 7.4.1). In both cases every entry of $\operatorname{adj}(x\mathbf{I}-\mathbf{A})$ lies in $F[x]_{\leq n-1}$. For $0\leq k\leq n-1$, let $\mathbf{B}_k\in M_n(F)$ be the matrix whose (i,j)-entry is the coefficient of x^k in the (i,j)-entry of $\operatorname{adj}(x\mathbf{I}-\mathbf{A})$. Then, entry by entry,

$$\operatorname{adj}(x\mathbf{I}-\mathbf{A})=\sum_{k=0}^{n-1}x^k\mathbf{B}_k.$$

Put $\mathbf{B}_{-1}=\mathbf{B}_n=\mathbf{O}$ for convenience.

Step 2: compare coefficients. By (1), which holds by Theorem 7.6.2 over the commutative ring $F[x]$ (the remark after it), the (i,j)-entries of both sides of (1) agree. On the left, by Definition 0.9.9,

$$\sum_{l=1}^{n}(x\delta_{il}-a_{il})\sum_{k=0}^{n-1}(\mathbf{B}_k)_{lj}\,x^k=\sum_{k=0}^{n-1}(\mathbf{B}_k)_{ij}\,x^{k+1}-\sum_{k=0}^{n-1}(\mathbf{AB}_k)_{ij}\,x^k,$$

where $\delta_{il}=1$ if $i=l$ and 0 otherwise. So for $0\leq k\leq n$, the coefficient of x^k is $(\mathbf{B}_{k-1})_{ij}-(\mathbf{AB}_k)_{ij}$. On the right the (i,j)-entry is $p_{\mathbf{A}}\,\delta_{ij}$, whose coefficient of x^k is $c_k\delta_{ij}$. Two polynomials are equal exactly when their coefficients are equal (Definition 0.8.1). Since this holds for every (i,j),

$$\mathbf{B}_{k-1}-\mathbf{AB}_k=c_k\,\mathbf{I}\qquad(0\leq k\leq n).\tag{2}$$

Step 3: telescope. Multiply equation (2) for k on the left by $\mathbf{A}^k$, and add the $n+1$ equations. On the right we get $\sum_{k=0}^{n}c_k\mathbf{A}^k=p_{\mathbf{A}}(\mathbf{A})$ by Definition 0.9.14. On the left, writing $\mathbf{S}_k=\mathbf{A}^k\mathbf{B}_{k-1}$ for $0\leq k\leq n+1$,

$$\sum_{k=0}^{n}\left(\mathbf{A}^k\mathbf{B}_{k-1}-\mathbf{A}^{k+1}\mathbf{B}_k\right)=\sum_{k=0}^{n}(\mathbf{S}_k-\mathbf{S}_{k+1})=\mathbf{S}_0-\mathbf{S}_{n+1}=\mathbf{B}_{-1}-\mathbf{A}^{n+1}\mathbf{B}_n=\mathbf{O},$$

since $\mathbf{B}_{-1}=\mathbf{B}_n=\mathbf{O}$. Hence $p_{\mathbf{A}}(\mathbf{A})=\mathbf{O}$.

The operator version. Let $\mathcal{B}$ be a basis of V and $\mathbf{A}=[T]_{\mathcal{B}}^{\mathcal{B}}$. By Definition 7.10.2, $p_T=p_{\mathbf{A}}$. By Corollary 4.1.3, $[p_T(T)]_{\mathcal{B}}^{\mathcal{B}}=p_T(\mathbf{A})=p_{\mathbf{A}}(\mathbf{A})=\mathbf{O}$, so $p_T(T)=0$ (Theorem 4.1.4). This proves the theorem.

■

Compare with the fake proof. There, $\mathbf{A}$ was pushed into the entries of $x\mathbf{I}-\mathbf{A}$. Here, (1) is first unpacked into the $n+1$ matrix equations (2), which contain no x at all, and $\mathbf{A}$ enters only when we multiply those equations by $\mathbf{A}^k$. Every multiplication is on the **left**, so no commutativity of matrices was used; the only commutativity used is inside $F[x]$, in the adjugate identity.

Nothing in the proof divides. The determinant, the adjugate identity, coefficient comparison, and matrix multiplication all make sense over a commutative ring, and Exercise (C1: Cayley–Hamilton over a commutative ring) asks you to check that the theorem holds, with the same proof, for matrices with entries in a commutative ring such as $\mathbb{Z}$ or $\mathbb{R}[t]$.

The proof also produced more than it was asked for. Equation (2) with $k=n$ says $\mathbf{B}_{n-1}=\mathbf{I}$, and the others say $\mathbf{B}_{k-1}=\mathbf{A}\mathbf{B}_k+c_k\mathbf{I}$. Working down from $\mathbf{B}_{n-1}$, every $\mathbf{B}_k$ is a polynomial in $\mathbf{A}$; Exercise (C2: The adjugate is a polynomial in the matrix) turns this into a formula for $\operatorname{adj}\mathbf{A}$.

9.6.3 Consequences

The first consequence is the promised comparison of the two polynomials.

Corollary 9.6.2. The Minimal Polynomial Divides the Characteristic Polynomial

Let V be finite-dimensional with $\dim V=n\geq 1$, and let $T\in\mathcal{L}(V)$.

(a) $m_T \mid p_T$; in particular $1\leq \deg m_T\leq n$.

(b) If p_T splits over F, say $p_T=(x-\lambda_1)^{a_1}\cdots(x-\lambda_k)^{a_k}$ with distinct λ_i, then

$$m_T=(x-\lambda_1)^{s_1}\cdots(x-\lambda_k)^{s_k}\qquad \text{with } 1\leq s_i\leq a_i=a_T(\lambda_i).$$

The same holds for $\mathbf{A}\in M_n(F)$ with $m_{\mathbf{A}}$ and $p_{\mathbf{A}}$. In particular, over $\mathbb{C}$, (b) applies to every $\mathbf{A}\in M_n(\mathbb{C})$.

Proof.

(a) By Theorem 9.6.1, $p_T(T)=0$, so $m_T\mid p_T$ by Theorem 9.5.1. Since p_T is monic of degree n (Theorem 7.10.2), $\deg m_T\leq n$ by Proposition 6.1.2 (c), and $\deg m_T\geq 1$ by Theorem 9.5.1, as $V\neq\{0\}$.

(b) A polynomial $x-\lambda$ of degree 1 is irreducible: if $x-\lambda=gh$, then $\deg g+\deg h=1$ (Theorem 0.8.3), so one factor is constant. The $x-\lambda_i$ are distinct and monic. By (a) and Lemma 9.5.2, $m_T=\prod_i(x-\lambda_i)^{s_i}$ with $0\leq s_i\leq a_i$. Each λ_i is a root of p_T, hence an eigenvalue (Theorem 9.2.1), hence a root of m_T (Theorem 9.5.3), so $s_i\geq 1$. Finally $a_i=\operatorname{mult}_{\lambda_i}(p_T)=a_T(\lambda_i)$ by Lemma 6.4.1, since the other factors do not vanish at λ_i. The matrix case is the operator $T_{\mathbf{A}}$, and over $\mathbb{C}$ every $p_{\mathbf{A}}$ splits (Corollary 6.5.5). ∎

This pays off a promise from Chapter 6: the bound n^2 of Theorem 6.7.2 improves to n, and p_T is an explicit annihilating polynomial of degree exactly n. It also makes Method 2 of the previous section a finite search that always succeeds. For a matrix with $p_{\mathbf{A}}=(x-2)^2(x-3)$, the corollary leaves exactly two candidates, $(x-2)(x-3)$ and $(x-2)^2(x-3)$, and one matrix product decides between them.

Warning. **Cayley–Hamilton gives an annihilating polynomial, not the minimal one.** From $p_{\mathbf{A}}(\mathbf{A}) = \mathbf{O}$ we may **not** conclude $m_{\mathbf{A}} = p_{\mathbf{A}}$; only $m_{\mathbf{A}} \mid p_{\mathbf{A}}$. For $\mathbf{A} = \operatorname{diag}(1,1,2)$, $p_{\mathbf{A}} = (x-1)^2(x-2)$ kills $\mathbf{A}$, and so does the proper divisor $(x-1)(x-2) = m_{\mathbf{A}}$. Nor may we conclude that every divisor of $p_{\mathbf{A}}$ with the right roots kills $\mathbf{A}$: for $\mathbf{J} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$, $x-1$ has the right root and does not.

The second consequence expresses inverses through powers.

Corollary 9.6.3. The Inverse Is a Polynomial in the Matrix

Let $n \geq 1$, $\mathbf{A} \in M_n(F)$, and $p_{\mathbf{A}} = x^n + c_{n-1}x^{n-1} + \cdots + c_1 x + c_0$. Then $c_0 = (-1)^n \det \mathbf{A}$. If $\mathbf{A}$ is invertible, then $c_0 \neq 0$ and

$$\mathbf{A}^{-1} = -\frac{1}{c_0}\left(\mathbf{A}^{n-1} + c_{n-1}\mathbf{A}^{n-2} + \cdots + c_2\mathbf{A} + c_1\mathbf{I}\right).$$

(For $n = 1$ the bracket is $\mathbf{I}$ alone, and the formula reads $\mathbf{A}^{-1} = -c_0^{-1}\mathbf{I}$.) The same holds for an invertible operator T on a space of dimension $n \geq 1$, with id_V in place of $\mathbf{I}$.

Proof. By Theorem 7.10.2, $c_0 = (-1)^n \det \mathbf{A}$, which is nonzero when $\mathbf{A}$ is invertible (Theorem 7.5.3). By Theorem 9.6.1,

$$\mathbf{A}\left(\mathbf{A}^{n-1} + c_{n-1}\mathbf{A}^{n-2} + \cdots + c_1\mathbf{I}\right) = \mathbf{A}^n + c_{n-1}\mathbf{A}^{n-1} + \cdots + c_1\mathbf{A} = -c_0\mathbf{I}.$$

Multiplying by $-c_0^{-1}$, the matrix in parentheses times $-c_0^{-1}$ is a right inverse of $\mathbf{A}$, hence the inverse by Theorem 3.4.5. For an operator, apply this to $[T]_{\mathcal{B}}^{\mathcal{B}}$ and use Corollary 4.1.3. ∎

So inverting an $n \times n$ matrix never requires more than its first $n-1$ powers and the coefficients of $p_{\mathbf{A}}$. The theorem also reduces **every** power of $\mathbf{A}$ to the first $n-1$. Divide x^N by $p_{\mathbf{A}}$ (Theorem 6.1.1): $x^N = q\,p_{\mathbf{A}} + r$ with $\deg r < n$. Then by Theorem 6.7.1 and Theorem 9.6.1,

$$\mathbf{A}^N = q(\mathbf{A})\,p_{\mathbf{A}}(\mathbf{A}) + r(\mathbf{A}) = r(\mathbf{A}).$$

When $p_{\mathbf{A}}$ has n distinct roots $\lambda_1, \ldots, \lambda_n$, the remainder can be found without long division. Substituting $x = \lambda_i$ in $x^N = q\,p_{\mathbf{A}} + r$ gives $r(\lambda_i) = \lambda_i^N$ for each i, and these n conditions determine $r \in F[x]_{\leq n-1}$ by Lagrange interpolation (Theorem 6.6.1). This substitution is legal: λ_i is a **scalar** plugged into an identity of polynomials.

Example. Cayley–Hamilton for a 3×3 Matrix

Let

$$\mathbf{A} = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix} \in M_3(\mathbb{Q}).$$

Compute $p_{\mathbf{A}}$ and verify $p_{\mathbf{A}}(\mathbf{A}) = \mathbf{O}$. Then use it to find $\mathbf{A}^{-1}$ and $\mathbf{A}^5$, and write down the matrices $\mathbf{B}_k$ of the proof.

Solution. Characteristic polynomial. Expand $\det(x\mathbf{I} - \mathbf{A})$ along the first row (Theorem 7.6.1 over $\mathbb{Q}[x]$):

$$p_{\mathbf{A}} = \det \begin{pmatrix} x-1 & -1 & 0 \\ 0 & x & -1 \\ -1 & 0 & x-1 \end{pmatrix} = (x-1)\det \begin{pmatrix} x & -1 \\ 0 & x-1 \end{pmatrix} + (-1)(-1)^{1+2} \det \begin{pmatrix} 0 & -1 \\ -1 & x-1 \end{pmatrix}.$$

The two minors are $x(x-1)$ and -1, so $p_{\mathbf{A}} = x(x-1)^2 - 1 = x^3 - 2x^2 + x - 1$. As a check, $\operatorname{tr}\mathbf{A} = 2$ matches the coefficient -2 of x^2, and $c_0 = -1 = (-1)^3 \det\mathbf{A}$ says $\det\mathbf{A} = 1$.

Verification. Multiplying out,

$$\mathbf{A}^2 = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 0 & 1 \\ 2 & 1 & 1 \end{pmatrix}, \qquad \mathbf{A}^3 = \mathbf{A}\,\mathbf{A}^2 = \begin{pmatrix} 2 & 1 & 2 \\ 2 & 1 & 1 \\ 3 & 2 & 2 \end{pmatrix}.$$

Then $\mathbf{A}^3 - 2\mathbf{A}^2 + \mathbf{A} - \mathbf{I}$ has first row $(2-2+1-1,\ 1-2+1,\ 2-2+0) = (0,0,0)$, second row $(2-2+0,\ 1-0+0-1,\ 1-2+1) = (0,0,0)$, and third row $(3-4+1,\ 2-2+0,\ 2-2+1-1) = (0,0,0)$. So $p_{\mathbf{A}}(\mathbf{A}) = \mathbf{O}$, as the theorem predicts.

Inverse. Here $c_0 = -1$, $c_1 = 1$, $c_2 = -2$. By Corollary 9.6.3,

$$\mathbf{A}^{-1} = \mathbf{A}^2 - 2\mathbf{A} + \mathbf{I} = \begin{pmatrix} 0 & -1 & 1 \\ 1 & 1 & -1 \\ 0 & 1 & 0 \end{pmatrix}.$$

Check: the first row of $\mathbf{A}\,\mathbf{A}^{-1}$ is $(0+1,\ -1+1,\ 1-1) = (1,0,0)$, and the other rows work the same way. The entries are integers, as they must be for an integer matrix of determinant 1 (Corollary 7.6.5).

Fifth power. The relation $\mathbf{A}^3 = 2\mathbf{A}^2 - \mathbf{A} + \mathbf{I}$ lets us lower degrees step by step:

$$\mathbf{A}^4 = 2\mathbf{A}^3 - \mathbf{A}^2 + \mathbf{A} = 2(2\mathbf{A}^2 - \mathbf{A} + \mathbf{I}) - \mathbf{A}^2 + \mathbf{A} = 3\mathbf{A}^2 - \mathbf{A} + 2\mathbf{I},$$

$$\mathbf{A}^5 = 3\mathbf{A}^3 - \mathbf{A}^2 + 2\mathbf{A} = 3(2\mathbf{A}^2 - \mathbf{A} + \mathbf{I}) - \mathbf{A}^2 + 2\mathbf{A} = 5\mathbf{A}^2 - \mathbf{A} + 3\mathbf{I}.$$

(Equivalently, $x^5 = (x^2 + 2x + 3)\,p_{\mathbf{A}} + (5x^2 - x + 3)$.) Hence

$$\mathbf{A}^5 = 5\begin{pmatrix} 1 & 1 & 1 \\ 1 & 0 & 1 \\ 2 & 1 & 1 \end{pmatrix} - \begin{pmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix} + 3\mathbf{I} = \begin{pmatrix} 7 & 4 & 5 \\ 5 & 3 & 4 \\ 9 & 5 & 7 \end{pmatrix}.$$

The matrices of the proof. By (2), $\mathbf{B}_2 = \mathbf{I}$, $\mathbf{B}_1 = \mathbf{A}\mathbf{B}_2 + c_2\mathbf{I} = \mathbf{A} - 2\mathbf{I}$, and $\mathbf{B}_0 = \mathbf{A}\mathbf{B}_1 + c_1\mathbf{I} =$

$\mathbf{A}^2 - 2\mathbf{A} + \mathbf{I}$. So

$$\operatorname{adj}(x\mathbf{I} - \mathbf{A}) = x^2\mathbf{I} + x(\mathbf{A} - 2\mathbf{I}) + (\mathbf{A}^2 - 2\mathbf{A} + \mathbf{I}) = \begin{pmatrix} x^2 - x & x - 1 & 1 \\ 1 & x^2 - 2x + 1 & x - 1 \\ x & 1 & x^2 - x \end{pmatrix},$$

which one can confirm entry by entry from cofactors of $x\mathbf{I} - \mathbf{A}$; for instance the $(1,3)$-entry is the cofactor $C_{31} = \det \begin{pmatrix} -1 & 0 \\ x & -1 \end{pmatrix} = 1$. The last equation of (2), $-\mathbf{A}\mathbf{B}_0 = c_0\mathbf{I} = -\mathbf{I}$, is the statement $\mathbf{A}^{-1} = \mathbf{B}_0$ found above.

Quick check. Let $\mathbf{A} \in M_2(\mathbb{R})$ with $\operatorname{tr}\mathbf{A} = 5$ and $\det\mathbf{A} = 6$. Write $\mathbf{A}^2$ and $\mathbf{A}^3$ as combinations of $\mathbf{A}$ and $\mathbf{I}$, and write $\mathbf{A}^{-1}$ the same way.

Solution. $p_{\mathbf{A}} = x^2 - 5x + 6$, so by Theorem 9.6.1 $\mathbf{A}^2 = 5\mathbf{A} - 6\mathbf{I}$. Then $\mathbf{A}^3 = 5\mathbf{A}^2 - 6\mathbf{A} = 5(5\mathbf{A} - 6\mathbf{I}) - 6\mathbf{A} = 19\mathbf{A} - 30\mathbf{I}$. By Corollary 9.6.3 with $c_0 = 6$ and $c_1 = -5$, $\mathbf{A}^{-1} = -\frac{1}{6}(\mathbf{A} - 5\mathbf{I}) = \frac{1}{6}(5\mathbf{I} - \mathbf{A})$.

Alternatively. When p_T splits, there is a second proof that avoids determinants of polynomial matrices entirely. The next section shows that such an operator has an upper triangular matrix, and a triangular matrix reveals directly why the product $(T - \lambda_1 \operatorname{id}_V)\cdots(T - \lambda_n \operatorname{id}_V)$ is zero. That proof is set out as an exercise there. The adjugate proof has the advantage of working over every field at once, with no splitting hypothesis.

9.6.4 Exercises

9.6.4.1 A. Check your understanding

Exercise. A1

(a) State the Cayley–Hamilton Theorem for $\mathbf{A} \in M_n(F)$.

(b) True or false: $\deg m_{\mathbf{A}} \le n$ for every $\mathbf{A} \in M_n(F)$, $n \ge 1$. Justify your answer.

(c) In one sentence, why is "$p_{\mathbf{A}}(\mathbf{A}) = \det(\mathbf{A}\mathbf{I} - \mathbf{A}) = 0$" not a proof?

(d) Suppose $p_{\mathbf{A}} = (x-1)^2(x+2)$. Which polynomials can $m_{\mathbf{A}}$ be?

(e) True or false: if $p_{\mathbf{A}}(\mathbf{A}) = \mathbf{O}$, then $p_{\mathbf{A}}$ is the minimal polynomial of $\mathbf{A}$. Justify your answer.

(f) How do you write $\mathbf{A}^{-1}$ as a polynomial in $\mathbf{A}$ when $\mathbf{A}$ is invertible?

Solution.

(a) If $n \ge 1$ and $\mathbf{A} \in M_n(F)$, then $p_{\mathbf{A}}(\mathbf{A}) = \mathbf{O}$ (Theorem 9.6.1).

(b) True. By Corollary 9.6.2, $m_{\mathbf{A}} \mid p_{\mathbf{A}}$, and $\deg p_{\mathbf{A}} = n$.

(c) It substitutes a matrix into the entries of $x\mathbf{I} - \mathbf{A}$ and then equates the matrix $p_{\mathbf{A}}(\mathbf{A})$ with the scalar $\det\mathbf{O}$, neither of which is legitimate.

(d) By Corollary 9.6.2 (b), $m_{\mathbf{A}} = (x-1)^s(x+2)^t$ with $1 \le s \le 2$ and $t = 1$: so $(x-1)(x+2)$ or $(x-1)^2(x+2)$. Both occur, for $\operatorname{diag}(1,1,-2)$ and for $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \oplus (-2)$ (Proposition 9.5.6).

(e) False. $p_{\mathbf{I}_2} = (x-1)^2$ kills $\mathbf{I}_2$, but $m_{\mathbf{I}_2} = x - 1$.

(f) If $p_{\mathbf{A}} = x^n + c_{n-1}x^{n-1} + \cdots + c_0$, then $c_0 = (-1)^n \det \mathbf{A} \neq 0$ and $\mathbf{A}^{-1} = -c_0^{-1}(\mathbf{A}^{n-1} + c_{n-1}\mathbf{A}^{n-2} + \cdots + c_1\mathbf{I})$ (Corollary 9.6.3).

9.6.4.2 B. Practice

Exercise. B1: Inverse and a hundredth power

Let $\mathbf{A} = \begin{pmatrix} 3 & -2 \\ 1 & 0 \end{pmatrix} \in M_2(\mathbb{R})$.

(a) Find $p_{\mathbf{A}}$, and use the Cayley–Hamilton Theorem to compute $\mathbf{A}^{-1}$.

(b) Hence find $\mathbf{A}^{100}$ as a combination $a\mathbf{A} + b\mathbf{I}$, and write it as a matrix.

Solution.

(a) $\operatorname{tr} \mathbf{A} = 3$ and $\det \mathbf{A} = 2$, so $p_{\mathbf{A}} = x^2 - 3x + 2 = (x-1)(x-2)$. By Theorem 9.6.1, $\mathbf{A}^2 - 3\mathbf{A} + 2\mathbf{I} = \mathbf{O}$, that is, $\mathbf{A}(3\mathbf{I} - \mathbf{A}) = 2\mathbf{I}$. By Corollary 9.6.3,

$$\mathbf{A}^{-1} = \tfrac{1}{2}(3\mathbf{I} - \mathbf{A}) = \begin{pmatrix} 0 & 1 \\ -\frac{1}{2} & \frac{3}{2} \end{pmatrix}.$$

Check: $\mathbf{A}\,\mathbf{A}^{-1} = \begin{pmatrix} 0+1 & 3-3 \\ 0 & 1 \end{pmatrix} = \mathbf{I}$.

(b) By Theorem 6.1.1, $x^{100} = q\,p_{\mathbf{A}} + (ax + b)$ for some q and scalars a, b. Substituting the roots $x = 1$ and $x = 2$ of $p_{\mathbf{A}}$ gives $1 = a + b$ and $2^{100} = 2a + b$. So $a = 2^{100} - 1$ and $b = 2 - 2^{100}$. By Theorem 9.6.1, $\mathbf{A}^{100} = q(\mathbf{A})p_{\mathbf{A}}(\mathbf{A}) + a\mathbf{A} + b\mathbf{I} = a\mathbf{A} + b\mathbf{I}$, that is,

$$\mathbf{A}^{100} = (2^{100} - 1)\mathbf{A} + (2 - 2^{100})\mathbf{I} = \begin{pmatrix} 2^{101} - 1 & 2 - 2^{101} \\ 2^{100} - 1 & 2 - 2^{100} \end{pmatrix}.$$

For $k = 1$ in place of 100, the same recipe gives $a = 1$, $b = 0$, and returns $\mathbf{A}$, as it should.

Exercise. B2: Trace zero in size two

(a) Let $\mathbf{A} \in M_2(F)$ with $\operatorname{tr} \mathbf{A} = 0$. Prove that $\mathbf{A}^2 = -(\det \mathbf{A})\,\mathbf{I}$.

(b) Deduce that for all $\mathbf{X}, \mathbf{Y} \in M_2(F)$, the matrix $(\mathbf{XY} - \mathbf{YX})^2$ is a scalar multiple of $\mathbf{I}$.

(c) Show that the converse of (a) fails: give $\mathbf{A} \in M_2(\mathbb{R})$ with $\mathbf{A}^2$ scalar and $\operatorname{tr} \mathbf{A} \neq 0$.

Solution.

(a) By Theorem 9.6.1, $\mathbf{A}^2 - (\operatorname{tr} \mathbf{A})\mathbf{A} + (\det \mathbf{A})\mathbf{I} = \mathbf{O}$. With $\operatorname{tr} \mathbf{A} = 0$, this is $\mathbf{A}^2 =$

$-(\det \mathbf{A})\mathbf{I}$.

(b) By Theorem 4.5.3 applied to $T_{\mathbf{X}}, T_{\mathbf{Y}}$, or directly by $\operatorname{tr}(\mathbf{XY}) = \operatorname{tr}(\mathbf{YX})$, we have $\operatorname{tr}(\mathbf{XY}-\mathbf{YX}) = 0$. By (a), $(\mathbf{XY}-\mathbf{YX})^2 = -\det(\mathbf{XY}-\mathbf{YX})\,\mathbf{I}$.

(c) $\mathbf{A} = \mathbf{I}_2$ has $\mathbf{A}^2 = \mathbf{I}_2$ and $\operatorname{tr}\mathbf{A} = 2 \neq 0$.

Exercise. B3: Minimal polynomial from one product

Each of the real matrices

$$\mathbf{M}_1 = \begin{pmatrix} 2 & 1 & 0 \\ 0 & 2 & 0 \\ -1 & 1 & 3 \end{pmatrix}, \qquad \mathbf{M}_2 = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ -1 & 0 & 3 \end{pmatrix}$$

has characteristic polynomial $(x-2)^2(x-3)$. Find $m_{\mathbf{M}_1}$ and $m_{\mathbf{M}_2}$. Are $\mathbf{M}_1$ and $\mathbf{M}_2$ similar?

Solution. By Corollary 9.6.2 (b), each minimal polynomial is $(x-2)(x-3)$ or $(x-2)^2(x-3)$, and only the first needs testing.

For $\mathbf{M}_1$: $\mathbf{M}_1 - 2\mathbf{I} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ -1 & 1 & 1 \end{pmatrix}$ and $\mathbf{M}_1 - 3\mathbf{I} = \begin{pmatrix} -1 & 1 & 0 \\ 0 & -1 & 0 \\ -1 & 1 & 0 \end{pmatrix}$. The first row of the product $(\mathbf{M}_1 - 2\mathbf{I})(\mathbf{M}_1 - 3\mathbf{I})$ is row 2 of $\mathbf{M}_1 - 3\mathbf{I}$, namely $(0,-1,0) \neq \mathbf{0}$. So $m_{\mathbf{M}_1} = (x-2)^2(x-3)$.

For $\mathbf{M}_2$: $\mathbf{M}_2 - 2\mathbf{I} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ -1 & 0 & 1 \end{pmatrix}$ and $\mathbf{M}_2 - 3\mathbf{I} = \begin{pmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ -1 & 0 & 0 \end{pmatrix}$. The only nonzero row of the product is $-(\text{row } 1) + (\text{row } 3)$ of $\mathbf{M}_2 - 3\mathbf{I}$, which is $(1,0,0) + (-1,0,0) = \mathbf{0}$. So the product is $\mathbf{O}$, and $m_{\mathbf{M}_2} = (x-2)(x-3)$.

The minimal polynomials differ, so $\mathbf{M}_1 \not\sim \mathbf{M}_2$ by Theorem 9.5.5.

9.6.4.3 C. Going deeper

Exercise. C1: Cayley–Hamilton over a commutative ring

Let R be a commutative ring (the remark in Chapter 7, §4), and suppose, as for $F[x]$, that the polynomials in x with coefficients in R form a commutative ring $R[x]$ in which two polynomials are equal exactly when all their coefficients are equal. For $\mathbf{A} \in M_n(R)$, define $p_{\mathbf{A}} = \det(x\mathbf{I} - \mathbf{A}) \in R[x]$.

(a) Go through the proof of Theorem 9.6.1 and name, for each step, the fact it uses. Explain why each fact holds over R and $R[x]$, and conclude that $p_{\mathbf{A}}(\mathbf{A}) = \mathbf{O}$ for every $\mathbf{A} \in M_n(R)$.

(b) Let $R = \mathbb{R}[t]$ and $\mathbf{A} = \begin{pmatrix} t & 1 \\ t^2 & 0 \end{pmatrix}$. Compute $p_{\mathbf{A}}$ and verify $p_{\mathbf{A}}(\mathbf{A}) = \mathbf{O}$ directly.

(c) Let $\mathbf{A} \in M_n(\mathbb{Z})$ with $\det\mathbf{A} = \pm 1$. Using (a) with $R = \mathbb{Z}$, show that $\mathbf{A}^{-1}$ has integer entries.

Solution.

(a) *The shape of $p_{\mathbf{A}}$.* In the Leibniz sum for $\det(x\mathbf{I}-\mathbf{A})$, the identity permutation contributes $(x-a_{11})\cdots(x-a_{nn})$, of degree n with leading coefficient 1, and every other term omits at least two diagonal entries, so it has degree at most $n-2$. Hence $p_{\mathbf{A}} = x^n + c_{n-1}x^{n-1} + \cdots + c_0$ with $c_k \in R$, which is what the proof uses in place of Theorem 7.10.2; only sums and products of entries appear. *Step 1* uses the Leibniz formula over $R[x]$ (which only multiplies and adds entries) and the degree bounds for sums and products, which hold because the degree of a product of polynomials is at most the sum of the degrees; no inverse is needed for the upper bound. *Step 2* uses the adjugate identity over the commutative ring $R[x]$, which holds by the remark after Theorem 7.6.2, and coefficient comparison in $R[x]$, which is the assumption on $R[x]$. *Step 3* uses only associativity and distributivity of matrix multiplication over R and a telescoping sum. No step divides, so each works verbatim over R, and $p_{\mathbf{A}}(\mathbf{A}) = \mathbf{O}$.

(b) $x\mathbf{I}-\mathbf{A} = \begin{pmatrix} x-t & -1 \\ -t^2 & x \end{pmatrix}$, so $p_{\mathbf{A}} = x(x-t)-t^2 = x^2-tx-t^2$. Then

$$\mathbf{A}^2 = \begin{pmatrix} t^2+t^2 & t \\ t^3 & t^2 \end{pmatrix}, \qquad \mathbf{A}^2 - t\mathbf{A} - t^2\mathbf{I} = \begin{pmatrix} 2t^2-t^2-t^2 & t-t \\ t^3-t^3 & t^2-0-t^2 \end{pmatrix} = \mathbf{O}.$$

(c) The Leibniz formula for $\det(x\mathbf{I}-\mathbf{A})$ only adds and multiplies integers and x, so the coefficients c_k of $p_{\mathbf{A}}$ are integers. Regarding $\mathbf{A}$ as a rational matrix, Theorem 7.10.2 gives $c_0 = (-1)^n \det \mathbf{A} = \pm 1$. By (a), $\mathbf{A}(\mathbf{A}^{n-1} + c_{n-1}\mathbf{A}^{n-2} + \cdots + c_1\mathbf{I}) = -c_0\mathbf{I}$. Since $c_0^2 = 1$, multiplying by $-c_0$ gives $\mathbf{A}\,\mathbf{B} = \mathbf{I}$ with $\mathbf{B} = -c_0(\mathbf{A}^{n-1} + \cdots + c_1\mathbf{I}) \in M_n(\mathbb{Z})$. Regarding $\mathbf{A}$ and $\mathbf{B}$ as rational matrices, Theorem 3.4.5 gives $\mathbf{B} = \mathbf{A}^{-1}$, which therefore has integer entries. (This reproves part of Corollary 7.6.5.)

Exercise. C2: The adjugate is a polynomial in the matrix

Let $n \geq 2$, $\mathbf{A} \in M_n(F)$, $p_{\mathbf{A}} = x^n + c_{n-1}x^{n-1} + \cdots + c_0$, and let $\mathbf{B}_0, \ldots, \mathbf{B}_{n-1}$ be as in the proof of Theorem 9.6.1.

(a) Using equation (2) of the proof, show that $\mathbf{B}_0 = \mathbf{A}^{n-1} + c_{n-1}\mathbf{A}^{n-2} + \cdots + c_2\mathbf{A} + c_1\mathbf{I}$.

(b) Show that the constant matrix $\mathbf{B}_0$ equals $\operatorname{adj}(-\mathbf{A})$, and that $\operatorname{adj}(-\mathbf{A}) = (-1)^{n-1}\operatorname{adj}\mathbf{A}$.

(c) Deduce that for **every** $\mathbf{A} \in M_n(F)$, invertible or not,

$$\operatorname{adj}\mathbf{A} = (-1)^{n-1}(\mathbf{A}^{n-1} + c_{n-1}\mathbf{A}^{n-2} + \cdots + c_2\mathbf{A} + c_1\mathbf{I}).$$

(d) Check (c) for $\mathbf{S} = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \\ 1 & 0 & 1 \end{pmatrix}$, whose adjugate Chapter 7 computed (the warning after Corollary 7.6.3), given that $p_{\mathbf{S}} = x^3 - 6x^2 + 2x$.

Solution.

(a) Equation (2) with $k = n$ gives $\mathbf{B}_{n-1} = c_n\mathbf{I} = \mathbf{I}$, and for $1 \leq k \leq n-1$ it gives $\mathbf{B}_{k-1} = \mathbf{A}\mathbf{B}_k + c_k\mathbf{I}$. By downward induction on k, $\mathbf{B}_{k-1} = \mathbf{A}^{n-k} + c_{n-1}\mathbf{A}^{n-k-1} + \cdots + c_k\mathbf{I}$: this holds for $k = n$, and if it holds for k, then $\mathbf{B}_{k-2} = \mathbf{A}\mathbf{B}_{k-1} + c_{k-1}\mathbf{I} = \mathbf{A}^{n-k+1} +$

$c_{n-1}\mathbf{A}^{n-k} + \cdots + c_k\mathbf{A} + c_{k-1}\mathbf{I}$. The case $k = 1$ is the claim.

(b) Each entry of $\operatorname{adj}(x\mathbf{I} - \mathbf{A})$ is a cofactor $\pm \det \mathbf{N}$ of $x\mathbf{I} - \mathbf{A}$, and its constant coefficient is its value at 0. By Lemma 8.3.7, $(\det \mathbf{N})(0) = \det \mathbf{N}(0)$, and $\mathbf{N}(0)$ is the corresponding submatrix of $-\mathbf{A}$. So the matrix of constant coefficients, $\mathbf{B}_0$, is $\operatorname{adj}(-\mathbf{A})$. Each cofactor of $-\mathbf{A}$ is $\pm$ the determinant of an $(n-1)\times(n-1)$ submatrix of $-\mathbf{A}$, which is -1 times the corresponding submatrix of $\mathbf{A}$. Multiplying each of the $n-1$ rows by -1 multiplies the determinant by $(-1)^{n-1}$ (Theorem 7.5.1). Hence $\operatorname{adj}(-\mathbf{A}) = (-1)^{n-1}\operatorname{adj}\mathbf{A}$.

(c) By (a) and (b), $(-1)^{n-1}\operatorname{adj}\mathbf{A} = \mathbf{B}_0 = \mathbf{A}^{n-1} + c_{n-1}\mathbf{A}^{n-2} + \cdots + c_1\mathbf{I}$; multiply by $(-1)^{n-1}$. Nothing assumed $\det \mathbf{A} \neq 0$.

(d) Here $n = 3$, $c_2 = -6$, $c_1 = 2$, so the formula predicts $\operatorname{adj}\mathbf{S} = \mathbf{S}^2 - 6\mathbf{S} + 2\mathbf{I}$. Now

$$\mathbf{S}^2 = \begin{pmatrix} 8 & 10 & 18 \\ 16 & 20 & 36 \\ 2 & 2 & 4 \end{pmatrix},$$

$$\mathbf{S}^2 - 6\mathbf{S} + 2\mathbf{I} = \begin{pmatrix} 8-6+2 & 10-12 & 18-18 \\ 16-12 & 20-24+2 & 36-36 \\ 2-6 & 2 & 4-6+2 \end{pmatrix} = \begin{pmatrix} 4 & -2 & 0 \\ 4 & -2 & 0 \\ -4 & 2 & 0 \end{pmatrix},$$

which is the adjugate found in Chapter 7.

Exercise. C3: Nilpotent matrices of size n

Let F be any field, $n \geq 1$, and $\mathbf{A} \in M_n(F)$ with $\mathbf{A}^k = \mathbf{O}$ for some $k \geq 1$.

(a) Prove that $m_{\mathbf{A}} = x^s$ for some s with $1 \leq s \leq n$.

(b) Deduce that $\mathbf{A}^n = \mathbf{O}$.

(c) Give a 3×3 example with $\mathbf{A}^2 \neq \mathbf{O}$, and explain why no 3×3 matrix has $\mathbf{A}^3 \neq \mathbf{O} = \mathbf{A}^4$.

Solution.

(a) x^k annihilates $\mathbf{A}$, so $m_{\mathbf{A}} \mid x^k$ (Theorem 9.5.1). The polynomial x is monic and irreducible (degree 1), so by Lemma 9.5.2, $m_{\mathbf{A}} = x^s$ with $0 \leq s \leq k$. By Corollary 9.6.2 (a), $1 \leq s = \deg m_{\mathbf{A}} \leq n$.

(b) $x^n = x^{n-s}\,x^s$, so $m_{\mathbf{A}} \mid x^n$ and $\mathbf{A}^n = \mathbf{O}$ by Theorem 9.5.1.

(c) The matrix $\mathbf{N}$ of the Quick check in Section 5 has $\mathbf{N}^2 \neq \mathbf{O} = \mathbf{N}^3$. If $\mathbf{A} \in M_3(F)$ satisfies $\mathbf{A}^4 = \mathbf{O}$, then $\mathbf{A}$ is nilpotent and $\mathbf{A}^3 = \mathbf{O}$ by (b), so $\mathbf{A}^3 \neq \mathbf{O} = \mathbf{A}^4$ is impossible. Note that no splitting or complex numbers were needed; compare Exercise (C2: Nilpotent operators), which worked over $\mathbb{C}$.

9.7 Triangularization

Section 4 found two ways for diagonalization to fail. The field can be too small, as for a rotation of $\mathbb{R}^2$. Or there can be too few eigenvectors, as for $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$, and no change of field helps. In the second case we should ask for less. The next simplest shape after a diagonal matrix is an upper triangular one, and triangular matrices are almost as good for computation: their eigenvalues, trace, and determinant are visible on the diagonal. This section proves that an operator has a triangular matrix exactly when its characteristic polynomial splits, so every complex matrix is similar to a triangular one. The proof is the induction on dimension that Section 1 promised: take one eigenvector, pass to the quotient, and repeat.

9.7.1 Triangular matrices and chains of invariant subspaces

As in Section 2, start with the wish and unwind it. Let $\mathcal{B} = (v_1, \dots, v_n)$ be a basis of V and $T \in \mathcal{L}(V)$. Column j of $[T]_{\mathcal{B}}^{\mathcal{B}}$ holds the coordinates of Tv_j (Definition 4.1.1). The matrix is upper triangular when every entry below the diagonal is 0 (Definition 0.9.5), that is, when Tv_j is a combination of $v_1, \dots, v_j$ only. So the first basis vector must be sent to a multiple of itself, the second into the plane of the first two, and so on. In the language of Section 1, each of the growing spans must be invariant. Chapter 2 proved this as an exercise (Exercise (C1: Upper triangular matrices and invariant subspaces)), where the word *invariant* was introduced just for that problem; here it is again, with one more piece of information about the diagonal.

Theorem 9.7.1. Triangular Matrices and Chains of Invariant Subspaces

Let V be finite-dimensional with basis $\mathcal{B} = (v_1, \dots, v_n)$, let $T \in \mathcal{L}(V)$, and put $U_0 = \{0\}$ and $U_k = \operatorname{span}(v_1, \dots, v_k)$ for $1 \le k \le n$. The following are equivalent:

(a) $[T]_{\mathcal{B}}^{\mathcal{B}}$ is upper triangular;

(b) $Tv_k \in U_k$ for every $k = 1, \dots, n$;

(c) each U_k is T-invariant.

In that case, if $d_1, \dots, d_n$ are the diagonal entries of $[T]_{\mathcal{B}}^{\mathcal{B}}$, then $(T - d_k \operatorname{id}_V)(U_k) \subseteq U_{k-1}$ for every k.

Proof. Write $[T]_{\mathcal{B}}^{\mathcal{B}} = (a_{ij})$, so that $Tv_j = \sum_i a_{ij} v_i$ with unique coefficients (Definition 4.1.1, Theorem 1.5.1).

(a) $\Leftrightarrow$ (b). If $a_{ij} = 0$ for $i > j$, then $Tv_j = \sum_{i \le j} a_{ij} v_i \in U_j$. Conversely, if $Tv_j \in U_j$, it has a representation using only $v_1, \dots, v_j$; by uniqueness, $a_{ij} = 0$ for $i > j$.

(b) $\Rightarrow$ (c). For $j \le k$, $Tv_j \in U_j \subseteq U_k$, so U_k is invariant by Lemma 9.1.1.

(c) $\Rightarrow$ (b). $v_k \in U_k$, so $Tv_k \in U_k$ by invariance.

Finally, suppose (a) holds and let $u = \sum_{j \le k} b_j v_j \in U_k$. For $j < k$, $(T - d_k \operatorname{id}_V)v_j = Tv_j - d_k v_j \in U_j \subseteq U_{k-1}$. For $j = k$, $(T - d_k \operatorname{id}_V)v_k = \sum_{i \le k} a_{ik} v_i - a_{kk} v_k = \sum_{i < k} a_{ik} v_i \in U_{k-1}$, since $d_k = a_{kk}$. By linearity, $(T - d_k \operatorname{id}_V)u \in U_{k-1}$.

■

A chain $U_1 \subset U_2 \subset \cdots \subset U_n = V$ of subspaces with $\dim U_k = k$ is often called a **flag**, after the picture of a point on a line in a plane in space. The theorem says: an upper triangular matrix is the same thing as a flag of invariant subspaces, together with a basis adapted to it. The last statement says what the diagonal entry d_k measures. Modulo the smaller space U_{k-1}, the operator T acts on the new basis vector v_k as multiplication by d_k.

The operators for which such a basis exists deserve a name.

Definition 9.7.1. Triangularizable

Let V be a finite-dimensional vector space over F. An operator $T \in \mathcal{L}(V)$ is **triangularizable over** F if there **exists** a basis $\mathcal{B}$ of V such that $[T]_{\mathcal{B}}^{\mathcal{B}}$ is upper triangular. A matrix $\mathbf{A} \in M_n(F)$ is **triangularizable over** F if there **exist** an **invertible** $\mathbf{P} \in M_n(F)$ and an upper triangular $\mathbf{U} \in M_n(F)$ with $\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \mathbf{U}$.

As for diagonalizability, $\mathbf{A}$ is triangularizable over F exactly when $T_{\mathbf{A}}$ is, by Theorem 4.2.4.

Examples.

- **Diagonalizable operators** are triangularizable, since a diagonal matrix is upper triangular. So are all operators on a space of dimension 1, and the zero and identity operators.
- **Differentiation.** For D on $\mathbb{R}[x]_{\leq n}$ in the basis $(1, x, \ldots, x^n)$, the matrix is upper triangular with zero diagonal: $D(x^k) = kx^{k-1} \in \operatorname{span}(1, \ldots, x^k)$. The flag is $\mathbb{R}[x]_{\leq 0} \subset \mathbb{R}[x]_{\leq 1} \subset \cdots \subset \mathbb{R}[x]_{\leq n}$, a chain of invariant subspaces we met in Section 1. D is not diagonalizable (its only eigenvectors are the constants), but it is triangularizable.
- **Too few eigenvectors.** $\mathbf{J} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ is already triangular, with flag $\operatorname{span}(\mathbf{e}_1) \subset F^2$.

Nonexample by minimal change. Reverse the basis for D on $\mathbb{R}[x]_{\leq 2}$: take $(x^2, x, 1)$. The spans are $\operatorname{span}(x^2) \subset \operatorname{span}(x^2, x) \subset \mathbb{R}[x]_{\leq 2}$, still a chain of subspaces of dimensions $1, 2, 3$. What fails is invariance of the first one: $D(x^2) = 2x \notin \operatorname{span}(x^2)$. Accordingly, the matrix in this basis is

$$\begin{pmatrix} 0 & 0 & 0 \\ 2 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix},$$

which is lower triangular, not upper. Being triangularizable is a property of T; being triangular is a property of T **and** a basis.

A triangularizable operator on $V \neq \{0\}$ must have an invariant line $U_1 = \operatorname{span}(v_1)$, hence an eigenvector (Proposition 9.1.2). So the rotation R of $\mathbb{R}^2$ by a right angle, which has no invariant line (Example (A Rotation Has No Invariant Lines)), is **not** triangularizable over $\mathbb{R}$. An eigenvector to start the flag is exactly what is needed, and the next theorem shows that it is also enough, provided eigenvectors keep existing at every stage.

Quick check. Is $\begin{pmatrix} 2 & 0 \\ 1 & 2 \end{pmatrix} \in M_2(\mathbb{R})$ triangularizable over $\mathbb{R}$? If so, find a basis in which its matrix is upper triangular.

Solution. Yes. We need a first basis vector that is an eigenvector. The matrix sends $\mathbf{e}_2$ to $(0, 2) = 2\mathbf{e}_2$, so start with $\mathbf{e}_2$, and complete with $\mathbf{e}_1$. In the basis $(\mathbf{e}_2, \mathbf{e}_1)$, the image of $\mathbf{e}_2$

is $2\mathbf{e}_2$ and the image of $\mathbf{e}_1$ is $(2,1) = \mathbf{e}_2 + 2\mathbf{e}_1$, so the matrix is $\begin{pmatrix} 2 & 1 \\ 0 & 2 \end{pmatrix}$, which is upper triangular. (The standard basis gives a lower triangular matrix: $\operatorname{span}(\mathbf{e}_1)$ is not invariant.)

9.7.2 The Triangularization Theorem

If T has a triangular matrix, its characteristic polynomial is a product of linear factors, one for each diagonal entry. The converse is the real content: splitting of p_T is enough to build the flag.

Theorem 9.7.2. Triangularization Theorem

Let V be a finite-dimensional vector space over F with $\dim V = n \geq 1$, and let $T \in \mathcal{L}(V)$. Then T is triangularizable over F if and only if p_T splits over F. The same holds for $\mathbf{A} \in M_n(F)$ and $p_{\mathbf{A}}$.

Idea. Take one eigenvector, extend, recurse. Since p_T splits, it has a root in F, so T has an eigenvector v_1. Put $U = \operatorname{span}(v_1)$ and extend v_1 to a basis $\mathcal{B}$, with $\bar{\mathcal{B}}$ the basis it induces on V/U. By Theorem 9.1.3 the matrix looks like

$$[T]_{\mathcal{B}}^{\mathcal{B}} = \begin{pmatrix} \lambda & * \\ \mathbf{0} & [\bar{T}]_{\bar{\mathcal{B}}}^{\bar{\mathcal{B}}} \end{pmatrix},$$

where $\bar{T}$ is the induced operator on the smaller space V/U. The first column is already right. If $[\bar{T}]_{\bar{\mathcal{B}}}^{\bar{\mathcal{B}}}$ could be made triangular, the whole matrix would be. Now $\bar{T}$ lives on a space of dimension $n-1$, and $p_T = (x-\lambda)\,p_{\bar{T}}$, so $p_{\bar{T}}$ splits too: this is the induction on dimension. The only technical point is that a basis of V/U lifts to vectors that, together with v_1, form a basis of V.

Proof. ($\Rightarrow$) Let $[T]_{\mathcal{B}}^{\mathcal{B}}$ be upper triangular with diagonal entries $d_1, \ldots, d_n$. The matrix $x\mathbf{I}_n - [T]_{\mathcal{B}}^{\mathcal{B}}$ is upper triangular with diagonal entries $x - d_i$, and Theorem 7.4.6 holds over $F[x]$ (the remark at the end of Chapter 7, §4). So $p_T = (x-d_1)\cdots(x-d_n)$ by Definition 7.10.2, which splits (Definition 6.4.2).

($\Leftarrow$) We prove, by induction on $n \geq 1$, that every operator on an n-dimensional space whose characteristic polynomial splits is triangularizable. For $n = 1$, every 1×1 matrix is upper triangular.

Let $n \geq 2$, and assume the statement for spaces of dimension $n-1$. Since p_T is monic of degree n (Theorem 7.10.2) and splits, $p_T = (x-c_1)(x-c_2)\cdots(x-c_n)$ with $c_i \in F$. As $p_T(c_1) = 0$, Theorem 9.2.1 gives an eigenvector v_1 with $Tv_1 = c_1 v_1$. Let $U = \operatorname{span}(v_1)$. It is T-invariant (Proposition 9.1.2), $\dim U = 1$, and $U \neq V$ since $n \geq 2$. Let $\bar{T}$ be the induced operator on V/U (Definition 9.1.2), a space of dimension $n - 1 \geq 1$ (Theorem 4.4.5).

$p_{\bar{T}}$ splits. The restriction $T|_U$ is $c_1\,\mathrm{id}_U$, whose matrix in the basis (v_1) is (c_1), so $p_{T|_U} = x - c_1$. By Theorem 9.1.3 (b),

$$(x-c_1)(x-c_2)\cdots(x-c_n) = p_T = (x-c_1)\,p_{\bar{T}}.$$

Canceling the nonzero factor $x - c_1$ (Corollary 0.8.4 (2)) gives $p_{\bar{T}} = (x-c_2)\cdots(x-c_n)$,

which splits.

Apply the induction hypothesis. There is a basis $\bar{\mathcal{C}} = (w_2 + U, \dots, w_n + U)$ of V/U with $[\bar{T}]_{\bar{\mathcal{C}}}^{\bar{\mathcal{C}}}$ upper triangular, for some $w_2, \dots, w_n \in V$.

Lift the basis. Let $\mathcal{B} = (v_1, w_2, \dots, w_n)$. Suppose $av_1 + b_2 w_2 + \cdots + b_n w_n = 0$. Passing to cosets, and using $v_1 \in U$, gives $b_2(w_2 + U) + \cdots + b_n(w_n + U) = 0 + U$ (Theorem 4.4.4). Since $\bar{\mathcal{C}}$ is a basis, all $b_i = 0$. Then $av_1 = 0$ with $v_1 \neq 0$, so $a = 0$ (Theorem 1.1.9). Thus $\mathcal{B}$ is a linearly independent list of n vectors in V, hence a basis (Theorem 1.6.4).

The matrix. $\mathcal{B}$ extends the basis (v_1) of U, and the cosets of the added vectors form $\bar{\mathcal{C}}$. By Theorem 9.1.3 (b),

$$[T]_{\mathcal{B}}^{\mathcal{B}} = \begin{pmatrix} c_1 & * \\ \mathbf{0} & [\bar{T}]_{\bar{\mathcal{C}}}^{\bar{\mathcal{C}}} \end{pmatrix},$$

where the $\mathbf{0}$ is a column of $n-1$ zeros. The block $[\bar{T}]_{\bar{\mathcal{C}}}^{\bar{\mathcal{C}}}$ is upper triangular, so $[T]_{\mathcal{B}}^{\mathcal{B}}$ is upper triangular. This completes the induction.

For $\mathbf{A} \in M_n(F)$, apply the result to $T_{\mathbf{A}}$, using Theorem 4.2.4.

■

The proof is an algorithm, and its output is a flag. The first invariant subspace is the eigenline U; the k-th is spanned by v_1 and the lifts $w_2, \dots, w_k$ of the first $k-1$ vectors of the flag for $\bar{T}$. The eigenvalue found at each stage is the next diagonal entry.

Over the complex numbers the hypothesis is automatic.

Corollary 9.7.3. Complex Operators Are Triangularizable

Every operator on a finite-dimensional complex vector space $V \neq \{0\}$ is triangularizable over $\mathbb{C}$. Every $\mathbf{A} \in M_n(\mathbb{C})$, $n \geq 1$, is similar to an upper triangular matrix in $M_n(\mathbb{C})$.

Proof. By Theorem 7.10.2, p_T is nonzero, and by Corollary 6.5.5 (a) it splits over $\mathbb{C}$. Apply Theorem 9.7.2.

■

So over $\mathbb{C}$ the question of Chapter 4 always has at least this answer: a basis exists in which the matrix has zeros below the diagonal. A real matrix can always be triangularized as a complex matrix, but perhaps only with complex entries.

Warning. **Triangularizing may require leaving the field, and the triangular form is far from unique.** The rotation $\mathbf{R} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ has $p_{\mathbf{R}} = x^2 + 1$, which does not split over $\mathbb{R}$, so no **real** invertible $\mathbf{P}$ makes $\mathbf{P}^{-1}\mathbf{R}\mathbf{P}$ triangular; over $\mathbb{C}$ it becomes $\operatorname{diag}(i, -i)$. And a triangularizable matrix has many triangular forms. $\begin{pmatrix} 1 & 1 \\ 0 & 2 \end{pmatrix}$ and $\begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}$ are both triangular forms of the same operator (the first has distinct eigenvalues $1, 2$, so it is diagonalizable, and similar to the second). The off-diagonal entries change with the flag and the basis, and even the order of the diagonal entries is not fixed. Only the diagonal entries, as a list with

repetitions, are determined, as the next theorem shows.

9.7.3 What the diagonal records

A diagonalization displays the eigenvalues. A triangularization does too, with their algebraic multiplicities.

Theorem 9.7.4. The Diagonal of a Triangular Form

Let V be finite-dimensional with $\dim V = n \geq 1$, $T \in \mathcal{L}(V)$, and let $\mathcal{B}$ be a basis with $[T]_{\mathcal{B}}^{\mathcal{B}}$ upper triangular, with diagonal entries $d_1, \ldots, d_n$. Then:

(a) $p_T = (x - d_1)(x - d_2)\cdots(x - d_n)$;

(b) $\operatorname{spec}(T) = \{d_1, \ldots, d_n\}$, and each eigenvalue λ occurs among $d_1, \ldots, d_n$ exactly $a_T(\lambda)$ times;

(c) $\operatorname{tr} T = d_1 + \cdots + d_n$ and $\det T = d_1 \cdots d_n$.

In particular, any two triangular forms of T have the same diagonal entries, up to order.

Proof.

(a) was shown in the proof of Theorem 9.7.2, ($\Rightarrow$).

(b) By Theorem 9.2.1, the eigenvalues are the roots of p_T in F, and λ is a root of $\prod_i (x - d_i)$ exactly when $\prod_i (\lambda - d_i) = 0$, that is, when $\lambda = d_i$ for some i (Theorem 0.7.1). By Theorem 6.4.2, $a_T(\lambda) = \operatorname{mult}_\lambda(p_T) = \sum_i \operatorname{mult}_\lambda(x - d_i)$. Each term is 1 if $d_i = \lambda$ and 0 otherwise, by Theorem 6.1.3 (b), since $x - d_i$ is not divisible by $(x - \lambda)^2$ for degree reasons. So the sum counts the indices with $d_i = \lambda$.

(c) By Definition 4.5.2, $\operatorname{tr} T = \operatorname{tr}[T]_{\mathcal{B}}^{\mathcal{B}} = \sum_i d_i$. By Definition 7.5.1 and Theorem 7.4.6, $\det T = \det[T]_{\mathcal{B}}^{\mathcal{B}} = \prod_i d_i$.

The final statement follows from (b): the number of times λ occurs on the diagonal is $a_T(\lambda)$, which does not depend on the basis.

■

Part (c) recovers Theorem 9.3.3. In Section 3 it came from comparing coefficients of p_T; here it is read off a matrix. For instance, a complex 3×3 matrix with $p_{\mathbf{A}} = (x-1)^2(x+4)$ is similar to a triangular matrix with diagonal $1, 1, -4$ in some order, so $\operatorname{tr} \mathbf{A} = -2$ and $\det \mathbf{A} = -4$, however complicated $\mathbf{A}$ looks.

The theorem does **not** say that a triangular form with **distinct** diagonal entries is itself diagonal: $\begin{pmatrix} 1 & 1 \\ 0 & 2 \end{pmatrix}$ is a counterexample. What does follow is that T is then diagonalizable (Corollary 9.4.2), so some **other** basis makes its matrix diagonal. And an operator with repeated diagonal entries may or may not be diagonalizable: compare $\mathbf{I}_2$ with $\mathbf{J}$.

9.7.4 Triangularizing a matrix in practice

For a matrix $\mathbf{A} \in M_n(F)$ whose characteristic polynomial splits, the proof of Theorem 9.7.2 becomes a procedure.

1. Find an eigenvalue λ_1 and an eigenvector $\mathbf{v}_1$.
2. Extend $\mathbf{v}_1$ to a basis $(\mathbf{v}_1, \mathbf{w}_2, \dots, \mathbf{w}_n)$ of F^n, choosing the $\mathbf{w}_i$ as simple as possible, for example standard basis vectors.
3. Compute the matrix of the induced operator on $F^n/\operatorname{span}(\mathbf{v}_1)$ in the basis $(\mathbf{w}_2+U, \dots, \mathbf{w}_n+U)$. Its entries are the coordinates of $\mathbf{A}\mathbf{w}_j$ along $\mathbf{w}_2, \dots, \mathbf{w}_n$, ignoring the $\mathbf{v}_1$-coordinate.
4. Triangularize this $(n-1)\times(n-1)$ matrix in the same way, and lift each coset back to its representative.

Example. Triangularizing a Matrix with a Single Eigenvector Line

Let

$$\mathbf{A} = \begin{pmatrix} 2 & 0 & 1 \\ 0 & 2 & 1 \\ 1 & -1 & 2 \end{pmatrix} \in M_3(\mathbb{R}).$$

Show that $\mathbf{A}$ is not diagonalizable but is triangularizable over $\mathbb{R}$, and find an invertible $\mathbf{P}$ with $\mathbf{P}^{-1}\mathbf{A}\mathbf{P}$ upper triangular.

Solution. The characteristic polynomial. Expanding along the first row (Theorem 7.6.1 over $\mathbb{R}[x]$),

$$\begin{aligned} p_{\mathbf{A}} &= \det \begin{pmatrix} x-2 & 0 & -1 \\ 0 & x-2 & -1 \\ -1 & 1 & x-2 \end{pmatrix} \\ &= (x-2)\big[(x-2)^2+1\big] + (-1)\det \begin{pmatrix} 0 & x-2 \\ -1 & 1 \end{pmatrix}. \end{aligned}$$

The last determinant is $0\cdot 1-(x-2)(-1) = x-2$, so $p_{\mathbf{A}} = (x-2)\big[(x-2)^2+1-1\big] = (x-2)^3$. It splits, so $\mathbf{A}$ is triangularizable by Theorem 9.7.2.

Eigenvectors. $\mathbf{A}-2\mathbf{I} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & -1 & 0 \end{pmatrix}$ has rank 2, so $g(2) = 1 < 3 = a(2)$, and $\mathbf{A}$ is not diagonalizable (Theorem 9.4.1 (e)). The system gives $z = 0$ and $x = y$, so $E_2(\mathbf{A}) = \operatorname{span}(\mathbf{v}_1)$ with $\mathbf{v}_1 = (1,1,0)$: all eigenvectors lie on one line.

Step 1. Let $U = \operatorname{span}(\mathbf{v}_1)$, and extend to the basis $(\mathbf{v}_1, \mathbf{e}_2, \mathbf{e}_3)$ of $\mathbb{R}^3$ (the matrix with these columns has determinant 1).

Step 2: the induced operator. Express the images of $\mathbf{e}_2$ and $\mathbf{e}_3$ in the basis:

$$\mathbf{A}\mathbf{e}_2 = (0,2,-1) = 0\,\mathbf{v}_1 + 2\mathbf{e}_2 - \mathbf{e}_3, \qquad \mathbf{A}\mathbf{e}_3 = (1,1,2) = 1\,\mathbf{v}_1 + 0\,\mathbf{e}_2 + 2\mathbf{e}_3.$$

Dropping the $\mathbf{v}_1$-coordinates, the induced operator $\bar{\mathbf{A}}$ on $\mathbb{R}^3/U$ has, in the basis $\bar{\mathcal{B}} = (\mathbf{e}_2+U, \mathbf{e}_3+U)$, the matrix

$$[\bar{\mathbf{A}}]_{\bar{\mathcal{B}}}^{\bar{\mathcal{B}}} = \begin{pmatrix} 2 & 0 \\ -1 & 2 \end{pmatrix},$$

with $p_{\bar{\mathbf{A}}} = (x-2)^2$, as the proof predicts.

Step 3: triangularize $\bar{\mathbf{A}}$. Its second column says $\bar{\mathbf{A}}(\mathbf{e}_3 + U) = 2(\mathbf{e}_3 + U)$, so $\mathbf{e}_3 + U$ is an eigenvector. In the basis $(\mathbf{e}_3 + U, \mathbf{e}_2 + U)$, $\bar{\mathbf{A}}(\mathbf{e}_2 + U) = -(\mathbf{e}_3 + U) + 2(\mathbf{e}_2 + U)$, and the matrix becomes $\begin{pmatrix} 2 & -1 \\ 0 & 2 \end{pmatrix}$.

Step 4: lift. The basis $\mathcal{B} = (\mathbf{v}_1, \mathbf{e}_3, \mathbf{e}_2)$ of $\mathbb{R}^3$ gives

$$\mathbf{P} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}, \qquad \mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \begin{pmatrix} 2 & 1 & 0 \\ 0 & 2 & -1 \\ 0 & 0 & 2 \end{pmatrix}.$$

Check. $\mathbf{A}\mathbf{v}_1 = (2,2,0) = 2\mathbf{v}_1$; $\mathbf{A}\mathbf{e}_3 = (1,1,2) = \mathbf{v}_1 + 2\mathbf{e}_3$; $\mathbf{A}\mathbf{e}_2 = (0,2,-1) = -\mathbf{e}_3 + 2\mathbf{e}_2$. These are the columns $(2,0,0)$, $(1,2,0)$, $(0,-1,2)$ of the triangular matrix, read in the basis $\mathcal{B}$. Also $\det \mathbf{P} = -1 \neq 0$.

The flag is $\operatorname{span}((1,1,0)) \subset \{(a,a,b) : a,b \in \mathbb{R}\} \subset \mathbb{R}^3$. One can check the plane directly: $\mathbf{A}(a,a,b) = (2a+b,\ 2a+b,\ 2b)$ lies in it again. The diagonal is $2,2,2$, as Theorem 9.7.4 requires.

Different choices in Step 2 give different triangular matrices, with the same diagonal. Chapter 10 shows how to choose the basis so that the entries above the diagonal are as simple as possible (the Jordan form). Chapter 12 shows that over $\mathbb{C}$ with an inner product, the basis can even be chosen orthonormal (Schur's theorem).

9.7.5 Exercises

9.7.5.1 A. Check your understanding

Exercise. A1

(a) Define what it means for $\mathbf{A} \in M_n(F)$ to be triangularizable over F.

(b) State the Triangularization Theorem.

(c) For a basis $(v_1, \ldots, v_n)$, which subspaces must be invariant for $[T]_{\mathcal{B}}^{\mathcal{B}}$ to be upper triangular?

(d) True or false: every $\mathbf{A} \in M_n(\mathbb{R})$ is triangularizable over $\mathbb{R}$. Justify your answer.

(e) True or false: every triangularizable matrix is diagonalizable. Justify your answer.

(f) Let $\mathbf{A} \in M_4(\mathbb{C})$ have $p_{\mathbf{A}} = (x-3)^3(x+1)$. What are the diagonal entries of any upper triangular matrix similar to $\mathbf{A}$, and what are $\operatorname{tr}\mathbf{A}$ and $\det\mathbf{A}$?

Solution.

(a) There are an invertible $\mathbf{P} \in M_n(F)$ and an upper triangular $\mathbf{U} \in M_n(F)$ with $\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \mathbf{U}$ (Definition 9.7.1).

(b) An operator on a space of dimension $n \geq 1$ over F is triangularizable over F if and only if its characteristic polynomial splits over F (Theorem 9.7.2).

(c) All the spans $\operatorname{span}(v_1, \ldots, v_k)$, $1 \leq k \leq n$ (Theorem 9.7.1).

(d) False. The rotation $\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ has $p = x^2 + 1$, which does not split over $\mathbb{R}$.

(e) False. $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ is triangular but not diagonalizable.

(f) By Theorem 9.7.4, the diagonal is $3, 3, 3, -1$ in some order, so $\operatorname{tr} \mathbf{A} = 8$ and $\det \mathbf{A} = -27$.

9.7.5.2 B. Practice

Exercise. B1: Triangularizing matrices

For each real matrix, find an invertible $\mathbf{P}$ such that $\mathbf{P}^{-1}\mathbf{AP}$ is upper triangular, and write down $\mathbf{P}^{-1}\mathbf{AP}$.

(a) $\mathbf{A} = \begin{pmatrix} 3 & 1 \\ -1 & 1 \end{pmatrix}$.

(b) $\mathbf{A} = \begin{pmatrix} 2 & 0 & 1 \\ 1 & 3 & 1 \\ 1 & 0 & 2 \end{pmatrix}$, with $p_{\mathbf{A}} = (x-1)(x-3)^2$ (Example (A 3×3 Characteristic Polynomial)).

Solution.

(a) $\operatorname{tr} \mathbf{A} = 4$, $\det \mathbf{A} = 4$, so $p_{\mathbf{A}} = (x-2)^2$, which splits. $\mathbf{A} - 2\mathbf{I} = \begin{pmatrix} 1 & 1 \\ -1 & -1 \end{pmatrix}$ gives the eigenvector $\mathbf{v}_1 = (1, -1)$. Extend with $\mathbf{e}_2$: $\mathbf{A}\mathbf{e}_2 = (1, 1) = \mathbf{v}_1 + 2\mathbf{e}_2$. So with

$$\mathbf{P} = \begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix}, \qquad \mathbf{P}^{-1}\mathbf{AP} = \begin{pmatrix} 2 & 1 \\ 0 & 2 \end{pmatrix}.$$

(In size 2, one eigenvector already fixes the flag.)

(b) $\mathbf{A}\mathbf{e}_2 = (0, 3, 0) = 3\mathbf{e}_2$, so $\mathbf{v}_1 = \mathbf{e}_2$ is an eigenvector for 3. Let $U = \operatorname{span}(\mathbf{e}_2)$ and extend to $(\mathbf{e}_2, \mathbf{e}_1, \mathbf{e}_3)$. Then $\mathbf{A}\mathbf{e}_1 = (2, 1, 1) = \mathbf{e}_2 + 2\mathbf{e}_1 + \mathbf{e}_3$ and $\mathbf{A}\mathbf{e}_3 = (1, 1, 2) = \mathbf{e}_2 + \mathbf{e}_1 + 2\mathbf{e}_3$, so the induced operator has matrix $\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$ in the basis $(\mathbf{e}_1 + U, \mathbf{e}_3 + U)$. Its characteristic polynomial is $x^2 - 4x + 3 = (x-3)(x-1)$, with eigenvectors $(1, 1)$ for 3 and $(1, -1)$ for 1, that is, $\mathbf{e}_1 + \mathbf{e}_3 + U$ and $\mathbf{e}_1 - \mathbf{e}_3 + U$. Lifting gives the basis $(\mathbf{e}_2,\ \mathbf{e}_1 + \mathbf{e}_3,\ \mathbf{e}_1 - \mathbf{e}_3)$. Now

$$\mathbf{A}(\mathbf{e}_1 + \mathbf{e}_3) = (3, 2, 3) = 2\mathbf{e}_2 + 3(\mathbf{e}_1 + \mathbf{e}_3),$$
$$\mathbf{A}(\mathbf{e}_1 - \mathbf{e}_3) = (1, 0, -1) = \mathbf{e}_1 - \mathbf{e}_3,$$

so

$$\mathbf{P} = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & -1 \end{pmatrix}, \qquad \mathbf{P}^{-1}\mathbf{AP} = \begin{pmatrix} 3 & 2 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

with $\det \mathbf{P} = 2 \neq 0$ (expand along the second row). The diagonal $3, 3, 1$ agrees with $p_{\mathbf{A}}$.

Exercise. B2: Nilpotent matrices are strictly triangular

Let $\mathbf{A} \in M_n(\mathbb{C})$, $n \geq 1$, with $\mathbf{A}^k = \mathbf{O}$ for some $k \geq 1$. Prove that $\mathbf{A}$ is similar to a **strictly** upper triangular matrix, that is, an upper triangular matrix with zero diagonal.

Solution. By Corollary 9.7.3, $\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \mathbf{U}$ is upper triangular for some invertible $\mathbf{P}$. By Theorem 9.7.4 (b), each diagonal entry of $\mathbf{U}$ is an eigenvalue of $\mathbf{A}$. By Exercise (C2: Nilpotent operators) (a), the only eigenvalue of a nilpotent operator on a nonzero space is 0. So every diagonal entry of $\mathbf{U}$ is 0, and $\mathbf{U}$ is strictly upper triangular. (The statement holds over every field, with a different proof: use the chain $\ker \mathbf{A} \subseteq \ker \mathbf{A}^2 \subseteq \cdots \subseteq \ker \mathbf{A}^k = F^n$ to build the basis.)

Exercise. B3: Eigenvalues of powers from a triangular form

Let $\mathbf{A} \in M_n(\mathbb{C})$, $n \geq 1$, and let $\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \mathbf{U}$ be upper triangular with diagonal entries $d_1, \ldots, d_n$.

(a) Show that the product of two upper triangular matrices is upper triangular, with diagonal entries the products of the corresponding diagonal entries.

(b) Deduce that $p_{\mathbf{A}^k} = (x - d_1^k) \cdots (x - d_n^k)$ and $\operatorname{tr}(\mathbf{A}^k) = d_1^k + \cdots + d_n^k$ for every $k \geq 1$.

(c) Suppose $p_{\mathbf{A}} = (x - 1)(x + 1)(x - 2)$. Find $p_{\mathbf{A}^2}$, $\operatorname{tr} \mathbf{A}^2$, and $\det \mathbf{A}^2$.

Solution.

(a) Let $\mathbf{U}, \mathbf{U}'$ be upper triangular. The (i, j)-entry of $\mathbf{U}\mathbf{U}'$ is $\sum_l u_{il} u'_{lj}$. A term can be nonzero only if $u_{il} \neq 0$ and $u'_{lj} \neq 0$, which forces $i \leq l \leq j$. So the entry is 0 when $i > j$, and for $i = j$ only $l = i$ survives, giving $u_{ii} u'_{ii}$.

(b) $\mathbf{P}^{-1}\mathbf{A}^k\mathbf{P} = (\mathbf{P}^{-1}\mathbf{A}\mathbf{P})^k = \mathbf{U}^k$, by induction on k. By (a) and induction, $\mathbf{U}^k$ is upper triangular with diagonal entries d_i^k. By Theorem 9.7.4 (a) and (c), applied to $T_{\mathbf{A}^k}$ in the basis given by the columns of $\mathbf{P}$, $p_{\mathbf{A}^k} = \prod_i (x - d_i^k)$ and $\operatorname{tr} \mathbf{A}^k = \sum_i d_i^k$.

(c) By Theorem 9.7.4, the diagonal of $\mathbf{U}$ is $1, -1, 2$ in some order. By (b), $p_{\mathbf{A}^2} = (x - 1)(x - 1)(x - 4) = (x - 1)^2(x - 4)$, $\operatorname{tr} \mathbf{A}^2 = 1 + 1 + 4 = 6$, and $\det \mathbf{A}^2 = 1 \cdot 1 \cdot 4 = 4$. The last agrees with $\det \mathbf{A}^2 = (\det \mathbf{A})^2 = (-2)^2$.

9.7.5.3 C. Going deeper

Exercise. C1: A second proof of Cayley–Hamilton

Let V be finite-dimensional with $\dim V = n \geq 1$, and let $T \in \mathcal{L}(V)$ have a basis $\mathcal{B} = (v_1, \ldots, v_n)$ with $[T]_{\mathcal{B}}^{\mathcal{B}}$ upper triangular, with diagonal entries $d_1, \ldots, d_n$. Let $U_0 = \{0\}$ and $U_k = \operatorname{span}(v_1, \ldots, v_k)$.

(a) Let $S_k = (T - d_1 \operatorname{id}_V)(T - d_2 \operatorname{id}_V) \cdots (T - d_k \operatorname{id}_V)$. Prove by induction on k that $S_k(U_k) = \{0\}$ for $1 \leq k \leq n$.

(b) Deduce that $p_T(T) = 0$, without using the adjugate.

(c) Deduce the Cayley–Hamilton Theorem for every $\mathbf{A} \in M_n(\mathbb{C})$, and then for every $\mathbf{A} \in$

$M_n(\mathbb{R})$.

(d) Why does this argument, on its own, not prove the theorem for $\mathbf{A} \in M_2(\mathbb{F}_2)$ with $p_{\mathbf{A}} = x^2 + x + 1$?

Hint: for (a), use the last statement of Theorem 9.7.1.

Solution.

(a) By Theorem 9.7.1, each U_k is invariant and $(T - d_k \operatorname{id}_V)(U_k) \subseteq U_{k-1}$. For $k = 1$, $S_1(U_1) = (T - d_1 \operatorname{id}_V)(U_1) \subseteq U_0 = \{0\}$. Suppose $S_{k-1}(U_{k-1}) = \{0\}$ for some $k \geq 2$. Since $S_k = S_{k-1}(T - d_k \operatorname{id}_V)$, for $u \in U_k$ we get $S_k u = S_{k-1}\big((T - d_k \operatorname{id}_V)u\big) \in S_{k-1}(U_{k-1}) = \{0\}$.

(b) By Theorem 9.7.4 (a), $p_T = \prod_{i=1}^{n}(x - d_i)$, so $p_T(T) = S_n$ by Theorem 6.7.1 (b). By (a), S_n kills $U_n = V$, so $p_T(T) = 0$.

(c) Let $\mathbf{A} \in M_n(\mathbb{C})$. By Corollary 9.7.3, $T_{\mathbf{A}}$ has a basis in which its matrix is upper triangular, and (b) gives $p_{\mathbf{A}}(T_{\mathbf{A}}) = 0$, so $p_{\mathbf{A}}(\mathbf{A}) = \mathbf{O}$ by Corollary 4.1.3. For $\mathbf{A} \in M_n(\mathbb{R})$, regard $\mathbf{A}$ as an element of $M_n(\mathbb{C})$. The Leibniz formula gives the same polynomial $p_{\mathbf{A}}$ with real coefficients, and $p_{\mathbf{A}}(\mathbf{A})$ is computed from the same entries by the same sums and products; it is $\mathbf{O}$ by the complex case.

(d) Over $\mathbb{F}_2$, $x^2 + x + 1$ has no root (its values at 0 and 1 are both 1), so it does not split, and the matrix is not triangularizable over $\mathbb{F}_2$ by Theorem 9.7.2. To run the argument we would need a larger field containing $\mathbb{F}_2$ in which the polynomial splits, and we have not constructed one. The adjugate proof of Theorem 9.6.1 needs no such field. (For instance $\mathbf{A} = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}$ has this characteristic polynomial, and indeed $\mathbf{A}^2 + \mathbf{A} + \mathbf{I} = \mathbf{O}$ over $\mathbb{F}_2$.)

Exercise. C2: Real operators have small invariant subspaces

Let $\mathbf{A} \in M_n(\mathbb{R})$, $n \geq 1$, regarded also as a complex matrix.

(a) Let $\lambda = a + bi$ $(a, b \in \mathbb{R})$ be a complex eigenvalue of $\mathbf{A}$ with eigenvector $\mathbf{z} = \mathbf{x} + i\mathbf{y}$, where $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$. Show that $\mathbf{A}\mathbf{x} = a\mathbf{x} - b\mathbf{y}$ and $\mathbf{A}\mathbf{y} = b\mathbf{x} + a\mathbf{y}$.

(b) Deduce that $\operatorname{span}(\mathbf{x}, \mathbf{y}) \subseteq \mathbb{R}^n$ is $\mathbf{A}$-invariant, and has dimension 1 or 2.

(c) Deduce that every operator on a real vector space V with $1 \leq \dim V < \infty$ has an invariant subspace of dimension 1 or 2.

(d) For $\mathbf{C} = \begin{pmatrix} 0 & 0 & 2 \\ 1 & 0 & -1 \\ 0 & 1 & 2 \end{pmatrix}$, check that $\mathbf{z} = (0, -2, 1) + i(-2, 1, 0)$ is an eigenvector for i, and find the invariant plane of (b).

Solution.

(a) The eigenvalue exists by Theorem 9.2.4. Since $\mathbf{A}$ has real entries, $\mathbf{A}\mathbf{z} = \mathbf{A}\mathbf{x} + i\mathbf{A}\mathbf{y}$ with $\mathbf{A}\mathbf{x}, \mathbf{A}\mathbf{y} \in \mathbb{R}^n$. On the other hand, $\lambda\mathbf{z} = (a + bi)(\mathbf{x} + i\mathbf{y}) = (a\mathbf{x} - b\mathbf{y}) + i(b\mathbf{x} + a\mathbf{y})$. A vector of $\mathbb{C}^n$ determines its real and imaginary parts entry by entry, so $\mathbf{A}\mathbf{x} = a\mathbf{x} - b\mathbf{y}$ and $\mathbf{A}\mathbf{y} = b\mathbf{x} + a\mathbf{y}$.

(b) By (a), $\mathbf{Ax}$ and $\mathbf{Ay}$ lie in $W = \operatorname{span}(\mathbf{x}, \mathbf{y})$, so W is invariant by Lemma 9.1.1. Since $\mathbf{z} \neq \mathbf{0}$, at least one of $\mathbf{x}, \mathbf{y}$ is nonzero, so $\dim W \geq 1$, and $\dim W \leq 2$ since W is spanned by two vectors.

(c) Choose a basis $\mathcal{B}$ of V, let $\mathbf{A} = [T]_{\mathcal{B}}^{\mathcal{B}} \in M_n(\mathbb{R})$, and let W be as in (b). Let $\Phi\colon V \to \mathbb{R}^n$ be the coordinate isomorphism $v \mapsto [v]_{\mathcal{B}}$ (Corollary 2.5.5). Then $\Phi^{-1}(W)$ is a subspace of the same dimension as W, and for $v \in \Phi^{-1}(W)$, $\Phi(Tv) = \mathbf{A}\Phi(v) \in W$ (Theorem 4.1.1), so $Tv \in \Phi^{-1}(W)$.

(d) With $\mathbf{x} = (0, -2, 1)$ and $\mathbf{y} = (-2, 1, 0)$: $\mathbf{Cx} = (2, -1, 0) = -\mathbf{y}$ and $\mathbf{Cy} = (0, -2, 1) = \mathbf{x}$. So $\mathbf{Cz} = \mathbf{Cx} + i\mathbf{Cy} = -\mathbf{y} + i\mathbf{x} = i(\mathbf{x} + i\mathbf{y}) = i\mathbf{z}$, and $\mathbf{z}$ is an eigenvector for i (here $a = 0$, $b = 1$, matching (a)). The invariant plane is $\operatorname{span}((0, -2, 1), (-2, 1, 0))$, which is the plane $x + 2y + 4z = 0$. Over $\mathbb{R}$, $\mathbf{C}$ also has the eigenvalue 2 with eigenvector $(1, 0, 1)$, an invariant line.

Exercise. C3: Traces of powers detect nilpotency

Let $\mathbf{C} \in M_n(\mathbb{C})$, $n \geq 1$, with $\operatorname{tr}(\mathbf{C}^k) = 0$ for $k = 1, \ldots, n$. The goal is to prove that $\mathbf{C}^n = \mathbf{O}$.

(a) Let $d_1, \ldots, d_n$ be the diagonal entries of a triangular form of $\mathbf{C}$. Show that $d_1^k + \cdots + d_n^k = 0$ for $k = 1, \ldots, n$.

(b) Suppose some $d_i \neq 0$. Let $\mu_1, \ldots, \mu_r$ ($r \geq 1$) be the distinct nonzero values among the d_i, and let $m_j \geq 1$ be the number of indices i with $d_i = \mu_j$. Show that $\sum_{j=1}^{r} m_j \mu_j^k = 0$ for $k = 1, \ldots, r$.

(c) Show that the $r \times r$ matrix with (k, j)-entry μ_j^k is invertible, and derive a contradiction from (b).

(d) Conclude that every $d_i = 0$, that $p_{\mathbf{C}} = x^n$, and that $\mathbf{C}^n = \mathbf{O}$.

(e) Show that the statement fails over $\mathbb{F}_2$.

Hint: for (c), factor the matrix as a transposed Vandermonde matrix times a diagonal matrix.

Solution.

(a) By Corollary 9.7.3 a triangular form exists, and by Exercise (B3: Eigenvalues of powers from a triangular form) (b), $\operatorname{tr} \mathbf{C}^k = \sum_i d_i^k$, which is 0 by hypothesis.

(b) Indices with $d_i = 0$ contribute nothing to $\sum_i d_i^k$ for $k \geq 1$. Grouping the other indices by their value, $\sum_i d_i^k = \sum_j m_j \mu_j^k$. Since $r \leq n$, (a) applies for $k = 1, \ldots, r$.

(c) Let $\mathbf{M}$ be the matrix with (k, j)-entry μ_j^k. The Vandermonde matrix $\mathbf{V} = \mathbf{V}(\mu_1, \ldots, \mu_r)$ has (j, k)-entry μ_j^{k-1}, so $\mathbf{V}^\top$ has (k, j)-entry μ_j^{k-1}, and multiplying column j by μ_j gives $\mathbf{M} = \mathbf{V}^\top \operatorname{diag}(\mu_1, \ldots, \mu_r)$. Since the μ_j are distinct, $\mathbf{V}$ is invertible (Corollary 7.8.2), hence so is $\mathbf{V}^\top$ (Theorem 7.4.5); since the μ_j are nonzero, the diagonal matrix is invertible. So $\mathbf{M}$ is invertible. By (b), $\mathbf{M}(m_1, \ldots, m_r) = \mathbf{0}$, so $(m_1, \ldots, m_r) = \mathbf{0}$ in $\mathbb{C}^r$. But each m_j is a positive integer, which is nonzero in $\mathbb{C}$. This contradiction shows that no d_i is nonzero.

(d) By (c), all $d_i = 0$, so $p_{\mathbf{C}} = x^n$ by Theorem 9.7.4 (a). By Theorem 9.6.1, $\mathbf{C}^n = p_{\mathbf{C}}(\mathbf{C}) = \mathbf{O}$.

(e) Over $\mathbb{F}_2$, $\mathbf{C} = \mathbf{I}_2$ has $\operatorname{tr} \mathbf{C}^k = \operatorname{tr} \mathbf{I}_2 = 1 + 1 = 0$ for every k, but $\mathbf{C}^k = \mathbf{I}_2 \neq \mathbf{O}$. The step that fails is (c): the positive integer $m_1 = 2$ is 0 in $\mathbb{F}_2$. Compare Exercise (C2: When $\mathbf{A}$ commutes with $[\mathbf{A}, \mathbf{B}]$), where in characteristic 0 the same conclusion was reached

for a commutator by a route that avoids eigenvalues.

9.8 Primary Decomposition

Chapter 6 ended with a lemma about any polynomial that kills an operator: if it factors into coprime pieces, the space splits into the kernels of the pieces. The previous sections produced the best polynomial to feed it, the minimal polynomial, which factors into powers of distinct irreducibles. This section runs the lemma on that factorization. The result, the primary decomposition, cuts V into invariant pieces, one for each irreducible factor of m_T, and on each piece the operator has a minimal polynomial that is a single prime power. Two answers follow at once. An operator is diagonalizable exactly when its minimal polynomial is a product of distinct linear factors. And every complex matrix with $\mathbf{A}^k = \mathbf{I}$ is diagonalizable, a promise from Section 4. Finally, when m_T splits, the pieces are the spaces on which Chapter 10 builds the Jordan form.

9.8.1 Splitting the space along the minimal polynomial

Here is the question. Diagonalization splits V into eigenspaces, and it can fail. What splitting survives for **every** operator? The minimal polynomial suggests one. By Theorem 6.3.5, if $V \neq \{0\}$, then m_T (monic, of degree at least 1 by Theorem 9.5.1) factors as

$$m_T = p_1^{e_1} p_2^{e_2} \cdots p_k^{e_k},$$

with $p_1, \dots, p_k$ **distinct** monic irreducible polynomials and exponents $e_i \geq 1$. Powers of distinct irreducibles have no common factor, so the kernel splitting lemma of Chapter 6 (Theorem 6.7.5) applies to them, peeling off one factor at a time. It gives a direct sum. The new information is what happens on the pieces.

The primary decomposition splits V into one invariant piece for each irreducible factor of m_T, and on that piece the operator is killed by exactly that factor's power.

Theorem 9.8.1. Primary Decomposition Theorem

Let $V \neq \{0\}$ be a finite-dimensional vector space over F, let $T \in \mathcal{L}(V)$, and let $m_T = p_1^{e_1} \cdots p_k^{e_k}$, where $p_1, \dots, p_k$ are distinct monic irreducible polynomials and $e_i \geq 1$. Put

$$V_i := \ker p_i(T)^{e_i} \qquad (i = 1, \dots, k).$$

Then:

(a) $V = V_1 \oplus V_2 \oplus \cdots \oplus V_k$;

(b) each V_i is T-invariant and nonzero;

(c) the restriction $T|_{V_i}$ has minimal polynomial $m_{T|_{V_i}} = p_i^{e_i}$;

(d) for each i there is $h_i \in F[x]$ such that $h_i(T)$ is the projection of V onto V_i along $\bigoplus_{j \neq i} V_j$.

In words: (a) every vector is **uniquely** a sum of pieces, one from each V_i. (b) T maps each piece into itself, so T is determined by the k restrictions, and in a basis made of bases of the V_i its matrix is block diagonal (Theorem 9.1.5). (c) Nothing is wasted: on V_i the operator is killed

by $p_i^{e_i}$, and by no proper divisor of it. (d) The pieces can be computed with polynomials in T alone. The word *primary* refers to the prime powers $p_i^{e_i}$.

Idea. ① The splitting is Chapter 6's lemma: distinct irreducible powers are pairwise coprime, and their product m_T kills T. ② Invariance holds for every kernel of a polynomial in T. ③ For (c), each restriction is killed by $p_i^{e_i}$, so its minimal polynomial is some $p_i^{f_i}$ with $f_i \le e_i$. Conversely, the product of these minimal polynomials kills T piece by piece, hence on all of V, so it is a multiple of m_T. A degree count then forces $f_i = e_i$.

Proof. Coprime. Let $i \neq j$ and $d = \gcd(p_i^{e_i}, p_j^{e_j})$, which is monic. By Lemma 9.5.2, $d = p_i^f$ and $d = p_j^g$ for some $f, g \ge 0$. By the uniqueness in Theorem 6.3.5, and since $p_i \neq p_j$, this forces $f = g = 0$, so $d = 1$.

(a) and (d). These follow from the kernel splitting lemma, applied one factor at a time.
Claim. Let U be a vector space over F, let $R \in \mathcal{L}(U)$, and let $q_1, \dots, q_r \in F[x]$ be pairwise coprime with $(q_1 \cdots q_r)(R) = 0$. Then $U = \ker q_1(R) \oplus \cdots \oplus \ker q_r(R)$, and for each i the projection of U onto $\ker q_i(R)$ along the sum of the other summands is $g_i(R)$ for some $g_i \in F[x]$.
Proof. Induction on r, with U and R allowed to vary at each step. For $r = 1$, $q_1(R) = 0$ gives $\ker q_1(R) = U$, and $g_1 = 1$ works.
Let $r \ge 2$ and put $q = q_2 \cdots q_r$. Since q_1 is coprime to each of $q_2, \dots, q_r$, it is coprime to q, and $aq_1 + bq = 1$ for some $a, b \in F[x]$ (Lemma 6.6.3 (a)). As $(q_1 q)(R) = 0$, Theorem 6.7.5 (c) gives

$$U = \ker q_1(R) \oplus W, \qquad W \coloneqq \ker q(R),$$

with the projections onto the two summands along each other equal to $(bq)(R)$ and $(aq_1)(R)$.
The subspace W is R-invariant (Theorem 6.7.4 (b)); write $S = R|_W$. Because W is R-invariant, $S^j w = R^j w$ for every $j \ge 0$ and $w \in W$, hence $g(S)w = g(R)w$ for every $g \in F[x]$. In particular $q(S) = 0$. By the induction hypothesis applied to W, $S \in \mathcal{L}(W)$, and $q_2, \dots, q_r$,

$$W = \ker q_2(S) \oplus \cdots \oplus \ker q_r(S),$$

with the projection of W onto its i-th summand equal to $h_i(S)$ for some $h_i \in F[x]$. For $i \ge 2$ we have $q_i \mid q$, so $\ker q_i(R) \subseteq \ker q(R) = W$, and therefore

$$\ker q_i(S) = \{\, w \in W : q_i(R)w = 0 \,\} = \ker q_i(R) \cap W = \ker q_i(R).$$

So $U = \ker q_1(R) + \cdots + \ker q_r(R)$, and the sum is direct: if $0 = u_1 + \cdots + u_r$ with $u_i \in \ker q_i(R)$, then $u_2 + \cdots + u_r \in W$, so $u_1 = 0$ and $u_2 + \cdots + u_r = 0$ by the directness of $U = \ker q_1(R) \oplus W$, and then every $u_i = 0$ by the directness of W (Theorem 1.7.5, (a) $\Rightarrow$ (b) and back).
Finally the projections. For $v \in U$, the component in $\ker q_1(R)$ is $(bq)(R)v$. For $i \ge 2$, the component of v in $\ker q_i(R)$ is the component in $\ker q_i(S)$ of the W-part $(aq_1)(R)v$, namely $h_i(S)(aq_1)(R)v = (h_i\, aq_1)(R)v$, the middle step because $g(S)$ and $g(R)$ agree on W. Each projection is therefore a polynomial in R.

■

The polynomials $p_1^{e_1}, \dots, p_k^{e_k}$ are pairwise coprime by the paragraph above, and their product m_T annihilates T. The Claim, applied to V, T, and these polynomials, gives $V = V_1 \oplus \cdots \oplus V_k$ with each projection $h_i(T)$ a polynomial in T. This proves (a) and (d).

(b) Each V_i is the kernel of a polynomial in T, so it is T-invariant by Theorem 6.7.4 (b). That $V_i \neq \{0\}$ will follow from (c): the minimal polynomial of an operator on the zero space is 1, and $p_i^{e_i} \neq 1$.

(c) Each V_i is T-invariant by (b), so $(T|_{V_i})^j v = T^j v$ for every $j \geq 0$ and $v \in V_i$, and hence

$$g(T|_{V_i}) = g(T)|_{V_i} \qquad \text{for every } g \in F[x]. \tag{$*$}$$

Let $q_i = m_{T|_{V_i}}$. By $(*)$, $p_i^{e_i}(T|_{V_i}) = p_i^{e_i}(T)|_{V_i} = 0$, so $q_i \mid p_i^{e_i}$ (Theorem 9.5.1), and $q_i = p_i^{f_i}$ with $0 \leq f_i \leq e_i$ by Lemma 9.5.2. Put $q = q_1 \cdots q_k$. For $v_i \in V_i$, polynomials in T commute (Theorem 6.7.1 (c)), so

$$q(T)v_i = \Big(\prod_{j \neq i} q_j\Big)(T)\; q_i(T)v_i = \Big(\prod_{j \neq i} q_j\Big)(T)\; q_i(T|_{V_i})v_i = 0,$$

using $(*)$ again. By (a), every $v \in V$ is a sum of such v_i, so $q(T) = 0$, and $m_T \mid q$ by Theorem 9.5.1. Comparing degrees (Proposition 6.1.2 (c), Theorem 0.8.3),

$$\sum_{i=1}^{k} e_i \deg p_i = \deg m_T \leq \deg q = \sum_{i=1}^{k} f_i \deg p_i.$$

Since $f_i \leq e_i$ and $\deg p_i \geq 1$ for every i, each term on the right is at most the matching term on the left, so equality holds term by term, and $f_i = e_i$. Hence $m_{T|_{V_i}} = p_i^{e_i}$. This proves the theorem.

■

The theorem reduces the study of T to the study of operators whose minimal polynomial is a power of **one** irreducible. For a matrix, choose a basis of each V_i and put them together: $\mathbf{A}$ is similar to a block diagonal matrix $\mathbf{A}_1 \oplus \cdots \oplus \mathbf{A}_k$ with $m_{\mathbf{A}_i} = p_i^{e_i}$, and then $m_{\mathbf{A}} = \operatorname{lcm}(m_{\mathbf{A}_1}, \dots, m_{\mathbf{A}_k})$ by Proposition 9.5.6, applied $k-1$ times, consistent with (c).

Example. A Primary Decomposition with an Irreducible Quadratic

Let $\mathbf{C} = \begin{pmatrix} 0 & 0 & 2 \\ 1 & 0 & -1 \\ 0 & 1 & 2 \end{pmatrix} \in M_3(\mathbb{R})$. Chapter 6 showed that its minimal polynomial is $m_{\mathbf{C}} = (x-2)(x^2+1)$ (Exercise (B2: An annihilating polynomial by elimination)). Find the primary decomposition of $\mathbb{R}^3$, the projections as polynomials in $\mathbf{C}$, and the matrix of $\mathbf{C}$ in an adapted basis.

Solution. The factorization. $x - 2$ is irreducible (degree 1), and $x^2 + 1$ is irreducible over $\mathbb{R}$ because it has no real root (Theorem 6.3.1 (b)). So $k = 2$, $p_1 = x - 2$, $p_2 = x^2 + 1$, $e_1 = e_2 = 1$.

The pieces. $V_1 = \ker(\mathbf{C} - 2\mathbf{I})$: the matrix $\mathbf{C} - 2\mathbf{I} = \begin{pmatrix} -2 & 0 & 2 \\ 1 & -2 & -1 \\ 0 & 1 & 0 \end{pmatrix}$ gives $y = 0$ and $x = z$, so $V_1 = \operatorname{span}((1,0,1))$. For $V_2 = \ker(\mathbf{C}^2 + \mathbf{I})$, multiply out:

$$\mathbf{C}^2 = \begin{pmatrix} 0 & 2 & 4 \\ 0 & -1 & 0 \\ 1 & 2 & 3 \end{pmatrix}, \qquad \mathbf{C}^2 + \mathbf{I} = \begin{pmatrix} 1 & 2 & 4 \\ 0 & 0 & 0 \\ 1 & 2 & 4 \end{pmatrix}.$$

So V_2 is the plane $x + 2y + 4z = 0$, with basis $(-2,1,0)$, $(-4,0,1)$. As the theorem predicts, $\dim V_1 + \dim V_2 = 3$, and $(1,0,1) \notin V_2$ since $1 + 0 + 4 \neq 0$.

The projections. Bézout's identity for p_1, p_2: since $x^2 + 1 = (x+2)(x-2) + 5$,

$$\tfrac{1}{5}(x^2+1) - \tfrac{1}{5}(x+2)(x-2) = 1.$$

Following Theorem 6.7.5 (c) with $p = x - 2$, $q = x^2 + 1$, $a = -\frac{1}{5}(x+2)$, and $b = \frac{1}{5}$, the projection onto V_1 along V_2 is $h_1(\mathbf{C}) = b(\mathbf{C})q(\mathbf{C}) = \frac{1}{5}(\mathbf{C}^2 + \mathbf{I})$, and the projection onto V_2 along V_1 is $h_2(\mathbf{C}) = \mathbf{I} - h_1(\mathbf{C})$. Concretely,

$$h_1(\mathbf{C}) \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \frac{x + 2y + 4z}{5} \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix},$$

which lies in V_1, is $\mathbf{0}$ on V_2, and fixes $(1,0,1)$.

The adapted basis. $\mathbf{C}(1,0,1) = (2,0,2)$. On V_2, $\mathbf{C}(-2,1,0) = (0,-2,1) = -2(-2,1,0) + (-4,0,1)$ and $\mathbf{C}(-4,0,1) = (2,-5,2) = -5(-2,1,0) + 2(-4,0,1)$. So in the basis $\mathcal{B} = ((1,0,1),(-2,1,0),(-4,0,1))$,

$$[T_{\mathbf{C}}]_{\mathcal{B}}^{\mathcal{B}} = (2) \oplus \begin{pmatrix} -2 & -5 \\ 1 & 2 \end{pmatrix}.$$

The second block has trace 0 and determinant $-4 + 5 = 1$, so its characteristic polynomial is $x^2 + 1$; it is not a scalar matrix, so its minimal polynomial is also $x^2 + 1$, as (c) says. Over $\mathbb{R}$, no further splitting of V_2 is possible, since $\mathbf{C}$ has no eigenvector in V_2. Over $\mathbb{C}$, the factor $x^2 + 1 = (x - i)(x + i)$ splits, and so does the plane.

Warning. **The pieces are kernels of the powers $p_i^{e_i}$, not of the p_i.** Replacing $\ker p_i(T)^{e_i}$ by $\ker p_i(T)$ can lose most of the space. For $\mathbf{J} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$, $m_{\mathbf{J}} = (x-1)^2$, and the primary decomposition has the single piece $\ker(\mathbf{J} - \mathbf{I})^2 = F^2$; but $\ker(\mathbf{J} - \mathbf{I}) = \operatorname{span}(\mathbf{e}_1)$ is only a line. The sum of the kernels $\ker p_i(T)$ is still direct, but it equals V only when every $e_i = 1$, which is the subject of the next subsection.

Quick check. Let $T \in \mathcal{L}(V)$, $V \neq \{0\}$ finite-dimensional over $\mathbb{Q}$, satisfy $T^3 = T^2$ but $T^2 \neq T$. Which polynomials can m_T be? What are the pieces of the primary decomposition?

Solution. $x^3 - x^2 = x^2(x-1)$ annihilates T, so by Lemma 9.5.2 $m_T = x^a(x-1)^b$ with $a \le 2$, $b \le 1$, not both 0. Since $x(x-1)$ does not annihilate T, and neither do its divisors x and $x-1$ (they would give $T = 0$ or $T = \mathrm{id}_V$, both satisfying $T^2 = T$), we need $a = 2$. So $m_T = x^2$ or $m_T = x^2(x-1)$. In the first case the only piece is $\ker T^2 = V$; in the second, $V = \ker T^2 \oplus \ker(T - \mathrm{id}_V)$.

9.8.2 Diagonalizability and the minimal polynomial

When every p_i is linear and every $e_i = 1$, the pieces are kernels of $T - \lambda_i \mathrm{id}_V$, that is, eigenspaces. That is exactly diagonalizability, and it gives the criterion Chapter 6 promised.

Theorem 9.8.2. Diagonalizability via the Minimal Polynomial

Let V be a finite-dimensional vector space over F with $\dim V \ge 1$, and let $T \in \mathcal{L}(V)$. The following are equivalent:

(a) T is diagonalizable;

(b) $m_T = (x-\lambda_1)(x-\lambda_2)\cdots(x-\lambda_k)$ for some **distinct** $\lambda_1, \dots, \lambda_k \in F$, that is, m_T splits over F with no repeated root;

(c) some polynomial $(x-\mu_1)\cdots(x-\mu_r)$ with **distinct** $\mu_1, \dots, \mu_r \in F$ annihilates T.

In that case $\lambda_1, \dots, \lambda_k$ are the distinct eigenvalues of T. The same holds for $\mathbf{A} \in M_n(F)$.

Idea.

(a) $\Rightarrow$ (b): the polynomial $\prod_i (x - \lambda_i)$ over the distinct eigenvalues kills each vector of an eigenbasis, hence kills T; and m_T must contain every factor $x - \lambda_i$, because its roots are the eigenvalues. (b) $\Rightarrow$ (c) is free. (c) $\Rightarrow$ (a): m_T divides a product of distinct linear factors, so it is one, and the primary decomposition with every $e_i = 1$ splits V into eigenspaces.

Proof.

(a) $\Rightarrow$ (b). Let $\lambda_1, \dots, \lambda_k$ be the distinct eigenvalues of T, let $\mathcal{B}$ be a basis of eigenvectors (Theorem 9.4.1 (b)), and put $q = (x-\lambda_1)\cdots(x-\lambda_k)$. For $v \in \mathcal{B}$ with eigenvalue λ_j, induction on i gives $T^i v = \lambda_j^i v$, hence $q(T)v = q(\lambda_j)v = 0$, as $x - \lambda_j$ is a factor of q. So $q(T)$ and the zero operator agree on a basis, and $q(T) = 0$ by the uniqueness in Theorem 2.1.5. Hence $m_T \mid q$, and by Lemma 9.5.2 m_T is a product of some of the factors $x - \lambda_i$. Each λ_i is a root of m_T by Theorem 9.5.3, so no factor is missing, and $m_T = q$.

(b) $\Rightarrow$ (c). Take $q = m_T$.

(c) $\Rightarrow$ (a). Let $q = (x-\mu_1)\cdots(x-\mu_r)$ with distinct μ_j annihilate T. Each factor $x - \mu_j$ is monic irreducible, being of degree 1 (Theorem 0.8.3: a factorization would leave one factor constant). By Theorem 9.5.1 and Lemma 9.5.2, $m_T = (x-\lambda_1)\cdots(x-\lambda_k)$ for some distinct λ_i among the μ_j, and $k \ge 1$ since $V \neq \{0\}$. By Theorem 9.5.3, $\lambda_1, \dots, \lambda_k$

are exactly the eigenvalues of T. Apply Theorem 9.8.1 with $p_i = x - \lambda_i$ and $e_i = 1$:

$$V = \ker(T - \lambda_1 \mathrm{id}_V) \oplus \cdots \oplus \ker(T - \lambda_k \mathrm{id}_V) = E_{\lambda_1}(T) \oplus \cdots \oplus E_{\lambda_k}(T).$$

This is condition (c) of Theorem 9.4.1, so T is diagonalizable. The statement for matrices follows by applying this to $T_{\mathbf{A}}$, since $m_{\mathbf{A}} = m_{T_{\mathbf{A}}}$.

■

This is the promise of Chapter 6 kept: **diagonalizable means the minimal polynomial splits with distinct roots.** The criterion is often the fastest test, because it needs no eigenvectors, only one polynomial identity. Chapter 6 also supplied a way to detect repeated factors without finding roots: if $\gcd(m_T, m_T') = 1$, then no irreducible factor of m_T is repeated (Corollary 6.4.6 (a)). So T is diagonalizable as soon as m_T splits and is coprime to its derivative.

Condition (c) is especially useful because it does not require knowing m_T at all.

- **Projections.** If $P^2 = P$, then $x(x-1)$ annihilates P, and $0 \neq 1$ in every field. So every projection on a finite-dimensional space is diagonalizable, over every field.
- **Involutions.** If $T^2 = \mathrm{id}_V$ and the characteristic of F is not 2, then $(x-1)(x+1)$ annihilates T, and $1 \neq -1$ because $2 \neq 0$. So T is diagonalizable. In characteristic 2 this fails: over $\mathbb{F}_2$, $\mathbf{J} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ has $\mathbf{J}^2 = \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix} = \mathbf{I}$, but $m_{\mathbf{J}} = (x-1)^2$.
- **Too few eigenvectors.** $\mathbf{J}$ over any field has $m_{\mathbf{J}} = (x-1)^2$, with a repeated root, so it is not diagonalizable, confirming Section 4 without any rank computation. Likewise the matrix of Example (The Minimal Polynomial from the Eigenvalues), with $m = (x-1)(x-3)^2$, is not diagonalizable, while that of Example (The Minimal Polynomial from the First Dependence), with $m = (x-1)(x-3)$, is.

The most striking use settles the question left open in Section 4 (Exercise (C1: Matrices with $\mathbf{A}^k = \mathbf{I}$)).

Example. Complex Matrices of Finite Order Are Diagonalizable

Let $\mathbf{A} \in M_n(\mathbb{C})$, $n \geq 1$, satisfy $\mathbf{A}^k = \mathbf{I}$ for some $k \geq 1$. Show that $\mathbf{A}$ is diagonalizable, with eigenvalues among the k-th roots of unity. Show that both "over $\mathbb{C}$" and "characteristic 0" matter.

Solution. Diagonalizable. The polynomial $q = x^k - 1$ annihilates $\mathbf{A}$. By Corollary 6.5.5, $q = (x - \zeta_1) \cdots (x - \zeta_k)$ for some $\zeta_j \in \mathbb{C}$. The ζ_j are distinct: if c were a multiple root, then $q(c) = 0$ and $q'(c) = kc^{k-1} = 0$ by Theorem 6.4.5 (a). But $q(c) = 0$ means $c^k = 1$, so $c \neq 0$; and $k \neq 0$ in $\mathbb{C}$; so $kc^{k-1} \neq 0$, a contradiction. By Theorem 9.8.2 ((c) $\Rightarrow$ (a)), $\mathbf{A}$ is diagonalizable. Its eigenvalues are roots of $m_{\mathbf{A}}$, which divides q, so each eigenvalue λ satisfies $\lambda^k = 1$. (Explicitly, the ζ_j are $e^{2\pi ij/k}$, $j = 0, \ldots, k-1$.)

Over $\mathbb{R}$ it fails. The rotation $\mathbf{R} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ has $\mathbf{R}^4 = \mathbf{I}$ but is not diagonalizable over $\mathbb{R}$, since $m_{\mathbf{R}} = x^2 + 1$ does not split. The step that breaks is the splitting of $x^4 - 1 = (x-1)(x+1)(x^2+1)$.

In characteristic p it fails. Over $\mathbb{F}_p$, $\mathbf{J} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ has $\mathbf{J}^p = \begin{pmatrix} 1 & p \\ 0 & 1 \end{pmatrix} = \mathbf{I}$ (the powers of $\mathbf{J}$ computed in Section 4), but $m_{\mathbf{J}} = (x-1)^2$. The step that breaks is distinctness: the derivative of $x^p - 1$ is $px^{p-1} = 0$ in $\mathbb{F}_p$, and indeed 1 is a multiple root of $x^p - 1$.

9.8.3 Restrictions and polynomials of diagonalizable operators

The criterion in terms of a single annihilating polynomial passes easily to smaller spaces.

Corollary 9.8.3. Restrictions of Diagonalizable Operators

Let V be finite-dimensional, $T \in \mathcal{L}(V)$ diagonalizable, and U a T-invariant subspace. Then $T|_U$ is diagonalizable. Moreover $U = \bigoplus_{\lambda \in \operatorname{spec}(T)} (U \cap E_\lambda(T))$.

Proof. If $U = \{0\}$, the empty basis gives the empty (diagonal) matrix, and both statements are trivial. Let $U \neq \{0\}$. By Theorem 9.8.2 ((a) $\Rightarrow$ (b)), $m_T = (x-\lambda_1)\cdots(x-\lambda_k)$ with distinct λ_i, the eigenvalues of T. Since U is T-invariant, $(T|_U)^j u = T^j u$ for every $j \geq 0$ and $u \in U$, so $m_T(T|_U) = m_T(T)|_U = 0$ and m_T is a product of distinct linear factors annihilating $T|_U$. By Theorem 9.8.2 ((c) $\Rightarrow$ (a)), $T|_U$ is diagonalizable.

For the decomposition, Theorem 9.4.1 (c) applied to $T|_U$ gives $U = \bigoplus_\mu E_\mu(T|_U)$, over the eigenvalues μ of $T|_U$. An eigenvector of $T|_U$ is an eigenvector of T lying in U, so $E_\mu(T|_U) = U \cap E_\mu(T)$, and every such μ is in $\operatorname{spec}(T)$. For $\lambda \in \operatorname{spec}(T)$ that is not an eigenvalue of $T|_U$, $U \cap E_\lambda(T) = \{0\}$, and adding zero summands does not change a direct sum.

■

The second statement says that an invariant subspace of a diagonalizable operator is spanned by the eigenvectors it contains. It will be the key step in Section 9, where two commuting diagonalizable operators are diagonalized by a single basis.

Diagonalizability is also preserved by polynomials: if T is diagonalizable and $p \in F[x]$, then every eigenvector of T for λ is an eigenvector of $p(T)$ for $p(\lambda)$ (Exercise (B3: Eigenvalues of powers and inverses) (a)), so a basis of eigenvectors of T is also one for $p(T)$. This is the operator form of Exercise (C2: Polynomials of diagonalizable matrices). The converse fails, as that exercise shows.

9.8.4 When the minimal polynomial splits: a first look at Chapter 10

Suppose p_T splits, $p_T = (x-\lambda_1)^{a_1}\cdots(x-\lambda_k)^{a_k}$ with distinct λ_i; this always happens over $\mathbb{C}$. By Corollary 9.6.2, $m_T = (x-\lambda_1)^{s_1}\cdots(x-\lambda_k)^{s_k}$ with $1 \leq s_i \leq a_i$, and the primary decomposition reads

$$V = \ker(T - \lambda_1 \operatorname{id}_V)^{s_1} \oplus \cdots \oplus \ker(T - \lambda_k \operatorname{id}_V)^{s_k}.$$

On the i-th piece, $T|_{V_i} = \lambda_i \operatorname{id}_{V_i} + N_i$, where $N_i = (T - \lambda_i \operatorname{id}_V)|_{V_i}$ satisfies $N_i^{s_i} = 0$, because $V_i = \ker(T - \lambda_i \operatorname{id}_V)^{s_i}$. So every operator whose characteristic polynomial splits is, piece by piece, a scalar plus a nilpotent operator. Chapter 10 calls the pieces **generalized eigenspaces** and finds the best basis for each nilpotent part, which leads to the Jordan form. Here we record only their dimensions.

Corollary 9.8.4. Dimensions of the Primary Components

Let V be finite-dimensional with $\dim V \geq 1$, $T \in \mathcal{L}(V)$, and suppose p_T splits over F, with distinct eigenvalues $\lambda_1, \ldots, \lambda_k$ and $m_T = \prod_i (x - \lambda_i)^{s_i}$. Let $V_i = \ker(T - \lambda_i \operatorname{id}_V)^{s_i}$. Then

$$p_{T|_{V_i}} = (x - \lambda_i)^{\dim V_i} \qquad \text{and} \qquad \dim V_i = a_T(\lambda_i).$$

Moreover $E_{\lambda_i}(T) \subseteq V_i$, with equality if and only if $s_i = 1$.

Idea. The decomposition makes p_T the product of the characteristic polynomials of the restrictions. Each of those factors is a monic divisor of p_T, and its roots are the eigenvalues of the restriction, which the minimal polynomial $(x - \lambda_i)^{s_i}$ pins down to the single value λ_i. So the i-th factor is $(x - \lambda_i)^{\dim V_i}$, and multiplying the factors back together identifies $\dim V_i$ as the multiplicity of λ_i in p_T.

Proof. By Theorem 9.8.1, $V = V_1 \oplus \cdots \oplus V_k$ with each V_i invariant, nonzero, and $m_{T|_{V_i}} = (x - \lambda_i)^{s_i}$. By Theorem 9.1.5, $p_T = p_{T|_{V_1}} \cdots p_{T|_{V_k}}$. So each $p_{T|_{V_i}}$ is a monic divisor of p_T, and by Lemma 9.5.2 it is a product of powers of the $x - \lambda_j$. Its roots are the eigenvalues of $T|_{V_i}$ (Theorem 9.2.1), which are the roots of $m_{T|_{V_i}} = (x - \lambda_i)^{s_i}$ (Theorem 9.5.3), that is, only λ_i. So $p_{T|_{V_i}} = (x - \lambda_i)^{d_i}$, with $d_i = \deg p_{T|_{V_i}} = \dim V_i$ (Theorem 7.10.2). Then $p_T = \prod_i (x - \lambda_i)^{d_i}$, and since the λ_i are distinct, Lemma 6.4.1 gives $a_T(\lambda_i) = d_i$.

If $Tv = \lambda_i v$, then $(T - \lambda_i \operatorname{id}_V)^{s_i} v = 0$ since $s_i \geq 1$, so $E_{\lambda_i}(T) \subseteq V_i$. If $s_i = 1$, the two spaces are equal by definition. Conversely, if $E_{\lambda_i}(T) = V_i$, then $x - \lambda_i$ annihilates $T|_{V_i}$, so $m_{T|_{V_i}} = (x - \lambda_i)^{s_i}$ divides $x - \lambda_i$, and $s_i = 1$.

■

So, when p_T splits, the dimension of the i-th piece is the algebraic multiplicity $a(\lambda_i)$, and inside it the eigenspace has dimension $g(\lambda_i)$. The operator is diagonalizable exactly when all these pieces are eigenspaces, which ties together the three criteria of this chapter: $g(\lambda_i) = a(\lambda_i)$ for all i (Section 4), every $s_i = 1$ (this section), and every piece $\ker(T - \lambda_i \operatorname{id}_V)^{s_i}$ equal to $\ker(T - \lambda_i \operatorname{id}_V)$.

Warning. **A single eigenvalue gives a single piece, whatever the operator looks like.** If $p_T = (x - \lambda)^n$, then the primary decomposition is $V = \ker(T - \lambda \operatorname{id}_V)^s$, all of V, and says nothing new. For instance $2\mathbf{I}_3$, $\begin{pmatrix} 2 & 1 \\ 0 & 2 \end{pmatrix} \oplus (2)$, and the matrix $\mathbf{A}$ of Example (Triangularizing a Matrix with a Single Eigenvector Line) all have $p = (x - 2)^3$ and a primary decomposition with one piece, yet they are pairwise nonsimilar: their minimal polynomials are $x - 2$, $(x - 2)^2$

(Proposition 9.5.6), and $(x-2)^3$, the last because $(\mathbf{A}-2\mathbf{I})^2 = \begin{pmatrix} 1 & -1 & 0 \\ 1 & -1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \neq \mathbf{O}$, and equal minimal polynomials are necessary for similarity (Theorem 9.5.5). Distinguishing operators within one piece is the work of Chapter 10.

9.8.5 Exercises

9.8.5.1 A. Check your understanding

Exercise. A1

(a) State the Primary Decomposition Theorem.

(b) State the criterion for diagonalizability in terms of the minimal polynomial.

(c) True or false: if $m_{\mathbf{A}} = x(x-1)(x+1)$ over $\mathbb{Q}$, then $\mathbf{A}$ is diagonalizable over $\mathbb{Q}$. Justify your answer.

(d) True or false: if $\mathbf{A} \in M_n(\mathbb{R})$ and $\mathbf{A}^k = \mathbf{I}$ for some $k \geq 1$, then $\mathbf{A}$ is diagonalizable over $\mathbb{R}$. Justify your answer.

(e) Let $m_T = (x-2)^2(x+1)$. What are the summands of the primary decomposition, and is T diagonalizable?

(f) True or false: the restriction of a diagonalizable operator to an invariant subspace is diagonalizable. Justify your answer.

Solution.

(a) If $V \neq \{0\}$ is finite-dimensional and $m_T = p_1^{e_1} \cdots p_k^{e_k}$ with distinct monic irreducible p_i, then $V = \bigoplus_i \ker p_i(T)^{e_i}$; the summands are invariant and nonzero, T restricted to the i-th has minimal polynomial $p_i^{e_i}$, and the projections are polynomials in T (Theorem 9.8.1).

(b) T is diagonalizable if and only if m_T is a product of distinct monic linear factors over F (Theorem 9.8.2).

(c) True: $m_{\mathbf{A}}$ splits over $\mathbb{Q}$ with distinct roots $0, 1, -1$.

(d) False. The rotation by a right angle satisfies $\mathbf{R}^4 = \mathbf{I}$ but has no real eigenvalue (Example (Complex Matrices of Finite Order Are Diagonalizable)).

(e) $\ker(T - 2\,\mathrm{id}_V)^2$ and $\ker(T + \mathrm{id}_V)$. T is not diagonalizable, since m_T has the repeated root 2.

(f) True, by Corollary 9.8.3.

9.8.5.2 B. Practice

Exercise. B1: Diagonalizable or not?

Determine whether each matrix is diagonalizable over the given field. Justify your answer.

(a) $\mathbf{A}_1 = \begin{pmatrix} 2 & 1 & 1 \\ -1 & 0 & -1 \\ 1 & 1 & 2 \end{pmatrix}$ over $\mathbb{R}$ (see Exercise (B1: Computing minimal polynomials)).

(b) $\mathbf{M}_1 = \begin{pmatrix} 2 & 1 & 0 \\ 0 & 2 & 0 \\ -1 & 1 & 3 \end{pmatrix}$ over $\mathbb{R}$ (see Exercise (B3: Minimal polynomial from one product)).

(c) Any $\mathbf{A} \in M_5(\mathbb{Q})$ with $\mathbf{A}^3 = 4\mathbf{A}$.

(d) Any $\mathbf{A} \in M_2(\mathbb{R})$ with $\mathbf{A}^2 + \mathbf{I} = \mathbf{O}$, over $\mathbb{R}$ and over $\mathbb{C}$.

Solution.

(a) By Exercise (B1: Computing minimal polynomials) (d), $m_{\mathbf{A}_1} = (x-1)(x-2)$, which has distinct linear factors. Diagonalizable, by Theorem 9.8.2.

(b) By Exercise (B3: Minimal polynomial from one product), $m_{\mathbf{M}_1} = (x-2)^2(x-3)$, with a repeated root. Not diagonalizable.

(c) $x^3 - 4x = x(x-2)(x+2)$ annihilates $\mathbf{A}$, and $0, 2, -2$ are distinct in $\mathbb{Q}$. Diagonalizable, by condition (c) of Theorem 9.8.2.

(d) $m_{\mathbf{A}} \mid x^2 + 1$. Over $\mathbb{R}$, $x^2 + 1$ has no root, so it is irreducible (Theorem 6.3.1), and $m_{\mathbf{A}} = x^2 + 1$ (it is not 1 since $\mathbf{A}$ acts on $\mathbb{R}^2 \neq \{\mathbf{0}\}$). It does not split over $\mathbb{R}$, so $\mathbf{A}$ is not diagonalizable over $\mathbb{R}$. Over $\mathbb{C}$, $x^2 + 1 = (x-i)(x+i)$ with $i \neq -i$, so condition (c) holds and $\mathbf{A}$ is diagonalizable over $\mathbb{C}$.

Exercise. B2: Decomposing $\mathbb{R}^3$

Let $\mathbf{A} = \begin{pmatrix} 2 & 1 & 1 \\ 0 & 1 & -1 \\ -1 & -1 & 1 \end{pmatrix} \in M_3(\mathbb{R})$, which has $m_{\mathbf{A}} = (x-1)^2(x-2)$ (Exercise (B1: Computing minimal polynomials) (c)).

(a) Find bases of $V_1 = \ker(\mathbf{A} - \mathbf{I})^2$ and $V_2 = \ker(\mathbf{A} - 2\mathbf{I})$, and check that $\mathbb{R}^3 = V_1 \oplus V_2$.

(b) Compare V_1 with $E_1(\mathbf{A})$.

(c) Using $(x-1)^2 - x(x-2) = 1$, write the two projections as polynomials in $\mathbf{A}$, and compute them.

(d) Write the matrix of $T_{\mathbf{A}}$ in the basis formed by your bases of V_1 and V_2.

Solution.

(a) $(\mathbf{A} - \mathbf{I})^2 = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 0 & -1 \\ -1 & -1 & 0 \end{pmatrix}^2 = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 1 & 0 \\ -1 & -1 & 0 \end{pmatrix}$, whose kernel is $x + y = 0$, with basis $(-1, 1, 0)$, $(0, 0, 1)$. $\mathbf{A} - 2\mathbf{I} = \begin{pmatrix} 0 & 1 & 1 \\ 0 & -1 & -1 \\ -1 & -1 & -1 \end{pmatrix}$ gives $y + z = 0$ and $x = -y - z = 0$, so $V_2 = \operatorname{span}((0, -1, 1))$. The three vectors form a basis of $\mathbb{R}^3$: the matrix with these columns has determinant $-1 \cdot (0 \cdot 1 - 1 \cdot (-1)) - 0 + 0 = -1 \neq 0$, expanding along the

first row. So $\mathbb{R}^3 = V_1 \oplus V_2$, as Theorem 9.8.1 guarantees.

(b) $\mathbf{A} - \mathbf{I}$ has rank 2 (its third row is minus the sum of the first two, and the first two are independent), so $E_1(\mathbf{A})$ is a line, $\operatorname{span}((-1,1,0))$, strictly inside the plane V_1. This matches Corollary 9.8.4: $\dim V_1 = a(1) = 2$, $g(1) = 1$, $s_1 = 2$.

(c) With $p = (x-1)^2$, $q = x-2$, $a = 1$, $b = -x$ in Theorem 6.7.5 (c), the projection onto V_1 along V_2 is $b(\mathbf{A})q(\mathbf{A}) = -\mathbf{A}(\mathbf{A} - 2\mathbf{I})$, and onto V_2 along V_1 it is $a(\mathbf{A})p(\mathbf{A}) = (\mathbf{A} - \mathbf{I})^2$. Computing,

$$-\mathbf{A}(\mathbf{A} - 2\mathbf{I}) = \begin{pmatrix} 1 & 0 & 0 \\ -1 & 0 & 0 \\ 1 & 1 & 1 \end{pmatrix}, \qquad (\mathbf{A} - \mathbf{I})^2 = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 1 & 0 \\ -1 & -1 & 0 \end{pmatrix}.$$

They add up to $\mathbf{I}$. The columns of the first satisfy $x + y = 0$, so they lie in V_1; the columns of the second are multiples of $(0,-1,1)$, so they lie in V_2.

(d) $\mathbf{A}(-1,1,0) = (-1,1,0)$, $\mathbf{A}(0,0,1) = (1,-1,1) = -(-1,1,0)+(0,0,1)$, and $\mathbf{A}(0,-1,1) = (0,-2,2) = 2(0,-1,1)$. So the matrix is

$$\begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix} \oplus (2),$$

block diagonal as Theorem 9.1.5 predicts, with first block of minimal polynomial $(x-1)^2$.

Exercise. B3: Matrices with $\mathbf{A}^3 = \mathbf{A}$

Let F have characteristic not 2, and let $\mathbf{A} \in M_n(F)$, $n \geq 1$, satisfy $\mathbf{A}^3 = \mathbf{A}$.

(a) Prove that $\mathbf{A}$ is diagonalizable, with $\operatorname{spec}(\mathbf{A}) \subseteq \{0, 1, -1\}$.

(b) Now let $F = \mathbb{R}$. Prove that $\operatorname{rank} \mathbf{A} = \operatorname{tr}(\mathbf{A}^2)$.

Solution.

(a) $x^3 - x = x(x-1)(x+1)$ annihilates $\mathbf{A}$. The roots $0, 1, -1$ are distinct: $1 \neq 0$, $-1 \neq 0$, and $1 \neq -1$ because $2 \neq 0$. By Theorem 9.8.2 ((c) $\Rightarrow$ (a)), $\mathbf{A}$ is diagonalizable, and its eigenvalues are roots of $m_{\mathbf{A}} \mid x^3 - x$, so they lie in $\{0, 1, -1\}$.

(b) Write $\mathbf{A} = \mathbf{P}\mathbf{D}\mathbf{P}^{-1}$ with $\mathbf{D}$ diagonal, entries in $\{0, 1, -1\}$. Then $\operatorname{rank} \mathbf{A} = \operatorname{rank} \mathbf{D}$ (Proposition 4.2.6 (a)), which is the number r of nonzero diagonal entries. Also $\mathbf{A}^2 = \mathbf{P}\mathbf{D}^2\mathbf{P}^{-1}$ (Theorem 9.4.3), and $\mathbf{D}^2$ is diagonal with r entries 1 and the rest 0. By Theorem 4.2.5, $\operatorname{tr} \mathbf{A}^2 = \operatorname{tr} \mathbf{D}^2 = r$.

9.8.5.3 C. Going deeper

Exercise. C1: $\mathbf{A}^p = \mathbf{A}$ over $\mathbb{F}_p$

Let p be a prime and $\mathbf{A} \in M_n(\mathbb{F}_p)$, $n \geq 1$, with $\mathbf{A}^p = \mathbf{A}$.

(a) Let $a \in \mathbb{F}_p$ be nonzero. Show that $b \mapsto ab$ is a bijection of $\mathbb{F}_p \setminus \{0\}$, and deduce $a^{p-1} = 1$. Conclude that $a^p = a$ for **every** $a \in \mathbb{F}_p$.

(b) Deduce that $x^p - x = \prod_{a \in \mathbb{F}_p}(x - a)$ in $\mathbb{F}_p[x]$.

(c) Prove that $\mathbf{A}$ is diagonalizable over $\mathbb{F}_p$.

(d) Show that $\mathbf{J} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ satisfies $\mathbf{J}^p = \mathbf{I}$ over $\mathbb{F}_p$ but is not diagonalizable, and explain why this does not contradict (c).

Hint: for (a), compare the product of all nonzero elements with the product of all their multiples by a.

Solution.

(a) The map $b \mapsto ab$ sends nonzero elements to nonzero elements (Theorem 0.7.1), and it is injective, since $ab = ab'$ gives $b = b'$ after multiplying by a^{-1}. An injective map from the finite set $\mathbb{F}_p \setminus \{0\}$ to itself is bijective (Theorem 0.4.9). So the elements ab, for $b \neq 0$, are the nonzero elements in another order. Let c be the product of all nonzero elements; it is nonzero. Multiplying the $p - 1$ elements ab gives $a^{p-1}c = c$, by commutativity, so $a^{p-1} = 1$. Multiplying by a gives $a^p = a$, which also holds for $a = 0$.

(b) By (a), each of the p distinct elements of $\mathbb{F}_p$ is a root of $f = x^p - x$. By Theorem 6.4.3, the multiplicities of these roots, each at least 1, add up to at most $\deg f = p$. So each multiplicity is 1, the sum equals p, and f splits; as f is monic, $f = \prod_{a \in \mathbb{F}_p}(x - a)$.

(c) By (b), $x^p - x$ is a product of distinct monic linear factors, and it annihilates $\mathbf{A}$. By Theorem 9.8.2 ((c) $\Rightarrow$ (a)), $\mathbf{A}$ is diagonalizable, with eigenvalues in $\mathbb{F}_p$.

(d) As in Example (Complex Matrices of Finite Order Are Diagonalizable), $\mathbf{J}^p = \begin{pmatrix} 1 & p \\ 0 & 1 \end{pmatrix} = \mathbf{I}$ in $\mathbb{F}_p$, while $m_{\mathbf{J}} = (x - 1)^2$, so $\mathbf{J}$ is not diagonalizable. There is no contradiction: $\mathbf{J}$ satisfies $\mathbf{J}^p = \mathbf{I}$, not $\mathbf{J}^p = \mathbf{J}$ (for $p \geq 2$, $\mathbf{J}^p = \mathbf{I} \neq \mathbf{J}$). The polynomial $x^p - 1$ has the multiple root 1, because its derivative px^{p-1} is 0 in $\mathbb{F}_p$ (Theorem 6.4.5 (a)).

Exercise. C2: Irreducible minimal polynomials

Let $V \neq \{0\}$ be finite-dimensional, $T \in \mathcal{L}(V)$, and suppose m_T is **irreducible**, of degree d. For $v \in V$, let $W_v = \{q(T)v : q \in F[x]\}$.

(a) Show that W_v is the smallest T-invariant subspace of V containing v.

(b) Let $v \neq 0$. Prove that $(v, Tv, \ldots, T^{d-1}v)$ is a basis of W_v. In particular $\dim W_v = d$.

(c) Deduce that the only T-invariant subspaces of W_v are $\{0\}$ and W_v.

(d) Prove that V is a direct sum of subspaces of the form W_v, and hence that d divides $\dim V$.

(e) Illustrate with the rotation $\mathbf{R}$ of $\mathbb{R}^2$ and with $\mathbf{R} \oplus \mathbf{R}$ on $\mathbb{R}^4$.

Hint: for (b), use the gcd of m_T with a polynomial of degree less than d. For (d), take a direct sum $W_{v_1} \oplus \cdots \oplus W_{v_s}$ that is as large as possible.

Solution.

(a) W_v is the image of the linear map $F[x] \to V$, $q \mapsto q(T)v$ (linear by Theorem 6.7.1 (a)), so it is a subspace (Theorem 2.2.2). It contains $v = 1(T)v$, and $T(q(T)v) = (xq)(T)v \in$

W_v, so it is invariant. If W is any invariant subspace containing v, then $T^k v \in W$ for all k by induction, so every $q(T)v \in W$, and $W_v \subseteq W$.

(b) *Independence.* Suppose $q(T)v = 0$ for some nonzero q of degree less than d. Let $g = \gcd(q, m_T)$. It is a monic divisor of the irreducible m_T, so $g = 1$ or $g = m_T$ (Definition 6.3.1). It is not m_T, since $m_T \nmid q$ for degree reasons (Proposition 6.1.2 (c)). So $g = 1$, and $aq + bm_T = 1$ for some a, b (Corollary 6.2.4). Then $v = a(T)q(T)v + b(T)m_T(T)v = 0$, a contradiction. So no nontrivial combination of $v, \dots, T^{d-1}v$ is 0. *Spanning.* For $q \in F[x]$, divide: $q = s\, m_T + r$ with $\deg r < d$ (Theorem 6.1.1). Then $q(T)v = r(T)v$, a combination of $v, \dots, T^{d-1}v$.

(c) Let $W \subseteq W_v$ be invariant and nonzero, and take $w \in W$ with $w \neq 0$. By (a), $W_w \subseteq W$. By (b), $\dim W_w = d = \dim W_v$, so $W_w = W_v$ (Theorem 1.6.8), and hence $W = W_v$.

(d) Among all families $v_1, \dots, v_s$ of nonzero vectors for which the sum $S = W_{v_1} + \cdots + W_{v_s}$ is direct, choose one with s as large as possible; this is possible since $\dim S = sd \leq \dim V$ by (b) and Theorem 1.7.5. Suppose $S \neq V$, and take $v \in V \setminus S$. The subspace $W_v \cap S$ is invariant (Proposition 9.1.4) and lies in W_v, but does not contain v; by (c) it is $\{0\}$. Then the sum $S + W_v$ is direct: if $u_1 + \cdots + u_s + w = 0$ with $u_j \in W_{v_j}$ and $w \in W_v$, then $w \in W_v \cap S = \{0\}$, and directness of S gives all $u_j = 0$. This contradicts the maximality of s. So $V = W_{v_1} \oplus \cdots \oplus W_{v_s}$, and $\dim V = sd$.

(e) For the rotation $\mathbf{R}$, $m_{\mathbf{R}} = x^2 + 1$ is irreducible over $\mathbb{R}$, $d = 2$, and $W_{\mathbf{v}} = \mathbb{R}^2$ for every $\mathbf{v} \neq \mathbf{0}$: no line is invariant (Example (A Rotation Has No Invariant Lines)). For $\mathbf{R} \oplus \mathbf{R}$, $m = \operatorname{lcm}(x^2 + 1, x^2 + 1) = x^2 + 1$ (Proposition 9.5.6), and $\mathbb{R}^4 = W_{\mathbf{e}_1} \oplus W_{\mathbf{e}_3}$ with $W_{\mathbf{e}_1} = \operatorname{span}(\mathbf{e}_1, \mathbf{e}_2)$ and $W_{\mathbf{e}_3} = \operatorname{span}(\mathbf{e}_3, \mathbf{e}_4)$; indeed $2 \mid 4$.

9.9 Commuting Operators

So far we have diagonalized or triangularized one operator at a time. Many problems hand us several at once: two matrices that describe the same system, or a matrix together with all its powers. Can a single basis make **all** of them diagonal, or all of them triangular? A necessary condition is easy to spot, because diagonal matrices commute. This section shows that commuting is also enough, as long as each operator can be handled on its own. The whole argument rests on one observation: an operator that commutes with T maps each eigenspace of T into itself, so we can search for common eigenvectors inside an eigenspace.

9.9.1 When can two operators share a basis?

Let V be finite-dimensional and $S, T \in \mathcal{L}(V)$. Suppose one basis $\mathcal{B}$ makes both $[S]_{\mathcal{B}}^{\mathcal{B}}$ and $[T]_{\mathcal{B}}^{\mathcal{B}}$ diagonal. Diagonal matrices multiply entry by entry on the diagonal, so $[S]_{\mathcal{B}}^{\mathcal{B}}[T]_{\mathcal{B}}^{\mathcal{B}} = [T]_{\mathcal{B}}^{\mathcal{B}}[S]_{\mathcal{B}}^{\mathcal{B}}$. By Corollary 4.1.3 this says $[ST]_{\mathcal{B}}^{\mathcal{B}} = [TS]_{\mathcal{B}}^{\mathcal{B}}$, and since $T \mapsto [T]_{\mathcal{B}}^{\mathcal{B}}$ is injective (Theorem 4.1.4), $ST = TS$. So **sharing a diagonalizing basis forces commuting**. The same is not true for triangular matrices: $\begin{pmatrix} 1 & 1 \\ 0 & 2 \end{pmatrix}$ and $\begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$ are both upper triangular, and their products in the two orders are $\begin{pmatrix} 1 & 2 \\ 0 & 4 \end{pmatrix}$ and $\begin{pmatrix} 1 & 1 \\ 0 & 4 \end{pmatrix}$. We will see, though, that commuting is a convenient sufficient condition in the triangular case as well.

The question for this section is the converse: if $ST = TS$, can we find a common basis? Here are the commuting pairs to keep in mind.

- **Polynomials in one operator.** For every $p, q \in F[x]$, $p(T)q(T) = q(T)p(T)$ (Theo-

rem 6.7.1 (c)). In particular T commutes with T^2, with T^{-1} when it exists (both products are id_V), and with every scalar operator $c\,\mathrm{id}_V$.

- **Diagonal matrices** commute with each other, and so do block diagonal matrices whose corresponding blocks commute (Theorem 8.1.4).
- **A noncommuting pair.** $\mathbf{E}_{12} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$ and $\mathbf{E}_{21} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}$ give $\mathbf{E}_{12}\mathbf{E}_{21} = \mathbf{E}_{11} \neq \mathbf{E}_{22} = \mathbf{E}_{21}\mathbf{E}_{12}$.

The last pair already shows that something must be assumed. Over any field, the eigenvectors of $\mathbf{E}_{12}$ are the nonzero multiples of $\mathbf{e}_1$, and those of $\mathbf{E}_{21}$ are the nonzero multiples of $\mathbf{e}_2$. They share no eigenvector at all, let alone a basis of them.

9.9.2 Commuting operators preserve eigenspaces

How does the hypothesis $ST = TS$ interact with eigenvectors? Take $Tv = \lambda v$ and ask what T does to Sv. There is no S in the eigenvector equation yet, so we apply S to both sides.

Theorem 9.9.1. Commuting Operators Preserve Eigenspaces

Let V be a vector space over F, and let $S, T \in \mathcal{L}(V)$ with $ST = TS$. Then for every $\lambda \in F$, the eigenspace $E_\lambda(T)$ is S-invariant.

Proof. Let $v \in E_\lambda(T)$, so $Tv = \lambda v$. Since $ST = TS$ and S is linear,

$$T(Sv) = (TS)v = (ST)v = S(\lambda v) = \lambda\, Sv.$$

Hence $Sv \in E_\lambda(T)$ by Definition 9.2.2. This shows that S maps $E_\lambda(T)$ into itself. ∎

No finite dimension is needed, and λ need not be an eigenvalue (then $E_\lambda(T) = \{0\}$ and there is nothing to prove). By symmetry, every eigenspace of S is T-invariant too.

In matrix language the theorem is a statement about block shapes. Take $T = \mathrm{diag}(1, 1, 2)$ on F^3, with $E_1(T) = \mathrm{span}(\mathbf{e}_1, \mathbf{e}_2)$ and $E_2(T) = \mathrm{span}(\mathbf{e}_3)$, and let S be any matrix with $ST = TS$. Both eigenspaces are S-invariant, and $F^3 = E_1(T) \oplus E_2(T)$, so by Theorem 9.1.5 the matrix S is block diagonal:

$$S = \begin{pmatrix} s_{11} & s_{12} & 0 \\ s_{21} & s_{22} & 0 \\ 0 & 0 & s_{33} \end{pmatrix}.$$

Conversely, every matrix of this shape commutes with T, since $T = \mathbf{I}_2 \oplus (2)$ and $\mathbf{I}_2$ commutes with every 2×2 block (Theorem 8.1.4). Inside the repeated eigenspace S can be anything; this freedom is what the rest of the section has to deal with.

Nonexample. Take $T = \mathrm{diag}(1, 2)$ and the swap $S = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$. Then $S\mathbf{e}_1 = \mathbf{e}_2 \notin E_1(T) = \mathrm{span}(\mathbf{e}_1)$, so $E_1(T)$ is not S-invariant. The theorem is not contradicted: $ST = \begin{pmatrix} 0 & 2 \\ 1 & 0 \end{pmatrix}$ and $TS = \begin{pmatrix} 0 & 1 \\ 2 & 0 \end{pmatrix}$ differ.

Quick check. Let $T = \mathbf{I}_2$ and $S = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ over $\mathbb{R}$. Do S and T commute? Is every eigenvector of T an eigenvector of S? Do S and T have a common eigenvector?

Solution. Yes, $\mathbf{I}_2$ commutes with every 2×2 matrix. No: every nonzero vector is an eigenvector of $\mathbf{I}_2$, but $S\mathbf{e}_2 = (1, 1)$ is not a multiple of $\mathbf{e}_2$. Yes: $\mathbf{e}_1$ is an eigenvector of both, for the eigenvalues 1 and 1. The theorem says S maps $E_1(T) = \mathbb{R}^2$ into itself, which is true but empty of information; the common eigenvector must be **searched for** inside that eigenspace, and here it is an eigenvector of S.

9.9.3 A common eigenvector

Here is the plan for finding a common eigenvector of commuting S and T. Pick an eigenvalue λ of T and put $U = E_\lambda(T) \neq \{0\}$. By Theorem 9.9.1, S restricts to an operator $S|_U$ on U. If $S|_U$ has an eigenvector u, then $Su = \mu u$ and, since $u \in U$, also $Tu = \lambda u$. So everything reduces to whether $S|_U$ has an eigenvalue. Over $\mathbb{C}$ it does (Theorem 9.2.4). Over a general field we need to know that the characteristic polynomial of a restriction splits when that of S does. We record this, together with the same fact for the induced operator on the quotient, since triangularization will need both.

Lemma 9.9.2. Restrictions and Quotients Inherit Splitting

Let V be a finite-dimensional vector space over F, let $T \in \mathcal{L}(V)$, and let U be a T-invariant subspace with $\{0\} \neq U \neq V$. If p_T splits over F, then $p_{T|_U}$ and $p_{\bar{T}}$ split over F, where $\bar{T}$ is the induced operator on V/U.

Idea. Section 1 factors p_T as $p_{T|_U}\, p_{\bar{T}}$, so the two factors can only have roots that p_T has. A polynomial splits exactly when its root multiplicities add up to its degree (Theorem 6.4.3). Count: the multiplicities of the two factors add up to those of p_T, which use up the whole degree; so neither factor can be short of its own degree.

Proof. By Theorem 9.1.3, $p_T = fg$ with $f = p_{T|_U}$ and $g = p_{\bar{T}}$, both monic (Theorem 7.10.2). Let $c_1, \dots, c_k$ be the distinct roots of p_T in F. Every root of f or of g in F is a root of p_T (Theorem 0.8.5), so it is among the c_i, and a c_i that is not a root of f has $\operatorname{mult}_{c_i}(f) = 0$. Hence, by Theorem 6.4.3 applied to f and to g,

$$\sum_{i=1}^{k} \operatorname{mult}_{c_i}(f) \leq \deg f, \qquad \sum_{i=1}^{k} \operatorname{mult}_{c_i}(g) \leq \deg g.$$

Adding, and using Theorem 6.4.2 and then Theorem 6.4.3 for the split polynomial p_T,

$$\begin{aligned}\deg f + \deg g = \deg p_T &= \sum_{i=1}^{k} \operatorname{mult}_{c_i}(p_T) \\ &= \sum_{i=1}^{k} \operatorname{mult}_{c_i}(f) + \sum_{i=1}^{k} \operatorname{mult}_{c_i}(g) \\ &\leq \deg f + \deg g.\end{aligned}$$

So both inequalities above are equalities, and f and g split over F by Theorem 6.4.3.

■

Now the common eigenvector. We prove it for a whole family of commuting operators at once, possibly infinite, since the proof costs no more and the next theorems need it. A family $\mathcal{F} \subseteq \mathcal{L}(V)$ is **commuting** if $ST = TS$ for all $S, T \in \mathcal{F}$.

Theorem 9.9.3. Commuting Operators Have a Common Eigenvector

Let $V \neq \{0\}$ be a finite-dimensional vector space over F, and let $\mathcal{F} \subseteq \mathcal{L}(V)$ be a nonempty commuting family such that p_T splits over F for every $T \in \mathcal{F}$. Then there is a nonzero $v \in V$ that is an eigenvector of **every** $T \in \mathcal{F}$. This applies to every commuting family when $F = \mathbb{C}$.

Idea. For two operators, the plan above works: an eigenvector of $S|_U$ with $U = E_\lambda(T)$. For many operators we would like to repeat it, shrinking U each time, and a family may be infinite. So we jump straight to the end of the shrinking: take a nonzero subspace W, invariant under the whole family, that is as small as possible. If some $T \in \mathcal{F}$ were not a scalar on W, an eigenspace of $T|_W$ would be a smaller such subspace. So every operator in the family is a scalar on W, and any nonzero vector of W is a common eigenvector.

Proof. Call a subspace **$\mathcal{F}$-invariant** if it is T-invariant for every $T \in \mathcal{F}$. The space V is a nonzero $\mathcal{F}$-invariant subspace, so among all nonzero $\mathcal{F}$-invariant subspaces there is one, W, of smallest dimension.

Let $T \in \mathcal{F}$. Then $T|_W$ is an operator on $W \neq \{0\}$, and $p_{T|_W}$ splits: if $W = V$ this is the hypothesis, and otherwise it is Lemma 9.9.2. Since $p_{T|_W}$ is monic of degree $\dim W \geq 1$, it has a root $\lambda \in F$, which is an eigenvalue of $T|_W$ by Theorem 9.2.1. Put

$$W' = \{w \in W : Tw = \lambda w\} = W \cap E_\lambda(T),$$

which is nonzero because it contains an eigenvector of $T|_W$. For every $S \in \mathcal{F}$, both W and $E_\lambda(T)$ are S-invariant, the first by the choice of W and the second by Theorem 9.9.1; so W' is S-invariant by Proposition 9.1.4. Hence W' is a nonzero $\mathcal{F}$-invariant subspace contained in W, and by the minimality of $\dim W$, $\dim W' = \dim W$. By Theorem 1.6.8, $W' = W$, that is, $Tw = \lambda w$ for every $w \in W$.

So every $T \in \mathcal{F}$ acts on W as multiplication by a scalar λ_T. Any nonzero $v \in W$ then satisfies $Tv = \lambda_T v$ for every $T \in \mathcal{F}$, as claimed. Over $\mathbb{C}$ every characteristic polynomial splits (Corollary 6.5.5), so the hypothesis holds automatically.

■

The proof never had to name the eigenvalues. It found a subspace on which every operator of the family is a scalar, and the scalars came for free.

Warning. **The splitting hypothesis cannot be dropped, and neither can commuting.** Over $\mathbb{R}$, the rotation $\mathbf{R} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ commutes with itself, yet the family $\{\mathbf{R}\}$ has no real eigenvector at all, since $p_{\mathbf{R}} = x^2 + 1$ has no real root. Over $\mathbb{C}$, the noncommuting pair $\mathbf{E}_{12}, \mathbf{E}_{21}$ has no common eigenvector, as we saw above.

9.9.4 Simultaneous triangularization

Section 7 triangularized one operator by splitting off an eigenvector and passing to the quotient. For a commuting family the same recursion works, because the eigenvector we split off can be chosen common to all.

Theorem 9.9.4. Simultaneous Triangularization

Let V be a finite-dimensional vector space over F with $\dim V = n \geq 1$, and let $\mathcal{F} \subseteq \mathcal{L}(V)$ be a nonempty commuting family such that p_T splits over F for every $T \in \mathcal{F}$. Then there is **one** basis $\mathcal{B}$ of V such that $[T]_{\mathcal{B}}^{\mathcal{B}}$ is upper triangular for **every** $T \in \mathcal{F}$.

In particular, if $\mathbf{A}_1, \dots, \mathbf{A}_r \in M_n(\mathbb{C})$ commute pairwise, there is an invertible $\mathbf{P} \in M_n(\mathbb{C})$ such that $\mathbf{P}^{-1}\mathbf{A}_i\mathbf{P}$ is upper triangular for every i.

Idea. **Take one common eigenvector, extend, recurse.** ① Theorem 9.9.3 gives v_1, an eigenvector of every $T \in \mathcal{F}$, so $U = \operatorname{span}(v_1)$ is invariant under the family. ② On V/U, the induced operators still commute, and their characteristic polynomials still split. ③ By induction on the dimension, one basis of V/U triangularizes all of them. ④ Lift it back and put v_1 in front; the block triangular form of Section 1 does the rest.

Proof. We use induction on n. If $n = 1$, every 1×1 matrix is upper triangular, so any basis works.

Let $n \geq 2$, and assume the theorem for spaces of dimension $n - 1$. By Theorem 9.9.3 there is $v_1 \neq 0$ with $Tv_1 = \lambda_T v_1$ for every $T \in \mathcal{F}$. Let $U = \operatorname{span}(v_1)$; it is T-invariant for every $T \in \mathcal{F}$ (Proposition 9.1.2), and $\{0\} \neq U \neq V$ since $\dim U = 1 < n$. For $T \in \mathcal{F}$ let $\bar{T}$ be the induced operator on V/U (Definition 9.1.2), and put $\bar{\mathcal{F}} = \{\bar{T} : T \in \mathcal{F}\}$.

The family $\bar{\mathcal{F}}$ is commuting: for $S, T \in \mathcal{F}$ and $v \in V$,

$$\bar{S}\bar{T}(v + U) = \bar{S}(Tv + U) = STv + U = TSv + U = \bar{T}\bar{S}(v + U).$$

Each $p_{\bar{T}}$ splits by Lemma 9.9.2, and $\dim V/U = n - 1$ by Theorem 4.4.5. By the induction hypothesis there are $w_2, \dots, w_n \in V$ such that $\bar{\mathcal{B}} = (w_2 + U, \dots, w_n + U)$ is a basis of V/U in which every $[\bar{T}]_{\bar{\mathcal{B}}}^{\bar{\mathcal{B}}}$ is upper triangular.

Let $\mathcal{B} = (v_1, w_2, \ldots, w_n)$. It is linearly independent: if $a_1 v_1 + a_2 w_2 + \cdots + a_n w_n = 0$, then passing to V/U, where $v_1 + U$ is the zero coset, gives $a_2(w_2 + U) + \cdots + a_n(w_n + U) = 0$, so $a_2 = \cdots = a_n = 0$ by the independence of $\bar{\mathcal{B}}$; then $a_1 v_1 = 0$ with $v_1 \neq 0$ gives $a_1 = 0$. A linearly independent list of length $n = \dim V$ is a basis (Theorem 1.6.4). Now $\mathcal{B}$ is a basis of U extended to a basis of V, and $\bar{\mathcal{B}}$ is exactly the list of cosets in Theorem 9.1.3. Hence for every $T \in \mathcal{F}$,

$$[T]_{\mathcal{B}}^{\mathcal{B}} = \begin{pmatrix} \lambda_T & * \\ \mathbf{0} & [\bar{T}]_{\bar{\mathcal{B}}}^{\bar{\mathcal{B}}} \end{pmatrix},$$

with a 1×1 upper-left block, a zero column below it, and an upper triangular lower-right block. Such a matrix is upper triangular. This completes the induction.

For the matrix statement, apply this to the family $\{T_{\mathbf{A}_1}, \ldots, T_{\mathbf{A}_r}\}$ on $\mathbb{C}^n$, whose characteristic polynomials split by Corollary 6.5.5, and let $\mathbf{P}$ have the vectors of $\mathcal{B}$ as columns; then $\mathbf{P}^{-1}\mathbf{A}_i\mathbf{P} = [T_{\mathbf{A}_i}]_{\mathcal{B}}^{\mathcal{B}}$ (Theorem 4.2.3).

■

A first use: if S and T commute and both characteristic polynomials split, the diagonal entries of $[S+T]_{\mathcal{B}}^{\mathcal{B}}$ and $[ST]_{\mathcal{B}}^{\mathcal{B}}$ are the sums and products of the diagonal entries of $[S]_{\mathcal{B}}^{\mathcal{B}}$ and $[T]_{\mathcal{B}}^{\mathcal{B}}$, position by position. For the sum this is entrywise; for the product, the (i,i)-entry of $\mathbf{XY}$ is $\sum_k x_{ik}y_{ki}$ (Definition 0.9.9), and when $\mathbf{X}$ and $\mathbf{Y}$ are upper triangular only $k = i$ can contribute, since $x_{ik} = 0$ for $k < i$ and $y_{ki} = 0$ for $k > i$. By Theorem 9.7.4, those diagonal entries are the eigenvalues. So the eigenvalues of $S + T$ are of the form $\lambda + \mu$, with λ an eigenvalue of S and μ one of T, **paired by position in a common triangular form**. Section 2 warned that nothing like this holds without commuting.

9.9.5 Simultaneous diagonalization

Diagonalization is the stronger wish, and it needs a stronger hypothesis: each operator must be diagonalizable on its own. The name for the goal:

Definition 9.9.1. Simultaneously Diagonalizable

Let V be finite-dimensional. A family $\mathcal{F} \subseteq \mathcal{L}(V)$ is **simultaneously diagonalizable** if there is **one** basis $\mathcal{B}$ of V such that $[T]_{\mathcal{B}}^{\mathcal{B}}$ is diagonal for **every** $T \in \mathcal{F}$. Matrices $\mathbf{A}_1, \ldots, \mathbf{A}_r \in M_n(F)$ are simultaneously diagonalizable over F if there is **one** invertible $\mathbf{P} \in M_n(F)$ with every $\mathbf{P}^{-1}\mathbf{A}_i\mathbf{P}$ diagonal.

Equivalently, by the unwinding in Section 4, V has a basis whose vectors are eigenvectors of every $T \in \mathcal{F}$. The opening of this section showed that a simultaneously diagonalizable family is commuting. For diagonalizable operators the converse holds.

Theorem 9.9.5. Simultaneous Diagonalization

Let V be a finite-dimensional vector space over F with $\dim V \geq 1$, and let $S, T \in \mathcal{L}(V)$ be diagonalizable. Then S and T are simultaneously diagonalizable if and only if $ST = TS$.

For matrices: if $\mathbf{A}, \mathbf{B} \in M_n(F)$ are diagonalizable over F, there is an invertible $\mathbf{P} \in M_n(F)$ with $\mathbf{P}^{-1}\mathbf{AP}$ and $\mathbf{P}^{-1}\mathbf{BP}$ both diagonal if and only if $\mathbf{AB} = \mathbf{BA}$.

Idea. Split V into the eigenspaces of T. On each piece T is a scalar, so **any** basis of the piece consists of eigenvectors of T; we are free to choose it to suit S. The piece is S-invariant because S commutes with T, and S restricted to it is diagonalizable because S is (Section 8). So each piece has a basis of eigenvectors of S, and gluing these bases together gives the common basis.

Proof. ($\Rightarrow$) Let $\mathcal{B}$ be a basis with $[S]_{\mathcal{B}}^{\mathcal{B}}$ and $[T]_{\mathcal{B}}^{\mathcal{B}}$ diagonal. Diagonal matrices commute, so by Corollary 4.1.3, $[ST]_{\mathcal{B}}^{\mathcal{B}} = [S]_{\mathcal{B}}^{\mathcal{B}}[T]_{\mathcal{B}}^{\mathcal{B}} = [T]_{\mathcal{B}}^{\mathcal{B}}[S]_{\mathcal{B}}^{\mathcal{B}} = [TS]_{\mathcal{B}}^{\mathcal{B}}$. The map $R \mapsto [R]_{\mathcal{B}}^{\mathcal{B}}$ is injective (Theorem 4.1.4), so $ST = TS$.

($\Leftarrow$) Suppose $ST = TS$. Let $\lambda_1, \dots, \lambda_k$ be the distinct eigenvalues of T, and put $U_i = E_{\lambda_i}(T)$. Since T is diagonalizable, Theorem 9.4.1 (c) gives

$$V = U_1 \oplus \cdots \oplus U_k.$$

Each U_i is S-invariant (Theorem 9.9.1), and S is diagonalizable, so $S|_{U_i}$ is diagonalizable by Corollary 9.8.3. Hence U_i has a basis $\mathcal{B}_i$ of eigenvectors of $S|_{U_i}$ (Theorem 9.4.1 (b)), which are eigenvectors of S. They are also eigenvectors of T, being nonzero vectors of $U_i = E_{\lambda_i}(T)$. Let $\mathcal{B}$ be the list $\mathcal{B}_1, \dots, \mathcal{B}_k$. Since the sum is direct, $\mathcal{B}$ is a basis of V by Theorem 1.7.5 ((a) $\Rightarrow$ (d)). Every vector of $\mathcal{B}$ is an eigenvector of both S and T, so $[S]_{\mathcal{B}}^{\mathcal{B}}$ and $[T]_{\mathcal{B}}^{\mathcal{B}}$ are diagonal (Theorem 9.4.1, (a) $\Leftrightarrow$ (b)).

For matrices, apply this to $T_{\mathbf{A}}$ and $T_{\mathbf{B}}$ on F^n, which are diagonalizable because $\mathbf{A}$ and $\mathbf{B}$ are (Definition 9.4.1), and commute because $\mathbf{AB} = \mathbf{BA}$; let $\mathbf{P}$ have the common eigenvectors as columns, as in Section 4.

■

The proof is also an algorithm. Diagonalize $\mathbf{A}$; in each eigenspace of $\mathbf{A}$ of dimension at least 2, write down the matrix of $\mathbf{B}$ restricted to it and diagonalize that small matrix; eigenspaces of dimension 1 need no work, since their basis vector is automatically an eigenvector of $\mathbf{B}$.

Example. Diagonalizing Two Commuting Matrices Together

Let

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 \\ -2 & 1 & 2 \\ -2 & 0 & 3 \end{pmatrix}, \qquad \mathbf{B} = \begin{pmatrix} 1 & -2 & 2 \\ 0 & 1 & 0 \\ 0 & -2 & 3 \end{pmatrix} \in M_3(\mathbb{R}).$$

Check that $\mathbf{AB} = \mathbf{BA}$, and find an invertible $\mathbf{P}$ such that $\mathbf{P}^{-1}\mathbf{AP}$ and $\mathbf{P}^{-1}\mathbf{BP}$ are both diagonal.

Solution. They commute. Multiplying out,

$$\mathbf{AB} = \begin{pmatrix} 1 & -2 & 2 \\ -2 & 1 & 2 \\ -2 & -2 & 5 \end{pmatrix} = \mathbf{BA}.$$

Diagonalize $\mathbf{A}$. Expanding $\det(x\mathbf{I}-\mathbf{A})$ along the first row, whose only nonzero entry is $x-1$,

$$p_{\mathbf{A}}(x) = (x-1)\det\begin{pmatrix} x-1 & -2 \\ 0 & x-3 \end{pmatrix} = (x-1)^2(x-3).$$

For $\lambda = 1$, $\mathbf{A}-\mathbf{I} = \begin{pmatrix} 0 & 0 & 0 \\ -2 & 0 & 2 \\ -2 & 0 & 2 \end{pmatrix}$ has rank 1 and reduces to $x = z$, so $E_1(\mathbf{A})$ is the plane with basis $\mathbf{u}_1 = (0,1,0)$, $\mathbf{u}_2 = (1,0,1)$. For $\lambda = 3$, $\mathbf{A}-3\mathbf{I} = \begin{pmatrix} -2 & 0 & 0 \\ -2 & -2 & 2 \\ -2 & 0 & 0 \end{pmatrix}$ gives $x = 0$ and $y = z$, so $E_3(\mathbf{A}) = \operatorname{span}((0,1,1))$. Since $g(1) = 2 = a(1)$ and $g(3) = 1 = a(3)$, $\mathbf{A}$ is diagonalizable. The same computation for $\mathbf{B}$ gives $p_{\mathbf{B}} = (x-1)^2(x-3)$, with $\mathbf{B}-\mathbf{I} = \begin{pmatrix} 0 & -2 & 2 \\ 0 & 0 & 0 \\ 0 & -2 & 2 \end{pmatrix}$ of rank 1 and $\mathbf{B}-3\mathbf{I}$ of rank 2, so $\mathbf{B}$ is diagonalizable as well.

A basis of eigenvectors of $\mathbf{A}$ need not work for $\mathbf{B}$: $\mathbf{B}\mathbf{u}_1 = (-2,1,-2)$ is not a multiple of $\mathbf{u}_1$. This is where the search inside the eigenspace comes in.

The one-dimensional eigenspace. $E_3(\mathbf{A})$ is $\mathbf{B}$-invariant (Theorem 9.9.1) and one-dimensional, so $\mathbf{v}_3 = (0,1,1)$ is an eigenvector of $\mathbf{B}$: indeed $\mathbf{B}\mathbf{v}_3 = (0,1,1) = 1\cdot\mathbf{v}_3$.

The plane $E_1(\mathbf{A})$. It is $\mathbf{B}$-invariant, and

$$\mathbf{B}\mathbf{u}_1 = (-2,1,-2) = \mathbf{u}_1 - 2\mathbf{u}_2, \qquad \mathbf{B}\mathbf{u}_2 = (3,0,3) = 3\mathbf{u}_2.$$

So the matrix of $\mathbf{B}|_{E_1(\mathbf{A})}$ in the basis $(\mathbf{u}_1,\mathbf{u}_2)$ is $\mathbf{M} = \begin{pmatrix} 1 & 0 \\ -2 & 3 \end{pmatrix}$, whose columns are the coordinates just found. It is lower triangular with eigenvalues 1 and 3. From $\mathbf{M}-\mathbf{I} = \begin{pmatrix} 0 & 0 \\ -2 & 2 \end{pmatrix}$ the eigenvector is $(1,1)$, which stands for $\mathbf{u}_1+\mathbf{u}_2 = (1,1,1)$; from $\mathbf{M}-3\mathbf{I} = \begin{pmatrix} -2 & 0 \\ -2 & 0 \end{pmatrix}$ it is $(0,1)$, which stands for $\mathbf{u}_2 = (1,0,1)$.

The common basis. Put $\mathbf{v}_1 = (1,0,1)$, $\mathbf{v}_2 = (1,1,1)$, $\mathbf{v}_3 = (0,1,1)$ as columns:

$$\mathbf{P} = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}, \qquad \mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \operatorname{diag}(1,1,3), \qquad \mathbf{P}^{-1}\mathbf{B}\mathbf{P} = \operatorname{diag}(3,1,1).$$

Check. $\mathbf{A}\mathbf{v}_1 = \mathbf{v}_1$, $\mathbf{A}\mathbf{v}_2 = \mathbf{v}_2$, $\mathbf{A}\mathbf{v}_3 = (0,3,3) = 3\mathbf{v}_3$; $\mathbf{B}\mathbf{v}_1 = (3,0,3) = 3\mathbf{v}_1$, $\mathbf{B}\mathbf{v}_2 = \mathbf{v}_2$, $\mathbf{B}\mathbf{v}_3 = \mathbf{v}_3$. Expanding along the first row, $\det\mathbf{P} = 1(1-1) - 1(0-1) + 0 = 1 \neq 0$, so $\mathbf{P}$ is invertible, with $\mathbf{P}^{-1} = \begin{pmatrix} 0 & -1 & 1 \\ 1 & 1 & -1 \\ -1 & 0 & 1 \end{pmatrix}$.

Notice that neither matrix alone determines $\mathbf{P}$: $\mathbf{A}$ cannot tell $\mathbf{v}_1$ from $\mathbf{v}_2$, and $\mathbf{B}$ cannot tell $\mathbf{v}_2$ from $\mathbf{v}_3$. Together they pin down all three lines.

Warning. **Commuting does not make an operator diagonalizable.** The theorem assumes that both S and T are diagonalizable. $\mathbf{J} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ commutes with $\mathbf{I}_2$, and $\mathbf{I}_2$ is diagonal, but $\mathbf{J}$ is not diagonalizable (Section 4), so no basis makes both diagonal. What survives is Theorem 9.9.4: $\mathbf{J}$ and $\mathbf{I}_2$ are already upper triangular together.

The theorem extends from two operators to any number, and even to infinitely many. An infinite family looks hopeless for an induction, but it lives inside the finite-dimensional space $\mathcal{L}(V)$, so finitely many of its members already span everything the family spans.

Theorem 9.9.6. Simultaneous Diagonalization of a Commuting Family

Let V be a finite-dimensional vector space over F with $\dim V \geq 1$, and let $\mathcal{F} \subseteq \mathcal{L}(V)$ be a family of diagonalizable operators. Then $\mathcal{F}$ is simultaneously diagonalizable if and only if it is commuting.

Idea. ① **Finite first.** For finitely many commuting diagonalizable $T_1, \ldots, T_r$, induct on r: split V into eigenspaces of T_1, and on each piece apply the case $r-1$ to the restrictions of $T_2, \ldots, T_r$. ② **Reduce to finite.** A longest linearly independent list $T_1, \ldots, T_r$ taken from $\mathcal{F}$ spans every member of $\mathcal{F}$, and has length at most $\dim \mathcal{L}(V)$. ③ A basis that diagonalizes $T_1, \ldots, T_r$ diagonalizes every linear combination of them.

Proof. ($\Rightarrow$) If one basis makes every $[T]_{\mathcal{B}}^{\mathcal{B}}$ diagonal, any two of these matrices commute, and the argument of Theorem 9.9.5 ($\Rightarrow$) gives $ST = TS$ for all $S, T \in \mathcal{F}$.

($\Leftarrow$) Suppose $\mathcal{F}$ is commuting. If $\mathcal{F}$ is empty, any basis works; assume not.

Step 1: finitely many operators. We prove by induction on $r \geq 1$: for every nonzero finite-dimensional space W and all pairwise commuting diagonalizable $T_1, \ldots, T_r \in \mathcal{L}(W)$, some basis of W consists of eigenvectors of every T_i. For $r = 1$ this is Theorem 9.4.1. Let $r \geq 2$, and let $\lambda_1, \ldots, \lambda_k$ be the distinct eigenvalues of T_1, with $U_j = E_{\lambda_j}(T_1) \neq \{0\}$. By Theorem 9.4.1 (c), $W = U_1 \oplus \cdots \oplus U_k$. Each U_j is T_i-invariant for every i (Theorem 9.9.1), each $T_i|_{U_j}$ is diagonalizable (Corollary 9.8.3), and the restrictions $T_2|_{U_j}, \ldots, T_r|_{U_j}$ commute pairwise, since they agree with $T_2, \ldots, T_r$ on U_j. By the induction hypothesis, U_j has a basis $\mathcal{B}_j$ of common eigenvectors of $T_2, \ldots, T_r$; they are eigenvectors of T_1 too, as nonzero vectors of U_j. The list $\mathcal{B}_1, \ldots, \mathcal{B}_k$ is a basis of W by Theorem 1.7.5 ((a) $\Rightarrow$ (d)). This completes the induction.

Step 2: a finite spanning subfamily. Let $N = \dim \mathcal{L}(V) = (\dim V)^2$ (Theorem 4.1.4). Every linearly independent list in $\mathcal{L}(V)$ has length at most N (Theorem 1.6.5), so among the linearly independent lists of members of $\mathcal{F}$ there is one of greatest length, say $(T_1, \ldots, T_r)$, with $r \geq 1$ because a single nonzero member is independent (if every member of $\mathcal{F}$ is 0, any basis works). For every $T \in \mathcal{F}$, the list $(T_1, \ldots, T_r, T)$ is longer, hence dependent, so $T \in \operatorname{span}(T_1, \ldots, T_r)$ by Lemma 1.5.3.

Step 3. By Step 1 there is a basis $\mathcal{B}$ of V with every $[T_i]_{\mathcal{B}}^{\mathcal{B}}$ diagonal. For $T \in \mathcal{F}$, write $T = c_1 T_1 + \cdots + c_r T_r$; then $[T]_{\mathcal{B}}^{\mathcal{B}} = c_1 [T_1]_{\mathcal{B}}^{\mathcal{B}} + \cdots + c_r [T_r]_{\mathcal{B}}^{\mathcal{B}}$ (Corollary 4.1.3), a linear combination of diagonal matrices, hence diagonal. So $\mathcal{B}$ diagonalizes every member of $\mathcal{F}$.

■

A typical infinite family is $\{p(T) : p \in F[x]\}$ for a diagonalizable T. It is commuting, each $p(T)$ is diagonalizable (Section 8), and here one basis works for the whole family without any of the theorem's machinery, namely an eigenbasis of T: an eigenvector of T for λ is an eigenvector of $p(T)$ for $p(\lambda)$ (Exercise (B3: Eigenvalues of powers and inverses)). The theorem says the same for families that are not generated by one operator in so visible a way.

9.9.6 Exercises

9.9.6.1 A. Check your understanding

Exercise. A1

(a) State the theorem relating commuting operators and eigenspaces.

(b) True or false: if $ST = TS$, then every eigenvector of T is an eigenvector of S. Justify your answer.

(c) True or false: any two commuting matrices in $M_2(\mathbb{R})$ have a common eigenvector in $\mathbb{R}^2$. Justify your answer.

(d) True or false: if $\mathbf{A}, \mathbf{B} \in M_n(F)$ are simultaneously diagonalizable, then $\mathbf{AB} = \mathbf{BA}$. Justify your answer.

(e) True or false: if $\mathbf{AB} = \mathbf{BA}$ and $\mathbf{A}$ is diagonalizable, then $\mathbf{B}$ is diagonalizable. Justify your answer.

(f) In the proof of Theorem 9.9.5, which hypothesis makes each eigenspace of T invariant under S, and which makes the restriction of S diagonalizable?

Solution.

(a) If $S, T \in \mathcal{L}(V)$ and $ST = TS$, then $E_\lambda(T)$ is S-invariant for every $\lambda \in F$ (Theorem 9.9.1).

(b) False. $T = \mathbf{I}_2$ commutes with $S = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$; the vector $\mathbf{e}_2$ is an eigenvector of T, but $S\mathbf{e}_2 = (1, 1)$ is not a multiple of $\mathbf{e}_2$.

(c) False. The rotation $\mathbf{R} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ commutes with itself and has no real eigenvector, since $p_{\mathbf{R}} = x^2 + 1$ has no real root.

(d) True. If $\mathbf{P}^{-1}\mathbf{AP} = \mathbf{D}$ and $\mathbf{P}^{-1}\mathbf{BP} = \mathbf{D}'$ are diagonal, then $\mathbf{AB} = \mathbf{PDP}^{-1}\mathbf{PD}'\mathbf{P}^{-1} = \mathbf{PDD}'\mathbf{P}^{-1} = \mathbf{PD}'\mathbf{DP}^{-1} = \mathbf{BA}$, since diagonal matrices commute.

(e) False. $\mathbf{A} = \mathbf{I}_2$ is diagonalizable and commutes with $\mathbf{B} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$, which is not diagonalizable.

(f) The hypothesis $ST = TS$ gives the invariance (Theorem 9.9.1). The hypothesis that S is diagonalizable gives, through Corollary 9.8.3, that $S|_{E_\lambda(T)}$ is diagonalizable.

9.9.6.2 B. Practice

Exercise. B1: Common eigenvectors

Let

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & -1 & 2 \end{pmatrix}, \qquad \mathbf{B} = \begin{pmatrix} 2 & 1 & 0 \\ -1 & 4 & 0 \\ 2 & -2 & 5 \end{pmatrix} \in M_3(\mathbb{R}).$$

(a) Verify that $\mathbf{AB} = \mathbf{BA}$.

(b) Find every common eigenvector of $\mathbf{A}$ and $\mathbf{B}$.

(c) Hence show that $\mathbf{A}$ and $\mathbf{B}$ are not simultaneously diagonalizable, and explain which hypothesis of Theorem 9.9.5 fails.

Solution.

(a) Multiplying out,

$$\mathbf{AB} = \begin{pmatrix} 2 & 1 & 0 \\ -1 & 4 & 0 \\ 7 & -7 & 10 \end{pmatrix} = \mathbf{BA}.$$

For instance, the $(3,1)$-entry of $\mathbf{AB}$ is row $(1,-1,2)$ of $\mathbf{A}$ times column $(2,-1,2)$ of $\mathbf{B}$, namely $2+1+4=7$, and that of $\mathbf{BA}$ is row $(2,-2,5)$ of $\mathbf{B}$ times column $(1,0,1)$ of $\mathbf{A}$, namely $2+0+5=7$.

(b) $\mathbf{A}$ is lower triangular, so $p_{\mathbf{A}} = (x-1)^2(x-2)$. $\mathbf{A}-\mathbf{I} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & -1 & 1 \end{pmatrix}$ has rank 1, so $E_1(\mathbf{A}) = \{(x,y,z) : x-y+z=0\}$ with basis $\mathbf{u}_1 = (1,1,0)$, $\mathbf{u}_2 = (-1,0,1)$. $\mathbf{A}-2\mathbf{I} = \begin{pmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 1 & -1 & 0 \end{pmatrix}$ gives $E_2(\mathbf{A}) = \operatorname{span}(\mathbf{e}_3)$.

A common eigenvector is an eigenvector of $\mathbf{A}$, so it lies in $E_1(\mathbf{A})$ or $E_2(\mathbf{A})$, and it must be an eigenvector of $\mathbf{B}$ there. On the line $E_2(\mathbf{A})$: $\mathbf{Be}_3 = (0,0,5) = 5\mathbf{e}_3$, so every nonzero multiple of $\mathbf{e}_3$ is a common eigenvector. On the plane $E_1(\mathbf{A})$, which is $\mathbf{B}$-invariant by Theorem 9.9.1:

$$\mathbf{Bu}_1 = (3,3,0) = 3\mathbf{u}_1, \qquad \mathbf{Bu}_2 = (-2,1,3) = \mathbf{u}_1 + 3\mathbf{u}_2.$$

So in the basis $\mathcal{B} = (\mathbf{u}_1, \mathbf{u}_2)$, $[\mathbf{B}|_{E_1(\mathbf{A})}]_{\mathcal{B}}^{\mathcal{B}} = \begin{pmatrix} 3 & 1 \\ 0 & 3 \end{pmatrix}$. Its only eigenvalue is 3, and its eigenvectors are the nonzero multiples of $(1,0)$, that is, of $\mathbf{u}_1$. Hence the common eigenvectors are exactly the nonzero multiples of $(1,1,0)$ and the nonzero multiples of $(0,0,1)$.

(c) By (b), any list of common eigenvectors lies in $\operatorname{span}((1,1,0), \mathbf{e}_3)$, a plane, so no three of them form a basis of $\mathbb{R}^3$. Hence no basis diagonalizes both. The failing hypothesis is that $\mathbf{B}$ is diagonalizable: $p_{\mathbf{B}} = (x-3)^2(x-5)$ (from the block computation, $p_{\mathbf{B}} = p_{\mathbf{B}|_{E_1(\mathbf{A})}}\, p_{\bar{\mathbf{B}}}$, or directly), and $\mathbf{B}-3\mathbf{I} = \begin{pmatrix} -1 & 1 & 0 \\ -1 & 1 & 0 \\ 2 & -2 & 2 \end{pmatrix}$ has rank 2, so $g_{\mathbf{B}}(3) = 1 < 2 = a_{\mathbf{B}}(3)$.

Exercise. B2: Decide simultaneous diagonalizability

For each pair of real matrices, determine whether they are simultaneously diagonalizable over $\mathbb{R}$. If they are, find a suitable $\mathbf{P}$. Justify your answer.

(a) $\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}$, $\mathbf{B} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$.

(b) $\mathbf{A} = \begin{pmatrix} 1 & 1 \\ 0 & 2 \end{pmatrix}$, $\mathbf{B} = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$.

(c) $\mathbf{A} = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$, $\mathbf{B} = \begin{pmatrix} 2 & 1 \\ 0 & 2 \end{pmatrix}$.

Solution.

(a) Yes. $\mathbf{AB} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} = \mathbf{BA}$. Both are diagonalizable: $\mathbf{A}$ has $p_{\mathbf{A}} = x^2 - 2x - 3 = (x-3)(x+1)$ and $\mathbf{B}$ has $p_{\mathbf{B}} = x^2 - 1 = (x-1)(x+1)$, each with distinct eigenvalues (Corollary 9.4.2). By Theorem 9.9.5 a common $\mathbf{P}$ exists. The eigenspaces of $\mathbf{A}$ are lines, so their spanning vectors $(1,1)$ and $(1,-1)$ work: $\mathbf{P} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$, $\mathbf{P}^{-1}\mathbf{AP} = \operatorname{diag}(3,-1)$, $\mathbf{P}^{-1}\mathbf{BP} = \operatorname{diag}(1,-1)$, since $\mathbf{B}(1,1) = (1,1)$ and $\mathbf{B}(1,-1) = (-1,1)$.

(b) No. Both are diagonalizable (distinct eigenvalues $1, 2$), but $\mathbf{AB} = \begin{pmatrix} 1 & 2 \\ 0 & 4 \end{pmatrix} \neq \begin{pmatrix} 1 & 1 \\ 0 & 4 \end{pmatrix} = \mathbf{BA}$, so by Theorem 9.9.5 ($\Rightarrow$) no common $\mathbf{P}$ exists.

(c) No. They commute, since $\mathbf{A} = 2\mathbf{I}$. But $\mathbf{B}$ is not diagonalizable: $p_{\mathbf{B}} = (x-2)^2$ and $\mathbf{B} - 2\mathbf{I}$ has rank 1, so $g(2) = 1 < 2$. If $\mathbf{P}^{-1}\mathbf{BP}$ were diagonal, $\mathbf{B}$ would be diagonalizable (Definition 9.4.1).

Exercise. B3: Matrices commuting with a diagonal matrix

(a) Let $\mathbf{D} = \operatorname{diag}(1,2,3) \in M_3(\mathbb{R})$. Prove that a matrix $\mathbf{B} \in M_3(\mathbb{R})$ commutes with $\mathbf{D}$ if and only if $\mathbf{B}$ is diagonal.

(b) Let $\mathbf{D}' = \operatorname{diag}(1,1,2) \in M_3(\mathbb{R})$. Describe all $\mathbf{B} \in M_3(\mathbb{R})$ with $\mathbf{BD}' = \mathbf{D}'\mathbf{B}$, and find one that is not diagonalizable.

Solution.

(a) ($\Leftarrow$) Diagonal matrices commute. ($\Rightarrow$) Suppose $\mathbf{BD} = \mathbf{DB}$. The eigenspaces of $\mathbf{D}$ are $E_1(\mathbf{D}) = \operatorname{span}(\mathbf{e}_1)$, $E_2(\mathbf{D}) = \operatorname{span}(\mathbf{e}_2)$, and $E_3(\mathbf{D}) = \operatorname{span}(\mathbf{e}_3)$. By Theorem 9.9.1 each is $\mathbf{B}$-invariant, so $\mathbf{B}\mathbf{e}_j \in \operatorname{span}(\mathbf{e}_j)$ for each j. The j-th column of $\mathbf{B}$ is $\mathbf{B}\mathbf{e}_j$, so it has zeros off the diagonal, and $\mathbf{B}$ is diagonal. (Directly: $(\mathbf{DB})_{ij} = i\,b_{ij}$ and $(\mathbf{BD})_{ij} = b_{ij}\,j$, so $(i-j)b_{ij} = 0$, and $b_{ij} = 0$ for $i \neq j$.)

(b) As computed in the text for $\operatorname{diag}(1,1,2)$, $\mathbf{BD}' = \mathbf{D}'\mathbf{B}$ exactly when $\mathbf{B} = \begin{pmatrix} b_{11} & b_{12} & 0 \\ b_{21} & b_{22} & 0 \\ 0 & 0 & b_{33} \end{pmatrix}$ with arbitrary entries: $E_1(\mathbf{D}') = \operatorname{span}(\mathbf{e}_1, \mathbf{e}_2)$ and $E_2(\mathbf{D}') = \operatorname{span}(\mathbf{e}_3)$ are $\mathbf{B}$-invariant,

which by Theorem 9.1.5 forces this shape, and conversely $\mathbf{D}' = \mathbf{I}_2 \oplus (2)$ commutes with every such block matrix (Theorem 8.1.4). For example $\mathbf{B} = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$ commutes with $\mathbf{D}'$, and it is not diagonalizable: $p_{\mathbf{B}} = x(x-1)^2$ and $\mathbf{B} - \mathbf{I}$ has rank 2, so $g(1) = 1 < 2 = a(1)$.

9.9.6.3 C. Going deeper

Exercise. C1: The commutant of a matrix with distinct eigenvalues

Let $\mathbf{A} \in M_n(F)$, $n \geq 1$, have n distinct eigenvalues $\lambda_1, \ldots, \lambda_n \in F$.

(a) Prove that if $\mathbf{AB} = \mathbf{BA}$, then $\mathbf{B}$ is diagonalizable and every eigenvector of $\mathbf{A}$ is an eigenvector of $\mathbf{B}$.

(b) Prove that if $\mathbf{AB} = \mathbf{BA}$, then $\mathbf{B} = q(\mathbf{A})$ for some $q \in F[x]$ of degree at most $n-1$.

(c) Deduce that $\{\mathbf{B} \in M_n(F) : \mathbf{AB} = \mathbf{BA}\}$ is a subspace of $M_n(F)$ of dimension n, with basis $(\mathbf{I}, \mathbf{A}, \ldots, \mathbf{A}^{n-1})$.

Hint: for (b), use Lagrange interpolation.

Solution.

(a) By Corollary 9.4.2, $\mathbf{A}$ is diagonalizable, say $\mathbf{P}^{-1}\mathbf{AP} = \mathbf{D}$ with $\mathbf{D}$ diagonal. A diagonal matrix is upper triangular, so its diagonal entries are the eigenvalues of $\mathbf{A}$, each occurring as often as its algebraic multiplicity (Theorem 9.7.4 (b)); here there are n distinct eigenvalues and n diagonal places, so each occurs once, and after permuting the columns of $\mathbf{P}$ and the diagonal entries of $\mathbf{D}$ together we may assume $\mathbf{D} = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$. Let $\mathbf{v}_i$ be column i of $\mathbf{P}$. Each $E_{\lambda_i}(\mathbf{A})$ has dimension at least 1, and the dimensions add up to at most n (Corollary 9.2.3), so each has dimension exactly 1 and $E_{\lambda_i}(\mathbf{A}) = \operatorname{span}(\mathbf{v}_i)$. Every eigenvector of $\mathbf{A}$ for λ_i is a nonzero multiple of $\mathbf{v}_i$. By Theorem 9.9.1, $\mathbf{Bv}_i \in \operatorname{span}(\mathbf{v}_i)$, so $\mathbf{Bv}_i = \mu_i \mathbf{v}_i$ for some $\mu_i \in F$, and $\mathbf{v}_i$, with every nonzero multiple of it, is an eigenvector of $\mathbf{B}$. The columns of $\mathbf{P}$ form a basis of eigenvectors of $\mathbf{B}$, so $\mathbf{P}^{-1}\mathbf{BP} = \operatorname{diag}(\mu_1, \ldots, \mu_n)$ and $\mathbf{B}$ is diagonalizable.

(b) The nodes $\lambda_1, \ldots, \lambda_n$ are distinct, so by Theorem 6.6.1 there is $q \in F[x]$ with $\deg q \leq n-1$ and $q(\lambda_i) = \mu_i$ for every i. By Theorem 9.4.3,

$$q(\mathbf{A}) = \mathbf{P} \operatorname{diag}\left(q(\lambda_1), \ldots, q(\lambda_n)\right) \mathbf{P}^{-1} = \mathbf{P} \operatorname{diag}(\mu_1, \ldots, \mu_n)\, \mathbf{P}^{-1} = \mathbf{B}.$$

(c) If $\mathbf{AB} = \mathbf{BA}$ and $\mathbf{AB}' = \mathbf{B}'\mathbf{A}$, then $\mathbf{A}(c\mathbf{B} + \mathbf{B}') = c\mathbf{BA} + \mathbf{B}'\mathbf{A} = (c\mathbf{B} + \mathbf{B}')\mathbf{A}$, and $\mathbf{AO} = \mathbf{OA}$; so the set $\mathcal{C}$ is a subspace. Each $\mathbf{A}^k$ commutes with $\mathbf{A}$, so $\operatorname{span}(\mathbf{I}, \mathbf{A}, \ldots, \mathbf{A}^{n-1}) \subseteq \mathcal{C}$, and (b) gives the reverse inclusion. For independence, let $c_0\mathbf{I} + c_1\mathbf{A} + \cdots + c_{n-1}\mathbf{A}^{n-1} = \mathbf{O}$ and put $q = \sum_k c_k x^k$. By Theorem 9.4.3, $\mathbf{O} = q(\mathbf{A}) = \mathbf{P} \operatorname{diag}(q(\lambda_1), \ldots, q(\lambda_n))\, \mathbf{P}^{-1}$, so $q(\lambda_i) = 0$ for all i. A polynomial of degree at most $n-1$ with n distinct roots is zero (Corollary 6.1.4), so every $c_k = 0$. Hence $(\mathbf{I}, \mathbf{A}, \ldots, \mathbf{A}^{n-1})$ is a basis of $\mathcal{C}$, and $\dim \mathcal{C} = n$.

Exercise. C2: Commuting nilpotent matrices

Recall that $\mathbf{N} \in M_n(F)$ is **nilpotent** if $\mathbf{N}^a = \mathbf{O}$ for some $a \geq 1$ (Exercise (C2: Nilpotent operators)). Let $\mathbf{N}, \mathbf{M} \in M_n(F)$ with $\mathbf{NM} = \mathbf{MN}$, $\mathbf{N}^a = \mathbf{O}$, and $\mathbf{M}^b = \mathbf{O}$ for some $a, b \geq 1$.

(a) Prove that $\mathbf{NM}$ is nilpotent.

(b) Prove that $(\mathbf{N} + \mathbf{M})^{a+b-1} = \mathbf{O}$, so that $\mathbf{N} + \mathbf{M}$ is nilpotent.

(c) Show that both conclusions can fail without $\mathbf{NM} = \mathbf{MN}$.

(d) Now let $F = \mathbb{C}$. Give a second proof of (b) using Theorem 9.9.4.

Hint: for (b), expand the power as a sum of products of $a+b-1$ factors, each equal to $\mathbf{N}$ or $\mathbf{M}$.

Solution.

(a) Since $\mathbf{NM} = \mathbf{MN}$, induction on k gives $(\mathbf{NM})^k = \mathbf{N}^k\mathbf{M}^k$: if it holds for k, then $(\mathbf{NM})^{k+1} = \mathbf{N}^k\mathbf{M}^k\mathbf{NM} = \mathbf{N}^k\mathbf{NM}^k\mathbf{M} = \mathbf{N}^{k+1}\mathbf{M}^{k+1}$, where $\mathbf{M}^k\mathbf{N} = \mathbf{NM}^k$ follows from $\mathbf{MN} = \mathbf{NM}$ by induction as well. With $k = a$, $(\mathbf{NM})^a = \mathbf{N}^a\mathbf{M}^a = \mathbf{O} \cdot \mathbf{M}^a = \mathbf{O}$.

(b) Let $s = a+b-1$. By distributivity, $(\mathbf{N}+\mathbf{M})^s$ is the sum of all 2^s products $\mathbf{X}_1\mathbf{X}_2\cdots\mathbf{X}_s$ with each $\mathbf{X}_t \in \{\mathbf{N}, \mathbf{M}\}$. Since $\mathbf{N}$ and $\mathbf{M}$ commute, each such product can be rearranged, one adjacent swap at a time, into $\mathbf{N}^i\mathbf{M}^{s-i}$, where i is the number of factors equal to $\mathbf{N}$. If $i \geq a$, then $\mathbf{N}^i = \mathbf{N}^a\mathbf{N}^{i-a} = \mathbf{O}$. If $i \leq a-1$, then $s - i \geq s - a + 1 = b$, so $\mathbf{M}^{s-i} = \mathbf{O}$. Either way the product is $\mathbf{O}$, so $(\mathbf{N} + \mathbf{M})^s = \mathbf{O}$.

(c) Let $\mathbf{N} = \mathbf{E}_{12}$ and $\mathbf{M} = \mathbf{E}_{21}$ in $M_2(F)$. Both square to $\mathbf{O}$. But $\mathbf{NM} = \mathbf{E}_{11}$ satisfies $\mathbf{E}_{11}^k = \mathbf{E}_{11} \neq \mathbf{O}$ for all $k \geq 1$, and $\mathbf{N}+\mathbf{M} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ satisfies $(\mathbf{N}+\mathbf{M})^2 = \mathbf{I}$, so $(\mathbf{N}+\mathbf{M})^k$ is $\mathbf{I}$ or $\mathbf{N} + \mathbf{M}$, never $\mathbf{O}$. Neither is nilpotent, and indeed $\mathbf{NM} = \mathbf{E}_{11} \neq \mathbf{E}_{22} = \mathbf{MN}$.

(d) By Theorem 9.9.4 there is an invertible $\mathbf{P} \in M_n(\mathbb{C})$ with $\mathbf{U} = \mathbf{P}^{-1}\mathbf{NP}$ and $\mathbf{U}' = \mathbf{P}^{-1}\mathbf{MP}$ upper triangular. Since $\mathbf{N}$ is nilpotent, $\operatorname{spec}(\mathbf{N}) = \{0\}$ (Exercise (C2: Nilpotent operators)), so every diagonal entry of $\mathbf{U}$ is 0 by Theorem 9.7.4; likewise for $\mathbf{U}'$. So $\mathbf{U}+\mathbf{U}' = \mathbf{P}^{-1}(\mathbf{N}+\mathbf{M})\mathbf{P}$ is upper triangular with zero diagonal. A matrix $\mathbf{X}$ of this kind is nilpotent: $\mathbf{Xe}_1 = \mathbf{0}$ and $\mathbf{Xe}_j \in \operatorname{span}(\mathbf{e}_1, \ldots, \mathbf{e}_{j-1})$, so by induction $\mathbf{X}^j\mathbf{e}_j = \mathbf{0}$ and $\mathbf{X}^n = \mathbf{O}$. Hence $(\mathbf{N} + \mathbf{M})^n = \mathbf{P}(\mathbf{U} + \mathbf{U}')^n\mathbf{P}^{-1} = \mathbf{O}$.

9.10 Spectral Mapping, Products in Both Orders, Left Eigenvectors

Once the eigenvalues of T are known, what can we say about the eigenvalues of T^2, of $T^2 - 3T + I$, or of any polynomial in T? And when two matrices are multiplied in the two possible orders, how are the eigenvalues of $\mathbf{AB}$ and $\mathbf{BA}$ related? This section answers both questions using the triangular forms of Section 7 and the block matrices of Chapter 8. Along the way it settles two debts: Chapter 7 showed $p_{\mathbf{AB}} = p_{\mathbf{BA}}$ only when $\mathbf{A}$ is invertible, and Chapter 8 proved that vanishing traces of all powers force a matrix to be nilpotent by a long computation, with no explanation of why. We end with left eigenvectors, the eigenvectors of the transpose, which pair with ordinary eigenvectors in a useful way and prepare the Markov chains of Section 11.

9.10.1 Eigenvalues of a polynomial in T

One half of the answer needs no theory. If $Tv = \lambda v$ with $v \neq 0$, then $T^k v = \lambda^k v$ for every k, and so $q(T)v = q(\lambda)v$ for every $q \in F[x]$ (Exercise (B3: Eigenvalues of powers and inverses)). So q **maps eigenvalues of T to eigenvalues of $q(T)$**, over every field and in every dimension.

The other half, that **every** eigenvalue of $q(T)$ arises this way, can fail. Part (c) of Exercise (B3: Eigenvalues of powers and inverses) already met the rotation $\mathbf{R} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ on $\mathbb{R}^2$: it has no real eigenvalue, while $\mathbf{R}^2 = -\mathbf{I}$ has the eigenvalue -1. The eigenvalues of $\mathbf{R}$ that would map to -1, namely $\pm i$, are missing from $\mathbb{R}$. So we expect the full statement to need a characteristic polynomial that splits. With that hypothesis we even get the multiplicities, because a triangular form shows all the eigenvalues at once on its diagonal.

The eigenvalues of $q(T)$ are the values of q at the eigenvalues of T, counted with multiplicity.

Theorem 9.10.1. Spectral Mapping Theorem

Let V be a finite-dimensional vector space over F with $\dim V = n \geq 1$, and let $T \in \mathcal{L}(V)$. Suppose p_T splits over F, say

$$p_T(x) = (x - \lambda_1)(x - \lambda_2) \cdots (x - \lambda_n),$$

where each eigenvalue appears as often as its algebraic multiplicity. Then for every $q \in F[x]$,

$$p_{q(T)}(x) = \big(x - q(\lambda_1)\big)\big(x - q(\lambda_2)\big) \cdots \big(x - q(\lambda_n)\big),$$

and consequently

$$\operatorname{spec}\big(q(T)\big) = q\big(\operatorname{spec}(T)\big) = \{q(\lambda) : \lambda \in \operatorname{spec}(T)\}.$$

In particular this holds for every operator on a nonzero finite-dimensional complex space, and for every $\mathbf{A} \in M_n(\mathbb{C})$.

Idea. Choose a basis $\mathcal{B}$ in which T is upper triangular (Section 7). Its diagonal is $\lambda_1, \ldots, \lambda_n$. Products and sums of upper triangular matrices are upper triangular, and their diagonals are computed position by position, so $q([T]_{\mathcal{B}}^{\mathcal{B}})$ is upper triangular with diagonal $q(\lambda_1), \ldots, q(\lambda_n)$. The characteristic polynomial of a triangular matrix is read off its diagonal.

Proof. By Theorem 9.7.2 there is a basis $\mathcal{B}$ of V such that $\mathbf{U} = [T]_{\mathcal{B}}^{\mathcal{B}}$ is upper triangular. By Theorem 9.7.4, its diagonal entries are $\lambda_1, \ldots, \lambda_n$ in some order; renumbering the λ_i, we may assume $u_{ii} = \lambda_i$.

Claim. If $\mathbf{X}, \mathbf{Y} \in M_n(F)$ are upper triangular, then so are $\mathbf{X} + \mathbf{Y}$, $c\mathbf{X}$ for $c \in F$, and $\mathbf{XY}$, with diagonal entries $x_{ii} + y_{ii}$, cx_{ii}, and $x_{ii}y_{ii}$.

Proof. The statements for $\mathbf{X} + \mathbf{Y}$ and $c\mathbf{X}$ are entrywise. For the product, $(\mathbf{XY})_{ij} = \sum_{k=1}^{n} x_{ik}y_{kj}$ (Definition 0.9.9). If $i > j$, every k has $k < i$ or $k \geq i > j$, so $x_{ik} = 0$ or $y_{kj} = 0$, and the sum is 0. If $i = j$, the only term that can be nonzero has $k \geq i$ and $k \leq i$, so $(\mathbf{XY})_{ii} = x_{ii}y_{ii}$.

■

By the claim and induction on k, $\mathbf{U}^k$ is upper triangular with diagonal entries $\lambda_1^k, \dots, \lambda_n^k$ (for $k = 0$, $\mathbf{U}^0 = \mathbf{I}$). Writing $q = \sum_k a_k x^k$, the claim again shows that $q(\mathbf{U}) = \sum_k a_k \mathbf{U}^k$ is upper triangular with diagonal entries $\sum_k a_k \lambda_i^k = q(\lambda_i)$. By Corollary 4.1.3, $[q(T)]_{\mathcal{B}}^{\mathcal{B}} = q(\mathbf{U})$, so by Definition 7.10.2 and the examples after Definition 7.10.1,

$$p_{q(T)} = p_{q(\mathbf{U})} = \prod_{i=1}^{n} (x - q(\lambda_i)).$$

For the spectra: by Theorem 9.2.1, $\mu \in F$ is an eigenvalue of $q(T)$ if and only if $p_{q(T)}(\mu) = \prod_i (\mu - q(\lambda_i)) = 0$, that is, if and only if $\mu = q(\lambda_i)$ for some i, since a product of nonzero field elements is nonzero. The λ_i are exactly the eigenvalues of T, each listed at least once (Theorem 9.2.1). Hence $\operatorname{spec}(q(T)) = \{q(\lambda) : \lambda \in \operatorname{spec}(T)\}$. Over $\mathbb{C}$, p_T always splits by Corollary 6.5.5.

■

The formula for $p_{q(T)}$ carries more than the spectrum. The algebraic multiplicity of μ as an eigenvalue of $q(T)$ is the number of indices i with $q(\lambda_i) = \mu$ (Lemma 6.4.1), so several eigenvalues of T can merge into one eigenvalue of $q(T)$, and their multiplicities add. Combined with Theorem 9.3.3, applied to $q(T)$, whose characteristic polynomial splits by the theorem, we get the most used consequence.

Corollary 9.10.2. Trace and Determinant of a Polynomial in T

Under the hypotheses of Theorem 9.10.1, for every $q \in F[x]$,

$$\operatorname{tr} q(T) = q(\lambda_1) + \cdots + q(\lambda_n), \qquad \det q(T) = q(\lambda_1) \cdots q(\lambda_n).$$

In particular $\operatorname{tr} T^k = \lambda_1^k + \cdots + \lambda_n^k$ for every $k \in \mathbb{N}$.

Proof. By Theorem 9.10.1, $p_{q(T)} = \prod_i (x - q(\lambda_i))$ splits. Group equal values: if $\mu_1, \dots, \mu_s$ are the distinct values among the $q(\lambda_i)$, then μ_j occurs exactly $a_{q(T)}(\mu_j)$ times in the list (Lemma 6.4.1). By Theorem 9.3.3 applied to $q(T)$, $\operatorname{tr} q(T) = \sum_j a_{q(T)}(\mu_j)\, \mu_j = \sum_i q(\lambda_i)$ and $\det q(T) = \prod_j \mu_j^{a_{q(T)}(\mu_j)} = \prod_i q(\lambda_i)$. Take $q = x^k$ for the last statement.

■

Example. A Polynomial in a Non-Diagonalizable Matrix

Let $\mathbf{A} = \begin{pmatrix} 1 & 1 & 1 \\ 2 & 0 & 1 \\ -3 & 3 & 2 \end{pmatrix} \in M_3(\mathbb{R})$, for which Exercise (B1: Diagonalize or show impossible) found $p_{\mathbf{A}} = (x+1)(x-2)^2$ and showed that $\mathbf{A}$ is not diagonalizable. Without computing any eigenvectors, find $p_{\mathbf{A}^2}$, $\operatorname{tr} \mathbf{A}^2$, and $\det \mathbf{A}^2$. Then find $p_{q(\mathbf{A})}$ for $q = x^2 - x - 2$, and decide whether $q(\mathbf{A}) = \mathbf{O}$.

Solution. *The square.* $p_{\mathbf{A}}$ splits over $\mathbb{R}$ with the list $\lambda_1 = -1$, $\lambda_2 = \lambda_3 = 2$. By Theorem 9.10.1 with $q = x^2$, the list for $\mathbf{A}^2$ is $1, 4, 4$, so

$$p_{\mathbf{A}^2} = (x-1)(x-4)^2, \qquad \operatorname{tr}\mathbf{A}^2 = 1+4+4 = 9, \qquad \det\mathbf{A}^2 = 1\cdot 4\cdot 4 = 16$$

by Corollary 9.10.2. *Check.* Multiplying out,

$$\mathbf{A}^2 = \begin{pmatrix} 0 & 4 & 4 \\ -1 & 5 & 4 \\ -3 & 3 & 4 \end{pmatrix},$$

whose trace is $0+5+4 = 9$; and $\det\mathbf{A} = (-1)\cdot 2\cdot 2 = -4$ by Theorem 9.3.3, so $\det\mathbf{A}^2 = (\det\mathbf{A})^2 = 16$ by Theorem 7.5.4.

The polynomial q. Here $q = (x-2)(x+1)$, so $q(-1) = 0$ and $q(2) = 0$. The list for $q(\mathbf{A})$ is $0, 0, 0$, and $p_{q(\mathbf{A})} = x^3$: the only eigenvalue of $q(\mathbf{A})$ is 0. Yet $q(\mathbf{A}) \neq \mathbf{O}$. Indeed

$$q(\mathbf{A}) = \mathbf{A}^2 - \mathbf{A} - 2\mathbf{I} = \begin{pmatrix} -3 & 3 & 3 \\ -3 & 3 & 3 \\ 0 & 0 & 0 \end{pmatrix},$$

and one checks $q(\mathbf{A})^2 = \mathbf{O}$. The spectral mapping theorem sees only eigenvalues, and a nonzero matrix can have 0 as its only eigenvalue. That $q(\mathbf{A}) \neq \mathbf{O}$ here is no accident: if $q(\mathbf{A}) = \mathbf{O}$, the minimal polynomial of $\mathbf{A}$ would divide (Theorem 9.5.1) $(x-2)(x+1)$, which has distinct roots, and $\mathbf{A}$ would be diagonalizable by Theorem 9.8.2.

Warning. **Without splitting, spectral mapping fails.** For the real rotation $\mathbf{R}$ above, $\operatorname{spec}(\mathbf{R}) = \varnothing$ in $\mathbb{R}$, so $q(\operatorname{spec}(\mathbf{R})) = \varnothing$ for $q = x^2$, while $\operatorname{spec}(\mathbf{R}^2) = \operatorname{spec}(-\mathbf{I}) = \{-1\}$. Over $\mathbb{C}$ the theorem is restored: $\operatorname{spec}(\mathbf{R}) = \{i, -i\}$, both square to -1, and $p_{\mathbf{R}^2} = (x+1)^2$. Also, the theorem is about eigenvalues and algebraic multiplicities only: $\mathbf{A} = \operatorname{diag}(1, -1)$ has two one-dimensional eigenspaces, while $\mathbf{A}^2 = \mathbf{I}$ has the single eigenspace $\mathbb{R}^2$.

Quick check. Let $\mathbf{A} \in M_2(\mathbb{C})$ have eigenvalues i and $-i$. Find $p_{\mathbf{A}^2}$ and $\det(\mathbf{A}^2 + \mathbf{I})$.

Solution. The list for $\mathbf{A}$ is $i, -i$ (two distinct roots of the degree-2 polynomial $p_{\mathbf{A}}$). By Theorem 9.10.1 with $q = x^2$, the list for $\mathbf{A}^2$ is $-1, -1$, so $p_{\mathbf{A}^2} = (x+1)^2$. With $q = x^2 + 1$, Corollary 9.10.2 gives $\det(\mathbf{A}^2 + \mathbf{I}) = q(i)\,q(-i) = 0 \cdot 0 = 0$.

9.10.2 Traces of powers detect nilpotency

Chapter 8 proved, in Exercise (C2: When $\mathbf{A}$ commutes with $[\mathbf{A}, \mathbf{B}]$), that a matrix $\mathbf{C}$ over a field of characteristic 0 with $\operatorname{tr}\mathbf{C}^k = 0$ for every $k \geq 1$ is nilpotent, by a careful argument with annihilating polynomials that gave no reason for the result. Here is the view through eigenvalues, for complex matrices. The traces of the powers are the **power sums** $\lambda_1^k + \cdots + \lambda_n^k$ of the eigenvalues (Corollary 9.10.2), and a nilpotent matrix is one whose eigenvalues are all 0.

So the question becomes: if all power sums of n complex numbers vanish, must the numbers vanish? A Vandermonde determinant says yes.

Theorem 9.10.3. Nilpotent Matrices via Eigenvalues and Traces

Let $n \geq 1$ and $\mathbf{A} \in M_n(\mathbb{C})$. The following are equivalent:

(a) $\mathbf{A}$ is nilpotent, that is, $\mathbf{A}^s = \mathbf{O}$ for some $s \geq 1$;
(b) $\operatorname{spec}(\mathbf{A}) = \{0\}$;
(c) $p_{\mathbf{A}} = x^n$;
(d) $\mathbf{A}^n = \mathbf{O}$;
(e) $\operatorname{tr} \mathbf{A}^k = 0$ for $k = 1, 2, \ldots, n$.

The same holds for $\mathbf{A} \in M_n(F)$ when F is a subfield of $\mathbb{C}$, such as $\mathbb{Q}$ or $\mathbb{R}$, with $\operatorname{spec}(\mathbf{A})$ in (b) taken over $\mathbb{C}$.

Idea. **Plan:** (a) $\Rightarrow$ (b) $\Rightarrow$ (c) $\Rightarrow$ (d) $\Rightarrow$ (a), and (c) $\Rightarrow$ (e) $\Rightarrow$ (b). All but one arrow are short. For (e) $\Rightarrow$ (b), suppose some eigenvalue is nonzero. Group the nonzero eigenvalues into distinct values $\mu_1, \ldots, \mu_r$ with multiplicities $m_1, \ldots, m_r \geq 1$. The vanishing traces say $\sum_j m_j \mu_j^k = 0$ for $k = 1, \ldots, r$: a square linear system for the vector $(m_1, \ldots, m_r)$, whose matrix is a Vandermonde matrix times the invertible $\operatorname{diag}(\mu_j)$. So the m_j would be 0, but they are positive integers, and positive integers are nonzero in $\mathbb{C}$.

Proof.

(a) $\Rightarrow$ (b). Let $\mathbf{A}^s = \mathbf{O}$ with $s \geq 1$. If $\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$ with $\mathbf{v} \neq \mathbf{0}$, then $\mathbf{A}^s\mathbf{v} = \lambda^s\mathbf{v}$ by induction on s, so $\lambda^s\mathbf{v} = \mathbf{0}$ and $\lambda^s = 0$ by Theorem 1.1.9; hence $\lambda = 0$, since a product of nonzero scalars is nonzero (Theorem 0.7.1 (f)). So $\operatorname{spec}(\mathbf{A}) \subseteq \{0\}$. Conversely $\mathbf{A}$ is not injective: if it were, its s-fold composite $\mathbf{A}^s$ would be injective too (Theorem 0.4.3 (a)), whereas $\mathbf{A}^s = \mathbf{O}$ is not injective on $\mathbb{C}^n \neq \{\mathbf{0}\}$. Hence $0 \in \operatorname{spec}(\mathbf{A})$ by Theorem 9.2.1 ((b) $\Rightarrow$ (a)), and $\operatorname{spec}(\mathbf{A}) = \{0\}$.

(b) $\Rightarrow$ (c). By Corollary 6.5.5, $p_{\mathbf{A}} = (x - z_1)\cdots(x - z_n)$ with $z_i \in \mathbb{C}$. Each z_i is an eigenvalue (Theorem 9.2.1), hence 0, so $p_{\mathbf{A}} = x^n$.

(c) $\Rightarrow$ (d). By Corollary 9.7.3, $\mathbf{A} = \mathbf{P}\mathbf{U}\mathbf{P}^{-1}$ with $\mathbf{P}$ invertible and $\mathbf{U}$ upper triangular, and by Theorem 9.7.4 the diagonal entries of $\mathbf{U}$ are the roots of $p_{\mathbf{A}} = x^n$, all 0. Then $\mathbf{U}\mathbf{e}_1 = \mathbf{0}$ and $\mathbf{U}\mathbf{e}_j \in \operatorname{span}(\mathbf{e}_1, \ldots, \mathbf{e}_{j-1})$ for $j \geq 2$, since column j of $\mathbf{U}$ has zeros from row j down. By induction on j, $\mathbf{U}^j\mathbf{e}_j = \mathbf{0}$, hence $\mathbf{U}^n\mathbf{e}_j = \mathbf{U}^{n-j}\mathbf{U}^j\mathbf{e}_j = \mathbf{0}$ for every j, and $\mathbf{U}^n = \mathbf{O}$. By Proposition 4.2.6 (c), $\mathbf{A}^n = \mathbf{P}\mathbf{U}^n\mathbf{P}^{-1} = \mathbf{O}$.

(d) $\Rightarrow$ (a). Take $s = n$.

(e) $\Rightarrow$ (e). The list of eigenvalues of $\mathbf{A}$ with multiplicity is $0, \ldots, 0$, so by Corollary 9.10.2, $\operatorname{tr} \mathbf{A}^k = 0^k + \cdots + 0^k = 0$ for every $k \geq 1$.

(f) $\Rightarrow$ (b). By Corollary 6.5.5, $p_{\mathbf{A}}$ splits; let $\lambda_1, \ldots, \lambda_n$ be the list of its roots with multiplicity. Since $n \geq 1$, $\operatorname{spec}(\mathbf{A})$ is nonempty, so it suffices to show that every $\lambda_i = 0$. Suppose not, let $\mu_1, \ldots, \mu_r$ $(1 \leq r \leq n)$ be the distinct nonzero values among the λ_i, and let $m_j \geq 1$ be the number of indices i with $\lambda_i = \mu_j$. For $1 \leq k \leq r$, the zero

eigenvalues contribute $0^k = 0$, so by Corollary 9.10.2 and (e),

$$\sum_{j=1}^{r} m_j \mu_j^k = \operatorname{tr} \mathbf{A}^k = 0.$$

In matrix form, $\mathbf{W}\mathbf{m} = \mathbf{0}$, where $\mathbf{m} = (m_1, \dots, m_r) \in \mathbb{C}^r$ and $\mathbf{W} \in M_r(\mathbb{C})$ has (k, j)-entry $\mu_j^k = \mu_j^{k-1}\mu_j$. Hence $\mathbf{W} = V(\mu_1, \dots, \mu_r)^\top \operatorname{diag}(\mu_1, \dots, \mu_r)$, where $V(\mu_1, \dots, \mu_r)$ is the Vandermonde matrix with (j, k)-entry μ_j^{k-1} (Theorem 7.8.1). The μ_j are distinct, so V is invertible (Corollary 7.8.2), and so is $V^\top$, with inverse $(V^{-1})^\top$ (Theorem 0.9.5). The μ_j are nonzero, so $\operatorname{diag}(\mu_1, \dots, \mu_r)$ is invertible. Therefore $\mathbf{W}$ is invertible, and $\mathbf{m} = \mathbf{W}^{-1}\mathbf{0} = \mathbf{0}$. But m_1 is a positive integer, and $m_1 \neq 0$ in $\mathbb{C}$. This contradiction shows that every $\lambda_i = 0$, so $\operatorname{spec}(\mathbf{A}) = \{0\}$.

For a subfield $F \subseteq \mathbb{C}$, regard $\mathbf{A}$ as a complex matrix. Its powers, their traces, its characteristic polynomial, and the condition $\mathbf{A}^s = \mathbf{O}$ are computed by the same arithmetic in F and in $\mathbb{C}$, so (a), (c), (d), and (e) mean the same thing over F and over $\mathbb{C}$, and the equivalences carry over.

■

In the setting of Exercise (C2: When **A** commutes with [**A**, **B**]), where $\mathbf{C} = [\mathbf{A}, \mathbf{B}]$ commutes with $\mathbf{A}$, part (a) of that exercise gives $\operatorname{tr} \mathbf{C}^k = 0$ for all $k \geq 1$, and for complex matrices the theorem now shows at once that $\mathbf{C}^n = \mathbf{O}$. That exercise remains the route for a general field of characteristic 0.

Warning. **The characteristic matters.** Over $\mathbb{F}_p$, the identity $\mathbf{I}_p \in M_p(\mathbb{F}_p)$ has $\operatorname{tr} \mathbf{I}_p^k = p \cdot 1 = 0$ for every k, yet $\mathbf{I}_p^k = \mathbf{I}_p \neq \mathbf{O}$. The proof above breaks exactly where it used that the positive integer m_1 is nonzero in the field: here the multiplicity $m_1 = p$ of the eigenvalue 1 is 0 in $\mathbb{F}_p$.

9.10.3 AB and BA

Chapter 7, in Exercise (B3: **AB** and **BA**), showed that $p_{\mathbf{AB}} = p_{\mathbf{BA}}$ for square $\mathbf{A}$ and $\mathbf{B}$ **when $\mathbf{A}$ is invertible**, because then $\mathbf{BA} = \mathbf{A}^{-1}(\mathbf{AB})\mathbf{A}$ is similar to $\mathbf{AB}$. Without invertibility the two products need not be similar, and for rectangular $\mathbf{A}$ and $\mathbf{B}$ they do not even have the same size: if $\mathbf{A}$ is $m \times n$ and $\mathbf{B}$ is $n \times m$, then $\mathbf{AB}$ is $m \times m$ and $\mathbf{BA}$ is $n \times n$. Their characteristic polynomials have degrees m and n. Multiplying by powers of x repairs the degrees, and the identity then holds in general.

Theorem 9.10.4. AB and BA Have the Same Nonzero Eigenvalues

Let $m, n \geq 1$, $\mathbf{A} \in M_{m\times n}(F)$, and $\mathbf{B} \in M_{n\times m}(F)$. Then

$$x^n\, p_{\mathbf{AB}}(x) = x^m\, p_{\mathbf{BA}}(x).$$

Consequently:

(a) for every $\lambda \neq 0$ in F, $a_{\mathbf{AB}}(\lambda) = a_{\mathbf{BA}}(\lambda)$; in particular $\mathbf{AB}$ and $\mathbf{BA}$ have the same nonzero eigenvalues;

(b) if $m = n$, then $p_{\mathbf{AB}} = p_{\mathbf{BA}}$; and if $m \geq n$, then $p_{\mathbf{AB}} = x^{m-n}\, p_{\mathbf{BA}}$.

Idea. We want one big matrix that "contains" $\mathbf{AB}$ and another that "contains" $\mathbf{BA}$, and that are similar to each other. Put them in $(m+n) \times (m+n)$ block form:

$$\mathbf{M} = \begin{pmatrix} \mathbf{AB} & \mathbf{O} \\ \mathbf{B} & \mathbf{O} \end{pmatrix}, \qquad \mathbf{N} = \begin{pmatrix} \mathbf{O} & \mathbf{O} \\ \mathbf{B} & \mathbf{BA} \end{pmatrix}.$$

Both are block triangular, so their characteristic polynomials are $p_{\mathbf{AB}} \cdot x^n$ and $x^m \cdot p_{\mathbf{BA}}$. The invertible block matrix $\mathbf{S} = \begin{pmatrix} \mathbf{I}_m & \mathbf{A} \\ \mathbf{O} & \mathbf{I}_n \end{pmatrix}$ turns one into the other: $\mathbf{MS}$ and $\mathbf{SN}$ are both $\begin{pmatrix} \mathbf{AB} & \mathbf{ABA} \\ \mathbf{B} & \mathbf{BA} \end{pmatrix}$.

Proof. Let $\mathbf{M}, \mathbf{N}, \mathbf{S} \in M_{m+n}(F)$ be the block matrices displayed in the Idea, partitioned by m, n in rows and columns. By Theorem 8.1.1,

$$\mathbf{MS} = \begin{pmatrix} \mathbf{AB} \cdot \mathbf{I}_m + \mathbf{O} & \mathbf{AB} \cdot \mathbf{A} + \mathbf{O} \\ \mathbf{B} \cdot \mathbf{I}_m + \mathbf{O} & \mathbf{B} \cdot \mathbf{A} + \mathbf{O} \end{pmatrix} = \begin{pmatrix} \mathbf{AB} & \mathbf{ABA} \\ \mathbf{B} & \mathbf{BA} \end{pmatrix},$$
$$\mathbf{SN} = \begin{pmatrix} \mathbf{O} + \mathbf{AB} & \mathbf{O} + \mathbf{A} \cdot \mathbf{BA} \\ \mathbf{O} + \mathbf{B} & \mathbf{O} + \mathbf{BA} \end{pmatrix} = \begin{pmatrix} \mathbf{AB} & \mathbf{ABA} \\ \mathbf{B} & \mathbf{BA} \end{pmatrix},$$

using associativity (Theorem 0.9.4). So $\mathbf{MS} = \mathbf{SN}$. The matrix $\mathbf{S}$ is invertible: by Theorem 8.1.1, $\begin{pmatrix} \mathbf{I}_m & \mathbf{A} \\ \mathbf{O} & \mathbf{I}_n \end{pmatrix} \begin{pmatrix} \mathbf{I}_m & -\mathbf{A} \\ \mathbf{O} & \mathbf{I}_n \end{pmatrix} = \begin{pmatrix} \mathbf{I}_m & -\mathbf{A} + \mathbf{A} \\ \mathbf{O} & \mathbf{I}_n \end{pmatrix} = \mathbf{I}_{m+n}$, and Theorem 3.4.5 applies. Hence $\mathbf{S}^{-1}\mathbf{MS} = \mathbf{N}$, and $p_{\mathbf{M}} = p_{\mathbf{N}}$ by Theorem 7.10.3.

Now compute both sides. The matrix $x\mathbf{I}_{m+n} - \mathbf{M} = \begin{pmatrix} x\mathbf{I}_m - \mathbf{AB} & \mathbf{O} \\ -\mathbf{B} & x\mathbf{I}_n \end{pmatrix}$ is lower block triangular with square diagonal blocks, so by Theorem 7.8.4, which holds over the commutative ring $F[x]$ (the remarks before Example (A 3×3 Characteristic Polynomial)), and by Theorem 7.4.6 for $x\mathbf{I}_n$,

$$p_{\mathbf{M}}(x) = \det(x\mathbf{I}_m - \mathbf{AB}) \det(x\mathbf{I}_n) = p_{\mathbf{AB}}(x)\, x^n.$$

Likewise $x\mathbf{I}_{m+n} - \mathbf{N} = \begin{pmatrix} x\mathbf{I}_m & \mathbf{O} \\ -\mathbf{B} & x\mathbf{I}_n - \mathbf{BA} \end{pmatrix}$ gives $p_{\mathbf{N}}(x) = x^m\, p_{\mathbf{BA}}(x)$. Since $p_{\mathbf{M}} = p_{\mathbf{N}}$, this proves $x^n p_{\mathbf{AB}} = x^m p_{\mathbf{BA}}$.

(a) Let $\lambda \neq 0$. Then λ is not a root of x^n, so $\mathrm{mult}_\lambda(x^n) = 0$, and similarly for x^m. By Theorem 6.4.2,

$$a_{\mathbf{AB}}(\lambda) = \mathrm{mult}_\lambda(x^n p_{\mathbf{AB}}) = \mathrm{mult}_\lambda(x^m p_{\mathbf{BA}}) = a_{\mathbf{BA}}(\lambda).$$

A scalar is an eigenvalue exactly when its algebraic multiplicity is positive (Theorem 9.2.1, Definition 6.4.1), so the nonzero eigenvalues agree.

(b) If $m \geq n$, then $x^n p_{\mathbf{AB}} = x^n(x^{m-n} p_{\mathbf{BA}})$, and canceling the nonzero polynomial x^n (Corollary 0.8.4) gives $p_{\mathbf{AB}} = x^{m-n} p_{\mathbf{BA}}$. For $m = n$ this is $p_{\mathbf{AB}} = p_{\mathbf{BA}}$.

■

This pays off the promise of Exercise (B3: **AB** and **BA**): $p_{\mathbf{AB}} = p_{\mathbf{BA}}$ for **all** square $\mathbf{A}$ and $\mathbf{B}$, invertible or not. Comparing coefficients (Theorem 7.10.2) recovers $\operatorname{tr}(\mathbf{AB}) = \operatorname{tr}(\mathbf{BA})$ and $\det(\mathbf{AB}) = \det(\mathbf{BA})$ once more.

The eigenvectors travel between the two products as well, and by hand. If $\mathbf{ABw} = \lambda\mathbf{w}$ with $\lambda \neq 0$, then

$$\mathbf{BA}(\mathbf{Bw}) = \mathbf{B}(\mathbf{ABw}) = \lambda\,\mathbf{Bw},$$

and $\mathbf{Bw} \neq \mathbf{0}$, since $\mathbf{A}(\mathbf{Bw}) = \lambda\mathbf{w} \neq \mathbf{0}$. So $\mathbf{Bw}$ is an eigenvector of $\mathbf{BA}$ for the same λ: multiplying an eigenvector of $\mathbf{AB}$ by $\mathbf{B}$ produces one of $\mathbf{BA}$. For $\lambda = 0$ the recipe can fail, because $\mathbf{Bw}$ may be $\mathbf{0}$ — which is the warning below in another guise.

A rank-one example. Let $\mathbf{u}, \mathbf{v} \in F^n$ with $n \geq 2$. Take $\mathbf{A} = \mathbf{u} \in M_{n\times 1}(F)$ and $\mathbf{B} = \mathbf{v}^\top \in M_{1\times n}(F)$. Then $\mathbf{AB} = \mathbf{uv}^\top$ is $n \times n$, and $\mathbf{BA} = \mathbf{v}^\top\mathbf{u}$ is the 1×1 matrix whose entry is the number $c = \mathbf{v}^\top\mathbf{u}$, so $p_{\mathbf{BA}} = x - c$. Applying the theorem with its m equal to our n, and its n equal to 1, gives $x\,p_{\mathbf{uv}^\top} = x^n(x - c)$, and canceling x,

$$p_{\mathbf{uv}^\top}(x) = x^{n-1}\,(x - \mathbf{v}^\top\mathbf{u}).$$

For instance, the all-ones matrix $\mathbf{1}\mathbf{1}^\top \in M_n(\mathbb{R})$, with $\mathbf{1} = (1, \ldots, 1)$, has $p = x^{n-1}(x - n)$: eigenvalue 0 with multiplicity $n-1$, and eigenvalue n. Evaluating at -1 recovers Theorem 7.8.5:

$$\begin{aligned}\det(\mathbf{I} + \mathbf{uv}^\top) &= (-1)^n \det(-\mathbf{I} - \mathbf{uv}^\top) = (-1)^n\, p_{\mathbf{uv}^\top}(-1) \\ &= (-1)^n(-1)^{n-1}(-1 - \mathbf{v}^\top\mathbf{u}) = 1 + \mathbf{v}^\top\mathbf{u}.\end{aligned}$$

Warning. **The zero eigenvalue behaves differently, and AB need not be similar to BA.** Take $\mathbf{A} = \mathbf{E}_{12}$ and $\mathbf{B} = \mathbf{E}_{11}$ in $M_2(F)$. Then $\mathbf{AB} = \mathbf{E}_{12}\mathbf{E}_{11} = \mathbf{O}$ and $\mathbf{BA} = \mathbf{E}_{11}\mathbf{E}_{12} = \mathbf{E}_{12}$. Both have $p = x^2$, as the theorem demands. But $g_{\mathbf{AB}}(0) = 2$ while $g_{\mathbf{BA}}(0) = 1$, and $\mathbf{AB} = \mathbf{O}$ is similar only to $\mathbf{O}$, so $\mathbf{AB} \not\sim \mathbf{BA}$. For rectangular matrices even the algebraic multiplicities of 0 differ, by $|m - n|$.

Quick check. Let $\mathbf{A} \in M_{2\times 5}(\mathbb{R})$ and $\mathbf{B} \in M_{5\times 2}(\mathbb{R})$ with $p_{\mathbf{AB}} = x^2 - 3x + 2$. What is $p_{\mathbf{BA}}$, and what are $\operatorname{tr}(\mathbf{BA})$ and $\det(\mathbf{BA})$?

Solution. By Theorem 9.10.4 with $m = 2$, $n = 5$, $x^5 p_{\mathbf{AB}} = x^2 p_{\mathbf{BA}}$, so $p_{\mathbf{BA}} = x^3(x^2 - 3x + 2) = x^5 - 3x^4 + 2x^3$. By Theorem 7.10.2, $\operatorname{tr}(\mathbf{BA}) = 3$ (equal to $\operatorname{tr}(\mathbf{AB})$), and $\det(\mathbf{BA}) = (-1)^5 \cdot 0 = 0$, as it must be, since $\mathbf{BA}$ has rank at most $2 < 5$.

9.10.4 Left eigenvectors

An eigenvector multiplies $\mathbf{A}$ from the right: $\mathbf{Av} = \lambda\mathbf{v}$. Row vectors can multiply from the left, and in applications, especially Markov chains, the row version is often the one with a meaning.

A left eigenvector is a row that $\mathbf{A}$ *only stretches when it acts from the right of the row.*

Definition 9.10.1. Left Eigenvector

Let $\mathbf{A} \in M_n(F)$ and $\lambda \in F$. A **left eigenvector** of $\mathbf{A}$ for λ is a **nonzero** $\mathbf{y} \in F^n$ with

$$\mathbf{y}^\top \mathbf{A} = \lambda \mathbf{y}^\top.$$

To distinguish them, the eigenvectors of Definition 9.2.1 are also called **right eigenvectors**.

In words: $\mathbf{y}^\top$ is a $1 \times n$ row, $\mathbf{y}^\top \mathbf{A}$ is again a row, and we ask that it be a multiple of $\mathbf{y}^\top$. For example, $\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 3 & 2 \end{pmatrix}$ and $\mathbf{y} = (1, 1)$ give $\mathbf{y}^\top \mathbf{A} = \begin{pmatrix} 4 & 4 \end{pmatrix} = 4\mathbf{y}^\top$, so $(1, 1)$ is a left eigenvector for 4, whereas the right eigenvectors for 4 are the multiples of $(2, 3)$ (Example (Eigenvalues and Eigenspaces of a 2×2 Matrix)). For a diagonal matrix, $\mathbf{e}_i$ is both a left and a right eigenvector for the i-th diagonal entry. And $\mathbf{y} = \mathbf{0}$ always satisfies the equation, which is why it is excluded, exactly as for right eigenvectors.

Nonexample by minimal change. For the same $\mathbf{A}$, change $\mathbf{y} = (1, 1)$ to $\mathbf{y} = (1, -1)$. Then $\mathbf{y}^\top \mathbf{A} = \begin{pmatrix} 1-3 & 2-2 \end{pmatrix} = \begin{pmatrix} -2 & 0 \end{pmatrix}$, which is not a multiple of $\begin{pmatrix} 1 & -1 \end{pmatrix}$: the row is nonzero, but the defining equation $\mathbf{y}^\top \mathbf{A} = \lambda \mathbf{y}^\top$ has no solution λ. Note that $(1, -1)$ **is** a right eigenvector of $\mathbf{A}$, for -1; the two notions do not agree.

Transposing turns the definition into one we know.

Proposition 9.10.5. Left Eigenvectors Are Eigenvectors of the Transpose

Let $\mathbf{A} \in M_n(F)$ and $\lambda \in F$.

(a) $\mathbf{y}$ is a left eigenvector of $\mathbf{A}$ for λ if and only if $\mathbf{y}$ is an eigenvector of $\mathbf{A}^\top$ for λ.

(b) $\mathbf{A}$ has a left eigenvector for λ if and only if $\lambda \in \operatorname{spec}(\mathbf{A})$. Moreover $p_{\mathbf{A}^\top} = p_{\mathbf{A}}$, and $a_{\mathbf{A}^\top}(\lambda) = a_{\mathbf{A}}(\lambda)$, $g_{\mathbf{A}^\top}(\lambda) = g_{\mathbf{A}}(\lambda)$.

Proof.

(a) By Theorem 0.9.5, $(\mathbf{y}^\top \mathbf{A})^\top = \mathbf{A}^\top \mathbf{y}$ and $(\lambda \mathbf{y}^\top)^\top = \lambda \mathbf{y}$. Since transposing twice gives back the original, $\mathbf{y}^\top \mathbf{A} = \lambda \mathbf{y}^\top$ if and only if $\mathbf{A}^\top \mathbf{y} = \lambda \mathbf{y}$.

(b) First, $p_{\mathbf{A}^\top} = p_{\mathbf{A}}$: by Theorem 0.9.5, $(x\mathbf{I} - \mathbf{A})^\top = x\mathbf{I} - \mathbf{A}^\top$, and transposing does not change a determinant (Theorem 7.4.5, which holds over the commutative ring $F[x]$ by the remark after it), so

$$p_{\mathbf{A}^\top} = \det(x\mathbf{I} - \mathbf{A}^\top) = \det\left((x\mathbf{I} - \mathbf{A})^\top\right) = \det(x\mathbf{I} - \mathbf{A}) = p_{\mathbf{A}}.$$

Hence $a_{\mathbf{A}^\top}(\lambda) = a_{\mathbf{A}}(\lambda)$, and $\operatorname{spec}(\mathbf{A}^\top) = \operatorname{spec}(\mathbf{A})$ by Theorem 9.2.1. With (a), $\mathbf{A}$ has a left eigenvector for λ if and only if $\lambda \in \operatorname{spec}(\mathbf{A}^\top) = \operatorname{spec}(\mathbf{A})$. Finally, $\mathbf{A}^\top - \lambda\mathbf{I} = (\mathbf{A} - \lambda\mathbf{I})^\top$, which has the same rank as $\mathbf{A} - \lambda\mathbf{I}$ (Theorem 3.5.5), so by Theorem 3.5.7, $g_{\mathbf{A}^\top}(\lambda) = n - \operatorname{rank}(\mathbf{A} - \lambda\mathbf{I})^\top = n - \operatorname{rank}(\mathbf{A} - \lambda\mathbf{I}) = g_{\mathbf{A}}(\lambda)$.

■

Remark. For an operator T on V, left eigenvectors correspond to eigenvectors of the dual map T' on V^* (Chapter 5), whose matrix in the dual bases is the transpose (Theorem 5.4.2):

an eigenvector of T' is a functional $\varphi \neq 0$ with $\varphi \circ T = \lambda\varphi$.

Left and right eigenvectors for the **same** eigenvalue are usually unrelated, as $(1,1)$ and $(2,3)$ show. For **different** eigenvalues, though, they are always orthogonal in the sense of the plain product $\mathbf{y}^{\top}\mathbf{v}$.

Theorem 9.10.6. Left and Right Eigenvectors Are Biorthogonal

Let $\mathbf{A} \in M_n(F)$, let $\mathbf{y}$ be a left eigenvector of $\mathbf{A}$ for μ, and let $\mathbf{v}$ be a right eigenvector of $\mathbf{A}$ for λ. If $\lambda \neq \mu$, then $\mathbf{y}^{\top}\mathbf{v} = 0$.

Proof. Compute the 1×1 matrix $\mathbf{y}^{\top}\mathbf{A}\mathbf{v}$ in two ways, using associativity (Theorem 0.9.4):

$$\mu\,\mathbf{y}^{\top}\mathbf{v} = (\mathbf{y}^{\top}\mathbf{A})\mathbf{v} = \mathbf{y}^{\top}\mathbf{A}\mathbf{v} = \mathbf{y}^{\top}(\mathbf{A}\mathbf{v}) = \lambda\,\mathbf{y}^{\top}\mathbf{v}.$$

Hence $(\lambda - \mu)\,\mathbf{y}^{\top}\mathbf{v} = 0$. Since $\lambda - \mu \neq 0$, this forces $\mathbf{y}^{\top}\mathbf{v} = 0$ (Theorem 0.7.1). ∎

For the example, $(1,1)$ is a left eigenvector for 4 and $(1,-1)$ a right eigenvector for -1, and indeed $1 - 1 = 0$.

For a diagonalizable matrix the left eigenvectors come for free with the diagonalization, and they give a formula for all powers that we use in Section 11.

Proposition 9.10.7. Expansion in Left and Right Eigenvectors

Let $\mathbf{A} = \mathbf{P}\mathbf{D}\mathbf{P}^{-1}$ with $\mathbf{P} \in M_n(F)$ invertible and $\mathbf{D} = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$. Let $\mathbf{v}_1, \dots, \mathbf{v}_n$ be the columns of $\mathbf{P}$, and let $\mathbf{y}_1^{\top}, \dots, \mathbf{y}_n^{\top}$ be the rows of $\mathbf{P}^{-1}$. Then:

(a) $\mathbf{v}_i$ is a right eigenvector and $\mathbf{y}_i$ a left eigenvector of $\mathbf{A}$ for λ_i;
(b) $\mathbf{y}_i^{\top}\mathbf{v}_j = 1$ if $i = j$ and 0 if $i \neq j$;
(c) for every $q \in F[x]$,

$$q(\mathbf{A}) = \sum_{i=1}^{n} q(\lambda_i)\,\mathbf{v}_i\mathbf{y}_i^{\top}, \qquad \text{in particular} \qquad \mathbf{A}^k = \sum_{i=1}^{n} \lambda_i^k\,\mathbf{v}_i\mathbf{y}_i^{\top} \quad (k \in \mathbb{N}).$$

Proof.

(a) From $\mathbf{A}\mathbf{P} = \mathbf{P}\mathbf{D}$, column i reads $\mathbf{A}\mathbf{v}_i = \lambda_i\mathbf{v}_i$ (Section 4), and $\mathbf{v}_i \neq \mathbf{0}$ because $\mathbf{P}$ is invertible. From $\mathbf{P}^{-1}\mathbf{A} = \mathbf{D}\mathbf{P}^{-1}$, row i reads $\mathbf{y}_i^{\top}\mathbf{A} = \lambda_i\mathbf{y}_i^{\top}$ (Theorem 0.9.2), and $\mathbf{y}_i \neq \mathbf{0}$ because $\mathbf{P}^{-1}$ is invertible.

(b) The (i,j)-entry of $\mathbf{P}^{-1}\mathbf{P} = \mathbf{I}$ is row i of $\mathbf{P}^{-1}$ times column j of $\mathbf{P}$, that is, $\mathbf{y}_i^{\top}\mathbf{v}_j$.

(c) By Theorem 9.4.3, $q(\mathbf{A}) = \mathbf{P}\,q(\mathbf{D})\,\mathbf{P}^{-1}$. The columns of $\mathbf{P}\,q(\mathbf{D})$ are $q(\lambda_i)\mathbf{v}_i$, and the rows of $\mathbf{P}^{-1}$ are $\mathbf{y}_i^{\top}$, so Corollary 8.1.3 gives $q(\mathbf{A}) = \sum_i q(\lambda_i)\mathbf{v}_i\mathbf{y}_i^{\top}$. Take $q = x^k$. ∎

Part (b) is biorthogonality with a normalization: $\mathbf{y}_i^{\top}\mathbf{v}_i = 1$. Part (c) splits $\mathbf{A}$ into rank-one pieces $\mathbf{v}_i\mathbf{y}_i^{\top}$, one per eigenvalue, and each piece is simply multiplied by λ_i^k when $\mathbf{A}$ is raised to the power k. In the running example, $\mathbf{P} = \begin{pmatrix} 2 & 1 \\ 3 & -1 \end{pmatrix}$ and $\mathbf{D} = \operatorname{diag}(4, -1)$ give $\mathbf{P}^{-1} = \frac{1}{5}\begin{pmatrix} 1 & 1 \\ 3 & -2 \end{pmatrix}$, so $\mathbf{y}_1 = \frac{1}{5}(1, 1)$ and $\mathbf{y}_2 = \frac{1}{5}(3, -2)$, and

$$\mathbf{A}^k = \frac{4^k}{5}\begin{pmatrix} 2 & 2 \\ 3 & 3 \end{pmatrix} + \frac{(-1)^k}{5}\begin{pmatrix} 3 & -2 \\ -3 & 2 \end{pmatrix}.$$

For $k = 1$ this is $\frac{1}{5}\begin{pmatrix} 8-3 & 8+2 \\ 12+3 & 12-2 \end{pmatrix} = \begin{pmatrix} 1 & 2 \\ 3 & 2 \end{pmatrix}$, as it should be.

9.10.5 Exercises

9.10.5.1 A. Check your understanding

Exercise. A1

(a) State the Spectral Mapping Theorem, including its hypothesis on p_T.

(b) True or false: for every $\mathbf{A} \in M_n(\mathbb{R})$, every real eigenvalue of $\mathbf{A}^2$ is the square of a real eigenvalue of $\mathbf{A}$. Justify your answer.

(c) True or false: if $\mathbf{A} \in M_{2\times 3}(F)$ and $\mathbf{B} \in M_{3\times 2}(F)$, then $p_{\mathbf{AB}} = p_{\mathbf{BA}}$. Justify your answer.

(d) True or false: for all $\mathbf{A}, \mathbf{B} \in M_n(F)$, $\mathbf{AB}$ is similar to $\mathbf{BA}$. Justify your answer.

(e) Define a left eigenvector of $\mathbf{A} \in M_n(F)$, and say how left eigenvectors of $\mathbf{A}$ are related to $\mathbf{A}^{\top}$.

(f) True or false: a matrix $\mathbf{A} \in M_3(\mathbb{C})$ with $\operatorname{tr}\mathbf{A} = \operatorname{tr}\mathbf{A}^2 = \operatorname{tr}\mathbf{A}^3 = 0$ satisfies $\mathbf{A}^3 = \mathbf{O}$. Justify your answer.

Solution.

(a) If V is finite-dimensional with $\dim V = n \geq 1$, $T \in \mathcal{L}(V)$, and $p_T = \prod_{i=1}^{n}(x - \lambda_i)$ splits over F, then for every $q \in F[x]$, $p_{q(T)} = \prod_i (x - q(\lambda_i))$ and $\operatorname{spec}(q(T)) = q(\operatorname{spec}(T))$ (Theorem 9.10.1).

(b) False. The rotation $\mathbf{R} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ has no real eigenvalue, while $\mathbf{R}^2 = -\mathbf{I}$ has the real eigenvalue -1.

(c) False. They have different degrees, 2 and 3. The correct identity is $x^3 p_{\mathbf{AB}} = x^2 p_{\mathbf{BA}}$, that is, $p_{\mathbf{BA}} = x\, p_{\mathbf{AB}}$ (Theorem 9.10.4).

(d) False. $\mathbf{A} = \mathbf{E}_{12}$, $\mathbf{B} = \mathbf{E}_{11}$ give $\mathbf{AB} = \mathbf{O}$ and $\mathbf{BA} = \mathbf{E}_{12} \neq \mathbf{O}$, and the only matrix similar to $\mathbf{O}$ is $\mathbf{O}$.

(e) A nonzero $\mathbf{y} \in F^n$ with $\mathbf{y}^{\top}\mathbf{A} = \lambda\mathbf{y}^{\top}$ for some $\lambda \in F$. Equivalently, $\mathbf{y}$ is an eigenvector of $\mathbf{A}^{\top}$ for λ (Proposition 9.10.5).

(f) True. By Theorem 9.10.3 with $n = 3$, (e) $\Rightarrow$ (d).

9.10.5.2 B. Practice

Exercise. B1: Eigenvalues of a polynomial in A

Let $\mathbf{A} \in M_3(\mathbb{C})$ have $p_{\mathbf{A}} = (x-1)^2(x+2)$.

(a) Find $p_{\mathbf{A}^2+\mathbf{A}+\mathbf{I}}$, $\operatorname{tr}(\mathbf{A}^2+\mathbf{A}+\mathbf{I})$, and $\det(\mathbf{A}^2+\mathbf{A}+\mathbf{I})$. Is $\mathbf{A}^2+\mathbf{A}+\mathbf{I}$ invertible?

(b) Find $p_{\mathbf{A}^2+\mathbf{A}-2\mathbf{I}}$. Hence show that $\mathbf{A}^2+\mathbf{A}-2\mathbf{I}$ is nilpotent.

(c) Find $\operatorname{tr}\mathbf{A}^5$ and $\det(\mathbf{A}^{-1})$.

Solution. The list of eigenvalues of $\mathbf{A}$ with multiplicity is $1, 1, -2$, since $p_{\mathbf{A}}$ splits.

(a) With $q = x^2+x+1$, $q(1) = 3$ and $q(-2) = 4-2+1 = 3$. By Theorem 9.10.1, $p_{q(\mathbf{A})} = (x-3)^3$, and by Corollary 9.10.2, $\operatorname{tr} q(\mathbf{A}) = 9$ and $\det q(\mathbf{A}) = 27 \neq 0$. So $\mathbf{A}^2+\mathbf{A}+\mathbf{I}$ is invertible (Theorem 7.5.3).

(b) With $r = x^2+x-2 = (x-1)(x+2)$, $r(1) = r(-2) = 0$, so $p_{r(\mathbf{A})} = x^3$ by Theorem 9.10.1. By Theorem 9.10.3 ((c) $\Rightarrow$ (d)), $r(\mathbf{A})^3 = \mathbf{O}$, so $\mathbf{A}^2+\mathbf{A}-2\mathbf{I}$ is nilpotent.

(c) By Corollary 9.10.2, $\operatorname{tr}\mathbf{A}^5 = 1+1+(-2)^5 = -30$. Also $\det\mathbf{A} = 1\cdot 1\cdot(-2) = -2 \neq 0$ (Theorem 9.3.3), so $\mathbf{A}$ is invertible and $\det(\mathbf{A}^{-1}) = (\det\mathbf{A})^{-1} = -\frac{1}{2}$ (Corollary 7.5.6).

Exercise. B2: AB and BA for a 2×3 example

Let $\mathbf{A} = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 2 & 0 \end{pmatrix}$ and $\mathbf{B} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \\ 1 & 1 \end{pmatrix}$.

(a) Compute $\mathbf{AB}$, $\mathbf{BA}$, $p_{\mathbf{AB}}$, and $p_{\mathbf{BA}}$, and verify that $x^3 p_{\mathbf{AB}} = x^2 p_{\mathbf{BA}}$.

(b) Hence list the eigenvalues of $\mathbf{BA}$ without further computation, and find an eigenvector of $\mathbf{BA}$ for each nonzero eigenvalue by applying $\mathbf{B}$ to eigenvectors of $\mathbf{AB}$.

Solution.

(a) By row-times-column,

$$\mathbf{AB} = \begin{pmatrix} 2 & 1 \\ 0 & -2 \end{pmatrix}, \qquad \mathbf{BA} = \begin{pmatrix} 1 & 0 & 1 \\ 0 & -2 & 0 \\ 1 & 2 & 1 \end{pmatrix}.$$

$\mathbf{AB}$ is upper triangular, so $p_{\mathbf{AB}} = (x-2)(x+2) = x^2-4$. Expanding $\det(x\mathbf{I}-\mathbf{BA})$ along the second row, whose only nonzero entry is $x+2$ in position $(2,2)$,

$$\begin{aligned} p_{\mathbf{BA}} &= (x+2)\det\begin{pmatrix} x-1 & -1 \\ -1 & x-1 \end{pmatrix} \\ &= (x+2)\big((x-1)^2-1\big) = (x+2)\,x\,(x-2) = x^3-4x. \end{aligned}$$

Then $x^3 p_{\mathbf{AB}} = x^5-4x^3 = x^2(x^3-4x) = x^2 p_{\mathbf{BA}}$.

(b) The eigenvalues of $\mathbf{BA}$ are the roots of $x(x-2)(x+2)$: $0, 2, -2$, matching Theorem 9.10.4 (the nonzero ones are those of $\mathbf{AB}$). If $\mathbf{ABw} = \lambda\mathbf{w}$, then $\mathbf{BA}(\mathbf{Bw}) = \mathbf{B}(\mathbf{ABw}) = \lambda\mathbf{Bw}$, so $\mathbf{Bw}$ is an eigenvector of $\mathbf{BA}$ as long as $\mathbf{Bw} \neq \mathbf{0}$. For $\lambda = 2$:

$\mathbf{w} = (1,0)$, $\mathbf{Bw} = (1,0,1)$, and $\mathbf{BA}(1,0,1) = (2,0,2)$. For $\lambda = -2$: $\mathbf{AB}+2\mathbf{I} = \begin{pmatrix} 4 & 1 \\ 0 & 0 \end{pmatrix}$ gives $\mathbf{w} = (1,-4)$, $\mathbf{Bw} = (1,4,-3)$, and $\mathbf{BA}(1,4,-3) = (1-3,-8,1+8-3) = (-2,-8,6) = -2(1,4,-3)$.

Exercise. B3: Left eigenvectors of a 2×2 matrix

Let $\mathbf{A} = \begin{pmatrix} 0 & 1 \\ -6 & 5 \end{pmatrix} \in M_2(\mathbb{R})$, whose right eigenvectors $(1,2)$ for 2 and $(1,3)$ for 3 were found in Exercise (B1: Diagonalize or show impossible).

(a) Find a left eigenvector of $\mathbf{A}$ for each eigenvalue.

(b) Verify Theorem 9.10.6 for these vectors, and rescale the left eigenvectors so that $\mathbf{y}_i^\top \mathbf{v}_i = 1$.

(c) Hence write $\mathbf{A}^k$ as a combination of two rank-one matrices, and check the formula for $k = 0$ and $k = 1$.

Solution.

(a) By Proposition 9.10.5, we need eigenvectors of $\mathbf{A}^\top = \begin{pmatrix} 0 & -6 \\ 1 & 5 \end{pmatrix}$. For 2: $\mathbf{A}^\top - 2\mathbf{I} = \begin{pmatrix} -2 & -6 \\ 1 & 3 \end{pmatrix}$ gives $y_1 = -3y_2$, so $(3,-1)$. For 3: $\mathbf{A}^\top - 3\mathbf{I} = \begin{pmatrix} -3 & -6 \\ 1 & 2 \end{pmatrix}$ gives $y_1 = -2y_2$, so $(2,-1)$. Check: $\begin{pmatrix} 3 & -1 \end{pmatrix}\mathbf{A} = \begin{pmatrix} 6 & -2 \end{pmatrix}$ and $\begin{pmatrix} 2 & -1 \end{pmatrix}\mathbf{A} = \begin{pmatrix} 6 & -3 \end{pmatrix}$.

(b) $(3,-1)\cdot(1,3) = 0$ and $(2,-1)\cdot(1,2) = 0$, as the theorem predicts for different eigenvalues. Also $(3,-1)\cdot(1,2) = 1$ and $(2,-1)\cdot(1,3) = -1$, so take $\mathbf{y}_1 = (3,-1)$ and $\mathbf{y}_2 = (-2,1)$.

(c) With $\mathbf{P} = \begin{pmatrix} 1 & 1 \\ 2 & 3 \end{pmatrix}$, the rows of $\mathbf{P}^{-1} = \begin{pmatrix} 3 & -1 \\ -2 & 1 \end{pmatrix}$ are exactly $\mathbf{y}_1^\top, \mathbf{y}_2^\top$, and by Proposition 9.10.7,

$$\mathbf{A}^k = 2^k \begin{pmatrix} 1 \\ 2 \end{pmatrix} \begin{pmatrix} 3 & -1 \end{pmatrix} + 3^k \begin{pmatrix} 1 \\ 3 \end{pmatrix} \begin{pmatrix} -2 & 1 \end{pmatrix} = 2^k \begin{pmatrix} 3 & -1 \\ 6 & -2 \end{pmatrix} + 3^k \begin{pmatrix} -2 & 1 \\ -6 & 3 \end{pmatrix}.$$

For $k = 0$: $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$. For $k = 1$: $\begin{pmatrix} 6-6 & -2+3 \\ 12-18 & -4+9 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -6 & 5 \end{pmatrix} = \mathbf{A}$.

9.10.5.3 C. Going deeper

Exercise. C1: Traces of powers of AB and BA

Let $\mathbf{A} \in M_{m\times n}(F)$ and $\mathbf{B} \in M_{n\times m}(F)$.

(a) Prove that $\mathbf{B}(\mathbf{AB})^{k-1}\mathbf{A} = (\mathbf{BA})^k$ for every $k \geq 1$, and deduce that $\operatorname{tr}\left((\mathbf{AB})^k\right) = \operatorname{tr}\left((\mathbf{BA})^k\right)$.

(b) Now let $F = \mathbb{C}$. Give a second proof of the trace identity from Theorem 9.10.4 and Corollary 9.10.2.

Solution.

(a) By induction on k. For $k = 1$ both sides are $\mathbf{BA}$. If $\mathbf{B}(\mathbf{AB})^{k-1}\mathbf{A} = (\mathbf{BA})^k$, then by associativity

$$\begin{aligned}\mathbf{B}(\mathbf{AB})^k\mathbf{A} &= \mathbf{B}(\mathbf{AB})^{k-1}(\mathbf{AB})\mathbf{A} = (\mathbf{B}(\mathbf{AB})^{k-1}\mathbf{A})(\mathbf{BA}) \\ &= (\mathbf{BA})^k(\mathbf{BA}) = (\mathbf{BA})^{k+1}.\end{aligned}$$

Now $(\mathbf{AB})^k = \mathbf{A}\,(\mathbf{B}(\mathbf{AB})^{k-1})$, with $\mathbf{A}$ of size $m \times n$ and $\mathbf{B}(\mathbf{AB})^{k-1}$ of size $n \times m$. By Theorem 0.9.6 (3),

$$\operatorname{tr}\left((\mathbf{AB})^k\right) = \operatorname{tr}\left(\mathbf{A} \cdot \mathbf{B}(\mathbf{AB})^{k-1}\right) = \operatorname{tr}\left(\mathbf{B}(\mathbf{AB})^{k-1} \cdot \mathbf{A}\right) = \operatorname{tr}\left((\mathbf{BA})^k\right).$$

(b) Over $\mathbb{C}$, $p_{\mathbf{AB}}$ and $p_{\mathbf{BA}}$ split. Suppose $m \geq n$ (otherwise swap the roles of $\mathbf{A}$ and $\mathbf{B}$). By Theorem 9.10.4, $p_{\mathbf{AB}} = x^{m-n}p_{\mathbf{BA}}$, so if $\lambda_1, \ldots, \lambda_n$ is the list of eigenvalues of $\mathbf{BA}$ with multiplicity, then the list for $\mathbf{AB}$ is $\lambda_1, \ldots, \lambda_n$ together with $m - n$ zeros. For $k \geq 1$ the zeros contribute $0^k = 0$, so by Corollary 9.10.2, $\operatorname{tr}\left((\mathbf{AB})^k\right) = \sum_i \lambda_i^k = \operatorname{tr}\left((\mathbf{BA})^k\right)$.

Exercise. C2: Adding a commuting nilpotent matrix

Let $\mathbf{A}, \mathbf{N} \in M_n(\mathbb{C})$, $n \geq 1$, with $\mathbf{AN} = \mathbf{NA}$ and $\mathbf{N}$ nilpotent.

(a) Prove that $p_{\mathbf{A}+\mathbf{N}} = p_{\mathbf{A}}$; in particular $\operatorname{spec}(\mathbf{A}+\mathbf{N}) = \operatorname{spec}(\mathbf{A})$ and $\det(\mathbf{A}+\mathbf{N}) = \det \mathbf{A}$.

(b) Deduce that if $\mathbf{A}$ is invertible, so is $\mathbf{A} + \mathbf{N}$.

(c) Show that (a) fails without $\mathbf{AN} = \mathbf{NA}$.

Hint: triangularize $\mathbf{A}$ *and* $\mathbf{N}$ *together.*

Solution.

(a) By Theorem 9.9.4 there is an invertible $\mathbf{P}$ with $\mathbf{U} = \mathbf{P}^{-1}\mathbf{AP}$ and $\mathbf{U}' = \mathbf{P}^{-1}\mathbf{NP}$ upper triangular. Since $\mathbf{N}$ is nilpotent, $\operatorname{spec}(\mathbf{N}) = \{0\}$ (Exercise (C2: Nilpotent operators)), so every diagonal entry of $\mathbf{U}'$ is 0 by Theorem 9.7.4. Then $\mathbf{P}^{-1}(\mathbf{A}+\mathbf{N})\mathbf{P} = \mathbf{U} + \mathbf{U}'$ is upper triangular with the same diagonal as $\mathbf{U}$. By Theorem 7.10.3 and the examples after Definition 7.10.1,

$$p_{\mathbf{A}+\mathbf{N}} = p_{\mathbf{U}+\mathbf{U}'} = \prod_i (x - u_{ii}) = p_{\mathbf{U}} = p_{\mathbf{A}}.$$

Equal characteristic polynomials have the same roots (Theorem 9.2.1) and the same constant term $(-1)^n \det$ (Theorem 7.10.2).

(b) By (a), $\det(\mathbf{A} + \mathbf{N}) = \det \mathbf{A} \neq 0$, so $\mathbf{A} + \mathbf{N}$ is invertible (Theorem 7.5.3).

(c) Let $\mathbf{A} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}$ and $\mathbf{N} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$. Both are nilpotent, so $p_{\mathbf{A}} = x^2$, but $\mathbf{A} + \mathbf{N} =$

$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ has $p = x^2 - 1$ and eigenvalues ± 1. Here $\mathbf{AN} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \neq \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} = \mathbf{NA}$.

Exercise. C3: How many traces, and which field?

(a) Let $\omega = e^{2\pi i/3} \in \mathbb{C}$ and $\mathbf{C} = \operatorname{diag}(1, \omega, \omega^2)$. Show that $\operatorname{tr} \mathbf{C} = \operatorname{tr} \mathbf{C}^2 = 0$ but $\mathbf{C}$ is not nilpotent. Explain why this does not contradict Theorem 9.10.3.

(b) Let $n \geq 2$. Give a matrix $\mathbf{C} \in M_n(\mathbb{C})$ with $\operatorname{tr} \mathbf{C}^k = 0$ for $k = 1, \dots, n-1$ that is not nilpotent.

(c) Let $\mathbf{A}, \mathbf{B} \in M_n(\mathbb{C})$ and suppose $\mathbf{C} = \mathbf{AB} - \mathbf{BA}$ commutes with $\mathbf{A}$. Prove that $\mathbf{C}^n = \mathbf{O}$.

(d) Over $\mathbb{F}_2$, find a nonnilpotent $\mathbf{C} \in M_2(\mathbb{F}_2)$ with $\operatorname{tr} \mathbf{C}^k = 0$ for every $k \geq 1$.

Solution.

(a) $\omega^3 = 1$ and $\omega \neq 1$, so $0 = \omega^3 - 1 = (\omega - 1)(\omega^2 + \omega + 1)$ gives $1 + \omega + \omega^2 = 0$. Hence $\operatorname{tr} \mathbf{C} = 1 + \omega + \omega^2 = 0$, and $\operatorname{tr} \mathbf{C}^2 = 1 + \omega^2 + \omega^4 = 1 + \omega^2 + \omega = 0$. But $\mathbf{C}^k = \operatorname{diag}(1, \omega^k, \omega^{2k}) \neq \mathbf{O}$ for every k. No contradiction: the theorem, for $n = 3$, requires $\operatorname{tr} \mathbf{C}^k = 0$ for $k = 1, 2, 3$, and $\operatorname{tr} \mathbf{C}^3 = 1 + 1 + 1 = 3 \neq 0$.

(b) Let $\zeta = e^{2\pi i/n}$ and $\mathbf{C} = \operatorname{diag}(1, \zeta, \dots, \zeta^{n-1})$. For $1 \leq k \leq n-1$, $\zeta^k \neq 1$ and $(\zeta^k)^n = 1$, so $\operatorname{tr} \mathbf{C}^k = \sum_{j=0}^{n-1} (\zeta^k)^j = \frac{(\zeta^k)^n - 1}{\zeta^k - 1} = 0$ by the geometric sum. But $\mathbf{C}^k$ has the entry 1 in position $(1,1)$ for all k, so $\mathbf{C}$ is not nilpotent.

(c) By Exercise (C2: When $\mathbf{A}$ commutes with $[\mathbf{A}, \mathbf{B}]$) (a), $\operatorname{tr} \mathbf{C}^k = 0$ for every $k \geq 1$, in particular for $k = 1, \dots, n$. By Theorem 9.10.3 ((e) $\Rightarrow$ (d)), $\mathbf{C}^n = \mathbf{O}$.

(d) $\mathbf{C} = \mathbf{I}_2$: $\operatorname{tr} \mathbf{C}^k = 1 + 1 = 0$ in $\mathbb{F}_2$ for every k, and $\mathbf{C}^k = \mathbf{I}_2 \neq \mathbf{O}$.

9.11 Powers, Recurrences, and Markov Chains

Section 4 promised that diagonalization is a tool for computation, and the Fibonacci numbers gave a first taste. This section collects three situations in which a quantity evolves in discrete steps by a fixed linear rule: a vector recurrence $\mathbf{x}_{k+1} = \mathbf{A}\mathbf{x}_k$, a scalar recurrence such as $a_{k+3} = 2a_{k+2} + a_{k+1} - 2a_k$, and a system that moves at random between finitely many states. In each case the state after k steps is a matrix power applied to the starting state, and the eigenvalues decide what happens in the long run. We use only the explicit formula $\mathbf{A}^k = \mathbf{P}\mathbf{D}^k\mathbf{P}^{-1}$ and limits of sequences of numbers, one entry at a time.

9.11.1 Vector recurrences

A **linear recurrence** of first order is a rule $\mathbf{x}_{k+1} = \mathbf{A}\mathbf{x}_k$ for $k \in \mathbb{N}$, with $\mathbf{A} \in M_n(F)$ fixed and a starting vector $\mathbf{x}_0 \in F^n$. By induction on k, the unique sequence obeying the rule is $\mathbf{x}_k = \mathbf{A}^k\mathbf{x}_0$. The question is how to see this power, and diagonalization answers it by changing to coordinates in which $\mathbf{A}$ acts coordinate by coordinate.

Theorem 9.11.1. Solving a Diagonalizable Recurrence

Let $\mathbf{A} \in M_n(F)$ be diagonalizable, with $\mathbf{A} = \mathbf{PDP}^{-1}$, $\mathbf{D} = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$, and let $\mathbf{v}_1, \dots, \mathbf{v}_n$ be the columns of $\mathbf{P}$. Let $\mathbf{x}_0 \in F^n$, and let $(c_1, \dots, c_n) = \mathbf{P}^{-1}\mathbf{x}_0$, so that $\mathbf{x}_0 = c_1\mathbf{v}_1 + \cdots + c_n\mathbf{v}_n$. Then the sequence defined by $\mathbf{x}_{k+1} = \mathbf{A}\mathbf{x}_k$ is

$$\mathbf{x}_k = \mathbf{A}^k\mathbf{x}_0 = c_1\lambda_1^k\,\mathbf{v}_1 + c_2\lambda_2^k\,\mathbf{v}_2 + \cdots + c_n\lambda_n^k\,\mathbf{v}_n \qquad (k \in \mathbb{N}).$$

Proof. By induction on k, $\mathbf{x}_k = \mathbf{A}^k\mathbf{x}_0$: this holds for $k = 0$ since $\mathbf{A}^0 = \mathbf{I}$, and $\mathbf{x}_{k+1} = \mathbf{A}\mathbf{x}_k = \mathbf{A}\cdot\mathbf{A}^k\mathbf{x}_0 = \mathbf{A}^{k+1}\mathbf{x}_0$. By Theorem 9.4.3, $\mathbf{A}^k\mathbf{x}_0 = \mathbf{PD}^k\mathbf{P}^{-1}\mathbf{x}_0 = \mathbf{PD}^k\mathbf{c}$ with $\mathbf{c} = (c_1, \dots, c_n)$. Now $\mathbf{D}^k\mathbf{c} = (\lambda_1^k c_1, \dots, \lambda_n^k c_n)$, and by Theorem 0.9.3, $\mathbf{P}$ times this vector is $\sum_i \lambda_i^k c_i\,\mathbf{v}_i$. Finally, $\mathbf{Pc} = \mathbf{x}_0$ says $\mathbf{x}_0 = \sum_i c_i\mathbf{v}_i$, by the same result. ∎

In words: expand the starting vector in a basis of eigenvectors; each coordinate is then multiplied by its own eigenvalue at every step, independently of the others. Over $\mathbb{R}$ or $\mathbb{C}$, where absolute values are available, the long-run behavior is visible at once. Terms with $|\lambda_i| < 1$ die out, terms with $|\lambda_i| > 1$ grow, and the largest $|\lambda_i|$ with $c_i \neq 0$ wins. The worked Fibonacci example (Example (A Closed Formula for the Fibonacci Numbers)) is the case $n = 2$.

9.11.2 Scalar recurrences and companion matrices

A recurrence such as $a_{k+3} = 2a_{k+2} + a_{k+1} - 2a_k$ involves one number sequence but several previous terms. The trick of Example (A Closed Formula for the Fibonacci Numbers) turns it into a vector recurrence: remember the last few terms as one vector.

Let $n \geq 1$ and $c_0, \dots, c_{n-1} \in F$, and consider sequences $(a_k)_{k\in\mathbb{N}}$ in F with

$$a_{k+n} + c_{n-1}a_{k+n-1} + \cdots + c_1a_{k+1} + c_0a_k = 0 \qquad \text{for every } k \in \mathbb{N}. \tag{$*$}$$

Such a sequence is determined by its **initial values** $a_0, \dots, a_{n-1}$, because $(*)$ computes each later term from the n before it. The polynomial

$$p(x) = x^n + c_{n-1}x^{n-1} + \cdots + c_1x + c_0$$

is called the **characteristic polynomial of the recurrence**; the name is justified below. Put $\mathbf{x}_k = (a_k, a_{k+1}, \dots, a_{k+n-1}) \in F^n$. Then $(*)$ says exactly $\mathbf{x}_{k+1} = \mathbf{C}\mathbf{x}_k$, where

$$\mathbf{C} = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & & & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -c_0 & -c_1 & -c_2 & \cdots & -c_{n-1} \end{pmatrix}:$$

the first $n-1$ rows shift the window by one, and the last row is $(*)$ solved for a_{k+n}. For $n \geq 2$, $\mathbf{C}$ is the transpose of the companion matrix $\mathbf{C}(p)$ of Exercise (C1: Companion matrices), so $p_{\mathbf{C}} = p_{\mathbf{C}(p)} = p$ by Exercise (B2: Transpose) and Exercise (C1: Companion matrices) (for $n = 1$, $\mathbf{C} = (-c_0)$ and $p_{\mathbf{C}} = x + c_0 = p$). The eigenvalues of $\mathbf{C}$ are the roots of p.

Theorem 9.11.2. Scalar Recurrences with Distinct Roots

Let $n \geq 1$, $c_0, \ldots, c_{n-1} \in F$, and suppose $p = x^n + c_{n-1}x^{n-1} + \cdots + c_0$ has n **distinct** roots $\lambda_1, \ldots, \lambda_n \in F$. Then a sequence $(a_k)_{k\in\mathbb{N}}$ in F satisfies $(*)$ if and only if there are $b_1, \ldots, b_n \in F$ with

$$a_k = b_1\lambda_1^k + b_2\lambda_2^k + \cdots + b_n\lambda_n^k \qquad \text{for every } k \in \mathbb{N}.$$

In that case $b_1, \ldots, b_n$ are uniquely determined by the initial values $a_0, \ldots, a_{n-1}$.

Idea. Why powers of the roots? A geometric sequence $a_k = \lambda^k$ satisfies $(*)$ exactly when $\lambda^k p(\lambda) = 0$, so the roots of p give n solutions. To see that there are no others, use the matrix $\mathbf{C}$: its eigenvector for λ_i is the window $(1, \lambda_i, \ldots, \lambda_i^{n-1})$ of the geometric sequence, these eigenvectors form an invertible Vandermonde matrix, and Theorem 9.11.1 does the rest.

Proof. ($\Leftarrow$) For each i and k, $\lambda_i^{k+n} + c_{n-1}\lambda_i^{k+n-1} + \cdots + c_0\lambda_i^k = \lambda_i^k\, p(\lambda_i) = 0$. If $a_k = \sum_i b_i\lambda_i^k$, then the left side of $(*)$ is $\sum_i b_i \cdot 0 = 0$, after regrouping the finite sums.

($\Rightarrow$) Let (a_k) satisfy $(*)$, and let $\mathbf{x}_k$ and $\mathbf{C}$ be as above, so $\mathbf{x}_{k+1} = \mathbf{C}\mathbf{x}_k$. For each i, let $\mathbf{w}_i = (1, \lambda_i, \lambda_i^2, \ldots, \lambda_i^{n-1})$. Then $\mathbf{C}\mathbf{w}_i = \lambda_i\mathbf{w}_i$: for $r \leq n-1$, entry r of $\mathbf{C}\mathbf{w}_i$ is entry $r+1$ of $\mathbf{w}_i$, which is $\lambda_i^r = \lambda_i \cdot \lambda_i^{r-1}$; and the last entry is $-(c_0 + c_1\lambda_i + \cdots + c_{n-1}\lambda_i^{n-1}) = \lambda_i^n - p(\lambda_i) = \lambda_i^n = \lambda_i \cdot \lambda_i^{n-1}$. Let $\mathbf{W}$ be the matrix with columns $\mathbf{w}_1, \ldots, \mathbf{w}_n$. Its (r, i)-entry is λ_i^{r-1}, so $\mathbf{W} = \mathbf{V}(\lambda_1, \ldots, \lambda_n)^\top$ in the notation of Theorem 7.8.1, and $\mathbf{W}$ is invertible, since the λ_i are distinct: $\det \mathbf{W} = \det \mathbf{V}(\lambda_1, \ldots, \lambda_n) \neq 0$ by Theorem 7.4.5 and Theorem 7.8.1, and Theorem 7.5.2 applies. The columns of $\mathbf{C}\mathbf{W}$ are $\lambda_i\mathbf{w}_i$, which are the columns of $\mathbf{W}\,\mathrm{diag}(\lambda_1, \ldots, \lambda_n)$, so $\mathbf{C} = \mathbf{W}\,\mathrm{diag}(\lambda_1, \ldots, \lambda_n)\,\mathbf{W}^{-1}$.

By Theorem 9.11.1, with $(b_1, \ldots, b_n) = \mathbf{W}^{-1}\mathbf{x}_0$,

$$\mathbf{x}_k = b_1\lambda_1^k\,\mathbf{w}_1 + \cdots + b_n\lambda_n^k\,\mathbf{w}_n.$$

The first entry of $\mathbf{x}_k$ is a_k, and the first entry of each $\mathbf{w}_i$ is 1, so $a_k = \sum_i b_i\lambda_i^k$.

For uniqueness, suppose $a_k = \sum_i b_i\lambda_i^k$ for all k. Taking $k = 0, \ldots, n-1$ gives $\mathbf{W}\mathbf{b} = (a_0, \ldots, a_{n-1}) = \mathbf{x}_0$, so $\mathbf{b} = \mathbf{W}^{-1}\mathbf{x}_0$ is determined by the initial values. ∎

So solving a recurrence with distinct roots comes down to three steps: find the roots of p, write the general form $\sum b_i\lambda_i^k$, and fit $b_1, \ldots, b_n$ to the initial values by solving a Vandermonde system.

Example. A Third-Order Recurrence

Find a formula for the sequence with $a_0 = 3$, $a_1 = 2$, $a_2 = 6$ and

$$a_{k+3} = 2a_{k+2} + a_{k+1} - 2a_k \qquad (k \in \mathbb{N}).$$

Solution. In the form $(*)$, $a_{k+3} - 2a_{k+2} - a_{k+1} + 2a_k = 0$, so $p(x) = x^3 - 2x^2 - x + 2$. Grouping, $p = x^2(x-2) - (x-2) = (x-2)(x-1)(x+1)$, with the distinct roots $1, -1, 2$. The window matrix is

$$\mathbf{C} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -2 & 1 & 2 \end{pmatrix},$$

with eigenvectors $(1,1,1)$, $(1,-1,1)$, $(1,2,4)$; for instance $\mathbf{C}(1,2,4) = (2,4,-2+2+8) = 2(1,2,4)$.

By Theorem 9.11.2, $a_k = b_1 + b_2(-1)^k + b_3 2^k$. The initial values give the Vandermonde system

$$b_1 + b_2 + b_3 = 3, \qquad b_1 - b_2 + 2b_3 = 2, \qquad b_1 + b_2 + 4b_3 = 6.$$

Subtracting the first equation from the third gives $3b_3 = 3$, so $b_3 = 1$. Then $b_1 + b_2 = 2$ and $b_1 - b_2 = 0$, so $b_1 = b_2 = 1$. Hence

$$a_k = 1 + (-1)^k + 2^k.$$

Check. The recurrence gives $a_3 = 2 \cdot 6 + 2 - 2 \cdot 3 = 8$ and $a_4 = 2 \cdot 8 + 6 - 2 \cdot 2 = 18$; the formula gives $1 - 1 + 8 = 8$ and $1 + 1 + 16 = 18$. For large k, $a_k \approx 2^k$: the root of largest absolute value dominates.

Repeated roots. When p has a repeated root, $\mathbf{C}$ is not diagonalizable: for a root λ, every eigenvector $(w_1, \dots, w_n)$ of $\mathbf{C}$ satisfies $w_{r+1} = \lambda w_r$, so it is a multiple of $(1, \lambda, \dots, \lambda^{n-1})$, and $g_{\mathbf{C}}(\lambda) = 1$ whatever $a_{\mathbf{C}}(\lambda)$ is. The general answer involves terms $k\lambda^k, k^2\lambda^k, \dots$, and it follows from the Jordan canonical form of Chapter 10. The 2×2 case can be done by hand.

Example. A Repeated Root

Over $\mathbb{R}$, find all sequences with $a_{k+2} = 4a_{k+1} - 4a_k$ for every $k \in \mathbb{N}$.

Solution. Here $p = x^2 - 4x + 4 = (x-2)^2$, and the window matrix is $\mathbf{C} = \begin{pmatrix} 0 & 1 \\ -4 & 4 \end{pmatrix}$. It is not diagonalizable, since $\mathbf{C} - 2\mathbf{I} = \begin{pmatrix} -2 & 1 \\ -4 & 2 \end{pmatrix}$ has rank 1, so $g(2) = 1 < 2 = a(2)$. But $\mathbf{N} = \mathbf{C} - 2\mathbf{I}$ satisfies

$$\mathbf{N}^2 = \begin{pmatrix} 4-4 & -2+2 \\ 8-8 & -4+4 \end{pmatrix} = \mathbf{O}.$$

Claim: $\mathbf{C}^k = 2^k\mathbf{I} + k2^{k-1}\mathbf{N}$ for all $k \in \mathbb{N}$. For $k = 0$ both sides are $\mathbf{I}$. If it holds for k, then, since $\mathbf{C} = 2\mathbf{I} + \mathbf{N}$ and $\mathbf{N}^2 = \mathbf{O}$,

$$\begin{aligned} \mathbf{C}^{k+1} &= (2^k\mathbf{I} + k2^{k-1}\mathbf{N})(2\mathbf{I} + \mathbf{N}) \\ &= 2^{k+1}\mathbf{I} + 2^k\mathbf{N} + k2^k\mathbf{N} + k2^{k-1}\mathbf{N}^2 = 2^{k+1}\mathbf{I} + (k+1)2^k\mathbf{N}. \end{aligned}$$

Now $\mathbf{x}_k = \mathbf{C}^k\mathbf{x}_0$ with $\mathbf{x}_0 = (a_0, a_1)$, and a_k is the first entry. Row 1 of $\mathbf{C}^k$ is $(2^k - 2k2^{k-1},\ k2^{k-1}) = ((1-k)2^k,\ k2^{k-1})$, so

$$a_k = (1-k)2^k a_0 + k2^{k-1}a_1 = \Big(a_0 + (\tfrac{a_1}{2} - a_0)k\Big)2^k.$$

So the solutions are exactly the sequences $a_k = (b_1 + b_2 k)2^k$ with $b_1, b_2 \in \mathbb{R}$: every solution has this form with $b_1 = a_0$, $b_2 = a_1/2 - a_0$, and every such sequence is the solution with $a_0 = b_1$, $a_1 = 2b_1 + 2b_2$. For example, $a_0 = 1$, $a_1 = 4$ gives $a_k = (1 + k)2^k$, and indeed $a_2 = 4 \cdot 4 - 4 \cdot 1 = 12 = 3 \cdot 4$. The new term $k2^k$ is the footprint of the missing eigenvector, just as the entry k was for $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}^k$ in Section 4.

9.11.3 Stochastic matrices

Now a system that moves at random. A shared bicycle is parked each night at one of a few stations; a machine is either working or under repair. We do not know the state for sure, but we know the probability of each state, and a fixed table of probabilities for moving from one state to another in one step.

Take states $1, \ldots, n$. Record the probabilities of the states at step k as a column $\mathbf{x}_k \in \mathbb{R}^n$, and let a_{ij} be the probability of moving to state i in one step, given that the system is in state j. The probability of being in state i at step $k + 1$ is the sum over the previous state j of "was in j" times "moved from j to i", that is, $\sum_j a_{ij}(\mathbf{x}_k)_j$. So $\mathbf{x}_{k+1} = \mathbf{A}\mathbf{x}_k$: a linear recurrence. The matrix $\mathbf{A}$ has two features: its entries are probabilities, and from each state j the system must go **somewhere**, so column j adds up to 1.

A stochastic matrix is a table of transition probabilities, one column for each starting state.

Definition 9.11.1. Stochastic Matrix, Probability Vector

Let $n \geq 1$.

(a) A vector $\mathbf{x} \in \mathbb{R}^n$ is a **probability vector** if $x_i \geq 0$ for every i and $x_1 + \cdots + x_n = 1$.

(b) A matrix $\mathbf{A} \in M_n(\mathbb{R})$ is **stochastic** if every column of $\mathbf{A}$ is a probability vector: $a_{ij} \geq 0$ for **all** i, j, and $\sum_{i=1}^n a_{ij} = 1$ for **every** j.

In words: the entries are nonnegative and each **column** sums to 1. With $\mathbf{1} = (1, \ldots, 1) \in \mathbb{R}^n$, the column condition reads $\mathbf{1}^\top \mathbf{A} = \mathbf{1}^\top$, and $\mathbf{1}^\top \mathbf{x} = 1$ says that the entries of $\mathbf{x}$ add up to 1.

Examples.

- $\begin{pmatrix} 9/10 & 1/5 \\ 1/10 & 4/5 \end{pmatrix}$: columns $(9/10, 1/10)$ and $(1/5, 4/5)$ are probability vectors.
- **Degenerate cases.** $\mathbf{I}_n$ is stochastic: every state stays put. The 1×1 matrix (1) is the only 1×1 stochastic matrix. Every **permutation matrix** is stochastic, since each column is a standard basis vector; it moves the system deterministically.
- $\frac{1}{4}\begin{pmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 1 & 1 & 3 \end{pmatrix}$: the columns sum to $\frac{4}{4}$, and all entries are nonnegative.

Nonexample by minimal change. Transpose the first example: $\begin{pmatrix} 9/10 & 1/10 \\ 1/5 & 4/5 \end{pmatrix}$ still has nonnegative entries, and its **rows** sum to 1, but its first column sums to $11/10$. It fails the column condition. Likewise $\begin{pmatrix} 3/2 & 0 \\ -1/2 & 1 \end{pmatrix}$ has columns summing to 1 but a negative entry.

Why this convention. Many books write probability vectors as **rows** and multiply on the left, $\mathbf{x}_{k+1}^{\top} = \mathbf{x}_k^{\top}\mathbf{M}$; their "stochastic" matrices have rows summing to 1, and they are the transposes of ours. We keep column vectors and $\mathbf{x}_{k+1} = \mathbf{A}\mathbf{x}_k$, in line with the rest of the book, so our stochastic matrices are **column-stochastic**. Translating between the conventions is a transpose, which changes neither eigenvalues nor their algebraic multiplicities (Proposition 9.10.5).

The first properties follow from $\mathbf{1}^{\top}\mathbf{A} = \mathbf{1}^{\top}$.

Proposition 9.11.3. Basic Properties of Stochastic Matrices

Let $\mathbf{A}, \mathbf{B} \in M_n(\mathbb{R})$ be stochastic and $\mathbf{x} \in \mathbb{R}^n$ a probability vector. Then:

(a) $\mathbf{A}\mathbf{x}$ is a probability vector;

(b) $\mathbf{A}\mathbf{B}$ is stochastic, and so is $\mathbf{A}^k$ for every $k \in \mathbb{N}$;

(c) $\mathbf{1}$ is a left eigenvector of $\mathbf{A}$ for the eigenvalue 1; in particular $1 \in \operatorname{spec}(\mathbf{A})$.

Proof.

(a) Each entry $(\mathbf{A}\mathbf{x})_i = \sum_j a_{ij}x_j$ is a sum of products of nonnegative numbers, so it is nonnegative. By associativity, $\mathbf{1}^{\top}(\mathbf{A}\mathbf{x}) = (\mathbf{1}^{\top}\mathbf{A})\mathbf{x} = \mathbf{1}^{\top}\mathbf{x} = 1$.

(b) By Theorem 0.9.2, column j of $\mathbf{A}\mathbf{B}$ is $\mathbf{A}\mathbf{b}_j$, where $\mathbf{b}_j$ is column j of $\mathbf{B}$, a probability vector. By (a), $\mathbf{A}\mathbf{b}_j$ is a probability vector, so $\mathbf{A}\mathbf{B}$ is stochastic. By induction on k, $\mathbf{A}^k$ is stochastic: $\mathbf{A}^0 = \mathbf{I}$ is, and $\mathbf{A}^{k+1} = \mathbf{A}^k\mathbf{A}$.

(c) $\mathbf{1}^{\top}\mathbf{A} = \mathbf{1}^{\top}$ with $\mathbf{1} \neq \mathbf{0}$ is Definition 9.10.1 with $\lambda = 1$, and Proposition 9.10.5 gives $1 \in \operatorname{spec}(\mathbf{A})$.

■

Warning. **The all-ones vector is a left eigenvector, not a right one.** For $\mathbf{A} = \begin{pmatrix} 9/10 & 1/5 \\ 1/10 & 4/5 \end{pmatrix}$, $\mathbf{A}\mathbf{1} = (11/10, 9/10) \neq \mathbf{1}$. The right eigenvectors for 1 are the multiples of $(2, 1)$, and they carry the interesting information: the long-run probabilities. Only when the **rows** also sum to 1 is $\mathbf{1}$ a right eigenvector.

9.11.4 Markov chains

Definition 9.11.2. Markov Chain, Steady State

Let $\mathbf{A} \in M_n(\mathbb{R})$ be stochastic. The **Markov chain** with **transition matrix** $\mathbf{A}$ and initial probability vector $\mathbf{x}_0$ is the sequence of probability vectors $\mathbf{x}_{k+1} = \mathbf{A}\mathbf{x}_k$, that is, $\mathbf{x}_k = \mathbf{A}^k\mathbf{x}_0$. A **steady state** of $\mathbf{A}$ is a probability vector $\mathbf{v}$ with $\mathbf{A}\mathbf{v} = \mathbf{v}$.

Each $\mathbf{x}_k$ is a probability vector by Proposition 9.11.3. A steady state is a right eigenvector for the eigenvalue 1, scaled so that its entries add up to 1 and required to have nonnegative entries; if the chain starts there, it stays there.

To talk about the long run we need limits, and we use the simplest kind. A sequence of complex

matrices $\mathbf{M}_k \in M_{m\times n}(\mathbb{C})$ **converges entrywise** to $\mathbf{M}$ if, for every position (r,s), the number sequence $(\mathbf{M}_k)_{rs}$ converges to $\mathbf{M}_{rs}$. We use four facts about sequences of complex numbers from calculus: sums and constant multiples of convergent sequences converge to the sums and multiples of the limits; $z_k \to 0$ if and only if $|z_k| \to 0$; $r^k \to 0$ for real $0 \le r < 1$; and a convergent sequence of real numbers ≥ 0 has a real limit ≥ 0. In particular, if $|\lambda| < 1$, then $|\lambda^k| = |\lambda|^k \to 0$ (Theorem 0.6.2), so $\lambda^k \to 0$.

Example. A Two-State Chain

A machine is either working (state 1) or under repair (state 2) each day. A working machine is still working the next day with probability $9/10$; a machine under repair is back to work the next day with probability $1/5$. So the transition matrix is

$$\mathbf{A} = \begin{pmatrix} 9/10 & 1/5 \\ 1/10 & 4/5 \end{pmatrix}.$$

Find $\mathbf{A}^k$, its entrywise limit, and the long-run probability that the machine is working, if it is working on day 0.

Solution. Eigenvalues. $\operatorname{tr}\mathbf{A} = 17/10$ and $\det\mathbf{A} = 36/50 - 1/50 = 7/10$, so, by the 2×2 formula $p_{\mathbf{A}} = x^2 - (\operatorname{tr}\mathbf{A})x + \det\mathbf{A}$ (the examples after Definition 7.10.1), $p_{\mathbf{A}} = x^2 - \frac{17}{10}x + \frac{7}{10} = (x-1)(x-\frac{7}{10})$. The eigenvalue 1 was guaranteed by Proposition 9.11.3.

Eigenvectors. $\mathbf{A} - \mathbf{I} = \begin{pmatrix} -1/10 & 1/5 \\ 1/10 & -1/5 \end{pmatrix}$ gives $x = 2y$, eigenvector $(2,1)$. $\mathbf{A} - \frac{7}{10}\mathbf{I} = \begin{pmatrix} 1/5 & 1/5 \\ 1/10 & 1/10 \end{pmatrix}$ gives $y = -x$, eigenvector $(1,-1)$. The eigenvalues are distinct, so

$$\mathbf{P} = \begin{pmatrix} 2 & 1 \\ 1 & -1 \end{pmatrix}, \qquad \mathbf{D} = \operatorname{diag}(1, \tfrac{7}{10}), \qquad \mathbf{P}^{-1} = \frac{1}{3}\begin{pmatrix} 1 & 1 \\ 1 & -2 \end{pmatrix},$$

using Theorem 0.9.8 with $\det\mathbf{P} = -3$.

Powers. By Theorem 9.4.3, writing $t = (7/10)^k$,

$$\mathbf{A}^k = \mathbf{P}\begin{pmatrix} 1 & 0 \\ 0 & t \end{pmatrix}\mathbf{P}^{-1} = \frac{1}{3}\begin{pmatrix} 2 & t \\ 1 & -t \end{pmatrix}\begin{pmatrix} 1 & 1 \\ 1 & -2 \end{pmatrix} = \frac{1}{3}\begin{pmatrix} 2+t & 2-2t \\ 1-t & 1+2t \end{pmatrix}.$$

Check: for $k=1$, $t = 7/10$ and $\frac{1}{3}\begin{pmatrix} 27/10 & 6/10 \\ 3/10 & 24/10 \end{pmatrix} = \mathbf{A}$; each column of $\mathbf{A}^k$ sums to 1, as Proposition 9.11.3 requires.

The limit. Since $0 \le 7/10 < 1$, $t \to 0$, so entrywise

$$\mathbf{A}^k \to \frac{1}{3}\begin{pmatrix} 2 & 2 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 2/3 \\ 1/3 \end{pmatrix}\begin{pmatrix} 1 & 1 \end{pmatrix}.$$

Starting from $\mathbf{x}_0 = \mathbf{e}_1$, $\mathbf{x}_k = \mathbf{A}^k\mathbf{e}_1 = (\frac{2+t}{3}, \frac{1-t}{3}) \to (\frac{2}{3}, \frac{1}{3})$. In the long run the machine works on two days out of three. The limit does not depend on the start: for any probability vector $\mathbf{x}_0 = (s, 1-s)$, $\mathbf{A}^k\mathbf{x}_0 \to \frac{1}{3}(2s + 2(1-s),\ s + (1-s)) = (\frac{2}{3}, \frac{1}{3})$. The vector $\mathbf{v} = (\frac{2}{3}, \frac{1}{3})$ is the steady state, the eigenvector $(2,1)$ scaled to sum 1.

The example shows the general mechanism. The eigenvalue 1 survives in $\mathbf{D}^k$, all others fade, and the surviving rank-one piece is (right eigenvector for 1) times (left eigenvector for 1), which for a stochastic matrix is the steady state times $\mathbf{1}^\top$.

Theorem 9.11.4. Long-Run Behavior of a Diagonalizable Markov Chain

Let $\mathbf{A} \in M_n(\mathbb{R})$, $n \geq 1$, be stochastic. Suppose that $\mathbf{A}$ is diagonalizable over $\mathbb{C}$, that $a_{\mathbf{A}}(1) = 1$, and that every other eigenvalue $\lambda \in \mathbb{C}$ of $\mathbf{A}$ has $|\lambda| < 1$. Then:

(a) $\mathbf{A}$ has exactly one steady state $\mathbf{v}$;

(b) $\mathbf{A}^k \to \mathbf{v}\mathbf{1}^\top$ entrywise as $k \to \infty$;

(c) for every probability vector $\mathbf{x}_0$, the Markov chain $\mathbf{x}_k = \mathbf{A}^k\mathbf{x}_0$ converges entrywise to $\mathbf{v}$.

Idea. Use the expansion $\mathbf{A}^k = \sum_i \lambda_i^k \mathbf{v}_i\mathbf{y}_i^\top$ of Section 10. The terms with $|\lambda_i| < 1$ tend to $\mathbf{O}$, which leaves the single term for $\lambda_1 = 1$. Its left eigenvector $\mathbf{y}_1$ must be a multiple of $\mathbf{1}$, because the left eigenvectors for 1 form a line and $\mathbf{1}$ is one of them. Rescaling turns $\mathbf{v}_1\mathbf{y}_1^\top$ into $\mathbf{v}\mathbf{1}^\top$, and the nonnegativity of $\mathbf{v}$ comes for free as a limit of nonnegative entries.

Proof. Since $\mathbf{A}$ is diagonalizable over $\mathbb{C}$, there are an invertible $\mathbf{P} \in M_n(\mathbb{C})$ and $\mathbf{D} = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ with $\mathbf{A} = \mathbf{P}\mathbf{D}\mathbf{P}^{-1}$. A diagonal matrix is upper triangular, so each eigenvalue occurs on the diagonal of $\mathbf{D}$ as often as its algebraic multiplicity (Theorem 9.7.4 (b)); in particular 1 occurs exactly once, because $a_{\mathbf{A}}(1) = 1$. Permuting the columns of $\mathbf{P}$ and the diagonal entries of $\mathbf{D}$ together leaves $\mathbf{A} = \mathbf{P}\mathbf{D}\mathbf{P}^{-1}$ intact, so we may assume $\lambda_1 = 1$, and then $|\lambda_i| < 1$ for $i \geq 2$ by hypothesis. Let $\mathbf{v}_i$ be the columns of $\mathbf{P}$ and $\mathbf{y}_i^\top$ the rows of $\mathbf{P}^{-1}$. By Proposition 9.10.7, for every k,

$$\mathbf{A}^k = \mathbf{v}_1\mathbf{y}_1^\top + \sum_{i=2}^{n} \lambda_i^k \mathbf{v}_i\mathbf{y}_i^\top.$$

Each entry of the sum is $\sum_{i\geq 2} \lambda_i^k z_i$ for fixed complex numbers z_i, and $\lambda_i^k \to 0$ for $i \geq 2$. Hence $\mathbf{A}^k \to Ł = \mathbf{v}_1\mathbf{y}_1^\top$ entrywise.

The shape of Ł. By Proposition 9.10.7, $\mathbf{y}_1$ is a left eigenvector for 1, and so is $\mathbf{1}$ (Proposition 9.11.3). Both lie in $E_1(\mathbf{A}^\top) \subseteq \mathbb{C}^n$ (Proposition 9.10.5 (a)), which has dimension $g_{\mathbf{A}^\top}(1) = g_{\mathbf{A}}(1) \leq a_{\mathbf{A}}(1) = 1$ by Proposition 9.10.5 (b) and Theorem 9.3.1. Since $\mathbf{1} \neq \mathbf{0}$, $\mathbf{y}_1 = c\mathbf{1}$ for some $c \in \mathbb{C}$. Put $\mathbf{v} = c\mathbf{v}_1$. Then $Ł = \mathbf{v}\mathbf{1}^\top$, $\mathbf{A}\mathbf{v} = c\,\mathbf{A}\mathbf{v}_1 = \mathbf{v}$, and $\mathbf{1}^\top\mathbf{v} = \mathbf{y}_1^\top\mathbf{v}_1 = 1$ by Proposition 9.10.7 (b).

$\mathbf{v}$ *is a steady state.* Column 1 of $Ł = \mathbf{v}\mathbf{1}^\top$ is $\mathbf{v}$. It is the limit of column 1 of $\mathbf{A}^k$, and $\mathbf{A}^k$ is stochastic (Proposition 9.11.3), so every entry of $\mathbf{v}$ is a limit of real numbers ≥ 0, hence real and ≥ 0. With $\mathbf{1}^\top\mathbf{v} = 1$, $\mathbf{v}$ is a probability vector, and $\mathbf{A}\mathbf{v} = \mathbf{v}$.

Uniqueness. If $\mathbf{w}$ is a steady state, then $\mathbf{w} \in E_1(\mathbf{A})$, a subspace of $\mathbb{C}^n$ of dimension $g_{\mathbf{A}}(1) \leq 1$ that contains $\mathbf{v} \neq \mathbf{0}$. So $\mathbf{w} = s\mathbf{v}$ for some $s \in \mathbb{C}$, and $1 = \mathbf{1}^\top\mathbf{w} = s\,\mathbf{1}^\top\mathbf{v} = s$. Hence $\mathbf{w} = \mathbf{v}$. This proves (a) and (b).

(c) Entry r of $\mathbf{A}^k\mathbf{x}_0$ is $\sum_s (\mathbf{A}^k)_{rs}(\mathbf{x}_0)_s$, which by (b) converges to $\sum_s v_r(\mathbf{x}_0)_s = v_r\,(\mathbf{1}^\top\mathbf{x}_0) = v_r$.

■

In practice: find the eigenvalues, check the hypotheses, and compute $\mathbf{v}$ by solving $(\mathbf{A}-\mathbf{I})\mathbf{v}=\mathbf{0}$ and scaling. Then (c) gives the long run **without** computing $\mathbf{P}^{-1}$ at all.

Example. Bicycles at Three Stations

A shared bicycle is parked each night at station 1, 2, or 3. The transition matrix is

$$\mathbf{A}=\frac{1}{4}\begin{pmatrix}2&1&0\\1&2&1\\1&1&3\end{pmatrix};$$

for instance a bicycle at station 3 stays there with probability $3/4$ and never goes to station 1 in one night. Find $\mathbf{A}^k$, its limit, and the long-run distribution of the bicycle.

Solution. Eigenvectors. We check three candidates:

$$\mathbf{A}\begin{pmatrix}1\\2\\3\end{pmatrix}=\frac{1}{4}\begin{pmatrix}4\\8\\12\end{pmatrix},\qquad \mathbf{A}\begin{pmatrix}-1\\0\\1\end{pmatrix}=\frac{1}{4}\begin{pmatrix}-2\\0\\2\end{pmatrix},\qquad \mathbf{A}\begin{pmatrix}-1\\1\\0\end{pmatrix}=\frac{1}{4}\begin{pmatrix}-1\\1\\0\end{pmatrix}.$$

So $1,\frac{1}{2},\frac{1}{4}$ are eigenvalues, with eigenvectors $\mathbf{v}_1=(1,2,3)$, $\mathbf{v}_2=(-1,0,1)$, $\mathbf{v}_3=(-1,1,0)$. A 3×3 matrix has at most three eigenvalues (Corollary 9.2.3), so these are all; they are distinct, so $\mathbf{A}$ is diagonalizable (Corollary 9.4.2) and $a_{\mathbf{A}}(1)=1$. As a check, $\operatorname{tr}\mathbf{A}=\frac{7}{4}=1+\frac{1}{2}+\frac{1}{4}$. The hypotheses of Theorem 9.11.4 hold.

The expansion. With $\mathbf{P}=\begin{pmatrix}1&-1&-1\\2&0&1\\3&1&0\end{pmatrix}$, whose determinant is -6,

$$\mathbf{P}^{-1}=\begin{pmatrix}1/6&1/6&1/6\\-1/2&-1/2&1/2\\-1/3&2/3&-1/3\end{pmatrix},$$

as multiplying out $\mathbf{P}^{-1}\mathbf{P}=\mathbf{I}$ confirms. The first row is $\frac{1}{6}\mathbf{1}^{\top}$, a multiple of $\mathbf{1}^{\top}$ as the proof of the theorem predicts. By Proposition 9.10.7,

$$\mathbf{A}^k=\frac{1}{6}\begin{pmatrix}1&1&1\\2&2&2\\3&3&3\end{pmatrix}+\frac{1}{2^k}\cdot\frac{1}{2}\begin{pmatrix}1&1&-1\\0&0&0\\-1&-1&1\end{pmatrix}+\frac{1}{4^k}\cdot\frac{1}{3}\begin{pmatrix}1&-2&1\\-1&2&-1\\0&0&0\end{pmatrix}.$$

For example, the $(1,1)$-entry of $\mathbf{A}^k$ is $\frac{1}{6}+\frac{1}{2}\cdot2^{-k}+\frac{1}{3}\cdot4^{-k}$, which is 1 for $k=0$ and $\frac{1}{6}+\frac{1}{4}+\frac{1}{12}=\frac{1}{2}$ for $k=1$, matching $a_{11}=2/4$.

The limit. As $k\to\infty$, $2^{-k}\to0$ and $4^{-k}\to0$, so

$$\mathbf{A}^k\to\frac{1}{6}\begin{pmatrix}1&1&1\\2&2&2\\3&3&3\end{pmatrix}=\mathbf{v}\mathbf{1}^{\top},\qquad \mathbf{v}=\left(\frac{1}{6},\frac{1}{3},\frac{1}{2}\right).$$

Wherever the bicycle starts, in the long run it is at station 1, 2, 3 with probabilities $\frac{1}{6}, \frac{1}{3}, \frac{1}{2}$. Directly, $\mathbf{v} = \frac{1}{6}\mathbf{v}_1$ and $\mathbf{A}\mathbf{v} = \mathbf{v}$. The slowest-fading term is the one with $\frac{1}{2}$: each entry's gap to its limit roughly halves every night.

Warning. **A stochastic matrix need not have convergent powers.** The swap $\mathbf{S} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ is stochastic, and $\mathbf{S}^k$ is $\mathbf{I}$ for even k and $\mathbf{S}$ for odd k, so $\mathbf{S}^k$ does not converge. Starting from $\mathbf{x}_0 = \mathbf{e}_1$, the chain alternates $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_1, \dots$ forever, although $\mathbf{S}$ has the steady state $(\frac{1}{2}, \frac{1}{2})$. The theorem does not apply: $\mathbf{S}$ has the eigenvalue -1, with $|-1| = 1$. At the other extreme, $\mathbf{I}_2$ has convergent powers, but $a(1) = 2$, every probability vector is a steady state, and the limit of $\mathbf{x}_k$ depends on $\mathbf{x}_0$.

Chapter 19 proves the Perron–Frobenius theorem, which gives the conclusions of Theorem 9.11.4 for every stochastic matrix with some power having all entries positive, without assuming diagonalizability and without computing eigenvalues. The exercises prove the first ingredient: no eigenvalue of a stochastic matrix has $|\lambda| > 1$.

Quick check. Is $\mathbf{A} = \begin{pmatrix} 1/2 & 1/3 \\ 1/2 & 2/3 \end{pmatrix}$ stochastic? If so, find its steady state.

Solution. Yes: the entries are nonnegative and the columns sum to 1. From $(\mathbf{A} - \mathbf{I})\mathbf{v} = \mathbf{0}$, the first row gives $-\frac{1}{2}v_1 + \frac{1}{3}v_2 = 0$, so $v_2 = \frac{3}{2}v_1$. Scaling to sum 1: $v_1 + \frac{3}{2}v_1 = 1$ gives $\mathbf{v} = (\frac{2}{5}, \frac{3}{5})$. (The eigenvalues are 1 and $\operatorname{tr}\mathbf{A} - 1 = \frac{1}{6}$; they are distinct, so $\mathbf{A}$ is diagonalizable with $a_{\mathbf{A}}(1) = 1$, and $\mathbf{v}$ is the only steady state and the long-run distribution, by Theorem 9.11.4.)

9.11.5 A population model

Not every matrix of this kind is stochastic. In a population divided into age classes, individuals are born, age, and die, and the total need not be preserved. A **Leslie matrix** records, in its first row, the average number of offspring per individual of each age class, and just below the diagonal, the fraction surviving into the next class. The population vector evolves by $\mathbf{x}_{k+1} = Ł\mathbf{x}_k$, and Theorem 9.11.1 describes the long run.

Example. Three Age Classes

A species is counted once a year in three age classes: young, juvenile, and adult. Half of the young survive to become juveniles, and half of the juveniles survive to become adults; no adult survives another year. Each juvenile has on average 7/2 offspring and each adult 3. The Leslie matrix is

$$Ł = \begin{pmatrix} 0 & 7/2 & 3 \\ 1/2 & 0 & 0 \\ 0 & 1/2 & 0 \end{pmatrix}.$$

Starting from $\mathbf{x}_0 = (20, 10, 5)$ (in hundreds of animals), find $\mathbf{x}_k$ and describe the long-run growth and age distribution.

Solution. Characteristic polynomial. Expanding $\det(x\mathbf{I} - Ł)$ along the first column,

$$p_Ł = x\det\begin{pmatrix} x & 0 \\ -1/2 & x \end{pmatrix} + \frac{1}{2}\det\begin{pmatrix} -7/2 & -3 \\ -1/2 & x \end{pmatrix}$$
$$= x^3 + \frac{1}{2}\Big(-\frac{7}{2}x - \frac{3}{2}\Big) = x^3 - \frac{7}{4}x - \frac{3}{4}.$$

Trying simple values, $p_Ł(-1) = -1 + \frac{7}{4} - \frac{3}{4} = 0$, and dividing, $p_Ł = (x+1)(x^2 - x - \frac{3}{4}) = (x+1)(x - \frac{3}{2})(x + \frac{1}{2})$. The eigenvalues $\frac{3}{2}, -1, -\frac{1}{2}$ are distinct, so Ł is diagonalizable.

Eigenvectors. $Ł(w_1, w_2, w_3) = \lambda(w_1, w_2, w_3)$ forces $\frac{1}{2}w_1 = \lambda w_2$ and $\frac{1}{2}w_2 = \lambda w_3$, so taking $w_3 = 1$ gives $w_2 = 2\lambda$, $w_1 = 4\lambda^2$. Hence, after scaling,

$$\lambda = \tfrac{3}{2}:\ (9,3,1), \qquad \lambda = -1:\ (4,-2,1), \qquad \lambda = -\tfrac{1}{2}:\ (1,-1,1).$$

For instance $Ł(9,3,1) = (\frac{21}{2} + 3, \frac{9}{2}, \frac{3}{2}) = \frac{3}{2}(9,3,1)$.

The initial vector. Solving $c_1(9,3,1) + c_2(4,-2,1) + c_3(1,-1,1) = (20,10,5)$ gives $c_1 = 3$, $c_2 = -3$, $c_3 = 5$: indeed $(27 - 12 + 5,\ 9 + 6 - 5,\ 3 - 3 + 5) = (20,10,5)$. By Theorem 9.11.1,

$$\mathbf{x}_k = 3\Big(\frac{3}{2}\Big)^k \begin{pmatrix} 9 \\ 3 \\ 1 \end{pmatrix} - 3(-1)^k \begin{pmatrix} 4 \\ -2 \\ 1 \end{pmatrix} + 5\Big(-\frac{1}{2}\Big)^k \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix}.$$

Check. For $k = 1$: $(\frac{81}{2} + 12 - \frac{5}{2},\ \frac{27}{2} - 6 + \frac{5}{2},\ \frac{9}{2} + 3 - \frac{5}{2}) = (50, 10, 5)$, and directly $Ł\mathbf{x}_0 = (35 + 15, 10, 5)$.

The long run. Divide by $(3/2)^k$:

$$\Big(\frac{2}{3}\Big)^k \mathbf{x}_k = 3\begin{pmatrix} 9 \\ 3 \\ 1 \end{pmatrix} - 3\Big(-\frac{2}{3}\Big)^k \begin{pmatrix} 4 \\ -2 \\ 1 \end{pmatrix} + 5\Big(-\frac{1}{3}\Big)^k \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix} \longrightarrow \begin{pmatrix} 27 \\ 9 \\ 3 \end{pmatrix},$$

since $|-\frac{2}{3}| < 1$ and $|-\frac{1}{3}| < 1$. So the population eventually grows by the factor $\frac{3}{2}$ each year, and the age classes approach the proportions $9:3:1$, the entries of the eigenvector for the largest eigenvalue. The eigenvalue -1 makes the population oscillate from year to year, but its effect is bounded while the dominant term grows, so its share fades.

9.11.6 Exercises

9.11.6.1 A. Check your understanding

Exercise. A1

(a) Define a stochastic matrix, and say which convention (rows or columns) this book uses.

(b) True or false: for every stochastic $\mathbf{A}$, the powers $\mathbf{A}^k$ converge entrywise. Justify your answer.

(c) True or false: $\mathbf{1}$ is a right eigenvector of every stochastic matrix. Justify your answer.

(d) True or false: every stochastic matrix is invertible. Justify your answer.

(e) Write down the general solution of $a_{k+2} = 5a_{k+1} - 6a_k$ over $\mathbb{R}$.

(f) In Theorem 9.11.1, what are the numbers $c_1, \dots, c_n$?

Solution.

(a) $\mathbf{A} \in M_n(\mathbb{R})$ is stochastic if all entries are nonnegative and every **column** sums to 1 (Definition 9.11.1). This book uses columns, with $\mathbf{x}_{k+1} = \mathbf{A}\mathbf{x}_k$.

(b) False. The swap $\mathbf{S} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ has $\mathbf{S}^k$ alternating between $\mathbf{I}$ and $\mathbf{S}$.

(c) False. $\mathbf{A} = \begin{pmatrix} 1 & 1/2 \\ 0 & 1/2 \end{pmatrix}$ is stochastic, and $\mathbf{A}\mathbf{1} = (3/2, 1/2)$ is not a multiple of $\mathbf{1}$. It is a **left** eigenvector (Proposition 9.11.3).

(d) False. $\begin{pmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{pmatrix}$ is stochastic with determinant 0.

(e) $p = x^2 - 5x + 6 = (x-2)(x-3)$, so by Theorem 9.11.2 the solutions are $a_k = b_1 2^k + b_2 3^k$ with $b_1, b_2 \in \mathbb{R}$.

(f) The coordinates of $\mathbf{x}_0$ in the basis of eigenvectors $\mathbf{v}_1, \dots, \mathbf{v}_n$: $(c_1, \dots, c_n) = \mathbf{P}^{-1}\mathbf{x}_0$.

9.11.6.2 B. Practice

Exercise. B1: Two recurrences

(a) Solve $a_{k+2} = a_{k+1} + 2a_k$ with $a_0 = 0$, $a_1 = 3$.

(b) Solve $a_{k+2} = 6a_{k+1} - 9a_k$ with $a_0 = 1$, $a_1 = 6$, by the method of Example (A Repeated Root).

Solution.

(a) $p = x^2 - x - 2 = (x-2)(x+1)$ has distinct roots, so by Theorem 9.11.2, $a_k = b_1 2^k + b_2(-1)^k$. From $a_0 = b_1 + b_2 = 0$ and $a_1 = 2b_1 - b_2 = 3$, $3b_1 = 3$, so $b_1 = 1$, $b_2 = -1$. Hence $a_k = 2^k - (-1)^k$. Check: $a_2 = 3+0 = 3 = 4-1$, $a_3 = 3+6 = 9 = 8+1$.

(b) The window matrix is $\mathbf{C} = \begin{pmatrix} 0 & 1 \\ -9 & 6 \end{pmatrix}$, with $p = (x-3)^2$. Put $\mathbf{N} = \mathbf{C} - 3\mathbf{I} = \begin{pmatrix} -3 & 1 \\ -9 & 3 \end{pmatrix}$; then $\mathbf{N}^2 = \begin{pmatrix} 9-9 & -3+3 \\ 27-27 & -9+9 \end{pmatrix} = \mathbf{O}$. As in the example, by induction $\mathbf{C}^k = 3^k\mathbf{I} + k3^{k-1}\mathbf{N}$: the step is $(3^k\mathbf{I} + k3^{k-1}\mathbf{N})(3\mathbf{I} + \mathbf{N}) = 3^{k+1}\mathbf{I} + 3^k\mathbf{N} + k3^k\mathbf{N}$. Row 1 of $\mathbf{C}^k$ is $(3^k - 3k3^{k-1},\ k3^{k-1}) = ((1-k)3^k,\ k3^{k-1})$, so

$$a_k = (1-k)3^k \cdot 1 + k3^{k-1} \cdot 6 = (1 - k + 2k)3^k = (1+k)3^k.$$

Check: $a_2 = 36 - 9 = 27 = 3 \cdot 9$, $a_3 = 6 \cdot 27 - 9 \cdot 6 = 108 = 4 \cdot 27$.

Exercise. B2: A three-state chain

Let

$$\mathbf{A} = \begin{pmatrix} 1/2 & 1/2 & 1/4 \\ 1/4 & 0 & 1/4 \\ 1/4 & 1/2 & 1/2 \end{pmatrix}.$$

(a) Verify that $\mathbf{A}$ is stochastic, and find its steady state.

(b) Verify that $(-1, 0, 1)$ and $(1, -2, 1)$ are eigenvectors of $\mathbf{A}$, and find their eigenvalues.

(c) Hence find the entrywise limit of $\mathbf{A}^k$, and the limit of the chain started at $\mathbf{x}_0 = \mathbf{e}_2$.

Solution.

(a) All entries are nonnegative; the columns sum to $\frac{1}{2} + \frac{1}{4} + \frac{1}{4} = 1$, $\frac{1}{2} + 0 + \frac{1}{2} = 1$, and $\frac{1}{4} + \frac{1}{4} + \frac{1}{2} = 1$. For the steady state, solve $(\mathbf{A} - \mathbf{I})\mathbf{v} = \mathbf{0}$:

$$\mathbf{A} - \mathbf{I} = \begin{pmatrix} -1/2 & 1/2 & 1/4 \\ 1/4 & -1 & 1/4 \\ 1/4 & 1/2 & -1/2 \end{pmatrix}.$$

Multiplying the rows by 4: $-2v_1 + 2v_2 + v_3 = 0$, $v_1 - 4v_2 + v_3 = 0$, $v_1 + 2v_2 - 2v_3 = 0$. Subtracting the second from the third gives $6v_2 - 3v_3 = 0$, so $v_3 = 2v_2$; then the second gives $v_1 = 4v_2 - 2v_2 = 2v_2$, and the first holds: $-4v_2 + 2v_2 + 2v_2 = 0$. So $E_1(\mathbf{A}) = \operatorname{span}((2, 1, 2))$, and scaling to sum 1, the steady state is $\mathbf{v} = (\frac{2}{5}, \frac{1}{5}, \frac{2}{5})$.

(b) $\mathbf{A}(-1, 0, 1) = (-\frac{1}{2} + \frac{1}{4},\ -\frac{1}{4} + \frac{1}{4},\ -\frac{1}{4} + \frac{1}{2}) = (-\frac{1}{4}, 0, \frac{1}{4}) = \frac{1}{4}(-1, 0, 1)$, eigenvalue $\frac{1}{4}$. $\mathbf{A}(1, -2, 1) = (\frac{1}{2} - 1 + \frac{1}{4},\ \frac{1}{4} + \frac{1}{4},\ \frac{1}{4} - 1 + \frac{1}{2}) = (-\frac{1}{4}, \frac{1}{2}, -\frac{1}{4}) = -\frac{1}{4}(1, -2, 1)$, eigenvalue $-\frac{1}{4}$.

(c) The eigenvalues $1, \frac{1}{4}, -\frac{1}{4}$ are distinct, so they are all the eigenvalues of $\mathbf{A}$, $\mathbf{A}$ is diagonalizable (Corollary 9.4.2), $a_{\mathbf{A}}(1) = 1$, and the other two have absolute value $\frac{1}{4} < 1$. By Theorem 9.11.4,

$$\mathbf{A}^k \to \mathbf{v}\mathbf{1}^\top = \frac{1}{5}\begin{pmatrix} 2 & 2 & 2 \\ 1 & 1 & 1 \\ 2 & 2 & 2 \end{pmatrix},$$

and every chain, in particular the one started at $\mathbf{e}_2$, converges to $\mathbf{v} = (\frac{2}{5}, \frac{1}{5}, \frac{2}{5})$.

Exercise. B3: The limit of a matrix power

Let $\mathbf{A} = \begin{pmatrix} 0 & 1 \\ -1/2 & 3/2 \end{pmatrix} \in M_2(\mathbb{R})$.

(a) Diagonalize $\mathbf{A}$, and find a formula for $\mathbf{A}^k$.

(b) Find the entrywise limit of $\mathbf{A}^k$.

(c) Hence find $\lim_{k\to\infty} a_k$ for the sequence with $a_{k+2} = \frac{3}{2}a_{k+1} - \frac{1}{2}a_k$, in terms of a_0 and a_1.

Solution.

(a) $\operatorname{tr}\mathbf{A} = \frac{3}{2}$ and $\det\mathbf{A} = \frac{1}{2}$, so $p_{\mathbf{A}} = x^2 - \frac{3}{2}x + \frac{1}{2} = (x-1)(x-\frac{1}{2})$. $\mathbf{A}-\mathbf{I} = \begin{pmatrix} -1 & 1 \\ -1/2 & 1/2 \end{pmatrix}$ gives the eigenvector $(1,1)$; $\mathbf{A} - \frac{1}{2}\mathbf{I} = \begin{pmatrix} -1/2 & 1 \\ -1/2 & 1 \end{pmatrix}$ gives $(2,1)$. So $\mathbf{P} = \begin{pmatrix} 1 & 2 \\ 1 & 1 \end{pmatrix}$, $\mathbf{D} = \operatorname{diag}(1, \frac{1}{2})$, $\mathbf{P}^{-1} = \begin{pmatrix} -1 & 2 \\ 1 & -1 \end{pmatrix}$ ($\det\mathbf{P} = -1$, Theorem 0.9.8). By Theorem 9.4.3, with $t = 2^{-k}$,

$$\mathbf{A}^k = \begin{pmatrix} 1 & 2t \\ 1 & t \end{pmatrix}\begin{pmatrix} -1 & 2 \\ 1 & -1 \end{pmatrix} = \begin{pmatrix} -1+2t & 2-2t \\ -1+t & 2-t \end{pmatrix}.$$

For $k=1$, $t=\frac{1}{2}$ gives $\begin{pmatrix} 0 & 1 \\ -1/2 & 3/2 \end{pmatrix} = \mathbf{A}$.

(b) Since $t = 2^{-k} \to 0$, $\mathbf{A}^k \to \begin{pmatrix} -1 & 2 \\ -1 & 2 \end{pmatrix}$ entrywise.

(c) $\mathbf{A}$ is the window matrix of this recurrence, so $(a_k, a_{k+1}) = \mathbf{A}^k(a_0, a_1)$ and $a_k = (-1+2t)a_0 + (2-2t)a_1$. Hence $a_k \to 2a_1 - a_0$. (For instance $a_0 = 0$, $a_1 = 1$ gives $0, 1, \frac{3}{2}, \frac{7}{4}, \cdots \to 2$.)

9.11.6.3 C. Going deeper

Exercise. C1: Eigenvalues of a stochastic matrix

Let $\mathbf{A} \in M_n(\mathbb{R})$ be stochastic, regarded as a complex matrix, and let $\lambda \in \mathbb{C}$ be an eigenvalue of $\mathbf{A}$.

(a) Prove that $|\lambda| \le 1$.

(b) Refine the argument: prove that $|\lambda - a_{ii}| \le 1 - a_{ii}$ for some index i.

(c) Deduce that if every diagonal entry of $\mathbf{A}$ is positive, then -1 is not an eigenvalue of $\mathbf{A}$.

(d) Show by an example that (c) fails if a diagonal entry is 0.

Hint: take a left eigenvector $\mathbf{y}$ *for* λ *and an index* i *with* $|y_i|$ *as large as possible.*

Solution.

(a) By Proposition 9.10.5 (over $\mathbb{C}$), there is $\mathbf{y} \in \mathbb{C}^n$, $\mathbf{y} \neq \mathbf{0}$, with $\mathbf{A}^{\top}\mathbf{y} = \lambda\mathbf{y}$. Choose i with $|y_i| \ge |y_j|$ for all j; then $|y_i| > 0$, since $\mathbf{y} \neq \mathbf{0}$. Entry i of $\mathbf{A}^{\top}\mathbf{y} = \lambda\mathbf{y}$ reads $\sum_j a_{ji}y_j = \lambda y_i$, because row i of $\mathbf{A}^{\top}$ is column i of $\mathbf{A}$. By Theorem 0.6.2 (d) and Theorem 0.6.4, applied to the n-term sum by induction on n, and since $a_{ji} \ge 0$ and column i sums to 1,

$$|\lambda|\,|y_i| = \Big|\sum_j a_{ji}y_j\Big| \le \sum_j a_{ji}|y_j| \le \sum_j a_{ji}|y_i| = |y_i|.$$

Dividing by $|y_i| > 0$ gives $|\lambda| \le 1$.

(b) With the same i, move the diagonal term across: $(\lambda - a_{ii})y_i = \sum_{j\neq i} a_{ji}y_j$. Hence, as

in (a),

$$|\lambda - a_{ii}|\,|y_i| \le \sum_{j\neq i} a_{ji}|y_j| \le \Big(\sum_{j\neq i} a_{ji}\Big)|y_i| = (1-a_{ii})|y_i|,$$

and dividing by $|y_i|$ gives $|\lambda - a_{ii}| \le 1 - a_{ii}$.

(c) Suppose $\lambda = -1$. By (b), some i has $|-1-a_{ii}| \le 1-a_{ii}$. Since $a_{ii} \ge 0$, $|-1-a_{ii}| = 1+a_{ii}$, so $1 + a_{ii} \le 1 - a_{ii}$ and $a_{ii} \le 0$. This contradicts $a_{ii} > 0$.

(d) The swap $\mathbf{S} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ is stochastic with zero diagonal, and $\mathbf{S}(1,-1) = (-1,1)$, so -1 is an eigenvalue.

Exercise. C2: Two-state chains

Let $0 \le p, q \le 1$ and $\mathbf{A} = \begin{pmatrix} 1-p & q \\ p & 1-q \end{pmatrix}$.

(a) Show that $\mathbf{A}$ is stochastic, with eigenvalues 1 and $1 - p - q$.

(b) Suppose $0 < p + q < 2$. Prove that $\mathbf{A}^k$ converges entrywise, and find the limit and the steady state.

(c) In particular, if $\mathbf{A}$ has no zero entry, then $\mathbf{A}^k$ converges. What happens when $p+q = 0$ and when $p + q = 2$?

Solution.

(a) The entries lie in $[0,1]$, and the columns sum to $(1-p)+p = 1$ and $q+(1-q) = 1$. Since $\operatorname{tr}\mathbf{A} = 2-p-q$ and $\det\mathbf{A} = (1-p)(1-q) - pq = 1-p-q$, $p_{\mathbf{A}} = x^2 - (2-p-q)x + (1-p-q) = (x-1)(x-(1-p-q))$, as multiplying out confirms.

(b) Put $r = 1-p-q$. Since $0 < p+q < 2$, $-1 < r < 1$, so $r \neq 1$ and $|r| < 1$. The eigenvalues $1 \neq r$ are distinct, so $\mathbf{A}$ is diagonalizable (Corollary 9.4.2). From $\mathbf{A} - \mathbf{I} = \begin{pmatrix} -p & q \\ p & -q \end{pmatrix}$, (q,p) is an eigenvector for 1, nonzero because $p+q > 0$; from $\mathbf{A} - r\mathbf{I} = \begin{pmatrix} q & q \\ p & p \end{pmatrix}$, $(1,-1)$ is one for r. So

$$\mathbf{P} = \begin{pmatrix} q & 1 \\ p & -1 \end{pmatrix}, \qquad \mathbf{P}^{-1} = \frac{1}{p+q}\begin{pmatrix} 1 & 1 \\ p & -q \end{pmatrix},$$

with $\det\mathbf{P} = -(p+q) \neq 0$. By Theorem 9.4.3,

$$\mathbf{A}^k = \frac{1}{p+q}\begin{pmatrix} q & r^k \\ p & -r^k \end{pmatrix}\begin{pmatrix} 1 & 1 \\ p & -q \end{pmatrix} = \frac{1}{p+q}\begin{pmatrix} q+pr^k & q-qr^k \\ p-pr^k & p+qr^k \end{pmatrix}.$$

Since $|r| < 1$, $r^k \to 0$, and

$$\mathbf{A}^k \to \frac{1}{p+q}\begin{pmatrix} q & q \\ p & p \end{pmatrix}, \qquad \mathbf{v} = \Big(\frac{q}{p+q}, \frac{p}{p+q}\Big).$$

The limit is $\mathbf{v}\mathbf{1}^{\top}$, and $\mathbf{v}$ is the steady state, in agreement with Theorem 9.11.4.

(c) If no entry is zero, then $1-p > 0$ and $p > 0$, so $0 < p < 1$, and likewise $0 < q < 1$; hence $0 < p+q < 2$ and (b) applies. If $p+q = 0$, then $p = q = 0$ and $\mathbf{A} = \mathbf{I}$: the

powers converge (to $\mathbf{I}$), but every probability vector is a steady state. If $p + q = 2$, then $p = q = 1$ and $\mathbf{A}$ is the swap, whose powers do not converge.

Chapter 10

Canonical Forms

Chapter 9 left one question open. Two matrices represent the same operator in different bases exactly when they are similar, and we can now decide similarity in easy cases: diagonalizable matrices are similar precisely when they have the same eigenvalues with the same multiplicities. But the matrices $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ and I_2 share their characteristic and minimal polynomials, their rank, their trace, and their determinant, and they are still not similar. Something is missing.

This chapter supplies it. When the characteristic polynomial splits, every operator has a **Jordan form**: a direct sum of blocks with one eigenvalue on the diagonal and ones just above it, unique up to the order of the blocks. Two matrices are similar exactly when their Jordan forms agree, so the similarity problem is solved. Over a field where the polynomial does not split, the **rational canonical form** does the same job, and the Smith normal form of $x\mathbf{I} - \mathbf{A}$ computes it. The chapter closes with what these forms are for: functions of matrices, the exponential $e^{t\mathbf{A}}$ and linear differential equations, and a complete answer to when the powers $\mathbf{A}^m$ converge.

What you need. From Chapter 0, well-ordering (Theorem 0.2.11) and field arithmetic (Theorem 0.7.1) for the induction that produces each form, degree (Theorem 0.8.3), the complex numbers with conjugation (Definition 0.6.1, Theorem 0.6.2) for the real forms, that a composition of bijections is a bijection (Theorem 0.4.3), and the matrix product, the transpose, and similarity of matrices (Definition 0.9.9, Definition 0.9.10, Example (Similarity)). From Chapter 1, direct sums, basis extension, and complexification (Theorem 1.7.5, Theorem 1.6.5, Definition 1.9.2). From Chapter 2, the workhorse: Rank–Nullity (Theorem 2.3.1), images (Theorem 2.2.4), and polynomials of an operator (Definition 2.4.2). From Chapter 3, row reduction and what it computes — the reduced row echelon form with its existence and uniqueness (Definition 3.2.4, Theorem 3.2.3, Theorem 3.3.4, Theorem 3.4.1, Corollary 3.5.6), Rank–Nullity for matrices (Theorem 3.5.7), and that a one-sided inverse of a square matrix is two-sided (Theorem 3.4.5). From Chapter 4, the matrix of a map with its rank and the isomorphism $\mathcal{L}(V, W) \cong M_{m\times n}(F)$ (Definition 4.1.1, Theorem 4.1.4, Theorem 4.1.5), projections along a complement (Theorem 4.5.1), and, above all, similarity with the invariants it preserves (Theorem 4.2.4, Proposition 4.2.6, Theorem 4.2.3, Theorem 4.2.5), which is what a canonical form is a canonical form for. From Chapter 6, the arithmetic of $F[x]$: division, gcds, Bézout, and unique factorization (Theorem 6.1.1, Corollary 6.2.4, Theorem 6.3.5), and interpolation (Theorem 6.6.4). From Chapter 7, determinants of block triangular matrices, Cauchy–Binet, and the companion matrix of a monic polynomial (Theorem 7.8.4, Theorem 7.7.2, Exercise (C1: Companion matrices)). From Chapter 8, block arithmetic, which is how every form here is manipulated once it is written down (Theorem 8.1.4, Theorem 8.1.1, Theorem 8.2.1). From Chapter 9, the minimal polynomial, primary decomposition, and triangularization (Definition 9.5.1, Theorem 9.8.1, Theorem 9.7.2).

10.0.1 Roadmap

- **Generalized eigenspaces.** Enlarging an eigenspace until it is big enough, and the decomposition it produces.
- **Nilpotent operators.** Chains of vectors, the block structure they give, and how ranks of powers determine it.
- **The Jordan canonical form.** Existence, uniqueness, and the complete answer to the similarity question over a splitting field.
- **The real Jordan form and the Weyr form.** Avoiding complex numbers for a real matrix, and a rearrangement of the Jordan form with its own uses.
- **Cyclic subspaces and companion matrices.** One vector and its images can generate everything.
- **The rational canonical form.** A canonical form over every field, with invariant factors as the complete invariants.
- **Polynomial matrices and the Smith normal form.** Row and column operations on $x\mathbf{I} - \mathbf{A}$ compute the invariant factors.
- **Semisimple operators and the Jordan–Chevalley decomposition.** Every operator splits uniquely into a diagonalizable part plus a commuting nilpotent part.
- **Functions of matrices, the exponential, and linear differential equations.** $f(\mathbf{A})$ from the Jordan form, and $x' = \mathbf{A}x$ solved by $e^{t\mathbf{A}}$.
- **Convergent and power-bounded matrices.** Exactly when $\mathbf{A}^m$ converges, and to what.

10.0.2 Named moves

- **Enlarge the kernel.** When $\ker(T - \lambda I)$ is too small, pass to $\ker(T - \lambda I)^k$ until the chain stops growing.
- **Reduce to the nilpotent case.** On each generalized eigenspace, $T - \lambda I$ is nilpotent, so the whole structure question is about nilpotent operators.
- **Read the chain off the columns.** The columns of a Jordan block say $Nv_1 = 0$ and $Nv_j = v_{j-1}$: a chain.
- **Count ranks of powers.** The block sizes are second differences of $\operatorname{rank} N^j$, so they can be found without finding a single chain.
- **A cyclic vector gives a companion matrix.** If $v, Tv, \dots$ span, the matrix of T is built from one polynomial.

10.1 Generalized Eigenspaces

Chapter 9 left one gap wide open. When an operator has too few eigenvectors, its eigenspaces are too small to fill the space, and diagonalization fails. The remedy of this chapter is to enlarge each eigenspace until the enlarged pieces do fill the space. This section builds the enlargement, shows that the enlarging process stops after finitely many steps, and proves that the pieces split V whenever the characteristic polynomial splits.

10.1.1 Enlarging an eigenspace

Take the standing counterexample of Chapter 9,

$$\mathbf{J} = \begin{pmatrix} 2 & 1 \\ 0 & 2 \end{pmatrix} \in M_2(F),$$

and let $\mathbf{N} = \mathbf{J} - 2\mathbf{I}$. The only eigenvalue is 2, and $E_2(\mathbf{J}) = \ker \mathbf{N} = \operatorname{span}(\mathbf{e}_1)$ is a line: one eigenvector direction where we need two, so $\mathbf{J}$ is not diagonalizable over any field (Theorem 9.4.1 (e)).

Now look at what $\mathbf{N}$ does to the missing direction $\mathbf{e}_2$. One application does not kill $\mathbf{e}_2$, but two do:

$$\mathbf{N}\mathbf{e}_2 = \mathbf{e}_1, \qquad \mathbf{N}^2\mathbf{e}_2 = \mathbf{N}\mathbf{e}_1 = \mathbf{0}.$$

So $\mathbf{e}_2$ is killed by $\mathbf{N}^2$, and $\ker \mathbf{N}^2 = F^2$ is everything. The eigenspace was too small, but the kernel of the **square** is as big as we could wish.

This is the whole idea. Instead of asking for vectors killed by $T - \lambda \operatorname{id}_V$, ask for vectors killed by **some power** of $T - \lambda \operatorname{id}_V$. The eigenvectors are still there (they are killed at the first step), and there may be more.

A generalized eigenvector is a vector that some power of $T - \lambda \operatorname{id}_V$ kills, not necessarily the first power.

10.1.2 Generalized eigenvectors and generalized eigenspaces

Definition 10.1.1. Generalized Eigenvector

Let V be a vector space over F, let $T \in \mathcal{L}(V)$ and $\lambda \in F$. A **generalized eigenvector** of T for λ is a vector $v \in V$ with $v \neq 0$ such that

$$(T - \lambda \operatorname{id}_V)^k v = 0 \qquad \text{for some integer } k \geq 1.$$

In words: the vector must be **nonzero**, exactly as for an eigenvector; the scalar λ must lie **in** F; and the power k is not fixed in advance — it may depend on v. Taking $k = 1$ recovers the eigenvector condition, so **every eigenvector for λ is a generalized eigenvector for λ.**

Collecting all of them, together with 0, gives the space we are after.

Definition 10.1.2. Generalized Eigenspace

Let V be a vector space over F, let $T \in \mathcal{L}(V)$ and $\lambda \in F$. The **generalized eigenspace** of T for λ is

$$\begin{aligned} G_\lambda(T) &\coloneqq \{\, v \in V : (T - \lambda \operatorname{id}_V)^k v = 0 \text{ for some integer } k \geq 1 \,\} \\ &= \bigcup_{k \geq 1} \ker(T - \lambda \operatorname{id}_V)^k. \end{aligned}$$

For $\mathbf{A} \in M_n(F)$ we write $G_\lambda(\mathbf{A}) = G_\lambda(T_\mathbf{A})$.

So $G_\lambda(T)$ consists of 0 together with all generalized eigenvectors for λ. Writing $N = T - \lambda \operatorname{id}_V$ for the rest of this subsection, the definition says $G_\lambda(T) = \ker N \cup \ker N^2 \cup \ker N^3 \cup \dots$.

Well-definedness: it really is a subspace. A union of subspaces is almost never a subspace (Proposition 1.2.3), so this needs a check. What saves us is that the union is **increasing**: if $N^k v = 0$, then $N^{k+1} v = N(N^k v) = N0 = 0$, so

$$\ker N \subseteq \ker N^2 \subseteq \ker N^3 \subseteq \dots$$

Now use the subspace test (Theorem 1.2.1). We have $0 \in \ker N \subseteq G_\lambda(T)$, so the set is nonempty. If $u, v \in G_\lambda(T)$, say $N^j u = 0$ and $N^k v = 0$, put $m = \max(j,k)$; then $u, v \in \ker N^m$ by the inclusions above, and $\ker N^m$ is a subspace, so $u+v$ and cu lie in $\ker N^m \subseteq G_\lambda(T)$ for every $c \in F$. Hence $G_\lambda(T)$ is a subspace of V.

Examples.

- **The standing counterexample.** For $\mathbf{J} = \begin{pmatrix} 2 & 1 \\ 0 & 2 \end{pmatrix}$, we computed $\ker(\mathbf{J} - 2\mathbf{I}) = \operatorname{span}(\mathbf{e}_1)$ and $\ker(\mathbf{J} - 2\mathbf{I})^2 = F^2$. So $G_2(\mathbf{J}) = F^2$, while $E_2(\mathbf{J})$ is only a line. The generalized eigenspace is strictly bigger than the eigenspace, and it is all of F^2.
- **A diagonalizable operator gains nothing.** Let $\mathbf{A} = \operatorname{diag}(2,5)$ over $\mathbb{Q}$. Then $(\mathbf{A} - 2\mathbf{I})^k = \operatorname{diag}(0, 3^k)$, whose kernel is $\operatorname{span}(\mathbf{e}_1)$ for every $k \geq 1$. So $G_2(\mathbf{A}) = E_2(\mathbf{A}) = \operatorname{span}(\mathbf{e}_1)$, and likewise $G_5(\mathbf{A}) = E_5(\mathbf{A}) = \operatorname{span}(\mathbf{e}_2)$. Enlarging an eigenspace of a diagonalizable operator does nothing, a fact we prove in general below.
- **Differentiation.** Let $D \in \mathcal{L}(F[x]_{\leq n})$ be differentiation. Then $D^{n+1} = 0$ (Example (Differentiation Is Nilpotent)), so **every** polynomial is killed by a power of D, that is, of $D - 0\,\mathrm{id}$, and $G_0(D) = F[x]_{\leq n}$. By contrast $E_0(D) = \ker D$ is the line of constants when $F = \mathbb{R}$. Here the generalized eigenspace for the single eigenvalue 0 is the whole space, of dimension $n+1$.
- **A scalar not among the eigenvalues.** For the same $\mathbf{A} = \operatorname{diag}(2,5)$ and $\lambda = 7$, the matrix $(\mathbf{A} - 7\mathbf{I})^k = \operatorname{diag}((-5)^k, (-2)^k)$ is invertible, so its kernel is $\{\mathbf{0}\}$ for every k. Hence $G_7(\mathbf{A}) = \{\mathbf{0}\}$. The definition makes sense for every $\lambda \in F$, and gives the zero space unless λ is an eigenvalue.
- **A degenerate case.** On a space of dimension 1, every operator is λ id for one scalar λ, and $G_\lambda = V$, $G_\mu = \{0\}$ for $\mu \neq \lambda$. Small as it is, this is the model for a 1×1 Jordan block.
- **No eigenvalues at all.** The rotation R of $\mathbb{R}^2$ by a right angle has no real eigenvalue (Example (A Rotation Has No Invariant Lines)), and indeed $G_\lambda(R) = \{\mathbf{0}\}$ for every $\lambda \in \mathbb{R}$: if $(R - \lambda\mathbf{I})^k \mathbf{v} = \mathbf{0}$ with $\mathbf{v} \neq \mathbf{0}$, then $R - \lambda\mathbf{I}$ would not be injective, so λ would be an eigenvalue. Over $\mathbb{C}$, where $p_R = x^2 + 1 = (x - i)(x + i)$ splits, $G_i(R)$ and $G_{-i}(R)$ are the two eigenlines.

Nonexample by minimal change. Keep $\mathbf{A} = \operatorname{diag}(2,5)$ over $\mathbb{Q}$ and take $\mathbf{v} = (1,1)$. This vector **is** killed by a polynomial in $\mathbf{A}$ whose roots are eigenvalues:

$$(\mathbf{A} - 2\mathbf{I})(\mathbf{A} - 5\mathbf{I})\mathbf{v} = \mathbf{0},$$

as one checks entry by entry. But it is a generalized eigenvector for **neither** eigenvalue: $(\mathbf{A} - 2\mathbf{I})^k \mathbf{v} = (0, 3^k) \neq \mathbf{0}$ and $(\mathbf{A} - 5\mathbf{I})^k \mathbf{v} = ((-3)^k, 0) \neq \mathbf{0}$ for every $k \geq 1$. The failing clause is that the definition demands a power of a **single** factor $T - \lambda\,\mathrm{id}_V$, not a product of different factors. (Of course $\mathbf{v} = \mathbf{e}_1 + \mathbf{e}_2$ is a *sum* of two generalized eigenvectors; that is the decomposition theorem below, not the definition.)

Why this definition. Three clauses are doing work. The vector must be nonzero, or every G_λ would contain a "generalized eigenvector" and the word would carry no information. The exponent must be allowed to grow, because nothing tells us in advance how many steps a vector needs: for D on $\mathbb{R}[x]_{\leq n}$ the polynomial x^n needs $n+1$ steps. And the power is a power of one factor, which is what makes G_λ one piece attached to one eigenvalue. The name is the honest one: these are the vectors that behave like eigenvectors after enough applications of $T - \lambda\,\mathrm{id}_V$.

Warning. **$G_\lambda(T)$ is usually bigger than $E_\lambda(T)$, and "generalized eigenvector" includes the ordinary ones.** The inclusion $E_\lambda(T) \subseteq G_\lambda(T)$ always holds, and it is strict

exactly when T has too few eigenvectors for λ: for $\mathbf{J} = \begin{pmatrix} 2 & 1 \\ 0 & 2 \end{pmatrix}$, $\dim E_2 = 1$ and $\dim G_2 = 2$. So a generalized eigenvector need not satisfy $Tv = \lambda v$: the vector $\mathbf{e}_2$ satisfies $\mathbf{J}\mathbf{e}_2 = \mathbf{e}_1 + 2\mathbf{e}_2$. Conversely, do not read "generalized" as "not an eigenvector": every eigenvector is a generalized eigenvector.

10.1.3 The kernel chain stabilizes

The definition of $G_\lambda(T)$ is an infinite union, which is awkward: to decide membership we would have to try every k. In finite dimension this is an illusion. The chain $\ker N \subseteq \ker N^2 \subseteq \dots$ is a chain of subspaces of V, and their dimensions cannot keep growing. Once the chain stops growing, it stops forever.

Theorem 10.1.1. The Kernel Chain Stabilizes

Let V be a finite-dimensional vector space over F with $\dim V = n$, let $N \in \mathcal{L}(V)$, and write $K_k = \ker N^k$ for $k \geq 0$, so that $K_0 = \{0\}$. Then:

(a) $K_0 \subseteq K_1 \subseteq K_2 \subseteq \dots$, and every K_k is N-invariant;

(b) if $K_m = K_{m+1}$ for some $m \geq 0$, then $K_{m+j} = K_m$ for **every** $j \geq 0$;

(c) there is a **least** index $s \geq 0$ with $K_s = K_{s+1}$; it satisfies $s \leq n$, the inclusions

$$K_0 \subsetneq K_1 \subsetneq \cdots \subsetneq K_s$$

are **strict**, and $K_k = K_s$ for every $k \geq s$;

(d) in particular $\ker N^n = \ker N^{n+j}$ for every $j \geq 0$.

Idea. The only real point is (b): equality at one step propagates. Read it backwards. To put $v \in K_{m+j+1}$ into K_{m+j}, look at $w = N^j v$. It lies in K_{m+1}, which by hypothesis is K_m, and that is exactly the statement $N^{m+j}v = 0$. Then (c) and (d) are bookkeeping: while the inclusions are strict the dimensions go up by at least 1, so strictness cannot survive past step n.

Proof.

(a) If $N^k v = 0$, then $N^{k+1}v = N(N^k v) = 0$, so $K_k \subseteq K_{k+1}$. Each $K_k = \ker N^k$ is the kernel of a polynomial in N, hence N-invariant by Theorem 6.7.4 (b).

(b) Suppose $K_m = K_{m+1}$. We prove $K_{m+j} = K_m$ by induction on $j \geq 0$. For $j = 0$ there is nothing to prove. Assume $K_{m+j} = K_m$, and let $v \in K_{m+j+1}$. Then

$$N^{m+1}(N^j v) = N^{m+j+1}v = 0,$$

so $N^j v \in K_{m+1} = K_m$, which says $N^{m+j}v = N^m(N^j v) = 0$. Hence $v \in K_{m+j} = K_m$. Together with $K_m \subseteq K_{m+j+1}$ from (a), this gives $K_{m+j+1} = K_m$.

(c) First, some m with $0 \leq m \leq n$ satisfies $K_m = K_{m+1}$. Otherwise $K_m \subsetneq K_{m+1}$ for every $m = 0, 1, \dots, n$, so $\dim K_{m+1} \geq \dim K_m + 1$ for those m, and hence $\dim K_{n+1} \geq n + 1$, contradicting $\dim K_{n+1} \leq \dim V = n$ (Theorem 1.6.7). By the well-ordering principle (Theorem 0.2.11) there is a least such index; call it s, and $s \leq n$. For $k < s$ minimality

gives $K_k \neq K_{k+1}$, so the inclusion $K_k \subseteq K_{k+1}$ of (a) is strict. For $k \geq s$, (b) gives $K_k = K_s$.

(d) Since $s \leq n$, (c) gives $K_n = K_s = K_{n+j}$ for every $j \geq 0$.

■

Quick check. Can $\ker N \subsetneq \ker N^2 = \ker N^3$ happen? Can $\ker N = \ker N^2 \subsetneq \ker N^3$ happen?

Solution. The first can: for $\mathbf{N} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$ on F^2 we have $\ker \mathbf{N} = \operatorname{span}(\mathbf{e}_1)$ and $\ker \mathbf{N}^2 = F^2 = \ker \mathbf{N}^3$, since $\mathbf{N}^2 = \mathbf{O}$. Here $s = 2$.

The second cannot, by Theorem 10.1.1 (b) with $m = 1$: equality at one step forces equality at every later step. This is why the theorem says the strict inclusions come **first** and the equalities **after**; the chain never resumes growing.

The index s will come up often enough to name. We call the least $s \geq 0$ with $\ker N^s = \ker N^{s+1}$ the **stabilization index** of N. Part (d) is the useful form: the chain has certainly stopped by step $n = \dim V$, whatever N is. Applying this to $N = T - \lambda \operatorname{id}_V$ turns the infinite union in Definition 10.1.2 into a single kernel.

Corollary 10.1.2. The Generalized Eigenspace Is One Kernel

Let V be a finite-dimensional vector space over F with $\dim V = n$, let $T \in \mathcal{L}(V)$ and $\lambda \in F$, and put $N = T - \lambda \operatorname{id}_V$ with stabilization index s. Then:

(a) $G_\lambda(T) = \ker N^s = \ker N^n$;

(b) $G_\lambda(T)$ is T-invariant, and $E_\lambda(T) \subseteq G_\lambda(T)$;

(c) $G_\lambda(T) \neq \{0\}$ if and only if λ is an eigenvalue of T;

(d) if $v \in G_\lambda(T)$ is nonzero and $k \geq 1$ is **least** with $N^k v = 0$, then $N^{k-1} v$ is an eigenvector of T for λ.

Proof.

(a) Every $\ker N^k$ with $k \geq 1$ is contained in $\ker N^{\max(k,n)} = \ker N^n$ by Theorem 10.1.1 (a) and (d), so the union $G_\lambda(T)$ equals $\ker N^n$; and $\ker N^n = \ker N^s$ by (c) of the same theorem, since $s \leq n$.

(b) By (a), $G_\lambda(T) = \ker N^n$ is the kernel of a polynomial in T, so T maps it into itself (Theorem 6.7.4 (b)). And $E_\lambda(T) = \ker N \subseteq \ker N^n$.

(c) ($\Leftarrow$) If λ is an eigenvalue, then $E_\lambda(T) \neq \{0\}$ (Theorem 9.2.1), and $E_\lambda(T) \subseteq G_\lambda(T)$ by (b). ($\Rightarrow$) If $v \in G_\lambda(T)$ is nonzero, take $k \geq 1$ least with $N^k v = 0$, which exists by the definition of $G_\lambda(T)$ and Theorem 0.2.11. Then $w = N^{k-1} v \neq 0$ by minimality of k (for $k = 1$ this reads $w = v \neq 0$), while $Nw = N^k v = 0$. So $w \in \ker N$ is nonzero and λ is an eigenvalue.

(d) This is the computation just made: $w = N^{k-1} v$ is nonzero and lies in $\ker N = E_\lambda(T)$.

■

Part (d) is a move worth remembering: **to get a genuine eigenvector out of a generalized one, apply $T-\lambda\,\mathrm{id}_V$ as many times as possible without reaching 0.** It is the reason the enlargement never produces eigenvalues out of thin air, and it will start every Jordan chain in Section 2.

10.1.4 The generalized eigenspace decomposition

We can now state the payoff. Chapter 9's primary decomposition (Theorem 9.8.1) already splits V into invariant pieces, one for each irreducible factor of m_T. When p_T splits, all those factors are linear, the pieces are kernels of powers of $T-\lambda_i\,\mathrm{id}_V$ — and those are exactly the generalized eigenspaces. The extra information here is that the piece attached to λ_i has dimension $a_T(\lambda_i)$, so the sizes are read off the characteristic polynomial.

When the characteristic polynomial splits, the generalized eigenspaces fill the space, and the one for λ has dimension the algebraic multiplicity of λ.

Theorem 10.1.3. Generalized Eigenspace Decomposition

Let V be a finite-dimensional vector space over F with $\dim V = n \geq 1$, let $T \in \mathcal{L}(V)$, and suppose p_T **splits** over F, with distinct eigenvalues $\lambda_1, \ldots, \lambda_k$ and

$$p_T = (x-\lambda_1)^{a_1}\cdots(x-\lambda_k)^{a_k}, \qquad m_T = (x-\lambda_1)^{s_1}\cdots(x-\lambda_k)^{s_k},$$

where $a_i = a_T(\lambda_i)$ and $1 \leq s_i \leq a_i$. Write $G_i = G_{\lambda_i}(T)$ and $N_i = (T-\lambda_i\,\mathrm{id}_V)|_{G_i}$. Then:

(a) $G_i = \ker(T-\lambda_i\,\mathrm{id}_V)^{s_i}$, and

$$V = G_1 \oplus G_2 \oplus \cdots \oplus G_k;$$

(b) each G_i is T-invariant and nonzero, with $E_{\lambda_i}(T) \subseteq G_i$, and $E_{\lambda_i}(T) = G_i$ if and only if $s_i = 1$;

(c) $\dim G_i = a_i$ and $p_{T|_{G_i}} = (x-\lambda_i)^{a_i}$;

(d) $N_i^{a_i} = 0$; that is, $T|_{G_i} = \lambda_i\,\mathrm{id}_{G_i} + N_i$ with N_i nilpotent, of stabilization index s_i;

(e) if $\mathcal{B}$ is a basis of V obtained by writing bases $\mathcal{B}_1, \ldots, \mathcal{B}_k$ of $G_1, \ldots, G_k$ one after another, then

$$[T]_{\mathcal{B}}^{\mathcal{B}} = [T|_{G_1}]_{\mathcal{B}_1}^{\mathcal{B}_1} \;\oplus\; \cdots \;\oplus\; [T|_{G_k}]_{\mathcal{B}_k}^{\mathcal{B}_k}$$

is block diagonal, with the i-th block of size a_i.

Idea. Almost everything is already proved; the one new step is (a), the identification $G_i = \ker(T-\lambda_i\,\mathrm{id}_V)^{s_i}$. One inclusion is free, since $\ker(T-\lambda_i)^{s_i}$ sits inside the union defining G_i. For the other, count. Chapter 9 says the primary piece $\ker(T-\lambda_i)^{s_i}$ has dimension a_i, so it is enough to show $\dim G_i \leq a_i$. Now G_i is invariant, and the only eigenvalue of T on G_i is λ_i: an eigenvector in G_i for μ is killed by $(\mu-\lambda_i)^n$, which forces $\mu = \lambda_i$. So the characteristic polynomial of the restriction is a power of $x-\lambda_i$ dividing p_T, and its degree $\dim G_i$ is at most a_i.

Proof. Since p_T splits and $m_T \mid p_T$, Corollary 9.6.2 (b) gives the displayed factorization of m_T with $1 \le s_i \le a_i$. Put $V_i = \ker(T - \lambda_i \operatorname{id}_V)^{s_i}$. The polynomials $x - \lambda_i$ are distinct monic irreducibles, so Theorem 9.8.1 applies to $m_T = \prod_i (x - \lambda_i)^{s_i}$ and gives

$$V = V_1 \oplus \cdots \oplus V_k,$$

with each V_i invariant and nonzero. By Corollary 9.8.4, $\dim V_i = a_i$, $p_{T|_{V_i}} = (x - \lambda_i)^{a_i}$, and $E_{\lambda_i}(T) \subseteq V_i$ with equality if and only if $s_i = 1$.

Claim: $G_i = V_i$. Fix i and write $G = G_i$, $\lambda = \lambda_i$, $a = a_i$, $N = T - \lambda \operatorname{id}_V$.

($\supseteq$) $V_i = \ker N^{s_i} \subseteq \bigcup_{j \ge 1} \ker N^j = G$.

($\subseteq$) It suffices to prove $\dim G \le a$: then $V_i \subseteq G$ and $\dim V_i = a \ge \dim G$ force $\dim V_i = \dim G$, and Theorem 1.6.8, applied to the subspace V_i of G, gives $V_i = G$.

First, *the only eigenvalue of* $T|_G$ *is* λ. Indeed, G is T-invariant (Corollary 10.1.2 (b)), so the restriction makes sense (Definition 9.1.2). Let $\mu \in F$ be an eigenvalue of $T|_G$, with eigenvector $v \in G$, $v \neq 0$. From $Tv = \mu v$ we get $Nv = (\mu - \lambda)v$, hence $N^n v = (\mu - \lambda)^n v$ by induction. But $G = \ker N^n$ by Corollary 10.1.2 (a), so $(\mu - \lambda)^n v = 0$ with $v \neq 0$, which forces the scalar $(\mu - \lambda)^n$ to be 0 (Theorem 1.1.9), and so $\mu = \lambda$, since a field has no zero divisors (Theorem 0.7.1).

Now compare characteristic polynomials. $G \neq \{0\}$, since λ is an eigenvalue of T (Corollary 10.1.2 (c)). If $G = V$, then every root of p_T is an eigenvalue of $T = T|_G$ (Theorem 9.2.1), so λ is the only root; as p_T splits and is monic of degree n (Theorem 7.10.2), $p_T = (x - \lambda)^n$ and $a = n = \dim G$. If $G \neq V$, then Theorem 9.1.3 (b) applies to the invariant subspace G and gives

$$p_T = p_{T|_G}\, p_{\bar{T}},$$

so $p_{T|_G}$ is a monic divisor of $p_T = \prod_j (x - \lambda_j)^{a_j}$. By Lemma 9.5.2, $p_{T|_G} = \prod_j (x - \lambda_j)^{c_j}$ with $0 \le c_j \le a_j$. Its roots are the eigenvalues of $T|_G$ (Theorem 9.2.1, applied to $T|_G$), which we just showed to be λ alone, so $c_j = 0$ for $j \neq i$. Hence $p_{T|_G} = (x - \lambda)^{c_i}$, and comparing degrees (Theorem 7.10.2) gives $\dim G = c_i \le a$. This proves the claim, and with it (a), (b), and (c), since $G_i = V_i$.

(d) By Theorem 9.8.1 (c) and the claim, $m_{T|_{G_i}} = (x - \lambda_i)^{s_i}$. For an integer $j \ge 0$, the operator N_i^j is $(x - \lambda_i)^j$ evaluated at $T|_{G_i}$, so $N_i^j = 0$ if and only if $(x - \lambda_i)^{s_i}$ divides $(x - \lambda_i)^j$ (Theorem 9.5.1), that is, if and only if $j \ge s_i$. Since $s_i \le a_i$, this gives $N_i^{a_i} = 0$, and it identifies s_i as the least exponent killing N_i. Hence $\ker N_i^{s_i} = G_i$ while $\ker N_i^{s_i - 1} \subsetneq G_i$, so the stabilization index of N_i is s_i.

(e) Each G_i is invariant and nonzero, and $V = \bigoplus_i G_i$, so Theorem 9.1.5 gives the block diagonal shape, the i-th block being $[T|_{G_i}]_{\mathcal{B}_i}^{\mathcal{B}_i}$, of size $\dim G_i = a_i$.

■

Read (d) again, since it is the plan for the rest of the chapter. **On its generalized eigenspace, T is a scalar plus a nilpotent operator.** The scalar part is as simple as an operator can be. So the only question left is what a nilpotent operator looks like in a well-chosen basis, and that is the subject of the next section. Once it is answered, (e) assembles the answers for the different eigenvalues into one matrix for T.

Two consistency checks. Summing dimensions in (a) and using (c) gives $\sum_i a_i = n$, which is

the degree count for a split p_T. And if T is diagonalizable, then every $s_i = 1$ (Theorem 9.8.2), so (b) says $G_i = E_{\lambda_i}(T)$ and (a) becomes the eigenspace decomposition $V = \bigoplus_i E_{\lambda_i}(T)$ of Theorem 9.4.1 (c). Nothing is lost, and nothing new is gained, in the diagonalizable case.

Example. Generalized Eigenspaces of a 4×4 Matrix

Let

$$\mathbf{A} = \begin{pmatrix} 2 & 1 & -1 & -2 \\ -3 & 3 & -1 & 1 \\ 3 & -1 & 3 & -1 \\ -3 & 1 & -1 & 3 \end{pmatrix} \in M_4(\mathbb{Q}), \qquad p_{\mathbf{A}} = (x-2)^3(x-5).$$

Find $E_2(\mathbf{A})$, $G_2(\mathbf{A})$, $G_5(\mathbf{A})$, the stabilization index of $\mathbf{A} - 2\mathbf{I}$, and the minimal polynomial. Verify Theorem 10.1.3 on this matrix.

Solution. The eigenspace. Write $\mathbf{N} = \mathbf{A} - 2\mathbf{I}$, so

$$\mathbf{N} = \begin{pmatrix} 0 & 1 & -1 & -2 \\ -3 & 1 & -1 & 1 \\ 3 & -1 & 1 & -1 \\ -3 & 1 & -1 & 1 \end{pmatrix}.$$

Rows 2 and 4 are equal and row 3 is their negative, so the row space is spanned by rows 1 and 2, which are not proportional; hence $\operatorname{rank}\mathbf{N} = 2$ and $\operatorname{nullity}\mathbf{N} = 2$ (Theorem 3.5.7). The kernel is cut out by

$$x_2 - x_3 - 2x_4 = 0, \qquad -3x_1 + x_2 - x_3 + x_4 = 0.$$

Subtracting the first equation from the second gives $-3x_1 + 3x_4 = 0$, so $x_1 = x_4$, and then $x_2 = x_3 + 2x_4$. With $x_3 = s$ and $x_4 = t$ free,

$$E_2(\mathbf{A}) = \{(t,\, s+2t,\, s,\, t)\} = \operatorname{span}\left((0,1,1,0),\ (1,2,0,1)\right), \qquad g_{\mathbf{A}}(2) = 2.$$

Since $a_{\mathbf{A}}(2) = 3 > 2$, the matrix is not diagonalizable (Theorem 9.4.1 (e)).

The generalized eigenspace. Squaring,

$$\mathbf{N}^2 = 9\begin{pmatrix} 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 1 \\ 1 & 0 & 0 & -1 \\ -1 & 0 & 0 & 1 \end{pmatrix},$$

which has rank 1, so $\ker\mathbf{N}^2$ has dimension 3 and is the hyperplane

$$\ker\mathbf{N}^2 = \{\,\mathbf{x} : x_1 = x_4\,\} = \operatorname{span}\left(\mathbf{e}_2,\ \mathbf{e}_3,\ \mathbf{e}_1 + \mathbf{e}_4\right).$$

The chain of kernels has dimensions $0, 2, 3$ so far. Since $\ker\mathbf{N}^2 \subseteq G_2(\mathbf{A})$ and $\dim G_2(\mathbf{A}) = a_{\mathbf{A}}(2) = 3 = \dim\ker\mathbf{N}^2$ by part (c) of the theorem, we get $G_2(\mathbf{A}) = \ker\mathbf{N}^2$, and the stabilization index is $s = 2$. One can see the stabilization by hand as well: multiplying out gives $\mathbf{N}^3 = 3\mathbf{N}^2$, so $\ker\mathbf{N}^3 = \ker\mathbf{N}^2$ directly.

The second generalized eigenspace. Here $a_{\mathbf{A}}(5) = 1$, so $\dim G_5(\mathbf{A}) = 1$ and $G_5(\mathbf{A}) = E_5(\mathbf{A})$ by part (b) — its s must be 1. Solving $(\mathbf{A} - 5\mathbf{I})\mathbf{x} = \mathbf{0}$ gives $G_5(\mathbf{A}) = E_5(\mathbf{A}) =$

$\operatorname{span}\big((0,1,-1,1)\big)$. As a check, $\mathbf{A}(0,1,-1,1) = (1+1-2,\ 3+1+1,\ -1-3-1,\ 1+1+3) = (0,5,-5,5)$.

The decomposition. $\mathbb{Q}^4 = G_2(\mathbf{A}) \oplus G_5(\mathbf{A})$, of dimensions $3+1=4$, and the sum is direct because $(0,1,-1,1) \notin G_2(\mathbf{A})$: its first and last coordinates are $0 \neq 1$.

The minimal polynomial. The stabilization indices are $s_1 = 2$ at $\lambda = 2$ and $s_2 = 1$ at $\lambda = 5$, and part (d) of the theorem identifies these indices as the exponents in $m_{\mathbf{A}}$, so $m_{\mathbf{A}} = (x-2)^2(x-5)$. Consistently, $(\mathbf{A}-2\mathbf{I})(\mathbf{A}-5\mathbf{I}) \neq \mathbf{O}$: the two factors are coprime, so the kernel of the product is $\ker(\mathbf{A}-2\mathbf{I}) \oplus \ker(\mathbf{A}-5\mathbf{I})$ (Theorem 6.7.5 (a)), of dimension $2+1=3<4$.

A basis adapted to the splitting. Take $\mathcal{B} = (\mathbf{e}_2,\ \mathbf{e}_3,\ \mathbf{e}_1+\mathbf{e}_4,\ (0,1,-1,1))$, a basis of G_2 followed by one of G_5. By part (e), $[T_{\mathbf{A}}]_{\mathcal{B}}^{\mathcal{B}}$ is block diagonal with a 3×3 block and the 1×1 block (5).

Warning. **The dimension of G_λ is the algebraic multiplicity, not the number of steps the chain takes.** In the example, $\dim G_2 = 3$ while the chain $\{\mathbf{0}\} \subsetneq \ker \mathbf{N} \subsetneq \ker \mathbf{N}^2 = G_2$ stabilizes after $s=2$ steps. The two numbers measure different things: $a(\lambda) = \dim G_\lambda$ is the total size of the piece, and s is the exponent of $x-\lambda$ in the minimal polynomial. All that holds in general is $1 \le s \le a(\lambda)$ and $g(\lambda) \le a(\lambda)$ (Theorem 9.3.1).

Finally, a word on what we may not yet assume. The decomposition needs p_T to split, which is automatic over $\mathbb{C}$ (Corollary 6.5.5) but fails, for instance, for a rotation of $\mathbb{R}^2$, whose generalized eigenspaces over $\mathbb{R}$ are all $\{\mathbf{0}\}$ and certainly do not fill $\mathbb{R}^2$. Operators over an arbitrary field need the coarser primary decomposition of Chapter 9, whose pieces we return to in Section 6.

10.1.5 Exercises

10.1.5.1 A. Check your understanding

Exercise. A1

(a) Define a generalized eigenvector and the generalized eigenspace $G_\lambda(T)$.

(b) State what Theorem 10.1.1 says about the chain $\ker N \subseteq \ker N^2 \subseteq \dots$, including the bound on where it stabilizes.

(c) True or false: every generalized eigenvector of T for λ satisfies $Tv = \lambda v$. Justify your answer.

(d) True or false: if λ is not an eigenvalue of T, then $G_\lambda(T) = \{0\}$. Justify your answer.

(e) Let $\dim V = 6$ and let $p_T = (x-1)^4(x-3)^2$ split over F. What are $\dim G_1(T)$ and $\dim G_3(T)$?

(f) In the situation of (e), what can you say about $\dim E_1(T)$?

Solution.

(a) A generalized eigenvector of T for $\lambda \in F$ is a **nonzero** $v \in V$ with $(T - \lambda\,\mathrm{id}_V)^k v = 0$ for **some** $k \ge 1$ (Definition 10.1.1); $G_\lambda(T)$ is the set of all of them together with 0,

that is, $\bigcup_{k\geq 1}\ker(T-\lambda\,\mathrm{id}_V)^k$ (Definition 10.1.2).

(b) The kernels increase; once two consecutive ones are equal, all later ones equal them; the strict inclusions come first, and the chain has stabilized by step $\dim V$, so $G_\lambda(T)=\ker(T-\lambda\,\mathrm{id}_V)^{\dim V}$.

(c) False. For $\mathbf{J}=\begin{pmatrix}2&1\\0&2\end{pmatrix}$, the vector $\mathbf{e}_2$ lies in $G_2(\mathbf{J})$ but $\mathbf{J}\mathbf{e}_2=\mathbf{e}_1+2\mathbf{e}_2\neq 2\mathbf{e}_2$. Only the vectors of $E_\lambda(T)\subseteq G_\lambda(T)$ satisfy the eigenvector equation.

(d) True, by Corollary 10.1.2 (c).

(e) $\dim G_1(T)=a_T(1)=4$ and $\dim G_3(T)=a_T(3)=2$ (Theorem 10.1.3 (c)).

(f) Only $1\leq\dim E_1(T)\leq 4$, since $E_1(T)\subseteq G_1(T)$ and $1\leq g_T(1)\leq a_T(1)$ (Theorem 9.3.1). All four values occur, as the Jordan forms of Section 3 will show.

10.1.5.2 B. Practice

Exercise. B1: Computing generalized eigenspaces

For each operator, find all generalized eigenspaces, and compare each with the corresponding eigenspace.

(a) $\mathbf{A}_1=\begin{pmatrix}3&1\\0&3\end{pmatrix}\in M_2(\mathbb{Q})$.

(b) $\mathbf{A}_2=\begin{pmatrix}2&1&1\\0&1&-1\\-1&-1&1\end{pmatrix}\in M_3(\mathbb{R})$, which has $m_{\mathbf{A}_2}=(x-1)^2(x-2)$ (Exercise (B2: Decomposing $\mathbb{R}^3$)).

(c) $D\in\mathcal{L}(\mathbb{R}[x]_{\leq 3})$, differentiation.

(d) The rotation $\mathbf{R}=\begin{pmatrix}0&-1\\1&0\end{pmatrix}$, first over $\mathbb{R}$ and then over $\mathbb{C}$.

Solution.

(a) $p_{\mathbf{A}_1}=(x-3)^2$, so 3 is the only eigenvalue, and $G_\lambda=\{\mathbf{0}\}$ for $\lambda\neq 3$ (Corollary 10.1.2 (c)). Since $(\mathbf{A}_1-3\mathbf{I})^2=\mathbf{O}$, $G_3(\mathbf{A}_1)=\mathbb{Q}^2$, while $E_3(\mathbf{A}_1)=\ker\begin{pmatrix}0&1\\0&0\end{pmatrix}=\mathrm{span}(\mathbf{e}_1)$ is a line.

(b) From $m_{\mathbf{A}_2}=(x-1)^2(x-2)$ and Theorem 10.1.3 (a), $G_1=\ker(\mathbf{A}_2-\mathbf{I})^2$ and $G_2=\ker(\mathbf{A}_2-2\mathbf{I})$. Chapter 9 computed these: $G_1=\mathrm{span}((-1,1,0),(0,0,1))$, the plane $x+y=0$, and $G_2=\mathrm{span}((0,-1,1))$. The eigenspace $E_1(\mathbf{A}_2)=\mathrm{span}((-1,1,0))$ is a line strictly inside G_1, while $E_2(\mathbf{A}_2)=G_2$ because $s_2=1$. All other G_λ are $\{\mathbf{0}\}$.

(c) $D^4=0$ (Example (Differentiation Is Nilpotent)), so $G_0(D)=\mathbb{R}[x]_{\leq 3}$, of dimension 4, while $E_0(D)=\ker D=\mathbb{R}$ is the line of constants. Since $p_D=x^4$ has 0 as its only root, every other generalized eigenspace is $\{0\}$.

(d) Over $\mathbb{R}$, $p_{\mathbf{R}}=x^2+1$ has no root, so $\mathbf{R}$ has no eigenvalue and $G_\lambda(\mathbf{R})=\{\mathbf{0}\}$ for every $\lambda\in\mathbb{R}$. The generalized eigenspaces do **not** fill $\mathbb{R}^2$; the hypothesis that $p_{\mathbf{R}}$ splits fails. Over $\mathbb{C}$, $p_{\mathbf{R}}=(x-i)(x+i)$ splits with $a(i)=a(-i)=1$, so $G_i(\mathbf{R})=E_i(\mathbf{R})=$

$\operatorname{span}((1,-i))$ and $G_{-i}(\mathbf{R}) = E_{-i}(\mathbf{R}) = \operatorname{span}((1,i))$, and $\mathbb{C}^2 = G_i \oplus G_{-i}$. (Check: $\mathbf{R}(1,-i) = (i,1) = i(1,-i)$.)

Exercise. B2: The image chain stabilizes too

Let V be finite-dimensional with $\dim V = n$, and let $N \in \mathcal{L}(V)$.

(a) Prove that $\operatorname{im} N \supseteq \operatorname{im} N^2 \supseteq \operatorname{im} N^3 \supseteq \dots$, and that each $\operatorname{im} N^k$ is N-invariant.

(b) Prove that $\operatorname{im} N^m = \operatorname{im} N^{m+1}$ if and only if $\ker N^m = \ker N^{m+1}$.

(c) Deduce that $\operatorname{im} N^n = \operatorname{im} N^{n+j}$ for every $j \geq 0$, and that the image chain stabilizes at exactly the same index s as the kernel chain.

Solution.

(a) $\operatorname{im} N^{k+1} = \{N^{k+1}v : v \in V\} = \{N^k(Nv) : v \in V\} \subseteq \operatorname{im} N^k$. Each $\operatorname{im} N^k$ is the image of a polynomial in N, hence N-invariant (Theorem 6.7.4 (b)).

(b) By Theorem 2.3.1, $\dim \operatorname{im} N^k + \dim \ker N^k = n$ for every k. The inclusions $\operatorname{im} N^{m+1} \subseteq \operatorname{im} N^m$ and $\ker N^m \subseteq \ker N^{m+1}$ are equalities exactly when the dimensions agree (Theorem 1.6.8), and $\dim \operatorname{im} N^m = \dim \operatorname{im} N^{m+1}$ if and only if $\dim \ker N^m = \dim \ker N^{m+1}$ by the displayed identity.

(c) By Theorem 10.1.1 (d), $\ker N^n = \ker N^{n+j}$; iterating (b) at each step from n to $n+j$ gives $\operatorname{im} N^n = \operatorname{im} N^{n+j}$. And by (b), the least index at which two consecutive images agree is the least index at which two consecutive kernels agree, namely s.

Exercise. B3: Reading the dimensions off a triangular matrix

Let

$$\mathbf{A} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 2 & 1 \\ 0 & 0 & 0 & 2 \end{pmatrix} \in M_4(\mathbb{Q}).$$

(a) Write down $p_{\mathbf{A}}$ and the algebraic multiplicities.

(b) Compute $E_1(\mathbf{A})$, $G_1(\mathbf{A})$, $E_2(\mathbf{A})$, and $G_2(\mathbf{A})$, and check $\dim G_\lambda = a_{\mathbf{A}}(\lambda)$.

(c) Verify that $\mathbb{Q}^4 = G_1(\mathbf{A}) \oplus G_2(\mathbf{A})$, and find $m_{\mathbf{A}}$.

Solution.

(a) $\mathbf{A}$ is upper triangular, so $p_{\mathbf{A}} = (x-1)^2(x-2)^2$ (Theorem 7.4.6), and $a_{\mathbf{A}}(1) = a_{\mathbf{A}}(2) = 2$.

(b) $\mathbf{A} - \mathbf{I} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}$ has rank 3 (rows $1, 2, 4$ are independent and row 3 is the sum

of rows 2 and 4), so $E_1(\mathbf{A}) = \operatorname{span}(\mathbf{e}_1)$. Squaring,

$$(\mathbf{A}-\mathbf{I})^2 = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

of rank 2, with kernel $\operatorname{span}(\mathbf{e}_1, \mathbf{e}_2)$; so $G_1(\mathbf{A}) = \operatorname{span}(\mathbf{e}_1, \mathbf{e}_2)$, of dimension $2 = a_{\mathbf{A}}(1)$. (Membership can be seen directly: $(\mathbf{A}-\mathbf{I})\mathbf{e}_2 = \mathbf{e}_1$ and $(\mathbf{A}-\mathbf{I})\mathbf{e}_1 = \mathbf{0}$.)

Similarly $\mathbf{A} - 2\mathbf{I} = \begin{pmatrix} -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$ has rank 3, so $E_2(\mathbf{A})$ is a line; solving gives $E_2(\mathbf{A}) = \operatorname{span}((1,1,1,0))$. And $(\mathbf{A}-2\mathbf{I})^2 = \begin{pmatrix} 1 & -2 & 1 & 0 \\ 0 & 1 & -1 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$ has rank 2; its kernel is cut out by $x_1 - 2x_2 + x_3 = 0$ and $x_2 - x_3 + x_4 = 0$, so $G_2(\mathbf{A}) = \operatorname{span}((1,1,1,0), (-2,-1,0,1))$, of dimension $2 = a_{\mathbf{A}}(2)$. (Check: $(\mathbf{A}-2\mathbf{I})(-2,-1,0,1) = (1,1,1,0)$, the eigenvector.)

(c) The four vectors $\mathbf{e}_1, \mathbf{e}_2, (1,1,1,0), (-2,-1,0,1)$ form a basis of $\mathbb{Q}^4$: the matrix with these columns is upper triangular with diagonal $1,1,1,1$, hence invertible. So the sum is direct and equals $\mathbb{Q}^4$, as Theorem 10.1.3 (a) predicts. Both stabilization indices are 2 (the eigenspaces are lines, the generalized eigenspaces planes), so $m_{\mathbf{A}} = (x-1)^2(x-2)^2 = p_{\mathbf{A}}$.

10.1.5.3 C. Going deeper

Exercise. C1: The Fitting decomposition

Let V be finite-dimensional with $\dim V = n$, and let $N \in \mathcal{L}(V)$ be **any** operator. Put $U = \ker N^n$ and $W = \operatorname{im} N^n$.

(a) Prove that $U \cap W = \{0\}$, and deduce $V = U \oplus W$.

(b) Prove that U and W are N-invariant, that $N|_U$ is nilpotent, and that $N|_W$ is invertible.

(c) Identify U as a generalized eigenspace of N.

(d) Compute U and W for $\mathbf{N} = \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix} \in M_2(\mathbb{Q})$.

Hint: for (a), if $v = N^n w$ and $N^n v = 0$, look at which kernel w lies in.

Solution.

(a) Let $v \in U \cap W$, say $v = N^n w$ with $N^n v = 0$. Then $N^{2n} w = N^n v = 0$, so $w \in \ker N^{2n} = \ker N^n$ by Theorem 10.1.1 (d), and therefore $v = N^n w = 0$. By Theorem 2.3.1, $\dim U + \dim W = n$, so $\dim(U+W) = \dim U + \dim W - \dim(U \cap W) = n$ (Theorem 1.7.3), whence $U + W = V$ (Theorem 1.6.8) and the sum is direct (Theorem 1.7.4).

(b) Both are invariant by Theorem 6.7.4 (b). On U we have $(N|_U)^n = 0$ by definition

of U, so $N|_U$ is nilpotent. On W, the map $N|_W\colon W \to W$ is surjective: $N(W) = N(\operatorname{im} N^n) = \operatorname{im} N^{n+1} = \operatorname{im} N^n = W$ by Exercise (B2: The image chain stabilizes too) (c). A surjective operator on a finite-dimensional space is invertible (Corollary 2.3.2 (e)).

(c) $U = \ker N^n = G_0(N)$, by Corollary 10.1.2 (a) with $\lambda = 0$. So the Fitting decomposition splits V into the generalized eigenspace for 0 and a part on which N is invertible.

(d) Here $n = 2$ and $\mathbf{N}^2 = \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix} = \mathbf{N}$, so $U = \ker \mathbf{N} = \operatorname{span}(\mathbf{e}_1)$ and $W = \operatorname{im} \mathbf{N} = \operatorname{span}((1,1))$. Indeed $\mathbb{Q}^2 = \operatorname{span}(\mathbf{e}_1) \oplus \operatorname{span}((1,1))$, $\mathbf{N}$ is zero on U, and $\mathbf{N}(1,1) = (1,1)$, so $\mathbf{N}|_W$ is the identity.

Exercise. C2: Commuting operators preserve generalized eigenspaces

Let V be finite-dimensional, and let $S, T \in \mathcal{L}(V)$ with $ST = TS$.

(a) Prove that $S(G_\lambda(T)) \subseteq G_\lambda(T)$ for every $\lambda \in F$.

(b) Deduce that if p_T splits, then every generalized eigenspace of T is invariant under every polynomial in T and under S, and that V has a basis in which both S and T have block diagonal matrices with blocks of sizes $a_T(\lambda_1), \dots, a_T(\lambda_k)$.

(c) Show by example that part (a) can be vacuous: find commuting S, T on $\mathbb{Q}^2$ and a λ with $E_\lambda(T)$ a line, $G_\lambda(T) = \mathbb{Q}^2$, and S not a scalar multiple of the identity.

Solution.

(a) Since $ST = TS$, S commutes with every polynomial in T (Theorem 6.7.1 (d)), in particular with N^k, where $N = T - \lambda \operatorname{id}_V$. Let $v \in G_\lambda(T)$, say $N^k v = 0$. Then

$$N^k(Sv) = S(N^k v) = S0 = 0,$$

so $Sv \in \ker N^k \subseteq G_\lambda(T)$. (This is Theorem 6.7.4 (a) with $R = N^k$.)

(b) Invariance under polynomials in T is Theorem 6.7.4 (b), and invariance under S is (a). If p_T splits, then $V = \bigoplus_i G_{\lambda_i}(T)$ with $\dim G_{\lambda_i}(T) = a_T(\lambda_i)$ (Theorem 10.1.3), and a basis assembled from bases of the summands makes both matrices block diagonal, by Theorem 9.1.5 applied to T and to S.

(c) Take $T = \begin{pmatrix} 2 & 1 \\ 0 & 2 \end{pmatrix}$ and $S = T$. They commute, $E_2(T) = \operatorname{span}(\mathbf{e}_1)$ is a line, $G_2(T) = \mathbb{Q}^2$, and S is not a scalar matrix. Consistently with Theorem 9.9.1, S maps the line $\operatorname{span}(\mathbf{e}_1)$ into itself, and by (a) it maps $G_2(T) = \mathbb{Q}^2$ into itself, which says nothing. The lesson is that (a) is informative only when $G_\lambda(T) \neq V$.

10.2 Nilpotent Operators

Section 1 reduced our problem to one piece at a time: on the generalized eigenspace $G_\lambda(T)$, the operator is $\lambda \operatorname{id} + N$ with some power of N equal to zero (Theorem 10.1.3 (d)). The scalar part needs no work, so everything now depends on a single question: what does an operator all of whose powers eventually vanish look like in the right basis? This section answers it completely,

and the answer is a list of numbers — the sizes of the staircases the operator builds.

10.2.1 Operators killed by a power

The objects have already appeared several times. Differentiation on $F[x]_{\le n}$ satisfies $D^{n+1} = 0$ (Example (Differentiation Is Nilpotent)). The matrix $\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$ has square $\mathbf{O}$. Over a field of characteristic 0, a commutator $[\mathbf{A}, \mathbf{B}]$ that commutes with $\mathbf{A}$ has some power equal to $\mathbf{O}$ (Exercise (C2: When $\mathbf{A}$ commutes with $[\mathbf{A}, \mathbf{B}]$)). Chapter 2 gave these operators their name in passing; here is the definition in full, with the one extra piece of data we shall need.

A nilpotent operator is one that some power annihilates: repeated application destroys every vector.

Definition 10.2.1. Nilpotent Operator, Index of Nilpotency

Let V be a vector space over F and $N \in \mathcal{L}(V)$. We say N is **nilpotent** if

$$N^s = 0 \qquad \text{for some integer } s \ge 1.$$

If $V \neq \{0\}$ and N is nilpotent, the **index of nilpotency** of N is the **least** such s; equivalently, $N^s = 0$ and $N^{s-1} \neq 0$. A matrix $\mathbf{A} \in M_n(F)$ is **nilpotent** if $\mathbf{A}^s = \mathbf{O}$ for some $s \ge 1$, that is, if $T_{\mathbf{A}}$ is nilpotent.

In words: one **single** exponent must kill **every** vector at once, and the exponent is at least 1, so the zero operator is nilpotent (index 1) but the identity on $V \neq \{0\}$ is not. The index is well defined: the exponents $s \ge 1$ with $N^s = 0$ form a nonempty set of positive integers, which has a least element (Theorem 0.2.11).

Examples.

- **The shift.** On F^k, let $N\mathbf{e}_1 = \mathbf{0}$ and $N\mathbf{e}_j = \mathbf{e}_{j-1}$ for $j \ge 2$. Then N^k kills every $\mathbf{e}_j$, while $N^{k-1}\mathbf{e}_k = \mathbf{e}_1 \neq \mathbf{0}$, so N is nilpotent of index k. Its matrix in the standard basis has a 1 in each position $(j-1, j)$ and zeros elsewhere. This matrix is the model for everything in this section, and we write it

$$\mathbf{J}_k(0) \coloneqq \begin{pmatrix} 0 & 1 & & \\ & 0 & \ddots & \\ & & \ddots & 1 \\ & & & 0 \end{pmatrix} \in M_k(F),$$

  with 1 in every **superdiagonal** entry $(j, j+1)$ and 0 everywhere else. (Section 3 writes $\mathbf{J}_k(\lambda) = \lambda\mathbf{I}_k + \mathbf{J}_k(0)$ and calls it a Jordan block.) For $k = 1$, $\mathbf{J}_1(0) = \mathbf{O}_1$.
- **Differentiation.** D on $F[x]_{\le n}$ has $D^{n+1} = 0$, and over $\mathbb{R}$ the index is exactly $n+1$, since $D^n x^n = n! \neq 0$ (Example (Differentiation Is Nilpotent)). Over $\mathbb{R}$, in the basis $(1, x, x^2/2!, \dots, x^n/n!)$, its matrix is exactly $\mathbf{J}_{n+1}(0)$, because $D(x^j/j!) = x^{j-1}/(j-1)!$.
- **All entries nonzero.** $\mathbf{A} = \begin{pmatrix} 2 & 4 \\ -1 & -2 \end{pmatrix}$ has $\mathbf{A}^2 = \begin{pmatrix} 4-4 & 8-8 \\ -2+2 & -4+4 \end{pmatrix} = \mathbf{O}$, so $\mathbf{A}$ is nilpotent of index 2, although no entry of $\mathbf{A}$ is 0. Nilpotency is not visible from the shape of the entries.
- **The degenerate case.** On a space of dimension 1 the only nilpotent operator is 0: if $N = c\,\mathrm{id}$ and $N^s = 0$, then $c^s = 0$, so $c = 0$. Small as it is, this is why a 1×1 block in a

nilpotent normal form carries a 0 and nothing else.

Nonexample by minimal change. Move the 1 in $\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$ to the other off-diagonal position and take $\mathbf{S} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$, the swap of the two coordinates. It is just as sparse, but $\mathbf{S}^2 = \mathbf{I}$, so $\mathbf{S}^{2k} = \mathbf{I} \neq \mathbf{O}$ and $\mathbf{S}^{2k+1} = \mathbf{S} \neq \mathbf{O}$ for every k: **no** power vanishes, and the defining clause fails outright. A gentler nonexample is $\operatorname{diag}(0, 1)$, which is singular and has 0 as an eigenvalue, yet $\operatorname{diag}(0, 1)^k = \operatorname{diag}(0, 1) \neq \mathbf{O}$; being noninvertible is far from being nilpotent.

Why this definition. The clause "for some s" cannot be replaced by a fixed exponent without losing the class, since on $\mathbb{R}[x]_{\leq n}$ differentiation needs $n + 1$ steps. On the other hand the exponent may always be taken to be $\dim V$, as the next proposition shows, so the definition is not as open-ended as it looks. The name says "zero power": *nil* (nothing) and *potens* (power).

Warning. **Nilpotent is a statement about one exponent killing everything, not about each vector separately, and not about the entries.** Differentiation on $\mathbb{R}[x]$ (**all** polynomials) kills every **single** polynomial after enough steps, yet no power of it is 0, because the number of steps needed is unbounded; the space is infinite-dimensional. And the matrix $\begin{pmatrix} 2 & 4 \\ -1 & -2 \end{pmatrix}$ is nilpotent with no zero entry, while $\operatorname{diag}(0, 1)$ has a zero entry, a zero eigenvalue, and is not nilpotent.

The first payoff collects the facts we shall use constantly. Note the interplay with Section 1: a nilpotent operator is one whose generalized eigenspace for 0 is everything.

Proposition 10.2.1. Basic Properties of Nilpotent Operators

Let V be a finite-dimensional vector space over F with $\dim V = n \geq 1$, and let $N \in \mathcal{L}(V)$. Then:

(a) N is nilpotent if and only if $N^n = 0$, if and only if $G_0(N) = V$;
(b) if N is nilpotent with index s, then $s \leq n$, $m_N = x^s$, and $p_N = x^n$;
(c) if N is nilpotent, then $\operatorname{spec}(N) = \{0\}$, and there is a basis in which the matrix of N is upper triangular with zero diagonal;
(d) conversely, if $p_N = x^n$, then N is nilpotent.

Idea. For (a) and (b), the kernel chain of Section 1 does the work: $N^s = 0$ says $\ker N^s = V$, and the chain has stopped growing by step n. For the triangular matrix in (c), build the basis along the chain: take a basis of $\ker N$, extend it to one of $\ker N^2$, and so on. A vector first appearing at level k is sent by N into $\ker N^{k-1}$, which is spanned by vectors listed **earlier**, so every column has zeros from the diagonal down.

Proof.

(a) If $N^s = 0$ for some $s \geq 1$, then $\ker N^s = V$, so $V = \ker N^{\max(s,n)} = \ker N^n$ by Theorem 10.1.1 (a) and (d), that is, $N^n = 0$. The converse is the definition with $s = n$.

And $G_0(N) = \ker N^n$ by Corollary 10.1.2 (a) with $\lambda = 0$, so $G_0(N) = V$ if and only if $N^n = 0$.

(b) Let s be the index and put $K_k = \ker N^k$, so $\{0\} = K_0 \subseteq K_1 \subseteq \cdots \subseteq K_s = V$ by (a) and Theorem 10.1.1 (a). Choose a basis $\mathcal{B}_1$ of K_1; extend it to a basis $\mathcal{B}_2$ of K_2; and continuing, extend $\mathcal{B}_{k-1}$ to a basis $\mathcal{B}_k$ of K_k for $k = 2, \dots, s$ (Theorem 1.6.5, legitimate because $\mathcal{B}_{k-1}$ is an independent list in K_k). Then $\mathcal{B} = \mathcal{B}_s = (v_1, \dots, v_n)$ is a basis of V, and for each k the first $\dim K_k$ of its vectors form a basis of K_k. Let $1 \le j \le n$ and let k be the level at which v_j was added, so $v_j \in K_k$ and $\dim K_{k-1} < j$. Then $Nv_j \in K_{k-1} = \operatorname{span}(v_1, \dots, v_{\dim K_{k-1}}) \subseteq \operatorname{span}(v_1, \dots, v_{j-1})$. So the j-th column of $[N]_{\mathcal{B}}^{\mathcal{B}}$ has zeros in rows $j, j+1, \dots, n$ (Definition 4.1.1, Theorem 1.5.1): the matrix is upper triangular with zero diagonal. Its diagonal entries are, with multiplicity, the eigenvalues of N (Theorem 9.7.4 (b)), and they are all 0, so $\operatorname{spec}(N) = \{0\}$.

(c) By (a), $N^n = 0$, so $s \le n$. The polynomial x^s annihilates N, and x^{s-1} does not, so $m_N \mid x^s$ and $m_N \nmid x^{s-1}$ (Theorem 9.5.1). Every monic divisor of x^s is x^j with $0 \le j \le s$ (Lemma 9.5.2), so $m_N = x^s$. For p_N, use the basis $\mathcal{B}$ of (c): there $x\mathbf{I}_n - [N]_{\mathcal{B}}^{\mathcal{B}}$ is upper triangular with every diagonal entry equal to x, so $p_N = x^n$ by Theorem 7.4.6 and Definition 7.10.2.

(d) If $p_N = x^n$, then $p_N(N) = N^n = 0$ by Theorem 9.6.1.

■

So nilpotency is detected by the characteristic polynomial: N is nilpotent exactly when $p_N = x^n$, which over $\mathbb{C}$ is exactly $\operatorname{spec}(N) = \{0\}$, as Chapter 9 found by a different route (Theorem 9.10.3). Part (c) also follows abstractly from Theorem 9.7.2, since x^n splits over every field; the proof above builds the basis explicitly. But a strictly upper triangular matrix still has $n(n-1)/2$ entries to describe, and almost all of them can be cleared away. That is the content of the structure theorem.

10.2.2 What one chain looks like

Look again at $\mathbf{J}_k(0)$ and read its columns, which is where a matrix always tells us what the operator does to basis vectors. With $\mathcal{B} = (\mathbf{b}_1, \dots, \mathbf{b}_k)$ the standard basis of F^k, the matrix $\mathbf{J}_k(0)$ says

$$N\mathbf{b}_1 = \mathbf{0}, \qquad N\mathbf{b}_2 = \mathbf{b}_1, \qquad N\mathbf{b}_3 = \mathbf{b}_2, \quad \dots, \quad N\mathbf{b}_k = \mathbf{b}_{k-1}.$$

Reading this from the right, the last basis vector $\mathbf{b}_k$ generates all the others: $\mathbf{b}_{k-1} = N\mathbf{b}_k$, $\mathbf{b}_{k-2} = N^2\mathbf{b}_k$, and so on down to $\mathbf{b}_1 = N^{k-1}\mathbf{b}_k$, which is an eigenvector, killed at the next step. So one Jordan block is one vector together with its successive images, a staircase descending to $\mathbf{0}$. Such a list deserves a name.

Definition 10.2.2. Jordan Chain

Let V be a vector space over F, let $N \in \mathcal{L}(V)$, and let $u \in V$ with $u \neq 0$. Suppose $N^k u = 0$ and $N^{k-1}u \neq 0$ for an integer $k \ge 1$. The **Jordan chain generated by** u is the ordered list

$$(N^{k-1}u,\ N^{k-2}u,\ \dots,\ Nu,\ u),$$

of **length** k, and u is called the **top** of the chain. For $T \in \mathcal{L}(V)$ and $\lambda \in F$, a Jordan chain of T for λ means a Jordan chain of $N = T - \lambda\,\mathrm{id}_V$.

The order matters, and we always write it with the **eigenvector first** and the top last, so that

N moves each entry to the one before it and kills the first. The first entry $N^{k-1}u$ is indeed an eigenvector for 0, by the argument of Corollary 10.1.2 (d). For example, the chain generated by $\mathbf{e}_2$ under $\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$ is $(\mathbf{e}_1, \mathbf{e}_2)$, of length 2.

A chain is automatically independent, which is the first sign that chains are good building blocks.

Lemma 10.2.2. A Jordan Chain Is Independent

Let $N \in \mathcal{L}(V)$ and let $(N^{k-1}u, \dots, Nu, u)$ be a Jordan chain, so $N^k u = 0 \neq N^{k-1}u$. Then:

(a) the list is linearly independent, so $k \leq \dim V$ when V is finite-dimensional;

(b) $C = \operatorname{span}(u, Nu, \dots, N^{k-1}u)$ is N-invariant of dimension k, and the matrix of $N|_C$ in the chain, taken in the order of Definition 10.2.2, is $\mathbf{J}_k(0)$.

Proof.

(a) Suppose $a_0 u + a_1 Nu + \cdots + a_{k-1}N^{k-1}u = 0$ with $a_j \in F$. Apply N^{k-1}. Every term with $j \geq 1$ becomes $a_j N^{k-1+j}u = 0$, since $N^k u = 0$ and $k - 1 + j \geq k$. What remains is $a_0 N^{k-1}u = 0$, and $N^{k-1}u \neq 0$, so $a_0 = 0$ (Theorem 1.1.9). Now apply N^{k-2} to the relation with $a_0 = 0$: by the same cancellation only $a_1 N^{k-1}u = 0$ survives, so $a_1 = 0$. Repeating this for $N^{k-3}, \dots, N^0$ gives $a_2 = \cdots = a_{k-1} = 0$. Hence the list is independent, and an independent list has at most $\dim V$ entries (Theorem 1.6.5).

(b) By (a) the k spanning vectors are independent, so $\dim C = k$. Write $b_j = N^{k-j}u$ for $j = 1, \dots, k$, the chain in order. Then $Nb_1 = N^k u = 0$ and $Nb_j = N^{k-j+1}u = b_{j-1}$ for $j \geq 2$, so all images lie in C and C is N-invariant (Lemma 9.1.1). The j-th column of the matrix of $N|_C$ is the coordinate vector of Nb_j, which is $\mathbf{0}$ for $j = 1$ and $\mathbf{e}_{j-1}$ for $j \geq 2$ (Definition 4.1.1). That is precisely $\mathbf{J}_k(0)$.

■

10.2.3 The structure theorem

Now the goal. If a nilpotent operator could be split into chains, its matrix would be a direct sum of blocks $\mathbf{J}_{k_i}(0)$, and nothing would be left to describe except the list of lengths. That is exactly what happens.

Every nilpotent operator is a disjoint union of staircases: the space breaks into chains, and the operator shifts each chain one step down.

Theorem 10.2.3. Structure Theorem for Nilpotent Operators

Let V be a finite-dimensional vector space over F with $\dim V = n \geq 1$, and let $N \in \mathcal{L}(V)$ be nilpotent. Then there are an integer $r \geq 1$, vectors $u_1, \dots, u_r \in V$, and integers

$$k_1 \geq k_2 \geq \cdots \geq k_r \geq 1, \qquad k_1 + k_2 + \cdots + k_r = n,$$

such that:

(a) $N^{k_i}u_i = 0$ for each i, and the list $\mathcal{B}$ obtained by writing the r Jordan chains one after

another,

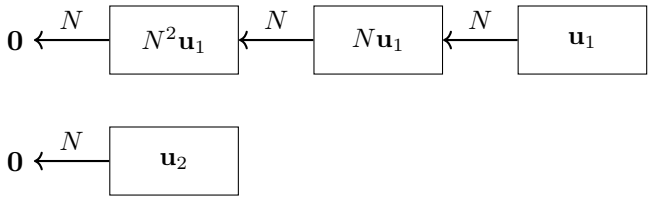
$$\mathcal{B} = (\underbrace{N^{k_1-1}u_1, \dots, Nu_1, u_1}_{\text{chain } 1}, \ \dots, \ \underbrace{N^{k_r-1}u_r, \dots, Nu_r, u_r}_{\text{chain } r}),$$

is a basis of V;

(b) with $C_i = \operatorname{span}(u_i, Nu_i, \dots, N^{k_i-1}u_i)$, each C_i is N-invariant, $V = C_1 \oplus \cdots \oplus C_r$, and

$$[N]_{\mathcal{B}}^{\mathcal{B}} = \mathbf{J}_{k_1}(0) \oplus \mathbf{J}_{k_2}(0) \oplus \cdots \oplus \mathbf{J}_{k_r}(0).$$

A basis as in (a) is called a **Jordan basis** for N. Before the proof, here is the picture the proof builds. Each chain is a row of boxes, the top of the chain on the right, and N moves one box to the left, off the diagram at the left edge. For lengths $(3, 1)$:



the boxes in the left column, $N^2\mathbf{u}_1$ and $\mathbf{u}_2$, form a basis of $\ker N$

Idea. Induction on $\dim V$, with the smaller object found inside: $W = \operatorname{im} N$. It is N-invariant, $N|_W$ is nilpotent, and $\dim W < \dim V$ because a nilpotent operator on a nonzero space is not injective. So W already has chains. Each chain of W is one box short of what we want: its top w_i lies in $\operatorname{im} N$, so $w_i = Nu_i$ for some $u_i \in V$, and we can extend the chain by one box on the right, gaining u_i. What is still missing is counted by $\ker N$: the ends $N^{l_i}u_i$ of the extended chains are independent vectors of $\ker N$, and any basis extension of them inside $\ker N$ supplies the chains of length 1. Then the count comes out exactly right, by Rank–Nullity (Theorem 2.3.1), and independence follows by applying N once and using the basis of W.

Proof. We argue by induction on $n = \dim V \geq 1$, the statement being: for every nilpotent N on a space of dimension n, chains as in (a) exist.

Base case $n = 1$. Here $N = 0$, as noted after Definition 10.2.1. Take $r = 1$, $k_1 = 1$, and u_1 any nonzero vector; then $N^1u_1 = 0$ and $\mathcal{B} = (u_1)$ is a basis.

Induction step. Let $n \geq 2$ and assume the statement for all spaces of dimension less than n. Let $N \in \mathcal{L}(V)$ be nilpotent with $\dim V = n$.

N is not injective: if it were, then N^s would be injective for every s (a composition of injective maps is injective, Theorem 0.4.3 (a)), contradicting $N^s = 0$ on $V \neq \{0\}$. So $\operatorname{nullity} N \geq 1$, and with $W = \operatorname{im} N$ Rank–Nullity gives $\dim W = n - \operatorname{nullity} N \leq n - 1$ (Theorem 2.3.1).

Step 1. Chains inside $\operatorname{im} N$**.** If $W = \{0\}$, then $N = 0$; take $r = n$, all $k_i = 1$, and $(u_1, \dots, u_n)$ any basis of V, and (a) holds. So assume $W \neq \{0\}$. Then W is N-invariant (Theorem 6.7.4 (b)), $N|_W$ is nilpotent ($(N|_W)^s = (N^s)|_W = 0$), and $1 \leq \dim W < n$. By the induction hypothesis there are $w_1, \dots, w_q \in W$ and $l_1 \geq \cdots \geq l_q \geq 1$ with $\sum_i l_i = \dim W$, $(N|_W)^{l_i} w_i = 0$, and

$$\mathcal{C} = (N^{l_1-1}w_1, \dots, w_1, \ \dots, \ N^{l_q-1}w_q, \dots, w_q)$$

a basis of W.

Step 2. Lift each chain by one box. Each w_i lies in $W = \operatorname{im} N$, so we may choose $u_i \in V$ with $Nu_i = w_i$ for $i = 1, \dots, q$. Put $k_i = l_i + 1$. Then

$$N^j u_i = N^{j-1} w_i \quad (j \geq 1), \qquad \text{so} \qquad N^{k_i} u_i = N^{l_i} w_i = 0.$$

Step 3. Fill up $\ker N$**.** The q vectors $N^{l_i} u_i = N^{l_i - 1} w_i$ $(i = 1, \dots, q)$ are members of the basis $\mathcal{C}$, hence linearly independent, and each lies in $\ker N$, because $N(N^{l_i} u_i) = N^{k_i} u_i = 0$. By Theorem 1.6.5 applied inside $\ker N$, there are vectors $u_{q+1}, \dots, u_r \in \ker N$ such that

$$(N^{l_1} u_1, \dots, N^{l_q} u_q,\ u_{q+1}, \dots, u_r)$$

is a basis of $\ker N$; in particular $r = \operatorname{nullity} N$. Put $k_i = 1$ for $q < i \leq r$, so that $N^{k_i} u_i = Nu_i = 0$ for those i as well.

Step 4. The combined list is a basis. Let $\mathcal{B}$ be the list of all $N^j u_i$ with $1 \leq i \leq r$ and $0 \leq j \leq k_i - 1$, the chains written one after another as in (a). Its length is

$$\sum_{i=1}^{q} (l_i + 1) + (r - q) = \dim W + q + (r - q) = \operatorname{rank} N + \operatorname{nullity} N = n$$

by Step 1 and Theorem 2.3.1. So by Theorem 1.6.4 (a) it is enough to prove independence.

Suppose

$$\sum_{i=1}^{r} \sum_{j=0}^{k_i - 1} a_{ij} N^j u_i = 0 \tag{$*$}$$

with $a_{ij} \in F$. Apply N to $(*)$. For $i > q$ we have $k_i = 1$ and $Nu_i = 0$, so those terms die. For $i \leq q$ and $j = k_i - 1 = l_i$ we get $N^{l_i + 1} u_i = 0$ by Step 2, so those terms die too. The rest is, using $N^{j+1} u_i = N^j w_i$,

$$\sum_{i=1}^{q} \sum_{j=0}^{l_i - 1} a_{ij} N^j w_i = 0.$$

This is a linear combination of the members of the basis $\mathcal{C}$ of W, so

$$a_{ij} = 0 \qquad \text{for } 1 \leq i \leq q,\ 0 \leq j \leq l_i - 1.$$

Substituting these zeros back into $(*)$ leaves only the terms with $i \leq q,\ j = l_i$ and those with $i > q,\ j = 0$:

$$\sum_{i=1}^{q} a_{i l_i} N^{l_i} u_i + \sum_{i=q+1}^{r} a_{i0} u_i = 0.$$

This is a linear combination of the members of the basis of $\ker N$ from Step 3, so all the remaining coefficients vanish as well. Hence $\mathcal{B}$ is independent, and therefore a basis of V. The lengths satisfy $\sum_i k_i = n$ by the count above, and reindexing the chains so that the lengths decrease — a permutation of the chains — arranges $k_1 \geq \cdots \geq k_r$. This proves (a), and completes the induction.

(b) Each C_i is N-invariant of dimension k_i, and writing $\mathcal{B}_{C_i} = (N^{k_i - 1} u_i, \dots, Nu_i, u_i)$ for its chain, $[N|_{C_i}]_{\mathcal{B}_{C_i}}^{\mathcal{B}_{C_i}} = \mathbf{J}_{k_i}(0)$ by Lemma 10.2.2 (b). The chains together form the basis $\mathcal{B}$ of V, so $V = C_1 + \cdots + C_r$. The sum is direct: if $c_1 + \cdots + c_r = 0$ with $c_i \in C_i$, then writing each c_i in terms of its own chain turns this into a linear combination of the

members of the basis $\mathcal{B}$ equal to 0, so every coefficient is 0 and every $c_i = 0$; now apply Theorem 1.7.5 ((b) $\Rightarrow$ (a)). Finally Theorem 9.1.5 gives that $[N]_{\mathcal{B}}^{\mathcal{B}}$ is block diagonal with blocks $[N|_{C_i}]_{\mathcal{B}_{C_i}}^{\mathcal{B}_{C_i}} = \mathbf{J}_{k_i}(0)$.

■

The theorem says that a nilpotent operator carries no information beyond the list of chain lengths. In matrix terms: **every nilpotent $\mathbf{A} \in M_n(F)$ is similar to $\mathbf{J}_{k_1}(0) \oplus \cdots \oplus \mathbf{J}_{k_r}(0)$ for some $k_1 \geq \cdots \geq k_r \geq 1$ summing to n** (apply the theorem to $T_{\mathbf{A}}$ and use Theorem 4.2.4). The index of nilpotency is k_1, the largest block (it is also the stabilization index of the kernel chain of Theorem 10.1.1, since $\ker N^{k_1} = V$): $\mathbf{J}_k(0)^k = \mathbf{O}$ and $\mathbf{J}_k(0)^{k-1} \neq \mathbf{O}$, and powers of a block diagonal matrix act blockwise (Theorem 8.1.4 (b)). Two questions remain: are the lengths unique, and how do we find them without finding the chains? One computation answers both.

10.2.4 Block sizes from the ranks of the powers

The lengths are visible in the ranks of the powers of N, which are similarity invariants. This is the practical algorithm and the uniqueness proof at once.

Theorem 10.2.4. Block Sizes from Ranks of Powers

Let V be finite-dimensional with $\dim V = n \geq 1$, let $N \in \mathcal{L}(V)$ be nilpotent, and suppose $\mathcal{B}$ is a Jordan basis for N with chain lengths $k_1 \geq \cdots \geq k_r \geq 1$, as in Theorem 10.2.3. Write $r_j = \operatorname{rank} N^j$ for $j \geq 0$, so $r_0 = n$. Then:

(a) for every integer $j \geq 0$,

$$\operatorname{rank} N^j = \sum_{i=1}^{r} \max(k_i - j,\, 0) \qquad \text{and} \qquad \dim \ker N^j = \sum_{i=1}^{r} \min(k_i,\, j);$$

(b) for every integer $j \geq 1$, the number of chains of length **at least** j is

$$\#\{\, i : k_i \geq j \,\} = r_{j-1} - r_j;$$

(c) for every integer $j \geq 1$, the number of chains of length **exactly** j is

$$\#\{\, i : k_i = j \,\} = r_{j-1} - 2r_j + r_{j+1};$$

(d) consequently the list $k_1 \geq \cdots \geq k_r$ is determined by N alone: any two Jordan bases for N have the same chain lengths. Moreover $r = n - \operatorname{rank} N = \operatorname{nullity} N$ and k_1 is the index of nilpotency of N.

Idea. Everything follows from what N^j does to the basis: it shifts each chain j boxes to the left, so the surviving images are the boxes at least j steps from the left edge. They are members of the basis, hence independent, and we can just count them: chain i contributes $k_i - j$ of them, or none if it is too short. Then (b) is a telescoping difference, because lengthening the shift by one step loses one box from every chain that still had one.

Proof.

(a) The basis $\mathcal{B}$ consists of the vectors $N^m u_i$ with $1 \le i \le r$ and $0 \le m \le k_i - 1$. Fix $j \ge 0$. Applying N^j to a member gives $N^j(N^m u_i) = N^{m+j} u_i$, which is 0 when $m + j \ge k_i$, because $N^{k_i} u_i = 0$, and is a member of $\mathcal{B}$, hence nonzero, when $m + j \le k_i - 1$. Since the image is spanned by the images of a basis (Theorem 2.2.4),

$$\operatorname{im} N^j = \operatorname{span}\left(N^m u_i : 1 \le i \le r,\ j \le m \le k_i - 1 \right).$$

This spanning list is a sublist of the basis $\mathcal{B}$, hence linearly independent, so it is a basis of $\operatorname{im} N^j$ and $\operatorname{rank} N^j$ is its length: for each i, the number of integers m with $j \le m \le k_i - 1$ is $k_i - j$ if $k_i > j$, and 0 otherwise, that is $\max(k_i - j, 0)$. This gives the first formula, and the second follows from Theorem 2.3.1 and $\sum_i k_i = n$:

$$\dim \ker N^j = n - \operatorname{rank} N^j = \sum_{i=1}^{r} \left(k_i - \max(k_i - j, 0)\right) = \sum_{i=1}^{r} \min(k_i, j).$$

(b) By (a), $r_{j-1} - r_j = \sum_i \left(\max(k_i - j + 1, 0) - \max(k_i - j, 0) \right)$. For a single i, the bracket is $(k_i - j + 1) - (k_i - j) = 1$ if $k_i \ge j$, and $0 - 0 = 0$ if $k_i \le j - 1$. Summing over i counts the indices with $k_i \ge j$.

(c) Subtract the count in (b) for $j+1$ from the count for j: a chain has length exactly j if and only if its length is at least j but not at least $j+1$. This gives $(r_{j-1} - r_j) - (r_j - r_{j+1})$.

(d) The numbers $r_j = \operatorname{rank} N^j$ are defined from N with no reference to a basis, so by (c) the number of chains of each length is determined by N; hence so is the list of lengths in decreasing order. Taking $j = 1$ in (b) gives $r = \#\{i : k_i \ge 1\} = r_0 - r_1 = n - \operatorname{rank} N$, which is nullity N by Theorem 2.3.1. Finally $N^{k_1} = 0$ and $N^{k_1 - 1} \ne 0$, since by (a) $\operatorname{rank} N^{k_1} = 0$ and $\operatorname{rank} N^{k_1 - 1} \ge k_1 - (k_1 - 1) = 1$; so k_1 is the index.

■

Corollary 10.2.5. Similarity of Nilpotent Matrices

Let $\mathbf{A}, \mathbf{B} \in M_n(F)$, $n \ge 1$, both nilpotent. Then $\mathbf{A} \sim \mathbf{B}$ if and only if

$$\operatorname{rank} \mathbf{A}^j = \operatorname{rank} \mathbf{B}^j \qquad \text{for every } j = 1, 2, \ldots, n - 1.$$

Equivalently, $\mathbf{A} \sim \mathbf{B}$ if and only if their chain lengths, as lists in decreasing order, agree.

Proof. ($\Rightarrow$) If $\mathbf{A} \sim \mathbf{B}$, say $\mathbf{B} = \mathbf{P}^{-1}\mathbf{A}\mathbf{P}$, then $\mathbf{B}^j = \mathbf{P}^{-1}\mathbf{A}^j\mathbf{P}$ (Proposition 4.2.6 (c)), so $\operatorname{rank} \mathbf{A}^j = \operatorname{rank} \mathbf{B}^j$ (Proposition 4.2.6 (a)).

($\Leftarrow$) Since $\operatorname{rank} T_{\mathbf{A}}^j = \operatorname{rank} \mathbf{A}^j$ and $\operatorname{rank} T_{\mathbf{B}}^j = \operatorname{rank} \mathbf{B}^j$ for every j (Theorem 4.1.5 (a)), the hypothesis is a statement about the operators. By Theorem 10.2.3 applied to $T_{\mathbf{A}}$ and to $T_{\mathbf{B}}$, together with Theorem 4.2.4, $\mathbf{A}$ is similar to $\mathbf{J}_{k_1}(0) \oplus \cdots \oplus \mathbf{J}_{k_r}(0)$ and $\mathbf{B}$ to $\mathbf{J}_{k'_1}(0) \oplus \cdots \oplus \mathbf{J}_{k'_{r'}}(0)$, with both lists decreasing. Equal ranks of all powers force equal lists, by Theorem 10.2.4 (c) (the ranks for $j \ge n$ are all 0, so the listed range suffices). Hence $\mathbf{A}$ and $\mathbf{B}$ are similar to the same matrix, and $\mathbf{A} \sim \mathbf{B}$ because similarity is an equivalence relation (Example (Similarity)).

■

Warning. **The successive differences count the blocks of size at least j, not the blocks of size exactly j; to get the exact counts, subtract twice.** Suppose $n = 5$ and $\operatorname{rank} N = 3$, $\operatorname{rank} N^2 = 1$, $\operatorname{rank} N^3 = 0$. Then $r_0 - r_1 = 2$ chains have length ≥ 1, $r_1 - r_2 = 2$ have length ≥ 2, and $r_2 - r_3 = 1$ has length ≥ 3. Reading these as exact counts would suggest two chains of length 1, two of length 2, and one of length 3, that is, lengths $1, 1, 2, 2, 3$, summing to 9 instead of 5. The correct counts are the second differences of Theorem 10.2.4 (c): exactly $2 - 2 = 0$ chains of length 1, exactly $2 - 1 = 1$ of length 2, and exactly $1 - 0 = 1$ of length 3. The partition is $(3, 2)$, and $3 + 2 = 5$ as it must be.

Quick check. A nilpotent N on a space of dimension 6 has $\operatorname{rank} N = 3$ and $\operatorname{rank} N^2 = 1$. What are its chain lengths?

Solution. Here $r_0 = 6$, $r_1 = 3$, and $r_2 = 1$, so there are $r_0 - r_1 = 3$ chains by part (b) of Theorem 10.2.4. Also $r_3 = 0$: a chain of length ≥ 4 would contribute at least 2 to $r_2 = \sum_i \max(k_i - 2, 0) = 1$, so every $k_i \leq 3$ and $r_3 = \sum_i \max(k_i - 3, 0) = 0$. By part (c) the exact counts are $r_0 - 2r_1 + r_2 = 1$ chain of length 1, $r_1 - 2r_2 + r_3 = 1$ of length 2, and $r_2 - 2r_3 + r_4 = 1$ of length 3. So the lengths are $(3, 2, 1)$, which indeed sum to 6.

10.2.5 Partitions and Young diagrams

The data left over is a list of decreasing positive integers with a given sum. Such lists have a name.

Definition 10.2.3. Partition of an Integer, Young Diagram, Dual Partition

Let $n \geq 1$. A **partition of** n is a finite list of integers

$$k_1 \geq k_2 \geq \cdots \geq k_r \geq 1 \qquad \text{with} \qquad k_1 + k_2 + \cdots + k_r = n.$$

Its **Young diagram** is the left-justified array of n boxes with k_i boxes in row i. The **dual partition** $k_1^* \geq k_2^* \geq \ldots$ is the list of **column** lengths of the diagram,

$$k_j^* = \#\{\, i : k_i \geq j \,\},$$

which is again a partition of n.

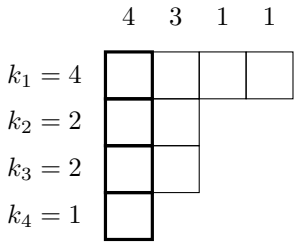
(The word *partition* is also used for a partition of a **set** into blocks (Definition 0.5.4) and for a partition of the row and column indices of a block matrix (Definition 8.1.1). Those are different notions; here the blocks are boxes in a diagram and the data is a list of numbers.)

For a nilpotent N on a space of dimension n, the chain lengths form a partition of n, and Theorem 10.2.4 (b) identifies the dual partition:

$$k_j^* = \#\{i : k_i \geq j\} = r_{j-1} - r_j = \dim \ker N^j - \dim \ker N^{j-1},$$

using Theorem 2.3.1 for the last equality. **The rows of the diagram are the chains; the columns are the jumps of the kernel chain.** In particular the first column is a basis of $\ker N$ (its boxes are the vectors $N^{k_i - 1} u_i$), the first two columns span $\ker N^2$, and so on. Here is

the diagram for the partition $(4,2,2,1)$ of 9, with the columns marked:



column lengths, above: the dual partition $(4,3,1,1)$
rows: the chain lengths $(4,2,2,1)$
the thick first column: a basis of $\ker N$
$\dim \ker N^j = 4, 7, 8, 9$ for $j = 1, 2, 3, 4$

Reading the diagram by rows gives the Jordan blocks; reading it by columns gives the dimensions $\dim \ker N^j = k_1^* + \cdots + k_j^*$. Transposing the diagram exchanges a partition with its dual.

Remark. For each n, the nilpotent operators on an n-dimensional space, up to similarity, correspond exactly to the partitions of n, and the correspondence does not depend on the field. For $n = 4$ the partitions are

$$(4), \quad (3,1), \quad (2,2), \quad (2,1,1), \quad (1,1,1,1),$$

so there are exactly five similarity classes of nilpotent 4×4 matrices over any field, with indices of nilpotency $4, 3, 2, 2, 1$.

10.2.6 Finding a Jordan basis

The proof of Theorem 10.2.3 is an algorithm, but run from the inside out. For small matrices it is easier to work from the longest chain down, using the rank data to know what to look for.

Algorithm. **Jordan basis of a nilpotent operator.**

1. Compute $r_j = \operatorname{rank} N^j$ until $r_j = 0$, and read off the chain lengths by Theorem 10.2.4 (c).
2. For each length in decreasing order, choose a top u with $N^{k-1}u \neq 0$, keeping the chains collected so far independent, and write down its chain.
3. Chains of length 1 are vectors completing a basis of $\ker N$ together with the ends $N^{k_i-1}u_i$ of the longer chains. Once n vectors are collected, Theorem 1.6.4 confirms a basis.

Example. A Jordan Basis for a 4×4 Nilpotent Matrix

Let

$$\mathbf{N} = \begin{pmatrix} 1 & 0 & 1 & 1 \\ 1 & -1 & 1 & 1 \\ 1 & -1 & 1 & 1 \\ -1 & 0 & -1 & -1 \end{pmatrix} \in M_4(\mathbb{Q}).$$

Show that $\mathbf{N}$ is nilpotent, find its chain lengths, and find a Jordan basis together with an invertible $\mathbf{P}$ with $\mathbf{P}^{-1}\mathbf{N}\mathbf{P}$ a direct sum of blocks $\mathbf{J}_k(0)$.

Solution. Nilpotency and ranks. Multiplying out,

$$\mathbf{N}^2 = \begin{pmatrix} 1 & -1 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ -1 & 1 & -1 & -1 \end{pmatrix}, \qquad \mathbf{N}^3 = \mathbf{O},$$

the last because every row of $\mathbf{N}^2$ is a multiple of $(1,-1,1,1)$, while

$$\begin{aligned}(1,-1,1,1)\,\mathbf{N} &= (1-1+1-1,\ \ 0+1-1+0,\ \ 1-1+1-1,\ \ 1-1+1-1) \\ &= \mathbf{0}.\end{aligned}$$

Since $\mathbf{N}^2 \neq \mathbf{O}$, the index is 3. Rows 2 and 3 of $\mathbf{N}$ are equal and row 4 = −row 1, so $\operatorname{rank}\mathbf{N} = 2$, and $\operatorname{rank}\mathbf{N}^2 = 1$. Thus

$$r_0 = 4, \quad r_1 = 2, \quad r_2 = 1, \quad r_3 = 0.$$

By Theorem 10.2.4 (c), the number of chains of length 1 is $r_0 - 2r_1 + r_2 = 4 - 4 + 1 = 1$, of length 2 is $r_1 - 2r_2 + r_3 = 2 - 2 + 0 = 0$, and of length 3 is $r_2 - 2r_3 + r_4 = 1 - 0 + 0 = 1$. The partition is $(3,1)$: one chain of length 3 and one of length 1, matching the picture drawn after Theorem 10.2.3.

The long chain. We need $\mathbf{u}_1 \notin \ker\mathbf{N}^2$. Try $\mathbf{u}_1 = \mathbf{e}_1$: the first column of $\mathbf{N}^2$ is $(1,0,0,-1) \neq \mathbf{0}$, so $\mathbf{N}^2\mathbf{e}_1 \neq \mathbf{0}$ and $\mathbf{e}_1$ works. Its chain, read off the first columns of $\mathbf{N}$ and $\mathbf{N}^2$, is

$$\mathbf{N}^2\mathbf{e}_1 = (1,0,0,-1), \qquad \mathbf{N}\mathbf{e}_1 = (1,1,1,-1), \qquad \mathbf{e}_1.$$

The short chain. $\ker\mathbf{N}$ has dimension 2 (Theorem 3.5.7). The equations $x_1 + x_3 + x_4 = 0$ (row 1) and $x_1 - x_2 + x_3 + x_4 = 0$ (row 2) give $x_2 = 0$ and $x_1 = -x_3 - x_4$, so

$$\ker\mathbf{N} = \operatorname{span}\big((-1,0,1,0),\ (-1,0,0,1)\big).$$

The end of the long chain, $\mathbf{N}^2\mathbf{e}_1 = (1,0,0,-1)$, is the second of these basis vectors with a minus sign, so we complete it to a basis of $\ker\mathbf{N}$ with $\mathbf{u}_2 = (-1,0,1,0)$.

The basis and the matrix. Take

$$\mathcal{B} = \big((1,0,0,-1),\ (1,1,1,-1),\ \mathbf{e}_1,\ (-1,0,1,0)\big),$$

$$\mathbf{P} = \begin{pmatrix} 1 & 1 & 1 & -1 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ -1 & -1 & 0 & 0 \end{pmatrix},$$

the matrix whose columns are the vectors of $\mathcal{B}$. Its third column has a single nonzero entry, the 1 in row 1, so expanding along that column (Theorem 7.6.1) gives

$$\det\mathbf{P} = \det\begin{pmatrix} 0 & 1 & 0 \\ 0 & 1 & 1 \\ -1 & -1 & 0 \end{pmatrix} = -1,$$

the last determinant computed by expanding along its own first column. So $\mathbf{P}$ is invertible and $\mathcal{B}$ is a basis. Then

$$\mathbf{P}^{-1}\mathbf{N}\mathbf{P} = \mathbf{J}_3(0) \oplus \mathbf{J}_1(0) = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

There is no need to invert $\mathbf{P}$ to check this: the claim $\mathbf{P}^{-1}N\mathbf{P} = \mathbf{J}$ is the same as $N\mathbf{P} = \mathbf{P}\mathbf{J}$, which reads column by column as

$$\begin{aligned} \mathbf{N}(1,0,0,-1) &= \mathbf{0}, \\ \mathbf{N}(1,1,1,-1) &= (1,0,0,-1), \\ \mathbf{N}\mathbf{e}_1 &= (1,1,1,-1), \\ \mathbf{N}(-1,0,1,0) &= \mathbf{0}, \end{aligned}$$

and these four statements are exactly how the chains were built (the first and last say that the two chain ends lie in $\ker \mathbf{N}$). As a sample verification, the second row of $\mathbf{N}(1,1,1,-1)$ is $1-1+1-1=0$, matching the second coordinate of $(1,0,0,-1)$.

10.2.7 Exercises

10.2.7.1 A. Check your understanding

Exercise. A1

(a) Define a nilpotent operator and its index of nilpotency.

(b) State the structure theorem for nilpotent operators, and say what a Jordan basis is.

(c) Write down the formula for the number of chains of length at least j in terms of the ranks of the powers.

(d) True or false: a singular matrix is nilpotent. Justify your answer.

(e) True or false: two nilpotent matrices in $M_n(F)$ with the same rank are similar. Justify your answer.

(f) How many similarity classes of nilpotent matrices are there in $M_3(F)$, and what are their indices of nilpotency?

Solution.

(a) $N \in \mathcal{L}(V)$ is nilpotent if $N^s = 0$ for **some** integer $s \geq 1$; the index is the least such s (Definition 10.2.1).

(b) For nilpotent N on V with $\dim V = n \geq 1$, there are chains generated by $u_1, \ldots, u_r$ of lengths $k_1 \geq \cdots \geq k_r \geq 1$ summing to n whose concatenation is a basis $\mathcal{B}$ of V — a Jordan basis — and $[N]_{\mathcal{B}}^{\mathcal{B}} = \mathbf{J}_{k_1}(0) \oplus \cdots \oplus \mathbf{J}_{k_r}(0)$ (Theorem 10.2.3).

(c) $\#\{i : k_i \geq j\} = \operatorname{rank} N^{j-1} - \operatorname{rank} N^j$ (Theorem 10.2.4 (b)).

(d) False. $\operatorname{diag}(0,1)$ is singular, and $\operatorname{diag}(0,1)^k = \operatorname{diag}(0,1) \neq \mathbf{O}$ for every $k \geq 1$. What is true is the converse: a nilpotent matrix on a nonzero space is singular, since $0 \in \operatorname{spec}(N)$ (Proposition 10.2.1 (c)).

(e) False for $n \geq 4$. The partitions $(3,1)$ and $(2,2)$ of 4 both give rank 2, but $\operatorname{rank} N^2$ is 1 and 0 respectively, so the matrices are not similar (Corollary 10.2.5). (For $n \leq 3$ the rank does determine the partition.)

(f) Three, one for each partition of 3: $(3), (2,1), (1,1,1)$, with indices $3, 2, 1$.

10.2.7.2 B. Practice

Exercise. B1: Block sizes from rank data

In each case N is nilpotent on a space of dimension n over F. Determine the chain lengths, or explain why no such operator exists.

(a) $n = 7$, $\operatorname{rank} N = 4$, $\operatorname{rank} N^2 = 2$, $\operatorname{rank} N^3 = 1$, $\operatorname{rank} N^4 = 0$.

(b) $n = 6$, $\operatorname{rank} N = 3$, $\operatorname{rank} N^2 = 1$, $\operatorname{rank} N^3 = 0$.

(c) $n = 6$, $\operatorname{rank} N = 4$, $\operatorname{rank} N^2 = 3$, $\operatorname{rank} N^3 = 0$.

Solution.

(a) With $r_0, \dots, r_4 = 7, 4, 2, 1, 0$, the exact counts from Theorem 10.2.4 (c) are: length 1: $7 - 8 + 2 = 1$; length 2: $4 - 4 + 1 = 1$; length 3: $2 - 2 + 0 = 0$; length 4: $1 - 0 + 0 = 1$. The lengths are $(4, 2, 1)$, and $4 + 2 + 1 = 7$.

(b) With $r_0, \dots, r_3 = 6, 3, 1, 0$: length 1: $6 - 6 + 1 = 1$; length 2: $3 - 2 + 0 = 1$; length 3: $1 - 0 + 0 = 1$. The lengths are $(3, 2, 1)$, summing to 6.

(c) No such operator exists. By (b) of the theorem, the number of chains of length ≥ 2 would be $r_1 - r_2 = 1$, while the number of length ≥ 3 would be $r_2 - r_3 = 3$, which is impossible: a chain of length ≥ 3 has length ≥ 2, so the counts must be nonincreasing in j. (Equivalently, Theorem 10.2.4 (c) would give $r_1 - 2r_2 + r_3 = 4 - 6 + 0 = -2$ chains of length 2, which is absurd.)

Exercise. B2: The five classes in dimension four

Let F be any field.

(a) List the five partitions of 4, and for each write down the corresponding nilpotent matrix in $M_4(F)$ as a direct sum of blocks $\mathbf{J}_k(0)$.

(b) For each, compute $\operatorname{rank} \mathbf{N}$, $\operatorname{rank} \mathbf{N}^2$, $\operatorname{rank} \mathbf{N}^3$, and the index of nilpotency.

(c) Which pairs among them share the same rank? Which invariant separates each such pair?

(d) Deduce that $\operatorname{rank} \mathbf{N}$ alone is not a complete invariant of nilpotent matrices, but the full list of ranks of powers is.

Solution.

(a) (4): $\mathbf{J}_4(0)$; $(3,1)$: $\mathbf{J}_3(0) \oplus \mathbf{J}_1(0)$; $(2,2)$: $\mathbf{J}_2(0) \oplus \mathbf{J}_2(0)$; $(2,1,1)$: $\mathbf{J}_2(0) \oplus \mathbf{J}_1(0) \oplus \mathbf{J}_1(0)$; $(1,1,1,1)$: the zero matrix.

(b) By Theorem 10.2.4 (a),

$$\operatorname{rank} \mathbf{N}^j = \sum_i \max(k_i - j, 0).$$

The ranks $(\operatorname{rank} \mathbf{N}, \operatorname{rank} \mathbf{N}^2, \operatorname{rank} \mathbf{N}^3)$ and the index are:

- (4): ranks $3, 2, 1$; index 4.
- $(3, 1)$: ranks $2, 1, 0$; index 3.
- $(2, 2)$: ranks $2, 0, 0$; index 2.
- $(2, 1, 1)$: ranks $1, 0, 0$; index 2.
- $(1, 1, 1, 1)$: ranks $0, 0, 0$; index 1.

(c) Only $(3, 1)$ and $(2, 2)$ share a rank, namely 2. They are separated by $\operatorname{rank} \mathbf{N}^2$ (1 against 0), equivalently by the index (3 against 2) or the minimal polynomial (x^3 against x^2, Proposition 10.2.1 (b)).

(d) The pair in (c) shows that equal rank does not imply similarity, while equality of $\operatorname{rank} \mathbf{N}^j$ for all j does (Corollary 10.2.5).

Exercise. B3: A Jordan basis by hand

Let

$$\mathbf{M} = \begin{pmatrix} -1 & -1 & 1 & -1 \\ 1 & 2 & -2 & 1 \\ 1 & 2 & -2 & 1 \\ 1 & 1 & -1 & 1 \end{pmatrix} \in M_4(\mathbb{Q}).$$

(a) Show that $\mathbf{M}^2 = \mathbf{O}$ and compute $\operatorname{rank} \mathbf{M}$.

(b) Determine the chain lengths of $\mathbf{M}$.

(c) Find a Jordan basis and an invertible $\mathbf{P}$ with $\mathbf{P}^{-1}\mathbf{M}\mathbf{P}$ a direct sum of blocks $\mathbf{J}_k(0)$. Verify your answer by checking $\mathbf{M}\mathbf{P} = \mathbf{P}\mathbf{J}$ column by column.

Hint: for (c), the tops of the chains may be taken among the standard basis vectors.

Solution.

(a) Write $\mathbf{m}_1, \dots, \mathbf{m}_4$ for the columns of $\mathbf{M}$, so $\mathbf{m}_1 = (-1, 1, 1, 1)$, $\mathbf{m}_2 = (-1, 2, 2, 1)$, $\mathbf{m}_3 = -\mathbf{m}_2$, and $\mathbf{m}_4 = \mathbf{m}_1$. The columns of $\mathbf{M}^2$ are $\mathbf{M}\mathbf{m}_1, \dots, \mathbf{M}\mathbf{m}_4$, so it is enough to check the first two. Row by row,

$$\begin{aligned} \mathbf{M}\mathbf{m}_1 &= (1 - 1 + 1 - 1,\ -1 + 2 - 2 + 1,\ -1 + 2 - 2 + 1,\ -1 + 1 - 1 + 1) \\ &= \mathbf{0}, \end{aligned}$$

$$\begin{aligned} \mathbf{M}\mathbf{m}_2 &= (1 - 2 + 2 - 1,\ -1 + 4 - 4 + 1,\ -1 + 4 - 4 + 1,\ -1 + 2 - 2 + 1) \\ &= \mathbf{0}. \end{aligned}$$

Hence $\mathbf{M}\mathbf{m}_3 = -\mathbf{M}\mathbf{m}_2 = \mathbf{0}$ and $\mathbf{M}\mathbf{m}_4 = \mathbf{M}\mathbf{m}_1 = \mathbf{0}$, so $\mathbf{M}^2 = \mathbf{O}$. Also $\operatorname{im} \mathbf{M} = \operatorname{span}(\mathbf{m}_1, \mathbf{m}_2)$ by the relations among the columns, and $\mathbf{m}_1, \mathbf{m}_2$ are not proportional, so $\operatorname{rank} \mathbf{M} = 2$.

(b) $r_0, r_1, r_2 = 4, 2, 0$. Chains of length 1: $r_0 - 2r_1 + r_2 = 4 - 4 + 0 = 0$; of length 2: $r_1 - 2r_2 + r_3 = 2$. So the partition is $(2, 2)$: two chains of length 2.

(c) We need two tops $\mathbf{u}, \mathbf{w}$ with $\mathbf{Mu}, \mathbf{Mw}$ independent; since $\operatorname{rank}\mathbf{M} = 2$ and $\operatorname{im}\mathbf{M} = \operatorname{span}(\mathbf{m}_1, \mathbf{m}_2)$, the vectors $\mathbf{u} = \mathbf{e}_1$ and $\mathbf{w} = \mathbf{e}_2$ do it, with $\mathbf{Me}_1 = \mathbf{m}_1 = (-1,1,1,1)$ and $\mathbf{Me}_2 = \mathbf{m}_2 = (-1,2,2,1)$. Take

$$\mathcal{B} = \big((-1,1,1,1),\ \mathbf{e}_1,\ (-1,2,2,1),\ \mathbf{e}_2\big), \qquad \mathbf{P} = \begin{pmatrix} -1 & 1 & -1 & 0 \\ 1 & 0 & 2 & 1 \\ 1 & 0 & 2 & 0 \\ 1 & 0 & 1 & 0 \end{pmatrix}.$$

The last column of $\mathbf{P}$ has a single nonzero entry, the 1 in row 2, so expanding along it (Theorem 7.6.1) gives

$$\det\mathbf{P} = \det\begin{pmatrix} -1 & 1 & -1 \\ 1 & 0 & 2 \\ 1 & 0 & 1 \end{pmatrix} = -1\cdot(1\cdot 1 - 2\cdot 1) = 1 \neq 0,$$

where the second determinant was expanded along its second column. So $\mathcal{B}$ is a basis, and

$$\mathbf{P}^{-1}\mathbf{MP} = \mathbf{J}_2(0)\oplus\mathbf{J}_2(0).$$

Column by column, $\mathbf{MP} = \mathbf{PJ}$ says: $\mathbf{M}(-1,1,1,1) = \mathbf{0}$ (first column of $\mathbf{PJ}$ is $\mathbf{0}$); $\mathbf{Me}_1 = (-1,1,1,1)$ (second column of $\mathbf{PJ}$ is the first column of $\mathbf{P}$); $\mathbf{M}(-1,2,2,1) = \mathbf{0}$; and $\mathbf{Me}_2 = (-1,2,2,1)$. All four were verified in (a) and (c).

10.2.7.3 C. Going deeper

Exercise. C1: A nilpotent matrix is similar to its transpose

Let $\mathbf{A} \in M_n(F)$ be nilpotent.

(a) Prove that $\mathbf{A}^\top$ is nilpotent and that $\operatorname{rank}(\mathbf{A}^\top)^j = \operatorname{rank}\mathbf{A}^j$ for every $j \geq 1$.

(b) Deduce that $\mathbf{A} \sim \mathbf{A}^\top$.

(c) For $\mathbf{A} = \mathbf{J}_3(0)$, find an explicit $\mathbf{P}$ with $\mathbf{P}^{-1}\mathbf{AP} = \mathbf{A}^\top$.

Solution.

(a) $(\mathbf{A}^\top)^j = (\mathbf{A}^j)^\top$ (Theorem 0.9.5), so $(\mathbf{A}^\top)^n = (\mathbf{A}^n)^\top = \mathbf{O}$ by Proposition 10.2.1 (a), and $\mathbf{A}^\top$ is nilpotent. A matrix and its transpose have the same rank (Theorem 3.5.5), so $\operatorname{rank}(\mathbf{A}^\top)^j = \operatorname{rank}(\mathbf{A}^j)^\top = \operatorname{rank}\mathbf{A}^j$.

(b) Both matrices are nilpotent with equal ranks of all powers, so Corollary 10.2.5 gives $\mathbf{A} \sim \mathbf{A}^\top$.

(c) $\mathbf{J}_3(0)$ sends $\mathbf{e}_1 \mapsto \mathbf{0}$, $\mathbf{e}_2 \mapsto \mathbf{e}_1$, $\mathbf{e}_3 \mapsto \mathbf{e}_2$, while $\mathbf{J}_3(0)^\top$ sends $\mathbf{e}_1 \mapsto \mathbf{e}_2 \mapsto \mathbf{e}_3 \mapsto \mathbf{0}$, that is, it is the shift in the reversed basis. So take the reversal permutation matrix

$$\mathbf{P} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \qquad \mathbf{P}^{-1} = \mathbf{P}.$$

Then $\mathbf{P}^{-1}\mathbf{J}_3(0)\mathbf{P} = \mathbf{J}_3(0)^\top$, as one checks column by column: $\mathbf{J}_3(0)\mathbf{P}$ has columns

$\mathbf{J}_3(0)\mathbf{e}_3 = \mathbf{e}_2$, $\mathbf{J}_3(0)\mathbf{e}_2 = \mathbf{e}_1$, $\mathbf{J}_3(0)\mathbf{e}_1 = \mathbf{0}$, and $\mathbf{P}\,\mathbf{J}_3(0)^{\top}$ has columns $\mathbf{P}\mathbf{e}_2 = \mathbf{e}_2$, $\mathbf{P}\mathbf{e}_3 = \mathbf{e}_1$, $\mathbf{P}\mathbf{0} = \mathbf{0}$. (The same reversal works for every $\mathbf{J}_k(0)$, and hence, blockwise, for every nilpotent matrix in Jordan form.)

Exercise. C2: One long chain

Let V be finite-dimensional with $\dim V = n \geq 1$, and let $N \in \mathcal{L}(V)$ be nilpotent with $N^{n-1} \neq 0$.

(a) Prove that there is a basis of V in which the matrix of N is $\mathbf{J}_n(0)$.

(b) Deduce that $m_N = p_N = x^n$, and that nullity $N = 1$.

(c) Show that the conclusion of (a) fails for $\mathbf{N} = \mathbf{J}_2(0) \oplus \mathbf{J}_1(0)$ on a space of dimension 3, and say which hypothesis fails.

Hint: for (a), take a vector outside $\ker N^{n-1}$*, which exists precisely because* $N^{n-1} \neq 0$.

Solution.

(a) Since $N^{n-1} \neq 0$, there is $u \in V$ with $N^{n-1}u \neq 0$, while $N^n u = 0$ by Proposition 10.2.1 (a). So u generates a Jordan chain of length n (Definition 10.2.2), which is a linearly independent list of $n = \dim V$ vectors (Lemma 10.2.2 (a)), hence a basis of V (Theorem 1.6.4 (a)). In it the matrix of N is $\mathbf{J}_n(0)$, by Lemma 10.2.2 (b) with $C = V$.

(b) In that basis $\operatorname{rank} N^j = n - j$ for $0 \leq j \leq n$ (Theorem 10.2.4 (a) with the single length $k_1 = n$), so nullity $N = 1$ by Theorem 2.3.1 and the index of nilpotency is n. Hence $m_N = x^n$ by Proposition 10.2.1 (b), and $p_N = x^n$ as well; they coincide, which by Corollary 9.6.2 is the extreme case $\deg m_N = n$.

(c) For $\mathbf{N} = \mathbf{J}_2(0) \oplus \mathbf{J}_1(0)$ on F^3 we have $\mathbf{N}^2 = \mathbf{O}$, so $\mathbf{N}^{n-1} = \mathbf{N}^2 = \mathbf{O}$ and the hypothesis $\mathbf{N}^{n-1} \neq \mathbf{O}$ fails. Accordingly $\mathbf{N}$ is not similar to $\mathbf{J}_3(0)$: its rank is 1, while $\operatorname{rank} \mathbf{J}_3(0) = 2$, and similar matrices have equal ranks (Proposition 4.2.6 (a)). Its partition is $(2, 1)$, not (3).

10.3 The Jordan Canonical Form

Chapter 4 asked how simple the matrix of an operator can be made by choosing the basis well, and Chapter 9 answered it only for operators with enough eigenvectors. The two previous sections have quietly finished the job: the space splits into generalized eigenspaces, and on each of them the operator is a scalar plus a nilpotent operator, which is a disjoint union of chains. Assembling these pieces gives the best matrix there is, and it is unique. This section states that theorem, proves it, reads the familiar invariants off it, and turns it into a similarity test.

10.3.1 The shape we are aiming for

Here is the goal, stated before its parts are named. **If p_T splits, then V has a basis in which the matrix of T is diagonal except for some ones immediately above the diagonal, and that matrix is unique up to the order of its diagonal blocks.** A diagonal matrix is the special case with no ones at all. So the dream of Chapter 4 comes true, with one extra ingredient beyond a diagonal matrix: the superdiagonal ones, one for each step of a Jordan

chain.

The building block is the matrix of $\lambda\,\mathrm{id}+N$ on a single chain. By Lemma 10.2.2 (b), the nilpotent part contributes $\mathbf{J}_k(0)$, so the block is $\lambda\mathbf{I}_k+\mathbf{J}_k(0)$.

Definition 10.3.1. Jordan Block

Let $k\geq 1$ and $\lambda\in F$. The **Jordan block** of size k for λ is the matrix

$$\mathbf{J}_k(\lambda)\coloneqq\lambda\mathbf{I}_k+\mathbf{J}_k(0)=\begin{pmatrix}\lambda & 1 & & \\ & \lambda & \ddots & \\ & & \ddots & 1\\ & & & \lambda\end{pmatrix}\in M_k(F),$$

with λ in every diagonal entry, 1 in every **superdiagonal** entry $(j,j+1)$, and 0 elsewhere. In particular $\mathbf{J}_1(\lambda)=(\lambda)$.

Reading its columns: if $[T]_{\mathcal{B}}^{\mathcal{B}}=\mathbf{J}_k(\lambda)$ for a basis $\mathcal{B}=(b_1,\dots,b_k)$, then $Tb_1=\lambda b_1$ and $Tb_j=b_{j-1}+\lambda b_j$ for $j\geq 2$; equivalently $(T-\lambda\,\mathrm{id}_V)b_j=b_{j-1}$ and $(T-\lambda\,\mathrm{id}_V)b_1=0$, so $\mathcal{B}$ is the Jordan chain generated by b_k (Definition 10.2.2). The single eigenvalue is λ, with $a(\lambda)=k$ and $g(\lambda)=1$, since $\mathbf{J}_k(\lambda)-\lambda\mathbf{I}_k=\mathbf{J}_k(0)$ has rank $k-1$.

A Jordan matrix is a diagonal matrix with some chains glued in: blocks along the diagonal, each block one eigenvalue with ones above it.

Definition 10.3.2. Jordan Matrix, Jordan Basis, Jordan Form

A **Jordan matrix** is a matrix of the form

$$\mathbf{J}=\mathbf{J}_{k_1}(\lambda_1)\oplus\mathbf{J}_{k_2}(\lambda_2)\oplus\cdots\oplus\mathbf{J}_{k_m}(\lambda_m)$$

for some $m\geq 1$, sizes $k_i\geq 1$, and scalars $\lambda_i\in F$, **not** assumed distinct. Let V be a finite-dimensional vector space over F with $\dim V\geq 1$ and $T\in\mathcal{L}(V)$. A **Jordan form** of T is a Jordan matrix $\mathbf{J}$ with $[T]_{\mathcal{B}}^{\mathcal{B}}=\mathbf{J}$ for **some** basis $\mathcal{B}$ of V; such a basis is called a **Jordan basis** for T. For $\mathbf{A}\in M_n(F)$, a **Jordan form** of $\mathbf{A}$ is a Jordan matrix $\mathbf{J}$ with $\mathbf{A}\sim\mathbf{J}$.

In words: the matrix is block diagonal (Definition 8.1.2), and each diagonal block is a single Jordan block. Different blocks may carry the same eigenvalue or different ones, and the sizes may repeat. A Jordan basis is a basis that is a concatenation of Jordan chains, one chain per block. The two descriptions match for $\mathbf{A}\in M_n(F)$ and $T=T_{\mathbf{A}}$, by Theorem 4.2.4.

The order of the blocks never matters. Reordering the blocks of $\mathbf{J}$ permutes the segments of the basis $\mathcal{B}$, so the reordered matrix is the matrix of the **same** operator in the permuted basis, hence similar to $\mathbf{J}$ (Theorem 4.2.4 (a)). This is why the uniqueness statement below can only be "up to the order of the blocks", and why it costs nothing to list the blocks in any convenient order.

Examples.

- **Diagonal matrices.** $\mathrm{diag}(d_1,\dots,d_n)=\mathbf{J}_1(d_1)\oplus\cdots\oplus\mathbf{J}_1(d_n)$ is a Jordan matrix with all blocks of size 1. So every diagonalizable operator has a Jordan form, namely a diagonal one.

- **The standing counterexample.** $\begin{pmatrix} 2 & 1 \\ 0 & 2 \end{pmatrix} = \mathbf{J}_2(2)$ is its own Jordan form, and $(\mathbf{e}_1, \mathbf{e}_2)$ is a Jordan basis: $\mathbf{e}_1$ is an eigenvector and $(\mathbf{J} - 2\mathbf{I})\mathbf{e}_2 = \mathbf{e}_1$.
- **Nilpotent operators.** By Theorem 10.2.3, a nilpotent operator has the Jordan form $\mathbf{J}_{k_1}(0) \oplus \cdots \oplus \mathbf{J}_{k_r}(0)$, all eigenvalues 0. Section 2 is exactly the case $\lambda = 0$ of this section.
- **Mixed blocks.** $\mathbf{J} = \mathbf{J}_2(3) \oplus \mathbf{J}_1(3) \oplus \mathbf{J}_2(1) \in M_5(F)$ has $p_{\mathbf{J}} = (x-3)^3(x-1)^2$, $g(3) = 2$, and $g(1) = 1$. It is the answer to the worked example at the end of this section.
- **The degenerate case.** On a space of dimension 1, every operator is $(\lambda) = \mathbf{J}_1(\lambda)$. One block, size one, no superdiagonal to speak of.

Nonexample by minimal change. The matrix $\begin{pmatrix} 2 & 1 \\ 0 & 3 \end{pmatrix}$ is upper triangular with a superdiagonal 1, but it is **not** a Jordan matrix: a 1 above the diagonal may only join two **equal** diagonal entries, because it records the step of a chain for a single eigenvalue. (Indeed this matrix has distinct eigenvalues, so it is diagonalizable, and its Jordan form is $\operatorname{diag}(2, 3)$.) Changing the superdiagonal entry instead, $\begin{pmatrix} 2 & 2 \\ 0 & 2 \end{pmatrix}$ is not a Jordan matrix either, since the entry above the diagonal must be 1; it is similar to $\mathbf{J}_2(2)$, by rescaling the second basis vector.

Why this definition. Two conventions are being fixed. First, the ones sit **above** the diagonal, which matches the chain order "eigenvector first" of Definition 10.2.2; putting them below the diagonal gives the transposed matrix, which is just as good and appears in some books. Second, the off-diagonal entries are 1 rather than arbitrary nonzero scalars: a nonzero entry c can be turned into a 1 by scaling one basis vector, so nothing is lost, and the normalization is what makes the form unique. Nothing else is allowed: if any further entry were permitted, the form would no longer be determined by the operator.

Warning. **A Jordan form exists only when p_T splits, so it depends on the field.** The rotation $\mathbf{R} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \in M_2(\mathbb{R})$ has no Jordan form over $\mathbb{R}$, since $p_{\mathbf{R}} = x^2 + 1$ has no real root, and a Jordan matrix is triangular with its eigenvalues on the diagonal. Over $\mathbb{C}$ its Jordan form is $\operatorname{diag}(i, -i)$. Over $\mathbb{Q}$ the matrix $\begin{pmatrix} 0 & 2 \\ 1 & 0 \end{pmatrix}$ has no Jordan form either, and over $\mathbb{R}$ it becomes $\operatorname{diag}(\sqrt{2}, -\sqrt{2})$. Section 6 gives a canonical form valid over **every** field.

10.3.2 The theorem

Every operator whose characteristic polynomial splits is a direct sum of Jordan blocks, in exactly one way.

Theorem 10.3.1. Jordan Canonical Form Theorem

Let V be a finite-dimensional vector space over F with $\dim V = n \geq 1$, and let $T \in \mathcal{L}(V)$.

(a) **(Existence)** T has a Jordan form if and only if p_T splits over F.

(b) **(Uniqueness)** Suppose $\mathbf{J}$ is a Jordan form of T. Then for every $\lambda \in F$ and every integer $j \geq 1$, the number of blocks $\mathbf{J}_k(\lambda)$ in $\mathbf{J}$ with $k \geq j$ equals

$$\operatorname{rank}(T - \lambda \operatorname{id}_V)^{j-1} - \operatorname{rank}(T - \lambda \operatorname{id}_V)^{j},$$

a number depending only on T, λ, and j. Consequently any two Jordan forms of T have the same blocks, up to the order in which they are listed.

Equivalently, for $\mathbf{A} \in M_n(F)$ with $p_{\mathbf{A}}$ split: $\mathbf{A}$ is similar to a Jordan matrix, and that matrix is unique up to the order of its blocks. We call it **the** Jordan form of $\mathbf{A}$.

Idea. Existence is an assembly of the two previous sections, in three steps. ① Split $V = \bigoplus_i G_{\lambda_i}(T)$ (Theorem 10.1.3). ② On G_{λ_i} write $T = \lambda_i \,\mathrm{id} + N_i$ with N_i nilpotent, and take a Jordan basis for N_i (Theorem 10.2.3); adding $\lambda_i \mathbf{I}$ turns each $\mathbf{J}_k(0)$ into $\mathbf{J}_k(\lambda_i)$. ③ Put the bases together: the matrix is block diagonal. For uniqueness, count ranks. A block $\mathbf{J}_k(\mu) - \lambda\mathbf{I}$ with $\mu \neq \lambda$ is invertible, so it contributes its full size to every power and drops out of differences; a block with $\mu = \lambda$ becomes $\mathbf{J}_k(0)$, whose j-th power has rank $\max(k-j, 0)$. So the difference of two consecutive ranks counts the λ-blocks that are still long enough.

Proof.

(a) ($\Leftarrow$) Suppose p_T splits, with distinct eigenvalues $\lambda_1, \dots, \lambda_d$.
Step 1. Split the space. By Theorem 10.1.3, $V = G_1 \oplus \cdots \oplus G_d$ with $G_i = G_{\lambda_i}(T)$ nonzero and T-invariant, $\dim G_i = a_i := a_T(\lambda_i)$, and

$$T|_{G_i} = \lambda_i \,\mathrm{id}_{G_i} + N_i, \qquad N_i = (T - \lambda_i \,\mathrm{id}_V)|_{G_i} \text{ nilpotent.}$$

Step 2. One eigenvalue at a time. Fix i. By Theorem 10.2.3 applied to $N_i \in \mathcal{L}(G_i)$, there is a basis $\mathcal{B}_i$ of G_i, a concatenation of Jordan chains of N_i, with

$$[N_i]_{\mathcal{B}_i}^{\mathcal{B}_i} = \mathbf{J}_{k_{i1}}(0) \oplus \cdots \oplus \mathbf{J}_{k_{ir_i}}(0), \qquad \textstyle\sum_l k_{il} = a_i.$$

Since $[\,\cdot\,]_{\mathcal{B}_i}^{\mathcal{B}_i}$ is linear (Theorem 4.1.4) and $[\mathrm{id}_{G_i}]_{\mathcal{B}_i}^{\mathcal{B}_i} = \mathbf{I}_{a_i} = \mathbf{I}_{k_{i1}} \oplus \cdots \oplus \mathbf{I}_{k_{ir_i}}$,

$$\begin{aligned} [T|_{G_i}]_{\mathcal{B}_i}^{\mathcal{B}_i} &= \lambda_i \mathbf{I}_{a_i} + [N_i]_{\mathcal{B}_i}^{\mathcal{B}_i} \\ &= (\lambda_i \mathbf{I}_{k_{i1}} + \mathbf{J}_{k_{i1}}(0)) \oplus \cdots \oplus (\lambda_i \mathbf{I}_{k_{ir_i}} + \mathbf{J}_{k_{ir_i}}(0)) \\ &= \mathbf{J}_{k_{i1}}(\lambda_i) \oplus \cdots \oplus \mathbf{J}_{k_{ir_i}}(\lambda_i), \end{aligned}$$

where the middle equality is the blockwise addition of Theorem 8.1.4 (a) and the last is Definition 10.3.1.
Step 3. Assemble. Let $\mathcal{B}$ be the basis of V obtained by writing $\mathcal{B}_1, \dots, \mathcal{B}_d$ one after another. By Theorem 9.1.5,

$$[T]_{\mathcal{B}}^{\mathcal{B}} = [T|_{G_1}]_{\mathcal{B}_1}^{\mathcal{B}_1} \oplus \cdots \oplus [T|_{G_d}]_{\mathcal{B}_d}^{\mathcal{B}_d},$$

which by Step 2 is a direct sum of Jordan blocks, that is, a Jordan matrix. So T has a Jordan form, and $\mathcal{B}$ is a Jordan basis.
($\Rightarrow$) Suppose $[T]_{\mathcal{B}}^{\mathcal{B}} = \mathbf{J}$ is a Jordan matrix. Each Jordan block is upper triangular, so $\mathbf{J}$ is upper triangular, and T is triangularizable (Definition 9.7.1). By Theorem 9.7.2, p_T splits over F.

(b) Let $\mathbf{J} = \mathbf{J}_{k_1}(\mu_1) \oplus \cdots \oplus \mathbf{J}_{k_m}(\mu_m) = [T]_{\mathcal{B}}^{\mathcal{B}}$, with $\mu_1, \dots, \mu_m \in F$ not necessarily distinct; fix $\lambda \in F$ and let $j \geq 0$. Then $[T - \lambda \,\mathrm{id}_V]_{\mathcal{B}}^{\mathcal{B}} = \mathbf{J} - \lambda\mathbf{I}_n$, which is block diagonal with

blocks $\mathbf{J}_{k_i}(\mu_i) - \lambda \mathbf{I}_{k_i}$, so by Theorem 8.1.4 (b),

$$[(T - \lambda \operatorname{id}_V)^j]_{\mathcal{B}}^{\mathcal{B}} = (\mathbf{J}_{k_1}(\mu_1) - \lambda \mathbf{I}_{k_1})^j \oplus \cdots \oplus (\mathbf{J}_{k_m}(\mu_m) - \lambda \mathbf{I}_{k_m})^j,$$

and the rank of a block diagonal matrix is the sum of the ranks of its blocks (Theorem 8.2.1 (b), applied $m - 1$ times). Also $\operatorname{rank}(T - \lambda \operatorname{id}_V)^j$ equals the rank of this matrix (Theorem 4.1.5 (a)). Now compute block by block.

Blocks with $\mu_i \neq \lambda$. Here $\mathbf{J}_{k_i}(\mu_i) - \lambda \mathbf{I}_{k_i}$ is upper triangular with every diagonal entry equal to $\mu_i - \lambda \neq 0$, so its determinant is $(\mu_i - \lambda)^{k_i} \neq 0$ (Theorem 7.4.6) and it is invertible (Theorem 7.5.2). Hence its j-th power is invertible too, of rank k_i, for every $j \geq 0$.

Blocks with $\mu_i = \lambda$. Here $\mathbf{J}_{k_i}(\lambda) - \lambda \mathbf{I}_{k_i} = \mathbf{J}_{k_i}(0)$, whose standard basis is a single Jordan chain of length k_i, so $\operatorname{rank} \mathbf{J}_{k_i}(0)^j = \max(k_i - j, 0)$ by Theorem 10.2.4 (a) with the one-element list of lengths (k_i).

Adding up, with $c = \sum_{\mu_i \neq \lambda} k_i$ independent of j,

$$\operatorname{rank}(T - \lambda \operatorname{id}_V)^j = c + \sum_{\mu_i = \lambda} \max(k_i - j,\, 0).$$

Subtracting the values at j and $j - 1$, the constant c cancels and each remaining bracket $\max(k_i - j + 1, 0) - \max(k_i - j, 0)$ equals 1 if $k_i \geq j$ and 0 otherwise, exactly as in the proof of Theorem 10.2.4 (b). Therefore

$$\operatorname{rank}(T - \lambda \operatorname{id}_V)^{j-1} - \operatorname{rank}(T - \lambda \operatorname{id}_V)^j = \#\{\, i : \mu_i = \lambda \text{ and } k_i \geq j \,\}.$$

The left-hand side is defined from T alone. So for each λ and each j the number of λ-blocks of size at least j, and hence by subtraction the number of size exactly j, is the same for every Jordan form of T. Two Jordan forms therefore consist of the same blocks with the same multiplicities, and differ only in the order in which the blocks are listed.

For a matrix $\mathbf{A} \in M_n(F)$ with $p_{\mathbf{A}}$ split, apply all of this to $T_{\mathbf{A}}$, using Theorem 4.2.4 to pass between bases and similarity.

■

Step 1 → Step 2 → Step 3, and the dream comes true: **for every complex square matrix there is a basis in which the matrix is diagonal apart from some ones on the superdiagonal**, since $p_{\mathbf{A}}$ always splits over $\mathbb{C}$ (Corollary 6.5.5). The proof also tells us where the basis comes from: one chain per block, found inside one generalized eigenspace at a time.

Quick check. How many Jordan forms, up to the order of the blocks, are possible for an operator with $p_T = (x - 2)^2(x - 3)$? Write them down, and say which invariant tells them apart.

Solution. The 3-blocks must have sizes summing to $a(3) = 1$, so there is exactly one, of size 1. The 2-blocks have sizes summing to $a(2) = 2$, so they form a partition of 2: either (2) or $(1, 1)$. The two possibilities are

$$\mathbf{J}_2(2) \oplus \mathbf{J}_1(3) \qquad \text{and} \qquad \mathbf{J}_1(2) \oplus \mathbf{J}_1(2) \oplus \mathbf{J}_1(3) = \operatorname{diag}(2, 2, 3).$$

They are told apart by $g(2)$, which is the number of 2-blocks: 1 in the first case, 2 in the second. Equivalently by $m_T = (x-2)^2(x-3)$ against $(x-2)(x-3)$, or by $\operatorname{rank}(T - 2\,\mathrm{id})$, which is 2 against 1.

10.3.3 Reading the invariants off the blocks

The Jordan form makes every invariant of Chapter 9 visible at a glance.

Corollary 10.3.2. Invariants from the Jordan Form

Let V be finite-dimensional with $\dim V = n \geq 1$, let $T \in \mathcal{L}(V)$ with p_T split, let $\mathbf{J}$ be a Jordan form of T, and let $\lambda \in F$. Then:

(a) $a_T(\lambda)$ is the **sum of the sizes** of the λ-blocks of $\mathbf{J}$; in particular $\lambda \in \operatorname{spec}(T)$ if and only if $\mathbf{J}$ has a λ-block;

(b) $g_T(\lambda)$ is the **number** of λ-blocks of $\mathbf{J}$;

(c) the exponent of $x - \lambda$ in m_T is the **largest size** of a λ-block of $\mathbf{J}$ (and 0 when $\mathbf{J}$ has no λ-block);

(d) $\dim \ker(T - \lambda\,\mathrm{id}_V)^j = \sum \min(k_i, j)$, summed over the sizes k_i of the λ-blocks, for every $j \geq 0$;

(e) T is diagonalizable if and only if every block of $\mathbf{J}$ has size 1, that is, if and only if $\mathbf{J}$ is diagonal.

Proof. Write $\mathbf{J} = \mathbf{J}_{k_1}(\mu_1) \oplus \cdots \oplus \mathbf{J}_{k_m}(\mu_m)$, with $\mu_i \in F$ not necessarily distinct, and $\mathcal{B}$ for a Jordan basis, so $[T]_{\mathcal{B}}^{\mathcal{B}} = \mathbf{J}$.

(a) $\mathbf{J}$ is upper triangular with diagonal entries μ_i repeated k_i times, so $p_T = \prod_{i=1}^{m}(x-\mu_i)^{k_i}$ (Theorem 7.4.6, Definition 7.10.2). Grouping equal μ_i and using that the multiplicity of λ as a root of a product of powers of distinct linear factors is the exponent it carries (Lemma 6.4.1), $a_T(\lambda) = \sum_{\mu_i = \lambda} k_i$. The last claim follows since $\operatorname{spec}(T)$ is the set of roots of p_T (Theorem 9.2.1).

(b) Take $j = 1$ in Theorem 10.3.1 (b): the number of λ-blocks of size ≥ 1, which is all of them, equals

$$\begin{aligned} \operatorname{rank}(T - \lambda\,\mathrm{id}_V)^0 - \operatorname{rank}(T - \lambda\,\mathrm{id}_V) &= n - \operatorname{rank}(T - \lambda\,\mathrm{id}_V) \\ &= \dim\ker(T - \lambda\,\mathrm{id}_V) = g_T(\lambda), \end{aligned}$$

by Theorem 2.3.1 and Definition 9.3.2.

(c) First, $m_{\mathbf{J}_k(\nu)} = (x-\nu)^k$ for every $\nu \in F$: the polynomial $(x-\nu)^k$ annihilates $\mathbf{J}_k(\nu)$ because $\mathbf{J}_k(\nu) - \nu\mathbf{I} = \mathbf{J}_k(0)$ and $\mathbf{J}_k(0)^k = \mathbf{O}$, while $\mathbf{J}_k(0)^{k-1} \neq \mathbf{O}$, so no smaller power of $x - \nu$ annihilates it; and $m_{\mathbf{J}_k(\nu)}$ divides $p_{\mathbf{J}_k(\nu)} = (x-\nu)^k$ (Corollary 9.6.2), hence is a power of $x - \nu$ (Lemma 9.5.2). Applying Proposition 9.5.6 $m-1$ times,

$$m_T = m_{\mathbf{J}} = \operatorname{lcm}\left((x-\mu_1)^{k_1}, \ldots, (x-\mu_m)^{k_m}\right) = \prod_{\nu \in \operatorname{spec}(T)} (x-\nu)^{\max\{k_i \,:\, \mu_i = \nu\}},$$

the last equality because the least common multiple of powers of linear factors takes, for each factor, the largest exponent occurring (Definition 6.2.4, Theorem 6.3.5).

(d) The proof of Theorem 10.3.1 (b) computed

$$\operatorname{rank}(T-\lambda\,\mathrm{id}_V)^j = c + \sum_{\mu_i=\lambda} \max(k_i - j,\,0), \qquad c \coloneqq \sum_{\mu_i\neq\lambda} k_i.$$

Since $\sum_{i=1}^{m} k_i = n$, we have $n - c = \sum_{\mu_i=\lambda} k_i$, so by Theorem 2.3.1

$$\begin{aligned}\dim\ker(T-\lambda\,\mathrm{id}_V)^j &= n - c - \sum_{\mu_i=\lambda} \max(k_i - j, 0)\\ &= \sum_{\mu_i=\lambda} (k_i - \max(k_i - j, 0))\\ &= \sum_{\mu_i=\lambda} \min(k_i, j).\end{aligned}$$

(e) If every block has size 1, then $\mathbf{J}$ is diagonal and T is diagonalizable (Definition 9.4.1). Conversely, if T is diagonalizable, then m_T is a product of distinct linear factors (Theorem 9.8.2), so every exponent in (c) is 1, which says that every block has size 1.

■

Parts (a) and (b) give the practical reading: **the total size of the λ-blocks is $a(\lambda)$, and their number is $g(\lambda)$.** The gap $a(\lambda) - g(\lambda)$, which Chapter 9 could only measure, is now explained: it counts the superdiagonal ones belonging to λ.

Warning. **The multiplicities $a(\lambda)$ and $g(\lambda)$ do not always determine the blocks.** For $n = 7$ and one eigenvalue λ with $a(\lambda) = 7$ and $g(\lambda) = 3$, the block sizes form a partition of 7 into 3 parts, and there are four of those: $(5,1,1)$, $(4,2,1)$, $(3,3,1)$, $(3,2,2)$. Knowing $a(\lambda)$ and $g(\lambda)$ means knowing $\dim\ker(T-\lambda\,\mathrm{id}_V)^j$ for $j = 1$ and for j large; the intermediate values $j = 2, 3, \ldots$ are genuinely new information, and Theorem 10.3.1 (b) needs them all.

10.3.4 Deciding similarity

We can now settle the question Chapter 4 raised and Chapter 7 sharpened: which data decide whether two matrices are similar?

Corollary 10.3.3. Similarity via the Jordan Form

Let $\mathbf{A}, \mathbf{B} \in M_n(F)$, $n \geq 1$, and suppose $p_{\mathbf{A}}$ splits over F. The following are equivalent:

(a) $\mathbf{A} \sim \mathbf{B}$;

(b) $p_{\mathbf{B}}$ splits over F, and $\mathbf{A}$ and $\mathbf{B}$ have the same Jordan form, up to the order of the blocks;

(c) $p_{\mathbf{A}} = p_{\mathbf{B}}$ and $\operatorname{rank}(\mathbf{A}-\lambda\mathbf{I})^j = \operatorname{rank}(\mathbf{B}-\lambda\mathbf{I})^j$ for every $\lambda \in \operatorname{spec}(\mathbf{A})$ and every $j = 1, \ldots, n$.

Proof.

(a) $\Rightarrow$ (c). Similar matrices have the same characteristic polynomial (Theorem 7.10.3). If $\mathbf{B} = \mathbf{P}^{-1}\mathbf{A}\mathbf{P}$, then $(\mathbf{B}-\lambda\mathbf{I})^j = \mathbf{P}^{-1}(\mathbf{A}-\lambda\mathbf{I})^j\mathbf{P}$ for every λ and j (Proposition 4.2.6

(c) applied to the polynomial $(x-\lambda)^j$), so the ranks agree (Proposition 4.2.6 (a)).

(b) $\Rightarrow$ (b). From $p_{\mathbf{B}} = p_{\mathbf{A}}$ we get that $p_{\mathbf{B}}$ splits and that $\operatorname{spec}(\mathbf{B}) = \operatorname{spec}(\mathbf{A})$ (Theorem 9.2.1). By Theorem 10.3.1 (a) both matrices have Jordan forms, and by (b) of that theorem the number of λ-blocks of each size in each of them is determined by the ranks $\operatorname{rank}(\cdot - \lambda\mathbf{I})^j$, with $j \le n$ sufficing because $(\mathbf{A}-\lambda\mathbf{I})^n$ and all higher powers have the same rank (Theorem 10.1.1 (d), Theorem 2.3.1). These numbers agree by hypothesis for $\lambda \in \operatorname{spec}(\mathbf{A}) = \operatorname{spec}(\mathbf{B})$, and for $\lambda \notin \operatorname{spec}(\mathbf{A})$ neither form has a λ-block (Corollary 10.3.2 (a)). So the two Jordan forms have the same blocks.

(c) $\Rightarrow$ (a). Let $\mathbf{J}$ be a Jordan form of $\mathbf{A}$ and $\mathbf{J}'$ one of $\mathbf{B}$, with the same blocks in possibly different orders. Then $\mathbf{J} \sim \mathbf{J}'$, because reordering blocks is a change of basis, as noted after Definition 10.3.2. Hence $\mathbf{A} \sim \mathbf{J} \sim \mathbf{J}' \sim \mathbf{B}$, and similarity is an equivalence relation (Example (Similarity)).

■

This is the promise of Chapter 4 kept, and of Chapter 7: when $p_{\mathbf{A}}$ splits — always, over $\mathbb{C}$ — the characteristic polynomial together with the ranks of the powers $(\mathbf{A}-\lambda\mathbf{I})^j$ is a **complete** set of similarity invariants. Chapter 4's warning pair is decided instantly: $\mathbf{I}_2$ has Jordan form $\operatorname{diag}(1,1)$ and $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} = \mathbf{J}_2(1)$ is its own, so they are not similar. And the sharper failure of Chapter 9 becomes transparent.

Remark. The matrices $\mathbf{J}_2(1) \oplus \mathbf{J}_2(1)$ and $\mathbf{J}_2(1) \oplus \mathbf{J}_1(1) \oplus \mathbf{J}_1(1)$ in $M_4(F)$ have the same characteristic polynomial $(x-1)^4$ **and** the same minimal polynomial $(x-1)^2$, yet they are not similar: their Jordan forms differ, and $\operatorname{rank}(\mathbf{A}-\mathbf{I})$ is 2 against 1. So no pair of polynomials can decide similarity; the block sizes carry strictly more information.

When $p_{\mathbf{A}}$ does not split, the Jordan form does not exist and this test says nothing. Section 6 replaces it by the rational canonical form, which works over every field.

Warning. **The Jordan form is not stable under small changes of the matrix.** For $\varepsilon \neq 0$, $\mathbf{A}_\varepsilon = \begin{pmatrix} 0 & 1 \\ \varepsilon & 0 \end{pmatrix}$ over $\mathbb{C}$ has $p = x^2 - \varepsilon$ with two distinct roots $\pm\sqrt{\varepsilon}$, so $\mathbf{A}_\varepsilon$ is diagonalizable and its Jordan form is $\operatorname{diag}(\sqrt{\varepsilon}, -\sqrt{\varepsilon})$, with **no** superdiagonal one. At $\varepsilon = 0$ the Jordan form jumps to $\mathbf{J}_2(0)$. So arbitrarily small perturbations change the form completely, which is why numerical work avoids it; Chapter 20 returns to this point.

10.3.5 Computing a Jordan form

The proof is an algorithm: it says to work one eigenvalue at a time and to find chains inside each generalized eigenspace. The rank data tell us the shape before we look for a single vector.

Algorithm. **Jordan form and a Jordan basis of $\mathbf{A} \in M_n(F)$.**

1. Factor $p_{\mathbf{A}}$. If it does not split over F, stop: there is no Jordan form.
2. For each eigenvalue λ, compute $\operatorname{rank}(\mathbf{A}-\lambda\mathbf{I})^j$ for $j = 0, 1, 2, \ldots$ until the rank stops changing, and get the λ-block sizes from Theorem 10.3.1 (b), subtracting twice for the

exact counts.

3. For each λ, find a Jordan basis of the nilpotent $(\mathbf{A}-\lambda\mathbf{I})$ restricted to $G_\lambda(\mathbf{A})$, as in Section 2: tops of long chains first, then vectors of $\ker(\mathbf{A}-\lambda\mathbf{I})$ to complete.
4. Let $\mathbf{P}$ be the matrix whose columns are all the chains, in order; then $\mathbf{P}^{-1}\mathbf{A}\mathbf{P}=\mathbf{J}$. Check the answer as $\mathbf{A}\mathbf{P}=\mathbf{P}\mathbf{J}$, column by column.

Example. The Jordan Form of a 5×5 Matrix

Let

$$\mathbf{A}=\begin{pmatrix} 3 & 1 & 0 & 0 & -1 \\ -2 & 2 & 2 & -2 & -1 \\ 1 & 0 & 2 & -1 & 1 \\ 1 & 0 & -1 & 2 & 1 \\ -2 & -1 & 2 & -2 & 2 \end{pmatrix} \in M_5(\mathbb{Q}), \qquad p_{\mathbf{A}}=(x-3)^3(x-1)^2.$$

Find the Jordan form $\mathbf{J}$ of $\mathbf{A}$, a Jordan basis, and an invertible $\mathbf{P}$ with $\mathbf{P}^{-1}\mathbf{A}\mathbf{P}=\mathbf{J}$. Verify the answer by hand.

Solution. The eigenvalue 3. Put $\mathbf{N}=\mathbf{A}-3\mathbf{I}$, so

$$\mathbf{N}=\begin{pmatrix} 0 & 1 & 0 & 0 & -1 \\ -2 & -1 & 2 & -2 & -1 \\ 1 & 0 & -1 & -1 & 1 \\ 1 & 0 & -1 & -1 & 1 \\ -2 & -1 & 2 & -2 & -1 \end{pmatrix}.$$

Rows 3 and 4 are equal, and so are rows 2 and 5, so the row space is spanned by rows $1,2,3$. Those three are independent: their entries in columns $1,2,4$ form the matrix $\begin{pmatrix} 0 & 1 & 0 \\ -2 & -1 & -2 \\ 1 & 0 & -1 \end{pmatrix}$, of determinant $-4\neq 0$. Hence $\operatorname{rank}\mathbf{N}=3$ and $\operatorname{nullity}\mathbf{N}=2$ (Theorem 3.5.7). Solving $\mathbf{N}\mathbf{x}=\mathbf{0}$ means solving

$$x_2-x_5=0, \qquad -2x_1-x_2+2x_3-2x_4-x_5=0, \qquad x_1-x_3-x_4+x_5=0.$$

The first gives $x_2=x_5$; substituting into the second gives $-2x_1+2x_3-2x_4-2x_5=0$, that is, $x_1=x_3-x_4-x_5$; and the third then reads $(x_3-x_4-x_5)-x_3-x_4+x_5=-2x_4=0$, so $x_4=0$. Hence

$$E_3(\mathbf{A})=\ker\mathbf{N}=\operatorname{span}\big((1,0,1,0,0),\ (-1,1,0,0,1)\big), \qquad g_{\mathbf{A}}(3)=2.$$

Squaring,

$$\mathbf{N}^2=4\begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & -1 & 1 & 1 \\ -1 & 0 & 1 & 0 & -1 \\ -1 & 0 & 1 & 0 & -1 \\ 1 & 0 & -1 & 1 & 1 \end{pmatrix},$$

whose nonzero rows span a 2-dimensional space, so $\operatorname{rank}\mathbf{N}^2=2$. Since $\dim G_3(\mathbf{A})=a_{\mathbf{A}}(3)=3$ and $\dim\ker\mathbf{N}^2=3$ (Theorem 3.5.7), we have $G_3(\mathbf{A})=\ker\mathbf{N}^2$, so $\operatorname{rank}\mathbf{N}^3=\operatorname{rank}\mathbf{N}^2=2$.

The block sizes for $\lambda = 3$ follow from Theorem 10.3.1 (b) with $r_0, r_1, r_2, r_3 = 5, 3, 2, 2$: blocks of size ≥ 1: $5 - 3 = 2$; of size ≥ 2: $3 - 2 = 1$; of size ≥ 3: $2 - 2 = 0$. So the 3-blocks have sizes 2 and 1.

The chain for 3. We need a top $\mathbf{u} \in \ker \mathbf{N}^2 \setminus \ker \mathbf{N}$. The second column of $\mathbf{N}^2$ is $\mathbf{0}$, so $\mathbf{e}_2 \in \ker \mathbf{N}^2$, while the second column of $\mathbf{N}$ is $(1, -1, 0, 0, -1) \neq \mathbf{0}$, so $\mathbf{e}_2 \notin \ker \mathbf{N}$. Take $\mathbf{u} = \mathbf{e}_2$; its chain is

$$\mathbf{N}\mathbf{u} = (1, -1, 0, 0, -1), \qquad \mathbf{u} = \mathbf{e}_2.$$

For the block of size 1 we need a vector of $\ker \mathbf{N}$ completing $\{\mathbf{N}\mathbf{u}\}$ to a basis of $\ker \mathbf{N}$. The vectors $\mathbf{N}\mathbf{u} = (1, -1, 0, 0, -1)$ and $\mathbf{v} = (1, 0, 1, 0, 0)$ are not proportional (their second entries are -1 and 0, their first entries both 1), so, $\ker \mathbf{N}$ being 2-dimensional, they form a basis of it (Theorem 1.6.4 (a)) and $\mathbf{v}$ does it.

The eigenvalue 1. Put $\mathbf{M} = \mathbf{A} - \mathbf{I}$ and solve $\mathbf{M}\mathbf{x} = \mathbf{0}$. Rows 3 and 4 of $\mathbf{M}$ read $x_1 + x_3 - x_4 + x_5 = 0$ and $x_1 - x_3 + x_4 + x_5 = 0$; adding and subtracting them gives $x_1 + x_5 = 0$ and $x_3 = x_4$. Row 1 reads $2x_1 + x_2 - x_5 = 0$, and row 2, after using $x_3 = x_4$, reads $-2x_1 + x_2 - x_5 = 0$; adding and subtracting these two gives $x_2 = x_5$ and $x_1 = 0$, hence $x_5 = 0$ and $x_2 = 0$. So

$$E_1(\mathbf{A}) = \ker \mathbf{M} = \operatorname{span}\big((0, 0, 1, 1, 0)\big), \qquad g_{\mathbf{A}}(1) = 1.$$

Since $a_{\mathbf{A}}(1) = 2$, the only partition of 2 with one part is (2): there is a single 1-block, of size 2. Its chain needs a top $\mathbf{w} \in \ker \mathbf{M}^2 \setminus \ker \mathbf{M}$. Computing,

$$\mathbf{M}^2 = 4 \begin{pmatrix} 1 & 1 & 0 & 0 & -1 \\ -1 & 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ -1 & -1 & 1 & -1 & 1 \end{pmatrix},$$

whose kernel is cut out by $x_1 + x_2 - x_5 = 0$, $-x_1 + x_3 - x_4 = 0$, and $x_3 - x_4 = 0$, giving $x_1 = 0$, $x_3 = x_4$, $x_2 = x_5$; so $\ker \mathbf{M}^2 = \operatorname{span}\big((0, 0, 1, 1, 0),\ (0, 1, 0, 0, 1)\big)$, of dimension $2 = a_{\mathbf{A}}(1)$ as expected. Take $\mathbf{w} = (0, 1, 0, 0, 1)$; then

$$\begin{aligned} \mathbf{M}\mathbf{w} &= (\text{column } 2 + \text{column } 5 \text{ of } \mathbf{M}) \\ &= (1, 1, 0, 0, -1) + (-1, -1, 1, 1, 1) = (0, 0, 1, 1, 0), \end{aligned}$$

which is the eigenvector, as it must be.

The answer. Writing the three chains one after another,

$$\mathcal{B} = \big((1, -1, 0, 0, -1),\ \mathbf{e}_2,\ (1, 0, 1, 0, 0),\ (0, 0, 1, 1, 0),\ (0, 1, 0, 0, 1) \big),$$

$$\mathbf{P} = \begin{pmatrix} 1 & 0 & 1 & 0 & 0 \\ -1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ -1 & 0 & 0 & 0 & 1 \end{pmatrix},$$

$$\mathbf{J} = \mathbf{J}_2(3) \oplus \mathbf{J}_1(3) \oplus \mathbf{J}_2(1) = \begin{pmatrix} 3 & 1 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 \\ 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

$\mathbf{P}$ is invertible: its fourth row is $(0,0,0,1,0)$, so expanding $\det \mathbf{P}$ along that row (Theorem 7.6.1) leaves

$$\det \mathbf{P} = \det \begin{pmatrix} 1 & 0 & 1 & 0 \\ -1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ -1 & 0 & 0 & 1 \end{pmatrix} = \det \begin{pmatrix} 1 & 0 & 0 \\ -1 & 1 & 1 \\ -1 & 0 & 1 \end{pmatrix} = 1,$$

expanding along the third row and then along the first row.

The check, column by column. The claim $\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \mathbf{J}$ is the same as $\mathbf{A}\mathbf{P} = \mathbf{P}\mathbf{J}$. Writing $\mathbf{p}_1, \dots, \mathbf{p}_5$ for the columns of $\mathbf{P}$, the five columns of $\mathbf{P}\mathbf{J}$ are

$$3\mathbf{p}_1, \quad \mathbf{p}_1 + 3\mathbf{p}_2, \quad 3\mathbf{p}_3, \quad \mathbf{p}_4, \quad \mathbf{p}_4 + \mathbf{p}_5,$$

read straight off the columns of $\mathbf{J}$. So the identity $\mathbf{A}\mathbf{P} = \mathbf{P}\mathbf{J}$ is the list of five statements

$$\mathbf{A}\mathbf{p}_1 = 3\mathbf{p}_1, \quad \mathbf{A}\mathbf{p}_2 = \mathbf{p}_1 + 3\mathbf{p}_2, \quad \mathbf{A}\mathbf{p}_3 = 3\mathbf{p}_3, \quad \mathbf{A}\mathbf{p}_4 = \mathbf{p}_4, \quad \mathbf{A}\mathbf{p}_5 = \mathbf{p}_4 + \mathbf{p}_5,$$

which are exactly how the chains were constructed: $\mathbf{p}_1, \mathbf{p}_3 \in \ker(\mathbf{A} - 3\mathbf{I})$, $(\mathbf{A} - 3\mathbf{I})\mathbf{p}_2 = \mathbf{p}_1$, $\mathbf{p}_4 \in \ker(\mathbf{A} - \mathbf{I})$, and $(\mathbf{A} - \mathbf{I})\mathbf{p}_5 = \mathbf{p}_4$. Two of them verified in full:

$$\mathbf{A}\mathbf{p}_2 = \mathbf{A}\mathbf{e}_2 = (1,2,0,0,-1) = (1,-1,0,0,-1) + 3(0,1,0,0,0) = \mathbf{p}_1 + 3\mathbf{p}_2,$$

reading the second column of $\mathbf{A}$; and

$$\begin{aligned} \mathbf{A}\mathbf{p}_4 = \mathbf{A}(0,0,1,1,0) &= (0+0,\ 2-2,\ 2-1,\ -1+2,\ 2-2) \\ &= (0,0,1,1,0) = \mathbf{p}_4, \end{aligned}$$

adding the third and fourth columns of $\mathbf{A}$. The remaining three are the same kind of one-line computation.

Cross-checks. The diagonal of $\mathbf{J}$ gives $\operatorname{tr}\mathbf{A} = 3+3+3+1+1 = 11$, and adding the diagonal of $\mathbf{A}$ gives $3+2+2+2+2 = 11$. By Corollary 10.3.2, $a(3) = 2+1 = 3$, $g(3) = 2$, $a(1) = 2$, $g(1) = 1$, and $m_{\mathbf{A}} = (x-3)^2(x-1)^2$, all matching the computations above.

10.3.6 Exercises

10.3.6.1 A. Check your understanding

Exercise. A1

(a) Define a Jordan block and a Jordan matrix, and state what a Jordan form of T is.

(b) State the Jordan Canonical Form Theorem, including the uniqueness clause.

(c) How do you read $a(\lambda)$, $g(\lambda)$, and the exponent of $x-\lambda$ in m_T off a Jordan form?

(d) True or false: every matrix in $M_n(\mathbb{R})$ has a Jordan form over $\mathbb{R}$. Justify your answer.

(e) True or false: if $\mathbf{A}$ and $\mathbf{B}$ have the same characteristic and the same minimal polynomial, they are similar. Justify your answer.

(f) Is $\begin{pmatrix} 5 & 1 \\ 0 & 4 \end{pmatrix}$ a Jordan matrix? What is its Jordan form?

Solution.

(a) $\mathbf{J}_k(\lambda) = \lambda \mathbf{I}_k + \mathbf{J}_k(0)$ has λ on the diagonal and 1 on the superdiagonal (Definition 10.3.1); a Jordan matrix is a direct sum of Jordan blocks; a Jordan form of T is a Jordan matrix equal to $[T]_{\mathcal{B}}^{\mathcal{B}}$ for some basis $\mathcal{B}$ (Definition 10.3.2).

(b) T has a Jordan form if and only if p_T splits, and then the number of λ-blocks of size at least j is $\operatorname{rank}(T-\lambda \operatorname{id}_V)^{j-1} - \operatorname{rank}(T-\lambda \operatorname{id}_V)^{j}$, so the blocks are unique up to their order (Theorem 10.3.1).

(c) $a(\lambda)$ is the total size of the λ-blocks, $g(\lambda)$ their number, and the exponent in m_T the largest of their sizes (Corollary 10.3.2).

(d) False. A Jordan matrix is triangular, so its existence forces p_T to split; the rotation by a right angle has $p = x^2+1$, which does not split over $\mathbb{R}$.

(e) False. $\mathbf{J}_2(1) \oplus \mathbf{J}_2(1)$ and $\mathbf{J}_2(1) \oplus \mathbf{J}_1(1) \oplus \mathbf{J}_1(1)$ share $p = (x-1)^4$ and $m = (x-1)^2$ but have different Jordan forms, hence are not similar (Corollary 10.3.3).

(f) No: a superdiagonal 1 may only stand between two equal diagonal entries. The eigenvalues 5 and 4 are distinct, so the matrix is diagonalizable (Corollary 9.4.2) and its Jordan form is $\operatorname{diag}(5,4)$.

10.3.6.2 B. Practice

Exercise. B1: Jordan forms from polynomial data

In each case T is an operator on a complex vector space with the given characteristic and minimal polynomials. List all possible Jordan forms, up to the order of the blocks, and say what extra data would pin the answer down.

(a) $p_T = (x-2)^4$, $m_T = (x-2)^2$.

(b) $p_T = (x-1)^3(x+1)^2$, $m_T = (x-1)^3(x+1)$.

(c) $p_T = (x-5)^5$, $m_T = (x-5)^3$.

Solution. By Corollary 10.3.2, the λ-block sizes form a partition of $a(\lambda)$ whose largest part is the exponent of $x-\lambda$ in m_T.

(a) Partitions of 4 with largest part 2: $(2,2)$ and $(2,1,1)$. So $\mathbf{J}_2(2) \oplus \mathbf{J}_2(2)$ or $\mathbf{J}_2(2) \oplus \mathbf{J}_1(2) \oplus \mathbf{J}_1(2)$. The value of $g(2)$ — that is, 2 against 3 — decides.

(b) For 1: partitions of 3 with largest part 3, only (3). For -1: partitions of 2 with largest part 1, only $(1,1)$. So the Jordan form is $\mathbf{J}_3(1) \oplus \mathbf{J}_1(-1) \oplus \mathbf{J}_1(-1)$, and no extra data is needed.

(c) Partitions of 5 with largest part 3: $(3,2)$ and $(3,1,1)$. So $\mathbf{J}_3(5) \oplus \mathbf{J}_2(5)$ or $\mathbf{J}_3(5) \oplus \mathbf{J}_1(5) \oplus \mathbf{J}_1(5)$, told apart by $g(5) = 2$ against 3, or by $\dim \ker(T - 5\,\mathrm{id})^2 = 4$ against 5.

Exercise. B2: Jordan forms and Jordan bases

For each matrix, find the Jordan form $\mathbf{J}$ and an invertible $\mathbf{P}$ with $\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \mathbf{J}$. Verify your $\mathbf{P}$ by checking $\mathbf{A}\mathbf{P} = \mathbf{P}\mathbf{J}$ column by column.

(a) $\mathbf{A}_1 = \begin{pmatrix} -1 & -1 & -2 \\ 1 & 2 & 1 \\ 1 & 0 & 2 \end{pmatrix} \in M_3(\mathbb{Q})$, with $p_{\mathbf{A}_1} = (x-1)^3$.

(b) $\mathbf{A}_2 = \begin{pmatrix} 2 & -1 & -1 \\ -3 & 4 & -1 \\ 3 & -2 & 3 \end{pmatrix} \in M_3(\mathbb{Q})$, with $p_{\mathbf{A}_2} = (x-2)^2(x-5)$.

(c) $\mathbf{A}_3 = \begin{pmatrix} 2 & 1 & 1 & -2 \\ -3 & 1 & -1 & -1 \\ -3 & -1 & 1 & -1 \\ -3 & -1 & -1 & 1 \end{pmatrix} \in M_4(\mathbb{Q})$, with $p_{\mathbf{A}_3} = (x-2)^3(x+1)$.

Solution.

(a) $\mathbf{A}_1 - \mathbf{I} = \begin{pmatrix} -2 & -1 & -2 \\ 1 & 1 & 1 \\ 1 & 0 & 1 \end{pmatrix}$ has rank 2, so $g(1) = 1$ and there is one block, of size 3: $\mathbf{J} = \mathbf{J}_3(1)$. A top must satisfy $(\mathbf{A}_1 - \mathbf{I})^2\mathbf{u} \neq \mathbf{0}$, and $(\mathbf{A}_1 - \mathbf{I})^2 = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ -1 & -1 & -1 \end{pmatrix}$ has nonzero first column, so $\mathbf{u} = \mathbf{e}_1$ works. The chain is

$$(\mathbf{A}_1 - \mathbf{I})^2\mathbf{e}_1 = (1,0,-1), \qquad (\mathbf{A}_1 - \mathbf{I})\mathbf{e}_1 = (-2,1,1), \qquad \mathbf{e}_1,$$

so $\mathbf{P} = \begin{pmatrix} 1 & -2 & 1 \\ 0 & 1 & 0 \\ -1 & 1 & 0 \end{pmatrix}$, with $\det \mathbf{P} = 1$. The check $\mathbf{A}_1\mathbf{P} = \mathbf{P}\mathbf{J}_3(1)$ reads $\mathbf{A}_1\mathbf{p}_1 = \mathbf{p}_1$, $\mathbf{A}_1\mathbf{p}_2 = \mathbf{p}_1 + \mathbf{p}_2$, $\mathbf{A}_1\mathbf{p}_3 = \mathbf{p}_2 + \mathbf{p}_3$, which are the three chain relations.

(b) $\mathbf{A}_2 - 2\mathbf{I} = \begin{pmatrix} 0 & -1 & -1 \\ -3 & 2 & -1 \\ 3 & -2 & 1 \end{pmatrix}$ has rank 2, so $g(2) = 1$ and the single 2-block has size $a(2) = 2$; the eigenvalue 5 gives one block of size 1. So $\mathbf{J} = \mathbf{J}_2(2) \oplus \mathbf{J}_1(5)$. For the chain, the third column of $\mathbf{A}_2 - 2\mathbf{I}$ is $(-1,-1,1) \neq \mathbf{0}$ and $(\mathbf{A}_2 - 2\mathbf{I})(-1,-1,1) = \mathbf{0}$, so $\mathbf{u} = \mathbf{e}_3$ is a top with $(\mathbf{A}_2 - 2\mathbf{I})\mathbf{e}_3 = (-1,-1,1)$. Solving $(\mathbf{A}_2 - 5\mathbf{I})\mathbf{x} = \mathbf{0}$ gives $\mathrm{span}((0,-1,1))$.

Hence

$$\mathbf{P} = \begin{pmatrix} -1 & 0 & 0 \\ -1 & 0 & -1 \\ 1 & 1 & 1 \end{pmatrix}, \qquad \det \mathbf{P} = -1,$$

and $\mathbf{A}_2\mathbf{P} = \mathbf{P}\mathbf{J}$ reads $\mathbf{A}_2\mathbf{p}_1 = 2\mathbf{p}_1$, $\mathbf{A}_2\mathbf{p}_2 = \mathbf{p}_1 + 2\mathbf{p}_2$, $\mathbf{A}_2\mathbf{p}_3 = 5\mathbf{p}_3$.

(c) $\mathbf{A}_3 - 2\mathbf{I} = \begin{pmatrix} 0 & 1 & 1 & -2 \\ -3 & -1 & -1 & -1 \\ -3 & -1 & -1 & -1 \\ -3 & -1 & -1 & -1 \end{pmatrix}$ has rank 2 (rows $2, 3, 4$ are equal), so $g(2) = 2$: with $a(2) = 3$ the 2-blocks have sizes 2 and 1. The eigenvalue -1 has $a = 1$, so $\mathbf{J} = \mathbf{J}_2(2) \oplus \mathbf{J}_1(2) \oplus \mathbf{J}_1(-1)$. Now $\ker(\mathbf{A}_3 - 2\mathbf{I}) = \operatorname{span}\big((0, -1, 1, 0), (-1, 2, 0, 1)\big)$, and the second column of $\mathbf{A}_3 - 2\mathbf{I}$ is $(1, -1, -1, -1)$, which is nonzero and lies in that kernel, so $\mathbf{u} = \mathbf{e}_2$ is the top of the 2-chain. Completing $\{(1, -1, -1, -1)\}$ to a basis of $\ker(\mathbf{A}_3 - 2\mathbf{I})$ with $(0, -1, 1, 0)$, and solving $(\mathbf{A}_3 + \mathbf{I})\mathbf{x} = \mathbf{0}$ to get $\operatorname{span}((0, 1, 1, 1))$, we obtain

$$\mathbf{P} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -1 & 1 & -1 & 1 \\ -1 & 0 & 1 & 1 \\ -1 & 0 & 0 & 1 \end{pmatrix}, \qquad \det \mathbf{P} = 1,$$

and $\mathbf{A}_3\mathbf{P} = \mathbf{P}\mathbf{J}$ reads $\mathbf{A}_3\mathbf{p}_1 = 2\mathbf{p}_1$, $\mathbf{A}_3\mathbf{p}_2 = \mathbf{p}_1 + 2\mathbf{p}_2$, $\mathbf{A}_3\mathbf{p}_3 = 2\mathbf{p}_3$, $\mathbf{A}_3\mathbf{p}_4 = -\mathbf{p}_4$. For instance the second of these is $\mathbf{A}_3\mathbf{e}_2 = (1, 1, -1, -1) = (1, -1, -1, -1) + 2(0, 1, 0, 0)$, read off the second column of $\mathbf{A}_3$.

Exercise. B3: Similar or not?

Determine whether each pair of matrices is similar. Justify your answer.

(a) $\mathbf{J}_2(4) \oplus \mathbf{J}_1(4)$ and $\mathbf{J}_1(4) \oplus \mathbf{J}_2(4)$.

(b) $\mathbf{J}_3(0) \oplus \mathbf{J}_1(0)$ and $\mathbf{J}_2(0) \oplus \mathbf{J}_2(0)$.

(c) $\mathbf{J}_2(1) \oplus \mathbf{J}_2(1)$ and $\mathbf{J}_2(1) \oplus \mathbf{J}_1(1) \oplus \mathbf{J}_1(1)$.

(d) $\operatorname{diag}(2, 2, 5)$ and the matrix $\mathbf{A}_2$ of Exercise (B2: Jordan forms and Jordan bases) (b).

Solution.

(a) Similar: the blocks are the same and only their order differs, which is a change of basis, as noted after Definition 10.3.2.

(b) Not similar. Both are nilpotent of rank 2, but $\operatorname{rank}(\mathbf{J}_3(0) \oplus \mathbf{J}_1(0))^2 = 1$ while $(\mathbf{J}_2(0) \oplus \mathbf{J}_2(0))^2 = \mathbf{O}$ (Theorem 8.1.4 (b)). By Corollary 10.3.3 (or Corollary 10.2.5) they are not similar; their partitions are $(3, 1)$ and $(2, 2)$.

(c) Not similar, although p and m agree, as in the remark after Corollary 10.3.3: $\operatorname{rank}(\mathbf{A} - \mathbf{I}) = 2$ for the first and 1 for the second, so $g(1)$ is 2 against 3.

(d) Not similar. $\operatorname{diag}(2, 2, 5)$ is diagonal, so $g(2) = 2$, while $\mathbf{A}_2$ has $g(2) = 1$ by Exercise (B2: Jordan forms and Jordan bases) (b); its Jordan form is $\mathbf{J}_2(2) \oplus \mathbf{J}_1(5)$, which is not diagonal. The characteristic polynomials agree, so this pair is another witness that $p_{\mathbf{A}}$ alone decides nothing.

10.3.6.3 C. Going deeper

Exercise. C1: A matrix is similar to its transpose

Let $\mathbf{A} \in M_n(F)$ with $p_{\mathbf{A}}$ split over F.

(a) Let $\mathbf{R} \in M_k(F)$ be the reversal permutation matrix, with 1 in each position $(j, k+1-j)$ and 0 elsewhere. Prove that $\mathbf{R}^{-1} = \mathbf{R}$ and $\mathbf{R}^{-1}\mathbf{J}_k(\lambda)\mathbf{R} = \mathbf{J}_k(\lambda)^{\top}$.

(b) Deduce that every Jordan matrix $\mathbf{J}$ satisfies $\mathbf{J} \sim \mathbf{J}^{\top}$.

(c) Deduce that $\mathbf{A} \sim \mathbf{A}^{\top}$.

Hint: for (c), transpose a similarity $\mathbf{A} = \mathbf{P}\mathbf{J}\mathbf{P}^{-1}$. The conclusion holds over every field; Section 7 removes the hypothesis that $p_{\mathbf{A}}$ splits.

Solution.

(a) $\mathbf{R}\mathbf{e}_j = \mathbf{e}_{k+1-j}$, so $\mathbf{R}^2\mathbf{e}_j = \mathbf{e}_j$ for every j and $\mathbf{R}^2 = \mathbf{I}_k$, that is, $\mathbf{R}^{-1} = \mathbf{R}$. Since $\mathbf{J}_k(\lambda) = \lambda\mathbf{I}_k + \mathbf{J}_k(0)$ and $\mathbf{R}^{-1}(\lambda\mathbf{I}_k)\mathbf{R} = \lambda\mathbf{I}_k$, it suffices to prove $\mathbf{R}^{-1}\mathbf{J}_k(0)\mathbf{R} = \mathbf{J}_k(0)^{\top}$. Both sides are computed on basis vectors: $\mathbf{J}_k(0)\mathbf{e}_1 = \mathbf{0}$ and $\mathbf{J}_k(0)\mathbf{e}_j = \mathbf{e}_{j-1}$ for $j \geq 2$, so

$$\mathbf{R}\mathbf{J}_k(0)\mathbf{R}\,\mathbf{e}_j = \mathbf{R}\,\mathbf{J}_k(0)\,\mathbf{e}_{k+1-j} = \begin{cases} \mathbf{R}\,\mathbf{e}_{k-j} = \mathbf{e}_{j+1}, & j \leq k-1, \\ \mathbf{R}\,\mathbf{0} = \mathbf{0}, & j = k, \end{cases}$$

which is exactly the action of $\mathbf{J}_k(0)^{\top}$, whose columns satisfy $\mathbf{J}_k(0)^{\top}\mathbf{e}_j = \mathbf{e}_{j+1}$ for $j \leq k-1$ and $\mathbf{J}_k(0)^{\top}\mathbf{e}_k = \mathbf{0}$. Hence $\mathbf{R}^{-1}\mathbf{J}_k(\lambda)\mathbf{R} = \mathbf{J}_k(\lambda)^{\top}$.

(b) Let $\mathbf{J} = \mathbf{J}_{k_1}(\lambda_1) \oplus \cdots \oplus \mathbf{J}_{k_m}(\lambda_m)$ and let $\mathbf{Q} = \mathbf{R}_{k_1} \oplus \cdots \oplus \mathbf{R}_{k_m}$, the direct sum of the reversal matrices of the block sizes. Then $\mathbf{Q}$ is invertible with $\mathbf{Q}^{-1} = \mathbf{R}_{k_1}^{-1} \oplus \cdots \oplus \mathbf{R}_{k_m}^{-1}$ (Theorem 8.1.4 (c)), and multiplying blockwise (Theorem 8.1.4 (a)),

$$\mathbf{Q}^{-1}\mathbf{J}\mathbf{Q} = (\mathbf{R}_{k_1}^{-1}\mathbf{J}_{k_1}(\lambda_1)\mathbf{R}_{k_1}) \oplus \cdots = \mathbf{J}_{k_1}(\lambda_1)^{\top} \oplus \cdots \oplus \mathbf{J}_{k_m}(\lambda_m)^{\top} = \mathbf{J}^{\top},$$

the last equality because transposing a block diagonal matrix transposes each block (Proposition 8.1.2).

(c) By Theorem 10.3.1 there are a Jordan matrix $\mathbf{J}$ and an invertible $\mathbf{P}$ with $\mathbf{A} = \mathbf{P}\mathbf{J}\mathbf{P}^{-1}$. Transposing and using $(\mathbf{X}\mathbf{Y})^{\top} = \mathbf{Y}^{\top}\mathbf{X}^{\top}$ and $(\mathbf{P}^{-1})^{\top} = (\mathbf{P}^{\top})^{-1}$ (Theorem 0.9.5, Theorem 0.9.7),

$$\mathbf{A}^{\top} = (\mathbf{P}^{-1})^{\top}\mathbf{J}^{\top}\mathbf{P}^{\top} = (\mathbf{P}^{\top})^{-1}\mathbf{J}^{\top}\mathbf{P}^{\top},$$

so $\mathbf{A}^{\top} \sim \mathbf{J}^{\top}$. By (b), $\mathbf{J}^{\top} \sim \mathbf{J} \sim \mathbf{A}$, and similarity is an equivalence relation (Example (Similarity)), so $\mathbf{A} \sim \mathbf{A}^{\top}$.

Exercise. C2: How much do p and m decide?

Work over $\mathbb{C}$.

(a) Find all Jordan forms with $p_T = (x-1)^3(x-2)^2$ and $m_T = (x-1)^2(x-2)$.

(b) Find all Jordan forms with $p_T = (x-1)^4(x-2)^2$ and $m_T = (x-1)^2(x-2)$.

(c) In (b), what is the least amount of extra information that determines the Jordan form?

(d) Prove that p_T and m_T determine the Jordan form whenever $a(\lambda) \leq 3$ for every eigen-

value λ.

Solution.

(a) For $\lambda = 1$: partitions of 3 with largest part 2, so $(2,1)$ only. For $\lambda = 2$: partitions of 2 with largest part 1, so $(1,1)$ only. There is exactly one Jordan form,

$$\mathbf{J}_2(1) \oplus \mathbf{J}_1(1) \oplus \mathbf{J}_1(2) \oplus \mathbf{J}_1(2).$$

(b) For $\lambda = 1$: partitions of 4 with largest part 2, namely $(2,2)$ and $(2,1,1)$. For $\lambda = 2$: $(1,1)$ as before. So there are two Jordan forms,

$$\begin{aligned} &\mathbf{J}_2(1) \oplus \mathbf{J}_2(1) \oplus \mathbf{J}_1(2) \oplus \mathbf{J}_1(2) \\ \text{and} \quad &\mathbf{J}_2(1) \oplus \mathbf{J}_1(1) \oplus \mathbf{J}_1(1) \oplus \mathbf{J}_1(2) \oplus \mathbf{J}_1(2). \end{aligned}$$

(c) The single number $g(1)$, which is 2 for the first and 3 for the second; equivalently $\operatorname{rank}(T - \mathrm{id}_V)$ or $\dim \ker(T - \mathrm{id}_V)$ (Corollary 10.3.2 (b)).

(d) Fix λ with $a = a(\lambda) \leq 3$ and let s be the exponent of $x - \lambda$ in m_T, which is the largest block size (Corollary 10.3.2 (c)). The λ-blocks form a partition of a with largest part s. For $a = 1$ the only partition is (1). For $a = 2$: $s = 2$ gives (2), and $s = 1$ gives $(1,1)$. For $a = 3$: $s = 3$ gives (3), $s = 2$ gives $(2,1)$, and $s = 1$ gives $(1,1,1)$. In every case the pair (a, s) determines the partition, so the blocks for each eigenvalue, and hence the whole Jordan form, are determined by p_T and m_T. The first ambiguity appears at $a = 4$, $s = 2$, which is part (a) of Exercise (B1: Jordan forms from polynomial data).

10.4 The Real Jordan Form and the Weyr Form

The Jordan form of Section 3 exists only when p_T splits, and over $\mathbb{R}$ that often fails: a rotation of the plane has no real eigenvalue at all. But a real matrix is also a complex matrix, and its complex eigenvalues come in conjugate pairs (Theorem 9.2.5). This section turns a complex Jordan basis into a **real** one by pairing each chain with its conjugate; the price for a conjugate pair is a 2×2 rotation-scaling block instead of a scalar, and the result is the best real matrix there is. The second half rearranges the same data in another way, the Weyr form, which is the form to reach for when a second operator commutes with the first.

10.4.1 A block for a conjugate pair

The obstruction is visible in the smallest case. The rotation $\mathbf{R} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \in M_2(\mathbb{R})$ has $p_{\mathbf{R}} = x^2 + 1$, so $\mathbf{R}$ has no real eigenvalue and no real eigenvector (Example (A Rotation Has No Invariant Lines)). Every real matrix similar to $\mathbf{R}$ therefore has no real eigenvalue either, so it cannot be triangular: whatever basis of $\mathbb{R}^2$ we choose, the matrix of $\mathbf{R}$ keeps two off-diagonal entries. So $\mathbf{R}$ is already as simple as it will ever get over $\mathbb{R}$, and the honest thing to do is to promote it to a building block.

More generally, for $\lambda = a + bi \in \mathbb{C}$ with $b \neq 0$, consider

$$\mathbf{\Lambda} = \begin{pmatrix} a & -b \\ b & a \end{pmatrix} \in M_2(\mathbb{R}), \qquad p_{\mathbf{\Lambda}} = (x-a)^2 + b^2 = \big(x - (a+bi)\big)\big(x - (a-bi)\big).$$

This matrix carries the conjugate pair $\lambda, \overline{\lambda}$ as its complex eigenvalues, has trace $2a = 2\operatorname{Re}\lambda$ and determinant $a^2 + b^2 = |\lambda|^2$, and is a scaled rotation: it stretches by $|\lambda|$ and turns by the argument of λ. It is the real substitute for the 1×1 block (λ), and it costs one extra dimension.

A real Jordan block is a Jordan block with the scalar λ replaced by the 2×2 rotation-scaling matrix it needs over $\mathbb{R}$, and the superdiagonal 1 replaced by $\mathbf{I}_2$.

Definition 10.4.1. Rotation-Scaling Block, Real Jordan Block

Let $\lambda = a + bi \in \mathbb{C}$ with $a, b \in \mathbb{R}$ and $b \neq 0$. The **rotation-scaling block** of the conjugate pair $\{\lambda, \overline{\lambda}\}$ is

$$\mathbf{\Lambda}(\lambda) \coloneqq \begin{pmatrix} a & -b \\ b & a \end{pmatrix} \in M_2(\mathbb{R}).$$

For an integer $k \geq 1$, the **real Jordan block** of size k for the pair $\{\lambda, \overline{\lambda}\}$ is the matrix $\mathbf{C}_k(\lambda) \in M_{2k}(\mathbb{R})$, partitioned into $k \times k$ blocks of size 2×2, with $\mathbf{\Lambda}(\lambda)$ in every diagonal block, $\mathbf{I}_2$ in every **superdiagonal** block, and the zero matrix elsewhere:

$$\mathbf{C}_k(\lambda) \coloneqq \begin{pmatrix} \mathbf{\Lambda}(\lambda) & \mathbf{I}_2 & & \\ & \mathbf{\Lambda}(\lambda) & \ddots & \\ & & \ddots & \mathbf{I}_2 \\ & & & \mathbf{\Lambda}(\lambda) \end{pmatrix}.$$

A **real Jordan matrix** is a direct sum of ordinary Jordan blocks $\mathbf{J}_k(\mu)$ with $\mu \in \mathbb{R}$ and real Jordan blocks $\mathbf{C}_k(\lambda)$ with $\lambda \notin \mathbb{R}$ and $\operatorname{Im}\lambda > 0$.

In words: $\mathbf{C}_k(\lambda)$ is built exactly like $\mathbf{J}_k(\lambda)$ (Definition 10.3.1), but every entry of $\mathbf{J}_k(\lambda)$ is replaced by a 2×2 block — the diagonal λ by $\mathbf{\Lambda}(\lambda)$, the superdiagonal 1 by $\mathbf{I}_2$, and each 0 by the 2×2 zero matrix. It has size $2k$, not k. The last clause matters: $\mathbf{\Lambda}(\overline{\lambda}) = \mathbf{\Lambda}(\lambda)^{\top}$, so the pair $\{\lambda, \overline{\lambda}\}$ offers two candidate blocks of each size, and a real Jordan matrix is required to use the member with $b > 0$, which makes $\mathbf{\Lambda}(\lambda)$ a rotation through a positive angle. Without that normalization the uniqueness below would fail, since $\mathbf{C}_k(\lambda)$ and $\mathbf{C}_k(\overline{\lambda})$ are similar over $\mathbb{R}$.

Examples.

- **Size one.** $\mathbf{C}_1(\lambda) = \mathbf{\Lambda}(\lambda)$. For $\lambda = i$ this is the rotation $\mathbf{R}$ above; for $\lambda = 2 + 3i$ it is $\begin{pmatrix} 2 & -3 \\ 3 & 2 \end{pmatrix}$.
- **Size two.** For $\lambda = 1 + i$,

$$\mathbf{C}_2(1+i) = \begin{pmatrix} 1 & -1 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 1 & 1 \end{pmatrix} \in M_4(\mathbb{R}).$$

  Its characteristic polynomial is $\big((x-1)^2 + 1\big)^2 = (x^2 - 2x + 2)^2$, by Theorem 7.8.4 applied to the block upper triangular shape.

- **A real Jordan matrix mixing the two kinds.** $\mathbf{J}_2(3)\oplus\mathbf{C}_1(i) \in M_4(\mathbb{R})$ has characteristic polynomial $(x-3)^2(x^2+1)$.

Nonexample by minimal change. Change the sign of the lower-left entry: $\begin{pmatrix} a & -b \\ -b & a \end{pmatrix}$ is still real, still has trace $2a$, and looks just as harmless. But it is symmetric, its characteristic polynomial is $(x-a)^2-b^2$, and its eigenvalues $a\pm b$ are **real**; for $a=0, b=1$ it is the coordinate swap, which is diagonalizable over $\mathbb{R}$. The clause that fails is that the two off-diagonal entries must have **opposite** signs; that is what makes the eigenvalues nonreal.

Why this definition. Two conventions are being fixed, and both are only conventions. First, the sign of b: replacing $\mathbf{\Lambda}(\lambda)$ by $\mathbf{\Lambda}(\overline{\lambda}) = \mathbf{\Lambda}(\lambda)^\top$ amounts to changing the sign of every second basis vector, so the two choices give similar matrices. Second, the ones: a real Jordan block could be written with $\mathbf{I}_2$ below the diagonal instead, giving the transpose. What is **not** a convention is the size: a conjugate pair cannot be carried by a 1×1 real block, because a 1×1 real matrix has a real eigenvalue.

The following computation is the whole idea of the section, and we will use it twice: once to see what $\mathbf{C}_k(\lambda)$ is over $\mathbb{C}$, and once to build the real basis.

Lemma 10.4.1. Realifying a Conjugate Pair of Chains

Let $\mathbf{A} \in M_n(\mathbb{R})$, let $\lambda = a+bi$ with $a, b \in \mathbb{R}$ and $b \neq 0$, and let $(\mathbf{w}_1, \dots, \mathbf{w}_k)$ be a Jordan chain of $\mathbf{A}$ for λ in $\mathbb{C}^n$, so that

$$(\mathbf{A}-\lambda\mathbf{I})\mathbf{w}_1 = \mathbf{0} \qquad \text{and} \qquad (\mathbf{A}-\lambda\mathbf{I})\mathbf{w}_j = \mathbf{w}_{j-1} \quad (2 \le j \le k).$$

Write $\mathbf{w}_j = \mathbf{u}_j - i\mathbf{v}_j$ with $\mathbf{u}_j, \mathbf{v}_j \in \mathbb{R}^n$, and put $\mathbf{u}_0 = \mathbf{v}_0 = \mathbf{0}$. Then:

(a) $\mathbf{A}\mathbf{u}_j = a\mathbf{u}_j + b\mathbf{v}_j + \mathbf{u}_{j-1}$ and $\mathbf{A}\mathbf{v}_j = -b\mathbf{u}_j + a\mathbf{v}_j + \mathbf{v}_{j-1}$ for $1 \le j \le k$;

(b) $(\overline{\mathbf{w}_1}, \dots, \overline{\mathbf{w}_k})$ is a Jordan chain of $\mathbf{A}$ for $\overline{\lambda}$;

(c) the $2k$ vectors $\mathbf{u}_1, \mathbf{v}_1, \dots, \mathbf{u}_k, \mathbf{v}_k$ span, over $\mathbb{C}$, the same subspace of $\mathbb{C}^n$ as the $2k$ vectors $\mathbf{w}_1, \dots, \mathbf{w}_k, \overline{\mathbf{w}_1}, \dots, \overline{\mathbf{w}_k}$, and one of the two lists is linearly independent over $\mathbb{C}$ if and only if the other is;

(d) if the $2k$ vectors $\mathbf{u}_1, \mathbf{v}_1, \dots, \mathbf{u}_k, \mathbf{v}_k$ are linearly independent over $\mathbb{R}$, then $U = \operatorname{span}(\mathbf{u}_1, \mathbf{v}_1, \dots, \mathbf{u}_k, \mathbf{v}_k) \subseteq \mathbb{R}^n$ is invariant under $\mathbf{A}$, and the matrix of $T_{\mathbf{A}}|_U$ in the ordered basis $(\mathbf{u}_1, \mathbf{v}_1, \dots, \mathbf{u}_k, \mathbf{v}_k)$ is $\mathbf{C}_k(\lambda)$.

Idea. For (a), write the chain relation as $\mathbf{A}\mathbf{w}_j = \lambda\mathbf{w}_j + \mathbf{w}_{j-1}$ and compare real and imaginary parts; since $\mathbf{A}$ is real, $\mathbf{A}$ acts on the two parts separately, and the only mixing comes from multiplying by $a+bi$. The minus sign in $\mathbf{w}_j = \mathbf{u}_j - i\mathbf{v}_j$ is chosen so that the answer comes out with the signs of $\mathbf{\Lambda}(\lambda)$ rather than its transpose. Part (c) is bookkeeping: $(\mathbf{u}_j, \mathbf{v}_j)$ and $(\mathbf{w}_j, \overline{\mathbf{w}_j})$ are obtained from each other by an invertible 2×2 matrix of scalars. Part (d) just reads off the columns from (a).

Proof.

(a) Every $\mathbf{z} \in \mathbb{C}^n$ is uniquely $\mathbf{x}+i\mathbf{y}$ with $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, by Definition 0.6.1 applied entry by entry; so $\mathbf{u}_j, \mathbf{v}_j$ are well defined. The chain relation says $\mathbf{A}\mathbf{w}_j = \lambda\mathbf{w}_j + \mathbf{w}_{j-1}$ for

$1 \le j \le k$. Expanding both sides with $\mathbf{w}_j = \mathbf{u}_j - i\mathbf{v}_j$ and $\lambda = a + bi$,

$$\begin{aligned}\mathbf{A}\mathbf{u}_j - i\mathbf{A}\mathbf{v}_j &= (a+bi)(\mathbf{u}_j - i\mathbf{v}_j) + \mathbf{u}_{j-1} - i\mathbf{v}_{j-1} \\ &= (a\mathbf{u}_j + b\mathbf{v}_j + \mathbf{u}_{j-1}) + i(b\mathbf{u}_j - a\mathbf{v}_j - \mathbf{v}_{j-1}),\end{aligned}$$

where we used $i^2 = -1$ and, on the left, that $\mathbf{A}$ has real entries, so $\mathbf{A}(\mathbf{u}_j - i\mathbf{v}_j) = \mathbf{A}\mathbf{u}_j - i\mathbf{A}\mathbf{v}_j$ with $\mathbf{A}\mathbf{u}_j, \mathbf{A}\mathbf{v}_j \in \mathbb{R}^n$. Comparing real parts gives the first identity, and comparing imaginary parts gives $-\mathbf{A}\mathbf{v}_j = b\mathbf{u}_j - a\mathbf{v}_j - \mathbf{v}_{j-1}$, that is, the second.

(b) Conjugating the identity $(\mathbf{A} - \lambda\mathbf{I})\mathbf{w}_j = \mathbf{w}_{j-1}$ entry by entry and using that $\mathbf{A}$ is real, $\overline{\mathbf{A}\mathbf{z}} = \mathbf{A}\overline{\mathbf{z}}$ and $\overline{\lambda\mathbf{z}} = \overline{\lambda}\,\overline{\mathbf{z}}$ for every $\mathbf{z} \in \mathbb{C}^n$ (Theorem 0.6.2 (a), (b)). Hence $(\mathbf{A} - \overline{\lambda}\mathbf{I})\overline{\mathbf{w}_j} = \overline{\mathbf{w}_{j-1}}$ for $j \ge 2$ and $(\mathbf{A} - \overline{\lambda}\mathbf{I})\overline{\mathbf{w}_1} = \mathbf{0}$, which is the chain condition for $\overline{\lambda}$. (That $\overline{\mathbf{w}_1} \neq \mathbf{0}$ follows from $\mathbf{w}_1 \neq \mathbf{0}$.)

(c) Since $\mathbf{w}_j = \mathbf{u}_j - i\mathbf{v}_j$ and $\overline{\mathbf{w}_j} = \mathbf{u}_j + i\mathbf{v}_j$, we have

$$\mathbf{u}_j = \tfrac{1}{2}(\mathbf{w}_j + \overline{\mathbf{w}_j}), \qquad \mathbf{v}_j = \tfrac{i}{2}(\mathbf{w}_j - \overline{\mathbf{w}_j}).$$

So each vector of either list is a $\mathbb{C}$-combination of the other list, and the two spans agree (Theorem 1.4.2). The passage from $(\mathbf{w}_j, \overline{\mathbf{w}_j})$ to $(\mathbf{u}_j, \mathbf{v}_j)$ is given, for each j separately, by the matrix $\begin{pmatrix} 1/2 & 1/2 \\ i/2 & -i/2 \end{pmatrix}$ of determinant $-i/2 \neq 0$, which is invertible; so the change of lists is an invertible linear substitution of the coefficients, and a nontrivial vanishing combination of one list yields one of the other. Hence one list is independent if and only if the other is.

(d) By (a), every $\mathbf{A}\mathbf{u}_j$ and every $\mathbf{A}\mathbf{v}_j$ lies in U, so U is $\mathbf{A}$-invariant (Lemma 9.1.1). Order the basis as $(\mathbf{u}_1, \mathbf{v}_1, \dots, \mathbf{u}_k, \mathbf{v}_k)$ and read off the columns (Definition 4.1.1). By (a), the column of $\mathbf{u}_j$ has entry a at $\mathbf{u}_j$, b at $\mathbf{v}_j$, 1 at $\mathbf{u}_{j-1}$, and 0 elsewhere; the column of $\mathbf{v}_j$ has $-b$ at $\mathbf{u}_j$, a at $\mathbf{v}_j$, 1 at $\mathbf{v}_{j-1}$, and 0 elsewhere. Grouping the basis into the consecutive pairs $(\mathbf{u}_j, \mathbf{v}_j)$, the (j,j) block is $\begin{pmatrix} a & -b \\ b & a \end{pmatrix} = \mathbf{\Lambda}(\lambda)$, the $(j-1, j)$ block is $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \mathbf{I}_2$, and all other blocks are zero. That is $\mathbf{C}_k(\lambda)$ (Definition 10.4.1).

■

Applying the lemma to the block itself tells us what a real Jordan block looks like over $\mathbb{C}$.

Corollary 10.4.2. A Real Jordan Block Splits into Two Complex Ones

Let $\lambda \notin \mathbb{R}$ and $k \ge 1$. Then, as matrices over $\mathbb{C}$,

$$\mathbf{C}_k(\lambda) \sim \mathbf{J}_k(\lambda) \oplus \mathbf{J}_k(\overline{\lambda}).$$

In particular $p_{\mathbf{C}_k(\lambda)} = \left((x-\lambda)(x-\overline{\lambda})\right)^k$, the polynomial $\left((x-a)^2 + b^2\right)^k \in \mathbb{R}[x]$, and $m_{\mathbf{C}_k(\lambda)}$ is the same polynomial.

Proof. Put $\mathbf{A} = \mathbf{C}_k(\lambda) \in M_{2k}(\mathbb{R})$ and let $(\mathbf{u}_1, \mathbf{v}_1, \dots, \mathbf{u}_k, \mathbf{v}_k)$ be the standard basis of $\mathbb{R}^{2k}$ in this order. Reading the columns of $\mathbf{C}_k(\lambda)$ gives exactly the identities of Lemma 10.4.1 (a).

Put $\mathbf{w}_j = \mathbf{u}_j - i\mathbf{v}_j$, so that $\mathbf{w}_1 = \mathbf{e}_1 - i\mathbf{e}_2 \neq \mathbf{0}$. Combining the two real identities,

$$\begin{aligned}\mathbf{A}\mathbf{w}_j = \mathbf{A}\mathbf{u}_j - i\mathbf{A}\mathbf{v}_j &= (a\mathbf{u}_j + b\mathbf{v}_j + \mathbf{u}_{j-1}) - i(-b\mathbf{u}_j + a\mathbf{v}_j + \mathbf{v}_{j-1}) \\ &= (a+bi)\mathbf{u}_j + (b-ai)\mathbf{v}_j + \mathbf{w}_{j-1} \\ &= \lambda\mathbf{w}_j + \mathbf{w}_{j-1},\end{aligned}$$

since $\lambda(-i\mathbf{v}_j) = (b-ai)\mathbf{v}_j$. So $(\mathbf{w}_1, \dots, \mathbf{w}_k)$ is a Jordan chain of $\mathbf{A}$ for λ, and by (b) so is $(\overline{\mathbf{w}_1}, \dots, \overline{\mathbf{w}_k})$ for $\overline{\lambda}$. By (c), the $2k$ vectors $\mathbf{w}_1, \dots, \mathbf{w}_k, \overline{\mathbf{w}_1}, \dots, \overline{\mathbf{w}_k}$ are independent over $\mathbb{C}$, because the standard basis is; being $2k$ independent vectors in $\mathbb{C}^{2k}$, they form a basis (Theorem 1.6.4 (a)). On the span of one chain, $\mathbf{A} - \lambda\mathbf{I}$ has matrix $\mathbf{J}_k(0)$ by Lemma 10.2.2 (b), so $\mathbf{A}$ has matrix $\lambda\mathbf{I}_k + \mathbf{J}_k(0) = \mathbf{J}_k(\lambda)$ there, the matrix of a sum of operators being the sum of their matrices (Theorem 4.1.4, Definition 10.3.1); likewise $\mathbf{J}_k(\overline{\lambda})$ on the span of the conjugate chain. By Theorem 9.1.5 the matrix of $T_{\mathbf{A}}$ in the combined basis is $\mathbf{J}_k(\lambda) \oplus \mathbf{J}_k(\overline{\lambda})$. Hence $\mathbf{C}_k(\lambda) \sim \mathbf{J}_k(\lambda) \oplus \mathbf{J}_k(\overline{\lambda})$ over $\mathbb{C}$ (Theorem 4.2.4).

The characteristic polynomial of $\mathbf{J}_k(\mu)$ is $(x-\mu)^k$ (Theorem 7.4.6), and similar matrices have the same characteristic polynomial (Theorem 7.10.3), so $p_{\mathbf{C}_k(\lambda)} = (x-\lambda)^k(x-\overline{\lambda})^k$. For the minimal polynomial, $m_{\mathbf{J}_k(\mu)} = (x-\mu)^k$ by Corollary 10.3.2 (c), a single Jordan block being a Jordan form of itself with one block, of size k; so $m_{\mathbf{C}_k(\lambda)} = \operatorname{lcm}\left((x-\lambda)^k, (x-\overline{\lambda})^k\right) = (x-\lambda)^k(x-\overline{\lambda})^k$ by Proposition 9.5.6 and Theorem 9.5.5, the factors being powers of distinct irreducibles.

■

Quick check. What are the complex eigenvalues of $\mathbf{C}_2(1+i)$, with their algebraic and geometric multiplicities?

Solution. By Corollary 10.4.2, $\mathbf{C}_2(1+i) \sim \mathbf{J}_2(1+i) \oplus \mathbf{J}_2(1-i)$ over $\mathbb{C}$. So the eigenvalues are $1 \pm i$, each with $a(\lambda) = 2$ (the total size of its blocks) and $g(\lambda) = 1$ (the number of its blocks), by Corollary 10.3.2 (a), (b).

10.4.2 The real Jordan form

Two facts about real matrices are needed before the theorem. The first is that a real matrix has the same kernel over $\mathbb{R}$ as over $\mathbb{C}$, in the sense that a real basis of the one is a complex basis of the other. Once one knows that $\tilde{U}$ is the complexification of U inside $\mathbb{C}^n = (\mathbb{R}^n)_{\mathbb{C}}$ (Definition 1.9.2, Example (The Complexification of $\mathbb{R}^n$ Is $\mathbb{C}^n$)), this is Theorem 1.9.3. It is worth proving directly, because the identification is itself part of what has to be checked, and because the statement is used both to find the real chains and to compare ranks.

Lemma 10.4.3. Real and Complex Null Spaces of a Real Matrix

Let $\mathbf{M} \in M_n(\mathbb{R})$, let $U = \{\, \mathbf{x} \in \mathbb{R}^n : \mathbf{M}\mathbf{x} = \mathbf{0} \,\}$ be its null space over $\mathbb{R}$, and let $\tilde{U} = \{\, \mathbf{z} \in \mathbb{C}^n : \mathbf{M}\mathbf{z} = \mathbf{0} \,\}$ be its null space over $\mathbb{C}$. Then:

(a) for $\mathbf{z} = \mathbf{x} + i\mathbf{y}$ with $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$: $\mathbf{z} \in \tilde{U}$ if and only if $\mathbf{x} \in U$ and $\mathbf{y} \in U$;

(b) every basis of U over $\mathbb{R}$ is a basis of $\tilde{U}$ over $\mathbb{C}$; in particular $\dim_{\mathbb{R}} U = \dim_{\mathbb{C}} \tilde{U}$ and

rank $\mathbf{M}$ is the same computed over $\mathbb{R}$ and over $\mathbb{C}$.

Proof.

(a) Since $\mathbf{M}$ is real, $\mathbf{Mz} = \mathbf{Mx} + i\mathbf{My}$ with $\mathbf{Mx}, \mathbf{My} \in \mathbb{R}^n$. By the uniqueness of real and imaginary parts (Definition 0.6.1), $\mathbf{Mz} = \mathbf{0}$ if and only if $\mathbf{Mx} = \mathbf{0}$ and $\mathbf{My} = \mathbf{0}$.

(b) Let $(\mathbf{x}_1, \dots, \mathbf{x}_r)$ be a basis of U over $\mathbb{R}$. *Spanning.* Let $\mathbf{z} = \mathbf{x} + i\mathbf{y} \in \tilde{U}$. By (a), $\mathbf{x}, \mathbf{y} \in U$, so $\mathbf{x} = \sum_j c_j \mathbf{x}_j$ and $\mathbf{y} = \sum_j d_j \mathbf{x}_j$ with $c_j, d_j \in \mathbb{R}$; hence $\mathbf{z} = \sum_j (c_j + id_j)\mathbf{x}_j$. *Independence.* Let $\sum_j \alpha_j \mathbf{x}_j = \mathbf{0}$ with $\alpha_j = c_j + id_j \in \mathbb{C}$. Separating real and imaginary parts, which is legitimate because the $\mathbf{x}_j$ are real, gives $\sum_j c_j \mathbf{x}_j = \mathbf{0}$ and $\sum_j d_j \mathbf{x}_j = \mathbf{0}$; independence over $\mathbb{R}$ forces every $c_j = d_j = 0$. So $(\mathbf{x}_1, \dots, \mathbf{x}_r)$ is a basis of $\tilde{U}$ and $\dim_{\mathbb{C}} \tilde{U} = r = \dim_{\mathbb{R}} U$. The rank statement follows from Theorem 3.5.7 over each field.

■

Now the theorem. Read it as the Jordan form with one amendment: a conjugate pair of complex blocks of size k is replaced by a single real block of size $2k$.

Every real matrix is similar, over the reals, to a direct sum of Jordan blocks for its real eigenvalues and real Jordan blocks for its conjugate pairs, and that matrix is unique up to the order of its blocks.

Theorem 10.4.4. Real Jordan Form

Let $\mathbf{A} \in M_n(\mathbb{R})$ with $n \geq 1$. List the eigenvalues of $\mathbf{A}$ in $\mathbb{C}$ as

$$\begin{aligned} &\mu_1, \dots, \mu_p \in \mathbb{R} \quad \text{(the real ones, distinct)}, \\ &\lambda_1, \overline{\lambda_1}, \dots, \lambda_q, \overline{\lambda_q} \quad \text{(the conjugate pairs, with } \operatorname{Im} \lambda_t > 0). \end{aligned}$$

(a) **(Existence)** There is an invertible $\mathbf{P} \in M_n(\mathbb{R})$ such that $\mathbf{P}^{-1}\mathbf{AP}$ is a real Jordan matrix (Definition 10.4.1), in which the blocks for a real eigenvalue μ_l are ordinary Jordan blocks $\mathbf{J}_k(\mu_l)$ and the blocks for a pair $\{\lambda_t, \overline{\lambda_t}\}$ are real Jordan blocks $\mathbf{C}_k(\lambda_t)$.

(b) **(Uniqueness)** In any such matrix, the number of blocks $\mathbf{J}_k(\mu)$ with $k \geq j$, for $\mu \in \mathbb{R}$, and the number of blocks $\mathbf{C}_k(\lambda)$ with $k \geq j$, for $\lambda \notin \mathbb{R}$, both equal

$$\operatorname{rank}(\mathbf{A} - \nu\mathbf{I})^{j-1} - \operatorname{rank}(\mathbf{A} - \nu\mathbf{I})^{j} \qquad (\nu = \mu \text{ resp. } \nu = \lambda),$$

the ranks being computed over $\mathbb{C}$ in the second case. Consequently the real Jordan matrix is determined by $\mathbf{A}$ up to the order of its blocks.

Idea. Work over $\mathbb{C}$, then come home. ① Split $\mathbb{C}^n = \bigoplus G_\nu(\mathbf{A})$ over all complex eigenvalues ν (Theorem 10.1.3). ② For a **real** eigenvalue, the generalized eigenspace is the complexification of a real subspace, by Lemma 10.4.3, and Section 2's structure theorem runs over $\mathbb{R}$ on that real subspace: nothing complex is needed. ③ For a **pair**, choose the chains in G_{λ_t} freely, and take their conjugates in $G_{\overline{\lambda_t}}$ — no choice there. ④ Replace each conjugate pair of chains by its real and imaginary parts; Lemma 10.4.1 says the new matrix is $\mathbf{C}_k(\lambda_t)$, and a count shows the real vectors collected are exactly n in number and independent. Uniqueness is the uniqueness of the complex Jordan form, since by Corollary 10.4.2 a real Jordan matrix has a known complex Jordan form.

Proof. Since $p_{\mathbf{A}} \in \mathbb{R}[x] \subseteq \mathbb{C}[x]$ has degree $n \geq 1$, it splits over $\mathbb{C}$ (Corollary 6.5.5), and its nonreal roots come in conjugate pairs with equal multiplicities (Theorem 9.2.5 (a)); so the list of eigenvalues has the displayed shape. Write $a(\nu)$ for the algebraic multiplicity of ν in $p_{\mathbf{A}}$, so $a(\lambda_t) = a(\overline{\lambda_t})$ and

$$\sum_{l=1}^{p} a(\mu_l) + \sum_{t=1}^{q} 2a(\lambda_t) = n. \tag{$*$}$$

By Theorem 10.1.3 applied to $T_{\mathbf{A}}$ over $\mathbb{C}$,

$$\mathbb{C}^n = \bigoplus_{l=1}^{p} G_{\mu_l}(\mathbf{A}) \;\oplus\; \bigoplus_{t=1}^{q} \left(G_{\lambda_t}(\mathbf{A}) \oplus G_{\overline{\lambda_t}}(\mathbf{A})\right),$$

with $\dim_{\mathbb{C}} G_\nu(\mathbf{A}) = a(\nu)$ and $G_\nu(\mathbf{A}) = \ker(\mathbf{A} - \nu\mathbf{I})^n$ (Corollary 10.1.2 (a)).

Step 1. Real eigenvalues, handled entirely over $\mathbb{R}$. Fix a real eigenvalue $\mu = \mu_l$. The matrix $(\mathbf{A} - \mu\mathbf{I})^n$ is real, so by Lemma 10.4.3 its real null space

$$U_\mu \coloneqq \{\, \mathbf{x} \in \mathbb{R}^n : (\mathbf{A} - \mu\mathbf{I})^n \mathbf{x} = \mathbf{0} \,\}$$

has a real basis which is a $\mathbb{C}$-basis of $G_\mu(\mathbf{A})$; in particular $\dim_{\mathbb{R}} U_\mu = a(\mu)$. The subspace U_μ is $\mathbf{A}$-invariant, since $(\mathbf{A}-\mu\mathbf{I})^n(\mathbf{A}\mathbf{x}) = \mathbf{A}(\mathbf{A}-\mu\mathbf{I})^n\mathbf{x} = \mathbf{0}$ for $\mathbf{x} \in U_\mu$ (Theorem 6.7.4 (b)), and $N_\mu \coloneqq (\mathbf{A}-\mu\mathbf{I})|_{U_\mu}$ satisfies $N_\mu^n = 0$, so N_μ is nilpotent (Definition 10.2.1). By Theorem 10.2.3 there is a basis $\mathcal{B}_\mu$ of the **real** space U_μ, a concatenation of Jordan chains of N_μ, with $[N_\mu]_{\mathcal{B}_\mu}^{\mathcal{B}_\mu} = \mathbf{J}_{k_1}(0) \oplus \cdots \oplus \mathbf{J}_{k_s}(0)$. Since $T_{\mathbf{A}}|_{U_\mu} = \mu\,\mathrm{id}_{U_\mu} + N_\mu$ and $[\mathrm{id}_{U_\mu}]_{\mathcal{B}_\mu}^{\mathcal{B}_\mu} = \mathbf{I}_{k_1} \oplus \cdots \oplus \mathbf{I}_{k_s}$, adding the matrices (Theorem 4.1.4) blockwise (Theorem 8.1.4 (a), Definition 10.3.1) gives

$$[T_{\mathbf{A}}|_{U_\mu}]_{\mathcal{B}_\mu}^{\mathcal{B}_\mu} = \mathbf{J}_{k_1}(\mu) \oplus \cdots \oplus \mathbf{J}_{k_s}(\mu),$$

a direct sum of ordinary Jordan blocks. Moreover $\mathcal{B}_\mu$, viewed in $\mathbb{C}^n$, is a Jordan basis of $G_\mu(\mathbf{A})$ for $T_{\mathbf{A}}$ over $\mathbb{C}$: the chain relations are the same equations, and $\mathcal{B}_\mu$ is a $\mathbb{C}$-basis of $G_\mu(\mathbf{A})$ by Lemma 10.4.3 (b).

Step 2. Conjugate pairs, chosen on one side. Fix t and write $\lambda = \lambda_t$, $G = G_\lambda(\mathbf{A})$, $\overline{G} = G_{\overline{\lambda}}(\mathbf{A})$. By Theorem 10.1.3 (d), $(\mathbf{A} - \lambda\mathbf{I})|_G$ is nilpotent, so Theorem 10.2.3 provides chains whose concatenation is a $\mathbb{C}$-basis of G; as in Step 1 these are Jordan chains of $\mathbf{A}$ for λ, of lengths $k_1, \ldots, k_m$ summing to $a(\lambda)$.

Claim: conjugating gives a Jordan basis of $\overline{G}$. Indeed, conjugation maps G onto $\overline{G}$: if $(\mathbf{A} - \lambda\mathbf{I})^n\mathbf{z} = \mathbf{0}$, then $(\mathbf{A} - \overline{\lambda}\mathbf{I})^n\overline{\mathbf{z}} = \mathbf{0}$ by the computation in Lemma 10.4.1 (b), and conjugating twice returns $\mathbf{z}$. If $(\mathbf{z}_1, \ldots, \mathbf{z}_d)$ is a $\mathbb{C}$-basis of G, then $(\overline{\mathbf{z}_1}, \ldots, \overline{\mathbf{z}_d})$ spans $\overline{G}$ — any $\mathbf{z} \in \overline{G}$ has $\overline{\mathbf{z}} \in G$, say $\overline{\mathbf{z}} = \sum_j c_j\mathbf{z}_j$, whence $\mathbf{z} = \sum_j \overline{c_j}\,\overline{\mathbf{z}_j}$ — and it has $d = \dim_{\mathbb{C}} \overline{G}$ entries, so it is a basis (Theorem 1.6.4 (b)). Applying this to the concatenated chains, and using Lemma 10.4.1 (b) chain by chain, the conjugated chains form a Jordan basis of $\overline{G}$ consisting of chains for $\overline{\lambda}$ of the same lengths $k_1, \ldots, k_m$.

Step 3. Realify each pair of conjugate chains. Keep t fixed and let $(\mathbf{w}_1, \ldots, \mathbf{w}_k)$ be one of the chains of Step 2, with $\mathbf{w}_j = \mathbf{u}_j - i\mathbf{v}_j$, $\mathbf{u}_j, \mathbf{v}_j \in \mathbb{R}^n$. By Lemma 10.4.1 (c), the $2k$ real vectors $\mathbf{u}_1, \mathbf{v}_1, \ldots, \mathbf{u}_k, \mathbf{v}_k$ have the same $\mathbb{C}$-span as the $2k$ chain vectors $\mathbf{w}_1, \ldots, \mathbf{w}_k, \overline{\mathbf{w}_1}, \ldots, \overline{\mathbf{w}_k}$. Do this for each of the m chains attached to λ_t, and for each t.

Step 4. The collected real vectors form a basis of $\mathbb{R}^n$. Let $\mathcal{B}$ be the list obtained by writing, one after another: the real Jordan bases $\mathcal{B}_{\mu_1}, \ldots, \mathcal{B}_{\mu_p}$ of Step 1, and then, for each

t and each chain of Step 2, the block of $2k$ real vectors of Step 3. Its length is

$$\sum_{l=1}^{p} a(\mu_l) + \sum_{t=1}^{q} \sum_{\text{chains}} 2k = \sum_{l=1}^{p} a(\mu_l) + \sum_{t=1}^{q} 2a(\lambda_t) = n$$

by $(*)$, since the chain lengths for λ_t sum to $a(\lambda_t)$. Its $\mathbb{C}$-span contains a $\mathbb{C}$-basis of each $G_{\mu_l}(\mathbf{A})$ (Step 1) and, for each t, all the vectors $\mathbf{w}_j$ and $\overline{\mathbf{w}_j}$ (Step 3), which together form a $\mathbb{C}$-basis of $G_{\lambda_t}(\mathbf{A}) \oplus G_{\overline{\lambda_t}}(\mathbf{A})$ by Step 2. So the $\mathbb{C}$-span of $\mathcal{B}$ is all of $\mathbb{C}^n$, and $\mathcal{B}$ has exactly n entries, so $\mathcal{B}$ is a $\mathbb{C}$-basis of $\mathbb{C}^n$ (Theorem 1.6.4 (b)). In particular $\mathcal{B}$ is linearly independent over $\mathbb{C}$, hence over $\mathbb{R}$, and being a list of n independent vectors in $\mathbb{R}^n$ it is a basis of $\mathbb{R}^n$ (Theorem 1.6.4 (a)).

Step 5. The matrix. Each segment of $\mathcal{B}$ spans an $\mathbf{A}$-invariant real subspace: the spaces U_{μ_l} by Step 1, and the spaces U of Lemma 10.4.1 (d) by Step 3 and the independence just proved. Their direct sum is $\mathbb{R}^n$, since $\mathcal{B}$ is a basis. By Theorem 9.1.5, $[T_{\mathbf{A}}]_{\mathcal{B}}^{\mathcal{B}}$ is block diagonal with the matrices of the restrictions as diagonal blocks, which are $\mathbf{J}_k(\mu_l)$'s by Step 1 and $\mathbf{C}_k(\lambda_t)$'s by Lemma 10.4.1 (d). Taking $\mathbf{P}$ to be the real matrix whose columns are the vectors of $\mathcal{B}$ gives $\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = [T_{\mathbf{A}}]_{\mathcal{B}}^{\mathcal{B}}$ (Theorem 4.2.4), which proves (a).

(b) Let $\mathbf{B} = \mathbf{P}^{-1}\mathbf{A}\mathbf{P}$ be any real Jordan matrix similar to $\mathbf{A}$ over $\mathbb{R}$, and regard the similarity over $\mathbb{C}$. By Corollary 10.4.2 and Theorem 8.1.4, replacing each block $\mathbf{C}_k(\lambda)$ of $\mathbf{B}$ by $\mathbf{J}_k(\lambda) \oplus \mathbf{J}_k(\overline{\lambda})$ produces a Jordan matrix similar to $\mathbf{B}$, hence to $\mathbf{A}$, over $\mathbb{C}$. By Theorem 10.3.1 (b), for every $\nu \in \mathbb{C}$ and $j \geq 1$ the number of blocks $\mathbf{J}_k(\nu)$ with $k \geq j$ in that Jordan matrix equals $\operatorname{rank}(\mathbf{A} - \nu\mathbf{I})^{j-1} - \operatorname{rank}(\mathbf{A} - \nu\mathbf{I})^{j}$, computed over $\mathbb{C}$. For a real μ, the blocks $\mathbf{J}_k(\mu)$ of the Jordan matrix are exactly the blocks $\mathbf{J}_k(\mu)$ of $\mathbf{B}$, since the replaced blocks contribute only nonreal eigenvalues; and the ranks over $\mathbb{C}$ agree with the ranks over $\mathbb{R}$ by Lemma 10.4.3 (b), applied to the real matrix $(\mathbf{A} - \mu\mathbf{I})^j$. For $\lambda \notin \mathbb{R}$ with $\operatorname{Im}\lambda > 0$, the blocks $\mathbf{J}_k(\lambda)$ of the Jordan matrix come one each from the blocks $\mathbf{C}_k(\lambda)$ of $\mathbf{B}$: a block $\mathbf{C}_k(\lambda')$ of $\mathbf{B}$ with $\lambda' \neq \lambda$ contributes $\mathbf{J}_k(\lambda')$ and $\mathbf{J}_k(\overline{\lambda'})$, and $\overline{\lambda'} \neq \lambda$ because λ and λ' both have **positive** imaginary part (Definition 10.4.1). In both cases the count is determined by $\mathbf{A}$, so any two real Jordan matrices similar to $\mathbf{A}$ have the same blocks up to order.

■

So over $\mathbb{R}$ the dream of Chapter 2 comes true in the amended form: **every real square matrix is similar over $\mathbb{R}$ to a matrix that is block diagonal with blocks $\mathbf{J}_k(\mu)$ and $\mathbf{C}_k(\lambda)$**, and the only new feature is the 2×2 rotation-scaling block. Counting: a real eigenvalue μ contributes $a(\mu)$ to the size and $g(\mu)$ blocks; a pair $\{\lambda, \overline{\lambda}\}$ contributes $2a(\lambda)$ to the size and $g(\lambda)$ real blocks. For an operator T on a real vector space V the same statement holds, applied to $[T]_{\mathcal{C}}^{\mathcal{C}}$ for any basis $\mathcal{C}$ of V: the theorem produces a real Jordan matrix similar to it, hence equal to $[T]_{\mathcal{B}}^{\mathcal{B}}$ for another basis $\mathcal{B}$ of V (Theorem 4.2.4). The complexification $V_{\mathbb{C}}$ of Definition 1.9.2 is where the complex chains live; for $V = \mathbb{R}^n$ it is $\mathbb{C}^n$ (Example (The Complexification of $\mathbb{R}^n$ Is $\mathbb{C}^n$)), which is the case written out above.

Warning. **A real Jordan matrix is not triangular, and the real form is not obtained by "ignoring" the imaginary parts.** The real Jordan form of a matrix with a nonreal eigenvalue **cannot** be triangular, since a triangular real matrix has its diagonal entries as eigenvalues, all real (Theorem 9.7.4). Nor is $\mathbf{\Lambda}(\lambda)$ a diagonal matrix in disguise: it has no real eigenvector. And the real basis is not the real part of the complex one: each complex

chain vector $\mathbf{w}_j$ contributes **two** real vectors, $\operatorname{Re}\mathbf{w}_j$ and $-\operatorname{Im}\mathbf{w}_j$, and dropping either of them destroys the invariance.

Example. The Real Jordan Form of a 4×4 Matrix

Let

$$\mathbf{A}=\begin{pmatrix} 0 & -1 & 1 & 0\\ 1 & 0 & 0 & 1\\ -1 & 0 & 2 & -1\\ 0 & -1 & 1 & 2 \end{pmatrix}\in M_4(\mathbb{R}),\qquad p_{\mathbf{A}}=(x^2-2x+2)^2.$$

Find the real Jordan form of $\mathbf{A}$ (Theorem 10.4.4) and an invertible $\mathbf{P}\in M_4(\mathbb{R})$ with $\mathbf{P}^{-1}\mathbf{AP}$ in that form.

Solution. The eigenvalues. $x^2-2x+2=(x-1)^2+1$ has roots $1\pm i$, so over $\mathbb{C}$ the eigenvalues are $\lambda=1+i$ and $\overline{\lambda}=1-i$, each with $a(\lambda)=2$. There is no real eigenvalue, so the real Jordan form consists of real blocks $\mathbf{C}_k(1+i)$ only, with sizes k summing to $a(\lambda)=2$.

A chain for $\lambda=1+i$. Put $\mathbf{M}=\mathbf{A}-(1+i)\mathbf{I}$, so that

$$\mathbf{M}=\begin{pmatrix} -1-i & -1 & 1 & 0\\ 1 & -1-i & 0 & 1\\ -1 & 0 & 1-i & -1\\ 0 & -1 & 1 & 1-i \end{pmatrix}.$$

Solve $\mathbf{Mz}=\mathbf{0}$ using rows 2, 1, and 4 in that order. Row 2 gives $z_1=(1+i)z_2-z_4$; row 1 gives $z_3=(1+i)z_1+z_2$. Substituting the first into the second, and writing $s=z_2$, $u=z_4$,

$$z_1=(1+i)s-u,\qquad z_3=(1+i)\big((1+i)s-u\big)+s=(1+2i)s-(1+i)u,$$

using $(1+i)^2=2i$. Row 4 reads $-z_2+z_3+(1-i)z_4=0$, which becomes

$$-s+(1+2i)s-(1+i)u+(1-i)u=2is-2iu=0,$$

so $s=u$. Row 3 is then automatic: with $s=u=1$ we get $\mathbf{z}=(i,1,i,1)$, and row 3 gives $-i+(1-i)i-1=-i+i+1-1=0$. So

$$\ker\mathbf{M}=\operatorname{span}\big((i,1,i,1)\big),\qquad g(\lambda)=1.$$

Hence there is exactly one λ-block over $\mathbb{C}$ (Corollary 10.3.2 (b)), of size $a(\lambda)=2$, and the real Jordan form is $\mathbf{C}_2(1+i)$. For the basis we need the chain: take $\mathbf{w}_1=(i,1,i,1)$ and solve $\mathbf{Mw}_2=\mathbf{w}_1$. Try $\mathbf{w}_2=(0,0,z_3,z_4)$, so that only columns 3 and 4 of $\mathbf{M}$ are used. Row 1 then reads $z_3=i$ and row 2 reads $z_4=1$, so $\mathbf{w}_2=(0,0,i,1)$ is the only candidate of this shape. Rows 3 and 4 confirm it:

$$(1-i)i-1=(i+1)-1=i\quad\text{and}\quad i+(1-i)=1,$$

which are the third and fourth entries of $\mathbf{w}_1$. So $(\mathbf{w}_1,\mathbf{w}_2)$ is a Jordan chain of $\mathbf{A}$ for $1+i$.

Normalizing the chain. If $(\mathbf{w}_1, \mathbf{w}_2)$ is a chain, so is $(c\mathbf{w}_1, c\mathbf{w}_2)$ for any $c \neq 0$, since the chain relations are linear. Taking $c = -i$ clears the denominators of the realification:

$$\mathbf{w}_1' = -i(i, 1, i, 1) = (1, -i, 1, -i), \qquad \mathbf{w}_2' = -i(0, 0, i, 1) = (0, 0, 1, -i).$$

Now read off $\mathbf{w}_j' = \mathbf{u}_j - i\mathbf{v}_j$ with $\mathbf{u}_j, \mathbf{v}_j$ real, as Lemma 10.4.1 requires:

$$\mathbf{u}_1 = (1, 0, 1, 0), \quad \mathbf{v}_1 = (0, 1, 0, 1), \qquad \mathbf{u}_2 = (0, 0, 1, 0), \quad \mathbf{v}_2 = (0, 0, 0, 1).$$

The answer. With $\mathcal{B} = (\mathbf{u}_1, \mathbf{v}_1, \mathbf{u}_2, \mathbf{v}_2)$,

$$\mathbf{P} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}, \qquad \mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \mathbf{C}_2(1+i) = \begin{pmatrix} 1 & -1 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 1 & 1 \end{pmatrix}.$$

$\mathbf{P}$ is lower triangular with ones on the diagonal, so $\det \mathbf{P} = 1$ (Theorem 7.4.6) and $\mathbf{P}$ is invertible.

The check, column by column. Writing $\mathbf{c}_1, \dots, \mathbf{c}_4$ for the columns of $\mathbf{P}$, the identity $\mathbf{AP} = \mathbf{P}\,\mathbf{C}_2(1+i)$ says

$$\begin{aligned} \mathbf{A}\mathbf{c}_1 &= \mathbf{c}_1 + \mathbf{c}_2, \\ \mathbf{A}\mathbf{c}_2 &= -\mathbf{c}_1 + \mathbf{c}_2, \\ \mathbf{A}\mathbf{c}_3 &= \mathbf{c}_1 + \mathbf{c}_3 + \mathbf{c}_4, \\ \mathbf{A}\mathbf{c}_4 &= \mathbf{c}_2 - \mathbf{c}_3 + \mathbf{c}_4, \end{aligned}$$

read off the four columns of $\mathbf{C}_2(1+i)$; these are exactly the identities of Lemma 10.4.1 (a) with $a = b = 1$. Two of them in full:

$$\begin{aligned} \mathbf{A}\mathbf{c}_1 = \mathbf{A}(1, 0, 1, 0) &= (0+1,\ 1+0,\ -1+2,\ 0+1) = (1, 1, 1, 1) \\ &= (1, 0, 1, 0) + (0, 1, 0, 1) = \mathbf{c}_1 + \mathbf{c}_2, \end{aligned}$$

adding columns 1 and 3 of $\mathbf{A}$; and

$$\begin{aligned} \mathbf{A}\mathbf{c}_3 = \mathbf{A}(0, 0, 1, 0) &= (1, 0, 2, 1) = (1, 0, 1, 0) + (0, 0, 1, 0) + (0, 0, 0, 1) \\ &= \mathbf{c}_1 + \mathbf{c}_3 + \mathbf{c}_4, \end{aligned}$$

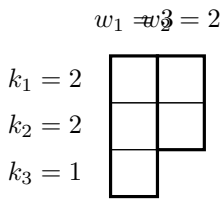
the third column of $\mathbf{A}$. The remaining two are the same kind of one-line computation.

Cross-checks. $\operatorname{tr} \mathbf{A} = 0+0+2+2 = 4$ and $\operatorname{tr} \mathbf{C}_2(1+i) = 4 \cdot 1 = 4$. Over $\mathbb{C}$, Corollary 10.4.2 gives the complex Jordan form $\mathbf{J}_2(1+i) \oplus \mathbf{J}_2(1-i)$, whose characteristic polynomial is $\big((x-1)^2 + 1\big)^2 = p_{\mathbf{A}}$.

10.4.3 The Weyr form

The Jordan form of a nilpotent operator is the Young diagram of Section 2 read **by rows**: one row per chain, one block per row. The same diagram can be read **by columns**, and the columns are the jumps of the kernel chain (Definition 10.2.3 and the discussion after Theorem 10.2.4). Reordering a Jordan basis accordingly costs nothing and gives a second normal form, the Weyr form. It carries exactly the same information — a partition for each eigenvalue — but it wears it differently, and for one purpose it is strictly better: the matrices that commute with a Weyr matrix are block upper triangular, which is false for the Jordan form.

Here is the diagram for the partition $(2,2,1)$ of 5, whose dual partition is $(3,2)$:



rows: the Jordan block sizes $(2,2,1)$
columns: the Weyr structure $(3,2)$
the operator moves each box one step left

The basis is the same; only its order changes. Instead of listing chain by chain, list the first column (a basis of the kernel), then the second column, and so on. Since the operator moves each box one step to the left, it maps the j-th group of basis vectors into the $(j-1)$-th, injectively, and the matrix becomes block upper triangular with a very rigid pattern of ones.

Definition 10.4.2. Basic Weyr Matrix, Weyr Structure, Weyr Form

Let $\lambda \in F$ and let $w_1 \geq w_2 \geq \cdots \geq w_s \geq 1$ be a partition of an integer $m \geq 1$ (Definition 10.2.3). The **basic Weyr matrix** with eigenvalue λ and **Weyr structure** $(w_1, \ldots, w_s)$ is the matrix $\mathbf{W} \in M_m(F)$, partitioned into blocks according to $m = w_1 + \cdots + w_s$, whose blocks are

$$\mathbf{W}_{jj} = \lambda \mathbf{I}_{w_j}, \qquad \mathbf{W}_{j,j+1} = \begin{pmatrix} \mathbf{I}_{w_{j+1}} \\ \mathbf{O} \end{pmatrix} \in M_{w_j \times w_{j+1}}(F), \qquad \mathbf{W}_{jl} = \mathbf{O} \text{ otherwise.}$$

A **Weyr matrix** is a direct sum of basic Weyr matrices. A **Weyr form** of $T \in \mathcal{L}(V)$ is a Weyr matrix $\mathbf{W}$ with $[T]_{\mathcal{B}}^{\mathcal{B}} = \mathbf{W}$ for some basis $\mathcal{B}$ of V; for $\mathbf{A} \in M_n(F)$, a Weyr form of $\mathbf{A}$ is a Weyr matrix similar to $\mathbf{A}$.

In words: the diagonal blocks are scalar, of the sizes given by the partition, and the only other nonzero blocks sit **immediately to the right** of the diagonal ones, each being an identity matrix stacked on top of a zero block — so $\mathbf{W}_{j,j+1}$ has full column rank w_{j+1}, which is possible exactly because $w_j \geq w_{j+1}$. The matrix is upper triangular, with λ repeated m times on the diagonal.

Examples.

- **One column.** If $s = 1$, then $\mathbf{W} = \lambda \mathbf{I}_{w_1}$: a scalar matrix. This is the diagonalizable case.
- **All parts equal to 1.** If $w_1 = \cdots = w_s = 1$, then $\mathbf{W} = \mathbf{J}_s(\lambda)$: a single Jordan block. So Jordan blocks are the basic Weyr matrices of the thinnest possible structure.
- **The structure $(3,2)$ with $\lambda = 5$.** Here $m = 5$ and

$$\mathbf{W} = \begin{pmatrix} 5 & 0 & 0 & 1 & 0 \\ 0 & 5 & 0 & 0 & 1 \\ 0 & 0 & 5 & 0 & 0 \\ 0 & 0 & 0 & 5 & 0 \\ 0 & 0 & 0 & 0 & 5 \end{pmatrix},$$

  with the 3×2 block $\begin{pmatrix} \mathbf{I}_2 \\ \mathbf{O} \end{pmatrix}$ in the upper right corner. As we will see, this is the Weyr form of $\mathbf{J}_2(5) \oplus \mathbf{J}_2(5) \oplus \mathbf{J}_1(5)$, the diagram drawn above.

Nonexample by minimal change. Put the identity block in the **lower** part of $\mathbf{W}_{12}$ instead:

with structure $(2,1)$ and $\lambda = 0$, compare

$$\mathbf{W} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \qquad \text{and} \qquad \mathbf{W}' = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}.$$

Both are nilpotent of rank 1, and indeed $\mathbf{W} \sim \mathbf{W}'$ (Corollary 10.2.5): the shape is not an extra invariant, and the two matrices represent the same operator in different bases. But only $\mathbf{W}$ is a Weyr matrix. The clause that fails for $\mathbf{W}'$ is that the block $\mathbf{W}_{j,j+1}$ must be an identity matrix stacked **on top of** a zero block, not below it. The ordering is a normalization, and it is the one the commutant property below is proved with: there the block $\mathbf{W}_{j,j+1}$ is canceled on the left, which uses that its columns are the **first** w_{j+1} standard basis vectors, in order. Concretely, every matrix commuting with $\mathbf{W}$ turns out to be upper triangular, while the commutant of $\mathbf{W}'$ contains the matrix unit $\mathbf{E}_{21}$.

The existence and uniqueness of the form are now a repackaging of Section 3.

Theorem 10.4.5. Weyr Form

Let V be a finite-dimensional vector space over F with $\dim V = n \geq 1$, and let $T \in \mathcal{L}(V)$.

(a) **(Existence)** T has a Weyr form if and only if p_T splits over F. One can be arranged with one basic Weyr block per eigenvalue.

(b) **(Uniqueness)** In any Weyr form of T with one basic block per eigenvalue, the block for λ has Weyr structure

$$w_j = \dim\ker(T - \lambda\ \mathrm{id}_V)^j - \dim\ker(T - \lambda\ \mathrm{id}_V)^{j-1} \qquad (j \geq 1),$$

the nonzero terms of this sequence; equivalently $(w_1, w_2, \dots)$ is the **dual partition** of the partition formed by the sizes of the λ-blocks of the Jordan form of T. So the Weyr form is determined by T up to the order of its basic blocks.

Idea. Both parts come from one observation: a Weyr basis is a Jordan basis in a different order. Arrange the λ-chains of a Jordan basis as the rows of the Young diagram, sorted by decreasing length, and read the diagram **column by column**. The operator $T - \lambda$ id moves each box one step to the left, so it carries the j-th column into the $(j-1)$-th, matching the i-th box of one with the i-th box of the other — which is exactly the block $\begin{pmatrix} \mathbf{I} \\ \mathbf{O} \end{pmatrix}$. For uniqueness, count the boxes: the first j columns of the diagram are a basis of $\ker(T - \lambda\ \mathrm{id})^j$, so the column lengths are the jumps of the kernel chain.

Proof. ($\Rightarrow$) A Weyr matrix is upper triangular, so if T has a Weyr form then T is triangularizable and p_T splits (Definition 9.7.1, Theorem 9.7.2).

($\Leftarrow$) Suppose p_T splits, and let λ be an eigenvalue. By Theorem 10.3.1 there is a Jordan basis; fix its part inside $G_\lambda(T)$, a concatenation of Jordan chains of $N = T - \lambda\ \mathrm{id}_V$ of lengths $k_1 \geq k_2 \geq \cdots \geq k_r \geq 1$ summing to $a(\lambda)$ (Theorem 10.2.3, after reindexing so that the lengths decrease). Write $b_{i,1}, \dots, b_{i,k_i}$ for the i-th chain in the order of Definition 10.2.2, so that

$$Nb_{i,1} = 0, \qquad Nb_{i,j} = b_{i,j-1} \quad (2 \leq j \leq k_i).$$

Put $s = k_1$ and, for $1 \le j \le s$, let $w_j = \#\{\, i : k_i \ge j \,\}$, the length of the j-th column of the Young diagram; then $w_1 \ge \cdots \ge w_s \ge 1$ is the dual partition of $(k_1, \dots, k_r)$ (Definition 10.2.3), a partition of $a(\lambda)$. Because the k_i decrease, $\{\, i : k_i \ge j \,\} = \{1, 2, \dots, w_j\}$. Let

$$\mathcal{C}_j = (b_{1,j}, b_{2,j}, \dots, b_{w_j,j})$$

be the j-th column, and let $\mathcal{B}_\lambda$ be the list $\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_s$ written one after another: a reordering of the chosen basis of $G_\lambda(T)$, hence a basis of it.

Read the columns of $[T|_{G_\lambda}]^{\mathcal{B}_\lambda}_{\mathcal{B}_\lambda}$. For $j = 1$, $Tb_{i,1} = \lambda b_{i,1}$. For $j \ge 2$, $Tb_{i,j} = \lambda b_{i,j} + b_{i,j-1}$, and $b_{i,j-1}$ is the i-th entry of $\mathcal{C}_{j-1}$, which exists because $i \le w_j \le w_{j-1}$. In the block partition of $\mathcal{B}_\lambda$ by the groups $\mathcal{C}_1, \dots, \mathcal{C}_s$, this says: the (j, j) block is $\lambda \mathbf{I}_{w_j}$; the $(j-1, j)$ block sends the i-th basis vector of group j to the i-th of group $j-1$ for $i = 1, \dots, w_j$, so it is $\begin{pmatrix} \mathbf{I}_{w_j} \\ \mathbf{O} \end{pmatrix}$; and every other block is $\mathbf{O}$. That is the basic Weyr matrix with eigenvalue λ and structure $(w_1, \dots, w_s)$ (Definition 10.4.2). Writing the lists $\mathcal{B}_\lambda$ for the distinct eigenvalues one after another and using $V = \bigoplus_\lambda G_\lambda(T)$ (Theorem 10.1.3) with Theorem 9.1.5, the resulting basis $\mathcal{B}$ makes $[T]^{\mathcal{B}}_{\mathcal{B}}$ a direct sum of basic Weyr matrices, one per eigenvalue.

(b) Let $\mathbf{W}$ be a Weyr form of T with one basic block per eigenvalue, and fix λ with basic block of structure $(w_1, \dots, w_s)$ and size m. Let $\mathcal{B}$ be a basis with $[T]^{\mathcal{B}}_{\mathcal{B}} = \mathbf{W}$, and name the basis vectors inside the λ-block $c_{i,j}$, where j indexes the groups and $i = 1, \dots, w_j$ the entries within a group. The blocks of Definition 10.4.2 say precisely that

$$Nc_{i,1} = 0, \qquad Nc_{i,j} = c_{i,j-1} \quad (j \ge 2), \qquad N = T - \lambda \operatorname{id}_V,$$

while on the span V' of the basis vectors belonging to the other basic blocks the operator N is invertible: V' is T-invariant and the matrix of $N|_{V'}$ is $\mathbf{W}' - \lambda \mathbf{I}$, where $\mathbf{W}'$ is a direct sum of basic Weyr matrices with eigenvalues $\mu \ne \lambda$, hence upper triangular with all diagonal entries $\mu - \lambda \ne 0$, so of nonzero determinant (Theorem 7.4.6, Theorem 7.5.2). Hence N^j kills $c_{i,l}$ for $l \le j$ and sends $c_{i,l} \mapsto c_{i,l-j}$ for $l > j$. Write $V = U \oplus V'$ with $U = \operatorname{span}(c_{i,l})$; both summands are N-invariant, so $\ker N^j = (\ker N^j \cap U) \oplus (\ker N^j \cap V')$, and the second summand is $\{0\}$ because $N^j|_{V'}$ is invertible. Inside U the displayed action gives

$$\ker N^j = \operatorname{span}\,(c_{i,l} : l \le j), \qquad \dim \ker N^j = w_1 + \cdots + w_j$$

for $0 \le j \le s$, the spanning vectors being members of the basis $\mathcal{B}$ and hence independent; and $\ker N^j = \ker N^s$ for $j \ge s$. Subtracting consecutive values gives $w_j = \dim \ker N^j - \dim \ker N^{j-1}$ as claimed, a number depending only on T and λ. By Corollary 10.3.2 (d) this is also $\#\{i : k_i \ge j\}$ for the λ-block sizes k_i of the Jordan form, that is, the dual partition. Since the structures are determined, so is $\mathbf{W}$ up to the order in which its basic blocks are listed.

■

Existence and uniqueness are thus free; what makes the form worth the trouble is the next property. Here $\mathbf{K} \in M_n(F)$ is called **block upper triangular** for a partition $n = n_1 + \cdots + n_s$, taken as both the row and the column partition (Definition 8.1.1), when every block $\mathbf{K}_{pq}$ with $p > q$ is zero; for $s = 2$ this is the shape of Chapter 8.

Proposition 10.4.6. The Commutant of a Basic Weyr Matrix Is Block Triangular

Let $\mathbf{W} \in M_m(F)$ be a basic Weyr matrix with eigenvalue λ and Weyr structure $(w_1, \dots, w_s)$, and let $\mathbf{K} \in M_m(F)$ satisfy $\mathbf{WK} = \mathbf{KW}$. Then $\mathbf{K}$ is block upper triangular for the partition $m = w_1 + \cdots + w_s$: its blocks $\mathbf{K}_{pq}$ vanish for $p > q$.

Idea. Write both products blockwise. Because $\mathbf{W}$ has only one nonzero block in each block row, each block equation is short: $\mathbf{E}_p \mathbf{K}_{p+1,q} = \mathbf{K}_{p,q-1} \mathbf{E}_{q-1}$, where $\mathbf{E}_j = \begin{pmatrix} \mathbf{I}_{w_{j+1}} \\ \mathbf{O} \end{pmatrix}$. Every $\mathbf{E}_j$ has full column rank, so it can be canceled on the left. Now sweep the block columns from left to right: the first block column forces $\mathbf{K}_{p1} = \mathbf{O}$ for $p > 1$, and each new column is forced by the previous one.

Proof. Put $\mathbf{N} = \mathbf{W} - \lambda \mathbf{I}_m$ and $\mathbf{E}_j = \begin{pmatrix} \mathbf{I}_{w_{j+1}} \\ \mathbf{O} \end{pmatrix} \in M_{w_j \times w_{j+1}}(F)$ for $1 \le j \le s-1$, so that $\mathbf{N}$ has blocks $\mathbf{N}_{j,j+1} = \mathbf{E}_j$ and all other blocks zero (Definition 10.4.2). Since $\lambda \mathbf{I}_m$ commutes with everything, $\mathbf{WK} = \mathbf{KW}$ is equivalent to $\mathbf{NK} = \mathbf{KN}$. Comparing the (p, q) blocks with Theorem 8.1.1, and reading an undefined symbol as $\mathbf{O}$,

$$\mathbf{E}_p \, \mathbf{K}_{p+1,q} = (\mathbf{NK})_{pq} = (\mathbf{KN})_{pq} = \mathbf{K}_{p,q-1} \, \mathbf{E}_{q-1} \qquad \text{for all } 1 \le p, q \le s. \tag{$*$}$$

Each $\mathbf{E}_j$ is injective as a map $F^{w_{j+1}} \to F^{w_j}$, so $\mathbf{E}_j \mathbf{X} = \mathbf{O}$ forces $\mathbf{X} = \mathbf{O}$ for every matrix $\mathbf{X}$ with w_{j+1} rows.

We prove $\mathbf{K}_{pq} = \mathbf{O}$ for $p > q$ by induction on $q \ge 1$.

Case $q = 1$. Let $p > 1$. Taking the indices $(p-1, 1)$ in $(*)$, the right-hand side involves $\mathbf{K}_{p-1,0}$, which is undefined and therefore read as $\mathbf{O}$. So $\mathbf{E}_{p-1} \mathbf{K}_{p,1} = \mathbf{O}$, and $p - 1 \le s - 1$ makes $\mathbf{E}_{p-1}$ one of the listed blocks; injectivity gives $\mathbf{K}_{p,1} = \mathbf{O}$.

Induction step. Let $q \ge 2$ and assume $\mathbf{K}_{p'q'} = \mathbf{O}$ whenever $p' > q'$ and $q' < q$. Let $p > q$. Taking the indices $(p-1, q)$ in $(*)$,

$$\mathbf{E}_{p-1} \mathbf{K}_{pq} = \mathbf{K}_{p-1,q-1} \mathbf{E}_{q-1}.$$

Here $p - 1 > q - 1$ and $q - 1 < q$, so $\mathbf{K}_{p-1,q-1} = \mathbf{O}$ by the induction hypothesis, and the right-hand side is $\mathbf{O}$. Again $1 \le p - 1 \le s - 1$, so injectivity of $\mathbf{E}_{p-1}$ gives $\mathbf{K}_{pq} = \mathbf{O}$. This completes the induction and the proof. ∎

Remark. The Jordan form has no such property, which is the reason the Weyr form exists at all. Take $\mathbf{J} = \mathbf{J}_2(0) \oplus \mathbf{J}_1(0)$ and $\mathbf{W}$ the Weyr matrix of the same operator, of structure $(2, 1)$:

$$\mathbf{J} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \qquad \mathbf{W} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Solving $\mathbf{JK} = \mathbf{KJ}$ entry by entry gives the commutant $\left\{ \begin{pmatrix} a & b & c \\ 0 & a & 0 \\ 0 & h & i \end{pmatrix} \right\}$, which contains matrices with a nonzero entry **below** the diagonal, in position $(3,2)$: the commutant of $\mathbf{J}$ is not contained in the upper triangular matrices, and in particular it is not block upper triangular for the Weyr partition $3 = 2+1$ of this operator. Solving $\mathbf{WK} = \mathbf{KW}$ gives $\left\{ \begin{pmatrix} a & b & c \\ 0 & e & f \\ 0 & 0 & a \end{pmatrix} \right\}$, every member of which is upper triangular, in agreement with Proposition 10.4.6. This is the property that makes the Weyr form the right one when two commuting operators must be brought to normal form at the same time.

Quick check. An operator T on a 7-dimensional complex space has one eigenvalue λ, with Jordan block sizes $(3,2,1,1)$. What is the Weyr structure of its Weyr form, and what is $\dim\ker(T-\lambda\,\mathrm{id})^2$?

Solution. The Weyr structure is the dual partition: $w_1 = \#\{i : k_i \geq 1\} = 4$, $w_2 = \#\{i : k_i \geq 2\} = 2$, $w_3 = \#\{i : k_i \geq 3\} = 1$, so $(4,2,1)$, a partition of 7 as it must be. By Theorem 10.4.5 (b), $\dim\ker(T-\lambda\,\mathrm{id})^2 = w_1 + w_2 = 6$. (Equivalently, $\sum_i \min(k_i, 2) = 2+2+1+1 = 6$, by Corollary 10.3.2 (d).)

10.4.4 Exercises

10.4.4.1 A. Check your understanding

Exercise. A1

(a) Write down $\mathbf{\Lambda}(\lambda)$ for $\lambda = a + bi$ and $\mathbf{C}_2(\lambda)$, and say what their sizes are.

(b) State the Real Jordan Form Theorem.

(c) True or false: every matrix in $M_n(\mathbb{R})$ is similar over $\mathbb{R}$ to an upper triangular real matrix. Justify your answer.

(d) Let $\mathbf{A} \in M_6(\mathbb{R})$ have complex eigenvalues $2, 2, 1+i, 1-i, 1+i, 1-i$. What are the possible real Jordan forms?

(e) Define the Weyr structure of an eigenvalue and say how it is computed from T.

(f) What is the Weyr form of $\mathbf{J}_3(7)$? And of $7\mathbf{I}_3$?

Solution.

(a) $\mathbf{\Lambda}(\lambda) = \begin{pmatrix} a & -b \\ b & a \end{pmatrix} \in M_2(\mathbb{R})$ and $\mathbf{C}_2(\lambda) = \begin{pmatrix} \mathbf{\Lambda}(\lambda) & \mathbf{I}_2 \\ \mathbf{O} & \mathbf{\Lambda}(\lambda) \end{pmatrix} \in M_4(\mathbb{R})$; the real Jordan block $\mathbf{C}_k(\lambda)$ has size $2k$ (Definition 10.4.1).

(b) Every $\mathbf{A} \in M_n(\mathbb{R})$ is similar, by a **real** invertible matrix, to a direct sum of blocks $\mathbf{J}_k(\mu)$ for the real eigenvalues μ and $\mathbf{C}_k(\lambda)$ for the conjugate pairs $\{\lambda, \overline{\lambda}\}$, and that matrix is unique up to the order of the blocks (Theorem 10.4.4).

(c) False. A triangular real matrix has real eigenvalues (Theorem 9.7.4), and similar matrices have the same characteristic polynomial (Theorem 7.10.3); so a real matrix with a nonreal eigenvalue, such as the rotation by a right angle, is similar to no real triangular matrix.

(d) The real eigenvalue 2 has $a(2) = 2$, so its blocks are $\mathbf{J}_2(2)$ or $\mathbf{J}_1(2) \oplus \mathbf{J}_1(2)$. The pair $\{1 \pm i\}$ has $a(1+i) = 2$, so its real blocks are $\mathbf{C}_2(1+i)$ (size 4) or $\mathbf{C}_1(1+i) \oplus \mathbf{C}_1(1+i)$. That gives four possible real Jordan forms, of total size $2 + 4 = 6$ in each case.

(e) The Weyr structure of λ is the partition $(w_1, w_2, \dots)$ with $w_j = \dim \ker(T - \lambda \operatorname{id}_V)^j - \dim \ker(T - \lambda \operatorname{id}_V)^{j-1}$, that is, the dual partition of the λ-block sizes of the Jordan form (Theorem 10.4.5).

(f) $\mathbf{J}_3(7)$ has one 7-block of size 3, so the Weyr structure is $(1, 1, 1)$ and the Weyr form is $\mathbf{J}_3(7)$ itself. For $7\mathbf{I}_3$ the block sizes are $(1, 1, 1)$, the dual partition is (3), and the Weyr form is $7\mathbf{I}_3$.

10.4.4.2 B. Practice

Exercise. B1: Real forms of two by two matrices

For each real matrix, find the complex eigenvalues, the real Jordan form (Theorem 10.4.4), and an invertible $\mathbf{P} \in M_2(\mathbb{R})$ putting it in that form. Verify $\mathbf{AP} = \mathbf{P\Lambda}$ column by column.

(a) $\mathbf{A}_1 = \begin{pmatrix} 1 & 2 \\ -2 & 1 \end{pmatrix}$.

(b) $\mathbf{A}_2 = \begin{pmatrix} 3 & -5 \\ 1 & -1 \end{pmatrix}$.

Solution.

(a) $p_{\mathbf{A}_1} = x^2 - (\operatorname{tr} \mathbf{A}_1)x + \det \mathbf{A}_1 = x^2 - 2x + 5 = (x-1)^2 + 4$, so the eigenvalues are $1 \pm 2i$ and the real Jordan form is $\mathbf{\Lambda}(1+2i) = \begin{pmatrix} 1 & -2 \\ 2 & 1 \end{pmatrix}$. For $\lambda = 1 + 2i$, $\mathbf{A}_1 - \lambda\mathbf{I} = \begin{pmatrix} -2i & 2 \\ -2 & -2i \end{pmatrix}$, and the first row gives $z_2 = iz_1$; take $\mathbf{w} = (1, i)$. Then $\mathbf{w} = \mathbf{u} - i\mathbf{v}$ with $\mathbf{u} = (1, 0)$ and $\mathbf{v} = (0, -1)$, so

$$\mathbf{P} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \qquad \mathbf{P}^{-1}\mathbf{A}_1\mathbf{P} = \begin{pmatrix} 1 & -2 \\ 2 & 1 \end{pmatrix}.$$

Check: $\mathbf{A}_1\mathbf{u} = (1, -2) = \mathbf{u} + 2\mathbf{v}$ and $\mathbf{A}_1\mathbf{v} = (-2, -1) = -2\mathbf{u} + \mathbf{v}$, which are the two columns of $\mathbf{\Lambda}(1 + 2i)$ as Lemma 10.4.1 (a) predicts with $a = 1, b = 2$.

(b) $p_{\mathbf{A}_2} = x^2 - 2x + (-3 + 5) = x^2 - 2x + 2$, with roots $1 \pm i$; the real Jordan form is $\mathbf{\Lambda}(1 + i) = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}$. For $\lambda = 1 + i$, the second row of $\mathbf{A}_2 - \lambda\mathbf{I} = \begin{pmatrix} 2 - i & -5 \\ 1 & -2 - i \end{pmatrix}$ gives $z_1 = (2 + i)z_2$; take $\mathbf{w} = (2 + i, 1)$. Then $\mathbf{u} = (2, 1)$ and $\mathbf{v} = -\operatorname{Im} \mathbf{w} = (-1, 0)$, so

$$\mathbf{P} = \begin{pmatrix} 2 & -1 \\ 1 & 0 \end{pmatrix}, \qquad \det \mathbf{P} = 1, \qquad \mathbf{P}^{-1}\mathbf{A}_2\mathbf{P} = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}.$$

Check: $\mathbf{A}_2\mathbf{u} = (6 - 5, 2 - 1) = (1, 1) = \mathbf{u} + \mathbf{v}$ and $\mathbf{A}_2\mathbf{v} = (-3, -1) = -\mathbf{u} + \mathbf{v}$.

Exercise. B2: Weyr forms of Jordan matrices

For each Jordan matrix, write down the block sizes, the Weyr structure of each eigenvalue, and the Weyr form (Theorem 10.4.5).

(a) $\mathbf{J}_3(0) \oplus \mathbf{J}_2(0) \oplus \mathbf{J}_1(0) \in M_6(F)$.
(b) $\mathbf{J}_2(4) \oplus \mathbf{J}_1(4) \oplus \mathbf{J}_1(9) \in M_4(F)$.
(c) $\mathbf{J}_2(0) \oplus \mathbf{J}_2(0) \in M_4(F)$.

Solution.

(a) Block sizes $(3,2,1)$, a partition of 6. Dual partition: $w_1 = 3$, $w_2 = 2$, $w_3 = 1$. So the Weyr form is the basic Weyr matrix with $\lambda = 0$ and structure $(3,2,1)$,

$$\mathbf{W} = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

with the 3×2 block $\begin{pmatrix} \mathbf{I}_2 \\ \mathbf{O} \end{pmatrix}$ and the 2×1 block $\begin{pmatrix} \mathbf{I}_1 \\ \mathbf{O} \end{pmatrix}$ above the diagonal. As a check, $\operatorname{rank} \mathbf{W} = 3 = \operatorname{rank} \mathbf{J}$ and $\operatorname{rank} \mathbf{W}^2 = 1 = \operatorname{rank} \mathbf{J}^2$.

(b) For $\lambda = 4$: sizes $(2,1)$, dual $(2,1)$. For $\lambda = 9$: sizes (1), dual (1). The Weyr form is the direct sum of the basic blocks,

$$\begin{pmatrix} 4 & 0 & 1 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{pmatrix} \oplus (9).$$

(c) Sizes $(2,2)$, dual $w_1 = 2$, $w_2 = 2$. The basic Weyr matrix with $\lambda = 0$ and structure $(2,2)$ is

$$\mathbf{W} = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

which is $\mathbf{J}_2(0) \oplus \mathbf{J}_2(0)$ with its basis reordered as $(\mathbf{e}_1, \mathbf{e}_3, \mathbf{e}_2, \mathbf{e}_4)$.

Exercise. B3: Real forms of three by three matrices from the characteristic polynomial

For each characteristic polynomial, list all possible real Jordan forms (Theorem 10.4.4) of a matrix $\mathbf{A} \in M_3(\mathbb{R})$ with that polynomial.

(a) $p_{\mathbf{A}} = (x-2)(x^2+1)$.
(b) $p_{\mathbf{A}} = (x+1)(x^2-2x+5)$.
(c) $p_{\mathbf{A}} = (x-1)^3$.

(d) $p_{\mathbf{A}} = (x^2+1)(x-1)$ but with $\mathbf{A}$ required to be diagonalizable over $\mathbb{R}$. Is this possible?

Solution.

(a) The real eigenvalue 2 has $a(2) = 1$, giving the block $\mathbf{J}_1(2) = (2)$. The pair $\{i, -i\}$ has $a(i) = 1$, giving $\mathbf{C}_1(i) = \mathbf{\Lambda}(i) = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$. So the real Jordan form is $(2) \oplus \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$, the only possibility.

(b) $x^2 - 2x + 5$ has roots $1 \pm 2i$, so the form is $(-1) \oplus \mathbf{\Lambda}(1+2i) = (-1) \oplus \begin{pmatrix} 1 & -2 \\ 2 & 1 \end{pmatrix}$, again the only possibility.

(c) All eigenvalues are real, so the real Jordan form is an ordinary Jordan form: the 1-block sizes form a partition of 3, giving $\mathbf{J}_3(1)$, $\mathbf{J}_2(1) \oplus \mathbf{J}_1(1)$, or $\mathbf{I}_3$.

(d) No. A diagonalizable $\mathbf{A} \in M_3(\mathbb{R})$ is similar over $\mathbb{R}$ to a diagonal real matrix, whose characteristic polynomial splits over $\mathbb{R}$ (Theorem 7.4.6); but x^2+1 has no real root. In fact whenever $\mathbf{A}$ has a nonreal eigenvalue its real Jordan form contains a block $\mathbf{\Lambda}(\lambda)$, which is not diagonal.

10.4.4.3 C. Going deeper

Exercise. C1: The real form of a companion matrix

Let $p = (x^2+1)^2 = x^4 + 2x^2 + 1 \in \mathbb{R}[x]$ and let $\mathbf{C} = \mathbf{C}(p) \in M_4(\mathbb{R})$ be its companion matrix, with 1 in each position just below the diagonal and last column $(-1, 0, -2, 0)$ (Exercise (C1: Companion matrices)).

(a) Write down $\mathbf{C}$, and explain why $p_{\mathbf{C}} = m_{\mathbf{C}} = p$.

(b) Determine the complex Jordan form of $\mathbf{C}$.

(c) Deduce the real Jordan form of $\mathbf{C}$.

(d) Verify your answer by comparing traces and characteristic polynomials.

Hint: for (b), the exponent of $x - i$ in $m_{\mathbf{C}}$ is the largest size of an i-block.

Solution.

(a) With $a_0 = 1$, $a_1 = 0$, $a_2 = 2$, $a_3 = 0$,

$$\mathbf{C} = \begin{pmatrix} 0 & 0 & 0 & -1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & -2 \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

By Exercise (C1: Companion matrices) (b), $p_{\mathbf{C}} = p$, and by Exercise (B3: Companion matrices) (d), $m_{\mathbf{C}} = p$ as well.

(b) Over $\mathbb{C}$, $p = (x-i)^2(x+i)^2$, so $a(i) = a(-i) = 2$. By Corollary 10.3.2 (c), the exponent 2 of $x - i$ in $m_{\mathbf{C}}$ is the largest size of an i-block; the sizes form a partition of 2 with largest part 2, so there is one block $\mathbf{J}_2(i)$. Likewise one block $\mathbf{J}_2(-i)$. The complex Jordan form is $\mathbf{J}_2(i) \oplus \mathbf{J}_2(-i)$.

(c) By Theorem 10.4.4, the real Jordan form has no ordinary Jordan blocks (there is no real eigenvalue) and one real block per complex i-block, of the same size parameter. So it is

$$\mathbf{C}_2(i) = \begin{pmatrix} 0 & -1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

(Consistently, Corollary 10.4.2 turns $\mathbf{C}_2(i)$ back into $\mathbf{J}_2(i) \oplus \mathbf{J}_2(-i)$ over $\mathbb{C}$.)

(d) $\operatorname{tr} \mathbf{C} = 0 = \operatorname{tr} \mathbf{C}_2(i)$, and both matrices have characteristic polynomial $(x^2+1)^2$: for $\mathbf{C}$ by (a), and for $\mathbf{C}_2(i)$ by Corollary 10.4.2. Their minimal polynomials agree too, both being $(x^2+1)^2$.

Exercise. C2: Commuting with a Weyr matrix, and with a Jordan matrix

Let F be a field.

(a) Let $\mathbf{W} \in M_3(F)$ be the basic Weyr matrix with $\lambda = 0$ and structure $(2,1)$ (Definition 10.4.2). Find all $\mathbf{K} \in M_3(F)$ with $\mathbf{WK} = \mathbf{KW}$, and check that they are block upper triangular for the partition $3 = 2+1$.

(b) Let $\mathbf{J} = \mathbf{J}_2(0) \oplus \mathbf{J}_1(0)$. Find all $\mathbf{K} \in M_3(F)$ with $\mathbf{JK} = \mathbf{KJ}$, and exhibit one that is **not** upper triangular. Which partition of 3 does it violate?

(c) Compute the dimension of each of the two commutants, and explain why they are equal.

Solution.

(a) $\mathbf{W} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$. Write $\mathbf{K} = (k_{ij})$. Then $\mathbf{WK}$ has first row equal to the third row of $\mathbf{K}$ and other rows zero, while $\mathbf{KW}$ has third column equal to the first column of $\mathbf{K}$ and other columns zero. Equating,

$$(k_{31}, k_{32}, k_{33}) = (0, 0, k_{11}), \qquad k_{21} = k_{31} = 0.$$

So the commutant is

$$\left\{ \begin{pmatrix} a & b & c \\ 0 & e & f \\ 0 & 0 & a \end{pmatrix} : a, b, c, e, f \in F \right\},$$

of dimension 5. The lower-left 1×2 block is $(k_{31}, k_{32}) = (0,0)$, so every such $\mathbf{K}$ is block upper triangular for $3 = 2+1$, as Proposition 10.4.6 asserts.

(b) $\mathbf{J} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$. Now $\mathbf{JK}$ has first row equal to the second row of $\mathbf{K}$ and other rows zero, while $\mathbf{KJ}$ has second column equal to the first column of $\mathbf{K}$ and other columns

zero. Equating gives $k_{21} = 0$, $k_{22} = k_{11}$, $k_{23} = 0$, and $k_{31} = 0$, so the commutant is

$$\left\{ \begin{pmatrix} a & b & c \\ 0 & a & 0 \\ 0 & h & i \end{pmatrix} \right\},$$

also of dimension 5. Take $\mathbf{K}_0 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$, which commutes with $\mathbf{J}$ (it is the case $a = b = c = i = 0$, $h = 1$). Its only nonzero entry sits in position $(3, 2)$, below the diagonal, so $\mathbf{K}_0$ is not upper triangular. The partition it violates is $3 = 2 + 1$, the Weyr partition of this operator: the lower-left 1×2 block of $\mathbf{K}_0$ is $(0, 1) \neq \mathbf{O}$. So the conclusion of Proposition 10.4.6 genuinely fails for the Jordan form, while (a) confirms it for the Weyr form (Theorem 10.4.5) of the same operator.

(c) Both have dimension 5. They must agree, because $\mathbf{J}$ and $\mathbf{W}$ are similar (both are nilpotent with rank = 1 and $\text{rank}^2 = 0$, so Corollary 10.2.5 applies): if $\mathbf{W} = \mathbf{P}^{-1}\mathbf{J}\mathbf{P}$, then $\mathbf{K} \mapsto \mathbf{P}^{-1}\mathbf{K}\mathbf{P}$ is a bijection from the commutant of $\mathbf{J}$ to that of $\mathbf{W}$, and it is linear and invertible, hence preserves dimension.

10.5 Cyclic Subspaces and Companion Matrices

The Jordan form of Section 3 exists only when the characteristic polynomial splits, because its building block $\mathbf{J}_k(\lambda)$ needs an eigenvalue λ in F. To build a normal form over **every** field we need a block that does not name a scalar. Look again at what a Jordan block really is: reading its columns, $\mathbf{J}_k(\lambda)$ is the matrix of T on the span of $u, Tu, \dots, T^{k-1}u$ for the top u of a chain. That description — *one vector and everything T can make of it* — mentions no eigenvalue at all. This section studies those subspaces, computes their dimension and their matrix, and decides when a single vector generates the whole space.

10.5.1 The subspace generated by one vector

Chapter 9 called a subspace U **invariant** when $Tu \in U$ for every $u \in U$ (Definition 9.1.1). Given one vector v, what is the smallest invariant subspace containing it? It must contain v, hence Tv, hence T^2v, and so on, hence every linear combination of these. That list is already closed under T, so nothing more is forced. So the answer is a span, and it deserves a name.

The cyclic subspace generated by a vector is everything that can be reached from it by applying the operator and taking combinations.

Definition 10.5.1. Cyclic Subspace

Let V be a vector space over F, let $T \in \mathcal{L}(V)$, and let $v \in V$. The **cyclic subspace generated by** v is

$$Z(v; T) \coloneqq \operatorname{span}\left(v,\ Tv,\ T^2v,\ \dots\right) = \operatorname{span}\left(\{\, T^j v : j \geq 0 \,\}\right),$$

where $T^0 = \mathrm{id}_V$. For $\mathbf{A} \in M_n(F)$ we write $Z(v; \mathbf{A}) = Z(v; T_{\mathbf{A}})$.

In words: we apply T to v over and over, collect all the vectors obtained, and take their span — so $Z(v;T)$ is a subspace by Theorem 1.3.1, whatever v is. The spanning list may be infinite, but we shall see in a moment that in finite dimension its first few entries already suffice. Note $Z(0;T) = \operatorname{span}(\emptyset) = \{0\}$ and $v \in Z(v;T)$.

Proposition 10.5.1. The Cyclic Subspace Is the Smallest Invariant Subspace Containing the Vector

Let V be a vector space over F, $T \in \mathcal{L}(V)$, and $v \in V$. Then:

(a) $Z(v;T) = \{\, p(T)v : p \in F[x] \,\}$;
(b) $Z(v;T)$ is T-invariant;
(c) if U is a T-invariant subspace of V with $v \in U$, then $Z(v;T) \subseteq U$.

Proof.

(a) A linear combination of the vectors $T^j v$ is $\sum_j a_j T^j v = p(T)v$ with $p = \sum_j a_j x^j$, and conversely (Definition 2.4.2). By Definition 1.3.2, these combinations are exactly the elements of $Z(v;T)$.

(b) For $p \in F[x]$, $T\big(p(T)v\big) = (xp)(T)v \in Z(v;T)$ by Theorem 6.7.1 (b) and (a). (Or: T maps each spanning vector $T^j v$ to $T^{j+1}v \in Z(v;T)$, so Lemma 9.1.1 applies.)

(c) Let U be invariant with $v \in U$. By induction on j, $T^j v \in U$: this holds for $j = 0$, and if $T^j v \in U$, then $T^{j+1}v = T(T^j v) \in U$ by invariance. So U contains the spanning list of $Z(v;T)$, and hence its span (Theorem 1.3.1, Proposition 1.3.4).

■

Examples. Throughout, V is finite-dimensional over F.

- **An eigenvector generates a line.** If $Tv = \lambda v$ with $v \neq 0$, then $T^j v = \lambda^j v$, so $Z(v;T) = \operatorname{span}(v)$, of dimension 1. The smallest invariant subspace containing an eigenvector is its own line (Proposition 9.1.2).
- **The top of a Jordan chain generates the block.** Let T have matrix $\mathbf{J}_k(\lambda)$ in the basis $\mathcal{B} = (b_1, \dots, b_k)$. Then b_k is the top of the chain (Definition 10.2.2), and $b_{k-1} = (T - \lambda \operatorname{id}_V)b_k = Tb_k - \lambda b_k \in \operatorname{span}(b_k, Tb_k)$; continuing downwards, every b_j lies in $\operatorname{span}(b_k, Tb_k, \dots, T^{k-j}b_k)$. So $Z(b_k;T) = V$: a single Jordan block is one cyclic subspace.
- **Differentiation.** For differentiation $D \in \mathcal{L}(\mathbb{R}[x]_{\leq 3})$, $Z(x^3;D) = \operatorname{span}(x^3, 3x^2, 6x, 6)$ is everything, while $Z(x^2;D) = \operatorname{span}(x^2, 2x, 2) = \mathbb{R}[x]_{\leq 2}$ is a proper subspace.
- **A companion matrix.** Let $\mathbf{A} \in M_n(F)$ have 1 in each position just below the diagonal, so that $\mathbf{A}\mathbf{e}_j = \mathbf{e}_{j+1}$ for $j \leq n-1$ (Exercise (C1: Companion matrices)). Then $\mathbf{A}^j\mathbf{e}_1 = \mathbf{e}_{j+1}$ for $j \leq n-1$, so $Z(\mathbf{e}_1;\mathbf{A})$ contains every standard basis vector and equals F^n.
- **The degenerate case.** $Z(0;T) = \{0\}$, the smallest invariant subspace of all.

Nonexample by minimal change. Truncate the spanning list: $\operatorname{span}(\mathbf{v}, T\mathbf{v})$ need **not** be invariant. For $\mathbf{A} = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$ and $\mathbf{v} = \mathbf{e}_1$ we get $\operatorname{span}(\mathbf{e}_1, \mathbf{A}\mathbf{e}_1) = \operatorname{span}(\mathbf{e}_1, \mathbf{e}_2)$, and $\mathbf{A}\mathbf{e}_2 = \mathbf{e}_3$ is outside it. What fails is closure: cutting the list short loses exactly the vectors that invariance demands. The correct statement is that the list may be cut short at the right place, which is the subject of the next subsection.

Warning. $Z(v;T)$ **depends on** T **as much as on** v**, and the spanning list** $(v, Tv, T^2v, \dots)$ **is almost never independent.** For $\mathbf{v} = \mathbf{e}_1$ and $T = \mathrm{id}$ the list is $(\mathbf{e}_1, \mathbf{e}_1, \mathbf{e}_1, \dots)$ and the cyclic subspace is a line; for the same $\mathbf{v}$ and the companion matrix above it is all of F^3. In finite dimension the list must become dependent by step $\dim V$ at the latest (Theorem 1.6.3), and where it first does is the only thing that matters.

10.5.2 The annihilator of a vector

The examples suggest measuring v by the first dependence in the list $v, Tv, T^2v, \dots$. A dependence is a polynomial relation: $\sum_j a_j T^j v = 0$ says that the polynomial $p = \sum_j a_j x^j$ kills v, not necessarily all of V. Chapter 6 taught us what to do with such a set of polynomials.

The T-annihilator of v is the minimal polynomial of T as seen by the single vector v.

Definition 10.5.2. The T-Annihilator of a Vector

Let V be a **finite-dimensional** vector space over F, let $T \in \mathcal{L}(V)$, and let $v \in V$. The set

$$I_v := \{\, p \in F[x] : p(T)v = 0 \,\}$$

is a nonzero ideal of $F[x]$, and the **T-annihilator** of v, written $m_{T,v}$, is its unique **monic** generator:

$$I_v = \langle m_{T,v} \rangle.$$

Equivalently, $m_{T,v}$ is the monic polynomial of **least degree** with $m_{T,v}(T)v = 0$. For $\mathbf{A} \in M_n(F)$ we write $m_{\mathbf{A},v}$.

Well-definedness. Three checks. *Ideal:* if $p, q \in I_v$, then $(p+q)(T)v = p(T)v + q(T)v = 0$, and for any $h \in F[x]$, $(hp)(T)v = h(T)(p(T)v) = 0$, using Theorem 6.7.1 (a), (b); so I_v is an ideal (Definition 6.2.1). *Nonzero:* $m_T \in I_v$, since $m_T(T) = 0$ as an operator (Definition 9.5.1), and $m_T \neq 0$. *Monic generator:* a nonzero ideal of $F[x]$ has exactly one monic generator, which is its nonzero element of least degree (Theorem 6.2.2). As with the minimal polynomial, the divisibility statement is the one we use:

$$p(T)v = 0 \iff m_{T,v} \mid p. \tag{$*$}$$

In words: $m_{T,v}$ records which polynomials in T kill this one vector. Two degenerate cases: $m_{T,0} = 1$, because every polynomial kills 0; and if $v \neq 0$, then $\deg m_{T,v} \geq 1$, since the constant 1 does not kill v. If v is an eigenvector for λ, then $m_{T,v} = x - \lambda$.

Why this definition. The minimal polynomial m_T is the "annihilator of all of V at once": m_T kills every vector, so $m_{T,v} \mid m_T$ for every v by $(*)$. The individual annihilators are finer, and they differ from vector to vector: for $\mathbf{J} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ we have $m_{\mathbf{J},\mathbf{e}_1} = x - 1$ but $m_{\mathbf{J},\mathbf{e}_2} = (x-1)^2 = m_{\mathbf{J}}$. This spread is exactly what will let us split V into cyclic pieces in Section 6.

Now the payoff: the annihilator computes the dimension of the cyclic subspace, hands us a basis, and hands us the matrix.

Theorem 10.5.2. Basis and Matrix of a Cyclic Subspace

Let V be a finite-dimensional vector space over F, let $T \in \mathcal{L}(V)$, and let $v \in V$ with $v \neq 0$. Write $Z = Z(v;T)$ and

$$m_{T,v} = x^k + a_{k-1}x^{k-1} + \cdots + a_1 x + a_0, \qquad k = \deg m_{T,v} \geq 1.$$

Then:

(a) $(v,\ Tv,\ \ldots,\ T^{k-1}v)$ is a basis of Z; in particular $\dim Z = \deg m_{T,v}$;

(b) $T^k v = -(a_0 v + a_1 Tv + \cdots + a_{k-1}T^{k-1}v)$, and the matrix of $T|_Z$ in the basis of (a) is the companion matrix of $m_{T,v}$ (Exercise (C1: Companion matrices); the definition is repeated in full as Definition 10.5.3 below);

(c) $m_{T|_Z} = p_{T|_Z} = m_{T,v}$;

(d) $m_{T,v} \mid m_T$, and $\deg m_{T,v} \leq \deg m_T \leq \dim V$.

Idea. For spanning, divide: any p is $q\,m_{T,v} + r$ with $\deg r < k$, and the $q\,m_{T,v}$ part dies on v, so $p(T)v = r(T)v$ uses only the first k powers. For independence, a shorter dependence would be a nonzero polynomial of degree $< k$ in I_v, and $m_{T,v}$ has the least degree there. Part (b) is the definition of $m_{T,v}$ rearranged, read as the last column of a matrix; (c) says that on Z the vector v already sees everything, so its annihilator is the minimal polynomial of the restriction, and a degree count makes it the characteristic polynomial too.

Proof.

(a) *Spanning.* Let $z \in Z$, say $z = p(T)v$ with $p \in F[x]$ (Proposition 10.5.1 (a)). Divide by $m_{T,v}$: there are $q, r \in F[x]$ with $p = q\,m_{T,v} + r$ and $\deg r < k$ (Theorem 6.1.1). Then

$$z = p(T)v = q(T)\big(m_{T,v}(T)v\big) + r(T)v = r(T)v,$$

by Theorem 6.7.1 (a), (b) and $m_{T,v}(T)v = 0$. As $\deg r < k$, the vector $r(T)v$ is a combination of $v, Tv, \ldots, T^{k-1}v$.
Independence. Suppose $c_0 v + c_1 Tv + \cdots + c_{k-1}T^{k-1}v = 0$ with the $c_j \in F$ not all zero. Then $q = \sum_{j<k} c_j x^j$ is a nonzero element of I_v of degree less than k, contradicting the fact that $m_{T,v}$ has least degree among the nonzero elements of I_v (Theorem 6.2.2). So all $c_j = 0$.

(b) The identity $m_{T,v}(T)v = 0$ reads $T^k v + a_{k-1}T^{k-1}v + \cdots + a_0 v = 0$, which is the display. Write $z_j = T^{j-1}v$ for $j = 1, \ldots, k$, the basis of (a). For $j \leq k-1$, $Tz_j = T^j v = z_{j+1}$, so the j-th column of the matrix is $\mathbf{e}_{j+1}$; and $Tz_k = T^k v = -a_0 z_1 - \cdots - a_{k-1}z_k$, so the last column is $(-a_0, -a_1, \ldots, -a_{k-1})$. That is exactly the companion matrix of $m_{T,v}$ in the sense of Exercise (C1: Companion matrices), whose definition is recalled in full in the next subsection. For $k = 1$ the matrix is the 1×1 matrix $(-a_0)$.

(c) Let $p \in F[x]$. If $p(T)v = 0$, then for every $q \in F[x]$,

$$p(T)\big(q(T)v\big) = q(T)\big(p(T)v\big) = 0$$

by Theorem 6.7.1 (c), so $p(T)$ kills all of Z by Proposition 10.5.1 (a). Since Z is T-invariant, $(T|_Z)^j z = T^j z$ for every $j \geq 0$ and $z \in Z$, hence $g(T|_Z) = g(T)|_Z$ for every $g \in F[x]$; with $g = p$ this reads $p(T|_Z) = 0$. Conversely, if $p(T|_Z) = 0$, then in

particular $p(T)v = 0$, as $v \in Z$. So the annihilating ideals of v and of $T|_Z$ coincide, and their monic generators agree: $m_{T|_Z} = m_{T,v}$ (Definition 9.5.1, Theorem 6.2.2). Finally $m_{T|_Z} \mid p_{T|_Z}$ (Corollary 9.6.2), both are monic, and $\deg p_{T|_Z} = \dim Z = k = \deg m_{T|_Z}$ by (a) and Theorem 7.10.2; so they are equal (Proposition 6.1.2 (c), (d)).

(d) $m_T \in I_v$ as noted after Definition 10.5.2, so $m_{T,v} \mid m_T$ by $(*)$, and the degrees compare by Proposition 6.1.2 (c). And $\deg m_T \le \deg p_T = \dim V$ by Corollary 9.6.2 and Theorem 7.10.2.

■

Quick check. Let T be the operator on $\mathbb{Q}^2$ with matrix $\operatorname{diag}(2,5)$, and let $\mathbf{v} = (1,1)$. Find $m_{T,\mathbf{v}}$ and $Z(\mathbf{v};T)$. Then do the same for $\mathbf{v} = \mathbf{e}_1$.

Solution. $(T - 2\,\mathrm{id})(T - 5\,\mathrm{id})\mathbf{v} = \mathbf{0}$, as one checks entry by entry, so $m_{T,\mathbf{v}} \mid (x-2)(x-5)$. Neither factor alone kills $\mathbf{v}$: $(T - 2\,\mathrm{id})\mathbf{v} = (0,3) \neq \mathbf{0}$ and $(T - 5\,\mathrm{id})\mathbf{v} = (-3,0) \neq \mathbf{0}$. So $m_{T,\mathbf{v}} = (x-2)(x-5)$, of degree 2, and $Z(\mathbf{v};T) = \mathbb{Q}^2$ by Theorem 10.5.2 (a); the basis it provides is $(\mathbf{v}, T\mathbf{v}) = ((1,1),(2,5))$. For $\mathbf{v} = \mathbf{e}_1$, an eigenvector, $m_{T,\mathbf{e}_1} = x - 2$ and $Z(\mathbf{e}_1;T) = \operatorname{span}(\mathbf{e}_1)$ is a line.

10.5.3 Companion matrices

The matrix in Theorem 10.5.2 (b) appeared in Chapter 7 as an exercise; it is the block we were looking for, and it deserves a definition of its own.

Definition 10.5.3. Companion Matrix

Let $n \ge 1$ and let $p = x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0 \in F[x]$ be **monic** of degree n. The **companion matrix** of p is the matrix $\mathbf{C}(p) \in M_n(F)$ with 1 in each position $(j+1, j)$ just **below** the diagonal, last column $(-a_0, -a_1, \ldots, -a_{n-1})$, and 0 elsewhere:

$$\mathbf{C}(p) = \begin{pmatrix} 0 & 0 & \cdots & 0 & -a_0 \\ 1 & 0 & \cdots & 0 & -a_1 \\ 0 & 1 & \cdots & 0 & -a_2 \\ \vdots & & \ddots & & \vdots \\ 0 & 0 & \cdots & 1 & -a_{n-1} \end{pmatrix}.$$

For $n = 1$ and $p = x + a_0$ this reads $\mathbf{C}(p) = (-a_0)$.

This is the convention of Exercise (C1: Companion matrices), where the $n \ge 2$ case was introduced: the subdiagonal ones shift the standard basis forward, $\mathbf{C}(p)\mathbf{e}_j = \mathbf{e}_{j+1}$ for $j \le n-1$, and the last column records the coefficients of p with their signs changed. Some books use the transpose, or put the coefficients in the last row; nothing changes but the bookkeeping, since a matrix and its transpose have the same characteristic polynomial (Exercise (B2: Transpose)).

Theorem 10.5.3. The Companion Matrix Realizes Its Polynomial Twice

Let $p \in F[x]$ be monic of degree $n \geq 1$ and $\mathbf{C} = \mathbf{C}(p)$. Then

$$p_{\mathbf{C}} = m_{\mathbf{C}} = p,$$

and $\mathbf{e}_1$ satisfies $Z(\mathbf{e}_1; \mathbf{C}) = F^n$ with $m_{\mathbf{C},\mathbf{e}_1} = p$. Consequently every monic polynomial over F is both the characteristic and the minimal polynomial of some square matrix over F.

Proof. For $n = 1$ the matrix is $(-a_0)$, with $p_{\mathbf{C}} = m_{\mathbf{C}} = x + a_0 = p$. Let $n \geq 2$. Since $\mathbf{C}\mathbf{e}_j = \mathbf{e}_{j+1}$ for $j \leq n-1$, induction gives $\mathbf{C}^j\mathbf{e}_1 = \mathbf{e}_{j+1}$ for $0 \leq j \leq n-1$. Hence $Z(\mathbf{e}_1; \mathbf{C})$ contains $\mathbf{e}_1, \dots, \mathbf{e}_n$, so $Z(\mathbf{e}_1; \mathbf{C}) = F^n$, and by Theorem 10.5.2 (a),

$$\deg m_{\mathbf{C},\mathbf{e}_1} = \dim Z(\mathbf{e}_1; \mathbf{C}) = n.$$

Applying Theorem 10.5.2 (c) with $Z = F^n$, so that $T_{\mathbf{C}}|_Z = T_{\mathbf{C}}$,

$$m_{\mathbf{C}} = m_{T_{\mathbf{C}}} = m_{\mathbf{C},\mathbf{e}_1} = p_{T_{\mathbf{C}}} = p_{\mathbf{C}},$$

where $m_{\mathbf{C}} = m_{T_{\mathbf{C}}}$ and $p_{T_{\mathbf{C}}} = p_{\mathbf{C}}$ hold by definition (Definition 9.5.1, Definition 7.10.2), the matrix of $T_{\mathbf{C}}$ in the standard basis being $\mathbf{C}$. Finally $m_{\mathbf{C},\mathbf{e}_1} = p$. Indeed, the basis of $Z(\mathbf{e}_1; \mathbf{C})$ supplied by Theorem 10.5.2 (a) is $(\mathbf{e}_1, \mathbf{C}\mathbf{e}_1, \dots, \mathbf{C}^{n-1}\mathbf{e}_1) = (\mathbf{e}_1, \dots, \mathbf{e}_n)$, the standard basis, so by Theorem 10.5.2 (b) the matrix of $T_{\mathbf{C}}$ in the standard basis is the companion matrix of $m_{\mathbf{C},\mathbf{e}_1}$; and that matrix is $\mathbf{C} = \mathbf{C}(p)$. Two companion matrices of monic polynomials of degree n are equal only if the polynomials are, since the last column of $\mathbf{C}(f)$ lists the coefficients of f with their signs changed (Definition 10.5.3). Hence $m_{\mathbf{C},\mathbf{e}_1} = p$, and $m_{\mathbf{C}} = p_{\mathbf{C}} = p$. The last sentence follows by taking $\mathbf{C}(p)$ for the given p.

■

So the companion matrix is the field-free analogue of a Jordan block: $\mathbf{J}_k(\lambda)$ and $\mathbf{C}((x-\lambda)^k)$ both have characteristic and minimal polynomial $(x-\lambda)^k$, and they are similar, as we will see in Section 6. The difference is that $\mathbf{C}(p)$ is available for **every** monic p, whether or not p has a root in F.

10.5.4 Cyclic operators

One block is one cyclic subspace. Before assembling many of them, we should know when a single one already fills the space.

Definition 10.5.4. Cyclic Vector, Cyclic Operator

Let V be a finite-dimensional vector space over F with $\dim V = n \geq 1$, and let $T \in \mathcal{L}(V)$. A vector $v \in V$ is a **cyclic vector** for T if

$$Z(v; T) = V.$$

We call T **cyclic** if it has a cyclic vector, and $\mathbf{A} \in M_n(F)$ **cyclic** if $T_{\mathbf{A}}$ is.

By Theorem 10.5.2 (a), v is cyclic exactly when $\deg m_{T,v} = n$, and then $(v, Tv, \dots, T^{n-1}v)$ is a basis of V. That is the definition used in Exercise (C1: Cyclic vectors), so the two agree.

Reading Theorem 10.5.2 (b) with $Z = V$: **in the basis generated by a cyclic vector, the matrix of T is a companion matrix.**

The key to the existence of cyclic vectors is a lemma about adding vectors whose annihilators have no common factor. It says that annihilators multiply, exactly as one hopes.

Lemma 10.5.4. Annihilator of a Sum with Coprime Annihilators

Let V be finite-dimensional, $T \in \mathcal{L}(V)$, and $u, w \in V$ with $f = m_{T,u}$ and $g = m_{T,w}$ **coprime**. Then

$$m_{T,u+w} = fg.$$

More generally, if $v_1, \dots, v_k \in V$ have pairwise coprime T-annihilators $f_1, \dots, f_k$, then $m_{T,v_1+\cdots+v_k} = f_1 f_2 \cdots f_k$.

Idea. Write $h = m_{T,u+w}$. That h divides fg is immediate, since fg kills both summands. For the other direction, multiply the relation $h(T)(u+w) = 0$ by $g(T)$: the w part dies, leaving $(hg)(T)u = 0$, so $f \mid hg$; and since f and g are coprime we may cancel g and conclude $f \mid h$. Symmetrically $g \mid h$, and two coprime divisors multiply.

Proof. Polynomials in T commute (Theorem 6.7.1 (c)). Hence

$$(fg)(T)(u+w) = g(T)\big(f(T)u\big) + f(T)\big(g(T)w\big) = 0,$$

so $h \mid fg$, where $h = m_{T,u+w}$. Conversely, applying $g(T)$ to $h(T)(u+w) = 0$ and using $g(T)w = 0$,

$$0 = g(T)h(T)(u+w) = (hg)(T)u + h(T)\big(g(T)w\big) = (hg)(T)u.$$

Therefore $f \mid hg$, and since $\gcd(f, g) = 1$, Theorem 6.2.8 gives $f \mid h$. Swapping the roles of u and w gives $g \mid h$. As f and g are coprime and both divide h, their product divides h (Lemma 6.6.3 (b)). With $h \mid fg$, and all three polynomials monic, $h = fg$ (Proposition 6.1.2 (d)).

For k vectors, induct on k. The case $k = 1$ is trivial. Assume $m_{T,v_1+\cdots+v_{k-1}} = f_1 \cdots f_{k-1}$. Each f_i with $i \le k-1$ is coprime to f_k, so $f_1 \cdots f_{k-1}$ is coprime to f_k (Lemma 6.6.3 (a)), and the two-vector case applies to $u = v_1 + \cdots + v_{k-1}$ and $w = v_k$.

■

Quick check. Let $\mathbf{A} = \operatorname{diag}(2, 2, 5)$ on $\mathbb{Q}^3$. Find $m_{\mathbf{A},\mathbf{e}_1}$, $m_{\mathbf{A},\mathbf{e}_3}$, and $m_{\mathbf{A},\mathbf{e}_1+\mathbf{e}_3}$. Is $\mathbf{e}_1 + \mathbf{e}_3$ a cyclic vector for $\mathbf{A}$?

Solution. The vectors $\mathbf{e}_1$ and $\mathbf{e}_3$ are eigenvectors, for 2 and for 5, so $m_{\mathbf{A},\mathbf{e}_1} = x - 2$ and $m_{\mathbf{A},\mathbf{e}_3} = x - 5$. These are distinct monic linear polynomials, hence coprime, so Lemma 10.5.4 gives

$$m_{\mathbf{A},\mathbf{e}_1+\mathbf{e}_3} = (x-2)(x-5),$$

of degree 2. By Theorem 10.5.2 (a), $\dim Z(\mathbf{e}_1+\mathbf{e}_3;\mathbf{A}) = 2 < 3$, so $\mathbf{e}_1+\mathbf{e}_3$ is **not** cyclic; indeed $Z(\mathbf{e}_1+\mathbf{e}_3;\mathbf{A}) = \operatorname{span}(\mathbf{e}_1,\mathbf{e}_3)$. The lemma makes the annihilator as large as the two pieces allow, and here that is still not all of $\mathbb{Q}^3$ — the eigenvalue 2 is repeated, so $m_{\mathbf{A}} = (x-2)(x-5)$ has degree 2 and no vector can do better.

The lemma turns the primary decomposition into a supply of vectors with the largest possible annihilator.

Proposition 10.5.5. Some Vector Has the Minimal Polynomial as Its Annihilator

Let $V \neq \{0\}$ be a finite-dimensional vector space over F and $T \in \mathcal{L}(V)$. Then there is $v \in V$ with

$$m_{T,v} = m_T.$$

Idea. The minimal polynomial factors as $m_T = p_1^{e_1}\cdots p_k^{e_k}$, and the primary decomposition supplies an invariant piece V_i on which T has minimal polynomial exactly $p_i^{e_i}$. Inside V_i some vector must need the full exponent e_i, or $p_i^{e_i-1}$ would kill the whole piece. Then add up one such vector from each piece: the annihilators are powers of distinct irreducibles, hence pairwise coprime, so Lemma 10.5.4 multiplies them back into m_T.

Proof. Write $m_T = p_1^{e_1}\cdots p_k^{e_k}$ with $p_1,\dots,p_k$ distinct monic irreducibles and $e_i \geq 1$ (Theorem 6.3.5; $\deg m_T \geq 1$ because $V \neq \{0\}$, by Theorem 9.5.1). By Theorem 9.8.1, $V = V_1 \oplus \cdots \oplus V_k$ with $V_i = \ker p_i(T)^{e_i}$ invariant and $m_{T|_{V_i}} = p_i^{e_i}$.

Fix i. *Claim: some $v_i \in V_i$ has $m_{T,v_i} = p_i^{e_i}$.* For $v \in V_i$ we have $p_i^{e_i}(T)v = 0$ by the definition of V_i, so $m_{T,v} \mid p_i^{e_i}$ and hence $m_{T,v} = p_i^{f}$ with $0 \leq f \leq e_i$ (Lemma 9.5.2). Suppose no $v \in V_i$ has $f = e_i$. Then $m_{T,v} \mid p_i^{e_i-1}$ for every $v \in V_i$, so $p_i^{e_i-1}(T)v = 0$ for every $v \in V_i$; as V_i is T-invariant, $(T|_{V_i})^j v = T^j v$ for every $j \geq 0$ and $v \in V_i$, so $g(T|_{V_i}) = g(T)|_{V_i}$ for every $g \in F[x]$, and therefore $p_i^{e_i-1}(T|_{V_i}) = 0$. By Theorem 9.5.1 this forces $p_i^{e_i} = m_{T|_{V_i}}$ to divide $p_i^{e_i-1}$, which is false by degrees (Proposition 6.1.2 (c)). This proves the claim.

Pick such a v_i for each i and put $v = v_1 + \cdots + v_k$. The polynomials $p_1^{e_1},\dots,p_k^{e_k}$ are pairwise coprime: a monic common divisor of $p_i^{e_i}$ and $p_j^{e_j}$ with $i \neq j$ is both a power of p_i and a power of p_j (Lemma 9.5.2), hence is 1 by the uniqueness of factorization (Theorem 6.3.5 (b)). So Lemma 10.5.4 gives

$$m_{T,v} = p_1^{e_1}\cdots p_k^{e_k} = m_T,$$

as claimed.

■

An operator is generated by one vector exactly when its minimal polynomial is as large as it can be.

Theorem 10.5.6. Cyclic Operators

Let V be a finite-dimensional vector space over F with $\dim V = n \geq 1$, and let $T \in \mathcal{L}(V)$. The following are equivalent:

(a) T is cyclic, that is, T has a cyclic vector;
(b) $\deg m_T = n$;
(c) $m_T = p_T$.

In that case v is a cyclic vector if and only if $m_{T,v} = m_T$, and in the basis $(v, Tv, \dots, T^{n-1}v)$ the matrix of T is $\mathbf{C}(m_T)$. The same holds for $\mathbf{A} \in M_n(F)$.

Proof.

(b) $\Leftrightarrow$ (c). Always $m_T \mid p_T$ with both monic and $\deg p_T = n$ (Corollary 9.6.2, Theorem 7.10.2). So $\deg m_T = n$ if and only if $m_T = p_T$ (Proposition 6.1.2 (c), (d)).

(c) $\Rightarrow$ (b). Let v be cyclic. Then $v \neq 0$, since $Z(0;T) = \{0\} \neq V$, and by Theorem 10.5.2 (a), (d),

$$n = \dim V = \dim Z(v;T) = \deg m_{T,v} \leq \deg m_T \leq n.$$

So all these numbers are equal; in particular $\deg m_T = n$ and $m_{T,v} = m_T$, the two being monic with $m_{T,v} \mid m_T$ of equal degree.

(d) $\Rightarrow$ (a). By Proposition 10.5.5 there is v with $m_{T,v} = m_T$, and then $\dim Z(v;T) = \deg m_T = n$ by Theorem 10.5.2 (a). A subspace of the same dimension as V is V (Theorem 1.6.8), so $Z(v;T) = V$ and v is cyclic.

For the final statements, the computation in (a) $\Rightarrow$ (b) shows a cyclic vector has $m_{T,v} = m_T$; conversely if $m_{T,v} = m_T$ and (b) holds, the argument of (b) $\Rightarrow$ (a) shows v is cyclic. The matrix is then $\mathbf{C}(m_{T,v}) = \mathbf{C}(m_T)$ by Theorem 10.5.2 (b). For a matrix $\mathbf{A}$, apply all of this to $T_{\mathbf{A}}$, using $m_{\mathbf{A}} = m_{T_{\mathbf{A}}}$ and $p_{\mathbf{A}} = p_{T_{\mathbf{A}}}$.

■

The implication (a) $\Rightarrow$ (b) was Exercise (C1: Cyclic vectors) (a) in Chapter 9, where the converse was promised for this chapter; the promise is now paid off, and the proof shows what the missing vector is made of — one contribution from each primary piece, chosen to need the full exponent. Two consequences are worth naming. First, an operator with n distinct eigenvalues is cyclic, since then $m_T = p_T$ has n distinct roots (Theorem 9.8.2). Second, **a matrix is similar to a companion matrix if and only if $m_{\mathbf{A}} = p_{\mathbf{A}}$**: one direction is the last sentence of the theorem, and the other follows from Theorem 10.5.3 together with the fact that similar matrices share both polynomials. Section 6 makes this the first case of a general normal form.

Warning. **A cyclic vector need not exist, and when it exists it is far from unique.** For $T = \mathrm{id}$ on F^2 we have $m_T = x - 1$ of degree $1 < 2$, so no vector is cyclic: indeed $Z(\mathbf{v}; \mathrm{id}) = \mathrm{span}(\mathbf{v})$ is a line for every $\mathbf{v} \neq \mathbf{0}$. At the other extreme, if v is cyclic, then so is cv for every $c \neq 0$, and usually many more vectors are: for $\mathbf{J} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ over $\mathbb{Q}$, a vector (s,t) is cyclic exactly when $t \neq 0$, since $((s,t),(s+t,t))$ has determinant $-t^2$. So "the" cyclic vector is never a well-defined object; only the polynomial m_T is.

Example. A Cyclic Subspace and a Cyclic Vector

Let
$$\mathbf{A} = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 2 \end{pmatrix} \in M_3(\mathbb{Q}).$$

(a) Find $p_{\mathbf{A}}$ and $m_{\mathbf{A}}$.

(b) Find $m_{\mathbf{A},\mathbf{e}_1}$, the subspace $Z(\mathbf{e}_1;\mathbf{A})$, and the matrix of $T_{\mathbf{A}}$ restricted to it.

(c) Find a cyclic vector for $\mathbf{A}$ and an invertible $\mathbf{P}$ with $\mathbf{P}^{-1}\mathbf{A}\mathbf{P}$ a companion matrix.

Solution.

(a) Expanding $\det(x\mathbf{I} - \mathbf{A})$ along the second row, whose only nonzero entry is $x - 1$ in position $(2,2)$ (Theorem 7.6.1),
$$p_{\mathbf{A}} = (x-1)\det\begin{pmatrix} x-1 & 0 \\ -1 & x-2 \end{pmatrix} = (x-1)^2(x-2).$$
For $m_{\mathbf{A}}$, the roots 1 and 2 must both occur (Theorem 9.5.3), so the first candidate is $(x-1)(x-2)$. With
$$\mathbf{A} - \mathbf{I} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 1 \end{pmatrix}, \qquad \mathbf{A} - 2\mathbf{I} = \begin{pmatrix} -1 & 1 & 0 \\ 0 & -1 & 0 \\ 1 & 0 & 0 \end{pmatrix},$$
the product $(\mathbf{A} - \mathbf{I})(\mathbf{A} - 2\mathbf{I})$ has rows: row 1 of $\mathbf{A} - \mathbf{I}$ is $(0,1,0)$, which picks out row 2 of $\mathbf{A} - 2\mathbf{I}$, namely $(0,-1,0)$; row 2 is zero; row 3 is $(1,0,1)$, which gives row 1 plus row 3 of $\mathbf{A} - 2\mathbf{I}$, namely $(0,1,0)$. So
$$(\mathbf{A} - \mathbf{I})(\mathbf{A} - 2\mathbf{I}) = \begin{pmatrix} 0 & -1 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \neq \mathbf{O},$$
and by Lemma 9.5.2 the only remaining candidate is $m_{\mathbf{A}} = (x-1)^2(x-2) = p_{\mathbf{A}}$. So $\mathbf{A}$ is cyclic by Theorem 10.5.6.

(b) $\mathbf{A}\mathbf{e}_1 = (1,0,1)$, the first column, and
$$\mathbf{A}^2\mathbf{e}_1 = \mathbf{A}(1,0,1) = (\text{column } 1) + (\text{column } 3) = (1,0,1) + (0,0,2) = (1,0,3).$$
Now $(1,0,3) = -2\mathbf{e}_1 + 3(1,0,1)$, so $(\mathbf{A}^2 - 3\mathbf{A} + 2\mathbf{I})\mathbf{e}_1 = \mathbf{0}$ and $m_{\mathbf{A},\mathbf{e}_1} \mid x^2 - 3x + 2 = (x-1)(x-2)$. Neither factor kills $\mathbf{e}_1$: $(\mathbf{A} - \mathbf{I})\mathbf{e}_1 = (0,0,1) \neq \mathbf{0}$ and $(\mathbf{A} - 2\mathbf{I})\mathbf{e}_1 = (-1,0,1) \neq \mathbf{0}$. Hence
$$m_{\mathbf{A},\mathbf{e}_1} = (x-1)(x-2) = x^2 - 3x + 2,$$
and by Theorem 10.5.2, $Z(\mathbf{e}_1;\mathbf{A}) = \operatorname{span}(\mathbf{e}_1, (1,0,1)) = \operatorname{span}(\mathbf{e}_1, \mathbf{e}_3)$, of dimension 2, with basis $\mathcal{B}_1 = (\mathbf{e}_1, \mathbf{A}\mathbf{e}_1)$ and
$$[T_{\mathbf{A}}|_{Z(\mathbf{e}_1;\mathbf{A})}]_{\mathcal{B}_1}^{\mathcal{B}_1} = \mathbf{C}(x^2 - 3x + 2) = \begin{pmatrix} 0 & -2 \\ 1 & 3 \end{pmatrix}.$$
In particular $\mathbf{e}_1$ is **not** cyclic for $\mathbf{A}$: its cyclic subspace is a plane, not $\mathbb{Q}^3$.

(c) Try $\mathbf{v} = \mathbf{e}_2$. Then

$$\mathbf{A}\mathbf{e}_2 = (1,1,0), \qquad \mathbf{A}^2\mathbf{e}_2 = \mathbf{A}(1,1,0) = (1,0,1) + (1,1,0) = (2,1,1).$$

The matrix with these three columns is

$$\mathbf{P} = \begin{pmatrix} 0 & 1 & 2 \\ 1 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}, \qquad \det \mathbf{P} = -\det \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix} = -1 \neq 0,$$

expanding along the first column, whose only nonzero entry is the 1 in position $(2,1)$. So $(\mathbf{e}_2, \mathbf{A}\mathbf{e}_2, \mathbf{A}^2\mathbf{e}_2)$ is a basis and $\mathbf{e}_2$ is a cyclic vector. By Theorem 10.5.6, $m_{\mathbf{A},\mathbf{e}_2} = m_{\mathbf{A}} = (x-1)^2(x-2) = x^3 - 4x^2 + 5x - 2$ and

$$\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \mathbf{C}(m_{\mathbf{A}}) = \begin{pmatrix} 0 & 0 & 2 \\ 1 & 0 & -5 \\ 0 & 1 & 4 \end{pmatrix}.$$

The check. The identity $\mathbf{A}\mathbf{P} = \mathbf{P}\,\mathbf{C}(m_{\mathbf{A}})$ reads, column by column, $\mathbf{A}\mathbf{e}_2 = \mathbf{A}\mathbf{e}_2$, $\mathbf{A}(\mathbf{A}\mathbf{e}_2) = \mathbf{A}^2\mathbf{e}_2$ — both true by construction — and

$$\mathbf{A}(\mathbf{A}^2\mathbf{e}_2) = 2\mathbf{e}_2 - 5\mathbf{A}\mathbf{e}_2 + 4\mathbf{A}^2\mathbf{e}_2,$$

which is the annihilator relation. Verifying it: $\mathbf{A}(2,1,1) = 2(1,0,1) + (1,1,0) + (0,0,2) = (3,1,4)$, while

$$2(0,1,0) - 5(1,1,0) + 4(2,1,1) = (0-5+8,\ 2-5+4,\ 4) = (3,1,4). \checkmark$$

10.5.5 Exercises

10.5.5.1 A. Check your understanding

Exercise. A1

(a) Define $Z(v;T)$ and the T-annihilator $m_{T,v}$.

(b) State the formula for $\dim Z(v;T)$ and say which basis of $Z(v;T)$ the theorem provides.

(c) Write down $\mathbf{C}(p)$ for $p = x^3 - 2x + 5$, and say what $p_{\mathbf{C}(p)}$ and $m_{\mathbf{C}(p)}$ are.

(d) True or false: every operator on a space of dimension $n \geq 2$ has a cyclic vector. Justify your answer.

(e) What is $m_{T,v}$ when v is an eigenvector of T for λ? What is $Z(v;T)$ then?

(f) True or false: if $m_{T,v} = m_T$ for some v, then v is a cyclic vector. Justify your answer.

Solution.

(a) $Z(v;T) = \operatorname{span}(v, Tv, T^2v, \dots) = \{p(T)v : p \in F[x]\}$ (Definition 10.5.1, Proposition 10.5.1); $m_{T,v}$ is the monic generator of the ideal $\{p : p(T)v = 0\}$, equivalently the monic polynomial of least degree killing v (Definition 10.5.2).

(b) $\dim Z(v;T) = \deg m_{T,v} = k$, with basis $(v, Tv, \dots, T^{k-1}v)$ (Theorem 10.5.2 (a)).

(c) Here $a_0 = 5$, $a_1 = -2$, $a_2 = 0$, so

$$\mathbf{C}(p) = \begin{pmatrix} 0 & 0 & -5 \\ 1 & 0 & 2 \\ 0 & 1 & 0 \end{pmatrix},$$

and $p_{\mathbf{C}(p)} = m_{\mathbf{C}(p)} = x^3 - 2x + 5$ (Theorem 10.5.3).

(d) False. id_V has $m_T = x - 1$ of degree $1 < n$, so no cyclic vector (Theorem 10.5.6). More generally any T with $\deg m_T < n$ fails.

(e) $m_{T,v} = x - \lambda$, and $Z(v;T) = \mathrm{span}(v)$ is the eigenline.

(f) False. By Theorem 10.5.2 (a), $\dim Z(v;T) = \deg m_{T,v} = \deg m_T$, which need not be n. For $T = \mathrm{id}$ on F^2 and $\mathbf{v} = \mathbf{e}_1$ we have $m_{T,\mathbf{v}} = x - 1 = m_T$, yet $Z(\mathbf{v};T) = \mathrm{span}(\mathbf{e}_1)$ is a line. The statement is true when $\deg m_T = \dim V$ (Theorem 10.5.6).

10.5.5.2 B. Practice

Exercise. B1: Annihilators of vectors

Let

$$\mathbf{A} = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 2 \end{pmatrix} \in M_3(\mathbb{Q}).$$

(a) Write down $p_{\mathbf{A}}$.

(b) Find $m_{\mathbf{A},\mathbf{e}_1}$, $m_{\mathbf{A},\mathbf{e}_2}$, and $m_{\mathbf{A},\mathbf{e}_3}$.

(c) For each of the three vectors, give $\dim Z(\mathbf{e}_j;\mathbf{A})$ and identify the subspace.

(d) Write down the matrix of $T_{\mathbf{A}}$ restricted to $Z(\mathbf{e}_2;\mathbf{A})$ in the basis $(\mathbf{e}_2, \mathbf{A}\mathbf{e}_2)$, and say which companion matrix it is.

Solution.

(a) $\mathbf{A}$ is upper triangular, so $p_{\mathbf{A}} = (x-1)^2(x-2)$ (Theorem 7.4.6).

(b) $\mathbf{A}\mathbf{e}_1 = \mathbf{e}_1$, so $\mathbf{e}_1$ is an eigenvector and $m_{\mathbf{A},\mathbf{e}_1} = x - 1$.
$\mathbf{A}\mathbf{e}_2 = (1,1,0) = \mathbf{e}_1 + \mathbf{e}_2$, so $(\mathbf{A} - \mathbf{I})\mathbf{e}_2 = \mathbf{e}_1 \neq \mathbf{0}$ and $(\mathbf{A} - \mathbf{I})^2\mathbf{e}_2 = (\mathbf{A} - \mathbf{I})\mathbf{e}_1 = \mathbf{0}$. Hence $(x-1)^2$ kills $\mathbf{e}_2$ but $x - 1$ does not, and $m_{\mathbf{A},\mathbf{e}_2} = (x-1)^2$ by Lemma 9.5.2.
$\mathbf{A}\mathbf{e}_3 = (1,1,2)$ and $\mathbf{A}^2\mathbf{e}_3 = \mathbf{A}(1,1,2) = (1,0,0) + (1,1,0) + 2(1,1,2) = (4,3,4)$. The matrix with columns $\mathbf{e}_3, (1,1,2), (4,3,4)$ is $\begin{pmatrix} 0 & 1 & 4 \\ 0 & 1 & 3 \\ 1 & 2 & 4 \end{pmatrix}$, of determinant $\det \begin{pmatrix} 1 & 4 \\ 1 & 3 \end{pmatrix} = -1 \neq 0$ by expansion along the first column. So the three vectors are independent, $\deg m_{\mathbf{A},\mathbf{e}_3} = 3$, and since $m_{\mathbf{A},\mathbf{e}_3} \mid m_{\mathbf{A}} \mid p_{\mathbf{A}}$ with $\deg p_{\mathbf{A}} = 3$ we get $m_{\mathbf{A},\mathbf{e}_3} = p_{\mathbf{A}} = (x-1)^2(x-2)$.

(c) $\dim Z(\mathbf{e}_1;\mathbf{A}) = 1$ and $Z(\mathbf{e}_1;\mathbf{A}) = \mathrm{span}(\mathbf{e}_1)$. $\dim Z(\mathbf{e}_2;\mathbf{A}) = 2$ and $Z(\mathbf{e}_2;\mathbf{A}) = \mathrm{span}(\mathbf{e}_2, \mathbf{e}_1 + \mathbf{e}_2) = \mathrm{span}(\mathbf{e}_1, \mathbf{e}_2)$. $\dim Z(\mathbf{e}_3;\mathbf{A}) = 3$, so $Z(\mathbf{e}_3;\mathbf{A}) = \mathbb{Q}^3$ and $\mathbf{e}_3$ is a cyclic vector.

(d) $(x-1)^2 = x^2 - 2x + 1$, so by Theorem 10.5.2 (b) the matrix is

$$\mathbf{C}((x-1)^2) = \begin{pmatrix} 0 & -1 \\ 1 & 2 \end{pmatrix}.$$

Exercise. B2: Cyclic or not?

Determine whether each matrix is cyclic. For those that are, find a cyclic vector; for those that are not, say which condition fails.

(a) $\operatorname{diag}(1,2,3) \in M_3(\mathbb{Q})$.

(b) $\operatorname{diag}(1,1,2) \in M_3(\mathbb{Q})$.

(c) $\mathbf{B} = \begin{pmatrix} 2 & 1 & 0 \\ 0 & 2 & 0 \\ 1 & 1 & 1 \end{pmatrix} \in M_3(\mathbb{Q})$, for which $p_{\mathbf{B}} = m_{\mathbf{B}} = (x-2)^2(x-1)$. Check first that none of $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$ is cyclic.

(d) $\mathbf{J}_2(0) \oplus \mathbf{J}_1(0) \in M_3(\mathbb{Q})$.

Hint: for (c), compute the three annihilators $m_{\mathbf{B},\mathbf{e}_j}$ and use Lemma 10.5.4.

Solution.

(a) Cyclic. The three eigenvalues are distinct, so $m = (x-1)(x-2)(x-3) = p$ (Theorem 9.8.2), of degree 3. A cyclic vector is $\mathbf{v} = (1,1,1)$: the matrix with columns $\mathbf{v}, \mathbf{A}\mathbf{v} = (1,2,3), \mathbf{A}^2\mathbf{v} = (1,4,9)$ is a Vandermonde matrix with distinct nodes $1,2,3$, hence invertible (Corollary 7.8.2).

(b) Not cyclic. Here $m = (x-1)(x-2)$ has degree $2 < 3$, so condition (b) of Theorem 10.5.6 fails.

(c) *The standard basis vectors.* $\mathbf{B}\mathbf{e}_1 = (2,0,1)$ and $\mathbf{B}^2\mathbf{e}_1 = \mathbf{B}(2,0,1) = 2(2,0,1) + (0,0,1) = (4,0,3)$. Since $(4,0,3) = -2\mathbf{e}_1 + 3(2,0,1)$, the vector $\mathbf{e}_1$ is killed by $x^2 - 3x + 2 = (x-1)(x-2)$, and by neither factor alone ($(\mathbf{B}-\mathbf{I})\mathbf{e}_1 = (1,0,1) \neq \mathbf{0}$, $(\mathbf{B}-2\mathbf{I})\mathbf{e}_1 = (0,0,1) \neq \mathbf{0}$); so $m_{\mathbf{B},\mathbf{e}_1} = (x-1)(x-2)$, of degree 2.
$\mathbf{B}\mathbf{e}_2 = (1,2,1)$ and $\mathbf{B}^2\mathbf{e}_2 = \mathbf{B}(1,2,1) = (2,0,1) + 2(1,2,1) + (0,0,1) = (4,4,4)$. As $(4,4,4) = -4\mathbf{e}_2 + 4(1,2,1)$, the vector $\mathbf{e}_2$ is killed by $x^2 - 4x + 4 = (x-2)^2$, and not by $x-2$; so $m_{\mathbf{B},\mathbf{e}_2} = (x-2)^2$.
$\mathbf{B}\mathbf{e}_3 = \mathbf{e}_3$, so $m_{\mathbf{B},\mathbf{e}_3} = x-1$.
All three degrees are less than 3, so none of $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$ is cyclic (Theorem 10.5.2 (a)).
A cyclic vector. The annihilators $(x-2)^2$ of $\mathbf{e}_2$ and $x-1$ of $-\mathbf{e}_3$ — note $m_{T,-\mathbf{v}} = m_{T,\mathbf{v}}$, since $p(T)(-\mathbf{v}) = -p(T)\mathbf{v}$ — are coprime, so Lemma 10.5.4 gives

$$m_{\mathbf{B},\mathbf{e}_2-\mathbf{e}_3} = (x-2)^2(x-1) = m_{\mathbf{B}}.$$

By Theorem 10.5.6, $\mathbf{v} = \mathbf{e}_2 - \mathbf{e}_3 = (0,1,-1)$ is a cyclic vector. As a check, $\mathbf{B}\mathbf{v} = (1,2,1) - (0,0,1) = (1,2,0)$ and $\mathbf{B}^2\mathbf{v} = \mathbf{B}(1,2,0) = (2,0,1) + 2(1,2,1) = (4,4,3)$, and the matrix with columns $(0,1,-1), (1,2,0), (4,4,3)$ has determinant

$$-\det\begin{pmatrix} 1 & 4 \\ -1 & 3 \end{pmatrix} + 4\det\begin{pmatrix} 1 & 2 \\ -1 & 0 \end{pmatrix} = -(3+4) + 4 \cdot 2 = 1 \neq 0$$

by expansion along the first row.

(d) Not cyclic: $m = x^2$ (Proposition 9.5.6) has degree $2 < 3$.

Exercise. B3: Cyclic subspaces for differentiation

Let $V = \mathbb{R}[x]_{\le 3}$ and let $D \in \mathcal{L}(V)$ be differentiation.

(a) Find $m_{D,q}$ for $q = 1$, $q = x$, $q = x^2$, and $q = x^3$.

(b) Find $Z(x^2; D)$ and the matrix of D restricted to it in the basis (x^2, Dx^2, D^2x^2).

(c) Is x^3 a cyclic vector for D? What is m_D?

Solution.

(a) $D1 = 0$, so $m_{D,1} = x$. $Dx = 1 \neq 0$ and $D^2x = 0$, so $m_{D,x} = x^2$. Likewise $D^2x^2 = 2 \neq 0$ and $D^3x^2 = 0$, so $m_{D,x^2} = x^3$; and $D^3x^3 = 6 \neq 0$ with $D^4x^3 = 0$, so $m_{D,x^3} = x^4$. (In each case the annihilator is a monic divisor of x^4, hence a power of x, by Lemma 9.5.2.)

(b) $Z(x^2; D) = \operatorname{span}(x^2, 2x, 2) = \mathbb{R}[x]_{\le 2}$, of dimension $3 = \deg m_{D,x^2}$. With $z_1 = x^2$, $z_2 = Dx^2 = 2x$, $z_3 = D^2x^2 = 2$ we have $Dz_1 = z_2$, $Dz_2 = z_3$, $Dz_3 = 0$, so the matrix is

$$\mathbf{C}(x^3) = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} = \mathbf{J}_3(0)^{\top}.$$

(c) Yes: $\dim Z(x^3; D) = \deg m_{D,x^3} = 4 = \dim V$. Hence $m_D = m_{D,x^3} = x^4 = p_D$ by Theorem 10.5.6, which agrees with Example (Differentiation Is Nilpotent).

10.5.5.3 C. Going deeper

Exercise. C1: What commutes with a cyclic operator

Let V be finite-dimensional with $\dim V = n \ge 1$ and let $T \in \mathcal{L}(V)$ satisfy $m_T = p_T$.

(a) Prove that every $S \in \mathcal{L}(V)$ with $ST = TS$ is a polynomial in T, that is, $S = q(T)$ for some $q \in F[x]$.

(b) Deduce that the set $\{\, S \in \mathcal{L}(V) : ST = TS \,\}$ is a subspace of $\mathcal{L}(V)$ of dimension exactly n.

(c) Show that (a) fails without the hypothesis $m_T = p_T$, by taking $T = \mathrm{id}_V$ with $n \ge 2$.

Hint for (a): apply S to a cyclic vector.

Solution.

(a) By Theorem 10.5.6 there is a cyclic vector v, and $\mathcal{B} = (v, Tv, \dots, T^{n-1}v)$ is a basis of V. Since $\mathcal{B}$ spans V, there is $q \in F[x]$ with $Sv = q(T)v$ (Proposition 10.5.1 (a)). For each $j \ge 0$, using $ST = TS$ and hence $ST^j = T^jS$ (induction on j), and that

polynomials in T commute with T (Theorem 6.7.1 (c)),

$$S(T^j v) = T^j(Sv) = T^j q(T)v = q(T)(T^j v).$$

So S and $q(T)$ agree on the basis $\mathcal{B}$, hence $S = q(T)$ (Theorem 2.1.5).

(b) The set is a subspace: it contains 0, and if $S_1T = TS_1$ and $S_2T = TS_2$, then $(S_1 + cS_2)T = T(S_1 + cS_2)$. By (a) it equals $\{q(T) : q \in F[x]\} = \operatorname{span}(\mathrm{id}_V, T, \dots, T^{n-1})$, where we used Theorem 6.1.1 to reduce q modulo m_T, whose degree is n. That spanning list is independent: a vanishing combination would give a nonzero polynomial of degree less than $n = \deg m_T$ annihilating T, contradicting Theorem 9.5.1. So the dimension is n.

(c) For $T = \mathrm{id}_V$ every S commutes with T, so the commutant is all of $\mathcal{L}(V)$, of dimension $n^2 \geq 4$. But the polynomials in id_V are just the scalar operators $c\,\mathrm{id}_V$, a subspace of dimension 1. Here $m_T = x - 1 \neq (x-1)^n = p_T$.

Exercise. C2: How many cyclic vectors?

Let V be finite-dimensional with $\dim V = n \geq 1$ and $T \in \mathcal{L}(V)$.

(a) Prove that v is a cyclic vector for T if and only if v lies in no T-invariant subspace U with $U \neq V$.

(b) Prove that if v is cyclic, then so is cv for every $c \in F$ with $c \neq 0$. Deduce that over an infinite field a cyclic operator has infinitely many cyclic vectors.

(c) Let $F = \mathbb{F}_2$ and $T = T_{\mathbf{J}}$ with $\mathbf{J} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$. Find **all** cyclic vectors of T.

(d) Deduce from (c) that the set of cyclic vectors, together with 0, need not be a subspace.

Solution.

(a) ($\Rightarrow$) Let v be cyclic and let U be T-invariant with $v \in U$. By Proposition 10.5.1 (c), $V = Z(v;T) \subseteq U$, so $U = V$. ($\Leftarrow$) If v is not cyclic, then $U = Z(v;T)$ is itself a T-invariant subspace containing v with $U \neq V$ (Proposition 10.5.1 (b)).

(b) For $c \neq 0$, $p(T)(cv) = c\,p(T)v$, so $p(T)(cv) = 0$ if and only if $p(T)v = 0$ (as $c \neq 0$ and by Theorem 1.1.9). Hence $m_{T,cv} = m_{T,v}$, and Theorem 10.5.6 makes cv cyclic as well. Over an infinite field the vectors cv, $c \neq 0$, are pairwise distinct (if $cv = c'v$ with $v \neq 0$, then $c = c'$), so there are infinitely many.

(c) Write $\mathbf{v} = (s,t)$ with $s, t \in \mathbb{F}_2$. Then $\mathbf{Jv} = (s+t, t)$, and $(\mathbf{v}, \mathbf{Jv})$ is a basis of $\mathbb{F}_2^2$ if and only if

$$\det \begin{pmatrix} s & s+t \\ t & t \end{pmatrix} = st - t(s+t) = -t^2 = t^2 \neq 0,$$

that is, if and only if $t = 1$. So the cyclic vectors are exactly $(0,1)$ and $(1,1)$, two of the four vectors of $\mathbb{F}_2^2$.

(d) The set $\{\mathbf{0}, (0,1), (1,1)\}$ has 3 elements, and a subspace of $\mathbb{F}_2^2$ has 1, 2, or 4 elements. Concretely, $(0,1) + (1,1) = (1,0)$, which is an eigenvector of $\mathbf{J}$ and hence not cyclic. So the set is not closed under addition.

10.6 The Rational Canonical Form

Everything so far has needed the characteristic polynomial to split. Section 5 removed that need from the building block: a companion matrix $\mathbf{C}(p)$ exists for every monic p, and it is the matrix of an operator on a cyclic subspace. What is left is to cut an arbitrary space into cyclic pieces. This section does that, over an **arbitrary field**, and the result is a genuine canonical form: a list of polynomials attached to T, from which the matrix is written down mechanically, and which two operators share exactly when they are similar. Along the way we learn that similarity does not change if the field is enlarged.

10.6.1 The plan

The goal, stated before its parts are named:

Every operator on a nonzero finite-dimensional space decomposes the space into cyclic subspaces $Z(v_1;T),\dots,Z(v_r;T)$ whose annihilators form a divisibility chain $d_1 \mid d_2 \mid \cdots \mid d_r = m_T$, and that chain is determined by T.

In the adapted basis the matrix is then $\mathbf{C}(d_1)\oplus\cdots\oplus\mathbf{C}(d_r)$, by Theorem 10.5.2 (b) and Theorem 9.1.5. So the whole content is the decomposition, and it is the hardest theorem of the chapter. The route has three steps, and each one is a tool we already have.

- **Step 1 (one prime at a time).** Chapter 9's primary decomposition splits $V = V_1\oplus\cdots\oplus V_k$, where $m_{T|_{V_i}} = p_i^{e_i}$ is a power of a single monic irreducible (Theorem 9.8.1). Nothing more is needed from the general theory.
- **Step 2 (inside one prime power).** On a piece with $m_T = p^e$, the operator $P = p(T)$ is nilpotent, and the structure theorem for nilpotent operators was proved by induction, with $\operatorname{im} N$ as the smaller space to which the hypothesis applies (Theorem 10.2.3). We run the same induction with P in place of N and cyclic subspaces $Z(u;T)$ in place of the spans of chains. The outcome is a decomposition into cyclic pieces whose annihilators are powers $p^{e_1}\ge p^{e_2}\ge\dots$ of the one prime.
- **Step 3 (regroup across primes).** A cyclic piece for p_1 and a cyclic piece for p_2 can be **merged**: the sum of their generators has annihilator the product, by Lemma 10.5.4. Merging the largest piece of every prime gives $d_r = m_T$, the next largest give d_{r-1}, and so on; the divisibility chain is automatic because within one prime the exponents decrease.

Uniqueness is a count, exactly as for the nilpotent case: the dimensions $\dim\ker p(T)^k$ are visible in the decomposition and see nothing but the exponents.

10.6.2 One prime at a time

Here is Step 2. Compare it with Theorem 10.2.3, which is the special case $p = x$: there the pieces are spanned by Jordan chains $u, Nu,\dots$, and here by $u, Tu, T^2u,\dots$.

Theorem 10.6.1. Cyclic Decomposition for a Single Prime Power

Let $V \neq \{0\}$ be a finite-dimensional vector space over F, let $T\in\mathcal{L}(V)$, and suppose

$$m_T = p^e, \qquad p\in F[x] \text{ monic irreducible of degree } d, \quad e\ge 1.$$

Then there are an integer $m\ge 1$, vectors $u_1,\dots,u_m\in V$, and exponents

$$e = k_1 \ge k_2 \ge \cdots \ge k_m \ge 1$$

such that

$$V = Z(u_1;T) \oplus Z(u_2;T) \oplus \cdots \oplus Z(u_m;T), \qquad m_{T,u_i} = p^{k_i} \quad (1 \le i \le m).$$

In particular $\dim V = d(k_1 + \cdots + k_m)$.

Idea. Induction on $\dim V$, with the smaller object found inside: $W = \operatorname{im} p(T)$. It is T-invariant, the minimal polynomial of $T|_W$ is p^{e-1}, and $\dim W < \dim V$ because $p(T)$ is not injective. So W already breaks into cyclic pieces $Z(w_i;T)$. Each is one step short of what we want: $w_i = p(T)u_i$ for some $u_i \in V$, and $Z(u_i;T)$ is a longer piece, with annihilator p times that of w_i. What is still missing is counted by $U = \ker p(T)$, on which T has **irreducible** minimal polynomial p, so that U is a direct sum of cyclic pieces of dimension d each and no proper cyclic piece inside one of them exists — that is the case $e = 1$, settled first. The ends $p^{k_i-1}(T)u_i$ of the lengthened pieces sit inside U; complete them to a decomposition of U, and the extra generators supply the pieces with annihilator p. Then Rank–Nullity makes the dimensions add up exactly, and directness is proved by applying $p(T)$ once, as in the nilpotent case.

Proof. We argue by induction on $n = \dim V \ge 1$, the statement being: for every field F, every monic irreducible $p \in F[x]$, every $e \ge 1$, and every T on a space of dimension n with $m_T = p^e$, a decomposition as described exists.

Write $P = p(T)$ and $U = \ker P$, and recall that $Z(v;T|_{V'}) = Z(v;T)$ whenever V' is a T-invariant subspace containing v, since the two operators agree on v and on everything T produces from it.

The case $e = 1$. Then $m_T = p$ is irreducible. For $u \neq 0$ we have $m_{T,u} \mid m_T = p$ (Theorem 10.5.2 (d)) and $m_{T,u} \neq 1$, so $m_{T,u} = p$ and $\dim Z(u;T) = d$ by Theorem 10.5.2 (a). Two consequences.

Every cyclic piece is minimal. If $W \subseteq Z(u;T)$ is T-invariant and $W \neq \{0\}$, pick $w \in W$ with $w \neq 0$. Then $Z(w;T) \subseteq W$ (Proposition 10.5.1 (c)) and $\dim Z(w;T) = d = \dim Z(u;T)$, so $Z(w;T) = Z(u;T)$ by Theorem 1.6.8, and $W = Z(u;T)$. So the only T-invariant subspaces of $Z(u;T)$ are $\{0\}$ and $Z(u;T)$.

The pieces fill V. Among all lists $u_1, \dots, u_m$ of nonzero vectors for which $Z(u_1;T) + \cdots + Z(u_m;T)$ is direct, choose one with m as large as possible; this is possible because such a sum has dimension $md \le \dim V$ (Theorem 1.7.5). Write S for that sum. If $S \neq V$, pick $v \in V \setminus S$. The subspace $Z(v;T) \cap S$ is T-invariant (Proposition 9.1.4), lies inside $Z(v;T)$ and does not contain v, so it is $\{0\}$ by the previous paragraph. Then $S + Z(v;T)$ is direct — a relation $c_1 + \cdots + c_m + c = 0$ with $c_j \in Z(u_j;T)$ and $c \in Z(v;T)$ puts $c \in Z(v;T) \cap S = \{0\}$, and directness of S kills the rest (Theorem 1.7.5) — contradicting the maximality of m. Hence $V = Z(u_1;T) \oplus \cdots \oplus Z(u_m;T)$ with every $m_{T,u_i} = p$, and every $k_i = 1 = e$.

This settles $e = 1$, and in particular the base case $n = 1$, where m_T is linear.

The case $e \ge 2$. Let $n \ge 2$ and assume the statement for all dimensions less than n.

Step 1. Cyclic pieces inside $\operatorname{im} P$. Put $W = \operatorname{im} P$. Since $e \ge 2$, the polynomial p does not annihilate T (else $m_T = p^e$ would divide p, Theorem 9.5.1), so $P \neq 0$ and $W \neq \{0\}$. Also P is not injective: otherwise $P^e = m_T(T) = 0$ would be injective on $V \neq \{0\}$ (Theorem 0.4.3),

which is false. So $1 \le \dim W < n$ by Theorem 2.3.1. The subspace W is T-invariant (Theorem 6.7.4 (b)), and

$$m_{T|_W} = p^{e-1}.$$

Indeed $p^{e-1}(T)$ kills W, because $p^{e-1}(T)Pv = m_T(T)v = 0$ for every v; and $p^{e-2}(T)$ does not, because $p^{e-2}(T)P \neq 0$ — that operator is $p^{e-1}(T)$, which is nonzero since $m_T = p^e \nmid p^{e-1}$ — so some $Pv \in W$ survives it. Hence $m_{T|_W}$ is a monic divisor of p^{e-1} but not of p^{e-2}, and Lemma 9.5.2 leaves only p^{e-1}. By the induction hypothesis applied to $T|_W$,

$$W = Z(w_1;T) \oplus \cdots \oplus Z(w_q;T), \qquad m_{T,w_i} = p^{l_i}, \qquad e-1 = l_1 \ge \cdots \ge l_q \ge 1.$$

Step 2. Lengthen each piece by one step. Each w_i lies in $W = \operatorname{im} P$, so we may choose $u_i \in V$ with $Pu_i = w_i$, for $i = 1, \dots, q$. Put $k_i = l_i + 1$. Then $p^{k_i}(T)u_i = p^{l_i}(T)w_i = 0$, so $m_{T,u_i} \mid p^{k_i}$; and $p^{l_i}(T)u_i \neq 0$, since otherwise $p^{l_i-1}(T)w_i = p^{l_i}(T)u_i = 0$, contradicting $m_{T,w_i} = p^{l_i}$. By Lemma 9.5.2,

$$m_{T,u_i} = p^{k_i} = p^{l_i+1} \qquad (1 \le i \le q).$$

Step 3. Fill up $\ker P$**.** Put $z_i = p^{l_i}(T)u_i$ for $i \le q$. Each is nonzero by Step 2, lies in $U = \ker P$ because $Pz_i = p^{k_i}(T)u_i = 0$, and lies in $Z(w_i;T)$, since $l_i \ge 1$ gives $z_i = p^{l_i-1}(T)p(T)u_i = p^{l_i-1}(T)w_i$. Consequently $Z(z_i;T) \subseteq Z(w_i;T)$ (Proposition 10.5.1 (b), (c)), so the sum $Z(z_1;T)+\cdots+Z(z_q;T)$ is direct: a vanishing sum of vectors $c_i \in Z(z_i;T) \subseteq Z(w_i;T)$ forces every $c_i = 0$ by the directness of Step 1 (Theorem 1.7.5).

Now $U \neq \{0\}$, U is T-invariant (Theorem 6.7.4 (b)), and $p(T|_U) = 0$ with p irreducible, so $m_{T|_U} = p$ as in the case $e = 1$. Among all lists $(z_1, \dots, z_q, z_{q+1}, \dots, z_s)$ of nonzero vectors of U extending $z_1, \dots, z_q$ and with $Z(z_1;T)+\cdots+Z(z_s;T)$ direct, choose one with s as large as possible; this is possible because each summand has dimension d, by the case $e = 1$ applied to $T|_U$, so $sd \le \dim U$. Let S be that direct sum. If $S \neq U$, pick $y \in U \setminus S$. The subspace $Z(y;T) \cap S$ is T-invariant (Proposition 9.1.4), lies inside $Z(y;T)$, and does not contain y; by the case $e = 1$, applied to $T|_U$, it is therefore $\{0\}$. Then $S + Z(y;T)$ is direct — a relation $c_1 + \cdots + c_s + c = 0$ with $c_j \in Z(z_j;T)$ and $c \in Z(y;T)$ puts $c \in Z(y;T) \cap S = \{0\}$, and directness of S kills the rest (Theorem 1.7.5) — contradicting the maximality of s. Hence

$$U = Z(z_1;T) \oplus \cdots \oplus Z(z_s;T), \qquad m_{T,z_j} = p \ \text{ for every } j.$$

Put $m = s$, and for $q < i \le m$ put $u_i = z_i$ and $k_i = 1$, so that $m_{T,u_i} = p^{k_i}$ for those i as well.

Step 4. The pieces fill V**.** First count. By Theorem 10.5.2 (a), $\dim Z(u_i;T) = \deg p^{k_i} = dk_i$. Hence

$$\begin{aligned}\sum_{i=1}^{m} \dim Z(u_i;T) &= \sum_{i=1}^{q} d(l_i+1) + (m-q)d \\ &= \Big(\sum_{i=1}^{q} dl_i\Big) + md = \dim W + \dim U = n,\end{aligned}$$

where $\sum_i dl_i = \sum_i \dim Z(w_i;T) = \dim W$ by Step 1, $md = \sum_j \dim Z(z_j;T) = \dim U$ by Step 3, and the last equality is Theorem 2.3.1.

Next, directness. Let $y_1+\cdots+y_m = 0$ with $y_i \in Z(u_i;T)$, say $y_i = f_i(T)u_i$ (Proposition 10.5.1 (a)). Apply P. For $i > q$ we get $Py_i = f_i(T)Pu_i = 0$, since $u_i \in U$. For $i \le q$ we get

$Py_i = f_i(T)w_i \in Z(w_i;T)$. So

$$\sum_{i=1}^{q} f_i(T)w_i = 0,$$

and the directness of Step 1 forces $f_i(T)w_i = 0$ for every $i \le q$, that is, $p^{l_i} \mid f_i$ (Definition 10.5.2). Write $f_i = p^{l_i} g_i$ for $i \le q$; then

$$y_i = g_i(T)\,p^{l_i}(T)u_i = g_i(T)z_i \in Z(z_i;T) \qquad (i \le q).$$

For $i > q$, $y_i \in Z(u_i;T) = Z(z_i;T)$ by definition. So the relation $y_1 + \cdots + y_m = 0$ is a relation between vectors of the independent subspaces $Z(z_1;T), \ldots, Z(z_m;T)$ of Step 3, and every $y_i = 0$. By Theorem 1.7.5 ((b) $\Rightarrow$ (a)) the sum $\sum_i Z(u_i;T)$ is direct, and by the count its dimension is n, so it is all of V (Theorem 1.6.8).

Finally the exponents. We have $k_1 = l_1 + 1 = e$, the k_i with $i \le q$ decrease because the l_i do, and the remaining ones equal $1 \le k_q$; so after listing the pieces in this order, $e = k_1 \ge \cdots \ge k_m \ge 1$. The dimension formula $\dim V = d\sum_i k_i$ is the count above. This completes the induction.

■

Two sanity checks. For $p = x$, so that T is nilpotent with $m_T = x^e$, the theorem returns Theorem 10.2.3: $Z(u;T) = \operatorname{span}(u, Tu, \ldots, T^{k-1}u)$ is the span of a Jordan chain written backwards, and $d = 1$, so $\dim V = \sum_i k_i$. And for $e = 1$ it returns Exercise (C2: Irreducible minimal polynomials): all pieces have dimension d, so $d \mid \dim V$.

10.6.3 Invariant factors

Now Steps 1 and 3 of the plan. Merging pieces across different primes is where Lemma 10.5.4 earns its keep.

Theorem 10.6.2. Cyclic Decomposition Theorem

Let $V \neq \{0\}$ be a finite-dimensional vector space over F and let $T \in \mathcal{L}(V)$. Then there are an integer $r \ge 1$, vectors $v_1, \ldots, v_r \in V$, and monic **nonconstant** polynomials $d_1, \ldots, d_r \in F[x]$ such that

$$V = Z(v_1;T) \oplus \cdots \oplus Z(v_r;T), \qquad m_{T,v_j} = d_j,$$

and

(a) $d_1 \mid d_2 \mid \cdots \mid d_r$;

(b) $d_r = m_T$;

(c) $d_1 d_2 \cdots d_r = p_T$; in particular $\sum_j \deg d_j = \dim V$;

(d) there is a **finer** decomposition $V = \bigoplus_{i,j} Z(v_{ij};T)$ into cyclic subspaces whose annihilators $m_{T,v_{ij}}$ are powers of monic irreducibles, and the list of these prime powers, with repetitions, consists exactly of the prime powers occurring in the factorizations of $d_1, \ldots, d_r$.

Idea. Run the plan. Primary decomposition gives one piece per prime; Theorem 10.6.1 breaks each piece into cyclic subspaces with annihilators $p_i^{e_{i1}}, p_i^{e_{i2}}, \ldots$, the exponents decreasing. Arrange these in an array, one row per prime, with the largest exponent of each prime in the

first column. Multiplying down a column gives one member of the chain, and the columns are numbered **backwards**: the first column, which carries the largest exponent of every prime, gives $d_r = m_T$; the second gives d_{r-1}; the last gives d_1. The chain $d_1 \mid \cdots \mid d_r$ comes out for free, because the exponents in a row decrease from left to right. The generator for a column is the sum of the generators in it, whose annihilator is the product by Lemma 10.5.4.

Proof. **Step 1.** By Theorem 6.3.5, $m_T = p_1^{e_1} \cdots p_k^{e_k}$ with $p_1, \dots, p_k$ distinct monic irreducibles and $e_i \geq 1$; here $\deg m_T \geq 1$ since $V \neq \{0\}$ (Theorem 9.5.1). By Theorem 9.8.1, $V = V_1 \oplus \cdots \oplus V_k$ with V_i nonzero and T-invariant and $m_{T|_{V_i}} = p_i^{e_i}$.

Step 2. Fix i and apply Theorem 10.6.1 to $T|_{V_i}$. Since $Z(v; T|_{V_i}) = Z(v; T)$ for $v \in V_i$, it gives vectors $v_{i1}, \dots, v_{im_i} \in V_i$ and exponents $e_i = e_{i1} \geq \cdots \geq e_{im_i} \geq 1$ with

$$V_i = Z(v_{i1}; T) \oplus \cdots \oplus Z(v_{im_i}; T), \qquad m_{T,v_{ij}} = p_i^{e_{ij}}.$$

Combining with Step 1, V is the direct sum of **all** the subspaces $Z(v_{ij}; T)$: a basis of V is obtained by concatenating bases of the V_i, each of which is a concatenation of bases of its own pieces, so Theorem 1.7.5 ((d) $\Rightarrow$ (a)) applies.

Step 3. Let $r = \max_i m_i$ and, for $1 \leq j \leq r$, put $j' = r + 1 - j$ and

$$d_j = \prod_{i\,:\,m_i \geq j'} p_i^{e_{ij'}}, \qquad v_j = \sum_{i\,:\,m_i \geq j'} v_{ij'}.$$

So d_r collects the largest exponent of every prime and d_1 the smallest ones still present.

The divisibility chain. The exponent of p_i in d_j is $e_{ij'}$ if $m_i \geq j'$ and 0 otherwise. As j increases, j' decreases, so this exponent is nondecreasing in j (the exponents $e_{i1} \geq e_{i2} \geq \dots$ decrease in the second index, and a 0 is followed only by values ≥ 0). Hence $d_j \mid d_{j+1}$ for every j (Lemma 9.5.2), which is (a). Taking $j = r$, so $j' = 1$, gives $d_r = \prod_i p_i^{e_{i1}} = \prod_i p_i^{e_i} = m_T$, which is (b). Each d_j is nonconstant: choosing i_0 with $m_{i_0} = r$, we have $m_{i_0} \geq j'$ for every j, so the factor $p_{i_0}^{e_{i_0 j'}}$ occurs in d_j, with exponent at least 1.

The annihilators and the pieces. For fixed j, the polynomials $p_i^{e_{ij'}}$ appearing in d_j are pairwise coprime, being powers of distinct monic irreducibles: a monic common divisor of p_i^a and $p_{i'}^b$ with $i \neq i'$ is a power of each, hence 1 by the uniqueness of factorization (Lemma 9.5.2, Theorem 6.3.5 (b)). So Lemma 10.5.4 gives $m_{T,v_j} = d_j$. Moreover

$$Z(v_j; T) = \bigoplus_{i\,:\,m_i \geq j'} Z(v_{ij'}; T)\,:$$

the left side is contained in the right, because the right side is a T-invariant subspace containing v_j (Proposition 9.1.4, Proposition 10.5.1 (c)), and the two have the same dimension,

$$\dim Z(v_j; T) = \deg d_j = \sum_i \deg p_i^{e_{ij'}} = \sum_i \dim Z(v_{ij'}; T),$$

by Theorem 10.5.2 (a) and Theorem 1.7.5 ((a) $\Rightarrow$ (e)); so they are equal (Theorem 1.6.8).

The decomposition. Each subspace $Z(v_{ij_0}; T)$, with $1 \leq j_0 \leq m_i$, occurs in exactly one of the sums just displayed, namely the one with $j = r + 1 - j_0$; so every pair (i, j_0) is used exactly

once and

$$\sum_{j=1}^{r} Z(v_j;T) = \sum_{i,j} Z(v_{ij};T) = V,$$

and $\sum_j \dim Z(v_j;T) = \sum_{i,j} \dim Z(v_{ij};T) = \dim V$. By Theorem 1.7.5 ((e) $\Rightarrow$ (a)) the sum over j is direct, and it equals V.

(c) Each $Z(v_j;T)$ is T-invariant and nonzero, so choose a basis $\mathcal{B}_j$ of $Z(v_j;T)$ for each j and let $\mathcal{B}$ be $\mathcal{B}_1,\dots,\mathcal{B}_r$ written one after another. By Theorem 9.1.5, $\mathcal{B}$ is a basis of V and $[T]_{\mathcal{B}}^{\mathcal{B}} = [T|_{Z(v_1;T)}]_{\mathcal{B}_1}^{\mathcal{B}_1} \oplus \cdots \oplus [T|_{Z(v_r;T)}]_{\mathcal{B}_r}^{\mathcal{B}_r}$. Then $x\mathbf{I} - [T]_{\mathcal{B}}^{\mathcal{B}}$ is block diagonal with the blocks $x\mathbf{I} - [T|_{Z(v_j;T)}]_{\mathcal{B}_j}^{\mathcal{B}_j}$, so Theorem 7.8.4, applied $r-1$ times, gives

$$p_T = \det\left(x\mathbf{I} - [T]_{\mathcal{B}}^{\mathcal{B}}\right) = \prod_{j=1}^{r} p_{T|_{Z(v_j;T)}}$$

(Definition 7.10.2). By Theorem 10.5.2 (c), $p_{T|_{Z(v_j;T)}} = m_{T,v_j} = d_j$. Hence $p_T = d_1 \cdots d_r$, and comparing degrees gives $\sum_j \deg d_j = \deg p_T = \dim V$ (Theorem 7.10.2).

(d) Step 2 already produced the finer decomposition $V = \bigoplus_{i,j} Z(v_{ij};T)$, with $m_{T,v_{ij}} = p_i^{e_{ij}}$ a prime power. By the definition of d_j in Step 3, the exponent of p_i in d_j is $e_{i,r+1-j}$ when $m_i \geq r+1-j$ and 0 otherwise, and each pair (i,j) with $1 \leq j \leq m_i$ is used exactly once. So the prime powers occurring in $d_1,\dots,d_r$ are exactly the polynomials $p_i^{e_{ij}}$, with the same repetitions.

■

The polynomials in the chain, and the prime powers they are built from, are the data we are after — but first we must know that they do not depend on the decomposition we happened to construct. The count that settles this is the count of Section 2 in disguise: there the chain lengths were read off from the ranks of the powers of N, and here the exponents are read off from the dimensions of the kernels of the powers of $p(T)$. Write $v_p(f)$ for the exponent of a monic irreducible p in the factorization of a nonzero $f \in F[x]$, so that $v_p(f) \geq 1$ exactly when $p \mid f$.

Proposition 10.6.3. Counting the Exponents

Let $V \neq \{0\}$ be finite-dimensional over F, let $T \in \mathcal{L}(V)$, and suppose

$$V = Z(u_1;T) \oplus \cdots \oplus Z(u_N;T), \qquad m_{T,u_i} = g_i,$$

with every g_i monic and nonconstant. Then for every monic irreducible $p \in F[x]$:

(a) $\dim \ker p(T)^k = \deg p \cdot \sum_{i=1}^{N} \min\left(k, v_p(g_i)\right)$ for every integer $k \geq 0$;

(b) for every $k \geq 1$,

$$\#\{\, i : v_p(g_i) \geq k \,\} = \frac{\dim \ker p(T)^k - \dim \ker p(T)^{k-1}}{\deg p},$$

a number depending only on T, p, and k.

Idea. Kernels respect the decomposition, so we may work inside one piece. There a vector is $h(T)u$, and it is killed by $p(T)^k$ exactly when $g \mid p^k h$, which after canceling the common factor $\gcd(g, p^k)$ says that h is a multiple of $g/\gcd(g, p^k)$. The surviving vectors are therefore a cyclic subspace again, generated by that multiple of u, and its annihilator is $\gcd(g, p^k) = p^{\min(k, v_p(g))}$. Summing over the pieces and taking differences in k then counts exponents.

Proof.

(a) Write $Z_i = Z(u_i; T)$ and $S = p(T)^k$. Each Z_i is T-invariant, hence S-invariant (Theorem 6.7.4 (b)). If $v = v_1 + \cdots + v_N$ with $v_i \in Z_i$, then $Sv = Sv_1 + \cdots + Sv_N$ with $Sv_i \in Z_i$, so $Sv = 0$ if and only if every $Sv_i = 0$, by the uniqueness of the splitting (Theorem 1.7.5). Therefore

$$\ker S = (\ker S \cap Z_1) \oplus \cdots \oplus (\ker S \cap Z_N),$$

and it suffices to compute each $\dim(\ker S \cap Z_i)$. Fix i and write $u = u_i$, $g = g_i$, $Z = Z_i$. Put

$$c = \gcd(g, p^k) = p^{\min(k,\, v_p(g))} \qquad \text{and} \qquad t = g/c,$$

the formula for c holding because a monic common divisor of g and p^k is a power of p (Lemma 9.5.2) dividing both, and $p^s \mid g$ exactly when $s \le v_p(g)$ (Theorem 6.3.5). Every element of Z is $h(T)u$ for some $h \in F[x]$ (Proposition 10.5.1 (a)), and

$$p(T)^k h(T)u = 0 \iff g \mid p^k h \iff t \mid h,$$

where the second equivalence holds because $g = ct$ and $p^k = c\,(p^k/c)$ with $\gcd(t, p^k/c) = 1$: from $ct \mid c(p^k/c)h$ we get $t \mid (p^k/c)h$, hence $t \mid h$ (Theorem 6.2.8), and conversely $t \mid h$ gives $g = ct \mid c\,h \mid p^k h$ since $c \mid p^k$. So

$$\ker S \cap Z = \{\, h(T)u : t \mid h \,\} = \{\, s(T)\big(t(T)u\big) : s \in F[x] \,\} = Z\big(t(T)u;\, T\big).$$

Its generator has annihilator c: for $s \in F[x]$, $s(T)t(T)u = 0$ if and only if $g \mid st$, that is, $c \mid s$, because $g = ct$. Taking $s = 1$ shows that $t(T)u \neq 0$ unless $c = 1$, so Theorem 10.5.2 (a) applies whenever $c \neq 1$ and gives

$$\dim(\ker S \cap Z_i) = \deg c = \min\big(k, v_p(g_i)\big) \deg p,$$

where for $c = 1$ the cyclic subspace is $\{0\}$ of dimension 0, in agreement. Adding over i gives (a).

(b) Subtract the values of (a) at k and $k-1$. Each term $\min(k, v_p(g_i)) - \min(k-1, v_p(g_i))$ equals 1 if $v_p(g_i) \ge k$ and 0 otherwise, so the difference is $\deg p$ times the number of indices i with $v_p(g_i) \ge k$. The left-hand side is defined from T, p, and k alone.

■

Corollary 10.6.4. Uniqueness of the Chain and of the Prime Powers

Let $V \neq \{0\}$ be finite-dimensional over F and $T \in \mathcal{L}(V)$.

(a) Any two decompositions as in Theorem 10.6.2 give the **same** ordered list $d_1 \mid \cdots \mid d_r$.

(b) Any decomposition $V = Z(u_1;T)\oplus\cdots\oplus Z(u_N;T)$ whose annihilators m_{T,u_i} are powers of monic irreducibles gives the same list of annihilators, up to order, namely the prime powers occurring in $d_1,\dots,d_r$.

(c) Explicitly, writing

$$c_p(k) = \frac{\dim\ker p(T)^k - \dim\ker p(T)^{k-1}}{\deg p} \qquad (p \text{ monic irreducible},\ k\ge 1),$$

the list is recovered by $r = \max_p c_p(1)$ and

$$v_p(d_j) = \#\{\, k\ge 1 : c_p(k)\ge r+1-j\,\} \qquad (1\le j\le r).$$

In particular the invariant factors of T are determined by the numbers $\dim\ker p(T)^k$ alone.

Idea. Divisibility puts the zeros first: in the list $v_p(d_1)\le\cdots\le v_p(d_r)$, the indices where p actually occurs are a block at the **right** end. So a count of how many d_j are divisible by p^k pins down *which* ones they are, and therefore each individual exponent. The counts are supplied by Proposition 10.6.3, which sees only T.

Proof.

(a) and (c). Let $V=\bigoplus_{j=1}^{r} Z(v_j;T)$ be such a decomposition, with annihilators $d_1\mid\cdots\mid d_r$, all monic and nonconstant. Fix a monic irreducible p. By Proposition 10.6.3 (b), applied to this decomposition,

$$c_p(k) = \#\{\, j : v_p(d_j)\ge k\,\} \qquad (k\ge 1),$$

so the left-hand side, which is defined from T, p, and k alone, computes the right-hand side. Divisibility makes the exponents nondecreasing, $v_p(d_1)\le\cdots\le v_p(d_r)$, so for each k the set $\{\, j : v_p(d_j)\ge k\,\}$ is the set of the **last** $c_p(k)$ indices:

$$v_p(d_j)\ge k \iff j\ge r+1-c_p(k) \iff c_p(k)\ge r+1-j. \qquad (*)$$

First, r is determined. Each d_j is nonconstant, so $v_p(d_1)\ge 1$ for some monic irreducible p; for that p, $(*)$ with $k=1$ and $j=1$ gives $c_p(1)\ge r$, while $c_p(1)\le r$ always. Hence $r=\max_p c_p(1)$. Next, for each p and each j, counting the $k\ge 1$ for which $(*)$ holds gives

$$v_p(d_j) = \#\{\, k\ge 1 : c_p(k)\ge r+1-j\,\},$$

a finite number, since $c_p(k)=0$ for $k>\deg m_T$. This proves (c). Finally each d_j is monic with determined exponents $v_p(d_j)$ for every p, so d_j is determined (Theorem 6.3.5), which proves (a).

(b) Let $V=\bigoplus_i Z(u_i;T)$ with $m_{T,u_i}=q_i^{f_i}$, q_i monic irreducible, and $f_i\ge 1$. For a monic irreducible p we have $v_p(q_i^{f_i}) = f_i$ if $q_i=p$ and 0 otherwise. Applying Proposition 10.6.3 (b) to this decomposition and to the one in (a) gives, for every p and every $k\ge 1$,

$$\#\{\, i : q_i=p,\ f_i\ge k\,\} = c_p(k) = \#\{\, j : v_p(d_j)\ge k\,\}.$$

Hence, for each p, the exponents f_i with $q_i=p$ and the positive exponents $v_p(d_j)$ form the same multiset. So the list $(q_i^{f_i})_i$ is exactly the list of prime powers occurring in

$d_1, \dots, d_r$, with the same repetitions. By Theorem 10.6.2 (d) such a decomposition exists, so both lists are nonempty and agree.

■

Definition 10.6.1. Invariant Factors, Elementary Divisors

Let $V \neq \{0\}$ be finite-dimensional over F and $T \in \mathcal{L}(V)$. The **invariant factors** of T are the monic nonconstant polynomials

$$d_1 \mid d_2 \mid \cdots \mid d_r$$

of Theorem 10.6.2, which by Corollary 10.6.4 (a) do not depend on the decomposition. The **elementary divisors** of T are the prime powers occurring in the factorizations of $d_1, \dots, d_r$, listed with repetition; by Corollary 10.6.4 (b) they are the annihilators of the pieces in any decomposition of V into cyclic subspaces with prime-power annihilators. For $\mathbf{A} \in M_n(F)$ the invariant factors and elementary divisors are those of $T_{\mathbf{A}}$.

In words: the invariant factors are the annihilators of the cyclic pieces, ordered so that each divides the next; the largest is m_T, and their product is p_T. The elementary divisors are the same information cut along the primes. Each list determines the other: one by factoring, the other by the regrouping recipe of Step 3 in the proof of Theorem 10.6.2, which Corollary 10.6.4 (c) shows to be forced.

Nonexample by minimal change. Drop the divisibility clause and keep only the product. For an operator on a $\mathbb{Q}$-space of dimension 3 with $p_T = (x-1)(x-2)^2$, the pair

$$d_1 = x - 1, \qquad d_2 = (x-2)^2$$

multiplies to p_T and consists of monic nonconstant polynomials of the right total degree. It is still not a list of invariant factors of anything, because $d_1 \nmid d_2$. The clause that fails is (a) of Theorem 10.6.2, and with it (b): the largest member of the list has to be m_T, and $m_T(T)$ kills every vector of V, so m_T is a multiple of the annihilator of every piece (Theorem 10.5.2 (d)). The genuine lists here are $((x-1)(x-2)^2)$ and $(x-2,\ (x-1)(x-2))$.

Quick check. Let T be an operator on a $\mathbb{Q}$-vector space with $p_T = (x^2+1)^2$. How many possibilities are there for the list of elementary divisors, and what are the corresponding invariant factors?

Solution. The product of the elementary divisors is p_T (by Theorem 10.6.2 (c) and Definition 10.6.1), and x^2+1 is irreducible over $\mathbb{Q}$ (no rational root, Theorem 6.3.1 (b)). So the elementary divisors are powers of x^2+1 with exponents forming a partition of 2: either $(x^2+1)^2$ alone, or x^2+1 twice. The invariant factors are $d_1 = (x^2+1)^2$ in the first case (one piece, $r = 1$) and $d_1 = d_2 = x^2+1$ in the second. The two are told apart by m_T, which is d_r: $(x^2+1)^2$ against x^2+1.

10.6.4 The rational canonical form

Everything is now assembled. Writing the matrix down is Theorem 10.5.2 (b), one piece at a time.

Every operator over every field has a block diagonal matrix whose blocks are companion matrices of a divisibility chain, and the chain is unique.

Theorem 10.6.5. Rational Canonical Form

Let $V \neq \{0\}$ be a finite-dimensional vector space over **any** field F, and let $T \in \mathcal{L}(V)$ have invariant factors $d_1 \mid d_2 \mid \cdots \mid d_r$. Then there is a basis $\mathcal{B}$ of V with

$$[T]_{\mathcal{B}}^{\mathcal{B}} = \mathbf{C}(d_1) \oplus \mathbf{C}(d_2) \oplus \cdots \oplus \mathbf{C}(d_r),$$

and this matrix is determined by T. It is called the **rational canonical form** of T. Equivalently, every $\mathbf{A} \in M_n(F)$ with $n \geq 1$ is similar to exactly one matrix of the form $\mathbf{C}(d_1) \oplus \cdots \oplus \mathbf{C}(d_r)$ with $d_1 \mid \cdots \mid d_r$ monic and nonconstant.

Proof. By Theorem 10.6.2, $V = Z(v_1;T) \oplus \cdots \oplus Z(v_r;T)$ with $m_{T,v_j} = d_j$. Each piece is T-invariant and nonzero, and by Theorem 10.5.2 (a), (b) the list $\mathcal{B}_j = (v_j, Tv_j, \ldots, T^{\deg d_j - 1}v_j)$ is a basis of $Z(v_j;T)$ in which the matrix of the restriction is $\mathbf{C}(d_j)$. Let $\mathcal{B}$ be $\mathcal{B}_1, \ldots, \mathcal{B}_r$ written one after another; by Theorem 9.1.5, $\mathcal{B}$ is a basis of V and $[T]_{\mathcal{B}}^{\mathcal{B}} = \mathbf{C}(d_1) \oplus \cdots \oplus \mathbf{C}(d_r)$. The blocks are determined by T, since the d_j are (Corollary 10.6.4 (a)).

For a matrix $\mathbf{A} \in M_n(F)$, apply this to $T_{\mathbf{A}}$ and use Theorem 4.2.4. For the uniqueness in the matrix form, suppose $\mathbf{A} \sim \mathbf{C}(f_1) \oplus \cdots \oplus \mathbf{C}(f_s)$ with $f_1 \mid \cdots \mid f_s$ monic and nonconstant. By Theorem 4.2.4 there is a basis $\mathcal{C}$ of F^n with $[T_{\mathbf{A}}]_{\mathcal{C}}^{\mathcal{C}} = \mathbf{C}(f_1) \oplus \cdots \oplus \mathbf{C}(f_s)$. Let $\mathcal{C}_j$ be the segment of $\mathcal{C}$ belonging to the j-th block and let U_j be its span; then $F^n = U_1 \oplus \cdots \oplus U_s$ with each U_j invariant and $[T_{\mathbf{A}}|_{U_j}]_{\mathcal{C}_j}^{\mathcal{C}_j} = \mathbf{C}(f_j)$ (Theorem 9.1.5). By Theorem 10.5.3, Theorem 9.5.5, and Definition 7.10.2, $m_{T_{\mathbf{A}}|_{U_j}} = p_{T_{\mathbf{A}}|_{U_j}} = f_j$, so $T_{\mathbf{A}}|_{U_j}$ is cyclic (Theorem 10.5.6) and $U_j = Z(\mathbf{u}_j; T_{\mathbf{A}})$ for some $\mathbf{u}_j$ with $m_{T_{\mathbf{A}},\mathbf{u}_j} = f_j$ (Theorem 10.5.2 (c)). This is a decomposition as in Theorem 10.6.2, so $(f_1, \ldots, f_s) = (d_1, \ldots, d_r)$ by Corollary 10.6.4 (a). Hence the matrix is the rational canonical form of $\mathbf{A}$.

■

This pays off the promise made in Section 3: **there is a canonical form over every field.** No hypothesis on F was used, and none is needed. The form is canonical in the strong sense: the blocks come in a prescribed order, dictated by divisibility, so two operators have the *same* form, not merely similar forms.

Corollary 10.6.6. Similarity via the Invariant Factors

Let $\mathbf{A}, \mathbf{B} \in M_n(F)$ with $n \geq 1$. The following are equivalent:

(a) $\mathbf{A} \sim \mathbf{B}$;

(b) $\mathbf{A}$ and $\mathbf{B}$ have the same invariant factors;

(c) $\mathbf{A}$ and $\mathbf{B}$ have the same elementary divisors, up to order;
(d) $\dim\ker p(\mathbf{A})^k = \dim\ker p(\mathbf{B})^k$ for every monic irreducible $p \in F[x]$ and every $k \geq 1$.

Proof.

(a) $\Rightarrow$ (d). If $\mathbf{B} = \mathbf{P}^{-1}\mathbf{A}\mathbf{P}$, then $p(\mathbf{B})^k = \mathbf{P}^{-1}p(\mathbf{A})^k\mathbf{P}$ (Proposition 4.2.6 (c)), so the two have the same rank (Proposition 4.2.6 (a)) and hence the same nullity (Theorem 3.5.7).

(b) $\Rightarrow$ (b). By Corollary 10.6.4 (c), the invariant factors of a matrix are computed from the numbers $\dim\ker p(\cdot)^k$ alone: those numbers give the counts $c_p(k)$, which give the length r and the exponents $v_p(d_j)$. Equal numbers therefore give equal invariant factors.

(c) $\Leftrightarrow$ (c). The elementary divisors are the prime powers in the factorizations of the invariant factors (Definition 10.6.1), so (b) $\Rightarrow$ (c). Conversely, $c_p(k)$ is the number of elementary divisors of the form p^f with $f \geq k$ (Corollary 10.6.4 (b) and (c)), so the list of elementary divisors determines every $c_p(k)$, hence the invariant factors by Corollary 10.6.4 (c); so (c) $\Rightarrow$ (b).

(d) $\Rightarrow$ (a). Both matrices are similar to $\mathbf{C}(d_1)\oplus\cdots\oplus\mathbf{C}(d_r)$ (Theorem 10.6.5), and similarity is an equivalence relation (Example (Similarity)).

■

So the invariant factors are a **complete set of similarity invariants over every field**, which is the promise of Chapter 7 kept in full; Corollary 10.3.3 was the special case in which $p_{\mathbf{A}}$ splits. Counting similarity classes becomes counting divisibility chains: over any field, the classes of $n \times n$ matrices with a given characteristic polynomial p correspond to the ways of writing $p = d_1 \cdots d_r$ with $d_1 \mid \cdots \mid d_r$ monic and nonconstant.

Warning. **The invariant factors are not the factors of $p_{\mathbf{A}}$ in any naive sense, and $m_{\mathbf{A}}$ alone is only the last of them.** For $\mathbf{A} = \operatorname{diag}(1, 1, 2)$ over $\mathbb{Q}$ the invariant factors are $d_1 = x - 1$ and $d_2 = (x-1)(x-2) = m_{\mathbf{A}}$, and $d_1 d_2 = (x-1)^2(x-2) = p_{\mathbf{A}}$; the factorization of $p_{\mathbf{A}}$ into linear factors plays no role in the chain. Nor does the chain have length $\deg p_{\mathbf{A}}$: here $r = 2$ for a 3×3 matrix, and $r = 1$ exactly when $\mathbf{A}$ is cyclic (Theorem 10.5.6).

10.6.5 Elementary divisors, and the way back to Jordan

Cutting the pieces as finely as possible gives a second form, usually with more and smaller blocks.

Theorem 10.6.7. Elementary Divisor Form

Let $V \neq \{0\}$ be finite-dimensional over F, let $T \in \mathcal{L}(V)$, and let $q_1, \ldots, q_N$ be the elementary divisors of T. Then there is a basis $\mathcal{B}$ of V with

$$[T]_{\mathcal{B}}^{\mathcal{B}} = \mathbf{C}(q_1) \oplus \mathbf{C}(q_2) \oplus \cdots \oplus \mathbf{C}(q_N),$$

and the blocks are determined by T up to their order.

Proof. By Theorem 10.6.2 (d) there is a decomposition $V = \bigoplus_{i,j} Z(v_{ij};T)$ into cyclic subspaces whose annihilators are prime powers, and by Corollary 10.6.4 (b) those annihilators are exactly the elementary divisors $q_1, \dots, q_N$ of T, up to order. Take the basis of Theorem 10.5.2 (a) in each piece and concatenate; Theorem 10.5.2 (b) and Theorem 9.1.5 give the displayed matrix. The list of blocks depends on T alone, again by Corollary 10.6.4 (b). ∎

When the characteristic polynomial splits, this is the Jordan form in disguise, because a companion block for a power of a linear polynomial is a Jordan block.

Corollary 10.6.8. Companion Blocks of Linear Powers Are Jordan Blocks

Let $\lambda \in F$ and $k \geq 1$. Then

$$\mathbf{C}((x-\lambda)^k) \sim \mathbf{J}_k(\lambda).$$

Consequently, if p_T splits over F, then the elementary divisors of T are the polynomials $(x-\lambda)^k$ for the Jordan blocks $\mathbf{J}_k(\lambda)$ of T, and the elementary divisor form of Theorem 10.6.7 and the Jordan form of T are obtained from one another by replacing each block $\mathbf{C}((x-\lambda)^k)$ by $\mathbf{J}_k(\lambda)$ and conversely.

Proof. Put $\mathbf{J} = \mathbf{J}_k(\lambda)$. Then $\mathbf{J} - \lambda\mathbf{I} = \mathbf{J}_k(0)$ satisfies $\mathbf{J}_k(0)^k = \mathbf{O} \neq \mathbf{J}_k(0)^{k-1}$, so $(x-\lambda)^k$ annihilates $\mathbf{J}$ and no smaller power does; by Lemma 9.5.2 and Theorem 9.5.1, $m_{\mathbf{J}} = (x-\lambda)^k$. Also $p_{\mathbf{J}} = (x-\lambda)^k$ (Theorem 7.4.6), so $m_{\mathbf{J}} = p_{\mathbf{J}}$ and $\mathbf{J}$ is cyclic (Theorem 10.5.6). By the last sentence of that theorem, $\mathbf{J}$ is similar to $\mathbf{C}(m_{\mathbf{J}}) = \mathbf{C}((x-\lambda)^k)$.

Now suppose p_T splits and let $\mathcal{B}$ be a Jordan basis, with $[T]_{\mathcal{B}}^{\mathcal{B}} = \mathbf{J}_{k_1}(\mu_1) \oplus \cdots \oplus \mathbf{J}_{k_m}(\mu_m)$ (Theorem 10.3.1). Each block corresponds to a T-invariant subspace U_l, spanned by the segment $\mathcal{B}_l$ of $\mathcal{B}$ belonging to it, with $V = U_1 \oplus \cdots \oplus U_m$ and $[T|_{U_l}]_{\mathcal{B}_l}^{\mathcal{B}_l} = \mathbf{J}_{k_l}(\mu_l)$. By the first paragraph, $m_{T|_{U_l}} = p_{T|_{U_l}} = (x-\mu_l)^{k_l}$, so $T|_{U_l}$ is cyclic: there is $v_l \in U_l$ with $U_l = Z(v_l;T)$ and $m_{T,v_l} = (x-\mu_l)^{k_l}$ (Theorem 10.5.6, Theorem 10.5.2 (c)). Since $x-\mu_l$ is irreducible, being of degree 1, this exhibits V as a direct sum of cyclic subspaces with prime-power annihilators, so by Corollary 10.6.4 (b) the elementary divisors of T are exactly the polynomials $(x-\mu_l)^{k_l}$. Replacing each Jordan block by the similar companion block, or conversely, therefore turns one form into the other. ∎

Remark. The two forms carry the same information in different shapes, and each is better for something. The rational canonical form of Theorem 10.6.5 has the fewest blocks and reads off $m_T = d_r$ at a glance; the elementary divisor form has the most blocks and reads off the primes, and over a splitting field it is the Jordan form. The rational form is also called the **Frobenius normal form**. The word *rational* refers to the field: the form is computed by rational operations in F, with no root extraction, which is exactly why it exists over every field.

10.6.6 Similarity does not depend on the field

Here is a striking consequence, and one that would be hard to guess. Two rational matrices might conceivably be conjugate by a complex matrix and by no rational one. They cannot be.

Lemma 10.6.9. Rank Does Not Change with the Field

Let K be a subfield of a field L and let $\mathbf{M} \in M_{m\times n}(K)$. Then the rank of $\mathbf{M}$ computed over K equals its rank computed over L.

Proof. If $\mathbf{M} = \mathbf{O}$, both ranks are 0. Otherwise, by Theorem 7.7.1 (a), each rank is the largest k such that some $k \times k$ minor of $\mathbf{M}$ is nonzero. The minors are determinants of submatrices of $\mathbf{M}$, computed by the Leibniz formula from the entries of $\mathbf{M}$ (Definition 7.4.1), so they are the **same elements of** K whichever field we regard $\mathbf{M}$ in. And an element of K is zero in K if and only if it is zero in L. So the two largest such k agree. ∎

Corollary 10.6.10. Similarity Is Independent of the Field

Let K be a subfield of L and let $\mathbf{A}, \mathbf{B} \in M_n(K)$, $n \geq 1$. If $\mathbf{A}$ and $\mathbf{B}$ are similar over L, then they are similar over K:

$$\mathbf{A} = \mathbf{Q}^{-1}\mathbf{B}\mathbf{Q} \text{ for some invertible } \mathbf{Q} \in M_n(L) \implies \mathbf{A} = \mathbf{P}^{-1}\mathbf{B}\mathbf{P} \text{ for some invertible } \mathbf{P} \in M_n(K).$$

Idea. Similarity over K is decided by the numbers $\dim\ker p(\mathbf{A})^k$ for the monic irreducibles p of $K[x]$ (Corollary 10.6.6). Each such number is a nullity of a matrix with entries in K, and nullities do not notice a larger field. Similarity over L makes the numbers agree when computed over L, hence over K.

Proof. Let $p \in K[x]$ be monic irreducible and $k \geq 1$. The matrices $p(\mathbf{A})^k$ and $p(\mathbf{B})^k$ have entries in K. Since $\mathbf{A}$ and $\mathbf{B}$ are similar over L, say $\mathbf{A} = \mathbf{Q}^{-1}\mathbf{B}\mathbf{Q}$, we have $p(\mathbf{A})^k = \mathbf{Q}^{-1}p(\mathbf{B})^k\mathbf{Q}$ (Proposition 4.2.6 (c)), so these two matrices have the same rank over L (Proposition 4.2.6 (a)). By Lemma 10.6.9 their ranks over K are the same numbers, hence

$$\dim\ker p(\mathbf{A})^k = n - \operatorname{rank} p(\mathbf{A})^k = n - \operatorname{rank} p(\mathbf{B})^k = \dim\ker p(\mathbf{B})^k$$

over K, by Theorem 3.5.7. As p and k were arbitrary, Corollary 10.6.6 ((d) ⇒ (a)) gives $\mathbf{A} \sim \mathbf{B}$ over K. ∎

The invariant factors are computed in $K[x]$ and never need a root; that is the whole reason. Note what the corollary does **not** say: the invariant factors themselves do change with the field, since the irreducible polynomials do. Over $\mathbb{Q}$ the matrix $\mathbf{C}((x^2+1)^2)$ has the single invariant factor

$(x^2+1)^2$ and the single elementary divisor $(x^2+1)^2$; over $\mathbb{C}$ its invariant factor is unchanged, but its elementary divisors are $(x-i)^2$ and $(x+i)^2$, and its Jordan form is $\mathbf{J}_2(i)\oplus\mathbf{J}_2(-i)$. What is field-independent is the **answer to the similarity question**, not the bookkeeping.

10.6.7 Computing the forms

Algorithm. **Invariant factors, elementary divisors, and the two forms of $\mathbf{A}\in M_n(F)$.**

1. Compute $p_{\mathbf{A}}$ and factor it into monic irreducibles over F: $p_{\mathbf{A}}=p_1^{a_1}\cdots p_k^{a_k}$.
2. For each i and each $k\geq 1$, compute $\dim\ker p_i(\mathbf{A})^k$ until it stops growing, and read off the exponents of the elementary divisors for p_i from Proposition 10.6.3, subtracting twice for exact counts. (Equivalently: find $m_{\mathbf{A}}$ and, in small cases, use $\sum_{i,j} e_{ij}\deg p_i = n$ with $\max_j e_{ij}=$ the exponent of p_i in $m_{\mathbf{A}}$.)
3. Group the exponents of each prime into a row, in decreasing order, and multiply down the columns, numbering them backwards: the first column gives $d_r=m_{\mathbf{A}}$, the second d_{r-1}, the last d_1.
4. Write $\mathbf{C}(d_1)\oplus\cdots\oplus\mathbf{C}(d_r)$, or $\bigoplus_{i,j}\mathbf{C}(p_i^{e_{ij}})$.
5. For a basis: find a cyclic generator for each piece, as in Section 5.

Example. Invariant Factors from Polynomial Data

Let T be an operator on a 6-dimensional $\mathbb{Q}$-vector space with

$$p_T=(x-1)^2(x^2+1)^2,\qquad m_T=(x-1)(x^2+1)^2.$$

Find the elementary divisors, the invariant factors, the rational canonical form (Theorem 10.6.5), and the elementary divisor form (Theorem 10.6.7).

Solution. The primes. $x-1$ and x^2+1 are monic irreducible over $\mathbb{Q}$ (the second has no rational root, Theorem 6.3.1 (b)), and $\deg p_T=2+4=6$ as it must be.

The exponents. The elementary divisors are powers $(x-1)^a$ and $(x^2+1)^b$. For each prime, the largest exponent occurring is its exponent in $m_T=d_r$ (Theorem 10.6.2 (b), Definition 10.6.1), and the exponents weighted by the degrees sum to $\deg p_T$ (Theorem 10.6.2 (c)). For $x-1$: the largest exponent is 1 and the total contribution is 2, so there are two elementary divisors $x-1$. For x^2+1: the largest exponent is 2 and the total contribution is $4=2\cdot 2$, so there is exactly one elementary divisor $(x^2+1)^2$. Hence the elementary divisors are

$$x-1,\qquad x-1,\qquad (x^2+1)^2.$$

The invariant factors. Arrange the exponents by prime in decreasing order:

$$x-1:\ 1,1;\qquad x^2+1:\ 2.$$

So $r=\max(2,1)=2$, and multiplying down the columns, the first giving d_2 and the second d_1,

$$d_2=(x-1)(x^2+1)^2=m_T,\qquad d_1=x-1.$$

Check: $d_1\mid d_2$ and $d_1d_2=(x-1)^2(x^2+1)^2=p_T$.

The forms. With $d_2 = (x-1)(x^2+1)^2 = (x-1)(x^4+2x^2+1) = x^5 - x^4 + 2x^3 - 2x^2 + x - 1$,

$$\mathbf{C}(d_1) \oplus \mathbf{C}(d_2) = (1) \oplus \begin{pmatrix} 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & -1 \\ 0 & 1 & 0 & 0 & 2 \\ 0 & 0 & 1 & 0 & -2 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix},$$

of total size $1 + 5 = 6$. The elementary divisor form is smaller-blocked:

$$\mathbf{C}(x-1) \oplus \mathbf{C}(x-1) \oplus \mathbf{C}((x^2+1)^2) = (1) \oplus (1) \oplus \begin{pmatrix} 0 & 0 & 0 & -1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & -2 \\ 0 & 0 & 1 & 0 \end{pmatrix},$$

of total size $1+1+4 = 6$. Read as a complex matrix, the second form splits further: over $\mathbb{C}$ we have $x^2+1 = (x-i)(x+i)$, so the elementary divisors become $x-1, x-1, (x-i)^2, (x+i)^2$ — the invariant factors d_1, d_2 are unchanged — and Corollary 10.6.8 turns the elementary divisor form into the Jordan form $(1) \oplus (1) \oplus \mathbf{J}_2(i) \oplus \mathbf{J}_2(-i)$. Over $\mathbb{Q}$ no further splitting is possible, since x^2+1 is irreducible there.

Example. The Rational Canonical Form of a Matrix over the Rationals

Let

$$\mathbf{A} = \begin{pmatrix} 0 & -1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 \end{pmatrix} \in M_4(\mathbb{Q}), \qquad p_{\mathbf{A}} = (x^2+1)^2.$$

Find the invariant factors of $\mathbf{A}$, its rational canonical form (Theorem 10.6.5), and an invertible $\mathbf{P} \in M_4(\mathbb{Q})$ realizing it.

Solution. The characteristic polynomial. In 2×2 blocks, $\mathbf{A} = \begin{pmatrix} \mathbf{R} & \mathbf{I}_2 \\ \mathbf{O} & \mathbf{R} \end{pmatrix}$ with $\mathbf{R} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$, so $x\mathbf{I} - \mathbf{A}$ is block upper triangular and $p_{\mathbf{A}} = (\det(x\mathbf{I}_2 - \mathbf{R}))^2 = (x^2+1)^2$ (Theorem 7.8.4), as stated.

The minimal polynomial. By Theorem 9.5.1 and Lemma 9.5.2, $m_{\mathbf{A}}$ is x^2+1 or $(x^2+1)^2$. Compute $\mathbf{A}^2$: its columns are $\mathbf{A}$ applied to the columns of $\mathbf{A}$. With $\mathbf{c}_1, \dots, \mathbf{c}_4$ the columns of $\mathbf{A}$,

$$\mathbf{A}\mathbf{c}_1 = \mathbf{A}(0,1,0,0) = (-1,0,0,0), \qquad \mathbf{A}\mathbf{c}_2 = \mathbf{A}(-1,0,0,0) = (0,-1,0,0),$$

$$\mathbf{A}\mathbf{c}_3 = \mathbf{A}(1,0,0,1) = (0,1,0,0) + (0,1,-1,0) = (0,2,-1,0),$$
$$\mathbf{A}\mathbf{c}_4 = \mathbf{A}(0,1,-1,0) = (-1,0,0,0) + (-1,0,0,-1) = (-2,0,0,-1).$$

So

$$\mathbf{A}^2 = \begin{pmatrix} -1 & 0 & 0 & -2 \\ 0 & -1 & 2 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}, \qquad \mathbf{A}^2 + \mathbf{I} = \begin{pmatrix} 0 & 0 & 0 & -2 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \neq \mathbf{O}.$$

Hence $m_{\mathbf{A}} = (x^2+1)^2 = p_{\mathbf{A}}$, and $\mathbf{A}$ is cyclic (Theorem 10.5.6). So $r = 1$, the single invariant factor is $d_1 = (x^2+1)^2 = x^4 + 2x^2 + 1$, and the rational canonical form is

$$\mathbf{C}(d_1) = \begin{pmatrix} 0 & 0 & 0 & -1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & -2 \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

The elementary divisor form (Theorem 10.6.7) is the same matrix, since x^2+1 is irreducible over $\mathbb{Q}$ and d_1 is already a prime power.

A cyclic vector. By Theorem 10.5.6 we need $\mathbf{v}$ with $m_{\mathbf{A},\mathbf{v}} = m_{\mathbf{A}}$; since $\mathbf{A}^2 + \mathbf{I} \neq \mathbf{O}$, pick $\mathbf{v}$ with $(\mathbf{A}^2 + \mathbf{I})\mathbf{v} \neq \mathbf{0}$. The third column of $\mathbf{A}^2 + \mathbf{I}$ is nonzero, so $\mathbf{v} = \mathbf{e}_3$ works: then $(\mathbf{A}^2 + \mathbf{I})\mathbf{v} \neq \mathbf{0}$, so $m_{\mathbf{A},\mathbf{v}} \nmid x^2 + 1$, and as $m_{\mathbf{A},\mathbf{v}}$ divides $(x^2+1)^2$ it must equal $(x^2+1)^2$ (Lemma 9.5.2). Its chain is

$$\mathbf{v} = (0,0,1,0), \quad \mathbf{A}\mathbf{v} = (1,0,0,1), \quad \mathbf{A}^2\mathbf{v} = (0,2,-1,0),$$
$$\mathbf{A}^3\mathbf{v} = \mathbf{A}(0,2,-1,0) = (-2,0,0,0) + (-1,0,0,-1) = (-3,0,0,-1),$$

reading columns of $\mathbf{A}$ as before. Hence

$$\mathbf{P} = \begin{pmatrix} 0 & 1 & 0 & -3 \\ 0 & 0 & 2 & 0 \\ 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{pmatrix}, \qquad \mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \mathbf{C}((x^2+1)^2).$$

Invertibility of $\mathbf{P}$. Expanding $\det \mathbf{P}$ along the third row, whose entries are $(1,0,-1,0)$ (Theorem 7.6.1), and then each 3×3 determinant along its own first column,

$$\begin{aligned} \det \mathbf{P} &= 1 \cdot \det \begin{pmatrix} 1 & 0 & -3 \\ 0 & 2 & 0 \\ 1 & 0 & -1 \end{pmatrix} + (-1)(-1)^{3+3} \det \begin{pmatrix} 0 & 1 & -3 \\ 0 & 0 & 0 \\ 0 & 1 & -1 \end{pmatrix} \\ &= 1 \cdot (2(-1+3)) - 0 = 4 \neq 0, \end{aligned}$$

the second determinant vanishing because of a zero row, and the first being $2 \det \begin{pmatrix} 1 & -3 \\ 1 & -1 \end{pmatrix} = 2 \cdot 2 = 4$ by expansion along its second row.

The check. The identity $\mathbf{A}\mathbf{P} = \mathbf{P}\,\mathbf{C}(d_1)$ reads column by column as $\mathbf{A}\mathbf{v} = \mathbf{A}\mathbf{v}$, $\mathbf{A}(\mathbf{A}\mathbf{v}) = \mathbf{A}^2\mathbf{v}$, $\mathbf{A}(\mathbf{A}^2\mathbf{v}) = \mathbf{A}^3\mathbf{v}$ — true by construction — and

$$\mathbf{A}(\mathbf{A}^3\mathbf{v}) = -\mathbf{v} - 2\mathbf{A}^2\mathbf{v},$$

which is the annihilator relation $\mathbf{A}^4 + 2\mathbf{A}^2 + \mathbf{I} = \mathbf{O}$ applied to $\mathbf{v}$. Verifying it directly: $\mathbf{A}(-3,0,0,-1) = -3(0,1,0,0) - (0,1,-1,0) = (0,-4,1,0)$, while

$$-\mathbf{v} - 2\mathbf{A}^2\mathbf{v} = -(0,0,1,0) - 2(0,2,-1,0) = (0,-4,2-1,0) = (0,-4,1,0). \checkmark$$

10.6.8 Exercises

10.6.8.1 A. Check your understanding

Exercise. A1

(a) State the Cyclic Decomposition Theorem, including the divisibility chain and what d_r and $\prod_j d_j$ are.

(b) Define the invariant factors and the elementary divisors of T.

(c) What is the rational canonical form of T, and over which fields does it exist?

(d) True or false: two matrices in $M_n(\mathbb{Q})$ that are similar over $\mathbb{C}$ are similar over $\mathbb{Q}$. Justify your answer.

(e) An operator on a 5-dimensional $\mathbb{Q}$-space has $p_T = (x-3)^5$ and $m_T = (x-3)^2$. What are its elementary divisors and invariant factors?

(f) True or false: $r = 1$ if and only if T is cyclic. Justify your answer.

Solution.

(a) For $V \neq \{0\}$ finite-dimensional and $T \in \mathcal{L}(V)$, there are $v_1, \dots, v_r$ with $V = Z(v_1;T) \oplus \cdots \oplus Z(v_r;T)$ and $m_{T,v_j} = d_j$, where $d_1 \mid d_2 \mid \cdots \mid d_r$ are monic and nonconstant, $d_r = m_T$, and $d_1 \cdots d_r = p_T$ (Theorem 10.6.2).

(b) The invariant factors are the polynomials $d_1 \mid \cdots \mid d_r$ of that theorem; the elementary divisors are the prime powers occurring in their factorizations, listed with repetition (Definition 10.6.1). Both are determined by T (Corollary 10.6.4).

(c) The matrix $\mathbf{C}(d_1) \oplus \cdots \oplus \mathbf{C}(d_r)$, which is $[T]_{\mathcal{B}}^{\mathcal{B}}$ for a suitable basis; it exists over **every** field (Theorem 10.6.5).

(d) True, by Corollary 10.6.10 with $K = \mathbb{Q}$ and $L = \mathbb{C}$.

(e) Elementary divisors are powers of $x-3$ with largest exponent 2 and exponents summing to 5: so $(x-3)^2, (x-3)^2, x-3$. The exponents, in decreasing order, are $2, 2, 1$; numbering the columns backwards gives $d_3 = (x-3)^2 = m_T$, $d_2 = (x-3)^2$, and $d_1 = x-3$.

(f) True. $r = 1$ means $V = Z(v_1;T)$, that is, v_1 is a cyclic vector; conversely if v is cyclic, then $d_1 = m_{T,v} = m_T = p_T$ is the only invariant factor, by Theorem 10.5.6 and the uniqueness in Corollary 10.6.4.

10.6.8.2 B. Practice

Exercise. B1: Invariant factors from given data

In each case T is an operator on a $\mathbb{Q}$-vector space with the stated characteristic and minimal polynomials. Find the elementary divisors, the invariant factors, and the rational canonical form (Theorem 10.6.5).

(a) $p_T = (x-2)^4$, $m_T = (x-2)^2$.
(b) $p_T = (x^2-2)^3$, $m_T = (x^2-2)^2$.
(c) $p_T = (x-1)^2(x+1)^2$, $m_T = (x-1)(x+1)$.

Solution.

(a) Powers of $x-2$ with largest exponent 2 and exponents summing to 4: either $2,2$ or $2,1,1$. So the data do **not** determine the answer; both

$$\text{elementary divisors } (x-2)^2, (x-2)^2 \quad \text{and} \quad (x-2)^2, x-2, x-2$$

are possible, with invariant factors $((x-2)^2, (x-2)^2)$ and $(x-2, x-2, (x-2)^2)$ and rational forms

$$\mathbf{C}((x-2)^2) \oplus \mathbf{C}((x-2)^2), \qquad (2) \oplus (2) \oplus \mathbf{C}((x-2)^2),$$

where $\mathbf{C}((x-2)^2) = \mathbf{C}(x^2-4x+4) = \begin{pmatrix} 0 & -4 \\ 1 & 4 \end{pmatrix}$. (The extra datum that decides is $\dim\ker(T-2\,\mathrm{id})$, which is 2 in the first case and 3 in the second, by Proposition 10.6.3.)

(b) x^2-2 is irreducible over $\mathbb{Q}$, since $\sqrt{2} \notin \mathbb{Q}$ (Theorem 0.2.3, Theorem 6.3.1 (b)). The exponents are powers of x^2-2 with largest 2 summing to 3: only $2,1$. So the elementary divisors are $(x^2-2)^2$ and x^2-2, the invariant factors are $d_1 = x^2-2$ and $d_2 = (x^2-2)^2 = x^4-4x^2+4$, and the rational form is

$$\mathbf{C}(x^2-2) \oplus \mathbf{C}((x^2-2)^2) = \begin{pmatrix} 0 & 2 \\ 1 & 0 \end{pmatrix} \oplus \begin{pmatrix} 0 & 0 & 0 & -4 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 4 \\ 0 & 0 & 1 & 0 \end{pmatrix},$$

of size $2+4 = 6 = \deg p_T$.

(c) Both primes have largest exponent 1 and total contribution 2, so the elementary divisors are $x-1, x-1, x+1, x+1$. Arranging by prime, $x-1 : 1,1$ and $x+1 : 1,1$, so $r=2$ and $d_2 = (x-1)(x+1) = m_T$, $d_1 = (x-1)(x+1)$. The rational form is

$$\mathbf{C}(x^2-1) \oplus \mathbf{C}(x^2-1) = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \oplus \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

(This operator is diagonalizable, with $\mathrm{diag}(1,1,-1,-1)$ as its Jordan form; the rational form is a different matrix representing the same operator.)

Exercise. B2: Rational forms of matrices

For each matrix, find $p_{\mathbf{A}}$, $m_{\mathbf{A}}$, the invariant factors, the rational canonical form (Theorem 10.6.5), and the elementary divisor form (Theorem 10.6.7).

(a) $\mathbf{A}_1 = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} \in M_3(\mathbb{Q})$.
(b) $\mathbf{A}_2 = \mathrm{diag}(1,1,2) \in M_3(\mathbb{Q})$.

(c) $\mathbf{A}_3 = \begin{pmatrix} 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 \end{pmatrix} \in M_4(\mathbb{Q})$.

Solution.

(a) Reading the columns, $\mathbf{A}_1\mathbf{e}_1 = \mathbf{e}_3$, $\mathbf{A}_1\mathbf{e}_3 = \mathbf{e}_2$, and $\mathbf{A}_1\mathbf{e}_2 = \mathbf{e}_1$, so $\mathbf{A}_1^3 = \mathbf{I}$ and the polynomial $x^3 - 1$ annihilates $\mathbf{A}_1$. Also $\mathbf{A}_1^2\mathbf{e}_1 = \mathbf{A}_1\mathbf{e}_3 = \mathbf{e}_2$, so $(\mathbf{e}_1, \mathbf{A}_1\mathbf{e}_1, \mathbf{A}_1^2\mathbf{e}_1) = (\mathbf{e}_1, \mathbf{e}_3, \mathbf{e}_2)$ is a basis of $\mathbb{Q}^3$ and $\mathbf{e}_1$ is a cyclic vector. Hence $\deg m_{\mathbf{A}_1} = 3$ (Theorem 10.5.6), and since $m_{\mathbf{A}_1}$ is monic of degree 3 dividing $x^3 - 1$, we get $m_{\mathbf{A}_1} = p_{\mathbf{A}_1} = x^3 - 1$. The single invariant factor is $d_1 = x^3 - 1$, and the rational canonical form is

$$\mathbf{C}(x^3 - 1) = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}.$$

Over $\mathbb{Q}$, $x^3 - 1 = (x-1)(x^2+x+1)$ with both factors irreducible (the quadratic has no rational root). So the elementary divisors are $x - 1$ and $x^2 + x + 1$, and the elementary divisor form is

$$(1) \oplus \mathbf{C}(x^2 + x + 1) = (1) \oplus \begin{pmatrix} 0 & -1 \\ 1 & -1 \end{pmatrix}.$$

(b) $p_{\mathbf{A}_2} = (x-1)^2(x-2)$ and $m_{\mathbf{A}_2} = (x-1)(x-2)$ (Proposition 9.5.6). Elementary divisors: $x-1, x-1, x-2$. Invariant factors: $d_2 = (x-1)(x-2) = m_{\mathbf{A}_2}$ and $d_1 = x - 1$. Rational form

$$(1) \oplus \mathbf{C}((x-1)(x-2)) = (1) \oplus \begin{pmatrix} 0 & -2 \\ 1 & 3 \end{pmatrix},$$

elementary divisor form $(1) \oplus (1) \oplus (2) = \operatorname{diag}(1, 1, 2) = \mathbf{A}_2$ itself.

(c) $\mathbf{A}_3 = \mathbf{R} \oplus \mathbf{R}$ with $\mathbf{R} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$, and $p_{\mathbf{R}} = x^2 + 1$, $m_{\mathbf{R}} = x^2 + 1$, so $p_{\mathbf{A}_3} = (x^2+1)^2$ (Theorem 7.8.4 applied to $x\mathbf{I} - \mathbf{A}_3$, which is block diagonal over $\mathbb{Q}[x]$) and $m_{\mathbf{A}_3} = \operatorname{lcm}(x^2+1, x^2+1) = x^2 + 1$ (Proposition 9.5.6). Since $x^2 + 1$ is irreducible over $\mathbb{Q}$, the elementary divisors are $x^2 + 1$ and $x^2 + 1$; $r = 2$ and $d_1 = d_2 = x^2 + 1$. Both forms are

$$\mathbf{C}(x^2+1) \oplus \mathbf{C}(x^2+1) = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \oplus \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} = \mathbf{A}_3.$$

Compare Example (The Rational Canonical Form of a Matrix over the Rationals): that matrix has the same characteristic polynomial but a different minimal polynomial, so the two are **not** similar.

Exercise. B3: Counting similarity classes over a finite field

Work over $F = \mathbb{F}_2$. For each polynomial p, find the number of similarity classes of matrices $\mathbf{A} \in M_3(\mathbb{F}_2)$ with $p_{\mathbf{A}} = p$, and list the invariant factors of each class.

(a) $p = (x+1)^3$.

(b) $p = x(x+1)^2$.

(c) $p = x^3 + x + 1$.

Hint: by Corollary 10.6.6, a class is a way of writing $p = d_1 \cdots d_r$ with $d_1 \mid \cdots \mid d_r$.

Solution.

(a) The only prime involved is $x+1$, so the elementary divisors are powers $(x+1)^a$ with exponents forming a partition of 3. The three partitions (3), (2, 1), (1, 1, 1) give

$$((x+1)^3), \qquad (x+1,\ (x+1)^2), \qquad (x+1,\ x+1,\ x+1)$$

as lists of invariant factors. So there are 3 classes, with representatives $\mathbf{C}((x+1)^3)$, $(1) \oplus \mathbf{C}((x+1)^2)$, and $\mathbf{I}_3$ (note $-1 = 1$ in $\mathbb{F}_2$).

(b) Now two primes, x with total exponent 1 and $x+1$ with total exponent 2. For x the only option is the single elementary divisor x; for $x+1$ the options are $(x+1)^2$ or $x+1, x+1$. So there are 2 classes, with invariant factors

$$(x(x+1)^2) \qquad \text{and} \qquad (x+1,\ x(x+1)).$$

In the first case $r = 1$ and the matrix is cyclic; in the second $r = 2$.

(c) $x^3 + x + 1$ has no root in $\mathbb{F}_2$ (its values at 0 and 1 are 1 and 1), so it is irreducible (Theorem 6.3.1 (b)). A prime of degree 3 in a characteristic polynomial of degree 3 can occur only with exponent 1, so the only possible list of elementary divisors is (x^3+x+1) itself: 1 class, represented by $\mathbf{C}(x^3 + x + 1)$.

10.6.8.3 C. Going deeper

Exercise. C1: Which matrices are companion matrices

Let $\mathbf{A} \in M_n(F)$ with $n \geq 1$.

(a) Prove that $\mathbf{A}$ is similar to a companion matrix if and only if $m_{\mathbf{A}} = p_{\mathbf{A}}$, and that the companion matrix is then $\mathbf{C}(p_{\mathbf{A}})$.

(b) Deduce that $\mathbf{A}$ is similar to a companion matrix if and only if its rational canonical form has exactly one block.

(c) Give an example of a matrix in $M_2(F)$ that is not similar to any companion matrix.

Solution.

(a) ($\Leftarrow$) If $m_{\mathbf{A}} = p_{\mathbf{A}}$, then $\mathbf{A}$ is cyclic (Theorem 10.5.6) and, in the basis generated by a cyclic vector, its matrix is $\mathbf{C}(m_{\mathbf{A}}) = \mathbf{C}(p_{\mathbf{A}})$; so $\mathbf{A} \sim \mathbf{C}(p_{\mathbf{A}})$ (Theorem 4.2.4).
($\Rightarrow$) Suppose $\mathbf{A} \sim \mathbf{C}(q)$ for some monic q. By Theorem 10.5.3, $m_{\mathbf{C}(q)} = p_{\mathbf{C}(q)} = q$. Similar matrices share both polynomials (Theorem 7.10.3, Theorem 9.5.5), so $m_{\mathbf{A}} = p_{\mathbf{A}} = q$, and the companion matrix is $\mathbf{C}(p_{\mathbf{A}})$.

(b) By Theorem 10.6.5 the form has one block exactly when $r = 1$, that is, when $d_1 = m_{\mathbf{A}} = p_{\mathbf{A}}$ (Theorem 10.6.2 (b), (c)). Now apply (a).

(c) $\mathbf{I}_2$: here $m = x - 1 \neq (x-1)^2 = p$. Concretely, $\mathbf{I}_2$ is similar only to itself, and no companion matrix of a monic quadratic is $\mathbf{I}_2$, since the $(2,1)$ entry of $\mathbf{C}(q)$ is $1 \neq 0$.

Exercise. C2: All similarity classes of two by two matrices over the smallest field

Work over $F = \mathbb{F}_2$, so that $M_2(\mathbb{F}_2)$ has 16 elements.

(a) List the four monic polynomials of degree 2 over $\mathbb{F}_2$ and factor each into irreducibles.

(b) For each, list the possible lists of invariant factors of a matrix $\mathbf{A} \in M_2(\mathbb{F}_2)$ with that characteristic polynomial.

(c) Conclude the total number of similarity classes in $M_2(\mathbb{F}_2)$, and give a representative of each.

(d) Check your count against the orbit sizes: $|\mathrm{GL}_2(\mathbb{F}_2)| = 6$, and the class of $\mathbf{A}$ has $6/|\{\mathbf{P} : \mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \mathbf{A}\}|$ elements.

Solution.

(a) x^2; $x^2 + 1 = (x+1)^2$; $x^2 + x = x(x+1)$; and $x^2 + x + 1$, which has no root in $\mathbb{F}_2$ and is therefore irreducible (Theorem 6.3.1 (b)).

(b) For $p = x^2$: elementary divisors x^2 (invariant factors (x^2)) or x, x (invariant factors (x, x)) — two options. For $p = (x+1)^2$: likewise two, $((x+1)^2)$ and $(x+1, x+1)$. For $p = x(x+1)$: each prime occurs once, so the elementary divisors are $x, x+1$ and there is a single list of invariant factors, $(x(x+1))$. For the irreducible $p = x^2 + x + 1$: one option, $(x^2 + x + 1)$.

(c) Total $2 + 2 + 1 + 1 = 6$ classes, with representatives

$$\mathbf{C}(x^2) = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}, \quad \mathbf{O}, \quad \mathbf{C}((x+1)^2) = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

$$\mathbf{I}_2, \quad \mathrm{diag}(0,1), \quad \mathbf{C}(x^2 + x + 1) = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix},$$

using $-1 = 1$ in $\mathbb{F}_2$. (For $(x+1)^2 = x^2 + 1$ the coefficients are $a_0 = 1, a_1 = 0$, so $\mathbf{C} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$, the coordinate swap.)

(d) The two scalar matrices $\mathbf{O}$ and $\mathbf{I}_2$ commute with everything, so their classes have $6/6 = 1$ element each. For $\mathrm{diag}(0,1)$, a matrix commutes with it exactly when it is diagonal (compare entries), and the invertible diagonal matrices over $\mathbb{F}_2$ are only $\mathbf{I}_2$, so the class has 6 elements. For $\mathbf{C}(x^2)$ and $\mathbf{C}(x^2+1)$ the commutant inside GL_2 has 2 elements — by Exercise (C1: What commutes with a cyclic operator) the commutant of a cyclic matrix is $\{a\mathbf{I} + b\mathbf{A}\}$, which here has 4 elements, of which 2 are invertible — so each class has 3 elements. For $\mathbf{C}(x^2 + x + 1)$, whose order is 3 in $\mathrm{GL}_2(\mathbb{F}_2)$, the commutant is $\{a\mathbf{I} + b\mathbf{A}\}$ again, with 3 invertible elements, so the class has 2 elements. Total:

$$1 + 1 + 6 + 3 + 3 + 2 = 16 = |M_2(\mathbb{F}_2)|. \checkmark$$

Exercise. C3: A matrix over the rationals is similar to its transpose

(a) Let $\mathbf{A}, \mathbf{B} \in M_n(\mathbb{Q})$ have the same Jordan form over $\mathbb{C}$. Prove that $\mathbf{A}$ and $\mathbf{B}$ are similar over $\mathbb{Q}$.

(b) Deduce that $\mathbf{A} \sim \mathbf{A}^\top$ over $\mathbb{Q}$ for every $\mathbf{A} \in M_n(\mathbb{Q})$.

(c) Which step of (b) would fail for a general field K in place of $\mathbb{Q}$?

Hint for (b): *Exercise (C1: A matrix is similar to its transpose).*

Solution.

(a) A matrix in $M_n(\mathbb{Q})$ is also a matrix in $M_n(\mathbb{C})$, and $p_{\mathbf{A}}$ splits over $\mathbb{C}$ (Corollary 6.5.5), so both have Jordan forms there. Having the same Jordan form, they are similar over $\mathbb{C}$ (Corollary 10.3.3). By Corollary 10.6.10 with $K = \mathbb{Q}$ and $L = \mathbb{C}$, they are similar over $\mathbb{Q}$.

(b) Let $\mathbf{A} \in M_n(\mathbb{Q})$. Over $\mathbb{C}$, $p_{\mathbf{A}}$ splits, so Exercise (C1: A matrix is similar to its transpose) (c) gives $\mathbf{A} \sim \mathbf{A}^{\top}$ over $\mathbb{C}$. Both matrices lie in $M_n(\mathbb{Q})$, so Corollary 10.6.10 gives $\mathbf{A} \sim \mathbf{A}^{\top}$ over $\mathbb{Q}$.

(c) The step that fails is the splitting: over a general field K there is no reason for $p_{\mathbf{A}}$ to split, and we have not constructed a larger field where it does. (The conclusion is nevertheless true over every field; Section 7 proves it without any splitting, using the Smith normal form of $x\mathbf{I} - \mathbf{A}$.) Note that Corollary 10.6.10 itself needs no hypothesis on K — only the existence of a suitable L does.

10.7 Polynomial Matrices and the Smith Normal Form

The rational canonical form of the previous section is a complete answer to the similarity problem, but the proof did not tell us how to *find* the invariant factors of a given matrix. This section supplies the missing algorithm, and it comes from an unexpected place: instead of working with $\mathbf{A}$, we work with the single polynomial matrix $x\mathbf{I} - \mathbf{A}$, and we reduce it by row and column operations whose scalars are allowed to be polynomials. The reduced form is diagonal, its nonconstant diagonal entries are the invariant factors, and two matrices are similar exactly when their reduced forms agree. On the way we finally pay off a promise from Chapter 4: every square matrix is similar to its transpose, over every field.

10.7.1 The space as a module over the polynomials

Everything in this chapter is built by applying **polynomials** in T to vectors: the T-annihilator of a vector, the cyclic subspace $Z(v;T)$, the kernels $\ker p(T)$. Name that structure once. For V a vector space over F and $T \in \mathcal{L}(V)$, the product

$$F[x] \times V \to V, \qquad p \cdot v := p(T)\, v$$

satisfies, by Theorem 6.7.1, exactly the vector space axioms with F replaced by $F[x]$: it is additive in each factor, $(pq)\cdot v = p\cdot(q\cdot v)$, and $1\cdot v = v$. A set with an addition and a product by the elements of a commutative ring R obeying these rules is an R**-module**, so the pair (V,T) is the same thing as an $F[x]$-module that is finite-dimensional over the constants F. We develop no general theory of modules; the point is the dictionary:

$F[x]$-module language	operator language
submodule of V	T-invariant subspace (Definition 9.1.1)
submodule generated by v, that is $F[x] \cdot v$	cyclic subspace $Z(v;T)$ (Definition 10.5.1)
the annihilator of v in $F[x]$	the T-annihilator of v
V is generated by one element	T has a cyclic vector

A subspace closed under multiplication by x is closed under multiplication by every polynomial, which is why the first two lines match.

Where does $x\mathbf{I} - \mathbf{A}$ come from here? Fix $\mathcal{B} = (b_1, \dots, b_n)$ with $\mathbf{A} = [T]_{\mathcal{B}}^{\mathcal{B}}$. As an $F[x]$-module, V is generated by $b_1, \dots, b_n$, subject to the n relations $x \cdot b_j = \sum_i a_{ij} b_i$, whose coefficient vectors are the columns of $x\mathbf{I} - \mathbf{A}$. So $x\mathbf{I} - \mathbf{A}$ is the **matrix of relations**, and changing generators or replacing one relation by a combination of relations should not change V. That is the intuition behind the operations we now allow; the proofs below are matrix proofs and do not rely on it.

10.7.2 Row and column operations with polynomial multipliers

A matrix in $M_n(F[x])$ will be called a **polynomial matrix**. Over a field we reduced a matrix by three kinds of row operation (Definition 3.2.1). Over $F[x]$ we must be careful with one of them: multiplying a row by x cannot be undone, because x has no inverse in $F[x]$. So the scaling operation keeps its scalar in F, while the operation that adds a multiple of one row to another may use any polynomial multiplier.

Definition 10.7.1. Elementary Operations over $F[x]$

Let $\mathbf{M} \in M_n(F[x])$. An **elementary row operation over** $F[x]$ on $\mathbf{M}$ is one of:

(E1) interchanging two rows;

(E2) multiplying one row by a **nonzero scalar** $c \in F$;

(E3) adding h times one row to a **different** row, where $h \in F[x]$ is **any** polynomial.

Elementary column operations over $F[x]$ are defined the same way with "column" in place of "row".

In words: (E1) and (E3) are as over a field, except that (E3) now permits a polynomial multiplier; (E2) is restricted, because only the nonzero constants have inverses in $F[x]$. Each of the three is reversible by an operation of the same kind: swap back, scale by c^{-1}, add $-h$ times the row back.

Two polynomial matrices reachable from one another by such operations should count as the same. As over a field (Definition 4.3.1), the clean way to say this is with invertible matrices on both sides — but "invertible" must now mean invertible *inside* $M_n(F[x])$.

Proposition 10.7.1. Invertible Polynomial Matrices

Let $\mathbf{U} \in M_n(F[x])$. Then $\mathbf{U}$ has an inverse in $M_n(F[x])$ if and only if $\det \mathbf{U}$ is a **nonzero constant**.

Proof. ($\Rightarrow$) Suppose $\mathbf{UW} = \mathbf{I}$ with $\mathbf{W} \in M_n(F[x])$. Taking determinants and using multiplicativity over the commutative ring $F[x]$ (Theorem 7.5.4 and the remark on commutative rings after it), $\det \mathbf{U} \cdot \det \mathbf{W} = 1$. By Theorem 0.8.3, $\deg(\det \mathbf{U}) + \deg(\det \mathbf{W}) = \deg 1 = 0$, and degrees of nonzero polynomials are ≥ 0, so $\deg(\det \mathbf{U}) = 0$; that is, $\det \mathbf{U} \in F \setminus \{0\}$.

($\Leftarrow$) Suppose $\det \mathbf{U} = c \in F \setminus \{0\}$. The adjugate identity

$$\mathbf{U} \operatorname{adj} \mathbf{U} = \operatorname{adj} \mathbf{U}\ \mathbf{U} = (\det \mathbf{U})\, \mathbf{I}$$

holds over any commutative ring (Theorem 7.6.2 and the remark on commutative rings following it), and $\operatorname{adj} \mathbf{U} \in M_n(F[x])$, since its entries are minors of $\mathbf{U}$. So $\mathbf{W} = c^{-1} \operatorname{adj} \mathbf{U} \in M_n(F[x])$ satisfies $\mathbf{UW} = \mathbf{WU} = \mathbf{I}$.

■

Definition 10.7.2. Equivalence over $F[x]$

Let $\mathbf{M}, \mathbf{M}' \in M_n(F[x])$. We say $\mathbf{M}$ and $\mathbf{M}'$ are **equivalent over** $F[x]$, written $\mathbf{M} \approx \mathbf{M}'$, if there are $\mathbf{U}, \mathbf{V} \in M_n(F[x])$, **each invertible in** $M_n(F[x])$, with

$$\mathbf{M}' = \mathbf{UMV}.$$

Proposition 10.7.2. Operations and Equivalence

Let $\mathbf{M}, \mathbf{M}' \in M_n(F[x])$.

(a) $\approx$ is an equivalence relation on $M_n(F[x])$.

(b) If $\mathbf{M}'$ is obtained from $\mathbf{M}$ by a finite sequence of elementary row and column operations over $F[x]$, then $\mathbf{M} \approx \mathbf{M}'$.

Proof.

(a) Take $\mathbf{U} = \mathbf{V} = \mathbf{I}$ for reflexivity. If $\mathbf{M}' = \mathbf{UMV}$, then $\mathbf{M} = \mathbf{U}^{-1}\mathbf{M}'\mathbf{V}^{-1}$, and $\mathbf{U}^{-1}, \mathbf{V}^{-1}$ are invertible in $M_n(F[x])$; this gives symmetry. If $\mathbf{M}' = \mathbf{UMV}$ and $\mathbf{M}'' = \mathbf{U}'\mathbf{M}'\mathbf{V}'$, then $\mathbf{M}'' = (\mathbf{U}'\mathbf{U})\mathbf{M}(\mathbf{VV}')$, and a product of invertible matrices is invertible, which gives transitivity.

(b) Each elementary row operation over $F[x]$ is left multiplication by the matrix obtained by performing that operation on $\mathbf{I}_n$, exactly as over a field (Theorem 3.4.1, whose proof uses only the ring operations). Those three matrices are $\mathbf{P}_{ij}$, $\mathbf{D}_i(c)$, and $\mathbf{I} + h\mathbf{E}_{ij}$ with $i \neq j$, of determinants -1, $c \neq 0$, and 1 respectively (Theorem 7.4.6 and Theorem 7.5.1, valid over a commutative ring), hence invertible in $M_n(F[x])$ by Proposition 10.7.1. Column operations are right multiplication by the same matrices. Now use (a).

■

Warning. **Multiplying a row by a nonconstant polynomial is not an allowed operation.** It changes the answer. The matrix $\mathbf{M} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ becomes $\begin{pmatrix} x & 0 \\ 0 & 1 \end{pmatrix}$ after "multiplying

row 1 by x", and the two are not equivalent: their determinants are 1 and x, and by Proposition 10.7.1 the first is invertible over $F[x]$ while the second is not. The rule to remember is that only the **units** of $F[x]$ — the nonzero constants — may be used as scaling factors.

10.7.3 The Smith normal form

Every polynomial matrix can be brought to a diagonal form whose diagonal entries divide one another in turn, and that form is unique.

The existence proof is Euclidean division, run on whichever nonzero entry has the smallest degree: if that entry fails to divide one of its neighbors in the same row or column, the remainder gives a nonzero entry of strictly smaller degree, and degrees cannot decrease forever.

Theorem 10.7.3. Smith Normal Form

Let F be a field, $n \geq 1$, and $\mathbf{M} \in M_n(F[x])$. Then:

(a) **(Existence)** $\mathbf{M}$ is equivalent over $F[x]$, by a finite sequence of elementary row and column operations, to a diagonal matrix

$$\mathbf{D} = \operatorname{diag}(d_1, d_2, \ldots, d_n)$$

in which every d_i is **monic or zero** and $d_i \mid d_{i+1}$ for $i = 1, \ldots, n-1$.

(b) **(Uniqueness)** The list $(d_1, \ldots, d_n)$ is determined by $\mathbf{M}$: if $\mathbf{M} \approx \operatorname{diag}(d_1, \ldots, d_n) \approx \operatorname{diag}(e_1, \ldots, e_n)$ with both lists as in (a), then $d_i = e_i$ for every i.

The matrix $\mathbf{D}$ is called the **Smith normal form** of $\mathbf{M}$, and $d_1, \ldots, d_n$ its **invariant divisors**.

Since $0 \mid f$ holds only for $f = 0$, the divisibility chain forces the zero entries, if any, to be the last ones.

Idea. **Step roadmap for existence.** ① Among all matrices reachable from $\mathbf{M}$ by operations, pick one whose smallest nonzero entry degree is least, and move that entry to the corner. ② That corner entry f must divide every other entry of its row and column: otherwise divide with remainder and produce an entry of smaller degree, contradicting the choice in ①. ③ Clear the first row and column with f; the same remainder trick, after adding a row to the first one, shows f divides every entry of what is left. ④ Induct on the size. The divisibility $d_1 \mid d_2$ survives the induction because operations replace entries by $F[x]$-combinations of entries.

For uniqueness, find a quantity that can be read off both $\mathbf{M}$ and $\mathbf{D}$. The greatest common divisor of all $k \times k$ minors is such a quantity: minors of a product are $F[x]$-combinations of minors of the factors (Cauchy–Binet), so the gcd cannot change under equivalence, and for a diagonal matrix with a divisibility chain it is simply $d_1 d_2 \cdots d_k$.

Proof.

(a) If $\mathbf{M} = \mathbf{O}$, it is already of the stated form. So assume $\mathbf{M} \neq \mathbf{O}$.

For a nonzero $\mathbf{N} \in M_n(F[x])$ put

$$\delta(\mathbf{N}) := \min\{\, \deg n_{ij} : n_{ij} \neq 0 \,\} \in \mathbb{N},$$

the least degree of a nonzero entry. Let $\mathcal{S}$ be the set of all nonzero matrices obtainable from $\mathbf{M}$ by finite sequences of elementary row and column operations over $F[x]$; it contains $\mathbf{M}$. The set $\{\delta(\mathbf{N}) : \mathbf{N} \in \mathcal{S}\}$ is a nonempty subset of $\mathbb{N}$, so it has a least element; **fix $\mathbf{N} \in \mathcal{S}$ with $\delta(\mathbf{N})$ least**.

Step 1. Put a least-degree entry in the corner, monic. Choose (i, j) with $n_{ij} \neq 0$ and $\deg n_{ij} = \delta(\mathbf{N})$. Swapping row i with row 1 and column j with column 1 (operations (E1)), and then scaling the new first row by the inverse of the leading coefficient of the corner entry (operation (E2)), we obtain a matrix in $\mathcal{S}$ whose $(1,1)$-entry f is **monic** of degree $\delta(\mathbf{N})$. These operations permute the entries and rescale some of them, so the new matrix still has least nonzero entry degree $\delta(\mathbf{N})$. Rename it $\mathbf{N}$.

Step 2. f divides every entry of the first row and the first column. Suppose $f \nmid n_{1j}$ for some $j \geq 2$. By Theorem 6.1.1, $n_{1j} = qf + r$ with $r \neq 0$ and $\deg r < \deg f$. Adding $-q$ times column 1 to column j (operation (E3)) produces a matrix $\mathbf{N}' \in \mathcal{S}$ whose $(1,j)$-entry is $r \neq 0$ with $\deg r < \deg f = \delta(\mathbf{N})$. Then $\delta(\mathbf{N}') < \delta(\mathbf{N})$, contradicting the minimality of $\delta(\mathbf{N})$. Hence $f \mid n_{1j}$ for every j, and the same argument with rows in place of columns gives $f \mid n_{i1}$ for every i.

Step 3. Clear the first row and column. For $j \geq 2$ add $-(n_{1j}/f)$ times column 1 to column j; the multipliers are polynomials by Step 2, and after these operations the first row is $(f, 0, \dots, 0)$ while the first column is unchanged. Then for $i \geq 2$ add $-(n_{i1}/f)$ times row 1 to row i; since the first row is now $(f, 0, \dots, 0)$, this clears the first column and changes nothing else. The result is

$$\begin{pmatrix} f & \mathbf{0}^{\top} \\ \mathbf{0} & \mathbf{M}' \end{pmatrix} \in \mathcal{S}, \qquad \mathbf{M}' \in M_{n-1}(F[x]),$$

and its least nonzero entry degree is still $\delta(\mathbf{N})$: it is $\geq \delta(\mathbf{N})$ by minimality, and f is an entry of degree $\delta(\mathbf{N})$.

Step 4. f divides every entry of $\mathbf{M}'$. Suppose $f \nmid (\mathbf{M}')_{kl}$ for some k, l. Add row $k+1$ to row 1. The new first row is $(f, (\mathbf{M}')_{k1}, \dots, (\mathbf{M}')_{k,n-1})$, so the $(1, l+1)$-entry of the resulting matrix $\mathbf{N}'' \in \mathcal{S}$ is $(\mathbf{M}')_{kl}$, which is not divisible by f, while the $(1,1)$-entry is still f, monic of degree $\delta(\mathbf{N})$. By minimality $\delta(\mathbf{N}'') = \delta(\mathbf{N})$, so Step 2 applies to $\mathbf{N}''$ and says that f divides every entry of the first row of $\mathbf{N}''$ — a contradiction.

Step 5. Induct on n. We prove (a) by induction on n. For $n = 1$, scale the single entry to be monic (or leave it 0). For $n \geq 2$, Steps 1–4 give a sequence of operations taking $\mathbf{M}$ to $f \oplus \mathbf{M}'$ with f monic and f dividing every entry of $\mathbf{M}'$. By the induction hypothesis there are operations taking $\mathbf{M}'$ to $\mathbf{D}' = \operatorname{diag}(d_2, \dots, d_n)$ with every d_i monic or zero and $d_2 \mid \cdots \mid d_n$; performing the corresponding operations on rows and columns $2, \dots, n$ of $f \oplus \mathbf{M}'$ takes it to $f \oplus \mathbf{D}' = \operatorname{diag}(f, d_2, \dots, d_n)$.

It remains to see that $f \mid d_2$. By Proposition 10.7.2, $\mathbf{D}' = \mathbf{U}'\mathbf{M}'\mathbf{V}'$ for some $\mathbf{U}', \mathbf{V}' \in M_{n-1}(F[x])$, so every entry of $\mathbf{D}'$ is a sum of products $u'_{ik}(\mathbf{M}')_{kl}v'_{lj}$ (Definition 0.9.9) and is therefore divisible by f, since every $(\mathbf{M}')_{kl}$ is. In particular $f \mid d_2$. Setting $d_1 = f$ finishes the induction.

(b) Call a set $I \subseteq [n]$ with $|I| = k$ a **k-set**. For $\mathbf{M} \in M_n(F[x])$ and $1 \leq k \leq n$ define the k-th **determinantal divisor**

$$\Delta_k(\mathbf{M}) := \text{the monic generator of the ideal } \langle\, \det \mathbf{M}_{I,J} : I, J \text{ } k\text{-sets} \,\rangle \subseteq F[x],$$

and $\Delta_k(\mathbf{M}) := 0$ if all the $k \times k$ minors of $\mathbf{M}$ vanish. This is well defined by Proposition 6.2.1 and Theorem 6.2.2: the ideal generated by the finitely many $k \times k$ minors is $\langle \Delta_k(\mathbf{M}) \rangle$ for exactly one monic $\Delta_k(\mathbf{M})$ unless the ideal is $\{0\}$. Equivalently, $\Delta_k(\mathbf{M})$ is the monic greatest common divisor of the $k \times k$ minors: it divides each of them, and every common divisor of them divides it (Theorem 6.2.3, extended from two polynomials to finitely many by the same ideal argument).

Claim 1. $\mathbf{M} \approx \mathbf{M}'$ implies $\Delta_k(\mathbf{M}) = \Delta_k(\mathbf{M}')$ for every k.

Proof. Let $\mathbf{M}' = \mathbf{UMV}$ with $\mathbf{U}, \mathbf{V}$ invertible over $F[x]$, and let I, J be k-sets. The submatrix of a product on rows I and columns J is the product of the submatrix of the left factor on rows I and all columns with the submatrix of the right factor on all rows and columns J (Definition 0.9.9, read off entry by entry). Applying this twice and then Theorem 7.7.2 over the commutative ring $F[x]$ — legitimate by the remark after that theorem — to the $k \times n$ times $n \times k$ products involved,

$$\det \mathbf{M}'_{I,J} = \sum_S \det \mathbf{U}_{I,S} \, \det(\mathbf{MV})_{S,J} = \sum_S \sum_R \det \mathbf{U}_{I,S} \, \det \mathbf{M}_{S,R} \, \det \mathbf{V}_{R,J},$$

with S and R running over the k-sets. So every $k \times k$ minor of $\mathbf{M}'$ is an $F[x]$-combination of $k \times k$ minors of $\mathbf{M}$, hence lies in $\langle \det \mathbf{M}_{S,R} \rangle$. Therefore $\Delta_k(\mathbf{M})$ divides every $k \times k$ minor of $\mathbf{M}'$, and so $\Delta_k(\mathbf{M}) \mid \Delta_k(\mathbf{M}')$. Exchanging the roles of $\mathbf{M}$ and $\mathbf{M}'$ using $\mathbf{M} = \mathbf{U}^{-1}\mathbf{M}'\mathbf{V}^{-1}$ gives $\Delta_k(\mathbf{M}') \mid \Delta_k(\mathbf{M})$. Two monic polynomials dividing each other are equal (Proposition 6.1.2), and if one is 0 so is the other.

■

Claim 2. If $\mathbf{D} = \operatorname{diag}(d_1, \dots, d_n)$ with $d_1 \mid d_2 \mid \cdots \mid d_n$, then $\Delta_k(\mathbf{D}) = d_1 d_2 \cdots d_k$.

Proof. Let I, J be k-sets. If $I \neq J$, pick $i \in I \setminus J$; row i of $\mathbf{D}$ has its only possibly nonzero entry in column $i \notin J$, so the submatrix $\mathbf{D}_{I,J}$ has a zero row and $\det \mathbf{D}_{I,J} = 0$. If $I = J = \{i_1 < \cdots < i_k\}$, then $\mathbf{D}_{I,I} = \operatorname{diag}(d_{i_1}, \dots, d_{i_k})$ and $\det \mathbf{D}_{I,I} = d_{i_1} \cdots d_{i_k}$ (Theorem 7.4.6, valid over the commutative ring $F[x]$ by the remark after it). Since $i_j \geq j$, the chain gives $d_j \mid d_{i_j}$ for each j, hence

$$d_1 d_2 \cdots d_k \mid d_{i_1} d_{i_2} \cdots d_{i_k}.$$

So $d_1 \cdots d_k$ divides every $k \times k$ minor of $\mathbf{D}$, and it is itself the minor for $I = J = \{1, \dots, k\}$. Hence the ideal generated by the $k \times k$ minors is $\langle d_1 \cdots d_k \rangle$. If $d_1 \cdots d_k \neq 0$ it is monic, being a product of monic polynomials, and equals $\Delta_k(\mathbf{D})$. If $d_1 \cdots d_k = 0$, say $d_m = 0$ with $m \leq k$, then for every k-set I we have $i_k \geq k \geq m$, so $d_m \mid d_{i_k}$ forces $d_{i_k} = 0$ and the minor vanishes; thus $\Delta_k(\mathbf{D}) = 0 = d_1 \cdots d_k$.

■

Now let $\mathbf{M} \approx \operatorname{diag}(d_1, \dots, d_n)$ with the d_i as in (a). By Claims 1 and 2,

$$\Delta_k(\mathbf{M}) = d_1 d_2 \cdots d_k \qquad (k = 1, \dots, n),$$

a quantity depending only on $\mathbf{M}$. Reading this backwards recovers the d_i: first $d_1 = \Delta_1(\mathbf{M})$; and for $k \geq 2$, if $\Delta_{k-1}(\mathbf{M}) \neq 0$ then $d_k = \Delta_k(\mathbf{M}) / \Delta_{k-1}(\mathbf{M})$, while if $\Delta_{k-1}(\mathbf{M}) = 0$ then $d_1 \cdots d_{k-1} = 0$, so some $d_m = 0$ with $m \leq k-1$, and the chain forces $d_k = 0$. Hence any second list $(e_1, \dots, e_n)$ as in (a) satisfies the same recipe and equals $(d_1, \dots, d_n)$. This proves the theorem.

■

The proof is an algorithm, and it is the one to use by hand: hunt for the entry of least degree, divide, and repeat.

Algorithm. **Smith normal form of $\mathbf{M} \in M_n(F[x])$.**

1. If $\mathbf{M} = \mathbf{O}$, stop. Otherwise find a nonzero entry of least degree and move it to position $(1,1)$ by a row and a column swap; scale row 1 so that the entry is monic. Call it f.
2. For each other entry g in row 1 or column 1 with $f \nmid g$, replace g by its remainder on division by f (subtract a multiple of column 1, resp. row 1), move the remainder to the corner, and go back to step 1.
3. When f divides all of row 1 and column 1, clear them. If some entry g of the remaining block is not divisible by f, add its row to row 1 and go back to step 2.
4. Repeat on the remaining $(n-1)\times(n-1)$ block.
5. Check the answer: Δ_k must equal $d_1 \cdots d_k$. In particular $d_1 \cdots d_n$ is $\det \mathbf{M}$ rescaled by a nonzero constant to be monic, and d_1 is the monic gcd of the entries.

Quick check. Two polynomial matrices in $M_2(F[x])$ have determinants x^2 and x^2, and in both the greatest common divisor of the four entries is 1. Are they equivalent over $F[x]$?

Solution. Yes. Here $\Delta_1 = 1$ (the gcd of the 1×1 minors, that is, of the entries) and Δ_2 is the determinant rescaled to be monic, so $\Delta_2 = x^2$ for both. By Theorem 10.7.3 both have Smith normal form $\operatorname{diag}(1, x^2)$, and since $\approx$ is an equivalence relation (Proposition 10.7.2 (a)) they are equivalent to each other. For 2×2 matrices the pair (Δ_1, Δ_2) is therefore a complete invariant.

Example. A Smith Normal Form Computed by Hand

Let

$$\mathbf{A} = \begin{pmatrix} 2 & 1 & -2 \\ 1 & 2 & -2 \\ 1 & 1 & -1 \end{pmatrix} \in M_3(\mathbb{Q}).$$

Find the Smith normal form of $x\mathbf{I} - \mathbf{A}$ over $\mathbb{Q}[x]$, naming every operation, and check the answer with determinantal divisors.

Solution. We have

$$x\mathbf{I} - \mathbf{A} = \begin{pmatrix} x-2 & -1 & 2 \\ -1 & x-2 & 2 \\ -1 & -1 & x+1 \end{pmatrix}.$$

Find the corner. The entries of least degree are the three entries equal to -1. Swap columns 1 and 2, then multiply the first row by -1:

$$\begin{pmatrix} -1 & x-2 & 2 \\ x-2 & -1 & 2 \\ -1 & -1 & x+1 \end{pmatrix} \longrightarrow \begin{pmatrix} 1 & -(x-2) & -2 \\ x-2 & -1 & 2 \\ -1 & -1 & x+1 \end{pmatrix}.$$

The corner is now 1, which divides everything, so no division step is needed.

Clear the first column. Add $-(x-2)$ times row 1 to row 2, and add row 1 to row 3. In row 2 this gives entries $-1+(x-2)^2=(x-1)(x-3)$ and $2+2(x-2)=2(x-1)$; in row 3 it gives $-1-(x-2)=-(x-1)$ and $x+1-2=x-1$:

$$\begin{pmatrix} 1 & -(x-2) & -2 \\ 0 & (x-1)(x-3) & 2(x-1) \\ 0 & -(x-1) & x-1 \end{pmatrix}.$$

Clear the first row. Add $(x-2)$ times column 1 to column 2, and 2 times column 1 to column 3. Since the first column below the corner is zero, nothing else changes:

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & (x-1)(x-3) & 2(x-1) \\ 0 & -(x-1) & x-1 \end{pmatrix} = \operatorname{diag}(1) \oplus \begin{pmatrix} (x-1)(x-3) & 2(x-1) \\ -(x-1) & x-1 \end{pmatrix}.$$

Every entry of the 2×2 block is divisible by 1, so we move on to the block.

The block. Its entries of least degree are the three of degree 1. Swap rows 2 and 3 and scale the new row 2 by -1, putting the monic $x-1$ in the corner of the block:

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & x-1 & -(x-1) \\ 0 & (x-1)(x-3) & 2(x-1) \end{pmatrix}.$$

Now $x-1$ divides every entry of its row and column. Add column 2 to column 3, which clears the block's first row and turns the $(3,3)$-entry into $2(x-1)+(x-1)(x-3)=(x-1)(x-1)$:

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & x-1 & 0 \\ 0 & (x-1)(x-3) & (x-1)^2 \end{pmatrix}.$$

Add $-(x-3)$ times row 2 to row 3 to clear the block's first column:

$$\mathbf{D} = \operatorname{diag}\big(1,\ x-1,\ (x-1)^2\big),$$

and $1 \mid x-1 \mid (x-1)^2$, so by the uniqueness in Theorem 10.7.3 this is **the** Smith normal form.

The check. The matrix $x\mathbf{I}-\mathbf{A}$ has an entry -1, so $\Delta_1=1$. Its determinant is $p_{\mathbf{A}}=(x-1)^3$ — expanding along the first row,

$$\begin{aligned} &(x-2)\big[(x-2)(x+1)+2\big] + \big[-(x+1)+2\big] + 2\big[1+(x-2)\big] \\ &\qquad = (x-2)(x^2-x) + (1-x) + 2(x-1), \end{aligned}$$

which is $(x-2)x(x-1)+(x-1)=(x-1)(x^2-2x+1)=(x-1)^3$ — so $\Delta_3=(x-1)^3$. For Δ_2, the minor on rows $\{2,3\}$ and columns $\{1,2\}$ is $\det\begin{pmatrix} -1 & x-2 \\ -1 & -1 \end{pmatrix} = 1+(x-2)=x-1$,

so $\Delta_2 \mid x-1$; and every 2×2 minor is divisible by $x-1$, as Claim 2 predicts from $\mathbf{D}$. Hence $\Delta_2 = x-1$, and

$$(d_1,d_2,d_3) = \left(\Delta_1,\ \tfrac{\Delta_2}{\Delta_1},\ \tfrac{\Delta_3}{\Delta_2}\right) = (1,\ x-1,\ (x-1)^2),$$

in agreement. As a sanity check on $\mathbf{A}$ itself: $\mathbf{A}-\mathbf{I}$ has rank 1, so $g_{\mathbf{A}}(1)=2$ and the Jordan form of $\mathbf{A}$ is $\mathbf{J}_2(1)\oplus\mathbf{J}_1(1)$ (Corollary 10.3.2).

10.7.4 Similarity read off from a polynomial matrix

We can now state the theorem that makes all of this useful. It converts a question about $\mathbf{A}$ and $\mathbf{B}$ — is there an invertible $\mathbf{P}\in M_n(F)$ with $\mathbf{B}=\mathbf{P}^{-1}\mathbf{A}\mathbf{P}$? — into a question about the two polynomial matrices $x\mathbf{I}-\mathbf{A}$ and $x\mathbf{I}-\mathbf{B}$, which the Smith algorithm answers by computation.

One direction is immediate. The other needs a tool: polynomial matrices can be divided by $x\mathbf{I}-\mathbf{A}$, with a **constant** remainder.

Lemma 10.7.4. Division of a Polynomial Matrix by $x\mathbf{I}-\mathbf{A}$

Let $\mathbf{A}\in M_n(F)$ and $\mathbf{M}\in M_n(F[x])$. Then there are $\mathbf{Q}\in M_n(F[x])$ and $\mathbf{R}\in M_n(F)$ with

$$\mathbf{M} = (x\mathbf{I}-\mathbf{A})\,\mathbf{Q} + \mathbf{R}.$$

Idea. Write $\mathbf{M}$ as a polynomial $\sum_k \mathbf{M}_k x^k$ with **constant matrix** coefficients. The scalar identity $x^k - a^k = (x-a)(x^{k-1}+\cdots+a^{k-1})$ has a matrix version $x^k\mathbf{I}-\mathbf{A}^k = (x\mathbf{I}-\mathbf{A})\mathbf{S}_k$, in which the second factor is a polynomial in x and $\mathbf{A}$ and so commutes with $x\mathbf{I}-\mathbf{A}$. Replacing each $x^k\mathbf{I}$ by $(x^k\mathbf{I}-\mathbf{A}^k)+\mathbf{A}^k$ puts a factor $x\mathbf{I}-\mathbf{A}$ in front of everything except a constant matrix.

Proof. Collecting the coefficients of each power of x in the entries of $\mathbf{M}$, write $\mathbf{M}=\sum_{k=0}^{m} x^k\mathbf{M}_k$ with $\mathbf{M}_0,\dots,\mathbf{M}_m\in M_n(F)$; note that x^k is a scalar polynomial, so $x^k\mathbf{M}_k = \mathbf{M}_k x^k$. For $k\ge 1$ put

$$\mathbf{S}_k := \sum_{j=0}^{k-1} x^{\,k-1-j}\mathbf{A}^{\,j} \in M_n(F[x]), \qquad \mathbf{S}_0 := \mathbf{O}.$$

Expanding the product and telescoping,

$$(x\mathbf{I}-\mathbf{A})\mathbf{S}_k = \sum_{j=0}^{k-1} x^{\,k-j}\mathbf{A}^{\,j} - \sum_{j=0}^{k-1} x^{\,k-1-j}\mathbf{A}^{\,j+1} = x^k\mathbf{I}-\mathbf{A}^k$$

for every $k \geq 0$. Therefore

$$\begin{aligned}\mathbf{M} &= \sum_{k=0}^{m} (x^k\mathbf{I})\mathbf{M}_k = \sum_{k=0}^{m} (x^k\mathbf{I} - \mathbf{A}^k)\mathbf{M}_k + \sum_{k=0}^{m} \mathbf{A}^k\mathbf{M}_k \\ &= (x\mathbf{I} - \mathbf{A}) \underbrace{\sum_{k=0}^{m} \mathbf{S}_k\mathbf{M}_k}_{=\mathbf{Q}} + \underbrace{\sum_{k=0}^{m} \mathbf{A}^k\mathbf{M}_k}_{=\mathbf{R}},\end{aligned}$$

with $\mathbf{Q} \in M_n(F[x])$ and $\mathbf{R} \in M_n(F)$, as claimed.

■

Theorem 10.7.5. Similarity as Equivalence of $x\mathbf{I} - \mathbf{A}$ and $x\mathbf{I} - \mathbf{B}$

Let F be a field, $n \geq 1$, and $\mathbf{A}, \mathbf{B} \in M_n(F)$. Then

$$\mathbf{A} \sim \mathbf{B} \quad \text{over } F \qquad \Longleftrightarrow \qquad x\mathbf{I} - \mathbf{A} \approx x\mathbf{I} - \mathbf{B} \quad \text{over } F[x].$$

Idea. ($\Rightarrow$) is a one-line computation: a constant invertible $\mathbf{P}$ is also an invertible polynomial matrix.

($\Leftarrow$) is the substance, and the whole difficulty is that the matrices $\mathbf{U}, \mathbf{V}$ witnessing the equivalence have polynomial entries, while we must produce a **constant** $\mathbf{P}$. The trick is to divide $\mathbf{U}$ by $x\mathbf{I} - \mathbf{B}$ and keep only the constant remainder $\mathbf{P}$. Comparing degrees in the resulting identity shows that $\mathbf{P}$ alone already conjugates: $\mathbf{P}(x\mathbf{I} - \mathbf{A}) = (x\mathbf{I} - \mathbf{B})\mathbf{P}$, that is, $\mathbf{PA} = \mathbf{BP}$. Running the same argument on the inverted relation produces a second constant $\mathbf{P}'$ with $\mathbf{P}'\mathbf{B} = \mathbf{AP}'$, and multiplying the two expressions for $\mathbf{V}^{-1}$ and $\mathbf{V}$ shows $\mathbf{PP}' = \mathbf{I}$, so $\mathbf{P}$ is invertible.

Proof. ($\Rightarrow$) Suppose $\mathbf{B} = \mathbf{P}^{-1}\mathbf{AP}$ with $\mathbf{P} \in M_n(F)$ invertible. Then, inside $M_n(F[x])$,

$$x\mathbf{I} - \mathbf{B} = \mathbf{P}^{-1}(x\mathbf{I})\mathbf{P} - \mathbf{P}^{-1}\mathbf{AP} = \mathbf{P}^{-1}(x\mathbf{I} - \mathbf{A})\mathbf{P},$$

and $\mathbf{P}^{-1}, \mathbf{P} \in M_n(F) \subseteq M_n(F[x])$ have nonzero constant determinants, hence are invertible over $F[x]$ (Proposition 10.7.1). So $x\mathbf{I} - \mathbf{A} \approx x\mathbf{I} - \mathbf{B}$.

($\Leftarrow$) Suppose

$$x\mathbf{I} - \mathbf{B} = \mathbf{U}(x\mathbf{I} - \mathbf{A})\mathbf{V} \tag{$*$}$$

with $\mathbf{U}, \mathbf{V} \in M_n(F[x])$ invertible over $F[x]$.

Step 1. Divide $\mathbf{U}$ by $x\mathbf{I} - \mathbf{B}$. By Lemma 10.7.4 there are $\mathbf{Q} \in M_n(F[x])$ and $\mathbf{P} \in M_n(F)$ with $\mathbf{U} = (x\mathbf{I} - \mathbf{B})\mathbf{Q} + \mathbf{P}$. Substituting into ($*$),

$$x\mathbf{I} - \mathbf{B} = (x\mathbf{I} - \mathbf{B})\,\mathbf{Q}(x\mathbf{I} - \mathbf{A})\mathbf{V} + \mathbf{P}(x\mathbf{I} - \mathbf{A})\mathbf{V},$$

so $\mathbf{P}(x\mathbf{I} - \mathbf{A})\mathbf{V} = (x\mathbf{I} - \mathbf{B})\big[\mathbf{I} - \mathbf{Q}(x\mathbf{I} - \mathbf{A})\mathbf{V}\big]$. Multiplying on the right by $\mathbf{V}^{-1}$ and putting

$$\mathbf{S} := \mathbf{V}^{-1} - \mathbf{Q}(x\mathbf{I} - \mathbf{A}) \in M_n(F[x]),$$

we get

$$\mathbf{P}(x\mathbf{I}-\mathbf{A}) = (x\mathbf{I}-\mathbf{B})\,\mathbf{S}. \tag{1}$$

Step 2. $\mathbf{S}$ is a constant matrix, and $\mathbf{PA} = \mathbf{BP}$. Suppose $\mathbf{S} \neq \mathbf{O}$ and write $\mathbf{S} = \sum_{k=0}^{d} x^k \mathbf{S}_k$ with $\mathbf{S}_k \in M_n(F)$ and $\mathbf{S}_d \neq \mathbf{O}$. The coefficient of x^{d+1} in $(x\mathbf{I}-\mathbf{B})\mathbf{S}$ is $\mathbf{S}_d \neq \mathbf{O}$, while the left-hand side of (1) is $x\mathbf{P} - \mathbf{PA}$, whose coefficients vanish beyond x^1. Hence $d+1 \leq 1$, that is $d = 0$. So in every case $\mathbf{S} \in M_n(F)$. Comparing the coefficients of x^1 in (1) gives $\mathbf{P} = \mathbf{S}$, and comparing the constant terms gives $-\mathbf{PA} = -\mathbf{BP}$. Therefore

$$\mathbf{PA} = \mathbf{BP}, \qquad \text{equivalently} \qquad \mathbf{P}(x\mathbf{I}-\mathbf{A}) = (x\mathbf{I}-\mathbf{B})\mathbf{P}, \tag{2}$$

the second form because $\mathbf{P}(x\mathbf{I}-\mathbf{A}) = x\mathbf{P} - \mathbf{PA} = x\mathbf{P} - \mathbf{BP} = (x\mathbf{I}-\mathbf{B})\mathbf{P}$. Moreover $\mathbf{S} = \mathbf{P}$ turns the definition of $\mathbf{S}$ into

$$\mathbf{V}^{-1} = \mathbf{P} + \mathbf{Q}(x\mathbf{I}-\mathbf{A}). \tag{3}$$

Step 3. The same argument, inverted. From $(*)$, $x\mathbf{I}-\mathbf{A} = \mathbf{U}^{-1}(x\mathbf{I}-\mathbf{B})\mathbf{V}^{-1}$, which is $(*)$ with $\mathbf{A}, \mathbf{B}$ interchanged and $\mathbf{U}, \mathbf{V}$ replaced by the invertible polynomial matrices $\mathbf{U}^{-1}, \mathbf{V}^{-1}$. So Steps 1 and 2 apply and give $\mathbf{Q}' \in M_n(F[x])$ and $\mathbf{P}' \in M_n(F)$ with

$$\mathbf{P}'(x\mathbf{I}-\mathbf{B}) = (x\mathbf{I}-\mathbf{A})\mathbf{P}' \qquad \text{and} \qquad \mathbf{V} = \mathbf{P}' + \mathbf{Q}'(x\mathbf{I}-\mathbf{B}). \tag{4}$$

Step 4. $\mathbf{P}$ is invertible. Multiplying (3) by (4),

$$\begin{aligned} \mathbf{I} = \mathbf{V}^{-1}\mathbf{V} &= \big[\mathbf{P} + \mathbf{Q}(x\mathbf{I}-\mathbf{A})\big]\big[\mathbf{P}' + \mathbf{Q}'(x\mathbf{I}-\mathbf{B})\big] \\ &= \mathbf{PP}' + \mathbf{PQ}'(x\mathbf{I}-\mathbf{B}) + \mathbf{Q}(x\mathbf{I}-\mathbf{A})\mathbf{P}' \\ &\quad + \mathbf{Q}(x\mathbf{I}-\mathbf{A})\mathbf{Q}'(x\mathbf{I}-\mathbf{B}). \end{aligned}$$

By the first identity in (4), $(x\mathbf{I}-\mathbf{A})\mathbf{P}' = \mathbf{P}'(x\mathbf{I}-\mathbf{B})$, so the third summand is $\mathbf{QP}'(x\mathbf{I}-\mathbf{B})$ and

$$\mathbf{I} - \mathbf{PP}' = \mathbf{Y}(x\mathbf{I}-\mathbf{B}), \qquad \mathbf{Y} \coloneqq \mathbf{PQ}' + \mathbf{QP}' + \mathbf{Q}(x\mathbf{I}-\mathbf{A})\mathbf{Q}' \in M_n(F[x]).$$

If $\mathbf{Y} \neq \mathbf{O}$, write $\mathbf{Y} = \sum_{k=0}^{e} x^k \mathbf{Y}_k$ with $\mathbf{Y}_e \neq \mathbf{O}$; the coefficient of x^{e+1} on the right is $\mathbf{Y}_e \neq \mathbf{O}$, while the left-hand side is constant. Hence $\mathbf{Y} = \mathbf{O}$ and $\mathbf{PP}' = \mathbf{I}$. By Theorem 3.4.5, $\mathbf{P}$ is invertible with $\mathbf{P}^{-1} = \mathbf{P}'$.

Step 5. Conclude. By (2), $\mathbf{PA} = \mathbf{BP}$, so $\mathbf{B} = \mathbf{PAP}^{-1}$ and $\mathbf{A} \sim \mathbf{B}$. This proves the theorem.

■

Two by-products are worth recording. Whenever $x\mathbf{I}-\mathbf{B} = \mathbf{U}(x\mathbf{I}-\mathbf{A})\mathbf{V}$ with $\mathbf{U}, \mathbf{V}$ invertible over $F[x]$, taking determinants gives $p_{\mathbf{B}} = (\det \mathbf{U})(\det \mathbf{V})\, p_{\mathbf{A}}$ with $\det \mathbf{U}, \det \mathbf{V}$ nonzero constants (Proposition 10.7.1); both $p_{\mathbf{A}}$ and $p_{\mathbf{B}}$ are monic of degree n, so

$$p_{\mathbf{A}} = p_{\mathbf{B}} \qquad \text{and} \qquad (\det \mathbf{U})(\det \mathbf{V}) = 1,$$

which recovers the invariance of the characteristic polynomial. And, by Theorem 10.7.3, the test is now decidable by a finite computation in $F[x]$: reduce $x\mathbf{I}-\mathbf{A}$ and $x\mathbf{I}-\mathbf{B}$ to Smith normal form and compare.

Corollary 10.7.6. The Smith Form of $x\mathbf{I}-\mathbf{A}$ Carries the Invariant Factors

Let $\mathbf{A} \in M_n(F)$ with invariant factors $d_1 \mid d_2 \mid \cdots \mid d_r$ as in Theorem 10.6.5. Then the Smith normal form of $x\mathbf{I} - \mathbf{A}$ is

$$\operatorname{diag}\big(\underbrace{1, \ldots, 1}_{n-r},\, d_1,\, d_2,\, \ldots,\, d_r\big).$$

In particular the invariant factors of $\mathbf{A}$ are exactly the nonconstant diagonal entries of the Smith normal form of $x\mathbf{I} - \mathbf{A}$, read from the top down, and two matrices in $M_n(F)$ are similar if and only if $x\mathbf{I} - \mathbf{A}$ and $x\mathbf{I} - \mathbf{B}$ have the same Smith normal form.

Proof. By Theorem 10.6.5, $\mathbf{A} \sim \mathbf{C}(d_1) \oplus \cdots \oplus \mathbf{C}(d_r)$ with d_i monic, $\deg d_i = k_i \geq 1$, and $\sum k_i = n$. By the ($\Rightarrow$) direction of Theorem 10.7.5 it is enough to find the Smith normal form of

$$x\mathbf{I}_n - \big(\mathbf{C}(d_1) \oplus \cdots \oplus \mathbf{C}(d_r)\big) = \big(x\mathbf{I}_{k_1} - \mathbf{C}(d_1)\big) \oplus \cdots \oplus \big(x\mathbf{I}_{k_r} - \mathbf{C}(d_r)\big),$$

the equality by Theorem 8.1.4 (a).

One block. Let d be monic of degree $k \geq 1$. By Definition 10.5.3, the only nonzero entries of $x\mathbf{I}_k - \mathbf{C}(d)$ are x on the diagonal (with $x + a_{k-1}$ in the corner), -1 in each position $(i+1, i)$, and the entries of the last column. Delete row 1 and column k. In the remaining $(k-1) \times (k-1)$ matrix, the entry in position (p, q) is the $(p+1, q)$-entry of $x\mathbf{I}_k - \mathbf{C}(d)$: it is -1 when $q = p$, it is x when $q = p+1$, and it is 0 when $q < p$. So that matrix is upper triangular with every diagonal entry -1, and its determinant is $(-1)^{k-1} \neq 0$ by Theorem 7.4.6, again over $F[x]$. Hence $\Delta_{k-1}(x\mathbf{I}_k - \mathbf{C}(d)) = 1$. By Claim 2 in the proof of Theorem 10.7.3, the determinantal divisors of any matrix are the partial products $d_1 \cdots d_j$ of its Smith diagonal, so $\Delta_j \mid \Delta_{k-1} = 1$ and therefore $\Delta_j = 1$ for every $j \leq k-1$. Also $\Delta_k = \det(x\mathbf{I}_k - \mathbf{C}(d)) = p_{\mathbf{C}(d)} = d$ by Theorem 10.5.3. By the uniqueness part of Theorem 10.7.3, the Smith normal form of $x\mathbf{I}_k - \mathbf{C}(d)$ is $\operatorname{diag}(1, \ldots, 1, d)$. (For $k = 1$ read this as $\operatorname{diag}(d)$.)

Assembling. If $\mathbf{M}_i \approx \mathbf{M}_i'$ with witnesses $\mathbf{U}_i, \mathbf{V}_i$, then $\bigoplus_i \mathbf{M}_i \approx \bigoplus_i \mathbf{M}_i'$ with witnesses $\bigoplus_i \mathbf{U}_i$ and $\bigoplus_i \mathbf{V}_i$, which are invertible with inverses $\bigoplus_i \mathbf{U}_i^{-1}$, $\bigoplus_i \mathbf{V}_i^{-1}$ (Theorem 8.1.4 (a), (c)). So $x\mathbf{I}_n - \mathbf{A}$ is equivalent to the direct sum of the matrices $\operatorname{diag}(1, \ldots, 1, d_i)$, a diagonal matrix whose diagonal entries are $n - r$ ones and the polynomials $d_1, \ldots, d_r$. Simultaneous row and column swaps — each a pair of elementary operations — reorder these entries at will, so $x\mathbf{I}_n - \mathbf{A}$ is equivalent to

$$\operatorname{diag}(1, \ldots, 1, d_1, d_2, \ldots, d_r),$$

and this is already in the shape required by Theorem 10.7.3 (a), because $1 \mid d_1 \mid d_2 \mid \cdots \mid d_r$ and all the entries are monic. By uniqueness it *is* the Smith normal form of $x\mathbf{I} - \mathbf{A}$. The final claim combines this with Theorem 10.7.5 and the uniqueness in Theorem 10.7.3.

■

So the determinantal divisors of $x\mathbf{I} - \mathbf{A}$ compute the invariant factors of $\mathbf{A}$ without any factorization of polynomials: Δ_k is a gcd, obtained from the Euclidean algorithm alone. Moreover $\Delta_n(x\mathbf{I} - \mathbf{A}) = p_{\mathbf{A}}$ and $\Delta_n / \Delta_{n-1} = d_r = m_{\mathbf{A}}$.

10.7.5 Every matrix is similar to its transpose

Chapter 4 asked whether $\mathbf{A}$ and $\mathbf{A}^{\top}$ are always similar and postponed the answer; Exercise (C1: A matrix is similar to its transpose) settled it whenever $p_{\mathbf{A}}$ splits. The Smith form removes that hypothesis, because transposing a matrix does not change its minors.

Corollary 10.7.7. A Matrix Is Similar to Its Transpose

Let F be any field and $\mathbf{A} \in M_n(F)$. Then $\mathbf{A} \sim \mathbf{A}^\top$.

Proof. Let $\mathbf{M} = x\mathbf{I} - \mathbf{A}$, so that $\mathbf{M}^\top = x\mathbf{I} - \mathbf{A}^\top$. For k-sets I, J we have $(\mathbf{M}^\top)_{I,J} = (\mathbf{M}_{J,I})^\top$ by Definition 0.9.10, so $\det(\mathbf{M}^\top)_{I,J} = \det \mathbf{M}_{J,I}$ by Theorem 7.4.5, valid over the commutative ring $F[x]$. Hence the $k \times k$ minors of $\mathbf{M}^\top$ are exactly the $k \times k$ minors of $\mathbf{M}$, listed in a different order, and

$$\Delta_k(x\mathbf{I} - \mathbf{A}^\top) = \Delta_k(x\mathbf{I} - \mathbf{A}) \qquad (k = 1, \dots, n).$$

By Theorem 10.7.3, each of $x\mathbf{I} - \mathbf{A}$ and $x\mathbf{I} - \mathbf{A}^\top$ is equivalent over $F[x]$ to the diagonal matrix built from its own determinantal divisors, and these two diagonal matrices coincide. Since $\approx$ is an equivalence relation (Proposition 10.7.2 (a)), $x\mathbf{I} - \mathbf{A} \approx x\mathbf{I} - \mathbf{A}^\top$. By Theorem 10.7.5, $\mathbf{A} \sim \mathbf{A}^\top$.

■

Two sentences of consequence. An operator and its dual have the same invariants: by Theorem 5.4.2 the matrix of T' in the dual basis is $([T]_{\mathcal{B}}^{\mathcal{B}})^\top$, so T and T' share their characteristic polynomial, their minimal polynomial, and their invariant factors. And a matrix and its transpose always have the same Jordan form when one exists, which says that the *rows* of $\mathbf{A} - \lambda\mathbf{I}$ are exactly as informative as its columns.

Warning. **Similar is not the same as "equivalent over $F[x]$" for the matrices themselves.** The theorem compares $x\mathbf{I} - \mathbf{A}$ with $x\mathbf{I} - \mathbf{B}$, not $\mathbf{A}$ with $\mathbf{B}$. As polynomial matrices, any two invertible $\mathbf{A}, \mathbf{B} \in M_n(F)$ are equivalent over $F[x]$ — take $\mathbf{U} = \mathbf{A}^{-1}$, $\mathbf{V} = \mathbf{B}$ — whether or not they are similar. The whole content of Theorem 10.7.5 is that inserting the variable x on the diagonal remembers the conjugation.

10.7.6 Exercises

10.7.6.1 A. Check your understanding

Exercise. A1

(a) List the three elementary operations over $F[x]$, and say which one differs from its counterpart over a field and why.

(b) State a criterion for a matrix $\mathbf{U} \in M_n(F[x])$ to be invertible in $M_n(F[x])$.

(c) State the Smith Normal Form theorem, including the uniqueness clause.

(d) Define the k-th determinantal divisor of a polynomial matrix, and say why it is unchanged by equivalence.

(e) True or false: if $\det(x\mathbf{I} - \mathbf{A}) = \det(x\mathbf{I} - \mathbf{B})$, then $x\mathbf{I} - \mathbf{A} \approx x\mathbf{I} - \mathbf{B}$. Justify your answer.

(f) How do you read $m_{\mathbf{A}}$ off the Smith normal form of $x\mathbf{I} - \mathbf{A}$?

Solution.

(a) Swap two rows; multiply a row by a nonzero **scalar** of F; add a **polynomial** multiple of one row to a different row (Definition 10.7.1). The scaling operation is restricted, because only the nonzero constants are invertible in $F[x]$; multiplying a row by x is not reversible.

(b) $\mathbf{U}$ is invertible over $F[x]$ if and only if $\det \mathbf{U}$ is a nonzero constant (Proposition 10.7.1).

(c) Every $\mathbf{M} \in M_n(F[x])$ is equivalent over $F[x]$, by elementary operations, to a diagonal $\operatorname{diag}(d_1, \dots, d_n)$ with each d_i monic or zero and $d_i \mid d_{i+1}$; this diagonal is unique (Theorem 10.7.3).

(d) $\Delta_k(\mathbf{M})$ is the monic generator of the ideal generated by all $k \times k$ minors of $\mathbf{M}$, equivalently their monic gcd. It is unchanged by equivalence because every $k \times k$ minor of $\mathbf{UMV}$ is an $F[x]$-combination of $k \times k$ minors of $\mathbf{M}$, by the Cauchy–Binet formula (Theorem 7.7.2; Claim 1 in the proof of Theorem 10.7.3).

(e) False. Equality of determinants says only $\Delta_n(x\mathbf{I} - \mathbf{A}) = \Delta_n(x\mathbf{I} - \mathbf{B})$; the lower determinantal divisors can differ. For $\mathbf{A} = \mathbf{J}_2(0) \oplus \mathbf{J}_2(0)$ and $\mathbf{B} = \mathbf{J}_3(0) \oplus \mathbf{J}_1(0)$ in $M_4(F)$, both determinants are x^4, but the Smith normal forms are $\operatorname{diag}(1, 1, x^2, x^2)$ and $\operatorname{diag}(1, 1, x, x^3)$. (Equivalently, the matrices are not similar, by Corollary 10.3.3.)

(f) $m_{\mathbf{A}}$ is the **last** diagonal entry of the Smith normal form of $x\mathbf{I} - \mathbf{A}$, that is $\Delta_n/\Delta_{n-1} = p_{\mathbf{A}}/\Delta_{n-1}$ (Corollary 10.7.6, since the last diagonal entry is $d_r = m_{\mathbf{A}}$ in the notation there).

10.7.6.2 B. Practice

Exercise. B1: Smith normal forms

Find the Smith normal form over $\mathbb{Q}[x]$ of each of the following, naming the operations you use.

(a) $\mathbf{M}_1 = \begin{pmatrix} x & x+1 \\ x-1 & x \end{pmatrix}$.

(b) $\mathbf{M}_2 = \begin{pmatrix} x^2-1 & 0 \\ 0 & x^2-x \end{pmatrix}$.

(c) $\mathbf{M}_3 = \begin{pmatrix} x & 0 & 1 \\ -1 & x & 0 \\ 0 & -1 & x \end{pmatrix}$.

Solution.

(a) $\det \mathbf{M}_1 = x^2 - (x+1)(x-1) = 1$, so $\mathbf{M}_1$ is invertible over $\mathbb{Q}[x]$ (Proposition 10.7.1) and $\Delta_1 = \Delta_2 = 1$. The Smith normal form is $\operatorname{diag}(1, 1) = \mathbf{I}_2$. (Concretely: subtract column 1 from column 2 to get $\begin{pmatrix} x & 1 \\ x-1 & 1 \end{pmatrix}$, then subtract row 1 from row 2, swap columns, and clear.)

(b) Here $\Delta_1 = \gcd(x^2-1, x^2-x) = \gcd\big((x-1)(x+1), x(x-1)\big) = x-1$ and $\Delta_2 = \det \mathbf{M}_2 = x(x-1)^2(x+1)$. Reading the determinantal divisors backwards as in the

proof of Theorem 10.7.3, the Smith normal form is

$$\operatorname{diag}\left(x-1,\ \tfrac{x(x-1)^2(x+1)}{x-1}\right) = \operatorname{diag}\left(x-1,\ x(x-1)(x+1)\right),$$

and indeed $x-1$ divides $x(x-1)(x+1)$. By operations: add column 2 to column 1, giving first column (x^2-1, x^2-x); subtract row 1 from row 2 to get $(x^2-1, 1-x)$ in column 1; the entry $1-x$ has least degree, so swap it into the corner, scale by -1, and continue.

(c) The 2×2 minor on rows $\{2,3\}$ and columns $\{1,2\}$ is $\det\begin{pmatrix}-1 & x\\ 0 & -1\end{pmatrix} = 1$, so $\Delta_2 = 1$, and then $\Delta_1 = 1$ as well (it divides Δ_2). Expanding $\det \mathbf{M}_3$ along the first row, $x(x^2-0) + 1\cdot(1-0) = x^3+1$, so $\Delta_3 = x^3+1$ and the Smith normal form is $\operatorname{diag}(1,1,x^3+1)$. Note $\mathbf{M}_3 = x\mathbf{I} - \mathbf{C}$ with $\mathbf{C} = \begin{pmatrix}0 & 0 & -1\\ 1 & 0 & 0\\ 0 & 1 & 0\end{pmatrix} = \mathbf{C}(x^3+1)$, and Corollary 10.7.6 predicts exactly this, since a companion matrix has the single invariant factor x^3+1.

Exercise. B2: Invariant factors from determinantal divisors

For each matrix $\mathbf{A}$, compute $\Delta_1, \dots, \Delta_n$ of $x\mathbf{I} - \mathbf{A}$, write down the Smith normal form, and read off the invariant factors, $m_{\mathbf{A}}$, and the Jordan form.

(a) $\mathbf{A} = \mathbf{J}_2(0) \oplus \mathbf{J}_2(0) \in M_4(\mathbb{Q})$.

(b) $\mathbf{A} = \mathbf{J}_3(0) \oplus \mathbf{J}_1(0) \in M_4(\mathbb{Q})$.

(c) $\mathbf{A} = \begin{pmatrix}3 & 1\\ 0 & 3\end{pmatrix} \in M_2(\mathbb{Q})$.

Solution. One remark makes the minors easy. If $\mathbf{M} = \mathbf{M}_1 \oplus \mathbf{M}_2$ with square blocks, and I, J are k-sets, split $I = I_1 \cup I_2$ and $J = J_1 \cup J_2$ according to the blocks. If $|I_1| \neq |J_1|$, then $\mathbf{M}_{I,J}$ has a zero row or a zero column and $\det \mathbf{M}_{I,J} = 0$; if $|I_1| = |J_1|$, then $\mathbf{M}_{I,J} = (\mathbf{M}_1)_{I_1,J_1} \oplus (\mathbf{M}_2)_{I_2,J_2}$ and the determinant is the product of the two, by Theorem 7.8.4, whose proof uses only sums and products of entries and so applies over $F[x]$ (the remark after that theorem). So **the minors of a block diagonal matrix are the products of matching-size minors of the blocks.** Here $x\mathbf{I} - \mathbf{A}$ is the direct sum of the blocks $x\mathbf{I}_k - \mathbf{J}_k(0)$, each upper triangular with x on the diagonal and -1 on the superdiagonal.

(a) $x\mathbf{I} - \mathbf{A} = \mathbf{N} \oplus \mathbf{N}$ with $\mathbf{N} = \begin{pmatrix}x & -1\\ 0 & x\end{pmatrix}$, whose minors are the entries $x, -1, 0$ for $k=1$ and x^2 for $k=2$. By the remark, the minors of $\mathbf{N}\oplus\mathbf{N}$ are $\{x,-1,0\}$ for $k=1$; the products of two 1×1 minors together with x^2 for $k=2$; x^2 times a 1×1 minor for $k=3$; and x^4 for $k=4$. Hence

$$\Delta_1 = 1, \qquad \Delta_2 = 1 \ \ (\text{from } (-1)(-1)=1), \qquad \Delta_3 = x^2, \qquad \Delta_4 = x^4,$$

so by Theorem 10.7.3 the Smith normal form is $\operatorname{diag}(1,1,x^2,x^2)$, the invariant factors are $x^2 \mid x^2$, $m_{\mathbf{A}} = x^2$ is the last of them, and the Jordan form is $\mathbf{J}_2(0)\oplus\mathbf{J}_2(0)$.

(b) Now $x\mathbf{I}-\mathbf{A}=\mathbf{N}_3\oplus(x)$ with $\mathbf{N}_3=\begin{pmatrix}x&-1&0\\0&x&-1\\0&0&x\end{pmatrix}$. The minors of $\mathbf{N}_3$ are $\{x,-1,0\}$ for $k=1$; $\{x^2,-x,1,0\}$ for $k=2$, the value 1 coming from rows $\{1,2\}$ and columns $\{2,3\}$; and x^3 for $k=3$. By the remark, the $k\times k$ minors of $x\mathbf{I}-\mathbf{A}$ are those of $\mathbf{N}_3$ together with x times its $(k-1)\times(k-1)$ minors, so

$$\begin{aligned}\Delta_1&=1,\\ \Delta_2&=\gcd(x^2,-x,1,x\cdot x,-x)=1,\\ \Delta_3&=\gcd(x^3,\ x\cdot x^2,\ -x\cdot x,\ x\cdot 1)=x,\\ \Delta_4&=x^4.\end{aligned}$$

The Smith normal form is $\operatorname{diag}(1,1,x,x^3)$: the invariant factors are $x\mid x^3$, $m_{\mathbf{A}}=x^3$ is the last of them, and the Jordan form is $\mathbf{J}_3(0)\oplus\mathbf{J}_1(0)$. Compared with (a) the determinants agree but Δ_3 does not, which is exactly why the two matrices are not similar.

(c) $x\mathbf{I}-\mathbf{A}=\begin{pmatrix}x-3&-1\\0&x-3\end{pmatrix}$. Then $\Delta_1=1$ and $\Delta_2=(x-3)^2$, so the Smith normal form is $\operatorname{diag}(1,(x-3)^2)$, the only invariant factor is $(x-3)^2=m_{\mathbf{A}}=p_{\mathbf{A}}$, and the Jordan form is $\mathbf{J}_2(3)$. Consistently, $\mathbf{A}$ is a companion matrix up to similarity, since $m_{\mathbf{A}}=p_{\mathbf{A}}$ (Theorem 10.5.6).

Exercise. B3: Deciding similarity

Determine whether each pair is similar over the stated field. Justify your answer using Theorem 10.7.5.

(a) $\mathbf{A}=\begin{pmatrix}0&-1\\1&0\end{pmatrix}$ and $\mathbf{B}=\begin{pmatrix}0&1\\-1&0\end{pmatrix}$ over $\mathbb{Q}$.

(b) $\mathbf{A}=\begin{pmatrix}2&1&-2\\1&2&-2\\1&1&-1\end{pmatrix}$ and $\mathbf{B}=\mathbf{J}_2(1)\oplus\mathbf{J}_1(1)$ over $\mathbb{Q}$.

(c) $\mathbf{A}=\begin{pmatrix}0&-2\\1&0\end{pmatrix}$ and $\mathbf{B}=\begin{pmatrix}0&-1\\2&0\end{pmatrix}$ over $\mathbb{Q}$.

Solution.

(a) Similar. Here $\mathbf{B}=\mathbf{A}^{\top}$, so Corollary 10.7.7 applies. (Explicitly, $\Delta_1=1$ and $\Delta_2=x^2+1$ for both, so both Smith normal forms are $\operatorname{diag}(1,x^2+1)$.)

(b) Similar. By Example (A Smith Normal Form Computed by Hand) the Smith normal form of $x\mathbf{I}-\mathbf{A}$ is $\operatorname{diag}(1,x-1,(x-1)^2)$. For $\mathbf{B}$, $x\mathbf{I}-\mathbf{B}=\begin{pmatrix}x-1&-1\\0&x-1\end{pmatrix}\oplus(x-1)$. Its entry -1 gives $\Delta_1=1$. By the block-minor rule in the solution to Exercise (B2: Invariant factors from determinantal divisors), the 2×2 minors are $(x-1)^2$, from the 2×2 block, together with the products of a 1×1 minor of each block, that is $(x-1)^2,-(x-1)$, and 0; so $\Delta_2=x-1$. Finally $\Delta_3=\det(x\mathbf{I}-\mathbf{B})=(x-1)^3$. The Smith normal forms agree, so $x\mathbf{I}-\mathbf{A}\approx x\mathbf{I}-\mathbf{B}$ and $\mathbf{A}\sim\mathbf{B}$.

(c) Similar. Here

$$x\mathbf{I}-\mathbf{A}=\begin{pmatrix} x & 2 \\ -1 & x \end{pmatrix}, \qquad x\mathbf{I}-\mathbf{B}=\begin{pmatrix} x & 1 \\ -2 & x \end{pmatrix},$$

and each has an entry that is a nonzero constant, so $\Delta_1 = 1$ in both cases; and $\Delta_2 = \det(x\mathbf{I}-\mathbf{A}) = x^2+2 = \det(x\mathbf{I}-\mathbf{B})$. So both Smith normal forms are $\operatorname{diag}(1, x^2+2)$, hence $x\mathbf{I}-\mathbf{A} \approx x\mathbf{I}-\mathbf{B}$ and $\mathbf{A} \sim \mathbf{B}$ by Theorem 10.7.5. Note that x^2+2 is irreducible over $\mathbb{Q}$, so no Jordan form exists over $\mathbb{Q}$ and Exercise (C1: A matrix is similar to its transpose) would not have decided this pair. (Here $\mathbf{B} \neq \mathbf{A}^\top$, so Corollary 10.7.7 does not apply directly; the Smith forms settle it anyway.)

10.7.6.3 C. Going deeper

Exercise. C1: The same algorithm over the integers

The proof of Theorem 10.7.3 used only one property of $F[x]$: division with remainder, where the "size" of a nonzero element is its degree. The integers $\mathbb{Z}$ have the same property, with $|a|$ in place of deg.

(a) State the version of Theorem 10.7.3 for $\mathbf{M} \in M_n(\mathbb{Z})$ that the same proof gives, saying what replaces "monic".

(b) Compute the Smith normal form over $\mathbb{Z}$ of $\mathbf{M} = \begin{pmatrix} 2 & 4 \\ 6 & 4 \end{pmatrix}$, naming the operations.

(c) Every finite abelian group G with n generators is $\mathbb{Z}^n$ modulo the column span of some $\mathbf{M} \in M_n(\mathbb{Z})$, the matrix of relations, exactly as V was $F[x]^n$ modulo the columns of $x\mathbf{I}-\mathbf{A}$. Using (b), identify the group with relation matrix $\mathbf{M}$ as a product of cyclic groups.

Hint: for (c), a diagonal relation matrix $\operatorname{diag}(d_1, \dots, d_n)$ *presents the product of the cyclic groups* $\mathbb{Z}/d_i\mathbb{Z}$.

Solution.

(a) Every $\mathbf{M} \in M_n(\mathbb{Z})$ can be brought by integer elementary operations — swap two rows or columns, multiply a row or column by ± 1 (the units of $\mathbb{Z}$), add an **integer** multiple of one row or column to another — to a diagonal matrix $\operatorname{diag}(d_1, \dots, d_n)$ with every $d_i \geq 0$ and $d_i \mid d_{i+1}$, and this diagonal is unique. "Monic" is replaced by "nonnegative", which is the normalization available because the units of $\mathbb{Z}$ are ± 1. Uniqueness again follows from the determinantal divisors Δ_k, now nonnegative integer gcds of the $k \times k$ minors, using the Cauchy–Binet formula (Theorem 7.7.2, valid over any commutative ring by the remark after it) over the commutative ring $\mathbb{Z}$.

(b) The least absolute value among the entries is 2, already in the corner. It divides 4 and 6. Subtract 2 times column 1 from column 2, and 3 times row 1 from row 2:

$$\begin{pmatrix} 2 & 4 \\ 6 & 4 \end{pmatrix} \longrightarrow \begin{pmatrix} 2 & 0 \\ 6 & -8 \end{pmatrix} \longrightarrow \begin{pmatrix} 2 & 0 \\ 0 & -8 \end{pmatrix} \longrightarrow \begin{pmatrix} 2 & 0 \\ 0 & 8 \end{pmatrix},$$

the last step multiplying row 2 by -1. Since $2 \mid 8$, the Smith normal form is $\operatorname{diag}(2, 8)$. Check: $\Delta_1 = \gcd(2,4,6,4) = 2$ and $\Delta_2 = |\det \mathbf{M}| = |8-24| = 16 = 2 \cdot 8$.

(c) By (b) and the hint, the group is $\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/8\mathbb{Z}$, of order $16 = |\det \mathbf{M}|$. This is

the structure theorem for finite abelian groups in miniature: the invariant divisors $d_1 \mid \cdots \mid d_n$ of the relation matrix are the orders of the cyclic factors.

Exercise. C2: Uniqueness by hand in the smallest case

Let $\mathbf{M} \in M_2(F[x])$ be nonzero.

(a) Prove directly, without using determinantal divisors, that $\Delta_1(\mathbf{M})$ — the monic gcd of the four entries of $\mathbf{M}$ — is unchanged by each of the six elementary operations of Definition 10.7.1.

(b) Prove that $\det(\mathbf{UMV}) = c \det \mathbf{M}$ for a nonzero constant c whenever $\mathbf{U}, \mathbf{V}$ are invertible over $F[x]$.

(c) Deduce that if $\operatorname{diag}(d_1, d_2) \approx \operatorname{diag}(e_1, e_2)$ with all four entries monic or zero, $d_1 \mid d_2$, and $e_1 \mid e_2$, then $d_1 = e_1$ and $d_2 = e_2$.

Solution.

(a) Write $d = \Delta_1(\mathbf{M})$, the monic gcd of the entries, that is, the monic generator of the ideal $\langle m_{11}, m_{12}, m_{21}, m_{22} \rangle$ (Theorem 6.2.2). A swap permutes the entries, so the ideal is unchanged. Scaling a row by $c \in F \setminus \{0\}$ replaces two generators g by cg, and $\langle cg \rangle = \langle g \rangle$ since c is a unit, so the ideal is unchanged. Adding h times row 1 to row 2 replaces m_{2j} by $m_{2j} + h m_{1j}$, which lies in the old ideal; so the new ideal is contained in the old one, and the operation is reversed by adding $-h$ times row 1 to row 2, which gives the reverse containment. The same three arguments apply to columns. Hence d is unchanged.

(b) By multiplicativity of the determinant over the commutative ring $F[x]$ (Theorem 7.5.4 and the remark on commutative rings), $\det(\mathbf{UMV}) = \det \mathbf{U} \cdot \det \mathbf{M} \cdot \det \mathbf{V}$, and $\det \mathbf{U}, \det \mathbf{V}$ are nonzero constants by Proposition 10.7.1. Take $c = \det \mathbf{U} \det \mathbf{V}$.

(c) By Proposition 10.7.2 and (a), Δ_1 is an invariant of equivalence as far as elementary operations go; and for a general equivalence $\mathbf{M}' = \mathbf{UMV}$ the argument of (a) applies verbatim with the generators of the ideal replaced using $m'_{ij} = \sum_{k,l} u_{ik} m_{kl} v_{lj} \in \langle m_{kl} \rangle$, and symmetrically. For the two diagonal matrices, $\Delta_1(\operatorname{diag}(d_1, d_2)) = \gcd(d_1, d_2) = d_1$ since $d_1 \mid d_2$ and d_1 is monic (or $d_1 = d_2 = 0$); likewise $\Delta_1(\operatorname{diag}(e_1, e_2)) = e_1$. Hence $d_1 = e_1$. By (b), $d_1 d_2 = c\, e_1 e_2$ for some nonzero constant c; if $d_1 = e_1 \neq 0$ then $d_2 = c\, e_2$, and two monic polynomials (or two zeros) differing by a constant factor are equal, so $d_2 = e_2$. If $d_1 = e_1 = 0$, then $d_2 = e_2 = 0$ by the divisibility chain.

10.8 Semisimple Operators and the Jordan–Chevalley Decomposition

The Jordan canonical form says that an operator whose characteristic polynomial splits is "a diagonal matrix plus a nilpotent one" — but only after a basis has been chosen. This section turns that sentence into a statement about the operator itself: $T = S + N$ with S diagonalizable, N nilpotent and, crucially, $SN = NS$; and there is exactly one such pair. Before that we need a name for the operators that deserve to be called diagonalizable even when the field is too small to see it.

10.8.1 Diagonalizable once the field is big enough

Two operators fail to be diagonalizable for completely different reasons. The rotation $\mathbf{R} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \in M_2(\mathbb{R})$ has $p_{\mathbf{R}} = m_{\mathbf{R}} = x^2 + 1$, with no real root, so it has no eigenvectors at all (Example (A Real Matrix with Non-Real Eigenvalues)). Yet over $\mathbb{C}$ it becomes $\operatorname{diag}(i, -i)$: the operator was fine, the field was too small. The shear $\mathbf{J}_2(0)$ has $m = x^2$ and is not diagonalizable over **any** field, however large: its single eigenvalue 0 has $a = 2$ and $g = 1$, and enlarging the field changes neither.

What separates the two cases is visible in the minimal polynomial. For $\mathbf{R}$ it is x^2+1, irreducible over $\mathbb{R}$ and appearing to the first power. For $\mathbf{J}_2(0)$ it is x^2, the **square** of the irreducible x. Chapter 9 proved that T is diagonalizable exactly when m_T is a product of **distinct linear** factors (Theorem 9.8.2). Deleting the word "linear" from that criterion gives the notion we want.

A semisimple operator is one whose minimal polynomial is squarefree: no irreducible factor is repeated.

Definition 10.8.1. Semisimple Operator

Let V be a finite-dimensional vector space over F with $\dim V \geq 1$, and let $T \in \mathcal{L}(V)$. We say T is **semisimple** if

$$m_T = p_1 p_2 \cdots p_k$$

for **distinct** monic irreducible $p_1, \ldots, p_k \in F[x]$; equivalently, if **no** irreducible polynomial $p \in F[x]$ satisfies $p^2 \mid m_T$. A matrix $\mathbf{A} \in M_n(F)$ is **semisimple over** F if $T_{\mathbf{A}}$ is.

In words: factor m_T into irreducibles over F (Theorem 6.3.5). Semisimplicity says every exponent in that factorization is 1. The two descriptions agree: the exponents are all 1 exactly when no p^2 divides m_T, by uniqueness of factorization. Note the dependence on the field, which is built into the word "irreducible over F".

Examples.

- **Diagonalizable operators.** If T is diagonalizable, m_T is a product of distinct linear factors (Theorem 9.8.2), and linear polynomials are irreducible (Definition 6.3.1). So T is semisimple.
- **The rotation.** $\mathbf{R} \in M_2(\mathbb{R})$ above has $m_{\mathbf{R}} = x^2 + 1$, irreducible over $\mathbb{R}$ since it has no real root (Theorem 6.3.1). One factor, exponent 1: semisimple over $\mathbb{R}$, though not diagonalizable over $\mathbb{R}$.
- **Any operator with irreducible minimal polynomial**, for the same reason. In particular the companion matrix $\mathbf{C}(p)$ of an irreducible p is semisimple, since $m_{\mathbf{C}(p)} = p$ (Theorem 10.5.3).
- **The degenerate case.** On a space of dimension 1, $T = \lambda \operatorname{id}_V$ has $m_T = x - \lambda$: semisimple. So is id_V on any V, with $m = x - 1$.

Nonexample by minimal change. Over $\mathbb{R}$, replace $m = x^2 + 1$ by $m = (x^2 + 1)^2$: the companion matrix $\mathbf{C}((x^2+1)^2) \in M_4(\mathbb{R})$ has $m = (x^2+1)^2$ (Theorem 10.5.3), so the irreducible $x^2 + 1$ is repeated and the matrix is **not** semisimple. The clause that fails is "distinct": there is one irreducible factor, but it occurs twice.

Why this definition. The word *simple* names the case in which V has no T-invariant subspace

except $\{0\}$ and V. That happens exactly when m_T is irreducible **and** $\dim V = \deg m_T$, for then every nonzero v has T-annihilator m_T and so $Z(v;T) = V$ (Theorem 10.5.2). *Semisimple* then means "a direct sum of simple pieces", which is what Theorem 9.8.1 followed by a cyclic decomposition (Theorem 10.6.2) delivers. The definition is the right one because it is exactly diagonalizability after the field has been enlarged enough to split m_T — provided the field is not pathological. Making that precise needs one small fact: the minimal polynomial does not change when the field grows.

Lemma 10.8.1. The Minimal Polynomial Does Not Notice a Larger Field

Let $K \subseteq L$ be fields and $\mathbf{A} \in M_n(K)$, $n \geq 1$. Then the minimal polynomial of $\mathbf{A}$ computed in $K[x]$ equals the minimal polynomial of $\mathbf{A}$ computed in $L[x]$.

Idea. The degree of $m_{\mathbf{A}}$ is the position of the first linear dependence in the list $\mathbf{I}, \mathbf{A}, \mathbf{A}^2, \ldots$ (Proposition 9.5.4). Whether a list of matrices with entries in K is dependent is decided by row reducing one matrix with entries in K, and row reduction does not care that a bigger field is available.

Proof. Write m^K and m^L for the two minimal polynomials, and $d = \deg m^K$, $e = \deg m^L$.

Since $m^K \in K[x] \subseteq L[x]$ annihilates $\mathbf{A}$, Theorem 9.5.1 gives $m^L \mid m^K$, so $e \leq d$.

For the reverse inequality, list the entries of $\mathbf{I}, \mathbf{A}, \ldots, \mathbf{A}^e$ as the columns of a matrix $\mathbf{B} \in M_{n^2 \times (e+1)}(K)$, one column per power. A linear combination $\sum_{i=0}^{e} c_i \mathbf{A}^i$ is the zero matrix exactly when $\mathbf{Bc} = \mathbf{0}$ for $\mathbf{c} = (c_0, \ldots, c_e)$. By Proposition 9.5.4 applied over L, the list $(\mathbf{I}, \mathbf{A}, \ldots, \mathbf{A}^e)$ is dependent over L, so $\mathbf{Bc} = \mathbf{0}$ has a nontrivial solution in L^{e+1}, and hence $\operatorname{rank}_L \mathbf{B} < e + 1$ (Theorem 3.5.7). Now reduce $\mathbf{B}$ to reduced row echelon form by row operations performed inside K (Theorem 3.2.3); the resulting matrix $\mathbf{R}$ has entries in K, and the definition of reduced row echelon form (Definition 3.2.4) makes no reference to the field, so $\mathbf{R}$ is also **the** reduced row echelon form of $\mathbf{B}$ over L, by the uniqueness in Theorem 3.3.4. The rank is the number of pivots of $\mathbf{R}$ (Corollary 3.5.6), a number read off $\mathbf{R}$ alone. Hence

$$\operatorname{rank}_K \mathbf{B} = \operatorname{rank}_L \mathbf{B} < e + 1,$$

so $\mathbf{Bc} = \mathbf{0}$ has a nontrivial solution in K^{e+1} as well. That solution gives a nonzero $f \in K[x]$ of degree at most e with $f(\mathbf{A}) = \mathbf{O}$, whence $d = \deg m^K \leq e$ by Theorem 9.5.1 and Proposition 6.1.2.

So $d = e$, and $m^L \mid m^K$ with both monic of the same degree forces $m^L = m^K$.

■

Proposition 10.8.2. Semisimple Means Diagonalizable After Enlarging the Field

Let K be a field of characteristic 0, let $\mathbf{A} \in M_n(K)$ with $n \geq 1$, and let $L \supseteq K$ be a field over which $m_{\mathbf{A}}$ splits. Then

$$\mathbf{A} \text{ is semisimple over } K \iff \mathbf{A} \text{ is diagonalizable over } L.$$

Proof. By Lemma 10.8.1 the minimal polynomial is the same polynomial $m = m_{\mathbf{A}}$ whether computed over K or over L. By Theorem 9.8.2, $\mathbf{A}$ is diagonalizable over L if and only if m is a product of distinct linear factors in $L[x]$, that is, if and only if m splits over L — which it does by hypothesis — with **no repeated root**.

($\Leftarrow$) Suppose m has no repeated root in L. If some irreducible $p \in K[x]$ satisfied $m = p^2 h$, then $p \mid m$ and m splits over L, so p has a root $\lambda \in L$; by Theorem 6.4.2, $\operatorname{mult}_\lambda(m) \geq 2\operatorname{mult}_\lambda(p) \geq 2$, and λ would be a repeated root of m. So no such p exists and $\mathbf{A}$ is semisimple over K.

($\Rightarrow$) Suppose $m = p_1 \cdots p_k$ with $p_1, \ldots, p_k \in K[x]$ distinct monic irreducibles. The formal derivative is computed from the coefficients (Definition 6.4.3), so each p_i' is the same polynomial whether read in $K[x]$ or in $L[x]$.

No p_i has a repeated root in L. Since K has characteristic 0 and $\deg p_i \geq 1$, the derivative p_i' is nonzero of degree $\deg p_i - 1$, so $p_i \nmid p_i'$; as p_i is irreducible, Lemma 6.3.2 gives $\gcd(p_i, p_i') = 1$, and by Corollary 6.2.4 there are $a, b \in K[x]$ with $ap_i + bp_i' = 1$. If $\lambda \in L$ were a repeated root of p_i, then $p_i(\lambda) = p_i'(\lambda) = 0$ by Theorem 6.4.5 (a), applied in $L[x]$; evaluating the identity at λ would give $1 = 0$.

No two p_i share a root in L. For $i \neq j$, a monic divisor of the irreducible p_j is 1 or p_j, so $p_i \nmid p_j$, and Lemma 6.3.2 gives $\gcd(p_i, p_j) = 1$. Again by Corollary 6.2.4, $ap_i + bp_j = 1$ for some $a, b \in K[x]$, and a common root in L would give $1 = 0$.

Therefore, by Theorem 6.4.2, every root of m in L has multiplicity 1: it is a root of exactly one p_i, and a simple one. So m is a product of distinct linear factors over L, and $\mathbf{A}$ is diagonalizable over L.

■

Remark. A field over which every irreducible polynomial has distinct roots in a splitting field is called **perfect**. The proof above shows that every field of characteristic 0 is perfect, and every finite field is perfect as well, although that needs an argument we do not give here. Over a nonperfect field Proposition 10.8.2 can fail: in characteristic p there are irreducible polynomials of the form $q(x^p)$, whose derivative vanishes identically and which acquire repeated roots in a splitting field. Everything in the rest of this section takes place over a field where p_T splits, so this subtlety will not arise again.

Quick check. Over $\mathbb{Q}$, which of $\mathbf{A} = \mathbf{C}(x^2+1) \oplus \mathbf{C}(x^2+1) \in M_4(\mathbb{Q})$ and $\mathbf{B} = \mathbf{C}((x^2+1)^2) \in M_4(\mathbb{Q})$ is semisimple? Is either diagonalizable over $\mathbb{Q}$?

Solution. $\mathbf{A}$ is semisimple. Each block has minimal polynomial x^2+1 (Theorem 10.5.3), so $m_{\mathbf{A}} = \operatorname{lcm}(x^2+1, x^2+1) = x^2+1$ (Proposition 9.5.6), an irreducible polynomial to the first power. $\mathbf{B}$ is not: $m_{\mathbf{B}} = (x^2+1)^2$ repeats the irreducible x^2+1. Neither is diagonalizable over $\mathbb{Q}$, since x^2+1 has no rational root, so neither minimal polynomial is a product of linear factors. Over $\mathbb{C}$, however, $\mathbf{A}$ becomes $\operatorname{diag}(i, -i, i, -i)$, while $\mathbf{B}$ becomes $\mathbf{J}_2(i) \oplus \mathbf{J}_2(-i)$, which is not diagonal.

10.8.2 Splitting an operator in two

Here is the goal. **When p_T splits, T is the sum of a diagonalizable operator and a nilpotent operator that commute, in exactly one way, and both summands are polynomials in T.** The Jordan form already shows *a* splitting: in a Jordan basis, $\mathbf{J} = \mathbf{D} + \mathbf{M}$ where $\mathbf{D}$ is the diagonal of $\mathbf{J}$ and $\mathbf{M}$ carries the superdiagonal ones, and these commute because $\mathbf{D}$ is constant on each block. What the theorem adds is that the two pieces do not depend on which Jordan basis was used, and that they can be written down without finding one.

Theorem 10.8.3. Jordan–Chevalley Decomposition

Let V be a finite-dimensional vector space over F with $\dim V = n \geq 1$, and let $T \in \mathcal{L}(V)$ with p_T split over F. Then:

(a) **(Existence)** There are $S, N \in \mathcal{L}(V)$ with

$$T = S + N, \qquad S \text{ diagonalizable}, \qquad N \text{ nilpotent}, \qquad SN = NS;$$

moreover $S = g(T)$ and $N = h(T)$ for some $g, h \in F[x]$.

(b) **(Uniqueness)** If $T = S' + N'$ with S' diagonalizable, N' nilpotent, and $S'N' = N'S'$, then $S' = S$ and $N' = N$.

We call S the **semisimple part** and N the **nilpotent part** of T.

Idea. **Existence.** The generalized eigenspace decomposition already puts the answer in front of us: on G_λ the operator is λ id +nilpotent. Declare S to be λ id on G_λ for each λ and let N be the rest. Then S is diagonalizable because a basis assembled from bases of the G_λ consists of eigenvectors of S, and N is nilpotent because it is nilpotent on each piece. That S and N are polynomials in T is the extra information carried by the primary decomposition: its projections are $h_i(T)$.

Uniqueness. A second decomposition $S' + N'$ has S' commuting with T, hence with every polynomial in T, hence with S and with N; the same for N'. Then $S - S' = N' - N$ is simultaneously diagonalizable and nilpotent, and the only such operator is 0.

Proof.

(a) Let $\lambda_1, \dots, \lambda_k$ be the distinct eigenvalues of T and write

$$m_T = (x - \lambda_1)^{s_1} \cdots (x - \lambda_k)^{s_k}, \qquad G_i = G_{\lambda_i}(T),$$

as in Theorem 10.1.3, so that $V = G_1 \oplus \cdots \oplus G_k$ with $G_i = \ker(T - \lambda_i \operatorname{id}_V)^{s_i}$. The factors $(x - \lambda_i)^{s_i}$ are exactly the primary factors of m_T, since $x - \lambda_i$ is irreducible, so G_i is the i-th primary component of Theorem 9.8.1. By part (d) of that theorem there are $h_1, \dots, h_k \in F[x]$ such that

$$\pi_i \coloneqq h_i(T) \in \mathcal{L}(V)$$

is the projection onto G_i along $\bigoplus_{j\neq i} G_j$. Put

$$g \coloneqq \sum_{i=1}^{k} \lambda_i h_i \in F[x],$$
$$S \coloneqq g(T) = \sum_{i=1}^{k} \lambda_i \pi_i,$$
$$N \coloneqq T - S = h(T) \text{ with } h = x - g.$$

S is diagonalizable. Fix i and $v \in G_i$. Then $\pi_i v = v$ and $\pi_j v = 0$ for $j \neq i$ (Theorem 4.5.1), so $Sv = \lambda_i v$. Choose a basis of each G_i and let $\mathcal{B}$ be the concatenation, a basis of V because $V = \bigoplus_i G_i$. Every vector of $\mathcal{B}$ is an eigenvector of S, so $[S]_{\mathcal{B}}^{\mathcal{B}}$ is diagonal and S is diagonalizable (Theorem 9.4.1).
N is nilpotent. Fix i and $v \in G_i$. Then $Nv = Tv - \lambda_i v = (T - \lambda_i \operatorname{id}_V)v$, which lies in G_i because G_i is T-invariant, so by induction $N^j v = (T - \lambda_i \operatorname{id}_V)^j v$ for every $j \geq 0$. With $s = \max_i s_i$ we get $N^s v = (T - \lambda_i \operatorname{id}_V)^s v = 0$, since $(T - \lambda_i \operatorname{id}_V)^{s_i} v = 0$ and $s \geq s_i$. As every $v \in V$ is a sum of such vectors, $N^s = 0$ and N is nilpotent (Definition 10.2.1).
They commute. Both $S = g(T)$ and $N = h(T)$ are polynomials in T, and any two polynomials in T commute (Theorem 6.7.1 (c)). This proves (a).

(b) Suppose $T = S' + N'$ as stated. First, S' commutes with T:

$$S'T = S'(S' + N') = S'S' + S'N' = S'S' + N'S' = (S' + N')S' = TS',$$

where the middle equality is the hypothesis $S'N' = N'S'$. Hence S' commutes with every power of T and so with every polynomial in T, in particular with $S = g(T)$ and $N = h(T)$. The same computation gives $N'T = TN'$, so N' too commutes with S and N.

Now set $X \coloneqq S - S'$. From $S + N = T = S' + N'$ we get

$$X = S - S' = N' - N.$$

The operators S and S' are diagonalizable and commute, so they are simultaneously diagonalizable (Theorem 9.9.5): there is a basis $\mathcal{C}$ in which $[S]_{\mathcal{C}}^{\mathcal{C}}$ and $[S']_{\mathcal{C}}^{\mathcal{C}}$ are both diagonal. Then $[X]_{\mathcal{C}}^{\mathcal{C}} = [S]_{\mathcal{C}}^{\mathcal{C}} - [S']_{\mathcal{C}}^{\mathcal{C}}$ is diagonal, so X is diagonalizable. On the other hand N and N' are nilpotent and commute, hence so do N' and $-N$. Say $N^a = 0$ and $(N')^b = 0$ with $a, b \geq 1$, and put $s = a + b - 1$. By distributivity, $X^s = (N' - N)^s$ is the sum of the 2^s products of s factors, each factor N' or $-N$; since the two commute, each product can be rearranged, one adjacent swap at a time, into $\pm (N')^i N^{s-i}$, where i counts the factors equal to N'. If $i \geq b$, then $(N')^i = 0$; and if $i \leq b - 1$, then $s - i \geq s - b + 1 = a$, so $N^{s-i} = 0$. Either way the product is 0, so $X^s = 0$ and X is nilpotent.

So X is diagonalizable and nilpotent. Being nilpotent, $\operatorname{spec}(X) = \{0\}$ (Proposition 10.2.1 (c)); being diagonalizable, $[X]_{\mathcal{D}}^{\mathcal{D}}$ is diagonal with the eigenvalues of X on the diagonal for some basis $\mathcal{D}$ (Theorem 9.4.1), so $[X]_{\mathcal{D}}^{\mathcal{D}} = \mathbf{O}$ and $X = 0$. Hence $S' = S$ and, subtracting from T, $N' = N$. This proves the theorem.

■

Warning. **Splitting a matrix into its diagonal part and its off-diagonal part is**

***not* the Jordan–Chevalley decomposition.** For $\mathbf{A} = \begin{pmatrix} 1 & 1 \\ 0 & 2 \end{pmatrix}$, the "obvious" split is $\mathbf{D} = \operatorname{diag}(1,2)$ plus $\mathbf{E} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$. Here $\mathbf{D}$ is diagonal and $\mathbf{E}$ is nilpotent, but

$$\mathbf{DE} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \neq \begin{pmatrix} 0 & 2 \\ 0 & 0 \end{pmatrix} = \mathbf{ED},$$

so the commuting requirement fails, and the split is not the one the theorem produces. Indeed $\mathbf{A}$ has distinct eigenvalues, so it is diagonalizable (Corollary 9.4.2) and its actual decomposition is $\mathbf{S} = \mathbf{A}$, $\mathbf{N} = \mathbf{O}$. The off-diagonal part of a matrix is a feature of the basis; the nilpotent part is a feature of the operator.

The formulas in the proof are also the recipe: build the projections π_i onto the generalized eigenspaces — by Bézout, as in Example (A Primary Decomposition with an Irreducible Quadratic) — and set $S = \sum_i \lambda_i \pi_i$.

Example. The Semisimple and Nilpotent Parts of a 3×3 Matrix

Let

$$\mathbf{A} = \begin{pmatrix} 2 & 2 & -1 \\ -1 & 3 & -1 \\ -1 & -1 & 2 \end{pmatrix} \in M_3(\mathbb{Q}), \qquad p_{\mathbf{A}} = (x-2)^2(x-3).$$

Find the Jordan–Chevalley decomposition $\mathbf{A} = \mathbf{S} + \mathbf{N}$, and exhibit $\mathbf{S}$ and $\mathbf{N}$ as polynomials in $\mathbf{A}$.

Solution. The minimal polynomial. Put $\mathbf{B} = \mathbf{A} - 2\mathbf{I} = \begin{pmatrix} 0 & 2 & -1 \\ -1 & 1 & -1 \\ -1 & -1 & 0 \end{pmatrix}$. Then

$$\mathbf{B}(\mathbf{A} - 3\mathbf{I}) = \begin{pmatrix} 0 & 2 & -1 \\ -1 & 1 & -1 \\ -1 & -1 & 0 \end{pmatrix} \begin{pmatrix} -1 & 2 & -1 \\ -1 & 0 & -1 \\ -1 & -1 & -1 \end{pmatrix} = \begin{pmatrix} -1 & 1 & -1 \\ 1 & -1 & 1 \\ 2 & -2 & 2 \end{pmatrix} \neq \mathbf{O},$$

so $m_{\mathbf{A}} \neq (x-2)(x-3)$; since $m_{\mathbf{A}} \mid p_{\mathbf{A}}$ and $m_{\mathbf{A}}$ has both 2 and 3 as roots (Theorem 9.5.3), $m_{\mathbf{A}} = (x-2)^2(x-3)$. The primary factors are $(x-2)^2$ and $x-3$.

Bézout. Dividing, $(x-2)^2 = x^2 - 4x + 4 = (x-3)(x-1) + 1$, so

$$1 \cdot (x-2)^2 - (x-1)(x-3) = 1.$$

Following Theorem 6.7.5 with $p = (x-2)^2$ and $q = x-3$: the projection onto $\ker q(\mathbf{A}) = G_3(\mathbf{A})$ along $\ker p(\mathbf{A}) = G_2(\mathbf{A})$ is $\pi_3 = (\mathbf{A} - 2\mathbf{I})^2 = \mathbf{B}^2$, and $\pi_2 = \mathbf{I} - \mathbf{B}^2$.

Compute. Squaring,

$$\mathbf{B}^2 = \begin{pmatrix} -1 & 3 & -2 \\ 0 & 0 & 0 \\ 1 & -3 & 2 \end{pmatrix}, \qquad \text{and } \mathbf{B}^3 = \mathbf{B}^2 \text{ (as } \mathbf{B}^2(\mathbf{B} - \mathbf{I}) = \mathbf{O}).$$

Hence

$$\mathbf{S} = 2\pi_2 + 3\pi_3 = 2\mathbf{I} + \mathbf{B}^2 = \begin{pmatrix} 1 & 3 & -2 \\ 0 & 2 & 0 \\ 1 & -3 & 4 \end{pmatrix},$$

$$\mathbf{N} = \mathbf{A} - \mathbf{S} = \mathbf{B} - \mathbf{B}^2 = \begin{pmatrix} 1 & -1 & 1 \\ -1 & 1 & -1 \\ -2 & 2 & -2 \end{pmatrix}.$$

The checks. First $\mathbf{S} + \mathbf{N} = \mathbf{A}$ by construction. Next $\mathbf{N}^2 = \mathbf{O}$: reading the rows of $\mathbf{N}$ as multiples of $(1, -1, 1)$ shows that $\mathbf{N} = \mathbf{u}\mathbf{w}^\top$ with $\mathbf{u} = (1, -1, -2)$ and $\mathbf{w} = (1, -1, 1)$, and $\mathbf{w}^\top\mathbf{u} = 1+1-2 = 0$, so $\mathbf{N}^2 = \mathbf{u}(\mathbf{w}^\top\mathbf{u})\mathbf{w}^\top = \mathbf{O}$. Next $\mathbf{S}$ is diagonalizable: $(\mathbf{S}-2\mathbf{I})(\mathbf{S}-3\mathbf{I}) = \mathbf{B}^2(\mathbf{B}^2 - \mathbf{I}) = \mathbf{B}^4 - \mathbf{B}^2 = \mathbf{O}$, because $\mathbf{B}^3 = \mathbf{B}^2$ gives $\mathbf{B}^4 = \mathbf{B}^2$; so $m_{\mathbf{S}}$ divides a product of distinct linear factors and $\mathbf{S}$ is diagonalizable (Theorem 9.8.2). Finally $\mathbf{SN} = \mathbf{NS}$, since both are polynomials in $\mathbf{A}$: explicitly $\mathbf{S} = 2\mathbf{I} + (\mathbf{A} - 2\mathbf{I})^2$ and $\mathbf{N} = (\mathbf{A} - 2\mathbf{I}) - (\mathbf{A} - 2\mathbf{I})^2$, that is,

$$\mathbf{S} = g(\mathbf{A}) \text{ with } g = x^2 - 4x + 6,$$
$$\mathbf{N} = h(\mathbf{A}) \text{ with } h = -x^2 + 5x - 6 = -(x-2)(x-3).$$

As a cross-check, $g + h = x$, so $g(\mathbf{A}) + h(\mathbf{A}) = \mathbf{A}$; and $\operatorname{tr}\mathbf{S} = 1 + 2 + 4 = 7 = 2 + 2 + 3$ is the sum of the eigenvalues of $\mathbf{A}$, while $\operatorname{tr}\mathbf{N} = 1 + 1 - 2 = 0$, as it must be for a nilpotent matrix (Proposition 10.2.1 (c) and Theorem 9.3.3).

Two consequences are worth naming now. Because S and N are polynomials in T, **every operator commuting with T automatically commutes with S and with N.** And the decomposition is exactly what makes functions of matrices computable in the next section: any reasonable function f will satisfy $f(T) = f(S + N)$ with the two parts handled separately, S by its eigenvalues and N by finitely many terms, since its powers stop.

Remark. The splitting exists in greater generality. Over a perfect field, every T can be written uniquely as $T = S + N$ with S **semisimple** in the sense of Definition 10.8.1, N nilpotent, and $SN = NS$; the proof runs through the primary decomposition of m_T into powers of general irreducibles instead of linear ones. We prove only the split case, which is all we use.

10.8.3 Exercises

10.8.3.1 A. Check your understanding

Exercise. A1

(a) Define a semisimple operator, and say how the definition specializes to Chapter 9's criterion for diagonalizability.

(b) State the Jordan–Chevalley Decomposition Theorem, including all three conditions on S and N and the uniqueness clause.

(c) True or false: $\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \in M_2(\mathbb{R})$ is semisimple. Justify your answer.

(d) True or false: if $\mathbf{A} = \mathbf{D} + \mathbf{E}$ with $\mathbf{D}$ the diagonal part of $\mathbf{A}$ and $\mathbf{E}$ the off-diagonal

part, then $\mathbf{D}$ is the semisimple part of $\mathbf{A}$. Justify your answer.

(e) Which hypothesis of the uniqueness clause does the pair $(\mathbf{D}, \mathbf{E})$ of part (d) typically violate?

(f) Why does every operator commuting with T commute with the semisimple part of T?

Solution.

(a) T is semisimple if m_T is a product of **distinct monic irreducible** polynomials over F (Definition 10.8.1). Requiring those irreducibles to be **linear** gives Theorem 9.8.2, the criterion for diagonalizability. So diagonalizable $\Rightarrow$ semisimple, and the converse holds when m_T splits.

(b) If p_T splits, then $T = S + N$ with S diagonalizable, N nilpotent, and $SN = NS$; such S and N are unique, and both are polynomials in T (Theorem 10.8.3).

(c) True. Its minimal polynomial is $x^2 + 1$, which has no real root and so is irreducible over $\mathbb{R}$ (Theorem 6.3.1); one irreducible factor, exponent 1.

(d) False. For $\mathbf{A} = \begin{pmatrix} 1 & 1 \\ 0 & 2 \end{pmatrix}$ the semisimple part is $\mathbf{A}$ itself, since $\mathbf{A}$ is diagonalizable, whereas $\mathbf{D} = \operatorname{diag}(1, 2) \neq \mathbf{A}$.

(e) The commuting condition $\mathbf{DE} = \mathbf{ED}$; see the warning after Theorem 10.8.3.

(f) Because $S = g(T)$ for some polynomial g (Theorem 10.8.3 (a)), and an operator commuting with T commutes with every power of T, hence with every polynomial in T.

10.8.3.2 B. Practice

Exercise. B1: Semisimple and nilpotent parts

For each matrix over $\mathbb{Q}$, find the Jordan–Chevalley decomposition $\mathbf{A} = \mathbf{S} + \mathbf{N}$ and express $\mathbf{S}$ as a polynomial in $\mathbf{A}$.

(a) $\mathbf{A}_1 = \begin{pmatrix} 3 & 1 \\ 0 & 3 \end{pmatrix}$.

(b) $\mathbf{A}_2 = \begin{pmatrix} 1 & 2 \\ 0 & 3 \end{pmatrix}$.

(c) $\mathbf{A}_3 = \begin{pmatrix} 2 & 1 & 3 \\ 0 & 2 & 0 \\ 0 & 0 & 5 \end{pmatrix}$.

Solution.

(a) $m_{\mathbf{A}_1} = (x - 3)^2$, so there is one eigenvalue and $G_3 = \mathbb{Q}^2$; the projection onto it is $\mathbf{I}$. Hence $\mathbf{S} = 3\mathbf{I}$ and $\mathbf{N} = \mathbf{A}_1 - 3\mathbf{I} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$, with $\mathbf{N}^2 = \mathbf{O}$. As a polynomial, $\mathbf{S} = 3\mathbf{I} = g(\mathbf{A}_1)$ with $g = 3$ the constant polynomial. (Every scalar multiple of $\mathbf{I}$ is diagonalizable, and $\mathbf{S}$ commutes with everything.)

(b) $\mathbf{A}_2$ has the distinct eigenvalues 1 and 3, so it is diagonalizable (Corollary 9.4.2). By the uniqueness in Theorem 10.8.3, $\mathbf{S} = \mathbf{A}_2$ and $\mathbf{N} = \mathbf{O}$; here $g = x$. The naive split

into $\operatorname{diag}(1,3)$ and $\begin{pmatrix} 0 & 2 \\ 0 & 0 \end{pmatrix}$ is **not** the Jordan–Chevalley decomposition.

(c) $p_{\mathbf{A}_3} = (x-2)^2(x-5)$. Put $\mathbf{M} = \mathbf{A}_3 - 2\mathbf{I} = \begin{pmatrix} 0 & 1 & 3 \\ 0 & 0 & 0 \\ 0 & 0 & 3 \end{pmatrix}$. Then $\mathbf{M}(\mathbf{A}_3 - 5\mathbf{I})$ has $(1,2)$-entry $-3 \neq 0$, so $m_{\mathbf{A}_3} = (x-2)^2(x-5)$. Bézout: $(x-2)^2 = (x-5)(x+1) + 9$, so $\frac{1}{9}(x-2)^2 - \frac{1}{9}(x+1)(x-5) = 1$, and by Theorem 6.7.5 the projection onto G_5 along G_2 is $\pi_5 = \frac{1}{9}\mathbf{M}^2$. Computing $\mathbf{M}^2 = \begin{pmatrix} 0 & 0 & 9 \\ 0 & 0 & 0 \\ 0 & 0 & 9 \end{pmatrix}$,

$$\mathbf{S} = 2\mathbf{I} + 3\pi_5 = 2\mathbf{I} + \tfrac{1}{3}\mathbf{M}^2 = \begin{pmatrix} 2 & 0 & 3 \\ 0 & 2 & 0 \\ 0 & 0 & 5 \end{pmatrix}, \qquad \mathbf{N} = \mathbf{A}_3 - \mathbf{S} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Checks: $\mathbf{N}^2 = \mathbf{O}$; $\mathbf{S} - 2\mathbf{I}$ has rank 1, so $g_{\mathbf{S}}(2) = 2 = a_{\mathbf{S}}(2)$ and $\mathbf{S}$ is diagonalizable (Theorem 9.4.1); and $\mathbf{SN} = \begin{pmatrix} 0 & 2 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} = \mathbf{NS}$. As a polynomial, $\mathbf{S} = 2\mathbf{I} + \frac{1}{3}(\mathbf{A}_3 - 2\mathbf{I})^2 = g(\mathbf{A}_3)$ with $g = \frac{1}{3}x^2 - \frac{4}{3}x + \frac{10}{3}$. Note that $\mathbf{S}$ keeps the entry 3 in position $(1,3)$: the diagonal part of $\mathbf{A}_3$ is the wrong answer, and it does not commute with the rest, since $\operatorname{diag}(2,2,5)(\mathbf{A}_3 - \operatorname{diag}(2,2,5))$ and the reverse product differ in position $(1,3)$ by -9.

Exercise. B2: When is the nilpotent part zero?

Let V be finite-dimensional with $\dim V \geq 1$ and let $T \in \mathcal{L}(V)$ with p_T split, with Jordan–Chevalley decomposition $T = S + N$. Prove that $N = 0$ if and only if T is diagonalizable, and that $S = 0$ if and only if T is nilpotent.

Solution. If $N = 0$, then $T = S$ is diagonalizable. Conversely, if T is diagonalizable, then $T = T + 0$ is a decomposition into a diagonalizable and a nilpotent part that commute ($T \cdot 0 = 0 = 0 \cdot T$), so by the uniqueness in Theorem 10.8.3, $S = T$ and $N = 0$.

If $S = 0$, then $T = N$ is nilpotent. Conversely, if T is nilpotent, then $T = 0 + T$ is such a decomposition, since the zero operator is diagonalizable and commutes with T; uniqueness gives $S = 0$ and $N = T$.

Exercise. B3: Block diagonal matrices

Let $\mathbf{A}_1 \in M_p(F)$ and $\mathbf{A}_2 \in M_q(F)$ have split characteristic polynomials, with Jordan–Chevalley decompositions $\mathbf{A}_i = \mathbf{S}_i + \mathbf{N}_i$. Prove that the Jordan–Chevalley decomposition of $\mathbf{A}_1 \oplus \mathbf{A}_2$ is

$$\mathbf{A}_1 \oplus \mathbf{A}_2 = (\mathbf{S}_1 \oplus \mathbf{S}_2) + (\mathbf{N}_1 \oplus \mathbf{N}_2).$$

Solution. Write $\mathbf{S} = \mathbf{S}_1 \oplus \mathbf{S}_2$ and $\mathbf{N} = \mathbf{N}_1 \oplus \mathbf{N}_2$. By Theorem 8.1.4 (a), $\mathbf{S} + \mathbf{N} = (\mathbf{S}_1 + \mathbf{N}_1) \oplus (\mathbf{S}_2 + \mathbf{N}_2) = \mathbf{A}_1 \oplus \mathbf{A}_2$. Also $p_{\mathbf{A}_1 \oplus \mathbf{A}_2} = p_{\mathbf{A}_1} p_{\mathbf{A}_2}$ splits (Theorem 7.8.4, applied over $F[x]$ as the remark after it allows), so Theorem 10.8.3 applies to $\mathbf{A}_1 \oplus \mathbf{A}_2$ and it suffices, by uniqueness, to check the three conditions.

$\mathbf{S}$ *is diagonalizable.* Pick invertible $\mathbf{P}_i$ with $\mathbf{P}_i^{-1}\mathbf{S}_i\mathbf{P}_i$ diagonal. Then $\mathbf{P} = \mathbf{P}_1 \oplus \mathbf{P}_2$ is invertible with $\mathbf{P}^{-1} = \mathbf{P}_1^{-1} \oplus \mathbf{P}_2^{-1}$, and $\mathbf{P}^{-1}\mathbf{S}\mathbf{P} = (\mathbf{P}_1^{-1}\mathbf{S}_1\mathbf{P}_1) \oplus (\mathbf{P}_2^{-1}\mathbf{S}_2\mathbf{P}_2)$ is diagonal (Theorem 8.1.4 (a), (c)).

$\mathbf{N}$ *is nilpotent.* If $\mathbf{N}_1^a = \mathbf{O}$ and $\mathbf{N}_2^b = \mathbf{O}$, then $\mathbf{N}^{\max(a,b)} = \mathbf{N}_1^{\max(a,b)} \oplus \mathbf{N}_2^{\max(a,b)} = \mathbf{O}$ by Theorem 8.1.4 (b).

They commute. $\mathbf{S}\mathbf{N} = (\mathbf{S}_1\mathbf{N}_1) \oplus (\mathbf{S}_2\mathbf{N}_2) = (\mathbf{N}_1\mathbf{S}_1) \oplus (\mathbf{N}_2\mathbf{S}_2) = \mathbf{N}\mathbf{S}$, again by Theorem 8.1.4 (a).

By the uniqueness in Theorem 10.8.3, $\mathbf{S}$ and $\mathbf{N}$ are the semisimple and nilpotent parts of $\mathbf{A}_1 \oplus \mathbf{A}_2$.

10.8.3.3 C. Going deeper

Exercise. C1: The multiplicative version

Call $U \in \mathcal{L}(V)$ **unipotent** if $U - \mathrm{id}_V$ is nilpotent. Let $T \in \mathcal{L}(V)$ be **invertible** with p_T split, and let $T = S + N$ be its Jordan–Chevalley decomposition.

(a) Prove that S is invertible.

(b) Prove that S^{-1} commutes with N, and that $S^{-1}N$ is nilpotent.

(c) Deduce that $T = SU$ with S diagonalizable, U unipotent, and $SU = US$.

(d) Prove that this multiplicative decomposition is unique: if $T = S'U'$ with S' diagonalizable, U' unipotent, and $S'U' = U'S'$, then $S' = S$ and $U' = U$.

Hint: for (d), turn the multiplicative statement into an additive one by writing $T = S' + S'(U' - \mathrm{id}_V)$.

Solution.

(a) By the proof of Theorem 10.8.3, $Sv = \lambda_i v$ for every $v \in G_{\lambda_i}(T)$, so $\mathrm{spec}(S) = \mathrm{spec}(T)$ — a basis of eigenvectors of S with these eigenvalues was exhibited there. Since T is invertible, $0 \notin \mathrm{spec}(T)$ (Theorem 9.2.6), so $0 \notin \mathrm{spec}(S)$ and S is invertible.

(b) From $SN = NS$, multiply on both sides by S^{-1}: $S^{-1}(SN)S^{-1} = S^{-1}(NS)S^{-1}$ gives $NS^{-1} = S^{-1}N$. Hence $(S^{-1}N)^j = S^{-j}N^j$ for every $j \geq 1$, by induction using $NS^{-1} = S^{-1}N$. With $N^s = 0$ we get $(S^{-1}N)^s = 0$.

(c) Put $U = \mathrm{id}_V + S^{-1}N$. Then $SU = S + N = T$, and $U - \mathrm{id}_V = S^{-1}N$ is nilpotent by (b), so U is unipotent. They commute: $SU = S + N = US$, since $S(S^{-1}N) = N = (S^{-1}N)S$ by (b).

(d) Suppose $T = S'U'$ with S' diagonalizable, U' unipotent, and $S'U' = U'S'$. Put $N' = S'(U' - \mathrm{id}_V)$, so that $T = S' + N'$. Then N' is nilpotent: S' commutes with U', hence with $U' - \mathrm{id}_V$, so $(N')^j = (S')^j(U' - \mathrm{id}_V)^j = 0$ once j is large. And $S'N' = S'S'(U' - \mathrm{id}_V) = S'(U' - \mathrm{id}_V)S' = N'S'$. So $T = S' + N'$ satisfies the hypotheses of Theorem 10.8.3 (b), giving $S' = S$ and $N' = N$. Then $U' - \mathrm{id}_V = (S')^{-1}N' = S^{-1}N =$

$U - \mathrm{id}_V$, so $U' = U$.

Exercise. C2: The parts inherit every symmetry

Let V be finite-dimensional with $\dim V \geq 1$, let $T \in \mathcal{L}(V)$ with p_T split, and let $T = S + N$ be its Jordan–Chevalley decomposition.

(a) Prove that if $X \in \mathcal{L}(V)$ satisfies $XT = TX$, then $XS = SX$ and $XN = NX$.

(b) Deduce that every T-invariant subspace is S-invariant and N-invariant, and that S and N map each $G_\lambda(T)$ into itself.

(c) Give an example of a diagonalizable S and a nilpotent N with $SN \neq NS$ for which $S + N$ is **not** the operator whose semisimple part is S. (So the commuting hypothesis cannot be dropped from the uniqueness statement.)

Solution.

(a) By Theorem 10.8.3 (a), $S = g(T)$ and $N = h(T)$ for some $g, h \in F[x]$. From $XT = TX$ we get $XT^j = T^jX$ for all $j \geq 0$ by induction, hence $Xf(T) = f(T)X$ for every $f \in F[x]$ by taking linear combinations. Applying this to $f = g$ and $f = h$ gives $XS = SX$ and $XN = NX$.

(b) Let $U \subseteq V$ be T-invariant. Then U is T^j-invariant for every j, hence $f(T)$-invariant for every $f \in F[x]$ (a sum of vectors of U lies in U); taking $f = g, h$ gives that U is S- and N-invariant. In particular $G_\lambda(T) = \ker(T - \lambda\,\mathrm{id}_V)^n$ is T-invariant (Corollary 10.1.2 (b)), so S and N map it into itself.

(c) Take $\mathbf{S} = \mathrm{diag}(1, 2)$ and $\mathbf{N} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$ in $M_2(\mathbb{Q})$. Then $\mathbf{S}$ is diagonal, $\mathbf{N}^2 = \mathbf{O}$, and $\mathbf{SN} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \neq \begin{pmatrix} 0 & 2 \\ 0 & 0 \end{pmatrix} = \mathbf{NS}$. Their sum $\mathbf{A} = \begin{pmatrix} 1 & 1 \\ 0 & 2 \end{pmatrix}$ has distinct eigenvalues, so it is diagonalizable, and its semisimple part is $\mathbf{A}$ itself, not $\mathbf{S}$, with nilpotent part $\mathbf{O}$, not $\mathbf{N}$ (Exercise (B2: When is the nilpotent part zero?)). Without the commuting hypothesis, an operator can be written as "diagonalizable plus nilpotent" in many ways: for any $c \in \mathbb{Q}$, $\mathbf{A} = \begin{pmatrix} 1 & c \\ 0 & 2 \end{pmatrix} + \begin{pmatrix} 0 & 1 - c \\ 0 & 0 \end{pmatrix}$ is such a splitting, the first summand being diagonalizable because its eigenvalues $1, 2$ are distinct.

10.9 Functions of Matrices, the Exponential, and Linear Differential Equations

We have known since Chapter 2 what $p(\mathbf{A})$ means for a polynomial p. This section gives a meaning to $f(\mathbf{A})$ for a much larger class of functions — square roots, logarithms, sines, and above all $e^{\mathbf{A}}$ — and the Jordan form is what makes it possible. The definition needs a well-definedness proof, which turns out to be Hermite interpolation from Chapter 6: whatever f is, $f(\mathbf{A})$ is really $p(\mathbf{A})$ for a polynomial p that agrees with f to the right order at each eigenvalue.

With $e^{t\mathbf{A}}$ in hand, every linear system of differential equations with constant coefficients is solved in one formula.

Throughout this section $F = \mathbb{C}$ (or $\mathbb{R}$, when we say so), so that every characteristic polynomial splits (Corollary 6.5.5) and Jordan forms exist.

10.9.1 What a polynomial does to a Jordan block

Start with what we already have. A Jordan block is a scalar plus a nilpotent matrix whose powers stop, and a polynomial applied to it can be read off from a Taylor expansion.

Lemma 10.9.1. A Polynomial of a Jordan Block

Let F have characteristic 0, let $k \geq 1$, $\lambda \in F$, and $p \in F[x]$. Then $p(\mathbf{J}_k(\lambda))$ is the upper triangular matrix whose $(r, r+j)$-entry is

$$\bigl(p(\mathbf{J}_k(\lambda))\bigr)_{r,r+j} = \frac{p^{(j)}(\lambda)}{j!} \qquad (0 \leq j \leq k-1,\ 1 \leq r \leq k-j),$$

with all entries below the diagonal equal to 0. In particular the value $p(\lambda)$ fills the diagonal, $p'(\lambda)$ the superdiagonal, and only the first k derivatives of p at λ are involved.

Proof. Write $\mathbf{J} = \mathbf{J}_k(\lambda) = \lambda\mathbf{I}_k + \mathbf{M}$ with $\mathbf{M} = \mathbf{J}_k(0)$ (Definition 10.3.1). Reading off the columns of $\mathbf{M}$, we have $\mathbf{M}\mathbf{e}_1 = \mathbf{0}$ and $\mathbf{M}\mathbf{e}_r = \mathbf{e}_{r-1}$ for $r \geq 2$, so by induction $\mathbf{M}^j\mathbf{e}_r = \mathbf{e}_{r-j}$ when $r > j$ and $\mathbf{M}^j\mathbf{e}_r = \mathbf{0}$ when $r \leq j$. Hence, for $0 \leq j \leq k-1$, the matrix $\mathbf{M}^j$ has 1 in each position $(r, r+j)$ and 0 elsewhere, and $\mathbf{M}^k = \mathbf{O}$. Let $d = \deg p$. By Theorem 6.4.7 (b), applied with $c = \lambda$ and $n = d$,

$$p(x) = \sum_{j=0}^{d} \frac{p^{(j)}(\lambda)}{j!}\,(x-\lambda)^j.$$

Substituting $\mathbf{J}$ for x — legitimate because evaluation at $\mathbf{J}$ is a ring homomorphism (Theorem 6.7.1) — and using $\mathbf{J} - \lambda\mathbf{I}_k = \mathbf{M}$,

$$p(\mathbf{J}) = \sum_{j=0}^{d} \frac{p^{(j)}(\lambda)}{j!}\,\mathbf{M}^j = \sum_{j=0}^{\min(d,\,k-1)} \frac{p^{(j)}(\lambda)}{j!}\,\mathbf{M}^j,$$

since $\mathbf{M}^j = \mathbf{O}$ for $j \geq k$. Each $\mathbf{M}^j$ contributes its coefficient to every position $(r, r+j)$ and to no other position, which is the claim. ∎

For $p = x^m$ this reads $p^{(j)}(\lambda)/j! = \binom{m}{j}\lambda^{m-j}$, so

$$\bigl(\mathbf{J}_k(\lambda)^m\bigr)_{r,\,r+j} = \binom{m}{j}\lambda^{\,m-j},$$

which is the formula behind everything that follows, here and in the next section.

10.9.2 Functions of a matrix

Reading Lemma 10.9.1 backwards suggests the definition. The formula for $p(\mathbf{J}_k(\lambda))$ uses nothing about p except the numbers $p(\lambda), p'(\lambda), \dots, p^{(k-1)}(\lambda)$. Any function that has those derivatives can be inserted instead.

To apply f to a matrix, apply it to each Jordan block, putting $f^{(j)}(\lambda)/j!$ on the j-th superdiagonal.

Definition 10.9.1. Function of a Matrix

Let $\mathbf{A} \in M_n(\mathbb{C})$ with Jordan form $\mathbf{J} = \mathbf{J}_{k_1}(\lambda_1) \oplus \cdots \oplus \mathbf{J}_{k_m}(\lambda_m)$ and $\mathbf{A} = \mathbf{P}\mathbf{J}\mathbf{P}^{-1}$. Let f be a complex-valued function that is defined and **at least $k_i - 1$ times differentiable at λ_i** for every i — for instance, a function given by a convergent power series on a disc around each eigenvalue. Define

$$f(\mathbf{J}_k(\lambda)) \coloneqq \begin{pmatrix} f(\lambda) & f'(\lambda) & \frac{f''(\lambda)}{2!} & \cdots & \frac{f^{(k-1)}(\lambda)}{(k-1)!} \\ & f(\lambda) & f'(\lambda) & \ddots & \vdots \\ & & \ddots & \ddots & \frac{f''(\lambda)}{2!} \\ & & & f(\lambda) & f'(\lambda) \\ & & & & f(\lambda) \end{pmatrix} \in M_k(\mathbb{C}),$$

the upper triangular matrix with $f^{(j)}(\lambda)/j!$ in every position $(r, r+j)$, and

$$f(\mathbf{A}) \coloneqq \mathbf{P}\left(f(\mathbf{J}_{k_1}(\lambda_1)) \oplus \cdots \oplus f(\mathbf{J}_{k_m}(\lambda_m))\right)\mathbf{P}^{-1}.$$

In words: the diagonal of $f(\mathbf{J})$ records the values $f(\lambda_i)$, and each further superdiagonal records the next derivative, divided by a factorial. For a diagonalizable $\mathbf{A}$ every block has size 1, no derivatives are needed, and the definition says the familiar thing: $f(\mathbf{A}) = \mathbf{P}\,\mathrm{diag}(f(\lambda_1), \dots, f(\lambda_n))\mathbf{P}^{-1}$.

Well-definedness. Two choices were made: the invertible $\mathbf{P}$ and the order of the blocks. The Jordan form itself is unique only up to the order of the blocks (Theorem 10.3.1), and even for a fixed $\mathbf{J}$ there are many $\mathbf{P}$ with $\mathbf{A} = \mathbf{P}\mathbf{J}\mathbf{P}^{-1}$. The next theorem removes both worries at once, by identifying $f(\mathbf{A})$ with $p(\mathbf{A})$ for a suitable **polynomial** p, an expression in which no choice appears.

Theorem 10.9.2. Functions of a Matrix Are Polynomials in It

Let $\mathbf{A} \in M_n(\mathbb{C})$ with distinct eigenvalues $\mu_1, \dots, \mu_d$, and write

$$m_{\mathbf{A}} = (x - \mu_1)^{s_1} \cdots (x - \mu_d)^{s_d},$$

so that s_l is the largest size of a μ_l-block of the Jordan form of $\mathbf{A}$ (Corollary 10.3.2 (c)). Let f be as in Definition 10.9.1. Then:

(a) there is a polynomial $p \in \mathbb{C}[x]$ with $\deg p < \deg m_{\mathbf{A}}$ and

$$p^{(j)}(\mu_l) = f^{(j)}(\mu_l) \qquad \text{for all } l = 1, \dots, d \text{ and } j = 0, \dots, s_l - 1,$$

and p is the **only** such polynomial of degree less than $\deg m_{\mathbf{A}}$;

(b) for **every** polynomial $q \in \mathbb{C}[x]$ satisfying those $\deg m_{\mathbf{A}}$ equations, $q(\mathbf{A}) = p(\mathbf{A})$;

(c) $f(\mathbf{A}) = p(\mathbf{A})$. In particular $f(\mathbf{A})$ does not depend on the choice of $\mathbf{P}$ or on the order of the blocks, and $f(\mathbf{A})$ is determined by the finitely many numbers $f^{(j)}(\mu_l)$.

The polynomial p of (a) is the **Hermite interpolating polynomial** of f on the spectrum of $\mathbf{A}$.

Idea. The list of conditions "match f and its first $s_l - 1$ derivatives at μ_l" is exactly a congruence modulo $(x-\mu_l)^{s_l}$, by the Taylor expansion; and the moduli $(x-\mu_l)^{s_l}$ are pairwise coprime. So the Chinese Remainder Theorem produces p and shows it is unique below the degree of the product $m_{\mathbf{A}}$ — this is Example (Values and Slopes by CRT) with more nodes and higher orders. Once p exists, Lemma 10.9.1 says that p and f do the same thing to every Jordan block of $\mathbf{A}$, because a μ_l-block has size at most s_l and only derivatives up to order $s_l - 1$ are consulted.

Proof.

(a) For each l let $r_l := \sum_{j=0}^{s_l-1} \frac{f^{(j)}(\mu_l)}{j!}(x-\mu_l)^j \in \mathbb{C}[x]$. For $g \in \mathbb{C}[x]$, the Taylor expansion of Theorem 6.4.7 (b) at μ_l, together with the uniqueness of the remainder on division by $(x-\mu_l)^{s_l}$ (Theorem 6.1.1), gives

$$g \equiv r_l \pmod{(x-\mu_l)^{s_l}} \qquad \Longleftrightarrow \qquad g^{(j)}(\mu_l) = f^{(j)}(\mu_l) \text{ for } j = 0, \dots, s_l - 1,$$

exactly as in Example (Values and Slopes by CRT) (a) for $s_l = 2$: the remainder of g on division by $(x-\mu_l)^{s_l}$ is $\sum_{j<s_l} \frac{g^{(j)}(\mu_l)}{j!}(x-\mu_l)^j$, and two polynomials of degree less than s_l are equal precisely when their coefficients in the basis $(1, x-\mu_l, \dots, (x-\mu_l)^{s_l-1})$ agree (Theorem 6.4.7 (a)).
The moduli $(x-\mu_1)^{s_1}, \dots, (x-\mu_d)^{s_d}$ are pairwise coprime: for $l \neq l'$, any common divisor divides both, and a monic common divisor would be a power of $x-\mu_l$ and a power of $x-\mu_{l'}$ with $\mu_l \neq \mu_{l'}$, hence 1 (Lemma 9.5.2, Theorem 6.3.5). Their product is $m_{\mathbf{A}}$. So Theorem 6.6.4 (c) provides exactly one p with $\deg p < \deg m_{\mathbf{A}}$ and $p \equiv r_l \pmod{(x-\mu_l)^{s_l}}$ for every l, which by the displayed equivalence is (a).

(b) Let q satisfy the same equations. By the equivalence again, $q \equiv r_l \equiv p \pmod{(x-\mu_l)^{s_l}}$ for every l, so Theorem 6.6.4 (b) gives $m_{\mathbf{A}} \mid q - p$. Writing $q - p = m_{\mathbf{A}} h$ and evaluating at $\mathbf{A}$, $q(\mathbf{A}) - p(\mathbf{A}) = m_{\mathbf{A}}(\mathbf{A})h(\mathbf{A}) = \mathbf{O}$ since $m_{\mathbf{A}}(\mathbf{A}) = \mathbf{O}$ (Definition 9.5.1). Hence $q(\mathbf{A}) = p(\mathbf{A})$.

(c) Let $\mathbf{J} = \mathbf{J}_{k_1}(\lambda_1) \oplus \cdots \oplus \mathbf{J}_{k_m}(\lambda_m)$ be any Jordan form of $\mathbf{A}$ and $\mathbf{A} = \mathbf{P}\mathbf{J}\mathbf{P}^{-1}$. Fix i and let l be the index with $\lambda_i = \mu_l$. By Corollary 10.3.2 (c), $k_i \leq s_l$, so Lemma 10.9.1 and the equations in (a) give

$$\left(p(\mathbf{J}_{k_i}(\lambda_i))\right)_{r,r+j} = \frac{p^{(j)}(\lambda_i)}{j!} = \frac{f^{(j)}(\lambda_i)}{j!} = \left(f(\mathbf{J}_{k_i}(\lambda_i))\right)_{r,r+j} \qquad (0 \leq j \leq k_i - 1),$$

the derivatives involved having order at most $k_i - 1 \leq s_l - 1$. So $p(\mathbf{J}_{k_i}(\lambda_i)) = f(\mathbf{J}_{k_i}(\lambda_i))$ for every i. Since evaluating a polynomial blockwise gives the polynomial of the block

diagonal matrix (Theorem 8.1.4 (a), (b)),

$$p(\mathbf{A}) = p(\mathbf{P}\mathbf{J}\mathbf{P}^{-1}) = \mathbf{P}\,p(\mathbf{J})\,\mathbf{P}^{-1} = \mathbf{P}\Big(\bigoplus_i p(\mathbf{J}_{k_i}(\lambda_i))\Big)\mathbf{P}^{-1}$$
$$= \mathbf{P}\Big(\bigoplus_i f(\mathbf{J}_{k_i}(\lambda_i))\Big)\mathbf{P}^{-1} = f(\mathbf{A}),$$

the second equality by Proposition 4.2.6 (c). The left-hand side mentions neither $\mathbf{P}$ nor the order of the blocks, so neither does the right-hand side. This proves the theorem.

■

So the definition is consistent with the old one: if f happens to be a polynomial, then f itself satisfies the interpolation conditions, and (b) gives $f(\mathbf{A})$ in the new sense $= f(\mathbf{A})$ in the old sense.

Warning. **$f(\mathbf{A})$ is not f applied to the entries of $\mathbf{A}$.** Even for $f(x) = x^2$ this fails: with $\mathbf{A} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$, squaring each entry gives $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$, while $\mathbf{A}^2 = \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix}$. The definition is designed so that $f(\mathbf{A})$ is a polynomial in $\mathbf{A}$, and entrywise application is not.

Example. Square Roots of Two Matrices

Using Definition 10.9.1 with $f(x) = \sqrt{x}$, the branch with $\sqrt{4} = 2$ and $\sqrt{9} = 3$, compute a square root of

(a) $\mathbf{A} = \begin{pmatrix} 5 & 4 \\ 4 & 5 \end{pmatrix}$;

(b) $\mathbf{B} = \begin{pmatrix} 4 & 1 \\ 0 & 4 \end{pmatrix}$.

Solution.

(a) $p_{\mathbf{A}} = (x-5)^2 - 16 = x^2 - 10x + 9 = (x-1)(x-9)$, so the eigenvalues are 1 and 9, distinct, and $\mathbf{A}$ is diagonalizable. Solving $(\mathbf{A} - \mathbf{I})\mathbf{x} = \mathbf{0}$ gives $\operatorname{span}((1,-1))$ and $(\mathbf{A} - 9\mathbf{I})\mathbf{x} = \mathbf{0}$ gives $\operatorname{span}((1,1))$. With $\mathbf{P} = \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}$, $\det \mathbf{P} = 2$, and $\mathbf{P}^{-1} = \frac{1}{2}\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}$, we have $\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \operatorname{diag}(1, 9)$ and

$$f(\mathbf{A}) = \mathbf{P}\operatorname{diag}(1,3)\mathbf{P}^{-1} = \tfrac{1}{2}\begin{pmatrix} 1 & 3 \\ -1 & 3 \end{pmatrix}\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} = \tfrac{1}{2}\begin{pmatrix} 4 & 2 \\ 2 & 4 \end{pmatrix} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}.$$

Check. $\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}^2 = \begin{pmatrix} 5 & 4 \\ 4 & 5 \end{pmatrix} = \mathbf{A}$.

(b) $\mathbf{B} = \mathbf{J}_2(4)$ is its own Jordan form, with $\mathbf{P} = \mathbf{I}$. Here $f(4) = 2$ and $f'(x) = \frac{1}{2\sqrt{x}}$, so

$f'(4) = \frac{1}{4}$, and Definition 10.9.1 gives

$$f(\mathbf{B}) = \begin{pmatrix} 2 & \frac{1}{4} \\ 0 & 2 \end{pmatrix}.$$

Check. $\begin{pmatrix} 2 & 1/4 \\ 0 & 2 \end{pmatrix}^2 = \begin{pmatrix} 4 & 2 \cdot \frac{1}{4} + \frac{1}{4} \cdot 2 \\ 0 & 4 \end{pmatrix} = \begin{pmatrix} 4 & 1 \\ 0 & 4 \end{pmatrix} = \mathbf{B}.$

The Hermite polynomial. For (b), $m_{\mathbf{B}} = (x-4)^2$, and the interpolating polynomial of degree < 2 with $p(4) = 2$, $p'(4) = \frac{1}{4}$ is $p = 2 + \frac{1}{4}(x-4) = \frac{1}{4}x + 1$. Indeed $p(\mathbf{B}) = \frac{1}{4}\mathbf{B} + \mathbf{I} = \begin{pmatrix} 2 & 1/4 \\ 0 & 2 \end{pmatrix}$, as Theorem 10.9.2 (c) predicts.

Quick check. A matrix $\mathbf{A} \in M_5(\mathbb{C})$ has Jordan form $\mathbf{J}_3(2) \oplus \mathbf{J}_1(2) \oplus \mathbf{J}_1(7)$. Which values of f and its derivatives does $f(\mathbf{A})$ depend on?

Solution. On $f(2), f'(2), f''(2)$, and $f(7)$: four numbers. Indeed $m_{\mathbf{A}} = (x-2)^3(x-7)$ by Corollary 10.3.2 (c), so $s_1 = 3$ for $\mu_1 = 2$ and $s_2 = 1$ for $\mu_2 = 7$, and Theorem 10.9.2 lists the derivatives of order $0, 1, 2$ at 2 and order 0 at 7. Note $\deg m_{\mathbf{A}} = 4$, matching the number of conditions and the degree bound on the interpolating polynomial. Nothing about f away from $\{2, 7\}$ matters, and no third derivative is consulted even though $\dim = 5$.

10.9.3 The matrix exponential

The most important function of a matrix is the exponential, and it has a second description, as the sum of a series. We check that the series converges, and that its sum is the matrix $\exp(\mathbf{A})$ that Definition 10.9.1 produces. Convergence is meant **entrywise**, as in Chapter 9: a sequence $\mathbf{M}_1, \mathbf{M}_2, \ldots$ in $M_n(\mathbb{C})$ converges to $\mathbf{M}$ if for each position (r, s) the number sequence $(\mathbf{M}_m)_{rs}$ converges to $\mathbf{M}_{rs}$, and a series of matrices converges if its sequence of partial sums does. No notion of length of a matrix is used anywhere in this section.

Remark. **What we assume from analysis.** Everything below uses exactly the following facts about real and complex numbers, and nothing else. Chapter 16 will develop the tools that make such statements routine; here they are quoted.

(A1) **The scalar exponential.** For every $z \in \mathbb{C}$ the series $\sum_{i \geq 0} z^i/i!$ converges; its sum is written e^z. It satisfies $e^{z+w} = e^z e^w$, $e^0 = 1$, and $e^{i\theta} = \cos\theta + i\sin\theta$ for real θ. For real $s \geq 0$, $e^s \geq 1$.

(A2) **Termwise differentiation.** If a power series $\sum_{m \geq 0} c_m t^m$ with $c_m \in \mathbb{C}$ converges for **every** real t, then the function $t \mapsto \sum_m c_m t^m$ is differentiable on $\mathbb{R}$ with derivative $\sum_{m \geq 1} m c_m t^{m-1}$.

(A3) **Limits and products.** Limits of convergent sequences of complex numbers respect sums, scalar multiples, and products. The derivative of a product of two differentiable functions $\mathbb{R} \to \mathbb{C}$ obeys the product rule, and differentiation is linear.

(A4) **Zero derivative.** A differentiable function $\mathbb{R} \to \mathbb{C}$ whose derivative vanishes every-

where is constant.

(A5) **Decay.** For every integer $j \geq 0$ and every real $c < 0$, $t^j e^{ct} \to 0$ as $t \to \infty$.

Fact (A5) is used only in the last exercise of this section. Nothing else about exp, cos, or sin is assumed: the derivative $\frac{d}{dt} e^{zt} = z\, e^{zt}$ for $z \in \mathbb{C}$ is the 1×1 case of Theorem 10.9.4 (b) below, proved from (A2), and the derivatives of cos and sin used in the examples follow from it through Euler's formula in (A1).

Definition 10.9.2. Matrix Exponential

Let $\mathbf{A} \in M_n(\mathbb{C})$. The **matrix exponential** of $\mathbf{A}$ is

$$e^{\mathbf{A}} := \sum_{m=0}^{\infty} \frac{\mathbf{A}^m}{m!} = \lim_{\mathbf{M}\to\infty} \sum_{m=0}^{\mathbf{M}} \frac{\mathbf{A}^m}{m!},$$

the limit taken entrywise, provided it exists; $\mathbf{A}^0 = \mathbf{I}$.

That the limit does exist, and what it is, is the content of the next theorem. The Jordan form reduces the question to a single block, where the series can be summed exactly.

Theorem 10.9.3. The Exponential Series Converges to the Jordan Definition

Let $\mathbf{A} \in M_n(\mathbb{C})$. Then the series $\sum_m \mathbf{A}^m/m!$ converges entrywise, and its sum is the matrix $\exp(\mathbf{A})$ given by Definition 10.9.1 for the function $f = \exp$. In particular, for a single Jordan block,

$$e^{\mathbf{J}_k(\lambda)} = e^{\lambda} \begin{pmatrix} 1 & 1 & \frac{1}{2!} & \cdots & \frac{1}{(k-1)!} \\ & 1 & 1 & \ddots & \vdots \\ & & \ddots & \ddots & \frac{1}{2!} \\ & & & 1 & 1 \\ & & & & 1 \end{pmatrix},$$

the upper triangular matrix with $e^{\lambda}/j!$ in every position $(r, r+j)$.

Idea. For one block the entries of the partial sums are themselves partial sums of a scalar series, and that series can be recognized: the terms $\binom{m}{j}\lambda^{m-j}/m!$ simplify to $\lambda^{m-j}/(j!\,(m-j)!)$, which is $1/j!$ times the $(m-j)$-th term of the series for e^{λ}. For general $\mathbf{A}$, conjugation by a fixed $\mathbf{P}$ turns each entry of a partial sum into a fixed finite linear combination of entries of the block partial sums, and limits pass through finite linear combinations.

Proof. **Step 1. One block.** Fix $k \geq 1$, $\lambda \in \mathbb{C}$, and a position $(r, r+j)$ with $0 \leq j \leq k-1$. By Lemma 10.9.1 applied to $p = x^m$,

$$\Big(\frac{\mathbf{J}_k(\lambda)^m}{m!}\Big)_{r,\,r+j} = \frac{1}{m!}\binom{m}{j}\lambda^{m-j} = \begin{cases} \dfrac{\lambda^{m-j}}{j!\,(m-j)!}, & m \geq j, \\ 0, & m < j, \end{cases}$$

using $\binom{m}{j}/m! = 1/(j!\,(m-j)!)$ for $m \ge j$ and $\binom{m}{j} = 0$ for $m < j$. Hence the partial sums satisfy

$$\sum_{m=0}^{\mathbf{M}} \Big(\frac{\mathbf{J}_k(\lambda)^m}{m!}\Big)_{r,\,r+j} = \frac{1}{j!}\sum_{i=0}^{\mathbf{M}-j}\frac{\lambda^i}{i!} \qquad (\mathbf{M} \ge j),$$

after the substitution $i = m - j$. By (A1) the right-hand side converges to $e^{\lambda}/j!$ as $\mathbf{M} \to \infty$. Entries with $j < 0$, that is entries below the diagonal, are 0 in every term. So the series converges entrywise and

$$e^{\mathbf{J}_k(\lambda)} = \Big(\text{upper triangular with } \tfrac{e^{\lambda}}{j!} \text{ on the } j\text{-th superdiagonal}\Big) = \exp(\mathbf{J}_k(\lambda)),$$

the last equality being Definition 10.9.1 for $f = \exp$, whose derivatives are $f^{(j)} = \exp$, so $f^{(j)}(\lambda)/j! = e^{\lambda}/j!$.

Step 2. A Jordan matrix. If $\mathbf{J} = \mathbf{J}_{k_1}(\lambda_1) \oplus \cdots \oplus \mathbf{J}_{k_m}(\lambda_m)$, then $\mathbf{J}^m/m!$ is the direct sum of the corresponding blocks (Theorem 8.1.4 (b)), so each entry of a partial sum of $\sum_m \mathbf{J}^m/m!$ is either identically 0 or an entry of a partial sum from one block. By Step 1 all of them converge, and $e^{\mathbf{J}} = \bigoplus_i e^{\mathbf{J}_{k_i}(\lambda_i)}$.

Step 3. General A. By Theorem 10.3.1 there are a Jordan matrix $\mathbf{J}$ and an invertible $\mathbf{P}$ with $\mathbf{A} = \mathbf{P}\mathbf{J}\mathbf{P}^{-1}$. Then $\mathbf{A}^m = \mathbf{P}\mathbf{J}^m\mathbf{P}^{-1}$ (Proposition 4.2.6 (c)), so

$$\sum_{m=0}^{\mathbf{M}} \frac{\mathbf{A}^m}{m!} = \mathbf{P}\Big(\sum_{m=0}^{\mathbf{M}} \frac{\mathbf{J}^m}{m!}\Big)\mathbf{P}^{-1}.$$

Each entry of the left-hand side is a fixed finite linear combination — with coefficients built from the entries of $\mathbf{P}$ and $\mathbf{P}^{-1}$, independent of $\mathbf{M}$ — of the entries of the middle partial sum. By Step 2 those entries converge, so by (A3) the left-hand side converges entrywise, with

$$e^{\mathbf{A}} = \mathbf{P}\,e^{\mathbf{J}}\,\mathbf{P}^{-1} = \mathbf{P}\Big(\bigoplus_i e^{\mathbf{J}_{k_i}(\lambda_i)}\Big)\mathbf{P}^{-1} = \exp(\mathbf{A}),$$

the last equality by Definition 10.9.1. This proves the theorem.

■

From now on we write $e^{\mathbf{A}}$ for this matrix and use both descriptions freely. Note one consequence of Theorem 10.9.2: $e^{\mathbf{A}}$ is a **polynomial** in $\mathbf{A}$, namely $p(\mathbf{A})$ for the Hermite interpolant of exp on $\operatorname{spec}(\mathbf{A})$. Everything that commutes with $\mathbf{A}$ therefore commutes with $e^{\mathbf{A}}$.

Theorem 10.9.4. Properties of the Matrix Exponential

Let $\mathbf{A}, \mathbf{B} \in M_n(\mathbb{C})$ and let $\mathbf{P} \in M_n(\mathbb{C})$ be invertible.

(a) $e^{\mathbf{O}} = \mathbf{I}$ and $e^{\mathbf{P}^{-1}\mathbf{A}\mathbf{P}} = \mathbf{P}^{-1}e^{\mathbf{A}}\mathbf{P}$.

(b) The function $\mathbb{R} \to M_n(\mathbb{C})$, $t \mapsto e^{t\mathbf{A}}$, is differentiable entrywise, with

$$\frac{d}{dt}\,e^{t\mathbf{A}} = \mathbf{A}e^{t\mathbf{A}} = e^{t\mathbf{A}}\mathbf{A}.$$

(c) $e^{\mathbf{A}}$ is invertible, with $(e^{\mathbf{A}})^{-1} = e^{-\mathbf{A}}$.

(d) If $\mathbf{A}\mathbf{B} = \mathbf{B}\mathbf{A}$, then $e^{\mathbf{A}+\mathbf{B}} = e^{\mathbf{A}}e^{\mathbf{B}}$.

(e) $\det e^{\mathbf{A}} = e^{\operatorname{tr}\mathbf{A}}$.

Idea. Part (b) is termwise differentiation, entry by entry. Everything after it is the same two-line trick: to prove $\mathbf{X}(t) = \mathbf{Y}(t)$ for two matrix functions, show that $\mathbf{Z}(t) = e^{-t\mathbf{C}}\mathbf{X}(t)$ has derivative $\mathbf{O}$ for a suitable $\mathbf{C}$, so $\mathbf{Z}$ is constant and equals $\mathbf{Z}(0)$. That is why (c) and (d) come after (b), even though the series would also give them. Part (e) is read off the Jordan form: $e^{\mathbf{J}}$ is triangular with e^{λ_i} on the diagonal.

Proof.

(a) $\mathbf{O}^m = \mathbf{O}$ for $m \geq 1$ and $\mathbf{O}^0 = \mathbf{I}$, so the series is $\mathbf{I}$. For the second claim, Step 3 of the previous proof applies verbatim with $\mathbf{J}$ replaced by $\mathbf{P}^{-1}\mathbf{A}\mathbf{P}$: the partial sums of $\sum_m (\mathbf{P}^{-1}\mathbf{A}\mathbf{P})^m/m!$ are $\mathbf{P}^{-1}(\sum_m \mathbf{A}^m/m!)\mathbf{P}$, and (A3) lets the limit pass through.

(b) Fix a position (r, s). By Theorem 10.9.3 the series $\sum_m t^m(\mathbf{A}^m)_{rs}/m!$ converges for every real t, since $(t\mathbf{A})^m = t^m\mathbf{A}^m$. So (A2) applies with $c_m = (\mathbf{A}^m)_{rs}/m!$ and gives

$$\begin{aligned}\frac{d}{dt}\left(e^{t\mathbf{A}}\right)_{rs} &= \sum_{m\geq 1}\frac{m\,t^{m-1}}{m!}(\mathbf{A}^m)_{rs} = \sum_{m\geq 1}\frac{t^{m-1}}{(m-1)!}(\mathbf{A}\cdot\mathbf{A}^{m-1})_{rs}\\ &= \Big(\mathbf{A}\sum_{l\geq 0}\frac{t^l\mathbf{A}^l}{l!}\Big)_{rs},\end{aligned}$$

where the last step moves the constant matrix $\mathbf{A}$ outside the limit: each entry of $\mathbf{A}(\sum_{l\leq L} t^l\mathbf{A}^l/l!)$ is a fixed linear combination of entries of the partial sum, so (A3) applies. Hence $\frac{d}{dt}e^{t\mathbf{A}} = \mathbf{A}e^{t\mathbf{A}}$. The same computation with $\mathbf{A}^m = \mathbf{A}^{m-1}\mathbf{A}$ gives $e^{t\mathbf{A}}\mathbf{A}$.

(c) Put $\mathbf{H}(t) = e^{t\mathbf{A}}e^{-t\mathbf{A}}$ for $t \in \mathbb{R}$. Each entry of $\mathbf{H}$ is a finite sum of products of differentiable functions, so by (A3) and (b),

$$\mathbf{H}'(t) = (\mathbf{A}e^{t\mathbf{A}})e^{-t\mathbf{A}} + e^{t\mathbf{A}}(-\mathbf{A}e^{-t\mathbf{A}}) = \mathbf{A}e^{t\mathbf{A}}e^{-t\mathbf{A}} - \mathbf{A}e^{t\mathbf{A}}e^{-t\mathbf{A}} = \mathbf{O},$$

where the second equality uses $e^{t\mathbf{A}}\mathbf{A} = \mathbf{A}e^{t\mathbf{A}}$ from (b). By (A4) every entry of $\mathbf{H}$ is constant, so $\mathbf{H}(t) = \mathbf{H}(0) = e^{\mathbf{O}}e^{\mathbf{O}} = \mathbf{I}$ for all t. At $t = 1$, $e^{\mathbf{A}}e^{-\mathbf{A}} = \mathbf{I}$, and Theorem 3.4.5 finishes it.

(d) Suppose $\mathbf{A}\mathbf{B} = \mathbf{B}\mathbf{A}$. Then $\mathbf{B}$ commutes with every partial sum of $\sum_m t^m\mathbf{A}^m/m!$, so by (A3) it commutes with the limit: $\mathbf{B}e^{t\mathbf{A}} = e^{t\mathbf{A}}\mathbf{B}$ for every t. Put

$$\mathbf{G}(t) = e^{-t(\mathbf{A}+\mathbf{B})}e^{t\mathbf{A}}e^{t\mathbf{B}}.$$

Differentiating entrywise with the product rule (A3) and (b),

$$\mathbf{G}'(t) = -(\mathbf{A}+\mathbf{B})e^{-t(\mathbf{A}+\mathbf{B})}e^{t\mathbf{A}}e^{t\mathbf{B}} + e^{-t(\mathbf{A}+\mathbf{B})}\mathbf{A}e^{t\mathbf{A}}e^{t\mathbf{B}} + e^{-t(\mathbf{A}+\mathbf{B})}e^{t\mathbf{A}}\mathbf{B}e^{t\mathbf{B}}.$$

In the third term replace $e^{t\mathbf{A}}\mathbf{B}$ by $\mathbf{B}e^{t\mathbf{A}}$; in the first, use $(\mathbf{A}+\mathbf{B})e^{-t(\mathbf{A}+\mathbf{B})} = e^{-t(\mathbf{A}+\mathbf{B})}(\mathbf{A}+\mathbf{B})$ from (b). The three terms become

$$e^{-t(\mathbf{A}+\mathbf{B})}(-(\mathbf{A}+\mathbf{B}) + \mathbf{A} + \mathbf{B})e^{t\mathbf{A}}e^{t\mathbf{B}} = \mathbf{O}.$$

By (A4), $\mathbf{G}$ is constant, so $\mathbf{G}(t) = \mathbf{G}(0) = \mathbf{I}$. Multiplying on the left by $e^{t(\mathbf{A}+\mathbf{B})}$ and using (c), $e^{t\mathbf{A}}e^{t\mathbf{B}} = e^{t(\mathbf{A}+\mathbf{B})}$; put $t = 1$.

(e) Write $\mathbf{A} = \mathbf{P}\mathbf{J}\mathbf{P}^{-1}$ with $\mathbf{J} = \bigoplus_i \mathbf{J}_{k_i}(\lambda_i)$ a Jordan matrix. By (a) and Theorem 10.9.3,

$$\det e^{\mathbf{A}} = \det(\mathbf{P} e^{\mathbf{J}} \mathbf{P}^{-1}) = \det e^{\mathbf{J}}$$

(Corollary 7.5.7), and $e^{\mathbf{J}} = \bigoplus_i e^{\mathbf{J}_{k_i}(\lambda_i)}$ is upper triangular with e^{λ_i} repeated k_i times along the diagonal. By Theorem 7.4.6 and (A1),

$$\det e^{\mathbf{J}} = \prod_i (e^{\lambda_i})^{k_i} = e^{\sum_i k_i \lambda_i} = e^{\operatorname{tr} \mathbf{J}} = e^{\operatorname{tr} \mathbf{A}},$$

the last equality because similar matrices have equal traces (Theorem 4.2.5). This proves the theorem.

■

Part (e) says something striking: $\det e^{\mathbf{A}} \neq 0$ always, which re-proves (c), and if $\operatorname{tr} \mathbf{A} = 0$ then $e^{\mathbf{A}} \in \mathrm{SL}_n(\mathbb{C})$.

The following corollary is the formula actually used in computations. Note that $t\mathbf{J}_k(\lambda)$ is **not** a Jordan block for $k \geq 2$ and $t \neq 1$, since its superdiagonal entries are t; the commuting law does the work instead.

Corollary 10.9.5. Exponential of a Jordan Block, with a Parameter

Let $k \geq 1$, $\lambda \in \mathbb{C}$, and $t \in \mathbb{R}$. Then $e^{t\mathbf{J}_k(\lambda)}$ is the upper triangular matrix with

$$\left(e^{t\mathbf{J}_k(\lambda)}\right)_{r,\,r+j} = \frac{t^j e^{\lambda t}}{j!} \qquad (0 \leq j \leq k-1)$$

and zeros below the diagonal. Consequently, if $\mathbf{A} = \mathbf{P}\mathbf{J}\mathbf{P}^{-1}$ with $\mathbf{J} = \mathbf{J}_{k_1}(\lambda_1) \oplus \cdots \oplus \mathbf{J}_{k_m}(\lambda_m)$, then

$$e^{t\mathbf{A}} = \mathbf{P}\left(e^{t\mathbf{J}_{k_1}(\lambda_1)} \oplus \cdots \oplus e^{t\mathbf{J}_{k_m}(\lambda_m)}\right)\mathbf{P}^{-1}.$$

In particular, for a diagonal $\mathbf{D} = \operatorname{diag}(d_1, \ldots, d_n)$, $e^{t\mathbf{D}} = \operatorname{diag}(e^{td_1}, \ldots, e^{td_n})$.

Proof. Write $t\mathbf{J}_k(\lambda) = (t\lambda)\mathbf{I}_k + t\mathbf{M}$ with $\mathbf{M} = \mathbf{J}_k(0)$. The two summands commute, since $\mathbf{I}_k$ commutes with everything, so Theorem 10.9.4 (d) gives $e^{t\mathbf{J}_k(\lambda)} = e^{(t\lambda)\mathbf{I}_k} e^{t\mathbf{M}}$. For the first factor, $\left((t\lambda)\mathbf{I}_k\right)^m = (t\lambda)^m \mathbf{I}_k$, so the partial sums are $\left(\sum_{m \leq \mathbf{M}} (t\lambda)^m/m!\right)\mathbf{I}_k$ and $e^{(t\lambda)\mathbf{I}_k} = e^{\lambda t}\mathbf{I}_k$ by (A1). For the second, $\mathbf{M}^j = \mathbf{O}$ for $j \geq k$, so the series terminates:

$$e^{t\mathbf{M}} = \sum_{j=0}^{k-1} \frac{t^j}{j!} \mathbf{M}^j,$$

and $\mathbf{M}^j$ has 1 in each position $(r, r+j)$, as computed in the proof of Lemma 10.9.1. Multiplying by the scalar $e^{\lambda t}$ gives the entry formula. The second display follows from $t\mathbf{A} = \mathbf{P}(t\mathbf{J})\mathbf{P}^{-1}$ with Theorem 10.9.4 (a) and Theorem 8.1.4 (b), exactly as in Step 2 of the proof of Theorem 10.9.3. The diagonal case is the case in which every block has size 1.

■

Warning. $e^{\mathbf{A}+\mathbf{B}} = e^{\mathbf{A}}e^{\mathbf{B}}$ **fails when $\mathbf{A}$ and $\mathbf{B}$ do not commute.** Take

$$\mathbf{A} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \qquad \mathbf{B} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}.$$

Both square to $\mathbf{O}$, so the series stop after two terms: $e^{\mathbf{A}} = \mathbf{I} + \mathbf{A} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ and $e^{\mathbf{B}} = \mathbf{I} + \mathbf{B} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$. Then

$$e^{\mathbf{A}}e^{\mathbf{B}} = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix} \neq \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix} = e^{\mathbf{B}}e^{\mathbf{A}}.$$

Since $\mathbf{A}+\mathbf{B} = \mathbf{B}+\mathbf{A}$, at most one of the two products can equal $e^{\mathbf{A}+\mathbf{B}}$, so the identity must fail for one of the pairs $(\mathbf{A}, \mathbf{B})$, $(\mathbf{B}, \mathbf{A})$. (In fact it fails for both: $(\mathbf{A}+\mathbf{B})^2 = \mathbf{I}$, so $e^{\mathbf{A}+\mathbf{B}}$ has all four entries nonzero and none of them equal to 2.) The moral: check $\mathbf{AB} = \mathbf{BA}$ before splitting an exponential.

10.9.4 Linear systems of differential equations

A system of n linear differential equations with constant coefficients,

$$x_1' = a_{11}x_1 + \cdots + a_{1n}x_n, \quad \ldots, \quad x_n' = a_{n1}x_1 + \cdots + a_{nn}x_n,$$

is the single equation $\mathbf{x}' = \mathbf{A}\mathbf{x}$ for a vector-valued function $\mathbf{x}\colon \mathbb{R} \to \mathbb{C}^n$, differentiated entrywise. In one dimension the answer is $x(t) = e^{ta}x_0$; the matrix exponential makes the general answer look exactly the same.

Theorem 10.9.6. Solution of a Linear System

Let $\mathbf{A} \in M_n(\mathbb{C})$ and $\mathbf{x}_0 \in \mathbb{C}^n$. Then the function

$$\mathbf{x}(t) = e^{t\mathbf{A}}\mathbf{x}_0 \qquad (t \in \mathbb{R})$$

is differentiable and satisfies $\mathbf{x}'(t) = \mathbf{A}\mathbf{x}(t)$ for all t and $\mathbf{x}(0) = \mathbf{x}_0$; and it is the **only** differentiable function $\mathbb{R} \to \mathbb{C}^n$ with these two properties.

Idea. Existence is Theorem 10.9.4 (b) applied to a fixed vector. For uniqueness, take any solution $\mathbf{y}$ and multiply it by $e^{-t\mathbf{A}}$, the one factor that can undo the growth: the product has derivative $\mathbf{0}$, hence is constant. This is the same move as in the proofs of parts (c) and (d) above.

Proof. Existence. Each entry of $\mathbf{x}(t) = e^{t\mathbf{A}}\mathbf{x}_0$ is a linear combination of entries of $e^{t\mathbf{A}}$ with constant coefficients, so $\mathbf{x}$ is differentiable with $\mathbf{x}'(t) = (\frac{d}{dt}e^{t\mathbf{A}})\mathbf{x}_0$ by the linearity of differentiation (A3). By Theorem 10.9.4 (b), $\mathbf{x}'(t) = \mathbf{A}e^{t\mathbf{A}}\mathbf{x}_0 = \mathbf{A}\mathbf{x}(t)$. And $\mathbf{x}(0) = e^{\mathbf{O}}\mathbf{x}_0 = \mathbf{x}_0$ by Theorem 10.9.4 (a).

Uniqueness. Let $\mathbf{y}\colon \mathbb{R} \to \mathbb{C}^n$ be differentiable with $\mathbf{y}' = \mathbf{A}\mathbf{y}$ and $\mathbf{y}(0) = \mathbf{x}_0$. Put $\mathbf{z}(t) =$

$e^{-t\mathbf{A}}\mathbf{y}(t)$. By the product rule (A3) and Theorem 10.9.4 (b) applied to $-\mathbf{A}$,

$$\mathbf{z}'(t) = (-\mathbf{A}e^{-t\mathbf{A}})\mathbf{y}(t) + e^{-t\mathbf{A}}\mathbf{y}'(t) = -\mathbf{A}e^{-t\mathbf{A}}\mathbf{y}(t) + e^{-t\mathbf{A}}\mathbf{A}\mathbf{y}(t) = \mathbf{0},$$

where the last step uses $e^{-t\mathbf{A}}\mathbf{A} = \mathbf{A}e^{-t\mathbf{A}}$ from Theorem 10.9.4 (b). By (A4) each entry of $\mathbf{z}$ is constant, so $\mathbf{z}(t) = \mathbf{z}(0) = \mathbf{x}_0$ for all t. Multiplying by $e^{t\mathbf{A}}$ and using Theorem 10.9.4 (c),

$$\mathbf{y}(t) = e^{t\mathbf{A}}\mathbf{z}(t) = e^{t\mathbf{A}}\mathbf{x}_0 = \mathbf{x}(t).$$

This proves the theorem.

■

The formula is only as useful as our ability to compute $e^{t\mathbf{A}}$, and the Jordan form does that. Two shapes of answer occur, and they are worth seeing side by side.

Example. Two Two-Dimensional Systems

Solve $\mathbf{x}' = \mathbf{A}\mathbf{x}$ with $\mathbf{x}(0) = (1, 0)$ for

(a) $\mathbf{A} = \begin{pmatrix} 1 & -2 \\ 2 & 1 \end{pmatrix}$ (complex eigenvalues);

(b) $\mathbf{A} = \begin{pmatrix} 2 & 1 \\ 0 & 2 \end{pmatrix}$ (one eigenvalue, one block of size 2).

Solution.

(a) $p_{\mathbf{A}} = (x-1)^2+4$, with roots $\lambda = 1 \pm 2i$, distinct, so $\mathbf{A}$ is diagonalizable over $\mathbb{C}$ and every block has size 1. Solving $(\mathbf{A} - (1+2i)\mathbf{I})\mathbf{x} = \mathbf{0}$: the first row reads $-2ix_1 - 2x_2 = 0$, so $x_2 = -ix_1$ and $(1, -i)$ is an eigenvector; $(1, i)$ is one for $1 - 2i$. With $\mathbf{P} = \begin{pmatrix} 1 & 1 \\ -i & i \end{pmatrix}$ we get $\det \mathbf{P} = 2i$ and $\mathbf{P}^{-1} = \frac{1}{2i}\begin{pmatrix} i & -1 \\ i & 1 \end{pmatrix} = \frac{1}{2}\begin{pmatrix} 1 & i \\ 1 & -i \end{pmatrix}$, so $\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \operatorname{diag}(1+2i, 1-2i)$. By Corollary 10.9.5,

$$\begin{aligned} e^{t\mathbf{A}} &= \mathbf{P}\operatorname{diag}(e^{(1+2i)t}, e^{(1-2i)t})\mathbf{P}^{-1} \\ &= \tfrac{1}{2}\begin{pmatrix} e^{(1+2i)t} + e^{(1-2i)t} & i(e^{(1+2i)t} - e^{(1-2i)t}) \\ -i(e^{(1+2i)t} - e^{(1-2i)t}) & e^{(1+2i)t} + e^{(1-2i)t} \end{pmatrix}. \end{aligned}$$

By (A1), $e^{(1\pm 2i)t} = e^t(\cos 2t \pm i \sin 2t)$, so $e^{(1+2i)t} + e^{(1-2i)t} = 2e^t \cos 2t$ and $e^{(1+2i)t} - e^{(1-2i)t} = 2ie^t \sin 2t$. Substituting, and using $i \cdot 2i = -2$,

$$e^{t\mathbf{A}} = e^t\begin{pmatrix} \cos 2t & -\sin 2t \\ \sin 2t & \cos 2t \end{pmatrix}, \qquad \mathbf{x}(t) = e^{t\mathbf{A}}\begin{pmatrix} 1 \\ 0 \end{pmatrix} = e^t\begin{pmatrix} \cos 2t \\ \sin 2t \end{pmatrix}.$$

The solution spirals outwards: rotation at angular speed 2, growth at rate 1. The entries are real, as they must be for a real $\mathbf{A}$ and a real $\mathbf{x}_0$.
Check. $\mathbf{x}'(t) = e^t(\cos 2t - 2\sin 2t,\ \sin 2t + 2\cos 2t)$, and $\mathbf{A}\mathbf{x}(t) = e^t(\cos 2t - 2\sin 2t,\ 2\cos 2t + \sin 2t)$. They agree, and $\mathbf{x}(0) = (1, 0)$.

(b) $\mathbf{A} = \mathbf{J}_2(2)$ is its own Jordan form. Writing $\mathbf{A} = 2\mathbf{I} + \mathbf{M}$ with $\mathbf{M} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$, the two summands commute and $\mathbf{M}^2 = \mathbf{O}$, so Corollary 10.9.5 with $k = 2$ gives

$$e^{t\mathbf{A}} = e^{2t}(\mathbf{I} + t\mathbf{M}) = e^{2t}\begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix}.$$

So $\mathbf{x}(t) = e^{2t}(1, 0)$. A different initial vector shows the new phenomenon: $\mathbf{x}(0) = (0, 1)$ gives $\mathbf{x}(t) = e^{2t}(t, 1)$, whose first entry is te^{2t}. **A repeated eigenvalue with a block of size k produces solutions containing $t^j e^{\lambda t}$ for $j < k$.** This is the same $t^j\lambda^m$ pattern that Chapter 9 met for recurrences with repeated roots, now in continuous time.

A scalar equation of higher order is a system in disguise, and the translation is the companion matrix of Section 5.

Example. A Second-Order Equation with a Repeated Root

Solve $y'' - 4y' + 4y = 0$ with $y(0) = 1$ and $y'(0) = 3$.

Solution. Make it a system. Put $\mathbf{u} = (y, y')$. Then $\mathbf{u}' = (y', y'') = (y', 4y' - 4y)$, so $\mathbf{u}' = \mathbf{M}\mathbf{u}$ with

$$\mathbf{M} = \begin{pmatrix} 0 & 1 \\ -4 & 4 \end{pmatrix}, \qquad \mathbf{u}(0) = (1, 3).$$

Note that $p_{\mathbf{M}} = x(x - 4) + 4 = x^2 - 4x + 4$ is the polynomial of the equation, and that $\mathbf{M} = \mathbf{C}(p_{\mathbf{M}})^{\top}$ for the companion matrix $\mathbf{C}(p_{\mathbf{M}}) = \begin{pmatrix} 0 & -4 \\ 1 & 4 \end{pmatrix}$ of Definition 10.5.3; by Corollary 10.7.7 the two are similar, so which of them one uses is immaterial.

Exponentiate. $p_{\mathbf{M}} = (x-2)^2$. Put $\mathbf{N} = \mathbf{M} - 2\mathbf{I} = \begin{pmatrix} -2 & 1 \\ -4 & 2 \end{pmatrix}$; then $\mathbf{N}^2 = \begin{pmatrix} 4-4 & -2+2 \\ 8-8 & -4+4 \end{pmatrix} = \mathbf{O}$, so $2\mathbf{I}$ and $\mathbf{N}$ commute and Theorem 10.9.4 (d) gives, as in the previous example,

$$e^{t\mathbf{M}} = e^{2t}(\mathbf{I} + t\mathbf{N}) = e^{2t}\begin{pmatrix} 1 - 2t & t \\ -4t & 1 + 2t \end{pmatrix}.$$

(No basis change was needed: $\mathbf{M}$ is a scalar plus a nilpotent matrix on the nose.)

Read off the answer.

$$\mathbf{u}(t) = e^{t\mathbf{M}}\begin{pmatrix} 1 \\ 3 \end{pmatrix} = e^{2t}\begin{pmatrix} (1 - 2t) + 3t \\ -4t + 3(1 + 2t) \end{pmatrix} = e^{2t}\begin{pmatrix} 1 + t \\ 3 + 2t \end{pmatrix},$$

so $y(t) = (1 + t)e^{2t}$.

Check. $y' = e^{2t} + 2(1 + t)e^{2t} = (3 + 2t)e^{2t}$, which is the second entry of $\mathbf{u}$, as it should be, and $y'(0) = 3$. Then $y'' = 2e^{2t} + 2(3 + 2t)e^{2t} = (8 + 4t)e^{2t}$, and

$$y'' - 4y' + 4y = \big[(8 + 4t) - 4(3 + 2t) + 4(1 + t)\big]e^{2t} = 0.$$

The repeated root 2 produces the term te^{2t}, exactly as the size-2 block predicts.

Chapter 16 will return to $e^{\mathbf{A}}$ with norms available, which give quantitative bounds instead of exact formulas. The next section stays with exact formulas and asks the discrete-time question: what happens to $\mathbf{A}^m$?

10.9.5 Exercises

10.9.5.1 A. Check your understanding

Exercise. A1

(a) State the formula for $p(\mathbf{J}_k(\lambda))$ and say where the factorials come from.

(b) Define $f(\mathbf{A})$ for $\mathbf{A} \in M_n(\mathbb{C})$ with a given Jordan form, and list the two choices whose irrelevance has to be checked.

(c) Which values of f and its derivatives does $f(\mathbf{A})$ depend on? How many numbers is that?

(d) State the five properties of $e^{\mathbf{A}}$ proved in Theorem 10.9.4, with their hypotheses.

(e) True or false: $e^{\mathbf{A}+\mathbf{B}} = e^{\mathbf{A}}e^{\mathbf{B}}$ for all $\mathbf{A}, \mathbf{B} \in M_n(\mathbb{C})$. Justify your answer.

(f) True or false: $e^{\mathbf{A}}$ can be singular. Justify your answer.

Solution.

(a) $p(\mathbf{J}_k(\lambda))$ is upper triangular with $p^{(j)}(\lambda)/j!$ in every position $(r, r+j)$ (Lemma 10.9.1). The factorials come from the Taylor expansion $p = \sum_j \frac{p^{(j)}(\lambda)}{j!}(x-\lambda)^j$ (Theorem 6.4.7), whose j-th term contributes $\mathbf{J}_k(0)^j$, the matrix with ones on the j-th superdiagonal.

(b) $f(\mathbf{A}) = \mathbf{P}(\bigoplus_i f(\mathbf{J}_{k_i}(\lambda_i)))\mathbf{P}^{-1}$ with $f(\mathbf{J}_k(\lambda))$ as in (a) (Definition 10.9.1). The choices are the matrix $\mathbf{P}$ and the order in which the blocks are listed; both are irrelevant because $f(\mathbf{A}) = p(\mathbf{A})$ for the Hermite interpolating polynomial p (Theorem 10.9.2).

(c) On $f^{(j)}(\mu)$ for each eigenvalue μ and each $j < s_\mu$, where s_μ is the exponent of $x - \mu$ in $m_{\mathbf{A}}$, that is, the largest μ-block size. The total number is $\sum_\mu s_\mu = \deg m_{\mathbf{A}}$ (Theorem 10.9.2).

(d) $e^{\mathbf{P}^{-1}\mathbf{A}\mathbf{P}} = \mathbf{P}^{-1}e^{\mathbf{A}}\mathbf{P}$ for invertible $\mathbf{P}$; $\frac{d}{dt}e^{t\mathbf{A}} = \mathbf{A}e^{t\mathbf{A}} = e^{t\mathbf{A}}\mathbf{A}$; $(e^{\mathbf{A}})^{-1} = e^{-\mathbf{A}}$; $e^{\mathbf{A}+\mathbf{B}} = e^{\mathbf{A}}e^{\mathbf{B}}$ **provided** $\mathbf{AB} = \mathbf{BA}$; and $\det e^{\mathbf{A}} = e^{\operatorname{tr}\mathbf{A}}$ with no hypothesis.

(e) False. With $\mathbf{A} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$ and $\mathbf{B} = \mathbf{A}^\top$, the matrices $e^{\mathbf{A}}e^{\mathbf{B}}$ and $e^{\mathbf{B}}e^{\mathbf{A}}$ differ, so the identity cannot hold for both orders; see the warning after Theorem 10.9.4.

(f) False. $\det e^{\mathbf{A}} = e^{\operatorname{tr}\mathbf{A}} \neq 0$ by Theorem 10.9.4 (e) and (A1), so $e^{\mathbf{A}}$ is always invertible (Theorem 7.5.2).

10.9.5.2 B. Practice

Exercise. B1: Computing exponentials

Compute $e^{t\mathbf{A}}$ for each of the following.

(a) $\mathbf{A} = \begin{pmatrix} 1 & 3 \\ 3 & 1 \end{pmatrix}$.

(b) $\mathbf{A} = \begin{pmatrix} 5 & 1 \\ 0 & 5 \end{pmatrix}$.

(c) $\mathbf{A} = \begin{pmatrix} 3 & 1 & 0 \\ 0 & 3 & 1 \\ 0 & 0 & 3 \end{pmatrix} = \mathbf{J}_3(3)$.

Solution.

(a) $p_{\mathbf{A}} = (x-1)^2 - 9$, with roots 4 and -2. Eigenvectors: $(\mathbf{A} - 4\mathbf{I})\mathbf{x} = \mathbf{0}$ gives $\operatorname{span}((1,1))$; $(\mathbf{A} + 2\mathbf{I})\mathbf{x} = \mathbf{0}$ gives $\operatorname{span}((1,-1))$. With $\mathbf{P} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$ we have $\det \mathbf{P} = -2$, $\mathbf{P}^{-1} = \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$, and $\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \operatorname{diag}(4,-2)$, so by Corollary 10.9.5

$$e^{t\mathbf{A}} = \mathbf{P}\operatorname{diag}(e^{4t}, e^{-2t})\mathbf{P}^{-1} = \frac{1}{2}\begin{pmatrix} e^{4t} + e^{-2t} & e^{4t} - e^{-2t} \\ e^{4t} - e^{-2t} & e^{4t} + e^{-2t} \end{pmatrix}.$$

Check at $t = 0$: both diagonal entries give 1 and both off-diagonal ones 0, so $e^{\mathbf{O}} = \mathbf{I}$.

(b) $\mathbf{A} = \mathbf{J}_2(5)$, so Corollary 10.9.5 with $k = 2$ gives

$$e^{t\mathbf{A}} = e^{5t}(\mathbf{I} + t\mathbf{M}) = e^{5t}\begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix}.$$

(c) By Corollary 10.9.5 with $k = 3$ and $\lambda = 3$, or directly from $\mathbf{A} = 3\mathbf{I} + \mathbf{M}$ with $\mathbf{M} = \mathbf{J}_3(0)$ and $\mathbf{M}^3 = \mathbf{O}$, giving $e^{t\mathbf{A}} = e^{3t}(\mathbf{I} + t\mathbf{M} + \frac{t^2}{2}\mathbf{M}^2)$:

$$e^{t\mathbf{A}} = e^{3t}\begin{pmatrix} 1 & t & \frac{t^2}{2} \\ 0 & 1 & t \\ 0 & 0 & 1 \end{pmatrix}.$$

This is Definition 10.9.1 for $f(x) = e^{tx}$: the j-th superdiagonal carries $f^{(j)}(3)/j! = t^j e^{3t}/j!$.

Exercise. B2: Solving systems

Solve each initial value problem, and verify your answer by differentiating.

(a) $x' = x + 3y$, $y' = 3x + y$, with $x(0) = 2$, $y(0) = 0$.

(b) $\mathbf{x}' = \begin{pmatrix} 0 & -3 \\ 3 & 0 \end{pmatrix}\mathbf{x}$ with $\mathbf{x}(0) = (0,1)$.

(c) $y''' = 0$ with $y(0) = 1$, $y'(0) = 0$, $y''(0) = 2$, by turning it into a system.

Solution.

(a) This is $\mathbf{x}' = \mathbf{A}\mathbf{x}$ for the matrix of Exercise (B1: Computing exponentials) (a). By Theorem 10.9.6,

$$\mathbf{x}(t) = e^{t\mathbf{A}}\begin{pmatrix}2\\0\end{pmatrix} = \begin{pmatrix}e^{4t}+e^{-2t}\\e^{4t}-e^{-2t}\end{pmatrix},$$

that is, $x = e^{4t}+e^{-2t}$ and $y = e^{4t}-e^{-2t}$. *Check.* $x' = 4e^{4t}-2e^{-2t}$ and $x+3y = e^{4t}+e^{-2t}+3e^{4t}-3e^{-2t} = 4e^{4t}-2e^{-2t}$; similarly for y'. And $x(0)=2$, $y(0)=0$.

(b) Write $\mathbf{A} = 3\mathbf{R}$ with $\mathbf{R} = \begin{pmatrix}0&-1\\1&0\end{pmatrix}$, whose eigenvalues are $\pm i$ (Example (A Real Matrix with Non-Real Eigenvalues)), so those of $\mathbf{A}$ are $\pm 3i$, distinct. Repeating the computation of Example (Two Two-Dimensional Systems) (a) with $1\pm 2i$ replaced by $\pm 3i$ — eigenvectors $(1,-i)$ for $3i$ and $(1,i)$ for $-3i$ — gives

$$e^{t\mathbf{A}} = \begin{pmatrix}\cos 3t & -\sin 3t\\ \sin 3t & \cos 3t\end{pmatrix}, \qquad \mathbf{x}(t) = \begin{pmatrix}-\sin 3t\\ \cos 3t\end{pmatrix}.$$

Check. $\mathbf{x}'(t) = (-3\cos 3t, -3\sin 3t)$ and $\mathbf{A}\mathbf{x}(t) = (-3\cos 3t, -3\sin 3t)$; and $\mathbf{x}(0) = (0,1)$. The solution runs around the unit circle: no growth, since the eigenvalues are purely imaginary.

(c) Put $\mathbf{u} = (y, y', y'')$. Then $\mathbf{u}' = (y', y'', 0) = \mathbf{M}\mathbf{u}$ with $\mathbf{M} = \mathbf{J}_3(0)$, and $\mathbf{u}(0) = (1,0,2)$. By Corollary 10.9.5 with $\lambda = 0$ and $k = 3$,

$$e^{t\mathbf{M}} = \begin{pmatrix}1 & t & t^2/2\\ 0 & 1 & t\\ 0 & 0 & 1\end{pmatrix}, \qquad \mathbf{u}(t) = e^{t\mathbf{M}}\begin{pmatrix}1\\0\\2\end{pmatrix} = \begin{pmatrix}1+t^2\\2t\\2\end{pmatrix}.$$

So $y(t) = 1+t^2$. *Check.* $y' = 2t$, $y'' = 2$, $y''' = 0$, and the three initial values are $1, 0, 2$. The eigenvalue 0 with one block of size 3 gives polynomial solutions $t^j e^{0t} = t^j$ for $j < 3$.

Exercise. B3: A square root and a logarithm

(a) Let $\mathbf{A} = \begin{pmatrix}2&1\\1&2\end{pmatrix}$. Compute $f(\mathbf{A})$ for $f(x) = \sqrt{x}$ (positive branch) and verify that $f(\mathbf{A})^2 = \mathbf{A}$.

(b) Let $\mathbf{B} = \begin{pmatrix}1&2\\0&1\end{pmatrix}$. Compute $f(\mathbf{B})$ for $f(x) = \log x$ with $\log 1 = 0$, and verify that $e^{f(\mathbf{B})} = \mathbf{B}$.

(c) Explain why the same method produces **no** square root of $\begin{pmatrix}0&1\\0&0\end{pmatrix}$.

Solution.

(a) $p_{\mathbf{A}} = (x-2)^2 - 1 = (x-1)(x-3)$, eigenvalues $1, 3$. Eigenvectors $(1,-1)$ for 1 and

$(1,1)$ for 3. With $\mathbf{P} = \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}$ and $\mathbf{P}^{-1} = \frac{1}{2}\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}$,

$$f(\mathbf{A}) = \mathbf{P}\,\mathrm{diag}(1,\sqrt{3})\mathbf{P}^{-1} = \frac{1}{2}\begin{pmatrix} 1+\sqrt{3} & -1+\sqrt{3} \\ -1+\sqrt{3} & 1+\sqrt{3} \end{pmatrix}.$$

Squaring, the diagonal entry is $\frac{1}{4}[(1+\sqrt{3})^2 + (\sqrt{3}-1)^2] = \frac{1}{4}(4+2\sqrt{3}+4-2\sqrt{3}) = 2$, and the off-diagonal entry is $\frac{1}{4}\cdot 2(1+\sqrt{3})(\sqrt{3}-1) = \frac{1}{2}(3-1) = 1$. So $f(\mathbf{A})^2 = \mathbf{A}$.

(b) Here $m_{\mathbf{B}} = (x-1)^2$, since $(\mathbf{B}-\mathbf{I})^2 = \mathbf{O} \neq \mathbf{B}-\mathbf{I}$, so by Theorem 10.9.2 the answer is $p(\mathbf{B})$ for the polynomial of degree < 2 with $p(1) = f(1) = 0$ and $p'(1) = f'(1) = 1$, namely $p = x - 1$. Hence

$$f(\mathbf{B}) = \mathbf{B} - \mathbf{I} = \begin{pmatrix} 0 & 2 \\ 0 & 0 \end{pmatrix}.$$

Its square is $\mathbf{O}$, so $e^{f(\mathbf{B})} = \mathbf{I} + f(\mathbf{B}) = \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix} = \mathbf{B}$. The Jordan route gives the same thing: with $\mathbf{Q} = \mathrm{diag}(2,1)$ we get $\mathbf{Q}^{-1}\mathbf{B}\mathbf{Q} = \mathbf{J}_2(1)$, and $f(\mathbf{J}_2(1)) = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$, so $f(\mathbf{B}) = \mathbf{Q} f(\mathbf{J}_2(1))\mathbf{Q}^{-1} = \begin{pmatrix} 0 & 2 \\ 0 & 0 \end{pmatrix}$.

(c) The matrix is $\mathbf{J}_2(0)$, with the single eigenvalue 0 and a block of size 2. Definition 10.9.1 needs f and f' at 0, and $f(x) = \sqrt{x}$ is not differentiable at 0 — the entry $f'(0)$ does not exist. In fact $\mathbf{J}_2(0)$ has no square root at all: if $\mathbf{X}^2 = \mathbf{J}_2(0)$, then $\mathbf{X}^4 = \mathbf{J}_2(0)^2 = \mathbf{O}$, so $\mathbf{X}$ is nilpotent, so $\mathbf{X}^2 = \mathbf{O}$ by Proposition 10.2.1 (a) with $n = 2$, contradicting $\mathbf{X}^2 = \mathbf{J}_2(0) \neq \mathbf{O}$.

10.9.5.3 C. Going deeper

Exercise. C1: How badly the exponential law fails

Let $\mathbf{A} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$ and $\mathbf{B} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}$ in $M_2(\mathbb{R})$.

(a) Compute $e^{\mathbf{A}}$, $e^{\mathbf{B}}$, $e^{\mathbf{A}}e^{\mathbf{B}}$, and $e^{\mathbf{B}}e^{\mathbf{A}}$.

(b) Compute $e^{\mathbf{A}+\mathbf{B}}$ in closed form. *Hint:* $(\mathbf{A}+\mathbf{B})^2 = \mathbf{I}$, *so split the series into even and odd terms.*

(c) Verify $\det e^{\mathbf{A}+\mathbf{B}} = e^{\mathrm{tr}(\mathbf{A}+\mathbf{B})}$ and $\det(e^{\mathbf{A}}e^{\mathbf{B}}) = e^{\mathrm{tr}\,\mathbf{A}}e^{\mathrm{tr}\,\mathbf{B}}$, and explain why equal determinants do not rescue the identity.

(d) Show that $e^{\mathbf{A}}e^{\mathbf{B}} = e^{\mathbf{C}}$ for **some** $\mathbf{C} \in M_2(\mathbb{R})$, and find $\mathrm{tr}\,\mathbf{C}$.

Solution.

(a) $\mathbf{A}^2 = \mathbf{B}^2 = \mathbf{O}$, so $e^{\mathbf{A}} = \mathbf{I} + \mathbf{A} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ and $e^{\mathbf{B}} = \mathbf{I} + \mathbf{B} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$. Then

$$e^{\mathbf{A}}e^{\mathbf{B}} = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}, \qquad e^{\mathbf{B}}e^{\mathbf{A}} = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}.$$

(b) Let $\mathbf{K} = \mathbf{A} + \mathbf{B} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$. Then $\mathbf{K}^2 = \mathbf{I}$, so $\mathbf{K}^{2l} = \mathbf{I}$ and $\mathbf{K}^{2l+1} = \mathbf{K}$. Splitting the partial sums by parity, which is legitimate because each entry sequence is a sum of two convergent sequences (A3),

$$e^{\mathbf{K}} = \Big(\sum_{l \ge 0} \frac{1}{(2l)!}\Big)\mathbf{I} + \Big(\sum_{l \ge 0} \frac{1}{(2l+1)!}\Big)\mathbf{K} = \frac{e + e^{-1}}{2}\mathbf{I} + \frac{e - e^{-1}}{2}\mathbf{K} = \begin{pmatrix} \frac{e+e^{-1}}{2} & \frac{e-e^{-1}}{2} \\ \frac{e-e^{-1}}{2} & \frac{e+e^{-1}}{2} \end{pmatrix},$$

where the two scalar sums were identified by adding and subtracting the series for e^1 and e^{-1} term by term (A1). Numerically the diagonal entries are about 1.543, not 2, so $e^{\mathbf{A}+\mathbf{B}} \neq e^{\mathbf{A}}e^{\mathbf{B}}$.

(c) $\operatorname{tr}(\mathbf{A} + \mathbf{B}) = 0$ and $\det e^{\mathbf{A}+\mathbf{B}} = \big(\frac{e+e^{-1}}{2}\big)^2 - \big(\frac{e-e^{-1}}{2}\big)^2 = \frac{1}{4}\big[(e + e^{-1})^2 - (e - e^{-1})^2\big] = \frac{1}{4} \cdot 4 = 1 = e^0$. Also $\det(e^{\mathbf{A}}e^{\mathbf{B}}) = 2 - 1 = 1 = e^0e^0$. The determinants agree because both matrices are products of exponentials of trace-zero matrices; a determinant is one number and cannot distinguish two matrices, so it gives no information about the identity either way. The entries do.

(d) $\mathbf{X} = e^{\mathbf{A}}e^{\mathbf{B}} = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$ is symmetric with $\det \mathbf{X} = 1$ and $\operatorname{tr} \mathbf{X} = 3$, so $p_{\mathbf{X}} = x^2 - 3x + 1$, whose roots $\frac{3 \pm \sqrt{5}}{2}$ are real, distinct, and positive. So $\mathbf{X} = \mathbf{P}\operatorname{diag}(\alpha, \beta)\mathbf{P}^{-1}$ with $\alpha\beta = 1$, $\alpha, \beta > 0$. Put $\mathbf{C} = \mathbf{P}\operatorname{diag}(\log \alpha, \log \beta)\mathbf{P}^{-1}$, a real matrix because $\mathbf{P}$ can be taken real (the eigenvalues are real) and the logarithms are real. By Definition 10.9.1 applied to $f = \exp$, $e^{\mathbf{C}} = \mathbf{P}\operatorname{diag}(\alpha, \beta)\mathbf{P}^{-1} = \mathbf{X}$. Finally $\operatorname{tr} \mathbf{C} = \log \alpha + \log \beta = \log(\alpha\beta) = \log 1 = 0$, as Theorem 10.9.4 (e) requires, since $\det \mathbf{X} = 1$.

Exercise. C2: When does every solution die out?

Let $\mathbf{A} \in M_n(\mathbb{C})$. Prove that the following are equivalent:

(i) every solution of $\mathbf{x}' = \mathbf{A}\mathbf{x}$ satisfies $\mathbf{x}(t) \to \mathbf{0}$ entrywise as $t \to \infty$;

(ii) every eigenvalue λ of $\mathbf{A}$ has $\operatorname{Re} \lambda < 0$.

Hint: compute the entries of $e^{t\mathbf{J}_k(\lambda)}$ and use (A5). For the failing direction, an eigenvector gives a solution you can write down.

Solution.

(ii) $\Rightarrow$ (i). By Theorem 10.9.6 every solution is $\mathbf{x}(t) = e^{t\mathbf{A}}\mathbf{x}_0$, so it suffices to show $e^{t\mathbf{A}} \to \mathbf{O}$ entrywise. Write $\mathbf{A} = \mathbf{P}\mathbf{J}\mathbf{P}^{-1}$ with $\mathbf{J} = \bigoplus_i \mathbf{J}_{k_i}(\lambda_i)$ (Theorem 10.3.1). Each entry of $e^{t\mathbf{A}} = \mathbf{P}e^{t\mathbf{J}}\mathbf{P}^{-1}$ is a fixed linear combination of entries of $e^{t\mathbf{J}}$, so by (A3) it is enough that $e^{t\mathbf{J}} \to \mathbf{O}$. By Corollary 10.9.5, the nonzero entries of $e^{t\mathbf{J}_k(\lambda)}$ are

$$\frac{t^j e^{\lambda t}}{j!} \qquad (0 \le j \le k - 1).$$

Writing $\lambda = c + i\omega$ with $c = \operatorname{Re} \lambda < 0$, we get $|t^j e^{\lambda t}| = t^j e^{ct}|e^{i\omega t}| = t^j e^{ct}$ for $t \ge 0$, since $|e^{i\omega t}| = |\cos \omega t + i \sin \omega t| = 1$ by (A1). By (A5) this tends to 0, and a complex sequence tends to 0 exactly when its moduli do. Hence $e^{t\mathbf{J}} \to \mathbf{O}$.

(iii) $\Rightarrow$ (ii). Suppose some eigenvalue λ has $\operatorname{Re}\lambda = c \geq 0$, and let $\mathbf{v} \neq \mathbf{0}$ be an eigenvector. Then $\mathbf{x}(t) = e^{\lambda t}\mathbf{v}$ is a solution: it is differentiable with $\mathbf{x}'(t) = \lambda e^{\lambda t}\mathbf{v} = e^{\lambda t}\mathbf{A}\mathbf{v} = \mathbf{A}\mathbf{x}(t)$, and $\mathbf{x}(0) = \mathbf{v}$. Pick a coordinate with $v_r \neq 0$; then

$$|x_r(t)| = |e^{\lambda t}|\,|v_r| = e^{ct}|v_r| \geq |v_r| > 0$$

for every $t \geq 0$, because $e^{ct} \geq 1$ when $c \geq 0$. So $x_r(t)$ does not tend to 0, and (i) fails. This proves the equivalence.

10.10 Convergent and Power-Bounded Matrices

Chapter 9 asked what happens to $\mathbf{A}^m$ as m grows, and could answer only for diagonalizable matrices: the powers of $\mathbf{A}$ behave like the powers of its eigenvalues. For a Markov chain that left a gap, because a transition matrix need not be diagonalizable. The Jordan form closes the gap completely. This section characterizes, in terms of the eigenvalues and the block sizes alone, when $\mathbf{A}^m$ tends to zero, when it stays bounded, when it converges, and when the geometric series $\mathbf{I} + \mathbf{A} + \mathbf{A}^2 + \cdots$ converges.

Convergence here means what it meant in Chapter 9: **entrywise**. No notion of the size of a matrix is used. Chapter 16 introduces norms and re-proves several of these results in a way that also gives numerical bounds; the price of doing without them is that we must look at one entry at a time, and the Jordan form is what makes that possible.

10.10.1 Entrywise limits and bounds

Definition 10.10.1. Entrywise Convergence and Boundedness

Let $\mathbf{M}_1, \mathbf{M}_2, \ldots$ be matrices in $M_{p\times q}(\mathbb{C})$ and $\mathbf{M} \in M_{p\times q}(\mathbb{C})$.

(a) The sequence $(\mathbf{M}_m)$ **converges** to $\mathbf{M}$, written $\mathbf{M}_m \to \mathbf{M}$, if for **every** position (r, s) the sequence of complex numbers $(\mathbf{M}_m)_{rs}$ converges to $\mathbf{M}_{rs}$. We then write $\mathbf{M} = \lim_m \mathbf{M}_m$.

(b) The sequence $(\mathbf{M}_m)$ is **bounded** if there is a real $c \geq 0$ with $|(\mathbf{M}_m)_{rs}| \leq c$ for **all** m and **all** positions (r, s).

(c) A series $\sum_{k\geq 0} \mathbf{M}_k$ **converges** to $\mathbf{M}$ if its sequence of partial sums $\sum_{k=0}^{K} \mathbf{M}_k$ converges to $\mathbf{M}$.

In words: nothing new is happening. Each of the pq positions is watched separately, and the words mean for matrices exactly what they mean for the finitely many number sequences at those positions. Since a finite set of bounds can be replaced by their maximum, (b) is the same as asking each entry sequence to be bounded.

Remark. **What we assume from analysis.** Beyond the four facts quoted in Chapter 9 §11, this section uses exactly the following.

(B1) **Polynomial against geometric decay.** For every integer $j \geq 0$ and every real r with $0 \leq r < 1$, $m^j r^m \to 0$ as $m \to \infty$.

(B2) **Growth.** For real $r > 1$, the sequence (r^m) is unbounded.

(B3) **Limits.** Limits of convergent sequences of complex numbers respect sums, scalar multiples, and products; a convergent sequence is bounded; every subsequence of a convergent sequence converges to the same limit; and $z_m \to z$ implies $|z_m| \to |z|$.

(B4) **Terms of a convergent series.** If $\sum_k z_k$ converges, then $z_k \to 0$.

Two reductions make every proof below short. The first moves the question along a similarity, so that we may assume $\mathbf{A}$ is a Jordan matrix. The second computes the powers of a single Jordan block.

Lemma 10.10.1. Convergence and Boundedness Survive a Similarity

Let $\mathbf{A}, \mathbf{B} \in M_n(\mathbb{C})$ with $\mathbf{B} = \mathbf{P}^{-1}\mathbf{A}\mathbf{P}$ for an invertible $\mathbf{P} \in M_n(\mathbb{C})$, and let $Ł \in M_n(\mathbb{C})$. Then:

(a) $\mathbf{A}^m \to Ł$ if and only if $\mathbf{B}^m \to \mathbf{P}^{-1}Ł\mathbf{P}$;

(b) $(\mathbf{A}^m)$ is bounded if and only if $(\mathbf{B}^m)$ is bounded.

Proof. By Proposition 4.2.6 (c), $\mathbf{B}^m = \mathbf{P}^{-1}\mathbf{A}^m\mathbf{P}$ for every $m \geq 0$. Writing out the products (Definition 0.9.9), the (r, s)-entry of $\mathbf{B}^m$ is

$$(\mathbf{B}^m)_{rs} = \sum_{u,v} (\mathbf{P}^{-1})_{ru}\,(\mathbf{A}^m)_{uv}\,\mathbf{P}_{vs},$$

a linear combination of the entries of $\mathbf{A}^m$ with coefficients that do not depend on m. So if $\mathbf{A}^m \to Ł$, then by (B3) each entry of $\mathbf{B}^m$ converges to the matching combination of the entries of $Ł$, that is, $\mathbf{B}^m \to \mathbf{P}^{-1}Ł\mathbf{P}$; and if $|(\mathbf{A}^m)_{uv}| \leq c$ for all m, u, v, then $|(\mathbf{B}^m)_{rs}| \leq c\sum_{u,v} |(\mathbf{P}^{-1})_{ru}|\,|\mathbf{P}_{vs}|$, a bound independent of m, so $(\mathbf{B}^m)$ is bounded. The converses follow by exchanging the roles of $\mathbf{A}$ and $\mathbf{B}$, using $\mathbf{A} = \mathbf{P}\mathbf{B}\mathbf{P}^{-1}$.

■

For the second reduction, recall from Lemma 10.9.1, applied to the polynomial x^m, that

$$\left(\mathbf{J}_k(\lambda)^m\right)_{r,\,r+j} = \binom{m}{j}\lambda^{m-j} \qquad (0 \leq j \leq k-1),$$

with zeros below the diagonal. Everything in this section is read off this one formula. Two remarks about it: the binomial coefficient is bounded by m^j, since

$$\binom{m}{j} = \frac{m(m-1)\cdots(m-j+1)}{j!} \leq \frac{m^j}{j!} \leq m^j \qquad (m \geq j \geq 1),$$

and $\binom{m}{1} = m$ grows without bound, while $\binom{m}{0} = 1$ is constant. So the only question is whether the geometric factor λ^m beats the polynomial factor m^j.

Quick check. Compute $|(\mathbf{J}_3(\lambda)^m)_{1,3}|$ for $\lambda = \frac{1}{2}$ and for $\lambda = 1$, and say in each case whether it tends to 0, stays bounded, or grows without bound.

Solution. Here $j = 2$, so the entry is $\binom{m}{2}\lambda^{m-2}$, of modulus $\frac{m(m-1)}{2}|\lambda|^{m-2}$. For $\lambda = \frac{1}{2}$ this is $2 \cdot \frac{m(m-1)}{2} 2^{-m} \le 2m^2 2^{-m}$, which tends to 0 by (B1) with $j = 2$ and $r = \frac{1}{2}$. For $\lambda = 1$ it is $\frac{m(m-1)}{2}$, which grows without bound. The modulus of λ decides, and when $|\lambda| = 1$ the block size decides.

10.10.2 Powers tending to zero

The powers of $\mathbf{A}$ *die out exactly when every eigenvalue has modulus less than one.*

Theorem 10.10.2. Powers Tending to Zero

Let $\mathbf{A} \in M_n(\mathbb{C})$, $n \ge 1$. Then $\mathbf{A}^m \to \mathbf{O}$ if and only if every eigenvalue λ of $\mathbf{A}$ satisfies $|\lambda| < 1$.

Idea. ($\Leftarrow$) Pass to the Jordan form and look at one entry: it is $\binom{m}{j}\lambda^{m-j}$, a polynomial in m times a geometric sequence that decays, and geometric decay wins. ($\Rightarrow$) The other direction needs no Jordan form at all: an eigenvector turns $\mathbf{A}^m$ into λ^m.

Proof. ($\Leftarrow$) Suppose $|\lambda| < 1$ for every $\lambda \in \operatorname{spec}(\mathbf{A})$. By Theorem 10.3.1, $\mathbf{A} = \mathbf{P}\mathbf{J}\mathbf{P}^{-1}$ with $\mathbf{J} = \mathbf{J}_{k_1}(\lambda_1) \oplus \cdots \oplus \mathbf{J}_{k_d}(\lambda_d)$ a Jordan matrix, whose diagonal entries λ_i are the eigenvalues of $\mathbf{A}$ (Corollary 10.3.2 (a)). By Theorem 8.1.4 (b), $\mathbf{J}^m$ is the direct sum of the blocks $\mathbf{J}_{k_i}(\lambda_i)^m$, so each entry of $\mathbf{J}^m$ is either 0 for every m or of the form $\binom{m}{j}\lambda_i^{m-j}$ with $0 \le j \le k_i - 1$. For the latter, with $r = |\lambda_i| < 1$, if $r = 0$ the entry is 0 for $m > j$; otherwise, for $m \ge j$,

$$\left|\binom{m}{j}\lambda_i^{m-j}\right| \le m^j r^{m-j} = r^{-j}\, m^j r^m \longrightarrow 0$$

by (B1), the factor r^{-j} being a constant. So $\mathbf{J}^m \to \mathbf{O}$, and $\mathbf{A}^m \to \mathbf{P}\,\mathbf{O}\,\mathbf{P}^{-1} = \mathbf{O}$ by Lemma 10.10.1 (a).

($\Rightarrow$) Suppose $\mathbf{A}^m \to \mathbf{O}$ and let $\lambda \in \operatorname{spec}(\mathbf{A})$ with eigenvector $\mathbf{v} \neq \mathbf{0}$. Then $\mathbf{A}^m\mathbf{v} = \lambda^m\mathbf{v}$ for every m (induction on m). Pick a coordinate r with $v_r \neq 0$. The r-th entry of $\mathbf{A}^m\mathbf{v}$ is a linear combination of entries of $\mathbf{A}^m$ with coefficients $v_1, \ldots, v_n$, so it tends to 0 by (B3); that entry is $\lambda^m v_r$, so $\lambda^m \to 0$ and hence $|\lambda|^m = |\lambda^m| \to 0$ by (B3). If $|\lambda| \ge 1$, then $|\lambda|^m \ge 1$ for every m and the limit cannot be 0. Hence $|\lambda| < 1$. This proves the theorem. ∎

Note the asymmetry: the hard direction needs the Jordan form, the easy one needs only an eigenvector. That pattern repeats in the next two theorems.

Warning. **The condition is on the eigenvalues, not on the entries.** Every entry of $\mathbf{A} = \begin{pmatrix} 0.9 & 0.9 \\ 0.9 & 0.9 \end{pmatrix}$ has modulus less than 1, yet $\mathbf{A}\mathbf{1} = 1.8\,\mathbf{1}$, so 1.8 is an eigenvalue and $\mathbf{A}^m\mathbf{1} = 1.8^m\mathbf{1}$ grows without bound. Conversely, a matrix with huge entries can have all

eigenvalues 0: the matrix $\begin{pmatrix} 0 & 10^6 \\ 0 & 0 \end{pmatrix}$ squares to $\mathbf{O}$. Nothing can be concluded from the size of the entries alone.

10.10.3 Powers that stay bounded

Dropping "tends to zero" to "stays bounded" allows eigenvalues on the unit circle, but only if they carry no chains: a block of size ≥ 2 contributes an entry $\binom{m}{1}\lambda^{m-1}$ of modulus m.

Theorem 10.10.3. Power-Bounded Matrices

Let $\mathbf{A} \in M_n(\mathbb{C})$, $n \geq 1$, with Jordan form $\mathbf{J}$. The sequence $(\mathbf{A}^m)_{m\geq 0}$ is bounded if and only if

(i) every eigenvalue λ of $\mathbf{A}$ satisfies $|\lambda| \leq 1$, and

(ii) every eigenvalue λ with $|\lambda| = 1$ has **all** of its blocks in $\mathbf{J}$ of size 1, equivalently $a_{\mathbf{A}}(\lambda) = g_{\mathbf{A}}(\lambda)$, equivalently $x - \lambda$ occurs to the first power in $m_{\mathbf{A}}$.

Proof. Write $\mathbf{A} = \mathbf{PJP}^{-1}$ with $\mathbf{J} = \mathbf{J}_{k_1}(\lambda_1) \oplus \cdots \oplus \mathbf{J}_{k_d}(\lambda_d)$ (Theorem 10.3.1). By Lemma 10.10.1 (b), $(\mathbf{A}^m)$ is bounded if and only if $(\mathbf{J}^m)$ is, and by Theorem 8.1.4 (b) the latter holds if and only if every block sequence $(\mathbf{J}_{k_i}(\lambda_i)^m)$ is bounded. The three reformulations in (ii) are Corollary 10.3.2 (a), (b), and (c): all λ-blocks have size 1 exactly when their number equals their total size, and exactly when the largest of their sizes is 1.

($\Leftarrow$) Assume (i) and (ii), and fix a block $\mathbf{J}_k(\lambda)$ of $\mathbf{J}$. If $|\lambda| < 1$, then $\mathbf{J}_k(\lambda)^m \to \mathbf{O}$ by Theorem 10.10.2 applied to the single block $\mathbf{J}_k(\lambda)$, whose only eigenvalue is λ, hence $(\mathbf{J}_k(\lambda)^m)$ is bounded by (B3). If $|\lambda| = 1$, then $k = 1$ by (ii), so $\mathbf{J}_k(\lambda)^m = (\lambda^m)$ is a 1×1 matrix of modulus $|\lambda|^m = 1$. Either way the block sequence is bounded, so $(\mathbf{J}^m)$ is bounded and so is $(\mathbf{A}^m)$.

($\Rightarrow$) Assume $(\mathbf{A}^m)$ is bounded, hence so is $(\mathbf{J}^m)$, hence so is every block sequence.

Proof of (i). Let λ be an eigenvalue and $\mathbf{J}_k(\lambda)$ one of its blocks. The $(1,1)$-entry of $\mathbf{J}_k(\lambda)^m$ is λ^m, so $(|\lambda|^m)$ is bounded; by (B2) this forces $|\lambda| \leq 1$.

Proof of (ii). Suppose some eigenvalue λ with $|\lambda| = 1$ had a block $\mathbf{J}_k(\lambda)$ with $k \geq 2$. Its $(1,2)$-entry at the m-th power is $\binom{m}{1}\lambda^{m-1} = m\lambda^{m-1}$, of modulus $m|\lambda|^{m-1} = m$, which is unbounded — contradicting the boundedness of that block sequence. So every such block has size 1. This proves the theorem.

■

10.10.4 Powers that converge

Boundedness is not convergence: $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}^m$ alternates between two matrices forever. Among the eigenvalues of modulus 1, only $\lambda = 1$ has convergent powers, because λ^m has to sit still.

Theorem 10.10.4. Convergent Matrices

Let $\mathbf{A} \in M_n(\mathbb{C})$, $n \geq 1$, with Jordan form $\mathbf{J}$. Then $(\mathbf{A}^m)$ converges if and only if every eigenvalue λ of $\mathbf{A}$ satisfies

(i) $|\lambda| < 1$, or

(ii) $\lambda = 1$ and all of its blocks in $\mathbf{J}$ have size 1.

In that case

$$\lim_{m\to\infty} \mathbf{A}^m = \mathbf{\Pi},$$

where $\mathbf{\Pi} \in M_n(\mathbb{C})$ is the matrix, in the standard basis, of the projection onto $E_1(\mathbf{A})$ along $\bigoplus_{\lambda\neq 1} G_\lambda(\mathbf{A})$ if $1 \in \operatorname{spec}(\mathbf{A})$, and $\mathbf{\Pi} = \mathbf{O}$ otherwise.

Idea. For the shape of the answer, take the Jordan form and let $m \to \infty$ in each block: the blocks with $|\lambda| < 1$ go to $\mathbf{O}$, the blocks $\mathbf{J}_1(1) = (1)$ stay at 1. So $\mathbf{J}^m$ tends to the 0/1 diagonal matrix that marks the 1-blocks, which is the matrix of a projection; conjugating back identifies the two subspaces. For the converse, the $(1,1)$-entry of a block is λ^m, and if that converges to z then so does $\lambda^{m+1} = \lambda \cdot \lambda^m$, giving $z = \lambda z$ and forcing $\lambda = 1$ or $z = 0$.

Proof. Write $\mathbf{A} = \mathbf{P}\mathbf{J}\mathbf{P}^{-1}$ with $\mathbf{J} = \mathbf{J}_{k_1}(\lambda_1) \oplus \cdots \oplus \mathbf{J}_{k_d}(\lambda_d)$ (Theorem 10.3.1), and let $\mathcal{B} = (\mathbf{b}_1, \ldots, \mathbf{b}_n)$ be the corresponding Jordan basis, the columns of $\mathbf{P}$. For each i let $C_i \subseteq \mathbb{C}^n$ be the span of the k_i basis vectors belonging to block i, so that $\mathbb{C}^n = C_1 \oplus \cdots \oplus C_d$.

($\Leftarrow$) Suppose every eigenvalue satisfies (i) or (ii). Fix a block. If $|\lambda_i| < 1$, then $\mathbf{J}_{k_i}(\lambda_i)^m \to \mathbf{O}$ by Theorem 10.10.2 applied to the block $\mathbf{J}_{k_i}(\lambda_i)$, whose only eigenvalue is λ_i. If $\lambda_i = 1$, then $k_i = 1$ by (ii) and $\mathbf{J}_1(1)^m = (1)$ for every m. By Theorem 8.1.4 (b), $\mathbf{J}^m$ is the direct sum of these, so

$$\mathbf{J}^m \longrightarrow \mathbf{E},$$

where $\mathbf{E}$ is the diagonal matrix carrying 1 in the positions of the blocks with $\lambda_i = 1$ and 0 everywhere else. By Lemma 10.10.1 (a), $\mathbf{A}^m \to \mathbf{P}\mathbf{E}\mathbf{P}^{-1}$.

It remains to identify $\mathbf{P}\mathbf{E}\mathbf{P}^{-1}$. The matrix $\mathbf{E}$ is the matrix in the basis $\mathcal{B}$ of the projection onto

$$U := \bigoplus_{\lambda_i=1} C_i \qquad \text{along} \qquad W := \bigoplus_{\lambda_i\neq 1} C_i,$$

because $\mathbf{E}$ fixes each basis vector of the first family and kills each of the second (Theorem 4.5.1). Since the columns of $\mathbf{P}$ are the vectors of $\mathcal{B}$, $\mathbf{P}$ is the change-of-coordinates matrix from $\mathcal{B}$ to the standard basis, so $\mathbf{P}\mathbf{E}\mathbf{P}^{-1}$ is the matrix of that projection in the standard basis (Theorem 4.2.3). Now, for any $\mu \in \operatorname{spec}(\mathbf{A})$,

$$\bigoplus_{\lambda_i=\mu} C_i = G_\mu(\mathbf{A}).$$

Indeed each C_i with $\lambda_i = \mu$ is killed by $(\mathbf{A} - \mu\mathbf{I})^{k_i}$, hence lies in $G_\mu(\mathbf{A})$ (Corollary 10.1.2 (a)); the sum on the left is direct and has dimension $\sum_{\lambda_i=\mu} k_i = a_{\mathbf{A}}(\mu)$ by Corollary 10.3.2 (a); and $\dim G_\mu(\mathbf{A}) = a_{\mathbf{A}}(\mu)$ by Theorem 10.1.3 (c). An inclusion of subspaces with equal dimensions is an equality (Theorem 1.6.8, applied inside $G_\mu(\mathbf{A})$). Hence $W = \bigoplus_{\lambda\neq 1} G_\lambda(\mathbf{A})$, and $U = G_1(\mathbf{A})$ if $1 \in \operatorname{spec}(\mathbf{A})$, with $U = \{\mathbf{0}\}$ otherwise. Finally, when $1 \in \operatorname{spec}(\mathbf{A})$,

condition (ii) says all 1-blocks have size 1, so $g_{\mathbf{A}}(1) = a_{\mathbf{A}}(1)$ by Corollary 10.3.2 (a), (b); since $E_1(\mathbf{A}) \subseteq G_1(\mathbf{A})$ with $\dim E_1(\mathbf{A}) = g_{\mathbf{A}}(1) = a_{\mathbf{A}}(1) = \dim G_1(\mathbf{A})$, we get $G_1(\mathbf{A}) = E_1(\mathbf{A})$. So $\lim_m \mathbf{A}^m = \mathbf{\Pi}$ as stated.

($\Rightarrow$) Suppose $(\mathbf{A}^m)$ converges. By Lemma 10.10.1 (a) so does $(\mathbf{J}^m)$, hence so does each block sequence. Fix a block $\mathbf{J}_k(\lambda)$.

Case $|\lambda| > 1$. Impossible: a convergent sequence is bounded (B3), so this is excluded by Theorem 10.10.3.

Case $|\lambda| = 1$ *and* $\lambda \neq 1$. The $(1,1)$-entries give a convergent sequence $\lambda^m \to z$. The sequence $(\lambda^{m+1})_{m\geq 0}$ is a tail of the same sequence, so it also converges to z; on the other hand $\lambda^{m+1} = \lambda \cdot \lambda^m \to \lambda z$ by (B3). Hence $z = \lambda z$, that is, $(1-\lambda)z = 0$, and $\lambda \neq 1$ forces $z = 0$. But $|\lambda^m| = 1$ for every m, so $|z| = 1$ by (B3), a contradiction. So this case does not occur either.

Case $\lambda = 1$. If $k \geq 2$, the $(1,2)$-entry of $\mathbf{J}_k(1)^m$ is $\binom{m}{1}1^{m-1} = m$, which does not converge. So $k = 1$.

The three cases leave exactly (i) and (ii). This proves the theorem.

■

The limit deserves a sentence. It is a projection, so $\mathbf{\Pi}^2 = \mathbf{\Pi}$ — which had to be true, since $\mathbf{\Pi}^2 = \lim_m \mathbf{A}^m \lim_m \mathbf{A}^m = \lim_m \mathbf{A}^{2m} = \mathbf{\Pi}$ by (B3) — and it acts as the identity on the eigenvectors for 1 and annihilates everything belonging to the other eigenvalues. In the long run, $\mathbf{A}$ forgets all of $\mathbb{C}^n$ except the fixed space $E_1(\mathbf{A})$.

Example. Deciding and Computing Limits of Powers

For each matrix, decide whether $\mathbf{A}^m$ tends to $\mathbf{O}$, stays bounded, converges, and compute $\lim \mathbf{A}^m$ when it exists.

(a) $\mathbf{A}_1 = \begin{pmatrix} 2 & 1 \\ -1 & 0 \end{pmatrix}$.

(b) $\mathbf{A}_2 = \begin{pmatrix} 0 & -\frac{1}{2} \\ \frac{1}{2} & 0 \end{pmatrix}$.

(c) $\mathbf{A}_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{3} & 1 \\ 0 & 0 & \frac{1}{3} \end{pmatrix}$.

Solution.

(a) $p_{\mathbf{A}_1} = x(x-2)+1 = (x-1)^2$, so the only eigenvalue is 1. Put $\mathbf{N} = \mathbf{A}_1 - \mathbf{I} = \begin{pmatrix} 1 & 1 \\ -1 & -1 \end{pmatrix}$; then $\mathbf{N} \neq \mathbf{O}$ and $\mathbf{N}^2 = \begin{pmatrix} 1-1 & 1-1 \\ -1+1 & -1+1 \end{pmatrix} = \mathbf{O}$, so $g_{\mathbf{A}_1}(1) = 2 - \operatorname{rank}\mathbf{N} = 1$ and the Jordan form is $\mathbf{J}_2(1)$. Condition (ii) of Theorem 10.10.3 fails, so $(\mathbf{A}_1^m)$ is **unbounded**, and in particular does not converge. Explicitly, $\mathbf{I}$ and $\mathbf{N}$ commute and $\mathbf{N}^2 = \mathbf{O}$, so the

binomial theorem gives

$$\mathbf{A}_1^m = (\mathbf{I}+\mathbf{N})^m = \mathbf{I} + m\mathbf{N} = \begin{pmatrix} 1+m & m \\ -m & 1-m \end{pmatrix},$$

which grows linearly.

(b) $p_{\mathbf{A}_2} = x^2 + \frac{1}{4}$, with roots $\pm\frac{i}{2}$, both of modulus $\frac{1}{2} < 1$. By Theorem 10.10.2, $\mathbf{A}_2^m \to \mathbf{O}$ (so also bounded and convergent). Explicitly $\mathbf{A}_2^2 = -\frac{1}{4}\mathbf{I}$, so $\mathbf{A}_2^{2l} = (-\frac{1}{4})^l\mathbf{I}$ and $\mathbf{A}_2^{2l+1} = (-\frac{1}{4})^l\mathbf{A}_2$, and every entry has modulus at most $4^{-l} \to 0$.

(c) $\mathbf{A}_3 = (1) \oplus \mathbf{J}_2(\frac{1}{3})$ is already a Jordan matrix. Its eigenvalues are 1, with a single block of size 1, and $\frac{1}{3}$, with $|\frac{1}{3}| < 1$. So condition (i) or (ii) holds for each eigenvalue and $\mathbf{A}_3^m$ converges by Theorem 10.10.4. With $\mathbf{P} = \mathbf{I}$, the matrix $\mathbf{E}$ of the proof is $\operatorname{diag}(1,0,0)$, so

$$\lim_m \mathbf{A}_3^m = \operatorname{diag}(1,0,0),$$

the projection onto $E_1(\mathbf{A}_3) = \operatorname{span}(\mathbf{e}_1)$ along $G_{1/3}(\mathbf{A}_3) = \operatorname{span}(\mathbf{e}_2, \mathbf{e}_3)$. Explicitly, by Lemma 10.9.1,

$$\mathbf{A}_3^m = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 3^{-m} & m\,3^{-(m-1)} \\ 0 & 0 & 3^{-m} \end{pmatrix},$$

and $m\,3^{-(m-1)} = 3m\,3^{-m} \to 0$ by (B1).

10.10.5 Markov chains, finished

Chapter 9 proved that a Markov chain settles down under three hypotheses: $\mathbf{A}$ diagonalizable over $\mathbb{C}$, $a_{\mathbf{A}}(1) = 1$, and $|\lambda| < 1$ for every other eigenvalue (Theorem 9.11.4). The first hypothesis can now be dropped, and more: for a stochastic matrix, the block condition of Theorem 10.10.3 holds automatically, because the powers of a stochastic matrix have entries in $[0,1]$.

Corollary 10.10.5. Convergence of a Markov Chain

Let $\mathbf{A} \in M_n(\mathbb{R})$ be stochastic (Definition 9.11.1), $n \geq 1$. Then:

(a) $(\mathbf{A}^m)$ is bounded; consequently every eigenvalue $\lambda \in \mathbb{C}$ of $\mathbf{A}$ satisfies $|\lambda| \leq 1$, and every eigenvalue with $|\lambda| = 1$ has only 1×1 blocks. In particular $a_{\mathbf{A}}(1) = g_{\mathbf{A}}(1)$.

(b) $(\mathbf{A}^m)$ converges if and only if 1 is the **only** eigenvalue of $\mathbf{A}$ of modulus 1; and then $\lim_m \mathbf{A}^m$ is the projection onto $E_1(\mathbf{A})$ along $\bigoplus_{\lambda\neq 1} G_\lambda(\mathbf{A})$.

Proof.

(a) By Proposition 9.11.3 (b), every $\mathbf{A}^m$ is stochastic, so its entries are nonnegative and each of its columns sums to 1; hence every entry lies in $[0,1]$ and $(\mathbf{A}^m)$ is bounded with $c = 1$. Now apply the ($\Rightarrow$) direction of Theorem 10.10.3. Since $1 \in \operatorname{spec}(\mathbf{A})$ by Proposition 9.11.3 (c), the 1-blocks exist and all have size 1, so $a_{\mathbf{A}}(1) = g_{\mathbf{A}}(1)$ by Corollary 10.3.2 (a), (b).

(b) By (a), every eigenvalue has $|\lambda| \leq 1$ and every eigenvalue of modulus 1 has only 1×1 blocks. So the criterion of Theorem 10.10.4 — each eigenvalue has $|\lambda| < 1$, or equals 1 with only 1×1 blocks — holds precisely when no eigenvalue $\lambda \neq 1$ has $|\lambda| = 1$. The

limit is the one given by Theorem 10.10.4.

■

So the only obstruction to a Markov chain settling down is an eigenvalue of modulus 1 other than 1 itself — periodicity, in the language of probability. Chapter 9's warning example, the swap $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$, has the eigenvalue -1 and is exactly this obstruction.

Example. A Two-State Chain

Let $\mathbf{A} = \begin{pmatrix} 3/4 & 1/2 \\ 1/4 & 1/2 \end{pmatrix}$. Decide whether $\mathbf{A}^m$ converges and find the limit.

Solution. $\mathbf{A}$ is stochastic: the entries are nonnegative, and each column sums to 1. Its characteristic polynomial is

$$p_{\mathbf{A}} = (x - \tfrac{3}{4})(x - \tfrac{1}{2}) - \tfrac{1}{8} = x^2 - \tfrac{5}{4}x + \tfrac{1}{4} = (x - 1)(x - \tfrac{1}{4}),$$

so $\operatorname{spec}(\mathbf{A}) = \{1, \tfrac{1}{4}\}$ and the only eigenvalue of modulus 1 is 1. By Corollary 10.10.5 (b), $\mathbf{A}^m$ converges.

For the limit, $E_1(\mathbf{A}) = \ker(\mathbf{A} - \mathbf{I})$, and $\mathbf{A} - \mathbf{I} = \begin{pmatrix} -1/4 & 1/2 \\ 1/4 & -1/2 \end{pmatrix}$ gives $\operatorname{span}((2, 1))$; the steady state is $\mathbf{v} = (\tfrac{2}{3}, \tfrac{1}{3})$, the multiple of $(2, 1)$ whose entries sum to 1. Also $E_{1/4}(\mathbf{A}) = \ker(\mathbf{A} - \tfrac{1}{4}\mathbf{I})$, and $\mathbf{A} - \tfrac{1}{4}\mathbf{I} = \begin{pmatrix} 1/2 & 1/2 \\ 1/4 & 1/4 \end{pmatrix}$ gives $\operatorname{span}((1, -1))$. Since the two eigenvalues are distinct, $G_{1/4}(\mathbf{A}) = E_{1/4}(\mathbf{A})$ and the limit $\mathbf{\Pi}$ is the projection onto $\operatorname{span}((2, 1))$ along $\operatorname{span}((1, -1))$. Computing it in the basis $\mathcal{C} = ((2, 1), (1, -1))$: with $\mathbf{Q} = \begin{pmatrix} 2 & 1 \\ 1 & -1 \end{pmatrix}$, $\det \mathbf{Q} = -3$, and $\mathbf{Q}^{-1} = \tfrac{1}{3}\begin{pmatrix} 1 & 1 \\ 1 & -2 \end{pmatrix}$, so

$$\mathbf{\Pi} = \mathbf{Q} \operatorname{diag}(1, 0) \mathbf{Q}^{-1} = \begin{pmatrix} 2 & 0 \\ 1 & 0 \end{pmatrix} \cdot \tfrac{1}{3}\begin{pmatrix} 1 & 1 \\ 1 & -2 \end{pmatrix} = \tfrac{1}{3}\begin{pmatrix} 2 & 2 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 2/3 & 2/3 \\ 1/3 & 1/3 \end{pmatrix}.$$

Check. $\mathbf{\Pi} = \mathbf{v}\mathbf{1}^{\top}$, as Theorem 9.11.4 predicts; $\mathbf{\Pi}^2 = \mathbf{\Pi}$ since $\mathbf{1}^{\top}\mathbf{v} = 1$; the columns of $\mathbf{\Pi}$ are the steady state, so every initial probability vector $\mathbf{x}_0$ gives $\mathbf{x}_m = \mathbf{A}^m \mathbf{x}_0 \to \mathbf{\Pi}\mathbf{x}_0 = \mathbf{v}(\mathbf{1}^{\top}\mathbf{x}_0) = \mathbf{v}$. And $\mathbf{A}^2 = \begin{pmatrix} 11/16 & 5/8 \\ 5/16 & 3/8 \end{pmatrix}$ is already close to $\mathbf{\Pi}$, the error decaying like 4^{-m}.

10.10.6 The geometric series of a matrix

The scalar identity $\sum_k a^k = (1 - a)^{-1}$ for $|a| < 1$ has a matrix version, with the same proof and the same hypothesis read off the eigenvalues.

Theorem 10.10.6. Neumann Series

Let $\mathbf{A} \in M_n(\mathbb{C})$, $n \geq 1$. Then the series $\sum_{k\geq 0} \mathbf{A}^k$ converges if and only if every eigenvalue λ of $\mathbf{A}$ satisfies $|\lambda| < 1$. In that case $\mathbf{I} - \mathbf{A}$ is invertible and

$$\sum_{k=0}^{\infty} \mathbf{A}^k = (\mathbf{I} - \mathbf{A})^{-1}.$$

Idea. Telescoping gives $(\mathbf{I} - \mathbf{A})(\mathbf{I} + \mathbf{A} + \cdots + \mathbf{A}^K) = \mathbf{I} - \mathbf{A}^{K+1}$, an exact identity for every K. If $\mathbf{I} - \mathbf{A}$ is invertible, solve it for the partial sum and let $K \to \infty$, where Theorem 10.10.2 kills $\mathbf{A}^{K+1}$. Conversely, the terms of a convergent series tend to $\mathbf{O}$, which is the hypothesis of that theorem again.

Proof. ($\Leftarrow$) Suppose $|\lambda| < 1$ for every $\lambda \in \operatorname{spec}(\mathbf{A})$. In particular $1 \notin \operatorname{spec}(\mathbf{A})$, so $0 \notin \operatorname{spec}(\mathbf{I} - \mathbf{A})$ — a vector with $(\mathbf{I} - \mathbf{A})\mathbf{v} = \mathbf{0}$ is a vector with $\mathbf{A}\mathbf{v} = \mathbf{v}$ — and $\mathbf{I} - \mathbf{A}$ is invertible (Theorem 9.2.6). Put $\mathbf{S}_K = \sum_{k=0}^{K} \mathbf{A}^k$. Multiplying out and canceling,

$$(\mathbf{I} - \mathbf{A})\mathbf{S}_K = \sum_{k=0}^{K} \mathbf{A}^k - \sum_{k=1}^{K+1} \mathbf{A}^k = \mathbf{I} - \mathbf{A}^{K+1},$$

hence $\mathbf{S}_K = (\mathbf{I} - \mathbf{A})^{-1}(\mathbf{I} - \mathbf{A}^{K+1})$. By Theorem 10.10.2, $\mathbf{A}^{K+1} \to \mathbf{O}$. Each entry of $\mathbf{S}_K$ is a fixed linear combination of the entries of $\mathbf{I} - \mathbf{A}^{K+1}$, with coefficients from $(\mathbf{I} - \mathbf{A})^{-1}$, so by (B3) it converges to the corresponding combination for $\mathbf{I}$, that is, $\mathbf{S}_K \to (\mathbf{I} - \mathbf{A})^{-1}\mathbf{I} = (\mathbf{I} - \mathbf{A})^{-1}$.

($\Rightarrow$) Suppose $\sum_k \mathbf{A}^k$ converges. Fix a position (r, s); the series of complex numbers $\sum_k (\mathbf{A}^k)_{rs}$ converges, so $(\mathbf{A}^k)_{rs} \to 0$ by (B4). This holds at every position, so $\mathbf{A}^k \to \mathbf{O}$, and Theorem 10.10.2 gives $|\lambda| < 1$ for every eigenvalue. This proves the theorem. ∎

Example. Inverting by Summing a Series

Let $\mathbf{A} = \begin{pmatrix} 0 & 1/2 \\ 1/2 & 0 \end{pmatrix}$. Verify that $\sum_k \mathbf{A}^k$ converges, sum it, and compare with $(\mathbf{I} - \mathbf{A})^{-1}$.

Solution. $p_{\mathbf{A}} = x^2 - \frac{1}{4}$, so $\operatorname{spec}(\mathbf{A}) = \{\frac{1}{2}, -\frac{1}{2}\}$ and both eigenvalues have modulus $\frac{1}{2} < 1$. By Theorem 10.10.6 the series converges.

Summing it. $\mathbf{A}^2 = \frac{1}{4}\mathbf{I}$, so $\mathbf{A}^{2l} = 4^{-l}\mathbf{I}$ and $\mathbf{A}^{2l+1} = 4^{-l}\mathbf{A}$. Grouping the partial sums by parity, which is legitimate because each entry sequence splits into two convergent pieces (B3),

$$\sum_{k\geq 0} \mathbf{A}^k = \Big(\sum_{l\geq 0} 4^{-l}\Big)(\mathbf{I} + \mathbf{A}) = \frac{1}{1 - \frac{1}{4}}(\mathbf{I} + \mathbf{A}) = \tfrac{4}{3}\begin{pmatrix} 1 & 1/2 \\ 1/2 & 1 \end{pmatrix} = \begin{pmatrix} 4/3 & 2/3 \\ 2/3 & 4/3 \end{pmatrix},$$

the scalar sum being the case $n = 1$ of Theorem 10.10.6 applied to the 1×1 matrix $(\frac{1}{4})$.

The inverse. $\mathbf{I}-\mathbf{A}=\begin{pmatrix}1 & -1/2\\ -1/2 & 1\end{pmatrix}$ has determinant $1-\frac{1}{4}=\frac{3}{4}$, so by Theorem 0.9.8

$$(\mathbf{I}-\mathbf{A})^{-1}=\tfrac{4}{3}\begin{pmatrix}1 & 1/2\\ 1/2 & 1\end{pmatrix},$$

in agreement.

Remark. The first thing every criterion in this section looks at is the largest modulus of an eigenvalue,

$$\rho(\mathbf{A})=\max\{|\lambda|:\lambda\in\operatorname{spec}(\mathbf{A})\},$$

called the **spectral radius** of $\mathbf{A}$. Two of the criteria need nothing else: $\mathbf{A}^m\to\mathbf{O}$ if and only if $\rho(\mathbf{A})<1$, and $\sum_k\mathbf{A}^k$ converges if and only if $\rho(\mathbf{A})<1$. The other two need more when $\rho(\mathbf{A})=1$: boundedness also asks for the block sizes at modulus 1, and convergence asks in addition that 1 be the only eigenvalue of modulus 1. Chapter 16 measures matrices by norms, proves that the spectral radius never exceeds the norm of $\mathbf{A}$ induced by any vector norm, and recovers $\rho(\mathbf{A})$ as a limit built from the norms of the powers $\mathbf{A}^m$; that is what turns these exact criteria into usable estimates.

10.10.7 Exercises

10.10.7.1 A. Check your understanding

Exercise. A1

(a) Define entrywise convergence and entrywise boundedness for a sequence of matrices.

(b) Write down the entries of $\mathbf{J}_k(\lambda)^m$, and say which one shows that a block of size ≥ 2 with $|\lambda|=1$ has unbounded powers.

(c) State the three criteria: $\mathbf{A}^m\to\mathbf{O}$, $(\mathbf{A}^m)$ bounded, $(\mathbf{A}^m)$ convergent.

(d) True or false: if every entry of $\mathbf{A}$ has modulus less than 1, then $\mathbf{A}^m\to\mathbf{O}$. Justify your answer.

(e) True or false: if $(\mathbf{A}^m)$ is bounded then it converges. Justify your answer.

(f) For which $\mathbf{A}$ does $\sum_k\mathbf{A}^k$ converge, and what is the sum?

Solution.

(a) $\mathbf{M}_m\to\mathbf{M}$ if $(\mathbf{M}_m)_{rs}\to\mathbf{M}_{rs}$ for every position (r,s); $(\mathbf{M}_m)$ is bounded if there is one constant c with $|(\mathbf{M}_m)_{rs}|\leq c$ for all m and all (r,s) (Definition 10.10.1).

(b) $\left(\mathbf{J}_k(\lambda)^m\right)_{r,r+j}=\binom{m}{j}\lambda^{m-j}$ for $0\leq j\leq k-1$, with zeros below the diagonal (Lemma 10.9.1 with $p=x^m$). The entry with $j=1$ is $m\lambda^{m-1}$, of modulus m when $|\lambda|=1$.

(c) $\mathbf{A}^m\to\mathbf{O}$ if and only if every eigenvalue has $|\lambda|<1$ (Theorem 10.10.2); $(\mathbf{A}^m)$ is bounded if and only if every eigenvalue has $|\lambda|\leq 1$ and those with $|\lambda|=1$ have only 1×1 blocks (Theorem 10.10.3); $(\mathbf{A}^m)$ converges if and only if every eigenvalue has $|\lambda|<1$ or equals 1 with only 1×1 blocks (Theorem 10.10.4).

(d) False. $\mathbf{A} = \begin{pmatrix} 0.9 & 0.9 \\ 0.9 & 0.9 \end{pmatrix}$ has the eigenvalue 1.8, and $\mathbf{A}^m\mathbf{1} = 1.8^m\mathbf{1}$ is unbounded; see the warning after Theorem 10.10.2.

(e) False. $\mathbf{A} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ satisfies $\mathbf{A}^2 = \mathbf{I}$, so $\mathbf{A}^m$ alternates between $\mathbf{I}$ and $\mathbf{A}$ and is bounded but not convergent. Its eigenvalues are ± 1, and the eigenvalue -1 is what Theorem 10.10.4 excludes.

(f) Exactly when every eigenvalue has $|\lambda| < 1$, and then the sum is $(\mathbf{I} - \mathbf{A})^{-1}$ (Theorem 10.10.6).

10.10.7.2 B. Practice

Exercise. B1: Decide the behavior

For each matrix over $\mathbb{C}$, decide whether $\mathbf{A}^m \to \mathbf{O}$, whether $(\mathbf{A}^m)$ is bounded, and whether $(\mathbf{A}^m)$ converges. Justify your answer.

(a) $\mathbf{A}_1 = \begin{pmatrix} 1/2 & 1 \\ 0 & 1/2 \end{pmatrix}$.

(b) $\mathbf{A}_2 = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$.

(c) $\mathbf{A}_3 = \mathbf{J}_2(1) \oplus \mathbf{J}_1(\frac{1}{2}) \in M_3(\mathbb{C})$.

(d) $\mathbf{A}_4 = \begin{pmatrix} 0 & 0 & 5 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$.

Solution.

(a) The only eigenvalue is $\frac{1}{2}$, of modulus < 1, so $\mathbf{A}_1^m \to \mathbf{O}$ by Theorem 10.10.2; hence it is bounded and convergent. (The superdiagonal 1 is irrelevant: by Lemma 10.9.1 the $(1,2)$-entry of $\mathbf{A}_1^m$ is $m\,2^{-(m-1)} \to 0$ by (B1).)

(b) $p_{\mathbf{A}_2} = x^2 + 1$, so $\operatorname{spec}(\mathbf{A}_2) = \{i, -i\}$, both of modulus 1, and the eigenvalues are distinct, so every block has size 1. By Theorem 10.10.3, $(\mathbf{A}_2^m)$ is **bounded**; it does not tend to $\mathbf{O}$ (Theorem 10.10.2) and does not converge, since the eigenvalue i has modulus 1 and is not 1 (Theorem 10.10.4). Concretely $\mathbf{A}_2^4 = \mathbf{I}$, so the powers cycle through four matrices.

(c) The eigenvalue 1 has a block of size 2, so Theorem 10.10.3 (ii) fails: $(\mathbf{A}_3^m)$ is **unbounded**, hence neither convergent nor tending to $\mathbf{O}$. The $(1,2)$-entry of $\mathbf{A}_3^m$ is m.

(d) $\mathbf{A}_4^2 = \mathbf{O}$, so $\mathbf{A}_4^m = \mathbf{O}$ for $m \geq 2$ and $\mathbf{A}_4^m \to \mathbf{O}$ trivially: all eigenvalues are 0. The large entry 5 does not matter, illustrating the warning after Theorem 10.10.2.

Exercise. B2: Computing limits

(a) Let $\mathbf{A} = \begin{pmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{pmatrix}$. Show that $\mathbf{A}^m$ converges and find the limit, both by Theorem 10.10.4 and by a direct computation of $\mathbf{A}^m$.

(b) Let $\mathbf{B} = \begin{pmatrix} 1 & 1/3 \\ 0 & 1/3 \end{pmatrix}$. Find $\lim_m \mathbf{B}^m$ and identify it as a projection.

Solution.

(a) $p_{\mathbf{A}} = x^2 - x = x(x-1)$, so $\operatorname{spec}(\mathbf{A}) = \{0, 1\}$. The eigenvalue 0 has $|0| < 1$ and the eigenvalue 1 has $a_{\mathbf{A}}(1) = 1$, so its single block has size 1; by Theorem 10.10.4 the powers converge, and the limit is the projection onto $E_1(\mathbf{A})$ along $G_0(\mathbf{A}) = E_0(\mathbf{A})$. Now $\mathbf{A} - \mathbf{I} = \begin{pmatrix} -1/2 & 1/2 \\ 1/2 & -1/2 \end{pmatrix}$ gives $E_1(\mathbf{A}) = \operatorname{span}((1,1))$ and $E_0(\mathbf{A}) = \ker \mathbf{A} = \operatorname{span}((1,-1))$. With $\mathbf{Q} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$ and $\mathbf{Q}^{-1} = \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$,

$$\lim_m \mathbf{A}^m = \mathbf{Q}\operatorname{diag}(1,0)\mathbf{Q}^{-1} = \tfrac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} = \mathbf{A}.$$

Directly. $\mathbf{A}^2 = \mathbf{A}$, so $\mathbf{A}^m = \mathbf{A}$ for every $m \geq 1$ and the limit is $\mathbf{A}$. ($\mathbf{A}$ is itself a projection, and the theorem returns it unchanged.)

(b) $\mathbf{B}$ is upper triangular, so $\operatorname{spec}(\mathbf{B}) = \{1, \frac{1}{3}\}$ (Theorem 7.4.6). Both criteria hold: $|\frac{1}{3}| < 1$, and $a_{\mathbf{B}}(1) = 1$ so the 1-block has size 1. Solving $(\mathbf{B}-\mathbf{I})\mathbf{x} = \mathbf{0}$ with $\mathbf{B}-\mathbf{I} = \begin{pmatrix} 0 & 1/3 \\ 0 & -2/3 \end{pmatrix}$ gives $E_1(\mathbf{B}) = \operatorname{span}(\mathbf{e}_1)$; solving $(\mathbf{B} - \frac{1}{3}\mathbf{I})\mathbf{x} = \mathbf{0}$ with $\mathbf{B} - \frac{1}{3}\mathbf{I} = \begin{pmatrix} 2/3 & 1/3 \\ 0 & 0 \end{pmatrix}$ gives $E_{1/3}(\mathbf{B}) = \operatorname{span}((1,-2))$. With $\mathbf{Q} = \begin{pmatrix} 1 & 1 \\ 0 & -2 \end{pmatrix}$ we have $\det \mathbf{Q} = -2$ and $\mathbf{Q}^{-1} = \frac{1}{-2}\begin{pmatrix} -2 & -1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 1/2 \\ 0 & -1/2 \end{pmatrix}$. Multiplying in two steps, $\mathbf{Q}\operatorname{diag}(1,0) = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$ and then

$$\lim_m \mathbf{B}^m = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}\mathbf{Q}^{-1} = \begin{pmatrix} 1 & 1/2 \\ 0 & 0 \end{pmatrix}.$$

This is the projection onto $\operatorname{span}(\mathbf{e}_1)$ along $\operatorname{span}((1,-2))$: it fixes $\mathbf{e}_1$ and sends $(1,-2)$ to $(1-1, 0) = \mathbf{0}$. *Check.* $\mathbf{B}^m = \begin{pmatrix} 1 & c_m \\ 0 & 3^{-m} \end{pmatrix}$ with $c_1 = \frac{1}{3}$ and $c_{m+1} = \frac{1}{3}(1 + c_m)$, read off $\mathbf{B}^{m+1} = \mathbf{B}^m\mathbf{B}$; the first values are $\frac{1}{3}, \frac{4}{9}, \frac{13}{27}$, climbing towards the fixed point $c = \frac{1}{3}(1+c)$, that is $c = \frac{1}{2}$.

Exercise. B3: A Neumann series

Let $\mathbf{A} = \begin{pmatrix} 1/3 & 1/3 \\ 0 & 1/3 \end{pmatrix}$.

(a) Explain why $\sum_k \mathbf{A}^k$ converges.

(b) Compute $(\mathbf{I} - \mathbf{A})^{-1}$ directly.

(c) Compute the partial sum $\mathbf{I} + \mathbf{A} + \mathbf{A}^2$ and compare it with (b).

Solution.

(a) $\mathbf{A}$ is upper triangular with both diagonal entries $\frac{1}{3}$, so $\operatorname{spec}(\mathbf{A}) = \{\frac{1}{3}\}$ (Theorem 7.4.6) and $|\frac{1}{3}| < 1$. By Theorem 10.10.6 the series converges, with sum $(\mathbf{I}-\mathbf{A})^{-1}$.

(b) $\mathbf{I}-\mathbf{A} = \begin{pmatrix} 2/3 & -1/3 \\ 0 & 2/3 \end{pmatrix}$, of determinant $\frac{4}{9}$. By Theorem 0.9.8,

$$(\mathbf{I}-\mathbf{A})^{-1} = \tfrac{9}{4}\begin{pmatrix} 2/3 & 1/3 \\ 0 & 2/3 \end{pmatrix} = \begin{pmatrix} 3/2 & 3/4 \\ 0 & 3/2 \end{pmatrix}.$$

(c) Write $\mathbf{A} = \frac{1}{3}(\mathbf{I}+\mathbf{M})$ with $\mathbf{M} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$, $\mathbf{M}^2 = \mathbf{O}$; then $\mathbf{A}^k = 3^{-k}(\mathbf{I}+k\mathbf{M})$ by the binomial theorem, the two summands commuting. So

$$\mathbf{I}+\mathbf{A}+\mathbf{A}^2 = (1+\tfrac{1}{3}+\tfrac{1}{9})\mathbf{I} + (\tfrac{1}{3}+\tfrac{2}{9})\mathbf{M} = \tfrac{13}{9}\mathbf{I} + \tfrac{5}{9}\mathbf{M} = \begin{pmatrix} 13/9 & 5/9 \\ 0 & 13/9 \end{pmatrix}.$$

Summing the whole series the same way, $(\mathbf{I}-\mathbf{A})^{-1} = (\sum_{k\geq 0} 3^{-k})\mathbf{I} + (\sum_{k\geq 0} k3^{-k})\mathbf{M}$, so comparing with (b) reads off $\sum_k 3^{-k} = \frac{3}{2}$ and $\sum_k k3^{-k} = \frac{3}{4}$. The two partial sums $\frac{13}{9}$ and $\frac{5}{9}$, about 1.44 and 0.56, are on their way to $\frac{3}{2}$ and $\frac{3}{4}$; the convergence is geometric, as the factor 3^{-k} shows.

10.10.7.3 C. Going deeper

Exercise. C1: Bounded but not convergent

(a) Give an example of $\mathbf{A} \in M_2(\mathbb{R})$ with $(\mathbf{A}^m)$ bounded but not convergent, and one with $(\mathbf{A}^m)$ convergent but $\mathbf{A}^m \not\to \mathbf{O}$.

(b) Prove that if $\mathbf{A}^m \to Ł$, then $Ł^2 = Ł$ and $\mathbf{A}Ł = Ł\mathbf{A} = Ł$.

(c) Deduce that if $\mathbf{A}^m \to Ł$ with $Ł$ **invertible**, then $Ł = \mathbf{I}$ and $\mathbf{A} = \mathbf{I}$.

(d) Give an invertible $\mathbf{A}$ with $\mathbf{A}^m \to Ł$ and $Ł \notin \{\mathbf{O}, \mathbf{I}\}$, so that the invertibility of $Ł$, not of $\mathbf{A}$, is what part (c) needs.

Solution.

(a) For the first, $\mathbf{A} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$: $\mathbf{A}^2 = \mathbf{I}$, so $\mathbf{A}^m$ alternates between $\mathbf{I}$ and $\mathbf{A}$ and is bounded but not convergent; its eigenvalues are ± 1. For the second, $\mathbf{A} = \begin{pmatrix} 1 & 0 \\ 0 & 1/2 \end{pmatrix}$: $\mathbf{A}^m = \operatorname{diag}(1, 2^{-m}) \to \operatorname{diag}(1,0) \neq \mathbf{O}$.

(b) Each entry of $\mathbf{A}^m\mathbf{A}^m = \mathbf{A}^{2m}$ is a sum of products of entries of $\mathbf{A}^m$, so by (B3) it converges to the corresponding expression in the entries of $Ł$; hence $\mathbf{A}^{2m} \to Ł^2$. But $(\mathbf{A}^{2m})_m$ is a subsequence of $(\mathbf{A}^m)_m$ — more precisely, for each position the sequence $(\mathbf{A}^{2m})_{rs}$ is a subsequence of the convergent sequence $(\mathbf{A}^m)_{rs}$, so it has the same limit $Ł_{rs}$. Hence $Ł^2 = Ł$. Similarly $\mathbf{A}^{m+1} = \mathbf{A}\mathbf{A}^m = \mathbf{A}^m\mathbf{A}$, and the left side is a tail of the convergent sequence, so it tends to $Ł$, while the two right sides tend to $\mathbf{A}Ł$ and $Ł\mathbf{A}$ by (B3). Hence $\mathbf{A}Ł = Ł\mathbf{A} = Ł$.

(c) Suppose $\mathbf{A}^m \to Ł$ with $Ł$ invertible. By (b), $Ł^2 = Ł$; multiplying by $Ł^{-1}$ gives $Ł = \mathbf{I}$.

Then $\mathbf{A}Ł = Ł$ reads $\mathbf{A} = \mathbf{I}$. (Conversely $\mathbf{A} = \mathbf{I}$ gives $\mathbf{A}^m = \mathbf{I} \to \mathbf{I}$, so this case really occurs.)

(d) $\mathbf{A} = \operatorname{diag}(1, \frac{1}{2})$ is invertible, and $\mathbf{A}^m = \operatorname{diag}(1, 2^{-m}) \to \operatorname{diag}(1, 0) = Ł$, which is neither $\mathbf{O}$ nor $\mathbf{I}$ — and not invertible. So no hypothesis on $\mathbf{A}$ alone forces $Ł \in \{\mathbf{O}, \mathbf{I}\}$: by Theorem 10.10.4 the limit is the projection onto $E_1(\mathbf{A})$ along the other generalized eigenspaces, and that projection is $\mathbf{O}$ or $\mathbf{I}$ only in the extreme cases $1 \notin \operatorname{spec}(\mathbf{A})$ and $E_1(\mathbf{A}) = \mathbb{C}^n$.

Exercise. C2: Chapter 9's Markov theorem without diagonalizability

Let $\mathbf{A} \in M_n(\mathbb{R})$ be stochastic with $a_{\mathbf{A}}(1) = 1$ and $|\lambda| < 1$ for every eigenvalue $\lambda \neq 1$.

(a) Prove that $\mathbf{1}^\top \mathbf{w} = 0$ for every $\mathbf{w} \in G_\lambda(\mathbf{A})$ with $\lambda \neq 1$. *Hint:* $\mathbf{1}^\top(\mathbf{A} - \lambda\mathbf{I}) = (1 - \lambda)\mathbf{1}^\top$.

(b) Prove that $(\mathbf{A}^m)$ converges to a **stochastic** matrix $\mathbf{\Pi}$, and deduce that $\mathbf{A}$ has a steady state $\mathbf{v}$ with $E_1(\mathbf{A}) = \operatorname{span}(\mathbf{v})$.

(c) Deduce that $\mathbf{\Pi} = \mathbf{v}\mathbf{1}^\top$ and that $\mathbf{A}^m\mathbf{x}_0 \to \mathbf{v}$ for every probability vector $\mathbf{x}_0$, **without** assuming that $\mathbf{A}$ is diagonalizable.

Solution.

(a) Since $\mathbf{A}$ is stochastic, $\mathbf{1}^\top\mathbf{A} = \mathbf{1}^\top$ (Proposition 9.11.3 (c)), so $\mathbf{1}^\top(\mathbf{A} - \lambda\mathbf{I}) = \mathbf{1}^\top - \lambda\mathbf{1}^\top = (1 - \lambda)\mathbf{1}^\top$. Iterating n times, $\mathbf{1}^\top(\mathbf{A} - \lambda\mathbf{I})^n = (1 - \lambda)^n\mathbf{1}^\top$. Let $\mathbf{w} \in G_\lambda(\mathbf{A})$, so $(\mathbf{A} - \lambda\mathbf{I})^n\mathbf{w} = \mathbf{0}$ (Corollary 10.1.2 (a)). Then

$$0 = \mathbf{1}^\top(\mathbf{A} - \lambda\mathbf{I})^n\mathbf{w} = (1 - \lambda)^n\,\mathbf{1}^\top\mathbf{w},$$

and $\lambda \neq 1$ gives $(1 - \lambda)^n \neq 0$, so $\mathbf{1}^\top\mathbf{w} = 0$.

(b) By hypothesis 1 is the only eigenvalue of modulus 1, so $(\mathbf{A}^m)$ converges by Corollary 10.10.5 (b); call the limit $\mathbf{\Pi}$. Every $\mathbf{A}^m$ is stochastic (Proposition 9.11.3 (b)), so all its entries are ≥ 0 and each of its columns sums to 1. Both properties pass to the limit: a limit of real numbers ≥ 0 is ≥ 0, and a column sum is a finite sum of entries, so it converges to the sum of the limits (Chapter 9 §11). Hence $\mathbf{\Pi}$ is stochastic, and in particular $\mathbf{\Pi} \neq \mathbf{O}$.
By Theorem 10.10.4, $\mathbf{\Pi}$ is the projection onto $E_1(\mathbf{A})$ along $\bigoplus_{\lambda \neq 1} G_\lambda(\mathbf{A})$, so $\operatorname{im} T_{\mathbf{\Pi}} = E_1(\mathbf{A})$. Since $a_{\mathbf{A}}(1) = 1$ and $1 \leq g_{\mathbf{A}}(1) \leq a_{\mathbf{A}}(1)$ (Theorem 9.3.1), $E_1(\mathbf{A})$ is a line. Each column of $\mathbf{\Pi}$ lies in $\operatorname{im} T_{\mathbf{\Pi}} = E_1(\mathbf{A})$ (Definition 0.9.9: column j is $\mathbf{\Pi}\mathbf{e}_j$) and is a probability vector, hence is fixed by $\mathbf{A}$ and is a steady state. Any two probability vectors on the same line are equal: if $\mathbf{w} = c\mathbf{v}$ with both entry sums equal to 1, then $1 = c$. So all the columns of $\mathbf{\Pi}$ are one and the same steady state $\mathbf{v}$, and $E_1(\mathbf{A}) = \operatorname{span}(\mathbf{v})$.

(c) All columns of $\mathbf{\Pi}$ equal $\mathbf{v}$, which is exactly the statement $\mathbf{\Pi} = \mathbf{v}\mathbf{1}^\top$. (Independently of (b), part (a) confirms it: $\mathbf{v}\mathbf{1}^\top$ fixes $\mathbf{v}$ because $\mathbf{1}^\top\mathbf{v} = 1$, and kills every $\mathbf{w} \in G_\lambda(\mathbf{A})$ with $\lambda \neq 1$ because $\mathbf{1}^\top\mathbf{w} = 0$; by Theorem 10.1.3 (a) those two conditions determine the projection onto $E_1(\mathbf{A}) = G_1(\mathbf{A})$ along the rest, using $\dim G_1(\mathbf{A}) = a_{\mathbf{A}}(1) = 1 = \dim E_1(\mathbf{A})$ and Theorem 1.6.8.) For a probability vector $\mathbf{x}_0$, each entry of $\mathbf{A}^m\mathbf{x}_0$ is a fixed linear combination of entries of $\mathbf{A}^m$, so by (B3)

$$\mathbf{A}^m\mathbf{x}_0 \longrightarrow \mathbf{\Pi}\mathbf{x}_0 = \mathbf{v}\,(\mathbf{1}^\top\mathbf{x}_0) = \mathbf{v}\,(x_{01} + \cdots + x_{0n}) = \mathbf{v}.$$

No diagonalizability was used: the blocks for the eigenvalues $\lambda \neq 1$ may have any sizes, since Theorem 10.10.2 does not care. This removes the diagonalizability hypothesis from Theorem 9.11.4.

Chapter 11

Inner Product Spaces

Nine chapters have described vector spaces without measuring anything in them. We can say that a list spans, that a subspace has dimension four, that two matrices are similar; we cannot say that a vector has length 3, that two vectors are perpendicular, or that one vector is the closest point of a plane to another. Every picture in the book so far has secretly used lengths and angles that the theory did not supply. This chapter supplies them, with a single piece of extra structure: an **inner product**, abstracted from the dot product on $\mathbb{R}^n$ by keeping exactly the properties that make the law of cosines work.

The gain is out of all proportion to the cost. Length gives the Cauchy–Schwarz and triangle inequalities. Perpendicularity gives orthonormal bases, in which coordinates are read off by a single inner product and Gram–Schmidt produces one from any basis at all. Splitting a space into a subspace and its orthogonal complement gives the nearest point of a subspace to a given vector, and with it least squares, the standard answer to an inconsistent system. The Riesz representation theorem pays off a debt from Chapter 5: an inner product is precisely the extra data that makes V and V^* the same space. From Riesz comes the **adjoint** T^*, which turns the four fundamental subspaces into two orthogonal decompositions, and the isometries, which are the maps that preserve the whole geometry. The chapter ends with the computational and classical faces of the same ideas: the QR factorization, Fourier series, and orthogonal polynomials. **Throughout, the field is $\mathbb{R}$ or $\mathbb{C}$**, because the construction needs both an ordering, so that a length squared can be positive, and complex conjugation, so that it can be real.

What you need. From Chapter 0, the complex conjugate and modulus with their arithmetic (Definition 0.6.2, Theorem 0.6.2) and the trace (Definition 0.9.12, Theorem 0.9.6). From Chapter 1, independence, bases, dimension, basis extension, and direct sums (Definition 1.4.1, Theorem 1.6.4, Theorem 1.6.5, Theorem 1.7.4). From Chapter 2, the coordinate isomorphism and Rank–Nullity (Corollary 2.5.5, Theorem 2.3.1). From Chapter 3, the Invertible Matrix Theorem (Theorem 3.4.4). From Chapter 4, the matrix of a linear map (Definition 4.1.1). From Chapter 5, the dual space, the annihilator, and the dual map, together with the warning that an isomorphism $V \cong V^*$ needs a choice (Definition 5.1.2, Definition 5.2.1, Definition 5.4.1, Theorem 5.4.2). From Chapter 6, the root bound and Lagrange interpolation (Corollary 6.1.4, Theorem 6.6.1). From Chapter 7, multiplicativity of the determinant and the nonnegativity of Gram determinants (Theorem 7.5.4, Corollary 7.7.3). From Chapter 8 §01, one fact, used four times in §08: a product of partitioned matrices multiplies blockwise (Theorem 8.1.1). Chapter 9 is used only in examples and in one classification (Definition 9.2.1); nothing before Section 7 needs an eigenvalue.

11.0.1 Roadmap

- **Inner products and norms.** The axioms, abstracted from the dot product; length, orthogonality, Pythagoras, Cauchy–Schwarz, and the triangle inequality.
- **Orthonormal bases and Gram–Schmidt.** Coordinates that cost one inner product each, the process that builds such a basis from any basis, and Gram matrices.
- **Orthogonal complements and projections.** The splitting $V = U \oplus U^{\perp}$, the projection P_U, and the fact that $P_U v$ is the point of U nearest to v.
- **Least squares and minimum-norm solutions.** What to do with an inconsistent system: the normal equations, and the smallest solution of an underdetermined one.
- **The Riesz representation theorem.** Every linear functional is an inner product with one fixed vector, which is the chosen isomorphism $V \cong V^*$ that Chapter 5 could not supply.
- **The adjoint.** Moving an operator to the other side of an inner product, the matrix $\mathbf{A}^*$, and the four fundamental subspaces in orthogonal form.
- **Isometries and unitary maps.** The maps that preserve all lengths, their many equivalent descriptions, the Frobenius norm and its blindness to unitary factors, and the classification of O(2).
- **QR, Householder, and Hessenberg.** Gram–Schmidt as a matrix factorization, reflections that do the same job more robustly, and the first step of a serious eigenvalue algorithm.
- **Bessel, Parseval, and Fourier series.** What an orthonormal list captures of a vector, and the classical Fourier coefficients as an instance.
- **Orthogonal polynomials.** Gram–Schmidt on $1, x, x^2, \ldots$, the three-term recurrence, real roots, and Gaussian quadrature.

11.0.2 Named moves

- **To show a vector is zero, pair it with itself.** Only 0 has $\langle z, z\rangle = 0$, so an identity of that form ends the argument.
- **To move T across an inner product, pay with T^*.** The adjoint is the price of the transfer, and most proofs about special operators compute one inner product in two ways.
- **Work with the norm squared, then expand.** Square roots do not expand; $\|u+v\|^2 = \langle u+v, u+v\rangle$ does.
- **Fix an orthonormal basis.** In its coordinates every inner product becomes the dot product, and an abstract statement becomes a matrix statement.
- **Project, then subtract.** The same move orthogonalizes a list (Gram–Schmidt), finds the nearest point of a subspace, and solves a least-squares problem.

11.1 Inner Products and Norms

Nine chapters, and we have not measured anything. We can say that a list is independent, that a subspace has dimension four, that an operator is diagonalizable. We cannot yet say that a vector has length 3, or that two vectors meet at a right angle. Geometry has appeared only in pictures, never in the theory. This section puts it into the theory, by abstracting the one operation in $\mathbb{R}^n$ that already computes lengths and angles: the dot product.

11.1.1 The field for this chapter

Throughout Chapter 11 the field F is $\mathbb{R}$ or $\mathbb{C}$. This is not laziness; two features of these fields are used on every page, and no other field has both.

The first is **positivity**. The point of the whole construction is that $\langle v, v\rangle$ should be the square of a length, so it must be a number we can compare with 0. Over $\mathbb{F}_p$ there is no ordering compatible with the arithmetic at all: in $\mathbb{F}_5$, adding 1 to itself five times returns to 0, so no nonzero element can be called "positive" in any useful sense. Over $\mathbb{Q}$ there is an ordering, but lengths would not exist, since $\sqrt{2} \notin \mathbb{Q}$. Over $\mathbb{R}$ every sum of squares is ≥ 0 and every nonnegative number has a square root, and that is exactly what we need.

The second is **conjugation**. Over $\mathbb{C}$ the naive sum of squares fails immediately: $1^2 + i^2 = 0$, so "$z_1^2 + z_2^2$" is not a length for $(1, i) \neq \mathbf{0}$. The repair is to use $z\overline{z} = |z|^2$ instead of z^2 (Theorem 0.6.2 (c)), and this needs the complex conjugate. We write $\overline{a}$ for the conjugate of $a \in F$, with the convention that $\overline{a} = a$ when $F = \mathbb{R}$; every formula below then reads correctly over both fields at once.

11.1.2 From the dot product to the axioms

In $\mathbb{R}^n$ the dot product is

$$\mathbf{x} \cdot \mathbf{y} = x_1 y_1 + \cdots + x_n y_n.$$

It computes lengths, because $\mathbf{x} \cdot \mathbf{x} = x_1^2 + \cdots + x_n^2$ is the square of the Euclidean length by Pythagoras. It computes angles too. For nonzero $\mathbf{x}, \mathbf{y} \in \mathbb{R}^2$ at an angle θ, the law of cosines applied to the triangle with sides $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{x} - \mathbf{y}$ says

$$\|\mathbf{x} - \mathbf{y}\|^2 = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 - 2\|\mathbf{x}\|\|\mathbf{y}\| \cos\theta,$$

while expanding the left side entry by entry gives $\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 - 2(\mathbf{x} \cdot \mathbf{y})$. Comparing the two, $\mathbf{x} \cdot \mathbf{y} = \|\mathbf{x}\|\|\mathbf{y}\| \cos\theta$. So one bilinear expression carries both length and angle, and perpendicularity is the single equation $\mathbf{x} \cdot \mathbf{y} = 0$.

Which properties of the dot product did that argument use? Only three: it is linear in each slot separately, it is symmetric, and $\mathbf{x} \cdot \mathbf{x} > 0$ for $\mathbf{x} \neq \mathbf{0}$. Those three we can demand of an abstract space. Over $\mathbb{C}$, as we saw, symmetry has to be weakened to conjugate symmetry, and the price is small: a conjugate bar in the second slot.

An inner product is a way of multiplying two vectors into a scalar so that a vector paired with itself has positive "length squared".

Definition 11.1.1. Inner Product

Let F be $\mathbb{R}$ or $\mathbb{C}$, and let V be a vector space **over** F. An **inner product** on V is a map

$$\langle \cdot, \cdot \rangle : V \times V \to F$$

such that for all $u, v, w \in V$ and all $a, b \in F$:

(IP1) $\langle au + bv, w\rangle = a\langle u, w\rangle + b\langle v, w\rangle$ (**linear in the first slot**);
(IP2) $\langle v, u\rangle = \overline{\langle u, v\rangle}$ (**conjugate symmetry**);
(IP3) $\langle v, v\rangle > 0$ for **every nonzero** $v \in V$ (**positive definiteness**).

A vector space equipped with an inner product is an **inner product space**. When $F = \mathbb{R}$ it is a **real** inner product space, when $F = \mathbb{C}$ a **complex** one.

In words, clause by clause. (IP1) says that if we freeze the second slot, what is left is a linear functional on V: the inner product distributes over sums and lets scalars out of the **first** argument untouched. (IP2) says that swapping the two arguments conjugates the answer; over

$\mathbb{R}$ it is plain symmetry, $\langle v, u\rangle = \langle u, v\rangle$. (IP3) says that a nonzero vector pairs with itself to give a **strictly** positive number. For the inequality in (IP3) to mean anything, $\langle v, v\rangle$ must be a **real** number; it is, and the reason is (IP2), as the next result shows. Note also what (IP3) does **not** say: it says nothing about $\langle u, v\rangle$ for $u \neq v$, which may be negative, or not even real.

The first consequences are small, and we will use all of them constantly.

Theorem 11.1.1. First Properties of an Inner Product

Let V be an inner product space over F. For all $u, v, w \in V$ and $a, b \in F$:

(a) $\langle v, v\rangle$ is a **real** number, and $\langle v, v\rangle \geq 0$ with equality if and only if $v = 0$;
(b) $\langle u, av + bw\rangle = \overline{a}\langle u, v\rangle + \overline{b}\langle u, w\rangle$ (**conjugate-linear in the second slot**);
(c) $\langle 0, v\rangle = \langle v, 0\rangle = 0$;
(d) if $\langle u, w\rangle = \langle v, w\rangle$ for **every** $w \in V$, then $u = v$.

Idea. Each part is one application of an axiom. For (a), a number equal to its own conjugate is real. For (b), swap slots with (IP2), use (IP1), and swap back. For (d), the only vector orthogonal to everything is the zero vector, and we find that out by pairing the difference **with itself**: that is the move to remember.

Proof.

(a) By (IP2) with $u = v$, $\langle v, v\rangle = \overline{\langle v, v\rangle}$, and a complex number equal to its own conjugate has zero imaginary part, so $\langle v, v\rangle \in \mathbb{R}$. If $v \neq 0$, then $\langle v, v\rangle > 0$ by (IP3). If $v = 0$, then $\langle 0, 0\rangle = \langle 0 \cdot 0, 0\rangle = 0\langle 0, 0\rangle = 0$ by (IP1). Hence $\langle v, v\rangle \geq 0$ always, with equality exactly when $v = 0$.

(b) Using (IP2), then (IP1), then (IP2) again,

$$\begin{aligned}\langle u, av + bw\rangle &= \overline{\langle av + bw, u\rangle} \\ &= \overline{a\langle v, u\rangle + b\langle w, u\rangle} \\ &= \overline{a}\,\overline{\langle v, u\rangle} + \overline{b}\,\overline{\langle w, u\rangle} \\ &= \overline{a}\langle u, v\rangle + \overline{b}\langle u, w\rangle,\end{aligned}$$

where the third equality uses Theorem 0.6.2 (a) and (b).

(c) $\langle 0, v\rangle = \langle 0 \cdot 0, v\rangle = 0\langle 0, v\rangle = 0$ by (IP1), and $\langle v, 0\rangle = \overline{\langle 0, v\rangle} = 0$ by (IP2).

(d) Put $z = u - v$. For every w, (IP1) gives $\langle z, w\rangle = \langle u, w\rangle - \langle v, w\rangle = 0$. Take $w = z$: then $\langle z, z\rangle = 0$, so $z = 0$ by (a), that is, $u = v$.

■

Part (d) is the first appearance of the move that runs through the whole chapter: **to show a vector is zero, pair it with itself.** There is exactly one vector of length zero, so an identity of the form $\langle z, z\rangle = 0$ is a proof that $z = 0$.

Part (b) is where the conjugate becomes visible in practice. An inner product over $\mathbb{C}$ is **not** bilinear: scalars leave the second slot wearing a bar. A map with properties (IP1) and (b) is called **sesquilinear**, from the Latin for "one and a half times linear".

11.1.3 Examples

The dot product on $\mathbb{R}^n$. Define $\langle \mathbf{x}, \mathbf{y} \rangle = x_1 y_1 + \cdots + x_n y_n$. (IP1) holds because each term is linear in x_i; (IP2) is the commutativity of multiplication in $\mathbb{R}$, and conjugation is the identity; (IP3) holds because $\langle \mathbf{x}, \mathbf{x} \rangle = x_1^2 + \cdots + x_n^2$ is a sum of squares of real numbers, which is 0 only when every $x_i = 0$. This is the **standard** inner product on $\mathbb{R}^n$, and it is the model for everything else.

The standard inner product on $\mathbb{C}^n$. Define

$$\langle \mathbf{x}, \mathbf{y} \rangle = x_1 \overline{y_1} + \cdots + x_n \overline{y_n}.$$

(IP1) is clear. For (IP2), $\overline{\langle \mathbf{x}, \mathbf{y} \rangle} = \sum \overline{x_i} y_i = \langle \mathbf{y}, \mathbf{x} \rangle$ by Theorem 0.6.2 (a), (b). For (IP3), $\langle \mathbf{x}, \mathbf{x} \rangle = \sum x_i \overline{x_i} = \sum |x_i|^2$ by Theorem 0.6.2 (c), a sum of nonnegative **real** numbers, zero only when every $x_i = 0$. The conjugates are doing exactly one job: making the diagonal values real and nonnegative.

Weighted inner products. Fix real numbers $w_1, \ldots, w_n > 0$ and put $\langle \mathbf{x}, \mathbf{y} \rangle = \sum_i w_i x_i \overline{y_i}$ on F^n. The same three checks go through, with $\langle \mathbf{x}, \mathbf{x} \rangle = \sum w_i |x_i|^2$; the weights must be **strictly** positive, or (IP3) fails for a standard basis vector. In the degenerate case $n = 1$ and $w_1 = 1$, this is $\langle x, y \rangle = x\overline{y}$ on F itself, the smallest nonzero inner product space there is. It matters because it is where the axioms are just the arithmetic of $|\cdot|$.

Continuous functions. Let $V = C[0,1]$ be the space of continuous functions $f\colon [0,1] \to F$, and set

$$\langle f, g \rangle = \int_0^1 f(t)\overline{g(t)}\, \mathrm{d}t.$$

(IP1) is linearity of the integral, and (IP2) follows because conjugation commutes with integration. For (IP3), $\langle f, f \rangle = \int_0^1 |f(t)|^2 \, \mathrm{d}t \geq 0$, and we must rule out equality for $f \neq 0$. Here is the one fact from analysis we assume: if f is continuous and $f(c) \neq 0$ for some c, then $|f| \geq |f(c)|/2$ on some interval of positive length δ around c, so $\int_0^1 |f|^2 \geq \delta |f(c)|^2/4 > 0$. Continuity is not decoration; the nonexample below shows what happens without it.

Matrices: the Frobenius inner product. First a piece of notation we will need all chapter.

Definition 11.1.2. Conjugate Transpose

For $\mathbf{A} \in M_{m\times n}(F)$, the **conjugate transpose** $\mathbf{A}^* \in M_{n\times m}(F)$ is the matrix with entries $(\mathbf{A}^*)_{ij} = \overline{a_{ji}}$. A square matrix with $\mathbf{A}^* = \mathbf{A}$ is **Hermitian**.

Over $\mathbb{R}$, $\mathbf{A}^* = \mathbf{A}^\top$ and Hermitian means symmetric (Definition 0.9.11). Now on $V = M_{m\times n}(F)$ define

$$\langle \mathbf{A}, \mathbf{B} \rangle = \operatorname{tr}(\mathbf{B}^* \mathbf{A}).$$

Writing out the trace, $\operatorname{tr}(\mathbf{B}^*\mathbf{A}) = \sum_j \sum_i \overline{b_{ij}} a_{ij}$, so this is the standard inner product of F^{mn} after listing the entries in a column. In particular (IP1)–(IP3) hold, with $\langle \mathbf{A}, \mathbf{A} \rangle = \sum_{i,j} |a_{ij}|^2$. This is the **Frobenius inner product**; for $n = 1$ it is the standard inner product on F^m.

Polynomials sampled at nodes. Fix $n+1$ **distinct** scalars $c_0, \ldots, c_n \in F$ and define, on $V = F[x]_{\leq n}$,

$$\langle p, q \rangle = \sum_{i=0}^{n} p(c_i)\overline{q(c_i)}.$$

(IP1) and (IP2) hold because evaluation at c_i is linear (Theorem 0.8.5) and the standard inner product on F^{n+1} has them; indeed $\langle p, q \rangle$ is the standard inner product of the sam-

ple vectors $(p(c_0), \dots, p(c_n))$ and $(q(c_0), \dots, q(c_n))$. Positivity is the interesting clause. If $\langle p, p\rangle = \sum |p(c_i)|^2 = 0$, then p vanishes at all $n+1$ distinct nodes. A nonzero polynomial of degree at most n has at most n distinct roots (Corollary 6.1.4), so $p = 0$. The **number** of nodes is what makes this work, and the nonexample below removes one.

11.1.4 Nonexamples, by minimal change

Change a sign: the hyperbolic form on $\mathbb{R}^2$. Let $\beta(\mathbf{x}, \mathbf{y}) = x_1y_1 - x_2y_2$. Bilinearity survives, since each term is still linear in each slot, and so does symmetry. What fails is (IP3), and the witness is $\mathbf{e}_2$: $\beta(\mathbf{e}_2, \mathbf{e}_2) = -1 < 0$. The vector $(1, 1)$ is worse still: it is nonzero, yet $\beta((1, 1), (1, 1)) = 0$, so it would be a nonzero vector of length 0. Forms like β are useful (they are the subject of the chapter on bilinear forms), but they are not inner products.

Drop the conjugate on $\mathbb{C}^2$. Let $\langle \mathbf{u}, \mathbf{v}\rangle = \mathbf{u}^\top \mathbf{v} = u_1v_1 + u_2v_2$ on $\mathbb{C}^2$. This map is linear in **both** slots and symmetric, so it satisfies (IP1) and looks tidier than the standard product. Take $\mathbf{u} = (1, i)$. Then

$$\langle \mathbf{u}, \mathbf{u}\rangle = 1^2 + i^2 = 1 - 1 = 0,$$

with $\mathbf{u} \neq \mathbf{0}$. Clause (IP3) fails. This one example is the whole reason for the conjugate: over $\mathbb{C}$, sums of squares vanish on nonzero vectors, sums of squared **moduli** do not.

Enlarge the space: drop continuity. Keep the formula $\langle f, g\rangle = \int_0^1 fg$, but let V be all Riemann-integrable functions $[0, 1] \to \mathbb{R}$. Linearity and symmetry are untouched. Let f be 1 at $t = \frac{1}{2}$ and 0 elsewhere. It is integrable, nonzero as a function, and $\int_0^1 f^2 = 0$. Positivity fails, so $\langle \cdot, \cdot\rangle$ is only **positive semidefinite** here. There are two standard repairs: restrict to continuous functions, as above, or regard two functions as equal when they differ on a negligible set. We take the first, and every function space in this chapter will be a space of continuous functions.

Remove a node. On $\mathbb{R}[x]_{\leq 2}$, put $\langle p, q\rangle = p(0)q(0) + p(1)q(1)$: the node example with one node too few. Bilinearity and symmetry hold. But $p = x^2 - x$ is a nonzero polynomial of degree 2 with $p(0) = p(1) = 0$, so $\langle p, p\rangle = 0$. The clause that fails is (IP3), and it fails because two nodes cannot pin down a quadratic.

Why this definition. Each clause earns its place. Without (IP3) there can be nonzero vectors v with $\langle v, v\rangle = 0$, and the proof of Theorem 11.1.1 (d) breaks down, because pairing a vector with itself no longer detects zero. (The form $x_1y_1 - x_2y_2$ below shows that the conclusion of (d) may survive anyway; what is lost is the argument for it, and with it every later proof that runs through $\langle v, v\rangle = 0 \Rightarrow v = 0$.) Without (IP2) the diagonal values need not be real, and "$\langle v, v\rangle > 0$" is meaningless. Linearity in the first slot rather than the second is a convention, but once chosen it must be kept: with (IP2) in place, linearity in one slot forces conjugate-linearity in the other, so a map cannot be linear in both unless $F = \mathbb{R}$.

Warning. **The conjugate goes in the second slot here, and books differ.** With our convention, $\langle au, v\rangle = a\langle u, v\rangle$ and $\langle u, av\rangle = \overline{a}\langle u, v\rangle$. Many books, especially in physics, put the conjugate in the first slot, so that their $\langle u, v\rangle$ is our $\langle v, u\rangle$. Every formula involving a lone conjugate then flips. When you read a formula elsewhere, check the convention first; when you write one, say which you use.

Quick check. Is $\langle \mathbf{x}, \mathbf{y}\rangle = x_1y_1 + x_2y_2 + x_1y_2$ an inner product on $\mathbb{R}^2$? If not, name the clause that fails and give a witness.

Solution. No. It is bilinear, and $\langle \mathbf{x}, \mathbf{x} \rangle = x_1^2 + x_2^2 + x_1 x_2 = (x_1 + \tfrac{1}{2}x_2)^2 + \tfrac{3}{4}x_2^2 > 0$ for $\mathbf{x} \neq \mathbf{0}$, so (IP1) and (IP3) both hold. The clause that fails is (IP2): with $\mathbf{u} = \mathbf{e}_1$ and $\mathbf{v} = \mathbf{e}_2$ we get $\langle \mathbf{u}, \mathbf{v} \rangle = 1$ but $\langle \mathbf{v}, \mathbf{u} \rangle = 0$. Symmetrizing the cross term, to $x_1 y_1 + x_2 y_2 + \tfrac{1}{2}(x_1 y_2 + x_2 y_1)$, repairs it.

11.1.5 Norms

By Theorem 11.1.1 (a), the number $\langle v, v \rangle$ is real and ≥ 0, so it has a unique nonnegative real square root. That square root is the length we were after.

Definition 11.1.3. Induced Norm

Let V be an inner product space. The **norm** of $v \in V$ is

$$\|v\| \coloneqq \sqrt{\langle v, v \rangle},$$

the **nonnegative** real square root. A vector with $\|v\| = 1$ is a **unit vector**, and replacing a nonzero v by $v/\|v\|$ is called **normalizing** it.

On $\mathbb{R}^n$ with the dot product this is the Euclidean length $\sqrt{x_1^2 + \cdots + x_n^2}$, and on $\mathbb{C}^n$ it is $\sqrt{\sum |x_i|^2}$. On $C[0,1]$ it is $\left(\int_0^1 |f|^2\right)^{1/2}$, and on matrices it is the **Frobenius norm** $\left(\sum |a_{ij}|^2\right)^{1/2}$. By definition $\|v\|^2 = \langle v, v \rangle$ always, and this identity, read from right to left, is the most-used step in the chapter: **work with the norm squared, then expand.**

Theorem 11.1.2. Properties of the Induced Norm

Let V be an inner product space over F, $u, v \in V$ and $c \in F$. Then:

(a) $\|v\| \geq 0$, with $\|v\| = 0$ if and only if $v = 0$;
(b) $\|cv\| = |c|\|v\|$;
(c) $\|u + v\|^2 = \|u\|^2 + 2\,\mathrm{Re}\langle u, v \rangle + \|v\|^2$.

Proof.

(a) Immediate from Theorem 11.1.1 (a) and the fact that a nonnegative real number has square root 0 only when it is 0.

(b) By (IP1) and Theorem 11.1.1 (b), $\langle cv, cv \rangle = c\overline{c}\langle v, v \rangle = |c|^2\|v\|^2$ using Theorem 0.6.2 (c). Taking nonnegative square roots gives $\|cv\| = |c|\|v\|$, since $|c| \geq 0$.

(c) Expand one slot at a time, by (IP1) and then Theorem 11.1.1 (b):

$$\begin{aligned} \|u + v\|^2 &= \langle u + v, u + v \rangle \\ &= \langle u, u \rangle + \langle u, v \rangle + \langle v, u \rangle + \langle v, v \rangle. \end{aligned}$$

By (IP2), $\langle v, u \rangle = \overline{\langle u, v \rangle}$, and $z + \overline{z} = 2\,\mathrm{Re}\, z$ by Theorem 0.6.2 (c). This gives the stated identity.

■

Part (c) is the workhorse. Over $\mathbb{R}$ the middle term is just $2\langle u, v\rangle$, and the identity is the familiar $(a+b)^2 = a^2+2ab+b^2$ for vectors. Replacing v by $-v$ gives $\|u-v\|^2 = \|u\|^2 - 2\,\mathrm{Re}\langle u, v\rangle + \|v\|^2$, and adding the two versions gives the **parallelogram law**

$$\|u+v\|^2 + \|u-v\|^2 = 2\|u\|^2 + 2\|v\|^2. \tag{$*$}$$

We return to $(*)$ in the last warning of this section.

Warning. **A squared norm is real and nonnegative, never complex.** If a computation in a complex space produces $\|\mathbf{v}\|^2 = 2i$, or a negative number, something has gone wrong — most often a missing conjugate. For $\mathbf{v} = (1, i) \in \mathbb{C}^2$, the correct value is $\|\mathbf{v}\|^2 = |1|^2+|i|^2 = 2$, while the conjugate-free expression $\mathbf{v}^{\top}\mathbf{v}$ gives 0. Use the answer as a check on the work.

11.1.6 Orthogonality and Pythagoras

With a norm in hand, the special relation $\langle u, v\rangle = 0$ deserves its name.

Definition 11.1.4. Orthogonal Vectors

Let V be an inner product space. Vectors $u, v \in V$ are **orthogonal**, written $u \perp v$, if $\langle u, v\rangle = 0$. A list $(v_1, \dots, v_k)$ is **orthogonal** if $v_i \perp v_j$ whenever $i \neq j$.

The relation is symmetric: $\langle u, v\rangle = 0$ if and only if $\langle v, u\rangle = \overline{\langle u, v\rangle} = 0$, so "$u \perp v$" needs no order. The zero vector is orthogonal to everything (Theorem 11.1.1 (c)), and by part (a) of the same result it is the **only** vector orthogonal to itself. In $\mathbb{R}^n$ this is perpendicularity in the usual sense, by the law of cosines computation above: $\cos\theta = 0$ exactly when $\mathbf{x}\cdot\mathbf{y} = 0$. In $\mathbb{C}^2$, the vectors $(1, i)$ and $(1, -i)$ are orthogonal, since $1\cdot\overline{1} + i\cdot\overline{(-i)} = 1 + i\cdot i = 0$.

Theorem 11.1.3. Pythagorean Theorem

Let $(v_1, \dots, v_k)$ be an orthogonal list in an inner product space. Then

$$\|v_1 + \cdots + v_k\|^2 = \|v_1\|^2 + \cdots + \|v_k\|^2.$$

Proof. Expand the left side by (IP1) and Theorem 11.1.1 (b):

$$\Big\langle \sum_i v_i,\ \sum_j v_j \Big\rangle = \sum_i \sum_j \langle v_i, v_j\rangle.$$

Every term with $i \neq j$ is 0, because the list is orthogonal. What survives is $\sum_i \langle v_i, v_i\rangle = \sum_i \|v_i\|^2$, as claimed. ∎

So orthogonality is exactly the condition under which squared lengths add. This is why the relation is worth naming, and why the rest of the chapter works so hard to produce orthogonal vectors.

Warning. **Over $\mathbb{C}$, the converse of Pythagoras fails.** By Theorem 11.1.2 (c), $\|u+v\|^2 = \|u\|^2 + \|v\|^2$ says only that $\operatorname{Re}\langle u, v\rangle = 0$, not that $\langle u, v\rangle = 0$. In $\mathbb{C}$ with $\langle z, w\rangle = z\overline{w}$, take $u = 1$ and $v = i$: then $\langle u, v\rangle = \bar{i} = -i \neq 0$, yet $\|1 + i\|^2 = 2 = \|1\|^2 + \|i\|^2$. Over $\mathbb{R}$ the converse does hold, since there Re does nothing.

Theorem 11.1.4. Orthogonal Nonzero Vectors Are Independent

Let $(v_1, \dots, v_k)$ be an orthogonal list of **nonzero** vectors in an inner product space. Then the list is linearly independent.

Idea. Start a dependence relation in the standard way, and then do the one thing an inner product allows: pair the relation with a single v_j. Orthogonality kills every term except the j-th, and what is left is $a_j\|v_j\|^2 = 0$ with $\|v_j\| \neq 0$.

Proof. Let $a_1v_1 + \cdots + a_kv_k = 0$ with $a_1, \dots, a_k \in F$. Fix j and take the inner product of both sides with v_j. By (IP1) and Theorem 11.1.1 (c),

$$0 = \Big\langle \sum_i a_i v_i,\ v_j \Big\rangle = \sum_i a_i \langle v_i, v_j\rangle = a_j\|v_j\|^2,$$

where the last equality holds because $\langle v_i, v_j\rangle = 0$ for $i \neq j$. Since $v_j \neq 0$, Theorem 11.1.1 (a) gives $\|v_j\|^2 > 0$, so $a_j = 0$. As j was arbitrary, all the coefficients vanish, and the list is linearly independent by Definition 1.4.1.

■

The hypothesis "nonzero" cannot be dropped: the list $(\mathbf{e}_1, \mathbf{0})$ is orthogonal and dependent. With it, the result is a free supply of independent lists, and this is what makes the next section possible.

11.1.7 The Cauchy–Schwarz inequality

Angles came from the formula $\cos\theta = (\mathbf{x}\cdot\mathbf{y})/(\|\mathbf{x}\|\|\mathbf{y}\|)$. To define an angle by that formula in an abstract space, we must know that the right-hand side lies in $[-1, 1]$. That is exactly the following inequality, the most used inequality in the subject.

Theorem 11.1.5. Cauchy–Schwarz Inequality

Let V be an inner product space and $u, v \in V$. Then

$$|\langle u, v\rangle| \leq \|u\|\,\|v\|,$$

with equality if and only if the list (u, v) is linearly dependent.

Idea. Picture u and v in $\mathbb{R}^2$. Drop a perpendicular from the tip of u onto the line through v. The foot of the perpendicular is a multiple cv, and the two legs cv and $u - cv$ of the

resulting right triangle have squared lengths adding to $\|u\|^2$. Since one leg has squared length $|\langle u, v\rangle|^2/\|v\|^2$, the inequality is just "a leg is no longer than the hypotenuse", and equality means the other leg is 0. The picture proves nothing in an abstract space, so we check algebraically that $u - cv \perp v$ for the right c, and then let Pythagoras do the work.

Proof. If $v = 0$, both sides are 0 by Theorem 11.1.1 (c), and $(u, 0)$ is dependent, so the statement holds. Suppose $v \neq 0$, so $\|v\| > 0$, and put

$$c = \frac{\langle u, v\rangle}{\|v\|^2}, \qquad z = u - cv.$$

Then $z \perp v$, since by (IP1)

$$\langle z, v\rangle = \langle u, v\rangle - c\langle v, v\rangle = \langle u, v\rangle - \langle u, v\rangle = 0,$$

and hence also $z \perp cv$. Now $u = z + cv$ is a sum of two orthogonal vectors, so Theorem 11.1.3 and Theorem 11.1.2 (b) give

$$\|u\|^2 = \|z\|^2 + |c|^2\|v\|^2 = \|z\|^2 + \frac{|\langle u, v\rangle|^2}{\|v\|^2}. \tag{$\dagger$}$$

Since $\|z\|^2 \geq 0$, dropping it gives $\|u\|^2\|v\|^2 \geq |\langle u, v\rangle|^2$, and taking nonnegative square roots proves the inequality.

Equality holds in ($\dagger$) exactly when $\|z\| = 0$, that is (Theorem 11.1.2 (a)) exactly when $u = cv$. If $u = cv$, the list (u, v) is dependent. Conversely, suppose (u, v) is dependent, say $au + bv = 0$ with a, b not both 0. Here $a \neq 0$, since $a = 0$ would give $bv = 0$ with $b \neq 0$ and hence $v = 0$, which we have excluded. So $u = dv$ with $d = -b/a$, and then $c = d\langle v, v\rangle/\|v\|^2 = d$, so $z = 0$. This proves the equality case.

■

The proof is worth rereading for its shape: the vector z was **found** from a picture and then **checked** from the axioms. That is the standard way to work in an abstract inner product space.

Corollary 11.1.6. Triangle Inequality

Let V be an inner product space and $u, v \in V$. Then

$$\|u + v\| \leq \|u\| + \|v\|.$$

Proof. For any complex number z we have $\operatorname{Re} z \leq |z|$, as in the proof of Theorem 0.6.4. Hence, by Theorem 11.1.2 (c) and then Theorem 11.1.5,

$$\begin{aligned} \|u + v\|^2 &= \|u\|^2 + 2\operatorname{Re}\langle u, v\rangle + \|v\|^2 \\ &\leq \|u\|^2 + 2|\langle u, v\rangle| + \|v\|^2 \\ &\leq \|u\|^2 + 2\|u\|\|v\| + \|v\|^2 \\ &= \big(\|u\| + \|v\|\big)^2. \end{aligned}$$

Both $\|u+v\|$ and $\|u\|+\|v\|$ are ≥ 0, so taking square roots preserves the inequality.

■

Together, Theorem 11.1.2 (a), (b), and Corollary 11.1.6 say that $\|\cdot\|$ is a **norm** in the general sense studied in Chapter 16. Chapter 16 also shows that not every norm arises from an inner product; the last warning of this section explains how to tell.

Definition 11.1.5. Angle Between Vectors

Let V be a **real** inner product space and let $u, v \in V$ be **nonzero**. The **angle** between u and v is the unique $\theta \in [0, \pi]$ with

$$\cos\theta = \frac{\langle u, v\rangle}{\|u\|\,\|v\|}.$$

The definition needs Theorem 11.1.5 twice over. Cauchy–Schwarz says the quotient has absolute value at most 1, so it lies in $[-1, 1]$; and cos is a strictly decreasing bijection from $[0, \pi]$ onto $[-1, 1]$, so exactly one θ qualifies. Without the inequality the formula could ask for the angle whose cosine is 7. Note also that $\theta = \pi/2$ if and only if $\langle u, v\rangle = 0$, so the definition is consistent with Definition 11.1.4.

Over $\mathbb{C}$ there is no angle, because $\langle u, v\rangle$ is not real. The quantity $|\langle u, v\rangle|/(\|u\|\|v\|) \in [0, 1]$ still makes sense, and it still measures how far the two vectors are from being orthogonal.

Example. Three Angles

Compute the angle between $\mathbf{u} = (1, 1, 0)$ and $\mathbf{v} = (1, 0, 1)$ in $\mathbb{R}^3$ with the dot product; between $p = 1$ and $q = x$ in $\mathbb{R}[x]_{\leq 1}$ with $\langle p, q\rangle = \int_0^1 pq$; and between $\mathbf{A} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$ and $\mathbf{B} = \begin{pmatrix} 1 & -1 \\ 1 & -1 \end{pmatrix}$ in $M_2(\mathbb{R})$ with the Frobenius inner product.

Solution. In $\mathbb{R}^3$: $\langle \mathbf{u}, \mathbf{v}\rangle = 1$, $\|\mathbf{u}\| = \|\mathbf{v}\| = \sqrt{2}$, so $\cos\theta = 1/2$ and $\theta = \pi/3$.

In $\mathbb{R}[x]_{\leq 1}$: $\langle 1, x\rangle = \int_0^1 t\,\mathrm{d}t = \tfrac{1}{2}$, $\|1\|^2 = \int_0^1 1 = 1$, and $\|x\|^2 = \int_0^1 t^2\,\mathrm{d}t = \tfrac{1}{3}$. Hence

$$\cos\theta = \frac{1/2}{1 \cdot 1/\sqrt{3}} = \frac{\sqrt{3}}{2},$$

so $\theta = \pi/6$. The constant 1 and the function x are far from orthogonal on $[0, 1]$, which is believable: both are positive there.

In $M_2(\mathbb{R})$: $\langle \mathbf{A}, \mathbf{B}\rangle = 1 - 1 + 1 - 1 = 0$, so $\theta = \pi/2$. The two matrices are orthogonal, with $\|\mathbf{A}\| = \|\mathbf{B}\| = 2$. As a check on Pythagoras, $\mathbf{A} + \mathbf{B} = \begin{pmatrix} 2 & 0 \\ 2 & 0 \end{pmatrix}$ has $\|\mathbf{A} + \mathbf{B}\|^2 = 8 = 4 + 4$.

Quick check. In $\mathbb{C}^2$ with the standard inner product, compute $\langle (1,i),(1,1)\rangle$ and $\langle (1,1),(1,i)\rangle$. Check the answers against (IP2), and against Cauchy–Schwarz.

Solution. $\langle (1,i),(1,1)\rangle = 1\cdot\overline{1}+i\cdot\overline{1} = 1+i$, and $\langle (1,1),(1,i)\rangle = 1\cdot\overline{1}+1\cdot\overline{i} = 1-i$. They are conjugates, as (IP2) requires. For Cauchy–Schwarz, $|1+i| = \sqrt{2}$, while $\|(1,i)\| = \sqrt{2}$ and $\|(1,1)\| = \sqrt{2}$, so the inequality reads $\sqrt{2} \le 2$. It is strict, which is right: $(1,i)$ and $(1,1)$ are independent.

Warning. **Not every norm comes from an inner product.** A function $\|\cdot\|$ can satisfy positivity, homogeneity, and the inequality of Corollary 11.1.6 without being $\sqrt{\langle v,v\rangle}$ for any inner product. The test is the parallelogram law $(*)$: every induced norm satisfies it, as we derived from Theorem 11.1.2 (c), so a norm that violates it is induced by nothing. Take $\|\mathbf{x}\|_\infty = \max(|x_1|,|x_2|)$ on $\mathbb{R}^2$ with $\mathbf{u}=\mathbf{e}_1$, $\mathbf{v}=\mathbf{e}_2$. Then $\|\mathbf{u}+\mathbf{v}\|_\infty = \|\mathbf{u}-\mathbf{v}\|_\infty = 1$, so the left side of $(*)$ is 2, while the right side is 4. Remarkably, the law is not only necessary but sufficient; Exercise (C1: The parallelogram law characterizes inner-product norms) proves the substance of that in $\mathbb{R}^2$, and Chapter 16 takes up general norms.

11.1.8 Exercises

11.1.8.1 A. Check your understanding

Exercise. A1

(a) State the three axioms of an inner product on a vector space over $F = \mathbb{R}$ or $\mathbb{C}$, and say in which slot the map is linear.

(b) Why must the field be $\mathbb{R}$ or $\mathbb{C}$? Name the two features used.

(c) Explain why $\langle v,v\rangle$ is always a real number.

(d) True or false: an inner product on a complex space is linear in the second slot. Justify your answer.

(e) State the Cauchy–Schwarz inequality, its equality case, and one place where the equality case is used.

(f) True or false: if $\|u+v\|^2 = \|u\|^2+\|v\|^2$ in a complex inner product space, then $u \perp v$. Justify your answer.

Solution.

(a) (IP1) $\langle au+bv,w\rangle = a\langle u,w\rangle + b\langle v,w\rangle$; (IP2) $\langle v,u\rangle = \overline{\langle u,v\rangle}$; (IP3) $\langle v,v\rangle > 0$ for $v \neq 0$ (Definition 11.1.1). It is linear in the **first** slot.

(b) Positivity, so that $\langle v,v\rangle > 0$ has meaning and square roots exist, and conjugation, so that $\langle v,v\rangle$ is real and nonnegative over $\mathbb{C}$.

(c) By (IP2) with $u = v$, the number equals its own conjugate, so its imaginary part is 0 (Theorem 11.1.1 (a)).

(d) False. It is **conjugate**-linear in the second slot: $\langle u,av\rangle = \overline{a}\langle u,v\rangle$ (Theorem 11.1.1 (b)).

For example, in $\mathbb{C}$ with $\langle z, w\rangle = z\overline{w}$, $\langle 1, i\rangle = -i$ while $i\langle 1, 1\rangle = i$.

(e) $|\langle u, v\rangle| \leq \|u\|\|v\|$, with equality if and only if (u, v) is dependent (Theorem 11.1.5). The inequality is what makes Definition 11.1.5 well posed: it puts the quotient $\langle u, v\rangle/(\|u\|\|v\|)$ inside $[-1, 1]$, where arccos is defined. The equality case then says which vectors sit at the ends of that interval: $\cos\theta = \pm 1$ exactly for dependent pairs, that is, $\theta = 0$ or $\theta = \pi$.

(f) False. It says only $\mathrm{Re}\langle u, v\rangle = 0$. In $\mathbb{C}$ with $\langle z, w\rangle = z\overline{w}$, take $u = 1$, $v = i$: then $\langle u, v\rangle = -i \neq 0$ but $\|1 + i\|^2 = 2 = 1 + 1$.

11.1.8.2 B. Practice

Exercise. B1: Determine which are inner products

Determine which of the following are inner products on the stated space. For those that are, verify the axioms; for those that are not, name the clause that fails and give an explicit witness.

(a) On $\mathbb{R}^2$: $\langle \mathbf{x}, \mathbf{y}\rangle = x_1y_1 + 3x_2y_2$.

(b) On $\mathbb{R}^2$: $\langle \mathbf{x}, \mathbf{y}\rangle = x_1y_1 + x_1y_2 + x_2y_1 + 2x_2y_2$.

(c) On $\mathbb{C}^2$: $\langle \mathbf{x}, \mathbf{y}\rangle = x_1\overline{y_1} + x_2y_2$.

(d) On $M_2(\mathbb{R})$: $\langle \mathbf{A}, \mathbf{B}\rangle = \mathrm{tr}(\mathbf{AB})$.

(e) On $\mathbb{R}[x]_{\leq 2}$: $\langle p, q\rangle = p(0)q(0) + p(1)q(1) + p(2)q(2)$.

Solution.

(a) An inner product: it is the weighted product with weights $1, 3 > 0$. Bilinearity and symmetry are clear, and $\langle \mathbf{x}, \mathbf{x}\rangle = x_1^2 + 3x_2^2 = 0$ forces $x_1 = x_2 = 0$.

(b) An inner product. It is bilinear term by term, and symmetric because the two cross terms are interchanged by swapping $\mathbf{x}$ and $\mathbf{y}$. For positivity, complete the square:

$$\langle \mathbf{x}, \mathbf{x}\rangle = x_1^2 + 2x_1x_2 + 2x_2^2 = (x_1 + x_2)^2 + x_2^2,$$

which is 0 only if $x_2 = 0$ and $x_1 + x_2 = 0$, that is $\mathbf{x} = \mathbf{0}$.

(c) Not an inner product. (IP2) fails: with $\mathbf{x} = (0, 1)$ and $\mathbf{y} = (0, i)$ we get $\langle \mathbf{x}, \mathbf{y}\rangle = i$ and $\langle \mathbf{y}, \mathbf{x}\rangle = i$, while (IP2) would require $\langle \mathbf{y}, \mathbf{x}\rangle = \overline{i} = -i$. Positivity fails too, at $\mathbf{x} = (0, i)$: $\langle \mathbf{x}, \mathbf{x}\rangle = i^2 = -1$.

(d) Not an inner product. It is bilinear, and symmetric by Theorem 0.9.6 (3). But (IP3) fails for $\mathbf{N} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \neq \mathbf{O}$, since $\mathbf{N}^2 = \mathbf{O}$ and so $\langle \mathbf{N}, \mathbf{N}\rangle = \mathrm{tr}(\mathbf{N}^2) = 0$. Replacing $\mathrm{tr}(\mathbf{AB})$ by $\mathrm{tr}(\mathbf{B}^\top\mathbf{A})$ repairs it: that is the Frobenius inner product.

(e) An inner product: the node example with the three distinct nodes $0, 1, 2$ on $\mathbb{R}[x]_{\leq 2}$. Bilinearity and symmetry come from linearity of evaluation. If $\langle p, p\rangle = p(0)^2 + p(1)^2 + p(2)^2 = 0$, then p has three distinct roots and degree at most 2, so $p = 0$ by Corollary 6.1.4.

Exercise. B2: Norms and angles

(a) In $\mathbb{R}^3$ with the dot product, find $\|\mathbf{u}\|$, $\|\mathbf{v}\|$, and the cosine of the angle between $\mathbf{u} = (1, 2, 2)$ and $\mathbf{v} = (2, -1, 2)$.

(b) In $\mathbb{R}[x]_{\le 2}$ with $\langle p, q\rangle = \int_{-1}^{1} pq$, compute $\|1\|$, $\|x\|$, and $\langle 1, x\rangle$. What is the angle between 1 and x?

(c) In $M_2(\mathbb{C})$ with the Frobenius inner product, compute $\|\mathbf{A}\|$ for $\mathbf{A} = \begin{pmatrix} 1 & i \\ 0 & 1+i \end{pmatrix}$.

Solution.

(a) $\|\mathbf{u}\| = \sqrt{1+4+4} = 3$ and $\|\mathbf{v}\| = \sqrt{4+1+4} = 3$. Also $\langle \mathbf{u}, \mathbf{v}\rangle = 2 - 2 + 4 = 4$, so $\cos\theta = 4/9$. Cauchy–Schwarz reads $4 \le 9$, and it is strict, as it must be for an independent pair.

(b) $\|1\|^2 = \int_{-1}^{1} 1\,dt = 2$, so $\|1\| = \sqrt{2}$; $\|x\|^2 = \int_{-1}^{1} t^2\,dt = \frac{2}{3}$, so $\|x\| = \sqrt{2/3}$; and $\langle 1, x\rangle = \int_{-1}^{1} t\,dt = 0$. The angle is $\pi/2$: on a symmetric interval, the constants are orthogonal to the odd functions. Compare this with Example (Three Angles), where the interval $[0, 1]$ gave the angle $\pi/6$ instead. **The geometry depends on the inner product, not only on the vectors.**

(c) $\|\mathbf{A}\|^2 = |1|^2 + |i|^2 + 0 + |1+i|^2 = 1 + 1 + 2 = 4$, so $\|\mathbf{A}\| = 2$.

Exercise. B3: Cauchy–Schwarz, concretely

(a) Verify Theorem 11.1.5 for $\mathbf{a} = (1, 2, 3)$ and $\mathbf{b} = (3, 2, 1)$ in $\mathbb{R}^3$ with the dot product, and say whether the inequality is strict.

(b) Deduce from Theorem 11.1.5 the classical inequality

$$\Big(\sum_{i=1}^{n} a_i b_i\Big)^2 \le \Big(\sum_{i=1}^{n} a_i^2\Big)\Big(\sum_{i=1}^{n} b_i^2\Big)$$

for all real a_i, b_i.

(c) Hence show that $(a_1 + \cdots + a_n)^2 \le n(a_1^2 + \cdots + a_n^2)$ for all real a_i.

Solution.

(a) $\langle \mathbf{a}, \mathbf{b}\rangle = 3 + 4 + 3 = 10$, so $|\langle \mathbf{a}, \mathbf{b}\rangle|^2 = 100$, while $\|\mathbf{a}\|^2\|\mathbf{b}\|^2 = 14 \cdot 14 = 196$. So $10 \le 14$, strictly. It must be strict, since $\mathbf{a}$ and $\mathbf{b}$ are not multiples of each other.

(b) Apply Theorem 11.1.5 on $\mathbb{R}^n$ with the dot product to $\mathbf{a} = (a_1, \dots, a_n)$ and $\mathbf{b} = (b_1, \dots, b_n)$. Then $\langle \mathbf{a}, \mathbf{b}\rangle = \sum a_i b_i$, $\|\mathbf{a}\|^2 = \sum a_i^2$, and $\|\mathbf{b}\|^2 = \sum b_i^2$, and squaring the inequality $|\langle \mathbf{a}, \mathbf{b}\rangle| \le \|\mathbf{a}\|\|\mathbf{b}\|$ gives the claim.

(c) Take $b_i = 1$ for all i in (b). Then $\sum a_i b_i = \sum a_i$ and $\sum b_i^2 = n$, so $(\sum a_i)^2 \le n \sum a_i^2$. Equality holds exactly when $\mathbf{a}$ is a multiple of $(1, \dots, 1)$, that is, when all the a_i are equal.

Exercise. B4: A lower bound for a difference of norms

Let V be an inner product space and $u, v \in V$. Prove that

$$\big|\|u\| - \|v\|\big| \le \|u - v\|.$$

Solution. Write $u = (u - v) + v$. By Corollary 11.1.6,

$$\|u\| \le \|u - v\| + \|v\|,$$

so $\|u\| - \|v\| \le \|u - v\|$. Swapping the roles of u and v gives $\|v\| - \|u\| \le \|v - u\|$, and $\|v - u\| = |-1|\|u - v\| = \|u - v\|$ by Theorem 11.1.2 (b). The quantity $\big|\|u\| - \|v\|\big|$ is one of the two numbers just bounded, so it is at most $\|u - v\|$. This proves the inequality.

11.1.8.3 C. Going deeper

Exercise. C1: The parallelogram law characterizes inner-product norms

Let $\|\cdot\|$ be a norm on $\mathbb{R}^2$, that is, a function $\mathbb{R}^2 \to \mathbb{R}$ satisfying the three properties of Theorem 11.1.2 (a), (b), and Corollary 11.1.6. Suppose it satisfies the parallelogram law

$$\|\mathbf{u} + \mathbf{v}\|^2 + \|\mathbf{u} - \mathbf{v}\|^2 = 2\|\mathbf{u}\|^2 + 2\|\mathbf{v}\|^2$$

for all $\mathbf{u}, \mathbf{v}$, and define $\beta(\mathbf{u}, \mathbf{v}) = \frac{1}{4}\big(\|\mathbf{u} + \mathbf{v}\|^2 - \|\mathbf{u} - \mathbf{v}\|^2\big)$.

(a) Show that $\beta(\mathbf{u}, \mathbf{u}) = \|\mathbf{u}\|^2$, that $\beta(\mathbf{u}, \mathbf{v}) = \beta(\mathbf{v}, \mathbf{u})$, and that $\beta(\mathbf{u}, \mathbf{0}) = 0$.

(b) Prove that $\beta(\mathbf{u}, \mathbf{v}) + \beta(\mathbf{u}, \mathbf{w}) = 2\beta\big(\mathbf{u}, \frac{1}{2}(\mathbf{v} + \mathbf{w})\big)$ for all $\mathbf{u}, \mathbf{v}, \mathbf{w}$.

(c) Deduce that $\beta(\mathbf{u}, \mathbf{v}) + \beta(\mathbf{u}, \mathbf{w}) = \beta(\mathbf{u}, \mathbf{v} + \mathbf{w})$.

(d) Granting that $\beta(\mathbf{u}, t\mathbf{v}) = t\beta(\mathbf{u}, \mathbf{v})$ for all $t \in \mathbb{R}$, conclude that β is an inner product on $\mathbb{R}^2$ whose induced norm is $\|\cdot\|$.

Hint: for (b), apply the parallelogram law to the pair $\mathbf{u} + \frac{1}{2}(\mathbf{v} + \mathbf{w})$, $\frac{1}{2}(\mathbf{v} - \mathbf{w})$, *and then to the pair* $\mathbf{u} - \frac{1}{2}(\mathbf{v} + \mathbf{w})$, $\frac{1}{2}(\mathbf{w} - \mathbf{v})$.

Solution.

(a) $\beta(\mathbf{u}, \mathbf{u}) = \frac{1}{4}\big(\|2\mathbf{u}\|^2 - \|\mathbf{0}\|^2\big) = \frac{1}{4} \cdot 4\|\mathbf{u}\|^2 = \|\mathbf{u}\|^2$, using homogeneity. Symmetry holds because $\|\mathbf{v} - \mathbf{u}\| = \|\mathbf{u} - \mathbf{v}\|$, again by homogeneity with $c = -1$. Finally $\beta(\mathbf{u}, \mathbf{0}) = \frac{1}{4}\big(\|\mathbf{u}\|^2 - \|\mathbf{u}\|^2\big) = 0$.

(b) Put $\mathbf{m} = \frac{1}{2}(\mathbf{v} + \mathbf{w})$ and $\mathbf{d} = \frac{1}{2}(\mathbf{v} - \mathbf{w})$, so that $\mathbf{v} = \mathbf{m} + \mathbf{d}$ and $\mathbf{w} = \mathbf{m} - \mathbf{d}$. Apply the parallelogram law to $\mathbf{u} + \mathbf{m}$ and $\mathbf{d}$: since $(\mathbf{u} + \mathbf{m}) + \mathbf{d} = \mathbf{u} + \mathbf{v}$ and $(\mathbf{u} + \mathbf{m}) - \mathbf{d} = \mathbf{u} + \mathbf{w}$,

$$\|\mathbf{u} + \mathbf{v}\|^2 + \|\mathbf{u} + \mathbf{w}\|^2 = 2\|\mathbf{u} + \mathbf{m}\|^2 + 2\|\mathbf{d}\|^2.$$

Apply it to $\mathbf{u} - \mathbf{m}$ and $\mathbf{d}$: since $(\mathbf{u} - \mathbf{m}) - \mathbf{d} = \mathbf{u} - \mathbf{v}$ and $(\mathbf{u} - \mathbf{m}) + \mathbf{d} = \mathbf{u} - \mathbf{w}$,

$$\|\mathbf{u} - \mathbf{v}\|^2 + \|\mathbf{u} - \mathbf{w}\|^2 = 2\|\mathbf{u} - \mathbf{m}\|^2 + 2\|\mathbf{d}\|^2.$$

Subtracting the second from the first, the terms $2\|\mathbf{d}\|^2$ cancel, and dividing by 4 gives $\beta(\mathbf{u}, \mathbf{v}) + \beta(\mathbf{u}, \mathbf{w}) = 2\beta(\mathbf{u}, \mathbf{m})$, which is the claim.

(c) Take $\mathbf{w} = \mathbf{0}$ in (b). By (a), $\beta(\mathbf{u}, \mathbf{0}) = 0$, so $\beta(\mathbf{u}, \mathbf{v}) = 2\beta(\mathbf{u}, \frac{1}{2}\mathbf{v})$ for all $\mathbf{v}$. Applying this with $\mathbf{v} + \mathbf{w}$ in place of $\mathbf{v}$, we get $2\beta(\mathbf{u}, \frac{1}{2}(\mathbf{v} + \mathbf{w})) = \beta(\mathbf{u}, \mathbf{v} + \mathbf{w})$, and the left side is $\beta(\mathbf{u}, \mathbf{v}) + \beta(\mathbf{u}, \mathbf{w})$ by (b).

(d) By (c) and the granted homogeneity, β is linear in its second slot, and by the symmetry in (a) it is then linear in the first as well, so (IP1) and (IP2) hold (conjugation is trivial over $\mathbb{R}$). For (IP3), $\beta(\mathbf{u}, \mathbf{u}) = \|\mathbf{u}\|^2 > 0$ for $\mathbf{u} \neq \mathbf{0}$, since a norm vanishes only at $\mathbf{0}$. So β is an inner product, and its induced norm is $\sqrt{\beta(\mathbf{u}, \mathbf{u})} = \|\mathbf{u}\|$.

Remark. The granted step in (d) is where analysis enters: additivity from (c) gives $\beta(\mathbf{u}, t\mathbf{v}) = t\beta(\mathbf{u}, \mathbf{v})$ for every **rational** t by induction, and the extension to real t uses that $t \mapsto \beta(\mathbf{u}, t\mathbf{v})$ is continuous, which follows from continuity of the norm. Nothing else in the argument needs anything beyond the three norm properties.

Exercise. C2: Recovering the inner product from the norm

Let V be an inner product space over F.

(a) For $F = \mathbb{R}$, prove the polarization identity $\langle u, v\rangle = \frac{1}{4}(\|u + v\|^2 - \|u - v\|^2)$.

(b) For $F = \mathbb{C}$, prove that

$$\langle u, v\rangle = \frac{1}{4}\sum_{k=0}^{3} i^k \, \|u + i^k v\|^2.$$

(c) Deduce that two inner products on V with the same induced norm are equal.

Solution.

(a) By Theorem 11.1.2 (c) and its version with $-v$, and since Re is the identity on $\mathbb{R}$,

$$\|u + v\|^2 - \|u - v\|^2 = 4\langle u, v\rangle.$$

Dividing by 4 gives the identity.

(b) Write $z = \langle u, v\rangle$. By Theorem 11.1.2 (c) and Theorem 11.1.1 (b),

$$\|u + i^k v\|^2 = \|u\|^2 + \|v\|^2 + 2\,\mathrm{Re}\,(\overline{i^k} z),$$

where we used $\|i^k v\| = |i^k|\|v\| = \|v\|$. Now $\sum_{k=0}^{3} i^k = 1 + i - 1 - i = 0$, so the terms $\|u\|^2 + \|v\|^2$ contribute nothing. Writing $z = a + bi$ with $a, b \in \mathbb{R}$, the four remaining terms $2i^k\,\mathrm{Re}(\overline{i^k} z)$ are, for $k = 0, 1, 2, 3$,

$$2a, \quad 2ib, \quad 2a, \quad 2ib,$$

since $\mathrm{Re}(z) = a$, $\mathrm{Re}(-iz) = b$, $\mathrm{Re}(-z) = -a$ and $\mathrm{Re}(iz) = -b$. Their sum is $4a + 4bi = 4z$, and dividing by 4 gives the identity.

(c) Both (a) and (b) express $\langle u, v\rangle$ using only the norm. So if two inner products induce the same norm, the right-hand sides agree for all u, v, hence so do the left-hand sides.

Exercise. C3: The equality cases

Let V be an inner product space.

(a) Prove that $\|u+v\| = \|u\|+\|v\|$ if and only if one of u, v is a **nonnegative real** multiple of the other.

(b) Deduce the equality condition in the classical inequality of Exercise (B3: Cauchy–Schwarz, concretely) (b): for real a_i, b_i, equality holds if and only if one of the vectors $(a_1, \dots, a_n)$, $(b_1, \dots, b_n)$ is a multiple of the other.

(c) Give real numbers a_i, b_i, not all zero, with $\sum a_i b_i < 0$ and equality in the classical inequality, and explain why this does **not** contradict (a).

Solution.

(a) ($\Leftarrow$) If $v = 0$ both sides are $\|u\|$. If $u = tv$ with $t \geq 0$, then $\|u + v\| = (t + 1)\|v\| = t\|v\| + \|v\| = \|u\| + \|v\|$ by Theorem 11.1.2 (b).
($\Rightarrow$) Suppose $\|u + v\| = \|u\| + \|v\|$. Squaring and using Theorem 11.1.2 (c),

$$2\,\mathrm{Re}\langle u, v\rangle = 2\|u\|\|v\|.$$

The chain in the proof of Corollary 11.1.6 is therefore an equality throughout, so $\mathrm{Re}\langle u, v\rangle = |\langle u, v\rangle| = \|u\|\|v\|$. The second equality and Theorem 11.1.5 make (u, v) dependent. Assume $v \neq 0$ (otherwise $v = 0u$ and we are done) and write $u = cv$. Then $\langle u, v\rangle = c\|v\|^2$, and the first equality says $\mathrm{Re}(c)\|v\|^2 = |c|\|v\|^2$, so $\mathrm{Re}\, c = |c|$, which forces c to be real and ≥ 0.

(b) Apply Theorem 11.1.5 on $\mathbb{R}^n$ as in Exercise (B3: Cauchy–Schwarz, concretely) (b). Squaring removes the sign, so equality in the squared inequality is equality in $|\langle \mathbf{a}, \mathbf{b}\rangle| = \|\mathbf{a}\|\|\mathbf{b}\|$, which by Theorem 11.1.5 holds exactly when $(\mathbf{a}, \mathbf{b})$ is dependent, that is, when one of the two vectors is a multiple of the other.

(c) Take $n = 1$, $a_1 = 1$, $b_1 = -1$: then $\sum a_i b_i = -1$ and both sides of the classical inequality equal 1. There is no contradiction with (a), because (a) is about equality in Corollary 11.1.6, which needs a nonnegative multiple, while the classical inequality is the **squared** Cauchy–Schwarz inequality, which is blind to the sign. Here $\|\mathbf{a} + \mathbf{b}\| = 0 \neq 2 = \|\mathbf{a}\| + \|\mathbf{b}\|$.

11.2 Orthonormal Bases and Gram–Schmidt

A basis lets us write every vector in coordinates, but finding those coordinates costs work: in Definition 1.5.2 the scalars $a_1, \dots, a_n$ come from solving a linear system. In an inner product space one kind of basis makes that cost disappear. If the basis vectors are mutually orthogonal unit vectors, then the i-th coordinate of v is just $\langle v, e_i\rangle$, one inner product, computed without reference to the others. This section defines such bases, shows that every finite-dimensional inner product space has one, and gives the algorithm that produces them.

11.2.1 Orthonormal lists

The standard basis of $\mathbb{R}^n$ has two features beyond being a basis: distinct vectors are orthogonal, and each has length 1. The two conditions can be packed into a single formula using the Kronecker delta δ_{ij}, which is 1 when $i = j$ and 0 otherwise.

An orthonormal list is a list of unit vectors that are pairwise perpendicular.

Definition 11.2.1. Orthonormal List

Let V be an inner product space. A list $(e_1, \dots, e_k)$ of vectors of V is **orthonormal** if

$$\langle e_i, e_j \rangle = \delta_{ij} \qquad \text{for all } i, j \in \{1, \dots, k\},$$

that is, if $\langle e_i, e_j \rangle = 0$ whenever $i \neq j$ (**orthogonal**) and $\|e_i\| = 1$ for every i (**normalized**).

In words: the single equation $\langle e_i, e_j \rangle = \delta_{ij}$ carries two conditions at once, because the case $i \neq j$ says the vectors are orthogonal and the case $i = j$ says $\|e_i\|^2 = 1$. The list must be **ordered**, as every list in this book is, but nothing in the definition depends on the order. Note that no vector of an orthonormal list is 0, since $\|0\| = 0 \neq 1$; so by Theorem 11.1.4, **every orthonormal list is linearly independent**. That single sentence is why orthonormality is so useful: independence, which normally takes an argument, comes for free.

Definition 11.2.2. Orthonormal Basis

An **orthonormal basis** of an inner product space V is an orthonormal list that is a basis of V.

Examples.

- **The standard basis.** In F^n with the standard inner product, $\langle \mathbf{e}_i, \mathbf{e}_j \rangle = \delta_{ij}$ by inspection. This is the model, and the notation $\mathbf{e}_i$ is deliberately reused for orthonormal vectors in a general space.
- **A rotated basis of $\mathbb{R}^2$.** The list $\left(\frac{1}{\sqrt{2}}(1,1),\ \frac{1}{\sqrt{2}}(1,-1)\right)$ is orthonormal: the inner product of the two vectors is $\frac{1}{2}(1-1) = 0$, and each has norm $\frac{1}{\sqrt{2}}\sqrt{2} = 1$. Geometrically it is the standard basis turned through $\pi/4$.
- **A basis of $\mathbb{R}^3$.** The columns of $\frac{1}{3}\begin{pmatrix} 1 & 2 & 2 \\ 2 & 1 & -2 \\ 2 & -2 & 1 \end{pmatrix}$ form an orthonormal list: each has norm $\frac{1}{3}\sqrt{1+4+4} = 1$, and the three pairwise inner products are $\frac{1}{9}(2+2-4)$, $\frac{1}{9}(2-4+2)$ and $\frac{1}{9}(4-2-2)$, all 0. Being independent and of length $3 = \dim \mathbb{R}^3$, it is a basis (Theorem 1.6.4 (a)).
- **Matrix units.** In $M_{m\times n}(F)$ with the Frobenius inner product, the matrix units $\mathbf{E}_{ij}$ satisfy $\langle \mathbf{E}_{ij}, \mathbf{E}_{pq} \rangle = 1$ if $(i,j) = (p,q)$ and 0 otherwise, since only one entry of each is nonzero. So the standard basis of $M_{m\times n}(F)$ is orthonormal.
- **The Lagrange basis.** On $F[x]_{\le n}$ with the node inner product $\langle p, q \rangle = \sum_{i=0}^{n} p(c_i)\overline{q(c_i)}$ of the previous section, the Lagrange polynomials $\ell_0, \dots, \ell_n$ of Definition 6.6.1 vanish at every node but their own, where they take the value 1; that is, $\ell_i(c_j) = \delta_{ij}$. Hence $\langle \ell_i, \ell_j \rangle = \sum_k \ell_i(c_k)\overline{\ell_j(c_k)} = \delta_{ij}$: the Lagrange basis is an orthonormal basis for this inner product. An interpolation formula and an orthonormal expansion turn out to be the same statement.
- **The degenerate case.** The empty list is orthonormal, vacuously, and it is an orthonormal basis of $V = \{0\}$. This is not a joke: it is the base case of every induction below.

Nonexample by minimal change. In $\mathbb{R}^2$, the list $((1,1),(1,-1))$ is orthogonal, and it is a basis. It is **not** orthonormal, because $\|(1,1)\| = \sqrt{2} \neq 1$. The failure is only in the normalization, and it is repaired by dividing each vector by its norm, which changes neither the spans nor the

orthogonality (Theorem 11.1.2 (b)). Scaling is always available; orthogonality is the real content.

11.2.2 Coordinates in an orthonormal basis

Here is the payoff promised in the opening.

Theorem 11.2.1. Coordinates, Inner Products, and Norms in an Orthonormal Basis

Let $(e_1, \dots, e_n)$ be an orthonormal basis of an inner product space V, and let $u, v \in V$. Then:

(a) $v = \sum_{i=1}^{n} \langle v, e_i \rangle \, e_i$;

(b) $\langle u, v \rangle = \sum_{i=1}^{n} \langle u, e_i \rangle \, \overline{\langle v, e_i \rangle}$;

(c) $\|v\|^2 = \sum_{i=1}^{n} |\langle v, e_i \rangle|^2$ (**Parseval's identity** for a finite orthonormal basis).

Idea. For (a) we already know that $v = \sum a_i e_i$ for unique scalars a_i; the only question is what they are. Pair both sides with e_j: orthonormality kills every term except the j-th, and that one survives with coefficient 1. Parts (b) and (c) then follow by expanding, since the coordinates are now known.

Proof.

(a) Since $(e_1, \dots, e_n)$ is a basis, Theorem 1.5.1 gives scalars $a_1, \dots, a_n \in F$ with $v = \sum_i a_i e_i$. Fix j. Pairing with e_j and using (IP1),

$$\langle v, e_j \rangle = \sum_{i=1}^{n} a_i \langle e_i, e_j \rangle = \sum_{i=1}^{n} a_i \delta_{ij} = a_j,$$

by Definition 11.2.1. So $a_j = \langle v, e_j \rangle$ for every j, which is (a).

(b) Write $u = \sum_i \langle u, e_i \rangle e_i$ and $v = \sum_j \langle v, e_j \rangle e_j$ by (a). Expanding the first slot with (IP1) and the second with Theorem 11.1.1 (b),

$$\langle u, v \rangle = \sum_i \sum_j \langle u, e_i \rangle \, \overline{\langle v, e_j \rangle} \, \langle e_i, e_j \rangle.$$

Only the terms with $i = j$ survive, and there $\langle e_i, e_i \rangle = 1$. This is (b).

(c) Take $u = v$ in (b): the i-th term becomes $\langle v, e_i \rangle \overline{\langle v, e_i \rangle} = |\langle v, e_i \rangle|^2$ by Theorem 0.6.2 (c), and the left side is $\|v\|^2$.

■

Read (b) again: in the coordinates supplied by an orthonormal basis, **every** inner product becomes the standard inner product of F^n. So an orthonormal basis does not merely simplify coordinates; it turns an abstract inner product space into F^n with the dot product. We record that as a corollary below, and use it repeatedly: **to make an abstract statement concrete, fix an orthonormal basis.**

Example. Reading Off Coordinates

Let $\mathbf{e}_1 = \frac{1}{3}(1,2,2)$, $\mathbf{e}_2 = \frac{1}{3}(2,1,-2)$, $\mathbf{e}_3 = \frac{1}{3}(2,-2,1)$, the orthonormal basis of $\mathbb{R}^3$ found above. Write $\mathbf{v} = (3,0,3)$ in this basis, and check Parseval's identity.

Solution. By Theorem 11.2.1 (a), the coordinates are

$$\begin{aligned}\langle \mathbf{v}, \mathbf{e}_1\rangle &= \tfrac{1}{3}(3+0+6) = 3,\\ \langle \mathbf{v}, \mathbf{e}_2\rangle &= \tfrac{1}{3}(6+0-6) = 0,\\ \langle \mathbf{v}, \mathbf{e}_3\rangle &= \tfrac{1}{3}(6+0+3) = 3,\end{aligned}$$

so $\mathbf{v} = 3\mathbf{e}_1 + 0\mathbf{e}_2 + 3\mathbf{e}_3$. No system was solved. Checking directly, $3\mathbf{e}_1 + 3\mathbf{e}_3 = (1,2,2) + (2,-2,1) = (3,0,3)$. For Parseval, $\|\mathbf{v}\|^2 = 9+0+9 = 18$, and indeed $3^2 + 0^2 + 3^2 = 18$.

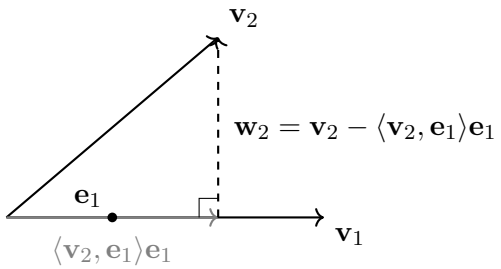
Quick check. With the same $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$, find the coordinates of $\mathbf{v} = (1,1,1)$ and verify Parseval's identity.

Solution. $\langle \mathbf{v}, \mathbf{e}_1\rangle = \frac{1}{3}(1+2+2) = \frac{5}{3}$, $\langle \mathbf{v}, \mathbf{e}_2\rangle = \frac{1}{3}(2+1-2) = \frac{1}{3}$, $\langle \mathbf{v}, \mathbf{e}_3\rangle = \frac{1}{3}(2-2+1) = \frac{1}{3}$. Then $\frac{25}{9} + \frac{1}{9} + \frac{1}{9} = \frac{27}{9} = 3 = \|\mathbf{v}\|^2$, as Parseval requires.

11.2.3 The Gram–Schmidt process

Everything so far assumed that an orthonormal basis was handed to us. Does one always exist? The answer is yes in finite dimension, and the proof is an algorithm that we will use constantly.

The idea comes from a picture in $\mathbb{R}^3$. Given independent $\mathbf{v}_1, \mathbf{v}_2$, normalize $\mathbf{v}_1$ to get $\mathbf{e}_1$. The vector $\mathbf{v}_2$ is generally not perpendicular to $\mathbf{e}_1$; but its "shadow" on the line through $\mathbf{e}_1$ is the multiple $\langle \mathbf{v}_2, \mathbf{e}_1\rangle \mathbf{e}_1$, and subtracting the shadow leaves a vector $\mathbf{w}_2$ perpendicular to $\mathbf{e}_1$. Normalize it. Then repeat with $\mathbf{v}_3$, subtracting its shadows on both $\mathbf{e}_1$ and $\mathbf{e}_2$.



Project, subtract, normalize. The right angle at the foot is the only thing that has to be checked, and one line of algebra checks it.

Theorem 11.2.2. Gram–Schmidt Process

Let V be an inner product space and let $(v_1, \dots, v_m)$ be a **linearly independent** list in V.

Define recursively

$$w_k = v_k - \sum_{i=1}^{k-1} \langle v_k, e_i \rangle \, e_i, \qquad e_k = \frac{w_k}{\|w_k\|} \qquad (k = 1, \dots, m).$$

Then every w_k is nonzero, so the recursion is defined, and:

(a) $(e_1, \dots, e_m)$ is an orthonormal list;
(b) $\operatorname{span}(e_1, \dots, e_k) = \operatorname{span}(v_1, \dots, v_k)$ for every $k = 1, \dots, m$;
(c) if $(v_1, \dots, v_r)$ is already orthonormal for some $r \le m$, then $e_i = v_i$ for $i \le r$.

In particular, every finite-dimensional inner product space has an orthonormal basis.

Idea. Induction on k. Hypothesis: $(e_1, \dots, e_k)$ is orthonormal and spans $\operatorname{span}(v_1, \dots, v_k)$. Two things need care. First, $w_k \neq 0$: if it were 0, then v_k would lie in $\operatorname{span}(e_1, \dots, e_{k-1}) = \operatorname{span}(v_1, \dots, v_{k-1})$, contradicting independence. That is the only place independence is used, and it is used exactly there. Second, $w_k \perp e_j$ for $j < k$: pair and watch the sum collapse.

Proof. We prove by induction on k that $w_k \neq 0$, that $(e_1, \dots, e_k)$ is orthonormal, and that $\operatorname{span}(e_1, \dots, e_k) = \operatorname{span}(v_1, \dots, v_k)$.

Base case $k = 1$. The empty sum is 0, so $w_1 = v_1$, which is nonzero because a list containing 0 is dependent. Then $e_1 = v_1/\|v_1\|$ has norm 1 by Theorem 11.1.2 (b), and $\operatorname{span}(e_1) = \operatorname{span}(v_1)$ since each vector is a nonzero multiple of the other.

Induction step. Let $2 \le k \le m$ and suppose the claim holds for $k - 1$. Put $U = \operatorname{span}(e_1, \dots, e_{k-1}) = \operatorname{span}(v_1, \dots, v_{k-1})$.

First, $w_k \neq 0$. If $w_k = 0$, then $v_k = \sum_{i<k} \langle v_k, e_i \rangle e_i \in U = \operatorname{span}(v_1, \dots, v_{k-1})$, so v_k would be a linear combination of the earlier v_i, making $(v_1, \dots, v_m)$ dependent, contrary to hypothesis. Hence $\|w_k\| > 0$ and e_k is defined, with $\|e_k\| = 1$.

Next, $e_k \perp e_j$ for every $j < k$. By (IP1) and the induction hypothesis $\langle e_i, e_j \rangle = \delta_{ij}$,

$$\langle w_k, e_j \rangle = \langle v_k, e_j \rangle - \sum_{i=1}^{k-1} \langle v_k, e_i \rangle \, \delta_{ij} = \langle v_k, e_j \rangle - \langle v_k, e_j \rangle = 0,$$

and e_k is a scalar multiple of w_k, so $\langle e_k, e_j \rangle = 0$ as well. Together with the induction hypothesis, $(e_1, \dots, e_k)$ is orthonormal, which is (a).

For (b), e_k is a linear combination of v_k and $e_1, \dots, e_{k-1}$, hence lies in $\operatorname{span}(v_1, \dots, v_k)$ by the induction hypothesis; so $\operatorname{span}(e_1, \dots, e_k) \subseteq \operatorname{span}(v_1, \dots, v_k)$. Conversely $v_k = \|w_k\| e_k + \sum_{i<k} \langle v_k, e_i \rangle e_i \in \operatorname{span}(e_1, \dots, e_k)$, and the earlier v_i lie there by the induction hypothesis. The two spans are equal, completing the induction.

For (c), suppose $(v_1, \dots, v_r)$ is orthonormal and that $e_i = v_i$ for all $i < k$, where $k \le r$. Then $\langle v_k, e_i \rangle = \langle v_k, v_i \rangle = 0$ for $i < k$, so $w_k = v_k$ and $\|w_k\| = 1$, giving $e_k = v_k$. Induction on k proves (c).

Finally, let V be finite-dimensional. If $V = \{0\}$, the empty list is an orthonormal basis. Otherwise take any basis $(v_1, \dots, v_n)$ of V and apply the process: by (a) the output is

orthonormal, and by (b) it spans $\operatorname{span}(v_1, \dots, v_n) = V$, so it is an orthonormal basis.

■

Two remarks on the formulas. The vector subtracted from v_k is exactly the expansion of v_k in the orthonormal basis $(e_1, \dots, e_{k-1})$ of U, which by Theorem 11.2.1 (a) is the best available "copy of v_k inside U"; the next section will call it the orthogonal projection of v_k onto U. And the subtraction is the same move that produced the vector z in the proof of Theorem 11.1.5. **Project, then subtract** is one move used twice.

Corollary 11.2.3. Extending an Orthonormal List

Let V be a finite-dimensional inner product space and let $(e_1, \dots, e_r)$ be an orthonormal list in V. Then there are vectors $e_{r+1}, \dots, e_n \in V$ such that $(e_1, \dots, e_n)$ is an orthonormal basis of V.

Proof. The list $(e_1, \dots, e_r)$ is linearly independent by Theorem 11.1.4, since its vectors are orthogonal and nonzero. By Theorem 1.6.5 there are $v_{r+1}, \dots, v_n \in V$ making $(e_1, \dots, e_r, v_{r+1}, \dots, v_n)$ a basis of V. Apply Theorem 11.2.2 to that basis. By part (c) the first r output vectors are $e_1, \dots, e_r$ unchanged, and by (a) and (b) the whole output is an orthonormal basis of V.

■

11.2.4 Two worked examples

Example. Gram–Schmidt in $\mathbb{R}^3$

Apply the Gram–Schmidt process of Theorem 11.2.2 to $\mathbf{v}_1 = (1,1,1)$, $\mathbf{v}_2 = (0,1,1)$, $\mathbf{v}_3 = (0,0,1)$ in $\mathbb{R}^3$ with the dot product.

Solution. Step 1. $\|\mathbf{v}_1\|^2 = 3$, so $\mathbf{e}_1 = \frac{1}{\sqrt{3}}(1,1,1)$.

Step 2. $\langle \mathbf{v}_2, \mathbf{e}_1 \rangle = \frac{1}{\sqrt{3}}(0+1+1) = \frac{2}{\sqrt{3}}$, so

$$\mathbf{w}_2 = (0,1,1) - \tfrac{2}{3}(1,1,1) = \tfrac{1}{3}(-2,1,1).$$

Then $\|\mathbf{w}_2\|^2 = \frac{1}{9}(4+1+1) = \frac{2}{3}$, and $\mathbf{e}_2 = \frac{1}{\sqrt{6}}(-2,1,1)$.

Step 3. $\langle \mathbf{v}_3, \mathbf{e}_1 \rangle = \frac{1}{\sqrt{3}}$ and $\langle \mathbf{v}_3, \mathbf{e}_2 \rangle = \frac{1}{\sqrt{6}}$, so

$$\mathbf{w}_3 = (0,0,1) - \tfrac{1}{3}(1,1,1) - \tfrac{1}{6}(-2,1,1) = (0, -\tfrac{1}{2}, \tfrac{1}{2}).$$

Then $\|\mathbf{w}_3\|^2 = \frac{1}{2}$ and $\mathbf{e}_3 = \frac{1}{\sqrt{2}}(0,-1,1)$.

Check. The three vectors have norm 1, and $\langle \mathbf{e}_1, \mathbf{e}_2 \rangle = \frac{1}{\sqrt{18}}(-2+1+1) = 0$, $\langle \mathbf{e}_1, \mathbf{e}_3 \rangle =$

$\frac{1}{\sqrt{6}}(0-1+1)=0$, $\langle \mathbf{e}_2, \mathbf{e}_3\rangle = \frac{1}{\sqrt{12}}(0-1+1)=0$. The span condition is visible too: $\mathbf{e}_1$ is a multiple of $\mathbf{v}_1$, and $\mathbf{e}_2$ is a combination of $\mathbf{v}_1$ and $\mathbf{v}_2$.

The next example runs the same three steps in a function space, where no picture is available and only the algebra is left.

Example. Gram–Schmidt on $1, x, x^2$

Apply the Gram–Schmidt process of Theorem 11.2.2 to $(1, x, x^2)$ in $\mathbb{R}[x]_{\le 2}$ with

$$\langle p, q\rangle = \int_{-1}^{1} p(t)q(t)\,\mathrm{d}t.$$

Solution. Step 1. $\|1\|^2 = \int_{-1}^{1} 1\,\mathrm{d}t = 2$, so $e_1 = 1/\sqrt{2}$.

Step 2. $\langle x, e_1\rangle = \frac{1}{\sqrt{2}}\int_{-1}^{1} t\,\mathrm{d}t = 0$, since the integrand is odd. So $w_2 = x$, and $\|x\|^2 = \int_{-1}^{1} t^2\,\mathrm{d}t = \frac{2}{3}$, giving $e_2 = \sqrt{3/2}\,x$.

Step 3. $\langle x^2, e_1\rangle = \frac{1}{\sqrt{2}}\cdot\frac{2}{3}$ and $\langle x^2, e_2\rangle = \sqrt{3/2}\int_{-1}^{1} t^3\,\mathrm{d}t = 0$, again by oddness. Hence

$$w_3 = x^2 - \tfrac{1}{\sqrt{2}}\cdot\tfrac{2}{3}\cdot\tfrac{1}{\sqrt{2}} = x^2 - \tfrac{1}{3}.$$

Its squared norm is

$$\int_{-1}^{1}\left(t^2-\tfrac{1}{3}\right)^2\mathrm{d}t = \tfrac{2}{5} - \tfrac{4}{9} + \tfrac{2}{9} = \tfrac{8}{45},$$

so $e_3 = \frac{\sqrt{10}}{4}(3x^2-1)$.

The monic polynomials produced along the way are 1, x, and $x^2-\frac{1}{3}$. These are, up to scaling, the first three **Legendre polynomials**; the usual normalization takes $P_2 = \frac{1}{2}(3x^2-1) = \frac{3}{2}(x^2-\frac{1}{3})$, fixed by the convention $P_k(1)=1$ rather than by the norm. Section 10 of this chapter takes up such families in general.

Notice how much work the oddness of the integrand did: two of the three inner products were 0 before any computation. Symmetry of the weight is worth looking for.

Warning. **Gram–Schmidt depends on the order of the list.** Run the process on $((0,1),(1,1))$ in $\mathbb{R}^2$: it returns $((0,1),(1,0))$. Run it on $((1,1),(0,1))$, the same two vectors in the other order: it returns $(\frac{1}{\sqrt{2}}(1,1), \frac{1}{\sqrt{2}}(-1,1))$. Both outputs are orthonormal bases of $\mathbb{R}^2$, and they are different. An orthonormal basis is never unique (for $\dim V \ge 1$ one can always replace e_1 by $-e_1$); the process returns **an** orthonormal basis adapted to the given order, not **the** orthonormal basis.

Quick check. Run Gram–Schmidt on $((1,0),(1,1))$ in $\mathbb{R}^2$, and then on $((1,1),(1,0))$. Do you get the same list?

Solution. First order: $\mathbf{e}_1 = (1,0)$; $\langle (1,1), \mathbf{e}_1 \rangle = 1$, so $\mathbf{w}_2 = (1,1) - (1,0) = (0,1)$ and $\mathbf{e}_2 = (0,1)$. Output $((1,0),(0,1))$.

Second order: $\mathbf{e}_1 = \frac{1}{\sqrt{2}}(1,1)$; $\langle (1,0), \mathbf{e}_1 \rangle = \frac{1}{\sqrt{2}}$, so $\mathbf{w}_2 = (1,0) - \frac{1}{2}(1,1) = \frac{1}{2}(1,-1)$ and $\mathbf{e}_2 = \frac{1}{\sqrt{2}}(1,-1)$. Output $(\frac{1}{\sqrt{2}}(1,1), \frac{1}{\sqrt{2}}(1,-1))$. Not the same list, and not even the same set of lines.

Warning. **The classical process is numerically fragile.** Computed with rounding, the vectors $\mathbf{e}_1, \mathbf{e}_2, \ldots$ drift out of orthogonality, and the drift grows with k and with how nearly dependent the input list is. Nothing above is wrong — the fault is in the arithmetic, not the mathematics — but in practice one uses a rearranged version or reflections instead. Section 8 of this chapter gives the reflection-based method, and Chapter 24 explains why it is the stable one.

11.2.5 All inner product spaces of the same dimension look alike

Corollary 11.2.4. Classification of Finite-Dimensional Inner Product Spaces

Let V be an inner product space over F with $\dim V = n$. Then there is an isomorphism $T\colon V \to F^n$ with

$$\langle Tu, Tv \rangle = \langle u, v \rangle \qquad \text{for all } u, v \in V,$$

where the right-hand inner product is that of V and the left-hand one is the standard inner product of F^n. An isomorphism with this property is an **isometric isomorphism**, so every n-dimensional inner product space over F is isometrically isomorphic to F^n.

Proof. By Theorem 11.2.2, V has an orthonormal basis $\mathcal{E} = (e_1, \ldots, e_n)$. Let $T\colon V \to F^n$ be the coordinate map $Tv = [v]_{\mathcal{E}}$, an isomorphism by Corollary 2.5.5. By Theorem 11.2.1 (a), the i-th entry of Tv is $\langle v, e_i \rangle$. Hence the standard inner product of Tu and Tv in F^n is $\sum_i \langle u, e_i \rangle \overline{\langle v, e_i \rangle}$, which equals $\langle u, v \rangle$ by Theorem 11.2.1 (b). ∎

So dimension is a complete invariant of finite-dimensional inner product spaces over a fixed field: there is nothing to classify. This should be compared with Theorem 2.5.4, which said the same for vector spaces; adding an inner product adds no new invariant. What becomes interesting is the interaction of the inner product with a given operator, and that is the subject of Chapter 12. Isometries in general are the subject of Section 7 of this chapter.

11.2.6 Gram matrices

Given a list of vectors, all of its inner products can be collected into one matrix, and that matrix turns out to answer questions about the list.

Definition 11.2.3. Gram Matrix

Let $(v_1, \dots, v_k)$ be a list in an inner product space V over F. Its **Gram matrix** is $\mathbf{G} \in M_k(F)$ with entries

$$(\mathbf{G})_{ij} = \langle v_j, v_i \rangle \qquad (1 \le i, j \le k).$$

The order of the subscripts looks backwards and is chosen on purpose: it makes $\mathbf{G}$ act on coefficient vectors the right way round, as part (b) below shows. For a real list the distinction disappears, since then $\langle v_j, v_i \rangle = \langle v_i, v_j \rangle$.

For example, the Gram matrix of $((1,1,0),(1,0,1),(0,1,1))$ in $\mathbb{R}^3$ is $\begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix}$, with the squared norms on the diagonal.

Over $\mathbb{C}$ the order of the subscripts becomes visible. For $\mathbf{v}_1 = (1,0)$ and $\mathbf{v}_2 = (i,1)$ in $\mathbb{C}^2$, the off-diagonal entries are

$$(\mathbf{G})_{12} = \langle \mathbf{v}_2, \mathbf{v}_1 \rangle = i, \qquad (\mathbf{G})_{21} = \langle \mathbf{v}_1, \mathbf{v}_2 \rangle = \bar{i} = -i,$$

so $\mathbf{G} = \begin{pmatrix} 1 & i \\ -i & 2 \end{pmatrix}$, which is Hermitian but not symmetric. Part (b) below says $\mathbf{x}^*\mathbf{G}\mathbf{x} = \|x_1\mathbf{v}_1 + x_2\mathbf{v}_2\|^2$; with $\mathbf{x} = (1,1)$ both sides are 3, since $\mathbf{v}_1 + \mathbf{v}_2 = (1+i, 1)$. Writing $(\mathbf{G})_{ij} = \langle \mathbf{v}_i, \mathbf{v}_j \rangle$ instead would transpose $\mathbf{G}$ and leave that identity reading $\mathbf{x}^*\mathbf{G}^\top\mathbf{x}$.

Theorem 11.2.5. Properties of the Gram Matrix

Let $(v_1, \dots, v_k)$ be a list in an inner product space V over F, with Gram matrix $\mathbf{G}$. Then:

(a) $\mathbf{G}$ is **Hermitian**, $\mathbf{G}^* = \mathbf{G}$, and its diagonal entries are the real numbers $\|v_i\|^2 \ge 0$;

(b) $\mathbf{x}^*\mathbf{G}\mathbf{x} = \left\| \sum_j x_j v_j \right\|^2$ for every $\mathbf{x} \in F^k$;

(c) $\mathbf{G}$ is invertible if and only if $(v_1, \dots, v_k)$ is linearly independent;

(d) $\det \mathbf{G}$ is a real number with $\det \mathbf{G} \ge 0$, and $\det \mathbf{G} > 0$ exactly when the list is linearly independent.

Idea. Part (b) is the whole theorem: the quadratic expression $\mathbf{x}^*\mathbf{G}\mathbf{x}$ is a squared length, so it vanishes only for the coefficient vectors of genuine dependence relations. That identifies $\operatorname{nul}(\mathbf{G})$ with the set of dependence relations and gives (c) at once. For (d), fix an orthonormal basis of the span and write $\mathbf{G} = \mathbf{A}^*\mathbf{A}$ for a coordinate matrix $\mathbf{A}$; a determinant of that shape is a squared modulus.

Proof.

(a) By (IP2), $(\mathbf{G}^*)_{ij} = \overline{(\mathbf{G})_{ji}} = \overline{\langle v_i, v_j \rangle} = \langle v_j, v_i \rangle = (\mathbf{G})_{ij}$. The diagonal entries are $\langle v_i, v_i \rangle = \|v_i\|^2$, real and ≥ 0 by Theorem 11.1.1 (a).

(b) Write $u = \sum_j x_j v_j$. Expanding by (IP1) and Theorem 11.1.1 (b),

$$\|u\|^2 = \sum_i \sum_j x_j \overline{x_i} \langle v_j, v_i \rangle = \sum_i \overline{x_i} \sum_j (\mathbf{G})_{ij} x_j,$$

and the right-hand side is exactly the entry of the 1×1 matrix $\mathbf{x}^{*}\mathbf{G}\mathbf{x}$.

(c) We show $\operatorname{nul}(\mathbf{G}) = \{\mathbf{x} \in F^{k} : \sum_{j} x_{j} v_{j} = 0\}$. If $\sum_{j} x_{j} v_{j} = 0$, then for each i, $(\mathbf{G}\mathbf{x})_{i} = \sum_{j} \langle v_{j}, v_{i} \rangle x_{j} = \langle \sum_{j} x_{j} v_{j}, v_{i} \rangle = 0$ by (IP1), so $\mathbf{G}\mathbf{x} = \mathbf{0}$. Conversely, if $\mathbf{G}\mathbf{x} = \mathbf{0}$, then $\mathbf{x}^{*}\mathbf{G}\mathbf{x} = 0$, so $\left\| \sum_{j} x_{j} v_{j} \right\| = 0$ by (b) and hence $\sum_{j} x_{j} v_{j} = 0$ by Theorem 11.1.2 (a). Now $\mathbf{G}$ is invertible if and only if $\mathbf{G}\mathbf{x} = \mathbf{0}$ forces $\mathbf{x} = \mathbf{0}$ (Theorem 3.4.4, (a) $\Leftrightarrow$ (b)), which by the displayed equality of sets says exactly that the only dependence relation among the v_{j} is the trivial one.

(d) Let $U = \operatorname{span}(v_{1}, \dots, v_{k})$, let $r = \dim U$, and let $(e_{1}, \dots, e_{r})$ be an orthonormal basis of U, which exists by Theorem 11.2.2. Let $\mathbf{A} \in M_{r \times k}(F)$ have as its j-th column the coordinate vector of v_{j}, so that $v_{j} = \sum_{p} a_{pj} e_{p}$. By Theorem 11.2.1 (b),

$$(\mathbf{G})_{ij} = \langle v_{j}, v_{i} \rangle = \sum_{p=1}^{r} a_{pj} \overline{a_{pi}} = (\mathbf{A}^{*}\mathbf{A})_{ij},$$

so $\mathbf{G} = \mathbf{A}^{*}\mathbf{A}$.

If the list is dependent, then $\mathbf{G}$ is singular by (c) and $\det \mathbf{G} = 0$. If it is independent, then $r = k$ and $\mathbf{A}$ is square. By Theorem 7.4.5, $\det(\mathbf{A}^{*}) = \det(\overline{\mathbf{A}})$, and $\det(\overline{\mathbf{A}}) = \overline{\det \mathbf{A}}$, because by the Leibniz formula (Theorem 7.4.3) the determinant is a sum of products of entries and conjugation preserves sums and products (Theorem 0.6.2). Hence, by Theorem 7.5.4,

$$\det \mathbf{G} = \det(\mathbf{A}^{*}) \det(\mathbf{A}) = \overline{\det \mathbf{A}} \, \det \mathbf{A} = |\det \mathbf{A}|^{2}.$$

This is real and ≥ 0, and it is > 0 because $\mathbf{A}$ is the matrix of an isomorphism onto its coordinates, hence invertible with $\det \mathbf{A} \neq 0$ (Theorem 7.5.2).

■

Over $\mathbb{R}$, with $V = \mathbb{R}^{n}$ and the dot product, part (d) is the statement Chapter 7 reached from Cauchy–Binet: Corollary 7.7.3 says that $\det(\mathbf{A}\mathbf{A}^{\top}) \geq 0$ for a real $\mathbf{A}$, with equality exactly when the rows of $\mathbf{A}$ are dependent. Taking the rows of $\mathbf{A}$ to be $\mathbf{v}_{1}, \dots, \mathbf{v}_{k}$ makes $\mathbf{A}\mathbf{A}^{\top}$ the Gram matrix, so that corollary is the special case of (d) in which the ambient space is $\mathbb{R}^{n}$. The pointer left there is now paid: a Gram matrix is exactly a matrix of pairwise inner products, and its determinant is nonnegative for that reason.

Example. Deciding Independence with a Gram Matrix

Decide whether $((1,1,0),(1,0,1),(0,1,-1))$ is independent in $\mathbb{R}^{3}$, using its Gram matrix.

Solution. The pairwise dot products give

$$\mathbf{G} = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & -1 \\ 1 & -1 & 2 \end{pmatrix},$$

and expanding along the first row, $\det \mathbf{G} = 2(4-1) - 1(2+1) + 1(-1-2) = 6 - 3 - 3 = 0$. By Theorem 11.2.5 (d) the list is dependent. Indeed $\mathbf{G}\mathbf{x} = \mathbf{0}$ for $\mathbf{x} = (1,-1,-1)$, and Theorem 11.2.5 (c) says this is a dependence relation: $(1,1,0) - (1,0,1) - (0,1,-1) = \mathbf{0}$.

Two features of this method are worth noting. It never needs coordinates in the ambient space, so it works verbatim for polynomials and functions, where writing the vectors as columns is not an option. And the answer comes with a witness: a null vector of $\mathbf{G}$ is a dependence relation, ready to read off.

11.2.7 Exercises

11.2.7.1 A. Check your understanding

Exercise. A1

(a) Define an orthonormal list, using the Kronecker delta, and say what an orthonormal basis is.

(b) Why is every orthonormal list linearly independent? Name the result used.

(c) Write down the coordinates of v in an orthonormal basis $(e_1, \dots, e_n)$, and state Parseval's identity.

(d) True or false: the Gram–Schmidt process (Theorem 11.2.2) applied to a list and to the same list in reverse order produces the same orthonormal list. Justify your answer.

(e) State the criterion for linear independence in terms of the Gram matrix.

(f) True or false: every 3-dimensional real inner product space is isometrically isomorphic to $\mathbb{R}^3$ with the dot product. Justify your answer.

Solution.

(a) A list $(e_1, \dots, e_k)$ with $\langle e_i, e_j \rangle = \delta_{ij}$ for all i, j (Definition 11.2.1); an orthonormal basis is such a list that is also a basis (Definition 11.2.2).

(b) Its vectors are pairwise orthogonal and nonzero (each has norm 1), so Theorem 11.1.4 applies.

(c) $v = \sum_i \langle v, e_i \rangle e_i$, and $\|v\|^2 = \sum_i |\langle v, e_i \rangle|^2$ (Theorem 11.2.1).

(d) False. On $((0,1),(1,1))$ the process returns $((0,1),(1,0))$, while on $((1,1),(0,1))$ it returns $(\frac{1}{\sqrt{2}}(1,1), \frac{1}{\sqrt{2}}(-1,1))$.

(e) The list is independent if and only if its Gram matrix is invertible, if and only if $\det \mathbf{G} > 0$ (Theorem 11.2.5 (c), (d)).

(f) True. By Corollary 11.2.4, the coordinate map with respect to any orthonormal basis is an isometric isomorphism onto $\mathbb{R}^3$.

11.2.7.2 B. Practice

Exercise. B1: Gram–Schmidt with numbers

Apply the Gram–Schmidt process of Theorem 11.2.2 to each list, with the dot product.

(a) $((1,0,1),(1,1,0),(0,1,1))$ in $\mathbb{R}^3$.

(b) $((1,1,0,0),(1,0,1,0),(0,0,1,1))$ in $\mathbb{R}^4$.

Solution.

(a) $\|(1,0,1)\|^2 = 2$, so $\mathbf{e}_1 = \frac{1}{\sqrt{2}}(1,0,1)$. Next $\langle (1,1,0), \mathbf{e}_1\rangle = \frac{1}{\sqrt{2}}$, so

$$\mathbf{w}_2 = (1,1,0) - \tfrac{1}{2}(1,0,1) = \tfrac{1}{2}(1,2,-1),$$

with $\|\mathbf{w}_2\|^2 = \frac{1}{4}\cdot 6 = \frac{3}{2}$, giving $\mathbf{e}_2 = \frac{1}{\sqrt{6}}(1,2,-1)$. Finally $\langle (0,1,1), \mathbf{e}_1\rangle = \frac{1}{\sqrt{2}}$ and $\langle (0,1,1), \mathbf{e}_2\rangle = \frac{1}{\sqrt{6}}$, so

$$\mathbf{w}_3 = (0,1,1) - \tfrac{1}{2}(1,0,1) - \tfrac{1}{6}(1,2,-1) = \tfrac{2}{3}(-1,1,1),$$

with $\|\mathbf{w}_3\|^2 = \frac{4}{9}\cdot 3 = \frac{4}{3}$, giving $\mathbf{e}_3 = \frac{1}{\sqrt{3}}(-1,1,1)$.

(b) $\mathbf{e}_1 = \frac{1}{\sqrt{2}}(1,1,0,0)$. Then $\langle (1,0,1,0), \mathbf{e}_1\rangle = \frac{1}{\sqrt{2}}$, so

$$\mathbf{w}_2 = (1,0,1,0) - \tfrac{1}{2}(1,1,0,0) = \tfrac{1}{2}(1,-1,2,0),$$

with $\|\mathbf{w}_2\|^2 = \frac{1}{4}\cdot 6 = \frac{3}{2}$ and $\mathbf{e}_2 = \frac{1}{\sqrt{6}}(1,-1,2,0)$. Finally $\langle (0,0,1,1), \mathbf{e}_1\rangle = 0$ and $\langle (0,0,1,1), \mathbf{e}_2\rangle = \frac{2}{\sqrt{6}}$, so

$$\mathbf{w}_3 = (0,0,1,1) - \tfrac{1}{3}(1,-1,2,0) = \tfrac{1}{3}(-1,1,1,3),$$

with $\|\mathbf{w}_3\|^2 = \frac{1}{9}\cdot 12 = \frac{4}{3}$ and $\mathbf{e}_3 = \frac{1}{2\sqrt{3}}(-1,1,1,3)$.

Exercise. B2: Gram–Schmidt on polynomials

In $\mathbb{R}[x]_{\le 2}$ with $\langle p, q\rangle = \int_0^1 p(t)q(t)\,\mathrm{d}t$, apply the Gram–Schmidt process of Theorem 11.2.2 to $(1, x, x^2)$. Compare the monic outputs with those of Example (Gram–Schmidt on $1, x, x^2$) and explain the difference.

Solution. $\|1\|^2 = 1$, so $e_1 = 1$. Then $\langle x, 1\rangle = \int_0^1 t\,\mathrm{d}t = \frac{1}{2}$, so $w_2 = x - \frac{1}{2}$ and

$$\|w_2\|^2 = \int_0^1 \left(t - \tfrac{1}{2}\right)^2 \mathrm{d}t = \tfrac{1}{3} - \tfrac{1}{2} + \tfrac{1}{4} = \tfrac{1}{12},$$

giving $e_2 = \sqrt{3}\,(2x-1)$. Next $\langle x^2, e_1\rangle = \frac{1}{3}$ and

$$\langle x^2, e_2\rangle = \sqrt{3}\int_0^1 (2t^3 - t^2)\,\mathrm{d}t = \sqrt{3}\left(\tfrac{1}{2} - \tfrac{1}{3}\right) = \tfrac{\sqrt{3}}{6},$$

so $w_3 = x^2 - \frac{1}{3} - \frac{\sqrt{3}}{6}\cdot\sqrt{3}(2x-1) = x^2 - x + \frac{1}{6}$. Its squared norm is

$$\int_0^1 \left(t^2 - t + \tfrac{1}{6}\right)^2 \mathrm{d}t = \tfrac{1}{180},$$

so $e_3 = \sqrt{5}\,(6x^2 - 6x + 1)$.

The monic outputs are 1, $x - \frac{1}{2}$, $x^2 - x + \frac{1}{6}$, whereas on $[-1,1]$ they were 1, x, $x^2 - \frac{1}{3}$. Orthogonality is a property of the inner product, not of the polynomials: changing

the interval changes which polynomials are orthogonal. These are the shifted Legendre polynomials, and the substitution $t \mapsto 2t - 1$ carries one family to the other.

Exercise. B3: Coordinates without solving a system

Let $\mathbf{e}_1 = \frac{1}{3}(1,2,2)$, $\mathbf{e}_2 = \frac{1}{3}(2,1,-2)$, $\mathbf{e}_3 = \frac{1}{3}(2,-2,1)$ in $\mathbb{R}^3$.

(a) Write $\mathbf{v} = (1,-1,4)$ in this basis.

(b) Compute $\|\mathbf{v}\|^2$ in two ways, and state which theorem says the answers agree.

(c) Compute $\langle \mathbf{v}, \mathbf{u} \rangle$ for $\mathbf{u} = (3,0,3)$ from the coordinates of $\mathbf{v}$ and $\mathbf{u}$, using Example (Reading Off Coordinates) for $\mathbf{u}$.

Solution.

(a) $\langle \mathbf{v}, \mathbf{e}_1 \rangle = \frac{1}{3}(1-2+8) = \frac{7}{3}$, $\langle \mathbf{v}, \mathbf{e}_2 \rangle = \frac{1}{3}(2-1-8) = -\frac{7}{3}$, $\langle \mathbf{v}, \mathbf{e}_3 \rangle = \frac{1}{3}(2+2+4) = \frac{8}{3}$. So $\mathbf{v} = \frac{7}{3}\mathbf{e}_1 - \frac{7}{3}\mathbf{e}_2 + \frac{8}{3}\mathbf{e}_3$.

(b) Directly, $\|\mathbf{v}\|^2 = 1+1+16 = 18$. From the coordinates, $\frac{49}{9} + \frac{49}{9} + \frac{64}{9} = \frac{162}{9} = 18$. They agree by Theorem 11.2.1 (c).

(c) The coordinates of $\mathbf{u}$ are $(3,0,3)$ by Example (Reading Off Coordinates), so by Theorem 11.2.1 (b),

$$\langle \mathbf{v}, \mathbf{u} \rangle = \tfrac{7}{3} \cdot 3 + \left(-\tfrac{7}{3}\right) \cdot 0 + \tfrac{8}{3} \cdot 3 = 7 + 8 = 15.$$

Directly, $\langle \mathbf{v}, \mathbf{u} \rangle = 3 + 0 + 12 = 15$.

Exercise. B4: Gram matrices

For each list, write down the Gram matrix and use it to decide whether the list is independent.

(a) $((1,1,0),(1,0,1),(0,1,1))$ in $\mathbb{R}^3$ with the dot product.

(b) $((1,i),(i,1))$ in $\mathbb{C}^2$ with the standard inner product.

(c) $(1, x, x^2 - \frac{1}{3})$ in $\mathbb{R}[x]_{\le 2}$ with $\langle p, q \rangle = \int_{-1}^{1} pq$.

Solution.

(a) $\mathbf{G} = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix}$, and expanding along the first row, $\det \mathbf{G} = 2(4-1) - 1(2-1) + 1(1-2) = 6 - 1 - 1 = 4 > 0$. By Theorem 11.2.5 (d) the list is independent.

(b) $\langle (1,i),(1,i) \rangle = 1 + 1 = 2$ and $\langle (i,1),(1,i) \rangle = i\overline{1} + 1\overline{i} = i - i = 0$, so $\mathbf{G} = 2\mathbf{I}_2$ by Theorem 11.2.5 (a). Then $\det \mathbf{G} = 4 > 0$ and the list is independent; in fact it is orthogonal, so Theorem 11.1.4 already gives the answer.

(c) These are the monic outputs of Example (Gram–Schmidt on $1, x, x^2$), so they are pairwise orthogonal, and $\mathbf{G} = \operatorname{diag}(2, \frac{2}{3}, \frac{8}{45})$. Its determinant is $2 \cdot \frac{2}{3} \cdot \frac{8}{45} = \frac{32}{135} > 0$, so the list is independent. A Gram matrix is diagonal exactly when the list is orthogonal.

11.2.7.3 C. Going deeper

Exercise. C1: The Gram matrix determines the list

Let $(v_1, \dots, v_k)$ be a list in an inner product space V and $(w_1, \dots, w_k)$ a list in an inner product space W, both over F, with the **same** Gram matrix: $\langle v_j, v_i \rangle = \langle w_j, w_i \rangle$ for all i, j.

(a) Prove that $\sum_j x_j v_j = 0$ if and only if $\sum_j x_j w_j = 0$, for every $\mathbf{x} \in F^k$.

(b) Deduce that there is an isomorphism $T\colon \operatorname{span}(v_1, \dots, v_k) \to \operatorname{span}(w_1, \dots, w_k)$, well defined by $T v_j = w_j$ for every j.

(c) Prove that T is an isometric isomorphism, that is, $\langle Tu, Tu' \rangle = \langle u, u' \rangle$ for all u, u' in the domain.

Solution.

(a) Let $\mathbf{G}$ be the common Gram matrix. By Theorem 11.2.5 (b), $\left\| \sum_j x_j v_j \right\|^2 = \mathbf{x}^* \mathbf{G} \mathbf{x} = \left\| \sum_j x_j w_j \right\|^2$. By Theorem 11.1.2 (a), one side is 0 if and only if the other is, and a norm is 0 only at the zero vector.

(b) Every element of $U = \operatorname{span}(v_1, \dots, v_k)$ is $\sum_j x_j v_j$ for some $\mathbf{x}$. Define $T\big(\sum_j x_j v_j\big) = \sum_j x_j w_j$. This is well defined: if $\sum_j x_j v_j = \sum_j y_j v_j$, then $\sum_j (x_j - y_j) v_j = 0$, so $\sum_j (x_j - y_j) w_j = 0$ by (a), that is $\sum_j x_j w_j = \sum_j y_j w_j$. It is linear because the formula is linear in $\mathbf{x}$, it is surjective onto $\operatorname{span}(w_1, \dots, w_k)$ by construction, and it is injective: if $T(\sum x_j v_j) = 0$, then $\sum x_j w_j = 0$, so $\sum x_j v_j = 0$ by (a).

(c) Let $u = \sum_j x_j v_j$ and $u' = \sum_i y_i v_i$. Expanding by (IP1) and Theorem 11.1.1 (b),

$$\langle u, u' \rangle = \sum_i \sum_j x_j \overline{y_i} \langle v_j, v_i \rangle,$$

and the same computation in W gives $\langle Tu, Tu' \rangle = \sum_i \sum_j x_j \overline{y_i} \langle w_j, w_i \rangle$. The two sums are equal term by term, since the Gram matrices agree.

Exercise. C2: Gram–Schmidt on a dependent list

Let $(v_1, \dots, v_m)$ be a list in an inner product space, **not** assumed independent, and run the recursion of Theorem 11.2.2.

(a) Let k be the smallest index with $v_k \in \operatorname{span}(v_1, \dots, v_{k-1})$. Prove that the recursion is defined up to e_{k-1} and that $w_k = 0$, so the step $e_k = w_k / \|w_k\|$ fails.

(b) Repair the process: show that if, whenever $w_k = 0$, we discard v_k and continue with v_{k+1}, the output is an orthonormal basis of $\operatorname{span}(v_1, \dots, v_m)$.

(c) Run the repaired process on $((1,1,0), (2,2,0), (1,0,1))$ in $\mathbb{R}^3$.

Solution.

(a) The list $(v_1, \dots, v_{k-1})$ is independent, since a dependence among its members would, by Theorem 1.4.1, put some v_j with $j < k$ in the span of its predecessors, contradicting the minimality of k. So Theorem 11.2.2 applies to it and produces $e_1, \dots, e_{k-1}$ with $\operatorname{span}(e_1, \dots, e_{k-1}) = \operatorname{span}(v_1, \dots, v_{k-1}) =: U$. Since $v_k \in U$, Theorem 11.2.1 (a)

applied inside U gives $v_k = \sum_{i<k} \langle v_k, e_i \rangle e_i$, which is precisely the vector subtracted, so $w_k = 0$. Dividing by $\|w_k\| = 0$ is impossible.

(b) Let $e_1, \dots, e_s$ be the vectors that the repaired process outputs. The proof of Theorem 11.2.2 shows, unchanged, that after processing $v_1, \dots, v_k$ the accumulated list is orthonormal and spans $\operatorname{span}(v_1, \dots, v_k)$: the only step that used independence was the claim $w_k \neq 0$, and in the repaired process a step with $w_k = 0$ discards a v_k that lies in the span already built, so the span does not change. Taking $k = m$, the output is an orthonormal list spanning $\operatorname{span}(v_1, \dots, v_m)$, hence an orthonormal basis of it.

(c) $\mathbf{e}_1 = \frac{1}{\sqrt{2}}(1,1,0)$. For $\mathbf{v}_2 = (2,2,0)$: $\langle \mathbf{v}_2, \mathbf{e}_1 \rangle = \frac{4}{\sqrt{2}} = 2\sqrt{2}$, so $\mathbf{w}_2 = (2,2,0) - 2\sqrt{2} \cdot \frac{1}{\sqrt{2}}(1,1,0) = \mathbf{0}$, and $\mathbf{v}_2$ is discarded. For $\mathbf{v}_3 = (1,0,1)$: $\langle \mathbf{v}_3, \mathbf{e}_1 \rangle = \frac{1}{\sqrt{2}}$, so $\mathbf{w}_3 = (1,0,1) - \frac{1}{2}(1,1,0) = \frac{1}{2}(1,-1,2)$, with $\|\mathbf{w}_3\|^2 = \frac{6}{4} = \frac{3}{2}$ and $\mathbf{e}_2 = \frac{1}{\sqrt{6}}(1,-1,2)$. The output $\left(\frac{1}{\sqrt{2}}(1,1,0), \frac{1}{\sqrt{6}}(1,-1,2)\right)$ is an orthonormal basis of the plane spanned by the three given vectors.

11.3 Orthogonal Complements and Projections

Chapter 1 proved that every subspace of a finite-dimensional space has a complement, and then warned that it has many: the x-axis in $\mathbb{R}^2$ is complemented by every other line through the origin. An inner product breaks the tie. Among all the complements of U there is now one built from U alone, the set of vectors perpendicular to everything in U, and the projection onto U along it turns out to solve an approximation problem: it finds the vector of U closest to a given v. This section constructs that complement, that projection, and that answer.

Throughout, V is an inner product space over $F = \mathbb{R}$ or $F = \mathbb{C}$, with inner product $\langle \cdot, \cdot \rangle$ (Definition 11.1.1) and induced norm $\|\cdot\|$ (Definition 11.1.3). We do **not** assume V is finite-dimensional; what we assume, each time, is that the subspace we project onto is.

11.3.1 Orthogonal complements

In $\mathbb{R}^3$, a plane U through the origin has a distinguished line attached to it: the normal line, the set of vectors perpendicular to the whole plane. Nothing about the plane had to be chosen to produce it. That construction is available in any inner product space, and it is available for any subset, not only for subspaces.

The orthogonal complement of U collects every vector that is perpendicular to all of U at once.

Definition 11.3.1. Orthogonal complement

Let V be an inner product space over F and let $S \subseteq V$ be **any** subset. The **orthogonal complement** of S is

$$S^{\perp} = \{\, v \in V : \langle v, s \rangle = 0 \text{ for every } s \in S \,\},$$

read "S perp".

In words: v belongs to $S^{\perp}$ when v is orthogonal (Definition 11.1.4) to **every** single element of S, not merely to some of them. The quantifier is what makes the definition useful, and it is

also what makes $S^\perp$ a subspace even when S is not. Note that the condition could equally be written $\langle s, v\rangle = 0$ for every $s \in S$, since $\langle s, v\rangle = \overline{\langle v, s\rangle}$ and a complex number vanishes exactly when its conjugate does. We will use whichever slot is convenient.

Proposition 11.3.1. Perp is always a subspace

For every subset S of an inner product space V, the set $S^\perp$ is a subspace of V. Moreover $S^\perp = \operatorname{span}(S)^\perp$.

Proof. We apply the Subspace Test (Theorem 1.2.1). First, $\langle 0, s\rangle = 0$ for every s, because $\langle \cdot, s\rangle$ is linear in the first slot and a linear map sends 0 to 0 (Theorem 2.1.1); so $0 \in S^\perp$. Next, let $v, w \in S^\perp$ and $a \in F$. For every $s \in S$,

$$\langle v + w, s\rangle = \langle v, s\rangle + \langle w, s\rangle = 0, \qquad \langle av, s\rangle = a\langle v, s\rangle = 0,$$

both by linearity in the first slot. Hence $v + w \in S^\perp$ and $av \in S^\perp$.

For the second claim, $S \subseteq \operatorname{span}(S)$ gives $\operatorname{span}(S)^\perp \subseteq S^\perp$. Conversely let $v \in S^\perp$ and let $u = a_1 s_1 + \cdots + a_k s_k$ with $s_i \in S$. Then

$$\langle u, v\rangle = a_1\langle s_1, v\rangle + \cdots + a_k\langle s_k, v\rangle = 0,$$

since each $\langle s_i, v\rangle = \overline{\langle v, s_i\rangle} = 0$. So $\langle v, u\rangle = \overline{\langle u, v\rangle} = 0$, and $v \in \operatorname{span}(S)^\perp$.

■

The second half is what makes computation possible: to test membership in $U^\perp$ it is enough to test against a spanning list of U, not against all of U.

Examples.

- **The extremes.** $\{0\}^\perp = V$, since $\langle v, 0\rangle = 0$ always. And $V^\perp = \{0\}$: if $v \in V^\perp$, then in particular $\langle v, v\rangle = 0$, and positive definiteness forces $v = 0$. This is the "pair it with itself" move, and it is the shape in which positivity almost always enters; it is the only place where positivity is used; we will use it again and again.
- **A plane in $\mathbb{R}^3$.** Let $U = \{(x, y, z) : x + y + z = 0\}$ with the dot product. The condition $x + y + z = 0$ says exactly $\langle (x, y, z), (1, 1, 1)\rangle = 0$, so $U = \{(1, 1, 1)\}^\perp$. The Orthogonal Decomposition Theorem below will give the converse, $U^\perp = \operatorname{span}((1, 1, 1))$.
- **A subset that is not a subspace.** In $\mathbb{R}^3$, take $S = \{(1, 0, 0), (1, 1, 0)\}$, two vectors and nothing else. Then $S^\perp = \{(x, y, z) : x = 0,\ x + y = 0\} = \operatorname{span}((0, 0, 1))$, a subspace, as Proposition 11.3.1 promises. It equals $\operatorname{span}(S)^\perp$, the perp of the xy-plane.
- **Polynomials.** On $\mathbb{R}[x]_{\le 2}$ with $\langle p, q\rangle = \int_{-1}^{1} p(t)q(t)\,dt$, let $U = \operatorname{span}(1)$ be the constants. Then $p \in U^\perp$ means $\int_{-1}^{1} p(t)\,dt = 0$: the polynomials of average value zero. For instance x and $x^2 - \frac{1}{3}$ lie in $U^\perp$, and they span it, since the Orthogonal Decomposition Theorem below gives $\dim U^\perp = 3 - 1 = 2$.

Nonexample by minimal change. Weaken "for every $\mathbf{s} \in S$" to "for some $\mathbf{s} \in S$". With $S = \{\mathbf{e}_1, \mathbf{e}_2\}$ in $\mathbb{R}^2$, the weakened set is $\{\mathbf{v} : \langle \mathbf{v}, \mathbf{e}_1\rangle = 0 \text{ or } \langle \mathbf{v}, \mathbf{e}_2\rangle = 0\}$, the union of the two axes. Closure under scalars still holds, and $\mathbf{0}$ still belongs; what fails is closure under addition, since $(1, 0) + (0, 1) = (1, 1)$ lies on neither axis. This is the union of Example (The Union of the Two Axes) again, in new clothing. The universal quantifier is doing real work.

Why this definition. One could try to define "the" complement of U by choosing a basis of U, extending it, and spanning the new vectors, which is what Theorem 1.7.6 does. That answer depends on every choice made along the way. The definition above depends only on U and on the inner product, and the next theorem shows that it really is a complement.

11.3.2 The orthogonal decomposition

Here is the theorem that makes $U^{\perp}$ more than a definition. Note the hypothesis: U must be finite-dimensional, but V need not be. The reason is that the proof needs an orthonormal basis of U, and Gram–Schmidt supplies one only for a finite list.

Theorem 11.3.2. Orthogonal Decomposition Theorem

Let V be an inner product space over F and let U be a **finite-dimensional** subspace of V. Then:

(a) $V = U \oplus U^{\perp}$;
(b) $(U^{\perp})^{\perp} = U$;
(c) if in addition V is finite-dimensional, then $\dim U^{\perp} = \dim V - \dim U$.

Idea. For (a) we must produce, from an arbitrary v, a piece inside U and a piece perpendicular to U. Fix an orthonormal basis of U. Theorem 11.2.1 says that a vector **of** U is the sum of its coordinates against that basis; so for a v that is not in U, the same sum is the natural guess for the U-piece. Then we subtract and check. That is the move of Gram–Schmidt — project, then subtract — used once instead of repeatedly. Uniqueness of the splitting is the "pair with itself" move. For (b), one inclusion is free and the other is (a) applied twice.

Proof.

(a) Since U is finite-dimensional, it has an orthonormal basis $(e_1, \dots, e_k)$ by Theorem 11.2.2. Let $v \in V$ and put

$$u = \sum_{i=1}^{k} \langle v, e_i \rangle e_i \in U, \qquad w = v - u.$$

For each j,

$$\langle w, e_j \rangle = \langle v, e_j \rangle - \sum_{i=1}^{k} \langle v, e_i \rangle \langle e_i, e_j \rangle = \langle v, e_j \rangle - \langle v, e_j \rangle = 0,$$

where the middle step uses $\langle e_i, e_j \rangle = \delta_{ij}$. So w is orthogonal to each e_j, hence to $\operatorname{span}(e_1, \dots, e_k) = U$ by Proposition 11.3.1. Therefore $v = u + w \in U + U^{\perp}$.
The sum is direct: if $x \in U \cap U^{\perp}$, then $\langle x, x \rangle = 0$ because $x \in U^{\perp}$ and $x \in U$, so $x = 0$ by positive definiteness. By Theorem 1.7.4, $V = U \oplus U^{\perp}$.

(b) ($\subseteq$) Let $u \in U$. For every $w \in U^{\perp}$ we have $\langle w, u \rangle = 0$, hence $\langle u, w \rangle = \overline{\langle w, u \rangle} = 0$. So $u \in (U^{\perp})^{\perp}$.
($\supseteq$) Let $v \in (U^{\perp})^{\perp}$. By (a), write $v = u + w$ with $u \in U$ and $w \in U^{\perp}$. By the inclusion just proved, $u \in (U^{\perp})^{\perp}$, so $w = v - u \in (U^{\perp})^{\perp}$ as well. Thus w lies in both $U^{\perp}$ and $(U^{\perp})^{\perp}$, giving $\langle w, w \rangle = 0$ and $w = 0$. Hence $v = u \in U$.

(c) By (a), $U^\perp$ is a complement of U in V, so $\dim U^\perp = \dim V - \dim U$ by Theorem 1.7.6.

■

Part (b) is worth pausing on. For a general subspace of a general vector space there is no operation "$\perp$" to apply twice, and even for the annihilator of Chapter 5 the analogous statement needed the double dual. Here it is a two-line consequence of (a). Part (c) says that $\perp$ exchanges dimension k with dimension $n - k$, just as the annihilator does (Theorem 5.2.2); Section 5 explains why the two facts are the same fact.

Quick check. Positive definiteness was used twice in the proof of Theorem 11.3.2, in the same way each time. Where? To see that it cannot be dropped, take $\mathbb{R}^2$ with the indefinite form $\beta(\mathbf{x}, \mathbf{y}) = x_1y_1 - x_2y_2$, let $U = \operatorname{span}((1, 1))$, and write $U^\perp$ for the set of $\mathbf{v}$ with $\beta(\mathbf{v}, \mathbf{u}) = 0$ for all $\mathbf{u} \in U$. Compute $U \cap U^\perp$.

Solution. It was used to show $U \cap U^\perp = \{0\}$ in (a), and again for $w = 0$ in (b); both times through "$\langle x, x\rangle = 0$ forces $x = 0$".

With β, the vector $\mathbf{u} = (1, 1)$ satisfies $\beta(\mathbf{u}, \mathbf{u}) = 1 - 1 = 0$, so $\mathbf{u}$ is "orthogonal to itself" and $\mathbf{u} \in U^\perp$. Hence $U \cap U^\perp = U \neq \{\mathbf{0}\}$, and the sum $U + U^\perp$ is not direct. Indeed $U^\perp = \{(x_1, x_2) : x_1 - x_2 = 0\} = U$ here, so $U + U^\perp = U$ is not all of $\mathbb{R}^2$. Both conclusions of the theorem fail.

11.3.3 Orthogonal projection

Once $V = U \oplus U^\perp$, Chapter 2 hands us an operator: the projection onto U along $U^\perp$ (Theorem 4.5.1). It deserves its own name and symbol, because the complement is no longer a choice.

Definition 11.3.2. Orthogonal projection

Let U be a finite-dimensional subspace of an inner product space V. By Theorem 11.3.2 (a), every $v \in V$ has a unique splitting $v = u + w$ with $u \in U$ and $w \in U^\perp$. The **orthogonal projection onto** U is the operator $P_U \in \mathcal{L}(V)$ defined by

$$P_U v = u.$$

The definition makes sense because the splitting exists and is unique, and P_U is linear by Theorem 4.5.1 (b), which also gives $P_U^2 = P_U$, $\operatorname{im} P_U = U$, and $\ker P_U = U^\perp$. Nothing so far tells us how to compute $P_U v$. The proof of Theorem 11.3.2 does, and that is the next theorem.

Theorem 11.3.3. Projection formula

Let U be a finite-dimensional subspace of an inner product space V, with orthonormal basis $(e_1, \dots, e_k)$. Then for every $v \in V$:

(a) $P_U v = \sum_{i=1}^{k} \langle v, e_i \rangle e_i$;

(b) $\langle P_U u, v \rangle = \langle u, P_U v \rangle$ for all $u, v \in V$;

(c) $\|P_U v\| \le \|v\|$, with equality if and only if $v \in U$.

Idea. Part (a) is not a new computation: the vector u built in the proof of Theorem 11.3.2 (a) was exactly this sum, and by uniqueness of the splitting it must be $P_U v$. For (b), split **both** u and v, and watch the cross terms die. For (c), split v and use Pythagoras.

Proof.

(a) Put $u = \sum_i \langle v, e_i \rangle e_i$ and $w = v - u$. The proof of Theorem 11.3.2 (a) showed $u \in U$ and $w \in U^{\perp}$. Since the splitting of v in $U \oplus U^{\perp}$ is unique (Definition 1.7.2), $P_U v = u$.

(b) Write $u = u_1 + u_2$ and $v = v_1 + v_2$ with $u_1, v_1 \in U$ and $u_2, v_2 \in U^{\perp}$. Then $\langle u_1, v_2 \rangle = 0$ because $v_2 \in U^{\perp}$ and $u_1 \in U$, and $\langle u_2, v_1 \rangle = 0$ for the same reason with the roles swapped. Hence

$$\langle P_U u, v \rangle = \langle u_1, v_1 + v_2 \rangle = \langle u_1, v_1 \rangle = \langle u_1 + u_2, v_1 \rangle = \langle u, P_U v \rangle.$$

(c) With $v = v_1 + v_2$ as above, v_1 and v_2 are orthogonal, so $\|v\|^2 = \|v_1\|^2 + \|v_2\|^2$ by Theorem 11.1.3. Since $\|v_2\|^2 \ge 0$, this gives $\|P_U v\|^2 = \|v_1\|^2 \le \|v\|^2$. Equality holds if and only if $\|v_2\| = 0$, that is $v_2 = 0$, that is $v = v_1 \in U$.

■

Two comments. First, the right-hand side of (a) visibly depends on the orthonormal basis chosen, while the left-hand side does not; so the sum $\sum_i \langle v, e_i \rangle e_i$ is the same for every orthonormal basis of U. That is a genuine piece of information, obtained for free. Second, property (b) says that P_U can be moved from one slot of the inner product to the other at no cost. Section 6 will call such an operator **self-adjoint** and write the property as $P_U^* = P_U$; we proved the property here so that nothing later depends on the order of the sections.

Orthonormal bases are convenient but not obligatory, and by hand it is usually faster to skip the square roots.

Corollary 11.3.4. Projecting with an orthogonal basis

Let U be a finite-dimensional subspace of an inner product space V, and let $(u_1, \dots, u_k)$ be an **orthogonal** basis of U, so that $\langle u_i, u_j \rangle = 0$ for $i \ne j$ and every $u_i \ne 0$. Then

$$P_U v = \sum_{i=1}^{k} \frac{\langle v, u_i \rangle}{\langle u_i, u_i \rangle} u_i \qquad \text{for every } v \in V.$$

Proof. Each u_i is nonzero, so $\|u_i\| > 0$ and $e_i = u_i / \|u_i\|$ is defined. Each e_i is a nonzero multiple of u_i, so $(e_1, \dots, e_k)$ is again a basis of U, and it is orthonormal: $\langle e_i, e_j \rangle =$

$\langle u_i, u_j\rangle/(\|u_i\|\|u_j\|) = \delta_{ij}$. By Theorem 11.3.3 (a),

$$P_U v = \sum_{i=1}^{k} \langle v, e_i\rangle\, e_i = \sum_{i=1}^{k} \frac{\langle v, u_i\rangle}{\|u_i\|^2}\, u_i,$$

where the two factors $1/\|u_i\|$ come out unchanged, the one in the second slot because $\|u_i\|$ is real. Finally $\|u_i\|^2 = \langle u_i, u_i\rangle$ by Definition 11.1.3.

■

Each summand $\langle v, u_i\rangle u_i/\langle u_i, u_i\rangle$ is the projection of v onto the line $\operatorname{span}(u_i)$: the case $k = 1$ of the corollary.

11.3.4 Best approximation and distance

Now the payoff. Of all the vectors of U, which is nearest to a given v? In $\mathbb{R}^3$, dropping a perpendicular from a point to a plane is the familiar recipe, and the recipe is correct in every inner product space.

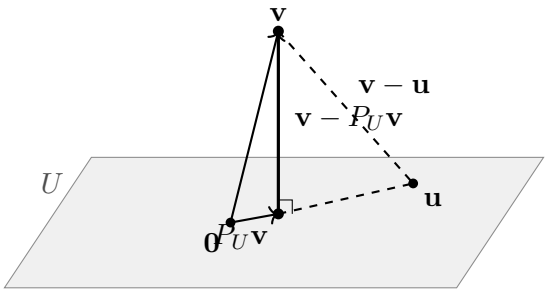
Theorem 11.3.5. Best Approximation Theorem

Let U be a finite-dimensional subspace of an inner product space V, and let $v \in V$. Then

$$\|v - P_U v\| \le \|v - u\| \qquad \text{for every } u \in U,$$

and equality holds **only** for $u = P_U v$.

Idea. Insert $P_U v$ into the difference: $v - u = (v - P_U v) + (P_U v - u)$. The first bracket is perpendicular to U and the second lies in U, so the two are orthogonal, and Pythagoras turns the comparison into "a sum of two squares is at least one of them". The picture is a right triangle with hypotenuse $v - u$.



Proof. Let $u \in U$ and write

$$v - u = (v - P_U v) + (P_U v - u).$$

The first bracket lies in $U^\perp$, since $v - P_U v$ is the $U^\perp$-part of v in Theorem 11.3.2 (a). The second lies in U, since $P_U v \in U$ and $u \in U$. Hence the two brackets are orthogonal, and Theorem 11.1.3 gives

$$\|v - u\|^2 = \|v - P_U v\|^2 + \|P_U v - u\|^2 \ge \|v - P_U v\|^2.$$

Taking square roots proves the inequality, since the norm is nonnegative. Equality forces $\|P_U v - u\|^2 = 0$, hence $P_U v - u = 0$ by positive definiteness, that is $u = P_U v$. This proves the theorem.

■

So the minimum is attained, and attained at exactly one point. That lets us define a distance.

Definition 11.3.3. Distance to a subspace

Let U be a finite-dimensional subspace of an inner product space V and let $v \in V$. The **distance from v to U** is

$$d(v, U) = \min_{u \in U} \|v - u\| = \|v - P_U v\|.$$

The word **minimum** is justified, not sloppy: Theorem 11.3.5 says the values $\|v - u\|$ really do have a smallest one. Note also that $d(v, U) = 0$ if and only if $v \in U$, and that, when V is finite-dimensional, $d(v, U) = \|P_{U^\perp} v\|$, since $v - P_U v$ is the $U^\perp$-component of v. (The hypothesis is needed for $P_{U^\perp}$ to be defined at all: Definition 11.3.2 asks the subspace projected onto to be finite-dimensional, and $U^\perp$ is not, in the infinite-dimensional setting this section is careful to allow.) When $U^\perp$ is the smaller of the two, computing that component is the faster route.

Example. Distance from a point to a plane

In $\mathbb{R}^3$ with the dot product, let $U = \{(x, y, z) : x + y + z = 0\}$ and $\mathbf{v} = (1, 2, 3)$. Find $P_U \mathbf{v}$ and $d(\mathbf{v}, U)$, twice: once through $U^\perp$, and once through an orthonormal basis of U.

Solution. Through $U^\perp$. Put $\mathbf{n} = (1, 1, 1)$. Every $(x, y, z) \in U$ satisfies $\langle (x, y, z), \mathbf{n} \rangle = x + y + z = 0$, so $\mathbf{n} \in U^\perp$. Since U is the solution set of one nontrivial equation, $\dim U = 2$, so $\dim U^\perp = 3 - 2 = 1$ by Theorem 11.3.2 (c), and therefore $U^\perp = \operatorname{span}(\mathbf{n})$. By Corollary 11.3.4 with the one-element basis $(\mathbf{n})$,

$$P_{U^\perp} \mathbf{v} = \frac{\langle \mathbf{v}, \mathbf{n} \rangle}{\langle \mathbf{n}, \mathbf{n} \rangle} \mathbf{n} = \frac{1 + 2 + 3}{3} (1, 1, 1) = (2, 2, 2),$$

so $P_U \mathbf{v} = \mathbf{v} - (2, 2, 2) = (-1, 0, 1)$ and $d(\mathbf{v}, U) = \|(2, 2, 2)\| = 2\sqrt{3}$.

Through an orthonormal basis of U. The vectors $(1, -1, 0)$ and $(1, 1, -2)$ lie in U and are orthogonal to each other, so

$$\mathbf{e}_1 = \tfrac{1}{\sqrt{2}}(1, -1, 0), \qquad \mathbf{e}_2 = \tfrac{1}{\sqrt{6}}(1, 1, -2)$$

is an orthonormal basis of U. Then $\langle \mathbf{v}, \mathbf{e}_1 \rangle = (1 - 2)/\sqrt{2} = -1/\sqrt{2}$ and $\langle \mathbf{v}, \mathbf{e}_2 \rangle = (1 + 2 - 6)/\sqrt{6} = -3/\sqrt{6}$, so Theorem 11.3.3 (a) gives

$$P_U \mathbf{v} = -\tfrac{1}{\sqrt{2}} \mathbf{e}_1 - \tfrac{3}{\sqrt{6}} \mathbf{e}_2 = -\tfrac{1}{2}(1, -1, 0) - \tfrac{1}{2}(1, 1, -2) = (-1, 0, 1),$$

the same answer. *Check:* $(-1) + 0 + 1 = 0$, so $P_U \mathbf{v} \in U$; and $\mathbf{v} - P_U \mathbf{v} = (2, 2, 2)$ is a multiple of $\mathbf{n}$, so it lies in $U^\perp$. The two conditions together identify the splitting, by uniqueness.

The next example is the same theorem in a space of functions, where "closest" is no longer something one can see.

Example. The best quadratic approximation to a cubic

On $V = \mathbb{R}[x]_{\le 3}$ with $\langle p, q\rangle = \int_{-1}^{1} p(t)q(t)\,dt$, let $U = \mathbb{R}[x]_{\le 2}$. Find the polynomial of U closest to x^3, and the distance.

Solution. The list $(1, x, x^2 - \frac{1}{3})$ is a basis of U, being three independent polynomials of degrees $0, 1, 2$ (Theorem 1.4.4) in a space of dimension 3. It is **orthogonal**: $\langle 1, x\rangle = 0$ and $\langle x, x^2 - \frac{1}{3}\rangle = 0$ because the integrands are odd, and

$$\langle 1, x^2 - \tfrac{1}{3}\rangle = \int_{-1}^{1} \left(t^2 - \tfrac{1}{3}\right) dt = \tfrac{2}{3} - \tfrac{2}{3} = 0.$$

Now apply Corollary 11.3.4 to $v = x^3$. Two of the three coefficients vanish because t^3 is odd: $\langle x^3, 1\rangle = 0$ and $\langle x^3, x^2 - \frac{1}{3}\rangle = 0$. The remaining one is

$$\frac{\langle x^3, x\rangle}{\langle x, x\rangle} = \frac{\int_{-1}^{1} t^4\,dt}{\int_{-1}^{1} t^2\,dt} = \frac{2/5}{2/3} = \frac{3}{5}.$$

Hence $P_U(x^3) = \frac{3}{5}x$. The error is $x^3 - \frac{3}{5}x$, with

$$\begin{aligned}\|x^3 - \tfrac{3}{5}x\|^2 &= \int_{-1}^{1} \left(t^6 - \tfrac{6}{5}t^4 + \tfrac{9}{25}t^2\right) dt \\ &= \tfrac{2}{7} - \tfrac{6}{5}\cdot\tfrac{2}{5} + \tfrac{9}{25}\cdot\tfrac{2}{3} = \tfrac{2}{7} - \tfrac{6}{25} = \tfrac{8}{175},\end{aligned}$$

so $d(x^3, U) = \sqrt{8/175} = \frac{2}{35}\sqrt{14} \approx 0.214$.

Check. The error should be orthogonal to U. Against 1 and $x^2 - \frac{1}{3}$ this holds because $t^3 - \frac{3}{5}t$ is odd; against x it reads $\int_{-1}^{1}(t^4 - \frac{3}{5}t^2)\,dt = \frac{2}{5} - \frac{3}{5}\cdot\frac{2}{3} = 0$. Notice that the answer is **not** the Taylor polynomial of x^3 at 0, which is 0: Taylor approximates near one point, while P_U spreads the error over the whole interval.

11.3.5 Three traps

Warning. $U^{\perp}$ **is not the annihilator** U^0**.** Both objects consist of things that "kill" U, but they live in different places: $U^{\perp} \subseteq V$ is made of **vectors** (Definition 11.3.1), while $U^0 \subseteq V^*$ is made of **functionals** (Definition 5.2.1). The annihilator exists over every field and needs no extra structure; the orthogonal complement needs an inner product, so it needs $F = \mathbb{R}$ or $F = \mathbb{C}$. They do have the same dimension, $\dim V - \dim U$ (Theorem 5.2.2, Theorem 11.3.2 (c)), and that is not a coincidence: Section 5 produces an explicit bijection $U^{\perp} \to U^0$. Until then, keep them apart.

Warning. **"Orthogonal projection" is a special case of Chapter 2's projections, not a synonym.** An operator with $P^2 = P$ is a projection onto $\operatorname{im} P$ **along** $\ker P$ (Def-

inition 4.5.1, Theorem 4.5.1), and Chapter 4, §05 stressed that the direction matters: on $\mathbb{R}^2$, both $\begin{pmatrix}1&0\\0&0\end{pmatrix}$ and $\begin{pmatrix}1&-1\\0&0\end{pmatrix}$ are projections onto the x-axis. Only the first is *orthogonal*, because only for the first is the kernel the perpendicular line; the second projects along $\operatorname{span}((1,1))$ and is called **oblique**. So P_U carries more information than "$P^2 = P$ and $\operatorname{im} P = U$": it also says which complement is used. Theorem 11.3.3 (b) is the extra property, and Exercise C3 below shows that it singles out the orthogonal projections among all projections.

Warning. $U^\perp$ **depends on the inner product, not only on** U**.** In $\mathbb{R}^2$ let $U = \operatorname{span}((1,1))$. With the dot product, $U^\perp = \operatorname{span}((1,-1))$. With the weighted inner product $\langle \mathbf{x}, \mathbf{y}\rangle = 2x_1y_1 + x_2y_2$, a vector (a,b) is orthogonal to $(1,1)$ when $2a + b = 0$, so $U^\perp = \operatorname{span}((1,-2))$. Same subspace, different perpendicular. The symbol $\perp$ always hides a choice of $\langle\cdot,\cdot\rangle$, and when two inner products are in play the notation must say which.

11.3.6 Exercises

11.3.6.1 A. Check your understanding

Exercise. A1

(a) Define the orthogonal complement $S^\perp$ of a subset S of an inner product space.

(b) True or false: $S^\perp$ is a subspace only when S is a subspace. Justify your answer.

(c) State the Orthogonal Decomposition Theorem, including every hypothesis.

(d) Let U be a 3-dimensional subspace of $\mathbb{R}^7$. What is $\dim U^\perp$, and what is $\dim (U^\perp)^\perp$?

(e) What quantity does $P_U v$ minimize, and over what set?

(f) Name two differences between $U^\perp$ and the annihilator U^0.

Solution.

(a) $S^\perp = \{v \in V : \langle v, s\rangle = 0 \text{ for every } s \in S\}$ (Definition 11.3.1).

(b) False. $S^\perp$ is a subspace for **every** subset S, by Proposition 11.3.1, and it equals $\operatorname{span}(S)^\perp$. For instance $\{(1,0,0),(1,1,0)\}^\perp = \operatorname{span}((0,0,1))$ in $\mathbb{R}^3$.

(c) If V is an inner product space over $\mathbb{R}$ or $\mathbb{C}$ and U is a **finite-dimensional** subspace, then $V = U \oplus U^\perp$ and $(U^\perp)^\perp = U$; if V is also finite-dimensional, then $\dim U^\perp = \dim V - \dim U$ (Theorem 11.3.2). V itself need not be finite-dimensional.

(d) $\dim U^\perp = 7 - 3 = 4$ and $\dim (U^\perp)^\perp = \dim U = 3$, by Theorem 11.3.2 (c) and (b).

(e) It minimizes $\|v - u\|$ over $u \in U$, and it is the **only** minimizer (Theorem 11.3.5).

(f) $U^\perp$ consists of vectors of V, U^0 of functionals in V^*; and $U^\perp$ needs an inner product (so $F = \mathbb{R}$ or $\mathbb{C}$), while U^0 is defined over any field. A third difference: $U^\perp$ is a complement of U inside V, whereas U^0 is not a subspace of V at all.

11.3.6.2 B. Practice

Exercise. B1: A complement and a projection in $\mathbb{R}^4$

In $\mathbb{R}^4$ with the dot product, let $U = \operatorname{span}\big((1,1,0,0),\ (0,0,1,1)\big)$.

(a) Find a basis of $U^\perp$ and verify the dimension count.

(b) Compute $P_U\mathbf{v}$ and $d(\mathbf{v}, U)$ for $\mathbf{v} = (1,2,3,4)$.

(c) Verify directly that $\mathbf{v} - P_U\mathbf{v} \in U^\perp$.

Solution.

(a) By Proposition 11.3.1 it is enough to be orthogonal to the two spanning vectors, so $(a,b,c,d) \in U^\perp$ means $a+b=0$ and $c+d=0$. Hence

$$U^\perp = \operatorname{span}\big((1,-1,0,0),\ (0,0,1,-1)\big),$$

these two being independent (look at the first and third entries). So $\dim U^\perp = 2 = 4-2$, as Theorem 11.3.2 (c) requires.

(b) The spanning list of U is orthogonal, so Corollary 11.3.4 applies with $\mathbf{u}_1 = (1,1,0,0)$, $\mathbf{u}_2 = (0,0,1,1)$, both of squared norm 2:

$$P_U\mathbf{v} = \frac{1+2}{2}\mathbf{u}_1 + \frac{3+4}{2}\mathbf{u}_2 = \left(\tfrac{3}{2}, \tfrac{3}{2}, \tfrac{7}{2}, \tfrac{7}{2}\right).$$

Then $\mathbf{v} - P_U\mathbf{v} = (-\tfrac{1}{2}, \tfrac{1}{2}, -\tfrac{1}{2}, \tfrac{1}{2})$ and $d(\mathbf{v}, U) = \sqrt{4 \cdot \tfrac{1}{4}} = 1$.

(c) The vector $(-\tfrac{1}{2}, \tfrac{1}{2}, -\tfrac{1}{2}, \tfrac{1}{2})$ satisfies $a+b=0$ and $c+d=0$, so it lies in $U^\perp$ by (a).

Exercise. B2: Distance from a point to a plane

In $\mathbb{R}^3$ with the dot product, let $U = \{(x,y,z) : 2x - y + 2z = 0\}$ and $\mathbf{v} = (1,2,3)$. Compute $d(\mathbf{v}, U)$ and $P_U\mathbf{v}$, and check that $P_U\mathbf{v}$ lies in U. Hence write down the general formula for the distance from $\mathbf{v}$ to the plane $\{\mathbf{x} : \langle \mathbf{x}, \mathbf{n}\rangle = 0\}$, $\mathbf{n} \neq \mathbf{0}$.

Solution. The plane is $\{\mathbf{n}\}^\perp$ for $\mathbf{n} = (2,-1,2)$, and $\mathbf{n} \in U^\perp$ with $\dim U^\perp = 3 - 2 = 1$ (Theorem 11.3.2 (c)), so $U^\perp = \operatorname{span}(\mathbf{n})$. Here $\langle \mathbf{v}, \mathbf{n}\rangle = 2-2+6 = 6$ and $\langle \mathbf{n}, \mathbf{n}\rangle = 4+1+4 = 9$, so

$$P_{U^\perp}\mathbf{v} = \tfrac{6}{9}\mathbf{n} = \left(\tfrac{4}{3}, -\tfrac{2}{3}, \tfrac{4}{3}\right), \qquad P_U\mathbf{v} = \mathbf{v} - P_{U^\perp}\mathbf{v} = \left(-\tfrac{1}{3}, \tfrac{8}{3}, \tfrac{5}{3}\right).$$

Check: $2(-\tfrac{1}{3}) - \tfrac{8}{3} + 2 \cdot \tfrac{5}{3} = \tfrac{-2-8+10}{3} = 0$, so $P_U\mathbf{v} \in U$. The distance is $d(\mathbf{v}, U) = \|\tfrac{6}{9}\mathbf{n}\| = \tfrac{6}{9} \cdot 3 = 2$.

In general $d(\mathbf{v}, U) = \|P_{U^\perp}\mathbf{v}\| = |\langle \mathbf{v}, \mathbf{n}\rangle|\,\|\mathbf{n}\|/\|\mathbf{n}\|^2 = |\langle \mathbf{v}, \mathbf{n}\rangle|/\|\mathbf{n}\|$, which for $\mathbf{n} = (2,-1,2)$ and $\mathbf{v} = (1,2,3)$ gives $6/3 = 2$.

Exercise. B3: The best quadratic approximation to a quartic

On $\mathbb{R}[x]_{\le 4}$ with $\langle p,q\rangle = \int_{-1}^{1} p(t)q(t)\,\mathrm{d}t$, let $U = \mathbb{R}[x]_{\le 2}$. Find $P_U(x^4)$ and $d(x^4, U)$. *Hint: the orthogonal basis of Example (The best quadratic approximation to a cubic) still works.*

Solution. Use the orthogonal basis $(1, x, x^2 - \frac{1}{3})$ of U from Example (The best quadratic approximation to a cubic), with $\langle 1,1\rangle = 2$ and

$$\langle x^2 - \tfrac{1}{3}, x^2 - \tfrac{1}{3}\rangle = \int_{-1}^{1} \Big(t^4 - \tfrac{2}{3}t^2 + \tfrac{1}{9}\Big)\mathrm{d}t = \tfrac{2}{5} - \tfrac{4}{9} + \tfrac{2}{9} = \tfrac{8}{45}.$$

The coefficients against x^4 are $\langle x^4, 1\rangle = \frac{2}{5}$, $\langle x^4, x\rangle = 0$ (odd integrand) and

$$\langle x^4, x^2 - \tfrac{1}{3}\rangle = \int_{-1}^{1} \Big(t^6 - \tfrac{1}{3}t^4\Big)\mathrm{d}t = \tfrac{2}{7} - \tfrac{1}{3}\cdot\tfrac{2}{5} = \tfrac{16}{105}.$$

By Corollary 11.3.4,

$$P_U(x^4) = \frac{2/5}{2}\cdot 1 + \frac{16/105}{8/45}\Big(x^2 - \tfrac{1}{3}\Big) = \tfrac{1}{5} + \tfrac{6}{7}\Big(x^2 - \tfrac{1}{3}\Big) = \tfrac{6}{7}x^2 - \tfrac{3}{35}.$$

For the distance, Theorem 11.1.3 applied to $x^4 = P_U(x^4) + (x^4 - P_U(x^4))$ gives $d^2 = \|x^4\|^2 - \|P_U(x^4)\|^2$. Here $\|x^4\|^2 = \int_{-1}^{1} t^8\,\mathrm{d}t = \frac{2}{9}$ and, the basis being orthogonal,

$$\|P_U(x^4)\|^2 = \Big(\tfrac{1}{5}\Big)^2\cdot 2 + \Big(\tfrac{6}{7}\Big)^2\cdot\tfrac{8}{45} = \tfrac{2}{25} + \tfrac{32}{245} = \tfrac{258}{1225}.$$

Hence $d^2 = \frac{2}{9} - \frac{258}{1225} = \frac{2450-2322}{11025} = \frac{128}{11025}$ and $d(x^4, U) = \frac{8\sqrt{2}}{105} \approx 0.108$.

11.3.6.3 C. Going deeper

Exercise. C1: Perp turns sums into intersections

Let U and W be subspaces of an inner product space V.

(a) Prove that $(U+W)^{\perp} = U^{\perp} \cap W^{\perp}$. (No finiteness is needed.)

(b) Assume now that V is finite-dimensional. Deduce that $(U \cap W)^{\perp} = U^{\perp} + W^{\perp}$.

Solution.

(a) ($\subseteq$) Let $v \in (U+W)^{\perp}$. Since $U \subseteq U+W$ (Theorem 1.7.1), v is orthogonal to every element of U, so $v \in U^{\perp}$; likewise $v \in W^{\perp}$.
($\supseteq$) Let $v \in U^{\perp} \cap W^{\perp}$ and let $u + w \in U + W$ with $u \in U$, $w \in W$. Then $\langle v, u+w\rangle = \langle v, u\rangle + \langle v, w\rangle = 0$, using additivity in the second slot. Hence $v \in (U+W)^{\perp}$.

(b) Apply (a) to the subspaces $U^{\perp}$ and $W^{\perp}$:

$$(U^{\perp} + W^{\perp})^{\perp} = (U^{\perp})^{\perp} \cap (W^{\perp})^{\perp} = U \cap W,$$

where the second equality is Theorem 11.3.2 (b), available because V is finite-dimensional, so all subspaces are (Theorem 1.6.7). Now take $\perp$ of both ends and

use Theorem 11.3.2 (b) once more, this time on the subspace $U^\perp + W^\perp$:

$$U^\perp + W^\perp = \big((U^\perp + W^\perp)^\perp\big)^\perp = (U \cap W)^\perp.$$

This proves the second identity.

Exercise. C2: A subspace with no orthogonal complement

Let $V = \mathbb{R}[x]$, with the **coefficient inner product**

$$\langle p, q\rangle = \sum_{k\ge 0} a_k b_k \quad \text{for } p = \sum_k a_k x^k,\ q = \sum_k b_k x^k,$$

a finite sum, since a polynomial has only finitely many nonzero coefficients. Let $U = \{\, p \in V : p(1) = 0 \,\}$.

(a) Verify that $\langle \cdot, \cdot \rangle$ is an inner product on V and that U is a subspace with $U \neq V$.

(b) Prove that $U^\perp = \{0\}$. Conclude that $U \oplus U^\perp \neq V$, so Theorem 11.3.2 fails without the hypothesis that U is finite-dimensional.

(c) Explain what goes wrong geometrically by exhibiting, for each $m \ge 1$, an element $p_m \in U$ with $\|p_m - 1\| = 1/\sqrt{m}$.

Solution.

(a) The form is symmetric and bilinear over $\mathbb{R}$ because each term $a_k b_k$ is, and $\langle p, p\rangle = \sum_k a_k^2 > 0$ whenever some $a_k \neq 0$, that is whenever $p \neq 0$. So it is an inner product. The map $\varepsilon\colon p \mapsto p(1)$ is linear (Theorem 0.8.5), so $U = \ker \varepsilon$ is a subspace (Theorem 2.2.1). It is proper, since the constant polynomial 1 has $\varepsilon(1) = 1 \neq 0$.

(b) Note first that $x^i - x^j \in U$ for all $i, j \ge 0$, since $1 - 1 = 0$. Let $q = \sum_k b_k x^k \in U^\perp$. Pairing with $x^i - x^j$ gives $b_i - b_j = 0$, so all the coefficients of q are equal, say to c. But only finitely many of them are nonzero, and there are infinitely many indices, so $c = 0$ and $q = 0$. Hence $U^\perp = \{0\}$ and $U + U^\perp = U \neq V$.

(c) Take $p_m = 1 - \frac{1}{m}(x + x^2 + \cdots + x^m)$. Then $p_m(1) = 1 - m \cdot \frac{1}{m} = 0$, so $p_m \in U$, while

$$\|p_m - 1\|^2 = \left\| \tfrac{1}{m}(x + \cdots + x^m) \right\|^2 = m \cdot \frac{1}{m^2} = \frac{1}{m}.$$

So the polynomial $1 \notin U$ has elements of U arbitrarily close to it: U leaves no room outside itself for a perpendicular direction, and the best approximation to 1 from U does not exist (the distances $\|1 - p\|$ get arbitrarily close to 0 without reaching it). The obstruction is not the dimension of V as such but the fact that U is not *closed*: filling in the missing limits is exactly what the theory of Hilbert spaces does, and there Theorem 11.3.2 is recovered for closed subspaces. That is analysis, not linear algebra, and we do not pursue it.

Exercise. C3: Which projections are orthogonal

Let V be a finite-dimensional inner product space and let $P \in \mathcal{L}(V)$ satisfy $P^2 = P$. Prove

that the following are equivalent:

(a) $P = P_U$ for some subspace U;
(b) $\langle Pu, v\rangle = \langle u, Pv\rangle$ for all $u, v \in V$;
(c) $\ker P = (\operatorname{im} P)^{\perp}$.

Hint: for (b) $\Rightarrow$ (c), pair a vector of $\ker P$ *with* Pv.

Solution.

(a) $\Rightarrow$ (b) is Theorem 11.3.3 (b).

(b) $\Rightarrow$ (c). ($\subseteq$) Let $x \in \ker P$ and let $Pv \in \operatorname{im} P$ be arbitrary. Then

$$\langle x, Pv\rangle = \overline{\langle Pv, x\rangle} = \overline{\langle v, Px\rangle} = \overline{\langle v, 0\rangle} = 0,$$

using (b) in the second equality. So $x \in (\operatorname{im} P)^{\perp}$. ($\supseteq$) Let $x \in (\operatorname{im} P)^{\perp}$. Then

$$\langle Px, Px\rangle = \langle x, P(Px)\rangle = \langle x, Px\rangle = 0,$$

the first equality by (b), the second by $P^2 = P$, and the third because $Px \in \operatorname{im} P$ while $x \in (\operatorname{im} P)^{\perp}$. Positive definiteness gives $Px = 0$, so $x \in \ker P$.

(c) $\Rightarrow$ (a). Put $U = \operatorname{im} P$. By Theorem 4.5.1 (a), $V = \operatorname{im} P \oplus \ker P = U \oplus U^{\perp}$, and P is the identity on U and zero on $\ker P = U^{\perp}$. By the uniqueness clause of Theorem 4.5.1 (b), P is the projection onto U along $U^{\perp}$, which is P_U by Definition 11.3.2.

11.4 Least Squares and Minimum-Norm Solutions

Chapter 3 taught us to decide whether $\mathbf{Ax} = \mathbf{b}$ has a solution, and to find them all when it does. Real data rarely cooperates. Measure a quantity that depends linearly on two unknowns, take ten measurements, and you have ten equations in two unknowns; each measurement carries a little error, and the ten equations are almost certainly inconsistent. Refusing to answer is not useful. This section replaces the question "which $\mathbf{x}$ solves $\mathbf{Ax} = \mathbf{b}$?" with "which $\mathbf{x}$ comes closest?", and shows that the new question always has an answer, computed by solving a square system.

The other half of the section handles the opposite complaint. When a system is consistent but underdetermined, there are too many solutions, and we want a canonical one. Orthogonality picks it: exactly one solution is perpendicular to the kernel, and it is the shortest.

Throughout, $F = \mathbb{R}$ or $F = \mathbb{C}$, and F^n and F^m carry the standard inner product $\langle \mathbf{x}, \mathbf{y}\rangle = \sum_i x_i \overline{y_i}$, with $\|\mathbf{x}\|^2 = \sum_i |x_i|^2$.

11.4.1 The residual, and a piece of notation

Given $\mathbf{A} \in M_{m\times n}(F)$ and $\mathbf{b} \in F^m$, write $\mathbf{r} = \mathbf{b} - \mathbf{Ax}$ for the **residual** of a candidate $\mathbf{x}$. The system is consistent exactly when some $\mathbf{x}$ makes $\mathbf{r} = \mathbf{0}$. When none does, we ask for the $\mathbf{x}$

making $\mathbf{r}$ as short as possible, that is, minimizing

$$\|\mathbf{A}\mathbf{x}-\mathbf{b}\|^2=\sum_{i=1}^{m}\bigl|(\mathbf{A}\mathbf{x})_i-b_i\bigr|^2,$$

a **sum of squares**; hence the name **least squares**.

Two observations turn this into geometry. First, as $\mathbf{x}$ ranges over F^n, the vector $\mathbf{A}\mathbf{x}$ ranges over exactly the column space $\operatorname{col}(\mathbf{A})\subseteq F^m$ (Theorem 0.9.3), a finite-dimensional subspace. Second, minimizing $\|\mathbf{A}\mathbf{x}-\mathbf{b}\|$ over all $\mathbf{x}$ is therefore the same as finding the point of $\operatorname{col}(\mathbf{A})$ nearest to $\mathbf{b}$. We already know that point: it is $P_{\operatorname{col}(\mathbf{A})}\mathbf{b}$, by Theorem 11.3.5. All that remains is to turn "the residual is perpendicular to $\operatorname{col}(\mathbf{A})$" into equations, and for that we need one piece of notation.

Recall the conjugate transpose $\mathbf{A}^*\in M_{n\times m}(F)$ of Definition 11.1.2: the matrix with entries $(\mathbf{A}^*)_{ij}=\overline{a_{ji}}$, obtained by transposing and then conjugating every entry, so that $\mathbf{A}^*=\mathbf{A}^{\top}$ over $\mathbb{R}$. Comparing entries and using Theorem 0.9.5 together with the rules for conjugation (Theorem 0.6.2) gives

$$(\mathbf{A}+\mathbf{B})^*=\mathbf{A}^*+\mathbf{B}^*,\quad (c\mathbf{A})^*=\overline{c}\,\mathbf{A}^*,\quad (\mathbf{A}\mathbf{B})^*=\mathbf{B}^*\mathbf{A}^*,\quad \mathbf{A}^{**}=\mathbf{A},$$

whenever the sizes match. The third is the one worth seeing done: $((\mathbf{A}\mathbf{B})^*)_{ij}=\overline{(\mathbf{A}\mathbf{B})_{ji}}=\overline{\sum_k a_{jk}b_{ki}}=\sum_k\overline{b_{ki}}\,\overline{a_{jk}}=\sum_k(\mathbf{B}^*)_{ik}(\mathbf{A}^*)_{kj}=(\mathbf{B}^*\mathbf{A}^*)_{ij}$, which is where the order reverses. The one consequence we really use is the following.

Lemma 11.4.1. Moving a matrix across the inner product

Let $\mathbf{A}\in M_{m\times n}(F)$. Then for all $\mathbf{x}\in F^n$ and $\mathbf{y}\in F^m$,

$$\langle\mathbf{A}\mathbf{x},\mathbf{y}\rangle=\langle\mathbf{x},\mathbf{A}^*\mathbf{y}\rangle,\qquad \langle\mathbf{y},\mathbf{A}\mathbf{x}\rangle=\langle\mathbf{A}^*\mathbf{y},\mathbf{x}\rangle.$$

Proof. Regarding vectors as columns, $\langle\mathbf{u},\mathbf{w}\rangle=\mathbf{w}^*\mathbf{u}$, a 1×1 matrix identified with its entry. Hence

$$\langle\mathbf{A}\mathbf{x},\mathbf{y}\rangle=\mathbf{y}^*(\mathbf{A}\mathbf{x})=(\mathbf{y}^*\mathbf{A})\mathbf{x}=((\mathbf{A}^*\mathbf{y})^*)\mathbf{x}=\langle\mathbf{x},\mathbf{A}^*\mathbf{y}\rangle,$$

where the third equality uses $(\mathbf{A}^*\mathbf{y})^*=\mathbf{y}^*\mathbf{A}^{**}=\mathbf{y}^*\mathbf{A}$. The second identity follows by conjugating the first, since $\langle\mathbf{y},\mathbf{A}\mathbf{x}\rangle=\overline{\langle\mathbf{A}\mathbf{x},\mathbf{y}\rangle}$ and $\overline{\langle\mathbf{x},\mathbf{A}^*\mathbf{y}\rangle}=\langle\mathbf{A}^*\mathbf{y},\mathbf{x}\rangle$. ∎

The lemma is the matrix case of a general principle: *to move a map across an inner product, pay with its adjoint.* Section 6 develops that principle for linear maps between abstract inner product spaces; here the two lines above are all we need.

11.4.2 The normal equations

Theorem 11.4.2. Least Squares Theorem

Let $\mathbf{A}\in M_{m\times n}(F)$ and $\mathbf{b}\in F^m$, and write $U=\operatorname{col}(\mathbf{A})$. For $\mathbf{x}\in F^n$, the following are equivalent:

(a) $\|\mathbf{A}\mathbf{x}-\mathbf{b}\|\le\|\mathbf{A}\mathbf{z}-\mathbf{b}\|$ for every $\mathbf{z}\in F^n$;

(b) $\mathbf{A}\mathbf{x}=P_U\mathbf{b}$;

(c) $\mathbf{A}^*\mathbf{A}\mathbf{x} = \mathbf{A}^*\mathbf{b}$.

Moreover such an $\mathbf{x}$ always exists. A vector satisfying these conditions is called a **least-squares solution** of $\mathbf{A}\mathbf{x} = \mathbf{b}$, and the equations in (c) are the **normal equations**.

Idea. Condition (a) is a minimization over $\mathbf{x}$; condition (b) is the same minimization rewritten in F^m, where Theorem 11.3.5 has already solved it. So (a) $\Leftrightarrow$ (b) is a translation, not an argument. The content is (b) $\Leftrightarrow$ (c): saying $\mathbf{A}\mathbf{x} = P_U\mathbf{b}$ is the same as saying the residual is perpendicular to U, and "perpendicular to every column of $\mathbf{A}$" is what $\mathbf{A}^*(\mathbf{b} - \mathbf{A}\mathbf{x}) = \mathbf{0}$ says. Existence is free, because $P_U\mathbf{b}$ lies in $U = \mathrm{col}(\mathbf{A})$ and so is $\mathbf{A}\mathbf{x}$ for some $\mathbf{x}$.

Proof. Since $\{\mathbf{A}\mathbf{z} : \mathbf{z} \in F^n\} = \mathrm{col}(\mathbf{A}) = U$ (Theorem 0.9.3), condition (a) says that $\mathbf{A}\mathbf{x}$ is a vector of U at least as close to $\mathbf{b}$ as every other vector of U. By Theorem 11.3.5 the unique such vector is $P_U\mathbf{b}$. Hence (a) $\Leftrightarrow$ (b).

(b) $\Leftrightarrow$ (c). Write $\mathbf{b} = P_U\mathbf{b} + (\mathbf{b} - P_U\mathbf{b})$, the splitting of Theorem 11.3.2 (a) with $\mathbf{b} - P_U\mathbf{b} \in U^\perp$. For $\mathbf{x} \in F^n$ we have $\mathbf{A}\mathbf{x} \in U$, so $\mathbf{b} - \mathbf{A}\mathbf{x} \in U^\perp$ holds if and only if $\mathbf{A}\mathbf{x}$ is the U-part of $\mathbf{b}$, that is, if and only if $\mathbf{A}\mathbf{x} = P_U\mathbf{b}$. It remains to see that

$$\mathbf{b} - \mathbf{A}\mathbf{x} \in U^\perp \iff \mathbf{A}^*\mathbf{A}\mathbf{x} = \mathbf{A}^*\mathbf{b}.$$

Put $\mathbf{r} = \mathbf{b} - \mathbf{A}\mathbf{x}$. By Proposition 11.3.1, $\mathbf{r} \in U^\perp$ if and only if $\mathbf{r}$ is orthogonal to every $\mathbf{A}\mathbf{z}$, and by Lemma 11.4.1,

$$\langle \mathbf{r}, \mathbf{A}\mathbf{z} \rangle = \langle \mathbf{A}^*\mathbf{r}, \mathbf{z} \rangle \qquad \text{for every } \mathbf{z} \in F^n.$$

If $\mathbf{A}^*\mathbf{r} = \mathbf{0}$, all these vanish. Conversely, if all vanish, take $\mathbf{z} = \mathbf{A}^*\mathbf{r}$ to get $\langle \mathbf{A}^*\mathbf{r}, \mathbf{A}^*\mathbf{r} \rangle = 0$, hence $\mathbf{A}^*\mathbf{r} = \mathbf{0}$ by positive definiteness. So $\mathbf{r} \in U^\perp$ if and only if $\mathbf{A}^*\mathbf{r} = \mathbf{0}$, which rearranges to $\mathbf{A}^*\mathbf{A}\mathbf{x} = \mathbf{A}^*\mathbf{b}$.

Existence. $P_U\mathbf{b} \in U = \mathrm{col}(\mathbf{A})$, so $P_U\mathbf{b} = \mathbf{A}\mathbf{x}$ for some $\mathbf{x} \in F^n$, and that $\mathbf{x}$ satisfies (b).

■

Three things are worth extracting. The normal equations are a **square** $n \times n$ system, whatever the shape of $\mathbf{A}$. They are **always consistent**, by the existence clause, even when $\mathbf{A}\mathbf{x} = \mathbf{b}$ is not. And the minimizing *value* $\mathbf{A}\mathbf{x}$ is unique — it is $P_U\mathbf{b}$ — even if the minimizing $\mathbf{x}$ is not.

Warning. **"Closest" is measured by the inner product, so the answer moves when the inner product does.** Least squares is not a coordinate-free notion: it minimizes $\|\mathbf{A}\mathbf{x} - \mathbf{b}\|$ for a particular norm on F^m, and rescaling one coordinate — reporting one measurement in centimeters instead of meters, say — is a change of inner product. Concretely, for the data $(0, 2), (1, 1), (2, 8)$ the least-squares line is $y = \frac{2}{3} + 3t$; if the middle point is trusted twice as much, so that the second coordinate of F^3 is weighted by 2, the best line becomes $y = 3t$. Both are correct answers to their own question, and Exercise C1 works the general case out. Before fitting, decide which errors count equally.

Quick check. The system $\mathbf{Ax} = \mathbf{b}$ may have no solution at all, yet the normal equations $\mathbf{A}^*\mathbf{Ax} = \mathbf{A}^*\mathbf{b}$ always have one. Where does the proof arrange that, and what stops the same argument from solving $\mathbf{Ax} = \mathbf{b}$ itself?

Solution. The existence clause uses that $P_U\mathbf{b}$ lies in $U = \operatorname{col}(\mathbf{A})$, and every vector of $\operatorname{col}(\mathbf{A})$ is $\mathbf{Ax}$ for some $\mathbf{x}$. So the normal equations ask $\mathbf{Ax}$ to hit $P_U\mathbf{b}$, a target that is inside the column space by construction. The original system asks $\mathbf{Ax}$ to hit $\mathbf{b}$, which need not be in $\operatorname{col}(\mathbf{A})$ at all. Projecting first is exactly the step that moves the target into reach; what is given up is the requirement $\mathbf{r} = \mathbf{0}$, which becomes $\mathbf{r} \perp \operatorname{col}(\mathbf{A})$.

11.4.3 When is the least-squares solution unique?

The minimizers of $\|\mathbf{Ax}-\mathbf{b}\|$ are the solutions of the normal equations, so they form the solution set of a linear system: a coset $\mathbf{x}_0 + \operatorname{nul}(\mathbf{A}^*\mathbf{A})$ (Theorem 3.1.3). To read off uniqueness we therefore need $\operatorname{nul}(\mathbf{A}^*\mathbf{A})$, and the following small lemma computes it.

Lemma 11.4.3. The null space of $\mathbf{A}^*\mathbf{A}$

Let $\mathbf{A} \in M_{m\times n}(F)$. Then $\operatorname{nul}(\mathbf{A}^*\mathbf{A}) = \operatorname{nul}(\mathbf{A})$.

Proof. ($\supseteq$) If $\mathbf{Ax} = \mathbf{0}$, then $\mathbf{A}^*\mathbf{Ax} = \mathbf{A}^*\mathbf{0} = \mathbf{0}$.

($\subseteq$) Suppose $\mathbf{A}^*\mathbf{Ax} = \mathbf{0}$. Pair with $\mathbf{x}$ and move $\mathbf{A}^*$ across using Lemma 11.4.1:

$$0 = \langle \mathbf{A}^*(\mathbf{Ax}), \mathbf{x}\rangle = \langle \mathbf{Ax}, \mathbf{Ax}\rangle = \|\mathbf{Ax}\|^2.$$

Hence $\mathbf{Ax} = \mathbf{0}$ by positive definiteness. This proves the lemma. ∎

This is the "pair it with itself" move once more, and it is the whole reason the normal equations behave well. It will return in Section 6 in the form $\ker(T^*T) = \ker T$ for a linear map T between inner product spaces, with the same two-line proof; the matrix statement above is the case we need now.

Corollary 11.4.4. Uniqueness of the least-squares solution

Let $\mathbf{A} \in M_{m\times n}(F)$ and $\mathbf{b} \in F^m$. The following are equivalent:

(a) $\mathbf{Ax} = \mathbf{b}$ has exactly one least-squares solution (for one, equivalently for every, $\mathbf{b}$);
(b) the columns of $\mathbf{A}$ are linearly independent, that is $\operatorname{rank}\mathbf{A} = n$;
(c) $\mathbf{A}^*\mathbf{A}$ is invertible.

In that case the least-squares solution is $\mathbf{x} = (\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^*\mathbf{b}$.

Proof. By Theorem 11.4.2 the least-squares solutions are the solutions of $\mathbf{A}^*\mathbf{A}\mathbf{x} = \mathbf{A}^*\mathbf{b}$, a consistent system; by Theorem 3.1.3 its solution set is a coset of $\operatorname{nul}(\mathbf{A}^*\mathbf{A})$, which is $\operatorname{nul}(\mathbf{A})$ by Lemma 11.4.3. So the solution is unique, for one $\mathbf{b}$ or for all, exactly when $\operatorname{nul}(\mathbf{A}) = \{\mathbf{0}\}$. Now $\operatorname{nul}(\mathbf{A}) = \{\mathbf{0}\}$ says precisely that the only vanishing linear combination of the columns of $\mathbf{A}$ is the trivial one, that is, that the columns are independent; and by Theorem 3.5.7 this is the same as $\operatorname{rank}\mathbf{A} = n$. This proves (a) $\Leftrightarrow$ (b).

(b) $\Leftrightarrow$ (c). The matrix $\mathbf{A}^*\mathbf{A}$ is $n \times n$, so it is invertible if and only if $\operatorname{nul}(\mathbf{A}^*\mathbf{A}) = \{\mathbf{0}\}$ (Theorem 3.4.4). By Lemma 11.4.3 this happens exactly when $\operatorname{nul}(\mathbf{A}) = \{\mathbf{0}\}$.

Finally, if $\mathbf{A}^*\mathbf{A}$ is invertible, multiplying the normal equations on the left by $(\mathbf{A}^*\mathbf{A})^{-1}$ gives the stated formula.

■

Warning. **$(\mathbf{A}^*\mathbf{A})^{-1}$ is not $\mathbf{A}^{-1}(\mathbf{A}^*)^{-1}$.** The rule $(\mathbf{BC})^{-1} = \mathbf{C}^{-1}\mathbf{B}^{-1}$ needs $\mathbf{B}$ and $\mathbf{C}$ to be invertible, and in least squares $\mathbf{A}$ is usually not even square. Take

$$\mathbf{A} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 0 & 1 \end{pmatrix}, \qquad \mathbf{A}^{\top}\mathbf{A} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}, \qquad (\mathbf{A}^{\top}\mathbf{A})^{-1} = \frac{1}{3}\begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}.$$

The left side exists; the symbols $\mathbf{A}^{-1}$ and $(\mathbf{A}^{\top})^{-1}$ do not exist at all, since $\mathbf{A}$ is 3×2. The formula $\mathbf{x} = (\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^*\mathbf{b}$ must be read as one object $(\mathbf{A}^*\mathbf{A})^{-1}$ applied to $\mathbf{A}^*\mathbf{b}$, never simplified.

11.4.4 Fitting a line and a parabola

Example. The least-squares line through four points

Find the line $y = c + dt$ that minimizes the sum of the squared vertical errors at the four data points

$$(1,3), \quad (2,1), \quad (3,5), \quad (4,5).$$

Compute the residual and the minimum value of the sum of squares.

Solution. Asking for $c + dt_i = y_i$ at the four points is the system $\mathbf{A}\mathbf{x} = \mathbf{b}$ with

$$\mathbf{A} = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{pmatrix}, \qquad \mathbf{x} = \begin{pmatrix} c \\ d \end{pmatrix}, \qquad \mathbf{b} = \begin{pmatrix} 3 \\ 1 \\ 5 \\ 5 \end{pmatrix},$$

and $\|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2$ is exactly the sum of the squared vertical errors. The columns of $\mathbf{A}$ are independent (they are not multiples of each other), so Corollary 11.4.4 gives a unique answer.

Over $\mathbb{R}$, $\mathbf{A}^* = \mathbf{A}^\top$, and

$$\mathbf{A}^\top\mathbf{A} = \begin{pmatrix} 4 & 10 \\ 10 & 30 \end{pmatrix}, \qquad \mathbf{A}^\top\mathbf{b} = \begin{pmatrix} 3+1+5+5 \\ 3+2+15+20 \end{pmatrix} = \begin{pmatrix} 14 \\ 40 \end{pmatrix}.$$

The normal equations $4c + 10d = 14$, $10c + 30d = 40$ simplify to $2c + 5d = 7$ and $c + 3d = 4$; subtracting twice the second from the first gives $-d = -1$, so $d = 1$ and $c = 1$. The least-squares line is

$$y = 1 + t.$$

Its values at $t = 1, 2, 3, 4$ are $2, 3, 4, 5$, so the residual is

$$\mathbf{r} = \mathbf{b} - \mathbf{A}\mathbf{x} = (3, 1, 5, 5) - (2, 3, 4, 5) = (1, -2, 1, 0),$$

and the minimum sum of squares is $\|\mathbf{r}\|^2 = 1 + 4 + 1 + 0 = 6$.

Check. The residual must be orthogonal to both columns of $\mathbf{A}$: $1 - 2 + 1 + 0 = 0$ and $1 - 4 + 3 + 0 = 0$. Both hold, which confirms the normal equations.

The same machinery fits any linear combination of prescribed functions; only the columns of $\mathbf{A}$ change. "Linear" refers to the unknown coefficients, not to the shape of the curve.

Example. The least-squares parabola

Find the quadratic $y = a + bt + ct^2$ minimizing the sum of squared errors at

$$(-2, 7), \quad (-1, 6), \quad (0, 2), \quad (1, 0), \quad (2, 5).$$

Solution. Now $\mathbf{A}$ has columns $(1, 1, 1, 1, 1)$, (t_i) and (t_i^2):

$$\mathbf{A} = \begin{pmatrix} 1 & -2 & 4 \\ 1 & -1 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \end{pmatrix}, \qquad \mathbf{b} = \begin{pmatrix} 7 \\ 6 \\ 2 \\ 0 \\ 5 \end{pmatrix}.$$

Because the nodes are symmetric about 0, every odd power sums to zero: $\sum t_i = 0$ and $\sum t_i^3 = 0$. With $\sum t_i^2 = 10$ and $\sum t_i^4 = 34$,

$$\mathbf{A}^\top\mathbf{A} = \begin{pmatrix} 5 & 0 & 10 \\ 0 & 10 & 0 \\ 10 & 0 & 34 \end{pmatrix}, \qquad \mathbf{A}^\top\mathbf{b} = \begin{pmatrix} 20 \\ -10 \\ 54 \end{pmatrix}.$$

The middle equation decouples: $10b = -10$, so $b = -1$. The outer two are $5a + 10c = 20$ and $10a + 34c = 54$; the first gives $a = 4 - 2c$, and substituting gives $40 + 14c = 54$, so $c = 1$ and $a = 2$. The least-squares parabola is

$$y = 2 - t + t^2.$$

Its values at the nodes are $8, 4, 2, 2, 4$, so $\mathbf{r} = (-1, 2, 0, -2, 1)$ and the minimum sum of squares is $\|\mathbf{r}\|^2 = 1 + 4 + 0 + 4 + 1 = 10$.

Check. $\sum r_i = 0$, $\sum t_i r_i = 2 - 2 + 0 - 2 + 2 = 0$, and $\sum t_i^2 r_i = -4 + 2 + 0 - 2 + 4 = 0$: the residual is orthogonal to all three columns.

Two practical remarks. Choosing the nodes symmetrically, as above, made $\mathbf{A}^{\top}\mathbf{A}$ nearly diagonal and the arithmetic painless; the same trick appears again in Section 9 when the columns are trigonometric functions. And although the normal equations are the right way to *understand* least squares, they are not always the right way to *compute* it: Section 8 solves the same problem through a QR factorization, which avoids forming $\mathbf{A}^{*}\mathbf{A}$ at all.

11.4.5 The minimum-norm solution

Now the opposite situation: $\mathbf{Ax} = \mathbf{b}$ is consistent, but $\mathbf{A}$ has a nontrivial null space, so the solutions form an infinite family $\mathbf{x}_0 + \operatorname{nul}(\mathbf{A})$ (Theorem 3.1.3). Which one should we report? Orthogonality answers.

Theorem 11.4.5. Minimum-Norm Solution

Let $\mathbf{A} \in M_{m\times n}(F)$ and let $\mathbf{b} \in F^m$ be such that $\mathbf{Ax} = \mathbf{b}$ is consistent. Write $K = \operatorname{nul}(\mathbf{A})$. Then:

(a) exactly one solution $\mathbf{x}_{\min}$ of $\mathbf{Ax} = \mathbf{b}$ lies in $K^{\perp}$;

(b) $\|\mathbf{x}_{\min}\| < \|\mathbf{x}\|$ for every other solution $\mathbf{x}$.

We call $\mathbf{x}_{\min}$ the **minimum-norm solution**.

Idea. The solution set is a coset $\mathbf{x}_0 + K$, and $F^n = K \oplus K^{\perp}$. Throwing away the K-part of $\mathbf{x}_0$ keeps it a solution, because that part is in the kernel, and lands it in $K^{\perp}$. Once there, Pythagoras compares it with every other solution: the others differ from it by a kernel vector, which is perpendicular to it and only adds length.

Proof.

(a) *Existence.* Let $\mathbf{x}_0$ be any solution. Since K is a subspace of the finite-dimensional space F^n, Theorem 11.3.2 (a) gives $F^n = K \oplus K^{\perp}$; write $\mathbf{x}_0 = \mathbf{k} + \mathbf{x}_{\min}$ with $\mathbf{k} \in K$ and $\mathbf{x}_{\min} \in K^{\perp}$. Then $\mathbf{Ax}_{\min} = \mathbf{Ax}_0 - \mathbf{Ak} = \mathbf{b} - \mathbf{0} = \mathbf{b}$, so $\mathbf{x}_{\min}$ is a solution lying in $K^{\perp}$.
Uniqueness. If $\mathbf{x}_{\min}$ and $\mathbf{y}$ are solutions in $K^{\perp}$, then $\mathbf{A}(\mathbf{x}_{\min} - \mathbf{y}) = \mathbf{0}$, so $\mathbf{x}_{\min} - \mathbf{y} \in K \cap K^{\perp} = \{\mathbf{0}\}$ by Theorem 11.3.2 (a).

(b) Every solution has the form $\mathbf{x} = \mathbf{x}_{\min} + \mathbf{k}$ with $\mathbf{k} \in K$, by Theorem 3.1.3. Since $\mathbf{x}_{\min} \in K^{\perp}$ and $\mathbf{k} \in K$, the two are orthogonal, and Theorem 11.1.3 gives

$$\|\mathbf{x}\|^2 = \|\mathbf{x}_{\min}\|^2 + \|\mathbf{k}\|^2.$$

If $\mathbf{x} \neq \mathbf{x}_{\min}$, then $\mathbf{k} \neq \mathbf{0}$, so $\|\mathbf{k}\|^2 > 0$ and $\|\mathbf{x}\| > \|\mathbf{x}_{\min}\|$. This proves the theorem.

∎

The subspace $K^\perp$ has a matrix description that makes $\mathbf{x}_{\min}$ computable. For $\mathbf{x} \in F^n$,

$$\begin{aligned} \mathbf{x} \in \operatorname{col}(\mathbf{A}^*)^\perp &\iff \langle \mathbf{x}, \mathbf{A}^*\mathbf{y} \rangle = 0 \quad \text{for all } \mathbf{y} \in F^m \\ &\iff \langle \mathbf{A}\mathbf{x}, \mathbf{y} \rangle = 0 \quad \text{for all } \mathbf{y} \in F^m \\ &\iff \mathbf{A}\mathbf{x} = \mathbf{0}, \end{aligned}$$

where the first step is Definition 11.3.1 together with $\operatorname{col}(\mathbf{A}^*) = \{\mathbf{A}^*\mathbf{y} : \mathbf{y} \in F^m\}$ (Theorem 0.9.3), the second is Lemma 11.4.1, and the last uses the choice $\mathbf{y} = \mathbf{A}\mathbf{x}$ with positive definiteness. So $\operatorname{nul}(\mathbf{A}) = \operatorname{col}(\mathbf{A}^*)^\perp$, and applying $\perp$ once more (Theorem 11.3.2 (b)) gives

$$\big(\operatorname{nul}(\mathbf{A})\big)^\perp = \operatorname{col}(\mathbf{A}^*).$$

In words: the minimum-norm solution is the unique solution that is a combination of the columns of $\mathbf{A}^*$, that is, of the conjugated rows of $\mathbf{A}$. Section 6 proves the same statement for linear maps, together with the three companion identities that make up the four fundamental subspaces.

Example. A minimum-norm solution

Find the solution of smallest norm of

$$\begin{cases} x_1 + x_2 + x_3 = 3, \\ x_1 + 2x_2 + 3x_3 = 7. \end{cases}$$

Solution. Here $\mathbf{A} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \end{pmatrix}$. Subtracting the first equation from the second gives $x_2 + 2x_3 = 4$, so $x_2 = 4 - 2x_3$ and then $x_1 = 3 - x_2 - x_3 = x_3 - 1$. With $t = x_3$ the solutions are

$$\mathbf{x}(t) = (-1, 4, 0) + t(1, -2, 1), \qquad t \in \mathbb{R},$$

so $K = \operatorname{nul}(\mathbf{A}) = \operatorname{span}((1, -2, 1))$. By Theorem 11.4.5 the answer is the unique t with $\mathbf{x}(t) \perp (1, -2, 1)$:

$$\langle \mathbf{x}(t), (1, -2, 1) \rangle = (-1 + t) - 2(4 - 2t) + t = 6t - 9 = 0, \qquad t = \tfrac{3}{2}.$$

Hence $\mathbf{x}_{\min} = (\tfrac{1}{2}, 1, \tfrac{3}{2})$, with $\|\mathbf{x}_{\min}\|^2 = \tfrac{1}{4} + 1 + \tfrac{9}{4} = \tfrac{7}{2}$.

Check. It is a solution, since $\tfrac{1}{2} + 1 + \tfrac{3}{2} = 3$ and $\tfrac{1}{2} + 2 + \tfrac{9}{2} = 7$, and it is orthogonal to the kernel, since $\tfrac{1}{2} - 2 + \tfrac{3}{2} = 0$. By the identity $(\operatorname{nul}\mathbf{A})^\perp = \operatorname{col}(\mathbf{A}^\top)$ above it must also be a combination of the rows of $\mathbf{A}$, and it is: $\mathbf{x}_{\min} = \tfrac{1}{2}(1, 2, 3)$, half the second row. Finally, the solution $\mathbf{x}(0) = (-1, 4, 0)$ has $\|\mathbf{x}(0)\|^2 = 17 > \tfrac{7}{2}$, as Theorem 11.4.5 (b) requires.

The two answers of this section — the least-squares solution and the minimum-norm solution — are usually wanted together: given any $\mathbf{A}$ and $\mathbf{b}$, first minimize $\|\mathbf{A}\mathbf{x} - \mathbf{b}\|$, then among the minimizers take the shortest. Chapter 13 packages the combined answer as a single matrix, the **pseudoinverse** $\mathbf{A}^+$, with $\mathbf{x} = \mathbf{A}^+\mathbf{b}$ in all cases at once.

11.4.6 Exercises

11.4.6.1 A. Check your understanding

Exercise. A1

(a) Define a least-squares solution of $\mathbf{Ax} = \mathbf{b}$, and write down the normal equations.

(b) True or false: if $\mathbf{Ax} = \mathbf{b}$ is consistent, then its solutions are exactly its least-squares solutions. Justify your answer.

(c) Explain in one sentence why the normal equations are always consistent.

(d) Give a condition on the columns of $\mathbf{A}$ equivalent to uniqueness of the least-squares solution.

(e) Where does the minimum-norm solution of a consistent system live, and why is it unique there?

(f) True or false: $(\mathbf{A}^*\mathbf{A})^{-1} = \mathbf{A}^{-1}(\mathbf{A}^*)^{-1}$. Justify your answer.

Solution.

(a) A vector $\mathbf{x} \in F^n$ with $\|\mathbf{Ax} - \mathbf{b}\| \le \|\mathbf{Az} - \mathbf{b}\|$ for every $\mathbf{z} \in F^n$. The normal equations are $\mathbf{A}^*\mathbf{Ax} = \mathbf{A}^*\mathbf{b}$ (Theorem 11.4.2).

(b) True. If $\mathbf{Ax}_0 = \mathbf{b}$ then $\|\mathbf{Ax}_0 - \mathbf{b}\| = 0$, which is the smallest possible value, so $\mathbf{x}_0$ is a least-squares solution; conversely a least-squares solution $\mathbf{x}$ has $\|\mathbf{Ax} - \mathbf{b}\| \le 0$, hence $\mathbf{Ax} = \mathbf{b}$. Least squares extends the notion of a solution without changing it when solutions exist.

(c) Because the target $P_{\mathrm{col}(\mathbf{A})}\mathbf{b}$ lies in $\mathrm{col}(\mathbf{A})$ by construction, so $\mathbf{Ax} = P_{\mathrm{col}(\mathbf{A})}\mathbf{b}$ is solvable (Theorem 11.4.2, existence).

(d) The columns are linearly independent, equivalently $\operatorname{rank}\mathbf{A} = n$, equivalently $\mathbf{A}^*\mathbf{A}$ is invertible (Corollary 11.4.4).

(e) In $(\operatorname{nul}\mathbf{A})^{\perp} = \mathrm{col}(\mathbf{A}^*)$. Two solutions there would differ by an element of $\mathrm{nul}(\mathbf{A}) \cap (\operatorname{nul}\mathbf{A})^{\perp} = \{\mathbf{0}\}$ (Theorem 11.4.5).

(f) False, and usually meaningless: $\mathbf{A}$ need not be square, so $\mathbf{A}^{-1}$ need not exist, while $(\mathbf{A}^*\mathbf{A})^{-1}$ can exist perfectly well — see the 3×2 example in the warning after Corollary 11.4.4.

11.4.6.2 B. Practice

Exercise. B1: A least-squares line

Find the least-squares line $y = c + dt$ for the data $(0, 3), (1, 2), (2, 3), (3, 6)$. Compute the residual, verify that it is orthogonal to both columns of $\mathbf{A}$, and give the minimum sum of squares.

Solution. With $\mathbf{A}$ having columns $(1, 1, 1, 1)$ and $(0, 1, 2, 3)$ and $\mathbf{b} = (3, 2, 3, 6)$,

$$\mathbf{A}^{\top}\mathbf{A} = \begin{pmatrix} 4 & 6 \\ 6 & 14 \end{pmatrix}, \qquad \mathbf{A}^{\top}\mathbf{b} = \begin{pmatrix} 14 \\ 0 + 2 + 6 + 18 \end{pmatrix} = \begin{pmatrix} 14 \\ 26 \end{pmatrix}.$$

The normal equations are $2c + 3d = 7$ and $3c + 7d = 13$. Multiplying the first by 3 and the second by 2 and subtracting gives $9d - 14d = 21 - 26$, so $d = 1$, and then $c = 2$. The line is $y = 2 + t$.

Its values at $t = 0, 1, 2, 3$ are $2, 3, 4, 5$, so $\mathbf{r} = (1, -1, -1, 1)$. Orthogonality: $1 - 1 - 1 + 1 = 0$ against the first column and $0 - 1 - 2 + 3 = 0$ against the second. The minimum sum of squares is $\|\mathbf{r}\|^2 = 4$.

Exercise. B2: A least-squares parabola

Find the least-squares quadratic $y = a + bt + ct^2$ for the data $(-1, 2), (0, 8), (1, 6), (2, 16)$, and the minimum sum of squares. *Hint: write down* $\mathbf{A}^\top\mathbf{A}$ *from the power sums* $\sum t_i^k$, $k = 0, \dots, 4$.

Solution. The power sums are $\sum 1 = 4$, $\sum t_i = 2$, $\sum t_i^2 = 6$, $\sum t_i^3 = 8$, $\sum t_i^4 = 18$, so

$$\mathbf{A}^\top\mathbf{A} = \begin{pmatrix} 4 & 2 & 6 \\ 2 & 6 & 8 \\ 6 & 8 & 18 \end{pmatrix}, \qquad \mathbf{A}^\top\mathbf{b} = \begin{pmatrix} 2+8+6+16 \\ -2+0+6+32 \\ 2+0+6+64 \end{pmatrix} = \begin{pmatrix} 32 \\ 36 \\ 72 \end{pmatrix}.$$

Solving: from the first two equations, $2a + b + 3c = 16$ and $a + 3b + 4c = 18$; eliminating a gives $5b + 5c = 20$, so $b + c = 4$. The third equation is $3a + 4b + 9c = 36$. Substituting $b = 4 - c$ into $2a + b + 3c = 16$ gives $2a + 2c = 12$, so $a = 6 - c$; then $3(6 - c) + 4(4 - c) + 9c = 36$ becomes $34 + 2c = 36$, so $c = 1$, $b = 3$, $a = 5$. The parabola is

$$y = 5 + 3t + t^2.$$

Its values at $t = -1, 0, 1, 2$ are $3, 5, 9, 15$, so $\mathbf{r} = (-1, 3, -3, 1)$ and the minimum sum of squares is $1 + 9 + 9 + 1 = 20$. *Check:* $\sum r_i = 0$, $\sum t_i r_i = 1 + 0 - 3 + 2 = 0$, $\sum t_i^2 r_i = -1 + 0 - 3 + 4 = 0$.

Exercise. B3: A minimum-norm solution

Find the minimum-norm solution of

$$\begin{cases} x_1 + x_3 = 2, \\ x_2 + x_3 = 3, \end{cases}$$

and verify that it is orthogonal to $\mathrm{nul}(\mathbf{A})$.

Solution. With $x_3 = t$ the solutions are $\mathbf{x}(t) = (2 - t,\ 3 - t,\ t) = (2, 3, 0) + t(-1, -1, 1)$, so $\mathrm{nul}(\mathbf{A}) = \mathrm{span}((-1, -1, 1))$. By Theorem 11.4.5 we need $\langle \mathbf{x}(t), (-1, -1, 1) \rangle = 0$:

$$-(2 - t) - (3 - t) + t = 3t - 5 = 0, \qquad t = \tfrac{5}{3}.$$

So $\mathbf{x}_{\min} = (\tfrac{1}{3}, \tfrac{4}{3}, \tfrac{5}{3})$, with $\|\mathbf{x}_{\min}\|^2 = \frac{1+16+25}{9} = \frac{14}{3}$. It solves the system, since $\frac{1}{3} + \frac{5}{3} = 2$ and $\frac{4}{3} + \frac{5}{3} = 3$, and $-\frac{1}{3} - \frac{4}{3} + \frac{5}{3} = 0$ confirms the orthogonality. For comparison, $\|(2, 3, 0)\|^2 = 13 > \frac{14}{3}$.

11.4.6.3 C. Going deeper

Exercise. C1: Weighted least squares

Let $\mathbf{A} \in M_{m\times n}(\mathbb{R})$, $\mathbf{b} \in \mathbb{R}^m$, and let $w_1, \dots, w_m > 0$ be weights, with $\mathbf{W} = \operatorname{diag}(w_1, \dots, w_m)$.

(a) Show that $\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{W}} = \mathbf{y}^{\top}\mathbf{W}\mathbf{x}$ is an inner product on $\mathbb{R}^m$.

(b) Prove that the minimizers of $\sum_{i=1}^{m} w_i((\mathbf{A}\mathbf{x})_i - b_i)^2$ are exactly the solutions of $\mathbf{A}^{\top}\mathbf{W}\mathbf{A}\mathbf{x} = \mathbf{A}^{\top}\mathbf{W}\mathbf{b}$.

(c) For the data $(0,2), (1,1), (2,8)$ with weights $(1,2,1)$, find the weighted least-squares line, and compare it with the unweighted one.

Solution.

(a) It is bilinear and symmetric because $\mathbf{y}^{\top}\mathbf{W}\mathbf{x} = \sum_i w_i x_i y_i$, and $\langle \mathbf{x}, \mathbf{x} \rangle_{\mathbf{W}} = \sum_i w_i x_i^2 > 0$ whenever some $x_i \neq 0$, because every $w_i > 0$. (This is where positivity of the weights is used; a zero weight would destroy definiteness.)

(b) Write $\|\cdot\|_{\mathbf{W}}$ for the norm induced by $\langle \cdot, \cdot \rangle_{\mathbf{W}}$. The quantity to minimize is exactly $\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_{\mathbf{W}}^2$. The proof of Theorem 11.4.2 used nothing about the inner product except its axioms, so it applies verbatim with $\langle \cdot, \cdot \rangle_{\mathbf{W}}$ in place of the standard one: the minimizers are the $\mathbf{x}$ for which $\mathbf{A}\mathbf{x} - \mathbf{b}$ is $\langle \cdot, \cdot \rangle_{\mathbf{W}}$-orthogonal to $\operatorname{col}(\mathbf{A})$, that is, for which $\langle \mathbf{A}\mathbf{x} - \mathbf{b}, \mathbf{A}\mathbf{z} \rangle_{\mathbf{W}} = 0$ for all $\mathbf{z} \in \mathbb{R}^n$. Now

$$\langle \mathbf{A}\mathbf{x} - \mathbf{b}, \mathbf{A}\mathbf{z} \rangle_{\mathbf{W}} = (\mathbf{A}\mathbf{z})^{\top}\mathbf{W}(\mathbf{A}\mathbf{x} - \mathbf{b}) = \mathbf{z}^{\top}(\mathbf{A}^{\top}\mathbf{W}\mathbf{A}\mathbf{x} - \mathbf{A}^{\top}\mathbf{W}\mathbf{b}),$$

and this vanishes for all $\mathbf{z} \in \mathbb{R}^n$ exactly when $\mathbf{A}^{\top}\mathbf{W}\mathbf{A}\mathbf{x} - \mathbf{A}^{\top}\mathbf{W}\mathbf{b} = \mathbf{0}$, by taking $\mathbf{z}$ equal to that vector.

(c) With $\mathbf{A}$ having columns $(1,1,1)$ and $(0,1,2)$, $\mathbf{b} = (2,1,8)$ and $\mathbf{W} = \operatorname{diag}(1,2,1)$,

$$\mathbf{A}^{\top}\mathbf{W}\mathbf{A} = \begin{pmatrix} 4 & 4 \\ 4 & 6 \end{pmatrix}, \qquad \mathbf{A}^{\top}\mathbf{W}\mathbf{b} = \begin{pmatrix} 2+2+8 \\ 0+2+16 \end{pmatrix} = \begin{pmatrix} 12 \\ 18 \end{pmatrix},$$

so $4c + 4d = 12$ and $4c + 6d = 18$, giving $d = 3$ and $c = 0$: the weighted line is $y = 3t$. Unweighted, $\mathbf{A}^{\top}\mathbf{A} = \begin{pmatrix} 3 & 3 \\ 3 & 5 \end{pmatrix}$ and $\mathbf{A}^{\top}\mathbf{b} = (11, 17)$, giving $3c + 3d = 11$, $3c + 5d = 17$, so $d = 3$ and $c = \frac{2}{3}$: the unweighted line is $y = \frac{2}{3} + 3t$. The middle point is trusted twice as much in the weighted fit, and the line moves down to pass closer to it. The weighted fit has weighted error $4 + 2 \cdot 4 + 4 = 16$, smaller than the weighted error $\frac{160}{9}$ of the unweighted line, while the unweighted fit has plain error $\frac{32}{3}$, smaller than the plain error 12 of the weighted line: each line wins the contest it was designed for.

Exercise. C2: The residual, two ways

Let $\mathbf{x}$ be a least-squares solution of $\mathbf{A}\mathbf{x} = \mathbf{b}$ and let $\mathbf{r} = \mathbf{b} - \mathbf{A}\mathbf{x}$.

(a) Prove that $\|\mathbf{r}\|^2 = \|\mathbf{b}\|^2 - \|\mathbf{A}\mathbf{x}\|^2$, and also that $\|\mathbf{r}\|^2 = \langle \mathbf{b}, \mathbf{r} \rangle = \|\mathbf{b}\|^2 - \langle \mathbf{A}^{*}\mathbf{b}, \mathbf{x} \rangle$.

(b) Use the second formula to recompute the minimum sum of squares in Example (The least-squares line through four points) without finding the residual vector.

Solution.

(a) By Theorem 11.4.2, $\mathbf{Ax} = P_U\mathbf{b}$ with $U = \operatorname{col}(\mathbf{A})$, so $\mathbf{r} = \mathbf{b} - P_U\mathbf{b} \in U^{\perp}$ while $\mathbf{Ax} \in U$. The two are orthogonal, and $\mathbf{b} = \mathbf{Ax}+\mathbf{r}$, so Theorem 11.1.3 gives $\|\mathbf{b}\|^2 = \|\mathbf{Ax}\|^2+\|\mathbf{r}\|^2$, which is the first identity.
For the second, $\langle \mathbf{b}, \mathbf{r}\rangle = \langle \mathbf{Ax}+\mathbf{r}, \mathbf{r}\rangle = \langle \mathbf{Ax}, \mathbf{r}\rangle + \|\mathbf{r}\|^2 = \|\mathbf{r}\|^2$, since $\langle \mathbf{Ax}, \mathbf{r}\rangle = \overline{\langle \mathbf{r}, \mathbf{Ax}\rangle} = 0$ by orthogonality. Expanding the other way, and using Lemma 11.4.1,

$$\langle \mathbf{b}, \mathbf{r}\rangle = \langle \mathbf{b}, \mathbf{b} - \mathbf{Ax}\rangle = \|\mathbf{b}\|^2 - \langle \mathbf{b}, \mathbf{Ax}\rangle = \|\mathbf{b}\|^2 - \langle \mathbf{A}^*\mathbf{b}, \mathbf{x}\rangle.$$

(b) In Example (The least-squares line through four points), $\|\mathbf{b}\|^2 = 9+1+25+25 = 60$, $\mathbf{A}^{\top}\mathbf{b} = (14, 40)$ and $\mathbf{x} = (1, 1)$, so

$$\|\mathbf{r}\|^2 = 60 - \langle (14, 40), (1, 1)\rangle = 60 - 54 = 6,$$

matching the value found there. Only the data already assembled for the normal equations is needed.

Exercise. C3: A rank-deficient least-squares problem

Let

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 2 \end{pmatrix}, \qquad \mathbf{b} = \begin{pmatrix} 1 \\ 2 \\ 0 \end{pmatrix}.$$

(a) Show that $\mathbf{Ax} = \mathbf{b}$ is inconsistent and that $\operatorname{rank}\mathbf{A} = 2$.

(b) Describe **all** least-squares solutions.

(c) Find the one of smallest norm, and the minimum sum of squares.

Solution.

(a) The third column is the sum of the first two, so the columns are dependent and $\operatorname{rank}\mathbf{A} \le 2$; the first two columns are independent, so $\operatorname{rank}\mathbf{A} = 2$ and $\operatorname{col}(\mathbf{A}) = \operatorname{span}\big((1, 0, 1), (0, 1, 1)\big)$. A vector (b_1, b_2, b_3) lies in $\operatorname{col}(\mathbf{A})$ exactly when $b_3 = b_1 + b_2$; here $0 \neq 1 + 2$, so $\mathbf{b} \notin \operatorname{col}(\mathbf{A})$ and the system is inconsistent.

(b) By Theorem 11.4.2 the least-squares solutions solve $\mathbf{A}^{\top}\mathbf{Ax} = \mathbf{A}^{\top}\mathbf{b}$. Here

$$\mathbf{A}^{\top}\mathbf{A} = \begin{pmatrix} 2 & 1 & 3 \\ 1 & 2 & 3 \\ 3 & 3 & 6 \end{pmatrix}, \qquad \mathbf{A}^{\top}\mathbf{b} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}.$$

Row reducing, the third row is the sum of the first two, and subtracting suitable multiples gives $x_1 + x_3 = 0$ and $x_2 + x_3 = 1$. Hence

$$\mathbf{x}(t) = (0, 1, 0) + t(-1, -1, 1), \qquad t \in \mathbb{R},$$

a coset of $\operatorname{nul}(\mathbf{A}) = \operatorname{span}\big((-1, -1, 1)\big) = \operatorname{nul}(\mathbf{A}^{\top}\mathbf{A})$, as Lemma 11.4.3 predicts.

(c) The minimizers form a coset of $\operatorname{nul}(\mathbf{A})$, so the argument of Theorem 11.4.5 applies to it verbatim: the shortest is the one orthogonal to $(-1, -1, 1)$. From $-(-t) - (1-t) + t =$

$3t - 1 = 0$ we get $t = \frac{1}{3}$ and

$$\mathbf{x}_{\min} = \left(-\tfrac{1}{3}, \tfrac{2}{3}, \tfrac{1}{3}\right), \qquad \|\mathbf{x}_{\min}\|^2 = \tfrac{1+4+1}{9} = \tfrac{2}{3}.$$

Every minimizer gives the same $\mathbf{Ax} = (0, 1, 1) = P_{\mathrm{col}(\mathbf{A})}\mathbf{b}$, with residual $\mathbf{r} = (1, 1, -1)$ and minimum sum of squares $\|\mathbf{r}\|^2 = 3$. *Check:* $\mathbf{r}$ is orthogonal to $(1, 0, 1)$ and to $(0, 1, 1)$, since $1 - 1 = 0$ and $1 - 1 = 0$.

11.5 The Riesz Representation Theorem

Chapter 5 built the dual space V^* of all linear functionals on V, proved that $\dim V^* = \dim V$ in finite dimension, and then delivered a warning: the resulting isomorphism $V \cong V^*$ is not canonical. Every basis of V gives one, different bases give different ones, and no argument singles out a preferred choice (Corollary 5.1.4, and the contrast with the natural map $\mathrm{ev}_V : V \to V^{**}$ of Theorem 5.4.6). The chapter closed by promising that an inner product would supply the missing ingredient. This section keeps the promise.

The mechanism is simple to state. Fix $w \in V$ and let $\varphi_w(v) = \langle v, w\rangle$. Because the inner product is linear in the **first** slot, φ_w is a linear functional. The Riesz Representation Theorem says that in finite dimension there are no others.

Throughout, V is a finite-dimensional inner product space over $F = \mathbb{R}$ or $F = \mathbb{C}$, unless the text says otherwise.

11.5.1 Every functional is an inner product

In F^n, Theorem 5.1.1 already told us that every functional has the form $\mathbf{x} \mapsto a_1x_1 + \cdots + a_nx_n$ for a unique row of coefficients. With the standard inner product $\langle \mathbf{x}, \mathbf{y}\rangle = \sum_i x_i\overline{y_i}$, that reads $\varphi(\mathbf{x}) = \langle \mathbf{x}, \mathbf{w}\rangle$ with $\mathbf{w} = (\overline{a_1}, \ldots, \overline{a_n})$. So in coordinates the theorem is a restatement of a fact we already have; the content is that it survives the removal of coordinates.

Theorem 11.5.1. Riesz Representation Theorem

Let V be a finite-dimensional inner product space over F and let $\varphi \in V^*$. Then there is a **unique** $w \in V$ such that

$$\varphi(v) = \langle v, w\rangle \qquad \text{for every } v \in V.$$

If $(e_1, \ldots, e_n)$ is any orthonormal basis of V, then $w = \sum_{i=1}^{n} \overline{\varphi(e_i)}\, e_i$.

Idea. Fix an orthonormal basis. Then a linear functional is determined by the n numbers $\varphi(e_i)$, and a vector is determined by its n coordinates, which Theorem 11.2.1 reads off as inner products. Matching the two lists tells us what w must be, and the only care needed is the conjugate: w sits in the second slot, so its coordinates come out conjugated. Uniqueness is the standard move — if two vectors pair identically with everything, subtract and pair the difference with itself.

Proof. *Existence.* Let $(e_1, \dots, e_n)$ be an orthonormal basis of V, which exists by Theorem 11.2.2, and put

$$w = \sum_{i=1}^{n} \overline{\varphi(e_i)}\, e_i.$$

For each j, conjugate-linearity in the second slot and $\langle e_j, e_i \rangle = \delta_{ij}$ give

$$\langle e_j, w \rangle = \sum_{i=1}^{n} \overline{\overline{\varphi(e_i)}}\, \langle e_j, e_i \rangle = \varphi(e_j).$$

So the two linear functionals $v \mapsto \langle v, w \rangle$ and φ agree on a basis, hence are equal by Theorem 2.1.5. (The first is linear because the inner product is linear in the first slot.)

Uniqueness. Suppose $\langle v, w \rangle = \langle v, w' \rangle$ for every v. Then $\langle v, w - w' \rangle = 0$ for every v, by additivity in the second slot. Taking $v = w - w'$ gives $\|w - w'\|^2 = 0$, so $w = w'$ by positive definiteness.

■

Call w the **Riesz vector** of φ. The formula is worth reading twice: the coordinates of w in an orthonormal basis are the conjugates of the values of φ on that basis. Over $\mathbb{R}$ the conjugation is invisible, which is why real textbooks can omit it and complex ones cannot.

Example. Two Riesz vectors

(a) On $V = \mathbb{R}[x]_{\le 1}$ with $\langle p, q \rangle = \int_0^1 p(t)q(t)\,\mathrm{d}t$, find the Riesz vector of $\varphi(p) = p(0)$.

(b) On $V = M_n(\mathbb{C})$ with the Frobenius inner product $\langle \mathbf{A}, \mathbf{B} \rangle = \operatorname{tr}(\mathbf{B}^*\mathbf{A})$, find the Riesz vector of $\varphi(\mathbf{A}) = \operatorname{tr}\mathbf{A}$.

Solution.

(a) We look for $w = a + bx$ with $\langle p, w \rangle = p(0)$ for all p. By linearity it is enough to impose this on the basis $(1, x)$:

$$\int_0^1 (a + bt)\,\mathrm{d}t = a + \tfrac{b}{2} = 1, \qquad \int_0^1 t(a + bt)\,\mathrm{d}t = \tfrac{a}{2} + \tfrac{b}{3} = 0.$$

The second equation gives $a = -\tfrac{2}{3}b$; substituting into the first gives $-\tfrac{2}{3}b + \tfrac{1}{2}b = -\tfrac{1}{6}b = 1$, so $b = -6$ and $a = 4$. Hence $w = 4 - 6x$.
Check. $\int_0^1 (4 - 6t)\,\mathrm{d}t = 4 - 3 = 1$ and $\int_0^1 t(4 - 6t)\,\mathrm{d}t = 2 - 2 = 0$, as required; and for a general $p = c + dx$ the pairing gives $c \cdot 1 + d \cdot 0 = c = p(0)$. Note how little w resembles "evaluation at 0": the Riesz vector is manufactured by the inner product, not by the point 0.

(b) We look for $\mathbf{W}$ with $\operatorname{tr}(\mathbf{W}^*\mathbf{A}) = \operatorname{tr}\mathbf{A}$ for every $\mathbf{A}$. Taking $\mathbf{W} = \mathbf{I}_n$ works, since $\mathbf{I}_n^* = \mathbf{I}_n$ and $\operatorname{tr}(\mathbf{I}_n\mathbf{A}) = \operatorname{tr}\mathbf{A}$. By uniqueness, $\mathbf{W} = \mathbf{I}_n$ is the answer. (This example was not solved by the basis formula, but by guessing and then invoking uniqueness — a legitimate and often faster route.)

11.5.2 The Riesz map, and the data Chapter 5 was missing

Uniqueness in Theorem 11.5.1 lets us read the theorem as a statement about a map.

Corollary 11.5.2. The Riesz map

Let V be a finite-dimensional inner product space over F. Define

$$\Phi\colon V \to V^*, \qquad \Phi(w) = \langle \cdot, w\rangle.$$

Then Φ is a bijection, and for all $w, w' \in V$ and $c \in F$,

$$\Phi(w + w') = \Phi(w) + \Phi(w'), \qquad \Phi(cw) = \overline{c}\,\Phi(w).$$

In particular Φ is an isomorphism of vector spaces when $F = \mathbb{R}$, and is **conjugate-linear** when $F = \mathbb{C}$.

Proof. Each $\Phi(w)$ lies in V^* because the inner product is linear in the first slot. Surjectivity is the existence half of Theorem 11.5.1 and injectivity is the uniqueness half (if $\Phi(w) = \Phi(w')$, then $w = w'$). For the two rules, evaluate at an arbitrary v: additivity in the second slot gives $\langle v, w + w'\rangle = \langle v, w\rangle + \langle v, w'\rangle$, and conjugate-homogeneity in the second slot gives $\langle v, cw\rangle = \overline{c}\langle v, w\rangle$. Over $\mathbb{R}$ we have $\overline{c} = c$, so Φ is linear, and a linear bijection is an isomorphism (Theorem 0.4.7). ∎

This is the payoff Chapter 5 promised. There, $V \cong V^*$ was proved by choosing a basis $\mathcal{B}$ and sending $v_i \mapsto \varphi_i$, and the warning was that the isomorphism changes when $\mathcal{B}$ does: a vector does not come with "its" functional. Here Φ is written down with no basis at all. It is not free of choices — it depends on $\langle \cdot, \cdot\rangle$ — but the choice has been made once, in advance, and it is a choice of *geometry* rather than of coordinates. **The inner product is exactly the extra data that turns $V \cong V^*$ from a counting fact into a formula.**

The two descriptions fit together neatly. Let $\mathcal{B} = (v_1, \dots, v_n)$ be an **orthonormal** basis of V, with dual basis $(\varphi_1, \dots, \varphi_n)$ (Theorem 5.1.3). Then $\Phi(v_i)$ is the functional $v \mapsto \langle v, v_i\rangle$, which sends v_j to δ_{ij}; that is,

$$\Phi(v_i) = \varphi_i \qquad (i = 1, \dots, n),$$

over either field. Over $\mathbb{R}$, where Φ is linear, this says that Φ **is** Chapter 5's basis-dependent isomorphism $v_i \mapsto \varphi_i$, for the inner product that declares $\mathcal{B}$ orthonormal. Choosing a basis and choosing an inner product are two ways of supplying the same missing data.

Quick check. Let $\mathcal{B} = (v_1, \dots, v_n)$ be an orthonormal basis of a **complex** inner product space and let $(\varphi_1, \dots, \varphi_n)$ be its dual basis. What is the Riesz vector of $i\varphi_1$?

Solution. By the identity $\Phi(v_i) = \varphi_i$ above, which holds over $\mathbb{C}$ as well, φ_1 has Riesz vector v_1. For $i\varphi_1$, use conjugate-homogeneity in reverse: $\Phi(-iv_1) = \overline{(-i)}\Phi(v_1) = i\varphi_1$. So the Riesz vector of $i\varphi_1$ is $-iv_1$, not iv_1. Alternatively, apply the formula of Theorem 11.5.1: the

coordinates of the Riesz vector are $\overline{(i\varphi_1)(v_j)} = \overline{i\delta_{1j}} = -i\delta_{1j}$.

Warning. **Over $\mathbb{C}$, the Riesz map is not linear, so it is not an isomorphism of complex vector spaces.** It is a bijection, and it is linear over $\mathbb{R}$, but $\Phi(iw) = -i\Phi(w)$. Complex spaces V and V^* are still isomorphic — they have the same dimension — but the isomorphism produced by an inner product is the conjugate-linear Φ. Most of what we do with Φ (Section 6 builds the adjoint from it) is unaffected, because the conjugation cancels when Φ is used twice.

Warning. **The Riesz vector depends on the inner product, not on φ alone.** On $\mathbb{R}^2$ let $\varphi(\mathbf{x}) = x_1 + x_2$. With the dot product, $\varphi(\mathbf{x}) = \langle \mathbf{x}, (1,1)\rangle$, so $\mathbf{w} = (1,1)$. With the weighted inner product $\langle \mathbf{x}, \mathbf{y}\rangle = 2x_1y_1 + x_2y_2$, we need $2w_1 = 1$ and $w_2 = 1$, so $\mathbf{w} = (\frac{1}{2}, 1)$. Same functional, different representing vector. Whenever two inner products are in play, the notation must say which Φ is meant.

11.5.3 Annihilators become orthogonal complements

Section 3 warned that the annihilator $U^0 \subseteq V^*$ and the orthogonal complement $U^\perp \subseteq V$ are different objects that happen to have the same dimension. The Riesz map explains the coincidence: it carries one onto the other.

Theorem 11.5.3. Perp corresponds to the annihilator

Let V be a finite-dimensional inner product space over F, let U be a subspace, and let Φ be the Riesz map. Then

$$\Phi(U^\perp) = U^0.$$

Equivalently, a functional $\varphi \in V^*$ annihilates U if and only if its Riesz vector lies in $U^\perp$.

Proof. Let $w \in V$. Then

$$\Phi(w) \in U^0 \iff \langle u, w\rangle = 0 \text{ for every } u \in U \iff w \in U^\perp,$$

the first equivalence being the definition of the annihilator (Definition 5.2.1) applied to the functional $\Phi(w) = \langle \cdot, w\rangle$, and the second the definition of the orthogonal complement (Definition 11.3.1). Since Φ is a bijection (Corollary 11.5.2), the set $\Phi(U^\perp)$ is therefore exactly U^0.

■

The dimension count now has a reason rather than a coincidence: $\dim U^0 = \dim V - \dim U$ (Theorem 5.2.2) and $\dim U^\perp = \dim V - \dim U$ (Theorem 11.3.2 (c)) agree because the two spaces correspond under a bijection. (Over $\mathbb{C}$ the bijection is only conjugate-linear, but a conjugate-linear bijection still carries subspaces to subspaces of the same dimension, since $c\Phi(w) = \Phi(\overline{c}w)$.)

Here is the correspondence in $\mathbb{R}^3$. Let $U = \operatorname{span}((1,2,0))$ with the dot product. Then $U^\perp = \{(a,b,c) : a+2b = 0\} = \operatorname{span}((2,-1,0),(0,0,1))$, and U^0 consists of the functionals $\mathbf{x} \mapsto \alpha x_1 + \beta x_2 + \gamma x_3$ with $\alpha + 2\beta = 0$, spanned by $2x_1 - x_2$ and x_3. These are precisely $\Phi((2,-1,0))$ and $\Phi((0,0,1))$.

Section 6 uses Theorem 11.5.3 in the same way for maps: the dual map $T'\colon W^* \to V^*$ of Chapter 5 and the adjoint $T^*\colon W \to V$ are the same map, read through the two Riesz maps.

11.5.4 What happens in infinite dimension

The proof used an orthonormal basis of the whole of V, so finite dimension was essential, and the theorem genuinely fails without it.

Example. A functional with no Riesz vector

Let $V = \mathbb{R}[x]$ with the coefficient inner product $\langle p, q\rangle = \sum_{k\geq 0} a_k b_k$ for $p = \sum_k a_k x^k$ and $q = \sum_k b_k x^k$ (a finite sum, and an inner product by Exercise (C2: A subspace with no orthogonal complement) (a)). Show that $\varphi(p) = p(1)$ has no Riesz vector.

Solution. Suppose $w = \sum_k b_k x^k$ satisfied $\langle p, w\rangle = p(1)$ for every $p \in V$. Taking $p = x^i$ gives

$$b_i = \langle x^i, w\rangle = \varphi(x^i) = 1^i = 1 \qquad \text{for every } i \geq 0.$$

So every coefficient of w equals 1. But a polynomial has only finitely many nonzero coefficients, so no such w exists.

What fails is not "infinite dimension" as a slogan. Two things are missing at once, and the same example shows both. First, V has gaps: Exercise (C2: A subspace with no orthogonal complement) exhibits polynomials $p_m \in \ker\varphi$ with $\|p_m - 1\| = 1/\sqrt{m}$, so the vectors p_m crowd around the polynomial 1 without any limit being forced to lie in $\ker\varphi$. Second, φ is wildly discontinuous with respect to $\|\cdot\|$: along the same sequence, $\varphi(p_m) = 0$ for every m, while $\varphi(1) = 1$.

Repairing both defects — completing the space, and restricting attention to functionals φ with $|\varphi(v)| \leq C\|v\|$ — restores the theorem. That statement is the Riesz representation theorem for Hilbert spaces, and it belongs to analysis, not to linear algebra; we record its existence and go no further. What survives here with no hypotheses at all is the easy direction: for every $w \in V$, the map $\langle \cdot, w\rangle$ *is* a linear functional. Only the converse needs finite dimension.

11.5.5 Exercises

11.5.5.1 A. Check your understanding

Exercise. A1

(a) State the Riesz Representation Theorem, including every hypothesis.

(b) Write down the formula for the Riesz vector of φ in terms of an orthonormal basis, and say where the conjugate comes from.

(c) True or false: over $\mathbb{C}$, the Riesz map $\Phi\colon V \to V^*$ is a linear isomorphism. Justify your

answer.

(d) In what sense does an inner product answer the question Chapter 5 left open about $V \cong V^*$?

(e) Under Φ, which subspace of V^* corresponds to $U^\perp$?

(f) Give an example showing that the theorem fails in infinite dimension.

Solution.

(a) If V is a **finite-dimensional** inner product space over $\mathbb{R}$ or $\mathbb{C}$ and $\varphi \in V^*$, then there is a unique $w \in V$ with $\varphi(v) = \langle v, w\rangle$ for all $v \in V$ (Theorem 11.5.1).

(b) $w = \sum_i \overline{\varphi(e_i)} e_i$ for any orthonormal basis $(e_1, \dots, e_n)$. The conjugate appears because w occupies the second slot, where scalars come out conjugated: to make $\langle e_j, w\rangle$ equal $\varphi(e_j)$, the coefficient of e_j in w must be $\overline{\varphi(e_j)}$.

(c) False. Φ is a bijection, but $\Phi(cw) = \overline{c}\Phi(w)$, so it is conjugate-linear and not $\mathbb{C}$-linear (Corollary 11.5.2). It is linear over $\mathbb{R}$, and over $\mathbb{R}$ it is an isomorphism.

(d) Chapter 5 showed $V \cong V^*$ by counting dimensions, with an isomorphism depending on a chosen basis and no preferred choice available (Corollary 5.1.4). An inner product supplies a specific map $\Phi(w) = \langle \cdot, w\rangle$, written without reference to any basis. The inner product is the extra data; picking an orthonormal basis and picking an inner product that makes a given basis orthonormal are the same act.

(e) U^0, the annihilator of U (Theorem 11.5.3).

(f) On $\mathbb{R}[x]$ with the coefficient inner product, $\varphi(p) = p(1)$ has no Riesz vector: such a vector would need every coefficient equal to 1 (Example (A functional with no Riesz vector)).

11.5.5.2 B. Practice

Exercise. B1: Riesz vectors for evaluation

On $V = \mathbb{R}[x]_{\le 2}$ use the node inner product $\langle p, q\rangle = p(-1)q(-1) + p(0)q(0) + p(1)q(1)$. Find the Riesz vectors of $\varphi(p) = p(0)$ and of $\psi(p) = p(1)$.

Solution. Since $-1, 0, 1$ are three distinct nodes and $\dim V = 3$, a polynomial of V is determined by its values there (Theorem 3.7.3). Writing w for a candidate Riesz vector of φ, the requirement $\langle p, w\rangle = p(0)$ reads

$$p(-1)w(-1) + p(0)w(0) + p(1)w(1) = p(0) \qquad \text{for all } p \in V.$$

Taking for p the three Lagrange polynomials at the nodes, which have value 1 at one node and 0 at the others (Definition 6.6.1), the three equations become $w(-1) = 0$, $w(0) = 1$, $w(1) = 0$. The unique quadratic with those values is $w = 1 - x^2$.

For ψ the same argument gives $w(-1) = 0$, $w(0) = 0$, $w(1) = 1$, whose interpolant is $w = \frac{1}{2}x(x+1)$.

Check. For $p = 3 + 2x - x^2$ we get $p(-1) = 0$, $p(0) = 3$, $p(1) = 4$. Pairing with $1 - x^2$,

whose values are $0, 1, 0$, gives $3 = p(0)$; pairing with $\frac{1}{2}x(x+1)$, whose values are $0, 0, 1$, gives $4 = p(1)$.

Exercise. B2: Riesz vectors on matrices

On $M_n(\mathbb{C})$ with the Frobenius inner product $\langle \mathbf{A}, \mathbf{B} \rangle = \operatorname{tr}(\mathbf{B}^*\mathbf{A})$, find the Riesz vector of:

(a) $\varphi(\mathbf{A}) = a_{12}$ (assume $n \geq 2$);

(b) $\varphi_{\mathbf{C}}(\mathbf{A}) = \operatorname{tr}(\mathbf{C}\mathbf{A})$, for a fixed $\mathbf{C} \in M_n(\mathbb{C})$.

Solution.

(a) Let $\mathbf{E}_{12}$ be the matrix unit with a 1 in position $(1, 2)$ and zeros elsewhere. Then $\mathbf{E}_{12}^* = \mathbf{E}_{21}$, and $(\mathbf{E}_{21}\mathbf{A})_{kk} = \sum_j (\mathbf{E}_{21})_{kj} a_{jk}$, which is 0 unless $k = 2$, where it equals a_{12}. Hence $\operatorname{tr}(\mathbf{E}_{12}^*\mathbf{A}) = a_{12} = \varphi(\mathbf{A})$, and by uniqueness in Theorem 11.5.1 the Riesz vector is $\mathbf{E}_{12}$.

(b) We need $\mathbf{W}$ with $\operatorname{tr}(\mathbf{W}^*\mathbf{A}) = \operatorname{tr}(\mathbf{C}\mathbf{A})$ for all $\mathbf{A}$. Taking $\mathbf{W}^* = \mathbf{C}$, that is $\mathbf{W} = \mathbf{C}^*$, works. By uniqueness it is the only answer. Part (a) is the case $\mathbf{C} = \mathbf{E}_{21}$, and the example in the text is the case $\mathbf{C} = \mathbf{I}_n$.

Exercise. B3: The correspondence in $\mathbb{R}^4$

In $\mathbb{R}^4$ with the dot product, let $U = \operatorname{span}\big((1, 0, 1, 0), (0, 1, 0, 1)\big)$. Find a basis of $U^\perp$ and a basis of U^0, and verify that Φ carries the first to the second.

Solution. A vector (a, b, c, d) lies in $U^\perp$ exactly when $a + c = 0$ and $b + d = 0$ (Proposition 11.3.1, testing against the two spanning vectors). So

$$U^\perp = \operatorname{span}\big((1, 0, -1, 0),\ (0, 1, 0, -1)\big),$$

of dimension $2 = 4 - 2$.

A functional $\varphi(\mathbf{x}) = \alpha x_1 + \beta x_2 + \gamma x_3 + \delta x_4$ lies in U^0 exactly when it kills both spanning vectors, that is, when $\alpha + \gamma = 0$ and $\beta + \delta = 0$. So $U^0 = \operatorname{span}(\varphi_1, \varphi_2)$ with

$$\varphi_1(\mathbf{x}) = x_1 - x_3, \qquad \varphi_2(\mathbf{x}) = x_2 - x_4.$$

Over $\mathbb{R}$, $\Phi(\mathbf{w})(\mathbf{x}) = \langle \mathbf{x}, \mathbf{w} \rangle = \sum_i x_i w_i$, so $\Phi((1, 0, -1, 0)) = \varphi_1$ and $\Phi((0, 1, 0, -1)) = \varphi_2$. Since Φ is linear over $\mathbb{R}$ and bijective, it carries the span of the two vectors onto the span of the two functionals, confirming Theorem 11.5.3.

11.5.5.3 C. Going deeper

Exercise. C1: Riesz without an orthonormal basis

Let V be a finite-dimensional inner product space and let $\varphi \in V^*$ be nonzero. Write $K = \ker \varphi$.

(a) Prove that $\dim K^{\perp} = 1$.

(b) Let u span $K^{\perp}$ with $\|u\| = 1$. Prove that $w = \overline{\varphi(u)}\,u$ satisfies $\varphi(v) = \langle v, w\rangle$ for all $v \in V$.

(c) Deduce Theorem 11.5.1, including the case $\varphi = 0$.

Hint for (b): split v using $V = K \oplus K^{\perp}$.

Solution.

(a) Since $\varphi \neq 0$, its image is a nonzero subspace of the 1-dimensional space F, so $\operatorname{im}\varphi = F$ and $\operatorname{rank}\varphi = 1$. By Theorem 2.3.1, $\dim K = \dim V - 1$, and by Theorem 11.3.2 (c), $\dim K^{\perp} = \dim V - \dim K = 1$.

(b) By Theorem 11.3.2 (a), $V = K \oplus K^{\perp}$, so each $v \in V$ is $v = k + cu$ with $k \in K$ and $c \in F$. Then $\varphi(v) = \varphi(k) + c\varphi(u) = c\varphi(u)$. On the other side,

$$\langle v, w\rangle = \langle k + cu, \overline{\varphi(u)}u\rangle = \varphi(u)\big(\langle k, u\rangle + c\langle u, u\rangle\big) = \varphi(u)\,c,$$

using conjugate-homogeneity in the second slot (which conjugates $\overline{\varphi(u)}$ back to $\varphi(u)$), then $\langle k, u\rangle = 0$ because $u \in K^{\perp}$ and $k \in K$, and $\langle u, u\rangle = \|u\|^2 = 1$. The two sides agree.

(c) If $\varphi = 0$, then $w = 0$ represents it. If $\varphi \neq 0$, part (a) supplies a unit vector u spanning the line $K^{\perp}$ — any nonzero vector of that line divided by its norm — and part (b) supplies w. Uniqueness is unchanged: if $\langle v, w\rangle = \langle v, w'\rangle$ for all v, take $v = w - w'$ to get $\|w - w'\|^2 = 0$. Note that this proof never mentions a basis of V; it uses only the orthogonal decomposition along the hyperplane $\ker\varphi$.

Exercise. C2: Evaluation at 0 is not represented

Let $V = \mathbb{R}[x]$ with $\langle p, q\rangle = \int_0^1 p(t)q(t)\,\mathrm{d}t$, and let $\varphi(p) = p(0)$. Suppose, for a contradiction, that $w = \sum_{j=0}^{d} c_j x^j$ satisfied $\langle p, w\rangle = p(0)$ for every $p \in V$.

(a) Show that $\displaystyle\sum_{j=0}^{d} \frac{c_j}{k + j + 1} = 0$ for every integer $k \geq 1$, and that the same sum equals 1 for $k = 0$.

(b) Let $N(s) = \sum_{j=0}^{d} c_j \prod_{i \neq j}(s + i + 1)$, the product running over $i \in \{0, \ldots, d\} \setminus \{j\}$. Show that $N(k) = 0$ for every integer $k \geq 1$, and deduce that N is the zero polynomial.

(c) Deduce that every $c_j = 0$, and hence a contradiction.

Solution.

(a) Take $p = x^k$. Then $p(0) = 0$ for $k \geq 1$ and $p(0) = 1$ for $k = 0$, while

$$\langle x^k, w\rangle = \int_0^1 t^k \sum_{j=0}^{d} c_j t^j\,\mathrm{d}t = \sum_{j=0}^{d} c_j \int_0^1 t^{k+j}\,\mathrm{d}t = \sum_{j=0}^{d} \frac{c_j}{k + j + 1}.$$

(b) For an integer $k \geq 1$, multiply the identity of (a) by $\prod_{i=0}^{d}(k + i + 1)$, a nonzero number. Each term $c_j/(k + j + 1)$ becomes $c_j \prod_{i \neq j}(k + i + 1)$, so the result is exactly

$N(k) = 0$. Thus the polynomial N, of degree at most d, has the infinitely many distinct roots $1, 2, 3, \ldots$. A nonzero polynomial of degree d has at most d distinct roots (Corollary 6.1.4), so $N = 0$.

(c) Evaluate N at $s = -(j_0 + 1)$ for a fixed $j_0 \in \{0, \ldots, d\}$. Every product $\prod_{i \neq j}(s + i + 1)$ with $j \neq j_0$ contains the factor $s + j_0 + 1 = 0$ and so vanishes; the term $j = j_0$ contributes $c_{j_0} \prod_{i \neq j_0}(i - j_0)$, and that product is nonzero because the factors $i - j_0$ are nonzero integers. Since $N = 0$, we get $c_{j_0} = 0$. This holds for every j_0, so $w = 0$. But then the case $k = 0$ of (a) reads $0 = 1$, a contradiction. Hence no Riesz vector exists.

The moral matches Example (A functional with no Riesz vector): point evaluation is a perfectly good linear functional, but on an infinite-dimensional space it is not represented by any vector, whichever of these two inner products we use.

11.6 The Adjoint

Chapter 5 attached to every linear map $T\colon V \to W$ a dual map $T'\colon W^* \to V^*$ running backwards between the dual spaces (Definition 5.4.1). An inner product puts a copy of the dual space back inside the space itself, by the Riesz correspondence of Theorem 11.5.1. Composing the two turns the backward map on functionals into a backward map on vectors, $W \to V$. That map is the **adjoint** T^*, and it is the single most useful construction in this chapter. This section builds it, computes its matrix, uses it to split both V and W into orthogonal pieces, and names the two classes of operators it singles out.

Throughout, F is $\mathbb{R}$ or $\mathbb{C}$, every inner product is linear in the first slot and conjugate-linear in the second, and every inner product space is finite-dimensional over F unless we say otherwise.

11.6.1 Paying to cross the inner product

Here is the situation the adjoint answers. We have a linear map T and an inner product, and an expression like $\langle Tv, w\rangle$ in which T sits in the first slot. Very often the argument needs T out of the way: we know something about w, or about the second slot, and the T is in the wrong place. Can we move it across?

For matrices the answer is already on the table. Lemma 11.4.1, proved in Section 4, says that for $\mathbf{A} \in M_{m\times n}(F)$, $\mathbf{x} \in F^n$ and $\mathbf{y} \in F^m$,

$$\langle \mathbf{A}\mathbf{x}, \mathbf{y}\rangle = \langle \mathbf{x}, \mathbf{A}^*\mathbf{y}\rangle,$$

with $\mathbf{A}^*$ the conjugate transpose of Definition 11.1.2. The one-line reason is worth repeating, because it is the model for the whole section: writing $\mathbf{y}^*$ for the row vector $\overline{\mathbf{y}}^{\top}$, the standard inner product is the matrix product $\langle \mathbf{x}, \mathbf{y}\rangle = \mathbf{y}^*\mathbf{x}$, so $\langle \mathbf{A}\mathbf{x}, \mathbf{y}\rangle = (\mathbf{y}^*\mathbf{A})\mathbf{x} = (\mathbf{A}^*\mathbf{y})^*\mathbf{x} = \langle \mathbf{x}, \mathbf{A}^*\mathbf{y}\rangle$. So a matrix does cross the inner product, and the price is a conjugate transpose. Over $\mathbb{R}$ the price is a transpose; over $\mathbb{C}$ a transpose and a conjugation.

To move a map from the first slot to the second slot, you pay with a new map going the other way.

That statement is about matrices and the standard inner product. The next theorem says the same map exists for every linear map between finite-dimensional inner product spaces, with no coordinates at all. Riesz does the work: the expression $\langle Tv, w\rangle$, read as a function of v with w frozen, is a linear functional on V, and Theorem 11.5.1 says that every such functional is pairing against one fixed vector.

Theorem 11.6.1. Existence and Uniqueness of the Adjoint

Let V and W be finite-dimensional inner product spaces over F, and let $T \in \mathcal{L}(V, W)$. Then there is exactly one map $T^*\colon W \to V$ such that

$$\langle Tv, w\rangle_W = \langle v, T^*w\rangle_V \qquad \text{for all } v \in V,\ w \in W,$$

and this T^* is linear.

Idea. Freeze w. Then $v \mapsto \langle Tv, w\rangle$ is a functional on V, so Riesz hands us one vector that represents it; call it T^*w. That is the whole of existence, and it also gives uniqueness, since the Riesz vector is unique. Linearity is not part of the construction and has to be checked, and the way to check an identity between vectors is the signature move of this chapter: show the difference pairs to zero with everything, then pair it with itself.

Proof. **Existence.** Fix $w \in W$ and define $\varphi_w\colon V \to F$ by $\varphi_w(v) \coloneqq \langle Tv, w\rangle$. It is linear: for $v_1, v_2 \in V$ and $c \in F$,

$$\varphi_w(v_1 + cv_2) = \langle Tv_1 + cTv_2, w\rangle = \langle Tv_1, w\rangle + c\langle Tv_2, w\rangle,$$

using the linearity of T and linearity in the **first** slot (Definition 11.1.1). Since V is finite-dimensional, Theorem 11.5.1 gives a unique $u \in V$ with $\varphi_w(v) = \langle v, u\rangle$ for all $v \in V$. Define $T^*w \coloneqq u$. By construction $\langle Tv, w\rangle = \langle v, T^*w\rangle$ for all v, which is the required identity.

Uniqueness. Suppose $S\colon W \to V$ also satisfies $\langle Tv, w\rangle = \langle v, Sw\rangle$ for all v, w. Fix w. Then $\langle v, T^*w - Sw\rangle = 0$ for every $v \in V$, by conjugate-additivity in the second slot. Taking $v = T^*w - Sw$ gives $\|T^*w - Sw\|^2 = 0$, so $T^*w = Sw$ by positive definiteness (Definition 11.1.1). As w was arbitrary, $S = T^*$.

Linearity. Let $w_1, w_2 \in W$ and $c \in F$. For every $v \in V$,

$$\begin{aligned}\langle v, T^*(w_1 + cw_2)\rangle &= \langle Tv, w_1 + cw_2\rangle = \langle Tv, w_1\rangle + \overline{c}\langle Tv, w_2\rangle \\ &= \langle v, T^*w_1\rangle + \overline{c}\langle v, T^*w_2\rangle = \langle v, T^*w_1 + cT^*w_2\rangle,\end{aligned}$$

where the second and the last equalities use conjugate-linearity in the second slot. Hence $\langle v, T^*(w_1 + cw_2) - T^*w_1 - cT^*w_2\rangle = 0$ for every v, and pairing the right-hand vector with itself makes it 0. So T^* is linear. This proves the theorem. ∎

Notice where the two conjugations went. The scalar c came out of the second slot as $\overline{c}$ and went back in as c; the two conjugations cancel, which is exactly why T^* is linear and not conjugate-linear. Notice also that only V had to be finite-dimensional: the proof never uses a basis of W. We keep both spaces finite-dimensional because the rest of the section does need it.

11.6.2 The adjoint

The theorem produced an object attached to T and to the two inner products, and it is unique, so it deserves a name and a notation.

The adjoint of T is the map you must apply on the other side of the inner product to leave the value unchanged.

Definition 11.6.1. Adjoint

Let V and W be finite-dimensional inner product spaces over F and let $T \in \mathcal{L}(V, W)$. The **adjoint** of T is the unique linear map $T^* \in \mathcal{L}(W, V)$ satisfying

$$\langle Tv, w\rangle = \langle v, T^*w\rangle \qquad \text{for all } v \in V \text{ and all } w \in W.$$

In words, clause by clause:

- T^* goes **backwards**: its domain is the codomain of T. If $T\colon V \to W$ then $T^*\colon W \to V$, like the dual map and unlike anything else we have attached to T.
- The defining equation is an identity in **two** variables, and it must hold for **all** v and **all** w. By linearity in the first slot and conjugate-linearity in the second, it is enough to check it when v and w range over bases of V and of W, and that is usually how T^* gets computed.
- The left inner product lives in W, the right one in V. They are different inner products; only the notation hides it.
- The definition mentions no basis, but it does mention the two inner products. Change an inner product and T^* changes.

Example. Four Adjoints

Compute the adjoint in each case. In (a) and (b) the spaces carry the standard inner product; in (c) and (d), V and W are arbitrary finite-dimensional inner product spaces.

(a) $T_{\mathbf{A}}\colon \mathbb{C}^3 \to \mathbb{C}^2$ for $\mathbf{A} = \begin{pmatrix} 1+i & 0 & 2 \\ i & 1 & -i \end{pmatrix}$.

(b) The forward shift $S\colon F^3 \to F^3$, $S(x_1, x_2, x_3) = (0, x_1, x_2)$.

(c) id_V and the zero map $V \to W$.

(d) The orthogonal projection $P_U\colon V \to V$ onto a subspace $U \subseteq V$.

Solution.

(a) By Lemma 11.4.1, $\langle \mathbf{Ax}, \mathbf{y}\rangle = \langle \mathbf{x}, \mathbf{A}^*\mathbf{y}\rangle$ for all $\mathbf{x}, \mathbf{y}$, which is the defining identity of the adjoint; so $(T_{\mathbf{A}})^* = T_{\mathbf{A}^*}$ for every matrix, by uniqueness in Theorem 11.6.1. Here

$$\mathbf{A}^* = \begin{pmatrix} 1-i & -i \\ 0 & 1 \\ 2 & i \end{pmatrix},$$

so $T^*_{\mathbf{A}}(y_1, y_2) = \big((1-i)y_1 - iy_2,\ y_2,\ 2y_1 + iy_2\big)$.

(b) Directly from the definition, for $\mathbf{x}, \mathbf{y} \in F^3$,

$$\langle S\mathbf{x}, \mathbf{y}\rangle = 0 \cdot \overline{y_1} + x_1\overline{y_2} + x_2\overline{y_3}, \qquad \langle \mathbf{x}, (y_2, y_3, 0)\rangle = x_1\overline{y_2} + x_2\overline{y_3} + x_3 \cdot 0.$$

The two agree for all $\mathbf{x}$ and $\mathbf{y}$. So S^* is the **backward** shift $S^*(y_1, y_2, y_3) = (y_2, y_3, 0)$. The forward shift pushes coordinates up; its adjoint pushes them down and drops the one that falls off the end.

(c) $\langle \mathrm{id}_V\, v, w\rangle = \langle v, w\rangle = \langle v, \mathrm{id}_V\, w\rangle$, so $(\mathrm{id}_V)^* = \mathrm{id}_V$. For the zero map, $\langle 0v, w\rangle = 0 = \langle v, 0\rangle$ for every w, so $0^* = 0$, the zero map $W \to V$. These are the degenerate cases, and they say the adjoint fixes what should be fixed.

(d) By Theorem 11.3.3 (b), $\langle P_U u, v\rangle = \langle u, P_U v\rangle$ for all $u, v \in V$. Since P_U is an operator on V, this is exactly the defining identity for its own adjoint, so $P_U^* = P_U$ by uniqueness in Theorem 11.6.1.

Part (d) is the first place where uniqueness earns its keep: we did not construct P_U^*, we recognized an identity we already had and read the answer off it. That move recurs throughout the section.

Example. Differentiation on a Polynomial Space

On $V = \mathbb{R}[x]_{\le 2}$ define $\langle p, q\rangle := a_0b_0 + a_1b_1 + a_2b_2$ for $p = a_0 + a_1x + a_2x^2$ and $q = b_0 + b_1x + b_2x^2$. (This is the standard inner product of $\mathbb{R}^3$ transported by the coordinate map, so it is an inner product, and $\mathcal{E} = (1, x, x^2)$ is an orthonormal basis.) Find the adjoint of $D\colon V \to V$, $D(p) = p'$.

Solution. For p, q as above, $D(p) = a_1 + 2a_2x$, so

$$\langle Dp, q\rangle = a_1b_0 + 2a_2b_1 + 0 \cdot b_2.$$

We need $D^*q = c_0 + c_1x + c_2x^2$ with $\langle p, D^*q\rangle = a_0c_0 + a_1c_1 + a_2c_2$ equal to the line above for **all** a_0, a_1, a_2. Comparing coefficients of a_0, a_1, a_2 gives $c_0 = 0$, $c_1 = b_0$, $c_2 = 2b_1$. Hence

$$D^*(b_0 + b_1x + b_2x^2) = b_0x + 2b_1x^2.$$

So D^* raises degrees where D lowers them. It is not an antiderivative: the coefficient 2 and the discarded b_2 both say so.

Here is a nonexample by minimal change, and it explains the conjugate. On $\mathbb{C}^2$, keep the standard inner product but try the plain transpose: is $\langle \mathbf{Ax}, \mathbf{y}\rangle = \langle \mathbf{x}, \mathbf{A}^\top \mathbf{y}\rangle$? Take $\mathbf{A} = \begin{pmatrix} i & 0 \\ 0 & 0 \end{pmatrix}$, which is symmetric, so $\mathbf{A}^\top = \mathbf{A}$, and take $\mathbf{x} = \mathbf{y} = \mathbf{e}_1$. Then $\langle \mathbf{Ae}_1, \mathbf{e}_1\rangle = \langle (i, 0), (1, 0)\rangle = i$, while $\langle \mathbf{e}_1, \mathbf{A}^\top \mathbf{e}_1\rangle = \langle (1, 0), (i, 0)\rangle = 1 \cdot \bar{i} = -i$. The two sides differ. Everything about the transpose still works — it is linear and it reverses products — but it fails the one clause that matters, the defining identity, and it fails it by a conjugation. Over $\mathbb{R}$ the conjugation is invisible and the transpose is correct.

Why this definition. We could have asked instead for a map R with $\langle Rw, v\rangle = \langle w, Tv\rangle$. Conjugating Definition 11.6.1 shows that $R = T^*$ as well, so nothing new appears; the convention above is the one that reads left to right. The notation T^* matches the matrix notation $\mathbf{A}^*$ of Example (Four Adjoints) (a), which is the point of the star. The word *adjoint* is unrelated to the *adjugate* of Definition 7.6.2; the collision of names is historical and the two objects have nothing to do with each other.

Warning. **The adjoint depends on the inner product, not on T alone.** On $\mathbb{R}^2$ take $T(x_1, x_2) = (x_2, 0)$. With the standard inner product, $T^*(y_1, y_2) = (0, y_1)$. With the weighted inner product $\langle \mathbf{x}, \mathbf{y}\rangle_w := x_1y_1 + 2x_2y_2$, which is also an inner product, we need

$\langle T\mathbf{x}, \mathbf{y}\rangle_w = x_2 y_1$ to equal $\langle \mathbf{x}, S\mathbf{y}\rangle_w = x_1(S\mathbf{y})_1 + 2x_2(S\mathbf{y})_2$ for all $\mathbf{x}$, which forces $(S\mathbf{y})_1 = 0$ and $(S\mathbf{y})_2 = \frac{1}{2}y_1$. So the adjoint is now $\mathbf{y} \mapsto (0, \frac{1}{2}y_1)$, a different map. Writing T^* without saying which inner product is meant is as incomplete as writing a matrix of T without saying which basis.

Quick check. Let $T \in \mathcal{L}(V, W)$ with $T \neq 0$. Can $T^* = 0$?

Solution. No. If $T^* = 0$ then $\langle Tv, w\rangle = \langle v, 0\rangle = 0$ for all v, w. Taking $w = Tv$ gives $\|Tv\|^2 = 0$, so $Tv = 0$ for every v, that is $T = 0$. This is the "pair it with itself" move again.

Finally, the finite-dimensional hypothesis is not decoration.

Example. A Map With No Adjoint

On $V = \mathbb{R}[x]$, all polynomials, define $\langle p, q\rangle := \sum_{k\geq 0} a_k b_k$, a finite sum since the coefficients vanish eventually. Show that $T\colon V \to V$, $T(p) := p(1) \cdot 1$, has no adjoint.

Solution. T is linear, being evaluation at 1 followed by multiplication by the constant polynomial 1. Suppose T^* existed. Apply the defining identity with $p = x^k$ and with $q = 1$, the constant polynomial: since $T(x^k) = 1$,

$$\langle x^k, T^*(1)\rangle = \langle T(x^k), 1\rangle = \langle 1, 1\rangle = 1 \qquad \text{for every } k \geq 0.$$

But $\langle x^k, T^*(1)\rangle$ is the coefficient of x^k in $T^*(1)$, so every coefficient of the polynomial $T^*(1)$ equals 1. No polynomial has infinitely many nonzero coefficients, a contradiction. So T^* does not exist.

What failed is Riesz: the functional $p \mapsto p(1)$ on $\mathbb{R}[x]$ is not pairing against any vector, precisely as in the infinite-dimensional discussion of Section 5.

11.6.3 Arithmetic of the adjoint

Starring behaves like transposing: it is additive, conjugate-homogeneous, order-reversing on products, and an involution.

Theorem 11.6.2. Properties of the Adjoint

Let U, V, W be finite-dimensional inner product spaces over F.

(a) For $S, T \in \mathcal{L}(V, W)$, $(S + T)^* = S^* + T^*$.
(b) For $T \in \mathcal{L}(V, W)$ and $c \in F$, $(cT)^* = \overline{c}\, T^*$.
(c) For $T \in \mathcal{L}(U, V)$ and $S \in \mathcal{L}(V, W)$, $(ST)^* = T^* S^*$.
(d) $T^{**} = T$ for every $T \in \mathcal{L}(V, W)$, and $(\mathrm{id}_V)^* = \mathrm{id}_V$.
(e) If $T \in \mathcal{L}(V, W)$ is invertible, then T^* is invertible and $(T^*)^{-1} = (T^{-1})^*$.

Idea. Each claim says that two maps $W \to V$ agree. By the uniqueness in Theorem 11.6.1, it is enough to check that the proposed right-hand side satisfies the defining identity of the left-hand side. So every proof is one chain of inner products, and nothing is ever solved for. Part (e) needs no chain at all: star the two equations $T^{-1}T = \mathrm{id}_V$ and $TT^{-1} = \mathrm{id}_W$ and use (c) and (d).

Proof. Throughout, v and w denote arbitrary vectors of the relevant spaces, and each verification ends by invoking uniqueness in Theorem 11.6.1.

(a) $\langle (S+T)v, w\rangle = \langle Sv, w\rangle + \langle Tv, w\rangle = \langle v, S^*w\rangle + \langle v, T^*w\rangle = \langle v, (S^*+T^*)w\rangle$, the last step by additivity in the second slot.

(b) $\langle (cT)v, w\rangle = c\langle Tv, w\rangle = c\langle v, T^*w\rangle = \langle v, \overline{c}\,T^*w\rangle$, where the last equality puts $\overline{c}$ into the second slot, which conjugates it back to c.

(c) For $u \in U$ and $w \in W$, using the defining identity for S and then for T,

$$\langle (ST)u, w\rangle = \langle S(Tu), w\rangle = \langle Tu, S^*w\rangle = \langle u, T^*(S^*w)\rangle = \langle u, (T^*S^*)w\rangle.$$

(d) The map $T^* \in \mathcal{L}(W, V)$ has its own adjoint $T^{**} \in \mathcal{L}(V, W)$, determined by $\langle T^*w, v\rangle = \langle w, T^{**}v\rangle$. Conjugating the defining identity of T^* and using the conjugate symmetry of the inner product (Definition 11.1.1),

$$\langle T^*w, v\rangle = \overline{\langle v, T^*w\rangle} = \overline{\langle Tv, w\rangle} = \langle w, Tv\rangle.$$

So T satisfies the identity that defines T^{**}, and $T^{**} = T$. That $(\mathrm{id}_V)^* = \mathrm{id}_V$ is Example (Four Adjoints) (c).

(e) Suppose T is invertible, with $T^{-1} \in \mathcal{L}(W, V)$. Applying (c) and (d) to $T^{-1}T = \mathrm{id}_V$ and to $TT^{-1} = \mathrm{id}_W$,

$$T^*(T^{-1})^* = (T^{-1}T)^* = (\mathrm{id}_V)^* = \mathrm{id}_V, \qquad (T^{-1})^*T^* = (TT^{-1})^* = \mathrm{id}_W\,.$$

Both composites are identities, so T^* is invertible with inverse $(T^{-1})^*$ (Definition 2.5.1). This proves the theorem.

■

Parts (a) and (b) say that $T \mapsto T^*$ is **conjugate**-linear, not linear: over $\mathbb{C}$ it is the same half-twist that the Riesz map carries in Corollary 11.5.2, and for the same reason. Part (c) is the order reversal familiar from transposes and from dual maps (Theorem 5.4.1 (b)). Part (d) says that starring is an involution, so no new maps appear by starring twice.

11.6.4 The matrix of the adjoint

Now fix bases and compute. The conclusion is the one the opening computation suggests, **provided the bases are orthonormal**; the hypothesis is where all the content sits.

Theorem 11.6.3. The Matrix of the Adjoint Is the Conjugate Transpose

Let V and W be finite-dimensional inner product spaces over F, let $\mathcal{B} = (v_1, \ldots, v_n)$ be an **orthonormal** basis of V and $\mathcal{C} = (w_1, \ldots, w_m)$ an **orthonormal** basis of W, and let

$T \in \mathcal{L}(V, W)$. Then

$$[T^*]_{\mathcal{C}}^{\mathcal{B}} = \left([T]_{\mathcal{B}}^{\mathcal{C}}\right)^*.$$

Idea. In an orthonormal basis a coordinate is an inner product (Theorem 11.2.1), so every matrix entry is an inner product, and the defining identity of T^* says precisely how those inner products are related. Write both entries out and compare.

Proof. Write $\mathbf{A} = [T]_{\mathcal{B}}^{\mathcal{C}}$ and $\mathbf{B} = [T^*]_{\mathcal{C}}^{\mathcal{B}}$, of sizes $m \times n$ and $n \times m$. By Definition 4.1.1, the j-th column of $\mathbf{A}$ holds the coordinates of Tv_j in $\mathcal{C}$, and by Theorem 11.2.1 those coordinates are inner products:

$$a_{ij} = \langle Tv_j, w_i \rangle, \qquad b_{ji} = \langle T^* w_i, v_j \rangle.$$

By conjugate symmetry and Definition 11.6.1,

$$b_{ji} = \langle T^* w_i, v_j \rangle = \overline{\langle v_j, T^* w_i \rangle} = \overline{\langle Tv_j, w_i \rangle} = \overline{a_{ij}}.$$

So $\mathbf{B}$ is the matrix whose (j, i) entry is $\overline{a_{ij}}$, which is $\mathbf{A}^*$. This proves the theorem.

■

Both hypotheses were used: orthonormality of $\mathcal{C}$ to read off a_{ij}, orthonormality of $\mathcal{B}$ to read off b_{ji}. Drop either one and the conclusion fails.

Warning. **The rule $[T^*]_{\mathcal{B}}^{\mathcal{B}} = \left([T]_{\mathcal{B}}^{\mathcal{B}}\right)^*$ is false in a basis that is not orthonormal.** The next example exhibits a two-dimensional counterexample with integer entries. The correct general statement involves the Gram matrix and is recorded in the remark after it.

Example. The Transpose Rule Breaks

Work in $\mathbb{R}^2$ with the standard inner product and let $T(x_1, x_2) = (x_2, 0)$. Let $\mathcal{B} = (\mathbf{b}_1, \mathbf{b}_2)$ with $\mathbf{b}_1 = (1, 0)$ and $\mathbf{b}_2 = (1, 1)$. Compute $[T]_{\mathcal{B}}^{\mathcal{B}}$ and $[T^*]_{\mathcal{B}}^{\mathcal{B}}$, and compare the second with the transpose of the first.

Solution. First, $\mathcal{B}$ is a basis (two independent vectors in $\mathbb{R}^2$) but it is not orthonormal: $\langle \mathbf{b}_1, \mathbf{b}_2 \rangle = 1 \neq 0$ and $\|\mathbf{b}_2\| = \sqrt{2}$.

By Example (Four Adjoints) (a) with $\mathbf{A} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$, the adjoint is $T^*(y_1, y_2) = (0, y_1)$.

The matrix of T. $T\mathbf{b}_1 = (0, 0) = 0\mathbf{b}_1 + 0\mathbf{b}_2$ and $T\mathbf{b}_2 = (1, 0) = 1\mathbf{b}_1 + 0\mathbf{b}_2$, so

$$[T]_{\mathcal{B}}^{\mathcal{B}} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \qquad \left([T]_{\mathcal{B}}^{\mathcal{B}}\right)^{\top} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}.$$

The matrix of T^*. $T^*\mathbf{b}_1 = (0, 1)$ and $T^*\mathbf{b}_2 = (0, 1)$. Solving $(0, 1) = c_1(1, 0) + c_2(1, 1)$ gives

$c_2 = 1$ and $c_1 = -1$, so both columns are $(-1, 1)$:

$$[T^*]_{\mathcal{B}}^{\mathcal{B}} = \begin{pmatrix} -1 & -1 \\ 1 & 1 \end{pmatrix}.$$

The two displayed matrices are different: the transpose rule fails. A quick sanity check that nothing else went wrong: the traces are 0 and 0, and both matrices have rank 1, as they must, since $\operatorname{rank} T = \operatorname{rank} T^* = 1$.

Remark. The correct rule. Let $\mathbf{G}$ be the Gram matrix of $\mathcal{B}$, so $g_{ij} = \langle \mathbf{b}_j, \mathbf{b}_i \rangle$ (Definition 11.2.3), which is invertible because $\mathcal{B}$ is independent (Theorem 11.2.5 (c)). Then $[T^*]_{\mathcal{B}}^{\mathcal{B}} = \mathbf{G}^{-1} \left([T]_{\mathcal{B}}^{\mathcal{B}} \right)^* \mathbf{G}$. In the example, $\mathbf{G} = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}$ and $\mathbf{G}^{-1} = \begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}$, and indeed

$$\begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix} = \begin{pmatrix} -1 & -1 \\ 1 & 1 \end{pmatrix}.$$

When $\mathcal{B}$ is orthonormal, $\mathbf{G} = \mathbf{I}$ and the rule collapses to Theorem 11.6.3. Exercise (C3: The adjoint in an arbitrary basis) asks for the general rule.

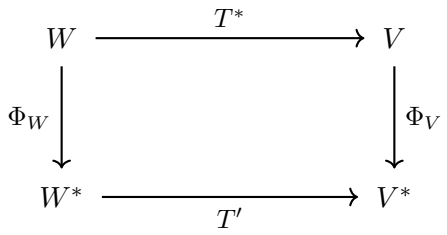
11.6.5 The adjoint and the dual map

Two backward maps are now attached to T: the dual map $T' \colon W^* \to V^*$ of Definition 5.4.1, which needs no inner product and sends a functional f to $f \circ T$, and the adjoint $T^* \colon W \to V$, which does need one. The Riesz map converts one into the other. This subsection is a connection, not a tool: nothing later in the book is proved with it, so a reader who has not met Chapter 5 can go straight to the next subsection and lose nothing.

Recall from Corollary 11.5.2 the Riesz map, the conjugate-linear bijection

$$\Phi_V \colon V \to V^*, \qquad \Phi_V(u) \coloneqq \langle \cdot, u \rangle,$$

and likewise Φ_W for W. The following square commutes: the two backward maps run horizontally, the two Riesz maps vertically, and the two routes from W to V^* agree.



the vertical maps are the Riesz maps, one for each space

Theorem 11.6.4. The Adjoint Is the Dual Map Read Through Riesz

Optional: nothing later depends on this.

Let V, W be finite-dimensional inner product spaces over F and $T \in \mathcal{L}(V, W)$. Then

$$\Phi_V \circ T^* = T' \circ \Phi_W.$$

Proof. Both sides send $w \in W$ to a functional on V, so evaluate them at an arbitrary $v \in V$. By Definition 5.4.1 and the definition of Φ_W,

$$(T'(\Phi_W w))(v) = (\Phi_W w)(Tv) = \langle Tv, w\rangle,$$

while by Definition 11.6.1 and the definition of Φ_V,

$$(\Phi_V(T^* w))(v) = \langle v, T^* w\rangle = \langle Tv, w\rangle.$$

The two agree for all v and all w, so the two maps are equal.

■

Read backwards, the identity says $T^* = \Phi_V^{-1} T' \Phi_W$: transport a vector of W into W^*, pull it back with the dual map, transport it back into V. So the adjoint carries no information that T' does not already carry; what it carries is the inner product used to transport. The matrix statements match too. In dual bases the matrix of T' is the transpose of the matrix of T (Theorem 5.4.2); in orthonormal bases the matrix of T^* is the conjugate transpose (Theorem 11.6.3). The extra conjugation is the price of Φ_V being conjugate-linear, and over $\mathbb{R}$ it disappears, which is why the two constructions are so easily confused there.

11.6.6 The four fundamental subspaces

Chapter 3 attached four subspaces to a matrix: $\mathrm{col}(\mathbf{A})$ and $\mathrm{nul}(\mathbf{A})$ (Definition 3.5.1, Definition 3.5.3) together with $\mathrm{row}(\mathbf{A})$ (Definition 3.5.2) and the null space of the transpose. Chapter 5 related two of them, with Theorem 5.6.5: the row space of $\mathbf{A}$ **is** the annihilator of $\mathrm{nul}(\mathbf{A})$, so every linear equation satisfied by the solution set is a combination of the given ones. That statement lived in the dual space, and Definition 5.2.1 carried a warning with it: over a general field the annihilator has no right to be pictured as a perpendicular. With an inner product, it does. The four subspaces become two orthogonal splittings, one in each of the two spaces.

Theorem 11.6.5. The Four Fundamental Subspaces

Let V, W be finite-dimensional inner product spaces over F and let $T \in \mathcal{L}(V, W)$. Then

(a) $\ker T^* = (\operatorname{im} T)^\perp$;
(b) $\operatorname{im} T^* = (\ker T)^\perp$;
(c) $V = \ker T \oplus \operatorname{im} T^*$;
(d) $W = \operatorname{im} T \oplus \ker T^*$.

Idea. Only (a) needs an argument, and it is one chain of equivalences: $T^* w = 0$ says a vector is zero, so pair it with everything, and the defining identity turns "everything" into "everything in the image of T". Then (b) is (a) applied to T^*, using $T^{**} = T$, and complemented with $(U^\perp)^\perp = U$. Parts (c) and (d) are then just the orthogonal decomposition theorem.

Proof.

(a) Let $w \in W$. Then

$$w \in \ker T^* \iff T^*w = 0 \iff \langle v, T^*w\rangle = 0 \text{ for all } v \in V$$
$$\iff \langle Tv, w\rangle = 0 \text{ for all } v \in V \iff w \in (\operatorname{im} T)^\perp.$$

The second equivalence is nondegeneracy: forwards it is immediate, and backwards we take $v = T^*w$ and get $\|T^*w\|^2 = 0$. The third is Definition 11.6.1. The fourth is Definition 11.3.1 together with conjugate symmetry, since $\langle Tv, w\rangle = 0$ if and only if $\langle w, Tv\rangle = 0$, and the vectors Tv are exactly the elements of $\operatorname{im} T$. Every step is reversible, so both inclusions come at once.

(b) Apply (a) to $T^* \in \mathcal{L}(W, V)$ in place of T. Since $T^{**} = T$ by Theorem 11.6.2 (d), this reads $\ker T = (\operatorname{im} T^*)^\perp$. Taking orthogonal complements and using $(U^\perp)^\perp = U$ for a subspace U of a finite-dimensional space (Theorem 11.3.2 (b)),

$$(\ker T)^\perp = ((\operatorname{im} T^*)^\perp)^\perp = \operatorname{im} T^*.$$

(c) $\ker T$ is a subspace of V (Theorem 2.2.1), so $V = \ker T \oplus (\ker T)^\perp$ by Theorem 11.3.2 (a), and $(\ker T)^\perp = \operatorname{im} T^*$ by (b).

(d) Likewise $W = \operatorname{im} T \oplus (\operatorname{im} T)^\perp$ by Theorem 11.3.2 (a), and $(\operatorname{im} T)^\perp = \ker T^*$ by (a). This proves the theorem.

■

For a matrix $\mathbf{A} \in M_{m\times n}(F)$ with the standard inner products, $T^*_{\mathbf{A}} = T_{\mathbf{A}^*}$, so the theorem reads

$$\operatorname{nul}(\mathbf{A}^*) = \operatorname{col}(\mathbf{A})^\perp, \qquad \operatorname{col}(\mathbf{A}^*) = \operatorname{nul}(\mathbf{A})^\perp,$$

with $F^m = \operatorname{col}(\mathbf{A}) \oplus \operatorname{nul}(\mathbf{A}^*)$ and $F^n = \operatorname{nul}(\mathbf{A}) \oplus \operatorname{col}(\mathbf{A}^*)$. Over $\mathbb{R}$ the columns of $\mathbf{A}^\top$ are the rows of $\mathbf{A}$ turned into vectors, so the second identity is the promised statement that **the row space and the null space are orthogonal complements of each other** in $\mathbb{R}^n$. This is Theorem 5.6.5 with the annihilator replaced by the orthogonal complement, exactly as Corollary 11.5.2 licenses; and taking dimensions returns Rank–Nullity for matrices (Theorem 3.5.7) one more time. Section 4 proved the identity $\operatorname{nul}(\mathbf{A})^\perp = \operatorname{col}(\mathbf{A}^*)$ directly, because the minimum-norm solution needed it; the theorem above is that computation done once for all four subspaces and for maps rather than matrices.

Example. All Four Subspaces of a 3 × 3 Matrix

For $\mathbf{A} = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 2 & 1 \end{pmatrix}$ over $\mathbb{R}$, find bases of the four subspaces and verify the two orthogonality statements.

Solution. Row three is the sum of rows one and two, and rows one and two are independent, so $\operatorname{rank} \mathbf{A} = 2$ and $\operatorname{row}(\mathbf{A})$ has the basis $(1, 1, 0), (0, 1, 1)$ (written as vectors).

Null space. $\mathbf{Ax} = \mathbf{0}$ reduces to $x_1 + x_2 = 0$ and $x_2 + x_3 = 0$, so $\operatorname{nul}(\mathbf{A}) = \operatorname{span}((1, -1, 1))$, of dimension $1 = 3 - 2$, as Theorem 3.5.7 requires.

Orthogonality in the domain. $(1,1,0)\cdot(1,-1,1)=0$ and $(0,1,1)\cdot(1,-1,1)=0$. So $\operatorname{row}(\mathbf{A})\perp\operatorname{nul}(\mathbf{A})$, and since $2+1=3$ the two are complementary: $\operatorname{nul}(\mathbf{A})^{\perp}=\operatorname{row}(\mathbf{A})$.

Column space. Columns one and two are independent and column three is column two minus column one, so $\operatorname{col}(\mathbf{A})=\operatorname{span}\big((1,0,1),(1,1,2)\big)$.

Null space of the transpose. $\mathbf{A}^{\top}\mathbf{y}=\mathbf{0}$ reads $y_1+y_3=0$, $y_1+y_2+2y_3=0$, $y_2+y_3=0$, giving $\operatorname{nul}(\mathbf{A}^{\top})=\operatorname{span}\big((-1,-1,1)\big)$.

Orthogonality in the codomain. $(1,0,1)\cdot(-1,-1,1)=0$ and $(1,1,2)\cdot(-1,-1,1)=0$, and $2+1=3$, so $\operatorname{col}(\mathbf{A})^{\perp}=\operatorname{nul}(\mathbf{A}^{\top})$.

Corollary 11.6.6. Rank of the Adjoint

Let V, W be finite-dimensional inner product spaces over F and $T\in\mathcal{L}(V,W)$. Then

(a) $\operatorname{rank}T^{*}=\operatorname{rank}T$;

(b) $\ker(T^{*}T)=\ker T$, and consequently $\operatorname{rank}(T^{*}T)=\operatorname{rank}T$.

Proof.

(a) By Theorem 11.6.5 (b) and a dimension count,

$$\operatorname{rank}T^{*}=\dim\operatorname{im}T^{*}=\dim(\ker T)^{\perp}=\dim V-\operatorname{nullity}T=\operatorname{rank}T,$$

the middle equality by Theorem 11.3.2 (c) and the last by the Rank–Nullity Theorem (Theorem 2.3.1).

(b) ($\supseteq$) If $Tv=0$ then $T^{*}Tv=T^{*}0=0$. ($\subseteq$) Suppose $T^{*}Tv=0$. Then

$$\|Tv\|^{2}=\langle Tv,Tv\rangle=\langle v,T^{*}Tv\rangle=\langle v,0\rangle=0,$$

so $Tv=0$. Hence the kernels agree. Both $T^{*}T$ and T are defined on V, so Rank–Nullity gives $\operatorname{rank}(T^{*}T)=\dim V-\operatorname{nullity}(T^{*}T)=\dim V-\operatorname{nullity}T=\operatorname{rank}T$.

■

Part (a) is a fourth proof that row rank equals column rank, after the three of Chapters 3, 4, and 5: for a real matrix it says $\operatorname{rank}\mathbf{A}^{\top}=\operatorname{rank}\mathbf{A}$. Part (b) is the engine of the normal equations: it is why $\mathbf{A}^{*}\mathbf{A}$ is invertible exactly when $\mathbf{A}$ has independent columns, which is the hypothesis in Corollary 11.4.4.

Quick check. Let $T\in\mathcal{L}(V,W)$ with $\dim V=4$, $\dim W=7$, and $\operatorname{rank}T=3$. What are $\dim\ker T^{*}$ and $\dim\operatorname{im}T^{*}$?

Solution. $\operatorname{rank}T^{*}=3$ by Corollary 11.6.6 (a), so $\dim\operatorname{im}T^{*}=3$; and $T^{*}\colon W\to V$ with $\dim W=7$, so $\dim\ker T^{*}=7-3=4$. The same number comes from Theorem 11.6.5 (a): $\dim(\operatorname{im}T)^{\perp}=7-3=4$.

11.6.7 Self-adjoint and normal operators

For an operator $T \in \mathcal{L}(V)$, the adjoint lives in the same space $\mathcal{L}(V)$, so we may compare T with T^*. Two comparisons matter, and everything in the next chapter rests on them.

A self-adjoint operator is one that costs nothing to move across the inner product.

Definition 11.6.2. Self-adjoint Operator

An operator $T \in \mathcal{L}(V)$ on a finite-dimensional inner product space is **self-adjoint** if $T^* = T$, that is, if

$$\langle Tu, v\rangle = \langle u, Tv\rangle \qquad \text{for all } u, v \in V.$$

The matrix counterpart is the **Hermitian** matrix of Definition 11.1.2, $\mathbf{A}^* = \mathbf{A}$; over $\mathbb{R}$ this says $\mathbf{A}^\top = \mathbf{A}$, and the matrix is **symmetric**.

In words: T^* is not a new map but T over again, so T must be an **operator**, with the same space and the same inner product on both sides; and the displayed identity is the defining identity of Definition 11.6.1 with T^* rewritten as T, still required for **all** u and **all** v.

By Theorem 11.6.3, an operator on a finite-dimensional inner product space is self-adjoint exactly when its matrix in one (equivalently every) orthonormal basis is Hermitian. Three examples to keep in mind: the orthogonal projection P_U onto any subspace, self-adjoint by Example (Four Adjoints) (d), which is the geometric case; on $\mathbb{R}^2$, the map of the symmetric $\mathbf{A} = \begin{pmatrix} 2 & 1 \\ 1 & 3 \end{pmatrix}$, where conjugation does nothing and $\mathbf{A}^* = \mathbf{A}^\top = \mathbf{A}$; and on $\mathbb{C}^2$, the map of the Hermitian $\mathbf{A} = \begin{pmatrix} 2 & 1+i \\ 1-i & 3 \end{pmatrix}$, where transposing swaps the two off-diagonal entries and conjugating swaps them back, while the diagonal is real and survives both.

Here is a nonexample by minimal change: keep the shape and move the conjugate. The matrix $\mathbf{B} = \begin{pmatrix} 2 & 1+i \\ 1+i & 3 \end{pmatrix}$ is **symmetric**, $\mathbf{B}^\top = \mathbf{B}$, but $\mathbf{B}^* = \begin{pmatrix} 2 & 1-i \\ 1-i & 3 \end{pmatrix} \neq \mathbf{B}$, so the clause $T^* = T$ fails. It fails visibly at $\mathbf{u} = \mathbf{e}_1$ and $\mathbf{v} = \mathbf{e}_2$: $\langle \mathbf{B}\mathbf{e}_1, \mathbf{e}_2\rangle = 1 + i$, while $\langle \mathbf{e}_1, \mathbf{B}\mathbf{e}_2\rangle = \overline{1+i} = 1 - i$. Over $\mathbb{C}$, self-adjointness means Hermitian, never merely symmetric.

Two more self-adjoint operators come for free: the combination $D + D^*$ of Example (Differentiation on a Polynomial Space), and T^*T for any T, since $(T^*T)^* = T^*T^{**} = T^*T$ by Theorem 11.6.2 (c) and (d).

An operator is normal when it and its adjoint may be applied in either order.

Definition 11.6.3. Normal Operator

An operator $T \in \mathcal{L}(V)$ on a finite-dimensional inner product space is **normal** if it commutes with its adjoint:

$$T^*T = TT^*.$$

In words: normality is one equation between two operators on V, with no vectors and no quantifiers in it. It does not ask that T and T^* be equal, only that their order may be swapped, so it is strictly weaker than self-adjointness.

Four examples, widening as they go. Every self-adjoint operator is normal, since then both sides

are T^2. So is every invertible T with $T^{-1} = T^*$ — Section 7 calls these unitary — since then both sides are id_V; the quarter-turn of $\mathbb{R}^2$, with matrix $\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ in the standard basis, is one, and it is not self-adjoint. On $\mathbb{C}^2$ the operator with matrix $i\mathbf{I}$ satisfies $(i\mathbf{I})^* = -i\mathbf{I}$ by Theorem 11.6.2 (b), which commutes with $i\mathbf{I}$ but is not equal to it: normal and not self-adjoint. Widest of all, every diagonal matrix $\mathbf{D} = \mathrm{diag}(d_1, \dots, d_n)$ is normal, since $\mathbf{D}^*\mathbf{D} = \mathbf{D}\mathbf{D}^* = \mathrm{diag}(|d_1|^2, \dots, |d_n|^2)$.

A nonexample by minimal change: take the diagonal matrix $\mathrm{diag}(1, 0)$ and slide its nonzero entry off the diagonal. The Jordan block $\mathbf{J}_2(0) = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$ on $\mathbb{C}^2$ has

$$\mathbf{J}_2(0)^*\mathbf{J}_2(0) = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}, \qquad \mathbf{J}_2(0)\mathbf{J}_2(0)^* = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix},$$

and the two differ. The $(1,1)$ entries say what went wrong. Writing T for the map of $\mathbf{J}_2(0)$, the $(1,1)$ entry of the first matrix is $\langle T^*T\mathbf{e}_1, \mathbf{e}_1\rangle = \|T\mathbf{e}_1\|^2 = 0$ and that of the second is $\langle TT^*\mathbf{e}_1, \mathbf{e}_1\rangle = \|T^*\mathbf{e}_1\|^2 = 1$, so $\|T\mathbf{e}_1\| \neq \|T^*\mathbf{e}_1\|$: the operator kills $\mathbf{e}_1$ and its adjoint stretches it to a unit vector.

Over $\mathbb{C}$, normality is the weakest hypothesis under which the spectral theorem holds, and that is its reason for existing. Chapter 12 will show that the diagonal example is already the general one.

The facts available now, before any eigenvalue theory, are the ones that follow straight from the definition.

Proposition 11.6.7. First Properties of Self-adjoint Operators

Let $T \in \mathcal{L}(V)$ be self-adjoint on a finite-dimensional inner product space V over F.

(a) $\langle Tv, v\rangle \in \mathbb{R}$ for every $v \in V$, even when $F = \mathbb{C}$.

(b) If $\mathcal{B}$ is an orthonormal basis of V, the diagonal entries of $[T]_{\mathcal{B}}^{\mathcal{B}}$ are real.

(c) T is normal.

Proof.

(a) By conjugate symmetry (Definition 11.1.1) and self-adjointness,

$$\overline{\langle Tv, v\rangle} = \langle v, Tv\rangle = \langle Tv, v\rangle.$$

A complex number equal to its own conjugate is real.

(b) Let $\mathcal{B} = (v_1, \dots, v_n)$ be orthonormal. By Theorem 11.2.1 the i-th coordinate of Tv_i is $\langle Tv_i, v_i\rangle$, so the (i,i) entry of $[T]_{\mathcal{B}}^{\mathcal{B}}$ is $a_{ii} = \langle Tv_i, v_i\rangle$, which is real by (a).

(c) $T^*T = TT = TT^*$.

■

Part (a) is the first hint of the spectral theorem. If $Tv = \lambda v$ with $v \neq 0$, then $\langle Tv, v\rangle = \lambda\|v\|^2$, and since the left side is real and $\|v\|^2$ is a positive real, λ must be real: a self-adjoint operator has no nonreal eigenvalues. Chapter 12 turns this observation into the spectral theorems: a self-adjoint operator has an orthonormal basis of eigenvectors, with real eigenvalues; on a complex space the operators with an orthonormal basis of eigenvectors are exactly the normal ones, and

on a real space exactly the self-adjoint ones. None of that is proved here, and none of it is used before then; but all of it is built from the adjoint and the orthogonal geometry of this chapter.

Warning. **"Self-adjoint" cannot be read off an arbitrary basis.** Being self-adjoint is a property of T together with the inner product, and the Hermitian test applies only in an **orthonormal** basis. The operator $T(x_1, x_2) = (x_2, 0)$ of Example (The Transpose Rule Breaks) is not self-adjoint, and its matrix in the standard orthonormal basis, $\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$, is correctly nonsymmetric. But the self-adjoint operator $P(x_1, x_2) = (x_1, 0)$, the orthogonal projection onto the first axis, has the **non**-symmetric matrix $\begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}$ in the basis $\mathcal{B}$ of that example, since $P\mathbf{b}_1 = \mathbf{b}_1$ and $P\mathbf{b}_2 = (1, 0) = \mathbf{b}_1$. A nonsymmetric matrix is evidence of nothing until the basis is known to be orthonormal.

11.6.8 Exercises

11.6.8.1 A. Check your understanding

Exercise. A1

(a) State the defining property of the adjoint T^* of $T \in \mathcal{L}(V, W)$, including the domain and codomain of T^* and all quantifiers.

(b) Which hypothesis of Theorem 11.6.1 is used to produce T^*, and which theorem supplies it?

(c) True or false: $(cT)^* = cT^*$ for every scalar c. Justify your answer.

(d) State the hypothesis needed for $[T^*]_{\mathcal{C}}^{\mathcal{B}} = ([T]_{\mathcal{B}}^{\mathcal{C}})^*$, and say what goes wrong without it.

(e) True or false: $\ker T^* = (\ker T)^{\perp}$. Justify your answer.

(f) Give an example of a normal operator that is not self-adjoint.

Solution.

(a) For $T\colon V \to W$, the adjoint is the unique $T^*\colon W \to V$ with $\langle Tv, w\rangle = \langle v, T^*w\rangle$ for **all** $v \in V$ and **all** $w \in W$ (Definition 11.6.1).

(b) Finite-dimensionality of V, which is what Theorem 11.5.1 needs in order to represent the functional $v \mapsto \langle Tv, w\rangle$.

(c) False: $(cT)^* = \overline{c}\,T^*$ (Theorem 11.6.2 (b)). Over $\mathbb{R}$ the two agree; over $\mathbb{C}$ take $T = \mathrm{id}$ and $c = i$, so that $(i\ \mathrm{id})^* = -i\ \mathrm{id} \neq i\ \mathrm{id}$.

(d) Both bases must be orthonormal (Theorem 11.6.3). Without it the rule fails: Example (The Transpose Rule Breaks) has $([T]_{\mathcal{B}}^{\mathcal{B}})^{\top} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}$ but $[T^*]_{\mathcal{B}}^{\mathcal{B}} = \begin{pmatrix} -1 & -1 \\ 1 & 1 \end{pmatrix}$.

(e) False. The correct statements are $\ker T^* = (\operatorname{im} T)^{\perp}$ and $\operatorname{im} T^* = (\ker T)^{\perp}$ (Theorem 11.6.5). The two sides need not even live in the same space when $V \neq W$.

(f) $T = i\ \mathrm{id}$ on $\mathbb{C}^2$: $T^* = -i\ \mathrm{id}$ commutes with T, but $T^* \neq T$.

11.6.8.2 B. Practice

Exercise. B1: Adjoints of matrix maps

Let $\mathbf{A} = \begin{pmatrix} 1+i & 0 & 2 \\ i & 1 & -i \end{pmatrix}$ and $\mathbf{B} = \begin{pmatrix} 0 & 1 \\ i & 0 \end{pmatrix}$, acting between the complex spaces $\mathbb{C}^3 \xrightarrow{T_{\mathbf{A}}} \mathbb{C}^2 \xrightarrow{T_{\mathbf{B}}} \mathbb{C}^2$ with the standard inner products.

(a) Write down $T_{\mathbf{A}}^*$ and $T_{\mathbf{B}}^*$ as matrix maps.

(b) Compute the matrices of $T_{\mathbf{B}}T_{\mathbf{A}}$ and of $(T_{\mathbf{B}}T_{\mathbf{A}})^*$, and verify $(T_{\mathbf{B}}T_{\mathbf{A}})^* = T_{\mathbf{A}}^* T_{\mathbf{B}}^*$ by direct computation.

(c) Verify that $T_{\mathbf{B}}$ is not self-adjoint, and determine whether it is normal.

Solution.

(a) $T_{\mathbf{A}}^* = T_{\mathbf{A}^*}$ and $T_{\mathbf{B}}^* = T_{\mathbf{B}^*}$ (Example (Four Adjoints) (a)), with

$$\mathbf{A}^* = \begin{pmatrix} 1-i & -i \\ 0 & 1 \\ 2 & i \end{pmatrix}, \qquad \mathbf{B}^* = \begin{pmatrix} 0 & -i \\ 1 & 0 \end{pmatrix}.$$

(b) $\mathbf{BA} = \begin{pmatrix} i & 1 & -i \\ i-1 & 0 & 2i \end{pmatrix}$, so

$$(\mathbf{BA})^* = \begin{pmatrix} -i & -i-1 \\ 1 & 0 \\ i & -2i \end{pmatrix}.$$

On the other side,

$$\mathbf{A}^*\mathbf{B}^* = \begin{pmatrix} 1-i & -i \\ 0 & 1 \\ 2 & i \end{pmatrix} \begin{pmatrix} 0 & -i \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} -i & -i(1-i) \\ 1 & 0 \\ i & -2i \end{pmatrix},$$

and $-i(1-i) = -i + i^2 = -1 - i$. The two matrices agree, as Theorem 11.6.2 (c) predicts.

(c) $\mathbf{B}^* \neq \mathbf{B}$, since the $(1,2)$ entries are $-i$ and 1. But

$$\mathbf{B}^*\mathbf{B} = \begin{pmatrix} 0 & -i \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ i & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ i & 0 \end{pmatrix} \begin{pmatrix} 0 & -i \\ 1 & 0 \end{pmatrix} = \mathbf{B}\mathbf{B}^*,$$

so $T_{\mathbf{B}}$ is normal (Definition 11.6.3) and not self-adjoint.

Exercise. B2: Adjoints on a polynomial space

Let $V = \mathbb{R}[x]_{\leq 2}$ and $W = \mathbb{R}[x]_{\leq 3}$, each with the coefficient inner product of Example (Differentiation on a Polynomial Space), so that $\mathcal{E} = (1, x, x^2)$ and $\mathcal{F} = (1, x, x^2, x^3)$ are orthonormal bases.

(a) Find the adjoint of $M\colon V \to W$, $M(p) = xp$.

(b) Find the adjoint of $E\colon V \to \mathbb{R}$, $E(p) = p(1)$, where $\mathbb{R}$ carries $\langle s, t\rangle = st$. (Compare Theorem 11.5.1.)

(c) Verify $\operatorname{rank} M^* = \operatorname{rank} M$ directly.

Solution.

(a) $M(1) = x$, $M(x) = x^2$, $M(x^2) = x^3$, so in the two orthonormal bases

$$[M]_{\mathcal{E}}^{\mathcal{F}} = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \qquad [M^*]_{\mathcal{F}}^{\mathcal{E}} = \left([M]_{\mathcal{E}}^{\mathcal{F}}\right)^{\top} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

by Theorem 11.6.3. Hence $M^*(c_0 + c_1x + c_2x^2 + c_3x^3) = c_1 + c_2x + c_3x^2$: delete the constant term and divide by x.

(b) E is linear and $E(a_0{+}a_1x{+}a_2x^2) = a_0{+}a_1{+}a_2$. For $t \in \mathbb{R}$ we need $\langle p, E^*t\rangle = \langle Ep, t\rangle = (a_0{+}a_1{+}a_2)t$ for all p, so comparing coefficients of a_0, a_1, a_2 gives $E^*(t) = t(1{+}x{+}x^2)$. In particular $E^*(1) = 1 + x + x^2$ is the Riesz vector of the evaluation functional at 1.

(c) M is injective, since $xp = 0$ forces $p = 0$, so $\operatorname{rank} M = \dim V = 3$ by Theorem 2.3.1. And M^* is surjective: $M^*(xq) = q$ for every $q \in V$. So $\operatorname{rank} M^* = 3$ too, agreeing with Corollary 11.6.6 (a).

Exercise. B3: Locating the kernel of the adjoint

Let $\mathbf{A} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 0 & 1 \end{pmatrix} \in M_{3\times 2}(\mathbb{R})$, and let $T = T_{\mathbf{A}}\colon \mathbb{R}^2 \to \mathbb{R}^3$ with the standard inner products.

(a) Find a basis of $\operatorname{im} T$ and a basis of $\ker T^*$.

(b) Verify directly that every vector of $\ker T^*$ is orthogonal to every vector of $\operatorname{im} T$, and that the dimensions add to 3.

(c) Find $\ker T$ and $\operatorname{im} T^*$, and check Theorem 11.6.5 (b) in this case.

Solution.

(a) The two columns $(1, 1, 0)$ and $(0, 1, 1)$ are independent, so they are a basis of $\operatorname{im} T = \operatorname{col}(\mathbf{A})$ and $\operatorname{rank} T = 2$. Since $T^* = T_{\mathbf{A}^\top}$, the kernel of T^* is $\operatorname{nul}(\mathbf{A}^\top)$: the equations $y_1 + y_2 = 0$ and $y_2 + y_3 = 0$ give $\ker T^* = \operatorname{span}\big((1, -1, 1)\big)$.

(b) $(1, 1, 0)\cdot(1, -1, 1) = 1{-}1{+}0 = 0$ and $(0, 1, 1)\cdot(1, -1, 1) = 0{-}1{+}1 = 0$. Since these two vectors span $\operatorname{im} T$, orthogonality extends to all of $\operatorname{im} T$ by linearity. The dimensions are $2 + 1 = 3 = \dim \mathbb{R}^3$, as Theorem 11.6.5 (d) requires.

(c) $\mathbf{Ax} = \mathbf{0}$ gives $x_1 = 0$ and $x_2 = 0$, so $\ker T = \{\mathbf{0}\}$ and $(\ker T)^{\perp} = \mathbb{R}^2$. On the other side, $\operatorname{rank} T^* = \operatorname{rank} T = 2$ by Corollary 11.6.6 (a), and $\operatorname{im} T^* \subseteq \mathbb{R}^2$, so $\operatorname{im} T^* = \mathbb{R}^2$. The two agree.

11.6.8.3 C. Going deeper

Exercise. C1: The adjoint of integration

Let $V = \mathbb{R}[x]_{\le 1}$ and $W = \mathbb{R}[x]_{\le 2}$ with the coefficient inner products of Exercise (B2: Adjoints on a polynomial space), and let

$$J\colon V \to W, \qquad J(p)(x) \coloneqq \int_0^x p(t)\,\mathrm{d}t.$$

(a) Find J^*.

(b) Let $D\colon W \to V$ be differentiation. Show that $J^* \neq D$, although $DJ = \mathrm{id}_V$.

(c) Explain in one sentence why (b) does not contradict Theorem 11.6.2 (e).

Solution.

(a) $J(1) = x$ and $J(x) = \frac{1}{2}x^2$, so in the orthonormal bases $\mathcal{D} = (1, x)$ of V and $\mathcal{E} = (1, x, x^2)$ of W,

$$[J]_{\mathcal{D}}^{\mathcal{E}} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & \frac{1}{2} \end{pmatrix}, \qquad [J^*]_{\mathcal{E}}^{\mathcal{D}} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & \frac{1}{2} \end{pmatrix}$$

by Theorem 11.6.3. Hence $J^*(c_0 + c_1 x + c_2 x^2) = c_1 + \frac{1}{2}c_2 x$.

(b) $D(c_0 + c_1 x + c_2 x^2) = c_1 + 2c_2 x$, so J^* and D differ on x^2: $J^*(x^2) = \frac{1}{2}x$ while $D(x^2) = 2x$. Yet $D(J(p)) = p$ for every $p \in V$, by differentiating the integral, or by multiplying the two matrices: $[D]_{\mathcal{E}}^{\mathcal{D}}\,[J]_{\mathcal{D}}^{\mathcal{E}} = \mathbf{I}_2$.

(c) Part (e) of Theorem 11.6.2 applies to an **invertible** map, and J is not invertible: it is injective but not surjective, since $\dim V = 2 < 3 = \dim W$. A one-sided inverse carries no information about the adjoint.

Exercise. C2: Skew-adjoint operators

Let V be a finite-dimensional inner product space over F and let $T \in \mathcal{L}(V)$ satisfy $T^* = -T$ (such a T is called **skew-adjoint**).

(a) Prove that $\langle Tv, v\rangle$ is purely imaginary for every $v \in V$ when $F = \mathbb{C}$, and that $\langle Tv, v\rangle = 0$ for every $v \in V$ when $F = \mathbb{R}$.

(b) Prove that every skew-adjoint T is normal.

(c) Suppose $F = \mathbb{C}$. Prove that T is skew-adjoint if and only if iT is self-adjoint, and deduce that every $S \in \mathcal{L}(V)$ can be written $S = A + iB$ with A, B self-adjoint.

Solution.

(a) By conjugate symmetry and $T^* = -T$,

$$\overline{\langle Tv, v\rangle} = \langle v, Tv\rangle = \langle T^*v, v\rangle = -\langle Tv, v\rangle.$$

A complex number equal to minus its own conjugate has zero real part, so $\langle Tv, v\rangle$ is purely imaginary. When $F = \mathbb{R}$ the inner product is real, so the same equation reads

$\langle Tv, v\rangle = -\langle Tv, v\rangle$, forcing $\langle Tv, v\rangle = 0$.

(b) $T^*T = (-T)T = -T^2$ and $TT^* = T(-T) = -T^2$, so the two agree.

(c) By Theorem 11.6.2 (b), $(iT)^* = \bar{i}T^* = -iT^*$. If $T^* = -T$ then $(iT)^* = iT$, so iT is self-adjoint; conversely, if $(iT)^* = iT$ then $-iT^* = iT$, so $T^* = -T$. For the decomposition put

$$A := \tfrac{1}{2}(S + S^*), \qquad B := \tfrac{1}{2i}(S - S^*).$$

Then $A+iB = \frac{1}{2}(S+S^*)+\frac{1}{2}(S-S^*) = S$. Both are self-adjoint: $A^* = \frac{1}{2}(S^*+S) = A$ by Theorem 11.6.2 (a), (b), and (d), and $\frac{1}{2i}(S - S^*)$ is $-i$ times the skew-adjoint operator $\frac{1}{2}(S - S^*)$, hence self-adjoint by the first part of (c). This is the operator version of writing a complex number as $a + bi$.

Exercise. C3: The adjoint in an arbitrary basis

Let V be a finite-dimensional inner product space, let $\mathcal{B} = (b_1, \dots, b_n)$ be **any** ordered basis of V, let $\mathbf{G}$ be its Gram matrix, $g_{ij} = \langle b_j, b_i\rangle$, and let $T \in \mathcal{L}(V)$.

(a) Prove that $\langle u, v\rangle = [v]_{\mathcal{B}}^{*}\, \mathbf{G}\, [u]_{\mathcal{B}}$ for all $u, v \in V$.

(b) Deduce that $[T^*]_{\mathcal{B}}^{\mathcal{B}} = \mathbf{G}^{-1} \left([T]_{\mathcal{B}}^{\mathcal{B}}\right)^{*} \mathbf{G}$.

(c) Check that (b) returns Theorem 11.6.3 when $\mathcal{B}$ is orthonormal, and verify it on Example (The Transpose Rule Breaks).

Hint: for (b), write the defining identity of T^* in the coordinates of (a) and use that $\mathbf{x}^*\mathbf{M}\mathbf{y} = \mathbf{x}^*\mathbf{N}\mathbf{y}$ for all $\mathbf{x}, \mathbf{y}$ forces $\mathbf{M} = \mathbf{N}$.

Solution.

(a) Write $u = \sum_j x_j b_j$ and $v = \sum_i y_i b_i$, so $\mathbf{x} = [u]_{\mathcal{B}}$ and $\mathbf{y} = [v]_{\mathcal{B}}$. Expanding in the first slot and then in the second,

$$\langle u, v\rangle = \sum_{i,j} x_j \overline{y_i} \langle b_j, b_i\rangle = \sum_{i,j} \overline{y_i}\, g_{ij}\, x_j = \mathbf{y}^*\mathbf{G}\mathbf{x}.$$

(b) Put $\mathbf{A} = [T]_{\mathcal{B}}^{\mathcal{B}}$ and $\mathbf{B} = [T^*]_{\mathcal{B}}^{\mathcal{B}}$. By Theorem 4.1.1, $[Tu]_{\mathcal{B}} = \mathbf{A}\mathbf{x}$ and $[T^*v]_{\mathcal{B}} = \mathbf{B}\mathbf{y}$. So (a) turns $\langle Tu, v\rangle = \langle u, T^*v\rangle$ into

$$\mathbf{y}^*\mathbf{G}\mathbf{A}\mathbf{x} = (\mathbf{B}\mathbf{y})^*\mathbf{G}\mathbf{x} = \mathbf{y}^*\mathbf{B}^*\mathbf{G}\mathbf{x} \qquad \text{for all } \mathbf{x}, \mathbf{y} \in F^n.$$

Taking $\mathbf{x} = \mathbf{e}_j$ and $\mathbf{y} = \mathbf{e}_i$ reads off the (i, j) entries, so $\mathbf{G}\mathbf{A} = \mathbf{B}^*\mathbf{G}$. The Gram matrix is invertible (Theorem 11.2.5 (c)), so $\mathbf{B}^* = \mathbf{G}\mathbf{A}\mathbf{G}^{-1}$, and starring both sides gives $\mathbf{B} = (\mathbf{G}^{-1})^*\mathbf{A}^*\mathbf{G}^* = \mathbf{G}^{-1}\mathbf{A}^*\mathbf{G}$, using $\mathbf{G}^* = \mathbf{G}$ (Theorem 11.2.5 (a)) and hence $(\mathbf{G}^{-1})^* = \mathbf{G}^{-1}$.

(c) If $\mathcal{B}$ is orthonormal then $g_{ij} = \delta_{ij}$, so $\mathbf{G} = \mathbf{I}$ and the formula reads $\mathbf{B} = \mathbf{A}^*$, which is Theorem 11.6.3. In Example (The Transpose Rule Breaks) the computation in the remark after it is exactly this formula, with $\mathbf{G} = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}$.

11.7 Isometries and Unitary Maps

Chapter 2 asked which linear maps are invertible. Now that we can measure, we can ask a sharper question: which linear maps leave the measurements alone? A map that never changes a length never changes an angle either, and the two facts turn out to be the same fact. This section identifies these maps in five equivalent ways, collects them into the orthogonal and unitary groups, and classifies them completely in dimensions two and three, where they are exactly the rotations and reflections that the words suggest. At the end we drop linearity from the hypothesis and get it back as a conclusion.

As throughout the chapter, F is $\mathbb{R}$ or $\mathbb{C}$, inner products are linear in the first slot, and spaces are finite-dimensional unless stated otherwise.

11.7.1 Maps that do not change length

Take the plane with its usual inner product and ask which linear maps move the picture rigidly. Turning the plane about the origin clearly does not change any length. Flipping it across a line through the origin does not either. Scaling by 2 obviously does. Shearing, $(x_1, x_2) \mapsto (x_1 + x_2, x_2)$, also does: it fixes $\mathbf{e}_1$ but sends $\mathbf{e}_2$ to $(1, 1)$, of length $\sqrt{2}$. So "does not change length" is a genuine restriction, and it is worth a name.

An isometry is a linear map that every ruler agrees with: the image of a vector is exactly as long as the vector.

Definition 11.7.1. Isometry

Let V and W be inner product spaces over F. A linear map $T \in \mathcal{L}(V, W)$ is an **isometry** if

$$\|Tv\| = \|v\| \qquad \text{for every } v \in V.$$

In words: the condition is required for **every** vector, not merely for the vectors of a basis, and it compares a norm in W with a norm in V. The word comes from *iso* (equal) and *metron* (measure). Some authors reserve "isometry" for a bijective such map; we do not, and Definition 11.7.2 below names the bijective case separately.

Two consequences are immediate. An isometry is injective, since $Tv = 0$ forces $\|v\| = \|Tv\| = 0$ and hence $v = 0$ (Theorem 2.2.3). And if $Tv = \lambda v$ with $v \neq 0$, then $\|v\| = \|\lambda v\| = |\lambda|\,\|v\|$ by Theorem 11.1.2 (b), so $|\lambda| = 1$. We record the second observation, which the classification below will use twice.

Proposition 11.7.1. Eigenvalues of an Isometry Have Modulus 1

Let $T \in \mathcal{L}(V)$ be an isometry of an inner product space over F, and let $\lambda \in F$ be an eigenvalue of T. Then $|\lambda| = 1$. In particular, if $F = \mathbb{R}$ then $\lambda = 1$ or $\lambda = -1$.

Proof. Let $v \neq 0$ satisfy $Tv = \lambda v$ (Definition 9.2.1). Then $\|v\| = \|Tv\| = \|\lambda v\| = |\lambda|\,\|v\|$ by Theorem 11.1.2 (b), and $\|v\| \neq 0$, so dividing gives $|\lambda| = 1$. The only real numbers of modulus 1 are ± 1.

■

Example. Isometries and Near-Isometries

Decide which of the following linear maps are isometries, with the standard inner products.

(a) $T\colon \mathbb{R}^2 \to \mathbb{R}^3$, $T(x_1, x_2) = (x_1, x_2, 0)$.

(b) $T_{\mathbf{P}}\colon F^n \to F^n$ for a permutation matrix $\mathbf{P}$ (Definition 3.4.2).

(c) $T\colon \mathbb{C} \to \mathbb{C}$, $T(z) = \alpha z$, where $|\alpha| = 1$.

(d) $T\colon \mathbb{R}^2 \to \mathbb{R}^2$, $T(x_1, x_2) = (x_1 + x_2, x_2)$.

(e) The orthogonal projection P_U onto a subspace U with $\{0\} \neq U \neq V$.

Solution.

(a) Yes: $\|T\mathbf{x}\|^2 = x_1^2 + x_2^2 + 0 = \|\mathbf{x}\|^2$. Note that T is not surjective, so an isometry need not be invertible.

(b) Yes. A permutation matrix has exactly one 1 in each row and column, so $\mathbf{Px}$ is the list $x_1, \dots, x_n$ in a different order, and $\|\mathbf{Px}\|^2 = \sum_i |x_{\sigma(i)}|^2 = \sum_i |x_i|^2$, the same sum reordered.

(c) Yes: $\|\alpha z\| = |\alpha z| = |\alpha||z| = |z|$.

(d) No: $\|T\mathbf{e}_2\| = \|(1,1)\| = \sqrt{2} \neq 1 = \|\mathbf{e}_2\|$. One vector is enough to refute.

(e) No: pick $v \in U^\perp$ with $v \neq 0$, which exists because $U \neq V$ (Theorem 11.3.2 (a)). Then $P_U v = 0$ (Theorem 11.3.3 (a)), so $\|P_U v\| = 0 \neq \|v\|$. A projection destroys length by design.

11.7.2 Five ways to say the same thing

The definition mentions only norms, but norms and inner products determine each other. In one direction this is Definition 11.1.3. In the other it is the following identity, which reconstructs the inner product from lengths alone; it is the reason that preserving length automatically preserves angle.

Lemma 11.7.2. Polarization Identities

Let V be an inner product space over F and let $u, v \in V$.

(a) If $F = \mathbb{R}$, then $\langle u, v\rangle = \frac{1}{4}(\|u + v\|^2 - \|u - v\|^2)$.

(b) If $F = \mathbb{C}$, then $\langle u, v\rangle = \frac{1}{4}\sum_{k=0}^{3} i^k \, \|u + i^k v\|^2$.

Idea. Expand every norm squared with $\|u+w\|^2 = \|u\|^2 + 2\,\mathrm{Re}\langle u, w\rangle + \|w\|^2$ (Theorem 11.1.2 (c)). In the real case the two expansions differ only in the middle term, which survives doubled. In the complex case the four terms $\|u\|^2 + \|w\|^2$ are the same in all four summands and are killed by $\sum_k i^k = 0$; what is left is a sum in which the $\langle u, v\rangle$ parts reinforce and the $\overline{\langle u, v\rangle}$ parts cancel.

Proof.

(a) By Theorem 11.1.2 (c), $\|u \pm v\|^2 = \|u\|^2 \pm 2\langle u, v\rangle + \|v\|^2$, since over $\mathbb{R}$ the inner product is its own real part. Subtracting, $\|u+v\|^2 - \|u-v\|^2 = 4\langle u, v\rangle$.

(b) Write $z = \langle u, v\rangle$. Since $|i^k| = 1$, Theorem 11.1.2 (b) and (c) give $\|i^k v\| = \|v\|$ and

$$\|u + i^k v\|^2 = \|u\|^2 + \|v\|^2 + 2\,\mathrm{Re}\langle u, i^k v\rangle = \|u\|^2 + \|v\|^2 + 2\,\mathrm{Re}(i^{-k} z),$$

where $\langle u, i^k v\rangle = \overline{i^k}\, z = i^{-k} z$ by conjugate-linearity in the second slot. Multiply by i^k and sum over $k = 0, 1, 2, 3$. The first two terms contribute $(\|u\|^2 + \|v\|^2)\sum_k i^k = 0$, because $1 + i - 1 - i = 0$. For the remaining terms use $2\,\mathrm{Re}(w) = w + \overline{w}$:

$$\sum_{k=0}^{3} i^k \cdot 2\,\mathrm{Re}(i^{-k} z) = \sum_{k=0}^{3} i^k (i^{-k} z + i^k \overline{z}) = \sum_{k=0}^{3} z + \overline{z} \sum_{k=0}^{3} i^{2k} = 4z + 0,$$

since $\sum_k i^{2k} = 1 - 1 + 1 - 1 = 0$. Dividing by 4 gives the identity.

■

Now the main theorem. It says that length, angle, the adjoint, and orthonormal bases all detect the same class of maps, and it is the reason isometries are easy to work with: whichever of the five descriptions a problem hands you, you may switch to whichever one you want.

Theorem 11.7.3. Characterizations of Isometries

Let V and W be finite-dimensional inner product spaces over F, and let $T \in \mathcal{L}(V, W)$. The following are equivalent.

(a) T is an isometry: $\|Tv\| = \|v\|$ for every $v \in V$.

(b) T preserves inner products: $\langle Tu, Tv\rangle = \langle u, v\rangle$ for all $u, v \in V$.

(c) $T^*T = \mathrm{id}_V$.

(d) T sends **every** orthonormal list in V to an orthonormal list in W.

(e) T sends **some** orthonormal basis of V to an orthonormal list in W.

(f) For one (equivalently every) pair of orthonormal bases $\mathcal{B}$ of V and $\mathcal{C}$ of W, the matrix $\mathbf{A} = [T]_{\mathcal{B}}^{\mathcal{C}} \in M_{m\times n}(F)$, where $n = \dim V$ and $m = \dim W$, satisfies $\mathbf{A}^*\mathbf{A} = \mathbf{I}_n$; that is, the n columns of $\mathbf{A}$ are orthonormal vectors of F^m.

If in addition $\dim W = \dim V$, then (a)–(f) are further equivalent to each of

(g1) T is invertible with $T^{-1} = T^*$;

(g2) $TT^* = \mathrm{id}_W$.

Idea. Run the cycle (a) ⇒ (b) ⇒ (c) ⇒ (a). The first step is polarization: lengths are all we are given, and Lemma 11.7.2 converts them into an inner product. The second is the adjoint doing its job, $\langle Tu, Tv\rangle = \langle u, T^*Tv\rangle$, plus nondegeneracy. The third is the same identity read backwards with $u = v$. Then (b) ⇒ (d) ⇒ (e) ⇒ (b) is a second, shorter cycle, and (c) ⇔ (f) is Theorem 11.6.3. The last two need a dimension count: $T^*T = \mathrm{id}$ makes T injective, and in equal dimensions injective means invertible.

Proof.

(a) $\Rightarrow$ (b). Suppose $\|Tv\| = \|v\|$ for all v. If $F = \mathbb{R}$, then by Lemma 11.7.2 (a) applied in W and then in V, and by the linearity of T,

$$\begin{aligned}\langle Tu, Tv\rangle &= \tfrac{1}{4}(\|T(u+v)\|^2 - \|T(u-v)\|^2)\\ &= \tfrac{1}{4}(\|u+v\|^2 - \|u-v\|^2) = \langle u, v\rangle.\end{aligned}$$

If $F = \mathbb{C}$, the same two lines work with Lemma 11.7.2 (b), since $T(u+i^k v) = Tu + i^k Tv$.

(b) $\Rightarrow$ (c). For all $u, v \in V$, Definition 11.6.1 gives $\langle u, T^*Tv\rangle = \langle Tu, Tv\rangle = \langle u, v\rangle$, so $\langle u, T^*Tv - v\rangle = 0$ for every u. Taking $u = T^*Tv - v$ makes its norm zero, so $T^*Tv = v$ for every v.

(c) $\Rightarrow$ (a). $\|Tv\|^2 = \langle Tv, Tv\rangle = \langle v, T^*Tv\rangle = \langle v, v\rangle = \|v\|^2$, and norms are nonnegative, so $\|Tv\| = \|v\|$.

(d) $\Rightarrow$ (d). Let $(u_1, \dots, u_k)$ be orthonormal in V, so $\langle u_i, u_j\rangle = \delta_{ij}$. Then $\langle Tu_i, Tu_j\rangle = \langle u_i, u_j\rangle = \delta_{ij}$, which is exactly what it means for $(Tu_1, \dots, Tu_k)$ to be orthonormal (Definition 11.2.1).

(e) $\Rightarrow$ (e). V is finite-dimensional, so it has an orthonormal basis by Theorem 11.2.2, and (d) applies to it.

(f) $\Rightarrow$ (b). Let $\mathcal{B} = (v_1, \dots, v_n)$ be an orthonormal basis of V with $(Tv_1, \dots, Tv_n)$ orthonormal. Given $u, v \in V$, expand them in $\mathcal{B}$ as $u = \sum_i a_i v_i$ and $v = \sum_j b_j v_j$. Expanding both slots and using orthonormality twice,

$$\langle Tu, Tv\rangle = \sum_{i,j} a_i \overline{b_j}\langle Tv_i, Tv_j\rangle = \sum_i a_i \overline{b_i} = \sum_{i,j} a_i \overline{b_j}\langle v_i, v_j\rangle = \langle u, v\rangle,$$

since in both double sums only the terms with $i = j$ survive.

(g) $\Leftrightarrow$ (f). Fix orthonormal bases $\mathcal{B}$ and $\mathcal{C}$ and put $\mathbf{A} = [T]_{\mathcal{B}}^{\mathcal{C}}$. By Theorem 11.6.3, $[T^*]_{\mathcal{C}}^{\mathcal{B}} = \mathbf{A}^*$, and by Theorem 4.1.2, $[T^*T]_{\mathcal{B}}^{\mathcal{B}} = \mathbf{A}^*\mathbf{A}$. Since a map is determined by its matrix in fixed bases, $T^*T = \mathrm{id}_V$ if and only if $\mathbf{A}^*\mathbf{A} = \mathbf{I}_n$. The (i, j) entry of $\mathbf{A}^*\mathbf{A}$ is the standard inner product of column j with column i, so this says the columns are orthonormal. As (c) does not mention the bases, the condition holds for one pair of orthonormal bases if and only if it holds for every pair.

Now assume $\dim W = \dim V$.

(c) $\Rightarrow$ (g1). By (c) and Theorem 2.2.3, T is injective, so $\operatorname{rank} T = \dim V = \dim W$ by Theorem 2.3.1, and T is surjective. Hence T is invertible, and multiplying $T^*T = \mathrm{id}_V$ on the right by T^{-1} gives $T^* = T^{-1}$.

(g1) $\Rightarrow$ (g2). $TT^* = TT^{-1} = \mathrm{id}_W$.

(g2) $\Rightarrow$ (c). $TT^* = \mathrm{id}_W$ makes T surjective, hence invertible by the equal dimensions and Theorem 2.3.1, and multiplying on the left by T^{-1} gives $T^* = T^{-1}$, so $T^*T = \mathrm{id}_V$. This proves the theorem.

■

Condition (f) is the one to use on a concrete matrix: **check that the columns are orthonormal**. Condition (c) is the one to use in a proof. Condition (e) is the cheapest to verify on an abstract map, since one basis suffices, and (d) is what you get out afterwards.

Warning. $T^*T = \mathrm{id}$ **does not give** $TT^* = \mathrm{id}$ **when the dimensions differ.** Let $T\colon \mathbb{R}^2 \to \mathbb{R}^3$ be $T(x_1, x_2) = (x_1, x_2, 0)$, with matrix $\mathbf{A} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix}$. Then $\mathbf{A}^\top \mathbf{A} = \mathbf{I}_2$, so T is an isometry; but

$$\mathbf{A}\mathbf{A}^\top = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \neq \mathbf{I}_3.$$

Ask *identity on what?* The composite TT^* is the orthogonal projection of $\mathbb{R}^3$ onto $\operatorname{im} T$, the x_1x_2-plane, and it is the identity there and nowhere else. Only when T is surjective can both products be identities.

Quick check. Is $\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$ the matrix of an isometry of $\mathbb{R}^2$? If not, is some scalar multiple of it?

Solution. No: the first column has norm $\sqrt{2} \neq 1$, so the columns are not orthonormal and condition (f) of Theorem 11.7.3 fails. The columns are, however, orthogonal, so dividing by $\sqrt{2}$ normalizes both and $\frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$ is an isometry.

11.7.3 Unitary and orthogonal maps

Conditions (g1) and (g2) describe the case that matters most, an isometry of a space with itself. It is the case with a name.

Definition 11.7.2. Unitary and Orthogonal

Let V be a finite-dimensional inner product space over F. An operator $T \in \mathcal{L}(V)$ is **unitary** if

$$T^*T = TT^* = \mathrm{id}_V,$$

equivalently if T is an invertible isometry. When $F = \mathbb{R}$, such a T is also called **orthogonal**.

A matrix $\mathbf{A} \in M_n(\mathbb{C})$ with $\mathbf{A}^*\mathbf{A} = \mathbf{I}_n$ is a **unitary matrix**; a matrix $\mathbf{A} \in M_n(\mathbb{R})$ with $\mathbf{A}^\top\mathbf{A} = \mathbf{I}_n$ is an **orthogonal matrix**.

For a **square** matrix the one-sided condition $\mathbf{A}^*\mathbf{A} = \mathbf{I}$ already forces $\mathbf{A}\mathbf{A}^* = \mathbf{I}$, by part (g2) of Theorem 11.7.3, or directly by Theorem 3.4.5. The warning above shows that squareness is not a formality.

Definition 11.7.3. The Unitary and Orthogonal Groups

For $n \geq 1$ set

$$\mathrm{U}(n) \coloneqq \{\mathbf{A} \in M_n(\mathbb{C}) : \mathbf{A}^*\mathbf{A} = \mathbf{I}_n\}, \qquad \mathrm{O}(n) \coloneqq \{\mathbf{A} \in M_n(\mathbb{R}) : \mathbf{A}^\top\mathbf{A} = \mathbf{I}_n\},$$

and $\mathrm{SO}(n) \coloneqq \{\mathbf{A} \in \mathrm{O}(n) : \det \mathbf{A} = 1\}$, the **special orthogonal group**.

Proposition 11.7.4. The Groups and Their Determinants

(a) $\mathrm{U}(n)$ and $\mathrm{O}(n)$ are groups under matrix multiplication, and $\mathrm{O}(n) = \mathrm{U}(n) \cap M_n(\mathbb{R})$.

(b) If $\mathbf{A} \in \mathrm{U}(n)$ then $|\det \mathbf{A}| = 1$. If $\mathbf{A} \in \mathrm{O}(n)$ then $\det \mathbf{A} = 1$ or $\det \mathbf{A} = -1$.

(c) $\mathrm{SO}(n)$ is a subgroup of $\mathrm{O}(n)$.

Proof.

(a) $\mathbf{I}^*\mathbf{I} = \mathbf{I}$, so $\mathbf{I} \in \mathrm{U}(n)$. If $\mathbf{A}, \mathbf{B} \in \mathrm{U}(n)$ then $(\mathbf{AB})^*(\mathbf{AB}) = \mathbf{B}^*\mathbf{A}^*\mathbf{AB} = \mathbf{B}^*\mathbf{B} = \mathbf{I}$, so $\mathbf{AB} \in \mathrm{U}(n)$. Each $\mathbf{A} \in \mathrm{U}(n)$ is invertible with $\mathbf{A}^{-1} = \mathbf{A}^*$, and $(\mathbf{A}^*)^*\mathbf{A}^* = \mathbf{A}\mathbf{A}^* = \mathbf{I}$, so $\mathbf{A}^{-1} \in \mathrm{U}(n)$. Associativity is inherited from matrix multiplication. The same three lines prove the claim for $\mathrm{O}(n)$, with $\mathbf{A}^* = \mathbf{A}^\top$ for a real matrix, which is also the last assertion.

(b) The determinant is the sum $\sum_\sigma \operatorname{sgn}(\sigma) a_{\sigma(1)1} \cdots a_{\sigma(n)n}$ of Definition 7.4.1, and conjugation of complex numbers respects sums and products, so $\det \overline{\mathbf{A}} = \overline{\det \mathbf{A}}$. With Theorem 7.4.5 this gives $\det \mathbf{A}^* = \det(\overline{\mathbf{A}}^\top) = \det \overline{\mathbf{A}} = \overline{\det \mathbf{A}}$. Hence, by Theorem 7.5.4,

$$1 = \det \mathbf{I}_n = \det(\mathbf{A}^*\mathbf{A}) = \overline{\det \mathbf{A}}\, \det \mathbf{A} = |\det \mathbf{A}|^2.$$

For real $\mathbf{A}$ the determinant is real, and the real numbers of modulus 1 are ± 1.

(c) $\det \mathbf{I} = 1$, and det is multiplicative (Theorem 7.5.4), so $\det(\mathbf{AB}) = 1$ and $\det(\mathbf{A}^{-1}) = (\det \mathbf{A})^{-1} = 1$ whenever $\det \mathbf{A} = \det \mathbf{B} = 1$.

■

Warning. $|\lambda| = 1$ **does not mean** $\lambda = \pm 1$**, and** $\det \mathbf{A} = 1$ **does not mean "rotation".** The matrix $-\mathbf{I}_4 \in \mathrm{SO}(4)$ has determinant $(-1)^4 = 1$, but it fixes no nonzero vector, so it is not a rotation about an axis in any sense. Likewise $-\mathbf{I}_3 \in \mathrm{O}(3)$ has determinant -1 but is not a reflection in a plane: a reflection fixes a plane pointwise, and $-\mathbf{I}_3$ fixes only $\mathbf{0}$. The clean classification below is a feature of dimensions 2 and 3, not a general theorem.

11.7.4 The Frobenius norm and unitary factors

Section 1 made $M_{m\times n}(F)$ into an inner product space by declaring

$$\langle \mathbf{A}, \mathbf{B} \rangle = \operatorname{tr}(\mathbf{B}^*\mathbf{A}),$$

the **Frobenius inner product**, and observing that it is nothing but the standard inner product of F^{mn} after the entries are listed in a single column. Its induced norm (Definition 11.1.3) was named there but never given a symbol. This is the place to give it one, because the single property that makes it useful is a statement about unitary factors, and unitary factors have only just arrived.

For $\mathbf{A} \in M_{m\times n}(\mathbb{C})$, the **Frobenius norm** of $\mathbf{A}$ is

$$\|\mathbf{A}\|_F := \sqrt{\langle \mathbf{A}, \mathbf{A}\rangle} = \Big(\sum_{i,j} |a_{ij}|^2\Big)^{1/2}.$$

The subscript F is there because $\|\cdot\|$ with no subscript already means the norm of a *vector*, and a matrix will later be asked to be both. Two remarks on the formula. First, it treats $\mathbf{A}$ as a list of mn numbers and forgets that they are arranged in a rectangle: it is blind to the difference between $\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$ and $\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$. Second, it is computable with no eigenvalues, no row reduction, and no choices: add up mn squared moduli.

The trace form of the same quantity is the one that does the work.

Quick check. Why is $\|\mathbf{A}\|_F^2 = \operatorname{tr}(\mathbf{A}^*\mathbf{A})$?

Solution. The (j,j) entry of $\mathbf{A}^*\mathbf{A}$ is $\sum_i \overline{a_{ij}} a_{ij} = \sum_i |a_{ij}|^2$, the squared length of the j-th column. Summing over j sweeps every entry of $\mathbf{A}$ exactly once, so $\operatorname{tr}(\mathbf{A}^*\mathbf{A}) = \sum_{i,j} |a_{ij}|^2$. Equivalently, it is $\langle \mathbf{A}, \mathbf{A}\rangle$ read off the definition of the Frobenius inner product.

Now the property that makes this the right yardstick whenever orthonormal bases are in play: unitary factors are invisible to it.

Lemma 11.7.5. Unitary Invariance of the Frobenius Norm

Let $\mathbf{A} \in M_{m\times n}(\mathbb{C})$, let $\mathbf{U} \in \mathrm{U}(m)$ and let $\mathbf{V} \in \mathrm{U}(n)$. Then

$$\|\mathbf{UA}\|_F = \|\mathbf{A}\|_F = \|\mathbf{AV}\|_F.$$

In particular, if $\mathbf{A}$ is square and $\mathbf{U}$ is unitary then $\|\mathbf{U}^*\mathbf{AU}\|_F = \|\mathbf{A}\|_F$.

Proof. Using $\|\mathbf{B}\|_F^2 = \operatorname{tr}(\mathbf{B}^*\mathbf{B})$, then $(\mathbf{UA})^* = \mathbf{A}^*\mathbf{U}^*$ and $\mathbf{U}^*\mathbf{U} = \mathbf{I}_m$,

$$\|\mathbf{UA}\|_F^2 = \operatorname{tr}(\mathbf{A}^*\mathbf{U}^*\mathbf{UA}) = \operatorname{tr}(\mathbf{A}^*\mathbf{A}) = \|\mathbf{A}\|_F^2.$$

For the right-hand factor, Theorem 0.9.6 (3) lets the $\mathbf{V}$ travel around the trace, and $\mathbf{VV}^* = \mathbf{I}_n$:

$$\|\mathbf{AV}\|_F^2 = \operatorname{tr}(\mathbf{V}^*\mathbf{A}^*\mathbf{AV}) = \operatorname{tr}(\mathbf{A}^*\mathbf{AVV}^*) = \|\mathbf{A}\|_F^2.$$

Both norms are nonnegative reals, so equality of the squares gives equality of the norms. For the last statement, $\mathbf{U}^*$ is unitary too, since $(\mathbf{U}^*)^*\mathbf{U}^* = \mathbf{UU}^* = \mathbf{I}$; apply the first equality to the left factor $\mathbf{U}^*$ and the second to the right factor $\mathbf{U}$. ∎

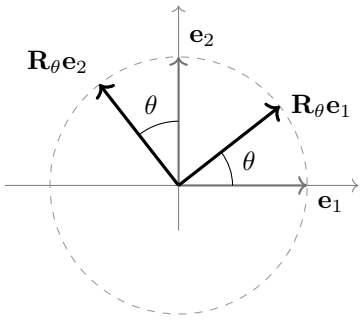
The consequence to keep in mind is this. If $T \in \mathcal{L}(V)$ is an operator on a finite-dimensional complex inner product space and we compute $\|[T]_{\mathcal{B}}^{\mathcal{B}}\|_F$ in an **orthonormal** basis $\mathcal{B}$, the answer does not depend on which orthonormal basis we chose, because two such matrices differ by a unitary similarity (Theorem 4.2.3). So the Frobenius norm is a property of T, not of our

coordinates — exactly like the trace and the determinant, and unlike, say, the largest entry.

11.7.5 Rotations and reflections of the plane

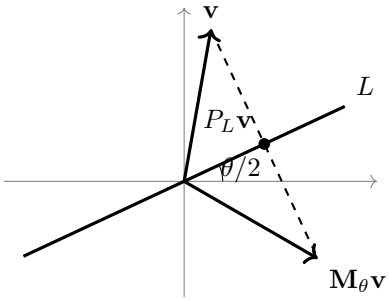
In dimension 2 the group $\mathrm{O}(2)$ can be written down completely, and the two determinant values split it into exactly the two families the words suggest. For $\theta \in \mathbb{R}$ put

$$\mathbf{R}_\theta := \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}, \qquad \mathbf{M}_\theta := \begin{pmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{pmatrix}.$$



the columns of $\mathbf{R}_\theta$ are the rotated axes

The first matrix rotates the plane by θ: its columns are $\mathbf{e}_1$ and $\mathbf{e}_2$ turned through θ. The second is a reflection, and the picture says in which mirror.



$\mathbf{M}_\theta$ reflects in the line L at angle $\theta/2$

Theorem 11.7.6. Classification of the Plane Isometries

Let $\mathbf{A} \in \mathrm{O}(2)$.

(a) If $\det \mathbf{A} = 1$, then $\mathbf{A} = \mathbf{R}_\theta$ for a unique $\theta \in [0, 2\pi)$.

(b) If $\det \mathbf{A} = -1$, then $\mathbf{A} = \mathbf{M}_\theta$ for a unique $\theta \in [0, 2\pi)$. In this case $\mathbf{A}^2 = \mathbf{I}$, the eigenvalues of $\mathbf{A}$ are 1 and -1, and $\mathbf{A}$ is the reflection in the line $L = \mathrm{span}\big((\cos\frac{\theta}{2}, \sin\frac{\theta}{2})\big)$: it fixes L pointwise and negates $L^\perp$.

Consequently $\mathrm{O}(2) = \{\mathbf{R}_\theta\} \cup \{\mathbf{M}_\theta\}$, and $\mathrm{SO}(2) = \{\mathbf{R}_\theta : \theta \in [0, 2\pi)\}$.

Idea. Condition (f) of Theorem 11.7.3 says the two columns are orthonormal. A unit vector of $\mathbb{R}^2$ is $(\cos\theta, \sin\theta)$ for exactly one $\theta \in [0, 2\pi)$, and there are exactly two unit vectors orthogonal to it. So there are only two candidates once the first column is chosen, and the determinant tells them apart. The geometric description of $\mathbf{M}_\theta$ is then a two-line verification:

guess the fixed vector from the picture and check it.

Proof. Write $\mathbf{A} = \begin{pmatrix} a & c \\ b & d \end{pmatrix} \in \mathrm{O}(2)$. By Theorem 11.7.3 (f), the columns are orthonormal: $a^2+b^2=1$, $c^2+d^2=1$, and $ac+bd=0$. Since (a,b) is a unit vector of $\mathbb{R}^2$, there is a unique $\theta \in [0, 2\pi)$ with $a = \cos\theta$ and $b = \sin\theta$. The vectors orthogonal to (a,b) form a line, spanned by $(-b,a)$, which is already a unit vector; so the unit vectors on that line are $\pm(-b,a)$, and $(c,d) = \pm(-\sin\theta, \cos\theta)$. The two choices give

$$\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} = \mathbf{R}_\theta \quad \text{and} \quad \begin{pmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{pmatrix} = \mathbf{M}_\theta,$$

with determinants $\cos^2\theta + \sin^2\theta = 1$ and $-\cos^2\theta - \sin^2\theta = -1$. This proves (a) and the first sentence of (b), and uniqueness of θ in both cases, since θ is already determined by the first column.

For the rest of (b), put $\mathbf{u} = (\cos\frac{\theta}{2}, \sin\frac{\theta}{2})$ and $\mathbf{w} = (-\sin\frac{\theta}{2}, \cos\frac{\theta}{2})$, two orthogonal unit vectors. Writing $\theta = \frac{\theta}{2} + \frac{\theta}{2}$ and using the subtraction formulas for cosine and sine,

$$\mathbf{M}_\theta \mathbf{u} = \begin{pmatrix} \cos\theta\cos\frac{\theta}{2} + \sin\theta\sin\frac{\theta}{2} \\ \sin\theta\cos\frac{\theta}{2} - \cos\theta\sin\frac{\theta}{2} \end{pmatrix} = \begin{pmatrix} \cos\frac{\theta}{2} \\ \sin\frac{\theta}{2} \end{pmatrix} = \mathbf{u},$$

$$\mathbf{M}_\theta \mathbf{w} = \begin{pmatrix} -\cos\theta\sin\frac{\theta}{2} + \sin\theta\cos\frac{\theta}{2} \\ -\sin\theta\sin\frac{\theta}{2} - \cos\theta\cos\frac{\theta}{2} \end{pmatrix} = \begin{pmatrix} \sin\frac{\theta}{2} \\ -\cos\frac{\theta}{2} \end{pmatrix} = -\mathbf{w}.$$

So $(\mathbf{u}, \mathbf{w})$ is an orthonormal basis of eigenvectors with eigenvalues 1 and -1, and $\mathbf{M}_\theta^2$ fixes both of them, hence $\mathbf{M}_\theta^2 = \mathbf{I}$. Since $L = \mathrm{span}(\mathbf{u})$, we have $L^\perp = \mathrm{span}(\mathbf{w})$, because $\mathbf{w} \perp \mathbf{u}$ and $\dim L^\perp = 1$ (Theorem 11.3.2 (c)), so $\mathbf{M}_\theta$ fixes L pointwise and negates $L^\perp$, which is the description of a reflection. This proves the theorem. ∎

So $\mathrm{SO}(2)$ is the circle of rotations, and the other half of $\mathrm{O}(2)$ is the circle of reflections. Multiplying two rotations adds the angles, $\mathbf{R}_\alpha \mathbf{R}_\beta = \mathbf{R}_{\alpha+\beta}$, so $\mathrm{SO}(2)$ is abelian; Exercise (C3: Two reflections make a rotation) shows what happens when two reflections are multiplied.

Example. Naming Two Plane Isometries

Classify each of the following as a rotation or a reflection, and give the angle or the mirror line.

(a) $\mathbf{A} = \frac{1}{2}\begin{pmatrix} 1 & -\sqrt{3} \\ \sqrt{3} & 1 \end{pmatrix}$.

(b) $\mathbf{B} = \frac{1}{5}\begin{pmatrix} 3 & 4 \\ 4 & -3 \end{pmatrix}$.

Solution.

(a) The columns $\frac{1}{2}(1,\sqrt{3})$ and $\frac{1}{2}(-\sqrt{3},1)$ have norm $\frac{1}{2}\sqrt{1+3} = 1$ and inner product $\frac{1}{4}(-\sqrt{3}+\sqrt{3}) = 0$, so $\mathbf{A} \in \mathrm{O}(2)$. Its determinant is $\frac{1}{4}(1+3) = 1$, so $\mathbf{A} = \mathbf{R}_\theta$ with $\cos\theta = \frac{1}{2}$ and $\sin\theta = \frac{\sqrt{3}}{2}$: rotation by $\theta = \pi/3$.

(b) The columns $\frac{1}{5}(3,4)$ and $\frac{1}{5}(4,-3)$ are orthonormal, and $\det \mathbf{B} = \frac{1}{25}(-9-16) = -1$, so $\mathbf{B} = \mathbf{M}_\theta$ with $\cos\theta = \frac{3}{5}$ and $\sin\theta = \frac{4}{5}$. The mirror is $L = \operatorname{span}\big((\cos\frac{\theta}{2}, \sin\frac{\theta}{2})\big)$, and rather than halve the angle we find the fixed line directly: $\mathbf{Bx} = \mathbf{x}$ reads $3x_1 + 4x_2 = 5x_1$, that is $x_1 = 2x_2$. So $L = \operatorname{span}\big((2,1)\big)$. Check: $\mathbf{B}(2,1) = \frac{1}{5}(6+4, 8-3) = (2,1)$, and $\mathbf{B}(-1,2) = \frac{1}{5}(-3+8, -4-6) = (1,-2) = -(-1,2)$, as a reflection must do to $L^\perp$.

11.7.6 Rotations of space

In dimension 3 the answer is the one physics assumes: every orientation-preserving isometry of $\mathbb{R}^3$ is a rotation about a line. The proof is where Chapter 9 is needed, for one reason only: to produce an eigenvector.

Theorem 11.7.7. Every Element of SO(3) Is a Rotation

Let $\mathbf{A} \in \mathrm{SO}(3)$. Then there is a unit vector $\mathbf{u} \in \mathbb{R}^3$ with $\mathbf{Au} = \mathbf{u}$, and an orthonormal basis $\mathcal{B} = (\mathbf{u}, \mathbf{w}_1, \mathbf{w}_2)$ of $\mathbb{R}^3$ in which

$$[T_{\mathbf{A}}]_{\mathcal{B}}^{\mathcal{B}} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta \\ 0 & \sin\theta & \cos\theta \end{pmatrix}$$

for some θ. That is, $\mathbf{A}$ fixes the **axis** $\operatorname{span}(\mathbf{u})$ and rotates the plane $\operatorname{span}(\mathbf{u})^\perp$ by θ. Moreover $\operatorname{tr}\mathbf{A} = 1 + 2\cos\theta$.

Idea. Three steps. ① Get a real eigenvalue: the characteristic polynomial is real of degree 3, and an odd-degree real polynomial has a real root. ② That eigenvalue is ± 1, since $T_{\mathbf{A}}$ is an isometry. ③ Split off the eigenline: its orthogonal complement is invariant, the restriction there lies in $\mathrm{O}(2)$, and Theorem 11.7.6 finishes. If the eigenvalue was -1 the restriction has determinant -1, so it is a reflection and supplies a $+1$ eigenvector inside the plane; feed that back into step ③.

Proof. Write $T = T_{\mathbf{A}}$, an isometry of $\mathbb{R}^3$ by Theorem 11.7.3.

Step 1: a real eigenvalue exists. By Theorem 7.10.2, $p_{\mathbf{A}} \in \mathbb{R}[x]$ is monic of degree 3, which is odd, so $p_{\mathbf{A}}$ has a real root λ by Corollary 6.5.9, and λ is an eigenvalue of $\mathbf{A}$ by Theorem 9.2.1. By Proposition 11.7.1, $\lambda = 1$ or $\lambda = -1$. Fix a unit eigenvector $\mathbf{u}_0$ for λ, possible after dividing by its norm.

Step 2: the orthogonal complement of an invariant line is invariant.

Claim. If $U \subseteq \mathbb{R}^3$ is a subspace with $T(U) \subseteq U$, then $T(U^\perp) \subseteq U^\perp$.

Proof. $T|_U : U \to U$ is injective, because T is, so it is surjective by Theorem 2.5.7 applied on the finite-dimensional space U. Let $\mathbf{w} \in U^\perp$ and $\mathbf{u} \in U$. Write $\mathbf{u} = T\mathbf{u}'$ with $\mathbf{u}' \in U$. Then $\langle T\mathbf{w}, \mathbf{u}\rangle = \langle T\mathbf{w}, T\mathbf{u}'\rangle = \langle \mathbf{w}, \mathbf{u}'\rangle = 0$, the middle step by Theorem 11.7.3 (b) and the

last because $\mathbf{u}' \in U$. So $T\mathbf{w} \in U^{\perp}$.

■

Step 3: reduce to the plane. Put $U = \operatorname{span}(\mathbf{u}_0)$, which is T-invariant. By the claim, $P := U^{\perp}$ is T-invariant, and $\dim P = 2$ by Theorem 11.3.2 (c). Choose an orthonormal basis $(\mathbf{w}_1, \mathbf{w}_2)$ of P (Theorem 11.2.2); then $\mathcal{B}_0 = (\mathbf{u}_0, \mathbf{w}_1, \mathbf{w}_2)$ is an orthonormal basis of $\mathbb{R}^3$, and in it

$$[T]_{\mathcal{B}_0}^{\mathcal{B}_0} = \begin{pmatrix} \lambda & 0 & 0 \\ 0 & c_{11} & c_{12} \\ 0 & c_{21} & c_{22} \end{pmatrix}, \qquad \mathbf{C} := \begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{pmatrix},$$

the first column because $T\mathbf{u}_0 = \lambda \mathbf{u}_0$, and the two zeros in the first row because $T\mathbf{w}_1, T\mathbf{w}_2 \in P$ have no $\mathbf{u}_0$-component. This matrix is orthogonal, being the matrix of an isometry in orthonormal bases (Theorem 11.7.3 (f)), so its second and third columns are orthonormal, which says $\mathbf{C}^{\top}\mathbf{C} = \mathbf{I}_2$ and $\mathbf{C} \in \mathrm{O}(2)$; and $\mathbf{C}$ is the matrix of $T|_P$ in the basis $(\mathbf{w}_1, \mathbf{w}_2)$. Expanding the determinant along the first column, and using that similar matrices have equal determinants (Corollary 7.5.7, with Theorem 4.2.4 (a)), $1 = \det \mathbf{A} = \lambda \det \mathbf{C}$, so $\det \mathbf{C} = \lambda$.

Case $\lambda = 1$. Then $\det \mathbf{C} = 1$, so $\mathbf{C} = \mathbf{R}_{\theta}$ for some θ by Theorem 11.7.6 (a), and $\mathcal{B} = \mathcal{B}_0$ and $\mathbf{u} = \mathbf{u}_0$ are as required.

Case $\lambda = -1$. Then $\det \mathbf{C} = -1$, so by Theorem 11.7.6 (b) the restriction $T|_P$ is a reflection of the plane P and has 1 as an eigenvalue. Let $\mathbf{u} \in P$ be a unit eigenvector for 1, so $T\mathbf{u} = \mathbf{u}$. Now rerun Steps 2 and 3 with $\mathbf{u}$ in place of $\mathbf{u}_0$. Because $\mathbf{u}$ was chosen with $T\mathbf{u} = \mathbf{u}$, the eigenvalue this time is 1, so the rerun lands in the previous case and cannot return here.

Finally, $\mathbf{A}$ and $[T]_{\mathcal{B}}^{\mathcal{B}}$ are matrices of the same operator, hence similar (Theorem 4.2.4 (a)), and similar matrices have equal traces (Theorem 4.2.5), so $\operatorname{tr} \mathbf{A} = 1 + 2\cos\theta$. This proves the theorem.

■

The angle is determined by $\mathbf{A}$ up to sign, since reversing the order of $\mathbf{w}_1, \mathbf{w}_2$ replaces θ by $-\theta$; the trace formula $\cos\theta = \frac{1}{2}(\operatorname{tr} \mathbf{A} - 1)$ gives $|\theta|$ with no computation at all. This settles $\mathrm{O}(3)$ as well: if $\det \mathbf{A} = -1$ then $-\mathbf{A} \in \mathrm{SO}(3)$, so $\mathbf{A}$ is a rotation composed with the antipodal map.

Example. Axis and Angle of a Rotation

Let $\mathbf{A} = \frac{1}{3}\begin{pmatrix} 2 & -2 & 1 \\ 1 & 2 & 2 \\ -2 & -1 & 2 \end{pmatrix}$. Verify that $\mathbf{A} \in \mathrm{SO}(3)$ and find its axis and angle.

Solution. Orthogonality. The three columns are $\frac{1}{3}(2,1,-2)$, $\frac{1}{3}(-2,2,-1)$ and $\frac{1}{3}(1,2,2)$. Their squared norms are $\frac{1}{9}(4+1+4) = 1$, $\frac{1}{9}(4+4+1) = 1$ and $\frac{1}{9}(1+4+4) = 1$, and their pairwise inner products are $\frac{1}{9}(-4+2+2) = 0$, $\frac{1}{9}(2+2-4) = 0$ and $\frac{1}{9}(-2+4-2) = 0$. So the columns are orthonormal and $\mathbf{A} \in \mathrm{O}(3)$ by Theorem 11.7.3 (f).

Determinant. Expanding along the first row, $27 \det \mathbf{A} = 2(4+2) - (-2)(2+4) + 1(-1+4) = 12 + 12 + 3 = 27$, so $\det \mathbf{A} = 1$ and $\mathbf{A} \in \mathrm{SO}(3)$.

Axis. Solve $\mathbf{Au} = \mathbf{u}$, that is $(3\mathbf{I} - 3\mathbf{A})\mathbf{u} = \mathbf{0}$:

$$\begin{pmatrix} 1 & 2 & -1 \\ -1 & 1 & -2 \\ 2 & 1 & 1 \end{pmatrix} \mathbf{u} = \mathbf{0}.$$

Adding row one to row two gives $3u_2 - 3u_3 = 0$, so $u_2 = u_3$; then row one gives $u_1 = -u_2$. The axis is $\operatorname{span}((-1,1,1))$, and indeed $\mathbf{A}(-1,1,1) = \frac{1}{3}(-2-2+1, -1+2+2, 2-1+2) = (-1,1,1)$.

Angle. $\operatorname{tr}\mathbf{A} = \frac{1}{3}(2+2+2) = 2$, so $\cos\theta = \frac{1}{2}(2-1) = \frac{1}{2}$ and $|\theta| = \pi/3$. The map is a rotation by 60° about the line through $(-1,1,1)$.

Quick check. A matrix $\mathbf{A} \in \mathrm{SO}(3)$ has $\operatorname{tr}\mathbf{A} = -1$. What is its angle, and what is $\mathbf{A}^2$?

Solution. $\cos\theta = \frac{1}{2}(-1-1) = -1$, so $\theta = \pi$: a half-turn about the axis. Then $\mathbf{A}^2$ is the rotation by 2π about the same axis, which is $\mathbf{I}_3$. (In the basis of Theorem 11.7.7 the matrix is $\operatorname{diag}(1,-1,-1)$, whose square is $\mathbf{I}_3$.)

11.7.7 Rigid motions

We assumed linearity from the start. It is worth knowing that we did not have to: for maps of $\mathbb{R}^n$ that fix the origin, preserving distance forces linearity.

Theorem 11.7.8. Distance-Preserving Maps Are Affine

Let $f\colon \mathbb{R}^n \to \mathbb{R}^n$ be any function, not assumed linear, with

$$\|f(\mathbf{x}) - f(\mathbf{y})\| = \|\mathbf{x} - \mathbf{y}\| \qquad \text{for all } \mathbf{x}, \mathbf{y} \in \mathbb{R}^n.$$

(a) If $f(\mathbf{0}) = \mathbf{0}$, then f is linear and orthogonal.

(b) In general, $f(\mathbf{x}) = \mathbf{Ax} + \mathbf{b}$ for a unique $\mathbf{A} \in \mathrm{O}(n)$ and a unique $\mathbf{b} \in \mathbb{R}^n$.

Idea. Setting $\mathbf{y} = \mathbf{0}$ turns the hypothesis into "preserves norms", and expanding the squared distance then turns it into "preserves inner products" without any polarization at all. Once inner products are preserved, f sends the standard basis to an orthonormal basis, and Theorem 11.2.1 writes $f(\mathbf{x})$ as an explicit linear expression in the coordinates of $\mathbf{x}$. Part (b) is part (a) applied to $\mathbf{x} \mapsto f(\mathbf{x}) - f(\mathbf{0})$.

Proof.

(a) Assume $f(\mathbf{0}) = \mathbf{0}$. Taking $\mathbf{y} = \mathbf{0}$ gives $\|f(\mathbf{x})\| = \|\mathbf{x}\|$ for all $\mathbf{x}$. Expanding both sides of the hypothesis with Theorem 11.1.2 (c),

$$\|f(\mathbf{x})\|^2 - 2\langle f(\mathbf{x}), f(\mathbf{y})\rangle + \|f(\mathbf{y})\|^2 = \|\mathbf{x}\|^2 - 2\langle \mathbf{x}, \mathbf{y}\rangle + \|\mathbf{y}\|^2,$$

and the four outer terms cancel in pairs, leaving

$$\langle f(\mathbf{x}), f(\mathbf{y})\rangle = \langle \mathbf{x}, \mathbf{y}\rangle \qquad \text{for all } \mathbf{x}, \mathbf{y}. \tag{$*$}$$

Put $\mathbf{f}_i := f(\mathbf{e}_i)$. By $(*)$, $\langle \mathbf{f}_i, \mathbf{f}_j\rangle = \langle \mathbf{e}_i, \mathbf{e}_j\rangle = \delta_{ij}$, so $(\mathbf{f}_1, \dots, \mathbf{f}_n)$ is an orthonormal list of n vectors in $\mathbb{R}^n$, hence an orthonormal basis (it is independent by Theorem 11.1.4 and has $n = \dim \mathbb{R}^n$ elements). Now let $\mathbf{x} = (x_1, \dots, x_n)$ be arbitrary. By Theorem 11.2.1 applied to the basis $(\mathbf{f}_1, \dots, \mathbf{f}_n)$, and then by $(*)$,

$$f(\mathbf{x}) = \sum_{i=1}^{n} \langle f(\mathbf{x}), \mathbf{f}_i\rangle\, \mathbf{f}_i = \sum_{i=1}^{n} \langle f(\mathbf{x}), f(\mathbf{e}_i)\rangle\, \mathbf{f}_i = \sum_{i=1}^{n} \langle \mathbf{x}, \mathbf{e}_i\rangle\, \mathbf{f}_i = \sum_{i=1}^{n} x_i\, \mathbf{f}_i.$$

The right-hand side is linear in $\mathbf{x}$, so f is linear; and it preserves norms, so it is an isometry, hence orthogonal by Definition 11.7.2, since it maps $\mathbb{R}^n$ to $\mathbb{R}^n$.

(b) Put $\mathbf{b} := f(\mathbf{0})$ and $g(\mathbf{x}) := f(\mathbf{x}) - \mathbf{b}$. Then g preserves distances, since the translation cancels, and $g(\mathbf{0}) = \mathbf{0}$; by (a), g is orthogonal, say $g(\mathbf{x}) = \mathbf{A}\mathbf{x}$ with $\mathbf{A} \in \mathrm{O}(n)$, and $f(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{b}$. For uniqueness, evaluating at $\mathbf{0}$ recovers $\mathbf{b}$, and then $\mathbf{A}\mathbf{x} = f(\mathbf{x}) - \mathbf{b}$ determines $\mathbf{A}$. This proves the theorem.

■

So the rigid motions of $\mathbb{R}^n$ are exactly the maps "rotate or reflect, then translate", and the linear ones are exactly the orthogonal maps. That is the sense in which $\mathrm{O}(n)$ is the symmetry group of Euclidean geometry, and it is why the same group reappears whenever a problem has a distance in it.

11.7.8 Exercises

11.7.8.1 A. Check your understanding

Exercise. A1

(a) Define an isometry $T \in \mathcal{L}(V, W)$, and state two conditions equivalent to it.

(b) What must be checked about a square matrix $\mathbf{A}$ to decide whether it is unitary? State the check in terms of its columns.

(c) True or false: every isometry is invertible. Justify your answer.

(d) True or false: if $\det \mathbf{A} = 1$ for $\mathbf{A} \in \mathrm{O}(n)$, then $\mathbf{A}$ fixes a nonzero vector. Justify your answer.

(e) Give the two families of matrices making up $\mathrm{O}(2)$, and say how the determinant distinguishes them.

(f) An operator T on a complex inner product space is unitary. Why must every eigenvalue of T lie on the unit circle?

Solution.

(a) $\|Tv\| = \|v\|$ for every $v \in V$ (Definition 11.7.1). Equivalent: $\langle Tu, Tv\rangle = \langle u, v\rangle$ for all u, v; and $T^*T = \mathrm{id}_V$ (Theorem 11.7.3).

(b) That its columns are orthonormal in F^n, that is $\mathbf{A}^*\mathbf{A} = \mathbf{I}_n$ (Theorem 11.7.3 (f)). For a square matrix this already gives $\mathbf{A}\mathbf{A}^* = \mathbf{I}_n$.

(c) False. $T\colon \mathbb{R}^2 \to \mathbb{R}^3$, $T(x_1, x_2) = (x_1, x_2, 0)$, is an isometry and is not surjective (Example (Isometries and Near-Isometries) (a)). An isometry is always injective, and it is invertible exactly when $\dim W = \dim V$.

(d) False for even n: $-\mathbf{I}_4 \in \mathrm{SO}(4)$ has determinant 1 and fixes only $\mathbf{0}$. It is true for odd n, though not by the argument of Theorem 11.7.7, which uses the $\mathrm{O}(2)$ classification of a 2-dimensional complement. The general reason is a count: the nonreal eigenvalues of a real $\mathbf{A}$ come in conjugate pairs $\lambda, \overline{\lambda}$ with $\lambda\overline{\lambda} = |\lambda|^2 = 1$ for an isometry, so they contribute 1 to the determinant and an even number to the count. With n odd, an odd number of eigenvalues are real, each ± 1, and their product is $\det \mathbf{A} = 1$; an odd number of factors ± 1 with product 1 cannot all be -1, so 1 is an eigenvalue and some nonzero vector is fixed.

(e) The rotations $\mathbf{R}_\theta$, with determinant 1, and the reflections $\mathbf{M}_\theta$, with determinant -1 (Theorem 11.7.6).

(f) A unitary operator is an isometry, so Proposition 11.7.1 applies: $|\lambda| = 1$.

11.7.8.2 B. Practice

Exercise. B1: Determine which are unitary

Determine which of the following matrices are orthogonal or unitary. Justify your answer; for those that are not, name the failing condition.

(a) $\frac{1}{3}\begin{pmatrix} 2 & -2 & 1 \\ 1 & 2 & 2 \\ -2 & -1 & 2 \end{pmatrix}$ over $\mathbb{R}$.

(b) $\frac{1}{\sqrt{2}}\begin{pmatrix} 1 & i \\ i & 1 \end{pmatrix}$ over $\mathbb{C}$.

(c) $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ over $\mathbb{R}$.

(d) $\frac{1}{5}\begin{pmatrix} 3 & 4 \\ 4 & 3 \end{pmatrix}$ over $\mathbb{R}$.

Solution.

(a) Orthogonal, as computed in Example (Axis and Angle of a Rotation): the three columns are orthonormal.

(b) Unitary. The columns are $\frac{1}{\sqrt{2}}(1, i)$ and $\frac{1}{\sqrt{2}}(i, 1)$, of norm $\frac{1}{\sqrt{2}}\sqrt{1+1} = 1$, and their inner product is

$$\tfrac{1}{2}(1 \cdot \overline{i} + i \cdot \overline{1}) = \tfrac{1}{2}(-i + i) = 0.$$

So $\mathbf{A}^*\mathbf{A} = \mathbf{I}_2$. (It is unitary but not orthogonal, since it is not real.)

(c) Not orthogonal. The second column $(1, 1)$ has norm $\sqrt{2} \neq 1$, so the columns are not orthonormal. Equivalently, the map stretches $\mathbf{e}_2$.

(d) Not orthogonal. Both columns are unit vectors, $\frac{1}{5}(3, 4)$ and $\frac{1}{5}(4, 3)$, but their inner product is $\frac{1}{25}(12 + 12) = \frac{24}{25} \neq 0$. Unit columns are not enough; they must also be orthogonal. (Its determinant is $\frac{1}{25}(9 - 16) = -\frac{7}{25}$, neither 1 nor -1, which confirms it by Proposition 11.7.4.)

Exercise. B2: Name the plane isometry

Each of the following lies in $\mathrm{O}(2)$. Decide whether it is a rotation or a reflection, and give the angle of rotation or the mirror line.

(a) $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$.

(b) $\frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}$.

(c) $\frac{1}{13}\begin{pmatrix} 5 & 12 \\ 12 & -5 \end{pmatrix}$.

Solution.

(a) Determinant -1, so a reflection. $\mathbf{Ax} = \mathbf{x}$ reads $x_2 = x_1$, so the mirror is $L = \operatorname{span}((1,1))$, the line $x_2 = x_1$. Check: $\mathbf{A}(1,-1) = (-1,1) = -(1,-1)$.

(b) Determinant $\frac{1}{2}(1+1) = 1$, so a rotation $\mathbf{R}_\theta$. Comparing with $\mathbf{R}_\theta$, $\cos\theta = \frac{1}{\sqrt{2}}$ and $\sin\theta = -\frac{1}{\sqrt{2}}$, so $\theta = -\pi/4$: a clockwise half of a right angle.

(c) Determinant $\frac{1}{169}(-25-144) = -1$, so a reflection. $\mathbf{Ax} = \mathbf{x}$ reads $5x_1 + 12x_2 = 13x_1$, that is $3x_2 = 2x_1$, so $L = \operatorname{span}((3,2))$. Check: $\mathbf{A}(3,2) = \frac{1}{13}(15+24, 36-10) = (3,2)$.

Exercise. B3: Permutations are orthogonal

Let $\sigma \in S_n$ and let $\mathbf{P}_\sigma \in M_n(\mathbb{R})$ be its permutation matrix (Definition 3.4.2), whose j-th column is $\mathbf{e}_{\sigma(j)}$.

(a) Prove that $\mathbf{P}_\sigma \in \mathrm{O}(n)$.

(b) Deduce that $\det \mathbf{P}_\sigma = \pm 1$, and identify the sign.

(c) Show that $\mathbf{P}_\sigma^{-1} = \mathbf{P}_\sigma^{\top} = \mathbf{P}_{\sigma^{-1}}$.

Solution.

(a) Column j of $\mathbf{P}_\sigma$ is $\mathbf{e}_{\sigma(j)}$, by Definition 3.4.2. So each column is a unit vector, and for $i \neq j$ the columns $\mathbf{e}_{\sigma(i)}$ and $\mathbf{e}_{\sigma(j)}$ are distinct standard basis vectors, because σ is injective, hence orthogonal. By Theorem 11.7.3 (f), $\mathbf{P}_\sigma \in \mathrm{O}(n)$.

(b) By Proposition 11.7.4 (b), $\det \mathbf{P}_\sigma = \pm 1$. The sign is $\operatorname{sgn}(\sigma)$: permuting the columns of $\mathbf{I}_n$ by σ multiplies the determinant by $\operatorname{sgn}(\sigma)$ (Theorem 7.2.1), and $\det \mathbf{I}_n = 1$.

(c) $\mathbf{P}_\sigma^{-1} = \mathbf{P}_\sigma^{\top}$ holds because $\mathbf{P}_\sigma$ is orthogonal, by (a). For the second equality, $(\mathbf{P}_\sigma)_{ij} = 1$ exactly when $i = \sigma(j)$, so $(\mathbf{P}_\sigma^{\top})_{ij} = (\mathbf{P}_\sigma)_{ji} = 1$ exactly when $j = \sigma(i)$, that is when $i = \sigma^{-1}(j)$; and that is the defining condition for $(\mathbf{P}_{\sigma^{-1}})_{ij} = 1$.

11.7.8.3 C. Going deeper

Exercise. C1: Determinant as a homomorphism

(a) Prove that $\det\colon \mathrm{O}(n) \to \{1,-1\}$ is a surjective group homomorphism with kernel $\mathrm{SO}(n)$, and deduce that exactly half of $\mathrm{O}(n)$ has determinant -1, in the sense that $\mathrm{O}(n)$ is the disjoint union of $\mathrm{SO}(n)$ and one coset $\mathbf{A}_0\,\mathrm{SO}(n)$.

(b) Prove that $\det\colon \mathrm{U}(n) \to \{z \in \mathbb{C} : |z| = 1\}$ is a surjective group homomorphism.

(c) Describe $\mathrm{O}(1)$ and $\mathrm{U}(1)$ completely.

Solution.

(a) $\det$ is multiplicative (Theorem 7.5.4) and takes only the values ± 1 on $\mathrm{O}(n)$ (Proposition 11.7.4 (b)), so it is a homomorphism into $\{1,-1\}$. It is surjective: $\det \mathbf{I}_n = 1$, and $\mathbf{A}_0 \coloneqq \mathrm{diag}(-1,1,\dots,1)$ is orthogonal, since its columns are $\pm\mathbf{e}_i$, with $\det \mathbf{A}_0 = -1$. Its kernel is $\{\mathbf{A} : \det \mathbf{A} = 1\} = \mathrm{SO}(n)$ by definition. If $\det \mathbf{A} = -1$ then $\det(\mathbf{A}_0^{-1}\mathbf{A}) = 1$, so $\mathbf{A} \in \mathbf{A}_0\,\mathrm{SO}(n)$; and no matrix has both determinants, so the union is disjoint.

(b) Multiplicativity again, and $|\det \mathbf{A}| = 1$ on $\mathrm{U}(n)$ by Proposition 11.7.4 (b). For surjectivity, given z with $|z| = 1$, the matrix $\mathrm{diag}(z,1,\dots,1)$ has orthonormal columns, hence lies in $\mathrm{U}(n)$, and its determinant is z (Theorem 7.4.6).

(c) A 1×1 matrix (a) satisfies $\overline{a}a = 1$ exactly when $|a| = 1$. So $\mathrm{U}(1)$ is the unit circle of $\mathbb{C}$ and $\mathrm{O}(1) = \{(1),(-1)\}$. Consistently, $\mathrm{SO}(1) = \{(1)\}$ is trivial.

Exercise. C2: Linearity is a hypothesis, not a consequence

(a) Let $c\colon \mathbb{C} \to \mathbb{C}$ be $c(z) = \overline{z}$, where $\mathbb{C}$ carries $\langle z, w\rangle = z\overline{w}$. Show that $\|c(z)\| = \|z\|$ for every z, that c is not linear over $\mathbb{C}$, and that c does not preserve the inner product. Which step of Theorem 11.7.3 fails?

(b) Let V be a **complex** inner product space and let $T \in \mathcal{L}(V)$ satisfy $\langle Tv, v\rangle = 0$ for every $v \in V$. Prove that $T = 0$.

(c) Show that (b) fails over $\mathbb{R}$ by exhibiting a nonzero $T \in \mathcal{L}(\mathbb{R}^2)$ with $\langle T\mathbf{v}, \mathbf{v}\rangle = 0$ for all $\mathbf{v}$.

Hint for (b): apply the hypothesis to $v = u + w$ and to $v = u + iw$.

Solution.

(a) $\|c(z)\| = |\overline{z}| = |z| = \|z\|$. But $c(iz) = \overline{iz} = -i\overline{z} = -ic(z) \neq ic(z)$ unless $z = 0$, so c is not $\mathbb{C}$-linear. And $\langle c(1), c(i)\rangle = \langle 1, -i\rangle = 1 \cdot \overline{-i} = i$, while $\langle 1, i\rangle = -i$. The proof of (a) $\Rightarrow$ (b) in Theorem 11.7.3 uses $T(u + i^k v) = Tu + i^k Tv$, which is exactly what c does not satisfy.

(b) Let $u, w \in V$. Applying the hypothesis to $u + w$ and expanding both slots,

$$\begin{aligned} 0 = \langle T(u+w), u+w\rangle &= \langle Tu, u\rangle + \langle Tu, w\rangle + \langle Tw, u\rangle + \langle Tw, w\rangle \\ &= \langle Tu, w\rangle + \langle Tw, u\rangle. \end{aligned} \tag{1}$$

Applying it to $u + iw$, and using $\langle Tu, iw\rangle = -i\langle Tu, w\rangle$, $\langle T(iw), u\rangle = i\langle Tw, u\rangle$ and $\langle T(iw), iw\rangle = i\overline{i}\langle Tw, w\rangle = 0$,

$$0 = -i\langle Tu, w\rangle + i\langle Tw, u\rangle.$$

Multiplying by i gives

$$\langle Tu, w\rangle - \langle Tw, u\rangle = 0. \tag{2}$$

Adding (1) and (2), $2\langle Tu, w\rangle = 0$, so $\langle Tu, w\rangle = 0$ for all u, w. Taking $w = Tu$ gives $\|Tu\|^2 = 0$, so $Tu = 0$ for every u, that is $T = 0$.

(c) Let T be the rotation $\mathbf{R}_{\pi/2}$, that is $T(x_1, x_2) = (-x_2, x_1)$. Then $\langle T\mathbf{x}, \mathbf{x}\rangle = -x_2x_1 + x_1x_2 = 0$ for every $\mathbf{x}$, and $T \neq 0$. The complex proof breaks down because the second equation is unavailable: there is no i to substitute.

Exercise. C3: Two reflections make a rotation

Let $\alpha, \beta \in \mathbb{R}$ and let $\mathbf{M}_\alpha, \mathbf{M}_\beta \in \mathrm{O}(2)$ be the reflections of Theorem 11.7.6, in the lines L_α and L_β at angles $\alpha/2$ and $\beta/2$.

(a) Prove that $\mathbf{M}_\alpha\mathbf{M}_\beta = \mathbf{R}_{\alpha-\beta}$.

(b) Deduce that the composite of the reflections in two lines through the origin is the rotation by **twice** the angle from the second line to the first, and that every rotation of the plane is a product of two reflections.

(c) Deduce that $\mathrm{O}(2)$ is not abelian, by computing $\mathbf{M}_\beta\mathbf{M}_\alpha$.

Solution.

(a) Multiplying out,

$$\mathbf{M}_\alpha\mathbf{M}_\beta = \begin{pmatrix} \cos\alpha & \sin\alpha \\ \sin\alpha & -\cos\alpha \end{pmatrix}\begin{pmatrix} \cos\beta & \sin\beta \\ \sin\beta & -\cos\beta \end{pmatrix} = \begin{pmatrix} c & -s \\ s & c \end{pmatrix},$$

where $c = \cos\alpha\cos\beta + \sin\alpha\sin\beta = \cos(\alpha-\beta)$ and $s = \sin\alpha\cos\beta - \cos\alpha\sin\beta = \sin(\alpha-\beta)$, by the subtraction formulas. The top-right entry is $\cos\alpha\sin\beta - \sin\alpha\cos\beta = -\sin(\alpha-\beta)$, as displayed. So $\mathbf{M}_\alpha\mathbf{M}_\beta = \mathbf{R}_{\alpha-\beta}$.

(b) The angle from L_β to L_α is $\frac{\alpha}{2} - \frac{\beta}{2}$, and (a) says the composite is the rotation by $\alpha - \beta$, twice that. Conversely, given θ, choose $\alpha = \theta$ and $\beta = 0$; then $\mathbf{R}_\theta = \mathbf{M}_\theta\mathbf{M}_0$, a product of two reflections.

(c) By (a) with the roles exchanged, $\mathbf{M}_\beta\mathbf{M}_\alpha = \mathbf{R}_{\beta-\alpha} = \mathbf{R}_{\alpha-\beta}^{-1}$. These two agree only when $\mathbf{R}_{\alpha-\beta}^2 = \mathbf{I}$, that is when $\alpha - \beta$ is a multiple of π. Taking $\alpha = \pi/2$ and $\beta = 0$ gives $\mathbf{M}_{\pi/2}\mathbf{M}_0 = \mathbf{R}_{\pi/2} \neq \mathbf{R}_{-\pi/2} = \mathbf{M}_0\mathbf{M}_{\pi/2}$, so $\mathrm{O}(2)$ is not abelian.

11.8 QR, Householder, and Hessenberg

Section 2 turned a basis into an orthonormal basis one vector at a time, and every step was bookkeeping: each new vector recorded in terms of those already produced. Collect that bookkeeping into a matrix and the Gram–Schmidt process becomes a **factorization $\mathbf{A} = \mathbf{QR}$** — an orthogonal counterpart of the **ŁU** factorization of Definition 3.6.1. This section proves it and its uniqueness, then builds it a second time out of reflections, a route that survives dependent columns. The same reflections, applied on **both** sides, bring any square matrix to an almost-triangular form by a unitary similarity. Throughout, F is $\mathbb{R}$ or $\mathbb{C}$, and F^m carries the standard inner product, linear in the first slot.

11.8.1 Gram–Schmidt in matrix form

Let $\mathbf{A} \in M_{m\times n}(F)$ have columns $\mathbf{a}_1, \dots, \mathbf{a}_n$, and suppose they are linearly independent. Apply Theorem 11.2.2 to the list $(\mathbf{a}_1, \dots, \mathbf{a}_n)$. It returns an orthonormal list $(\mathbf{q}_1, \dots, \mathbf{q}_n)$ in F^m with

$$\operatorname{span}(\mathbf{q}_1, \dots, \mathbf{q}_k) = \operatorname{span}(\mathbf{a}_1, \dots, \mathbf{a}_k)$$

for every k. In particular $\mathbf{a}_k$ lies in $\operatorname{span}(\mathbf{q}_1, \dots, \mathbf{q}_k)$, and Theorem 11.2.1 (a), applied inside that subspace, expands it:

$$\mathbf{a}_k = \sum_{i=1}^{k} \langle \mathbf{a}_k, \mathbf{q}_i \rangle \, \mathbf{q}_i \qquad (k = 1, \dots, n). \tag{$*$}$$

Now read $(*)$ as a matrix product. Put $\mathbf{Q} = (\mathbf{q}_1 \mid \cdots \mid \mathbf{q}_n) \in M_{m\times n}(F)$ and let $\mathbf{R} \in M_n(F)$ have entries

$$r_{ik} = \begin{cases} \langle \mathbf{a}_k, \mathbf{q}_i \rangle, & i \le k, \\ 0, & i > k. \end{cases}$$

Then $\mathbf{R}$ is upper triangular (Definition 0.9.5), and $(*)$ says that the k-th column of $\mathbf{A}$ equals $\mathbf{Q}$ times the k-th column of $\mathbf{R}$. That is exactly $\mathbf{A} = \mathbf{QR}$. **Nothing has been computed; the factorization is the Gram–Schmidt output, rewritten.**

Two features of the rewriting matter later. First the diagonal: in the notation of Theorem 11.2.2, $\mathbf{q}_k = \mathbf{w}_k / \|\mathbf{w}_k\|$ with $\mathbf{w}_k = \mathbf{a}_k - \sum_{i<k} \langle \mathbf{a}_k, \mathbf{q}_i \rangle \mathbf{q}_i$, so comparing with $(*)$ gives

$$r_{kk} = \langle \mathbf{a}_k, \mathbf{q}_k \rangle = \|\mathbf{w}_k\| > 0,$$

strictly positive because $\mathbf{w}_k \neq \mathbf{0}$. Second, $\mathbf{R}$ is recoverable from $\mathbf{Q}$ and $\mathbf{A}$: since the columns of $\mathbf{Q}$ are orthonormal,

$$(\mathbf{Q}^*\mathbf{Q})_{ij} = \sum_{p=1}^{m} \overline{q_{pi}}\, q_{pj} = \langle \mathbf{q}_j, \mathbf{q}_i \rangle = \delta_{ij},$$

by Definition 11.1.2, so $\mathbf{Q}^*\mathbf{Q} = \mathbf{I}_n$ and $\mathbf{R} = \mathbf{Q}^*\mathbf{A}$.

Theorem 11.8.1. QR Factorization

Let $\mathbf{A} \in M_{m\times n}(F)$ have **linearly independent** columns. Then there are matrices $\mathbf{Q} \in M_{m\times n}(F)$ and $\mathbf{R} \in M_n(F)$ with

(a) the columns of $\mathbf{Q}$ orthonormal, equivalently $\mathbf{Q}^*\mathbf{Q} = \mathbf{I}_n$;

(b) $\mathbf{R}$ upper triangular with **positive real** diagonal entries;

(c) $\mathbf{A} = \mathbf{QR}$.

Moreover $\mathbf{Q}$ and $\mathbf{R}$ are **unique**: any pair satisfying (a), (b), (c) is this one. The pair $(\mathbf{Q}, \mathbf{R})$ is the **QR factorization** of $\mathbf{A}$.

Idea. Existence is the paragraph above: run Gram–Schmidt and read off the coefficients. For uniqueness, suppose $\mathbf{Q}_1\mathbf{R}_1 = \mathbf{Q}_2\mathbf{R}_2$. We cannot cancel $\mathbf{Q}$, which is not square, but $\mathbf{Q}^*\mathbf{Q} = \mathbf{I}$ turns $\mathbf{A}^*\mathbf{A}$ into $\mathbf{R}^*\mathbf{R}$ and makes the $\mathbf{Q}$'s vanish. That leaves $\mathbf{R}_1^*\mathbf{R}_1 = \mathbf{R}_2^*\mathbf{R}_2$, so the matrix $\mathbf{S}$ carrying $\mathbf{R}_1$ to $\mathbf{R}_2$ is upper triangular with positive diagonal and satisfies $\mathbf{S}^*\mathbf{S} = \mathbf{I}$. Such a matrix has no room to be anything but $\mathbf{I}$: column by column, orthogonality to the earlier columns kills every entry above the diagonal.

Proof. Existence. Apply Theorem 11.2.2 to the columns of $\mathbf{A}$, which is legitimate because they are independent, and define $\mathbf{Q}$ and $\mathbf{R}$ as above. Then (a) holds by orthonormality of the output, (b) holds because $r_{kk} = \|\mathbf{w}_k\| > 0$, and (c) is $(*)$.

Uniqueness. Suppose $\mathbf{A} = \mathbf{Q}_1\mathbf{R}_1 = \mathbf{Q}_2\mathbf{R}_2$ with both pairs satisfying (a) and (b). Each $\mathbf{R}_t$ has nonzero diagonal entries, hence is invertible by Lemma 3.6.2. Using $\mathbf{Q}_t^*\mathbf{Q}_t = \mathbf{I}_n$ twice,

$$\mathbf{R}_1^*\mathbf{R}_1 = \mathbf{R}_1^*\mathbf{Q}_1^*\mathbf{Q}_1\mathbf{R}_1 = \mathbf{A}^*\mathbf{A} = \mathbf{R}_2^*\mathbf{Q}_2^*\mathbf{Q}_2\mathbf{R}_2 = \mathbf{R}_2^*\mathbf{R}_2.$$

Put $\mathbf{S} = \mathbf{R}_2\mathbf{R}_1^{-1}$, so that $\mathbf{R}_2 = \mathbf{S}\mathbf{R}_1$. The matrix $\mathbf{R}_1^{-1}$ is upper triangular: its j-th column is the solution $\mathbf{x}$ of $\mathbf{R}_1\mathbf{x} = \mathbf{e}_j$, and solving from the bottom row upward gives $x_i = 0$ for every $i > j$, exactly as in the proof of Lemma 3.6.2. Hence $\mathbf{S}$ is upper triangular, being a product of upper triangular matrices: in $(\mathbf{S})_{ij} = \sum_k (\mathbf{R}_2)_{ik}(\mathbf{R}_1^{-1})_{kj}$ a term can be nonzero only when $i \le k$ and $k \le j$, which is impossible for $i > j$, and for $i = j$ only the term $k = i$ survives. So its diagonal entries are $(\mathbf{S})_{kk} = (\mathbf{R}_2)_{kk}(\mathbf{R}_1)_{kk}^{-1} > 0$, a quotient of positive reals. Substituting $\mathbf{R}_2 = \mathbf{S}\mathbf{R}_1$ into the display gives $\mathbf{R}_1^*\mathbf{R}_1 = \mathbf{R}_1^*\mathbf{S}^*\mathbf{S}\mathbf{R}_1$, and multiplying by $(\mathbf{R}_1^*)^{-1} = (\mathbf{R}_1^{-1})^*$ on the left and $\mathbf{R}_1^{-1}$ on the right leaves

$$\mathbf{S}^*\mathbf{S} = \mathbf{I}_n.$$

Claim. An upper triangular $\mathbf{S} \in M_n(F)$ with $\mathbf{S}^*\mathbf{S} = \mathbf{I}_n$ and positive real diagonal entries is $\mathbf{I}_n$.

Proof. Write $\mathbf{s}_1, \dots, \mathbf{s}_n$ for the columns of $\mathbf{S}$. As computed before the theorem, $\mathbf{S}^*\mathbf{S} = \mathbf{I}_n$ says $\langle \mathbf{s}_j, \mathbf{s}_i \rangle = \delta_{ij}$. We show $\mathbf{s}_j = \mathbf{e}_j$ by induction on j. Let $1 \le j \le n$ and suppose $\mathbf{s}_i = \mathbf{e}_i$ for every $i < j$ (no assumption when $j = 1$). For $i < j$,

$$0 = \langle \mathbf{s}_j, \mathbf{s}_i \rangle = \langle \mathbf{s}_j, \mathbf{e}_i \rangle = s_{ij},$$

so every entry of $\mathbf{s}_j$ above the diagonal vanishes; and $s_{ij} = 0$ for $i > j$ because $\mathbf{S}$ is upper triangular. Hence $\mathbf{s}_j = s_{jj}\mathbf{e}_j$, and $1 = \langle \mathbf{s}_j, \mathbf{s}_j \rangle = |s_{jj}|^2$ forces $|s_{jj}| = 1$. Since $s_{jj} > 0$, this gives $s_{jj} = 1$ and $\mathbf{s}_j = \mathbf{e}_j$.

■

By the Claim, $\mathbf{S} = \mathbf{I}_n$, so $\mathbf{R}_2 = \mathbf{R}_1$ and therefore $\mathbf{Q}_2 = \mathbf{A}\mathbf{R}_2^{-1} = \mathbf{A}\mathbf{R}_1^{-1} = \mathbf{Q}_1$. This proves uniqueness.

■

The positivity in (b) is a normalization and nothing deeper: without it, scaling $\mathbf{q}_k$ by any c of modulus 1 and row k of $\mathbf{R}$ by $\bar{c}$ gives another factorization. Exercise (C1: All the QR factorizations of a matrix) describes all of them.

Warning. **In $\mathbf{A} = \mathbf{QR}$ the matrix $\mathbf{Q}$ is usually not square, and $\mathbf{QQ}^* \neq \mathbf{I}$.** Only $\mathbf{Q}^*\mathbf{Q} = \mathbf{I}_n$ holds. For the 3×2 matrix of Example (A 3×2 QR, and a least-squares problem)

below,

$$\mathbf{Q}\mathbf{Q}^* = \tfrac{1}{18}\begin{pmatrix} 2 & 4 & 4 \\ 4 & 17 & -1 \\ 4 & -1 & 17 \end{pmatrix},$$

whose trace is 2, not 3. This is the "identity on what?" trap of Theorem 11.7.3: the columns of $\mathbf{Q}$ are an orthonormal basis of a **plane**, and $\mathbf{Q}\mathbf{Q}^*$ is the orthogonal projection onto it (Theorem 11.3.3).

Quick check. Each factor of a QR factorization is recoverable from the other: $\mathbf{R} = \mathbf{Q}^*\mathbf{A}$ and $\mathbf{Q} = \mathbf{A}\mathbf{R}^{-1}$. One of the two recoveries needs an invertibility hypothesis and the other does not. Which, and why?

Solution. The second. Multiplying $\mathbf{A} = \mathbf{Q}\mathbf{R}$ on the left by $\mathbf{Q}^*$ gives $\mathbf{Q}^*\mathbf{Q}\mathbf{R} = \mathbf{R}$, which uses only $\mathbf{Q}^*\mathbf{Q} = \mathbf{I}_n$: nothing is inverted, and $\mathbf{Q}$ is not even square. Multiplying on the right by $\mathbf{R}^{-1}$ presupposes that $\mathbf{R}$ is invertible — true by Lemma 3.6.2, but only because clause (b) makes the diagonal positive.

11.8.2 Least squares through QR

An inconsistent system $\mathbf{A}\mathbf{x} = \mathbf{b}$ was handled in Section 4 by the normal equations $\mathbf{A}^*\mathbf{A}\mathbf{x} = \mathbf{A}^*\mathbf{b}$ (Theorem 11.4.2). When the columns of $\mathbf{A}$ are independent, the QR factorization simplifies those equations to a triangular system.

Substituting $\mathbf{A} = \mathbf{Q}\mathbf{R}$ and using $\mathbf{Q}^*\mathbf{Q} = \mathbf{I}_n$,

$$\mathbf{A}^*\mathbf{A} = \mathbf{R}^*\mathbf{Q}^*\mathbf{Q}\mathbf{R} = \mathbf{R}^*\mathbf{R}, \qquad \mathbf{A}^*\mathbf{b} = \mathbf{R}^*\mathbf{Q}^*\mathbf{b},$$

so the normal equations read $\mathbf{R}^*\mathbf{R}\mathbf{x} = \mathbf{R}^*\mathbf{Q}^*\mathbf{b}$. Now $\mathbf{R}$ is invertible by Lemma 3.6.2, hence so is $\mathbf{R}^*$, with $(\mathbf{R}^*)^{-1} = (\mathbf{R}^{-1})^*$ — conjugate-transpose $\mathbf{R}\mathbf{R}^{-1} = \mathbf{I}_n$. So $\mathbf{R}^*$ cancels:

$$\mathbf{R}\mathbf{x} = \mathbf{Q}^*\mathbf{b}. \tag{$\dagger$}$$

By Corollary 11.4.4 the least-squares solution is unique here, and ($\dagger$) is a triangular system, solved by back substitution. The deeper reason to prefer it is not the operation count: forming $\mathbf{A}^*\mathbf{A}$ squares the sensitivity of the problem to rounding, while ($\dagger$) never forms that product at all. Chapter 24 makes "sensitivity" precise and proves the comparison; here we record only the practice.

Example. A 3×2 QR, and a least-squares problem

Let

$$\mathbf{A} = \begin{pmatrix} 1 & 1 \\ 2 & 1 \\ 2 & 3 \end{pmatrix}, \qquad \mathbf{b} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}.$$

(a) Find the QR factorization of $\mathbf{A}$.

(b) Find the $\mathbf{x} \in \mathbb{R}^2$ minimizing $\|\mathbf{A}\mathbf{x} - \mathbf{b}\|$, using (a).

Solution.

(a) Write $\mathbf{a}_1 = (1,2,2)$ and $\mathbf{a}_2 = (1,1,3)$. They are independent, since neither is a multiple of the other.
Step 1. $\|\mathbf{a}_1\|^2 = 1+4+4 = 9$, so $r_{11} = 3$ and $\mathbf{q}_1 = \frac{1}{3}(1,2,2)$.
Step 2. $r_{12} = \langle \mathbf{a}_2, \mathbf{q}_1 \rangle = \frac{1}{3}(1+2+6) = 3$, so

$$\mathbf{w}_2 = (1,1,3) - 3 \cdot \tfrac{1}{3}(1,2,2) = (0,-1,1),$$

with $\|\mathbf{w}_2\| = \sqrt{2} = r_{22}$ and $\mathbf{q}_2 = \frac{1}{\sqrt{2}}(0,-1,1)$. Hence

$$\mathbf{Q} = \begin{pmatrix} \frac{1}{3} & 0 \\ \frac{2}{3} & -\frac{1}{\sqrt{2}} \\ \frac{2}{3} & \frac{1}{\sqrt{2}} \end{pmatrix}, \qquad \mathbf{R} = \begin{pmatrix} 3 & 3 \\ 0 & \sqrt{2} \end{pmatrix}.$$

Check. $\mathbf{QR}$ has first column $3\mathbf{q}_1 = (1,2,2)$ and second column $3\mathbf{q}_1 + \sqrt{2}\,\mathbf{q}_2 = (1,2,2) + (0,-1,1) = (1,1,3)$, which is $\mathbf{A}$.

(b) By (†) we need $\mathbf{Q}^*\mathbf{b}$. Here $\langle \mathbf{b}, \mathbf{q}_1 \rangle = \frac{1}{3}(1+2+2) = \frac{5}{3}$ and $\langle \mathbf{b}, \mathbf{q}_2 \rangle = \frac{1}{\sqrt{2}}(0-1+1) = 0$, so the system is

$$\begin{pmatrix} 3 & 3 \\ 0 & \sqrt{2} \end{pmatrix} \mathbf{x} = \begin{pmatrix} \frac{5}{3} \\ 0 \end{pmatrix}.$$

Back substitution gives $\sqrt{2}\,x_2 = 0$, so $x_2 = 0$, and then $3x_1 = \frac{5}{3}$, so $x_1 = \frac{5}{9}$. The minimizer is $\mathbf{x} = (\frac{5}{9}, 0)$.

Check against the normal equations. $\mathbf{A}^\top\mathbf{A} = \begin{pmatrix} 9 & 9 \\ 9 & 11 \end{pmatrix}$ and $\mathbf{A}^\top\mathbf{b} = (5,5)$. Subtracting the first equation from the second gives $2x_2 = 0$, and then $9x_1 = 5$. The same answer, with the same arithmetic done in a worse order.

11.8.3 Reflections

Gram–Schmidt builds $\mathbf{Q}$ out of the columns of $\mathbf{A}$. There is an entirely different route to a triangular $\mathbf{R}$, modeled on elimination: hit $\mathbf{A}$ on the left with simple matrices that create zeros, one column at a time, and collect the inverses. Chapter 3 used shears and got $\mathbf{A} = \text{Ł}\mathbf{U}$; if every matrix we hit $\mathbf{A}$ with is **unitary** instead, the collected product is unitary and we get $\mathbf{A} = \mathbf{QR}$. Which unitary matrices are simple enough? A rotation in the plane of two coordinates is one answer, taken up below. The more efficient answer is a reflection, because one reflection can flatten an entire column at once.

A reflection in a hyperplane sends a vector to its mirror image: keep the part perpendicular to $\mathbf{w}$*, and reverse the part along* $\mathbf{w}$*.*

Definition 11.8.1. Householder Reflection

Let $\mathbf{w} \in F^m$ be **nonzero**. The **Householder reflection** determined by $\mathbf{w}$ is

$$\mathbf{H}_{\mathbf{w}} = \mathbf{I}_m - \frac{2\,\mathbf{w}\mathbf{w}^*}{\|\mathbf{w}\|^2} \in M_m(F).$$

In words: $\mathbf{w}\mathbf{w}^*$ is an $m \times m$ matrix (a column times a row), and for any $\mathbf{v} \in F^m$ the product $\mathbf{w}^*\mathbf{v}$ is the scalar $\sum_p \overline{w_p} v_p = \langle \mathbf{v}, \mathbf{w} \rangle$. So the definition says

$$\mathbf{H}_{\mathbf{w}}\mathbf{v} = \mathbf{v} - 2\,\frac{\langle \mathbf{v}, \mathbf{w} \rangle}{\|\mathbf{w}\|^2}\,\mathbf{w} = \mathbf{v} - 2P_{\mathrm{span}(\mathbf{w})}\mathbf{v}, \tag{‡}$$

where the last equality is Theorem 11.3.3 with the orthonormal basis $(\mathbf{w}/\|\mathbf{w}\|)$ of the line $\mathrm{span}(\mathbf{w})$. Writing $\mathbf{v} = P_{\mathrm{span}(\mathbf{w})}\mathbf{v} + (\mathbf{v} - P_{\mathrm{span}(\mathbf{w})}\mathbf{v})$ as in Theorem 11.3.2, we see that $\mathbf{H}_{\mathbf{w}}$ keeps the component of $\mathbf{v}$ in $\mathrm{span}(\mathbf{w})^\perp$ and reverses the component along $\mathbf{w}$. That is a mirror reflection in the hyperplane $\mathrm{span}(\mathbf{w})^\perp$.

Note at once that $\mathbf{H}_{c\mathbf{w}} = \mathbf{H}_{\mathbf{w}}$ for every nonzero scalar c, since $(c\mathbf{w})(c\mathbf{w})^* = |c|^2\mathbf{w}\mathbf{w}^*$ and $\|c\mathbf{w}\|^2 = |c|^2\|\mathbf{w}\|^2$. **Only the line through $\mathbf{w}$ matters.**

Examples.

- **The smallest case.** In F^1 take $\mathbf{w} = 1$: then $\mathbf{H}_{\mathbf{w}} = -1$, the map $x \mapsto -x$. The "hyperplane" is $\{0\}$, and this is the only nontrivial isometry of the line.
- **A coordinate reflection.** In $\mathbb{R}^2$ take $\mathbf{w} = \mathbf{e}_2$, so $\mathbf{w}\mathbf{w}^\top = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$ and $\mathbf{H}_{\mathbf{w}} = \mathrm{diag}(1, -1)$: the reflection in the x-axis.
- **A diagonal reflection.** In $\mathbb{R}^2$ take $\mathbf{w} = (1, -1)$, so $\|\mathbf{w}\|^2 = 2$ and $\mathbf{H}_{\mathbf{w}} = \mathbf{I}_2 - \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$, which swaps the coordinates. It is the reflection in the line $y = x$, and $\mathrm{span}(\mathbf{w})^\perp$ is that line.
- **A complex one.** In $\mathbb{C}^m$ take $\mathbf{w} = \mathbf{e}_1$: then $\mathbf{H}_{\mathbf{w}} = \mathrm{diag}(-1, 1, \ldots, 1)$, over either field.

Nonexample by minimal change. Drop the factor 2: the matrix $\mathbf{I}_m - \mathbf{w}\mathbf{w}^*/\|\mathbf{w}\|^2$ is $\mathbf{I}_m - P_{\mathrm{span}(\mathbf{w})}$, the orthogonal projection onto $\mathrm{span}(\mathbf{w})^\perp$. It is Hermitian, as $\mathbf{H}_{\mathbf{w}}$ is, but **not** unitary and not invertible: it sends $\mathbf{w}$ to $\mathbf{0}$. Halving the coefficient turns "reverse the component along $\mathbf{w}$" into "delete it", an isometry into a shadow.

Proposition 11.8.2. Properties of a Householder Reflection

Let $\mathbf{w} \in F^m$ be nonzero and $\mathbf{H} = \mathbf{H}_{\mathbf{w}}$. Then

(a) $\mathbf{H}^* = \mathbf{H}$ ($\mathbf{H}$ is Hermitian);
(b) $\mathbf{H}^2 = \mathbf{I}_m$, so $\mathbf{H}^{-1} = \mathbf{H}$;
(c) $\mathbf{H}$ is unitary;
(d) $\mathbf{H}\mathbf{w} = -\mathbf{w}$, and $\mathbf{H}\mathbf{v} = \mathbf{v}$ for every $\mathbf{v} \perp \mathbf{w}$;
(e) $\det \mathbf{H} = -1$.

Proof.

(a) By Definition 11.1.2, $(\mathbf{w}\mathbf{w}^*)^* = \mathbf{w}^{**}\mathbf{w}^* = \mathbf{w}\mathbf{w}^*$, and $\|\mathbf{w}\|^2$ is a real number (Theorem 11.1.1 (a)), so $\mathbf{H}^* = \mathbf{I}_m - 2\mathbf{w}\mathbf{w}^*/\|\mathbf{w}\|^2 = \mathbf{H}$.

(b) Since $\mathbf{w}^*\mathbf{w} = \langle \mathbf{w}, \mathbf{w} \rangle = \|\mathbf{w}\|^2$,

$$\mathbf{H}^2 = \mathbf{I}_m - \frac{4\mathbf{w}\mathbf{w}^*}{\|\mathbf{w}\|^2} + \frac{4\,\mathbf{w}(\mathbf{w}^*\mathbf{w})\mathbf{w}^*}{\|\mathbf{w}\|^4} = \mathbf{I}_m,$$

because the last two terms cancel.

(c) By (a) and (b), $\mathbf{H}^*\mathbf{H} = \mathbf{H}^2 = \mathbf{I}_m$, which is the criterion of Definition 11.7.2.

(d) Both statements are read off (‡): $P_{\mathrm{span}(\mathbf{w})}\mathbf{w} = \mathbf{w}$ gives $\mathbf{H}\mathbf{w} = \mathbf{w} - 2\mathbf{w} = -\mathbf{w}$, and $\langle \mathbf{v}, \mathbf{w} \rangle = 0$ gives $\mathbf{H}\mathbf{v} = \mathbf{v}$.

(e) Extend the unit vector $\mathbf{w}/\|\mathbf{w}\|$ to an orthonormal basis $(\mathbf{f}_1, \ldots, \mathbf{f}_m)$ of F^m with $\mathbf{f}_1 = \mathbf{w}/\|\mathbf{w}\|$, which is possible by Corollary 11.2.3, and let $\mathbf{P} = (\mathbf{f}_1 \mid \cdots \mid \mathbf{f}_m)$. By (d), $\mathbf{H}\mathbf{f}_1 = -\mathbf{f}_1$ and $\mathbf{H}\mathbf{f}_i = \mathbf{f}_i$ for $i \geq 2$, that is, $\mathbf{H}\mathbf{P} = \mathbf{P}\mathbf{D}$ with $\mathbf{D} = \mathrm{diag}(-1, 1, \ldots, 1)$. The matrix $\mathbf{P}$ is unitary, hence invertible, so $\mathbf{H} = \mathbf{P}\mathbf{D}\mathbf{P}^{-1}$. Taking determinants, $\det \mathbf{H} = \det \mathbf{D}$ by Theorem 7.5.4 and Corollary 7.5.6, and $\det \mathbf{D} = -1$ by Theorem 7.4.6.

■

So a Householder reflection is a unitary involution of determinant -1: cheap to store (one vector) and cheap to apply, by (‡). What makes it useful is the following choice of $\mathbf{w}$.

Theorem 11.8.3. A Reflection Flattening a Vector

Let $\mathbf{x} \in F^m$ be nonzero, and let

$$\alpha = \begin{cases} x_1/|x_1|, & x_1 \neq 0, \\ 1, & x_1 = 0, \end{cases} \qquad \mathbf{w} = \mathbf{x} + \alpha\|\mathbf{x}\|\,\mathbf{e}_1.$$

Then $\mathbf{w} \neq \mathbf{0}$ and

$$\mathbf{H}_{\mathbf{w}}\,\mathbf{x} = -\alpha\|\mathbf{x}\|\,\mathbf{e}_1.$$

In particular $\mathbf{H}_{\mathbf{w}}\mathbf{x}$ is a multiple of $\mathbf{e}_1$ of length $\|\mathbf{x}\|$. Over $\mathbb{R}$, α is the sign of x_1 (or 1 if $x_1 = 0$), and $\mathbf{H}_{\mathbf{w}}\mathbf{x} = \mp\|\mathbf{x}\|\mathbf{e}_1$ according to that sign.

Idea. Picture the real plane. We want a mirror line taking $\mathbf{x}$ to a point on the first axis at the same distance from the origin. There are two such points, $\pm\|\mathbf{x}\|\mathbf{e}_1$, and for each the mirror is the perpendicular bisector of the segment joining $\mathbf{x}$ to it — whose normal direction is the **difference** $\mathbf{x} \mp \|\mathbf{x}\|\mathbf{e}_1$. So the vector $\mathbf{w}$ is forced, up to which of the two targets we aim at, and all that remains is to check that the coefficient $2\langle \mathbf{x}, \mathbf{w} \rangle/\|\mathbf{w}\|^2$ in (‡) comes out exactly 1, so that $\mathbf{H}_{\mathbf{w}}\mathbf{x} = \mathbf{x} - \mathbf{w}$. The role of α is to choose the target on the far side from $\mathbf{x}$.

Proof. Note first that $\overline{\alpha}x_1 = |x_1|$ in both cases, and $|\alpha| = 1$. Expanding with (IP1) and Theorem 11.1.1 (b),

$$\begin{aligned} \|\mathbf{w}\|^2 &= \|\mathbf{x}\|^2 + \overline{\alpha}\|\mathbf{x}\|\langle \mathbf{x}, \mathbf{e}_1 \rangle + \alpha\|\mathbf{x}\|\langle \mathbf{e}_1, \mathbf{x} \rangle + |\alpha|^2\|\mathbf{x}\|^2 \\ &= 2\|\mathbf{x}\|^2 + 2\|\mathbf{x}\|\,|x_1|, \end{aligned}$$

where the second line uses $\langle \mathbf{x}, \mathbf{e}_1 \rangle = x_1$, $\langle \mathbf{e}_1, \mathbf{x} \rangle = \overline{x_1}$ and $\overline{\alpha}x_1 + \alpha\overline{x_1} = 2|x_1|$. Since $\mathbf{x} \neq \mathbf{0}$ we

have $\|\mathbf{x}\| > 0$, so $\|\mathbf{w}\|^2 > 0$ and $\mathbf{w} \neq \mathbf{0}$. Likewise

$$\langle \mathbf{x}, \mathbf{w} \rangle = \|\mathbf{x}\|^2 + \overline{\alpha}\|\mathbf{x}\|\, x_1 = \|\mathbf{x}\|^2 + \|\mathbf{x}\||x_1|,$$

so that $2\langle \mathbf{x}, \mathbf{w} \rangle = \|\mathbf{w}\|^2$. Hence by (‡),

$$\mathbf{H}_{\mathbf{w}}\mathbf{x} = \mathbf{x} - \frac{2\langle \mathbf{x}, \mathbf{w} \rangle}{\|\mathbf{w}\|^2}\,\mathbf{w} = \mathbf{x} - \mathbf{w} = -\alpha\|\mathbf{x}\|\,\mathbf{e}_1.$$

Finally $\|-\alpha\|\mathbf{x}\|\mathbf{e}_1\| = |\alpha|\|\mathbf{x}\| = \|\mathbf{x}\|$ by Theorem 11.1.2 (b), as it must be, since $\mathbf{H}_{\mathbf{w}}$ is unitary. This proves the theorem.

■

Why that sign. The opposite choice $\mathbf{w}' = \mathbf{x} - \alpha\|\mathbf{x}\|\mathbf{e}_1$ also works whenever $\mathbf{w}' \neq \mathbf{0}$, giving $\mathbf{H}_{\mathbf{w}'}\mathbf{x} = +\alpha\|\mathbf{x}\|\mathbf{e}_1$; in exact arithmetic the two are equally good, and in floating point they are not. The first entry of $\mathbf{w}'$ is $x_1 - \alpha\|\mathbf{x}\|$, a difference of two numbers of modulus $|x_1|$ and $\|\mathbf{x}\|$ — and when $\mathbf{x}$ is close to a multiple of $\mathbf{e}_1$ those two are nearly equal, so the leading digits cancel and the computed first entry keeps only the few digits they did not share. The other entries of $\mathbf{w}'$ are copied from $\mathbf{x}$ and stay exact, so the computed $\mathbf{w}'$ points in a noticeably wrong direction, and so does the reflection built from it. With $\mathbf{w} = \mathbf{x} + \alpha\|\mathbf{x}\|\mathbf{e}_1$ the first entry has modulus $|x_1| + \|\mathbf{x}\|$: like-signed quantities, nothing to cancel. (The same computation shows $\mathbf{w} \neq \mathbf{0}$ always, whereas $\mathbf{w}' = \mathbf{0}$ exactly when $\mathbf{x} = \alpha\|\mathbf{x}\|\mathbf{e}_1$, the extreme case of that cancellation.) Chapter 24 quantifies the loss; the summary here is that both signs are correct mathematics and only one is correct arithmetic.

Warning. **Do not build the matrix $\mathbf{H}_{\mathbf{w}}$ in order to use it.** Forming it and multiplying costs m^2 operations per column, while (‡) applies the same reflection with one inner product and one scaled subtraction, about $4m$. Every "multiply by $\mathbf{H}$" below is really an instance of (‡); the matrix is a proof device.

Quick check. Let $\mathbf{x} = (0, 3, 4) \in \mathbb{R}^3$. Find $\mathbf{w}$ and $\mathbf{H}_{\mathbf{w}}\mathbf{x}$ as in Theorem 11.8.3, and say why the sign question does not arise here.

Solution. $\|\mathbf{x}\| = 5$ and $x_1 = 0$, so $\alpha = 1$ and $\mathbf{w} = (0, 3, 4) + 5(1, 0, 0) = (5, 3, 4)$. Then $\mathbf{H}_{\mathbf{w}}\mathbf{x} = -5\mathbf{e}_1 = (-5, 0, 0)$. With $x_1 = 0$ there is no cancellation to avoid: the first entry of $\mathbf{w}$ is 0 ± 5 either way, and both choices are equally accurate. The convention $\alpha = 1$ is only a convention.

11.8.4 QR by reflections

One reflection makes the first column a multiple of $\mathbf{e}_1$ — exactly the zeros an upper triangular matrix needs in column 1. Repeating on the trailing submatrix triangularizes everything.

Theorem 11.8.4. QR by Householder Reflections

Let $\mathbf{A} \in M_{m\times n}(F)$. Then there are a **unitary** $\mathbf{Q} \in M_m(F)$ and an $\mathbf{R} \in M_{m\times n}(F)$ with $r_{ij} = 0$ whenever $i > j$, such that

$$\mathbf{A} = \mathbf{QR}.$$

Moreover $\mathbf{Q}$ can be taken to be a product of at most $\min(n, m-1)$ Householder reflections. **No hypothesis on the columns of A is needed.**

Idea. Induction on the number of columns. Step one is Theorem 11.8.3 applied to the first column: after it, the first column is $\rho\mathbf{e}_1$ and what is left to triangularize is the trailing $(m-1)\times(n-1)$ block. Feed that block to the induction hypothesis, and lift the unitary matrix it returns to F^m by putting a 1 in the top-left corner — a block matrix of that shape does not touch the first row or the first column, so it cannot spoil the zeros just created.

Proof. We induct on n, the statement being asserted for every $m \geq 1$ at once.

If $m = 1$, then $\mathbf{A}$ has no entries below its diagonal, so $\mathbf{Q} = \mathbf{I}_1$ and $\mathbf{R} = \mathbf{A}$ work. Assume from now on that $m \geq 2$.

Let $\mathbf{a}$ be the first column of $\mathbf{A}$. If $\mathbf{a} = \mathbf{0}$, put $\mathbf{H} = \mathbf{I}_m$ and $\rho = 0$; otherwise let $\mathbf{H} = \mathbf{H}_{\mathbf{w}}$ be the reflection of Theorem 11.8.3 for $\mathbf{x} = \mathbf{a}$, and $\rho = -\alpha\|\mathbf{a}\|$. In both cases $\mathbf{H}$ is unitary and Hermitian with $\mathbf{H}^{-1} = \mathbf{H}$ — by Proposition 11.8.2 (a)–(c) for a reflection, and because $\mathbf{I}_m^* = \mathbf{I}_m = \mathbf{I}_m^{-1}$ in the degenerate case — and $\mathbf{Ha} = \rho\mathbf{e}_1$. Hence, by Theorem 8.1.1,

$$\mathbf{HA} = \begin{pmatrix} \rho & \mathbf{b}^* \\ \mathbf{0} & \mathbf{A}' \end{pmatrix}$$

for some $\mathbf{b} \in F^{n-1}$ and $\mathbf{A}' \in M_{(m-1)\times(n-1)}(F)$, where the $\mathbf{0}$ is a column of $m-1$ zeros.

Base case $n = 1$. There is no block $\mathbf{A}'$, and $\mathbf{HA} = \rho\mathbf{e}_1$ already has $r_{i1} = 0$ for $i > 1$. So $\mathbf{A} = \mathbf{H}(\mathbf{HA})$ is of the required form with $\mathbf{Q} = \mathbf{H}$.

Induction step. Let $n \geq 2$ and assume the statement for $n-1$ columns. Applied to $\mathbf{A}'$, which has $m - 1 \geq 1$ rows, it gives a unitary $\mathbf{Q}' \in M_{m-1}(F)$ and an $\mathbf{R}'$ with zeros below the diagonal and $\mathbf{A}' = \mathbf{Q}'\mathbf{R}'$. Put

$$\mathbf{U} = \begin{pmatrix} 1 & \mathbf{0}^* \\ \mathbf{0} & \mathbf{Q}' \end{pmatrix}, \qquad \mathbf{R} = \begin{pmatrix} \rho & \mathbf{b}^* \\ \mathbf{0} & \mathbf{R}' \end{pmatrix}.$$

Then $\mathbf{U}^*\mathbf{U} = \mathbf{I}_m$ by Theorem 8.1.1, so $\mathbf{U}$ is unitary, and $\mathbf{UR} = \mathbf{HA}$ by the same computation. Also $\mathbf{R}$ has zeros below its diagonal: in column 1 this is the block $\mathbf{0}$; and for $j \geq 2$ and $i > j$, the (i, j)-entry of $\mathbf{R}$ is the $(i-1, j-1)$-entry of $\mathbf{R}'$, which vanishes because $i - 1 > j - 1$.

Therefore $\mathbf{A} = \mathbf{HUR}$, and $\mathbf{Q} = \mathbf{HU}$ is unitary as a product of unitary matrices (Proposition 11.7.4). Unwinding the induction, $\mathbf{Q}$ is a product of one reflection (or $\mathbf{I}_m$) for each of the columns $1, \dots, \min(n, m-1)$, since the recursion stops once one row is left. This proves the theorem.

■

Example. A 3×3 QR by reflections

Triangularize

$$\mathbf{A} = \begin{pmatrix} 2 & 0 & 1 \\ 1 & 5 & 2 \\ 2 & 5 & 4 \end{pmatrix}$$

by Householder reflections, and read off $\mathbf{Q}$ and $\mathbf{R}$.

Solution. Step 1. The first column is $\mathbf{a} = (2, 1, 2)$ with $\|\mathbf{a}\| = 3$ and $a_1 = 2 > 0$, so $\alpha = 1$ and

$$\mathbf{w}_1 = (2, 1, 2) + 3(1, 0, 0) = (5, 1, 2), \qquad \|\mathbf{w}_1\|^2 = 30.$$

Apply (‡) to each column: $\mathbf{H}_1\mathbf{v} = \mathbf{v} - \frac{1}{15}\langle \mathbf{v}, \mathbf{w}_1 \rangle \mathbf{w}_1$. All three columns pair with $\mathbf{w}_1$ to give 15, so each loses exactly one copy of $\mathbf{w}_1$:

$$\mathbf{H}_1\mathbf{A} = \begin{pmatrix} -3 & -5 & -4 \\ 0 & 4 & 1 \\ 0 & 3 & 2 \end{pmatrix}.$$

(As predicted, the first column is $-\|\mathbf{a}\|\mathbf{e}_1$.)

Step 2. Work now in the trailing 2×2 block, whose first column is $\mathbf{y} = (4, 3)$ with $\|\mathbf{y}\| = 5$ and $y_1 > 0$. So $\mathbf{w}_2 = (9, 3)$, and inside $\mathbb{R}^3$ we reflect in $\widetilde{\mathbf{w}}_2 = (0, 9, 3)$, which fixes $\mathbf{e}_1$ and so leaves column 1 alone; here $\mathbf{H}_2\mathbf{v} = \mathbf{v} - \frac{1}{45}\langle \mathbf{v}, \widetilde{\mathbf{w}}_2 \rangle \widetilde{\mathbf{w}}_2$. The second and third columns of $\mathbf{H}_1\mathbf{A}$ pair with $\widetilde{\mathbf{w}}_2$ to give $36 + 9 = 45$ and $9 + 6 = 15$, so they lose one copy and a third of a copy respectively:

$$\mathbf{R} = \mathbf{H}_2\mathbf{H}_1\mathbf{A} = \begin{pmatrix} -3 & -5 & -4 \\ 0 & -5 & -2 \\ 0 & 0 & 1 \end{pmatrix}.$$

The unitary factor. By Proposition 11.8.2 (b), $\mathbf{A} = \mathbf{H}_1\mathbf{H}_2\mathbf{R}$, so $\mathbf{Q} = \mathbf{H}_1\mathbf{H}_2$. Computing the product,

$$\mathbf{Q} = \tfrac{1}{3}\begin{pmatrix} -2 & 2 & -1 \\ -1 & -2 & -2 \\ -2 & -1 & 2 \end{pmatrix},$$

whose columns have length 1 and are pairwise orthogonal, as one checks from the integer entries: for instance $(-2)(2) + (-1)(-2) + (-2)(-1) = 0$.

Check. $\det \mathbf{A} = 15$ by expansion along the first row, and $\det \mathbf{R} = (-3)(-5)(1) = 15$ by Theorem 7.4.6, consistent with $\det \mathbf{Q} = (-1)(-1) = 1$ from Proposition 11.8.2 (e).

The diagonal of $\mathbf{R}$ here is $(-3, -5, 1)$, not positive: the reflection route produces **a** factorization, not **the** one of Theorem 11.8.1. The two are related by flipping the sign of each row of $\mathbf{R}$ with a negative diagonal entry, and of the matching column of $\mathbf{Q}$; Exercise (B2: The same matrix by reflections) carries this out on the matrix of Example (A 3×2 QR, and a least-squares problem).

Warning. **"The QR factorization" is definite only with the positivity normalization.** Theorem 11.8.1 pins the pair down by demanding $r_{kk} > 0$; the reflection algorithm

ignores that demand and chooses signs for stability instead. So two correct computations of "the QR factorization of $\mathbf{A}$" can differ by signs. Compare $|r_{kk}|$, and the lines spanned by the columns of $\mathbf{Q}$.

Where Gram–Schmidt cannot follow. Theorem 11.8.4 assumed nothing about the columns, and that is a genuine gain. Let the columns of $\mathbf{A}$ be dependent, and let k be the first index with $\mathbf{a}_k \in \operatorname{span}(\mathbf{a}_1, \dots, \mathbf{a}_{k-1})$. The Gram–Schmidt recursion then produces $\mathbf{w}_k = \mathbf{0}$ and stops, because the next step divides by $\|\mathbf{w}_k\|$; Exercise (C2: Gram–Schmidt on a dependent list) works this out. Nor is the failure an artifact of the method: if $\mathbf{A} = \mathbf{QR}$ with $\mathbf{Q}^*\mathbf{Q} = \mathbf{I}_n$ and $\mathbf{R}$ upper triangular with positive diagonal, then $\mathbf{R}$ is invertible by Lemma 3.6.2, so $\operatorname{rank}\mathbf{A} = \operatorname{rank}\mathbf{Q} = n$ and the columns were independent after all. **A QR factorization in the sense of Theorem 11.8.1 exists exactly for matrices of full column rank.** The reflection route is untroubled: a dependent column simply puts a zero on the diagonal of $\mathbf{R}$ where the dependence appears, and the algorithm continues (Exercise (C2: Reflections do not mind dependence)).

11.8.5 Givens rotations

A reflection zeroes a whole column at once. Sometimes that is too much: if $\mathbf{A}$ is nearly triangular already, we would rather kill the few offending entries one at a time and leave the rest alone.

Definition 11.8.2. Givens Rotation

Let $1 \le i < j \le m$ and let $c, s \in \mathbb{R}$ with $c^2 + s^2 = 1$. The **Givens rotation** $\mathbf{G}(i, j; c, s) \in M_m(\mathbb{R})$ is the identity matrix with the four entries in rows and columns i, j replaced by

$$\begin{pmatrix} g_{ii} & g_{ij} \\ g_{ji} & g_{jj} \end{pmatrix} = \begin{pmatrix} c & s \\ -s & c \end{pmatrix}.$$

The two columns involved are orthonormal, since $c^2 + s^2 = 1$ and $cs - sc = 0$, and every other column of $\mathbf{G}$ is a standard basis vector, orthogonal to both; so $\mathbf{G}^\top\mathbf{G} = \mathbf{I}_m$ and $\mathbf{G}$ is orthogonal, with $\det\mathbf{G} = c^2 + s^2 = 1$ by expansion along those other columns (Theorem 7.6.1). Moreover $\mathbf{G}$ acts as the identity on every coordinate other than i and j. Given a vector $\mathbf{v}$ with $(v_i, v_j) \neq (0, 0)$, set

$$r = \sqrt{v_i^2 + v_j^2}, \qquad c = \frac{v_i}{r}, \qquad s = \frac{v_j}{r}.$$

Then the i-th entry of $\mathbf{Gv}$ is $(v_i^2 + v_j^2)/r = r$ and the j-th is $(-v_j v_i + v_i v_j)/r = 0$, while every other entry is unchanged. **One rotation, one new zero.**

Here is one rotation at work. To clear the $(3,1)$-entry of the matrix $\mathbf{A}$ below, take $i = 1$, $j = 3$, and for $\mathbf{v}$ its first column, so that $(v_1, v_3) = (3, 4)$ and $r = 5$, $c = \frac{3}{5}$, $s = \frac{4}{5}$. Then $\mathbf{G} = \mathbf{G}(1, 3; \frac{3}{5}, \frac{4}{5})$ replaces row 1 by $\frac{3}{5}$(row 1) + $\frac{4}{5}$(row 3) and row 3 by $-\frac{4}{5}$(row 1) + $\frac{3}{5}$(row 3), and leaves row 2 alone:

$$\mathbf{A} = \begin{pmatrix} 3 & 5 & 0 \\ 0 & 2 & 1 \\ 4 & 0 & 5 \end{pmatrix}, \qquad \mathbf{GA} = \begin{pmatrix} 5 & 3 & 4 \\ 0 & 2 & 1 \\ 0 & -4 & 3 \end{pmatrix}.$$

The new $(1,1)$-entry is $\frac{1}{5}(3 \cdot 3 + 4 \cdot 4) = 5 = r$ and the new $(3,1)$-entry is $\frac{1}{5}(-4 \cdot 3 + 3 \cdot 4) = 0$, as designed; the zero already sitting in position $(2,1)$ survives because row 2 was never touched. A reflection would have flattened the whole first column in one step, at the price of rewriting row 2 as well.

Applied in a suitable order, rotations triangularize any real matrix, which is a second proof of Theorem 11.8.4 over $\mathbb{R}$ at the cost of more steps: a full $m \times n$ matrix needs about mn rotations where reflections need n. For a nearly triangular matrix the count reverses, which is why the Hessenberg matrices below are handled with rotations in practice. Over $\mathbb{C}$ one takes $c \in \mathbb{R}$, $s \in \mathbb{C}$ with $c^2 + |s|^2 = 1$, and $-\overline{s}$ in place of $-s$; to zero the j-th entry of $\mathbf{v}$ one puts $r = \sqrt{|v_i|^2 + |v_j|^2}$ and takes $c = |v_i|/r$ and $s = (v_i/|v_i|)\overline{v_j}/r$ when $v_i \neq 0$, and $c = 0$, $s = 1$ when $v_i = 0$.

11.8.6 Hessenberg form

So far the unitary matrices acted on one side. A similarity needs both: to preserve eigenvalues we must pass from $\mathbf{A}$ to $\mathbf{Q}^*\mathbf{A}\mathbf{Q}$, not to $\mathbf{Q}^*\mathbf{A}$. What can reflections achieve under that constraint?

Not triangularity, and the reason is worth seeing before the theorem. Suppose we reflect so that the first column becomes $\rho\mathbf{e}_1$. Multiplying on the right must not disturb column 1, so the right factor must fix $\mathbf{e}_1$ — and so must the left factor, since they are the same matrix. But a reflection fixing $\mathbf{e}_1$ acts only on the coordinates $2, \dots, n$, so in column 1 the best it can do is compress rows $2, \dots, n$ into row 2. **The entry just below the diagonal survives, by necessity.** That is exactly the pattern we can reach.

Theorem 11.8.5. Reduction to Hessenberg Form

Let $\mathbf{A} \in M_n(F)$, $n \geq 1$. Then there is a unitary $\mathbf{Q} \in M_n(F)$ such that $\mathbf{Z} = \mathbf{Q}^*\mathbf{A}\mathbf{Q}$ is **upper Hessenberg**, that is,

$$z_{ij} = 0 \qquad \text{whenever } i > j + 1.$$

The matrix $\mathbf{Q}$ can be taken to be a product of $n - 2$ factors, each a Householder reflection or an identity matrix (and $\mathbf{Q} = \mathbf{I}_n$ when $n \leq 2$). For $n = 4$ the pattern is

$$\mathbf{Z} = \begin{pmatrix} * & * & * & * \\ * & * & * & * \\ 0 & * & * & * \\ 0 & 0 & * & * \end{pmatrix}.$$

Idea. Work column by column, as in Theorem 11.8.4, but apply each reflection on both sides. At step k it is chosen to fix $\mathbf{e}_1, \dots, \mathbf{e}_k$ and to flatten the part of column k in rows $k+1, \dots, n$. Two observations do the bookkeeping. On the **left** it only recombines rows $k+1, \dots, n$, and in the earlier columns those rows are already zero, so a combination of them is still zero. On the **right** it only recombines columns $k+1, \dots, n$, so the columns already in shape are untouched.

Proof. For $0 \leq k \leq n-2$ say that $\mathbf{B} \in M_n(F)$ satisfies $(*_k)$ if $b_{ij} = 0$ whenever $j \leq k$ and $i \geq j + 2$. Every matrix satisfies $(*_0)$ vacuously, and $(*_{n-2})$ says precisely that $\mathbf{B}$ is upper Hessenberg, since the conditions with $j = n - 1$ and $j = n$ are empty.

Put $\mathbf{A}_0 = \mathbf{A}$. Let $1 \leq k \leq n - 2$ and suppose $\mathbf{A}_{k-1}$ satisfies $(*_{k-1})$. Let $\mathbf{a} \in F^{n-k}$ be the vector of entries in rows $k+1, \dots, n$ of column k of $\mathbf{A}_{k-1}$. If $\mathbf{a} = \mathbf{0}$, put $\mathbf{H}_k = \mathbf{I}_{n-k}$; otherwise let $\mathbf{H}_k \in M_{n-k}(F)$ be the reflection of Theorem 11.8.3 with $\mathbf{H}_k\mathbf{a} = \rho\mathbf{e}_1$, where

$\rho = -\alpha\|\mathbf{a}\|$. Set

$$\mathbf{Q}_k = \begin{pmatrix} \mathbf{I}_k & \mathbf{O} \\ \mathbf{O} & \mathbf{H}_k \end{pmatrix} \in M_n(F), \qquad \mathbf{A}_k = \mathbf{Q}_k \mathbf{A}_{k-1} \mathbf{Q}_k.$$

The block $\mathbf{H}_k$ is unitary and Hermitian with $\mathbf{H}_k^{-1} = \mathbf{H}_k$ — by Proposition 11.8.2 (a)–(c) for a reflection, and because $\mathbf{I}_{n-k}^* = \mathbf{I}_{n-k} = \mathbf{I}_{n-k}^{-1}$ otherwise — so by Theorem 8.1.1 $\mathbf{Q}_k$ is unitary and Hermitian with $\mathbf{Q}_k^{-1} = \mathbf{Q}_k$, and $\mathbf{Q}_k \mathbf{e}_j = \mathbf{e}_j$ for every $j \le k$.

Claim 1. $\mathbf{Q}_k \mathbf{A}_{k-1}$ satisfies $(*_{k-1})$, and its entries in rows $i \ge k+2$ of column k are 0.

Proof. By Theorem 8.1.1, the rows $1, \dots, k$ of $\mathbf{Q}_k \mathbf{A}_{k-1}$ are those of $\mathbf{A}_{k-1}$, and each of the rows $k+1, \dots, n$ is a linear combination of the rows $k+1, \dots, n$ of $\mathbf{A}_{k-1}$. Fix a column $j \le k-1$. By $(*_{k-1})$, the entries of that column in rows $i \ge j+2$ vanish, and $k+1 \ge j+2$, so all the entries being combined are 0 and the combination is 0. Hence $(*_{k-1})$ survives. In column k, the entries in rows $k+1, \dots, n$ form the vector $\mathbf{H}_k \mathbf{a} = \rho \mathbf{e}_1$, whose entries after the first are 0; these sit in rows $k+2, \dots, n$.

■

Claim 2. $\mathbf{A}_k = (\mathbf{Q}_k \mathbf{A}_{k-1}) \mathbf{Q}_k$ has the same first k columns as $\mathbf{Q}_k \mathbf{A}_{k-1}$.

Proof. For $j \le k$, the j-th column of $\mathbf{B}\mathbf{Q}_k$ is $\mathbf{B}(\mathbf{Q}_k \mathbf{e}_j) = \mathbf{B}\mathbf{e}_j$, the j-th column of $\mathbf{B}$.

■

Combining the two claims, $\mathbf{A}_k$ satisfies $(*_k)$: columns $1, \dots, k-1$ keep the zeros guaranteed by Claim 1, and column k has zeros in rows $k+2, \dots, n$, also by Claim 1, and both survive the right multiplication by Claim 2.

Running $k = 1, \dots, n-2$ produces $\mathbf{A}_{n-2}$ satisfying $(*_{n-2})$, that is, upper Hessenberg. Put $\mathbf{Q} = \mathbf{Q}_1 \mathbf{Q}_2 \cdots \mathbf{Q}_{n-2}$, unitary as a product of unitary matrices (Proposition 11.7.4). Since each $\mathbf{Q}_k^* = \mathbf{Q}_k$,

$$\mathbf{Q}^* \mathbf{A} \mathbf{Q} = \mathbf{Q}_{n-2} \cdots \mathbf{Q}_1 \, \mathbf{A} \, \mathbf{Q}_1 \cdots \mathbf{Q}_{n-2} = \mathbf{A}_{n-2},$$

which proves the theorem. (For $n \le 2$ there are no steps and every matrix is already Hessenberg.)

■

Because $\mathbf{Q}^{-1} = \mathbf{Q}^*$, the matrices $\mathbf{A}$ and $\mathbf{Z}$ are similar (Definition 4.2.2), so they share their characteristic polynomial, their eigenvalues, and their trace. The reduction is not merely tidying: it removes $\binom{n-1}{2}$ entries and keeps **the whole spectrum**.

Example. Hessenberg form of a 3×3 matrix

Reduce

$$\mathbf{A} = \begin{pmatrix} 2 & 1 & 0 \\ 0 & 3 & 1 \\ 1 & 0 & 4 \end{pmatrix}$$

to upper Hessenberg form by a unitary similarity.

Solution. Only $k = 1$ occurs, and only the entry a_{31} has to be killed. The relevant vector is

$\mathbf{a} = (a_{21}, a_{31}) = (0,1) \in \mathbb{R}^2$, with $\|\mathbf{a}\| = 1$ and $a_1 = 0$, so $\alpha = 1$ and

$$\mathbf{w} = (0,1) + (1,0) = (1,1), \qquad \|\mathbf{w}\|^2 = 2.$$

Hence $\mathbf{H}_1 = \mathbf{I}_2 - \mathbf{w}\mathbf{w}^\top = \begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix}$, and

$$\mathbf{Q}_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & -1 & 0 \end{pmatrix}.$$

Left multiplication by $\mathbf{Q}_1$ exchanges rows 2 and 3 and negates them, and right multiplication does the same to columns 2 and 3:

$$\mathbf{Q}_1\mathbf{A} = \begin{pmatrix} 2 & 1 & 0 \\ -1 & 0 & -4 \\ 0 & -3 & -1 \end{pmatrix}, \qquad \mathbf{Q}_1\mathbf{A}\mathbf{Q}_1 = \begin{pmatrix} 2 & 0 & -1 \\ -1 & 4 & 0 \\ 0 & 1 & 3 \end{pmatrix}.$$

Entry $(3,1)$ is 0 and entry $(2,1)$ is $-1 = -\|\mathbf{a}\|$, as Theorem 11.8.3 predicts. Watch column 1 through the second multiplication: it is untouched, because $\mathbf{Q}_1\mathbf{e}_1 = \mathbf{e}_1$. Both matrices have trace 9, as similarity requires.

Warning. **Hessenberg form is not one small step from triangular form.** A few more reflections will not clear the subdiagonal, and not merely because the argument above breaks. The diagonal of a triangular matrix similar to $\mathbf{A}$ lists the eigenvalues of $\mathbf{A}$ (Theorem 9.7.4), so a finite procedure using only arithmetic and square roots that always produced one would extract the roots of an arbitrary polynomial from its coefficients by those operations alone. No such procedure exists: from degree five on, the roots of a general polynomial cannot be written with arithmetic and radicals at all. That is a theorem of Abel and Ruffini, which this book does not prove and which nothing later depends on; we quote it only to explain why the algorithms of Chapter 24 iterate instead of terminating. Over $\mathbb{C}$ a triangularizing unitary does exist (Theorem 9.7.2); what does not exist is a finite algorithm of this kind producing it. Over $\mathbb{R}$ even the existence fails: a rotation in $\mathrm{O}(2)$ with no real eigenvalue is already Hessenberg and is orthogonally similar to no real triangular matrix.

Hessenberg form is where the practical computation of eigenvalues begins. One applies the reduction once, then iterates a cheap step that preserves the pattern: factor $\mathbf{Z} = \mathbf{Q}\mathbf{R}$ and replace $\mathbf{Z}$ by $\mathbf{R}\mathbf{Q} = \mathbf{Q}^*\mathbf{Z}\mathbf{Q}$, again a unitary similarity. Chapter 24 develops that iteration and explains why every implementation starts with the reduction proved here.

Quick check. In the reduction of a 5×5 matrix, how many reflections are used, and how many entries does the second one set to zero?

Solution. Three, namely $k = 1, 2, 3$. The step $k = 2$ chooses a reflection acting on the coordinates $3, 4, 5$ and sets the entries in rows 4 and 5 of column 2 to zero: two entries. (Step 1 sets three, step 3 sets one, for $\binom{4}{2} = 6$ in total.)

11.8.7 Exercises

11.8.7.1 A. Check your understanding

Exercise. A1

(a) State Theorem 11.8.1 in full, including the two normalizations that make $\mathbf{Q}$ and $\mathbf{R}$ unique.

(b) Which hypothesis on $\mathbf{A}$ does Theorem 11.8.1 need, and which does Theorem 11.8.4 need? Explain the difference in one sentence.

(c) Write down the Householder reflection determined by $\mathbf{w}$, and list three of its properties.

(d) In Theorem 11.8.3, which of the two possible targets $\pm\|\mathbf{x}\|\mathbf{e}_1$ is chosen, and why?

(e) True or false: applying reflections on both sides, as in Theorem 11.8.5, can be pushed to make every complex matrix unitarily similar to an upper triangular one. Justify your answer.

(f) Why must the reflection used at step k of Theorem 11.8.5 fix $\mathbf{e}_1, \dots, \mathbf{e}_k$?

Solution.

(a) If the columns of $\mathbf{A} \in M_{m\times n}(F)$ are independent, then $\mathbf{A} = \mathbf{QR}$ with $\mathbf{R}$ upper triangular; the normalizations are that the columns of $\mathbf{Q}$ are **orthonormal** and that the diagonal of $\mathbf{R}$ is **positive real**. With both, the pair is unique.

(b) Theorem 11.8.1 needs independent columns; Theorem 11.8.4 needs nothing. A positive diagonal forces $\mathbf{R}$ to be invertible and hence $\operatorname{rank}\mathbf{A} = n$, while the reflection route is content to put a zero on the diagonal where a dependence occurs.

(c) $\mathbf{H}_{\mathbf{w}} = \mathbf{I}_m - 2\mathbf{w}\mathbf{w}^*/\|\mathbf{w}\|^2$ (Definition 11.8.1); it is Hermitian, unitary, equal to its own inverse, and has determinant -1 (Proposition 11.8.2).

(d) The target $-\alpha\|\mathbf{x}\|\mathbf{e}_1$, where α is the phase (over $\mathbb{R}$, the sign) of x_1 — that is, the target on the **opposite** side from x_1. The reason is arithmetic: this makes the first entry of $\mathbf{w} = \mathbf{x} + \alpha\|\mathbf{x}\|\mathbf{e}_1$ a sum of like-signed quantities, with no cancellation of leading digits.

(e) False. A reflection preserving the zeros already created must fix $\mathbf{e}_1, \dots, \mathbf{e}_k$, and then it cannot touch the entry in row $k+1$ of column k. More decisively, such a finite procedure would compute the roots of an arbitrary polynomial by arithmetic and square roots alone, which the theorem of Abel and Ruffini quoted in the warning rules out.

(f) Because the same matrix multiplies on the right, and right multiplication by $\mathbf{Q}_k$ changes column j only if $\mathbf{Q}_k\mathbf{e}_j \neq \mathbf{e}_j$. Fixing $\mathbf{e}_1, \dots, \mathbf{e}_k$ is exactly what leaves columns $1, \dots, k$, and the zeros in them, alone.

11.8.7.2 B. Practice

Exercise. B1: QR by Gram–Schmidt

Find the QR factorization of

$$\mathbf{A} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}$$

in the sense of Theorem 11.8.1, and verify that $\mathbf{R} = \mathbf{Q}^{\top}\mathbf{A}$ for the top-left 2×2 entries.

Solution. Step 1. $\mathbf{a}_1 = (1,1,1)$ has $\|\mathbf{a}_1\| = \sqrt{3}$, so $r_{11} = \sqrt{3}$ and $\mathbf{q}_1 = \frac{1}{\sqrt{3}}(1,1,1)$.

Step 2. $r_{12} = \langle \mathbf{a}_2, \mathbf{q}_1 \rangle = \frac{2}{\sqrt{3}}$, so

$$\mathbf{w}_2 = (1,1,0) - \tfrac{2}{3}(1,1,1) = \tfrac{1}{3}(1,1,-2),$$

with $\|\mathbf{w}_2\|^2 = \frac{1}{9} \cdot 6 = \frac{2}{3}$. Hence $r_{22} = \sqrt{2/3}$ and $\mathbf{q}_2 = \frac{1}{\sqrt{6}}(1,1,-2)$.

Step 3. $r_{13} = \langle \mathbf{a}_3, \mathbf{q}_1 \rangle = \frac{1}{\sqrt{3}}$ and $r_{23} = \langle \mathbf{a}_3, \mathbf{q}_2 \rangle = \frac{1}{\sqrt{6}}$, so

$$\mathbf{w}_3 = (1,0,0) - \tfrac{1}{3}(1,1,1) - \tfrac{1}{6}(1,1,-2) = (\tfrac{1}{2}, -\tfrac{1}{2}, 0),$$

with $\|\mathbf{w}_3\| = \frac{1}{\sqrt{2}} = r_{33}$ and $\mathbf{q}_3 = \frac{1}{\sqrt{2}}(1,-1,0)$. Therefore

$$\mathbf{Q} = \begin{pmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{3}} & -\frac{2}{\sqrt{6}} & 0 \end{pmatrix}, \qquad \mathbf{R} = \begin{pmatrix} \sqrt{3} & \frac{2}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ 0 & \frac{2}{\sqrt{6}} & \frac{1}{\sqrt{6}} \\ 0 & 0 & \frac{1}{\sqrt{2}} \end{pmatrix},$$

where $r_{22} = \sqrt{2/3} = 2/\sqrt{6}$. All three diagonal entries are positive, as required.

For the verification, $(\mathbf{Q}^{\top}\mathbf{A})_{11} = \langle \mathbf{a}_1, \mathbf{q}_1 \rangle = 3/\sqrt{3} = \sqrt{3} = r_{11}$ and $(\mathbf{Q}^{\top}\mathbf{A})_{12} = \langle \mathbf{a}_2, \mathbf{q}_1 \rangle = 2/\sqrt{3} = r_{12}$; also $(\mathbf{Q}^{\top}\mathbf{A})_{21} = \langle \mathbf{a}_1, \mathbf{q}_2 \rangle = (1+1-2)/\sqrt{6} = 0 = r_{21}$ and $(\mathbf{Q}^{\top}\mathbf{A})_{22} = \langle \mathbf{a}_2, \mathbf{q}_2 \rangle = 2/\sqrt{6} = r_{22}$.

Exercise. B2: The same matrix by reflections

Let $\mathbf{A} = \begin{pmatrix} 1 & 1 \\ 2 & 1 \\ 2 & 3 \end{pmatrix}$, the matrix of Example (A 3×2 QR, and a least-squares problem).

(a) Find $\mathbf{w}_1$ and the reflection $\mathbf{H}_1$ of Theorem 11.8.3 for the first column of $\mathbf{A}$, and compute $\mathbf{H}_1\mathbf{A}$.

(b) Find the second reflection and hence a factorization $\mathbf{A} = \mathbf{QR}$ with $\mathbf{Q} \in M_3(\mathbb{R})$ orthogonal and $\mathbf{R} \in M_{3\times 2}(\mathbb{R})$ upper triangular.

(c) Compare the $\mathbf{R}$ of (b) with the $\mathbf{R}$ of Example (A 3×2 QR, and a least-squares problem), and say exactly how the two factorizations differ.

Solution.

(a) The first column is $\mathbf{a} = (1,2,2)$ with $\|\mathbf{a}\| = 3$ and $a_1 = 1 > 0$, so $\alpha = 1$ and $\mathbf{w}_1 = (1,2,2) + 3\mathbf{e}_1 = (4,2,2)$, with $\|\mathbf{w}_1\|^2 = 24$. By (‡), $\mathbf{H}_1\mathbf{v} = \mathbf{v} - \frac{1}{12}\langle \mathbf{v}, \mathbf{w}_1 \rangle \mathbf{w}_1$. Since

$\langle \mathbf{a}, \mathbf{w}_1 \rangle = 4+4+4 = 12$ and $\langle (1,1,3), \mathbf{w}_1 \rangle = 4+2+6 = 12$, each column loses one copy of $\mathbf{w}_1$:

$$\mathbf{H}_1\mathbf{A} = \begin{pmatrix} -3 & -3 \\ 0 & -1 \\ 0 & 1 \end{pmatrix}.$$

(b) The trailing vector is $\mathbf{y} = (-1, 1)$, with $\|\mathbf{y}\| = \sqrt{2}$ and $y_1 = -1 < 0$, so $\alpha = -1$ and $\mathbf{w}_2 = (-1,1) - \sqrt{2}(1,0) = (-1-\sqrt{2}, 1)$. By Theorem 11.8.3, the reflection sends $\mathbf{y}$ to $+\sqrt{2}\,\mathbf{e}_1$. Embedding it as $\mathbf{H}_2 = \mathbf{I}_1 \oplus \mathbf{H}_{\mathbf{w}_2}$, we get

$$\mathbf{R} = \mathbf{H}_2\mathbf{H}_1\mathbf{A} = \begin{pmatrix} -3 & -3 \\ 0 & \sqrt{2} \\ 0 & 0 \end{pmatrix}, \qquad \mathbf{Q} = \mathbf{H}_1\mathbf{H}_2,$$

and $\mathbf{A} = \mathbf{QR}$ because each $\mathbf{H}_i$ is its own inverse (Proposition 11.8.2 (b)).

(c) In Example (A 3×2 QR, and a least-squares problem), $\mathbf{R}_0 = \begin{pmatrix} 3 & 3 \\ 0 & \sqrt{2} \end{pmatrix}$. Deleting the zero row of $\mathbf{R}$ leaves $\mathbf{R}_0$ with its **first row negated**, and correspondingly the first column of $\mathbf{Q}$ is $-\mathbf{q}_1$ while the second is $\mathbf{q}_2$. The reflection route did not aim for a positive diagonal, and by the uniqueness in Theorem 11.8.1 signs are the only discrepancy possible.

Exercise. B3: A least-squares fit through QR

Fit a line $y = c_0 + c_1 t$ to the data $(t, y) = (0,1), (1,3), (2,2)$ in the least-squares sense, by finding the QR factorization of the design matrix and solving $\mathbf{Rx} = \mathbf{Q}^{\top}\mathbf{b}$. Compute the residual and check the answer against the normal equations.

Solution. The system is $\mathbf{Ax} = \mathbf{b}$ with

$$\mathbf{A} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{pmatrix}, \qquad \mathbf{b} = \begin{pmatrix} 1 \\ 3 \\ 2 \end{pmatrix}, \qquad \mathbf{x} = (c_0, c_1).$$

Gram–Schmidt: $\|\mathbf{a}_1\| = \sqrt{3}$, so $r_{11} = \sqrt{3}$ and $\mathbf{q}_1 = \frac{1}{\sqrt{3}}(1,1,1)$. Then $r_{12} = \langle \mathbf{a}_2, \mathbf{q}_1 \rangle = 3/\sqrt{3} = \sqrt{3}$ and

$$\mathbf{w}_2 = (0,1,2) - (1,1,1) = (-1,0,1),$$

so $r_{22} = \sqrt{2}$ and $\mathbf{q}_2 = \frac{1}{\sqrt{2}}(-1,0,1)$. Now $\langle \mathbf{b}, \mathbf{q}_1 \rangle = 6/\sqrt{3} = 2\sqrt{3}$ and $\langle \mathbf{b}, \mathbf{q}_2 \rangle = (-1+0+2)/\sqrt{2} = 1/\sqrt{2}$, so (†) reads

$$\begin{pmatrix} \sqrt{3} & \sqrt{3} \\ 0 & \sqrt{2} \end{pmatrix} \mathbf{x} = \begin{pmatrix} 2\sqrt{3} \\ 1/\sqrt{2} \end{pmatrix}.$$

Back substitution gives $c_1 = \frac{1}{2}$ and then $\sqrt{3}(c_0 + \frac{1}{2}) = 2\sqrt{3}$, so $c_0 = \frac{3}{2}$. The fitted line is $y = \frac{3}{2} + \frac{1}{2}t$, with residual $\mathbf{Ax} - \mathbf{b} = (\frac{1}{2}, -1, \frac{1}{2})$ and $\|\mathbf{Ax} - \mathbf{b}\|^2 = \frac{3}{2}$.

Check. $\mathbf{A}^{\top}\mathbf{A} = \begin{pmatrix} 3 & 3 \\ 3 & 5 \end{pmatrix}$ and $\mathbf{A}^{\top}\mathbf{b} = (6, 7)$. Subtracting the first normal equation from the second gives $2c_1 = 1$, and then $3c_0 + \frac{3}{2} = 6$, so $c_0 = \frac{3}{2}$. The residual is orthogonal to both

columns of $\mathbf{A}$: $\frac{1}{2} - 1 + \frac{1}{2} = 0$ and $0 \cdot \frac{1}{2} - 1 \cdot 1 + 2 \cdot \frac{1}{2} = 0$, which is what the normal equations of Theorem 11.4.2 say.

Exercise. B4: Determine which are QR factorizations

Determine which of the following are QR factorizations of the matrix on the left in the sense of Theorem 11.8.1. Justify your answer; for those that are not, name the clause that fails and repair it if possible.

(a) $\begin{pmatrix} 3 & 1 \\ 4 & 7 \end{pmatrix} = \frac{1}{5}\begin{pmatrix} 3 & -4 \\ 4 & 3 \end{pmatrix}\begin{pmatrix} 5 & \frac{31}{5} \\ 0 & \frac{17}{5} \end{pmatrix}$.

(b) $\begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} 1 & \frac{1}{2} \\ 0 & \frac{1}{2} \end{pmatrix}$.

(c) $\begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}$.

Solution.

(a) Yes. The product is

$$\frac{1}{5}\begin{pmatrix} 15 & \frac{93}{5} - \frac{68}{5} \\ 20 & \frac{124}{5} + \frac{51}{5} \end{pmatrix} = \frac{1}{5}\begin{pmatrix} 15 & 5 \\ 20 & 35 \end{pmatrix} = \begin{pmatrix} 3 & 1 \\ 4 & 7 \end{pmatrix},$$

the columns of the left factor are $\frac{1}{5}(3,4)$ and $\frac{1}{5}(-4,3)$, both of length 1 and orthogonal, and the diagonal entries 5 and $\frac{17}{5}$ are positive. All three clauses hold, so by the uniqueness in Theorem 11.8.1 this is **the** QR factorization.

(b) No. The identity is correct, and $\mathbf{R}$ is upper triangular with positive diagonal, but the columns of the left factor have length $\sqrt{2}$, not 1, so clause (a) of Theorem 11.8.1 fails. Repair it by dividing each column of $\mathbf{Q}$ by $\sqrt{2}$ and multiplying the matching row of $\mathbf{R}$ by $\sqrt{2}$:

$$\mathbf{Q} = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}, \qquad \mathbf{R} = \begin{pmatrix} \sqrt{2} & \frac{1}{\sqrt{2}} \\ 0 & \frac{1}{\sqrt{2}} \end{pmatrix}.$$

(c) No. The identity is correct and the columns of $\mathbf{Q} = \mathbf{I}_2$ are orthonormal, but $r_{11} = -1 < 0$, so clause (b) fails. Repair it by moving the sign across: $\mathbf{Q} = \operatorname{diag}(-1, 1)$, whose columns are still orthonormal, and $\mathbf{R} = \mathbf{I}_2$, whose diagonal is positive. This is the move described in Exercise (C1: All the QR factorizations of a matrix) with $\mathbf{D} = \operatorname{diag}(-1, 1)$.

11.8.7.3 C. Going deeper

Exercise. C1: All the QR factorizations of a matrix

Let $\mathbf{A} \in M_{m\times n}(F)$ have independent columns, with QR factorization $\mathbf{A} = \mathbf{QR}$ as in Theorem 11.8.1.

(a) Prove that an upper triangular $\mathbf{S} \in M_n(F)$ with $\mathbf{S}^*\mathbf{S} = \mathbf{I}_n$ is diagonal, with diagonal entries of modulus 1.

(b) Deduce that the pairs $(\mathbf{Q}_1, \mathbf{R}_1)$ with $\mathbf{Q}_1^*\mathbf{Q}_1 = \mathbf{I}_n$, $\mathbf{R}_1$ upper triangular and invertible, and $\mathbf{A} = \mathbf{Q}_1\mathbf{R}_1$, are exactly the pairs $(\mathbf{Q}\mathbf{D}^{-1}, \mathbf{D}\mathbf{R})$ for $\mathbf{D} = \operatorname{diag}(d_1, \dots, d_n)$ with $|d_k| = 1$.

(c) Hence give a second proof of the uniqueness statement of Theorem 11.8.1.

Solution.

(a) Let $\mathbf{s}_1, \dots, \mathbf{s}_n$ be the columns of $\mathbf{S}$; as in the proof of Theorem 11.8.1, $\mathbf{S}^*\mathbf{S} = \mathbf{I}_n$ says $\langle \mathbf{s}_j, \mathbf{s}_i \rangle = \delta_{ij}$. We show $\mathbf{s}_j = s_{jj}\mathbf{e}_j$ with $|s_{jj}| = 1$, by induction on j. Suppose this holds for every $i < j$. For such i,

$$0 = \langle \mathbf{s}_j, \mathbf{s}_i \rangle = \overline{s_{ii}}\langle \mathbf{s}_j, \mathbf{e}_i \rangle = \overline{s_{ii}}\, s_{ij},$$

and $s_{ii} \neq 0$, so $s_{ij} = 0$. Together with upper triangularity this gives $\mathbf{s}_j = s_{jj}\mathbf{e}_j$, and $\|\mathbf{s}_j\| = 1$ gives $|s_{jj}| = 1$.

(b) ($\Leftarrow$) If $|d_k| = 1$ for all k, then $\mathbf{D}^*\mathbf{D} = \mathbf{I}_n$, so $(\mathbf{Q}\mathbf{D}^{-1})^*(\mathbf{Q}\mathbf{D}^{-1}) = (\mathbf{D}^{-1})^*\mathbf{Q}^*\mathbf{Q}\mathbf{D}^{-1} = (\mathbf{D}\mathbf{D}^*)^{-1} = \mathbf{I}_n$; the matrix $\mathbf{D}\mathbf{R}$ is upper triangular and invertible; and $(\mathbf{Q}\mathbf{D}^{-1})(\mathbf{D}\mathbf{R}) = \mathbf{Q}\mathbf{R} = \mathbf{A}$.
($\Rightarrow$) Suppose $\mathbf{A} = \mathbf{Q}_1\mathbf{R}_1$ as stated. Exactly as in the proof of Theorem 11.8.1, $\mathbf{R}_1^*\mathbf{R}_1 = \mathbf{A}^*\mathbf{A} = \mathbf{R}^*\mathbf{R}$, and $\mathbf{S} = \mathbf{R}_1\mathbf{R}^{-1}$ is upper triangular with $\mathbf{S}^*\mathbf{S} = \mathbf{I}_n$. By (a), $\mathbf{S} = \mathbf{D}$ is diagonal with $|d_k| = 1$. Then $\mathbf{R}_1 = \mathbf{D}\mathbf{R}$ and $\mathbf{Q}_1 = \mathbf{A}\mathbf{R}_1^{-1} = \mathbf{Q}\mathbf{R}\mathbf{R}^{-1}\mathbf{D}^{-1} = \mathbf{Q}\mathbf{D}^{-1}$.

(c) If in addition both $\mathbf{R}$ and $\mathbf{R}_1 = \mathbf{D}\mathbf{R}$ have positive real diagonal entries, then $d_k = (\mathbf{R}_1)_{kk}/(\mathbf{R})_{kk}$ is a positive real number of modulus 1, so $d_k = 1$ for every k. Hence $\mathbf{D} = \mathbf{I}_n$, $\mathbf{R}_1 = \mathbf{R}$, and $\mathbf{Q}_1 = \mathbf{Q}$.

Exercise. C2: Reflections do not mind dependence

Let

$$\mathbf{A} = \begin{pmatrix} 1 & 1 & 2 \\ 2 & 1 & 3 \\ 2 & 3 & 5 \end{pmatrix},$$

whose third column is the sum of the first two.

(a) Run the Gram–Schmidt process of Theorem 11.2.2 on the columns of $\mathbf{A}$ and say precisely which step fails.

(b) Carry out the Householder reduction of Theorem 11.8.4 on $\mathbf{A}$ and exhibit $\mathbf{R}$.

(c) Explain, using Theorem 11.8.1, why no factorization $\mathbf{A} = \mathbf{Q}\mathbf{R}$ with $\mathbf{Q}^*\mathbf{Q} = \mathbf{I}_3$ and $\mathbf{R}$ upper triangular with positive diagonal can exist.

Solution.

(a) Steps 1 and 2 go through: $\mathbf{q}_1 = \frac{1}{3}(1,2,2)$ and, since $\langle \mathbf{a}_2, \mathbf{q}_1 \rangle = \frac{1}{3}(1+2+6) = 3$, $\mathbf{w}_2 = (1,1,3) - (1,2,2) = (0,-1,1)$ and $\mathbf{q}_2 = \frac{1}{\sqrt{2}}(0,-1,1)$. At step 3, $\mathbf{a}_3 = \mathbf{a}_1 + \mathbf{a}_2$ lies in $\operatorname{span}(\mathbf{a}_1, \mathbf{a}_2) = \operatorname{span}(\mathbf{q}_1, \mathbf{q}_2)$, so by Theorem 11.2.1 (a) the vector subtracted from $\mathbf{a}_3$ is $\mathbf{a}_3$ itself and $\mathbf{w}_3 = \mathbf{0}$. The normalization $\mathbf{q}_3 = \mathbf{w}_3/\|\mathbf{w}_3\|$ divides by 0 and fails. This is the situation of Exercise (C2: Gram–Schmidt on a dependent list).

(b) The first column is $(1,2,2)$, so as in Exercise (B2: The same matrix by reflections) we

take $\mathbf{w}_1 = (4,2,2)$ with $\|\mathbf{w}_1\|^2 = 24$, and $\mathbf{H}_1\mathbf{v} = \mathbf{v} - \frac{1}{12}\langle \mathbf{v}, \mathbf{w}_1\rangle \mathbf{w}_1$. The inner products of the three columns with $\mathbf{w}_1$ are 12, 12, and 24, so

$$\mathbf{H}_1\mathbf{A} = \begin{pmatrix} -3 & -3 & -6 \\ 0 & -1 & -1 \\ 0 & 1 & 1 \end{pmatrix}.$$

The trailing vector of column 2 is $(-1,1)$, which by Exercise (B2: The same matrix by reflections) (b) the second reflection sends to $\sqrt{2}\,\mathbf{e}_1$; it sends the trailing vector of column 3, which is the same vector $(-1,1)$, to the same place. Hence

$$\mathbf{R} = \begin{pmatrix} -3 & -3 & -6 \\ 0 & \sqrt{2} & \sqrt{2} \\ 0 & 0 & 0 \end{pmatrix},$$

upper triangular, with a zero in position $(3,3)$ exactly where the dependence sits. No step of the algorithm divided by zero.

(c) Such an $\mathbf{R}$ would be invertible by Lemma 3.6.2, so $\operatorname{rank}\mathbf{A} = \operatorname{rank}(\mathbf{QR}) = \operatorname{rank}\mathbf{Q} = 3$. But $\mathbf{a}_3 = \mathbf{a}_1 + \mathbf{a}_2$, so the columns are dependent and $\operatorname{rank}\mathbf{A} \le 2$, a contradiction. The positive-diagonal factorization exists precisely for matrices of full column rank.

Exercise. C3: Triangularizing with rotations

(a) Let $(a,b) \in \mathbb{R}^2$ be nonzero and $r = \sqrt{a^2+b^2}$. Verify that the Givens rotation $\mathbf{G}$ of Definition 11.8.2 with $c = a/r$ and $s = b/r$ satisfies $\mathbf{G}(a,b) = (r,0)$, and that $\mathbf{G} \in \mathrm{SO}(2)$.

(b) Use (a) to write $\mathbf{A} = \begin{pmatrix} 3 & 1 \\ 4 & 2 \end{pmatrix}$ as $\mathbf{QR}$ with $\mathbf{Q} \in \mathrm{SO}(2)$ and $\mathbf{R}$ upper triangular with positive diagonal.

(c) How many Givens rotations does it take to triangularize a general 3×3 real matrix, and which entries does each one kill? Compare with the number of Householder reflections.

Solution.

(a) $\mathbf{G}(a,b) = (ca+sb,\ -sa+cb) = ((a^2+b^2)/r,\ (-ab+ab)/r) = (r,0)$. Also $c^2+s^2 = 1$, so the columns of $\mathbf{G}$ are orthonormal and $\det\mathbf{G} = c^2+s^2 = 1$; hence $\mathbf{G} \in \mathrm{SO}(2)$.

(b) The first column is $(3,4)$, so $r = 5$, $c = \frac{3}{5}$, $s = \frac{4}{5}$, and $\mathbf{G} = \frac{1}{5}\begin{pmatrix} 3 & 4 \\ -4 & 3 \end{pmatrix}$. Then

$$\mathbf{GA} = \tfrac{1}{5}\begin{pmatrix} 9+16 & 3+8 \\ -12+12 & -4+6 \end{pmatrix} = \begin{pmatrix} 5 & \frac{11}{5} \\ 0 & \frac{2}{5} \end{pmatrix} =: \mathbf{R},$$

whose diagonal entries 5 and $\frac{2}{5}$ are positive. Since $\mathbf{G}^{-1} = \mathbf{G}^{\top} \in \mathrm{SO}(2)$, we get $\mathbf{A} = \mathbf{QR}$ with $\mathbf{Q} = \mathbf{G}^{\top} = \frac{1}{5}\begin{pmatrix} 3 & -4 \\ 4 & 3 \end{pmatrix}$. As a check, $\det\mathbf{A} = 2 = 5\cdot\frac{2}{5} = \det\mathbf{R}$.

(c) Three: one for a_{21} (rows $1,2$), one for a_{31} (rows $1,3$), and one for the new $(3,2)$-entry (rows $2,3$). The last acts on rows 2 and 3, which by then have zeros in column 1, so it does not undo the earlier work. Householder needs two reflections, one per column,

since each clears a whole column at once.

11.9 Bessel, Parseval, and Fourier Series

Section 3 measured a vector against a finite-dimensional subspace and produced the nearest point in it. Read that answer coefficient by coefficient and it becomes an inequality: the numbers $\langle v, e_i \rangle$ collected against an orthonormal list can never carry more squared length than v itself has. That is Bessel's inequality. It costs nothing in dimension, it holds in spaces far too large for Chapter 1's machinery, and its equality case is Parseval's identity. This section proves both, applies them to the sines and cosines, and then stops at the exact place where linear algebra stops and analysis begins.

Throughout, V is an inner product space over $F = \mathbb{R}$ or $F = \mathbb{C}$, with inner product $\langle \cdot, \cdot \rangle$ (Definition 11.1.1) and induced norm $\|\cdot\|$ (Definition 11.1.3). **We do not assume that V is finite-dimensional.** What we assume each time is that the *list* we compute against is finite.

11.9.1 Coefficients against an orthonormal list

In Theorem 11.2.1 the list was an orthonormal **basis**, and the coefficients $\langle v, e_i \rangle$ recovered v exactly. Suppose now the list is only a list: too short to be a basis, or sitting inside a space that has no finite basis at all, like the continuous functions on an interval. The numbers $\langle v, e_i \rangle$ are still there. What can be said about them when they no longer determine v?

Against a shorter list you can only lose length, never gain it.

Definition 11.9.1. Fourier Coefficients With Respect to an Orthonormal List

Let $(e_1, \dots, e_k)$ be an orthonormal list in an inner product space V and let $v \in V$. The scalars

$$\langle v, e_1 \rangle, \ \dots, \ \langle v, e_k \rangle \in F$$

are the **Fourier coefficients** of v with respect to the list.

The name is borrowed from the classical case below, where the e_i are sines and cosines; nothing in the definition mentions trigonometry. When the list happens to be a basis, these are exactly the coordinates of v, by Theorem 11.2.1 (a). When it is not, they are the coordinates of a *shadow* of v, and the next theorem measures that shadow.

Theorem 11.9.1. Bessel's Inequality

Let V be an inner product space over F, **not assumed finite-dimensional**, and let $(e_1, \dots, e_k)$ be an orthonormal list in V. Then for **every** $v \in V$,

$$\sum_{i=1}^{k} |\langle v, e_i \rangle|^2 \ \leq \ \|v\|^2.$$

Idea. The left-hand side is not a quantity out of nowhere. Let U be the span of the list. By the projection formula, $P_U v$ is exactly $\sum_i \langle v, e_i \rangle e_i$, and its summands are pairwise orthogonal, so Pythagoras turns its squared length into precisely that sum. The inequality then says no

more than "a shadow is no longer than what casts it", which we already proved. The one point needing care is that V may be infinite-dimensional; it does not matter, because Section 3 only ever asked the **subspace** to be finite-dimensional.

Proof. Put $U = \operatorname{span}(e_1, \dots, e_k)$. The vectors e_i are orthogonal and nonzero, so the list is linearly independent by Theorem 11.1.4; being a spanning list of U as well, it is an orthonormal basis of U, and $\dim U = k$ is finite. Hence Theorem 11.3.3 applies to U even though V need not be finite-dimensional, and part (a) of it gives

$$P_U v = \sum_{i=1}^{k} \langle v, e_i \rangle \, e_i .$$

The summands $\langle v, e_i \rangle e_i$ are pairwise orthogonal, since $e_i \perp e_j$ for $i \neq j$ and scalars come out of both slots. So Theorem 11.1.3 and Theorem 11.1.2 (b) give

$$\|P_U v\|^2 = \sum_{i=1}^{k} |\langle v, e_i \rangle|^2 \|e_i\|^2 = \sum_{i=1}^{k} |\langle v, e_i \rangle|^2 ,$$

using $\|e_i\| = 1$. By Theorem 11.3.3 (c), $\|P_U v\| \leq \|v\|$. Both sides are nonnegative, so squaring preserves the inequality, and this proves the theorem.

■

The proof gives more than the inequality, and the surplus is worth recording. The vector $v - P_U v$ lies in $U^{\perp}$ by Theorem 11.3.2, so it is orthogonal to $P_U v \in U$, and Theorem 11.1.3 turns $v = P_U v + (v - P_U v)$ into $\|v\|^2 = \|P_U v\|^2 + \|v - P_U v\|^2$. Rearranging, and substituting $\|P_U v\|^2 = \sum_{i=1}^{k} |\langle v, e_i \rangle|^2$ from the proof above,

$$\|v - P_U v\|^2 = \|v\|^2 - \sum_{i=1}^{k} |\langle v, e_i \rangle|^2 . \tag{$*$}$$

In words: **the deficit in Bessel's inequality is the squared distance from v to the span of the list** (Definition 11.3.3). An inequality with a known defect is much more useful than an inequality, and we will read $(*)$ from both sides.

Remark. For an infinite orthonormal list $(e_1, e_2, e_3, \dots)$, apply the theorem to the first k terms for each k. The partial sums $\sum_{i \leq k} |\langle v, e_i \rangle|^2$ increase with k and are all bounded above by $\|v\|^2$, so they converge, and $\sum_{i=1}^{\infty} |\langle v, e_i \rangle|^2 \leq \|v\|^2$. This is the only fact from analysis used in the present paragraph: a bounded increasing sequence of real numbers converges.

Now the equality case, which is where the word *basis* returns.

Theorem 11.9.2. Parseval's Identity

Let $(e_1, \dots, e_k)$ be an orthonormal list in an inner product space V and let $v \in V$. Then

$$\sum_{i=1}^{k} |\langle v, e_i \rangle|^2 = \|v\|^2 \quad \text{if and only if} \quad v \in \operatorname{span}(e_1, \dots, e_k).$$

In particular, if $(e_1, \dots, e_k)$ is an orthonormal **basis** of V, the identity holds for every $v \in V$.

Proof. Keep the notation of the proof of Theorem 11.9.1, where we showed $\sum_i |\langle v, e_i\rangle|^2 = \|P_U v\|^2$ with $U = \operatorname{span}(e_1, \dots, e_k)$. So the stated equality says $\|P_U v\| = \|v\|$, and by Theorem 11.3.3 (c) this holds if and only if $v \in U$. If the list is a basis of V, then $U = V$ and every $v \in V$ lies in U.

■

For an orthonormal basis of a finite-dimensional space this is Theorem 11.2.1 (c) again; what is new is the equality **case**, which identifies exactly the vectors for which no length is lost. Keep the two statements apart:

- **Bessel** is about an orthonormal **list**, in **any** inner product space, and it is an inequality.
- **Parseval** is the equality, and it holds exactly for the vectors the list already spans.

The whole difference is whether the list reaches v.

Warning. **An orthonormal list is not an orthonormal basis, and the deficit in Bessel can be everything.** In $C[-\pi, \pi]$ below, the list $(\cos x, \sin x)$ is orthonormal, and for the constant function $f = 1$ both coefficients are 0, so the left-hand side of Bessel is 0 while $\|f\|^2 = 2$. Nothing at all is captured. A short orthonormal list is not a bad basis; it is not a basis.

Quick check. In $\mathbb{R}^4$ with the dot product, let $\mathbf{e}_1 = \frac{1}{2}(1,1,1,1)$ and $\mathbf{e}_2 = \frac{1}{2}(1,1,-1,-1)$, and let $\mathbf{v} = (1,2,3,4)$. Check that $(\mathbf{e}_1, \mathbf{e}_2)$ is orthonormal, compute both sides of Bessel's inequality, and say what the deficit equals.

Solution. Each has norm $\frac{1}{2}\sqrt{4} = 1$, and $\langle \mathbf{e}_1, \mathbf{e}_2\rangle = \frac{1}{4}(1+1-1-1) = 0$, so the list is orthonormal. The coefficients are $\langle \mathbf{v}, \mathbf{e}_1\rangle = \frac{1}{2}(1+2+3+4) = 5$ and $\langle \mathbf{v}, \mathbf{e}_2\rangle = \frac{1}{2}(1+2-3-4) = -2$, so the left-hand side is $25 + 4 = 29$, while $\|\mathbf{v}\|^2 = 1 + 4 + 9 + 16 = 30$. The inequality is strict, and by $(*)$ the deficit 1 is the squared distance from $\mathbf{v}$ to $\operatorname{span}(\mathbf{e}_1, \mathbf{e}_2)$. Indeed $P_U\mathbf{v} = 5\mathbf{e}_1 - 2\mathbf{e}_2 = (\frac{3}{2}, \frac{3}{2}, \frac{7}{2}, \frac{7}{2})$ and $\mathbf{v} - P_U\mathbf{v} = \frac{1}{2}(-1,1,-1,1)$, whose squared norm is $4 \cdot \frac{1}{4} = 1$.

11.9.2 The trigonometric system

The classical home of these two statements is a space of functions, and the first thing to fix is the inner product, constant included.

Let $V = C[-\pi, \pi]$ be the space of continuous functions $f\colon [-\pi, \pi] \to \mathbb{R}$, a real vector space, and define

$$\langle f, g\rangle = \frac{1}{\pi} \int_{-\pi}^{\pi} f(t)\, g(t)\, \mathrm{d}t.$$

This is the continuous-function inner product of Definition 11.1.1 on a different interval and rescaled by the positive constant $1/\pi$; rescaling by a positive constant preserves all three axioms,

and positive definiteness is the argument given there. The constant is chosen for one reason only: it makes the sines and cosines unit vectors. With $1/(2\pi)$ in front they would have norm $1/\sqrt{2}$, and with no constant at all, norm $\sqrt{\pi}$.

Warning. **Always state the normalizing constant before computing.** Formulas for Fourier coefficients differ between books by exactly such a factor, and a misplaced π or 2 changes every number below. In this section $\langle f, g\rangle = \frac{1}{\pi}\int_{-\pi}^{\pi} fg$, and every coefficient, norm, and inequality is written for that choice.

Here are the facts we assume. From calculus: the integral of a finite sum is the sum of the integrals; $\int_{-\pi}^{\pi} \mathrm{d}t = 2\pi$; $\int_{-\pi}^{\pi} \sin(mt)\,\mathrm{d}t = 0$ for every integer m and $\int_{-\pi}^{\pi} \cos(mt)\,\mathrm{d}t = 0$ for every **nonzero** integer m (for $m \neq 0$ evaluate the antiderivative $-\cos(mt)/m$, respectively $\sin(mt)/m$, each of which takes the same value at $t = -\pi$ as at $t = \pi$; for $m = 0$ the sine integrand is 0); integration by parts, and the substitution $t \mapsto -t$, both of which we use later in the section when computing particular Fourier coefficients. From trigonometry: the product-to-sum identities. **No theorem about convergence is used anywhere in this section**; every integral we compute is the integral of a finite sum of sines and cosines.

Example. The Trigonometric System Is Orthonormal

Fix $n \geq 1$. Show that the list of $2n + 1$ functions

$$\frac{1}{\sqrt{2}},\ \cos x,\ \sin x,\ \cos 2x,\ \sin 2x,\ \ldots,\ \cos nx,\ \sin nx$$

is orthonormal in $V = C[-\pi, \pi]$ with $\langle f, g\rangle = \frac{1}{\pi}\int_{-\pi}^{\pi} fg$.

Solution. There are five kinds of pair. Write j, k for indices in $\{1, \ldots, n\}$.

The constant with itself. $\|1/\sqrt{2}\|^2 = \frac{1}{\pi}\int_{-\pi}^{\pi} \frac{1}{2}\,\mathrm{d}t = \frac{1}{\pi}\cdot\pi = 1$.

Constant against $\cos kx$ *or* $\sin kx$. These are $\frac{1}{\pi\sqrt{2}}\int_{-\pi}^{\pi} \cos(kt)\,\mathrm{d}t = 0$ and $\frac{1}{\pi\sqrt{2}}\int_{-\pi}^{\pi} \sin(kt)\,\mathrm{d}t = 0$, since $k \neq 0$.

Two cosines. By $\cos A\cos B = \frac{1}{2}[\cos(A-B) + \cos(A+B)]$,

$$\langle \cos jx, \cos kx\rangle = \frac{1}{2\pi}\int_{-\pi}^{\pi} \big[\cos((j-k)t) + \cos((j+k)t)\big]\mathrm{d}t.$$

If $j \neq k$, both $j - k$ and $j + k$ are nonzero integers, so both integrals vanish and the value is 0. If $j = k$, the first integrand is the constant 1 and the second has nonzero frequency $2k$, so the value is $\frac{1}{2\pi}(2\pi + 0) = 1$.

Two sines. By $\sin A\sin B = \frac{1}{2}[\cos(A-B) - \cos(A+B)]$, the same two integrals appear with a minus sign, giving 0 for $j \neq k$ and $\frac{1}{2\pi}(2\pi - 0) = 1$ for $j = k$.

A sine against a cosine. By $\sin A\cos B = \frac{1}{2}[\sin(A+B) + \sin(A-B)]$,

$$\langle \sin jx, \cos kx\rangle = \frac{1}{2\pi}\int_{-\pi}^{\pi} \big[\sin((j+k)t) + \sin((j-k)t)\big]\mathrm{d}t = 0,$$

because every integral of a sine over $[-\pi,\pi]$ vanishes, including the case $j=k$ where the second integrand is $\sin 0 = 0$.

So each of the $2n+1$ functions is a unit vector and distinct ones are orthogonal, which is Definition 11.2.1.

Write $W_n = \operatorname{span}(1, \cos x, \sin x, \dots, \cos nx, \sin nx)$, the space of **trigonometric polynomials of degree at most** n. The list above is an orthonormal list, hence independent (Theorem 11.1.4), and it spans W_n, since $1 = \sqrt{2}\cdot\frac{1}{\sqrt{2}}$. So it is an orthonormal basis of W_n and $\dim W_n = 2n+1$.

For $f \in V$ the classical **Fourier coefficients** of f are

$$a_k = \frac{1}{\pi}\int_{-\pi}^{\pi} f(t)\cos(kt)\,dt, \qquad b_k = \frac{1}{\pi}\int_{-\pi}^{\pi} f(t)\sin(kt)\,dt,$$

for $k \geq 0$ and $k \geq 1$ respectively. Comparing with the inner product, $a_k = \langle f, \cos kx\rangle$ and $b_k = \langle f, \sin kx\rangle$, while the coefficient against the constant of the orthonormal list is $\langle f, 1/\sqrt{2}\rangle = a_0/\sqrt{2}$. These are exactly the Fourier coefficients of Definition 11.9.1 for this list.

Theorem 11.9.3. Fourier Partial Sums Are Best Approximations

Let $f \in C[-\pi,\pi]$, let $n \geq 1$, and let a_k, b_k be the Fourier coefficients of f. Put

$$S_n f = \frac{a_0}{2} + \sum_{k=1}^{n} (a_k \cos kx + b_k \sin kx).$$

Then, with $\langle f, g\rangle = \frac{1}{\pi}\int_{-\pi}^{\pi} fg$:

(a) $S_n f = P_{W_n} f$;

(b) $\|f - S_n f\| \leq \|f - g\|$ for every $g \in W_n$, with equality **only** for $g = S_n f$;

(c) $\dfrac{a_0^2}{2} + \displaystyle\sum_{k=1}^{n} (a_k^2 + b_k^2) \leq \frac{1}{\pi}\int_{-\pi}^{\pi} f(t)^2\,dt.$

Idea. Nothing new happens here. The subspace W_n is finite-dimensional and we have an orthonormal basis of it, so Section 3 answers the approximation question and Theorem 11.9.1 answers the coefficient question; the work is only to match the classical notation to ours. The one thing to watch is the factor $\frac{1}{2}$ in front of a_0: it is not a convention pulled from the air, it is what $\langle f, 1/\sqrt{2}\rangle \cdot \frac{1}{\sqrt{2}}$ equals.

Proof.

(a) By Example (The Trigonometric System Is Orthonormal) the list

$$\left(\tfrac{1}{\sqrt{2}}, \cos x, \sin x, \dots, \cos nx, \sin nx\right)$$

is an orthonormal basis of W_n. By Theorem 11.3.3 (a),

$$P_{W_n} f = \left\langle f, \tfrac{1}{\sqrt{2}}\right\rangle \tfrac{1}{\sqrt{2}} + \sum_{k=1}^{n} (\langle f, \cos kx\rangle \cos kx + \langle f, \sin kx\rangle \sin kx).$$

The first term is $\frac{a_0}{\sqrt{2}} \cdot \frac{1}{\sqrt{2}} = \frac{a_0}{2}$, and the remaining terms are $a_k \cos kx + b_k \sin kx$. Hence $P_{W_n} f = S_n f$.

(b) This is Theorem 11.3.5 applied to the subspace $U = W_n$, which is legitimate because $\dim W_n = 2n + 1$ is finite, together with (a).

(c) This is Theorem 11.9.1 for the same orthonormal list, since $|\langle f, 1/\sqrt{2}\rangle|^2 = a_0^2/2$ and $\|f\|^2 = \frac{1}{\pi}\int_{-\pi}^{\pi} f^2$.

■

So the Fourier partial sum is not a guess at f that happens to work: it **is** the orthogonal projection of f onto the trigonometric polynomials of degree at most n, and by Theorem 11.3.5 it is the unique trigonometric polynomial of that degree closest to f in this norm. The mysterious $a_0/2$ of the classical formula is an artifact of normalizing the constant function.

11.9.3 A worked computation

Example. The Fourier Coefficients of $f(x) = x$

In $C[-\pi, \pi]$ with $\langle f, g\rangle = \frac{1}{\pi}\int_{-\pi}^{\pi} fg$, let $f(x) = x$. Find all a_k and b_k, write down $S_n f$, and say what Theorem 11.9.3 (c) gives.

Solution. We use one further calculus fact: a continuous **odd** function integrates to 0 over $[-\pi, \pi]$, and a continuous **even** one integrates to twice its integral over $[0, \pi]$. Both come from the substitution $t \mapsto -t$.

The cosine coefficients. For every $k \geq 0$ the function $t\cos(kt)$ is odd, so $a_k = 0$. In particular $a_0 = 0$.

The sine coefficients. The function $t\sin(kt)$ is even, so for $k \geq 1$,

$$b_k = \frac{1}{\pi}\int_{-\pi}^{\pi} t\sin(kt)\,\mathrm{d}t = \frac{2}{\pi}\int_{0}^{\pi} t\sin(kt)\,\mathrm{d}t.$$

Integrating by parts, with $\mathrm{d}(-\cos(kt)/k) = \sin(kt)\,\mathrm{d}t$,

$$\int_0^{\pi} t\sin(kt)\,\mathrm{d}t = \Big[\frac{-t\cos(kt)}{k}\Big]_0^{\pi} + \frac{1}{k}\int_0^{\pi}\cos(kt)\,\mathrm{d}t = \frac{-\pi\cos(k\pi)}{k},$$

since the last integral is $[\sin(kt)/k^2]_0^{\pi} = 0$. As $\cos(k\pi) = (-1)^k$, this equals $\pi(-1)^{k+1}/k$, and therefore

$$b_k = \frac{2(-1)^{k+1}}{k}.$$

The partial sums. Hence

$$S_n f = \sum_{k=1}^{n} \frac{2(-1)^{k+1}}{k}\sin kx = 2\sin x - \sin 2x + \tfrac{2}{3}\sin 3x - \dots,$$

and by Theorem 11.9.3 (b) this is the closest trigonometric polynomial of degree at most n to the function x in this norm.

What Bessel's inequality (Theorem 11.9.1) gives. First $\|f\|^2 = \frac{1}{\pi}\int_{-\pi}^{\pi} t^2\,\mathrm{d}t = \frac{1}{\pi}\cdot\frac{2\pi^3}{3} = \frac{2\pi^2}{3}$. Since $a_k = 0$ and $b_k^2 = 4/k^2$, part (c) of Theorem 11.9.3 reads

$$\sum_{k=1}^{n}\frac{4}{k^2} \le \frac{2\pi^2}{3}, \qquad \text{that is} \qquad \sum_{k=1}^{n}\frac{1}{k^2} \le \frac{\pi^2}{6}$$

for **every** n. The partial sums on the left increase and are bounded above, so they converge, and $\sum_{k=1}^{\infty} 1/k^2 \le \pi^2/6$.

That bound is sharp: the true value of the sum is $\pi^2/6$. But sharpness is not something Bessel can deliver. By Theorem 11.9.2, equality for a fixed finite list holds only when f lies in the span of that list, and the function x lies in no W_n — it is not a trigonometric polynomial, since every element of W_n takes the same value at $-\pi$ and at π. Getting the equality requires a statement about the **infinite** system, and that statement is an input from analysis; Exercise (C1: Parseval for the full trigonometric system) carries it out and flags it.

Quick check. Let $g(x) = \cos^2 x$ in $C[-\pi,\pi]$. Using $\cos^2 x = \frac{1}{2} + \frac{1}{2}\cos 2x$, write down all Fourier coefficients of g, find $S_1 g$, and compute the deficit in Bessel's inequality for $n = 1$.

Solution. The right-hand side $\frac{1}{2} + \frac{1}{2}\cos 2x$ already lies in W_2, and coefficients against an orthonormal basis are unique (Theorem 11.2.1 (a)), so, applying that result inside W_2 and reading the coefficients off: $a_0 = 1$, $a_2 = \frac{1}{2}$, and all other a_k and all b_k are 0. Hence $S_1 g = a_0/2 = \frac{1}{2}$, a constant. For the deficit, $\|g\|^2 = \frac{1}{\pi}\int_{-\pi}^{\pi}\cos^4 t\,\mathrm{d}t = \frac{1}{\pi}\cdot\frac{3\pi}{4} = \frac{3}{4}$, while the Bessel sum for $n = 1$ is $a_0^2/2 = \frac{1}{2}$. The deficit is $\frac{1}{4}$, and by $(*)$ it should be $\|g - S_1 g\|^2 = \left\|\frac{1}{2}\cos 2x\right\|^2 = \frac{1}{4}\cdot 1 = \frac{1}{4}$. It is.

11.9.4 What best approximation does and does not say

It is worth being exact about the boundary here, because it is easy to read more into Theorem 11.9.3 than it contains.

Identity $(*)$ applied to $U = W_n$ says

$$\|f - S_n f\|^2 = \|f\|^2 - \frac{a_0^2}{2} - \sum_{k=1}^{n}(a_k^2 + b_k^2).$$

The subtracted sum never decreases as n grows, so the numbers $\|f - S_n f\|^2$ never increase, and they are bounded below by 0. Hence they converge to some limit $L \ge 0$. Everything up to this sentence is linear algebra plus one fact about monotone bounded sequences.

Whether $L = 0$ is a different kind of question. It asks whether the trigonometric polynomials come arbitrarily close, in this norm, to every continuous function, and no argument in this book decides it. The statement is true, and it is the analytic content of the theory of Fourier series; it is what makes the *infinite* trigonometric system behave like an orthonormal basis even though it is not a basis in the sense of Chapter 1.

And even $L = 0$ would not be a statement about individual points. Our norm averages the error

over the whole interval, so it tolerates a large error on a short interval. The function $f(x) = x$ shows this at once:

$$S_n f(\pi) = \sum_{k=1}^{n} \frac{2(-1)^{k+1}}{k} \sin(k\pi) = 0$$

for every n, while $f(\pi) = \pi$. The error at the single point $x = \pi$ never shrinks at all.

Warning. **Mean-square best approximation is not pointwise approximation.** Theorem 11.9.3 says that $S_n f$ minimizes $\int_{-\pi}^{\pi}(f-g)^2$ over $g \in W_n$. It does not say that $S_n f(x)$ is close to $f(x)$ for a given x, and for $f(x) = x$ at $x = \pi$ it is not. Whether a Fourier series converges at a point is a question in analysis with a delicate answer; it is not decided by anything in this chapter, and a projection theorem cannot decide it.

11.9.5 Legendre and Chebyshev

Two families of orthogonal polynomials show the same two theorems at work with a different weight.

Legendre. Example (Gram–Schmidt on $1, x, x^2$) ran Gram–Schmidt on $(1, x, x^2)$ in $\mathbb{R}[x]_{\le 2}$ with $\langle p, q\rangle = \int_{-1}^{1} pq$, producing the orthogonal list $1,\ x,\ x^2 - \frac{1}{3}$ and, after normalizing, the orthonormal basis

$$e_1 = \tfrac{1}{\sqrt{2}}, \qquad e_2 = \sqrt{3/2}\,x, \qquad e_3 = \tfrac{\sqrt{10}}{4}(3x^2 - 1).$$

Inside $\mathbb{R}[x]_{\le 2}$ this is a **basis**, so Theorem 11.9.2 gives $\|p\|^2 = \sum_{i=1}^{3} |\langle p, e_i\rangle|^2$ for every quadratic p. Viewed inside the larger space $C[-1,1]$ of continuous functions, the same three vectors are only a **list**, and for an f outside $\mathbb{R}[x]_{\le 2}$ Theorem 11.9.1 is strict, the deficit being the squared distance from f to $\mathbb{R}[x]_{\le 2}$ by $(*)$. That deficit is what Example (The best quadratic approximation to a cubic) computed for $f = x^3$. One list, two readings, and which reading applies depends on the ambient space, not on the list. In practice one projects with the orthogonal list $1,\ x,\ x^2 - \frac{1}{3}$ through Corollary 11.3.4 and avoids the square roots entirely.

Chebyshev. Change the weight. On $\mathbb{R}[x]$ define

$$\langle p, q\rangle_w = \int_{-1}^{1} \frac{p(t)\,q(t)}{\sqrt{1-t^2}}\,\mathrm{d}t.$$

The integrand is unbounded at $t = \pm 1$, so this is an improper integral; it converges for polynomials, because $|pq|$ is bounded on $[-1,1]$ and $\int_{-1}^{1} \mathrm{d}t/\sqrt{1-t^2} = \pi$. The three axioms are checked as in Definition 11.1.1, positivity because $1/\sqrt{1-t^2} > 0$ on $(-1,1)$.

The **Chebyshev polynomials** are defined by the requirement

$$T_n(\cos\theta) = \cos n\theta \qquad \text{for all } \theta,$$

which does determine a polynomial of degree n, and determines it uniquely, since $\cos\theta$ runs over the whole of $[-1,1]$ and two polynomials agreeing on an infinite set are equal (Corollary 6.1.4). The first few are $T_0 = 1$, $T_1 = x$, $T_2 = 2x^2 - 1$ (from $\cos 2\theta = 2\cos^2\theta - 1$), and Exercise (C2: Chebyshev orthogonality by substitution) handles the general case. They are orthogonal for $\langle\cdot,\cdot\rangle_w$:

$$\langle T_m, T_n\rangle_w = 0 \quad (m \neq n), \qquad \langle T_n, T_n\rangle_w = \begin{cases} \pi, & n = 0, \\ \pi/2, & n \ge 1. \end{cases}$$

For $m, n \le 2$ this can be checked by hand, using $\int_{-1}^{1} \mathrm{d}t/\sqrt{1-t^2} = \pi$ and $\int_{-1}^{1} t^2\,\mathrm{d}t/\sqrt{1-t^2} = \pi/2$ (both come from $t = \cos\theta$). The pairs (T_0, T_1) and (T_1, T_2) have odd integrands on a symmetric interval, so those integrals are 0 by the substitution $t \mapsto -t$. For the remaining pair,

$$\langle T_0, T_2\rangle_w = \int_{-1}^{1} \frac{2t^2-1}{\sqrt{1-t^2}}\,\mathrm{d}t = 2\cdot\frac{\pi}{2} - \pi = 0.$$

Why should such a weight produce orthogonality? Because it is the trigonometric system in disguise. Substituting $t = \cos\theta$ turns $\langle p, q\rangle_w$ into $\int_0^{\pi} p(\cos\theta)q(\cos\theta)\,\mathrm{d}\theta$, and on the Chebyshev polynomials that integrand is $\cos m\theta\cos n\theta$ — the very product whose integral we computed in Example (The Trigonometric System Is Orthonormal). Exercise (C2: Chebyshev orthogonality by substitution) asks for this in full.

Since the T_n are orthogonal but not normalized, projections onto their span are computed with Corollary 11.3.4 rather than Theorem 11.3.3. Section 10 of this chapter takes up orthogonal polynomials as a subject in their own right, including the recurrence that generates the T_n and the quadrature rules they produce.

11.9.6 Exercises

11.9.6.1 A. Check your understanding

Exercise. A1

(a) State Bessel's inequality, saying exactly what is assumed about the space and what is assumed about the list.

(b) State Parseval's identity, including the condition under which equality holds for a given vector.

(c) True or false: Bessel's inequality requires the ambient space to be finite-dimensional. Justify your answer.

(d) Write down the inner product used on $C[-\pi, \pi]$ in this section, including its normalizing constant, and say what the constant is for.

(e) State what Theorem 11.9.3 gives about $S_n f$, and name one thing it does not give.

(f) True or false: if $(e_1, \dots, e_k)$ is an orthonormal list and $\sum_i |\langle v, e_i\rangle|^2 = \|v\|^2$ for one particular $v \neq 0$, then the list is an orthonormal basis of V. Justify your answer.

Solution.

(a) If $(e_1, \dots, e_k)$ is an orthonormal list in an inner product space V over $\mathbb{R}$ or $\mathbb{C}$, then $\sum_{i=1}^{k} |\langle v, e_i\rangle|^2 \le \|v\|^2$ for every $v \in V$ (Theorem 11.9.1). The list is assumed **finite and orthonormal**; the space is assumed only to be an inner product space, with no dimension hypothesis.

(b) Equality $\sum_i |\langle v, e_i\rangle|^2 = \|v\|^2$ holds if and only if $v \in \operatorname{span}(e_1, \dots, e_k)$; in particular it holds for every v when the list is an orthonormal basis of V (Theorem 11.9.2).

(c) False. Only the span of the list has to be finite-dimensional, and it automatically is, being spanned by k vectors. The proof uses Theorem 11.3.3, whose hypothesis is that the **subspace** projected onto is finite-dimensional. The application to $C[-\pi, \pi]$ is exactly a case where V is not finite-dimensional.

(d) $\langle f, g\rangle = \frac{1}{\pi}\int_{-\pi}^{\pi} f(t)g(t)\,\mathrm{d}t$. The constant $1/\pi$ makes $\cos kx$ and $\sin kx$ unit vectors for every $k \ge 1$; without it their norms would be $\sqrt{\pi}$.

(e) It gives that $S_n f$ is the orthogonal projection of f onto W_n, hence the unique element of W_n minimizing $\|f - g\|$, and it gives Bessel's inequality for the coefficients. It does **not** give that $S_n f(x) \to f(x)$ at any particular x, nor even that $\|f - S_n f\| \to 0$.

(f) False. Equality for one v says only that $v \in \operatorname{span}(e_1, \dots, e_k)$ (Theorem 11.9.2). In $\mathbb{R}^3$ take the list $(\mathbf{e}_1)$ and $\mathbf{v} = \mathbf{e}_1$: equality holds, and the list is not a basis.

11.9.6.2 B. Practice

Exercise. B1: Fourier coefficients

Work in $C[-\pi, \pi]$ with $\langle f, g\rangle = \frac{1}{\pi}\int_{-\pi}^{\pi} fg$. Find all Fourier coefficients a_k and b_k of each of the following, and in each case write down $S_2 f$.

(a) $f(x) = |x|$.

(b) $f(x) = x^2$.

(c) $f(x) = 3 - 2\sin x + \sin 3x$.

Solution.

(a) $|t|\sin(kt)$ is odd, so $b_k = 0$ for all k. For the cosines, $|t|\cos(kt)$ is even, so $a_k = \frac{2}{\pi}\int_0^{\pi} t\cos(kt)\,dt$. For $k = 0$ this is $\frac{2}{\pi}\cdot\frac{\pi^2}{2} = \pi$. For $k \geq 1$, integrating by parts,

$$\int_0^{\pi} t\cos(kt)\,dt = \Big[\frac{t\sin(kt)}{k}\Big]_0^{\pi} - \frac{1}{k}\int_0^{\pi}\sin(kt)\,dt = \frac{(-1)^k - 1}{k^2},$$

since the bracket vanishes and $\int_0^{\pi}\sin(kt)\,dt = (1 - (-1)^k)/k$. Hence

$$a_k = \frac{2((-1)^k - 1)}{\pi k^2} = \begin{cases} -\dfrac{4}{\pi k^2}, & k \text{ odd}, \\ 0, & k \text{ even}, \end{cases}$$

and $S_2 f = \dfrac{\pi}{2} - \dfrac{4}{\pi}\cos x$.

(b) $t^2\sin(kt)$ is odd, so $b_k = 0$. Also $a_0 = \frac{2}{\pi}\int_0^{\pi} t^2\,dt = \frac{2\pi^2}{3}$. For $k \geq 1$, integrating by parts twice gives

$$\int_0^{\pi} t^2\cos(kt)\,dt = \Big[\frac{t^2\sin(kt)}{k}\Big]_0^{\pi} - \frac{2}{k}\int_0^{\pi} t\sin(kt)\,dt = \frac{2\pi(-1)^k}{k^2},$$

using $\int_0^{\pi} t\sin(kt)\,dt = \pi(-1)^{k+1}/k$ from Example (The Fourier Coefficients of $f(x) = x$). Hence $a_k = \frac{2}{\pi}\cdot\frac{2\pi(-1)^k}{k^2} = \frac{4(-1)^k}{k^2}$, so $a_1 = -4$, $a_2 = 1$, and

$$S_2 f = \frac{\pi^2}{3} - 4\cos x + \cos 2x.$$

(c) The function already lies in W_3. Coefficients against an orthonormal basis are unique (Theorem 11.2.1 (a)), so, applying that result inside W_3, we may read them off: $a_0/2 = 3$, that is $a_0 = 6$, together with $b_1 = -2$ and $b_3 = 1$; every other coefficient is 0. Hence

$S_2 f = 3 - 2\sin x$, which is **not** f: the term $\sin 3x$ lies outside W_2 and is discarded by the projection. (From $n = 3$ on, $S_n f = f$, as Theorem 11.3.3 (a) predicts, since $f \in W_3$ and a projection fixes its own subspace.)

Exercise. B2: Verifying Bessel for a truncation

(a) In $\mathbb{R}^4$ with the dot product let $\mathbf{e}_1 = \frac{1}{2}(1,1,1,1)$, $\mathbf{e}_2 = \frac{1}{2}(1,-1,1,-1)$ and $\mathbf{v} = (2,0,0,2)$. Verify Theorem 11.9.1 for $\mathbf{v}$ and this list, compute the deficit, and check against $(*)$ by computing $\|\mathbf{v} - P_U\mathbf{v}\|$ directly, where $U = \operatorname{span}(\mathbf{e}_1, \mathbf{e}_2)$.

(b) For $f(x) = |x|$ and $n = 1$, write out both sides of Theorem 11.9.3 (c) using Exercise (B1: Fourier coefficients) (a), and show that the resulting inequality is equivalent to $96 \le \pi^4$.

Solution.

(a) Both vectors have norm $\frac{1}{2}\sqrt{4} = 1$ and $\langle \mathbf{e}_1, \mathbf{e}_2 \rangle = \frac{1}{4}(1 - 1 + 1 - 1) = 0$, so the list is orthonormal. The coefficients are $\langle \mathbf{v}, \mathbf{e}_1 \rangle = \frac{1}{2}(2 + 0 + 0 + 2) = 2$ and $\langle \mathbf{v}, \mathbf{e}_2 \rangle = \frac{1}{2}(2 - 0 + 0 - 2) = 0$, so the left-hand side of Bessel is 4, while $\|\mathbf{v}\|^2 = 4 + 4 = 8$. The inequality $4 \le 8$ holds, with deficit 4. By Theorem 11.3.3 (a), $P_U\mathbf{v} = 2\mathbf{e}_1 = (1,1,1,1)$, so $\mathbf{v} - P_U\mathbf{v} = (1,-1,-1,1)$ and $\|\mathbf{v} - P_U\mathbf{v}\|^2 = 4$, matching $(*)$.

(b) With $a_0 = \pi$ and $a_1 = -4/\pi$, and $b_1 = 0$, the left-hand side of (c) for $n = 1$ is

$$\frac{a_0^2}{2} + a_1^2 = \frac{\pi^2}{2} + \frac{16}{\pi^2}.$$

The right-hand side is $\|f\|^2 = \frac{1}{\pi}\int_{-\pi}^{\pi} t^2\,dt = \frac{2\pi^2}{3}$. So the inequality reads

$$\frac{\pi^2}{2} + \frac{16}{\pi^2} \le \frac{2\pi^2}{3},$$

that is $16/\pi^2 \le 2\pi^2/3 - \pi^2/2 = \pi^2/6$, that is $96 \le \pi^4$. Since $\pi > 3.14$ and $3.14^4 > 97$, the inequality is true, and it is tight: the deficit is $\pi^2/6 - 16/\pi^2 \approx 0.024$, so the single cosine $\cos x$ already captures almost all of $|x|$.

Exercise. B3: The best approximation by $\operatorname{span}(1, \cos x)$

In $C[-\pi, \pi]$ with $\langle f, g \rangle = \frac{1}{\pi}\int_{-\pi}^{\pi} fg$, let $U = \operatorname{span}(1, \cos x)$ and $f(x) = x^2$. Find $P_U f$ and $d(f, U)^2$. *Hint: the list $(1, \cos x)$ is orthogonal but not orthonormal.*

Solution. The two functions are orthogonal, since $\langle 1, \cos x \rangle = \frac{1}{\pi}\int_{-\pi}^{\pi} \cos t\,dt = 0$, and both are nonzero, with $\langle 1, 1 \rangle = 2$ and $\langle \cos x, \cos x \rangle = 1$ by Example (The Trigonometric System Is Orthonormal). So $(1, \cos x)$ is an orthogonal basis of U, and Corollary 11.3.4 gives

$$P_U f = \frac{\langle f, 1 \rangle}{2} \cdot 1 + \frac{\langle f, \cos x \rangle}{1} \cdot \cos x.$$

By Exercise (B1: Fourier coefficients) (b), $\langle f, 1 \rangle = a_0 = 2\pi^2/3$ and $\langle f, \cos x \rangle = a_1 = -4$.

Hence

$$P_U f = \frac{\pi^2}{3} - 4\cos x.$$

For the distance, normalize: $(1/\sqrt{2},\ \cos x)$ is an orthonormal basis of the same U, with coefficients $\langle f, 1/\sqrt{2}\rangle = a_0/\sqrt{2}$ and $\langle f, \cos x\rangle = a_1$. Now $\|f\|^2 = \frac{1}{\pi}\int_{-\pi}^{\pi} t^4\,\mathrm{d}t = \frac{1}{\pi}\cdot\frac{2\pi^5}{5} = \frac{2\pi^4}{5}$ and the Bessel sum is $a_0^2/2 + a_1^2 = 2\pi^4/9 + 16$, so $(*)$ gives

$$d(f,U)^2 = \frac{2\pi^4}{5} - \frac{2\pi^4}{9} - 16 = \frac{8\pi^4}{45} - 16 \approx 1.32.$$

It is positive, as it must be, since $x^2 \notin U$.

11.9.6.3 C. Going deeper

Exercise. C1: Parseval for the full trigonometric system

This exercise uses an **input from analysis that this book does not prove**: the trigonometric system is *complete* in $C[-\pi,\pi]$, meaning that for every $f \in C[-\pi,\pi]$,

$$\frac{a_0^2}{2} + \sum_{k=1}^{\infty}(a_k^2 + b_k^2) = \frac{1}{\pi}\int_{-\pi}^{\pi} f(t)^2\,\mathrm{d}t.$$

(a) Apply this to $f(x) = x$ and deduce that $\sum_{k=1}^{\infty} 1/k^2 = \pi^2/6$.

(b) Hence deduce that $\sum_{k=1}^{\infty} 1/(2k-1)^2 = \pi^2/8$.

(c) Say precisely which step used the analysis input, and explain why no result of this section can supply it.

Solution.

(a) By Example (The Fourier Coefficients of $f(x) = x$), $a_k = 0$ for all k and $b_k = 2(-1)^{k+1}/k$, so $b_k^2 = 4/k^2$; and $\|f\|^2 = 2\pi^2/3$. The stated identity becomes

$$\sum_{k=1}^{\infty}\frac{4}{k^2} = \frac{2\pi^2}{3},$$

so $\sum_{k=1}^{\infty} 1/k^2 = \pi^2/6$.

(b) All terms are positive and the series converges, so it may be split into its even-index and odd-index parts. The even part is

$$\sum_{j=1}^{\infty}\frac{1}{(2j)^2} = \frac{1}{4}\sum_{j=1}^{\infty}\frac{1}{j^2} = \frac{\pi^2}{24}.$$

Hence the odd part is $\pi^2/6 - \pi^2/24 = (4\pi^2 - \pi^2)/24 = \pi^2/8$, which is the claim.

(c) The analysis input is used exactly once, at the equality sign in (a). This section proves only the inequality $\sum_{k\le n} 4/k^2 \le 2\pi^2/3$ for each n (Theorem 11.9.1), and Theorem 11.9.2 turns that inequality into an equality only when f lies in the span of the finitely many functions used. The function x lies in no W_n, since every element of W_n takes equal values at $-\pi$ and π while $f(-\pi) = -\pi \neq \pi = f(\pi)$. Passing to the infinite system is a statement about a limit of functions, and nothing in this book proves such a statement.

Remark. The classical route to $\sum 1/(2k-1)^2 = \pi^2/8$ applies the same identity to the square wave, the function equal to -1 on $[-\pi, 0)$ and 1 on $[0, \pi]$, whose coefficients are $b_k = 4/(\pi k)$ for odd k and 0 otherwise. That function is not continuous, so it does not live in the space $C[-\pi, \pi]$ of this section, and enlarging the space to all integrable functions destroys positive definiteness, as the third nonexample of Section 1 shows. Working with $f(x) = x$ keeps everything inside a genuine inner product space.

Exercise. C2: Chebyshev orthogonality by substitution

Let $\langle p, q\rangle_w = \int_{-1}^{1} p(t)q(t)(1-t^2)^{-1/2}\,\mathrm{d}t$ on $\mathbb{R}[x]$, the Chebyshev inner product of the text.

(a) Prove that for every $n \geq 0$ there is a polynomial T_n of degree n with $T_n(\cos\theta) = \cos n\theta$ for all θ, and find T_0, T_1, T_2, T_3.

(b) Prove that $\langle p, q\rangle_w = \int_0^{\pi} p(\cos\theta)\, q(\cos\theta)\,\mathrm{d}\theta$ for all polynomials p, q.

(c) Deduce that $\langle T_m, T_n\rangle_w = 0$ for $m \neq n$, and compute $\langle T_n, T_n\rangle_w$.

Hint for (a): the identity $\cos((n+1)\theta) + \cos((n-1)\theta) = 2\cos\theta\cos n\theta$.

Solution.

(a) Induction on n. Take $T_0 = 1$ and $T_1 = x$, which satisfy $T_0(\cos\theta) = 1 = \cos 0\theta$ and $T_1(\cos\theta) = \cos\theta$, of degrees 0 and 1. Suppose T_{n-1} and T_n exist, of degrees $n-1$ and n. Put $T_{n+1} = 2xT_n - T_{n-1}$. Then for every θ,

$$T_{n+1}(\cos\theta) = 2\cos\theta\cos n\theta - \cos((n-1)\theta) = \cos((n+1)\theta)$$

by the stated identity. Its degree is $n+1$, since the leading term of $2xT_n$ has degree $n+1$ and T_{n-1} has degree $n-1$. This completes the induction, and the recurrence gives $T_2 = 2x^2 - 1$ and $T_3 = 2x(2x^2-1) - x = 4x^3 - 3x$.

(b) The integral is improper at both endpoints, so it means the limit as $\delta \to 0^+$ of the integrals over $[-1+\delta,\ 1-\delta]$. On each such interval the integrand is continuous, and $t = \cos\theta$ is a continuously differentiable bijection from the corresponding θ-interval, so the change-of-variables rule applies there. With $t = \cos\theta$, θ running from π down to 0, we have $\mathrm{d}t = -\sin\theta\,\mathrm{d}\theta$ and $\sqrt{1-t^2} = \sin\theta$, which is ≥ 0 on $[0, \pi]$. The two minus signs cancel when the limits are reversed, and letting $\delta \to 0^+$,

$$\int_{-1}^{1} \frac{p(t)q(t)}{\sqrt{1-t^2}}\,\mathrm{d}t = \int_0^{\pi} p(\cos\theta)\, q(\cos\theta)\,\mathrm{d}\theta.$$

The right-hand side is a proper integral of a continuous function, so both sides converge, and the limits agree.

(c) By (a) and (b), $\langle T_m, T_n\rangle_w = \int_0^{\pi} \cos m\theta\cos n\theta\,\mathrm{d}\theta$. Using $\cos A\cos B = \frac{1}{2}[\cos(A-B) + \cos(A+B)]$ and $\int_0^{\pi}\cos(p\theta)\,\mathrm{d}\theta = [\sin(p\theta)/p]_0^{\pi} = 0$ for every nonzero integer p:

if $m \neq n$, then $m - n \neq 0$, and $m + n \geq 1$ because the two indices differ, so both integrals vanish and the value is 0;

if $m = n \geq 1$, the value is $\frac{1}{2}\big(\int_0^{\pi} 1\,\mathrm{d}\theta + 0\big) = \pi/2$;

if $m = n = 0$, the integrand is 1 and the value is π.

So the T_n are orthogonal for $\langle\cdot,\cdot\rangle_w$, with $\langle T_0, T_0\rangle_w = \pi$ and $\langle T_n, T_n\rangle_w = \pi/2$ for $n \geq 1$. Consequently $(T_0/\sqrt{\pi},\ T_1\sqrt{2/\pi},\ \ldots,\ T_n\sqrt{2/\pi})$ is an orthonormal list, and Theorem 11.9.1 applies to it verbatim.

11.10 Orthogonal Polynomials

Section 2 ran Gram–Schmidt on $(1, x, x^2)$ with $\langle p, q\rangle = \int_{-1}^{1} pq$, and stopped at degree 2 only because the space did (Example (Gram–Schmidt on $1, x, x^2$)). Nothing in the process cared. Feed it $(1, x, x^2, x^3, \ldots)$ and it returns an infinite orthogonal sequence, one polynomial of each degree. Three things make such sequences worth a section. Each member is computable from the two before it. Each member of degree k has k distinct real roots inside the interval. And those roots are the best possible points at which to sample a function to estimate its integral. The first fact rests on one small identity; the other two rest on the first.

Throughout this section the field is $\mathbb{R}$.

11.10.1 Two ways to make an inner product on polynomials

Both constructions we need appeared in Section 1.

From a weight. Fix an interval with endpoints $-\infty \leq a < b \leq \infty$ and a **weight** w: a function continuous and **strictly positive** on the open interval (a, b), with $\int_a^b |t|^k w(t)\,\mathrm{d}t$ finite for every $k \in \mathbb{N}$. Define

$$\langle p, q\rangle = \int_a^b p(t)\,q(t)\,w(t)\,\mathrm{d}t.$$

Finiteness of the moments makes this a real number for all $p, q \in \mathbb{R}[x]$, and bilinearity and symmetry come from linearity of the integral. For positivity, let $p \neq 0$. Then $p^2 w$ is continuous and ≥ 0 on (a, b), and > 0 off the finitely many roots of p (Corollary 6.1.4), so $\langle p, p\rangle > 0$. That last step is the one calculus fact this section leans on: a continuous nonnegative function on an interval whose integral is 0 vanishes identically. As in Definition 11.1.1 it follows from the local fact that a continuous function nonzero at a point stays bounded away from 0 on a small interval around it, which then contributes a positive amount to the integral.

From nodes. Fix distinct $c_0 < \cdots < c_n$ in $\mathbb{R}$ and put $\langle p, q\rangle = \sum_{i=0}^{n} p(c_i)q(c_i)$. Section 1 checked that this is an inner product **on** $\mathbb{R}[x]_{\leq n}$, positivity coming from the root bound. It is not one on $\mathbb{R}[x]$: the polynomial $(x - c_0)\cdots(x - c_n)$ is nonzero and pairs with itself to give 0. So in the node case everything below stops at degree n.

The two constructions share one property, and it is the engine of the section.

Proposition 11.10.1. Multiplication by x is symmetric

Let $\langle\cdot,\cdot\rangle$ be an inner product on $\mathbb{R}[x]$ coming from a weight, or the node inner product on $\mathbb{R}[x]_{\leq n}$. Then

$$\langle xp, q\rangle = \langle p, xq\rangle$$

for all polynomials p, q in the space (in the node case, for all p, q of degree at most $n - 1$, so that xp and xq still lie in it).

Proof. In the weight case both sides equal $\int_a^b t\,p(t)q(t)w(t)\,dt$, and in the node case both sides equal $\sum_{i=0}^{n} c_i\,p(c_i)q(c_i)$.

■

The proof is one line, but the statement is not a triviality: multiplication by x moves from one slot to the other for free. That is exactly the identity $\langle Tu, v\rangle = \langle u, Tv\rangle$ defining a **self-adjoint** operator (Definition 11.6.2), and it is the reason the recurrence below has three terms instead of $k+2$.

Warning. **Multiplication by x is not an operator on $\mathbb{R}[x]_{\le n}$.** It raises degree, so it maps that space out of itself; on $\mathbb{R}[x]$ it is an operator, but there the space is infinite-dimensional, and Definition 11.6.2 was stated for a finite-dimensional one. "Self-adjoint" is here a name for Proposition 11.10.1, not a literal application of the definition. It is the identity, not the name, that the proofs use.

11.10.2 Orthogonal polynomial sequences

Applied to the independent list $(1, x, x^2, \dots)$, Gram–Schmidt never stalls. What comes out is one polynomial of each degree, pairwise orthogonal, and that combination deserves a name.

An orthogonal polynomial sequence is one polynomial of every degree, no two of them pairing to anything but zero.

Definition 11.10.1. Orthogonal Polynomial Sequence

Let $\langle\cdot,\cdot\rangle$ be an inner product on $\mathbb{R}[x]$, or on $\mathbb{R}[x]_{\le n}$. An **orthogonal polynomial sequence** for it is a sequence $p_0, p_1, p_2, \dots$ of polynomials (a finite list $p_0, \dots, p_n$ in the second case) such that

(OP1) $\deg p_k = k$ for **every** index k;

(OP2) $\langle p_j, p_k\rangle = 0$ whenever $j \neq k$.

The sequence is **monic** if every p_k has leading coefficient 1.

In words. (OP1) says one polynomial of each degree: no gaps, no repeats, no accidental drop in degree. (OP2) says the members are pairwise orthogonal. Together they say more than either alone, because by (OP1) the list $(p_0, \dots, p_{k-1})$ is a basis of $\mathbb{R}[x]_{\le k-1}$, so (OP2) upgrades to

$$\langle p_k, q\rangle = 0 \qquad \text{for every } q \in \mathbb{R}[x]_{\le k-1},$$

and that is the form in which the condition always gets used.

Two things need checking: that such a sequence exists, and that fixing the scale pins it down.

Proposition 11.10.2. Existence and uniqueness of the monic sequence

Let $\langle\cdot,\cdot\rangle$ be an inner product on $\mathbb{R}[x]$. For each $k \in \mathbb{N}$ there is **exactly one** monic p_k of degree k with $\langle p_k, q\rangle = 0$ for every $q \in \mathbb{R}[x]_{\le k-1}$. These p_k form the unique monic orthogonal polynomial sequence, and $(p_0, \dots, p_k)$ is an orthogonal basis of $\mathbb{R}[x]_{\le k}$ for every k. The same

holds on $\mathbb{R}[x]_{\le n}$ for $k \le n$.

Idea. Existence is Gram–Schmidt, once we know the k-th output really has degree k; it does, because it lies in $\mathbb{R}[x]_{\le k}$ but not in the span of its predecessors, which is $\mathbb{R}[x]_{\le k-1}$. Uniqueness is the standard move: two candidates differ by a polynomial of lower degree that is orthogonal to everything of lower degree, hence to itself.

Proof. *Existence.* The list $(1, x, \dots, x^k)$ is independent, so Theorem 11.2.2 applies and returns an orthonormal list $(e_0, \dots, e_k)$ with $\operatorname{span}(e_0, \dots, e_j) = \operatorname{span}(1, x, \dots, x^j) = \mathbb{R}[x]_{\le j}$ for every $j \le k$. Fix j. Then $e_j \in \mathbb{R}[x]_{\le j}$, and $e_j \notin \mathbb{R}[x]_{\le j-1} = \operatorname{span}(e_0, \dots, e_{j-1})$, since an orthonormal list is independent (Theorem 11.1.4). Hence $\deg e_j = j$ exactly. Let $a_j \neq 0$ be its leading coefficient and put $p_j = e_j/a_j$, monic of degree j. For $q \in \mathbb{R}[x]_{\le j-1} = \operatorname{span}(e_0, \dots, e_{j-1})$ we get $\langle p_j, q\rangle = 0$ by orthogonality of the list.

Uniqueness. Suppose p and p' are both monic of degree k and orthogonal to every element of $\mathbb{R}[x]_{\le k-1}$. Their leading terms cancel, so $p - p' \in \mathbb{R}[x]_{\le k-1}$. Therefore $\langle p - p', p - p'\rangle = \langle p, p - p'\rangle - \langle p', p - p'\rangle = 0$, and positive definiteness gives $p = p'$.

Finally, $(p_0, \dots, p_k)$ is an orthogonal list of nonzero vectors, hence independent (Theorem 11.1.4), with $k + 1 = \dim \mathbb{R}[x]_{\le k}$ members, hence a basis (Theorem 1.6.4).

■

Examples. For $\langle p, q\rangle = \int_{-1}^{1} pq$ the first three monic members are 1, x, $x^2 - \frac{1}{3}$ (Example (Gram–Schmidt on $1, x, x^2$)); these are the monic **Legendre polynomials**. For the node inner product with nodes $-1, 0, 1$ on $\mathbb{R}[x]_{\le 2}$ we have $\langle 1, 1\rangle = 3$, $\langle x, 1\rangle = 0$, $\langle x, x\rangle = 2$, $\langle x^2, 1\rangle = 2$, $\langle x^2, x\rangle = 0$, so $p_0 = 1$, $p_1 = x$, $p_2 = x^2 - \frac{2}{3}$ — and there the sequence ends, since $\mathbb{R}[x]_{\le 2}$ holds nothing of degree 3. The degenerate member is always $p_0 = 1$, the only monic polynomial of degree 0, with an empty condition on it.

A nonexample by minimal change. Replace the Legendre $p_2 = x^2 - \frac{1}{3}$ by $\tilde{p}_2 = x^2 + x - \frac{1}{3}$. Still monic of degree 2, so (OP1) survives; but $\langle \tilde{p}_2, x\rangle = \langle x, x\rangle = \frac{2}{3} \neq 0$, so (OP2) fails. Adding a multiple of a lower member is exactly the freedom orthogonality removes.

Why monic? Rescaling each p_k by a nonzero λ_k gives another orthogonal polynomial sequence, so a normalization has to be chosen, and traditions differ: the classical Legendre polynomials P_k are fixed by $P_k(1) = 1$, and Gram–Schmidt as run in Section 2 fixes $\|p_k\| = 1$. Monic is the convenient choice here, because it is the one for which the recurrence below needs no extra factor.

Warning. **The sequence belongs to the inner product, not to the polynomials.** "The orthogonal polynomials" is meaningless without a weight and an interval. Exercise B2 of Section 2 (Exercise (B2: Gram–Schmidt on polynomials)) ran the same process with $\int_0^1$ instead of $\int_{-1}^1$ and got 1, $x - \frac{1}{2}$, $x^2 - x + \frac{1}{6}$: same degrees, different polynomials.

Quick check. For $\langle p, q\rangle = \int_{-1}^{1} pq$, how many polynomials of degree 2 with $\|p\| = 1$ are orthogonal to $\mathbb{R}[x]_{\le 1}$? Write them down.

Solution. Exactly two. If p has degree 2 and $p \perp \mathbb{R}[x]_{\le 1}$, let $\lambda \neq 0$ be its leading coefficient; then p/λ is monic of degree 2 and still orthogonal to $\mathbb{R}[x]_{\le 1}$, so $p/\lambda = x^2 - \frac{1}{3}$ by Proposition 11.10.2. Thus $p = \lambda(x^2 - \frac{1}{3})$, and $\|p\| = 1$ leaves $\lambda = \pm 1/\|x^2 - \frac{1}{3}\|$. Since Example (Gram–Schmidt on $1, x, x^2$) computed $\|x^2 - \frac{1}{3}\|^2 = \frac{8}{45}$, the two answers are $\pm\frac{\sqrt{10}}{4}(3x^2 - 1)$. Fixing the scale by the norm leaves a sign; fixing it by "monic" leaves nothing.

11.10.3 The three-term recurrence

Computing p_n by Gram–Schmidt costs one inner product for every earlier member, so reaching degree n costs about $n^2/2$ integrals. Almost all of them return 0. The next theorem says which ones do not, and reduces the cost to two integrals per step.

Theorem 11.10.3. Three-Term Recurrence

Let (p_k) be the monic orthogonal polynomial sequence of Proposition 11.10.2, for an inner product satisfying Proposition 11.10.1. For a weight on an interval this covers every k; for the node inner product on n nodes, where Proposition 11.10.1 holds only up to degree $n-1$, read the statement for $k+1 \le n$. Put $p_{-1} = 0$ and, for $k \ge 0$,

$$a_k = \frac{\langle xp_k, p_k\rangle}{\langle p_k, p_k\rangle}, \qquad b_k = \frac{\langle p_k, p_k\rangle}{\langle p_{k-1}, p_{k-1}\rangle} \quad (k \ge 1).$$

Then $p_0 = 1$, $p_1 = x - a_0$, and

$$p_{k+1} = (x - a_k)\,p_k - b_k\,p_{k-1} \qquad \text{for every } k \ge 1.$$

Moreover $b_k > 0$ for every $k \ge 1$.

Idea. The polynomial xp_k is monic of degree $k+1$, so it is p_{k+1} plus something of degree at most k, and that something expands in the orthogonal basis $(p_0, \dots, p_k)$ with coefficients $\langle xp_k, p_j\rangle/\langle p_j, p_j\rangle$ (Corollary 11.3.4). The point is that almost every coefficient is 0: move the x across with Proposition 11.10.1, and the coefficient becomes $\langle p_k, xp_j\rangle$, which vanishes as soon as $\deg(xp_j) = j+1$ is still below k. Only $j = k$ and $j = k-1$ can survive.

Proof. $p_0 = 1$ is the unique monic polynomial of degree 0, and $p_1 = x - a_0$ is the unique monic polynomial of degree 1 with $\langle p_1, p_0\rangle = 0$, since $\langle x - a_0, 1\rangle = \langle x, 1\rangle - a_0\langle 1, 1\rangle = 0$ exactly for $a_0 = \langle x \cdot 1, 1\rangle/\langle 1, 1\rangle$.

Now fix $k \ge 1$. By Proposition 11.10.2, $(p_0, \dots, p_{k+1})$ is an orthogonal basis of $\mathbb{R}[x]_{\le k+1}$,

and xp_k lies in that space. By Corollary 11.3.4 applied inside $\mathbb{R}[x]_{\le k+1}$,

$$xp_k = \sum_{j=0}^{k+1} c_j\, p_j, \qquad c_j = \frac{\langle xp_k, p_j\rangle}{\langle p_j, p_j\rangle}.$$

Both xp_k and p_{k+1} are monic of degree $k+1$, and every other term of the sum has degree at most k; comparing coefficients of x^{k+1} gives $c_{k+1} = 1$.

Let $j \le k-2$. By Proposition 11.10.1, $\langle xp_k, p_j\rangle = \langle p_k, xp_j\rangle$, and $\deg(xp_j) = j+1 \le k-1$, so $xp_j \in \mathbb{R}[x]_{\le k-1}$ and the pairing is 0 by Proposition 11.10.2. Hence $c_j = 0$ for every $j \le k-2$, and

$$xp_k = p_{k+1} + c_k\, p_k + c_{k-1}\, p_{k-1},$$

which rearranges to $p_{k+1} = (x - c_k)p_k - c_{k-1}p_{k-1}$.

It remains to identify the two coefficients. By definition $c_k = \langle xp_k, p_k\rangle / \langle p_k, p_k\rangle = a_k$. For the other, xp_{k-1} is monic of degree k, so $xp_{k-1} - p_k$ has degree at most $k-1$ and therefore pairs to 0 with p_k. Hence, using Proposition 11.10.1 again,

$$\langle xp_k, p_{k-1}\rangle = \langle p_k, xp_{k-1}\rangle = \langle p_k, p_k\rangle,$$

so $c_{k-1} = \langle p_k, p_k\rangle / \langle p_{k-1}, p_{k-1}\rangle = b_k$. Finally b_k is a quotient of $\|p_k\|^2$ by $\|p_{k-1}\|^2$, both strictly positive because p_k and p_{k-1} are nonzero, so $b_k > 0$.

■

Two inner products per step generate the whole sequence. The positivity of b_k is not decoration either: it is the sign that drives the interlacing argument of Exercise (C1: Interlacing roots).

Remark. If the interval is symmetric, $(a, b) = (-c, c)$, and the weight is even, $w(-t) = w(t)$, then $a_k = 0$ for every k and $p_k(-x) = (-1)^k p_k(x)$. Induct on k: $p_0 = 1$ is even, and $a_0 = \langle x, 1\rangle / \langle 1, 1\rangle = 0$ because $t\, w(t)$ is odd, so $p_1 = x$ is odd. If p_{k-1} and p_k have parities $(-1)^{k-1}$ and $(-1)^k$, then $t\, p_k(t)^2 w(t)$ is odd, so $a_k = 0$, and $p_{k+1} = xp_k - b_k p_{k-1}$ is a combination of two polynomials of parity $(-1)^{k+1}$.

Example. The monic Legendre polynomials, continued

For $\langle p, q\rangle = \int_{-1}^{1} pq$, Example (Gram–Schmidt on $1, x, x^2$) gave $p_0 = 1$, $p_1 = x$, $p_2 = x^2 - \frac{1}{3}$, with $\langle p_0, p_0\rangle = 2$, $\langle p_1, p_1\rangle = \frac{2}{3}$, and $\langle p_2, p_2\rangle = \frac{8}{45}$. Use Theorem 11.10.3 to obtain p_3, and check the recurrence reproduces p_2.

Solution. The weight $w = 1$ is even on $(-1, 1)$, so every $a_k = 0$ by the remark above. Then

$$b_1 = \frac{\langle p_1, p_1\rangle}{\langle p_0, p_0\rangle} = \frac{2/3}{2} = \frac{1}{3}, \qquad b_2 = \frac{\langle p_2, p_2\rangle}{\langle p_1, p_1\rangle} = \frac{8/45}{2/3} = \frac{4}{15}.$$

The recurrence at $k = 1$ gives $p_2 = x \cdot x - \frac{1}{3} \cdot 1 = x^2 - \frac{1}{3}$, which is what Gram–Schmidt produced. At $k = 2$,

$$p_3 = x\Big(x^2 - \tfrac{1}{3}\Big) - \tfrac{4}{15}\, x = x^3 - \tfrac{3}{5}x.$$

Check. $\langle p_3, 1\rangle = \langle p_3, x^2\rangle = 0$ because the integrands are odd, and

$$\langle p_3, x\rangle = \int_{-1}^{1} \left(t^4 - \tfrac{3}{5}t^2\right)\mathrm{d}t = \tfrac{2}{5} - \tfrac{3}{5}\cdot\tfrac{2}{3} = 0.$$

So $p_3 \perp \mathbb{R}[x]_{\le 2}$, as it should be. Two integrals produced a polynomial that Gram–Schmidt would have charged three for, and the saving grows with the degree.

11.10.4 Where the roots are

The roots of $p_2 = x^2 - \frac{1}{3}$ are $\pm 1/\sqrt{3} \approx \pm 0.577$, and those of $p_3 = x^3 - \frac{3}{5}x$ are 0 and $\pm\sqrt{3/5} \approx \pm 0.775$. All five lie inside $(-1, 1)$, where the weight lives. No coincidence. The argument first needs one fact about real polynomials, in which $\mathrm{mult}_r(f)$ is the multiplicity of r as a root of f (Definition 6.4.1).

Lemma 11.10.4. Even multiplicities mean no sign change

Let $I \subseteq \mathbb{R}$ be an open interval and let $f \in \mathbb{R}[x]$ be nonzero. If $\mathrm{mult}_r(f)$ is even for every $r \in I$, then $f \ge 0$ throughout I, or $f \le 0$ throughout I.

Proof. By Theorem 6.4.3 over $\mathbb{R}$, write $f = (x - d_1)^{m_1} \cdots (x - d_l)^{m_l}\, g$, where $d_1, \ldots, d_l$ are the distinct real roots of f, $m_i = \mathrm{mult}_{d_i}(f)$, and g has no real root. Separate the factors according to whether $d_i \in I$:

$$f = h \cdot \left(\textstyle\prod_{d_i \in I} (x - d_i)^{m_i}\right), \qquad h = g \cdot \textstyle\prod_{d_i \notin I} (x - d_i)^{m_i}.$$

Each m_i with $d_i \in I$ is even by hypothesis, so the second factor is the square of a real polynomial and is ≥ 0 everywhere. The polynomial h has no root in I: its roots are the d_i outside I, and g has none. Being continuous and nowhere zero on the interval I, h has constant sign there by the intermediate value theorem. Hence f has that same sign throughout I, weakly, with equality only at the d_i inside I.

■

Theorem 11.10.5. Roots of an Orthogonal Polynomial

Let w be a weight on (a, b), let $\langle p, q\rangle = \int_a^b pqw$, and let (p_k) be the monic orthogonal polynomial sequence. For every $k \ge 1$, the polynomial p_k has k **distinct real** roots, and every one of them lies in the **open** interval (a, b).

Idea. Suppose p_k changed sign at only a few points $c_1, \ldots, c_m$ of (a, b), with $m < k$. Build the polynomial $q = (x - c_1)\cdots(x - c_m)$. It is cheap — degree at most $k - 1$ — so p_k is orthogonal to it. But q was designed to change sign at exactly the points where p_k does, so the product $p_k q$ changes sign nowhere, and a function of one sign that is not identically zero cannot integrate to zero against a positive weight. The contradiction forces $m \ge k$, and k distinct roots is all a polynomial of degree k can hold.

Proof. Fix $k \geq 1$. Let $c_1 < \cdots < c_m$ be the roots of p_k in (a, b) whose multiplicity is **odd** (possibly $m = 0$), and set $q = (x - c_1) \cdots (x - c_m)$, with $q = 1$ when $m = 0$.

Suppose, for contradiction, that $m \leq k - 1$. Then $q \in \mathbb{R}[x]_{\leq k-1}$, so $\langle p_k, q \rangle = 0$ by Proposition 11.10.2.

Put $f = p_k q$, a nonzero polynomial (Theorem 0.8.3). Let $r \in (a, b)$ be a root of f. By Theorem 6.4.2, $\operatorname{mult}_r(f) = \operatorname{mult}_r(p_k) + \operatorname{mult}_r(q)$. If $r = c_i$ for some i, then $\operatorname{mult}_r(q) = 1$ and $\operatorname{mult}_r(p_k)$ is odd by the choice of the c_i, so the sum is even. If r is none of the c_i, then $q(r) \neq 0$, so $\operatorname{mult}_r(q) = 0$ and $\operatorname{mult}_r(p_k)$ is even, again by the choice of the c_i. Either way $\operatorname{mult}_r(f)$ is even.

By Lemma 11.10.4, $f \geq 0$ throughout (a, b) or $f \leq 0$ throughout (a, b); replacing q by $-q$ if necessary, which changes neither $\langle p_k, q \rangle = 0$ nor $\deg q$, we may assume $f \geq 0$ there. Then fw is continuous and ≥ 0 on (a, b), and it is > 0 off the finitely many roots of f, because $w > 0$. By the calculus fact recalled when the inner product was set up — a continuous nonnegative function with integral 0 vanishes identically — an integrand of that shape cannot integrate to 0. Therefore

$$\langle p_k, q \rangle = \int_a^b f(t)\, w(t)\, \mathrm{d}t > 0,$$

contradicting $\langle p_k, q \rangle = 0$.

Hence $m \geq k$. Each c_i is a root of p_k, and p_k is nonzero of degree k, so $m \leq k$ by Corollary 6.1.4; thus $m = k$ and p_k has k distinct roots $c_1, \ldots, c_k$, all in (a, b). By Corollary 6.1.4 again, $(x - c_1) \cdots (x - c_k)$ divides p_k, and both are monic of degree k, so $p_k = (x - c_1) \cdots (x - c_k)$. In particular each root is simple and p_k has no roots outside (a, b), real or complex. This proves the theorem. ∎

So the weight hands us k points inside its interval, chosen by the inner product rather than by us. The next subsection says what they are good for.

Quick check. The monic Laguerre polynomial of degree 2 for the weight $w(t) = e^{-t}$ on $(0, \infty)$ is $p_2 = x^2 - 4x + 2$. Check that Theorem 11.10.5 is not lying.

Solution. The roots are $2 \pm \sqrt{2}$, that is ≈ 0.586 and ≈ 3.414: real, distinct, and both in $(0, \infty)$. An unbounded interval is no obstacle, since the theorem asks only that the weight be positive with finite moments, and $\int_0^\infty t^k e^{-t}\, \mathrm{d}t = k!$.

11.10.5 Gaussian quadrature

Chapter 6 wrote $\int_0^1 p$ as a fixed combination of $p(0)$, $p(\frac{1}{2})$ and $p(1)$, valid for every $p \in \mathbb{R}[x]_{\leq 2}$ (Example (Integrals and Slopes from Values)). The construction works for any prescribed nodes: integrate the Lagrange interpolation formula (Theorem 6.6.1) and k distinct nodes give a rule exact on $\mathbb{R}[x]_{\leq k-1}$.

But a k-point rule has $2k$ numbers at its disposal, k nodes and k coefficients, and only k of

them were spent. Spending the nodes too should buy degree $2k-1$. Where to put them is the question, and the answer is the one place we have not chosen: the roots of p_k.

Theorem 11.10.6. Gaussian Quadrature

Let w be a weight on (a,b) with $\langle p,q\rangle = \int_a^b pqw$, let $k \geq 1$, and let $t_1 < \cdots < t_k$ be the roots of the monic orthogonal polynomial p_k, which lie in (a,b) by Theorem 11.10.5. Let $\ell_1, \ldots, \ell_k \in \mathbb{R}[x]_{\leq k-1}$ be the Lagrange polynomials for these nodes (Definition 6.6.1) and set

$$\alpha_i = \int_a^b \ell_i(t)\, w(t)\, \mathrm{d}t \qquad (i = 1, \ldots, k).$$

Then every α_i is **strictly positive**, and

$$\int_a^b f(t)\, w(t)\, \mathrm{d}t = \sum_{i=1}^{k} \alpha_i\, f(t_i)$$

for **every** $f \in \mathbb{R}[x]_{\leq 2k-1}$.

Idea. Two moves, one for each side of the identity. On the left, divide f by p_k: $f = qp_k + r$. The degree bookkeeping — $\deg f \leq 2k-1$ against $\deg p_k = k$ — forces $\deg q \leq k-1$, which is exactly the condition under which $\int qp_k w = \langle p_k, q\rangle$ vanishes. So the integral of f is the integral of its remainder. On the right, p_k vanishes at every node, so the sum also only sees the remainder. And the remainder has degree at most $k-1$, which is precisely the range on which k nodes reproduce a polynomial exactly. The two sides meet there.

Proof. Let $f \in \mathbb{R}[x]_{\leq 2k-1}$. Since $p_k \neq 0$, Theorem 6.1.1 gives $q, r \in \mathbb{R}[x]$ with

$$f = qp_k + r, \qquad \deg r < k.$$

If $q = 0$ then $q \in \mathbb{R}[x]_{\leq k-1}$ trivially. Otherwise $qp_k = f - r$ has degree at most $2k-1$, while $\deg(qp_k) = \deg q + k$ by Theorem 0.8.3; hence $\deg q \leq k-1$. Either way $q \in \mathbb{R}[x]_{\leq k-1}$, so Proposition 11.10.2 gives $\langle p_k, q\rangle = 0$, that is, $\int_a^b qp_k w = 0$. Therefore

$$\int_a^b fw = \int_a^b qp_k w + \int_a^b rw = \int_a^b rw. \tag{$*$}$$

At each node, $p_k(t_i) = 0$, so

$$f(t_i) = q(t_i)p_k(t_i) + r(t_i) = r(t_i) \qquad (i = 1, \ldots, k). \tag{$**$}$$

Since $\deg r \leq k-1$ and $t_1, \ldots, t_k$ are k distinct points, Theorem 6.6.1 (b) gives $r = \sum_i r(t_i)\ell_i$. Integrating against w and using the definition of α_i,

$$\int_a^b rw = \sum_{i=1}^{k} r(t_i)\, \alpha_i = \sum_{i=1}^{k} \alpha_i f(t_i),$$

the last step by $(**)$. Combining with $(*)$ proves the quadrature identity.

For positivity, fix j and apply the identity to $f = \ell_j^2$, which is legitimate because $\deg \ell_j^2 = 2k-2 \leq 2k-1$. By Theorem 6.6.1 (a), $\ell_j(t_i) = \delta_{ij}$, so the right-hand side collapses to

α_j, while the left-hand side is $\langle \ell_j, \ell_j \rangle = \|\ell_j\|^2$. Since $\ell_j \neq 0$, positive definiteness gives $\alpha_j = \|\ell_j\|^2 > 0$.

■

Each hypothesis did one job. Orthogonality to lower degrees killed the quotient on the left; the nodes being roots of p_k killed it again on the right. That is also why the nodes cannot be moved: shift one and the second cancellation fails. Exercise (C2: The ceiling is 2k − 1) shows $2k-1$ is a ceiling, not just what this proof happens to reach.

Example. Two-point Gauss–Legendre

Take $w = 1$ on $(-1, 1)$ and $k = 2$. Find the nodes and weights, and test the resulting rule on $f = 1 + 2x + 3x^2 + 4x^3$ and on $f = x^4$.

Solution. The monic p_2 is $x^2 - \frac{1}{3}$ (Example (Gram–Schmidt on $1, x, x^2$)), so the nodes are $t_1 = -1/\sqrt{3}$ and $t_2 = 1/\sqrt{3}$. The Lagrange polynomials are $\ell_1 = (x - t_2)/(t_1 - t_2)$ and $\ell_2 = (x - t_1)/(t_2 - t_1)$, with $t_1 - t_2 = -2/\sqrt{3}$. Since $\int_{-1}^{1} t\,\mathrm{d}t = 0$ and $\int_{-1}^{1} 1\,\mathrm{d}t = 2$,

$$\alpha_1 = \frac{0 - 2t_2}{t_1 - t_2} = \frac{-2/\sqrt{3}}{-2/\sqrt{3}} = 1,$$

and $\alpha_2 = 1$ by the same computation with the roles swapped. The rule is

$$\int_{-1}^{1} f(t)\,\mathrm{d}t = f\Big(-\tfrac{1}{\sqrt{3}}\Big) + f\Big(\tfrac{1}{\sqrt{3}}\Big) \qquad \text{for } f \in \mathbb{R}[x]_{\leq 3}.$$

Two evaluations, no coefficients to remember, every cubic exactly.

The cubic test. Exactly, $\int_{-1}^{1}(1 + 2t + 3t^2 + 4t^3)\mathrm{d}t = 2 + 0 + 3 \cdot \frac{2}{3} + 0 = 4$. By the rule, $f(-1/\sqrt{3}) = 2 - \frac{10\sqrt{3}}{9}$ and $f(1/\sqrt{3}) = 2 + \frac{10\sqrt{3}}{9}$, whose sum is 4. The odd parts cancel and the constant and quadratic parts contribute $1 + 1$ and $1 + 1$.

The quartic test. Exactly, $\int_{-1}^{1} t^4\,\mathrm{d}t = \frac{2}{5}$. By the rule, $(1/\sqrt{3})^4 + (1/\sqrt{3})^4 = \frac{2}{9}$. These differ, so degree $4 = 2k$ is already out of reach. Compare Example (Integrals and Slopes from Values), where the nodes were fixed in advance and **three** of them guaranteed exactness only up to degree 2 — guaranteed by the construction, that is; that particular symmetric rule happens to integrate cubics exactly as well, which is luck of the symmetry and not something the construction promises.

Warning. **"Exact" means exact on polynomials.** Theorem 11.10.6 says nothing about $\int_a^b fw$ for a general f, and nothing about what happens as k grows; on a nonpolynomial the rule returns an approximation whose error is a question for analysis. What linear algebra delivers is exactness on a subspace, and the reason that subspace is the largest available.

11.10.6 The classical families

Four weights account for most of the orthogonal polynomials that occur in practice. The table records them; the entries are statements, and Exercise (B2: Two rows of the table) asks you to verify two of the rows by integration.

Family	Interval	Weight $w(t)$	p_1	p_2
Legendre	$(-1,1)$	1	x	$x^2-\frac{1}{3}$
Chebyshev	$(-1,1)$	$(1-t^2)^{-1/2}$	x	$x^2-\frac{1}{2}$
Hermite	$(-\infty,\infty)$	e^{-t^2}	x	$x^2-\frac{1}{2}$
Laguerre	$(0,\infty)$	e^{-t}	$x-1$	x^2-4x+2

The last two columns give the monic members of degrees 1 and 2; $p_0=1$ in every row. The Chebyshev weight is unbounded at ±1 and the Hermite and Laguerre intervals are unbounded, so in those three rows the integrals defining the inner product are improper; they converge, which is the finiteness condition our definition of a weight demanded. The first three weights are even on symmetric intervals, so their a_k all vanish and their members alternate in parity; the Laguerre weight is not, and $a_0=1$ there. Section 9 of this chapter meets the Chebyshev family again from the Fourier side, where the substitution $t=\cos\theta$ turns its awkward weight into $\mathrm{d}\theta$.

11.10.7 Exercises

11.10.7.1 A. Check your understanding

Exercise. A1

(a) Define what it means for $p_0,p_1,p_2,\dots$ to be a **monic orthogonal polynomial sequence** for an inner product on $\mathbb{R}[x]$.

(b) State the three-term recurrence, including the formulas for a_k and b_k and the starting values.

(c) True or false: $b_k>0$ for every $k\geq1$. Justify your answer.

(d) For $\langle p,q\rangle=\int_{-1}^{1}pq$, how many roots does p_7 have, of what kind, and where?

(e) True or false: for **every** inner product on $\mathbb{R}[x]$ one has $\langle xp,q\rangle=\langle p,xq\rangle$. Justify your answer.

(f) How many nodes does a Gaussian rule need in order to integrate every polynomial of degree at most 11 exactly?

Solution.

(a) A sequence with $\deg p_k=k$ and leading coefficient 1 for every k, and $\langle p_j,p_k\rangle=0$ whenever $j\neq k$ (Definition 11.10.1).

(b) With $p_{-1}=0$, $p_0=1$, and $p_1=x-a_0$, one has $p_{k+1}=(x-a_k)p_k-b_kp_{k-1}$ for $k\geq1$, where $a_k=\langle xp_k,p_k\rangle/\langle p_k,p_k\rangle$ and $b_k=\langle p_k,p_k\rangle/\langle p_{k-1},p_{k-1}\rangle$ (Theorem 11.10.3).

(c) True. Both $\langle p_k,p_k\rangle=\|p_k\|^2$ and $\langle p_{k-1},p_{k-1}\rangle=\|p_{k-1}\|^2$ are strictly positive, since p_k and p_{k-1} are nonzero and the inner product is positive definite.

(d) Seven roots, all real, all distinct, all in the open interval $(-1,1)$ (Theorem 11.10.5). In

particular p_7 has no complex roots and no repeated ones.

(e) False. Take the coefficient inner product $\langle p, q\rangle = \sum_k u_k v_k$ on $\mathbb{R}[x]$, where $p = \sum_k u_k x^k$ and $q = \sum_k v_k x^k$ (Example (A functional with no Riesz vector)). Then $\langle x \cdot 1, x\rangle = \langle x, x\rangle = 1$, while $\langle 1, x \cdot x\rangle = \langle 1, x^2\rangle = 0$. The identity is a property of the weight and node constructions (Proposition 11.10.1), not of inner products in general.

(f) Six. A k-point rule is exact up to degree $2k-1$, and $2 \cdot 6 - 1 = 11$.

11.10.7.2 B. Practice

Exercise. B1: A weighted interval

On $\mathbb{R}[x]$ put $\langle p, q\rangle = \int_0^1 p(t)q(t)\,t\,dt$. Compute the monic p_0, p_1, and p_2, and verify directly that $p_2 \perp \mathbb{R}[x]_{\le 1}$.

Solution. The moments are $\langle x^k, 1\rangle = \int_0^1 t^{k+1}dt = 1/(k+2)$, so $\frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}$ for $k = 0, 1, 2, 3$.

$p_0 = 1$. Then $a_0 = \langle x, 1\rangle/\langle 1, 1\rangle = (1/3)/(1/2) = \frac{2}{3}$, so $p_1 = x - \frac{2}{3}$, and

$$\langle p_1, p_1\rangle = \tfrac{1}{4} - \tfrac{4}{3}\cdot\tfrac{1}{3} + \tfrac{4}{9}\cdot\tfrac{1}{2} = \tfrac{1}{4} - \tfrac{2}{9} = \tfrac{1}{36}.$$

Next $\langle xp_1, p_1\rangle = \int_0^1 (t^4 - \frac{4}{3}t^3 + \frac{4}{9}t^2)dt = \frac{1}{5} - \frac{1}{3} + \frac{4}{27} = \frac{2}{135}$, so

$$a_1 = \frac{2/135}{1/36} = \frac{8}{15}, \qquad b_1 = \frac{1/36}{1/2} = \frac{1}{18},$$

and Theorem 11.10.3 gives

$$p_2 = \left(x - \tfrac{8}{15}\right)\left(x - \tfrac{2}{3}\right) - \tfrac{1}{18} = x^2 - \tfrac{6}{5}x + \tfrac{3}{10}.$$

Verification. $\langle p_2, 1\rangle = \int_0^1 (t^3 - \frac{6}{5}t^2 + \frac{3}{10}t)dt = \frac{1}{4} - \frac{2}{5} + \frac{3}{20} = 0$, and $\langle p_2, x\rangle = \int_0^1 (t^4 - \frac{6}{5}t^3 + \frac{3}{10}t^2)dt = \frac{1}{5} - \frac{3}{10} + \frac{1}{10} = 0$. Since $(1, x)$ spans $\mathbb{R}[x]_{\le 1}$, this is the whole check.

Exercise. B2: Two rows of the table

Verify the degree-1 and degree-2 entries of the table for two of the classical families.

(a) Chebyshev: $w(t) = (1-t^2)^{-1/2}$ on $(-1, 1)$. You may use, without proof, that $\int_{-1}^1 (1-t^2)^{-1/2}dt = \pi$ and $\int_{-1}^1 t^2(1-t^2)^{-1/2}dt = \pi/2$.

(b) Laguerre: $w(t) = e^{-t}$ on $(0, \infty)$. You may use, without proof, that $\int_0^\infty t^k e^{-t}\,dt = k!$.

Solution.

(a) The weight is even, so $\langle x^j, x^i\rangle = 0$ whenever $i + j$ is odd. Hence $\langle x, 1\rangle = 0$, and x is monic of degree 1 and orthogonal to $\mathbb{R}[x]_{\le 0}$, so $p_1 = x$ by Proposition 11.10.2. Write $p_2 = x^2 + \beta x + \gamma$. Then $\langle p_2, x\rangle = \langle x^2, x\rangle + \beta\langle x, x\rangle + \gamma\langle 1, x\rangle = 0 + \beta \cdot \frac{\pi}{2} + 0$, forcing $\beta = 0$; and $\langle p_2, 1\rangle = \frac{\pi}{2} + 0 + \gamma\pi = 0$, forcing $\gamma = -\frac{1}{2}$. So $p_2 = x^2 - \frac{1}{2}$.

(b) The moments are $\langle x^k, 1\rangle = k!$, so $1, 1, 2, 6$ for $k = 0, 1, 2, 3$. For $p_1 = x + \beta$: $\langle p_1, 1\rangle =$

$1+\beta=0$, so $p_1=x-1$. For $p_2=x^2+\beta x+\gamma$, the two conditions $\langle p_2,1\rangle=0$ and $\langle p_2,x\rangle=0$ read

$$2+\beta+\gamma=0, \qquad 6+2\beta+\gamma=0.$$

Subtracting gives $\beta=-4$, then $\gamma=2$. So $p_2=x^2-4x+2$, whose roots $2\pm\sqrt{2}$ we already located in the Quick check above.

Exercise. B3: One more Legendre polynomial

Continue Example (The monic Legendre polynomials, continued): compute $\langle p_3,p_3\rangle$ for $p_3=x^3-\frac{3}{5}x$, then use Theorem 11.10.3 to find p_4. Verify that $\langle p_4,1\rangle=\langle p_4,x^2\rangle=0$.

Solution.

$$\langle p_3,p_3\rangle=\int_{-1}^{1}\Big(t^6-\tfrac{6}{5}t^4+\tfrac{9}{25}t^2\Big)\mathrm{d}t=\tfrac{2}{7}-\tfrac{6}{5}\cdot\tfrac{2}{5}+\tfrac{9}{25}\cdot\tfrac{2}{3}=\tfrac{2}{7}-\tfrac{6}{25}=\tfrac{8}{175}.$$

Since the weight is even, $a_3=0$, and $b_3=\langle p_3,p_3\rangle/\langle p_2,p_2\rangle=(8/175)/(8/45)=\frac{9}{35}$. Hence

$$p_4=x\Big(x^3-\tfrac{3}{5}x\Big)-\tfrac{9}{35}\Big(x^2-\tfrac{1}{3}\Big)=x^4-\tfrac{6}{7}x^2+\tfrac{3}{35}.$$

Verification. $\langle p_4,1\rangle=\frac{2}{5}-\frac{6}{7}\cdot\frac{2}{3}+\frac{3}{35}\cdot 2=\frac{14-20+6}{35}=0$, and

$$\langle p_4,x^2\rangle=\tfrac{2}{7}-\tfrac{6}{7}\cdot\tfrac{2}{5}+\tfrac{3}{35}\cdot\tfrac{2}{3}=\tfrac{10-12+2}{35}=0.$$

Orthogonality to x and x^3 is automatic by parity.

Exercise. B4: A two-point rule on the unit interval

Work on $(0,1)$ with $w=1$. Exercise B2 of Section 2 (Exercise (B2: Gram–Schmidt on polynomials)) found the monic orthogonal polynomials 1, $x-\frac{1}{2}$, $x^2-x+\frac{1}{6}$.

(a) Write down the two-point Gaussian rule for $\int_0^1 f$.

(b) Check it on $f=x^3$, and show that it fails on $f=x^4$.

Solution.

(a) The nodes are the roots of $x^2-x+\frac{1}{6}$, namely

$$t_{1,2}=\tfrac{1}{2}\mp\tfrac{\sqrt{3}}{6}, \qquad t_2-t_1=\tfrac{\sqrt{3}}{3},$$

both in $(0,1)$, as Theorem 11.10.5 promises. With $\ell_1=(x-t_2)/(t_1-t_2)$ and $\int_0^1 t\,\mathrm{d}t=\frac{1}{2}$,

$$\alpha_1=\frac{\frac{1}{2}-t_2}{t_1-t_2}=\frac{-\sqrt{3}/6}{-\sqrt{3}/3}=\tfrac{1}{2},$$

and $\alpha_2=\frac{1}{2}$ likewise. So $\int_0^1 f=\frac{1}{2}(f(t_1)+f(t_2))$ for $f\in\mathbb{R}[x]_{\le 3}$.

(b) Write $t_{1,2}=\frac{1}{2}\mp u$ with $u=\sqrt{3}/6$, so $u^2=\frac{1}{12}$. The odd powers of u cancel in a sum

over t_1 and t_2, so

$$t_1^3 + t_2^3 = 2\Big(\tfrac{1}{8} + 3 \cdot \tfrac{1}{2} \cdot \tfrac{1}{12}\Big) = 2 \cdot \tfrac{1}{4} = \tfrac{1}{2},$$

and half of that is $\frac{1}{4} = \int_0^1 t^3 \, dt$. For the quartic,

$$t_1^4 + t_2^4 = 2\Big(\tfrac{1}{16} + 6 \cdot \tfrac{1}{4} \cdot \tfrac{1}{12} + \tfrac{1}{144}\Big) = 2 \cdot \tfrac{28}{144} = \tfrac{7}{18},$$

so the rule returns $\frac{7}{36}$, whereas $\int_0^1 t^4 \, dt = \frac{1}{5}$. The rule is exact to degree $3 = 2k - 1$ and no further.

11.10.7.3 C. Going deeper

Exercise. C1: Interlacing roots

Let w be a weight on (a, b) and let (p_k) be its monic orthogonal polynomial sequence. The roots of p_k are said to **interlace** those of p_{k+1} when exactly one root of p_k lies strictly between each pair of consecutive roots of p_{k+1}. This is true for every k; prove the first two cases by hand.

(a) Let r be the root of p_1 and $s_1 < s_2$ the roots of p_2. Prove that $s_1 < r < s_2$.

(b) Let $u_1 < u_2 < u_3$ be the roots of p_3. Prove that $u_1 < s_1 < u_2 < s_2 < u_3$.

Hint: evaluate the three-term recurrence at a root.

Solution.

(a) By Theorem 11.10.3 with $k = 1$, $p_2 = (x - a_1)p_1 - b_1 p_0 = (x - a_1)(x - r) - b_1$. Evaluating at r gives $p_2(r) = -b_1 < 0$, since $b_1 > 0$. By Theorem 11.10.5, $p_2 = (x - s_1)(x - s_2)$ with $s_1 < s_2$ real, and this product is negative exactly on (s_1, s_2). Hence $s_1 < r < s_2$.

(b) By Theorem 11.10.3 with $k = 2$, $p_3 = (x - a_2)p_2 - b_2 p_1$. At a root of p_2 the first term dies, so

$$p_3(s_i) = -b_2 \, p_1(s_i) = -b_2(s_i - r) \qquad (i = 1, 2).$$

By (a), $s_1 - r < 0$ and $s_2 - r > 0$, and $b_2 > 0$, so $p_3(s_1) > 0$ and $p_3(s_2) < 0$.

By Theorem 11.10.5, $p_3 = (x - u_1)(x - u_2)(x - u_3)$ with $u_1 < u_2 < u_3$. Such a product is negative on $(-\infty, u_1)$, positive on (u_1, u_2), negative on (u_2, u_3), and positive on (u_3, ∞). So $p_3(s_1) > 0$ puts s_1 in (u_1, u_2) or in (u_3, ∞), and $p_3(s_2) < 0$ puts s_2 in $(-\infty, u_1)$ or in (u_2, u_3). The option $s_1 > u_3$ is impossible: it would give $s_2 > s_1 > u_3$, where p_3 is positive, contradicting $p_3(s_2) < 0$. The option $s_2 < u_1$ is impossible too: it would give $s_1 < s_2 < u_1$, where p_3 is negative, contradicting $p_3(s_1) > 0$. Hence $u_1 < s_1 < u_2$ and $u_2 < s_2 < u_3$, which is the assertion.

The general case is not proved this way. The coefficients a_k and b_k assemble into a symmetric tridiagonal matrix whose characteristic polynomial is p_k, and the interlacing then follows from the interlacing theorem for the eigenvalues of a symmetric matrix and its leading submatrices, which is Chapter 17's business.

Exercise. C2: The ceiling is 2k − 1

Let w be a weight on (a, b) and let $k \geq 1$.

(a) Let $t_1, \dots, t_k$ be **any** k distinct real numbers and $\alpha_1, \dots, \alpha_k$ **any** real numbers. Prove that there is an $f \in \mathbb{R}[x]_{\leq 2k}$ with $\int_a^b f w \neq \sum_i \alpha_i f(t_i)$.

(b) Hence deduce that the Gaussian rule of Theorem 11.10.6 is exact on $\mathbb{R}[x]_{\leq 2k-1}$ and on no larger space $\mathbb{R}[x]_{\leq d}$.

Solution.

(a) Put $g = (x - t_1) \cdots (x - t_k)$ and $f = g^2$, a polynomial of degree $2k$. Every node is a root of g, so $f(t_i) = 0$ for every i and the right-hand side is 0, whatever the α_i are. The left-hand side is $\int_a^b g^2 w = \langle g, g \rangle$, and $g \neq 0$, so $\langle g, g \rangle > 0$ by positive definiteness. The two sides differ.

(b) Exactness on $\mathbb{R}[x]_{\leq 2k-1}$ is Theorem 11.10.6. Suppose the rule were exact on $\mathbb{R}[x]_{\leq d}$ for some $d \geq 2k$. Since $\mathbb{R}[x]_{\leq 2k} \subseteq \mathbb{R}[x]_{\leq d}$, it would then be exact on every polynomial of degree at most $2k$. But its nodes are k distinct real numbers (Theorem 11.10.5), so (a) supplies a polynomial of degree $2k$ on which it is not. Hence no such d exists, and $2k - 1$ is the largest degree for which the rule is exact. The concrete instance $k = 2$, $f = x^4$ is the failure computed in Example (Two-point Gauss–Legendre).

Chapter 12

Spectral Theory

Take the unit circle of $\mathbb{R}^2$ and push it through a linear map. What comes out is an ellipse. The ellipse has two axes, at right angles to each other, and it is natural to ask whether those axes sit along directions that the map merely stretches — whether, in other words, the map's eigenvectors are perpendicular. Usually they are not. Chapter 9 found eigenvectors and asked only that they be independent; a basis of eigenvectors can be as skewed as it likes, and for a matrix like $\begin{pmatrix} 1 & 1 \\ 0 & 2 \end{pmatrix}$ it is. Perpendicular eigenvectors are a strictly stronger demand, and it is the demand that matters in practice: coordinates in an orthonormal basis are read off by one inner product each, the change of basis is an isometry, and nothing is distorted along the way.

This chapter answers the question completely. There is an orthonormal basis of eigenvectors exactly when the operator commutes with its adjoint over $\mathbb{C}$, and exactly when it equals its adjoint over $\mathbb{R}$. Those two statements are the **spectral theorems**, and they are the most used theorems in applied linear algebra. Everything else here grows out of them: the triangular form that every complex matrix has in an orthonormal basis (Schur), the shape of a rotation in three dimensions, several commuting operators diagonalized at once, functions of an operator computed on its spectrum, the discrete Fourier transform as one instance of the whole picture, and finally a number that measures how far a matrix is from being normal. **The field is $\mathbb{R}$ or $\mathbb{C}$ throughout**, as in Chapter 11, and every statement says which one it needs. The difference between the two is not cosmetic: the Fundamental Theorem of Algebra hands $\mathbb{C}$ an eigenvalue for free, and over $\mathbb{R}$ each eigenvalue has to be earned.

What you need. From Chapter 6, the Fundamental Theorem of Algebra and the splitting it gives (Theorem 6.5.1, Corollary 6.5.5), together with Lagrange interpolation (Theorem 6.6.1). From Chapter 9, eigenvalues and eigenspaces, the characteristic polynomial and multiplicities, the existence of an eigenvalue over $\mathbb{C}$ (Definition 9.2.1, Theorem 9.2.1, Theorem 9.2.4), triangularization and what its diagonal records (Theorem 9.7.2, Corollary 9.7.3, Theorem 9.7.4, Theorem 9.3.3), the conjugate-pair behavior of real matrices (Theorem 9.2.5), and the two results on commuting operators (Theorem 9.9.3, Theorem 9.9.5). From Chapter 1, the complexification of a real vector space (Definition 1.9.2, Theorem 1.9.3). From Chapter 0, the trace with its cyclic invariance, the transpose, polynomials of a matrix, and the roots of unity that §09's Fourier matrix is built from (Theorem 0.9.6, Theorem 0.9.5, Theorem 0.9.9, Example (Roots of Unity)). From Chapter 2, the bookkeeping every proof here does silently: a linear map is free on a basis (Theorem 2.1.5), polynomials of an operator (Definition 2.4.2), and invertibility (Theorem 2.5.7). From Chapter 3, permutation matrices (Lemma 3.4.7) and the Invertible Matrix Theorem (Theorem 3.4.4). From Chapter 4, the matrix of a map and of a composition with the isomorphism onto matrices (Definition 4.1.1, Theorem 4.1.2, Theorem 4.1.4), and change of coordinates with

similarity (Definition 4.2.1, Theorem 4.2.3, Theorem 4.2.4). From Chapter 7, determinants: the triangular and block triangular rules, multiplicativity, the invertibility criterion, and the characteristic polynomial with its similarity invariance (Theorem 7.4.6, Theorem 7.8.4, Theorem 7.5.4, Theorem 7.5.2, Corollary 7.5.7, Definition 7.10.2, Theorem 7.10.2, Theorem 7.10.3). From Chapter 10, the Jordan form and the rotation-scaling block (Theorem 10.3.1, Definition 10.4.1). From Chapter 8, the Kronecker product and the vec operator (Definition 8.4.1, Theorem 8.4.1, Definition 8.4.2, Theorem 8.4.4). Above all, from Chapter 11: orthonormal bases and Gram–Schmidt (Theorem 11.2.2, Corollary 11.2.3, Theorem 11.2.1), orthogonal complements and projections (Theorem 11.3.2, Definition 11.3.2, Theorem 11.3.3), the adjoint with its arithmetic and its matrix (Definition 11.6.1, Theorem 11.6.2, Theorem 11.6.3, Theorem 11.6.5), the two classes of operators it singles out (Definition 11.6.2, Definition 11.6.3, Proposition 11.6.7), and isometries (Definition 11.7.1, Theorem 11.7.3, Definition 11.7.2, Definition 11.7.3).

12.0.1 Roadmap

- **Schur's unitary triangularization.** Every complex matrix is upper triangular in some orthonormal basis; the workhorse the rest of the chapter runs on, and the real block version alongside it.
- **Self-adjoint operators.** Real eigenvalues over both fields, orthogonal eigenspaces, invariant complements, and the test $\langle Tv, v\rangle = 0$ that forces $T = 0$.
- **Normal operators.** The weakest hypothesis that works over $\mathbb{C}$: four equivalent descriptions, $\ker T = \ker T^*$, and why a normal operator has no Jordan blocks.
- **The complex spectral theorem.** Normal is exactly the same as having an orthonormal eigenbasis; Schur plus one entry-by-entry computation.
- **The real spectral theorem.** Self-adjoint is exactly the same as having an orthonormal eigenbasis over $\mathbb{R}$; the principal axis theorem, and the rotation that shows why "normal" is not enough here.
- **The shape of an orthogonal matrix.** Blocks $[1]$, $[-1]$, and plane rotations; every element of SO(3) is a turn about an axis, and reflections generate.
- **Commuting families.** One orthonormal basis that diagonalizes every member of a commuting family of normal operators at once.
- **Spectral resolutions and functions of a normal operator.** Writing $T = \sum \lambda_i P_i$ with orthogonal projections, and defining $f(T)$ by applying f to the spectrum.
- **Circulants and the discrete Fourier transform.** One unitary matrix diagonalizes every circulant, and multiplication becomes convolution.
- **The Sylvester equation and block diagonalization.** When $\mathbf{X} \mapsto \mathbf{AX} - \mathbf{XB}$ is invertible, and what that buys for block triangular matrices.
- **How far from normal?** Schur's inequality turns normality from a yes-or-no property into a measurement.

12.0.2 Named moves

- **Take one eigenvector, then pass to its orthogonal complement.** As soon as the operator is normal or self-adjoint, that complement is invariant, so the induction on dimension runs inside the space instead of in a quotient. This replaces Chapter 9's "extend a basis and pass to the quotient".
- **Compute $\langle Tv, v\rangle$ two ways.** Once with the eigenvalue equation, once by moving T across to T^*. Comparing the two is almost the whole of the elementary theory.
- **Work with $\|Tv\|^2 = \langle T^*Tv, v\rangle$.** A statement about lengths becomes a statement about the single operator T^*T, and a difference of two such operators is tested by the zero test.
- **Triangular plus normal forces diagonal.** Compare one diagonal entry of $\mathbf{A}^*\mathbf{A}$ with

the same entry of $\mathbf{A}\mathbf{A}^*$, and a row of off-diagonal entries has to vanish.
- **Split into real and imaginary parts.** Write $T = A + iB$ with A and B self-adjoint; then T is normal exactly when A and B commute, and questions about T become questions about a commuting pair.
- **Diagonalize, then read the answer off the diagonal.** Powers, square roots, exponentials, and the whole functional calculus are one line once an orthonormal eigenbasis is in hand.

12.1 Schur Triangularization

Chapter 9 proved that every operator on a nonzero complex vector space has an upper triangular matrix in **some** basis (Corollary 9.7.3). That basis was built by extending an eigenvector and passing to a quotient, and it came out as skewed as the operator happened to be. Now we have lengths and angles, so we can be fussier and ask for an **orthonormal** basis. The remarkable thing is that the answer is still yes, at no extra cost in hypotheses, and the proof is the same induction run inside an orthogonal complement rather than in a quotient. This section proves it, reads off the first consequences, and gives the real version, where a conjugate pair of eigenvalues costs one 2×2 block.

Throughout, V is a finite-dimensional inner product space over $F = \mathbb{R}$ or $F = \mathbb{C}$, and every matrix statement uses the standard inner product on F^n.

12.1.1 Why the obvious route fails

Here is the natural first attempt, and it is worth watching it fail, because the repair is the whole idea of the proof.

We want an orthonormal basis $\mathcal{B} = (v_1, \dots, v_n)$ with $[T]_{\mathcal{B}}^{\mathcal{B}}$ upper triangular, that is, with

$$Tv_j \in \operatorname{span}(v_1, \dots, v_j) \qquad \text{for every } j.$$

For $j = 1$ this says v_1 is an eigenvector of T, and over $\mathbb{C}$ one exists (Theorem 9.2.4). Normalize it, set $U = (\operatorname{span}(v_1))^{\perp}$, and try to finish by induction on U. The attempt stops at once: to speak of T restricted to U we need U to be T-invariant, and it need not be.

Example. An Eigenline Whose Complement Is Not Invariant

Let $\mathbf{A} = \begin{pmatrix} 1 & 1 \\ 0 & 2 \end{pmatrix}$ act on $\mathbb{C}^2$. Check that $\mathbf{e}_1$ is an eigenvector and decide whether $(\operatorname{span}(\mathbf{e}_1))^{\perp}$ is invariant.

Solution. $\mathbf{A}\mathbf{e}_1 = (1, 0) = \mathbf{e}_1$, so $\mathbf{e}_1$ is an eigenvector for the eigenvalue 1. The orthogonal complement of $\operatorname{span}(\mathbf{e}_1)$ is $\operatorname{span}(\mathbf{e}_2)$, and $\mathbf{A}\mathbf{e}_2 = (1, 2)$, which is not a multiple of $\mathbf{e}_2$. So $(\operatorname{span}(\mathbf{e}_1))^{\perp}$ is **not** invariant, and the induction has nothing to run on.

So the first basis vector is the wrong end to start from. Look at the target instead. Suppose $\mathcal{B}$ is orthonormal and $[T]_{\mathcal{B}}^{\mathcal{B}} = \mathbf{T}$ is upper triangular. By Theorem 11.6.3 the matrix of T^* in the same basis is $\mathbf{T}^*$, which is **lower** triangular; and the last column of a lower triangular matrix has only its bottom entry. Reading that column off,

$$T^* v_n = \overline{t_{nn}}\, v_n.$$

In the basis we are trying to build, the last vector is an eigenvector of T^* — not of T.

That is the vector to produce first, and it is exactly the one whose orthogonal complement behaves.

Lemma 12.1.1. Eigenvectors of the Adjoint Cut Out Invariant Subspaces

Let V be a finite-dimensional inner product space over F, let $T \in \mathcal{L}(V)$, and let $W \subseteq V$ be a T^*-invariant subspace. Then $W^\perp$ is T-invariant.

Proof. Let $u \in W^\perp$ and $w \in W$. Since W is T^*-invariant, $T^*w \in W$, so by Definition 11.6.1

$$\langle Tu, w\rangle = \langle u, T^*w\rangle = 0.$$

As $w \in W$ was arbitrary, $Tu \in W^\perp$. Hence $T(W^\perp) \subseteq W^\perp$.

■

The special case we need is $W = \operatorname{span}(v)$ with $T^*v = \mu v$: one vector, one line, and an invariant complement of dimension $n-1$ to recurse into. And over $\mathbb{C}$ such a v always exists, because T^* is an operator on a nonzero complex space and the Fundamental Theorem of Algebra gives it an eigenvalue. That is where the field enters, and it is the only place it enters.

12.1.2 The theorem

Theorem 12.1.2. Schur's Unitary Triangularization

Let V be a finite-dimensional inner product space over $\mathbb{C}$ with $\dim V = n \geq 1$, and let $T \in \mathcal{L}(V)$.

(a) There is an **orthonormal** basis $\mathcal{B}$ of V such that $[T]_{\mathcal{B}}^{\mathcal{B}}$ is upper triangular.

(b) More precisely, let $\lambda_1, \ldots, \lambda_n$ be the eigenvalues of T, each listed as often as its algebraic multiplicity, in **any** order we please. Then $\mathcal{B}$ can be chosen so that the i-th diagonal entry of $[T]_{\mathcal{B}}^{\mathcal{B}}$ is λ_i.

Idea. Induction on $\dim V$, building the basis from its **last** vector backwards. ① Pick the eigenvalue λ_n we want in the bottom-right corner; the conjugate $\overline{\lambda_n}$ is then an eigenvalue of T^*, because starring turns a noninvertible operator into a noninvertible operator. ② Take a unit eigenvector e of T^* for $\overline{\lambda_n}$ and put $U = (\operatorname{span}(e))^\perp$; by Lemma 12.1.1 it is T-invariant, so $T|_U$ exists and lives on a space of dimension $n-1$. ③ Check that the characteristic polynomial of $T|_U$ is p_T with the factor $x - \lambda_n$ removed, so the inductive hypothesis may be applied with the shortened list. ④ Append e to the basis it returns.

Proof. We induct on $n = \dim V$. Since $\mathbb{C}$ is algebraically closed, p_T splits (Corollary 6.5.5), so a list $\lambda_1, \ldots, \lambda_n$ as in (b) exists and $p_T = (x - \lambda_1)\cdots(x - \lambda_n)$; proving (b) proves (a).

For $n = 1$, any unit vector v_1 is a basis, every 1×1 matrix is upper triangular, and its single entry is the single eigenvalue λ_1.

Let $n \geq 2$ and assume the statement for all inner product spaces of dimension $n-1$. Fix a list $\lambda_1, \dots, \lambda_n$ as in (b).

$\overline{\lambda_n}$ is an eigenvalue of T^*. Since λ_n is an eigenvalue of T, the operator $T - \lambda_n \operatorname{id}_V$ is not invertible (Theorem 9.2.1). By Theorem 11.6.2 (a), (b), and (d), its adjoint is $T^* - \overline{\lambda_n} \operatorname{id}_V$, and the adjoint of that is $T - \lambda_n \operatorname{id}_V$ again. If $T^* - \overline{\lambda_n} \operatorname{id}_V$ were invertible, then Theorem 11.6.2 (e) would make its adjoint $T - \lambda_n \operatorname{id}_V$ invertible too. So $T^* - \overline{\lambda_n} \operatorname{id}_V$ is not invertible, and $\overline{\lambda_n} \in \operatorname{spec}(T^*)$.

Choose an eigenvector of T^* for $\overline{\lambda_n}$; it is nonzero, so dividing by its norm gives a unit vector e with $T^* e = \overline{\lambda_n} e$. Put $U = (\operatorname{span}(e))^{\perp}$. By Theorem 11.3.2 (c), $\dim U = n - 1 \geq 1$, and U is T-invariant by Lemma 12.1.1. Write $S = T|_U \in \mathcal{L}(U)$, where U carries the inner product of V.

The bottom-right entry is λ_n. Using Definition 11.6.1 and conjugate-linearity in the second slot,

$$\langle Te, e \rangle = \langle e, T^* e \rangle = \langle e, \overline{\lambda_n} e \rangle = \lambda_n \|e\|^2 = \lambda_n.$$

The shortened list belongs to S. Take any orthonormal basis $(f_1, \dots, f_{n-1})$ of U (Theorem 11.2.2) and set $\mathcal{C} = (f_1, \dots, f_{n-1}, e)$, an orthonormal list of n vectors, hence an orthonormal basis of V by Theorem 11.1.4 and Theorem 1.6.4. Since U is T-invariant, Theorem 9.1.3 (b) applies with the basis $\mathcal{C}$ and gives $p_T = p_S \, p_{\bar{T}}$, where $\bar{T}$ is the induced operator on the one-dimensional quotient V/U. Its matrix is the single bottom-right entry of $[T]_{\mathcal{C}}^{\mathcal{C}}$, which is $\langle Te, e \rangle = \lambda_n$ by Theorem 11.2.1; so $p_{\bar{T}} = x - \lambda_n$. Canceling that factor in $\mathbb{C}[x]$,

$$p_S = (x - \lambda_1) \cdots (x - \lambda_{n-1}).$$

So $\lambda_1, \dots, \lambda_{n-1}$ is a list of the eigenvalues of S with their algebraic multiplicities.

Induct. By the inductive hypothesis there is an orthonormal basis $(v_1, \dots, v_{n-1})$ of U in which the matrix of S is upper triangular with diagonal $\lambda_1, \dots, \lambda_{n-1}$. Put $\mathcal{B} = (v_1, \dots, v_{n-1}, e)$, again an orthonormal basis of V for the reason given above. For $j \leq n-1$ we have $Tv_j = Sv_j \in \operatorname{span}(v_1, \dots, v_j)$, so the j-th column of $[T]_{\mathcal{B}}^{\mathcal{B}}$ vanishes below row j and has λ_j in row j. The n-th column has nothing below row n, and its entry there is $\langle Te, e \rangle = \lambda_n$. Hence $[T]_{\mathcal{B}}^{\mathcal{B}}$ is upper triangular with diagonal $\lambda_1, \dots, \lambda_n$. This proves the theorem.

■

Two hypotheses did all the work, and it is worth naming which. **Complex** produced the eigenvalue of T^*, through the Fundamental Theorem of Algebra; over $\mathbb{R}$ the theorem is false as stated, and the last part of this section says what survives. **Inner product** produced the invariant complement, through Lemma 12.1.1; without one we would be back in Chapter 9's quotient. Notice also what the theorem does **not** claim: nothing was assumed about T, and nothing beyond triangularity is promised. Diagonal is a different and much stronger request, and Sections 4 and 5 say exactly which operators grant it.

12.1.3 The matrix form

Translating through an orthonormal basis turns the theorem into a factorization. Recall that a square complex matrix is **unitary** when $\mathbf{U}^*\mathbf{U} = \mathbf{I}$ (Definition 11.7.2), and call two matrices **unitarily similar** when $\mathbf{B} = \mathbf{U}^*\mathbf{A}\mathbf{U}$ for some unitary $\mathbf{U}$; this is similarity by a matrix that is also an isometry, so it preserves the inner product as well as the operator.

Corollary 12.1.3. Schur Factorization

Let $\mathbf{A} \in M_n(\mathbb{C})$ with $n \geq 1$, and let $\lambda_1, \dots, \lambda_n$ be the eigenvalues of $\mathbf{A}$, each listed as often as its algebraic multiplicity, in any prescribed order. Then there are a unitary $\mathbf{U} \in M_n(\mathbb{C})$ and an upper triangular $\mathbf{T} \in M_n(\mathbb{C})$ with

$$\mathbf{A} = \mathbf{U}\mathbf{T}\mathbf{U}^*, \qquad t_{ii} = \lambda_i \quad (1 \leq i \leq n).$$

Proof. Apply Theorem 12.1.2 to $T_{\mathbf{A}}$ on $\mathbb{C}^n$ with the standard inner product and the given list: there is an orthonormal basis $\mathcal{B} = (\mathbf{u}_1, \dots, \mathbf{u}_n)$ of $\mathbb{C}^n$ with $\mathbf{T} \coloneqq [T_{\mathbf{A}}]_{\mathcal{B}}^{\mathcal{B}}$ upper triangular and $t_{ii} = \lambda_i$. Let $\mathbf{U}$ be the matrix whose j-th column is $\mathbf{u}_j$. Its columns are orthonormal, so $\mathbf{U}$ is unitary (Theorem 11.7.3 (f), Definition 11.7.2), and $\mathbf{U} = [\mathrm{id}]_{\mathcal{B}}^{\mathcal{E}}$ for the standard basis $\mathcal{E}$ (Definition 4.2.1). Since $[T_{\mathbf{A}}]_{\mathcal{E}}^{\mathcal{E}} = \mathbf{A}$, Theorem 4.2.3 gives $\mathbf{T} = \mathbf{U}^{-1}\mathbf{A}\mathbf{U} = \mathbf{U}^*\mathbf{A}\mathbf{U}$, the last step because $\mathbf{U}^{-1} = \mathbf{U}^*$ for a unitary matrix. Rearranging gives $\mathbf{A} = \mathbf{U}\mathbf{T}\mathbf{U}^*$. ∎

The order of the diagonal is free because at each step of the induction we chose which eigenvalue to send to the bottom-right corner, and any of them was available. That freedom is used constantly: it lets us put a distinguished eigenvalue first, or collect equal eigenvalues into a block.

Remark. **How to compute one.** The induction above runs from the bottom-right corner, because that is the corner where an operator on a subspace is available. For a **matrix** there is a second route, running from the top-left, and it is the one to use by hand. Let $\mathbf{q}_1$ be a unit eigenvector of $\mathbf{A}$ for an eigenvalue λ, extend it to an orthonormal basis of $\mathbb{C}^n$ (Corollary 11.2.3) and let $\mathbf{Q}$ be the unitary matrix with those columns. Then $\mathbf{Q}^*\mathbf{A}\mathbf{q}_1 = \lambda\mathbf{Q}^*\mathbf{q}_1 = \lambda\mathbf{e}_1$, so

$$\mathbf{Q}^*\mathbf{A}\mathbf{Q} = \begin{pmatrix} \lambda & \mathbf{b}^* \\ \mathbf{0} & \mathbf{A}_1 \end{pmatrix}$$

for some $\mathbf{b} \in \mathbb{C}^{n-1}$ and $\mathbf{A}_1 \in M_{n-1}(\mathbb{C})$. Triangularize $\mathbf{A}_1 = \mathbf{V}\mathbf{T}_1\mathbf{V}^*$ and put $\mathbf{U} = \mathbf{Q}(1 \oplus \mathbf{V})$; the product is unitary and $\mathbf{U}^*\mathbf{A}\mathbf{U}$ is upper triangular. This is legitimate for matrices precisely because we may replace $\mathbf{A}$ by the compressed block $\mathbf{A}_1$, which is not the restriction of any operator: $(\mathrm{span}(\mathbf{q}_1))^{\perp}$ is not invariant, as Example (An Eigenline Whose Complement Is Not Invariant) showed.

Example. A Two-by-Two Schur Factorization

Find a unitary $\mathbf{U}$ and an upper triangular $\mathbf{T}$ with $\mathbf{U}^*\mathbf{A}\mathbf{U} = \mathbf{T}$, as in Corollary 12.1.3, for

$$\mathbf{A} = \begin{pmatrix} 2 & -1 \\ 1 & 4 \end{pmatrix}.$$

Solution. Here $\operatorname{tr}\mathbf{A} = 6$ and $\det\mathbf{A} = 8 + 1 = 9$, so $p_{\mathbf{A}} = x^2 - 6x + 9 = (x-3)^2$ and the only eigenvalue is 3, with algebraic multiplicity 2.

Solving $(\mathbf{A}-3\mathbf{I})\mathbf{x}=\mathbf{0}$ with $\mathbf{A}-3\mathbf{I}=\begin{pmatrix}-1 & -1\\ 1 & 1\end{pmatrix}$ gives the single eigenline $\operatorname{span}\big((1,-1)\big)$. Take $\mathbf{q}_1=\frac{1}{\sqrt{2}}(1,-1)$ and complete with $\mathbf{q}_2=\frac{1}{\sqrt{2}}(1,1)$, a unit vector orthogonal to $\mathbf{q}_1$. With $\mathbf{U}$ the matrix of those columns,

$$\mathbf{U}=\tfrac{1}{\sqrt{2}}\begin{pmatrix}1 & 1\\ -1 & 1\end{pmatrix},\qquad \mathbf{U}^*\mathbf{A}\mathbf{U}=\begin{pmatrix}3 & -2\\ 0 & 3\end{pmatrix}.$$

Both diagonal entries are 3, as they must be. The off-diagonal -2 cannot be removed: a matrix unitarily similar to $\operatorname{diag}(3,3)=3\mathbf{I}$ would equal $\mathbf{U}(3\mathbf{I})\mathbf{U}^*=3\mathbf{I}$, and $\mathbf{A}\neq 3\mathbf{I}$.

Example. A Three-by-Three Schur Factorization

Find a unitary $\mathbf{U}$ and an upper triangular $\mathbf{T}$ with $\mathbf{U}^*\mathbf{A}\mathbf{U}=\mathbf{T}$, as in Corollary 12.1.3, for

$$\mathbf{A}=\begin{pmatrix}1 & 0 & 1\\ 0 & 1 & -2\\ 4 & -3 & -2\end{pmatrix}.$$

Solution. Step 1: peel off one eigenvector. Expanding $\det(x\mathbf{I}-\mathbf{A})$ gives $p_{\mathbf{A}}=(x-1)(x-3)(x+4)$, so 1 is an eigenvalue. Solving $(\mathbf{A}-\mathbf{I})\mathbf{x}=\mathbf{0}$ by row reduction gives the eigenline spanned by $(3,4,0)$, of norm 5. Put

$$\mathbf{q}_1=\tfrac{1}{5}(3,4,0),\quad \mathbf{q}_2=\tfrac{1}{5}(4,-3,0),\quad \mathbf{q}_3=(0,0,1),$$

an orthonormal basis of $\mathbb{C}^3$ (each has norm 1 and the three pairwise inner products are 0; the entries are real, so no conjugates intrude). With $\mathbf{Q}=\begin{pmatrix}\mathbf{q}_1 & \mathbf{q}_2 & \mathbf{q}_3\end{pmatrix}$,

$$\mathbf{Q}^*\mathbf{A}\mathbf{Q}=\begin{pmatrix}1 & 0 & -1\\ 0 & 1 & 2\\ 0 & 5 & -2\end{pmatrix},\qquad \mathbf{A}_1=\begin{pmatrix}1 & 2\\ 5 & -2\end{pmatrix}.$$

Step 2: triangularize the block. Here $\operatorname{tr}\mathbf{A}_1=-1$ and $\det\mathbf{A}_1=-12$, so $p_{\mathbf{A}_1}=x^2+x-12=(x-3)(x+4)$. For the eigenvalue 3, $\mathbf{A}_1-3\mathbf{I}=\begin{pmatrix}-2 & 2\\ 5 & -5\end{pmatrix}$ gives the eigenvector $(1,1)$. Take $\mathbf{p}_1=\frac{1}{\sqrt{2}}(1,1)$ and $\mathbf{p}_2=\frac{1}{\sqrt{2}}(1,-1)$, and let $\mathbf{V}$ be the matrix of those columns. Then

$$\mathbf{V}^*\mathbf{A}_1\mathbf{V}=\begin{pmatrix}3 & 3\\ 0 & -4\end{pmatrix}.$$

Step 3: assemble. Put $\mathbf{U}=\mathbf{Q}(1\oplus\mathbf{V})$, whose columns are $\mathbf{q}_1$, $(\mathbf{q}_2+\mathbf{q}_3)/\sqrt{2}$, and $(\mathbf{q}_2-\mathbf{q}_3)/\sqrt{2}$:

$$\mathbf{U}=\tfrac{1}{5\sqrt{2}}\begin{pmatrix}3\sqrt{2} & 4 & 4\\ 4\sqrt{2} & -3 & -3\\ 0 & 5 & -5\end{pmatrix},\qquad \mathbf{U}^*\mathbf{A}\mathbf{U}=\begin{pmatrix}1 & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}}\\ 0 & 3 & 3\\ 0 & 0 & -4\end{pmatrix}.$$

The diagonal reads $1, 3, -4$, and indeed $p_{\mathbf{A}} = (x-1)(x-3)(x+4)$, with trace $0 = 1+3-4$ and determinant $-12 = 1 \cdot 3 \cdot (-4)$.

Quick check. Why does Theorem 12.1.2 not prove that every complex matrix is unitarily similar to a **diagonal** matrix?

Solution. Because the proof controls only the entries below the diagonal. The entries above it are whatever they happen to be, and Example (A Two-by-Two Schur Factorization) shows they need not vanish: the matrix there has a -2 in the corner, and no unitary can remove it. A matrix unitarily similar to a diagonal one satisfies strong conditions — Section 4 identifies them exactly — and most matrices do not.

12.1.4 Reading the diagonal

A triangular form with an explicit factorization makes two Chapter 9 facts fall out in a line each.

Corollary 12.1.4. Trace Is the Sum of the Eigenvalues

Let $\mathbf{A} \in M_n(\mathbb{C})$ have eigenvalues $\lambda_1, \dots, \lambda_n$, listed with algebraic multiplicity. Then $\operatorname{tr} \mathbf{A} = \lambda_1 + \cdots + \lambda_n$.

Proof. Write $\mathbf{A} = \mathbf{U}\mathbf{T}\mathbf{U}^*$ as in Corollary 12.1.3. By Theorem 0.9.6 (3) applied to the pair $\mathbf{U}$ and $\mathbf{T}\mathbf{U}^*$,

$$\operatorname{tr} \mathbf{A} = \operatorname{tr}\big(\mathbf{U}(\mathbf{T}\mathbf{U}^*)\big) = \operatorname{tr}\big((\mathbf{T}\mathbf{U}^*)\mathbf{U}\big) = \operatorname{tr} \mathbf{T},$$

since $\mathbf{U}^*\mathbf{U} = \mathbf{I}$. The diagonal of $\mathbf{T}$ is the list $\lambda_1, \dots, \lambda_n$.

■

Corollary 12.1.5. Determinant Is the Product of the Eigenvalues

Let $\mathbf{A} \in M_n(\mathbb{C})$ have eigenvalues $\lambda_1, \dots, \lambda_n$, listed with algebraic multiplicity. Then $\det \mathbf{A} = \lambda_1 \cdots \lambda_n$.

Proof. With $\mathbf{A} = \mathbf{U}\mathbf{T}\mathbf{U}^*$ as above, Theorem 7.5.4 gives $\det \mathbf{A} = \det \mathbf{U} \det \mathbf{T} \det \mathbf{U}^*$, and $\det \mathbf{U} \det \mathbf{U}^* = \det(\mathbf{U}\mathbf{U}^*) = \det \mathbf{I} = 1$. So $\det \mathbf{A} = \det \mathbf{T}$, which is the product of the diagonal entries by Theorem 7.4.6.

■

Both statements are Theorem 9.3.3 over $\mathbb{C}$, recovered from the diagonal of $\mathbf{T}$ instead of from the coefficients of $p_{\mathbf{A}}$. Nothing is new; what is new is how short the argument became once a triangular form was available in hand.

The next consequence is not a restatement of anything. It says that the matrices we know how to handle sit arbitrarily close to the ones we do not.

Corollary 12.1.6. Diagonalizable Matrices Are Everywhere

Let $\mathbf{A} \in M_n(\mathbb{C})$ and let $\varepsilon > 0$. Then there is a **diagonalizable** $\mathbf{A}' \in M_n(\mathbb{C})$ with

$$|a_{ij} - a'_{ij}| < \varepsilon \qquad \text{for all } i, j.$$

In this sense every complex square matrix is a limit of diagonalizable ones.

Idea. Diagonalizability is hard to arrange directly, but n distinct eigenvalues force it (Corollary 9.4.2), and in a triangular form the eigenvalues are sitting on the diagonal where we can nudge them apart. The only thing to check is that a small nudge inside the triangular form is still a small nudge after conjugating back by $\mathbf{U}$, and that is Cauchy–Schwarz on the rows of a unitary matrix.

Proof. Write $\mathbf{A} = \mathbf{UTU}^*$ as in Corollary 12.1.3. Choose $\delta_1, \dots, \delta_n \in \mathbb{C}$ with $|\delta_k| < \varepsilon$ and with $t_{11} + \delta_1, \dots, t_{nn} + \delta_n$ pairwise distinct. This is possible one index at a time: having chosen $\delta_1, \dots, \delta_{k-1}$, the forbidden values of δ_k are the $k-1$ numbers $t_{jj} + \delta_j - t_{kk}$, and the disc $\{|z| < \varepsilon\}$ has infinitely many points.

Put $\mathbf{D} = \operatorname{diag}(\delta_1, \dots, \delta_n)$ and $\mathbf{A}' = \mathbf{U}(\mathbf{T} + \mathbf{D})\mathbf{U}^*$. The matrix $\mathbf{T} + \mathbf{D}$ is upper triangular with n distinct diagonal entries, so by Theorem 9.7.4 (b) it has n distinct eigenvalues; $\mathbf{A}'$ is similar to it, hence has the same eigenvalues, and is diagonalizable by Corollary 9.4.2.

For the entries, $\mathbf{A} - \mathbf{A}' = -\mathbf{UDU}^*$, whose (i, j) entry is $-\sum_k u_{ik}\delta_k\overline{u_{jk}}$. Hence

$$|a_{ij} - a'_{ij}| \le \max_k |\delta_k| \sum_{k=1}^{n} |u_{ik}|\,|u_{jk}| \le \max_k |\delta_k| < \varepsilon,$$

where the middle inequality is Theorem 11.1.5 applied to the vectors of moduli, together with the fact that the rows of $\mathbf{U}$ are unit vectors, which is $\mathbf{UU}^* = \mathbf{I}$. This proves the corollary. ∎

This is the tool behind a standard strategy: *prove it for diagonalizable matrices, then take a limit.* Chapter 16, where matrices acquire norms and limits can be spoken of properly, runs arguments of exactly this shape; and Exercise (C1: Cayley–Hamilton from the triangular factors) uses the triangular form directly, with no limits at all, to reprove the Cayley–Hamilton theorem over $\mathbb{C}$.

12.1.5 The real version

Over $\mathbb{R}$ the theorem fails, and it fails for the reason it always does: a rotation has no real eigenvector, so no real basis makes its matrix triangular (Theorem 9.7.2). The most one can hope for is to let a conjugate pair occupy a 2×2 block, and that much is true.

Theorem 12.1.7. Real Schur Form

Let $\mathbf{A} \in M_n(\mathbb{R})$ with $n \geq 1$. Then there are an orthogonal $\mathbf{Q} \in \mathrm{O}(n)$ and a matrix $\mathbf{T} \in M_n(\mathbb{R})$ with

$$\mathbf{A} = \mathbf{Q}\mathbf{T}\mathbf{Q}^\top,$$

where $\mathbf{T}$ is **block** upper triangular with diagonal blocks of size 1×1 and 2×2, and every 2×2 diagonal block has a conjugate pair of nonreal eigenvalues.

Idea. The same induction, with one change. The eigenvector of T^* may fail to exist over $\mathbb{R}$, but an invariant subspace of dimension 1 or 2 always does — that is Exercise (C2: Real operators have small invariant subspaces), proved in Chapter 9 from a complex eigenvector $\mathbf{x} + i\mathbf{y}$ by taking $\operatorname{span}(\mathbf{x}, \mathbf{y})$. Apply it to T^*, take the orthogonal complement, and induct. Dimension 1 occurs exactly when T^* has a real eigenvalue, and that is what keeps the 2×2 blocks honest.

Proof. Let $T = T_{\mathbf{A}}$ on $\mathbb{R}^n$ with the standard inner product, so that $T^* = T_{\mathbf{A}^\top}$ by Theorem 11.6.3. We induct on n.

Choose a T^*-invariant subspace W as follows. If T^* has a real eigenvalue, let W be a corresponding eigenline, so $d := \dim W = 1$. If it does not, take a complex eigenvalue $\mu = a + bi$ of $\mathbf{A}^\top$ with $b \neq 0$, which exists by Theorem 9.2.4 applied to $\mathbf{A}^\top$ over $\mathbb{C}$, and a complex eigenvector $\mathbf{x} + i\mathbf{y}$ with $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$. Comparing real and imaginary parts of $\mathbf{A}^\top(\mathbf{x} + i\mathbf{y}) = \mu(\mathbf{x} + i\mathbf{y})$ gives $\mathbf{A}^\top\mathbf{x} = a\mathbf{x} - b\mathbf{y}$ and $\mathbf{A}^\top\mathbf{y} = b\mathbf{x} + a\mathbf{y}$, so $W := \operatorname{span}(\mathbf{x}, \mathbf{y})$ is T^*-invariant. It has dimension 2: were $\mathbf{x}, \mathbf{y}$ dependent, W would be a T^*-invariant line and T^* would have a real eigenvalue, contrary to the case we are in. So $d = 2$. (Exercise (C2: Real operators have small invariant subspaces) runs the same argument in Chapter 9.)

Put $U = W^\perp$, which is T-invariant by Lemma 12.1.1 and has dimension $n-d$ (Theorem 11.3.2 (c)). Pick an orthonormal basis $(\mathbf{v}_1, \dots, \mathbf{v}_{n-d})$ of U and one $(\mathbf{v}_{n-d+1}, \dots, \mathbf{v}_n)$ of W (Theorem 11.2.2); together they form an orthonormal basis $\mathcal{B}$ of $\mathbb{R}^n$, since the two blocks are orthogonal to each other. Because $T(U) \subseteq U$,

$$[T]_{\mathcal{B}}^{\mathcal{B}} = \begin{pmatrix} \mathbf{C} & * \\ \mathbf{O} & \mathbf{B} \end{pmatrix}$$

with $\mathbf{C} \in M_{n-d}(\mathbb{R})$ the matrix of $T|_U$ and $\mathbf{B} \in M_d(\mathbb{R})$. Writing $\mathbf{Q}_{\mathcal{B}}$ for the matrix whose columns are the vectors of $\mathcal{B}$, which is orthogonal because those columns are orthonormal, Theorem 4.2.3 gives $\mathbf{Q}_{\mathcal{B}}^\top \mathbf{A} \mathbf{Q}_{\mathcal{B}} = [T]_{\mathcal{B}}^{\mathcal{B}}$.

The block $\mathbf{B}$ is of the allowed kind. If $d = 1$ it is 1×1. If $d = 2$, then by Theorem 11.6.3 the matrix of T^* in $\mathcal{B}$ is the transpose of the display, whose bottom-right block is $\mathbf{B}^\top$; as W is T^*-invariant, $\mathbf{B}^\top$ is the matrix of $T^*|_W$. A real eigenvalue of $\mathbf{B}^\top$ would be a real eigenvalue of T^*, which this case excludes. So $p_{\mathbf{B}} = p_{\mathbf{B}^\top}$ has no real root, and its two complex roots form a conjugate pair by Theorem 9.2.5.

Close the induction. If $U = \{\mathbf{0}\}$ then $n = d \leq 2$, the list $\mathcal{B}$ is a basis of W, and $\mathbf{Q}_{\mathcal{B}}^\top \mathbf{A} \mathbf{Q}_{\mathcal{B}} = \mathbf{B}$ is already a single allowed block; take $\mathbf{Q} = \mathbf{Q}_{\mathcal{B}}$. Otherwise $1 \leq n - d < n$, and by the inductive hypothesis $\mathbf{C} = \mathbf{Q}_1 \mathbf{T}_1 \mathbf{Q}_1^\top$ with $\mathbf{Q}_1 \in \mathrm{O}(n-d)$ and $\mathbf{T}_1$ of the required shape. Put $\mathbf{Q} = \mathbf{Q}_{\mathcal{B}}(\mathbf{Q}_1 \oplus \mathbf{I}_d)$, a product of orthogonal matrices and so orthogonal (Proposition 11.7.4 (a)). Then

$$\mathbf{Q}^\top \mathbf{A} \mathbf{Q} = (\mathbf{Q}_1 \oplus \mathbf{I}_d)^\top \begin{pmatrix} \mathbf{C} & * \\ \mathbf{O} & \mathbf{B} \end{pmatrix} (\mathbf{Q}_1 \oplus \mathbf{I}_d) = \begin{pmatrix} \mathbf{T}_1 & * \\ \mathbf{O} & \mathbf{B} \end{pmatrix},$$

which is block upper triangular with diagonal blocks those of $\mathbf{T}_1$ followed by $\mathbf{B}$. This proves the theorem.

■

Two remarks on the shape. The 1×1 blocks carry the real eigenvalues of $\mathbf{A}$ and the 2×2 blocks carry the nonreal ones in conjugate pairs, so the block sizes are forced by the spectrum, even though the blocks themselves are not. And when all the eigenvalues of $\mathbf{A}$ are real, every block is 1×1 and the real Schur form is an honest orthogonal triangularization.

Warning. **Neither T nor U is unique; only the diagonal multiset is.** For $\mathbf{A}=\begin{pmatrix}1 & 1\\ 0 & 2\end{pmatrix}$, taking $\mathbf{U}=\mathbf{I}$ gives $\mathbf{T}=\mathbf{A}$, with diagonal $(1,2)$. Taking instead $\mathbf{U}=\frac{1}{\sqrt{2}}\begin{pmatrix}1 & 1\\ 1 & -1\end{pmatrix}$ gives

$$\mathbf{U}^*\mathbf{A}\mathbf{U}=\begin{pmatrix}2 & -1\\ 0 & 1\end{pmatrix},$$

with diagonal $(2,1)$ and a different corner entry. So a Schur form is **not** a canonical form: it does not decide similarity, and two matrices can have the same triangular form without being equal. The canonical form for similarity is Jordan's (Theorem 10.3.1), and Jordan's cannot in general be reached unitarily — Example (A Two-by-Two Schur Factorization) is already a Jordan block in disguise whose off-diagonal entry is -2, not 1. What the Schur form gives up in uniqueness it buys back in stability: the change of basis is an isometry.

12.1.6 Exercises

12.1.6.1 A. Check your understanding

Exercise. A1

(a) State Theorem 12.1.2, including the field and every hypothesis.

(b) In the proof, an eigenvector of T^* is taken rather than one of T. Say why, in one sentence, and name the property of $(\operatorname{span}(v))^{\perp}$ that is wanted.

(c) True or false: if $\mathbf{A}\in M_n(\mathbb{C})$ is upper triangular in some orthonormal basis, then its eigenvalues are the diagonal entries of that triangular matrix. Justify your answer.

(d) Which step of the proof of Theorem 12.1.2 fails over $\mathbb{R}$?

(e) Give a matrix in $M_2(\mathbb{C})$ that is upper triangular in an orthonormal basis but is **not** diagonal in any.

Solution.

(a) Let V be a finite-dimensional inner product space over $\mathbb{C}$ with $\dim V\geq 1$ and let $T\in\mathcal{L}(V)$. Then V has an orthonormal basis $\mathcal{B}$ with $[T]_{\mathcal{B}}^{\mathcal{B}}$ upper triangular, and the diagonal may be made to read the eigenvalues of T, with algebraic multiplicity, in any prescribed order.

(b) Because the induction needs $(\operatorname{span}(v))^{\perp}$ to be T**-invariant**, and Lemma 12.1.1 gives

that exactly when $\operatorname{span}(v)$ is T^*-invariant, that is, when v is an eigenvector of T^*.

(c) True. The matrix of T in **any** basis has the same characteristic polynomial (Theorem 7.10.3), and for a triangular matrix that polynomial is the product of $x - t_{ii}$ (Theorem 9.7.4 (a)). Orthonormality is irrelevant to this particular point.

(d) The Fundamental Theorem of Algebra, twice: once so that p_T splits into the list of n roots the statement speaks of (Corollary 6.5.5), and once for the existence of an eigenvalue of T^* at each stage of the induction. It came from Theorem 9.2.4, whose proof rests on the Fundamental Theorem of Algebra (Theorem 6.5.1); over $\mathbb{R}$ the rotation $\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ has no eigenvalue at all. Every other step is valid over $\mathbb{R}$.

(e) $\mathbf{A} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$. It is already upper triangular in the standard orthonormal basis. It is not diagonal in any basis, orthonormal or not: its only eigenvalue is 0, so a diagonal form would be the zero matrix, and $\mathbf{A} \neq \mathbf{O}$.

12.1.6.2 B. Practice

Exercise. B1: A Schur form with complex entries

Let $\mathbf{A} = \begin{pmatrix} 1 & 2 \\ -1 & 3 \end{pmatrix}$, regarded as a complex matrix.

(a) Find the eigenvalues of $\mathbf{A}$.

(b) Find a unitary $\mathbf{U} \in M_2(\mathbb{C})$ with $\mathbf{U}^*\mathbf{A}\mathbf{U}$ upper triangular, and compute that triangular matrix.

(c) Explain why no **real** orthogonal $\mathbf{Q}$ makes $\mathbf{Q}^\top\mathbf{A}\mathbf{Q}$ upper triangular.

Solution.

(a) $\operatorname{tr}\mathbf{A} = 4$ and $\det\mathbf{A} = 3 + 2 = 5$, so $p_{\mathbf{A}} = x^2 - 4x + 5 = (x-2)^2 + 1$, with roots $2 + i$ and $2 - i$.

(b) For $\lambda = 2 + i$,

$$\mathbf{A} - \lambda\mathbf{I} = \begin{pmatrix} -1 - i & 2 \\ -1 & 1 - i \end{pmatrix}.$$

The second row gives $-x + (1 - i)y = 0$, so $(1 - i, 1)$ is an eigenvector; its norm is $\sqrt{2 + 1} = \sqrt{3}$. Put $\mathbf{q}_1 = \frac{1}{\sqrt{3}}(1 - i, 1)$. A unit vector orthogonal to it is $\mathbf{q}_2 = \frac{1}{\sqrt{3}}(1, -1 - i)$, since

$$\langle \mathbf{q}_2, \mathbf{q}_1 \rangle = \tfrac{1}{3}\big(1 \cdot \overline{1 - i} + (-1 - i) \cdot \overline{1}\big) = \tfrac{1}{3}\big((1 + i) - (1 + i)\big) = 0$$

and $\|\mathbf{q}_2\|^2 = \frac{1}{3}(1 + 2) = 1$. With $\mathbf{U} = \begin{pmatrix} \mathbf{q}_1 & \mathbf{q}_2 \end{pmatrix}$,

$$\mathbf{U}^*\mathbf{A}\mathbf{U} = \begin{pmatrix} 2 + i & -1 - 2i \\ 0 & 2 - i \end{pmatrix}.$$

(c) A real upper triangular matrix has its diagonal entries as eigenvalues (Theorem 9.7.4), and those entries are real. But $\mathbf{Q}^\top\mathbf{A}\mathbf{Q}$ is similar to $\mathbf{A}$, whose eigenvalues $2 \pm i$ are not

real. So no such $\mathbf{Q}$ exists. (This is Theorem 12.1.7 doing the only thing it can: here $\mathbf{T}$ is a single 2×2 block, and $\mathbf{Q} = \mathbf{I}$ already works.)

Exercise. B2: A three-by-three Schur form

Let $\mathbf{A} = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 3 \\ 0 & 0 & 4 \end{pmatrix}$.

(a) Verify that $(1, -1, 0)$ is an eigenvector and find the corresponding eigenvalue.

(b) Extend it to an orthonormal basis of $\mathbb{R}^3$ and compute $\mathbf{Q}^\top \mathbf{A} \mathbf{Q}$ for the resulting $\mathbf{Q}$.

(c) Hence write down the eigenvalues of $\mathbf{A}$ with their algebraic multiplicities.

Solution.

(a) $\mathbf{A}(1, -1, 0) = (2 - 1,\ 1 - 2,\ 0) = (1, -1, 0)$, so the eigenvalue is 1.

(b) Take $\mathbf{q}_1 = \frac{1}{\sqrt{2}}(1, -1, 0)$, $\mathbf{q}_2 = \frac{1}{\sqrt{2}}(1, 1, 0)$, and $\mathbf{q}_3 = (0, 0, 1)$. These are pairwise orthogonal unit vectors, hence an orthonormal basis of $\mathbb{R}^3$. With $\mathbf{Q} = \begin{pmatrix} \mathbf{q}_1 & \mathbf{q}_2 & \mathbf{q}_3 \end{pmatrix}$,

$$\mathbf{Q}^\top \mathbf{A} \mathbf{Q} = \begin{pmatrix} 1 & 0 & -\sqrt{2} \\ 0 & 3 & 2\sqrt{2} \\ 0 & 0 & 4 \end{pmatrix}.$$

This is already upper triangular, so a single step sufficed and $\mathbf{U} = \mathbf{Q}$. (In general the lower-right 2×2 block would need a second step, as in Example (A Three-by-Three Schur Factorization).)

(c) The diagonal is $1, 3, 4$, so by Theorem 9.7.4 these are the eigenvalues, each of algebraic multiplicity 1. As a check, $\operatorname{tr} \mathbf{A} = 8 = 1 + 3 + 4$ and $\det \mathbf{A} = 4(4 - 1) = 12 = 1 \cdot 3 \cdot 4$.

Exercise. B3: The field matters

Let $\mathbf{A} = \begin{pmatrix} a & -b \\ b & a \end{pmatrix} \in M_2(\mathbb{R})$ with $b \neq 0$.

(a) Find the eigenvalues of $\mathbf{A}$ over $\mathbb{C}$ and show that $\mathbf{A}$ is unitarily similar to a diagonal matrix in $M_2(\mathbb{C})$.

(b) Prove that there is no invertible $\mathbf{P} \in M_2(\mathbb{R})$ at all, orthogonal or otherwise, with $\mathbf{P}^{-1} \mathbf{A} \mathbf{P}$ upper triangular.

(c) Say what Theorem 12.1.7 produces for this $\mathbf{A}$.

Solution.

(a) $\operatorname{tr} \mathbf{A} = 2a$ and $\det \mathbf{A} = a^2 + b^2$, so $p_{\mathbf{A}} = x^2 - 2ax + a^2 + b^2$ with roots $a \pm bi$, distinct because $b \neq 0$. For $\lambda = a + bi$, $\mathbf{A}(1, -i) = (a + bi,\ b - ai) = (a + bi)(1, -i)$, so $\mathbf{q}_1 = \frac{1}{\sqrt{2}}(1, -i)$ is a unit eigenvector; likewise $\mathbf{q}_2 = \frac{1}{\sqrt{2}}(1, i)$ is a unit eigenvector for

$a - bi$, and

$$\langle \mathbf{q}_1, \mathbf{q}_2 \rangle = \tfrac{1}{2}(1 \cdot 1 + (-i)\bar{i}) = \tfrac{1}{2}(1 - 1) = 0.$$

So $\mathbf{U} = \begin{pmatrix} \mathbf{q}_1 & \mathbf{q}_2 \end{pmatrix}$ is unitary and $\mathbf{U}^*\mathbf{A}\mathbf{U} = \operatorname{diag}(a + bi,\ a - bi)$.

(b) Suppose $\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \mathbf{T}$ were upper triangular with $\mathbf{P} \in M_2(\mathbb{R})$ invertible. Then $\mathbf{T} \in M_2(\mathbb{R})$ is similar to $\mathbf{A}$, so $p_{\mathbf{T}} = p_{\mathbf{A}}$ (Theorem 7.10.3); but $p_{\mathbf{T}} = (x - t_{11})(x - t_{22})$ has the real roots t_{11}, t_{22}, while $p_{\mathbf{A}}$ has none. Contradiction.

(c) Since $\mathbf{A}$ has no real eigenvalue, the real Schur form of $\mathbf{A}$ is a single 2×2 diagonal block; $\mathbf{Q} = \mathbf{I}_2$ and $\mathbf{T} = \mathbf{A}$ already work. There is nothing to simplify: over $\mathbb{R}$ this matrix is as reduced as it gets.

12.1.6.3 C. Going deeper

Exercise. C1: Cayley–Hamilton from the triangular factors

Let $\mathbf{T} \in M_n(\mathbb{C})$ be upper triangular with diagonal entries $\lambda_1, \dots, \lambda_n$, and set $\mathbf{N}_k = (\mathbf{T} - \lambda_1\mathbf{I})(\mathbf{T} - \lambda_2\mathbf{I}) \cdots (\mathbf{T} - \lambda_k\mathbf{I})$ for $1 \le k \le n$.

(a) Prove by induction on k that the first k columns of $\mathbf{N}_k$ are zero.

(b) Deduce that $p_{\mathbf{T}}(\mathbf{T}) = \mathbf{O}$.

(c) Deduce the Cayley–Hamilton Theorem (Theorem 9.6.1) for every $\mathbf{A} \in M_n(\mathbb{C})$, using Corollary 12.1.3.

Hint for (a): the j-th column of $\mathbf{T} - \lambda_j\mathbf{I}$ has its only possibly nonzero entries in rows $1, \dots, j - 1$.

Solution.

(a) For $k = 1$: the first column of $\mathbf{T} - \lambda_1\mathbf{I}$ is $(\lambda_1 - \lambda_1, 0, \dots, 0) = \mathbf{0}$, since $\mathbf{T}$ is upper triangular.
Assume the first $k - 1$ columns of $\mathbf{N}_{k-1}$ are zero, where $2 \le k \le n$. Then $\mathbf{N}_k = \mathbf{N}_{k-1}(\mathbf{T} - \lambda_k\mathbf{I})$, so the j-th column of $\mathbf{N}_k$ is $\mathbf{N}_{k-1}\mathbf{c}_j$ with $\mathbf{c}_j$ the j-th column of $\mathbf{T} - \lambda_k\mathbf{I}$. For $j \le k - 1$ we get $\mathbf{N}_k\mathbf{e}_j = \mathbf{N}_{k-1}(\mathbf{T} - \lambda_k\mathbf{I})\mathbf{e}_j$, and since $(\mathbf{T} - \lambda_k\mathbf{I})\mathbf{e}_j \in \operatorname{span}(\mathbf{e}_1, \dots, \mathbf{e}_j)$ by upper triangularity, this is a combination of the first $j \le k - 1$ columns of $\mathbf{N}_{k-1}$, all zero. For $j = k$ the column $\mathbf{c}_k$ of $\mathbf{T} - \lambda_k\mathbf{I}$ has its (k, k) entry equal to $\lambda_k - \lambda_k = 0$ and zeros below, so $\mathbf{c}_k \in \operatorname{span}(\mathbf{e}_1, \dots, \mathbf{e}_{k-1})$, and $\mathbf{N}_{k-1}\mathbf{c}_k$ is a combination of the first $k - 1$ columns of $\mathbf{N}_{k-1}$, again zero. So all of the first k columns of $\mathbf{N}_k$ vanish.

(b) By Theorem 9.7.4 (a), $p_{\mathbf{T}} = (x - \lambda_1) \cdots (x - \lambda_n)$, so $p_{\mathbf{T}}(\mathbf{T}) = \mathbf{N}_n$. By (a) with $k = n$, all n columns of $\mathbf{N}_n$ are zero, that is $p_{\mathbf{T}}(\mathbf{T}) = \mathbf{O}$.

(c) Write $\mathbf{A} = \mathbf{U}\mathbf{T}\mathbf{U}^*$ by Corollary 12.1.3. Similar matrices have the same characteristic polynomial (Theorem 7.10.3), so $p_{\mathbf{A}} = p_{\mathbf{T}}$, and $\mathbf{A}^k = \mathbf{U}\mathbf{T}^k\mathbf{U}^*$ for every $k \ge 0$ because the inner factors $\mathbf{U}^*\mathbf{U}$ cancel. Hence

$$p_{\mathbf{A}}(\mathbf{A}) = \mathbf{U}\, p_{\mathbf{T}}(\mathbf{T})\, \mathbf{U}^* = \mathbf{U}\,\mathbf{O}\,\mathbf{U}^* = \mathbf{O}.$$

Chapter 9 reached the same conclusion from the flag of invariant subspaces (Exercise (C1: A second proof of Cayley–Hamilton)); here the whole argument happens inside one matrix product.

Exercise. C2: The Frobenius norm is unitarily invariant

Chapter 11 §07 introduced the Frobenius norm $\|\mathbf{A}\|_F = \left(\sum_{i,j}|a_{ij}|^2\right)^{1/2}$ and proved (a) and (b) below as Lemma 11.7.5. Prove them again here without looking, since the argument is two lines and (c) is what this section is for.

(a) Prove that $\|\mathbf{A}\|_F^2 = \operatorname{tr}(\mathbf{A}^*\mathbf{A})$.

(b) Prove that $\|\mathbf{UA}\|_F = \|\mathbf{A}\|_F = \|\mathbf{AV}\|_F$ whenever $\mathbf{U} \in M_m(\mathbb{C})$ and $\mathbf{V} \in M_n(\mathbb{C})$ are unitary.

(c) Deduce that if $\mathbf{A} = \mathbf{UTU}^*$ is a Schur factorization (Corollary 12.1.3), then $\|\mathbf{A}\|_F = \|\mathbf{T}\|_F$.

Solution.

(a) The (j,j) entry of $\mathbf{A}^*\mathbf{A}$ is $\sum_i \overline{a_{ij}}a_{ij} = \sum_i |a_{ij}|^2$. Summing over j gives $\operatorname{tr}(\mathbf{A}^*\mathbf{A}) = \sum_{i,j}|a_{ij}|^2 = \|\mathbf{A}\|_F^2$.

(b) By (a) and $\mathbf{U}^*\mathbf{U} = \mathbf{I}_m$,

$$\|\mathbf{UA}\|_F^2 = \operatorname{tr}((\mathbf{UA})^*(\mathbf{UA})) = \operatorname{tr}(\mathbf{A}^*\mathbf{U}^*\mathbf{UA}) = \operatorname{tr}(\mathbf{A}^*\mathbf{A}) = \|\mathbf{A}\|_F^2.$$

For the other side, use $\operatorname{tr}(\mathbf{BC}) = \operatorname{tr}(\mathbf{CB})$ (Theorem 0.9.6 (3)) and $\mathbf{VV}^* = \mathbf{I}_n$:

$$\|\mathbf{AV}\|_F^2 = \operatorname{tr}(\mathbf{V}^*\mathbf{A}^*\mathbf{AV}) = \operatorname{tr}(\mathbf{A}^*\mathbf{AVV}^*) = \operatorname{tr}(\mathbf{A}^*\mathbf{A}) = \|\mathbf{A}\|_F^2.$$

Norms are nonnegative, so the squares being equal gives the norms equal.

(c) Apply (b) twice: $\|\mathbf{A}\|_F = \|\mathbf{UTU}^*\|_F = \|\mathbf{TU}^*\|_F = \|\mathbf{T}\|_F$, using that $\mathbf{U}^*$ is unitary as well. So a Schur factorization redistributes the size of $\mathbf{A}$ between the diagonal and the strictly upper triangle without changing the total. Section 11 turns that observation into a measurement.

Exercise. C3: Distinct eigenvalues are dense, but not over the reals

(a) Sharpen Corollary 12.1.6: prove that for every $\mathbf{A} \in M_n(\mathbb{C})$ and every $\varepsilon > 0$ there is $\mathbf{A}'$ with n **distinct** eigenvalues and $|a_{ij} - a'_{ij}| < \varepsilon$ for all i, j.

(b) Show that the analogous statement fails over $\mathbb{R}$: for $\mathbf{R} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ there is an $\varepsilon > 0$ such that **no** real matrix within ε of $\mathbf{R}$, entry by entry, has a real eigenvalue at all.

Hint for (b): a real 2×2 matrix with entries a, b, c, d has real eigenvalues exactly when $(a-d)^2 + 4bc \geq 0$.

Solution.

(a) This is exactly what the proof of Corollary 12.1.6 produced, before the conclusion was weakened. There the δ_k were chosen so that the diagonal entries of $\mathbf{T} + \mathbf{D}$ are pairwise distinct, and by Theorem 9.7.4 (b) those n distinct numbers are the eigenvalues of $\mathbf{T} + \mathbf{D}$, hence of the similar matrix $\mathbf{A}' = \mathbf{U}(\mathbf{T} + \mathbf{D})\mathbf{U}^*$. The entrywise estimate is unchanged.

(b) The characteristic polynomial of $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ is $x^2-(a+d)x+(ad-bc)$, whose discriminant is

$$(a+d)^2 - 4(ad-bc) = (a-d)^2 + 4bc,$$

so the eigenvalues are real exactly when this is ≥ 0. Take $\varepsilon = \frac{1}{4}$ and let $\mathbf{A}'$ have entries within ε of those of $\mathbf{R}$, so $|a|, |d| \leq \frac{1}{4}$, $b \in [-\frac{5}{4}, -\frac{3}{4}]$, and $c \in [\frac{3}{4}, \frac{5}{4}]$. Then $(a-d)^2 \leq \frac{1}{4}$ and $bc \leq -\frac{9}{16}$, so

$$(a-d)^2 + 4bc \leq \tfrac{1}{4} - \tfrac{9}{4} = -2 < 0.$$

Hence $\mathbf{A}'$ has no real eigenvalue, let alone two distinct ones. The reason is the one this section keeps meeting: over $\mathbb{C}$ the eigenvalues move continuously and can always be separated, while over $\mathbb{R}$ a conjugate pair cannot be pushed onto the real line by a small change.

12.2 Self-Adjoint Operators

Among complex numbers, the real ones are those fixed by conjugation: $\overline{z} = z$. Among operators on an inner product space, conjugation is replaced by the adjoint, and the operators fixed by it are the **self-adjoint** ones, $T^* = T$. The analogy is not decoration. Self-adjoint operators have real eigenvalues, they come with a real-valued quantity $\langle Tv, v\rangle$ attached to every vector, they can be compared and ordered, and over $\mathbb{R}$ they are exactly the operators the spectral theorem applies to. This section collects everything about them that does not need a spectral theorem, and in doing so assembles every ingredient the real spectral theorem of Section 5 will use.

Throughout, V is a finite-dimensional inner product space over $F = \mathbb{R}$ or $F = \mathbb{C}$, and $T \in \mathcal{L}(V)$.

12.2.1 The operators fixed by starring

Chapter 11 already made the definition, so we only recall it: $T \in \mathcal{L}(V)$ is **self-adjoint** when $T^* = T$, equivalently when

$$\langle Tu, v\rangle = \langle u, Tv\rangle \qquad \text{for all } u, v \in V$$

(Definition 11.6.2). In an orthonormal basis this says the matrix is **Hermitian**, $\mathbf{A}^* = \mathbf{A}$, which over $\mathbb{R}$ is **symmetric**, $\mathbf{A}^\top = \mathbf{A}$ (Theorem 11.6.3). Three facts were proved there and will be used without further comment: $\langle Tv, v\rangle$ is real for every v, even over $\mathbb{C}$; the diagonal entries of the matrix in an orthonormal basis are real; and every self-adjoint operator is normal (Proposition 11.6.7).

Deciding self-adjointness is always the same job — compute T^*, or check the displayed identity on a basis — but the traps differ by example.

Example. Four Tests

Determine which of the following are self-adjoint. Justify your answer; for those that are not, name the failing pair of vectors.

(a) $T_{\mathbf{A}}$ on $\mathbb{C}^2$ with the standard inner product, $\mathbf{A} = \begin{pmatrix} 2 & 1-i \\ 1+i & 3 \end{pmatrix}$.

(b) $T_{\mathbf{B}}$ on $\mathbb{C}^2$ with the standard inner product, $\mathbf{B} = \begin{pmatrix} 1 & i \\ i & 1 \end{pmatrix}$.

(c) Differentiation D on $V = \mathbb{R}[x]_{\le 2}$ with $\langle p, q\rangle = \int_0^1 pq$.

(d) The orthogonal projection P_U onto a subspace U of any V.

Solution.

(a) Self-adjoint. Transposing swaps the off-diagonal entries and conjugating swaps them back, while the diagonal entries 2 and 3 are real; so $\mathbf{A}^* = \mathbf{A}$. The standard basis of $\mathbb{C}^2$ is orthonormal, so Theorem 11.6.3 turns this into $T_{\mathbf{A}}^* = T_{\mathbf{A}}$.

(b) Not self-adjoint. It is **symmetric**, $\mathbf{B}^\top = \mathbf{B}$, but $\mathbf{B}^* = \begin{pmatrix} 1 & -i \\ -i & 1 \end{pmatrix} \neq \mathbf{B}$. The failing pair is $\mathbf{e}_1, \mathbf{e}_2$: $\langle \mathbf{B}\mathbf{e}_1, \mathbf{e}_2\rangle = i$, while $\langle \mathbf{e}_1, \mathbf{B}\mathbf{e}_2\rangle = \overline{i} = -i$.

(c) Not self-adjoint. Take $p = x$ and $q = 1$. Then $Dp = 1$ and $Dq = 0$, so

$$\langle Dp, q\rangle = \int_0^1 1 \,\mathrm{d}x = 1, \qquad \langle p, Dq\rangle = 0.$$

The obstruction is visible in general: integration by parts gives

$$\langle Dp, q\rangle + \langle p, Dq\rangle = p(1)q(1) - p(0)q(0),$$

so D is as far from self-adjoint as it could be — the two sides add up rather than agreeing, apart from boundary terms.

(d) Self-adjoint, for every U. By Theorem 11.3.3 (b), $\langle P_U u, v\rangle = \langle u, P_U v\rangle$ for all u, v, which is the displayed identity.

Warning. **Over $\mathbb{C}$, symmetric is not the right notion; Hermitian is.** Part (b) above is symmetric and its eigenvalues are not even real: $\operatorname{tr}\mathbf{B} = 2$ and $\det\mathbf{B} = 1 - i^2 = 2$, so $p_{\mathbf{B}} = x^2 - 2x + 2$ and the eigenvalues are $1+i$ and $1-i$. Every theorem in this section would fail for it. The word "symmetric matrix" belongs to the **real** case, where conjugation does nothing and $\mathbf{A}^* = \mathbf{A}^\top$; over $\mathbb{C}$ always write, and check, $\mathbf{A}^* = \mathbf{A}$.

12.2.2 Eigenvalues are real

The first theorem is the operator version of "a number equal to its own conjugate is real", and its proof is the move that runs this whole chapter: compute one inner product two different ways.

Theorem 12.2.1. Self-adjoint Operators Have Real Eigenvalues

Let V be a finite-dimensional inner product space over $F = \mathbb{R}$ or $F = \mathbb{C}$ and let $T \in \mathcal{L}(V)$ be self-adjoint. Then every eigenvalue of T is real.

Idea. Take $Tv = \lambda v$ with $v \neq 0$ and pair it with v. Computing $\langle Tv, v\rangle$ straight off the eigenvalue equation produces λ; computing it by pushing T into the second slot produces $\overline{\lambda}$, because a scalar leaving the second slot picks up a conjugate. The two answers must agree, and $\|v\|^2 \neq 0$ lets us cancel.

Proof. Let λ be an eigenvalue of T, with eigenvector $v \neq 0$. On one side,

$$\langle Tv, v\rangle = \langle \lambda v, v\rangle = \lambda \|v\|^2,$$

by linearity in the first slot. On the other, using Definition 11.6.2 and then conjugate-linearity in the second slot,

$$\langle Tv, v\rangle = \langle v, Tv\rangle = \langle v, \lambda v\rangle = \overline{\lambda}\|v\|^2.$$

So $\lambda\|v\|^2 = \overline{\lambda}\|v\|^2$. Since $v \neq 0$ we have $\|v\|^2 \neq 0$ by positive definiteness (Definition 11.1.1), and dividing gives $\lambda = \overline{\lambda}$. A scalar equal to its own conjugate is real.

■

Chapter 11 already proved this, in the remark after Proposition 11.6.7. It is promoted to a theorem here because Sections 3 to 5 lean on it constantly, and a result cited that often should have a number. Over $\mathbb{R}$ the conclusion sounds empty, because the eigenvalues of a real operator are real by definition. What is not empty over $\mathbb{R}$ is the corresponding statement about the **roots of the characteristic polynomial**, and that is the next theorem: it says the roots do not escape into $\mathbb{C}$, which is exactly what guarantees that a real self-adjoint operator has an eigenvector at all.

12.2.3 Existence over the real numbers

Over $\mathbb{C}$ an eigenvalue is free (Theorem 9.2.4). Over $\mathbb{R}$ it is not, and the rotation of $\mathbb{R}^2$ by a quarter turn has none. Self-adjointness is what rules that out.

Theorem 12.2.2. Real Self-adjoint Operators Have an Eigenvalue

Let $V \neq \{0\}$ be a finite-dimensional inner product space over $\mathbb{R}$ and let $T \in \mathcal{L}(V)$ be self-adjoint. Then:

(a) every root of p_T in $\mathbb{C}$ is real, so p_T splits over $\mathbb{R}$;

(b) T has an eigenvalue.

Idea. There is no complex number in sight, so we put one there. Fix an orthonormal basis; the matrix $\mathbf{A}$ of T is real symmetric. The same array of numbers, read as a complex matrix, is Hermitian, and it acts on $\mathbb{C}^n$, which is the complexification of $\mathbb{R}^n$ (Definition 1.9.2, Example (The Complexification of $\mathbb{R}^n$ Is $\mathbb{C}^n$)) — the arena Chapter 1 built, and Chapter 9 put to work, precisely so that a real matrix could borrow complex eigenvalues. Over $\mathbb{C}$ an eigenvalue exists, and Theorem 12.2.1 says it is real. A real root of p_T is an eigenvalue of T itself, and we are back where we started, one eigenvalue richer.

Proof. Let $n = \dim V \geq 1$. By Theorem 11.2.2, V has an orthonormal basis $\mathcal{B}$; put $\mathbf{A} = [T]_{\mathcal{B}}^{\mathcal{B}} \in M_n(\mathbb{R})$. By Theorem 11.6.3 and Definition 11.6.2, $\mathbf{A}^* = \mathbf{A}$, and since $\mathbf{A}$ is real this says $\mathbf{A}^{\top} = \mathbf{A}$.

Now read $\mathbf{A}$ as an element of $M_n(\mathbb{C})$, acting on $\mathbb{C}^n$ with the standard inner product; this is the extension of the real map $T_{\mathbf{A}}$ to the complexification $(\mathbb{R}^n)_{\mathbb{C}} = \mathbb{C}^n$, which acts on $\mathbf{x}+i\mathbf{y}$ by

$\mathbf{Ax} + i\mathbf{Ay}$ (Chapter 9, Section 2). Because $\mathbf{A}$ is real and symmetric, $\mathbf{A}^* = \overline{\mathbf{A}}^{\top} = \mathbf{A}^{\top} = \mathbf{A}$, so $T_{\mathbf{A}}$ is self-adjoint as an operator on $\mathbb{C}^n$ (Theorem 11.6.3 again, the standard basis of $\mathbb{C}^n$ being orthonormal).

(a) By Definition 7.10.2, $p_T = p_{\mathbf{A}}$, and the entries of $x\mathbf{I} - \mathbf{A}$ are real, so this is the same polynomial whether the determinant is expanded over $\mathbb{R}$ or over $\mathbb{C}$; it lies in $\mathbb{R}[x] \subseteq \mathbb{C}[x]$. Let $\lambda \in \mathbb{C}$ be a root of $p_{\mathbf{A}}$. By Theorem 9.2.1, λ is an eigenvalue of $\mathbf{A}$ over $\mathbb{C}$, and $\mathbf{A}$ is self-adjoint there, so $\lambda \in \mathbb{R}$ by Theorem 12.2.1. Since every root of p_T in $\mathbb{C}$ lies in $\mathbb{R}$, and p_T splits over $\mathbb{C}$ by Corollary 6.5.5, it splits over $\mathbb{R}$.

(b) p_T is monic of degree $n \geq 1$ (Theorem 7.10.2), so by (a) it has a root $\lambda \in \mathbb{R}$. By Theorem 9.2.1 ((e) $\Rightarrow$ (a)), λ is an eigenvalue of T. This proves the theorem.

■

Remark. There is a second, very different route to (b): the function $v \mapsto \langle Tv, v\rangle$ is real-valued and continuous on the unit sphere of V, which is closed and bounded, so it attains a maximum, and a vector where it does so turns out to be an eigenvector for the largest eigenvalue. That argument is shorter to state but it is analysis, not algebra, and it appeals to the extreme value theorem directly. We keep to the algebraic route, which reaches that theorem only at one remove: the proof just given rests on Corollary 6.5.5, hence on the fundamental theorem of algebra (Theorem 6.5.1), and the proof of that in Chapter 6 uses the extreme value theorem on a closed disc. The maximization picture returns in Chapter 17, where the Rayleigh quotient and the Courant–Fischer theorem describe every eigenvalue of a self-adjoint operator by optimization.

Quick check. The rotation $\mathbf{R} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ on $\mathbb{R}^2$ has no eigenvalue, and over $\mathbb{C}$ its eigenvalues $\pm i$ are not real. Which hypothesis of Theorem 12.2.2 does it fail, and how do you see it in one line?

Solution. Self-adjointness. In the standard orthonormal basis the matrix of $\mathbf{R}^*$ is $\mathbf{R}^{\top} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} = -\mathbf{R}$, which is not $\mathbf{R}$. So $\mathbf{R}$ is skew-adjoint, not self-adjoint, and neither theorem applies. (It is, however, normal, since $\mathbf{R}^{\top}\mathbf{R} = \mathbf{R}\mathbf{R}^{\top} = \mathbf{I}$ — which is why Section 5 has to insist on self-adjoint rather than normal over $\mathbb{R}$.)

12.2.4 Eigenvectors for different eigenvalues are orthogonal

The second theorem is the same move with two vectors instead of one, and it is the geometric half of the spectral theorem: the eigenspaces are not merely independent, as in Chapter 9, but perpendicular.

Theorem 12.2.3. Distinct Eigenvalues Give Orthogonal Eigenvectors

Let $T \in \mathcal{L}(V)$ be self-adjoint, and let $u, v \in V$ satisfy $Tu = \lambda u$ and $Tv = \mu v$ with $\lambda \neq \mu$. Then $\langle u, v\rangle = 0$. Consequently $E_\lambda(T) \perp E_\mu(T)$ for distinct eigenvalues λ, μ.

Proof. Compute $\langle Tu, v\rangle$ two ways. First, $\langle Tu, v\rangle = \langle \lambda u, v\rangle = \lambda\langle u, v\rangle$. Second, by Definition 11.6.2 and conjugate-linearity in the second slot,

$$\langle Tu, v\rangle = \langle u, Tv\rangle = \langle u, \mu v\rangle = \overline{\mu}\langle u, v\rangle = \mu\langle u, v\rangle,$$

the last step because μ is real by Theorem 12.2.1. Subtracting, $(\lambda - \mu)\langle u, v\rangle = 0$, and $\lambda - \mu \neq 0$, so $\langle u, v\rangle = 0$. Since every element of $E_\lambda(T)$ is such a u and every element of $E_\mu(T)$ such a v, the two eigenspaces are orthogonal.

■

Note where the previous theorem was spent: it made $\overline{\mu} = \mu$, so that the conjugation which appears on the way out of the second slot does no damage. Without real eigenvalues the computation would end at $\lambda\langle u, v\rangle = \overline{\mu}\langle u, v\rangle$, which says nothing when $\lambda = \overline{\mu}$.

Example. A Hermitian Two-by-Two

For $\mathbf{A} = \begin{pmatrix} 2 & 1-i \\ 1+i & 3 \end{pmatrix}$ on $\mathbb{C}^2$, find the eigenvalues and an eigenvector for each, and verify that the two eigenvectors are orthogonal.

Solution. Here $\operatorname{tr}\mathbf{A} = 5$ and $\det\mathbf{A} = 6 - (1-i)(1+i) = 6 - 2 = 4$, so

$$p_{\mathbf{A}} = x^2 - 5x + 4 = (x-1)(x-4).$$

Both eigenvalues are real, as Theorem 12.2.1 promised.

For $\lambda = 1$: $\mathbf{A} - \mathbf{I} = \begin{pmatrix} 1 & 1-i \\ 1+i & 2 \end{pmatrix}$, and the first row gives $x + (1-i)y = 0$, so $\mathbf{u} = (-1+i, 1)$. Check the second row: $(1+i)(-1+i) + 2 = (-1 + i - i + i^2) + 2 = -2 + 2 = 0$.

For $\lambda = 4$: $\mathbf{A} - 4\mathbf{I} = \begin{pmatrix} -2 & 1-i \\ 1+i & -1 \end{pmatrix}$, and the first row gives $-2x + (1-i)y = 0$, so $\mathbf{v} = (1-i, 2)$.

Finally

$$\langle \mathbf{u}, \mathbf{v}\rangle = (-1+i)\overline{(1-i)} + 1 \cdot \overline{2} = (-1+i)(1+i) + 2 = -2 + 2 = 0,$$

so $\mathbf{u} \perp \mathbf{v}$, as Theorem 12.2.3 requires. Normalizing, $\|\mathbf{u}\|^2 = 2 + 1 = 3$ and $\|\mathbf{v}\|^2 = 2 + 4 = 6$, so

$$\left(\tfrac{1}{\sqrt{3}}(-1+i, 1),\ \tfrac{1}{\sqrt{6}}(1-i, 2)\right)$$

is an orthonormal basis of $\mathbb{C}^2$ consisting of eigenvectors of $\mathbf{A}$. Section 4 proves that this always happens, for exactly the operators it happens for.

12.2.5 Invariant complements

Section 5 will prove the real spectral theorem by the move named at the front of the chapter: find one eigenvector, pass to its orthogonal complement, induct. For that to work the complement must be invariant, and the restriction must again be self-adjoint so that the induction can be applied to it. Both hold, and neither needs an eigenvector — the statement is about any invariant subspace.

Theorem 12.2.4. Self-adjointness Passes to Orthogonal Complements

Let $T \in \mathcal{L}(V)$ be self-adjoint and let $U \subseteq V$ be a T-invariant subspace. Then:

(a) $U^{\perp}$ is T-invariant;

(b) $T|_U \in \mathcal{L}(U)$ and $T|_{U^{\perp}} \in \mathcal{L}(U^{\perp})$ are self-adjoint, for the inner products inherited from V.

Proof.

(a) Since T is self-adjoint, $T^* = T$, so U is T^*-invariant. By Lemma 12.1.1, $U^{\perp}$ is T-invariant.

(b) U is T-invariant, so $T|_U$ maps U to U and is an operator on U. For $u_1, u_2 \in U$,

$$\langle T|_U u_1, u_2\rangle = \langle Tu_1, u_2\rangle = \langle u_1, Tu_2\rangle = \langle u_1, T|_U u_2\rangle,$$

where the inner products are those of V, which is what U inherits. By Definition 11.6.2, $T|_U$ is self-adjoint. The same three equalities with $U^{\perp}$ in place of U give the second claim, using (a) to know that $T|_{U^{\perp}}$ is an operator on $U^{\perp}$.

■

So the pair $(U, U^{\perp})$ splits a self-adjoint operator into two self-adjoint operators on smaller spaces, and nothing is lost: $V = U \oplus U^{\perp}$ by Theorem 11.3.2 (a). Combined with Theorem 12.2.2, which guarantees that a nonzero space always offers an eigenvector to start from, this is a complete induction, and Section 5 will do no more than write it out.

12.2.6 Two zero tests, and the gap between the fields

An operator carries the function $v \mapsto \langle Tv, v\rangle$, sometimes called its quadratic form. How much of T does that one function remember? The answer depends on the field, and the difference is the sharpest illustration in this book of why $\mathbb{C}$ is the easier place to work.

Theorem 12.2.5. The Zero Test for Self-adjoint Operators

Let V be a finite-dimensional inner product space over $F = \mathbb{R}$ or $F = \mathbb{C}$ and let $T \in \mathcal{L}(V)$ be **self-adjoint**. If $\langle Tv, v\rangle = 0$ for every $v \in V$, then $T = 0$.

Idea. The hypothesis only ever feeds T one vector, while the conclusion is about $\langle Tu, w\rangle$ for **two**. So manufacture two out of one: apply the hypothesis to $u + w$ and expand. The

diagonal terms die by hypothesis and what is left is a statement about the cross terms, which self-adjointness turns into a statement about $\langle Tu, w\rangle$ alone.

Proof. Let $u, w \in V$. Expanding $\langle T(u+w), u+w\rangle$ in both slots,

$$\begin{aligned} 0 &= \langle Tu, u\rangle + \langle Tu, w\rangle + \langle Tw, u\rangle + \langle Tw, w\rangle \\ &= \langle Tu, w\rangle + \langle Tw, u\rangle, \end{aligned}$$

the two outer terms vanishing by hypothesis. Since T is self-adjoint, $\langle Tw, u\rangle = \langle w, Tu\rangle = \overline{\langle Tu, w\rangle}$ by conjugate symmetry. Writing $z = \langle Tu, w\rangle$, the display says $z + \overline{z} = 0$, that is

$$\operatorname{Re}\langle Tu, w\rangle = 0 \qquad \text{for all } u, w \in V. \tag{$*$}$$

If $F = \mathbb{R}$ this already says $\langle Tu, w\rangle = 0$ for all u, w. If $F = \mathbb{C}$, apply $(*)$ with iw in place of w: since $\langle Tu, iw\rangle = \bar{i}\, z = -iz$ and $\operatorname{Re}(-iz) = \operatorname{Im}(z)$, we also get $\operatorname{Im}\langle Tu, w\rangle = 0$. Either way $\langle Tu, w\rangle = 0$ for all u, w.

Taking $w = Tu$ gives $\|Tu\|^2 = 0$, so $Tu = 0$ for every u, that is $T = 0$. This proves the theorem.

■

Now drop the hypothesis. Over $\mathbb{C}$ nothing is lost.

Theorem 12.2.6. The Zero Test over the Complex Numbers

Let V be a finite-dimensional **complex** inner product space and let $T \in \mathcal{L}(V)$ be any operator. If $\langle Tv, v\rangle = 0$ for every $v \in V$, then $T = 0$.

Idea. The previous proof used self-adjointness in exactly one place: to convert $\langle Tw, u\rangle$ into $\overline{\langle Tu, w\rangle}$, which gave the real part only. Without it we must extract the imaginary part some other way, and over $\mathbb{C}$ there is another way — substitute $u + iw$ as well as $u + w$, and take a combination of the two identities. There is no third substitution available over $\mathbb{R}$, which is precisely why the real statement is false.

Proof. Let $u, w \in V$. Applying the hypothesis to $u + w$ and expanding as before,

$$\langle Tu, w\rangle + \langle Tw, u\rangle = 0. \tag{1}$$

Applying it to $u + iw$ and using $\langle Tu, iw\rangle = -i\langle Tu, w\rangle$, $\langle T(iw), u\rangle = i\langle Tw, u\rangle$, and $\langle T(iw), iw\rangle = i\bar{i}\langle Tw, w\rangle = 0$,

$$-i\langle Tu, w\rangle + i\langle Tw, u\rangle = 0.$$

Multiplying that by i gives

$$\langle Tu, w\rangle - \langle Tw, u\rangle = 0. \tag{2}$$

Adding (1) and (2), $2\langle Tu, w\rangle = 0$, so $\langle Tu, w\rangle = 0$ for all u, w. Taking $w = Tu$ gives $\|Tu\|^2 = 0$, hence $T = 0$.

■

This statement appeared once already, as Exercise (C2: Linearity is a hypothesis, not a consequence) (b), where it was an exercise in the arithmetic of complex inner products. It is restated and proved here because Section 3 needs to cite it as a theorem: the four characterizations of a normal operator all come from applying it to $T^*T - TT^*$. The proof is the same one; what is new is the company it now keeps.

Warning. **Over $\mathbb{R}$, the quadratic form does not determine the operator.** Let T be the quarter turn of $\mathbb{R}^2$, $T(x_1, x_2) = (-x_2, x_1)$. Then

$$\langle T\mathbf{x}, \mathbf{x}\rangle = -x_2x_1 + x_1x_2 = 0 \qquad \text{for every } \mathbf{x} \in \mathbb{R}^2,$$

and yet $T \neq 0$. So Theorem 12.2.6 genuinely fails over $\mathbb{R}$, and Theorem 12.2.5 genuinely needs its hypothesis: this T is not self-adjoint, since $T^* = -T$. The general real statement is that $\langle Tv, v\rangle = 0$ for all v forces $T^* = -T$ and no more — Exercise (C3: What the real quadratic form does determine).

Quick check. Over $\mathbb{C}$, suppose $\langle Tv, v\rangle$ is **real** for every $v \in V$. Must T be self-adjoint?

Solution. Yes. Put $S = T - T^*$. For every v, $\langle T^*v, v\rangle = \langle v, Tv\rangle = \overline{\langle Tv, v\rangle} = \langle Tv, v\rangle$, the last step because the number is real. Hence $\langle Sv, v\rangle = 0$ for all v, and Theorem 12.2.6 gives $S = 0$, that is $T^* = T$. Combined with Proposition 11.6.7 (a), this says that over $\mathbb{C}$ "self-adjoint" and "$\langle Tv, v\rangle$ always real" are the same condition. Over $\mathbb{R}$ the criterion is useless, since $\langle Tv, v\rangle$ is automatically real for every operator.

12.2.7 Exercises

12.2.7.1 A. Check your understanding

Exercise. A1

(a) State what it means for $T \in \mathcal{L}(V)$ to be self-adjoint, in terms of T^* and in terms of inner products, and say what the matrix condition is and in which bases it may be used.

(b) True or false: a symmetric matrix in $M_2(\mathbb{C})$ has real eigenvalues. Justify your answer.

(c) Which field does Theorem 12.2.2 concern, and which earlier theorem supplies the eigenvalue it borrows?

(d) True or false: if T is self-adjoint and U is any subspace, then $U^\perp$ is T-invariant. Justify your answer.

(e) State the difference between Theorem 12.2.5 and Theorem 12.2.6, and give the standard example showing the second has no real analogue.

(f) Give a normal operator on $\mathbb{R}^2$ that is not self-adjoint.

Solution.

(a) $T^* = T$; equivalently $\langle Tu, v\rangle = \langle u, Tv\rangle$ for **all** $u, v \in V$ (Definition 11.6.2). The matrix condition is $\mathbf{A}^* = \mathbf{A}$, and by Theorem 11.6.3 it may be used only in an **orthonormal** basis.

(b) False. $\mathbf{B} = \begin{pmatrix} 1 & i \\ i & 1 \end{pmatrix}$ is symmetric with eigenvalues $1 \pm i$ (Example (Four Tests) (b) and the warning after it). Over $\mathbb{C}$ the relevant condition is $\mathbf{A}^* = \mathbf{A}$, not $\mathbf{A}^\top = \mathbf{A}$.

(c) The real field: it says a self-adjoint operator on a nonzero **real** inner product space has an eigenvalue. The eigenvalue is borrowed from Theorem 9.2.4, applied to the matrix read as a complex matrix on the complexification $\mathbb{C}^n$ of $\mathbb{R}^n$, and is then shown to be real by Theorem 12.2.1.

(d) False as stated: U must be T**-invariant** (Theorem 12.2.4). For example, $T = T_{\mathbf{A}}$ on $\mathbb{R}^2$ with $\mathbf{A} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ is self-adjoint, and $U = \operatorname{span}(\mathbf{e}_1)$ has $U^\perp = \operatorname{span}(\mathbf{e}_2)$ with $T\mathbf{e}_2 = \mathbf{e}_1 \notin U^\perp$. Here U was not invariant either, since $T\mathbf{e}_1 = \mathbf{e}_2$.

(e) Theorem 12.2.5 holds over $\mathbb{R}$ and $\mathbb{C}$ but assumes T self-adjoint; Theorem 12.2.6 drops that assumption but holds only over $\mathbb{C}$. The quarter turn $T(x_1, x_2) = (-x_2, x_1)$ of $\mathbb{R}^2$ satisfies $\langle T\mathbf{x}, \mathbf{x}\rangle = 0$ for every $\mathbf{x}$ and is not zero.

(f) The same quarter turn: $T^* = -T$, so $T^*T = TT^* = \mathrm{id}$ and T is normal, while $T^* \neq T$.

12.2.7.2 B. Practice

Exercise. B1: Determine which are self-adjoint

Determine which of the following operators are self-adjoint. Justify your answer; for those that are not, exhibit a pair u, v with $\langle Tu, v\rangle \neq \langle u, Tv\rangle$, or compute T^*.

(a) $T_{\mathbf{A}}$ on $\mathbb{C}^2$, $\mathbf{A} = \begin{pmatrix} 1 & 2+i \\ 2-i & -5 \end{pmatrix}$, standard inner product.

(b) $T_{\mathbf{B}}$ on $\mathbb{C}^2$, $\mathbf{B} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$, standard inner product.

(c) $T_{\mathbf{C}}$ on $\mathbb{R}^3$, $\mathbf{C} = \begin{pmatrix} 1 & 2 & 0 \\ 2 & 1 & 3 \\ 0 & 3 & 1 \end{pmatrix}$, standard inner product.

(d) S^*S, for an arbitrary $S \in \mathcal{L}(V, W)$ between finite-dimensional inner product spaces.

(e) $P_U - P_{U^\perp}$, for a subspace U of a finite-dimensional V.

(f) $T(x_1, x_2) = (x_2, x_1)$ on $\mathbb{R}^2$ with the standard inner product, whose matrix in the basis $\mathcal{C} = ((1,0),(1,2))$ is $\begin{pmatrix} -\frac{1}{2} & \frac{3}{2} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix}$, which is not symmetric.

Solution.

(a) Self-adjoint. The diagonal entries 1 and -5 are real and the off-diagonal pair $2+i$,

$2-i$ are conjugates, so $\mathbf{A}^* = \mathbf{A}$; the standard basis is orthonormal, so Theorem 11.6.3 applies.

(b) Not self-adjoint. $\mathbf{B}^* = \mathbf{B}^\top = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} = -\mathbf{B}$, so $T_{\mathbf{B}}$ is skew-adjoint. Concretely $\langle \mathbf{B}\mathbf{e}_1, \mathbf{e}_2 \rangle = -1$ while $\langle \mathbf{e}_1, \mathbf{B}\mathbf{e}_2 \rangle = 1$.

(c) Self-adjoint. $\mathbf{C}$ is real and $\mathbf{C}^\top = \mathbf{C}$, and the standard basis of $\mathbb{R}^3$ is orthonormal.

(d) Self-adjoint, for every S. By Theorem 11.6.2 (c) and (d), $(S^*S)^* = S^*S^{**} = S^*S$.

(e) Self-adjoint. Each of P_U and $P_{U^\perp}$ is self-adjoint (Example (Four Tests) (d)), and by Theorem 11.6.2 (a) and (b) the difference of self-adjoint operators is self-adjoint: $(P_U - P_{U^\perp})^* = P_U^* - P_{U^\perp}^* = P_U - P_{U^\perp}$.

(f) Self-adjoint, in spite of the nonsymmetric matrix. The basis $\mathcal{C}$ is **not** orthonormal — $\langle (1,0), (1,2) \rangle = 1 \neq 0$ — so Theorem 11.6.3 does not apply to it, and the shape of $[T]_{\mathcal{C}}^{\mathcal{C}}$ says nothing either way. Test the definition instead: for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^2$,

$$\langle T\mathbf{x}, \mathbf{y} \rangle = x_2 y_1 + x_1 y_2 = \langle \mathbf{x}, T\mathbf{y} \rangle,$$

so $T^* = T$. In the standard basis, which *is* orthonormal, the matrix is $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$, symmetric as it must be. The moral is the one Chapter 11, Section 6 closes with: self-adjointness is a property of the operator and the inner product, never of a matrix taken on its own.

Exercise. B2: Eigenvalues and orthogonality

(a) For $\mathbf{A} = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix} \in M_2(\mathbb{C})$, check that $\mathbf{A}$ is Hermitian, find its eigenvalues and an eigenvector for each, and verify that the eigenvectors are orthogonal.

(b) For $\mathbf{S} = \begin{pmatrix} 3 & 1 & 1 \\ 1 & 3 & 1 \\ 1 & 1 & 3 \end{pmatrix} \in M_3(\mathbb{R})$, find the eigenvalues with their multiplicities and bases of the eigenspaces, and verify that the two eigenspaces are orthogonal.

(c) Hence write down an orthonormal basis of $\mathbb{R}^3$ consisting of eigenvectors of $\mathbf{S}$.

Solution.

(a) $\mathbf{A}^* = \overline{\mathbf{A}}^\top = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix} = \mathbf{A}$, so $\mathbf{A}$ is Hermitian. Then $\operatorname{tr}\mathbf{A} = 0$ and $\det\mathbf{A} = 0 - (-i)(i) = -1$, so $p_{\mathbf{A}} = x^2 - 1$ and the eigenvalues are 1 and -1, both real.
$\mathbf{A}(1,i) = (-i \cdot i,\ i \cdot 1) = (1,i)$, so $\mathbf{u} = (1,i)$ works for $\lambda = 1$. $\mathbf{A}(1,-i) = (-1,\ i) = -(1,-i)$, so $\mathbf{v} = (1,-i)$ works for $\lambda = -1$. Finally

$$\langle \mathbf{u}, \mathbf{v} \rangle = 1 \cdot \overline{1} + i\,\overline{(-i)} = 1 + i \cdot i = 0.$$

(b) Expanding the determinant gives $p_{\mathbf{S}} = (x-5)(x-2)^2$, so $\operatorname{spec}(\mathbf{S}) = \{5, 2\}$ with algebraic multiplicities 1 and 2. The eigenspaces are read off from $\mathbf{S} = 2\mathbf{I} + \mathbf{J}$, where $\mathbf{J}$ is the all-ones matrix: since $\mathbf{J}(1,1,1) = (3,3,3)$ and $\mathbf{J}\mathbf{x} = \mathbf{0}$ exactly when $x_1 + x_2 + x_3 = 0$,

$$E_5(\mathbf{S}) = \operatorname{span}\big((1,1,1)\big), \quad E_2(\mathbf{S}) = \operatorname{span}\big((1,-1,0), (1,0,-1)\big),$$

of dimensions 1 and 2. Orthogonality: $(1,1,1)\cdot(1,-1,0)=0$ and $(1,1,1)\cdot(1,0,-1)=0$, so $E_5(\mathbf{S}) \perp E_2(\mathbf{S})$, as Theorem 12.2.3 requires.

(c) The two vectors spanning $E_2(\mathbf{S})$ are not orthogonal to each other, so run Gram–Schmidt inside that eigenspace (Theorem 11.2.2). With $\mathbf{w}_1 = (1,-1,0)$,

$$(1,0,-1) - \tfrac{1}{2}(1,-1,0) = \tfrac{1}{2}(1,1,-2).$$

Normalizing all three vectors gives the orthonormal eigenbasis

$$\tfrac{1}{\sqrt{3}}(1,1,1), \quad \tfrac{1}{\sqrt{2}}(1,-1,0), \quad \tfrac{1}{\sqrt{6}}(1,1,-2).$$

The eigenvalues are $5, 2, 2$ in that order.

Exercise. B3: A self-adjoint square root of zero

Let $T \in \mathcal{L}(V)$ be self-adjoint with $T^2 = 0$. Prove that $T = 0$. Deduce that a self-adjoint T with $T^k = 0$ for some $k \geq 1$ is zero.

Solution. For every $v \in V$, by Definition 11.6.2,

$$\|Tv\|^2 = \langle Tv, Tv\rangle = \langle T^2 v, v\rangle = \langle 0, v\rangle = 0,$$

so $Tv = 0$. As v was arbitrary, $T = 0$.

For the deduction, use strong induction on k (Theorem 0.2.10). If $k = 1$ there is nothing to prove. Let $k \geq 2$ with $T^k = 0$, and put $m = \lceil k/2 \rceil$, so that $1 \leq m < k$ and $2m \geq k$. Then $(T^m)^2 = T^{2m} = T^{2m-k}T^k = 0$, and T^m is self-adjoint, since $(T^m)^* = (T^*)^m = T^m$ by Theorem 11.6.2 (c). By the first part, $T^m = 0$, and the inductive hypothesis applied to the exponent m gives $T = 0$.

12.2.7.3 C. Going deeper

Exercise. C1: The operator S^*S

Let $S \in \mathcal{L}(V, W)$ between finite-dimensional inner product spaces over F, and set $T = S^*S \in \mathcal{L}(V)$.

(a) Prove that T is self-adjoint and that $\langle Tv, v\rangle \geq 0$ for every $v \in V$, with equality if and only if $Sv = 0$.

(b) Deduce that every eigenvalue of T is a nonnegative real number.

(c) Prove that T is invertible if and only if S is injective.

Hint for (a): the quantity $\langle Tv, v\rangle$ is a norm squared in disguise.

Solution.

(a) $T^* = (S^*S)^* = S^*S^{**} = S^*S = T$, by Theorem 11.6.2 (c) and (d). For the second

claim, use conjugate symmetry and then Definition 11.6.1:

$$\langle Tv, v\rangle = \overline{\langle v, S^*Sv\rangle} = \overline{\langle Sv, Sv\rangle} = \|Sv\|^2,$$

the last step because a norm squared is a nonnegative real number and is its own conjugate. So $\langle Tv, v\rangle \geq 0$, with equality exactly when $\|Sv\| = 0$, that is $Sv = 0$ (Definition 11.1.1).

(b) Let $Tv = \lambda v$ with $v \neq 0$. By Theorem 12.2.1, $\lambda \in \mathbb{R}$. By (a), $0 \leq \langle Tv, v\rangle = \lambda\|v\|^2$, and $\|v\|^2 > 0$, so $\lambda \geq 0$.

(c) ($\Rightarrow$) If $Sv = 0$ then $Tv = 0$, so injectivity of T forces $v = 0$; and an invertible operator is injective. ($\Leftarrow$) If S is injective and $Tv = 0$, then by (a) $\|Sv\|^2 = \langle Tv, v\rangle = 0$, so $Sv = 0$ and $v = 0$. Thus $\ker T = \{0\}$, so T is injective, hence bijective by Corollary 2.3.2 (e). This is Corollary 11.6.6 (b) again, seen from the quadratic form.

Operators with the property in (a) are called **positive**, and they are the subject of Chapter 13; part (b) is the first half of the characterization proved there.

Exercise. C2: Sums and products

Let $S, T \in \mathcal{L}(V)$ be self-adjoint.

(a) Prove that $S + T$ and cT are self-adjoint for every **real** scalar c, and show by example that cT need not be self-adjoint for $c \in \mathbb{C}$.

(b) Prove that ST is self-adjoint if and only if $ST = TS$.

(c) Give self-adjoint $S, T \in \mathcal{L}(\mathbb{C}^2)$ with ST not self-adjoint.

Solution.

(a) By Theorem 11.6.2 (a), $(S+T)^* = S^* + T^* = S + T$. By (b) of the same theorem, $(cT)^* = \bar{c}T^* = \bar{c}T$, which equals cT exactly when $\bar{c} = c$ or $T = 0$. For a complex counterexample take $T = \mathrm{id}$ on $\mathbb{C}^2$ and $c = i$: then $(i\,\mathrm{id})^* = -i\,\mathrm{id} \neq i\,\mathrm{id}$.

(b) By Theorem 11.6.2 (c), $(ST)^* = T^*S^* = TS$. So $(ST)^* = ST$ if and only if $TS = ST$.

(c) Take $S = T_{\mathbf{A}}$ and $T = T_{\mathbf{B}}$ on $\mathbb{C}^2$ with

$$\mathbf{A} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \qquad \mathbf{B} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

Both are Hermitian. But $\mathbf{AB} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$ and $\mathbf{BA} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}$, which differ, so by (b) the product is not self-adjoint. Indeed $(\mathbf{AB})^* = \mathbf{BA} \neq \mathbf{AB}$.

Exercise. C3: What the real quadratic form does determine

Let V be a finite-dimensional inner product space over $\mathbb{R}$ and $T \in \mathcal{L}(V)$. Call T **skew-adjoint** if $T^* = -T$.

(a) Prove that $\langle Tv, v\rangle = 0$ for every $v \in V$ if and only if T is skew-adjoint.

(b) Deduce that every $T \in \mathcal{L}(V)$ can be written uniquely as $T = A + B$ with A self-adjoint and B skew-adjoint, and that $\langle Tv, v\rangle = \langle Av, v\rangle$ for every v.

(c) Conclude that over $\mathbb{R}$ the function $v \mapsto \langle Tv, v\rangle$ determines A and nothing about B. Contrast this with the complex case.

Solution.

(a) ($\Leftarrow$) If $T^* = -T$, then for every v, $\langle Tv, v\rangle = \langle v, T^*v\rangle = -\langle v, Tv\rangle = -\langle Tv, v\rangle$, the last step by symmetry of a real inner product. A real number equal to its own negative is 0.
($\Rightarrow$) Suppose $\langle Tv, v\rangle = 0$ for all v. Expanding $\langle T(u+w), u+w\rangle = 0$ and canceling the two diagonal terms,

$$\langle Tu, w\rangle + \langle Tw, u\rangle = 0.$$

Over $\mathbb{R}$, $\langle Tw, u\rangle = \langle u, Tw\rangle$, so $\langle Tu, w\rangle = \langle u, -Tw\rangle$ for all u, w. By the uniqueness in Theorem 11.6.1, $T^* = -T$.

(b) Put $A = \frac{1}{2}(T + T^*)$ and $B = \frac{1}{2}(T - T^*)$. Then $A + B = T$, and by Theorem 11.6.2 (a), (b), and (d), $A^* = A$ and $B^* = -B$. For uniqueness, if $T = A' + B'$ is another such splitting, then $T^* = A' - B'$, so $A' = \frac{1}{2}(T + T^*) = A$ and $B' = B$. Finally $\langle Bv, v\rangle = 0$ for every v by (a), so $\langle Tv, v\rangle = \langle Av, v\rangle + \langle Bv, v\rangle = \langle Av, v\rangle$.

(c) By (b), two operators T and T' with the same self-adjoint part have the same quadratic form; and conversely, if $\langle Tv, v\rangle = \langle T'v, v\rangle$ for all v, then $A - A'$ is self-adjoint with vanishing quadratic form, so $A = A'$ by Theorem 12.2.5. So the quadratic form sees exactly A. The skew part is invisible: the quarter turn of $\mathbb{R}^2$ is skew-adjoint, and adding it to any T changes nothing. Over $\mathbb{C}$, by contrast, Theorem 12.2.6 says the quadratic form determines T completely — there is no invisible part at all, which is why the complex theory is the easier one.

12.3 Normal Operators

Section 2 collected everything a self-adjoint operator gives us: real eigenvalues, orthogonal eigenspaces, an invariant orthogonal complement. That is exactly the equipment an orthonormal basis of eigenvectors needs. But over $\mathbb{C}$ the hypothesis is too heavy, and this section finds the lighter one that still carries the load. It is a condition already met in Chapter 11, and the rest of the chapter runs on it.

As throughout, F is $\mathbb{R}$ or $\mathbb{C}$, inner products are linear in the first slot, and every inner product space is finite-dimensional unless we say otherwise.

12.3.1 The hypothesis that self-adjointness overshoots

Take the operator on $\mathbb{C}^2$ whose matrix in the standard basis is $\mathbf{D} = \operatorname{diag}(1, i)$. The standard basis is orthonormal and consists of eigenvectors of $\mathbf{D}$, so this operator has everything the spectral theorem promises. Yet it is not self-adjoint: one of its eigenvalues is i, and a self-adjoint operator has only real eigenvalues (Theorem 12.2.1). So over $\mathbb{C}$ self-adjointness cannot be the condition we are looking for. It is sufficient and not necessary, and we are throwing away every operator with a nonreal eigenvalue — including every unitary operator other than the trivial ones.

What survives? Compare the two computations. For a self-adjoint T we have $T^*T = TT = TT^*$. For $\mathbf{D} = \operatorname{diag}(1, i)$ we have $\mathbf{D}^* = \operatorname{diag}(1, -i)$, which is a different matrix, but

$$\mathbf{D}^*\mathbf{D} = \operatorname{diag}(1, 1) = \mathbf{D}\mathbf{D}^*.$$

Neither operator equals its adjoint; both **commute** with their adjoints. That is the common feature, and Chapter 11 already gave it a name.

A normal operator is one that never gets in its adjoint's way: the two may be applied in either order.

Recall Definition 11.6.3: an operator $T \in \mathcal{L}(V)$ is **normal** when $T^*T = TT^*$. That is a single equation between two operators, with no vectors and no quantifiers in it, and it is strictly weaker than $T^* = T$. Section 6 of Chapter 11 checked the definition on self-adjoint operators, on unitary ones, on $i\mathbf{I}$, and on diagonal matrices, and found the 2×2 Jordan block $\mathbf{J}_2(0)$ failing it. We take those for granted and go one step further: normality is best understood as a statement about **lengths**, and that is the content of the first theorem below.

Before it, a wider catalog, since the class is larger than Chapter 11's four examples suggest.

Example. Which of these are normal?

Decide which of the following matrices are normal, with the standard inner product on $\mathbb{C}^n$ or $\mathbb{R}^n$. For those that are, say which familiar family they belong to.

(a) $\mathbf{A} = \begin{pmatrix} 2 & 1+i \\ 1-i & 3 \end{pmatrix}$ over $\mathbb{C}$.

(b) $\mathbf{B} = \begin{pmatrix} 0 & 2 \\ -2 & 0 \end{pmatrix}$ over $\mathbb{R}$.

(c) $\mathbf{U} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & i \\ i & 1 \end{pmatrix}$ over $\mathbb{C}$.

(d) $\mathbf{C} = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \\ 2 & 3 & 1 \end{pmatrix}$ over $\mathbb{C}$.

(e) $\mathbf{M} = \begin{pmatrix} 2 & 1 \\ -1 & 2 \end{pmatrix}$ over $\mathbb{R}$.

(f) $\mathbf{N} = \begin{pmatrix} 1 & 1 \\ 0 & 2 \end{pmatrix}$ over $\mathbb{R}$.

Solution.

(a) Normal. $\mathbf{A}^* = \mathbf{A}$, since transposing swaps the off-diagonal entries and conjugating swaps them back, so $\mathbf{A}$ is Hermitian and both products equal $\mathbf{A}^2$. Every **self-adjoint** operator is normal.

(b) Normal. $\mathbf{B}^\top = -\mathbf{B}$, so $\mathbf{B}^\top\mathbf{B} = -\mathbf{B}^2 = \mathbf{B}\mathbf{B}^\top$. An operator with $T^* = -T$ is called **skew-adjoint**, and the same one-line computation makes every such operator normal.

(c) Normal. $\mathbf{U}^*\mathbf{U} = \frac{1}{2}\begin{pmatrix} 1 & -i \\ -i & 1 \end{pmatrix}\begin{pmatrix} 1 & i \\ i & 1 \end{pmatrix} = \frac{1}{2}\begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} = \mathbf{I}$, so $\mathbf{U}$ is **unitary** (Definition 11.7.2), and then $\mathbf{U}^*\mathbf{U} = \mathbf{I} = \mathbf{U}\mathbf{U}^*$.

(d) Normal. Each row of $\mathbf{C}$ is the previous one shifted cyclically to the right; such a matrix is called a **circulant**. Writing out the products,

$$\mathbf{C}^{\top}\mathbf{C} = \mathbf{C}\mathbf{C}^{\top} = \begin{pmatrix} 14 & 11 & 11 \\ 11 & 14 & 11 \\ 11 & 11 & 14 \end{pmatrix}.$$

Both entries in position (i, j) are the dot product of two rows of $\mathbf{C}$, or of two columns, and cyclic shifting makes those two lists of dot products agree. Section 9 explains circulants properly and diagonalizes all of them at once.

(e) Normal, and in none of the families above: $\mathbf{M}^{\top} \neq \pm\mathbf{M}$ and $\mathbf{M}^{\top}\mathbf{M} = 5\mathbf{I} \neq \mathbf{I}$. Still,

$$\mathbf{M}^{\top}\mathbf{M} = \mathbf{M}\mathbf{M}^{\top} = \begin{pmatrix} 5 & 0 \\ 0 & 5 \end{pmatrix}.$$

This is $\sqrt{5}$ times a rotation of the plane, which is the rotation-scaling block $\mathbf{\Lambda}(2 + i)$ transposed (Definition 10.4.1). Normality is genuinely wider than "self-adjoint or unitary".

(f) **Not** normal:

$$\mathbf{N}^{\top}\mathbf{N} = \begin{pmatrix} 1 & 1 \\ 1 & 5 \end{pmatrix}, \qquad \mathbf{N}\mathbf{N}^{\top} = \begin{pmatrix} 2 & 2 \\ 2 & 4 \end{pmatrix}.$$

The $(1, 1)$ entries already disagree, $1 \neq 2$. This matrix is diagonalizable, with eigenvalues 1 and 2; Section 4 shows what it costs to be diagonalizable and not normal.

12.3.2 Normality is a statement about lengths

The definition compares two operators. The next theorem turns it into a comparison of two numbers, one vector at a time, and that is the form every later proof uses. The engine is the zero test of Section 2: a self-adjoint operator S with $\langle Sv, v\rangle = 0$ for all v is zero.

Theorem 12.3.1. Characterizations of Normality

Let V be a finite-dimensional inner product space over F and let $T \in \mathcal{L}(V)$. The following are equivalent.

(a) T is normal.
(b) $\|Tv\| = \|T^*v\|$ for every $v \in V$.
(c) $\langle Tu, Tv\rangle = \langle T^*u, T^*v\rangle$ for all $u, v \in V$.

Suppose in addition that $F = \mathbb{C}$, and set

$$A = \tfrac{1}{2}(T + T^*), \qquad B = \tfrac{1}{2i}(T - T^*),$$

so that A and B are self-adjoint and $T = A + iB$. Then (a)–(c) are further equivalent to

(d) $AB = BA$.

Idea. Condition (a) says that the operator $S = T^*T - TT^*$ is zero. That operator is self-adjoint, so by the zero test of Theorem 12.2.5 it is enough to know $\langle Sv, v\rangle = 0$ for every

v — and $\langle Sv, v\rangle$ is exactly $\|Tv\|^2 - \|T^*v\|^2$. So (a) and (b) are the same statement read at two different levels of resolution. Condition (c) is (b) with the two slots unglued, and it costs nothing in one direction and a specialization in the other. Condition (d) is the operator version of writing a complex number as $a+bi$; that decomposition appeared in Exercise (C2: Skew-adjoint operators) (c), and multiplying out $TT^* - T^*T$ in terms of A and B leaves only the commutator.

Proof.

(a) $\Rightarrow$ (c). For all $u, v \in V$, by Definition 11.6.1 and $T^{**} = T$ (Theorem 11.6.2 (d)),

$$\langle Tu, Tv\rangle = \langle T^*Tu, v\rangle, \qquad \langle T^*u, T^*v\rangle = \langle TT^*u, v\rangle.$$

Since $T^*T = TT^*$, the two right-hand sides agree.

(b) $\Rightarrow$ (b). Take $u = v$. Then $\|Tv\|^2 = \|T^*v\|^2$, and both norms are nonnegative reals, so they are equal.

(c) $\Rightarrow$ (a). Put $S = T^*T - TT^*$. Then S is self-adjoint: by Theorem 11.6.2 (a), (c), and (d),

$$S^* = (T^*T)^* - (TT^*)^* = T^*T^{**} - T^{**}T^* = T^*T - TT^* = S.$$

For any $v \in V$, using Definition 11.6.1 twice,

$$\begin{aligned}\langle Sv, v\rangle &= \langle T^*Tv, v\rangle - \langle TT^*v, v\rangle \\ &= \langle Tv, Tv\rangle - \langle T^*v, T^*v\rangle \\ &= \|Tv\|^2 - \|T^*v\|^2 = 0\end{aligned}$$

by (b). Since S is self-adjoint, Theorem 12.2.5 gives $S = 0$, that is $T^*T = TT^*$.

Now let $F = \mathbb{C}$. By Theorem 11.6.2 (a), (b), and (d),

$$\begin{aligned}A^* &= \tfrac{1}{2}(T^* + T) = A, \\ B^* &= \overline{\left(\tfrac{1}{2i}\right)}(T^* - T) = \tfrac{1}{2i}(T - T^*) = B,\end{aligned}$$

so both are self-adjoint, and $A + iB = \frac{1}{2}(T + T^*) + \frac{1}{2}(T - T^*) = T$; this is the decomposition of Exercise (C2: Skew-adjoint operators) (c). Expanding $T = A + iB$ and $T^* = A - iB$,

$$\begin{aligned}TT^* &= (A + iB)(A - iB) = A^2 + B^2 + i(BA - AB), \\ T^*T &= (A - iB)(A + iB) = A^2 + B^2 + i(AB - BA).\end{aligned}$$

Subtracting, $TT^* - T^*T = 2i(BA - AB)$. So T is normal exactly when $AB = BA$, which is (a) $\Leftrightarrow$ (d). This proves the theorem.

■

Condition (b) is the one to remember. It says that T and T^* stretch every single vector by the same amount, even though they need not move it in the same direction. Everything in the next three sections is squeezed out of that sentence. Condition (d) records that normality over $\mathbb{C}$ is the operator analogue of the fact that any two real numbers commute: split T into a "real part" and an "imaginary part", and normality is precisely the failure of those two parts to interfere.

Quick check. The symmetric matrix $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ and the skew-symmetric matrix $\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$ are both normal, the first being self-adjoint and the second skew-adjoint. Is their sum normal?

Solution. No. The sum is $\begin{pmatrix} 0 & 2 \\ 0 & 0 \end{pmatrix} = 2\mathbf{J}_2(0)$, which is twice the standard nonnormal example: it kills $\mathbf{e}_1$ and sends $\mathbf{e}_2$ to a vector of length 2, while its transpose does the opposite, so condition (b) fails at $\mathbf{v} = \mathbf{e}_1$. The normal operators do not form a subspace of $\mathcal{L}(V)$, and they are not closed under composition either.

12.3.3 What normality buys

Four consequences, all short, all obtained by turning some statement about T into the same statement about T^*.

Corollary 12.3.2. Kernel, Image, and an Orthogonal Splitting

Let $T \in \mathcal{L}(V)$ be normal on a finite-dimensional inner product space V. Then:

(a) $\ker T = \ker T^*$;
(b) $\operatorname{im} T = \operatorname{im} T^*$;
(c) $V = \ker T \oplus \operatorname{im} T$, and the two summands are orthogonal to each other.

Proof.

(a) For $v \in V$, Theorem 12.3.1 (b) gives $\|Tv\| = \|T^*v\|$. So $Tv = 0$ if and only if $\|Tv\| = 0$, if and only if $\|T^*v\| = 0$, if and only if $T^*v = 0$.

(b) By Theorem 11.6.5 (b) applied to T^* in place of T, and using $T^{**} = T$ (Theorem 11.6.2 (d)),
$$\operatorname{im} T = \operatorname{im} T^{**} = (\ker T^*)^{\perp} = (\ker T)^{\perp} = \operatorname{im} T^*,$$
where the third equality is (a) and the fourth is Theorem 11.6.5 (b) for T.

(c) By (b), $\operatorname{im} T = (\ker T)^{\perp}$, and $V = \ker T \oplus (\ker T)^{\perp}$ by Theorem 11.3.2 (a). Orthogonality of the summands is the definition of the orthogonal complement.

■

Part (c) fails badly without normality, and the failure is not subtle: for a general operator $\ker T$ and $\operatorname{im} T$ can be **equal**. Take T with matrix $\mathbf{J}_2(0) = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$ on $\mathbb{C}^2$. Then $T\mathbf{e}_1 = \mathbf{0}$ and $T\mathbf{e}_2 = \mathbf{e}_1$, so
$$\ker T = \operatorname{span}(\mathbf{e}_1) = \operatorname{im} T.$$
The sum $\ker T + \operatorname{im} T$ is one-dimensional, so it is neither direct nor all of $\mathbb{C}^2$. Normality is exactly what rules this out.

Theorem 12.3.3. An Eigenvector of a Normal Operator Is an Eigenvector of Its Adjoint

Let $T \in \mathcal{L}(V)$ be normal, let $\lambda \in F$, and let $v \in V$. Then

$$Tv = \lambda v \quad \text{if and only if} \quad T^*v = \overline{\lambda}\, v.$$

In particular $E_\lambda(T) = E_{\overline{\lambda}}(T^*)$, and $\lambda \in \operatorname{spec}(T)$ if and only if $\overline{\lambda} \in \operatorname{spec}(T^*)$.

Idea. Both sides say that v lies in a kernel: the left in $\ker(T - \lambda \operatorname{id}_V)$, the right in $\ker(T^* - \overline{\lambda} \operatorname{id}_V)$. Those two operators are adjoint to each other, so Corollary 12.3.2 (a) closes the gap — provided $T - \lambda \operatorname{id}_V$ is itself normal, which is the only thing to check.

Proof. Write $S = T - \lambda \operatorname{id}_V$. By Theorem 11.6.2 (a), (b), and (d), $S^* = T^* - \overline{\lambda} \operatorname{id}_V$. Since $T^*T = TT^*$ and id_V commutes with everything,

$$\begin{aligned} S^*S &= T^*T - \overline{\lambda}T - \lambda T^* + |\lambda|^2 \operatorname{id}_V \\ &= TT^* - \lambda T^* - \overline{\lambda}T + |\lambda|^2 \operatorname{id}_V = SS^*, \end{aligned}$$

so S is normal. By Corollary 12.3.2 (a), $\ker S = \ker S^*$. Now $Tv = \lambda v$ says $v \in \ker S$, and $T^*v = \overline{\lambda}v$ says $v \in \ker S^*$, so the two conditions agree. The statement about eigenspaces is the equality $\ker S = \ker S^*$ rewritten, and the statement about spectra follows because an eigenspace is nonzero exactly when the scalar is an eigenvalue (Theorem 9.2.1).

■

The theorem is the reason normal operators are so easy to handle: T and T^* have the *same* eigenvectors, with conjugate eigenvalues. The proof of the next corollary is the classical self-adjoint argument (Theorem 12.2.3) with one conjugate bar inserted, and that bar is exactly what Theorem 12.3.3 supplies.

Corollary 12.3.4. Eigenvectors for Distinct Eigenvalues Are Orthogonal

Let $T \in \mathcal{L}(V)$ be normal, let $\lambda \neq \mu$ be eigenvalues of T, and let $Tu = \lambda u$ and $Tv = \mu v$. Then $\langle u, v \rangle = 0$. Hence $E_\lambda(T) \perp E_\mu(T)$ whenever $\lambda \neq \mu$.

Proof. Compute $\langle Tu, v \rangle$ in two ways. First, $\langle Tu, v \rangle = \langle \lambda u, v \rangle = \lambda \langle u, v \rangle$. Second, by Definition 11.6.1 and then Theorem 12.3.3 applied to v,

$$\langle Tu, v \rangle = \langle u, T^*v \rangle = \langle u, \overline{\mu}\, v \rangle = \mu \langle u, v \rangle,$$

the last step because a scalar leaves the second slot conjugated. Subtracting, $(\lambda - \mu)\langle u, v \rangle = 0$, and $\lambda - \mu \neq 0$, so $\langle u, v \rangle = 0$. Every vector of $E_\lambda(T)$ is orthogonal to every vector of $E_\mu(T)$ by the same computation.

■

Theorem 12.3.5. A Normal Operator Has No Hidden Kernel

Let $T \in \mathcal{L}(V)$ be normal. Then $\ker T^k = \ker T$ for every integer $k \geq 1$. In particular, if $T^k = 0$ for some $k \geq 1$ — that is, if T is **nilpotent** (Definition 10.2.1) — then $T = 0$.

Idea. The inclusion $\ker T \subseteq \ker T^2$ is free, and the whole content is the reverse one. If $T^2 v = 0$, then the vector $w = Tv$ lies in $\ker T$, hence also in $\ker T^*$ by Corollary 12.3.2; and a vector that is killed by T^* and is itself in the image of T must be 0, because pairing it with itself moves T across the inner product and hits the T^*.

Proof. First take $k = 2$. The inclusion $\ker T \subseteq \ker T^2$ holds for every operator. Conversely, suppose $T^2 v = 0$ and put $w = Tv$. Then $Tw = 0$, so $T^* w = 0$ by Corollary 12.3.2 (a). Therefore, by Definition 11.6.1,

$$\begin{aligned} \|w\|^2 &= \langle w, w \rangle = \langle Tv, w \rangle \\ &= \langle v, T^* w \rangle = \langle v, 0 \rangle = 0, \end{aligned}$$

so $w = Tv = 0$ and $v \in \ker T$. Hence $\ker T^2 = \ker T$.

Now induct on k. The cases $k = 1, 2$ are done. Let $k \geq 2$ and assume $\ker T^k = \ker T$. If $T^{k+1} v = 0$, then $T^2(T^{k-1} v) = 0$, so $T^k v = T(T^{k-1} v) = 0$ by the case $k = 2$, and then $Tv = 0$ by the inductive hypothesis. So $\ker T^{k+1} \subseteq \ker T$, and the reverse inclusion is again free.

Finally, if $T^k = 0$ for some $k \geq 1$, then $\ker T = \ker T^k = V$, so $T = 0$.

■

This is the sentence that explains why normal operators never produce Jordan blocks. Over $\mathbb{C}$, a Jordan block of size $m \geq 2$ for the eigenvalue λ contributes a vector killed by $(T - \lambda \operatorname{id}_V)^m$ but not by $T - \lambda \operatorname{id}_V$ (Definition 10.2.2). For a normal T the operator $T - \lambda \operatorname{id}_V$ is normal, as the proof of Theorem 12.3.3 showed, so Theorem 12.3.5 forbids such a vector. Every block in the Jordan form of a normal operator therefore has size 1, and a normal operator on a complex space is diagonalizable (Theorem 10.3.1, Definition 9.4.1).

Remark. That is diagonalizable, not yet **orthogonally** diagonalizable. The eigenspaces of distinct eigenvalues are already mutually orthogonal by Corollary 12.3.4, so all that is missing is an orthonormal basis inside each eigenspace, which Gram–Schmidt supplies (Theorem 11.2.2). Section 4 proves the whole statement in one stroke instead, from Schur's theorem, because that proof also delivers the converse and works without Chapter 10.

Two converses are worth flagging now, so that you know they are coming and do not try to prove them here. A normal operator whose eigenvalues are all real is self-adjoint, and a normal operator whose eigenvalues all have modulus 1 is unitary. Both are false without normality, and neither can be reached with the tools of this section: they need an orthonormal basis of eigenvectors, which is the business of Section 4.

Before the warning, one structural fact deserves to be stated where later sections can lean on it. It is the reason a normal operator can be taken apart orthogonally at all.

Proposition 12.3.6. An Invariant Subspace of a Normal Operator Is Reducing

Let $T \in \mathcal{L}(V)$ be normal on a finite-dimensional inner product space, and let $U \subseteq V$ be T-invariant. Then U is T^*-invariant, $U^\perp$ is T-invariant, and $T|_U$ is a normal operator on U.

Idea. Write the matrix in an orthonormal basis adapted to U. Invariance kills the block below the diagonal, and normality, read on the top-left block alone, kills the block above it. The trace is what converts "these two blocks agree" into "that block is zero".

Proof. Choose an orthonormal basis $\mathcal{B}_U = (e_1, \dots, e_r)$ of U and extend it to an orthonormal basis $\mathcal{B}$ of V (Corollary 11.2.3). Since $Te_j \in U$ for $j \le r$, the first r columns of $\mathbf{A} = [T]_{\mathcal{B}}^{\mathcal{B}}$ have zeros below row r, so

$$\mathbf{A} = \begin{pmatrix} \mathbf{B} & \mathbf{C} \\ \mathbf{O} & \mathbf{D} \end{pmatrix},$$

with $\mathbf{B}$ of size $r \times r$. Because $\mathcal{B}$ is orthonormal, $[T^*]_{\mathcal{B}}^{\mathcal{B}} = \mathbf{A}^*$ (Theorem 11.6.3), and T normal means $\mathbf{A}^*\mathbf{A} = \mathbf{A}\mathbf{A}^*$. Comparing the $(1,1)$ blocks (Theorem 8.1.1),

$$(\mathbf{A}^*\mathbf{A})_{11} = \mathbf{B}^*\mathbf{B}, \qquad (\mathbf{A}\mathbf{A}^*)_{11} = \mathbf{B}\mathbf{B}^* + \mathbf{C}\mathbf{C}^*.$$

Taking traces and using $\operatorname{tr}(\mathbf{B}^*\mathbf{B}) = \operatorname{tr}(\mathbf{B}\mathbf{B}^*)$ (Theorem 0.9.6 (3)) gives $\operatorname{tr}(\mathbf{C}\mathbf{C}^*) = 0$. But $\operatorname{tr}(\mathbf{C}\mathbf{C}^*) = \sum_{i,j} |c_{ij}|^2$, a sum of nonnegative reals, so $\mathbf{C} = \mathbf{O}$.

Hence $\mathbf{A} = \mathbf{B} \oplus \mathbf{D}$. Reading the last $n - r$ columns, $Te_j \in U^\perp$ for $j > r$, so $U^\perp$ is T-invariant; reading the first r columns of $\mathbf{A}^*$, U is T^*-invariant. Finally $[T|_U]_{\mathcal{B}_U}^{\mathcal{B}_U} = \mathbf{B}$ and the $(1,1)$ blocks above now read $\mathbf{B}^*\mathbf{B} = \mathbf{B}\mathbf{B}^*$, so $T|_U$ is normal. This shows all three claims. ∎

Warning. **Normality cannot be read off a corner.** A principal submatrix of a normal matrix need not be normal. The 3×3 cyclic shift

$$\mathbf{S} = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$$

is a permutation matrix, hence orthogonal (Exercise (B3: Permutations are orthogonal)), hence normal; but deleting its last row and column leaves $\begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}$, the transpose of $\mathbf{J}_2(0)$, which is not normal. Equivalently: the operator that $\mathbf{S}$ compresses to $\operatorname{span}(\mathbf{e}_1, \mathbf{e}_2)$ is not normal, because that subspace is not $\mathbf{S}$-invariant. What **is** true is that an invariant subspace of a normal operator is automatically invariant under the adjoint too, and the restriction there is normal; that is Proposition 12.3.6, proved just above.

12.3.4 Exercises

12.3.4.1 A. Check your understanding

Exercise. A1

(a) State the definition of a normal operator, and state one equivalent condition phrased in terms of norms.

(b) True or false: every self-adjoint operator is normal, and every normal operator is self-adjoint. Justify both halves.

(c) Let T be normal with $Tv = 3v$. What is T^*v? What if $Tv = 3iv$?

(d) Give a normal operator on $\mathbb{R}^2$ that is neither self-adjoint nor unitary.

(e) True or false: if S and T are normal then $S + T$ is normal. Justify your answer.

(f) Explain in one sentence why a normal operator cannot be nilpotent unless it is zero.

Solution.

(a) T is normal when $T^*T = TT^*$ (Definition 11.6.3). Equivalently, $\|Tv\| = \|T^*v\|$ for every $v \in V$ (Theorem 12.3.1 (b)).

(b) The first half is true: if $T^* = T$ then both products are T^2. The second half is false: $\operatorname{diag}(1, i)$ on $\mathbb{C}^2$ is normal and has a nonreal eigenvalue, so it is not self-adjoint (Theorem 12.2.1).

(c) $T^*v = 3v$ in the first case and $T^*v = -3iv$ in the second, by Theorem 12.3.3: the adjoint has the conjugate eigenvalue on the same eigenvector.

(d) $\begin{pmatrix} 2 & 1 \\ -1 & 2 \end{pmatrix}$ works (Example (Which of these are normal?) (e)): it is not symmetric, and $\mathbf{M}^\top \mathbf{M} = 5\mathbf{I} \neq \mathbf{I}$. Any $\mathbf{\Lambda}(\lambda)$ with $|\lambda| \neq 1$ and $\operatorname{Im} \lambda \neq 0$ does.

(e) False. $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 2 \\ 0 & 0 \end{pmatrix}$ is a sum of two normal matrices and is not normal.

(f) If $T^k = 0$ then $\ker T = \ker T^k = V$ by Theorem 12.3.5, so $T = 0$.

12.3.4.2 B. Practice

Exercise. B1: Determine which are normal

Determine which of the following matrices are normal, with the standard inner product. Justify your answer; for those that are not, exhibit a vector $\mathbf{v}$ with $\|\mathbf{Av}\| \neq \|\mathbf{A}^*\mathbf{v}\|$.

(a) $\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$ over $\mathbb{R}$.

(b) $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ over $\mathbb{R}$.

(c) $\begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix}$ over $\mathbb{C}$.

(d) $\begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix}$ over $\mathbb{R}$.

(e) $\begin{pmatrix} i & 1 \\ 0 & i \end{pmatrix}$ over $\mathbb{C}$.

Solution.

(a) Normal: the matrix is symmetric, so it is self-adjoint over $\mathbb{R}$, and self-adjoint implies normal.

(b) Not normal. $\mathbf{A}^\top\mathbf{A} = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}$ and $\mathbf{A}\mathbf{A}^\top = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$. With $\mathbf{v} = \mathbf{e}_1$, $\|\mathbf{A}\mathbf{e}_1\| = \|(1,0)\| = 1$ while $\|\mathbf{A}^\top\mathbf{e}_1\| = \|(1,1)\| = \sqrt{2}$.

(c) Normal. $\mathbf{A}^* = \begin{pmatrix} 0 & -i \\ -i & 0 \end{pmatrix} = -\mathbf{A}$, so $\mathbf{A}$ is skew-adjoint and $\mathbf{A}^*\mathbf{A} = -\mathbf{A}^2 = \mathbf{A}\mathbf{A}^*$. (Concretely both products are $\mathbf{I}$.)

(d) Normal: the matrix is $(1) \oplus \mathbf{\Lambda}(i)$, a direct sum of a 1×1 block and a rotation by a quarter turn, and its columns are orthonormal, so it is orthogonal (Theorem 11.7.3 (f)).

(e) Not normal. $\mathbf{A}^*\mathbf{A} = \begin{pmatrix} 1 & -i \\ i & 2 \end{pmatrix}$ and $\mathbf{A}\mathbf{A}^* = \begin{pmatrix} 2 & -i \\ i & 1 \end{pmatrix}$; with $\mathbf{v} = \mathbf{e}_1$ we get $\|\mathbf{A}\mathbf{e}_1\| = |i| = 1$ and $\|\mathbf{A}^*\mathbf{e}_1\| = \|(-i,1)\| = \sqrt{2}$. Adding $i\mathbf{I}$ to a nonnormal matrix does not repair it: $\mathbf{A} = i\mathbf{I} + \mathbf{J}_2(0)$ and $i\mathbf{I}$ is normal, but sums of normal matrices need not be normal.

Exercise. B2: The splitting of a singular normal operator

Let $\mathbf{C} = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \\ -1 & 0 & 1 \end{pmatrix} \in M_3(\mathbb{R})$, and let $T = T_{\mathbf{C}}$ be the operator it defines on $\mathbb{R}^3$.

(a) Verify that T is normal.

(b) Compute $\ker T$ and $\ker T^*$ and check that they agree.

(c) Compute $\operatorname{im} T$, and verify directly that $\mathbb{R}^3 = \ker T \oplus \operatorname{im} T$ with the two summands orthogonal, as Corollary 12.3.2 (c) predicts.

Solution.

(a) Each row of $\mathbf{C}$ is the previous one shifted cyclically, so $\mathbf{C}$ is a circulant. Explicitly,

$$\mathbf{C}^\top\mathbf{C} = \mathbf{C}\mathbf{C}^\top = \begin{pmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{pmatrix},$$

so T is normal.

(b) $\mathbf{C}\mathbf{x} = \mathbf{0}$ reads $x_1 = x_2 = x_3$, because the first two rows give $x_1 = x_2$ and $x_2 = x_3$. So

$\ker T = \operatorname{span}((1,1,1))$. The transpose $\mathbf{C}^\top$ has rows $(1,0,-1)$, $(-1,1,0)$, $(0,-1,1)$, and $\mathbf{C}^\top\mathbf{x} = \mathbf{0}$ again reads $x_1 = x_3$ and $x_1 = x_2$. The two kernels agree, as Corollary 12.3.2 (a) requires.

(c) The rank is $3 - 1 = 2$ by Theorem 2.3.1, and the first two columns $(1,0,-1)$ and $(-1,1,0)$ are independent, so

$$\operatorname{im} T = \operatorname{span}((1,0,-1),\ (-1,1,0)).$$

Both spanning vectors have coordinate sum 0, hence are orthogonal to $(1,1,1)$, so $\operatorname{im} T \perp \ker T$. Since $\dim \ker T + \dim \operatorname{im} T = 1 + 2 = 3$ and the two subspaces meet only in $\mathbf{0}$ (being orthogonal), $\mathbb{R}^3 = \ker T \oplus \operatorname{im} T$. Indeed $\operatorname{im} T = \{\mathbf{x} : x_1 + x_2 + x_3 = 0\} = (\ker T)^\perp$.

Exercise. B3: Triangular and normal, by hand

Let $\mathbf{A} = \begin{pmatrix} a & b \\ 0 & d \end{pmatrix} \in M_2(\mathbb{C})$.

(a) Compute the $(1,1)$ entries of $\mathbf{A}^*\mathbf{A}$ and of $\mathbf{A}\mathbf{A}^*$.

(b) Hence prove that $\mathbf{A}$ is normal if and only if $b = 0$, that is, if and only if $\mathbf{A}$ is diagonal.

(c) Which upper triangular matrices in $M_2(\mathbb{C})$ are unitary?

Hint: for (a), you only need one entry of each product.

Solution.

(a) The $(1,1)$ entry of $\mathbf{A}^*\mathbf{A}$ is the squared norm of the first **column** of $\mathbf{A}$, namely $|a|^2$. The $(1,1)$ entry of $\mathbf{A}\mathbf{A}^*$ is the squared norm of the first **row**, namely $|a|^2 + |b|^2$. In full,

$$\mathbf{A}^*\mathbf{A} = \begin{pmatrix} |a|^2 & \overline{a}b \\ a\overline{b} & |b|^2 + |d|^2 \end{pmatrix}, \qquad \mathbf{A}\mathbf{A}^* = \begin{pmatrix} |a|^2 + |b|^2 & b\overline{d} \\ \overline{b}d & |d|^2 \end{pmatrix}.$$

(b) ($\Rightarrow$) If $\mathbf{A}$ is normal, comparing the $(1,1)$ entries gives $|a|^2 = |a|^2 + |b|^2$, so $|b|^2 = 0$ and $b = 0$. ($\Leftarrow$) A diagonal matrix is normal, since $\mathbf{D}^*\mathbf{D} = \mathbf{D}\mathbf{D}^* = \operatorname{diag}(|a|^2, |d|^2)$.

(c) A unitary matrix is normal, so by (b) it must be diagonal, $\operatorname{diag}(a,d)$, and then $\mathbf{A}^*\mathbf{A} = \operatorname{diag}(|a|^2, |d|^2) = \mathbf{I}$ forces $|a| = |d| = 1$. Conversely every such matrix is unitary. So the upper triangular unitary matrices are exactly $\operatorname{diag}(a,d)$ with $|a| = |d| = 1$.

12.3.4.3 C. Going deeper

Exercise. C1: A normal operator with no proper invariant subspace

Let V be a finite-dimensional complex inner product space with $V \neq \{0\}$, and let $T \in \mathcal{L}(V)$ be normal. Suppose the only T-invariant subspaces of V are $\{0\}$ and V.

(a) Show that $\dim V = 1$.

(b) Show that the conclusion fails over $\mathbb{R}$, by exhibiting a normal operator on $\mathbb{R}^2$ whose only invariant subspaces are $\{\mathbf{0}\}$ and $\mathbb{R}^2$.

(c) Where exactly does the argument of (a) use that the field is $\mathbb{C}$?

Hint: an eigenvector spans an invariant subspace.

Solution.

(a) Since $V \neq \{0\}$ and the field is $\mathbb{C}$, T has an eigenvector $v \neq 0$ (Theorem 9.2.4). Then $\operatorname{span}(v)$ is T-invariant and nonzero, so by hypothesis $\operatorname{span}(v) = V$ and $\dim V = 1$. Normality is not needed for this direction.

(b) The quarter turn $\mathbf{R}_{\pi/2} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ is orthogonal, hence normal. A 1-dimensional invariant subspace would be spanned by a real eigenvector, and $\mathbf{R}_{\pi/2}$ has none: its characteristic polynomial is $x^2 + 1$, with no real root. So its only invariant subspaces are $\{\mathbf{0}\}$ and $\mathbb{R}^2$, while $\dim \mathbb{R}^2 = 2$.

(c) Only in the existence of an eigenvector. Over $\mathbb{C}$ that is the Fundamental Theorem of Algebra (Theorem 6.5.1) reaching the characteristic polynomial; over $\mathbb{R}$ the polynomial may have no root, which is exactly what (b) exploits. Note that Proposition 12.3.6 still holds over $\mathbb{R}$ — it says nothing about invariant subspaces *existing*.

Exercise. C2: Arithmetic of normal operators

Let $S, T \in \mathcal{L}(V)$ be normal.

(a) Prove that T^* is normal and that cT is normal for every $c \in F$.

(b) Prove that $p(T)$ is normal for every polynomial $p \in F[x]$.

(c) Give a concrete S and T, both normal, with ST not normal. (Part (b) shows that this cannot happen when S is a polynomial in T.)

Solution.

(a) $(T^*)^* T^* = TT^* = T^*T = T^*(T^*)^*$, using $T^{**} = T$ (Theorem 11.6.2 (d)). For cT, by Theorem 11.6.2 (b) we have $(cT)^* = \overline{c}T^*$, so

$$(cT)^*(cT) = \overline{c}c\, T^*T = |c|^2\, TT^* = (cT)(cT)^*.$$

(b) Write $p(x) = \sum_k a_k x^k$. Then $p(T)^* = \sum_k \overline{a_k}(T^*)^k$ by Theorem 11.6.2 (a), (b), and (c). Every power of T commutes with every power of T^*: from $TT^* = T^*T$ one moves a T^* past one T at a time. Hence $p(T)$ and $p(T)^*$, being sums of products of such powers, commute, and $p(T)$ is normal.

(c) Take $S = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$ and $T = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$, both symmetric and hence normal. Then

$$ST = \begin{pmatrix} 0 & 1 \\ 2 & 0 \end{pmatrix},$$

$$(ST)^\top (ST) = \begin{pmatrix} 4 & 0 \\ 0 & 1 \end{pmatrix}, \qquad (ST)(ST)^\top = \begin{pmatrix} 1 & 0 \\ 0 & 4 \end{pmatrix},$$

so ST is not normal.

Exercise. C3: Normality depends on the inner product

Normality is a property of T **together with** an inner product, not of T alone. Work on $V = \mathbb{R}^2$ with the operator T whose standard matrix is $\mathbf{A} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$.

(a) Verify that T is normal for the standard inner product.

(b) Now use $\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{G}} = \mathbf{y}^\top \mathbf{G} \mathbf{x}$ with $\mathbf{G} = \operatorname{diag}(1, 2)$, which is an inner product on $\mathbb{R}^2$. Using the rule $[T^*]_{\mathcal{E}}^{\mathcal{E}} = \mathbf{G}^{-1} \mathbf{A}^\top \mathbf{G}$ in the standard basis $\mathcal{E}$ (Exercise (C3: The adjoint in an arbitrary basis)), compute the matrix of the adjoint of T for this inner product.

(c) Show that T is **not** normal for $\langle \cdot, \cdot \rangle_{\mathbf{G}}$, and reconcile this with (a).

Solution.

(a) $\mathbf{A}^\top \mathbf{A} = \mathbf{A} \mathbf{A}^\top = \mathbf{I}$, so T is even orthogonal, hence normal.

(b) $\mathbf{G}^{-1} = \operatorname{diag}(1, \frac{1}{2})$, so

$$\mathbf{G}^{-1} \mathbf{A}^\top \mathbf{G} = \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{2} \end{pmatrix} \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} = \begin{pmatrix} 0 & 2 \\ -\frac{1}{2} & 0 \end{pmatrix}.$$

Call this matrix $\mathbf{A}^\star$.

(c) Computing both products,

$$\mathbf{A} \mathbf{A}^\star = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & 2 \end{pmatrix}, \qquad \mathbf{A}^\star \mathbf{A} = \begin{pmatrix} 2 & 0 \\ 0 & \frac{1}{2} \end{pmatrix},$$

which differ, so T is not normal for $\langle \cdot, \cdot \rangle_{\mathbf{G}}$. There is no contradiction: the adjoint is built from the inner product (Definition 11.6.1), so changing the inner product changes T^*, and with it the equation $T^* T = T T^*$. The standard basis is orthonormal for the first inner product and not for the second, which is exactly the hypothesis Theorem 11.6.3 needs.

12.4 The Complex Spectral Theorem

Two things are now on the table. Section 1 put every complex operator into upper triangular form in an **orthonormal** basis, and Section 3 identified the operators that commute with their adjoints. Put them together and the chapter's main theorem falls out in a few lines: a triangular matrix that is normal has nothing above the diagonal. This section proves that, reads off the consequences, and says exactly how much of the resulting picture is unique.

The field is $\mathbb{C}$ throughout this section, and every inner product space is finite-dimensional. The reason for insisting on $\mathbb{C}$ is Schur's theorem, and behind it the Fundamental Theorem of Algebra (Theorem 6.5.1): over $\mathbb{R}$ an operator can fail to have a single eigenvector, and Section 5 shows what has to change.

12.4.1 Triangular plus normal forces diagonal

Start with the smallest case, where the mechanism is visible at a glance. Let

$$\mathbf{A} = \begin{pmatrix} a & b \\ 0 & d \end{pmatrix} \in M_2(\mathbb{C})$$

be upper triangular. The $(1,1)$ entry of $\mathbf{A}^*\mathbf{A}$ is the squared length of the first **column** of $\mathbf{A}$, and the $(1,1)$ entry of $\mathbf{A}\mathbf{A}^*$ is the squared length of the first **row**:

$$(\mathbf{A}^*\mathbf{A})_{11} = |a|^2, \qquad (\mathbf{A}\mathbf{A}^*)_{11} = |a|^2 + |b|^2.$$

If $\mathbf{A}$ is normal these are equal, so $|b|^2 = 0$ and $b = 0$. One entry of one product did all the work. (This is Exercise (B3: Triangular and normal, by hand), which you may already have done.)

The general case is the same comparison run down the diagonal, one row at a time. Nothing is clever; the only care needed is to use, at step j, the zeros that steps $1, \dots, j-1$ have already produced.

Lemma 12.4.1. A Normal Triangular Matrix Is Diagonal

Let $\mathbf{T} = (t_{ij}) \in M_n(\mathbb{C})$ be upper triangular and normal. Then $\mathbf{T}$ is diagonal.

Idea. For any square matrix, the j-th diagonal entry of $\mathbf{T}^*\mathbf{T}$ is the squared length of column j, and the j-th diagonal entry of $\mathbf{T}\mathbf{T}^*$ is the squared length of row j. Normality says those two lengths agree for every j. For an upper triangular matrix, column j reaches only down to the diagonal and row j only rightwards from it, so the equation compares "what sits above t_{jj}" with "what sits to the right of t_{jj}". Row 1 has nothing above it, so the right side must be empty too; that clears row 1, which empties the column sums one step further down, and the argument walks down the diagonal.

Proof. For each j, computing the (j,j) entries of the two products entry by entry,

$$(\mathbf{T}^*\mathbf{T})_{jj} = \sum_{i=1}^{n} \overline{t_{ij}}\, t_{ij} = \sum_{i=1}^{n} |t_{ij}|^2, \qquad (\mathbf{T}\mathbf{T}^*)_{jj} = \sum_{k=1}^{n} |t_{jk}|^2.$$

Since $\mathbf{T}$ is upper triangular, $t_{ij} = 0$ whenever $i > j$, and $t_{jk} = 0$ whenever $k < j$. So normality of $\mathbf{T}$ says

$$\sum_{i\le j}|t_{ij}|^2=\sum_{k\ge j}|t_{jk}|^2 \qquad (j=1,\dots,n). \tag{12.4.1}$$

We prove by induction on j that $t_{jk}=0$ for every $k>j$.

Base case $j=1$. The left-hand side of Eq. 12.4.1 has the single term $|t_{11}|^2$, so

$$|t_{11}|^2=|t_{11}|^2+\sum_{k>1}|t_{1k}|^2.$$

Hence $\sum_{k>1}|t_{1k}|^2=0$, a sum of nonnegative reals, so $t_{1k}=0$ for every $k>1$.

Inductive step. Let $1<j\le n$ and suppose $t_{mk}=0$ whenever $m<j$ and $k>m$. In the left-hand side of Eq. 12.4.1, every term with $i<j$ is $|t_{ij}|^2$ with $i<j$, hence with $j>i$; by the inductive hypothesis, taking $m=i$ and $k=j$, each such t_{ij} is 0. So the left-hand side reduces to $|t_{jj}|^2$ and Eq. 12.4.1 becomes

$$|t_{jj}|^2=|t_{jj}|^2+\sum_{k>j}|t_{jk}|^2,$$

whence $t_{jk}=0$ for every $k>j$, again because a sum of nonnegative reals vanishes only when every term does.

By induction, every entry strictly above the diagonal is zero. Every entry strictly below it is zero because $\mathbf{T}$ is upper triangular. So $\mathbf{T}$ is diagonal, as claimed.

■

Note what the proof did **not** need: no eigenvector was chosen, no dimension was inducted on, no subspace was split off. The whole argument is an identity between two lists of nonnegative numbers. That is why it also survives, unchanged in spirit, into the block version used in Section 5 and the quantitative version in Section 11.

12.4.2 The theorem

Theorem 12.4.2. Complex Spectral Theorem

Let V be a finite-dimensional inner product space over $\mathbb{C}$ and let $T\in\mathcal{L}(V)$. The following are equivalent.

(a) T is normal.

(b) V has an orthonormal basis consisting of eigenvectors of T.

Idea. ($\Leftarrow$) is a one-line check: in such a basis the matrix of T is diagonal, the matrix of T^* is its conjugate transpose, and two diagonal matrices commute. ($\Rightarrow$) is Schur plus the lemma: take an orthonormal basis making the matrix triangular, observe that normality of T is normality of that matrix because the basis is orthonormal, and apply Lemma 12.4.1. A diagonal matrix in the basis $\mathcal{B}$ says precisely that each vector of $\mathcal{B}$ is an eigenvector.

Proof. ($\Leftarrow$) Suppose $\mathcal{B} = (e_1, \dots, e_n)$ is an orthonormal basis with $Te_j = \lambda_j e_j$ for each j. Then $\mathbf{D} := [T]_{\mathcal{B}}^{\mathcal{B}} = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$, and since $\mathcal{B}$ is orthonormal, $[T^*]_{\mathcal{B}}^{\mathcal{B}} = \mathbf{D}^* = \operatorname{diag}(\overline{\lambda_1}, \dots, \overline{\lambda_n})$ by Theorem 11.6.3. Both products are the same diagonal matrix:

$$\mathbf{D}^*\mathbf{D} = \mathbf{D}\mathbf{D}^* = \operatorname{diag}(|\lambda_1|^2, \dots, |\lambda_n|^2).$$

Taking the matrix of an operator in a fixed basis turns composition into matrix multiplication and is injective (Theorem 4.1.2, Theorem 4.1.4), so $T^*T = TT^*$.

($\Rightarrow$) Suppose T is normal. By Schur's theorem (Theorem 12.1.2) there is an orthonormal basis $\mathcal{B} = (e_1, \dots, e_n)$ of V in which $\mathbf{T} := [T]_{\mathcal{B}}^{\mathcal{B}}$ is upper triangular. Because $\mathcal{B}$ is orthonormal, Theorem 11.6.3 gives $[T^*]_{\mathcal{B}}^{\mathcal{B}} = \mathbf{T}^*$, so the equation $T^*T = TT^*$ reads $\mathbf{T}^*\mathbf{T} = \mathbf{T}\mathbf{T}^*$: the matrix $\mathbf{T}$ is normal. By Lemma 12.4.1, $\mathbf{T}$ is diagonal, say $\mathbf{T} = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$. Reading off the j-th column (Definition 4.1.1), $Te_j = \lambda_j e_j$, and $e_j \neq 0$ because $\|e_j\| = 1$. So $\mathcal{B}$ is an orthonormal basis of eigenvectors of T. This proves the theorem.

■

Corollary 12.4.3. Unitary Diagonalization

Let $\mathbf{A} \in M_n(\mathbb{C})$. Then $\mathbf{A}$ is normal if and only if there exist a unitary $\mathbf{U} \in \mathrm{U}(n)$ and a diagonal $\mathbf{D} \in M_n(\mathbb{C})$ with

$$\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{U}^*.$$

A matrix admitting such a factorization is called **unitarily diagonalizable**. The diagonal of $\mathbf{D}$ is the list of eigenvalues of $\mathbf{A}$ with multiplicity, in an order that may be prescribed in advance, and the columns of $\mathbf{U}$ are a corresponding orthonormal basis of eigenvectors.

Proof. Apply Theorem 12.4.2 to the operator $T_{\mathbf{A}}$ on $\mathbb{C}^n$ with its standard inner product, for which the standard basis is orthonormal.

($\Rightarrow$) Let $(\mathbf{u}_1, \dots, \mathbf{u}_n)$ be an orthonormal basis of eigenvectors, $\mathbf{A}\mathbf{u}_j = \lambda_j \mathbf{u}_j$, and let $\mathbf{U}$ be the matrix with these columns; it is unitary because its columns are orthonormal (Theorem 11.7.3 (f)). Then $\mathbf{A}\mathbf{U}$ has j-th column $\lambda_j \mathbf{u}_j$, which is the j-th column of $\mathbf{U}\mathbf{D}$ for $\mathbf{D} = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$. So $\mathbf{A}\mathbf{U} = \mathbf{U}\mathbf{D}$ and $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{U}^{-1} = \mathbf{U}\mathbf{D}\mathbf{U}^*$.

($\Leftarrow$) If $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{U}^*$ then $\mathbf{A}^* = \mathbf{U}\mathbf{D}^*\mathbf{U}^*$, and

$$\mathbf{A}^*\mathbf{A} = \mathbf{U}\mathbf{D}^*\mathbf{D}\mathbf{U}^* = \mathbf{U}\mathbf{D}\mathbf{D}^*\mathbf{U}^* = \mathbf{A}\mathbf{A}^*,$$

using $\mathbf{U}^*\mathbf{U} = \mathbf{I}$ twice and the fact that diagonal matrices commute.

Finally, $\mathbf{A}$ is similar to $\mathbf{D}$, so they have the same characteristic polynomial (Theorem 7.10.3), and the roots of $\det(x\mathbf{I} - \mathbf{D}) = \prod_j (x - \lambda_j)$ are the diagonal entries with multiplicity. Reordering the eigenvectors reorders the columns of $\mathbf{U}$ and the diagonal of $\mathbf{D}$ in the same way, so any order may be prescribed.

■

Example. Unitarily diagonalizing a Hermitian matrix

Find a unitary $\mathbf{U}$ and a diagonal $\mathbf{D}$ with $\mathbf{U}^*\mathbf{A}\mathbf{U} = \mathbf{D}$, for

$$\mathbf{A} = \begin{pmatrix} 2 & 1-i \\ 1+i & 3 \end{pmatrix}.$$

Solution. $\mathbf{A}^* = \mathbf{A}$, so $\mathbf{A}$ is Hermitian, hence normal, and Corollary 12.4.3 applies. Its characteristic polynomial is

$$\begin{aligned} x^2 - (\operatorname{tr}\mathbf{A})x + \det\mathbf{A} &= x^2 - 5x + (6 - (1-i)(1+i)) \\ &= x^2 - 5x + 4, \end{aligned}$$

since $(1-i)(1+i) = 1 - i^2 = 2$. So the eigenvalues are 4 and 1, both real as Theorem 12.2.1 promises.

For $\lambda = 4$, the first row of $\mathbf{A} - 4\mathbf{I} = \begin{pmatrix} -2 & 1-i \\ 1+i & -1 \end{pmatrix}$ gives $-2z_1 + (1-i)z_2 = 0$; taking $z_2 = 2$ gives $\mathbf{u} = (1-i,\, 2)$, with $\|\mathbf{u}\|^2 = 2 + 4 = 6$.

For $\lambda = 1$, the first row of $\mathbf{A} - \mathbf{I} = \begin{pmatrix} 1 & 1-i \\ 1+i & 2 \end{pmatrix}$ gives $z_1 = -(1-i)z_2$; taking $z_2 = 1$ gives $\mathbf{w} = (-1+i,\, 1)$, with $\|\mathbf{w}\|^2 = 2 + 1 = 3$.

The two are orthogonal, as Corollary 12.3.4 predicts:

$$\begin{aligned} \langle \mathbf{u}, \mathbf{w} \rangle &= (1-i)\overline{(-1+i)} + 2 \cdot \overline{1} \\ &= (1-i)(-1-i) + 2 = -2 + 2 = 0. \end{aligned}$$

Normalizing and using them as columns,

$$\mathbf{U} = \begin{pmatrix} \frac{1-i}{\sqrt{6}} & \frac{-1+i}{\sqrt{3}} \\ \frac{2}{\sqrt{6}} & \frac{1}{\sqrt{3}} \end{pmatrix}, \qquad \mathbf{D} = \begin{pmatrix} 4 & 0 \\ 0 & 1 \end{pmatrix},$$

and $\mathbf{U}^*\mathbf{U} = \mathbf{I}$, $\mathbf{U}^*\mathbf{A}\mathbf{U} = \mathbf{D}$. As a check on the arithmetic, $\operatorname{tr}\mathbf{D} = 5 = \operatorname{tr}\mathbf{A}$ and $\det\mathbf{D} = 4 = \det\mathbf{A}$.

12.4.3 Reading the class off the spectrum

Once the eigenvalues sit on a diagonal, every question about T that is invariant under a unitary change of basis becomes a question about a list of complex numbers. Here are three answers.

Corollary 12.4.4. The Spectrum Decides

Let $T \in \mathcal{L}(V)$ be normal on a finite-dimensional complex inner product space.

(a) T is self-adjoint if and only if every eigenvalue of T is real.

(b) T is unitary if and only if every eigenvalue of T has modulus 1.

(c) $\langle Tv, v \rangle$ is a nonnegative real number for every $v \in V$ if and only if every eigenvalue of

T is a nonnegative real.

Idea. By Theorem 12.4.2 there is an orthonormal basis $(e_1, \dots, e_n)$ with $Te_j = \lambda_j e_j$, and then $T^*e_j = \overline{\lambda_j} e_j$ by Theorem 12.3.3. Two operators that agree on a basis are equal, so each of the three conditions can be tested one basis vector at a time, where it becomes a statement about the single number λ_j.

Proof. Fix an orthonormal basis $(e_1, \dots, e_n)$ of eigenvectors, $Te_j = \lambda_j e_j$ (Theorem 12.4.2), and recall $T^*e_j = \overline{\lambda_j} e_j$ (Theorem 12.3.3). Every eigenvalue of T occurs among $\lambda_1, \dots, \lambda_n$, because $\operatorname{diag}(\lambda_1, \dots, \lambda_n)$ is the matrix of T and its characteristic polynomial is $\prod_j (x - \lambda_j)$.

(a) $T = T^*$ holds if and only if $Te_j = T^*e_j$ for every j (Theorem 2.1.5), that is $\lambda_j e_j = \overline{\lambda_j} e_j$, that is $\lambda_j = \overline{\lambda_j}$ since $e_j \neq 0$. A complex number equals its conjugate exactly when it is real.

(b) $T^*Te_j = T^*(\lambda_j e_j) = \lambda_j \overline{\lambda_j} e_j = |\lambda_j|^2 e_j$, so $T^*T = \mathrm{id}_V$ if and only if $|\lambda_j|^2 = 1$ for every j. That is what it means for T to be unitary (Definition 11.7.2): the equation $T^*T = \mathrm{id}_V$ makes T injective, hence invertible on a finite-dimensional space (Theorem 2.5.7), with $T^{-1} = T^*$ and therefore $TT^* = \mathrm{id}_V$ as well.

(c) Write $v = \sum_j c_j e_j$. Expanding both slots and using $\langle e_j, e_i \rangle = \delta_{ij}$,

$$\langle Tv, v \rangle = \Big\langle \sum_j c_j \lambda_j e_j, \sum_i c_i e_i \Big\rangle = \sum_j \lambda_j |c_j|^2.$$

($\Leftarrow$) If every λ_j is a nonnegative real, this is a sum of nonnegative reals. ($\Rightarrow$) Taking $v = e_j$ gives $\langle Te_j, e_j \rangle = \lambda_j$, which is then a nonnegative real by hypothesis.

■

Part (c) names a class that Chapter 13 takes as its subject: the **positive** operators. Over $\mathbb{C}$ the hypothesis of normality is not really needed in (c), because $\langle Tv, v \rangle$ real for every v already forces $T = T^*$: apply Theorem 12.2.6 to the operator $T - T^*$, for which $\langle (T - T^*)v, v \rangle = \langle Tv, v \rangle - \overline{\langle Tv, v \rangle} = 0$. Over $\mathbb{R}$ that shortcut is unavailable, and Section 5 shows the quarter turn defeating it.

The next corollary is a small shock. The adjoint is defined through the inner product and has nothing to do with polynomials. For a normal operator it is one anyway.

Corollary 12.4.5. The Adjoint of a Normal Operator Is a Polynomial in It

Let $T \in \mathcal{L}(V)$ be normal on a finite-dimensional complex inner product space, with distinct eigenvalues $\lambda_1, \dots, \lambda_k$. Then there is a polynomial $p \in \mathbb{C}[x]$ of degree less than k with $T^* = p(T)$.

Proof. The scalars $\lambda_1, \dots, \lambda_k$ are distinct, so by Lagrange interpolation (Theorem 6.6.1) there is a polynomial $p \in \mathbb{C}[x]$ of degree less than k with

$$p(\lambda_i) = \overline{\lambda_i} \qquad (i = 1, \dots, k).$$

Let $(e_1, \dots, e_n)$ be an orthonormal basis of eigenvectors (Theorem 12.4.2), with $Te_j = \mu_j e_j$ where each μ_j is one of the λ_i. Then $p(T)e_j = p(\mu_j)e_j = \overline{\mu_j}e_j = T^*e_j$, the first equality because $T^i e_j = \mu_j^i e_j$ for every $i \geq 0$, by induction from $Te_j = \mu_j e_j$, and the last by Theorem 12.3.3. The operators $p(T)$ and T^* agree on a basis, hence are equal (Theorem 2.1.5).

■

One consequence is worth stating on its own. Any operator S that commutes with T also commutes with every polynomial in T, hence with T^*. So for normal operators the apparently stronger requirement "commutes with T and with T^*" is no requirement at all, and Section 7 will use this to diagonalize whole commuting families at once. The polynomial depends on T and not only on its size: for $T = \mathrm{id}_V$ one may take the constant $p = 1$, while for the matrix $\mathrm{diag}(1, i)$ the conditions $p(1) = 1$ and $p(i) = -i$ are met by

$$p(x) = ix + (1 - i),$$

as one checks by substituting $x = 1$ and $x = i$.

12.4.4 Exactly what is unique

The factorization $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{U}^*$ is not unique, and it is worth being precise about which parts of it are.

Unique: the **set** of eigenvalues $\mathrm{spec}(T)$, and for each $\lambda \in \mathrm{spec}(T)$ the eigenspace $E_\lambda(T)$ with its dimension. These are defined from T alone, without any choice. Consequently the diagonal of $\mathbf{D}$ is unique **as a list with multiplicities up to reordering**: each λ appears exactly $\dim E_\lambda(T)$ times.

Not unique: the order of the diagonal, and the matrix $\mathbf{U}$. Two freedoms create all the others. First, permuting the eigenvectors permutes the columns of $\mathbf{U}$ and the diagonal of $\mathbf{D}$ in step. Second, within a single eigenspace $E_\lambda(T)$ of dimension m, **any** orthonormal basis of that eigenspace will do, and there are many when $m \geq 2$: the choices are related by the unitary maps of $E_\lambda(T)$. For $\mathbf{A} = \mathbf{I}_n$, for instance, every unitary $\mathbf{U}$ satisfies $\mathbf{U}^*\mathbf{A}\mathbf{U} = \mathbf{I}_n$. What no freedom can change is that eigenvectors for distinct eigenvalues stay orthogonal (Corollary 12.3.4), so the decomposition

$$V = E_{\lambda_1}(T) \oplus \cdots \oplus E_{\lambda_k}(T)$$

into mutually orthogonal eigenspaces is itself unique. Section 8 turns that decomposition into the spectral resolution and shows it determines T completely.

Quick check. The matrix $\mathbf{A} = \mathrm{diag}(1, 1, 2)$ is normal. Write down two different unitary matrices $\mathbf{U}$ with $\mathbf{U}^*\mathbf{A}\mathbf{U} = \mathbf{A}$, and explain why, for $\mathrm{diag}(1, 2, 3)$ with the order of the diagonal fixed, the freedom shrinks to a single unit-modulus scalar on each column.

Solution. Take $\mathbf{U}_1 = \mathbf{I}_3$ and $\mathbf{U}_2 = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$; more generally $\mathbf{R} \oplus (1)$ works for **any** $\mathbf{R} \in \mathrm{U}(2)$, because $\mathbf{A} = \mathbf{I}_2 \oplus (2)$ and $\mathbf{R}^*\mathbf{I}_2\mathbf{R} = \mathbf{I}_2$. The freedom comes from the two-dimensional eigenspace $E_1(\mathbf{A}) = \mathrm{span}(\mathbf{e}_1, \mathbf{e}_2)$, inside which every orthonormal basis is as good as any other. For $\mathrm{diag}(1, 2, 3)$ each eigenspace is a line, and a line has exactly two unit vectors over $\mathbb{R}$ and a circle of them over $\mathbb{C}$; the residual freedom is only the scalar factor of

modulus 1 in each column, which changes $\mathbf{U}$ but not the decomposition into eigenspaces.

Warning. **"Diagonalizable" and "unitarily diagonalizable" are different conditions.** Every unitarily diagonalizable matrix is diagonalizable, because $\mathbf{U}$ is in particular invertible. The converse fails, and it fails for matrices as small and as ordinary as $\begin{pmatrix} 1 & 1 \\ 0 & 2 \end{pmatrix}$. Having n distinct eigenvalues guarantees diagonalizability (Corollary 9.4.2) and guarantees nothing about orthogonality.

Example. Diagonalizable, and no orthonormal eigenbasis

For $\mathbf{A} = \begin{pmatrix} 1 & 1 \\ 0 & 2 \end{pmatrix} \in M_2(\mathbb{R})$, find an invertible $\mathbf{P}$ with $\mathbf{P}^{-1}\mathbf{AP}$ diagonal, and show that no **orthogonal** $\mathbf{Q}$ makes $\mathbf{Q}^{\top}\mathbf{AQ}$ diagonal.

Solution. The matrix is upper triangular, so its eigenvalues are the diagonal entries 1 and 2 (Theorem 7.4.6). For $\lambda = 1$, $\mathbf{A} - \mathbf{I} = \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix}$ has kernel spanned by $(1, 0)$. For $\lambda = 2$, $\mathbf{A} - 2\mathbf{I} = \begin{pmatrix} -1 & 1 \\ 0 & 0 \end{pmatrix}$ has kernel spanned by $(1, 1)$. Two distinct eigenvalues in dimension 2 give a basis of eigenvectors (Corollary 9.4.2), so with

$$\mathbf{P} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \quad \text{we get} \quad \mathbf{P}^{-1}\mathbf{AP} = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}.$$

Now suppose $\mathbf{Q}^{\top}\mathbf{AQ} = \mathbf{D}$ were diagonal with $\mathbf{Q}$ orthogonal. Then $\mathbf{A} = \mathbf{QDQ}^{\top}$ would be normal by Corollary 12.4.3. But

$$\mathbf{A}^{\top}\mathbf{A} = \begin{pmatrix} 1 & 1 \\ 1 & 5 \end{pmatrix} \neq \begin{pmatrix} 2 & 2 \\ 2 & 4 \end{pmatrix} = \mathbf{AA}^{\top},$$

so $\mathbf{A}$ is not normal and no such $\mathbf{Q}$ exists. The obstruction is visible without any theory: the eigenvectors are $(1, 0)$ and $(1, 1)$, the eigenspaces are lines and therefore the *only* candidates for eigenvector directions, and $(1, 0) \cdot (1, 1) = 1 \neq 0$. There simply is no orthogonal pair of eigenvectors to build $\mathbf{Q}$ from.

12.4.5 Exercises

12.4.5.1 A. Check your understanding

Exercise. A1

(a) State the complex spectral theorem, including every hypothesis, and state its matrix form.

(b) What goes wrong with the proof if the field is $\mathbb{R}$? Name the exact step.

(c) True or false: a matrix in $M_n(\mathbb{C})$ with n distinct eigenvalues is unitarily diagonalizable. Justify your answer.

(d) Let $\mathbf{A} \in M_3(\mathbb{C})$ be normal with $\operatorname{spec}(\mathbf{A}) = \{2, 2, -1\}$ as a list with multiplicity. Is the unitary $\mathbf{U}$ in $\mathbf{A} = \mathbf{UDU}^*$ determined by $\mathbf{A}$ and the order of $\mathbf{D}$? What is determined?

(e) A normal operator T satisfies $T^* = T^3$. What can you say about its eigenvalues?

(f) Explain why Lemma 12.4.1 cannot be weakened to "a normal matrix with some of its below-diagonal entries zero is diagonal".

Solution.

(a) For a finite-dimensional inner product space V over $\mathbb{C}$ and $T \in \mathcal{L}(V)$: T is normal if and only if V has an orthonormal basis of eigenvectors of T (Theorem 12.4.2). Matrix form: $\mathbf{A} \in M_n(\mathbb{C})$ is normal if and only if $\mathbf{A} = \mathbf{UDU}^*$ with $\mathbf{U}$ unitary and $\mathbf{D}$ diagonal (Corollary 12.4.3).

(b) Schur's theorem. It needs an eigenvalue at every stage of its induction, which over $\mathbb{C}$ comes from Theorem 6.5.1; over $\mathbb{R}$ an operator may have none at all, and then there is no triangular form to start from.

(c) False. $\begin{pmatrix} 1 & 1 \\ 0 & 2 \end{pmatrix}$ has the distinct eigenvalues 1 and 2 and is not normal, hence is not unitarily diagonalizable (Example (Diagonalizable, and no orthonormal eigenbasis)).

(d) No. The eigenspace $E_2(\mathbf{A})$ is two-dimensional, and replacing its orthonormal basis by any other one changes $\mathbf{U}$ while leaving $\mathbf{D}$ alone. What is determined: $\operatorname{spec}(\mathbf{A}) = \{2, -1\}$, the eigenspaces $E_2(\mathbf{A})$ and $E_{-1}(\mathbf{A})$, their dimensions 2 and 1, and their mutual orthogonality.

(e) By Theorem 12.3.3, $Tv = \lambda v$ gives $T^*v = \overline{\lambda}v$, while $T^* = T^3$ gives $T^*v = \lambda^3 v$. Hence $\overline{\lambda} = \lambda^3$. Taking moduli, $|\lambda| = |\lambda|^3$, so $|\lambda|$ is 0 or 1. If $\lambda \neq 0$, multiplying $\overline{\lambda} = \lambda^3$ by λ gives $|\lambda|^2 = \lambda^4$, that is $\lambda^4 = 1$. So every eigenvalue lies in $\{0, 1, -1, i, -i\}$.

(f) The lemma uses triangularity at **every** index, not at a few places: the comparison of row and column lengths at step j is informative only because the whole of column j below the diagonal is already known to vanish. Zeros in some of those positions buy nothing. The circulant $\begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{pmatrix}$ is normal and has zeros in positions $(2, 1)$ and $(3, 2)$, two of the three places below its diagonal, and it is not diagonal.

12.4.5.2 B. Practice

Exercise. B1: Unitarily diagonalize

For each matrix, find a unitary $\mathbf{U}$ and a diagonal $\mathbf{D}$ with $\mathbf{U}^*\mathbf{AU} = \mathbf{D}$.

(a) $\mathbf{A} = \begin{pmatrix} 3 & -i \\ i & 3 \end{pmatrix}$.

(b) $\mathbf{A} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & i \\ i & 1 \end{pmatrix}$.

Solution.

(a) $\mathbf{A}^* = \mathbf{A}$, so $\mathbf{A}$ is Hermitian and normal. Its characteristic polynomial is $x^2 - 6x + (9 - (-i)(i)) = x^2 - 6x + 8$, with roots 4 and 2.

For $\lambda = 4$: $\mathbf{A} - 4\mathbf{I} = \begin{pmatrix} -1 & -i \\ i & -1 \end{pmatrix}$, and the first row gives $z_1 = -iz_2$; take $\mathbf{u} = (-i, 1)$, $\|\mathbf{u}\|^2 = 2$. For $\lambda = 2$: $\mathbf{A} - 2\mathbf{I} = \begin{pmatrix} 1 & -i \\ i & 1 \end{pmatrix}$, and the first row gives $z_1 = iz_2$; take $\mathbf{w} = (i, 1)$, $\|\mathbf{w}\|^2 = 2$. They are orthogonal: $\langle \mathbf{u}, \mathbf{w} \rangle = (-i)\bar{i} + 1 = (-i)(-i) + 1 = -1 + 1 = 0$. Hence

$$\mathbf{U} = \tfrac{1}{\sqrt{2}} \begin{pmatrix} -i & i \\ 1 & 1 \end{pmatrix}, \qquad \mathbf{D} = \begin{pmatrix} 4 & 0 \\ 0 & 2 \end{pmatrix}.$$

(b) $\mathbf{A}$ is unitary (Example (Which of these are normal?) (c)), hence normal. Here $\operatorname{tr} \mathbf{A} = \sqrt{2}$ and $\det \mathbf{A} = \tfrac{1}{2}(1 - i^2) = 1$, so the characteristic polynomial is $x^2 - \sqrt{2}x + 1$, with roots

$$x = \frac{\sqrt{2} \pm \sqrt{2 - 4}}{2} = \frac{1 \pm i}{\sqrt{2}},$$

both of modulus 1, as Corollary 12.4.4 (b) requires. For $\lambda = (1 + i)/\sqrt{2}$, the equation $\mathbf{Az} = \lambda \mathbf{z}$ reads $\tfrac{1}{\sqrt{2}}(z_1 + iz_2) = \tfrac{1+i}{\sqrt{2}} z_1$, that is $iz_2 = iz_1$, so $\mathbf{u} = (1, 1)$. For $\lambda = (1 - i)/\sqrt{2}$ the same line gives $iz_2 = -iz_1$, so $\mathbf{w} = (1, -1)$. These are orthogonal, so

$$\mathbf{U} = \tfrac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}, \qquad \mathbf{D} = \tfrac{1}{\sqrt{2}} \begin{pmatrix} 1 + i & 0 \\ 0 & 1 - i \end{pmatrix}.$$

Here $\mathbf{U}$ is real orthogonal even though $\mathbf{A}$ is not real.

Exercise. B2: A circulant, diagonalized

Let $\omega = e^{2\pi i/3}$, so $\omega^3 = 1$ and $1 + \omega + \omega^2 = 0$, and let

$$\mathbf{C} = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{pmatrix} \in M_3(\mathbb{C}).$$

(a) Verify that $\mathbf{C}$ is normal.

(b) Show that $\mathbf{f}_j = \tfrac{1}{\sqrt{3}}(1, \omega^j, \omega^{2j})$ is an eigenvector of $\mathbf{C}$ for $j = 0, 1, 2$, and find the eigenvalue.

(c) Deduce a unitary $\mathbf{F}$ with $\mathbf{F}^*\mathbf{CF}$ diagonal, and write down the diagonal explicitly.

Solution.

(a) Both products are the same matrix:

$$\mathbf{C}^*\mathbf{C} = \mathbf{CC}^* = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix}.$$

(The entries are dot products of pairs of rows, respectively columns, of $\mathbf{C}$; cyclic shifting makes the two lists agree.)

(b) Write $\mathbf{f} = (1, \omega^j, \omega^{2j})$ before normalizing. Then

$$\mathbf{C}\mathbf{f} = (1+\omega^j,\ \omega^j+\omega^{2j},\ 1+\omega^{2j}).$$

Using $\omega^{3j}=1$, the second entry is $\omega^j(1+\omega^j)$ and the third is $\omega^{2j}(\omega^{-2j}+1) = \omega^{2j}(\omega^j+1)$, since $\omega^{-2j}=\omega^j$. So $\mathbf{C}\mathbf{f} = (1+\omega^j)\mathbf{f}$, and the eigenvalue is $1+\omega^j$.

(c) The three vectors $\mathbf{f}_0, \mathbf{f}_1, \mathbf{f}_2$ are orthonormal: each has norm $\frac{1}{\sqrt{3}}\sqrt{3}=1$, and for $j \neq l$,

$$\langle \mathbf{f}_j, \mathbf{f}_l \rangle = \tfrac{1}{3}\sum_{m=0}^{2} \omega^{mj}\overline{\omega^{ml}} = \tfrac{1}{3}\sum_{m=0}^{2} (\omega^{j-l})^m = 0,$$

because $\omega^{j-l} \neq 1$ is a cube root of unity and $1+z+z^2=0$ for such z. Taking $\mathbf{F}$ with columns $\mathbf{f}_0, \mathbf{f}_1, \mathbf{f}_2$,

$$\begin{aligned}\mathbf{F}^*\mathbf{C}\mathbf{F} &= \operatorname{diag}(2,\ 1+\omega,\ 1+\omega^2)\\ &= \operatorname{diag}\Big(2,\ \tfrac{1}{2}+\tfrac{\sqrt{3}}{2}i,\ \tfrac{1}{2}-\tfrac{\sqrt{3}}{2}i\Big),\end{aligned}$$

using $\omega = -\frac{1}{2}+\frac{\sqrt{3}}{2}i$. Section 9 shows that this one matrix $\mathbf{F}$ diagonalizes **every** 3×3 circulant.

Exercise. B3: Decide unitary diagonalizability

Determine which of the following matrices are unitarily diagonalizable over $\mathbb{C}$, and which are merely diagonalizable. Justify your answer.

(a) $\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$.

(b) $\begin{pmatrix} 2 & 3 \\ 0 & 2 \end{pmatrix}$.

(c) $\begin{pmatrix} 1 & 2 \\ 0 & 3 \end{pmatrix}$.

(d) $\begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$.

Solution.

(a) Unitarily diagonalizable: the matrix is skew-symmetric, hence skew-adjoint over $\mathbb{C}$ and normal, so Corollary 12.4.3 applies. (Its eigenvalues are $\pm i$.)

(b) Neither. Comparing $(1,1)$ entries, $(\mathbf{A}^*\mathbf{A})_{11} = 4$ and $(\mathbf{A}\mathbf{A}^*)_{11} = 4+9 = 13$, so $\mathbf{A}$ is not normal. It is not even diagonalizable: the only eigenvalue is 2 and $\mathbf{A}-2\mathbf{I} = 3\mathbf{J}_2(0) \neq \mathbf{O}$ has rank 1, so $\dim E_2(\mathbf{A}) = 1 < 2$ (Theorem 9.4.1).

(c) Diagonalizable but not unitarily: the eigenvalues 1 and 3 are distinct (Corollary 9.4.2), while $(\mathbf{A}^*\mathbf{A})_{11} = 1 \neq 5 = (\mathbf{A}\mathbf{A}^*)_{11}$, so $\mathbf{A}$ is not normal.

(d) Unitarily diagonalizable: the matrix is a permutation matrix, so its columns are or-

thonormal and it is unitary, hence normal. (Its eigenvalues are the cube roots of unity.)

12.4.5.3 C. Going deeper

Exercise. C1: Operators of finite order

Let $T \in \mathcal{L}(V)$ be normal on a finite-dimensional complex inner product space, and suppose $T^k = \mathrm{id}_V$ for some integer $k \geq 1$.

(a) Prove that every eigenvalue of T is a k-th root of unity.

(b) Deduce that T is unitary and unitarily diagonalizable.

(c) Show that the hypothesis "normal" cannot be dropped from (b) by exhibiting a T with $T^2 = \mathrm{id}_V$ that is not unitary.

Solution.

(a) If $Tv = \lambda v$ with $v \neq 0$, then $T^k v = \lambda^k v$ (Exercise (B3: Eigenvalues of powers and inverses) (a)) and $T^k v = v$, so $(\lambda^k - 1)v = 0$ and $\lambda^k = 1$.

(b) Every k-th root of unity has modulus 1, so T is unitary by Corollary 12.4.4 (b), and it is unitarily diagonalizable by Theorem 12.4.2 because it is normal.

(c) On $\mathbb{C}^2$ take $\mathbf{A} = \begin{pmatrix} 1 & 1 \\ 0 & -1 \end{pmatrix}$. Then $\mathbf{A}^2 = \mathbf{I}$, so $T_{\mathbf{A}}$ has order 2; but its second column $(1, -1)$ has norm $\sqrt{2} \neq 1$, so $\mathbf{A}$ is not unitary (Theorem 11.7.3 (f)). It is still diagonalizable, with eigenvalues 1 and -1; what fails is only the orthogonality of the eigenvectors.

Exercise. C2: Same spectrum, same multiplicities

Let $\mathbf{A}, \mathbf{B} \in M_n(\mathbb{C})$ be normal, and suppose they have the same eigenvalues with the same multiplicities, that is, the same characteristic polynomial.

(a) Prove that there is a unitary $\mathbf{W}$ with $\mathbf{B} = \mathbf{W}\mathbf{A}\mathbf{W}^*$ (one says $\mathbf{A}$ and $\mathbf{B}$ are **unitarily similar**).

(b) Give two matrices in $M_2(\mathbb{C})$ with the same characteristic polynomial that are not unitarily similar, and say which hypothesis of (a) they violate.

(c) Deduce that two normal matrices are unitarily similar if and only if they are similar.

Solution.

(a) By Corollary 12.4.3 write $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{U}^*$ and $\mathbf{B} = \mathbf{V}\mathbf{D}'\mathbf{V}^*$ with $\mathbf{U}, \mathbf{V}$ unitary and $\mathbf{D}, \mathbf{D}'$ diagonal. The diagonals of $\mathbf{D}$ and $\mathbf{D}'$ are the eigenvalue lists of $\mathbf{A}$ and $\mathbf{B}$ with multiplicity, which agree as multisets; and the order on the diagonal may be prescribed, so we may choose the eigenvector orderings to make $\mathbf{D}' = \mathbf{D}$. Then

$$\mathbf{B} = \mathbf{V}\mathbf{D}\mathbf{V}^* = \mathbf{V}\mathbf{U}^*\mathbf{A}\mathbf{U}\mathbf{V}^* = \mathbf{W}\mathbf{A}\mathbf{W}^* \quad \text{with} \quad \mathbf{W} = \mathbf{V}\mathbf{U}^*,$$

and $\mathbf{W}$ is unitary because $\mathrm{U}(n)$ is closed under products and inverses (Proposi-

tion 11.7.4).

(b) $\mathbf{A} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$ and $\mathbf{B} = \mathbf{O}$ both have characteristic polynomial x^2; they are not even similar, since similar matrices have equal rank. A pair that is similar but not unitarily similar: $\begin{pmatrix} 1 & 1 \\ 0 & 2 \end{pmatrix}$ and $\operatorname{diag}(1, 2)$. In both cases the violated hypothesis is normality of one of the two matrices.

(c) ($\Rightarrow$) A unitary similarity is a similarity. ($\Leftarrow$) Similar matrices have the same characteristic polynomial (Theorem 7.10.3), so if both are normal, (a) upgrades the similarity to a unitary one.

Exercise. C3: Normality cannot be dropped

Corollary 12.4.4 reads the class of a **normal** operator off its spectrum. This exercise shows that each part fails without that hypothesis.

(a) Give $\mathbf{A} \in M_2(\mathbb{C})$ with both eigenvalues real that is not Hermitian.

(b) Give $\mathbf{A} \in M_2(\mathbb{C})$ with both eigenvalues of modulus 1 that is not unitary.

(c) Over $\mathbb{R}$, give $\mathbf{A} \in M_2(\mathbb{R})$ with $\langle \mathbf{Ax}, \mathbf{x} \rangle \geq 0$ for every $\mathbf{x} \in \mathbb{R}^2$ whose eigenvalues are not nonnegative reals. Why does no such example exist over $\mathbb{C}$?

Solution.

(a) $\mathbf{A} = \begin{pmatrix} 1 & 1 \\ 0 & 2 \end{pmatrix}$ has eigenvalues 1 and 2 and is not Hermitian, since $\mathbf{A}^* = \begin{pmatrix} 1 & 0 \\ 1 & 2 \end{pmatrix} \neq \mathbf{A}$. It is not normal (Example (Diagonalizable, and no orthonormal eigenbasis)).

(b) $\mathbf{A} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ has 1 as its only eigenvalue, of modulus 1, and is not unitary: its second column has norm $\sqrt{2}$. Again it is not normal, the $(1,1)$ entries of $\mathbf{A}^*\mathbf{A}$ and $\mathbf{AA}^*$ being 1 and 2.

(c) Take the quarter turn $\mathbf{A} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$. For every $\mathbf{x} = (x_1, x_2) \in \mathbb{R}^2$,

$$\langle \mathbf{Ax}, \mathbf{x} \rangle = -x_2 x_1 + x_1 x_2 = 0 \geq 0,$$

yet $\mathbf{A}$ has no real eigenvalue at all; over $\mathbb{C}$ its eigenvalues are $\pm i$, which are not nonnegative reals. Over $\mathbb{C}$ no such example exists: if $\langle Tv, v \rangle = 0$ for every v then $T = 0$ by Theorem 12.2.6, and more generally $\langle Tv, v \rangle$ real for every v forces T to be self-adjoint, hence normal, by applying Theorem 12.2.6 to $T - T^*$. Real scalars are not enough to detect an operator over $\mathbb{R}$, and they are over $\mathbb{C}$.

12.5 The Real Spectral Theorem

Over $\mathbb{C}$ the answer was clean: an orthonormal basis of eigenvectors exists exactly for the normal operators. Over $\mathbb{R}$ the same question has a different answer, and a smaller one. The class shrinks from normal to self-adjoint, and the reason is a single example that we have met since

Chapter 9 and can finally name as the obstruction: a rotation of the plane is as normal as an operator can be and has no eigenvector at all. This section proves the real theorem, examines that obstruction, and then rescues what can be rescued for real normal operators. It ends with the application that made the theorem famous outside linear algebra: the axes of a conic.

Throughout, V is a finite-dimensional inner product space over $\mathbb{R}$ unless stated otherwise, and "symmetric" means $\mathbf{A}^{\top} = \mathbf{A}$.

12.5.1 One eigenvector, then its orthogonal complement

The proof is the signature move of the chapter, run as an induction. Section 2 did all the work: it produced an eigenvector, and it showed that the orthogonal complement of an invariant subspace is again invariant with a self-adjoint restriction. Those two facts fit together into an induction on dimension with nothing left over.

Theorem 12.5.1. Real Spectral Theorem

Let V be a finite-dimensional inner product space over $\mathbb{R}$ and let $T \in \mathcal{L}(V)$. The following are equivalent.

(a) T is self-adjoint.

(b) V has an orthonormal basis consisting of eigenvectors of T.

Idea. ($\Leftarrow$) In such a basis the matrix is diagonal, and a real diagonal matrix is its own transpose. ($\Rightarrow$) Induct on $\dim V$. Peel off one unit eigenvector e_1, which exists by Theorem 12.2.2; the line $U = \operatorname{span}(e_1)$ is invariant, so Theorem 12.2.4 makes $U^{\perp}$ invariant with a self-adjoint restriction. That restriction lives on a space of dimension $\dim V - 1$, so the inductive hypothesis applies to it, and the basis it returns is already orthogonal to e_1.

Proof. ($\Leftarrow$) Let $\mathcal{B}$ be an orthonormal basis of eigenvectors, so $\mathbf{D} := [T]_{\mathcal{B}}^{\mathcal{B}}$ is diagonal with real entries. Since $\mathcal{B}$ is orthonormal, $[T^*]_{\mathcal{B}}^{\mathcal{B}} = \mathbf{D}^* = \mathbf{D}^{\top} = \mathbf{D}$ by Theorem 11.6.3, the last equality because a diagonal matrix equals its transpose. Two operators with the same matrix in one basis are equal (Theorem 4.1.4), so $T^* = T$.

($\Rightarrow$) Induct on $n = \dim V$. If $n = 0$ the empty list is an orthonormal basis and there is nothing to prove. Let $n \geq 1$ and suppose the statement holds for every real inner product space of dimension $n - 1$.

Since $V \neq \{0\}$ and T is self-adjoint, T has an eigenvalue $\lambda \in \mathbb{R}$ by Theorem 12.2.2, say $Tv = \lambda v$ with $v \neq 0$. Put $e_1 = v/\|v\|$, a unit eigenvector, and $U = \operatorname{span}(e_1)$. Then U is T-invariant, because $T(ce_1) = c\lambda e_1 \in U$.

By Theorem 12.2.4, $U^{\perp}$ is T-invariant and the restriction $T|_{U^{\perp}}$ is a self-adjoint operator on $U^{\perp}$, which is an inner product space in its own right under the restricted inner product. Its dimension is $n - 1$ by Theorem 11.3.2 (c). By the inductive hypothesis, $U^{\perp}$ has an orthonormal basis $(e_2, \ldots, e_n)$ of eigenvectors of $T|_{U^{\perp}}$; each e_j satisfies $Te_j = \mu_j e_j$ in V as well, since the restriction is the same map.

Finally $(e_1, \dots, e_n)$ is orthonormal: the last $n-1$ vectors are orthonormal among themselves and each lies in $U^{\perp}$, hence is orthogonal to e_1, which has norm 1. An orthonormal list is linearly independent (Theorem 11.1.4), and it has $n = \dim V$ members, so it is a basis (Theorem 1.6.4). Every member is an eigenvector of T. This proves the theorem.

■

Remark. No complexification appears anywhere in that induction, which is worth noticing: the argument is entirely real. The one place where complex numbers were needed is inside Theorem 12.2.2, whose proof complexifies once to borrow an eigenvalue from the Fundamental Theorem of Algebra (Theorem 6.5.1) and then observes that the eigenvalue is real. So $\mathbb{C}$ is used exactly once, to start the machine, and never again.

Corollary 12.5.2. Orthogonal Diagonalization

Let $\mathbf{A} \in M_n(\mathbb{R})$. Then $\mathbf{A}$ is symmetric if and only if there are an orthogonal $\mathbf{Q} \in \mathrm{O}(n)$ and a **real** diagonal $\mathbf{D}$ with

$$\mathbf{A} = \mathbf{Q}\mathbf{D}\mathbf{Q}^{\top}.$$

The diagonal of $\mathbf{D}$ is the eigenvalue list of $\mathbf{A}$ with multiplicity, in any prescribed order, and the columns of $\mathbf{Q}$ are a corresponding orthonormal basis of eigenvectors.

Proof. Apply Theorem 12.5.1 to $T_{\mathbf{A}}$ on $\mathbb{R}^n$ with the standard inner product, for which the standard basis is orthonormal, so that $\mathbf{A}^{\top} = \mathbf{A}$ is exactly self-adjointness (Theorem 11.6.3).

($\Rightarrow$) Let $(\mathbf{q}_1, \dots, \mathbf{q}_n)$ be an orthonormal basis with $\mathbf{A}\mathbf{q}_j = \lambda_j \mathbf{q}_j$, and let $\mathbf{Q}$ have these columns; it is orthogonal by Theorem 11.7.3 (f). Column j of $\mathbf{A}\mathbf{Q}$ is $\lambda_j \mathbf{q}_j$, which is column j of $\mathbf{Q}\mathbf{D}$ for $\mathbf{D} = \mathrm{diag}(\lambda_1, \dots, \lambda_n)$. So $\mathbf{A}\mathbf{Q} = \mathbf{Q}\mathbf{D}$ and $\mathbf{A} = \mathbf{Q}\mathbf{D}\mathbf{Q}^{-1} = \mathbf{Q}\mathbf{D}\mathbf{Q}^{\top}$. The eigenvalues are real because $\mathbf{A}$ is self-adjoint (Theorem 12.2.1), so $\mathbf{D}$ is a real matrix.

($\Leftarrow$) If $\mathbf{A} = \mathbf{Q}\mathbf{D}\mathbf{Q}^{\top}$ with $\mathbf{D}$ real diagonal, then

$$\mathbf{A}^{\top} = (\mathbf{Q}\mathbf{D}\mathbf{Q}^{\top})^{\top} = \mathbf{Q}\mathbf{D}^{\top}\mathbf{Q}^{\top} = \mathbf{Q}\mathbf{D}\mathbf{Q}^{\top} = \mathbf{A}.$$

The claims about the order and the columns are as in Corollary 12.4.3.

■

12.5.2 The negative result: normal is not enough over the reals

Over $\mathbb{C}$ the class with an orthonormal eigenbasis was the normal operators. Over $\mathbb{R}$ Theorem 12.5.1 says it is the self-adjoint ones, which is a strictly smaller class: every self-adjoint operator is normal, and the converse now fails. It is worth seeing exactly how it fails, because the failure is not a technicality about multiplicities or about a badly chosen basis. It is the complete absence of an eigenvector.

Example. A normal real operator with no eigenvector

Let $T \in \mathcal{L}(\mathbb{R}^2)$ be the quarter turn, with matrix

$$\mathbf{R} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$$

in the standard basis. Show that T is normal, and that T has no eigenvalue and no eigenvector in $\mathbb{R}^2$.

Solution. Normality: $\mathbf{R}^\top \mathbf{R} = \mathbf{R}\mathbf{R}^\top = \mathbf{I}$, so T is even orthogonal, hence normal. (One may also read it off the shape: $\mathbf{R} = \mathbf{\Lambda}(i)$ is the rotation-scaling block of Definition 10.4.1.)

No eigenvalue: $p_{\mathbf{R}}(x) = x^2 + 1$, which has no root in $\mathbb{R}$. So $\mathbf{R} - \lambda\mathbf{I}$ is invertible for every real λ, and $\ker(T - \lambda\,\mathrm{id}) = \{\mathbf{0}\}$ (Theorem 9.2.1). Geometrically, T turns every nonzero vector by a right angle, so $T\mathbf{v}$ is never a real multiple of $\mathbf{v}$; the operator maps no line to itself.

That example rules out every possible improvement of Theorem 12.5.1 to normal operators over $\mathbb{R}$, and not by a narrow margin. An orthonormal basis of eigenvectors needs n eigenvectors; the quarter turn supplies none. So the two spectral theorems are not two versions of one statement with a technical difference: the real one has a genuinely smaller hypothesis, and the only reason it is not smaller still is Theorem 12.2.2, which manufactures the eigenvector that the rotation lacks.

Two further remarks keep the picture honest. First, the quarter turn is not exceptional. Over $\mathbb{R}$, *every* normal operator that is not self-adjoint fails to have an orthonormal eigenbasis, because Theorem 12.5.1 is an equivalence: having such a basis and being self-adjoint are the same condition. Second, nothing is wrong with the operator; what is missing is the field. Regard $\mathbf{R}$ as a complex matrix and it is unitarily diagonalizable, with eigenvalues $\pm i$ and eigenvectors $(1, \mp i)$ in $\mathbb{C}^2$ (the sign of the eigenvalue and of the entry being opposite). The complex spectral theorem never fails; it simply answers a question about $\mathbb{C}^n$.

Quick check. Is $\mathbf{R} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ orthogonally diagonalizable over $\mathbb{R}$? Is it unitarily diagonalizable over $\mathbb{C}$? Are the two answers consistent?

Solution. Not over $\mathbb{R}$: an orthogonal diagonalization $\mathbf{R} = \mathbf{Q}\mathbf{D}\mathbf{Q}^\top$ would make $\mathbf{R}$ symmetric by Corollary 12.5.2, and $\mathbf{R}^\top = -\mathbf{R} \neq \mathbf{R}$. Over $\mathbb{C}$ it is unitarily diagonalizable, because it is normal (Corollary 12.4.3); explicitly $\mathbf{U}^*\mathbf{R}\mathbf{U} = \mathrm{diag}(i, -i)$ for $\mathbf{U} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ -i & i \end{pmatrix}$. There is no conflict: the two statements concern different spaces, $\mathbb{R}^2$ and $\mathbb{C}^2$, and the diagonalizing basis of the second one is not real.

12.5.3 What a real normal operator does look like

The rotation is not a defect to be apologized for; it is a building block. Chapter 10 already promoted it to one when it built the real Jordan form, and the same 2×2 block returns here as the only thing that can obstruct a real diagonalization of a normal operator.

Theorem 12.5.3. Real Normal Form of a Normal Operator

Let V be a finite-dimensional inner product space over $\mathbb{R}$ and let $T \in \mathcal{L}(V)$ be normal. Then V has an orthonormal basis in which the matrix of T is block diagonal,

$$(c_1) \oplus \cdots \oplus (c_p) \oplus \mathbf{\Lambda}(\lambda_1) \oplus \cdots \oplus \mathbf{\Lambda}(\lambda_q),$$

with $c_1, \dots, c_p \in \mathbb{R}$ and, for each j, $\lambda_j = a_j + b_j i$ with $b_j > 0$ and

$$\mathbf{\Lambda}(\lambda_j) = \begin{pmatrix} a_j & -b_j \\ b_j & a_j \end{pmatrix}$$

the rotation-scaling block of Definition 10.4.1. In matrix form: $\mathbf{A} \in M_n(\mathbb{R})$ is normal if and only if $\mathbf{A} = \mathbf{Q}\mathbf{B}\mathbf{Q}^\top$ with $\mathbf{Q} \in \mathrm{O}(n)$ and $\mathbf{B}$ of the displayed shape.

Idea. The real Schur form of Section 1 gets us to a **block** triangular matrix with 1×1 and 2×2 diagonal blocks, and the corner blocks above the diagonal are the only obstruction left. The argument that killed them in the complex case, comparing a diagonal entry of $\mathbf{T}^\top\mathbf{T}$ with the same entry of $\mathbf{T}\mathbf{T}^\top$, works verbatim on blocks once "entry" is replaced by "trace of a block": the identity $\operatorname{tr}(\mathbf{X}\mathbf{X}^\top) = \sum_{i,j} x_{ij}^2$ turns the comparison into a sum of squares that has to vanish. What remains is to see what a 2×2 real normal matrix can be, and there are exactly two answers: symmetric, which splits further, or a rotation-scaling block.

Proof. Fix an orthonormal basis of V and let $\mathbf{A} \in M_n(\mathbb{R})$ be the matrix of T in it; then $\mathbf{A}^\top$ is the matrix of T^* (Theorem 11.6.3), and T normal means $\mathbf{A}^\top\mathbf{A} = \mathbf{A}\mathbf{A}^\top$. Every orthogonal change of orthonormal basis replaces $\mathbf{A}$ by $\mathbf{Q}^\top\mathbf{A}\mathbf{Q}$, which is again normal:

$$(\mathbf{Q}^\top\mathbf{A}\mathbf{Q})^\top(\mathbf{Q}^\top\mathbf{A}\mathbf{Q}) = \mathbf{Q}^\top\mathbf{A}^\top\mathbf{A}\mathbf{Q} = \mathbf{Q}^\top\mathbf{A}\mathbf{A}^\top\mathbf{Q},$$

and the same computation in the other order gives $(\mathbf{Q}^\top\mathbf{A}\mathbf{Q})(\mathbf{Q}^\top\mathbf{A}\mathbf{Q})^\top$. So we may replace $\mathbf{A}$ by any orthogonally similar matrix at will.

Step 1: a block triangular form. By the real Schur form (Theorem 12.1.7) there is $\mathbf{Q}_1 \in \mathrm{O}(n)$ with $\mathbf{T} = \mathbf{Q}_1^\top\mathbf{A}\mathbf{Q}_1$ block upper triangular, its diagonal blocks of size 1×1 and 2×2. Write the block partition as $\mathbf{T} = (\mathbf{T}_{kl})_{1 \le k,l \le m}$ with $\mathbf{T}_{kl} = \mathbf{O}$ for $k > l$. By the previous paragraph $\mathbf{T}$ is normal.

Step 2: the corners vanish.

Claim. A normal, block upper triangular real matrix is block diagonal.

Proof. Induct on the number m of diagonal blocks; $m = 1$ is trivial. Comparing the $(1,1)$

blocks of the two products (Theorem 8.1.1) and using $\mathbf{T}_{k1} = \mathbf{O}$ for $k > 1$,

$$(\mathbf{T}^\top\mathbf{T})_{11} = \mathbf{T}_{11}^\top\mathbf{T}_{11},$$
$$(\mathbf{T}\mathbf{T}^\top)_{11} = \mathbf{T}_{11}\mathbf{T}_{11}^\top + \sum_{l>1}\mathbf{T}_{1l}\mathbf{T}_{1l}^\top.$$

Take traces. Since $\operatorname{tr}(\mathbf{XY}) = \operatorname{tr}(\mathbf{YX})$ (Theorem 0.9.6), the two terms $\operatorname{tr}(\mathbf{T}_{11}^\top\mathbf{T}_{11})$ and $\operatorname{tr}(\mathbf{T}_{11}\mathbf{T}_{11}^\top)$ are equal and cancel, leaving

$$\sum_{l>1}\operatorname{tr}(\mathbf{T}_{1l}\mathbf{T}_{1l}^\top) = 0.$$

For a real matrix $\mathbf{X}$, $\operatorname{tr}(\mathbf{X}\mathbf{X}^\top) = \sum_{i,j} x_{ij}^2 \geq 0$, with equality only for $\mathbf{X} = \mathbf{O}$. So $\mathbf{T}_{1l} = \mathbf{O}$ for every $l > 1$: the first block row is clear, and $\mathbf{T} = \mathbf{T}_{11} \oplus \mathbf{T}'$ where $\mathbf{T}'$ is the trailing block upper triangular matrix with $m - 1$ diagonal blocks. Then

$$\mathbf{T}^\top\mathbf{T} = \mathbf{T}_{11}^\top\mathbf{T}_{11} \oplus \mathbf{T}'^\top\mathbf{T}', \qquad \mathbf{T}\mathbf{T}^\top = \mathbf{T}_{11}\mathbf{T}_{11}^\top \oplus \mathbf{T}'\mathbf{T}'^\top,$$

so normality of $\mathbf{T}$ gives normality of $\mathbf{T}'$, and the inductive hypothesis applies.

■

So after Step 2 the matrix is $\mathbf{D}_1 \oplus \cdots \oplus \mathbf{D}_m$ with each $\mathbf{D}_k$ of size 1 or 2. Both products are now computed blockwise, so normality of the whole reads $\mathbf{D}_k^\top\mathbf{D}_k = \mathbf{D}_k\mathbf{D}_k^\top$ for each k: every diagonal block is normal.

Step 3: the shape of a 2×2 real normal block. Let $\mathbf{D} = \begin{pmatrix} p & q \\ r & s \end{pmatrix}$ be real and normal. Expanding both products,

$$\mathbf{D}\mathbf{D}^\top - \mathbf{D}^\top\mathbf{D} = \begin{pmatrix} q^2 - r^2 & (r-q)(p-s) \\ (r-q)(p-s) & r^2 - q^2 \end{pmatrix},$$

so normality says $q^2 = r^2$ and $(r - q)(p - s) = 0$. The first equation gives $q = r$ or $q = -r$, and when $q = -r = 0$ both hold, so the following two cases are exhaustive.

Case 1: $q = r$. Then $\mathbf{D}$ is symmetric, and Corollary 12.5.2 supplies an orthogonal 2×2 matrix turning it into a real diagonal matrix, which is two 1×1 blocks.

Case 2: $q = -r$ and $q \neq 0$. Then $r - q = -2q \neq 0$, so $p = s$ and

$$\mathbf{D} = \begin{pmatrix} p & q \\ -q & p \end{pmatrix}.$$

If $q < 0$, this is $\mathbf{\Lambda}(p - qi)$ with $-q > 0$, as required. If $q > 0$, conjugate by the orthogonal matrix $\operatorname{diag}(1, -1)$:

$$\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}\begin{pmatrix} p & q \\ -q & p \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} = \begin{pmatrix} p & -q \\ q & p \end{pmatrix},$$

which is $\mathbf{\Lambda}(p + qi)$ with $q > 0$.

Step 4: assembling. Apply Step 3 inside each diagonal block, using the direct sum $\mathbf{Q}_2$ of the resulting 1×1 and 2×2 orthogonal matrices, which is orthogonal because its columns are orthonormal. Then $\mathbf{Q} = \mathbf{Q}_1\mathbf{Q}_2$ is orthogonal and $\mathbf{Q}^\top\mathbf{A}\mathbf{Q}$ has the stated shape; reordering

the blocks by a permutation matrix, itself orthogonal, puts the 1×1 blocks first. The columns of $\mathbf{Q}$ are the required orthonormal basis.

For the converse, a block diagonal matrix $\mathbf{B}$ of the displayed shape is normal, since $\mathbf{B}^\top\mathbf{B}$ and $\mathbf{B}\mathbf{B}^\top$ are computed blockwise and every block is normal: $(c)^\top(c) = (c^2) = (c)(c)^\top$, and $\mathbf{\Lambda}(\lambda)^\top\mathbf{\Lambda}(\lambda) = |\lambda|^2\mathbf{I}_2 = \mathbf{\Lambda}(\lambda)\mathbf{\Lambda}(\lambda)^\top$. Then $\mathbf{A} = \mathbf{Q}\mathbf{B}\mathbf{Q}^\top$ is normal by the first paragraph. This proves the theorem.

■

The two theorems now fit together. A real normal operator is self-adjoint exactly when no $\mathbf{\Lambda}$ block occurs, and then Theorem 12.5.3 is Theorem 12.5.1. The eigenvalues of $\mathbf{A}$ are the numbers c_i together with the conjugate pairs $\lambda_j, \overline{\lambda_j}$, so the $\mathbf{\Lambda}$ blocks are precisely the nonreal part of the spectrum, packaged two real dimensions at a time. Section 6 specializes this to orthogonal operators, where $|\lambda| = 1$ forces every 2×2 block to be a plane rotation.

12.5.4 A symmetric matrix as a sum of projections

Writing $\mathbf{A} = \mathbf{Q}\mathbf{D}\mathbf{Q}^\top$ and expanding the product column by column turns the diagonalization into a sum of very simple matrices.

Corollary 12.5.4. Spectral Decomposition of a Symmetric Matrix

Let $\mathbf{A} \in M_n(\mathbb{R})$ be symmetric, and let $(\mathbf{q}_1, \dots, \mathbf{q}_n)$ be an orthonormal basis of $\mathbb{R}^n$ with $\mathbf{A}\mathbf{q}_i = \lambda_i\mathbf{q}_i$. Then

$$\mathbf{A} = \sum_{i=1}^{n} \lambda_i\, \mathbf{q}_i\mathbf{q}_i^\top,$$

and each $\mathbf{q}_i\mathbf{q}_i^\top$ is the matrix of the orthogonal projection of $\mathbb{R}^n$ onto the line $\operatorname{span}(\mathbf{q}_i)$.

Proof. Let $\mathbf{Q}$ have columns $\mathbf{q}_1, \dots, \mathbf{q}_n$ and $\mathbf{D} = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$, so $\mathbf{A} = \mathbf{Q}\mathbf{D}\mathbf{Q}^\top$ by Corollary 12.5.2. The columns of $\mathbf{Q}\mathbf{D}$ are $\lambda_i\mathbf{q}_i$, and the rows of $\mathbf{Q}^\top$ are $\mathbf{q}_i^\top$, so the column-times-row expansion of the product (Corollary 8.1.3) gives $\mathbf{A} = \sum_i (\lambda_i\mathbf{q}_i)\mathbf{q}_i^\top$, which is the stated sum.

For the second claim, $\mathbf{q}_i$ is a unit vector, so the orthogonal projection onto $\operatorname{span}(\mathbf{q}_i)$ is $\mathbf{x} \mapsto \langle \mathbf{x}, \mathbf{q}_i\rangle\mathbf{q}_i$ by Theorem 11.3.3. Since $\langle \mathbf{x}, \mathbf{q}_i\rangle = \mathbf{q}_i^\top\mathbf{x}$ is a scalar, $\langle \mathbf{x}, \mathbf{q}_i\rangle\mathbf{q}_i = \mathbf{q}_i(\mathbf{q}_i^\top\mathbf{x}) = (\mathbf{q}_i\mathbf{q}_i^\top)\mathbf{x}$.

■

Example. Orthogonally diagonalizing a 3 by 3 matrix

Orthogonally diagonalize

$$\mathbf{A} = \begin{pmatrix} 3 & 1 & 1 \\ 1 & 3 & 1 \\ 1 & 1 & 3 \end{pmatrix},$$

and write it as a sum of rank-one orthogonal projections.

Solution. The rows all sum to 5, so $\mathbf{A}(1,1,1) = 5(1,1,1)$ and 5 is an eigenvalue. Also $\mathbf{A} = 2\mathbf{I} + \mathbf{J}$ where $\mathbf{J}$ is the all-ones matrix, and $\mathbf{Jx} = \mathbf{0}$ exactly when $x_1 + x_2 + x_3 = 0$; on that plane $\mathbf{Ax} = 2\mathbf{x}$. So

$$E_5(\mathbf{A}) = \operatorname{span}\big((1,1,1)\big), \qquad E_2(\mathbf{A}) = \{\mathbf{x} : x_1 + x_2 + x_3 = 0\},$$

of dimensions 1 and 2, which already account for all of $\mathbb{R}^3$. The two eigenspaces are orthogonal, as Theorem 12.2.3 requires.

Inside $E_2(\mathbf{A})$ any orthonormal basis will do. Start from the independent pair $(1,-1,0)$ and $(1,0,-1)$, both in the plane, and run one step of Gram–Schmidt (Theorem 11.2.2): the second vector minus its component along the first is $(1,0,-1) - \frac{1}{2}(1,-1,0) = \frac{1}{2}(1,1,-2)$, a multiple of $(1,1,-2)$. Normalizing all three,

$$\mathbf{q}_1 = \tfrac{1}{\sqrt{3}}(1,1,1), \quad \mathbf{q}_2 = \tfrac{1}{\sqrt{2}}(1,-1,0), \quad \mathbf{q}_3 = \tfrac{1}{\sqrt{6}}(1,1,-2).$$

With $\mathbf{Q} = (\mathbf{q}_1 \mid \mathbf{q}_2 \mid \mathbf{q}_3)$ we get $\mathbf{Q}^\top\mathbf{Q} = \mathbf{I}$ and $\mathbf{Q}^\top\mathbf{AQ} = \operatorname{diag}(5,2,2)$. By Corollary 12.5.4,

$$\mathbf{A} = 5\,\mathbf{q}_1\mathbf{q}_1^\top + 2\,\mathbf{q}_2\mathbf{q}_2^\top + 2\,\mathbf{q}_3\mathbf{q}_3^\top.$$

As a check, $\mathbf{q}_2\mathbf{q}_2^\top + \mathbf{q}_3\mathbf{q}_3^\top = \mathbf{I} - \mathbf{q}_1\mathbf{q}_1^\top$ because the three projections sum to the identity, so the right-hand side is $2\mathbf{I} + 3\,\mathbf{q}_1\mathbf{q}_1^\top = 2\mathbf{I} + \mathbf{J}$, which is $\mathbf{A}$. Notice that $\mathbf{q}_2$ and $\mathbf{q}_3$ were not forced on us, while the two projections $\mathbf{q}_1\mathbf{q}_1^\top$ and $\mathbf{q}_2\mathbf{q}_2^\top + \mathbf{q}_3\mathbf{q}_3^\top$ were: they are the projections onto the two eigenspaces. Section 8 makes that observation into a theorem.

12.5.5 The principal axis theorem

Here is the application. A **quadratic form** on $\mathbb{R}^n$ is a function

$$q(\mathbf{x}) = \mathbf{x}^\top\mathbf{Ax} = \sum_{i,j} a_{ij}x_ix_j$$

for some $\mathbf{A} \in M_n(\mathbb{R})$. Only the symmetric part of $\mathbf{A}$ matters, since $\mathbf{x}^\top\mathbf{Ax} = \mathbf{x}^\top\mathbf{A}^\top\mathbf{x}$ for a 1×1 matrix, so we always take $\mathbf{A}$ symmetric; then $\mathbf{A}$ is determined by q, with a_{ii} the coefficient of x_i^2 and a_{ij} **half** the coefficient of x_ix_j for $i \neq j$. Chapter 14 studies such forms in their own right.

Theorem 12.5.5. Principal Axis Theorem

Let $\mathbf{A} \in M_n(\mathbb{R})$ be symmetric, let $q(\mathbf{x}) = \mathbf{x}^\top\mathbf{Ax}$, and let $\mathbf{Q} \in \mathrm{O}(n)$ with $\mathbf{Q}^\top\mathbf{AQ} = \mathbf{D} = \operatorname{diag}(\lambda_1,\dots,\lambda_n)$, the columns $\mathbf{q}_1,\dots,\mathbf{q}_n$ of $\mathbf{Q}$ being an orthonormal basis of eigenvectors. Substituting $\mathbf{x} = \mathbf{Qy}$, that is, taking $y_i = \langle \mathbf{x}, \mathbf{q}_i\rangle$ to be the coordinates of $\mathbf{x}$ in that basis,

$$q(\mathbf{x}) = \sum_{i=1}^{n} \lambda_i y_i^2.$$

In particular, for $n = 2$ with $\lambda_1, \lambda_2 > 0$, the level set $\{q = 1\}$ is an ellipse whose axes lie along $\mathbf{q}_1$ and $\mathbf{q}_2$, with half-lengths $1/\sqrt{\lambda_1}$ and $1/\sqrt{\lambda_2}$.

Proof. Since $\mathbf{Q}$ is invertible, $\mathbf{x} = \mathbf{Qy}$ sets up a bijection between $\mathbf{x}$ and $\mathbf{y} = \mathbf{Q}^{-1}\mathbf{x} = \mathbf{Q}^\top\mathbf{x}$,

whose i-th entry is $\mathbf{q}_i^{\top}\mathbf{x} = \langle \mathbf{x}, \mathbf{q}_i \rangle$. Substituting,

$$q(\mathbf{Q}\mathbf{y}) = (\mathbf{Q}\mathbf{y})^{\top}\mathbf{A}(\mathbf{Q}\mathbf{y}) = \mathbf{y}^{\top}(\mathbf{Q}^{\top}\mathbf{A}\mathbf{Q})\mathbf{y} = \mathbf{y}^{\top}\mathbf{D}\mathbf{y} = \sum_i \lambda_i y_i^2,$$

using $(\mathbf{Q}\mathbf{y})^{\top} = \mathbf{y}^{\top}\mathbf{Q}^{\top}$ (Theorem 0.9.5).

For $n = 2$ with both $\lambda_i > 0$, the condition $q(\mathbf{x}) = 1$ becomes $\lambda_1 y_1^2 + \lambda_2 y_2^2 = 1$, the standard equation of an ellipse in the $\mathbf{y}$ coordinates, meeting the y_1-axis where $y_1 = \pm 1/\sqrt{\lambda_1}$ and the y_2-axis where $y_2 = \pm 1/\sqrt{\lambda_2}$. Those axes are the lines $\operatorname{span}(\mathbf{q}_1)$ and $\operatorname{span}(\mathbf{q}_2)$ of the original picture, because $\mathbf{y} = \mathbf{e}_1$ corresponds to $\mathbf{x} = \mathbf{Q}\mathbf{e}_1 = \mathbf{q}_1$. Since $\mathbf{Q}$ is orthogonal, it preserves lengths (Theorem 11.7.3), so the half-lengths are unchanged by the substitution.

■

Example. Finding the axes of a conic

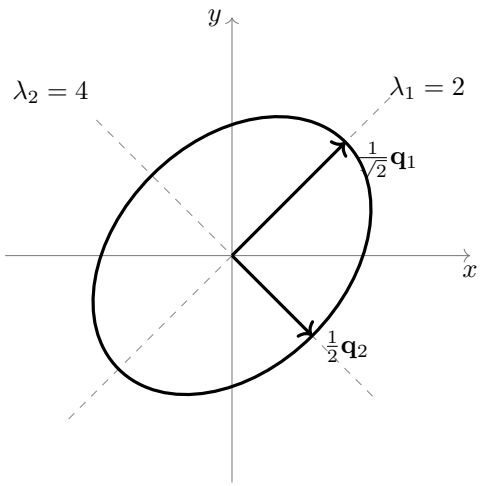
Identify the curve $3x^2 - 2xy + 3y^2 = 1$ in $\mathbb{R}^2$, and find its axes and their half-lengths.

Solution. The symmetric matrix of the form is

$$\mathbf{A} = \begin{pmatrix} 3 & -1 \\ -1 & 3 \end{pmatrix},$$

the off-diagonal entries being half of -2. Its characteristic polynomial is $x^2 - 6x + 8$, so the eigenvalues are 2 and 4. For $\lambda = 2$, $\mathbf{A} - 2\mathbf{I} = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$ gives $x = y$ and the unit eigenvector $\mathbf{q}_1 = \frac{1}{\sqrt{2}}(1, 1)$. For $\lambda = 4$, $\mathbf{A} - 4\mathbf{I} = \begin{pmatrix} -1 & -1 \\ -1 & -1 \end{pmatrix}$ gives $x = -y$ and $\mathbf{q}_2 = \frac{1}{\sqrt{2}}(1, -1)$.

In the coordinates $y_1 = \langle \mathbf{x}, \mathbf{q}_1 \rangle$, $y_2 = \langle \mathbf{x}, \mathbf{q}_2 \rangle$, the equation is $2y_1^2 + 4y_2^2 = 1$. Both eigenvalues are positive, so the curve is an **ellipse**. Its major axis lies along $\mathbf{q}_1$, the line $y = x$, with half-length $1/\sqrt{2}$; its minor axis lies along $\mathbf{q}_2$, the line $y = -x$, with half-length $1/2$. The check is that $\frac{1}{\sqrt{2}}(1, 1)$ scaled to length $1/\sqrt{2}$ is the point $(\frac{1}{2}, \frac{1}{2})$, and $3 \cdot \frac{1}{4} - 2 \cdot \frac{1}{4} + 3 \cdot \frac{1}{4} = 1$.



$3x^2 - 2xy + 3y^2 = 1$: the axes lie along the eigenvectors, with half-lengths $1/\sqrt{\lambda_i}$

Two features of the picture are worth naming. The axes are **orthogonal** because the eigenvectors are, and that is the whole content of the theorem; a nonsymmetric matrix in the same role would produce a pair of distinguished directions that need not be perpendicular. And the bigger eigenvalue gives the **shorter** axis, since the half-length is $1/\sqrt{\lambda}$: a large λ makes q reach the value 1 sooner.

Remark. Chapter 14 will diagonalize quadratic forms again, but by **congruence**: replacing $\mathbf{A}$ by $\mathbf{P}^\top\mathbf{AP}$ for an arbitrary invertible $\mathbf{P}$, not necessarily orthogonal. That is a weaker relation, and it destroys the eigenvalues while preserving only their signs, which is Sylvester's law of inertia. What makes the present theorem special is that $\mathbf{Q}^\top = \mathbf{Q}^{-1}$, so $\mathbf{Q}^\top\mathbf{AQ} = \mathbf{Q}^{-1}\mathbf{AQ}$ and the two relations coincide: an orthogonal change of variables is a similarity *and* a congruence, so it keeps the eigenvalues and the geometry at once.

Warning. **The theorem needs $\mathbb{R}$, and symmetry is not the same as having real eigenvalues.** The matrix $\begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}$ is symmetric with rational entries, but its eigenvalues are $(1 \pm \sqrt{5})/2$, which are irrational; over $\mathbb{Q}$ there is no eigenvector at all, and no version of Theorem 12.5.1 survives. Conversely $\begin{pmatrix} 1 & 1 \\ 0 & 2 \end{pmatrix}$ has the real eigenvalues 1 and 2 and is not symmetric; real eigenvalues alone buy nothing, since they do not make the eigenvectors orthogonal.

12.5.6 Exercises

12.5.6.1 A. Check your understanding

Exercise. A1

(a) State the real spectral theorem and its matrix form, with every hypothesis.

(b) Which two results from Section 2 does the induction use, and what does each one supply?

(c) True or false: a normal operator on a real inner product space has an orthonormal basis of eigenvectors. Justify your answer.

(d) What are the possible diagonal blocks in the real normal form of a normal operator, and what does the condition $b > 0$ rule out?

(e) Given a symmetric $\mathbf{A}$ with orthonormal eigenvectors $\mathbf{q}_1, \mathbf{q}_2$ and eigenvalues 9 and 1, describe the curve $\mathbf{x}^\top \mathbf{A}\mathbf{x} = 1$.

(f) True or false: a real matrix with real eigenvalues is symmetric. Justify your answer.

Solution.

(a) For a finite-dimensional real inner product space V and $T \in \mathcal{L}(V)$: T is self-adjoint if and only if V has an orthonormal basis of eigenvectors (Theorem 12.5.1). Matrix form: $\mathbf{A} \in M_n(\mathbb{R})$ is symmetric if and only if $\mathbf{A} = \mathbf{Q}\mathbf{D}\mathbf{Q}^\top$ with $\mathbf{Q}$ orthogonal and $\mathbf{D}$ real diagonal (Corollary 12.5.2).

(b) Theorem 12.2.2 supplies one eigenvector to peel off, and Theorem 12.2.4 supplies the smaller object to induct on, namely the self-adjoint restriction to the orthogonal complement.

(c) False. The quarter turn of $\mathbb{R}^2$ is normal and has no eigenvector at all (Example (A normal real operator with no eigenvector)).

(d) Blocks (c) with $c \in \mathbb{R}$, and 2×2 blocks $\mathbf{\Lambda}(\lambda) = \begin{pmatrix} a & -b \\ b & a \end{pmatrix}$ with $b > 0$ (Theorem 12.5.3). The condition $b > 0$ rules out the transposed block $\mathbf{\Lambda}(\overline{\lambda})$, which is orthogonally similar to it via $\mathrm{diag}(1, -1)$; fixing the sign is what makes the list of blocks well defined.

(e) An ellipse with axes along $\mathbf{q}_1$ and $\mathbf{q}_2$, of half-lengths $1/3$ and 1 (Theorem 12.5.5). The longer axis is along $\mathbf{q}_2$, the eigenvector for the **smaller** eigenvalue.

(f) False. $\begin{pmatrix} 1 & 1 \\ 0 & 2 \end{pmatrix}$ has eigenvalues $1, 2$ and is not symmetric. The correct statement is the converse: a symmetric real matrix has real eigenvalues.

12.5.6.2 B. Practice

Exercise. B1: Orthogonally diagonalize

For each symmetric matrix, find an orthogonal $\mathbf{Q}$ and a diagonal $\mathbf{D}$ with $\mathbf{Q}^\top \mathbf{A}\mathbf{Q} = \mathbf{D}$.

(a) $\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}$.

(b) $\mathbf{A} = \begin{pmatrix} 4 & 0 & 1 \\ 0 & 2 & 0 \\ 1 & 0 & 4 \end{pmatrix}$.

Solution.

(a) $p_{\mathbf{A}}(x) = x^2 - 2x - 3 = (x - 3)(x + 1)$, so the eigenvalues are 3 and -1. For $\lambda = 3$, $\mathbf{A} - 3\mathbf{I} = \begin{pmatrix} -2 & 2 \\ 2 & -2 \end{pmatrix}$ gives $x_1 = x_2$ and $\mathbf{q}_1 = \frac{1}{\sqrt{2}}(1, 1)$; for $\lambda = -1$, $\mathbf{q}_2 = \frac{1}{\sqrt{2}}(1, -1)$.

Hence

$$\mathbf{Q} = \tfrac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}, \qquad \mathbf{D} = \begin{pmatrix} 3 & 0 \\ 0 & -1 \end{pmatrix}.$$

(b) The second coordinate is untouched: $\mathbf{A}\mathbf{e}_2 = 2\mathbf{e}_2$. On $\operatorname{span}(\mathbf{e}_1, \mathbf{e}_3)$ the matrix acts as $\begin{pmatrix} 4 & 1 \\ 1 & 4 \end{pmatrix}$, with eigenvalues 5 and 3 and eigenvectors $(1,1)$ and $(1,-1)$ there. Lifting back,

$$\mathbf{q}_1 = \tfrac{1}{\sqrt{2}}(1,0,1), \quad \mathbf{q}_2 = \tfrac{1}{\sqrt{2}}(-1,0,1), \quad \mathbf{q}_3 = (0,1,0),$$

with eigenvalues $5, 3, 2$. These are orthonormal, so $\mathbf{Q} = (\mathbf{q}_1 \mid \mathbf{q}_2 \mid \mathbf{q}_3)$ is orthogonal and $\mathbf{Q}^\top \mathbf{A}\mathbf{Q} = \operatorname{diag}(5,3,2)$. As a check, $5+3+2 = 10 = \operatorname{tr}\mathbf{A}$.

Exercise. B2: Identify the conic

Identify the curve $5x^2 + 4xy + 5y^2 = 1$: say what kind of conic it is, find its axes, and find their half-lengths.

Solution. The symmetric matrix of the form is $\mathbf{A} = \begin{pmatrix} 5 & 2 \\ 2 & 5 \end{pmatrix}$, the off-diagonal entries being half of 4. Its characteristic polynomial is $x^2 - 10x + 21 = (x-7)(x-3)$, so the eigenvalues are 7 and 3. For $\lambda = 7$, $\mathbf{A} - 7\mathbf{I} = \begin{pmatrix} -2 & 2 \\ 2 & -2 \end{pmatrix}$ gives $\mathbf{q}_1 = \tfrac{1}{\sqrt{2}}(1,1)$; for $\lambda = 3$, $\mathbf{q}_2 = \tfrac{1}{\sqrt{2}}(1,-1)$.

By Theorem 12.5.5 the equation becomes $7y_1^2 + 3y_2^2 = 1$. Both eigenvalues are positive, so the curve is an ellipse. Its axes are the lines $y = x$ and $y = -x$, with half-lengths $1/\sqrt{7}$ along $\mathbf{q}_1$ and $1/\sqrt{3}$ along $\mathbf{q}_2$; the major axis is the second one. Check: the point $\tfrac{1}{\sqrt{3}}\mathbf{q}_2 = \tfrac{1}{\sqrt{6}}(1,-1)$ gives $\tfrac{5}{6} - \tfrac{4}{6} + \tfrac{5}{6} = 1$.

Exercise. B3: A sum of rank-one projections

Using your answer to Exercise (B1: Orthogonally diagonalize) (b), write

$$\mathbf{A} = \begin{pmatrix} 4 & 0 & 1 \\ 0 & 2 & 0 \\ 1 & 0 & 4 \end{pmatrix}$$

as $\sum_i \lambda_i \mathbf{q}_i \mathbf{q}_i^\top$, with the three matrices $\mathbf{q}_i \mathbf{q}_i^\top$ written out, and verify the identity by adding them.

Solution. With $\mathbf{q}_1 = \tfrac{1}{\sqrt{2}}(1,0,1)$, $\mathbf{q}_2 = \tfrac{1}{\sqrt{2}}(-1,0,1)$, $\mathbf{q}_3 = (0,1,0)$, and $\lambda = 5, 3, 2$,

$$\mathbf{q}_1\mathbf{q}_1^\top = \tfrac{1}{2}\begin{pmatrix} 1 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 1 \end{pmatrix}, \qquad \mathbf{q}_2\mathbf{q}_2^\top = \tfrac{1}{2}\begin{pmatrix} 1 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 1 \end{pmatrix},$$

and $\mathbf{q}_3\mathbf{q}_3^\top$ is the matrix with a single 1 in position $(2,2)$. Then

$$5\,\mathbf{q}_1\mathbf{q}_1^\top + 3\,\mathbf{q}_2\mathbf{q}_2^\top = \begin{pmatrix} 4 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 4 \end{pmatrix},$$

since $\frac{1}{2}(5+3) = 4$ on the diagonal and $\frac{1}{2}(5-3) = 1$ off it. Adding $2\,\mathbf{q}_3\mathbf{q}_3^\top$ inserts the 2 in position $(2,2)$ and gives $\mathbf{A}$, as Corollary 12.5.4 requires.

12.5.6.3 C. Going deeper

Exercise. C1: One eigenvalue only

(a) Let $\mathbf{A} \in M_n(\mathbb{R})$ be symmetric with every eigenvalue equal to λ. Prove that $\mathbf{A} = \lambda\mathbf{I}$.

(b) Show that (a) fails without symmetry, by exhibiting a matrix in $M_2(\mathbb{R})$ whose only eigenvalue is 0 and which is not $\mathbf{O}$.

(c) Deduce that a symmetric $\mathbf{A}$ with $\mathbf{A}^2 = \mathbf{A}$ is the matrix of an orthogonal projection.

Solution.

(a) By Corollary 12.5.2, $\mathbf{A} = \mathbf{Q}\mathbf{D}\mathbf{Q}^\top$ with $\mathbf{D}$ diagonal carrying the eigenvalues, so $\mathbf{D} = \lambda\mathbf{I}$. Hence $\mathbf{A} = \mathbf{Q}(\lambda\mathbf{I})\mathbf{Q}^\top = \lambda\mathbf{Q}\mathbf{Q}^\top = \lambda\mathbf{I}$.

(b) $\mathbf{J}_2(0) = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$ has characteristic polynomial x^2, so 0 is its only eigenvalue, and it is not the zero matrix. It is of course not symmetric.

(c) If $\mathbf{A}^2 = \mathbf{A}$ and $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$ with $\mathbf{x} \neq \mathbf{0}$, then $\lambda^2\mathbf{x} = \mathbf{A}^2\mathbf{x} = \mathbf{A}\mathbf{x} = \lambda\mathbf{x}$, so $\lambda^2 = \lambda$ and $\lambda \in \{0, 1\}$. By Corollary 12.5.2, $\mathbf{A} = \mathbf{Q}\mathbf{D}\mathbf{Q}^\top$ with $\mathbf{D}$ diagonal of zeros and ones, so $\mathbf{A} = \sum_{i \in S} \mathbf{q}_i\mathbf{q}_i^\top$ where S indexes the eigenvalue 1 (Corollary 12.5.4). That is the sum of the projections onto an orthonormal basis of $U = E_1(\mathbf{A})$, which is the orthogonal projection P_U by Theorem 11.3.3.

Exercise. C2: Diagonalizing two matrices at once

Let $\mathbf{A}, \mathbf{B} \in M_n(\mathbb{R})$ be symmetric with $\mathbf{A}\mathbf{B} = \mathbf{B}\mathbf{A}$.

(a) Prove that each eigenspace $E_\lambda(\mathbf{A})$ is $\mathbf{B}$-invariant, and that the restriction of $T_{\mathbf{B}}$ to it is self-adjoint.

(b) Prove that there is a single $\mathbf{Q} \in \mathrm{O}(n)$ with $\mathbf{Q}^\top\mathbf{A}\mathbf{Q}$ and $\mathbf{Q}^\top\mathbf{B}\mathbf{Q}$ both diagonal.

(c) Show that the conclusion fails without commutativity, using $\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$ and $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$.

Hint for (b): apply Theorem 12.5.1 inside each eigenspace of $\mathbf{A}$.

Solution.

(a) Invariance is Theorem 9.9.1: if $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$ then $\mathbf{A}(\mathbf{B}\mathbf{x}) = \mathbf{B}(\mathbf{A}\mathbf{x}) = \lambda\mathbf{B}\mathbf{x}$. For self-adjointness of the restriction, take $\mathbf{u}, \mathbf{v} \in E_\lambda(\mathbf{A})$; then $\langle \mathbf{B}\mathbf{u}, \mathbf{v}\rangle = \langle \mathbf{u}, \mathbf{B}\mathbf{v}\rangle$ because $\mathbf{B}$

is symmetric on all of $\mathbb{R}^n$, and both $\mathbf{Bu}$ and $\mathbf{Bv}$ lie in $E_\lambda(\mathbf{A})$ by the first part, so the identity holds inside the subspace with its restricted inner product.

(b) Let $\lambda_1, \dots, \lambda_k$ be the distinct eigenvalues of $\mathbf{A}$. By Theorem 12.5.1, $\mathbb{R}^n$ is the direct sum of the eigenspaces $E_{\lambda_i}(\mathbf{A})$, and these are mutually orthogonal (Theorem 12.2.3). By (a) and Theorem 12.5.1 applied to the self-adjoint operator $T_{\mathbf{B}}|_{E_{\lambda_i}(\mathbf{A})}$, each $E_{\lambda_i}(\mathbf{A})$ has an orthonormal basis of eigenvectors of $\mathbf{B}$. Every vector of $E_{\lambda_i}(\mathbf{A})$ is an eigenvector of $\mathbf{A}$, so these basis vectors are common eigenvectors. Stringing the k bases together gives an orthonormal basis of $\mathbb{R}^n$, orthonormal across the blocks because distinct eigenspaces of $\mathbf{A}$ are orthogonal. Let $\mathbf{Q}$ have those vectors as columns; then $\mathbf{Q} \in \mathrm{O}(n)$ and both $\mathbf{Q}^\top\mathbf{AQ}$ and $\mathbf{Q}^\top\mathbf{BQ}$ are diagonal.

(c) These two do not commute: $\mathbf{AB} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$ while $\mathbf{BA} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}$. And no common eigenbasis exists: the eigenvectors of $\mathbf{A}$ are the multiples of $\mathbf{e}_1$ and of $\mathbf{e}_2$, while $\mathbf{Be}_1 = \mathbf{e}_2$ is not a multiple of $\mathbf{e}_1$. Commutativity is necessary as well as sufficient, since two simultaneously diagonalized matrices commute.

Exercise. C3: A rotation-scaling in normal form

Let $\mathbf{A} = \begin{pmatrix} 3 & 4 \\ -4 & 3 \end{pmatrix} \in M_2(\mathbb{R})$.

(a) Verify that $\mathbf{A}$ is normal and not symmetric, and find its complex eigenvalues.

(b) Find an orthogonal $\mathbf{Q}$ with $\mathbf{Q}^\top\mathbf{AQ} = \mathbf{\Lambda}(\lambda)$ for the eigenvalue λ with positive imaginary part.

(c) Explain why no orthogonal $\mathbf{Q}$ makes $\mathbf{Q}^\top\mathbf{AQ}$ diagonal, and why this does not contradict Theorem 12.4.2.

Solution.

(a) $\mathbf{A}^\top\mathbf{A} = \mathbf{AA}^\top = 25\mathbf{I}$, so $\mathbf{A}$ is normal; $\mathbf{A}^\top \neq \mathbf{A}$ since $4 \neq -4$. Its characteristic polynomial is $x^2 - 6x + 25 = (x-3)^2 + 16$, with roots $3 \pm 4i$.

(b) The eigenvalue with positive imaginary part is $\lambda = 3 + 4i$, so $\mathbf{\Lambda}(\lambda) = \begin{pmatrix} 3 & -4 \\ 4 & 3 \end{pmatrix}$. Take $\mathbf{Q} = \mathrm{diag}(1, -1)$, which is orthogonal. Then

$$\mathbf{Q}^\top\mathbf{AQ} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}\begin{pmatrix} 3 & 4 \\ -4 & 3 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} = \begin{pmatrix} 3 & -4 \\ 4 & 3 \end{pmatrix},$$

which is $\mathbf{\Lambda}(3+4i)$. This is Case 2 of the proof of Theorem 12.5.3 with $p = 3$ and $q = 4$.

(c) If $\mathbf{Q}^\top\mathbf{AQ}$ were diagonal for some $\mathbf{Q} \in \mathrm{O}(2)$, then $\mathbf{A} = \mathbf{QDQ}^\top$ would be symmetric by Corollary 12.5.2, and it is not. There is no contradiction with Theorem 12.4.2, which concerns a **complex** inner product space: over $\mathbb{C}^2$ the matrix $\mathbf{A}$ is indeed unitarily diagonalizable, to $\mathrm{diag}(3+4i, 3-4i)$, but the diagonalizing basis $\frac{1}{\sqrt{2}}(1, \pm i)$ is not real, so it is unavailable inside $\mathbb{R}^2$.

12.6 The Shape of an Orthogonal Matrix

An orthogonal matrix is normal, so the real normal form of the previous section applies to it — and the extra information that an orthogonal matrix carries, that every eigenvalue has modulus 1, collapses that form into something very small. What comes out is a complete description of the rigid motions of $\mathbb{R}^n$ fixing the origin: a few axes reversed, and the rest of the space turned in independent two-dimensional planes. In dimension 3 this is the statement that every rotation has an axis, and we push it as far as an explicit formula for the rotation matrix in terms of its axis and angle.

Throughout, $\mathbb{R}^n$ carries the standard inner product, and $\mathbf{Q}$ is a real matrix.

12.6.1 From the real normal form to the canonical form

Two pieces of notation, both recalled. For $\theta \in \mathbb{R}$, Chapter 11 wrote

$$\mathbf{R}_\theta = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

for the matrix that rotates the plane by θ (Theorem 11.7.6). It is exactly the rotation-scaling block of Chapter 10 for a complex number on the unit circle: $\mathbf{R}_\theta = \mathbf{\Lambda}(e^{i\theta})$ in the notation of Definition 10.4.1, since $e^{i\theta} = \cos\theta + i\sin\theta$. And $\mathbf{A} \oplus \mathbf{B}$ is the block diagonal matrix with diagonal blocks $\mathbf{A}$ and $\mathbf{B}$.

Now the observation that starts everything. If $\mathbf{Q} \in \mathrm{O}(n)$ then $\mathbf{Q}^\top\mathbf{Q} = \mathbf{I}_n = \mathbf{Q}\mathbf{Q}^\top$, so $\mathbf{Q}$ commutes with its own adjoint: **every orthogonal matrix is normal**. So Theorem 12.5.3 applies. All that remains is to say what the blocks can be, and for that we use the one thing an orthogonal matrix has that a general normal matrix does not.

Theorem 12.6.1. Canonical Form of an Orthogonal Matrix

Let $\mathbf{Q} \in \mathrm{O}(n)$. Then there are integers $p, q, m \geq 0$ with $p+q+2m = n$, and angles $\theta_1, \ldots, \theta_m$ with $0 < \theta_j < \pi$, and a matrix $\mathbf{P} \in \mathrm{O}(n)$, such that

$$\mathbf{P}^\top\mathbf{Q}\mathbf{P} = \mathbf{I}_p \oplus (-\mathbf{I}_q) \oplus \mathbf{R}_{\theta_1} \oplus \cdots \oplus \mathbf{R}_{\theta_m}.$$

Moreover $\det \mathbf{Q} = (-1)^q$, and p, q, and the multiset $\{\theta_1, \ldots, \theta_m\}$ are determined by $\mathbf{Q}$.

Idea. $\mathbf{Q}$ is normal, so Theorem 12.5.3 already hands us a block diagonal matrix with 1×1 blocks $[a]$ and 2×2 blocks $\mathbf{\Lambda}(\lambda)$. The blocks are unconstrained for a general normal matrix; here they are not, because the eigenvalues of an isometry lie on the unit circle. A real eigenvalue on the unit circle is ± 1, and a conjugate pair on the unit circle is $e^{\pm i\theta}$, which turns the scaling block into a pure rotation. Then sort the ± 1s to the front.

Proof. The operator $T_{\mathbf{Q}}$ on $\mathbb{R}^n$ satisfies $T_{\mathbf{Q}}^* = T_{\mathbf{Q}^\top}$ by Theorem 11.6.3, since the standard basis is orthonormal, so $T_{\mathbf{Q}}^* T_{\mathbf{Q}} = T_{\mathbf{Q}^\top\mathbf{Q}} = \mathrm{id}$ and likewise in the other order. Hence $T_{\mathbf{Q}}$ is normal (Definition 11.6.3), and by Theorem 12.5.3 there is an orthonormal basis $\mathcal{B}$ of $\mathbb{R}^n$ in

which

$$[T_{\mathbf{Q}}]_{\mathcal{B}}^{\mathcal{B}} = [a_1] \oplus \cdots \oplus [a_r] \oplus \mathbf{\Lambda}(\lambda_1) \oplus \cdots \oplus \mathbf{\Lambda}(\lambda_m),$$

where $a_i \in \mathbb{R}$ and $\lambda_j = c_j + d_j i$ with $d_j > 0$. Let $\mathbf{P}$ be the matrix whose columns are the vectors of $\mathcal{B}$. Its columns are orthonormal, so $\mathbf{P} \in \mathrm{O}(n)$ by Theorem 11.7.3 (f), and $\mathbf{P}^{\top}\mathbf{Q}\mathbf{P} = \mathbf{P}^{-1}\mathbf{Q}\mathbf{P} = [T_{\mathbf{Q}}]_{\mathcal{B}}^{\mathcal{B}}$ by Theorem 4.2.3.

The blocks. A block diagonal matrix is triangular by blocks, so the eigenvalues of $[T_{\mathbf{Q}}]_{\mathcal{B}}^{\mathcal{B}}$ over $\mathbb{C}$ are the a_i together with the roots of each $\det(x\mathbf{I}_2 - \mathbf{\Lambda}(\lambda_j)) = (x - c_j)^2 + d_j^2$, namely λ_j and $\overline{\lambda_j}$. Now $\mathrm{O}(n) = \mathrm{U}(n) \cap M_n(\mathbb{R})$ by Proposition 11.7.4 (a), so $\mathbf{Q}$ is unitary, and every complex eigenvalue λ of $\mathbf{Q}$ satisfies $|\lambda| = 1$ by Proposition 11.7.1 applied to $T_{\mathbf{Q}}$ on $\mathbb{C}^n$. Therefore $a_i = \pm 1$ for each i, and $c_j^2 + d_j^2 = 1$ for each j. Since $c_j^2 + d_j^2 = 1$ and $d_j > 0$, there is a unique $\theta_j \in (0, \pi)$ with $c_j = \cos\theta_j$ and $d_j = \sin\theta_j$, and then $\mathbf{\Lambda}(\lambda_j) = \mathbf{R}_{\theta_j}$.

Sorting. Let p be the number of i with $a_i = 1$ and q the number with $a_i = -1$, so $p + q + 2m = n$. Reordering the vectors of $\mathcal{B}$ so that the $+1$ blocks come first, then the -1 blocks, then the rotation blocks in pairs, leaves the basis orthonormal and permutes the diagonal blocks; the new $\mathbf{P}$ is the old one with its columns permuted, still orthogonal. This gives the displayed form.

The determinant and uniqueness. By Theorem 7.5.4 and Corollary 7.5.7, $\det\mathbf{Q} = \det(\mathbf{P}^{\top}\mathbf{Q}\mathbf{P})$, and the determinant of a block diagonal matrix is the product of the determinants of its blocks (Theorem 7.8.4), which are 1, -1, and $\det\mathbf{R}_{\theta_j} = \cos^2\theta_j + \sin^2\theta_j = 1$. Hence $\det\mathbf{Q} = (-1)^q$. Finally, similar matrices have the same eigenvalues with the same multiplicities, and reading the eigenvalues off the displayed form gives 1 with multiplicity p, -1 with multiplicity q, and the pairs $e^{\pm i\theta_j}$; so p, q, and the θ_j are recovered from $\mathbf{Q}$ alone. This proves the theorem.

■

So an orthogonal matrix is, up to an orthogonal change of coordinates, nothing but a list of angles together with a count of reversed axes. The number q of reversed axes is not itself an invariant of the *geometry* — a -1 block and a -1 block together make $-\mathbf{I}_2 = \mathbf{R}_{\pi}$, a rotation by π — which is why the theorem demands $\theta_j \neq \pi$ and keeps the -1s separate. What the determinant sees is only the parity of q.

Quick check. What is the canonical form of a $\mathbf{Q} \in \mathrm{O}(2)$ with $\det\mathbf{Q} = -1$? Compare with Theorem 11.7.6.

Solution. Here $p + q + 2m = 2$ and $(-1)^q = -1$, so q is odd, forcing $q = 1$, $p = 1$, $m = 0$. The canonical form is $\mathbf{I}_1 \oplus (-\mathbf{I}_1) = \mathrm{diag}(1, -1)$. Chapter 11 found that such a $\mathbf{Q}$ is the reflection $\mathbf{M}_{\theta}$ in a line L, fixing L and negating $L^{\perp}$; in the orthonormal basis of eigenvectors of Theorem 11.7.6 (b) its matrix is exactly $\mathrm{diag}(1, -1)$. Note what the canonical form drops: the angle of the mirror. All reflections of the plane are orthogonally similar to one another, and Exercise (C1: The plane, re-derived) makes this comparison precise.

12.6.2 Rotations of space

In dimension 3 the arithmetic $p+q+2m=3$ leaves very few possibilities, and they all look the same.

Corollary 12.6.2. Orthogonal Matrices in Dimension Three

Let $\mathbf{Q} \in \mathrm{O}(3)$. Then there are an orthonormal basis $(\mathbf{u}, \mathbf{w}_1, \mathbf{w}_2)$ of $\mathbb{R}^3$ and an angle $\theta \in [0, \pi]$ such that, in that basis, the matrix of $T_{\mathbf{Q}}$ is

$$[1] \oplus \mathbf{R}_\theta \quad \text{if } \det \mathbf{Q} = 1, \qquad [-1] \oplus \mathbf{R}_\theta \quad \text{if } \det \mathbf{Q} = -1.$$

In particular every $\mathbf{Q} \in \mathrm{SO}(3)$ fixes the line $\mathrm{span}(\mathbf{u})$ pointwise and rotates the plane $\mathrm{span}(\mathbf{u})^\perp$ by θ.

Proof. Write the canonical form of Theorem 12.6.1, with $p+q+2m=3$ and $\det \mathbf{Q} = (-1)^q$.

Case $\det \mathbf{Q} = 1$. Then q is even, so $q \in \{0, 2\}$. If $q = 0$ then $p+2m=3$, giving $(p, m) = (3, 0)$ or $(1, 1)$; if $q = 2$ then $p = 1$ and $m = 0$. The three forms are

$$\mathbf{I}_3 = [1] \oplus \mathbf{R}_0, \qquad [1] \oplus \mathbf{R}_{\theta_1}, \qquad [1] \oplus (-\mathbf{I}_2) = [1] \oplus \mathbf{R}_\pi,$$

and in every case $p \geq 1$, so the form is $[1] \oplus \mathbf{R}_\theta$ with $\theta \in [0, \pi]$.

Case $\det \mathbf{Q} = -1$. Then q is odd, so $q \in \{1, 3\}$. If $q = 1$ then $p+2m=2$, giving $(p, m) = (2, 0)$ or $(0, 1)$; if $q = 3$ then $p = m = 0$. The three forms are $[-1] \oplus \mathbf{I}_2 = [-1] \oplus \mathbf{R}_0$, then $[-1] \oplus \mathbf{R}_{\theta_1}$, and $-\mathbf{I}_3 = [-1] \oplus \mathbf{R}_\pi$.

In both cases the first basis vector $\mathbf{u}$ spans the ± 1 eigenline and $(\mathbf{w}_1, \mathbf{w}_2)$ is an orthonormal basis of $\mathrm{span}(\mathbf{u})^\perp$, which is what the last sentence asserts when $\det \mathbf{Q} = 1$.

■

Chapter 11 proved the SO(3) statement directly, as Theorem 11.7.7, by producing a real eigenvalue from the odd degree of the characteristic polynomial and then invoking the classification of O(2). We do not repeat that argument: the point here is that the same conclusion falls out of the canonical form with no work at all, and that the same three lines also settle $\det \mathbf{Q} = -1$. An orthogonal matrix of determinant -1 in dimension 3 is a **rotatory reflection**: a rotation about an axis followed by the reflection in the plane perpendicular to it. Its two extreme cases are the plane reflection ($\theta = 0$) and the antipodal map $-\mathbf{I}_3$ ($\theta = \pi$).

Remark. The angle θ is determined by $\mathbf{Q}$ only up to sign, because exchanging $\mathbf{w}_1$ and $\mathbf{w}_2$ replaces $\mathbf{R}_\theta$ by $\mathbf{R}_{-\theta}$. Restricting to $\theta \in [0, \pi]$, as above, makes it unique; to speak of a *signed* angle one must also fix which of the two unit vectors on the axis is called $\mathbf{u}$.

12.6.3 The cross-product matrix

The canonical form tells us that a rotation of $\mathbb{R}^3$ exists in a suitable basis. For computation we want it in the standard basis, written directly in terms of the axis. The bookkeeping is done by one matrix.

Recall from Chapter 7 that the cross product of $\mathbf{a}, \mathbf{b} \in \mathbb{R}^3$ is

$$\mathbf{a} \times \mathbf{b} = (a_2b_3 - a_3b_2,\ a_3b_1 - a_1b_3,\ a_1b_2 - a_2b_1).$$

For fixed $\mathbf{a}$ this is linear in $\mathbf{b}$, so it is given by a matrix.

Definition 12.6.1. Cross-product Matrix

For $\mathbf{n} = (n_1, n_2, n_3) \in \mathbb{R}^3$, the **cross-product matrix** of $\mathbf{n}$ is

$$\mathbf{K_n} \coloneqq \begin{pmatrix} 0 & -n_3 & n_2 \\ n_3 & 0 & -n_1 \\ -n_2 & n_1 & 0 \end{pmatrix} \in M_3(\mathbb{R}),$$

so that $\mathbf{K_n v} = \mathbf{n} \times \mathbf{v}$ for every $\mathbf{v} \in \mathbb{R}^3$.

The last clause is a check of the three entries against the formula above, with $\mathbf{a} = \mathbf{n}$ and $\mathbf{b} = \mathbf{v}$. The map $\mathbf{n} \mapsto \mathbf{K_n}$ is linear and injective, and its image is exactly the set of skew-symmetric 3×3 matrices: a skew matrix has zero diagonal and three free entries below it, and those three entries are $n_3, -n_2, n_1$. We record the facts we need, all of them for a **unit** vector.

Lemma 12.6.3. Properties of the Cross-product Matrix

Let $\mathbf{n} \in \mathbb{R}^3$ with $\|\mathbf{n}\| = 1$, and write $\mathbf{K} = \mathbf{K_n}$. Then

(a) $\mathbf{K}^\top = -\mathbf{K}$ and $\mathbf{Kn} = \mathbf{0}$;
(b) $\mathbf{K}^2 = \mathbf{n}\mathbf{n}^\top - \mathbf{I}_3$, so $\mathbf{K}^2$ is symmetric and $\mathbf{K}^2\mathbf{v} = -\mathbf{v}$ for every $\mathbf{v} \perp \mathbf{n}$;
(c) $\mathbf{K}^3 = -\mathbf{K}$;
(d) for every $\mathbf{v} \perp \mathbf{n}$: $\|\mathbf{Kv}\| = \|\mathbf{v}\|$, $\mathbf{Kv} \perp \mathbf{v}$, and $\mathbf{Kv} \perp \mathbf{n}$.

Idea. Everything follows from (b), and (b) is one entry of a matrix product together with $n_1^2 + n_2^2 + n_3^2 = 1$. Part (c) is (b) multiplied by $\mathbf{K}$, where the term $\mathbf{K}\mathbf{n}\mathbf{n}^\top$ dies by (a). For (d), a norm is computed through its square, and $\|\mathbf{Kv}\|^2 = \mathbf{v}^\top\mathbf{K}^\top\mathbf{Kv} = -\mathbf{v}^\top\mathbf{K}^2\mathbf{v}$ turns (d) into (b) as well.

Proof.

(a) Transposing $\mathbf{K_n}$ changes the sign of each of the three off-diagonal pairs, and the diagonal is zero. For the second claim, $\mathbf{Kn} = \mathbf{n} \times \mathbf{n} = \mathbf{0}$ directly from the formula, each coordinate being of the form $n_in_j - n_jn_i$.

(b) The $(1,1)$ entry of $\mathbf{K}^2$ is $0 \cdot 0 + (-n_3)n_3 + n_2(-n_2) = -n_2^2 - n_3^2 = n_1^2 - 1$, using $\|\mathbf{n}\|^2 = 1$; and $(\mathbf{n}\mathbf{n}^\top - \mathbf{I}_3)_{11} = n_1^2 - 1$. The $(1,2)$ entry of $\mathbf{K}^2$ is $0(-n_3) + (-n_3)0 + n_2n_1 = n_1n_2 = (\mathbf{n}\mathbf{n}^\top)_{12}$, and the $(1,3)$ entry is $0n_2 + (-n_3)(-n_1) + n_20 = n_1n_3$. The remaining six entries are obtained by cycling the indices $1 \to 2 \to 3 \to 1$, under which both $\mathbf{K}$ and $\mathbf{n}\mathbf{n}^\top - \mathbf{I}_3$ are unchanged in form. Hence $\mathbf{K}^2 = \mathbf{n}\mathbf{n}^\top - \mathbf{I}_3$, which is symmetric since $(\mathbf{n}\mathbf{n}^\top)^\top = \mathbf{n}\mathbf{n}^\top$. If $\mathbf{v} \perp \mathbf{n}$ then $\mathbf{n}^\top\mathbf{v} = 0$, so $\mathbf{K}^2\mathbf{v} = \mathbf{n}(\mathbf{n}^\top\mathbf{v}) - \mathbf{v} = -\mathbf{v}$.

(c) By (b) and (a), $\mathbf{K}^3 = \mathbf{K}(\mathbf{n}\mathbf{n}^\top - \mathbf{I}_3) = (\mathbf{Kn})\mathbf{n}^\top - \mathbf{K} = -\mathbf{K}$.

(d) Let $\mathbf{v} \perp \mathbf{n}$. Using (a) and then (b),

$$\|\mathbf{K}\mathbf{v}\|^2 = (\mathbf{K}\mathbf{v})^\top(\mathbf{K}\mathbf{v}) = \mathbf{v}^\top\mathbf{K}^\top\mathbf{K}\mathbf{v} = -\mathbf{v}^\top\mathbf{K}^2\mathbf{v} = \mathbf{v}^\top\mathbf{v} = \|\mathbf{v}\|^2.$$

Next, $\langle \mathbf{K}\mathbf{v}, \mathbf{v}\rangle = \mathbf{v}^\top\mathbf{K}\mathbf{v}$ is a 1×1 matrix, hence equal to its own transpose $\mathbf{v}^\top\mathbf{K}^\top\mathbf{v} = -\mathbf{v}^\top\mathbf{K}\mathbf{v}$; a number equal to its own negative is 0, so $\mathbf{K}\mathbf{v} \perp \mathbf{v}$. Finally $\langle \mathbf{K}\mathbf{v}, \mathbf{n}\rangle = \mathbf{n}^\top\mathbf{K}\mathbf{v} = (\mathbf{K}^\top\mathbf{n})^\top\mathbf{v} = -(\mathbf{K}\mathbf{n})^\top\mathbf{v} = 0$ by (a). This proves the lemma.

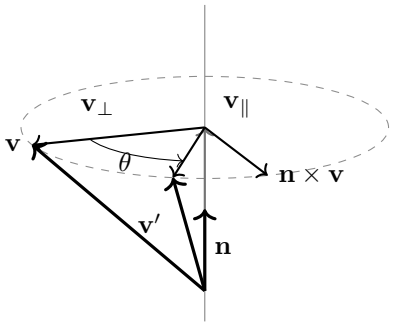
■

Part (c) is the identity that makes every power of $\mathbf{K}$ reduce to $\mathbf{K}$ or $\mathbf{K}^2$, and it is the reason Rodrigues' formula has only three terms and no more.

12.6.4 Rodrigues' formula

Fix a unit vector $\mathbf{n}$ and an angle θ. We want a formula for "turn space by θ about the axis $\operatorname{span}(\mathbf{n})$". The construction is forced by the picture: split a vector into the part along the axis, which does not move, and the part perpendicular to the axis, which turns inside a single plane.

Write $\mathbf{v}_\parallel = \langle \mathbf{v}, \mathbf{n}\rangle\mathbf{n} = \mathbf{n}\mathbf{n}^\top\mathbf{v}$ for the component of $\mathbf{v}$ along the axis and $\mathbf{v}_\perp = \mathbf{v} - \mathbf{v}_\parallel$ for the component perpendicular to it; this is the splitting of Theorem 11.3.2 for $U = \operatorname{span}(\mathbf{n})$, and $\mathbf{v}_\parallel = P_U\mathbf{v}$ by Theorem 11.3.3 (a) with the orthonormal basis $(\mathbf{n})$ of U. Write $\mathbf{v}'$ for the vector we want, the image of $\mathbf{v}$ under the turn.



the axis part stays, the perpendicular part turns

By Lemma 12.6.3 (d), the vector $\mathbf{K}_\mathbf{n}\mathbf{v}_\perp$ has the same length as $\mathbf{v}_\perp$ and is perpendicular to both $\mathbf{v}_\perp$ and $\mathbf{n}$: it is $\mathbf{v}_\perp$ turned by a quarter turn inside the plane $\mathbf{n}^\perp$. So turning $\mathbf{v}_\perp$ by θ means taking $\cos\theta\,\mathbf{v}_\perp + \sin\theta\,\mathbf{K}_\mathbf{n}\mathbf{v}_\perp$, and $\mathbf{v}' = \mathbf{v}_\parallel + \cos\theta\,\mathbf{v}_\perp + \sin\theta\,\mathbf{K}_\mathbf{n}\mathbf{v}_\perp$. Finally $\mathbf{K}_\mathbf{n}\mathbf{v}_\perp = \mathbf{K}_\mathbf{n}\mathbf{v}$, since $\mathbf{K}_\mathbf{n}\mathbf{v}_\parallel$ is a multiple of $\mathbf{K}_\mathbf{n}\mathbf{n} = \mathbf{0}$. That is the definition we take, and the theorem says it is the right one.

Theorem 12.6.4. Rodrigues' Rotation Formula

Let $\mathbf{n} \in \mathbb{R}^3$ be a unit vector, let $\theta \in \mathbb{R}$, and write $\mathbf{K} = \mathbf{K}_\mathbf{n}$. Define

$$\mathbf{R}_{\mathbf{n},\theta} \coloneqq \mathbf{I}_3 + \sin\theta\,\mathbf{K} + (1 - \cos\theta)\,\mathbf{K}^2.$$

(a) For every $\mathbf{v} \in \mathbb{R}^3$,

$$\mathbf{R}_{\mathbf{n},\theta}\mathbf{v} = \langle \mathbf{v}, \mathbf{n}\rangle\mathbf{n} + \cos\theta\,(\mathbf{v} - \langle \mathbf{v}, \mathbf{n}\rangle\mathbf{n}) + \sin\theta\,(\mathbf{n} \times \mathbf{v}).$$

(b) $\mathbf{R}_{\mathbf{n},\theta}\mathbf{n} = \mathbf{n}$, and for every unit $\mathbf{w} \perp \mathbf{n}$ the list $\mathcal{B} = (\mathbf{n}, \mathbf{w}, \mathbf{n} \times \mathbf{w})$ is an orthonormal basis of $\mathbb{R}^3$ with $[T_{\mathbf{R}_{\mathbf{n},\theta}}]_{\mathcal{B}}^{\mathcal{B}} = [1] \oplus \mathbf{R}_\theta$. In particular $\mathbf{R}_{\mathbf{n},\theta} \in \mathrm{SO}(3)$.

(c) Conversely, every $\mathbf{Q} \in \mathrm{SO}(3)$ equals $\mathbf{R}_{\mathbf{n},\theta}$ for some unit vector $\mathbf{n}$ and some $\theta \in [0, \pi]$.

Idea. Part (a) is the paragraph above, read backwards: substitute $\mathbf{n}\mathbf{n}^\top = \mathbf{K}^2 + \mathbf{I}_3$ from Lemma 12.6.3 (b) into the geometric expression and collect terms. Part (b) then needs only the images of three basis vectors, and every one of them is a two-line computation with (a) and the lemma; the matrix that comes out is the canonical form of Corollary 12.6.2, which also gives orthogonality and the determinant for free. Part (c) runs Corollary 12.6.2 in reverse: it hands us a basis $(\mathbf{u}, \mathbf{w}_1, \mathbf{w}_2)$, and the only question is whether $\mathbf{w}_2$ is $\mathbf{u} \times \mathbf{w}_1$ or its negative — and if it is the negative, we change the sign of the angle.

Proof.

(a) By Lemma 12.6.3 (b), $\mathbf{n}\mathbf{n}^\top = \mathbf{K}^2 + \mathbf{I}_3$. Hence, writing $\langle \mathbf{v}, \mathbf{n} \rangle \mathbf{n} = \mathbf{n}\mathbf{n}^\top \mathbf{v}$ and $\mathbf{n} \times \mathbf{v} = \mathbf{K}\mathbf{v}$,

$$\begin{aligned} &\mathbf{n}\mathbf{n}^\top \mathbf{v} + \cos\theta\,(\mathbf{v} - \mathbf{n}\mathbf{n}^\top \mathbf{v}) + \sin\theta\,\mathbf{K}\mathbf{v} \\ &\quad = \cos\theta\,\mathbf{v} + (1 - \cos\theta)\mathbf{n}\mathbf{n}^\top \mathbf{v} + \sin\theta\,\mathbf{K}\mathbf{v} \\ &\quad = \cos\theta\,\mathbf{v} + (1 - \cos\theta)(\mathbf{K}^2 + \mathbf{I}_3)\mathbf{v} + \sin\theta\,\mathbf{K}\mathbf{v} \\ &\quad = (\mathbf{I}_3 + \sin\theta\,\mathbf{K} + (1 - \cos\theta)\mathbf{K}^2)\mathbf{v}, \end{aligned}$$

since $\cos\theta + (1 - \cos\theta) = 1$. The right-hand side is $\mathbf{R}_{\mathbf{n},\theta}\mathbf{v}$.

(b) First $\mathbf{K}\mathbf{n} = \mathbf{0}$ and $\mathbf{K}^2\mathbf{n} = \mathbf{K}(\mathbf{K}\mathbf{n}) = \mathbf{0}$ by Lemma 12.6.3 (a), so $\mathbf{R}_{\mathbf{n},\theta}\mathbf{n} = \mathbf{n}$.
Let $\mathbf{w}$ be a unit vector with $\mathbf{w} \perp \mathbf{n}$, and put $\mathbf{z} = \mathbf{K}\mathbf{w} = \mathbf{n} \times \mathbf{w}$. By Lemma 12.6.3 (d), $\|\mathbf{z}\| = 1$, $\mathbf{z} \perp \mathbf{w}$, and $\mathbf{z} \perp \mathbf{n}$, so $\mathcal{B} = (\mathbf{n}, \mathbf{w}, \mathbf{z})$ is an orthonormal list of three vectors in $\mathbb{R}^3$, hence an orthonormal basis (Theorem 11.1.4 and Theorem 1.6.4). Now compute the three images. We have $\mathbf{R}_{\mathbf{n},\theta}\mathbf{n} = \mathbf{n}$. Next, $\mathbf{K}^2\mathbf{w} = -\mathbf{w}$ by Lemma 12.6.3 (b), since $\mathbf{w} \perp \mathbf{n}$, so

$$\mathbf{R}_{\mathbf{n},\theta}\mathbf{w} = \mathbf{w} + \sin\theta\,\mathbf{z} - (1 - \cos\theta)\mathbf{w} = \cos\theta\,\mathbf{w} + \sin\theta\,\mathbf{z}.$$

Finally $\mathbf{z} \perp \mathbf{n}$, so $\mathbf{K}^2\mathbf{z} = -\mathbf{z}$, and $\mathbf{K}\mathbf{z} = \mathbf{K}^2\mathbf{w} = -\mathbf{w}$; hence

$$\mathbf{R}_{\mathbf{n},\theta}\mathbf{z} = \mathbf{z} - \sin\theta\,\mathbf{w} - (1 - \cos\theta)\mathbf{z} = -\sin\theta\,\mathbf{w} + \cos\theta\,\mathbf{z}.$$

Reading off the columns (Definition 4.1.1), $[T_{\mathbf{R}_{\mathbf{n},\theta}}]_{\mathcal{B}}^{\mathcal{B}} = [1] \oplus \mathbf{R}_\theta$. This matrix is orthogonal, and it is the matrix of $T_{\mathbf{R}_{\mathbf{n},\theta}}$ in an orthonormal basis, so $T_{\mathbf{R}_{\mathbf{n},\theta}}$ is an isometry and $\mathbf{R}_{\mathbf{n},\theta} \in \mathrm{O}(3)$ by Theorem 11.7.3 (f) and Definition 11.7.2. The two matrices are matrices of the same operator, hence similar (Theorem 4.2.4 (a)), so they have the same determinant (Corollary 7.5.7), namely 1; so $\mathbf{R}_{\mathbf{n},\theta} \in \mathrm{SO}(3)$.

(c) Let $\mathbf{Q} \in \mathrm{SO}(3)$. By Corollary 12.6.2 there are an orthonormal basis $\mathcal{C} = (\mathbf{u}, \mathbf{w}_1, \mathbf{w}_2)$ of $\mathbb{R}^3$ and $\theta \in [0, \pi]$ with $[T_{\mathbf{Q}}]_{\mathcal{C}}^{\mathcal{C}} = [1] \oplus \mathbf{R}_\theta$. The vector $\mathbf{u} \times \mathbf{w}_1$ is a unit vector orthogonal to $\mathbf{u}$ and $\mathbf{w}_1$ by Lemma 12.6.3 (d), and so is $\mathbf{w}_2$; since $\mathrm{span}(\mathbf{u}, \mathbf{w}_1)^\perp$ is a line (Theorem 11.3.2 (c)), its unit vectors are $\pm\mathbf{u} \times \mathbf{w}_1$, so $\mathbf{w}_2 = \varepsilon\,\mathbf{u} \times \mathbf{w}_1$ with $\varepsilon = \pm 1$.

Case $\varepsilon = 1$. Then (b), applied with $\mathbf{n} = \mathbf{u}$ and $\mathbf{w} = \mathbf{w}_1$, gives $[T_{\mathbf{R}_{\mathbf{u},\theta}}]_{\mathcal{C}}^{\mathcal{C}} = [1] \oplus \mathbf{R}_\theta = [T_{\mathbf{Q}}]_{\mathcal{C}}^{\mathcal{C}}$. Operators with the same matrix in one basis are equal, so $\mathbf{Q} = \mathbf{R}_{\mathbf{u},\theta}$.

Case $\varepsilon = -1$. Then $\mathbf{u} \times \mathbf{w}_1 = -\mathbf{w}_2$, and the matrix of $T_{\mathbf{Q}}$ in the basis $(\mathbf{u}, \mathbf{w}_1, -\mathbf{w}_2)$ is obtained from $[1] \oplus \mathbf{R}_\theta$ by changing the sign of the third basis vector, which changes the

sign of the two off-diagonal entries of the $\mathbf{R}_\theta$ block: it is $[1] \oplus \mathbf{R}_{-\theta}$. By the previous case, $\mathbf{Q} = \mathbf{R}_{\mathbf{u},-\theta}$. Finally $\mathbf{K}_{-\mathbf{u}} = -\mathbf{K}_{\mathbf{u}}$, so

$$\mathbf{R}_{-\mathbf{u},\theta} = \mathbf{I}_3 - \sin\theta\,\mathbf{K}_{\mathbf{u}} + (1-\cos\theta)\mathbf{K}_{\mathbf{u}}^2 = \mathbf{R}_{\mathbf{u},-\theta},$$

and $\mathbf{Q} = \mathbf{R}_{-\mathbf{u},\theta}$ with $-\mathbf{u}$ a unit vector and $\theta \in [0,\pi]$. This proves the theorem.

■

Part (c) is worth restating on its own: **a rotation of $\mathbb{R}^3$ is nothing but a unit vector and an angle**. That is three numbers of data, one of them redundant, for an object that started life as nine matrix entries subject to six equations.

Quick check. Compute $\mathbf{R}_{\mathbf{n},\pi}$ from Rodrigues' formula and identify the map geometrically.

Solution. With $\sin\pi = 0$ and $1-\cos\pi = 2$, the formula gives $\mathbf{R}_{\mathbf{n},\pi} = \mathbf{I}_3 + 2\mathbf{K}^2 = \mathbf{I}_3 + 2(\mathbf{n}\mathbf{n}^\top - \mathbf{I}_3) = 2\mathbf{n}\mathbf{n}^\top - \mathbf{I}_3$, using Lemma 12.6.3 (b). It fixes $\mathbf{n}$ and negates every vector perpendicular to $\mathbf{n}$: the half-turn about the axis. Equivalently $\mathbf{R}_{\mathbf{n},\pi} = -(\mathbf{I}_3 - 2\mathbf{n}\mathbf{n}^\top) = -\mathbf{H}_{\mathbf{n}}$, minus the Householder reflection of Definition 11.8.1 in the plane $\mathbf{n}^\perp$.

12.6.5 Reading off the axis and the angle

Rodrigues' formula runs one way: from $(\mathbf{n},\theta)$ to $\mathbf{Q}$. The inverse direction is just as cheap, because the formula splits $\mathbf{Q}$ into a symmetric part and a skew part that carry the two pieces of data separately.

Theorem 12.6.5. Axis and Angle from the Matrix

Let $\mathbf{Q} = \mathbf{R}_{\mathbf{n},\theta} \in \mathrm{SO}(3)$ with $\mathbf{n}$ a unit vector and $\theta \in [0,\pi]$. Then

(a) $\operatorname{tr}\mathbf{Q} = 1 + 2\cos\theta$, so $\cos\theta = \frac{1}{2}(\operatorname{tr}\mathbf{Q} - 1)$ and θ is determined by $\mathbf{Q}$;

(b) $\mathbf{Q} - \mathbf{Q}^\top = 2\sin\theta\,\mathbf{K}_{\mathbf{n}}$;

(c) if $0 < \theta < \pi$, then $\sin\theta > 0$ and

$$\mathbf{n} = \frac{1}{2\sin\theta}\,(q_{32} - q_{23},\; q_{13} - q_{31},\; q_{21} - q_{12});$$

if $\theta = 0$ then $\mathbf{Q} = \mathbf{I}_3$ and the axis is arbitrary; and if $\theta = \pi$ then $\mathbf{Q} + \mathbf{I}_3 = 2\mathbf{n}\mathbf{n}^\top$, whose nonzero columns are the multiples $2n_i\mathbf{n}$ of $\mathbf{n}$.

Proof. Write $\mathbf{K} = \mathbf{K}_{\mathbf{n}}$. By Lemma 12.6.3 (b), $\mathbf{K}^2 = \mathbf{n}\mathbf{n}^\top - \mathbf{I}_3$, so $\operatorname{tr}\mathbf{K}^2 = \operatorname{tr}(\mathbf{n}\mathbf{n}^\top) - 3 = \|\mathbf{n}\|^2 - 3 = -2$; and $\operatorname{tr}\mathbf{K} = 0$, the diagonal of $\mathbf{K}$ being zero.

(a) Taking traces in Rodrigues' formula, which is legitimate because the trace is linear (Theorem 0.9.6 (1)),

$$\operatorname{tr}\mathbf{Q} = 3 + \sin\theta\cdot 0 + (1-\cos\theta)(-2) = 1 + 2\cos\theta.$$

Since cos is injective on $[0,\pi]$, this determines θ.

(b) By Lemma 12.6.3 (a) and (b), $\mathbf{K}^\top = -\mathbf{K}$ and $(\mathbf{K}^2)^\top = \mathbf{K}^2$. Transposing Rodrigues' formula therefore gives $\mathbf{Q}^\top = \mathbf{I}_3 - \sin\theta\,\mathbf{K} + (1-\cos\theta)\mathbf{K}^2$, and subtracting, $\mathbf{Q} - \mathbf{Q}^\top = 2\sin\theta\,\mathbf{K}$.

(c) For $\theta \in (0,\pi)$ we have $\sin\theta > 0$, so (b) may be divided by $2\sin\theta$, and the three independent entries of $\mathbf{K_n}$ are $(\mathbf{K_n})_{32} = n_1$, $(\mathbf{K_n})_{13} = n_2$, and $(\mathbf{K_n})_{21} = n_3$ by Definition 12.6.1; the displayed formula reads these off $\mathbf{Q} - \mathbf{Q}^\top$. If $\theta = 0$ then Rodrigues' formula gives $\mathbf{Q} = \mathbf{I}_3$. If $\theta = \pi$ then $\mathbf{Q} = 2\mathbf{n}\mathbf{n}^\top - \mathbf{I}_3$, as computed in the Quick check above, so $\mathbf{Q} + \mathbf{I}_3 = 2\mathbf{n}\mathbf{n}^\top$, whose i-th column is $2n_i\mathbf{n}$; at least one $n_i \neq 0$, so at least one column is a nonzero multiple of $\mathbf{n}$. This proves the theorem.

■

The two degenerate angles are exactly the two at which the skew part $\mathbf{Q} - \mathbf{Q}^\top$ vanishes, and they are exactly the two at which the axis is not determined by a *signed* angle: at $\theta = 0$ there is no rotation to speak of, and at $\theta = \pi$ turning by $+\pi$ and $-\pi$ about the same axis give the same map. Everywhere else, part (c) is a two-subtraction recipe.

Example. Axis and Angle of a Space Rotation

Let

$$\mathbf{Q} = \tfrac{1}{9}\begin{pmatrix} 1 & -4 & 8 \\ 8 & 4 & 1 \\ -4 & 7 & 4 \end{pmatrix}.$$

Verify that $\mathbf{Q} \in \mathrm{SO}(3)$, find its axis and angle, and confirm the answer with Rodrigues' formula.

Solution. Orthogonality. The columns are $\frac{1}{9}(1,8,-4)$, $\frac{1}{9}(-4,4,7)$, and $\frac{1}{9}(8,1,4)$. Each has squared norm $\frac{1}{81}(1+64+16) = \frac{1}{81}(16+16+49) = \frac{1}{81}(64+1+16) = 1$, and the three pairwise inner products are

$$\tfrac{1}{81}(-4+32-28) = \tfrac{1}{81}(8+8-16) = \tfrac{1}{81}(-32+4+28) = 0.$$

So $\mathbf{Q} \in \mathrm{O}(3)$ by Theorem 11.7.3 (f).

Determinant. Expanding along the first row, $729\det\mathbf{Q} = 1(16-7) - (-4)(32+4) + 8(56+16) = 9 + 144 + 576 = 729$, so $\det\mathbf{Q} = 1$ and $\mathbf{Q} \in \mathrm{SO}(3)$.

Angle. $\operatorname{tr}\mathbf{Q} = \frac{1}{9}(1+4+4) = 1$, so $\cos\theta = \frac{1}{2}(1-1) = 0$ and $\theta = \pi/2$ by Theorem 12.6.5 (a).

Axis. Here $2\sin\theta = 2$. The three entries needed by Theorem 12.6.5 (c) are

$$q_{32} - q_{23} = \tfrac{7-1}{9} = \tfrac{6}{9},$$
$$q_{13} - q_{31} = q_{21} - q_{12} = \tfrac{8+4}{9} = \tfrac{12}{9},$$

so $\mathbf{n} = \frac{1}{2}(\frac{6}{9}, \frac{12}{9}, \frac{12}{9}) = \frac{1}{3}(1,2,2)$, which is a unit vector since $1+4+4 = 9$.

Check. $\mathbf{Q}(1,2,2) = \frac{1}{9}(1-8+16,\ 8+8+2,\ -4+14+8) = \frac{1}{9}(9,18,18) = (1,2,2)$, so the

axis is fixed. And Rodrigues' formula with $\theta = \pi/2$ reads $\mathbf{R}_{\mathbf{n},\pi/2} = \mathbf{I}_3 + \mathbf{K} + \mathbf{K}^2$, where

$$\mathbf{K} = \tfrac{1}{3}\begin{pmatrix} 0 & -2 & 2 \\ 2 & 0 & -1 \\ -2 & 1 & 0 \end{pmatrix}, \qquad \mathbf{K}^2 = \mathbf{n}\mathbf{n}^\top - \mathbf{I}_3 = \tfrac{1}{9}\begin{pmatrix} -8 & 2 & 2 \\ 2 & -5 & 4 \\ 2 & 4 & -5 \end{pmatrix}.$$

Adding $\mathbf{I}_3 + \mathbf{K} + \mathbf{K}^2$ entry by entry, for instance $1 + 0 - \frac{8}{9} = \frac{1}{9}$ in position $(1,1)$ and $0 - \frac{2}{3} + \frac{2}{9} = -\frac{4}{9}$ in position $(1,2)$, returns $\mathbf{Q}$.

12.6.6 Reflections generate

The canonical form is one way to say that an orthogonal matrix is built from simple pieces. Here is another, and the pieces are even simpler: mirrors. Recall from Chapter 11 the Householder reflection $\mathbf{H}_\mathbf{w} = \mathbf{I}_n - 2\mathbf{w}\mathbf{w}^\top/\|\mathbf{w}\|^2$ determined by a nonzero $\mathbf{w}$ (Definition 11.8.1); it is orthogonal, equal to its own inverse, and has determinant -1 (Proposition 11.8.2).

Lemma 12.6.6. A Mirror Between Two Vectors of Equal Length

Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ with $\|\mathbf{x}\| = \|\mathbf{y}\|$ and $\mathbf{x} \neq \mathbf{y}$. Then $\mathbf{w} \coloneqq \mathbf{x} - \mathbf{y}$ is nonzero and $\mathbf{H}_\mathbf{w}\mathbf{x} = \mathbf{y}$.

Proof. $\mathbf{w} \neq \mathbf{0}$ because $\mathbf{x} \neq \mathbf{y}$. Expanding the two inner products and using $\|\mathbf{y}\|^2 = \|\mathbf{x}\|^2$,

$$\begin{aligned} \|\mathbf{w}\|^2 &= \|\mathbf{x}\|^2 - 2\langle \mathbf{x}, \mathbf{y}\rangle + \|\mathbf{y}\|^2 = 2(\|\mathbf{x}\|^2 - \langle \mathbf{x}, \mathbf{y}\rangle), \\ \langle \mathbf{x}, \mathbf{w}\rangle &= \|\mathbf{x}\|^2 - \langle \mathbf{x}, \mathbf{y}\rangle, \end{aligned}$$

so $2\langle \mathbf{x}, \mathbf{w}\rangle = \|\mathbf{w}\|^2$. Hence, by the formula $\mathbf{H}_\mathbf{w}\mathbf{v} = \mathbf{v} - 2(\langle \mathbf{v}, \mathbf{w}\rangle/\|\mathbf{w}\|^2)\mathbf{w}$ of Definition 11.8.1,

$$\mathbf{H}_\mathbf{w}\mathbf{x} = \mathbf{x} - \mathbf{w} = \mathbf{x} - (\mathbf{x} - \mathbf{y}) = \mathbf{y},$$

as claimed. ∎

This is Theorem 11.8.3 with the target chosen for geometry rather than for arithmetic: there, $\mathbf{y}$ was a prescribed multiple of $\mathbf{e}_1$ and the sign was picked to avoid cancellation; here $\mathbf{y}$ is whatever we need, and the mirror is the perpendicular bisector of the segment from $\mathbf{x}$ to $\mathbf{y}$.

Theorem 12.6.7. Reflections Generate the Orthogonal Group

Let $n \geq 1$ and $\mathbf{Q} \in \mathrm{O}(n)$. Then $\mathbf{Q}$ is a product of at most n Householder reflections. (The identity is the empty product, of zero reflections.)

Idea. Induct on n, and make the induction bite by fixing one coordinate at a time. If $\mathbf{Q}$ already fixes $\mathbf{e}_1$, then it maps $\mathbf{e}_1^\perp$ into itself, and the inductive hypothesis applied inside that $(n-1)$-dimensional space produces at most $n-1$ reflections, each of which extends to a reflection of $\mathbb{R}^n$ by taking the same $\mathbf{w}$. If $\mathbf{Q}$ does not fix $\mathbf{e}_1$, spend one reflection making it

do so: $\mathbf{Q}\mathbf{e}_1$ and $\mathbf{e}_1$ are two distinct vectors of the same length, and Lemma 12.6.6 is exactly the tool that swaps them.

Proof. Induction on n. For $n = 1$, a matrix $[a] \in \mathrm{O}(1)$ satisfies $a^2 = [a]^\top[a] = 1$, so $(a-1)(a+1) = 0$ and $a = \pm 1$; thus $\mathrm{O}(1) = \{[1], [-1]\}$, where $[1]$ is the empty product and $[-1] = \mathbf{H}_{\mathbf{e}_1}$, one reflection.

Let $n \geq 2$ and assume the statement in dimension $n-1$. Let $\mathbf{Q} \in \mathrm{O}(n)$ and set $\mathbf{x} = \mathbf{Q}\mathbf{e}_1$, a unit vector since $T_{\mathbf{Q}}$ is an isometry (Theorem 11.7.3).

Case 1: $\mathbf{x} = \mathbf{e}_1$. Put $U = \mathrm{span}(\mathbf{e}_1)^\perp$, of dimension $n-1$ by Theorem 11.3.2 (c). For $\mathbf{v} \in U$ we have $\langle \mathbf{Q}\mathbf{v}, \mathbf{e}_1 \rangle = \langle \mathbf{Q}\mathbf{v}, \mathbf{Q}\mathbf{e}_1 \rangle = \langle \mathbf{v}, \mathbf{e}_1 \rangle = 0$, the middle equality by Theorem 11.7.3 (b); so U is $T_{\mathbf{Q}}$-invariant. Identify U with $\mathbb{R}^{n-1}$ through the orthonormal basis $(\mathbf{e}_2, \dots, \mathbf{e}_n)$. The restriction $T_{\mathbf{Q}}|_U$ is an isometry of U, so its matrix $\mathbf{Q}'$ in that basis lies in $\mathrm{O}(n-1)$ (Theorem 11.7.3 (f)), and by the inductive hypothesis $\mathbf{Q}' = \mathbf{H}_{\mathbf{w}'_1} \cdots \mathbf{H}_{\mathbf{w}'_k}$ with $k \leq n-1$ and $\mathbf{w}'_i \in \mathbb{R}^{n-1}$. Let $\mathbf{w}_i \in U \subseteq \mathbb{R}^n$ be the vector with coordinates $\mathbf{w}'_i$ in the basis $(\mathbf{e}_2, \dots, \mathbf{e}_n)$. Then $\mathbf{H}_{\mathbf{w}_i}$ fixes $\mathbf{e}_1$, because $\mathbf{e}_1 \perp \mathbf{w}_i$ (Proposition 11.8.2 (d)), and on U it acts as $\mathbf{H}_{\mathbf{w}'_i}$, because the defining formula only involves inner products, which agree. Hence $\mathbf{Q}$ and $\mathbf{H}_{\mathbf{w}_1} \cdots \mathbf{H}_{\mathbf{w}_k}$ agree on $\mathbf{e}_1$ and on U, so they agree on $\mathbb{R}^n = \mathrm{span}(\mathbf{e}_1) \oplus U$, and $\mathbf{Q}$ is a product of $k \leq n-1 \leq n$ reflections.

Case 2: $\mathbf{x} \neq \mathbf{e}_1$. Then $\|\mathbf{x}\| = 1 = \|\mathbf{e}_1\|$, so $\mathbf{w} := \mathbf{x} - \mathbf{e}_1$ is nonzero and $\mathbf{H}_{\mathbf{w}}\mathbf{x} = \mathbf{e}_1$ by Lemma 12.6.6. The matrix $\mathbf{H}_{\mathbf{w}}\mathbf{Q}$ is orthogonal, being a product of orthogonal matrices (Proposition 11.7.4 (a)), and $\mathbf{H}_{\mathbf{w}}\mathbf{Q}\mathbf{e}_1 = \mathbf{H}_{\mathbf{w}}\mathbf{x} = \mathbf{e}_1$. By Case 1, $\mathbf{H}_{\mathbf{w}}\mathbf{Q}$ is a product of at most $n-1$ reflections. Multiplying on the left by $\mathbf{H}_{\mathbf{w}}$ and using $\mathbf{H}_{\mathbf{w}}^2 = \mathbf{I}_n$ (Proposition 11.8.2 (b)),

$$\mathbf{Q} = \mathbf{H}_{\mathbf{w}}(\mathbf{H}_{\mathbf{w}}\mathbf{Q}),$$

a product of at most n reflections. This completes the induction. ∎

Counting reflections recovers the determinant: each factor contributes -1 (Proposition 11.8.2 (e)), so a product of k reflections has determinant $(-1)^k$, and an element of $\mathrm{SO}(n)$ needs an even number of them. The bound n is sharp — $-\mathbf{I}_n$ is a product of n reflections and no fewer, since a product of k reflections fixes the intersection of k hyperplanes, a subspace of dimension at least $n-k$, while $-\mathbf{I}_n$ fixes only $\mathbf{0}$.

Remark. The general theorem behind this one, due to Cartan and Dieudonné, says the same for the orthogonal group of any nondegenerate symmetric bilinear form over any field of characteristic $\neq 2$. Only the Euclidean case is proved here, and only the Euclidean case is used.

12.6.7 Composing rotations

Two rotations of the plane about the origin compose into a rotation, and the angles add: $\mathbf{R}_\alpha \mathbf{R}_\beta = \mathbf{R}_{\alpha+\beta}$, so $\mathrm{SO}(2)$ is abelian (Chapter 11). The reason is visible in the canonical form: in dimension 2 there is only one rotation block, so a rotation is a single number, and numbers commute.

In dimension 3 the picture changes completely, because there is an axis to move as well as an angle to add.

Example. Two Quarter Turns in Two Orders

Let $\mathbf{A} = \mathbf{R}_{\mathbf{e}_1,\pi/2}$ and $\mathbf{B} = \mathbf{R}_{\mathbf{e}_3,\pi/2}$, the quarter turns about the first and third coordinate axes. Compute $\mathbf{AB}$ and $\mathbf{BA}$, and find the axis and angle of each.

Solution. From Theorem 12.6.4 (b), $\mathbf{A}$ fixes $\mathbf{e}_1$ and turns the $(\mathbf{e}_2, \mathbf{e}_3)$-plane by $\pi/2$, and $\mathbf{B}$ fixes $\mathbf{e}_3$ and turns the $(\mathbf{e}_1, \mathbf{e}_2)$-plane by $\pi/2$:

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix}, \qquad \mathbf{B} = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Multiplying,

$$\mathbf{AB} = \begin{pmatrix} 0 & -1 & 0 \\ 0 & 0 & -1 \\ 1 & 0 & 0 \end{pmatrix}, \qquad \mathbf{BA} = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}.$$

These are different matrices, so $\mathrm{SO}(3)$ is not abelian. Both have trace 0, so both are rotations by the angle θ with $\cos\theta = \frac{1}{2}(0-1) = -\frac{1}{2}$, that is $\theta = 2\pi/3$. Their axes differ. Solving $\mathbf{ABv} = \mathbf{v}$ gives $-v_2 = v_1$, $-v_3 = v_2$, $v_1 = v_3$, so the axis of $\mathbf{AB}$ is $\mathrm{span}((1,-1,1))$; solving $\mathbf{BAv} = \mathbf{v}$ gives $v_3 = v_1$, $v_1 = v_2$, $v_2 = v_3$, so the axis of $\mathbf{BA}$ is $\mathrm{span}((1,1,1))$. Same angle, different axes: the order of two quarter turns changes where the resulting axis points.

That the product is *again* a rotation is not an accident of this example: $\mathrm{SO}(3)$ is a group (Proposition 11.7.4), so any product of rotations is orthogonal with determinant 1, and Corollary 12.6.2 makes it a rotation. What no formula makes obvious is which one; Exercise (C2: Products of rotations) pursues this.

Warning. **In dimension 4 and above, $\det \mathbf{Q} = 1$ does not mean "rotation about an axis".** Take

$$\mathbf{Q} = \mathbf{R}_{\pi/2} \oplus \mathbf{R}_{\pi/3} \in \mathrm{O}(4), \qquad \det \mathbf{Q} = 1 \cdot 1 = 1.$$

Its canonical form is itself, with $p = q = 0$ and $m = 2$, so 1 is not an eigenvalue and $\mathbf{Q}$ fixes no nonzero vector at all: there is no axis. Worse, the two angles are different, so $\mathbf{Q}$ is not a rotation "in a plane" either — it turns two orthogonal planes at two different speeds, and no single angle describes it. Chapter 11 made the same point with $-\mathbf{I}_4 \in \mathrm{SO}(4)$; the canonical form explains it, since for even n the counts $p = q = 0$ are allowed. The clean statement "determinant 1 means a rotation about an axis" is a feature of dimension 3, where $p + q + 2m = 3$ is odd and forces $p \geq 1$ when q is even.

12.6.8 Exercises

12.6.8.1 A. Check your understanding

Exercise. A1

(a) State the canonical form of an element of $\mathrm{O}(n)$, saying what the blocks are and which angles are allowed.

(b) Which hypothesis on $\mathbf{Q}$ turns the rotation-scaling blocks of the real normal form into rotation blocks, and why?

(c) Write down the cross-product matrix $\mathbf{K_n}$ and state the identity $\mathbf{K_n^3} =?$ for a unit $\mathbf{n}$.

(d) Given $\mathbf{Q} \in \mathrm{SO}(3)$, how do you find its angle, and how do you find its axis? Name the two degenerate cases.

(e) True or false: if $\mathbf{Q} \in \mathrm{O}(n)$ and $\det \mathbf{Q} = 1$, then $\mathbf{Q}$ fixes a nonzero vector. Justify your answer.

(f) How many Householder reflections can be needed to write an element of $\mathrm{O}(n)$, and what does the parity of that number say?

Solution.

(a) By Theorem 12.6.1, $\mathbf{P}^\top \mathbf{Q} \mathbf{P} = \mathbf{I}_p \oplus (-\mathbf{I}_q) \oplus \mathbf{R}_{\theta_1} \oplus \cdots \oplus \mathbf{R}_{\theta_m}$ for some $\mathbf{P} \in \mathrm{O}(n)$, with $p + q + 2m = n$ and $0 < \theta_j < \pi$. The angles 0 and π are excluded because $\mathbf{R}_0 = \mathbf{I}_2$ and $\mathbf{R}_\pi = -\mathbf{I}_2$ are already accounted for by the ± 1 blocks.

(b) Orthogonality. An orthogonal matrix is unitary as a complex matrix (Proposition 11.7.4 (a)), so each of its eigenvalues has modulus 1 (Proposition 11.7.1). A block $\mathbf{\Lambda}(\lambda)$ has eigenvalues $\lambda, \overline{\lambda}$, so $|\lambda| = 1$ and the scaling factor is 1.

(c) $\mathbf{K_n}$ has rows $(0, -n_3, n_2)$, $(n_3, 0, -n_1)$, $(-n_2, n_1, 0)$ (Definition 12.6.1), and $\mathbf{K_n^3} = -\mathbf{K_n}$ when $\|\mathbf{n}\| = 1$ (Lemma 12.6.3 (c)).

(d) The angle from the trace, $\cos\theta = \frac{1}{2}(\operatorname{tr}\mathbf{Q} - 1)$ with $\theta \in [0, \pi]$; the axis from the skew part, $\mathbf{Q} - \mathbf{Q}^\top = 2\sin\theta\, \mathbf{K_n}$ (Theorem 12.6.5). The degenerate cases are $\theta = 0$, where $\mathbf{Q} = \mathbf{I}_3$ and there is no axis to find, and $\theta = \pi$, where the skew part vanishes and one reads $\mathbf{n}$ off $\mathbf{Q} + \mathbf{I}_3 = 2\mathbf{n}\mathbf{n}^\top$ instead.

(e) False for even n. The matrix $\mathbf{R}_{\pi/2} \oplus \mathbf{R}_{\pi/3} \in \mathrm{SO}(4)$ has no eigenvalue 1. It is true for odd n: then $p + q + 2m = n$ is odd and q is even, so $p = n - q - 2m$ is odd, hence $p \geq 1$, and the $\mathbf{I}_p$ block supplies a fixed vector.

(f) At most n (Theorem 12.6.7). Each reflection has determinant -1, so the number used has the same parity in every such factorization as the sign of $\det \mathbf{Q}$ dictates: even for $\det \mathbf{Q} = 1$, odd for $\det \mathbf{Q} = -1$.

12.6.8.2 B. Practice

Exercise. B1: Axis and angle

For each of the following, decide whether it lies in $\mathrm{SO}(3)$, $\mathrm{O}(3) \setminus \mathrm{SO}(3)$, or neither. For those in $\mathrm{SO}(3)$, find the axis and the angle; for those in $\mathrm{O}(3)$ with determinant -1, give the canonical form $[-1] \oplus \mathbf{R}_\theta$ and say what the map is geometrically.

(a) $\frac{1}{9}\begin{pmatrix} 4 & -8 & 1 \\ 4 & 1 & -8 \\ 7 & 4 & 4 \end{pmatrix}$.

(b) $\frac{1}{9}\begin{pmatrix} -7 & 4 & 4 \\ 4 & -1 & 8 \\ 4 & 8 & -1 \end{pmatrix}$.

(c) $\frac{1}{9}\begin{pmatrix} 7 & -4 & -4 \\ -4 & 1 & -8 \\ -4 & -8 & 1 \end{pmatrix}$.

(d) $\frac{1}{9}\begin{pmatrix} 4 & -8 & 1 \\ 4 & 1 & -8 \\ 7 & 4 & 5 \end{pmatrix}$.

Solution.

(a) The columns $\frac{1}{9}(4,4,7)$, $\frac{1}{9}(-8,1,4)$, $\frac{1}{9}(1,-8,4)$ have squared norms $\frac{1}{81}(16+16+49) = \frac{1}{81}(64+1+16) = \frac{1}{81}(1+64+16) = 1$ and pairwise inner products $\frac{1}{81}(-32+4+28) = \frac{1}{81}(4-32+28) = \frac{1}{81}(-8-8+16) = 0$, so the matrix is orthogonal. Expanding along the first row, $729\det = 4(4+32)+8(16+56)+1(16-7) = 144+576+9 = 729$, so it lies in $\mathrm{SO}(3)$. Its trace is $\frac{1}{9}(4+1+4) = 1$, so $\cos\theta = 0$ and $\theta = \pi/2$. For the axis, $q_{32}-q_{23} = \frac{1}{9}(4+8) = \frac{12}{9}$, $q_{13}-q_{31} = \frac{1}{9}(1-7) = -\frac{6}{9}$, and $q_{21}-q_{12} = \frac{1}{9}(4+8) = \frac{12}{9}$; dividing by $2\sin\theta = 2$ gives $\mathbf{n} = \frac{1}{3}(2,-1,2)$. Check: $\mathbf{Q}(2,-1,2) = \frac{1}{9}(8+8+2,\ 8-1-16,\ 14-4+8) = (2,-1,2)$.

(b) The matrix is symmetric. Each column has entries $\pm 7, \pm 4, \pm 4$ in some order, so each has squared norm $\frac{1}{81}(49+16+16) = 1$, and the pairwise inner products are $\frac{1}{81}(-28-4+32) = \frac{1}{81}(-28+32-4) = \frac{1}{81}(16-8-8) = 0$; so it is orthogonal. Expanding along the first row, $729\det = -7(1-64)-4(-4-32)+4(32+4) = 441+144+144 = 729$, so it lies in $\mathrm{SO}(3)$. Its trace is $\frac{1}{9}(-7-1-1) = -1$, so $\cos\theta = -1$ and $\theta = \pi$: a half-turn. The skew part is $\mathbf{O}$, so use $\mathbf{Q}+\mathbf{I}_3 = \frac{1}{9}\begin{pmatrix} 2 & 4 & 4 \\ 4 & 8 & 8 \\ 4 & 8 & 8 \end{pmatrix} = 2\mathbf{n}\mathbf{n}^{\top}$ with $\mathbf{n} = \frac{1}{3}(1,2,2)$. Check: $\mathbf{Q}(1,2,2) = \frac{1}{9}(-7+8+8,\ 4-2+16,\ 4+16-2) = (1,2,2)$.

(c) This is the negative of (b), so it is orthogonal with $\det = (-1)^3 \cdot 1 = -1$. Its trace is 1, and for the form $[-1]\oplus\mathbf{R}_\theta$ the trace is $-1+2\cos\theta$, so $\cos\theta = 1$ and $\theta = 0$: the canonical form is $[-1]\oplus\mathbf{I}_2 = \mathrm{diag}(-1,1,1)$, a reflection. Indeed it equals $\mathbf{I}_3 - 2\mathbf{n}\mathbf{n}^{\top} = \mathbf{H}_{\mathbf{n}}$ with $\mathbf{n} = \frac{1}{3}(1,2,2)$, the reflection in the plane $x_1+2x_2+2x_3 = 0$.

(d) Neither. This is (a) with the $(3,3)$ entry changed from 4 to 5. The third column $\frac{1}{9}(1,-8,5)$ has squared norm $\frac{1}{81}(1+64+25) = \frac{90}{81} = \frac{10}{9} \neq 1$, so the columns are not orthonormal and the matrix is not orthogonal (Theorem 11.7.3 (f)). Its determinant is $\frac{85}{81}$, neither 1 nor -1, which settles it a second way (Proposition 11.7.4). Note that checking the determinant alone is never enough: $\det = 1$ does not make a matrix orthogonal.

Exercise. B2: Mirrors

(a) Write the matrix $\mathbf{M} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}$ as a Householder reflection $\mathbf{H}_{\mathbf{w}}$, and say which plane is its mirror.

(b) Following the proof of Theorem 12.6.7, write $\mathbf{Q} = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$ as a product of Householder reflections, and check the count against $\det \mathbf{Q}$.

Solution.

(a) $\mathbf{M}$ exchanges $\mathbf{e}_1$ and $\mathbf{e}_3$ and fixes $\mathbf{e}_2$. By Lemma 12.6.6 with $\mathbf{x} = \mathbf{e}_1$ and $\mathbf{y} = \mathbf{e}_3$, take $\mathbf{w} = \mathbf{e}_1 - \mathbf{e}_3 = (1, 0, -1)$; then $\|\mathbf{w}\|^2 = 2$ and

$$\mathbf{H}_{\mathbf{w}} = \mathbf{I}_3 - \mathbf{w}\mathbf{w}^{\top} = \mathbf{I}_3 - \begin{pmatrix} 1 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 1 \end{pmatrix} = \mathbf{M}.$$

The mirror is $\operatorname{span}(\mathbf{w})^{\perp}$, the plane $x_1 = x_3$.

(b) $\mathbf{Q}\mathbf{e}_1 = \mathbf{e}_2 \neq \mathbf{e}_1$, so we are in Case 2 with $\mathbf{w}_1 = \mathbf{Q}\mathbf{e}_1 - \mathbf{e}_1 = (-1, 1, 0)$ and $\|\mathbf{w}_1\|^2 = 2$:

$$\mathbf{H}_{\mathbf{w}_1} = \mathbf{I}_3 - \mathbf{w}_1\mathbf{w}_1^{\top} = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Then $\mathbf{H}_{\mathbf{w}_1}\mathbf{Q} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$, which fixes $\mathbf{e}_1$ and exchanges $\mathbf{e}_2$ and $\mathbf{e}_3$; the computation of (a), with the indices shifted, shows that it is $\mathbf{H}_{\mathbf{w}_2}$ for $\mathbf{w}_2 = \mathbf{e}_2 - \mathbf{e}_3 = (0, 1, -1)$. Hence $\mathbf{Q} = \mathbf{H}_{\mathbf{w}_1}\mathbf{H}_{\mathbf{w}_2}$, a product of two reflections. The count is even, matching $\det \mathbf{Q} = (-1)^2 = 1$; and indeed $\mathbf{Q}$ is the rotation by $2\pi/3$ about $(1, 1, 1)$ found in Example (Two Quarter Turns in Two Orders).

Exercise. B3: Rodrigues in an example

(a) Compute $\mathbf{R}_{\mathbf{e}_3,\theta}$ from Rodrigues' formula and check that it is $\mathbf{R}_{\theta} \oplus [1]$.

(b) Compute $\mathbf{R}_{\mathbf{n},2\pi/3}$ for $\mathbf{n} = \frac{1}{\sqrt{3}}(1, 1, 1)$, and identify the resulting matrix.

Solution.

(a) With $\mathbf{n} = \mathbf{e}_3$ we have $\mathbf{K} = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$ and $\mathbf{K}^2 = \mathbf{e}_3\mathbf{e}_3^{\top} - \mathbf{I}_3 = \operatorname{diag}(-1, -1, 0)$. Hence

$$\mathbf{R}_{\mathbf{e}_3,\theta} = \mathbf{I}_3 + \sin\theta\,\mathbf{K} + (1 - \cos\theta)\mathbf{K}^2 = \begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

the diagonal entries being $1 - (1 - \cos\theta) = \cos\theta$ twice and 1 once. This is $\mathbf{R}_{\theta} \oplus [1]$, the rotation of the $(\mathbf{e}_1, \mathbf{e}_2)$-plane, as it must be.

(b) Here $\cos\theta = -\frac{1}{2}$ and $\sin\theta = \frac{\sqrt{3}}{2}$. Writing $\mathbf{1} = (1, 1, 1)$, so that $\mathbf{n} = \mathbf{1}/\sqrt{3}$ and

$\mathbf{n}\mathbf{n}^\top = \frac{1}{3}\mathbf{1}\mathbf{1}^\top$,

$$\sin\theta\,\mathbf{K} = \tfrac{\sqrt{3}}{2}\cdot\tfrac{1}{\sqrt{3}}\,\mathbf{K_1} = \tfrac{1}{2}\begin{pmatrix} 0 & -1 & 1 \\ 1 & 0 & -1 \\ -1 & 1 & 0 \end{pmatrix},$$

$$(1-\cos\theta)\mathbf{K}^2 = \tfrac{3}{2}\left(\tfrac{1}{3}\mathbf{1}\mathbf{1}^\top - \mathbf{I}_3\right) = \tfrac{1}{2}\mathbf{1}\mathbf{1}^\top - \tfrac{3}{2}\mathbf{I}_3.$$

Adding $\mathbf{I}_3$ to these two gives $\frac{1}{2}\mathbf{1}\mathbf{1}^\top - \frac{1}{2}\mathbf{I}_3$ plus the skew part. The first piece has 0 on the diagonal and $\frac{1}{2}$ elsewhere, and adding the skew part turns each off-diagonal pair $\frac{1}{2}, \frac{1}{2}$ into $0, 1$:

$$\mathbf{R}_{\mathbf{n},2\pi/3} = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}.$$

This is the cyclic shift $\mathbf{e}_1 \mapsto \mathbf{e}_2 \mapsto \mathbf{e}_3 \mapsto \mathbf{e}_1$, which is the matrix $\mathbf{BA}$ of Example (Two Quarter Turns in Two Orders): same axis $(1,1,1)$, same angle $2\pi/3$.

12.6.8.3 C. Going deeper

Exercise. C1: The plane, re-derived

(a) Use Theorem 12.6.1 to prove that every $\mathbf{Q} \in \mathrm{O}(2)$ is either orthogonally similar to $\mathbf{R}_\theta$ for a unique $\theta \in [0, \pi]$, or orthogonally similar to $\mathrm{diag}(1, -1)$.

(b) Chapter 11 proved more: $\mathbf{Q}$ **equals** $\mathbf{R}_\theta$ or $\mathbf{M}_\theta$ (Theorem 11.7.6). Explain exactly what information (a) loses, and why the canonical form must lose it.

(c) Deduce that any two reflections of $\mathbb{R}^2$ are orthogonally similar, and that $\mathbf{R}_\alpha$ and $\mathbf{R}_\beta$ are orthogonally similar if and only if $\beta = \pm\alpha$.

Solution.

(a) With $n = 2$, $p + q + 2m = 2$. If $m = 1$ then $p = q = 0$ and the form is $\mathbf{R}_{\theta_1}$ with $\theta_1 \in (0, \pi)$. If $m = 0$ then $(p, q) \in \{(2, 0), (1, 1), (0, 2)\}$, giving $\mathbf{I}_2 = \mathbf{R}_0$, $\mathrm{diag}(1, -1)$, and $-\mathbf{I}_2 = \mathbf{R}_\pi$. So the possibilities are $\mathbf{R}_\theta$ with $\theta \in [0, \pi]$ and $\mathrm{diag}(1, -1)$, and they are distinguished by the determinant. Uniqueness of θ follows from Theorem 12.6.1, or directly from $\operatorname{tr}\mathbf{R}_\theta = 2\cos\theta$ and the injectivity of cos on $[0, \pi]$.

(b) Part (a) determines $\mathbf{Q}$ only up to orthogonal similarity, that is, up to a change of orthonormal coordinates. It therefore cannot see *which* line is the mirror of a reflection, nor the sign of the angle of a rotation. It must lose the mirror, because the mirror moves: if $\mathbf{P}$ is the rotation by φ then $\mathbf{P}\mathbf{M}_\theta\mathbf{P}^\top = \mathbf{M}_{\theta+2\varphi}$, so all the $\mathbf{M}_\theta$ are conjugate. It must lose the sign of the angle, because conjugating by $\mathrm{diag}(1, -1)$ sends $\mathbf{R}_\theta$ to $\mathbf{R}_{-\theta}$.

(c) Every reflection has determinant -1, so by (a) each is orthogonally similar to $\mathrm{diag}(1, -1)$, and orthogonal similarity is an equivalence relation, so any two of them are similar to each other. For rotations: if $\beta = \pm\alpha$ then $\mathbf{R}_\beta = \mathbf{R}_\alpha$ or $\mathbf{R}_\beta = \mathrm{diag}(1, -1)\mathbf{R}_\alpha\,\mathrm{diag}(1, -1)$, so they are orthogonally similar. Conversely, similar matrices have equal traces (Theorem 4.2.5), so $\cos\alpha = \cos\beta$, which for real angles means $\beta \equiv \pm\alpha$ modulo 2π.

Exercise. C2: Products of rotations

(a) Prove that the product of two rotations of $\mathbb{R}^3$ is a rotation of $\mathbb{R}^3$, and explain which theorem supplies the axis.

(b) Let $\mathbf{A}$ and $\mathbf{B}$ be the half-turns about two **perpendicular** axes $\operatorname{span}(\mathbf{a})$ and $\operatorname{span}(\mathbf{b})$, with $\mathbf{a} \perp \mathbf{b}$ unit vectors. Prove that $\mathbf{AB}$ is the half-turn about the axis $\operatorname{span}(\mathbf{a} \times \mathbf{b})$.

(c) Deduce that every rotation of $\mathbb{R}^3$ is a product of two half-turns.

Hint for (c): use Theorem 12.6.7, or the description of a half-turn in the Quick check after Theorem 12.6.4.

Solution.

(a) $\mathrm{SO}(3)$ is a group (Proposition 11.7.4 (c)), so if $\mathbf{A}, \mathbf{B} \in \mathrm{SO}(3)$ then $\mathbf{AB} \in \mathrm{SO}(3)$. By Corollary 12.6.2, $\mathbf{AB}$ fixes a line pointwise and rotates the perpendicular plane; that corollary is what supplies the axis, and it does so through the count $p \geq 1$ in the canonical form.

(b) Extend $(\mathbf{a}, \mathbf{b})$ to an orthonormal basis $(\mathbf{a}, \mathbf{b}, \mathbf{c})$ of $\mathbb{R}^3$, where we may take $\mathbf{c} = \mathbf{a} \times \mathbf{b}$: it is a unit vector orthogonal to $\mathbf{a}$ and $\mathbf{b}$ by Lemma 12.6.3 (d). By the Quick check after Theorem 12.6.4, $\mathbf{A} = 2\mathbf{a}\mathbf{a}^{\top} - \mathbf{I}_3$ fixes $\mathbf{a}$ and negates $\mathbf{a}^{\perp}$, so in the basis $(\mathbf{a}, \mathbf{b}, \mathbf{c})$, $\mathbf{A}$ is $\operatorname{diag}(1, -1, -1)$ and $\mathbf{B}$ is $\operatorname{diag}(-1, 1, -1)$. Their product is $\operatorname{diag}(-1, -1, 1)$, which fixes $\mathbf{c}$ and negates $\mathbf{c}^{\perp}$: the half-turn about $\operatorname{span}(\mathbf{c}) = \operatorname{span}(\mathbf{a} \times \mathbf{b})$. (Its trace is -1, confirming $\theta = \pi$ through Theorem 12.6.5 (a).)

(c) Let $\mathbf{Q} = \mathbf{R}_{\mathbf{n},\theta} \in \mathrm{SO}(3)$, with $\mathbf{n}$ a unit vector and $\theta \in [0, \pi]$ (Theorem 12.6.4 (c)). Choose a unit $\mathbf{w} \perp \mathbf{n}$, and work throughout in the orthonormal basis $\mathcal{B} = (\mathbf{n}, \mathbf{w}, \mathbf{K}_{\mathbf{n}}\mathbf{w})$ of Theorem 12.6.4 (b), in which $\mathbf{Q}$ has matrix $[1] \oplus \mathbf{R}_{\theta}$. Put $\mathbf{w}' = \mathbf{R}_{\mathbf{n},\theta/2}\mathbf{w}$, again a unit vector perpendicular to $\mathbf{n}$, since $\mathbf{R}_{\mathbf{n},\theta/2}$ fixes $\mathbf{n}$ and is an isometry; by Theorem 12.6.4 (b) its coordinate vector in $\mathcal{B}$ is $\mathbf{d} = (0, \cos\frac{\theta}{2}, \sin\frac{\theta}{2})$. Let $\mathbf{A}$ and $\mathbf{B}$ be the half-turns about $\operatorname{span}(\mathbf{w}')$ and $\operatorname{span}(\mathbf{w})$, so that in $\mathcal{B}$ their matrices are $2\mathbf{d}\mathbf{d}^{\top} - \mathbf{I}_3$ and $\operatorname{diag}(-1, 1, -1)$ respectively, by the Quick check after Theorem 12.6.4. Using $2\cos^2\frac{\theta}{2} - 1 = \cos\theta$, $2\sin^2\frac{\theta}{2} - 1 = -\cos\theta$, and $2\cos\frac{\theta}{2}\sin\frac{\theta}{2} = \sin\theta$,

$$2\mathbf{d}\mathbf{d}^{\top} - \mathbf{I}_3 = [-1] \oplus \begin{pmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{pmatrix}.$$

Multiplying the two matrices blockwise, $\mathbf{AB}$ has matrix

$$[1] \oplus \begin{pmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} = [1] \oplus \mathbf{R}_{\theta},$$

which is the matrix of $\mathbf{Q}$ in the same basis. Hence $\mathbf{Q} = \mathbf{AB}$.

Exercise. C3: Moving one point of the sphere to another

Let $S = \{\mathbf{x} \in \mathbb{R}^3 : \|\mathbf{x}\| = 1\}$.

(a) Prove that for all $\mathbf{u}, \mathbf{v} \in S$ there is $\mathbf{Q} \in \mathrm{SO}(3)$ with $\mathbf{Q}\mathbf{u} = \mathbf{v}$. (One says that $\mathrm{SO}(3)$ **acts transitively** on S.)

(b) Prove that $\{\mathbf{Q} \in \mathrm{SO}(3) : \mathbf{Q}\mathbf{u} = \mathbf{u}\} = \{\mathbf{R}_{\mathbf{u},\theta} : \theta \in \mathbb{R}\}$, and that this set is an abelian

subgroup of SO(3).

(c) Is the analogue of (a) true for SO(2) acting on the unit circle of $\mathbb{R}^2$? And for SO(4) acting on the unit sphere of $\mathbb{R}^4$? Justify your answers.

Solution.

(a) Extend $\mathbf{u}$ to an orthonormal basis $(\mathbf{u}, \mathbf{a}_2, \mathbf{a}_3)$ of $\mathbb{R}^3$ (Corollary 11.2.3), and likewise extend $\mathbf{v}$ to $(\mathbf{v}, \mathbf{b}_2, \mathbf{b}_3)$. Let $\mathbf{A} = (\mathbf{u} \mid \mathbf{a}_2 \mid \mathbf{a}_3)$ and $\mathbf{B} = (\mathbf{v} \mid \mathbf{b}_2 \mid \mathbf{b}_3)$; both are orthogonal, since their columns are orthonormal (Theorem 11.7.3 (f)). Replacing $\mathbf{a}_3$ by $-\mathbf{a}_3$ changes the sign of $\det \mathbf{A}$ and keeps the basis orthonormal, so we may assume $\det \mathbf{A} = 1$, and likewise $\det \mathbf{B} = 1$. Put $\mathbf{Q} = \mathbf{B}\mathbf{A}^\top$. Then $\mathbf{Q} \in \mathrm{O}(3)$ and $\det \mathbf{Q} = \det \mathbf{B} \det \mathbf{A}^\top = 1$ by Theorem 7.5.4 and Theorem 7.4.5, so $\mathbf{Q} \in \mathrm{SO}(3)$; and $\mathbf{A}^\top \mathbf{u} = \mathbf{e}_1$, because the entries of $\mathbf{A}^\top \mathbf{u}$ are the inner products of $\mathbf{u}$ with the columns of $\mathbf{A}$, so $\mathbf{Q}\mathbf{u} = \mathbf{B}\mathbf{e}_1 = \mathbf{v}$.

(b) ($\supseteq$) Each $\mathbf{R}_{\mathbf{u},\theta}$ lies in SO(3) and fixes $\mathbf{u}$, by Theorem 12.6.4 (b). ($\subseteq$) If $\mathbf{Q} \in \mathrm{SO}(3)$ fixes $\mathbf{u}$, write $\mathbf{Q} = \mathbf{R}_{\mathbf{n},\theta}$ with $\theta \in [0, \pi]$ (Theorem 12.6.4 (c)). If $\theta = 0$ then $\mathbf{Q} = \mathbf{I}_3 = \mathbf{R}_{\mathbf{u},0}$. Otherwise the fixed space of $\mathbf{Q}$ is exactly the line $\operatorname{span}(\mathbf{n})$, since in the basis of Theorem 12.6.4 (b) the matrix is $[1] \oplus \mathbf{R}_\theta$ and $\mathbf{R}_\theta$ fixes no nonzero vector for $\theta \in (0, \pi]$; so $\mathbf{u} = \pm\mathbf{n}$, and $\mathbf{R}_{-\mathbf{n},\theta} = \mathbf{R}_{\mathbf{n},-\theta}$ gives $\mathbf{Q} = \mathbf{R}_{\mathbf{u},\pm\theta}$. For the group claim: all these matrices are simultaneously $[1] \oplus \mathbf{R}_\theta$ in the *same* orthonormal basis $(\mathbf{u}, \mathbf{w}, \mathbf{K}_\mathbf{u}\mathbf{w})$, and $\mathbf{R}_\alpha \mathbf{R}_\beta = \mathbf{R}_{\alpha+\beta} = \mathbf{R}_\beta \mathbf{R}_\alpha$, so the set is closed under products and inverses and is abelian.

(c) Yes for SO(2): given unit $\mathbf{u}, \mathbf{v} \in \mathbb{R}^2$, write $\mathbf{u} = (\cos\alpha, \sin\alpha)$ and $\mathbf{v} = (\cos\beta, \sin\beta)$ and take $\mathbf{Q} = \mathbf{R}_{\beta-\alpha}$, which is in SO(2) and sends $\mathbf{u}$ to $\mathbf{v}$. Yes for SO(4) as well, by the same argument as (a): extend $\mathbf{u}$ and $\mathbf{v}$ to orthonormal bases of $\mathbb{R}^4$, fix the determinants to 1 by flipping the sign of the last vector if necessary, and take $\mathbf{Q} = \mathbf{B}\mathbf{A}^\top$. Nothing in that argument used the dimension 3 — transitivity holds for $\mathrm{SO}(n)$ with $n \geq 2$. What is special to $n = 3$ is the classification of the maps themselves, not their ability to move one unit vector to another.

12.7 Commuting Families

Chapter 9 asked when several operators can be diagonalized in one basis and answered: exactly when they commute, provided each of them is diagonalizable on its own. With an inner product in hand the question sharpens in the same way the spectral theorem sharpened diagonalization. Can one **orthonormal** basis diagonalize all of them at once? For normal operators over $\mathbb{C}$ the answer is yes, and the proof is the chapter's signature move applied family-wide: cut the space into the eigenspaces of one member and recurse.

Throughout, V is a finite-dimensional inner product space over $F = \mathbb{R}$ or $\mathbb{C}$, and a family $\mathcal{F} \subseteq \mathcal{L}(V)$ is called **commuting** if $ST = TS$ for all $S, T \in \mathcal{F}$, as in Chapter 9.

12.7.1 What Chapter 9 already gives, and what is missing

Three results from Chapter 9 do the bookkeeping, and it is worth being exact about what each one delivers.

- Theorem 9.9.1: if $ST = TS$ then every eigenspace $E_\lambda(T)$ is S-invariant. No hypothesis at all beyond commuting — not even finite dimension. This is what lets us look for common eigenvectors *inside* an eigenspace.
- Theorem 9.9.3: a commuting family whose characteristic polynomials all split has a common eigenvector. Over $\mathbb{C}$ the splitting is automatic.
- Theorem 9.9.5 and Theorem 9.9.6: a commuting family of **diagonalizable** operators is simultaneously diagonalizable (Definition 9.9.1), and conversely.

Two things are missing from that list. First, diagonalizability is a hypothesis there, and it has to be checked separately for each member. Second, and more to the point here, the common basis produced is just a basis: nothing says its vectors are orthogonal, and in general they are not.

Normality repairs both at once, over $\mathbb{C}$. It repairs the first because a normal operator is automatically diagonalizable, by Theorem 12.4.2. It repairs the second because that same theorem produces an *orthonormal* eigenbasis, so the pieces we cut the space into are mutually orthogonal.

One more ingredient is needed before the recursion can run: after cutting V into the eigenspaces of one member, the other members must still be normal on each piece. Section 3 supplies exactly this. An invariant subspace of a normal operator is automatically invariant under the adjoint as well, and the restriction to it is again normal: that is Proposition 12.3.6, proved there by a trace computation on the block form. Over $\mathbb{R}$ the self-adjoint version is Theorem 12.2.4. (What fails, as the warning of Section 3 records, is the analogous statement for a subspace that is **not** invariant: a principal submatrix of a normal matrix need not be normal.) With that recalled, nothing stands in the way.

12.7.2 One orthonormal basis for the whole family

Theorem 12.7.1. One Orthonormal Eigenbasis for a Commuting Family

Let $V \neq \{0\}$ be a finite-dimensional **complex** inner product space and let $\mathcal{F} \subseteq \mathcal{L}(V)$ be a nonempty commuting family of **normal** operators. Then V has an orthonormal basis $\mathcal{B}$ each of whose vectors is an eigenvector of every $T \in \mathcal{F}$; equivalently, $[T]_{\mathcal{B}}^{\mathcal{B}}$ is diagonal for every $T \in \mathcal{F}$.

Conversely, if some orthonormal basis $\mathcal{B}$ makes every $[T]_{\mathcal{B}}^{\mathcal{B}}$ diagonal, then $\mathcal{F}$ is commuting and every member of $\mathcal{F}$ is normal.

Idea. Induction on $\dim V$, with the two cases that every such induction has. If every member of the family is a scalar multiple of the identity, any orthonormal basis will do and there is nothing to arrange. Otherwise some $T_0 \in \mathcal{F}$ has at least two distinct eigenvalues, and Theorem 12.4.2 splits V into the eigenspaces of T_0 — **orthogonally**, and into pieces strictly smaller than V. Each piece is invariant under the whole family because the operators commute, and the restriction of a normal operator to an invariant subspace is normal, so each restricted family is again a commuting family of normal operators on a smaller space. Run the induction on each piece and glue.

Proof. Induction on $n = \dim V \geq 1$.

If $n = 1$, any unit vector v of V is a basis, and $Tv \in V = \operatorname{span}(v)$ makes v an eigenvector of

every $T \in \mathcal{L}(V)$.

Let $n \geq 2$ and assume the theorem for all complex inner product spaces of dimension less than n.

Case 1: every $T \in \mathcal{F}$ is a scalar multiple of id_V. Then every nonzero vector is an eigenvector of every $T \in \mathcal{F}$, and V has an orthonormal basis by Theorem 11.2.2.

Case 2: some $T_0 \in \mathcal{F}$ is not a scalar multiple of id_V. Since T_0 is normal, Theorem 12.4.2 gives an orthonormal basis $\mathcal{C} = (c_1, \ldots, c_n)$ of V with $T_0 c_j = \mu_j c_j$. Let $\lambda_1, \ldots, \lambda_k$ be the distinct values among $\mu_1, \ldots, \mu_n$, and put

$$V_i = \operatorname{span}\{c_j : \mu_j = \lambda_i\}, \qquad i = 1, \ldots, k.$$

The V_i are spanned by disjoint sub-lists of an orthonormal basis, so they are pairwise orthogonal, $V = V_1 \oplus \cdots \oplus V_k$, and $\dim V_1 + \cdots + \dim V_k = n$.

Claim. $V_i = E_{\lambda_i}(T_0)$ for each i.

Proof. The inclusion $V_i \subseteq E_{\lambda_i}(T_0)$ holds because every spanning vector c_j of V_i satisfies $T_0 c_j = \lambda_i c_j$. Conversely let $v \in E_{\lambda_i}(T_0)$ and write $v = v_1 + \cdots + v_k$ with $v_j \in V_j$. Applying T_0, which acts as λ_j on V_j, gives $\sum_j \lambda_j v_j = T_0 v = \lambda_i v = \sum_j \lambda_i v_j$, so $\sum_j (\lambda_j - \lambda_i) v_j = 0$. The sum being direct, each term vanishes, and $\lambda_j \neq \lambda_i$ for $j \neq i$ forces $v_j = 0$. Hence $v = v_i \in V_i$. ∎

Because T_0 is not a scalar multiple of id_V, we have $k \geq 2$: if $k = 1$ then $V = V_1 = E_{\lambda_1}(T_0)$ and $T_0 = \lambda_1 \mathrm{id}_V$. Hence $\dim V_i < n$ for every i.

Fix i and let $S \in \mathcal{F}$. By Theorem 9.9.1, $V_i = E_{\lambda_i}(T_0)$ is S-invariant, and therefore $S|_{V_i}$ is a normal operator on V_i by Proposition 12.3.6. Moreover the family

$$\mathcal{F}_i = \{\, S|_{V_i} : S \in \mathcal{F} \,\}$$

is commuting, since $(S|_{V_i})(T|_{V_i}) = (ST)|_{V_i} = (TS)|_{V_i} = (T|_{V_i})(S|_{V_i})$ for $S, T \in \mathcal{F}$, all four restrictions being defined because V_i is invariant.

By the inductive hypothesis applied to $V_i \neq \{0\}$ and $\mathcal{F}_i$, there is an orthonormal basis $\mathcal{B}_i$ of V_i whose vectors are eigenvectors of every $S|_{V_i}$; such a vector $v \in V_i$ satisfies $Sv = (S|_{V_i})v = \mu v$, so it is an eigenvector of S itself. Let $\mathcal{B}$ be the concatenation $\mathcal{B}_1, \ldots, \mathcal{B}_k$. It is orthonormal, because each $\mathcal{B}_i$ is and because $V_i \perp V_j$ for $i \neq j$, and it has $\sum_i \dim V_i = n$ vectors, so it is an orthonormal basis of V (Theorem 11.1.4 and Theorem 1.6.4). Every vector of $\mathcal{B}$ is an eigenvector of every $T \in \mathcal{F}$. This completes the induction.

For the converse, suppose every $[T]_{\mathcal{B}}^{\mathcal{B}}$ is diagonal with $\mathcal{B}$ orthonormal. Diagonal matrices commute, so $[ST]_{\mathcal{B}}^{\mathcal{B}} = [TS]_{\mathcal{B}}^{\mathcal{B}}$ and hence $ST = TS$ (Theorem 4.1.4). And $[T^*]_{\mathcal{B}}^{\mathcal{B}} = ([T]_{\mathcal{B}}^{\mathcal{B}})^*$ by Theorem 11.6.3 is diagonal as well, so $[T^*T]_{\mathcal{B}}^{\mathcal{B}}$ and $[TT^*]_{\mathcal{B}}^{\mathcal{B}}$ are products of diagonal matrices in the two orders, hence equal; so T is normal. This proves the theorem. ∎

Notice which hypothesis did which job. *Commuting* made the eigenspaces of T_0 invariant under the rest of the family; *normal* made those eigenspaces orthogonal and made the restrictions normal again; and *complex* is what makes T_0 have any eigenvalues at all. Drop any one of the three and the argument stops.

In matrix language:

Corollary 12.7.2. Commuting Normal Matrices

Let $\mathbf{A}_1, \dots, \mathbf{A}_r \in M_n(\mathbb{C})$ be normal and pairwise commuting. Then there is a unitary $\mathbf{U} \in \mathrm{U}(n)$ such that $\mathbf{U}^*\mathbf{A}_i\mathbf{U}$ is diagonal for every i. The same holds for an arbitrary commuting family of normal matrices in $M_n(\mathbb{C})$.

Proof. Apply Theorem 12.7.1 to the family $\{T_{\mathbf{A}_1}, \dots, T_{\mathbf{A}_r}\}$ on $\mathbb{C}^n$. Each $T_{\mathbf{A}_i}$ is normal because $T_{\mathbf{A}_i}^* = T_{\mathbf{A}_i^*}$ (Theorem 11.6.3, the standard basis being orthonormal) and $\mathbf{A}_i^*\mathbf{A}_i = \mathbf{A}_i\mathbf{A}_i^*$; and the operators commute because the matrices do. Let $\mathbf{U}$ have the resulting orthonormal basis $\mathcal{B}$ as its columns; then $\mathbf{U} \in \mathrm{U}(n)$ by Theorem 11.7.3 (f), and $\mathbf{U}^*\mathbf{A}_i\mathbf{U} = \mathbf{U}^{-1}\mathbf{A}_i\mathbf{U} = [T_{\mathbf{A}_i}]_{\mathcal{B}}^{\mathcal{B}}$ is diagonal by Theorem 4.2.3. Nothing in the argument used that the family is finite. ∎

Corollary 12.7.3. Commuting Symmetric Matrices

Let V be a finite-dimensional **real** inner product space and let $\mathcal{F} \subseteq \mathcal{L}(V)$ be a nonempty commuting family of self-adjoint operators. Then V has an orthonormal basis of common eigenvectors. Equivalently, if $\mathbf{A}_1, \dots, \mathbf{A}_r \in M_n(\mathbb{R})$ are symmetric and pairwise commuting, there is $\mathbf{Q} \in \mathrm{O}(n)$ with every $\mathbf{Q}^{\top}\mathbf{A}_i\mathbf{Q}$ diagonal.

Proof. Repeat the proof of Theorem 12.7.1 word for word, with two substitutions: use Theorem 12.5.1 in place of Theorem 12.4.2 to split V into the eigenspaces of a member T_0 that is not a scalar multiple of id_V, and use Theorem 12.2.4 (b) in place of Proposition 12.3.6 to see that each restriction $S|_{V_i}$ is again self-adjoint. Every other step is unchanged, and the only property of $\mathbb{C}$ used in the original was the existence of an eigenvalue, which Theorem 12.5.1 supplies over $\mathbb{R}$ from self-adjointness. The matrix statement follows as in Corollary 12.7.2, using Corollary 12.5.2 and $\mathbf{Q}^{\top} = \mathbf{Q}^{-1}$. ∎

Example. Two Commuting Symmetric Matrices, Diagonalized Together

Let

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}, \qquad \mathbf{B} = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{pmatrix} \in M_3(\mathbb{R}).$$

Check that $\mathbf{AB} = \mathbf{BA}$ and find $\mathbf{Q} \in \mathrm{O}(3)$ with $\mathbf{Q}^{\top}\mathbf{AQ}$ and $\mathbf{Q}^{\top}\mathbf{BQ}$ both diagonal.

Solution. They commute. Multiplying out,

$$\mathbf{AB} = \begin{pmatrix} 0 & 2 & 2 \\ 2 & 1 & 1 \\ 2 & 1 & 1 \end{pmatrix} = \mathbf{BA}.$$

Both matrices are symmetric, so Corollary 12.7.3 applies.

Diagonalize $\mathbf{A}$. Each row of $\mathbf{A}$ sums to 2, so $\mathbf{A}(1,1,1) = (2,2,2)$ and 2 is an eigenvalue with eigenvector $(1,1,1)$. Also $\mathbf{A} = \mathbf{1}\mathbf{1}^{\top} - \mathbf{I}_3$, where $\mathbf{1} = (1,1,1)$; any $\mathbf{v}$ with $\mathbf{1}^{\top}\mathbf{v} = 0$, that is with $v_1 + v_2 + v_3 = 0$, satisfies $\mathbf{Av} = -\mathbf{v}$. So $E_2(\mathbf{A}) = \operatorname{span}((1,1,1))$ and $E_{-1}(\mathbf{A})$ is the plane $v_1 + v_2 + v_3 = 0$, and these account for all of $\mathbb{R}^3$.

The one-dimensional eigenspace. By Theorem 9.9.1 the line $E_2(\mathbf{A})$ is $\mathbf{B}$-invariant, so its spanning vector must already be an eigenvector of $\mathbf{B}$. Indeed $\mathbf{B}(1,1,1) = (2,2,2)$.

The plane. Inside $E_{-1}(\mathbf{A})$ we must search. The matrix $\mathbf{B}$ has $\mathbf{B}(0,1,-1) = (0,0,0)$ and $\mathbf{B}(-2,1,1) = (-4,2,2) = 2(-2,1,1)$, and both $(0,1,-1)$ and $(-2,1,1)$ have coordinate sum 0, so they lie in $E_{-1}(\mathbf{A})$. They are orthogonal, and independent, so they are a basis of the plane.

The common basis. Normalizing the three vectors,

$$\mathbf{Q} = \begin{pmatrix} 1/\sqrt{3} & -2/\sqrt{6} & 0 \\ 1/\sqrt{3} & 1/\sqrt{6} & 1/\sqrt{2} \\ 1/\sqrt{3} & 1/\sqrt{6} & -1/\sqrt{2} \end{pmatrix},$$

whose columns are orthonormal, so $\mathbf{Q} \in \mathrm{O}(3)$ by Theorem 11.7.3 (f). Then

$$\mathbf{Q}^{\top}\mathbf{AQ} = \operatorname{diag}(2,-1,-1), \qquad \mathbf{Q}^{\top}\mathbf{BQ} = \operatorname{diag}(2,2,0),$$

the diagonal entries being the eigenvalues found above, column by column. Neither matrix alone forces the choice: $\mathbf{A}$ cannot tell the second column from the third, and $\mathbf{B}$ cannot tell the first from the second. Together they pin down all three lines.

Warning. **Neither hypothesis can be dropped.**

Commuting is not enough. The matrices $\mathbf{A} = \begin{pmatrix} 1 & 1 \\ 0 & 2 \end{pmatrix}$ and $\mathbf{B} = \begin{pmatrix} 1 & 3 \\ 0 & 4 \end{pmatrix}$ satisfy $\mathbf{AB} = \begin{pmatrix} 1 & 7 \\ 0 & 8 \end{pmatrix} = \mathbf{BA}$, and each is diagonalizable, having two distinct eigenvalues. They even share the eigenbasis $\mathbf{e}_1, (1,1)$: $\mathbf{A}(1,1) = 2(1,1)$ and $\mathbf{B}(1,1) = 4(1,1)$. But that basis is not orthogonal, and no orthonormal one exists, because an orthonormal eigenbasis would make $\mathbf{A}$ normal (Corollary 12.4.3), whereas

$$\mathbf{A}^{*}\mathbf{A} = \begin{pmatrix} 1 & 1 \\ 1 & 5 \end{pmatrix} \neq \begin{pmatrix} 2 & 2 \\ 2 & 4 \end{pmatrix} = \mathbf{A}\mathbf{A}^{*}.$$

Normal is not enough. The matrices $\mathbf{X} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ and $\mathbf{Z} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$ are real symmetric,

hence normal, and each is orthogonally diagonalizable on its own. They do not commute:

$$\mathbf{XZ} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \neq \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} = \mathbf{ZX}.$$

In fact they share no eigenvector at all: the eigenvectors of $\mathbf{Z}$ are the multiples of $\mathbf{e}_1$ and of $\mathbf{e}_2$, and $\mathbf{X}\mathbf{e}_1 = \mathbf{e}_2$ is a multiple of neither.

12.7.3 Sums and products stay normal

Normality is not preserved by addition or multiplication. It is preserved by both as soon as the two operators commute, and the common eigenbasis makes this a one-line matter.

Theorem 12.7.4. Sums and Products of Commuting Normal Operators

Let V be a finite-dimensional complex inner product space and let $S, T \in \mathcal{L}(V)$ be normal with $ST = TS$. Then $S + T$ and ST are normal.

Proof. If $V = \{0\}$ there is nothing to prove. Otherwise apply Theorem 12.7.1 to the family $\{S, T\}$: there is an orthonormal basis $\mathcal{B}$ with $[S]_{\mathcal{B}}^{\mathcal{B}} = \mathbf{D}$ and $[T]_{\mathcal{B}}^{\mathcal{B}} = \mathbf{D}'$ both diagonal. Then $[S+T]_{\mathcal{B}}^{\mathcal{B}} = \mathbf{D} + \mathbf{D}'$ and $[ST]_{\mathcal{B}}^{\mathcal{B}} = \mathbf{D}\mathbf{D}'$ are diagonal, so every vector of $\mathcal{B}$ is an eigenvector of both operators, and both are normal by Theorem 12.4.2. ∎

Remark. Commutativity is essential, not a convenience. Take $\mathbf{M} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$, which is symmetric, and $\mathbf{N} = \operatorname{diag}(1, 2i)$, which is diagonal; both are normal and they do not commute. Then

$$\mathbf{M} + \mathbf{N} = \begin{pmatrix} 1 & 1 \\ 1 & 2i \end{pmatrix}, \qquad \mathbf{MN} = \begin{pmatrix} 0 & 2i \\ 1 & 0 \end{pmatrix},$$

and neither is normal: for the sum, $(\mathbf{M}+\mathbf{N})(\mathbf{M}+\mathbf{N})^*$ has $(1,2)$ entry $1 - 2i$ while $(\mathbf{M}+\mathbf{N})^*(\mathbf{M}+\mathbf{N})$ has $1 + 2i$ there; for the product, $(\mathbf{MN})(\mathbf{MN})^* = \operatorname{diag}(4, 1)$ while $(\mathbf{MN})^*(\mathbf{MN}) = \operatorname{diag}(1, 4)$.

Quick check. Let $\mathbf{A} \in M_n(\mathbb{C})$ be normal and let $p \in \mathbb{C}[x]$. Explain, without using Theorem 12.7.1, why $\mathbf{A}$ and $p(\mathbf{A})$ are simultaneously unitarily diagonalizable.

Solution. By Corollary 12.4.3, $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{U}^*$ with $\mathbf{U}$ unitary and $\mathbf{D}$ diagonal. Then $\mathbf{A}^m = \mathbf{U}\mathbf{D}^m\mathbf{U}^*$ for every $m \geq 0$, since the inner factors $\mathbf{U}^*\mathbf{U}$ cancel, so $p(\mathbf{A}) = \mathbf{U}p(\mathbf{D})\mathbf{U}^*$, and $p(\mathbf{D}) = \operatorname{diag}(p(d_1), \dots, p(d_n))$ is diagonal. The *same* $\mathbf{U}$ diagonalizes both. This is consistent with the theorem, since $\mathbf{A}$ and $p(\mathbf{A})$ commute and $p(\mathbf{A})$ is normal, being unitarily diagonalizable.

12.7.4 Without normality: simultaneous triangularization

If we ask only for triangular matrices, normality may be dropped entirely. Chapter 9 proved this with ordinary bases (Theorem 9.9.4); over $\mathbb{C}$ an inner product upgrades the basis to an orthonormal one, exactly as Theorem 12.1.2 upgraded Theorem 9.7.2.

Theorem 12.7.5. Simultaneous Unitary Triangularization

Let $V \neq \{0\}$ be a finite-dimensional complex inner product space and let $\mathcal{F} \subseteq \mathcal{L}(V)$ be a nonempty commuting family. Then there is an orthonormal basis $\mathcal{B}$ of V such that $[T]_{\mathcal{B}}^{\mathcal{B}}$ is upper triangular for **every** $T \in \mathcal{F}$. In particular, if $\mathbf{A}_1, \dots, \mathbf{A}_r \in M_n(\mathbb{C})$ commute pairwise, there is $\mathbf{U} \in \mathrm{U}(n)$ with every $\mathbf{U}^*\mathbf{A}_i\mathbf{U}$ upper triangular.

Idea. The same swap that proves Schur's theorem (Theorem 12.1.2). A common eigenvector v of $\mathcal{F}$ gives an invariant line, but $\operatorname{span}(v)^\perp$ need not be invariant, so the recursion has nowhere to go. Take instead a common eigenvector u of the **adjoints** T^*: they commute too, and $\operatorname{span}(u)^\perp$ *is* invariant under every $T \in \mathcal{F}$. Recurse there, and put u at the **end** of the basis, where an upper triangular matrix allows an arbitrary column.

Proof. Induction on $n = \dim V \geq 1$. For $n = 1$ every 1×1 matrix is upper triangular, and a unit vector is an orthonormal basis.

Let $n \geq 2$ and assume the theorem in dimension $n - 1$. The family $\mathcal{F}^* = \{T^* : T \in \mathcal{F}\}$ is commuting: for $S, T \in \mathcal{F}$, $S^*T^* = (TS)^* = (ST)^* = T^*S^*$ by Theorem 11.6.2. Since V is a complex space, Theorem 9.9.3 provides $u \neq 0$ with $T^*u = \nu_T u$ for every $T \in \mathcal{F}$; dividing by $\|u\|$, take $\|u\| = 1$.

Put $W = \operatorname{span}(u)$ and $U = W^\perp$, of dimension $n - 1$ by Theorem 11.3.2 (c). For each $T \in \mathcal{F}$ the line W is T^*-invariant, since $T^*u = \nu_T u$, so U is T-invariant by Lemma 12.1.1. The family $\{T|_U : T \in \mathcal{F}\}$ is commuting, by the same computation on restrictions as in the proof of Theorem 12.7.1. By the inductive hypothesis there is an orthonormal basis $\mathcal{B}_U = (v_1, \dots, v_{n-1})$ of U such that $[T|_U]_{\mathcal{B}_U}^{\mathcal{B}_U}$ is upper triangular for every $T \in \mathcal{F}$.

Let $\mathcal{B} = (v_1, \dots, v_{n-1}, u)$. It is orthonormal, since u is a unit vector orthogonal to U, and it has n vectors, so it is an orthonormal basis of V. Fix $T \in \mathcal{F}$. For $j \leq n - 1$, $Tv_j = T|_U v_j \in \operatorname{span}(v_1, \dots, v_j)$, because $[T|_U]_{\mathcal{B}_U}^{\mathcal{B}_U}$ is upper triangular; so the j-th column of $[T]_{\mathcal{B}}^{\mathcal{B}}$ has zeros below row j. The last column is unconstrained, and an upper triangular matrix permits any entries there. Hence $[T]_{\mathcal{B}}^{\mathcal{B}}$ is upper triangular, which completes the induction.

For the matrix statement, apply this to the operators $T_{\mathbf{A}_i}$ on $\mathbb{C}^n$ and let $\mathbf{U}$ have the vectors of $\mathcal{B}$ as columns, as in Corollary 12.7.2.

■

The order matters: the common eigenvector of the adjoints goes last, and the triangular shape is built from the bottom right corner upward. Running the same recursion with a common eigenvector of $\mathcal{F}$ itself and putting it first would work only if $\operatorname{span}(v)^\perp$ were invariant, which is the extra fact that normality supplies and a bare commuting family does not.

12.7.5 Exercises

12.7.5.1 A. Check your understanding

Exercise. A1

(a) State Theorem 12.7.1, naming every hypothesis.

(b) Which result of Chapter 9 makes the eigenspaces of one member invariant under the others, and which hypothesis does it use?

(c) Why is the restriction of a normal operator to an invariant subspace again normal, and which earlier result says so? Why does the same fail for a subspace that is not invariant?

(d) True or false: two commuting diagonalizable matrices are simultaneously **unitarily** diagonalizable. Justify your answer.

(e) True or false: if S and T are normal then $S+T$ is normal. Justify your answer.

(f) Which hypothesis of Theorem 12.7.1 is dropped in Theorem 12.7.5, and what is lost in the conclusion?

Solution.

(a) Let $V \neq \{0\}$ be a finite-dimensional **complex** inner product space and $\mathcal{F} \subseteq \mathcal{L}(V)$ a **nonempty** family of **normal** operators that **commute pairwise**. Then V has an orthonormal basis of common eigenvectors.

(b) Theorem 9.9.1: if $ST = TS$, then $E_\lambda(T)$ is S-invariant. The only hypothesis is that the two operators commute.

(c) Proposition 12.3.6: in an orthonormal basis adapted to U the matrix of a normal T is $\mathbf{A} = \begin{pmatrix} \mathbf{B} & \mathbf{C} \\ \mathbf{O} & \mathbf{D} \end{pmatrix}$, and comparing the traces of the $(1,1)$ blocks of $\mathbf{A}^*\mathbf{A}$ and $\mathbf{A}\mathbf{A}^*$ forces $\mathbf{C} = \mathbf{O}$; the matrix is then block diagonal and its first block is normal. For a subspace that is not invariant the zero block below is absent and the argument collapses: the Section 3 warning compresses the 3×3 cyclic shift to $\operatorname{span}(\mathbf{e}_1, \mathbf{e}_2)$ and gets a nonnormal 2×2 matrix.

(d) False. $\begin{pmatrix} 1 & 1 \\ 0 & 2 \end{pmatrix}$ and $\begin{pmatrix} 1 & 3 \\ 0 & 4 \end{pmatrix}$ commute and are diagonalizable, but the first is not normal, so it has no orthonormal eigenbasis at all (Corollary 12.4.3). Simultaneous diagonalization by an invertible matrix is all that Chapter 9 promises (Theorem 9.9.5).

(e) False in general, true when they commute (Theorem 12.7.4). For a counterexample take $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ and $\operatorname{diag}(1, 2i)$, as in the Remark after that theorem.

(f) Normality. What is lost is that the common matrices are only upper triangular, not diagonal.

12.7.5.2 B. Practice

Exercise. B1: Diagonalize two at once

Let

$$\mathbf{A} = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix}, \qquad \mathbf{B} = \begin{pmatrix} 2 & 0 & -2 \\ 0 & 2 & -2 \\ -2 & -2 & 4 \end{pmatrix} \in M_3(\mathbb{R}).$$

(a) Verify that $\mathbf{AB} = \mathbf{BA}$ and that both are symmetric.

(b) Find $\mathbf{Q} \in \mathrm{O}(3)$ with $\mathbf{Q}^{\top}\mathbf{AQ}$ and $\mathbf{Q}^{\top}\mathbf{BQ}$ both diagonal, and write down the two diagonal matrices.

(c) Explain why $\mathbf{A}$ on its own does not determine $\mathbf{Q}$ up to permuting and rescaling columns, but the pair does.

Solution.

(a) Both are equal to their transposes by inspection. Each row of $\mathbf{B}$ sums to 0 and each column too, so $\mathbf{1}\mathbf{1}^{\top}\mathbf{B} = \mathbf{O} = \mathbf{B}\mathbf{1}\mathbf{1}^{\top}$ where $\mathbf{1} = (1,1,1)$. Since $\mathbf{A} = \mathbf{1}\mathbf{1}^{\top} + \mathbf{I}_3$, we get $\mathbf{AB} = \mathbf{B} = \mathbf{BA}$. (Direct multiplication confirms $\mathbf{AB} = \mathbf{B}$.)

(b) The eigenvectors of $\mathbf{A} = \mathbf{1}\mathbf{1}^{\top} + \mathbf{I}_3$ are $\mathbf{1}$, with $\mathbf{A}\mathbf{1} = 3\mathbf{1} + \mathbf{1} = 4\mathbf{1}$, and every $\mathbf{v}$ with $v_1 + v_2 + v_3 = 0$, with $\mathbf{Av} = \mathbf{v}$. Inside that plane we search with $\mathbf{B}$:

$$\mathbf{B}(1,-1,0) = (2,-2,0) = 2(1,-1,0),$$
$$\mathbf{B}(1,1,-2) = (6,6,-12) = 6(1,1,-2),$$

and both vectors have coordinate sum 0. Also $\mathbf{B}\mathbf{1} = \mathbf{0}$. The three vectors $\mathbf{1}$, $(1,-1,0)$, $(1,1,-2)$ are pairwise orthogonal, as one checks from $1-1+0=0$, $1+1-2=0$, and $1-1+0=0$. Normalizing,

$$\mathbf{Q} = \begin{pmatrix} 1/\sqrt{3} & 1/\sqrt{2} & 1/\sqrt{6} \\ 1/\sqrt{3} & -1/\sqrt{2} & 1/\sqrt{6} \\ 1/\sqrt{3} & 0 & -2/\sqrt{6} \end{pmatrix},$$

and $\mathbf{Q}^{\top}\mathbf{AQ} = \mathrm{diag}(4,1,1)$, $\mathbf{Q}^{\top}\mathbf{BQ} = \mathrm{diag}(0,2,6)$.

(c) The eigenvalue 1 of $\mathbf{A}$ has a two-dimensional eigenspace, so *every* orthonormal basis of that plane serves $\mathbf{A}$ equally well, and there are infinitely many, not obtained from one another by permuting and rescaling columns. The matrix $\mathbf{B}$ has three distinct eigenvalues, so its eigenlines are determined, and the pair therefore determines the three lines of $\mathbf{Q}$ up to order and sign.

Exercise. B2: Check the two warnings

(a) For $\mathbf{A} = \begin{pmatrix} 1 & 1 \\ 0 & 2 \end{pmatrix}$ and $\mathbf{B} = \begin{pmatrix} 1 & 3 \\ 0 & 4 \end{pmatrix}$, find an invertible $\mathbf{P}$ with $\mathbf{P}^{-1}\mathbf{AP}$ and $\mathbf{P}^{-1}\mathbf{BP}$ diagonal, and show that no **unitary** $\mathbf{P}$ does this.

(b) For $\mathbf{X} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ and $\mathbf{Z} = \mathrm{diag}(1,-1)$, write down an orthogonal $\mathbf{Q}_1$ diagonalizing $\mathbf{X}$ and an orthogonal $\mathbf{Q}_2$ diagonalizing $\mathbf{Z}$, and prove that no single invertible matrix diagonalizes both.

Solution.

(a) The common eigenvectors are $\mathbf{e}_1$ (eigenvalues 1 and 1) and $(1,1)$ (eigenvalues 2 and 4): $\mathbf{A}(1,1) = (2,2)$ and $\mathbf{B}(1,1) = (4,4)$. So $\mathbf{P} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ works, with $\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \operatorname{diag}(1,2)$ and $\mathbf{P}^{-1}\mathbf{B}\mathbf{P} = \operatorname{diag}(1,4)$. If a unitary $\mathbf{P}$ did the same, then $\mathbf{A} = \mathbf{P}\mathbf{D}\mathbf{P}^*$ with $\mathbf{D}$ diagonal would be normal by Corollary 12.4.3; but $\mathbf{A}^*\mathbf{A} = \begin{pmatrix} 1 & 1 \\ 1 & 5 \end{pmatrix}$ and $\mathbf{A}\mathbf{A}^* = \begin{pmatrix} 2 & 2 \\ 2 & 4 \end{pmatrix}$ differ in every entry.

(b) $\mathbf{X}$ has eigenvectors $(1,1)$ and $(1,-1)$ for the eigenvalues 1 and -1, so $\mathbf{Q}_1 = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$ gives $\mathbf{Q}_1^{\top}\mathbf{X}\mathbf{Q}_1 = \operatorname{diag}(1,-1)$. $\mathbf{Z}$ is already diagonal, so $\mathbf{Q}_2 = \mathbf{I}_2$. No single invertible $\mathbf{P}$ diagonalizes both: if it did, then $\mathbf{P}^{-1}\mathbf{X}\mathbf{P}$ and $\mathbf{P}^{-1}\mathbf{Z}\mathbf{P}$ would be diagonal, hence commuting, and multiplying by $\mathbf{P}$ and $\mathbf{P}^{-1}$ would give $\mathbf{X}\mathbf{Z} = \mathbf{Z}\mathbf{X}$. But $\mathbf{X}\mathbf{Z} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ and $\mathbf{Z}\mathbf{X} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$.

Exercise. B3: Triangularize two at once

Let $\mathbf{A} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ and $\mathbf{B} = \begin{pmatrix} 2 & 3 \\ 0 & 2 \end{pmatrix}$ in $M_2(\mathbb{C})$.

(a) Verify that $\mathbf{A}\mathbf{B} = \mathbf{B}\mathbf{A}$.

(b) Determine whether $\mathbf{A}$ and $\mathbf{B}$ are simultaneously diagonalizable, and justify your answer.

(c) Exhibit a unitary $\mathbf{U}$ with $\mathbf{U}^*\mathbf{A}\mathbf{U}$ and $\mathbf{U}^*\mathbf{B}\mathbf{U}$ both upper triangular, as Theorem 12.7.5 promises.

Solution.

(a) $\mathbf{A}\mathbf{B} = \begin{pmatrix} 2 & 5 \\ 0 & 2 \end{pmatrix} = \mathbf{B}\mathbf{A}$.

(b) No. If one basis made both diagonal, then $\mathbf{A}$ alone would be diagonalizable; but $p_{\mathbf{A}} = (x-1)^2$ and $\mathbf{A} - \mathbf{I}_2 = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \neq \mathbf{O}$, so $\mathbf{A}$ has the single eigenvalue 1 with a one-dimensional eigenspace, and is not diagonalizable. This is the hypothesis that Theorem 9.9.5 needs and that commuting does not supply.

(c) $\mathbf{U} = \mathbf{I}_2$: both matrices are already upper triangular in the standard basis, which is orthonormal. The content of the theorem is that such a basis always exists, not that it is hard to find in this example.

12.7.5.3 C. Going deeper

Exercise. C1: A special case of a theorem of Fuglede

Let V be a finite-dimensional complex inner product space and let $S, T \in \mathcal{L}(V)$ be **both normal**, with $ST = TS$.

(a) Prove that $ST^* = T^*S$ and $S^*T = TS^*$.

(b) Deduce that $S + T^*$ and ST^* are normal.

(c) Give an example of two commuting operators, **not** both normal, for which the conclusion of (a) fails as stated for the pair — that is, exhibit commuting S, T with $ST^* \neq T^*S$.

Note on the hypotheses: a theorem of Fuglede says that in (a) the operator S need not be assumed normal — it is enough that T be normal and that S commute with T. That stronger statement is true and is not proved in this book; (a) is only the case in which both are normal, and its proof uses the normality of S in an essential way.

Solution.

(a) By Theorem 12.7.1 applied to the commuting family $\{S, T\}$ of normal operators, there is an orthonormal basis $\mathcal{B}$ with $[S]_{\mathcal{B}}^{\mathcal{B}} = \mathbf{D}$ and $[T]_{\mathcal{B}}^{\mathcal{B}} = \mathbf{D}'$ diagonal. By Theorem 11.6.3, $[T^*]_{\mathcal{B}}^{\mathcal{B}} = (\mathbf{D}')^*$, which is the diagonal matrix with the conjugated entries. Diagonal matrices commute, so

$$[ST^*]_{\mathcal{B}}^{\mathcal{B}} = \mathbf{D}(\mathbf{D}')^* = (\mathbf{D}')^*\mathbf{D} = [T^*S]_{\mathcal{B}}^{\mathcal{B}},$$

and $ST^* = T^*S$ because a map is determined by its matrix (Theorem 4.1.4). Exchanging the roles of S and T gives $TS^* = S^*T$.

(b) T^* is normal, since $(T^*)^*T^* = TT^* = T^*T = T^*(T^*)^*$. By (a), S and T^* commute, and both are normal, so Theorem 12.7.4 applies to the pair (S, T^*) and gives that $S + T^*$ and ST^* are normal.

(c) Take $S = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$ and $T = \mathbf{I}_2$. They commute, T is normal, and S is not. Here $ST^* = S = T^*S$, so this pair does not refute anything — as the theorem of Fuglede predicts, since T is normal. Take instead $T = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$ and $S = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$, which commute, since both are polynomials in T. Then

$$ST^* = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}, \qquad T^*S = \begin{pmatrix} 0 & 0 \\ 1 & 1 \end{pmatrix},$$

which differ. Here it is T, not S, that fails to be normal, and that is exactly the hypothesis the theorem of Fuglede keeps.

Exercise. C2: A commuting family of orthogonal projections

Let $V \neq \{0\}$ be a finite-dimensional inner product space over $F = \mathbb{R}$ or $\mathbb{C}$, and let $P_1, \dots, P_r \in \mathcal{L}(V)$ be the orthogonal projections onto subspaces $U_1, \dots, U_r$, pairwise commuting.

(a) Prove that there is an orthonormal basis $\mathcal{B}$ of V such that every vector of $\mathcal{B}$ lies in U_i or in $U_i^{\perp}$, for every i.

(b) Deduce that P_iP_j is the orthogonal projection onto $U_i \cap U_j$.

(c) Prove that $P_i + P_j$ is an orthogonal projection if and only if $U_i \perp U_j$, and identify the subspace in that case.

Hint for (a): the eigenvalues of an idempotent operator.

Solution.

(a) Each P_i is self-adjoint, by Example (Four Adjoints) (d). Over $\mathbb{R}$, Corollary 12.7.3 applies directly to the commuting family $\{P_1, \dots, P_r\}$; over $\mathbb{C}$ a self-adjoint operator is normal, since $P^*P = P^2 = PP^*$, so Theorem 12.7.1 applies. Either way there is an orthonormal basis $\mathcal{B}$ of common eigenvectors. Let $v \in \mathcal{B}$ and fix i. If $P_i v = \mu v$, then $\mu^2 v = P_i^2 v = P_i v = \mu v$, so $\mu \in \{0, 1\}$, since $v \neq 0$. If $\mu = 1$ then $v = P_i v \in \operatorname{im} P_i = U_i$; if $\mu = 0$ then $v \in \ker P_i = U_i^\perp$ (Definition 11.3.2).

(b) Write $\mathcal{B} = (v_1, \dots, v_n)$ and let $W = U_i \cap U_j$. By (a) each v_t satisfies $P_i v_t = a_t v_t$ and $P_j v_t = b_t v_t$ with $a_t, b_t \in \{0, 1\}$, so $P_i P_j v_t = a_t b_t v_t$, which is v_t when $v_t \in U_i \cap U_j$ and 0 otherwise. Let $L = \operatorname{span}\{v_t : a_t b_t = 1\}$; then $L \subseteq W$. Conversely, if $w \in W$, expand $w = \sum_t c_t v_t$; applying P_i gives $w = \sum_t a_t c_t v_t$, so $c_t = a_t c_t$ and hence $c_t = 0$ whenever $a_t = 0$, and likewise for b_t. So $w \in L$ and $W = L$. Now $P_i P_j$ fixes each $v_t \in L$ and kills the rest, and by Theorem 11.3.3 (a) with the orthonormal basis $\{v_t : a_t b_t = 1\}$ of W, so does P_W; two operators agreeing on a basis are equal, so $P_i P_j = P_W$.

(c) Write $Q = P_i + P_j$.

($\Rightarrow$) Suppose Q is the orthogonal projection onto some subspace W. Then $Q^2 = Q$ (Definition 11.3.2), and expanding,

$$P_i + P_j = Q^2 = P_i^2 + P_i P_j + P_j P_i + P_j^2 = P_i + P_j + 2P_i P_j,$$

using $P_i^2 = P_i$, $P_j^2 = P_j$, and $P_j P_i = P_i P_j$. Hence $2P_i P_j = 0$, so $P_i P_j = 0$. Now let $v \in U_j$; then $P_i v = P_i(P_j v) = 0$, so $v \in \ker P_i = U_i^\perp$. Therefore $U_j \subseteq U_i^\perp$, which says $U_i \perp U_j$.

($\Leftarrow$) Suppose $U_i \perp U_j$, so $U_j \subseteq U_i^\perp = \ker P_i$ and $U_i \subseteq \ker P_j$. For $v \in U_i$ we get $Qv = v + 0 = v$, and symmetrically $Qv = v$ for $v \in U_j$, so Q is the identity on $U_i \oplus U_j$. For $v \in (U_i \oplus U_j)^\perp = U_i^\perp \cap U_j^\perp$ we get $Qv = 0 + 0 = 0$. Since $V = (U_i \oplus U_j) \oplus (U_i \oplus U_j)^\perp$ by Theorem 11.3.2 (a), Q agrees with $P_{U_i \oplus U_j}$ on both summands, hence everywhere. So $P_i + P_j = P_{U_i \oplus U_j}$.

12.8 Spectral Resolutions and Functions of a Normal Operator

The last two sections produce an orthonormal basis of eigenvectors. A basis, though, is a choice: permute it, or rotate inside the eigenspace of a repeated eigenvalue, and the basis changes while the operator does not. This section rewrites the same information with nothing chosen — a list of eigenvalues, and the orthogonal projections onto their eigenspaces — and then spends it. Once an operator is written as $\lambda_1 P_1 + \cdots + \lambda_k P_k$, any function whatsoever on those k numbers produces a new operator, and powers, square roots, and exponentials all become one line of arithmetic on the spectrum.

Throughout, V is a **nonzero** finite-dimensional inner product space over $F = \mathbb{R}$ or $F = \mathbb{C}$, and $T \in \mathcal{L}(V)$. We say that T satisfies

$(*)$ V has an orthonormal basis of eigenvectors of T.

By Theorem 12.4.2 this holds over $\mathbb{C}$ exactly for the **normal** operators, and by Theorem 12.5.1 it holds over $\mathbb{R}$ exactly for the **self-adjoint** ones. Stating everything for an operator satisfying $(*)$ covers both fields at once, and keeps in view the only property we ever use.

12.8.1 Grouping an eigenbasis by eigenvalue

Here is the observation the section runs on. Let $(e_1, \dots, e_n)$ be an orthonormal eigenbasis of T, with $Te_j = \mu_j e_j$. The list $\mu_1, \dots, \mu_n$ repeats: an eigenvalue of geometric multiplicity 3 appears three times. Sort the basis vectors into bundles by the eigenvalue they carry, and each bundle spans one eigenspace. Summing the coordinates of a vector inside one bundle is exactly projecting it onto that eigenspace — orthogonally, because the basis is orthonormal. So T is a weighted sum of perpendicular projections, one weight per **distinct** eigenvalue, and the weights and the projections no longer remember the basis.

A normal operator is a sum of perpendicular projections, weighted by its distinct eigenvalues.

Definition 12.8.1. Spectral Resolution

Let $T \in \mathcal{L}(V)$ satisfy $(*)$, let $\lambda_1, \dots, \lambda_k$ be the **distinct** eigenvalues of T, and write $E_i = E_{\lambda_i}(T)$ for the corresponding eigenspaces. Let

$$P_i := P_{E_i} \in \mathcal{L}(V)$$

be the **orthogonal** projection onto E_i (Definition 11.3.2). The identity

$$T = \lambda_1 P_1 + \lambda_2 P_2 + \cdots + \lambda_k P_k$$

is the **spectral resolution** of T, and the family $(P_1, \dots, P_k)$ is the **resolution of the identity** attached to T.

Two clauses carry weight. The eigenvalues are listed **once each**, so k is the number of distinct eigenvalues and not $\dim V$; and the projections are **orthogonal**, which is the whole content of the definition, since every diagonalizable operator has *some* such expression with oblique projections. That the displayed identity is true, and that the P_i fit together as neatly as one could want, is the next theorem.

Theorem 12.8.1. The Spectral Resolution and Its Projections

Let $T \in \mathcal{L}(V)$ satisfy $(*)$, with distinct eigenvalues $\lambda_1, \dots, \lambda_k$, eigenspaces $E_i = E_{\lambda_i}(T)$, and $P_i = P_{E_i}$. Then:

(a) $V = E_1 \oplus \cdots \oplus E_k$, and $E_i \perp E_j$ whenever $i \neq j$;
(b) $P_i^* = P_i$ and $P_i^2 = P_i$ for every i, and $P_i P_j = 0$ whenever $i \neq j$;
(c) $P_1 + \cdots + P_k = \mathrm{id}_V$;
(d) $T = \lambda_1 P_1 + \cdots + \lambda_k P_k$;
(e) $TP_i = P_i T = \lambda_i P_i$ for every i.

Idea. Everything is read off one orthonormal eigenbasis. Sorting its vectors by eigenvalue cuts it into k bundles; the only thing to check is that the i-th bundle spans the **whole** of E_i, not just part of it, and that is a coordinate comparison. After that, (a) is the orthonormality of the basis, (c) and (d) are the coordinate formula $v = \sum_j \langle v, e_j\rangle e_j$ with its terms grouped, and (b) and (e) are bookkeeping.

Proof. Let $\mathcal{B} = (e_1, \dots, e_n)$ be an orthonormal basis of V with $Te_j = \mu_j e_j$, which exists by $(*)$. Each e_j is nonzero, so each μ_j is an eigenvalue of T and therefore equals exactly one λ_i. Put

$$\mathcal{S}_i := \{\, j : \mu_j = \lambda_i \,\},$$

so that $\{1, \dots, n\}$ is the disjoint union of $\mathcal{S}_1, \dots, \mathcal{S}_k$.

Step 1. $E_i = \operatorname{span}(e_j : j \in \mathcal{S}_i)$. The inclusion $(\supseteq)$ is immediate, since $Te_j = \lambda_i e_j$ for $j \in \mathcal{S}_i$. For $(\subseteq)$, let $v \in E_i$ and write $v = \sum_j c_j e_j$ in the basis $\mathcal{B}$. Applying T,

$$\sum_j c_j \mu_j e_j = Tv = \lambda_i v = \sum_j c_j \lambda_i e_j.$$

Coordinates in a basis are unique, so $c_j(\mu_j - \lambda_i) = 0$ for every j, and hence $c_j = 0$ unless $j \in \mathcal{S}_i$. In particular $\mathcal{S}_i \neq \varnothing$, because $E_i \neq \{0\}$.

Step 2 (a). By Step 1, $V = \operatorname{span}(\mathcal{B}) = E_1 + \cdots + E_k$. If $i \neq j$, then $\mathcal{S}_i$ and $\mathcal{S}_j$ are disjoint and $\mathcal{B}$ is orthonormal, so every spanning vector of E_i is orthogonal to every spanning vector of E_j; hence $E_i \perp E_j$. Pairwise orthogonal subspaces are independent: if $u_1 + \cdots + u_k = 0$ with $u_i \in E_i$, then pairing with u_l kills every term but one and leaves $\|u_l\|^2 = 0$, so $u_l = 0$. By Theorem 1.7.4 the sum is direct.

Step 3 (b). Each P_i is self-adjoint by Theorem 11.3.3 (b), and satisfies $P_i^2 = P_i$ with $\operatorname{im} P_i = E_i$ and $\ker P_i = E_i^{\perp}$ (Definition 11.3.2 and the discussion following it). For $i \neq j$, Step 2 gives $E_j \subseteq E_i^{\perp}$, so $\operatorname{im} P_j \subseteq \ker P_i$ and $P_i P_j = 0$.

Step 4 (c) and (d). By Step 1, $(e_j : j \in \mathcal{S}_i)$ is an orthonormal basis of E_i, so Theorem 11.3.3 (a) gives

$$P_i v = \sum_{j \in \mathcal{S}_i} \langle v, e_j\rangle e_j \qquad (v \in V).$$

Summing over i collects every index j exactly once, so by Theorem 11.2.1,

$$\sum_{i=1}^{k} P_i v = \sum_{j=1}^{n} \langle v, e_j\rangle e_j = v,$$

which is (c). Multiplying the i-th group by λ_i instead,

$$\begin{aligned} \sum_{i=1}^{k} \lambda_i P_i v &= \sum_{i=1}^{k} \sum_{j \in \mathcal{S}_i} \langle v, e_j\rangle\, \lambda_i e_j \\ &= \sum_{j=1}^{n} \langle v, e_j\rangle\, \mu_j e_j = T\Big(\sum_{j=1}^{n} \langle v, e_j\rangle e_j\Big) = Tv, \end{aligned}$$

where the second equality uses $\mu_j = \lambda_i$ for $j \in \mathcal{S}_i$, and the third the linearity of T together with $Te_j = \mu_j e_j$. This is (d).

Step 5 (e). For $v \in V$ we have $P_i v \in E_i$, so $T(P_i v) = \lambda_i P_i v$; thus $TP_i = \lambda_i P_i$. In the other order, (d) and (b) give

$$P_i T = P_i\Big(\sum_{j=1}^{k} \lambda_j P_j\Big) = \lambda_i P_i^2 = \lambda_i P_i.$$

This proves the theorem.

■

For matrices the picture is the familiar one, packaged differently. If $\mathbf{A} \in M_n(\mathbb{C})$ is normal, write $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{U}^*$ with $\mathbf{U}$ unitary and $\mathbf{D}$ diagonal (Corollary 12.4.3). Grouping the columns $\mathbf{q}_1, \dots, \mathbf{q}_n$ of $\mathbf{U}$ by their eigenvalue turns this into

$$\mathbf{A} = \sum_{i=1}^{k} \lambda_i \mathbf{P}_i, \qquad \mathbf{P}_i = \sum_{j \in \mathcal{S}_i} \mathbf{q}_j \mathbf{q}_j^*,$$

so each $\mathbf{P}_i$ is a sum of rank-one projections and $\operatorname{rank} \mathbf{P}_i = \dim E_i$.

Example. A Symmetric 3×3 Matrix

Let $\mathbf{A} = \begin{pmatrix} 1 & 2 & 2 \\ 2 & 1 & 2 \\ 2 & 2 & 1 \end{pmatrix} \in M_3(\mathbb{R})$. Find the spectral resolution of $\mathbf{A}$.

Solution. Write $\mathbf{A} = 2\mathbf{J} - \mathbf{I}$, where $\mathbf{J}$ is the all-ones matrix. Since $\mathbf{J}\mathbf{x} = (x_1 + x_2 + x_3)\mathbf{1}$, we get $\mathbf{J}\mathbf{1} = 3\mathbf{1}$ and $\mathbf{J}\mathbf{x} = \mathbf{0}$ exactly when $x_1 + x_2 + x_3 = 0$. So $\mathbf{J}$ has eigenvalues 3 and 0, and $\mathbf{A}$ has eigenvalues

$$\lambda_1 = 2 \cdot 3 - 1 = 5, \qquad \lambda_2 = 2 \cdot 0 - 1 = -1,$$

with $E_5 = \operatorname{span}(\mathbf{1})$ of dimension 1 and $E_{-1} = \mathbf{1}^{\perp}$ of dimension 2. Since $\|\mathbf{1}\|^2 = 3$, the orthogonal projection onto E_5 is $\mathbf{P}_1 = \frac{1}{3}\mathbf{1}\mathbf{1}^{\top} = \frac{1}{3}\mathbf{J}$, and $\mathbf{P}_2 = \mathbf{I} - \mathbf{P}_1$ by Theorem 12.8.1 (c). Concretely,

$$\mathbf{P}_1 = \frac{1}{3}\begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}, \qquad \mathbf{P}_2 = \frac{1}{3}\begin{pmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{pmatrix}.$$

Check. Both are symmetric, both satisfy $\mathbf{P}_i^2 = \mathbf{P}_i$, their product is $\mathbf{O}$, they sum to $\mathbf{I}$, and

$$5\mathbf{P}_1 - \mathbf{P}_2 = \tfrac{5}{3}\mathbf{J} - \mathbf{I} + \tfrac{1}{3}\mathbf{J} = 2\mathbf{J} - \mathbf{I} = \mathbf{A}.$$

Note that the resolution has two terms, not three: $k = 2$ counts **distinct** eigenvalues, while $\operatorname{rank} \mathbf{P}_2 = 2$ records the multiplicity of -1.

12.8.2 Orthogonal projections versus oblique ones

An expression $T = \sum \lambda_i \pi_i$ with idempotents summing to the identity is not by itself news. In the proof of Theorem 10.8.3 the semisimple part of T is written as $\sum_i \lambda_i \pi_i$, where $\pi_i = h_i(T)$ projects onto the generalized eigenspace $G_{\lambda_i}(T)$ along the sum of the others (Theorem 9.8.1 (d)); when T is diagonalizable, that is a decomposition of T itself, and those π_i satisfy $\pi_i^2 = \pi_i$, $\pi_i \pi_j = 0$ for $i \neq j$, and $\sum_i \pi_i = \mathrm{id}_V$, exactly as in Theorem 12.8.1 (b) and (c). What they do

not satisfy is $\pi_i^* = \pi_i$, and that single missing condition is the whole difference between Chapter 10 and this section.

Example. Two Projections That Are Not Orthogonal

Let $\mathbf{A} = \begin{pmatrix} 1 & 1 \\ 0 & 2 \end{pmatrix} \in M_2(\mathbb{R})$, which is diagonalizable with eigenvalues 1 and 2. Compute the projections $\boldsymbol{\pi}_1, \boldsymbol{\pi}_2$ of Theorem 9.8.1 for $\mathbf{A}$, verify that $\mathbf{A} = \boldsymbol{\pi}_1 + 2\boldsymbol{\pi}_2$, and decide whether they are orthogonal projections.

Solution. The eigenspaces are $E_1 = \operatorname{span}(\mathbf{e}_1)$ and $E_2 = \operatorname{span}((1,1))$, read off from $\mathbf{A} - \mathbf{I} = \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix}$ and $\mathbf{A} - 2\mathbf{I} = \begin{pmatrix} -1 & 1 \\ 0 & 0 \end{pmatrix}$. Since the eigenvalues are distinct, the generalized eigenspaces are the eigenspaces. With $\mathbf{P} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ we have $\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \operatorname{diag}(1,2)$, so

$$\boldsymbol{\pi}_1 = \mathbf{P}\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}\mathbf{P}^{-1} = \begin{pmatrix} 1 & -1 \\ 0 & 0 \end{pmatrix}, \qquad \boldsymbol{\pi}_2 = \mathbf{I} - \boldsymbol{\pi}_1 = \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix}.$$

Both square to themselves, their products vanish in both orders, they sum to $\mathbf{I}$, and

$$1 \cdot \boldsymbol{\pi}_1 + 2 \cdot \boldsymbol{\pi}_2 = \begin{pmatrix} 1 & -1 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 2 \\ 0 & 2 \end{pmatrix} = \mathbf{A}.$$

They are **not** orthogonal projections: $\boldsymbol{\pi}_1^{\top} = \begin{pmatrix} 1 & 0 \\ -1 & 0 \end{pmatrix} \neq \boldsymbol{\pi}_1$. Geometrically, $\boldsymbol{\pi}_1$ projects onto $\operatorname{span}(\mathbf{e}_1)$ **along** $\operatorname{span}((1,1))$, and those two lines meet at $45°$, not at a right angle. Consistently, $\mathbf{A}$ is not normal:

$$\mathbf{A}^{\top}\mathbf{A} = \begin{pmatrix} 1 & 1 \\ 1 & 5 \end{pmatrix} \neq \begin{pmatrix} 2 & 2 \\ 2 & 4 \end{pmatrix} = \mathbf{A}\mathbf{A}^{\top}.$$

Warning. **"Diagonalizable" gives projections; only "normal" makes them orthogonal.** Every diagonalizable operator splits as $\sum \lambda_i \pi_i$ with idempotents that annihilate each other and sum to the identity, and Example (Two Projections That Are Not Orthogonal) is such a splitting for a matrix that is not normal. The projections there are not self-adjoint, so they are not the projections of Definition 12.8.1, and the eigenspaces they project onto are not perpendicular. Whenever an argument needs $\langle P_i u, v\rangle = \langle u, P_i v\rangle$ — and most of this section does — normality is what pays for it.

12.8.3 What makes a resolution the resolution

The projections in Definition 12.8.1 were built from the eigenspaces, so they are visibly determined by T. It is much more useful to know the converse: that *any* decomposition with the right formal properties has to be that one. This is what lets us recognize a spectral resolution when it turns up by accident, and it is the engine of the uniqueness proof for square roots below.

Theorem 12.8.2. Uniqueness of the Spectral Resolution

Let V be a nonzero finite-dimensional inner product space over F and let $T \in \mathcal{L}(V)$. Suppose

$$T = \mu_1 Q_1 + \mu_2 Q_2 + \cdots + \mu_m Q_m,$$

where $\mu_1, \ldots, \mu_m \in F$ are **distinct** and $Q_1, \ldots, Q_m \in \mathcal{L}(V)$ are **nonzero** operators satisfying

(R1) $Q_i^* = Q_i$ and $Q_i^2 = Q_i$ for every i;
(R2) $Q_i Q_j = 0$ whenever $i \neq j$;
(R3) $Q_1 + \cdots + Q_m = \mathrm{id}_V$.

Then T satisfies $(*)$, $\mathrm{spec}(T) = \{\mu_1, \ldots, \mu_m\}$, and Q_i is the orthogonal projection onto $E_{\mu_i}(T)$ for every i. In particular m is the number of distinct eigenvalues of T, and the given decomposition is the spectral resolution of T.

Idea. Three moves. ① A self-adjoint idempotent is the orthogonal projection onto its own image: self-adjointness is used here and nowhere else, which is exactly why Example (Two Projections That Are Not Orthogonal) escapes the theorem. ② Conditions (R2) and (R3) then break V into the orthogonal direct sum of the images $U_i = \operatorname{im} Q_i$. ③ On U_i the operator T is multiplication by μ_i, so $U_i \subseteq E_{\mu_i}(T)$; and because the μ_i are distinct, an eigenvector of T cannot spread across two pieces, which upgrades the inclusion to an equality.

Proof. Write $U_i = \operatorname{im} Q_i$, which is nonzero because $Q_i \neq 0$.

Step 1. $Q_i = P_{U_i}$. Fix i and drop the subscript. If $u \in U$, say $u = Qx$, then $Qu = Q^2 x = Qx = u$, so Q is the identity on U. Next, $\ker Q = U^\perp$. Indeed, if $Qw = 0$, then for every x,

$$\langle w, Qx \rangle = \langle Qw, x \rangle = 0$$

by Definition 11.6.1 and (R1), so $w \in U^\perp$; conversely if $w \in U^\perp$, then $Qw \in U$ and

$$\|Qw\|^2 = \langle Qw, Qw \rangle = \langle w, Q^2 w \rangle = \langle w, Qw \rangle = 0,$$

using (R1) twice and then $Qw \in U \perp w$; so $Qw = 0$. Finally, for any $v \in V$,

$$v = Qv + (v - Qv), \qquad Qv \in U, \quad Q(v - Qv) = Qv - Q^2 v = 0,$$

so $v - Qv \in \ker Q = U^\perp$. This is the splitting of v in $U \oplus U^\perp$ (Theorem 11.3.2), so $Qv = P_U v$ by Definition 11.3.2.

Step 2. $V = U_1 \oplus \cdots \oplus U_m$, *orthogonally.* By (R3), every v equals $\sum_i Q_i v \in \sum_i U_i$. For $i \neq j$, $u = Q_i x$ and $w = Q_j y$ satisfy

$$\langle u, w \rangle = \langle Q_i x, Q_j y \rangle = \langle Q_j Q_i x, y \rangle = 0$$

by (R1) and (R2), so $U_i \perp U_j$. As in Step 2 of the proof of Theorem 12.8.1, pairwise orthogonal subspaces are independent, so the sum is direct.

Step 3. *T acts on U_i as multiplication by μ_i.* Let $u \in U_i$. By Step 1, $Q_i u = u$, and for $j \neq i$, $Q_j u = Q_j Q_i u = 0$ by (R2). Hence

$$Tu = \sum_{l=1}^{m} \mu_l Q_l u = \mu_i u.$$

Since $U_i \neq \{0\}$, this makes μ_i an eigenvalue of T, with $U_i \subseteq E_{\mu_i}(T)$.

Step 4. *$\operatorname{spec}(T) = \{\mu_1, \dots, \mu_m\}$ and $U_i = E_{\mu_i}(T)$.* Let λ be an eigenvalue of T with eigenvector $v \neq 0$. By (R3), $v = \sum_i Q_i v$ with $Q_i v \in U_i$, and by Step 3,

$$\sum_{i=1}^{m} \mu_i Q_i v = Tv = \lambda v = \sum_{i=1}^{m} \lambda Q_i v.$$

Subtracting, $\sum_i (\mu_i - \lambda) Q_i v = 0$ with the i-th term in U_i, so by the directness of Step 2 every term vanishes: $Q_i v = 0$ whenever $\mu_i \neq \lambda$. As $v \neq 0$, some $Q_i v \neq 0$, and for that i we get $\lambda = \mu_i$. The μ_l being distinct, i is the only index with $\mu_i = \lambda$, so $v = Q_i v \in U_i$. Therefore $\operatorname{spec}(T) \subseteq \{\mu_1, \dots, \mu_m\}$ and $E_{\mu_i}(T) \subseteq U_i$; with Step 3 both are equalities.

Step 5. By Steps 1 and 4, $Q_i = P_{U_i} = P_{E_{\mu_i}(T)}$. Choosing an orthonormal basis of each U_i (Theorem 11.2.2) and concatenating gives an orthonormal basis of V by Step 2, made of eigenvectors of T by Step 3; so T satisfies $(*)$. This proves the theorem.

■

Read the other way round, the theorem is the complex spectral theorem again with no basis in sight: over $\mathbb{C}$, an operator is normal **if and only if** it can be written as $\sum \mu_i Q_i$ with distinct μ_i and nonzero self-adjoint idempotents satisfying (R2) and (R3) (Theorem 12.4.2). Conditions (R1)–(R3) are the entire content of the theorem, compressed into three lines of algebra.

Quick check. The matrices $\boldsymbol{\pi}_1, \boldsymbol{\pi}_2$ of Example (Two Projections That Are Not Orthogonal) satisfy $\boldsymbol{\pi}_i^2 = \boldsymbol{\pi}_i$, $\boldsymbol{\pi}_1 \boldsymbol{\pi}_2 = \boldsymbol{\pi}_2 \boldsymbol{\pi}_1 = \mathbf{O}$, and $\boldsymbol{\pi}_1 + \boldsymbol{\pi}_2 = \mathbf{I}$, and the coefficients 1 and 2 are distinct. Which hypothesis of Theorem 12.8.2 fails, and what does its failure cost?

Solution. Self-adjointness in (R1): $\boldsymbol{\pi}_1^{\top} \neq \boldsymbol{\pi}_1$. Everything else holds. The cost is Step 1 of the proof: $\boldsymbol{\pi}_1$ is a projection onto $\operatorname{span}(\mathbf{e}_1)$, but along $\operatorname{span}((1,1))$ rather than along $\operatorname{span}(\mathbf{e}_1)^{\perp} = \operatorname{span}(\mathbf{e}_2)$, so it is not $P_{\operatorname{span}(\mathbf{e}_1)}$. Consequently the two images are not perpendicular, and the conclusion fails at its strongest point: $\begin{pmatrix} 1 & 1 \\ 0 & 2 \end{pmatrix}$ does not satisfy $(*)$, since it is not normal.

12.8.4 Functions of an operator

Powers are now trivial. Multiplying the resolution by itself and using $P_i P_j = 0$ for $i \neq j$ leaves only the diagonal terms:

$$T^2 = \Big(\sum_i \lambda_i P_i\Big)\Big(\sum_j \lambda_j P_j\Big) = \sum_i \lambda_i^2 P_i,$$

and the same computation repeated gives $T^m = \sum_i \lambda_i^m P_i$. So $p(T) = \sum_i p(\lambda_i) P_i$ for every polynomial p. Stare at that formula and notice what it does **not** use: nothing about p except its k values on the spectrum. Any function at all supplies those.

Definition 12.8.2. Function of an Operator

Let $T \in \mathcal{L}(V)$ satisfy $(*)$, with distinct eigenvalues $\lambda_1, \dots, \lambda_k$ and spectral resolution $T = \sum_i \lambda_i P_i$. For **any** function $f\colon \operatorname{spec}(T) \to F$, define

$$f(T) \coloneqq \sum_{i=1}^{k} f(\lambda_i)\, P_i \ \in \mathcal{L}(V).$$

In words: apply f to each eigenvalue and reweight the same projections. The function needs no continuity, no differentiability, and no formula — it is k numbers. And the definition is unambiguous: the λ_i and the P_i are determined by T alone, by Theorem 12.8.2, so no choice of basis or of eigenvectors enters.

Four instances show the range. Taking $f(\lambda) = \lambda$ returns T. Taking $f \equiv 1$ returns id_V, by Theorem 12.8.1 (c). Taking $f(\lambda) = \overline{\lambda}$ returns T^*, since $P_i^* = P_i$. And if $0 \notin \operatorname{spec}(T)$, taking $f(\lambda) = \lambda^{-1}$ returns T^{-1}, because

$$\Big(\sum_i \lambda_i P_i\Big)\Big(\sum_j \lambda_j^{-1} P_j\Big) = \sum_i P_i = \mathrm{id}_V\,.$$

The notation is, on its face, dangerous. For a polynomial p the symbol $p(T)$ already means something — the operator $a_0\,\mathrm{id}_V + a_1 T + \cdots + a_N T^N$ of Definition 2.4.2, which Chapter 6 studied at length. If Definition 12.8.2 disagreed with it, the whole notation would have to be abandoned. Part (a) below says it does not, and that is the reason the rest of the theorem is safe to use.

Theorem 12.8.3. Properties of the Functional Calculus

Let $T \in \mathcal{L}(V)$ satisfy $(*)$, with distinct eigenvalues $\lambda_1, \dots, \lambda_k$ and resolution of the identity $(P_1, \dots, P_k)$. Let $f, g\colon \operatorname{spec}(T) \to F$ and $c \in F$, with $f+g$, cf, and fg taken pointwise.

(a) **(Compatibility)** For every $p \in F[x]$, the operator $p(T)$ of Definition 2.4.2 equals $\sum_i p(\lambda_i) P_i$; that is, it agrees with Definition 12.8.2 applied to the restriction of p to $\operatorname{spec}(T)$.

(b) **(Algebra homomorphism)** $(f+g)(T) = f(T) + g(T)$, $(cf)(T) = c\,f(T)$, $(fg)(T) = f(T)g(T) = g(T)f(T)$, and $1(T) = \mathrm{id}_V$.

(c) **(Adjoints)** $f(T)^* = \overline{f}(T)$, where $\overline{f}(\lambda) = \overline{f(\lambda)}$. In particular $f(T)$ is self-adjoint when f is real-valued, and unitary when $|f(\lambda_i)| = 1$ for every i.

(d) **(Spectral mapping)** $f(T)$ satisfies $(*)$, with the same orthonormal eigenbasis as T, and $\operatorname{spec}(f(T)) = f(\operatorname{spec}(T))$.

(e) **(Commutation)** If $S \in \mathcal{L}(V)$ commutes with T, then S commutes with every P_i and hence with $f(T)$. In particular T and $f(T)$ commute.

Idea. Everything rests on one multiplication table, $P_i P_j = \delta_{ij} P_i$. It says that the operators $\sum_i c_i P_i$ multiply exactly the way the coefficient lists $(c_1, \dots, c_k)$ multiply coordinatewise, so $\mathcal{L}(V)$ contains a copy of the algebra of all F-valued functions on $\operatorname{spec}(T)$, and $f \mapsto f(T)$ is

the obvious map into it. That makes (b) and (c) one-line computations. The one statement needing an argument is (a) — that the copy sitting inside is the *same* copy Chapter 2 built by substituting T into polynomials — and the argument is to compute the powers of T. For (e), a commuting S preserves each eigenspace, and an operator preserving all the summands of an orthogonal direct sum commutes with the projections onto them.

Proof.

(a) We first show that

$$T^m = \sum_{i=1}^{k} \lambda_i^m P_i \qquad \text{for every integer } m \geq 0,$$

with the usual convention $\lambda^0 = 1$. For $m = 0$ both sides equal id_V, by Theorem 12.8.1 (c). Assuming the formula for m, Theorem 12.8.1 (d) and (b) give

$$\begin{aligned} T^{m+1} = T \cdot T^m &= \Big(\sum_i \lambda_i P_i\Big)\Big(\sum_j \lambda_j^m P_j\Big) \\ &= \sum_{i,j} \lambda_i \lambda_j^m \, P_i P_j = \sum_i \lambda_i^{m+1} P_i, \end{aligned}$$

since $P_i P_j = 0$ for $i \neq j$ and $P_i^2 = P_i$. Now let $p = a_0 + a_1 x + \cdots + a_N x^N$. By Definition 2.4.2 and the formula just proved,

$$\begin{aligned} p(T) = \sum_{m=0}^{N} a_m T^m &= \sum_{m=0}^{N} a_m \sum_{i=1}^{k} \lambda_i^m P_i \\ &= \sum_{i=1}^{k} \Big(\sum_{m=0}^{N} a_m \lambda_i^m \Big) P_i = \sum_{i=1}^{k} p(\lambda_i) P_i, \end{aligned}$$

the middle step being a rearrangement of a finite double sum.

(b) The first two identities are immediate from Definition 12.8.2. For the product, Theorem 12.8.1 (b) gives

$$f(T)g(T) = \sum_{i,j} f(\lambda_i) g(\lambda_j) P_i P_j = \sum_i f(\lambda_i) g(\lambda_i) P_i,$$

which is $(fg)(T)$. The right-hand side is symmetric in f and g, so $f(T)g(T) = g(T)f(T)$. Finally $1(T) = \sum_i P_i = \mathrm{id}_V$ by Theorem 12.8.1 (c).

(c) By Theorem 11.6.2 (a) and (b), and $P_i^* = P_i$,

$$f(T)^* = \sum_{i=1}^{k} \overline{f(\lambda_i)}\, P_i^* = \overline{f}(T).$$

If f is real-valued then $\overline{f} = f$ and $f(T)^* = f(T)$. If $|f(\lambda_i)| = 1$ for all i, then $\overline{f} f \equiv 1$, so by (b) $f(T)^* f(T) = (\overline{f} f)(T) = \mathrm{id}_V$, and likewise $f(T) f(T)^* = \mathrm{id}_V$; so $f(T)$ is unitary (Definition 11.7.2).

(d) Let $\mathcal{B} = (e_1, \dots, e_n)$ be an orthonormal eigenbasis of T, with $Te_j = \mu_j e_j$, and fix j. Let i be the index with $\mu_j = \lambda_i$. Then $e_j \in E_i$, so $P_i e_j = e_j$, while $P_l e_j = 0$ for $l \neq i$ because $E_i \subseteq E_l^{\perp} = \ker P_l$, by Theorem 12.8.1 (a) and Definition 11.3.2. Hence

$$f(T)e_j = f(\lambda_i)e_j = f(\mu_j)\, e_j.$$

So $\mathcal{B}$ is an orthonormal basis of eigenvectors of $f(T)$, which is $(*)$ for $f(T)$. The matrix of $f(T)$ in $\mathcal{B}$ is $\operatorname{diag}(f(\mu_1), \dots, f(\mu_n))$, so its eigenvalues are exactly the numbers $f(\mu_j)$, that is the numbers $f(\lambda_i)$; hence $\operatorname{spec}(f(T)) = f(\operatorname{spec}(T))$.

(e) Suppose $ST = TS$. Each E_i is S-invariant: if $Tv = \lambda_i v$, then $T(Sv) = S(Tv) = \lambda_i Sv$. Let $v \in V$ and write $v = \sum_i P_i v$ with $P_i v \in E_i$ (Theorem 12.8.1 (c)). Applying S gives $Sv = \sum_i SP_i v$ with $SP_i v \in E_i$, and this is the decomposition of Sv in $E_1 \oplus \cdots \oplus E_k$, which is unique. Comparing it with $Sv = \sum_i P_i(Sv)$ gives $P_i Sv = SP_i v$ for every i, that is $SP_i = P_i S$. Hence S commutes with $f(T) = \sum_i f(\lambda_i) P_i$. Since T commutes with itself, the last claim follows. This proves the theorem.

■

Part (a) is what makes the notation honest. Writing $\sqrt{T}$, e^T, and $p(T)$ in the same paragraph is legitimate precisely because the polynomial case is common ground: the new definition extends the old one instead of competing with it. It also reconciles this section with Chapter 10. An operator satisfying $(*)$ is diagonalizable, so every Jordan block of its matrix has size 1, and Definition 10.9.1 consults no derivatives at all — it too returns $\sum_i f(\lambda_i)\mathbf{P}_i$. The two constructions of $f(\mathbf{A})$ agree for normal $\mathbf{A}$, and Chapter 10's extra machinery is precisely the price of dropping normality.

Warning. **Outside the normal case, f on the spectrum is not enough data.** For $\mathbf{A} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$ the only eigenvalue is 0, so the recipe "$f(\mathbf{A}) = \sum_i f(\lambda_i)\mathbf{P}_i$" would read $f(\mathbf{A}) = f(0)\mathbf{I}$ for every f — and with $f(x) = x$ it would announce $\mathbf{A} = \mathbf{O}$. The recipe is not wrong, it is inapplicable: $\mathbf{A}$ is not normal, so there is no resolution of the identity to expand in. Chapter 10's definition of $f(\mathbf{A})$ has to consult $f'(0)$ as well (Definition 10.9.1), and that derivative is exactly what the missing orthogonality would have spared us.

Quick check. Let $T \in \mathcal{L}(V)$ satisfy $(*)$ with $\operatorname{spec}(T) = \{0, 1\}$. Identify T.

Solution. T is the orthogonal projection onto $E_1(T)$. Indeed the spectral resolution reads $T = 0 \cdot P_1 + 1 \cdot P_2 = P_2 = P_{E_1(T)}$. So the operators satisfying $(*)$ with spectrum in $\{0, 1\}$ are exactly the orthogonal projections, together with 0 and id_V for the degenerate spectra $\{0\}$ and $\{1\}$.

12.8.5 Square roots

The function $\lambda \mapsto \sqrt{\lambda}$ is defined on $[0, \infty)$, so Definition 12.8.2 produces a square root of T as soon as the spectrum lies there. The interesting part is that the square root is **unique** once we

ask it to be of the same kind, and the uniqueness proof is a direct application of Theorem 12.8.2: a square root's own resolution, squared, has to be T's.

Note first that over $\mathbb{C}$ the hypothesis is stronger than it looks. If T satisfies $(*)$ and $\operatorname{spec}(T) \subseteq \mathbb{R}$, then $f(\lambda) = \lambda$ is real-valued on the spectrum, so $T = f(T)$ is self-adjoint by Theorem 12.8.3 (c). A normal operator with real spectrum is self-adjoint.

Theorem 12.8.4. The Square Root of a Nonnegative Normal Operator

Let V be a nonzero finite-dimensional inner product space over F and let $T \in \mathcal{L}(V)$ satisfy $(*)$ with $\operatorname{spec}(T) \subseteq [0, \infty)$. Then there is **exactly one** operator $S \in \mathcal{L}(V)$ such that

(i) S satisfies $(*)$;
(ii) $\operatorname{spec}(S) \subseteq [0, \infty)$;
(iii) $S^2 = T$.

It is $S = \sum_i \sqrt{\lambda_i}\, P_i$, where $T = \sum_i \lambda_i P_i$ is the spectral resolution of T. This S is self-adjoint, and it commutes with every operator that commutes with T.

Idea. Existence is the functional calculus applied to $\sqrt{\ }$: squaring a function squares its values. Uniqueness is where the work is, and the trick is not to compare S with the candidate directly but to *square S's own resolution.* If $S = \sum_j \mu_j Q_j$, then $S^2 = \sum_j \mu_j^2 Q_j$, and the μ_j^2 are still distinct because the μ_j are distinct **and nonnegative**. So $\sum_j \mu_j^2 Q_j$ meets every condition of Theorem 12.8.2 for T, which forces it to be T's spectral resolution term by term.

Proof. Existence. Let $f(\lambda) = \sqrt{\lambda}$ on $\operatorname{spec}(T) \subseteq [0, \infty)$, which takes values in $[0, \infty) \subseteq F$, and put $S = f(T) = \sum_i \sqrt{\lambda_i} P_i$. Then f^2 is the function $\lambda \mapsto \lambda$, so by Theorem 12.8.3 (b) and (a),

$$S^2 = f(T) f(T) = (f^2)(T) = T,$$

giving (iii). By Theorem 12.8.3 (d), S satisfies $(*)$ and $\operatorname{spec}(S) = \{\sqrt{\lambda_1}, \dots, \sqrt{\lambda_k}\} \subseteq [0, \infty)$, giving (i) and (ii). Since f is real-valued, S is self-adjoint by Theorem 12.8.3 (c), and it commutes with everything commuting with T by (e) of the same theorem.

Uniqueness. Let S satisfy (i), (ii), and (iii), and let

$$S = \mu_1 Q_1 + \cdots + \mu_l Q_l$$

be its spectral resolution (Theorem 12.8.1): the μ_j are the distinct eigenvalues of S, and the Q_j are nonzero self-adjoint idempotents with $Q_j Q_{j'} = 0$ for $j \neq j'$ and $\sum_j Q_j = \mathrm{id}_V$. By Theorem 12.8.3 (a) applied to S and $p = x^2$,

$$T = S^2 = \sum_{j=1}^{l} \mu_j^2\, Q_j.$$

The scalars μ_j^2 are **distinct**: the μ_j are distinct by construction, and nonnegative by (ii), so $\mu_j^2 = \mu_{j'}^2$ would force $\mu_j = \mu_{j'}$. The family $(Q_1, \dots, Q_l)$ satisfies (R1), (R2), and (R3) of Theorem 12.8.2, and every Q_j is nonzero. That theorem therefore identifies the displayed expression as the spectral resolution of T: after renumbering, $l = k$, $\mu_j^2 = \lambda_j$, and $Q_j = P_j$

for every j. Since $\mu_j \geq 0$ and $\mu_j^2 = \lambda_j$, we get $\mu_j = \sqrt{\lambda_j}$, and hence

$$S = \sum_{j=1}^{k} \sqrt{\lambda_j}\, P_j,$$

which is the operator constructed above. This proves the theorem.

■

The square root is written $\sqrt{T}$, or $T^{1/2}$. Chapter 13 takes the hypothesis $\operatorname{spec}(T) \subseteq [0, \infty)$ as the definition of a **positive** operator, proves that it is equivalent to $\langle Tv, v\rangle \geq 0$ for all v, and makes this theorem the tool behind polar decomposition and the singular value decomposition; the notation $\mathbf{A}^{1/2}$ in the notation table refers to exactly the operator built here.

Example. A Square Root, and Three Impostors

Let $\mathbf{A} = \begin{pmatrix} 10 & 6 \\ 6 & 10 \end{pmatrix} \in M_2(\mathbb{R})$. Find its spectral resolution, compute $\sqrt{\mathbf{A}}$ and $\mathbf{A}^m$, and find a symmetric square root of $\mathbf{A}$ other than $\sqrt{\mathbf{A}}$.

Solution. The resolution. $\mathbf{A}$ is symmetric, hence self-adjoint on $\mathbb{R}^2$. Since $\mathbf{A}(1,1) = (16,16)$ and $\mathbf{A}(1,-1) = (4,-4)$, the eigenvalues are 16 and 4, with eigenlines $\operatorname{span}((1,1))$ and $\operatorname{span}((1,-1))$. Both spanning vectors have squared length 2, so

$$\mathbf{P}_1 = \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}, \qquad \mathbf{P}_2 = \frac{1}{2}\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}, \qquad \mathbf{A} = 16\mathbf{P}_1 + 4\mathbf{P}_2.$$

The square root. Both eigenvalues are positive, so Theorem 12.8.4 applies:

$$\sqrt{\mathbf{A}} = 4\mathbf{P}_1 + 2\mathbf{P}_2 = \begin{pmatrix} 2 & 2 \\ 2 & 2 \end{pmatrix} + \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} = \begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix},$$

and indeed $\begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix}^2 = \begin{pmatrix} 10 & 6 \\ 6 & 10 \end{pmatrix}$.

Powers. By Theorem 12.8.3 (a), $\mathbf{A}^m = 16^m\mathbf{P}_1 + 4^m\mathbf{P}_2$, that is

$$\mathbf{A}^m = \frac{1}{2}\begin{pmatrix} 16^m + 4^m & 16^m - 4^m \\ 16^m - 4^m & 16^m + 4^m \end{pmatrix}.$$

For $m = 3$ this reads $\begin{pmatrix} 2080 & 2016 \\ 2016 & 2080 \end{pmatrix}$, which is $\mathbf{A}^3$.

An impostor. Changing one sign gives $\mathbf{B} = 4\mathbf{P}_1 - 2\mathbf{P}_2 = \begin{pmatrix} 1 & 3 \\ 3 & 1 \end{pmatrix}$, and $\mathbf{B}^2 = \mathbf{A}$ as well. So $\mathbf{A}$ has more than one symmetric square root; $\mathbf{B}$ fails clause (ii) of Theorem 12.8.4, since $\operatorname{spec}(\mathbf{B}) = \{4, -2\}$. Flipping the other sign, and both, gives $-\mathbf{B}$ and $-\sqrt{\mathbf{A}}$: four symmetric square roots, and exactly one of them has nonnegative spectrum.

Warning. **"A square root" and "the square root" are different things.** Uniqueness in Theorem 12.8.4 needs all three clauses. Dropping (ii) admits the three sign-flipped roots of Example (A Square Root, and Three Impostors). Dropping (i) admits more still, and the cheapest witness is the zero operator on $\mathbb{C}^2$. Take

$$\mathbf{S} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \qquad \mathbf{S}^2 = \mathbf{O}.$$

Here $\operatorname{spec}(\mathbf{S}) = \{0\} \subseteq [0, \infty)$, so (ii) holds, and (iii) holds by the display, yet $\mathbf{S} \neq \mathbf{O} = \sqrt{\mathbf{O}}$. Only (i) fails: $\mathbf{S}^*\mathbf{S} = \operatorname{diag}(0, 1)$ while $\mathbf{S}\mathbf{S}^* = \operatorname{diag}(1, 0)$. The phrase "**the** square root of T" always means the one satisfying all three.

12.8.6 The exponential and the unitary group

Chapter 10 defined $e^{\mathbf{A}}$ for every $\mathbf{A} \in M_n(\mathbb{C})$ as the sum of the series $\sum_m \mathbf{A}^m/m!$ (Definition 10.9.2) and proved that the series converges (Theorem 10.9.3). We do not redevelop any of that. What the spectral resolution adds is that for a normal matrix the series can be summed by inspection, and that the answer identifies the unitary matrices as the exponentials of the Hermitian ones — the matrix version of $|e^{i\theta}| = 1$.

Theorem 12.8.5. Unitary Matrices Are Exponentials of Hermitian Matrices

Let $n \geq 1$.

(a) If $\mathbf{A} \in M_n(\mathbb{C})$ is normal with spectral resolution $\mathbf{A} = \sum_i \lambda_i \mathbf{P}_i$, then $e^{\mathbf{A}} = \sum_i e^{\lambda_i} \mathbf{P}_i$.

(b) If $\mathbf{H} \in M_n(\mathbb{C})$ is Hermitian, then $e^{i\mathbf{H}} \in \mathrm{U}(n)$, and $\det e^{i\mathbf{H}} = e^{i \operatorname{tr} \mathbf{H}}$.

(c) Conversely, every $\mathbf{U} \in \mathrm{U}(n)$ equals $e^{i\mathbf{H}}$ for some Hermitian $\mathbf{H}$, which may be chosen with $\operatorname{spec}(\mathbf{H}) \subseteq [0, 2\pi)$.

Idea. For (a), the partial sums of the series collapse: $\mathbf{A}^m = \sum_i \lambda_i^m \mathbf{P}_i$, so each partial sum is a fixed combination of the k scalar partial sums of $\sum_m \lambda_i^m/m!$, whose limits are the numbers e^{λ_i}. For (b) and (c), the resolution turns everything into a statement about k complex numbers: $\mathbf{H}$ Hermitian means real eigenvalues, $e^{i\mathbf{H}}$ then has eigenvalues on the unit circle, and a unitary matrix has eigenvalues on the unit circle, each of which is $e^{i\theta}$ for a unique $\theta \in [0, 2\pi)$. Reading the last sentence backwards constructs $\mathbf{H}$.

Proof.

(a) By Theorem 12.8.3 (a), $\mathbf{A}^m = \sum_i \lambda_i^m \mathbf{P}_i$ for every $m \geq 0$, so for every $N \geq 0$,

$$\sum_{m=0}^{N} \frac{\mathbf{A}^m}{m!} = \sum_{i=1}^{k} \Big(\sum_{m=0}^{N} \frac{\lambda_i^m}{m!} \Big) \mathbf{P}_i.$$

Fix a position (r, s). Its entry on the right is $\sum_i c_{N,i} \, (\mathbf{P}_i)_{rs}$, where $c_{N,i} = \sum_{m \leq N} \lambda_i^m/m!$ and the coefficients $(\mathbf{P}_i)_{rs}$ do not depend on N. Each $c_{N,i} \to e^{\lambda_i}$ as $N \to \infty$ by the scalar exponential series, so the entry converges to $\sum_i e^{\lambda_i} (\mathbf{P}_i)_{rs}$. Hence the series converges entrywise to $\sum_i e^{\lambda_i} \mathbf{P}_i$, and by Definition 10.9.2 this matrix is $e^{\mathbf{A}}$.

(b) $\mathbf{H}$ is normal, so it has a spectral resolution $\mathbf{H} = \sum_j \theta_j \mathbf{P}_j$, and its eigenvalues θ_j are real by Theorem 12.2.1. The scalars $i\theta_j$ are distinct, so $i\mathbf{H} = \sum_j i\theta_j \mathbf{P}_j$ is the spectral resolution of $i\mathbf{H}$ by Theorem 12.8.2, and (a) gives

$$e^{i\mathbf{H}} = \sum_j e^{\,i\theta_j} \mathbf{P}_j.$$

Each θ_j is real, so $\overline{e^{\,i\theta_j}}\, e^{\,i\theta_j} = |e^{\,i\theta_j}|^2 = 1$. Hence, using $\mathbf{P}_j^* = \mathbf{P}_j$ and $\mathbf{P}_j \mathbf{P}_{j'} = \delta_{jj'} \mathbf{P}_j$ from Theorem 12.8.1 (b),

$$(e^{i\mathbf{H}})^* e^{i\mathbf{H}} = \sum_j \overline{e^{\,i\theta_j}} e^{\,i\theta_j}\, \mathbf{P}_j = \sum_j \mathbf{P}_j = \mathbf{I},$$

so $e^{i\mathbf{H}} \in \mathrm{U}(n)$ by Definition 11.7.3. For the determinant, Theorem 10.9.4 (e) gives $\det e^{i\mathbf{H}} = e^{\operatorname{tr}(i\mathbf{H})} = e^{\,i \operatorname{tr} \mathbf{H}}$.

(c) Let $\mathbf{U} \in \mathrm{U}(n)$. Then $\mathbf{U}^*\mathbf{U} = \mathbf{U}\mathbf{U}^* = \mathbf{I}$, so $\mathbf{U}$ is normal; let $\mathbf{U} = \sum_j \nu_j \mathbf{P}_j$ be its spectral resolution. Every eigenvalue has $|\nu_j| = 1$: if $\mathbf{U}\mathbf{x} = \nu_j \mathbf{x}$ with $\mathbf{x} \neq \mathbf{0}$, then $\|\mathbf{x}\| = \|\mathbf{U}\mathbf{x}\| = |\nu_j|\, \|\mathbf{x}\|$ because $\mathbf{U}$ is an isometry (Theorem 11.7.3), and $\|\mathbf{x}\| \neq 0$. So there is a unique $\theta_j \in [0, 2\pi)$ with $\nu_j = e^{\,i\theta_j}$, and the θ_j are distinct because the ν_j are. Put

$$\mathbf{H} \coloneqq \sum_j \theta_j \mathbf{P}_j.$$

The θ_j are real, so $\mathbf{H}^* = \sum_j \theta_j \mathbf{P}_j^* = \mathbf{H}$ by Theorem 11.6.2 (a), (b) and Theorem 12.8.1 (b): $\mathbf{H}$ is Hermitian, with $\operatorname{spec}(\mathbf{H}) = \{\theta_j\} \subseteq [0, 2\pi)$ by Theorem 12.8.2. By (a) applied to $i\mathbf{H}$,

$$e^{i\mathbf{H}} = \sum_j e^{\,i\theta_j} \mathbf{P}_j = \sum_j \nu_j \mathbf{P}_j = \mathbf{U}.$$

This proves the theorem.

■

Part (c) says that the exponential map from Hermitian matrices to $\mathrm{U}(n)$ is **surjective**. One consequence belongs to topology rather than algebra, so we only record it: the path $t \mapsto e^{\,it\mathbf{H}}$, $t \in [0, 1]$, stays inside $\mathrm{U}(n)$ and runs from $\mathbf{I}$ to $\mathbf{U}$, so every unitary matrix can be deformed continuously to the identity. Nothing of the sort holds in $\mathrm{O}(n)$, where the determinant is ± 1 and cannot change sign along a path.

Example. Exponentiating a Hermitian Matrix, and Undoing It

Let $\mathbf{H} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$. Compute $e^{\,i\theta\mathbf{H}}$ for $\theta \in \mathbb{R}$. Then find a Hermitian $\mathbf{K}$ with $e^{i\mathbf{K}} = \mathbf{Q}$, where $\mathbf{Q} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}$.

Solution. Forwards. $\mathbf{H}$ is symmetric with $\mathbf{H}(1,1) = (1,1)$ and $\mathbf{H}(1,-1) = -(1,-1)$, so its eigenvalues are ± 1 and

$$\mathbf{H} = \mathbf{P}_+ - \mathbf{P}_-, \qquad \mathbf{P}_\pm = \frac{1}{2} \begin{pmatrix} 1 & \pm 1 \\ \pm 1 & 1 \end{pmatrix}.$$

Hence $i\theta\mathbf{H}$ has resolution $i\theta\mathbf{P}_+ - i\theta\mathbf{P}_-$ (for $\theta \neq 0$), and Theorem 12.8.5 (a) gives

$$\begin{aligned} e^{i\theta\mathbf{H}} &= e^{i\theta}\mathbf{P}_+ + e^{-i\theta}\mathbf{P}_- \\ &= \begin{pmatrix} \cos\theta & i\sin\theta \\ i\sin\theta & \cos\theta \end{pmatrix}, \end{aligned}$$

using $e^{\pm i\theta} = \cos\theta \pm i\sin\theta$. Each row has squared length $\cos^2\theta + \sin^2\theta = 1$ and the two rows are orthogonal, confirming that the result is unitary.

Backwards. $\mathbf{Q}$ is the rotation by $\pi/4$, hence unitary, with $\mathbf{Q}(1,-i) = e^{i\pi/4}(1,-i)$ and $\mathbf{Q}(1,i) = e^{-i\pi/4}(1,i)$. Both eigenvectors have squared length 2, so the spectral projections are

$$\mathbf{R}_+ = \frac{1}{2}\begin{pmatrix} 1 & i \\ -i & 1 \end{pmatrix}, \qquad \mathbf{R}_- = \frac{1}{2}\begin{pmatrix} 1 & -i \\ i & 1 \end{pmatrix}.$$

Writing $e^{-i\pi/4} = e^{i\cdot 7\pi/4}$ to land in $[0, 2\pi)$, the recipe of Theorem 12.8.5 (c) gives

$$\mathbf{K} = \frac{\pi}{4}\mathbf{R}_+ + \frac{7\pi}{4}\mathbf{R}_- = \begin{pmatrix} \pi & -\frac{3\pi i}{4} \\ \frac{3\pi i}{4} & \pi \end{pmatrix},$$

Hermitian with spectrum $\{\pi/4,\, 7\pi/4\}$. Leaving $[0, 2\pi)$, the smaller choice $\frac{\pi}{4}(\mathbf{R}_+ - \mathbf{R}_-) = \frac{\pi}{4}\begin{pmatrix} 0 & i \\ -i & 0 \end{pmatrix}$ works just as well: the Hermitian logarithm is far from unique.

12.8.7 Exercises

12.8.7.1 A. Check your understanding

Exercise. A1

(a) Define the spectral resolution of T, and say which operators have one over $\mathbb{C}$ and which over $\mathbb{R}$.

(b) List the properties of the family $(P_1, \dots, P_k)$ given by Theorem 12.8.1, and say how many terms the resolution of a self-adjoint $T \in \mathcal{L}(\mathbb{R}^5)$ with eigenvalues $2, 2, 2, 7, 7$ has.

(c) True or false: if $T = \mu_1 Q_1 + \mu_2 Q_2$ with $\mu_1 \neq \mu_2$ and Q_1, Q_2 nonzero operators satisfying $Q_i^2 = Q_i$, $Q_1 Q_2 = Q_2 Q_1 = 0$, and $Q_1 + Q_2 = \mathrm{id}_V$, then T is normal. Justify your answer.

(d) Define $f(T)$ for a function f on $\mathrm{spec}(T)$, and say exactly how much information about f it uses.

(e) State Theorem 12.8.3 (a) and explain in one sentence why the rest of the section would be unusable without it.

(f) Give two different self-adjoint square roots of id_V on $\mathbb{R}^2$, and say which one Theorem 12.8.4 singles out and why.

Solution.

(a) For T with an orthonormal basis of eigenvectors, distinct eigenvalues $\lambda_1, \dots, \lambda_k$, and $P_i = P_{E_{\lambda_i}(T)}$ the orthogonal projection onto the i-th eigenspace, the spectral resolution

is $T = \sum_i \lambda_i P_i$ (Definition 12.8.1). Over $\mathbb{C}$ the operators with one are exactly the normal operators (Theorem 12.4.2); over $\mathbb{R}$, exactly the self-adjoint ones (Theorem 12.5.1).

(b) Each P_i is self-adjoint and idempotent, $P_iP_j = 0$ for $i \neq j$, $\sum_i P_i = \mathrm{id}_V$, and $TP_i = P_iT = \lambda_i P_i$. The resolution has **two** terms, $T = 2P_1 + 7P_2$, with $\operatorname{rank} P_1 = 3$ and $\operatorname{rank} P_2 = 2$: the resolution counts distinct eigenvalues, not multiplicities.

(c) False. Self-adjointness of the Q_i is missing, and it is the one hypothesis of Theorem 12.8.2 that does the work. The matrices $\boldsymbol{\pi}_1, \boldsymbol{\pi}_2$ of Example (Two Projections That Are Not Orthogonal) satisfy everything listed, with $\mu_1 = 1$ and $\mu_2 = 2$, and $\boldsymbol{\pi}_1 + 2\boldsymbol{\pi}_2 = \begin{pmatrix} 1 & 1 \\ 0 & 2 \end{pmatrix}$, which is not normal.

(d) $f(T) = \sum_i f(\lambda_i) P_i$ (Definition 12.8.2). It uses exactly the k values $f(\lambda_1), \dots, f(\lambda_k)$; two functions agreeing on $\operatorname{spec}(T)$ give the same operator, and f need not be defined, let alone continuous, anywhere else.

(e) For $p \in F[x]$, the operator $\sum_i p(\lambda_i) P_i$ coincides with the operator $p(T)$ of Definition 2.4.2. Without it, the symbol $p(T)$ would carry two unrelated meanings at once, and no computation mixing polynomials with other functions — for instance $(\sqrt{T})^2 = T$ — could be trusted.

(f) id and the reflection $T_{\mathbf{D}}$ with $\mathbf{D} = \operatorname{diag}(1, -1)$, both self-adjoint with square id. Theorem 12.8.4 singles out id, because $\operatorname{spec}(\mathrm{id}) = \{1\} \subseteq [0, \infty)$ while $\operatorname{spec}(T_{\mathbf{D}}) = \{1, -1\}$ is not. (A third root is $-\mathrm{id}$, and a fourth is $T_{\mathbf{D}'}$ with $\mathbf{D}' = \operatorname{diag}(-1, 1)$.)

12.8.7.2 B. Practice

Exercise. B1: Write these as $\sum \lambda_i \mathbf{P}_i$

Find the spectral resolution of each of the following matrices, and check in each case that the projections are self-adjoint, idempotent, mutually annihilating, and sum to $\mathbf{I}$.

(a) $\mathbf{A} = \begin{pmatrix} 2 & 1-i \\ 1+i & 3 \end{pmatrix} \in M_2(\mathbb{C})$.

(b) $\mathbf{S} = \begin{pmatrix} 2 & 0 & 1 \\ 0 & 3 & 0 \\ 1 & 0 & 2 \end{pmatrix} \in M_3(\mathbb{R})$.

(c) $\mathbf{U} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} \in M_2(\mathbb{C})$.

Solution.

(a) $\mathbf{A}^* = \mathbf{A}$, $\operatorname{tr} \mathbf{A} = 5$, and $\det \mathbf{A} = 6 - (1-i)(1+i) = 4$, so $p_{\mathbf{A}} = x^2 - 5x + 4$ and $\operatorname{spec}(\mathbf{A}) = \{1, 4\}$. From $\mathbf{A} - 4\mathbf{I} = \begin{pmatrix} -2 & 1-i \\ 1+i & -1 \end{pmatrix}$ the first row gives the eigenvector $\mathbf{v}_4 = (1-i, 2)$, with $\|\mathbf{v}_4\|^2 = 2 + 4 = 6$; from $\mathbf{A} - \mathbf{I}$ likewise $\mathbf{v}_1 = (-1+i, 1)$, with $\|\mathbf{v}_1\|^2 = 3$. Hence

$$\mathbf{P}_4 = \frac{\mathbf{v}_4 \mathbf{v}_4^*}{6} = \frac{1}{3} \begin{pmatrix} 1 & 1-i \\ 1+i & 2 \end{pmatrix}, \qquad \mathbf{P}_1 = \frac{1}{3} \begin{pmatrix} 2 & -1+i \\ -1-i & 1 \end{pmatrix},$$

and $\mathbf{A} = 1 \cdot \mathbf{P}_1 + 4 \cdot \mathbf{P}_4$. Both are Hermitian, both square to themselves, $\mathbf{P}_1\mathbf{P}_4 = \mathbf{O}$,

and $\mathbf{P}_1 + \mathbf{P}_4 = \mathbf{I}$.

(b) $\mathbf{S}$ is symmetric, and $\mathbf{S}(1,0,1) = 3(1,0,1)$, $\mathbf{S}(0,1,0) = 3(0,1,0)$, $\mathbf{S}(1,0,-1) = (1,0,-1)$, so $\operatorname{spec}(\mathbf{S}) = \{3,1\}$ with $E_1 = \operatorname{span}((1,0,-1))$ and $E_3 = \operatorname{span}((1,0,1),(0,1,0))$. Hence

$$\mathbf{P}_1 = \frac{1}{2}\begin{pmatrix} 1 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 1 \end{pmatrix}, \qquad \mathbf{P}_3 = \mathbf{I} - \mathbf{P}_1 = \frac{1}{2}\begin{pmatrix} 1 & 0 & 1 \\ 0 & 2 & 0 \\ 1 & 0 & 1 \end{pmatrix},$$

and $\mathbf{S} = \mathbf{P}_1 + 3\mathbf{P}_3$. The stated properties follow from $\mathbf{P}_1^\top = \mathbf{P}_1$, $\mathbf{P}_1^2 = \mathbf{P}_1$, and $\mathbf{P}_3 = \mathbf{I} - \mathbf{P}_1$.

(c) $\mathbf{U}$ is real orthogonal, hence unitary and normal. Since $\mathbf{U}(1,-i) = \frac{1}{\sqrt{2}}(1+i, 1-i) = e^{i\pi/4}(1,-i)$, and similarly $\mathbf{U}(1,i) = e^{-i\pi/4}(1,i)$, we get $\operatorname{spec}(\mathbf{U}) = \{e^{\pm i\pi/4}\}$. Both eigenvectors have squared length 2, so with $\mathbf{v}_\pm = (1, \mp i)$,

$$\mathbf{P}_\pm = \frac{\mathbf{v}_\pm \mathbf{v}_\pm^*}{2} = \frac{1}{2}\begin{pmatrix} 1 & \pm i \\ \mp i & 1 \end{pmatrix}, \qquad \mathbf{U} = e^{i\pi/4}\mathbf{P}_+ + e^{-i\pi/4}\mathbf{P}_-.$$

Both are Hermitian and idempotent, $\mathbf{P}_+\mathbf{P}_- = \mathbf{O}$ and $\mathbf{P}_+ + \mathbf{P}_- = \mathbf{I}$.

Exercise. B2: Computing through the resolution

Let $\mathbf{C} = \begin{pmatrix} 13 & 12 \\ 12 & 13 \end{pmatrix} \in M_2(\mathbb{R})$.

(a) Find the spectral resolution of $\mathbf{C}$.

(b) Hence compute $\sqrt{\mathbf{C}}$, and verify by squaring.

(c) Hence give a closed formula for $\mathbf{C}^m$ and write down $\mathbf{C}^{100}$.

(d) Hence compute $e^{\mathbf{C}}$.

Solution.

(a) $\mathbf{C}(1,1) = 25(1,1)$ and $\mathbf{C}(1,-1) = (1,-1)$, so $\operatorname{spec}(\mathbf{C}) = \{25,1\}$ and, both spanning vectors having squared length 2,

$$\mathbf{P}_1 = \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}, \qquad \mathbf{P}_2 = \frac{1}{2}\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}, \qquad \mathbf{C} = 25\mathbf{P}_1 + \mathbf{P}_2.$$

(b) Both eigenvalues are positive, so by Theorem 12.8.4,

$$\sqrt{\mathbf{C}} = 5\mathbf{P}_1 + 1 \cdot \mathbf{P}_2 = \begin{pmatrix} 3 & 2 \\ 2 & 3 \end{pmatrix},$$

and $\begin{pmatrix} 3 & 2 \\ 2 & 3 \end{pmatrix}^2 = \begin{pmatrix} 13 & 12 \\ 12 & 13 \end{pmatrix} = \mathbf{C}$.

(c) By Theorem 12.8.3 (a), $\mathbf{C}^m = 25^m\mathbf{P}_1 + 1^m\mathbf{P}_2$, that is

$$\mathbf{C}^m = \frac{1}{2}\begin{pmatrix} 25^m + 1 & 25^m - 1 \\ 25^m - 1 & 25^m + 1 \end{pmatrix},$$

so $\mathbf{C}^{100} = \frac{1}{2}\left(\begin{smallmatrix} 25^{100}+1 & 25^{100}-1 \\ 25^{100}-1 & 25^{100}+1 \end{smallmatrix}\right)$. (As a check, $m = 2$ gives $\frac{1}{2}\left(\begin{smallmatrix} 626 & 624 \\ 624 & 626 \end{smallmatrix}\right) = \left(\begin{smallmatrix} 313 & 312 \\ 312 & 313 \end{smallmatrix}\right)$, which is $\mathbf{C}^2$.)

(d) By Theorem 12.8.5 (a),

$$e^{\mathbf{C}} = e^{25}\mathbf{P}_1 + e\,\mathbf{P}_2 = \frac{1}{2}\begin{pmatrix} e^{25}+e & e^{25}-e \\ e^{25}-e & e^{25}+e \end{pmatrix}.$$

Exercise. B3: Is it a spectral resolution?

Determine which of the following displays the spectral resolution of the operator $T_{\mathbf{A}}$ on the stated space. Justify your answer; for those that do not, name the condition of Theorem 12.8.2 that fails and say what the correct resolution is.

(a) On $\mathbb{R}^2$: $\mathbf{A} = \begin{pmatrix} 2 & 0 \\ 0 & -1 \end{pmatrix}$, with $\mu_1 = 2$, $\mathbf{Q}_1 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$, $\mu_2 = -1$, $\mathbf{Q}_2 = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$.

(b) On $\mathbb{R}^2$: $\mathbf{A} = 3\mathbf{I}$, with $\mu_1 = \mu_2 = 3$, $\mathbf{Q}_1 = \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$, $\mathbf{Q}_2 = \frac{1}{2}\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$.

(c) On $\mathbb{R}^2$: $\mathbf{A} = \begin{pmatrix} 3 & 1 \\ 0 & 1 \end{pmatrix}$, with $\mu_1 = 3$, $\mathbf{Q}_1 = \frac{1}{2}\begin{pmatrix} 2 & 1 \\ 0 & 0 \end{pmatrix}$, $\mu_2 = 1$, $\mathbf{Q}_2 = \frac{1}{2}\begin{pmatrix} 0 & -1 \\ 0 & 2 \end{pmatrix}$.

Solution.

(a) Yes. The two matrices are symmetric, idempotent, nonzero, annihilate each other, and sum to $\mathbf{I}$, and $2\mathbf{Q}_1 - \mathbf{Q}_2 = \mathbf{A}$; the coefficients 2 and -1 are distinct. By Theorem 12.8.2 this is the spectral resolution, and indeed $E_2 = \operatorname{span}(\mathbf{e}_1)$, $E_{-1} = \operatorname{span}(\mathbf{e}_2)$.

(b) No: the scalars μ_1, μ_2 are not **distinct**, so the hypothesis of Theorem 12.8.2 fails at once. The identity $3\mathbf{Q}_1 + 3\mathbf{Q}_2 = 3\mathbf{I} = \mathbf{A}$ is true but uninformative — the eigenvalue 3 has been split artificially across two projections. The spectral resolution of $3\mathbf{I}$ has a single term, $3\mathbf{I} = 3 \cdot \mathrm{id}$, since $\operatorname{spec}(3\mathbf{I}) = \{3\}$ and $E_3 = \mathbb{R}^2$.

(c) No: $\mathbf{Q}_1$ is not self-adjoint, since $\mathbf{Q}_1^{\top} = \frac{1}{2}\begin{pmatrix} 2 & 0 \\ 1 & 0 \end{pmatrix} \neq \mathbf{Q}_1$, so (R1) fails. Everything else holds: $\mathbf{Q}_1^2 = \mathbf{Q}_1$, $\mathbf{Q}_2 = \mathbf{I} - \mathbf{Q}_1$, $\mathbf{Q}_1\mathbf{Q}_2 = \mathbf{Q}_2\mathbf{Q}_1 = \mathbf{O}$, and $3\mathbf{Q}_1 + \mathbf{Q}_2 = \mathbf{A}$. There is no correct resolution to offer, because $\mathbf{A}$ is not normal: $\mathbf{A}^{\top}\mathbf{A} = \begin{pmatrix} 9 & 3 \\ 3 & 2 \end{pmatrix}$ while $\mathbf{A}\mathbf{A}^{\top} = \begin{pmatrix} 10 & 1 \\ 1 & 1 \end{pmatrix}$. The projections here are onto $E_3 = \operatorname{span}(\mathbf{e}_1)$ along $E_1 = \operatorname{span}((-1, 2))$, two lines that are not perpendicular.

12.8.7.3 C. Going deeper

Exercise. C1: The projections are polynomials in the operator

Let $T \in \mathcal{L}(V)$ satisfy $(*)$, with distinct eigenvalues $\lambda_1, \ldots, \lambda_k \in F$ and resolution of the identity $(P_1, \ldots, P_k)$.

(a) Prove that $P_i = \ell_i(T)$, where $\ell_i \in F[x]$ is the i-th Lagrange basis polynomial for the

nodes $\lambda_1, \dots, \lambda_k$.

(b) Deduce that for every $f\colon \operatorname{spec}(T) \to F$ there is a polynomial p of degree less than k with $f(T) = p(T)$.

(c) Deduce that every operator commuting with T commutes with each P_i, re-proving Theorem 12.8.3 (e) by a different route.

Hint for (a): Definition 6.6.1 produces a polynomial taking prescribed values at prescribed nodes.

Solution.

(a) The λ_i are k distinct elements of F, so Definition 6.6.1 gives $\ell_1, \dots, \ell_k \in F[x]$ of degree $k-1$, and Theorem 6.6.1 (a) gives $\ell_i(\lambda_j) = \delta_{ij}$. By Theorem 12.8.3 (a),

$$\ell_i(T) = \sum_{j=1}^{k} \ell_i(\lambda_j) P_j = \sum_{j=1}^{k} \delta_{ij} P_j = P_i.$$

(b) Let p be the interpolating polynomial of degree less than k with $p(\lambda_i) = f(\lambda_i)$ for every i, which exists by Theorem 6.6.1. By Theorem 12.8.3 (a),

$$p(T) = \sum_{i} p(\lambda_i) P_i = \sum_{i} f(\lambda_i) P_i = f(T).$$

Explicitly, $p = \sum_i f(\lambda_i)\ell_i$.

(c) If $ST = TS$, then S commutes with every polynomial in T by Theorem 6.7.1 (d), in particular with $\ell_i(T) = P_i$.

Exercise. C2: When the projections have rank one

Let $T \in \mathcal{L}(V)$ satisfy $(*)$, with $\dim V = n$ and resolution of the identity $(P_1, \dots, P_k)$.

(a) Prove that $\operatorname{rank} P_i = \dim E_{\lambda_i}(T)$ and that $\sum_i \operatorname{rank} P_i = n$.

(b) Deduce that every P_i has rank 1 if and only if T has n distinct eigenvalues.

(c) Suppose every P_i has rank 1. Prove that every $S \in \mathcal{L}(V)$ commuting with T is a polynomial in T.

Solution.

(a) $\operatorname{im} P_i = E_{\lambda_i}(T)$ by Definition 11.3.2, so $\operatorname{rank} P_i = \dim E_{\lambda_i}(T)$. By Theorem 12.8.1 (a), $V = E_1 \oplus \cdots \oplus E_k$, so the dimensions add to n.

(b) If every $\operatorname{rank} P_i = 1$, then (a) gives $k = n$, so T has n distinct eigenvalues. Conversely, if T has n distinct eigenvalues, then $k = n$ and (a) forces each of the n positive integers $\operatorname{rank} P_i$ to be 1.

(c) By (b), $k = n$ and each E_i is a line, say $E_i = \operatorname{span}(e_i)$ with $(e_1, \dots, e_n)$ an orthonormal eigenbasis. Let $ST = TS$. As in the proof of Theorem 12.8.3 (e), S maps E_i into itself, so $Se_i = c_i e_i$ for some $c_i \in F$. Define $f\colon \operatorname{spec}(T) \to F$ by $f(\lambda_i) = c_i$. Then $f(T)e_i = c_i e_i = Se_i$ for every i, and two operators agreeing on a basis are equal, so $S = f(T)$. By Exercise (C1: The projections are polynomials in the operator) (b),

$f(T)$ is a polynomial in T.

Exercise. C3: What the functional calculus forgets

Let $T \in \mathcal{L}(V)$ satisfy $(*)$, with distinct eigenvalues $\lambda_1, \dots, \lambda_k$, and let $f, g\colon \operatorname{spec}(T) \to F$.

(a) Prove that $f(T) = 0$ if and only if $f(\lambda_i) = 0$ for every i. Deduce that $f(T) = g(T)$ if and only if $f = g$.

(b) Deduce that $\{p(T) : p \in F[x]\} = \operatorname{span}(P_1, \dots, P_k)$, a subspace of $\mathcal{L}(V)$ of dimension exactly k.

(c) Deduce that $m_T = (x - \lambda_1)(x - \lambda_2)\cdots(x - \lambda_k)$, so that the minimal polynomial of an operator satisfying $(*)$ has no repeated root.

Solution.

(a) $(\Leftarrow)$ is immediate from Definition 12.8.2. For $(\Rightarrow)$, fix i and pick $v \in E_{\lambda_i}(T)$ with $v \neq 0$, which is possible because λ_i is an eigenvalue. Then $P_i v = v$ and $P_j v = 0$ for $j \neq i$, since $E_i \subseteq \ker P_j$ (Theorem 12.8.1 (a), (b)). Hence

$$0 = f(T)v = \sum_j f(\lambda_j) P_j v = f(\lambda_i) v,$$

and $v \neq 0$ forces $f(\lambda_i) = 0$. For the second claim, apply this to $f - g$, using Theorem 12.8.3 (b).

(b) By Theorem 12.8.3 (a), every $p(T)$ equals $\sum_i p(\lambda_i) P_i$, so the left-hand side is contained in $\operatorname{span}(P_1, \dots, P_k)$; conversely every $\sum_i c_i P_i$ is $f(T)$ for the function $f(\lambda_i) = c_i$, hence is a polynomial in T by Exercise (C1: The projections are polynomials in the operator) (b). The P_i are linearly independent: if $\sum_i c_i P_i = 0$, that operator is $f(T)$ for $f(\lambda_i) = c_i$, so every $c_i = 0$ by (a). Hence the dimension is k.

(c) Put $p = \prod_i (x - \lambda_i)$, monic of degree k. Since $p(\lambda_i) = 0$ for every i, part (a) gives $p(T) = 0$, so $m_T \mid p$ by Theorem 9.5.1. Conversely $m_T(T) = 0$, so $m_T(\lambda_i) = 0$ for every i by (a), and hence each $x - \lambda_i$ divides m_T (Theorem 6.1.3 (b)). Distinct monic linear polynomials are coprime, since a common divisor would divide their difference, a nonzero constant; so $p = \prod_i (x - \lambda_i)$ divides m_T by Lemma 6.6.3 (b). Two monic polynomials dividing each other are equal, so $m_T = p$, a product of distinct linear factors.

12.9 Circulants and the Discrete Fourier Transform

The spectral theorem of Section 4 gives every normal matrix a unitary of its own, found by solving that matrix's eigenvalue problem. This section does something better for one particular family: it writes down a **single** unitary matrix, before any matrix of the family is named, and that one matrix diagonalizes all of them at once. The family is the circulants, the unitary is built from the n-th roots of unity, and the resulting dictionary between multiplying matrices and multiplying numbers is the discrete Fourier transform.

Two earlier pages pointed here. In Example (Which of these are normal?) (d) a 3×3 matrix whose rows were cyclic shifts of one another turned out to be normal, with the row dot products and the column dot products agreeing entry for entry; Exercise (B2: A circulant, diagonalized) met the same phenomenon again and diagonalized one such matrix by hand. Both times the agreement was described and not explained. It is explained below, and the explanation is the first theorem of the section.

Throughout, the field is $\mathbb{C}$, and **rows, columns, and vector entries are numbered** $0, 1, \ldots, n-1$. That is a departure from the rest of the book, and it is worth it: every formula here involves an index read modulo n, and modular arithmetic starts at 0.

12.9.1 Matrices whose rows walk around

Fix $n \geq 1$ and let $\mathbf{S} \in M_n(\mathbb{C})$ be the matrix with $(\mathbf{S})_{jk} = 1$ when $k \equiv j+1 \pmod{n}$ and $(\mathbf{S})_{jk} = 0$ otherwise: a 1 in each position $(j, j+1)$ for $j \leq n-2$, one more in the corner position $(n-1, 0)$, and zeros elsewhere. For $n = 4$,

$$\mathbf{S} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{pmatrix}.$$

Applied to a vector, $\mathbf{S}$ moves every entry one place up and sends the top entry around to the bottom: $\mathbf{Sx} = (x_1, x_2, \ldots, x_{n-1}, x_0)$. We call $\mathbf{S}$ the **cyclic shift**. Each row and each column of $\mathbf{S}$ carries exactly one 1, so $\mathbf{S}$ is a permutation matrix (Definition 3.4.2), and shifting n times returns every entry to where it started, so $\mathbf{S}^n = \mathbf{I}$.

Now think of $\mathbf{S}$ as a building block. The powers $\mathbf{I}, \mathbf{S}, \mathbf{S}^2, \ldots, \mathbf{S}^{n-1}$ are the n distinct "rotations of the identity", and a general linear combination of them is the object of this section.

A circulant is a matrix that does not know where the list begins: shift the indices and nothing changes.

Definition 12.9.1. Circulant matrix

Let $n \geq 1$ and let $\mathbf{c} = (c_0, c_1, \ldots, c_{n-1}) \in \mathbb{C}^n$. The **circulant** with first row $\mathbf{c}$ is the matrix $\mathbf{C} \in M_n(\mathbb{C})$ whose entries are

$$(\mathbf{C})_{jk} = c_{k-j} \qquad (0 \leq j, k \leq n-1),$$

where the subscript $k - j$ is read **modulo** n. The polynomial

$$p_{\mathbf{c}}(x) = c_0 + c_1 x + c_2 x^2 + \cdots + c_{n-1} x^{n-1} \in \mathbb{C}[x]$$

is called the **symbol** of $\mathbf{C}$.

In words: the entry in a given position depends only on how far the column is to the right of the diagonal, counted cyclically. Written out,

$$\mathbf{C} = \begin{pmatrix} c_0 & c_1 & \cdots & c_{n-1} \\ c_{n-1} & c_0 & \cdots & c_{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ c_1 & c_2 & \cdots & c_0 \end{pmatrix},$$

so each row is the row above it moved one step to the right, with the entry that falls off the right-hand end reappearing on the left. The columns behave the same way going down. Note that the data is one vector of length n, not n^2 numbers: a circulant is a very small object wearing a large costume.

Three examples, simplest first. The vector $\mathbf{c} = (1, 0, \dots, 0)$ gives $\mathbf{C} = \mathbf{I}$, and $\mathbf{c} = (0, 1, 0, \dots, 0)$ gives $\mathbf{C} = \mathbf{S}$; in both cases the definition is checked by reading off $(\mathbf{C})_{jk} = c_{k-j}$, which is 1 exactly when $k = j$, respectively $k = j + 1$, modulo n. The vector $\mathbf{c} = (1, 1, \dots, 1)$ gives the all-ones matrix. The 3×3 matrix of Example (Which of these are normal?) (d) is the circulant with first row $(1, 2, 3)$, and the matrix of Exercise (B2: A circulant, diagonalized) is the circulant with first row $(1, 1, 0)$.

Our running example for the whole section is the 4×4 circulant with first row $(1, 2, 3, 4)$:

$$\mathbf{C} = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 4 & 1 & 2 & 3 \\ 3 & 4 & 1 & 2 \\ 2 & 3 & 4 & 1 \end{pmatrix}, \qquad p_{\mathbf{c}}(x) = 1 + 2x + 3x^2 + 4x^3.$$

A nonexample by minimal change. Keep the first row $(1, 2, 3, 4)$ but let the entries that fall off the end vanish instead of wrapping around, giving a matrix $\mathbf{B}$ with rows $(1, 2, 3, 4)$, $(0, 1, 2, 3)$, $(0, 0, 1, 2)$, $(0, 0, 0, 1)$. Every diagonal of $\mathbf{B}$ is still constant, and its first row is still $\mathbf{c}$, but $(\mathbf{B})_{10} = 0$ while the definition demands $c_{0-1} = c_3 = 4$. The clause that fails is the one that does all the work here: the subscript must be read modulo n. Matrices with constant diagonals and no wrap-around have their own name and their own theory, and the warning at the end of the section says why they do not fit into this one.

The reason a circulant is worth a name is that it is a polynomial in the shift.

Theorem 12.9.1. Circulants Are Exactly the Polynomials in the Shift

Let $\mathbf{c} \in \mathbb{C}^n$ and let $\mathbf{C}$ be the circulant with first row $\mathbf{c}$, with symbol $p_{\mathbf{c}}$. Then:

(a) $\mathbf{C} = p_{\mathbf{c}}(\mathbf{S})$; conversely, $q(\mathbf{S})$ is a circulant for every $q \in \mathbb{C}[x]$.

(b) $\mathbf{C}^*$ is the circulant with first row $(\overline{c_0}, \overline{c_{n-1}}, \overline{c_{n-2}}, \dots, \overline{c_1})$.

(c) $\mathbf{C}$ is normal, and any two circulants of the same size commute.

Idea. For (a), read off the entries of $\mathbf{S}^m$: it has a 1 in position (j, k) exactly when $k = j + m$ modulo n. So $\mathbf{S}^m$ is the circulant that picks out the coefficient c_m, and summing over m rebuilds $\mathbf{C}$ one coefficient at a time. Once (a) is in hand, (b) and (c) are arithmetic with the single relation $\mathbf{S}^* = \mathbf{S}^{n-1}$: the adjoint of a polynomial in $\mathbf{S}$ is again a polynomial in $\mathbf{S}$, and two polynomials in one matrix always commute (Theorem 0.9.9).

Proof.

(a) We claim that $(\mathbf{S}^m)_{jk} = 1$ if $k \equiv j + m \pmod{n}$ and $(\mathbf{S}^m)_{jk} = 0$ otherwise, for every $m \geq 0$. For $m = 0$ this says $\mathbf{S}^0 = \mathbf{I}$. Assuming it for m, expand one product: $(\mathbf{S}^{m+1})_{jk} = \sum_i (\mathbf{S})_{ji} (\mathbf{S}^m)_{ik}$, and the only nonzero factor $(\mathbf{S})_{ji}$ sits at $i \equiv j + 1$, so the sum is $(\mathbf{S}^m)_{j+1,k}$, which is 1 exactly when $k \equiv j + 1 + m$. This proves the claim by

induction. Hence, for each j and k,

$$\big(p_{\mathbf{c}}(\mathbf{S})\big)_{jk} = \sum_{m=0}^{n-1} c_m (\mathbf{S}^m)_{jk} = c_{k-j},$$

because exactly one value of m in $\{0, \dots, n-1\}$ satisfies $m \equiv k - j \pmod{n}$. So $p_{\mathbf{c}}(\mathbf{S}) = \mathbf{C}$. Conversely, let $q \in \mathbb{C}[x]$. Dividing by $x^n - 1$ (Theorem 6.1.1) gives $q = h \cdot (x^n - 1) + r$ with $\deg r < n$, and $\mathbf{S}^n = \mathbf{I}$ gives $q(\mathbf{S}) = h(\mathbf{S})(\mathbf{S}^n - \mathbf{I}) + r(\mathbf{S}) = r(\mathbf{S})$, which is the circulant whose first row is the coefficient list of r.

(b) $\mathbf{S}$ is a permutation matrix with real entries, so $\mathbf{S}^* = \mathbf{S}^{\top} = \mathbf{S}^{-1}$ (Lemma 3.4.7 (b)), and $\mathbf{S}^{-1} = \mathbf{S}^{n-1}$ because $\mathbf{S}^n = \mathbf{I}$. Therefore

$$\mathbf{C}^* = \sum_{m=0}^{n-1} \overline{c_m}\, (\mathbf{S}^*)^m = \sum_{m=0}^{n-1} \overline{c_m}\, \mathbf{S}^{-m}.$$

For $m = 0$ the term is $\overline{c_0}\mathbf{I}$, and for $1 \le m \le n-1$ we have $\mathbf{S}^{-m} = \mathbf{S}^{n-m}$. So the coefficient of $\mathbf{S}^r$ is $\overline{c_0}$ when $r = 0$ and $\overline{c_{n-r}}$ when $1 \le r \le n-1$, which is the stated first row.

(c) By (a) and (b), both $\mathbf{C}$ and $\mathbf{C}^*$ are polynomials in $\mathbf{S}$, and any two polynomials in the same matrix commute (Theorem 0.9.9). Hence $\mathbf{C}^*\mathbf{C} = \mathbf{C}\mathbf{C}^*$. The same argument applies to two circulants $\mathbf{C} = p(\mathbf{S})$ and $\mathbf{C}' = q(\mathbf{S})$.

■

That settles the coincidence noticed in Example (Which of these are normal?) (d) and again in Exercise (B2: A circulant, diagonalized) (a). It is worth seeing it once at the level of entries, because the entries are what those pages compared. For any $\mathbf{A} \in M_n(\mathbb{C})$, writing $\mathbf{a}_0, \dots, \mathbf{a}_{n-1}$ for its columns and $\mathbf{r}_0, \dots, \mathbf{r}_{n-1}$ for its rows, the definition of the matrix product gives

$$(\mathbf{A}^*\mathbf{A})_{jk} = \langle \mathbf{a}_k, \mathbf{a}_j \rangle, \qquad (\mathbf{A}\mathbf{A}^*)_{jk} = \langle \mathbf{r}_j, \mathbf{r}_k \rangle.$$

So "normal" says exactly that the table of column inner products and the table of row inner products agree. Now let $\mathbf{A} = \mathbf{C}$ be a circulant, so that $(\mathbf{C})_{ik} = c_{k-i}$. Re-index the first sum by $t = j - i$ and the second by $t = i - k$; as i runs through all residues modulo n, so does t in each case. This gives

$$\langle \mathbf{a}_k, \mathbf{a}_j \rangle = \sum_{t=0}^{n-1} c_{t+k-j}\, \overline{c_t} = \langle \mathbf{r}_j, \mathbf{r}_k \rangle,$$

all subscripts modulo n. Both tables depend only on $k - j$, and they are literally the same sum. Cyclic shifting is what makes the re-indexing legal, and that is the whole of the coincidence.

Quick check. Which vectors $\mathbf{c} \in \mathbb{C}^n$ give a circulant that is Hermitian? Answer for $n = 4$ with real entries.

Solution. By Theorem 12.9.1 (b), $\mathbf{C}^* = \mathbf{C}$ forces $\overline{c_0} = c_0$ and $\overline{c_{n-m}} = c_m$ for $1 \le m \le n-1$. For $n = 4$ with real entries this reads $c_1 = c_3$, with c_0 and c_2 unconstrained: the symmetric real 4×4 circulants are exactly those whose first row has the form (c_0, c_1, c_2, c_1).

12.9.2 The eigenvectors of the shift

Everything now follows from one eigenvalue problem, the one for $\mathbf{S}$. Since $\mathbf{S}^n = \mathbf{I}$, any eigenvalue λ satisfies $\lambda^n = 1$, so the candidates are the n-th roots of unity of Example (Roots of Unity). There are n of them and $\mathbf{S}$ is an $n \times n$ matrix, which is encouraging. The eigenvectors are equally explicit.

Fix

$$\omega := e^{2\pi i/n},$$

so that $1, \omega, \omega^2, \dots, \omega^{n-1}$ are the n distinct n-th roots of unity, and $\omega^r = 1$ if and only if n divides r (Example (Roots of Unity)).

Theorem 12.9.2. Eigenvectors of the Cyclic Shift

For $j = 0, 1, \dots, n-1$ let

$$\mathbf{f}_j := \tfrac{1}{\sqrt{n}}(1,\ \omega^j,\ \omega^{2j},\ \dots,\ \omega^{(n-1)j}) \in \mathbb{C}^n,$$

that is, $(\mathbf{f}_j)_m = \omega^{jm}/\sqrt{n}$. Then $\mathbf{S}\mathbf{f}_j = \omega^j \mathbf{f}_j$. Consequently $\mathbf{S}$ is a unitary matrix with the n distinct eigenvalues $1, \omega, \dots, \omega^{n-1}$, and each eigenspace $E_{\omega^j}(\mathbf{S})$ is the line $\operatorname{span}(\mathbf{f}_j)$.

Idea. The shift moves entry $m+1$ into slot m. If the entries are the successive powers of a fixed number z, that is the same as multiplying the whole vector by z — provided the wrap-around at the end also multiplies by z, which is exactly the condition $z^n = 1$. So the geometric vectors $(1, z, z^2, \dots)$ with $z^n = 1$ are forced on us, and there are n of them.

Proof. Fix j and let $0 \le m \le n-1$. By the description of $\mathbf{S}$, the m-th entry of $\mathbf{S}\mathbf{f}_j$ is the $(m+1)$-st entry of $\mathbf{f}_j$, indices modulo n. For $m < n-1$ this is

$$\frac{\omega^{j(m+1)}}{\sqrt{n}} = \omega^j \cdot \frac{\omega^{jm}}{\sqrt{n}} = \omega^j (\mathbf{f}_j)_m,$$

and for $m = n-1$ it is $(\mathbf{f}_j)_0 = 1/\sqrt{n}$, while $\omega^j (\mathbf{f}_j)_{n-1} = \omega^j \omega^{j(n-1)}/\sqrt{n} = \omega^{jn}/\sqrt{n} = 1/\sqrt{n}$ because $\omega^n = 1$. The two agree, so $\mathbf{S}\mathbf{f}_j = \omega^j \mathbf{f}_j$. Each $\mathbf{f}_j$ is nonzero, since its 0-th entry is $1/\sqrt{n}$, so ω^j is an eigenvalue.

The matrix $\mathbf{S}$ is a real permutation matrix, so $\mathbf{S}^\top \mathbf{S} = \mathbf{I}$ by Lemma 3.4.7 (b), and $\mathbf{S}^* = \mathbf{S}^\top$; hence $\mathbf{S}$ is unitary (Definition 11.7.2). The n numbers $1, \omega, \dots, \omega^{n-1}$ are distinct, so $\mathbf{S}$ has n distinct eigenvalues, and the sum of their eigenspaces is direct (Corollary 9.2.3). Each of those n eigenspaces has dimension at least 1 and their dimensions add up to at most $\dim \mathbb{C}^n = n$, so each has dimension exactly 1. Since $\mathbf{f}_j \in E_{\omega^j}(\mathbf{S})$ is nonzero, that eigenspace is $\operatorname{span}(\mathbf{f}_j)$.

■

The vectors $\mathbf{f}_0, \dots, \mathbf{f}_{n-1}$ are called the **Fourier vectors**. They are already known to be pairwise orthogonal: $\mathbf{S}$ is unitary, hence normal, and eigenvectors of a normal matrix for distinct eigenvalues are orthogonal (Corollary 12.3.4). But the direct computation is short, it produces the normalization at the same time, and it is the computation on which the rest of the section

runs. So we do it.

Theorem 12.9.3. The Fourier Matrix Is Unitary

The list $(\mathbf{f}_0, \mathbf{f}_1, \dots, \mathbf{f}_{n-1})$ is an orthonormal basis of $\mathbb{C}^n$. Equivalently, the **Fourier matrix**

$$\mathbf{F} \coloneqq \frac{1}{\sqrt{n}} \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & \omega & \cdots & \omega^{n-1} \\ \vdots & \vdots & & \vdots \\ 1 & \omega^{n-1} & \cdots & \omega^{(n-1)^2} \end{pmatrix}, \qquad (\mathbf{F})_{mj} = \frac{\omega^{mj}}{\sqrt{n}},$$

whose j-th column is $\mathbf{f}_j$, is unitary: $\mathbf{F}^*\mathbf{F} = \mathbf{F}\mathbf{F}^* = \mathbf{I}$.

Proof. Let $0 \le j, k \le n-1$. Since $|\omega| = 1$ we have $\overline{\omega} = \omega^{-1}$, hence $\overline{\omega^{km}} = \omega^{-km}$, and so

$$\langle \mathbf{f}_j, \mathbf{f}_k \rangle = \frac{1}{n} \sum_{m=0}^{n-1} \omega^{jm}\, \overline{\omega^{km}} = \frac{1}{n} \sum_{m=0}^{n-1} (\omega^{j-k})^m,$$

the inner product being linear in the first slot and conjugate-linear in the second.

Case $j = k$. Then $\omega^{j-k} = \omega^0 = 1$, so every one of the n terms equals 1 and the sum is n. Hence $\langle \mathbf{f}_j, \mathbf{f}_j \rangle = n/n = 1$, and $\|\mathbf{f}_j\| = 1$.

Case $j \neq k$. Put $z \coloneqq \omega^{j-k}$. Then $z^n = (\omega^n)^{j-k} = 1$. Also $z \neq 1$: by Example (Roots of Unity), $\omega^{j-k} = 1$ would force n to divide $j-k$, and $0 < |j-k| \le n-1$ makes that impossible. Now use the factorization

$$(z-1)(1 + z + z^2 + \cdots + z^{n-1}) = z^n - 1 = 0,$$

which is a polynomial identity valid for every complex z. Since $z - 1 \neq 0$, we may divide by it, and

$$\sum_{m=0}^{n-1} z^m = 0.$$

Hence $\langle \mathbf{f}_j, \mathbf{f}_k \rangle = 0$.

So the list is orthonormal; it has $n = \dim \mathbb{C}^n$ members and is independent (Theorem 11.1.4), hence a basis (Theorem 1.6.4). A matrix whose columns form an orthonormal basis of $\mathbb{C}^n$ is unitary (Theorem 11.7.3 (f)), which gives $\mathbf{F}^*\mathbf{F} = \mathbf{F}\mathbf{F}^* = \mathbf{I}$.

■

The second case is where the whole section lives. A sum of n unit complex numbers, spread evenly around the circle, cancels — and it cancels for the cheapest possible reason, that a geometric series with ratio $z \neq 1$ and $z^n = 1$ telescopes to zero.

Example. The Fourier matrix of size four

Write out $\mathbf{F}$ for $n = 4$, and check one orthogonality relation by hand.

Solution. For $n = 4$, $\omega = e^{2\pi i/4} = i$, and the powers cycle $i^0 = 1$, $i^1 = i$, $i^2 = -1$, $i^3 = -i$. So

$$\mathbf{F} = \frac{1}{2}\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & i & -1 & -i \\ 1 & -1 & 1 & -1 \\ 1 & -i & -1 & i \end{pmatrix}.$$

Take $j = 1$, $k = 3$. Then $\omega^{j-k} = i^{-2} = -1$, and

$$\langle \mathbf{f}_1, \mathbf{f}_3 \rangle = \tfrac{1}{4}(1 + (-1) + 1 + (-1)) = 0,$$

matching the geometric-series computation with $z = -1$, for which $1 + z + z^2 + z^3 = 0$. Note that $\mathbf{F}$ is symmetric, $\mathbf{F}^\top = \mathbf{F}$, but **not** Hermitian: $\mathbf{F}^* = \overline{\mathbf{F}} \neq \mathbf{F}$, since $(\mathbf{F})_{12} = i/2$.

12.9.3 One unitary for every circulant

Theorem 12.9.4. Circulant Diagonalization

Let $\mathbf{C} \in M_n(\mathbb{C})$ be the circulant with first row $\mathbf{c}$ and symbol $p_{\mathbf{c}}$. Then every Fourier vector is an eigenvector of $\mathbf{C}$,

$$\mathbf{C}\mathbf{f}_j = p_{\mathbf{c}}(\omega^j)\,\mathbf{f}_j \qquad (j = 0, 1, \dots, n-1),$$

and consequently

$$\mathbf{F}^*\mathbf{C}\mathbf{F} = \operatorname{diag}\big(p_{\mathbf{c}}(1),\ p_{\mathbf{c}}(\omega),\ \dots,\ p_{\mathbf{c}}(\omega^{n-1})\big).$$

The same unitary $\mathbf{F}$ works for every circulant of size n, and $\det \mathbf{C} = \prod_{j=0}^{n-1} p_{\mathbf{c}}(\omega^j)$.

Idea. There is nothing left to prove. A circulant is a polynomial in $\mathbf{S}$ by Theorem 12.9.1, and a polynomial in a matrix acts on an eigenvector of that matrix by the same polynomial applied to the eigenvalue. The point is not the proof but the statement: $\mathbf{F}$ was written down without reference to $\mathbf{c}$, so a single change of orthonormal basis handles the entire family at once.

Proof. By Theorem 12.9.1 (a), $\mathbf{C} = p_{\mathbf{c}}(\mathbf{S})$, and by Theorem 12.9.2, $\mathbf{S}\mathbf{f}_j = \omega^j \mathbf{f}_j$. Iterating the latter gives $\mathbf{S}^m \mathbf{f}_j = \omega^{jm}\mathbf{f}_j$ for every $m \geq 0$, so

$$\mathbf{C}\mathbf{f}_j = \sum_{m=0}^{n-1} c_m\, \mathbf{S}^m \mathbf{f}_j = \Big(\sum_{m=0}^{n-1} c_m \omega^{jm}\Big)\mathbf{f}_j = p_{\mathbf{c}}(\omega^j)\,\mathbf{f}_j.$$

The j-th column of $\mathbf{F}$ is $\mathbf{f}_j$, so the j-th column of $\mathbf{C}\mathbf{F}$ is $p_{\mathbf{c}}(\omega^j)\mathbf{f}_j$, which is the j-th column of $\mathbf{F}\mathbf{D}$ for $\mathbf{D} = \operatorname{diag}(p_{\mathbf{c}}(1), \dots, p_{\mathbf{c}}(\omega^{n-1}))$. Hence $\mathbf{C}\mathbf{F} = \mathbf{F}\mathbf{D}$, and multiplying on the left by $\mathbf{F}^* = \mathbf{F}^{-1}$ (Theorem 12.9.3) gives $\mathbf{F}^*\mathbf{C}\mathbf{F} = \mathbf{D}$. Taking determinants, $\det \mathbf{C} = \det \mathbf{D} = \prod_j p_{\mathbf{c}}(\omega^j)$, since $\det \mathbf{F}^* \det \mathbf{F} = \det(\mathbf{F}^*\mathbf{F}) = 1$ (Theorem 7.5.4).

■

This re-proves normality from the other side, and closes the account opened in Example (Which of these are normal?) (d): $\mathbf{C} = \mathbf{F}\mathbf{D}\mathbf{F}^*$ with $\mathbf{F}$ unitary and $\mathbf{D}$ diagonal is precisely the factorization that Corollary 12.4.3 reserves for normal matrices. The eigenvalues $p_{\mathbf{c}}(\omega^j)$ need no characteristic polynomial either: they are n evaluations of a polynomial of degree less than n.

Example. The running example, diagonalized

Find the eigenvalues and the determinant of the circulant $\mathbf{C}$ with first row $(1, 2, 3, 4)$, and write down $\mathbf{F}^*\mathbf{C}\mathbf{F}$.

Solution. Here $n = 4$, $\omega = i$, and $p_{\mathbf{c}}(x) = 1 + 2x + 3x^2 + 4x^3$. Evaluate at the four fourth roots of unity, using $i^2 = -1$ and $i^3 = -i$:

$$\begin{aligned} p_{\mathbf{c}}(1) &= 1 + 2 + 3 + 4 = 10, \\ p_{\mathbf{c}}(i) &= 1 + 2i - 3 - 4i = -2 - 2i, \\ p_{\mathbf{c}}(-1) &= 1 - 2 + 3 - 4 = -2, \\ p_{\mathbf{c}}(-i) &= 1 - 2i - 3 + 4i = -2 + 2i. \end{aligned}$$

So $\mathbf{F}^*\mathbf{C}\mathbf{F} = \operatorname{diag}(10, -2 - 2i, -2, -2 + 2i)$ with $\mathbf{F}$ as in Example (The Fourier matrix of size four), and

$$\det \mathbf{C} = 10 \cdot (-2 - 2i) \cdot (-2) \cdot (-2 + 2i) = -20(4 + 4) = -160,$$

using $(-2 - 2i)(-2 + 2i) = 4 + 4 = 8$. As a check on the first eigenvalue: every row of $\mathbf{C}$ sums to 10, so $\mathbf{C}\mathbf{1} = 10 \cdot \mathbf{1}$, and $\mathbf{f}_0 = \frac{1}{2}\mathbf{1}$.

Quick check. The matrix of Example (The running example, diagonalized) is real, and two of its eigenvalues are not. Why must those two be complex conjugates of each other, and which two are they in general?

Solution. If every c_m is real then $\overline{p_{\mathbf{c}}(z)} = p_{\mathbf{c}}(\overline{z})$ for every z. Taking $z = \omega^j$ and using $\overline{\omega^j} = \omega^{-j} = \omega^{n-j}$,

$$\overline{p_{\mathbf{c}}(\omega^j)} = p_{\mathbf{c}}(\omega^{n-j}),$$

so the eigenvalue attached to $\mathbf{f}_j$ and the one attached to $\mathbf{f}_{n-j}$ are conjugate. Here $n = 4$ and the pair is $j = 1, 3$: $\overline{-2 - 2i} = -2 + 2i$. The eigenvalues for $j = 0$ and, when n is even, $j = n/2$ are their own conjugates, hence real — which is why 10 and -2 came out real. (This is also Theorem 9.2.5, seen here with the eigenvectors named.)

Once every circulant is $\mathbf{F}\mathbf{D}\mathbf{F}^*$, the family closes up on itself.

Corollary 12.9.5. Circulants Form a Commutative Algebra

Let $\mathcal{C}_n \subseteq M_n(\mathbb{C})$ be the set of circulants of size n. Then:

(a) $\mathcal{C}_n$ is a subspace of $M_n(\mathbb{C})$ of dimension n, with basis $\mathbf{I}, \mathbf{S}, \dots, \mathbf{S}^{n-1}$;

(b) $\mathcal{C}_n$ is closed under products, and any two of its members commute;

(c) $\mathbf{A} \in \mathcal{C}_n$ if and only if $\mathbf{F}^*\mathbf{A}\mathbf{F}$ is diagonal;

(d) a circulant $\mathbf{C}$ is invertible if and only if $p_{\mathbf{c}}(\omega^j) \neq 0$ for every j, and then $\mathbf{C}^{-1}$ is again a circulant.

Proof.

(a) The map $\mathbf{c} \mapsto p_{\mathbf{c}}(\mathbf{S})$ from $\mathbb{C}^n$ to $M_n(\mathbb{C})$ is linear and, by Theorem 12.9.1 (a), its image is exactly $\mathcal{C}_n$. It is injective, because the first row of $p_{\mathbf{c}}(\mathbf{S})$ is $\mathbf{c}$. So $\mathcal{C}_n$ is a subspace of dimension n, and the images $\mathbf{I}, \mathbf{S}, \dots, \mathbf{S}^{n-1}$ of the standard basis form a basis of it.

(b) If $\mathbf{C} = p(\mathbf{S})$ and $\mathbf{C}' = q(\mathbf{S})$, then $\mathbf{C}\mathbf{C}' = (pq)(\mathbf{S})$ (Theorem 0.9.9), which is a circulant by Theorem 12.9.1 (a). Commutativity is part (c) of that theorem.

(c) ($\Rightarrow$) is Theorem 12.9.4. ($\Leftarrow$) Suppose $\mathbf{F}^*\mathbf{A}\mathbf{F} = \mathbf{D} = \operatorname{diag}(d_0, \dots, d_{n-1})$. The n numbers $1, \omega, \dots, \omega^{n-1}$ are distinct, so by Lagrange interpolation (Theorem 6.6.1) there is $q \in \mathbb{C}[x]$ with $q(\omega^j) = d_j$ for every j. By Theorem 12.9.4 applied to the circulant $q(\mathbf{S})$, which has symbol the remainder of q on division by $x^n - 1$,

$$\mathbf{F}^* q(\mathbf{S}) \mathbf{F} = \operatorname{diag}(q(1), \dots, q(\omega^{n-1})) = \mathbf{D}.$$

Multiplying both sides by $\mathbf{F}$ on the left and $\mathbf{F}^*$ on the right gives $q(\mathbf{S}) = \mathbf{A}$, so $\mathbf{A}$ is a circulant.

(d) From $\mathbf{C} = \mathbf{F}\mathbf{D}\mathbf{F}^*$, $\det \mathbf{C} = \det \mathbf{D}$, so $\mathbf{C}$ is invertible exactly when no diagonal entry $p_{\mathbf{c}}(\omega^j)$ vanishes (Theorem 7.5.2). In that case $\mathbf{C}^{-1} = \mathbf{F}\mathbf{D}^{-1}\mathbf{F}^*$, and $\mathbf{F}^*\mathbf{C}^{-1}\mathbf{F} = \mathbf{D}^{-1}$ is diagonal, so $\mathbf{C}^{-1}$ is a circulant by (c).

■

Part (c) is the sharpest statement of the section: being a circulant is not a statement about the pattern of the entries at all, but a statement about which orthonormal basis diagonalizes the matrix. The pattern was only ever a way of recognizing it.

12.9.4 Multiplying circulants is convolution

A circulant is its first row. So the product of two circulants must be computable from the two first rows, and it is worth knowing what that computation is.

Given $\mathbf{c}, \mathbf{d} \in \mathbb{C}^n$, define their **cyclic convolution** $\mathbf{c} * \mathbf{d} \in \mathbb{C}^n$ by

$$(\mathbf{c} * \mathbf{d})_r := \sum_{l=0}^{n-1} c_l \, d_{r-l} \qquad (r = 0, \dots, n-1),$$

subscripts modulo n. In words: to get entry r, slide the reversed list $\mathbf{d}$ along $\mathbf{c}$ so that the indices add up to r, and add the products. It is the multiplication of polynomials with the exponents read modulo n, and the next theorem says so.

Theorem 12.9.6. Convolution Theorem

Let $\mathbf{c}, \mathbf{d} \in \mathbb{C}^n$ with circulants $\mathbf{C}, \mathbf{C}'$ and symbols $p_{\mathbf{c}}, p_{\mathbf{d}}$. Then:

(a) $\mathbf{C}\mathbf{C}'$ is the circulant with first row $\mathbf{c} * \mathbf{d}$;

(b) for every j, $p_{\mathbf{c}*\mathbf{d}}(\omega^j) = p_{\mathbf{c}}(\omega^j)\, p_{\mathbf{d}}(\omega^j)$.

In words: convolving the first rows corresponds to multiplying the eigenvalue lists entry by entry.

Idea. Both parts are the identity $p_{\mathbf{c}}p_{\mathbf{d}} \equiv p_{\mathbf{c}*\mathbf{d}} \pmod{x^n - 1}$, looked at twice. Multiplying the two symbols produces a polynomial of degree up to $2n-2$; reducing the exponents modulo n folds the coefficients back into n slots, and the fold is precisely the convolution sum. Evaluating at a root of $x^n - 1$ makes the reduction invisible, which is (b); substituting $\mathbf{S}$, for which $\mathbf{S}^n = \mathbf{I}$, makes it invisible as well, which is (a).

Proof. Expand the product of the symbols and collect exponents modulo n:

$$p_{\mathbf{c}}(x)\, p_{\mathbf{d}}(x) = \sum_{l=0}^{n-1} \sum_{m=0}^{n-1} c_l d_m x^{l+m}.$$

For each $r \in \{0, \dots, n-1\}$, the pairs (l, m) with $l + m \equiv r \pmod{n}$ are exactly those with $m \equiv r - l$, one value of m for each l. Hence, working modulo $x^n - 1$ and using $x^{l+m} \equiv x^{(l+m) \bmod n}$,

$$p_{\mathbf{c}}(x)\, p_{\mathbf{d}}(x) \equiv \sum_{r=0}^{n-1} \Big(\sum_{l=0}^{n-1} c_l d_{r-l} \Big) x^r = p_{\mathbf{c}*\mathbf{d}}(x) \pmod{x^n - 1}.$$

(a) Substitute $\mathbf{S}$. Since $\mathbf{S}^n = \mathbf{I}$, the multiple of $x^n - 1$ contributes nothing, so

$$\mathbf{C}\mathbf{C}' = p_{\mathbf{c}}(\mathbf{S})\, p_{\mathbf{d}}(\mathbf{S}) = p_{\mathbf{c}*\mathbf{d}}(\mathbf{S}),$$

the first equality by Theorem 12.9.1 (a) and Theorem 0.9.9. By Theorem 12.9.1 (a) again, $p_{\mathbf{c}*\mathbf{d}}(\mathbf{S})$ is the circulant with first row $\mathbf{c} * \mathbf{d}$, the coefficient list having length n.

(b) Substitute $x = \omega^j$, a root of $x^n - 1$, so that the multiple of $x^n - 1$ again contributes nothing. This gives $p_{\mathbf{c}}(\omega^j) p_{\mathbf{d}}(\omega^j) = p_{\mathbf{c}*\mathbf{d}}(\omega^j)$.

■

Part (b) deserves its own sentence, because it is the reason the discrete Fourier transform is used at all. Define the **discrete Fourier transform** of $\mathbf{c} \in \mathbb{C}^n$ to be the vector of eigenvalues,

$$\widehat{\mathbf{c}} := \big(p_{\mathbf{c}}(1),\ p_{\mathbf{c}}(\omega),\ \dots,\ p_{\mathbf{c}}(\omega^{n-1})\big) = \sqrt{n}\, \mathbf{F}\mathbf{c},$$

the second expression because the j-th entry of $\mathbf{F}\mathbf{c}$ is $n^{-1/2} \sum_m \omega^{jm} c_m$. Then (b) reads

$$\widehat{(\mathbf{c} * \mathbf{d})}_j = \widehat{\mathbf{c}}_j\, \widehat{\mathbf{d}}_j \qquad (j = 0, \dots, n-1):$$

the transform turns convolution into entrywise multiplication. Convolution is a tangled sum of n^2 products; after one unitary change of basis it is n ordinary multiplications. Everything the transform buys comes from this line.

Example. Convolution in size four

Let $\mathbf{c} = (1, 2, 3, 4)$ and $\mathbf{d} = (1, 1, 0, 0)$. Compute $\mathbf{c} * \mathbf{d}$, and verify Theorem 12.9.6 (b) at all

four roots of unity.

Solution. Since $d_0 = d_1 = 1$ and $d_2 = d_3 = 0$, the sum $\sum_l c_l d_{r-l}$ keeps only the terms with $r - l \equiv 0$ or 1, that is $l = r$ and $l = r - 1$. So $(\mathbf{c} * \mathbf{d})_r = c_r + c_{r-1}$, subscripts modulo 4:

$$\mathbf{c} * \mathbf{d} = (1+4,\ 2+1,\ 3+2,\ 4+3) = (5, 3, 5, 7).$$

The three symbols are $p_{\mathbf{c}}(x) = 1 + 2x + 3x^2 + 4x^3$, $p_{\mathbf{d}}(x) = 1 + x$, and $p_{\mathbf{c}*\mathbf{d}}(x) = 5 + 3x + 5x^2 + 7x^3$. Evaluating each at $1, i, -1, -i$, with $p_{\mathbf{c}}$ already computed in Example (The running example, diagonalized):

$$\begin{aligned}
j = 0: &\quad 10 \cdot 2 = 20, && 5 + 3 + 5 + 7 = 20;\\
j = 1: &\quad (-2 - 2i)(1 + i) = -4i, && 5 + 3i - 5 - 7i = -4i;\\
j = 2: &\quad (-2) \cdot 0 = 0, && 5 - 3 + 5 - 7 = 0;\\
j = 3: &\quad (-2 + 2i)(1 - i) = 4i, && 5 - 3i - 5 + 7i = 4i.
\end{aligned}$$

All four agree. For the middle line, $(-2 - 2i)(1 + i) = -2 - 2i - 2i - 2i^2 = -4i$.

12.9.5 Solving a circulant system, and a word about speed

If $\mathbf{Cx} = \mathbf{b}$ with $\mathbf{C}$ an invertible circulant, then in the Fourier basis the system falls apart into n scalar equations. Writing $\mathbf{C} = \mathbf{F}\mathbf{D}\mathbf{F}^*$ and setting $\mathbf{y} = \mathbf{F}^*\mathbf{x}$, $\mathbf{g} = \mathbf{F}^*\mathbf{b}$, the system becomes $\mathbf{D}\mathbf{y} = \mathbf{g}$, that is $y_j = g_j/\lambda_j$ with $\lambda_j = p_{\mathbf{c}}(\omega^j)$. Then $\mathbf{x} = \mathbf{F}\mathbf{y}$. Three steps: transform, divide, transform back.

Example. A circulant system

Solve $\mathbf{Cx} = \mathbf{b}$ for the circulant $\mathbf{C}$ with first row $(1, 2, 3, 4)$ and $\mathbf{b} = (1, 1, -1, -1)$.

Solution. The eigenvalues are $\lambda = (10, -2 - 2i, -2, -2 + 2i)$ by Example (The running example, diagonalized). First transform $\mathbf{b}$: the j-th entry of $\mathbf{F}^*\mathbf{b}$ is $\frac{1}{2}\sum_m \overline{\omega^{jm}} b_m = \frac{1}{2}\sum_m (-i)^{jm} b_m$, so

$$\begin{aligned}
g_0 &= \tfrac{1}{2}(1 + 1 - 1 - 1) = 0,\\
g_1 &= \tfrac{1}{2}(1 + (-i) - (-1) - (i)) = 1 - i,\\
g_2 &= \tfrac{1}{2}(1 - 1 - 1 + 1) = 0,\\
g_3 &= \tfrac{1}{2}(1 + i + 1 + i) = 1 + i.
\end{aligned}$$

Now divide. The two zero entries stay zero, and

$$y_1 = \frac{1 - i}{-2 - 2i} = \frac{(1 - i)(-2 + 2i)}{8} = \frac{4i}{8} = \frac{i}{2}, \qquad y_3 = \overline{y_1} = -\frac{i}{2},$$

where the middle step multiplies numerator and denominator by $\overline{-2 - 2i} = -2 + 2i$, whose product with $-2 - 2i$ is 8. Finally $\mathbf{x} = \mathbf{F}\mathbf{y}$, whose m-th entry is $\frac{1}{2}(i^m y_1 + i^{3m} y_3) = \frac{1}{2} \cdot \frac{i}{2}(i^m - (-i)^m)$:

$$\mathbf{x} = \left(0,\ -\tfrac{1}{2},\ 0,\ \tfrac{1}{2}\right).$$

Check directly: the second row of $\mathbf{C}$ is $(4,1,2,3)$, and $-\frac{1}{2}+\frac{3}{2}=1=b_1$; the third row is $(3,4,1,2)$, and $-2+1=-1=b_2$. The answer is real, as it must be: $\mathbf{C}$ and $\mathbf{b}$ are real, so $\mathbf{x}=\mathbf{C}^{-1}\mathbf{b}$ is real, and this shows up in the transform as the conjugate symmetry $y_3=\overline{y_1}$.

Counted honestly, the recipe above saves something, but not yet very much. Forming $\mathbf{F}^*\mathbf{b}$ as a matrix–vector product costs about n^2 multiplications, and so does forming $\mathbf{Fy}$, so the whole solve costs on the order of n^2. That beats general elimination, which is on the order of n^3, but it is no better than multiplying $\mathbf{b}$ by $\mathbf{C}^{-1}$ once that inverse has been found. What changes the picture is that $\mathbf{F}$ does not have to be applied as a general matrix. The powers of ω repeat, and splitting the sum $\sum_m \omega^{jm}b_m$ into its even and odd m reduces one transform of length n to two of length $n/2$; iterating brings the cost down to about $n\log n$. That algorithm is the **fast Fourier transform**. It proves no new theorem: the matrix $\mathbf{F}$, the factorization $\mathbf{C}=\mathbf{FDF}^*$ and everything above are unchanged, and the fast Fourier transform is only a way of applying $\mathbf{F}$ quickly. Chapter 24 develops it, together with the cost accounting that makes "quickly" precise.

Warning. A **Toeplitz** matrix — one with constant diagonals, $(\mathbf{A})_{jk}=a_{k-j}$ with the subscript **not** reduced modulo n — is not a circulant, and $\mathbf{F}$ does not diagonalize it. Take

$$\mathbf{A}=\begin{pmatrix}0&1&0\\1&0&1\\0&1&0\end{pmatrix}.$$

Every diagonal of $\mathbf{A}$ is constant, so $\mathbf{A}$ is Toeplitz, but the circulant with first row $(0,1,0)$ is $\mathbf{S}$, whose second row is $(0,0,1)$, not $(1,0,1)$. And $\mathbf{F}$ fails at the very first column: $\mathbf{f}_0=\frac{1}{\sqrt{3}}(1,1,1)$ and $\mathbf{Af}_0=\frac{1}{\sqrt{3}}(1,2,1)$, which is not a multiple of $\mathbf{f}_0$, so $\mathbf{f}_0$ is not an eigenvector and $\mathbf{F}^*\mathbf{AF}$ is not diagonal. Normality is not the issue: $\mathbf{A}$ is symmetric, hence normal, and Theorem 12.5.1 does hand it an orthonormal eigenbasis. It is simply a different basis, with the irrational eigenvalues $0,\pm\sqrt{2}$ in place of evaluations of a polynomial at roots of unity. What the wrap-around buys is not diagonalizability; it is the *same* diagonalizing basis for everybody.

12.9.6 Exercises

12.9.6.1 A. Check your understanding

Exercise. A1

(a) Define the circulant with first row $\mathbf{c}=(c_0,\dots,c_{n-1})$, and write down its symbol.

(b) State the eigenvalues and the eigenvectors of the cyclic shift $\mathbf{S}$ on $\mathbb{C}^n$.

(c) Give the eigenvalues of the circulant with first row $\mathbf{c}$, and say how many multiplications are needed to compute all of them from $\mathbf{c}$ by direct evaluation.

(d) Determine whether the following are true, with a reason. (i) Every Toeplitz matrix is a circulant. (ii) Every circulant is normal. (iii) The Fourier matrix $\mathbf{F}$ is Hermitian.

(e) Explain in one sentence why two circulants of the same size commute, without computing any entries.

Solution.

(a) The matrix $\mathbf{C} \in M_n(\mathbb{C})$ with $(\mathbf{C})_{jk} = c_{k-j}$, subscripts modulo n and indices running from 0 to $n-1$ (Definition 12.9.1). Its symbol is $p_{\mathbf{c}}(x) = c_0 + c_1 x + \cdots + c_{n-1}x^{n-1}$.

(b) With $\omega = e^{2\pi i/n}$: the eigenvalues are $1, \omega, \ldots, \omega^{n-1}$, all distinct, and ω^j has the one-dimensional eigenspace spanned by $\mathbf{f}_j = n^{-1/2}(1, \omega^j, \ldots, \omega^{(n-1)j})$ (Theorem 12.9.2).

(c) The eigenvalues are $p_{\mathbf{c}}(\omega^j)$ for $j = 0, \ldots, n-1$ (Theorem 12.9.4). Each evaluation is a sum of n products, so all n of them cost about n^2 multiplications done directly.

(d) (i) False: the 3×3 matrix in the warning above has constant diagonals but is not a circulant. (ii) True, by Theorem 12.9.1 (c); equivalently by Corollary 12.4.3, since $\mathbf{C} = \mathbf{F}\mathbf{D}\mathbf{F}^*$. (iii) False. $\mathbf{F}$ is symmetric, but $\mathbf{F}^* = \overline{\mathbf{F}}$, and for $n = 4$ the entry $(\mathbf{F})_{12} = i/2$ is not real (Example (The Fourier matrix of size four)). $\mathbf{F}$ is Hermitian only for $n = 1, 2$, where all its entries are real.

(e) Both are polynomials in the single matrix $\mathbf{S}$ (Theorem 12.9.1 (a)), and two polynomials in one matrix always commute (Theorem 0.9.9).

12.9.6.2 B. Practice

Exercise. B1: Eigenvalues of small circulants

For each vector $\mathbf{c}$, write down the symbol, compute all eigenvalues of the circulant with first row $\mathbf{c}$, and give its determinant.

(a) $\mathbf{c} = (1, 1, 1)$, $n = 3$.

(b) $\mathbf{c} = (0, 1, 0, 0)$, $n = 4$.

(c) $\mathbf{c} = (2, -1, 0, -1)$, $n = 4$.

Solution.

(a) $p_{\mathbf{c}}(x) = 1 + x + x^2$. With $\omega = e^{2\pi i/3}$, the eigenvalues are $p_{\mathbf{c}}(1) = 3$ and $p_{\mathbf{c}}(\omega^j) = 1 + \omega^j + \omega^{2j}$ for $j = 1, 2$. For $j \neq 0$, $z = \omega^j$ satisfies $z \neq 1$ and $z^3 = 1$, so $1 + z + z^2 = (z^3 - 1)/(z - 1) = 0$ — the geometric-series computation of Theorem 12.9.3 again. The eigenvalues are $3, 0, 0$ and $\det \mathbf{C} = 0$. (This matrix is the all-ones matrix, of rank 1.)

(b) $p_{\mathbf{c}}(x) = x$, so $\mathbf{C} = \mathbf{S}$. The eigenvalues are $\omega^j = i^j$, namely $1, i, -1, -i$, and

$$\det \mathbf{S} = 1 \cdot i \cdot (-1) \cdot (-i) = \big(i \cdot (-i)\big)\big(1 \cdot (-1)\big) = -1.$$

This agrees with the sign of the permutation: $\mathbf{S}$ permutes the coordinates in one 4-cycle, which is odd.

(c) $p_{\mathbf{c}}(x) = 2 - x - x^3$. Then $p_{\mathbf{c}}(1) = 0$; $p_{\mathbf{c}}(i) = 2 - i - i^3 = 2 - i + i = 2$; $p_{\mathbf{c}}(-1) = 2 + 1 + 1 = 4$; $p_{\mathbf{c}}(-i) = 2 + i - i = 2$. The eigenvalues are $0, 2, 4, 2$, and $\det \mathbf{C} = 0$. The kernel contains $\mathbf{f}_0$, that is the all-ones vector, as one sees directly: every row of $\mathbf{C}$ sums to 0.

Exercise. B2: Solve a circulant system by transforming

Let $\mathbf{C}$ be the 4×4 circulant with first row $(3, 1, 0, 1)$ and let $\mathbf{b} = (1, 0, 0, 0)$. Compute the eigenvalues of $\mathbf{C}$, and solve $\mathbf{C}\mathbf{x} = \mathbf{b}$ by transforming, dividing, and transforming back. Check

your answer against the first two rows of $\mathbf{C}$.

Solution. The symbol is $p_{\mathbf{c}}(x) = 3 + x + x^3$, and with $\omega = i$,

$$p_{\mathbf{c}}(1) = 5, \quad p_{\mathbf{c}}(i) = 3 + i - i = 3, \quad p_{\mathbf{c}}(-1) = 1, \quad p_{\mathbf{c}}(-i) = 3.$$

All four are nonzero, so $\mathbf{C}$ is invertible (Corollary 12.9.5 (d)), with $\det \mathbf{C} = 5 \cdot 3 \cdot 1 \cdot 3 = 45$.

Transform: $(\mathbf{F}^*\mathbf{b})_j = \frac{1}{2}\sum_m \overline{\omega^{jm}} b_m = \frac{1}{2}$ for every j, since $\mathbf{b} = \mathbf{e}_0$. Divide: $y_j = \frac{1}{2}\lambda_j^{-1}$, giving $\mathbf{y} = \frac{1}{2}(\frac{1}{5}, \frac{1}{3}, 1, \frac{1}{3})$. Transform back: the m-th entry of $\mathbf{x} = \mathbf{F}\mathbf{y}$ is

$$x_m = \frac{1}{2}\sum_{j=0}^{3} i^{mj} y_j = \frac{1}{4}\Big(\tfrac{1}{5} + \tfrac{i^m}{3} + (-1)^m + \tfrac{i^{3m}}{3}\Big).$$

For $m = 1$ and $m = 3$ the two middle terms cancel, since $i^m + i^{3m} = i^m + \overline{i^m}$ is 0 when $i^m = \pm i$; the bracket is $\frac{1}{5} - 1$, so $x_1 = x_3 = -\frac{1}{5}$. For $m = 0$ the bracket is $\frac{1}{5} + \frac{1}{3} + 1 + \frac{1}{3} = \frac{28}{15}$, so $x_0 = \frac{7}{15}$. For $m = 2$ it is $\frac{1}{5} - \frac{1}{3} + 1 - \frac{1}{3} = \frac{8}{15}$, so $x_2 = \frac{2}{15}$. Hence

$$\mathbf{x} = \Big(\tfrac{7}{15}, \ -\tfrac{1}{5}, \ \tfrac{2}{15}, \ -\tfrac{1}{5}\Big).$$

Check: row 0 of $\mathbf{C}$ is $(3,1,0,1)$ and $3 \cdot \frac{7}{15} - \frac{1}{5} - \frac{1}{5} = \frac{21-3-3}{15} = 1$; row 1 is $(1,3,1,0)$ and $\frac{7}{15} - \frac{3}{5} + \frac{2}{15} = \frac{7-9+2}{15} = 0$.

Exercise. B3: The convolution theorem in an example

Let $\mathbf{c} = (1,0,2,0)$ and $\mathbf{d} = (0,1,0,3)$ in $\mathbb{C}^4$.

(a) Compute $\mathbf{c} * \mathbf{d}$.

(b) Verify $p_{\mathbf{c}*\mathbf{d}}(\omega^j) = p_{\mathbf{c}}(\omega^j)p_{\mathbf{d}}(\omega^j)$ for $j = 0, 1, 2, 3$.

(c) Hence write down the first row of the product of the two circulants, without multiplying any matrices.

Solution.

(a) Only $c_0 = 1$ and $c_2 = 2$ are nonzero, so $(\mathbf{c} * \mathbf{d})_r = d_r + 2d_{r-2}$, subscripts modulo 4. With $\mathbf{d} = (0,1,0,3)$:

$$\mathbf{c} * \mathbf{d} = (0 + 0, \ 1 + 6, \ 0 + 0, \ 3 + 2) = (0, 7, 0, 5).$$

(b) The symbols are $p_{\mathbf{c}}(x) = 1 + 2x^2$, $p_{\mathbf{d}}(x) = x + 3x^3$, and $p_{\mathbf{c}*\mathbf{d}}(x) = 7x + 5x^3$. With $\omega = i$:

$$\begin{aligned} j = 0: &\quad 3 \cdot 4 = 12, && 7 + 5 = 12; \\ j = 1: &\quad (-1)(-2i) = 2i, && 7i - 5i = 2i; \\ j = 2: &\quad 3 \cdot (-4) = -12, && -7 - 5 = -12; \\ j = 3: &\quad (-1)(2i) = -2i, && -7i + 5i = -2i. \end{aligned}$$

Here $p_{\mathbf{c}}(i) = 1 + 2i^2 = -1$ and $p_{\mathbf{d}}(i) = i + 3i^3 = i - 3i = -2i$, and similarly at the other roots. All four rows agree.

(c) By Theorem 12.9.6 (a), the product is the circulant with first row $\mathbf{c} * \mathbf{d} = (0, 7, 0, 5)$.

12.9.6.3 C. Going deeper

Exercise. C1: What commutes with the shift

Let $\mathbf{A} \in M_n(\mathbb{C})$ satisfy $\mathbf{AS} = \mathbf{SA}$. Prove that $\mathbf{A}$ is a circulant. *Hint: consider the eigenspaces of* $\mathbf{S}$.

Solution. By Theorem 12.9.2, $\mathbf{S}$ has the n distinct eigenvalues ω^j, and $E_{\omega^j}(\mathbf{S}) = \operatorname{span}(\mathbf{f}_j)$ is one-dimensional. Since $\mathbf{A}$ commutes with $\mathbf{S}$, each eigenspace of $\mathbf{S}$ is invariant under $\mathbf{A}$ (Theorem 9.9.1). So $\mathbf{A}\mathbf{f}_j \in \operatorname{span}(\mathbf{f}_j)$, that is $\mathbf{A}\mathbf{f}_j = d_j \mathbf{f}_j$ for some $d_j \in \mathbb{C}$.

Therefore the j-th column of $\mathbf{AF}$ is $d_j \mathbf{f}_j$, which is the j-th column of $\mathbf{FD}$ for $\mathbf{D} = \operatorname{diag}(d_0, \dots, d_{n-1})$. Hence $\mathbf{AF} = \mathbf{FD}$ and $\mathbf{F}^*\mathbf{AF} = \mathbf{D}$ is diagonal, so $\mathbf{A}$ is a circulant by Corollary 12.9.5 (c). This proves the converse of Theorem 12.9.1 (c) for the single matrix $\mathbf{S}$: the commutant of the cyclic shift is exactly the algebra of circulants.

Exercise. C2: Determinant and invertibility

Let $\mathbf{C}$ be the circulant with first row $\mathbf{c} \in \mathbb{C}^n$ and symbol $p_{\mathbf{c}}$.

(a) Prove that $\det \mathbf{C} = \prod_{j=0}^{n-1} p_{\mathbf{c}}(\omega^j)$, and use it to compute the determinant of the $n \times n$ all-ones matrix for every $n \geq 2$.

(b) Prove that $\mathbf{C}$ is invertible if and only if $p_{\mathbf{c}}$ and $x^n - 1$ are coprime in $\mathbb{C}[x]$.

(c) Give an example of a 4×4 circulant with integer entries, none of them 0, whose determinant is 0.

Solution.

(a) This is the last claim of Theorem 12.9.4: from $\mathbf{C} = \mathbf{FDF}^*$ and $\det(\mathbf{F}^*)\det(\mathbf{F}) = \det(\mathbf{F}^*\mathbf{F}) = 1$, multiplicativity of the determinant (Theorem 7.5.4) gives $\det \mathbf{C} = \det \mathbf{D}$, which is the product of the diagonal entries $p_{\mathbf{c}}(\omega^j)$.
For the all-ones matrix, $\mathbf{c} = (1, 1, \dots, 1)$ and $p_{\mathbf{c}}(x) = 1 + x + \cdots + x^{n-1}$. Then $p_{\mathbf{c}}(1) = n$, while for $1 \leq j \leq n-1$ the number $z = \omega^j$ satisfies $z^n = 1$ and $z \neq 1$, so $p_{\mathbf{c}}(z) = (z^n - 1)/(z - 1) = 0$. Hence $\det \mathbf{C} = n \cdot 0 \cdots 0 = 0$ for every $n \geq 2$, as it must be, since all the rows are equal.

(b) By (a), $\det \mathbf{C} \neq 0$ if and only if $p_{\mathbf{c}}(\omega^j) \neq 0$ for every j, that is, if and only if $p_{\mathbf{c}}$ and $x^n - 1$ have no common root, the roots of $x^n - 1$ being exactly $1, \omega, \dots, \omega^{n-1}$ (Example (Roots of Unity)). Over $\mathbb{C}$, having no common root is the same as having no common nonconstant factor: such a factor would have a root by Theorem 6.5.1, and that root would be common to both; conversely a common root α makes $x - \alpha$ a common factor. So invertibility of $\mathbf{C}$ is equivalent to $p_{\mathbf{c}}$ and $x^n - 1$ being coprime (Definition 6.2.3). Combined with Theorem 7.5.2, this is the statement asked for.

(c) Take $\mathbf{c} = (1, 2, 1, 2)$, so

$$\mathbf{C} = \begin{pmatrix} 1 & 2 & 1 & 2 \\ 2 & 1 & 2 & 1 \\ 1 & 2 & 1 & 2 \\ 2 & 1 & 2 & 1 \end{pmatrix}, \qquad p_{\mathbf{c}}(x) = 1 + 2x + x^2 + 2x^3.$$

Here $p_{\mathbf{c}}(i) = 1 + 2i - 1 - 2i = 0$, so $\det \mathbf{C} = 0$ by (a). The reason is visible without any of the theory — rows 0 and 2 are equal — and the Fourier picture names a kernel vector: $\mathbf{C}\mathbf{f}_1 = \mathbf{0}$, and since $\mathbf{C}$ is real, the real and imaginary parts of $\mathbf{f}_1$ are in the kernel too, giving $\mathbf{C}(1, 0, -1, 0) = \mathbf{0}$.

Exercise. C3: The real orthogonal form of a real circulant

Let $\mathbf{C} \in M_4(\mathbb{R})$ be the circulant with first row $(1, 2, 3, 4)$, with eigenvalues $10, -2 - 2i, -2, -2 + 2i$ as in Example (The running example, diagonalized). Put

$$\mathbf{g} := \sqrt{2}\,\operatorname{Re}\mathbf{f}_1, \qquad \mathbf{h} := \sqrt{2}\,\operatorname{Im}\mathbf{f}_1,$$

the real and imaginary parts taken entrywise.

(a) Write $\mathbf{g}$ and $\mathbf{h}$ out, and show that $(\mathbf{f}_0, \mathbf{f}_2, \mathbf{g}, \mathbf{h})$ is an orthonormal basis of $\mathbb{R}^4$.

(b) Writing $a + bi := -2 + 2i$, prove that $\mathbf{C}\mathbf{g} = a\mathbf{g} + b\mathbf{h}$ and $\mathbf{C}\mathbf{h} = -b\mathbf{g} + a\mathbf{h}$.

(c) Hence write down the matrix of $\mathbf{C}$ in the ordered basis of (a), and identify it as an instance of Theorem 12.5.3.

Hint for (b): apply $\mathbf{C}$ to $\mathbf{f}_3 = (\mathbf{g} - i\mathbf{h})/\sqrt{2}$ and compare real and imaginary parts.

Solution.

(a) With $\omega = i$, $\mathbf{f}_1 = \frac{1}{2}(1, i, -1, -i)$, so

$$\mathbf{g} = \tfrac{1}{\sqrt{2}}(1, 0, -1, 0), \qquad \mathbf{h} = \tfrac{1}{\sqrt{2}}(0, 1, 0, -1).$$

Also $\mathbf{f}_0 = \frac{1}{2}(1, 1, 1, 1)$ and $\mathbf{f}_2 = \frac{1}{2}(1, -1, 1, -1)$ are real, so all four vectors lie in $\mathbb{R}^4$. Each has norm 1: for instance $\|\mathbf{g}\|^2 = \frac{1}{2}(1 + 1) = 1$ and $\|\mathbf{f}_0\|^2 = \frac{1}{4} \cdot 4 = 1$. They are pairwise orthogonal for the real dot product. First $\mathbf{g} \cdot \mathbf{h} = 0$, because the nonzero entries sit in different slots, and $\mathbf{f}_0 \cdot \mathbf{f}_2 = \frac{1}{4}(1 - 1 + 1 - 1) = 0$. The four remaining pairings are

$$\mathbf{f}_0 \cdot \mathbf{g} = \tfrac{1}{2\sqrt{2}}(1 - 1) = 0, \quad \mathbf{f}_2 \cdot \mathbf{g} = \tfrac{1}{2\sqrt{2}}(1 - 1) = 0,$$
$$\mathbf{f}_0 \cdot \mathbf{h} = \tfrac{1}{2\sqrt{2}}(1 - 1) = 0, \quad \mathbf{f}_2 \cdot \mathbf{h} = \tfrac{1}{2\sqrt{2}}(-1 + 1) = 0.$$

Four orthonormal vectors are independent (Theorem 11.1.4) and there are $\dim \mathbb{R}^4 = 4$ of them, so they form a basis (Theorem 1.6.4).

(b) From $\mathbf{f}_1 = \frac{1}{\sqrt{2}}(\mathbf{g} + i\mathbf{h})$ and $\mathbf{f}_3 = \overline{\mathbf{f}_1}$ we get $\mathbf{f}_3 = \frac{1}{\sqrt{2}}(\mathbf{g} - i\mathbf{h})$. Its eigenvalue is $p_{\mathbf{c}}(\omega^3) = p_{\mathbf{c}}(-i) = -2 + 2i = a + bi$, so $a = -2$ and $b = 2$, and $\mathbf{C}\mathbf{f}_3 = (a + bi)\mathbf{f}_3$ reads

$$\tfrac{1}{\sqrt{2}}(\mathbf{C}\mathbf{g} - i\,\mathbf{C}\mathbf{h}) = \tfrac{1}{\sqrt{2}}(a + bi)(\mathbf{g} - i\mathbf{h}) = \tfrac{1}{\sqrt{2}}((a\mathbf{g} + b\mathbf{h}) + i(b\mathbf{g} - a\mathbf{h})).$$

Both $\mathbf{Cg}$ and $\mathbf{Ch}$ are real vectors, because $\mathbf{C}$, $\mathbf{g}$, and $\mathbf{h}$ are real. Comparing real parts gives $\mathbf{Cg} = a\mathbf{g} + b\mathbf{h}$, and comparing imaginary parts gives $-\mathbf{Ch} = b\mathbf{g} - a\mathbf{h}$, that is $\mathbf{Ch} = -b\mathbf{g} + a\mathbf{h}$. With $a = -2$ and $b = 2$,

$$\mathbf{Cg} = -2\mathbf{g} + 2\mathbf{h}, \qquad \mathbf{Ch} = -2\mathbf{g} - 2\mathbf{h}.$$

A direct check: $\sqrt{2}\,\mathbf{Cg} = \mathbf{C}(1, 0, -1, 0) = (1-3,\ 4-2,\ 3-1,\ 2-4) = (-2, 2, 2, -2)$, while $\sqrt{2}\,(-2\mathbf{g} + 2\mathbf{h}) = (-2, 2, 2, -2)$.

(c) Reading the coordinates of the images off (b), the matrix in the ordered basis $(\mathbf{f}_0, \mathbf{f}_2, \mathbf{g}, \mathbf{h})$ is

$$\begin{pmatrix} 10 & 0 & 0 & 0 \\ 0 & -2 & 0 & 0 \\ 0 & 0 & -2 & -2 \\ 0 & 0 & 2 & -2 \end{pmatrix} = (10) \oplus (-2) \oplus \mathbf{\Lambda}(-2 + 2i),$$

since $\mathbf{\Lambda}(a + bi)$ has rows $(a, -b)$ and (b, a). This is exactly the shape promised by Theorem 12.5.3 for a real normal operator: two 1×1 blocks for the real eigenvalues, and one rotation-scaling block for the conjugate pair, with $b = 2 > 0$. The change of basis is orthogonal, because the basis is orthonormal.

12.10 The Sylvester Equation

Chapter 8 wrote an operator with an invariant subspace as a block triangular matrix, with a corner block $\mathbf{C}$ recording how the subspace fails to have an invariant complement (Theorem 8.1.7). Chapter 10 spent its length on what to do when that corner cannot be cleared. The question of *when* it can be cleared has an exact answer, and the answer is a linear equation in the unknown matrix $\mathbf{X}$:

$$\mathbf{AX} - \mathbf{XB} = \mathbf{C}.$$

Chapter 8 already turned this equation into an ordinary linear system with the Kronecker product, and stopped one step short of the criterion because that step needs eigenvalues. We have eigenvalues now, and Schur's factorization of Section 1 to organize them. This section proves the criterion, uses it to clear corners, and reads off what it says about the matrices that commute with a given one.

Throughout, the field is $\mathbb{C}$ unless stated otherwise, and $\mathbf{A} \in M_m(\mathbb{C})$, $\mathbf{B} \in M_n(\mathbb{C})$, while the unknown $\mathbf{X}$ and the data $\mathbf{C}$ live in $M_{m\times n}(\mathbb{C})$.

12.10.1 The Sylvester operator

The left-hand side of the equation is linear in $\mathbf{X}$: replacing $\mathbf{X}$ by $\mathbf{X} + \mathbf{X}'$ or by $c\mathbf{X}$ does the same to $\mathbf{AX} - \mathbf{XB}$, because matrix multiplication is bilinear (Theorem 0.9.4). So the whole question is about one linear map, and it deserves a name.

Solving $\mathbf{AX} - \mathbf{XB} = \mathbf{C}$ *means inverting a single linear operator on the space of* $m \times n$ *matrices.*

Definition 12.10.1. Sylvester operator

Let $\mathbf{A} \in M_m(F)$ and $\mathbf{B} \in M_n(F)$. The **Sylvester operator** of the pair $(\mathbf{A}, \mathbf{B})$ is

$$\mathcal{S}_{\mathbf{A},\mathbf{B}} \colon M_{m\times n}(F) \to M_{m\times n}(F), \qquad \mathcal{S}_{\mathbf{A},\mathbf{B}}(\mathbf{X}) = \mathbf{AX} - \mathbf{XB}.$$

The equation $\mathbf{AX} - \mathbf{XB} = \mathbf{C}$, for a given $\mathbf{C} \in M_{m\times n}(F)$, is the **Sylvester equation**.

In words: $\mathcal{S}_{\mathbf{A},\mathbf{B}}$ multiplies by $\mathbf{A}$ on the left, multiplies by $\mathbf{B}$ on the right, and subtracts. It is an operator on a space of dimension mn, not on F^m or F^n, and that is the only thing that makes the equation look harder than $\mathbf{Ax} = \mathbf{b}$.

Three special cases are worth keeping in view.

1. $\mathbf{B} = \mathbf{O}$**.** Then $\mathcal{S}_{\mathbf{A},\mathbf{O}}(\mathbf{X}) = \mathbf{AX}$, and the equation is n copies of $\mathbf{Ax} = \mathbf{b}$, one for each column of $\mathbf{X}$ (Exercise (B2: One-sided equations)). It is uniquely solvable for every $\mathbf{C}$ exactly when $\mathbf{A}$ is invertible, that is, when $0 \notin \operatorname{spec}(\mathbf{A})$. Since $\operatorname{spec}(\mathbf{O}) = \{0\}$, this already matches the criterion we are heading for.
2. $\mathbf{B} = \mathbf{A}$ **and** $\mathbf{C} = \mathbf{O}$**.** Then $\ker \mathcal{S}_{\mathbf{A},\mathbf{A}}$ is the **commutant** of $\mathbf{A}$, the set of matrices commuting with $\mathbf{A}$. This is the question Chapter 8 opened with, and Chapter 10 §04 studied it for a single Jordan or Weyr block.
3. $m = n = 1$**.** Then $\mathcal{S}_{a,b}(x) = (a-b)x$, which is invertible exactly when $a \neq b$. The general criterion will be this one, run over all pairs of eigenvalues.

Nonexample by minimal change. The map $\mathbf{X} \mapsto \mathbf{AX} - \mathbf{XB} - \mathbf{X}^\top$ is still linear on $M_n(F)$, but it is not a Sylvester operator: the transpose is not of the form "multiply on one side". Everything below uses the two-sided product shape and nothing else.

Chapter 8 already wrote the matrix of $\mathcal{S}_{\mathbf{A},\mathbf{B}}$. Stacking the columns of $\mathbf{X}$ with the vectorization vec of Definition 8.4.2 and applying the vec identity Theorem 8.4.4 twice, once with $\mathbf{B} = \mathbf{I}_n$ and once with $\mathbf{A} = \mathbf{I}_m$,

$$\begin{aligned} \operatorname{vec}\left(\mathcal{S}_{\mathbf{A},\mathbf{B}}(\mathbf{X})\right) &= \mathbf{K}\,\operatorname{vec}\mathbf{X}, \\ \mathbf{K} &\coloneqq \mathbf{I}_n \otimes \mathbf{A} - \mathbf{B}^\top \otimes \mathbf{I}_m. \end{aligned} \tag{12.10.1}$$

Here $\mathbf{K} \in M_{mn}(F)$, and $\operatorname{vec} \colon M_{m\times n}(F) \to F^{mn}$ is an isomorphism. So Eq. 12.10.1 says that $\mathbf{K}$ *is* the matrix of $\mathcal{S}_{\mathbf{A},\mathbf{B}}$ with respect to the basis of matrix units ordered column by column. In particular $\mathcal{S}_{\mathbf{A},\mathbf{B}}$ is invertible if and only if $\mathbf{K}$ is.

That is exactly where Chapter 8 left the problem. What it could not do was compute the eigenvalues of $\mathbf{K}$.

12.10.2 The spectrum of a Kronecker sum

Chapter 8 proved how the Kronecker product treats rank, trace, and determinant (Theorem 8.4.3) and how it treats products, transposes, and inverses (Theorem 8.4.1). It proved nothing about eigenvalues, and Exercise (C2: Kronecker sums) got only as far as the case where both factors are already triangular. Schur closes the gap, because over $\mathbb{C}$ every matrix can be made triangular unitarily, and the mixed product rule carries a pair of unitary conjugations through the Kronecker product in one step.

Lemma 12.10.1. The Eigenvalues of a Kronecker Sum

Let $\mathbf{A} \in M_m(\mathbb{C})$ with eigenvalues $\lambda_1, \dots, \lambda_m$ and let $\mathbf{B} \in M_n(\mathbb{C})$ with eigenvalues $\mu_1, \dots, \mu_n$,

each list repeated according to algebraic multiplicity. Then the mn eigenvalues of

$$\mathbf{K} = \mathbf{I}_n \otimes \mathbf{A} - \mathbf{B}^\top \otimes \mathbf{I}_m \in M_{mn}(\mathbb{C}),$$

listed with algebraic multiplicity, are the mn differences $\lambda_i - \mu_j$ $(1 \le i \le m,\ 1 \le j \le n)$.

Idea. Triangularize both ingredients at once. Schur (Corollary 12.1.3) writes $\mathbf{A} = \mathbf{U}\mathbf{T}\mathbf{U}^*$ and $\mathbf{B}^\top = \mathbf{V}\mathbf{S}\mathbf{V}^*$ with $\mathbf{T}, \mathbf{S}$ upper triangular. The mixed product rule then conjugates *both* summands of $\mathbf{K}$ by the single unitary matrix $\mathbf{V} \otimes \mathbf{U}$, because $\mathbf{I}_n \otimes \mathbf{A}$ and $\mathbf{B}^\top \otimes \mathbf{I}_m$ each have one trivial factor to absorb the other conjugation. What is left is $\mathbf{I}_n \otimes \mathbf{T} - \mathbf{S} \otimes \mathbf{I}_m$, and a Kronecker product with one triangular and one identity factor is triangular, so the eigenvalues are visible on the diagonal.

Proof. By Corollary 12.1.3 there are a unitary $\mathbf{U} \in M_m(\mathbb{C})$ and an upper triangular $\mathbf{T} \in M_m(\mathbb{C})$ with $\mathbf{A} = \mathbf{U}\mathbf{T}\mathbf{U}^*$ and $t_{ii} = \lambda_i$. By Proposition 9.10.5 (b), $p_{\mathbf{B}^\top} = p_{\mathbf{B}}$, so $\mathbf{B}^\top$ has eigenvalue list $\mu_1, \dots, \mu_n$; applying Corollary 12.1.3 to $\mathbf{B}^\top$ gives a unitary $\mathbf{V} \in M_n(\mathbb{C})$ and an upper triangular $\mathbf{S} \in M_n(\mathbb{C})$ with $\mathbf{B}^\top = \mathbf{V}\mathbf{S}\mathbf{V}^*$ and $s_{jj} = \mu_j$.

Put $\mathbf{W} = \mathbf{V} \otimes \mathbf{U} \in M_{mn}(\mathbb{C})$. Conjugating every entry of $\mathbf{A} \otimes \mathbf{B}$ conjugates every block $a_{ij}\mathbf{B}$, so $\overline{\mathbf{A} \otimes \mathbf{B}} = \overline{\mathbf{A}} \otimes \overline{\mathbf{B}}$ (Definition 8.4.1), and combining this with Theorem 8.4.1 (d) gives $(\mathbf{A} \otimes \mathbf{B})^* = \mathbf{A}^* \otimes \mathbf{B}^*$ whenever the product is defined. Hence $\mathbf{W}^* = \mathbf{V}^* \otimes \mathbf{U}^*$, and by the mixed product rule Theorem 8.4.1 (c),

$$\mathbf{W}^*\mathbf{W} = (\mathbf{V}^*\mathbf{V}) \otimes (\mathbf{U}^*\mathbf{U}) = \mathbf{I}_n \otimes \mathbf{I}_m = \mathbf{I}_{mn},$$

so $\mathbf{W}$ is unitary. Using the mixed product rule twice more,

$$\mathbf{W}(\mathbf{I}_n \otimes \mathbf{T})\mathbf{W}^* = (\mathbf{V}\mathbf{I}_n\mathbf{V}^*) \otimes (\mathbf{U}\mathbf{T}\mathbf{U}^*) = \mathbf{I}_n \otimes \mathbf{A},$$
$$\mathbf{W}(\mathbf{S} \otimes \mathbf{I}_m)\mathbf{W}^* = (\mathbf{V}\mathbf{S}\mathbf{V}^*) \otimes (\mathbf{U}\mathbf{I}_m\mathbf{U}^*) = \mathbf{B}^\top \otimes \mathbf{I}_m.$$

Subtracting, $\mathbf{W}(\mathbf{I}_n \otimes \mathbf{T} - \mathbf{S} \otimes \mathbf{I}_m)\mathbf{W}^* = \mathbf{K}$, so $\mathbf{K}$ is similar to

$$\mathbf{R} \coloneqq \mathbf{I}_n \otimes \mathbf{T} - \mathbf{S} \otimes \mathbf{I}_m.$$

$\mathbf{R}$ is upper triangular. Label the rows and columns of an $mn \times mn$ matrix by pairs as in Definition 8.4.1: row $(j-1)m + i$ with $1 \le j \le n$, $1 \le i \le m$, and likewise column $(l-1)m + k$. In those labels the entries of the two summands are

$$(\mathbf{I}_n \otimes \mathbf{T})_{(j-1)m+i,\,(l-1)m+k} = \delta_{jl}\, t_{ik},$$
$$(\mathbf{S} \otimes \mathbf{I}_m)_{(j-1)m+i,\,(l-1)m+k} = s_{jl}\, \delta_{ik}.$$

The first is nonzero only when $j = l$ and $i \le k$, since $\mathbf{T}$ is upper triangular; then $(j-1)m+i \le (l-1)m + k$. The second is nonzero only when $j \le l$ and $i = k$; if $j = l$ the two labels are equal, and if $j < l$ then

$$(j-1)m + i \le (j-1)m + m = jm \le (l-1)m < (l-1)m + k,$$

using $i \le m$, $j \le l - 1$, and $k \ge 1$. So each summand vanishes below the diagonal, and so does $\mathbf{R}$.

The diagonal. Taking $j = l$ and $i = k$ above, the diagonal entry of $\mathbf{R}$ in position $(j-1)m+i$ is $t_{ii} - s_{jj} = \lambda_i - \mu_j$. As (i, j) runs over all mn pairs, so does the position $(j-1)m+i$, by division with remainder.

By Theorem 9.7.4 (a), applied to the operator $T_{\mathbf{R}}$ on $\mathbb{C}^{mn}$ and the standard basis, $p_{\mathbf{R}} = \prod_{i,j}(x - (\lambda_i - \mu_j))$. Since $\mathbf{K} \sim \mathbf{R}$, Theorem 7.10.3 gives $p_{\mathbf{K}} = p_{\mathbf{R}}$. This proves the lemma.

■

Two remarks on what the proof used. The hypothesis $F = \mathbb{C}$ entered once, in Schur, and through Schur it is the Fundamental Theorem of Algebra again. And the unitarity of $\mathbf{W}$ was a convenience, not a necessity: any similarity triangularizing $\mathbf{A}$ and $\mathbf{B}^{\top}$ would do, but Schur hands us one for free and the Kronecker product of two unitary matrices is unitary, which is worth recording on its own.

12.10.3 When the equation is uniquely solvable

The criterion now costs one line.

Theorem 12.10.2. Sylvester's Theorem

Let $\mathbf{A} \in M_m(\mathbb{C})$ and $\mathbf{B} \in M_n(\mathbb{C})$. The following are equivalent.

(a) $\operatorname{spec}(\mathbf{A}) \cap \operatorname{spec}(\mathbf{B}) = \varnothing$: the two matrices have **no common eigenvalue**.
(b) The Sylvester operator $\mathcal{S}_{\mathbf{A},\mathbf{B}} \colon \mathbf{X} \mapsto \mathbf{AX} - \mathbf{XB}$ is invertible on $M_{m\times n}(\mathbb{C})$.
(c) For every $\mathbf{C} \in M_{m\times n}(\mathbb{C})$, the equation $\mathbf{AX} - \mathbf{XB} = \mathbf{C}$ has **exactly one** solution $\mathbf{X}$.
(d) The only $\mathbf{X} \in M_{m\times n}(\mathbb{C})$ with $\mathbf{AX} = \mathbf{XB}$ is $\mathbf{X} = \mathbf{O}$.

Proof. By Eq. 12.10.1 the isomorphism vec turns $\mathcal{S}_{\mathbf{A},\mathbf{B}}$ into the matrix $\mathbf{K} = \mathbf{I}_n \otimes \mathbf{A} - \mathbf{B}^{\top} \otimes \mathbf{I}_m$: a matrix $\mathbf{X}$ solves $\mathbf{AX} - \mathbf{XB} = \mathbf{C}$ if and only if the vector $\operatorname{vec}\mathbf{X}$ solves $\mathbf{Ky} = \operatorname{vec}\mathbf{C}$, and every $\mathbf{y} \in \mathbb{C}^{mn}$ and every $\mathbf{z} \in \mathbb{C}^{mn}$ arise this way. Transported through vec, statements (b), (c), and (d) say respectively that $\mathbf{K}$ is invertible, that $\mathbf{Ky} = \mathbf{z}$ has exactly one solution for every $\mathbf{z}$, and that $\mathbf{Ky} = \mathbf{0}$ forces $\mathbf{y} = \mathbf{0}$. These three are equivalent by Theorem 3.4.4, parts (a), (e), and (b): (e) gives existence for every $\mathbf{z}$, and (b) upgrades it to uniqueness, since two solutions differ by a member of the null space.

It remains to match (a) with (b). By Theorem 9.2.6, $\mathbf{K}$ is invertible if and only if $0 \notin \operatorname{spec}(\mathbf{K})$. By Lemma 12.10.1, $\operatorname{spec}(\mathbf{K}) = \{\lambda_i - \mu_j\}$, where $\lambda_1, \dots, \lambda_m$ and $\mu_1, \dots, \mu_n$ list the eigenvalues of $\mathbf{A}$ and of $\mathbf{B}$. So $0 \in \operatorname{spec}(\mathbf{K})$ if and only if $\lambda_i = \mu_j$ for some pair (i, j), that is, if and only if $\mathbf{A}$ and $\mathbf{B}$ have a common eigenvalue. Negating both sides matches (a) with (b), and the proof is complete.

■

This is the statement Chapter 8 promised. At the end of Example (A Sylvester Equation as a Linear System) it was said that over $\mathbb{C}$ the equation $\mathbf{AX} - \mathbf{XB} = \mathbf{C}$ is uniquely solvable for every $\mathbf{C}$ exactly when $\mathbf{A}$ and $\mathbf{B}$ have no common eigenvalue, and that the criterion needs eigenvalues and would be proved in Chapter 12. Parts (a) and (c) of Theorem 12.10.2 are that sentence, and the eigenvalues it needed were supplied by Lemma 12.10.1. The quantitative version — *how badly* conditioned the equation is when the spectra are close but disjoint — is a separate

matter, and belongs with perturbation theory in Chapter 20.

Part (d) is the version used most often in practice, because it is a statement with no data in it: to show that some matrix is zero, exhibit it as a solution of $\mathbf{AX} = \mathbf{XB}$ for a pair with disjoint spectra. The corollary on commutants below is that move and nothing else.

Example. Solving a Sylvester Equation

Over $\mathbb{C}$, solve $\mathbf{AX} - \mathbf{XB} = \mathbf{C}$ for

$$\mathbf{A} = \begin{pmatrix} 2 & 1 \\ 0 & 3 \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} 1 & 0 \\ 2 & -1 \end{pmatrix}, \quad \mathbf{C} = \begin{pmatrix} 4 & -2 \\ 0 & 4 \end{pmatrix}.$$

Solution. The matrix $\mathbf{A}$ is upper triangular, so $\operatorname{spec}(\mathbf{A}) = \{2, 3\}$ (Theorem 9.7.4); $\mathbf{B}$ is lower triangular, so $\operatorname{spec}(\mathbf{B}) = \{1, -1\}$. The spectra are disjoint, and by Theorem 12.10.2 there is exactly one solution.

To find it, use Eq. 12.10.1. Here $\mathbf{B}^{\top} = \begin{pmatrix} 1 & 2 \\ 0 & -1 \end{pmatrix}$, so by Definition 8.4.1

$$\mathbf{I}_2 \otimes \mathbf{A} = \begin{pmatrix} 2 & 1 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 2 & 1 \\ 0 & 0 & 0 & 3 \end{pmatrix}, \qquad \mathbf{B}^{\top} \otimes \mathbf{I}_2 = \begin{pmatrix} 1 & 0 & 2 & 0 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix},$$

and their difference is

$$\mathbf{K} = \begin{pmatrix} 1 & 1 & -2 & 0 \\ 0 & 2 & 0 & -2 \\ 0 & 0 & 3 & 1 \\ 0 & 0 & 0 & 4 \end{pmatrix}.$$

Its diagonal entries are $1, 2, 3, 4$, which are the four differences $2-1$, $3-1$, $2-(-1)$, $3-(-1)$, as Lemma 12.10.1 predicts; none is zero, confirming invertibility.

With $\operatorname{vec}\mathbf{X} = (x_{11}, x_{21}, x_{12}, x_{22})$ and $\operatorname{vec}\mathbf{C} = (4, 0, -2, 4)$, back substitution in $\mathbf{K}\operatorname{vec}\mathbf{X} = \operatorname{vec}\mathbf{C}$ gives $4x_{22} = 4$, so $x_{22} = 1$; then $3x_{12} + 1 = -2$, so $x_{12} = -1$; then $2x_{21} - 2 = 0$, so $x_{21} = 1$; then $x_{11} + 1 + 2 = 4$, so $x_{11} = 1$. Hence

$$\mathbf{X} = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}.$$

Check: $\mathbf{AX} = \begin{pmatrix} 3 & -1 \\ 3 & 3 \end{pmatrix}$ and $\mathbf{XB} = \begin{pmatrix} -1 & 1 \\ 3 & -1 \end{pmatrix}$, and their difference is $\mathbf{C}$.

Quick check. Let $\mathbf{A} \in M_3(\mathbb{C})$ have eigenvalues $1, 2, 2$ and let $\mathbf{B} \in M_2(\mathbb{C})$ have eigenvalues $0, 3$. List the eigenvalues of $\mathcal{S}_{\mathbf{A},\mathbf{B}}$ with multiplicity, and decide whether $\mathbf{AX} - \mathbf{XB} = \mathbf{C}$ is uniquely solvable for every $\mathbf{C} \in M_{3\times 2}(\mathbb{C})$.

Solution. By Lemma 12.10.1 the six eigenvalues are the differences $\lambda_i - \mu_j$:

$$1-0=1, \quad 2-0=2, \quad 2-0=2, \quad 1-3=-2, \quad 2-3=-1, \quad 2-3=-1.$$

None is zero — equivalently $\operatorname{spec}(\mathbf{A}) = \{1,2\}$ and $\operatorname{spec}(\mathbf{B}) = \{0,3\}$ are disjoint — so by Theorem 12.10.2 the equation has exactly one solution for every $\mathbf{C}$.

Warning. **The criterion depends on the sign.** For the equation $\mathbf{AX}+\mathbf{XB}=\mathbf{C}$, which is the form Chapter 8 used in Example (A Sylvester Equation as a Linear System), the relevant operator is $\mathcal{S}_{\mathbf{A},-\mathbf{B}}$, and $\operatorname{spec}(-\mathbf{B}) = \{-\mu_j\}$. So the condition is $\lambda_i + \mu_j \neq 0$ for all i, j, **not** $\operatorname{spec}(\mathbf{A}) \cap \operatorname{spec}(\mathbf{B}) = \varnothing$. The two conditions are genuinely different: with $\mathbf{A} = \operatorname{diag}(1,2)$ and $\mathbf{B} = \begin{pmatrix} -1 & 5 \\ 0 & 3 \end{pmatrix}$ the spectra $\{1,2\}$ and $\{-1,3\}$ are disjoint, yet $1+(-1)=0$, so $\mathbf{AX}+\mathbf{XB}=\mathbf{C}$ is not solvable for every $\mathbf{C}$. Read the sign off the equation before quoting the criterion.

12.10.4 Clearing the corner

We can now answer the question the section opened with.

Corollary 12.10.3. Block Diagonalization by a Sylvester Solution

Let $\mathbf{A} \in M_m(F)$, $\mathbf{B} \in M_n(F)$, and $\mathbf{C} \in M_{m\times n}(F)$, and put

$$\mathbf{M} = \begin{pmatrix} \mathbf{A} & \mathbf{C} \\ \mathbf{O} & \mathbf{B} \end{pmatrix} \in M_{m+n}(F).$$

(a) If $\mathbf{X} \in M_{m\times n}(F)$ satisfies $\mathbf{AX} - \mathbf{XB} = -\mathbf{C}$, then with $\mathbf{P} = \begin{pmatrix} \mathbf{I}_m & \mathbf{X} \\ \mathbf{O} & \mathbf{I}_n \end{pmatrix}$ we have $\mathbf{P}^{-1}\mathbf{M}\mathbf{P} = \mathbf{A} \oplus \mathbf{B}$. In particular $\mathbf{M} \sim \mathbf{A} \oplus \mathbf{B}$.

(b) Conversely, if $\mathbf{P}^{-1}\mathbf{M}\mathbf{P} = \mathbf{A} \oplus \mathbf{B}$ for some $\mathbf{X} \in M_{m\times n}(F)$ and the matrix $\mathbf{P}$ built from it as in (a), then that $\mathbf{X}$ satisfies $\mathbf{AX} - \mathbf{XB} = -\mathbf{C}$.

(c) If $F = \mathbb{C}$ and $\operatorname{spec}(\mathbf{A}) \cap \operatorname{spec}(\mathbf{B}) = \varnothing$, then a solution $\mathbf{X}$ exists, and it is unique; so $\mathbf{M} \sim \mathbf{A} \oplus \mathbf{B}$ for **every** corner $\mathbf{C}$.

Idea. The matrices $\begin{pmatrix} \mathbf{I}_m & \mathbf{X} \\ \mathbf{O} & \mathbf{I}_n \end{pmatrix}$ are exactly the changes of basis that add multiples of the first block of basis vectors to the second and leave the first alone. Conjugating $\mathbf{M}$ by one of them cannot touch the two diagonal blocks or create anything below the diagonal; all it can do is move the corner, and the block multiplication shows it moves the corner by $\mathbf{AX} - \mathbf{XB}$. So the corner can be driven to $\mathbf{O}$ precisely when $-\mathbf{C}$ lies in the image of the Sylvester operator.

Proof.

(a) First, $\mathbf{P}$ is invertible with

$$\mathbf{P}^{-1} = \begin{pmatrix} \mathbf{I}_m & -\mathbf{X} \\ \mathbf{O} & \mathbf{I}_n \end{pmatrix},$$

since multiplying the two matrices in either order gives $\mathbf{I}_{m+n}$ by Theorem 8.1.1, the partitions being conformable. The same theorem gives

$$\mathbf{MP} = \begin{pmatrix} \mathbf{A} & \mathbf{AX}+\mathbf{C} \\ \mathbf{O} & \mathbf{B} \end{pmatrix},$$

and then

$$\mathbf{P}^{-1}\mathbf{MP} = \begin{pmatrix} \mathbf{A} & \mathbf{AX}+\mathbf{C}-\mathbf{XB} \\ \mathbf{O} & \mathbf{B} \end{pmatrix}.$$

By hypothesis $\mathbf{AX}-\mathbf{XB} = -\mathbf{C}$, so the corner is $\mathbf{O}$ and the result is $\mathbf{A} \oplus \mathbf{B}$ (Definition 8.1.2).

(b) The computation in (a) used nothing about $\mathbf{X}$ until the last sentence. So if that $\mathbf{P}$ satisfies $\mathbf{P}^{-1}\mathbf{MP} = \mathbf{A}\oplus\mathbf{B}$, comparing the corners of the two sides gives $\mathbf{AX}+\mathbf{C}-\mathbf{XB} = \mathbf{O}$, which is the equation.

(c) By Theorem 12.10.2 ((a) $\Rightarrow$ (c)), the equation $\mathbf{AX}-\mathbf{XB} = -\mathbf{C}$ has exactly one solution. Apply (a).

■

Example. Clearing a Corner

Block-diagonalize

$$\mathbf{M} = \begin{pmatrix} 2 & 1 & -4 & 2 \\ 0 & 3 & 0 & -4 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 2 & -1 \end{pmatrix} \in M_4(\mathbb{Q}).$$

Solution. Partition $\mathbf{M}$ into 2×2 blocks. The diagonal blocks are the matrices $\mathbf{A}$ and $\mathbf{B}$ of Example (Solving a Sylvester Equation), and the corner is

$$\begin{pmatrix} -4 & 2 \\ 0 & -4 \end{pmatrix} = -\mathbf{C}$$

for the $\mathbf{C}$ of that example. So the equation to solve is $\mathbf{AX}-\mathbf{XB} = -(-\mathbf{C}) = \mathbf{C}$, which is the one already solved there, with the unique solution $\mathbf{X} = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}$. By Corollary 12.10.3 (a), taking

$$\mathbf{P} = \begin{pmatrix} 1 & 0 & 1 & -1 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

gives $\mathbf{P}^{-1}\mathbf{M}\mathbf{P} = \mathbf{A} \oplus \mathbf{B}$, that is,

$$\mathbf{P}^{-1}\mathbf{M}\mathbf{P} = \begin{pmatrix} 2 & 1 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 2 & -1 \end{pmatrix}.$$

The eigenvalues $2, 3, 1, -1$ are of course unchanged; what has changed is that the space has been split into two invariant pieces of dimension 2.

Remark. Part (b) restricts the shape of $\mathbf{P}$, and one may ask whether that is a real restriction: if $\mathbf{M}$ is similar to $\mathbf{A} \oplus \mathbf{B}$ by *any* invertible matrix, must the Sylvester equation be solvable? The answer is yes over every field. This is **Roth's removal rule**, and the proof is not a block computation: it goes through the characteristic matrix and the criterion $\mathbf{A} \sim \mathbf{B}$ if and only if $x\mathbf{I} - \mathbf{A} \approx x\mathbf{I} - \mathbf{B}$ over $F[x]$ (Theorem 10.7.5). We do not need it here and do not prove it; nothing later in the book depends on it.

Primary decomposition, seen again. Put Corollary 12.10.3 next to Schur and a familiar theorem drops out with no extra work. Let $\mathbf{A} \in M_N(\mathbb{C})$ have distinct eigenvalues $\lambda_1, \dots, \lambda_k$ with algebraic multiplicities $a_1, \dots, a_k$. By Corollary 12.1.3 we may prescribe the order of the diagonal in a Schur factorization, so choose the order that lists all copies of λ_1 first, then all copies of λ_2, and so on. Partitioning the resulting triangular matrix along $N = a_1 + \cdots + a_k$ makes it block upper triangular with diagonal blocks $\mathbf{T}_1, \dots, \mathbf{T}_k$, where $\mathbf{T}_i$ is triangular with the single diagonal value λ_i, so $\operatorname{spec}(\mathbf{T}_i) = \{\lambda_i\}$. Now peel the blocks off one at a time: $\operatorname{spec}(\mathbf{T}_1)$ is disjoint from the spectrum of the rest, so Corollary 12.10.3 (c) splits off $\mathbf{T}_1$, and induction on k finishes. The conclusion is that $\mathbf{A} \sim \mathbf{T}_1 \oplus \cdots \oplus \mathbf{T}_k$ with $\operatorname{spec}(\mathbf{T}_i) = \{\lambda_i\}$ — the matrix form of the primary decomposition of Theorem 9.8.1, here obtained from a triangular form and a linear equation rather than from the kernel-splitting lemma.

12.10.5 Commuting with a matrix of disjoint spectra

Here is the promised use of Theorem 12.10.2 (d). It says that disjoint spectra force a block structure on anything that commutes.

Corollary 12.10.4. Disjoint Spectra Split the Commutant

Let $\mathbf{A}_1 \in M_{n_1}(\mathbb{C}), \dots, \mathbf{A}_r \in M_{n_r}(\mathbb{C})$ have **pairwise disjoint** spectra, and put $\mathbf{D} = \mathbf{A}_1 \oplus \cdots \oplus \mathbf{A}_r \in M_N(\mathbb{C})$, $N = n_1 + \cdots + n_r$. Let $\mathbf{Z} \in M_N(\mathbb{C})$ satisfy $\mathbf{D}\mathbf{Z} = \mathbf{Z}\mathbf{D}$, and partition $\mathbf{Z}$ into blocks $\mathbf{Z}_{pq} \in M_{n_p \times n_q}(\mathbb{C})$ for the partition $N = n_1 + \cdots + n_r$. Then

$$\mathbf{Z}_{pq} = \mathbf{O} \quad \text{for all } p \neq q,$$

so $\mathbf{Z} = \mathbf{Z}_{11} \oplus \cdots \oplus \mathbf{Z}_{rr}$ is block **diagonal**, and each diagonal block satisfies $\mathbf{A}_p\mathbf{Z}_{pp} = \mathbf{Z}_{pp}\mathbf{A}_p$.

Proof. By Theorem 8.1.1, the (p, q) block of $\mathbf{D}\mathbf{Z}$ is $\mathbf{A}_p\mathbf{Z}_{pq}$ and the (p, q) block of $\mathbf{Z}\mathbf{D}$ is $\mathbf{Z}_{pq}\mathbf{A}_q$, because $\mathbf{D}$ has only one nonzero block in each block row and in each block column.

So $\mathbf{DZ} = \mathbf{ZD}$ is equivalent to

$$\mathbf{A}_p\mathbf{Z}_{pq} = \mathbf{Z}_{pq}\mathbf{A}_q \qquad (1 \le p, q \le r).$$

Let $p \neq q$. Then $\operatorname{spec}(\mathbf{A}_p) \cap \operatorname{spec}(\mathbf{A}_q) = \varnothing$ by hypothesis, so Theorem 12.10.2 ((a) $\Rightarrow$ (d)) forces $\mathbf{Z}_{pq} = \mathbf{O}$. The diagonal equations $\mathbf{A}_p\mathbf{Z}_{pp} = \mathbf{Z}_{pp}\mathbf{A}_p$ are the remaining cases $p = q$, and they carry no information beyond commuting.

■

Read the statement twice, because a similar-sounding claim is false.

Warning. **It is the disjointness of the spectra that does the work, not the block diagonal shape.** Nothing here says that a matrix commuting with a block diagonal matrix is block diagonal, and nothing here says that a commutant consists of triangular matrices. Both fail as soon as two diagonal blocks share an eigenvalue. Chapter 10 §04 computes the commutant of $\mathbf{J} = \mathbf{J}_2(0) \oplus \mathbf{J}_1(0)$ explicitly and finds it to be

$$\left\{ \begin{pmatrix} a & b & c \\ 0 & a & 0 \\ 0 & h & i \end{pmatrix} : a, b, c, h, i \in F \right\},$$

which contains matrices with a nonzero entry **below** the diagonal, in position $(3, 2)$. That does not contradict Corollary 12.10.4: the two blocks $\mathbf{J}_2(0)$ and $\mathbf{J}_1(0)$ have the same spectrum $\{0\}$, so the corollary has nothing to say about them. What is true for a single repeated eigenvalue is weaker and needs a different normal form: the commutant of a basic **Weyr** matrix is block upper triangular for the Weyr partition (Proposition 10.4.6), and the Jordan form has no such property. Note too that even when Corollary 12.10.4 applies, it says nothing about the diagonal blocks $\mathbf{Z}_{pp}$ themselves, which range over the whole commutant of $\mathbf{A}_p$.

The corollary is the reason spectral separation is so useful. Combined with the primary decomposition above, it says that over $\mathbb{C}$ the commutant of any matrix splits along its distinct eigenvalues: after the change of basis that block-diagonalizes $\mathbf{A}$ by spectrum, everything commuting with $\mathbf{A}$ becomes block diagonal too, and the hard part of describing a commutant is always the single-eigenvalue part. The pleasant case of the same phenomenon appeared in Theorem 12.7.1, where the pieces are honest eigenspaces and the splitting is orthogonal.

12.10.6 Lyapunov's equation

One specialization is important enough to have its own name.

Definition 12.10.2. Lyapunov equation

Let $\mathbf{A} \in M_n(\mathbb{C})$ and $\mathbf{Q} \in M_n(\mathbb{C})$. The **Lyapunov equation** for the pair $(\mathbf{A}, \mathbf{Q})$ is

$$\mathbf{AX} + \mathbf{XA}^* = -\mathbf{Q},$$

with unknown $\mathbf{X} \in M_n(\mathbb{C})$. Call $\mathbf{A}$ **stable** when every $\lambda \in \operatorname{spec}(\mathbf{A})$ has $\operatorname{Re}\lambda < 0$.

This is the Sylvester equation with $\mathbf{B} = -\mathbf{A}^*$, since $\mathbf{AX} - \mathbf{X}(-\mathbf{A}^*) = \mathbf{AX} + \mathbf{XA}^*$. To apply

Theorem 12.10.2 we need $\operatorname{spec}(-\mathbf{A}^*)$.

The eigenvalues of $\mathbf{A}^*$ are the conjugates of those of $\mathbf{A}$. Indeed $\det(\mathbf{M}^*) = \overline{\det \mathbf{M}}$ for every $\mathbf{M} \in M_n(\mathbb{C})$, because $\det(\mathbf{M}^\top) = \det \mathbf{M}$ (Theorem 7.4.5) and the determinant is a sum of products of entries with coefficients ± 1 (Definition 7.4.1), so conjugating every entry conjugates the value. Applying this to $\mathbf{M} = \lambda \mathbf{I} - \mathbf{A}$, whose conjugate transpose is $\overline{\lambda}\mathbf{I} - \mathbf{A}^*$, gives $\det(\overline{\lambda}\mathbf{I} - \mathbf{A}^*) = \overline{\det(\lambda \mathbf{I} - \mathbf{A})}$. So one determinant vanishes exactly when the other does, and $\lambda \in \operatorname{spec}(\mathbf{A})$ if and only if $\overline{\lambda} \in \operatorname{spec}(\mathbf{A}^*)$ (Theorem 9.2.1).

Therefore $\operatorname{spec}(-\mathbf{A}^*) = \{-\overline{\lambda} : \lambda \in \operatorname{spec}(\mathbf{A})\}$, and Theorem 12.10.2 says the Lyapunov equation has exactly one solution for every $\mathbf{Q}$ precisely when

$$\lambda + \overline{\mu} \neq 0 \qquad \text{for all } \lambda, \mu \in \operatorname{spec}(\mathbf{A}).$$

If $\mathbf{A}$ is stable this holds, because then $\operatorname{Re}(\lambda + \overline{\mu}) = \operatorname{Re}\lambda + \operatorname{Re}\mu < 0$. So a stable matrix gives a uniquely solvable Lyapunov equation, whatever $\mathbf{Q}$ is.

What makes the equation important is a statement we are not yet equipped to prove. Call a Hermitian matrix $\mathbf{X}$ **positive definite** when $\mathbf{x}^*\mathbf{X}\mathbf{x} > 0$ for every nonzero $\mathbf{x} \in \mathbb{C}^n$. Then: *if $\mathbf{A}$ is stable and $\mathbf{Q}$ is positive definite, the unique solution $\mathbf{X}$ of $\mathbf{A}\mathbf{X} + \mathbf{X}\mathbf{A}^* = -\mathbf{Q}$ is positive definite; and conversely, if some positive definite $\mathbf{X}$ solves the equation for some positive definite $\mathbf{Q}$, then $\mathbf{A}$ is stable.* Positivity is the subject of Chapter 13, where the tools to handle $\mathbf{x}^*\mathbf{X}\mathbf{x}$ are built, and the consequence for differential equations — that a stable $\mathbf{A}$ makes every solution of $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$ decay, with $\mathbf{x}^*\mathbf{X}\mathbf{x}$ as the quantity that measures the decay — belongs to Chapter 16. We record the equation here because its *solvability*, which is the part Sylvester's theorem settles, is exactly what makes the rest possible.

12.10.7 Exercises

12.10.7.1 A. Check your understanding

Exercise. A1

(a) Define the Sylvester operator of a pair $(\mathbf{A}, \mathbf{B})$ with $\mathbf{A} \in M_m(F)$, $\mathbf{B} \in M_n(F)$, and state the dimension of the space it acts on.

(b) Write down the matrix of the Sylvester operator with respect to the vectorization vec, and say which Chapter 8 result produces it.

(c) State the criterion for $\mathbf{A}\mathbf{X} - \mathbf{X}\mathbf{B} = \mathbf{C}$ to have exactly one solution for every $\mathbf{C}$, over $\mathbb{C}$.

(d) True or false: if $\mathbf{A}\mathbf{X} - \mathbf{X}\mathbf{B} = \mathbf{C}$ has at least one solution for every $\mathbf{C}$, then that solution is unique. Justify your answer.

(e) True or false: if $\operatorname{spec}(\mathbf{A}) \cap \operatorname{spec}(\mathbf{B}) = \varnothing$, then $\mathbf{A}\mathbf{X} + \mathbf{X}\mathbf{B} = \mathbf{C}$ has exactly one solution for every $\mathbf{C}$. Justify your answer.

Solution.

(a) $\mathcal{S}_{\mathbf{A},\mathbf{B}}(\mathbf{X}) = \mathbf{A}\mathbf{X} - \mathbf{X}\mathbf{B}$ for $\mathbf{X} \in M_{m\times n}(F)$ (Definition 12.10.1). It acts on $M_{m\times n}(F)$, of dimension mn.

(b) $\mathbf{K} = \mathbf{I}_n \otimes \mathbf{A} - \mathbf{B}^\top \otimes \mathbf{I}_m$, by Eq. 12.10.1; it comes from the vec identity Theorem 8.4.4 applied twice, with $\operatorname{vec}(\mathbf{A}\mathbf{X}) = (\mathbf{I}_n \otimes \mathbf{A}) \operatorname{vec} \mathbf{X}$ and $\operatorname{vec}(\mathbf{X}\mathbf{B}) = (\mathbf{B}^\top \otimes \mathbf{I}_m) \operatorname{vec} \mathbf{X}$.

(c) Exactly when $\mathbf{A}$ and $\mathbf{B}$ have no common eigenvalue, that is $\operatorname{spec}(\mathbf{A}) \cap \operatorname{spec}(\mathbf{B}) = \varnothing$

(Theorem 12.10.2).

(d) True. Existence for every $\mathbf{C}$ says $\mathcal{S}_{\mathbf{A},\mathbf{B}}$ is surjective; an operator on the finite-dimensional space $M_{m\times n}(F)$ is surjective if and only if it is injective (Theorem 3.4.4 through vec, or Rank–Nullity), and injectivity is uniqueness.

(e) False. The criterion for the $+$ form is $\lambda+\mu\neq 0$ for all $\lambda\in\operatorname{spec}(\mathbf{A})$, $\mu\in\operatorname{spec}(\mathbf{B})$, since $\mathbf{AX}+\mathbf{XB}=\mathcal{S}_{\mathbf{A},-\mathbf{B}}(\mathbf{X})$. With $\mathbf{A}=\operatorname{diag}(1,2)$ and $\mathbf{B}=\operatorname{diag}(-1,3)$ the spectra are disjoint but $1+(-1)=0$, so the equation is not uniquely solvable.

12.10.7.2 B. Practice

Exercise. B1: A Sylvester equation by vectorization

Over $\mathbb{Q}$, let

$$\mathbf{A}=\begin{pmatrix}3&1\\0&1\end{pmatrix},\qquad \mathbf{B}=\begin{pmatrix}-1&0\\1&0\end{pmatrix},\qquad \mathbf{C}=\begin{pmatrix}2&1\\0&-2\end{pmatrix}.$$

(a) Compute $\operatorname{spec}(\mathbf{A})$ and $\operatorname{spec}(\mathbf{B})$, and decide whether $\mathbf{AX}-\mathbf{XB}=\mathbf{C}$ has exactly one solution.

(b) Write down $\mathbf{K}=\mathbf{I}_2\otimes\mathbf{A}-\mathbf{B}^{\top}\otimes\mathbf{I}_2$, and check that its diagonal entries are the four differences $\lambda_i-\mu_j$.

(c) Solve the equation.

Solution.

(a) $\mathbf{A}$ is upper triangular with diagonal $3,1$, so $\operatorname{spec}(\mathbf{A})=\{3,1\}$ (Theorem 9.7.4). For $\mathbf{B}$, $p_{\mathbf{B}}=\det\begin{pmatrix}x+1&0\\-1&x\end{pmatrix}=x(x+1)$, so $\operatorname{spec}(\mathbf{B})=\{0,-1\}$. The spectra are disjoint, so by Theorem 12.10.2 there is exactly one solution for this (or any) $\mathbf{C}$.

(b) $\mathbf{B}^{\top}=\begin{pmatrix}-1&1\\0&0\end{pmatrix}$, so

$$\mathbf{I}_2\otimes\mathbf{A}=\begin{pmatrix}3&1&0&0\\0&1&0&0\\0&0&3&1\\0&0&0&1\end{pmatrix},\qquad \mathbf{B}^{\top}\otimes\mathbf{I}_2=\begin{pmatrix}-1&0&1&0\\0&-1&0&1\\0&0&0&0\\0&0&0&0\end{pmatrix},$$

and

$$\mathbf{K}=\begin{pmatrix}4&1&-1&0\\0&2&0&-1\\0&0&3&1\\0&0&0&1\end{pmatrix}.$$

Its diagonal is $4,2,3,1$, and the differences $\lambda_i-\mu_j$ are $3-(-1)=4$, $1-(-1)=2$, $3-0=3$, $1-0=1$, as Lemma 12.10.1 asserts.

(c) With $\operatorname{vec}\mathbf{X}=(x_{11},x_{21},x_{12},x_{22})$ and $\operatorname{vec}\mathbf{C}=(2,0,1,-2)$, back substitution gives $x_{22}=-2$; then $3x_{12}+x_{22}=1$, so $x_{12}=1$; then $2x_{21}-x_{22}=0$, so $x_{21}=-1$; then

$4x_{11} + x_{21} - x_{12} = 2$, so $x_{11} = 1$. Hence

$$\mathbf{X} = \begin{pmatrix} 1 & 1 \\ -1 & -2 \end{pmatrix}.$$

Check: $\mathbf{AX} = \begin{pmatrix} 2 & 1 \\ -1 & -2 \end{pmatrix}$ and $\mathbf{XB} = \begin{pmatrix} 0 & 0 \\ -1 & 0 \end{pmatrix}$, and their difference is $\mathbf{C}$.

Exercise. B2: Block-diagonalizing a 3×3 matrix

Let

$$\mathbf{M} = \begin{pmatrix} 2 & 1 & 2 \\ 0 & 2 & 3 \\ 0 & 0 & 5 \end{pmatrix} \in M_3(\mathbb{Q}).$$

(a) Partition $\mathbf{M}$ as $3 = 2+1$, name the blocks $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, and explain why Corollary 12.10.3 applies.

(b) Solve the corresponding Sylvester equation.

(c) Write down $\mathbf{P}$ and verify that $\mathbf{P}^{-1}\mathbf{MP} = \mathbf{A} \oplus \mathbf{B}$.

Solution.

(a) The blocks are $\mathbf{A} = \begin{pmatrix} 2 & 1 \\ 0 & 2 \end{pmatrix}$, $\mathbf{B} = (5)$, and $\mathbf{C} = \begin{pmatrix} 2 \\ 3 \end{pmatrix}$. Both diagonal blocks are triangular, so $\operatorname{spec}(\mathbf{A}) = \{2\}$ and $\operatorname{spec}(\mathbf{B}) = \{5\}$ (Theorem 9.7.4); these are disjoint, so Corollary 12.10.3 (c) applies and a unique $\mathbf{X} \in M_{2\times 1}(\mathbb{Q})$ exists.

(b) With $\mathbf{X} = \mathbf{x} = (x_1, x_2)$, the equation $\mathbf{Ax} - 5\mathbf{x} = -\mathbf{C}$ reads $(\mathbf{A} - 5\mathbf{I}_2)\mathbf{x} = -\mathbf{C}$, that is

$$\begin{pmatrix} -3 & 1 \\ 0 & -3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} -2 \\ -3 \end{pmatrix}.$$

The second row gives $x_2 = 1$, and the first then gives $-3x_1 + 1 = -2$, so $x_1 = 1$. Hence $\mathbf{x} = (1, 1)$.

(c) Taking

$$\mathbf{P} = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}, \qquad \mathbf{P}^{-1} = \begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \\ 0 & 0 & 1 \end{pmatrix},$$

we get $\mathbf{MP} = \begin{pmatrix} 2 & 1 & 5 \\ 0 & 2 & 5 \\ 0 & 0 & 5 \end{pmatrix}$ and then

$$\mathbf{P}^{-1}\mathbf{MP} = \begin{pmatrix} 2 & 1 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 5 \end{pmatrix} = \mathbf{A} \oplus \mathbf{B},$$

as Corollary 12.10.3 (a) predicts.

Exercise. B3: Which equations are uniquely solvable?

Determine, for each of the following, whether the stated equation has exactly one solution $\mathbf{X} \in M_2(\mathbb{C})$ for **every** right-hand side $\mathbf{C} \in M_2(\mathbb{C})$. Justify your answer.

(a) $\mathbf{AX} - \mathbf{XB} = \mathbf{C}$ with $\mathbf{A} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$, $\mathbf{B} = \begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix}$.

(b) $\mathbf{AX} - \mathbf{XB} = \mathbf{C}$ with $\mathbf{A} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$, $\mathbf{B} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$.

(c) $\mathbf{AX} + \mathbf{XB} = \mathbf{C}$ with $\mathbf{A} = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$, $\mathbf{B} = \begin{pmatrix} -1 & 5 \\ 0 & 3 \end{pmatrix}$.

(d) $\mathbf{AX} - \mathbf{XB} = \mathbf{C}$ with $\mathbf{A} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$, $\mathbf{B} = \begin{pmatrix} 1 & 0 \\ 2 & 1 \end{pmatrix}$.

(e) $\mathbf{AX} - \mathbf{XB} = \mathbf{C}$ with $\mathbf{A} \in M_2(\mathbb{C})$ nilpotent and $\mathbf{B} \in M_2(\mathbb{C})$ invertible.

Solution.

(a) Yes. $\operatorname{spec}(\mathbf{A}) = \{1\}$ and $\operatorname{spec}(\mathbf{B}) = \{2, 3\}$ are disjoint (Theorem 9.7.4), so Theorem 12.10.2 applies.

(b) No. Both characteristic polynomials are $x^2 + 1$, so $\operatorname{spec}(\mathbf{A}) = \operatorname{spec}(\mathbf{B}) = \{i, -i\}$, which is as far from disjoint as possible. A witness is easy to produce here: $\mathbf{X} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$ gives $\mathbf{AX} = \begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix} = \mathbf{XB}$, so $\mathbf{X}$ is a nonzero member of $\ker \mathcal{S}_{\mathbf{A},\mathbf{B}}$ and condition (d) of Theorem 12.10.2 fails.

(c) No. This is the $+$ form, so the operator is $\mathcal{S}_{\mathbf{A},-\mathbf{B}}$ and the condition is $\lambda + \mu \neq 0$. Here $\operatorname{spec}(\mathbf{A}) = \{1, 2\}$ and $\operatorname{spec}(\mathbf{B}) = \{-1, 3\}$, and $1 + (-1) = 0$.

(d) No. Both are triangular with diagonal $1, 1$, so $\operatorname{spec}(\mathbf{A}) = \operatorname{spec}(\mathbf{B}) = \{1\}$.

(e) Yes. A nilpotent matrix has $\operatorname{spec}(\mathbf{A}) = \{0\}$, since $\mathbf{A}^k = \mathbf{O}$ forces $\lambda^k = 0$ for every eigenvalue λ; and $0 \notin \operatorname{spec}(\mathbf{B})$ because $\mathbf{B}$ is invertible (Theorem 9.2.6). So the spectra are disjoint.

12.10.7.3 C. Going deeper

Exercise. C1: Eigenvectors of the Sylvester operator

Let $\mathbf{A} \in M_m(\mathbb{C})$ and $\mathbf{B} \in M_n(\mathbb{C})$, and let $\mathcal{S} = \mathcal{S}_{\mathbf{A},\mathbf{B}}$.

(a) Suppose $\mathbf{Ax} = \lambda \mathbf{x}$ with $\mathbf{x} \in \mathbb{C}^m$ nonzero, and suppose $\mathbf{y} \in \mathbb{C}^n$ is a left eigenvector of $\mathbf{B}$ for μ, so that $\mathbf{y}^{\top}\mathbf{B} = \mu \mathbf{y}^{\top}$ (Definition 9.10.1). Prove that $\mathbf{x}\mathbf{y}^{\top} \neq \mathbf{O}$ and that $\mathcal{S}(\mathbf{x}\mathbf{y}^{\top}) = (\lambda - \mu)\, \mathbf{x}\mathbf{y}^{\top}$.

(b) Deduce that every difference $\lambda - \mu$ with $\lambda \in \operatorname{spec}(\mathbf{A})$ and $\mu \in \operatorname{spec}(\mathbf{B})$ is an eigenvalue of $\mathcal{S}$.

(c) Hence give a second proof that $\mathcal{S}$ is not invertible when $\mathbf{A}$ and $\mathbf{B}$ have a common eigenvalue, one that uses no Kronecker products.

Solution.

(a) Choose i with $x_i \neq 0$ and j with $y_j \neq 0$; these exist because $\mathbf{x}$ and $\mathbf{y}$ are nonzero. The (i,j) entry of $\mathbf{x}\mathbf{y}^\top$ is $x_i y_j \neq 0$, so $\mathbf{x}\mathbf{y}^\top \neq \mathbf{O}$. For the eigenvalue equation, associativity of matrix multiplication (Theorem 0.9.4) gives

$$\mathcal{S}(\mathbf{x}\mathbf{y}^\top) = (\mathbf{A}\mathbf{x})\mathbf{y}^\top - \mathbf{x}(\mathbf{y}^\top\mathbf{B}) = \lambda\mathbf{x}\mathbf{y}^\top - \mu\mathbf{x}\mathbf{y}^\top,$$

which is $(\lambda-\mu)\mathbf{x}\mathbf{y}^\top$.

(b) Let $\lambda \in \operatorname{spec}(\mathbf{A})$ and $\mu \in \operatorname{spec}(\mathbf{B})$. A right eigenvector $\mathbf{x}$ for λ exists by Definition 9.2.1, and a left eigenvector $\mathbf{y}$ for μ exists by Proposition 9.10.5 (b). By (a), $\mathbf{x}\mathbf{y}^\top$ is a nonzero matrix with $\mathcal{S}(\mathbf{x}\mathbf{y}^\top) = (\lambda-\mu)\mathbf{x}\mathbf{y}^\top$, so $\lambda-\mu$ is an eigenvalue of $\mathcal{S}$ with eigenvector $\mathbf{x}\mathbf{y}^\top$.

(c) If $\lambda \in \operatorname{spec}(\mathbf{A}) \cap \operatorname{spec}(\mathbf{B})$, apply (b) with $\mu = \lambda$: then 0 is an eigenvalue of $\mathcal{S}$, so $\ker \mathcal{S} \neq \{\mathbf{O}\}$ and $\mathcal{S}$ is not invertible. This shows one implication of Theorem 12.10.2 directly. (It does not give the converse: knowing every $\lambda-\mu$ is an eigenvalue does not by itself show there are no others.)

Exercise. C2: The kernel when the spectra meet

Let $\mathbf{A} \in M_m(\mathbb{C})$ and $\mathbf{B} \in M_n(\mathbb{C})$ have eigenvalue lists $\lambda_1, \dots, \lambda_m$ and $\mu_1, \dots, \mu_n$ with algebraic multiplicity, and let $d = \#\{(i,j) : \lambda_i = \mu_j\}$.

(a) Prove that $\dim\ker \mathcal{S}_{\mathbf{A},\mathbf{B}} \geq 1$ if and only if $d \geq 1$.

(b) Prove that $\dim\ker \mathcal{S}_{\mathbf{A},\mathbf{B}} \leq d$.

(c) Show that the inequality in (b) can be strict, by computing $\ker \mathcal{S}_{\mathbf{A},\mathbf{B}}$ for $\mathbf{A} = \mathbf{B} = \mathbf{J}_2(0)$.

Solution.

(a) $d \geq 1$ says exactly that $\mathbf{A}$ and $\mathbf{B}$ have a common eigenvalue, and by Theorem 12.10.2 ((a) $\Leftrightarrow$ (d)) that is exactly the failure of $\ker \mathcal{S}_{\mathbf{A},\mathbf{B}} = \{\mathbf{O}\}$.

(b) The operator $\mathcal{S}_{\mathbf{A},\mathbf{B}}$ has matrix $\mathbf{K}$ (Eq. 12.10.1), so $\dim\ker \mathcal{S}_{\mathbf{A},\mathbf{B}} = g_{\mathbf{K}}(0)$, the geometric multiplicity of 0 as an eigenvalue of $\mathbf{K}$ (when $0 \notin \operatorname{spec}(\mathbf{K})$ both sides are 0 and there is nothing to prove). By Lemma 12.10.1 the eigenvalue list of $\mathbf{K}$ consists of the mn differences $\lambda_i - \mu_j$, so the algebraic multiplicity is $a_{\mathbf{K}}(0) = d$. By Theorem 9.3.1, $g_{\mathbf{K}}(0) \leq a_{\mathbf{K}}(0) = d$.

(c) Here $m = n = 2$ and $\lambda_1 = \lambda_2 = \mu_1 = \mu_2 = 0$, so $d = 4$. The kernel is the commutant of $\mathbf{J} = \mathbf{J}_2(0)$: writing $\mathbf{X} = \begin{pmatrix} p & q \\ r & s \end{pmatrix}$,

$$\mathbf{J}\mathbf{X} = \begin{pmatrix} r & s \\ 0 & 0 \end{pmatrix}, \qquad \mathbf{X}\mathbf{J} = \begin{pmatrix} 0 & p \\ 0 & r \end{pmatrix},$$

so $\mathbf{J}\mathbf{X} = \mathbf{X}\mathbf{J}$ forces $r = 0$ and $s = p$, and leaves q free. Hence $\ker \mathcal{S}_{\mathbf{J},\mathbf{J}} = \operatorname{span}(\mathbf{I}_2, \mathbf{J})$, of dimension $2 < 4 = d$.

Exercise. C3: Only the scalars commute with everything

Let F be a field, $n \geq 2$, and let $\mathbf{E}_{ij} \in M_n(F)$ be the matrix unit with 1 in position (i, j) and 0 elsewhere.

(a) Suppose $\mathbf{A} \in M_n(F)$ satisfies $\mathbf{A}\mathbf{E}_{ij} = \mathbf{E}_{ij}\mathbf{A}$ for all $i \neq j$. Prove that $\mathbf{A} = c\mathbf{I}_n$ for some $c \in F$.

(b) Deduce that the only matrices commuting with every $\mathbf{X} \in M_n(F)$ are the scalar matrices, that is, $\ker \mathcal{S}_{\mathbf{A},\mathbf{A}} = M_n(F)$ forces $\mathbf{A}$ scalar.

(c) Deduce that the only matrices commuting with every **invertible** matrix are again the scalar matrices.

Hint for (c): $\mathbf{I}_n + \mathbf{E}_{ij}$ *is invertible when* $i \neq j$.

Solution.

(a) Fix $i \neq j$. By Theorem 0.9.2, $\mathbf{A}\mathbf{E}_{ij}$ has column j equal to $\mathbf{A}\mathbf{e}_i$, that is column i of $\mathbf{A}$, and all other columns zero; while $\mathbf{E}_{ij}\mathbf{A}$ has row i equal to row j of $\mathbf{A}$, and all other rows zero. Compare entries.
Take $k \neq i$. The (k, j) entry of $\mathbf{A}\mathbf{E}_{ij}$ is a_{ki}, and the (k, j) entry of $\mathbf{E}_{ij}\mathbf{A}$ is 0, because row $k \neq i$ of $\mathbf{E}_{ij}\mathbf{A}$ is zero. Hence $a_{ki} = 0$ whenever $k \neq i$. Since $n \geq 2$, every index i has some $j \neq i$ available, so this holds for every i, and $\mathbf{A}$ is diagonal.
Now the (i, j) entry: on the left it is a_{ii} (the i-th entry of column i of $\mathbf{A}$), and on the right it is a_{jj} (the j-th entry of row j of $\mathbf{A}$). So $a_{ii} = a_{jj}$ for all $i \neq j$. Writing c for this common value, $\mathbf{A} = c\mathbf{I}_n$.

(b) If $\mathbf{A}\mathbf{X} = \mathbf{X}\mathbf{A}$ for every $\mathbf{X} \in M_n(F)$, then in particular for $\mathbf{X} = \mathbf{E}_{ij}$ with $i \neq j$, so (a) gives $\mathbf{A} = c\mathbf{I}_n$. Conversely every scalar matrix commutes with everything. Hence $\ker \mathcal{S}_{\mathbf{A},\mathbf{A}} = M_n(F)$ if and only if $\mathbf{A}$ is scalar.

(c) For $i \neq j$ we have $\mathbf{E}_{ij}^2 = \mathbf{O}$, so $(\mathbf{I}_n + \mathbf{E}_{ij})(\mathbf{I}_n - \mathbf{E}_{ij}) = \mathbf{I}_n$ and $\mathbf{I}_n + \mathbf{E}_{ij}$ is invertible. If $\mathbf{A}$ commutes with every invertible matrix, it commutes with $\mathbf{I}_n + \mathbf{E}_{ij}$, hence with $\mathbf{E}_{ij}$ (subtract $\mathbf{A}\mathbf{I}_n = \mathbf{I}_n\mathbf{A}$), for all $i \neq j$. By (a), $\mathbf{A} = c\mathbf{I}_n$.

12.11 How Far from Normal

Every theorem so far in this chapter has been a yes-or-no statement. An operator either commutes with its adjoint or it does not; there either is an orthonormal basis of eigenvectors or there is not. That is unsatisfying, because the two matrices $\begin{pmatrix} 1 & 10^{-6} \\ 0 & 2 \end{pmatrix}$ and $\begin{pmatrix} 1 & 10^{6} \\ 0 & 2 \end{pmatrix}$ are both nonnormal, and no one would want to treat them alike. This section replaces the yes-or-no answer with a number. The number comes out of Schur's theorem almost for free, once we know that the Frobenius norm cannot tell one orthonormal basis from another, and it turns out to measure exactly the mass that a Schur factorization is forced to leave above the diagonal.

Throughout, $\mathbf{A} \in M_n(\mathbb{C})$ and $\mathbb{C}^n$ carries the standard inner product. Eigenvalues are always counted **in** $\mathbb{C}$ and **with algebraic multiplicity**, so that an $n \times n$ matrix always has a list of exactly n of them (Corollary 6.5.5).

12.11.1 The yardstick

Sizes here are measured in the **Frobenius norm**, which Chapter 11 §07 defined as the norm induced by the Frobenius inner product:

$$\|\mathbf{A}\|_F = \sqrt{\langle \mathbf{A}, \mathbf{A} \rangle} = \Big(\sum_{i,j} |a_{ij}|^2 \Big)^{1/2} = \sqrt{\operatorname{tr}(\mathbf{A}^* \mathbf{A})},$$

where $\langle \mathbf{A}, \mathbf{B} \rangle = \operatorname{tr}(\mathbf{B}^* \mathbf{A})$. It is the right instrument here for one reason, proved there: unitary factors are invisible to it, so $\|\mathbf{UA}\|_F = \|\mathbf{A}\|_F = \|\mathbf{AV}\|_F$ for unitary $\mathbf{U}$ and $\mathbf{V}$, and in particular a unitary similarity leaves it alone (Lemma 11.7.5). Everything below follows from that one sentence and from Schur's theorem.

12.11.2 Schur's inequality

Here is the observation the section is built on. Write a matrix in a Schur form $\mathbf{A} = \mathbf{UTU}^*$. The diagonal of $\mathbf{T}$ holds the eigenvalues, so $\sum_i |\lambda_i|^2$ is the Frobenius mass sitting *on* the diagonal of $\mathbf{T}$. The rest of $\mathbf{T}$ sits strictly above it. And by Lemma 11.7.5 the total mass of $\mathbf{T}$ is the total mass of $\mathbf{A}$. Rearranged, that is a theorem.

Theorem 12.11.1. Schur's Inequality

Let $\mathbf{A} \in M_n(\mathbb{C})$ have eigenvalues $\lambda_1, \dots, \lambda_n$, listed with algebraic multiplicity. Then

$$\sum_{i=1}^{n} |\lambda_i|^2 \;\le\; \|\mathbf{A}\|_F^2,$$

with **equality if and only if $\mathbf{A}$ is normal**.

Idea. Schur-triangularize, then count. ① $\mathbf{A} = \mathbf{UTU}^*$ with $\mathbf{T}$ upper triangular and the λ_i on its diagonal (Corollary 12.1.3). ② The Frobenius norm does not notice the $\mathbf{U}$'s, so $\|\mathbf{A}\|_F = \|\mathbf{T}\|_F$. ③ Splitting $\|\mathbf{T}\|_F^2$ into its diagonal part and its strictly upper part gives the inequality on the nose, with the strictly upper part as the exact surplus. ④ For the equality case, that surplus vanishes exactly when $\mathbf{T}$ is diagonal, and "triangular and normal" and "diagonal" were shown to be the same thing in Lemma 12.4.1.

Proof. By Corollary 12.1.3 there are a unitary $\mathbf{U}$ and an upper triangular $\mathbf{T} = (t_{ij})$ with $\mathbf{A} = \mathbf{UTU}^*$ and $t_{ii} = \lambda_i$ for every i. Since $t_{ij} = 0$ whenever $i > j$, the entries of $\mathbf{T}$ fall into the diagonal ones and the strictly upper ones, so by Lemma 11.7.5

$$\begin{aligned} \|\mathbf{A}\|_F^2 = \|\mathbf{T}\|_F^2 &= \sum_{i=1}^{n} |t_{ii}|^2 + \sum_{i<j} |t_{ij}|^2 \\ &= \sum_{i=1}^{n} |\lambda_i|^2 + \sum_{i<j} |t_{ij}|^2. \end{aligned} \tag{12.11.1}$$

The second sum is a sum of nonnegative reals, so it is ≥ 0, and the inequality follows. Moreover equality holds if and only if that sum is 0, which happens if and only if $t_{ij} = 0$ for all $i < j$, that is, if and only if **this $\mathbf{T}$** is diagonal.

($\Leftarrow$) Suppose $\mathbf{A}$ is normal. Then so is $\mathbf{T} = \mathbf{U}^*\mathbf{A}\mathbf{U}$: using $\mathbf{U}\mathbf{U}^* = \mathbf{I}$ in the middle of each product,

$$\mathbf{T}^*\mathbf{T} = \mathbf{U}^*\mathbf{A}^*\mathbf{A}\mathbf{U} = \mathbf{U}^*\mathbf{A}\mathbf{A}^*\mathbf{U} = \mathbf{T}\mathbf{T}^*.$$

A normal upper triangular matrix is diagonal (Lemma 12.4.1), so $\mathbf{T}$ is diagonal and Eq. 12.11.1 gives equality.

($\Rightarrow$) Suppose equality holds. Then, as noted, $\mathbf{T}$ is diagonal, so $\mathbf{A} = \mathbf{U}\mathbf{T}\mathbf{U}^*$ exhibits $\mathbf{A}$ as unitarily diagonalizable, and Corollary 12.4.3 makes $\mathbf{A}$ normal.

This proves the inequality and both halves of the equality case.

■

Notice that the proof never had to worry about *which* Schur factorization was used. The two sides of the inequality — the eigenvalue list and $\|\mathbf{A}\|_F$ — are determined by $\mathbf{A}$ alone, so the surplus $\sum_{i<j}|t_{ij}|^2$ in Eq. 12.11.1 comes out the same for every triangularization of $\mathbf{A}$, even though the individual entries t_{ij} certainly do not. That is worth a second look, and we will look at it in a moment.

Example. Both sides of Schur's inequality

Verify Theorem 12.11.1 for

$$\mathbf{A} = \begin{pmatrix} 0 & 1 \\ 2 & 0 \end{pmatrix},$$

and locate the surplus inside an explicit Schur factorization (Corollary 12.1.3).

Solution. Here $\operatorname{tr}\mathbf{A} = 0$ and $\det\mathbf{A} = -2$, so $p_{\mathbf{A}} = x^2 - 2$ and the eigenvalues are $\pm\sqrt{2}$. Thus

$$\sum_i |\lambda_i|^2 = 2 + 2 = 4, \qquad \|\mathbf{A}\|_F^2 = 0 + 1 + 4 + 0 = 5,$$

and indeed $4 < 5$. The inequality is strict, so Theorem 12.11.1 already says that $\mathbf{A}$ is not normal — no product of matrices needed.

For the factorization, solve $(\mathbf{A} - \sqrt{2}\mathbf{I})\mathbf{x} = \mathbf{0}$: the first row reads $-\sqrt{2}x_1 + x_2 = 0$, giving the eigenvector $(1, \sqrt{2})$ of norm $\sqrt{3}$. Take

$$\mathbf{q}_1 = \tfrac{1}{\sqrt{3}}(1, \sqrt{2}), \qquad \mathbf{q}_2 = \tfrac{1}{\sqrt{3}}(-\sqrt{2}, 1),$$

a pair of orthonormal vectors, and let $\mathbf{U}$ have them as columns. Then

$$\mathbf{U}^*\mathbf{A}\mathbf{U} = \begin{pmatrix} \sqrt{2} & -1 \\ 0 & -\sqrt{2} \end{pmatrix} = \mathbf{T}.$$

Its Frobenius mass is $2 + 1 + 2 = 5 = \|\mathbf{A}\|_F^2$, as Lemma 11.7.5 promises, and the split of Eq. 12.11.1 reads $5 = 4 + 1$: the whole surplus is the single entry -1 above the diagonal.

Quick check. The real matrix $\mathbf{A} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ has $\|\mathbf{A}\|_F^2 = 2$ and no real eigenvalue at all.

Does Theorem 12.11.1 fail for it?

Solution. No, because the eigenvalues are counted in $\mathbb{C}$. They are i and $-i$, so $\sum_i |\lambda_i|^2 = 1 + 1 = 2 = \|\mathbf{A}\|_F^2$: equality, which is exactly right, since $\mathbf{A}^\top\mathbf{A} = \mathbf{A}\mathbf{A}^\top = \mathbf{I}$ makes $\mathbf{A}$ normal. Restricting the eigenvalue list to $\mathbb{R}$ would leave the left-hand side empty and the statement meaningless; this is one more place where the Fundamental Theorem of Algebra (Theorem 6.5.1) is doing silent work.

12.11.3 The departure from normality

Schur's inequality has a slack term, and the slack term is the interesting object. It is nonnegative for every matrix; it vanishes exactly on the normal ones; and, both of its ingredients being unitary-similarity invariants, it survives a change of orthonormal basis. Those are precisely the properties one wants of a measurement, so we name it.

The departure from normality is the Frobenius mass that no Schur factorization can push onto the diagonal.

Definition 12.11.1. Departure from Normality

Let $\mathbf{A} \in M_n(\mathbb{C})$ have eigenvalues $\lambda_1, \dots, \lambda_n$, listed with algebraic multiplicity. The **departure from normality** of $\mathbf{A}$ is the nonnegative real number $\Delta(\mathbf{A})$ defined by

$$\Delta(\mathbf{A})^2 := \|\mathbf{A}\|_F^2 - \sum_{i=1}^{n} |\lambda_i|^2.$$

Clause by clause: the quantity on the right is a real number because $\|\mathbf{A}\|_F^2$ and each $|\lambda_i|^2$ are real; it is **nonnegative** by Theorem 12.11.1, which is what allows us to take its nonnegative square root and call the result $\Delta(\mathbf{A})$; and the eigenvalue list, though it comes with no preferred order, contributes a sum that does not care about the order. So $\Delta(\mathbf{A})$ is well defined, with no choices made anywhere.

Two facts come straight from Theorem 12.11.1 and its proof.

- $\Delta(\mathbf{A}) = 0$ **if and only if $\mathbf{A}$ is normal.** This is the equality case, restated.
- $\Delta(\mathbf{A}) = \left(\sum_{i<j} |t_{ij}|^2\right)^{1/2}$ **for every** Schur factorization $\mathbf{A} = \mathbf{U}\mathbf{T}\mathbf{U}^*$, by Eq. 12.11.1. So $\Delta(\mathbf{A})$ is the length of the strictly upper triangular part of $\mathbf{T}$, and the remark after Theorem 12.11.1 says that this length is the same for all of them, even though the entries are not.

And one more, which is the reason the number deserves to be called a measurement rather than a coincidence: $\Delta(\mathbf{U}^*\mathbf{A}\mathbf{U}) = \Delta(\mathbf{A})$ for every unitary $\mathbf{U}$. Both ingredients are unitary-similarity invariants — the Frobenius norm by Lemma 11.7.5, the eigenvalue list because similar matrices have the same characteristic polynomial (Theorem 7.10.3). Exercise (C1: Unitary similarity preserves the defect, ordinary similarity does not) asks for the details, and for an example showing that ordinary similarity wrecks the number completely.

Example. The defect of a Jordan block

Compute $\Delta(\mathbf{J}_k(\lambda))$ for the $k \times k$ Jordan block with eigenvalue λ.

Solution. The matrix $\mathbf{J}_k(\lambda)$ has λ in each of its k diagonal positions and 1 in each of the $k-1$ superdiagonal positions, and zeros elsewhere, so

$$\|\mathbf{J}_k(\lambda)\|_F^2 = k|\lambda|^2 + (k-1).$$

It is upper triangular, so its eigenvalue list is λ repeated k times (Theorem 9.7.4), giving $\sum_i |\lambda_i|^2 = k|\lambda|^2$. Subtracting,

$$\Delta(\mathbf{J}_k(\lambda))^2 = k-1, \qquad \Delta(\mathbf{J}_k(\lambda)) = \sqrt{k-1}.$$

Three things are worth reading off. The answer does not depend on λ: sliding a Jordan block up and down the complex plane does not make it more or less normal. It is 0 exactly when $k = 1$, agreeing with the fact that a 1×1 matrix is normal. And it grows with the block, which matches what Theorem 12.3.5 gave us in Section 3: a normal operator has no Jordan block of size larger than 1 (Theorem 10.3.1). Here the defect is counting exactly the superdiagonal ones that normality forbids.

A minimal-change nonexample makes the same point on a smaller scale. The matrix $\operatorname{diag}(1,3)$ is normal, so its defect is 0. Change the single entry in position $(1,2)$ from 0 to 2:

$$\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 0 & 3 \end{pmatrix}, \qquad \|\mathbf{A}\|_F^2 = 14, \qquad \sum_i |\lambda_i|^2 = 1 + 9 = 10,$$

so $\Delta(\mathbf{A}) = 2$. The eigenvalues did not move — a triangular matrix wears them on its diagonal — and the entire change went into the defect. That is the general picture for triangular matrices: for them, Δ is just the length of everything above the diagonal.

The defect also bounds something one might actually want to know. Fix a Schur factorization $\mathbf{A} = \mathbf{U}\mathbf{T}\mathbf{U}^*$ (Corollary 12.1.3), write $\mathbf{D}$ for the diagonal part of $\mathbf{T}$, and put $\mathbf{N} = \mathbf{U}\mathbf{D}\mathbf{U}^*$, which is normal by Corollary 12.4.3. Then

$$\|\mathbf{A} - \mathbf{N}\|_F = \|\mathbf{T} - \mathbf{D}\|_F = \Delta(\mathbf{A}),$$

the first step by Lemma 11.7.5. So $\Delta(\mathbf{A})$ is an upper bound for the Frobenius distance from $\mathbf{A}$ to the set of normal matrices. It is not that distance.

Warning. **Δ is a defect, not a distance.** It is tempting to read $\Delta(\mathbf{A})$ as "how far $\mathbf{A}$ is from the nearest normal matrix", and the displayed bound above makes the temptation worse. But the bound can be strict. Take $\mathbf{A} = \mathbf{J}_2(0) = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$, so that $\Delta(\mathbf{A}) = 1$ by Example (The defect of a Jordan block). The symmetric — hence normal — matrix $\mathbf{N} = \frac{1}{2}\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ satisfies

$$\|\mathbf{A} - \mathbf{N}\|_F = \left\| \tfrac{1}{2} \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \right\|_F = \tfrac{1}{\sqrt{2}} < 1.$$

So some normal matrix is strictly nearer to $\mathbf{A}$ than $\Delta(\mathbf{A})$. What $\Delta(\mathbf{A})$ measures is the mass that survives *triangularization*, which is a genuine and useful quantity, but it is an upper

bound for the distance and not the distance itself.

12.11.4 Eigenvalues that will not sit still

Here is why anyone outside this chapter cares. Eigenvalues are computed, in practice, from data that is only approximately right, so the question "if I move the matrix a little, how far do the eigenvalues move?" is not academic. For a normal matrix the answer is "just as little". For a matrix with a large defect it can be "enormously further". The following pair of 2×2 matrices shows the whole phenomenon with nothing but the quadratic formula.

Example. A perturbation that moves the eigenvalues a thousand times as far

Fix a real number $M > 0$ and a real ε with $0 < \varepsilon < M$. Compare the effect of adding ε to the lower-left entry of each of

$$\mathbf{A} = \begin{pmatrix} 1 & M \\ 0 & 1 \end{pmatrix} \qquad \text{and} \qquad \mathbf{N} = \begin{pmatrix} 1 & M \\ -M & 1 \end{pmatrix}.$$

Compute the defects of $\mathbf{A}$ and $\mathbf{N}$, the eigenvalues of both perturbed matrices exactly, and the distance each eigenvalue travels.

Solution. The two defects. The matrix $\mathbf{A}$ is upper triangular with both diagonal entries 1, so its eigenvalue list is $1, 1$ and

$$\Delta(\mathbf{A})^2 = (1 + M^2 + 1) - (1 + 1) = M^2, \qquad \Delta(\mathbf{A}) = M.$$

The matrix $\mathbf{N}$ is normal: $\mathbf{N}^\top\mathbf{N} = \mathbf{N}\mathbf{N}^\top = (1 + M^2)\mathbf{I}$, a direct multiplication. So $\Delta(\mathbf{N}) = 0$. Its eigenvalues are $1 \pm Mi$, since $p_{\mathbf{N}} = (x - 1)^2 + M^2$. The two matrices have the same diagonal and the same entry M in the corner we are leaving alone; where they differ is that one has defect M and the other defect 0.

The nonnormal one. Perturbing the corner gives $\mathbf{A}_\varepsilon = \begin{pmatrix} 1 & M \\ \varepsilon & 1 \end{pmatrix}$, with

$$p_{\mathbf{A}_\varepsilon} = (x - 1)^2 - M\varepsilon, \qquad \lambda_\pm = 1 \pm \sqrt{M\varepsilon}.$$

The perturbation has size $\|\mathbf{A}_\varepsilon - \mathbf{A}\|_F = \varepsilon$, and each eigenvalue has moved a distance $\sqrt{M\varepsilon}$, which is $\sqrt{M/\varepsilon}$ times the size of the perturbation. With $M = 100$ and $\varepsilon = 10^{-4}$, the perturbation is one ten-thousandth and the eigenvalues move by $\sqrt{100 \cdot 10^{-4}} = 1/10$: a thousandfold amplification.

The normal one. Perturbing the same corner gives $\mathbf{N}_\varepsilon = \begin{pmatrix} 1 & M \\ -M + \varepsilon & 1 \end{pmatrix}$, with

$$p_{\mathbf{N}_\varepsilon} = (x - 1)^2 + M(M - \varepsilon), \qquad \mu_\pm = 1 \pm i\sqrt{M^2 - M\varepsilon},$$

the square root being a real number because $\varepsilon < M$. The eigenvalue $1 + Mi$ has moved to $1 + i\sqrt{M^2 - M\varepsilon}$, a distance of

$$M - \sqrt{M^2 - M\varepsilon} = \frac{M\varepsilon}{M + \sqrt{M^2 - M\varepsilon}} \le \frac{M\varepsilon}{M} = \varepsilon,$$

where the middle equality multiplies top and bottom by $M + \sqrt{M^2 - M\varepsilon}$ and the last step uses $\sqrt{M^2 - M\varepsilon} \geq 0$. So here the eigenvalues move by **at most** the size of the perturbation. The same denominator is at most $2M$, so the movement is also at least $\varepsilon/2$: with $M = 100$ and $\varepsilon = 10^{-4}$ it is a shade over $\varepsilon/2$.

The contrast is not an accident of these two matrices, and it is not a theorem we can prove here either. The general statement — that for a diagonalizable $\mathbf{A}$ every eigenvalue of a perturbed $\mathbf{A} + \mathbf{E}$ lies within $\kappa\|\mathbf{E}\|$ of some eigenvalue of $\mathbf{A}$, where κ measures how far from unitary the diagonalizing matrix is, and that $\kappa = 1$ exactly in the normal case — is the Bauer–Fike theorem. It needs matrix norms and a little analysis, and Chapter 20 develops both and proves it. What the example above establishes on its own is modest and exact: of two 2×2 matrices with the same diagonal and the same large off-diagonal entry, the one with defect 0 amplifies a corner perturbation by a factor of at most 1, and the one with defect M amplifies it by $\sqrt{M/\varepsilon}$. The defect is not bookkeeping.

Remark. The amplification $\sqrt{M\varepsilon}$ in Example (A perturbation that moves the eigenvalues a thousand times as far) has a square root in it, and that square root is the signature of a repeated eigenvalue. The unperturbed $\mathbf{A}$ has the single eigenvalue 1 twice over with only one eigenvector, so it is a Jordan block in disguise; splitting a double root is what costs the square root. A matrix with well-separated eigenvalues and a large defect behaves less dramatically.

12.11.5 The characterizations collected

Schur's inequality gives a normality test of a new kind. Every earlier test compared two matrix products or ranged over all vectors; this one compares two numbers, and one of them is a trace.

Theorem 12.11.2. Normality as a Trace Identity

Let $\mathbf{A} \in M_n(\mathbb{C})$ have eigenvalues $\lambda_1, \dots, \lambda_n$, listed with algebraic multiplicity. Then $\mathbf{A}$ is normal if and only if

$$\operatorname{tr}(\mathbf{A}^*\mathbf{A}) = \sum_{i=1}^{n} |\lambda_i|^2.$$

Proof. We showed above that $\operatorname{tr}(\mathbf{A}^*\mathbf{A}) = \|\mathbf{A}\|_F^2$. So the displayed equation is the equality case of Theorem 12.11.1, which holds exactly when $\mathbf{A}$ is normal. ∎

Small as it is, this is the first criterion in the chapter that never asks for $\mathbf{A}\mathbf{A}^*$. If the spectrum is already known — from the characteristic polynomial, or because the matrix arrived in triangular form — normality costs one sum of n^2 squared moduli and one sum of n more.

We can now put every description of normality in one place. Sections 3 and 4 proved most of them; this section adds the last one; and one equivalence, the converse of Corollary 12.4.5, has not been proved anywhere and is proved here.

Theorem 12.11.3. Normality: The Complete List

Let $\mathbf{A} \in M_n(\mathbb{C})$ have eigenvalues $\lambda_1, \dots, \lambda_n$, listed with algebraic multiplicity, and put

$$\mathbf{H} = \tfrac{1}{2}(\mathbf{A} + \mathbf{A}^*), \qquad \mathbf{K} = \tfrac{1}{2i}(\mathbf{A} - \mathbf{A}^*),$$

both of which are Hermitian, with $\mathbf{A} = \mathbf{H} + i\mathbf{K}$. The following are equivalent.

(a) $\mathbf{A}^*\mathbf{A} = \mathbf{A}\mathbf{A}^*$, that is, $\mathbf{A}$ is normal.
(b) $\|\mathbf{A}\mathbf{x}\| = \|\mathbf{A}^*\mathbf{x}\|$ for every $\mathbf{x} \in \mathbb{C}^n$.
(c) $\langle \mathbf{A}\mathbf{x}, \mathbf{A}\mathbf{y} \rangle = \langle \mathbf{A}^*\mathbf{x}, \mathbf{A}^*\mathbf{y} \rangle$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{C}^n$.
(d) $\mathbf{H}\mathbf{K} = \mathbf{K}\mathbf{H}$.
(e) $\mathbb{C}^n$ has an orthonormal basis of eigenvectors of $\mathbf{A}$.
(f) $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{U}^*$ for some unitary $\mathbf{U}$ and some diagonal $\mathbf{D}$.
(g) $\mathbf{A}^* = p(\mathbf{A})$ for some polynomial $p \in \mathbb{C}[x]$.
(h) $\sum_{i=1}^{n} |\lambda_i|^2 = \|\mathbf{A}\|_F^2$.

Proof. Let $T = T_{\mathbf{A}}$ on $\mathbb{C}^n$ with its standard inner product, for which the standard basis is orthonormal; then $T^* = T_{\mathbf{A}^*}$ by Theorem 11.6.3, and T is normal exactly when $\mathbf{A}$ is.

(a), (b), (c), and (d) are equivalent by Theorem 12.3.1 applied to T, whose four conditions are these four read for T instead of $\mathbf{A}$. (a) $\Leftrightarrow$ (e) is Theorem 12.4.2, again for T. (a) $\Leftrightarrow$ (f) is Corollary 12.4.3. (a) $\Leftrightarrow$ (h) is Theorem 12.11.2, since $\|\mathbf{A}\|_F^2 = \operatorname{tr}(\mathbf{A}^*\mathbf{A})$. (a) $\Rightarrow$ (g) is Corollary 12.4.5, applied to T and read back as matrices.

It remains to prove (g) $\Rightarrow$ (a). Suppose $\mathbf{A}^* = p(\mathbf{A})$ with $p = c_0 + c_1 x + \cdots + c_d x^d$. Every power of $\mathbf{A}$ commutes with $\mathbf{A}$, hence so does any linear combination of powers:

$$p(\mathbf{A})\mathbf{A} = \sum_{k=0}^{d} c_k \mathbf{A}^{k+1} = \mathbf{A}\, p(\mathbf{A}).$$

Substituting $\mathbf{A}^*$ for $p(\mathbf{A})$ on both sides gives $\mathbf{A}^*\mathbf{A} = \mathbf{A}\mathbf{A}^*$, which is (a). This closes the loop and proves the theorem. ∎

This list is the summary result of the chapter, and it is worth reading as a whole rather than as eight separate facts. Conditions (a) and (d) are algebraic: one equation between products. Conditions (b) and (c) are metric: $\mathbf{A}$ and $\mathbf{A}^*$ distort lengths and angles identically. Conditions (e) and (f) are geometric: the eigenvectors can be chosen perpendicular. Condition (g) is a surprise, and says that for a normal matrix the adjoint carries no information beyond $\mathbf{A}$ itself. And condition (h) is quantitative: it is the only one on the list that comes with a *number* attached, so that when it fails one can ask by how much. That number is $\Delta(\mathbf{A})$, and it is the reason this section exists.

The operator versions of (a)–(g) hold verbatim on any finite-dimensional complex inner product space, by passing to the matrix in an orthonormal basis. So does (h), once $\|\cdot\|_F$ is read as the Frobenius norm of that matrix, which Lemma 11.7.5 showed does not depend on which orthonormal basis is used.

12.11.6 Exercises

12.11.6.1 A. Check your understanding

Exercise. A1

(a) Define the Frobenius norm of a matrix, and give the expression for $\|\mathbf{A}\|_F^2$ as a trace.

(b) State Theorem 12.11.1, including the equality case and how the eigenvalues are counted.

(c) Define $\Delta(\mathbf{A})$, and say why the definition needs Theorem 12.11.1 before it can be made.

(d) True or false: $\Delta(\mathbf{A})$ equals the Frobenius distance from $\mathbf{A}$ to the nearest normal matrix. Justify your answer.

(e) True or false: if $\mathbf{A}$ and $\mathbf{B}$ are similar then $\Delta(\mathbf{A}) = \Delta(\mathbf{B})$. Justify your answer.

(f) Name the two facts about the Frobenius norm and about triangular matrices that the proof of Theorem 12.11.1 rests on.

Solution.

(a) $\|\mathbf{A}\|_F = \left(\sum_{i,j}|a_{ij}|^2\right)^{1/2}$, the norm induced by the Frobenius inner product $\langle \mathbf{A}, \mathbf{B}\rangle = \operatorname{tr}(\mathbf{B}^*\mathbf{A})$ of Chapter 11. As a trace, $\|\mathbf{A}\|_F^2 = \operatorname{tr}(\mathbf{A}^*\mathbf{A})$.

(b) For $\mathbf{A} \in M_n(\mathbb{C})$ with eigenvalues $\lambda_1, \dots, \lambda_n$ listed **in $\mathbb{C}$ and with algebraic multiplicity**, $\sum_i |\lambda_i|^2 \le \|\mathbf{A}\|_F^2$, with equality if and only if $\mathbf{A}$ is normal.

(c) $\Delta(\mathbf{A})^2 = \|\mathbf{A}\|_F^2 - \sum_i |\lambda_i|^2$, and $\Delta(\mathbf{A})$ is its nonnegative square root. Without Theorem 12.11.1 the right-hand side might be negative, and there would be no real square root to take.

(d) False. It is an upper bound for that distance, as the paragraph before the warning shows, but the warning gives $\mathbf{A} = \mathbf{J}_2(0)$ with $\Delta(\mathbf{A}) = 1$ and a normal matrix at distance $1/\sqrt{2}$.

(e) False. The matrices $\begin{pmatrix} 1 & 1 \\ 0 & 2 \end{pmatrix}$ and $\operatorname{diag}(1, 2)$ are similar, since the first is diagonalizable with eigenvalues 1 and 2, yet their defects are 1 and 0. What is true is that Δ is unchanged by **unitary** similarity; see Exercise (C1: Unitary similarity preserves the defect, ordinary similarity does not).

(f) That the Frobenius norm is unchanged by unitary factors (Lemma 11.7.5), so that $\|\mathbf{A}\|_F = \|\mathbf{T}\|_F$; and that a normal upper triangular matrix is diagonal (Lemma 12.4.1), which is what makes the equality case work in the direction ($\Leftarrow$).

12.11.6.2 B. Practice

Exercise. B1: Computing the defect

For each matrix below, compute $\|\mathbf{A}\|_F^2$, the eigenvalue list, and $\Delta(\mathbf{A})$. State in each case whether the matrix is normal, and say how you know without multiplying $\mathbf{A}$ by $\mathbf{A}^*$.

(a) $\begin{pmatrix} 3 & 4 \\ 0 & 5 \end{pmatrix}$

(b) $\begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}$

(c) $\mathbf{J}_3(2)$

(d) $\begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$

Solution.

(a) $\|\mathbf{A}\|_F^2 = 9 + 16 + 25 = 50$. The matrix is upper triangular, so its eigenvalues are 3 and 5 (Theorem 9.7.4) and $\sum_i |\lambda_i|^2 = 34$. Hence $\Delta(\mathbf{A})^2 = 16$ and $\Delta(\mathbf{A}) = 4$. Since $\Delta(\mathbf{A}) \neq 0$, the matrix is not normal (Theorem 12.11.1).

(b) $\|\mathbf{A}\|_F^2 = 1 + 1 + 1 + 1 = 4$. Here $\operatorname{tr}\mathbf{A} = 2$ and $\det\mathbf{A} = 1 + 1 = 2$, so $p_{\mathbf{A}} = x^2 - 2x + 2$ and the eigenvalues are $1 \pm i$, each of modulus $\sqrt{2}$. Then $\sum_i |\lambda_i|^2 = 2 + 2 = 4$, so $\Delta(\mathbf{A}) = 0$ and the matrix **is** normal.

(c) $\|\mathbf{J}_3(2)\|_F^2 = 3 \cdot 4 + 2 = 14$, the eigenvalue is 2 three times, and $\sum_i |\lambda_i|^2 = 12$. So $\Delta^2 = 2$ and $\Delta = \sqrt{2}$, which is $\sqrt{k-1}$ with $k = 3$, as Example (The defect of a Jordan block) says. Not normal.

(d) $\|\mathbf{A}\|_F^2 = 3$. The matrix is the cyclic shift, with $p_{\mathbf{A}} = x^3 - 1$ (expand along the first row, or note $\mathbf{A}^3 = \mathbf{I}$ and $\mathbf{A} \neq \mathbf{I}$ with $\mathbf{A}$ having no repeated eigenvalue). Its eigenvalues are the three cube roots of unity, each of modulus 1, so $\sum_i |\lambda_i|^2 = 3$ and $\Delta(\mathbf{A}) = 0$: normal.

Exercise. B2: Verifying Schur's inequality

Let

$$\mathbf{A} = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 2 & 1 \\ 0 & 0 & 3 \end{pmatrix}.$$

(a) Compute both sides of Theorem 12.11.1 for $\mathbf{A}$ and verify the inequality.

(b) Hence determine whether $\mathbf{A}$ is normal, without computing $\mathbf{A}^*\mathbf{A}$ or $\mathbf{A}\mathbf{A}^*$.

(c) Write down a normal matrix $\mathbf{N}$ with $\|\mathbf{A} - \mathbf{N}\|_F = \Delta(\mathbf{A})$, and verify the equality.

Solution.

(a) $\|\mathbf{A}\|_F^2 = 1 + 1 + 1 + 4 + 1 + 9 = 17$. The matrix is upper triangular, so its eigenvalues are $1, 2, 3$ (Theorem 9.7.4) and $\sum_i |\lambda_i|^2 = 1 + 4 + 9 = 14$. Indeed $14 \leq 17$.

(b) The inequality is strict, so by the equality case of Theorem 12.11.1 the matrix is not normal. Equivalently $\Delta(\mathbf{A}) = \sqrt{17 - 14} = \sqrt{3} \neq 0$.

(c) Since $\mathbf{A}$ is already upper triangular, $\mathbf{A} = \mathbf{I}\mathbf{A}\mathbf{I}^*$ is a Schur factorization (Corollary 12.1.3) with $\mathbf{U} = \mathbf{I}$ and $\mathbf{T} = \mathbf{A}$. The construction before the warning takes $\mathbf{N}$ to be the diagonal part, $\mathbf{N} = \operatorname{diag}(1, 2, 3)$, which is normal. Then $\mathbf{A} - \mathbf{N}$ has the three entries 1 above the diagonal and zeros elsewhere, so $\|\mathbf{A} - \mathbf{N}\|_F = \sqrt{3} = \Delta(\mathbf{A})$.

Exercise. B3: The nearest symmetric matrix

Let $\mathbf{A} \in M_n(\mathbb{R})$ and write $\mathbf{S} = \frac{1}{2}(\mathbf{A} + \mathbf{A}^\top)$ and $\mathbf{W} = \frac{1}{2}(\mathbf{A} - \mathbf{A}^\top)$, so that $\mathbf{A} = \mathbf{S} + \mathbf{W}$ with $\mathbf{S}^\top = \mathbf{S}$ and $\mathbf{W}^\top = -\mathbf{W}$.

(a) Prove that $\langle \mathbf{B}, \mathbf{C} \rangle = 0$ in the Frobenius inner product whenever $\mathbf{B} \in M_n(\mathbb{R})$ is symmetric and $\mathbf{C} \in M_n(\mathbb{R})$ is skew-symmetric.

(b) Deduce that $\|\mathbf{A} - \mathbf{B}\|_F \geq \|\mathbf{W}\|_F$ for every symmetric $\mathbf{B} \in M_n(\mathbb{R})$, with equality only for $\mathbf{B} = \mathbf{S}$. Hence $\mathbf{S}$ is the nearest symmetric matrix to $\mathbf{A}$.

(c) Carry this out for $\mathbf{A} = \begin{pmatrix} 2 & 4 \\ 0 & -2 \end{pmatrix}$, and compare the resulting distance with $\Delta(\mathbf{A})$.

Solution.

(a) For real matrices $\langle \mathbf{B}, \mathbf{C} \rangle = \operatorname{tr}(\mathbf{C}^\top \mathbf{B})$. Since $\mathbf{C}^\top = -\mathbf{C}$, this is $-\operatorname{tr}(\mathbf{CB})$. On the other hand $\operatorname{tr}(\mathbf{CB}) = \operatorname{tr}((\mathbf{CB})^\top)$ by Theorem 0.9.6 (2), and $(\mathbf{CB})^\top = \mathbf{B}^\top \mathbf{C}^\top = -\mathbf{BC}$, so $\operatorname{tr}(\mathbf{CB}) = -\operatorname{tr}(\mathbf{BC}) = -\operatorname{tr}(\mathbf{CB})$, the last step by Theorem 0.9.6 (3). A number equal to its own negative is 0, so $\operatorname{tr}(\mathbf{CB}) = 0$ and $\langle \mathbf{B}, \mathbf{C} \rangle = 0$.

(b) Let $\mathbf{B}$ be symmetric. Then $\mathbf{A} - \mathbf{B} = (\mathbf{S} - \mathbf{B}) + \mathbf{W}$, where $\mathbf{S} - \mathbf{B}$ is symmetric and $\mathbf{W}$ is skew-symmetric, so the two are orthogonal by (a) and Pythagoras (Theorem 11.1.3) gives

$$\|\mathbf{A} - \mathbf{B}\|_F^2 = \|\mathbf{S} - \mathbf{B}\|_F^2 + \|\mathbf{W}\|_F^2 \geq \|\mathbf{W}\|_F^2.$$

Equality forces $\|\mathbf{S} - \mathbf{B}\|_F = 0$, that is $\mathbf{B} = \mathbf{S}$ (Theorem 11.1.2 (a)). So $\mathbf{S}$ is the unique nearest symmetric matrix, at distance $\|\mathbf{W}\|_F$.

(c) Here $\mathbf{S} = \begin{pmatrix} 2 & 2 \\ 2 & -2 \end{pmatrix}$ and $\mathbf{W} = \begin{pmatrix} 0 & 2 \\ -2 & 0 \end{pmatrix}$, so the distance from $\mathbf{A}$ to the symmetric matrices is $\|\mathbf{W}\|_F = \sqrt{8} = 2\sqrt{2}$. For the defect, $\|\mathbf{A}\|_F^2 = 4 + 16 + 4 = 24$, and $\mathbf{A}$ is triangular with eigenvalues 2 and -2, so $\sum_i |\lambda_i|^2 = 8$ and $\Delta(\mathbf{A}) = 4$. A symmetric matrix is normal, so this again exhibits a normal matrix strictly closer than $\Delta(\mathbf{A})$: $2\sqrt{2} \approx 2.83 < 4$.

12.11.6.3 C. Going deeper

Exercise. C1: Unitary similarity preserves the defect, ordinary similarity does not

(a) Prove that $\Delta(\mathbf{U}^* \mathbf{A} \mathbf{U}) = \Delta(\mathbf{A})$ for every $\mathbf{A} \in M_n(\mathbb{C})$ and every unitary $\mathbf{U}$.

(b) Give $\mathbf{A}, \mathbf{B} \in M_2(\mathbb{C})$ that are similar with $\Delta(\mathbf{A}) \neq \Delta(\mathbf{B})$, exhibiting an explicit $\mathbf{P}$ with $\mathbf{P}^{-1} \mathbf{A} \mathbf{P} = \mathbf{B}$.

(c) Deduce that two similar matrices with different defects cannot be unitarily similar, and use this to show that $\begin{pmatrix} 1 & 1 \\ 0 & 2 \end{pmatrix}$ and $\operatorname{diag}(1, 2)$ are not unitarily similar.

Solution.

(a) Put $\mathbf{B} = \mathbf{U}^* \mathbf{A} \mathbf{U}$. By Lemma 11.7.5, $\|\mathbf{B}\|_F = \|\mathbf{A}\|_F$. Also $\mathbf{B}$ is similar to $\mathbf{A}$, so $p_{\mathbf{B}} = p_{\mathbf{A}}$ (Theorem 7.10.3) and the two matrices have the same eigenvalue list with multiplicities. Both terms in Definition 12.11.1 therefore agree, so $\Delta(\mathbf{B})^2 = \Delta(\mathbf{A})^2$, and the defects are equal since both are nonnegative.

(b) Take $\mathbf{A} = \begin{pmatrix} 1 & 1 \\ 0 & 2 \end{pmatrix}$ and $\mathbf{B} = \operatorname{diag}(1,2)$. Then $\|\mathbf{A}\|_F^2 = 1+1+4 = 6$ with eigenvalues $1, 2$, so $\Delta(\mathbf{A})^2 = 6 - 5 = 1$ and $\Delta(\mathbf{A}) = 1$, while $\Delta(\mathbf{B}) = 0$ because $\mathbf{B}$ is diagonal, hence normal. They are similar: eigenvectors of $\mathbf{A}$ are $(1,0)$ for $\lambda = 1$ and $(1,1)$ for $\lambda = 2$, so with $\mathbf{P} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$, whose inverse is $\begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix}$, we get $\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \mathbf{B}$ by direct multiplication.

(c) By (a), unitarily similar matrices have equal defects. So if $\Delta(\mathbf{A}) \neq \Delta(\mathbf{B})$ then no unitary $\mathbf{U}$ can satisfy $\mathbf{U}^*\mathbf{A}\mathbf{U} = \mathbf{B}$. The pair in (b) has defects 1 and 0, so it is similar but not unitarily similar. (This is the failure of "diagonalizable" to imply "unitarily diagonalizable", seen through a number instead of through Example (Diagonalizable, and no orthonormal eigenbasis).)

Exercise. C2: Equality plus a single eigenvalue forces a scalar

Let $\mathbf{A} \in M_n(\mathbb{C})$ satisfy $\|\mathbf{A}\|_F^2 = \sum_{i=1}^n |\lambda_i|^2$, and suppose $\mathbf{A}$ has **only one** distinct eigenvalue λ, so that $p_{\mathbf{A}} = (x-\lambda)^n$.

(a) Prove that $\mathbf{A} = \lambda\mathbf{I}$.

(b) Deduce that a normal matrix whose characteristic polynomial is x^n is the zero matrix.

(c) Show that the hypothesis of equality cannot be dropped, by computing Δ for a matrix with $p_{\mathbf{A}} = (x-\lambda)^n$ that is not $\lambda\mathbf{I}$.

Solution.

(a) By the equality case of Theorem 12.11.1, $\mathbf{A}$ is normal. So by Corollary 12.4.3 there are a unitary $\mathbf{U}$ and a diagonal $\mathbf{D}$ with $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{U}^*$, the diagonal of $\mathbf{D}$ being the eigenvalue list of $\mathbf{A}$. That list is λ repeated n times, so $\mathbf{D} = \lambda\mathbf{I}$ and

$$\mathbf{A} = \mathbf{U}(\lambda\mathbf{I})\mathbf{U}^* = \lambda\mathbf{U}\mathbf{U}^* = \lambda\mathbf{I}.$$

(b) A normal matrix satisfies the equality hypothesis (Theorem 12.11.1 again), and $p_{\mathbf{A}} = x^n$ says its only eigenvalue is 0. By (a), $\mathbf{A} = 0 \cdot \mathbf{I} = \mathbf{O}$. In words: a normal nilpotent matrix is zero, which is Theorem 12.3.5 seen from the quantitative side.

(c) Take $\mathbf{A} = \mathbf{J}_n(\lambda)$ with $n \geq 2$. Its characteristic polynomial is $(x-\lambda)^n$ and it is not $\lambda\mathbf{I}$, and $\Delta(\mathbf{A}) = \sqrt{n-1} \neq 0$ by Example (The defect of a Jordan block). So equality in Theorem 12.11.1 fails, exactly as it must.

Exercise. C3: The commutator is at least as big as the defect squared

For $\mathbf{A} \in M_n(\mathbb{C})$ write $[\mathbf{A}, \mathbf{A}^*] = \mathbf{A}\mathbf{A}^* - \mathbf{A}^*\mathbf{A}$.

(a) Prove that $\|[\mathbf{A}, \mathbf{A}^*]\|_F$ is unchanged when $\mathbf{A}$ is replaced by $\mathbf{U}^*\mathbf{A}\mathbf{U}$ with $\mathbf{U}$ unitary.

(b) Let $\mathbf{T} = \begin{pmatrix} \lambda_1 & b \\ 0 & \lambda_2 \end{pmatrix}$. Compute $[\mathbf{T}, \mathbf{T}^*]$ and show that

$$\|[\mathbf{T}, \mathbf{T}^*]\|_F^2 = 2|b|^4 + 2|b|^2|\lambda_1 - \lambda_2|^2.$$

(c) Deduce that every $\mathbf{A} \in M_2(\mathbb{C})$ satisfies $\|[\mathbf{A}, \mathbf{A}^*]\|_F \geq \sqrt{2}\,\Delta(\mathbf{A})^2$, and determine exactly when equality holds.

Hint for (c): Schur-triangularize, and use (a) to move the computation to the triangular factor.

Solution.

(a) Put $\mathbf{B} = \mathbf{U}^*\mathbf{A}\mathbf{U}$, so $\mathbf{B}^* = \mathbf{U}^*\mathbf{A}^*\mathbf{U}$. Inserting $\mathbf{U}\mathbf{U}^* = \mathbf{I}$ between the factors,

$$[\mathbf{B}, \mathbf{B}^*] = \mathbf{U}^*(\mathbf{A}\mathbf{A}^* - \mathbf{A}^*\mathbf{A})\mathbf{U} = \mathbf{U}^*[\mathbf{A}, \mathbf{A}^*]\mathbf{U},$$

and Lemma 11.7.5 gives $\|[\mathbf{B}, \mathbf{B}^*]\|_F = \|[\mathbf{A}, \mathbf{A}^*]\|_F$.

(b) Multiplying out,

$$\mathbf{T}\mathbf{T}^* = \begin{pmatrix} |\lambda_1|^2 + |b|^2 & b\overline{\lambda_2} \\ \lambda_2\overline{b} & |\lambda_2|^2 \end{pmatrix},$$

$$\mathbf{T}^*\mathbf{T} = \begin{pmatrix} |\lambda_1|^2 & \overline{\lambda_1}b \\ \lambda_1\overline{b} & |b|^2 + |\lambda_2|^2 \end{pmatrix}.$$

Subtracting,

$$[\mathbf{T}, \mathbf{T}^*] = \begin{pmatrix} |b|^2 & b\,\overline{(\lambda_2 - \lambda_1)} \\ \overline{b}\,(\lambda_2 - \lambda_1) & -|b|^2 \end{pmatrix}.$$

Its four squared moduli are $|b|^4$, $|b|^2|\lambda_1 - \lambda_2|^2$, $|b|^2|\lambda_1 - \lambda_2|^2$, and $|b|^4$. Adding them gives the stated formula.

(c) Let $\mathbf{A} \in M_2(\mathbb{C})$ and write $\mathbf{A} = \mathbf{U}\mathbf{T}\mathbf{U}^*$ as in Corollary 12.1.3, with $\mathbf{T}$ upper triangular of the shape in (b) and λ_1, λ_2 the eigenvalues of $\mathbf{A}$. By Eq. 12.11.1, $\Delta(\mathbf{A})^2 = |b|^2$. By (a) and (b),

$$\|[\mathbf{A}, \mathbf{A}^*]\|_F^2 = 2|b|^4 + 2|b|^2|\lambda_1 - \lambda_2|^2 \geq 2|b|^4 = 2\,\Delta(\mathbf{A})^4,$$

and taking nonnegative square roots gives $\|[\mathbf{A}, \mathbf{A}^*]\|_F \geq \sqrt{2}\,\Delta(\mathbf{A})^2$.

Equality holds exactly when $|b|^2|\lambda_1 - \lambda_2|^2 = 0$, that is, when $b = 0$ or $\lambda_1 = \lambda_2$. The first case says $\Delta(\mathbf{A}) = 0$, that is, $\mathbf{A}$ is normal, and then both sides are 0. So equality holds if and only if $\mathbf{A}$ is normal or its two eigenvalues coincide.

Chapter 13

Positive Matrices and the Singular Value Decomposition

A matrix that is not square has no eigenvalues. There is no equation $\mathbf{Av} = \lambda\mathbf{v}$ to write down, because $\mathbf{Av}$ and $\mathbf{v}$ live in different spaces, and everything Chapters 9 to 12 built on that equation is unavailable. Yet the geometric picture that opened Chapter 12 survives untouched: push the unit sphere of F^n through any $\mathbf{A} \in M_{m\times n}(F)$, square or not, and what comes out is an ellipsoid. It has semi-axes, they are perpendicular, and the directions they came from were perpendicular too. Reading that picture as a factorization gives $\mathbf{A} = \mathbf{U\Sigma V}^*$, the **singular value decomposition**: every matrix whatsoever maps some orthonormal basis to an orthogonal set. It is the most used factorization in applied linear algebra, and this chapter's business is to earn it and then spend it.

The road runs through positivity. A nonsquare $\mathbf{A}$ has no eigenvalues, but $\mathbf{A}^*\mathbf{A}$ does, and $\mathbf{A}^*\mathbf{A}$ is self-adjoint with a nonnegative quadratic form — it is a **positive semidefinite** matrix, the analogue of a nonnegative number. Chapter 12 left that class unexamined; here it comes first, with its many equivalent tests, its square root, and the triangular factorization that elimination finds. From the square root comes the absolute value $|\mathbf{A}|$, from $|\mathbf{A}|$ the polar decomposition $\mathbf{A} = \mathbf{W}|\mathbf{A}|$, and from the spectral theorem applied to $\mathbf{A}^*\mathbf{A}$ the singular value decomposition itself. Once it is in hand, a series of questions that look unrelated turn out to have one answer: what are orthonormal bases of the four fundamental subspaces, what is the best rank-k approximation of a matrix, what should $\mathbf{A}^{-1}$ mean when $\mathbf{A}$ is not invertible, and when is a projection orthogonal. **The field is $\mathbb{R}$ or $\mathbb{C}$ throughout**, as in Chapters 11 and 12, and every statement says which one it needs.

What you need. From Chapter 12, the two spectral theorems and their matrix forms (Theorem 12.4.2, Corollary 12.4.3, Theorem 12.5.1, Corollary 12.5.2, Corollary 12.5.4, Theorem 12.5.5), the elementary theory of self-adjoint operators (Theorem 12.2.1, Theorem 12.2.3, Theorem 12.2.6, Exercise (C1: The operator S^*S)), commuting families (Corollary 12.7.2, Corollary 12.7.3), and the functional calculus with its square root (Definition 12.8.1, Theorem 12.8.1, Theorem 12.8.3, Theorem 12.8.4). From Chapter 12 §10, the Sylvester equation and the Lyapunov equation it specializes to (Theorem 12.10.2, Definition 12.10.2), used in §06. From Chapter 9, the eigenvalue form of the Invertible Matrix Theorem, which §§01 and 06 use to turn a positivity question into a statement about the spectrum (Theorem 9.2.6), and the trace and determinant as sum and product of the eigenvalues (Theorem 9.3.3). From Chapter 11, the Frobenius norm and its unitary invariance (Lemma 11.7.5), the adjoint and what it does to the four subspaces (Definition 11.6.1, Theorem 11.6.2, Theorem 11.6.3, Theorem 11.6.5, Corollary 11.6.6), orthogonal decompositions and projections (Theorem 11.3.2,

Definition 11.3.2, Theorem 11.3.3, Theorem 11.3.5), least squares (Theorem 11.4.2, Corollary 11.4.4, Theorem 11.4.5, Lemma 11.4.3), Gram matrices (Definition 11.2.3, Theorem 11.2.5), isometries (Definition 11.7.1, Theorem 11.7.3, Definition 11.7.2), and the QR factorization (Theorem 11.8.1). From Chapter 8, block multiplication and Schur complements (Theorem 8.1.1, Definition 8.3.1, Theorem 8.3.1, Theorem 8.3.2, Theorem 8.3.3, Theorem 8.3.4). From Chapter 7, multiplicativity of the determinant, triangular determinants, and the nonnegativity of Gram determinants (Theorem 7.5.4, Theorem 7.4.6, Corollary 7.7.3). From Chapter 6, Lagrange interpolation (Theorem 6.6.1). From Chapter 4, projections along a complement (Definition 4.5.1, Theorem 4.5.1). From Chapter 3, the $\text{Ł}\mathbf{D}\text{Ł}^\top$ factorization and the invertible matrix theorem it started (Theorem 3.6.8, Theorem 3.4.4). From Chapter 1, one fact, in §13: a subspace of the same finite dimension as a space containing it is the whole of it (Theorem 1.6.8).

13.0.1 Roadmap

- **Positive semidefinite and positive definite matrices.** The definition, the real-versus-complex asymmetry that trips everyone, and the full list of equivalent tests, ending in Sylvester's criterion and the trap that comes with it.
- **The square root and Cholesky.** The unique positive square root, why it is a polynomial in the matrix, the absolute value $|\mathbf{A}|$, and the factorization $\mathbf{A} = \text{Ł}\text{Ł}^*$ whose algorithm is itself the test for definiteness.
- **Gram matrices.** A matrix is a Gram matrix exactly when it is positive semidefinite; the Gram determinant as a squared volume, and distance to a subspace computed two ways.
- **Two forms at once.** A positive definite form and a self-adjoint one can be diagonalized simultaneously — by congruence, not by similarity — and the generalized eigenvalue problem $\mathbf{Bx} = \lambda\mathbf{Ax}$ follows.
- **The Loewner order.** What $\mathbf{A} \succeq \mathbf{B}$ means, what monotonicity survives it (traces, determinants, inverses), and the chapter's negative result: squaring does not.
- **Blocks and Schur complements.** Testing a block matrix for positive definiteness one block at a time, and the positive definite solution of the Lyapunov equation, which the same test delivers: a certificate of stability that computes no eigenvalue.
- **Determinant inequalities.** What the same block decomposition says about the size of a determinant: Fischer's inequality for two diagonal blocks, and Hadamard's for the diagonal entries and for the lengths of the columns.
- **Products entry by entry.** The Schur product theorem — the Hadamard product of positive matrices is positive — and Oppenheim's inequality.
- **The singular value decomposition.** Every matrix is $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$; the singular values are determined, the unitaries are not, and the compact form reads off orthonormal bases of all four fundamental subspaces.
- **Polar decomposition and Takagi.** $\mathbf{A} = \mathbf{W}|\mathbf{A}|$ is $z = e^{i\theta}r$ for matrices; $\mathbf{W}$ is the nearest unitary matrix to $\mathbf{A}$; and a complex symmetric matrix factors as $\mathbf{U}\mathbf{\Sigma}\mathbf{U}^\top$.
- **Low-rank approximation.** Truncating the singular value sum gives the best approximation of each rank, in the Frobenius norm, with an error one can write down.
- **The pseudoinverse.** Inverting the nonzero singular values defines $\mathbf{A}^+$; it is pinned down by four conditions, and $\mathbf{A}^+\mathbf{b}$ is the minimum-norm least-squares solution promised in Chapter 11.
- **Projections revisited.** An idempotent is an orthogonal projection exactly when it is self-adjoint; everything else is oblique, and the obliqueness is a measurable angle.

13.0.2 Named moves

- **Test with $\langle \mathbf{Ax}, \mathbf{x} \rangle$.** Positivity is a statement about one scalar function of one vector. Most proofs in the first half of the chapter consist of evaluating that function at a cleverly chosen $\mathbf{x}$ — a standard basis vector, a vector supported on a few coordinates, an eigenvector, or $\mathbf{Sy}$ for a substitution $\mathbf{S}$.
- **Factor as $\mathbf{B}^*\mathbf{B}$.** Every positive semidefinite matrix is one, and that form turns a statement about the matrix into a statement about lengths: $\langle \mathbf{B}^*\mathbf{Bx}, \mathbf{x} \rangle = \|\mathbf{Bx}\|^2$. It is also what makes every positive semidefinite matrix a Gram matrix.
- **Diagonalize, do it to the eigenvalues, come back.** The square root, the pseudoinverse, and the whole functional calculus are this one move. The work is always in checking that the result does not depend on the diagonalization chosen.
- **Apply the spectral theorem to $\mathbf{A}^*\mathbf{A}$.** When the matrix in front of you is not square, not normal, or not even an operator, this is the way in. The singular value decomposition is nothing but the spectral theorem for $\mathbf{A}^*\mathbf{A}$, read back through $\mathbf{A}$.
- **Truncate the sum.** Writing $\mathbf{A} = \sum_i \sigma_i \mathbf{u}_i \mathbf{v}_i^*$ turns a matrix into a sum of rank-one pieces ordered by importance. Cutting the sum short gives the compact form, the best low-rank approximation, and the pseudoinverse, three times from one idea.

13.1 Positive Semidefinite and Positive Definite Matrices

Chapter 12 sorted the operators that have an orthonormal eigenbasis. Among them the self-adjoint ones are the matrix analogue of the real numbers, since they are exactly the operators fixed by starring. Inside the real numbers sits a smaller and even more useful set, the nonnegative ones, and this section finds its matrix analogue. The answer turns out to have three faces — a sign condition on the spectrum, a sign condition on a scalar function, and a factorization — and the whole of the rest of the chapter is built on the fact that the three agree.

Throughout, $F = \mathbb{R}$ or $F = \mathbb{C}$, and F^n carries the standard inner product $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{y}^*\mathbf{x}$.

13.1.1 Which self-adjoint matrices deserve to be called nonnegative?

A real number t is nonnegative in three equivalent ways: $t \geq 0$; $tx^2 \geq 0$ for every real x; and $t = s^2$ for some real s. Each of the three has a matrix reading. "Every eigenvalue is ≥ 0" copies the first. "$\langle \mathbf{Ax}, \mathbf{x} \rangle \geq 0$ for every $\mathbf{x}$" copies the second, because $\langle \mathbf{Ax}, \mathbf{x} \rangle$ is the one scalar a matrix attaches to a single vector. "$\mathbf{A} = \mathbf{B}^*\mathbf{B}$" copies the third, with the star doing what nothing does for numbers but everything does for matrices.

It is not obvious that the three agree, and for a general matrix they do not: they all fail for $\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$ in different ways. What makes them agree is self-adjointness, and the middle condition is the one to take as the definition, because it is the one that says something about every vector at once.

A matrix is positive semidefinite when its quadratic form never goes negative.

Definition 13.1.1. Positive Semidefinite and Positive Definite

Let $\mathbf{A} \in M_n(F)$. Then $\mathbf{A}$ is **positive semidefinite**, written $\mathbf{A} \succeq \mathbf{O}$, if

(P1) $\mathbf{A}$ is **Hermitian**, that is $\mathbf{A}^* = \mathbf{A}$ (over $\mathbb{R}$: symmetric, $\mathbf{A}^\top = \mathbf{A}$); and
(P2) $\langle \mathbf{Ax}, \mathbf{x} \rangle \geq 0$ for **every** $\mathbf{x} \in F^n$.

It is **positive definite**, written $\mathbf{A} \succ \mathbf{O}$, if (P1) holds and $\langle \mathbf{Ax}, \mathbf{x} \rangle > 0$ for every **nonzero** $\mathbf{x} \in F^n$.

The same words apply to an operator $T \in \mathcal{L}(V)$ on a finite-dimensional inner product space: T is positive semidefinite if $T^* = T$ and $\langle Tv, v \rangle \geq 0$ for every $v \in V$, and positive definite if the inequality is strict for $v \neq 0$.

In words: (P1) puts $\mathbf{A}$ in the class where the analogy with real numbers is available at all, and (P2) says that the number $\langle \mathbf{Ax}, \mathbf{x} \rangle = \mathbf{x}^* \mathbf{Ax}$ is never negative. The word **every** in (P2) is the whole content; a single vector proves nothing. In the definite case the vector $\mathbf{x} = \mathbf{0}$ has to be excluded, since $\langle \mathbf{A0}, \mathbf{0} \rangle = 0$ for every $\mathbf{A}$ whatsoever.

Clause (P1) is also what makes clause (P2) meaningful. For a Hermitian $\mathbf{A}$, the number $\langle \mathbf{Ax}, \mathbf{x} \rangle$ is **real** for every $\mathbf{x}$, over $\mathbb{C}$ as well as over $\mathbb{R}$ (Proposition 11.6.7), so comparing it with 0 is legitimate. Without (P1) the number can be a genuine complex number, and "≥ 0" would be meaningless.

Positive definite implies positive semidefinite, since a condition that holds strictly for $\mathbf{x} \neq \mathbf{0}$ holds weakly for $\mathbf{x} = \mathbf{0}$ too. The reverse fails, and the smallest witness is $\operatorname{diag}(1, 0)$.

Before any examples, the trap.

Warning. **Over $\mathbb{R}$, clause (P1) has to be assumed; over $\mathbb{C}$ it comes free.** Let $\mathbf{A} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$, the quarter turn of $\mathbb{R}^2$. Then $\langle \mathbf{Ax}, \mathbf{x} \rangle = -x_2 x_1 + x_1 x_2 = 0$ for every $\mathbf{x} \in \mathbb{R}^2$, so (P2) holds in the strongest possible way, and yet $\mathbf{A}^\top = -\mathbf{A} \neq \mathbf{A}$. A real matrix can pass the quadratic-form test and be nowhere near symmetric, because over $\mathbb{R}$ the quadratic form cannot see the skew part of a matrix at all (Exercise (C3: What the real quadratic form does determine)). So a real definition that dropped (P1) would call the quarter turn positive semidefinite, and none of the theorems below would survive.

Over $\mathbb{C}$ the situation is the opposite, and it is worth recording once.

Proposition 13.1.1. Over $\mathbb{C}$, positivity implies Hermitian

Let $\mathbf{A} \in M_n(\mathbb{C})$ satisfy $\langle \mathbf{Ax}, \mathbf{x} \rangle \in \mathbb{R}$ for every $\mathbf{x} \in \mathbb{C}^n$ — in particular, let $\langle \mathbf{Ax}, \mathbf{x} \rangle \geq 0$ for every $\mathbf{x}$. Then $\mathbf{A}^* = \mathbf{A}$.

Proof. Put $\mathbf{S} = \mathbf{A} - \mathbf{A}^*$. For every $\mathbf{x}$, moving the star across (Lemma 11.4.1) and using conjugate symmetry,

$$\langle \mathbf{A}^* \mathbf{x}, \mathbf{x} \rangle = \langle \mathbf{x}, \mathbf{Ax} \rangle = \overline{\langle \mathbf{Ax}, \mathbf{x} \rangle} = \langle \mathbf{Ax}, \mathbf{x} \rangle,$$

the last equality because the number is real. Hence $\langle \mathbf{S}\mathbf{x}, \mathbf{x}\rangle = 0$ for every $\mathbf{x} \in \mathbb{C}^n$, and Theorem 12.2.6 gives $\mathbf{S} = \mathbf{O}$.

■

So over $\mathbb{C}$ one may state the definition with (P2) alone; over $\mathbb{R}$ one may not. We keep (P1) in the definition for both fields, so that a single sentence covers them and so that no reader has to remember which field licenses which shortcut.

Now the examples, all checked against (P1) and (P2) in turn.

Example. Five Positive Matrices

Decide which of the following are positive semidefinite, and which are positive definite.

(a) $\mathbf{I}_n$, and more generally $\mathbf{D} = \operatorname{diag}(d_1, \dots, d_n)$ with every d_i real.

(b) $\mathbf{x}\mathbf{x}^*$ for a fixed $\mathbf{x} \in F^n$.

(c) The Gram matrix $\mathbf{G}$ of a list $(v_1, \dots, v_k)$ in an inner product space.

(d) The zero matrix $\mathbf{O} \in M_n(F)$.

(e) $\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$.

Solution.

(a) $\mathbf{D}^* = \mathbf{D}$ because the d_i are real, so (P1) holds. For (P2),

$$\langle \mathbf{D}\mathbf{x}, \mathbf{x}\rangle = \mathbf{x}^*\mathbf{D}\mathbf{x} = \sum_{i=1}^{n} d_i |x_i|^2.$$

So $\mathbf{D} \succeq \mathbf{O}$ exactly when every $d_i \geq 0$, and $\mathbf{D} \succ \mathbf{O}$ exactly when every $d_i > 0$: if some $d_i < 0$, the vector $\mathbf{e}_i$ refutes (P2), and if some $d_i = 0$, the vector $\mathbf{e}_i$ refutes definiteness. In particular $\mathbf{I}_n \succ \mathbf{O}$, with $\langle \mathbf{x}, \mathbf{x}\rangle = \|\mathbf{x}\|^2$.

(b) $(\mathbf{x}\mathbf{x}^*)^* = \mathbf{x}^{**}\mathbf{x}^* = \mathbf{x}\mathbf{x}^*$, so (P1) holds. For (P2), using $\langle \mathbf{u}, \mathbf{w}\rangle = \mathbf{w}^*\mathbf{u}$ twice,

$$\mathbf{y}^*(\mathbf{x}\mathbf{x}^*)\mathbf{y} = (\mathbf{y}^*\mathbf{x})(\mathbf{x}^*\mathbf{y}) = \langle \mathbf{x}, \mathbf{y}\rangle \, \overline{\langle \mathbf{x}, \mathbf{y}\rangle} = |\langle \mathbf{x}, \mathbf{y}\rangle|^2 \geq 0.$$

So $\mathbf{x}\mathbf{x}^* \succeq \mathbf{O}$. It is never positive definite for $n \geq 2$, since any $\mathbf{y} \neq \mathbf{0}$ orthogonal to $\mathbf{x}$ makes the value 0; and for $\mathbf{x} = \mathbf{0}$ it is the zero matrix.

(c) $\mathbf{G}^* = \mathbf{G}$ by Theorem 11.2.5 (a), and $\mathbf{x}^*\mathbf{G}\mathbf{x} = \|\sum_j x_j v_j\|^2 \geq 0$ by part (b) of the same theorem. So $\mathbf{G} \succeq \mathbf{O}$ always, and by (c) of that theorem $\mathbf{G} \succ \mathbf{O}$ exactly when the list is linearly independent.

(d) $\mathbf{O}^* = \mathbf{O}$ and $\langle \mathbf{O}\mathbf{x}, \mathbf{x}\rangle = 0 \geq 0$, so $\mathbf{O} \succeq \mathbf{O}$. This is the degenerate case, and it matters: it is the bottom of the order introduced in Section 5, and it shows that (P2) can hold with equality everywhere.

(e) Symmetric, so (P1) holds; but $\langle \mathbf{A}\mathbf{e}_2, \mathbf{e}_2\rangle = -1 < 0$. Not positive semidefinite. This is (a) with $d_2 = -1$.

Example (e) is the nonexample by minimal change: keep (P1), change one diagonal entry, and

(P2) fails at one standard basis vector. The quarter turn of the warning is the other kind of nonexample, keeping (P2) and losing (P1).

Two remarks on the name and the convention. "Definite" is the older word, from the theory of quadratic forms: the form $q(\mathbf{x}) = \mathbf{x}^*\mathbf{A}\mathbf{x}$ has a *definite* sign. "Semi" weakens the sign to allow zeros. Some authors write "positive" for what we call positive semidefinite and "strictly positive" for positive definite; the words differ but the two classes are the same two classes everywhere.

Warning. **A positive matrix is not a matrix with positive entries.** The all-ones matrix $\mathbf{J} \in M_2(\mathbb{R})$ has every entry positive and is positive semidefinite, but $\begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}$ has every entry positive and is **not**: at $\mathbf{x} = (1, -1)$ its form is $1 - 2 - 2 + 1 = -2$. Conversely $\operatorname{diag}(1, 1)$ has a zero entry and is positive definite. Entrywise positivity is a different subject, the one Chapter 9 met through stochastic matrices, and it shares only a word with this one.

13.1.2 Positivity survives a change of variable

Before the equivalences, one small result that will be used in almost every proof of this chapter. The quadratic form $\mathbf{x} \mapsto \mathbf{x}^*\mathbf{A}\mathbf{x}$ is attached to $\mathbf{A}$; substituting $\mathbf{x} = \mathbf{S}\mathbf{y}$ turns it into the quadratic form of $\mathbf{S}^*\mathbf{A}\mathbf{S}$. The operation $\mathbf{A} \mapsto \mathbf{S}^*\mathbf{A}\mathbf{S}$ is called **congruence**, and Chapter 14 studies it for its own sake. Here we need only that it preserves the sign.

Proposition 13.1.2. Congruence preserves positivity

Let $\mathbf{A} \in M_n(F)$ be Hermitian and $\mathbf{S} \in M_{n\times m}(F)$. Then $\mathbf{S}^*\mathbf{A}\mathbf{S} \in M_m(F)$ is Hermitian, and:

(a) if $\mathbf{A} \succeq \mathbf{O}$, then $\mathbf{S}^*\mathbf{A}\mathbf{S} \succeq \mathbf{O}$;

(b) if $\mathbf{A} \succ \mathbf{O}$ and $\operatorname{nul}(\mathbf{S}) = \{\mathbf{0}\}$, then $\mathbf{S}^*\mathbf{A}\mathbf{S} \succ \mathbf{O}$. In particular, for $m = n$ and $\mathbf{S}$ invertible, $\mathbf{A} \succ \mathbf{O}$ if and only if $\mathbf{S}^*\mathbf{A}\mathbf{S} \succ \mathbf{O}$.

Proof. $(\mathbf{S}^*\mathbf{A}\mathbf{S})^* = \mathbf{S}^*\mathbf{A}^*\mathbf{S}^{**} = \mathbf{S}^*\mathbf{A}\mathbf{S}$, so the matrix is Hermitian. For every $\mathbf{y} \in F^m$,

$$\mathbf{y}^*(\mathbf{S}^*\mathbf{A}\mathbf{S})\mathbf{y} = (\mathbf{S}\mathbf{y})^*\mathbf{A}(\mathbf{S}\mathbf{y}) = \langle \mathbf{A}(\mathbf{S}\mathbf{y}), \mathbf{S}\mathbf{y} \rangle. \tag{$*$}$$

(a) If $\mathbf{A} \succeq \mathbf{O}$, the right-hand side of $(*)$ is ≥ 0 for every $\mathbf{y}$.

(b) If $\mathbf{A} \succ \mathbf{O}$ and $\operatorname{nul}(\mathbf{S}) = \{\mathbf{0}\}$, then $\mathbf{y} \neq \mathbf{0}$ forces $\mathbf{S}\mathbf{y} \neq \mathbf{0}$, so the right-hand side of $(*)$ is > 0. For the last claim, take $\mathbf{S}$ invertible: one direction is what we just proved, and the other is the same statement applied to $\mathbf{S}^{-1}$ and the matrix $\mathbf{S}^*\mathbf{A}\mathbf{S}$, since $(\mathbf{S}^{-1})^*(\mathbf{S}^*\mathbf{A}\mathbf{S})\mathbf{S}^{-1} = \mathbf{A}$.

■

13.1.3 The equivalent descriptions

The characterization needs one determinant identity, which is worth stating on its own because it also describes every coefficient of a characteristic polynomial. Recall from Definition 7.7.1 that $\mathbf{A}_{I,I}$ denotes the submatrix of $\mathbf{A}$ on the rows and columns indexed by I; its determinant is a **principal minor** of $\mathbf{A}$, and $\mathbf{A}_{I,I}$ is a **principal submatrix**.

Lemma 13.1.3. Shifting by $t\mathbf{I}$, minor by minor

Let $\mathbf{A} \in M_n(F)$ and $t \in F$. Then

$$\det(\mathbf{A} + t\mathbf{I}_n) = \sum_{k=0}^{n} E_k(\mathbf{A})\, t^{\,n-k},$$

where $E_0(\mathbf{A}) = 1$ and, for $1 \le k \le n$, $E_k(\mathbf{A}) = \sum_{|I|=k} \det \mathbf{A}_{I,I}$ is the sum of all $k \times k$ principal minors of $\mathbf{A}$.

Idea. Column j of $\mathbf{A} + t\mathbf{I}$ is $\mathbf{a}_j + t\mathbf{e}_j$, a sum of two columns, and the determinant is linear in each column separately. Expanding all n columns turns one determinant into 2^n of them, one for each choice of which columns contribute their $t\mathbf{e}_j$. A determinant with $\mathbf{e}_j$ in column j collapses under cofactor expansion to the determinant with row j and column j deleted, and doing that for every chosen column leaves exactly a principal submatrix.

Proof. Write $\mathbf{a}_1, \dots, \mathbf{a}_n$ for the columns of $\mathbf{A}$, so that column j of $\mathbf{A} + t\mathbf{I}$ is $\mathbf{a}_j + t\mathbf{e}_j$. The determinant is linear in each column with the others fixed (Theorem 7.4.3). Applying that in column 1, then in column 2, and so on through column n,

$$\det(\mathbf{A} + t\mathbf{I}) = \sum_{S \subseteq \{1,\dots,n\}} t^{|S|} \det \mathbf{M}_S,$$

where $\mathbf{M}_S$ is the matrix whose j-th column is $\mathbf{e}_j$ for $j \in S$ and $\mathbf{a}_j$ for $j \notin S$.

Fix S and let $j \in S$. Expanding $\det \mathbf{M}_S$ along column j (Theorem 7.6.1 (a)) leaves one term, since that column is $\mathbf{e}_j$: the entry is 1 in row j and 0 elsewhere, and its cofactor is $(-1)^{j+j} M_{jj} = M_{jj}$. So $\det \mathbf{M}_S$ equals the determinant of $\mathbf{M}_S$ with row j and column j deleted, which is a matrix of the same shape on the index set $\{1, \dots, n\} \setminus \{j\}$. Repeating this for each $j \in S$ deletes every row and column indexed by S and leaves

$$\det \mathbf{M}_S = \det \mathbf{A}_{S^c, S^c}, \qquad S^c = \{1, \dots, n\} \setminus S.$$

Collecting the subsets by $k = |S^c|$, so that $|S| = n - k$, gives the stated formula.

■

Remark. Running the same expansion over $F[x]$ with the indeterminate $-x$ in place of t, and multiplying by $(-1)^n$, turns this into $p_{\mathbf{A}}(x) = \sum_k (-1)^k E_k(\mathbf{A}) x^{n-k}$: the coefficients of the characteristic polynomial are, up to sign, the sums of principal minors. The cases $k = 1$ and $k = n$ recover $\operatorname{tr} \mathbf{A}$ and $\det \mathbf{A}$ from Theorem 7.10.2.

Here is the main theorem of the section.

Theorem 13.1.4. Characterizations of Positive Semidefiniteness

Let $\mathbf{A} \in M_n(F)$ be Hermitian. The following are equivalent.

(a) $\mathbf{A} \succeq \mathbf{O}$.
(b) Every eigenvalue of $\mathbf{A}$ is a nonnegative real number.
(c) $\mathbf{A} = \mathbf{B}^*\mathbf{B}$ for some $\mathbf{B} \in M_n(F)$.
(d) $\mathbf{A} = \mathbf{C}^2$ for exactly one $\mathbf{C} \in M_n(F)$ with $\mathbf{C} \succeq \mathbf{O}$.
(e) Every principal minor of $\mathbf{A}$ is ≥ 0.

Idea. The spectral theorem makes (a) and (b) two readings of one diagonal matrix, since an orthonormal change of basis does not change the sign of a quadratic form. From (b) the square root of Chapter 12 produces $\mathbf{C}$, and $\mathbf{C}$ is its own $\mathbf{B}$. The step $(c) \Rightarrow (a)$ is the identity $\langle \mathbf{B}^*\mathbf{B}\mathbf{x}, \mathbf{x}\rangle = \|\mathbf{B}\mathbf{x}\|^2$, one line of moving $\mathbf{B}^*$ across. For (e), a principal submatrix is what the form does to vectors supported on a few coordinates, which gives $(a) \Rightarrow (e)$ at once; the return trip uses Lemma 13.1.3 to show that $\det(\mathbf{A} + t\mathbf{I}) > 0$ for every $t > 0$, so that no $-t$ can be an eigenvalue.

Proof. **(a) $\Leftrightarrow$ (b).** ($\Rightarrow$) Let $\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$ with $\mathbf{v} \neq \mathbf{0}$. By Theorem 12.2.1, $\lambda \in \mathbb{R}$, and

$$0 \leq \langle \mathbf{A}\mathbf{v}, \mathbf{v}\rangle = \langle \lambda\mathbf{v}, \mathbf{v}\rangle = \lambda\|\mathbf{v}\|^2,$$

so $\lambda \geq 0$, as $\|\mathbf{v}\|^2 > 0$. ($\Leftarrow$) By Corollary 12.4.3 (over $\mathbb{C}$) or Corollary 12.5.2 (over $\mathbb{R}$), $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{U}^*$ with $\mathbf{U}$ unitary and $\mathbf{D} = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$ the eigenvalues. Put $\mathbf{y} = \mathbf{U}^*\mathbf{x}$. Then

$$\langle \mathbf{A}\mathbf{x}, \mathbf{x}\rangle = \mathbf{x}^*\mathbf{U}\mathbf{D}\mathbf{U}^*\mathbf{x} = \mathbf{y}^*\mathbf{D}\mathbf{y} = \sum_i \lambda_i |y_i|^2 \geq 0,$$

every term being a product of two nonnegative reals.

(b) $\Rightarrow$ (d). A Hermitian matrix has an orthonormal basis of eigenvectors, so the operator $T_{\mathbf{A}}$ on F^n satisfies the hypothesis of Theorem 12.8.4, whose spectral condition $\operatorname{spec}(T_{\mathbf{A}}) \subseteq [0, \infty)$ is exactly (b). That theorem gives a unique $\mathbf{C}$ with an orthonormal eigenbasis, nonnegative spectrum, and $\mathbf{C}^2 = \mathbf{A}$, and it is self-adjoint; by the equivalence just proved, applied to $\mathbf{C}$, this says $\mathbf{C} \succeq \mathbf{O}$. Conversely any $\mathbf{C} \succeq \mathbf{O}$ with $\mathbf{C}^2 = \mathbf{A}$ is Hermitian with nonnegative spectrum, hence is that one. So there is exactly one such $\mathbf{C}$.

(d) $\Rightarrow$ (c). Take $\mathbf{B} = \mathbf{C}$; then $\mathbf{B}^*\mathbf{B} = \mathbf{C}^*\mathbf{C} = \mathbf{C}^2 = \mathbf{A}$.

(c) $\Rightarrow$ (a). Conjugate-transposing twice, $(\mathbf{B}^*\mathbf{B})^* = \mathbf{B}^*(\mathbf{B}^*)^* = \mathbf{B}^*\mathbf{B}$, which is (P1). And for every $\mathbf{x}$,

$$\langle \mathbf{B}^*\mathbf{B}\mathbf{x}, \mathbf{x}\rangle = \mathbf{x}^*\mathbf{B}^*\mathbf{B}\mathbf{x} = (\mathbf{B}\mathbf{x})^*(\mathbf{B}\mathbf{x}) = \|\mathbf{B}\mathbf{x}\|^2 \;\geq\; 0,$$

which is (P2). Both facts are recorded in operator form in Exercise (C1: The operator S^*S) (a), applied to the map $\mathbf{x} \mapsto \mathbf{B}\mathbf{x}$.

(a) $\Rightarrow$ (e). Let $I = \{i_1 < \cdots < i_k\}$. Given $\mathbf{y} \in F^k$, let $\mathbf{x} \in F^n$ have $x_{i_p} = y_p$ and all other entries 0. Then $\mathbf{x}^*\mathbf{A}\mathbf{x} = \mathbf{y}^*\mathbf{A}_{I,I}\mathbf{y}$, because the terms $\overline{x_i}a_{ij}x_j$ with $i \notin I$ or $j \notin I$ vanish. So $\mathbf{A}_{I,I}$ is Hermitian and satisfies (P2), that is $\mathbf{A}_{I,I} \succeq \mathbf{O}$. By the equivalence (a) $\Leftrightarrow$ (b) applied to $\mathbf{A}_{I,I}$, all its eigenvalues are ≥ 0. Diagonalizing it as $\mathbf{A}_{I,I} = \mathbf{U}\mathbf{D}\mathbf{U}^*$ as above, with $\mathbf{U}^* = \mathbf{U}^{-1}$, and taking determinants (Theorem 7.5.4, Corollary 7.5.6),

$$\det \mathbf{A}_{I,I} = \det \mathbf{D} = \lambda_1 \cdots \lambda_k \geq 0.$$

(e) ⇒ (b). Let $t > 0$ be real. Every $E_k(\mathbf{A}) \geq 0$ by hypothesis and $E_0(\mathbf{A}) = 1$, so Lemma 13.1.3 gives

$$\det(\mathbf{A} + t\mathbf{I}) = \sum_{k=0}^{n} E_k(\mathbf{A}) t^{n-k} \geq t^n > 0.$$

In particular $\mathbf{A} + t\mathbf{I}$ is invertible (Theorem 7.5.3), so $-t$ is not an eigenvalue of $\mathbf{A}$ (Theorem 9.2.6 applied to $\mathbf{A} + t\mathbf{I}$). As $t > 0$ was arbitrary and every eigenvalue of $\mathbf{A}$ is real (Theorem 12.2.1), no eigenvalue is negative. This closes the cycle and proves the theorem.

■

Item (c) is the one to keep in mind: **every positive semidefinite matrix is a Gram matrix**, since $\mathbf{B}^*\mathbf{B}$ is the Gram matrix of the columns of $\mathbf{B}$ (Definition 11.2.3). Section 3 turns that remark into a theorem. Item (d) names the square root, which Section 2 develops. Item (b) is how one *thinks*; item (e) is, for small matrices, how one *computes*.

13.1.4 The definite case, and Sylvester's criterion

The definite versions of (a), (b) and (c) are proved by putting "strict" in front of the same arguments. The new item is a test that can be run by hand on a numerical matrix without finding a single eigenvalue.

Theorem 13.1.5. Characterizations of Positive Definiteness, with Sylvester's Criterion

Let $\mathbf{A} \in M_n(F)$ be Hermitian, and for $1 \leq k \leq n$ let $\mathbf{A}_k$ be its leading principal submatrix, the top-left $k \times k$ corner. The following are equivalent.

(a) $\mathbf{A} \succ \mathbf{O}$.

(b) Every eigenvalue of $\mathbf{A}$ is a positive real number.

(c) $\mathbf{A} = \mathbf{B}^*\mathbf{B}$ for some **invertible** $\mathbf{B} \in M_n(F)$.

(d) **(Sylvester's criterion)** $\det \mathbf{A}_k > 0$ for every $k = 1, \ldots, n$.

Idea. Steps (a) ⇔ (b) ⇔ (c) repeat the previous proof with strict inequalities. For (a) ⇒ (d), a leading principal submatrix inherits definiteness, and a determinant is a product of eigenvalues. The content is (d) ⇒ (a), and it is an induction on n that peels off the last row and column. The inductive hypothesis makes the corner $\mathbf{A}_{n-1}$ definite, hence invertible, and one shear — the congruence by $\mathbf{S} = \begin{pmatrix} \mathbf{I} & -\mathbf{A}_{n-1}^{-1}\mathbf{b} \\ \mathbf{0}^\top & 1 \end{pmatrix}$, which is exactly the block elimination of Chapter 8 — clears the border and turns $\mathbf{A}$ into $\mathbf{A}_{n-1} \oplus (s)$ with one scalar s left over. Determinants identify the sign of s, and Proposition 13.1.2 carries the verdict back.

Proof. **(a) ⇔ (b).** As in Theorem 13.1.4, with $\geq$ replaced by $>$ throughout: for (⇒), $0 < \lambda\|\mathbf{v}\|^2$ forces $\lambda > 0$; for (⇐), $\sum_i \lambda_i |y_i|^2 > 0$ whenever $\mathbf{y} = \mathbf{U}^*\mathbf{x} \neq \mathbf{0}$, which happens exactly when $\mathbf{x} \neq \mathbf{0}$ because $\mathbf{U}^*$ is invertible.

(b) ⇒ (c). Let $\mathbf{C} \succeq \mathbf{O}$ with $\mathbf{C}^2 = \mathbf{A}$ (Theorem 13.1.4 (d)). Since every eigenvalue of $\mathbf{A}$ is positive, $0 \notin \operatorname{spec}(\mathbf{A})$ and $\mathbf{A}$ is invertible (Theorem 9.2.6). If $\mathbf{C}\mathbf{x} = \mathbf{0}$ then $\mathbf{A}\mathbf{x} = \mathbf{C}^2\mathbf{x} = \mathbf{0}$, so $\mathbf{x} = \mathbf{0}$; hence $\mathbf{C}$ is invertible too (Theorem 3.4.4). Take $\mathbf{B} = \mathbf{C}$.

(c) ⇒ (a). For $\mathbf{x} \neq \mathbf{0}$, $\mathbf{Bx} \neq \mathbf{0}$, so $\langle \mathbf{B}^*\mathbf{Bx}, \mathbf{x}\rangle = \|\mathbf{Bx}\|^2 > 0$.

(a) ⇒ (d). As in the proof of (a) ⇒ (e) of Theorem 13.1.4 with $I = \{1, \dots, k\}$, the corner $\mathbf{A}_k$ satisfies $\mathbf{y}^*\mathbf{A}_k\mathbf{y} = \mathbf{x}^*\mathbf{Ax} > 0$ for $\mathbf{y} \neq \mathbf{0}$, so $\mathbf{A}_k \succ \mathbf{O}$. Its eigenvalues are positive, and $\det \mathbf{A}_k$ is their product, computed as in the proof of Theorem 13.1.4, so $\det \mathbf{A}_k > 0$.

(d) ⇒ (a). Induct on n. For $n = 1$, $\mathbf{A} = (a_{11})$ with $a_{11} = \det \mathbf{A}_1 > 0$, and $\langle \mathbf{A}x, x\rangle = a_{11}|x|^2 > 0$ for $x \neq 0$.

Let $n \geq 2$, and write

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}' & \mathbf{b} \\ \mathbf{b}^* & d \end{pmatrix}, \qquad \mathbf{A}' = \mathbf{A}_{n-1},\ \mathbf{b} \in F^{n-1},\ d \in \mathbb{R}.$$

The leading principal submatrices of $\mathbf{A}'$ are $\mathbf{A}_1, \dots, \mathbf{A}_{n-1}$, so the inductive hypothesis gives $\mathbf{A}' \succ \mathbf{O}$; in particular $\mathbf{A}'$ is invertible, since $\det \mathbf{A}' = \det \mathbf{A}_{n-1} > 0$. Put

$$\mathbf{S} = \begin{pmatrix} \mathbf{I}_{n-1} & -\mathbf{A}'^{-1}\mathbf{b} \\ \mathbf{0}^\top & 1 \end{pmatrix}, \qquad s = d - \mathbf{b}^*\mathbf{A}'^{-1}\mathbf{b}.$$

Block multiplication (Theorem 8.1.1), together with $(\mathbf{A}'^{-1})^* = \mathbf{A}'^{-1}$ because $\mathbf{A}'$ is Hermitian, gives

$$\mathbf{S}^*\mathbf{AS} = \begin{pmatrix} \mathbf{A}' & \mathbf{0} \\ \mathbf{0}^\top & s \end{pmatrix}. \tag{$*$}$$

Both $\mathbf{S}$ and $\mathbf{S}^*$ are triangular with 1s on the diagonal, so $\det \mathbf{S} = \det \mathbf{S}^* = 1$ by Theorem 7.4.6. Taking determinants in $(*)$ with Theorem 7.5.4 and Theorem 7.8.4,

$$0 < \det \mathbf{A} = \det(\mathbf{S}^*\mathbf{AS}) = s \det \mathbf{A}',$$

and $\det \mathbf{A}' > 0$, so $s > 0$. The right-hand side of $(*)$ is therefore positive definite: for $\mathbf{y} = (\mathbf{y}', y_n) \neq \mathbf{0}$,

$$\mathbf{y}^* \begin{pmatrix} \mathbf{A}' & \mathbf{0} \\ \mathbf{0}^\top & s \end{pmatrix} \mathbf{y} = \mathbf{y}'^*\mathbf{A}'\mathbf{y}' + s|y_n|^2 > 0,$$

since the two terms are ≥ 0 and not both zero. Finally $\mathbf{S}$ is invertible, so Proposition 13.1.2 (b) turns $\mathbf{S}^*\mathbf{AS} \succ \mathbf{O}$ back into $\mathbf{A} \succ \mathbf{O}$. This completes the induction and proves the theorem.

■

The number $s = d - \mathbf{b}^*\mathbf{A}'^{-1}\mathbf{b}$ in that proof is the Schur complement $\mathbf{A}/\mathbf{A}'$ of Definition 8.3.1, and $(*)$ is the block ŁDU factorization of Theorem 8.3.1 written as a congruence. Section 6 replaces the scalar s by a block and gets the general test.

Warning. **For semidefiniteness, the leading principal minors are not enough.** Sylvester's criterion has no "$\geq$" version. Take

$$\mathbf{A} = \begin{pmatrix} 0 & 0 \\ 0 & -1 \end{pmatrix}.$$

Its leading principal minors are $\det \mathbf{A}_1 = 0$ and $\det \mathbf{A}_2 = 0$, both ≥ 0, and yet $\langle \mathbf{A}\mathbf{e}_2, \mathbf{e}_2\rangle = -1$, so $\mathbf{A}$ is not positive semidefinite. What Theorem 13.1.4 (e) demands is **all** principal minors, and the one this matrix fails is $\det \mathbf{A}_{\{2\},\{2\}} = -1 < 0$. The leading corners simply never

look at the second coordinate on its own. This is the single most common error about the subject: for the strict test n minors suffice, for the weak test one needs all $2^n - 1$ of them.

Quick check. Is $\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix}$ positive semidefinite? Is it positive definite? Which minors did you use?

Solution. It is positive semidefinite but not positive definite. The principal minors are 1, 4, and $\det \mathbf{A} = 4 - 4 = 0$, all ≥ 0, so Theorem 13.1.4 (e) applies. It is not definite, since $\det \mathbf{A}_2 = 0$ violates Sylvester's criterion; concretely $\mathbf{A}(2, -1) = \mathbf{0}$, so $\langle \mathbf{Ax}, \mathbf{x} \rangle = 0$ at the nonzero $\mathbf{x} = (2, -1)$. One can also see everything at once from $\mathbf{A} = \mathbf{b}\mathbf{b}^\top$ with $\mathbf{b} = (1, 2)$, which is Example (Five Positive Matrices) (b).

13.1.5 Two by two, and where these matrices come from

The 2×2 case is worth memorizing, because it is the one that appears inside larger arguments.

Example. The 2×2 Test

Let $\mathbf{A} = \begin{pmatrix} a & b \\ \bar{b} & d \end{pmatrix}$ be Hermitian, so $a, d \in \mathbb{R}$ and $b \in F$. State and justify the tests for $\mathbf{A} \succ \mathbf{O}$ and for $\mathbf{A} \succeq \mathbf{O}$.

Solution. $\mathbf{A} \succ \mathbf{O}$ if and only if $a > 0$ and $ad - |b|^2 > 0$. This is Sylvester's criterion, Theorem 13.1.5 (d): the two leading principal minors are a and $\det \mathbf{A} = ad - |b|^2$.

$\mathbf{A} \succeq \mathbf{O}$ if and only if $a \geq 0$, $d \geq 0$, and $ad - |b|^2 \geq 0$. This is Theorem 13.1.4 (e): the principal minors are a, d, and $\det \mathbf{A}$, and all three are needed. Dropping $d \geq 0$ readmits $\operatorname{diag}(0, -1)$.

Note what the second condition says: $|b|^2 \leq ad$, so an off-diagonal entry can never be larger in modulus than the geometric mean of the two diagonal entries it sits between. For $\mathbf{A} = \begin{pmatrix} 2 & 1+i \\ 1-i & 3 \end{pmatrix}$ over $\mathbb{C}$: $a = 2 > 0$ and $ad - |b|^2 = 6 - 2 = 4 > 0$, so $\mathbf{A} \succ \mathbf{O}$.

Where do positive semidefinite matrices come from? Almost always from item (c) of Theorem 13.1.4, and the standard source is data.

Example. A Covariance-style Matrix

Three quantities are measured on four occasions, and each column of

$$\mathbf{X} = \begin{pmatrix} 2 & 1 & 0 \\ -1 & 1 & 1 \\ -1 & -1 & 0 \\ 0 & -1 & -1 \end{pmatrix} \in M_{4\times 3}(\mathbb{R})$$

has been adjusted to sum to 0. Show that $\mathbf{A} = \mathbf{X}^\top\mathbf{X}$ is positive definite, and check the verdict against Sylvester's criterion.

Solution. Multiplying out, a_{ij} is the dot product of columns i and j of $\mathbf{X}$:

$$\mathbf{A} = \begin{pmatrix} 6 & 2 & -1 \\ 2 & 4 & 2 \\ -1 & 2 & 2 \end{pmatrix}.$$

The columns of $\mathbf{X}$ are linearly independent: if $\mathbf{Xx} = \mathbf{0}$, rows 1 and 3 read $2x_1 + x_2 = 0$ and $-x_1 - x_2 = 0$, which force $x_1 = x_2 = 0$, and then row 2 reads $x_3 = 0$. So $\operatorname{nul}(\mathbf{X}) = \{\mathbf{0}\}$, and taking $\mathbf{S} = \mathbf{X}$ in Proposition 13.1.2 (b) with the positive definite matrix $\mathbf{I}_4$ gives $\mathbf{A} = \mathbf{X}^\top\mathbf{I}_4\mathbf{X} \succ \mathbf{O}$. Equivalently, $\langle \mathbf{Ax}, \mathbf{x}\rangle = \|\mathbf{Xx}\|^2 > 0$ for $\mathbf{x} \neq \mathbf{0}$.

Sylvester's criterion agrees: $\det \mathbf{A}_1 = 6$, $\det \mathbf{A}_2 = 24 - 4 = 20$, and expanding along the first row,

$$\det \mathbf{A} = 6(8-4) - 2(4+2) + (-1)(4+4) = 24 - 12 - 8 = 4,$$

all positive. The reason the two routes must agree is that $\mathbf{A}$ is the Gram matrix of the columns of $\mathbf{X}$, and Theorem 11.2.5 (d) already said that its determinant is positive exactly when those columns are independent.

13.1.6 The invertible matrix theorem grows

Chapter 2 collected the conditions equivalent to invertibility for an operator (Theorem 2.5.7); Chapter 3 gave the eight matrix conditions of the Invertible Matrix Theorem (Theorem 3.4.4); Chapter 7 added the determinant (Theorem 7.5.3) and Chapter 9 the eigenvalue 0 (Theorem 9.2.6). Positivity adds a pair, and they are the ones used whenever a square system is attacked through least squares.

Theorem 13.1.6. Invertible Matrix Theorem, with Positive Definiteness

Let $\mathbf{A} \in M_n(F)$ with $F = \mathbb{R}$ or $F = \mathbb{C}$. The following are equivalent, and each is equivalent to each of the conditions of Theorem 3.4.4, Theorem 7.5.3, and Theorem 9.2.6.

(a) $\mathbf{A}$ is invertible.
(b) $\mathbf{A}^*\mathbf{A} \succ \mathbf{O}$.
(c) $\mathbf{A}\mathbf{A}^* \succ \mathbf{O}$.

For a general $\mathbf{A} \in M_{m\times n}(F)$, one always has $\mathbf{A}^*\mathbf{A} \succeq \mathbf{O}$, with $\mathbf{A}^*\mathbf{A} \succ \mathbf{O}$ if and only if the columns of $\mathbf{A}$ are linearly independent.

Proof. Let $\mathbf{A} \in M_{m\times n}(F)$. Then $\mathbf{A}^*\mathbf{A}$ is positive semidefinite by Theorem 13.1.4 ((c) $\Rightarrow$ (a)). Whether it is definite is decided by the identity $\langle \mathbf{A}^*\mathbf{A}\mathbf{x}, \mathbf{x}\rangle = \mathbf{x}^*\mathbf{A}^*\mathbf{A}\mathbf{x} = \|\mathbf{A}\mathbf{x}\|^2$, computed in the proof of Theorem 13.1.4 ((c) $\Rightarrow$ (a)) and recorded in operator form in Exercise (C1: The operator S^*S) (a): the left side is > 0 for every $\mathbf{x} \neq \mathbf{0}$ exactly when $\mathbf{A}\mathbf{x} \neq \mathbf{0}$ for every $\mathbf{x} \neq \mathbf{0}$, that is, exactly when $\operatorname{nul}(\mathbf{A}) = \{\mathbf{0}\}$, that is, exactly when the columns of $\mathbf{A}$ are linearly independent. The same conclusion comes from Lemma 11.4.3, which says $\operatorname{nul}(\mathbf{A}^*\mathbf{A}) = \operatorname{nul}(\mathbf{A})$ and so lets one read the condition off $\mathbf{A}$ directly.

Now let $\mathbf{A}$ be square. Then "$\operatorname{nul}(\mathbf{A}) = \{\mathbf{0}\}$" is item (b) of Theorem 3.4.4, which gives (a) $\Leftrightarrow$ (b). Also $\mathbf{A}$ is invertible if and only if $\mathbf{A}^*$ is, since $(\mathbf{A}^{-1})^*$ is a two-sided inverse of $\mathbf{A}^*$ and conversely; applying (a) $\Leftrightarrow$ (b) to $\mathbf{A}^*$ therefore gives (a) $\Leftrightarrow$ (c).

■

The two new items look like a detour — why test $\mathbf{A}^*\mathbf{A}$ when $\mathbf{A}$ itself is available? Because $\mathbf{A}^*\mathbf{A}$ is Hermitian and positive, and everything in this chapter applies to it. Chapter 11 already used it once, in the normal equations of least squares; Section 9 will use it again, to build the singular value decomposition of a matrix that is not even square, and that decomposition supplies the last item of the list: a square matrix is invertible exactly when none of its singular values is zero.

13.1.7 Exercises

13.1.7.1 A. Check your understanding

Exercise. A1

(a) State the definition of $\mathbf{A} \succeq \mathbf{O}$ and of $\mathbf{A} \succ \mathbf{O}$, including every quantifier and the field.

(b) Explain why clause (P1) cannot be dropped from the real definition, with a concrete matrix.

(c) True or false: if $\mathbf{A}$ is Hermitian and every leading principal minor of $\mathbf{A}$ is ≥ 0, then $\mathbf{A} \succeq \mathbf{O}$. Justify your answer.

(d) Write down the three conditions of the 2×2 test for $\mathbf{A} \succeq \mathbf{O}$.

(e) Name the item of Theorem 13.1.4 that says every positive semidefinite matrix is a Gram matrix.

Solution.

(a) $\mathbf{A} \in M_n(F)$ with $F = \mathbb{R}$ or $\mathbb{C}$ is positive semidefinite if $\mathbf{A}^* = \mathbf{A}$ and $\langle \mathbf{A}\mathbf{x}, \mathbf{x}\rangle \geq 0$ for **every** $\mathbf{x} \in F^n$; positive definite if $\mathbf{A}^* = \mathbf{A}$ and $\langle \mathbf{A}\mathbf{x}, \mathbf{x}\rangle > 0$ for every **nonzero** $\mathbf{x} \in F^n$ (Definition 13.1.1).

(b) Because over $\mathbb{R}$ the quadratic form does not see the skew part. The quarter turn $\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ has $\langle \mathbf{A}\mathbf{x}, \mathbf{x}\rangle = 0$ for every $\mathbf{x} \in \mathbb{R}^2$ and is not symmetric.

(c) False. $\operatorname{diag}(0, -1)$ has leading principal minors 0 and 0, and $\langle \mathbf{A}\mathbf{e}_2, \mathbf{e}_2\rangle = -1$. All principal minors are needed (Theorem 13.1.4 (e)).

(d) With $\mathbf{A} = \begin{pmatrix} a & b \\ \bar{b} & d \end{pmatrix}$ Hermitian: $a \geq 0$, $d \geq 0$, and $ad - |b|^2 \geq 0$.

(e) Item (c): $\mathbf{A} = \mathbf{B}^*\mathbf{B}$, and $\mathbf{B}^*\mathbf{B}$ is the Gram matrix of the columns of $\mathbf{B}$.

13.1.7.2 B. Practice

Exercise. B1: Determine the definiteness

For each matrix below, determine whether it is positive definite, positive semidefinite but not positive definite, or neither. Justify your answer.

(a) $\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$ over $\mathbb{R}$.

(b) $\begin{pmatrix} 0 & 0 \\ 0 & -1 \end{pmatrix}$ over $\mathbb{R}$.

(c) $\begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix}$ over $\mathbb{R}$.

(d) The all-ones matrix $\mathbf{J} \in M_3(\mathbb{R})$.

(e) $\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$ over $\mathbb{R}$.

(f) $\begin{pmatrix} 2 & 1+i \\ 1-i & 1 \end{pmatrix}$ over $\mathbb{C}$.

Solution.

(a) Positive definite. Symmetric, with leading principal minors 2 and $4 - 1 = 3$, both positive (Theorem 13.1.5 (d)).

(b) Neither. The leading principal minors are 0 and 0, which tempts one into "semidefinite", but the principal minor $\det \mathbf{A}_{\{2\},\{2\}} = -1$ is negative, and $\langle \mathbf{A}\mathbf{e}_2, \mathbf{e}_2 \rangle = -1 < 0$.

(c) Positive definite. Symmetric, with leading principal minors 2, $4 - 1 = 3$, and $2 \cdot 3 - (-1)(-2) = 4$, all positive.

(d) Positive semidefinite, not positive definite. $\mathbf{J} = \mathbf{1}\mathbf{1}^\top$, which is Example (Five Positive Matrices) (b) with $\mathbf{x} = \mathbf{1}$; so $\langle \mathbf{J}\mathbf{x}, \mathbf{x} \rangle = |\langle \mathbf{x}, \mathbf{1} \rangle|^2 \geq 0$. It vanishes at the nonzero $\mathbf{x} = (1, -1, 0)$, so it is not definite. (Equivalently, $\det \mathbf{J} = 0$.)

(e) Neither: it is not symmetric, so (P1) fails and it is not in the class at all. Its quadratic form is identically 0, which shows how little the form sees over $\mathbb{R}$.

(f) Positive semidefinite, not positive definite. Hermitian, with $a = 2 \geq 0$, $d = 1 \geq 0$, and $ad - |b|^2 = 2 - 2 = 0$. The determinant is 0, so definiteness fails; indeed $\mathbf{A}(1+i, -2) = \mathbf{0}$.

Exercise. B2: A parameter

For which $t \in \mathbb{R}$ is

$$\mathbf{A}(t) = \begin{pmatrix} 1 & t & 0 \\ t & 2 & t \\ 0 & t & 3 \end{pmatrix}$$

positive definite? Justify your answer.

Solution. $\mathbf{A}(t)$ is symmetric for every real t, so Sylvester's criterion applies (Theorem 13.1.5 (d)). The leading principal minors are

$$\det \mathbf{A}_1 = 1, \qquad \det \mathbf{A}_2 = 2 - t^2, \qquad \det \mathbf{A}_3 = 6 - 4t^2,$$

the last by expansion along the first row: $1 \cdot (6 - t^2) - t \cdot (3t - 0) = 6 - 4t^2$. All three are positive exactly when $t^2 < 2$ and $t^2 < 3/2$, that is when $t^2 < 3/2$. Hence $\mathbf{A}(t) \succ \mathbf{O}$ if and only if $|t| < \sqrt{6}/2$.

At $t = \sqrt{6}/2$ the matrix is positive semidefinite but singular, and for $|t| > \sqrt{6}/2$ the determinant is negative, so $\mathbf{A}(t)$ has an odd number of negative eigenvalues and is not even semidefinite.

Exercise. B3: The inverse of a positive definite matrix

Let $\mathbf{A} \in M_n(F)$ with $\mathbf{A} \succ \mathbf{O}$. Prove that $\mathbf{A}$ is invertible and that $\mathbf{A}^{-1} \succ \mathbf{O}$.

Solution. By Theorem 13.1.5 (b), every eigenvalue of $\mathbf{A}$ is positive, so $0 \notin \operatorname{spec}(\mathbf{A})$ and $\mathbf{A}$ is invertible (Theorem 9.2.6). Then $(\mathbf{A}^{-1})^* = (\mathbf{A}^*)^{-1} = \mathbf{A}^{-1}$, so $\mathbf{A}^{-1}$ is Hermitian. Apply Proposition 13.1.2 (b) with $\mathbf{S} = \mathbf{A}^{-1}$, which is invertible:

$$\mathbf{S}^*\mathbf{A}\mathbf{S} = \mathbf{A}^{-1}\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1} \succ \mathbf{O}.$$

(Alternatively: the eigenvalues of $\mathbf{A}^{-1}$ are the reciprocals $1/\lambda_i$ of those of $\mathbf{A}$, hence positive, and Theorem 13.1.5 (b) applies again.)

13.1.7.3 C. Going deeper

Exercise. C1: Zero trace forces zero

Let $\mathbf{A} \in M_n(F)$ with $\mathbf{A} \succeq \mathbf{O}$.

(a) Prove that $a_{ii} \geq 0$ for every i, and that $\operatorname{tr} \mathbf{A} \geq 0$.

(b) Prove that if $\operatorname{tr} \mathbf{A} = 0$ then $\mathbf{A} = \mathbf{O}$.

(c) Deduce that if $\mathbf{A} \succeq \mathbf{O}$ and $-\mathbf{A} \succeq \mathbf{O}$, then $\mathbf{A} = \mathbf{O}$.

Solution.

(a) $a_{ii} = \mathbf{e}_i^* \mathbf{A} \mathbf{e}_i = \langle \mathbf{A}\mathbf{e}_i, \mathbf{e}_i \rangle \geq 0$ by (P2). Summing, $\operatorname{tr} \mathbf{A} = \sum_i a_{ii} \geq 0$.

(b) Write $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{U}^*$ with $\mathbf{U}$ unitary and $\mathbf{D} = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ the eigenvalues (Corollary 12.4.3, Corollary 12.5.2). Then $\operatorname{tr} \mathbf{A} = \operatorname{tr}(\mathbf{D}\mathbf{U}^*\mathbf{U}) = \operatorname{tr} \mathbf{D} = \sum_i \lambda_i$ by Theo-

rem 0.9.6, and each $\lambda_i \geq 0$ by Theorem 13.1.4 (b). A sum of nonnegative reals is 0 only if every term is 0, so $\mathbf{D} = \mathbf{O}$ and hence $\mathbf{A} = \mathbf{O}$.
Alternatively, without eigenvalues: by (a) each $a_{ii} \geq 0$, so $\operatorname{tr} \mathbf{A} = 0$ forces every $a_{ii} = 0$. Writing $\mathbf{A} = \mathbf{B}^*\mathbf{B}$ (Theorem 13.1.4 (c)), the i-th diagonal entry is $\|\mathbf{b}_i\|^2$ for the i-th column $\mathbf{b}_i$ of $\mathbf{B}$; so every column of $\mathbf{B}$ is $\mathbf{0}$, that is $\mathbf{B} = \mathbf{O}$ and $\mathbf{A} = \mathbf{O}$.

(c) If both $\mathbf{A} \succeq \mathbf{O}$ and $-\mathbf{A} \succeq \mathbf{O}$, then by (a) $\operatorname{tr} \mathbf{A} \geq 0$ and $\operatorname{tr}(-\mathbf{A}) = -\operatorname{tr} \mathbf{A} \geq 0$, so $\operatorname{tr} \mathbf{A} = 0$ and (b) gives $\mathbf{A} = \mathbf{O}$.

Exercise. C2: The cone of positive matrices

Let $\mathcal{P}_n \subseteq M_n(F)$ be the set of positive semidefinite matrices.

(a) Prove that if $\mathbf{A}, \mathbf{B} \in \mathcal{P}_n$ and $c \geq 0$ is real, then $\mathbf{A} + \mathbf{B} \in \mathcal{P}_n$ and $c\mathbf{A} \in \mathcal{P}_n$. (A set with these two properties is a **convex cone**: it contains $(1-t)\mathbf{A} + t\mathbf{B}$ for $0 \leq t \leq 1$ as well.)

(b) Prove that $\mathcal{P}_n$ contains no line through $\mathbf{O}$: if $\mathbf{A} \in \mathcal{P}_n$ and $-\mathbf{A} \in \mathcal{P}_n$ then $\mathbf{A} = \mathbf{O}$.

(c) Prove that $\mathcal{P}_n$ is closed under congruence: if $\mathbf{A} \in \mathcal{P}_n$ and $\mathbf{S} \in M_n(F)$, then $\mathbf{S}^*\mathbf{A}\mathbf{S} \in \mathcal{P}_n$. Give an example showing that $\mathcal{P}_n$ is **not** closed under multiplication.

Hint for (c): for the example, two positive semidefinite matrices whose product is not even Hermitian will do.

Solution.

(a) Sums and real multiples of Hermitian matrices are Hermitian. For every $\mathbf{x}$,

$$\langle (\mathbf{A} + \mathbf{B})\mathbf{x}, \mathbf{x} \rangle = \langle \mathbf{A}\mathbf{x}, \mathbf{x} \rangle + \langle \mathbf{B}\mathbf{x}, \mathbf{x} \rangle \geq 0,$$

and $\langle c\mathbf{A}\mathbf{x}, \mathbf{x} \rangle = c\langle \mathbf{A}\mathbf{x}, \mathbf{x} \rangle \geq 0$ since $c \geq 0$. Convexity follows by taking $c = 1 - t$ and $c = t$ and adding.

(b) This is Exercise (C1: Zero trace forces zero) (c).

(c) Immediate from Proposition 13.1.2 (a). For the failure of multiplicativity, take

$$\mathbf{A} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \qquad \mathbf{B} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix},$$

both positive semidefinite ($\mathbf{A} = \mathbf{e}_1\mathbf{e}_1^\top$ and $\mathbf{B} = \mathbf{1}\mathbf{1}^\top$). Then

$$\mathbf{A}\mathbf{B} = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix},$$

which is not symmetric, so it is not positive semidefinite — it fails (P1), never mind (P2). Section 2 shows what does survive: $\operatorname{tr}(\mathbf{A}\mathbf{B}) \geq 0$, and $\mathbf{A}\mathbf{B} \succeq \mathbf{O}$ whenever $\mathbf{A}$ and $\mathbf{B}$ commute.

Exercise. C3: The largest entry sits on the diagonal

Let $\mathbf{A} \in M_n(F)$ with $\mathbf{A} \succeq \mathbf{O}$.

(a) Prove that $|a_{ij}|^2 \le a_{ii}a_{jj}$ for all i, j.

(b) Deduce that if $a_{ii} = 0$ for some i, then the whole i-th row and i-th column of $\mathbf{A}$ are zero.

(c) Deduce that $\max_{i,j}|a_{ij}| = \max_i a_{ii}$.

Hint for (a): look at a 2×2 principal submatrix.

Solution.

(a) For $i = j$ the claim is $a_{ii}^2 \le a_{ii}^2$. For $i \ne j$, take $I = \{i, j\}$. By the proof of Theorem 13.1.4 ((a) $\Rightarrow$ (e)), $\mathbf{A}_{I,I} \succeq \mathbf{O}$, and Example (The 2×2 Test) gives $a_{ii}a_{jj} - |a_{ij}|^2 = \det \mathbf{A}_{I,I} \ge 0$.

(b) If $a_{ii} = 0$, then (a) gives $|a_{ij}|^2 \le 0 \cdot a_{jj} = 0$ for every j, so $a_{ij} = 0$; and $a_{ji} = \overline{a_{ij}} = 0$ since $\mathbf{A}$ is Hermitian.

(c) Let $M = \max_i a_{ii}$, a maximum of nonnegative reals by Exercise (C1: Zero trace forces zero) (a). By (a), $|a_{ij}| \le \sqrt{a_{ii}a_{jj}} \le M$ for all i, j, so $\max_{i,j}|a_{ij}| \le M$. The reverse inequality holds because the diagonal entries are among the entries. Hence the two maxima are equal.

13.2 Square Roots and the Cholesky Factorization

A nonnegative real number has exactly one nonnegative square root, and it is the reason "nonnegative" is a useful class of numbers at all: it lets us take lengths, define standard deviations, and undo a squaring. Section 1 produced the matrix analogue as one clause of a theorem. This section makes it an object with a name, adds the one property the clause did not mention — that the square root is a **polynomial** in the matrix — and then turns the same idea into the triangular factorization that elimination discovers, the one Chapter 3 promised and could not yet explain.

Throughout, $F = \mathbb{R}$ or $F = \mathbb{C}$, and all matrices are over F.

13.2.1 The positive square root

Theorem 13.1.4 (d) already says that a positive semidefinite $\mathbf{A}$ has exactly one positive semidefinite square root. That is an existence-and-uniqueness statement, so by the usual convention it deserves a notation.

The square root of a positive matrix is what you get by taking square roots of its eigenvalues and leaving the eigenvectors alone.

Theorem 13.2.1. The Positive Square Root

Let $\mathbf{A} \in M_n(F)$ with $\mathbf{A} \succeq \mathbf{O}$. Then there is exactly one $\mathbf{C} \in M_n(F)$ with $\mathbf{C} \succeq \mathbf{O}$ and $\mathbf{C}^2 = \mathbf{A}$. We write $\mathbf{C} = \mathbf{A}^{1/2}$, also $\sqrt{\mathbf{A}}$. Moreover:

(a) $\mathbf{A}^{1/2} = p(\mathbf{A})$ for some polynomial p with real coefficients (the nodes λ_i and the values $\sqrt{\lambda_i}$ are real, so the interpolation of Theorem 6.6.1 may be run over $\mathbb{R}$, and $\mathbb{R}[x] \subseteq \mathbb{C}[x]$);

(b) $\mathbf{A}^{1/2}\mathbf{B} = \mathbf{B}\mathbf{A}^{1/2}$ for **every** $\mathbf{B} \in M_n(F)$ with $\mathbf{A}\mathbf{B} = \mathbf{B}\mathbf{A}$;

(c) $\operatorname{nul}(\mathbf{A}^{1/2}) = \operatorname{nul}(\mathbf{A})$ and $\operatorname{rank}\mathbf{A}^{1/2} = \operatorname{rank}\mathbf{A}$; in particular $\mathbf{A}^{1/2}$ is invertible if and only if $\mathbf{A}$ is.

The eigenvalues of $\mathbf{A}^{1/2}$ are the nonnegative square roots of those of $\mathbf{A}$, with the same eigenvectors.

Idea. Existence and uniqueness are Section 1. What is new is (a), and it is the reason (b) is free. On the spectrum, the square root is just a function $\lambda \mapsto \sqrt{\lambda}$ defined at finitely many points, and a function defined at finitely many points is a polynomial — that is Lagrange interpolation from Chapter 6. The functional calculus of Chapter 12 says that applying a polynomial to $\mathbf{A}$ and applying the corresponding function to the spectrum are the same operation, so the interpolating polynomial evaluated at $\mathbf{A}$ *is* the square root. Once $\mathbf{A}^{1/2}$ is a polynomial in $\mathbf{A}$, anything commuting with $\mathbf{A}$ commutes with it, with no further work.

Proof. Existence and uniqueness are Theorem 13.1.4 (d).

(a) Let $\lambda_1, \dots, \lambda_k$ be the **distinct** eigenvalues of $\mathbf{A}$ and let $\mathbf{A} = \sum_i \lambda_i \mathbf{P}_i$ be its spectral resolution (Definition 12.8.1, Theorem 12.8.1), which exists because a Hermitian matrix has an orthonormal basis of eigenvectors. Every λ_i is a nonnegative real number by Theorem 13.1.4 (b), so the numbers $\sqrt{\lambda_i}$ are defined and real. Let $f(\lambda) = \sqrt{\lambda}$ on $\operatorname{spec}(\mathbf{A})$. By Theorem 12.8.3 (b) and (a),

$$f(\mathbf{A})^2 = (f^2)(\mathbf{A}) = \mathbf{A},$$

and by (c) and (d), $f(\mathbf{A})$ is Hermitian with spectrum $\{\sqrt{\lambda_1}, \dots, \sqrt{\lambda_k}\} \subseteq [0, \infty)$, so $f(\mathbf{A}) \succeq \mathbf{O}$ by Theorem 13.1.4 (b). Uniqueness gives $f(\mathbf{A}) = \mathbf{A}^{1/2}$.
The numbers $\lambda_1, \dots, \lambda_k$ are distinct, so by Lagrange interpolation (Theorem 6.6.1) there is a polynomial p of degree at most $k-1$, with real coefficients, such that $p(\lambda_i) = \sqrt{\lambda_i}$ for every i. By Theorem 12.8.3 (a), $p(\mathbf{A}) = \sum_i p(\lambda_i)\mathbf{P}_i = f(\mathbf{A}) = \mathbf{A}^{1/2}$.

(b) Suppose $\mathbf{A}\mathbf{B} = \mathbf{B}\mathbf{A}$. Then $\mathbf{B}$ commutes with every power of $\mathbf{A}$, hence with every polynomial in $\mathbf{A}$, hence with $p(\mathbf{A}) = \mathbf{A}^{1/2}$.

(c) If $\mathbf{A}^{1/2}\mathbf{x} = \mathbf{0}$ then $\mathbf{A}\mathbf{x} = \mathbf{A}^{1/2}(\mathbf{A}^{1/2}\mathbf{x}) = \mathbf{0}$. Conversely, if $\mathbf{A}\mathbf{x} = \mathbf{0}$, then since $\mathbf{A}^{1/2}$ is Hermitian,

$$\|\mathbf{A}^{1/2}\mathbf{x}\|^2 = \langle \mathbf{A}^{1/2}\mathbf{x}, \mathbf{A}^{1/2}\mathbf{x}\rangle = \langle \mathbf{A}\mathbf{x}, \mathbf{x}\rangle = 0,$$

so $\mathbf{A}^{1/2}\mathbf{x} = \mathbf{0}$. Hence the two null spaces agree, and the ranks agree by Theorem 3.5.7. Invertibility is the case $\operatorname{nul} = \{\mathbf{0}\}$ (Theorem 3.4.4).

The last sentence is the spectral mapping statement of Theorem 12.8.3 (d) for $f = \sqrt{}$. ∎

Two routes to the same matrix, then, and both are worth practicing: assemble $\sum_i \sqrt{\lambda_i}\mathbf{P}_i$ from the projections, or find one interpolating polynomial and substitute.

Example. Two Square Roots

Compute $\mathbf{A}^{1/2}$ for

(a) $\mathbf{A} = \begin{pmatrix} 13 & 12 \\ 12 & 13 \end{pmatrix}$;

(b) $\mathbf{B} = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix}$,

in each case by the spectral resolution and by a polynomial.

Solution.

(a) $\mathbf{A}$ is symmetric. Since $\mathbf{A}(1,1) = (25, 25)$ and $\mathbf{A}(1,-1) = (1,-1)$, the eigenvalues are 25 and 1, with eigenlines spanned by $(1,1)$ and $(1,-1)$. Both are positive, so $\mathbf{A} \succ \mathbf{O}$. The orthogonal projections onto the two eigenlines are

$$\mathbf{P}_1 = \tfrac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}, \qquad \mathbf{P}_2 = \tfrac{1}{2}\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix},$$

so $\mathbf{A} = 25\mathbf{P}_1 + \mathbf{P}_2$ and

$$\mathbf{A}^{1/2} = 5\mathbf{P}_1 + 1 \cdot \mathbf{P}_2 = \begin{pmatrix} 3 & 2 \\ 2 & 3 \end{pmatrix}.$$

By the polynomial route: we need p with $p(25) = 5$ and $p(1) = 1$. The line through the two points is $p(x) = \frac{1}{6}x + \frac{5}{6}$, and

$$\tfrac{1}{6}\mathbf{A} + \tfrac{5}{6}\mathbf{I} = \tfrac{1}{6}\begin{pmatrix} 18 & 12 \\ 12 & 18 \end{pmatrix} = \begin{pmatrix} 3 & 2 \\ 2 & 3 \end{pmatrix}.$$

Check: $\begin{pmatrix} 3 & 2 \\ 2 & 3 \end{pmatrix}^2 = \begin{pmatrix} 13 & 12 \\ 12 & 13 \end{pmatrix}$.

(b) Here $\mathbf{B} = \mathbf{I} + \mathbf{J}$ with $\mathbf{J}$ the all-ones matrix. Since $\mathbf{J1} = 3\mathbf{1}$ and $\mathbf{Jx} = \mathbf{0}$ for every $\mathbf{x}$ with entries summing to 0, the eigenvalues of $\mathbf{J}$ are 3 and 0, so those of $\mathbf{B}$ are 4 and 1; both are positive, so $\mathbf{B} \succ \mathbf{O}$. We need p with $p(4) = 2$ and $p(1) = 1$, namely $p(x) = \frac{1}{3}x + \frac{2}{3}$, giving

$$\mathbf{B}^{1/2} = \tfrac{1}{3}\mathbf{B} + \tfrac{2}{3}\mathbf{I} = \tfrac{1}{3}\begin{pmatrix} 4 & 1 & 1 \\ 1 & 4 & 1 \\ 1 & 1 & 4 \end{pmatrix}.$$

Check by squaring: the diagonal entry of the square is $\frac{1}{9}(16 + 1 + 1) = 2$ and an off-diagonal entry is $\frac{1}{9}(4 + 4 + 1) = 1$, which reproduces $\mathbf{B}$. Only two distinct eigenvalues appear, so a polynomial of degree 1 suffices even though $\mathbf{B}$ is 3×3.

Warning. **$\mathbf{A}^{1/2}$ is not the entrywise square root, and it is not the only square root.** For $\mathbf{A} = \begin{pmatrix} 13 & 12 \\ 12 & 13 \end{pmatrix}$ the entrywise roots $\begin{pmatrix} \sqrt{13} & \sqrt{12} \\ \sqrt{12} & \sqrt{13} \end{pmatrix}$ form a matrix whose square is not $\mathbf{A}$. And $\mathbf{A}$ has other square roots: $-\mathbf{A}^{1/2}$ is one, and $5\mathbf{P}_1 - \mathbf{P}_2 = \begin{pmatrix} 2 & 3 \\ 3 & 2 \end{pmatrix}$ is another symmetric one. The word **positive** in "positive square root" is what makes the object unique; drop it

and this $\mathbf{A}$, whose two eigenvalues are distinct and positive, has four symmetric square roots, one for each choice of sign on each eigenline. When an eigenvalue repeats there are infinitely many, symmetric ones included: every reflection $\begin{pmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{pmatrix}$ is a symmetric square root of $\mathbf{I}_2$.

Quick check. Let $\mathbf{A} \succeq \mathbf{O}$ and let $\mathbf{U}$ be unitary. What is $(\mathbf{U}^*\mathbf{A}\mathbf{U})^{1/2}$?

Solution. $\mathbf{U}^*\mathbf{A}^{1/2}\mathbf{U}$. It is positive semidefinite by Proposition 13.1.2 (a), since $\mathbf{A}^{1/2} \succeq \mathbf{O}$, and its square is $\mathbf{U}^*\mathbf{A}^{1/2}\mathbf{U}\mathbf{U}^*\mathbf{A}^{1/2}\mathbf{U} = \mathbf{U}^*\mathbf{A}\mathbf{U}$ because $\mathbf{U}\mathbf{U}^* = \mathbf{I}$. Uniqueness in Theorem 13.2.1 does the rest. In words: the square root does not care which orthonormal coordinates we use, which is exactly what one wants of an operation defined on the spectrum.

13.2.2 The absolute value of a matrix

A complex number has a modulus $|z| = \sqrt{\bar{z}z}$, and the polar form $z = e^{i\theta}|z|$ splits it into a rotation and a nonnegative stretch. Section 10 will split a matrix the same way, and the stretching factor it needs is built now. For **any** matrix, square or not, $\mathbf{A}^*\mathbf{A}$ is positive semidefinite (Theorem 13.1.6), so it has a square root.

Definition 13.2.1. Absolute Value of a Matrix

Let $\mathbf{A} \in M_{m\times n}(F)$. The **absolute value** of $\mathbf{A}$ is

$$|\mathbf{A}| := (\mathbf{A}^*\mathbf{A})^{1/2} \in M_n(F),$$

the unique positive semidefinite square root of $\mathbf{A}^*\mathbf{A}$.

Note the size: $|\mathbf{A}|$ is $n \times n$, not $m \times n$, because $\mathbf{A}^*\mathbf{A}$ acts on the domain of $\mathbf{A}$. It is well defined because $\mathbf{A}^*\mathbf{A} \succeq \mathbf{O}$ and Theorem 13.2.1 gives exactly one root.

Proposition 13.2.2. What the absolute value preserves

Let $\mathbf{A} \in M_{m\times n}(F)$. Then for every $\mathbf{x} \in F^n$,

$$\||\mathbf{A}|\mathbf{x}\| = \|\mathbf{A}\mathbf{x}\|.$$

Consequently $\operatorname{nul}|\mathbf{A}| = \operatorname{nul}\mathbf{A}$ and $\operatorname{rank}|\mathbf{A}| = \operatorname{rank}\mathbf{A}$.

Proof. Since $|\mathbf{A}|$ is Hermitian and $|\mathbf{A}|^2 = \mathbf{A}^*\mathbf{A}$,

$$\||\mathbf{A}|\mathbf{x}\|^2 = \langle |\mathbf{A}|^2\mathbf{x}, \mathbf{x}\rangle = \langle \mathbf{A}^*\mathbf{A}\mathbf{x}, \mathbf{x}\rangle = \|\mathbf{A}\mathbf{x}\|^2,$$

using Lemma 11.4.1 twice. Both sides are nonnegative reals, so the norms agree. Hence

$|\mathbf{A}|\mathbf{x} = \mathbf{0}$ if and only if $\mathbf{A}\mathbf{x} = \mathbf{0}$, and the rank statement follows by Theorem 3.5.7.

■

So $|\mathbf{A}|$ does exactly what $\mathbf{A}$ does to lengths, and nothing else. That single identity is the whole engine of the polar decomposition: the map $|\mathbf{A}|\mathbf{x} \mapsto \mathbf{A}\mathbf{x}$ preserves norms, so it extends to a unitary matrix, and $\mathbf{A} = \mathbf{W}|\mathbf{A}|$ follows.

Example. Two Absolute Values

Compute $|\mathbf{A}|$ for $\mathbf{A} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ and for $\mathbf{B} = \begin{pmatrix} 3 & 0 \\ 4 & 0 \end{pmatrix}$, both over $\mathbb{R}$.

Solution. For the quarter turn, $\mathbf{A}^{\top}\mathbf{A} = \mathbf{I}$, so $|\mathbf{A}| = \mathbf{I}^{1/2} = \mathbf{I}$. This is right: a rotation changes no length, so the matrix that does the same thing to lengths and nothing else is the identity.

For $\mathbf{B}$, $\mathbf{B}^{\top}\mathbf{B} = \begin{pmatrix} 25 & 0 \\ 0 & 0 \end{pmatrix}$, whose positive square root is $|\mathbf{B}| = \begin{pmatrix} 5 & 0 \\ 0 & 0 \end{pmatrix}$. Check Proposition 13.2.2: $\|\mathbf{B}\mathbf{x}\| = \|(3x_1, 4x_1)\| = 5|x_1| = \||\mathbf{B}|\mathbf{x}\|$. Both have rank 1 and null space $\operatorname{span}(\mathbf{e}_2)$.

Warning. **$|\mathbf{A}|$ is not the entrywise absolute value, and $|\mathbf{A}\mathbf{B}| \neq |\mathbf{A}||\mathbf{B}|$.** For the quarter turn above, the entrywise absolute value is $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$, while $|\mathbf{A}| = \mathbf{I}$. For the product rule, take

$$\mathbf{A} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \qquad \mathbf{B} = \begin{pmatrix} 0 & 0 \\ 1 & 1 \end{pmatrix}.$$

Then $\mathbf{A}\mathbf{B} = \mathbf{O}$, so $|\mathbf{A}\mathbf{B}| = \mathbf{O}$, while $|\mathbf{A}| = \mathbf{A}$ and $|\mathbf{B}| = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$, giving

$$|\mathbf{A}||\mathbf{B}| = \tfrac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix} \neq \mathbf{O}.$$

That product is not even Hermitian, so it could not be the absolute value of anything. The modulus of a product of **numbers** is the product of the moduli because numbers commute; matrices do not.

13.2.3 Cholesky: elimination on a positive definite matrix

Chapter 3 ran elimination on a symmetric matrix and found that the two triangular factors are secretly the same one: $\mathbf{A} = \text{Ł}\mathbf{D}\text{Ł}^{\top}$ (Theorem 3.6.8). It then observed that when every entry of $\mathbf{D}$ is positive one can split $\mathbf{D} = \mathbf{D}^{1/2}\mathbf{D}^{1/2}$ and write $\mathbf{A} = (\text{Ł}\mathbf{D}^{1/2})(\text{Ł}\mathbf{D}^{1/2})^{\top}$, called this the **Cholesky factorization**, and promised that Chapter 13 would connect it with positive definite

matrices. Here is the connection, and it is the best kind: the hypothesis "every entry of $\mathbf{D}$ is positive" is not an extra assumption at all. It holds exactly when $\mathbf{A}$ is positive definite, and it can be checked while the elimination runs.

Theorem 13.2.3. Cholesky Factorization

Let $\mathbf{A} \in M_n(F)$ be Hermitian. Then $\mathbf{A} \succ \mathbf{O}$ if and only if

$$\mathbf{A} = Ł Ł^*$$

for some lower triangular $Ł \in M_n(F)$ whose diagonal entries are **real and strictly positive**. In that case $Ł$ is unique.

Idea. ($\Leftarrow$) is one line: such an $Ł$ is invertible, and $Ł Ł^* = (Ł^*)^* Ł^*$ is Theorem 13.1.5 (c) with $\mathbf{B} = Ł^*$.

($\Rightarrow$) is Chapter 3's elimination with its hypothesis supplied, run symmetrically. One pass clears the first column and the first row at once and leaves a smaller matrix that is again positive definite, so an induction produces $\mathbf{A} = \mathbf{M}\mathbf{D}\mathbf{M}^*$ with $\mathbf{M}$ unit lower triangular and every pivot $d_k > 0$. Absorbing $\mathbf{D}^{1/2}$ into $\mathbf{M}$ finishes it. For uniqueness, strip the positive diagonal off each candidate $Ł$; what is left is an LU factorization of $\mathbf{A}$, and Chapter 3 already proved those unique.

Proof. ($\Leftarrow$) Suppose $\mathbf{A} = Ł Ł^*$ with $Ł$ lower triangular and every $\ell_{kk} > 0$. Then $Ł$ is invertible by Lemma 3.6.2, hence so is $Ł^*$, and $\mathbf{A} = (Ł^*)^* Ł^*$ is positive definite by Theorem 13.1.5 ((c) $\Rightarrow$ (a)).

($\Rightarrow$) Suppose $\mathbf{A} \succ \mathbf{O}$.

Claim. There are a unit lower triangular $\mathbf{M} \in M_n(F)$ and a diagonal $\mathbf{D} = \operatorname{diag}(d_1, \dots, d_n)$ whose entries are **real and strictly positive**, with $\mathbf{A} = \mathbf{M}\mathbf{D}\mathbf{M}^*$.

Proof. Induction on n. For $n = 1$, the single entry is $a_{11} = \langle \mathbf{A}\mathbf{e}_1, \mathbf{e}_1 \rangle > 0$ by (P2) of Definition 13.1.1 in its strict form, so $\mathbf{M} = (1)$ and $\mathbf{D} = (a_{11})$ will do.

Let $n \geq 2$. Since $\mathbf{A}^* = \mathbf{A}$, we may write

$$\mathbf{A} = \begin{pmatrix} \alpha & \mathbf{c}^* \\ \mathbf{c} & \mathbf{A}' \end{pmatrix}, \qquad \alpha = a_{11}, \quad \mathbf{c} \in F^{n-1}, \quad \mathbf{A}' \in M_{n-1}(F) \text{ Hermitian.}$$

Here $\alpha = \langle \mathbf{A}\mathbf{e}_1, \mathbf{e}_1 \rangle > 0$, so α^{-1} exists and is a positive real number. Put

$$\mathbf{S} \coloneqq \mathbf{A}' - \alpha^{-1} \mathbf{c}\mathbf{c}^*,$$

which is Hermitian, since $(\mathbf{c}\mathbf{c}^*)^* = \mathbf{c}\mathbf{c}^*$ and α is real.

The smaller matrix is again positive definite. Let $\mathbf{y} \in F^{n-1}$ with $\mathbf{y} \neq \mathbf{0}$, set $t = -\alpha^{-1} \mathbf{c}^* \mathbf{y}$ and $\mathbf{x} = (t, \mathbf{y}) \in F^n$, which is nonzero because $\mathbf{y}$ is. Multiplying out in blocks (Theorem 8.1.1),

$$\langle \mathbf{A}\mathbf{x}, \mathbf{x} \rangle = \mathbf{x}^* \mathbf{A} \mathbf{x} = \alpha |t|^2 + \bar{t}\, \mathbf{c}^* \mathbf{y} + \mathbf{y}^* \mathbf{c}\, t + \mathbf{y}^* \mathbf{A}' \mathbf{y}.$$

With $\mathbf{c}^*\mathbf{y} = -\alpha t$ and $\mathbf{y}^*\mathbf{c} = \overline{\mathbf{c}^*\mathbf{y}} = -\alpha\bar{t}$, the first three terms are

$$\alpha|t|^2 - \alpha|t|^2 - \alpha|t|^2 = -\alpha|t|^2 = -\alpha^{-1}|\mathbf{c}^*\mathbf{y}|^2,$$

so $\langle \mathbf{Ax}, \mathbf{x}\rangle = \mathbf{y}^*(\mathbf{A}' - \alpha^{-1}\mathbf{cc}^*)\mathbf{y} = \langle \mathbf{Sy}, \mathbf{y}\rangle$. The left side is > 0 because $\mathbf{A} \succ \mathbf{O}$ and $\mathbf{x} \neq \mathbf{0}$. Hence $\mathbf{S} \succ \mathbf{O}$.

The induction step. By the induction hypothesis, $\mathbf{S} = \mathbf{M}'\mathbf{D}'(\mathbf{M}')^*$ with $\mathbf{M}' \in M_{n-1}(F)$ unit lower triangular and $\mathbf{D}' = \operatorname{diag}(d_2, \ldots, d_n)$ real with positive entries. Put

$$\mathbf{M} = \begin{pmatrix} 1 & \mathbf{0}^\top \\ \alpha^{-1}\mathbf{c} & \mathbf{M}' \end{pmatrix}, \qquad \mathbf{D} = \begin{pmatrix} \alpha & \mathbf{0}^\top \\ \mathbf{0} & \mathbf{D}' \end{pmatrix}.$$

Then $\mathbf{M}$ is unit lower triangular and $\mathbf{D}$ is diagonal with the positive real entries $\alpha, d_2, \ldots, d_n$. Block multiplication (Theorem 8.1.1) gives

$$\mathbf{MDM}^* = \begin{pmatrix} \alpha & \mathbf{0}^\top \\ \mathbf{c} & \mathbf{M}'\mathbf{D}' \end{pmatrix} \begin{pmatrix} 1 & \alpha^{-1}\mathbf{c}^* \\ \mathbf{0} & (\mathbf{M}')^* \end{pmatrix} = \begin{pmatrix} \alpha & \mathbf{c}^* \\ \mathbf{c} & \alpha^{-1}\mathbf{cc}^* + \mathbf{M}'\mathbf{D}'(\mathbf{M}')^* \end{pmatrix},$$

and $\alpha^{-1}\mathbf{cc}^* + \mathbf{M}'\mathbf{D}'(\mathbf{M}')^* = \alpha^{-1}\mathbf{cc}^* + \mathbf{S} = \mathbf{A}'$. So $\mathbf{MDM}^* = \mathbf{A}$, which proves the claim.

■

The claim is one pass of elimination performed symmetrically, and $d_1 = \alpha, d_2, \ldots, d_n$ are its pivots. Over $\mathbb{R}$ it is the $\text{Ł}\mathbf{D}\text{Ł}^\top$ factorization of Theorem 3.6.8 with the hypothesis that makes every pivot positive.

Write $d_1, \ldots, d_n$ for the diagonal entries of $\mathbf{D}$, all of them positive real numbers. Then $\mathbf{D}^{1/2} \coloneqq \operatorname{diag}(\sqrt{d_1}, \ldots, \sqrt{d_n})$ is a real matrix, and putting

$$\text{Ł} = \mathbf{MD}^{1/2}$$

gives a lower triangular matrix — its (i,j)-entry is $m_{ij}\sqrt{d_j}$, which vanishes for $i < j$ — whose k-th diagonal entry is $\sqrt{d_k} > 0$, and

$$\text{Ł}\text{Ł}^* = \mathbf{MD}^{1/2}(\mathbf{D}^{1/2})^*\mathbf{M}^* = \mathbf{MDM}^* = \mathbf{A}.$$

Uniqueness. Suppose $\mathbf{A} = \text{Ł}_1\text{Ł}_1^* = \text{Ł}_2\text{Ł}_2^*$, with each Ł_i lower triangular with positive diagonal entries. Split off the diagonal: write $\mathbf{\Delta}_i = \operatorname{diag}((\text{Ł}_i)_{11}, \ldots, (\text{Ł}_i)_{nn})$, a real diagonal matrix with positive entries, and $\mathbf{M}_i = \text{Ł}_i\mathbf{\Delta}_i^{-1}$, which is lower triangular with every diagonal entry 1. Then $\text{Ł}_i = \mathbf{M}_i\mathbf{\Delta}_i$ and

$$\mathbf{A} = \mathbf{M}_i\mathbf{\Delta}_i^2\mathbf{M}_i^* = \mathbf{M}_i(\mathbf{\Delta}_i^2\mathbf{M}_i^*).$$

The second factor is upper triangular, its (k,l)-entry being $(\mathbf{\Delta}_i^2)_{kk}\overline{(\mathbf{M}_i)_{lk}}$, which vanishes for $k > l$. So for $i = 1$ and $i = 2$ this displays an LU factorization of $\mathbf{A}$, which is invertible: Ł_1 is triangular with nonzero diagonal, hence invertible by Lemma 3.6.2, and $\mathbf{A} = \text{Ł}_1\text{Ł}_1^*$ is a product of two invertible matrices. By Theorem 3.6.5 the two agree:

$$\mathbf{M}_1 = \mathbf{M}_2 \qquad \text{and} \qquad \mathbf{\Delta}_1^2\mathbf{M}_1^* = \mathbf{\Delta}_2^2\mathbf{M}_2^*.$$

Canceling the invertible $\mathbf{M}_1^* = \mathbf{M}_2^*$ gives $\mathbf{\Delta}_1^2 = \mathbf{\Delta}_2^2$, and since both have positive diagonal entries, $\mathbf{\Delta}_1 = \mathbf{\Delta}_2$. Hence $\text{Ł}_1 = \mathbf{M}_1\mathbf{\Delta}_1 = \mathbf{M}_2\mathbf{\Delta}_2 = \text{Ł}_2$. This proves the theorem.

■

Remark. The pivots are determinant ratios. For $i, j \le k$, a term of $(\mathbf{MDM}^*)_{ij} = \sum_l m_{il}(\mathbf{DM}^*)_{lj}$ with $l > k$ has $l > i$ and so $m_{il} = 0$; the sum therefore stops at $l = k$, which says that the leading principal submatrices multiply: $\mathbf{A}_k = \mathbf{M}_k\mathbf{D}_k(\mathbf{M}^*)_k = \mathbf{M}_k\mathbf{D}_k\mathbf{M}_k^*$. Taking determinants with Theorem 7.5.4 and Theorem 7.4.6 gives $\det \mathbf{A}_k = d_1 \cdots d_k$, so $d_1 = \det \mathbf{A}_1$ and $d_k = \det \mathbf{A}_k / \det \mathbf{A}_{k-1}$ for $k \ge 2$. Read the other way, this is Sylvester's criterion (Theorem 13.1.5 (d)) appearing inside the elimination: the leading minors are positive exactly when the pivots are.

Reading the equation $\mathbf{A} = Ł Ł^*$ entry by entry turns it into a recipe, and the recipe is the standard way to factor by hand. Comparing the (j, j) entry of both sides gives $a_{jj} = \sum_{k \le j} |\ell_{jk}|^2$, and comparing the (i, j) entry for $i > j$ gives $a_{ij} = \sum_{k \le j} \ell_{ik}\overline{\ell_{jk}}$. Solving each for its one new unknown, column by column:

$$\ell_{jj} = \Big(a_{jj} - \textstyle\sum_{k<j} |\ell_{jk}|^2\Big)^{1/2},$$
$$\ell_{ij} = \Big(a_{ij} - \textstyle\sum_{k<j} \ell_{ik}\overline{\ell_{jk}}\Big)/\ell_{jj} \qquad (i > j).$$

Example. A Cholesky Factorization, and One That Fails

Factor

$$\mathbf{A} = \begin{pmatrix} 4 & -2 & 6 \\ -2 & 5 & -1 \\ 6 & -1 & 26 \end{pmatrix}$$

as $Ł Ł^{\top}$, and run the same algorithm on the matrix $\mathbf{A}'$ obtained from $\mathbf{A}$ by changing the $(3, 3)$ entry to 9.

Solution. Column 1. $\ell_{11} = \sqrt{4} = 2$; then $\ell_{21} = -2/2 = -1$ and $\ell_{31} = 6/2 = 3$.

Column 2. $\ell_{22} = \sqrt{5 - (-1)^2} = \sqrt{4} = 2$; then

$$\ell_{32} = \frac{a_{32} - \ell_{31}\ell_{21}}{\ell_{22}} = \frac{-1 - 3(-1)}{2} = 1.$$

Column 3. $\ell_{33} = \sqrt{26 - 3^2 - 1^2} = \sqrt{16} = 4$. So

$$Ł = \begin{pmatrix} 2 & 0 & 0 \\ -1 & 2 & 0 \\ 3 & 1 & 4 \end{pmatrix},$$

and multiplying back confirms $Ł Ł^{\top} = \mathbf{A}$. The diagonal entries $2, 2, 4$ are positive, so $\mathbf{A} \succ \mathbf{O}$ by Theorem 13.2.3; their squares $4, 4, 16$ are the pivots $d_k = \det \mathbf{A}_k / \det \mathbf{A}_{k-1}$, consistent with $\det \mathbf{A}_1 = 4$, $\det \mathbf{A}_2 = 16$, $\det \mathbf{A} = 256$.

The failure. For $\mathbf{A}'$, columns 1 and 2 are unchanged, and the last step asks for

$$\ell_{33} = \sqrt{9 - 3^2 - 1^2} = \sqrt{-1},$$

which is not a positive real number. The algorithm stops, and by Theorem 13.2.3 its stopping is a proof: $\mathbf{A}'$ is not positive definite. Indeed $\det \mathbf{A}' = -16 < 0$, so Sylvester's criterion fails at $k = 3$.

The example shows the practical point. **The algorithm is the test.** To decide whether a numerical Hermitian matrix is positive definite, one does not compute eigenvalues and one does not compute n determinants; one starts the Cholesky recursion and watches whether every quantity under a square root stays positive. Sylvester's criterion guarantees that this verdict is correct, because $d_1 \cdots d_k = \det \mathbf{A}_k$.

Remark. Two remarks on cost, both of which belong properly to Chapter 24. First, Cholesky does about half the arithmetic of a general LU factorization, because it computes only Ł and never the second triangular factor. Second, and less obvious, it needs no pivoting: the recursion above never divides by something that positive definiteness allows to be zero, so the row swaps that Chapter 3 introduced for stability are simply not required here. Positive definiteness is the one structural hypothesis under which elimination is unconditionally safe.

Warning. **Cholesky needs definiteness, not just semidefiniteness, for uniqueness.** The matrix $\mathbf{A} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$ is positive semidefinite and does factor as $Ł Ł^{\top}$, for instance with $Ł = \mathbf{A}$. But the diagonal entry $\ell_{11} = 0$ is not positive, so Theorem 13.2.3 does not apply — and indeed uniqueness fails, since $\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$ and $\begin{pmatrix} 0 & 0 \\ 0 & -1 \end{pmatrix}$ are two different lower triangular matrices Ł with $Ł Ł^{\top} = \mathbf{A}$. Also note that the Ł of Cholesky is **not** the Ł of Chapter 3's LU factorization: the latter is unit lower triangular, and the two differ by the factor $\mathbf{D}^{1/2}$.

Quick check. Let $\mathbf{A} \succ \mathbf{O}$ with Cholesky factor Ł. What is $\det \mathbf{A}$ in terms of Ł, and what is the Cholesky factor of $\mathbf{A}^{-1}$?

Solution. $\det \mathbf{A} = \det Ł \det Ł^{*} = |\ell_{11} \cdots \ell_{nn}|^2 = (\ell_{11} \cdots \ell_{nn})^2$, by Theorem 7.5.4 and Theorem 7.4.6, the last step because the ℓ_{kk} are positive reals. For the inverse, $\mathbf{A}^{-1} = (Ł Ł^{*})^{-1} = (Ł^{*})^{-1} Ł^{-1} = \mathbf{M}\mathbf{M}^{*}$ with $\mathbf{M} = (Ł^{*})^{-1}$, which is **upper** triangular; so this is not itself a Cholesky factorization. The Cholesky factor of $\mathbf{A}^{-1}$ is a different matrix, which is why solving $\mathbf{A}\mathbf{x} = \mathbf{b}$ is done by two triangular substitutions with Ł rather than by inverting anything.

13.2.4 Exercises

13.2.4.1 A. Check your understanding

Exercise. A1

(a) State what $\mathbf{A}^{1/2}$ is, including the two conditions that pin it down.

(b) Define $|\mathbf{A}|$ for $\mathbf{A} \in M_{m\times n}(F)$ and say what size it is.

(c) State the Cholesky factorization theorem.

(d) True or false: every Hermitian matrix has exactly one Hermitian square root. Justify your answer.

(e) Explain in one sentence why the Cholesky algorithm is a test for positive definiteness.

Solution.

(a) For $\mathbf{A} \succeq \mathbf{O}$, $\mathbf{A}^{1/2}$ is the unique matrix $\mathbf{C}$ with $\mathbf{C} \succeq \mathbf{O}$ and $\mathbf{C}^2 = \mathbf{A}$ (Theorem 13.2.1). Both conditions are needed: positivity, and squaring to $\mathbf{A}$.

(b) $|\mathbf{A}| = (\mathbf{A}^*\mathbf{A})^{1/2}$, which is $n \times n$, the size of the domain of $\mathbf{A}$ (Definition 13.2.1).

(c) A Hermitian $\mathbf{A}$ is positive definite if and only if $\mathbf{A} = Ł Ł^*$ with Ł lower triangular with positive real diagonal entries, and then Ł is unique (Theorem 13.2.3).

(d) False, twice over. A Hermitian matrix with a negative eigenvalue has no Hermitian square root at all, since $\mathbf{C}^2 \succeq \mathbf{O}$ always. And a positive semidefinite $\mathbf{A}$ with two distinct eigenvalues has at least two Hermitian square roots, $\pm\mathbf{A}^{1/2}$. Uniqueness needs the root itself to be positive semidefinite.

(e) Because the algorithm computes $\ell_{jj}^2 = \det \mathbf{A}_j / \det \mathbf{A}_{j-1}$ one step at a time, so it produces a positive real ℓ_{jj} at every step exactly when Sylvester's criterion holds.

13.2.4.2 B. Practice

Exercise. B1: Square roots

Compute $\mathbf{A}^{1/2}$ for each of the following, and in each case give a polynomial p with $p(\mathbf{A}) = \mathbf{A}^{1/2}$.

(a) $\mathbf{A} = \begin{pmatrix} 5 & 4 \\ 4 & 5 \end{pmatrix}$.

(b) $\mathbf{A} = \begin{pmatrix} 10 & -6 \\ -6 & 10 \end{pmatrix}$.

Solution.

(a) $\mathbf{A}(1,1) = (9,9)$ and $\mathbf{A}(1,-1) = (1,-1)$, so the eigenvalues are 9 and 1 with eigenlines $\operatorname{span}((1,1))$ and $\operatorname{span}((1,-1))$; both positive, so $\mathbf{A} \succ \mathbf{O}$. With $\mathbf{P}_1, \mathbf{P}_2$ the projections onto those lines,

$$\mathbf{A}^{1/2} = 3\mathbf{P}_1 + \mathbf{P}_2 = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}.$$

For the polynomial, solve $p(9) = 3$, $p(1) = 1$: $p(x) = \frac{1}{4}x + \frac{3}{4}$, and $\frac{1}{4}\mathbf{A} + \frac{3}{4}\mathbf{I} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$.

Check: $\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}^2 = \begin{pmatrix} 5 & 4 \\ 4 & 5 \end{pmatrix}$.

(b) $\mathbf{A}(1,-1) = (16,-16)$ and $\mathbf{A}(1,1) = (4,4)$, so the eigenvalues are 16 and 4. Then

$$\mathbf{A}^{1/2} = 4 \cdot \tfrac{1}{2}\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} + 2 \cdot \tfrac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 3 & -1 \\ -1 & 3 \end{pmatrix}.$$

For the polynomial, $p(16) = 4$ and $p(4) = 2$ give $p(x) = \frac{1}{6}x + \frac{4}{3}$, and $\frac{1}{6}\mathbf{A} + \frac{4}{3}\mathbf{I} = \begin{pmatrix} 3 & -1 \\ -1 & 3 \end{pmatrix}$. Check: $\begin{pmatrix} 3 & -1 \\ -1 & 3 \end{pmatrix}^2 = \begin{pmatrix} 10 & -6 \\ -6 & 10 \end{pmatrix}$.

Exercise. B2: Cholesky, and a planted failure

(a) Find the Cholesky factor of $\mathbf{A} = \begin{pmatrix} 1 & 2 & -1 \\ 2 & 5 & 1 \\ -1 & 1 & 14 \end{pmatrix}$, and hence write down $\det \mathbf{A}$.

(b) Run the same algorithm on $\mathbf{A}' = \begin{pmatrix} 1 & 2 & -1 \\ 2 & 5 & 1 \\ -1 & 1 & 9 \end{pmatrix}$ and say what the outcome proves.

Solution.

(a) $\ell_{11} = \sqrt{1} = 1$, $\ell_{21} = 2$, $\ell_{31} = -1$. Then $\ell_{22} = \sqrt{5-4} = 1$ and

$$\ell_{32} = \frac{1 - (-1)(2)}{1} = 3.$$

Finally $\ell_{33} = \sqrt{14 - 1 - 9} = \sqrt{4} = 2$. So

$$Ł = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ -1 & 3 & 2 \end{pmatrix},$$

and $Ł Ł^{\top} = \mathbf{A}$. All diagonal entries are positive, so $\mathbf{A} \succ \mathbf{O}$. By the Quick check above, $\det \mathbf{A} = (1 \cdot 1 \cdot 2)^2 = 4$.

(b) Columns 1 and 2 are identical, and the last step asks for $\ell_{33} = \sqrt{9 - 1 - 9} = \sqrt{-1}$. The algorithm fails, so $\mathbf{A}'$ is not positive definite (Theorem 13.2.3). Confirming by Sylvester: $\det \mathbf{A}'_1 = 1$, $\det \mathbf{A}'_2 = 1$, and $\det \mathbf{A}' = 1 \cdot 44 - 2 \cdot 19 - 7 = -1 < 0$.

Exercise. B3: An absolute value

Let $\mathbf{A} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \in M_2(\mathbb{R})$. Compute $|\mathbf{A}|$, verify $\||\mathbf{A}|\mathbf{x}\| = \|\mathbf{A}\mathbf{x}\|$ directly, and determine whether $|\mathbf{A}| = \mathbf{A}$.

Solution. $\mathbf{A}^{\top}\mathbf{A} = \begin{pmatrix} 2 & 2 \\ 2 & 2 \end{pmatrix}$, whose eigenvalues are 4 and 0 with eigenlines $\operatorname{span}((1,1))$ and

$\operatorname{span}((1,-1))$. Hence

$$|\mathbf{A}| = 2 \cdot \tfrac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} + \mathbf{O} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} = \mathbf{A}.$$

Directly: $\mathbf{A}\mathbf{x} = (x_1+x_2)(1,1)$, so $\|\mathbf{A}\mathbf{x}\| = \sqrt{2}\,|x_1+x_2|$, and the same computation applies to $|\mathbf{A}| = \mathbf{A}$. The equality $|\mathbf{A}| = \mathbf{A}$ is no accident: $\mathbf{A} \succeq \mathbf{O}$ here, and for a positive semidefinite $\mathbf{A}$ one has $\mathbf{A}^*\mathbf{A} = \mathbf{A}^2$, so $|\mathbf{A}| = (\mathbf{A}^2)^{1/2} = \mathbf{A}$ by uniqueness in Theorem 13.2.1.

13.2.4.3 C. Going deeper

Exercise. C1: The trace of a product of positive matrices

Let $\mathbf{A}, \mathbf{B} \in M_n(F)$ with $\mathbf{A} \succeq \mathbf{O}$ and $\mathbf{B} \succeq \mathbf{O}$.

(a) Prove that $\operatorname{tr}(\mathbf{AB}) = \operatorname{tr}(\mathbf{A}^{1/2}\mathbf{B}\mathbf{A}^{1/2})$, and deduce that $\operatorname{tr}(\mathbf{AB}) \geq 0$.

(b) Prove that $\operatorname{tr}(\mathbf{AB}) = 0$ if and only if $\mathbf{AB} = \mathbf{O}$.

Hint for (b): a positive semidefinite matrix of trace 0 is $\mathbf{O}$.

Solution.

(a) By Theorem 0.9.6, $\operatorname{tr}(\mathbf{XY}) = \operatorname{tr}(\mathbf{YX})$. With $\mathbf{X} = \mathbf{A}^{1/2}$ and $\mathbf{Y} = \mathbf{A}^{1/2}\mathbf{B}$,

$$\operatorname{tr}(\mathbf{AB}) = \operatorname{tr}(\mathbf{A}^{1/2}\mathbf{A}^{1/2}\mathbf{B}) = \operatorname{tr}(\mathbf{A}^{1/2}\mathbf{B}\mathbf{A}^{1/2}).$$

Now $\mathbf{A}^{1/2}\mathbf{B}\mathbf{A}^{1/2} = (\mathbf{A}^{1/2})^*\mathbf{B}\mathbf{A}^{1/2} \succeq \mathbf{O}$ by Proposition 13.1.2 (a), since $\mathbf{A}^{1/2}$ is Hermitian. The trace of a positive semidefinite matrix is ≥ 0 by Exercise (C1: Zero trace forces zero) (a). Hence $\operatorname{tr}(\mathbf{AB}) \geq 0$.

(b) ($\Leftarrow$) is clear. ($\Rightarrow$) Suppose $\operatorname{tr}(\mathbf{AB}) = 0$. By (a), $\mathbf{C} \coloneqq \mathbf{A}^{1/2}\mathbf{B}\mathbf{A}^{1/2} \succeq \mathbf{O}$ and $\operatorname{tr}\mathbf{C} = 0$, so $\mathbf{C} = \mathbf{O}$ by Exercise (C1: Zero trace forces zero) (b). Put $\mathbf{M} = \mathbf{B}^{1/2}\mathbf{A}^{1/2}$. Then

$$\mathbf{M}^*\mathbf{M} = \mathbf{A}^{1/2}\mathbf{B}^{1/2}\mathbf{B}^{1/2}\mathbf{A}^{1/2} = \mathbf{A}^{1/2}\mathbf{B}\mathbf{A}^{1/2} = \mathbf{C} = \mathbf{O},$$

so $\mathbf{M} = \mathbf{O}$ by Lemma 11.4.3 (its null space is all of F^n, hence so is $\operatorname{nul}\mathbf{M}$). Therefore

$$\mathbf{AB} = \mathbf{A}^{1/2}(\mathbf{A}^{1/2}\mathbf{B}^{1/2})\mathbf{B}^{1/2} = \mathbf{A}^{1/2}\mathbf{M}^*\mathbf{B}^{1/2} = \mathbf{O}.$$

Exercise. C2: Commuting positive matrices

Let $\mathbf{A} \in M_n(F)$ with $\mathbf{A} \succeq \mathbf{O}$.

(a) Prove **without** using Theorem 13.2.1 (a) that if $\mathbf{BA} = \mathbf{AB}$, then $\mathbf{B}\mathbf{A}^{1/2} = \mathbf{A}^{1/2}\mathbf{B}$. *Hint: show that* $\mathbf{B}$ *maps each eigenspace of* $\mathbf{A}$ *into itself.*

(b) Let $\mathbf{B} \succeq \mathbf{O}$ with $\mathbf{AB} = \mathbf{BA}$. Prove that $\mathbf{AB} \succeq \mathbf{O}$ and that $(\mathbf{AB})^{1/2} = \mathbf{A}^{1/2}\mathbf{B}^{1/2}$.

Solution.

(a) Let $\lambda \in \operatorname{spec}(\mathbf{A})$ and $\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$. Then $\mathbf{A}(\mathbf{B}\mathbf{v}) = \mathbf{B}(\mathbf{A}\mathbf{v}) = \lambda\mathbf{B}\mathbf{v}$, so $\mathbf{B}$ maps $E_\lambda(\mathbf{A})$ into itself. Since $\mathbf{A} \succeq \mathbf{O}$, its distinct eigenvalues $\lambda_1, \ldots, \lambda_k$ are nonnegative reals,

and $\mathbf{A}^{1/2} = \sum_i \sqrt{\lambda_i}\,\mathbf{P}_i$ with $\mathbf{P}_i$ the orthogonal projection onto $E_{\lambda_i}(\mathbf{A})$ (proof of Theorem 13.2.1). For $\mathbf{v} \in E_{\lambda_j}(\mathbf{A})$ we have $\mathbf{P}_i\mathbf{v} = \mathbf{0}$ for $i \neq j$ and $\mathbf{P}_j\mathbf{v} = \mathbf{v}$, so $\mathbf{A}^{1/2}\mathbf{v} = \sqrt{\lambda_j}\,\mathbf{v}$. Now $F^n = E_{\lambda_1}(\mathbf{A}) \oplus \cdots \oplus E_{\lambda_k}(\mathbf{A})$ (Theorem 12.8.1), so it suffices to check $\mathbf{B}\mathbf{A}^{1/2}\mathbf{v} = \mathbf{A}^{1/2}\mathbf{B}\mathbf{v}$ for $\mathbf{v} \in E_{\lambda_j}(\mathbf{A})$. The left side is $\sqrt{\lambda_j}\,\mathbf{B}\mathbf{v}$; the right side is $\mathbf{A}^{1/2}(\mathbf{B}\mathbf{v}) = \sqrt{\lambda_j}\,\mathbf{B}\mathbf{v}$ as well, because $\mathbf{B}\mathbf{v} \in E_{\lambda_j}(\mathbf{A})$ by the first sentence. Hence $\mathbf{B}\mathbf{A}^{1/2} = \mathbf{A}^{1/2}\mathbf{B}$.

(b) By (a), $\mathbf{A}^{1/2}$ commutes with $\mathbf{B}$. Applying (a) once more, this time to $\mathbf{B} \succeq \mathbf{O}$ with the commuting matrix $\mathbf{A}^{1/2}$, gives that $\mathbf{A}^{1/2}$ commutes with $\mathbf{B}^{1/2}$. Put $\mathbf{C} = \mathbf{A}^{1/2}\mathbf{B}^{1/2}$. Then

$$\mathbf{C}^* = (\mathbf{B}^{1/2})^*(\mathbf{A}^{1/2})^* = \mathbf{B}^{1/2}\mathbf{A}^{1/2} = \mathbf{C},$$

so $\mathbf{C}$ is Hermitian, and $\mathbf{C}^2 = \mathbf{A}^{1/2}\mathbf{B}^{1/2}\mathbf{A}^{1/2}\mathbf{B}^{1/2} = \mathbf{A}\mathbf{B}$ after one commutation. It remains to see $\mathbf{C} \succeq \mathbf{O}$. The matrices $\mathbf{A}^{1/2}$ and $\mathbf{B}^{1/2}$ are commuting Hermitian matrices, so by Corollary 12.7.2 (over $\mathbb{C}$) or Corollary 12.7.3 (over $\mathbb{R}$) there is a unitary $\mathbf{U}$ with $\mathbf{U}^*\mathbf{A}^{1/2}\mathbf{U} = \operatorname{diag}(\alpha_i)$ and $\mathbf{U}^*\mathbf{B}^{1/2}\mathbf{U} = \operatorname{diag}(\beta_i)$, all $\alpha_i, \beta_i \geq 0$. Then $\mathbf{U}^*\mathbf{C}\mathbf{U} = \operatorname{diag}(\alpha_i\beta_i) \succeq \mathbf{O}$, so $\mathbf{C} \succeq \mathbf{O}$ by Proposition 13.1.2 (a). Hence $\mathbf{A}\mathbf{B} = \mathbf{C}^2 \succeq \mathbf{O}$ and, by uniqueness in Theorem 13.2.1, $(\mathbf{A}\mathbf{B})^{1/2} = \mathbf{C} = \mathbf{A}^{1/2}\mathbf{B}^{1/2}$.

Exercise. C3: Why definiteness is needed

(a) Give a positive semidefinite $\mathbf{A} \in M_2(\mathbb{R})$, not positive definite, together with two different lower triangular Ł satisfying $\mathbf{A} = \text{Ł}\text{Ł}^{\top}$. Which hypothesis of Theorem 13.2.3 fails?

(b) Prove that if $\mathbf{A} \succeq \mathbf{O}$ then $\mathbf{A} + t\mathbf{I} \succ \mathbf{O}$ for every real $t > 0$, and hence that every positive semidefinite matrix is a limit of positive definite ones in the entrywise sense.

Solution.

(a) Take $\mathbf{A} = \operatorname{diag}(0, 1)$. Then $\text{Ł}_1 = \operatorname{diag}(0, 1)$ and $\text{Ł}_2 = \operatorname{diag}(0, -1)$ are both lower triangular with $\text{Ł}_i\text{Ł}_i^{\top} = \mathbf{A}$. The hypothesis that fails is that the diagonal entries of Ł be strictly positive: $\ell_{11} = 0$ here, which is exactly the first pivot vanishing. Uniqueness in Theorem 13.2.3 is not available without it.

(b) $\mathbf{A} + t\mathbf{I}$ is Hermitian, and for $\mathbf{x} \neq \mathbf{0}$,

$$\langle (\mathbf{A} + t\mathbf{I})\mathbf{x}, \mathbf{x}\rangle = \langle \mathbf{A}\mathbf{x}, \mathbf{x}\rangle + t\|\mathbf{x}\|^2 \geq t\|\mathbf{x}\|^2 > 0,$$

so $\mathbf{A} + t\mathbf{I} \succ \mathbf{O}$. Each entry of $\mathbf{A} + t\mathbf{I}$ differs from the corresponding entry of $\mathbf{A}$ by t or by 0, so letting t decrease to 0 recovers $\mathbf{A}$ entry by entry. This is the standard device for extending a statement from positive definite to positive semidefinite matrices whenever the statement survives taking limits; Theorem 13.1.4 already used it in the form $\det(\mathbf{A} + t\mathbf{I}) > 0$.

13.3 Gram Matrices

Chapter 11 attached to every list of vectors the matrix of its pairwise inner products and read three facts off it: it is Hermitian, its quadratic form is a squared length, and it is invertible exactly when the list is independent (Definition 11.2.3, Theorem 11.2.5). One question was left

standing. Given a square matrix, is it the Gram matrix of *something*? The answer is the shortest possible one: exactly when it is positive semidefinite. This section proves that, and then spends it twice — the determinant of a Gram matrix is a squared volume, and a ratio of two Gram determinants is a distance to a subspace.

Throughout, $F = \mathbb{R}$ or $F = \mathbb{C}$ and all inner product spaces are over F. When the list must be named we write $\mathbf{G}(v_1, \dots, v_k)$ for its Gram matrix, whose entries are $\langle v_j, v_i \rangle$ as in Definition 11.2.3.

13.3.1 Which matrices are Gram matrices?

Half of the answer is already proved. By Theorem 11.2.5 (a) and (b), a Gram matrix $\mathbf{G}$ is Hermitian and satisfies

$$\mathbf{x}^* \mathbf{G} \mathbf{x} = \Big\| \sum_j x_j v_j \Big\|^2 \;\geq\; 0$$

for every $\mathbf{x} \in F^k$. That is word for word the definition of positive semidefiniteness (Definition 13.1.1). So being positive semidefinite is **necessary**. The content of the next theorem is that nothing else is needed, and this is where the chapter's second signature move earns its keep: a positive semidefinite matrix factors as $\mathbf{B}^*\mathbf{B}$, and the columns of $\mathbf{B}$ are then the list we were looking for.

Theorem 13.3.1. Gram Matrices Are Exactly the Positive Semidefinite Matrices

Let $\mathbf{G} \in M_k(F)$. The following are equivalent.

(a) $\mathbf{G}$ is the Gram matrix of some list of k vectors in some inner product space over F;

(b) $\mathbf{G}$ is the Gram matrix of some list of k vectors in F^r with the standard inner product, where $r = \operatorname{rank} \mathbf{G}$;

(c) $\mathbf{G} \succeq \mathbf{O}$.

Moreover the list is unique up to an isometry: if $(v_1, \dots, v_k)$ in V and $(w_1, \dots, w_k)$ in W have the same Gram matrix, then there is an isometry from $\operatorname{span}(v_1, \dots, v_k)$ onto $\operatorname{span}(w_1, \dots, w_k)$ carrying v_j to w_j for every j.

Idea.

(a) $\Rightarrow$ (c) is the display above. For (c) $\Rightarrow$ (b), diagonalize: the spectral theorem writes $\mathbf{G} = \mathbf{U}\mathbf{D}\mathbf{U}^*$ with $\mathbf{D}$ diagonal and, by Theorem 13.1.4, with nonnegative diagonal. Then $\mathbf{D}^{1/2}\mathbf{U}^*$ is a square root of $\mathbf{G}$ in the $\mathbf{B}^*\mathbf{B}$ sense, and its rows beyond the r-th are zero, so the columns already live in F^r. For the uniqueness clause, the Gram matrix *is* the recipe for every inner product among the vectors, so the obvious map $\sum a_j v_j \mapsto \sum a_j w_j$ has no choice but to preserve them; the only thing needing care is that it is well defined, and part (b) of Theorem 11.2.5 says precisely that.

Proof.

(b) $\Rightarrow$ (a) is immediate, since F^r with the standard inner product is an inner product

space.

(c) $\Rightarrow$ (c). Let $\mathbf{G}$ be the Gram matrix of $(v_1, \dots, v_k)$. By Theorem 11.2.5 (a), $\mathbf{G}^* = \mathbf{G}$, and by (b) of the same theorem $\mathbf{x}^*\mathbf{G}\mathbf{x} = \|\sum_j x_j v_j\|^2 \geq 0$ for every $\mathbf{x} \in F^k$. Hence $\mathbf{G} \succeq \mathbf{O}$ by Definition 13.1.1.

(d) $\Rightarrow$ (b). Since $\mathbf{G}$ is Hermitian, the spectral theorem for matrices gives a unitary $\mathbf{U}$ and a diagonal $\mathbf{D}$ with $\mathbf{G} = \mathbf{U}\mathbf{D}\mathbf{U}^*$ (Corollary 12.4.3 over $\mathbb{C}$; Corollary 12.5.2 over $\mathbb{R}$, with $\mathbf{U}$ orthogonal and $\mathbf{D}$ real). The diagonal entries of $\mathbf{D}$ are the eigenvalues of $\mathbf{G}$, so they are real and ≥ 0 by Theorem 13.1.4, and we may prescribe their order: let $d_1, \dots, d_r > 0$ and $d_{r+1} = \cdots = d_k = 0$. Here r is the number of nonzero eigenvalues, which is $\operatorname{rank} \mathbf{G}$, because $\mathbf{G}$ and $\mathbf{D}$ have the same rank.

Put $\mathbf{B} = \mathbf{D}^{1/2}\mathbf{U}^* \in M_k(F)$, where $\mathbf{D}^{1/2} = \operatorname{diag}(\sqrt{d_1}, \dots, \sqrt{d_k})$. Then

$$\mathbf{B}^*\mathbf{B} = \mathbf{U}\mathbf{D}^{1/2}\mathbf{D}^{1/2}\mathbf{U}^* = \mathbf{U}\mathbf{D}\mathbf{U}^* = \mathbf{G},$$

using $(\mathbf{D}^{1/2})^* = \mathbf{D}^{1/2}$. Rows $r+1, \dots, k$ of $\mathbf{B}$ are zero, since row p of $\mathbf{D}^{1/2}\mathbf{U}^*$ is $\sqrt{d_p}$ times row p of $\mathbf{U}^*$. Let $\mathbf{B}' \in M_{r\times k}(F)$ consist of the first r rows of $\mathbf{B}$; deleting zero rows changes no entry of $\mathbf{B}^*\mathbf{B}$, so $(\mathbf{B}')^*\mathbf{B}' = \mathbf{G}$ as well. Let $\mathbf{v}_1, \dots, \mathbf{v}_k \in F^r$ be the columns of $\mathbf{B}'$. With the standard inner product $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{y}^*\mathbf{x}$ on F^r,

$$\langle \mathbf{v}_j, \mathbf{v}_i \rangle = \mathbf{v}_i^*\mathbf{v}_j = \big((\mathbf{B}')^*\mathbf{B}'\big)_{ij} = (\mathbf{G})_{ij},$$

so $\mathbf{G}$ is the Gram matrix of $(\mathbf{v}_1, \dots, \mathbf{v}_k)$.

The uniqueness clause. Let $(v_1, \dots, v_k)$ in V and $(w_1, \dots, w_k)$ in W both have Gram matrix $\mathbf{G}$, and put $U = \operatorname{span}(v_1, \dots, v_k)$ and $U' = \operatorname{span}(w_1, \dots, w_k)$. Define $T\colon U \to U'$ by

$$T\Big(\sum_j a_j v_j\Big) = \sum_j a_j w_j.$$

This is well defined: if $\sum_j a_j v_j = \sum_j b_j v_j$, let $\mathbf{x} \in F^k$ be the vector with entries $x_j = a_j - b_j$; then Theorem 11.2.5 (b), applied to each list in turn, gives

$$\Big\| \sum_j x_j w_j \Big\|^2 = \mathbf{x}^*\mathbf{G}\mathbf{x} = \Big\| \sum_j x_j v_j \Big\|^2 = 0,$$

so $\sum_j a_j w_j = \sum_j b_j w_j$. The same display with every $b_j = 0$ shows $\|Tu\| = \|u\|$ for every $u \in U$, so T is an isometry (Definition 11.7.1); it is linear and surjective by its definition on spanning vectors. This proves the theorem.

■

Three consequences are worth naming at once.

First, the corollary promised in Chapter 7 is now proved from the right principle. Corollary 7.7.3 said that $\det(\mathbf{A}\mathbf{A}^{\top}) \geq 0$ for real $\mathbf{A}$, with equality exactly when the rows of $\mathbf{A}$ are dependent, and it proved this by writing the determinant as a sum of squares of minors. The structural reason is the theorem above: $\mathbf{A}\mathbf{A}^{\top}$ is a Gram matrix, hence positive semidefinite, hence has nonnegative eigenvalues, and its determinant is their product. Positivity of the determinant is not an accident of Cauchy–Binet; it is one shadow of a much stronger property, and the other shadows come free — **every** principal minor of a Gram matrix is ≥ 0 too, since a principal submatrix of $\mathbf{G}(v_1, \dots, v_k)$ is the Gram matrix of a sublist.

Second, the smallest space that can hold a realization of $\mathbf{G}$ has dimension rank $\mathbf{G}$, and no smaller one will do: a list of k vectors in a space of dimension s has a Gram matrix of rank at most s, because $\mathbf{G} = \mathbf{A}^*\mathbf{A}$ for the $s \times k$ coordinate matrix $\mathbf{A}$ of the list and $\operatorname{rank}(\mathbf{A}^*\mathbf{A}) = \operatorname{rank}\mathbf{A} \le s$ by Corollary 11.6.6 (b).

Third, the theorem is a free source of examples in both directions: any vectors at all produce a positive semidefinite table, and conversely positivity is a real constraint on a table of numbers — a stronger one than most readers expect.

Warning. **Plausible-looking symmetric tables need not be Gram matrices.** The condition is not "symmetric with nonnegative diagonal", and it is not even "every 2×2 principal minor is ≥ 0". Take

$$\mathbf{G} = \begin{pmatrix} 5 & -3 & -3 \\ -3 & 5 & -3 \\ -3 & -3 & 5 \end{pmatrix}.$$

Its three 2×2 principal minors are all $25 - 9 = 16 > 0$, so any two of the three vectors would be a legitimate pair. But $\mathbf{G}(1,1,1)^{\top} = (-1,-1,-1)^{\top}$, so -1 is an eigenvalue and $\mathbf{G} \not\succeq \mathbf{O}$: no three vectors in any inner product space have this table. In words, three vectors of squared length 5 cannot be **pairwise** at an angle with cosine $-3/5$; two of them can.

Quick check. Is the $k \times k$ matrix with every entry equal to 1 a Gram matrix? If so, exhibit a list; if not, say which clause fails.

Solution. It is. Take $v_1 = \cdots = v_k = v$ for any unit vector v: then $\langle v_j, v_i \rangle = 1$ for all i, j. Consistently, the matrix has rank 1, and Theorem 13.3.1 (b) predicts a realization in F^1; the list $(1, 1, \ldots, 1)$ in F is one.

13.3.2 The Gram determinant is a squared volume

Chapter 7 defined the volume of a parallelepiped spanned by n vectors of $\mathbb{R}^n$ as $|\det|$ of the matrix of their coordinates (Definition 7.9.1). That definition needs a **square** matrix, so it says nothing about two vectors in $\mathbb{R}^3$. The missing notion was already present in Chapter 7's motivation, though: volume is base times height. Turning that into a definition costs one line, now that we have distances to subspaces.

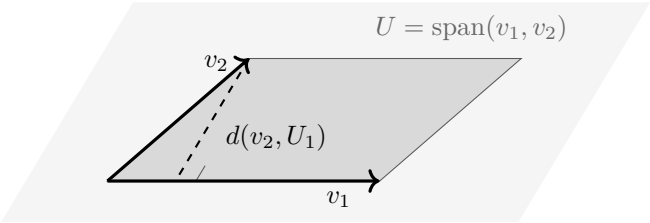
Definition 13.3.1. Volume of a list

Let $(v_1, \ldots, v_k)$ be a list in an inner product space V over F. Its **k-dimensional volume** $\operatorname{vol}(v_1, \ldots, v_k) \ge 0$ is defined recursively by

$$\operatorname{vol}(v_1) = \|v_1\|, \qquad \operatorname{vol}(v_1, \ldots, v_k) = \operatorname{vol}(v_1, \ldots, v_{k-1}) \cdot d(v_k, U_{k-1}),$$

where $U_{k-1} = \operatorname{span}(v_1, \ldots, v_{k-1})$ and $d(\cdot,\cdot)$ is the distance of Definition 11.3.3.

In words: the volume of the last vector's parallelepiped is the volume of the one below it, times the height of the last vector above the subspace the others span. For $k = 2$ this is base times height for a parallelogram, and it is 0 exactly when v_2 lies on the line through v_1.



base times height: $\operatorname{vol}(v_1, v_2) = \|v_1\|\, d(v_2, U_1)$

The recursion is easy to state and awkward to compute with, because each step needs a projection. The Gram determinant computes all of it at once.

Lemma 13.3.2. Bordering a Gram matrix

Let $(v_1, \dots, v_k)$ be a list in an inner product space V, let $U = \operatorname{span}(v_1, \dots, v_k)$, and let $w \in V$. Then

$$\det \mathbf{G}(v_1, \dots, v_k, w) = \det \mathbf{G}(v_1, \dots, v_k) \cdot d(w, U)^2.$$

Idea. Replace w by $w - P_U w$. That is a column operation on the *list*, and a column operation on the list is a congruence $\mathbf{G} \mapsto \mathbf{E}^* \mathbf{G} \mathbf{E}$ on the Gram matrix, with $\mathbf{E}$ unitriangular, so the determinant does not move. After the replacement the last vector is orthogonal to all the others, and the Gram matrix has become block diagonal.

Proof. First record how the Gram matrix responds to a change of list. If $u_j = \sum_t e_{tj} v_t$ for a matrix $\mathbf{E} = (e_{tj})$, then by sesquilinearity

$$\langle u_j, u_i \rangle = \sum_t \sum_s e_{tj} \overline{e_{si}} \langle v_t, v_s \rangle,$$

which is the (i, j)-entry of $\mathbf{E}^* \mathbf{G} \mathbf{E}$; that is, $\mathbf{G}(u_1, \dots, u_m) = \mathbf{E}^* \mathbf{G}(v_1, \dots, v_k)\mathbf{E}$.

Now write $w = p + h$ with $p = P_U w \in U$ and $h = w - p \in U^\perp$, which is possible by Theorem 11.3.2 (a). Since $p \in U$, there are scalars $c_1, \dots, c_k$ with $p = \sum_t c_t v_t$. Let $\mathbf{E} \in M_{k+1}(F)$ be the identity matrix with its last column replaced by $(-c_1, \dots, -c_k, 1)$. Applying the rule above to the list $(v_1, \dots, v_k, w)$ and this $\mathbf{E}$ turns it into $(v_1, \dots, v_k, h)$, so

$$\mathbf{G}(v_1, \dots, v_k, h) = \mathbf{E}^* \, \mathbf{G}(v_1, \dots, v_k, w) \, \mathbf{E}.$$

The matrix $\mathbf{E}$ is triangular with ones on the diagonal, so $\det \mathbf{E} = 1$ by Theorem 7.4.6, and $\det(\mathbf{E}^*) = \overline{\det \mathbf{E}} = 1$. By Theorem 7.5.4 the two Gram determinants are equal.

Finally, $h \in U^\perp$ gives $\langle h, v_i \rangle = \langle v_i, h \rangle = 0$ for every i, so

$$\mathbf{G}(v_1, \dots, v_k, h) = \begin{pmatrix} \mathbf{G}(v_1, \dots, v_k) & \mathbf{0} \\ \mathbf{0} & \|h\|^2 \end{pmatrix},$$

whose determinant is $\det \mathbf{G}(v_1, \dots, v_k) \cdot \|h\|^2$ by cofactor expansion along the last row. Since $\|h\| = \|w - P_U w\| = d(w, U)$ by Definition 11.3.3, this proves the lemma.

■

The recursion names the vectors in a particular order, and the last one plays a special role. The value does not depend on that choice: Theorem 13.3.3 below identifies $\operatorname{vol}(v_1, \dots, v_k)^2$ with $\det \mathbf{G}$, and permuting the list permutes the rows of $\mathbf{G}$ and the columns in the same way, which leaves the determinant unchanged (Theorem 7.5.1 (a) applied once to the rows and once to the columns, so the two sign changes cancel). So the volume is an invariant of the list as a set of vectors, and we are free to order it however a computation prefers.

Theorem 13.3.3. The Gram Determinant Is the Squared Volume

Let $(v_1, \dots, v_k)$ be a list in an inner product space over F. Then

$$\det \mathbf{G}(v_1, \dots, v_k) = \operatorname{vol}(v_1, \dots, v_k)^2.$$

In particular, if $\mathbf{A} \in M_{m \times k}(F)$ has columns $\mathbf{v}_1, \dots, \mathbf{v}_k \in F^m$, then $\operatorname{vol}(\mathbf{v}_1, \dots, \mathbf{v}_k) = \sqrt{\det(\mathbf{A}^* \mathbf{A})}$.

Proof. Induction on k. For $k = 1$, $\mathbf{G}(v_1) = (\|v_1\|^2)$ and $\operatorname{vol}(v_1) = \|v_1\|$.

Let $k \geq 2$ and assume the statement for $k-1$. With $U_{k-1} = \operatorname{span}(v_1, \dots, v_{k-1})$, Lemma 13.3.2 and then the induction hypothesis give

$$\begin{aligned} \det \mathbf{G}(v_1, \dots, v_k) &= \det \mathbf{G}(v_1, \dots, v_{k-1}) \cdot d(v_k, U_{k-1})^2 \\ &= \operatorname{vol}(v_1, \dots, v_{k-1})^2 \cdot d(v_k, U_{k-1})^2, \end{aligned}$$

and the right-hand side is $\operatorname{vol}(v_1, \dots, v_k)^2$ by Definition 13.3.1.

For the last claim, $\langle \mathbf{v}_j, \mathbf{v}_i \rangle = \mathbf{v}_i^* \mathbf{v}_j = (\mathbf{A}^* \mathbf{A})_{ij}$ in F^m with the standard inner product, so $\mathbf{G}(\mathbf{v}_1, \dots, \mathbf{v}_k) = \mathbf{A}^* \mathbf{A}$, and the volume is the nonnegative square root of its determinant.

■

Two checks that the definition is the right one. If $k = m$ and $F = \mathbb{R}$, then $\mathbf{A}$ is square and $\det(\mathbf{A}^\top \mathbf{A}) = (\det \mathbf{A})^2$ by Theorem 7.5.4 and Theorem 7.4.5, so $\operatorname{vol}(\mathbf{v}_1, \dots, \mathbf{v}_m) = |\det \mathbf{A}|$: Definition 13.3.1 agrees with Definition 7.9.1 whenever both apply. And for $k = 2$ in $\mathbb{R}^3$, Exercise (C1: Area of a triangle in space) *defined* the area of a parallelogram in space to be $\sqrt{\det(\mathbf{A}^\top \mathbf{A})}$ and promised a justification from inner products; the theorem is that justification, since base times height is what an area should be.

The volume vanishes exactly when the list is dependent, which is Theorem 11.2.5 (d) read through the theorem: a flat parallelepiped has no volume.

13.3.3 Distance to a subspace, by determinants

Dividing in Lemma 13.3.2 gives a formula for a distance with no projection in sight.

Theorem 13.3.4. Distance by Gram Determinants

Let U be a finite-dimensional subspace of an inner product space V, let $(v_1, \dots, v_k)$ be a **basis** of U, and let $w \in V$. Then

$$d(w, U)^2 = \frac{\det \mathbf{G}(v_1, \dots, v_k, w)}{\det \mathbf{G}(v_1, \dots, v_k)}.$$

Proof. The list $(v_1, \dots, v_k)$ is independent, so $\det \mathbf{G}(v_1, \dots, v_k) > 0$ by Theorem 11.2.5 (d) and the quotient is defined. Dividing the identity of Lemma 13.3.2 by it gives the formula. ∎

No orthonormal basis, no Gram–Schmidt, and no projection: two determinants and a division. The price is a $(k+1) \times (k+1)$ determinant in the numerator, so for large k the projection route of Chapter 11 is faster. The quotient is also a ratio of squared volumes, $\operatorname{vol}(v_1, \dots, v_k, w)^2 / \operatorname{vol}(v_1, \dots, v_k)^2$: volume over base area, which is the height.

Example. The same distance twice

In $\mathbb{R}^3$ with the dot product, let $U = \{(x, y, z) : x + y + z = 0\}$ and $\mathbf{w} = (1, 2, 3)$. Compute $d(\mathbf{w}, U)$ by Theorem 13.3.4, and compare with Example (Distance from a point to a plane).

Solution. The vectors $\mathbf{v}_1 = (1, -1, 0)$ and $\mathbf{v}_2 = (0, 1, -1)$ lie in U and are independent, and $\dim U = 2$, so they are a basis of U. Their inner products are $\langle \mathbf{v}_1, \mathbf{v}_1 \rangle = 2$, $\langle \mathbf{v}_2, \mathbf{v}_2 \rangle = 2$, $\langle \mathbf{v}_1, \mathbf{v}_2 \rangle = -1$, and with $\mathbf{w}$: $\langle \mathbf{w}, \mathbf{w} \rangle = 14$, $\langle \mathbf{w}, \mathbf{v}_1 \rangle = 1 - 2 = -1$, $\langle \mathbf{w}, \mathbf{v}_2 \rangle = 2 - 3 = -1$. Hence

$$\mathbf{G}(\mathbf{v}_1, \mathbf{v}_2) = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}, \qquad \mathbf{G}(\mathbf{v}_1, \mathbf{v}_2, \mathbf{w}) = \begin{pmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 14 \end{pmatrix}.$$

The first determinant is $4 - 1 = 3$. Expanding the second along the first row,

$$2(28 - 1) + 1(-14 - 1) - 1(1 + 2) = 54 - 15 - 3 = 36.$$

So $d(\mathbf{w}, U)^2 = 36/3 = 12$ and $d(\mathbf{w}, U) = 2\sqrt{3}$. Example (Distance from a point to a plane) obtained $\|(2, 2, 2)\| = 2\sqrt{3}$ by projecting onto the normal line. The two routes never meet on the page, and they agree.

Remark. The numerator 36 is the squared volume of the parallelepiped on $\mathbf{v}_1, \mathbf{v}_2, \mathbf{w}$, and the denominator 3 is the squared area of its base. The determinant of the coordinate matrix of $(\mathbf{v}_1, \mathbf{v}_2, \mathbf{w})$ is 6, and indeed $6^2 = 36$: Theorem 13.3.3 in the square case.

13.3.4 Exercises

13.3.4.1 A. Check your understanding

Exercise. A1

(a) State the criterion for a matrix $\mathbf{G} \in M_k(F)$ to be the Gram matrix of some list of vectors.

(b) Determine whether the following statement is correct, and justify your answer: every symmetric matrix with nonnegative entries is a Gram matrix.

(c) A list of 5 vectors in $\mathbb{R}^3$ has a 5×5 Gram matrix. What are the possible values of its rank, and what is its determinant?

(d) Write down the formula for $d(w, U)$ in terms of Gram determinants, saying what list is used.

(e) Define the k-dimensional volume of a list, and say what Theorem 13.3.3 asserts about it.

Solution.

(a) $\mathbf{G}$ is a Gram matrix if and only if $\mathbf{G} \succeq \mathbf{O}$ (Theorem 13.3.1). The list may then be taken in F^r with $r = \operatorname{rank} \mathbf{G}$.

(b) Incorrect. $\begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}$ has nonnegative entries and eigenvalues 3 and -1, so it is not positive semidefinite and by Theorem 13.3.1 it is no Gram matrix. (Concretely: two unit vectors cannot have inner product 2, by Cauchy–Schwarz, Theorem 11.1.5.)

(c) The rank is $\operatorname{rank} \mathbf{G} = \dim \operatorname{span}(v_1, \dots, v_5) \le 3$, so it is $0, 1, 2$ or 3; in particular $\mathbf{G}$ is singular and $\det \mathbf{G} = 0$. (A list of 5 vectors in a 3-dimensional space is dependent, so Theorem 11.2.5 (d) gives $\det \mathbf{G} = 0$ directly.)

(d) If $(v_1, \dots, v_k)$ is a **basis** of U, then

$$d(w, U)^2 = \frac{\det \mathbf{G}(v_1, \dots, v_k, w)}{\det \mathbf{G}(v_1, \dots, v_k)}$$

by Theorem 13.3.4.

(e) $\operatorname{vol}(v_1) = \|v_1\|$ and $\operatorname{vol}(v_1, \dots, v_k) = \operatorname{vol}(v_1, \dots, v_{k-1})\, d(v_k, \operatorname{span}(v_1, \dots, v_{k-1}))$ (Definition 13.3.1). Theorem 13.3.3 says $\det \mathbf{G}(v_1, \dots, v_k) = \operatorname{vol}(v_1, \dots, v_k)^2$.

13.3.4.2 B. Practice

Exercise. B1: Which are Gram matrices?

Determine which of the following are Gram matrices of some list of vectors in some real inner product space. For those that are, exhibit a list; for those that are not, give a reason.

(i) $\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$

(ii) $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$

(iii) $\begin{pmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{pmatrix}$

(iv) $\begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}$

Solution.

(i) Yes. It is symmetric with eigenvalues 1 and 3, both > 0, so it is positive definite and Theorem 13.3.1 applies. A list: $\mathbf{v}_1 = (1, 1, 0)$, $\mathbf{v}_2 = (1, 0, 1)$ in $\mathbb{R}^3$, whose Gram matrix is exactly this.

(ii) No. It is not symmetric, so it is not Hermitian and Definition 13.1.1 fails at once; a Gram matrix over $\mathbb{R}$ is symmetric by Theorem 11.2.5 (a).

(iii) Yes. Its rank is 2: the rows sum to zero, and the top-left 2×2 minor is $3 \neq 0$. Its eigenvalues are 0 (eigenvector $(1, 1, 1)$) and 3 (twice, on the plane $x + y + z = 0$), all ≥ 0. A list in $\mathbb{R}^2$, as Theorem 13.3.1 (b) promises with $r = 2$: the three vectors $(\sqrt{2}, 0)$, $(-\frac{1}{\sqrt{2}}, \sqrt{\frac{3}{2}})$, $(-\frac{1}{\sqrt{2}}, -\sqrt{\frac{3}{2}})$, three vectors of length $\sqrt{2}$ at 120° to each other.

(iv) No. It is symmetric, but its leading principal minors are $1, 1, -1$, and the determinant of a positive semidefinite matrix is ≥ 0 (its eigenvalues are ≥ 0). So it is not positive semidefinite, and by Theorem 13.3.1 it is not a Gram matrix. (Its eigenvalues are 1 and $1 \pm \sqrt{2}$, one of which is negative.)

Exercise. B2: A volume two ways

In $\mathbb{R}^3$, let $\mathbf{v}_1 = (1, 1, 0)$, $\mathbf{v}_2 = (1, 0, 1)$, $\mathbf{v}_3 = (0, 1, 1)$. Compute $\mathrm{vol}(\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3)$ with Theorem 13.3.3, and check the answer against Definition 7.9.1.

Solution. The pairwise dot products give

$$\mathbf{G} = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix},$$

and expanding along the first row, $\det \mathbf{G} = 2(4 - 1) - 1(2 - 1) + 1(1 - 2) = 6 - 1 - 1 = 4$. By Theorem 13.3.3, $\mathrm{vol}(\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3) = \sqrt{4} = 2$.

Directly, the matrix $\mathbf{A}$ with these columns has $\det \mathbf{A} = 1(0 - 1) - 1(1 - 0) + 0 = -2$, so Definition 7.9.1 gives $|-2| = 2$. The two agree, as they must, since $\det(\mathbf{A}^\top \mathbf{A}) = (\det \mathbf{A})^2$.

Exercise. B3: A distance in $\mathbb{R}^4$

In $\mathbb{R}^4$ with the dot product, let $U = \mathrm{span}\,((1, 1, 0, 0), (0, 1, 1, 0))$ and $\mathbf{w} = (1, 0, 0, 1)$. Find $d(\mathbf{w}, U)$ using Theorem 13.3.4.

Solution. Write $\mathbf{v}_1 = (1,1,0,0)$ and $\mathbf{v}_2 = (0,1,1,0)$; they are independent, hence a basis of U. The inner products are $\langle \mathbf{v}_1, \mathbf{v}_1\rangle = \langle \mathbf{v}_2, \mathbf{v}_2\rangle = 2$, $\langle \mathbf{v}_1, \mathbf{v}_2\rangle = 1$, $\langle \mathbf{w}, \mathbf{w}\rangle = 2$, $\langle \mathbf{w}, \mathbf{v}_1\rangle = 1$, $\langle \mathbf{w}, \mathbf{v}_2\rangle = 0$. So

$$\mathbf{G}(\mathbf{v}_1, \mathbf{v}_2) = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}, \qquad \mathbf{G}(\mathbf{v}_1, \mathbf{v}_2, \mathbf{w}) = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 0 \\ 1 & 0 & 2 \end{pmatrix},$$

with determinants 3 and $2(4) - 1(2) + 1(-2) = 4$. Hence $d(\mathbf{w}, U)^2 = 4/3$ and $d(\mathbf{w}, U) = 2/\sqrt{3}$.

Check. The projection is $P_U \mathbf{w} = \frac{1}{3}(2,1,-1,0)$, so $\mathbf{w} - P_U \mathbf{w} = \frac{1}{3}(1,-1,1,3)$, of norm $\frac{1}{3}\sqrt{12} = 2/\sqrt{3}$.

13.3.4.3 C. Going deeper

Exercise. C1: Rank of a Gram matrix

Let $(v_1, \dots, v_k)$ be a list in a finite-dimensional inner product space V over F, with Gram matrix $\mathbf{G}$.

(a) Prove that $\operatorname{rank} \mathbf{G} = \dim \operatorname{span}(v_1, \dots, v_k)$.

(b) Deduce Theorem 11.2.5 (c): $\mathbf{G}$ is invertible if and only if the list is independent.

Hint for (a): the coordinate matrix of the list in an orthonormal basis of the span.

Solution.

(a) Let $U = \operatorname{span}(v_1, \dots, v_k)$, $r = \dim U$, and let $(e_1, \dots, e_r)$ be an orthonormal basis of U, which exists by Theorem 11.2.2. Let $\mathbf{A} \in M_{r \times k}(F)$ have as its j-th column the coordinate vector of v_j. As computed in the proof of Theorem 11.2.5 (d), $\mathbf{G} = \mathbf{A}^* \mathbf{A}$. By Corollary 11.6.6 (b), $\operatorname{rank} \mathbf{G} = \operatorname{rank}(\mathbf{A}^* \mathbf{A}) = \operatorname{rank} \mathbf{A}$. Finally $\operatorname{rank} \mathbf{A} = r$: the columns of $\mathbf{A}$ span F^r, because the coordinate map $U \to F^r$ is an isomorphism carrying the spanning list $(v_1, \dots, v_k)$ of U to the columns of $\mathbf{A}$.

(b) $\mathbf{G}$ is a $k \times k$ matrix, so it is invertible if and only if $\operatorname{rank} \mathbf{G} = k$ (Theorem 7.5.3 (d)). By (a) that says $\dim \operatorname{span}(v_1, \dots, v_k) = k$, which holds exactly when the list is independent.

Exercise. C2: Why the conjugate is not optional

Over $\mathbb{C}$, one might try to build a table of inner products with the **bilinear** expression $\mathbf{v}_i^{\top} \mathbf{v}_j$, without any conjugate. Let $\mathbf{v} = (1, i) \in \mathbb{C}^2$.

(a) Compute $\mathbf{v}^{\top} \mathbf{v}$ and $\mathbf{v}^* \mathbf{v}$, and explain why only one of them is a legitimate squared length.

(b) The list $(\mathbf{v})$ is independent, yet the 1×1 "bilinear Gram matrix" $(\mathbf{v}^{\top} \mathbf{v})$ is singular. Which hypothesis of Theorem 13.3.1 is being violated, and which statement of Chapter 7 does this illustrate?

Solution.

(a) $\mathbf{v}^{\top}\mathbf{v} = 1^2 + i^2 = 0$, while $\mathbf{v}^{*}\mathbf{v} = |1|^2 + |i|^2 = 2$. A squared length must be ≥ 0 and must vanish only for the zero vector (Definition 11.1.1), and $\mathbf{v} \neq \mathbf{0}$; so $\mathbf{v}^{\top}\mathbf{v}$ cannot be one, and $\mathbf{v}^{*}\mathbf{v} = \|\mathbf{v}\|^2$ is the correct value.

(b) Theorem 13.3.1 is a statement about the Gram matrix of Definition 11.2.3, whose entries are $\langle \mathbf{v}_j, \mathbf{v}_i \rangle = \mathbf{v}_i^{*}\mathbf{v}_j$. The matrix $(\mathbf{v}^{\top}\mathbf{v}) = \mathbf{O}_1$ is not that matrix, so the theorem simply does not apply; the genuine Gram matrix here is (2), which is positive definite, as it must be for an independent list. This is exactly the failure recorded after Corollary 7.7.3: over $\mathbb{C}$, $\mathbf{A} = \begin{pmatrix} 1 & i \end{pmatrix}$ has rank 1 while $\mathbf{A}\mathbf{A}^{\top} = \mathbf{O}_1$, so the real equality case of that corollary has no complex analogue — unless the transpose is replaced by the conjugate transpose.

Exercise. C3: Height and volume

Let $(v_1, \dots, v_k)$ be an independent list in an inner product space over F. For each j, write U_j for the span of $v_1, \dots, v_j$.

(a) Prove that $\mathrm{vol}(v_1, \dots, v_k) \leq \|v_1\|\|v_2\| \cdots \|v_k\|$, with equality if and only if the list is orthogonal.

(b) Deduce that $\det \mathbf{G}(v_1, \dots, v_k) \leq \|v_1\|^2 \cdots \|v_k\|^2$.

Hint for (a): compare each height with the corresponding length.

Solution.

(a) Fix $j \geq 2$ and write $v_j = P_{U_{j-1}} v_j + h_j$ with $h_j \in U_{j-1}^{\perp}$, as in Theorem 11.3.2 (a). The two summands are orthogonal, so Theorem 11.1.3 gives

$$\|v_j\|^2 = \|P_{U_{j-1}} v_j\|^2 + \|h_j\|^2 \;\geq\; \|h_j\|^2 = d(v_j, U_{j-1})^2,$$

with equality if and only if $P_{U_{j-1}} v_j = 0$. Multiplying the k inequalities $d(v_j, U_{j-1}) \leq \|v_j\|$ (with $d(v_1, U_0)$ read as $\|v_1\|$) and using Definition 13.3.1 repeatedly gives

$$\mathrm{vol}(v_1, \dots, v_k) \leq \|v_1\| \cdots \|v_k\|.$$

All factors are > 0, since the list is independent, so equality holds if and only if it holds in every factor, that is $P_{U_{j-1}} v_j = 0$ for every $j \geq 2$. That says each v_j is orthogonal to U_{j-1}, hence to every earlier v_i; and conversely an orthogonal list has $P_{U_{j-1}} v_j = 0$ by Theorem 11.3.3 (a).

(b) Square (a) and apply Theorem 13.3.3.

13.4 Two Forms at Once

Chapter 12 asked when several operators can be diagonalized in one orthonormal basis, and answered: when they commute and each is normal (Theorem 12.7.1). That theorem is about **operators**, and its currency is the unitary similarity $\mathbf{U}^{*}T\mathbf{U}$. This section asks a question that looks similar and is not. Given two **forms** — two self-adjoint matrices, each thought of as

the recipe $\mathbf{x} \mapsto \mathbf{x}^*\mathbf{A}\mathbf{x}$ — can one change of variable make both of them diagonal at once? No commuting is assumed, and in general the two matrices do not commute. The price is that the change of variable is no longer unitary, and the hypothesis is that one of the two forms is positive definite.

Throughout, $F = \mathbb{R}$ or $F = \mathbb{C}$, and "self-adjoint" means $\mathbf{A}^* = \mathbf{A}$ (over $\mathbb{R}$: symmetric).

13.4.1 Similarity is for operators, congruence is for forms

A matrix can play two roles, and the two roles come with different change-of-variable rules.

As an **operator**, $\mathbf{A}$ eats a vector and returns a vector. Rewriting $\mathbf{x} = \mathbf{S}\mathbf{y}$ in the equation $\mathbf{z} = \mathbf{A}\mathbf{x}$ turns it into $\mathbf{S}^{-1}\mathbf{z} = (\mathbf{S}^{-1}\mathbf{A}\mathbf{S})\mathbf{y}$: the new matrix is $\mathbf{S}^{-1}\mathbf{A}\mathbf{S}$, and the relation is **similarity**. Eigenvalues, trace, determinant, and characteristic polynomial all survive it.

As a **form**, $\mathbf{A}$ eats a vector and returns the scalar $\mathbf{x}^*\mathbf{A}\mathbf{x}$. Rewriting $\mathbf{x} = \mathbf{S}\mathbf{y}$ turns that scalar into

$$(\mathbf{S}\mathbf{y})^*\mathbf{A}(\mathbf{S}\mathbf{y}) = \mathbf{y}^*(\mathbf{S}^*\mathbf{A}\mathbf{S})\mathbf{y},$$

so the new matrix is $\mathbf{S}^*\mathbf{A}\mathbf{S}$. Two matrices $\mathbf{A}$ and $\mathbf{S}^*\mathbf{A}\mathbf{S}$ with $\mathbf{S}$ invertible are called **congruent**; this is the operation of Proposition 13.1.2, where it was already shown to preserve the sign of a form. Congruence is the right relation here for the plain reason that it is the one the question is phrased in: "make both forms diagonal" means "make both $\mathbf{x}^*\mathbf{A}\mathbf{x}$ and $\mathbf{x}^*\mathbf{B}\mathbf{x}$ into sums of squares in the same new coordinates", and that is a statement about $\mathbf{S}^*\mathbf{A}\mathbf{S}$ and $\mathbf{S}^*\mathbf{B}\mathbf{S}$. Chapter 14 studies congruence in its own right, including what it does preserve; for now, note that it certainly does **not** preserve eigenvalues, since $\mathbf{S}^*\mathbf{I}\mathbf{S} = \mathbf{S}^*\mathbf{S}$ can be any positive definite matrix at all.

Congruence and similarity coincide in exactly one situation: when $\mathbf{S}$ is unitary, $\mathbf{S}^* = \mathbf{S}^{-1}$. That is the situation of Chapter 12, and it is what makes that chapter's theorems so strong and its hypotheses so restrictive.

Warning. **This is not Chapter 12's simultaneous diagonalization.** Theorem 12.7.1 needs the operators to **commute** and produces a **unitary** $\mathbf{U}$ with every $\mathbf{U}^*T\mathbf{U}$ diagonal; eigenvalues are preserved. The theorem below needs **no commuting** and one of the two matrices to be **positive definite**, and produces a merely invertible $\mathbf{S}$; the eigenvalues of $\mathbf{B}$ are destroyed, and even $\mathbf{A}$ comes out as $\mathbf{I}$. The two theorems have no common generalization: neither hypothesis implies the other. When the matrices both commute and one is positive definite, both theorems apply and give different factorizations.

13.4.2 Diagonalizing two forms at once

Theorem 13.4.1. Diagonalizing Two Forms by Congruence

Let $\mathbf{A}, \mathbf{B} \in M_n(F)$ be self-adjoint with $\mathbf{A} \succ \mathbf{O}$. Then there is an invertible $\mathbf{S} \in M_n(F)$ with

$$\mathbf{S}^*\mathbf{A}\mathbf{S} = \mathbf{I}_n \qquad \text{and} \qquad \mathbf{S}^*\mathbf{B}\mathbf{S} = \mathbf{D}$$

for some **real** diagonal $\mathbf{D}$. Over $\mathbb{R}$, with $\mathbf{A}, \mathbf{B}$ symmetric, $\mathbf{S}$ may be taken real.

Idea. Two steps, and the first one is the whole idea. ① *Whiten.* The hypothesis $\mathbf{A} \succ \mathbf{O}$ gives an invertible positive definite square root $\mathbf{A}^{1/2}$, and the congruence by $\mathbf{A}^{-1/2}$ turns $\mathbf{A}$ into $\mathbf{I}$. That uses up the definiteness of $\mathbf{A}$ and nothing else. ② Now $\mathbf{I}$ is invariant under every **unitary** congruence, because a unitary congruence is a similarity; so we still have the whole spectral theorem available for the transformed $\mathbf{B}$, and it is self-adjoint. Diagonalize it and compose the two changes of variable.

Proof. Since $\mathbf{A} \succ \mathbf{O}$, Theorem 13.2.1 gives a unique $\mathbf{A}^{1/2} \succeq \mathbf{O}$ with $(\mathbf{A}^{1/2})^2 = \mathbf{A}$. Its eigenvalues are the nonnegative square roots of the eigenvalues of $\mathbf{A}$, which are > 0 by Theorem 13.1.5; hence $\mathbf{A}^{1/2} \succ \mathbf{O}$ and in particular $\mathbf{A}^{1/2}$ is invertible. Write $\mathbf{A}^{-1/2} = (\mathbf{A}^{1/2})^{-1}$, which is again self-adjoint, since the inverse of an invertible self-adjoint matrix is self-adjoint.

Put $\mathbf{C} = \mathbf{A}^{-1/2}\mathbf{B}\mathbf{A}^{-1/2}$. Then $\mathbf{C}^* = \mathbf{A}^{-1/2}\mathbf{B}^*\mathbf{A}^{-1/2} = \mathbf{C}$, so $\mathbf{C}$ is self-adjoint. By the spectral theorem for matrices there is a unitary $\mathbf{U}$ and a real diagonal $\mathbf{D}$ with $\mathbf{C} = \mathbf{U}\mathbf{D}\mathbf{U}^*$ (Corollary 12.4.3 over $\mathbb{C}$; Corollary 12.5.2 over $\mathbb{R}$, with $\mathbf{U}$ orthogonal and real).

Set $\mathbf{S} = \mathbf{A}^{-1/2}\mathbf{U}$, a product of invertible matrices, hence invertible. Then

$$\mathbf{S}^*\mathbf{A}\mathbf{S} = \mathbf{U}^*\mathbf{A}^{-1/2}\mathbf{A}\mathbf{A}^{-1/2}\mathbf{U} = \mathbf{U}^*\mathbf{U} = \mathbf{I}_n,$$

where the middle equality uses $\mathbf{A}^{-1/2}\mathbf{A}\mathbf{A}^{-1/2} = \mathbf{A}^{-1/2}\mathbf{A}^{1/2}\mathbf{A}^{1/2}\mathbf{A}^{-1/2} = \mathbf{I}_n$. And

$$\mathbf{S}^*\mathbf{B}\mathbf{S} = \mathbf{U}^*\mathbf{A}^{-1/2}\mathbf{B}\mathbf{A}^{-1/2}\mathbf{U} = \mathbf{U}^*\mathbf{C}\mathbf{U} = \mathbf{D}.$$

Over $\mathbb{R}$ every matrix named here is real: $\mathbf{A}^{1/2}$ is real because it is a polynomial in $\mathbf{A}$ (Theorem 13.2.1), and $\mathbf{U}$ is real by Corollary 12.5.2. This proves the theorem. ∎

The theorem says more than "both become diagonal": the definite one becomes the **identity**, which is the strongest normalization available, and it costs nothing extra. Note also which hypothesis did which job. Definiteness of $\mathbf{A}$ was used exactly once, to invert $\mathbf{A}^{1/2}$. Self-adjointness of $\mathbf{B}$ was used exactly once, to apply the spectral theorem to $\mathbf{C}$. Nothing at all was assumed about $\mathbf{AB}$ versus $\mathbf{BA}$.

Quick check. Why can $\mathbf{S}$ not be taken unitary in Theorem 13.4.1?

Solution. A unitary $\mathbf{S}$ has $\mathbf{S}^* = \mathbf{S}^{-1}$, so $\mathbf{S}^*\mathbf{A}\mathbf{S} = \mathbf{I}$ would give $\mathbf{A} = \mathbf{S}\mathbf{I}\mathbf{S}^* = \mathbf{I}$. So a unitary $\mathbf{S}$ can only work when $\mathbf{A}$ is already the identity — and then the theorem is just the spectral theorem for $\mathbf{B}$.

13.4.3 The generalized eigenvalue problem

The matrix $\mathbf{S}$ of the theorem is not an abstract object: its columns solve an eigenvalue problem in which the identity has been replaced by $\mathbf{A}$.

Definition 13.4.1. Generalized eigenvalue

Let $\mathbf{A}, \mathbf{B} \in M_n(F)$. A scalar $\lambda \in F$ is a **generalized eigenvalue** of the **pair** $(\mathbf{A}, \mathbf{B})$ if

$$\mathbf{Bx} = \lambda \mathbf{Ax} \qquad \text{for some } \mathbf{x} \neq \mathbf{0},$$

and such an $\mathbf{x}$ is a **generalized eigenvector** for λ. The family of matrices $\mathbf{B} - \lambda \mathbf{A}$ is called the **pencil** of the pair.

Taking $\mathbf{A} = \mathbf{I}$ recovers Definition 9.2.1, so the word "generalized" is honest. The condition $\mathbf{Bx} = \lambda \mathbf{Ax}$ with $\mathbf{x} \neq \mathbf{0}$ says that $\mathbf{B} - \lambda \mathbf{A}$ is singular, that is $\det(\mathbf{B} - \lambda \mathbf{A}) = 0$; so the generalized eigenvalues are the roots of the polynomial $\det(\mathbf{B} - \lambda \mathbf{A})$ in λ, the **characteristic polynomial of the pencil**. Its degree is at most n, and exactly n when $\mathbf{A}$ is invertible, since the λ^n-coefficient is $(-1)^n \det \mathbf{A}$. When $\mathbf{A}$ is singular the degree can drop, and the polynomial can even be a nonzero constant, in which case the pair has no generalized eigenvalues at all; exercise C3 below is such a pair. From here on $\mathbf{A} \succ \mathbf{O}$, so this does not arise.

Nothing so far forces those roots to be real: for $\mathbf{A} = \mathbf{I}$ they are just eigenvalues, and a general $\mathbf{B}$ has complex ones. Self-adjointness of $\mathbf{B}$ alone is not enough either, as the warning at the end of this section shows. It is definiteness of $\mathbf{A}$ that makes the problem behave.

Theorem 13.4.2. The Definite Generalized Eigenvalue Problem

Let $\mathbf{A}, \mathbf{B} \in M_n(F)$ be self-adjoint with $\mathbf{A} \succ \mathbf{O}$, and let $\mathbf{S}$ and $\mathbf{D} = \operatorname{diag}(d_1, \dots, d_n)$ be as in Theorem 13.4.1, with columns $\mathbf{s}_1, \dots, \mathbf{s}_n$ of $\mathbf{S}$. Then:

(a) $\mathbf{Bs}_i = d_i \mathbf{As}_i$ for every i, and $\mathbf{s}_i^* \mathbf{As}_j = \delta_{ij}$: the columns of $\mathbf{S}$ are a basis of F^n of generalized eigenvectors, orthonormal for the inner product $\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{A}} = \mathbf{y}^* \mathbf{Ax}$;

(b) the generalized eigenvalues of $(\mathbf{A}, \mathbf{B})$ are exactly $d_1, \dots, d_n$, all **real**, and

$$\det(\mathbf{B} - \lambda \mathbf{A}) = \frac{1}{|\det \mathbf{S}|^2} \prod_{i=1}^{n} (d_i - \lambda);$$

(c) $\mathbf{A}^{-1}\mathbf{B} = \mathbf{SDS}^{-1}$; in particular $\mathbf{A}^{-1}\mathbf{B}$ is diagonalizable with real eigenvalues $d_1, \dots, d_n$, although it need not be self-adjoint.

Idea. Everything is already in the two equations $\mathbf{S}^*\mathbf{AS} = \mathbf{I}$ and $\mathbf{S}^*\mathbf{BS} = \mathbf{D}$. Substituting the first into the second turns $\mathbf{D}$ into $\mathbf{S}^*\mathbf{ASD}$, and canceling the invertible $\mathbf{S}^*$ leaves $\mathbf{BS} = \mathbf{ASD}$, which read column by column is (a). Then (b) and (c) are two ways of undoing the substitution.

Proof.

(a) From $\mathbf{S}^*\mathbf{AS} = \mathbf{I}$ we get $\mathbf{D} = \mathbf{ID} = \mathbf{S}^*\mathbf{ASD}$, and also $\mathbf{D} = \mathbf{S}^*\mathbf{BS}$. Therefore $\mathbf{S}^*(\mathbf{BS} - \mathbf{ASD}) = \mathbf{O}$, and $\mathbf{S}^*$ is invertible, so

$$\mathbf{BS} = \mathbf{ASD}.$$

Column i of the left side is $\mathbf{Bs}_i$ and of the right side is $d_i \mathbf{As}_i$, since $\mathbf{D}$ is diagonal. Each $\mathbf{s}_i \neq \mathbf{0}$ because $\mathbf{S}$ is invertible, so each d_i is a generalized eigenvalue with eigenvector $\mathbf{s}_i$.

The entry (i,j) of $\mathbf{S}^*\mathbf{A}\mathbf{S} = \mathbf{I}$ is $\mathbf{s}_i^*\mathbf{A}\mathbf{s}_j$, which is therefore δ_{ij}. Finally $\overline{\langle \mathbf{x},\mathbf{y}\rangle_{\mathbf{A}}} = \mathbf{y}^*\mathbf{A}\mathbf{x}$ is an inner product on F^n, exactly because $\mathbf{A} \succ \mathbf{O}$: it is sesquilinear, $\overline{\langle \mathbf{y},\mathbf{x}\rangle_{\mathbf{A}}} = \overline{\mathbf{x}^*\mathbf{A}\mathbf{y}} = \mathbf{y}^*\mathbf{A}^*\mathbf{x} = \langle \mathbf{x},\mathbf{y}\rangle_{\mathbf{A}}$, and $\langle \mathbf{x},\mathbf{x}\rangle_{\mathbf{A}} = \mathbf{x}^*\mathbf{A}\mathbf{x} > 0$ for $\mathbf{x} \neq \mathbf{0}$.

(b) From $\mathbf{S}^*\mathbf{A}\mathbf{S} = \mathbf{I}$ and $\mathbf{S}^*\mathbf{B}\mathbf{S} = \mathbf{D}$ we get $\mathbf{A} = (\mathbf{S}^*)^{-1}\mathbf{S}^{-1}$ and $\mathbf{B} = (\mathbf{S}^*)^{-1}\mathbf{D}\mathbf{S}^{-1}$, hence

$$\mathbf{B} - \lambda\mathbf{A} = (\mathbf{S}^*)^{-1}(\mathbf{D} - \lambda\mathbf{I})\mathbf{S}^{-1}.$$

Taking determinants with Theorem 7.5.4, and using $\det(\mathbf{S}^*) = \overline{\det \mathbf{S}}$,

$$\det(\mathbf{B} - \lambda\mathbf{A}) = \frac{\det(\mathbf{D} - \lambda\mathbf{I})}{\overline{\det \mathbf{S}}\,\det \mathbf{S}} = \frac{1}{|\det \mathbf{S}|^2}\prod_i (d_i - \lambda),$$

the last step by Theorem 7.4.6. Since $|\det \mathbf{S}|^2 \neq 0$, the roots of $\det(\mathbf{B} - \lambda\mathbf{A})$ are exactly $d_1, \dots, d_n$; and λ is a generalized eigenvalue if and only if $\mathbf{B} - \lambda\mathbf{A}$ is singular, that is if and only if $\det(\mathbf{B} - \lambda\mathbf{A}) = 0$ (Theorem 3.4.4). The d_i are real because $\mathbf{D} = \mathbf{S}^*\mathbf{B}\mathbf{S}$ is self-adjoint, so its diagonal entries equal their own conjugates.

(c) Inverting $\mathbf{A} = (\mathbf{S}^*)^{-1}\mathbf{S}^{-1}$ gives $\mathbf{A}^{-1} = \mathbf{S}\mathbf{S}^*$, so

$$\mathbf{A}^{-1}\mathbf{B} = \mathbf{S}\mathbf{S}^*(\mathbf{S}^*)^{-1}\mathbf{D}\mathbf{S}^{-1} = \mathbf{S}\mathbf{D}\mathbf{S}^{-1}.$$

This proves the theorem.

■

Part (c) explains the phenomenon. The matrix $\mathbf{A}^{-1}\mathbf{B}$ is what one would naively form to turn $\mathbf{B}\mathbf{x} = \lambda\mathbf{A}\mathbf{x}$ into an ordinary eigenvalue problem, and it is usually not self-adjoint, so there is no reason from Chapter 12 for its eigenvalues to be real. But it **is** self-adjoint with respect to the inner product $\langle \cdot,\cdot\rangle_{\mathbf{A}}$, and reality and the orthogonal eigenbasis come from that inner product rather than the standard one. The congruence $\mathbf{S}$ is exactly the change of variable that turns $\langle \cdot,\cdot\rangle_{\mathbf{A}}$ into the standard inner product.

Example. A pencil that does not commute

Let

$$\mathbf{A} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}, \qquad \mathbf{B} = \begin{pmatrix} 0 & 3 \\ 3 & 6 \end{pmatrix}.$$

Check that $\mathbf{A} \succ \mathbf{O}$, that $\mathbf{A}\mathbf{B} \neq \mathbf{B}\mathbf{A}$, and that $\mathbf{B}$ is indefinite. Find the generalized eigenvalues and an $\mathbf{A}$-orthonormal basis of generalized eigenvectors, and write down $\mathbf{S}$.

Solution. Setting up. $\mathbf{A}$ is symmetric with leading principal minors $2 > 0$ and $4 - 1 = 3 > 0$, so $\mathbf{A} \succ \mathbf{O}$ by Theorem 13.1.5. Its own eigenvalues are 1 and 3. The products are

$$\mathbf{A}\mathbf{B} = \begin{pmatrix} 3 & 12 \\ 6 & 15 \end{pmatrix}, \qquad \mathbf{B}\mathbf{A} = \begin{pmatrix} 3 & 6 \\ 12 & 15 \end{pmatrix},$$

which differ, so Chapter 12's theorem does not apply. Also $\det \mathbf{B} = -9 < 0$, so $\mathbf{B}$ has one positive and one negative eigenvalue and is neither positive nor negative semidefinite.

The pencil.

$$\mathbf{B} - \lambda\mathbf{A} = \begin{pmatrix} -2\lambda & 3-\lambda \\ 3-\lambda & 6-2\lambda \end{pmatrix},$$

so

$$\begin{aligned} \det(\mathbf{B} - \lambda\mathbf{A}) &= -2\lambda(6-2\lambda) - (3-\lambda)^2 \\ &= 4\lambda^2 - 12\lambda - 9 + 6\lambda - \lambda^2 \\ &= 3\lambda^2 - 6\lambda - 9 = 3(\lambda - 3)(\lambda + 1). \end{aligned}$$

The generalized eigenvalues are $\lambda = -1$ and $\lambda = 3$, both real, as Theorem 13.4.2 (b) promises.

Eigenvectors. For $\lambda = -1$, $\mathbf{B} + \mathbf{A} = \begin{pmatrix} 2 & 4 \\ 4 & 8 \end{pmatrix}$, whose kernel is spanned by $\mathbf{x}_1 = (2, -1)$. For $\lambda = 3$, $\mathbf{B} - 3\mathbf{A} = \begin{pmatrix} -6 & 0 \\ 0 & 0 \end{pmatrix}$, whose kernel is spanned by $\mathbf{x}_2 = (0, 1)$. Checking directly: $\mathbf{B}\mathbf{x}_1 = (-3, 0)$ and $\mathbf{A}\mathbf{x}_1 = (3, 0)$, so $\mathbf{B}\mathbf{x}_1 = -\mathbf{A}\mathbf{x}_1$; and $\mathbf{B}\mathbf{x}_2 = (3, 6) = 3(1, 2) = 3\mathbf{A}\mathbf{x}_2$.

Normalizing. $\mathbf{x}_1^\top \mathbf{A}\mathbf{x}_1 = (2, -1) \cdot (3, 0) = 6$ and $\mathbf{x}_2^\top \mathbf{A}\mathbf{x}_2 = (0, 1) \cdot (1, 2) = 2$, while $\mathbf{x}_1^\top \mathbf{A}\mathbf{x}_2 = (2, -1) \cdot (1, 2) = 0$: the two are already $\mathbf{A}$-orthogonal, as they must be, having different generalized eigenvalues. Scaling to $\mathbf{A}$-length 1,

$$\mathbf{S} = \begin{pmatrix} 2/\sqrt{6} & 0 \\ -1/\sqrt{6} & 1/\sqrt{2} \end{pmatrix}, \qquad \mathbf{S}^\top \mathbf{A}\mathbf{S} = \mathbf{I}_2, \qquad \mathbf{S}^\top \mathbf{B}\mathbf{S} = \begin{pmatrix} -1 & 0 \\ 0 & 3 \end{pmatrix}.$$

So in the coordinates $\mathbf{x} = \mathbf{S}\mathbf{y}$ the two forms read $\mathbf{x}^\top \mathbf{A}\mathbf{x} = y_1^2 + y_2^2$ and $\mathbf{x}^\top \mathbf{B}\mathbf{x} = -y_1^2 + 3y_2^2$. The signs $(-, +)$ match the signs of the eigenvalues of $\mathbf{B}$ itself, which are $3 \pm 3\sqrt{2}$; Chapter 14 shows that this is no coincidence.

The problem $\mathbf{B}\mathbf{x} = \lambda\mathbf{A}\mathbf{x}$ with $\mathbf{A} \succ \mathbf{O}$ is the one an engineer meets first: for a vibrating structure, $\mathbf{A}$ is the mass matrix and $\mathbf{B}$ the stiffness matrix, the generalized eigenvalues are the squared natural frequencies, and the $\mathbf{A}$-orthonormal eigenvectors are the normal modes, in which the motion decouples into independent one-dimensional oscillations.

13.4.4 When neither form is definite

Definiteness of one of the two matrices is not decoration.

Warning. **Two self-adjoint matrices, neither definite, need not be simultaneously diagonalizable by congruence.** Take the real symmetric pair

$$\mathbf{A} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \qquad \mathbf{B} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

Suppose $\mathbf{S}^\top \mathbf{A}\mathbf{S}$ and $\mathbf{S}^\top \mathbf{B}\mathbf{S}$ were both diagonal for some invertible real $\mathbf{S}$, say $\mathbf{D}_\mathbf{A}$ and $\mathbf{D}_\mathbf{B}$. Since $\mathbf{A}$ is invertible, so is $\mathbf{D}_\mathbf{A}$, and from $\mathbf{A} = (\mathbf{S}^\top)^{-1}\mathbf{D}_\mathbf{A}\mathbf{S}^{-1}$ and $\mathbf{B} = (\mathbf{S}^\top)^{-1}\mathbf{D}_\mathbf{B}\mathbf{S}^{-1}$ we get

$$\mathbf{A}^{-1}\mathbf{B} = \mathbf{S}\mathbf{D}_\mathbf{A}^{-1}\mathbf{S}^\top(\mathbf{S}^\top)^{-1}\mathbf{D}_\mathbf{B}\mathbf{S}^{-1} = \mathbf{S}(\mathbf{D}_\mathbf{A}^{-1}\mathbf{D}_\mathbf{B})\mathbf{S}^{-1},$$

so $\mathbf{A}^{-1}\mathbf{B}$ would be similar to a real diagonal matrix and would have real eigenvalues. But $\mathbf{A}^{-1} = \mathbf{A}$ here, and

$$\mathbf{A}^{-1}\mathbf{B} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$$

has characteristic polynomial $\lambda^2 + 1$, with roots $\pm i$. So no such $\mathbf{S}$ exists — and the pencil has **no** generalized eigenvalues in $\mathbb{R}$ at all.

The obstruction in the warning is exactly the failure of part (c) of Theorem 13.4.2. That part is in fact the right way to test a pair: if $\mathbf{A}$ is invertible, a necessary condition for simultaneous diagonalization by congruence is that $\mathbf{A}^{-1}\mathbf{B}$ be diagonalizable with real eigenvalues, and the theorem says that $\mathbf{A} \succ \mathbf{O}$ is enough to guarantee it.

13.4.5 Exercises

13.4.5.1 A. Check your understanding

Exercise. A1

(a) Define what it means for two matrices to be **congruent**, and say which of similarity and congruence is the natural relation for a form.

(b) State Theorem 13.4.1, with all hypotheses.

(c) Name two differences between Theorem 13.4.1 and Theorem 12.7.1.

(d) Determine whether the following statement is correct, and justify your answer: the generalized eigenvalues of $(\mathbf{A}, \mathbf{B})$ are eigenvalues of $\mathbf{B}$.

(e) For $\mathbf{A} \succ \mathbf{O}$ and $\mathbf{B}$ self-adjoint, what inner product makes $\mathbf{A}^{-1}\mathbf{B}$ self-adjoint?

Solution.

(a) $\mathbf{A}$ and $\mathbf{C}$ are congruent if $\mathbf{C} = \mathbf{S}^*\mathbf{A}\mathbf{S}$ for some invertible $\mathbf{S}$. Congruence is the natural relation for a form, since the substitution $\mathbf{x} = \mathbf{S}\mathbf{y}$ turns $\mathbf{x}^*\mathbf{A}\mathbf{x}$ into $\mathbf{y}^*(\mathbf{S}^*\mathbf{A}\mathbf{S})\mathbf{y}$; similarity $\mathbf{S}^{-1}\mathbf{A}\mathbf{S}$ is the natural relation for an operator.

(b) If $\mathbf{A}, \mathbf{B} \in M_n(F)$ are self-adjoint and $\mathbf{A} \succ \mathbf{O}$, there is an invertible $\mathbf{S}$ with $\mathbf{S}^*\mathbf{A}\mathbf{S} = \mathbf{I}_n$ and $\mathbf{S}^*\mathbf{B}\mathbf{S}$ real diagonal.

(c) Chapter 12's theorem assumes the operators commute and are normal, and delivers a **unitary** change of basis preserving eigenvalues; this one assumes nothing about commuting but assumes one matrix is positive definite, and delivers a merely invertible $\mathbf{S}$ under which eigenvalues are not preserved.

(d) Incorrect. In Example (A pencil that does not commute) the generalized eigenvalues are -1 and 3, while the eigenvalues of $\mathbf{B}$ are $3 \pm 3\sqrt{2}$. They agree when $\mathbf{A} = \mathbf{I}$, and not in general.

(e) $\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{A}} = \mathbf{y}^*\mathbf{A}\mathbf{x}$ (Theorem 13.4.2 (a)).

13.4.5.2 B. Practice

Exercise. B1: A generalized eigenvalue problem

Let

$$\mathbf{A} = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}, \qquad \mathbf{B} = \begin{pmatrix} 2 & 2 \\ 2 & 1 \end{pmatrix}.$$

(a) Verify that $\mathbf{A} \succ \mathbf{O}$ and find the generalized eigenvalues of $(\mathbf{A}, \mathbf{B})$.

(b) Find generalized eigenvectors and an invertible $\mathbf{S}$ with $\mathbf{S}^\top\mathbf{A}\mathbf{S} = \mathbf{I}_2$ and $\mathbf{S}^\top\mathbf{B}\mathbf{S}$ diagonal.

Solution.

(a) The leading principal minors of $\mathbf{A}$ are $1 > 0$ and $2 - 1 = 1 > 0$, so $\mathbf{A} \succ \mathbf{O}$ by Theorem 13.1.5. Then

$$\mathbf{B} - \lambda\mathbf{A} = \begin{pmatrix} 2-\lambda & 2-\lambda \\ 2-\lambda & 1-2\lambda \end{pmatrix},$$

so

$$\begin{aligned} \det(\mathbf{B} - \lambda\mathbf{A}) &= (2-\lambda)(1-2\lambda) - (2-\lambda)^2 \\ &= (2-\lambda)\big[(1-2\lambda) - (2-\lambda)\big] \\ &= (2-\lambda)(-1-\lambda) = (\lambda-2)(\lambda+1). \end{aligned}$$

The generalized eigenvalues are 2 and -1.

(b) For $\lambda = 2$: $\mathbf{B} - 2\mathbf{A} = \begin{pmatrix} 0 & 0 \\ 0 & -3 \end{pmatrix}$, with kernel spanned by $\mathbf{x}_1 = (1, 0)$. For $\lambda = -1$: $\mathbf{B} + \mathbf{A} = \begin{pmatrix} 3 & 3 \\ 3 & 3 \end{pmatrix}$, with kernel spanned by $\mathbf{x}_2 = (1, -1)$. Check: $\mathbf{B}\mathbf{x}_1 = (2, 2) = 2\mathbf{A}\mathbf{x}_1$ since $\mathbf{A}\mathbf{x}_1 = (1, 1)$; and $\mathbf{B}\mathbf{x}_2 = (0, 1)$ while $\mathbf{A}\mathbf{x}_2 = (0, -1)$, so $\mathbf{B}\mathbf{x}_2 = -\mathbf{A}\mathbf{x}_2$.

Now $\mathbf{x}_1^\top\mathbf{A}\mathbf{x}_1 = (1, 0) \cdot (1, 1) = 1$ and $\mathbf{x}_2^\top\mathbf{A}\mathbf{x}_2 = (1, -1) \cdot (0, -1) = 1$, and $\mathbf{x}_1^\top\mathbf{A}\mathbf{x}_2 = (1, 0) \cdot (0, -1) = 0$. Both are already $\mathbf{A}$-unit vectors, so

$$\mathbf{S} = \begin{pmatrix} 1 & 1 \\ 0 & -1 \end{pmatrix}, \qquad \mathbf{S}^\top\mathbf{A}\mathbf{S} = \mathbf{I}_2, \qquad \mathbf{S}^\top\mathbf{B}\mathbf{S} = \begin{pmatrix} 2 & 0 \\ 0 & -1 \end{pmatrix}.$$

Exercise. B2: Whitening by hand

Let $\mathbf{A} = \operatorname{diag}(4, 9)$ and $\mathbf{B} = \begin{pmatrix} 4 & 6 \\ 6 & 9 \end{pmatrix}$. Follow the two steps of the proof of Theorem 13.4.1: compute $\mathbf{A}^{-1/2}$, then $\mathbf{C} = \mathbf{A}^{-1/2}\mathbf{B}\mathbf{A}^{-1/2}$, diagonalize $\mathbf{C}$, and write down $\mathbf{S}$ and $\mathbf{S}^\top\mathbf{B}\mathbf{S}$.

Solution. $\mathbf{A}$ is diagonal with positive entries, so $\mathbf{A}^{1/2} = \operatorname{diag}(2, 3)$ and $\mathbf{A}^{-1/2} = \operatorname{diag}(\tfrac{1}{2}, \tfrac{1}{3})$. Then

$$\mathbf{C} = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{3} \end{pmatrix} \begin{pmatrix} 4 & 6 \\ 6 & 9 \end{pmatrix} \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{3} \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}.$$

Its eigenvalues are 2, with unit eigenvector $\frac{1}{\sqrt{2}}(1, 1)$, and 0, with unit eigenvector $\frac{1}{\sqrt{2}}(1, -1)$. So $\mathbf{U} = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$ and $\mathbf{D} = \operatorname{diag}(2, 0)$, and

$$\mathbf{S} = \mathbf{A}^{-1/2}\mathbf{U} = \frac{1}{\sqrt{2}}\begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{3} & -\frac{1}{3} \end{pmatrix}, \qquad \mathbf{S}^\top\mathbf{A}\mathbf{S} = \mathbf{I}_2, \qquad \mathbf{S}^\top\mathbf{B}\mathbf{S} = \begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix}.$$

As a check, $\det(\mathbf{B} - \lambda\mathbf{A}) = (4 - 4\lambda)(9 - 9\lambda) - 36 = 36\lambda^2 - 72\lambda = 36\lambda(\lambda - 2)$, whose roots 0 and 2 are the diagonal entries of $\mathbf{S}^\top\mathbf{B}\mathbf{S}$.

Exercise. B3: A pair that resists

Let $\mathbf{A} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ and $\mathbf{B} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$. Determine whether there is an invertible real $\mathbf{S}$ making $\mathbf{S}^{\top}\mathbf{A}\mathbf{S}$ and $\mathbf{S}^{\top}\mathbf{B}\mathbf{S}$ both diagonal. Justify your answer.

Solution. There is not. Both matrices are symmetric and $\mathbf{A}$ is invertible, with $\mathbf{A}^{-1} = \mathbf{A}$. If such an $\mathbf{S}$ existed, the computation in the warning above would give $\mathbf{A}^{-1}\mathbf{B} = \mathbf{S}(\mathbf{D}_{\mathbf{A}}^{-1}\mathbf{D}_{\mathbf{B}})\mathbf{S}^{-1}$, so $\mathbf{A}^{-1}\mathbf{B}$ would be diagonalizable. But

$$\mathbf{A}^{-1}\mathbf{B} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix},$$

which is nonzero with $(\mathbf{A}^{-1}\mathbf{B})^2 = \mathbf{O}$, hence has 0 as its only eigenvalue and is not diagonalizable (a diagonalizable matrix with only the eigenvalue 0 is $\mathbf{O}$). Here the generalized eigenvalues are both 0 and real, so reality of the spectrum is not the obstruction; diagonalizability is.

13.4.5.3 C. Going deeper

Exercise. C1: The A-inner product

Let $\mathbf{A} \succ \mathbf{O}$ and $\mathbf{B}$ be self-adjoint in $M_n(F)$, and write $\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{A}} = \mathbf{y}^*\mathbf{A}\mathbf{x}$.

(a) Prove that $\langle \cdot, \cdot \rangle_{\mathbf{A}}$ is an inner product on F^n.

(b) Prove that $T = T_{\mathbf{A}^{-1}\mathbf{B}}$ is self-adjoint on $(F^n, \langle \cdot, \cdot \rangle_{\mathbf{A}})$.

(c) Deduce Theorem 13.4.2 (a) and (b) from the spectral theorem applied to T, without using Theorem 13.4.1.

Solution.

(a) Sesquilinearity in the first slot is clear from the formula. Conjugate symmetry: $\overline{\langle \mathbf{y}, \mathbf{x} \rangle_{\mathbf{A}}} = \overline{\mathbf{x}^*\mathbf{A}\mathbf{y}} = \mathbf{y}^*\mathbf{A}^*\mathbf{x} = \mathbf{y}^*\mathbf{A}\mathbf{x} = \langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{A}}$, using $\mathbf{A}^* = \mathbf{A}$. Positive definiteness: $\langle \mathbf{x}, \mathbf{x} \rangle_{\mathbf{A}} = \mathbf{x}^*\mathbf{A}\mathbf{x} > 0$ for $\mathbf{x} \neq \mathbf{0}$, which is Definition 13.1.1 for $\mathbf{A} \succ \mathbf{O}$.

(b) For all $\mathbf{x}, \mathbf{y}$,

$$\langle T\mathbf{x}, \mathbf{y} \rangle_{\mathbf{A}} = \mathbf{y}^*\mathbf{A}\mathbf{A}^{-1}\mathbf{B}\mathbf{x} = \mathbf{y}^*\mathbf{B}\mathbf{x},$$

and

$$\langle \mathbf{x}, T\mathbf{y} \rangle_{\mathbf{A}} = (\mathbf{A}^{-1}\mathbf{B}\mathbf{y})^*\mathbf{A}\mathbf{x} = \mathbf{y}^*\mathbf{B}^*\mathbf{A}^{-1}\mathbf{A}\mathbf{x} = \mathbf{y}^*\mathbf{B}\mathbf{x},$$

using $(\mathbf{A}^{-1})^* = \mathbf{A}^{-1}$ and $\mathbf{B}^* = \mathbf{B}$. The two agree, so T is self-adjoint for this inner product (Definition 11.6.1).

(c) By Theorem 12.5.1 (over $\mathbb{R}$) or Theorem 12.4.2 (over $\mathbb{C}$) applied to the self-adjoint T on the inner product space $(F^n, \langle \cdot, \cdot \rangle_{\mathbf{A}})$, there is a basis $\mathbf{s}_1, \dots, \mathbf{s}_n$ of F^n, orthonormal for $\langle \cdot, \cdot \rangle_{\mathbf{A}}$, with $T\mathbf{s}_i = d_i\mathbf{s}_i$ and $d_i \in \mathbb{R}$ (Theorem 12.2.1). The equation $\mathbf{A}^{-1}\mathbf{B}\mathbf{s}_i = d_i\mathbf{s}_i$ is $\mathbf{B}\mathbf{s}_i = d_i\mathbf{A}\mathbf{s}_i$, and $\mathbf{s}_i^*\mathbf{A}\mathbf{s}_j = \langle \mathbf{s}_j, \mathbf{s}_i \rangle_{\mathbf{A}} = \delta_{ij}$; this is (a). For (b), λ is a generalized eigenvalue exactly when $\mathbf{A}^{-1}\mathbf{B} - \lambda\mathbf{I}$ is singular, that is when λ is an eigenvalue of $\mathbf{A}^{-1}\mathbf{B}$, and those are $d_1, \dots, d_n$.

Exercise. C2: Signs of the generalized eigenvalues

Let $\mathbf{A} \succ \mathbf{O}$ and $\mathbf{B}$ be self-adjoint in $M_n(F)$, with generalized eigenvalues $d_1, \dots, d_n$.

(a) Prove that $\mathbf{B} \succeq \mathbf{O}$ if and only if every $d_i \geq 0$, and $\mathbf{B} \succ \mathbf{O}$ if and only if every $d_i > 0$.

(b) Deduce that if $\mathbf{A} \succ \mathbf{O}$ and $\mathbf{B} \succ \mathbf{O}$, then $\det \mathbf{B} / \det \mathbf{A} = \prod_i d_i > 0$.

Solution.

(a) Let $\mathbf{S}$ be as in Theorem 13.4.1, so $\mathbf{S}^*\mathbf{B}\mathbf{S} = \mathbf{D} = \operatorname{diag}(d_1, \dots, d_n)$. For $\mathbf{x} \in F^n$ put $\mathbf{y} = \mathbf{S}^{-1}\mathbf{x}$; as $\mathbf{x}$ runs over F^n so does $\mathbf{y}$, and $\mathbf{x} \neq \mathbf{0}$ exactly when $\mathbf{y} \neq \mathbf{0}$. Then

$$\mathbf{x}^*\mathbf{B}\mathbf{x} = \mathbf{y}^*\mathbf{S}^*\mathbf{B}\mathbf{S}\mathbf{y} = \mathbf{y}^*\mathbf{D}\mathbf{y} = \sum_i d_i |y_i|^2.$$

If every $d_i \geq 0$, this is ≥ 0 for all $\mathbf{y}$, so $\mathbf{B} \succeq \mathbf{O}$. Conversely, taking $\mathbf{y} = \mathbf{e}_i$ gives $d_i \geq 0$. The same argument with strict inequalities gives the second claim, using that $\sum_i d_i |y_i|^2 > 0$ for $\mathbf{y} \neq \mathbf{0}$ when every $d_i > 0$.

(b) Taking determinants in $\mathbf{S}^*\mathbf{B}\mathbf{S} = \mathbf{D}$ and $\mathbf{S}^*\mathbf{A}\mathbf{S} = \mathbf{I}$ gives $|\det \mathbf{S}|^2 \det \mathbf{B} = \prod_i d_i$ and $|\det \mathbf{S}|^2 \det \mathbf{A} = 1$. Dividing, $\det \mathbf{B} / \det \mathbf{A} = \prod_i d_i$, which is > 0 by (a).

Exercise. C3: Dropping definiteness of A

Let $\mathbf{A} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$ and $\mathbf{B} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$, so that $\mathbf{A} \succeq \mathbf{O}$ but $\mathbf{A} \not\succ \mathbf{O}$.

(a) Compute $\det(\mathbf{B} - \lambda \mathbf{A})$ and deduce that the pair has no generalized eigenvalues at all.

(b) Prove that no invertible $\mathbf{S} \in M_2(\mathbb{R})$ makes $\mathbf{S}^{\top}\mathbf{A}\mathbf{S}$ and $\mathbf{S}^{\top}\mathbf{B}\mathbf{S}$ both diagonal.

Hint for (b): compare the degree in λ of $\det(\mathbf{B} - \lambda \mathbf{A})$ on both sides.

Solution.

(a) $\mathbf{B} - \lambda \mathbf{A} = \begin{pmatrix} -\lambda & 1 \\ 1 & 0 \end{pmatrix}$, with determinant -1. This is never 0, so $\mathbf{B} - \lambda \mathbf{A}$ is invertible for every λ and there is no $\mathbf{x} \neq \mathbf{0}$ with $\mathbf{B}\mathbf{x} = \lambda \mathbf{A}\mathbf{x}$.

(b) Suppose $\mathbf{S}^{\top}\mathbf{A}\mathbf{S} = \operatorname{diag}(a_1, a_2)$ and $\mathbf{S}^{\top}\mathbf{B}\mathbf{S} = \operatorname{diag}(b_1, b_2)$. As in the proof of Theorem 13.4.2 (b),

$$\det(\mathbf{B} - \lambda \mathbf{A}) = \frac{1}{(\det \mathbf{S})^2}(b_1 - \lambda a_1)(b_2 - \lambda a_2).$$

By (a) the left side is the constant -1, so the right side is a nonzero constant. It is nonzero, so $b_1 b_2 \neq 0$ and hence $\det(\mathbf{S}^{\top}\mathbf{B}\mathbf{S}) \neq 0$; it is constant in λ, so $a_1 = a_2 = 0$, which forces $\mathbf{S}^{\top}\mathbf{A}\mathbf{S} = \mathbf{O}$ and therefore $\mathbf{A} = \mathbf{O}$. That is false, so no such $\mathbf{S}$ exists. (Alternatively: $\operatorname{rank}(\mathbf{S}^{\top}\mathbf{A}\mathbf{S}) = \operatorname{rank} \mathbf{A} = 1$, so exactly one a_i is nonzero, and the right side then has degree 1 in λ, not 0.)

13.5 The Loewner Order

Every theorem so far in this chapter has compared a self-adjoint matrix with 0. Comparing it with 0 is only the first case of comparing it with something, and the analogy with real numbers suggests the general one: say that $\mathbf{A}$ is at least $\mathbf{B}$ when the difference is nonnegative. That is a genuine order, and a surprising amount of ordinary arithmetic survives it. A little does not, and what fails is the point of the section: squaring both sides of an inequality between matrices is not allowed.

Throughout, $F = \mathbb{R}$ or $F = \mathbb{C}$, all matrices are in $M_n(F)$, and $\lambda_1(\mathbf{A}) \geq \cdots \geq \lambda_n(\mathbf{A})$ are the eigenvalues of a Hermitian $\mathbf{A}$ in decreasing order.

13.5.1 Comparing two self-adjoint matrices

The relation $\mathbf{A} \succeq \mathbf{O}$ already behaves like "≥ 0" for numbers: it is preserved by sums, by nonnegative scalars, and by congruence (Proposition 13.1.2). For real numbers, $a \geq b$ means nothing other than $a - b \geq 0$. Copying that sentence is the whole definition.

One self-adjoint matrix dominates another when their difference is positive semidefinite.

Definition 13.5.1. The Loewner order

Let $\mathbf{A}, \mathbf{B} \in M_n(F)$ be Hermitian. Write

$$\mathbf{A} \succeq \mathbf{B} \quad \text{to mean} \quad \mathbf{A} - \mathbf{B} \succeq \mathbf{O},$$

and $\mathbf{A} \succ \mathbf{B}$ to mean $\mathbf{A} - \mathbf{B} \succ \mathbf{O}$. This relation on the set of Hermitian matrices is the **Loewner order**.

In words: $\mathbf{A} \succeq \mathbf{B}$ says that $\langle \mathbf{A}\mathbf{x}, \mathbf{x} \rangle \geq \langle \mathbf{B}\mathbf{x}, \mathbf{x} \rangle$ for **every** $\mathbf{x} \in F^n$, since $\langle (\mathbf{A} - \mathbf{B})\mathbf{x}, \mathbf{x} \rangle = \langle \mathbf{A}\mathbf{x}, \mathbf{x} \rangle - \langle \mathbf{B}\mathbf{x}, \mathbf{x} \rangle$. So the order compares the two quadratic forms pointwise, and it is exactly as strong as that comparison. Note also that the difference of two Hermitian matrices is Hermitian, so clause (P1) of Definition 13.1.1 is automatic here and only (P2) has to be checked. Taking $\mathbf{B} = \mathbf{O}$ recovers $\mathbf{A} \succeq \mathbf{O}$, so the notation is consistent.

Three examples, each of which will be used later.

- **Diagonal matrices.** $\operatorname{diag}(a_1, \ldots, a_n) \succeq \operatorname{diag}(b_1, \ldots, b_n)$ if and only if $a_i \geq b_i$ for every i, since the difference is diagonal and a diagonal Hermitian matrix is positive semidefinite exactly when its entries, which are its eigenvalues, are ≥ 0 (Theorem 13.1.4).
- **Eigenvalue bounds.** For Hermitian $\mathbf{A}$, $\lambda_1(\mathbf{A})\mathbf{I} \succeq \mathbf{A} \succeq \lambda_n(\mathbf{A})\mathbf{I}$. Indeed $\lambda_1(\mathbf{A})\mathbf{I} - \mathbf{A}$ is Hermitian with eigenvalues $\lambda_1(\mathbf{A}) - \lambda_i(\mathbf{A}) \geq 0$.
- **Projections.** If $\mathbf{P} \in M_n(F)$ is the matrix of an orthogonal projection onto a subspace of F^n, then $\mathbf{I} \succeq \mathbf{P} \succeq \mathbf{O}$: it is Hermitian, and its eigenvalues are 0 and 1.

A **nonexample by minimal change**: replace $\operatorname{diag}(1, 1)$ by $\operatorname{diag}(1, -1)$ in the first item. Then $\operatorname{diag}(1, -1) \succeq \operatorname{diag}(0, 0)$ fails, and it fails only in the second coordinate — the test is coordinate-wise for diagonal matrices, and one bad coordinate is enough.

Proposition 13.5.1. The Loewner order is a partial order

On the set of Hermitian matrices in $M_n(F)$, the relation $\succeq$ is reflexive, transitive, and

antisymmetric: $\mathbf{A} \succeq \mathbf{A}$; if $\mathbf{A} \succeq \mathbf{B}$ and $\mathbf{B} \succeq \mathbf{C}$ then $\mathbf{A} \succeq \mathbf{C}$; and if $\mathbf{A} \succeq \mathbf{B}$ and $\mathbf{B} \succeq \mathbf{A}$ then $\mathbf{A} = \mathbf{B}$.

Proof. Reflexivity: $\mathbf{A} - \mathbf{A} = \mathbf{O}$, and $\langle \mathbf{O}\mathbf{x}, \mathbf{x} \rangle = 0 \geq 0$.

Transitivity: $\mathbf{A} - \mathbf{C} = (\mathbf{A} - \mathbf{B}) + (\mathbf{B} - \mathbf{C})$, and for every $\mathbf{x}$,

$$\langle (\mathbf{A} - \mathbf{C})\mathbf{x}, \mathbf{x} \rangle = \langle (\mathbf{A} - \mathbf{B})\mathbf{x}, \mathbf{x} \rangle + \langle (\mathbf{B} - \mathbf{C})\mathbf{x}, \mathbf{x} \rangle \;\geq\; 0,$$

both summands being ≥ 0 by hypothesis.

Antisymmetry: put $\mathbf{S} = \mathbf{A} - \mathbf{B}$, which is Hermitian. From $\mathbf{A} \succeq \mathbf{B}$ we get $\langle \mathbf{S}\mathbf{x}, \mathbf{x} \rangle \geq 0$ for every $\mathbf{x}$, and from $\mathbf{B} \succeq \mathbf{A}$ we get $\langle -\mathbf{S}\mathbf{x}, \mathbf{x} \rangle \geq 0$, that is $\langle \mathbf{S}\mathbf{x}, \mathbf{x} \rangle \leq 0$. Hence $\langle \mathbf{S}\mathbf{x}, \mathbf{x} \rangle = 0$ for every $\mathbf{x} \in F^n$, and $\mathbf{S}$ is self-adjoint, so $\mathbf{S} = \mathbf{O}$ by Theorem 12.2.5. Thus $\mathbf{A} = \mathbf{B}$. ∎

Antisymmetry is the clause that needed an argument, and the argument is the one the chapter keeps reusing: a self-adjoint matrix is determined by its quadratic form. Over $\mathbb{R}$ this is genuinely a theorem — the quarter turn of Definition 13.1.1's warning has vanishing quadratic form and is not $\mathbf{O}$ — and it is self-adjointness that rules that out.

Warning. **The Loewner order is partial, and it is not the entrywise order.** Two things go wrong if one forgets this. First, most pairs are incomparable: $\mathbf{A} = \operatorname{diag}(1, 0)$ and $\mathbf{B} = \operatorname{diag}(0, 1)$ satisfy neither $\mathbf{A} \succeq \mathbf{B}$ nor $\mathbf{B} \succeq \mathbf{A}$, so an argument that splits into "either $\mathbf{A} \succeq \mathbf{B}$ or $\mathbf{B} \succeq \mathbf{A}$" is not an argument. Second, $\mathbf{A} \succeq \mathbf{B}$ says nothing entrywise, in either direction:

$$\begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} \succeq \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}, \qquad \text{yet the off-diagonal entries went down;}$$

the difference is $\left(\begin{smallmatrix} 1 & -1 \\ -1 & 1 \end{smallmatrix}\right)$, with eigenvalues 0 and 2. Conversely $\left(\begin{smallmatrix} 1 & 2 \\ 2 & 1 \end{smallmatrix}\right)$ has every entry ≥ 0 and is **not** $\succeq \mathbf{O}$, its eigenvalues being 3 and -1.

Quick check. Is the set of Hermitian $\mathbf{A}$ with $\mathbf{I} \succeq \mathbf{A} \succeq \mathbf{O}$ closed under multiplication?

Solution. No. Take $\mathbf{A} = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 0 \end{smallmatrix}\right)$ and $\mathbf{B} = \frac{1}{2}\left(\begin{smallmatrix} 1 & 1 \\ 1 & 1 \end{smallmatrix}\right)$, both orthogonal projections, so both lie in the set. Then $\mathbf{AB} = \frac{1}{2}\left(\begin{smallmatrix} 1 & 1 \\ 0 & 0 \end{smallmatrix}\right)$, which is not even Hermitian, so it is not in the set and the order does not apply to it. The order lives on Hermitian matrices, and Hermitian matrices are not closed under multiplication (Exercise (C2: Sums and products)).

13.5.2 What survives

One small fact is needed first, and it is worth stating on its own because it is the elementary half of a theme Chapter 17 takes up in full: the largest and smallest eigenvalues of a Hermitian matrix are the largest and smallest values of its quadratic form on the unit sphere.

Lemma 13.5.2. The extreme eigenvalues as extreme values

Let $\mathbf{A} \in M_n(F)$ be Hermitian. Then for every $\mathbf{x} \in F^n$ with $\|\mathbf{x}\| = 1$,

$$\lambda_n(\mathbf{A}) \;\le\; \langle \mathbf{A}\mathbf{x}, \mathbf{x}\rangle \;\le\; \lambda_1(\mathbf{A}),$$

and both bounds are attained, at unit eigenvectors for $\lambda_n(\mathbf{A})$ and $\lambda_1(\mathbf{A})$ respectively.

Proof. By Corollary 12.4.3 (over $\mathbb{C}$) or Corollary 12.5.2 (over $\mathbb{R}$), F^n has an orthonormal basis $(\mathbf{q}_1, \dots, \mathbf{q}_n)$ with $\mathbf{A}\mathbf{q}_i = \lambda_i(\mathbf{A})\mathbf{q}_i$. Write $\mathbf{x} = \sum_i c_i \mathbf{q}_i$; then $\sum_i |c_i|^2 = \|\mathbf{x}\|^2 = 1$ by Theorem 11.2.1, and

$$\langle \mathbf{A}\mathbf{x}, \mathbf{x}\rangle = \Big\langle \sum_i \lambda_i(\mathbf{A}) c_i \mathbf{q}_i, \sum_j c_j \mathbf{q}_j \Big\rangle = \sum_i \lambda_i(\mathbf{A}) |c_i|^2,$$

the cross terms vanishing by orthonormality. The right-hand side is a weighted average of the $\lambda_i(\mathbf{A})$ with nonnegative weights summing to 1, hence lies between the smallest and the largest. Taking $\mathbf{x} = \mathbf{q}_1$ gives $\lambda_1(\mathbf{A})$ and $\mathbf{x} = \mathbf{q}_n$ gives $\lambda_n(\mathbf{A})$.

■

Theorem 13.5.3. Basic Properties of the Loewner Order

Let $\mathbf{A}, \mathbf{B} \in M_n(F)$ be Hermitian with $\mathbf{A} \succeq \mathbf{B}$. Then:

(a) $\mathbf{S}^*\mathbf{A}\mathbf{S} \succeq \mathbf{S}^*\mathbf{B}\mathbf{S}$ for every $\mathbf{S} \in M_{n\times m}(F)$;
(b) $\lambda_1(\mathbf{A}) \ge \lambda_1(\mathbf{B})$ and $\lambda_n(\mathbf{A}) \ge \lambda_n(\mathbf{B})$;
(c) if moreover $\mathbf{B} \succeq \mathbf{O}$, then $\operatorname{tr}\mathbf{A} \ge \operatorname{tr}\mathbf{B}$ and $\det\mathbf{A} \ge \det\mathbf{B}$;
(d) if moreover $\mathbf{B} \succ \mathbf{O}$, then $\mathbf{A} \succ \mathbf{O}$ and $\mathbf{B}^{-1} \succeq \mathbf{A}^{-1}$.

Idea. Each clause has a one-line mechanism. (a) is the substitution $\mathbf{x} = \mathbf{S}\mathbf{y}$. (b) evaluates the quadratic form at the vector that is extremal for the *other* matrix. (c) splits: the trace is a sum of eigenvalues, and for the determinant we whiten by $\mathbf{B}^{-1/2}$ so that the comparison becomes "$\succeq \mathbf{I}$", where the determinant is visibly ≥ 1. (d) is the same whitening, plus the observation that inverting reverses the order for *numbers* ≥ 1.

Proof.

(a) This is Proposition 13.1.2 (a) applied to the Hermitian matrix $\mathbf{A} - \mathbf{B} \succeq \mathbf{O}$, together with $\mathbf{S}^*\mathbf{A}\mathbf{S} - \mathbf{S}^*\mathbf{B}\mathbf{S} = \mathbf{S}^*(\mathbf{A} - \mathbf{B})\mathbf{S}$.

(b) Let $\mathbf{x}$ be a unit eigenvector of $\mathbf{B}$ for $\lambda_1(\mathbf{B})$. By Lemma 13.5.2 applied twice, and $\mathbf{A} \succeq \mathbf{B}$ in between,

$$\lambda_1(\mathbf{B}) = \langle \mathbf{B}\mathbf{x}, \mathbf{x}\rangle \le \langle \mathbf{A}\mathbf{x}, \mathbf{x}\rangle \le \lambda_1(\mathbf{A}).$$

Now let $\mathbf{y}$ be a unit eigenvector of $\mathbf{A}$ for $\lambda_n(\mathbf{A})$. Then

$$\lambda_n(\mathbf{A}) = \langle \mathbf{A}\mathbf{y}, \mathbf{y}\rangle \ge \langle \mathbf{B}\mathbf{y}, \mathbf{y}\rangle \ge \lambda_n(\mathbf{B}).$$

(c) The matrix $\mathbf{A}-\mathbf{B}$ is Hermitian and $\succeq \mathbf{O}$, so all its eigenvalues are ≥ 0 by Theorem 13.1.4, and $\operatorname{tr}(\mathbf{A}-\mathbf{B})$, being their sum (Theorem 9.3.3), is ≥ 0. Hence $\operatorname{tr}\mathbf{A} \geq \operatorname{tr}\mathbf{B}$. For the determinant, first note $\mathbf{A} \succeq \mathbf{B} \succeq \mathbf{O}$ gives $\mathbf{A} \succeq \mathbf{O}$ by Proposition 13.5.1, so $\det\mathbf{A} \geq 0$ and $\det\mathbf{B} \geq 0$, both being products of nonnegative eigenvalues. If $\mathbf{B}$ is singular then $\det\mathbf{B} = 0 \leq \det\mathbf{A}$ and we are done. Otherwise $\mathbf{B} \succeq \mathbf{O}$ with no zero eigenvalue means $\mathbf{B} \succ \mathbf{O}$ (Theorem 13.1.5), so $\mathbf{B}^{1/2} \succ \mathbf{O}$ is invertible and we may set $\mathbf{C} = \mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2}$, where $\mathbf{B}^{-1/2} = (\mathbf{B}^{1/2})^{-1}$ is Hermitian. By (a) with $\mathbf{S} = \mathbf{B}^{-1/2}$,

$$\mathbf{C} = \mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2} \;\succeq\; \mathbf{B}^{-1/2}\mathbf{B}\mathbf{B}^{-1/2} = \mathbf{I}_n,$$

so $\mathbf{C}-\mathbf{I} \succeq \mathbf{O}$ and every eigenvalue of $\mathbf{C}$ is ≥ 1. Hence $\det\mathbf{C} \geq 1$. Since $\mathbf{A} = \mathbf{B}^{1/2}\mathbf{C}\mathbf{B}^{1/2}$, Theorem 7.5.4 gives

$$\det\mathbf{A} = \det(\mathbf{B}^{1/2})^2 \det\mathbf{C} = \det\mathbf{B}\cdot\det\mathbf{C} \;\geq\; \det\mathbf{B}.$$

(d) Suppose $\mathbf{B} \succ \mathbf{O}$. For $\mathbf{x} \neq \mathbf{0}$, $\langle\mathbf{A}\mathbf{x},\mathbf{x}\rangle \geq \langle\mathbf{B}\mathbf{x},\mathbf{x}\rangle > 0$, so $\mathbf{A} \succ \mathbf{O}$ and in particular $\mathbf{A}$ is invertible (Theorem 13.1.5). Let $\mathbf{C} = \mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2} \succeq \mathbf{I}$ as in (c). Then $\mathbf{C} \succ \mathbf{O}$, so $\mathbf{C}$ is invertible, and $\mathbf{C}^{-1}$ is Hermitian with eigenvalues the reciprocals of those of $\mathbf{C}$, all lying in $(0,1]$. Therefore $\mathbf{I}-\mathbf{C}^{-1}$ is Hermitian with nonnegative eigenvalues, that is $\mathbf{I} \succeq \mathbf{C}^{-1}$ (Theorem 13.1.4). Applying (a) with $\mathbf{S} = \mathbf{B}^{-1/2}$ to this inequality,

$$\mathbf{B}^{-1} = \mathbf{B}^{-1/2}\mathbf{I}\mathbf{B}^{-1/2} \;\succeq\; \mathbf{B}^{-1/2}\mathbf{C}^{-1}\mathbf{B}^{-1/2} = \mathbf{B}^{-1/2}\mathbf{B}^{1/2}\mathbf{A}^{-1}\mathbf{B}^{1/2}\mathbf{B}^{-1/2} = \mathbf{A}^{-1},$$

using $\mathbf{C}^{-1} = (\mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2})^{-1} = \mathbf{B}^{1/2}\mathbf{A}^{-1}\mathbf{B}^{1/2}$. This proves the theorem.

■

Clause (d) is the memorable one: **the order reverses under inversion**, exactly as $a \geq b > 0$ gives $1/b \geq 1/a$ for numbers. Clause (b) is the beginning of a longer story. The full statement is that $\lambda_i(\mathbf{A}) \geq \lambda_i(\mathbf{B})$ for **every** i, not only for $i = 1$ and $i = n$; but the intermediate eigenvalues have no description as a plain maximum over the whole unit sphere, and the min–max description that does the job is the Courant–Fischer theorem of Chapter 17. Nothing in this chapter uses the intermediate cases, so we prove the two extreme ones and leave the rest there.

Remark. Clause (c) does not extend to $\mathbf{A} \succeq \mathbf{B}$ without $\mathbf{B} \succeq \mathbf{O}$. Take $\mathbf{A} = \operatorname{diag}(1,-1)$ and $\mathbf{B} = \operatorname{diag}(0,-2)$: then $\mathbf{A}-\mathbf{B} = \operatorname{diag}(1,1) \succeq \mathbf{O}$, so $\mathbf{A} \succeq \mathbf{B}$, yet $\det\mathbf{A} = -1 < 0 = \det\mathbf{B}$. The hypothesis $\mathbf{B} \succeq \mathbf{O}$ is what lets "bigger" be read off the eigenvalues one at a time, and so multiplied up into a determinant.

13.5.3 What fails: squaring

For positive numbers, $a \geq b \geq 0$ implies $a^2 \geq b^2$. The matrix version is false, and the smallest counterexample has integer entries.

Example. Squaring does not preserve the Loewner order

Let

$$\mathbf{A} = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}, \qquad \mathbf{B} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}.$$

Show that $\mathbf{A} \succeq \mathbf{B} \succeq \mathbf{O}$ and that $\mathbf{A}^2 \succeq \mathbf{B}^2$ is false.

Solution. The hypotheses. Both matrices are real symmetric. $\mathbf{B} \succeq \mathbf{O}$: its leading principal minors are 1 and 0, and directly $\mathbf{x}^\top \mathbf{B}\mathbf{x} = (x_1 + x_2)^2 \geq 0$. Its eigenvalues are 0 and 2. Next,

$$\mathbf{A} - \mathbf{B} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \succeq \mathbf{O},$$

so $\mathbf{A} \succeq \mathbf{B}$. (In fact $\mathbf{A} \succ \mathbf{O}$: $\det \mathbf{A} = 1 > 0$ and $a_{11} = 2 > 0$, so Sylvester's criterion applies, Theorem 13.1.5.)

The squares.

$$\mathbf{A}^2 = \begin{pmatrix} 5 & 3 \\ 3 & 2 \end{pmatrix}, \qquad \mathbf{B}^2 = \begin{pmatrix} 2 & 2 \\ 2 & 2 \end{pmatrix}, \qquad \mathbf{A}^2 - \mathbf{B}^2 = \begin{pmatrix} 3 & 1 \\ 1 & 0 \end{pmatrix}.$$

Its determinant is $0 - 1 = -1 < 0$, so its two eigenvalues have opposite signs and it is not positive semidefinite (Theorem 13.1.4). A witness is available without computing eigenvalues: with $\mathbf{x} = (1, -2)$,

$$\mathbf{x}^\top (\mathbf{A}^2 - \mathbf{B}^2)\mathbf{x} = 3 - 4 + 0 = -1 < 0.$$

So $\mathbf{A}^2 \not\succeq \mathbf{B}^2$.

Where does the argument for numbers break? It uses $a^2 - b^2 = (a - b)(a + b)$, a product of two nonnegative numbers. For matrices, $(\mathbf{A} - \mathbf{B})(\mathbf{A} + \mathbf{B}) = \mathbf{A}^2 + \mathbf{A}\mathbf{B} - \mathbf{B}\mathbf{A} - \mathbf{B}^2$, which is $\mathbf{A}^2 - \mathbf{B}^2$ only when $\mathbf{A}$ and $\mathbf{B}$ commute — and in the example $\mathbf{A}\mathbf{B} = \left(\begin{smallmatrix} 3 & 3 \\ 2 & 2 \end{smallmatrix}\right)$ while $\mathbf{B}\mathbf{A} = \left(\begin{smallmatrix} 3 & 2 \\ 3 & 2 \end{smallmatrix}\right)$. Noncommutativity is the whole obstruction, and when it is removed the number argument goes through; that is exercise C1.

What *is* true is the same statement for the square **root**, and it is true with no commuting hypothesis at all.

Proposition 13.5.4. The square root is order preserving

Let $\mathbf{A}, \mathbf{B} \in M_n(F)$ with $\mathbf{A} \succeq \mathbf{B} \succeq \mathbf{O}$. Then $\mathbf{A}^{1/2} \succeq \mathbf{B}^{1/2}$.

Idea. Suppose not. Then $\mathbf{X} - \mathbf{Y}$, with $\mathbf{X} = \mathbf{A}^{1/2}$ and $\mathbf{Y} = \mathbf{B}^{1/2}$, has a negative eigenvalue μ, with unit eigenvector $\mathbf{u}$. Two different consequences of $\mathbf{X}\mathbf{u} = \mathbf{Y}\mathbf{u} + \mu\mathbf{u}$ then collide. Pairing with $\mathbf{u}$ says $\langle \mathbf{Y}\mathbf{u}, \mathbf{u} \rangle \geq |\mu|$, because $\langle \mathbf{X}\mathbf{u}, \mathbf{u} \rangle \geq 0$. Taking norms and using $\|\mathbf{X}\mathbf{u}\|^2 \geq \|\mathbf{Y}\mathbf{u}\|^2$, which is exactly the hypothesis $\mathbf{A} \succeq \mathbf{B}$, says $|\mu| \geq 2\langle \mathbf{Y}\mathbf{u}, \mathbf{u} \rangle$. Together: $|\mu| \geq 2|\mu|$.

Proof. Write $\mathbf{X} = \mathbf{A}^{1/2}$ and $\mathbf{Y} = \mathbf{B}^{1/2}$, both Hermitian and $\succeq \mathbf{O}$ with $\mathbf{X}^2 = \mathbf{A}$ and $\mathbf{Y}^2 = \mathbf{B}$ (Theorem 13.2.1). Then $\mathbf{X} - \mathbf{Y}$ is Hermitian. Suppose, for a contradiction, that $\mathbf{X} - \mathbf{Y} \not\succeq \mathbf{O}$.

By Theorem 13.1.4 it has a negative eigenvalue $\mu < 0$; choose a unit eigenvector $\mathbf{u}$, so that

$$\mathbf{X}\mathbf{u} = \mathbf{Y}\mathbf{u} + \mu\mathbf{u}.$$

Pairing this with $\mathbf{u}$ and using $\|\mathbf{u}\| = 1$,

$$\langle \mathbf{X}\mathbf{u}, \mathbf{u}\rangle = \langle \mathbf{Y}\mathbf{u}, \mathbf{u}\rangle + \mu.$$

All three numbers are real (Proposition 11.6.7), and $\langle \mathbf{X}\mathbf{u}, \mathbf{u}\rangle \geq 0$ because $\mathbf{X} \succeq \mathbf{O}$. Therefore

$$\langle \mathbf{Y}\mathbf{u}, \mathbf{u}\rangle \;\geq\; -\mu = |\mu|. \tag{$*$}$$

Now take squared norms. Since $\mathbf{X}$ is Hermitian, $\|\mathbf{X}\mathbf{u}\|^2 = \langle \mathbf{X}^2\mathbf{u}, \mathbf{u}\rangle = \langle \mathbf{A}\mathbf{u}, \mathbf{u}\rangle$, and likewise $\|\mathbf{Y}\mathbf{u}\|^2 = \langle \mathbf{B}\mathbf{u}, \mathbf{u}\rangle$; so $\mathbf{A} \succeq \mathbf{B}$ gives $\|\mathbf{X}\mathbf{u}\|^2 \geq \|\mathbf{Y}\mathbf{u}\|^2$. On the other hand, expanding $\mathbf{X}\mathbf{u} = \mathbf{Y}\mathbf{u} + \mu\mathbf{u}$,

$$\|\mathbf{X}\mathbf{u}\|^2 = \|\mathbf{Y}\mathbf{u}\|^2 + 2\mu\langle \mathbf{Y}\mathbf{u}, \mathbf{u}\rangle + \mu^2,$$

where the cross term is $2\mu\langle \mathbf{Y}\mathbf{u}, \mathbf{u}\rangle$ because μ and $\langle \mathbf{Y}\mathbf{u}, \mathbf{u}\rangle$ are real. Comparing the two,

$$2\mu\langle \mathbf{Y}\mathbf{u}, \mathbf{u}\rangle + \mu^2 \;\geq\; 0.$$

Substituting $\mu = -|\mu|$ turns this into $|\mu|^2 \geq 2|\mu|\langle \mathbf{Y}\mathbf{u}, \mathbf{u}\rangle$, and dividing by $|\mu| > 0$ gives

$$|\mu| \;\geq\; 2\langle \mathbf{Y}\mathbf{u}, \mathbf{u}\rangle. \tag{$**$}$$

Combining $(*)$ and $(**)$, $|\mu| \geq 2|\mu|$, so $|\mu| \leq 0$ and $\mu = 0$, contradicting $\mu < 0$. Hence $\mathbf{X} - \mathbf{Y} \succeq \mathbf{O}$, which is the assertion.

■

So among the powers $t \mapsto t^s$, the exponent $s = 1/2$ preserves the Loewner order while $s = 2$ does not. The dividing line is at $s = 1$: the functions preserving the order are called **operator monotone**, and $t \mapsto t^s$ is operator monotone on $[0, \infty)$ exactly for $0 \leq s \leq 1$. Chapter 21 proves this, along with the companion fact that $t \mapsto \log t$ is operator monotone while $t \mapsto e^t$ is not.

13.5.4 Exercises

13.5.4.1 A. Check your understanding

Exercise. A1

(a) Define $\mathbf{A} \succeq \mathbf{B}$, and say on which set of matrices the relation is defined.

(b) Determine whether the following statement is correct, and justify your answer: for Hermitian $\mathbf{A}, \mathbf{B}$, either $\mathbf{A} \succeq \mathbf{B}$ or $\mathbf{B} \succeq \mathbf{A}$.

(c) Determine whether the following statement is correct, and justify your answer: $\mathbf{A} \succeq \mathbf{B}$ implies $a_{ij} \geq b_{ij}$ for all i, j.

(d) State what fails when both sides of $\mathbf{A} \succeq \mathbf{B} \succeq \mathbf{O}$ are squared, and what happens instead when both sides have their square roots taken.

(e) Which clause of Theorem 13.5.3 reverses an inequality, and what is the analogous fact about real numbers?

Solution.

(a) For Hermitian $\mathbf{A}, \mathbf{B} \in M_n(F)$, $\mathbf{A} \succeq \mathbf{B}$ means $\mathbf{A} - \mathbf{B} \succeq \mathbf{O}$, equivalently $\langle \mathbf{A}\mathbf{x}, \mathbf{x} \rangle \geq \langle \mathbf{B}\mathbf{x}, \mathbf{x} \rangle$ for every $\mathbf{x} \in F^n$ (Definition 13.5.1). It is defined on the Hermitian matrices of a fixed size.

(b) Incorrect: the order is partial. $\operatorname{diag}(1, 0)$ and $\operatorname{diag}(0, 1)$ are incomparable, since $\operatorname{diag}(1, -1)$ and $\operatorname{diag}(-1, 1)$ are both indefinite.

(c) Incorrect. $\operatorname{diag}(2, 2) \succeq \left(\begin{smallmatrix} 1 & 1 \\ 1 & 1 \end{smallmatrix}\right)$, because the difference $\left(\begin{smallmatrix} 1 & -1 \\ -1 & 1 \end{smallmatrix}\right)$ has eigenvalues 0 and 2; yet the $(1, 2)$-entries are 0 and 1.

(d) Squaring fails: $\mathbf{A} \succeq \mathbf{B} \succeq \mathbf{O}$ does not give $\mathbf{A}^2 \succeq \mathbf{B}^2$ (Example (Squaring does not preserve the Loewner order)). Taking square roots is legitimate: $\mathbf{A}^{1/2} \succeq \mathbf{B}^{1/2}$ (Proposition 13.5.4).

(e) Clause (d): $\mathbf{A} \succeq \mathbf{B} \succ \mathbf{O}$ gives $\mathbf{B}^{-1} \succeq \mathbf{A}^{-1}$. For real numbers, $a \geq b > 0$ gives $1/b \geq 1/a$.

13.5.4.2 B. Practice

Exercise. B1: Decide the relation

For each pair of real symmetric matrices, determine whether $\mathbf{A} \succeq \mathbf{B}$, whether $\mathbf{B} \succeq \mathbf{A}$, or neither. Justify your answer.

(i) $\mathbf{A} = \begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix}$, $\mathbf{B} = \mathbf{I}_2$.

(ii) $\mathbf{A} = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}$, $\mathbf{B} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$.

(iii) $\mathbf{A} = \begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix}$, $\mathbf{B} = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}$.

Solution.

(i) $\mathbf{A} - \mathbf{B} = \left(\begin{smallmatrix} 2 & 1 \\ 1 & 2 \end{smallmatrix}\right)$, with leading principal minors 2 and 3, both > 0, so $\mathbf{A} - \mathbf{B} \succ \mathbf{O}$ by Theorem 13.1.5 and $\mathbf{A} \succ \mathbf{B}$. Then $\mathbf{B} \succeq \mathbf{A}$ would force $\mathbf{A} = \mathbf{B}$ by Proposition 13.5.1, which is false, so $\mathbf{B} \not\succeq \mathbf{A}$.

(ii) $\mathbf{A} - \mathbf{B} = \left(\begin{smallmatrix} 1 & -1 \\ -1 & 0 \end{smallmatrix}\right)$, with determinant $-1 < 0$, so it is indefinite and $\mathbf{A} \not\succeq \mathbf{B}$. Also $\mathbf{B} - \mathbf{A}$ has determinant -1, so $\mathbf{B} \not\succeq \mathbf{A}$. Neither.

(iii) $\mathbf{A} - \mathbf{B} = \operatorname{diag}(-1, 2)$, indefinite; $\mathbf{B} - \mathbf{A} = \operatorname{diag}(1, -2)$, indefinite. Neither. (For diagonal matrices the comparison is entrywise, and the two entries disagree.)

Exercise. B2: The order reverses

Let $\mathbf{A} = \begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix}$ and $\mathbf{B} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$. Verify that $\mathbf{A} \succeq \mathbf{B} \succ \mathbf{O}$, compute $\mathbf{A}^{-1}$ and $\mathbf{B}^{-1}$, and check Theorem 13.5.3 (d) directly. Check the determinant clause as well.

Solution. $\mathbf{A}-\mathbf{B}=\mathbf{I}_2\succ\mathbf{O}$, so $\mathbf{A}\succ\mathbf{B}$. And $\mathbf{B}\succ\mathbf{O}$, its leading principal minors being 2 and 3. Then $\det\mathbf{A}=8$, $\det\mathbf{B}=3$, so

$$\mathbf{A}^{-1}=\frac{1}{8}\begin{pmatrix}3&-1\\-1&3\end{pmatrix},\qquad \mathbf{B}^{-1}=\frac{1}{3}\begin{pmatrix}2&-1\\-1&2\end{pmatrix},$$

and

$$\mathbf{B}^{-1}-\mathbf{A}^{-1}=\frac{1}{24}\begin{pmatrix}16-9&-8+3\\-8+3&16-9\end{pmatrix}=\frac{1}{24}\begin{pmatrix}7&-5\\-5&7\end{pmatrix}.$$

Its leading principal minors are $7/24>0$ and $(49-25)/576=1/24>0$, so $\mathbf{B}^{-1}\succ\mathbf{A}^{-1}$, as Theorem 13.5.3 (d) predicts. For the determinant clause, $\det\mathbf{A}=8\geq 3=\det\mathbf{B}$; and for the trace, $6\geq 4$.

Exercise. B3: Another failure of squaring

Let $\mathbf{A}=\begin{pmatrix}3&1\\1&1\end{pmatrix}$ and $\mathbf{B}=\begin{pmatrix}1&1\\1&1\end{pmatrix}$. Verify that $\mathbf{A}\succeq\mathbf{B}\succeq\mathbf{O}$, and exhibit a vector $\mathbf{x}$ with $\mathbf{x}^{\top}(\mathbf{A}^2-\mathbf{B}^2)\mathbf{x}<0$.

Solution. $\mathbf{A}-\mathbf{B}=\operatorname{diag}(2,0)\succeq\mathbf{O}$, so $\mathbf{A}\succeq\mathbf{B}$; and $\mathbf{x}^{\top}\mathbf{B}\mathbf{x}=(x_1+x_2)^2\geq 0$, so $\mathbf{B}\succeq\mathbf{O}$. Squaring,

$$\mathbf{A}^2=\begin{pmatrix}10&4\\4&2\end{pmatrix},\qquad \mathbf{B}^2=\begin{pmatrix}2&2\\2&2\end{pmatrix},\qquad \mathbf{A}^2-\mathbf{B}^2=\begin{pmatrix}8&2\\2&0\end{pmatrix},$$

whose determinant is $-4<0$. With $\mathbf{x}=(1,-3)$,

$$\mathbf{x}^{\top}(\mathbf{A}^2-\mathbf{B}^2)\mathbf{x}=8-12+0=-4<0.$$

So $\mathbf{A}^2\not\succeq\mathbf{B}^2$. (Any $\mathbf{x}=(1,t)$ with $t<-2$ works, since the value is $8+4t$.)

13.5.4.3 C. Going deeper

Exercise. C1: When squaring does work

Let $\mathbf{A},\mathbf{B}\in M_n(F)$ with $\mathbf{A}\succeq\mathbf{B}\succeq\mathbf{O}$ and $\mathbf{AB}=\mathbf{BA}$.

(a) Prove that if $\mathbf{P},\mathbf{Q}\succeq\mathbf{O}$ and $\mathbf{PQ}=\mathbf{QP}$, then $\mathbf{PQ}\succeq\mathbf{O}$.

(b) Deduce that $\mathbf{A}^2\succeq\mathbf{B}^2$.

(c) Deduce that $\mathbf{A}^k\succeq\mathbf{B}^k$ for every integer $k\geq 1$.

Hint for (a): $\mathbf{P}^{1/2}$ *is a polynomial in* $\mathbf{P}$.

Solution.

(a) By Theorem 13.2.1, $\mathbf{P}^{1/2}$ is a polynomial in $\mathbf{P}$, so it commutes with every matrix that

commutes with $\mathbf{P}$; in particular $\mathbf{P}^{1/2}\mathbf{Q} = \mathbf{Q}\mathbf{P}^{1/2}$. Hence

$$\mathbf{P}\mathbf{Q} = \mathbf{P}^{1/2}\mathbf{P}^{1/2}\mathbf{Q} = \mathbf{P}^{1/2}\mathbf{Q}\mathbf{P}^{1/2} = (\mathbf{P}^{1/2})^{*}\mathbf{Q}\mathbf{P}^{1/2},$$

which is $\succeq \mathbf{O}$ by Proposition 13.1.2 (a), since $\mathbf{Q} \succeq \mathbf{O}$ and $\mathbf{P}^{1/2}$ is Hermitian.

(b) Put $\mathbf{P} = \mathbf{A} - \mathbf{B}$ and $\mathbf{Q} = \mathbf{A} + \mathbf{B}$. Both are $\succeq \mathbf{O}$: the first by hypothesis, the second as a sum of two matrices $\succeq \mathbf{O}$, using $\mathbf{A} \succeq \mathbf{B} \succeq \mathbf{O}$ and Proposition 13.5.1. They commute, because $\mathbf{AB} = \mathbf{BA}$ makes $\mathbf{PQ} = \mathbf{A}^2 + \mathbf{AB} - \mathbf{BA} - \mathbf{B}^2 = \mathbf{A}^2 - \mathbf{B}^2$ and $\mathbf{QP} = \mathbf{A}^2 - \mathbf{AB} + \mathbf{BA} - \mathbf{B}^2 = \mathbf{A}^2 - \mathbf{B}^2$ as well. By (a), $\mathbf{A}^2 - \mathbf{B}^2 = \mathbf{PQ} \succeq \mathbf{O}$.

(c) Induction on k. The case $k = 1$ is the hypothesis. Assume $\mathbf{A}^{k-1} \succeq \mathbf{B}^{k-1}$. Then

$$\mathbf{A}^k - \mathbf{B}^k = \mathbf{A}^{k-1}(\mathbf{A} - \mathbf{B}) + (\mathbf{A}^{k-1} - \mathbf{B}^{k-1})\mathbf{B},$$

and each term is $\succeq \mathbf{O}$ by (a). Indeed $\mathbf{A} \succeq \mathbf{O}$ is Hermitian with eigenvalues ≥ 0, so $\mathbf{A}^{k-1}$ is Hermitian with eigenvalues ≥ 0 and hence $\mathbf{A}^{k-1} \succeq \mathbf{O}$ (Theorem 13.1.4); it commutes with $\mathbf{A} - \mathbf{B}$, since $\mathbf{AB} = \mathbf{BA}$. Likewise $\mathbf{A}^{k-1} - \mathbf{B}^{k-1} \succeq \mathbf{O}$ by the induction hypothesis and it commutes with $\mathbf{B} \succeq \mathbf{O}$. Finally a sum of two matrices $\succeq \mathbf{O}$ is $\succeq \mathbf{O}$, directly from Definition 13.1.1.

Exercise. C2: Kernels and ranks

Let $\mathbf{A}, \mathbf{B} \in M_n(F)$ with $\mathbf{A} \succeq \mathbf{B} \succeq \mathbf{O}$.

(a) Prove that $\operatorname{nul}(\mathbf{A}) \subseteq \operatorname{nul}(\mathbf{B})$.

(b) Deduce that $\operatorname{rank}\mathbf{A} \geq \operatorname{rank}\mathbf{B}$.

(c) Give an example with $\mathbf{A} \succeq \mathbf{B} \succeq \mathbf{O}$ and $\operatorname{rank}\mathbf{A} = \operatorname{rank}\mathbf{B}$ but $\mathbf{A} \neq \mathbf{B}$.

Hint for (a): pair $\mathbf{Ax} = \mathbf{0}$ *with* $\mathbf{x}$*, then use* $\mathbf{B}^{1/2}$.

Solution.

(a) Let $\mathbf{x} \in \operatorname{nul}(\mathbf{A})$. Then $\langle \mathbf{Ax}, \mathbf{x}\rangle = 0$, and $0 \leq \langle \mathbf{Bx}, \mathbf{x}\rangle \leq \langle \mathbf{Ax}, \mathbf{x}\rangle = 0$, so $\langle \mathbf{Bx}, \mathbf{x}\rangle = 0$. Writing $\mathbf{B} = \mathbf{B}^{1/2}\mathbf{B}^{1/2}$ with $\mathbf{B}^{1/2}$ Hermitian (Theorem 13.2.1),

$$0 = \langle \mathbf{Bx}, \mathbf{x}\rangle = \langle \mathbf{B}^{1/2}\mathbf{x}, \mathbf{B}^{1/2}\mathbf{x}\rangle = \|\mathbf{B}^{1/2}\mathbf{x}\|^2,$$

so $\mathbf{B}^{1/2}\mathbf{x} = \mathbf{0}$ and hence $\mathbf{Bx} = \mathbf{B}^{1/2}(\mathbf{B}^{1/2}\mathbf{x}) = \mathbf{0}$.

(b) By (a), $\operatorname{nullity}\mathbf{A} \leq \operatorname{nullity}\mathbf{B}$, and by Rank–Nullity (Theorem 2.3.1) applied to both matrices on F^n, $\operatorname{rank}\mathbf{A} = n - \operatorname{nullity}\mathbf{A} \geq n - \operatorname{nullity}\mathbf{B} = \operatorname{rank}\mathbf{B}$.

(c) $\mathbf{A} = \operatorname{diag}(2, 2)$ and $\mathbf{B} = \operatorname{diag}(1, 1)$: both have rank 2, $\mathbf{A} - \mathbf{B} = \mathbf{I}_2 \succeq \mathbf{O}$, and $\mathbf{A} \neq \mathbf{B}$.

Exercise. C3: Largest eigenvalue of a sum

Let $\mathbf{A}, \mathbf{B} \in M_n(F)$ be Hermitian.

(a) Prove that $\lambda_1(\mathbf{A})\mathbf{I} \succeq \mathbf{A}$, and deduce $\lambda_1(\mathbf{A} + \mathbf{B}) \leq \lambda_1(\mathbf{A}) + \lambda_1(\mathbf{B})$.

(b) Deduce that if $\mathbf{A} \succeq \mathbf{B}$, then $0 \leq \lambda_1(\mathbf{A}) - \lambda_1(\mathbf{B}) \leq \lambda_1(\mathbf{A} - \mathbf{B})$.

Solution.

(a) The matrix $\lambda_1(\mathbf{A})\mathbf{I}-\mathbf{A}$ is Hermitian with eigenvalues $\lambda_1(\mathbf{A})-\lambda_i(\mathbf{A})\geq 0$, so it is $\succeq \mathbf{O}$ by Theorem 13.1.4; that is $\lambda_1(\mathbf{A})\mathbf{I}\succeq\mathbf{A}$. Adding the same inequality for $\mathbf{B}$, and using that a sum of two matrices $\succeq \mathbf{O}$ is again $\succeq \mathbf{O}$ (Exercise (C2: The cone of positive matrices) (a)),

$$(\lambda_1(\mathbf{A})+\lambda_1(\mathbf{B}))\mathbf{I}\ \succeq\ \mathbf{A}+\mathbf{B}.$$

By Theorem 13.5.3 (b) applied to this pair, $\lambda_1(\mathbf{A})+\lambda_1(\mathbf{B})=\lambda_1\big((\lambda_1(\mathbf{A})+\lambda_1(\mathbf{B}))\mathbf{I}\big)\geq \lambda_1(\mathbf{A}+\mathbf{B})$.

(b) Write $\mathbf{A}=\mathbf{B}+(\mathbf{A}-\mathbf{B})$. Applying (a) to the pair $\mathbf{B}$ and $\mathbf{A}-\mathbf{B}$,

$$\lambda_1(\mathbf{A})\leq\lambda_1(\mathbf{B})+\lambda_1(\mathbf{A}-\mathbf{B}),$$

so $\lambda_1(\mathbf{A})-\lambda_1(\mathbf{B})\leq\lambda_1(\mathbf{A}-\mathbf{B})$. For the lower bound, $\mathbf{A}\succeq\mathbf{B}$ gives $\lambda_1(\mathbf{A})\geq\lambda_1(\mathbf{B})$ by Theorem 13.5.3 (b). Note that (a) itself used no order hypothesis at all; only the lower bound needs $\mathbf{A}\succeq\mathbf{B}$.

13.6 Positive Matrices in Blocks

Chapter 8 cut a matrix into blocks, ran one step of elimination, and gave a name to what was left in the far corner: the Schur complement. Section 1 of this chapter met that same quantity as a single number, in the induction that proved Sylvester's criterion. This section puts the two together. For a Hermitian matrix split into blocks, positive definiteness can be decided one block at a time, and the deciding quantity is the Schur complement. The test is then spent on a question Chapter 12 left open: which matrices carry a positive definite solution of the Lyapunov equation, a certificate of stability that names no eigenvalue and computes none. What the same block decomposition says about determinants is the next section's subject.

Throughout, $F=\mathbb{R}$ or $F=\mathbb{C}$, and F^n carries the standard inner product.

13.6.1 What the blocks of a positive matrix must satisfy

A Hermitian matrix cut into blocks of sizes k and l has a shape forced on it. Suppose

$$\mathbf{M}=\begin{pmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{D} & \mathbf{C}\end{pmatrix}\in M_{k+l}(F),$$

with $\mathbf{A}\in M_k(F)$, $\mathbf{B}\in M_{k\times l}(F)$, $\mathbf{D}\in M_{l\times k}(F)$, and $\mathbf{C}\in M_l(F)$. Starring a block matrix transposes the arrangement and stars each block, so $\mathbf{M}^*=\mathbf{M}$ says exactly three things: $\mathbf{A}^*=\mathbf{A}$, $\mathbf{C}^*=\mathbf{C}$, and $\mathbf{D}=\mathbf{B}^*$. A Hermitian $\mathbf{M}$ is therefore always of the form

$$\mathbf{M}=\begin{pmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{B}^* & \mathbf{C}\end{pmatrix}$$

with $\mathbf{A}$ and $\mathbf{C}$ Hermitian and $\mathbf{B}$ unrestricted, and we write it this way from now on.

Two necessary conditions for positivity are immediate. The vectors of F^{k+l} whose last l coordinates vanish form a copy of F^k, and on them the quadratic form of $\mathbf{M}$ is the quadratic form of $\mathbf{A}$:

$$\begin{pmatrix}\mathbf{x}\\ \mathbf{0}\end{pmatrix}^{*}\mathbf{M}\begin{pmatrix}\mathbf{x}\\ \mathbf{0}\end{pmatrix}=\mathbf{x}^{*}\mathbf{A}\mathbf{x},\qquad \begin{pmatrix}\mathbf{0}\\ \mathbf{y}\end{pmatrix}^{*}\mathbf{M}\begin{pmatrix}\mathbf{0}\\ \mathbf{y}\end{pmatrix}=\mathbf{y}^{*}\mathbf{C}\mathbf{y}.$$

So $\mathbf{M}\succeq\mathbf{O}$ forces $\mathbf{A}\succeq\mathbf{O}$ and $\mathbf{C}\succeq\mathbf{O}$. This is the statement that a principal submatrix of a positive semidefinite matrix is positive semidefinite, used already in the proof of Theorem 13.1.4, here for the two index blocks $\{1,\dots,k\}$ and $\{k+1,\dots,k+l\}$.

The two conditions are not sufficient, and the smallest counterexample has $k=l=1$:

$$\mathbf{M}=\begin{pmatrix}0&1\\ 1&1\end{pmatrix},\qquad \mathbf{A}=\mathbf{O}_1\succeq\mathbf{O},\quad \mathbf{C}=(1)\succ\mathbf{O}.$$

Here $\det\mathbf{M}=-1<0$, so $\mathbf{M}\not\succeq\mathbf{O}$ by Theorem 13.1.4 (e). Something must couple $\mathbf{B}$ to $\mathbf{A}$ and $\mathbf{C}$, and the 2×2 case already says what. For $\mathbf{M}=\left(\begin{smallmatrix}a&b\\ \bar b&c\end{smallmatrix}\right)$ Hermitian, Example (The 2×2 Test) gives $\mathbf{M}\succ\mathbf{O}$ if and only if $a>0$ and $ac-|b|^2>0$. Divide the second condition by the positive number a:

$$c-\bar b\,a^{-1}b>0.$$

Written with $\mathbf{B}=(b)$, so that $\mathbf{B}^{*}=(\bar b)$, the left-hand side is the Schur complement $\mathbf{M}/\mathbf{A}$ of Definition 8.3.1. The general statement now writes itself: replace the scalars by blocks.

13.6.2 Testing one block at a time

Theorem 13.6.1. Positive Definiteness Block by Block

Let $k,l\geq 1$ and let

$$\mathbf{M}=\begin{pmatrix}\mathbf{A}&\mathbf{B}\\ \mathbf{B}^{*}&\mathbf{C}\end{pmatrix}\in M_{k+l}(F)$$

be Hermitian, with $\mathbf{A}\in M_k(F)$, $\mathbf{B}\in M_{k\times l}(F)$, and $\mathbf{C}\in M_l(F)$.

(a) If $\mathbf{M}\succeq\mathbf{O}$, then $\mathbf{A}\succeq\mathbf{O}$ and $\mathbf{C}\succeq\mathbf{O}$. If $\mathbf{M}\succ\mathbf{O}$, then $\mathbf{A}\succ\mathbf{O}$ and $\mathbf{C}\succ\mathbf{O}$.

(b) Suppose $\mathbf{A}\succ\mathbf{O}$. Then $\mathbf{A}$ is invertible, the Schur complement

$$\mathbf{M}/\mathbf{A}=\mathbf{C}-\mathbf{B}^{*}\mathbf{A}^{-1}\mathbf{B}\in M_l(F)$$

is Hermitian, and

$$\mathbf{M}\succ\mathbf{O}\iff\mathbf{M}/\mathbf{A}\succ\mathbf{O},\qquad \mathbf{M}\succeq\mathbf{O}\iff\mathbf{M}/\mathbf{A}\succeq\mathbf{O}.$$

Idea. Everything is one congruence. Chapter 8's block elimination clears $\mathbf{B}$ and $\mathbf{B}^*$ at once and leaves $\mathbf{A}$ and $\mathbf{M}/\mathbf{A}$ on the diagonal; for a Hermitian $\mathbf{M}$ the two outer factors of that elimination are adjoint to each other, because the row operation and the column operation are the same operation seen twice. So the elimination is a congruence $\mathbf{S}^*\mathbf{M}\mathbf{S}$, and Proposition 13.1.2 says that a congruence by an invertible matrix neither creates nor destroys positivity. All that remains is to read the verdict off a block **diagonal** matrix, where the quadratic form splits into two independent pieces. This is the step $(*)$ of Section 1's proof of Sylvester's criterion, with the last coordinate replaced by a block of l coordinates.

Proof.

(a) Let $\mathbf{x} \in F^k$ and put $\mathbf{v} = (\mathbf{x}, \mathbf{0}) \in F^{k+l}$. By Theorem 8.1.1, $\mathbf{M}\mathbf{v} = (\mathbf{A}\mathbf{x}, \mathbf{B}^*\mathbf{x})$, and hence $\mathbf{v}^*\mathbf{M}\mathbf{v} = \mathbf{x}^*\mathbf{A}\mathbf{x}$. If $\mathbf{M} \succeq \mathbf{O}$, the left side is ≥ 0 for every $\mathbf{x}$, so $\mathbf{A} \succeq \mathbf{O}$; if $\mathbf{M} \succ \mathbf{O}$, then $\mathbf{x} \neq \mathbf{0}$ forces $\mathbf{v} \neq \mathbf{0}$ and the left side is > 0, so $\mathbf{A} \succ \mathbf{O}$. The same argument with $\mathbf{v} = (\mathbf{0}, \mathbf{y})$ gives the statements for $\mathbf{C}$.

(b) Since $\mathbf{A} \succ \mathbf{O}$, every eigenvalue of $\mathbf{A}$ is positive (Theorem 13.1.5 (b)), so $0 \notin \operatorname{spec}(\mathbf{A})$ and $\mathbf{A}$ is invertible (Theorem 9.2.6). Also $(\mathbf{A}^{-1})^* = (\mathbf{A}^*)^{-1} = \mathbf{A}^{-1}$, so

$$(\mathbf{M}/\mathbf{A})^* = \mathbf{C}^* - \mathbf{B}^*(\mathbf{A}^{-1})^*\mathbf{B} = \mathbf{C} - \mathbf{B}^*\mathbf{A}^{-1}\mathbf{B} = \mathbf{M}/\mathbf{A},$$

and $\mathbf{M}/\mathbf{A}$ is Hermitian. Put

$$\mathbf{S} = \begin{pmatrix} \mathbf{I}_k & -\mathbf{A}^{-1}\mathbf{B} \\ \mathbf{O} & \mathbf{I}_l \end{pmatrix}, \qquad \mathbf{S}^* = \begin{pmatrix} \mathbf{I}_k & \mathbf{O} \\ -\mathbf{B}^*\mathbf{A}^{-1} & \mathbf{I}_l \end{pmatrix},$$

the second because $\mathbf{A}^{-1}$ is Hermitian. By Theorem 8.1.5, $\mathbf{S}$ is invertible. Two applications of Theorem 8.1.1 give

$$\begin{aligned} \mathbf{M}\mathbf{S} &= \begin{pmatrix} \mathbf{A} & \mathbf{B} - \mathbf{A}\mathbf{A}^{-1}\mathbf{B} \\ \mathbf{B}^* & \mathbf{C} - \mathbf{B}^*\mathbf{A}^{-1}\mathbf{B} \end{pmatrix} = \begin{pmatrix} \mathbf{A} & \mathbf{O} \\ \mathbf{B}^* & \mathbf{M}/\mathbf{A} \end{pmatrix}, \\ \mathbf{S}^*(\mathbf{M}\mathbf{S}) &= \begin{pmatrix} \mathbf{A} & \mathbf{O} \\ \mathbf{B}^* - \mathbf{B}^*\mathbf{A}^{-1}\mathbf{A} & \mathbf{M}/\mathbf{A} \end{pmatrix} = \begin{pmatrix} \mathbf{A} & \mathbf{O} \\ \mathbf{O} & \mathbf{M}/\mathbf{A} \end{pmatrix}. \end{aligned} \tag{$*$}$$

Write $\mathbf{N} = \mathbf{A} \oplus (\mathbf{M}/\mathbf{A})$ for that block diagonal matrix. For $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2) \in F^k \times F^l$,

$$\mathbf{y}^*\mathbf{N}\mathbf{y} = \mathbf{y}_1^*\mathbf{A}\mathbf{y}_1 + \mathbf{y}_2^*(\mathbf{M}/\mathbf{A})\mathbf{y}_2. \tag{$\dagger$}$$

Claim. Given $\mathbf{A} \succ \mathbf{O}$: $\mathbf{N} \succ \mathbf{O}$ if and only if $\mathbf{M}/\mathbf{A} \succ \mathbf{O}$, and $\mathbf{N} \succeq \mathbf{O}$ if and only if $\mathbf{M}/\mathbf{A} \succeq \mathbf{O}$.

Proof. Taking $\mathbf{y}_1 = \mathbf{0}$ in $(\dagger)$ shows that positivity of $\mathbf{N}$, weak or strict, passes to $\mathbf{M}/\mathbf{A}$, since $\mathbf{y}_2 \neq \mathbf{0}$ forces $\mathbf{y} \neq \mathbf{0}$. Conversely, suppose $\mathbf{M}/\mathbf{A} \succeq \mathbf{O}$; then both terms of $(\dagger)$ are ≥ 0, so $\mathbf{N} \succeq \mathbf{O}$. If moreover $\mathbf{M}/\mathbf{A} \succ \mathbf{O}$, then for $\mathbf{y} \neq \mathbf{0}$ at least one of $\mathbf{y}_1$, $\mathbf{y}_2$ is nonzero, so at least one term of $(\dagger)$ is > 0 while the other is ≥ 0; hence $\mathbf{N} \succ \mathbf{O}$.

■

Finally $\mathbf{S}$ is invertible, so Proposition 13.1.2 (b) gives $\mathbf{M} \succ \mathbf{O}$ if and only if $\mathbf{S}^{*}\mathbf{M}\mathbf{S} = \mathbf{N} \succ \mathbf{O}$. For the weak version, Proposition 13.1.2 (a) applied to $\mathbf{S}$ turns $\mathbf{M} \succeq \mathbf{O}$ into $\mathbf{N} \succeq \mathbf{O}$, and applied to $\mathbf{S}^{-1}$, together with $(\mathbf{S}^{-1})^{*}\mathbf{N}\mathbf{S}^{-1} = \mathbf{M}$, turns $\mathbf{N} \succeq \mathbf{O}$ back into $\mathbf{M} \succeq \mathbf{O}$. Combining with the claim proves (b).

■

The identity $(*)$ is Theorem 8.3.1 (a) in disguise. That theorem factors $\mathbf{M} = \text{Ł}\,\mathbf{N}\,\mathbf{U}$ with Ł lower and $\mathbf{U}$ upper block triangular; here Ł and $\mathbf{U}$ are $(\mathbf{S}^{-1})^{*}$ and $\mathbf{S}^{-1}$, so the two outer factors are adjoint to each other and the factorization is a **congruence**, not merely an equivalence. Taking determinants in $(*)$ recovers Theorem 8.3.2,

$$\det \mathbf{M} = \det \mathbf{A} \cdot \det(\mathbf{M}/\mathbf{A}),$$

which is the form the next section uses.

Chapter 8 closed its treatment of Schur complements by observing that when $\mathbf{M}$ is real symmetric the complement $\mathbf{M}/\mathbf{A}$ is symmetric again, and promising that "Chapter 13 uses Schur complements to test whether $\mathbf{M}$ is positive definite one block at a time". Theorem 13.6.1 is that test, and part (b) says exactly what the promise says: once the corner block $\mathbf{A}$ is known to be positive definite, the whole of $\mathbf{M}$ is positive definite precisely when the smaller matrix $\mathbf{M}/\mathbf{A}$ is. A question about $k + l$ coordinates has become a question about l of them.

It also finishes what Section 1 started. There, in the induction proving Sylvester's criterion, the matrix was split with $k = n - 1$ and $l = 1$; the shear $\mathbf{S}$ was the matrix above with the column $\mathbf{b}$ in place of $\mathbf{B}$; and the congruence came out as $\mathbf{S}^{*}\mathbf{A}\mathbf{S} = \mathbf{A}' \oplus (s)$ with the single number $s = d - \mathbf{b}^{*}\mathbf{A}'^{-1}\mathbf{b}$ in the corner. That section promised that Section 6 would replace the scalar by a block. The scalar has become $\mathbf{M}/\mathbf{A}$, and the argument is word for word the same one.

Example. A 4×4 Tested in 2×2 Blocks

Decide whether each of

$$\mathbf{M} = \left(\begin{array}{cc|cc} 2 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \\ \hline 1 & 1 & 3 & 2 \\ 0 & 1 & 2 & 4 \end{array}\right), \qquad \mathbf{M}' = \left(\begin{array}{cc|cc} 2 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \\ \hline 1 & 1 & 2 & 0 \\ 0 & 1 & 0 & 2 \end{array}\right)$$

is positive definite, using the displayed partition. For the one that is not, exhibit a vector on which the quadratic form is negative.

Solution. Both matrices are real symmetric with the same corner and border,

$$\mathbf{A} = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}, \qquad \mathbf{B} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}.$$

Step 1: the corner. The leading principal minors of $\mathbf{A}$ are 2 and $2 - 1 = 1$, both positive, so $\mathbf{A} \succ \mathbf{O}$ by Sylvester's criterion (Theorem 13.1.5 (d)) and Theorem 13.6.1 (b) applies. Since

$\det \mathbf{A} = 1$, Theorem 0.9.8 gives $\mathbf{A}^{-1} = \left(\begin{smallmatrix} 1 & -1 \\ -1 & 2 \end{smallmatrix}\right)$.

Step 2: the coupling term. It is the same for both matrices:

$$\mathbf{B}^\top \mathbf{A}^{-1} = \begin{pmatrix} 0 & 1 \\ -1 & 2 \end{pmatrix}, \qquad \mathbf{B}^\top \mathbf{A}^{-1} \mathbf{B} = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}.$$

Step 3: the complements. For $\mathbf{M}$, with $\mathbf{C} = \left(\begin{smallmatrix} 3 & 2 \\ 2 & 4 \end{smallmatrix}\right)$,

$$\mathbf{M}/\mathbf{A} = \begin{pmatrix} 3 & 2 \\ 2 & 4 \end{pmatrix} - \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix},$$

whose leading principal minors are 2 and 3. So $\mathbf{M}/\mathbf{A} \succ \mathbf{O}$, hence $\mathbf{M} \succ \mathbf{O}$. As a by-product, $\det \mathbf{M} = \det \mathbf{A} \cdot \det(\mathbf{M}/\mathbf{A}) = 1 \cdot 3 = 3$.

For $\mathbf{M}'$, with $\mathbf{C}' = \left(\begin{smallmatrix} 2 & 0 \\ 0 & 2 \end{smallmatrix}\right)$,

$$\mathbf{M}'/\mathbf{A} = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} - \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ -1 & 0 \end{pmatrix},$$

whose determinant is -1. So $\mathbf{M}'/\mathbf{A}$ is not even positive semidefinite, and $\mathbf{M}'$ is not positive definite.

Step 4: the witness. The congruence $(*)$ turns a bad vector for $\mathbf{M}'/\mathbf{A}$ into a bad vector for $\mathbf{M}'$. Here $\mathbf{y}_2 = (1, 1)$ gives $\mathbf{y}_2^\top (\mathbf{M}'/\mathbf{A}) \mathbf{y}_2 = 1 - 1 - 1 + 0 = -1$, so $\mathbf{y} = (0, 0, 1, 1)$ satisfies $\mathbf{y}^\top (\mathbf{S}^\top \mathbf{M}' \mathbf{S}) \mathbf{y} = -1$, and $\mathbf{x} = \mathbf{S}\mathbf{y}$ satisfies $\mathbf{x}^\top \mathbf{M}' \mathbf{x} = -1$. Since $-\mathbf{A}^{-1}\mathbf{B} = \left(\begin{smallmatrix} 0 & 1 \\ -1 & -2 \end{smallmatrix}\right)$, we get $\mathbf{x} = (1, -3, 1, 1)$. Checking directly, $\mathbf{M}'\mathbf{x} = (0, 0, 0, -1)$ and $\mathbf{x}^\top (\mathbf{M}'\mathbf{x}) = 0 + 0 + 0 - 1 = -1$.

One 2×2 inverse and one 2×2 product replaced four leading principal minors of a 4×4 matrix. And when the answer is "no", the congruence hands back an explicit vector, which a determinant test never does.

Quick check. Let $\mathbf{A} \succ \mathbf{O}$. What is $\mathbf{M}/\mathbf{A}$ when $\mathbf{B} = \mathbf{O}$, and what does Theorem 13.6.1 (b) say in that case?

Solution. When $\mathbf{B} = \mathbf{O}$ the Schur complement is $\mathbf{M}/\mathbf{A} = \mathbf{C} - \mathbf{O} = \mathbf{C}$, so the criterion reads: $\mathbf{M} \succ \mathbf{O}$ if and only if $\mathbf{C} \succ \mathbf{O}$. That is the right answer, since $\mathbf{M} = \mathbf{A} \oplus \mathbf{C}$ is then block diagonal and (†) says its quadratic form is the sum of the two separate quadratic forms. The general case is this case after the change of variable $\mathbf{S}$.

Warning. **The corner must be definite, not merely semidefinite.** If $\mathbf{A}$ is singular, then $\mathbf{A}^{-1}$ does not exist, $\mathbf{M}/\mathbf{A}$ is undefined, and Theorem 13.6.1 (b) says nothing whatsoever. This is a real obstruction, not a gap in the proof. The matrix $\left(\begin{smallmatrix} 0 & 1 \\ 1 & 1 \end{smallmatrix}\right)$ above has $\mathbf{A} \succeq \mathbf{O}$ and $\mathbf{C} \succ \mathbf{O}$ and is **not** positive semidefinite, while $\left(\begin{smallmatrix} 0 & 0 \\ 0 & 1 \end{smallmatrix}\right)$, with the same $\mathbf{A}$ and the same $\mathbf{C}$, **is**. The two differ only in $\mathbf{B}$, so no criterion that ignores $\mathbf{B}$ can separate them.

Remark. There is a criterion for the semidefinite case with a singular corner, and it has three clauses instead of one: for Hermitian $\mathbf{M}$, one has $\mathbf{M} \succeq \mathbf{O}$ exactly when $\mathbf{A} \succeq \mathbf{O}$, every column of $\mathbf{B}$ lies in the column space of $\mathbf{A}$, and $\mathbf{C} - \mathbf{B}^*\mathbf{A}^+\mathbf{B} \succeq \mathbf{O}$. The middle clause is what the warning's first matrix violates. The matrix $\mathbf{A}^+$ in the third clause is the Moore–Penrose pseudoinverse; it agrees with $\mathbf{A}^{-1}$ when $\mathbf{A}$ is invertible, and it is built from the singular values of $\mathbf{A}$ in Section 12 of this chapter. The statement and its proof belong there, with the tools they need. This section keeps the definite case, which is the one the rest of the chapter consumes.

13.6.3 A positive definite certificate for stability

Chapter 12 ended its section on the Sylvester equation with a specialization it could not finish. For $\mathbf{A} \in M_n(\mathbb{C})$ and a Hermitian $\mathbf{Q} \in M_n(\mathbb{C})$, the **Lyapunov equation** is $\mathbf{AX} + \mathbf{XA}^* = -\mathbf{Q}$, and $\mathbf{A}$ is called **stable** when every $\lambda \in \operatorname{spec}(\mathbf{A})$ has $\operatorname{Re}\lambda < 0$ (Definition 12.10.2). That section showed that a stable $\mathbf{A}$ makes the equation uniquely solvable whatever $\mathbf{Q}$ is, because the spectra of $\mathbf{A}$ and of $-\mathbf{A}^*$ are then disjoint and Theorem 12.10.2 applies. It then stated, without proof, that "if $\mathbf{A}$ is stable and $\mathbf{Q}$ is positive definite, the unique solution $\mathbf{X}$ of $\mathbf{AX} + \mathbf{XA}^* = -\mathbf{Q}$ is positive definite; and conversely, if some positive definite $\mathbf{X}$ solves the equation for some positive definite $\mathbf{Q}$, then $\mathbf{A}$ is stable", and added that "positivity is the subject of Chapter 13, where the tools to handle $\mathbf{x}^*\mathbf{Xx}$ are built".

They are built, and the one the proof below needs is the block test just proved. The difficulty is that $\mathbf{X}$ is given by an equation rather than by a formula: there is no factorization to look at and no spectrum to compute. What one can do is take a single coordinate off at a time and let a Schur complement carry the verdict — which is exactly Theorem 13.6.1 (b) with $l = 1$.

Everything rests on evaluating the equation at one well-chosen vector. Eigenvectors of $\mathbf{A}$ are the natural candidates, but $\mathbf{X}$ sits between $\mathbf{A}$ and $\mathbf{A}^*$ and blocks the usual move; the repair is to feed the equation the vector $\mathbf{X}^{-1}\mathbf{v}$ rather than $\mathbf{v}$.

Lemma 13.6.2. Reading the Lyapunov Equation at an Eigenvector

Let $\mathbf{A}, \mathbf{Q}, \mathbf{X} \in M_n(\mathbb{C})$ satisfy $\mathbf{AX} + \mathbf{XA}^* = -\mathbf{Q}$, with $\mathbf{X}$ Hermitian and invertible, and let $\mathbf{Av} = \lambda\mathbf{v}$ with $\mathbf{v} \neq \mathbf{0}$. Put $\mathbf{u} = \mathbf{X}^{-1}\mathbf{v}$. Then $\mathbf{u} \neq \mathbf{0}$, the number $\mathbf{v}^*\mathbf{X}^{-1}\mathbf{v}$ is real, and

$$2(\operatorname{Re}\lambda)\,\mathbf{v}^*\mathbf{X}^{-1}\mathbf{v} = -\,\mathbf{u}^*\mathbf{Qu}. \tag{13.6.1}$$

Proof. Since $\mathbf{Xu} = \mathbf{v} \neq \mathbf{0}$, we have $\mathbf{u} \neq \mathbf{0}$. The inverse of a Hermitian invertible matrix is Hermitian, because $(\mathbf{X}^{-1})^* = (\mathbf{X}^*)^{-1} = \mathbf{X}^{-1}$, so $\mathbf{v}^*\mathbf{X}^{-1}\mathbf{v} = \langle \mathbf{X}^{-1}\mathbf{v}, \mathbf{v}\rangle$ is a real number (Proposition 11.6.7 (a)). Write $s = \mathbf{v}^*\mathbf{X}^{-1}\mathbf{v} = \mathbf{v}^*\mathbf{u}$; then $\mathbf{u}^*\mathbf{v} = \overline{\mathbf{v}^*\mathbf{u}} = s$ as well, the last step because s is real.

Evaluate the quadratic form of each side of the equation at $\mathbf{u}$. Using $\mathbf{Xu} = \mathbf{v}$ and $\mathbf{u}^*\mathbf{X} = (\mathbf{Xu})^* = \mathbf{v}^*$, which is where the Hermitian hypothesis is spent,

$$\mathbf{u}^*\mathbf{AXu} = \mathbf{u}^*\mathbf{Av} = \lambda\,\mathbf{u}^*\mathbf{v} = \lambda s, \qquad \mathbf{u}^*\mathbf{XA}^*\mathbf{u} = \mathbf{v}^*\mathbf{A}^*\mathbf{u} = (\mathbf{Av})^*\mathbf{u} = \overline{\lambda}\,\mathbf{v}^*\mathbf{u} = \overline{\lambda}s.$$

Adding the two gives $\mathbf{u}^*(\mathbf{AX} + \mathbf{XA}^*)\mathbf{u} = (\lambda + \overline{\lambda})s = 2(\operatorname{Re}\lambda)s$, while the right-hand side of the equation gives $\mathbf{u}^*(-\mathbf{Q})\mathbf{u} = -\mathbf{u}^*\mathbf{Qu}$. This proves the identity.

■

The identity has three ingredients, and it fixes the sign of any one of them once the signs of the other two are known. Both halves of the theorem are that observation.

Theorem 13.6.3. Lyapunov's Theorem

Let $\mathbf{A} \in M_n(\mathbb{C})$ and let $\mathbf{Q} \in M_n(\mathbb{C})$ with $\mathbf{Q} \succ \mathbf{O}$.

(a) If $\mathbf{A}$ is stable, then $\mathbf{AX} + \mathbf{XA}^* = -\mathbf{Q}$ has exactly one solution $\mathbf{X} \in M_n(\mathbb{C})$, and that solution is positive definite.

(b) Conversely, if some $\mathbf{X} \succ \mathbf{O}$ satisfies $\mathbf{AX} + \mathbf{XA}^* = -\mathbf{Q}$, then $\mathbf{A}$ is stable.

Idea. Both halves are Eq. 13.6.1 read in different directions: it ties the sign of $\operatorname{Re}\lambda$ to the sign of $\mathbf{v}^*\mathbf{X}^{-1}\mathbf{v}$, the third factor $-\mathbf{u}^*\mathbf{Qu}$ being negative throughout. Knowing $\mathbf{X} \succ \mathbf{O}$ gives (b) in one line. For (a) the same identity controls the quadratic form of $\mathbf{X}^{-1}$ only along eigenvectors of $\mathbf{A}$, and the work is to get from there to every vector. Peel one eigenvector off at a time: choose an orthonormal basis whose **last** vector is a unit eigenvector of $\mathbf{A}$, so that $\mathbf{A}$ becomes block lower triangular with a 1×1 corner. The leading $(n-1) \times (n-1)$ block of the transformed equation is a Lyapunov equation again, for a stable matrix, so induction makes the leading block of $\mathbf{X}$ positive definite. One Schur complement is then all that stands between us and the verdict — and by Theorem 8.3.4 that number is the reciprocal of the last diagonal entry of $\mathbf{X}^{-1}$, which the identity has just shown to be positive.

Proof. **Existence and uniqueness.** Suppose $\mathbf{A}$ is stable. The eigenvalues of $-\mathbf{A}^*$ are the numbers $-\overline{\mu}$ with $\mu \in \operatorname{spec}(\mathbf{A})$, as Chapter 12 computes from $\det(\mathbf{M}^*) = \overline{\det \mathbf{M}}$ immediately after Definition 12.10.2. If some $\lambda \in \operatorname{spec}(\mathbf{A})$ were also an eigenvalue of $-\mathbf{A}^*$, then $\lambda = -\overline{\mu}$ for some $\mu \in \operatorname{spec}(\mathbf{A})$, so $\lambda + \overline{\mu} = 0$; but $\operatorname{Re}(\lambda + \overline{\mu}) = \operatorname{Re}\lambda + \operatorname{Re}\mu < 0$. So $\operatorname{spec}(\mathbf{A}) \cap \operatorname{spec}(-\mathbf{A}^*) = \varnothing$, and Theorem 12.10.2, (a) $\Rightarrow$ (c), applied to the pair $(\mathbf{A}, -\mathbf{A}^*)$, says that $\mathbf{AX} - \mathbf{X}(-\mathbf{A}^*) = -\mathbf{Q}$ has exactly one solution. That is the Lyapunov equation.

(b). Let $\mathbf{X} \succ \mathbf{O}$ satisfy the equation. Every eigenvalue of $\mathbf{X}$ is positive (Theorem 13.1.5 (b)), so $\mathbf{X}$ is invertible (Theorem 9.2.6), and $\mathbf{X}^{-1} \succ \mathbf{O}$: writing $\mathbf{X} = \mathbf{B}^*\mathbf{B}$ with $\mathbf{B}$ invertible (Theorem 13.1.5 (c)) and putting $\mathbf{C} = (\mathbf{B}^*)^{-1}$, which is invertible with $\mathbf{C}^* = \mathbf{B}^{-1}$, we get $\mathbf{C}^*\mathbf{C} = \mathbf{B}^{-1}(\mathbf{B}^*)^{-1} = (\mathbf{B}^*\mathbf{B})^{-1} = \mathbf{X}^{-1}$, and Theorem 13.1.5 (c) $\Rightarrow$ (a) applies to $\mathbf{X}^{-1}$.

Let $\lambda \in \operatorname{spec}(\mathbf{A})$ and let $\mathbf{v} \neq \mathbf{0}$ be an eigenvector for it. With $\mathbf{u} = \mathbf{X}^{-1}\mathbf{v}$, Lemma 13.6.2 gives $\mathbf{u} \neq \mathbf{0}$ and

$$2(\operatorname{Re}\lambda)\,\mathbf{v}^*\mathbf{X}^{-1}\mathbf{v} = -\,\mathbf{u}^*\mathbf{Qu} < 0,$$

the inequality because $\mathbf{Q} \succ \mathbf{O}$ and $\mathbf{u} \neq \mathbf{0}$. Also $\mathbf{v}^*\mathbf{X}^{-1}\mathbf{v} > 0$, since $\mathbf{X}^{-1} \succ \mathbf{O}$ and $\mathbf{v} \neq \mathbf{0}$. Dividing, $\operatorname{Re}\lambda < 0$. As λ was an arbitrary eigenvalue, $\mathbf{A}$ is stable.

(a). Let $\mathbf{A}$ be stable, $\mathbf{Q} \succ \mathbf{O}$, and let $\mathbf{X}$ be the unique solution.

Step 1: $\mathbf{X}$ is Hermitian. Starring the equation gives $(\mathbf{AX} + \mathbf{XA}^*)^* = \mathbf{AX}^* + \mathbf{X}^*\mathbf{A}^*$, while $(-\mathbf{Q})^* = -\mathbf{Q}$ because $\mathbf{Q}$ is Hermitian. So $\mathbf{X}^*$ solves the same equation, and uniqueness forces $\mathbf{X}^* = \mathbf{X}$.

Step 2: $\mathbf{X}$ is invertible. Let $\mathbf{v} \in \mathbb{C}^n$ with $\mathbf{Xv} = \mathbf{0}$. Then $\mathbf{v}^*\mathbf{X} = (\mathbf{Xv})^* = \mathbf{0}^*$ by Step 1, so

both terms of

$$\mathbf{v}^*(\mathbf{AX}+\mathbf{XA}^*)\mathbf{v}=\mathbf{v}^*\mathbf{A}(\mathbf{Xv})+(\mathbf{v}^*\mathbf{X})\mathbf{A}^*\mathbf{v}$$

vanish. The left-hand side is $-\mathbf{v}^*\mathbf{Qv}$, so $\mathbf{v}^*\mathbf{Qv}=0$, and $\mathbf{Q}\succ\mathbf{O}$ forces $\mathbf{v}=\mathbf{0}$. So $\operatorname{nul}(\mathbf{X})=\{\mathbf{0}\}$ and $\mathbf{X}$ is invertible (Theorem 3.4.4).

Step 3: induction on n. We prove by induction on $n\geq 1$ that for every stable $\mathbf{A}\in M_n(\mathbb{C})$ and every $\mathbf{Q}\succ\mathbf{O}$ the unique solution is positive definite.

For $n=1$, write $\mathbf{A}=(\alpha)$, $\mathbf{Q}=(q)$, and $\mathbf{X}=(x)$, with $\operatorname{Re}\alpha<0$, with $q>0$, and with x real by Step 1. The equation reads $2(\operatorname{Re}\alpha)x=-q$, so $x=q/(-2\operatorname{Re}\alpha)>0$ and $\mathbf{X}\succ\mathbf{O}$.

Let $n\geq 2$ and assume the statement for $n-1$. By Theorem 9.2.4 there are $\tau\in\operatorname{spec}(\mathbf{A})$ and $\mathbf{w}\in\mathbb{C}^n$ with $\mathbf{Aw}=\tau\mathbf{w}$ and $\|\mathbf{w}\|=1$, after dividing an eigenvector by its norm; and $\operatorname{Re}\tau<0$. Extend $\mathbf{w}$ to an orthonormal basis $(\mathbf{w},\mathbf{w}_2,\dots,\mathbf{w}_n)$ of $\mathbb{C}^n$ (Corollary 11.2.3) and let $\mathbf{W}$ be the matrix with columns $\mathbf{w}_2,\dots,\mathbf{w}_n,\mathbf{w}$ in that order. Reordering an orthonormal list leaves it orthonormal, so the columns of $\mathbf{W}$ are an orthonormal basis and $\mathbf{W}$ is unitary (Theorem 11.7.3 (f), Definition 11.7.2). Its last column is $\mathbf{w}$, so $\mathbf{We}_n=\mathbf{w}$ and $\mathbf{W}^*\mathbf{w}=\mathbf{W}^{-1}\mathbf{w}=\mathbf{e}_n$.

Put $\hat{\mathbf{A}}=\mathbf{W}^*\mathbf{AW}$, $\hat{\mathbf{X}}=\mathbf{W}^*\mathbf{XW}$, and $\hat{\mathbf{Q}}=\mathbf{W}^*\mathbf{QW}$. Inserting $\mathbf{WW}^*=\mathbf{I}$ between the factors of each product,

$$\hat{\mathbf{A}}\hat{\mathbf{X}}+\hat{\mathbf{X}}\hat{\mathbf{A}}^*=\mathbf{W}^*(\mathbf{AX}+\mathbf{XA}^*)\mathbf{W}=-\hat{\mathbf{Q}},$$

and $\hat{\mathbf{Q}}\succ\mathbf{O}$ by Proposition 13.1.2 (b), while $\hat{\mathbf{X}}$ is Hermitian and invertible because $\mathbf{X}$ is and $\mathbf{W}$ is unitary. The last column of $\hat{\mathbf{A}}$ is $\mathbf{W}^*\mathbf{AWe}_n=\mathbf{W}^*\mathbf{Aw}=\tau\mathbf{W}^*\mathbf{w}=\tau\mathbf{e}_n$, so in blocks of sizes $n-1$ and 1,

$$\hat{\mathbf{A}}=\begin{pmatrix}\mathbf{A}' & \mathbf{0}\\ \mathbf{g}^* & \tau\end{pmatrix},\qquad \hat{\mathbf{X}}=\begin{pmatrix}\mathbf{X}' & \mathbf{b}\\ \mathbf{b}^* & d\end{pmatrix},\qquad \hat{\mathbf{Q}}=\begin{pmatrix}\mathbf{Q}' & *\\ * & *\end{pmatrix}.$$

Claim. $\mathbf{X}'\succ\mathbf{O}$.

Proof. By Theorem 8.1.1 the leading $(n-1)\times(n-1)$ block of $\hat{\mathbf{A}}\hat{\mathbf{X}}$ is $\mathbf{A}'\mathbf{X}'+\mathbf{0}\mathbf{b}^*=\mathbf{A}'\mathbf{X}'$. Since $\hat{\mathbf{X}}$ is Hermitian, $\hat{\mathbf{X}}\hat{\mathbf{A}}^*=(\hat{\mathbf{A}}\hat{\mathbf{X}})^*$, whose leading block is $(\mathbf{A}'\mathbf{X}')^*=\mathbf{X}'\mathbf{A}'^*$. Reading off the leading block of the transformed equation,

$$\mathbf{A}'\mathbf{X}'+\mathbf{X}'\mathbf{A}'^*=-\mathbf{Q}'.$$

Here $\mathbf{Q}'\succ\mathbf{O}$ by Theorem 13.6.1 (a), applied to $\hat{\mathbf{Q}}\succ\mathbf{O}$ with this partition.

And $\mathbf{A}'$ is stable. Let $\lambda\in\operatorname{spec}(\mathbf{A}')$, so that $\det(\lambda\mathbf{I}_{n-1}-\mathbf{A}')=0$ by Theorem 9.2.1 (d). The matrix

$$\lambda\mathbf{I}_n-\hat{\mathbf{A}}=\begin{pmatrix}\lambda\mathbf{I}_{n-1}-\mathbf{A}' & \mathbf{0}\\ -\mathbf{g}^* & \lambda-\tau\end{pmatrix}$$

is block triangular with a zero block in the upper right, so Theorem 7.8.4 gives $\det(\lambda\mathbf{I}_n-\hat{\mathbf{A}})=\det(\lambda\mathbf{I}_{n-1}-\mathbf{A}')\cdot(\lambda-\tau)=0$, and $\lambda\in\operatorname{spec}(\hat{\mathbf{A}})$ by Theorem 9.2.1 again. Finally $\hat{\mathbf{A}}=\mathbf{W}^{-1}\mathbf{AW}$, so $p_{\hat{\mathbf{A}}}=p_{\mathbf{A}}$ by Theorem 7.10.3, and two matrices with the same characteristic polynomial have the same eigenvalues (Theorem 9.2.1 (e)). Hence $\lambda\in\operatorname{spec}(\mathbf{A})$ and $\operatorname{Re}\lambda<0$.

So $\mathbf{A}'\in M_{n-1}(\mathbb{C})$ is stable and $\mathbf{Q}'\succ\mathbf{O}$. By the existence and uniqueness step, applied to the pair $(\mathbf{A}',\mathbf{Q}')$, the equation $\mathbf{A}'\mathbf{Y}+\mathbf{Y}\mathbf{A}'^*=-\mathbf{Q}'$ has exactly one solution; $\mathbf{X}'$ is a solution, hence is that one. The inductive hypothesis gives $\mathbf{X}'\succ\mathbf{O}$.

■

Since $\mathbf{X}' \succ \mathbf{O}$ it is invertible, so the Schur complement $\hat{\mathbf{X}}/\mathbf{X}' = d - \mathbf{b}^*\mathbf{X}'^{-1}\mathbf{b}$ is defined, and it is a 1×1 Hermitian matrix by Theorem 13.6.1 (b), that is a real number s. The matrix $\hat{\mathbf{X}}$ is invertible, so Theorem 8.3.4 (a) says that $s \neq 0$ and that the lower-right entry of $\hat{\mathbf{X}}^{-1}$ is its inverse:

$$\mathbf{e}_n^*\hat{\mathbf{X}}^{-1}\mathbf{e}_n = \frac{1}{s}.$$

On the other hand $\hat{\mathbf{A}}\mathbf{e}_n = \tau\mathbf{e}_n$ with $\mathbf{e}_n \neq \mathbf{0}$ and $\operatorname{Re}\tau < 0$, so Lemma 13.6.2, applied to the triple $(\hat{\mathbf{A}}, \hat{\mathbf{Q}}, \hat{\mathbf{X}})$ and this eigenvector, gives

$$2(\operatorname{Re}\tau)\,\mathbf{e}_n^*\hat{\mathbf{X}}^{-1}\mathbf{e}_n = -\,\mathbf{u}^*\hat{\mathbf{Q}}\mathbf{u} < 0, \qquad \mathbf{u} = \hat{\mathbf{X}}^{-1}\mathbf{e}_n \neq \mathbf{0},$$

the inequality because $\hat{\mathbf{Q}} \succ \mathbf{O}$. Therefore $\mathbf{e}_n^*\hat{\mathbf{X}}^{-1}\mathbf{e}_n > 0$, that is $1/s > 0$, that is $s > 0$ and $\hat{\mathbf{X}}/\mathbf{X}' = (s) \succ \mathbf{O}$. Now Theorem 13.6.1 (b), with the corner block $\mathbf{X}' \succ \mathbf{O}$, gives $\hat{\mathbf{X}} \succ \mathbf{O}$. Finally $\mathbf{X} = \mathbf{W}\hat{\mathbf{X}}\mathbf{W}^* = (\mathbf{W}^*)^*\hat{\mathbf{X}}\mathbf{W}^*$ is a congruence of $\hat{\mathbf{X}}$ by the invertible matrix $\mathbf{W}^*$, so $\mathbf{X} \succ \mathbf{O}$ by Proposition 13.1.2 (b). This completes the induction and proves (a).

■

Part (b) is what makes the theorem useful, and part (a) is what makes part (b) usable. Reading (b) alone: to prove that every eigenvalue of $\mathbf{A}$ has negative real part, it is enough to exhibit **one** positive definite $\mathbf{X}$ with $\mathbf{AX}+\mathbf{XA}^*$ negative definite, and no eigenvalue need ever be computed. Such a certificate is a linear object — Chapter 12 §10 turned the equation into an ordinary $n^2 \times n^2$ linear system — whereas eigenvalues are roots of a polynomial. Reading (a) with it: the certificate always exists when the claim is true, so the method never fails for lack of a witness. Pick any $\mathbf{Q} \succ \mathbf{O}$, solve, and test the answer with Sylvester's criterion (Theorem 13.1.5 (d)).

Example. A Certificate for a Shear

Take $\mathbf{Q} = \mathbf{I}_2$ and solve $\mathbf{AX} + \mathbf{XA}^\top = -\mathbf{I}_2$ for

$$\mathbf{A} = \begin{pmatrix} -1 & 2 \\ 0 & -1 \end{pmatrix}, \qquad \text{and then for} \qquad \mathbf{A}_1 = \begin{pmatrix} 1 & 2 \\ 0 & 2 \end{pmatrix}.$$

In each case decide from the solution whether the matrix is stable.

Solution. Both matrices are real, so we look for a real symmetric $\mathbf{X} = \left(\begin{smallmatrix} x & y \\ y & z \end{smallmatrix}\right)$; the remark below says nothing is lost by doing so. For a real $\mathbf{A}$ we have $\mathbf{A}^* = \mathbf{A}^\top$, and $\mathbf{XA}^\top = (\mathbf{AX})^\top$ because $\mathbf{X}$ is symmetric, so the equation says that $\mathbf{AX} + (\mathbf{AX})^\top = -\mathbf{I}_2$.

For $\mathbf{A}$, a direct computation gives $\mathbf{AX} = \left(\begin{smallmatrix} -x+2y & -y+2z \\ -y & -z \end{smallmatrix}\right)$, so

$$\mathbf{AX} + (\mathbf{AX})^\top = \begin{pmatrix} -2x+4y & -2y+2z \\ -2y+2z & -2z \end{pmatrix} = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}.$$

The corner entry gives $z = \frac{1}{2}$, the off-diagonal entry gives $y = z = \frac{1}{2}$, and the first entry gives $-2x + 2 = -1$, so $x = \frac{3}{2}$. Hence

$$\mathbf{X} = \frac{1}{2}\begin{pmatrix} 3 & 1 \\ 1 & 1 \end{pmatrix},$$

and its leading principal minors are $\frac{3}{2}$ and $\frac{3}{4} - \frac{1}{4} = \frac{1}{2}$, both positive, so $\mathbf{X} \succ \mathbf{O}$ by Sylvester's criterion. By Theorem 13.6.3 (b), $\mathbf{A}$ is stable — as it is, its only eigenvalue being -1 twice.

For $\mathbf{A}_1$, the same computation with $\mathbf{A}_1\mathbf{X} = \left(\begin{smallmatrix} x+2y & y+2z \\ 2y & 2z \end{smallmatrix}\right)$ gives

$$\mathbf{A}_1\mathbf{X} + (\mathbf{A}_1\mathbf{X})^\top = \begin{pmatrix} 2x+4y & 3y+2z \\ 3y+2z & 4z \end{pmatrix} = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix},$$

so $z = -\frac{1}{4}$, then $y = -\frac{2}{3}z = \frac{1}{6}$, then $2x = -1 - \frac{2}{3}$ and $x = -\frac{5}{6}$. The equation is still uniquely solvable, since the eigenvalues 1 and 2 of $\mathbf{A}_1$ admit no relation $\lambda + \overline{\mu} = 0$, but the solution

$$\mathbf{X}_1 = \begin{pmatrix} -5/6 & 1/6 \\ 1/6 & -1/4 \end{pmatrix}$$

has $\mathbf{e}_1^\top \mathbf{X}_1 \mathbf{e}_1 = -5/6 < 0$ and is not positive definite. So Theorem 13.6.3 (a) is not contradicted, and indeed $\mathbf{A}_1$ is not stable, its eigenvalues being 1 and 2. In fact $-\mathbf{X}_1 \succ \mathbf{O}$, its leading principal minors being 5/6 and 13/72: replacing $\mathbf{A}_1$ by $-\mathbf{A}_1$, which is stable, replaces $\mathbf{X}_1$ by $-\mathbf{X}_1$, and Theorem 13.6.3 (a) applies to that pair.

Remark. Over $\mathbb{R}$ nothing changes. If $\mathbf{A}$ and $\mathbf{Q}$ are real with $\mathbf{Q} \succ \mathbf{O}$, then conjugating the equation entrywise turns the solution $\mathbf{X}$ into another solution $\overline{\mathbf{X}}$, because $\overline{\mathbf{A}} = \mathbf{A}$ and $\overline{\mathbf{A}^*} = \mathbf{A}^\top = \mathbf{A}^*$ and $\overline{\mathbf{Q}} = \mathbf{Q}$; by uniqueness $\overline{\mathbf{X}} = \mathbf{X}$, so $\mathbf{X}$ is a real symmetric matrix, positive definite in the sense of Definition 13.1.1 over $\mathbb{R}$ as well, since its quadratic form on $\mathbb{R}^n$ is the restriction of its quadratic form on $\mathbb{C}^n$.

Quick check. Why does Theorem 13.6.3 say nothing about the quarter turn $\mathbf{A} = \left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)$? Show that for this $\mathbf{A}$ the equation $\mathbf{AX} + \mathbf{XA}^* = -\mathbf{I}_2$ has no solution whatsoever.

Solution. The eigenvalues of $\mathbf{A}$ are $\pm i$, which have real part 0, so $\mathbf{A}$ is not stable and neither part of the theorem applies. Nothing weaker survives either: $\mathbf{A}$ is real with $\mathbf{A}^* = \mathbf{A}^\top = -\mathbf{A}$, so for **every** $\mathbf{X} \in M_2(\mathbb{C})$,

$$\operatorname{tr}(\mathbf{AX} + \mathbf{XA}^*) = \operatorname{tr}(\mathbf{AX}) + \operatorname{tr}(\mathbf{A}^*\mathbf{X}) = \operatorname{tr}\left((\mathbf{A} + \mathbf{A}^*)\mathbf{X}\right) = 0$$

by Theorem 0.9.6, while $\operatorname{tr}(-\mathbf{I}_2) = -2$. So no $\mathbf{X}$ solves the equation. This is Theorem 12.10.2 seen from the other side: $\lambda = i$ and $\mu = i$ give $\lambda + \overline{\mu} = 0$, the Sylvester operator is singular, and $-\mathbf{I}_2$ happens to lie outside its image.

The two theorems of this section answer different questions with the same computation. Theorem 13.6.1 decides a positivity question about $k + l$ coordinates by asking it of l of them; Theorem 13.6.3 turns a question about the spectrum of $\mathbf{A}$ into a positivity question about a

single matrix $\mathbf{X}$, which the block test then settles one coordinate at a time. One thing the congruence $(*)$ carries has gone unspent by both: the determinant identity $\det \mathbf{M} = \det \mathbf{A} \cdot \det(\mathbf{M}/\mathbf{A})$ recorded above. The next section reads two classical inequalities off it.

13.6.4 Exercises

13.6.4.1 A. Check your understanding

Exercise. A1

(a) Write down the general form of a Hermitian matrix in 2×2 block form, and say what each block must satisfy.

(b) State Theorem 13.6.1 (b) in full, including its hypothesis.

(c) True or false: if $\mathbf{M} = \left(\begin{smallmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^* & \mathbf{C} \end{smallmatrix}\right)$ is Hermitian with $\mathbf{A} \succeq \mathbf{O}$ and $\mathbf{C} \succeq \mathbf{O}$, then $\mathbf{M} \succeq \mathbf{O}$. Justify your answer.

(d) State both halves of Theorem 13.6.3, and say which of the two is the one that certifies stability.

(e) The induction in the proof of Theorem 13.6.3 (a) conjugates by a unitary matrix whose **last** column is an eigenvector of $\mathbf{A}$. What would go wrong if it were the first column?

Solution.

(a) $\mathbf{M} = \left(\begin{smallmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^* & \mathbf{C} \end{smallmatrix}\right)$ with $\mathbf{A} \in M_k(F)$ and $\mathbf{C} \in M_l(F)$ Hermitian and $\mathbf{B} \in M_{k\times l}(F)$ arbitrary. The condition $\mathbf{M}^* = \mathbf{M}$ forces the lower-left block to be $\mathbf{B}^*$ and both diagonal blocks to be Hermitian.

(b) If $\mathbf{A} \succ \mathbf{O}$, then $\mathbf{A}$ is invertible, $\mathbf{M}/\mathbf{A} = \mathbf{C} - \mathbf{B}^*\mathbf{A}^{-1}\mathbf{B}$ is Hermitian, $\mathbf{M} \succ \mathbf{O}$ if and only if $\mathbf{M}/\mathbf{A} \succ \mathbf{O}$, and $\mathbf{M} \succeq \mathbf{O}$ if and only if $\mathbf{M}/\mathbf{A} \succeq \mathbf{O}$.

(c) False. Take $\mathbf{A} = \mathbf{O}_1$, $\mathbf{B} = (1)$, and $\mathbf{C} = (1)$, so $\mathbf{M} = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 1 \end{smallmatrix}\right)$. Both diagonal blocks are positive semidefinite, but $\det \mathbf{M} = -1 < 0$, so $\mathbf{M} \not\succeq \mathbf{O}$ by Theorem 13.1.4 (e).

(d) Let $\mathbf{Q} \succ \mathbf{O}$. If $\mathbf{A}$ is stable, then $\mathbf{AX} + \mathbf{XA}^* = -\mathbf{Q}$ has exactly one solution and that solution is positive definite; conversely, if some $\mathbf{X} \succ \mathbf{O}$ satisfies the equation, then $\mathbf{A}$ is stable. The second half is the certificate: one positive definite $\mathbf{X}$ proves that every eigenvalue has negative real part, and no eigenvalue is computed.

(e) With the eigenvector last, $\hat{\mathbf{A}} = \mathbf{W}^*\mathbf{A}\mathbf{W}$ is block **lower** triangular with the 1×1 corner at the bottom right, so the leading $(n-1) \times (n-1)$ block of the transformed equation is again a Lyapunov equation and the inductive hypothesis applies to it; and the surviving coordinate is then handled by a Schur complement whose corner block is the leading $\mathbf{X}' \succ \mathbf{O}$, which is the shape Theorem 13.6.1 (b) requires. Putting the eigenvector first would make $\hat{\mathbf{A}}$ block upper triangular, leaving the Lyapunov sub-equation in the **trailing** block and the Schur complement with a corner the theorem says nothing about.

13.6.4.2 B. Practice

Exercise. B1: Definiteness by blocks

Each matrix below is real symmetric and is displayed with a $2, 2$ partition. Determine whether it is positive definite. Justify your answer, and for any that is not, give a vector $\mathbf{x}$

with $\mathbf{x}^\top \mathbf{M}\mathbf{x} < 0$.

(a)
$$\mathbf{M}_1 = \left(\begin{array}{cc|cc} 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & -1 \\ \hline 1 & 1 & 3 & 0 \\ 1 & -1 & 0 & 3 \end{array}\right)$$

(b)
$$\mathbf{M}_2 = \left(\begin{array}{cc|cc} 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & -1 \\ \hline 1 & 1 & 2 & 0 \\ 1 & -1 & 0 & 1 \end{array}\right)$$

Solution. In both cases $\mathbf{A} = \mathbf{I}_2 \succ \mathbf{O}$ and $\mathbf{B} = \left(\begin{smallmatrix} 1 & 1 \\ 1 & -1 \end{smallmatrix}\right)$, so Theorem 13.6.1 (b) applies with $\mathbf{A}^{-1} = \mathbf{I}_2$ and

$$\mathbf{B}^\top \mathbf{A}^{-1} \mathbf{B} = \mathbf{B}^\top \mathbf{B} = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}.$$

(a) $\mathbf{M}_1/\mathbf{A} = \left(\begin{smallmatrix} 3 & 0 \\ 0 & 3 \end{smallmatrix}\right) - \left(\begin{smallmatrix} 2 & 0 \\ 0 & 2 \end{smallmatrix}\right) = \mathbf{I}_2 \succ \mathbf{O}$. Hence $\mathbf{M}_1 \succ \mathbf{O}$, and $\det \mathbf{M}_1 = \det \mathbf{A} \det(\mathbf{M}_1/\mathbf{A}) = 1$.

(b) $\mathbf{M}_2/\mathbf{A} = \left(\begin{smallmatrix} 2 & 0 \\ 0 & 1 \end{smallmatrix}\right) - \left(\begin{smallmatrix} 2 & 0 \\ 0 & 2 \end{smallmatrix}\right) = \left(\begin{smallmatrix} 0 & 0 \\ 0 & -1 \end{smallmatrix}\right)$. This is the matrix of Section 1's warning: its leading principal minors are 0 and 0, and yet it is not positive semidefinite. So $\mathbf{M}_2$ is not positive definite, and not even positive semidefinite.

For the witness, $\mathbf{y}_2 = (0, 1)$ gives $\mathbf{y}_2^\top (\mathbf{M}_2/\mathbf{A})\mathbf{y}_2 = -1$. With $\mathbf{S} = \left(\begin{smallmatrix} \mathbf{I}_2 & -\mathbf{B} \\ \mathbf{O} & \mathbf{I}_2 \end{smallmatrix}\right)$ and $\mathbf{y} = (0, 0, 0, 1)$, the vector $\mathbf{x} = \mathbf{S}\mathbf{y} = (-1, 1, 0, 1)$ satisfies $\mathbf{x}^\top \mathbf{M}_2 \mathbf{x} = -1$.

Exercise. B2: Reading a certificate

(a) Let
$$\mathbf{A} = \begin{pmatrix} -3 & 4 \\ -2 & 2 \end{pmatrix}, \qquad \mathbf{X} = \frac{1}{4}\begin{pmatrix} 22 & 16 \\ 16 & 15 \end{pmatrix}.$$
Check that $\mathbf{A}\mathbf{X} + \mathbf{X}\mathbf{A}^\top = -\mathbf{I}_2$ and that $\mathbf{X} \succ \mathbf{O}$, and conclude that $\mathbf{A}$ is stable without solving a characteristic equation. Then compute the characteristic polynomial and confirm the verdict.

(b) Show that the hypothesis $\mathbf{Q} \succ \mathbf{O}$ in Theorem 13.6.3 (b) cannot be weakened to $\mathbf{Q} \succeq \mathbf{O}$: exhibit an $\mathbf{A}$ that is not stable together with an $\mathbf{X} \succ \mathbf{O}$ and a $\mathbf{Q} \succeq \mathbf{O}$ satisfying $\mathbf{A}\mathbf{X} + \mathbf{X}\mathbf{A}^* = -\mathbf{Q}$.

Solution.

(a) Multiplying,
$$\mathbf{A}\mathbf{X} = \frac{1}{4}\begin{pmatrix} -66 + 64 & -48 + 60 \\ -44 + 32 & -32 + 30 \end{pmatrix} = \frac{1}{4}\begin{pmatrix} -2 & 12 \\ -12 & -2 \end{pmatrix},$$

and since $\mathbf{X}$ is symmetric, $\mathbf{X}\mathbf{A}^\top = (\mathbf{A}\mathbf{X})^\top$. The off-diagonal entries of $\mathbf{A}\mathbf{X}$ are negatives of each other, so they cancel, and

$$\mathbf{A}\mathbf{X} + \mathbf{X}\mathbf{A}^\top = \frac{1}{4}\begin{pmatrix} -4 & 0 \\ 0 & -4 \end{pmatrix} = -\mathbf{I}_2.$$

The leading principal minors of $\mathbf{X}$ are $22/4 = 11/2 > 0$ and $\det \mathbf{X} = \frac{1}{16}(22 \cdot 15 - 16^2) = \frac{74}{16} = \frac{37}{8} > 0$, so $\mathbf{X} \succ \mathbf{O}$ by Sylvester's criterion (Theorem 13.1.5 (d)). With $\mathbf{Q} = \mathbf{I}_2 \succ \mathbf{O}$, Theorem 13.6.3 (b) gives that $\mathbf{A}$ is stable. Checking: $\operatorname{tr}\mathbf{A} = -1$ and $\det \mathbf{A} = -6 + 8 = 2$, so the characteristic polynomial is $\lambda^2 + \lambda + 2$, whose roots $(-1 \pm i\sqrt{7})/2$ have real part $-\frac{1}{2} < 0$.

(b) Take the quarter turn $\mathbf{A} = \left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)$ and $\mathbf{X} = \mathbf{I}_2 \succ \mathbf{O}$. Then $\mathbf{A}^* = \mathbf{A}^\top = -\mathbf{A}$, so $\mathbf{A}\mathbf{X} + \mathbf{X}\mathbf{A}^* = \mathbf{A} - \mathbf{A} = \mathbf{O}$, and the equation holds with $\mathbf{Q} = \mathbf{O} \succeq \mathbf{O}$. But the eigenvalues of $\mathbf{A}$ are $\pm i$, with real part 0, so $\mathbf{A}$ is not stable. A positive definite solution on its own certifies nothing; it is the strict positivity of $\mathbf{Q}$ that makes $-\mathbf{u}^*\mathbf{Q}\mathbf{u}$ strictly negative in Eq. 13.6.1, and that is what forces $\operatorname{Re}\lambda < 0$.

13.6.4.3 C. Going deeper

Exercise. C1: A zero corner kills its border

(a) Prove that if $\mathbf{P} \succeq \mathbf{O}$ and $\mathbf{x}^*\mathbf{P}\mathbf{x} = 0$, then $\mathbf{P}\mathbf{x} = \mathbf{0}$.

(b) Deduce that if $\mathbf{M} = \left(\begin{smallmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^* & \mathbf{C} \end{smallmatrix}\right) \succeq \mathbf{O}$ and $\mathbf{A} = \mathbf{O}$, then $\mathbf{B} = \mathbf{O}$.

(c) Deduce that if $\mathbf{M} \succeq \mathbf{O}$ and $\mathbf{A}$ is singular, then $\mathbf{M}$ is singular. What does Theorem 13.6.1 say about such an $\mathbf{M}$?

Hint for (a): write $\mathbf{P} = \mathbf{N}^*\mathbf{N}$.

Solution.

(a) By Theorem 13.1.4 (c), $\mathbf{P} = \mathbf{N}^*\mathbf{N}$ for some $\mathbf{N}$. Then $0 = \mathbf{x}^*\mathbf{P}\mathbf{x} = \|\mathbf{N}\mathbf{x}\|^2$, so $\mathbf{N}\mathbf{x} = \mathbf{0}$ and $\mathbf{P}\mathbf{x} = \mathbf{N}^*(\mathbf{N}\mathbf{x}) = \mathbf{0}$.

(b) Let $\mathbf{x} \in F^k$ and put $\mathbf{v} = (\mathbf{x}, \mathbf{0})$. As computed in the proof of Theorem 13.6.1 (a), $\mathbf{v}^*\mathbf{M}\mathbf{v} = \mathbf{x}^*\mathbf{A}\mathbf{x} = 0$. By (a), $\mathbf{M}\mathbf{v} = \mathbf{0}$. But $\mathbf{M}\mathbf{v} = (\mathbf{A}\mathbf{x}, \mathbf{B}^*\mathbf{x}) = (\mathbf{0}, \mathbf{B}^*\mathbf{x})$, so $\mathbf{B}^*\mathbf{x} = \mathbf{0}$ for every $\mathbf{x} \in F^k$, that is $\mathbf{B}^* = \mathbf{O}$ and $\mathbf{B} = \mathbf{O}$.

(c) Choose $\mathbf{x} \neq \mathbf{0}$ with $\mathbf{A}\mathbf{x} = \mathbf{0}$, which is possible since $\mathbf{A}$ is singular (Theorem 3.4.4), and put $\mathbf{v} = (\mathbf{x}, \mathbf{0}) \neq \mathbf{0}$. Then $\mathbf{v}^*\mathbf{M}\mathbf{v} = \mathbf{x}^*\mathbf{A}\mathbf{x} = 0$, so $\mathbf{M}\mathbf{v} = \mathbf{0}$ by (a), and $\mathbf{M}$ is singular. Theorem 13.6.1 says nothing at all about such an $\mathbf{M}$: part (a) is a one-way implication, and part (b) needs $\mathbf{A} \succ \mathbf{O}$, which a singular $\mathbf{A}$ is not. This is the situation the warning after the theorem describes, and the criterion that does cover it waits for Section 12.

Exercise. C2: A Cauchy–Schwarz inequality for blocks

Let $\mathbf{M} = \left(\begin{smallmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^* & \mathbf{C} \end{smallmatrix}\right) \succeq \mathbf{O}$ with $\mathbf{A} \in M_k(F)$ and $\mathbf{C} \in M_l(F)$.

(a) Prove that for all $\mathbf{x} \in F^k$ and $\mathbf{y} \in F^l$,

$$|\mathbf{x}^*\mathbf{B}\mathbf{y}|^2 \leq (\mathbf{x}^*\mathbf{A}\mathbf{x})\,(\mathbf{y}^*\mathbf{C}\mathbf{y}).$$

(b) Deduce the Cauchy–Schwarz inequality $|\langle \mathbf{y}, \mathbf{x}\rangle| \le \|\mathbf{x}\|\,\|\mathbf{y}\|$ in F^n.

Hint for (a): apply the definition of $\mathbf{M} \succeq \mathbf{O}$ to the vector $(\mathbf{x}, t\mathbf{y})$, and choose the scalar t well.

Solution.

(a) Write $p = \mathbf{x}^*\mathbf{A}\mathbf{x} \ge 0$, $q = \mathbf{y}^*\mathbf{C}\mathbf{y} \ge 0$, and $z = \mathbf{x}^*\mathbf{B}\mathbf{y}$. For a scalar $t \in F$, block multiplication (Theorem 8.1.1) gives, for $\mathbf{v} = (\mathbf{x}, t\mathbf{y})$,

$$\mathbf{v}^*\mathbf{M}\mathbf{v} = p + \bar{t}\,\bar{z} + tz + |t|^2 q = p + 2\,\mathrm{Re}(tz) + |t|^2 q \ge 0,$$

where the middle step uses $\bar{t}\,\bar{z} = \overline{tz}$ and $w + \overline{w} = 2\,\mathrm{Re}(w)$. Now take $t = r\omega$ with $r \ge 0$ real and ω a scalar of modulus 1 chosen so that $\omega z = -|z|$ (over $\mathbb{R}$, $\omega = \pm 1$). Then

$$p - 2r|z| + r^2 q \ge 0 \qquad \text{for every } r \ge 0. \tag{$*$}$$

If $q = 0$, then $(*)$ says $p \ge 2r|z|$ for all $r \ge 0$, which forces $|z| = 0$, and the claim reads $0 \le 0$. If $q > 0$, take $r = |z|/q \ge 0$ in $(*)$:

$$p - \frac{2|z|^2}{q} + \frac{|z|^2}{q} = p - \frac{|z|^2}{q} \ge 0,$$

so $|z|^2 \le pq$, as claimed.

(b) Take $k = l = n$ and $\mathbf{A} = \mathbf{B} = \mathbf{C} = \mathbf{I}_n$, so that $\mathbf{M} = \begin{pmatrix} \mathbf{I}_n & \mathbf{I}_n \\ \mathbf{I}_n & \mathbf{I}_n \end{pmatrix}$. This $\mathbf{M}$ is positive semidefinite, since $\mathbf{M} = \mathbf{W}^*\mathbf{W}$ for $\mathbf{W} = \begin{pmatrix} \mathbf{I}_n & \mathbf{I}_n \\ \mathbf{O} & \mathbf{O} \end{pmatrix}$ (Theorem 13.1.4 (c)). Part (a) then reads $|\mathbf{x}^*\mathbf{y}|^2 \le \|\mathbf{x}\|^2\|\mathbf{y}\|^2$, and $\mathbf{x}^*\mathbf{y} = \langle \mathbf{y}, \mathbf{x}\rangle$ by the definition of the standard inner product. This is Theorem 11.1.5 for F^n.

13.7 Determinant Inequalities for Positive Matrices

Section 6 cut a Hermitian matrix into blocks and read a verdict off the corner: once the corner block is positive definite, the whole matrix is positive definite exactly when one smaller matrix, the Schur complement, is. The same cut carries a second piece of information, which that verdict threw away. The congruence behind it is a factorization, so it computes a determinant as well, and what the off-diagonal block does to that determinant goes one way only: it can push the determinant down, and never up.

That single observation, applied to a two-block partition, is Fischer's inequality: the determinant of a positive matrix is at most the product of the determinants of its two diagonal blocks. Applied again and again, until every block is a single entry, it is Hadamard's inequality: the determinant is at most the product of the diagonal entries. The second has a picture attached, and it is the reason the inequality is remembered — among all parallelepipeds with prescribed edge lengths, the rectangular one has the largest volume.

Throughout, $F = \mathbb{R}$ or $F = \mathbb{C}$, and F^n carries the standard inner product. A Hermitian $\mathbf{M} \in M_{k+l}(F)$ is written in blocks as

$$\mathbf{M} = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^* & \mathbf{C} \end{pmatrix}, \qquad \mathbf{A} \in M_k(F), \quad \mathbf{B} \in M_{k\times l}(F), \quad \mathbf{C} \in M_l(F),$$

with $\mathbf{A}$ and $\mathbf{C}$ Hermitian, as in Section 6.

13.7.1 Fischer's inequality

When $\mathbf{A} \succ \mathbf{O}$, the congruence of Theorem 13.6.1 turns $\mathbf{M}$ into $\mathbf{A} \oplus (\mathbf{M}/\mathbf{A})$, and taking determinants in it gives Theorem 8.3.2,

$$\det \mathbf{M} = \det \mathbf{A} \cdot \det(\mathbf{M}/\mathbf{A}).$$

If the Schur complement could be compared with $\mathbf{C}$, that would bound $\det \mathbf{M}$ by $\det \mathbf{A} \det \mathbf{C}$ — and it can be, because $\mathbf{C}$ is the Schur complement plus something positive:

$$\mathbf{C} = (\mathbf{M}/\mathbf{A}) + \mathbf{B}^*\mathbf{A}^{-1}\mathbf{B}.$$

So what is needed is that adding a positive semidefinite matrix never decreases a determinant. That is the next lemma, and it is where the square root of Section 2 earns its keep.

Lemma 13.7.1. Adding a positive matrix does not decrease the determinant

Let $\mathbf{P}, \mathbf{Q} \in M_n(F)$ with $\mathbf{P} \succ \mathbf{O}$ and $\mathbf{Q} \succeq \mathbf{O}$. Then

$$\det(\mathbf{P} + \mathbf{Q}) \geq \det \mathbf{P} > 0,$$

with equality if and only if $\mathbf{Q} = \mathbf{O}$.

Idea. Whiten: divide the problem by $\mathbf{P}$ on both sides, using the inverse of $\mathbf{P}^{1/2}$. The sum $\mathbf{P} + \mathbf{Q}$ becomes $\mathbf{I} + \mathbf{K}$ with $\mathbf{K} \succeq \mathbf{O}$, and the determinant of $\mathbf{I} + \mathbf{K}$ is $\prod_i (1 + \mu_i)$ over the eigenvalues $\mu_i \geq 0$ of $\mathbf{K}$. A product of numbers ≥ 1 is ≥ 1, and it equals 1 only when every factor does.

Proof. By Theorem 13.2.1 the matrix $\mathbf{P}^{1/2} \succeq \mathbf{O}$ exists, and it is invertible because $\mathbf{P}$ is (Theorem 13.2.1 (c)). Write $\mathbf{R} = (\mathbf{P}^{1/2})^{-1}$, which is Hermitian, being the inverse of a Hermitian matrix. Since $\mathbf{P}^{1/2}\mathbf{R} = \mathbf{I}$ and $(\mathbf{P}^{1/2})^2 = \mathbf{P}$,

$$\mathbf{P}^{1/2}(\mathbf{I} + \mathbf{R}\mathbf{Q}\mathbf{R})\mathbf{P}^{1/2} = \mathbf{P} + \mathbf{Q}.$$

Put $\mathbf{K} = \mathbf{R}^*\mathbf{Q}\mathbf{R} = \mathbf{R}\mathbf{Q}\mathbf{R}$, so that $\mathbf{K} \succeq \mathbf{O}$ by Proposition 13.1.2 (a). Taking determinants with Theorem 7.5.4, and using $(\det \mathbf{P}^{1/2})^2 = \det \mathbf{P}$,

$$\det(\mathbf{P} + \mathbf{Q}) = \det \mathbf{P} \cdot \det(\mathbf{I} + \mathbf{K}).$$

Now $\mathbf{K}$ is Hermitian, so $\mathbf{K} = \mathbf{U}\,\mathrm{diag}(\mu_1, \dots, \mu_n)\mathbf{U}^*$ with $\mathbf{U}$ unitary, by Corollary 12.4.3 over $\mathbb{C}$ and Corollary 12.5.2 over $\mathbb{R}$; and every $\mu_i \geq 0$ by Theorem 13.1.4 (b). Then $\mathbf{I} + \mathbf{K} = \mathbf{U}\,\mathrm{diag}(1 + \mu_1, \dots, 1 + \mu_n)\mathbf{U}^*$, so

$$\det(\mathbf{I} + \mathbf{K}) = \prod_{i=1}^{n} (1 + \mu_i) \geq 1,$$

each factor being ≥ 1. Also $\det \mathbf{P} > 0$, since the eigenvalues of $\mathbf{P}$ are positive (Theorem 13.1.5 (b)) and the determinant is their product. Hence $\det(\mathbf{P} + \mathbf{Q}) \geq \det \mathbf{P} > 0$.

For the equality case, the product equals 1 exactly when every $\mu_i = 0$, that is when $\mathbf{K} = \mathbf{U}\,\mathbf{O}\,\mathbf{U}^* = \mathbf{O}$, and then $\mathbf{Q} = \mathbf{P}^{1/2}\mathbf{K}\mathbf{P}^{1/2} = \mathbf{O}$. Conversely $\mathbf{Q} = \mathbf{O}$ gives equality. This proves the lemma.

■

Remark. Section 5 states this in its natural generality: the determinant is monotone for the Loewner order, so $\mathbf{A} \succeq \mathbf{B} \succeq \mathbf{O}$ implies $\det \mathbf{A} \geq \det \mathbf{B}$. The special case proved here, in which the smaller of the two matrices is definite, is all this section and the next one need, and its proof is short enough to keep local.

Corollary 13.7.2. Fischer's Inequality

Let $\mathbf{M} = \left(\begin{smallmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^* & \mathbf{C} \end{smallmatrix}\right) \in M_{k+l}(F)$ with $\mathbf{M} \succeq \mathbf{O}$, where $\mathbf{A} \in M_k(F)$ and $\mathbf{C} \in M_l(F)$. Then

$$\det \mathbf{M} \leq \det \mathbf{A} \cdot \det \mathbf{C}.$$

If $\mathbf{M} \succ \mathbf{O}$, equality holds if and only if $\mathbf{B} = \mathbf{O}$.

Idea. Do the definite case first and reach the general case by a limit. For $\mathbf{M} \succ \mathbf{O}$ the Schur complement exists, the determinant splits as $\det \mathbf{A}\det(\mathbf{M}/\mathbf{A})$, and $\det \mathbf{C} \geq \det(\mathbf{M}/\mathbf{A})$ by Lemma 13.7.1 applied to $\mathbf{C} = (\mathbf{M}/\mathbf{A}) + \mathbf{B}^*\mathbf{A}^{-1}\mathbf{B}$. For $\mathbf{M} \succeq \mathbf{O}$, apply the definite case to $\mathbf{M} + t\mathbf{I}$ and let t shrink to 0: both sides are polynomials in t.

Proof. Case 1: $\mathbf{M} \succ \mathbf{O}$. By Theorem 13.6.1 (a), $\mathbf{A} \succ \mathbf{O}$, and by (b), $\mathbf{M}/\mathbf{A} \succ \mathbf{O}$. By Theorem 8.3.2,

$$\det \mathbf{M} = \det \mathbf{A} \cdot \det(\mathbf{M}/\mathbf{A}).$$

The matrix $\mathbf{A}^{-1}$ is positive definite. Indeed the eigenvalues of $\mathbf{A}$ are positive (Theorem 13.1.5 (b)), so $\mathbf{A}$ is invertible (Theorem 9.2.6); the identity $(\mathbf{A}^{-1})^* = (\mathbf{A}^*)^{-1} = \mathbf{A}^{-1}$ makes $\mathbf{A}^{-1}$ Hermitian; and Proposition 13.1.2 (b) with the invertible $\mathbf{S} = \mathbf{A}^{-1}$ turns $\mathbf{S}^*\mathbf{A}\mathbf{S} = \mathbf{A}^{-1}\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}$ into $\mathbf{A}^{-1} \succ \mathbf{O}$. (This is Exercise (B3: The inverse of a positive definite matrix).) So $\mathbf{Q} := \mathbf{B}^*\mathbf{A}^{-1}\mathbf{B} \succeq \mathbf{O}$ by Proposition 13.1.2 (a) with $\mathbf{S} = \mathbf{B}$. Since

$\mathbf{C} = (\mathbf{M}/\mathbf{A}) + \mathbf{Q}$, Lemma 13.7.1 with $\mathbf{P} = \mathbf{M}/\mathbf{A}$ gives $\det \mathbf{C} \geq \det(\mathbf{M}/\mathbf{A})$, with equality if and only if $\mathbf{Q} = \mathbf{O}$. Multiplying by $\det \mathbf{A} > 0$,

$$\det \mathbf{M} = \det \mathbf{A} \det(\mathbf{M}/\mathbf{A}) \leq \det \mathbf{A} \det \mathbf{C},$$

with equality if and only if $\mathbf{B}^* \mathbf{A}^{-1} \mathbf{B} = \mathbf{O}$.

It remains to see that $\mathbf{B}^* \mathbf{A}^{-1} \mathbf{B} = \mathbf{O}$ holds only for $\mathbf{B} = \mathbf{O}$. Let $\mathbf{R} = (\mathbf{A}^{1/2})^{-1}$, a Hermitian matrix with $\mathbf{R}^2 = \mathbf{A}^{-1}$, as in the proof of Lemma 13.7.1. Then $\mathbf{B}^* \mathbf{A}^{-1} \mathbf{B} = (\mathbf{RB})^* (\mathbf{RB})$, whose (j, j) entry is the squared norm of the j-th column of $\mathbf{RB}$. If the whole matrix is $\mathbf{O}$, every column of $\mathbf{RB}$ is $\mathbf{0}$, so $\mathbf{RB} = \mathbf{O}$ and $\mathbf{B} = \mathbf{A}^{1/2}(\mathbf{RB}) = \mathbf{O}$. The converse is clear.

Case 2: $\mathbf{M} \succeq \mathbf{O}$. For real $t > 0$ the matrix $\mathbf{M} + t\mathbf{I}_{k+l}$ is Hermitian with $\mathbf{v}^*(\mathbf{M} + t\mathbf{I})\mathbf{v} = \mathbf{v}^*\mathbf{M}\mathbf{v} + t\|\mathbf{v}\|^2 > 0$ for $\mathbf{v} \neq \mathbf{0}$, so $\mathbf{M} + t\mathbf{I} \succ \mathbf{O}$. Its blocks are $\mathbf{A} + t\mathbf{I}_k$, $\mathbf{B}$, and $\mathbf{C} + t\mathbf{I}_l$, so Case 1 gives

$$\det(\mathbf{M} + t\mathbf{I}) \leq \det(\mathbf{A} + t\mathbf{I}_k) \cdot \det(\mathbf{C} + t\mathbf{I}_l)$$

for every $t > 0$. A determinant is a polynomial in the entries (Theorem 7.4.3), and the entries here are polynomials in t, so both sides are polynomial, hence continuous, functions of t. Letting t decrease to 0 preserves the inequality, and at $t = 0$ it reads $\det \mathbf{M} \leq \det \mathbf{A} \det \mathbf{C}$. This proves the corollary.

■

The inequality says that for a positive matrix the off-diagonal blocks are a liability: they can only shrink the determinant, and they cost nothing only when they are absent. Applying it repeatedly, to finer and finer partitions, gives the classical case in which every block is 1×1. That is the next result.

Quick check. What does Corollary 13.7.2 say for $k = l = 1$, that is for $\mathbf{M} = \left(\begin{smallmatrix} a & b \\ b & c \end{smallmatrix}\right) \succeq \mathbf{O}$?

Solution. It says $ac - |b|^2 = \det \mathbf{M} \leq \det(a) \det(c) = ac$, that is $|b|^2 \geq 0$ — true for any b whatsoever, and with equality exactly when $b = 0$. So in the smallest case the inequality carries no information beyond its equality condition. The content is in the blocks: for larger k and l the statement is no longer a triviality about one number, and its proof needs the whole of Lemma 13.7.1.

13.7.2 Hadamard's inequality

Theorem 13.7.3. Hadamard's Inequality

(a) Let $\mathbf{A} \in M_n(F)$ with $\mathbf{A} \succeq \mathbf{O}$. Then

$$\det \mathbf{A} \leq a_{11} a_{22} \cdots a_{nn},$$

with equality if and only if $\mathbf{A}$ is diagonal or some $a_{ii} = 0$.

(b) Let $\mathbf{B} \in M_n(F)$ have columns $\mathbf{b}_1, \dots, \mathbf{b}_n$. Then

$$|\det \mathbf{B}| \le \|\mathbf{b}_1\| \, \|\mathbf{b}_2\| \cdots \|\mathbf{b}_n\|,$$

with equality if and only if the columns are pairwise orthogonal or some column is $\mathbf{0}$.

Idea. Part (a) is Fischer's inequality with $k = 1$, applied again and again: peel off the first row and column, bound $\det \mathbf{A}$ by $a_{11} \det \mathbf{A}'$, and induct on the size. Part (b) is part (a) for the Gram matrix $\mathbf{A} = \mathbf{B}^*\mathbf{B}$, whose diagonal entries are the squared lengths $\|\mathbf{b}_j\|^2$ and whose determinant is $|\det \mathbf{B}|^2$. The equality conditions transfer along the same dictionary: $\mathbf{A}$ diagonal says the columns of $\mathbf{B}$ are orthogonal, and $a_{jj} = 0$ says $\mathbf{b}_j = \mathbf{0}$.

Proof.

(a) Induct on n. For $n = 1$ both sides are a_{11}.
Let $n \ge 2$ and split $\mathbf{A}$ with $k = 1$:

$$\mathbf{A} = \begin{pmatrix} a_{11} & \mathbf{b}^* \\ \mathbf{b} & \mathbf{A}' \end{pmatrix}, \qquad \mathbf{b} \in F^{n-1}, \quad \mathbf{A}' \in M_{n-1}(F),$$

so that $\mathbf{A}'$ is the submatrix on rows and columns $2, \dots, n$. By Theorem 13.6.1 (a), $\mathbf{A}' \succeq \mathbf{O}$, and by Corollary 13.7.2,

$$\det \mathbf{A} \le a_{11} \det \mathbf{A}' \le a_{11} \, (a_{22} \cdots a_{nn}),$$

the second step being the inductive hypothesis applied to $\mathbf{A}'$. This is the inequality.
For equality, suppose first that some $a_{ii} = 0$. Fix $j \ne i$ and put $I = \{i, j\}$. By Theorem 13.1.4 ((a) $\Rightarrow$ (e)) the principal minor on the rows and columns I is ≥ 0, so

$$0 \le \det \mathbf{A}_{I,I} = a_{ii}a_{jj} - |a_{ij}|^2 = -|a_{ij}|^2,$$

forcing $a_{ij} = 0$; and $a_{ji} = \overline{a_{ij}} = 0$ because $\mathbf{A}$ is Hermitian. As $j \ne i$ was arbitrary, row i and column i of $\mathbf{A}$ are zero (which is Exercise (C3: The largest entry sits on the diagonal) (b)) and $\det \mathbf{A} = 0$; the right-hand side is 0 too, so equality holds. If $\mathbf{A}$ is diagonal, both sides are the product of the diagonal entries. Conversely, suppose equality holds and every $a_{ii} > 0$. Then $\det \mathbf{A} = a_{11} \cdots a_{nn} > 0$, so $\mathbf{A}$ is invertible, and a positive semidefinite matrix with no zero eigenvalue is positive definite (Theorem 13.1.5 (b)). Both inequalities in the display are therefore equalities. The first is the equality case of Corollary 13.7.2 for the positive definite $\mathbf{A}$, so $\mathbf{b} = \mathbf{0}$: the first row and column vanish off the diagonal. The second is the equality case of the inductive hypothesis for $\mathbf{A}'$, whose diagonal entries are positive, so $\mathbf{A}'$ is diagonal. Hence $\mathbf{A}$ is diagonal, and the induction is complete.

(b) Put $\mathbf{A} = \mathbf{B}^*\mathbf{B}$, which is positive semidefinite by Theorem 13.1.4 (c). Its (j, j) entry is $\mathbf{b}_j^*\mathbf{b}_j = \|\mathbf{b}_j\|^2$. By Theorem 7.5.4, and because $\det(\mathbf{B}^*) = \overline{\det \mathbf{B}}$ — the star transposes, which does not change the determinant (Theorem 7.4.5), and conjugates every entry, which conjugates the determinant —

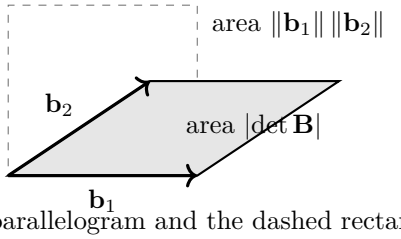
$$\det \mathbf{A} = \overline{\det \mathbf{B}} \, \det \mathbf{B} = |\det \mathbf{B}|^2.$$

Part (a) now reads $|\det \mathbf{B}|^2 \le \|\mathbf{b}_1\|^2 \cdots \|\mathbf{b}_n\|^2$, and taking nonnegative square roots gives the inequality.

For equality, note that $a_{jj} = 0$ says $\mathbf{b}_j = \mathbf{0}$, and that $\mathbf{A}$ is diagonal exactly when $\langle \mathbf{b}_j, \mathbf{b}_i \rangle = a_{ij} = 0$ for all $i \neq j$, that is, exactly when the columns are pairwise orthogonal. So the equality condition of (a) translates into the stated one. This proves the theorem.

■

Remark. Part (b) has a geometric reading, and it is the reason the inequality is remembered. Over $\mathbb{R}$, $|\det \mathbf{B}|$ is the volume of the parallelepiped with edges $\mathbf{b}_1, \dots, \mathbf{b}_n$ (Definition 7.9.1), and the right-hand side is the volume of the box with the same edge lengths. So among all parallelepipeds with prescribed edge lengths, the rectangular one has the largest volume, and up to degenerate cases it is the only one. Tilting an edge costs volume.



the shaded parallelogram and the dashed rectangle have the same two edge lengths

Example. Two Determinant Bounds

Bound $|\det \mathbf{B}|$ using Theorem 13.7.3 (b) for

$$\mathbf{B}_1 = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}, \qquad \mathbf{B}_2 = \begin{pmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{pmatrix},$$

and compare each bound with the exact value.

Solution. For $\mathbf{B}_1$, both columns have norm $\sqrt{2}$, so the bound is 2. The determinant is $-1 - 1 = -2$, so $|\det \mathbf{B}_1| = 2$ and the bound is attained. The equality case predicts this: $\langle (1,-1), (1,1) \rangle = 1 - 1 = 0$, so the columns are orthogonal.

For $\mathbf{B}_2$, the columns have norms $\sqrt{5}$, $\sqrt{6}$, and $\sqrt{5}$, so the bound is $5\sqrt{6} \approx 12.25$. Expanding along the first row, $\det \mathbf{B}_2 = 2(4-1) - 1(2-0) = 4$. The bound holds and is far from sharp, as it must be, since $\langle (1,2,1), (2,1,0) \rangle = 4 \neq 0$ means the columns are not orthogonal.

Warning. **Part (a) needs positivity; without it the inequality can fail in either direction.** The symmetric matrix

$$\mathbf{A} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & -1 & 0 \\ 1 & 0 & -1 \end{pmatrix}$$

has $\det \mathbf{A} = 3$, while $a_{11}a_{22}a_{33} = 1 \cdot (-1) \cdot (-1) = 1$. So $\det \mathbf{A} > \prod_i a_{ii}$, and a diagonal entry of an indefinite matrix carries no information about the size of the determinant. Part (b), by contrast, has no hypothesis at all — but it bounds by **column** norms. For rows, apply the same result to $\mathbf{B}^\top$, whose determinant is unchanged (Theorem 7.4.5).

13.7.3 Exercises

13.7.3.1 A. Check your understanding

Exercise. A1

(a) State Corollary 13.7.2, including its hypothesis, and say when equality holds.

(b) State both forms of Theorem 13.7.3, and say what hypothesis each one needs.

(c) In the proof of Fischer's inequality, which matrix plays the role of $\mathbf{Q}$ in Lemma 13.7.1, and which plays the role of $\mathbf{P}$?

(d) True or false: if $\mathbf{A} \succeq \mathbf{O}$ and $\det \mathbf{A} = a_{11}a_{22}\cdots a_{nn}$, then $\mathbf{A}$ is diagonal. Justify your answer.

(e) Theorem 13.7.3 (b) holds for **every** square $\mathbf{B}$, with no positivity assumed anywhere. Where did the hypothesis go?

Solution.

(a) For $\mathbf{M} = \left(\begin{smallmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^* & \mathbf{C} \end{smallmatrix}\right) \succeq \mathbf{O}$ with square diagonal blocks, $\det \mathbf{M} \le \det \mathbf{A} \cdot \det \mathbf{C}$. The hypothesis is that the whole of $\mathbf{M}$ is positive semidefinite. When $\mathbf{M} \succ \mathbf{O}$, equality holds exactly when $\mathbf{B} = \mathbf{O}$.

(b) For $\mathbf{A} \succeq \mathbf{O}$, $\det \mathbf{A} \le a_{11}\cdots a_{nn}$, and the hypothesis is positive semidefiniteness. For any square $\mathbf{B}$: $|\det \mathbf{B}| \le \|\mathbf{b}_1\| \cdots \|\mathbf{b}_n\|$ over the columns, with no hypothesis.

(c) $\mathbf{P} = \mathbf{M}/\mathbf{A}$ and $\mathbf{Q} = \mathbf{B}^*\mathbf{A}^{-1}\mathbf{B}$, so that $\mathbf{P} + \mathbf{Q} = \mathbf{C}$.

(d) False. Equality also holds whenever some $a_{ii} = 0$, and then $\mathbf{A}$ need not be diagonal: the matrix

$$\mathbf{A} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

is positive semidefinite, being $\mathbf{N}^*\mathbf{N}$ for $\mathbf{N}$ the matrix with first row $(1, 1, 0)$ and the rest zero, and both sides are 0. What Theorem 13.7.3 (a) says is that equality forces $\mathbf{A}$ to be diagonal **or** some diagonal entry to vanish, and the second alternative cannot be dropped.

(e) Into the Gram matrix. Part (b) is part (a) applied to $\mathbf{A} = \mathbf{B}^*\mathbf{B}$, which is positive semidefinite whatever $\mathbf{B}$ is (Theorem 13.1.4 (c)). The positivity is not assumed because it is manufactured.

13.7.3.2 B. Practice

Exercise. B1: A bordered matrix

Let

$$\mathbf{M} = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 0 \\ 1 & 0 & 2 \end{pmatrix}.$$

(a) Using the partition with $k = 1$, compute $\mathbf{M}/\mathbf{A}$ and decide whether $\mathbf{M} \succ \mathbf{O}$.

(b) Hence compute $\det \mathbf{M}$, and check Fischer's inequality for this partition.

Solution.

(a) Here $\mathbf{A} = (2) \succ \mathbf{O}$, $\mathbf{B} = \begin{pmatrix} 1 & 1 \end{pmatrix}$, and $\mathbf{C} = \left(\begin{smallmatrix} 2 & 0 \\ 0 & 2 \end{smallmatrix}\right)$. Then

$$\mathbf{B}^* \mathbf{A}^{-1} \mathbf{B} = \tfrac{1}{2} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \begin{pmatrix} 1 & 1 \end{pmatrix} = \tfrac{1}{2} \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}, \qquad \mathbf{M}/\mathbf{A} = \begin{pmatrix} 3/2 & -1/2 \\ -1/2 & 3/2 \end{pmatrix}.$$

The leading principal minors of $\mathbf{M}/\mathbf{A}$ are $3/2$ and $9/4 - 1/4 = 2$, both positive, so $\mathbf{M}/\mathbf{A} \succ \mathbf{O}$ and $\mathbf{M} \succ \mathbf{O}$ by Theorem 13.6.1 (b).

(b) By Theorem 8.3.2, $\det \mathbf{M} = 2 \cdot 2 = 4$. Fischer's inequality (Corollary 13.7.2) for this partition reads $\det \mathbf{M} \le \det \mathbf{A} \det \mathbf{C} = 2 \cdot 4 = 8$, and indeed $4 \le 8$. The inequality is strict, as it must be, since $\mathbf{B} \neq \mathbf{O}$.

Exercise. B2: Hadamard bounds

(a) Bound $|\det \mathbf{B}|$ for $\mathbf{B} = \left(\begin{smallmatrix} 1 & 2 \\ 3 & 4 \end{smallmatrix}\right)$, and compare with the exact value.

(b) Let

$$\mathbf{H} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix}.$$

Compute $\mathbf{H}^\top \mathbf{H}$, and deduce $|\det \mathbf{H}|$ without expanding the determinant.

Solution.

(a) The columns are $(1, 3)$ and $(2, 4)$, with norms $\sqrt{10}$ and $\sqrt{20}$, so the bound is $\sqrt{200} = 10\sqrt{2} \approx 14.14$. The determinant is $4 - 6 = -2$, so $|\det \mathbf{B}| = 2$. The bound is true and very weak, because the columns are far from orthogonal.

(b) Every column of $\mathbf{H}$ has norm 2, and any two distinct columns are orthogonal: each of the six pairs agrees in two positions and disagrees in the other two, so the dot product is $1 + 1 - 1 - 1 = 0$. Hence $\mathbf{H}^\top \mathbf{H} = 4\mathbf{I}_4$. Taking determinants with Theorem 7.5.4 and Theorem 7.4.5, $(\det \mathbf{H})^2 = 4^4 = 256$, so $|\det \mathbf{H}| = 16$. This is exactly the Hadamard bound $\|\mathbf{h}_1\| \cdots \|\mathbf{h}_4\| = 2^4 = 16$, as the equality case of Theorem 13.7.3 (b) predicts for orthogonal columns.

13.7.3.3 C. Going deeper

Exercise. C1: How large can a determinant be?

(a) Let $\mathbf{B} \in M_n(F)$ have every entry of modulus at most 1. Prove that $|\det \mathbf{B}| \le n^{n/2}$.

(b) Show that the bound is attained for $n = 1$, $n = 2$, and $n = 4$.

(c) Let $\mathbf{B} \in M_3(\mathbb{R})$ have every entry equal to 1 or -1. Prove that $\det \mathbf{B}$ is a multiple of 4, and deduce that $|\det \mathbf{B}| \le 4$. Hence the bound of (a) is **not** attained for $n = 3$.

Hint for (c): normalize the first row and the first column, then subtract the first column from the other two.

Solution.

(a) Each column $\mathbf{b}_j$ has n entries of modulus ≤ 1, so $\|\mathbf{b}_j\|^2 = \sum_i |b_{ij}|^2 \le n$ and $\|\mathbf{b}_j\| \le \sqrt{n}$. By Theorem 13.7.3 (b), $|\det \mathbf{B}| \le (\sqrt{n})^n = n^{n/2}$.

(b) For $n = 1$, $\mathbf{B} = (1)$ gives $1 = 1^{1/2}$. For $n = 2$, $\mathbf{B} = \left(\begin{smallmatrix} 1 & 1 \\ 1 & -1 \end{smallmatrix}\right)$ has $|\det \mathbf{B}| = 2 = 2^{2/2}$. For $n = 4$, the matrix $\mathbf{H}$ of Exercise (B2: Hadamard bounds) (b) has $|\det \mathbf{H}| = 16 = 4^{4/2}$. In each case all entries have modulus 1 and the columns are orthogonal, which is exactly what is needed for equality both in $\|\mathbf{b}_j\| \le \sqrt{n}$ and in Theorem 13.7.3 (b).

(c) Multiplying a row or a column by -1 multiplies $\det \mathbf{B}$ by -1 (Theorem 7.5.1 (b) and (d)) and keeps all entries ± 1, so we may assume that the first row **and** the first column both consist of 1s: first scale the three columns to make the first row $(1, 1, 1)$, then scale rows 2 and 3 to make the first column $(1, 1, 1)$, which leaves the first row alone. Subtracting column 1 from columns 2 and 3 leaves the determinant unchanged (Theorem 7.5.1 (c) and (d)) and gives a matrix whose first row is $(1, 0, 0)$ and whose remaining entries are $c_{ij} = b_{ij} - 1$ for $i, j \in \{2, 3\}$; each lies in $\{0, -2\}$, since $b_{ij} \in \{1, -1\}$. Expanding along the first row (Theorem 7.6.1),

$$\det \mathbf{B} = c_{22}c_{33} - c_{23}c_{32}$$

with every $c_{ij} \in \{0, -2\}$. Each product lies in $\{0, 4\}$, so $\det \mathbf{B} \in \{-4, 0, 4\}$ and in particular $|\det \mathbf{B}| \le 4$. Since $3^{3/2} = 3\sqrt{3} \approx 5.196 > 4$, the bound of (a) is not attained. The value 4 does occur, for instance for $\left(\begin{smallmatrix} 1 & 1 & 1 \\ 1 & -1 & 1 \\ 1 & 1 & -1 \end{smallmatrix}\right)$.

13.8 Products Entry by Entry

Matrix multiplication is not the only way to multiply two matrices of the same size. There is a much simpler product — multiply the entries in matching positions — which is useless for composing linear maps and extremely useful everywhere else. This section introduces it and proves the one theorem that makes it interesting: it preserves positivity. A determinant inequality follows, and it contains Hadamard's inequality of Section 7 as a special case.

Throughout, $F = \mathbb{R}$ or $F = \mathbb{C}$.

13.8.1 The entrywise product

Two matrices of the same shape can be multiplied the way two lists of numbers are: position by position. The operation has no interpretation in terms of linear maps, which is one reason it looks like a curiosity at first sight, and it is written with a small circle to keep it away from ordinary multiplication.

The Hadamard product multiplies matrices the way a spreadsheet would.

Definition 13.8.1. Hadamard Product

Let $\mathbf{A}, \mathbf{B} \in M_{m\times n}(F)$. Their **Hadamard product**, or **entrywise product**, is the matrix $\mathbf{A} \circ \mathbf{B} \in M_{m\times n}(F)$ with

$$(\mathbf{A} \circ \mathbf{B})_{ij} = a_{ij}b_{ij} \qquad \text{for all } i, j.$$

Both factors must have **the same shape**, and the result has that shape too; there is no rule like "columns of the first match rows of the second". The operation is commutative and associative, and it is bilinear — $(\mathbf{A} + \mathbf{A}') \circ \mathbf{B} = \mathbf{A} \circ \mathbf{B} + \mathbf{A}' \circ \mathbf{B}$ and $(c\mathbf{A}) \circ \mathbf{B} = c(\mathbf{A} \circ \mathbf{B})$ — because each of these is a statement about numbers, checked one entry at a time. Its identity element is the all-ones matrix $\mathbf{J}$, not $\mathbf{I}$.

Examples.

- $\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \circ \begin{pmatrix} 5 & 0 \\ -1 & 2 \end{pmatrix} = \begin{pmatrix} 5 & 0 \\ -3 & 8 \end{pmatrix}$.
- $\mathbf{A} \circ \mathbf{I} = \operatorname{diag}(a_{11}, \ldots, a_{nn})$: multiplying entrywise by the identity **erases** everything off the diagonal, where ordinary multiplication by $\mathbf{I}$ changes nothing.
- $\mathbf{A} \circ \mathbf{J} = \mathbf{A}$, so $\mathbf{J}$ is the identity for $\circ$.
- For two diagonal matrices, $\mathbf{D} \circ \mathbf{D}' = \mathbf{D}\mathbf{D}'$: both products are $\operatorname{diag}(d_1d'_1, \ldots, d_nd'_n)$. For matrices in general the two have nothing to do with each other, though they can coincide by accident: $\left(\begin{smallmatrix} 1 & 1 \\ 0 & 0 \end{smallmatrix}\right)$ and $\left(\begin{smallmatrix} 1 & 0 \\ 0 & 0 \end{smallmatrix}\right)$ have equal Hadamard and matrix products without either being diagonal.
- For column vectors $\mathbf{u}, \mathbf{w} \in F^n$, regarded as $n \times 1$ matrices, $\mathbf{u} \circ \mathbf{w} = (u_1w_1, \ldots, u_nw_n)$.

Nonexample by minimal change. Take $\mathbf{A} = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$ and $\mathbf{B} = \mathbf{A}$. Then $\mathbf{AB} = \mathbf{I}$, but $\mathbf{A} \circ \mathbf{B} = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right) \circ \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right) = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$. The two products differ, and they differ in the most basic way: $\mathbf{A}$ is its own inverse and its own Hadamard square.

Warning. **$\mathbf{A} \circ \mathbf{B}$ is not $\mathbf{AB}$, and almost nothing carries over.** Two invertible matrices can have a singular entrywise product:

$$\begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix} \circ \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} = \begin{pmatrix} 2 & 2 \\ 2 & 2 \end{pmatrix},$$

where the two factors have determinants -3 and 3 and the product has determinant 0. There is no formula for $\det(\mathbf{A} \circ \mathbf{B})$ in terms of $\det \mathbf{A}$ and $\det \mathbf{B}$, no relation between the eigenvalues, and no rule $(\mathbf{A} \circ \mathbf{B})^{-1} = \mathbf{A}^{-1} \circ \mathbf{B}^{-1}$. The one structural property that does survive is the subject of this section, and it is a genuine surprise.

The name is the same one that Section 7 attached to a determinant inequality. That is a coincidence of a busy mathematician, not of the mathematics: Hadamard's inequality and the Hadamard product are unrelated statements, and this section will nonetheless prove one from the other.

13.8.2 The Schur product theorem

Positivity is a statement about the quadratic form $\mathbf{x}^*\mathbf{A}\mathbf{x}$, and the quadratic form of $\mathbf{A} \circ \mathbf{B}$ has no visible relation to those of $\mathbf{A}$ and $\mathbf{B}$. So the following theorem cannot be proved by staring at the form. What works is to break both matrices into the simplest positive pieces there are — rank-one matrices $\mathbf{u}\mathbf{u}^*$ — and to notice that the entrywise product of two such pieces is again one.

Lemma 13.8.1. Entrywise product of two rank-one positive matrices

Let $\mathbf{u}, \mathbf{w} \in F^n$. Then

$$(\mathbf{u}\mathbf{u}^*) \circ (\mathbf{w}\mathbf{w}^*) = (\mathbf{u} \circ \mathbf{w})(\mathbf{u} \circ \mathbf{w})^*.$$

Proof. Compare the (i, j) entries. The (i, j) entry of $\mathbf{u}\mathbf{u}^*$ is $u_i\overline{u_j}$ and that of $\mathbf{w}\mathbf{w}^*$ is $w_i\overline{w_j}$, so the left-hand side has (i, j) entry

$$u_i\overline{u_j}\, w_i\overline{w_j} = (u_i w_i)\,\overline{(u_j w_j)},$$

the rearrangement using commutativity of F and $\overline{u_j}\,\overline{w_j} = \overline{u_j w_j}$. The vector $\mathbf{u} \circ \mathbf{w}$ has i-th entry $u_i w_i$, so the right-hand side has (i, j) entry $(u_i w_i)\overline{(u_j w_j)}$ as well. The two matrices agree entry by entry.

■

That one line is the whole theorem; everything else is bookkeeping over the spectral decompositions.

Theorem 13.8.2. Schur Product Theorem

Let $\mathbf{A}, \mathbf{B} \in M_n(F)$.

(a) If $\mathbf{A} \succeq \mathbf{O}$ and $\mathbf{B} \succeq \mathbf{O}$, then $\mathbf{A} \circ \mathbf{B} \succeq \mathbf{O}$.
(b) If $\mathbf{A} \succ \mathbf{O}$ and $\mathbf{B} \succ \mathbf{O}$, then $\mathbf{A} \circ \mathbf{B} \succ \mathbf{O}$.

Idea. For (a): the spectral theorem writes $\mathbf{A} = \sum_i \lambda_i \mathbf{u}_i\mathbf{u}_i^*$ and $\mathbf{B} = \sum_j \mu_j \mathbf{w}_j\mathbf{w}_j^*$ with all $\lambda_i, \mu_j \geq 0$. Expand $\mathbf{A} \circ \mathbf{B}$ by bilinearity into n^2 terms; by Lemma 13.8.1 each term is a nonnegative multiple of a matrix $\mathbf{z}\mathbf{z}^*$, and those are positive semidefinite. A sum of positive semidefinite matrices is positive semidefinite, so we are done.

For (b), do not redo the argument with strict inequalities: the vectors $\mathbf{u}_i \circ \mathbf{w}_j$ need not span, and the bookkeeping becomes unpleasant. Instead peel a multiple of the identity off $\mathbf{B}$. If

$\varepsilon > 0$ is the smallest eigenvalue of $\mathbf{B}$, then $\mathbf{B} = (\mathbf{B} - \varepsilon\mathbf{I}) + \varepsilon\mathbf{I}$ with $\mathbf{B} - \varepsilon\mathbf{I} \succeq \mathbf{O}$, so

$$\mathbf{A} \circ \mathbf{B} = \mathbf{A} \circ (\mathbf{B} - \varepsilon\mathbf{I}) + \varepsilon\,(\mathbf{A} \circ \mathbf{I}),$$

and $\mathbf{A} \circ \mathbf{I}$ is the diagonal matrix of the a_{ii}, which are strictly positive. Part (a) handles the first term, and a positive definite matrix plus a positive semidefinite one is positive definite.

Proof.

(a) Both $\mathbf{A}$ and $\mathbf{B}$ are Hermitian, so each has an orthonormal basis of eigenvectors. Let $(\mathbf{u}_1, \dots, \mathbf{u}_n)$ be one for $\mathbf{A}$, with $\mathbf{A}\mathbf{u}_i = \lambda_i \mathbf{u}_i$, and let $\mathbf{U}$ be the unitary matrix with these columns, so that $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{U}^*$ with $\mathbf{D} = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$ by Corollary 12.4.3 over $\mathbb{C}$ and Corollary 12.5.2 over $\mathbb{R}$. The columns of $\mathbf{U}\mathbf{D}$ are $\lambda_i \mathbf{u}_i$ and the rows of $\mathbf{U}^*$ are $\mathbf{u}_i^*$, so Corollary 8.1.3 gives

$$\mathbf{A} = \sum_{i=1}^{n} \lambda_i\, \mathbf{u}_i \mathbf{u}_i^*,$$

which over $\mathbb{R}$ is Corollary 12.5.4. In the same way $\mathbf{B} = \sum_j \mu_j \mathbf{w}_j \mathbf{w}_j^*$. Every $\lambda_i \geq 0$ and every $\mu_j \geq 0$, by Theorem 13.1.4 (b).
The Hadamard product is bilinear, so expanding both sums,

$$\mathbf{A} \circ \mathbf{B} = \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \mu_j \left[(\mathbf{u}_i \mathbf{u}_i^*) \circ (\mathbf{w}_j \mathbf{w}_j^*) \right] = \sum_{i,j} \lambda_i \mu_j\, \mathbf{z}_{ij} \mathbf{z}_{ij}^*,$$

where $\mathbf{z}_{ij} = \mathbf{u}_i \circ \mathbf{w}_j$ and the second equality is Lemma 13.8.1. Each $\mathbf{z}_{ij}\mathbf{z}_{ij}^*$ is positive semidefinite by Example (Five Positive Matrices) (b), and each coefficient $\lambda_i \mu_j$ is a nonnegative real number. A nonnegative combination of positive semidefinite matrices is positive semidefinite (Exercise (C2: The cone of positive matrices) (a)): the combination is Hermitian, being a real combination of Hermitian matrices, and for every $\mathbf{x}$,

$$\left\langle \left(\textstyle\sum_{i,j} \lambda_i \mu_j \mathbf{z}_{ij} \mathbf{z}_{ij}^* \right) \mathbf{x}, \mathbf{x} \right\rangle = \textstyle\sum_{i,j} \lambda_i \mu_j \langle \mathbf{z}_{ij} \mathbf{z}_{ij}^* \mathbf{x}, \mathbf{x} \rangle \;\geq\; 0,$$

every term being a product of nonnegative reals. So $\mathbf{A} \circ \mathbf{B} \succeq \mathbf{O}$.

(b) Let $\varepsilon > 0$ be the smallest eigenvalue of $\mathbf{B}$, which is positive by Theorem 13.1.5 (b). The matrix $\mathbf{B} - \varepsilon\mathbf{I}$ is Hermitian with eigenvalues $\mu_j - \varepsilon \geq 0$, so $\mathbf{B} - \varepsilon\mathbf{I} \succeq \mathbf{O}$ by Theorem 13.1.4 (b). Since $\circ$ is bilinear and $\mathbf{A} \circ \mathbf{I} = \operatorname{diag}(a_{11}, \dots, a_{nn})$,

$$\mathbf{A} \circ \mathbf{B} = \mathbf{A} \circ (\mathbf{B} - \varepsilon\mathbf{I}) \;+\; \varepsilon\, \operatorname{diag}(a_{11}, \dots, a_{nn}).$$

The first term is positive semidefinite by (a). In the second, $a_{ii} = \langle \mathbf{A}\mathbf{e}_i, \mathbf{e}_i \rangle > 0$ because $\mathbf{A} \succ \mathbf{O}$ and $\mathbf{e}_i \neq \mathbf{0}$, so $\varepsilon \operatorname{diag}(a_{11}, \dots, a_{nn}) \succ \mathbf{O}$ by Example (Five Positive Matrices) (a). For $\mathbf{x} \neq \mathbf{0}$, the quadratic form of the sum is ≥ 0 plus > 0, hence > 0. So $\mathbf{A} \circ \mathbf{B} \succ \mathbf{O}$, and the theorem is proved.

■

The theorem is a genuine addition to the list of operations that preserve positivity. Sums and nonnegative multiples were easy (Exercise (C2: The cone of positive matrices)), congruence was easy (Proposition 13.1.2), and the ordinary product is not on the list at all, since $\mathbf{AB}$ need

not even be Hermitian. The entrywise product is on the list, and nothing in the definition of positivity suggests it should be.

Example. A 3×3 Entrywise Product

Let

$$\mathbf{A} = \begin{pmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{pmatrix}, \qquad \mathbf{B} = \begin{pmatrix} 1 & -1 & 1 \\ -1 & 2 & -1 \\ 1 & -1 & 3 \end{pmatrix}.$$

Check that $\mathbf{A} \succ \mathbf{O}$ and $\mathbf{B} \succ \mathbf{O}$, compute $\mathbf{A} \circ \mathbf{B}$, and verify directly that it is positive definite.

Solution. The leading principal minors of $\mathbf{A}$ are 2, $4 - 1 = 3$ and, expanding along the first row, $2(4 - 1) - 1(2 - 0) = 4$. All are positive, so $\mathbf{A} \succ \mathbf{O}$ by Sylvester's criterion (Theorem 13.1.5 (d)).

Those of $\mathbf{B}$ are 1, $2 - 1 = 1$, and $1(6 - 1) + 1(-3 + 1) + 1(1 - 2) = 5 - 2 - 1 = 2$. All positive, so $\mathbf{B} \succ \mathbf{O}$.

Entry by entry,

$$\mathbf{A} \circ \mathbf{B} = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 6 \end{pmatrix},$$

whose leading principal minors are 2, $8 - 1 = 7$, and $2(24 - 1) + 1(-6 - 0) = 46 - 6 = 40$. All positive, so $\mathbf{A} \circ \mathbf{B} \succ \mathbf{O}$, as Theorem 13.8.2 (b) promised. Note that the eigenvalues of $\mathbf{A} \circ \mathbf{B}$ have nothing to do with those of $\mathbf{A}$ and $\mathbf{B}$, and that $\det(\mathbf{A} \circ \mathbf{B}) = 40$ is much larger than $\det \mathbf{A} \det \mathbf{B} = 4 \cdot 2 = 8$. The next theorem explains part of that: it forces $\det(\mathbf{A} \circ \mathbf{B}) \geq \det \mathbf{A} \prod_i b_{ii} = 24$. Getting down to the product $\det \mathbf{A} \det \mathbf{B}$ needs Hadamard's inequality (Theorem 13.7.3) as well, which is Exercise (C2: The product of the determinants).

Quick check. Let $\mathbf{A} = \left(\begin{smallmatrix} 1 & 1 \\ 1 & 1 \end{smallmatrix}\right)$ and $\mathbf{B} = \left(\begin{smallmatrix} 2 & -1 \\ -1 & 2 \end{smallmatrix}\right)$. Compute $\mathbf{A} \circ \mathbf{B}$ and decide whether it is positive definite. Does this contradict Theorem 13.8.2 (b)?

Solution. $\mathbf{A} \circ \mathbf{B} = \left(\begin{smallmatrix} 2 & -1 \\ -1 & 2 \end{smallmatrix}\right)$, with leading principal minors 2 and 3, so it is positive definite. There is no contradiction: Theorem 13.8.2 (b) is an implication, not an equivalence. Here $\mathbf{A}$ is only positive semidefinite ($\mathbf{A} = \mathbf{1}\mathbf{1}^{\top}$, so $\det \mathbf{A} = 0$), and the theorem's hypothesis fails — but the conclusion happens to hold anyway. Part (a) is what applies, and it correctly predicts $\mathbf{A} \circ \mathbf{B} \succeq \mathbf{O}$.

13.8.3 Oppenheim's inequality

Hadamard's inequality of Section 7 compared $\det \mathbf{A}$ with the product of the diagonal entries of $\mathbf{A}$. Oppenheim's inequality does the same for a Hadamard product, and it contains Hadamard's as the special case $\mathbf{B} = \mathbf{I}$.

Theorem 13.8.3. Oppenheim's Inequality

Let $\mathbf{A}, \mathbf{B} \in M_n(F)$ with $\mathbf{A} \succeq \mathbf{O}$ and $\mathbf{B} \succeq \mathbf{O}$. Then

$$\det(\mathbf{A} \circ \mathbf{B}) \;\geq\; \det \mathbf{A} \cdot b_{11} b_{22} \cdots b_{nn}.$$

Idea. Induct on n, taking both matrices positive definite first. Split off the $(1,1)$ entry of each: the Schur complements $\mathbf{A}/\alpha$ and $\mathbf{B}/\beta$ are positive definite of size $n-1$, and a short computation, whose only real step is Lemma 13.8.1 again, shows that the Schur complement of the corner in $\mathbf{A} \circ \mathbf{B}$ is

$$(\mathbf{A} \circ \mathbf{B})/(\alpha\beta) = (\mathbf{A}/\alpha) \circ \mathbf{B}' + \tfrac{1}{\alpha}\,(\mathbf{a}\mathbf{a}^*) \circ (\mathbf{B}/\beta),$$

a positive definite matrix plus a positive semidefinite one. So its determinant is at least that of the first term, by Lemma 13.7.1, and the inductive hypothesis handles that. The semidefinite case follows by replacing $\mathbf{A}$ and $\mathbf{B}$ with $\mathbf{A} + t\mathbf{I}$ and $\mathbf{B} + t\mathbf{I}$ and letting t shrink to 0.

Proof. Case 1: $\mathbf{A} \succ \mathbf{O}$ *and* $\mathbf{B} \succ \mathbf{O}$. Induct on n. For $n = 1$ both sides are $a_{11} b_{11}$.

Let $n \geq 2$, and split both matrices with a $1 \times (n-1)$ partition:

$$\mathbf{A} = \begin{pmatrix} \alpha & \mathbf{a}^* \\ \mathbf{a} & \mathbf{A}' \end{pmatrix}, \qquad \mathbf{B} = \begin{pmatrix} \beta & \mathbf{b}^* \\ \mathbf{b} & \mathbf{B}' \end{pmatrix},$$

with $\alpha = a_{11} > 0$, $\beta = b_{11} > 0$, and $\mathbf{a}, \mathbf{b} \in F^{n-1}$. The Schur complements are

$$\mathbf{A}/\alpha = \mathbf{A}' - \tfrac{1}{\alpha}\,\mathbf{a}\mathbf{a}^*, \qquad \mathbf{B}/\beta = \mathbf{B}' - \tfrac{1}{\beta}\,\mathbf{b}\mathbf{b}^*,$$

and both are positive definite by Theorem 13.6.1 (b). Since conjugation acts entrywise, $\mathbf{a}^* \circ \mathbf{b}^* = (\mathbf{a} \circ \mathbf{b})^*$, so

$$\mathbf{A} \circ \mathbf{B} = \begin{pmatrix} \alpha\beta & (\mathbf{a} \circ \mathbf{b})^* \\ \mathbf{a} \circ \mathbf{b} & \mathbf{A}' \circ \mathbf{B}' \end{pmatrix},$$

and $\alpha\beta > 0$, so this matrix too has a Schur complement at its corner:

$$(\mathbf{A} \circ \mathbf{B})/(\alpha\beta) = \mathbf{A}' \circ \mathbf{B}' - \tfrac{1}{\alpha\beta}\,(\mathbf{a} \circ \mathbf{b})(\mathbf{a} \circ \mathbf{b})^*.$$

By Lemma 13.8.1 the subtracted term is $\frac{1}{\alpha\beta}(\mathbf{a}\mathbf{a}^*)\circ(\mathbf{b}\mathbf{b}^*)$. Substituting $\mathbf{A}' = \mathbf{A}/\alpha + \frac{1}{\alpha}\mathbf{a}\mathbf{a}^*$ and expanding by bilinearity,

$$\begin{aligned}\mathbf{A}'\circ\mathbf{B}' &= (\mathbf{A}/\alpha)\circ\mathbf{B}' + \tfrac{1}{\alpha}\,(\mathbf{a}\mathbf{a}^*)\circ\mathbf{B}' \\ &= (\mathbf{A}/\alpha)\circ\mathbf{B}' + \tfrac{1}{\alpha}\,(\mathbf{a}\mathbf{a}^*)\circ\left(\mathbf{B}/\beta + \tfrac{1}{\beta}\mathbf{b}\mathbf{b}^*\right),\end{aligned}$$

and the last term of this is exactly $\frac{1}{\alpha\beta}(\mathbf{a}\mathbf{a}^*)\circ(\mathbf{b}\mathbf{b}^*)$. It therefore cancels, leaving

$$(\mathbf{A}\circ\mathbf{B})/(\alpha\beta) = \underbrace{(\mathbf{A}/\alpha)\circ\mathbf{B}'}_{\succ\,\mathbf{O}} + \underbrace{\tfrac{1}{\alpha}\,(\mathbf{a}\mathbf{a}^*)\circ(\mathbf{B}/\beta)}_{\succeq\,\mathbf{O}}. \tag{$*$}$$

The first term is positive definite by Theorem 13.8.2 (b), since $\mathbf{A}/\alpha \succ \mathbf{O}$ and $\mathbf{B}' \succ \mathbf{O}$ (the latter by Theorem 13.6.1 (a) applied to $\mathbf{B}$). The second is positive semidefinite by Theorem 13.8.2 (a), since $\mathbf{a}\mathbf{a}^* \succeq \mathbf{O}$ by Example (Five Positive Matrices) (b) and $\mathbf{B}/\beta \succ \mathbf{O}$, and $1/\alpha > 0$.

Now compute. By Theorem 8.3.2 applied to $\mathbf{A}\circ\mathbf{B}$, then $(*)$ with Lemma 13.7.1, then the inductive hypothesis applied to the pair $\mathbf{A}/\alpha$ and $\mathbf{B}'$:

$$\begin{aligned}\det(\mathbf{A}\circ\mathbf{B}) &= \alpha\beta\cdot\det\big[(\mathbf{A}\circ\mathbf{B})/(\alpha\beta)\big] \\ &\geq \alpha\beta\cdot\det\big[(\mathbf{A}/\alpha)\circ\mathbf{B}'\big] \\ &\geq \alpha\beta\cdot\det(\mathbf{A}/\alpha)\cdot b_{22}\cdots b_{nn}.\end{aligned}$$

By Theorem 8.3.2 again, now for $\mathbf{A}$, we have $\det(\mathbf{A}/\alpha) = \det\mathbf{A}/\alpha$. Substituting and canceling α,

$$\det(\mathbf{A}\circ\mathbf{B}) \;\geq\; \beta\,\det\mathbf{A}\cdot b_{22}\cdots b_{nn} = \det\mathbf{A}\cdot b_{11}b_{22}\cdots b_{nn},$$

which completes the induction.

Case 2: $\mathbf{A}\succeq\mathbf{O}$ *and* $\mathbf{B}\succeq\mathbf{O}$. For real $t>0$, the matrices $\mathbf{A}+t\mathbf{I}$ and $\mathbf{B}+t\mathbf{I}$ are positive definite, as in the proof of Corollary 13.7.2. Case 1 gives

$$\det\big[(\mathbf{A}+t\mathbf{I})\circ(\mathbf{B}+t\mathbf{I})\big] \;\geq\; \det(\mathbf{A}+t\mathbf{I})\cdot\prod_{i=1}^{n}(b_{ii}+t).$$

Every entry of $(\mathbf{A}+t\mathbf{I})\circ(\mathbf{B}+t\mathbf{I})$ is a polynomial in t, so both sides are polynomial, hence continuous, functions of t (Theorem 7.4.3). Letting t decrease to 0 preserves the inequality and gives the statement. This proves the theorem.

■

Remark. Taking $\mathbf{B}=\mathbf{I}$ recovers Hadamard's inequality. Indeed $\mathbf{A}\circ\mathbf{I} = \operatorname{diag}(a_{11},\dots,a_{nn})$ and $\prod_i b_{ii} = 1$, so the conclusion reads $a_{11}\cdots a_{nn}\geq\det\mathbf{A}$, which is Theorem 13.7.3 (a). Swapping the roles of $\mathbf{A}$ and $\mathbf{B}$ gives the companion bound $\det(\mathbf{A}\circ\mathbf{B})\geq\det\mathbf{B}\cdot\prod_i a_{ii}$, and combining either one with Hadamard's inequality gives $\det(\mathbf{A}\circ\mathbf{B})\geq\det\mathbf{A}\det\mathbf{B}$; that last

consequence is Exercise (C2: The product of the determinants).

13.8.4 Where this gets used

The natural home of the Hadamard product is data. A covariance matrix is a Gram matrix (Theorem 11.2.5), hence positive semidefinite, and rescaling each variable to unit variance turns it into a **correlation matrix**: positive semidefinite with every diagonal entry equal to 1. The Schur product theorem says that the entrywise product of two correlation matrices is another one — the diagonal entries stay $1 \cdot 1 = 1$ — so one can damp the correlations of one model by those of another and still have a legitimate covariance structure. Nothing in the definition of "positive semidefinite" makes that obvious, and it is the reason the theorem is quoted far outside linear algebra.

Warning. **Entrywise products preserve positivity; entrywise inverses do not.** Let $\mathbf{A} = \left(\begin{smallmatrix} 2 & 1 \\ 1 & 2 \end{smallmatrix}\right)$, which is positive definite, with leading principal minors 2 and 3. Replacing each entry by its reciprocal gives $\left(\begin{smallmatrix} 1/2 & 1 \\ 1 & 1/2 \end{smallmatrix}\right)$, whose determinant is $\frac{1}{4} - 1 = -\frac{3}{4} < 0$: not positive semidefinite. The entrywise reciprocal is not the Hadamard inverse of anything useful, and the matrix inverse $\mathbf{A}^{-1}$, which **is** positive definite (Exercise (B3: The inverse of a positive definite matrix)), has nothing to do with it.

13.8.5 Exercises

13.8.5.1 A. Check your understanding

Exercise. A1

(a) Define $\mathbf{A} \circ \mathbf{B}$, and say what shapes the two factors must have.

(b) What is $\mathbf{A} \circ \mathbf{I}$? What is $\mathbf{A} \circ \mathbf{J}$?

(c) State the Schur product theorem.

(d) True or false: if $\mathbf{A} \succ \mathbf{O}$, then the matrix with entries $1/a_{ij}$ (assuming no entry is 0) is positive definite. Justify your answer.

(e) What does Oppenheim's inequality say when $\mathbf{B} = \mathbf{I}$?

Solution.

(a) For $\mathbf{A}, \mathbf{B} \in M_{m\times n}(F)$ of the **same** shape, $(\mathbf{A} \circ \mathbf{B})_{ij} = a_{ij}b_{ij}$, and $\mathbf{A} \circ \mathbf{B}$ has that same shape (Definition 13.8.1).

(b) $\mathbf{A} \circ \mathbf{I} = \operatorname{diag}(a_{11}, \dots, a_{nn})$, and $\mathbf{A} \circ \mathbf{J} = \mathbf{A}$, since $\mathbf{J}$ has every entry 1.

(c) If $\mathbf{A} \succeq \mathbf{O}$ and $\mathbf{B} \succeq \mathbf{O}$, then $\mathbf{A} \circ \mathbf{B} \succeq \mathbf{O}$; if both are positive definite, so is $\mathbf{A} \circ \mathbf{B}$ (Theorem 13.8.2).

(d) False. For $\mathbf{A} = \left(\begin{smallmatrix} 2 & 1 \\ 1 & 2 \end{smallmatrix}\right) \succ \mathbf{O}$, the entrywise reciprocal is $\left(\begin{smallmatrix} 1/2 & 1 \\ 1 & 1/2 \end{smallmatrix}\right)$, with determinant $-3/4 < 0$.

(e) By Theorem 13.8.3 it says $\det(\mathbf{A}\circ\mathbf{I}) \geq \det \mathbf{A} \cdot 1$, that is $a_{11} \cdots a_{nn} \geq \det \mathbf{A}$: Hadamard's inequality, Theorem 13.7.3 (a).

13.8.5.2 B. Practice

Exercise. B1: Compute and decide

For each pair below, compute $\mathbf{A} \circ \mathbf{B}$ and determine whether it is positive definite, positive semidefinite but not definite, or neither. Say in each case whether Theorem 13.8.2 applies.

(a) $\mathbf{A} = \left(\begin{smallmatrix} 1 & 1 \\ 1 & 1 \end{smallmatrix}\right)$, $\mathbf{B} = \left(\begin{smallmatrix} 1 & 1 \\ 1 & 1 \end{smallmatrix}\right)$.

(b) $\mathbf{A} = \left(\begin{smallmatrix} 1 & 1 \\ 1 & 1 \end{smallmatrix}\right)$, $\mathbf{B} = \left(\begin{smallmatrix} 1 & 2 \\ 2 & 1 \end{smallmatrix}\right)$.

(c) $\mathbf{A} = \left(\begin{smallmatrix} 2 & 1 \\ 1 & 2 \end{smallmatrix}\right)$, $\mathbf{B} = \begin{pmatrix} 3 & -1 \\ -1 & 1 \end{pmatrix}$.

Solution.

(a) $\mathbf{A} \circ \mathbf{B} = \left(\begin{smallmatrix} 1 & 1 \\ 1 & 1 \end{smallmatrix}\right)$. Both factors are $\mathbf{1}\mathbf{1}^{\top} \succeq \mathbf{O}$ (Example (Five Positive Matrices) (b)), so Theorem 13.8.2 (a) applies and predicts $\succeq \mathbf{O}$. The product has determinant 0, so it is positive semidefinite but not definite.

(b) $\mathbf{A} \circ \mathbf{B} = \left(\begin{smallmatrix} 1 & 2 \\ 2 & 1 \end{smallmatrix}\right)$, whose determinant is $1 - 4 = -3 < 0$: neither. The theorem does not apply, since $\mathbf{B}$ is not positive semidefinite (its determinant is -3). This is what the hypothesis is for.

(c) $\mathbf{A} \succ \mathbf{O}$ (minors 2, 3) and $\mathbf{B} \succ \mathbf{O}$ (minors 3, 2), so Theorem 13.8.2 (b) applies. Indeed $\mathbf{A} \circ \mathbf{B} = \begin{pmatrix} 6 & -1 \\ -1 & 2 \end{pmatrix}$, with minors 6 and 11, is positive definite. Theorem 13.8.3 also checks out: $11 \geq \det \mathbf{A} \cdot b_{11} b_{22} = 3 \cdot 3 = 9$.

Exercise. B2: Entrywise squares

Let $\mathbf{A} = \left(\begin{smallmatrix} 1 & 1 & 1 \\ 1 & 2 & 2 \\ 1 & 2 & 3 \end{smallmatrix}\right)$.

(a) Show that $\mathbf{A} \succ \mathbf{O}$.

(b) Compute $\mathbf{A} \circ \mathbf{A}$, the matrix of squared entries, and verify that it is positive definite.

(c) Explain why (b) needed no computation.

Solution.

(a) The leading principal minors are 1, $2 - 1 = 1$ and, expanding along the first row, $1(6 - 4) - 1(3 - 2) + 1(2 - 2) = 1$. All positive, so $\mathbf{A} \succ \mathbf{O}$ by Theorem 13.1.5 (d).

(b) $\mathbf{A} \circ \mathbf{A} = \left(\begin{smallmatrix} 1 & 1 & 1 \\ 1 & 4 & 4 \\ 1 & 4 & 9 \end{smallmatrix}\right)$, with leading principal minors 1, $4 - 1 = 3$, and $1(36 - 16) - 1(9 - 4) + 1(4 - 4) = 20 - 5 = 15$. All positive, so it is positive definite.

(c) Theorem 13.8.2 (b) with $\mathbf{B} = \mathbf{A}$ gives $\mathbf{A} \circ \mathbf{A} \succ \mathbf{O}$ immediately.

13.8.5.3 C. Going deeper

Exercise. C1: Polynomials applied entrywise

Let $\mathbf{A} \in M_n(F)$ with $\mathbf{A} \succeq \mathbf{O}$.

(a) Prove that the entrywise power $\mathbf{A}^{\circ m}$, with (i, j) entry a_{ij}^m, is positive semidefinite for

every integer $m \geq 1$.

(b) Let $p(x) = c_0 + c_1 x + \cdots + c_d x^d$ have **nonnegative real** coefficients. Prove that the matrix $p^{\circ}(\mathbf{A})$ with (i,j) entry $p(a_{ij})$ is positive semidefinite.

(c) Give an example of $\mathbf{A} \succeq \mathbf{O}$ and a polynomial p with a negative coefficient for which $p^{\circ}(\mathbf{A})$ is not positive semidefinite.

Solution.

(a) Induct on m. For $m = 1$ the claim is the hypothesis. If $\mathbf{A}^{\circ m} \succeq \mathbf{O}$, then $\mathbf{A}^{\circ(m+1)} = \mathbf{A}^{\circ m} \circ \mathbf{A} \succeq \mathbf{O}$ by Theorem 13.8.2 (a).

(b) By definition of the entries, $p^{\circ}(\mathbf{A}) = c_0\mathbf{J} + c_1\mathbf{A} + c_2\mathbf{A}^{\circ 2} + \cdots + c_d\mathbf{A}^{\circ d}$, since each side has (i,j) entry $\sum_k c_k a_{ij}^k$ (with $a_{ij}^0 = 1$, which is the (i,j) entry of $\mathbf{J}$). Now $\mathbf{J} = \mathbf{1}\mathbf{1}^{\top} \succeq \mathbf{O}$ by Example (Five Positive Matrices) (b), each $\mathbf{A}^{\circ k} \succeq \mathbf{O}$ by (a), and each $c_k \geq 0$. A nonnegative combination of positive semidefinite matrices is positive semidefinite (Exercise (C2: The cone of positive matrices) (a)).

(c) Take $\mathbf{A} = \mathbf{J} \in M_2(\mathbb{R})$, which is positive semidefinite, and $p(x) = x - 2$. Then $p^{\circ}(\mathbf{A})$ has every entry -1, that is $p^{\circ}(\mathbf{A}) = -\mathbf{J}$, and $\langle -\mathbf{J}\mathbf{e}_1, \mathbf{e}_1 \rangle = -1 < 0$. So the hypothesis on the coefficients cannot be dropped.

Exercise. C2: The product of the determinants

Let $\mathbf{A}, \mathbf{B} \in M_n(F)$ with $\mathbf{A} \succeq \mathbf{O}$ and $\mathbf{B} \succeq \mathbf{O}$. Prove that

$$\det(\mathbf{A} \circ \mathbf{B}) \;\geq\; \det \mathbf{A} \cdot \det \mathbf{B}.$$

Is the inequality ever an equality with both matrices positive definite?

Solution. By Theorem 13.8.3, $\det(\mathbf{A} \circ \mathbf{B}) \geq \det \mathbf{A} \cdot \prod_i b_{ii}$. By Theorem 13.7.3 (a) applied to $\mathbf{B} \succeq \mathbf{O}$, $\prod_i b_{ii} \geq \det \mathbf{B}$. Both $\det \mathbf{A} \geq 0$ and $\prod_i b_{ii} \geq 0$ (the diagonal entries of a positive semidefinite matrix are ≥ 0 by Exercise (C1: Zero trace forces zero) (a)), so multiplying the second inequality by $\det \mathbf{A} \geq 0$ and chaining,

$$\det(\mathbf{A} \circ \mathbf{B}) \;\geq\; \det \mathbf{A} \prod_i b_{ii} \;\geq\; \det \mathbf{A} \det \mathbf{B}.$$

Yes, equality can occur: take $\mathbf{A} = \mathbf{B} = \mathbf{I}_n$, so that $\mathbf{A} \circ \mathbf{B} = \mathbf{I}_n$ and all three quantities are 1. More generally, take $\mathbf{A} \succ \mathbf{O}$ diagonal and $\mathbf{B} = \mathbf{I}_n$.

Exercise. C3: The rank of an entrywise product

Let $\mathbf{A}, \mathbf{B} \in M_n(F)$ with $\mathbf{A} \succeq \mathbf{O}$ and $\mathbf{B} \succeq \mathbf{O}$, of ranks r and s. Prove that

$$\operatorname{rank}(\mathbf{A} \circ \mathbf{B}) \leq rs.$$

Hint: how many terms in the proof of Theorem 13.8.2 (a) are actually nonzero?

Solution. Write $\mathbf{A} = \sum_i \lambda_i \mathbf{u}_i \mathbf{u}_i^*$ and $\mathbf{B} = \sum_j \mu_j \mathbf{w}_j \mathbf{w}_j^*$ as in the proof of Theorem 13.8.2 (a). The number of nonzero eigenvalues of $\mathbf{A}$ is $\operatorname{rank} \mathbf{A} = r$: diagonalizing, $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{U}^*$ with $\mathbf{U}$ invertible, so $\operatorname{rank} \mathbf{A} = \operatorname{rank} \mathbf{D}$ by Theorem 3.5.9, and $\operatorname{rank} \mathbf{D}$ counts the nonzero λ_i. Likewise exactly s of the μ_j are nonzero. Dropping the vanishing terms,

$$\mathbf{A} \circ \mathbf{B} = \sum_{\lambda_i \neq 0} \sum_{\mu_j \neq 0} \lambda_i \mu_j \, \mathbf{z}_{ij} \mathbf{z}_{ij}^*, \qquad \mathbf{z}_{ij} = \mathbf{u}_i \circ \mathbf{w}_j,$$

a sum of rs matrices, each of rank at most 1, since every column of $\mathbf{z}_{ij}\mathbf{z}_{ij}^*$ lies in $\operatorname{span}(\mathbf{z}_{ij})$. The rank of a sum of two matrices is at most the sum of their ranks (Exercise (C1: Rank of a sum)); applying that repeatedly to the rs summands gives $\operatorname{rank}(\mathbf{A} \circ \mathbf{B}) \leq rs$.

13.9 The Singular Value Decomposition

Chapter 12 answered one question completely: an operator has an orthonormal basis of eigenvectors exactly when it is normal. That answer is useless for a 3×5 matrix, which has no eigenvalues at all — the equation $\mathbf{A}\mathbf{x} = \lambda \mathbf{x}$ is not even type-correct, because $\mathbf{A}\mathbf{x}$ and $\mathbf{x}$ live in different spaces. This section shows that every matrix, of every shape, still admits an orthonormal-to-orthogonal description, and that the numbers in it are the eigenvalues of a positive semidefinite matrix we already understand. It is the most-used factorization in applied linear algebra, and the whole of it comes from applying the spectral theorem to $\mathbf{A}^*\mathbf{A}$.

Throughout, F is $\mathbb{R}$ or $\mathbb{C}$, $\mathbf{A}^*$ is the conjugate transpose (the transpose when $F = \mathbb{R}$), and *unitary* means unitary over $\mathbb{C}$ and orthogonal over $\mathbb{R}$, as in Definition 11.7.2. Both F^m and F^n carry their standard inner products.

13.9.1 Where to look when there are no eigenvalues

Let $\mathbf{A} \in M_{m \times n}(F)$. There is exactly one square matrix that can be built from $\mathbf{A}$ at no cost, and Chapter 11 has already used it twice: the $n \times n$ matrix $\mathbf{A}^*\mathbf{A}$ of the normal equations. It is self-adjoint, since $(\mathbf{A}^*\mathbf{A})^* = \mathbf{A}^*\mathbf{A}^{**} = \mathbf{A}^*\mathbf{A}$, and it is positive semidefinite for the reason that makes every argument in this chapter work:

$$\langle \mathbf{A}^*\mathbf{A}\mathbf{x}, \mathbf{x} \rangle = \langle \mathbf{A}\mathbf{x}, \mathbf{A}\mathbf{x} \rangle = \|\mathbf{A}\mathbf{x}\|^2 \geq 0. \tag{13.9.1}$$

So $\mathbf{A}^*\mathbf{A} \succeq \mathbf{O}$, and Theorem 13.1.4 says its eigenvalues are nonnegative real numbers, while the spectral theorem supplies an orthonormal basis of F^n consisting of its eigenvectors. Those eigenvalues are attached to $\mathbf{A}$, they exist for every shape of $\mathbf{A}$, and Eq. 13.9.1 says what they measure: how much $\mathbf{A}$ stretches.

A rectangular matrix has no eigenvalues of its own, but $\mathbf{A}^*\mathbf{A}$ does, and their square roots are the stretching factors of $\mathbf{A}$.

Definition 13.9.1. Singular Values

Let $\mathbf{A} \in M_{m\times n}(F)$ and put $p = \min(m, n)$. Let

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$$

be the eigenvalues of $\mathbf{A}^*\mathbf{A} \in M_n(F)$, listed **with multiplicity** in decreasing order. The **singular values** of $\mathbf{A}$ are

$$\sigma_i(\mathbf{A}) := \sqrt{\lambda_i} \qquad (i = 1, \ldots, p),$$

so that $\sigma_1(\mathbf{A}) \geq \sigma_2(\mathbf{A}) \geq \cdots \geq \sigma_p(\mathbf{A}) \geq 0$. We write σ_i when $\mathbf{A}$ is clear.

Clause by clause. The list has exactly $p = \min(m, n)$ entries, not n: when $n > m$ the matrix $\mathbf{A}^*\mathbf{A}$ has rank $\operatorname{rank}\mathbf{A} \leq m < n$ by Corollary 11.6.6 (b), so $\lambda_{m+1} = \cdots = \lambda_n = 0$ and the discarded eigenvalues are all zero. The square root is the ordinary nonnegative real square root, available because every $\lambda_i \geq 0$. The ordering is **decreasing**, matching the book's convention $\lambda_1(\mathbf{A}) \geq \cdots \geq \lambda_n(\mathbf{A})$ for the eigenvalues of a Hermitian matrix, and it is part of the definition: "the second singular value" means a definite number. Finally, the singular values are attached to $\mathbf{A}$ itself, with no choices made, because the eigenvalue list of $\mathbf{A}^*\mathbf{A}$ is.

Four examples, simplest first.

1. **The identity.** $\mathbf{A} = \mathbf{I}_n$ gives $\mathbf{A}^*\mathbf{A} = \mathbf{I}_n$, so every $\sigma_i = 1$. More generally a unitary $\mathbf{U} \in M_n(F)$ has $\mathbf{U}^*\mathbf{U} = \mathbf{I}$, so all its singular values are 1. A unitary matrix stretches nothing in any direction, and the list of singular values says exactly that.
2. **A rectangular diagonal matrix.** If $\mathbf{A} \in M_{m\times n}(F)$ has $a_{ii} = d_i$ for $i \leq p$ and all other entries 0, then $\mathbf{A}^*\mathbf{A} = \operatorname{diag}(|d_1|^2, \ldots)$ and the singular values are $|d_1|, \ldots, |d_p|$ **sorted downwards**. The absolute value matters: the singular values of $\operatorname{diag}(1, -2)$ are $2, 1$.
3. **A positive semidefinite matrix.** If $\mathbf{A} \succeq \mathbf{O}$ then $\mathbf{A}^*\mathbf{A} = \mathbf{A}^2$, whose eigenvalues are $\lambda_i(\mathbf{A})^2$ with $\lambda_i(\mathbf{A}) \geq 0$, so $\sigma_i(\mathbf{A}) = \lambda_i(\mathbf{A})$. For the matrices this chapter started with, singular values and eigenvalues coincide.
4. **A single row.** For $\mathbf{A} = \mathbf{a}^* \in M_{1\times n}(F)$ with $\mathbf{a} \neq \mathbf{0}$, $p = 1$ and $\mathbf{A}^*\mathbf{A} = \mathbf{a}\mathbf{a}^*$, whose nonzero eigenvalue is $\|\mathbf{a}\|^2$. So $\sigma_1 = \|\mathbf{a}\|$. The degenerate case $\mathbf{A} = \mathbf{O}$ has every $\sigma_i = 0$, and it matters: it is the only matrix with $\sigma_1 = 0$, since $\sigma_1 = 0$ forces $\mathbf{A}^*\mathbf{A} = \mathbf{O}$ and then $\|\mathbf{A}\mathbf{x}\|^2 = 0$ for every $\mathbf{x}$ by Eq. 13.9.1.

Now the nonexample, by minimal change. Replace "square roots of the eigenvalues of $\mathbf{A}^*\mathbf{A}$" by "absolute values of the eigenvalues of $\mathbf{A}$". For $\mathbf{A} \succeq \mathbf{O}$, and more generally for any normal $\mathbf{A}$, the two agree (Exercise (C3: Singular values equal to eigenvalue moduli)). But the second recipe is undefined for a nonsquare matrix, and for

$$\mathbf{N} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$$

it returns $0, 0$, while $\mathbf{N}^*\mathbf{N} = \operatorname{diag}(0, 1)$ gives the singular values $1, 0$. The clause that fails is the one that does all the work: $\mathbf{A}^*\mathbf{A}$ is positive semidefinite for **every** $\mathbf{A}$, whereas $\mathbf{A}$ itself may have no useful spectrum.

Warning. **Singular values are not eigenvalues.** The matrix $\mathbf{N}$ above has both eigenvalues equal to 0 and largest singular value 1: it annihilates $\mathbf{e}_1$ but sends $\mathbf{e}_2$ to a unit vector, so calling it "small" because its spectrum is $\{0\}$ is wrong. Singular values are always real and nonnegative; eigenvalues need not be either. The two lists agree, after taking absolute

values, exactly for normal matrices — this is Exercise (C3: Singular values equal to eigenvalue moduli), and Chapter 12's Schur inequality is what makes the converse true.

Quick check. What are the singular values of an orthogonal projection $\mathbf{P} \in M_n(F)$ onto an r-dimensional subspace?

Solution. $\mathbf{P}^*\mathbf{P} = \mathbf{P}^2 = \mathbf{P}$, because an orthogonal projection is self-adjoint and idempotent. So the eigenvalues of $\mathbf{P}^*\mathbf{P}$ are those of $\mathbf{P}$, namely 1 with multiplicity r and 0 with multiplicity $n-r$. Sorted downwards, the singular values are $\sigma_1 = \cdots = \sigma_r = 1$ and $\sigma_{r+1} = \cdots = \sigma_n = 0$. This is also Example 3 above, since $\mathbf{P} \succeq \mathbf{O}$.

13.9.2 The theorem

Here is the payoff. The eigenvectors of $\mathbf{A}^*\mathbf{A}$ are an orthonormal basis of the domain; the claim is that $\mathbf{A}$ carries them to an **orthogonal** list in the codomain, whose lengths are the singular values.

Theorem 13.9.1. Singular Value Decomposition

Let $\mathbf{A} \in M_{m\times n}(F)$, let $p = \min(m, n)$, and let $\sigma_1 \geq \cdots \geq \sigma_p \geq 0$ be its singular values. Then there are unitary matrices $\mathbf{U} \in M_m(F)$ and $\mathbf{V} \in M_n(F)$ such that

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*,$$

where $\mathbf{\Sigma} \in M_{m\times n}(\mathbb{R})$ has $(\mathbf{\Sigma})_{ii} = \sigma_i$ for $i = 1, \ldots, p$ and every other entry 0. Moreover the number of nonzero σ_i equals $\operatorname{rank} \mathbf{A}$.

Idea. Picture the 2×2 real case. A matrix carries the unit circle to an ellipse, and an ellipse has two perpendicular axes; the theorem says that the two unit vectors sent to the ends of those axes are themselves perpendicular. That is exactly what Chapter 12's principal axis theorem (Theorem 12.5.5) said about the level curves of a quadratic form, and the link is that $\|\mathbf{A}\mathbf{x}\|^2 = \mathbf{x}^*(\mathbf{A}^*\mathbf{A})\mathbf{x}$ **is** a quadratic form, with matrix $\mathbf{A}^*\mathbf{A}$. So the plan is: ① diagonalize $\mathbf{A}^*\mathbf{A}$ by the spectral theorem, getting an orthonormal basis $\mathbf{v}_1, \ldots, \mathbf{v}_n$ with $\mathbf{A}^*\mathbf{A}\mathbf{v}_i = \sigma_i^2\mathbf{v}_i$; ② compute $\langle \mathbf{A}\mathbf{v}_i, \mathbf{A}\mathbf{v}_j \rangle$ and watch the eigenvalue equation force it to be $\sigma_i^2\delta_{ij}$, so the images are already orthogonal; ③ normalize the nonzero ones to unit vectors $\mathbf{u}_i$, and, if there are fewer than m of them, pad the list out to an orthonormal basis of F^m; ④ read $\mathbf{A}\mathbf{V} = \mathbf{U}\mathbf{\Sigma}$ off column by column.

Proof. **Step 1: diagonalize $\mathbf{A}^*\mathbf{A}$.** The matrix $\mathbf{A}^*\mathbf{A} \in M_n(F)$ is self-adjoint, and Eq. 13.9.1 makes it positive semidefinite in the sense of Definition 13.1.1. By the spectral theorem in matrix form — Corollary 12.5.2 when $F = \mathbb{R}$, Corollary 12.4.3 when $F = \mathbb{C}$, each of which allows the eigenvalues to be listed in a prescribed order — there is a unitary $\mathbf{V} \in M_n(F)$

with columns $\mathbf{v}_1, \dots, \mathbf{v}_n$ such that

$$\mathbf{A}^*\mathbf{A}\,\mathbf{v}_i = \lambda_i \mathbf{v}_i, \qquad \lambda_1 \geq \cdots \geq \lambda_n \geq 0,$$

the eigenvalues being nonnegative by Theorem 13.1.4. Set $\sigma_i = \sqrt{\lambda_i}$ for all $i \leq n$; for $i \leq p$ these are the singular values of Definition 13.9.1.

Step 2: the images are orthogonal. For all i, j, using $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{y}^*\mathbf{x}$ and the eigenvalue equation,

$$\langle \mathbf{A}\mathbf{v}_i, \mathbf{A}\mathbf{v}_j \rangle = \langle \mathbf{A}^*\mathbf{A}\mathbf{v}_i, \mathbf{v}_j \rangle = \lambda_i \langle \mathbf{v}_i, \mathbf{v}_j \rangle = \lambda_i \delta_{ij},$$

the first equality by Definition 11.6.1 and the last because $\mathbf{V}$ is unitary, so its columns are orthonormal (Theorem 11.7.3 (f)). In particular $\|\mathbf{A}\mathbf{v}_i\|^2 = \lambda_i = \sigma_i^2$, so

$$\mathbf{A}\mathbf{v}_i = \mathbf{0} \iff \sigma_i = 0. \tag{13.9.2}$$

Step 3: normalize, and count. Let r be the number of indices with $\sigma_i > 0$; since the σ_i decrease, these are $i = 1, \dots, r$. For $i \leq r$ put

$$\mathbf{u}_i := \frac{1}{\sigma_i}\mathbf{A}\mathbf{v}_i,$$

which is legitimate because $\sigma_i \neq 0$. By Step 2, $\langle \mathbf{u}_i, \mathbf{u}_j \rangle = \lambda_i \delta_{ij}/(\sigma_i \sigma_j) = \delta_{ij}$ for $i, j \leq r$, so $(\mathbf{u}_1, \dots, \mathbf{u}_r)$ is an orthonormal list in F^m.

We claim $r = \operatorname{rank} \mathbf{A}$. Any $\mathbf{x} \in F^n$ is $\mathbf{x} = \sum_i c_i \mathbf{v}_i$ since $(\mathbf{v}_1, \dots, \mathbf{v}_n)$ is a basis, and then $\mathbf{A}\mathbf{x} = \sum_{i \leq r} c_i \sigma_i \mathbf{u}_i$ by Eq. 13.9.2. Hence $\operatorname{im} \mathbf{A} = \operatorname{span}(\mathbf{u}_1, \dots, \mathbf{u}_r)$, and the $\mathbf{u}_i$ are orthonormal, hence independent (Theorem 11.1.4), so $\operatorname{rank} \mathbf{A} = r$. In particular $r \leq \min(m, n) = p$, so every index i with $\sigma_i > 0$ satisfies $i \leq p$, and the nonzero singular values are exactly $\sigma_1, \dots, \sigma_r$.

Step 4: pad the list. If $r = m$ the list $(\mathbf{u}_1, \dots, \mathbf{u}_r)$ is already an orthonormal basis of F^m, and we set $\mathbf{U} = (\mathbf{u}_1 \mid \cdots \mid \mathbf{u}_m)$. If $r < m$, apply Corollary 11.2.3 to the orthonormal list $(\mathbf{u}_1, \dots, \mathbf{u}_r)$ in F^m: there are $\mathbf{u}_{r+1}, \dots, \mathbf{u}_m$ making $(\mathbf{u}_1, \dots, \mathbf{u}_m)$ an orthonormal basis of F^m. Either way let $\mathbf{U} \in M_m(F)$ have these columns; it is unitary because its columns are orthonormal (Theorem 11.7.3 (f)). Note that the padding vectors are **not** determined by $\mathbf{A}$, and nothing below uses any property of them beyond orthonormality.

Step 5: read off the columns. Let $\mathbf{\Sigma} \in M_{m \times n}(\mathbb{R})$ be as in the statement. Column i of $\mathbf{U}\mathbf{\Sigma}$ is $\mathbf{U}$ times column i of $\mathbf{\Sigma}$, which is $\sigma_i \mathbf{e}_i$ for $i \leq p$ and $\mathbf{0}$ for $i > p$; so column i of $\mathbf{U}\mathbf{\Sigma}$ is $\sigma_i \mathbf{u}_i$ for $i \leq p$ and $\mathbf{0}$ for $i > p$. Column i of $\mathbf{A}\mathbf{V}$ is $\mathbf{A}\mathbf{v}_i$. Compare:

- for $i \leq r$, $\mathbf{A}\mathbf{v}_i = \sigma_i \mathbf{u}_i$ by the definition of $\mathbf{u}_i$;
- for $r < i \leq p$, $\sigma_i = 0$, so both columns are $\mathbf{0}$ by Eq. 13.9.2;
- for $i > p$ (possible only when $n > m$), we have $i > p \geq r$, so $\sigma_i = 0$ and $\mathbf{A}\mathbf{v}_i = \mathbf{0}$, and both columns are $\mathbf{0}$.

Hence $\mathbf{A}\mathbf{V} = \mathbf{U}\mathbf{\Sigma}$. Since $\mathbf{V}$ is unitary, $\mathbf{V}^{-1} = \mathbf{V}^*$, so $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$. This proves the theorem.

■

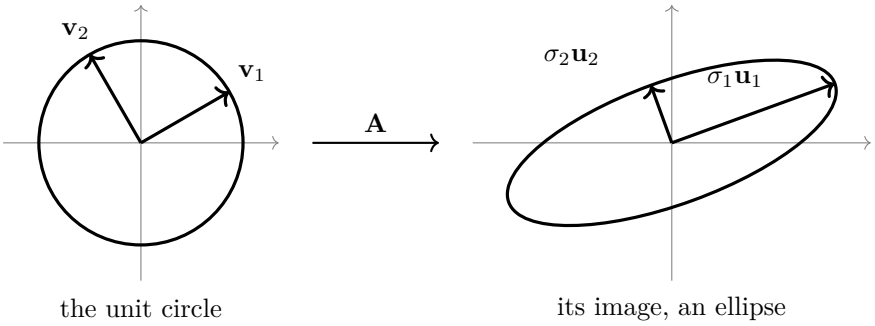
A factorization $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$ as in the theorem is called **a singular value decomposition** of $\mathbf{A}$, the columns $\mathbf{u}_i$ of $\mathbf{U}$ are **left singular vectors** and the columns $\mathbf{v}_i$ of $\mathbf{V}$ are **right singular vectors**. The two equations to keep in mind are the ones the proof produced,

$$\mathbf{A}\mathbf{v}_i = \sigma_i \mathbf{u}_i, \qquad \mathbf{A}^*\mathbf{u}_i = \sigma_i \mathbf{v}_i \qquad (i \le r), \tag{13.9.3}$$

the second obtained by applying $\mathbf{A}^*$ to the first and using $\mathbf{A}^*\mathbf{A}\mathbf{v}_i = \sigma_i^2 \mathbf{v}_i$. In words: $\mathbf{A}$ matches up two orthonormal bases, one in each space, stretching the i-th direction by σ_i, and $\mathbf{A}^*$ runs the same correspondence backwards with the same factors.

Remark. **The last item on the invertibility list.** For a square $\mathbf{A} \in M_n(F)$, the count at the end of the theorem says that all n singular values are nonzero exactly when $\operatorname{rank} \mathbf{A} = n$, which is invertibility (Theorem 7.5.3 (d)). This is the item Chapter 3 promised when it began the list (Theorem 3.4.4), later grown in Theorem 7.5.3, Theorem 9.2.6, and Theorem 13.1.6, and it is the last one the book adds.

Here is the picture the Idea described, in the real 2×2 case.



A carries a pair of perpendicular unit vectors to the semi-axes of an ellipse; their lengths are σ_1 and σ_2

Compare the principal axes figure of Chapter 12 §05. There the ellipse was a level curve $\mathbf{x}^\top \mathbf{A}\mathbf{x} = 1$ of a quadratic form and the axes lay along the eigenvectors of a symmetric $\mathbf{A}$; here the ellipse is the *image* of the unit circle under an arbitrary $\mathbf{A}$. The two pictures are the same picture, because $\|\mathbf{A}\mathbf{x}\|^2 = \mathbf{x}^*(\mathbf{A}^*\mathbf{A})\mathbf{x}$ is a quadratic form with matrix $\mathbf{A}^*\mathbf{A}$, and the right singular vectors of $\mathbf{A}$ are its principal axes. A large σ_i now means a **long** semi-axis, whereas in Theorem 12.5.5 a large λ_i meant a short one; the reason is that the two pictures are related by taking reciprocals of square roots.

13.9.3 What is unique and what is not

The singular values were defined before the theorem, so they cannot depend on any choice. What needs saying is the converse: any factorization of the stated shape recovers them.

Theorem 13.9.2. The Singular Values Are Determined

Let $\mathbf{A} \in M_{m\times n}(F)$ and suppose $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$ with $\mathbf{U}, \mathbf{V}$ unitary and $\mathbf{\Sigma} \in M_{m\times n}(\mathbb{R})$ of the shape in Theorem 13.9.1, with diagonal entries $s_1 \ge \cdots \ge s_p \ge 0$. Then $s_i = \sigma_i(\mathbf{A})$ for every i. Moreover:

(a) the columns of $\mathbf{V}$ are forced to be an orthonormal basis of F^n consisting of eigenvectors of $\mathbf{A}^*\mathbf{A}$, with $\mathbf{A}^*\mathbf{A}\mathbf{v}_i = \sigma_i^2 \mathbf{v}_i$ for every $i \le n$, where $\sigma_i := 0$ for $p < i \le n$; and any such basis occurs;

(b) once $\mathbf{V}$ is chosen, $\mathbf{u}_i = \mathbf{A}\mathbf{v}_i/\sigma_i$ is forced for every $i \le r = \operatorname{rank} \mathbf{A}$, while $\mathbf{u}_{r+1}, \dots, \mathbf{u}_m$

may be any orthonormal basis of $(\operatorname{im}\mathbf{A})^{\perp}$.

Proof. From $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$ and $\mathbf{U}^*\mathbf{U} = \mathbf{I}_m$,

$$\mathbf{A}^*\mathbf{A} = \mathbf{V}\mathbf{\Sigma}^*\mathbf{U}^*\mathbf{U}\mathbf{\Sigma}\mathbf{V}^* = \mathbf{V}(\mathbf{\Sigma}^*\mathbf{\Sigma})\mathbf{V}^*,$$

and $\mathbf{\Sigma}^*\mathbf{\Sigma} \in M_n(\mathbb{R})$ is the diagonal matrix with entries $s_1^2, \dots, s_p^2$ followed by zeros. So $\mathbf{A}^*\mathbf{A}$ is unitarily diagonalized with this diagonal, and its eigenvalue list in decreasing order is therefore $s_1^2 \geq \cdots \geq s_p^2 \geq 0 \geq \dots$ — already sorted, since the s_i are. Comparing with Definition 13.9.1 gives $s_i^2 = \sigma_i(\mathbf{A})^2$, hence $s_i = \sigma_i(\mathbf{A})$ because both are nonnegative.

(a) The displayed equation says the i-th column of $\mathbf{V}$ satisfies $\mathbf{A}^*\mathbf{A}\mathbf{v}_i = \sigma_i^2\mathbf{v}_i$, and the columns of a unitary matrix are orthonormal. Conversely, given any orthonormal basis of eigenvectors of $\mathbf{A}^*\mathbf{A}$ ordered by decreasing eigenvalue, Steps 2–5 of the proof of Theorem 13.9.1 build a factorization from it.

(b) For $i \leq r$ we have $\sigma_i \neq 0$, and comparing column i of $\mathbf{AV} = \mathbf{U}\mathbf{\Sigma}$ gives $\mathbf{A}\mathbf{v}_i = \sigma_i\mathbf{u}_i$, so $\mathbf{u}_i = \mathbf{A}\mathbf{v}_i/\sigma_i$. For $i > r$ the corresponding column of $\mathbf{\Sigma}$ is zero, so $\mathbf{u}_i$ is unconstrained by the equation $\mathbf{AV} = \mathbf{U}\mathbf{\Sigma}$; the only constraint left is that $(\mathbf{u}_1, \dots, \mathbf{u}_m)$ be orthonormal, which says exactly that $(\mathbf{u}_{r+1}, \dots, \mathbf{u}_m)$ is an orthonormal basis of $\operatorname{span}(\mathbf{u}_1, \dots, \mathbf{u}_r)^{\perp} = (\operatorname{im}\mathbf{A})^{\perp}$.

■

So $\mathbf{\Sigma}$ is unique and $\mathbf{U}, \mathbf{V}$ are not. The freedom is of two kinds, and both are visible in small cases. If $\mathbf{A}^*\mathbf{A}$ has a repeated eigenvalue, any orthonormal basis of that eigenspace may be used: $\mathbf{A} = \mathbf{I}_2$ has $\mathbf{\Sigma} = \mathbf{I}_2$ and $\mathbf{I}_2 = \mathbf{U}\mathbf{I}_2\mathbf{U}^*$ for **every** unitary $\mathbf{U}$, with $\mathbf{V} = \mathbf{U}$. If the eigenvalues are distinct, each $\mathbf{v}_i$ is still free up to a scalar of modulus 1, and replacing $\mathbf{v}_i$ by $c\mathbf{v}_i$ replaces $\mathbf{u}_i$ by $c\mathbf{u}_i$, leaving $\sigma_i\mathbf{u}_i\mathbf{v}_i^*$ unchanged because $c\bar{c} = 1$.

Remark. That last observation is why the factorization is worth having even though it is not unique: the *products* $\sigma_i\mathbf{u}_i\mathbf{v}_i^*$ are what later sections use, and those are unchanged by the phase freedom. Section 12 will define the pseudoinverse through a singular value decomposition and will have to check exactly this.

13.9.4 The compact form, and the four subspaces

Steps 3 and 5 of the proof produced more than the statement recorded: they located $\operatorname{im}\mathbf{A}$ and $\ker\mathbf{A}$ inside the two bases. Writing the factorization so that only the nonzero singular values appear makes this visible.

Theorem 13.9.3. Compact Singular Value Decomposition

Let $\mathbf{A} \in M_{m\times n}(F)$ have rank r and let $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$ be as in Theorem 13.9.1, with columns $\mathbf{u}_1, \dots, \mathbf{u}_m$ and $\mathbf{v}_1, \dots, \mathbf{v}_n$. Then

$$\mathbf{A} = \sum_{i=1}^{r} \sigma_i\, \mathbf{u}_i\mathbf{v}_i^*,$$

a sum of r matrices of rank 1. Equivalently $\mathbf{A} = \mathbf{U}_r\mathbf{\Sigma}_r\mathbf{V}_r^*$, where $\mathbf{U}_r \in M_{m\times r}(F)$ and

$\mathbf{V}_r \in M_{n\times r}(F)$ hold the first r columns of $\mathbf{U}$ and $\mathbf{V}$, and $\mathbf{\Sigma}_r = \operatorname{diag}(\sigma_1, \dots, \sigma_r)$ is square and **invertible**.

Proof. Write $\mathbf{A} = (\mathbf{U\Sigma})\mathbf{V}^*$. The columns of $\mathbf{U\Sigma}$ are $\sigma_1\mathbf{u}_1, \dots, \sigma_p\mathbf{u}_p$ followed by zero columns, as computed in Step 5 of Theorem 13.9.1, and the rows of $\mathbf{V}^*$ are $\mathbf{v}_1^*, \dots, \mathbf{v}_n^*$. By the outer product expansion of a matrix product (Corollary 8.1.3),

$$\mathbf{A} = \sum_{i=1}^{p} (\sigma_i \mathbf{u}_i)\mathbf{v}_i^* = \sum_{i=1}^{r} \sigma_i\, \mathbf{u}_i \mathbf{v}_i^*,$$

the terms with $i > r$ vanishing because $\sigma_i = 0$ there. Each $\mathbf{u}_i\mathbf{v}_i^*$ is nonzero with all columns multiples of $\mathbf{u}_i$, hence of rank 1. The same computation with the truncated matrices gives $\mathbf{U}_r\mathbf{\Sigma}_r\mathbf{V}_r^* = \sum_{i\le r} \sigma_i \mathbf{u}_i\mathbf{v}_i^*$, and $\mathbf{\Sigma}_r$ is invertible because its diagonal entries $\sigma_1, \dots, \sigma_r$ are nonzero.

■

Corollary 13.9.4. The Four Fundamental Subspaces from the Compact Form

With the notation of Theorem 13.9.3:

(a) $(\mathbf{u}_1, \dots, \mathbf{u}_r)$ is an orthonormal basis of $\operatorname{im}\mathbf{A} = \operatorname{col}(\mathbf{A})$;
(b) $(\mathbf{u}_{r+1}, \dots, \mathbf{u}_m)$ is an orthonormal basis of $\ker\mathbf{A}^*$;
(c) $(\mathbf{v}_1, \dots, \mathbf{v}_r)$ is an orthonormal basis of $\operatorname{im}\mathbf{A}^*$;
(d) $(\mathbf{v}_{r+1}, \dots, \mathbf{v}_n)$ is an orthonormal basis of $\ker\mathbf{A}$.

Idea. Two of the four were proved on the way to Theorem 13.9.1. The other two come for free by applying the same statements to $\mathbf{A}^*$, whose singular value decomposition is $\mathbf{A}^* = \mathbf{V\Sigma}^*\mathbf{U}^*$ — the roles of $\mathbf{u}$ and $\mathbf{v}$ simply swap.

Proof. Part (a). Step 3 of Theorem 13.9.1 showed $\operatorname{im}\mathbf{A} = \operatorname{span}(\mathbf{u}_1, \dots, \mathbf{u}_r)$, and these r vectors are orthonormal.

Part (d). By Eq. 13.9.2, $\mathbf{A}\mathbf{v}_i = \mathbf{0}$ exactly for $i > r$, and for $\mathbf{x} = \sum_i c_i\mathbf{v}_i$ we computed $\mathbf{Ax} = \sum_{i\le r} c_i\sigma_i\mathbf{u}_i$. Since the $\mathbf{u}_i$ are independent and the σ_i nonzero for $i \le r$, $\mathbf{Ax} = \mathbf{0}$ if and only if $c_1 = \cdots = c_r = 0$, that is, $\mathbf{x} \in \operatorname{span}(\mathbf{v}_{r+1}, \dots, \mathbf{v}_n)$.

Parts (b) and (c). Taking conjugate transposes, $\mathbf{A}^* = \mathbf{V\Sigma}^*\mathbf{U}^*$, and $\mathbf{\Sigma}^* \in M_{n\times m}(\mathbb{R})$ has the same diagonal $\sigma_1, \dots, \sigma_p$. So this is a singular value decomposition of $\mathbf{A}^*$, with the columns of $\mathbf{V}$ as left singular vectors and those of $\mathbf{U}$ as right singular vectors, and $\operatorname{rank}\mathbf{A}^* = \operatorname{rank}\mathbf{A} = r$ by Corollary 11.6.6 (a). Now the two parts just proved, applied to $\mathbf{A}^*$ instead of $\mathbf{A}$, give exactly (c) and (b).

■

This is the cleanest possible form of Theorem 11.6.5. That theorem said the four subspaces come in two orthogonal pairs, $F^n = \ker\mathbf{A} \oplus \operatorname{im}\mathbf{A}^*$ and $F^m = \operatorname{im}\mathbf{A} \oplus \ker\mathbf{A}^*$. The corollary

hands over **orthonormal bases of all four at once**, out of two matrices, and adds the missing piece of information: how $\mathbf{A}$ maps one pair to the other. Restricted to $\operatorname{im}\mathbf{A}^*$, the map $\mathbf{A}$ is a bijection onto $\operatorname{im}\mathbf{A}$, and in the bases $(\mathbf{v}_1,\dots,\mathbf{v}_r)$ and $(\mathbf{u}_1,\dots,\mathbf{u}_r)$ its matrix is the diagonal $\mathbf{\Sigma}_r$. Every matrix, after an orthonormal change of basis on each side, is a positive diagonal matrix padded with zeros.

Quick check. Let $\mathbf{A}\in M_{m\times n}(F)$ have rank r. Using Corollary 13.9.4, what is $\dim\ker\mathbf{A}^*$, and why does the answer not depend on which singular value decomposition was used?

Solution. The list $(\mathbf{u}_{r+1},\dots,\mathbf{u}_m)$ has $m-r$ members, so $\dim\ker\mathbf{A}^* = m-r$. It cannot depend on the factorization because $r=\operatorname{rank}\mathbf{A}$ is determined by $\mathbf{A}$ (Theorem 13.9.2), and in any case the answer agrees with $\dim(\operatorname{im}\mathbf{A})^\perp = m-\operatorname{rank}\mathbf{A}$ from Theorem 11.6.5 and Theorem 11.3.2.

13.9.5 Two computations and one comparison

The Frobenius norm. From $\|\mathbf{A}\|_F^2=\operatorname{tr}(\mathbf{A}^*\mathbf{A})$ and $\mathbf{A}^*\mathbf{A}=\mathbf{V}(\mathbf{\Sigma}^*\mathbf{\Sigma})\mathbf{V}^*$, or directly from Lemma 11.7.5,

$$\|\mathbf{A}\|_F^2=\|\mathbf{U}\mathbf{\Sigma}\mathbf{V}^*\|_F^2=\|\mathbf{\Sigma}\|_F^2=\sum_{i=1}^{p}\sigma_i^2.$$

So the Frobenius norm is the Euclidean length of the singular value list. Section 11 turns this one identity into the best low-rank approximation theorem.

The largest stretch. Let $\mathbf{x}=\sum_i c_i\mathbf{v}_i$ be a unit vector, so $\sum_i|c_i|^2=1$ by Theorem 11.9.2. Then $\mathbf{A}\mathbf{x}=\sum_{i\le r}c_i\sigma_i\mathbf{u}_i$, and since the $\mathbf{u}_i$ are orthonormal, Pythagoras (Theorem 11.1.3) gives

$$\|\mathbf{A}\mathbf{x}\|^2=\sum_{i\le r}\sigma_i^2|c_i|^2\le\sigma_1^2\sum_i|c_i|^2=\sigma_1^2,$$

with equality at $\mathbf{x}=\mathbf{v}_1$. So $\sigma_1=\max\{\|\mathbf{A}\mathbf{x}\|:\|\mathbf{x}\|=1\}$: the largest singular value is the largest factor by which $\mathbf{A}$ can lengthen a vector. Chapter 16 gives that quantity a name, the operator norm $\|\mathbf{A}\|_2$, and this computation is why $\|\mathbf{A}\|_2=\sigma_1$.

Against the eigenvalues. For a normal $\mathbf{A}\in M_n(\mathbb{C})$, write $\mathbf{A}=\mathbf{U}\mathbf{D}\mathbf{U}^*$ with $\mathbf{D}=\operatorname{diag}(\lambda_1,\dots,\lambda_n)$ by Corollary 12.4.3. Then

$$\mathbf{A}^*\mathbf{A}=\mathbf{U}\mathbf{D}^*\mathbf{D}\mathbf{U}^*=\mathbf{U}\operatorname{diag}(|\lambda_1|^2,\dots,|\lambda_n|^2)\mathbf{U}^*,$$

so the singular values of $\mathbf{A}$ are the moduli $|\lambda_i|$, sorted downwards. For a general matrix there is no relation at all in one direction — $\mathbf{N}=\begin{pmatrix}0&1\\0&0\end{pmatrix}$ has spectrum $\{0\}$ and $\sigma_1=1$ — and only the crude bound $|\lambda|\le\sigma_1$ in the other, since $\mathbf{A}\mathbf{x}=\lambda\mathbf{x}$ with $\|\mathbf{x}\|=1$ gives $|\lambda|=\|\mathbf{A}\mathbf{x}\|\le\sigma_1$.

Example. A singular value decomposition of a 2 by 3 matrix

Find a singular value decomposition of

$$\mathbf{A}=\begin{pmatrix}1&0&1\\0&1&-1\end{pmatrix}\in M_{2\times3}(\mathbb{R}),$$

and read off orthonormal bases of its four fundamental subspaces.

Solution. Here $m = 2$, $n = 3$, $p = 2$. The proof asks for the eigenvectors of the 3×3 matrix $\mathbf{A}^\top\mathbf{A}$, but there is a shortcut worth taking whenever $m < n$: run the proof on $\mathbf{A}^\top$ instead. It produces $\mathbf{A}^\top = \mathbf{V}\mathbf{\Sigma}^\top\mathbf{U}^\top$, and transposing that equation gives $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, which is what we want. The eigenvalue computation then lives in the smaller matrix

$$\mathbf{A}\mathbf{A}^\top = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}.$$

Step 1. The characteristic polynomial of $\mathbf{A}\mathbf{A}^\top$ is $x^2-4x+3 = (x-1)(x-3)$, so the eigenvalues are 3 and 1 and the singular values are $\sigma_1 = \sqrt{3}$, $\sigma_2 = 1$. For $\lambda = 3$, $\mathbf{A}\mathbf{A}^\top - 3\mathbf{I} = \begin{pmatrix} -1 & -1 \\ -1 & -1 \end{pmatrix}$ gives $x_1 = -x_2$, and for $\lambda = 1$ we get $x_1 = x_2$. Normalizing,

$$\mathbf{u}_1 = \tfrac{1}{\sqrt{2}}(1, -1), \qquad \mathbf{u}_2 = \tfrac{1}{\sqrt{2}}(1, 1).$$

Both singular values are nonzero, so $r = \operatorname{rank}\mathbf{A} = 2$.

Steps 2–4. The roles of $\mathbf{u}$ and $\mathbf{v}$ are swapped along with the roles of $\mathbf{A}$ and $\mathbf{A}^\top$, so here $\mathbf{v}_i = \mathbf{A}^\top\mathbf{u}_i/\sigma_i$:

$$\mathbf{v}_1 = \frac{\mathbf{A}^\top(1,-1)}{\sqrt{2}\sqrt{3}} = \tfrac{1}{\sqrt{6}}(1, -1, 2), \qquad \mathbf{v}_2 = \frac{\mathbf{A}^\top(1,1)}{\sqrt{2}} = \tfrac{1}{\sqrt{2}}(1, 1, 0).$$

This time the padding step is needed, since $r = 2 < 3$: a unit vector orthogonal to both is

$$\mathbf{v}_3 = \tfrac{1}{\sqrt{3}}(-1, 1, 1).$$

Step 5. Assembling,

$$\mathbf{U} = \tfrac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}, \qquad \mathbf{\Sigma} = \begin{pmatrix} \sqrt{3} & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix},$$

$$\mathbf{V} = \begin{pmatrix} \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{3}} \\ -\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{3}} \\ \frac{2}{\sqrt{6}} & 0 & \frac{1}{\sqrt{3}} \end{pmatrix}, \qquad \mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top.$$

As a check, $\|\mathbf{A}\|_F^2 = 1+0+1+0+1+1 = 4$ and $\sigma_1^2 + \sigma_2^2 = 3+1 = 4$.

The four subspaces (Corollary 13.9.4), with $r = 2$: $\operatorname{col}(\mathbf{A})$ has the orthonormal basis $\mathbf{u}_1, \mathbf{u}_2$, which is all of $\mathbb{R}^2$, so $\operatorname{nul}(\mathbf{A}^\top) = \{\mathbf{0}\}$; $\operatorname{row}(\mathbf{A}) = \operatorname{col}(\mathbf{A}^\top)$ has the orthonormal basis $\mathbf{v}_1, \mathbf{v}_2$; and $\operatorname{nul}(\mathbf{A}) = \operatorname{span}(\mathbf{v}_3)$ is the line through $(-1, 1, 1)$. The last is easy to verify directly: $\mathbf{A}(-1, 1, 1) = (-1+1, 1-1) = \mathbf{0}$.

13.9.6 Exercises

13.9.6.1 A. Check your understanding

Exercise. A1

(a) Define the singular values of a matrix $\mathbf{A} \in M_{m\times n}(F)$. How many are there?

(b) State the singular value decomposition, with the size and the defining property of each factor.

(c) True or false: the singular values of $\mathbf{A}$ are the absolute values of its eigenvalues. Justify your answer.

(d) How many nonzero singular values does $\mathbf{A}$ have, and why?

(e) Given a compact singular value decomposition $\mathbf{A} = \sum_{i\le r} \sigma_i \mathbf{u}_i \mathbf{v}_i^*$, write down orthonormal bases of all four fundamental subspaces of $\mathbf{A}$.

(f) True or false: in $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$ the three factors are determined by $\mathbf{A}$. Justify your answer.

Solution.

(a) They are $\sigma_i = \sqrt{\lambda_i}$ for $i = 1, \dots, \min(m,n)$, where $\lambda_1 \ge \cdots \ge \lambda_n \ge 0$ are the eigenvalues of $\mathbf{A}^*\mathbf{A}$ with multiplicity (Definition 13.9.1). There are $\min(m,n)$ of them.

(b) Theorem 13.9.1: $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$ with $\mathbf{U} \in M_m(F)$ unitary, $\mathbf{V} \in M_n(F)$ unitary, and $\mathbf{\Sigma} \in M_{m\times n}(\mathbb{R})$ having $\sigma_1 \ge \cdots \ge \sigma_p \ge 0$ in positions (i,i) and zeros elsewhere.

(c) False. $\mathbf{N} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$ has both eigenvalues 0 and singular values $1, 0$. The statement is true for normal matrices (Exercise (C3: Singular values equal to eigenvalue moduli)), and for a nonsquare matrix it does not even make sense.

(d) Exactly $\operatorname{rank} \mathbf{A}$, by Theorem 13.9.1. The reason is that $\mathbf{A}\mathbf{v}_i = \mathbf{0}$ precisely when $\sigma_i = 0$, so the $\mathbf{v}_i$ with $\sigma_i = 0$ span $\ker \mathbf{A}$.

(e) $(\mathbf{u}_1, \dots, \mathbf{u}_r)$ for $\operatorname{im} \mathbf{A}$; $(\mathbf{u}_{r+1}, \dots, \mathbf{u}_m)$ for $\ker \mathbf{A}^*$; $(\mathbf{v}_1, \dots, \mathbf{v}_r)$ for $\operatorname{im} \mathbf{A}^*$; $(\mathbf{v}_{r+1}, \dots, \mathbf{v}_n)$ for $\ker \mathbf{A}$ (Corollary 13.9.4).

(f) False for $\mathbf{U}$ and $\mathbf{V}$, true for $\mathbf{\Sigma}$ (Theorem 13.9.2). For instance $\mathbf{I}_2 = \mathbf{U}\mathbf{I}_2\mathbf{U}^*$ for every unitary $\mathbf{U}$. The freedom is: any orthonormal eigenbasis of $\mathbf{A}^*\mathbf{A}$ may be used for $\mathbf{V}$, and the last $m - r$ columns of $\mathbf{U}$ may be any orthonormal basis of $(\operatorname{im} \mathbf{A})^\perp$.

13.9.6.2 B. Practice

Exercise. B1: A 2 by 2 decomposition

Find a singular value decomposition $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ of

$$\mathbf{A} = \begin{pmatrix} 3 & 0 \\ 4 & 5 \end{pmatrix} \in M_2(\mathbb{R}),$$

and compare the singular values with the eigenvalues of $\mathbf{A}$.

Solution. $\mathbf{A}^\top\mathbf{A} = \begin{pmatrix} 25 & 20 \\ 20 & 25 \end{pmatrix}$, with characteristic polynomial $x^2 - 50x + (625 - 400) = x^2 - 50x + 225$, whose roots are 45 and 5. So $\sigma_1 = 3\sqrt{5}$ and $\sigma_2 = \sqrt{5}$.

For $\lambda = 45$ the equation $-20x_1 + 20x_2 = 0$ gives $\mathbf{v}_1 = \frac{1}{\sqrt{2}}(1,1)$; for $\lambda = 5$, $\mathbf{v}_2 = \frac{1}{\sqrt{2}}(1,-1)$.

Then
$$\mathbf{u}_1 = \frac{\mathbf{A}\mathbf{v}_1}{3\sqrt{5}} = \frac{(3,9)}{3\sqrt{10}} = \tfrac{1}{\sqrt{10}}(1,3), \qquad \mathbf{u}_2 = \frac{\mathbf{A}\mathbf{v}_2}{\sqrt{5}} = \frac{(3,-1)}{\sqrt{10}}.$$
Hence
$$\mathbf{U} = \tfrac{1}{\sqrt{10}}\begin{pmatrix} 1 & 3 \\ 3 & -1 \end{pmatrix}, \quad \mathbf{\Sigma} = \begin{pmatrix} 3\sqrt{5} & 0 \\ 0 & \sqrt{5} \end{pmatrix}, \quad \mathbf{V} = \tfrac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix},$$
and $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top = \mathbf{A}$. Check: $\sigma_1^2 + \sigma_2^2 = 45 + 5 = 50 = 9 + 16 + 25 = \|\mathbf{A}\|_F^2$.

The eigenvalues of $\mathbf{A}$ are 3 and 5, the diagonal entries of a triangular matrix. They are not the singular values, not even after reordering: $\mathbf{A}$ is not normal, since $\mathbf{A}^\top\mathbf{A} \neq \mathbf{A}\mathbf{A}^\top = \begin{pmatrix} 9 & 12 \\ 12 & 41 \end{pmatrix}$. Note, however, that $\sigma_1\sigma_2 = 15 = |\det \mathbf{A}|$, as Exercise (C2: The adjoint and the determinant) explains.

Exercise. B2: When the left singular vectors must be padded

Find a singular value decomposition of
$$\mathbf{A} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \end{pmatrix} \in M_{3\times 2}(\mathbb{R}),$$
saying explicitly where the padding step of Theorem 13.9.1 is used.

Solution. $\mathbf{A}^\top\mathbf{A} = \operatorname{diag}(2,1)$, already diagonal with decreasing diagonal, so $\mathbf{V} = \mathbf{I}_2$, $\mathbf{v}_1 = \mathbf{e}_1$, $\mathbf{v}_2 = \mathbf{e}_2$, and $\sigma_1 = \sqrt{2}$, $\sigma_2 = 1$. Both are nonzero, so $r = 2$. Then
$$\mathbf{u}_1 = \tfrac{1}{\sqrt{2}}\mathbf{A}\mathbf{e}_1 = \tfrac{1}{\sqrt{2}}(1,0,1), \qquad \mathbf{u}_2 = \mathbf{A}\mathbf{e}_2 = (0,1,0).$$
Here $r = 2 < 3 = m$, so Step 4 of Theorem 13.9.1 applies: extend $(\mathbf{u}_1, \mathbf{u}_2)$ to an orthonormal basis of $\mathbb{R}^3$. A vector orthogonal to both is $(1,0,-1)$, so take $\mathbf{u}_3 = \frac{1}{\sqrt{2}}(1,0,-1)$. Then
$$\mathbf{U} = \begin{pmatrix} \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \\ 0 & 1 & 0 \\ \frac{1}{\sqrt{2}} & 0 & -\frac{1}{\sqrt{2}} \end{pmatrix}, \qquad \mathbf{\Sigma} = \begin{pmatrix} \sqrt{2} & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix}, \qquad \mathbf{V} = \mathbf{I}_2,$$
and $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top = \mathbf{U}\mathbf{\Sigma} = \mathbf{A}$. The vector $\mathbf{u}_3$ is not determined by $\mathbf{A}$ beyond its line: $-\mathbf{u}_3$ would do just as well. By Corollary 13.9.4, $\operatorname{nul}(\mathbf{A}^\top) = \operatorname{span}(\mathbf{u}_3)$, which is confirmed by $\mathbf{A}^\top(1,0,-1) = (1-1,0) = \mathbf{0}$.

Exercise. B3: Compact form and the four subspaces

For
$$\mathbf{A} = \begin{pmatrix} 2 & 2 \\ 1 & 1 \\ 2 & 2 \end{pmatrix} \in M_{3\times 2}(\mathbb{R}),$$

find the compact singular value decomposition $\mathbf{A} = \sum_{i \le r} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$, and hence orthonormal bases of all four fundamental subspaces.

Solution. $\mathbf{A}^\top \mathbf{A} = \begin{pmatrix} 9 & 9 \\ 9 & 9 \end{pmatrix}$, with eigenvalues 18 and 0, so $\sigma_1 = 3\sqrt{2}$, $\sigma_2 = 0$, and $r = 1$. An eigenvector for 18 is $\mathbf{v}_1 = \frac{1}{\sqrt{2}}(1, 1)$, and for 0 it is $\mathbf{v}_2 = \frac{1}{\sqrt{2}}(1, -1)$. Then

$$\mathbf{u}_1 = \frac{\mathbf{A}\mathbf{v}_1}{3\sqrt{2}} = \frac{(4, 2, 4)}{\sqrt{2} \cdot 3\sqrt{2}} = \tfrac{1}{3}(2, 1, 2),$$

a unit vector since $4 + 1 + 4 = 9$. The compact form is $\mathbf{A} = 3\sqrt{2}\, \mathbf{u}_1 \mathbf{v}_1^\top$, and multiplying out gives back $\mathbf{A}$.

Four subspaces: $\operatorname{col}(\mathbf{A}) = \operatorname{span}(\frac{1}{3}(2, 1, 2))$ and $\operatorname{row}(\mathbf{A}) = \operatorname{span}(\frac{1}{\sqrt{2}}(1, 1))$, both one-dimensional; $\operatorname{nul}(\mathbf{A}) = \operatorname{span}(\frac{1}{\sqrt{2}}(1, -1))$; and $\operatorname{nul}(\mathbf{A}^\top)$ is the plane $\{\mathbf{y} : 2y_1 + y_2 + 2y_3 = 0\}$, for which an orthonormal basis is $\frac{1}{\sqrt{2}}(1, 0, -1)$ and $\frac{1}{3\sqrt{2}}(1, -4, 1)$. (Any orthonormal basis of that plane is a legitimate $(\mathbf{u}_2, \mathbf{u}_3)$, which is the freedom described in Theorem 13.9.2 (b).)

Exercise. B4: Determine whether

Determine which of the following statements are correct. For those that are correct, prove them; for those that are not, give a counterexample.

(a) If every eigenvalue of $\mathbf{A} \in M_n(\mathbb{C})$ is 0, then every singular value of $\mathbf{A}$ is 0.
(b) $\sigma_i(c\mathbf{A}) = |c|\, \sigma_i(\mathbf{A})$ for every scalar c.
(c) $\sigma_i(\mathbf{A}^2) = \sigma_i(\mathbf{A})^2$ for every square $\mathbf{A}$.
(d) $\mathbf{A}$ is invertible if and only if $n = m$ and every singular value of $\mathbf{A}$ is nonzero.

Solution.

(a) Incorrect. $\mathbf{N} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$ has spectrum $\{0\}$, while $\mathbf{N}^*\mathbf{N} = \operatorname{diag}(0, 1)$ gives $\sigma_1 = 1$.

(b) Correct. $(c\mathbf{A})^*(c\mathbf{A}) = \overline{c}c\, \mathbf{A}^*\mathbf{A} = |c|^2 \mathbf{A}^*\mathbf{A}$, whose eigenvalues are $|c|^2\lambda_i$; taking square roots and noting that multiplying by $|c|^2 \ge 0$ preserves the decreasing order gives the claim.

(c) Incorrect, with the same $\mathbf{N}$: $\mathbf{N}^2 = \mathbf{O}$ has both singular values 0, while $\sigma_1(\mathbf{N})^2 = 1$. (It is correct for $\mathbf{A} \succeq \mathbf{O}$, and more generally for normal $\mathbf{A}$, because then the singular values are the moduli of the eigenvalues and $\mathbf{A}^2$ has eigenvalues λ_i^2.)

(d) Correct. If $\mathbf{A} \in M_n(F)$ then, by Theorem 13.9.1, the number of nonzero singular values is $\operatorname{rank} \mathbf{A}$, and $\mathbf{A}$ is invertible exactly when $\operatorname{rank} \mathbf{A} = n$ (Theorem 7.5.3 (d)). A nonsquare matrix is never invertible.

13.9.6.3 C. Going deeper

Exercise. C1: The Frobenius norm and the largest singular value

Let $\mathbf{A} \in M_{m\times n}(F)$ with singular values $\sigma_1 \geq \cdots \geq \sigma_p$.

(a) The text obtained $\|\mathbf{A}\|_F^2 = \sum_{i=1}^{p} \sigma_i^2$ from unitary invariance. Prove it a second way, directly from $\|\mathbf{A}\|_F^2 = \operatorname{tr}(\mathbf{A}^*\mathbf{A})$ and the eigenvalues of $\mathbf{A}^*\mathbf{A}$, without invoking Lemma 11.7.5.

(b) Deduce that $\sigma_1 \leq \|\mathbf{A}\|_F$, and determine exactly when equality holds.

Solution.

(a) Take a singular value decomposition $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$ (Theorem 13.9.1). By Lemma 11.7.5, applied once on each side, $\|\mathbf{A}\|_F = \|\mathbf{U}\mathbf{\Sigma}\mathbf{V}^*\|_F = \|\mathbf{\Sigma}\|_F$. The only nonzero entries of $\mathbf{\Sigma}$ are $\sigma_1, \ldots, \sigma_p$ on the diagonal, so $\|\mathbf{\Sigma}\|_F^2 = \sum_i \sigma_i^2$.

(b) Every term of $\sum_i \sigma_i^2$ is nonnegative, so $\sigma_1^2 \leq \sum_i \sigma_i^2 = \|\mathbf{A}\|_F^2$, and taking nonnegative square roots gives $\sigma_1 \leq \|\mathbf{A}\|_F$. Equality holds if and only if $\sigma_2 = \cdots = \sigma_p = 0$, that is, if and only if $\operatorname{rank} \mathbf{A} \leq 1$ (Theorem 13.9.1).

Exercise. C2: The adjoint and the determinant

Let $\mathbf{A} \in M_{m\times n}(F)$.

(a) Prove that $\mathbf{A}$ and $\mathbf{A}^*$ have the same nonzero singular values, with the same multiplicities.

(b) Suppose $m = n$. Prove that $|\det \mathbf{A}| = \sigma_1\sigma_2\cdots\sigma_n$.

(c) Hence give a formula for $|\det \mathbf{A}|$ in terms of $\mathbf{A}^*\mathbf{A}$ only, and check it on $\mathbf{A} = \begin{pmatrix} 3 & 0 \\ 4 & 5 \end{pmatrix}$.

Solution.

(a) If $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$ is a singular value decomposition, then $\mathbf{A}^* = \mathbf{V}\mathbf{\Sigma}^*\mathbf{U}^*$ with $\mathbf{V}, \mathbf{U}$ unitary and $\mathbf{\Sigma}^* \in M_{n\times m}(\mathbb{R})$ carrying the same diagonal entries $\sigma_1, \ldots, \sigma_p$. By Theorem 13.9.2 this is a singular value decomposition of $\mathbf{A}^*$, so the singular value list of $\mathbf{A}^*$ is $\sigma_1, \ldots, \sigma_p$ as well. (The two lists have the same length $p = \min(m, n)$, so they are in fact equal, not merely equal in their nonzero parts.)

(b) Take determinants in $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$. By Theorem 7.5.4, $\det \mathbf{A} = \det \mathbf{U} \cdot \det \mathbf{\Sigma} \cdot \det \mathbf{V}^*$. A unitary $\mathbf{U}$ has $|\det \mathbf{U}| = 1$, since $\det(\mathbf{U}^*)\det(\mathbf{U}) = \det(\mathbf{I}) = 1$ and $\det(\mathbf{U}^*) = \overline{\det \mathbf{U}}$, so $|\det \mathbf{U}|^2 = 1$. The same holds for $\mathbf{V}^*$. And $\det \mathbf{\Sigma} = \sigma_1 \cdots \sigma_n$ by Theorem 7.4.6. Taking absolute values gives $|\det \mathbf{A}| = \sigma_1 \cdots \sigma_n$.

(c) Squaring, $|\det \mathbf{A}|^2 = \prod_i \sigma_i^2 = \det(\mathbf{A}^*\mathbf{A})$, so $|\det \mathbf{A}| = \sqrt{\det(\mathbf{A}^*\mathbf{A})}$. For the given $\mathbf{A}$, $\det \mathbf{A} = 15$ and $\det(\mathbf{A}^\top\mathbf{A}) = 625 - 400 = 225 = 15^2$.

Exercise. C3: Singular values equal to eigenvalue moduli

Let $\mathbf{A} \in M_n(\mathbb{C})$ have eigenvalues $\lambda_1, \ldots, \lambda_n$ listed with algebraic multiplicity and ordered so that $|\lambda_1| \geq \cdots \geq |\lambda_n|$.

(a) Prove that if $\mathbf{A}$ is normal, then $\sigma_i(\mathbf{A}) = |\lambda_i|$ for every i.

(b) Prove the converse: if $\sigma_i(\mathbf{A}) = |\lambda_i|$ for every i, then $\mathbf{A}$ is normal.

Hint for (b): compare two expressions for $\|\mathbf{A}\|_F^2$.

Solution.

(a) Suppose $\mathbf{A}$ is normal. By Corollary 12.4.3, $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{U}^*$ with $\mathbf{U}$ unitary and $\mathbf{D} = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$. Then

$$\mathbf{A}^*\mathbf{A} = \mathbf{U}\mathbf{D}^*\mathbf{U}^*\mathbf{U}\mathbf{D}\mathbf{U}^* = \mathbf{U}\operatorname{diag}(|\lambda_i|^2)\,\mathbf{U}^*,$$

so the eigenvalues of $\mathbf{A}^*\mathbf{A}$ are $|\lambda_1|^2, \dots, |\lambda_n|^2$ with multiplicity. These are already in decreasing order by hypothesis, so $\sigma_i(\mathbf{A}) = |\lambda_i|$ for every i by Definition 13.9.1.

(b) Suppose $\sigma_i(\mathbf{A}) = |\lambda_i|$ for every i. By Exercise (C1: The Frobenius norm and the largest singular value) (a),

$$\sum_{i=1}^{n} |\lambda_i|^2 = \sum_{i=1}^{n} \sigma_i^2 = \|\mathbf{A}\|_F^2.$$

Schur's inequality (Theorem 12.11.1) says $\sum_i |\lambda_i|^2 \le \|\mathbf{A}\|_F^2$ with equality **if and only if** $\mathbf{A}$ is normal. We have equality, so $\mathbf{A}$ is normal.

13.10 Polar Decomposition and Takagi's Factorization

Every nonzero complex number is a rotation times a stretch, $z = e^{i\theta} r$ with $r = |z| > 0$. Section 9 has put the matrix version within reach: if $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$, then inserting $\mathbf{V}^*\mathbf{V} = \mathbf{I}$ in the middle splits the factorization into a unitary part and a positive semidefinite part. This section carries that out, proves that the unitary part is the closest unitary matrix to $\mathbf{A}$, and then treats a factorization that looks like the singular value decomposition but is not: Takagi's, for complex **symmetric** matrices.

The field is $\mathbb{C}$ unless stated otherwise, and $|\mathbf{A}| = (\mathbf{A}^*\mathbf{A})^{1/2}$ is the matrix absolute value of Definition 13.2.1.

13.10.1 Rotation times stretch

Theorem 13.10.1. Polar Decomposition

Let $\mathbf{A} \in M_n(\mathbb{C})$. Then there is a unitary $\mathbf{W} \in M_n(\mathbb{C})$ with

$$\mathbf{A} = \mathbf{W}|\mathbf{A}|.$$

The positive semidefinite factor $|\mathbf{A}|$ is determined by $\mathbf{A}$. The unitary factor $\mathbf{W}$ is determined by $\mathbf{A}$ if and only if $\mathbf{A}$ is invertible, and then $\mathbf{W} = \mathbf{A}|\mathbf{A}|^{-1}$. Moreover the same $\mathbf{W}$ satisfies the right-handed identity

$$\mathbf{A} = |\mathbf{A}^*|\,\mathbf{W}, \qquad |\mathbf{A}^*| = \mathbf{W}|\mathbf{A}|\mathbf{W}^*.$$

Idea. The singular value decomposition already separates the two kinds of behavior: $\mathbf{U}$ and $\mathbf{V}^*$ rotate, and $\mathbf{\Sigma}$ stretches along the axes. Writing $\mathbf{U\Sigma V}^* = (\mathbf{UV}^*)(\mathbf{V\Sigma V}^*)$ puts all the rotating in the first bracket and all the stretching in the second, and the second bracket is a positive semidefinite matrix whose square is $\mathbf{A}^*\mathbf{A}$. Uniqueness of the positive semidefinite square root does the rest. For the uniqueness of $\mathbf{W}$, notice that $\mathbf{W}$ is pinned down only where $|\mathbf{A}|$ is invertible; when $\mathbf{A} = \mathbf{O}$, every unitary works.

Proof. **Existence.** By Theorem 13.9.1 write $\mathbf{A} = \mathbf{U\Sigma V}^*$ with $\mathbf{U}, \mathbf{V} \in M_n(\mathbb{C})$ unitary and $\mathbf{\Sigma} = \operatorname{diag}(\sigma_1, \dots, \sigma_n)$ with $\sigma_i \geq 0$. Put

$$\mathbf{W} \coloneqq \mathbf{UV}^*, \qquad \mathbf{P} \coloneqq \mathbf{V\Sigma V}^*.$$

Then $\mathbf{W}$ is unitary, being a product of unitary matrices, and

$$\mathbf{WP} = \mathbf{UV}^*\mathbf{V\Sigma V}^* = \mathbf{U\Sigma V}^* = \mathbf{A}.$$

The matrix $\mathbf{P}$ is self-adjoint, since $\mathbf{P}^* = \mathbf{V\Sigma}^*\mathbf{V}^* = \mathbf{P}$, and positive semidefinite: its eigenvalues are the $\sigma_i \geq 0$ because $\mathbf{V}^*\mathbf{PV} = \mathbf{\Sigma}$, so Theorem 13.1.4 applies. Finally

$$\mathbf{P}^2 = \mathbf{V\Sigma V}^*\mathbf{V\Sigma V}^* = \mathbf{V\Sigma}^2\mathbf{V}^* = \mathbf{A}^*\mathbf{A},$$

the last equality because $\mathbf{A}^*\mathbf{A} = \mathbf{V\Sigma U}^*\mathbf{U\Sigma V}^* = \mathbf{V\Sigma}^2\mathbf{V}^*$. A positive semidefinite matrix has exactly one positive semidefinite square root (Theorem 13.2.1), so $\mathbf{P} = (\mathbf{A}^*\mathbf{A})^{1/2} = |\mathbf{A}|$, and $\mathbf{A} = \mathbf{W}|\mathbf{A}|$.

The factor $|\mathbf{A}|$ is determined, since $\mathbf{A}^*\mathbf{A}$ is and the square root is unique.

Uniqueness of $\mathbf{W}$. ($\Rightarrow$) Suppose $\mathbf{A}$ is invertible. Then $|\mathbf{A}|^2 = \mathbf{A}^*\mathbf{A}$ is invertible, so $|\mathbf{A}|$ is, and any $\mathbf{W}$ with $\mathbf{A} = \mathbf{W}|\mathbf{A}|$ satisfies $\mathbf{W} = \mathbf{A}|\mathbf{A}|^{-1}$: there is only one. ($\Leftarrow$) Suppose $\mathbf{A}$ is not invertible, so there is a nonzero $\mathbf{z} \in \ker \mathbf{A}$. Let $\mathbf{H} = \mathbf{H}_{\mathbf{z}}$ be the Householder reflection of Definition 11.8.1, which is unitary with $\mathbf{Hz} = -\mathbf{z}$ and $\mathbf{Hv} = \mathbf{v}$ for every $\mathbf{v} \perp \mathbf{z}$ (Proposition 11.8.2 (c), (d)). By Proposition 13.2.2, $\operatorname{nul}|\mathbf{A}| = \operatorname{nul} \mathbf{A}$, and $|\mathbf{A}|$ is self-adjoint, so

$$\operatorname{col}|\mathbf{A}| = \operatorname{col}(|\mathbf{A}|^*) = (\operatorname{nul}|\mathbf{A}|)^{\perp} = (\operatorname{nul} \mathbf{A})^{\perp} \subseteq \mathbf{z}^{\perp}$$

by Theorem 11.6.5 (b). Hence $\mathbf{H}$ fixes every vector of $\operatorname{col}|\mathbf{A}|$, that is, $\mathbf{H}|\mathbf{A}| = |\mathbf{A}|$, and therefore $(\mathbf{WH})|\mathbf{A}| = \mathbf{A}$ as well, with $\mathbf{WH}$ unitary and $\mathbf{WH} \neq \mathbf{W}$ because $\mathbf{H} \neq \mathbf{I}$ and $\mathbf{W}$ is invertible. So $\mathbf{W}$ is not determined.

The right-handed form. Set $\mathbf{Q} \coloneqq \mathbf{W}|\mathbf{A}|\mathbf{W}^*$. It is self-adjoint and positive semidefinite, being a unitary congruence of a positive semidefinite matrix: $\mathbf{x}^*\mathbf{Qx} = (\mathbf{W}^*\mathbf{x})^*|\mathbf{A}|(\mathbf{W}^*\mathbf{x}) \geq 0$. And

$$\mathbf{Q}^2 = \mathbf{W}|\mathbf{A}|^2\mathbf{W}^* = \mathbf{WA}^*\mathbf{AW}^* = \mathbf{AA}^*,$$

using $\mathbf{A} = \mathbf{W}|\mathbf{A}|$, so $\mathbf{A}^* = |\mathbf{A}|\mathbf{W}^*$ and $\mathbf{AA}^* = \mathbf{W}|\mathbf{A}|^2\mathbf{W}^*$. By uniqueness of the square root again, $\mathbf{Q} = (\mathbf{AA}^*)^{1/2} = |\mathbf{A}^*|$. Therefore $|\mathbf{A}^*|\mathbf{W} = \mathbf{W}|\mathbf{A}|\mathbf{W}^*\mathbf{W} = \mathbf{W}|\mathbf{A}| = \mathbf{A}$. This proves the theorem.

■

Remark. There is a second route to existence, the one Proposition 13.2.2 was pointing at: the identity $\||\mathbf{A}|\mathbf{x}\| = \|\mathbf{A}\mathbf{x}\|$ says that $|\mathbf{A}|\mathbf{x} \mapsto \mathbf{A}\mathbf{x}$ is a well-defined isometry from $\operatorname{col}|\mathbf{A}|$ onto $\operatorname{col}\mathbf{A}$, and extending it by any isometry between the orthogonal complements produces $\mathbf{W}$ directly. That proof shows more clearly *where* the freedom in $\mathbf{W}$ lives, but it needs the extension step written out; the route above borrows all of that from Theorem 13.9.1.

For $n = 1$ the theorem is exactly $z = e^{i\theta}|z|$: a 1×1 unitary matrix is a complex number of modulus 1, and $|\mathbf{A}|$ is the modulus. The nonuniqueness for singular $\mathbf{A}$ is the matrix version of the fact that θ is undefined when $z = 0$.

Example. A polar decomposition

Find the left and right polar decompositions of $\mathbf{A} = \begin{pmatrix} 3 & 0 \\ 4 & 5 \end{pmatrix}$.

Solution. Exercise (B1: A 2 by 2 decomposition) gives the singular value decomposition

$$\mathbf{U} = \tfrac{1}{\sqrt{10}}\begin{pmatrix} 1 & 3 \\ 3 & -1 \end{pmatrix}, \quad \mathbf{\Sigma} = \sqrt{5}\begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix}, \quad \mathbf{V} = \tfrac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}.$$

Hence

$$\mathbf{W} = \mathbf{U}\mathbf{V}^{\top} = \tfrac{1}{\sqrt{20}}\begin{pmatrix} 1 & 3 \\ 3 & -1 \end{pmatrix}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} = \tfrac{1}{\sqrt{5}}\begin{pmatrix} 2 & -1 \\ 1 & 2 \end{pmatrix},$$

$$|\mathbf{A}| = \mathbf{V}\mathbf{\Sigma}\mathbf{V}^{\top} = \sqrt{5}\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}.$$

Check: $|\mathbf{A}|^2 = 5\begin{pmatrix} 5 & 4 \\ 4 & 5 \end{pmatrix} = \begin{pmatrix} 25 & 20 \\ 20 & 25 \end{pmatrix} = \mathbf{A}^{\top}\mathbf{A}$, the columns of $\mathbf{W}$ are orthonormal, and $\mathbf{W}|\mathbf{A}| = \mathbf{A}$. Here $\mathbf{W}$ is a rotation, since $\det\mathbf{W} = (4+1)/5 = 1$.

For the right-handed form, $\mathbf{A}\mathbf{A}^{\top} = \begin{pmatrix} 9 & 12 \\ 12 & 41 \end{pmatrix}$ and

$$|\mathbf{A}^*| = \mathbf{W}|\mathbf{A}|\mathbf{W}^{\top} = \tfrac{\sqrt{5}}{5}\begin{pmatrix} 6 & 3 \\ 3 & 14 \end{pmatrix},$$

whose square is indeed $\frac{1}{5}\begin{pmatrix} 45 & 60 \\ 60 & 205 \end{pmatrix} = \mathbf{A}\mathbf{A}^{\top}$. The two absolute values differ, and that is not an accident.

Remark. $|\mathbf{A}| = |\mathbf{A}^*|$ exactly when $\mathbf{A}$ is normal. Indeed $\mathbf{A}^*\mathbf{A} = \mathbf{A}\mathbf{A}^*$ if and only if the two positive semidefinite square roots agree, by uniqueness. So for a normal matrix — and only then — the stretching looks the same before and after the rotation, and $\mathbf{W}$ commutes with $|\mathbf{A}|$ by the identity $|\mathbf{A}^*| = \mathbf{W}|\mathbf{A}|\mathbf{W}^*$.

Quick check. What is the polar decomposition of a unitary matrix? Of a positive semidefinite matrix?

Solution. If $\mathbf{A}$ is unitary then $\mathbf{A}^*\mathbf{A} = \mathbf{I}$, so $|\mathbf{A}| = \mathbf{I}$ and $\mathbf{W} = \mathbf{A}$: all rotation, no stretch. If $\mathbf{A} \succeq \mathbf{O}$ then $\mathbf{A}^*\mathbf{A} = \mathbf{A}^2$ and $\mathbf{A}$ is itself a positive semidefinite square root of $\mathbf{A}^2$, so $|\mathbf{A}| = \mathbf{A}$ by uniqueness (Theorem 13.2.1) and $\mathbf{W} = \mathbf{I}$ whenever $\mathbf{A}$ is invertible: all stretch, no rotation.

13.10.2 The nearest unitary matrix

The polar factor is not merely *a* unitary matrix attached to $\mathbf{A}$. It is the one closest to $\mathbf{A}$, measured in the Frobenius norm of Chapter 12 §11. This is the reason the polar decomposition is used in practice: given a matrix that ought to be unitary but has drifted, $\mathbf{W}$ is how to put it back.

Corollary 13.10.2. The Polar Factor Is the Nearest Unitary Matrix

Let $\mathbf{A} \in M_n(\mathbb{C})$ with polar decomposition $\mathbf{A} = \mathbf{W}|\mathbf{A}|$. Then

$$\|\mathbf{A} - \mathbf{W}\|_F \le \|\mathbf{A} - \mathbf{Q}\|_F \qquad \text{for every unitary } \mathbf{Q} \in M_n(\mathbb{C}),$$

and the minimum value is $\left(\|\mathbf{A}\|_F^2 + n - 2\sum_i \sigma_i\right)^{1/2}$. If $\mathbf{A}$ is invertible, $\mathbf{W}$ is the **only** unitary matrix attaining the minimum.

Idea. Two moves. First, the Frobenius norm does not see unitary factors (Lemma 11.7.5), so sandwiching by $\mathbf{U}^*$ and $\mathbf{V}$ turns the problem into: among unitary $\mathbf{Z}$, which is closest to the diagonal matrix $\mathbf{\Sigma}$? Second, expanding $\|\mathbf{\Sigma} - \mathbf{Z}\|_F^2$ leaves only the diagonal entries of $\mathbf{Z}$ to be chosen, and each of them has modulus at most 1 because it sits in a unit column. So the best possible $\mathbf{Z}$ is $\mathbf{I}$, and translating back gives $\mathbf{Q} = \mathbf{U}\mathbf{V}^* = \mathbf{W}$.

Proof. Fix a singular value decomposition $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$ as in Theorem 13.9.1, so that $\mathbf{W} = \mathbf{U}\mathbf{V}^*$ as in the proof of Theorem 13.10.1. Let $\mathbf{Q}$ be unitary and put $\mathbf{Z} := \mathbf{U}^*\mathbf{Q}\mathbf{V}$, which is unitary; as $\mathbf{Q}$ runs over all unitary matrices so does $\mathbf{Z}$, since $\mathbf{Q} = \mathbf{U}\mathbf{Z}\mathbf{V}^*$. By Lemma 11.7.5, applied with the unitary factors $\mathbf{U}^*$ on the left and $\mathbf{V}$ on the right,

$$\|\mathbf{A} - \mathbf{Q}\|_F = \|\mathbf{U}^*(\mathbf{A} - \mathbf{Q})\mathbf{V}\|_F = \|\mathbf{\Sigma} - \mathbf{Z}\|_F.$$

Expanding the square in the Frobenius inner product $\langle \mathbf{X}, \mathbf{Y} \rangle = \operatorname{tr}(\mathbf{Y}^*\mathbf{X})$ of Chapter 11 §01,

$$\begin{aligned} \|\mathbf{\Sigma} - \mathbf{Z}\|_F^2 &= \|\mathbf{\Sigma}\|_F^2 + \|\mathbf{Z}\|_F^2 - 2\operatorname{Re}\operatorname{tr}(\mathbf{Z}^*\mathbf{\Sigma}) \\ &= \sum_{i=1}^{n} \sigma_i^2 + n - 2\sum_{i=1}^{n} \sigma_i \operatorname{Re} z_{ii}, \end{aligned} \tag{13.10.1}$$

where $\|\mathbf{Z}\|_F^2 = \operatorname{tr}(\mathbf{Z}^*\mathbf{Z}) = \operatorname{tr}\mathbf{I}_n = n$, and $\operatorname{tr}(\mathbf{Z}^*\mathbf{\Sigma}) = \sum_i (\mathbf{Z}^*)_{ii}\sigma_i = \sum_i \overline{z_{ii}}\sigma_i$, whose real part is $\sum_i \sigma_i \operatorname{Re} z_{ii}$.

Column i of $\mathbf{Z}$ is a unit vector, so $|z_{ii}| \le 1$ and hence $\operatorname{Re} z_{ii} \le 1$. Since every $\sigma_i \ge 0$,

$$\sum_{i=1}^{n} \sigma_i \operatorname{Re} z_{ii} \le \sum_{i=1}^{n} \sigma_i,$$

so Eq. 13.10.1 gives $\|\mathbf{A}-\mathbf{Q}\|_F^2 \ge \sum_i \sigma_i^2 + n - 2\sum_i \sigma_i$ for every unitary $\mathbf{Q}$. Taking $\mathbf{Q} = \mathbf{W} = \mathbf{U}\mathbf{V}^*$ makes $\mathbf{Z} = \mathbf{U}^*\mathbf{U}\mathbf{V}^*\mathbf{V} = \mathbf{I}_n$, for which $\operatorname{Re} z_{ii} = 1$ for every i and the bound is attained. Finally $\|\mathbf{A}\|_F = \|\mathbf{U}\mathbf{\Sigma}\mathbf{V}^*\|_F = \|\mathbf{\Sigma}\|_F$ by Lemma 11.7.5, applied once on each side, and the only nonzero entries of $\mathbf{\Sigma}$ are $\sigma_1, \dots, \sigma_n$ on its diagonal, so $\|\mathbf{A}\|_F^2 = \sum_i \sigma_i^2$ (Exercise (C1: The Frobenius norm and the largest singular value)). The minimum value is therefore as stated.

Now suppose $\mathbf{A}$ is invertible, so $\sigma_i > 0$ for every i, and suppose $\mathbf{Q}$ attains the minimum. Then $\sum_i \sigma_i(1 - \operatorname{Re} z_{ii}) = 0$ is a sum of nonnegative terms, so $\operatorname{Re} z_{ii} = 1$ for every i; with $|z_{ii}| \le 1$ this forces $z_{ii} = 1$. Column i of $\mathbf{Z}$ then has a coordinate equal to 1 and total squared length 1, so all its other coordinates vanish and the column is $\mathbf{e}_i$. Hence $\mathbf{Z} = \mathbf{I}_n$ and $\mathbf{Q} = \mathbf{U}\mathbf{Z}\mathbf{V}^* = \mathbf{W}$. This proves the corollary.

■

For $\mathbf{A} = \begin{pmatrix} 3 & 0 \\ 4 & 5 \end{pmatrix}$ of Example (A polar decomposition), with $\sigma_1 = 3\sqrt{5}$, $\sigma_2 = \sqrt{5}$, and $\|\mathbf{A}\|_F^2 = 50$, the nearest unitary matrix is $\mathbf{W} = \frac{1}{\sqrt{5}}\begin{pmatrix} 2 & -1 \\ 1 & 2 \end{pmatrix}$ and the distance to the unitary group is $\sqrt{52 - 8\sqrt{5}}$, about 5.84.

Warning. The corollary is a statement about the **Frobenius** norm. A different way of measuring distance can have a different minimizer, and "nearest" always needs its yardstick named. It is true, though not proved here, that $\mathbf{W}$ also minimizes the distance in every unitarily invariant norm; Chapter 21 takes up that family of norms.

13.10.3 Takagi's factorization

A complex matrix can be symmetric, $\mathbf{A}^\top = \mathbf{A}$, without being Hermitian, $\mathbf{A}^* = \mathbf{A}$. Over $\mathbb{R}$ the two coincide and Chapter 12 handled them; over $\mathbb{C}$ they are genuinely different conditions, and symmetry is much the weaker one. It does not imply normality, it does not imply diagonalizability, and it says nothing about the eigenvalues. Here is the standing example:

$$\mathbf{N} = \begin{pmatrix} 1 & i \\ i & -1 \end{pmatrix}, \qquad \mathbf{N}^\top = \mathbf{N}, \qquad \mathbf{N}^2 = \mathbf{O}.$$

A nonzero matrix with $\mathbf{N}^2 = \mathbf{O}$ has minimal polynomial x^2 and is therefore not diagonalizable, so no version of the spectral theorem can apply to $\mathbf{N}$. Nevertheless something survives, and it is a statement about singular values rather than eigenvalues.

Theorem 13.10.3. Takagi's Factorization

Let $\mathbf{A} \in M_n(\mathbb{C})$ be **symmetric**, $\mathbf{A}^\top = \mathbf{A}$, with singular values $\sigma_1 \ge \cdots \ge \sigma_n \ge 0$. Then

there is a unitary $\mathbf{U} \in M_n(\mathbb{C})$ with

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^{\top}, \qquad \mathbf{\Sigma} = \operatorname{diag}(\sigma_1, \dots, \sigma_n).$$

Read the statement against the two theorems it resembles.

- **Against the singular value decomposition.** Theorem 13.9.1 gives $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$ with two unrelated unitary matrices. Since $\mathbf{U}^{\top} = (\overline{\mathbf{U}})^*$, Takagi's identity says $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}(\overline{\mathbf{U}})^*$: it is a singular value decomposition in which the right factor is forced to be $\mathbf{V} = \overline{\mathbf{U}}$. So the content is not a new diagonal matrix but a relation between $\mathbf{U}$ and $\mathbf{V}$.
- **Against the spectral theorem.** Corollary 12.4.3 writes a normal $\mathbf{A}$ as $\mathbf{U}\mathbf{D}\mathbf{U}^*$, a *similarity*, with the eigenvalues of $\mathbf{A}$ on the diagonal. Takagi writes a symmetric $\mathbf{A}$ as $\mathbf{U}\mathbf{\Sigma}\mathbf{U}^{\top}$, which is a *congruence* $\mathbf{X} \mapsto \mathbf{U}\mathbf{X}\mathbf{U}^{\top}$, with the singular values on the diagonal. The hypotheses are different (symmetric, not normal), the diagonal holds different numbers, and the transpose is not the conjugate transpose.

The difference is visible even for a real matrix. Take $\mathbf{A} = \begin{pmatrix} 3 & 4 \\ 4 & -3 \end{pmatrix}$, which is real symmetric with eigenvalues 5 and -5. Corollary 12.5.2 gives $\mathbf{A} = \mathbf{Q}\operatorname{diag}(5, -5)\mathbf{Q}^{\top}$ with $\mathbf{Q}$ real orthogonal. Takagi insists on the **nonnegative** diagonal $\operatorname{diag}(5, 5)$, and no real $\mathbf{U}$ can achieve that, since $\mathbf{U}\operatorname{diag}(5, 5)\mathbf{U}^{\top} = 5\mathbf{U}\mathbf{U}^{\top} = 5\mathbf{I}$ for real orthogonal $\mathbf{U}$. A genuinely complex $\mathbf{U}$ is needed, and Example (Takagi for a real symmetric matrix) finds one.

Idea. Four steps, and the first two are the polar decomposition doing the work. ① Symmetry makes $\mathbf{A}$ intertwine $\mathbf{A}^*\mathbf{A}$ with its conjugate: $\mathbf{A}(\mathbf{A}^*\mathbf{A}) = \overline{(\mathbf{A}^*\mathbf{A})}\mathbf{A}$. Because the positive square root is a polynomial in the matrix, the same relation passes to $|\mathbf{A}|$. ② That relation, combined with $\mathbf{A}^{\top} = \mathbf{A}$, forces the polar factor $\mathbf{W}$ to be a **symmetric** unitary matrix. ③ A symmetric unitary matrix factors as $\mathbf{V}\mathbf{V}^{\top}$: split it into real and imaginary parts, which turn out to be commuting real symmetric matrices, and use Chapter 12's simultaneous diagonalization (Corollary 12.7.3). ④ Conjugating $\mathbf{A}$ by that $\mathbf{V}$ in the right way produces a matrix that is both symmetric and Hermitian, hence **real** symmetric, and the real spectral theorem in matrix form (Corollary 12.5.2) finishes the job. We prove all of this for invertible $\mathbf{A}$, where the polar factor is unique and $|\mathbf{A}|^{-1}$ exists.

Proof. We prove the theorem for **invertible** $\mathbf{A}$. Throughout, $\mathbf{A}^{\top} = \mathbf{A}$, so $\mathbf{A}^* = \overline{\mathbf{A}^{\top}} = \overline{\mathbf{A}}$. Write $\mathbf{M} \coloneqq \mathbf{A}^*\mathbf{A} = \overline{\mathbf{A}}\mathbf{A}$ and $\mathbf{P} \coloneqq |\mathbf{A}| = \mathbf{M}^{1/2}$. Since $\mathbf{A}$ is invertible, so are $\mathbf{M}$ and $\mathbf{P}$, and $\mathbf{M} \succ \mathbf{O}$.

Step 1: $\mathbf{A}\mathbf{M} = \overline{\mathbf{M}}\mathbf{A}$ and $\mathbf{A}\mathbf{P} = \overline{\mathbf{P}}\mathbf{A}$. For the first, $\overline{\mathbf{M}} = \overline{\overline{\mathbf{A}}\mathbf{A}} = \mathbf{A}\overline{\mathbf{A}}$, so

$$\mathbf{A}\mathbf{M} = \mathbf{A}\overline{\mathbf{A}}\mathbf{A} = \overline{\mathbf{M}}\mathbf{A}.$$

For the second we need a polynomial.

Claim. There is a polynomial p with **real** coefficients such that $p(\mathbf{M}) = \mathbf{P}$ and $p(\overline{\mathbf{M}}) = \overline{\mathbf{P}}$.

Proof. By Theorem 13.2.1 (a) there is a polynomial p with real coefficients such that $p(\mathbf{M}) = \mathbf{M}^{1/2} = \mathbf{P}$. Conjugating every entry of the identity $\mathbf{P} = p(\mathbf{M})$, and using that the coefficients of p are real so that conjugation passes through them, gives $\overline{\mathbf{P}} = p(\overline{\mathbf{M}})$.

■

Iterating Step 1's first identity gives $\mathbf{A}\mathbf{M}^j = \overline{\mathbf{M}}^j\mathbf{A}$ for every $j \geq 0$, hence $\mathbf{A}\,p(\mathbf{M}) = p(\overline{\mathbf{M}})\,\mathbf{A}$ for every polynomial p, and with the claim's p this reads

$$\mathbf{A}\mathbf{P} = \overline{\mathbf{P}}\mathbf{A}. \tag{13.10.2}$$

Step 2: the polar factor is symmetric. By Theorem 13.10.1, $\mathbf{A} = \mathbf{W}\mathbf{P}$ with $\mathbf{W} = \mathbf{A}\mathbf{P}^{-1}$ unitary. Substituting $\mathbf{A} = \mathbf{W}\mathbf{P}$ into Eq. 13.10.2 gives $\mathbf{W}\mathbf{P}^2 = \overline{\mathbf{P}}\mathbf{W}\mathbf{P}$, and canceling the invertible $\mathbf{P}$ on the right,

$$\mathbf{W}\mathbf{P} = \overline{\mathbf{P}}\,\mathbf{W}.$$

On the other hand $\mathbf{P}$ is Hermitian, so $\mathbf{P}^\top = \overline{\mathbf{P}}$, and transposing $\mathbf{A} = \mathbf{W}\mathbf{P}$ while using $\mathbf{A}^\top = \mathbf{A}$ gives

$$\mathbf{W}\mathbf{P} = \mathbf{A} = \mathbf{A}^\top = \mathbf{P}^\top\mathbf{W}^\top = \overline{\mathbf{P}}\,\mathbf{W}^\top.$$

Comparing the two displays, $\overline{\mathbf{P}}\mathbf{W} = \overline{\mathbf{P}}\mathbf{W}^\top$, and $\overline{\mathbf{P}}$ is invertible, so $\mathbf{W}^\top = \mathbf{W}$.

Step 3: a symmetric unitary matrix is $\mathbf{V}\mathbf{V}^\top$.

Claim. Let $\mathbf{W} \in M_n(\mathbb{C})$ be unitary with $\mathbf{W}^\top = \mathbf{W}$. Then $\mathbf{W} = \mathbf{V}\mathbf{V}^\top$ for some unitary $\mathbf{V}$.

Proof. From $\mathbf{W}^{-1} = \mathbf{W}^* = \overline{\mathbf{W}^\top} = \overline{\mathbf{W}}$ we get $\overline{\mathbf{W}}\mathbf{W} = \mathbf{I}$. Write $\mathbf{W} = \mathbf{X} + i\mathbf{Y}$ with $\mathbf{X} = \frac{1}{2}(\mathbf{W} + \overline{\mathbf{W}})$ and $\mathbf{Y} = \frac{1}{2i}(\mathbf{W} - \overline{\mathbf{W}})$, both **real** matrices, and both symmetric because $\mathbf{W}$ and $\overline{\mathbf{W}}$ are. Then

$$\mathbf{I} = \overline{\mathbf{W}}\mathbf{W} = (\mathbf{X} - i\mathbf{Y})(\mathbf{X} + i\mathbf{Y}) = \mathbf{X}^2 + \mathbf{Y}^2 + i(\mathbf{X}\mathbf{Y} - \mathbf{Y}\mathbf{X}),$$

and since $\mathbf{X}^2 + \mathbf{Y}^2$ and $\mathbf{X}\mathbf{Y} - \mathbf{Y}\mathbf{X}$ are real matrices, comparing real and imaginary parts gives

$$\mathbf{X}^2 + \mathbf{Y}^2 = \mathbf{I}, \qquad \mathbf{X}\mathbf{Y} = \mathbf{Y}\mathbf{X}.$$

So $\mathbf{X}$ and $\mathbf{Y}$ are commuting real symmetric matrices, and Corollary 12.7.3 supplies $\mathbf{Q} \in \mathrm{O}(n)$ with $\mathbf{Q}^\top\mathbf{X}\mathbf{Q} = \operatorname{diag}(x_1, \ldots, x_n)$ and $\mathbf{Q}^\top\mathbf{Y}\mathbf{Q} = \operatorname{diag}(y_1, \ldots, y_n)$. Hence $\mathbf{Q}^\top\mathbf{W}\mathbf{Q} = \operatorname{diag}(x_1 + iy_1, \ldots, x_n + iy_n)$, and $x_j^2 + y_j^2 = 1$ for every j by the first relation. Each $x_j + iy_j$ therefore has modulus 1. The polynomial $x^2 - (x_j + iy_j)$ has a root $c_j \in \mathbb{C}$ by Theorem 6.5.1, and $|c_j|^2 = |c_j^2| = 1$, so $|c_j| = 1$. Put $\mathbf{C} = \operatorname{diag}(c_1, \ldots, c_n)$ and $\mathbf{V} := \mathbf{Q}\mathbf{C}$, which is unitary as a product of unitary matrices. Since $\mathbf{Q}$ is real orthogonal, $\mathbf{Q}^\top = \mathbf{Q}^{-1}$, and $\mathbf{C}^\top = \mathbf{C}$, so

$$\mathbf{V}\mathbf{V}^\top = \mathbf{Q}\mathbf{C}\mathbf{C}^\top\mathbf{Q}^\top = \mathbf{Q}\mathbf{C}^2\mathbf{Q}^\top = \mathbf{Q}\,\operatorname{diag}(x_j + iy_j)\,\mathbf{Q}^\top = \mathbf{W}.$$

This proves the claim. ∎

Step 4: conjugate to a real matrix. Take $\mathbf{V}$ from Step 3, so $\mathbf{W} = \mathbf{V}\mathbf{V}^\top$, and set

$$\mathbf{B} := \mathbf{V}^*\mathbf{A}\overline{\mathbf{V}}.$$

Three observations. First, $\mathbf{B}^\top = \mathbf{B}$: using $(\overline{\mathbf{V}})^\top = \mathbf{V}^*$, $(\mathbf{V}^*)^\top = \overline{\mathbf{V}}$, and $\mathbf{A}^\top = \mathbf{A}$,

$$\mathbf{B}^\top = (\overline{\mathbf{V}})^\top\,\mathbf{A}^\top\,(\mathbf{V}^*)^\top = \mathbf{V}^*\mathbf{A}\overline{\mathbf{V}} = \mathbf{B}.$$

Second, $\mathbf{B} \succ \mathbf{O}$: since $\mathbf{A} = \mathbf{W}\mathbf{P} = \mathbf{V}\mathbf{V}^\top\mathbf{P}$ and $\mathbf{V}^*\mathbf{V} = \mathbf{I}$,

$$\mathbf{B} = \mathbf{V}^*\mathbf{V}\mathbf{V}^\top\mathbf{P}\overline{\mathbf{V}} = \mathbf{V}^\top\mathbf{P}\overline{\mathbf{V}} = (\overline{\mathbf{V}})^*\,\mathbf{P}\,(\overline{\mathbf{V}}),$$

a unitary congruence of the positive definite $\mathbf{P}$, so $\mathbf{x}^*\mathbf{B}\mathbf{x} = (\overline{\mathbf{V}}\mathbf{x})^*\mathbf{P}(\overline{\mathbf{V}}\mathbf{x}) > 0$ for $\mathbf{x} \neq \mathbf{0}$; in particular $\mathbf{B}^* = \mathbf{B}$. Third, $\mathbf{B}$ is **real**, because

$$\overline{\mathbf{B}} = \overline{\mathbf{B}^{\top}} = \mathbf{B}^* = \mathbf{B},$$

the first equality using $\mathbf{B}^{\top} = \mathbf{B}$, the second the definition of the conjugate transpose, and the third Hermitian symmetry.

So $\mathbf{B}$ is a real symmetric positive definite matrix. By Corollary 12.5.2 there is $\mathbf{Q}_0 \in \mathrm{O}(n)$ with $\mathbf{B} = \mathbf{Q}_0\mathbf{\Sigma}\mathbf{Q}_0^{\top}$, where $\mathbf{\Sigma}$ is the diagonal matrix of eigenvalues of $\mathbf{B}$, which we list in decreasing order. Finally, using $\mathbf{V}\mathbf{V}^* = \mathbf{I}$ and its conjugate $\overline{\mathbf{V}}\mathbf{V}^{\top} = \mathbf{I}$,

$$\mathbf{V}\mathbf{B}\mathbf{V}^{\top} = \mathbf{V}\mathbf{V}^*\mathbf{A}\overline{\mathbf{V}}\mathbf{V}^{\top} = \mathbf{A},$$

so with $\mathbf{U} \coloneqq \mathbf{V}\mathbf{Q}_0$, which is unitary,

$$\mathbf{A} = \mathbf{V}\mathbf{Q}_0\mathbf{\Sigma}\mathbf{Q}_0^{\top}\mathbf{V}^{\top} = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^{\top}.$$

It remains to identify the diagonal. Conjugate-transposing, $\mathbf{A}^* = \overline{\mathbf{U}}\mathbf{\Sigma}\mathbf{U}^*$, so

$$\mathbf{A}^*\mathbf{A} = \overline{\mathbf{U}}\mathbf{\Sigma}\mathbf{U}^*\mathbf{U}\mathbf{\Sigma}\mathbf{U}^{\top} = \overline{\mathbf{U}}\,\mathbf{\Sigma}^2\,(\overline{\mathbf{U}})^*.$$

This exhibits $\mathbf{\Sigma}^2$ as the eigenvalue list of $\mathbf{A}^*\mathbf{A}$ in decreasing order, so $\mathbf{\Sigma} = \operatorname{diag}(\sigma_1, \dots, \sigma_n)$ by Definition 13.9.1. This proves the theorem for invertible $\mathbf{A}$.

■

Remark. **What is left out.** The theorem as stated is true for every complex symmetric $\mathbf{A}$, singular ones included, and the example below exhibits a singular case. Our proof used invertibility twice: to know that $\mathbf{P}$ is invertible, so that $\mathbf{W} = \mathbf{A}\mathbf{P}^{-1}$ is determined and the cancellations of Step 2 are legal. One way to remove the hypothesis is to apply the proved case to $\mathbf{A} + t\mathbf{I}$, which is symmetric and invertible for all but finitely many t, and let $t \to 0$ along a sequence for which the unitary factors converge — a step that needs the compactness of the unitary group, which is analysis this book has not set up. We state the general theorem and prove the invertible case.

Example. Takagi for a real symmetric matrix

Find a Takagi factorization of $\mathbf{A} = \begin{pmatrix} 3 & 4 \\ 4 & -3 \end{pmatrix}$.

Solution. First the singular values: $\mathbf{A}^*\mathbf{A} = \mathbf{A}^{\top}\mathbf{A} = \mathbf{A}^2 = 25\mathbf{I}$, so $\sigma_1 = \sigma_2 = 5$ and $\mathbf{\Sigma} = 5\mathbf{I}$. We therefore need a unitary $\mathbf{U}$ with

$$\mathbf{A} = 5\,\mathbf{U}\mathbf{U}^{\top}, \qquad \text{that is} \qquad \mathbf{U}\mathbf{U}^{\top} = \tfrac{1}{5}\mathbf{A}.$$

The matrix $\frac{1}{5}\mathbf{A}$ is real symmetric and orthogonal, with eigenvalues 1 and -1. Its eigenvectors: $(\mathbf{A} - 5\mathbf{I})\mathbf{x} = \mathbf{0}$ reads $-2x_1 + 4x_2 = 0$, giving $\mathbf{q}_1 = \frac{1}{\sqrt{5}}(2, 1)$, and orthogonally $\mathbf{q}_2 = \frac{1}{\sqrt{5}}(1, -2)$.

So with $\mathbf{Q} = \frac{1}{\sqrt{5}}\begin{pmatrix} 2 & 1 \\ 1 & -2 \end{pmatrix} \in \mathrm{O}(2)$ we have $\frac{1}{5}\mathbf{A} = \mathbf{Q}\,\mathrm{diag}(1,-1)\mathbf{Q}^{\top}$. The diagonal entries 1 and -1 have square roots 1 and i, so, following Step 3 of Theorem 13.10.3, set

$$\mathbf{U} = \mathbf{Q}\,\mathrm{diag}(1,i) = \frac{1}{\sqrt{5}}\begin{pmatrix} 2 & i \\ 1 & -2i \end{pmatrix}.$$

Then $\mathbf{U}$ is unitary — its columns $\frac{1}{\sqrt{5}}(2,1)$ and $\frac{i}{\sqrt{5}}(1,-2)$ are orthonormal, since multiplying a real orthonormal pair by unit scalars keeps it orthonormal — and

$$\mathbf{U}\mathbf{\Sigma}\mathbf{U}^{\top} = 5\,\mathbf{U}\mathbf{U}^{\top} = \mathbf{Q}\,\mathrm{diag}(5,-5)\mathbf{Q}^{\top} = \mathbf{A}.$$

Note where the complex entries were needed: $\mathrm{diag}(1,i)$ converted the eigenvalue -5 of $\mathbf{A}$ into the singular value $+5$.

For a singular example, return to $\mathbf{N} = \begin{pmatrix} 1 & i \\ i & -1 \end{pmatrix}$. Here $\mathbf{N}^{*}\mathbf{N} = \begin{pmatrix} 2 & 2i \\ -2i & 2 \end{pmatrix}$, with eigenvalues 4 and 0, so $\sigma_1 = 2$ and $\sigma_2 = 0$. With $\mathbf{U} = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & i \\ i & 1 \end{pmatrix}$, which is unitary, one checks $\mathbf{U}\,\mathrm{diag}(2,0)\mathbf{U}^{\top} = 2\,\mathbf{u}_1\mathbf{u}_1^{\top} = \mathbf{N}$, where $\mathbf{u}_1 = \frac{1}{\sqrt{2}}(1,i)$ is the first column of $\mathbf{U}$. The theorem holds here even though our proof did not reach it, and even though $\mathbf{N}$ has no eigenbasis at all.

Warning. **Symmetric is not Hermitian, and $\mathbf{U}^{\top}$ is not $\mathbf{U}^{*}$.** Takagi's hypothesis is $\mathbf{A}^{\top} = \mathbf{A}$ with no conjugation, which over $\mathbb{C}$ is a different and weaker condition than $\mathbf{A}^{*} = \mathbf{A}$; the matrix $\mathbf{N}$ above is symmetric, is not Hermitian, is not normal, and is not diagonalizable. The conclusion likewise carries $\mathbf{U}^{\top}$, not $\mathbf{U}^{*}$: writing $\mathbf{U}\mathbf{\Sigma}\mathbf{U}^{*}$ by mistake would assert that $\mathbf{A}$ is positive semidefinite, which $\mathbf{N}$ certainly is not.

13.10.4 Exercises

13.10.4.1 A. Check your understanding

Exercise. A1

(a) State the polar decomposition, and say which factor is unique and when.

(b) What is $|\mathbf{A}|$, and what identity relates $\||\mathbf{A}|\mathbf{x}\|$ to $\|\mathbf{A}\mathbf{x}\|$?

(c) What does the polar decomposition say when $n = 1$?

(d) True or false: $|\mathbf{A}| = |\mathbf{A}^{*}|$ for every square $\mathbf{A}$. Justify your answer.

(e) State Takagi's factorization, and name two ways in which it differs from the singular value decomposition and from the complex spectral theorem.

(f) True or false: a complex symmetric matrix is diagonalizable. Justify your answer.

Solution.

(a) Every $\mathbf{A} \in M_n(\mathbb{C})$ is $\mathbf{A} = \mathbf{W}|\mathbf{A}|$ with $\mathbf{W}$ unitary (Theorem 13.10.1). The positive semidefinite factor $|\mathbf{A}|$ is always unique; $\mathbf{W}$ is unique exactly when $\mathbf{A}$ is invertible, and

then $\mathbf{W} = \mathbf{A}|\mathbf{A}|^{-1}$.

(b) $|\mathbf{A}| = (\mathbf{A}^*\mathbf{A})^{1/2}$, the unique positive semidefinite square root of $\mathbf{A}^*\mathbf{A}$ (Definition 13.2.1), and $\||\mathbf{A}|\mathbf{x}\| = \|\mathbf{A}\mathbf{x}\|$ for every $\mathbf{x}$.

(c) It says $z = e^{i\theta}|z|$: a complex number is a number of modulus 1 times its modulus, with θ undetermined exactly when $z = 0$.

(d) False in general, true exactly for normal $\mathbf{A}$. For $\mathbf{A} = \begin{pmatrix} 3 & 0 \\ 4 & 5 \end{pmatrix}$, $\mathbf{A}^\top\mathbf{A} = \begin{pmatrix} 25 & 20 \\ 20 & 25 \end{pmatrix} \neq \begin{pmatrix} 9 & 12 \\ 12 & 41 \end{pmatrix} = \mathbf{A}\mathbf{A}^\top$, so the two square roots differ.

(e) Theorem 13.10.3: a symmetric $\mathbf{A} \in M_n(\mathbb{C})$ is $\mathbf{U}\mathbf{\Sigma}\mathbf{U}^\top$ with $\mathbf{U}$ unitary and $\mathbf{\Sigma}$ the diagonal matrix of singular values. Against the singular value decomposition: the right factor is not free but forced to be $\overline{\mathbf{U}}$. Against the spectral theorem: the hypothesis is symmetry rather than normality, the operation is the congruence $\mathbf{X} \mapsto \mathbf{U}\mathbf{X}\mathbf{U}^\top$ rather than a similarity, and the diagonal carries singular values rather than eigenvalues.

(f) False. $\mathbf{N} = \begin{pmatrix} 1 & i \\ i & -1 \end{pmatrix}$ is symmetric and nonzero with $\mathbf{N}^2 = \mathbf{O}$, so its minimal polynomial is x^2, which has a repeated root; it is not diagonalizable.

13.10.4.2 B. Practice

Exercise. B1: A polar decomposition by hand

Find both polar decompositions $\mathbf{A} = \mathbf{W}|\mathbf{A}| = |\mathbf{A}^*|\mathbf{W}$ of

$$\mathbf{A} = \begin{pmatrix} 0 & -2 \\ 3 & 0 \end{pmatrix},$$

and say what $\mathbf{W}$ is geometrically.

Solution. $\mathbf{A}^\top\mathbf{A} = \begin{pmatrix} 9 & 0 \\ 0 & 4 \end{pmatrix}$, which is diagonal with nonnegative entries, so $|\mathbf{A}| = \operatorname{diag}(3, 2)$: this is positive semidefinite and squares to $\mathbf{A}^\top\mathbf{A}$, which by Theorem 13.2.1 identifies it. Since $\mathbf{A}$ is invertible,

$$\mathbf{W} = \mathbf{A}|\mathbf{A}|^{-1} = \begin{pmatrix} 0 & -2 \\ 3 & 0 \end{pmatrix}\begin{pmatrix} \frac{1}{3} & 0 \\ 0 & \frac{1}{2} \end{pmatrix} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix},$$

the quarter turn of $\mathbb{R}^2$. Check: $\mathbf{W}^\top\mathbf{W} = \mathbf{I}$ and $\mathbf{W}|\mathbf{A}| = \mathbf{A}$.

For the right-handed form, $\mathbf{A}\mathbf{A}^\top = \operatorname{diag}(4, 9)$, so $|\mathbf{A}^*| = \operatorname{diag}(2, 3)$, and indeed

$$|\mathbf{A}^*|\mathbf{W} = \begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix}\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 0 & -2 \\ 3 & 0 \end{pmatrix} = \mathbf{A}.$$

The two absolute values are different, which is consistent with $\mathbf{A}$ not being normal. In words: $\mathbf{A}$ stretches by 3 horizontally and 2 vertically, then rotates by a quarter turn; equivalently it rotates first and then stretches by 2 horizontally and 3 vertically.

Exercise. B2: The nearest unitary matrix

Let $\mathbf{A} = \operatorname{diag}(2, -3) \in M_2(\mathbb{R})$. Find the unitary matrix nearest to $\mathbf{A}$ in the Frobenius norm, compute the distance, and verify the value against the formula of Corollary 13.10.2.

Solution. $\mathbf{A}^*\mathbf{A} = \operatorname{diag}(4, 9)$, so $|\mathbf{A}| = \operatorname{diag}(2, 3)$ and $\mathbf{W} = \mathbf{A}|\mathbf{A}|^{-1} = \operatorname{diag}(1, -1)$, which is unitary. By Corollary 13.10.2, $\mathbf{W}$ is the nearest unitary matrix, and it is the only one since $\mathbf{A}$ is invertible. Directly,

$$\mathbf{A} - \mathbf{W} = \operatorname{diag}(1, -2), \qquad \|\mathbf{A} - \mathbf{W}\|_F^2 = 1 + 4 = 5.$$

The formula gives $\|\mathbf{A}\|_F^2 + n - 2(\sigma_1 + \sigma_2) = 13 + 2 - 2(3 + 2) = 5$, the same. So the distance is $\sqrt{5}$.

Exercise. B3: A Takagi factorization

Find a Takagi factorization of $\mathbf{A} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$, and explain why no **real** $\mathbf{U}$ can work.

Solution. $\mathbf{A}$ is symmetric and $\mathbf{A}^*\mathbf{A} = \mathbf{A}^2 = \mathbf{I}$, so $\sigma_1 = \sigma_2 = 1$ and $\mathbf{\Sigma} = \mathbf{I}$. We need a unitary $\mathbf{U}$ with $\mathbf{U}\mathbf{U}^\top = \mathbf{A}$.

Diagonalize $\mathbf{A}$ over $\mathbb{R}$: its eigenvalues are 1 and -1, with unit eigenvectors $\frac{1}{\sqrt{2}}(1, 1)$ and $\frac{1}{\sqrt{2}}(1, -1)$, so $\mathbf{A} = \mathbf{Q}\operatorname{diag}(1, -1)\mathbf{Q}^\top$ with $\mathbf{Q} = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$. Following Step 3 of Theorem 13.10.3, take square roots 1 and i of the diagonal entries:

$$\mathbf{U} = \mathbf{Q}\operatorname{diag}(1, i) = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & i \\ 1 & -i \end{pmatrix}, \qquad \mathbf{U}\mathbf{U}^\top = \mathbf{Q}\operatorname{diag}(1, -1)\mathbf{Q}^\top = \mathbf{A}.$$

$\mathbf{U}$ is unitary, and $\mathbf{U}\mathbf{\Sigma}\mathbf{U}^\top = \mathbf{U}\mathbf{U}^\top = \mathbf{A}$.

No real $\mathbf{U}$ works: a real $\mathbf{U}$ with $\mathbf{U}^*\mathbf{U} = \mathbf{I}$ is orthogonal, and then $\mathbf{U}\mathbf{\Sigma}\mathbf{U}^\top = \mathbf{U}\mathbf{U}^\top = \mathbf{I} \neq \mathbf{A}$.

13.10.4.3 C. Going deeper

Exercise. C1: The positive factor is forced

Let $\mathbf{A} \in M_n(\mathbb{C})$.

(a) Suppose $\mathbf{A} = \mathbf{Q}\mathbf{P}$ with $\mathbf{Q}$ unitary and $\mathbf{P} \succeq \mathbf{O}$. Prove that $\mathbf{P} = |\mathbf{A}|$. (Invertibility is **not** assumed.)

(b) Deduce that $\mathbf{A}$ is unitary if and only if $|\mathbf{A}| = \mathbf{I}$.

Solution.

(a) From $\mathbf{A} = \mathbf{Q}\mathbf{P}$ we get $\mathbf{A}^*\mathbf{A} = \mathbf{P}^*\mathbf{Q}^*\mathbf{Q}\mathbf{P} = \mathbf{P}^*\mathbf{P} = \mathbf{P}^2$, using $\mathbf{Q}^*\mathbf{Q} = \mathbf{I}$ and $\mathbf{P}^* = \mathbf{P}$.

So $\mathbf{P}$ is a positive semidefinite square root of $\mathbf{A}^*\mathbf{A}$. Such a square root is unique (Theorem 13.2.1), and $|\mathbf{A}|$ is one by definition (Definition 13.2.1). Hence $\mathbf{P} = |\mathbf{A}|$.

(b) ($\Rightarrow$) If $\mathbf{A}$ is unitary then $\mathbf{A}^*\mathbf{A} = \mathbf{I}$, and $\mathbf{I} \succeq \mathbf{O}$ with $\mathbf{I}^2 = \mathbf{I}$, so $|\mathbf{A}| = \mathbf{I}$ by uniqueness. ($\Leftarrow$) If $|\mathbf{A}| = \mathbf{I}$ then $\mathbf{A}^*\mathbf{A} = |\mathbf{A}|^2 = \mathbf{I}$, which for a square matrix says $\mathbf{A}$ is unitary (Definition 11.7.2).

Exercise. C2: Normality through the polar factors

Let $\mathbf{A} \in M_n(\mathbb{C})$ be invertible with polar decomposition $\mathbf{A} = \mathbf{W}|\mathbf{A}|$. Prove that the following are equivalent.

(a) $\mathbf{A}$ is normal.

(b) $|\mathbf{A}| = |\mathbf{A}^*|$.

(c) $\mathbf{W}|\mathbf{A}| = |\mathbf{A}|\mathbf{W}$.

Solution.

(a) $\Rightarrow$ (b). If $\mathbf{A}^*\mathbf{A} = \mathbf{A}\mathbf{A}^*$, then $|\mathbf{A}|$ and $|\mathbf{A}^*|$ are positive semidefinite square roots of the same matrix, hence equal by Theorem 13.2.1.

(b) $\Rightarrow$ (c). By Theorem 13.10.1, $|\mathbf{A}^*| = \mathbf{W}|\mathbf{A}|\mathbf{W}^*$. If this equals $|\mathbf{A}|$, multiply on the right by $\mathbf{W}$ to get $|\mathbf{A}|\mathbf{W} = \mathbf{W}|\mathbf{A}|$.

(c) $\Rightarrow$ (a). Write $\mathbf{P} = |\mathbf{A}|$, so $\mathbf{A} = \mathbf{W}\mathbf{P} = \mathbf{P}\mathbf{W}$. Then

$$\mathbf{A}\mathbf{A}^* = \mathbf{W}\mathbf{P}\mathbf{P}^*\mathbf{W}^* = \mathbf{W}\mathbf{P}^2\mathbf{W}^* = \mathbf{P}^2\mathbf{W}\mathbf{W}^* = \mathbf{P}^2,$$

using $\mathbf{P}^* = \mathbf{P}$, then $\mathbf{W}\mathbf{P}^2 = \mathbf{P}^2\mathbf{W}$ (which follows from $\mathbf{W}\mathbf{P} = \mathbf{P}\mathbf{W}$), then $\mathbf{W}\mathbf{W}^* = \mathbf{I}$. And $\mathbf{A}^*\mathbf{A} = \mathbf{P}^*\mathbf{W}^*\mathbf{W}\mathbf{P} = \mathbf{P}^2$. So $\mathbf{A}\mathbf{A}^* = \mathbf{A}^*\mathbf{A}$.

Exercise. C3: Complex symmetric matrices are transposed squares

Let $\mathbf{A} \in M_n(\mathbb{C})$ be symmetric and invertible.

(a) Using Theorem 13.10.3, prove that $\mathbf{A} = \mathbf{B}^\top\mathbf{B}$ for some invertible $\mathbf{B} \in M_n(\mathbb{C})$.

(b) Show that over $\mathbb{R}$ the analogous statement is false unless $\mathbf{A}$ is positive definite, and give a 1×1 witness.

(c) Explain in one sentence why (a) and (b) do not contradict each other.

Solution.

(a) By Theorem 13.10.3, $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^\top$ with $\mathbf{U}$ unitary and $\mathbf{\Sigma} = \operatorname{diag}(\sigma_1, \dots, \sigma_n)$, where every $\sigma_i > 0$ because $\mathbf{A}$ is invertible (Theorem 13.9.1). Let $\mathbf{\Sigma}^{1/2} = \operatorname{diag}(\sqrt{\sigma_1}, \dots, \sqrt{\sigma_n})$ and put $\mathbf{B} := \mathbf{\Sigma}^{1/2}\mathbf{U}^\top$, which is invertible as a product of invertible matrices. Then

$$\mathbf{B}^\top\mathbf{B} = \mathbf{U}\,\mathbf{\Sigma}^{1/2}\mathbf{\Sigma}^{1/2}\,\mathbf{U}^\top = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^\top = \mathbf{A},$$

using $(\mathbf{\Sigma}^{1/2})^\top = \mathbf{\Sigma}^{1/2}$ and $(\mathbf{U}^\top)^\top = \mathbf{U}$.

(b) Over $\mathbb{R}$, if $\mathbf{A} = \mathbf{B}^\top\mathbf{B}$ with $\mathbf{B} \in M_n(\mathbb{R})$ then $\mathbf{x}^\top\mathbf{A}\mathbf{x} = \|\mathbf{B}\mathbf{x}\|^2 \geq 0$ for every $\mathbf{x}$, so

$\mathbf{A} \succeq \mathbf{O}$, and $\mathbf{A} \succ \mathbf{O}$ when $\mathbf{B}$ is invertible. So no real $\mathbf{A}$ with a negative eigenvalue is of this form: for $n = 1$, $\mathbf{A} = (-1)$ is symmetric and invertible but $b^2 = -1$ has no real solution.

(c) The factor $\mathbf{B}$ produced in (a) is complex, and over $\mathbb{C}$ the quantity $\mathbf{x}^\top \mathbf{A}\mathbf{x}$ is not a sum of squared lengths — the transpose carries no conjugation, so $\mathbf{B}^\top \mathbf{B}$ says nothing about positivity. Indeed $b = i$ solves $b^2 = -1$.

13.11 Low-Rank Approximation

The compact singular value decomposition writes a matrix as a sum of rank-one pieces arranged in decreasing order of size. A sum can be cut short. This section shows that cutting it short after k terms produces the **best** approximation to $\mathbf{A}$ by a matrix of rank k, measured in the Frobenius norm, and that the error it leaves is a quantity we can write down before doing any work. One theorem answers three questions at once: how to compress a matrix, which k-dimensional subspace best fits a cloud of data, and how far a given matrix is from being singular.

Throughout, $F = \mathbb{R}$ or $F = \mathbb{C}$, $\mathbf{A} \in M_{m\times n}(F)$, $p = \min(m, n)$, $r = \operatorname{rank} \mathbf{A}$, and

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p \geq 0$$

are the singular values of $\mathbf{A}$ (Definition 13.9.1), so that $\sigma_i > 0$ exactly for $i \leq r$. We fix a singular value decomposition $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$ (Theorem 13.9.1) and write $\mathbf{u}_1, \ldots, \mathbf{u}_m$ for the columns of $\mathbf{U}$ and $\mathbf{v}_1, \ldots, \mathbf{v}_n$ for the columns of $\mathbf{V}$.

13.11.1 Cutting the sum short

Multiplying out $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$ column by column gives the rank-one expansion

$$\mathbf{A} = \sum_{i=1}^{p} \sigma_i \mathbf{u}_i \mathbf{v}_i^* = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^*,$$

the second equality because the terms with $i > r$ have $\sigma_i = 0$; this is the compact form Theorem 13.9.3. Each summand is a matrix of rank one, and the coefficients decrease. The natural thing to do with such a sum is to keep the large terms and discard the small ones.

Keep the first k rank-one pieces and throw the rest away.

Definition 13.11.1. Truncated singular value decomposition

For $0 \leq k \leq p$, the **rank-k truncation** of $\mathbf{A}$ relative to the chosen singular value decomposition $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$ (Theorem 13.9.1) is

$$\mathbf{A}_k := \sum_{i=1}^{k} \sigma_i \mathbf{u}_i \mathbf{v}_i^*,$$

with $\mathbf{A}_0 = \mathbf{O}$; for $k \geq p$ we set $\mathbf{A}_k = \mathbf{A}$.

Two clauses need comment. The subscript counts **terms kept**, not rows and columns: $\mathbf{A}_k$ is the same size as $\mathbf{A}$, not a submatrix of it. (Section 1 used $\mathbf{A}_k$ for the top-left $k \times k$ corner; that meaning is gone for the rest of this section.) And the definition says "relative to the chosen

decomposition", because Theorem 13.9.2 leaves the unitary factors some freedom, and different choices can give different truncations. Everything proved below holds for **every** choice; the warning at the end of the section shows that the dependence is real.

The rank of $\mathbf{A}_k$ is exactly $\min(k, r)$: the vectors $\mathbf{u}_1, \dots, \mathbf{u}_k$ are orthonormal, hence independent, so the columns of $\mathbf{A}_k$ span $\operatorname{span}(\mathbf{u}_1, \dots, \mathbf{u}_{\min(k,r)})$ once the vanishing terms are dropped.

Example. Truncating a diagonal matrix

Let $\mathbf{A} = \operatorname{diag}(5, 3, 1) \in M_3(\mathbb{R})$. Write down $\mathbf{A}_1$ and $\mathbf{A}_2$.

Solution. A diagonal matrix with positive decreasing entries is already in the form $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$ with $\mathbf{U} = \mathbf{V} = \mathbf{I}_3$, so $\sigma_i = 5, 3, 1$ and $\mathbf{u}_i = \mathbf{v}_i = \mathbf{e}_i$. Hence

$$\mathbf{A}_1 = 5\mathbf{e}_1\mathbf{e}_1^* = \operatorname{diag}(5, 0, 0), \qquad \mathbf{A}_2 = \operatorname{diag}(5, 3, 0).$$

Truncation of a diagonal matrix is just deleting its smallest diagonal entries. Everything in this section is this statement, conjugated by two unitary matrices.

The first thing to compute is how much was thrown away.

Lemma 13.11.1. The Frobenius norm counts singular values

Let $\mathbf{A} \in M_{m\times n}(F)$ have singular values $\sigma_1 \geq \cdots \geq \sigma_p$. Then (the first identity is the one Section 9 already obtained, recorded here with a label because the proof of the main theorem uses it repeatedly)

$$\|\mathbf{A}\|_F^2 = \sum_{i=1}^{p} \sigma_i^2, \qquad \|\mathbf{A} - \mathbf{A}_k\|_F^2 = \sum_{i>k} \sigma_i^2$$

for every $0 \leq k \leq p$.

Proof. The Frobenius norm does not see unitary factors (Lemma 11.7.5), so from $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$,

$$\|\mathbf{A}\|_F = \|\mathbf{\Sigma}\mathbf{V}^*\|_F = \|\mathbf{\Sigma}\|_F.$$

The only nonzero entries of $\mathbf{\Sigma}$ are $\sigma_1, \dots, \sigma_p$ on the diagonal, so $\|\mathbf{\Sigma}\|_F^2 = \sum_i \sigma_i^2$.

For the second identity, subtract the first k terms of the rank-one expansion:

$$\mathbf{A} - \mathbf{A}_k = \sum_{i>k} \sigma_i \mathbf{u}_i \mathbf{v}_i^* = \mathbf{U}\mathbf{\Sigma}^{(k)}\mathbf{V}^*,$$

where $\mathbf{\Sigma}^{(k)}$ is $\mathbf{\Sigma}$ with its first k diagonal entries replaced by 0. Applying the first identity to this matrix — the display is a singular value decomposition of it in the sense of Theorem 13.9.1, up to reordering the diagonal — gives $\|\mathbf{A} - \mathbf{A}_k\|_F^2 = \sum_{i>k} \sigma_i^2$.

■

So the truncation error is decided in advance by the singular values, and it is small exactly when the discarded σ_i are small. The content of the section is that no other matrix of rank k does

better.

Quick check. A matrix has singular values $5, 4, 3$. What is the smallest k for which the relative error $\|\mathbf{A} - \mathbf{A}_k\|_F / \|\mathbf{A}\|_F$ is at most $\frac{1}{2}$?

Solution. By Lemma 13.11.1, $\|\mathbf{A}\|_F^2 = 25 + 16 + 9 = 50$. For $k = 1$ the error squared is $16 + 9 = 25$, so the relative error is $\sqrt{25/50} = 1/\sqrt{2} \approx 0.71$, too big. For $k = 2$ the error squared is 9, and $\sqrt{9/50} \approx 0.42 \le \frac{1}{2}$. So $k = 2$. Note that the test is on the *squares*: halving the error means keeping three quarters of $\sum \sigma_i^2$.

13.11.2 How much a subspace can hold

The proof of the theorem needs one inequality, and it is the only place where any work happens. The question it answers is: if we are allowed to keep only an s-dimensional subspace of F^m, how much of $\mathbf{A}$ can we capture? The answer is $\sigma_1^2 + \cdots + \sigma_s^2$, and the best subspace is the span of the leading left singular vectors.

Lemma 13.11.2. An orthonormal list captures at most the leading singular values

Let $\mathbf{A} \in M_{m\times n}(F)$ with singular values $\sigma_1 \ge \cdots \ge \sigma_p$, extended by $\sigma_i = 0$ for $p < i \le m$. Let $\mathbf{q}_1, \dots, \mathbf{q}_s$ be an orthonormal list in F^m, so that $s \le m$. Then

$$\sum_{j=1}^{s} \|\mathbf{A}^* \mathbf{q}_j\|^2 \;\le\; \sum_{i=1}^{s} \sigma_i^2.$$

Idea. Write each $\mathbf{q}_j$ in the orthonormal basis $(\mathbf{u}_1, \dots, \mathbf{u}_m)$. The quantity $\|\mathbf{A}^*\mathbf{q}_j\|^2$ then becomes $\sum_i \sigma_i^2 |c_{ji}|^2$, so the total is $\sum_i \sigma_i^2 t_i$ with weights $t_i = \sum_j |c_{ji}|^2$. Two facts pin the weights down: each $t_i \le 1$, because the $\mathbf{q}_j$ are orthonormal and Bessel's inequality applies to $\mathbf{u}_i$; and $\sum_i t_i = s$, because each $\mathbf{q}_j$ is a unit vector. A weighted sum of a decreasing sequence with weights in $[0, 1]$ that sum to s is largest when the first s weights are 1 — which is the last display of the proof.

Proof. If $s = 0$ both sides are empty sums, so assume $s \ge 1$. Fix j and put $c_{ji} = \langle \mathbf{q}_j, \mathbf{u}_i \rangle$ for $1 \le i \le m$, so that the i-th entry of $\mathbf{U}^*\mathbf{q}_j$ is c_{ji}. Since $\mathbf{A}^* = \mathbf{V}\mathbf{\Sigma}^*\mathbf{U}^*$ and $\mathbf{V}$ is unitary, hence an isometry (Definition 11.7.2, Theorem 11.7.3),

$$\|\mathbf{A}^*\mathbf{q}_j\|^2 = \|\mathbf{\Sigma}^*\mathbf{U}^*\mathbf{q}_j\|^2 = \sum_{i=1}^{m} \sigma_i^2 |c_{ji}|^2,$$

because $\mathbf{\Sigma}^*$ multiplies the i-th entry by σ_i and discards the entries beyond the p-th, which is the same as using $\sigma_i = 0$ there. Summing over j and exchanging the two finite sums,

$$\sum_{j=1}^{s} \|\mathbf{A}^*\mathbf{q}_j\|^2 = \sum_{i=1}^{m} \sigma_i^2 t_i, \qquad t_i := \sum_{j=1}^{s} |c_{ji}|^2.$$

Now bound the weights. Since $|c_{ji}| = |\langle \mathbf{u}_i, \mathbf{q}_j \rangle|$ and $(\mathbf{q}_1, \dots, \mathbf{q}_s)$ is an orthonormal list, Bessel's inequality (Theorem 11.9.1) applied to the vector $\mathbf{u}_i$ gives

$$t_i = \sum_{j=1}^{s} |\langle \mathbf{u}_i, \mathbf{q}_j \rangle|^2 \;\le\; \|\mathbf{u}_i\|^2 = 1.$$

Since $(\mathbf{u}_1, \dots, \mathbf{u}_m)$ is an orthonormal **basis** of F^m, Parseval's identity (Theorem 11.9.2) gives $\sum_{i=1}^{m} |c_{ji}|^2 = \|\mathbf{q}_j\|^2 = 1$ for each j, and therefore

$$\sum_{i=1}^{m} t_i = \sum_{j=1}^{s} \sum_{i=1}^{m} |c_{ji}|^2 = s.$$

Finally, using $\sigma_i^2 \ge \sigma_s^2$ together with $t_i - 1 \le 0$ for $i \le s$, and $\sigma_i^2 \le \sigma_s^2$ together with $t_i \ge 0$ for $i > s$,

$$\begin{aligned} \sum_{i=1}^{m} \sigma_i^2 t_i - \sum_{i=1}^{s} \sigma_i^2 &= \sum_{i \le s} \sigma_i^2 (t_i - 1) + \sum_{i > s} \sigma_i^2 t_i \\ &\le \sigma_s^2 \Big(\sum_{i \le s} (t_i - 1) + \sum_{i > s} t_i \Big) \\ &= \sigma_s^2 \Big(\sum_{i=1}^{m} t_i - s \Big) = 0. \end{aligned}$$

This is the stated inequality.

■

Equality is attainable: take $\mathbf{q}_j = \mathbf{u}_j$ for $j \le s$. Then $\mathbf{A}^* \mathbf{u}_j = \sigma_j \mathbf{v}_j$ by the rank-one expansion, so $\|\mathbf{A}^* \mathbf{u}_j\|^2 = \sigma_j^2$ and the two sides agree. The lemma is the extreme case of a family of statements about sums of the largest eigenvalues of a Hermitian matrix; Chapter 17 develops the family through the Courant–Fischer min–max theorem, of which the inequality above is the crudest and most useful consequence.

13.11.3 The Eckart–Young theorem

Now the theorem. Read it as a statement about a minimization problem: over the set of matrices of rank at most k — which is not a subspace, so no earlier optimization theorem applies to it directly — the minimum of $\|\mathbf{A} - \mathbf{B}\|_F$ exists, is attained at a matrix we can write down, and has a value we already computed.

Theorem 13.11.3. Eckart–Young Theorem, Frobenius case

Let $\mathbf{A} \in M_{m \times n}(F)$ with singular values $\sigma_1 \ge \cdots \ge \sigma_p$, and let $0 \le k \le p$. Then for **every** $\mathbf{B} \in M_{m \times n}(F)$ with $\operatorname{rank} \mathbf{B} \le k$,

$$\|\mathbf{A} - \mathbf{B}\|_F \;\ge\; \Big(\sum_{i > k} \sigma_i^2 \Big)^{1/2},$$

and equality holds for $\mathbf{B} = \mathbf{A}_k$. In words: the truncation $\mathbf{A}_k$ minimizes $\|\mathbf{A} - \mathbf{B}\|_F$ over all matrices of rank at most k, and the minimum value is $\big(\sum_{i>k} \sigma_i^2\big)^{1/2}$.

Idea. The difficulty is that $\mathbf{B}$ is an arbitrary matrix of rank $\le k$, and there is no formula for such a thing. The rescue is to stop looking at $\mathbf{B}$ and look only at its column space S, a subspace of dimension $s \le k$. ① Every column of $\mathbf{B}$ lies in S, so column by column the Best Approximation Theorem says $\mathbf{B}$ cannot beat $P_S\mathbf{A}$, the matrix obtained by projecting every column of $\mathbf{A}$ onto S. So it suffices to minimize over subspaces. ② For a subspace, Pythagoras splits the error into "all of $\mathbf{A}$" minus "the part S captures", and the captured part is exactly the quantity Lemma 13.11.2 bounds. ③ The bound is $\sigma_1^2 + \cdots + \sigma_k^2$, and subtracting it from $\|\mathbf{A}\|_F^2 = \sum_i \sigma_i^2$ leaves $\sum_{i>k} \sigma_i^2$, which Lemma 13.11.1 says is exactly what $\mathbf{A}_k$ achieves.

Proof. **Step 1: reduce to a projection.** Let $\mathbf{B}$ have rank $\le k$ and put $S = \operatorname{col}(\mathbf{B})$, so $s := \dim S = \operatorname{rank}\mathbf{B} \le k$. Write $\mathbf{a}_1, \dots, \mathbf{a}_n$ and $\mathbf{b}_1, \dots, \mathbf{b}_n$ for the columns of $\mathbf{A}$ and of $\mathbf{B}$. By the definition of the Frobenius norm, summing squares of entries column by column,

$$\|\mathbf{A} - \mathbf{B}\|_F^2 = \sum_{j=1}^{n} \|\mathbf{a}_j - \mathbf{b}_j\|^2.$$

Each $\mathbf{b}_j$ lies in S, so Theorem 11.3.5 gives $\|\mathbf{a}_j - \mathbf{b}_j\| \ge \|\mathbf{a}_j - P_S\mathbf{a}_j\|$ for every j, where P_S is the orthogonal projection onto S (Definition 11.3.2), identified throughout with its matrix in the standard basis of F^m. Hence

$$\|\mathbf{A} - \mathbf{B}\|_F^2 \;\ge\; \sum_{j=1}^{n} \|\mathbf{a}_j - P_S\mathbf{a}_j\|^2 = \|\mathbf{A} - P_S\mathbf{A}\|_F^2,$$

the last equality because the j-th column of $P_S\mathbf{A}$ is $P_S\mathbf{a}_j$.

Step 2: what a subspace captures. Choose an orthonormal basis $(\mathbf{q}_1, \dots, \mathbf{q}_s)$ of S. For each j, the vectors $\mathbf{a}_j - P_S\mathbf{a}_j$ and $P_S\mathbf{a}_j$ are orthogonal, so Theorem 11.1.3 gives $\|\mathbf{a}_j - P_S\mathbf{a}_j\|^2 = \|\mathbf{a}_j\|^2 - \|P_S\mathbf{a}_j\|^2$, and summing over j,

$$\|\mathbf{A} - P_S\mathbf{A}\|_F^2 = \|\mathbf{A}\|_F^2 - \|P_S\mathbf{A}\|_F^2.$$

By Theorem 11.3.3 (a), $P_S\mathbf{a}_j = \sum_{l \le s} \langle \mathbf{a}_j, \mathbf{q}_l \rangle \mathbf{q}_l$, whose summands are pairwise orthogonal, so Theorem 11.1.3 gives $\|P_S\mathbf{a}_j\|^2 = \sum_{l=1}^{s} |\langle \mathbf{a}_j, \mathbf{q}_l \rangle|^2$. Exchanging the sums,

$$\|P_S\mathbf{A}\|_F^2 = \sum_{l=1}^{s} \sum_{j=1}^{n} |\langle \mathbf{a}_j, \mathbf{q}_l \rangle|^2 = \sum_{l=1}^{s} \|\mathbf{A}^* \mathbf{q}_l\|^2,$$

where the last equality holds because $\langle \mathbf{a}_j, \mathbf{q}_l \rangle = \mathbf{q}_l^* \mathbf{a}_j$ is the j-th entry of the row vector $\mathbf{q}_l^* \mathbf{A}$, whose entries are the conjugates of those of $\mathbf{A}^* \mathbf{q}_l$.

Step 3: apply the bound. By Lemma 13.11.2 and $s \le k$, and since all $\sigma_i^2 \ge 0$,

$$\|P_S\mathbf{A}\|_F^2 \;\le\; \sum_{i=1}^{s} \sigma_i^2 \;\le\; \sum_{i=1}^{k} \sigma_i^2.$$

Combining the three steps with $\|\mathbf{A}\|_F^2 = \sum_i \sigma_i^2$ (Lemma 13.11.1),

$$\|\mathbf{A} - \mathbf{B}\|_F^2 \;\ge\; \|\mathbf{A}\|_F^2 - \sum_{i=1}^{k} \sigma_i^2 = \sum_{i>k} \sigma_i^2.$$

Equality. The matrix $\mathbf{A}_k$ has rank $\min(k, r) \le k$, and $\|\mathbf{A} - \mathbf{A}_k\|_F^2 = \sum_{i>k} \sigma_i^2$ by Lemma 13.11.1. So the lower bound is attained, and it is the minimum. This proves the theorem.

■

Three remarks are worth extracting. First, the theorem is a genuine minimum, not an infimum: the minimizing matrix is exhibited. Second, the minimum value depends on $\mathbf{A}$ only through its singular values, so it is unchanged if $\mathbf{A}$ is replaced by $\mathbf{U}'\mathbf{A}\mathbf{V}'$ with $\mathbf{U}', \mathbf{V}'$ unitary. Third, the case $k = r - 1$ reads: the Frobenius distance from $\mathbf{A}$ to the matrices of rank $< r$ is σ_r, the smallest nonzero singular value. Exercise C1 turns that into a statement about how far an invertible matrix is from being singular.

Remark. The same matrix $\mathbf{A}_k$ is also optimal in the **spectral norm** $\|\cdot\|_2$, where the minimum value is σ_{k+1} rather than $\left(\sum_{i>k} \sigma_i^2\right)^{1/2}$. We cannot state that here, because the spectral norm is defined only in Chapter 16. In fact both statements are instances of one theorem: $\mathbf{A}_k$ is a best rank-k approximation in **every** unitarily invariant norm. Chapter 21 proves that, and the proof is genuinely different from the one above — it needs inequalities between the singular values of $\mathbf{A}$ and of $\mathbf{A} - \mathbf{B}$, not a projection argument.

13.11.4 A worked example

Example. The best rank-one approximation of a 3×2 matrix

Let

$$\mathbf{A} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{pmatrix} \in M_{3\times 2}(\mathbb{R}).$$

Find the best rank-one approximation of $\mathbf{A}$ in the Frobenius norm, and compute the error in two ways.

Solution. Singular values and vectors. Since $\mathbf{A}$ has more rows than columns, work with the 2×2 matrix

$$\mathbf{A}^{\top}\mathbf{A} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix},$$

whose eigenvalues are 3 and 1 with orthonormal eigenvectors

$$\mathbf{v}_1 = \tfrac{1}{\sqrt{2}}(1, 1), \qquad \mathbf{v}_2 = \tfrac{1}{\sqrt{2}}(1, -1).$$

Hence $\sigma_1 = \sqrt{3}$ and $\sigma_2 = 1$, and the left singular vectors are $\mathbf{u}_i = \mathbf{A}\mathbf{v}_i/\sigma_i$:

$$\mathbf{u}_1 = \tfrac{1}{\sqrt{6}}(1, 1, 2), \qquad \mathbf{u}_2 = \tfrac{1}{\sqrt{2}}(1, -1, 0).$$

(Both are unit vectors, as they must be: $1 + 1 + 4 = 6$ and $1 + 1 = 2$.)

The approximation. By Theorem 13.11.3 the answer is $\mathbf{A}_1 = \sigma_1 \mathbf{u}_1 \mathbf{v}_1^{\top}$. The scalar in front is

$\sqrt{3}/(\sqrt{6}\sqrt{2}) = \sqrt{3}/\sqrt{12} = \frac{1}{2}$, so

$$\mathbf{A}_1 = \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 2 & 2 \end{pmatrix}.$$

It has rank one, its two columns being equal and nonzero.

The error, twice. Directly,

$$\mathbf{A} - \mathbf{A}_1 = \frac{1}{2}\begin{pmatrix} 1 & -1 \\ -1 & 1 \\ 0 & 0 \end{pmatrix}, \qquad \|\mathbf{A} - \mathbf{A}_1\|_F^2 = 4 \cdot \tfrac{1}{4} = 1.$$

From the theorem, the error squared is $\sigma_2^2 = 1$. The two agree. As a further check, $\|\mathbf{A}\|_F^2 = 1+1+1+1 = 4$ and $\sigma_1^2 + \sigma_2^2 = 3 + 1 = 4$, as Lemma 13.11.1 requires; the relative error of the rank-one approximation is $\sqrt{1/4} = \frac{1}{2}$.

13.11.5 Best-fitting subspaces

Step 1 of the proof did more than its job. It showed that the optimal $\mathbf{B}$ may be taken of the form $P_S\mathbf{A}$, and Steps 2 and 3 identified the best S. That identification is a theorem in its own right, and it is the statement people mean when they speak of fitting a low-dimensional structure to data.

Corollary 13.11.4. The best-fitting subspace

Let $\mathbf{a}_1, \dots, \mathbf{a}_n \in F^m$ be the columns of $\mathbf{A}$, and let $1 \le k \le p$. Among all subspaces $S \subseteq F^m$ with $\dim S \le k$, the total squared distance

$$D(S) \coloneqq \sum_{j=1}^{n} d(\mathbf{a}_j, S)^2$$

is smallest for $S = \operatorname{span}(\mathbf{u}_1, \dots, \mathbf{u}_k)$, and its smallest value is $\sum_{i>k} \sigma_i^2$.

Proof. By Definition 11.3.3, $d(\mathbf{a}_j, S) = \|\mathbf{a}_j - P_S\mathbf{a}_j\|$, so $D(S) = \|\mathbf{A} - P_S\mathbf{A}\|_F^2$ as in Step 1 of Theorem 13.11.3. Since $\operatorname{rank}(P_S\mathbf{A}) \le \dim S \le k$, the theorem gives $D(S) \ge \sum_{i>k} \sigma_i^2$. For $S_0 = \operatorname{span}(\mathbf{u}_1, \dots, \mathbf{u}_k)$ the columns $\mathbf{u}_1, \dots, \mathbf{u}_k$ are orthonormal, so Theorem 11.3.3 (a) gives

$$P_{S_0}\mathbf{A} = \sum_{i=1}^{k} \mathbf{u}_i \mathbf{u}_i^* \mathbf{A} = \sum_{i=1}^{k} \sigma_i \mathbf{u}_i \mathbf{v}_i^* = \mathbf{A}_k,$$

using $\mathbf{u}_i^*\mathbf{A} = \sigma_i \mathbf{v}_i^*$, which is the rank-one expansion read one row at a time. Hence $D(S_0) = \|\mathbf{A} - \mathbf{A}_k\|_F^2 = \sum_{i>k} \sigma_i^2$.

■

What this is used for, honestly. Two applications are within reach of what we have proved, and a third is not.

The first is **compression**. Storing $\mathbf{A}$ costs mn numbers; storing $\mathbf{A}_k$ as $\sigma_1, \mathbf{u}_1, \mathbf{v}_1, \dots, \sigma_k, \mathbf{u}_k, \mathbf{v}_k$

costs $k(m+n+1)$. When k is much smaller than m and n that is a large saving, and Theorem 13.11.3 says no other rank-k object stored the same way is closer to $\mathbf{A}$ in the Frobenius norm. The relative error is $\left(\sum_{i>k}\sigma_i^2/\sum_i\sigma_i^2\right)^{1/2}$, computable before choosing k. Whether that error is acceptable is not a mathematical question.

The second is **principal components**. Take n data points in $\mathbb{R}^m$ and place them as the columns of $\mathbf{A}$. Then Corollary 13.11.4 says: the k-dimensional subspace through the origin that minimizes the total squared distance to the data is spanned by $\mathbf{u}_1,\dots,\mathbf{u}_k$, the leading left singular vectors. These directions are the **principal components** of the data, and the numbers σ_i^2 measure how much of $\sum_j\|\mathbf{a}_j\|^2$ each one accounts for. That is a theorem of linear algebra, and it is the whole of what this book has proved.

What is *not* proved here is the statistics. The usual procedure subtracts the mean of the data first, so that $\frac{1}{n}\mathbf{A}\mathbf{A}^\top$ becomes a sample covariance matrix, and then interprets $\sigma_i^2/\sum_l\sigma_l^2$ as a fraction of variance explained. Those steps need a probabilistic model, a reason to prefer squared error, and an argument that the sample answer says something about a larger population. Chapter 24 §10 supplies the first two — the model for the finite population of the data themselves, where expectation is an average and no measure theory is needed, and three reasons to prefer squared error — and says plainly that the third belongs to probability and statistics, which this book does not develop. It also says why a singular value decomposition is never computed by forming $\mathbf{A}^*\mathbf{A}$ and finding its eigenvalues, the route taken in every hand computation in this chapter; the algorithm that replaces it is assembled nowhere in this book.

Warning. **The truncation is optimal for this norm, and low rank does not mean small.** Two mistakes to avoid.

Other measures of error give other answers. Let $\mathbf{A}=\operatorname{diag}(2,1)$, so $\sigma_1=2$, $\sigma_2=1$, and $\mathbf{A}_1=\operatorname{diag}(2,0)$, whose error matrix $\operatorname{diag}(0,1)$ has a largest entry of modulus 1. The rank-one matrix

$$\mathbf{B}=\frac{1}{2}\begin{pmatrix}3&\sqrt{3}\\ \sqrt{3}&1\end{pmatrix},\qquad \mathbf{A}-\mathbf{B}=\frac{1}{2}\begin{pmatrix}1&-\sqrt{3}\\ -\sqrt{3}&1\end{pmatrix},$$

has $\det\mathbf{B}=\frac{3}{4}-\frac{3}{4}=0$, so $\operatorname{rank}\mathbf{B}=1$, and every entry of $\mathbf{A}-\mathbf{B}$ has modulus at most $\sqrt{3}/2\approx 0.87<1$. So $\mathbf{B}$ beats $\mathbf{A}_1$ if the error is measured entrywise. It loses in the Frobenius norm, as it must: $\|\mathbf{A}-\mathbf{B}\|_F^2=2>1=\|\mathbf{A}-\mathbf{A}_1\|_F^2$.

Small rank and small size are unrelated. The $n\times n$ all-ones matrix $\mathbf{J}$ has rank 1 and $\|\mathbf{J}\|_F=n$, while $\varepsilon\mathbf{I}_n$ has rank n and $\|\varepsilon\mathbf{I}_n\|_F=\varepsilon\sqrt{n}$, as small as we like. "Approximating by a matrix of low rank" is a statement about structure, never about magnitude.

13.11.6 Exercises

13.11.6.1 A. Check your understanding

Exercise. A1

(a) State the Eckart–Young Theorem, including the value of the minimum.

(b) Express $\|\mathbf{A}\|_F$ in terms of the singular values of $\mathbf{A}$.

(c) Let $\operatorname{rank}\mathbf{A}=r$. What is $\mathbf{A}_k$ for $k\ge r$, and what is the minimum of $\|\mathbf{A}-\mathbf{B}\|_F$ over $\operatorname{rank}\mathbf{B}\le k$ in that case?

(d) True or false: the truncation $\mathbf{A}_k$ is the *unique* matrix of rank at most k nearest to $\mathbf{A}$ in the Frobenius norm. Justify your answer.

(e) A matrix of rank 1 is close to $\mathbf{O}$ in the Frobenius norm. True or false? Justify your answer.

Solution.

(a) For $\mathbf{A} \in M_{m\times n}(F)$ with singular values $\sigma_1 \geq \cdots \geq \sigma_p$ and $0 \leq k \leq p$: every $\mathbf{B}$ with $\operatorname{rank}\mathbf{B} \leq k$ satisfies $\|\mathbf{A}-\mathbf{B}\|_F \geq \left(\sum_{i>k}\sigma_i^2\right)^{1/2}$, with equality for $\mathbf{B} = \mathbf{A}_k = \sum_{i\leq k}\sigma_i \mathbf{u}_i\mathbf{v}_i^*$.

(b) $\|\mathbf{A}\|_F = \left(\sum_{i=1}^{p}\sigma_i^2\right)^{1/2}$ (Lemma 13.11.1).

(c) $\mathbf{A}_k = \mathbf{A}$, since the terms with $i > r$ are zero; the minimum is 0, attained at $\mathbf{B} = \mathbf{A}$ itself, which has rank $r \leq k$.

(d) False in general. Take $\mathbf{A} = \mathbf{I}_2$, so $\sigma_1 = \sigma_2 = 1$. Both $\mathbf{B}_1 = \operatorname{diag}(1,0)$ and $\mathbf{B}_2 = \frac{1}{2}\left(\begin{smallmatrix}1&1\\1&1\end{smallmatrix}\right)$ have rank 1, and $\|\mathbf{I}_2-\mathbf{B}_1\|_F^2 = 1 = \|\mathbf{I}_2-\mathbf{B}_2\|_F^2$, which is the minimum σ_2^2. (Both are truncations of $\mathbf{I}_2$, for two different singular value decompositions.)

(e) False. Rank is unrelated to size: the all-ones matrix $\mathbf{J} \in M_n(\mathbb{R})$ has rank 1 and $\|\mathbf{J}\|_F = n$.

13.11.6.2 B. Practice

Exercise. B1: Errors from the singular values alone

A matrix $\mathbf{A}$ has singular values $6, 3, 2, 1$.

(a) Compute $\|\mathbf{A}\|_F$ and the minimum of $\|\mathbf{A}-\mathbf{B}\|_F$ over matrices $\mathbf{B}$ of rank at most k, for $k = 1, 2, 3$.

(b) Find the smallest k for which the relative error $\|\mathbf{A}-\mathbf{A}_k\|_F/\|\mathbf{A}\|_F$ is at most 0.3.

Solution.

(a) $\|\mathbf{A}\|_F^2 = 36+9+4+1 = 50$, so $\|\mathbf{A}\|_F = \sqrt{50} = 5\sqrt{2}$. By Theorem 13.11.3 the minima are

$$\sqrt{9+4+1} = \sqrt{14}, \qquad \sqrt{4+1} = \sqrt{5}, \qquad \sqrt{1} = 1$$

for $k = 1, 2, 3$.

(b) The condition is $\sum_{i>k}\sigma_i^2 \leq 0.09 \cdot 50 = 4.5$. For $k = 2$ the left side is $5 > 4.5$; for $k = 3$ it is $1 \leq 4.5$. So $k = 3$.

Exercise. B2: A best rank-one approximation

Let $\mathbf{A} = \begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix} \in M_2(\mathbb{R})$. Find the singular values of $\mathbf{A}$, the best rank-one approximation $\mathbf{A}_1$ in the Frobenius norm, and the error, computed both directly and from the theorem.

Solution. $\mathbf{A}$ is symmetric with eigenvalues 4 and 2 (trace 6, determinant 8) and orthonormal eigenvectors $\mathbf{q}_1 = \frac{1}{\sqrt{2}}(1,1)$, $\mathbf{q}_2 = \frac{1}{\sqrt{2}}(1,-1)$. Both eigenvalues are positive, so $\mathbf{A} \succ \mathbf{O}$, and $\mathbf{A}^\top \mathbf{A} = \mathbf{A}^2$ has eigenvalues 16 and 4; hence $\sigma_1 = 4$, $\sigma_2 = 2$, with $\mathbf{u}_i = \mathbf{v}_i = \mathbf{q}_i$.

Therefore

$$\mathbf{A}_1 = 4\mathbf{q}_1\mathbf{q}_1^\top = 4 \cdot \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 2 & 2 \\ 2 & 2 \end{pmatrix},$$

which has rank one. Directly,

$$\mathbf{A} - \mathbf{A}_1 = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}, \qquad \|\mathbf{A} - \mathbf{A}_1\|_F^2 = 4.$$

From Theorem 13.11.3 the error squared is $\sigma_2^2 = 4$. The two agree, and $\|\mathbf{A}\|_F^2 = 9+1+1+9 = 20 = 16 + 4$ checks Lemma 13.11.1.

Exercise. B3: Pythagoras for the truncation

Let $\mathbf{A} \in M_{m\times n}(F)$ and $0 \le k \le p$. Prove that

$$\|\mathbf{A}\|_F^2 = \|\mathbf{A}_k\|_F^2 + \|\mathbf{A} - \mathbf{A}_k\|_F^2,$$

and deduce that $k \mapsto \|\mathbf{A} - \mathbf{A}_k\|_F$ is nonincreasing.

Solution. By Lemma 13.11.1 applied to $\mathbf{A}$ and to $\mathbf{A} - \mathbf{A}_k$, and by the same lemma applied to $\mathbf{A}_k$, whose singular values are $\sigma_1, \dots, \sigma_k$ followed by zeros (it equals $\mathbf{U}\mathbf{\Sigma}'\mathbf{V}^*$ with $\mathbf{\Sigma}'$ the matrix keeping only the first k diagonal entries of $\mathbf{\Sigma}$),

$$\|\mathbf{A}_k\|_F^2 + \|\mathbf{A} - \mathbf{A}_k\|_F^2 = \sum_{i \le k} \sigma_i^2 + \sum_{i > k} \sigma_i^2 = \|\mathbf{A}\|_F^2.$$

Since $\|\mathbf{A} - \mathbf{A}_k\|_F^2 = \sum_{i>k} \sigma_i^2$ and the σ_i^2 are nonnegative, increasing k by one removes a nonnegative term, so the sequence is nonincreasing. This proves both claims.

13.11.6.3 C. Going deeper

Exercise. C1: How far is a matrix from being singular?

Let $\mathbf{A} \in M_n(F)$ be invertible, with singular values $\sigma_1 \ge \cdots \ge \sigma_n > 0$.

(a) Prove that $\|\mathbf{A} - \mathbf{B}\|_F \ge \sigma_n$ for every singular $\mathbf{B} \in M_n(F)$.

(b) Exhibit a singular $\mathbf{B}$ with $\|\mathbf{A} - \mathbf{B}\|_F = \sigma_n$.

(c) Deduce that the distance from $\mathbf{A}$ to the set of singular matrices, measured in the Frobenius norm, is exactly σ_n, and illustrate this with $\mathbf{A} = \begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix}$.

Solution.

(a) A singular $\mathbf{B} \in M_n(F)$ has $\operatorname{rank} \mathbf{B} \le n - 1$ (Theorem 7.5.3 (d)). By Theorem 13.11.3

with $k = n - 1$,

$$\|\mathbf{A} - \mathbf{B}\|_F \;\geq\; \Big(\sum_{i>n-1} \sigma_i^2\Big)^{1/2} = \sigma_n.$$

(b) Take $\mathbf{B} = \mathbf{A}_{n-1} = \sum_{i \leq n-1} \sigma_i \mathbf{u}_i \mathbf{v}_i^*$. Its rank is $n - 1 < n$, so it is singular, and $\|\mathbf{A} - \mathbf{B}\|_F = \sigma_n$ by Lemma 13.11.1.

(c) By (a) the distance is at least σ_n and by (b) it is at most σ_n, so it equals σ_n; in particular the minimum is attained. For the given $\mathbf{A}$, Exercise B2 found $\sigma_2 = 2$ and $\mathbf{A}_1 = \left(\begin{smallmatrix} 2 & 2 \\ 2 & 2 \end{smallmatrix}\right)$, which is singular and at Frobenius distance 2. So no singular matrix is closer to $\mathbf{A}$ than 2. The smallest singular value is therefore a quantitative version of invertibility: $\mathbf{A}$ is invertible exactly when $\sigma_n > 0$, and it is *far* from singular exactly when σ_n is large.

Exercise. C2: The best line through a cloud of points

The four points $(1, 1), (2, 2), (1, -1)$ and $(0, 0)$ of $\mathbb{R}^2$ are to be fitted by a line through the origin, minimizing the total squared distance from the points to the line.

(a) Write the points as the columns of a matrix $\mathbf{A}$ and compute $\mathbf{A}\mathbf{A}^\top$ and the singular values of $\mathbf{A}$.

(b) Use Corollary 13.11.4 to find the best line and the minimum total squared distance.

(c) Verify the answer by computing the four distances directly.

Solution.

(a) With the points as columns,

$$\mathbf{A} = \begin{pmatrix} 1 & 2 & 1 & 0 \\ 1 & 2 & -1 & 0 \end{pmatrix}, \qquad \mathbf{A}\mathbf{A}^\top = \begin{pmatrix} 6 & 4 \\ 4 & 6 \end{pmatrix}.$$

The eigenvalues of $\mathbf{A}\mathbf{A}^\top$ are 10 and 2 (trace 12, determinant 20), with orthonormal eigenvectors $\mathbf{u}_1 = \frac{1}{\sqrt{2}}(1, 1)$ and $\mathbf{u}_2 = \frac{1}{\sqrt{2}}(1, -1)$. The eigenvalues of $\mathbf{A}\mathbf{A}^\top$ are the squares of the singular values, since $\mathbf{A}\mathbf{A}^\top = \mathbf{U}\mathbf{\Sigma}\mathbf{\Sigma}^*\mathbf{U}^*$; hence $\sigma_1 = \sqrt{10}$ and $\sigma_2 = \sqrt{2}$.

(b) By Corollary 13.11.4 with $k = 1$, the best line is $S_0 = \operatorname{span}(\mathbf{u}_1)$, the line $y = x$, and the minimum total squared distance is $\sigma_2^2 = 2$.

(c) The points $(1, 1), (2, 2)$ and $(0, 0)$ lie on $y = x$, contributing 0. For $(1, -1)$, the projection onto $\operatorname{span}((1, 1))$ is $\frac{1-1}{2}(1, 1) = (0, 0)$, so the distance squared is $1 + 1 = 2$. The total is 2, as predicted. (The point $(0, 0)$ contributes nothing to any subspace through the origin; it is there as a reminder that the corollary counts all columns, including zero ones.)

Exercise. C3: Truncation respects unitaries but not sums

Let $\mathbf{A} \in M_{m \times n}(F)$, and let $\mathbf{Q} \in M_m(F)$ and $\mathbf{R} \in M_n(F)$ be unitary (orthogonal, when $F = \mathbb{R}$).

(a) Prove that $\mathbf{Q}\mathbf{A}_k\mathbf{R}^*$ is a best rank-k approximation of $\mathbf{Q}\mathbf{A}\mathbf{R}^*$ in the Frobenius norm.

(b) Deduce that the function $\mathbf{A} \mapsto \min\{\|\mathbf{A}-\mathbf{B}\|_F : \operatorname{rank}\mathbf{B} \le k\}$ is unchanged by $\mathbf{A} \mapsto \mathbf{QAR}^*$.

(c) Give $\mathbf{A}, \mathbf{B} \in M_2(\mathbb{R})$ with $(\mathbf{A}+\mathbf{B})_1 \neq \mathbf{A}_1 + \mathbf{B}_1$ for every choice of the truncations involved. *Hint: two rank-one matrices whose sum is a multiple of the identity.*

Solution.

(a) If $\mathbf{A} = \mathbf{U\Sigma V}^*$ is a singular value decomposition (Theorem 13.9.1), then

$$\mathbf{QAR}^* = (\mathbf{QU})\mathbf{\Sigma}(\mathbf{RV})^*$$

is one as well, since a product of unitary matrices is unitary (Proposition 11.7.4 (a)) and $\mathbf{\Sigma}$ is unchanged. Its left and right singular vectors are $\mathbf{Qu}_i$ and $\mathbf{Rv}_i$, so its rank-k truncation for this decomposition is

$$\sum_{i\le k} \sigma_i(\mathbf{Qu}_i)(\mathbf{Rv}_i)^* = \mathbf{Q}\Big(\sum_{i\le k}\sigma_i\mathbf{u}_i\mathbf{v}_i^*\Big)\mathbf{R}^* = \mathbf{QA}_k\mathbf{R}^*.$$

By Theorem 13.11.3 applied to $\mathbf{QAR}^*$, this is a best rank-k approximation.

(b) The two matrices have the same singular values, so by Theorem 13.11.3 both minima equal $\left(\sum_{i>k}\sigma_i^2\right)^{1/2}$.

(c) Take $\mathbf{A} = \operatorname{diag}(1,0)$ and $\mathbf{B} = \operatorname{diag}(0,1)$. Each has rank one, so $\mathbf{A}_1 = \mathbf{A}$ and $\mathbf{B}_1 = \mathbf{B}$ for every singular value decomposition, and $\mathbf{A}_1 + \mathbf{B}_1 = \mathbf{I}_2$. But $\mathbf{A}+\mathbf{B} = \mathbf{I}_2$, whose rank-one truncations all have rank one, hence are never $\mathbf{I}_2$. So $(\mathbf{A}+\mathbf{B})_1 \neq \mathbf{A}_1 + \mathbf{B}_1$. Truncation is not a linear operation, which is why "keep the biggest pieces" cannot be analyzed one summand at a time.

13.12 The Moore–Penrose Pseudoinverse

Chapter 11 solved $\mathbf{Ax} = \mathbf{b}$ twice over: when there is no solution it produced a least-squares solution, and when there are too many it produced the shortest one. It then left an IOU. The two answers "are usually wanted together: given any $\mathbf{A}$ and $\mathbf{b}$, first minimize $\|\mathbf{Ax}-\mathbf{b}\|$, then among the minimizers take the shortest. Chapter 13 packages the combined answer as a single matrix, the **pseudoinverse** $\mathbf{A}^+$, with $\mathbf{x} = \mathbf{A}^+\mathbf{b}$ in all cases at once." This section builds that matrix and pays the debt.

Throughout, $F = \mathbb{R}$ or $F = \mathbb{C}$, $\mathbf{A} \in M_{m\times n}(F)$, $p = \min(m,n)$, and $r = \operatorname{rank}\mathbf{A}$, with singular values $\sigma_1 \ge \cdots \ge \sigma_p \ge 0$ (Definition 13.9.1), so $\sigma_i > 0$ exactly for $i \le r$.

13.12.1 One matrix instead of three formulas

Here is the state of play at the end of Chapter 11. To answer "what is the best $\mathbf{x}$?" we had three different recipes, and choosing between them required knowing something about $\mathbf{A}$ first.

- If $\mathbf{A}$ is square and invertible, $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$.
- If the columns of $\mathbf{A}$ are independent, $\mathbf{x} = (\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^*\mathbf{b}$ (Corollary 11.4.4).
- If $\mathbf{Ax} = \mathbf{b}$ is consistent but $\mathbf{A}$ has a kernel, $\mathbf{x} = \mathbf{x}_{\min}$, the unique solution orthogonal to $\operatorname{nul}(\mathbf{A})$ (Theorem 11.4.5).

Three recipes, three hypotheses, and a gap: when $\mathbf{A}$ has both a kernel *and* a column space

missing $\mathbf{b}$, none of them applies. That is not an exotic case. It is what happens whenever a large table of measurements has both redundant columns and noise.

The goal, stated before we know how to reach it, is a single matrix $\mathbf{A}^+$, depending on $\mathbf{A}$ alone, for which $\mathbf{x} = \mathbf{A}^+\mathbf{b}$ is the right answer in every case. Reading the goal backwards tells us what $\mathbf{A}^+$ must do, and the singular value decomposition tells us how to build it.

13.12.2 Inverting what can be inverted

Why is there no inverse in general? Two obstructions, and they are the two halves of Theorem 11.6.5. The map $\mathbf{x} \mapsto \mathbf{A}\mathbf{x}$ may fail to be injective, because $\operatorname{nul}(\mathbf{A}) \neq \{\mathbf{0}\}$; and it may fail to be surjective, because $\operatorname{col}(\mathbf{A}) \neq F^m$. Neither obstruction is fatal if we are willing to sacrifice part of the domain and part of the codomain.

Restrict the map to $\operatorname{col}(\mathbf{A}^*) = \operatorname{nul}(\mathbf{A})^\perp$, which throws away the kernel. What is left is a bijection onto $\operatorname{col}(\mathbf{A})$, and a bijection has an inverse. Extend that inverse to all of F^m by declaring it $\mathbf{0}$ on $\operatorname{col}(\mathbf{A})^\perp$, where nothing in the image can tell us anything. That is the whole construction, and it is the same trick that turns sin into arcsin: shrink the domain until the function is injective, shrink the codomain until it is surjective, invert, and say clearly what was given up.

Invert what can be inverted, and send the rest to zero.

The singular value decomposition performs all three operations at once, because it presents $\mathbf{A}$ as a diagonal matrix in orthonormal coordinates, and a diagonal matrix is inverted one entry at a time.

Definition 13.12.1. Moore–Penrose pseudoinverse

Let $\mathbf{A} \in M_{m\times n}(F)$ and let $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$ be a singular value decomposition (Theorem 13.9.1), so that $\mathbf{U} \in M_m(F)$ and $\mathbf{V} \in M_n(F)$ are unitary and $\mathbf{\Sigma} \in M_{m\times n}(F)$ has the singular values $\sigma_1, \dots, \sigma_p$ on its diagonal and zeros elsewhere. Let $\mathbf{\Sigma}^+ \in M_{n\times m}(F)$ be the matrix with

$$(\mathbf{\Sigma}^+)_{ii} = \begin{cases} \sigma_i^{-1}, & \sigma_i > 0, \\ 0, & \sigma_i = 0, \end{cases}$$

for $1 \le i \le p$, and all other entries 0. The **Moore–Penrose pseudoinverse** of $\mathbf{A}$ is

$$\mathbf{A}^+ \coloneqq \mathbf{V}\mathbf{\Sigma}^+\mathbf{U}^* \in M_{n\times m}(F).$$

Clause by clause. The matrix $\mathbf{\Sigma}^+$ is **$\mathbf{\Sigma}$ transposed and then inverted where inversion is legal**: the shape flips from $m \times n$ to $n \times m$, each nonzero diagonal entry is replaced by its reciprocal, and each zero diagonal entry is left alone — **not** replaced by anything, since 0^{-1} does not exist. The outer factors are exchanged and starred, exactly as they would be in $(\mathbf{U}\mathbf{\Sigma}\mathbf{V}^*)^{-1} = \mathbf{V}\mathbf{\Sigma}^{-1}\mathbf{U}^*$ if $\mathbf{\Sigma}$ were invertible. And the result is $n \times m$, the shape an inverse of $\mathbf{A}$ would have to be.

Well-definedness is a real question here, not a formality. The singular *values* are determined by $\mathbf{A}$, but the unitary factors are not (Theorem 13.9.2), so the recipe above is applied to a choice, and nothing so far says that two choices give the same answer. They do.

Proposition 13.12.1. The pseudoinverse does not depend on the decomposition

Let $\mathbf{A} \in M_{m\times n}(F)$ and let $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$ be any singular value decomposition (Theo-

rem 13.9.1). Then $\mathbf{X} = \mathbf{V}\mathbf{\Sigma}^{+}\mathbf{U}^{*}$ is the unique matrix in $M_{n\times m}(F)$ such that

(i) for every $\mathbf{y} \in \operatorname{col}(\mathbf{A})$, $\mathbf{X}\mathbf{y}$ is the unique vector of $\operatorname{col}(\mathbf{A}^{*})$ that $\mathbf{A}$ sends to $\mathbf{y}$;
(ii) $\mathbf{X}\mathbf{y} = \mathbf{0}$ for every $\mathbf{y} \in \operatorname{col}(\mathbf{A})^{\perp}$.

In particular $\mathbf{A}^{+}$ depends on $\mathbf{A}$ alone.

Idea. Conditions (i) and (ii) mention only $\mathbf{A}$, so if one matrix satisfies them, that matrix is determined by $\mathbf{A}$. The work is to check that they can be satisfied at all — which is the "restrict, invert, extend" story of the previous page written out — and that the matrix built from an arbitrary decomposition is the one that satisfies them. For the second part, the singular vectors are orthonormal bases of the subspaces involved, so both conditions can be checked on basis vectors.

Proof. **The conditions determine at most one matrix.** First, $\mathbf{A}$ maps $\operatorname{col}(\mathbf{A}^{*})$ bijectively onto $\operatorname{col}(\mathbf{A})$. Injectivity: if $\mathbf{x} \in \operatorname{col}(\mathbf{A}^{*})$ and $\mathbf{A}\mathbf{x} = \mathbf{0}$, then $\mathbf{x} \in \operatorname{nul}(\mathbf{A}) \cap \operatorname{nul}(\mathbf{A})^{\perp} = \{\mathbf{0}\}$, since $\operatorname{col}(\mathbf{A}^{*}) = \operatorname{nul}(\mathbf{A})^{\perp}$ by Theorem 11.6.5 (b). Surjectivity: the image is contained in $\operatorname{col}(\mathbf{A})$, and by injectivity it has dimension $\dim \operatorname{col}(\mathbf{A}^{*}) = \operatorname{rank}\mathbf{A}^{*} = \operatorname{rank}\mathbf{A} = \dim \operatorname{col}(\mathbf{A})$, the middle equality by Corollary 11.6.6 (a) applied to $\mathbf{x} \mapsto \mathbf{A}\mathbf{x}$. So (i) prescribes $\mathbf{X}$ on $\operatorname{col}(\mathbf{A})$ and (ii) prescribes it on $\operatorname{col}(\mathbf{A})^{\perp}$; since $F^{m} = \operatorname{col}(\mathbf{A}) \oplus \operatorname{col}(\mathbf{A})^{\perp}$ by Theorem 11.3.2 (a), a linear map is determined by the two prescriptions.

The matrix $\mathbf{X} = \mathbf{V}\mathbf{\Sigma}^{+}\mathbf{U}^{*}$ satisfies them. Write $\mathbf{u}_1, \dots, \mathbf{u}_m$ and $\mathbf{v}_1, \dots, \mathbf{v}_n$ for the columns of $\mathbf{U}$ and $\mathbf{V}$. Since rank is unchanged by invertible factors (Theorem 3.5.9), $r = \operatorname{rank}\mathbf{A} = \operatorname{rank}\mathbf{\Sigma}$ is the number of nonzero σ_i. By Corollary 13.9.4, $(\mathbf{u}_1, \dots, \mathbf{u}_r)$ is an orthonormal basis of $\operatorname{col}(\mathbf{A})$, the remaining $\mathbf{u}_j$ are an orthonormal basis of $\operatorname{col}(\mathbf{A})^{\perp}$, and $(\mathbf{v}_1, \dots, \mathbf{v}_r)$ is an orthonormal basis of $\operatorname{col}(\mathbf{A}^{*})$. Multiplying out the definition,

$$\mathbf{X} = \sum_{i=1}^{r} \sigma_i^{-1} \mathbf{v}_i \mathbf{u}_i^{*},$$

so $\mathbf{X}\mathbf{u}_i = \sigma_i^{-1}\mathbf{v}_i$ for $i \le r$ and $\mathbf{X}\mathbf{u}_j = \mathbf{0}$ for $j > r$. The second statement gives (ii) by linearity.

For (i), let $\mathbf{y} = \sum_{i\le r} c_i \mathbf{u}_i \in \operatorname{col}(\mathbf{A})$. Then $\mathbf{X}\mathbf{y} = \sum_{i\le r} c_i \sigma_i^{-1}\mathbf{v}_i \in \operatorname{col}(\mathbf{A}^{*})$, and since $\mathbf{A}\mathbf{v}_i = \sigma_i \mathbf{u}_i$ (the rank-one expansion of Theorem 13.9.3 read one column at a time),

$$\mathbf{A}(\mathbf{X}\mathbf{y}) = \sum_{i\le r} c_i \sigma_i^{-1} \sigma_i \mathbf{u}_i = \mathbf{y}.$$

By the injectivity proved above, $\mathbf{X}\mathbf{y}$ is the *unique* vector of $\operatorname{col}(\mathbf{A}^{*})$ with this property. This proves (i), and with it the proposition.

■

So $\mathbf{A}^{+}$ is an honest function of $\mathbf{A}$. Five examples, simplest first.

Example. Five pseudoinverses

(a) **$\mathbf{A}$ invertible.** Then $m = n = r$, $\mathbf{\Sigma}$ is invertible and $\mathbf{\Sigma}^{+} = \mathbf{\Sigma}^{-1}$, so $\mathbf{A}^{+} = \mathbf{V}\mathbf{\Sigma}^{-1}\mathbf{U}^{*} =$

$(\mathbf{U}\mathbf{\Sigma}\mathbf{V}^*)^{-1} = \mathbf{A}^{-1}$. The pseudoinverse extends the inverse, which is the first thing it had to do.

(b) $\mathbf{A} = \mathbf{O} \in M_{m\times n}(F)$**.** Every singular value is 0, so $\mathbf{\Sigma}^+ = \mathbf{O}$ and $\mathbf{A}^+ = \mathbf{O} \in M_{n\times m}(F)$. The degenerate case is worth keeping in mind: the pseudoinverse of the worst possible matrix is not undefined, it is zero, which is what "invert nothing, kill everything" means.

(c) **Diagonal.** $\operatorname{diag}(2,0,5)^+ = \operatorname{diag}(\frac{1}{2},0,\frac{1}{5})$. The zero is left where it is.

(d) **A single nonzero column** $\mathbf{a} \in M_{n\times 1}(F)$**.** Here $\sigma_1 = \|\mathbf{a}\|$, $\mathbf{u}_1 = \mathbf{a}/\|\mathbf{a}\|$, and $\mathbf{v}_1 = 1$, so $\mathbf{a}^+ = \mathbf{a}^*/\|\mathbf{a}\|^2$, a row vector. Then $\mathbf{a}^+\mathbf{a} = 1$ but $\mathbf{a}\mathbf{a}^+ = \mathbf{a}\mathbf{a}^*/\|\mathbf{a}\|^2$ is the orthogonal projection onto $\operatorname{span}(\mathbf{a})$.

(e) $\mathbf{A} = \left(\begin{smallmatrix}1&1\\1&1\end{smallmatrix}\right)$**.** This is $\sigma_1\mathbf{u}_1\mathbf{v}_1^*$ with $\sigma_1 = 2$ and $\mathbf{u}_1 = \mathbf{v}_1 = \frac{1}{\sqrt{2}}(1,1)$, since $2\mathbf{u}_1\mathbf{v}_1^* = 2\cdot\frac{1}{2}\left(\begin{smallmatrix}1&1\\1&1\end{smallmatrix}\right) = \mathbf{A}$. Hence

$$\mathbf{A}^+ = \tfrac{1}{2}\mathbf{v}_1\mathbf{u}_1^* = \frac{1}{4}\begin{pmatrix}1&1\\1&1\end{pmatrix}.$$

A **nonexample by minimal change.** Item (a) of the list is often written $\mathbf{A}^+ = (\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^*$, and Theorem 13.12.3 below confirms that formula whenever the columns of $\mathbf{A}$ are independent. Change one column of a 3×2 matrix into a multiple of the other, and the formula stops existing: $\mathbf{A}^*\mathbf{A}$ is then singular (Lemma 11.4.3), so $(\mathbf{A}^*\mathbf{A})^{-1}$ is not a matrix at all. The pseudoinverse survives the change; the formula does not. That is the difference between a definition and a recipe.

Why this definition. It is forced. Any reasonable "inverse" must undo $\mathbf{A}$ where $\mathbf{A}$ can be undone; there is exactly one subspace on which $\mathbf{A}$ is a bijection and which meets $\operatorname{nul}(\mathbf{A})$ only in $\mathbf{0}$ while being determined by $\mathbf{A}$ alone, namely $\operatorname{nul}(\mathbf{A})^\perp$; and setting the answer to $\mathbf{0}$ on $\operatorname{col}(\mathbf{A})^\perp$ is the only choice that keeps the map linear without inventing information. The names are those of the two people who published the idea, thirty-five years apart: Moore in 1920 through a projection formula, Penrose in 1955 through the four conditions of the next theorem.

Warning. $\mathbf{A}^+\mathbf{A}$ **is not the identity, and** $\mathbf{A}^+$ **is** $n\times m$**.** Two habits from the invertible case have to go. First the shape: if $\mathbf{A}$ is 3×5 then $\mathbf{A}^+$ is 5×3, so $\mathbf{A}\mathbf{A}^+$ is 3×3 and $\mathbf{A}^+\mathbf{A}$ is 5×5 — they are not even the same size, let alone both $\mathbf{I}$. Second the identity: for $\mathbf{A} = \left(\begin{smallmatrix}1&1\\1&1\end{smallmatrix}\right)$ of Example (Five pseudoinverses) (e),

$$\mathbf{A}^+\mathbf{A} = \frac{1}{4}\begin{pmatrix}1&1\\1&1\end{pmatrix}\begin{pmatrix}1&1\\1&1\end{pmatrix} = \frac{1}{2}\begin{pmatrix}1&1\\1&1\end{pmatrix} \neq \mathbf{I}_2.$$

The right question is not "is it the identity?" but "the identity on *what*?", and Theorem 13.12.3 answers it: $\mathbf{A}^+\mathbf{A}$ is the identity on $\operatorname{col}(\mathbf{A}^*)$ and zero on $\operatorname{nul}(\mathbf{A})$.

Quick check. Let $\mathbf{A} = \left(\begin{smallmatrix}1&1\\0&0\end{smallmatrix}\right)$. Find $\mathbf{A}^+$, then compute $\mathbf{A}\mathbf{A}^+$ and $\mathbf{A}^+\mathbf{A}$ and identify each as a projection.

Solution. Here $\mathbf{A} = \mathbf{e}_1(1,1)$ has rank one, so $\sigma_1 = \|\mathbf{e}_1\|\,\|(1,1)\| = \sqrt{2}$ with $\mathbf{u}_1 = \mathbf{e}_1$ and

$\mathbf{v}_1 = \frac{1}{\sqrt{2}}(1,1)$, and

$$\mathbf{A}^+ = \sigma_1^{-1}\mathbf{v}_1\mathbf{u}_1^* = \frac{1}{2}\begin{pmatrix}1 & 0\\ 1 & 0\end{pmatrix}.$$

Therefore

$$\mathbf{A}\mathbf{A}^+ = \begin{pmatrix}1 & 0\\ 0 & 0\end{pmatrix}, \qquad \mathbf{A}^+\mathbf{A} = \frac{1}{2}\begin{pmatrix}1 & 1\\ 1 & 1\end{pmatrix}.$$

Neither is $\mathbf{I}_2$. The first is the orthogonal projection onto $\operatorname{span}(\mathbf{e}_1) = \operatorname{col}(\mathbf{A})$; the second is the orthogonal projection onto $\operatorname{span}\big((1,1)\big) = \operatorname{col}(\mathbf{A}^\top)$, the row space. Each is the identity exactly on the subspace named, and zero on its orthogonal complement — which is what "as close to the identity as the shapes allow" means.

13.12.3 The four Penrose conditions

The definition builds $\mathbf{A}^+$ out of a decomposition. The following theorem characterizes it by four equations that mention nothing but $\mathbf{A}$, $\mathbf{X}$, and the star. It is the second, and better, answer to the well-definedness question, and it is how one *verifies* a candidate pseudoinverse in practice: guess, then check four identities.

Theorem 13.12.2. The Penrose Conditions

Let $\mathbf{A} \in M_{m\times n}(F)$. Then $\mathbf{X} = \mathbf{A}^+$ is the **unique** matrix in $M_{n\times m}(F)$ satisfying all four of

(P1) $\mathbf{A}\mathbf{X}\mathbf{A} = \mathbf{A}$;
(P2) $\mathbf{X}\mathbf{A}\mathbf{X} = \mathbf{X}$;
(P3) $(\mathbf{A}\mathbf{X})^* = \mathbf{A}\mathbf{X}$;
(P4) $(\mathbf{X}\mathbf{A})^* = \mathbf{X}\mathbf{A}$.

Idea. Existence is a computation with $\mathbf{\Sigma}$: each condition collapses to an identity between $\mathbf{\Sigma}$, $\mathbf{\Sigma}^+$, and the diagonal matrix of r ones. Uniqueness is a puzzle. Given two solutions $\mathbf{X}_1$ and $\mathbf{X}_2$, the goal is a chain of equalities starting at $\mathbf{X}_1$ and ending at $\mathbf{X}_2$, so the plan is to walk to a symmetric midpoint: get $\mathbf{X}_1 = \mathbf{X}_1\mathbf{A}\mathbf{X}_2$ by pushing $\mathbf{X}_2$ in from the right, then run the mirror image of the same argument with $\mathbf{X}_2$ as the subject. Two preliminary identities make the walk possible, and both come from starring one of the conditions — the move that manufactures an $\mathbf{A}^*$ when the hypotheses offer only $\mathbf{A}$.

Proof. **Existence.** Let $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$ and $\mathbf{X} = \mathbf{V}\mathbf{\Sigma}^+\mathbf{U}^*$. Using $\mathbf{V}^*\mathbf{V} = \mathbf{I}_n$ and $\mathbf{U}^*\mathbf{U} = \mathbf{I}_m$,

$$\mathbf{A}\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{\Sigma}^+\mathbf{U}^*, \qquad \mathbf{X}\mathbf{A} = \mathbf{V}\mathbf{\Sigma}^+\mathbf{\Sigma}\mathbf{V}^*.$$

Now $\mathbf{\Sigma}\mathbf{\Sigma}^+ \in M_m(F)$ is diagonal with $\sigma_i\sigma_i^{-1} = 1$ in the first r diagonal places and 0 elsewhere, and $\mathbf{\Sigma}^+\mathbf{\Sigma} \in M_n(F)$ is diagonal with the same pattern. Both are real diagonal, hence Hermitian, so $\mathbf{A}\mathbf{X}$ and $\mathbf{X}\mathbf{A}$ are Hermitian, which is (P3) and (P4). Moreover $\mathbf{\Sigma}\mathbf{\Sigma}^+\mathbf{\Sigma} = \mathbf{\Sigma}$ and $\mathbf{\Sigma}^+\mathbf{\Sigma}\mathbf{\Sigma}^+ = \mathbf{\Sigma}^+$, because multiplying a diagonal matrix by a diagonal matrix of ones and zeros keeps exactly the entries in the positions marked 1, and those are the positions where

$\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}^{+}$ are nonzero anyway. Hence

$$\mathbf{A}\mathbf{X}\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{+}\boldsymbol{\Sigma}\mathbf{V}^{*} = \mathbf{A}, \qquad \mathbf{X}\mathbf{A}\mathbf{X} = \mathbf{V}\boldsymbol{\Sigma}^{+}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{+}\mathbf{U}^{*} = \mathbf{X},$$

which is (P1) and (P2).

Uniqueness. Suppose $\mathbf{X}$ satisfies (P1)–(P4). Two consequences:

$$\mathbf{X}\mathbf{X}^{*}\mathbf{A}^{*} = \mathbf{X}, \qquad \mathbf{A}\mathbf{A}^{*}\mathbf{X}^{*} = \mathbf{A}. \tag{$*$}$$

For the first, $\mathbf{X} = \mathbf{X}\mathbf{A}\mathbf{X}$ by (P2), and $\mathbf{A}\mathbf{X} = (\mathbf{A}\mathbf{X})^{*} = \mathbf{X}^{*}\mathbf{A}^{*}$ by (P3), so $\mathbf{X} = \mathbf{X}(\mathbf{A}\mathbf{X}) = \mathbf{X}\mathbf{X}^{*}\mathbf{A}^{*}$. For the second, $\mathbf{A} = \mathbf{A}\mathbf{X}\mathbf{A}$ by (P1), and $\mathbf{X}\mathbf{A} = (\mathbf{X}\mathbf{A})^{*} = \mathbf{A}^{*}\mathbf{X}^{*}$ by (P4), so $\mathbf{A} = \mathbf{A}(\mathbf{X}\mathbf{A}) = \mathbf{A}\mathbf{A}^{*}\mathbf{X}^{*}$. Starring the second identity of $(*)$ gives the form we also need,

$$\mathbf{X}\mathbf{A}\mathbf{A}^{*} = \mathbf{A}^{*}. \tag{$**$}$$

Now let $\mathbf{X}_1$ and $\mathbf{X}_2$ both satisfy (P1)–(P4). Starting from $\mathbf{X}_1$ and working towards a term containing $\mathbf{X}_2$,

$$\begin{aligned}\mathbf{X}_1 &= \mathbf{X}_1\mathbf{X}_1^{*}\mathbf{A}^{*} = \mathbf{X}_1\mathbf{X}_1^{*}(\mathbf{A}\mathbf{X}_2\mathbf{A})^{*}\\ &= \mathbf{X}_1\mathbf{X}_1^{*}\mathbf{A}^{*}(\mathbf{A}\mathbf{X}_2)^{*} = \mathbf{X}_1\mathbf{X}_1^{*}\mathbf{A}^{*}\mathbf{A}\mathbf{X}_2\\ &= \mathbf{X}_1\mathbf{A}\mathbf{X}_2.\end{aligned}$$

The first equality is $(*)$ for $\mathbf{X}_1$; the second replaces $\mathbf{A}$ by $\mathbf{A}\mathbf{X}_2\mathbf{A}$, which is (P1) for $\mathbf{X}_2$; the third reverses the product under the star; the fourth is (P3) for $\mathbf{X}_2$; and the last is $(*)$ for $\mathbf{X}_1$ again, applied to the first three factors.

Now the mirror argument, with $\mathbf{X}_2$ as the subject:

$$\begin{aligned}\mathbf{X}_1\mathbf{A}\mathbf{X}_2 &= \mathbf{X}_1\mathbf{A}\mathbf{A}^{*}\mathbf{X}_2^{*}\mathbf{X}_2 = \mathbf{A}^{*}\mathbf{X}_2^{*}\mathbf{X}_2\\ &= (\mathbf{X}_2\mathbf{A})^{*}\mathbf{X}_2 = \mathbf{X}_2\mathbf{A}\mathbf{X}_2 = \mathbf{X}_2.\end{aligned}$$

Here the first equality replaces $\mathbf{A}$ by $\mathbf{A}\mathbf{A}^{*}\mathbf{X}_2^{*}$, which is $(*)$ for $\mathbf{X}_2$; the second is $(**)$ for $\mathbf{X}_1$; the fourth is (P4) for $\mathbf{X}_2$; and the last is (P2) for $\mathbf{X}_2$.

Hence $\mathbf{X}_1 = \mathbf{X}_2$. Since $\mathbf{A}^{+}$ satisfies the four conditions, every matrix satisfying them equals $\mathbf{A}^{+}$. This proves the theorem.

■

Both halves of the argument are the same two moves: star a condition to trade $\mathbf{A}$ for $\mathbf{A}^{*}$, then use $(*)$ to collapse a triple product. The conditions are also exactly what one should check when an explicit candidate is on the table, and several of the proofs below do nothing else.

Remark. The four conditions are not independent decoration. Dropping (P2) leaves the **generalized inverses**, matrices with $\mathbf{A}\mathbf{X}\mathbf{A} = \mathbf{A}$, of which there are many when $\mathbf{A}$ is singular; dropping (P3) and (P4) leaves more still. Each condition removes freedom, and only all four together remove all of it. Exercise C3 identifies the matrices for which $\mathbf{A}^{*}$ itself already satisfies them.

13.12.4 What the pseudoinverse does

Theorem 13.12.3. Properties of the pseudoinverse

Let $\mathbf{A} \in M_{m\times n}(F)$. Then:

(a) $\mathbf{A}\mathbf{A}^+ = P_{\mathrm{col}(\mathbf{A})}$ and $\mathbf{A}^+\mathbf{A} = P_{\mathrm{col}(\mathbf{A}^*)}$, the orthogonal projections onto the column space of $\mathbf{A}$ in F^m and onto the column space of $\mathbf{A}^*$ in F^n;

(b) $(\mathbf{A}^+)^+ = \mathbf{A}$;

(c) $(\mathbf{A}^+)^* = (\mathbf{A}^*)^+$;

(d) $\operatorname{rank}\mathbf{A}^+ = \operatorname{rank}\mathbf{A}$;

(e) $\mathbf{A}^+ = \mathbf{A}^{-1}$ if $\mathbf{A}$ is invertible; $\mathbf{A}^+ = (\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^*$ if $\operatorname{rank}\mathbf{A} = n$; and $\mathbf{A}^+ = \mathbf{A}^*(\mathbf{A}\mathbf{A}^*)^{-1}$ if $\operatorname{rank}\mathbf{A} = m$.

Idea. Clause (a) is the computation already done in the existence half of Theorem 13.12.2, read through Theorem 11.3.3. Clauses (b) and (c) are one line each, because the Penrose conditions are almost symmetric in $\mathbf{A}$ and $\mathbf{X}$, and starring all four of them produces the four conditions for $\mathbf{A}^*$. Clause (e) is three checks of the same kind: write down the candidate and confirm (P1)–(P4).

Proof.

(a) With $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$ and the notation of Proposition 13.12.1, the existence computation gave $\mathbf{A}\mathbf{A}^+ = \mathbf{U}\mathbf{\Sigma}\mathbf{\Sigma}^+\mathbf{U}^*$, where $\mathbf{\Sigma}\mathbf{\Sigma}^+$ is diagonal with r ones followed by zeros. Multiplying out,

$$\mathbf{A}\mathbf{A}^+ = \sum_{i=1}^{r} \mathbf{u}_i\mathbf{u}_i^*.$$

For $\mathbf{y} \in F^m$ this sends $\mathbf{y}$ to $\sum_{i\le r}\langle \mathbf{y}, \mathbf{u}_i\rangle \mathbf{u}_i$, which by Theorem 11.3.3 (a) is the orthogonal projection of $\mathbf{y}$ onto $\operatorname{span}(\mathbf{u}_1,\dots,\mathbf{u}_r) = \mathrm{col}(\mathbf{A})$ (Corollary 13.9.4). The same computation with $\mathbf{\Sigma}^+\mathbf{\Sigma}$ gives $\mathbf{A}^+\mathbf{A} = \sum_{i\le r}\mathbf{v}_i\mathbf{v}_i^* = P_{\mathrm{col}(\mathbf{A}^*)}$.

(b) The four Penrose conditions for the pair $(\mathbf{A}, \mathbf{X})$ become, when the roles of $\mathbf{A}$ and $\mathbf{X}$ are exchanged, the list (P2), (P1), (P4), (P3) — the same four conditions. So $\mathbf{A}$ satisfies the Penrose conditions for $\mathbf{A}^+$, and Theorem 13.12.2 gives $(\mathbf{A}^+)^+ = \mathbf{A}$.

(c) Put $\mathbf{X} = (\mathbf{A}^+)^*$ and check the four conditions for $\mathbf{A}^*$. Using $(\mathbf{B}\mathbf{C})^* = \mathbf{C}^*\mathbf{B}^*$ throughout,

$$\mathbf{A}^*\mathbf{X}\mathbf{A}^* = (\mathbf{A}\mathbf{A}^+\mathbf{A})^* = \mathbf{A}^*, \qquad \mathbf{X}\mathbf{A}^*\mathbf{X} = (\mathbf{A}^+\mathbf{A}\mathbf{A}^+)^* = \mathbf{X},$$

which are (P1) and (P2). Next, $\mathbf{A}^*\mathbf{X} = (\mathbf{A}^+\mathbf{A})^* = \mathbf{A}^+\mathbf{A}$ by (P4) for $\mathbf{A}$, and $\mathbf{A}^+\mathbf{A}$ is Hermitian, so (P3) holds; and $\mathbf{X}\mathbf{A}^* = (\mathbf{A}\mathbf{A}^+)^* = \mathbf{A}\mathbf{A}^+$ by (P3) for $\mathbf{A}$, which is Hermitian, so (P4) holds. By uniqueness, $\mathbf{X} = (\mathbf{A}^*)^+$.

(d) $\mathbf{A}^+ = \mathbf{V}\mathbf{\Sigma}^+\mathbf{U}^*$ with $\mathbf{U}, \mathbf{V}$ invertible, so $\operatorname{rank}\mathbf{A}^+ = \operatorname{rank}\mathbf{\Sigma}^+ = r$ by Theorem 3.5.9, the last equality because $\mathbf{\Sigma}^+$ has exactly r nonzero entries, all in distinct rows and columns.

(e) The invertible case is Example (Five pseudoinverses) (a). Suppose $\operatorname{rank}\mathbf{A} = n$. Then $\mathbf{A}^*\mathbf{A}$ is invertible (Corollary 11.4.4), so $\mathbf{X} = (\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^*$ is defined, and $((\mathbf{A}^*\mathbf{A})^{-1})^* = (\mathbf{A}^*\mathbf{A})^{-1}$ since $\mathbf{A}^*\mathbf{A}$ is Hermitian. Then

$$\mathbf{X}\mathbf{A} = (\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^*\mathbf{A} = \mathbf{I}_n,$$

which is Hermitian, giving (P4); and (P1), (P2) follow at once, since $\mathbf{AXA} = \mathbf{AI}_n = \mathbf{A}$ and $\mathbf{XAX} = \mathbf{I}_n\mathbf{X} = \mathbf{X}$. For (P3), $\mathbf{AX} = \mathbf{A}(\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^*$, whose conjugate transpose is itself. So $\mathbf{X} = \mathbf{A}^+$.

For the last formula, suppose $\operatorname{rank}\mathbf{A} = m$. Then $\operatorname{rank}\mathbf{A}^* = m$ is the number of columns of $\mathbf{A}^*$, so the case just proved applies to $\mathbf{A}^*$ and gives $(\mathbf{A}^*)^+ = (\mathbf{A}\mathbf{A}^*)^{-1}\mathbf{A}$. By (c),

$$\mathbf{A}^+ = \left((\mathbf{A}^*)^+\right)^* = \left((\mathbf{A}\mathbf{A}^*)^{-1}\mathbf{A}\right)^* = \mathbf{A}^*(\mathbf{A}\mathbf{A}^*)^{-1},$$

using that $\mathbf{A}\mathbf{A}^*$ is Hermitian, hence so is its inverse. This proves the theorem.

■

Clause (a) is the one to remember, and it answers the question the warning raised. The products $\mathbf{A}\mathbf{A}^+$ and $\mathbf{A}^+\mathbf{A}$ are not identities; they are the identity **restricted to the subspaces where $\mathbf{A}$ is invertible**, and zero on the complements. Clause (e) shows that the three recipes of Chapter 11 were all the pseudoinverse in disguise, each written in the notation that its hypothesis made available.

Warning. **$(\mathbf{AB})^+ \neq \mathbf{B}^+\mathbf{A}^+$ in general.** The rule $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$ does not survive. Take

$$\mathbf{A} = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}, \qquad \mathbf{B} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}.$$

The Quick check above found $\mathbf{A}^+ = \frac{1}{2}\left(\begin{smallmatrix} 1 & 0 \\ 1 & 0 \end{smallmatrix}\right)$, and one may confirm it against Theorem 13.12.2: the two products computed there are Hermitian, and $\mathbf{A}\mathbf{A}^+\mathbf{A} = \mathbf{A}$, $\mathbf{A}^+\mathbf{A}\mathbf{A}^+ = \mathbf{A}^+$. Meanwhile $\mathbf{B}$ is Hermitian and idempotent, so $\mathbf{X} = \mathbf{B}$ satisfies (P1)–(P4) and $\mathbf{B}^+ = \mathbf{B}$. Now $\mathbf{AB} = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 0 \end{smallmatrix}\right) = \mathbf{B}$, so

$$(\mathbf{AB})^+ = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \qquad \mathbf{B}^+\mathbf{A}^+ = \frac{1}{2}\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}.$$

The two differ by a factor of 2. The reversal rule does hold under extra hypotheses — for instance when $\mathbf{A}$ has independent columns and $\mathbf{B}$ has independent rows, which is Exercise C2 — but never assume it.

13.12.5 Least squares and minimum norm, in one matrix

Everything is now in place to pay Chapter 11's debt.

Theorem 13.12.4. The pseudoinverse solves both problems at once

Let $\mathbf{A} \in M_{m\times n}(F)$ and $\mathbf{b} \in F^m$, and put $\mathbf{x}^+ = \mathbf{A}^+\mathbf{b}$. Then:

(a) $\mathbf{x}^+$ is a least-squares solution of $\mathbf{Ax} = \mathbf{b}$, that is, $\|\mathbf{A}\mathbf{x}^+ - \mathbf{b}\| \le \|\mathbf{A}\mathbf{z} - \mathbf{b}\|$ for every $\mathbf{z} \in F^n$;

(b) $\|\mathbf{x}^+\| < \|\mathbf{x}\|$ for every other least-squares solution $\mathbf{x}$;

(c) if $\mathbf{Ax} = \mathbf{b}$ is consistent, $\mathbf{x}^+$ is its minimum-norm solution in the sense of Theorem 11.4.5; if $\mathbf{A}$ is invertible, $\mathbf{x}^+ = \mathbf{A}^{-1}\mathbf{b}$.

Idea. Both halves are already proved; the pseudoinverse only has to be recognized inside them. For (a), $\mathbf{Ax}^+ = \mathbf{AA}^+\mathbf{b}$ is the projection of $\mathbf{b}$ onto $\mathrm{col}(\mathbf{A})$, which is precisely condition (b) of the Least Squares Theorem. For (b), the least-squares solutions are the solutions of one consistent system, so the Minimum-Norm Solution Theorem applies to that system and singles out the solution lying in $\mathrm{nul}(\mathbf{A})^\perp$; and $\mathbf{A}^+\mathbf{b}$ lies there, because every column of $\mathbf{A}^+$ does.

Proof.

(a) By Theorem 13.12.3 (a), $\mathbf{Ax}^+ = \mathbf{AA}^+\mathbf{b} = P_U\mathbf{b}$ with $U = \mathrm{col}(\mathbf{A})$. This is condition (b) of Theorem 11.4.2, so $\mathbf{x}^+$ is a least-squares solution.

(b) By Theorem 11.4.2, the least-squares solutions of $\mathbf{Ax} = \mathbf{b}$ are exactly the solutions of the system $\mathbf{Ax} = P_U\mathbf{b}$, which is consistent because $P_U\mathbf{b} \in U = \mathrm{col}(\mathbf{A})$. Apply Theorem 11.4.5 to that system: exactly one of its solutions lies in $\mathrm{nul}(\mathbf{A})^\perp$, and it is strictly shorter than all the others.
It remains to see that $\mathbf{x}^+$ is that one. By (P2), $\mathbf{A}^+ = \mathbf{A}^+\mathbf{AA}^+ = P_{\mathrm{col}(\mathbf{A}^*)}\mathbf{A}^+$ using Theorem 13.12.3 (a), so every column of $\mathbf{A}^+$, and hence every vector $\mathbf{A}^+\mathbf{b}$, lies in $\mathrm{col}(\mathbf{A}^*)$. By Theorem 11.6.5 (b), $\mathrm{col}(\mathbf{A}^*) = \mathrm{nul}(\mathbf{A})^\perp$. So $\mathbf{x}^+$ is the minimum-norm solution of $\mathbf{Ax} = P_U\mathbf{b}$, which by the previous paragraph is what (b) asserts.

(c) If $\mathbf{Ax} = \mathbf{b}$ is consistent then $\mathbf{b} \in \mathrm{col}(\mathbf{A})$, so $P_U\mathbf{b} = \mathbf{b}$ and the least-squares solutions are the actual solutions; (b) then says $\mathbf{x}^+$ is the shortest of them, which is Theorem 11.4.5's $\mathbf{x}_{\min}$. If $\mathbf{A}$ is invertible, $\mathbf{A}^+ = \mathbf{A}^{-1}$ by Theorem 13.12.3 (e). This proves the theorem. ■

The promise, discharged. Section 4 of Chapter 11 said that the two answers are usually wanted together — first minimize $\|\mathbf{Ax} - \mathbf{b}\|$, then among the minimizers take the shortest — and that Chapter 13 would package the combined answer as a single matrix $\mathbf{A}^+$, with $\mathbf{x} = \mathbf{A}^+\mathbf{b}$ in all cases at once. That is Theorem 13.12.4, word for word: clause (a) does the minimizing, clause (b) takes the shortest, and clause (c) plus Theorem 13.12.3 (e) checks that the three special recipes of Chapter 11 are recovered as special cases. There is no hypothesis on $\mathbf{A}$ anywhere in the statement.

The notation, settled. Chapter 11 wrote $\mathbf{x}_{\min}$ for the minimum-norm solution of a *consistent* system and left the symbol without a formula. From now on it has one, and its meaning extends to every system:

$$\mathbf{x}_{\min} = \mathbf{A}^+\mathbf{b},$$

the minimum-norm least-squares solution of $\mathbf{Ax} = \mathbf{b}$. For a consistent system this agrees with Chapter 11's $\mathbf{x}_{\min}$, by Theorem 13.12.4 (c); for an inconsistent one it is the shortest of the minimizers of $\|\mathbf{Ax} - \mathbf{b}\|$, by clauses (a) and (b). One symbol, one formula, no case distinction.

Example. Chapter 11's minimum-norm solution, recomputed

In Chapter 11 the system

$$x_1 + x_2 + x_3 = 3, \qquad x_1 + 2x_2 + 3x_3 = 7$$

was solved for its shortest solution by intersecting the solution line with $\operatorname{nul}(\mathbf{A})^{\perp}$, as Theorem 11.4.5 prescribes. Recompute $\mathbf{x}_{\min}$ as $\mathbf{A}^{+}\mathbf{b}$.

Solution. Here

$$\mathbf{A} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \end{pmatrix}, \qquad \mathbf{b} = (3, 7).$$

The two rows are independent, so $\operatorname{rank}\mathbf{A} = 2 = m$ and Theorem 13.12.3 (e) gives $\mathbf{A}^{+} = \mathbf{A}^{\top}(\mathbf{A}\mathbf{A}^{\top})^{-1}$. Now

$$\mathbf{A}\mathbf{A}^{\top} = \begin{pmatrix} 3 & 6 \\ 6 & 14 \end{pmatrix}, \qquad (\mathbf{A}\mathbf{A}^{\top})^{-1} = \frac{1}{6}\begin{pmatrix} 14 & -6 \\ -6 & 3 \end{pmatrix},$$

since the determinant is $42 - 36 = 6$. Therefore

$$\mathbf{A}^{+} = \frac{1}{6}\begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{pmatrix}\begin{pmatrix} 14 & -6 \\ -6 & 3 \end{pmatrix} = \frac{1}{6}\begin{pmatrix} 8 & -3 \\ 2 & 0 \\ -4 & 3 \end{pmatrix},$$

and

$$\mathbf{x}_{\min} = \mathbf{A}^{+}\mathbf{b} = \frac{1}{6}\begin{pmatrix} 24 - 21 \\ 6 + 0 \\ -12 + 21 \end{pmatrix} = \left(\tfrac{1}{2}, 1, \tfrac{3}{2}\right).$$

This is exactly the answer Chapter 11 obtained by hand, and $\|\mathbf{x}_{\min}\|^2 = \frac{1}{4} + 1 + \frac{9}{4} = \frac{7}{2}$ as before. The system is consistent — $\frac{1}{2} + 1 + \frac{3}{2} = 3$ and $\frac{1}{2} + 2 + \frac{9}{2} = 7$ — so no projection was needed; what the pseudoinverse did was replace a geometric argument by one matrix product.

The next example is the case no formula of Chapter 11 reaches: the system is inconsistent **and** underdetermined at the same time.

Example. Inconsistent and underdetermined at once

Find $\mathbf{x}_{\min}$ for

$$\mathbf{A} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}, \qquad \mathbf{b} = (1, 3),$$

and verify directly that it minimizes the residual and is the shortest minimizer.

Solution. The system reads $x_1 + x_2 = 1$ and $x_1 + x_2 = 3$, so it is inconsistent; and $\operatorname{nul}(\mathbf{A}) = \operatorname{span}((1, -1)) \neq \{\mathbf{0}\}$, so the columns are dependent. Neither $\mathbf{A}^{-1}$ nor $(\mathbf{A}^{*}\mathbf{A})^{-1}\mathbf{A}^{*}$ exists. From Example (Five pseudoinverses) (e), $\mathbf{A}^{+} = \frac{1}{4}\left(\begin{smallmatrix} 1 & 1 \\ 1 & 1 \end{smallmatrix}\right)$, so

$$\mathbf{x}_{\min} = \mathbf{A}^{+}\mathbf{b} = \frac{1}{4}(1 + 3, 1 + 3) = (1, 1).$$

Check. $\operatorname{col}(\mathbf{A}) = \operatorname{span}\big((1,1)\big)$, and the projection of $\mathbf{b} = (1,3)$ onto it is $\frac{1+3}{2}(1,1) = (2,2)$. Indeed $\mathbf{A}(1,1) = (2,2)$, so the residual is $\mathbf{b} - (2,2) = (-1,1)$, of squared length 2 — the smallest possible. The least-squares solutions are all $\mathbf{x}$ with $x_1 + x_2 = 2$, that is $(1,1) + t(1,-1)$, and

$$\|(1,1) + t(1,-1)\|^2 = 2 + 2t^2,$$

which is smallest exactly at $t = 0$. So $(1,1)$ is the shortest least-squares solution, as Theorem 13.12.4 (b) predicts.

13.12.6 The block criterion, finished

Section 6 tested a block matrix for positive definiteness one block at a time, and had to assume the corner block invertible: the Schur complement $\mathbf{M}/\mathbf{A} = \mathbf{C} - \mathbf{B}^*\mathbf{A}^{-1}\mathbf{B}$ is meaningless otherwise. The pseudoinverse removes that hypothesis, and the promise Section 6 made is paid here.

Proposition 13.12.5. The Semidefinite Block Criterion

Let $\mathbf{A} \in M_k(F)$ and $\mathbf{C} \in M_l(F)$ be self-adjoint and put

$$\mathbf{M} = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^* & \mathbf{C} \end{pmatrix} \in M_{k+l}(F).$$

Then $\mathbf{M} \succeq \mathbf{O}$ if and only if $\mathbf{A} \succeq \mathbf{O}$, every column of $\mathbf{B}$ lies in $\operatorname{col}(\mathbf{A})$, and $\mathbf{C} - \mathbf{B}^*\mathbf{A}^{+}\mathbf{B} \succeq \mathbf{O}$.

Idea. Section 6 cleared the corner with the shear $\left(\begin{smallmatrix} \mathbf{I} & -\mathbf{A}^{-1}\mathbf{B} \\ \mathbf{O} & \mathbf{I} \end{smallmatrix}\right)$. Write $\mathbf{A}^{+}$ in place of $\mathbf{A}^{-1}$ and the same shear still clears it — but only because $\mathbf{A}\mathbf{A}^{+}$ is the projection onto $\operatorname{col}(\mathbf{A})$, so $\mathbf{A}\mathbf{A}^{+}\mathbf{B} = \mathbf{B}$ exactly when the columns of $\mathbf{B}$ are already in $\operatorname{col}(\mathbf{A})$. That condition is not a technicality: it is what the two-sided test forces.

Proof. Suppose first that $\mathbf{M} \succeq \mathbf{O}$. Taking $\mathbf{v} = (\mathbf{x}, \mathbf{0})$ gives $\langle \mathbf{A}\mathbf{x}, \mathbf{x}\rangle \geq 0$, so $\mathbf{A} \succeq \mathbf{O}$. For the column condition, let $\mathbf{x} \in \operatorname{nul}(\mathbf{A})$ and $\mathbf{y} \in F^l$, and put $\mathbf{v}_t = (t\mathbf{x}, \mathbf{y})$ for $t \in \mathbb{R}$. Then

$$\langle \mathbf{M}\mathbf{v}_t, \mathbf{v}_t\rangle = t^2\langle \mathbf{A}\mathbf{x}, \mathbf{x}\rangle + 2t\,\Re\langle \mathbf{B}\mathbf{y}, \mathbf{x}\rangle + \langle \mathbf{C}\mathbf{y}, \mathbf{y}\rangle = 2t\,\Re\langle \mathbf{B}\mathbf{y}, \mathbf{x}\rangle + \langle \mathbf{C}\mathbf{y}, \mathbf{y}\rangle,$$

using $\mathbf{A}\mathbf{x} = \mathbf{0}$. A real affine function of t that is nonnegative for every t has zero slope, so $\Re\langle \mathbf{B}\mathbf{y}, \mathbf{x}\rangle = 0$ for all $\mathbf{y}$; replacing $\mathbf{y}$ by $i\mathbf{y}$ over $\mathbb{C}$ kills the imaginary part too. Hence $\langle \mathbf{B}\mathbf{y}, \mathbf{x}\rangle = 0$ for every $\mathbf{y}$, that is $\mathbf{B}^*\mathbf{x} = \mathbf{0}$. So $\operatorname{nul}(\mathbf{A}) \subseteq \operatorname{nul}(\mathbf{B}^*)$, and taking orthogonal complements with Theorem 11.6.5 gives $\operatorname{col}(\mathbf{B}) \subseteq \operatorname{col}(\mathbf{A})$, since $\mathbf{A}$ is self-adjoint.

Now assume $\mathbf{A} \succeq \mathbf{O}$ and $\operatorname{col}(\mathbf{B}) \subseteq \operatorname{col}(\mathbf{A})$; we show $\mathbf{M} \succeq \mathbf{O}$ is equivalent to $\mathbf{C} - \mathbf{B}^*\mathbf{A}^{+}\mathbf{B} \succeq \mathbf{O}$, which finishes both directions. By Theorem 13.12.3 (a), $\mathbf{A}\mathbf{A}^{+} = P_{\operatorname{col}(\mathbf{A})}$, so $\mathbf{A}\mathbf{A}^{+}\mathbf{B} = \mathbf{B}$; taking adjoints, and using that $\mathbf{A}$ and $\mathbf{A}^{+}$ are self-adjoint (Theorem 13.12.3 (c)), also $\mathbf{B}^*\mathbf{A}^{+}\mathbf{A} = \mathbf{B}^*$. Put

$$\mathbf{S} = \begin{pmatrix} \mathbf{I}_k & -\mathbf{A}^{+}\mathbf{B} \\ \mathbf{O} & \mathbf{I}_l \end{pmatrix},$$

which is invertible. Multiplying out and canceling with the two identities just recorded,

$$\mathbf{S}^{*}\mathbf{M}\mathbf{S} = \begin{pmatrix} \mathbf{A} & \mathbf{O} \\ \mathbf{O} & \mathbf{C} - \mathbf{B}^{*}\mathbf{A}^{+}\mathbf{B} \end{pmatrix}.$$

Since $\mathbf{S}$ is invertible, Proposition 13.1.2 (a) applied to $\mathbf{S}$ and then to $\mathbf{S}^{-1}$ shows that $\mathbf{M} \succeq \mathbf{O}$ if and only if $\mathbf{S}^{*}\mathbf{M}\mathbf{S} \succeq \mathbf{O}$, and a block diagonal matrix is positive semidefinite exactly when both blocks are. This shows the three conditions together are equivalent to $\mathbf{M} \succeq \mathbf{O}$.

■

The column condition cannot be dropped. With

$$\mathbf{A} = \begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix}, \qquad \mathbf{C} = (1),$$

the choice $\mathbf{B} = (1, 0)^{\top}$ has its column in $\operatorname{col}(\mathbf{A})$ and gives $\mathbf{C} - \mathbf{B}^{*}\mathbf{A}^{+}\mathbf{B} = (\tfrac{1}{2}) \succeq \mathbf{O}$, so $\mathbf{M} \succeq \mathbf{O}$. The choice $\mathbf{B} = (0, 1)^{\top}$ does not, and there $\mathbf{M}$ is not positive semidefinite even though $\mathbf{A} \succeq \mathbf{O}$, $\mathbf{C} \succeq \mathbf{O}$ and $\mathbf{B}^{*}\mathbf{A}^{+}\mathbf{B} = \mathbf{O}$ makes the third condition hold trivially.

13.12.7 Exercises

13.12.7.1 A. Check your understanding

Exercise. A1

(a) Define $\mathbf{A}^{+}$ in terms of a singular value decomposition of $\mathbf{A}$ (Theorem 13.9.1), and say what size it is.

(b) State the four Penrose conditions, and say what Theorem 13.12.2 asserts about them.

(c) What are $\mathbf{A}\mathbf{A}^{+}$ and $\mathbf{A}^{+}\mathbf{A}$? Under what condition on $\mathbf{A}$ is the second of them $\mathbf{I}_n$?

(d) Write down $\mathbf{O}^{+}$ for the zero matrix $\mathbf{O} \in M_{m\times n}(F)$, and $\mathbf{A}^{+}$ for an invertible $\mathbf{A}$.

(e) True or false: $(\mathbf{A}\mathbf{B})^{+} = \mathbf{B}^{+}\mathbf{A}^{+}$ whenever the product $\mathbf{A}\mathbf{B}$ is defined. Justify your answer.

Solution.

(a) If $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{*}$, then $\mathbf{A}^{+} = \mathbf{V}\mathbf{\Sigma}^{+}\mathbf{U}^{*}$, where $\mathbf{\Sigma}^{+} \in M_{n\times m}(F)$ has σ_i^{-1} in the (i,i) place for each $\sigma_i > 0$ and zeros elsewhere. It is $n \times m$, the transpose of the shape of $\mathbf{A}$.

(b) $\mathbf{A}\mathbf{X}\mathbf{A} = \mathbf{A}$, $\mathbf{X}\mathbf{A}\mathbf{X} = \mathbf{X}$, $(\mathbf{A}\mathbf{X})^{*} = \mathbf{A}\mathbf{X}$, $(\mathbf{X}\mathbf{A})^{*} = \mathbf{X}\mathbf{A}$. The theorem says $\mathbf{A}^{+}$ satisfies all four and is the only matrix that does; so the four conditions may be used as a definition, and checking them is enough to identify a candidate.

(c) $\mathbf{A}\mathbf{A}^{+} = P_{\operatorname{col}(\mathbf{A})}$ and $\mathbf{A}^{+}\mathbf{A} = P_{\operatorname{col}(\mathbf{A}^{*})}$ (Theorem 13.12.3 (a)). The second is $\mathbf{I}_n$ exactly when $\operatorname{col}(\mathbf{A}^{*}) = F^{n}$, that is when $\operatorname{rank}\mathbf{A} = n$, that is when the columns of $\mathbf{A}$ are independent.

(d) $\mathbf{O}^{+} = \mathbf{O} \in M_{n\times m}(F)$; $\mathbf{A}^{+} = \mathbf{A}^{-1}$ when $\mathbf{A}$ is invertible.

(e) False. With $\mathbf{A} = \left(\begin{smallmatrix} 1 & 1 \\ 0 & 0 \end{smallmatrix}\right)$ and $\mathbf{B} = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 0 \end{smallmatrix}\right)$ the two sides are $\left(\begin{smallmatrix} 1 & 0 \\ 0 & 0 \end{smallmatrix}\right)$ and $\frac{1}{2}\left(\begin{smallmatrix} 1 & 0 \\ 0 & 0 \end{smallmatrix}\right)$, as computed in the warning above.

13.12.7.2 B. Practice

Exercise. B1: Four pseudoinverses

Compute $\mathbf{A}^+$ for each of the following, stating which method you used.

(a) $\mathbf{A} = \operatorname{diag}(3, 0, 0, 4) \in M_4(\mathbb{R})$;

(b) $\mathbf{A} = (1, 2, 2)^\top \in M_{3\times 1}(\mathbb{R})$;

(c) $\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix} \in M_2(\mathbb{R})$;

(d) $\mathbf{A} = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix} \in M_{2\times 3}(\mathbb{R})$.

Solution.

(a) A diagonal matrix is its own singular value decomposition (Theorem 13.9.1) up to ordering, so invert the nonzero entries in place: $\mathbf{A}^+ = \operatorname{diag}(\tfrac{1}{3}, 0, 0, \tfrac{1}{4})$.

(b) A single nonzero column, so Example (Five pseudoinverses) (d) applies with $\|\mathbf{A}\|^2 = 1 + 4 + 4 = 9$:

$$\mathbf{A}^+ = \tfrac{1}{9}(1, 2, 2) \in M_{1\times 3}(\mathbb{R}).$$

(c) Rank one: $\mathbf{A} = \mathbf{u}\mathbf{v}^\top$ with $\mathbf{u} = (1, 2)$ and $\mathbf{v} = (1, 2)$. Then $\sigma_1 = \|\mathbf{u}\|\,\|\mathbf{v}\| = 5$ with $\mathbf{u}_1 = \mathbf{u}/\sqrt{5}$ and $\mathbf{v}_1 = \mathbf{v}/\sqrt{5}$, so

$$\mathbf{A}^+ = \tfrac{1}{5}\mathbf{v}_1\mathbf{u}_1^\top = \tfrac{1}{25}\begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix}.$$

(Check: $\mathbf{A}\mathbf{A}^+ = \frac{1}{25}\left(\begin{smallmatrix} 5 & 10 \\ 10 & 20 \end{smallmatrix}\right)$ is Hermitian and idempotent, and $\mathbf{A}\mathbf{A}^+\mathbf{A} = \mathbf{A}$.)

(d) The rows are independent, so $\operatorname{rank}\mathbf{A} = 2 = m$ and Theorem 13.12.3 (e) gives $\mathbf{A}^+ = \mathbf{A}^\top(\mathbf{A}\mathbf{A}^\top)^{-1}$. Here $\mathbf{A}\mathbf{A}^\top = \left(\begin{smallmatrix} 2 & 1 \\ 1 & 2 \end{smallmatrix}\right)$ with inverse $\frac{1}{3}\left(\begin{smallmatrix} 2 & -1 \\ -1 & 2 \end{smallmatrix}\right)$, so

$$\mathbf{A}^+ = \frac{1}{3}\begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix} = \frac{1}{3}\begin{pmatrix} 2 & -1 \\ -1 & 2 \\ 1 & 1 \end{pmatrix}.$$

Exercise. B2: A system that is both inconsistent and underdetermined

Let

$$\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 2 & 4 \\ 1 & 2 \end{pmatrix}, \qquad \mathbf{b} = (1, 4, 1).$$

Compute $\mathbf{A}^+$ and $\mathbf{x}_{\min} = \mathbf{A}^+\mathbf{b}$. Verify that $\mathbf{A}\mathbf{x}_{\min}$ is the orthogonal projection of $\mathbf{b}$ onto $\operatorname{col}(\mathbf{A})$, and compare $\|\mathbf{x}_{\min}\|$ with the norm of one other least-squares solution.

Solution. Both columns are multiples of $\mathbf{u} = (1, 2, 1)$, so $\mathbf{A} = \mathbf{u}\mathbf{v}^\top$ with $\mathbf{v} = (1, 2)$, and $\operatorname{rank}\mathbf{A} = 1$. Then $\|\mathbf{u}\|^2 = 6$, $\|\mathbf{v}\|^2 = 5$, and $\sigma_1 = \sqrt{30}$, with $\mathbf{u}_1 = \mathbf{u}/\sqrt{6}$ and $\mathbf{v}_1 = \mathbf{v}/\sqrt{5}$.

Hence

$$\mathbf{A}^+ = \sigma_1^{-1}\mathbf{v}_1\mathbf{u}_1^\top = \frac{1}{30}\mathbf{v}\mathbf{u}^\top = \frac{1}{30}\begin{pmatrix}1 & 2 & 1\\ 2 & 4 & 2\end{pmatrix}.$$

Therefore

$$\mathbf{x}_{\min} = \mathbf{A}^+\mathbf{b} = \frac{1}{30}(1+8+1,\ 2+16+2) = \left(\tfrac{1}{3}, \tfrac{2}{3}\right).$$

Projection. $\operatorname{col}(\mathbf{A}) = \operatorname{span}(\mathbf{u})$ and $\langle \mathbf{b}, \mathbf{u}\rangle = 1+8+1 = 10$, so $P_{\operatorname{col}(\mathbf{A})}\mathbf{b} = \frac{10}{6}\mathbf{u} = \frac{5}{3}(1,2,1)$. And $\mathbf{A}\mathbf{x}_{\min} = \mathbf{u}(\mathbf{v}^\top\mathbf{x}_{\min}) = \mathbf{u}(\frac{1}{3}+\frac{4}{3}) = \frac{5}{3}\mathbf{u}$. They agree.

Shortest. The least-squares solutions are the $\mathbf{x}$ with $\mathbf{v}^\top\mathbf{x} = x_1 + 2x_2 = \frac{5}{3}$. Another one is $(\frac{5}{3}, 0)$, of squared norm $\frac{25}{9}$, against $\|\mathbf{x}_{\min}\|^2 = \frac{1}{9}+\frac{4}{9} = \frac{5}{9}$. The pseudoinverse solution is shorter, as Theorem 13.12.4 (b) requires.

Exercise. B3: Checking a candidate

Let $\mathbf{A} = \begin{pmatrix}1 & 0\\ 1 & 1\\ 0 & 1\end{pmatrix}$ and $\mathbf{X} = \frac{1}{3}\begin{pmatrix}2 & 1 & -1\\ -1 & 1 & 2\end{pmatrix}$. Verify the four Penrose conditions for $\mathbf{X}$, and hence find the minimum-norm least-squares solution of $\mathbf{A}\mathbf{x} = (1,2,3)$.

Solution. Compute the two products first:

$$\mathbf{X}\mathbf{A} = \frac{1}{3}\begin{pmatrix}3 & 0\\ 0 & 3\end{pmatrix} = \mathbf{I}_2, \qquad \mathbf{A}\mathbf{X} = \frac{1}{3}\begin{pmatrix}2 & 1 & -1\\ 1 & 2 & 1\\ -1 & 1 & 2\end{pmatrix}.$$

(P4): $\mathbf{X}\mathbf{A} = \mathbf{I}_2$ is Hermitian. (P3): $\mathbf{A}\mathbf{X}$ is visibly symmetric. (P1): $\mathbf{A}\mathbf{X}\mathbf{A} = \mathbf{A}(\mathbf{X}\mathbf{A}) = \mathbf{A}\mathbf{I}_2 = \mathbf{A}$. (P2): $\mathbf{X}\mathbf{A}\mathbf{X} = \mathbf{I}_2\mathbf{X} = \mathbf{X}$. All four hold, so $\mathbf{X} = \mathbf{A}^+$ by Theorem 13.12.2.

Then

$$\mathbf{x}_{\min} = \mathbf{X}(1,2,3) = \frac{1}{3}(2+2-3,\ -1+2+6) = \left(\tfrac{1}{3}, \tfrac{7}{3}\right).$$

(The columns of $\mathbf{A}$ are independent, so this is also $(\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}^\top\mathbf{b}$, and it is the unique least-squares solution; the normal equations $\left(\begin{smallmatrix}2 & 1\\ 1 & 2\end{smallmatrix}\right)\mathbf{x} = (3,5)$ confirm it.)

13.12.7.3 C. Going deeper

Exercise. C1: Pseudoinverses of $\mathbf{A}^*\mathbf{A}$

Let $\mathbf{A} \in M_{m\times n}(F)$.

(a) Prove that $(\mathbf{A}^*\mathbf{A})^+ = \mathbf{A}^+(\mathbf{A}^+)^*$.

(b) Deduce that $\mathbf{A}^+ = (\mathbf{A}^*\mathbf{A})^+\mathbf{A}^*$, the formula $(\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^*$ with the inverse replaced by the pseudoinverse.

Hint: use a singular value decomposition and reduce each identity to an identity between diagonal matrices.

Solution. Write $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$.

(a) Then $\mathbf{A}^*\mathbf{A} = \mathbf{V}\mathbf{\Sigma}^*\mathbf{\Sigma}\mathbf{V}^*$, and $\mathbf{\Sigma}^*\mathbf{\Sigma} \in M_n(F)$ is diagonal with entries $\sigma_1^2, \dots, \sigma_p^2$ followed by zeros. That display is a singular value decomposition of $\mathbf{A}^*\mathbf{A}$ in the sense of Theorem 13.9.1 (both outer factors are $\mathbf{V}$, which is unitary, and the middle factor is diagonal with nonnegative entries), so

$$(\mathbf{A}^*\mathbf{A})^+ = \mathbf{V}(\mathbf{\Sigma}^*\mathbf{\Sigma})^+\mathbf{V}^*,$$

where $(\mathbf{\Sigma}^*\mathbf{\Sigma})^+$ is diagonal with entries σ_i^{-2} for $i \le r$ and 0 elsewhere. On the other side, $(\mathbf{A}^+)^* = (\mathbf{V}\mathbf{\Sigma}^+\mathbf{U}^*)^* = \mathbf{U}(\mathbf{\Sigma}^+)^*\mathbf{V}^*$, so

$$\mathbf{A}^+(\mathbf{A}^+)^* = \mathbf{V}\mathbf{\Sigma}^+\mathbf{U}^*\mathbf{U}(\mathbf{\Sigma}^+)^*\mathbf{V}^* = \mathbf{V}\mathbf{\Sigma}^+(\mathbf{\Sigma}^+)^*\mathbf{V}^*.$$

The middle factor $\mathbf{\Sigma}^+(\mathbf{\Sigma}^+)^*$ is diagonal with entries $\sigma_i^{-1}\sigma_i^{-1} = \sigma_i^{-2}$ for $i \le r$ and 0 elsewhere, which is $(\mathbf{\Sigma}^*\mathbf{\Sigma})^+$. Hence the two sides agree.

(b) Using (a) and $\mathbf{A}^* = \mathbf{V}\mathbf{\Sigma}^*\mathbf{U}^*$,

$$(\mathbf{A}^*\mathbf{A})^+\mathbf{A}^* = \mathbf{V}\mathbf{\Sigma}^+(\mathbf{\Sigma}^+)^*\mathbf{\Sigma}^*\mathbf{U}^*.$$

Now $(\mathbf{\Sigma}^+)^*\mathbf{\Sigma}^* = (\mathbf{\Sigma}\mathbf{\Sigma}^+)^* = \mathbf{\Sigma}\mathbf{\Sigma}^+$, which is the $m \times m$ diagonal matrix with r ones and then zeros; and $\mathbf{\Sigma}^+$ multiplied by it on the right is $\mathbf{\Sigma}^+$, since $\mathbf{\Sigma}^+$ has nonzero entries only in the first r columns. So the right-hand side is $\mathbf{V}\mathbf{\Sigma}^+\mathbf{U}^* = \mathbf{A}^+$.

Exercise. C2: When the reversal rule does hold

Let $\mathbf{A} \in M_{m\times n}(F)$ with $\operatorname{rank}\mathbf{A} = n$ and $\mathbf{B} \in M_{n\times q}(F)$ with $\operatorname{rank}\mathbf{B} = n$. Prove that

$$(\mathbf{AB})^+ = \mathbf{B}^+\mathbf{A}^+.$$

Hint: write both pseudoinverses with the formulas of Theorem 13.12.3 (e) and verify the Penrose conditions for the product.

Solution. By Theorem 13.12.3 (e), $\mathbf{A}^+ = (\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^*$ and $\mathbf{B}^+ = \mathbf{B}^*(\mathbf{B}\mathbf{B}^*)^{-1}$, the two inverses existing because $\operatorname{rank}\mathbf{A} = n$ and $\operatorname{rank}\mathbf{B}^* = n$. Put

$$\mathbf{X} = \mathbf{B}^+\mathbf{A}^+ = \mathbf{B}^*(\mathbf{B}\mathbf{B}^*)^{-1}(\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^*.$$

Throughout we use $\mathbf{A}^+\mathbf{A} = \mathbf{I}_n$ and $\mathbf{B}\mathbf{B}^+ = \mathbf{I}_n$, which the two formulas give at once.

(P1): $(\mathbf{AB})\mathbf{X}(\mathbf{AB}) = \mathbf{A}(\mathbf{B}\mathbf{B}^+)(\mathbf{A}^+\mathbf{A})\mathbf{B} = \mathbf{A}\mathbf{I}_n\mathbf{I}_n\mathbf{B} = \mathbf{AB}$.

(P2): $\mathbf{X}(\mathbf{AB})\mathbf{X} = \mathbf{B}^+(\mathbf{A}^+\mathbf{A})(\mathbf{B}\mathbf{B}^+)\mathbf{A}^+ = \mathbf{B}^+\mathbf{A}^+ = \mathbf{X}$.

(P3): $(\mathbf{AB})\mathbf{X} = \mathbf{A}(\mathbf{B}\mathbf{B}^+)\mathbf{A}^+ = \mathbf{A}\mathbf{A}^+ = \mathbf{A}(\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^*$, which equals its own conjugate transpose, since $\mathbf{A}^*\mathbf{A}$ is Hermitian and so is its inverse.

(P4): $\mathbf{X}(\mathbf{AB}) = \mathbf{B}^+(\mathbf{A}^+\mathbf{A})\mathbf{B} = \mathbf{B}^+\mathbf{B} = \mathbf{B}^*(\mathbf{B}\mathbf{B}^*)^{-1}\mathbf{B}$, Hermitian for the same reason.

All four hold, so $\mathbf{X} = (\mathbf{AB})^+$ by Theorem 13.12.2.

Exercise. C3: When the pseudoinverse is the adjoint

Let $\mathbf{A} \in M_{m\times n}(F)$.

(a) Prove that $\mathbf{A}^+ = \mathbf{A}^*$ if and only if $\mathbf{A}\mathbf{A}^*\mathbf{A} = \mathbf{A}$.

(b) Deduce that $\mathbf{A}^+ = \mathbf{A}^*$ if and only if every nonzero singular value of $\mathbf{A}$ equals 1.

(c) Give an example of such an $\mathbf{A}$ in $M_2(\mathbb{R})$ which is neither unitary nor an orthogonal projection.

Solution.

(a) ($\Rightarrow$) If $\mathbf{A}^+ = \mathbf{A}^*$, then (P1) reads $\mathbf{A}\mathbf{A}^*\mathbf{A} = \mathbf{A}$.
($\Leftarrow$) Suppose $\mathbf{A}\mathbf{A}^*\mathbf{A} = \mathbf{A}$ and put $\mathbf{X} = \mathbf{A}^*$. Then (P1) is the hypothesis; (P2) is $\mathbf{A}^*\mathbf{A}\mathbf{A}^* = (\mathbf{A}\mathbf{A}^*\mathbf{A})^* = \mathbf{A}^*$; (P3) is $(\mathbf{A}\mathbf{A}^*)^* = \mathbf{A}\mathbf{A}^*$, true for every $\mathbf{A}$; and (P4) is $(\mathbf{A}^*\mathbf{A})^* = \mathbf{A}^*\mathbf{A}$, likewise. By Theorem 13.12.2, $\mathbf{A}^+ = \mathbf{A}^*$.

(b) With $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$ we have $\mathbf{A}^* = \mathbf{V}\mathbf{\Sigma}^*\mathbf{U}^*$ and $\mathbf{A}^+ = \mathbf{V}\mathbf{\Sigma}^+\mathbf{U}^*$. Since $\mathbf{U}$ and $\mathbf{V}$ are invertible, $\mathbf{A}^+ = \mathbf{A}^*$ if and only if $\mathbf{\Sigma}^+ = \mathbf{\Sigma}^*$, that is if and only if $\sigma_i^{-1} = \sigma_i$ for every $i \le r$, that is if and only if every nonzero singular value is 1 (the equation $\sigma^2 = 1$ with $\sigma > 0$ forces $\sigma = 1$).

(c) Take

$$\mathbf{A} = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}.$$

Then $\mathbf{A}\mathbf{A}^\top = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 0 \end{smallmatrix}\right)$, so $\mathbf{A}\mathbf{A}^\top\mathbf{A} = \mathbf{A}$ and (a) applies. Its singular values are 1 and 0, matching (b). It is not unitary, since it is singular, and not an orthogonal projection, since $\mathbf{A}^\top \neq \mathbf{A}$. Such a matrix maps $\operatorname{col}(\mathbf{A}^*)$ isometrically onto $\operatorname{col}(\mathbf{A})$ and sends $\operatorname{nul}(\mathbf{A})$ to $\mathbf{0}$: a unitary matrix on the part of the space where it is invertible, and nothing at all elsewhere.

13.13 Projections Revisited

Chapter 4 built projections out of nothing but a direct sum: given $V = U \oplus W$, the operator that keeps the U-piece and discards the W-piece satisfies $P^2 = P$, and every operator satisfying $P^2 = P$ arises this way. Chapter 11 built a second kind by taking $W = U^\perp$, the orthogonal projection P_U. The second is a special case of the first, and Chapter 4 left the comparison open, saying only that its projections "are often called **oblique** projections, to distinguish them from the orthogonal projections met later, in the chapters on inner product spaces". This section makes the distinction exact, measures how far an oblique projection is from being orthogonal, and closes Part III.

Throughout, V is a finite-dimensional inner product space over $F = \mathbb{R}$ or $\mathbb{C}$, and $n = \dim V$.

13.13.1 Which projections are orthogonal?

Two recollections, because the section is about the gap between them. From Chapter 2: an operator $P \in \mathcal{L}(V)$ with $P^2 = P$ is a **projection**, also called an **idempotent** (Definition 4.5.1); it satisfies $V = \operatorname{im} P \oplus \ker P$, it is the identity on $\operatorname{im} P$ and zero on $\ker P$, and each splitting $V = U \oplus W$ has exactly one projection with image U and kernel W (Theorem 4.5.1). No inner product appears there, and the field was arbitrary.

From Chapter 11: the splitting $V = U \oplus U^{\perp}$ of Theorem 11.3.2 is one of the many available, and P_U is the projection belonging to it (Definition 11.3.2). Chapter 4 warned that a subspace is the image of infinitely many projections, one per complement. Which one does the inner product pick out, and how is it recognized without being told the kernel?

A projection is orthogonal exactly when it throws away perpendicularly — and that is the same as costing nothing to move across the inner product.

Theorem 13.13.1. Orthogonal Projections Among the Idempotents

Let V be a finite-dimensional inner product space over F and let $P \in \mathcal{L}(V)$ satisfy $P^2 = P$. The following are equivalent.

(a) P is the orthogonal projection onto its image: $P = P_U$ with $U = \operatorname{im} P$.
(b) $\ker P = (\operatorname{im} P)^{\perp}$.
(c) $P^* = P$.
(d) P is normal.
(e) $\|Pv\| \le \|v\|$ **for every** $v \in V$.

Idea. Conditions (a) and (b) say one thing twice, once about the operator and once about its two subspaces. Going from (b) to (c) is a computation Chapter 11 has already done for P_U. Coming back from (c) is the four-subspaces theorem: $\ker P^*$ is always $(\operatorname{im} P)^{\perp}$, so once $P^* = P$ the kernel is forced; normality is the same remark with $\ker P^*$ replaced by $\ker P$. The one direction with work in it is (e) $\Rightarrow$ (b), where a bound on lengths has to become a statement about angles. Feed P a vector $u + tw$ with $u \in \operatorname{im} P$ and $w \in \ker P$: its image is u whatever t is, so choosing t to make $\|u + tw\|$ as small as possible squeezes $\langle u, w\rangle$ to zero.

Proof. **(a) $\Leftrightarrow$ (b).** Write $U = \operatorname{im} P$. ($\Rightarrow$) If $P = P_U$, then $\ker P = \ker P_U = U^{\perp}$, since P_U is the projection onto U along $U^{\perp}$ (Definition 11.3.2 and Theorem 4.5.1 (b)). ($\Leftarrow$) If $\ker P = U^{\perp}$, then P is a projection with image U and kernel $U^{\perp}$, and so is P_U; by the uniqueness in Theorem 4.5.1 (b), applied to $V = U \oplus U^{\perp}$, the two are equal.

(b) $\Rightarrow$ (c). By the equivalence just proved, $P = P_U$ with $U = \operatorname{im} P$. By Theorem 11.3.3 (b), $\langle Pu, v\rangle = \langle u, Pv\rangle$ for all $u, v \in V$. That is the defining identity of the adjoint, so $P^* = P$ by the uniqueness in Definition 11.6.1.

(c) $\Rightarrow$ (d). If $P^* = P$ then $P^*P = P^2 = PP^*$, which is Definition 11.6.3.

(d) $\Rightarrow$ (b). Since P is normal, $\ker P = \ker P^*$ by Corollary 12.3.2 (a). By Theorem 11.6.5 (a), $\ker P^* = (\operatorname{im} P)^{\perp}$. Combining the two gives (b).

(b) $\Rightarrow$ (e). Let $v \in V$ and split it as $v = Pv + (v - Pv)$. The first piece lies in $\operatorname{im} P$, and the second lies in $\ker P$, because $P(v - Pv) = Pv - P^2v = 0$. By (b) the two pieces are orthogonal, so Theorem 11.1.3 gives

$$\|v\|^2 = \|Pv\|^2 + \|v - Pv\|^2 \ge \|Pv\|^2,$$

and taking nonnegative square roots gives $\|Pv\| \le \|v\|$.

(e) $\Rightarrow$ (b). We first show $\langle u, w\rangle = 0$ for every $u \in \operatorname{im} P$ and every $w \in \ker P$. There is

nothing to prove when $w = 0$, so assume $w \neq 0$, write $z = \langle u, w\rangle$, and put

$$t = -\frac{z}{\|w\|^2}, \qquad v = u + tw,$$

which is legitimate because $\|w\| \neq 0$. Then $Pv = Pu + tPw = u$, since P is the identity on $\operatorname{im} P$ and zero on $\ker P$ (Theorem 4.5.1 (a)). Expanding the norm squared and using conjugate symmetry,

$$\begin{aligned}\|v\|^2 &= \|u\|^2 + \bar{t}\langle u, w\rangle + t\langle w, u\rangle + |t|^2\|w\|^2 \\ &= \|u\|^2 - \frac{|z|^2}{\|w\|^2} - \frac{|z|^2}{\|w\|^2} + \frac{|z|^2}{\|w\|^2} = \|u\|^2 - \frac{|z|^2}{\|w\|^2}.\end{aligned}$$

Now (e) applied to this v says $\|u\|^2 = \|Pv\|^2 \leq \|v\|^2$, which forces $|z|^2 \leq 0$ and hence $z = 0$.

So $\ker P \subseteq (\operatorname{im} P)^\perp$. The two subspaces have the same dimension: $\dim \ker P = n - \operatorname{rank} P$ by Theorem 2.3.1, and $\dim(\operatorname{im} P)^\perp = n - \dim \operatorname{im} P = n - \operatorname{rank} P$ by Theorem 11.3.2 (c). A subspace contained in another of the same finite dimension equals it (Theorem 1.6.8), so $\ker P = (\operatorname{im} P)^\perp$. This closes the cycle and proves the theorem.

■

The theorem is the promised comparison. Every projection of Chapter 4 that is **not** self-adjoint is oblique, and now "oblique" has a precise meaning: its kernel is a complement of its image that is not the perpendicular one. Chapter 12 has already met a pair of such projections, in Example (Two Projections That Are Not Orthogonal), and the Quick check after Theorem 12.8.2 traced exactly what their obliqueness cost.

Clause (c) also settles Chapter 4's warning that "projection onto U" is incomplete information.

Corollary 13.13.2. One Orthogonal Projection per Subspace

Let U be a subspace of a finite-dimensional inner product space V. Among the projections of V with image U — one for each complement of U, by Theorem 4.5.1 — exactly one is self-adjoint, namely P_U.

Proof. P_U has image U and is self-adjoint by Theorem 13.13.1, (a) ⇒ (c). Conversely, let P be a self-adjoint projection with $\operatorname{im} P = U$. By the same theorem, (c) ⇒ (b), $\ker P = U^\perp$, which is also $\ker P_U$. Two projections with the same image and the same kernel are equal, by the uniqueness in Theorem 4.5.1 (b).

■

Chapter 4's own example is now decided by inspection. On $\mathbb{R}^2$ with the dot product, $P(x, y) = (x, 0)$ and $P'(x, y) = (x - y, 0)$ are both projections onto the x-axis, with standard matrices

$$\mathbf{P} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \qquad \mathbf{P}' = \begin{pmatrix} 1 & -1 \\ 0 & 0 \end{pmatrix}.$$

The standard basis is orthonormal, so Theorem 11.6.3 says self-adjointness is symmetry of these matrices. The first is symmetric, so $P = P_U$ for the x-axis U. The second is not, so P' is oblique — as the picture in Chapter 4 showed, it slides along the line $y = x$ rather than straight down.

Clause (e) fails visibly at $\mathbf{v} = (1, -1)$: there $P'\mathbf{v} = (2, 0)$, and $\|P'\mathbf{v}\| = 2$ while $\|\mathbf{v}\| = \sqrt{2}$.

Quick check. Does $\mathbf{A} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}$ act on $\mathbb{R}^2$ as a projection? If so, is it an orthogonal projection?

Solution. $\mathbf{A}^2 = \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix} = \mathbf{A}$, so yes, it is a projection. Its image is the span of its columns, $\operatorname{span}((1, 1))$, and its kernel is $\operatorname{span}(\mathbf{e}_2)$. Since $\langle (1, 1), \mathbf{e}_2 \rangle = 1 \neq 0$, the kernel is not $(\operatorname{im} \mathbf{A})^\perp$, so the projection is oblique by Theorem 13.13.1 (b). The quickest route is clause (c): $\mathbf{A}^\top \neq \mathbf{A}$.

Warning. **Clause (e) is a statement about every vector, and one vector proves nothing.** For the oblique $\mathbf{P}'$ above, $\|\mathbf{P}'\mathbf{e}_1\| = 1 = \|\mathbf{e}_1\|$ and $\|\mathbf{P}'(1, 1)\| = 0 \leq \sqrt{2}$; both are consistent with the bound, and both are worthless as evidence. The bound fails only in the direction $(1, -1)$. If you want to certify a projection as orthogonal by hand, check $\mathbf{P}^* = \mathbf{P}$ in an orthonormal basis, not the length of one image.

Warning. **An orthogonal projection is almost never an orthogonal matrix.** If $\mathbf{P}$ is unitary and idempotent, it is invertible, so multiplying $\mathbf{P}^2 = \mathbf{P}$ by $\mathbf{P}^{-1}$ gives $\mathbf{P} = \mathbf{I}$. Apart from the identity, every orthogonal projection is singular. "Orthogonal" in P_U names the direction of the discarding, not membership in $\mathrm{O}(n)$.

13.13.2 The block that measures obliqueness

Chapter 2 already knows the shape of a projection up to a change of basis: with a basis of $\operatorname{im} P$ followed by a basis of $\ker P$, the matrix is $\operatorname{diag}(1, \ldots, 1, 0, \ldots, 0)$ with $\operatorname{rank} P$ ones (Corollary 4.5.2). That basis is built to fit P and is usually not orthonormal, so the normal form erases exactly the information we now want. Insisting on an orthonormal basis costs one extra block, and that block is the answer.

Theorem 13.13.3. Canonical Form of an Idempotent in an Orthonormal Basis

Let V be an n-dimensional inner product space over F and let $P \in \mathcal{L}(V)$ be a projection of rank r. Let $(f_1, \ldots, f_r)$ be any orthonormal basis of $\operatorname{im} P$ and $(f_{r+1}, \ldots, f_n)$ any orthonormal basis of $(\operatorname{im} P)^\perp$. Then $\mathcal{B} = (f_1, \ldots, f_n)$ is an orthonormal basis of V, and

$$[P]_{\mathcal{B}}^{\mathcal{B}} = \begin{pmatrix} \mathbf{I}_r & \mathbf{X} \\ \mathbf{O} & \mathbf{O} \end{pmatrix}$$

for some $\mathbf{X} \in M_{r \times (n-r)}(F)$. Moreover P is an orthogonal projection if and only if $\mathbf{X} = \mathbf{O}$.

Idea. Only two facts are used, and both are already proved. First, P fixes its image, so the first r columns are forced to be $\mathbf{e}_1, \ldots, \mathbf{e}_r$. Second, every output of P lies in $\operatorname{im} P$, so the

last $n-r$ columns have nothing below row r. What sits **above** row r in those columns is unconstrained, and that is $\mathbf{X}$: it records where P sends the directions perpendicular to its image, which an orthogonal projection would send to 0.

Proof. Write $U = \operatorname{im} P$, so $\dim U = r$ and $\dim U^{\perp} = n - r$ by Theorem 11.3.2 (c). Each f_i with $i \le r$ is orthogonal to each f_j with $j > r$, since the first lies in U and the second in $U^{\perp}$; so $\mathcal{B}$ is an orthonormal list of n vectors, hence an orthonormal basis of V (Theorem 11.1.4).

For $i \le r$ we have $f_i \in \operatorname{im} P$, so $Pf_i = f_i$ by Theorem 4.5.1 (a), and column i of $[P]_{\mathcal{B}}^{\mathcal{B}}$ is $\mathbf{e}_i$. For $j > r$ we have $Pf_j \in \operatorname{im} P = U = \operatorname{span}(f_1, \dots, f_r)$, so $Pf_j = \sum_{i \le r} x_{ij} f_i$ for scalars $x_{ij} \in F$, and column j has zeros in rows $r+1, \dots, n$. Collecting the x_{ij} into $\mathbf{X}$ gives the displayed shape.

For the last claim, $\mathcal{B}$ is orthonormal, so $[P^*]_{\mathcal{B}}^{\mathcal{B}} = ([P]_{\mathcal{B}}^{\mathcal{B}})^*$ by Theorem 11.6.3, and

$$([P]_{\mathcal{B}}^{\mathcal{B}})^* = \begin{pmatrix} \mathbf{I}_r & \mathbf{O} \\ \mathbf{X}^* & \mathbf{O} \end{pmatrix}.$$

The two matrices agree if and only if $\mathbf{X} = \mathbf{O}$. Since an operator is determined by its matrix in a basis, $P^* = P$ if and only if $\mathbf{X} = \mathbf{O}$, and $P^* = P$ says P is an orthogonal projection by Theorem 13.13.1.

■

The block $\mathbf{X}$ is not unique, because the two orthonormal bases were not. Replacing them changes $\mathbf{X}$ into $\mathbf{A}^*\mathbf{X}\mathbf{B}$ for some unitary $\mathbf{A} \in M_r(F)$ and $\mathbf{B} \in M_{n-r}(F)$, and that leaves the singular values of $\mathbf{X}$ alone: since

$$(\mathbf{A}^*\mathbf{X}\mathbf{B})^*(\mathbf{A}^*\mathbf{X}\mathbf{B}) = \mathbf{B}^*(\mathbf{X}^*\mathbf{X})\mathbf{B},$$

the two matrices $(\mathbf{A}^*\mathbf{X}\mathbf{B})^*(\mathbf{A}^*\mathbf{X}\mathbf{B})$ and $\mathbf{X}^*\mathbf{X}$ are unitarily similar, so they have the same eigenvalues, and Definition 13.9.1 reads the singular values off those. So the list $\sigma_1(\mathbf{X}) \ge \sigma_2(\mathbf{X}) \ge \dots$ belongs to P itself, and P is an orthogonal projection exactly when that list is all zeros.

13.13.2.1 The two-dimensional case, in full

Take $n = 2$ and $r = 1$, so that $\mathbf{X}$ is a single scalar $x \in F$, and

$$[P]_{\mathcal{B}}^{\mathcal{B}} = \begin{pmatrix} 1 & x \\ 0 & 0 \end{pmatrix}, \qquad \mathcal{B} = (f_1, f_2),$$

with f_1 spanning the line $\operatorname{im} P$ and f_2 spanning the perpendicular line $(\operatorname{im} P)^{\perp}$. Everything about P can be read off this one number.

The kernel. $P(af_1 + bf_2) = (a + xb)f_1$, which is 0 exactly when $a = -xb$. So

$$\ker P = \operatorname{span}(-xf_1 + f_2),$$

a line that meets $(\operatorname{im} P)^{\perp} = \operatorname{span}(f_2)$ only at 0 unless $x = 0$.

The angle. Suppose $F = \mathbb{R}$. The two lines $\operatorname{im} P$ and $\ker P$ are distinct, since their intersection is $\{0\}$; let $\theta \in (0, \pi/2]$ be the angle between them, meaning the angle of Definition 11.1.5 between f_1 and whichever of $\pm(-xf_1 + f_2)$ makes it acute. With $k = -xf_1 + f_2$ we have $\|k\| = \sqrt{1 + x^2}$

and $\langle f_1, k\rangle = -x$, so

$$\cos\theta = \frac{|x|}{\sqrt{1+x^2}}, \qquad \sin\theta = \frac{1}{\sqrt{1+x^2}}, \qquad |x| = \cot\theta.$$

The dictionary is complete: $x = 0$ is $\theta = \pi/2$, the orthogonal case; and as the kernel tips toward the image, $\theta \to 0$ and $|x| \to \infty$. The single entry x **is** the obliqueness, in the units of a cotangent.

The stretching. For a unit vector $v = af_1 + bf_2$, we have $\|Pv\| = |a + xb| = |\langle (a,b), (1,\overline{x})\rangle|$ in F^2 with the standard inner product, so Theorem 11.1.5 gives

$$\|Pv\| \le \sqrt{1+|x|^2},$$

with equality exactly when (a,b) is a multiple of $(1,\overline{x})$, for instance $(a,b) = (1,\overline{x})/\sqrt{1+|x|^2}$. So the largest value of $\|Pv\|$ over unit vectors v is $\sqrt{1+|x|^2}$, which over $\mathbb{R}$ is $1/\sin\theta$. Chapter 16 will give this number a name, the operator norm $\|P\|$; here it is enough to say that an orthogonal projection never lengthens a vector, while an oblique one lengthens some vector by $1/\sin\theta$, a factor that is unbounded as the kernel closes in on the image.

Example. Two Oblique Projections, Measured

For each of

$$\mathbf{P}' = \begin{pmatrix} 1 & -1 \\ 0 & 0 \end{pmatrix}, \qquad \mathbf{Q} = \begin{pmatrix} 2 & -1 \\ 2 & -1 \end{pmatrix}$$

on $\mathbb{R}^2$, verify that it is a projection, put it in the form of Theorem 13.13.3, and find the angle between its image and its kernel and the largest length of an image of a unit vector.

Solution. The matrix $\mathbf{P}'$. $(\mathbf{P}')^2 = \mathbf{P}'$ by direct multiplication, and $\operatorname{im}\mathbf{P}' = \operatorname{span}(\mathbf{e}_1)$, $\ker\mathbf{P}' = \operatorname{span}((1,1))$. Here $\mathbf{f}_1 = \mathbf{e}_1$ already spans the image and $\mathbf{f}_2 = \mathbf{e}_2$ spans its perpendicular, and the standard basis is orthonormal, so the given matrix **is** the canonical form, with $x = -1$. Hence $\cot\theta = 1$, that is $\theta = \pi/4$, and the largest value of $\|\mathbf{P}'\mathbf{v}\|$ on the unit circle is $\sqrt{1+1} = \sqrt{2}$, attained at $\mathbf{v} = (1,-1)/\sqrt{2}$. Both readings match Chapter 4's figure, where $\mathbf{P}'$ projects along the line $y = x$.

The matrix $\mathbf{Q}$. $\mathbf{Q}^2 = \begin{pmatrix} 4-2 & -2+1 \\ 4-2 & -2+1 \end{pmatrix} = \mathbf{Q}$, so $\mathbf{Q}$ is a projection, of rank 1. Its image is $\operatorname{span}((1,1))$ and its kernel is $\operatorname{span}((1,2))$, since $\mathbf{Q}(1,2) = (0,0)$. Take

$$\mathbf{f}_1 = \tfrac{1}{\sqrt{2}}(1,1), \qquad \mathbf{f}_2 = \tfrac{1}{\sqrt{2}}(1,-1),$$

an orthonormal basis of $\operatorname{im}\mathbf{Q}$ and of $(\operatorname{im}\mathbf{Q})^\perp$ respectively. Then $\mathbf{Q}\mathbf{f}_1 = \mathbf{f}_1$ and $\mathbf{Q}\mathbf{f}_2 = \tfrac{1}{\sqrt{2}}(3,3) = 3\mathbf{f}_1$, so $x = 3$ and

$$[T_{\mathbf{Q}}]_{\mathcal{B}}^{\mathcal{B}} = \begin{pmatrix} 1 & 3 \\ 0 & 0 \end{pmatrix}.$$

Hence $\cot\theta = 3$. Checking against the two lines directly: the angle between $(1,1)$ and $(1,2)$ has $\cos\theta = 3/\sqrt{10}$ and $\sin\theta = 1/\sqrt{10}$, so $\cot\theta = 3$, and the largest value of $\|\mathbf{Q}\mathbf{v}\|$ on the unit circle is $\sqrt{1+9} = \sqrt{10} = 1/\sin\theta$. The two routes agree.

In higher rank the same block $\mathbf{X}$ does the same job, and Exercise (C3: How far an oblique projection can stretch) turns its largest singular value into the largest value of $\|Pv\|$ over unit

v, by the argument that just ran in the 2×2 case. The reason the singular values appear is the one this chapter has used throughout: $\mathbf{X}$ is a rectangular matrix, so its eigenvalues do not exist, and Theorem 13.9.1 is what measures it instead.

13.13.3 Looking back at Part III

Part III added one piece of structure to the vector spaces of Parts I and II, and spent three chapters finding out what it buys.

Chapter 11 added the inner product, and with it length, orthogonality, and angle. Orthonormal bases made coordinates free, orthogonal complements split every space canonically, and projection onto a subspace solved the nearest-point problem, hence least squares. Its deepest construction was the **adjoint**, the price of moving an operator from one slot of the inner product to the other; the adjoint is what turned the kernel and the image into two orthogonal pairs of subspaces.

Chapter 12 asked which operators the inner product can diagonalize, and the adjoint answered: over $\mathbb{C}$ the normal ones, over $\mathbb{R}$ the self-adjoint ones. Written without bases, the spectral theorem says $T = \sum_i \lambda_i P_i$ with the P_i orthogonal projections onto the eigenspaces — which is why the present section had to exist, and why Theorem 12.8.1 needed clause (b) at all. Once a matrix is diagonal in an orthonormal basis, anything can be done to its eigenvalues and undone afterwards, and that is the functional calculus.

Chapter 13 asked which self-adjoint operators behave like nonnegative numbers, collected the equivalent answers in Theorem 13.1.4, and used them twice: to build the square root and the absolute value $|\mathbf{A}| = (\mathbf{A}^*\mathbf{A})^{1/2}$ of Definition 13.2.1, and then to get past the restriction that had hung over the whole part. The spectral theorem needs a square operator, and most matrices are not square — but $\mathbf{A}^*\mathbf{A}$ always **is** square, self-adjoint, and positive semidefinite, whatever $\mathbf{A}$ is. Applying the spectral theorem to it and reading the result back through $\mathbf{A}$ gives Theorem 13.9.1, and with it orthonormal bases of all four fundamental subspaces (Corollary 13.9.4), the polar decomposition of Theorem 13.10.1, the best low-rank approximation, and the pseudoinverse. A rectangular matrix has no eigenvalues; it has singular values, and they do the work.

So the part is one chain, each link an application of the last: **inner product** → **adjoint** → **spectral theorem** → **singular value decomposition**. Its named moves travel. To show a vector is zero, pair it with itself. To move an operator across the inner product, pay with the adjoint. Work with the norm squared. Fix an orthonormal basis. Project, then subtract. Test positivity with the single scalar $\langle \mathbf{Ax}, \mathbf{x} \rangle$, and factor a positive matrix as $\mathbf{B}^*\mathbf{B}$. Diagonalize, work on the eigenvalues, come back. When the matrix is not square, apply the spectral theorem to $\mathbf{A}^*\mathbf{A}$. And when the answer is a sum ordered by size, truncate it.

What comes next takes the quadratic form $\mathbf{x} \mapsto \langle \mathbf{Ax}, \mathbf{x} \rangle$, which was only a test in this chapter, and makes it the object of study: Chapter 14 classifies forms up to congruence, by inertia, over $\mathbb{R}$ and over $\mathbb{C}$. Chapter 16 gives the operator norm promised twice above, and Chapter 17 characterizes the eigenvalues of a Hermitian matrix as maxima and minima of $\langle \mathbf{Ax}, \mathbf{x} \rangle$ — at which point the test becomes the theory.

13.13.4 Exercises

13.13.4.1 A. Check your understanding

Exercise. A1

(a) Define what it means for $P \in \mathcal{L}(V)$ to be a projection, and what it means for it to be the orthogonal projection onto a subspace U. Which of the two definitions needs an inner product?

(b) State three conditions on a projection P, each equivalent to "P is an orthogonal projection".

(c) Determine whether the following is correct, and justify your answer: if $\|Pv\| \le \|v\|$ for some nonzero v, then the projection P is orthogonal.

(d) Determine whether the following is correct, and justify your answer: every projection of rank r on an n-dimensional inner product space has matrix $\mathbf{I}_r \oplus \mathbf{O}$ in some orthonormal basis.

Solution.

(a) P is a projection when $P^2 = P$ (Definition 4.5.1); P is the orthogonal projection onto U when it is the projection onto U along $U^\perp$ (Definition 11.3.2). Only the second needs an inner product, since $U^\perp$ is undefined without one.

(b) By Theorem 13.13.1: $\ker P = (\operatorname{im} P)^\perp$; $P^* = P$; P is normal; $\|Pv\| \le \|v\|$ for every v. Any three of these will do.

(c) Incorrect. Take $\mathbf{P}' = \left(\begin{smallmatrix} 1 & -1 \\ 0 & 0 \end{smallmatrix}\right)$ on $\mathbb{R}^2$ and $\mathbf{v} = (1, 1)$. Then $\mathbf{P}'\mathbf{v} = \mathbf{0}$, so $\|\mathbf{P}'\mathbf{v}\| = 0 \le \sqrt{2}$, yet $\mathbf{P}'$ is oblique because $(\mathbf{P}')^\top \neq \mathbf{P}'$. Clause (e) of Theorem 13.13.1 quantifies over **every** $\mathbf{v}$.

(d) Incorrect in general, and correct exactly for the orthogonal projections. By Theorem 13.13.3 the matrix in a suitable orthonormal basis is $\left(\begin{smallmatrix} \mathbf{I}_r & \mathbf{X} \\ \mathbf{O} & \mathbf{O} \end{smallmatrix}\right)$, and it is $\mathbf{I}_r \oplus \mathbf{O}$ only when $\mathbf{X} = \mathbf{O}$. If some orthonormal basis gave $\mathbf{I}_r \oplus \mathbf{O}$, that matrix would be Hermitian, so P would be self-adjoint by Theorem 11.6.3. The matrix $\mathbf{P}'$ of part (c) is a counterexample. (In a basis that is merely a basis, $\mathbf{I}_r \oplus \mathbf{O}$ is always attainable: that is Corollary 4.5.2.)

13.13.4.2 B. Practice

Exercise. B1: Which are orthogonal projections?

Determine which of the following matrices are projections, and among those, which are orthogonal projections. Justify your answer. For the oblique ones over $\mathbb{R}$, find the angle between image and kernel.

(a) $\mathbf{A} = \frac{1}{5}\begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix}$ on $\mathbb{R}^2$.

(b) $\mathbf{B} = \begin{pmatrix} 1 & 0 \\ -2 & 0 \end{pmatrix}$ on $\mathbb{R}^2$.

(c) $\mathbf{C} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$ on $\mathbb{R}^2$.

(d) $\mathbf{D} = \begin{pmatrix} 1 & i \\ 0 & 0 \end{pmatrix}$ on $\mathbb{C}^2$.

Solution.

(a) $\mathbf{A}^2 = \frac{1}{25}\left(\begin{smallmatrix} 5 & 10 \\ 10 & 20 \end{smallmatrix}\right) = \mathbf{A}$, so $\mathbf{A}$ is a projection, and $\mathbf{A}^\top = \mathbf{A}$. By Theorem 13.13.1 it is the orthogonal projection onto $\operatorname{im}\mathbf{A} = \operatorname{span}((1,2))$.

(b) $\mathbf{B}^2 = \mathbf{B}$, so $\mathbf{B}$ is a projection, but $\mathbf{B}^\top \neq \mathbf{B}$, so it is oblique. Its image is $\operatorname{span}((1,-2))$ and its kernel is $\operatorname{span}(\mathbf{e}_2)$. The angle between these lines has $\cos\theta = |\langle(1,-2),(0,1)\rangle|/(\sqrt{5}\cdot 1) = 2/\sqrt{5}$, so $\sin\theta = 1/\sqrt{5}$ and $\cot\theta = 2$; that is $\theta = \arctan(1/2)$, about $26.6°$.

(c) $\mathbf{C}^2 = \left(\begin{smallmatrix} 2 & 2 \\ 2 & 2 \end{smallmatrix}\right) = 2\mathbf{C} \neq \mathbf{C}$, so $\mathbf{C}$ is not a projection at all, although it is symmetric. Symmetry alone is not idempotence; $\frac{1}{2}\mathbf{C}$ is the orthogonal projection onto $\operatorname{span}((1,1))$.

(d) $\mathbf{D}^2 = \mathbf{D}$, so $\mathbf{D}$ is a projection, and $\mathbf{D}^* = \left(\begin{smallmatrix} 1 & 0 \\ -i & 0 \end{smallmatrix}\right) \neq \mathbf{D}$, so it is oblique. It is already in the canonical form of Theorem 13.13.3, with $x = i$. There is no angle over $\mathbb{C}$ (Definition 11.1.5 is stated for real spaces), but $|x| = 1$, and the largest value of $\|\mathbf{D}\mathbf{v}\|$ over unit $\mathbf{v}$ is $\sqrt{1+1} = \sqrt{2}$, the same number the real matrix $\mathbf{P}'$ produced.

Exercise. B2: Putting a projection in canonical form

Let $\mathbf{A} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ act on $\mathbb{R}^3$ with the dot product.

(a) Verify that $\mathbf{A}$ is a projection and find $\operatorname{im}\mathbf{A}$ and $\ker\mathbf{A}$.

(b) Find an orthonormal basis $\mathcal{B}$ of $\mathbb{R}^3$ as in Theorem 13.13.3 and write down $[T_{\mathbf{A}}]_{\mathcal{B}}^{\mathcal{B}}$.

(c) Hence determine whether $\mathbf{A}$ is an orthogonal projection, and find $\operatorname{rank}\mathbf{A}$ without row reducing.

Solution.

(a) $\mathbf{A}^2 = \mathbf{A}$, by multiplying out: the second column of $\mathbf{A}^2$ is $\mathbf{A}(1,1,0) = (1,1,0)$, and the first and third columns are $\mathbf{0}$ and $\mathbf{e}_3$, unchanged. The columns of $\mathbf{A}$ span $\operatorname{im}\mathbf{A} = \operatorname{span}((1,1,0),\mathbf{e}_3)$, which is 2-dimensional. And $\mathbf{A}\mathbf{x} = (x_2,x_2,x_3)$ is $\mathbf{0}$ exactly when $x_2 = x_3 = 0$, so $\ker\mathbf{A} = \operatorname{span}(\mathbf{e}_1)$.

(b) $\operatorname{im}\mathbf{A}$ has the orthonormal basis $\mathbf{f}_1 = \frac{1}{\sqrt{2}}(1,1,0)$, $\mathbf{f}_2 = \mathbf{e}_3$, and $(\operatorname{im}\mathbf{A})^\perp = \operatorname{span}((1,-1,0))$ has the orthonormal basis $\mathbf{f}_3 = \frac{1}{\sqrt{2}}(1,-1,0)$. Now $\mathbf{A}\mathbf{f}_1 = \mathbf{f}_1$ and $\mathbf{A}\mathbf{f}_2 = \mathbf{f}_2$, while

$$\mathbf{A}\mathbf{f}_3 = \tfrac{1}{\sqrt{2}}(-1,-1,0) = -\mathbf{f}_1.$$

Hence

$$[T_{\mathbf{A}}]_{\mathcal{B}}^{\mathcal{B}} = \begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \qquad \mathbf{X} = \begin{pmatrix} -1 \\ 0 \end{pmatrix}.$$

(c) $\mathbf{X} \neq \mathbf{O}$, so $\mathbf{A}$ is oblique by Theorem 13.13.3; indeed $\ker\mathbf{A} = \operatorname{span}(\mathbf{e}_1)$ is not $(\operatorname{im}\mathbf{A})^\perp = \operatorname{span}((1,-1,0))$. The rank is the number of ones in the canonical form, namely 2; equivalently $\operatorname{rank}\mathbf{A} = \operatorname{tr}\mathbf{A} = 0+1+1 = 2$ by Theorem 4.5.4.

Exercise. B3: The complementary projection

Let $P \in \mathcal{L}(V)$ be a projection on a finite-dimensional inner product space, with $U = \operatorname{im} P$.

(a) Prove that P is an orthogonal projection if and only if $\mathrm{id}_V - P$ is.

(b) Prove that in that case $\mathrm{id}_V - P = P_{U^\perp}$.

Hint: Exercise (B2: The complementary projection) identifies the image and the kernel of $\mathrm{id}_V - P$.

Solution.

(a) By Exercise (B2: The complementary projection) (a), $\mathrm{id}_V - P$ is a projection. By Theorem 11.6.2 (a), (b) and (d), $(\mathrm{id}_V - P)^* = \mathrm{id}_V - P^*$. Hence $(\mathrm{id}_V - P)^* = \mathrm{id}_V - P$ if and only if $P^* = P$, and Theorem 13.13.1 converts each side into the corresponding statement about orthogonal projections.

(b) Suppose P is an orthogonal projection. By Exercise (B2: The complementary projection) (b), $\operatorname{im}(\mathrm{id}_V - P) = \ker P$, which is $U^\perp$ by Theorem 13.13.1 (b). By (a), $\mathrm{id}_V - P$ is an orthogonal projection, so by clause (a) of that theorem it is the orthogonal projection onto its own image, that is $P_{U^\perp}$.

13.13.4.3 C. Going deeper

Exercise. C1: Products of orthogonal projections

Let P and Q be orthogonal projections on a finite-dimensional inner product space V, with images U and W.

(a) Prove that PQ is an orthogonal projection if and only if $PQ = QP$.

(b) Suppose $PQ = QP$. Prove that $PQ = P_{U \cap W}$.

(c) Give an explicit example of two orthogonal projections on $\mathbb{R}^2$ whose product is not a projection.

Solution.

(a) ($\Leftarrow$) Suppose $PQ = QP$. Then $(PQ)^2 = P(QP)Q = P(PQ)Q = P^2Q^2 = PQ$, so PQ is a projection, and $(PQ)^* = Q^*P^* = QP = PQ$ by Theorem 11.6.2 (c) and self-adjointness of P and Q. By Theorem 13.13.1, PQ is an orthogonal projection.
($\Rightarrow$) Suppose PQ is an orthogonal projection. Then $PQ = (PQ)^* = Q^*P^* = QP$.

(b) Assume $PQ = QP$, so PQ is an orthogonal projection by (a), and by clause (a) of Theorem 13.13.1 it equals $P_{\operatorname{im}(PQ)}$. It remains to show $\operatorname{im}(PQ) = U \cap W$. ($\subseteq$) Every value PQv lies in $\operatorname{im} P = U$; and $PQv = QPv$ lies in $\operatorname{im} Q = W$. ($\supseteq$) If $v \in U \cap W$, then $Qv = v$ and $Pv = v$, since a projection fixes its image (Theorem 4.5.1 (a)), so $PQv = v$ and $v \in \operatorname{im}(PQ)$.

(c) Project onto the x-axis and onto the line $y = x$, with standard matrices

$$\mathbf{P} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \qquad \mathbf{Q} = \tfrac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}.$$

Both are symmetric and idempotent, hence orthogonal projections. But $\mathbf{PQ} = \frac{1}{2}\left(\begin{smallmatrix} 1 & 1 \\ 0 & 0 \end{smallmatrix}\right)$ and $(\mathbf{PQ})^2 = \frac{1}{4}\left(\begin{smallmatrix} 1 & 1 \\ 0 & 0 \end{smallmatrix}\right) \neq \mathbf{PQ}$, so the product is not even a projection. Consistently with (a), $\mathbf{QP} = \frac{1}{2}\left(\begin{smallmatrix} 1 & 0 \\ 1 & 0 \end{smallmatrix}\right) \neq \mathbf{PQ}$.

Exercise. C2: The Frobenius norm detects obliqueness

Let $\mathbf{P} \in M_n(F)$ be idempotent, of rank r.

(a) Prove that $\|\mathbf{P}\|_F^2 \geq r$.

(b) Prove that equality holds if and only if $\mathbf{P}$ is an orthogonal projection.

Hint: the canonical form of Theorem 13.13.3, together with the invariance of the Frobenius norm under a unitary change of orthonormal basis.

Solution. Apply Theorem 13.13.3 to the operator $\mathbf{x} \mapsto \mathbf{Px}$ on F^n, and let $\mathbf{U}$ be the matrix whose columns are the vectors of the orthonormal basis it produces. Then $\mathbf{U}$ is unitary (Theorem 11.7.3 (f)), and the change-of-basis formula gives

$$\mathbf{U}^*\mathbf{PU} = \mathbf{U}^{-1}\mathbf{PU} = \begin{pmatrix} \mathbf{I}_r & \mathbf{X} \\ \mathbf{O} & \mathbf{O} \end{pmatrix}.$$

By Lemma 11.7.5, $\|\mathbf{P}\|_F = \|\mathbf{U}^*\mathbf{PU}\|_F$.

(a) Summing the squared moduli of the entries of the block matrix,

$$\|\mathbf{P}\|_F^2 = \|\mathbf{I}_r\|_F^2 + \|\mathbf{X}\|_F^2 = r + \|\mathbf{X}\|_F^2 \geq r,$$

since $\|\mathbf{X}\|_F^2 \geq 0$.

(b) Equality holds if and only if $\|\mathbf{X}\|_F = 0$, that is $\mathbf{X} = \mathbf{O}$, which by Theorem 13.13.3 is exactly the condition that $\mathbf{P}$ be an orthogonal projection. (Note that $r = \operatorname{rank}\mathbf{P} = \operatorname{tr}\mathbf{P}$ by Theorem 4.5.4, so the criterion may also be written $\operatorname{tr}(\mathbf{P}^*\mathbf{P}) = \operatorname{tr}\mathbf{P}$.)

Exercise. C3: How far an oblique projection can stretch

Let P be a projection of rank r on an n-dimensional inner product space V over F, with $0 < r < n$, and let $\mathbf{X}$ be the block of Theorem 13.13.3, with largest singular value $\sigma_1 = \sigma_1(\mathbf{X})$.

(a) Prove that $\|\mathbf{Xz}\| \leq \sigma_1\|\mathbf{z}\|$ for every $\mathbf{z} \in F^{n-r}$, with equality for some $\mathbf{z} \neq \mathbf{0}$.

(b) Prove that the largest value of $\|Pv\|$ over unit vectors $v \in V$ is $\sqrt{1+\sigma_1^2}$.

(c) Deduce the two-dimensional formula $\sqrt{1+|x|^2}$ of the text, and state what the answer becomes when P is an orthogonal projection.

Solution.

(a) Write $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$ as in Theorem 13.9.1, with $\mathbf{U}, \mathbf{V}$ unitary, and let p be the length of the list of singular values. For $\mathbf{z} \in F^{n-r}$ put $\mathbf{w} = \mathbf{V}^*\mathbf{z}$, so that $\|\mathbf{w}\| = \|\mathbf{z}\|$ and $\|\mathbf{Xz}\| = \|\mathbf{\Sigma w}\|$, a unitary matrix preserving norms (Theorem 11.7.3). The vector $\mathbf{\Sigma w}$

has entries $\sigma_i w_i$ for $i \le p$ and 0 beyond, so

$$\|\mathbf{\Sigma w}\|^2 = \sum_{i \le p} \sigma_i^2 |w_i|^2 \le \sigma_1^2 \sum_{i \le p} |w_i|^2 \le \sigma_1^2 \|\mathbf{w}\|^2 = \sigma_1^2 \|\mathbf{z}\|^2,$$

since σ_1 is the largest of the σ_i. Taking $\mathbf{z} = \mathbf{V}\mathbf{e}_1$, a unit vector, gives $\mathbf{w} = \mathbf{e}_1$ and $\|\mathbf{Xz}\| = \sigma_1$, so the bound is attained.

(b) Use the orthonormal basis $\mathcal{B}$ of Theorem 13.13.3 and write a vector of V by its coordinate vector $(\mathbf{y}, \mathbf{z})$ with $\mathbf{y} \in F^r$ and $\mathbf{z} \in F^{n-r}$; since $\mathcal{B}$ is orthonormal, Theorem 11.2.1 (c) gives $\|v\|^2 = \|\mathbf{y}\|^2 + \|\mathbf{z}\|^2$, and Pv has coordinates $(\mathbf{y} + \mathbf{Xz}, \mathbf{0})$. By Corollary 11.1.6, then (a), then Theorem 11.1.5 in $\mathbb{R}^2$ applied to $(\|\mathbf{y}\|, \|\mathbf{z}\|)$ and $(1, \sigma_1)$,

$$\begin{aligned} \|Pv\| = \|\mathbf{y} + \mathbf{Xz}\| &\le \|\mathbf{y}\| + \sigma_1 \|\mathbf{z}\| \\ &\le \sqrt{1 + \sigma_1^2}\, \sqrt{\|\mathbf{y}\|^2 + \|\mathbf{z}\|^2} = \sqrt{1 + \sigma_1^2}\, \|v\|. \end{aligned}$$

For equality, let $\mathbf{z}_0$ be the unit vector of (a), so $\mathbf{Xz}_0 = \sigma_1 \mathbf{u}_0$ with $\|\mathbf{u}_0\| = 1$ (take $\mathbf{u}_0 = \mathbf{Xz}_0/\sigma_1$ if $\sigma_1 > 0$; if $\sigma_1 = 0$ take any unit $\mathbf{u}_0$). Put

$$\mathbf{y} = \frac{\mathbf{u}_0}{\sqrt{1 + \sigma_1^2}}, \qquad \mathbf{z} = \frac{\sigma_1 \mathbf{z}_0}{\sqrt{1 + \sigma_1^2}}.$$

Then $\|\mathbf{y}\|^2 + \|\mathbf{z}\|^2 = 1$ and $\mathbf{y} + \mathbf{Xz} = (1 + \sigma_1^2)\mathbf{u}_0 / \sqrt{1 + \sigma_1^2}$, whose norm is $\sqrt{1 + \sigma_1^2}$.

(c) When $n = 2$ and $r = 1$, $\mathbf{X}$ is the 1×1 matrix (x), whose only singular value is $|x|$; the formula becomes $\sqrt{1 + |x|^2}$. When P is an orthogonal projection, $\mathbf{X} = \mathbf{O}$, so $\sigma_1 = 0$ and the largest value is 1 — an orthogonal projection never lengthens a vector, which is clause (e) of Theorem 13.13.1 with the sharpest constant.

Chapter 14

Bilinear and Quadratic Forms

An inner product does two jobs at once. It **pairs** two vectors and returns a scalar, linearly in each of them; and it **measures**, because $\langle v, v\rangle$ is a positive number whenever $v \neq 0$. Chapters 11 to 13 leaned on the second job constantly: lengths, angles, orthonormal bases, least squares, positive definiteness, and the singular value decomposition are all consequences of positivity. This chapter keeps the pairing and throws the measuring away. What is left is a **bilinear form**, a function $\beta\colon V \times V \to F$ linear in each slot, with no sign condition, no symmetry required, and no field restriction. The question of the chapter is what survives.

A surprising amount does. Forms still have matrices, and the matrices still change by a rule when the basis changes — but the rule is $\mathbf{A} \mapsto \mathbf{P}^{\top}\mathbf{A}\mathbf{P}$, not $\mathbf{A} \mapsto \mathbf{P}^{-1}\mathbf{A}\mathbf{P}$. The organizing relation is therefore **congruence**, not similarity, and the two are genuinely different: congruent matrices need not share a single eigenvalue. What they do share is a rank, a symmetry type, and — over $\mathbb{R}$ — a triple of signs, which is Sylvester's law of inertia and the central classification theorem of the chapter. Beyond $\mathbb{R}$ the classification depends on the arithmetic of the field, and beyond symmetry there is a second world, the alternating forms, where every nondegenerate example looks like every other of the same size. Along the way we meet vectors on which a form vanishes without being zero, the groups that preserve a form, the geometry that physics runs on, and one algebra built to make a quadratic form into a square.

The field is a general F here. Since Chapter 10 the book has worked over $\mathbb{R}$ or $\mathbb{C}$, because inner products need them; this chapter opens the field back up, and that is the point rather than an afterthought. Chapters 11 to 13 turn out to have been the positive definite special case of one line of this theory. Two consequences of the general setting are worth flagging in advance. First, "positive" has no meaning over a field with no order, so the definiteness arguments of Chapter 13 simply cannot be run; what replaces them is either a square-class computation or nothing at all. Second, when $1 + 1 = 0$ — characteristic 2, Definition 0.7.2 — the identity $\beta(u, v) = \frac{1}{2}(q(u + v) - q(u) - q(v))$ is unavailable, and symmetric and alternating forms stop being complementary. Every statement below that needs a characteristic hypothesis carries one, and every statement without one holds over every field.

What you need. From Chapter 7, multilinearity and alternation, which this chapter reuses rather than redefines (Definition 7.2.1, Definition 7.2.2, Definition 7.2.3, Theorem 7.2.1, Theorem 7.2.2), together with the determinant and its multiplicativity (Definition 7.4.1, Theorem 7.5.4, Theorem 7.4.5). From Chapter 2, the rank of a linear map (Definition 2.3.2). From Chapter 4, the matrix of a map, change of coordinates, and similarity (Definition 4.1.1, Theorem 4.1.1, Definition 4.2.1, Theorem 4.2.1, Theorem 4.2.3, Definition 4.2.2), together with equivalence of matrices and the rank normal form (Definition 4.3.1, Theorem 4.3.3). From Chapter 5, the dual space and its bases (Definition 5.1.1, Definition 5.1.2, Theorem 5.1.3, Corollary 5.1.4,

Theorem 5.5.1, Theorem 5.3.2). From Chapter 0, fields and characteristic (Definition 0.7.1, Definition 0.7.2, Theorem 0.7.5), the transpose and symmetry (Definition 0.9.10, Definition 0.9.11, Theorem 0.9.5), and groups (Definition 0.10.1, Definition 0.10.2). From Chapter 1, the dimension counting that every classification here ends with: direct sums and complements (Theorem 1.7.4, Theorem 1.7.5, Theorem 1.7.6), the dimension formula for a sum of subspaces (Theorem 1.7.3), and the step from an inclusion to an equality (Theorem 1.6.8). From Chapter 3, the $\mathbf{LDL}^\top$ factorization (Theorem 3.6.8). From Chapters 11 to 13, everything used for the real and complex specializations: the orthogonal and unitary groups (Definition 11.7.2, Definition 11.7.3), the real spectral theorem and principal axes (Theorem 12.5.1, Corollary 12.5.2, Theorem 12.5.5), the orthogonal group in low dimension (Proposition 11.7.4, Theorem 12.6.1), and positivity (Definition 13.1.1, Theorem 13.1.4, Theorem 13.1.5, Proposition 13.1.2, Theorem 13.4.1). Those last results are used **only** where the field is $\mathbb{R}$ or $\mathbb{C}$; they are never hypotheses of the general theory.

14.0.1 Roadmap

- **Bilinear forms and their matrices.** What a form is, the Gram matrix that records it, the bijection between forms and matrices, and the form's companion map $V \to V^*$, whose kernel is the radical and whose invertibility is nondegeneracy.
- **Congruence.** Changing the basis replaces $\mathbf{A}$ by $\mathbf{P}^\top\mathbf{AP}$. What that relation preserves — rank, symmetry type, nondegeneracy, the square class of the determinant — and what it destroys, starting with the eigenvalues.
- **Three kinds of form.** Symmetric, alternating, skew-symmetric, and Hermitian; why the middle two coincide exactly when the characteristic is not 2; polarization, and the quadratic form.
- **Completing the square.** Over any field of characteristic $\neq 2$, every symmetric form has an orthogonal basis. The proof is an induction, and the algorithm is symmetric row-and-column elimination.
- **Inertia.** Over $\mathbb{R}$ the number of positive, negative, and zero diagonal entries does not depend on the diagonalization. This classifies real symmetric matrices up to congruence and reconciles congruence with the spectral theorem.
- **Definiteness revisited.** Chapter 13's tests, restated as statements about signatures, and the second-derivative test they decide.
- **Where the form vanishes.** Isotropic vectors, totally isotropic subspaces, hyperbolic planes, and the Witt decomposition of a nondegenerate space.
- **Witt's theorems.** An isometry between subspaces extends to the whole space; one summand of an orthogonal sum may be canceled from both sides.
- **Symplectic forms.** A nondegenerate alternating form lives only in even dimension and has exactly one shape, over every field, with no characteristic hypothesis anywhere.
- **The classical groups.** $\mathrm{O}(n)$, $\mathrm{O}(p,q)$, $\mathrm{U}(n)$, and the symplectic group, all read off one equation $\mathbf{P}^\top\mathbf{AP} = \mathbf{A}$.
- **Minkowski space.** The signature $(1,3)$ form, the light cone, hyperbolic rotations, and the triangle inequality running backwards.
- **A first look at Clifford algebras.** Making a quadratic form into a square, and how $\mathbb{C}$ and the quaternions come out of it.

14.0.2 Named moves

- **Polarize.** A symmetric form and its quadratic function $q(v) = \beta(v,v)$ carry the same information, because β can be recovered from q by expanding $q(u+v)$ and dividing by 2. That division is the whole reason characteristic 2 is excluded whenever it is excluded.

- **Complete the square.** Every diagonalization theorem for symmetric forms in this chapter is this one manipulation, applied by induction on the dimension: isolate a variable that actually appears, absorb its cross terms into a perfect square, and recurse on what is left.
- **Split off a hyperbolic plane.** For indefinite forms this replaces "peel off an eigenvector". A single nonzero isotropic vector forces a two-dimensional piece of standard shape to split off orthogonally, and the induction runs on the rest.
- **Count signs, not values.** The numbers on a diagonalized form are not invariants; their signs are. Whenever a quantity refuses to be an invariant, ask whether some coarser feature of it is.
- **Congruence, not similarity.** At every theorem, ask which relation is in play. If the matrix is eaten twice by the same vector, the relation is congruence and eigenvalues are not available; if it is eaten once and returns a vector, the relation is similarity and they are.

14.1 Bilinear Forms and Their Matrices

Chapter 7 needed a function of several vectors that was linear in each of them separately, and called it a multilinear form (Definition 7.2.1). Chapter 11 needed a function of two vectors that was linear in the first, conjugate-linear in the second, and positive on the diagonal, and called it an inner product. This section takes the common skeleton of the two — two slots, linear in each, nothing else assumed — and studies it on its own. The payoff is immediate: such a function is recorded by a square matrix, exactly as a linear map is, and the whole rest of the chapter is about how that matrix behaves.

Throughout, V is a vector space over a field F. From the second subsection on, V is finite-dimensional with $\dim V = n \geq 1$, and $\mathcal{B} = (v_1, \dots, v_n)$ is an ordered basis of V. **No hypothesis on the characteristic of F is used anywhere in this section**, and none is used in the next one either.

14.1.1 Pairing without measuring

The dot product on F^n is a machine with two inputs and one scalar output. Over $\mathbb{R}$ it does two different jobs. It **pairs**: it takes $\mathbf{x}$ and $\mathbf{y}$ and returns $\sum_i x_i y_i$, linearly in each of them. And it **measures**: $\mathbf{x} \cdot \mathbf{x}$ is a sum of squares, so it is positive unless $\mathbf{x} = \mathbf{0}$, and from that one fact Chapter 11 extracted lengths, angles, orthonormal bases, and everything built on them.

Measuring is the fragile half. It needs an order on F, so it is meaningless over $\mathbb{C}$ without a conjugation and meaningless over $\mathbb{F}_5$ altogether. Pairing needs nothing: $\sum_i x_i y_i$ makes sense over every field, and so do plenty of other two-slot expressions, such as $x_1y_2 - x_2y_1$ or $x_1y_1 - x_2y_2$. Chapter 7 already met the first of these as the signed area, and met the general notion as the case $k = 2$ of a multilinear form. We give that case its own name, because from here on it is the object of study rather than a step toward the determinant.

A bilinear form pairs two vectors of the same space into a scalar, linearly in each slot separately.

Definition 14.1.1. Bilinear Form

Let V be a vector space over a field F. A **bilinear form** on V is a function $\beta\colon V \times V \to F$ that is linear in each argument when the other is held fixed: **for all** $u, u', v, v' \in V$ and **all** $c \in F$,

(B1) $\beta(u+u',v)=\beta(u,v)+\beta(u',v)$ and $\beta(cu,v)=c\,\beta(u,v)$;
(B2) $\beta(u,v+v')=\beta(u,v)+\beta(u,v')$ and $\beta(u,cv)=c\,\beta(u,v)$.

Equivalently, β is a 2-linear form on V in the sense of Definition 7.2.1.

In words: (B1) says that for each fixed v, the function $u\mapsto\beta(u,v)$ is a linear functional on V (Definition 5.1.1), and (B2) says the same for $v\mapsto\beta(u,v)$ with u fixed. The word **fixed** is the whole content. Nothing is assumed about how the two slots compare: $\beta(u,v)$ and $\beta(v,u)$ may be unrelated, and $\beta(v,v)$ may be any scalar at all, including 0 for a nonzero v.

Two consequences come free from linearity in one slot. Taking $c=0$ in (B1) gives $\beta(0,v)=0$ for every v, and likewise $\beta(u,0)=0$. And by induction on the number of terms,

$$\beta\Big(\sum_i a_iu_i,\ \sum_j b_jv_j\Big)=\sum_{i,j}a_ib_j\,\beta(u_i,v_j),$$

which is the identity everything in this section rests on: a bilinear form is determined by its values on pairs drawn from a spanning set.

Example. Six two-slot functions

Decide which of the following are bilinear forms on the given space over the given field.

(a) The dot product on F^n: $\beta(\mathbf{x},\mathbf{y})=x_1y_1+\cdots+x_ny_n$.
(b) On F^n, for a fixed $\mathbf{A}\in M_n(F)$: $\beta_{\mathbf{A}}(\mathbf{x},\mathbf{y})=\mathbf{x}^{\top}\mathbf{A}\mathbf{y}$.
(c) On F^2: $\beta(\mathbf{x},\mathbf{y})=x_1y_2-x_2y_1$.
(d) On $\mathbb{R}[x]_{\le 2}$: $\beta(f,g)=\int_0^1 f(t)g(t)\,\mathrm{d}t$.
(e) On any V: the zero function $\beta(u,v)=0$.
(f) On $\mathbb{R}^n$: $\beta(\mathbf{x},\mathbf{y})=\|\mathbf{x}\|\,\|\mathbf{y}\|$, the product of the Euclidean lengths.

Solution.

(a) Fix $\mathbf{y}$. Then $\mathbf{x}\mapsto\sum_i x_iy_i$ is a linear functional on F^n, since it is a fixed linear combination of the coordinate functionals (Theorem 5.1.1). Fixing $\mathbf{x}$ gives the same statement in $\mathbf{y}$. So (B1) and (B2) hold and the dot product is bilinear.

(b) With $\mathbf{A}=\mathbf{I}_n$ this is (a). In general, write out the entries: $\mathbf{x}^{\top}\mathbf{A}\mathbf{y}=\sum_{i,j}a_{ij}x_iy_j$. For fixed $\mathbf{y}$ this is $\sum_i(\sum_j a_{ij}y_j)x_i$, a fixed linear combination of the coordinates of $\mathbf{x}$, hence a linear functional; symmetrically in $\mathbf{y}$. So $\beta_{\mathbf{A}}$ is bilinear for **every** $\mathbf{A}$, symmetric or not. Part (c) below is the case $\mathbf{A}=\left(\begin{smallmatrix}0&1\\-1&0\end{smallmatrix}\right)$.

(c) This is the signed area of Chapter 7, and it is bilinear by Example (Alternating Forms) (a). It is worth noticing what it does on the diagonal: $\beta(\mathbf{x},\mathbf{x})=x_1x_2-x_2x_1=0$ for **every** $\mathbf{x}$. A bilinear form may vanish identically on the diagonal without being the zero form.

(d) For fixed g, the map $f\mapsto\int_0^1 fg$ is linear, because the integral of a sum is the sum of the integrals and constants come out. Symmetrically in g. So β is bilinear. Here $\beta(f,f)=\int_0^1 f^2\ge 0$, so over $\mathbb{R}$ this form does measure: it is the inner product of Chapter 11.

(e) Both clauses read $0=0+0$ and $0=c\cdot 0$. The zero form is bilinear on every space over every field. It is the degenerate case, and it matters: it shows that a bilinear form can carry no information whatsoever, and it is the form whose matrix below is the zero

matrix.

(f) Not bilinear. Fix $\mathbf{y} = \mathbf{e}_1$. Then $\beta(\mathbf{e}_1, \mathbf{e}_1) + \beta(-\mathbf{e}_1, \mathbf{e}_1) = 1 + 1 = 2$, while $\beta(\mathbf{e}_1 + (-\mathbf{e}_1), \mathbf{e}_1) = \beta(\mathbf{0}, \mathbf{e}_1) = 0$. Additivity in the first slot fails, so (B1) fails. The homogeneity half of (B1) also fails, since $\beta(-\mathbf{e}_1, \mathbf{e}_1) = 1 \neq -\beta(\mathbf{e}_1, \mathbf{e}_1)$.

Item (f) is the nonexample by minimal change worth remembering: the Cauchy–Schwarz inequality says that $|\mathbf{x} \cdot \mathbf{y}| \leq \|\mathbf{x}\|\|\mathbf{y}\|$, so $\|\mathbf{x}\|\|\mathbf{y}\|$ sits right next to the dot product numerically, and yet it is not a form at all, because the absolute value hidden in the norm destroys additivity. A second minimal change makes the same point more cheaply: $\beta(\mathbf{x}, \mathbf{y}) = x_1 y_1 + 1$ fails (B1) because $\beta(\mathbf{0}, \mathbf{0}) = 1 \neq 0$.

Warning. **Bilinear is not linear.** A bilinear form is not a linear function of the pair (u, v), and the two properties are not even compatible in general. For the dot product on $\mathbb{R}^2$, doubling both arguments multiplies the value by 4: $\beta(2\mathbf{e}_1, 2\mathbf{e}_1) = 4$, not $2\beta(\mathbf{e}_1, \mathbf{e}_1) = 2$. Sums may only be split **one slot at a time**, with the other slot held frozen. Chapter 7 gave this warning for multilinear forms in general; it is repeated here because it is the mistake readers make most often when a form is written with the familiar-looking symbol $\mathbf{x}^\top \mathbf{A}\mathbf{y}$.

Remark. The same definition with two different spaces, $\beta\colon V \times W \to F$ linear in each slot, is called a **bilinear pairing** of V and W. The basic example is the evaluation pairing $V^* \times V \to F$, $(\varphi, v) \mapsto \varphi(v)$, which is linear in φ by the definition of the operations on V^* (Definition 5.1.2) and linear in v because φ is. We insist on $V = W$ in Definition 14.1.1 because the questions of this chapter — is $\beta(u, v)$ equal to $\beta(v, u)$, what is $\beta(v, v)$, which vectors are orthogonal to which — only make sense when both slots hold vectors of the same space.

14.1.2 The matrix of a form

The displayed expansion above says that a bilinear form on a finite-dimensional space is determined by finitely many numbers: the values $\beta(v_i, v_j)$ on pairs of basis vectors. There are n^2 of them, they are indexed by a row and a column, and that is a matrix. Chapter 2 recorded a linear map by a matrix in exactly the same spirit (Definition 4.1.1), and the only difference is where the bases sit. There the two bases belonged to the input space and the output space. A form has no output space to speak of, but it has two **slots**, and each slot has to be fed a basis before the table can be written down. Throughout this chapter both slots hold vectors of the same V and we feed them the same basis, so the two choices coincide; the notation still records both, with the subscript naming the basis used on the first slot and the superscript the basis used on the second.

The matrix of a form is its table of values on pairs of basis vectors.

Definition 14.1.2. Matrix of a Bilinear Form

Let β be a bilinear form on a finite-dimensional V with ordered basis $\mathcal{B} = (v_1, \ldots, v_n)$. The **matrix of β with respect to $\mathcal{B}$**, also called its **Gram matrix**, is the matrix $[\beta]_{\mathcal{B}}^{\mathcal{B}} \in M_n(F)$

with entries

$$\left([\beta]_{\mathcal{B}}^{\mathcal{B}}\right)_{ij} := \beta(v_i, v_j).$$

In words: the entry in row i and column j is the value of β with the i-th basis vector in the **left** slot and the j-th in the **right** slot. The subscript is the basis read into the left slot and the superscript the basis read into the right one; here they are the same $\mathcal{B}$, and the notation writes it twice rather than leaving a slot empty. The order matters, because $\beta(v_i, v_j)$ and $\beta(v_j, v_i)$ need not agree; the convention chosen here is the one that makes the recovery formula below carry no transpose on the **matrix**.

Chapter 11 also called a matrix of pairings a Gram matrix, and the two conventions differ, so it is worth saying how. Definition 11.2.3 set $(\mathbf{G})_{ij} = \langle v_j, v_i \rangle$, with the indices crossed, because an inner product is conjugate-linear in its second slot and the crossing is what makes $\mathbf{G}$ Hermitian and $\mathbf{x}^*\mathbf{G}\mathbf{x}$ come out as a squared norm. A bilinear form has no conjugation to accommodate, so here the natural order is the uncrossed one, which is what makes the recovery formula below free of transposes. For a real symmetric form the two agree; in general each is the transpose of the other.

Nothing needs checking for well-definedness: once $\mathcal{B}$ is fixed, each entry is a single value of a function. What does need saying is that no information is lost.

Theorem 14.1.1. Forms and Matrices Correspond

Let V be an n-dimensional vector space over F with ordered basis $\mathcal{B}$, and let β be a bilinear form on V with $\mathbf{A} = [\beta]_{\mathcal{B}}^{\mathcal{B}}$.

(a) For all $u, v \in V$,
$$\beta(u, v) = [u]_{\mathcal{B}}^{\top}\, \mathbf{A}\, [v]_{\mathcal{B}}.$$

(b) $\mathbf{A}$ is the **only** matrix in $M_n(F)$ with this property.

(c) The assignment $\beta \mapsto [\beta]_{\mathcal{B}}^{\mathcal{B}}$ is a bijection from the set of bilinear forms on V onto $M_n(F)$, and it respects sums and scalar multiples: $[\beta + \gamma]_{\mathcal{B}}^{\mathcal{B}} = [\beta]_{\mathcal{B}}^{\mathcal{B}} + [\gamma]_{\mathcal{B}}^{\mathcal{B}}$ and $[c\beta]_{\mathcal{B}}^{\mathcal{B}} = c\,[\beta]_{\mathcal{B}}^{\mathcal{B}}$.

Idea. Part (a) is the expansion $\beta(\sum a_i v_i, \sum b_j v_j) = \sum_{i,j} a_i b_j \beta(v_i, v_j)$ with the double sum recognized as a matrix product. For (b), feed the property a pair of basis vectors and watch it return a single entry. For (c), (b) gives injectivity, and surjectivity is Example (Six two-slot functions) (b) transported to V through coordinates.

Proof.

(a) Write $[u]_{\mathcal{B}} = (a_1, \dots, a_n)$ and $[v]_{\mathcal{B}} = (b_1, \dots, b_n)$, so that $u = \sum_i a_i v_i$ and $v = \sum_j b_j v_j$. Expanding one slot at a time, by (B1) and then (B2),

$$\beta(u, v) = \sum_{i=1}^{n} \sum_{j=1}^{n} a_i b_j\, \beta(v_i, v_j) = \sum_{i,j} a_i\, a_{ij}\, b_j,$$

writing $a_{ij} = \beta(v_i, v_j)$ for the entries of $\mathbf{A}$. The last sum is the 1×1 matrix product $[u]_{\mathcal{B}}^{\top}\mathbf{A}[v]_{\mathcal{B}}$, by the definition of matrix multiplication (Definition 0.9.9).

(b) Suppose $\mathbf{M} \in M_n(F)$ also satisfies $\beta(u, v) = [u]_{\mathcal{B}}^{\top}\mathbf{M}[v]_{\mathcal{B}}$ for all u, v. Take $u = v_i$ and $v = v_j$. Then $[v_i]_{\mathcal{B}} = \mathbf{e}_i$ and $[v_j]_{\mathcal{B}} = \mathbf{e}_j$, so the right-hand side is $\mathbf{e}_i^{\top}\mathbf{M}\mathbf{e}_j = m_{ij}$, while

the left-hand side is $\beta(v_i, v_j) = a_{ij}$. Hence $m_{ij} = a_{ij}$ for all i, j, that is, $\mathbf{M} = \mathbf{A}$.

(c) *Injective.* If $[\beta]_{\mathcal{B}}^{\mathcal{B}} = [\gamma]_{\mathcal{B}}^{\mathcal{B}}$, then by (a) the two forms agree at every pair (u, v), so $\beta = \gamma$.

Surjective. Given $\mathbf{M} \in M_n(F)$, define $\gamma(u, v) := [u]_{\mathcal{B}}^{\top} \mathbf{M} [v]_{\mathcal{B}}$. The coordinate map $v \mapsto [v]_{\mathcal{B}}$ is linear (Corollary 2.5.5), and $(\mathbf{x}, \mathbf{y}) \mapsto \mathbf{x}^{\top} \mathbf{M} \mathbf{y}$ is bilinear on F^n by Example (Six two-slot functions) (b); composing a bilinear form with a linear map in each slot leaves it bilinear, so γ is a bilinear form on V. By (b) its matrix is $\mathbf{M}$.

Linear. The forms $\beta + \gamma$ and $c\beta$ are defined pointwise, so their values at (v_i, v_j) are $\beta(v_i, v_j) + \gamma(v_i, v_j)$ and $c\,\beta(v_i, v_j)$. These are the entries of $[\beta]_{\mathcal{B}}^{\mathcal{B}} + [\gamma]_{\mathcal{B}}^{\mathcal{B}}$ and of $c\,[\beta]_{\mathcal{B}}^{\mathcal{B}}$. This proves the theorem.

■

So bilinear forms on an n-dimensional space are the same thing as $n \times n$ matrices, once a basis is fixed — and in particular they form a vector space of dimension n^2. The qualification "once a basis is fixed" is not decoration; the next section is entirely about what happens when the basis is changed.

Example. Four Gram matrices

Compute $[\beta]_{\mathcal{B}}^{\mathcal{B}}$ in each case.

(a) The dot product on F^n, with $\mathcal{B} = \mathcal{E}$ the standard basis.

(b) $\beta_{\mathbf{A}}(\mathbf{x}, \mathbf{y}) = \mathbf{x}^{\top} \mathbf{A} \mathbf{y}$ on F^n, with $\mathcal{B} = \mathcal{E}$.

(c) The signed area $\beta(\mathbf{x}, \mathbf{y}) = x_1 y_2 - x_2 y_1$ on F^2, with $\mathcal{B} = \mathcal{E}$.

(d) $\beta(f, g) = \int_0^1 f g$ on $\mathbb{R}[x]_{\le 2}$, with $\mathcal{B} = (1, x, x^2)$.

Solution.

(a) $\beta(\mathbf{e}_i, \mathbf{e}_j) = \delta_{ij}$, so $[\beta]_{\mathcal{E}}^{\mathcal{E}} = \mathbf{I}_n$. The dot product is the form whose matrix is the identity.

(b) $\beta_{\mathbf{A}}(\mathbf{e}_i, \mathbf{e}_j) = \mathbf{e}_i^{\top} \mathbf{A} \mathbf{e}_j = a_{ij}$, so $[\beta_{\mathbf{A}}]_{\mathcal{E}}^{\mathcal{E}} = \mathbf{A}$. Together with Theorem 14.1.1 (c) this says that on F^n with the standard basis, "form" and "matrix" are two words for the same data, and $\mathbf{x}^{\top} \mathbf{A} \mathbf{y}$ is the general form.

(c) $\beta(\mathbf{e}_1, \mathbf{e}_1) = 0$, $\beta(\mathbf{e}_1, \mathbf{e}_2) = 1$, $\beta(\mathbf{e}_2, \mathbf{e}_1) = -1$, $\beta(\mathbf{e}_2, \mathbf{e}_2) = 0$, so

$$[\beta]_{\mathcal{E}}^{\mathcal{E}} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}.$$

The matrix is not symmetric, and its diagonal is zero — the two features that will separate this form from the dot product throughout the chapter.

(d) With $\mathcal{B} = (1, x, x^2)$ the (i, j) entry is $\int_0^1 t^{i-1} t^{j-1} \mathrm{d}t = 1/(i + j - 1)$, so

$$[\beta]_{\mathcal{B}}^{\mathcal{B}} = \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{3} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} \end{pmatrix}.$$

This matrix is symmetric, as it must be, since $\beta(f, g) = \beta(g, f)$. Its determinant is

1/2160, so it is invertible; what that means for β is the subject of the last part of this section.

Quick check. A bilinear form β on F^2 satisfies $\beta(\mathbf{e}_1,\mathbf{e}_1)=2$, $\beta(\mathbf{e}_1,\mathbf{e}_2)=-1$, $\beta(\mathbf{e}_2,\mathbf{e}_1)=3$, and $\beta(\mathbf{e}_2,\mathbf{e}_2)=0$. What is $\beta((1,1),(1,-1))$?

Solution. The matrix is $\mathbf{A}=\left(\begin{smallmatrix}2&-1\\3&0\end{smallmatrix}\right)$, and by Theorem 14.1.1 (a),

$$\beta((1,1),(1,-1))=\begin{pmatrix}1&1\end{pmatrix}\begin{pmatrix}2&-1\\3&0\end{pmatrix}\begin{pmatrix}1\\-1\end{pmatrix}=6.$$

Expanding by hand gives the same. The pair is $(\mathbf{e}_1+\mathbf{e}_2,\ \mathbf{e}_1-\mathbf{e}_2)$, so the four terms are

$$\beta(\mathbf{e}_1,\mathbf{e}_1)-\beta(\mathbf{e}_1,\mathbf{e}_2)+\beta(\mathbf{e}_2,\mathbf{e}_1)-\beta(\mathbf{e}_2,\mathbf{e}_2),$$

which is $2-(-1)+3-0=6$. Note that the two off-diagonal values enter with **opposite** roles: $\beta(\mathbf{e}_1,\mathbf{e}_2)$ comes from the minus sign in the right slot, $\beta(\mathbf{e}_2,\mathbf{e}_1)$ from the plus sign in the left one. Since $\beta(\mathbf{e}_1,\mathbf{e}_2)\neq\beta(\mathbf{e}_2,\mathbf{e}_1)$ here, swapping them would change the answer.

Now the warning that the rest of the chapter exists to manage.

Warning. **The matrix of a form and the matrix of a map are different objects with the same name.** Both are square arrays attached to V and a basis, both are written $\mathbf{A}$, and both change when the basis changes — by **different rules**. If $\mathbf{P}$ is the change-of-coordinates matrix between the old and the new basis, then

$$\text{a map: } \mathbf{A}\mapsto\mathbf{P}^{-1}\mathbf{A}\mathbf{P}, \qquad \text{a form: } \mathbf{A}\mapsto\mathbf{P}^{\top}\mathbf{A}\mathbf{P}.$$

The first is Theorem 4.2.3; the second is proved in the next section. They agree only when $\mathbf{P}^{\top}=\mathbf{P}^{-1}$. Concretely, take $\mathbf{A}=\operatorname{diag}(1,2)$ and $\mathbf{P}=\left(\begin{smallmatrix}1&1\\0&1\end{smallmatrix}\right)$. Then

$$\mathbf{P}^{-1}\mathbf{A}\mathbf{P}=\begin{pmatrix}1&-1\\0&2\end{pmatrix}, \qquad \mathbf{P}^{\top}\mathbf{A}\mathbf{P}=\begin{pmatrix}1&1\\1&3\end{pmatrix}.$$

The first still has eigenvalues 1 and 2; the second has eigenvalues $2\pm\sqrt{2}$. Same $\mathbf{A}$, same $\mathbf{P}$, two answers with nothing in common. Before writing $\mathbf{A}$ down, ask what it eats: a map eats one vector and returns a vector, so $\mathbf{A}\mathbf{x}$ is meaningful; a form eats two vectors and returns a scalar, so $\mathbf{x}^{\top}\mathbf{A}\mathbf{y}$ is meaningful and $\mathbf{A}\mathbf{x}$ is a meaningless intermediate. Almost every error in this chapter is this confusion.

14.1.3 The form as a map into the dual space

Fixing the right slot of β leaves a function of the left slot, and clause (B1) says that function is a linear functional. So each $v\in V$ produces an element of V^* (Definition 5.1.2), and clause (B2) says the production is itself linear. That is a linear map $V\to V^*$, and it is the cleanest way to say what it means for a form to be nondegenerate.

Write $R_\beta\colon V \to V^*$ for the map sending v to the functional $\beta(\cdot, v)$, that is,

$$(R_\beta(v))(u) = \beta(u, v) \qquad (u, v \in V).$$

The letter is local to this section and the subscript names the form; R is for the **right** slot, the one that is fixed. That $R_\beta(v)$ lies in V^* is (B1), and that R_β is linear is (B2). Its kernel has a name.

Definition 14.1.3. Radical of a Form

Let β be a bilinear form on V. The **radical** of β is

$$\operatorname{rad}(\beta) \coloneqq \{\, v \in V : \beta(u, v) = 0 \text{ for every } u \in V \,\} = \ker R_\beta,$$

a subspace of V. Its elements are the vectors that **every** vector is orthogonal to.

Definition 14.1.4. Nondegenerate Form

A bilinear form β on V is **nondegenerate** if $\operatorname{rad}(\beta) = \{0\}$, that is, if for every **nonzero** $v \in V$ there **exists** $u \in V$ with $\beta(u, v) \neq 0$. Otherwise β is **degenerate**.

In words: nondegeneracy says that no nonzero vector is invisible to the form. It does not say that $\beta(v, v) \neq 0$, and it does not say that β is positive in any sense; it only says that a vector which the form cannot distinguish from 0 must be 0. On a finite-dimensional space it is a rank condition.

Proposition 14.1.2. Nondegeneracy is invertibility of the Gram matrix

Let β be a bilinear form on an n-dimensional V with ordered basis $\mathcal{B}$, and put $\mathbf{A} = [\beta]_{\mathcal{B}}^{\mathcal{B}}$. Then:

(a) $\mathbf{A}$ is the matrix of R_β from the basis $\mathcal{B}$ to the dual basis $\mathcal{B}^*$; that is, $\mathbf{A} = [R_\beta]_{\mathcal{B}}^{\mathcal{B}^*}$.

(b) $\dim \operatorname{rad}(\beta) = n - \operatorname{rank} \mathbf{A}$.

(c) The following are equivalent: β is nondegenerate; $\mathbf{A}$ is invertible; R_β is an isomorphism $V \to V^*$.

Idea. Only (a) has content, and it is one line once the right formula is recalled: the coordinates of a functional in the dual basis are its values on $\mathcal{B}$ (Theorem 5.1.3 (c)), and the values of $R_\beta(v_j)$ on $\mathcal{B}$ are the entries of the j-th column of $\mathbf{A}$. Parts (b) and (c) are then Rank–Nullity and the invertible matrix theorem, applied to a map we now know the matrix of.

Proof.

(a) By Definition 4.1.1, the j-th column of $[R_\beta]_{\mathcal{B}}^{\mathcal{B}^*}$ is $[R_\beta(v_j)]_{\mathcal{B}^*}$. By Theorem 5.1.3 (c), a functional $\varphi \in V^*$ has $\varphi = \sum_i \varphi(v_i)\varphi_i$ in the dual basis $\mathcal{B}^* = (\varphi_1, \dots, \varphi_n)$, so the i-th coordinate of $R_\beta(v_j)$ is $(R_\beta(v_j))(v_i) = \beta(v_i, v_j) = a_{ij}$. Hence the (i, j) entry of $[R_\beta]_{\mathcal{B}}^{\mathcal{B}^*}$ is a_{ij}, as claimed.

(b) By (a) and Theorem 4.1.5, $\operatorname{rank} R_\beta = \operatorname{rank} \mathbf{A}$. Since $\operatorname{rad}(\beta) = \ker R_\beta$, Rank–Nullity (Theorem 2.3.1) gives $\dim \operatorname{rad}(\beta) = n - \operatorname{rank} \mathbf{A}$.

(c) By (b), $\operatorname{rad}(\beta) = \{0\}$ if and only if $\operatorname{rank}\mathbf{A} = n$, which by Theorem 3.4.4 is exactly invertibility of $\mathbf{A}$. Since $\dim V^* = n = \dim V$ (Corollary 5.1.4), the map R_β is an isomorphism if and only if it is injective (Corollary 2.3.2 (e)), that is, if and only if its kernel is $\{0\}$ (Theorem 2.2.3). This proves the proposition.

■

Remark. Freezing the **left** slot instead gives a second map $u \mapsto \beta(u, \cdot)$, whose kernel is the **left radical** $\{u : \beta(u, v) = 0 \text{ for all } v\}$. The same computation shows its matrix is $\mathbf{A}^\top$, so by Theorem 3.5.5 and Theorem 2.3.1 the two radicals have the same dimension — although, as Exercise (B3: Two radicals) shows, they need not be the same subspace. In particular "nondegenerate" means the same thing read from either side, and for the symmetric forms that occupy most of this chapter the two radicals coincide.

Example. Three radicals

Find $\operatorname{rad}(\beta)$ and decide degeneracy.

(a) $\beta(\mathbf{x}, \mathbf{y}) = x_1 y_1$ on F^2.

(b) $\beta(\mathbf{x}, \mathbf{y}) = x_1 y_1 - x_2 y_2$ on $\mathbb{R}^2$.

(c) The zero form on any $V \neq \{0\}$.

Solution.

(a) $[\beta]_{\mathcal{E}}^{\mathcal{E}} = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 0 \end{smallmatrix}\right)$, of rank 1, so the radical is one-dimensional by Proposition 14.1.2 (b). Explicitly, $\beta(\mathbf{u}, \mathbf{v}) = u_1 v_1$ vanishes for all $\mathbf{u}$ exactly when $v_1 = 0$, so $\operatorname{rad}(\beta) = \operatorname{span}(\mathbf{e}_2)$. Degenerate.

(b) $[\beta]_{\mathcal{E}}^{\mathcal{E}} = \operatorname{diag}(1, -1)$, invertible, so β is nondegenerate. Yet $\beta\big((1,1),(1,1)\big) = 1 - 1 = 0$: a nonzero vector on which the form vanishes when paired **with itself**. Nondegeneracy forbids a vector orthogonal to everything; it says nothing about a vector orthogonal to itself. This form is the one this chapter returns to twice, in the sections on isotropic vectors and on Minkowski space.

(c) The matrix is 0, of rank 0, and $\operatorname{rad}(\beta) = V$. This is the extreme degenerate case: every vector is invisible.

Warning. **Nondegenerate does not mean definite, and the radical is not the set where $\beta(v, v) = 0$.** Example (b) above is nondegenerate, and the set $\{\mathbf{v} : \beta(\mathbf{v}, \mathbf{v}) = 0\}$ is the pair of lines $x_2 = \pm x_1$ — not a subspace, and not the radical, which is $\{\mathbf{0}\}$. Example (c) of Example (Six two-slot functions) has $\beta(\mathbf{v}, \mathbf{v}) = 0$ for **every** $\mathbf{v}$ and is still nondegenerate, since its matrix $\left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)$ is invertible. An inner-product-trained reader expects "$\beta(v, v) = 0 \Rightarrow v = 0$", and that expectation is a theorem about positive definite forms over $\mathbb{R}$, not about forms.

Quick check. Let β be a bilinear form on a finite-dimensional V, and let $U = \operatorname{rad}(\beta)$. Is β nondegenerate on the subspace U?

Solution. No, unless $U = \{0\}$ — and then only for a silly reason. The restriction of β to $U \times U$ is the **zero** form, since $\beta(u, v) = 0$ already for $u \in V$ and $v \in U$. The zero form on a nonzero space is degenerate, so the restriction is nondegenerate exactly when $U = \{0\}$, that is, exactly when β was nondegenerate to begin with. The point of the question is that "nondegenerate" is a property of a form **on a specified space**, and restricting changes the space.

14.1.4 Exercises

14.1.4.1 A. Check your understanding

Exercise. A1

(a) Define what it means for $\beta\colon V \times V \to F$ to be a bilinear form, stating both clauses in full.

(b) Define the matrix $[\beta]_{\mathcal{B}}^{\mathcal{B}}$ of a bilinear form with respect to an ordered basis $\mathcal{B}$, and state how $\beta(u, v)$ is recovered from it.

(c) Determine whether the following is correct, with a reason: every bilinear form on F^n is $\beta(\mathbf{x}, \mathbf{y}) = \mathbf{x}^{\top}\mathbf{A}\mathbf{y}$ for exactly one $\mathbf{A} \in M_n(F)$.

(d) Define the radical of β and state what it means for β to be nondegenerate.

(e) Determine whether the following is correct, with a reason: if $\beta(v, v) = 0$ for some $v \neq 0$, then β is degenerate.

Solution.

(a) Definition 14.1.1: linearity in the left slot for each fixed right argument, (B1), and linearity in the right slot for each fixed left argument, (B2).

(b) Definition 14.1.2: $\left([\beta]_{\mathcal{B}}^{\mathcal{B}}\right)_{ij} = \beta(v_i, v_j)$. Recovery is Theorem 14.1.1 (a): $\beta(u, v) = [u]_{\mathcal{B}}^{\top}[\beta]_{\mathcal{B}}^{\mathcal{B}}[v]_{\mathcal{B}}$.

(c) Correct. Existence and uniqueness are Theorem 14.1.1 (c) and (b) with $\mathcal{B} = \mathcal{E}$, for which $[\mathbf{x}]_{\mathcal{E}} = \mathbf{x}$.

(d) Definition 14.1.3 and Definition 14.1.4: $\operatorname{rad}(\beta) = \{v : \beta(u, v) = 0 \text{ for all } u\}$, and β is nondegenerate when this is $\{0\}$.

(e) Incorrect. The two conditions are unrelated. The signed area on F^2 has $\beta(\mathbf{v}, \mathbf{v}) = 0$ for every $\mathbf{v}$ and is nondegenerate, since its matrix $\left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)$ is invertible (Proposition 14.1.2).

14.1.4.2 B. Practice

Exercise. B1: Which are forms?

Determine which of the following are bilinear forms on the given space. Justify your answer;

for those that fail, name the clause that fails and give a witness.

(a) On $\mathbb{R}^2$: $\beta(\mathbf{x},\mathbf{y}) = x_1y_1 + 2x_1y_2 - 3x_2y_1$.
(b) On $\mathbb{R}^2$: $\beta(\mathbf{x},\mathbf{y}) = x_1y_1 + 1$.
(c) On $\mathbb{R}[x]_{\le 2}$: $\beta(f,g) = f(1)g(2) - f(0)g(0)$.
(d) On $\mathbb{R}^2$: $\beta(\mathbf{x},\mathbf{y}) = x_1^2y_1$.
(e) On $M_2(\mathbb{R})$: $\beta(\mathbf{X},\mathbf{Y}) = \operatorname{tr}(\mathbf{X}^\top\mathbf{Y})$.

Solution.

(a) Bilinear: it is $\mathbf{x}^\top\mathbf{A}\mathbf{y}$ with $\mathbf{A} = \left(\begin{smallmatrix} 1 & 2 \\ -3 & 0 \end{smallmatrix}\right)$, so Example (Six two-slot functions) (b) applies.
(b) Not bilinear. Clause (B1) fails: $\beta(\mathbf{0},\mathbf{0}) = 1$, whereas homogeneity with $c = 0$ forces $\beta(\mathbf{0},\mathbf{y}) = 0$ for any bilinear form.
(c) Bilinear. Evaluation at a point is linear (Theorem 0.8.5), so for fixed g the map $f \mapsto g(2)f(1) - g(0)f(0)$ is a linear combination of two linear functionals, hence linear; symmetrically in g.
(d) Not bilinear. Clause (B1) fails in the additive half: with $\mathbf{y} = \mathbf{e}_1$, $\beta(\mathbf{e}_1 + \mathbf{e}_1, \mathbf{e}_1) = 4$ while $\beta(\mathbf{e}_1,\mathbf{e}_1) + \beta(\mathbf{e}_1,\mathbf{e}_1) = 2$. The homogeneity half fails too, since $\beta(c\mathbf{x},\mathbf{y}) = c^2x_1^2y_1$.
(e) Bilinear. By Theorem 0.9.5 and linearity of the trace (Theorem 0.9.6), for fixed $\mathbf{Y}$ the map $\mathbf{X} \mapsto \operatorname{tr}(\mathbf{X}^\top\mathbf{Y})$ is linear, and for fixed $\mathbf{X}$ the map $\mathbf{Y} \mapsto \operatorname{tr}(\mathbf{X}^\top\mathbf{Y})$ is linear. This is the Frobenius inner product of Chapter 11, seen as a form.

Exercise. B2: A Gram matrix from sample values

On $V = \mathbb{R}[x]_{\le 2}$, define $\beta(f,g) = f(0)g(0) + f(1)g(1) + f(2)g(2)$.

(a) Verify that β is a bilinear form and compute $[\beta]_{\mathcal{B}}^{\mathcal{B}}$ for $\mathcal{B} = (1, x, x^2)$.
(b) Compute $\beta(1 + x,\ x^2 - 1)$ directly from the definition, and again from the matrix. Check that the answers agree.
(c) Hence decide whether β is nondegenerate.

Solution.

(a) Each of the three terms is a product of two evaluations, hence bilinear as in Exercise (B1: Which are forms?) (c), and a sum of bilinear forms is bilinear by Theorem 14.1.1 (c). Numbering rows and columns $1, 2, 3$ for $1, x, x^2$, the (i,j) entry is $\sum_{t\in\{0,1,2\}} t^{\,i-1}t^{\,j-1}$, with $0^0 = 1$. So

$$[\beta]_{\mathcal{B}}^{\mathcal{B}} = \begin{pmatrix} 3 & 3 & 5 \\ 3 & 5 & 9 \\ 5 & 9 & 17 \end{pmatrix}.$$

(b) Directly: $f = 1 + x$ takes the values $1, 2, 3$ at $0, 1, 2$, and $g = x^2 - 1$ takes the values $-1, 0, 3$. So $\beta(f,g) = 1(-1) + 2\cdot 0 + 3\cdot 3 = 8$. From the matrix, $[f]_{\mathcal{B}} = (1, 1, 0)$ and $[g]_{\mathcal{B}} = (-1, 0, 1)$, and

$$[\beta]_{\mathcal{B}}^{\mathcal{B}}[g]_{\mathcal{B}} = (2, 6, 12), \qquad (1, 1, 0)\cdot(2, 6, 12) = 8.$$

(c) Let $\mathbf{M}$ be the matrix with rows $(1, t, t^2)$ for $t = 0, 1, 2$. Then $[\beta]_{\mathcal{B}}^{\mathcal{B}} = \mathbf{M}^\top \mathbf{M}$, and $\det \mathbf{M} = 2$ by the Vandermonde formula (Theorem 7.8.1), so $\det [\beta]_{\mathcal{B}}^{\mathcal{B}} = 4 \neq 0$ by Theorem 7.5.4 and Theorem 7.4.5. By Proposition 14.1.2, β is nondegenerate.

Exercise. B3: Two radicals

On F^3, let $\beta(\mathbf{x}, \mathbf{y}) = \mathbf{x}^\top \mathbf{A} \mathbf{y}$ with

$$\mathbf{A} = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \\ 1 & 2 & 3 \end{pmatrix}.$$

Find $\operatorname{rad}(\beta)$ and the left radical, give a basis of each, and decide whether they are equal.

Solution. Every row of $\mathbf{A}$ is a multiple of $(1, 2, 3)$, so $\operatorname{rank} \mathbf{A} = 1$ and both radicals have dimension $3 - 1 = 2$, by Proposition 14.1.2 (b) and the remark following it.

The radical is $\operatorname{nul}(\mathbf{A}) = \{\mathbf{y} : y_1 + 2y_2 + 3y_3 = 0\}$, with basis $(-2, 1, 0)$ and $(-3, 0, 1)$.

The left radical is $\operatorname{nul}(\mathbf{A}^\top)$. Every row of $\mathbf{A}^\top$ is a multiple of $(1, 2, 1)$, so it is $\{\mathbf{x} : x_1 + 2x_2 + x_3 = 0\}$, with basis $(-2, 1, 0)$ and $(-1, 0, 1)$.

They are **not** equal: $(-3, 0, 1)$ lies in the radical but $-3 + 0 + 1 = -2 \neq 0$, so it is not in the left radical. Equal dimensions, different subspaces. In particular β is degenerate.

14.1.4.3 C. Going deeper

Exercise. C1: Over $\mathbb{F}_2$

Let $\beta(\mathbf{x}, \mathbf{y}) = x_1 y_1 + x_2 y_2$ on $V = \mathbb{F}_2^2$.

(a) Write down $[\beta]_{\mathcal{E}}^{\mathcal{E}}$ and prove that β is nondegenerate.

(b) Find every $\mathbf{v} \in V$ with $\beta(\mathbf{v}, \mathbf{v}) = 0$.

(c) Hence explain why the argument "$\beta(\mathbf{v}, \mathbf{v}) = 0$ only for $\mathbf{v} = \mathbf{0}$" cannot be used to prove nondegeneracy over a general field, even for the dot product.

Solution.

(a) $[\beta]_{\mathcal{E}}^{\mathcal{E}} = \mathbf{I}_2$, which is invertible, so β is nondegenerate by Proposition 14.1.2.

(b) $\beta(\mathbf{v}, \mathbf{v}) = v_1^2 + v_2^2$. In $\mathbb{F}_2$ every element satisfies $t^2 = t$, so this is $v_1 + v_2$, which vanishes for $\mathbf{v} = (0, 0)$ and $\mathbf{v} = (1, 1)$. So the nonzero vector $(1, 1)$ satisfies $\beta(\mathbf{v}, \mathbf{v}) = 0$.

(c) Over $\mathbb{R}$, $\beta(\mathbf{v}, \mathbf{v}) = v_1^2 + v_2^2$ is a sum of squares and vanishes only at $\mathbf{0}$, so the implication holds there; the proof uses that a sum of nonnegative reals is zero only when each term is. Over $\mathbb{F}_2$ there is no order, nothing is nonnegative, and the conclusion is false for $(1, 1)$ even though the form is the dot product and is nondegenerate. Nondegeneracy is a statement about the rank of a matrix and survives the change of field; the diagonal statement is a statement about positivity and does not.

Exercise. C2: Orthogonal complements for a form

Let β be a bilinear form on a vector space V with $\dim V = n$, and for a subspace $U \subseteq V$ put

$$U^{\perp_\beta} := \{\, v \in V : \beta(u, v) = 0 \text{ for every } u \in U \,\}.$$

(a) Prove that $U^{\perp_\beta}$ is a subspace and that $\dim U^{\perp_\beta} \geq n - \dim U$.

(b) Prove that equality holds when β is nondegenerate.

(c) Give an example with $\dim U = 1$ in which the inequality in (a) is strict.

Hint: consider the map $V \to U^*$ that sends v to the restriction of $\beta(\cdot, v)$ to U.

Solution.

(a) Let $\rho\colon V^* \to U^*$ be the restriction map and $S = \rho R_\beta \colon V \to U^*$, so $S(v) = \beta(\cdot, v)\big|_U$. Both factors are linear, so S is linear, and $U^{\perp_\beta} = \ker S$ by the definition of S; hence $U^{\perp_\beta}$ is a subspace. By Rank–Nullity (Theorem 2.3.1),

$$\dim U^{\perp_\beta} = n - \operatorname{rank} S \geq n - \dim U^* = n - \dim U,$$

using $\operatorname{im} S \subseteq U^*$ and $\dim U^* = \dim U$ (Corollary 5.1.4).

(b) If β is nondegenerate then R_β is an isomorphism (Proposition 14.1.2 (c)), and ρ is surjective (Theorem 5.5.1). So $S = \rho R_\beta$ is surjective, $\operatorname{rank} S = \dim U$, and the inequality in (a) is an equality.

(c) Take $\beta(\mathbf{x}, \mathbf{y}) = x_1 y_1$ on F^2 and $U = \operatorname{span}(\mathbf{e}_2)$. Then $\beta(\mathbf{e}_2, \mathbf{v}) = 0$ for every $\mathbf{v}$, so $U^{\perp_\beta} = F^2$ has dimension 2, while $n - \dim U = 1$. The form is degenerate, as (b) requires it to be.

Exercise. C3: The evaluation pairing

Let V be finite-dimensional and let $W = V^* \times V$. Define β on W by

$$\beta((\varphi, v), (\psi, w)) := \psi(v) + \varphi(w).$$

(a) Prove that β is a bilinear form on W.

(b) Prove that β is nondegenerate. *Hint: the two summands of an element of the radical can be killed separately.*

(c) Deduce that if $v \in V$ satisfies $\varphi(v) = 0$ for every $\varphi \in V^*$, then $v = 0$, and name the result of Chapter 5 this recovers.

Solution.

(a) Fix (ψ, w). The map $(\varphi, v) \mapsto \psi(v) + \varphi(w)$ is the sum of $(\varphi, v) \mapsto \psi(v)$, which is linear because ψ is and because projection onto the second summand is linear, and $(\varphi, v) \mapsto \varphi(w) = \operatorname{ev}_w(\varphi)$, which is linear because $\operatorname{ev}_w \in V^{**}$ (Theorem 5.3.1). So clause (B1) holds. Clause (B2) follows by the same argument with the roles of the two arguments exchanged, since the defining expression is symmetric in them.

(b) Suppose $(\psi, w) \in \operatorname{rad}(\beta)$, so $\psi(v) + \varphi(w) = 0$ for all $\varphi \in V^*$ and all $v \in V$. Taking

$\varphi = 0$ gives $\psi(v) = 0$ for every v, so $\psi = 0$. Taking $v = 0$ then gives $\varphi(w) = 0$ for every $\varphi \in V^*$, so $\mathrm{ev}_w = 0$ in V^{**}, and $w = 0$ because ev_V is injective (Theorem 5.3.1). Hence the radical is $\{0\}$.

(c) The second half of (b) is exactly that statement: $\varphi(v) = 0$ for all $\varphi \in V^*$ forces $v = 0$. It recovers the injectivity of the evaluation map $\mathrm{ev}_V : V \to V^{**}$ of Theorem 5.3.1, which is the first half of Theorem 5.3.2. So the nondegeneracy of this form is the statement that functionals separate points (Theorem 5.1.5).

14.2 Congruence

The matrix of a form in Definition 14.1.2 depends on a basis, and a basis is a choice. As with linear maps in Chapter 4, the useful question is therefore not "what is the matrix?" but "which matrices arise from the same form?". The answer here is $\mathbf{P}^{\top}\mathbf{A}\mathbf{P}$ rather than $\mathbf{P}^{-1}\mathbf{A}\mathbf{P}$, and one symbol's difference changes the whole theory: the new relation keeps the rank and the symmetry type of a matrix and throws its eigenvalues away. This section proves the change-of-basis rule, names the relation, sorts the invariants from the casualties, and settles the debt Chapters 12 and 13 ran up when they kept postponing all of this.

Throughout, V is a vector space over an arbitrary field F with $\dim V = n \geq 1$, and β is a bilinear form on V. **No hypothesis on the characteristic of F is needed in this section.**

14.2.1 The change-of-basis rule for a form

Two bases $\mathcal{B}$ and $\mathcal{B}'$ of V give two matrices for one form. The link between them is the change-of-coordinates matrix of Definition 4.2.1, and the computation is three lines, because Theorem 14.1.1 (a) already expresses the form through coordinates.

Theorem 14.2.1. Change of Basis for a Form

Let β be a bilinear form on a finite-dimensional V, let $\mathcal{B}$ and $\mathcal{B}'$ be ordered bases of V, and let $\mathbf{P} = [\mathrm{id}]_{\mathcal{B}'}^{\mathcal{B}}$ be the change-of-coordinates matrix from $\mathcal{B}'$ to $\mathcal{B}$. Then

$$[\beta]_{\mathcal{B}'}^{\mathcal{B}'} = \mathbf{P}^{\top}\,[\beta]_{\mathcal{B}}^{\mathcal{B}}\,\mathbf{P}.$$

Idea. There is nothing to discover: write the form in $\mathcal{B}$-coordinates, substitute $[v]_{\mathcal{B}} = \mathbf{P}[v]_{\mathcal{B}'}$ in both slots, and read off the matrix that appears between the two coordinate vectors. The uniqueness clause of Theorem 14.1.1 then says that matrix is the one we want, with no further checking.

Proof. Write $\mathbf{A} = [\beta]_{\mathcal{B}}^{\mathcal{B}}$. By Theorem 4.2.1 (a), $[v]_{\mathcal{B}} = \mathbf{P}[v]_{\mathcal{B}'}$ for every $v \in V$. So for all $u, v \in V$, using Theorem 14.1.1 (a) and then Theorem 0.9.5 (part 4),

$$\begin{aligned}
\beta(u,v) &= [u]_{\mathcal{B}}^{\top}\,\mathbf{A}\,[v]_{\mathcal{B}} \\
&= \left(\mathbf{P}[u]_{\mathcal{B}'}\right)^{\top}\mathbf{A}\left(\mathbf{P}[v]_{\mathcal{B}'}\right) \\
&= [u]_{\mathcal{B}'}^{\top}\left(\mathbf{P}^{\top}\mathbf{A}\mathbf{P}\right)[v]_{\mathcal{B}'}.
\end{aligned}$$

Thus $\mathbf{P}^{\top}\mathbf{A}\mathbf{P}$ is a matrix with the property that characterizes $[\beta]_{\mathcal{B}'}^{\mathcal{B}'}$, and by the uniqueness in Theorem 14.1.1 (b) it **is** $[\beta]_{\mathcal{B}'}^{\mathcal{B}'}$. This proves the theorem.

■

Notice that $\mathbf{P}$ is the same matrix as in Theorem 4.2.3: the change-of-coordinates matrix from the **new** basis to the **old** one, whose columns are the new basis vectors written in old coordinates. Only the way it acts has changed, from $\mathbf{P}^{-1}\mathbf{A}\mathbf{P}$ to $\mathbf{P}^{\top}\mathbf{A}\mathbf{P}$.

Example. Killing the cross term

On $\mathbb{R}^2$, let $\beta(\mathbf{x},\mathbf{y}) = x_1y_1 + 2x_1y_2 + 2x_2y_1 + x_2y_2$, so that

$$\mathbf{A} = [\beta]_{\mathcal{E}}^{\mathcal{E}} = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}.$$

Compute $[\beta]_{\mathcal{B}'}^{\mathcal{B}'}$ for $\mathcal{B}' = ((1,0),\ (-2,1))$, both by Theorem 14.2.1 and directly from Definition 14.1.2.

Solution. The columns of $\mathbf{P} = [\mathrm{id}]_{\mathcal{B}'}^{\mathcal{E}}$ are the vectors of $\mathcal{B}'$ in standard coordinates, so

$$\mathbf{P} = \begin{pmatrix} 1 & -2 \\ 0 & 1 \end{pmatrix}, \qquad \mathbf{A}\mathbf{P} = \begin{pmatrix} 1 & 0 \\ 2 & -3 \end{pmatrix},$$

and therefore

$$\mathbf{P}^{\top}\mathbf{A}\mathbf{P} = \begin{pmatrix} 1 & 0 \\ -2 & 1 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 2 & -3 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & -3 \end{pmatrix}.$$

Directly: with $\mathbf{v}_1' = (1,0)$ and $\mathbf{v}_2' = (-2,1)$, the four values are $\beta(\mathbf{v}_1',\mathbf{v}_1') = 1$, $\beta(\mathbf{v}_1',\mathbf{v}_2') = \beta(\mathbf{v}_2',\mathbf{v}_1') = 0$, and $\beta(\mathbf{v}_2',\mathbf{v}_2') = -3$. The two computations agree.

The new basis was not guessed. In coordinates the form reads $q(\mathbf{x}) = x_1^2 + 4x_1x_2 + x_2^2 = (x_1+2x_2)^2 - 3x_2^2$, a perfect square plus a leftover, and the substitution $y_1 = x_1 + 2x_2$, $y_2 = x_2$ — that is, $\mathbf{x} = \mathbf{P}\mathbf{y}$ — is what completing the square asks for. A later section turns this into a theorem and an algorithm.

14.2.2 Congruent matrices

Every invertible $\mathbf{P}$ is a change-of-coordinates matrix between some pair of bases (Proposition 4.2.2), so Theorem 14.2.1 says exactly this: the matrices of one form, over all bases, are the matrices $\mathbf{P}^{\top}\mathbf{A}\mathbf{P}$ with $\mathbf{P}$ invertible. That set deserves a name, and the name is the organizing idea of the chapter.

Two matrices are congruent when they record the same form in two bases.

Definition 14.2.1. Congruent Matrices

Let $\mathbf{A},\mathbf{B} \in M_n(F)$. We say $\mathbf{A}$ is **congruent** to $\mathbf{B}$, written $\mathbf{A} \simeq \mathbf{B}$, if there **exists** an **invertible** $\mathbf{P} \in M_n(F)$ such that

$$\mathbf{B} = \mathbf{P}^{\top}\mathbf{A}\mathbf{P}.$$

In words: congruence allows one matrix to be sandwiched by an invertible matrix and its transpose. The word **invertible** is essential; without it, $\mathbf{A} \mapsto \mathbf{P}^\top \mathbf{A}\mathbf{P}$ with $\mathbf{P} = \mathbf{O}$ would relate everything to $\mathbf{O}$. The **transpose**, not the inverse, is what distinguishes the relation from similarity (Definition 4.2.2), and the symbol $\simeq$ is chosen to sit near $\sim$ without being it.

Three immediate readings. First, by Theorem 14.2.1, $\mathbf{A} \simeq \mathbf{B}$ if and only if $\mathbf{A}$ and $\mathbf{B}$ are the matrices of a single bilinear form in two bases. Second, in the language of quadratic expressions on F^n, $\mathbf{A} \simeq \mathbf{B}$ says that the substitution $\mathbf{x} = \mathbf{P}\mathbf{y}$ turns $\mathbf{x}^\top \mathbf{A}\mathbf{x}$ into $\mathbf{y}^\top \mathbf{B}\mathbf{y}$: congruence is invertible change of variables in a quadratic expression. Third, over $\mathbb{R}$ it is the relation $\mathbf{A} \mapsto \mathbf{S}^* \mathbf{A}\mathbf{S}$ of Chapter 13, because there the star is the transpose.

Example. Congruences and a noncongruence

(a) $\mathbf{I}_2 \simeq \left(\begin{smallmatrix} 1 & 1 \\ 1 & 2 \end{smallmatrix}\right)$ over any field.
(b) $\mathbf{I}_2 \simeq \operatorname{diag}(4, 1)$ over $\mathbb{Q}$.
(c) $\mathbf{A} \simeq \mathbf{A}$ for every $\mathbf{A}$.
(d) $\mathbf{I}_2 \not\simeq \left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$ over any field.

Solution.

(a) Take $\mathbf{P} = \left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$, which is invertible with determinant 1. Then $\mathbf{P}^\top \mathbf{I}_2 \mathbf{P} = \mathbf{P}^\top \mathbf{P} = \left(\begin{smallmatrix} 1 & 1 \\ 1 & 2 \end{smallmatrix}\right)$.
(b) Take $\mathbf{P} = \operatorname{diag}(2, 1)$, invertible over $\mathbb{Q}$. Then $\mathbf{P}^\top \mathbf{I}_2 \mathbf{P} = \operatorname{diag}(4, 1)$. Scaling a basis vector by c scales the corresponding diagonal entry by c^2, which is the single most useful congruence to have in mind.
(c) Take $\mathbf{P} = \mathbf{I}_n$.
(d) Suppose $\left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right) = \mathbf{P}^\top \mathbf{I}_2 \mathbf{P} = \mathbf{P}^\top \mathbf{P}$ for some $\mathbf{P}$. Then, by Theorem 0.9.5, $(\mathbf{P}^\top \mathbf{P})^\top = \mathbf{P}^\top \mathbf{P}$, so the matrix would be symmetric; it is not. This is the nonexample by minimal change: one entry of $\left(\begin{smallmatrix} 1 & 1 \\ 1 & 2 \end{smallmatrix}\right)$ in (a) is altered, and congruence to $\mathbf{I}_2$ is lost — not for a subtle reason, but because the altered matrix is not the matrix of a symmetric form in any basis. The general statement is Proposition 14.2.4 below.

Before using the word "congruent" in both directions, we owe the usual check.

Proposition 14.2.2. Congruence is an equivalence relation

Congruence is an equivalence relation on $M_n(F)$ (Definition 0.5.2).

Proof. Reflexive: $\mathbf{A} = \mathbf{I}_n^\top \mathbf{A}\mathbf{I}_n$.

Symmetric: suppose $\mathbf{B} = \mathbf{P}^\top \mathbf{A}\mathbf{P}$ with $\mathbf{P}$ invertible, and put $\mathbf{Q} = \mathbf{P}^{-1}$, which is invertible. By Theorem 0.9.5 (part 4), $\mathbf{Q}^\top \mathbf{P}^\top = (\mathbf{P}\mathbf{Q})^\top = \mathbf{I}_n^\top = \mathbf{I}_n$, so

$$\mathbf{Q}^\top \mathbf{B}\mathbf{Q} = (\mathbf{Q}^\top \mathbf{P}^\top)\mathbf{A}(\mathbf{P}\mathbf{Q}) = \mathbf{I}_n \mathbf{A}\mathbf{I}_n = \mathbf{A}.$$

Transitive: if $\mathbf{B} = \mathbf{P}^\top \mathbf{A}\mathbf{P}$ and $\mathbf{C} = \mathbf{Q}^\top \mathbf{B}\mathbf{Q}$ with $\mathbf{P}, \mathbf{Q}$ invertible, then $\mathbf{P}\mathbf{Q}$ is invertible and

$\mathbf{C} = \mathbf{Q}^{\top}\mathbf{P}^{\top}\mathbf{A}\mathbf{P}\mathbf{Q} = (\mathbf{P}\mathbf{Q})^{\top}\mathbf{A}(\mathbf{P}\mathbf{Q})$. This proves the proposition.

■

So the matrices in $M_n(F)$ fall into congruence classes, and the classification problem of this chapter is to describe them. The rest of the section collects the tools for telling classes apart.

14.2.3 What congruence preserves

Three invariants come out of the definition with almost no work. The first is the rank, which we already know is the co-dimension of the radical (Proposition 14.1.2).

Theorem 14.2.3. Congruence preserves rank

Let $\mathbf{A}, \mathbf{B} \in M_n(F)$ with $\mathbf{A} \simeq \mathbf{B}$. Then $\operatorname{rank}\mathbf{A} = \operatorname{rank}\mathbf{B}$. Consequently, whether the associated form is nondegenerate is a property of the congruence class, and it is legitimate to speak of the **rank of a bilinear form**, meaning the rank of its matrix in any basis.

Proof. Write $\mathbf{B} = \mathbf{P}^{\top}\mathbf{A}\mathbf{P}$ with $\mathbf{P}$ invertible. Then $\mathbf{P}^{\top}$ is invertible, with inverse $(\mathbf{P}^{-1})^{\top}$ (Theorem 0.9.5). Multiplying by invertible matrices on either side does not change the rank (Theorem 3.5.9), so $\operatorname{rank}\mathbf{B} = \operatorname{rank}(\mathbf{A}\mathbf{P}) = \operatorname{rank}\mathbf{A}$. Nondegeneracy is the condition $\operatorname{rank}\mathbf{A} = n$ by Proposition 14.1.2 (c), so it too depends only on the class. The last sentence follows because, by Theorem 14.2.1, any two matrices of one form are congruent.

■

The second invariant is the symmetry type. Here "alternating" is used in the matrix sense that Definition 7.2.2 suggests: a **zero diagonal** on top of skew-symmetry. The next section separates the three conditions carefully; for now we only need that congruence respects each of them.

Proposition 14.2.4. Congruence preserves the symmetry type

Let $\mathbf{A}, \mathbf{B} \in M_n(F)$ with $\mathbf{B} = \mathbf{P}^{\top}\mathbf{A}\mathbf{P}$ for an invertible $\mathbf{P}$.

(a) If $\mathbf{A}^{\top} = \mathbf{A}$, then $\mathbf{B}^{\top} = \mathbf{B}$.
(b) If $\mathbf{A}^{\top} = -\mathbf{A}$, then $\mathbf{B}^{\top} = -\mathbf{B}$.
(c) If $\mathbf{A}^{\top} = -\mathbf{A}$ and every diagonal entry of $\mathbf{A}$ is 0, the same holds for $\mathbf{B}$.

Each implication is an equivalence, since congruence is symmetric (Proposition 14.2.2).

Proof. By Theorem 0.9.5, $\mathbf{B}^{\top} = (\mathbf{P}^{\top}\mathbf{A}\mathbf{P})^{\top} = \mathbf{P}^{\top}\mathbf{A}^{\top}\mathbf{P}$.

(a) If $\mathbf{A}^{\top} = \mathbf{A}$, this is $\mathbf{P}^{\top}\mathbf{A}\mathbf{P} = \mathbf{B}$.
(b) If $\mathbf{A}^{\top} = -\mathbf{A}$, it is $\mathbf{P}^{\top}(-\mathbf{A})\mathbf{P} = -\mathbf{B}$.
(c) Assume the hypothesis of (c). Then $\mathbf{B}$ is skew-symmetric by (b), so only the diagonal

is left to check. For any $\mathbf{x} \in F^n$, grouping the terms of the double sum in pairs,

$$\mathbf{x}^\top \mathbf{A}\mathbf{x} = \sum_{i<j} (a_{ij} + a_{ji}) x_i x_j + \sum_i a_{ii} x_i^2 = 0,$$

since $a_{ij} + a_{ji} = 0$ by skew-symmetry and $a_{ii} = 0$ by hypothesis. Writing $\mathbf{p}_1, \dots, \mathbf{p}_n$ for the columns of $\mathbf{P}$, the i-th diagonal entry of $\mathbf{B} = \mathbf{P}^\top \mathbf{A}\mathbf{P}$ is $b_{ii} = \mathbf{e}_i^\top \mathbf{P}^\top \mathbf{A}\mathbf{P}\mathbf{e}_i = \mathbf{p}_i^\top \mathbf{A}\mathbf{p}_i$, which is 0 by the display. This proves the proposition.

■

The third invariant is the determinant, but only in a weakened sense: congruence multiplies it by a square.

Theorem 14.2.5. Congruence changes the determinant by a square

Let $\mathbf{A}, \mathbf{B} \in M_n(F)$ with $\mathbf{B} = \mathbf{P}^\top \mathbf{A}\mathbf{P}$ for an invertible $\mathbf{P}$. Then

$$\det \mathbf{B} = (\det \mathbf{P})^2 \det \mathbf{A},$$

with $\det \mathbf{P} \neq 0$. In particular $\det \mathbf{A} = 0$ if and only if $\det \mathbf{B} = 0$, and when they are nonzero, $\det \mathbf{B} = c^2 \det \mathbf{A}$ for some $c \in F \setminus \{0\}$.

Proof. By Theorem 7.5.4 and Theorem 7.4.5,

$$\det \mathbf{B} = \det(\mathbf{P}^\top) \det \mathbf{A} \det \mathbf{P} = (\det \mathbf{P})^2 \det \mathbf{A}.$$

Since $\mathbf{P}$ is invertible, $\det \mathbf{P} \neq 0$ (Theorem 7.5.2), so $(\det \mathbf{P})^2 \neq 0$ and the two determinants vanish together. Taking $c = \det \mathbf{P}$ gives the last claim.

■

The determinant itself is therefore **not** a congruence invariant — Example (Congruences and a noncongruence) (b) moved it from 1 to 4. What is invariant is the determinant read modulo squares. Write $F^\times = F \setminus \{0\}$, and for $a, b \in F^\times$ declare a and b equivalent when $a = c^2 b$ for some $c \in F^\times$. This is an equivalence relation on $F^\times$ (Definition 0.5.2): it is reflexive with $c = 1$, symmetric with c^{-1}, and transitive by multiplying the two scalars. Its classes are called **square classes**, and the set of them is commonly written $F^\times/(F^\times)^2$.

Definition 14.2.2. Discriminant

Let β be a **nondegenerate** bilinear form on an n-dimensional V. The **discriminant** of β is the square class of $\det[\beta]_{\mathcal{B}}^{\mathcal{B}}$ in $F^\times$, for any ordered basis $\mathcal{B}$ of V.

This is well defined by Theorem 14.2.5 and Theorem 14.2.1: another basis changes the determinant by the square $(\det \mathbf{P})^2$, hence not its class. The discriminant is coarse but genuinely useful, and how much it sees depends entirely on the field. Over $\mathbb{C}$ every nonzero number is a square, so there is only one class and the discriminant says nothing. Over $\mathbb{R}$ there are two, $\{t > 0\}$ and $\{t < 0\}$, so the discriminant records one bit: the sign of the determinant. Over $\mathbb{Q}$ there are infinitely many, and the discriminant is a strong invariant.

Quick check. Over $\mathbb{R}$, are $\operatorname{diag}(1,1)$ and $\operatorname{diag}(-1,-1)$ distinguished by their discriminants?

Solution. No. Both determinants equal 1, so the discriminants agree. And yet the two matrices are not congruent over $\mathbb{R}$. Suppose $\operatorname{diag}(-1,-1) = \mathbf{P}^{\top}\mathbf{I}_2\mathbf{P} = \mathbf{P}^{\top}\mathbf{P}$ for an invertible $\mathbf{P}$. Evaluating both sides on $\mathbf{x} = \mathbf{e}_1$ gives

$$-1 = \mathbf{e}_1^{\top} \operatorname{diag}(-1,-1)\mathbf{e}_1 = \mathbf{e}_1^{\top}\mathbf{P}^{\top}\mathbf{P}\mathbf{e}_1 = \|\mathbf{P}\mathbf{e}_1\|^2 \geq 0,$$

a contradiction. So the discriminant is not a complete invariant, even over $\mathbb{R}$ and even for 2×2 matrices. Proving that the full "sign pattern" *is* an invariant, and that it is the complete one over $\mathbb{R}$, is the law of inertia, proved later in this chapter.

14.2.4 What congruence destroys

Now the negative result of the section, and the reason the previous three chapters could not have been this one.

Example. Congruent, with no eigenvalue in common

(a) $\mathbf{I}_2 \simeq \left(\begin{smallmatrix} 1 & 1 \\ 1 & 2 \end{smallmatrix}\right)$ by Example (Congruences and a noncongruence) (a). Compare their eigenvalues.

(b) $\left(\begin{smallmatrix} 1 & 2 \\ 2 & 1 \end{smallmatrix}\right) \simeq \operatorname{diag}(1,-3)$ by Example (Killing the cross term). Compare their eigenvalues.

Solution.

(a) $\mathbf{I}_2$ has the single eigenvalue 1, twice. The characteristic polynomial of $\left(\begin{smallmatrix} 1 & 1 \\ 1 & 2 \end{smallmatrix}\right)$ is x^2-3x+1, whose roots are $(3\pm\sqrt{5})/2$, that is, approximately 2.618 and 0.382. The two matrices share no eigenvalue. They do share a rank, 2; a symmetry type, symmetric; and a discriminant, since $\det = 1$ for both.

(b) The characteristic polynomial of $\left(\begin{smallmatrix} 1 & 2 \\ 2 & 1 \end{smallmatrix}\right)$ is $x^2-2x-3 = (x-3)(x+1)$, so its eigenvalues are 3 and -1. Those of $\operatorname{diag}(1,-3)$ are 1 and -3. Again no eigenvalue in common — but this time one positive and one negative on each side, and the determinants are -3 both times. The signs survived; the numbers did not. That is the pattern the law of inertia will explain.

Chapter 13 introduced this operation twice and both times deferred it to here. Section 1 of that chapter wrote that "the operation $\mathbf{A} \mapsto \mathbf{S}^*\mathbf{A}\mathbf{S}$ is called **congruence**, and Chapter 14 studies it for its own sake", keeping only that it preserves the sign of a positive form (Proposition 13.1.2). Section 4 repeated the promise in fuller form: "Chapter 14 studies congruence in its own right, including what it does preserve; for now, note that it certainly does **not** preserve eigenvalues, since $\mathbf{S}^*\mathbf{I}\mathbf{S} = \mathbf{S}^*\mathbf{S}$ can be any positive definite matrix at all."

Both debts are now paid, and it is worth saying the answer in one place. **What congruence preserves:** the rank (Theorem 14.2.3), and with it nondegeneracy and the dimension of the radical; the symmetry type (Proposition 14.2.4); and the square class of the determinant when that is nonzero (Theorem 14.2.5), the discriminant. Over $\mathbb{R}$ it preserves one thing more, the

triple of signs, and that is the deepest of the four; it is proved in the section on inertia. **What congruence destroys:** the eigenvalues, and therefore the characteristic polynomial, the trace, and the determinant itself. Not every similarity invariant is lost — rank and invertibility, parts (a) and (b) of Proposition 4.2.6, survive both relations — but every invariant that reads the eigenvalues is. Chapter 13's parenthetical remark is exactly right and can be made sharper: taking $\mathbf{A} = \mathbf{I}_n$ in Definition 14.2.1, the congruence class of $\mathbf{I}_n$ over $\mathbb{R}$ is the set of all matrices $\mathbf{P}^\top\mathbf{P}$ with $\mathbf{P}$ invertible, which by Theorem 13.1.5 (c) is the set of **all** positive definite matrices. One class, every spectrum with all eigenvalues positive.

Warning. **Congruent and similar are different relations, and neither implies the other.** By Example (Congruent, with no eigenvalue in common) (a), $\mathbf{I}_2 \simeq \left(\begin{smallmatrix} 1 & 1 \\ 1 & 2 \end{smallmatrix}\right)$, while the only matrix similar to $\mathbf{I}_2$ is $\mathbf{I}_2$ itself, since $\mathbf{P}^{-1}\mathbf{I}_2\mathbf{P} = \mathbf{I}_2$ for every invertible $\mathbf{P}$. In the other direction, $\operatorname{diag}(1, 2)$ is similar to $\left(\begin{smallmatrix} 1 & -1 \\ 0 & 2 \end{smallmatrix}\right)$ — conjugate by $\left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$ — but not congruent to it, because $\operatorname{diag}(1, 2)$ is symmetric and the other matrix is not (Proposition 14.2.4). So do not carry any habit across. In particular, "diagonalizable" means two different things in this chapter, and a matrix that is diagonalizable by congruence need not be diagonalizable by similarity, nor conversely.

14.2.5 Where the two relations meet

There is exactly one situation in which the confusion is harmless: when the change-of-basis matrix satisfies $\mathbf{P}^\top = \mathbf{P}^{-1}$, that is, when $\mathbf{P}$ is orthogonal (Definition 11.7.3).

Proposition 14.2.6. An orthogonal change of basis is both

Let $\mathbf{A} \in M_n(\mathbb{R})$ and $\mathbf{Q} \in \mathrm{O}(n)$. Then $\mathbf{Q}^\top\mathbf{A}\mathbf{Q} = \mathbf{Q}^{-1}\mathbf{A}\mathbf{Q}$, so this matrix is at once congruent and similar to $\mathbf{A}$. Consequently, if $\mathbf{A}$ is symmetric, then $\mathbf{A}$ is congruent to the diagonal matrix $\mathbf{D}$ of its eigenvalues, listed with multiplicity in any prescribed order.

Proof. The first statement is the definition of an orthogonal matrix, $\mathbf{Q}^\top\mathbf{Q} = \mathbf{I}_n$, which gives $\mathbf{Q}^\top = \mathbf{Q}^{-1}$; the matrix $\mathbf{Q}^\top\mathbf{A}\mathbf{Q}$ is then congruent to $\mathbf{A}$ by Definition 14.2.1, since $\mathbf{Q}$ is invertible, and similar to $\mathbf{A}$ by Definition 4.2.2.

For the second statement, let $\mathbf{A}$ be symmetric. By Corollary 12.5.2 there are $\mathbf{Q} \in \mathrm{O}(n)$ and a real diagonal $\mathbf{D}$, carrying the eigenvalues of $\mathbf{A}$ in any prescribed order, with $\mathbf{A} = \mathbf{Q}\mathbf{D}\mathbf{Q}^\top$. Multiplying on the left by $\mathbf{Q}^\top$ and on the right by $\mathbf{Q}$ gives $\mathbf{D} = \mathbf{Q}^\top\mathbf{A}\mathbf{Q}$, which is a congruence. This proves the proposition.

■

This is why the real spectral theorem is so strong, and it is what Chapter 12 meant when it said, of the principal axis theorem, that "an orthogonal change of variables is a similarity *and* a congruence, so it keeps the eigenvalues and the geometry at once". The general theory of this chapter has no orthogonal matrices to call on — a general field has no inner product, hence no orthogonal group in that sense — so the two relations part company and only congruence remains. What Proposition 14.2.6 gives over $\mathbb{R}$ is a bridge: it says that every real symmetric matrix is congruent to a diagonal matrix, and it identifies the diagonal entries as eigenvalues. The section on inertia will show that the entries are free to move but their signs are not, which

is how the two pictures are finally reconciled.

Quick check. Over $\mathbb{R}$, which matrices are congruent to $\mathbf{I}_n$?

Solution. Exactly the positive definite symmetric matrices. If $\mathbf{B} = \mathbf{P}^\top \mathbf{I}_n \mathbf{P} = \mathbf{P}^\top \mathbf{P}$ with $\mathbf{P}$ invertible, then $\mathbf{B}$ is symmetric and $\mathbf{x}^\top \mathbf{B}\mathbf{x} = \|\mathbf{P}\mathbf{x}\|^2 > 0$ for $\mathbf{x} \neq \mathbf{0}$, since $\mathbf{P}\mathbf{x} \neq \mathbf{0}$. Conversely every positive definite $\mathbf{B}$ is $\mathbf{P}^\top \mathbf{P}$ for an invertible $\mathbf{P}$ by Theorem 13.1.5. Compare the similarity class of $\mathbf{I}_n$, which contains $\mathbf{I}_n$ alone: one relation collapses this class to a point, the other spreads it over every positive definite matrix there is.

14.2.6 Exercises

14.2.6.1 A. Check your understanding

Exercise. A1

(a) State the change-of-basis rule for the matrix of a bilinear form, saying which matrix $\mathbf{P}$ appears.

(b) Define what it means for $\mathbf{A}, \mathbf{B} \in M_n(F)$ to be congruent.

(c) For each of rank, symmetry, determinant, eigenvalues, and the square class of a nonzero determinant, say whether it is a congruence invariant.

(d) Determine whether the following is correct, with a reason: congruent matrices are similar.

(e) State the one circumstance in which a change of basis is simultaneously a similarity and a congruence.

Solution.

(a) Theorem 14.2.1: $[\beta]_{\mathcal{B}'}^{\mathcal{B}'} = \mathbf{P}^\top [\beta]_{\mathcal{B}}^{\mathcal{B}} \mathbf{P}$, where $\mathbf{P} = [\mathrm{id}]_{\mathcal{B}'}^{\mathcal{B}}$ is the change-of-coordinates matrix from the new basis to the old one.

(b) Definition 14.2.1: $\mathbf{B} = \mathbf{P}^\top \mathbf{A}\mathbf{P}$ for some invertible $\mathbf{P} \in M_n(F)$.

(c) Invariant: rank (Theorem 14.2.3); symmetry (Proposition 14.2.4); the square class of a nonzero determinant (Theorem 14.2.5). Not invariant: the determinant itself, and the eigenvalues (Example (Congruent, with no eigenvalue in common)).

(d) Incorrect. $\mathbf{I}_2 \simeq \left(\begin{smallmatrix} 1 & 1 \\ 1 & 2 \end{smallmatrix}\right)$, but the only matrix similar to $\mathbf{I}_2$ is $\mathbf{I}_2$.

(e) When the change-of-coordinates matrix $\mathbf{P}$ satisfies $\mathbf{P}^\top = \mathbf{P}^{-1}$, that is, when $\mathbf{P}$ is orthogonal (Proposition 14.2.6).

14.2.6.2 B. Practice

Exercise. B1: A congruence in $\mathbb{Q}$

Let $\mathbf{A} = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$ and $\mathbf{P} = \left(\begin{smallmatrix} 1 & 1 \\ 1 & -1 \end{smallmatrix}\right)$, both over $\mathbb{Q}$.

(a) Compute $\mathbf{P}^\top \mathbf{A}\mathbf{P}$, and hence write down a matrix congruent to $\mathbf{A}$ that is diagonal.

(b) Find the eigenvalues of both matrices and compare them.

(c) Compute the discriminant of the form $\mathbf{x}^\top \mathbf{A}\mathbf{y}$ and check that it agrees for the two matrices.

Solution.

(a) $\mathbf{P}$ is invertible, with $\det \mathbf{P} = -2 \neq 0$. Then

$$\mathbf{AP} = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}, \qquad \mathbf{P}^\top \mathbf{AP} = \begin{pmatrix} 2 & 0 \\ 0 & -2 \end{pmatrix}.$$

So $\mathbf{A} \simeq \operatorname{diag}(2, -2)$ over $\mathbb{Q}$.

(b) The characteristic polynomial of $\mathbf{A}$ is $x^2 - 1$, so its eigenvalues are 1 and -1; those of $\operatorname{diag}(2, -2)$ are 2 and -2. No eigenvalue is shared, and each matrix has one positive and one negative eigenvalue.

(c) $\det \mathbf{A} = -1$ and $\det \operatorname{diag}(2, -2) = -4 = (-2)^2 \cdot (-1)$, which is the relation $\det \mathbf{B} = (\det \mathbf{P})^2 \det \mathbf{A}$ of Theorem 14.2.5. The two determinants differ by the square 4, so they lie in the same square class of $\mathbb{Q}^\times$ and the discriminants agree, as Definition 14.2.2 requires.

Exercise. B2: A noncongruence over $\mathbb{Q}$

(a) Prove that $\operatorname{diag}(1, 1)$ and $\operatorname{diag}(1, 2)$ are **not** congruent over $\mathbb{Q}$.

(b) Prove that they **are** congruent over $\mathbb{R}$, by exhibiting $\mathbf{P}$.

(c) Explain in one sentence why (a) and (b) are consistent.

Solution.

(a) Suppose $\operatorname{diag}(1, 2) = \mathbf{P}^\top \operatorname{diag}(1, 1)\mathbf{P}$ with $\mathbf{P} \in M_2(\mathbb{Q})$ invertible. By Theorem 14.2.5, $2 = (\det \mathbf{P})^2 \cdot 1$, so $\det \mathbf{P} \in \mathbb{Q}$ would satisfy $(\det \mathbf{P})^2 = 2$. No rational number squares to 2 (Theorem 0.2.3), a contradiction. Hence the two are not congruent over $\mathbb{Q}$.

(b) Take $\mathbf{P} = \operatorname{diag}(1, 1/\sqrt{2})$, which is invertible over $\mathbb{R}$. Then $\mathbf{P}^\top \operatorname{diag}(1, 2)\mathbf{P} = \operatorname{diag}(1, 1)$, and congruence is symmetric (Proposition 14.2.2).

(c) Congruence is a relation **over a specified field**: the matrix $\mathbf{P}$ is required to have entries in that field, and $1/\sqrt{2}$ is available in $\mathbb{R}$ but not in $\mathbb{Q}$. Equivalently, 2 is a square in $\mathbb{R}^\times$ and not in $\mathbb{Q}^\times$, so the discriminant separates the two matrices over $\mathbb{Q}$ and not over $\mathbb{R}$.

Exercise. B3: Completing the square, again

On $\mathbb{R}^2$, let $\beta(\mathbf{x}, \mathbf{y}) = x_1 y_1 - x_1 y_2 - x_2 y_1 + 2x_2 y_2$.

(a) Write down $\mathbf{A} = [\beta]_{\mathcal{E}}^{\mathcal{E}}$ and expand $q(\mathbf{x}) = \beta(\mathbf{x}, \mathbf{x})$ as a polynomial in x_1, x_2.

(b) Complete the square to find a basis $\mathcal{B}'$ with $[\beta]_{\mathcal{B}'}^{\mathcal{B}'} = \mathbf{I}_2$. Verify by computing $\mathbf{P}^\top \mathbf{AP}$.

(c) Hence find the eigenvalues of $\mathbf{A}$ and compare them with those of $\mathbf{I}_2$.

Solution.

(a) $\mathbf{A} = \left(\begin{smallmatrix} 1 & -1 \\ -1 & 2 \end{smallmatrix}\right)$, and $q(\mathbf{x}) = x_1^2 - 2x_1x_2 + 2x_2^2$.

(b) $q(\mathbf{x}) = (x_1 - x_2)^2 + x_2^2$. So the substitution $y_1 = x_1 - x_2$, $y_2 = x_2$, that is $x_1 = y_1 + y_2$ and $x_2 = y_2$, makes $q = y_1^2 + y_2^2$. This is $\mathbf{x} = \mathbf{P}\mathbf{y}$ with

$$\mathbf{P} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \qquad \mathbf{AP} = \begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix}, \qquad \mathbf{P}^{\top}\mathbf{AP} = \mathbf{I}_2.$$

By Theorem 14.2.1 the basis is $\mathcal{B}' = ((1,0),\ (1,1))$, the columns of $\mathbf{P}$.

(c) The characteristic polynomial of $\mathbf{A}$ is $x^2 - 3x + 1$, with roots $(3 \pm \sqrt{5})/2$; $\mathbf{I}_2$ has the eigenvalue 1 twice. The two matrices are congruent with no eigenvalue in common. What they share is that all eigenvalues are positive on both sides, which by the check above is no accident: the congruence class of $\mathbf{I}_2$ over $\mathbb{R}$ consists of exactly the positive definite matrices.

14.2.6.3 C. Going deeper

Exercise. C1: Congruence versus equivalence

(a) Prove that if $\mathbf{A} \simeq \mathbf{B}$ then $\mathbf{A}$ and $\mathbf{B}$ are equivalent in the sense of Definition 4.3.1.

(b) Give two matrices in $M_2(\mathbb{Q})$ that are equivalent but not congruent, and justify your example.

(c) Hence explain what "the congruence classes refine the equivalence classes" means, and say what extra information a congruence class carries that a rank does not.

Solution.

(a) If $\mathbf{B} = \mathbf{P}^{\top}\mathbf{AP}$ with $\mathbf{P}$ invertible, then $\mathbf{P}^{\top}$ is invertible (Theorem 0.9.5), so $\mathbf{B} = \mathbf{QAP}$ with $\mathbf{Q} = \mathbf{P}^{\top}$ and $\mathbf{P}$ both invertible, which is Definition 4.3.1.

(b) Take $\operatorname{diag}(1,1)$ and $\operatorname{diag}(1,2)$. Both have rank 2, so they are equivalent by Theorem 4.3.3 (c). They are not congruent over $\mathbb{Q}$ by Exercise (B2: A noncongruence over $\mathbb{Q}$) (a).

(c) By (a), each congruence class lies inside a single equivalence class, and by (b) an equivalence class may split into more than one congruence class; that is what "refine" means. An equivalence class is named by the rank alone (Theorem 4.3.3 (c)). A congruence class remembers the rank, and in addition the symmetry type and the discriminant — and over $\mathbb{R}$, the signs. Equivalence allows two independent changes of basis, one in each slot; congruence allows only one, used twice, and the extra invariants are the price of that constraint.

Exercise. C2: Orthogonal congruence

Let $\mathbf{A} \in M_n(\mathbb{R})$ be symmetric with eigenvalues $\lambda_1, \ldots, \lambda_n$, listed with multiplicity.

(a) Prove that $\mathbf{A} \simeq \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$.

(b) Deduce that if every λ_i is nonzero, then $\mathbf{A}$ is congruent to a diagonal matrix all of

whose entries are ± 1.

(c) Give two real symmetric 2×2 matrices that are congruent and have no eigenvalue in common.

Hint for (b): recall from Example (Congruences and a noncongruence) (b) what scaling a basis vector does to a diagonal entry.

Solution.

(a) This is Proposition 14.2.6: by Corollary 12.5.2 there is $\mathbf{Q} \in \mathrm{O}(n)$ with $\mathbf{Q}^{\top}\mathbf{A}\mathbf{Q} = \mathbf{D} = \mathrm{diag}(\lambda_1, \dots, \lambda_n)$, and $\mathbf{Q}$ is invertible, so $\mathbf{A} \simeq \mathbf{D}$.

(b) Let $\mathbf{S} = \mathrm{diag}(c_1, \dots, c_n)$ with $c_i = 1/\sqrt{|\lambda_i|}$, which is defined and nonzero because $\lambda_i \neq 0$. Then $\mathbf{S}$ is invertible and $\mathbf{S}^{\top}\mathbf{D}\mathbf{S} = \mathrm{diag}(c_1^2\lambda_1, \dots, c_n^2\lambda_n)$, whose i-th entry is $\lambda_i/|\lambda_i| = \pm 1$ according to the sign of λ_i. By (a) and transitivity (Proposition 14.2.2), $\mathbf{A}$ is congruent to that matrix.

(c) By (b), $\mathrm{diag}(1, -3)$ and $\mathrm{diag}(2, -1)$ are both congruent to $\mathrm{diag}(1, -1)$, hence to each other by Proposition 14.2.2; explicitly, $\mathbf{P} = \mathrm{diag}(\sqrt{2}, 1/\sqrt{3})$ gives $\mathbf{P}^{\top}\,\mathrm{diag}(1, -3)\mathbf{P} = \mathrm{diag}(2, -1)$. Their eigenvalues are $\{1, -3\}$ and $\{2, -1\}$, which are disjoint. Note that the **signs** match, one positive and one negative on each side; part (b) is the reason, and the law of inertia will say it is forced.

Exercise. C3: How much the discriminant sees

(a) Show that $\mathbb{C}^{\times}$ has exactly one square class, and that $\mathbb{R}^{\times}$ has exactly two.

(b) List the squares in $\mathbb{F}_5^{\times}$ and show that $\mathbb{F}_5^{\times}$ has exactly two square classes.

(c) Hence determine, for each of $\mathbb{C}$, $\mathbb{R}$, and $\mathbb{F}_5$, the largest number of pairwise noncongruent invertible matrices in M_n that the discriminant alone can distinguish.

Solution.

(a) Over $\mathbb{C}$, every $a \neq 0$ has a square root (Example (Square Roots of $5-12i$)), so $a = c^2 \cdot 1$ and every element is equivalent to 1: one class. Over $\mathbb{R}$, squares of nonzero reals are exactly the positive reals, so a and b are equivalent precisely when $a/b > 0$, which happens exactly when they have the same sign. The classes are the positive reals and the negative reals: two classes, and they are distinct since 1 and -1 are not related.

(b) The squares are $1^2 = 1$, $2^2 = 4$, $3^2 = 4$, $4^2 = 1$, so the set of nonzero squares is $\{1, 4\}$. Multiplying $\{1, 4\}$ by 2 gives $\{2, 3\}$, so the classes are $\{1, 4\}$ and $\{2, 3\}$: two classes, each of size 2, covering all four elements of $\mathbb{F}_5^{\times}$.

(c) The discriminant takes one value per square class, so it can separate at most as many classes as there are square classes: one over $\mathbb{C}$, two over $\mathbb{R}$, and two over $\mathbb{F}_5$. Over $\mathbb{C}$ it is therefore useless, and over $\mathbb{R}$ and $\mathbb{F}_5$ it sorts invertible matrices into at most two piles. By Exercise (C2: Orthogonal congruence) (b), over $\mathbb{R}$ every invertible symmetric matrix is congruent to one of the $n+1$ matrices $\mathbf{I}_p \oplus -\mathbf{I}_{n-p}$, whose determinant is $(-1)^{n-p}$; so the discriminant sees only the parity of $n - p$, and for $n \geq 2$ it cannot separate, say, $\mathbf{I}_2$ from $-\mathbf{I}_2$. That the $n+1$ matrices really are pairwise noncongruent is the law of inertia, not anything the discriminant can deliver.

14.3 Three Kinds of Form

A bilinear form has two slots, and so far nothing has related them. Almost every form that occurs in practice does relate them: the dot product does not care which vector goes first, while the determinant of a 2×2 matrix changes sign when its columns are swapped. This section names the three symmetry conditions the rest of the chapter runs on, shows that two of them are the same condition exactly when $1 + 1 \neq 0$ in the field, and extracts from a symmetric form the single-variable function $q(v) = \beta(v, v)$ that carries all of its information.

Throughout, V is a finite-dimensional vector space over a field F, and β is a bilinear form on V (Definition 14.1.1). The characteristic hypothesis is stated wherever it is needed and nowhere else.

14.3.1 Symmetric forms

The dot product on F^n satisfies $\mathbf{x} \cdot \mathbf{y} = \mathbf{y} \cdot \mathbf{x}$, and so does every integral pairing $\int_0^1 fg$, and so does $\mathbf{x}^\top \mathbf{A}\mathbf{y}$ whenever $\mathbf{A}$ happens to satisfy $\mathbf{A}^\top = \mathbf{A}$. The condition recurs, so it gets a name.

A symmetric form cannot tell its two arguments apart.

Definition 14.3.1. Symmetric Bilinear Form

A bilinear form $\beta\colon V \times V \to F$ is **symmetric** if

$$\beta(u, v) = \beta(v, u)$$

for all $u, v \in V$.

In words: the value is unchanged by swapping the two inputs. The condition is imposed for **every** pair, not just for pairs of basis vectors — although, as the next result shows, checking it on basis vectors is enough.

Three examples and one nonexample.

- **The dot product** on F^n, $\beta(\mathbf{x}, \mathbf{y}) = x_1 y_1 + \cdots + x_n y_n$. Each product $x_i y_i$ is unchanged by the swap because multiplication in F is commutative.
- **A symmetric matrix**, $\beta(\mathbf{x}, \mathbf{y}) = \mathbf{x}^\top \mathbf{A}\mathbf{y}$ with $\mathbf{A}^\top = \mathbf{A}$ (Definition 0.9.11). Here $\mathbf{x}^\top \mathbf{A}\mathbf{y}$ is a 1×1 matrix, so it equals its own transpose $\mathbf{y}^\top \mathbf{A}^\top \mathbf{x} = \mathbf{y}^\top \mathbf{A}\mathbf{x}$ by Theorem 0.9.5.
- **Evaluation pairings.** On $F[x]_{\leq n}$, fix scalars $t_1, \ldots, t_k \in F$ and set $\beta(p, q) = \sum_i p(t_i) q(t_i)$. Symmetry is again commutativity of F.
- **A nonexample by minimal change.** On F^2 put $\beta(\mathbf{x}, \mathbf{y}) = x_1 y_2$. It is bilinear, by the one-slot check of Definition 7.2.1. But $\beta(\mathbf{e}_1, \mathbf{e}_2) = 1$ and $\beta(\mathbf{e}_2, \mathbf{e}_1) = 0$, so the symmetry clause fails at that one pair. Exactly one clause has been broken: bilinearity survives untouched.

The matrix of a form (Definition 14.1.2) turns the condition into a matrix condition, and turns "for all u, v" into a finite check.

Proposition 14.3.1. Symmetry Is Visible in the Matrix

Let $\mathcal{B}$ be a basis of V and $\mathbf{A} = [\beta]_{\mathcal{B}}^{\mathcal{B}}$. Then β is symmetric if and only if $\mathbf{A}^\top = \mathbf{A}$.

Proof. ($\Rightarrow$) Suppose β is symmetric. Writing $\mathcal{B} = (v_1, \dots, v_n)$, the entries of $\mathbf{A}$ are $a_{ij} = \beta(v_i, v_j)$, so $a_{ij} = \beta(v_j, v_i) = a_{ji}$, which is Definition 0.9.11.

($\Leftarrow$) Suppose $\mathbf{A}^\top = \mathbf{A}$, and let $u, v \in V$. By Theorem 14.1.1,

$$\beta(u, v) = [u]_{\mathcal{B}}^\top \mathbf{A} [v]_{\mathcal{B}}.$$

The right-hand side is a 1×1 matrix, hence equal to its transpose, which by Theorem 0.9.5 is $[v]_{\mathcal{B}}^\top \mathbf{A}^\top [u]_{\mathcal{B}} = [v]_{\mathcal{B}}^\top \mathbf{A} [u]_{\mathcal{B}} = \beta(v, u)$. This proves the equivalence.

■

Because congruence sends $\mathbf{A}$ to $\mathbf{P}^\top \mathbf{A} \mathbf{P}$ and $(\mathbf{P}^\top \mathbf{A} \mathbf{P})^\top = \mathbf{P}^\top \mathbf{A}^\top \mathbf{P}$ by Theorem 0.9.5, a congruent copy of a symmetric matrix is symmetric. So the condition $\mathbf{A}^\top = \mathbf{A}$ does not depend on which basis was used to build $\mathbf{A}$ — as it must not, since by Proposition 14.3.1 it says something about β alone.

14.3.2 Alternating and skew-symmetric forms

Chapter 7 needed two ways of saying "the value collapses when two arguments coincide", and found that they are not the same condition. Both were defined there for k-linear forms, and a bilinear form is the case $k = 2$, so nothing needs redefining. Read at $k = 2$:

- β is **alternating** (Definition 7.2.2) if $\beta(v, v) = 0$ for **every** $v \in V$;
- β is **skew-symmetric** (Definition 7.2.3) if $\beta(u, v) = -\beta(v, u)$ for **all** $u, v \in V$.

The standard example of both is the signed area $\beta(\mathbf{x}, \mathbf{y}) = x_1 y_2 - x_2 y_1$ on F^2, which is the determinant of the matrix with columns $\mathbf{x}$ and $\mathbf{y}$. A second family, the one Section 9 will be about, is the **standard alternating form** on F^{2m},

$$\beta(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{m} (x_i y_{m+i} - x_{m+i} y_i),$$

which pairs the first block of coordinates against the second.

Each condition reads off the matrix, and the two readings differ by the diagonal.

Proposition 14.3.2. Alternation and Skewness in the Matrix

Let $\mathcal{B} = (v_1, \dots, v_n)$ be a basis of V and $\mathbf{A} = [\beta]_{\mathcal{B}}^{\mathcal{B}}$. Then:

(a) β is skew-symmetric if and only if $\mathbf{A}^\top = -\mathbf{A}$;
(b) β is alternating if and only if $\mathbf{A}^\top = -\mathbf{A}$ **and** $a_{ii} = 0$ for every i.

Proof.

(a) repeats the proof of Proposition 14.3.1 with a minus sign throughout.
(b) ($\Rightarrow$) If β is alternating then $a_{ii} = \beta(v_i, v_i) = 0$, and β is skew-symmetric by Theorem 7.2.1 (a), so $\mathbf{A}^\top = -\mathbf{A}$ by (a).

($\Leftarrow$) Suppose $\mathbf{A}^\top = -\mathbf{A}$ with zero diagonal, and let $v \in V$ have coordinates $[v]_{\mathcal{B}} =$

$(c_1, \dots, c_n)$. By Theorem 14.1.1 and bilinearity,

$$\beta(v,v) = \sum_{i,j} c_i c_j a_{ij} = \sum_i c_i^2 a_{ii} + \sum_{i<j} c_i c_j (a_{ij} + a_{ji}),$$

where the second sum collects the terms (i,j) and (j,i) together. Every a_{ii} is 0, and every $a_{ij} + a_{ji}$ is 0 because $\mathbf{A}^\top = -\mathbf{A}$. So $\beta(v,v) = 0$. This proves the equivalence.

■

Part (b) is the whole difficulty in one line: a skew-symmetric matrix has $a_{ii} = -a_{ii}$, and *usually* that forces $a_{ii} = 0$. Usually, but not always.

14.3.3 Where characteristic 2 earns its keep

Chapter 0 warned that in a field of characteristic 2 the argument "$x = -x$, so $2x = 0$, so $x = 0$" divides by $2 = 1+1$, which is 0 there, and it promised that this is why later results on symmetric bilinear forms would carry the hypothesis "characteristic $\neq 2$" (Definition 0.7.2). This is the theorem the promise was about.

Theorem 14.3.3. Alternating Versus Skew-Symmetric

Let F be a field and let V be a vector space over F.

(a) Every alternating bilinear form on V is skew-symmetric. No hypothesis on F or on V is needed.

(b) If $\operatorname{char} F \neq 2$, every skew-symmetric bilinear form on V is alternating. Again V is arbitrary.

(c) If $\operatorname{char} F = 2$ and $V \neq \{0\}$, there is a skew-symmetric bilinear form on V that is **not** alternating.

So the converse of (a) holds **if and only if** $\operatorname{char} F \neq 2$.

Idea. For (a), feed $u + v$ into both slots and expand: the two diagonal terms vanish by hypothesis, and what is left is exactly the skew identity. For (b), feed v into both slots of the skew identity, which gives $\beta(v,v) = -\beta(v,v)$; canceling needs $1+1$ to be invertible, and that is precisely what characteristic $\neq 2$ supplies. For (c), notice that in characteristic 2 the words "skew-symmetric" and "symmetric" mean the same thing, so any symmetric form that is nonzero on some vector **paired with itself** is a witness — and the dot product is one.

Proof.

(a) Let β be alternating and let $u, v \in V$. Expanding by bilinearity,

$$0 = \beta(u+v, u+v) = \beta(u,u) + \beta(u,v) + \beta(v,u) + \beta(v,v),$$

and the first and last terms are 0 because β is alternating. Hence $\beta(u,v) = -\beta(v,u)$. (This is Theorem 7.2.1 (a) at $k = 2$.)

(b) Let β be skew-symmetric and $v \in V$. Taking $u = v$ in the defining identity gives

$\beta(v,v) = -\beta(v,v)$, that is, $(1+1)\beta(v,v) = 0$. Since $\operatorname{char} F \neq 2$, we have $1+1 \neq 0$ in F: otherwise the smallest positive n with $n \cdot 1 = 0$ would be 2, which is Definition 0.7.2. A nonzero element of a field is invertible, so multiplying by $(1+1)^{-1}$ gives $\beta(v,v) = 0$. (This is Theorem 7.2.2 at $k = 2$.)

(c) Suppose $\operatorname{char} F = 2$, so that $-1 = 1$ in F. Since $V \neq \{0\}$ is finite-dimensional, it has a basis with at least one vector, and we may take $\varphi \in V^*$ to be the coordinate functional of its first vector v_1, so that $\varphi(v_1) = 1$. Define

$$\beta(u,v) = \varphi(u)\,\varphi(v).$$

This is bilinear, being a product of two linear functionals, and it is symmetric. Because $-1 = 1$, symmetric and skew-symmetric say the same thing, so β is skew-symmetric. But $\beta(v_1, v_1) = 1 \neq 0$, so β is not alternating.

Finally, (b) and (c) together say that the converse of (a) holds exactly when $\operatorname{char} F \neq 2$. This proves the theorem.

■

The concrete witness is worth writing down.

Example. The dot product over the field with two elements

Let $F = \mathbb{F}_2$ and $V = \mathbb{F}_2^2$, and let β be the dot product, $\beta(\mathbf{x},\mathbf{y}) = x_1y_1 + x_2y_2$. Decide whether β is symmetric, skew-symmetric, and alternating, and compute the function $\mathbf{v} \mapsto \beta(\mathbf{v},\mathbf{v})$ on all four vectors of V.

Solution. The matrix of β in the standard basis is $\mathbf{I}_2$. It is symmetric, so β is symmetric by Proposition 14.3.1. Over $\mathbb{F}_2$ we have $-\mathbf{I}_2 = \mathbf{I}_2$, so the matrix also satisfies $\mathbf{I}_2^{\top} = -\mathbf{I}_2$, and β is skew-symmetric by Proposition 14.3.2 (a). It is **not** alternating: the diagonal entries of $\mathbf{I}_2$ are $1 \neq 0$, and indeed $\beta(\mathbf{e}_1,\mathbf{e}_1) = 1$.

On the four vectors, $\beta(\mathbf{v},\mathbf{v}) = v_1^2 + v_2^2$. In $\mathbb{F}_2$ both elements satisfy $c^2 = c$, so this equals $v_1 + v_2$:

$$\beta(\mathbf{0},\mathbf{0}) = 0, \quad \beta(\mathbf{e}_1,\mathbf{e}_1) = 1, \quad \beta(\mathbf{e}_2,\mathbf{e}_2) = 1, \quad \beta(\mathbf{1},\mathbf{1}) = 0.$$

Notice what that list is: the function $\mathbf{v} \mapsto \beta(\mathbf{v},\mathbf{v})$ is the **linear** functional $v_1 + v_2$. In characteristic 2 a "quadratic" function coming from a bilinear form can be linear, which is the first sign that the notions below will not behave.

Warning. **In characteristic 2, "skew-symmetric" carries no information beyond "symmetric".** Since $-1 = 1$, the two conditions are literally the same condition, and the useful notion is the strictly stronger one, alternating. This is why Chapter 7 built the determinant on alternation rather than skewness, and why every statement below that pairs symmetric against alternating needs $\operatorname{char} F \neq 2$.

14.3.4 Splitting a form into two halves

In characteristic $\neq 2$ the two conditions are complementary in the strongest possible sense: every bilinear form is a symmetric one plus an alternating one, in exactly one way. The recipe is the one that splits a function on $\mathbb{R}$ into its even and odd parts.

Theorem 14.3.4. Symmetric Plus Alternating

Let $\operatorname{char} F \neq 2$ and let β be a bilinear form on V. Then there are a **unique** symmetric form β_s and a **unique** alternating form β_a on V with $\beta = \beta_s + \beta_a$, namely

$$\beta_s(u,v) = \tfrac{1}{2}(\beta(u,v) + \beta(v,u)),$$
$$\beta_a(u,v) = \tfrac{1}{2}(\beta(u,v) - \beta(v,u)).$$

In matrices: $\mathbf{A} = \frac{1}{2}(\mathbf{A} + \mathbf{A}^\top) + \frac{1}{2}(\mathbf{A} - \mathbf{A}^\top)$.

Proof. Since $\operatorname{char} F \neq 2$, the element $1+1$ is nonzero, hence invertible, and $\frac{1}{2}$ denotes $(1+1)^{-1}$. Both displayed formulas are bilinear, being combinations of the bilinear maps $(u,v) \mapsto \beta(u,v)$ and $(u,v) \mapsto \beta(v,u)$, and their sum is β. Swapping u and v leaves β_s unchanged, so β_s is symmetric; and $\beta_a(v,v) = \frac{1}{2}(\beta(v,v) - \beta(v,v)) = 0$, so β_a is alternating.

For uniqueness, suppose $\beta_s + \beta_a = \beta'_s + \beta'_a$ with β_s, β'_s symmetric and β_a, β'_a alternating, and put $\delta = \beta_s - \beta'_s = \beta'_a - \beta_a$. Then δ is symmetric, being a difference of symmetric forms, and alternating, being a difference of alternating forms. By Theorem 14.3.3 (a) it is also skew-symmetric, so for all u, v,

$$\delta(u,v) = \delta(v,u) = -\delta(u,v),$$

giving $(1+1)\delta(u,v) = 0$ and hence $\delta(u,v) = 0$. Therefore $\beta_s = \beta'_s$ and $\beta_a = \beta'_a$, as claimed. ∎

Both halves of the argument divided by $1+1$, and both failures are real. In characteristic 2 the formulas are meaningless, and more than that, the statement itself is false: alternating forms are then a *subset* of the symmetric ones, not a complement to them.

Warning. **In characteristic 2, symmetric and alternating forms are not complementary.** Over $\mathbb{F}_2$ take $\beta(\mathbf{x},\mathbf{y}) = x_1 y_2$ on $\mathbb{F}_2^2$. An alternating form is skew-symmetric (Theorem 14.3.3 (a)), and skew-symmetric means symmetric here, so **every** sum of a symmetric form and an alternating form is symmetric. But $\beta(\mathbf{e}_1, \mathbf{e}_2) = 1$ while $\beta(\mathbf{e}_2, \mathbf{e}_1) = 0$, so β is not symmetric and therefore cannot be written as such a sum at all. The decomposition does not merely lose uniqueness; existence fails.

14.3.5 The quadratic form

A bilinear form is a function of two vectors, which makes it awkward to plot, to differentiate, or to write on a page. Feeding the same vector into both slots turns it into a function of one vector. Chapter 12 met this function on $\mathbb{R}^n$ under the name **quadratic form** and used it to identify conics; here is the general definition, and the surprising fact that nothing is lost.

A quadratic form is what a bilinear form sees when both of its arguments are the same vector.

Definition 14.3.2. Quadratic Form

Let β be a bilinear form on V. The **quadratic form associated with** β is the function

$$q\colon V \to F, \qquad q(v) = \beta(v, v).$$

A function $q\colon V \to F$ is called **a quadratic form on** V if $q(v) = \beta(v, v)$ for some **symmetric** bilinear form β on V.

Two clauses deserve separate readings. The first says how to build q from a given β, for **any** β over **any** field. The second says which functions $V \to F$ are entitled to the name, and it insists on a **symmetric** β; the polarization theorem below will show that over a field of characteristic $\neq 2$ this symmetric β is then unique, so that "the" bilinear form of a quadratic form is well defined.

Insisting on symmetry costs nothing in characteristic $\neq 2$. If $q(v) = \beta(v, v)$ for some bilinear β, then by Theorem 14.3.4 we may split $\beta = \beta_s + \beta_a$, and $\beta_a(v, v) = 0$, so $q(v) = \beta_s(v, v)$ with β_s symmetric. The alternating part is invisible to q.

In coordinates, with $\mathbf{A} = [\beta]_{\mathcal{B}}^{\mathcal{B}}$ symmetric and $\mathbf{x} = [v]_{\mathcal{B}}$,

$$q(v) = \mathbf{x}^{\top}\mathbf{A}\mathbf{x} = \sum_{i} a_{ii}x_i^2 + 2\sum_{i<j} a_{ij}x_i x_j, \tag{14.3.1}$$

which is a homogeneous polynomial of degree 2 in the coordinates: hence the name. Reading Eq. 14.3.1 backwards recovers $\mathbf{A}$ from the polynomial: a_{ii} is the coefficient of x_i^2, and a_{ij} for $i \neq j$ is **half** the coefficient of $x_i x_j$. That is word for word the recipe Chapter 12 gave for the symmetric matrix of a real quadratic form, and the halving is the same division by $1 + 1$ as everywhere else in this section.

Example. From a polynomial to its matrix and back

Over $\mathbb{Q}$, let $q(x_1, x_2, x_3) = 2x_1^2 - x_2^2 + 6x_1x_2 - 4x_2x_3$. Find the symmetric matrix $\mathbf{A}$ with $q(\mathbf{x}) = \mathbf{x}^{\top}\mathbf{A}\mathbf{x}$, and evaluate $q(1, 1, 1)$ both from the polynomial and from $\mathbf{A}$.

Solution. The squared terms give $a_{11} = 2$, $a_{22} = -1$, and $a_{33} = 0$. The cross terms give $a_{12} = a_{21} = 3$ (half of 6), $a_{23} = a_{32} = -2$ (half of -4), and $a_{13} = a_{31} = 0$ since x_1x_3 does not appear. So

$$\mathbf{A} = \begin{pmatrix} 2 & 3 & 0 \\ 3 & -1 & -2 \\ 0 & -2 & 0 \end{pmatrix}.$$

From the polynomial, $q(1, 1, 1) = 2 - 1 + 6 - 4 = 3$. From the matrix, $\mathbf{A}(1, 1, 1) = (5, 0, -2)$ and $(1, 1, 1) \cdot (5, 0, -2) = 3$. The two agree.

Now the theorem that makes q as good as β.

Theorem 14.3.5. Polarization

Let $\operatorname{char} F \neq 2$. The map $\beta \mapsto q$, $q(v) = \beta(v, v)$, is a bijection from the set of symmetric bilinear forms on V onto the set of quadratic forms on V. Its inverse is given by either of

$$\begin{aligned}\beta(u, v) &= \tfrac{1}{2}(q(u + v) - q(u) - q(v)) \\ &= \tfrac{1}{4}(q(u + v) - q(u - v)).\end{aligned}$$

In particular a symmetric form is determined by its quadratic form: if $\beta(v, v) = \beta'(v, v)$ for all v, then $\beta = \beta'$.

Idea. Only one computation is needed, and it is the expansion of $q(u \pm v)$ by bilinearity. Symmetry merges the two cross terms into one, so the cross term appears with a factor $1 + 1$, and dividing by it is the whole trick. Surjectivity is the definition of "quadratic form"; injectivity is the displayed formula.

Proof. Let β be symmetric with quadratic form q, and let $u, v \in V$. By bilinearity and then symmetry,

$$q(u \pm v) = \beta(u, u) \pm \beta(u, v) \pm \beta(v, u) + \beta(v, v),$$

so $q(u \pm v) = q(u) + q(v) \pm (1 + 1)\beta(u, v)$. Since $\operatorname{char} F \neq 2$, the element $1 + 1$ is invertible, and taking the upper sign gives the first formula. Subtracting the two versions gives $q(u + v) - q(u - v) = (1 + 1)^2\beta(u, v)$, which is the second.

The map $\beta \mapsto q$ is surjective onto the set of quadratic forms by Definition 14.3.2, which defines that set as the image. It is injective because the first displayed formula recovers $\beta(u, v)$ from q alone; so if two symmetric forms have the same q, the formula returns the same value for each, and they are equal. Hence the map is a bijection with the stated inverse. This proves the theorem.

■

Quick check. Which bilinear forms are both symmetric and alternating, over a field of characteristic $\neq 2$?

Solution. Only the zero form. If β is alternating then $q = 0$, and if β is also symmetric then Theorem 14.3.5 recovers $\beta(u, v) = \frac{1}{2}(0 - 0 - 0) = 0$. (Over $\mathbb{F}_2$ the answer is different: every alternating form is symmetric there, as Theorem 14.3.3 shows, and there are plenty of nonzero ones.)

The second formula of Theorem 14.3.5 should look familiar. Over $\mathbb{R}$, an inner product is a symmetric bilinear form whose quadratic form is $q(v) = \|v\|^2$, and the formula reads

$$\langle u, v\rangle = \tfrac{1}{4}(\|u + v\|^2 - \|u - v\|^2),$$

which is exactly Lemma 11.7.2 (a). Chapter 11 proved it to show that an isometry automatically preserves angles; the content here is the same, stripped of positivity and of $\mathbb{R}$. What Chapter

11 needed positivity for was the *norm*, not the identity.

14.3.6 Hermitian forms

Over $\mathbb{C}$ the pattern changes, because $\langle v, v\rangle$ is required to be a positive real number and no bilinear form can deliver that: bilinearity gives $\beta(iv, iv) = i^2\beta(v, v) = -\beta(v, v)$, so the values at v and iv have opposite signs. Chapter 11 solved this by making the second slot conjugate-linear. That fix has nothing to do with positivity, so it can be made here too.

Definition 14.3.3. Hermitian Form

Let V be a vector space over $\mathbb{C}$. A function $\beta\colon V \times V \to \mathbb{C}$ is a **Hermitian form** if it is **linear in the first argument**, and

$$\beta(v, u) = \overline{\beta(u, v)} \qquad \text{for all } u, v \in V.$$

Two remarks on the clauses. First, the two conditions together force conjugate-linearity in the second argument: $\beta(u, cv) = \overline{\beta(cv, u)} = \overline{c\,\beta(v, u)} = \bar{c}\,\beta(u, v)$. So a Hermitian form is **not** bilinear, and Definition 14.1.1 does not cover it. Second, taking $u = v$ gives $\beta(v, v) = \overline{\beta(v, v)}$, so the associated function $q(v) = \beta(v, v)$ takes **real** values even though β does not.

The examples are the ones from Chapter 11 with positivity dropped: $\beta(\mathbf{x}, \mathbf{y}) = \mathbf{y}^*\mathbf{A}\mathbf{x}$ on $\mathbb{C}^n$ for any $\mathbf{A}$ with $\mathbf{A}^* = \mathbf{A}$, including indefinite choices such as $\mathbf{A} = \operatorname{diag}(1, -1)$. The matrix criterion is the same proof as Proposition 14.3.1 with stars in place of transposes.

Proposition 14.3.6. Hermitian Forms and Hermitian Matrices

Let β be linear in the first argument and conjugate-linear in the second, let $\mathcal{B} = (v_1, \dots, v_n)$ be a basis of V, and let $\mathbf{A}$ be the matrix with entries $a_{ij} = \beta(v_j, v_i)$. Then β is Hermitian if and only if $\mathbf{A}^* = \mathbf{A}$, and in that case $\beta(u, v) = [v]_{\mathcal{B}}^*\mathbf{A}[u]_{\mathcal{B}}$.

Proof. The index order $a_{ij} = \beta(v_j, v_i)$ is the one used for Gram matrices in Chapter 11 ($\mathbf{G}$ in Definition 11.2.3), chosen so that the displayed formula comes out with no transposing of the answer. Expanding $u = \sum_j c_j v_j$ and $v = \sum_i d_i v_i$ by linearity in the first slot and conjugate-linearity in the second,

$$\beta(u, v) = \sum_{i,j} c_j \overline{d_i}\,\beta(v_j, v_i) = \sum_{i,j} \overline{d_i}\, a_{ij} c_j,$$

which is the matrix product $[v]_{\mathcal{B}}^*\mathbf{A}[u]_{\mathcal{B}}$. If β is Hermitian then in particular $a_{ji} = \beta(v_i, v_j) = \overline{\beta(v_j, v_i)} = \overline{a_{ij}}$, which is $\mathbf{A}^* = \mathbf{A}$. Conversely, if $a_{ji} = \overline{a_{ij}}$ then conjugating the display gives

$$\overline{\beta(u, v)} = \sum_{i,j} d_i\, \overline{a_{ij}}\, \overline{c_j} = \sum_{i,j} \overline{c_j}\, a_{ji}\, d_i = \beta(v, u),$$

so β is Hermitian.

■

Polarization works for Hermitian forms too, and it is the four-term identity rather than the

two-term one: by Lemma 11.7.2 (b) with $q(v) = \beta(v, v)$,

$$\beta(u, v) = \tfrac{1}{4} \sum_{k=0}^{3} i^k \, q(u + i^k v).$$

The computation in Chapter 11 expanded each term using linearity in the first slot, conjugate-linearity in the second and $\beta(v, u) = \overline{\beta(u, v)}$, and used positivity nowhere, so it proves this identity too. Note the contrast with the real case: over $\mathbb{C}$ there is no characteristic to worry about, since $\operatorname{char} \mathbb{C} = 0$, and the reason a Hermitian form is recoverable from q is the extra room provided by i.

Warning. **A quadratic form over a field of characteristic 2 is a different object, and this book does not define it.** Two things go wrong there, and Theorem 14.3.5 repairs neither. *The form is not determined by q.* Over $\mathbb{F}_2$, the symmetric form $\beta(\mathbf{x}, \mathbf{y}) = x_1y_2 + x_2y_1$ has $q(\mathbf{x}) = 2x_1x_2 = 0$, the same q as the zero form, so $\beta \mapsto q$ is not injective on symmetric forms. *And Definition 14.3.2 names too few functions.* The function $q(\mathbf{x}) = x_1x_2$ on $\mathbb{F}_2^2$ is $\beta(\mathbf{x}, \mathbf{x})$ for the bilinear $\beta(\mathbf{x}, \mathbf{y}) = x_1y_2$, yet it is $\beta(\mathbf{x}, \mathbf{x})$ for **no symmetric** β: for a symmetric $\mathbf{A}$, Eq. 14.3.1 gives $a_{11}x_1^2 + a_{22}x_2^2 = a_{11}x_1 + a_{22}x_2$ over $\mathbb{F}_2$, whose values at $(1, 1)$, $(1, 0)$, and $(0, 1)$ are $a_{11} + a_{22}$, a_{11}, and a_{22}, and no choice makes those $1, 0, 0$. The theory of quadratic forms in characteristic 2 therefore defines q first, by the axioms $q(cv) = c^2q(v)$ and "$q(u+v) - q(u) - q(v)$ is bilinear", and treats β as derived data. Everything that follows assumes $\operatorname{char} F \neq 2$ precisely to stay out of that theory.

Three kinds of form, then, and three matrix shapes: symmetric, alternating, Hermitian. The next two sections take the first of them and classify it up to congruence, first over any field of characteristic $\neq 2$, then over $\mathbb{R}$.

14.3.7 Exercises

14.3.7.1 A. Check your understanding

Exercise. A1

(a) Define what it means for a bilinear form β on V to be **symmetric**, **alternating**, and **skew-symmetric**.

(b) State the relationship between alternating and skew-symmetric forms, including the exact role of the characteristic.

(c) Write down the matrix condition, in a fixed basis, for each of the three notions in (a).

(d) Determine whether the following statement is correct, and justify your answer: over $\mathbb{Q}$, a bilinear form is determined by the function $v \mapsto \beta(v, v)$.

(e) Why is a Hermitian form not an example of Definition 14.1.1?

Solution.

(a) Symmetric: $\beta(u, v) = \beta(v, u)$ for all u, v. Alternating: $\beta(v, v) = 0$ for all v. Skew-symmetric: $\beta(u, v) = -\beta(v, u)$ for all u, v.

(b) Alternating always implies skew-symmetric (Theorem 14.3.3 (a)). The converse holds if and only if $\operatorname{char} F \neq 2$; in characteristic 2 the dot product on $\mathbb{F}_2^2$ is skew-symmetric and not alternating.

(c) With $\mathbf{A} = [\beta]_{\mathcal{B}}^{\mathcal{B}}$: symmetric means $\mathbf{A}^\top = \mathbf{A}$; skew-symmetric means $\mathbf{A}^\top = -\mathbf{A}$; alternating means $\mathbf{A}^\top = -\mathbf{A}$ together with $a_{ii} = 0$ for all i (Proposition 14.3.1, Proposition 14.3.2).

(d) Incorrect. Since $\operatorname{char}\mathbb{Q} = 0 \neq 2$, only the **symmetric** part is determined: any alternating form has $\beta(v, v) = 0$ identically, so β and $\beta + \beta_a$ have the same quadratic form for every alternating β_a. What is true is Theorem 14.3.5, which determines β among *symmetric* forms.

(e) It is not linear in the second argument: $\beta(u, cv) = \overline{c}\,\beta(u, v)$, and $\overline{c} \neq c$ for $c \notin \mathbb{R}$.

14.3.7.2 B. Practice

Exercise. B1: Determine which of the following

Over $\mathbb{Q}$, with $V = \mathbb{Q}^2$ and $\mathbf{x} = (x_1, x_2)$, $\mathbf{y} = (y_1, y_2)$, determine for each of the following bilinear forms whether it is symmetric, whether it is alternating, and whether it is skew-symmetric. Justify each answer.

(a) $\beta(\mathbf{x}, \mathbf{y}) = x_1y_1 - x_2y_2$.

(b) $\beta(\mathbf{x}, \mathbf{y}) = 3x_1y_2 - 3x_2y_1$.

(c) $\beta(\mathbf{x}, \mathbf{y}) = x_1y_2 + 2x_2y_1$.

(d) $\beta(\mathbf{x}, \mathbf{y}) = 0$.

Solution. Write down the matrix $\mathbf{A}$ in the standard basis in each case and apply Proposition 14.3.1 and Proposition 14.3.2.

(a) $\mathbf{A} = \operatorname{diag}(1, -1)$, which satisfies $\mathbf{A}^\top = \mathbf{A}$: symmetric. Not skew-symmetric, since $\mathbf{A}^\top = -\mathbf{A}$ would need $1 = -1$ in $\mathbb{Q}$. Not alternating, since $\beta(\mathbf{e}_1, \mathbf{e}_1) = 1$.

(b) $\mathbf{A} = \left(\begin{smallmatrix} 0 & 3 \\ -3 & 0 \end{smallmatrix}\right)$, with $\mathbf{A}^\top = -\mathbf{A}$ and zero diagonal: skew-symmetric and alternating. Not symmetric, since $3 \neq -3$.

(c) $\mathbf{A} = \left(\begin{smallmatrix} 0 & 1 \\ 2 & 0 \end{smallmatrix}\right)$. Symmetry would need $1 = 2$ and skewness would need $1 = -2$, both false in $\mathbb{Q}$; so it is neither, and it is not alternating either, since an alternating form is skew-symmetric. (Its diagonal entries do vanish, so $\beta(\mathbf{e}_i, \mathbf{e}_i) = 0$ for the two basis vectors — but $\beta(\mathbf{1}, \mathbf{1}) = 1 + 2 = 3 \neq 0$. Checking the basis vectors alone is not enough.)

(d) The zero form is symmetric, skew-symmetric, and alternating at once; its matrix is $\mathbf{O}$, which satisfies every condition. By the Quick check above it is the only form that is both symmetric and alternating over $\mathbb{Q}$.

Exercise. B2: Polarizing a quadratic form

Over $\mathbb{R}$, let $q(x_1, x_2, x_3) = x_1^2 + 4x_1x_2 - 2x_2x_3 + 5x_3^2$.

(a) Find the symmetric matrix $\mathbf{A}$ with $q(\mathbf{x}) = \mathbf{x}^\top\mathbf{A}\mathbf{x}$.

(b) Compute $\beta(\mathbf{e}_1 + \mathbf{e}_2, \mathbf{e}_3)$ directly from $\mathbf{A}$, and again from the polarization formula $\frac{1}{2}(q(\mathbf{u} + \mathbf{v}) - q(\mathbf{u}) - q(\mathbf{v}))$.

Solution.

(a) The squared terms give $a_{11} = 1$, $a_{22} = 0$, $a_{33} = 5$; the cross terms give $a_{12} = a_{21} = 2$, $a_{23} = a_{32} = -1$, $a_{13} = a_{31} = 0$. So

$$\mathbf{A} = \begin{pmatrix} 1 & 2 & 0 \\ 2 & 0 & -1 \\ 0 & -1 & 5 \end{pmatrix}.$$

(b) Put $\mathbf{u} = (1, 1, 0)$ and $\mathbf{v} = (0, 0, 1)$. Directly, $\mathbf{Av} = (0, -1, 5)$ and $\mathbf{u}^\top(\mathbf{Av}) = 0 - 1 = -1$.

By polarization, $q(\mathbf{u}) = 1 + 4 = 5$ and $q(\mathbf{v}) = 5$, while $\mathbf{u} + \mathbf{v} = (1, 1, 1)$ gives $q(\mathbf{u} + \mathbf{v}) = 1 + 4 - 2 + 5 = 8$. Hence $\frac{1}{2}(8 - 5 - 5) = -1$. The two agree, as Theorem 14.3.5 says they must.

Exercise. B3: Splitting a form

Let $\mathbf{A} = \begin{pmatrix} 1 & 5 \\ -1 & 4 \end{pmatrix}$ over $\mathbb{Q}$, and let $\beta(\mathbf{x}, \mathbf{y}) = \mathbf{x}^\top \mathbf{A}\mathbf{y}$. Write $\beta = \beta_s + \beta_a$ as in Theorem 14.3.4, giving the matrix of each part, and verify that β and β_s have the same quadratic form.

Solution. We have $\mathbf{A}^\top = \left(\begin{smallmatrix} 1 & -1 \\ 5 & 4 \end{smallmatrix}\right)$, so

$$\tfrac{1}{2}(\mathbf{A} + \mathbf{A}^\top) = \begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix}, \qquad \tfrac{1}{2}(\mathbf{A} - \mathbf{A}^\top) = \begin{pmatrix} 0 & 3 \\ -3 & 0 \end{pmatrix}.$$

The first is symmetric and the second is alternating, and they sum to $\mathbf{A}$. For the quadratic forms, $\mathbf{x}^\top \mathbf{A}\mathbf{x} = x_1^2 + 5x_1x_2 - x_2x_1 + 4x_2^2 = x_1^2 + 4x_1x_2 + 4x_2^2$, and the symmetric part gives $x_1^2 + 2x_1x_2 + 2x_2x_1 + 4x_2^2$, the same polynomial. The alternating part contributes $3x_1x_2 - 3x_2x_1 = 0$.

14.3.7.3 C. Going deeper

Exercise. C1: Is this still true over $\mathbb{F}_2$?

Let V be a vector space over $\mathbb{F}_2$ with $\dim V = n \geq 1$.

(a) Prove that every alternating bilinear form on V is symmetric.

(b) Prove that the alternating forms on V form a subspace of the space of all bilinear forms, and compute its dimension.

(c) Deduce that for $n = 2$ there is exactly one nonzero alternating form on V, and write down its matrix.

Hint: for (b), use Proposition 14.3.2.

Solution.

(a) Let β be alternating. By Theorem 14.3.3 (a), $\beta(u, v) = -\beta(v, u)$. Over $\mathbb{F}_2$ we have

$-1 = 1$, so $-\beta(v,u) = \beta(v,u)$ and β is symmetric.

(b) Fix a basis $\mathcal{B}$. The map $\beta \mapsto [\beta]_{\mathcal{B}}^{\mathcal{B}}$ is a linear bijection from bilinear forms to $M_n(\mathbb{F}_2)$ (Theorem 14.1.1). By Proposition 14.3.2 (b), β is alternating exactly when its matrix satisfies $\mathbf{A}^\top = -\mathbf{A}$ with zero diagonal; over $\mathbb{F}_2$ the first condition is $\mathbf{A}^\top = \mathbf{A}$, so the image is the set of symmetric matrices with zero diagonal. Such a matrix is determined freely by its entries a_{ij} with $i < j$, of which there are $\binom{n}{2}$. These conditions are linear, so the set of alternating forms is a subspace, of dimension $n(n-1)/2$.

(c) For $n = 2$ the dimension is 1, so the space of alternating forms is $\{0, \beta\}$ for a single nonzero β, whose matrix must be $\left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$ — the only symmetric $\mathbb{F}_2$-matrix with zero diagonal other than $\mathbf{O}$. Concretely $\beta(\mathbf{x}, \mathbf{y}) = x_1 y_2 + x_2 y_1$, and indeed $\beta(\mathbf{x}, \mathbf{x}) = 2x_1 x_2 = 0$.

Exercise. C2: How much of a form does q see?

Let $\operatorname{char} F \neq 2$ and let β be any bilinear form on V, with $q(v) = \beta(v, v)$.

(a) Prove that two bilinear forms have the same quadratic form if and only if their difference is alternating.

(b) Deduce that the set of bilinear forms with a given quadratic form q contains exactly one symmetric form.

(c) For $\dim V = n$, how many bilinear forms share a given quadratic form when $F = \mathbb{F}_3$?

Solution.

(a) ($\Rightarrow$) Suppose $\beta_1(v,v) = \beta_2(v,v)$ for all v. Then $\delta = \beta_1 - \beta_2$ is bilinear with $\delta(v,v) = 0$ for all v, which is Definition 7.2.2 at $k = 2$.
($\Leftarrow$) If $\delta = \beta_1 - \beta_2$ is alternating then $\beta_1(v,v) - \beta_2(v,v) = \delta(v,v) = 0$ for every v.

(b) Existence: by Theorem 14.3.4 write $\beta = \beta_s + \beta_a$; then β_s is symmetric with the same quadratic form, since $\beta_a(v,v) = 0$. Uniqueness: two symmetric forms with the same q are equal by Theorem 14.3.5.

(c) By (a), the forms sharing q are exactly $\beta_s + \beta_a$ with β_a alternating, so there are as many as there are alternating forms. Over a field with 3 elements, an alternating form corresponds by Proposition 14.3.2 (b) to a matrix with $\mathbf{A}^\top = -\mathbf{A}$ and zero diagonal, which is determined freely by the $\binom{n}{2}$ entries above the diagonal. So there are $3^{n(n-1)/2}$ of them.

Exercise. C3: A Hermitian form is a real symmetric form in disguise

Let β be a Hermitian form on $\mathbb{C}^n$ with matrix $\mathbf{A} = \mathbf{A}^*$ as in Proposition 14.3.6, and regard $\mathbb{C}^n$ as a real vector space of dimension $2n$ by forgetting multiplication by i.

(a) Prove that $\tau(\mathbf{u}, \mathbf{v}) = \operatorname{Re} \beta(\mathbf{u}, \mathbf{v})$ is a symmetric $\mathbb{R}$-bilinear form on that real space, with $\tau(\mathbf{v}, \mathbf{v}) = \beta(\mathbf{v}, \mathbf{v})$.

(b) Prove that $\omega(\mathbf{u}, \mathbf{v}) = \operatorname{Im} \beta(\mathbf{u}, \mathbf{v})$ is an alternating $\mathbb{R}$-bilinear form.

(c) For $n = 1$ and $\mathbf{A} = (1)$, write both τ and ω as explicit forms on $\mathbb{R}^2$.

Solution.

(a) Taking real parts of an $\mathbb{R}$-linear identity is $\mathbb{R}$-linear, and β is additive in each slot and commutes with real scalars in each slot (conjugation fixes $\mathbb{R}$), so τ is $\mathbb{R}$-bilinear. Symmetry: $\beta(\mathbf{v},\mathbf{u}) = \overline{\beta(\mathbf{u},\mathbf{v})}$, and conjugate numbers have equal real parts, so $\tau(\mathbf{v},\mathbf{u}) = \tau(\mathbf{u},\mathbf{v})$. Finally $\beta(\mathbf{v},\mathbf{v})$ is real, as noted after Definition 14.3.3, so $\tau(\mathbf{v},\mathbf{v}) = \beta(\mathbf{v},\mathbf{v})$.

(b) $\mathbb{R}$-bilinearity is the same argument with imaginary parts. Conjugate numbers have opposite imaginary parts, so $\omega(\mathbf{v},\mathbf{u}) = -\omega(\mathbf{u},\mathbf{v})$: ω is skew-symmetric, and since $\operatorname{char}\mathbb{R} = 0 \neq 2$, Theorem 14.3.3 (b) makes it alternating. (Directly: $\beta(\mathbf{v},\mathbf{v})$ is real, so its imaginary part is 0.)

(c) Here $\beta(u,v) = u\overline{v}$ for $u, v \in \mathbb{C}$. Writing $u = x_1 + ix_2$ and $v = y_1 + iy_2$, we get $u\overline{v} = (x_1y_1 + x_2y_2) + i(x_2y_1 - x_1y_2)$. So τ is the dot product on $\mathbb{R}^2$ and $\omega(\mathbf{x},\mathbf{y}) = x_2y_1 - x_1y_2$, the negative of the signed area.

14.4 Completing the Square

Every symmetric matrix over $\mathbb{R}$ can be diagonalized by an orthogonal matrix, which is Chapter 12's spectral theorem. That theorem needs $\mathbb{R}$, an inner product, and the existence of eigenvalues. This section proves a weaker conclusion under far weaker hypotheses: over **any** field of characteristic $\neq 2$, with no inner product, no order, and no eigenvalues available, every symmetric form still becomes diagonal after a change of basis — provided we ask only for congruence, not for similarity. The proof is one algebraic manipulation, completing the square, run by induction; and the induction has a case that a careless proof loses.

Throughout, F is a field with $\operatorname{char} F \neq 2$ (Definition 0.7.2), V is a finite-dimensional vector space over F, β is a **symmetric** bilinear form on V (Definition 14.3.1), and $q(v) = \beta(v,v)$ is its quadratic form (Definition 14.3.2).

14.4.1 Orthogonality without an inner product

Chapter 11 called two vectors orthogonal when their inner product vanished. The word survives the loss of positivity unchanged, because only the vanishing was ever used.

Definition 14.4.1. Orthogonal Basis for a Form

Let β be a bilinear form on V. Vectors $u, v \in V$ are **orthogonal with respect to** β if $\beta(u,v) = 0$. An ordered basis $\mathcal{B} = (v_1, \ldots, v_n)$ of V is an **orthogonal basis for** β if

$$\beta(v_i, v_j) = 0 \qquad \text{whenever } i \neq j.$$

The definition is engineered to say exactly one thing about matrices: by Definition 14.1.2, $\mathcal{B}$ is an orthogonal basis for β precisely when $[\beta]_{\mathcal{B}}^{\mathcal{B}}$ is a **diagonal** matrix, with $q(v_1), \ldots, q(v_n)$ down the diagonal. Two warnings come free. The vectors of such a basis are not required to satisfy $q(v_i) = 1$, and over a general field they usually cannot be rescaled to. And a vector may be orthogonal to itself: $q(v_i) = 0$ is allowed, and for a degenerate form it is unavoidable.

What an orthogonal basis buys is that q becomes a sum of squares with no cross terms. If $v = c_1v_1 + \cdots + c_nv_n$ and $d_i = q(v_i)$, then expanding by bilinearity and deleting the $i \neq j$ terms,

$$q(v) = \sum_{i=1}^{n} d_i\, c_i^2. \tag{14.4.1}$$

That is the "sum of squares" a reader has met when completing the square in two variables, and Eq. 14.4.1 is the reason the whole section is called what it is.

14.4.2 The theorem, and the case that breaks it

Here is the naive argument, and it is worth seeing it fail. To make $q(\mathbf{v}) = ax_1^2 + 2bx_1x_2 + cx_2^2$ into a sum of squares, write

$$q = a\Big(x_1 + \frac{b}{a}x_2\Big)^2 + \Big(c - \frac{b^2}{a}\Big)x_2^2,$$

which is exactly a change of variable. The step divides by $a = q(\mathbf{e}_1)$. If $a = 0$ there is nothing to complete, and one may hope to start with x_2 instead — but $q(\mathbf{x}) = 2x_1x_2$ has $q(\mathbf{e}_1) = q(\mathbf{e}_2) = 0$, and neither variable offers a foothold. The repair is to change the vector, not the variable: $q(\mathbf{e}_1 + \mathbf{e}_2) = 2 \neq 0$. That repair is a lemma, and it is where $\operatorname{char} F \neq 2$ is spent.

Lemma 14.4.1. A Non-Zero Form Is Non-Zero Somewhere on the Diagonal

Let $\operatorname{char} F \neq 2$ and let β be a symmetric bilinear form on V with $\beta \neq 0$. Then there exists $v \in V$ with $q(v) \neq 0$.

Proof. Since $\beta \neq 0$, there are $u, w \in V$ with $a = \beta(u, w) \neq 0$. If $q(u) \neq 0$ or $q(w) \neq 0$ we are done. So suppose $q(u) = q(w) = 0$. Expanding by bilinearity and using symmetry,

$$q(u + w) = q(u) + 2\beta(u, w) + q(w) = (1 + 1)a.$$

Since $\operatorname{char} F \neq 2$, the element $1 + 1$ is nonzero, and a product of two nonzero elements of a field is nonzero, so $q(u + w) \neq 0$. This proves the lemma. ∎

Contrapositively: if q vanishes identically then $\beta = 0$. That is the case the induction must handle separately, and the lemma is what lets it be handled in one line.

Theorem 14.4.2. Diagonalization of a Symmetric Form

Let F be a field with $\operatorname{char} F \neq 2$, let V be a finite-dimensional vector space over F, and let β be a symmetric bilinear form on V. Then V has an orthogonal basis for β.

Equivalently: every symmetric $\mathbf{A} \in M_n(F)$ is congruent to a diagonal matrix, that is, there is an invertible $\mathbf{P} \in M_n(F)$ with $\mathbf{P}^{\top}\mathbf{A}\mathbf{P}$ diagonal.

Idea. Induction on $\dim V$. The step wants to peel off one basis vector and recurse on a complement, exactly as "peel off an eigenvector" does in Chapter 9 — but the complement here is the set of vectors orthogonal to the chosen one, not an invariant subspace, and it is available only if the chosen v_1 has $q(v_1) \neq 0$. So the proof splits into two steps: ① if q vanishes everywhere, the form is zero and **every** basis is orthogonal; ② otherwise

Lemma 14.4.1 supplies a v_1 with $q(v_1) \neq 0$, the linear functional $w \mapsto \beta(v_1, w)$ is nonzero, its kernel W is a hyperplane meeting $\operatorname{span}(v_1)$ trivially, and the induction hypothesis applies to β restricted to W. Step ① is not a formality: without it the argument has no base for the recursion when the form degenerates, and the word "equivalently" in the statement would be unearned.

Proof. We induct on $n = \dim V$. If $n = 0$ the empty basis is orthogonal, vacuously; if $n = 1$ any basis (v_1) is orthogonal, since there is no pair $i \neq j$. Let $n \geq 2$ and suppose the result holds for all spaces of dimension $n - 1$ over F.

Step 1: the form vanishes on the diagonal. Suppose $q(v) = 0$ for every $v \in V$. By Lemma 14.4.1, read contrapositively, $\beta = 0$. Then $\beta(v_i, v_j) = 0$ for every pair of vectors whatsoever, so **any** basis of V is an orthogonal basis for β, and its matrix is $\mathbf{O}$, which is diagonal.

Step 2: the form does not vanish on the diagonal. Otherwise there is $v_1 \in V$ with $c = q(v_1) \neq 0$; in particular $v_1 \neq 0$. Define

$$\varphi \colon V \to F, \qquad \varphi(w) = \beta(v_1, w).$$

Then φ is linear, because β is linear in its second argument, and $\varphi(v_1) = c \neq 0$, so $\varphi \neq 0$. A nonzero functional is surjective onto F (Proposition 5.1.2), so $\operatorname{rank}\varphi = 1$ and, by the Rank–Nullity Theorem (Theorem 2.3.1), $W = \ker\varphi$ has $\dim W = n - 1$.

Moreover $\operatorname{span}(v_1) \cap W = \{0\}$: if $av_1 \in W$ then $0 = \varphi(av_1) = ac$, and $c \neq 0$ forces $a = 0$. So the sum $\operatorname{span}(v_1) + W$ is direct, and $\dim(\operatorname{span}(v_1) + W) = 1 + (n - 1) = n$, both by Theorem 1.7.4. A subspace of V of dimension $\dim V$ is all of V (Theorem 1.6.8), so

$$V = \operatorname{span}(v_1) \oplus W.$$

The restriction of β to $W \times W$ is a bilinear form on W, and it is symmetric because β is. As $\dim W = n - 1$, the induction hypothesis provides an orthogonal basis $(v_2, \dots, v_n)$ of W for that restriction.

Concatenating bases of the summands of a direct sum gives a basis of it (Theorem 1.7.5, (a) $\Leftrightarrow$ (d)), so $\mathcal{B} = (v_1, v_2, \dots, v_n)$ is a basis of V. It is orthogonal for β: for $j \geq 2$ we have $v_j \in W = \ker\varphi$, so $\beta(v_1, v_j) = \varphi(v_j) = 0$, and $\beta(v_j, v_1) = 0$ by symmetry; and for $2 \leq i \neq j$ the value $\beta(v_i, v_j)$ is 0 by the choice of the basis of W. This completes the induction.

For the matrix statement, let $\mathbf{A} \in M_n(F)$ be symmetric and let $\beta(\mathbf{x}, \mathbf{y}) = \mathbf{x}^{\top}\mathbf{A}\mathbf{y}$ on $V = F^n$, which is symmetric by Proposition 14.3.1 and has $[\beta]_{\mathcal{E}}^{\mathcal{E}} = \mathbf{A}$ in the standard basis $\mathcal{E}$. Take an orthogonal basis $\mathcal{B}$ as above and let $\mathbf{P} = [\mathrm{id}]_{\mathcal{B}}^{\mathcal{E}}$, the matrix whose columns are the vectors of $\mathcal{B}$, which is invertible. By Theorem 14.2.1, $[\beta]_{\mathcal{B}}^{\mathcal{B}} = \mathbf{P}^{\top}\mathbf{A}\mathbf{P}$, and this matrix is diagonal. This proves the theorem.

■

Two hypotheses did visible work. Symmetry was used twice: to make $\beta(v_j, v_1)$ vanish along with $\beta(v_1, v_j)$, and to keep the restriction to W symmetric so the induction could run. The characteristic hypothesis was used exactly once, inside Lemma 14.4.1 — and that single use is essential: Exercise C2 below exhibits a symmetric matrix over $\mathbb{F}_2$ that is congruent to no

diagonal matrix at all.

Warning. **Diagonal does not mean "the diagonal entries are eigenvalues".** The matrix $\mathbf{P}^\top\mathbf{AP}$ of Theorem 14.4.2 is congruent to $\mathbf{A}$, not similar to it, and its entries are the values $q(v_i)$ at whatever basis the algorithm happened to produce. Over $\mathbb{R}$ they can be scaled at will: replacing v_i by tv_i replaces d_i by $t^2 d_i$. There is no canonical list of numbers here. The next section finds what there *is*.

Quick check. In Theorem 14.4.2, how many of the diagonal entries d_i are zero?

Solution. Exactly $n - \operatorname{rank}\mathbf{A}$ of them. A diagonal matrix has rank equal to its number of nonzero entries, and congruence preserves rank (Theorem 14.2.3), so the count is the same for every orthogonal basis even though the individual d_i are not. In the language of Section 1, those v_i with $d_i = 0$ span the radical $\operatorname{rad}(\beta)$ of Definition 14.1.3, since in an orthogonal basis such a v_i pairs to zero with every basis vector; and β is nondegenerate (Definition 14.1.4) exactly when no d_i is zero.

14.4.3 The algorithm: symmetric row and column operations

The proof is constructive, but running it by hand means chasing kernels of functionals. There is a bookkeeping device that does the same work on the matrix directly, and it is the practical content of this section.

The key observation is that a congruence by an elementary matrix is a **pair** of operations. Let $\mathbf{E}$ be an elementary matrix (Definition 3.4.1). Right multiplication $\mathbf{AE}$ performs a column operation, and left multiplication $\mathbf{E}^\top\mathbf{A}$ performs the corresponding row operation. So $\mathbf{A} \mapsto \mathbf{E}^\top\mathbf{AE}$ performs both, and the result is congruent to $\mathbf{A}$ and again symmetric. Doing the operations in pairs is exactly what keeps the matrix symmetric, and that is the whole idea.

Algorithm. To diagonalize a symmetric $\mathbf{A} \in M_n(F)$ by congruence, write $\mathbf{A}$ and $\mathbf{I}_n$ side by side and repeat, for $i = 1, 2, \dots, n$ in turn:

- **If** $a_{ii} = 0$ **but some** $a_{jj} \neq 0$ **with** $j > i$, swap rows i, j and then columns i, j.
- **If** $a_{ii} = 0$ **and every** $a_{jj} = 0$ **for** $j \geq i$, look for an entry $a_{ij} \neq 0$ with $j > i$. If there is none, then row i and column i are already zero in every position, so move on to $i + 1$. If there is one, add row j to row i, then column j to column i; the new a_{ii} is $2a_{ij} \neq 0$.
- **Now** $a_{ii} \neq 0$**.** For each $j > i$ with $a_{ij} \neq 0$, subtract (a_{ij}/a_{ii}) times row i from row j, then (a_{ij}/a_{ii}) times column i from column j.

Over a field where division is inconvenient — $\mathbb{Q}$ or $\mathbb{Z}$-valued data by hand — the clearing step may instead be done as row $j \mapsto a_{ii}$ row $j - a_{ij}$ row i, with the matching column operation. That is the same move composed with scaling row and column j by $a_{ii} \neq 0$, which is itself a congruence, so it is legitimate; it keeps the entries integral and changes the diagonal result only by square factors. The worked example below takes this route.

Perform only the **column** half of each operation on the matrix standing to the right. When

$\mathbf{A}$ has become diagonal, that matrix is $\mathbf{P}$, and $\mathbf{P}^{\top}\mathbf{A}\mathbf{P}$ is the diagonal result.

The second bullet is Lemma 14.4.1 made mechanical, and the factor $2 = 1 + 1$ in it is where the algorithm would stop in characteristic 2. The right-hand matrix works because applying the column operations $\mathbf{E}_1, \mathbf{E}_2, \dots, \mathbf{E}_k$ in order turns $\mathbf{I}$ into $\mathbf{E}_1\mathbf{E}_2 \cdots \mathbf{E}_k$, which is the product whose transpose and self sandwich $\mathbf{A}$.

Example. Diagonalizing a symmetric matrix with zero diagonal

Over $\mathbb{Q}$, diagonalize

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}$$

by congruence, producing an invertible $\mathbf{P}$ with $\mathbf{P}^{\top}\mathbf{A}\mathbf{P}$ diagonal.

Solution. The form is $q(\mathbf{x}) = 2x_1x_2 + 2x_1x_3 + 2x_2x_3$: every diagonal entry is 0, so there is no square to complete yet, and the second bullet applies.

Step 1 (create a pivot). Add row 2 to row 1, then column 2 to column 1. The matrix becomes

$$\begin{pmatrix} 1 & 1 & 2 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix} \longrightarrow \begin{pmatrix} 2 & 1 & 2 \\ 1 & 0 & 1 \\ 2 & 1 & 0 \end{pmatrix},$$

the left matrix after the row operation and the right one after the matching column operation. The new corner entry is $q(\mathbf{e}_1 + \mathbf{e}_2) = 2$.

Step 2 (clear the first row and column, in two paired moves). Rather than divide by 2, use the operation "row $2 \mapsto 2 \cdot$ row $2 -$ row 1", which is a scaling followed by an addition, hence a congruence by a product of two elementary matrices, and legitimate because $2 \neq 0$ in $\mathbb{Q}$. Apply it and its column partner. This keeps every entry an integer:

$$\begin{pmatrix} 2 & 1 & 2 \\ 0 & -1 & 0 \\ 2 & 1 & 0 \end{pmatrix} \longrightarrow \begin{pmatrix} 2 & 0 & 2 \\ 0 & -2 & 0 \\ 2 & 0 & 0 \end{pmatrix}.$$

Then "row $3 \mapsto$ row $3 -$ row 1" and its column partner:

$$\begin{pmatrix} 2 & 0 & 2 \\ 0 & -2 & 0 \\ 0 & 0 & -2 \end{pmatrix} \longrightarrow \begin{pmatrix} 2 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & -2 \end{pmatrix}.$$

The transforming matrix. Applying the three **column** operations to $\mathbf{I}_3$ in the same order gives

$$\mathbf{I}_3 \to \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \to \begin{pmatrix} 1 & -1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \to \begin{pmatrix} 1 & -1 & -1 \\ 1 & 1 & -1 \\ 0 & 0 & 1 \end{pmatrix} = \mathbf{P}.$$

Here $\det \mathbf{P} = 2 \neq 0$, so $\mathbf{P}$ is invertible, and

$$\mathbf{P}^\top \mathbf{A} \mathbf{P} = \operatorname{diag}(2, -2, -2).$$

The check is one column at a time: with $\mathbf{p}_1 = (1, 1, 0)$, $\mathbf{p}_2 = (-1, 1, 0)$, and $\mathbf{p}_3 = (-1, -1, 1)$, we get $q(\mathbf{p}_1) = 2$, $q(\mathbf{p}_2) = -2$, and $q(\mathbf{p}_3) = 2 - 2 - 2 = -2$, while $\beta(\mathbf{p}_1, \mathbf{p}_2) = \mathbf{p}_1^\top \mathbf{A} \mathbf{p}_2 = (1, 1, 0) \cdot (1, -1, 0) = 0$ and similarly for the other two pairs.

Two observations about that example. The determinants are $\det \mathbf{A} = 2$ and $\det(\mathbf{P}^\top \mathbf{A} \mathbf{P}) = 8$, and they differ by $(\det \mathbf{P})^2 = 4$: a congruence multiplies the determinant by a square, which is the discriminant statement of Section 2. And the eigenvalues of $\mathbf{A}$ are $2, -1, -1$, whose signs are $+, -, -$ — the same pattern of signs as $2, -2, -2$, with none of the same numbers. That coincidence is Section 5.

14.4.4 Over the complex numbers, and over a general field

The diagonal entries produced by the algorithm are not canonical, but there is a precise statement of how much freedom remains.

Proposition 14.4.3. Diagonal Entries Are Defined Only up to Squares

Let $\mathcal{B} = (v_1, \dots, v_n)$ be an orthogonal basis for β with $d_i = q(v_i)$, and let $t_1, \dots, t_n \in F$ be nonzero. Then $(t_1 v_1, \dots, t_n v_n)$ is again an orthogonal basis for β, with diagonal entries $t_i^2 d_i$.

Proof. Scaling basis vectors by nonzero scalars again gives a basis. For $i \neq j$, bilinearity gives $\beta(t_i v_i, t_j v_j) = t_i t_j \beta(v_i, v_j) = 0$, so the new basis is orthogonal; and $q(t_i v_i) = t_i^2 q(v_i) = t_i^2 d_i$. ∎

So each nonzero d_i can be changed to any other element of its coset in the group $F^\times / (F^\times)^2$ of **square classes**, and no further, by this move. How much that costs depends entirely on the arithmetic of F:

- Over $\mathbb{C}$, every element is a square, so there is exactly one nonzero square class and every nonzero d_i can be turned into 1.
- Over $\mathbb{R}$, the squares are the positive numbers, so there are two classes and every nonzero d_i can be turned into 1 or -1, according to its sign — the subject of the next section.
- Over $\mathbb{Q}$ there are infinitely many classes ($2, 3, 5, \dots$ are pairwise inequivalent), so the 1-dimensional forms $q(x) = x^2$ and $q(x) = 2x^2$ are already noncongruent.
- Over $\mathbb{F}_p$ with p odd there are exactly two classes, as for $\mathbb{R}$, but "which class" is not a matter of sign.

The complex case collapses completely.

Corollary 14.4.4. Classification of Complex Symmetric Matrices by Rank

Let $\mathbf{A} \in M_n(\mathbb{C})$ be symmetric of rank r. Then $\mathbf{A}$ is congruent to

$$\mathbf{I}_r \oplus \mathbf{O}_{n-r} = \operatorname{diag}(\underbrace{1, \dots, 1}_{r}, 0, \dots, 0).$$

Consequently two symmetric matrices in $M_n(\mathbb{C})$ are congruent if and only if they have the same rank, and there are exactly $n+1$ congruence classes.

Proof. By Theorem 14.4.2 (with $\operatorname{char}\mathbb{C} = 0$) there is an invertible $\mathbf{P}_1$ with $\mathbf{P}_1^{\top}\mathbf{A}\mathbf{P}_1 = \mathbf{D} = \operatorname{diag}(d_1, \dots, d_n)$ diagonal. Congruence preserves rank (Theorem 14.2.3), and the rank of a diagonal matrix is its number of nonzero entries, so exactly r of the d_i are nonzero. Permuting the basis vectors is a congruence by a permutation matrix $\mathbf{P}_2$, so we may assume $d_1, \dots, d_r \neq 0$ and $d_{r+1} = \cdots = d_n = 0$.

Every nonzero complex number has a square root, so choose $s_i \in \mathbb{C}$ with $s_i^2 = d_i$ for $i \leq r$, and set $t_i = 1/s_i$ for $i \leq r$ and $t_i = 1$ otherwise. By Proposition 14.4.3 the congruence by $\mathbf{P}_3 = \operatorname{diag}(t_1, \dots, t_n)$ replaces d_i by $t_i^2 d_i$, which is 1 for $i \leq r$ and 0 otherwise. So $\mathbf{P}^{\top}\mathbf{A}\mathbf{P} = \mathbf{I}_r \oplus \mathbf{O}_{n-r}$ with $\mathbf{P} = \mathbf{P}_1\mathbf{P}_2\mathbf{P}_3$.

For the classification: congruent matrices have equal rank by Theorem 14.2.3, and matrices of equal rank are both congruent to the same normal form, hence to each other by Proposition 14.2.2. The possible ranks are $0, 1, \dots, n$, giving $n+1$ classes. This proves the corollary. ∎

Contrast this with similarity, where complex symmetric matrices are not classified by anything so simple: the matrix $\left(\begin{smallmatrix} 1 & i \\ i & -1 \end{smallmatrix}\right)$ is symmetric, nonzero, and satisfies $\mathbf{A}^2 = \mathbf{O}$, so it is not even diagonalizable as an operator. Congruence is a coarser relation than similarity on symmetric matrices, and over $\mathbb{C}$ it is coarse enough to be finite.

14.4.5 Exercises

14.4.5.1 A. Check your understanding

Exercise. A1

(a) Define what it means for an ordered basis to be an **orthogonal basis for a bilinear form** β, and say what that means about $[\beta]_{\mathcal{B}}^{\mathcal{B}}$.

(b) State Theorem 14.4.2 with all of its hypotheses.

(c) Which step of the proof uses $\operatorname{char} F \neq 2$, and what exactly is divided by?

(d) Determine whether the following statement is correct, and justify your answer: if $\mathbf{P}^{\top}\mathbf{A}\mathbf{P} = \operatorname{diag}(d_1, \dots, d_n)$, then the d_i are the eigenvalues of $\mathbf{A}$.

(e) Describe the pair of operations that a congruence by an elementary matrix performs.

Solution.

(a) $\mathcal{B} = (v_1, \dots, v_n)$ is orthogonal for β if $\beta(v_i, v_j) = 0$ whenever $i \neq j$. Equivalently $[\beta]_{\mathcal{B}}^{\mathcal{B}}$ is diagonal, with entries $q(v_i)$.

(b) If $\operatorname{char} F \neq 2$, V is finite-dimensional over F and β is a symmetric bilinear form on V, then V has an orthogonal basis for β; equivalently every symmetric $\mathbf{A} \in M_n(F)$ is congruent to a diagonal matrix.

(c) Lemma 14.4.1, inside Step 2 of the proof. It produces $q(u+w) = (1+1)\beta(u,w)$ and needs $1+1 \neq 0$ to conclude that this is nonzero. Equivalently, the algorithm's repair move multiplies by 2, so it needs $2 \neq 0$.

(d) Incorrect. The relation is congruence, not similarity. In Example (Diagonalizing a symmetric matrix with zero diagonal) the diagonal entries are $2, -2, -2$ while the eigenvalues are $2, -1, -1$. Only the signs matched there, and Section 5 explains why.

(e) A column operation and the matching row operation: $\mathbf{A} \mapsto \mathbf{E}^{\top}\mathbf{A}\mathbf{E}$. Performing both keeps the matrix symmetric and keeps it in the same congruence class.

14.4.5.2 B. Practice

Exercise. B1: Symmetric elimination

Over $\mathbb{Q}$, let

$$\mathbf{A} = \begin{pmatrix} 1 & 2 & 0 \\ 2 & 1 & 3 \\ 0 & 3 & 1 \end{pmatrix}.$$

Diagonalize $\mathbf{A}$ by paired row and column operations, exhibiting $\mathbf{P}$ and $\mathbf{P}^{\top}\mathbf{A}\mathbf{P}$.

Solution. Here $a_{11} = 1 \neq 0$, so the pivot is already available.

Clear the first row and column. Subtract $2\times$ row 1 from row 2, then $2\times$ column 1 from column 2. The entry a_{13} is already 0, so nothing else is needed in this round:

$$\begin{pmatrix} 1 & 2 & 0 \\ 0 & -3 & 3 \\ 0 & 3 & 1 \end{pmatrix} \longrightarrow \begin{pmatrix} 1 & 0 & 0 \\ 0 & -3 & 3 \\ 0 & 3 & 1 \end{pmatrix}.$$

Clear the second row and column. The pivot is -3, and $a_{23} = 3$, so add row 2 to row 3 and then column 2 to column 3:

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & -3 & 3 \\ 0 & 0 & 4 \end{pmatrix} \longrightarrow \begin{pmatrix} 1 & 0 & 0 \\ 0 & -3 & 0 \\ 0 & 0 & 4 \end{pmatrix}.$$

The transforming matrix. Applying the two column operations to $\mathbf{I}_3$ gives

$$\mathbf{P} = \begin{pmatrix} 1 & -2 & -2 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}, \qquad \mathbf{P}^{\top}\mathbf{A}\mathbf{P} = \operatorname{diag}(1, -3, 4).$$

As a check, $\det \mathbf{P} = 1$ and $\det \mathbf{A} = -12 = \det \operatorname{diag}(1, -3, 4)$.

Exercise. B2: No pivot on the diagonal

Over $\mathbb{Q}$, let $\mathbf{A} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ and $q(\mathbf{x}) = 2x_1x_2$.

(a) Find a vector $\mathbf{v}$ with $q(\mathbf{v}) \neq 0$, as in Lemma 14.4.1.

(b) Diagonalize $\mathbf{A}$ by paired operations and exhibit $\mathbf{P}$.

(c) Write q as a difference of two squares in the new coordinates.

Solution.

(a) Both $\mathbf{e}_1$ and $\mathbf{e}_2$ give $q = 0$, so take $\mathbf{v} = \mathbf{e}_1 + \mathbf{e}_2$: $q(\mathbf{v}) = 2 \neq 0$.

(b) Add row 2 to row 1, then column 2 to column 1, reaching $\left(\begin{smallmatrix} 2 & 1 \\ 1 & 0 \end{smallmatrix}\right)$. Then apply "row $2 \mapsto 2 \cdot$ row $2 -$ row 1" and its column partner, reaching $\operatorname{diag}(2, -2)$. Applying the same two column operations to $\mathbf{I}_2$ gives

$$\mathbf{P} = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}, \qquad \mathbf{P}^{\top}\mathbf{A}\mathbf{P} = \operatorname{diag}(2, -2).$$

Check directly: $\mathbf{P}^{\top}\mathbf{A} = \left(\begin{smallmatrix} 1 & 1 \\ 1 & -1 \end{smallmatrix}\right)$ and multiplying by $\mathbf{P}$ gives $\left(\begin{smallmatrix} 2 & 0 \\ 0 & -2 \end{smallmatrix}\right)$.

(c) With $\mathbf{x} = \mathbf{P}\mathbf{y}$, that is $x_1 = y_1 - y_2$ and $x_2 = y_1 + y_2$, we get $q = 2(y_1 - y_2)(y_1 + y_2) = 2y_1^2 - 2y_2^2$. This is the familiar identity $ab = \left(\frac{a+b}{2}\right)^2 - \left(\frac{a-b}{2}\right)^2$ read as a change of variable, and it needs 2 to be invertible.

Exercise. B3: Reading off the rank

Let $\mathbf{A} \in M_4(\mathbb{C})$ be symmetric with $\operatorname{rank}\mathbf{A} = 3$. Write down the normal form of $\mathbf{A}$ under congruence, and say how many congruence classes of symmetric matrices there are in $M_4(\mathbb{C})$.

Solution. By Corollary 14.4.4, $\mathbf{A}$ is congruent to $\mathbf{I}_3 \oplus \mathbf{O}_1 = \operatorname{diag}(1, 1, 1, 0)$. Since rank is a complete invariant there, the classes correspond to the possible ranks $0, 1, 2, 3, 4$, so there are 5 of them.

14.4.5.3 C. Going deeper

Exercise. C1: Square classes over a finite field

Work over $\mathbb{F}_5$, whose nonzero squares are 1 and 4.

(a) Prove that the 1×1 matrices (1) and (2) are not congruent over $\mathbb{F}_5$.

(b) Prove that $\mathbf{I}_2$ and $2\mathbf{I}_2$ **are** congruent over $\mathbb{F}_5$, by exhibiting a $\mathbf{P}$.

(c) Explain why (a) and (b) are consistent with Proposition 14.4.3.

Hint: for (b), look for two vectors $\mathbf{p}_1, \mathbf{p}_2$ with $\mathbf{p}_i \cdot \mathbf{p}_i = 2$ and $\mathbf{p}_1 \cdot \mathbf{p}_2 = 0$.

Solution.

(a) A congruence of 1×1 matrices is $(a)\mapsto(p^2a)$ with $p\neq 0$. So (1) and (2) are congruent exactly when $2=p^2$ for some $p\in\mathbb{F}_5^\times$. The squares of $1,2,3,4$ are $1,4,4,1$, so 2 is not a square and the two are not congruent.

(b) Take $\mathbf{p}_1=(1,1)$ and $\mathbf{p}_2=(1,4)$. Then $\mathbf{p}_1\cdot\mathbf{p}_1=2$, $\mathbf{p}_2\cdot\mathbf{p}_2=1+16=17=2$, and $\mathbf{p}_1\cdot\mathbf{p}_2=1+4=5=0$ in $\mathbb{F}_5$. With

$$\mathbf{P}=\begin{pmatrix}1&1\\1&4\end{pmatrix},\qquad \det\mathbf{P}=4-1=3\neq 0,$$

we get $\mathbf{P}^\top\mathbf{I}_2\mathbf{P}=\mathbf{P}^\top\mathbf{P}=\operatorname{diag}(2,2)=2\mathbf{I}_2$.

(c) Proposition 14.4.3 only says what **rescaling a single basis vector** can do, and that move cannot change a square class. Part (b) does not contradict it, because the congruence there is not diagonal: it mixes the two coordinates, and both diagonal entries change square class at once. What is invariant is the class of their product, the **determinant**: $\det\mathbf{I}_2=1$ and $\det(2\mathbf{I}_2)=4$ differ by $4=2^2$, a square.

Exercise. C2: Is this still true over $\mathbb{F}_2$?

Let $\mathbf{A}=\begin{pmatrix}0&1\\1&0\end{pmatrix}$ over $\mathbb{F}_2$, and let $\beta(\mathbf{x},\mathbf{y})=\mathbf{x}^\top\mathbf{A}\mathbf{y}$ on $\mathbb{F}_2^2$.

(a) Prove that β is alternating.

(b) Prove that if $\mathbf{P}\in M_2(\mathbb{F}_2)$ is invertible, then every diagonal entry of $\mathbf{P}^\top\mathbf{A}\mathbf{P}$ is 0.

(c) Deduce that $\mathbf{A}$ is congruent to **no** diagonal matrix except the zero matrix, and hence that Theorem 14.4.2 fails over $\mathbb{F}_2$.

Solution.

(a) By Proposition 14.3.2 (b): over $\mathbb{F}_2$ we have $-\mathbf{A}=\mathbf{A}$, so $\mathbf{A}^\top=\mathbf{A}=-\mathbf{A}$, and the diagonal entries of $\mathbf{A}$ are 0. (Directly: $\beta(\mathbf{x},\mathbf{x})=2x_1x_2=0$.)

(b) The i-th diagonal entry of $\mathbf{P}^\top\mathbf{A}\mathbf{P}$ is $\mathbf{p}_i^\top\mathbf{A}\mathbf{p}_i=q(\mathbf{p}_i)$, where $\mathbf{p}_i$ is the i-th column of $\mathbf{P}$. By (a), q vanishes identically, so every diagonal entry is 0.

(c) A diagonal matrix congruent to $\mathbf{A}$ has all diagonal entries 0 by (b), hence is the zero matrix. But congruence preserves rank (Theorem 14.2.3) and $\operatorname{rank}\mathbf{A}=2\neq 0=\operatorname{rank}\mathbf{O}$. So no diagonal matrix is congruent to $\mathbf{A}$. Since $\mathbf{A}$ is symmetric over $\mathbb{F}_2$, the conclusion of Theorem 14.4.2 fails there, and the hypothesis $\operatorname{char}F\neq 2$ cannot be dropped.

Exercise. C3: Diagonalizing the restriction

Let $\operatorname{char}F\neq 2$, let β be a symmetric bilinear form on V, and let $U\subseteq V$ be a subspace.

(a) Prove that U has an orthogonal basis for the restriction of β to U.

(b) Give an example, with $\dim V=2$, in which β is nondegenerate but its restriction to some 1-dimensional U is the zero form.

(c) Deduce that an orthogonal basis of U cannot always be extended to an orthogonal basis of V by adding vectors of U alone, and say in one sentence what Theorem 14.4.2 does guarantee instead.

Solution.

(a) The restriction $\beta|_{U\times U}$ is bilinear, since it is the same function on a smaller domain, and symmetric for the same reason. U is finite-dimensional, being a subspace of a finite-dimensional space (Theorem 1.6.7). So Theorem 14.4.2 applies to $(U, \beta|_{U\times U})$ directly.

(b) Over $\mathbb{Q}$ take $\beta(\mathbf{x}, \mathbf{y}) = x_1y_2 + x_2y_1$ on $\mathbb{Q}^2$, with matrix $\left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$, which is invertible, so β is nondegenerate. Put $U = \operatorname{span}(\mathbf{e}_1)$. Then $\beta(a\mathbf{e}_1, b\mathbf{e}_1) = 0$ for all a, b, so the restriction is the zero form even though β is not.

(c) In (b) the single vector $\mathbf{e}_1$ is an orthogonal basis of U, but no vector of U can be added to it, and any orthogonal basis of V for this β must consist of vectors $\mathbf{v}$ with $q(\mathbf{v}) \neq 0$ (its matrix is diagonal and invertible), so $\mathbf{e}_1$ belongs to none of them. What the theorem guarantees is only the existence of *some* orthogonal basis of V, built from scratch; it promises nothing about extending a given orthogonal list, and Section 8 is where a correct extension statement finally appears.

14.5 Sylvester's Law of Inertia

The previous section produced a diagonal matrix and then took it away again: the diagonal entries are not invariants, because rescaling the i-th basis vector by t multiplies the i-th entry by t^2. Over $\mathbb{R}$ the squares are exactly the positive numbers, so rescaling can change the size of an entry but never its sign. This section proves that the signs are all that survives, and that they survive completely: the number of positive, negative, and zero entries is the same for every diagonalization. That is the classification of real symmetric matrices up to congruence, and it is also the missing link between congruence and the spectral theorem.

Throughout this section the field is $\mathbb{R}$. Every form is a symmetric bilinear form on a finite-dimensional real vector space V, and every matrix is a real symmetric matrix. The words "positive" and "negative" have no meaning over a general field, and almost nothing in this section survives the move to one; Corollary 14.4.4 already showed that over $\mathbb{C}$ the answer is different, since rank alone classifies there.

14.5.1 Normalizing the diagonal

Start from Theorem 14.4.2, which applies because $\operatorname{char}\mathbb{R} = 0$. It gives an orthogonal basis $(v_1, \dots, v_n)$ for β, with $q(v_i) = d_i$. Over $\mathbb{R}$ we may go one step further and scale.

By Proposition 14.4.3, replacing v_i by t_iv_i replaces d_i by $t_i^2d_i$. Choosing $t_i = 1/\sqrt{|d_i|}$ when $d_i \neq 0$, and $t_i = 1$ when $d_i = 0$, turns every diagonal entry into 1, -1, or 0. A permutation of the basis then sorts them, so after a further congruence by a permutation matrix the matrix of β is

$$\mathbf{I}_{n_+} \oplus (-\mathbf{I}_{n_-}) \oplus \mathbf{O}_{n_0}, \tag{14.5.1}$$

where n_+ counts the positive d_i, n_- the negative ones and n_0 the zeros, and $n_+ + n_- + n_0 = n$. Nothing yet says that these three counts do not depend on the orthogonal basis we started from. That is the theorem.

Quick check. One of the three counts is already known to be an invariant, by a result of Section 2. Which one, and why?

Solution. The count n_0. A diagonal matrix has rank equal to its number of nonzero entries, so $n_+ + n_- = \operatorname{rank}[\beta]_{\mathcal{B}}^{\mathcal{B}}$, and congruence preserves rank (Theorem 14.2.3). Hence $n_+ + n_-$ is the same for every orthogonal basis, and so is $n_0 = n - (n_+ + n_-)$. What is **not** yet known is how that total splits into positives and negatives — which is exactly the work left for the theorem below.

14.5.2 The law of inertia

Theorem 14.5.1. Sylvester's Law of Inertia

Let β be a symmetric bilinear form on a real vector space V with $\dim V = n$, and let $\mathcal{B}$ and $\mathcal{B}'$ be two orthogonal bases for β. Let (n_+, n_-, n_0) be the numbers of positive, negative, and zero diagonal entries of $[\beta]_{\mathcal{B}}^{\mathcal{B}}$, and (n'_+, n'_-, n'_0) the corresponding numbers for $[\beta]_{\mathcal{B}'}^{\mathcal{B}'}$. Then

$$n_+ = n'_+, \qquad n_- = n'_-, \qquad n_0 = n'_0.$$

Idea. Rank has already delivered $n_+ + n_- = n'_+ + n'_-$, so it is enough to prove $n_+ = n'_+$, and for that we may suppose $n_+ > n'_+$ and look for a disaster. The two bases carve out two subspaces with incompatible behavior: on $U = \operatorname{span}(v_1, \dots, v_{n_+})$ the form is **strictly positive** away from 0, because Eq. 14.4.1 has only positive coefficients there; on $U' = \operatorname{span}(v'_{n'_+ + 1}, \dots, v'_n)$ it is ≤ 0 everywhere, because there the coefficients are all ≤ 0. A nonzero vector in both would have $q(v) > 0$ and $q(v) \le 0$ at once, so the two subspaces meet only in 0. But their dimensions add up to more than n, and two such subspaces of an n-dimensional space cannot meet only in 0. Contradiction plus a dimension count.

Proof. Write $\mathcal{B} = (v_1, \dots, v_n)$ and $\mathcal{B}' = (v'_1, \dots, v'_n)$, and order each so that the positive entries come first, then the negative ones, then the zeros; reordering a basis does not change the three counts. Put $d_i = q(v_i)$ and $d'_j = q(v'_j)$, so that

$$d_1, \dots, d_{n_+} > 0, \qquad d'_{n'_+ + 1}, \dots, d'_n \le 0.$$

As in the Quick check, $n_+ + n_- = \operatorname{rank}[\beta]_{\mathcal{B}}^{\mathcal{B}} = \operatorname{rank}[\beta]_{\mathcal{B}'}^{\mathcal{B}'} = n'_+ + n'_-$ by Theorem 14.2.3, and $n_0 = n - (n_+ + n_-)$. So it suffices to prove $n_+ = n'_+$; the other two equalities then follow.

Suppose not. Swapping the names of the two bases if necessary, we may assume $n_+ > n'_+$. Put

$$U = \operatorname{span}(v_1, \dots, v_{n_+}), \qquad U' = \operatorname{span}(v'_{n'_+ + 1}, \dots, v'_n),$$

so that $\dim U = n_+$ and $\dim U' = n - n'_+$.

Claim: $U \cap U' = \{0\}$. Let $v \in U \cap U'$ with $v \neq 0$. Writing $v = \sum_{i \le n_+} c_i v_i$ and using Eq. 14.4.1 for the basis $\mathcal{B}$, whose later coordinates of v are zero,

$$q(v) = \sum_{i=1}^{n_+} d_i c_i^2 > 0,$$

because every d_i with $i \le n_+$ is positive, every $c_i^2 \ge 0$, and at least one $c_i \neq 0$ since $v \neq 0$. Writing instead $v = \sum_{j > n'_+} b_j v'_j$ and using Eq. 14.4.1 for $\mathcal{B}'$,

$$q(v) = \sum_{j=n'_+ + 1}^{n} d'_j b_j^2 \le 0,$$

because every d'_j with $j > n'_+$ is ≤ 0 and every $b_j^2 \ge 0$. A real number cannot be both > 0 and ≤ 0, so no such v exists, which proves the claim.

On the other hand, $U + U'$ is a subspace of V, so $\dim(U + U') \le n$, and the dimension formula (Theorem 1.7.3) gives

$$\begin{aligned} \dim(U \cap U') &= \dim U + \dim U' - \dim(U + U') \\ &\ge n_+ + (n - n'_+) - n \\ &= n_+ - n'_+ > 0. \end{aligned}$$

So $U \cap U'$ contains a nonzero vector, contradicting the claim. Therefore $n_+ = n'_+$, and with it $n_- = n'_-$ and $n_0 = n'_0$. This proves the theorem.

■

The name comes from mechanics, where the form is an inertia tensor and the theorem says that a change of coordinates cannot turn a direction of positive inertia into one of negative inertia. The invariant has a name of its own.

Definition 14.5.1. Inertia and Signature

Let β be a symmetric bilinear form on a real vector space V of dimension n, diagonalized in some orthogonal basis. The triple

$$(n_+, n_-, n_0)$$

of numbers of positive, negative, and zero diagonal entries is the **inertia** of β, and it is well defined by Theorem 14.5.1. The pair (n_+, n_-) is the **signature** of β, and $n_+ + n_- = \operatorname{rank} \beta$ while $n_0 = n - \operatorname{rank} \beta$. The same words are used for a real symmetric matrix $\mathbf{A}$, through the form $\beta(\mathbf{x}, \mathbf{y}) = \mathbf{x}^{\top} \mathbf{A} \mathbf{y}$, and we write $n_+(\mathbf{A})$, $n_-(\mathbf{A})$, $n_0(\mathbf{A})$.

Two conventions are in circulation and both are called the signature: the pair (n_+, n_-), used here, and the single integer $n_+ - n_-$. They carry the same information once the rank is known, and the difference is only ever notational; where a source writes "signature 2" it means $n_+ - n_- = 2$. The triple (n_+, n_-, n_0) is the safest thing to quote, since it also records the degeneracy.

14.5.3 The classification over the reals

Corollary 14.5.2. Classification of Real Symmetric Matrices by Signature

Let $\mathbf{A}, \mathbf{B} \in M_n(\mathbb{R})$ be symmetric. Then:

(a) $\mathbf{A}$ is congruent to $\mathbf{I}_{n_+} \oplus (-\mathbf{I}_{n_-}) \oplus \mathbf{O}_{n_0}$, where (n_+, n_-, n_0) is the inertia of $\mathbf{A}$, and to no other matrix of this shape;

(b) $\mathbf{A} \simeq \mathbf{B}$ if and only if $\mathbf{A}$ and $\mathbf{B}$ have the same inertia;

(c) the number of congruence classes of symmetric matrices in $M_n(\mathbb{R})$ is $\frac{1}{2}(n+1)(n+2)$.

Proof.

(a) The congruence to $\mathbf{I}_{n_+} \oplus (-\mathbf{I}_{n_-}) \oplus \mathbf{O}_{n_0}$ is Eq. 14.5.1, obtained from Theorem 14.4.2 followed by the scaling and sorting described above. That matrix is itself diagonal with inertia (n_+, n_-, n_0), so by Theorem 14.5.1 no matrix of the same shape with a different triple can be congruent to $\mathbf{A}$.

(b) ($\Rightarrow$) A congruence $\mathbf{B} = \mathbf{P}^\top \mathbf{A} \mathbf{P}$ with $\mathbf{P}$ invertible is a change of basis for the form (Theorem 14.2.1), so an orthogonal basis diagonalizing $\mathbf{B}$ yields one diagonalizing $\mathbf{A}$ with the same diagonal, and Theorem 14.5.1 makes the inertia equal.
($\Leftarrow$) If $\mathbf{A}$ and $\mathbf{B}$ have the same inertia, each is congruent to the same matrix $\mathbf{I}_{n_+} \oplus (-\mathbf{I}_{n_-}) \oplus \mathbf{O}_{n_0}$ by (a), hence to each other, since congruence is an equivalence relation (Proposition 14.2.2).

(c) By (b) the classes correspond to the triples (n_+, n_-, n_0) of nonnegative integers summing to n, equivalently to the pairs (n_+, n_-) with $n_+ + n_- \le n$. For each value $k = n_+ + n_- \in \{0, 1, \dots, n\}$ there are $k+1$ choices of n_+, so the total is $1 + 2 + \cdots + (n+1) = \frac{1}{2}(n+1)(n+2)$. This proves the corollary.

■

So congruence over $\mathbb{R}$ has a complete and finite list of invariants, in sharp contrast to similarity, where even 2×2 real symmetric matrices fall into infinitely many classes (one for each unordered pair of eigenvalues). Congruence forgets a great deal; the point of the law of inertia is that what it forgets is precisely everything except the signs.

14.5.4 Reconciling congruence with the spectral theorem

Chapter 12 diagonalized a real symmetric matrix a different way, by an **orthogonal** $\mathbf{Q}$, and got the eigenvalues on the diagonal. Chapter 12 also pointed out why that theorem is special: because $\mathbf{Q}^\top = \mathbf{Q}^{-1}$, the one equation $\mathbf{Q}^\top \mathbf{A} \mathbf{Q} = \mathbf{D}$ can be read twice, and "an orthogonal change of variables is a similarity *and* a congruence, so it keeps the eigenvalues and the geometry at once". Reading the same equation twice is all that is needed to identify the inertia.

Theorem 14.5.3. Inertia Is the Sign Pattern of the Spectrum

Let $\mathbf{A} \in M_n(\mathbb{R})$ be symmetric. Then $n_+(\mathbf{A})$ is the number of positive eigenvalues of $\mathbf{A}$, counted with multiplicity, $n_-(\mathbf{A})$ the number of negative ones, and $n_0(\mathbf{A})$ the multiplicity of 0.

Proof. By Corollary 12.5.2 there is an orthogonal $\mathbf{Q} \in \mathrm{O}(n)$ and a real diagonal $\mathbf{D} = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ with $\mathbf{A} = \mathbf{Q}\mathbf{D}\mathbf{Q}^\top$, where $\lambda_1, \ldots, \lambda_n$ is the eigenvalue list of $\mathbf{A}$ with multiplicity. Equivalently $\mathbf{D} = \mathbf{Q}^\top\mathbf{A}\mathbf{Q}$.

Read as a **similarity**: $\mathbf{Q}^\top = \mathbf{Q}^{-1}$, so $\mathbf{D} = \mathbf{Q}^{-1}\mathbf{A}\mathbf{Q}$ and the diagonal entries of $\mathbf{D}$ are the eigenvalues of $\mathbf{A}$.

Read as a **congruence**: $\mathbf{Q}$ is invertible, so $\mathbf{D} = \mathbf{Q}^\top\mathbf{A}\mathbf{Q}$ exhibits $\mathbf{D}$ as congruent to $\mathbf{A}$ (Definition 14.2.1). Being diagonal, $\mathbf{D}$ is the matrix of the form of $\mathbf{A}$ in an orthogonal basis, namely the columns of $\mathbf{Q}$. By Theorem 14.5.1 its numbers of positive, negative, and zero entries are the inertia of $\mathbf{A}$.

The same list of numbers is being counted both times, so the counts agree. This proves the theorem.

■

This is what Chapter 13 promised when, having diagonalized a pencil and found $\mathbf{S}^\top\mathbf{B}\mathbf{S} = \mathrm{diag}(-1, 3)$ while the eigenvalues of $\mathbf{B}$ were $3 \pm 3\sqrt{2}$, it observed that the signs $(-, +)$ "match the signs of the eigenvalues of $\mathbf{B}$ itself" and said that Chapter 14 would show this is no coincidence. It is not: the congruence $\mathbf{S}^\top\mathbf{B}\mathbf{S}$ has the inertia of $\mathbf{B}$ by Theorem 14.5.1, and that inertia is the sign pattern of the spectrum by Theorem 14.5.3. Here $3 + 3\sqrt{2} > 0$ and $3 - 3\sqrt{2} < 0$ — since $3\sqrt{2} > 4$ — so the inertia of $\mathbf{B}$ is $(1, 1, 0)$, and any congruent diagonal matrix must show exactly one positive and one negative entry, as $\mathrm{diag}(-1, 3)$ does.

The same reading settles the definiteness tests of Chapter 13 in one line: $\mathbf{A} \succ \mathbf{O}$ means $\mathbf{x}^\top\mathbf{A}\mathbf{x} > 0$ for all $\mathbf{x} \neq \mathbf{0}$, which by Theorem 13.1.5 is the statement that every eigenvalue is positive, which by Theorem 14.5.3 is the statement that the inertia is $(n, 0, 0)$. Section 6 takes that dictionary further.

Warning. **Congruent matrices have the same inertia and unrelated eigenvalues.** Take $\mathbf{A} = \mathrm{diag}(1, -1)$ and $\mathbf{P} = \mathrm{diag}(2, 3)$. Then

$$\mathbf{P}^\top\mathbf{A}\mathbf{P} = \mathrm{diag}(4, -9),$$

so $\mathbf{A} \simeq \mathrm{diag}(4, -9)$, yet the eigenvalues have moved from $1, -1$ to $4, -9$. Nothing about their sizes, their ratios, their sum, or their product is preserved — tr goes from 0 to -5, det from -1 to -36 — and only the pattern $(+, -)$ survives. For a second witness with no scaling in sight, $\mathbf{I}_2$ is congruent to $\mathbf{S}^\top\mathbf{S} = \left(\begin{smallmatrix} 1 & 1 \\ 1 & 2 \end{smallmatrix}\right)$ with $\mathbf{S} = \left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$; the eigenvalues change from $1, 1$ to $(3 \pm \sqrt{5})/2$, and the inertia stays $(2, 0, 0)$.

14.5.5 Computing a signature

Theorem 14.5.3 gives a method — find the eigenvalues and count signs — but it is the expensive one, since it needs the roots of a degree-n polynomial. The cheap method is the algorithm of the previous section, which needs only arithmetic: diagonalize by paired row and column operations and read off the signs. The eigenvalues are never computed and never needed.

Example. A signature by symmetric elimination

Find the inertia of

$$\mathbf{A} = \begin{pmatrix} 1 & 1 & 2 \\ 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix},$$

and say whether the form $q(\mathbf{x}) = \mathbf{x}^\top \mathbf{A}\mathbf{x}$ is positive definite.

Solution. The corner entry is 1, so no repair move is needed.

Clear the first row and column. Subtract row 1 from row 2 and 2× row 1 from row 3, then do the same two operations on columns:

$$\begin{pmatrix} 1 & 1 & 2 \\ 0 & 1 & 1 \\ 0 & 1 & -3 \end{pmatrix} \longrightarrow \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & -3 \end{pmatrix}.$$

Clear the second row and column. The pivot is 1 and $a_{23} = 1$, so subtract row 2 from row 3 and then column 2 from column 3:

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & -4 \end{pmatrix} \longrightarrow \operatorname{diag}(1, 1, -4).$$

So the inertia is $(2, 1, 0)$ and the signature is $(2, 1)$. The form is **not** positive definite: it is negative at the third basis vector of the new basis, namely the third column of

$$\mathbf{P} = \begin{pmatrix} 1 & -1 & -1 \\ 0 & 1 & -1 \\ 0 & 0 & 1 \end{pmatrix},$$

which is $\mathbf{p}_3 = (-1, -1, 1)$. Checking directly, $\mathbf{A}\mathbf{p}_3 = (0, 0, -4)$ and $\mathbf{p}_3 \cdot (0, 0, -4) = -4 < 0$.

The eigenvalue route would have to factor $\det(x\mathbf{I} - \mathbf{A}) = x^3 - 4x^2 - 9x + 4$, whose roots are irrational; the elimination never met them. As a partial check on the arithmetic, $\det \mathbf{A} = -4$ and $\det \operatorname{diag}(1, 1, -4) = -4$, consistent with $\det \mathbf{P} = 1$.

Over $\mathbb{R}$, then, a symmetric matrix carries exactly two pieces of congruence data, and both are cheap: its rank and how that rank splits into positive and negative directions. The next section reads the same two numbers as a statement about definiteness, and uses them to decide maxima and minima.

14.5.6 Exercises

14.5.6.1 A. Check your understanding

Exercise. A1

(a) Define the **inertia** and the **signature** of a real symmetric matrix.

(b) State Theorem 14.5.1, with all hypotheses.

(c) Which part of the inertia is already an invariant before the law of inertia is proved, and by which earlier result?

(d) Determine whether the following statement is correct, and justify your answer: congruent real symmetric matrices have the same trace.

(e) Explain in one sentence why an orthogonal change of basis lets the spectral theorem compute an inertia.

Solution.

(a) Diagonalize the associated form in an orthogonal basis. The inertia is the triple (n_+, n_-, n_0) of numbers of positive, negative, and zero diagonal entries, and the signature is the pair (n_+, n_-).

(b) If β is a symmetric bilinear form on a finite-dimensional real vector space and $\mathcal{B}, \mathcal{B}'$ are two orthogonal bases for β, then the two matrices $[\beta]_{\mathcal{B}}^{\mathcal{B}}$ and $[\beta]_{\mathcal{B}'}^{\mathcal{B}'}$ have the same numbers of positive, negative, and zero diagonal entries.

(c) The count n_0, equivalently the sum $n_+ + n_-$, because that sum is the rank and congruence preserves rank (Theorem 14.2.3).

(d) Incorrect. $\operatorname{diag}(1, -1)$ and $\operatorname{diag}(4, -9)$ are congruent, by the congruence with $\mathbf{P} = \operatorname{diag}(2, 3)$, and their traces are 0 and -5. Trace is a similarity invariant (Theorem 4.2.5), not a congruence invariant.

(e) For an orthogonal $\mathbf{Q}$ we have $\mathbf{Q}^{\top} = \mathbf{Q}^{-1}$, so the single equation $\mathbf{Q}^{\top}\mathbf{A}\mathbf{Q} = \mathbf{D}$ is simultaneously a similarity, which puts the eigenvalues on the diagonal, and a congruence, which makes that diagonal compute the inertia.

14.5.6.2 B. Practice

Exercise. B1: Signatures by elimination

Find the inertia of each of the following real symmetric matrices by paired row and column operations. Do not compute any eigenvalues.

(a) $\mathbf{A} = \begin{pmatrix} 1 & 2 & 0 \\ 2 & 1 & 3 \\ 0 & 3 & 1 \end{pmatrix}$.

(b) $\mathbf{B} = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}$.

(c) $\mathbf{C} = \begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix}$.

Solution.

(a) By Exercise (B1: Symmetric elimination), $\mathbf{A}$ is congruent to $\operatorname{diag}(1, -3, 4)$. Inertia $(2, 1, 0)$.

(b) By Example (Diagonalizing a symmetric matrix with zero diagonal), $\mathbf{B}$ is congruent to $\operatorname{diag}(2, -2, -2)$. Inertia $(1, 2, 0)$.

(c) Subtract $2\times$ row 1 from row 2, then $2\times$ column 1 from column 2:

$$\begin{pmatrix} 1 & 2 \\ 0 & 0 \end{pmatrix} \longrightarrow \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}.$$

Inertia $(1,0,1)$: the form $q(\mathbf{x}) = (x_1 + 2x_2)^2$ is degenerate, vanishing on the line $x_1 = -2x_2$.

Exercise. B2: Which of these are congruent?

Sort the following real symmetric matrices into congruence classes, justifying each answer.

$$\mathbf{A}_1 = \operatorname{diag}(1,1,-1), \quad \mathbf{A}_2 = \operatorname{diag}(5,2,-7), \quad \mathbf{A}_3 = \operatorname{diag}(1,-1,-1),$$

$$\mathbf{A}_4 = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \qquad \mathbf{A}_5 = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -3 \end{pmatrix}.$$

Solution. Compute each inertia and apply Corollary 14.5.2 (b).

$\mathbf{A}_1$: $(2,1,0)$. $\mathbf{A}_2$: $(2,1,0)$. $\mathbf{A}_3$: $(1,2,0)$. $\mathbf{A}_5$: $(1,1,1)$.

For $\mathbf{A}_4$, the top-left 2×2 block is the form $2x_1x_2$, which Exercise (B2: No pivot on the diagonal) diagonalizes to $\operatorname{diag}(2,-2)$; so $\mathbf{A}_4$ is congruent to $\operatorname{diag}(2,-2,1)$, with inertia $(2,1,0)$.

Therefore the classes are $\{\mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_4\}$, $\{\mathbf{A}_3\}$, and $\{\mathbf{A}_5\}$. Note that $\mathbf{A}_1$ and $\mathbf{A}_3$ have the same rank and are still not congruent, so rank alone does not classify over $\mathbb{R}$ — unlike over $\mathbb{C}$ (Corollary 14.4.4).

Exercise. B3: Inertia from a characteristic polynomial

Let $\mathbf{A} \in M_3(\mathbb{R})$ be symmetric with $\det(x\mathbf{I} - \mathbf{A}) = x^3 - 2x^2 - 5x + 6$.

(a) Find the inertia of $\mathbf{A}$.

(b) Is $\mathbf{A}$ congruent to $\operatorname{diag}(1,1,-1)$? Is it similar to it?

Solution.

(a) The polynomial factors as $(x-1)(x+2)(x-3)$, since 1 is a root and dividing gives $x^2 - x - 6$. The eigenvalues are $1, 3, -2$, so by Theorem 14.5.3 the inertia is $(2,1,0)$.

(b) Congruent: yes, since $\operatorname{diag}(1,1,-1)$ also has inertia $(2,1,0)$ and Corollary 14.5.2 (b) applies. Similar: no, because similar matrices have equal characteristic polynomials, and $\operatorname{diag}(1,1,-1)$ has $(x-1)^2(x+1) \neq (x-1)(x-3)(x+2)$.

14.5.6.3 C. Going deeper

Exercise. C1: An intrinsic description of n_+

Let β be a symmetric bilinear form on a real vector space V of dimension n, with quadratic form q. Call a subspace $U \subseteq V$ **positive** if $q(v) > 0$ for every $v \in U$ with $v \neq 0$.

(a) Prove that n_+ is the largest dimension of a positive subspace of V.

(b) Deduce Theorem 14.5.1 again from (a), in two lines.

(c) State and prove the corresponding description of n_-.

Hint: for (a), use an orthogonal basis to produce one positive subspace of dimension n_+, and the dimension formula to rule out a larger one.

Solution.

(a) Fix an orthogonal basis $(v_1, \dots, v_n)$ ordered so that $d_1, \dots, d_{n_+} > 0$ and $d_j \leq 0$ for $j > n_+$.
A positive subspace of dimension n_+. Put $U = \operatorname{span}(v_1, \dots, v_{n_+})$. For $v = \sum_{i \leq n_+} c_i v_i$ nonzero, Eq. 14.4.1 gives $q(v) = \sum_{i \leq n_+} d_i c_i^2 > 0$, since all $d_i > 0$ and some $c_i \neq 0$. So U is positive and $\dim U = n_+$.
No larger one. Let X be positive with $\dim X = k$, and put $W = \operatorname{span}(v_{n_+ + 1}, \dots, v_n)$, of dimension $n - n_+$. Every $w \in W$ has $q(w) = \sum_{j > n_+} d_j c_j^2 \leq 0$. If $k > n_+$ then by Theorem 1.7.3,

$$\dim(X \cap W) \geq k + (n - n_+) - n = k - n_+ > 0,$$

so there is a nonzero $v \in X \cap W$ with $q(v) > 0$ and $q(v) \leq 0$, which is impossible. Hence $k \leq n_+$, and the maximum is n_+.

(b) The right-hand side of (a) — the largest dimension of a positive subspace — is defined from β alone and mentions no basis. So the number n_+ computed from any orthogonal basis equals that basis-free quantity, hence is the same for all of them. Then $n_+ + n_- = \operatorname{rank} \beta$ is basis-free too (Theorem 14.2.3), so n_- and n_0 are determined as well.

(c) n_- is the largest dimension of a **negative** subspace, one on which $q(v) < 0$ for every $v \neq 0$. Apply (a) to the form $-\beta$, whose diagonal entries in any orthogonal basis are the negatives of those of β, so that $n_+(-\beta) = n_-(\beta)$, and whose positive subspaces are exactly the negative subspaces of β.

Exercise. C2: Congruence cannot be read off the eigenvalues, and inertia can

(a) Give two real symmetric 2×2 matrices that are congruent but have no eigenvalue in common.

(b) Prove that if $\mathbf{A}$ and $\mathbf{B}$ are real symmetric and **similar**, then they are congruent.

(c) Give two real symmetric 2×2 matrices that are congruent but not similar, so that the converse of (b) fails.

Solution.

(a) $\operatorname{diag}(1, -1)$ and $\operatorname{diag}(4, -9)$, congruent by $\mathbf{P} = \operatorname{diag}(2, 3)$. Their spectra are $\{1, -1\}$ and $\{4, -9\}$, which are disjoint.

(b) Similar matrices have the same characteristic polynomial (Theorem 7.10.3), hence the same eigenvalues with multiplicity. By Theorem 14.5.3 they then have the same inertia, and by Corollary 14.5.2 (b) they are congruent. (Alternatively and more directly: by Corollary 12.5.2 each is orthogonally congruent to the diagonal matrix of its eigenvalues, and those diagonal matrices agree up to a permutation.)

(c) The pair in (a) again: $\operatorname{diag}(1,-1)$ and $\operatorname{diag}(4,-9)$ are congruent, but their characteristic polynomials are x^2-1 and $x^2+5x-36$, so they are not similar (Theorem 7.10.3). Congruence is strictly coarser than similarity on real symmetric matrices.

Exercise. C3: Inertia of a block matrix

Let $\mathbf{A} \in M_m(\mathbb{R})$ and $\mathbf{B} \in M_k(\mathbb{R})$ be symmetric. Let also $\mathbf{D} \in M_n(\mathbb{R})$ be symmetric and invertible, let $\mathbf{b} \in \mathbb{R}^n$, let $c \in \mathbb{R}$, and put

$$\mathbf{M} = \begin{pmatrix} \mathbf{D} & \mathbf{b} \\ \mathbf{b}^{\top} & c \end{pmatrix} \in M_{n+1}(\mathbb{R}).$$

(a) Prove that the inertia of $\mathbf{A} \oplus \mathbf{B}$ is the entrywise sum of the inertias of $\mathbf{A}$ and $\mathbf{B}$.

(b) Prove that $\mathbf{M}$ is congruent to $\mathbf{D} \oplus (c - \mathbf{b}^{\top}\mathbf{D}^{-1}\mathbf{b})$.

(c) Hence find the inertia of $\mathbf{M}$ when $\mathbf{D} = \operatorname{diag}(1,-1)$, $\mathbf{b} = (1,1)$, and $c = 0$.

Hint: for (b), the congruence is the block elimination of Chapter 8, and the leftover scalar is a Schur complement.

Solution.

(a) Let $\mathbf{P}^{\top}\mathbf{A}\mathbf{P}$ and $\mathbf{Q}^{\top}\mathbf{B}\mathbf{Q}$ be the canonical forms of Corollary 14.5.2 (a). Then $\mathbf{P} \oplus \mathbf{Q}$ is invertible and

$$(\mathbf{P} \oplus \mathbf{Q})^{\top}(\mathbf{A} \oplus \mathbf{B})(\mathbf{P} \oplus \mathbf{Q}) = (\mathbf{P}^{\top}\mathbf{A}\mathbf{P}) \oplus (\mathbf{Q}^{\top}\mathbf{B}\mathbf{Q}),$$

a diagonal matrix whose positive, negative, and zero entries are those of the two blocks together. By Theorem 14.5.1 those counts are the inertia of $\mathbf{A} \oplus \mathbf{B}$.

(b) Put

$$\mathbf{S} = \begin{pmatrix} \mathbf{I}_n & -\mathbf{D}^{-1}\mathbf{b} \\ \mathbf{0}^{\top} & 1 \end{pmatrix},$$

which is invertible, being unit upper triangular. Multiplying, the first block row of $\mathbf{M}\mathbf{S}$ is $(\mathbf{D} \mid \mathbf{D}(-\mathbf{D}^{-1}\mathbf{b}) + \mathbf{b}) = (\mathbf{D} \mid \mathbf{0})$, and then

$$\mathbf{S}^{\top}\mathbf{M}\mathbf{S} = \begin{pmatrix} \mathbf{D} & \mathbf{0} \\ \mathbf{0}^{\top} & c - \mathbf{b}^{\top}\mathbf{D}^{-1}\mathbf{b} \end{pmatrix},$$

using $\mathbf{D}^{\top} = \mathbf{D}$. This is $\mathbf{D} \oplus (c - \mathbf{b}^{\top}\mathbf{D}^{-1}\mathbf{b})$, the Schur complement $\mathbf{M}/\mathbf{D}$ sitting in the corner.

(c) Here $\mathbf{D}^{-1} = \operatorname{diag}(1,-1)$, so $\mathbf{b}^{\top}\mathbf{D}^{-1}\mathbf{b} = 1 - 1 = 0$ and the leftover scalar is $c - 0 = 0$. By (b) and (a), the inertia of $\mathbf{M}$ is the inertia of $\operatorname{diag}(1,-1)$ plus that of the 1×1 zero matrix, namely $(1,1,0) + (0,0,1) = (1,1,1)$. In particular $\mathbf{M}$ is singular.

14.6 Definiteness and Second Derivatives

Chapter 13 studied positive definite matrices at length, and the previous section showed that the signature is what congruence remembers. Putting the two together costs almost nothing and buys two things: a vocabulary that covers the forms Chapter 13 had no name for, the indefinite and the semidefinite ones, and the classical application that made quadratic forms worth studying in the first place. Nothing from Chapter 13 is reproved here. The field is $\mathbb{R}$ throughout, since none of the five words below means anything without an order.

Throughout, V is a finite-dimensional real vector space, β is a **symmetric** bilinear form on V (Definition 14.3.1), $q(v) = \beta(v, v)$ is its quadratic form (Definition 14.3.2), and $\mathbf{A} = [\beta]_{\mathcal{B}}^{\mathcal{B}}$ is its matrix in a basis $\mathcal{B}$ (Definition 14.1.2), a real symmetric matrix.

14.6.1 Five words for the sign of a form

Chapter 13 named two of the possible behaviors of q: never negative, and always positive away from $\mathbf{0}$. There are three more, and the missing ones are not exotic. The form $q(x_1, x_2) = x_1^2 - x_2^2$ is positive at $\mathbf{e}_1$, negative at $\mathbf{e}_2$, and nothing in Chapter 13's vocabulary describes it.

A form is definite when q has one strict sign, semidefinite when it has one sign but is allowed to vanish, and indefinite when it has both signs.

Definition 14.6.1. Definiteness of a Real Symmetric Form

Let β be a symmetric bilinear form on a real vector space V, with quadratic form q. Then β is:

(a) **positive definite** if $q(v) > 0$ for **every nonzero** $v \in V$;
(b) **positive semidefinite** if $q(v) \geq 0$ for every $v \in V$;
(c) **negative definite** if $q(v) < 0$ for every nonzero $v \in V$;
(d) **negative semidefinite** if $q(v) \leq 0$ for every $v \in V$;
(e) **indefinite** if there exist $u, w \in V$ with $q(u) > 0$ and $q(w) < 0$.

The same five words apply to a real symmetric matrix $\mathbf{A}$, through the form $\beta(\mathbf{x}, \mathbf{y}) = \mathbf{x}^{\top}\mathbf{A}\mathbf{y}$ on $\mathbb{R}^n$.

The small words carry the weight. In (a) and (c) the inequality is **strict** and the vector must be **nonzero**; in (b) and (d) it is weak and $v = 0$ is allowed, where $q(0) = 0$ makes the condition automatic. Definite implies semidefinite of the same sign. Clause (e) is not the negation of (b): $q(x_1, x_2) = -x_1^2$ is neither positive semidefinite nor indefinite. For matrices, (a) and (b) are Definition 13.1.1 restricted to $F = \mathbb{R}$, so $\mathbf{A} \succ \mathbf{O}$ and $\mathbf{A} \succeq \mathbf{O}$ keep their meanings, and the negative versions are the positive ones for $-\mathbf{A}$.

Four examples on $\mathbb{R}^2$, all diagonal so that q can be read off Eq. 14.4.1: $x_1^2 + x_2^2$ is positive definite; x_1^2 is positive semidefinite but not definite, since it vanishes at $\mathbf{e}_2 \neq \mathbf{0}$; $-x_1^2 - x_2^2$ is negative definite; $x_1^2 - x_2^2$ is indefinite. The zero form is the degenerate case, satisfying (b) and (d) and nothing else — which is why the five classes overlap rather than partition. Changing one sign in the third gives the nonexample $-x_1^2 + x_2^2$: it is no longer negative semidefinite, and the clause that fails is (d) at $\mathbf{v} = \mathbf{e}_2$, where $q = 1 > 0$.

Now the reading through the signature, which makes all five conditions into statements about three integers.

Theorem 14.6.1. Definiteness Is the Signature

Let β be a symmetric bilinear form on a real vector space V of dimension n, with inertia (n_+, n_-, n_0) (Definition 14.5.1). Then:

(a) β is positive definite $\iff$ $(n_+, n_-, n_0) = (n, 0, 0)$;
(b) β is positive semidefinite $\iff$ $n_- = 0$;
(c) β is negative definite $\iff$ $(n_+, n_-, n_0) = (0, n, 0)$;
(d) β is negative semidefinite $\iff$ $n_+ = 0$;
(e) β is indefinite $\iff$ $n_+ > 0$ and $n_- > 0$.

Idea. Fix an orthogonal basis. Then q is $\sum_i d_i c_i^2$ with no cross terms, and each of the five conditions becomes a visible condition on the signs of the d_i: plugging in a basis vector reads off one d_i, and a general v mixes the d_i with nonnegative weights. Sylvester's law is what lets us fix one basis and still be talking about β.

Proof. By Theorem 14.4.2 there is an orthogonal basis $\mathcal{B} = (v_1, \dots, v_n)$ for β; put $d_i = q(v_i)$. By Theorem 14.5.1 the numbers of positive, negative, and zero d_i are n_+, n_-, and n_0, whichever orthogonal basis is used. For $v = \sum_i c_i v_i$, Eq. 14.4.1 gives

$$q(v) = \sum_{i=1}^{n} d_i c_i^2 .$$

(b) ($\Rightarrow$) If some $d_j < 0$ then $q(v_j) = d_j < 0$, so β is not positive semidefinite. ($\Leftarrow$) If no d_i is negative then every term $d_i c_i^2$ is ≥ 0, so $q(v) \geq 0$.

(c) ($\Rightarrow$) A positive definite β is positive semidefinite, so $n_- = 0$ by (b); and if some $d_j = 0$ then $q(v_j) = 0$ with $v_j \neq 0$, a contradiction. Hence $n_0 = 0$ and $n_+ = n$. ($\Leftarrow$) If every $d_i > 0$ and $v \neq 0$, some $c_j \neq 0$, so $q(v) \geq d_j c_j^2 > 0$, the other terms being ≥ 0.

(d) and (d) follow by applying (a) and (b) to $-\beta$, whose diagonal entries are $-d_i$ and whose inertia is therefore (n_-, n_+, n_0).

(e) ($\Rightarrow$) Suppose $q(u) > 0$ and $q(w) < 0$. Then β is neither positive semidefinite nor negative semidefinite, so $n_- > 0$ by (b) and $n_+ > 0$ by (d). ($\Leftarrow$) If $n_+ > 0$ and $n_- > 0$, some $d_i > 0$ and some $d_j < 0$, and the basis vectors v_i, v_j are the required u, w. This proves the theorem.

■

So the five words sort the $\frac{1}{2}(n+1)(n+2)$ congruence classes of Corollary 14.5.2 into named families, and every one of them is decided by counting signs. Two of these five verdicts were already settled in Chapter 13 by other means, and it is worth putting the tests side by side.

Proposition 14.6.2. Chapter 13's Tests, in the Language of Forms

Let β be a symmetric bilinear form on an n-dimensional real vector space, $\mathbf{A} = [\beta]_{\mathcal{B}}^{\mathcal{B}}$ its matrix in any basis, and $\mathbf{A}_k$ the top-left $k \times k$ corner of $\mathbf{A}$. Then:

(a) β is positive definite $\iff$ every eigenvalue of $\mathbf{A}$ is positive $\iff$ $\det \mathbf{A}_k > 0$ for

$k = 1, \dots, n$;

(b) β is negative definite $\iff$ every eigenvalue of $\mathbf{A}$ is negative $\iff$ $(-1)^k \det \mathbf{A}_k > 0$ for $k = 1, \dots, n$.

Proof.

(a) By Definition 14.1.2 and Theorem 14.1.1, $q(v) = \mathbf{x}^\top \mathbf{A}\mathbf{x}$ where $\mathbf{x} = [v]_{\mathcal{B}}$, and $v \neq 0$ exactly when $\mathbf{x} \neq \mathbf{0}$. So β is positive definite exactly when $\mathbf{A} \succ \mathbf{O}$, and the two stated tests are Theorem 13.1.5 (b) and (d), the second being Sylvester's criterion.

(b) Apply (a) to $-\beta$, whose matrix is $-\mathbf{A}$. The eigenvalues of $-\mathbf{A}$ are the negatives of those of $\mathbf{A}$, and $(-\mathbf{A})_k = -\mathbf{A}_k$ has $\det(-\mathbf{A}_k) = (-1)^k \det \mathbf{A}_k$, since scaling each of the k rows by -1 multiplies the determinant by -1 once per row (Theorem 7.5.1 (b)). This proves the proposition.

■

The eigenvalue test is the one to think with, the minor test the one to compute with; cheapest of all is the symmetric elimination of Section 4, which returns the whole signature rather than a verdict. Nothing is said about the semidefinite cases on purpose: Sylvester's criterion has no weak version, as Chapter 13 showed with $\operatorname{diag}(0, -1)$, whose leading minors are both 0.

Quick check. A real symmetric $\mathbf{A} \in M_3(\mathbb{R})$ has $\det \mathbf{A}_1 = -2$, $\det \mathbf{A}_2 = 5$, $\det \mathbf{A}_3 = -1$. What is its definiteness?

Solution. Negative definite. The signs alternate from a negative one, so $(-1)^k \det \mathbf{A}_k > 0$ for $k = 1, 2, 3$: indeed $(-1)(-2) = 2 > 0$, $(+1)(5) = 5 > 0$, $(-1)(-1) = 1 > 0$. By Proposition 14.6.2 (b) the form is negative definite, and by Theorem 14.6.1 (c) its inertia is $(0, 3, 0)$. Note that $\det \mathbf{A}_3 = \det \mathbf{A} < 0$ on its own would have been no evidence at all.

14.6.2 The second-derivative test

Here is the use. A differentiable function $f\colon \mathbb{R}^n \to \mathbb{R}$ is constant to first order at a critical point, so whether the point is a minimum, a maximum, or neither is decided — usually — by the second-order behavior. That behavior is a quadratic form, and Theorem 14.6.1 decides it. The linear algebra is the whole content; what has to be imported is one theorem of analysis, and we state it rather than hide it.

Definition 14.6.2. Hessian

Let f be a real-valued function on an open set $U \subseteq \mathbb{R}^n$, twice continuously differentiable on U, and let $\mathbf{a} \in U$. The **Hessian** of f at $\mathbf{a}$ is the matrix

$$\mathbf{H}_f(\mathbf{a}) \in M_n(\mathbb{R}), \qquad \big(\mathbf{H}_f(\mathbf{a})\big)_{ij} = \frac{\partial^2 f}{\partial x_i \, \partial x_j}(\mathbf{a}).$$

A point $\mathbf{a} \in U$ with $\nabla f(\mathbf{a}) = \mathbf{0}$ is a **critical point** of f.

Two facts about $\mathbf{H}_f(\mathbf{a})$ are needed, and neither is linear algebra. They are exactly the analysis this book imports, so we set them out as assumptions rather than let them pass unnoticed. Let f be twice continuously differentiable on an open $U \subseteq \mathbb{R}^n$ and let $\mathbf{a} \in U$.

(A1) *(Symmetry of the second derivatives.)* The mixed partial derivatives satisfy $\partial^2 f/\partial x_i \partial x_j = \partial^2 f/\partial x_j \partial x_i$ on U, so $\mathbf{H}_f(\mathbf{a})$ is a **symmetric** matrix.

(A2) *(Taylor's theorem with a second-order remainder.)* For every $\mathbf{h}$ small enough that $\mathbf{a}+\mathbf{h} \in U$,

$$f(\mathbf{a}+\mathbf{h}) = f(\mathbf{a}) + \nabla f(\mathbf{a}) \cdot \mathbf{h} + \tfrac{1}{2}\,\mathbf{h}^{\top}\mathbf{H}_f(\mathbf{a})\mathbf{h} + r(\mathbf{h}),$$

where the remainder satisfies $r(\mathbf{h})/\|\mathbf{h}\|^2 \to 0$ as $\mathbf{h} \to \mathbf{0}$.

These are two standard theorems of multivariable calculus. This book proves neither. They are quoted so that the theorem below has hypotheses to stand on, and Chapter 18 quotes them again for its second-derivative test of convexity, exactly as the theorem of Abel and Ruffini was quoted in Chapter 11. Everything after them is linear algebra.

By (A1) the Hessian is a real symmetric matrix, so it has an inertia, and by Theorem 14.6.1 that inertia decides the definiteness of the quadratic form $\mathbf{h} \mapsto \mathbf{h}^{\top}\mathbf{H}_f(\mathbf{a})\mathbf{h}$. One small lemma converts definiteness into a quantitative bound, and it is the only place where a real number is needed rather than a sign.

Lemma 14.6.3. A Positive Definite Form Is Bounded Below by a Multiple of the Norm Squared

Let $\mathbf{A} \in M_n(\mathbb{R})$ be symmetric and positive definite, and let λ be its smallest eigenvalue. Then $\lambda > 0$ and

$$\mathbf{x}^{\top}\mathbf{A}\mathbf{x} \;\geq\; \lambda\|\mathbf{x}\|^2 \qquad \text{for every } \mathbf{x} \in \mathbb{R}^n.$$

Proof. By Corollary 12.5.2 there are an orthogonal $\mathbf{Q} \in \mathrm{O}(n)$ and a diagonal $\mathbf{D} = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$ listing the eigenvalues of $\mathbf{A}$ with $\mathbf{A} = \mathbf{Q}\mathbf{D}\mathbf{Q}^{\top}$. By Theorem 13.1.5 (b) every $\lambda_i > 0$, so $\lambda = \min_i \lambda_i > 0$. Put $\mathbf{y} = \mathbf{Q}^{\top}\mathbf{x}$. Then

$$\mathbf{x}^{\top}\mathbf{A}\mathbf{x} = \mathbf{y}^{\top}\mathbf{D}\mathbf{y} = \sum_{i=1}^{n} \lambda_i y_i^2 \;\geq\; \lambda \sum_{i=1}^{n} y_i^2 = \lambda\|\mathbf{y}\|^2,$$

using $\lambda_i \geq \lambda$ and $y_i^2 \geq 0$ termwise. Finally $\mathbf{Q}\mathbf{Q}^{\top} = \mathbf{I}$, so $\|\mathbf{y}\|^2 = \mathbf{x}^{\top}\mathbf{Q}\mathbf{Q}^{\top}\mathbf{x} = \|\mathbf{x}\|^2$. This proves the lemma.

■

Theorem 14.6.4. The Second-Derivative Test

Let f be a real-valued function, twice continuously differentiable on an open set $U \subseteq \mathbb{R}^n$, let $\mathbf{a} \in U$ be a critical point of f, and let (n_+, n_-, n_0) be the inertia of $\mathbf{H}_f(\mathbf{a})$. Then:

(a) if $\mathbf{H}_f(\mathbf{a})$ is positive definite, that is $(n_+, n_-, n_0) = (n, 0, 0)$, then $\mathbf{a}$ is a strict local minimum of f;

(b) if $\mathbf{H}_f(\mathbf{a})$ is negative definite, that is $(n_+, n_-, n_0) = (0, n, 0)$, then $\mathbf{a}$ is a strict local maximum of f;

(c) if $\mathbf{H}_f(\mathbf{a})$ is indefinite, that is $n_+ > 0$ and $n_- > 0$, then $\mathbf{a}$ is neither a local minimum

nor a local maximum: every neighborhood of $\mathbf{a}$ contains points where $f > f(\mathbf{a})$ and points where $f < f(\mathbf{a})$. Such an $\mathbf{a}$ is a **saddle point**.

If $n_0 > 0$ and $\mathbf{a}$ is not covered by (a)–(c), the test gives no information.

Idea. Taylor says $f(\mathbf{a}+\mathbf{h}) - f(\mathbf{a})$ is $\frac{1}{2}q(\mathbf{h})$ plus a remainder that is smaller than $\|\mathbf{h}\|^2$ near $\mathbf{0}$, where $q(\mathbf{h}) = \mathbf{h}^\top \mathbf{H}_f(\mathbf{a})\mathbf{h}$. For (a) the quadratic term is at least $\frac{1}{2}\lambda\|\mathbf{h}\|^2$ by Lemma 14.6.3, which beats a remainder eventually smaller than $\frac{1}{4}\lambda\|\mathbf{h}\|^2$; so the difference is positive. For (c) there is no bound in every direction, and none is needed: walk along a single line $\mathbf{h} = t\mathbf{u}$ where $q(\mathbf{u}) > 0$, and the whole thing is t^2 times something tending to $\frac{1}{2}q(\mathbf{u}) > 0$. One direction up, one direction down, and no extremum.

Proof. Write $\mathbf{H} = \mathbf{H}_f(\mathbf{a})$ and $q(\mathbf{h}) = \mathbf{h}^\top \mathbf{H}\mathbf{h}$. Since $\mathbf{a}$ is critical, $\nabla f(\mathbf{a}) = \mathbf{0}$, so (A2) reads

$$f(\mathbf{a}+\mathbf{h}) - f(\mathbf{a}) = \tfrac{1}{2}q(\mathbf{h}) + r(\mathbf{h}), \qquad \frac{r(\mathbf{h})}{\|\mathbf{h}\|^2} \to 0. \tag{$*$}$$

(a) Let $\lambda > 0$ be the smallest eigenvalue of $\mathbf{H}$, so that $q(\mathbf{h}) \geq \lambda\|\mathbf{h}\|^2$ by Lemma 14.6.3. Since $r(\mathbf{h})/\|\mathbf{h}\|^2 \to 0$, there is $\delta > 0$ with $\mathbf{a}+\mathbf{h} \in U$ and $|r(\mathbf{h})| \leq \frac{1}{4}\lambda\|\mathbf{h}\|^2$ whenever $0 < \|\mathbf{h}\| < \delta$. For such $\mathbf{h}$, by $(*)$,

$$f(\mathbf{a}+\mathbf{h}) - f(\mathbf{a}) \;\geq\; \tfrac{1}{2}\lambda\|\mathbf{h}\|^2 - \tfrac{1}{4}\lambda\|\mathbf{h}\|^2 = \tfrac{1}{4}\lambda\|\mathbf{h}\|^2 > 0.$$

So $f(\mathbf{a}+\mathbf{h}) > f(\mathbf{a})$ for every $\mathbf{h} \neq \mathbf{0}$ with $\|\mathbf{h}\| < \delta$, which is what "strict local minimum" says.

(b) The Hessian of $-f$ at $\mathbf{a}$ is $-\mathbf{H}$, whose inertia is $(n, 0, 0)$ because that of $\mathbf{H}$ is $(0, n, 0)$; so $-\mathbf{H}$ is positive definite by Theorem 14.6.1 (a). Part (a) applied to $-f$ makes $\mathbf{a}$ a strict local minimum of $-f$, that is a strict local maximum of f.

(c) By Theorem 14.6.1 (e) there are $\mathbf{u}, \mathbf{w} \in \mathbb{R}^n$ with $q(\mathbf{u}) > 0$ and $q(\mathbf{w}) < 0$; both are nonzero, since $q(\mathbf{0}) = 0$. Fix $t \neq 0$ small and put $\mathbf{h} = t\mathbf{u}$. Then $q(t\mathbf{u}) = t^2 q(\mathbf{u})$, so $(*)$ gives

$$\frac{f(\mathbf{a}+t\mathbf{u}) - f(\mathbf{a})}{t^2} = \tfrac{1}{2}q(\mathbf{u}) + \frac{r(t\mathbf{u})}{t^2}.$$

Now $r(t\mathbf{u})/t^2 = \|\mathbf{u}\|^2 \cdot r(t\mathbf{u})/\|t\mathbf{u}\|^2 \to 0$ as $t \to 0$, because $\|t\mathbf{u}\| \to 0$ and $\|\mathbf{u}\|^2$ is a constant. So for all small enough $t \neq 0$ the right-hand side is at least $\frac{1}{4}q(\mathbf{u}) > 0$, and hence $f(\mathbf{a}+t\mathbf{u}) > f(\mathbf{a})$. The same computation with $\mathbf{w}$ in place of $\mathbf{u}$ gives $f(\mathbf{a}+t\mathbf{w}) < f(\mathbf{a})$ for all small enough $t \neq 0$. Every neighborhood of $\mathbf{a}$ contains points of both kinds, so $\mathbf{a}$ is neither a local minimum nor a local maximum. This proves the theorem.

■

Cases (a)–(c) exhaust the nondegenerate ones: if $n_0 = 0$ then $n_+ + n_- = n$, so either $n_+ = n$, or $n_- = n$, or both are positive. What is left undecided is exactly $n_0 > 0$, and there the test is silent because it has to be.

Example. Two critical points of a cubic surface

Find the critical points of $f(x,y) = x^3 - 3xy + y^3$ on $\mathbb{R}^2$ and classify each one.

Solution. The partial derivatives are $f_x = 3x^2 - 3y$ and $f_y = -3x + 3y^2$. Setting both to zero gives $y = x^2$ and $x = y^2$, hence $x = x^4$, so $x(x^3 - 1) = 0$ and $x \in \{0, 1\}$. The real critical points are $(0,0)$ and $(1,1)$.

Differentiating again,

$$\mathbf{H}_f(x,y) = \begin{pmatrix} 6x & -3 \\ -3 & 6y \end{pmatrix}.$$

At $(0,0)$ the Hessian is $\left(\begin{smallmatrix} 0 & -3 \\ -3 & 0 \end{smallmatrix}\right)$, with $\det = -9 < 0$. A 2×2 symmetric matrix with negative determinant has eigenvalue product -9, hence one eigenvalue of each sign, so its inertia is $(1,1,0)$ by Theorem 14.5.3 and the form is indefinite. By Theorem 14.6.4 (c), $(0,0)$ is a saddle point. Concretely $f(t,t) = 2t^3 - 3t^2 < 0$ and $f(t,-t) = 3t^2 > 0$ for small $t > 0$.

At $(1,1)$ the Hessian is $\left(\begin{smallmatrix} 6 & -3 \\ -3 & 6 \end{smallmatrix}\right)$, whose leading principal minors are $6 > 0$ and $36 - 9 = 27 > 0$. By Proposition 14.6.2 (a) it is positive definite, so $(1,1)$ is a strict local minimum, with value $f(1,1) = -1$.

Example. One Hessian, three answers

The three functions

$$g_1 = x^2 + y^4, \qquad g_2 = x^2 - y^4, \qquad g_3 = x^2$$

all have $\nabla g_i(0,0) = \mathbf{0}$ and the same Hessian at the origin. What is that Hessian, and what happens at the origin in each case?

Solution. Every second derivative of y^4 vanishes at $y = 0$, so all three have

$$\mathbf{H}_{g_i}(0,0) = \begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix},$$

with inertia $(1,0,1)$: positive semidefinite, not definite, not indefinite. The test is silent, and it must be, because the three answers differ. For g_1 the origin is a strict local minimum, since $x^2 + y^4 > 0$ away from $\mathbf{0}$. For g_2 it is a saddle: $g_2(t,0) = t^2 > 0$ and $g_2(0,t) = -t^4 < 0$ for $t \neq 0$. For g_3 it is a minimum but not a strict one, since g_3 vanishes along the whole y-axis. The quadratic term decided nothing here; the quartic term decided everything.

Warning. **The test reads the signature, not the determinant.** In two variables the familiar rule is "$\det \mathbf{H} > 0$ and $f_{xx} > 0$ means a minimum", and the second half is not decoration: $\operatorname{diag}(-1,-1)$ has $\det = 1 > 0$, and $-x^2 - y^2$ has a strict local **maximum** at the origin. All $\det \mathbf{H} > 0$ says in two variables is that $n_0 = 0$ and n_- is even — definite, of a sign the determinant cannot see. In three variables it says even less: $\operatorname{diag}(1,-1,-1)$ has $\det = 1 > 0$ and is indefinite.

14.6.3 Exercises

14.6.3.1 A. Check your understanding

Exercise. A1

(a) Define what it means for a symmetric bilinear form on a real vector space to be **indefinite**.

(b) State the signature condition equivalent to negative semidefiniteness, and the one equivalent to positive definiteness.

(c) Determine whether the following statement is correct, and justify your answer: a symmetric form that is not positive semidefinite is indefinite.

(d) Name the two results of analysis that Theorem 14.6.4 assumes, and say which part of its proof uses each.

(e) A critical point has $n_0 = 2$ and $n_- = 0$. What does Theorem 14.6.4 conclude?

Solution.

(a) β is indefinite if there exist u and w in V with $q(u) > 0$ and $q(w) < 0$, where $q(v) = \beta(v, v)$.

(b) Negative semidefinite $\iff$ $n_+ = 0$; positive definite $\iff$ $(n_+, n_-, n_0) = (n, 0, 0)$, equivalently $n_+ = n$. Both are Theorem 14.6.1.

(c) Incorrect. The form $q(x_1, x_2) = -x_1^2$ has inertia $(0, 1, 1)$; it is not positive semidefinite, since $q(\mathbf{e}_1) = -1$, and not indefinite, since q is never positive. Failing (b) of Definition 14.6.1 means only $n_- > 0$; being indefinite needs $n_+ > 0$ as well.

(d) Symmetry of the mixed second partial derivatives, which makes the Hessian symmetric so that it has an inertia at all; and Taylor's theorem with a second-order remainder, which is the identity $(*)$ that every part of the proof manipulates.

(e) Nothing. Here $n_+ = n - 2$ and $n_- = 0$, so the Hessian is positive semidefinite but not definite and not indefinite, and none of (a), (b), (c) applies. Example (One Hessian, three answers) shows that all of "strict minimum", "nonstrict minimum", and "saddle" really do occur in this situation.

14.6.3.2 B. Practice

Exercise. B1: Classify by the signature

Determine the definiteness of each of the following real symmetric matrices, and give the inertia of each. Justify your answers.

$$\mathbf{A}_1 = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}, \quad \mathbf{A}_2 = \begin{pmatrix} -1 & 2 \\ 2 & -5 \end{pmatrix}, \quad \mathbf{A}_3 = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}.$$

Solution. For $\mathbf{A}_1$: $\det(\mathbf{A}_1)_1 = 2 > 0$ and $\det \mathbf{A}_1 = 4 - 1 = 3 > 0$, so $\mathbf{A}_1$ is positive definite by Proposition 14.6.2 (a), with inertia $(2, 0, 0)$.

For $\mathbf{A}_2$: $(-1)^1 \det(\mathbf{A}_2)_1 = 1 > 0$ and $(-1)^2 \det \mathbf{A}_2 = 5 - 4 = 1 > 0$, so $\mathbf{A}_2$ is negative definite

by Proposition 14.6.2 (b), with inertia $(0, 2, 0)$.

For $\mathbf{A}_3$: $\det \mathbf{A}_3 = 1-4 = -3 < 0$, so the two eigenvalues have product -3 and hence opposite signs. The inertia is $(1, 1, 0)$ by Theorem 14.5.3, and $\mathbf{A}_3$ is indefinite by Theorem 14.6.1 (e). Explicitly $q(1, 1) = 6 > 0$ and $q(1, -1) = -2 < 0$.

Exercise. B2: Critical points of a quartic

Find all critical points of $f(x, y) = x^4 + y^4 - 4xy$ and classify each with Theorem 14.6.4.

Solution. $f_x = 4x^3 - 4y$ and $f_y = 4y^3 - 4x$, so $y = x^3$ and $x = y^3 = x^9$. Then $x(x^8 - 1) = 0$, so $x \in \{0, 1, -1\}$ and the critical points are $(0, 0)$, $(1, 1)$, $(-1, -1)$. The Hessian is

$$\mathbf{H}_f(x, y) = \begin{pmatrix} 12x^2 & -4 \\ -4 & 12y^2 \end{pmatrix}.$$

At $(0, 0)$ it is $\left(\begin{smallmatrix} 0 & -4 \\ -4 & 0 \end{smallmatrix}\right)$, with determinant $-16 < 0$, hence inertia $(1, 1, 0)$ and a saddle point. At $(\pm 1, \pm 1)$ it is $\left(\begin{smallmatrix} 12 & -4 \\ -4 & 12 \end{smallmatrix}\right)$, with leading minors $12 > 0$ and $144 - 16 = 128 > 0$, hence positive definite: both are strict local minima, each of value $f = 1 + 1 - 4 = -2$.

14.6.3.3 C. Going deeper

Exercise. C1: The signature measures positive subspaces

Let β be a symmetric bilinear form on a real vector space V of dimension n, with inertia (n_+, n_-, n_0). Call a subspace $U \subseteq V$ **positive** if $q(v) > 0$ for every nonzero $v \in U$.

(a) Prove that V has a positive subspace of dimension n_+.

(b) Prove that every positive subspace has dimension at most n_+. *Hint: consider a subspace on which $q \le 0$ of the largest available dimension.*

(c) Hence deduce that n_+ equals the largest dimension of a positive subspace, and that this gives a second proof that n_+ does not depend on the orthogonal basis.

Solution. Fix an orthogonal basis $\mathcal{B} = (v_1, \dots, v_n)$ for β, ordered so that $d_i = q(v_i) > 0$ for $i \le n_+$, then negative, then zero.

(a) Put $U = \operatorname{span}(v_1, \dots, v_{n_+})$. For a nonzero $v = \sum_{i \le n_+} c_i v_i$, Eq. 14.4.1 gives $q(v) = \sum_{i \le n_+} d_i c_i^2 > 0$, since every $d_i > 0$ and some $c_i \neq 0$. So U is positive of dimension n_+.

(b) Put $U' = \operatorname{span}(v_{n_+ + 1}, \dots, v_n)$, of dimension $n - n_+$. For $v \in U'$, Eq. 14.4.1 gives $q(v) = \sum_{i > n_+} d_i c_i^2 \le 0$, every d_i there being ≤ 0. Let W be any positive subspace. Then $W \cap U' = \{0\}$, since a nonzero vector in the intersection would have $q > 0$ and $q \le 0$. By the dimension formula (Theorem 1.7.3),

$$\dim W + (n - n_+) = \dim(W + U') \le n,$$

so $\dim W \le n_+$.

(c) By (a) and (b) the largest dimension of a positive subspace is exactly n_+. That

description mentions only β and V, no basis, so n_+ is an invariant of β. Together with the invariance of $\operatorname{rank}\beta = n_+ + n_-$ under congruence (Theorem 14.2.3) this recovers Theorem 14.5.1.

Exercise. C2: A saddle the determinant hides

Let $f(x, y, z) = x^2 - y^2 - z^2 + xyz$.

(a) Verify that the origin is a critical point and compute $\mathbf{H}_f(\mathbf{0})$ and its determinant.

(b) Classify the origin, and explain why the sign of $\det \mathbf{H}_f(\mathbf{0})$ alone would have been useless.

(c) Prove that no twice continuously differentiable $g\colon \mathbb{R}^3 \to \mathbb{R}$ has a strict local maximum at a critical point $\mathbf{0}$ with $\det \mathbf{H}_g(\mathbf{0}) > 0$, and give a $g\colon \mathbb{R}^2 \to \mathbb{R}$ that does. Which feature of the dimension is responsible?

Solution.

(a) $f_x = 2x + yz$, $f_y = -2y + xz$, $f_z = -2z + xy$, all zero at $\mathbf{0}$. The second derivatives at $\mathbf{0}$ are $f_{xx} = 2$, $f_{yy} = f_{zz} = -2$, and $f_{xy} = z$, $f_{xz} = y$, $f_{yz} = x$, all zero at $\mathbf{0}$. So $\mathbf{H}_f(\mathbf{0}) = \operatorname{diag}(2, -2, -2)$, with $\det = 8 > 0$.

(b) The inertia is $(1, 2, 0)$, so the Hessian is indefinite and $\mathbf{0}$ is a saddle point by Theorem 14.6.4 (c). The determinant is the product of the eigenvalues (Corollary 12.1.5) and therefore records only the parity of n_-; here $\det > 0$ says n_- is even, which is equally true of the negative definite $\operatorname{diag}(-2, -2, -2)$ and of the positive definite $\mathbf{I}_3$. It cannot distinguish them.

(c) Suppose $\mathbf{0}$ is a local maximum of g and write $q(\mathbf{h}) = \mathbf{h}^{\top}\mathbf{H}_g(\mathbf{0})\mathbf{h}$. If $q(\mathbf{u}) > 0$ for some $\mathbf{u}$, the computation in the proof of Theorem 14.6.4 (c) gives $g(t\mathbf{u}) > g(\mathbf{0})$ for all small $t \neq 0$, contradicting the maximum. So $q(\mathbf{u}) \leq 0$ for every $\mathbf{u}$, that is $\mathbf{H}_g(\mathbf{0})$ is negative semidefinite and $n_+ = 0$ by Theorem 14.6.1 (d). If moreover $\det \mathbf{H}_g(\mathbf{0}) > 0$ then 0 is not an eigenvalue, so $n_0 = 0$ and the inertia is $(0, 3, 0)$. But then $\det \mathbf{H}_g(\mathbf{0}) = \lambda_1\lambda_2\lambda_3$ is a product of three negative numbers, hence negative — a contradiction. On $\mathbb{R}^2$ there is no obstruction: $g(x, y) = -x^2 - y^2$ has $\mathbf{H}_g(\mathbf{0}) = -2\mathbf{I}_2$, determinant $4 > 0$, and a strict local maximum at $\mathbf{0}$. The responsible feature is the parity of the dimension: the determinant is the product of n eigenvalues, so a negative definite Hessian has det of sign $(-1)^n$.

14.7 Where the Form Vanishes

Sections 4 and 5 diagonalized a symmetric form and counted the signs on the diagonal. Over $\mathbb{R}$ that is the whole story, but it hides the feature that makes indefinite forms geometrically different from inner products: there are nonzero vectors on which the form vanishes. This section takes those vectors seriously. They are not defects to be quarantined; they organize the space, because each one forces a two-dimensional piece of completely standard shape to split off. The result is a decomposition of any nondegenerate space into a pile of identical planes plus a remainder on which the form vanishes nowhere.

Throughout, F is a field with $\operatorname{char} F \neq 2$ (Definition 0.7.2), V is a finite-dimensional vector

space over F with $\dim V = n$, and β is a **symmetric** bilinear form on V (Definition 14.3.1) with quadratic form $q(v) = \beta(v, v)$ (Definition 14.3.2). Two vectors are orthogonal when β pairs them to 0 (Definition 14.4.1). For a subspace $U \subseteq V$ we write, as in Exercise (C2: Orthogonal complements for a form),

$$U^{\perp_\beta} = \{\, v \in V : \beta(u, v) = 0 \text{ for every } u \in U \,\},$$

and we write $V = U \perp W$ to say that $V = U \oplus W$ with every vector of U orthogonal to every vector of W — an **orthogonal direct sum**. The subscript β is dropped when only one form is in play.

14.7.1 An intuition to give up

Thirteen chapters have trained one reflex above all others: $\langle v, v\rangle = 0$ forces $v = 0$. It is how we proved that an isometry is injective, that an orthogonal list is independent, that a Gram matrix determines its vectors. It is also the first thing that goes.

Take $V = \mathbb{R}^2$ with

$$q(x_1, x_2) = x_1^2 - x_2^2, \qquad \mathbf{A} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

The vector $\mathbf{v} = (1, 1)$ has $q(\mathbf{v}) = 1 - 1 = 0$. It is not $\mathbf{0}$, and the form is not defective: $\det \mathbf{A} = -1 \neq 0$, so β is nondegenerate (Proposition 14.1.2). Indeed $\beta(\mathbf{v}, (1, -1)) = 1 - (1)(-1) = 2 \neq 0$, so $\mathbf{v}$ is perfectly visible to β — it is only invisible to *itself*. The line $\operatorname{span}(\mathbf{v})$ is a line of "length zero" inside a space where length is otherwise perfectly well behaved.

It gets worse, and the worse case is the one worth internalizing. Take $V = \mathbb{R}^4$ with

$$q(\mathbf{x}) = x_1^2 + x_2^2 - x_3^2 - x_4^2,$$

of signature $(2, 2)$ and therefore nondegenerate. Let $W = \operatorname{span}(\mathbf{e}_1 + \mathbf{e}_3,\ \mathbf{e}_2 + \mathbf{e}_4)$, a plane. For $\mathbf{w} = a(\mathbf{e}_1 + \mathbf{e}_3) + b(\mathbf{e}_2 + \mathbf{e}_4)$,

$$q(\mathbf{w}) = a^2 + b^2 - a^2 - b^2 = 0,$$

and in fact β vanishes on every pair from W, since $\beta(\mathbf{e}_1 + \mathbf{e}_3, \mathbf{e}_2 + \mathbf{e}_4) = 0$ too. So β restricted to W is the **zero form** on a two-dimensional subspace of a four-dimensional space on which β is nondegenerate.

Say that again, because it is the sentence an inner-product-trained reader will refuse. A nondegenerate form can vanish identically on a subspace of half the dimension of the whole space. Nondegeneracy is a statement about V; it says that no vector is orthogonal to *everything*. It says nothing whatsoever about what β does inside a subspace. The vectors of W above are orthogonal to each other and to themselves; they are not orthogonal to $\mathbf{e}_1 - \mathbf{e}_3$, and that is all nondegeneracy ever promised.

Once that is accepted, the rest of the section is straightforward. The vanishing vectors get a name, and then a structure theorem.

14.7.2 Isotropic vectors and totally isotropic subspaces

An isotropic vector is a nonzero vector that the form cannot tell from zero when it looks at it twice.

Definition 14.7.1. Isotropic and Anisotropic

Let β be a symmetric bilinear form on V with quadratic form q.

(a) A vector $v \in V$ is **isotropic** if $v \neq 0$ and $q(v) = 0$, and **anisotropic** if $q(v) \neq 0$.

(b) The form β is **anisotropic** if it has no isotropic vectors: that is, if $q(v) = 0$ implies $v = 0$. Otherwise β is **isotropic**.

Both clauses hinge on small words. In (a) the vector is required to be **nonzero**, so 0 is neither isotropic nor anisotropic; the zero vector always has $q(0) = 0$ and carries no information. In (b), "anisotropic" is a property of the **form on a particular space**. Anisotropy passes down to subspaces, since a subspace has fewer vectors to test; isotropy does not, and a form isotropic on V is very often anisotropic on a subspace. The word comes from *iso* (equal) and *tropos* (direction): a vector along which the form does not distinguish one direction from another.

Examples, over several fields.

- **A positive definite form has no isotropic vectors.** If $q(v) > 0$ for $v \neq 0$ then $q(v) = 0$ forces $v = 0$. So every inner product is anisotropic, which is exactly the reflex described above. The same holds for negative definite forms.
- **Over $\mathbb{R}$, signature $(1,1)$.** For $q = x_1^2 - x_2^2$ the isotropic vectors are the nonzero multiples of $(1,1)$ and of $(1,-1)$, since $q = (x_1 - x_2)(x_1 + x_2)$. Two lines, and nothing else.
- **Over $\mathbb{C}$.** For $q = x_1^2 + x_2^2$ on $\mathbb{C}^2$ the factorization $q = (x_1 + ix_2)(x_1 - ix_2)$ is available, so $(1,i)$ and $(1,-i)$ are isotropic. Over $\mathbb{R}$ the same q is positive definite. **Isotropy depends on the field**, not only on the polynomial.
- **Over $\mathbb{F}_3$ and $\mathbb{F}_5$.** For $q = x_1^2 + x_2^2$ the squares in $\mathbb{F}_3$ are 0 and 1, so $q(\mathbf{x}) = 0$ needs $x_1 = x_2 = 0$ and the form is anisotropic. In $\mathbb{F}_5$, however, $1^2 + 2^2 = 5 = 0$, so $(1,2)$ is isotropic. The same formula, two finite fields, two answers.
- **The degenerate case.** Every nonzero vector of the radical is isotropic, since $v \in \operatorname{rad}(\beta)$ gives $\beta(u,v) = 0$ for every u, in particular for $u = v$. So a degenerate form is automatically isotropic. The converse fails, and the $\mathbb{R}^2$ example above is the witness.

When a whole subspace consists of isotropic vectors, more is true than the definition says, and it deserves its own name.

Definition 14.7.2. Totally Isotropic Subspace

A subspace $W \subseteq V$ is **totally isotropic** for β if $\beta(u,w) = 0$ for **all** $u, w \in W$; equivalently, if $W \subseteq W^{\perp_\beta}$. The zero subspace is totally isotropic, trivially.

The two formulations agree because $W \subseteq W^{\perp_\beta}$ says exactly that every $w \in W$ is orthogonal to every $u \in W$. A third formulation is available in characteristic $\neq 2$ and is the one to use in practice: W is totally isotropic if and only if $q(w) = 0$ for every $w \in W$. One direction is immediate; the other is Theorem 14.3.5 applied to the restricted form, which recovers β on W from q on W and so makes it zero. In the $\mathbb{R}^4$ example above we checked exactly this: q vanished on W, and therefore so did β.

Before the first theorem, one dimension count has to be on the table, because almost everything in this section and the next is a step away from it. Exercise (C2: Orthogonal complements for a form) set it as an exercise in Section 1; here it is with a proof, so that the results below rest on it rather than on the exercise.

Lemma 14.7.1. Dimension of an Orthogonal Complement

Let β be a **nondegenerate** bilinear form on V with $\dim V = n$, and let $U \subseteq V$ be a subspace. Then $U^{\perp_\beta}$ is a subspace of V, and

$$\dim U^{\perp_\beta} = n - \dim U.$$

Idea. Freezing the second slot of β turns it into a map $V \to V^*$, and restricting a functional to U turns V^* into U^*. The composite has kernel exactly $U^{\perp_\beta}$, and nondegeneracy makes both maps surjective, so Rank–Nullity converts "onto U^*" into the count.

Proof. Let $R\colon V \to V^*$ send v to the functional $\beta(\cdot, v)$ — the map called R_β in Section 1 — and let $\rho\colon V^* \to U^*$ be restriction to U. Both are linear, so $S \coloneqq \rho \circ R\colon V \to U^*$ is linear, and $S(v) = \beta(\cdot, v)\big|_U$. By the definition of $U^{\perp_\beta}$, a vector v lies in $U^{\perp_\beta}$ exactly when $\beta(u, v) = 0$ for every $u \in U$, that is, exactly when $S(v) = 0$. Hence $U^{\perp_\beta} = \ker S$, which is a subspace.

Since β is nondegenerate, R is an isomorphism (Proposition 14.1.2 (c)), and ρ is surjective (Theorem 5.5.1). A composite of two surjections is a surjection, so $\operatorname{rank} S = \dim U^* = \dim U$ (Corollary 5.1.4). By Rank–Nullity (Theorem 2.3.1),

$$\dim U^{\perp_\beta} = \dim \ker S = n - \operatorname{rank} S = n - \dim U,$$

as claimed. ∎

Totally isotropic subspaces cannot be too large.

Proposition 14.7.2. Half the Dimension, at Most

Let β be a nondegenerate symmetric form on V with $\dim V = n$, and let $W \subseteq V$ be totally isotropic. Then $\dim W \leq n/2$.

Proof. Since β is nondegenerate, $\dim W^{\perp_\beta} = n - \dim W$ by Lemma 14.7.1. Since W is totally isotropic, $W \subseteq W^{\perp_\beta}$, so $\dim W \leq n - \dim W$, which is the claim. ∎

Warning. **An isotropic vector is not a zero vector, and a form that vanishes on a subspace need not be degenerate.** These are two different errors and both are common. The first confuses $q(v) = 0$ with $v = 0$; over $\mathbb{R}^2$ with $q = x_1^2 - x_2^2$ the vector $(1, 1)$ separates them. The second confuses "β vanishes on $W \times W$" with "β vanishes on $\{v\} \times V$". Degeneracy is about the **second**: the radical consists of vectors killed by pairing with *everything*. The plane W of the $\mathbb{R}^4$ example carries the zero form and contains no radical vector at all.

Quick check. Is the plane $\operatorname{span}(\mathbf{e}_1, \mathbf{e}_2) \subseteq \mathbb{R}^4$ totally isotropic for $q(\mathbf{x}) = x_1^2 + x_2^2 - x_3^2 - x_4^2$? What about $\operatorname{span}(\mathbf{e}_1 + \mathbf{e}_3, \mathbf{e}_1 - \mathbf{e}_3)$?

Solution. No to both. On $\operatorname{span}(\mathbf{e}_1, \mathbf{e}_2)$ the form restricts to $x_1^2 + x_2^2$, which is positive definite, so that plane is as far from totally isotropic as possible: it is anisotropic. The second plane is $\operatorname{span}(\mathbf{e}_1, \mathbf{e}_3)$, on which q restricts to $x_1^2 - x_3^2$; its two spanning vectors $\mathbf{e}_1 \pm \mathbf{e}_3$ are indeed isotropic, but $\beta(\mathbf{e}_1 + \mathbf{e}_3, \mathbf{e}_1 - \mathbf{e}_3) = 1 - (-1) = 2 \neq 0$, so the plane is not totally isotropic. Containing isotropic vectors is not the same as being totally isotropic.

14.7.3 Splitting off a nondegenerate piece

Everything below rests on one mechanism: if a subspace carries a nondegenerate restriction, it splits off orthogonally and the rest of the space is again nondegenerate. That is the analogue of "restrict to an invariant complement", and it is what makes induction possible.

Proposition 14.7.3. A Nondegenerate Subspace Splits Off

Let β be a nondegenerate symmetric form on V with $\dim V = n$, and let $U \subseteq V$ be a subspace such that $\beta|_{U \times U}$ is nondegenerate. Then

$$V = U \perp U^{\perp_\beta},$$

and β restricted to $U^{\perp_\beta}$ is nondegenerate.

Idea. Two things to check and one dimension count. The intersection $U \cap U^{\perp_\beta}$ is precisely the radical of the restricted form, which the hypothesis kills; the dimension of $U^{\perp_\beta}$ is $n - \dim U$ because the global form is nondegenerate; and a vector of $U^{\perp_\beta}$ orthogonal to $U^{\perp_\beta}$ is orthogonal to all of V, hence zero.

Proof. Write $U^\perp = U^{\perp_\beta}$. A vector $v \in U \cap U^\perp$ lies in U and satisfies $\beta(u, v) = 0$ for every $u \in U$, so $v \in \operatorname{rad}(\beta|_{U \times U})$, which is $\{0\}$ by hypothesis (Definition 14.1.4). Hence $U \cap U^\perp = \{0\}$ and the sum $U + U^\perp$ is direct (Theorem 1.7.4).

Since β is nondegenerate on V, Lemma 14.7.1 gives $\dim U^\perp = n - \dim U$. So $\dim(U \oplus U^\perp) = n$, and a subspace of V of dimension n is V (Theorem 1.6.8). The sum is orthogonal by the definition of $U^\perp$ together with the symmetry of β, so $V = U \perp U^\perp$.

Finally let $w \in U^\perp$ satisfy $\beta(x, w) = 0$ for every $x \in U^\perp$. For $u \in U$ we also have $\beta(u, w) = 0$, since $w \in U^\perp$. Every $v \in V$ is $u + x$ with $u \in U$ and $x \in U^\perp$, so $\beta(v, w) = 0$ for every $v \in V$, that is $w \in \operatorname{rad}(\beta) = \{0\}$. This proves the proposition.

■

14.7.4 Hyperbolic planes

Which two-dimensional forms are isotropic? Over $\mathbb{R}$ with signature $(1,1)$ we found two isotropic lines, and the example $q = x_1^2 - x_2^2$ can be rewritten by the substitution $u = x_1 + x_2$, $w = (x_1 - x_2)/2$ as $q = 2uw$ — the cross term alone, with no squares at all. That shape is the model, and it is the same over every field.

Definition 14.7.3. Hyperbolic Pair, Hyperbolic Plane

Let β be a symmetric bilinear form on V. A **hyperbolic pair** is an ordered pair (e, f) of vectors of V with

$$q(e) = 0, \qquad q(f) = 0, \qquad \beta(e, f) = 1.$$

The subspace $H = \operatorname{span}(e, f)$ is then called a **hyperbolic plane**, and (e, f) is a **hyperbolic basis** of it.

Four things follow at once. First, e and f are linearly independent: if $f = ce$ then $\beta(e, f) = c\,q(e) = 0 \neq 1$. So H really is a plane, $\dim H = 2$. Second, in the basis (e, f) the Gram matrix of $\beta|_H$ is

$$[\beta|_H]_{(e,f)}^{(e,f)} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

whose determinant is $-1 \neq 0$, so $\beta|_H$ is **nondegenerate** (Proposition 14.1.2) and Proposition 14.7.3 applies to it. Third, $q(ae + bf) = 2ab$, so on a hyperbolic plane q takes **every** value in F: given $c \in F$, take $a = c/2$ and $b = 1$. Fourth, over $\mathbb{R}$ the Gram matrix has eigenvalues 1 and -1, so a hyperbolic plane has signature $(1,1)$ and is the unique real example up to congruence.

Now the workhorse of this section and the next. It says that one isotropic vector is enough to manufacture a whole hyperbolic plane around it.

Lemma 14.7.4. Every Isotropic Vector Lies in a Hyperbolic Plane

Let $\operatorname{char} F \neq 2$, let β be a **nondegenerate** symmetric bilinear form on V, and let $e \in V$ be isotropic. Then there is $f \in V$ such that (e, f) is a hyperbolic pair. In particular e lies in a hyperbolic plane $H = \operatorname{span}(e, f)$, and $\beta|_H$ is nondegenerate.

Idea. Nondegeneracy hands us some u with $\beta(e, u) \neq 0$, and rescaling makes $\beta(e, u) = 1$. Only one condition is still missing, $q(u) = 0$, and u can be corrected without disturbing what has been achieved: adding a multiple of e changes $q(u)$ but not $\beta(e, u)$, because $q(e) = 0$. Solving for the multiple is one line, and it divides by 2 — which is where the characteristic hypothesis is spent, for the third time in this chapter.

Proof. Since $e \neq 0$ and β is nondegenerate, there is $u_0 \in V$ with $c = \beta(e, u_0) \neq 0$ (Definition 14.1.4). Put $u = c^{-1}u_0$, so that

$$\beta(e, u) = 1.$$

Now set

$$f = u - \tfrac{1}{2}q(u)\,e,$$

which is legitimate because $\operatorname{char} F \neq 2$ makes $2 = 1 + 1$ invertible. Then, using bilinearity, symmetry, and $q(e) = 0$,

$$\beta(e, f) = \beta(e, u) - \tfrac{1}{2} q(u)\, q(e) = 1 - 0 = 1,$$

and

$$\begin{aligned} q(f) &= q(u) - 2 \cdot \tfrac{1}{2} q(u)\, \beta(u, e) + \tfrac{1}{4} q(u)^2\, q(e) \\ &= q(u) - q(u) + 0 = 0. \end{aligned}$$

Together with $q(e) = 0$, this shows that (e, f) is a hyperbolic pair. As noted after Definition 14.7.3, e and f are independent, so $H = \operatorname{span}(e, f)$ is a plane, and the Gram matrix $\left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$ is invertible, so $\beta|_H$ is nondegenerate. This proves the lemma.

■

Remark. Nondegeneracy of β is not decoration here. On $\mathbb{R}^2$ with $q(\mathbf{x}) = x_1^2$, the vector $\mathbf{e}_2$ is isotropic and lies in no hyperbolic plane, because $\beta(\mathbf{e}_2, \mathbf{v}) = 0$ for every $\mathbf{v}$ and a hyperbolic pair needs $\beta(\mathbf{e}, \mathbf{f}) = 1$. What fails is the very first line of the proof.

14.7.5 The Witt decomposition

Peeling off a hyperbolic plane and repeating is now a two-line induction, and it is the structural theorem of the section: it does for indefinite forms what "peel off an eigenvector" did for operators in Chapter 9.

Theorem 14.7.5. Witt Decomposition

Let $\operatorname{char} F \neq 2$ and let β be a nondegenerate symmetric bilinear form on a finite-dimensional V. Then there are an integer $m \geq 0$, hyperbolic planes $H_1, \dots, H_m \subseteq V$, and a subspace $V_0 \subseteq V$ such that

$$V = H_1 \perp H_2 \perp \cdots \perp H_m \perp V_0$$

and β restricted to V_0 is **anisotropic**. In particular $\dim V = 2m + \dim V_0$.

Idea. Induction on $\dim V$. Either β is anisotropic, in which case take $m = 0$ and $V_0 = V$ and stop; or there is an isotropic vector, in which case Lemma 14.7.4 wraps a hyperbolic plane H_1 around it, Proposition 14.7.3 peels H_1 off with a nondegenerate remainder $H_1^{\perp}$ of dimension $\dim V - 2$, and the induction hypothesis finishes the job there.

Proof. Induct on $\dim V$. If $\dim V = 0$, take $m = 0$ and $V_0 = \{0\}$, on which the form is anisotropic vacuously. Let $\dim V \geq 1$ and suppose the statement holds for every nondegenerate symmetric form on a space of smaller dimension.

If β is anisotropic, take $m = 0$ and $V_0 = V$.

Otherwise β has an isotropic vector e. By Lemma 14.7.4 there is f with (e, f) a hyperbolic pair, and $H_1 = \operatorname{span}(e, f)$ is a hyperbolic plane with $\beta|_{H_1}$ nondegenerate. By Proposi-

tion 14.7.3,

$$V = H_1 \perp H_1^{\perp_\beta},$$

and β restricted to $V' = H_1^{\perp_\beta}$ is nondegenerate, with $\dim V' = \dim V - 2 < \dim V$. The induction hypothesis applied to $(V', \beta|_{V'})$ gives hyperbolic planes $H_2, \dots, H_m \subseteq V'$ and an anisotropic $V_0 \subseteq V'$ with $V' = H_2 \perp \cdots \perp H_m \perp V_0$. Each H_i with $i \geq 2$ and V_0 lie inside $V' = H_1^{\perp_\beta}$, so all the summands are pairwise orthogonal, and

$$V = H_1 \perp H_2 \perp \cdots \perp H_m \perp V_0.$$

Counting dimensions, $\dim V = 2m + \dim V_0$. This completes the induction.

■

Two questions are now unavoidable: is m the same for every such decomposition, and is V_0 determined? The first has an answer here, and it comes from a quantity that mentions no decomposition at all.

Definition 14.7.4. Witt Index

Let β be a nondegenerate symmetric bilinear form on a finite-dimensional V. The **Witt index** of β is

$$\operatorname{ind}(\beta) := \max\{\, \dim W : W \subseteq V \text{ totally isotropic} \,\},$$

the largest dimension of a totally isotropic subspace. By Proposition 14.7.2 it satisfies $\operatorname{ind}(\beta) \leq n/2$, and it is 0 exactly when β is anisotropic.

The maximum exists because the dimensions in question form a nonempty finite set of integers: $W = \{0\}$ is always totally isotropic, and all dimensions are at most n. Nothing in the definition refers to a basis or a decomposition, so the Witt index is an invariant of (V, β) by construction. It is also exactly the number of planes.

Proposition 14.7.6. The Number of Hyperbolic Planes Is the Witt Index

In any decomposition $V = H_1 \perp \cdots \perp H_m \perp V_0$ as in Theorem 14.7.5, $m = \operatorname{ind}(\beta)$. In particular m is the same for every such decomposition, and so is $\dim V_0 = n - 2m$.

Idea. One inequality is a construction: the first vectors $e_1, \dots, e_m$ of the m hyperbolic bases span a totally isotropic subspace of dimension m, since each is isotropic and different planes are orthogonal. The other inequality is an induction that trades one plane for one dimension: given a totally isotropic W, the vectors of W orthogonal to e_1 project into $H_1^{\perp}$ as a totally isotropic subspace of dimension at least $\dim W - 1$. Repeating m times lands inside the anisotropic V_0, where the only totally isotropic subspace is $\{0\}$.

Proof. Let (e_i, f_i) be a hyperbolic basis of H_i for each i.

Step 1: $\operatorname{ind}(\beta) \geq m$. Put $W = \operatorname{span}(e_1, \dots, e_m)$. The e_i lie in different summands of a direct sum and are nonzero, so they are independent and $\dim W = m$. For $i \neq j$ we have $\beta(e_i, e_j) = 0$ because $H_i \perp H_j$, and $\beta(e_i, e_i) = q(e_i) = 0$. By bilinearity β vanishes on

$W \times W$, so W is totally isotropic of dimension m.

Step 2: a plane can be traded for a dimension.

Claim. Let β be nondegenerate on V, let $H = \operatorname{span}(e, f) \subseteq V$ be a hyperbolic plane, let $V' = H^{\perp_\beta}$, and let $W \subseteq V$ be totally isotropic. Then V' contains a totally isotropic subspace of dimension at least $\dim W - 1$.

Proof. By Proposition 14.7.3, $V = H \perp V'$; let $P\colon V \to V'$ be the projection with kernel H, so that $v - P(v) \in H$ for every v. Put

$$W_1 = \{\, w \in W : \beta(e, w) = 0 \,\},$$

the kernel of a linear functional on W, so $\dim W_1 \geq \dim W - 1$.

For $w \in V$ write $w - P(w) = ae + bf$. Since $e \in H$ is orthogonal to V',

$$\beta(e, w) = \beta(e, ae + bf) = a\, q(e) + b = b.$$

So $w \in W_1$ forces $w - P(w) \in \operatorname{span}(e)$. Consequently, for $w, w' \in W_1$ the vectors $w - P(w)$ and $w' - P(w')$ both lie in $\operatorname{span}(e)$, where β vanishes because $q(e) = 0$; hence

$$\beta(P(w), P(w')) = \beta(w, w') - \beta(w - P(w),\, w' - P(w')) = 0 - 0 = 0,$$

the first equality using the orthogonality of H and V'. So $P(W_1)$ is a totally isotropic subspace of V'.

It remains to bound its dimension. By Rank–Nullity (Theorem 2.3.1) applied to $P|_{W_1}$, whose kernel is $W_1 \cap H$,

$$\dim P(W_1) = \dim W_1 - \dim(W_1 \cap H).$$

If $W_1 \cap H = \{0\}$, then $\dim P(W_1) = \dim W_1 \geq \dim W - 1$ and we are done. Otherwise pick $x \neq 0$ in $W_1 \cap H$. Then $P(x) = 0$, so $x = x - P(x) \in \operatorname{span}(e)$ by the paragraph above, whence $e \in W$. But then $\beta(e, w) = 0$ for every $w \in W$, because W is totally isotropic and $e \in W$; so $W_1 = W$. Moreover every $x \in W_1 \cap H$ lies in $\operatorname{span}(e)$ by the same argument, so $W_1 \cap H = \operatorname{span}(e)$ has dimension 1, and $\dim P(W_1) = \dim W - 1$. Either way the claim holds.

■

Step 3: $\operatorname{ind}(\beta) \leq m$. Let $W \subseteq V$ be totally isotropic. Apply Step 2 to $H = H_1$: since $H_1^{\perp_\beta} = H_2 \perp \cdots \perp H_m \perp V_0$ carries a nondegenerate form, we obtain a totally isotropic subspace of it of dimension at least $\dim W - 1$. Repeating with $H_2, \ldots, H_m$ in turn produces a totally isotropic subspace of V_0 of dimension at least $\dim W - m$. But $\beta|_{V_0}$ is anisotropic, so a nonzero vector of V_0 has $q \neq 0$ and the only totally isotropic subspace of V_0 is $\{0\}$. Hence $\dim W - m \leq 0$.

Steps 1 and 3 give $\operatorname{ind}(\beta) = m$, and $\dim V_0 = n - 2m$ follows by counting. This proves the proposition.

■

So the number of planes is pinned down. What is not yet pinned down is V_0 itself: different decompositions produce different subspaces, and the claim one wants is that they all carry *the same form* in a suitable sense. Making that precise needs the word "isometry" for forms, and

proving it needs Witt's theorems; both are the business of the next section.

14.7.6 The two familiar fields

Over $\mathbb{R}$ and $\mathbb{C}$ everything can be computed, because the classification is already known.

Proposition 14.7.7. The Witt Index over the Reals and the Complexes

(a) Let β be a nondegenerate symmetric form on a real V with signature (n_+, n_-) (Definition 14.5.1), so $n_+ + n_- = n$. Then $\operatorname{ind}(\beta) = \min(n_+, n_-)$, and the anisotropic part has dimension $|n_+ - n_-|$ and is definite.

(b) Let β be a nondegenerate symmetric form on a complex V with $\dim V = n$. Then $\operatorname{ind}(\beta) = \lfloor n/2 \rfloor$. In particular every such form with $n \geq 2$ has an isotropic vector, and the anisotropic part has dimension 0 or 1.

Proof.

(a) Put $k = \min(n_+, n_-)$. By Theorem 14.4.2 and the rescaling that produced Eq. 14.5.1, together with $n_0 = 0$, there is an orthogonal basis $(v_1, \dots, v_n)$ with $q(v_i) = 1$ for $i \leq n_+$, $q(v_i) = -1$ for $n_+ < i \leq n$, and distinct basis vectors orthogonal. For $1 \leq i \leq k$ set

$$e_i = v_i + v_{n_+ + i}, \qquad f_i = \tfrac{1}{2}(v_i - v_{n_+ + i}).$$

Then $q(e_i) = 1 + (-1) = 0$ and $q(f_i) = \tfrac{1}{4}(1 + (-1)) = 0$, while

$$\beta(e_i, f_i) = \tfrac{1}{2}(q(v_i) - q(v_{n_+ + i})) = \tfrac{1}{2}(1 + 1) = 1,$$

so (e_i, f_i) is a hyperbolic pair and $H_i = \operatorname{span}(e_i, f_i)$ a hyperbolic plane. These planes are pairwise orthogonal, since for $i \neq j$ their spanning vectors involve disjoint sets of the v's. Let V_0 be the span of the remaining $|n_+ - n_-|$ basis vectors, all of the same sign; then $V = H_1 \perp \cdots \perp H_k \perp V_0$, and $\beta|_{V_0}$ is definite, hence anisotropic. By Proposition 14.7.6, $\operatorname{ind}(\beta) = k$.

(b) By Corollary 14.4.4 and nondegeneracy, there is a basis $(v_1, \dots, v_n)$ with $q(v_j) = 1$ for every j and distinct basis vectors orthogonal. Write $n = 2k + r$ with $r \in \{0, 1\}$, and let i denote the imaginary unit. For $1 \leq j \leq k$ set

$$e_j = v_{2j-1} + i\, v_{2j}, \qquad f_j = \tfrac{1}{2}(v_{2j-1} - i\, v_{2j}).$$

Then $q(e_j) = 1 + i^2 = 0$ and likewise $q(f_j) = \tfrac{1}{4}(1 + i^2) = 0$, while $\beta(e_j, f_j) = \tfrac{1}{2}(1 + 1) = 1$. As in (a) these span k pairwise orthogonal hyperbolic planes, and what is left is the span of v_n when $r = 1$, on which $q(c v_n) = c^2 \neq 0$ for $c \neq 0$, and $\{0\}$ when $r = 0$. So $\operatorname{ind}(\beta) = k = \lfloor n/2 \rfloor$ by Proposition 14.7.6. This proves the proposition.

■

Two readings are worth recording. Over $\mathbb{R}$ the Witt index is the smaller of the two counts of signs: a definite form has index 0, and a form of signature (m, m) has the largest possible index $m = n/2$ — the $\mathbb{R}^4$ example that opened the section, with $n_+ = n_- = 2$, sits exactly there. Over $\mathbb{C}$ the index is as large as it can be for every nondegenerate form, because -1 is a square and so a sum of two squares always factors; the only anisotropic complex spaces are $\{0\}$ and the lines.

Quick check. A nondegenerate real form has signature $(3, 1)$. What is its Witt index, and what is the anisotropic part?

Solution. The index is $\min(3, 1) = 1$, so exactly one hyperbolic plane splits off and $\dim V_0 = 4 - 2 = 2$. By Proposition 14.7.7 (a), V_0 carries a definite form of dimension $|3 - 1| = 2$, positive definite here since the positive signs are the ones in surplus. Concretely $\mathbb{R}^{3,1}$ with $q = x_1^2 + x_2^2 + x_3^2 - x_4^2$ splits as one hyperbolic plane, spanned by $\mathbf{e}_3 \pm \mathbf{e}_4$ after rescaling, orthogonal to the positive definite plane $\operatorname{span}(\mathbf{e}_1, \mathbf{e}_2)$. The largest totally isotropic subspace is a line, and the lines of this kind sweep out the light cone that Section 11 studies.

14.7.7 Exercises

14.7.7.1 A. Check your understanding

Exercise. A1

(a) Define **isotropic vector**, **totally isotropic subspace**, and **hyperbolic pair**.

(b) Determine whether the following statement is correct, and justify your answer: a nondegenerate form has no isotropic vectors.

(c) What is the Witt index of a positive definite form on a space of dimension n? Of a form of signature $(4, 4)$?

(d) State the Gram matrix of a hyperbolic plane in a hyperbolic basis, and say why it shows the restricted form is nondegenerate.

(e) Which hypothesis of Lemma 14.7.4 fails for the form $q(\mathbf{x}) = x_1^2$ on $\mathbb{R}^2$ at the isotropic vector $\mathbf{e}_2$?

Solution.

(a) A vector v is isotropic if $v \neq 0$ and $q(v) = 0$. A subspace W is totally isotropic if $\beta(u, w) = 0$ for all $u, w \in W$. A hyperbolic pair is a pair (e, f) with $q(e) = q(f) = 0$ and $\beta(e, f) = 1$.

(b) Incorrect. Nondegeneracy says $\operatorname{rad}(\beta) = \{0\}$, a statement about pairing with **every** vector. On $\mathbb{R}^2$ with $q = x_1^2 - x_2^2$ the form is nondegenerate, since its Gram matrix has determinant -1, and $(1, 1)$ is isotropic.

(c) 0 for a positive definite form, since it is anisotropic, so the only totally isotropic subspace is $\{0\}$. For signature $(4, 4)$ the index is $\min(4, 4) = 4$ by Proposition 14.7.7 (a), the largest value Proposition 14.7.2 permits in dimension 8.

(d) It is $\left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$, of determinant $-1 \neq 0$, and a form is nondegenerate exactly when its Gram matrix is invertible (Proposition 14.1.2).

(e) Nondegeneracy of β. Here $\mathbf{e}_2 \in \operatorname{rad}(\beta)$, so there is no $\mathbf{u}$ with $\beta(\mathbf{e}_2, \mathbf{u}) \neq 0$ and the first line of the proof has nothing to choose.

14.7.7.2 B. Practice

Exercise. B1: Find the isotropic vectors

For each of the following real quadratic forms, find all isotropic vectors, give the signature, and state the Witt index.

(a) $q(x_1, x_2) = x_1^2 + 2x_1x_2$ on $\mathbb{R}^2$.

(b) $q(x_1, x_2, x_3) = x_1x_2$ on $\mathbb{R}^3$. *(Is this form nondegenerate?)*

(c) $q(x_1, x_2, x_3) = x_1^2 + x_2^2 + x_3^2$ on $\mathbb{R}^3$.

Solution.

(a) $q = x_1(x_1+2x_2)$, so $q = 0$ exactly on the two lines $x_1 = 0$ and $x_1 = -2x_2$. The isotropic vectors are the nonzero multiples of $(0, 1)$ and of $(-2, 1)$. The Gram matrix is $\left(\begin{smallmatrix} 1 & 1 \\ 1 & 0 \end{smallmatrix}\right)$, of determinant $-1 < 0$, so the eigenvalues have opposite signs and the signature is $(1, 1)$. The Witt index is $\min(1, 1) = 1$; indeed $(0, 1)$ is isotropic and by Lemma 14.7.4 the whole space is one hyperbolic plane.

(b) The Gram matrix is $\left(\begin{smallmatrix} 0 & 1/2 & 0 \\ 1/2 & 0 & 0 \\ 0 & 0 & 0 \end{smallmatrix}\right)$, of rank 2, so the form is **degenerate** with $\operatorname{rad}(\beta) = \operatorname{span}(\mathbf{e}_3)$ and the notion of Witt index does not apply as defined. The isotropic vectors are those with $x_1x_2 = 0$, namely the union of the two planes $x_1 = 0$ and $x_2 = 0$, minus $\mathbf{0}$. On the nondegenerate part $\operatorname{span}(\mathbf{e}_1, \mathbf{e}_2)$ the signature is $(1, 1)$ and the index is 1.

(c) There are none: q is positive definite, so $q(\mathbf{x}) = 0$ forces $\mathbf{x} = \mathbf{0}$. The signature is $(3, 0)$ and the Witt index is 0.

Exercise. B2: Build a hyperbolic basis

(a) For $q(x_1, x_2) = x_1^2 + x_2^2$ on $\mathbb{C}^2$, find a hyperbolic pair.

(b) For $q(x_1, x_2, x_3) = x_1^2 + x_2^2 - x_3^2$ on $\mathbb{R}^3$, find a hyperbolic pair $(\mathbf{e}, \mathbf{f})$ with $\mathbf{e} = \mathbf{e}_1 + \mathbf{e}_3$, and identify the anisotropic part $\operatorname{span}(\mathbf{e}, \mathbf{f})^{\perp_\beta}$.

Solution.

(a) Take $\mathbf{e} = (1, i)$, so $q(\mathbf{e}) = 1+i^2 = 0$. The vector $\mathbf{u} = (1, -i)$ has $\beta(\mathbf{e}, \mathbf{u}) = 1\cdot 1+i\cdot(-i) = 2$, so rescale: $\mathbf{f} = \frac{1}{2}(1, -i)$ gives $\beta(\mathbf{e}, \mathbf{f}) = 1$ and $q(\mathbf{f}) = \frac{1}{4}(1+i^2) = 0$. So $((1, i), \frac{1}{2}(1, -i))$ is a hyperbolic pair, and $\mathbb{C}^2$ is a single hyperbolic plane.

(b) $q(\mathbf{e}) = 1 - 1 = 0$, so $\mathbf{e}$ is isotropic. Follow the proof of Lemma 14.7.4: $\mathbf{u}_0 = \mathbf{e}_1$ gives $\beta(\mathbf{e}, \mathbf{e}_1) = 1$, so take $\mathbf{u} = \mathbf{e}_1$, and $q(\mathbf{u}) = 1$, whence

$$\mathbf{f} = \mathbf{e}_1 - \tfrac{1}{2}(\mathbf{e}_1 + \mathbf{e}_3) = \tfrac{1}{2}(\mathbf{e}_1 - \mathbf{e}_3).$$

Check: $q(\mathbf{f}) = \frac{1}{4}(1 - 1) = 0$ and $\beta(\mathbf{e}, \mathbf{f}) = \frac{1}{2}(1 + 1) = 1$. For the complement, $\mathbf{v} = (v_1, v_2, v_3)$ is orthogonal to both $\mathbf{e}$ and $\mathbf{f}$ exactly when $v_1 - v_3 = 0$ and $v_1 + v_3 = 0$, that is $v_1 = v_3 = 0$. So $\operatorname{span}(\mathbf{e}, \mathbf{f})^{\perp_\beta} = \operatorname{span}(\mathbf{e}_2)$, on which $q = x_2^2$ is positive definite. This is the decomposition of Theorem 14.7.5 with $m = 1$, matching $\min(2, 1) = 1$.

Exercise. B3: A totally isotropic subspace of the largest size

Let $q(\mathbf{x}) = x_1x_4 + x_2x_3$ on $\mathbb{R}^4$.

(a) Write down the Gram matrix and verify that β is nondegenerate.

(b) Exhibit a totally isotropic subspace of dimension 2, and explain why none of dimension 3 exists.

(c) Determine the signature of q and confirm the Witt index by Proposition 14.7.7 (a).

Solution.

(a) Halving the cross-term coefficients (Eq. 14.3.1),

$$\mathbf{A} = \tfrac{1}{2}\begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix},$$

whose determinant is $\frac{1}{16}$ det of the reversal permutation matrix, namely $\frac{1}{16} \cdot 1 \neq 0$. So β is nondegenerate by Proposition 14.1.2.

(b) Take $W = \operatorname{span}(\mathbf{e}_1, \mathbf{e}_2)$. Every $\mathbf{w} = (a, b, 0, 0)$ has $q(\mathbf{w}) = 0$, and $\beta(\mathbf{e}_1, \mathbf{e}_2) = 0$ from the matrix, so W is totally isotropic of dimension 2. No totally isotropic subspace has dimension 3, because Proposition 14.7.2 caps it at $4/2 = 2$.

(c) The two planes $\operatorname{span}(\mathbf{e}_1, \mathbf{e}_4)$ and $\operatorname{span}(\mathbf{e}_2, \mathbf{e}_3)$ are orthogonal, and $(\mathbf{e}_1, 2\mathbf{e}_4)$ is a hyperbolic pair, since $q(\mathbf{e}_1) = q(2\mathbf{e}_4) = 0$ and $\beta(\mathbf{e}_1, 2\mathbf{e}_4) = 2 \cdot \frac{1}{2} = 1$; the same works for $(\mathbf{e}_2, 2\mathbf{e}_3)$. So each plane is hyperbolic, of signature $(1, 1)$. Hence q has signature $(2, 2)$ and, by Proposition 14.7.7 (a), Witt index $\min(2, 2) = 2$, attained by the W of part (b).

14.7.7.3 C. Going deeper

Exercise. C1: Isotropy makes the form surjective

Let β be a nondegenerate symmetric form on V over a field F with $\operatorname{char} F \neq 2$.

(a) Prove that if β has an isotropic vector, then $q\colon V \to F$ is surjective.

(b) Deduce that if some $c \in F$ is not of the form $q(v)$, then β is anisotropic.

(c) Use (b) to show that the form $q(x_1, x_2) = x_1^2 + x_2^2$ on $\mathbb{F}_3^2$ is anisotropic, and give its Witt index.

Solution.

(a) Let e be isotropic. By Lemma 14.7.4 there is f with (e, f) a hyperbolic pair. For $c \in F$, put $v = \frac{c}{2}e + f$; this is legitimate since $\operatorname{char} F \neq 2$. Then

$$q(v) = \tfrac{c^2}{4}q(e) + 2 \cdot \tfrac{c}{2}\beta(e, f) + q(f) = 0 + c + 0 = c.$$

So every $c \in F$ is a value of q.

(b) Contrapositive of (a): if q is not surjective then β has no isotropic vector.

(c) The squares in $\mathbb{F}_3$ are $0^2 = 0$, $1^2 = 1$, $2^2 = 1$, so the values of $x_1^2 + x_2^2$ are $0, 1, 2$ — which

is all of $\mathbb{F}_3$, so (b) gives nothing directly. Argue instead from the definition: $q(\mathbf{x}) = 0$ needs $x_1^2 = -x_2^2 = 2x_2^2$. If $x_2 \neq 0$ then $x_2^2 = 1$, so $x_1^2 = 2$, which is not a square in $\mathbb{F}_3$; hence $x_2 = 0$ and then $x_1 = 0$. The form is anisotropic, its Witt index is 0, and the Gram matrix $\mathbf{I}_2$ shows it is nondegenerate. Over $\mathbb{F}_5$ the same form is isotropic, since $1^2 + 2^2 = 0$.

Exercise. C2: A maximal totally isotropic subspace and its complement

Let β be nondegenerate symmetric on V with $\dim V = n$, and let $W \subseteq V$ be totally isotropic with $\dim W = \operatorname{ind}(\beta) = m$.

(a) Prove that $\dim W^{\perp_\beta} = n - m$ and $W \subseteq W^{\perp_\beta}$.

(b) Prove that the form induced by β on the quotient $W^{\perp_\beta}/W$ is well defined and anisotropic. *Hint: a coset with an isotropic representative would enlarge W.*

(c) Deduce that $n - 2m = \dim(W^{\perp_\beta}/W)$ agrees with the dimension of the anisotropic part in Theorem 14.7.5.

Solution.

(a) The dimension is Lemma 14.7.1, using nondegeneracy, and the inclusion is Definition 14.7.2.

(b) For $u, v \in W^{\perp_\beta}$ and $w, w' \in W$,

$$\beta(u + w, v + w') = \beta(u, v),$$

because each of the three remaining terms pairs a vector of W with a vector of $W^{\perp_\beta} \supseteq W$, hence vanishes. So $\bar{\beta}(u + W, v + W) := \beta(u, v)$ is well defined, and it is bilinear and symmetric. Suppose $\bar{q}(u + W) = 0$ with $u \notin W$, that is $q(u) = 0$ with $u \in W^{\perp_\beta} \setminus W$. Then $W + \operatorname{span}(u)$ is totally isotropic: it is spanned by isotropic vectors, $\beta(u, w) = 0$ for $w \in W$ because $u \in W^{\perp_\beta}$, and β vanishes on $W \times W$. Its dimension is $m + 1 > \operatorname{ind}(\beta)$, a contradiction. So $\bar{\beta}$ is anisotropic.

(c) By (a), $\dim(W^{\perp_\beta}/W) = (n - m) - m = n - 2m$. By Proposition 14.7.6 the anisotropic part V_0 of a Witt decomposition also has dimension $n - 2m$, since the index is m. (That the two anisotropic forms agree, and not merely their dimensions, is proved in the next section.)

Exercise. C3: Where characteristic 2 breaks

Work over $\mathbb{F}_2$ and let β be the dot product on $\mathbb{F}_2^2$, with Gram matrix $\mathbf{I}_2$.

(a) Verify that β is symmetric and nondegenerate, and find all $\mathbf{v}$ with $q(\mathbf{v}) = 0$.

(b) Show that β has a vector with $q(\mathbf{v}) = 0$ and $\mathbf{v} \neq \mathbf{0}$ that lies in no hyperbolic plane, and identify which line of the proof of Lemma 14.7.4 fails.

(c) Explain in one sentence why the third description of "totally isotropic" given after Definition 14.7.2 is unavailable here.

Solution.

(a) The Gram matrix $\mathbf{I}_2$ is symmetric and invertible, so β is symmetric and nondegenerate (Proposition 14.1.2). Here $q(\mathbf{x}) = x_1^2 + x_2^2 = (x_1 + x_2)^2$ over $\mathbb{F}_2$, which vanishes exactly when $x_1 = x_2$. So $q(\mathbf{v}) = 0$ for $\mathbf{v} = \mathbf{0}$ and for $\mathbf{v} = (1, 1)$.

(b) Take $\mathbf{v} = (1, 1)$. A hyperbolic plane containing it would need $\mathbf{f}$ with $q(\mathbf{f}) = 0$ and $\beta(\mathbf{v}, \mathbf{f}) = 1$. The only nonzero vector with $q = 0$ is $\mathbf{v}$ itself, and $\beta(\mathbf{v}, \mathbf{v}) = q(\mathbf{v}) = 0 \neq 1$. The step that fails is the correction $\mathbf{f} = \mathbf{u} - \frac{1}{2}q(\mathbf{u})\mathbf{e}$: the element $2 = 1 + 1$ is 0 in $\mathbb{F}_2$ and is not invertible, so $\frac{1}{2}$ does not exist.

(c) Because it is Theorem 14.3.5, which needs $\operatorname{char} F \neq 2$ to divide by 2; over $\mathbb{F}_2$ the line $\operatorname{span}((1, 1))$ has $q \equiv 0$ on it and is indeed totally isotropic here, but the implication "q vanishes on W implies β vanishes on $W \times W$" is not available in general.

14.8 Witt's Extension and Cancellation Theorems

The previous section left one question open. A nondegenerate space splits as hyperbolic planes plus an anisotropic remainder, and the number of planes is pinned down by the Witt index — but the remainder is a subspace chosen along the way, and different choices give different subspaces. What should be true is that they all carry the same form. Proving it needs a theorem with a much wider reach: any isometry between two subspaces of a nondegenerate space is the restriction of an isometry of the whole space. That theorem is the longest proof of the chapter, and the cancellation statement that finishes Section 7 falls out of it in four lines.

Throughout, F is a field with $\operatorname{char} F \neq 2$ (Definition 0.7.2), all spaces are finite-dimensional over F, and all forms are **symmetric** bilinear forms (Definition 14.3.1). Orthogonal complements $U^{\perp_\beta}$, orthogonal direct sums $U \perp W$, and isotropic and hyperbolic vocabulary are as in Section 7.

14.8.1 Isometries of forms

Chapter 11 called a linear map an isometry when it preserved the inner product (Definition 11.7.1). The same word works here, with the form in place of the inner product and with no positivity anywhere.

An isometry is a change of variables that the form does not notice.

Definition 14.8.1. Isometry of Bilinear Spaces

Let β be a bilinear form on V and β' one on V'. An **isometry** from (V, β) to (V', β') is a linear **bijection** $\sigma\colon V \to V'$ with

$$\beta'(\sigma(u), \sigma(v)) = \beta(u, v) \qquad \text{for all } u, v \in V.$$

The two spaces are **isometric**, written $(V, \beta) \cong (V', \beta')$ or just $V \cong V'$, if such a σ exists. An isometry of (V, β) with itself is an **isometry of** V.

When $U \subseteq V$ and $U' \subseteq V'$ are subspaces, "an isometry $U \to U'$" always means an isometry for the **restricted** forms $\beta|_{U\times U}$ and $\beta'|_{U'\times U'}$.

Two clauses need pointing at. The map must be a **bijection**, so that the inverse is again an isometry; and the identity must hold for **all** pairs, though by bilinearity it is enough to check it on a pair of bases. The symbol $\cong$ is the one used for isomorphism throughout the book, and no confusion arises: an isometry is an isomorphism that also respects the form, and within this chapter $\cong$ between spaces carrying forms always means isometric.

Three readings make the notion concrete.

- **Isometry is congruence in disguise.** If $\mathcal{B}$ is a basis of V and $\mathcal{B}'$ one of V', then σ is an isometry exactly when the Gram matrices satisfy $[\beta]_{\mathcal{B}}^{\mathcal{B}} = \mathbf{P}^{\top}\,[\beta']_{\mathcal{B}'}^{\mathcal{B}'}\,\mathbf{P}$ with $\mathbf{P} = [\sigma]_{\mathcal{B}}^{\mathcal{B}'}$. So $(V,\beta) \cong (V',\beta')$ if and only if their Gram matrices are congruent (Definition 14.2.1). Every classification theorem of this chapter is therefore a classification up to isometry.
- **Over $\mathbb{R}$.** Two real spaces with nondegenerate symmetric forms are isometric exactly when they have the same signature, by Corollary 14.5.2. So, writing $\mathbb{R}^{p,q}$ for $\mathbb{R}^{p+q}$ carrying the form of signature (p,q), we have $\mathbb{R}^{2,1} \ncong \mathbb{R}^{1,2}$ even though both are three-dimensional and both nondegenerate.
- **The degenerate case.** An isometry carries $\operatorname{rad}(\beta)$ onto $\operatorname{rad}(\beta')$, because $\beta'(\sigma u, \sigma v) = \beta(u,v)$ and σ is surjective. In particular isometric spaces have radicals of equal dimension.

The isometries of a fixed (V,β) form a group under composition; Section 10 names the classical examples. For now only two kinds are needed, and the second is built from the first.

14.8.2 Reflections

Chapter 11 §08 built Householder reflections out of an inner product, to clear a column. The same formula works for any symmetric form, provided the vector we reflect in is anisotropic, and it is the only explicit isometry this section needs.

Definition 14.8.2. Reflection in an Anisotropic Vector

Let β be a symmetric bilinear form on V and let $w \in V$ be anisotropic, that is $q(w) \neq 0$. The **reflection in** w is the map

$$\tau_w \colon V \to V, \qquad \tau_w(v) = v - \frac{2\beta(v,w)}{q(w)}\, w.$$

The formula is Definition 11.8.1 with $\langle v, w\rangle$ replaced by $\beta(v,w)$ and $\|w\|^2$ by $q(w)$; the division is legitimate precisely because w is anisotropic.

Proposition 14.8.1. Reflections Are Involutive Isometries

Let β be a symmetric bilinear form on V and $w \in V$ anisotropic. Then τ_w is linear, and:

(a) $\tau_w(w) = -w$, and $\tau_w(v) = v$ for every v with $\beta(v,w) = 0$;
(b) $\tau_w \circ \tau_w = \mathrm{id}_V$; in particular τ_w is bijective;
(c) $\beta(\tau_w u, \tau_w v) = \beta(u,v)$ for all $u, v \in V$.

Hence τ_w is an isometry of V.

Proof. Linearity holds because $v \mapsto \beta(v,w)$ is linear and $q(w)$ is a fixed nonzero scalar. Write $c(v) = 2\beta(v,w)/q(w)$.

(a) $c(w) = 2q(w)/q(w) = 2$, so $\tau_w(w) = w - 2w = -w$. If $\beta(v,w) = 0$ then $c(v) = 0$ and $\tau_w(v) = v$.

(b) By (a) and linearity, $\beta(\tau_w v, w) = \beta(v,w) - c(v)q(w) = -\beta(v,w)$, so $c(\tau_w v) = -c(v)$ and
$$\tau_w(\tau_w v) = \big(v - c(v)w\big) + c(v)w = v.$$

(c) Expanding by bilinearity and using symmetry,
$$\beta(\tau_w u, \tau_w v) = \beta(u,v) - c(v)\beta(u,w) - c(u)\beta(w,v) + c(u)c(v)q(w).$$
Now $c(u)c(v)q(w) = 4\beta(u,w)\beta(v,w)/q(w)$, while the two middle terms together contribute $-4\beta(u,w)\beta(v,w)/q(w)$. They cancel, leaving $\beta(u,v)$. Together with (b) this makes τ_w a bijective linear map preserving β, that is an isometry.

■

The use of reflections is that they move one anisotropic vector to any other of the same value — the one explicit construction the long proof below will need.

Lemma 14.8.2. Moving One Anisotropic Vector to Another

Let $\operatorname{char} F \neq 2$, let β be a symmetric bilinear form on V, and let $u, u' \in V$ satisfy
$$q(u) = q(u') \neq 0.$$
Then there is an isometry ρ of V with $\rho(u) = u'$, and ρ may be taken to be a reflection or a product of two reflections.

Idea. The reflection in $u - u'$ is the natural candidate, and it works whenever $u - u'$ is anisotropic. It need not be — but $u + u'$ then is, because the two values $q(u-u')$ and $q(u+u')$ add up to $4q(u) \neq 0$ and so cannot both vanish. Reflecting in $u+u'$ overshoots to $-u'$, and one more reflection, in u' itself, brings it back.

Proof. Expanding both by bilinearity and adding,
$$q(u-u') + q(u+u') = 2q(u) + 2q(u') = 4q(u) \neq 0,$$
where the last step uses $q(u) = q(u') \neq 0$. So at least one of $q(u-u')$ and $q(u+u')$ is nonzero.

Case 1: $q(u-u') \neq 0$. Put $w = u - u'$. Then
$$q(w) = q(u) - 2\beta(u,u') + q(u') = 2\big(q(u) - \beta(u,u')\big),$$
while $\beta(u,w) = q(u) - \beta(u,u')$. Hence $2\beta(u,w)/q(w) = 1$, and
$$\tau_w(u) = u - w = u'.$$

Take $\rho = \tau_w$.

Case 2: $q(u - u') = 0$. Then $q(u + u') = 4q(u) \neq 0$; put $w = u + u'$. The same computation with $-u'$ in place of u' gives $q(w) = 2(q(u) + \beta(u, u'))$ and $\beta(u, w) = q(u) + \beta(u, u')$, so again $2\beta(u, w)/q(w) = 1$ and

$$\tau_w(u) = u - w = -u'.$$

Since $q(u') \neq 0$, the reflection $\tau_{u'}$ is defined, and $\tau_{u'}(-u') = u'$ by Proposition 14.8.1 (a). Take $\rho = \tau_{u'} \circ \tau_w$, a composition of two isometries and hence an isometry. This proves the lemma.

■

Quick check. In $\mathbb{R}^3$ with $q(\mathbf{x}) = x_1^2 + x_2^2 - x_3^2$, find a reflection carrying $\mathbf{e}_1$ to $\mathbf{e}_2$.

Solution. Both have $q = 1 \neq 0$, so Lemma 14.8.2 applies. Case 1 works: $\mathbf{w} = \mathbf{e}_1 - \mathbf{e}_2$ has $q(\mathbf{w}) = 1 + 1 = 2 \neq 0$, so

$$\tau_{\mathbf{w}}(\mathbf{v}) = \mathbf{v} - \beta(\mathbf{v}, \mathbf{w})\,\mathbf{w},$$

since $2/q(\mathbf{w}) = 1$. Then $\beta(\mathbf{e}_1, \mathbf{w}) = 1$ and $\tau_{\mathbf{w}}(\mathbf{e}_1) = \mathbf{e}_1 - (\mathbf{e}_1 - \mathbf{e}_2) = \mathbf{e}_2$, as required. The reflection fixes $\mathbf{e}_1 + \mathbf{e}_2$ and $\mathbf{e}_3$, which are the vectors orthogonal to $\mathbf{w}$.

14.8.3 Witt's extension theorem

Here is the chapter's long theorem. Read the statement carefully: the two subspaces live in the **same** space V, and the form on V is required to be nondegenerate, but nothing at all is assumed about the restrictions to U and U' — they may be degenerate, they may be zero.

Theorem 14.8.3. Witt's Extension Theorem

Let F be a field with $\operatorname{char} F \neq 2$, let V be a finite-dimensional vector space over F, and let β be a **nondegenerate** symmetric bilinear form on V. Let $U, U' \subseteq V$ be subspaces and let

$$\sigma : U \to U'$$

be an isometry for the restricted forms. Then there is an isometry φ of V with $\varphi|_U = \sigma$.

Idea. Induction, but on what? Two different moves are available, and each improves a different number.

① If $\beta|_U$ is **nondegenerate**, then U contains an anisotropic vector u_1. A reflection moves $\sigma(u_1)$ back to u_1, so we may assume σ fixes u_1; then everything of U orthogonal to u_1 stays inside the hyperplane $u_1^{\perp}$, which is again nondegenerate and **one dimension smaller**. Recurse there and glue.

② If $\beta|_U$ is **degenerate**, pick a nonzero z in its radical. Nondegeneracy of β on V produces a partner w making (z, w) a hyperbolic pair orthogonal to the rest of U, and σ extends to $U \oplus Fw$ by sending w to the matching partner on the other side. The subspace has grown,

but its **radical has shrunk by one**.

So the induction runs on the pair $(\dim V,\ \dim \operatorname{rad}(\beta|_U))$ read lexicographically: ① lowers the first entry, ② keeps it and lowers the second. Step 0 disposes of $U = \{0\}$.

Proof. We prove the statement for every triple (V, U, σ) satisfying the hypotheses, by induction on the pair

$$(\dim V,\ \dim \operatorname{rad}(\beta|_{U\times U}))$$

ordered lexicographically: one pair precedes another if its first entry is smaller, or the first entries agree and the second is smaller. This order on pairs of nonnegative integers has no infinite descending chain, so the induction is legitimate. Write $R = \operatorname{rad}(\beta|_{U\times U}) = U \cap U^{\perp_\beta}$.

Step 0: $U = \{0\}$**.** Then $U' = \{0\}$ and $\varphi = \mathrm{id}_V$ works.

Step 1: $U \neq \{0\}$ **and** $R = \{0\}$**.** Since $\beta|_U$ is nondegenerate and $U \neq \{0\}$, the form $\beta|_U$ is not the zero form. By Lemma 14.4.1, applied to $\beta|_U$ on U and using $\operatorname{char} F \neq 2$, there is $u_1 \in U$ with $q(u_1) \neq 0$.

Claim. We may assume $\sigma(u_1) = u_1$.

Proof. Put $u_1' = \sigma(u_1)$. Then $q(u_1') = \beta(\sigma u_1, \sigma u_1) = q(u_1) \neq 0$, so Lemma 14.8.2 gives an isometry ρ of V with $\rho(u_1) = u_1'$. The composite $\tau = \rho^{-1} \circ \sigma$ is an isometry from U onto $\rho^{-1}(U')$ with $\tau(u_1) = u_1$, and if φ_0 is an isometry of V extending τ, then $\rho \circ \varphi_0$ is an isometry of V extending σ. So it suffices to treat τ.

■

Assume then that $\sigma(u_1) = u_1$. The map $u \mapsto \beta(u_1, u)$ is a linear functional on U taking the nonzero value $q(u_1)$ at u_1, so its kernel

$$U_1 = \{\, u \in U : \beta(u_1, u) = 0 \,\}$$

has $\dim U_1 = \dim U - 1$ by Rank–Nullity (Theorem 2.3.1), and $U = \operatorname{span}(u_1) \oplus U_1$, since $a u_1 \in U_1$ forces $a\, q(u_1) = 0$ and hence $a = 0$.

Because $q(u_1) \neq 0$, the form restricted to $\operatorname{span}(u_1)$ is nondegenerate, so Proposition 14.7.3 gives

$$V = \operatorname{span}(u_1) \perp W, \qquad W = \operatorname{span}(u_1)^{\perp_\beta},$$

with $\beta|_W$ nondegenerate and $\dim W = \dim V - 1$. Now $U_1 \subseteq W$ by definition, and $\sigma(U_1) \subseteq W$ as well: for $u \in U_1$,

$$\beta(u_1, \sigma u) = \beta(\sigma u_1, \sigma u) = \beta(u_1, u) = 0.$$

So $\sigma|_{U_1} : U_1 \to \sigma(U_1)$ is an isometry between two subspaces of W, and $(W, \beta|_W)$ satisfies the hypotheses of the theorem with $\dim W < \dim V$. By the induction hypothesis there is an isometry θ of W with $\theta|_{U_1} = \sigma|_{U_1}$.

Define $\varphi\colon V \to V$ by

$$\varphi(a u_1 + x) = a u_1 + \theta(x) \qquad (a \in F,\ x \in W),$$

which is well defined and linear because $V = \operatorname{span}(u_1) \oplus W$, and bijective because θ is. It is

an isometry: using the orthogonality of $\operatorname{span}(u_1)$ and W on both sides,

$$\begin{aligned}\beta(\varphi(au_1+x),\varphi(bu_1+y)) &= ab\,q(u_1)+\beta(\theta x,\theta y)\\ &= ab\,q(u_1)+\beta(x,y)\\ &= \beta(au_1+x,\ bu_1+y).\end{aligned}$$

Finally φ extends σ: a vector $u\in U$ is au_1+x with $x\in U_1\subseteq W$, and

$$\varphi(u)=au_1+\theta(x)=a\sigma(u_1)+\sigma(x)=\sigma(u).$$

Step 2: $R\neq\{0\}$. Choose $z\in R$ with $z\neq 0$ and a complement U_0 of $\operatorname{span}(z)$ in U, so that $U=\operatorname{span}(z)\oplus U_0$ (Theorem 1.7.6). Since σ is an isometry of U onto U', it carries R onto $\operatorname{rad}(\beta|_{U'})$; write $z'=\sigma(z)$ and $U_0'=\sigma(U_0)$, so that $U'=\operatorname{span}(z')\oplus U_0'$ and z' lies in the radical of $\beta|_{U'}$. Note $q(z)=0$, because z pairs to zero with everything in U, in particular with itself; likewise $q(z')=0$.

Claim. There is $w\in V$ with $q(w)=0$, $\beta(z,w)=1$, and $\beta(u_0,w)=0$ for every $u_0\in U_0$.

Proof. Consider the linear map $S\colon V\to U^*$ sending v to the restriction of $\beta(\cdot,v)$ to U. Its kernel is $U^{\perp_\beta}$, which has dimension $\dim V-\dim U$ by Lemma 14.7.1, since β is nondegenerate. By Rank–Nullity (Theorem 2.3.1), $\operatorname{rank}S=\dim U=\dim U^*$, so S is surjective. Let $\psi_z\in U^*$ be the functional with $\psi_z(z)=1$ and $\psi_z(U_0)=0$, which exists because $U=\operatorname{span}(z)\oplus U_0$. Pick $w_0\in V$ with $S(w_0)=\psi_z$, that is $\beta(z,w_0)=1$ and $\beta(u_0,w_0)=0$ for $u_0\in U_0$.

Put $w=w_0-\tfrac12 q(w_0)\,z$, which uses $\operatorname{char}F\neq 2$. Since $q(z)=0$ and $\beta(u,z)=0$ for every $u\in U$,

$$\beta(z,w)=1-\tfrac12 q(w_0)q(z)=1,\qquad \beta(u_0,w)=0,$$

and

$$q(w)=q(w_0)-q(w_0)\beta(w_0,z)+\tfrac14 q(w_0)^2q(z)=0,$$

using $\beta(w_0,z)=\beta(z,w_0)=1$.

■

Apply the claim on both sides: it gives w for (z,U_0) and, by the same argument with z' and U_0', a vector w' with $q(w')=0$, $\beta(z',w')=1$, and $\beta(U_0',w')=0$.

The vector w is not in U, since $\beta(z,w)=1$ while $\beta(z,u)=0$ for every $u\in U$; likewise $w'\notin U'$. So

$$\widetilde{U}=U\oplus Fw,\qquad \widetilde{U}'=U'\oplus Fw'$$

are subspaces of V of dimension $\dim U+1$. Define $\widetilde{\sigma}(u+cw)=\sigma(u)+cw'$, a linear bijection $\widetilde{U}\to\widetilde{U}'$. It is an isometry: for $u=az+u_0\in U$ we have

$$\beta(u,w)=a\beta(z,w)+\beta(u_0,w)=a,$$

and $\sigma(u)=az'+\sigma(u_0)$ with $\sigma(u_0)\in U_0'$, so $\beta(\sigma u,w')=a$ as well. Hence, expanding by bilinearity and using $q(w)=q(w')=0$,

$$\begin{aligned}\beta\big(\widetilde{\sigma}(u+cw),\,\widetilde{\sigma}(y+dw)\big) &= \beta(\sigma u,\sigma y)+d\,\beta(\sigma u,w')+c\,\beta(\sigma y,w')\\ &= \beta(u,y)+d\,\beta(u,w)+c\,\beta(y,w)\\ &= \beta(u+cw,\ y+dw).\end{aligned}$$

Claim. $\operatorname{rad}(\beta|_{\widetilde{U}}) = R \cap U_0$, and its dimension is $\dim R - 1$.

Proof. Let $x = az + u_0 + cw$ lie in $\operatorname{rad}(\beta|_{\widetilde{U}})$, with $a, c \in F$ and $u_0 \in U_0$. Pairing x with w gives $a\beta(z,w) + \beta(u_0,w) + c\,q(w) = a$, so $a = 0$. Pairing x with z gives $\beta(u_0,z) + c\,\beta(w,z) = c$, since $z \in R$ kills u_0; so $c = 0$. Hence $x = u_0 \in U_0$, and x pairs to zero with all of $U \subseteq \widetilde{U}$, so $x \in R \cap U_0$. Conversely a vector of $R \cap U_0$ pairs to zero with U and with w, hence with $\widetilde{U}$.

For the dimension: every $r \in R \subseteq U$ is $az + u_0$, and $u_0 = r - az \in R$ because $z \in R$; so $R = \operatorname{span}(z) \oplus (R \cap U_0)$ and $\dim(R \cap U_0) = \dim R - 1$.

■

The pair attached to the triple $(V, \widetilde{U}, \widetilde{\sigma})$ is therefore $(\dim V, \dim R - 1)$, which precedes $(\dim V, \dim R)$ in the lexicographic order. By the induction hypothesis there is an isometry φ of V with $\varphi|_{\widetilde{U}} = \widetilde{\sigma}$. Since $U \subseteq \widetilde{U}$ and $\widetilde{\sigma}|_U = \sigma$, this φ extends σ.

Steps 0, 1, and 2 cover every case, so the induction is complete and the theorem is proved.

■

Three places earned their hypotheses. Nondegeneracy of β on V was used twice, once through Proposition 14.7.3 in Step 1 and once through the surjectivity of S in Step 2. Characteristic $\neq 2$ was used twice as well, in Lemma 14.4.1 and in the correction $w = w_0 - \frac{1}{2}q(w_0)z$, which is the same correction that built a hyperbolic pair in Lemma 14.7.4. And symmetry is everywhere: without it, "the" radical and "the" orthogonal complement would each split into a left and a right version.

14.8.4 Cancellation

The corollary is what the theorem is usually used for, and it is where a careless statement goes wrong. Witt cancellation removes a summand from an **orthogonal** decomposition of two spaces that are **already known to be isometric**, and what it produces is an isometry of the complements.

Corollary 14.8.4. Witt's Cancellation Theorem

Let $\operatorname{char} F \neq 2$.

(a) Let β be a nondegenerate symmetric form on V, and let $U, U' \subseteq V$ be isometric subspaces. Then $U^{\perp_\beta} \cong U'^{\perp_\beta}$.

(b) Let (V, β) and (V', β') be finite-dimensional spaces with nondegenerate symmetric forms, and suppose

$$V = U \perp W, \qquad V' = U' \perp W'.$$

If $V \cong V'$ and $U \cong U'$, then $W \cong W'$.

Idea. For (a), extend the isometry $U \to U'$ to the whole of V and watch where the complement goes: an isometry of V carrying U onto U' has no choice but to carry $U^\perp$ onto $U'^\perp$. For (b), transport everything into the single space V along an isometry $V \to V'$, and apply (a).

Proof.

(a) Let $\sigma\colon U \to U'$ be an isometry. By Theorem 14.8.3 there is an isometry φ of V with $\varphi|_U = \sigma$; in particular $\varphi(U) = U'$. If $x \in U^{\perp_\beta}$ and $u' \in U'$, write $u' = \varphi(u)$ with $u \in U$; then

$$\beta(u', \varphi x) = \beta(\varphi u, \varphi x) = \beta(u, x) = 0,$$

so $\varphi(U^{\perp_\beta}) \subseteq U'^{\perp_\beta}$. Both subspaces have dimension $\dim V - \dim U$ by Lemma 14.7.1, using $\dim U = \dim U'$, and φ is injective, so the inclusion is an equality. Restricting φ gives an isometry $U^{\perp_\beta} \to U'^{\perp_\beta}$.

(b) First, $W = U^{\perp_\beta}$: the orthogonality in $V = U \perp W$ gives $W \subseteq U^{\perp_\beta}$, and both have dimension $\dim V - \dim U$, so they are equal (Theorem 1.6.8). Likewise $W' = U'^{\perp_{\beta'}}$.

Let $\psi\colon V \to V'$ be an isometry. Then $\psi^{-1}(U')$ is a subspace of V isometric to U', hence isometric to U. Moreover $\psi^{-1}(U'^{\perp_{\beta'}}) = (\psi^{-1}(U'))^{\perp_\beta}$, because ψ preserves the forms and is surjective. Applying (a) inside V to the isometric subspaces U and $\psi^{-1}(U')$,

$$W = U^{\perp_\beta} \cong (\psi^{-1}(U'))^{\perp_\beta} = \psi^{-1}(W') \cong W',$$

the last isometry being the restriction of ψ. This proves the corollary.

■

Warning. **Cancellation needs the whole spaces to be isometric, and the sums to be orthogonal.** Neither hypothesis can be dropped.

Drop the first: inside $\mathbb{R}^2$ with $q = x_1^2 + x_2^2$ and inside $\mathbb{R}^2$ with $q' = x_1^2 - x_2^2$, the lines $\operatorname{span}(\mathbf{e}_1)$ are isometric, each carrying the form $\langle 1\rangle$. Their complements are $\operatorname{span}(\mathbf{e}_2)$ with $q = x_2^2$ and $\operatorname{span}(\mathbf{e}_2)$ with $q' = -x_2^2$, which are **not** isometric: one is positive definite, the other negative definite. Isometric subspaces of **different** ambient spaces say nothing about complements.

Drop the second: in $\mathbb{R}^2$ with $q = x_1^2 - x_2^2$, both

$$V = \operatorname{span}(\mathbf{e}_1) \oplus \operatorname{span}(\mathbf{e}_1 + \mathbf{e}_2) \quad \text{and} \quad V = \operatorname{span}(\mathbf{e}_1) \oplus \operatorname{span}(\mathbf{e}_2)$$

are direct sums with the same first summand. The second summands carry $q(\mathbf{e}_1 + \mathbf{e}_2) = 0$ and $q(\mathbf{e}_2) = -1$, so they are not isometric. The first decomposition is not orthogonal, since $\beta(\mathbf{e}_1, \mathbf{e}_1 + \mathbf{e}_2) = 1 \neq 0$, and that is exactly what breaks.

14.8.5 The Witt decomposition is unique

We can now close the question Section 7 left open.

Corollary 14.8.5. The Anisotropic Part Is Determined

Let $\operatorname{char} F \neq 2$, let β be a nondegenerate symmetric form on a finite-dimensional V, and let

$$V = H_1 \perp \cdots \perp H_m \perp V_0 = H_1' \perp \cdots \perp H_{m'}' \perp V_0'$$

be two decompositions as in Theorem 14.7.5. Then $m = m' = \operatorname{ind}(\beta)$ and $V_0 \cong V_0'$.

Proof. That $m = m' = \operatorname{ind}(\beta)$ is Proposition 14.7.6. Put $U = H_1 \perp \cdots \perp H_m$ and $U' = H_1' \perp \cdots \perp H_m'$. Choosing a hyperbolic basis of each plane gives bases of U and of U' in which both Gram matrices are the same block diagonal matrix, namely m copies of $\left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$ down the diagonal. The linear bijection $U \to U'$ matching these bases in order is therefore an isometry.

Both V_0 and V_0' are complements: $V_0 \subseteq U^{\perp_\beta}$ by orthogonality, and the dimensions agree, so $V_0 = U^{\perp_\beta}$ (Theorem 1.6.8), and likewise $V_0' = U'^{\perp_\beta}$. By Corollary 14.8.4 (a), $V_0 \cong V_0'$. This proves the corollary.

■

So a nondegenerate symmetric form is completely described by two pieces of data: an integer, the Witt index, and an anisotropic form, determined up to isometry. Classifying all forms over a field F therefore reduces to classifying the **anisotropic** ones, which is where the arithmetic of F enters and where the subject stops being linear algebra. Over $\mathbb{R}$ the anisotropic forms are the definite ones, so the data $(\operatorname{ind}, V_0)$ is just the signature again. Over $\mathbb{C}$ the anisotropic forms have dimension 0 or 1, so the index and the parity of n are everything. Over $\mathbb{Q}$ the classification is a theorem of number theory and is not in this book.

14.8.6 Exercises

14.8.6.1 A. Check your understanding

Exercise. A1

(a) State Witt's extension theorem, with all its hypotheses.

(b) Define the reflection τ_w and say which vectors it fixes.

(c) Determine whether the following statement is correct, and justify your answer: if $U \subseteq V$ and $U' \subseteq V'$ are isometric subspaces of nondegenerate spaces, then $U^\perp$ and $U'^\perp$ are isometric.

(d) Where is $\operatorname{char} F \neq 2$ used in the proof of Theorem 14.8.3?

(e) Two real nondegenerate forms of dimension 6 both have Witt index 2. Must they be isometric?

Solution.

(a) Let $\operatorname{char} F \neq 2$, let β be a nondegenerate symmetric bilinear form on a finite-dimensional V over F, let $U, U' \subseteq V$ be subspaces and $\sigma\colon U \to U'$ an isometry for the restricted forms. Then σ extends to an isometry of V.

(b) $\tau_w(v) = v - (2\beta(v,w)/q(w))w$, defined for anisotropic w. It fixes every v with $\beta(v,w) = 0$ and sends w to $-w$.

(c) Incorrect. The warning after Corollary 14.8.4 gives the witness: the line $\operatorname{span}(\mathbf{e}_1)$ inside $(\mathbb{R}^2, x_1^2 + x_2^2)$ and inside $(\mathbb{R}^2, x_1^2 - x_2^2)$ are isometric, with complements of opposite definiteness. Cancellation requires the ambient spaces to be isometric too.

(d) Twice. In Step 1, through Lemma 14.4.1, which produces an anisotropic vector in a nonzero subspace where the form is not zero. In Step 2, through the correction $w = w_0 - \frac{1}{2}q(w_0)z$, which divides by 2.

(e) No. By Corollary 14.8.5 each is two hyperbolic planes orthogonal to an anisotropic part of dimension $6-4=2$, and over $\mathbb{R}$ an anisotropic form is definite, so that part has signature $(2,0)$ or $(0,2)$. The two possibilities give signatures $(4,2)$ and $(2,4)$ for the whole space, which are different, so the forms need not be isometric. What Corollary 14.8.5 says is that the anisotropic part is determined by the form, not that the index determines the form.

14.8.6.2 B. Practice

Exercise. B1: Extend an isometry by hand

Let $V=\mathbb{R}^3$ with $q(\mathbf{x})=x_1^2+x_2^2-x_3^2$, let $U=\operatorname{span}(\mathbf{e}_1)$, $U'=\operatorname{span}(\mathbf{e}_2)$, and let $\sigma(\mathbf{e}_1)=\mathbf{e}_2$.

(a) Verify that σ is an isometry of U onto U'.

(b) Extend σ to an isometry of V, and write down its matrix in the standard basis.

(c) Determine $U^{\perp_\beta}$ and $U'^{\perp_\beta}$ and exhibit the isometry between them produced by your extension.

Solution.

(a) $\beta(a\mathbf{e}_1,b\mathbf{e}_1)=ab$ and $\beta(\sigma(a\mathbf{e}_1),\sigma(b\mathbf{e}_1))=\beta(a\mathbf{e}_2,b\mathbf{e}_2)=ab$, and σ is a linear bijection $U\to U'$.

(b) Use the Quick check: $\mathbf{w}=\mathbf{e}_1-\mathbf{e}_2$ has $q(\mathbf{w})=2$, and $\tau_{\mathbf{w}}(\mathbf{v})=\mathbf{v}-\beta(\mathbf{v},\mathbf{w})\mathbf{w}$ sends $\mathbf{e}_1$ to $\mathbf{e}_2$. In coordinates $\beta(\mathbf{v},\mathbf{w})=v_1-v_2$, so

$$\tau_{\mathbf{w}}(\mathbf{v})=(v_2,\ v_1,\ v_3),\qquad \mathbf{P}=\begin{pmatrix}0&1&0\\1&0&0\\0&0&1\end{pmatrix}.$$

This swaps the first two coordinates and indeed satisfies $\mathbf{P}^{\top}\operatorname{diag}(1,1,-1)\mathbf{P}=\operatorname{diag}(1,1,-1)$.

(c) $U^{\perp_\beta}=\{\mathbf{v}:v_1=0\}=\operatorname{span}(\mathbf{e}_2,\mathbf{e}_3)$ and $U'^{\perp_\beta}=\operatorname{span}(\mathbf{e}_1,\mathbf{e}_3)$. The extension maps $\mathbf{e}_2\mapsto\mathbf{e}_1$ and $\mathbf{e}_3\mapsto\mathbf{e}_3$, an isometry between them; both carry the form $\operatorname{diag}(1,-1)$, as Corollary 14.8.4 (a) predicts.

Exercise. B2: Cancel a hyperbolic plane

Let H denote a hyperbolic plane over $\mathbb{R}$, and let A and B be nondegenerate real spaces with forms, with $\dim A=\dim B=3$. Suppose $H\perp A\cong H\perp B$.

(a) Prove that $A\cong B$.

(b) If A has signature $(2,1)$, what is the signature of B?

(c) Does $A\perp B\cong A\perp B$ with $A\cong A$ allow us to conclude $B\cong B$? Say which part of Corollary 14.8.4 is being used and why the statement is not vacuous.

Solution.

(a) Put $V=H\perp A$ and $V'=H\perp B$, both nondegenerate. The two copies of H are

isometric, and $V \cong V'$ by hypothesis, so Corollary 14.8.4 (b) gives $A \cong B$.

(b) By (a), $B \cong A$, so B has signature $(2,1)$ as well; isometric real spaces have equal signatures by Corollary 14.5.2.

(c) It uses (b) of Corollary 14.8.4 with $U = U' = A$ and $W = W' = B$; the conclusion $B \cong B$ is true but empty. The content of (b) appears only when the two decompositions are genuinely different, as in (a) above, where the hypothesis $V \cong V'$ is an assumption about the **whole** spaces and not about A and B.

Exercise. B3: Reflections in an indefinite space

In $\mathbb{R}^2$ with $q(\mathbf{x}) = x_1^2 - x_2^2$:

(a) Compute the matrix of $\tau_{\mathbf{e}_1}$ and of $\tau_{\mathbf{e}_2}$ in the standard basis, and check each satisfies $\mathbf{P}^\top \mathbf{A} \mathbf{P} = \mathbf{A}$ with $\mathbf{A} = \operatorname{diag}(1,-1)$.

(b) Explain why $\tau_{\mathbf{v}}$ is undefined for $\mathbf{v} = (1,1)$, and say what fails.

(c) Find an isometry carrying $(1,0)$ to $(-1,0)$ and one carrying $(2,\sqrt{3})$ to $(-2,\sqrt{3})$. *(Check the values of q first.)*

Solution.

(a) $q(\mathbf{e}_1) = 1$ and $\beta(\mathbf{v}, \mathbf{e}_1) = v_1$, so $\tau_{\mathbf{e}_1}(\mathbf{v}) = \mathbf{v} - 2v_1\mathbf{e}_1 = (-v_1, v_2)$, with matrix $\operatorname{diag}(-1,1)$. Similarly $q(\mathbf{e}_2) = -1$ and $\beta(\mathbf{v}, \mathbf{e}_2) = -v_2$, so the coefficient is $2(-v_2)/(-1) = 2v_2$ and $\tau_{\mathbf{e}_2}(\mathbf{v}) = \mathbf{v} - 2v_2\mathbf{e}_2 = (v_1, -v_2)$, with matrix $\operatorname{diag}(1,-1)$. Both matrices $\mathbf{P}$ are diagonal with entries ± 1, so $\mathbf{P}^\top \mathbf{A} \mathbf{P} = \mathbf{P}\mathbf{A}\mathbf{P} = \mathbf{A}$, as each sign is squared.

(b) $q(1,1) = 1 - 1 = 0$, so $(1,1)$ is isotropic and the denominator $q(\mathbf{w})$ in Definition 14.8.2 is zero. The formula divides by 0; there is no reflection in an isotropic vector.

(c) Both $(1,0)$ and $(-1,0)$ have $q = 1 \neq 0$. Here $\mathbf{u} - \mathbf{u}' = (2,0)$ with $q = 4 \neq 0$, so Case 1 of Lemma 14.8.2 applies: $\tau_{(2,0)} = \tau_{\mathbf{e}_1}$ works, and indeed $\tau_{\mathbf{e}_1}(1,0) = (-1,0)$. For the second pair, $q(2,\sqrt{3}) = 4 - 3 = 1 = q(-2,\sqrt{3})$. Now $\mathbf{u} - \mathbf{u}' = (4,0)$, with $q = 16 \neq 0$, so $\tau_{\mathbf{e}_1}$ again works: it negates the first coordinate and fixes the second.

14.8.6.3 C. Going deeper

Exercise. C1: Inertia from cancellation

Work over $\mathbb{R}$ and let V be an n-dimensional space with a nondegenerate symmetric form.

(a) Prove that if $\mathbb{R}^{p,q}$ denotes $\mathbb{R}^{p+q}$ with the form $x_1^2 + \cdots + x_p^2 - x_{p+1}^2 - \cdots - x_{p+q}^2$, then $\mathbb{R}^{p,q} \cong \mathbb{R}^{p',q'}$ forces $p = p'$ and $q = q'$, using only Corollary 14.8.4 and Proposition 14.7.7.

(b) Explain in one sentence how this compares with the proof of Theorem 14.5.1 in Section 5.

Solution.

(a) Suppose $\mathbb{R}^{p,q} \cong \mathbb{R}^{p',q'}$. Equal dimension gives $p+q = p'+q' = n$. By Proposition 14.7.7 (a) the Witt index is an isometry invariant, so $\min(p,q) = \min(p',q') = k$. By Corollary 14.8.5 each space is k hyperbolic planes orthogonal to an anisotropic part, and

those anisotropic parts are isometric. The anisotropic part of $\mathbb{R}^{p,q}$ is positive definite of dimension $p-q$ when $p \geq q$ and negative definite of dimension $q-p$ when $q > p$. A positive definite form is never isometric to a negative definite one of the same dimension, since $q > 0$ on one and $q < 0$ on the other away from 0 — unless both dimensions are 0. So the two anisotropic parts have the same dimension and the same sign, giving $p - q = p' - q'$. With $p + q = p' + q'$ this forces $p = p'$ and $q = q'$.

(b) Section 5 proved this directly, by producing a positive subspace and a nonpositive subspace whose dimensions are too large to intersect trivially; the route here instead invokes the whole Witt machinery, so it is longer but shows that inertia is the real instance of a statement that holds over every field of characteristic $\neq 2$.

Exercise. C2: The isometry group acts transitively

Let β be a nondegenerate symmetric form on V, with $\operatorname{char} F \neq 2$, and let G be the group of isometries of V.

(a) Prove that if $q(u) = q(u') \neq 0$ then some $\varphi \in G$ has $\varphi(u) = u'$.

(b) Prove that if u and u' are both isotropic and nonzero, then some $\varphi \in G$ has $\varphi(u) = u'$. *Hint: an isotropic line is a one-dimensional subspace on which the form is zero.*

(c) Deduce that G acts transitively on each "level set" $\{v \neq 0 : q(v) = c\}$ for $c \neq 0$, and on the set of nonzero isotropic vectors.

Solution.

(a) This is Lemma 14.8.2, which already produces such an isometry as a product of at most two reflections. Alternatively, $\operatorname{span}(u) \to \operatorname{span}(u')$, $u \mapsto u'$, is an isometry of one-dimensional subspaces because $q(u) = q(u')$, so Theorem 14.8.3 extends it.

(b) The map $\sigma\colon \operatorname{span}(u) \to \operatorname{span}(u')$ with $\sigma(au) = au'$ is a linear bijection, and

$$\beta(\sigma(au), \sigma(bu)) = ab\, q(u') = 0 = ab\, q(u) = \beta(au, bu),$$

so it is an isometry of the restricted forms — both of which are zero. By Theorem 14.8.3 it extends to $\varphi \in G$. Note the restricted forms here are **degenerate**, which is exactly the case Step 2 of the proof was built for.

(c) Immediate from (a) and (b): any two members of such a set are carried to one another by some $\varphi \in G$. In particular, over $\mathbb{R}^{1,3}$ the isometry group moves any nonzero null vector to any other, which is the statement that the light cone has no distinguished direction.

Exercise. C3: What extension does not say

(a) Give an example of a nondegenerate (V, β) and subspaces U, U' of the **same dimension** that are not isometric, and say why Theorem 14.8.3 does not apply.

(b) Let β be nondegenerate on V and let $U \subseteq V$ be totally isotropic with $\dim U = \operatorname{ind}(\beta)$. Prove that any two such U are carried onto one another by an isometry of V.

(c) Deduce that all maximal totally isotropic subspaces of a nondegenerate space have the same dimension, and compare with Proposition 14.7.2.

Solution.

(a) Take $V = \mathbb{R}^2$ with $q = x_1^2 - x_2^2$, $U = \operatorname{span}(\mathbf{e}_1)$, and $U' = \operatorname{span}(\mathbf{e}_2)$. Both are lines, but q is positive on $U \setminus \{\mathbf{0}\}$ and negative on $U' \setminus \{\mathbf{0}\}$, so no linear bijection can preserve the form. Theorem 14.8.3 has nothing to extend: its hypothesis is that an isometry $U \to U'$ is **given**, and here none exists.

(b) Let U and U' be totally isotropic of the same dimension k. Any linear bijection $\sigma\colon U \to U'$ is an isometry, since both restricted forms are identically zero, so $\beta(\sigma u, \sigma v) = 0 = \beta(u, v)$ for all $u, v \in U$. By Theorem 14.8.3, σ extends to an isometry φ of V, and $\varphi(U) = U'$.

(c) Let U be maximal among totally isotropic subspaces, and let U_1 be one of the largest dimension $m = \operatorname{ind}(\beta)$. If $\dim U < m$, choose any k-dimensional subspace $U_2 \subseteq U_1$ with $k = \dim U$; it is totally isotropic, so by (b) there is an isometry φ of V with $\varphi(U) = U_2$. Then $\varphi^{-1}(U_1)$ is totally isotropic of dimension $m > k$ and contains $\varphi^{-1}(U_2) = U$, contradicting maximality. So $\dim U = m$ for every maximal U. Proposition 14.7.2 says only $m \le n/2$; the present statement is sharper, since it says that no totally isotropic subspace can be maximal below the top dimension.

14.9 Alternating Forms and Darboux Bases

Almost every classification theorem so far in this chapter has carried the same asterisk — the exceptions being the crude invariants of Section 2, rank and the square class of the determinant, which need nothing of the field. Diagonalize a symmetric form: characteristic $\neq 2$. Split a form into a symmetric and an alternating half: characteristic $\neq 2$. Recover a form from its quadratic form: characteristic $\neq 2$. This section has no asterisk. Alternating forms admit a complete classification over **every** field, with no hypothesis on the characteristic at all, and the classification is as clean as a classification gets: one congruence class for each even rank.

Throughout, F is an arbitrary field, V is a finite-dimensional vector space over F, and β is an **alternating** bilinear form on V (Definition 7.2.2), meaning $\beta(v, v) = 0$ for **every** $v \in V$. Two facts from Section 3 are used constantly. First, β is then skew-symmetric, $\beta(u, v) = -\beta(v, u)$, by Theorem 14.3.3 (a) — and that direction needs no hypothesis on F. Second, by Proposition 14.3.2 (b) the matrix of β in any basis is skew-symmetric with **zero diagonal**, and we call such a matrix **alternating**. Symmetry of the situation is worth noting once: since $\beta(u, v) = 0$ if and only if $\beta(v, u) = 0$, "orthogonal" needs no left or right, and the radical $\operatorname{rad}(\beta)$ of Definition 14.1.3 is what one would want it to be from either side.

14.9.1 The two-dimensional model

A symmetric form has a one-dimensional model to build from: on a line $\operatorname{span}(v)$ it is $q(cv) = c^2 q(v)$, a scalar multiple of the squaring function, and Section 4 assembled a general symmetric form out of such lines. An alternating form offers nothing on a line at all. For any v and any scalars c, d,

$$\beta(cv, dv) = cd\,\beta(v, v) = 0,$$

so β restricted to any line is zero. The one-dimensional pieces are all identical and all trivial. The smallest place where an alternating form has something to say is a plane, and there is essentially only one thing it can say there.

Example. Alternating Forms on a Plane

Describe all alternating bilinear forms on F^2, and decide which are nondegenerate.

Solution. An alternating 2×2 matrix has zero diagonal and $a_{21} = -a_{12}$, so it is

$$\mathbf{A} = \begin{pmatrix} 0 & a \\ -a & 0 \end{pmatrix}$$

for a single scalar $a \in F$, and every such matrix is alternating. By Theorem 14.1.1 the corresponding form is $\beta(\mathbf{x}, \mathbf{y}) = a(x_1y_2 - x_2y_1)$. Its determinant is a^2, so by Proposition 14.1.2 the form is nondegenerate exactly when $a \neq 0$.

When $a \neq 0$, replacing the basis vector $\mathbf{e}_2$ by $a^{-1}\mathbf{e}_2$ replaces a by 1. So up to a change of basis there are exactly **two** alternating forms on F^2: the zero form and the one with matrix $\left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)$, whatever the field.

That last matrix is the atom of the theory, and the pair of vectors producing it deserves a name. It is the alternating counterpart of the hyperbolic plane of Section 7: a two-dimensional space carrying a nondegenerate form on which the associated quadratic function vanishes identically. Here the vanishing is automatic, because *every* alternating form vanishes on the diagonal, so the whole content is in the pairing.

A hyperbolic pair is two vectors that pair to 1 *and to nothing else.*

Definition 14.9.1. Hyperbolic Pair

Let β be an alternating form on V. A **hyperbolic pair** for β is an ordered pair (u, w) of vectors of V with

$$\beta(u, w) = 1.$$

Its **span** $H = \operatorname{span}(u, w)$ is called a **hyperbolic plane** for β.

The definition looks as though it asks for one condition, but it quietly supplies three more. Alternation already gives $\beta(u, u) = \beta(w, w) = 0$, and skewness gives $\beta(w, u) = -1$. So the matrix of β restricted to H in the ordered basis (u, w) is exactly $\left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)$. Also, u and w are automatically linearly independent: if $w = cu$ then $\beta(u, w) = c\,\beta(u, u) = 0 \neq 1$, and $u \neq 0$ for the same reason. So H really is a plane, $\dim H = 2$, and β restricted to H is nondegenerate.

Warning. **An alternating form is never diagonalizable, except when it is zero.** Section 4 found an orthogonal basis for every symmetric form in characteristic $\neq 2$; no such basis exists here. If $\mathcal{B}$ is any basis and $[\beta]_{\mathcal{B}}^{\mathcal{B}}$ is diagonal, then its off-diagonal entries vanish by assumption and its diagonal entries vanish because β is alternating, so $[\beta]_{\mathcal{B}}^{\mathcal{B}} = \mathbf{O}$ and $\beta = 0$. The standard form below is a block form, and it has to be: the 2×2 block is irreducible.

14.9.2 Splitting off a hyperbolic pair

The move that drives the whole section is the one Section 7 used for isotropic vectors: find a small nondegenerate piece, and cut the space along it. Here the small piece is a hyperbolic plane, and it can always be found as soon as β is not identically zero.

Lemma 14.9.1. Peeling Off a Hyperbolic Plane

Let β be an alternating form on a finite-dimensional V, and suppose $\beta \neq 0$.

(a) There is a hyperbolic pair (u, w) for β.

(b) For any such pair, with $H = \operatorname{span}(u, w)$ and

$$H^{\perp_\beta} = \{\, v \in V : \beta(u, v) = \beta(w, v) = 0 \,\},$$

we have $V = H \oplus H^{\perp_\beta}$ and $\dim H^{\perp_\beta} = \dim V - 2$.

(c) $\operatorname{rad}(\beta) = \operatorname{rad}(\beta|_{H^{\perp_\beta}})$. In particular, if β is nondegenerate then so is its restriction to $H^{\perp_\beta}$.

Idea. For (a), the hypothesis $\beta \neq 0$ hands us two vectors pairing to something nonzero; divide one of them by that scalar. For (b), the two conditions cutting out $H^{\perp_\beta}$ are two linear conditions, so the count should be $\dim V - 2$, and Rank–Nullity will say so once we check that the two conditions are independent — which is exactly what $\beta(u, w) = 1$ provides. Part (c) is the bookkeeping that lets the induction in the theorem keep its hypothesis.

Proof.

(a) Since $\beta \neq 0$ there are $u_0, w_0 \in V$ with $c := \beta(u_0, w_0) \neq 0$. Put $u = u_0$ and $w = c^{-1} w_0$, which is legitimate because $c \neq 0$ and F is a field. Then $\beta(u, w) = c^{-1}\beta(u_0, w_0) = 1$ by linearity in the second argument.

(b) Define $\varphi\colon V \to F^2$ by $\varphi(v) = (\beta(u, v),\, \beta(w, v))$. It is linear, because β is linear in its second argument, and $\ker \varphi = H^{\perp_\beta}$ by the definition of $H^{\perp_\beta}$. It is surjective:

$$\varphi(w) = (\beta(u, w), \beta(w, w)) = (1, 0), \qquad \varphi(u) = (0, -1),$$

using $\beta(w, w) = \beta(u, u) = 0$ and $\beta(w, u) = -1$, and these two values span F^2. Hence $\operatorname{rank} \varphi = 2$ and $\dim H^{\perp_\beta} = \dim V - 2$ by Rank–Nullity (Theorem 2.3.1).
Next, $H \cap H^{\perp_\beta} = \{0\}$. Let $v = au + bw$ lie in $H^{\perp_\beta}$. Then $0 = \beta(u, v) = a\beta(u, u) + b\beta(u, w) = b$, and $0 = \beta(w, v) = a\beta(w, u) = -a$, so $a = b = 0$. Hence the sum $H + H^{\perp_\beta}$ is direct (Theorem 1.7.4), and by Theorem 1.7.3 its dimension is $2 + (\dim V - 2) = \dim V$, so it is all of V (Theorem 1.6.8).

(c) ($\supseteq$) Let $v \in H^{\perp_\beta}$ satisfy $\beta(z, v) = 0$ for every $z \in H^{\perp_\beta}$. Any $y \in V$ decomposes as $y = h + z$ with $h \in H$ and $z \in H^{\perp_\beta}$ by (b), and $\beta(h, v) = -\beta(v, h) = 0$ because $v \in H^{\perp_\beta}$ and h is a combination of u and w. Hence $\beta(y, v) = 0$ for every $y \in V$, so $v \in \operatorname{rad}(\beta)$.

($\subseteq$) Let $v \in \operatorname{rad}(\beta)$. Then in particular $\beta(u, v) = \beta(w, v) = 0$, so $v \in H^{\perp_\beta}$, and $\beta(z, v) = 0$ for every z at all, so v lies in the radical of the restriction. This proves the lemma.

■

Notice what the proof did **not** use: it never divided by 2, never distinguished β from $-\beta$, and never needed a vector with $\beta(v,v) \neq 0$ — which is fortunate, since no such vector exists. Compare the corresponding step for symmetric forms, Lemma 14.4.1 in Section 4: there the whole difficulty was producing a vector on which the quadratic form did not vanish, and producing it cost the hypothesis $\operatorname{char} F \neq 2$. Here the quadratic form vanishes everywhere and we do not care.

14.9.3 The standard form

With the lemma in hand the classification is an induction that peels two dimensions at a time.

Theorem 14.9.2. Standard Form of an Alternating Form

Let F be **any** field, let V be a vector space over F with $\dim V = n$, and let β be an alternating bilinear form on V. Then there are an integer $m \geq 0$ with $2m \leq n$ and a basis

$$\mathcal{B} = (u_1, \dots, u_m, w_1, \dots, w_m, z_1, \dots, z_{n-2m})$$

of V such that, for all i, j,

$$\beta(u_i, w_j) = \delta_{ij}, \qquad \beta(u_i, u_j) = \beta(w_i, w_j) = 0,$$

and $\beta(z_k, v) = 0$ for every k and every $v \in V$. Equivalently, in this basis

$$[\beta]_{\mathcal{B}}^{\mathcal{B}} = \mathbf{\Omega}_{2m} \oplus \mathbf{O}_{n-2m}, \qquad \mathbf{\Omega}_{2m} \coloneqq \begin{pmatrix} \mathbf{O} & \mathbf{I}_m \\ -\mathbf{I}_m & \mathbf{O} \end{pmatrix}. \tag{14.9.1}$$

In particular:

(a) every alternating $\mathbf{A} \in M_n(F)$ is congruent to $\mathbf{\Omega}_{2m} \oplus \mathbf{O}_{n-2m}$ for exactly one m;
(b) if β is nondegenerate then $n = 2m$ is **even** and $[\beta]_{\mathcal{B}}^{\mathcal{B}} = \mathbf{\Omega}_n$.

Idea. Induction on n, two dimensions at a time. ① If $\beta = 0$, take $m = 0$ and any basis; this is the base case and also the case that stops the recursion. ② Otherwise Lemma 14.9.1 produces a hyperbolic pair (u_1, w_1) and splits $V = H \oplus H^{\perp_\beta}$ with $\dim H^{\perp_\beta} = n - 2$; apply the induction hypothesis to β restricted to $H^{\perp_\beta}$, and put the pieces together. ③ Finally reorder the basis so that all the u's come before all the w's, which turns a string of 2×2 blocks into the one block matrix displayed. The radical, by part (c) of the lemma, is exactly what the last stage of the recursion leaves behind.

Proof. **Existence, by induction on $n = \dim V$.** For $n = 0$ and $n = 1$ every alternating form is zero: in dimension 1, $\beta(cv, dv) = cd\,\beta(v,v) = 0$. Take $m = 0$.

Let $n \geq 2$ and assume the statement for all spaces of dimension less than n.

Case 1: $\beta = 0$. Take $m = 0$ and any basis of V; every required equation reads $0 = 0$.

Case 2: $\beta \neq 0$. By Lemma 14.9.1 (a) there is a hyperbolic pair (u_1, w_1), and by (b) we have $V = H \oplus W$ with $H = \operatorname{span}(u_1, w_1)$ and $W = H^{\perp_\beta}$ of dimension $n - 2$. The restriction $\beta|_W$ is again alternating, being a restriction of an alternating form, so the induction hypothesis

applies to it: there is a basis

$$(u_2, \dots, u_m,\ w_2, \dots, w_m,\ z_1, \dots, z_{n-2m})$$

of W in which $\beta|_W(u_i, w_j) = \delta_{ij}$, all other pairings among the u's and w's vanish, and each z_k lies in $\operatorname{rad}(\beta|_W)$.

Putting u_1 and w_1 in front gives a basis of $V = H \oplus W$. The required equations hold: within H they are the definition of a hyperbolic pair together with alternation; within W they are the induction hypothesis; and across the two, every pairing of u_1 or w_1 with a vector of W vanishes because $W = H^{\perp_\beta}$ and β is skew-symmetric. Finally each z_k lies in $\operatorname{rad}(\beta|_W) = \operatorname{rad}(\beta)$ by Lemma 14.9.1 (c), which is the last assertion.

The matrix. Order the basis as $\mathcal{B} = (u_1, \dots, u_m, w_1, \dots, w_m, z_1, \dots, z_{n-2m})$. By Definition 14.1.2 the (i, j) entry of $[\beta]_{\mathcal{B}}^{\mathcal{B}}$ is the pairing of the i-th and j-th basis vectors, and the equations above say precisely that the top-left $m \times m$ block and the w-w block are $\mathbf{O}$, the u-w block is $\mathbf{I}_m$, the w-u block is $-\mathbf{I}_m$, and every row and column indexed by a z_k is zero. That is Eq. 14.9.1.

(a). Given alternating $\mathbf{A} \in M_n(F)$, let $\beta(\mathbf{x}, \mathbf{y}) = \mathbf{x}^{\top}\mathbf{A}\mathbf{y}$ on F^n; it is alternating by Proposition 14.3.2 (b). The basis just constructed gives $\mathbf{P}^{\top}\mathbf{A}\mathbf{P} = \mathbf{\Omega}_{2m} \oplus \mathbf{O}_{n-2m}$ with $\mathbf{P}$ the matrix whose columns are the basis vectors, by Theorem 14.2.1, so $\mathbf{A} \simeq \mathbf{\Omega}_{2m} \oplus \mathbf{O}_{n-2m}$. Uniqueness of m: congruence preserves rank (Theorem 14.2.3), and $\operatorname{rank}(\mathbf{\Omega}_{2m} \oplus \mathbf{O}_{n-2m}) = 2m$ because $\mathbf{\Omega}_{2m}$ is invertible, its square being $-\mathbf{I}_{2m}$. So $2m = \operatorname{rank}\mathbf{A}$ is determined by $\mathbf{A}$.

(b). If β is nondegenerate then $\operatorname{rad}(\beta) = \{0\}$ by Definition 14.1.4, so there are no z_k, that is $n = 2m$. This proves the theorem.

■

Definition 14.9.2. Symplectic Basis

A basis of V as in Theorem 14.9.2 is a **symplectic basis**, or a **Darboux basis**, for the alternating form β. A nondegenerate alternating form on a finite-dimensional space is called a **symplectic form**, and the pair (V, β) a **symplectic vector space**.

Now read the hypotheses of the theorem again, and notice what is missing. There is no "char $F \neq 2$", and there is no place one could be inserted, because no step of the proof divided by 2. Compare the statements this chapter has accumulated:

- Theorem 14.3.3: skew implies alternating only when $\operatorname{char} F \neq 2$.
- Theorem 14.3.4: the splitting into halves exists only when $\operatorname{char} F \neq 2$.
- Theorem 14.3.5: a symmetric form is recovered from its quadratic form only when $\operatorname{char} F \neq 2$.
- Theorem 14.4.2: a symmetric form has an orthogonal basis only when $\operatorname{char} F \neq 2$.
- Theorem 14.9.2: no hypothesis.

Over $\mathbb{F}_2$, where every one of the first four statements is false, the last one is true and gives a complete classification. The reason is visible in the proof: the symmetric theory needs a vector with $q(v) \neq 0$, and that vector is manufactured by a polarization identity that divides by 2. The alternating theory needs a *pair* with $\beta(u, w) \neq 0$, and a pair is handed over by the assumption $\beta \neq 0$ itself, free of charge. Symmetric forms are hard in characteristic 2; alternating forms are not hard anywhere.

Quick check. Over $\mathbb{F}_2$ we have $-1 = 1$, so $\boldsymbol{\Omega}_{2m} = \left(\begin{smallmatrix} \mathbf{O} & \mathbf{I}_m \\ \mathbf{I}_m & \mathbf{O} \end{smallmatrix}\right)$, which is a **symmetric** matrix. Does that contradict the warning above, that an alternating form is never diagonalizable?

Solution. No. Over $\mathbb{F}_2$ every alternating matrix is symmetric, since skew-symmetric and symmetric are the same condition there (Theorem 14.3.3 and the warning after it). What distinguishes alternating from merely symmetric over $\mathbb{F}_2$ is the zero diagonal, and $\boldsymbol{\Omega}_{2m}$ does have a zero diagonal. The warning says a **diagonal** alternating matrix must be $\mathbf{O}$, and $\boldsymbol{\Omega}_{2m}$ is symmetric but not diagonal. Concretely, over $\mathbb{F}_2$ the matrix $\left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$ is the Gram matrix of the alternating form $\beta(\mathbf{x}, \mathbf{y}) = x_1 y_2 + x_2 y_1$, and no change of basis makes it diagonal.

14.9.4 Running the algorithm

The proof is an algorithm, and it is short enough to carry out by hand. Each round of the induction does three things: find a pair with $\beta(u, w) \neq 0$, rescale to make the value 1, and then correct every remaining basis vector so that it becomes orthogonal to both. The correction is a formula worth recording. If $\beta(u, w) = 1$, put

$$v' := v - \beta(v, w)\, u + \beta(v, u)\, w. \tag{14.9.2}$$

Expanding and using $\beta(u, u) = \beta(w, w) = 0$, $\beta(u, w) = 1$, and $\beta(w, u) = -1$,

$$\begin{aligned} \beta(v', u) &= \beta(v, u) - \beta(v, w) \cdot 0 + \beta(v, u) \cdot (-1) = 0, \\ \beta(v', w) &= \beta(v, w) - \beta(v, w) \cdot 1 + \beta(v, u) \cdot 0 = 0, \end{aligned}$$

so $v' \in H^{\perp_\beta}$. It is the alternating analogue of subtracting a projection in Gram–Schmidt, with one difference: nothing is divided by a length, so it works over any field.

Example. A Symplectic Basis for a 4×4 Form

Let $\beta(\mathbf{x}, \mathbf{y}) = \mathbf{x}^{\top} \mathbf{A} \mathbf{y}$ on $\mathbb{Q}^4$, where

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 2 & 3 \\ -1 & 0 & 1 & 1 \\ -2 & -1 & 0 & 0 \\ -3 & -1 & 0 & 0 \end{pmatrix}.$$

Find a symplectic basis, and the matrix $\mathbf{P}$ with $\mathbf{P}^{\top} \mathbf{A} \mathbf{P} = \boldsymbol{\Omega}_4$.

Solution. The matrix is skew-symmetric with zero diagonal, so β is alternating (Proposition 14.3.2 (b)).

Round 1. The entry $a_{12} = \beta(\mathbf{e}_1, \mathbf{e}_2) = 1$ is already 1, so $(\mathbf{u}_1, \mathbf{w}_1) = (\mathbf{e}_1, \mathbf{e}_2)$ is a hyperbolic pair and no rescaling is needed. Correct $\mathbf{e}_3$ and $\mathbf{e}_4$ by Eq. 14.9.2, reading the needed values off the matrix: $\beta(\mathbf{e}_3, \mathbf{e}_2) = a_{32} = -1$, $\beta(\mathbf{e}_3, \mathbf{e}_1) = a_{31} = -2$, $\beta(\mathbf{e}_4, \mathbf{e}_2) = a_{42} = -1$, $\beta(\mathbf{e}_4, \mathbf{e}_1) =$

$a_{41} = -3$. So

$$\mathbf{v}_3 = \mathbf{e}_3 + \mathbf{e}_1 - 2\mathbf{e}_2 = (1, -2, 1, 0),$$
$$\mathbf{v}_4 = \mathbf{e}_4 + \mathbf{e}_1 - 3\mathbf{e}_2 = (1, -3, 0, 1).$$

Round 2. Both $\mathbf{v}_3$ and $\mathbf{v}_4$ pair to zero with $\mathbf{e}_1$ and $\mathbf{e}_2$, as the formula guarantees, so the recursion continues inside $\operatorname{span}(\mathbf{v}_3, \mathbf{v}_4)$. Compute $\mathbf{A}\mathbf{v}_4$ row by row with $\mathbf{v}_4 = (1, -3, 0, 1)$: row 1 gives $0 - 3 + 0 + 3 = 0$, row 2 gives $-1 + 0 + 0 + 1 = 0$, row 3 gives $-2 + 3 + 0 + 0 = 1$, row 4 gives $-3 + 3 + 0 + 0 = 0$. So $\mathbf{A}\mathbf{v}_4 = (0, 0, 1, 0)^{\top}$, and

$$\beta(\mathbf{v}_3, \mathbf{v}_4) = \mathbf{v}_3^{\top}(\mathbf{A}\mathbf{v}_4) = 1 \cdot 0 + (-2) \cdot 0 + 1 \cdot 1 + 0 \cdot 0 = 1.$$

The pair $(\mathbf{u}_2, \mathbf{w}_2) = (\mathbf{v}_3, \mathbf{v}_4)$ is therefore hyperbolic, and the recursion stops: the two remaining dimensions are used up.

The answer. The symplectic basis is $\mathcal{B} = (\mathbf{e}_1, \mathbf{v}_3, \mathbf{e}_2, \mathbf{v}_4)$, listing the $\mathbf{u}$'s first and then the $\mathbf{w}$'s. The matrix whose columns are these four vectors is

$$\mathbf{P} = \begin{pmatrix} 1 & 1 & 0 & 1 \\ 0 & -2 & 1 & -3 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \qquad \mathbf{P}^{\top}\mathbf{A}\mathbf{P} = \mathbf{\Omega}_4.$$

A check of two entries confirms the pattern: the $(1,3)$ entry of $\mathbf{P}^{\top}\mathbf{A}\mathbf{P}$ is $\beta(\mathbf{e}_1, \mathbf{e}_2) = 1$ and the $(2,4)$ entry is $\beta(\mathbf{v}_3, \mathbf{v}_4) = 1$, while the $(1,2)$ entry is $\beta(\mathbf{e}_1, \mathbf{v}_3) = a_{11} + a_{12}(-2) + a_{13} = 0 - 2 + 2 = 0$.

14.9.5 What the standard form classifies

The theorem is a classification, and it is worth reading off exactly what it classifies and how coarsely.

Corollary 14.9.3. Alternating Matrices Are Classified by Rank

Let F be any field and let $\mathbf{A}, \mathbf{B} \in M_n(F)$ be alternating.

(a) $\operatorname{rank}\mathbf{A}$ is **even**.

(b) $\mathbf{A} \simeq \mathbf{B}$ if and only if $\operatorname{rank}\mathbf{A} = \operatorname{rank}\mathbf{B}$.

(c) There are exactly $\lfloor n/2 \rfloor + 1$ congruence classes of alternating matrices in $M_n(F)$, one for each even rank $0, 2, 4, \ldots$ up to n.

(d) If n is odd, every alternating $\mathbf{A} \in M_n(F)$ is singular.

Proof.

(a) By Theorem 14.9.2 (a), $\mathbf{A} \simeq \mathbf{\Omega}_{2m} \oplus \mathbf{O}_{n-2m}$ for some m, and congruence preserves rank (Theorem 14.2.3), so $\operatorname{rank}\mathbf{A} = 2m$.

(b) ($\Rightarrow$) is Theorem 14.2.3. ($\Leftarrow$) If $\operatorname{rank}\mathbf{A} = \operatorname{rank}\mathbf{B} = 2m$ then by (a) and Theorem 14.9.2 both are congruent to $\mathbf{\Omega}_{2m} \oplus \mathbf{O}_{n-2m}$, and congruence is an equivalence relation (Proposition 14.2.2), so $\mathbf{A} \simeq \mathbf{B}$.

(c) By (a) and (b) the classes are in bijection with the even integers $2m$ satisfying $0 \leq$

$2m \le n$, and there are $\lfloor n/2 \rfloor + 1$ of those.

(d) By (a), $\operatorname{rank} \mathbf{A}$ is even and at most n; since n is odd, $\operatorname{rank} \mathbf{A} \le n-1 < n$, so $\mathbf{A}$ is not invertible.

■

Part (d) deserves a second look, because there is a one-line proof of it that works only half the time. If $\mathbf{A}^\top = -\mathbf{A}$ with n odd, then $\det \mathbf{A} = \det(\mathbf{A}^\top) = \det(-\mathbf{A}) = (-1)^n \det \mathbf{A} = -\det \mathbf{A}$, so $(1+1)\det \mathbf{A} = 0$ and $\det \mathbf{A} = 0$ — provided $1 + 1 \neq 0$. In characteristic 2 that argument collapses, and it collapses for a real reason: over $\mathbb{F}_2$ the matrix $\operatorname{diag}(1,0,0)$ is skew-symmetric (because skew means symmetric there) and has odd rank 1. Part (d) survives because it is a statement about *alternating* matrices, and the standard form proves it without ever touching the determinant.

Warning. **"Skew-symmetric" is not enough.** Every statement in Corollary 14.9.3 is about alternating matrices: skew-symmetric **with zero diagonal**. Over $\mathbb{F}_2$, $\mathbf{I}_2$ is skew-symmetric and has rank 2, which is harmless; but $\operatorname{diag}(1,0) \in M_2(\mathbb{F}_2)$ is skew-symmetric of rank 1, and $\mathbf{I}_3 \in M_3(\mathbb{F}_2)$ is skew-symmetric and invertible. Both would contradict the corollary if "skew-symmetric" could be substituted for "alternating". In characteristic $\neq 2$ the two words agree (Theorem 14.3.3 (b)) and the distinction can be forgotten; nowhere else.

14.9.6 The determinant, and the Pfaffian

Part (a) of Theorem 14.9.2 says that in even size an invertible alternating matrix is congruent to $\mathbf{\Omega}_n$. Congruence multiplies the determinant by a square (Theorem 14.2.5), so the determinant of an alternating matrix is heavily constrained.

Corollary 14.9.4. The Determinant of an Alternating Matrix Is a Square

Let F be any field and let $\mathbf{A} \in M_n(F)$ be alternating. Then $\det \mathbf{A}$ is a square in F: there is $c \in F$ with $\det \mathbf{A} = c^2$.

Proof. If $\mathbf{A}$ is singular then $\det \mathbf{A} = 0 = 0^2$. So assume $\mathbf{A}$ is invertible. Then $\operatorname{rank} \mathbf{A} = n$ is even by Corollary 14.9.3 (a), say $n = 2m$, and Theorem 14.9.2 (a) supplies an invertible $\mathbf{P}$ with $\mathbf{P}^\top \mathbf{A} \mathbf{P} = \mathbf{\Omega}_{2m}$. Taking determinants and using Theorem 14.2.5,

$$\det \mathbf{\Omega}_{2m} = (\det \mathbf{P})^2 \det \mathbf{A}.$$

Now $\det \mathbf{\Omega}_{2m} = 1$. Indeed, performing the m row swaps $R_i \leftrightarrow R_{m+i}$ for $i = 1, \dots, m$ turns $\mathbf{\Omega}_{2m}$ into $(-\mathbf{I}_m) \oplus \mathbf{I}_m$, so by Theorem 7.5.1 (a) and Theorem 7.8.4,

$$(-1)^m \det \mathbf{\Omega}_{2m} = \det\big((-\mathbf{I}_m) \oplus \mathbf{I}_m\big) = (-1)^m,$$

and $(-1)^m$ is invertible, so $\det \mathbf{\Omega}_{2m} = 1$. Hence $\det \mathbf{A} = (\det \mathbf{P})^{-2} = \big((\det \mathbf{P})^{-1}\big)^2$, a square, since $\det \mathbf{P} \neq 0$. This proves the corollary.

■

The corollary says a square root of $\det \mathbf{A}$ exists; it does not produce one. In fact there is a canonical choice. For every even n there is a polynomial Pf in the entries a_{ij} with $i < j$, with integer coefficients, called the **Pfaffian**, satisfying $\det \mathbf{A} = \mathrm{Pf}(\mathbf{A})^2$ for every alternating $\mathbf{A}$. Constructing it in general is a piece of multilinear algebra that we do not carry out, so we record only the name and the two smallest cases, both of which can be checked by expanding the determinant.

Example. The Pfaffian in Sizes 2 and 4

For $n = 2$,

$$\mathbf{A} = \begin{pmatrix} 0 & a \\ -a & 0 \end{pmatrix}, \qquad \det \mathbf{A} = a^2, \qquad \mathrm{Pf}(\mathbf{A}) = a.$$

For $n = 4$, write the six independent entries as $a = a_{12}$, $b = a_{13}$, $c = a_{14}$, $d = a_{23}$, $e = a_{24}$, $f = a_{34}$. Expanding the 4×4 determinant gives

$$\det \mathbf{A} = (af - be + cd)^2, \qquad \mathrm{Pf}(\mathbf{A}) = af - be + cd.$$

The three terms pair up $\{1,2\}\{3,4\}$, $\{1,3\}\{2,4\}$, and $\{1,4\}\{2,3\}$: the Pfaffian sums over the ways of splitting $\{1,2,3,4\}$ into two pairs, with a sign. For the matrix $\mathbf{A}$ of Example (A Symplectic Basis for a 4×4 Form) this reads $\mathrm{Pf}(\mathbf{A}) = 1 \cdot 0 - 2 \cdot 1 + 3 \cdot 1 = 1$, and indeed $\det \mathbf{A} = 1$.

Remark. The Pfaffian is not merely a square root: it is the one that behaves well under congruence, satisfying $\mathrm{Pf}(\mathbf{P}^{\top}\mathbf{A}\mathbf{P}) = \det(\mathbf{P})\,\mathrm{Pf}(\mathbf{A})$. That identity is the reason the Pfaffian matters. Section 10 records one of its consequences for the symplectic group, without proof and without leaning on it. We do not prove the identity either.

14.9.7 Symmetric against alternating

Put the two halves of the chapter side by side. Both are theories of a bilinear form under congruence; the answers could hardly be more different.

	Symmetric forms	Alternating forms
Field hypothesis	$\mathrm{char}\, F \neq 2$ for every theorem	none
Canonical form	diagonal, entries not canonical	$\mathbf{\Omega}_{2m} \oplus \mathbf{O}_{n-2m}$
Complete invariants	rank **and** more: over $\mathbb{R}$ the signature, over $\mathbb{Q}$ the discriminant and local data	rank alone
Classes in $M_n(\mathbb{R})$	$\frac{1}{2}(n+1)(n+2)$	$\lfloor n/2 \rfloor + 1$
Classes in $M_n(\mathbb{C})$	$n+1$	$\lfloor n/2 \rfloor + 1$
Classes in $M_n(\mathbb{Q})$	infinitely many already for $n = 1$	$\lfloor n/2 \rfloor + 1$
Possible ranks	every r with $0 \le r \le n$	even r only
Effect of the field	decisive	none

Two entries carry the point. Over $\mathbb{Q}$, the 1×1 symmetric matrices $[1], [2], [3], [5], \ldots$ are pairwise

noncongruent, because congruence in size 1 multiplies by a nonzero square (Theorem 14.2.5) and distinct squarefree positive integers lie in distinct classes of $\mathbb{Q}^\times/(\mathbb{Q}^\times)^2$. So the symmetric classification over $\mathbb{Q}$ is an infinite problem in number theory — Section 4 said as much. The alternating classification over $\mathbb{Q}$ has four classes in size 6, and the same four over $\mathbb{F}_2$, over $\mathbb{R}$, and over any field whatsoever.

The reason, one last time, is the two-dimensional model. A symmetric form on a line is $c \mapsto ac^2$, and rescaling changes a by a square, so the diagonal entry carries the field's arithmetic with it. An alternating form on a plane is $\left(\begin{smallmatrix} 0 & a \\ -a & 0 \end{smallmatrix}\right)$, and rescaling *one* of the two basis vectors changes a to ca for any $c \neq 0$ — an arbitrary nonzero scalar, not a square. There is nothing left for the field to remember.

14.9.8 Exercises

14.9.8.1 A. Check your understanding

Exercise. A1

(a) Define a hyperbolic pair for an alternating form, and say why its two vectors are automatically linearly independent.

(b) State Theorem 14.9.2, and say which hypothesis on F it needs.

(c) Determine whether the following is correct: every skew-symmetric matrix over every field has even rank. Justify your answer.

(d) Why can an alternating form never have a diagonal Gram matrix, unless the form is zero?

(e) How many congruence classes of alternating matrices are there in $M_5(\mathbb{R})$? In $M_5(\mathbb{F}_7)$?

Solution.

(a) A pair (u, w) with $\beta(u, w) = 1$ (Definition 14.9.1). If $w = cu$, then $\beta(u, w) = c\,\beta(u, u) = 0 \neq 1$, so $w \notin \operatorname{span}(u)$; and $u \neq 0$ since $\beta(0, w) = 0$. Two vectors, neither in the span of the other, are independent.

(b) Every alternating form on an n-dimensional V has a basis in which its matrix is $\mathbf{\Omega}_{2m} \oplus \mathbf{O}_{n-2m}$, with $2m = \operatorname{rank}\beta$; equivalently every alternating matrix is congruent to exactly one such matrix. **No** hypothesis on F is needed — in particular none on $\operatorname{char} F$.

(c) Incorrect. Over $\mathbb{F}_2$, $\operatorname{diag}(1, 0)$ is skew-symmetric, because $-1 = 1$ makes skew-symmetric and symmetric the same condition, and it has rank 1. The correct statement is about **alternating** matrices, Corollary 14.9.3 (a). In characteristic $\neq 2$ the two notions coincide (Theorem 14.3.3 (b)) and the statement is true.

(d) The off-diagonal entries of a diagonal matrix are 0 by definition, and the diagonal entries of the Gram matrix of an alternating form are $\beta(v_i, v_i) = 0$. So the whole matrix is $\mathbf{O}$, and then $\beta = 0$ by Theorem 14.1.1 (b).

(e) $\lfloor 5/2 \rfloor + 1 = 3$ in both cases, by Corollary 14.9.3 (c): the ranks $0, 2, 4$. The field makes no difference.

14.9.8.2 B. Practice

Exercise. B1: A symplectic basis

Let $\beta(\mathbf{x},\mathbf{y}) = \mathbf{x}^{\top}\mathbf{B}\mathbf{y}$ on $\mathbb{Q}^4$, with

$$\mathbf{B} = \begin{pmatrix} 0 & 0 & 1 & 1 \\ 0 & 0 & 2 & 1 \\ -1 & -2 & 0 & 0 \\ -1 & -1 & 0 & 0 \end{pmatrix}.$$

Find a symplectic basis for β and an invertible $\mathbf{P} \in M_4(\mathbb{Q})$ with $\mathbf{P}^{\top}\mathbf{B}\mathbf{P} = \boldsymbol{\Omega}_4$.

Solution. $\mathbf{B}$ is skew-symmetric with zero diagonal, hence alternating (Proposition 14.3.2 (b)).

Round 1. Here $\beta(\mathbf{e}_1,\mathbf{e}_2) = 0$, so the first two standard basis vectors are **not** a hyperbolic pair; this is the step the worked example did not need. Look for any nonzero entry: $\beta(\mathbf{e}_1,\mathbf{e}_3) = 1$. Take $\mathbf{u}_1 = \mathbf{e}_1$, $\mathbf{w}_1 = \mathbf{e}_3$.

Correct the other two by Eq. 14.9.2, using $\beta(\mathbf{e}_2,\mathbf{e}_3) = 2$, $\beta(\mathbf{e}_2,\mathbf{e}_1) = 0$, $\beta(\mathbf{e}_4,\mathbf{e}_3) = 0$, $\beta(\mathbf{e}_4,\mathbf{e}_1) = -1$:

$$\mathbf{v}_2 = \mathbf{e}_2 - 2\mathbf{e}_1 + 0\cdot\mathbf{e}_3 = (-2,1,0,0),$$
$$\mathbf{v}_4 = \mathbf{e}_4 - 0\cdot\mathbf{e}_1 - \mathbf{e}_3 = (0,0,-1,1).$$

Round 2. Compute $\mathbf{B}\mathbf{v}_4$ with $\mathbf{v}_4 = (0,0,-1,1)$: row 1 gives $-1+1 = 0$, row 2 gives $-2+1 = -1$, rows 3 and 4 give 0. So $\mathbf{B}\mathbf{v}_4 = (0,-1,0,0)^{\top}$ and

$$\beta(\mathbf{v}_2,\mathbf{v}_4) = \mathbf{v}_2^{\top}(\mathbf{B}\mathbf{v}_4) = (-2)\cdot 0 + 1\cdot(-1) = -1.$$

This is nonzero but not 1, so rescale: replace $\mathbf{v}_4$ by $\mathbf{w}_2 := -\mathbf{v}_4 = (0,0,1,-1)$, giving $\beta(\mathbf{v}_2,\mathbf{w}_2) = 1$. Set $\mathbf{u}_2 = \mathbf{v}_2$.

The answer. The symplectic basis is $(\mathbf{u}_1,\mathbf{u}_2,\mathbf{w}_1,\mathbf{w}_2) = (\mathbf{e}_1,\ (-2,1,0,0),\ \mathbf{e}_3,\ (0,0,1,-1))$, so

$$\mathbf{P} = \begin{pmatrix} 1 & -2 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & -1 \end{pmatrix}, \qquad \mathbf{P}^{\top}\mathbf{B}\mathbf{P} = \boldsymbol{\Omega}_4.$$

$\mathbf{P}$ is invertible, being block upper triangular with invertible diagonal blocks, and $\det\mathbf{P} = -1 \neq 0$.

Exercise. B2: Pfaffians and determinants

For each matrix below, compute the Pfaffian using the formula of Example (The Pfaffian in Sizes 2 and 4), and verify that its square is the determinant.

$$\mathbf{A}_1 = \begin{pmatrix} 0 & 2 & 0 & 1 \\ -2 & 0 & 1 & 0 \\ 0 & -1 & 0 & 3 \\ -1 & 0 & -3 & 0 \end{pmatrix}, \qquad \mathbf{A}_2 = \begin{pmatrix} 0 & 1 & 1 & 1 \\ -1 & 0 & 1 & 1 \\ -1 & -1 & 0 & 1 \\ -1 & -1 & -1 & 0 \end{pmatrix}.$$

Hence write down, for each, an invertible matrix congruence class in the sense of Corollary 14.9.3.

Solution. For $\mathbf{A}_1$: $a = 2$, $b = 0$, $c = 1$, $d = 1$, $e = 0$, $f = 3$, so $\mathrm{Pf}(\mathbf{A}_1) = af - be + cd = 6 - 0 + 1 = 7$ and $\det \mathbf{A}_1 = 49$. Expanding the determinant along the first row confirms 49.

For $\mathbf{A}_2$: $a = b = c = d = e = f = 1$, so $\mathrm{Pf}(\mathbf{A}_2) = 1 - 1 + 1 = 1$ and $\det \mathbf{A}_2 = 1$.

Both determinants are nonzero, so both matrices have rank 4, and by Corollary 14.9.3 (b) both lie in the single congruence class of invertible alternating matrices in $M_4(\mathbb{Q})$: each is congruent to $\mathbf{\Omega}_4$, and therefore $\mathbf{A}_1 \simeq \mathbf{A}_2$. Note that the two have different determinants, 49 and 1; congruence changes the determinant by a square (Theorem 14.2.5), and $49 = 7^2$.

Exercise. B3: Determine which of the following

Determine which of the following matrices are alternating, and for those that are, give the rank and the standard form $\mathbf{\Omega}_{2m} \oplus \mathbf{O}_{n-2m}$ it is congruent to. Justify your answer.

(a) $\left(\begin{smallmatrix} 0 & 5 \\ -5 & 0 \end{smallmatrix}\right)$ over $\mathbb{Q}$.
(b) $\left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$ over $\mathbb{Q}$.
(c) $\left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$ over $\mathbb{F}_2$.
(d) The 5×5 matrix with $a_{ij} = j - i$ over $\mathbb{Q}$.

Solution.

(a) Alternating: zero diagonal and $a_{21} = -a_{12}$. Its determinant is $25 \neq 0$, so the rank is 2 and it is congruent to $\mathbf{\Omega}_2$. (Explicitly, scaling the second basis vector by $1/5$ does it.)

(b) Not alternating. It is symmetric with $a_{12} = 1 \neq -1 = -a_{21}$, so it is not even skew-symmetric over $\mathbb{Q}$.

(c) Alternating. Over $\mathbb{F}_2$ we have $1 = -1$, so the matrix is skew-symmetric, and its diagonal is zero. Its determinant is $-1 = 1 \neq 0$, so the rank is 2 and it is congruent to $\mathbf{\Omega}_2 = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$ — it **is** $\mathbf{\Omega}_2$, since $-1 = 1$. Comparing (b) and (c): the same array of symbols is alternating over one field and not over another.

(d) Alternating: $a_{ii} = 0$ and $a_{ji} = i - j = -a_{ij}$. Its rank is even by Corollary 14.9.3 (a), and it is not zero, so the rank is 2 or 4. Every row is a combination of $(1, 2, 3, 4, 5)$ and $(1, 1, 1, 1, 1)$: indeed row i is $(j - i)_j = (1, 2, 3, 4, 5) - i(1, 1, 1, 1, 1)$. So the row space has dimension at most 2, and rows 1 and 2 are independent, giving rank exactly 2. Hence the matrix is congruent to $\mathbf{\Omega}_2 \oplus \mathbf{O}_3$. In particular it is singular, as Corollary 14.9.3 (d) requires for odd size.

14.9.8.3 C. Going deeper

Exercise. C1: Lagrangian subspaces

Let β be a symplectic form on V with $\dim V = 2m$ (Definition 14.9.2), and call a subspace $U \subseteq V$ **totally isotropic** if $\beta(u, u') = 0$ for all $u, u' \in U$.

(a) Prove that $\operatorname{span}(u_1, \dots, u_m)$ is totally isotropic, for any symplectic basis.

(b) Prove that every totally isotropic subspace satisfies $\dim U \le m$.

(c) Deduce that a symplectic form on a space of dimension $2m$ has a totally isotropic subspace of dimension exactly m, and no larger one. (These are the **Lagrangian** subspaces.)

Hint for (b): compare U with $U^{\perp_\beta}$ of Exercise (C2: Orthogonal complements for a form).

Solution.

(a) By Theorem 14.9.2, $\beta(u_i, u_j) = 0$ for all i, j. A general pair of elements of the span is $\sum_i c_i u_i$ and $\sum_j d_j u_j$, and bilinearity gives $\sum_{i,j} c_i d_j \beta(u_i, u_j) = 0$.

(b) Let U be totally isotropic. By definition $\beta(u, v) = 0$ for all $u \in U$ and all $v \in U$, so $U \subseteq U^{\perp_\beta}$. Since β is nondegenerate, $\dim U^{\perp_\beta} = 2m - \dim U$ by Exercise (C2: Orthogonal complements for a form) (b). Therefore

$$\dim U \le 2m - \dim U,$$

that is $2 \dim U \le 2m$. Both sides are integers, so $\dim U \le m$. (No division by 2 inside the field occurs here: the inequality is between integers, not field elements, so the argument is valid in every characteristic.)

(c) The subspace of (a) has dimension m, since the u_i are part of a basis, and it is totally isotropic; by (b) no totally isotropic subspace has dimension greater than m.

Exercise. C2: The inverse of an alternating matrix

Let F be any field.

(a) Prove that every alternating matrix of **odd** size over F is singular, without using determinants.

(b) Let $\mathbf{A} \in M_n(F)$ be alternating and invertible. Prove that $\mathbf{A}^{-1}$ is alternating.

Hint for (b): the adjugate. The (i, i) entry of $\operatorname{adj} \mathbf{A}$ is a minor of $\mathbf{A}$ of size $n - 1$.

Solution.

(a) This is Corollary 14.9.3 (d): by Theorem 14.9.2 the rank is even, and an even number that is at most an odd n is at most $n - 1$, so the matrix is not invertible. The proof of the standard form uses no determinants.

(b) Since $\mathbf{A}$ is invertible, Theorem 7.6.2 gives $\mathbf{A}^{-1} = (\det \mathbf{A})^{-1} \operatorname{adj} \mathbf{A}$. First, skewness: $(\mathbf{A}^{-1})^\top = (\mathbf{A}^\top)^{-1} = (-\mathbf{A})^{-1} = -\mathbf{A}^{-1}$, using Theorem 0.9.5 and the fact that inversion turns $-\mathbf{A} = (-1)\mathbf{A}$ into $(-1)^{-1}\mathbf{A}^{-1} = -\mathbf{A}^{-1}$.

Now the diagonal. By Definition 7.6.2 the (i, i) entry of $\operatorname{adj} \mathbf{A}$ is the cofactor $C_{ii} = (-1)^{2i} M_{ii} = M_{ii}$, the determinant of the matrix obtained from $\mathbf{A}$ by deleting row i and column i. Deleting a row and the matching column of an alternating matrix leaves an alternating matrix, since the surviving entries keep their positions relative to the diagonal. Its size is $n - 1$, and n is even by Corollary 14.9.3 (a) applied to the invertible $\mathbf{A}$; so $n - 1$ is odd and $M_{ii} = 0$ by (a). Hence every diagonal entry of $\mathbf{A}^{-1}$ is 0, and with skewness this

makes $\mathbf{A}^{-1}$ alternating.

Exercise. C3: Counting over a finite field

Let q be a power of an **odd** prime.

(a) How many alternating bilinear forms are there on $\mathbb{F}_q^{2m}$? How many symmetric ones?

(b) How many of the alternating forms on $\mathbb{F}_q^2$ are nondegenerate, and how many congruence classes do they fall into?

(c) How many of the symmetric forms on $\mathbb{F}_q^1$ are nondegenerate, and how many congruence classes do they fall into? *Hint: the squaring map* $\mathbb{F}_q^\times \to \mathbb{F}_q^\times$ *has kernel* $\{1, -1\}$.

(d) Explain in one or two sentences why the answers to (b) and (c) differ, in terms of what a change of basis can do to the single free parameter in each case.

Solution.

(a) By Theorem 14.1.1 (c), forms correspond bijectively to matrices, so we count matrices. An alternating matrix of size $n = 2m$ is determined freely by its entries above the diagonal, of which there are $\binom{n}{2} = m(2m-1)$; so there are $q^{m(2m-1)}$ alternating forms. A symmetric matrix is determined freely by its entries on and above the diagonal, $\binom{n}{2} + n = m(2m+1)$ of them, giving $q^{m(2m+1)}$.

(b) The alternating forms on $\mathbb{F}_q^2$ have matrices $\left(\begin{smallmatrix} 0 & a \\ -a & 0 \end{smallmatrix}\right)$, so there are q of them, and the nondegenerate ones are those with $a \neq 0$, of which there are $q-1$. By Corollary 14.9.3 (b) all of these have rank 2 and so lie in a **single** congruence class.

(c) A symmetric form on $\mathbb{F}_q^1$ has matrix $[d]$ and is nondegenerate exactly when $d \neq 0$, so there are $q-1$ of them. Congruence in size 1 replaces d by $p^2 d$ with $p \neq 0$ (Definition 14.2.1), so the classes are the cosets of the subgroup of squares in $\mathbb{F}_q^\times$. The squaring homomorphism $\mathbb{F}_q^\times \to \mathbb{F}_q^\times$ has kernel $\{x : x^2 = 1\} = \{1, -1\}$, which has **two** elements because q is odd; so its image has index 2, and there are exactly **two** congruence classes. For $q = 5$, for instance, the squares are $\{1, 4\}$ and the nonsquares $\{2, 3\}$, and $[1] \not\simeq [2]$.

(d) In the alternating case a change of basis may replace a by ca for **any** $c \neq 0$, because the two slots of $\beta(u, w)$ hold different basis vectors and only one of them need be rescaled. In the symmetric case the parameter sits in $\beta(v, v)$, where both slots hold the same vector, so rescaling changes d to $c^2 d$ — by a **square** only. The square class then survives as an invariant, and over $\mathbb{F}_q$ it has two values.

14.10 Groups Defined by Forms

Chapter 11 defined the orthogonal and unitary groups by a sentence: they consist of the operators that preserve the inner product. Read that sentence again and notice that positivity never appears in it. "Preserves $\langle \cdot, \cdot \rangle$" makes sense for any bilinear or Hermitian form at all, and it always produces a group. This section takes the definition apart and puts it back together over an arbitrary form, so that the four families of **classical groups** — orthogonal, indefinite orthogonal, unitary, symplectic — become four instances of one construction, chosen by the four kinds of form this chapter has classified.

Everything is recalled, not reproved. $\mathrm{O}(n)$, $\mathrm{SO}(n)$, and $\mathrm{U}(n)$ were defined in Chapter 11 (Definition 11.7.2, Definition 11.7.3), shown there to be groups with the expected determinants (Proposition 11.7.4), and classified element by element in the real case in Chapter 12 (Theorem 12.6.1). None of that is repeated. What is new here is the pattern behind it.

14.10.1 The isometry group of a form

An operator preserves a form when applying it to both arguments changes nothing. This is the same idea as an isometry of an inner product space (Definition 11.7.1), with the norm removed from the statement because a general form has no norm.

The isometry group of a form is the set of invertible changes of coordinates that the form cannot detect.

Definition 14.10.1. Isometry Group of a Form

Let V be a finite-dimensional vector space over F and β a **nondegenerate** bilinear form on V (Definition 14.1.4). An operator $T \in \mathcal{L}(V)$ is an **isometry of** β if

$$\beta(Tu, Tv) = \beta(u, v) \qquad \text{for all } u, v \in V,$$

and the set of all of them is written $\mathrm{Isom}(\beta)$.

For a **nonsingular** $\mathbf{A} \in M_n(F)$, the corresponding matrix group is

$$\mathrm{Isom}(\mathbf{A}) \coloneqq \{\, \mathbf{P} \in M_n(F) : \mathbf{P}^{\top}\mathbf{A}\mathbf{P} = \mathbf{A} \,\}.$$

When $F = \mathbb{C}$ and $\mathbf{A}^{*} = \mathbf{A}$ is nonsingular Hermitian, $\mathrm{Isom}(\mathbf{A})$ means instead $\{\, \mathbf{P} : \mathbf{P}^{*}\mathbf{A}\mathbf{P} = \mathbf{A} \,\}$.

Read the clauses. The condition is required **for all** pairs, not just for a basis — though checking it on a basis is enough, since both sides are bilinear in u and in v. The matrix condition $\mathbf{P}^{\top}\mathbf{A}\mathbf{P} = \mathbf{A}$ is the matrix condition of Theorem 14.2.1 with the new Gram matrix required to be the *old* one: an isometry is a change of basis that leaves the Gram matrix alone. And the two versions match, because if $\mathbf{A} = [\beta]_{\mathcal{B}}^{\mathcal{B}}$ then $T \in \mathrm{Isom}(\beta)$ if and only if $[T]_{\mathcal{B}}^{\mathcal{B}} \in \mathrm{Isom}(\mathbf{A})$, by Theorem 14.1.1 (a) applied twice.

Two clauses need defending. First, **nondegeneracy**. Without it the definition is worthless: for $\mathbf{A} = \mathbf{O}$ the condition $\mathbf{P}^{\top}\mathbf{O}\mathbf{P} = \mathbf{O}$ holds for every $\mathbf{P}$, including $\mathbf{P} = \mathbf{O}$, and the "group" is all of $M_n(F)$, which has no inverses. Second, the definition does **not** require $\mathbf{P}$ to be invertible. It does not have to: invertibility comes free, as the next result records, and this is exactly what nondegeneracy buys.

Theorem 14.10.1. Isometry Groups Are Groups

Let $\mathbf{A} \in M_n(F)$ be nonsingular. Then every $\mathbf{P} \in \mathrm{Isom}(\mathbf{A})$ is invertible, and $\mathrm{Isom}(\mathbf{A})$ is a subgroup of $\mathrm{GL}_n(F)$. The same holds for $\mathrm{Isom}(\beta)$ inside the group of invertible operators on V, and for the Hermitian version.

Proof. Let $\mathbf{P} \in \mathrm{Isom}(\mathbf{A})$. Taking determinants in $\mathbf{P}^{\top}\mathbf{A}\mathbf{P} = \mathbf{A}$ and using Theorem 7.5.4 and Theorem 7.4.5,

$$(\det \mathbf{P})^2 \det \mathbf{A} = \det \mathbf{A}.$$

Since $\mathbf{A}$ is nonsingular, $\det \mathbf{A} \neq 0$ (Theorem 7.5.2), and canceling it in the field F gives $(\det \mathbf{P})^2 = 1$. Hence $\det \mathbf{P} \neq 0$ and $\mathbf{P}$ is invertible, again by Theorem 7.5.2.

Now the three subgroup checks. $\mathbf{I}_n^{\top}\mathbf{A}\mathbf{I}_n = \mathbf{A}$, so $\mathbf{I}_n \in \mathrm{Isom}(\mathbf{A})$. If $\mathbf{P}, \mathbf{Q} \in \mathrm{Isom}(\mathbf{A})$ then

$$(\mathbf{PQ})^{\top}\mathbf{A}(\mathbf{PQ}) = \mathbf{Q}^{\top}(\mathbf{P}^{\top}\mathbf{AP})\mathbf{Q} = \mathbf{Q}^{\top}\mathbf{AQ} = \mathbf{A},$$

using Theorem 0.9.5 for the first equality, so $\mathbf{PQ} \in \mathrm{Isom}(\mathbf{A})$. And if $\mathbf{P} \in \mathrm{Isom}(\mathbf{A})$, multiply $\mathbf{P}^{\top}\mathbf{AP} = \mathbf{A}$ on the left by $(\mathbf{P}^{\top})^{-1} = (\mathbf{P}^{-1})^{\top}$ and on the right by $\mathbf{P}^{-1}$, which is legitimate because $\mathbf{P}$ is invertible; this gives $\mathbf{A} = (\mathbf{P}^{-1})^{\top}\mathbf{A}\mathbf{P}^{-1}$, so $\mathbf{P}^{-1} \in \mathrm{Isom}(\mathbf{A})$. Associativity is inherited from $M_n(F)$.

For the Hermitian version replace $^{\top}$ by * throughout; the determinant step becomes $\overline{\det \mathbf{P}}\det \mathbf{P}\det \mathbf{A} = \det \mathbf{A}$, that is $|\det \mathbf{P}|^2 = 1$, which again forces $\det \mathbf{P} \neq 0$. The operator version is the matrix version read through a basis. This proves the theorem.

■

The proof is the one Chapter 11 gave for $\mathrm{O}(n)$ and $\mathrm{U}(n)$ in Proposition 11.7.4 (a), with $\mathbf{I}_n$ replaced by $\mathbf{A}$; nothing in it used what $\mathbf{A}$ was, only that it was nonsingular. That is the whole point of the section: the group structure never depended on positivity, on symmetry, or on the field.

14.10.2 Congruent forms give conjugate groups

Before listing the families, one observation makes the list a classification rather than a list of unrelated examples. Congruent matrices have the "same" isometry group, in the strongest sense available.

Proposition 14.10.2. Congruence Conjugates the Isometry Group

Let $\mathbf{A}, \mathbf{B} \in M_n(F)$ be nonsingular with $\mathbf{B} = \mathbf{Q}^{\top}\mathbf{AQ}$ for an invertible $\mathbf{Q}$ (Definition 14.2.1). Then

$$\mathrm{Isom}(\mathbf{B}) = \mathbf{Q}^{-1}\,\mathrm{Isom}(\mathbf{A})\,\mathbf{Q},$$

so the two groups are conjugate subgroups of $\mathrm{GL}_n(F)$; in particular they are isomorphic as groups, by $\mathbf{P} \mapsto \mathbf{QPQ}^{-1}$.

Proof. Let $\mathbf{P} \in M_n(F)$. Substituting $\mathbf{B} = \mathbf{Q}^{\top}\mathbf{AQ}$ and writing $\mathbf{R} = \mathbf{QPQ}^{-1}$,

$$\begin{aligned} \mathbf{P}^{\top}\mathbf{BP} = \mathbf{B} &\iff \mathbf{P}^{\top}\mathbf{Q}^{\top}\mathbf{AQP} = \mathbf{Q}^{\top}\mathbf{AQ} \\ &\iff (\mathbf{QPQ}^{-1})^{\top}\mathbf{A}(\mathbf{QPQ}^{-1}) = \mathbf{A}, \end{aligned}$$

where the second step multiplies on the left by $(\mathbf{Q}^{\top})^{-1}$ and on the right by $\mathbf{Q}^{-1}$ and uses Theorem 0.9.5; both operations are reversible because $\mathbf{Q}$ is invertible. So $\mathbf{P} \in \mathrm{Isom}(\mathbf{B})$ if and only if $\mathbf{R} = \mathbf{QPQ}^{-1} \in \mathrm{Isom}(\mathbf{A})$, which is the claimed equality of sets. Conjugation by $\mathbf{Q}$ is an isomorphism of groups, being a bijection that respects products.

■

So the isometry group depends only on the congruence class of the form, and this chapter has

already listed the congruence classes. Over $\mathbb{R}$, Corollary 14.5.2 says a nondegenerate symmetric form is congruent to exactly one $\mathbf{I}_{n_+} \oplus (-\mathbf{I}_{n_-})$; over $\mathbb{C}$, Corollary 14.4.4 says a nondegenerate symmetric form is congruent to $\mathbf{I}_n$ and nothing else; and Theorem 14.9.2 says a symplectic form is congruent to $\mathbf{\Omega}_{2m}$ and nothing else. Each classification theorem of this chapter therefore doubles as a list of groups.

14.10.3 The four families

Definition 14.10.2. The Indefinite Orthogonal and Symplectic Groups

For $p, q \geq 0$ with $p + q = n$, set $\mathbf{I}_{p,q} \coloneqq \mathbf{I}_p \oplus (-\mathbf{I}_q) \in M_n(\mathbb{R})$ and

$$\mathrm{O}(p,q) \coloneqq \mathrm{Isom}(\mathbf{I}_{p,q}) = \{\, \mathbf{P} \in M_n(\mathbb{R}) : \mathbf{P}^{\top}\mathbf{I}_{p,q}\mathbf{P} = \mathbf{I}_{p,q} \,\},$$

the **indefinite orthogonal group** of signature (p, q). For $m \geq 1$ and any field F, set

$$\mathrm{Sp}(2m, F) \coloneqq \mathrm{Isom}(\mathbf{\Omega}_{2m}) = \{\, \mathbf{P} \in M_{2m}(F) : \mathbf{P}^{\top}\mathbf{\Omega}_{2m}\mathbf{P} = \mathbf{\Omega}_{2m} \,\},$$

the **symplectic group**, where $\mathbf{\Omega}_{2m}$ is the standard alternating matrix of Eq. 14.9.1.

With Chapter 11's groups recalled, the four families line up as four choices of form.

Family	Form preserved	Kind of form	Section
$\mathrm{O}(n)$	$\mathbf{I}_n$ over $\mathbb{R}$	symmetric, positive definite	Ch 11 §07
$\mathrm{O}(p,q)$	$\mathbf{I}_{p,q}$ over $\mathbb{R}$	symmetric, signature (p,q)	§05
$\mathrm{U}(n)$	$\mathbf{I}_n$ over $\mathbb{C}$	Hermitian, positive definite	Ch 11 §07
$\mathrm{Sp}(2m, F)$	$\mathbf{\Omega}_{2m}$ over F	alternating, nondegenerate	§09

Two remarks on the list. $\mathrm{O}(n) = \mathrm{O}(n, 0)$, so the ordinary orthogonal group is the extreme case of the indefinite one; and $\mathrm{O}(p, q)$ and $\mathrm{O}(q, p)$ are conjugate, because negating a form does not change which maps preserve it, and $-\mathbf{I}_{p,q} = (-\mathbf{I}_p) \oplus \mathbf{I}_q$ becomes $\mathbf{I}_{q,p}$ after a permutation of the coordinates, which is a congruence. There is no Sp for odd size: a nondegenerate alternating form exists only in even dimension (Theorem 14.9.2 (b)).

Example. Members of Each Family

Verify the following memberships, with $\mathbf{\Omega}_{2m}$ as above.

(a) $\left(\begin{smallmatrix} \cosh t & \sinh t \\ \sinh t & \cosh t \end{smallmatrix}\right) \in \mathrm{O}(1,1)$ for every $t \in \mathbb{R}$.

(b) $\left(\begin{smallmatrix} \mathbf{M} & \mathbf{O} \\ \mathbf{O} & (\mathbf{M}^{\top})^{-1} \end{smallmatrix}\right) \in \mathrm{Sp}(2m, F)$ for every invertible $\mathbf{M} \in M_m(F)$.

(c) $\left(\begin{smallmatrix} \mathbf{I}_m & \mathbf{S} \\ \mathbf{O} & \mathbf{I}_m \end{smallmatrix}\right) \in \mathrm{Sp}(2m, F)$ if and only if $\mathbf{S}^{\top} = \mathbf{S}$.

Solution.

(a) With $\mathbf{A} = \mathbf{I}_{1,1} = \operatorname{diag}(1,-1)$, $c = \cosh t$, and $s = \sinh t$,

$$\begin{pmatrix} c & s \\ s & c \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}\begin{pmatrix} c & s \\ s & c \end{pmatrix} = \begin{pmatrix} c^2 - s^2 & 0 \\ 0 & s^2 - c^2 \end{pmatrix} = \mathbf{I}_{1,1},$$

using $\cosh^2 t - \sinh^2 t = 1$ and the symmetry of the outer matrix.

(b) Write $\mathbf{P} = \mathbf{M} \oplus \mathbf{N}$ with $\mathbf{N} = (\mathbf{M}^{\top})^{-1}$. Multiplying the blocks,

$$\mathbf{P}^{\top}\boldsymbol{\Omega}_{2m}\mathbf{P} = \begin{pmatrix} \mathbf{M}^{\top} & \mathbf{O} \\ \mathbf{O} & \mathbf{N}^{\top} \end{pmatrix}\begin{pmatrix} \mathbf{O} & \mathbf{N} \\ -\mathbf{M} & \mathbf{O} \end{pmatrix} = \begin{pmatrix} \mathbf{O} & \mathbf{M}^{\top}\mathbf{N} \\ -\mathbf{N}^{\top}\mathbf{M} & \mathbf{O} \end{pmatrix},$$

which equals $\boldsymbol{\Omega}_{2m}$ exactly when $\mathbf{M}^{\top}\mathbf{N} = \mathbf{I}_m$, that is $\mathbf{N} = (\mathbf{M}^{\top})^{-1}$.

(c) With $\mathbf{P} = \left(\begin{smallmatrix} \mathbf{I}_m & \mathbf{S} \\ \mathbf{O} & \mathbf{I}_m \end{smallmatrix}\right)$,

$$\mathbf{P}^{\top}\boldsymbol{\Omega}_{2m}\mathbf{P} = \begin{pmatrix} \mathbf{I}_m & \mathbf{O} \\ \mathbf{S}^{\top} & \mathbf{I}_m \end{pmatrix}\begin{pmatrix} \mathbf{O} & \mathbf{I}_m \\ -\mathbf{I}_m & -\mathbf{S} \end{pmatrix} = \begin{pmatrix} \mathbf{O} & \mathbf{I}_m \\ -\mathbf{I}_m & \mathbf{S}^{\top} - \mathbf{S} \end{pmatrix},$$

which equals $\boldsymbol{\Omega}_{2m}$ exactly when $\mathbf{S}^{\top} = \mathbf{S}$. Such a $\mathbf{P}$ is a **symplectic shear**; part (c) is the first sign that symmetric matrices appear inside the symplectic group, and the dimension count below will make that precise.

14.10.4 Determinants

For the orthogonal family the determinant argument is already inside the proof of Theorem 14.10.1, and it costs nothing to state it over any field.

Theorem 14.10.3. Determinants in the Classical Groups

(a) Let F be any field and $\mathbf{A} \in M_n(F)$ nonsingular. Then $\det \mathbf{P} = \pm 1$ for every $\mathbf{P} \in \operatorname{Isom}(\mathbf{A})$. In particular $\det \mathbf{P} = \pm 1$ for $\mathbf{P} \in \mathrm{O}(p,q)$, and both signs occur when $n \geq 1$.

(b) If $\mathbf{A} \in M_n(\mathbb{C})$ is Hermitian and nonsingular, then $|\det \mathbf{P}| = 1$ for every $\mathbf{P} \in \operatorname{Isom}(\mathbf{A})$; in particular for $\mathbf{P} \in \mathrm{U}(n)$.

(c) $\mathrm{Sp}(2, F) = \mathrm{SL}_2(F)$. In particular every $\mathbf{P} \in \mathrm{Sp}(2, F)$ has $\det \mathbf{P} = 1$ exactly, and -1 does not occur.

Proof.

(a) The proof of Theorem 14.10.1 established $(\det \mathbf{P})^2 = 1$. The polynomial $x^2 - 1 = (x-1)(x+1)$ has at most two roots in a field (Corollary 6.1.4), and 1 and -1 are roots, so $\det \mathbf{P} \in \{1, -1\}$. For $\mathrm{O}(p,q)$, the identity has determinant 1; and $\mathbf{D} := \operatorname{diag}(-1, 1, \ldots, 1)$ satisfies $\mathbf{D}^{\top}\mathbf{I}_{p,q}\mathbf{D} = \mathbf{I}_{p,q}$, since $\mathbf{D}$ is diagonal and squaring its entries gives $\mathbf{I}_n$, while $\det \mathbf{D} = -1$.

(b) Also established in that proof: $|\det \mathbf{P}|^2 = 1$, and a nonnegative real with square 1 is 1. This is Proposition 11.7.4 (b) for $\mathbf{A} = \mathbf{I}_n$.

(c) Write $\mathbf{J} = \boldsymbol{\Omega}_2 = \left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)$ and let $\mathbf{P} = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ be arbitrary. Multiplying out,

$$\mathbf{P}^{\top}\mathbf{J}\mathbf{P} = \begin{pmatrix} a & c \\ b & d \end{pmatrix}\begin{pmatrix} c & d \\ -a & -b \end{pmatrix} = \begin{pmatrix} 0 & ad-bc \\ bc-ad & 0 \end{pmatrix} = (\det \mathbf{P})\,\mathbf{J}.$$

Since $\mathbf{J} \neq \mathbf{O}$, the equation $\mathbf{P}^{\top}\mathbf{J}\mathbf{P} = \mathbf{J}$ holds if and only if $\det \mathbf{P} = 1$. This proves the theorem.

■

Warning. **Every symplectic matrix has determinant $+1$, not ± 1.** The general argument of part (a) applies to $\boldsymbol{\Omega}_{2m}$ as well and gives only $(\det \mathbf{P})^2 = 1$, so it leaves -1 open. It is nevertheless true that $\mathrm{Sp}(2m, F) \subseteq \mathrm{SL}_{2m}(F)$ for every m, so -1 never occurs. Part (c) proves this for $m = 1$; **the general case is not proved in this book.** The clean route runs through the Pfaffian of Section 9 and its transformation rule $\mathrm{Pf}(\mathbf{P}^{\top}\mathbf{A}\mathbf{P}) = \det(\mathbf{P})\,\mathrm{Pf}(\mathbf{A})$: applied to $\mathbf{A} = \boldsymbol{\Omega}_{2m}$, whose Pfaffian is 1, it turns $\mathbf{P}^{\top}\boldsymbol{\Omega}_{2m}\mathbf{P} = \boldsymbol{\Omega}_{2m}$ directly into $\det \mathbf{P} = 1$. We stated that rule without proof, so we state its consequence without proof too. Do not use the statement as a step in an argument elsewhere in this book; treat it as information about the wider subject, like the fact that the general polynomial of degree five has no solution by radicals.

There is a consequence worth noticing. $\mathrm{O}(n)$ has the proper subgroup $\mathrm{SO}(n)$ cut out by $\det = 1$, and Chapter 12 made heavy use of it: rotations against reflections. The symplectic group has no such subgroup, because — by the fact just recorded, which this book does not prove — the determinant condition is already satisfied by everything in it. A symplectic form has no notion of reflection to detect.

14.10.5 Counting parameters

How big is each group? "Big" has an honest answer in this book only as a count, so here is the count, and here is what it does and does not mean.

The equation $\mathbf{P}^{\top}\mathbf{A}\mathbf{P} = \mathbf{A}$ is a system of scalar equations in the n^2 unknown entries of $\mathbf{P}$. If the system consisted of k independent equations we would expect the solutions to form a family described by $n^2 - k$ free parameters, and that number is what everybody calls the **dimension** of the group. The key observation is that the n^2 equations are far from independent, because the matrix $\mathbf{P}^{\top}\mathbf{A}\mathbf{P} - \mathbf{A}$ is not an arbitrary matrix: it has the same symmetry type as $\mathbf{A}$, by Proposition 14.2.4. So the equations repeat, and only the independent ones should be counted.

- If $\mathbf{A}$ is **symmetric**, then $\mathbf{P}^{\top}\mathbf{A}\mathbf{P} - \mathbf{A}$ is symmetric, and a symmetric $n \times n$ matrix has $\frac{1}{2}n(n+1)$ independent entries. Expected dimension: $n^2 - \frac{1}{2}n(n+1) = \frac{1}{2}n(n-1)$.
- If $\mathbf{A}$ is **alternating**, then $\mathbf{P}^{\top}\mathbf{A}\mathbf{P} - \mathbf{A}$ is alternating, and an alternating $n \times n$ matrix has $\frac{1}{2}n(n-1)$ independent entries. Expected dimension: $n^2 - \frac{1}{2}n(n-1) = \frac{1}{2}n(n+1)$.
- If $\mathbf{A}$ is **Hermitian** over $\mathbb{C}$, count real parameters: $\mathbf{P}$ has $2n^2$ of them, and a Hermitian matrix has n real diagonal entries and $\frac{1}{2}n(n-1)$ complex entries above the diagonal, in total n^2 real conditions. Expected real dimension: $2n^2 - n^2 = n^2$.

Group	Entries in $\mathbf{P}$	Independent conditions	Dimension
$\mathrm{O}(n)$, $\mathrm{O}(p,q)$, $p+q=n$	n^2	$\frac{1}{2}n(n+1)$	$\frac{1}{2}n(n-1)$
$\mathrm{U}(n)$ (real parameters)	$2n^2$	n^2	n^2
$\mathrm{Sp}(2m,\mathbb{R})$	$4m^2$	$m(2m-1)$	$m(2m+1)$

Small cases are a sanity check. $\mathrm{O}(2)$ comes out at dimension 1, and Chapter 11 found that its elements are described by a single angle. $\mathrm{O}(3)$ comes out at 3, matching the axis-and-angle description of Corollary 12.6.2: two parameters for the axis direction, one for the angle. $\mathrm{U}(1)$ comes out at 1, and $\mathrm{U}(1)$ is the unit circle in $\mathbb{C}$. And $\mathrm{Sp}(2,\mathbb{R}) = \mathrm{SL}_2(\mathbb{R})$ comes out at 3, which is right: four entries subject to the single equation $ad-bc=1$.

Compare the orthogonal and symplectic rows of the table at $n=2m$. The orthogonal group is smaller than the symplectic group of the same size, by exactly $\frac{1}{2}n(n+1)-\frac{1}{2}n(n-1)=n$. The gap is the gap between symmetric and alternating matrices, which is the n diagonal entries — the same diagonal that Section 3 found separating the two notions in the first place.

Warning. **This count is a heuristic, not a theorem of this book.** "Dimension" here means the number of free parameters one expects a solution set to have, and the step from "k independent-looking equations" to "n^2-k parameters" is not justified by anything proved above; for a nonlinear system it can fail. Making it precise requires the notion of a smooth manifold and the implicit function theorem, which belong to analysis and differential geometry, not here. What is proved here is only the count of conditions. Every number in the table is nevertheless correct.

Quick check. The family $\mathrm{O}(p,q)$ is indexed by two integers, yet the table gives it a dimension depending only on $n=p+q$, not on how n splits. Does that mean $\mathrm{O}(3,0)$ and $\mathrm{O}(2,1)$ are the same group?

Solution. No. The count of conditions only sees that $\mathbf{I}_{p,q}$ is symmetric, not which symmetric matrix it is, so it cannot distinguish the two. They are genuinely different groups: $\mathrm{O}(3)=\mathrm{O}(3,0)$ consists of matrices all of whose entries have absolute value at most 1, since its columns are unit vectors, whereas $\mathrm{O}(2,1)$ contains matrices with arbitrarily large entries, as the next warning shows. Conjugate groups come from **congruent** forms (Proposition 14.10.2), and $\mathbf{I}_{3,0}$ and $\mathbf{I}_{2,1}$ are not congruent, because their signatures differ (Corollary 14.5.2).

14.10.6 Boundedness and its absence

Warning. $\mathrm{O}(p,q)$ **is not compact when** $p,q\geq 1$**, unlike** $\mathrm{O}(n)$**.** Compactness is a topological notion that this book does not develop, so the statement is **recorded here, not proved**. What can be seen without topology is the underlying reason, which is that the entries are

unbounded. Take the hyperbolic plane $\mathbf{A} = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$, a nondegenerate symmetric real form of signature $(1,1)$, hence congruent to $\mathbf{I}_{1,1}$. For every real $t \neq 0$,

$$\begin{pmatrix} t & 0 \\ 0 & t^{-1} \end{pmatrix}^{\top} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} t & 0 \\ 0 & t^{-1} \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

so $\operatorname{diag}(t, t^{-1})$ lies in $\operatorname{Isom}(\mathbf{A})$, and its entries are unbounded as $t \to \infty$. By Proposition 14.10.2 the same unbounded family sits inside $\mathrm{O}(1,1)$, and padding with an identity block puts it inside every $\mathrm{O}(p,q)$ with $p, q \geq 1$. By contrast every entry of a matrix in $\mathrm{O}(n)$ has absolute value at most 1, because each column is a unit vector for the Euclidean norm.

The same phenomenon shows up in $\mathrm{Sp}(2m, \mathbb{R})$, by Example (Members of Each Family) (b) with $\mathbf{M} = [t]$: the symplectic group is unbounded too. Only the two positive definite families, $\mathrm{O}(n)$ and $\mathrm{U}(n)$, are bounded, and positivity is exactly why — the form itself controls the size of the columns. Definiteness, which this chapter spent nine sections learning to do without, turns out to be what made Chapter 11's groups small.

Section 11 takes the indefinite case that physics uses, signature $(1,3)$, and works out what its isometry group does to the geometry of spacetime.

14.10.7 Exercises

14.10.7.1 A. Check your understanding

Exercise. A1

(a) Define $\operatorname{Isom}(\mathbf{A})$ for a nonsingular $\mathbf{A} \in M_n(F)$, and explain why its elements are automatically invertible.

(b) Which form does $\mathrm{O}(p,q)$ preserve, and which does $\mathrm{Sp}(2m, F)$ preserve?

(c) Determine whether the following is correct: if $\mathbf{A} \simeq \mathbf{B}$, then $\operatorname{Isom}(\mathbf{A}) = \operatorname{Isom}(\mathbf{B})$. Justify your answer.

(d) State what is known here about $\det \mathbf{P}$ for $\mathbf{P} \in \mathrm{Sp}(2m, F)$, distinguishing what is proved from what is only recorded.

(e) In the parameter count, what exactly does the word "dimension" mean, and what is not being claimed?

Solution.

(a) $\operatorname{Isom}(\mathbf{A}) = \{\mathbf{P} : \mathbf{P}^{\top}\mathbf{A}\mathbf{P} = \mathbf{A}\}$ (Definition 14.10.1). Taking determinants gives $(\det \mathbf{P})^2 \det \mathbf{A} = \det \mathbf{A}$, and $\det \mathbf{A} \neq 0$, so $(\det \mathbf{P})^2 = 1$ and $\det \mathbf{P} \neq 0$ (Theorem 14.10.1).

(b) $\mathrm{O}(p,q)$ preserves the real symmetric form with matrix $\mathbf{I}_{p,q} = \mathbf{I}_p \oplus (-\mathbf{I}_q)$, of signature (p,q). $\mathrm{Sp}(2m, F)$ preserves the standard alternating form with matrix $\mathbf{\Omega}_{2m}$ (Definition 14.10.2).

(c) Incorrect as stated: the two groups are **conjugate**, $\operatorname{Isom}(\mathbf{B}) = \mathbf{Q}^{-1} \operatorname{Isom}(\mathbf{A}) \mathbf{Q}$, not equal (Proposition 14.10.2). For a witness take $\mathbf{A} = \mathbf{I}_2$ and $\mathbf{B} = \operatorname{diag}(1,4) = \mathbf{Q}^{\top}\mathbf{A}\mathbf{Q}$ with $\mathbf{Q} = \operatorname{diag}(1,2)$. Then $\left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right) \in \mathrm{O}(2) = \operatorname{Isom}(\mathbf{A})$, but the corresponding element $\mathbf{Q}^{-1}\left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)\mathbf{Q} = \left(\begin{smallmatrix} 0 & 2 \\ -1/2 & 0 \end{smallmatrix}\right)$ of $\operatorname{Isom}(\mathbf{B})$ is not in $\mathrm{O}(2)$. What is always true is that the

two groups are isomorphic.

(d) Proved here: $(\det \mathbf{P})^2 = 1$, hence $\det \mathbf{P} = \pm 1$, for every m (Theorem 14.10.3 (a)); and $\det \mathbf{P} = 1$ exactly, for $m = 1$, since $\mathrm{Sp}(2, F) = \mathrm{SL}_2(F)$ (part (c)). Recorded but not proved: $\det \mathbf{P} = 1$ for every m, which follows from the transformation rule of the Pfaffian.

(e) The number of free parameters expected in the solution set of $\mathbf{P}^\top \mathbf{A} \mathbf{P} = \mathbf{A}$, computed as n^2 minus the number of independent scalar conditions. It is not claimed that the solution set really is a family described by that many free parameters; that statement needs the implicit function theorem and is not proved here.

14.10.7.2 B. Practice

Exercise. B1: Determine which of the following

For each matrix, determine whether it lies in $\mathrm{O}(2)$, in $\mathrm{O}(1,1)$, in $\mathrm{Sp}(2, \mathbb{R})$ — where the relevant forms are $\mathbf{I}_2$, $\mathbf{I}_{1,1} = \mathrm{diag}(1, -1)$, and $\mathbf{\Omega}_2$ — listing all that apply. Justify each answer.

(a) $\left(\begin{smallmatrix} 1 & 0 \\ 0 & -1 \end{smallmatrix}\right)$.

(b) $\left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$.

(c) $\left(\begin{smallmatrix} 3 & 0 \\ 0 & 1/3 \end{smallmatrix}\right)$.

(d) $\left(\begin{smallmatrix} 5/3 & 4/3 \\ 4/3 & 5/3 \end{smallmatrix}\right)$.

Solution. Throughout, membership in $\mathrm{Sp}(2, \mathbb{R})$ is decided by $\det \mathbf{P} = 1$ alone (Theorem 14.10.3 (c)), and membership in $\mathrm{O}(2)$ by orthonormality of the columns.

(a) $\mathrm{O}(2)$: yes, the columns are $\mathbf{e}_1$ and $-\mathbf{e}_2$, orthonormal. $\mathrm{O}(1,1)$: yes, $\mathbf{P}$ is diagonal with entries ± 1, so $\mathbf{P}^\top \mathbf{I}_{1,1} \mathbf{P} = \mathbf{I}_{1,1}$. $\mathrm{Sp}(2, \mathbb{R})$: no, $\det \mathbf{P} = -1$.

(b) $\mathrm{O}(2)$: yes, a rotation by $\pi/2$. $\mathrm{O}(1,1)$: no — $\mathbf{P}^\top \mathbf{I}_{1,1} \mathbf{P} = \left(\begin{smallmatrix} -1 & 0 \\ 0 & 1 \end{smallmatrix}\right) = -\mathbf{I}_{1,1} \neq \mathbf{I}_{1,1}$. $\mathrm{Sp}(2, \mathbb{R})$: yes, $\det \mathbf{P} = 1$.

(c) $\mathrm{O}(2)$: no, the first column has norm 3. $\mathrm{O}(1,1)$: no — for diagonal $\mathbf{P} = \mathrm{diag}(s, t)$ the condition is $\mathrm{diag}(s^2, -t^2) = \mathrm{diag}(1, -1)$, that is $s, t = \pm 1$, and $3 \neq \pm 1$. $\mathrm{Sp}(2, \mathbb{R})$: yes, $\det \mathbf{P} = 3 \cdot \frac{1}{3} = 1$.

(d) $\mathrm{O}(2)$: no, the first column has norm $\sqrt{25/9 + 16/9} = \sqrt{41}/3 \neq 1$. $\mathrm{O}(1,1)$: yes — this is $\left(\begin{smallmatrix} \cosh t & \sinh t \\ \sinh t & \cosh t \end{smallmatrix}\right)$ with $\cosh t = 5/3$, $\sinh t = 4/3$, and $(5/3)^2 - (4/3)^2 = 1$, so Example (Members of Each Family) (a) applies. $\mathrm{Sp}(2, \mathbb{R})$: yes, $\det \mathbf{P} = 25/9 - 16/9 = 1$.

Exercise. B2: A symplectic matrix of size four

Let $\mathbf{M} = \left(\begin{smallmatrix} 1 & 2 \\ 0 & 1 \end{smallmatrix}\right)$ and $\mathbf{S} = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$, and put

$$\mathbf{P}_1 = \begin{pmatrix} \mathbf{M} & \mathbf{O} \\ \mathbf{O} & (\mathbf{M}^\top)^{-1} \end{pmatrix}, \qquad \mathbf{P}_2 = \begin{pmatrix} \mathbf{I}_2 & \mathbf{S} \\ \mathbf{O} & \mathbf{I}_2 \end{pmatrix}.$$

(a) Write out $\mathbf{P}_1$ explicitly and verify $\mathbf{P}_1 \in \mathrm{Sp}(4, \mathbb{Q})$ directly from the definition.

(b) Explain why $\mathbf{P}_1 \mathbf{P}_2 \in \mathrm{Sp}(4, \mathbb{Q})$ without computing the product.

(c) Compute $\det \mathbf{P}_1$ and $\det \mathbf{P}_2$, and check them against Theorem 14.10.3.

Solution.

(a) $\mathbf{M}^\top = \left(\begin{smallmatrix} 1 & 0 \\ 2 & 1 \end{smallmatrix}\right)$, so $(\mathbf{M}^\top)^{-1} = \left(\begin{smallmatrix} 1 & 0 \\ -2 & 1 \end{smallmatrix}\right)$ and

$$\mathbf{P}_1 = \begin{pmatrix} 1 & 2 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & -2 & 1 \end{pmatrix}.$$

By the block computation of Example (Members of Each Family) (b), $\mathbf{P}_1^\top \mathbf{\Omega}_4 \mathbf{P}_1 = \left(\begin{smallmatrix} \mathbf{O} & \mathbf{M}^\top\mathbf{N} \\ -\mathbf{N}^\top\mathbf{M} & \mathbf{O} \end{smallmatrix}\right)$ with $\mathbf{N} = (\mathbf{M}^\top)^{-1}$, and $\mathbf{M}^\top\mathbf{N} = \mathbf{M}^\top(\mathbf{M}^\top)^{-1} = \mathbf{I}_2$. So $\mathbf{P}_1^\top \mathbf{\Omega}_4 \mathbf{P}_1 = \mathbf{\Omega}_4$.

(b) $\mathbf{S}$ is symmetric, so $\mathbf{P}_2 \in \mathrm{Sp}(4, \mathbb{Q})$ by Example (Members of Each Family) (c). A product of two elements of a group is again in the group, and $\mathrm{Sp}(4, \mathbb{Q})$ is a group by Theorem 14.10.1.

(c) $\mathbf{P}_1 = \mathbf{M} \oplus (\mathbf{M}^\top)^{-1}$ is block diagonal, so by Theorem 7.8.4 and Corollary 7.5.6,

$$\det \mathbf{P}_1 = \det \mathbf{M} \, \det((\mathbf{M}^\top)^{-1}) = \det \mathbf{M} \, (\det \mathbf{M})^{-1} = 1,$$

using $\det \mathbf{M}^\top = \det \mathbf{M}$ (Theorem 7.4.5). And $\mathbf{P}_2$ is block upper triangular with both diagonal blocks equal to $\mathbf{I}_2$, so $\det \mathbf{P}_2 = 1$ by Theorem 7.8.4. Both equal $+1$, consistent with the recorded fact $\mathrm{Sp}(2m, F) \subseteq \mathrm{SL}_{2m}(F)$; part (a) of the theorem only guarantees ± 1.

Exercise. B3: Counting parameters

(a) Compute the expected dimensions of $\mathrm{O}(4)$, $\mathrm{O}(2,2)$, $\mathrm{U}(3)$, and $\mathrm{Sp}(4, \mathbb{R})$.

(b) For which n and m do $\mathrm{O}(n)$ and $\mathrm{Sp}(2m, \mathbb{R})$ have the same expected dimension? Solve the resulting equation.

(c) The group $\mathrm{SO}(n)$ is cut out of $\mathrm{O}(n)$ by the single equation $\det \mathbf{P} = 1$. Why does this **not** reduce the expected dimension by one?

Solution.

(a) $\mathrm{O}(4)$ and $\mathrm{O}(2,2)$ both have $n = 4$, so $\frac{1}{2} \cdot 4 \cdot 3 = 6$. $\mathrm{U}(3)$: $3^2 = 9$ real parameters. $\mathrm{Sp}(4, \mathbb{R})$: $m = 2$, so $m(2m+1) = 10$.

(b) The equation is $\frac{1}{2}n(n-1) = m(2m+1)$, that is $n(n-1) = (2m+1)(2m)$. Reading the right-hand side as $n(n-1)$ with $n = 2m+1$ shows that $n = 2m+1$ is a solution for **every** m, and it is the only one: $n \mapsto n(n-1)$ is strictly increasing for $n \geq 1$, so it takes each value at most once. Hence

$$\dim \mathrm{O}(2m+1) = \dim \mathrm{Sp}(2m, \mathbb{R}) = m(2m+1),$$

and no other pairing of the two families matches. The smallest case is $\mathrm{O}(3)$ and $\mathrm{Sp}(2, \mathbb{R}) = \mathrm{SL}_2(\mathbb{R})$, both of dimension 3.

(c) Because $\det \mathbf{P} = \pm 1$ already, by Theorem 14.10.3 (a). The condition $\det \mathbf{P} = 1$ does not shrink a continuum of possibilities to a smaller one; it selects one of two discrete

alternatives. So $\mathrm{SO}(n)$ has the same expected dimension $\frac{1}{2}n(n-1)$ as $\mathrm{O}(n)$, and the two groups differ only in how many separate pieces they have — which is a topological statement of the kind this section does not prove.

14.10.7.3 C. Going deeper

Exercise. C1: Isometries of a degenerate form

Let $\mathbf{A} = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 0 \end{smallmatrix}\right) \in M_2(\mathbb{R})$ and let $G = \{\, \mathbf{P} \in M_2(\mathbb{R}) : \mathbf{P}^{\top}\mathbf{A}\mathbf{P} = \mathbf{A} \,\}$, the set produced by dropping the nondegeneracy hypothesis from Definition 14.10.1.

(a) Determine G explicitly.

(b) Exhibit an element of G that is not invertible, and explain which step of the proof of Theorem 14.10.1 it defeats.

(c) Prove that the invertible elements of G do form a group.

Solution.

(a) Write $\mathbf{P} = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$. Then

$$\mathbf{P}^{\top}\mathbf{A}\mathbf{P} = \begin{pmatrix} a & c \\ b & d \end{pmatrix}\begin{pmatrix} a & b \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} a^2 & ab \\ ab & b^2 \end{pmatrix},$$

so the condition is $a^2 = 1$, $ab = 0$, $b^2 = 0$. The last two say $b = 0$, and the first says $a = \pm 1$. Hence

$$G = \left\{ \begin{pmatrix} \pm 1 & 0 \\ c & d \end{pmatrix} : c, d \in \mathbb{R} \right\}.$$

(b) Take $c = d = 0$ and $a = 1$: the matrix $\left(\begin{smallmatrix} 1 & 0 \\ 0 & 0 \end{smallmatrix}\right)$ lies in G and has determinant 0. The step it defeats is the cancellation of $\det \mathbf{A}$ in $(\det \mathbf{P})^2 \det \mathbf{A} = \det \mathbf{A}$. Here $\det \mathbf{A} = 0$, the equation reads $0 = 0$, and it says nothing about $\det \mathbf{P}$.

(c) Let $G^{\times}$ be the set of invertible elements of G; from (a) these are the matrices above with $d \neq 0$. It contains $\mathbf{I}_2$. It is closed under products, since G is (the computation in Theorem 14.10.1 needs only $\mathbf{P}^{\top}\mathbf{A}\mathbf{P} = \mathbf{A}$ and $\mathbf{Q}^{\top}\mathbf{A}\mathbf{Q} = \mathbf{A}$, not invertibility) and a product of invertible matrices is invertible. And it is closed under inverses: if $\mathbf{P} \in G^{\times}$ then multiplying $\mathbf{P}^{\top}\mathbf{A}\mathbf{P} = \mathbf{A}$ by $(\mathbf{P}^{-1})^{\top}$ and $\mathbf{P}^{-1}$ gives $\mathbf{P}^{-1} \in G$, and $\mathbf{P}^{-1}$ is invertible. So $G^{\times}$ is a group — but it is *not* all of G, which is the reason Definition 14.10.1 assumes nondegeneracy rather than repairing the damage afterwards.

Exercise. C2: The symplectic group of a two-dimensional form, revisited

Let F be any field and let β be a nondegenerate alternating form on a two-dimensional vector space V over F.

(a) Prove that $\mathrm{Isom}(\beta) = \{\, T \in \mathcal{L}(V) : \det T = 1 \,\}$.

(b) Deduce that for $F = \mathbb{F}_2$ the group $\mathrm{Sp}(2, \mathbb{F}_2)$ has exactly 6 elements, and that $\mathrm{Sp}(2, \mathbb{F}_2) = \mathrm{GL}_2(\mathbb{F}_2)$.

(c) Is $\mathrm{Sp}(2, \mathbb{F}_3) = \mathrm{GL}_2(\mathbb{F}_3)$? Justify your answer by a count.

Hint for (b) and (c): count the invertible matrices over $\mathbb{F}_q$ by choosing the columns in turn.

Solution.

(a) Choose a symplectic basis $\mathcal{B}$ for β, which exists by Theorem 14.9.2 (b); then $[\beta]_{\mathcal{B}}^{\mathcal{B}} = \mathbf{\Omega}_2$. By the correspondence after Definition 14.10.1, $T \in \mathrm{Isom}(\beta)$ if and only if $\mathbf{P} \coloneqq [T]_{\mathcal{B}}^{\mathcal{B}}$ lies in $\mathrm{Isom}(\mathbf{\Omega}_2) = \mathrm{Sp}(2, F)$, which by Theorem 14.10.3 (c) says $\det \mathbf{P} = 1$. Finally $\det T = \det \mathbf{P}$ by Definition 7.5.1, since the determinant of an operator is that of its matrix in any basis.

(b) By (a), $\mathrm{Sp}(2, \mathbb{F}_2) = \mathrm{SL}_2(\mathbb{F}_2)$. Count $\mathrm{GL}_2(\mathbb{F}_2)$ first: the first column may be any of the $4 - 1 = 3$ nonzero vectors, and the second any of the $4 - 2 = 2$ vectors outside the span of the first, giving 6. Over $\mathbb{F}_2$ the only nonzero scalar is 1, so every invertible matrix has $\det = 1$; hence $\mathrm{SL}_2(\mathbb{F}_2) = \mathrm{GL}_2(\mathbb{F}_2)$ has 6 elements.

(c) No. The same count over $\mathbb{F}_3$ gives $|\mathrm{GL}_2(\mathbb{F}_3)| = (9-1)(9-3) = 48$. The determinant is a surjective homomorphism $\mathrm{GL}_2(\mathbb{F}_3) \to \mathbb{F}_3^{\times}$, whose target has 2 elements, and its fibers all have the same size; so $|\mathrm{SL}_2(\mathbb{F}_3)| = 24 \neq 48$. Concretely $\mathrm{diag}(1, -1)$ is invertible with determinant $-1 \neq 1$, so it lies in $\mathrm{GL}_2(\mathbb{F}_3)$ and not in $\mathrm{Sp}(2, \mathbb{F}_3)$.

Exercise. C3: Bounded and unbounded

(a) Prove that every entry of a matrix $\mathbf{P} \in \mathrm{O}(n)$ satisfies $|p_{ij}| \leq 1$.

(b) Prove that for every $n \geq 2$ and every real M, the group $\mathrm{O}(n-1, 1)$ contains a matrix with an entry larger than M.

(c) Explain in two sentences why (a) and (b) together make it plausible that $\mathrm{O}(n)$ is compact and $\mathrm{O}(n-1, 1)$ is not, and say precisely what has and has not been proved.

Solution.

(a) The columns of $\mathbf{P} \in \mathrm{O}(n)$ are orthonormal in $\mathbb{R}^n$ (Theorem 11.7.3 (f)), so for each j we have $\sum_i p_{ij}^2 = 1$. Each term of a sum of nonnegative reals is at most the sum, so $p_{ij}^2 \leq 1$ and $|p_{ij}| \leq 1$.

(b) Put $\mathbf{B}_t \coloneqq \left(\begin{smallmatrix} \cosh t & \sinh t \\ \sinh t & \cosh t \end{smallmatrix}\right)$, which lies in $\mathrm{O}(1, 1)$ for every real t by Example (Members of Each Family) (a), and consider

$$\mathbf{P}_t \coloneqq \mathbf{I}_{n-2} \oplus \mathbf{B}_t \in M_n(\mathbb{R}).$$

(For $n = 2$ read $\mathbf{P}_t = \mathbf{B}_t$.) As a block diagonal matrix $\mathbf{I}_{n-1,1} = \mathbf{I}_{n-2} \oplus \mathbf{I}_{1,1}$, and each block of $\mathbf{P}_t$ preserves its own block of the form: $\mathbf{I}_{n-2}^{\top} \mathbf{I}_{n-2} \mathbf{I}_{n-2} = \mathbf{I}_{n-2}$, and $\mathbf{B}_t^{\top} \mathbf{I}_{1,1} \mathbf{B}_t = \mathbf{I}_{1,1}$ by Example (Members of Each Family) (a). Multiplying block by block gives $\mathbf{P}_t^{\top} \mathbf{I}_{n-1,1} \mathbf{P}_t = \mathbf{I}_{n-1,1}$, so $\mathbf{P}_t \in \mathrm{O}(n-1, 1)$. Its (n, n) entry is $\cosh t = \frac{1}{2}(e^t + e^{-t}) \geq \frac{1}{2} e^t$, and e^t grows without bound, so some t makes this entry exceed the given M.

(c) A compact subset of $M_n(\mathbb{R}) \cong \mathbb{R}^{n^2}$ must be bounded, so (b) rules out compactness for $\mathrm{O}(n-1, 1)$, while (a) supplies half of what $\mathrm{O}(n)$ would need. What has been **proved** here is only the statement about entries: $\mathrm{O}(n)$ is bounded and $\mathrm{O}(n-1, 1)$ is not. That "bounded and closed implies compact" in $\mathbb{R}^{n^2}$, and that $\mathrm{O}(n)$ is closed, are theorems of

analysis that this book does not prove, so the compactness claim itself remains recorded rather than established.

14.11 Minkowski Space

Sylvester's law of inertia says that a real symmetric form is nothing but a pair of counts, and that any two forms with the same counts are the same form in disguise. One pair of counts has a claim on our attention that none of the others has: the form of signature $(1,3)$ is the one the physical world appears to run on. This section does the linear algebra of that form. Everything in it is this chapter's machinery specialized, and the payoff is a picture — a cone dividing space into three regions — together with a classification theorem that is the hyperbolic twin of a circular one we already proved.

14.11.1 The form of special relativity

Fix $n_+ = 1$ and $n_- = 3$ in the canonical form of Corollary 14.5.2, and there is only one form to write down.

Definition 14.11.1. Minkowski Space

Minkowski space $\mathbb{R}^{1,3}$ is the real vector space $\mathbb{R}^4$ equipped with the symmetric bilinear form

$$\eta(\mathbf{x},\mathbf{y}) := x_1y_1 - x_2y_2 - x_3y_3 - x_4y_4,$$

whose quadratic form is $q(\mathbf{x}) = x_1^2 - x_2^2 - x_3^2 - x_4^2$. The first coordinate is the **time** coordinate and the remaining three are the **space** coordinates. More generally, for $p, m \geq 0$, $\mathbb{R}^{p,m}$ is $\mathbb{R}^{p+m}$ with the form whose matrix in the standard basis is $\mathbf{I}_p \oplus (-\mathbf{I}_m)$.

In words: η is the dot product of $\mathbb{R}^4$ with three of its four signs reversed. Its matrix in the standard basis $\mathcal{E}$ is

$$[\eta]_{\mathcal{E}}^{\mathcal{E}} = \operatorname{diag}(1,-1,-1,-1),$$

which is symmetric, so η is a symmetric form (Definition 14.3.1), and it is diagonal, so $\mathcal{E}$ is an orthogonal basis for η (Definition 14.4.1). Reading off the signs, the signature of η is $(1,3)$ and its inertia is $(1,3,0)$ (Definition 14.5.1). The form is **nondegenerate**: $\eta(\mathbf{e}_1,\mathbf{x}) = x_1$ and $\eta(\mathbf{e}_j,\mathbf{x}) = -x_j$ for $j = 2,3,4$, so a vector paired to zero with everything has all four coordinates zero (Definition 14.1.4).

Warning. **Two sign conventions are in use, and both are called "the" Minkowski form.** We take $(+,-,-,-)$, signature $(1,3)$; many sources take $(-,+,+,+)$, signature $(3,1)$, which is our $-\eta$. The two are not congruent — congruence preserves the signature (Theorem 14.5.1) — but every statement below transfers by reversing each inequality on q, because q changes sign. The one thing both conventions agree on is the set $\{q = 0\}$, which is why the picture of this section is convention-free. State your convention before computing, and never mix the two.

14.11.2 Timelike, spacelike, and null

Nothing in this section is a new notion. Section 7 studied the vectors on which a form vanishes, and called them isotropic; over $\mathbb{R}$ a nondegenerate indefinite form has plenty of them, and the ones belonging to η have names of their own.

The form sorts nonzero vectors into three classes by the sign of q, and the middle class — where q vanishes — is exactly the isotropy of Section 7.

Definition 14.11.2. Timelike, Spacelike, Null

Let $\mathbf{x} \in \mathbb{R}^{1,3}$ with $\mathbf{x} \neq \mathbf{0}$. Then $\mathbf{x}$ is

(a) **timelike** if $q(\mathbf{x}) > 0$;
(b) **null**, or **lightlike**, if $q(\mathbf{x}) = 0$;
(c) **spacelike** if $q(\mathbf{x}) < 0$.

The set $\{\mathbf{x} \in \mathbb{R}^{1,3} : q(\mathbf{x}) = 0\}$ is the **light cone**. A timelike $\mathbf{x}$ is **future-pointing** if $x_1 > 0$ and **past-pointing** if $x_1 < 0$.

The word **null** is a second name for something already defined: a null vector is precisely a nonzero **isotropic** vector for η in the sense of Definition 14.7.1, and the light cone is the isotropy locus of η together with $\mathbf{0}$. So "the light cone" is not new geometry; it is the locus Section 7 already forced us to take seriously, drawn for one particular form. Timelike and spacelike are then just names for the two sides that the isotropy locus separates.

The last clause of the definition is legitimate because a timelike vector cannot have $x_1 = 0$: if it did, then $q(\mathbf{x}) = -(x_2^2+x_3^2+x_4^2) \leq 0$. So every timelike vector is future-pointing or past-pointing, and exactly one of the two.

Example. Reading off the three types

Classify $\mathbf{e}_1$, $\mathbf{e}_2$, $\mathbf{e}_1 + \mathbf{e}_2$, and $(2, 1, 1, 1)$ in $\mathbb{R}^{1,3}$, and decide which of them are isotropic.

Solution. $q(\mathbf{e}_1) = 1 > 0$, so $\mathbf{e}_1$ is timelike and future-pointing. $q(\mathbf{e}_2) = -1 < 0$, so $\mathbf{e}_2$ is spacelike. $q(\mathbf{e}_1 + \mathbf{e}_2) = 1 - 1 = 0$, so $\mathbf{e}_1 + \mathbf{e}_2$ is null, hence isotropic. Finally $q(2, 1, 1, 1) = 4 - 3 = 1 > 0$, timelike and future-pointing. Only $\mathbf{e}_1 + \mathbf{e}_2$ is isotropic; the other three are anisotropic.

Section 7's structure theorem applies verbatim. A nonzero isotropic vector sits in a hyperbolic plane, and here we can point to one: take $\mathbf{u} = \mathbf{e}_1+\mathbf{e}_2$ and $\mathbf{w} = \frac{1}{2}(\mathbf{e}_1-\mathbf{e}_2)$. Then $q(\mathbf{u}) = q(\mathbf{w}) = 0$ and

$$\eta(\mathbf{u}, \mathbf{w}) = \tfrac{1}{2}(1 - (-1)) = 1,$$

so the Gram matrix of $(\mathbf{u}, \mathbf{w})$ is $\left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$ and $\operatorname{span}(\mathbf{u}, \mathbf{w})$ is a hyperbolic plane (Definition 14.7.3). Splitting it off, Theorem 14.7.5 writes

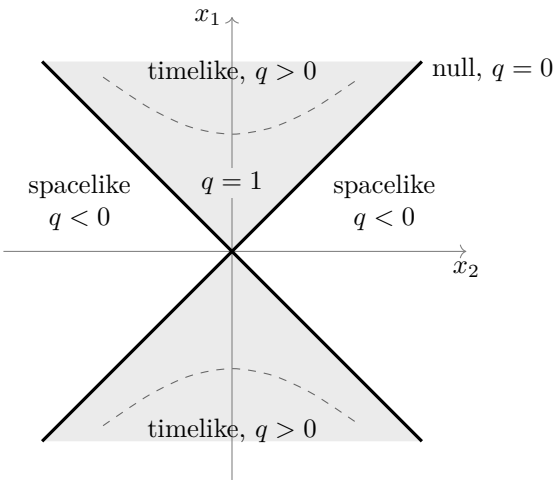
$$\mathbb{R}^{1,3} \;=\; H \perp W,$$

with H hyperbolic of signature $(1, 1)$ and W anisotropic of signature $(0, 2)$, that is, negative definite. Only one plane splits off, because the Witt index of a real form of signature (n_+, n_-) is $\min(n_+, n_-)$ (Proposition 14.7.7), which here is $\min(1, 3) = 1$.

Quick check. Is there a **plane** in $\mathbb{R}^{1,3}$ on which η vanishes identically?

Solution. No. Suppose $\mathbf{x} = (a, \mathbf{u})$ and $\mathbf{y} = (b, \mathbf{w})$ span such a plane, where $a, b \in \mathbb{R}$ and $\mathbf{u}, \mathbf{w} \in \mathbb{R}^3$. Then $q(\mathbf{x}) = q(\mathbf{y}) = 0$ gives $a^2 = \|\mathbf{u}\|^2$ and $b^2 = \|\mathbf{w}\|^2$, while $\eta(\mathbf{x}, \mathbf{y}) = 0$ gives $ab = \mathbf{u} \cdot \mathbf{w}$. Hence $|\mathbf{u} \cdot \mathbf{w}| = |a||b| = \|\mathbf{u}\|\|\mathbf{w}\|$, which is the equality case of Theorem 11.1.5, so $(\mathbf{u}, \mathbf{w})$ is dependent. If $\mathbf{u} = \mathbf{0}$ then $a = 0$ and $\mathbf{x} = \mathbf{0}$; otherwise $\mathbf{w} = \lambda \mathbf{u}$, so $b^2 = \lambda^2 a^2$ and $ab = \lambda a^2$, which force $b = \lambda a$ and $\mathbf{y} = \lambda \mathbf{x}$. Either way $(\mathbf{x}, \mathbf{y})$ is dependent, so no plane exists. Structurally this is the statement that the Witt index of $\mathbb{R}^{1,3}$ is $\min(1, 3) = 1$ (Proposition 14.7.7), since by Definition 14.7.4 no totally isotropic subspace has dimension larger than the index.

The picture is easiest to draw with one space coordinate instead of three, in $\mathbb{R}^{1,1}$, where the light cone degenerates to the two lines $x_2 = \pm x_1$.



The light cone $q = 0$ in the Minkowski plane separates the timelike vectors from the spacelike ones; the "unit sphere" $q = 1$ is the dashed hyperbola, and it is unbounded

Two features of the picture deserve to be said out loud. The shaded region is **not** connected: the future-pointing and past-pointing timelike vectors form two separate pieces, and no continuous path of timelike vectors joins them. That last claim is worth one line, since it is the only place in this section where a topological word is used. Suppose γ were a continuous path of timelike vectors from a future-pointing vector to a past-pointing one. Its first coordinate x_1 is a continuous real function that is positive at one end and negative at the other, so by the intermediate value theorem — an analysis input, quoted and not proved here, as in Section 6 — it vanishes somewhere along the path. But $x_1 = 0$ forces $q(\mathbf{x}) = -x_2^2 - x_3^2 - x_4^2 \leq 0$, so that point of the path is not timelike, a contradiction. And the level set $q = 1$, which plays the role the unit circle plays for the dot product, is unbounded. Both facts come straight from the mixed signs, and both are invisible to anyone whose intuition was trained on positive definite forms.

Warning. **Orthogonality for η is not the perpendicularity you can see.** A null vector is η-orthogonal to *itself*, and the consequence is that Chapter 11's decomposition $V = U \oplus U^{\perp}$ fails. Take $U = \operatorname{span}((1, 1, 0, 0))$. Since $\eta((1, 1, 0, 0), \mathbf{x}) = x_1 - x_2$, the η-orthogonal complement of U is the hyperplane $\{x_1 = x_2\}$, which **contains** U. So $U + U^{\perp} = U^{\perp}$ is

three-dimensional, not four, and $U \cap U^{\perp} = U \neq \{\mathbf{0}\}$. Nothing has gone wrong; positivity was doing that work in Chapter 11, and it is gone.

14.11.3 Hyperbolic rotations

Now the isometries. The group attached to η is the isometry group of its matrix (Definition 14.10.1), which is $\mathrm{O}(1,3)$ in the notation of Definition 14.10.2: the matrices $\mathbf{P}$ with $\mathbf{P}^{\top}[\eta]_{\mathcal{E}}^{\mathcal{E}}\mathbf{P} = [\eta]_{\mathcal{E}}^{\mathcal{E}}$. Its elements are the **Lorentz transformations**. We classify them in dimension 2, where η becomes

$$\eta_2(\mathbf{x},\mathbf{y}) = x_1y_1 - x_2y_2, \qquad [\eta_2]_{\mathcal{E}}^{\mathcal{E}} = \operatorname{diag}(1,-1).$$

The classification of $\mathrm{O}(2)$ in Theorem 11.7.6 rested on one geometric fact: a unit vector of the Euclidean plane is $(\cos\theta, \sin\theta)$ for exactly one $\theta \in [0, 2\pi)$. Replace the circle by the hyperbola and the same proof runs. We need the corresponding parametrization, and we take it from calculus:

$$\begin{aligned} \cosh t &= \tfrac{1}{2}(e^{t} + e^{-t}), \qquad \sinh t = \tfrac{1}{2}(e^{t} - e^{-t}), \\ \cosh^2 t - \sinh^2 t &= 1, \qquad \cosh t > 0, \\ \cosh(s+t) &= \cosh s \cosh t + \sinh s \sinh t, \\ \sinh(s+t) &= \sinh s \cosh t + \cosh s \sinh t, \end{aligned}$$

and $\sinh\colon \mathbb{R} \to \mathbb{R}$ is a strictly increasing bijection. Those facts are the only analysis used in this section.

Definition 14.11.3. Boost

For $t \in \mathbb{R}$, the **boost of rapidity** t is

$$\mathbf{B}_t \coloneqq \begin{pmatrix} \cosh t & \sinh t \\ \sinh t & \cosh t \end{pmatrix} \in M_2(\mathbb{R}).$$

A boost is the hyperbolic analogue of the rotation $\mathbf{R}_\theta$ of Chapter 11: same shape, hyperbolic functions in place of circular ones, and no minus sign in the corner.

Theorem 14.11.1. Classification of the Plane Lorentz Transformations

Let $\mathbf{P} \in \mathrm{O}(1,1)$, that is, $\mathbf{P} \in M_2(\mathbb{R})$ with $\mathbf{P}^{\top}\operatorname{diag}(1,-1)\mathbf{P} = \operatorname{diag}(1,-1)$. Then there are a unique $t \in \mathbb{R}$ and unique signs $\varepsilon, \mu \in \{1,-1\}$ with

$$\mathbf{P} = \begin{pmatrix} \varepsilon\cosh t & \mu\sinh t \\ \sinh t & \varepsilon\mu\cosh t \end{pmatrix},$$

and $\det \mathbf{P} = \mu$. Writing $\mathbf{D} = \operatorname{diag}(1,-1)$, this says

$$\mathrm{O}(1,1) = \{\, \mathbf{B}_t,\ \mathbf{B}_t\mathbf{D},\ -\mathbf{B}_t,\ -\mathbf{B}_t\mathbf{D} \ :\ t \in \mathbb{R} \,\},$$

a disjoint union of four families. Moreover $\mathbf{B}_s\mathbf{B}_t = \mathbf{B}_{s+t}$ for all s, t, so $\{\mathbf{B}_t : t \in \mathbb{R}\}$ is a

subgroup of $\mathrm{O}(1,1)$ isomorphic to $(\mathbb{R},+)$.

Idea. The congruence $\mathbf{P}^\top\mathbf{D}\mathbf{P}=\mathbf{D}$ says exactly that the two columns of $\mathbf{P}$ are an orthogonal basis for η_2 with q-values 1 and -1. So: parametrize the vectors with $q=1$, which is the hyperbola; then the second column is forced, because the η_2-orthogonal complement of a nonisotropic vector is a line, and on that line only two vectors have $q=-1$. This is Theorem 11.7.6's proof with the circle replaced by the hyperbola.

Proof. Write $\mathbf{P}=\begin{pmatrix} a & c \\ b & d \end{pmatrix}$ with columns $\mathbf{p}_1=(a,b)$ and $\mathbf{p}_2=(c,d)$. The (i,j) entry of $\mathbf{P}^\top\mathbf{D}\mathbf{P}$ is $\eta_2(\mathbf{p}_i,\mathbf{p}_j)$, so the hypothesis says

$$q(\mathbf{p}_1)=1, \qquad q(\mathbf{p}_2)=-1, \qquad \eta_2(\mathbf{p}_1,\mathbf{p}_2)=0.$$

The first column. Since sinh is a bijection of $\mathbb{R}$, there is a unique $t\in\mathbb{R}$ with $b=\sinh t$. Then $a^2=1+b^2=1+\sinh^2 t=\cosh^2 t$, so $a=\varepsilon\cosh t$ for a unique sign ε, unique because $\cosh t>0$.

The second column. The vector $\mathbf{z}=(\sinh t,\varepsilon\cosh t)$ satisfies $\eta_2(\mathbf{p}_1,\mathbf{z})=\varepsilon\cosh t\sinh t-\sinh t\,\varepsilon\cosh t=0$, and $\mathbf{z}\neq\mathbf{0}$ because $\varepsilon\cosh t\neq 0$. The set $\{\mathbf{y}:\eta_2(\mathbf{p}_1,\mathbf{y})=0\}$ is the null space of the nonzero row vector $(a,-b)$, hence a line, so it equals $\operatorname{span}(\mathbf{z})$. Therefore $\mathbf{p}_2=\mu'\mathbf{z}$ for some $\mu'\in\mathbb{R}$, and $q(\mathbf{z})=\sinh^2 t-\cosh^2 t=-1$ gives $-1=q(\mathbf{p}_2)=(\mu')^2 q(\mathbf{z})=-(\mu')^2$. Hence $\mu'=\mu$ is a sign, and it is determined by $\mathbf{P}$, since $d=\mu\varepsilon\cosh t$ and $\cosh t>0$. This is the displayed matrix, and expanding,

$$\det\mathbf{P}=\mu(\varepsilon^2\cosh^2 t-\sinh^2 t)=\mu.$$

The four families. For $\varepsilon=\mu=1$ the matrix is $\mathbf{B}_t$; for $\varepsilon=1,\mu=-1$ it is $\mathbf{B}_t\mathbf{D}$; for $\varepsilon=-1,\mu=1$ it is $-\mathbf{B}_{-t}$; and for $\varepsilon=\mu=-1$ it is $-\mathbf{B}_{-t}\mathbf{D}$, each by multiplying out and using $\sinh(-t)=-\sinh t$, $\cosh(-t)=\cosh t$. As t runs over $\mathbb{R}$ so does $-t$, which gives the displayed set equality. The families are disjoint because ε is the sign of the $(1,1)$ entry and $\mu=\det\mathbf{P}$, and both are determined by $\mathbf{P}$.

The subgroup. By the addition formulas for cosh and sinh,

$$\mathbf{B}_s\mathbf{B}_t=\begin{pmatrix} \cosh(s{+}t) & \sinh(s{+}t) \\ \sinh(s{+}t) & \cosh(s{+}t) \end{pmatrix}=\mathbf{B}_{s+t}.$$

So $t\mapsto\mathbf{B}_t$ is a group homomorphism $(\mathbb{R},+)\to\mathrm{O}(1,1)$; it is injective because sinh is, and its image is a subgroup by Definition 0.10.2. This proves the theorem.

■

The four families are sorted by two signs, and each sign has a reading. The determinant μ says whether the space coordinate is reversed, exactly as in $\mathrm{O}(2)$. The sign ε of the $(1,1)$ entry says whether the future cone is sent to the future cone: for $\mathbf{B}_t$ the image of $\mathbf{e}_1$ is $(\cosh t,\sinh t)$, with positive first coordinate, while $-\mathbf{B}_t$ sends it to $(-\cosh t,-\sinh t)$. So the identity component consists of the boosts alone, and the other three families are obtained from it by reversing space,

time, or both.

14.11.4 Circular and hyperbolic, side by side

The comparison is the point of this section, so here it is in one table. In each column, the isometry group of the named form on $\mathbb{R}^2$.

	$\mathrm{O}(2)$: form $\mathbf{I}_2$	$\mathrm{O}(1,1)$: form $\mathrm{diag}(1,-1)$
level set $q=1$	circle $x_1^2+x_2^2=1$	hyperbola $x_1^2-x_2^2=1$, two branches
canonical element	$\mathbf{R}_\theta=\left(\begin{smallmatrix}\cos\theta & -\sin\theta\\ \sin\theta & \cos\theta\end{smallmatrix}\right)$	$\mathbf{B}_t=\left(\begin{smallmatrix}\cosh t & \sinh t\\ \sinh t & \cosh t\end{smallmatrix}\right)$
parameter	$\theta\in[0,2\pi)$, periodic	$t\in\mathbb{R}$, unbounded
addition law	$\mathbf{R}_\alpha\mathbf{R}_\beta=\mathbf{R}_{\alpha+\beta}$	$\mathbf{B}_s\mathbf{B}_t=\mathbf{B}_{s+t}$
determinant $+1$ part	$\mathrm{SO}(2)=\{\mathbf{R}_\theta\}$	$\{\mathbf{B}_t\}\cup\{-\mathbf{B}_t\}$
determinant -1 part	reflections $\mathbf{M}_\theta$	$\{\mathbf{B}_t\mathbf{D}\}\cup\{-\mathbf{B}_t\mathbf{D}\}$
number of pieces	2	4
entries	all of modulus ≤ 1	unbounded

The two canonical forms differ in one place only: $\mathbf{R}_\theta$ carries a minus sign that $\mathbf{B}_t$ does not, and that single sign is the difference between $\cos^2+\sin^2=1$ and $\cosh^2-\sinh^2=1$. Everything else in the table follows from it. In particular the last row is the unboundedness that Section 10 recorded when it observed, without proof, that $\mathrm{O}(p,q)$ is not compact for $p,q\ge 1$ while $\mathrm{O}(n)$ is.

Remark. In the physical reading, $\tanh t$ is a velocity measured in units of the speed of light, and the additivity $\mathbf{B}_s\mathbf{B}_t=\mathbf{B}_{s+t}$ of rapidities becomes the relativistic law for adding velocities. Exercise (C1) derives the formula from the theorem. The book verifies the algebra and stops there.

14.11.5 The triangle inequality, reversed

For a timelike $\mathbf{x}$ the number $q(\mathbf{x})$ is positive, so it has a positive square root, and the analogy with a norm is irresistible. Write

$$\tau(\mathbf{x}) := \sqrt{q(\mathbf{x})} \qquad (\mathbf{x}\text{ timelike}).$$

The analogy holds, but with every inequality upside down. First the tool.

Lemma 14.11.2. Reversed Cauchy–Schwarz Inequality

Let $\mathbf{x},\mathbf{y}\in\mathbb{R}^{1,3}$ be timelike and both future-pointing. Then

$$\eta(\mathbf{x},\mathbf{y})\ge\tau(\mathbf{x})\tau(\mathbf{y})>0,$$

with equality if and only if $\mathbf{y}=\lambda\mathbf{x}$ for some $\lambda>0$.

Idea. Split each vector as a time coordinate plus a space part, $\mathbf{x} = (a, \mathbf{u})$. Timelike and future-pointing says $a > \|\mathbf{u}\| \geq 0$. Then $\eta(\mathbf{x}, \mathbf{y}) = ab - \mathbf{u} \cdot \mathbf{w}$, and Theorem 11.1.5 bounds $\mathbf{u} \cdot \mathbf{w}$ by $\|\mathbf{u}\|\|\mathbf{w}\|$ — in the helpful direction, since it appears with a minus sign. What is left is to compare $ab - \|\mathbf{u}\|\|\mathbf{w}\|$ with $\tau(\mathbf{x})\tau(\mathbf{y})$, and squaring both sides turns the difference into a perfect square.

Proof. Write $\mathbf{x} = (a, \mathbf{u})$ and $\mathbf{y} = (b, \mathbf{w})$ with $a, b \in \mathbb{R}$ and $\mathbf{u}, \mathbf{w} \in \mathbb{R}^3$, so that $q(\mathbf{x}) = a^2 - \|\mathbf{u}\|^2$ and $\eta(\mathbf{x}, \mathbf{y}) = ab - \mathbf{u} \cdot \mathbf{w}$. Timelike gives $a^2 > \|\mathbf{u}\|^2$ and $b^2 > \|\mathbf{w}\|^2$, and future-pointing gives $a, b > 0$; hence $a > \|\mathbf{u}\| \geq 0$ and $b > \|\mathbf{w}\| \geq 0$, so $ab > \|\mathbf{u}\|\|\mathbf{w}\| \geq 0$. By Theorem 11.1.5 in $\mathbb{R}^3$,

$$\eta(\mathbf{x}, \mathbf{y}) = ab - \mathbf{u} \cdot \mathbf{w} \geq ab - \|\mathbf{u}\|\|\mathbf{w}\| > 0.$$

Both sides of the inequality to be proved are now known positive, so it suffices to compare squares. Expanding,

$$\begin{aligned} &\left(ab - \|\mathbf{u}\|\|\mathbf{w}\|\right)^2 - q(\mathbf{x})q(\mathbf{y}) \\ &\quad = a^2\|\mathbf{w}\|^2 + b^2\|\mathbf{u}\|^2 - 2ab\|\mathbf{u}\|\|\mathbf{w}\| \\ &\quad = \left(a\|\mathbf{w}\| - b\|\mathbf{u}\|\right)^2 \geq 0. \end{aligned}$$

Therefore $ab - \|\mathbf{u}\|\|\mathbf{w}\| \geq \sqrt{q(\mathbf{x})q(\mathbf{y})} = \tau(\mathbf{x})\tau(\mathbf{y})$, and combining with the previous display gives $\eta(\mathbf{x}, \mathbf{y}) \geq \tau(\mathbf{x})\tau(\mathbf{y}) > 0$.

For the equality case, equality forces both steps to be equalities: $\mathbf{u} \cdot \mathbf{w} = \|\mathbf{u}\|\|\mathbf{w}\|$ and $a\|\mathbf{w}\| = b\|\mathbf{u}\|$. Suppose first $\mathbf{u} = \mathbf{0}$. Then $a\|\mathbf{w}\| = 0$ and $a > 0$, so $\mathbf{w} = \mathbf{0}$, and $\mathbf{y} = (b, \mathbf{0}) = (b/a)\mathbf{x}$ with $b/a > 0$. Suppose instead $\mathbf{u} \neq \mathbf{0}$. The equality case of Theorem 11.1.5 makes $(\mathbf{u}, \mathbf{w})$ dependent, so $\mathbf{w} = \lambda\mathbf{u}$ with $\lambda \in \mathbb{R}$, and then $\mathbf{u} \cdot \mathbf{w} = \lambda\|\mathbf{u}\|^2 = \|\mathbf{u}\|\|\mathbf{w}\| = |\lambda|\|\mathbf{u}\|^2$ forces $\lambda \geq 0$. Now $a\|\mathbf{w}\| = b\|\mathbf{u}\|$ reads $a\lambda\|\mathbf{u}\| = b\|\mathbf{u}\|$, so $b = \lambda a$ and $\mathbf{y} = \lambda\mathbf{x}$; and $\lambda > 0$ because $b > 0$. Conversely, if $\mathbf{y} = \lambda\mathbf{x}$ with $\lambda > 0$ then $\eta(\mathbf{x}, \mathbf{y}) = \lambda q(\mathbf{x})$ and $\tau(\mathbf{x})\tau(\mathbf{y}) = \sqrt{\lambda^2 q(\mathbf{x})^2} = \lambda q(\mathbf{x})$, so equality holds. This proves the lemma.

■

Theorem 14.11.3. Reversed Triangle Inequality

Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^{1,3}$ be timelike and both future-pointing. Then $\mathbf{x} + \mathbf{y}$ is timelike and future-pointing, and

$$\tau(\mathbf{x} + \mathbf{y}) \geq \tau(\mathbf{x}) + \tau(\mathbf{y}),$$

with equality if and only if $\mathbf{y}$ is a positive multiple of $\mathbf{x}$.

Proof. By Theorem 14.3.5, $q(\mathbf{x} + \mathbf{y}) = q(\mathbf{x}) + 2\eta(\mathbf{x}, \mathbf{y}) + q(\mathbf{y})$. By Lemma 14.11.2, $\eta(\mathbf{x}, \mathbf{y}) \geq \tau(\mathbf{x})\tau(\mathbf{y})$, so

$$q(\mathbf{x} + \mathbf{y}) \geq q(\mathbf{x}) + 2\tau(\mathbf{x})\tau(\mathbf{y}) + q(\mathbf{y}) = \left(\tau(\mathbf{x}) + \tau(\mathbf{y})\right)^2,$$

where the last step uses $q = \tau^2$ on timelike vectors. The right-hand side is positive, so $\mathbf{x} + \mathbf{y}$ is timelike; its first coordinate is a sum of two positive numbers, so it is future-pointing. Both sides of the last display are nonnegative, so taking square roots preserves the inequality and gives $\tau(\mathbf{x} + \mathbf{y}) \geq \tau(\mathbf{x}) + \tau(\mathbf{y})$. Equality holds exactly when it holds in Lemma 14.11.2, that is, exactly when $\mathbf{y} = \lambda\mathbf{x}$ with $\lambda > 0$. This proves the theorem.

■

Remark. Compare Corollary 11.1.6, which gives $\|u+v\| \le \|u\| + \|v\|$ in an inner product space, and Theorem 0.6.4, its ancestor in $\mathbb{C}$. Same shape, opposite direction: the reversal is produced by the single sign change in the form, and by nothing else.

Read as a statement about clocks, the reversed inequality says that a journey made in two straight legs accumulates *less* elapsed time than the single straight leg between the same endpoints, which is the algebra behind the twin paradox; whether that algebra describes the world is a question for physics, and this book answers only the algebra.

Quick check. Where exactly would the proof of Theorem 14.11.3 break if $\mathbf{x}$ were future-pointing and $\mathbf{y}$ past-pointing?

Solution. At Lemma 14.11.2, whose proof uses $a, b > 0$ to get $ab > \|\mathbf{u}\|\|\mathbf{w}\| \ge 0$. With $b < 0$ we have $ab < 0$, and $\eta(\mathbf{x}, \mathbf{y}) = ab - \mathbf{u}\cdot\mathbf{w}$ can be negative, so the chain collapses at its first step. The conclusion genuinely fails: take $\mathbf{x} = (1,0,0,0)$ and $\mathbf{y} = (-1,0,0,0)$, both timelike with $\tau = 1$, whose sum is $\mathbf{0}$ and is not timelike at all.

14.11.6 Exercises

14.11.6.1 A. Check your understanding

Exercise. A1

(a) State the sign convention used in this section, give the matrix of η in the standard basis, and say what its signature and inertia are.

(b) Define timelike, null, and spacelike, and say which earlier notion the word "null" renames.

(c) True or false: the sum of two spacelike vectors is spacelike. Justify your answer.

(d) Write down $\det \mathbf{B}_t$ and $\mathbf{B}_t^{-1}$, and say which row of the comparison table each answer sits in.

(e) True or false: $\mathrm{O}(1,1)$ is abelian. Justify your answer.

Solution.

(a) The convention is $(+,-,-,-)$, so $[\eta]_{\mathcal{E}}^{\mathcal{E}} = \mathrm{diag}(1,-1,-1,-1)$, with signature $(1,3)$ and inertia $(1,3,0)$ (Definition 14.5.1). The opposite convention $(-,+,+,+)$ uses $-\eta$ and has signature $(3,1)$.

(b) For $\mathbf{x} \neq \mathbf{0}$: timelike means $q(\mathbf{x}) > 0$, null means $q(\mathbf{x}) = 0$, spacelike means $q(\mathbf{x}) < 0$ (Definition 14.11.2). "Null" renames "nonzero isotropic vector" (Definition 14.7.1).

(c) False. Take $\mathbf{x} = (1,2,0,0)$ and $\mathbf{y} = (1,-2,0,0)$, with $q(\mathbf{x}) = q(\mathbf{y}) = 1 - 4 = -3 < 0$. Their sum is $(2,0,0,0)$, with $q = 4 > 0$, which is timelike.

(d) $\det \mathbf{B}_t = 1$ by Theorem 14.11.1 with $\mu = 1$, and $\mathbf{B}_t^{-1} = \mathbf{B}_{-t}$, since $\mathbf{B}_t\mathbf{B}_{-t} = \mathbf{B}_0 = \mathbf{I}_2$. The first sits in the "determinant $+1$ part" row, the second in the "addition law" row, which is what makes $t \mapsto \mathbf{B}_t$ a group homomorphism.

(e) False. $\mathbf{B}_1$ and $\mathbf{D} = \operatorname{diag}(1,-1)$ both lie in $\mathrm{O}(1,1)$, and $\mathbf{B}_1\mathbf{D}$ has $(1,2)$ entry $-\sinh 1$ while $\mathbf{D}\mathbf{B}_1$ has $(1,2)$ entry $\sinh 1$; these differ since $\sinh 1 \neq 0$.

14.11.6.2 B. Practice

Exercise. B1

Determine which of the following lie in $\mathrm{O}(1,1)$. For each one that does, give its rapidity t and the signs ε, μ of Theorem 14.11.1. Justify your answer.

$$\mathbf{A}_1 = \begin{pmatrix} 5/3 & 4/3 \\ 4/3 & 5/3 \end{pmatrix}, \quad \mathbf{A}_2 = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}, \quad \mathbf{A}_3 = \begin{pmatrix} 5/3 & -4/3 \\ 4/3 & -5/3 \end{pmatrix}.$$

Solution. Write $\mathbf{D} = \operatorname{diag}(1,-1)$ and test $\mathbf{P}^{\top}\mathbf{D}\mathbf{P} = \mathbf{D}$ column by column, using $q(\mathbf{p}_1) = 1$, $q(\mathbf{p}_2) = -1$, $\eta_2(\mathbf{p}_1, \mathbf{p}_2) = 0$.

$\mathbf{A}_1$: $(5/3)^2 - (4/3)^2 = (25-16)/9 = 1$; $(4/3)^2 - (5/3)^2 = -1$; $(5/3)(4/3) - (4/3)(5/3) = 0$. So $\mathbf{A}_1 \in \mathrm{O}(1,1)$. Here $\sinh t = 4/3$ and $\cosh t = 5/3$, so $\varepsilon = \mu = 1$ and $\mathbf{A}_1 = \mathbf{B}_t$ with $t = \log 3$, since $\cosh(\log 3) = \frac{1}{2}(3 + \frac{1}{3}) = 5/3$ and $\sinh(\log 3) = \frac{1}{2}(3 - \frac{1}{3}) = 4/3$.

$\mathbf{A}_2$: $2^2 - 1^2 = 3 \neq 1$, so the first column already fails and $\mathbf{A}_2 \notin \mathrm{O}(1,1)$. (Indeed $\mathbf{A}_2^{\top}\mathbf{D}\mathbf{A}_2 = \operatorname{diag}(3,-3) = 3\mathbf{D}$, so $\mathbf{A}_2$ scales the form rather than preserving it.)

$\mathbf{A}_3$: first column as for $\mathbf{A}_1$; second column $(-4/3)^2 - (-5/3)^2 = -1$; and $(5/3)(-4/3) - (4/3)(-5/3) = 0$. So $\mathbf{A}_3 \in \mathrm{O}(1,1)$, with $t = \log 3$, $\varepsilon = 1$, and $\mu = \det \mathbf{A}_3 = -25/9 + 16/9 = -1$; that is, $\mathbf{A}_3 = \mathbf{B}_{\log 3}\mathbf{D}$.

Exercise. B2

(a) Classify $\mathbf{x} = (3,1,2,2)$, $\mathbf{y} = (1,0,1,0)$, and $\mathbf{z} = (0,1,1,1)$ in $\mathbb{R}^{1,3}$ as timelike, null, or spacelike.

(b) The vector $\mathbf{y}$ is null. Find a $\mathbf{w} \in \mathbb{R}^{1,3}$ such that $(\mathbf{y}, \mathbf{w})$ is a basis of a hyperbolic plane, that is, $q(\mathbf{w}) = 0$ and $\eta(\mathbf{y}, \mathbf{w}) = 1$.

Solution.

(a) $q(\mathbf{x}) = 9 - 1 - 4 - 4 = 0$, so $\mathbf{x}$ is null. $q(\mathbf{y}) = 1 - 0 - 1 - 0 = 0$, so $\mathbf{y}$ is null. $q(\mathbf{z}) = 0 - 1 - 1 - 1 = -3 < 0$, so $\mathbf{z}$ is spacelike.

(b) Try $\mathbf{w} = (c, 0, -c, 0)$, which has $q(\mathbf{w}) = c^2 - c^2 = 0$ for every c. Then $\eta(\mathbf{y}, \mathbf{w}) = 1 \cdot c - 1 \cdot (-c) = 2c$, so $c = \frac{1}{2}$ works: $\mathbf{w} = (\frac{1}{2}, 0, -\frac{1}{2}, 0)$. The Gram matrix of $(\mathbf{y}, \mathbf{w})$ is $\left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$, so $\operatorname{span}(\mathbf{y}, \mathbf{w})$ is a hyperbolic plane (Definition 14.7.3).

Exercise. B3

Let $\mathbf{x} = (5,3,0,0)$ and $\mathbf{y} = (13,12,0,0)$ in $\mathbb{R}^{1,3}$. Compute $\tau(\mathbf{x})$, $\tau(\mathbf{y})$, $\eta(\mathbf{x},\mathbf{y})$, and $\tau(\mathbf{x}+\mathbf{y})$, and check Lemma 14.11.2 and Theorem 14.11.3 on this pair. Is either inequality an equality?

Solution. $q(\mathbf{x}) = 25 - 9 = 16$, so $\mathbf{x}$ is timelike and future-pointing with $\tau(\mathbf{x}) = 4$. $q(\mathbf{y}) = 169 - 144 = 25$, so $\tau(\mathbf{y}) = 5$. Next $\eta(\mathbf{x},\mathbf{y}) = 5\cdot 13 - 3\cdot 12 = 65 - 36 = 29$, and $\tau(\mathbf{x})\tau(\mathbf{y}) = 20$, so $29 \geq 20$ as Lemma 14.11.2 requires. Finally $\mathbf{x}+\mathbf{y} = (18,15,0,0)$ with $q = 324 - 225 = 99$, so $\tau(\mathbf{x}+\mathbf{y}) = \sqrt{99}$. Since $9^2 = 81 < 99$, we get $\tau(\mathbf{x}+\mathbf{y}) > 9 = \tau(\mathbf{x}) + \tau(\mathbf{y})$, as Theorem 14.11.3 requires. Neither is an equality, and indeed $\mathbf{y}$ is not a multiple of $\mathbf{x}$, since $13/5 \neq 12/3$.

14.11.6.3 C. Going deeper

Exercise. C1

For $t \in \mathbb{R}$ put $v(t) = \tanh t = \sinh t / \cosh t$.

(a) Prove that $v\colon \mathbb{R} \to (-1,1)$ is a bijection, and that $\mathbf{B}_t\mathbf{e}_1$ is a positive multiple of $(1, v(t), 0, 0)$ read in $\mathbb{R}^{1,1}$.

(b) Hence deduce that $v(s+t) = \dfrac{v(s)+v(t)}{1+v(s)v(t)}$, and say which part of Theorem 14.11.1 does the work.

Solution.

(a) Since $\cosh t > 0$ and $\cosh^2 t - \sinh^2 t = 1$, we have $|\sinh t| < \cosh t$, so $|v(t)| < 1$ and v maps into $(-1,1)$. For injectivity and surjectivity, solve $v(t) = c$ with $|c| < 1$: the equations $\sinh t = c\cosh t$ and $\cosh^2 t - \sinh^2 t = 1$ give $\cosh^2 t\,(1-c^2) = 1$, so $\cosh t = (1-c^2)^{-1/2}$, which is a legitimate value since it is ≥ 1, and then $\sinh t = c(1-c^2)^{-1/2}$ determines t uniquely because sinh is a bijection. So each c has exactly one preimage. In $\mathbb{R}^{1,1}$, $\mathbf{B}_t\mathbf{e}_1 = (\cosh t, \sinh t) = \cosh t\,(1, v(t))$, and $\cosh t > 0$.

(b) By Theorem 14.11.1, $\mathbf{B}_s\mathbf{B}_t = \mathbf{B}_{s+t}$. Comparing entries,

$$v(s+t) = \frac{\sinh(s+t)}{\cosh(s+t)} = \frac{\sinh s\cosh t + \cosh s \sinh t}{\cosh s\cosh t + \sinh s \sinh t},$$

and dividing numerator and denominator by $\cosh s \cosh t > 0$ gives $(v(s)+v(t))/(1+v(s)v(t))$. The work is done by the addition law $\mathbf{B}_s\mathbf{B}_t = \mathbf{B}_{s+t}$; the parametrization merely renames t as $v(t)$, and because v is a bijection onto $(-1,1)$ no information is lost in the renaming.

Exercise. C2

(a) Prove that $\{\mathbf{P}\mathbf{e}_1 : \mathbf{P} \in \mathrm{O}(1,1)\}$ is the set of all $\mathbf{x} \in \mathbb{R}^{1,1}$ with $q(\mathbf{x}) = 1$, and that this set is unbounded.

(b) Deduce that no bound $|p_{ij}| \leq c$ holds for all $\mathbf{P} \in \mathrm{O}(1,1)$, and contrast this with $\mathrm{O}(2)$, where $|p_{ij}| \leq 1$ always. Which hypothesis of Chapter 11's $\mathrm{O}(2)$ theory is responsible?

Solution.

(a) ($\subseteq$) If $\mathbf{P} \in \mathrm{O}(1,1)$ then $q(\mathbf{P}\mathbf{e}_1) = (\mathbf{P}\mathbf{e}_1)^{\top}\mathbf{D}(\mathbf{P}\mathbf{e}_1) = \mathbf{e}_1^{\top}\mathbf{P}^{\top}\mathbf{D}\mathbf{P}\mathbf{e}_1 = \mathbf{e}_1^{\top}\mathbf{D}\mathbf{e}_1 = 1$. ($\supseteq$) If $q(\mathbf{x}) = 1$, the proof of Theorem 14.11.1 shows $\mathbf{x} = (\varepsilon\cosh t, \sinh t)$ for some t and sign ε; then $\mathbf{x}$ is the first column of $\mathbf{B}_t$ when $\varepsilon = 1$, and of $-\mathbf{B}_{-t}$ when $\varepsilon = -1$, and both lie in $\mathrm{O}(1,1)$. The set is unbounded because $\cosh t \to \infty$ as $t \to \infty$.

(b) The first column of $\mathbf{P}$ ranges over an unbounded set by (a), so its entries are unbounded, and no such c exists; this is the unboundedness Section 10 recorded, obtained here from the orbit rather than from a one-parameter family. In $\mathrm{O}(2)$ the columns are unit vectors for the Euclidean norm, so every entry has modulus at most 1 and the group sits inside a bounded region. The responsible hypothesis is **positive definiteness** of the form: it is what makes $q(\mathbf{x}) = 1$ a bounded set, namely the unit sphere. Once one sign is reversed, the level set is a hyperbola and nothing is bounded.

Exercise. C3

The Quick check above showed that $\mathbb{R}^{1,3}$ has no plane on which the form vanishes identically. Show that $\mathbb{R}^{2,2}$ does have one, by exhibiting two vectors that span it, and reconcile the two answers with Theorem 14.7.5.

Hint: pair each positive coordinate with a negative one.

Solution. In $\mathbb{R}^{2,2}$ the form is $\eta'(\mathbf{x},\mathbf{y}) = x_1y_1 + x_2y_2 - x_3y_3 - x_4y_4$. Put $\mathbf{u} = (1,0,1,0)$ and $\mathbf{w} = (0,1,0,1)$. Then

$$q(\mathbf{u}) = 1 - 1 = 0, \quad q(\mathbf{w}) = 1 - 1 = 0, \quad \eta'(\mathbf{u},\mathbf{w}) = 0,$$

the last because every term pairs a zero with something. The list $(\mathbf{u},\mathbf{w})$ is independent, since a relation $a\mathbf{u} + b\mathbf{w} = \mathbf{0}$ reads $(a,b,a,b) = \mathbf{0}$. So $\operatorname{span}(\mathbf{u},\mathbf{w})$ is a totally isotropic plane, and by bilinearity η' vanishes on all of it.

The reconciliation: by Definition 14.7.4 the Witt index bounds the dimension of a totally isotropic subspace, it counts the hyperbolic planes of Theorem 14.7.5, and over $\mathbb{R}$ it equals $\min(n_+, n_-)$ (Proposition 14.7.7). For $\mathbb{R}^{1,3}$ this is $\min(1,3) = 1$, so totally isotropic subspaces are lines at most; for $\mathbb{R}^{2,2}$ it is $\min(2,2) = 2$, so a totally isotropic plane is exactly what one should expect, and the two vectors above realize it. What differs is not the dimension 4 but how evenly the four signs are split.

14.12 A First Look at Clifford Algebras

The chapter ends with a construction rather than a classification. Everything so far has taken a quadratic form apart; this section asks instead for an algebra in which the form becomes a *square*, so that $q(v)$ is literally v times v for a multiplication we invent. The answer is a Clifford algebra, and two of the most familiar algebras in mathematics are its first two examples. This is an introduction only: we define the object, prove what the definition alone gives, work three cases out completely, and stop.

14.12.1 Making a form into a square

Take $V = \mathbb{R}^2$ with the negative definite form $q(\mathbf{x}) = -(x_1^2 + x_2^2)$, and suppose we had an associative multiplication on some larger space containing V, with an identity 1, in which every vector satisfies $v^2 = q(v) \cdot 1$. Write ϵ_1, ϵ_2 for the two standard basis vectors sitting inside that algebra. Expanding $v = x_1\epsilon_1 + x_2\epsilon_2$ and using only bilinearity of the product,

$$v^2 = x_1^2\epsilon_1^2 + x_1x_2(\epsilon_1\epsilon_2 + \epsilon_2\epsilon_1) + x_2^2\epsilon_2^2.$$

We want this to equal $-(x_1^2 + x_2^2) \cdot 1$ for **every** x_1, x_2. Taking $x_2 = 0$ forces $\epsilon_1^2 = -1$, and taking $x_1 = 0$ forces $\epsilon_2^2 = -1$. What is left is $x_1x_2(\epsilon_1\epsilon_2 + \epsilon_2\epsilon_1) = 0$ for all x_1, x_2, which forces

$$\epsilon_1\epsilon_2 = -\epsilon_2\epsilon_1.$$

So the two basis vectors must square to -1 and must **anticommute**. That last relation hands us a third square root of -1 we did not ask for, since $(\epsilon_1\epsilon_2)^2 = \epsilon_1\epsilon_2\epsilon_1\epsilon_2 = -\epsilon_1^2\epsilon_2^2 = -1$. Three anticommuting square roots of -1: those are Hamilton's quaternions. Run the same argument on a one-dimensional V with $q(x) = -x^2$ and the single relation $\epsilon_1^2 = -1$ produces $\mathbb{C}$.

A Clifford algebra is the smallest place where the vectors of V can be multiplied and each one squares to its own q-value.

14.12.2 Algebras

The objects above are algebras, a word the book has used informally and now needs precisely.

Definition 14.12.1. Algebra over a Field

An **algebra over** F, or an F**-algebra**, is a vector space A over F together with a map $A \times A \to A$, written $(x, y) \mapsto xy$, which is

(A1) **bilinear**: linear in each argument with the other fixed;

(A2) **associative**: $(xy)z = x(yz)$ for all $x, y, z \in A$;

(A3) **unital**: there is an element $1_A \in A$ with $1_Ax = x1_A = x$ for every $x \in A$.

A **homomorphism of** F**-algebras** is an F-linear map $\varphi\colon A \to B$ with $\varphi(xy) = \varphi(x)\varphi(y)$ for all x, y and $\varphi(1_A) = 1_B$. An **isomorphism** is a bijective homomorphism.

Every algebra here is associative and unital, and we stop saying so. The examples are old friends: F itself; $M_n(F)$; $F[x]$; $\mathcal{L}(V)$ under composition; and $\mathbb{C}$ over $\mathbb{R}$, of dimension 2. A nonexample by minimal change: $\mathbb{R}^3$ with the cross product is bilinear but fails (A2) and (A3), since $(\mathbf{e}_1 \times \mathbf{e}_1) \times \mathbf{e}_2 = \mathbf{0}$ while $\mathbf{e}_1 \times (\mathbf{e}_1 \times \mathbf{e}_2) = -\mathbf{e}_2$, and no vector acts as an identity.

One consequence of (A1) is used constantly below: a bilinear map is determined by its values on a basis, so two products agree, and a linear map is multiplicative, as soon as this is checked on basis elements.

14.12.3 The definition

Definition 14.12.2. Clifford Algebra

Let F be a field with $\operatorname{char} F \neq 2$, let V be a finite-dimensional F-vector space, let β be a symmetric bilinear form on V and let $q(v) = \beta(v, v)$ be its quadratic form (Definition 14.3.2).

A **Clifford algebra** of (V, q) is a pair (C, ι), where C is an F-algebra and $\iota\colon V \to C$ is a linear map, such that

(C1) $\iota(v)^2 = q(v)1_C$ for **every** $v \in V$;

(C2) (**universal property**) for every F-algebra A and every linear map $f\colon V \to A$ with $f(v)^2 = q(v)1_A$ for every $v \in V$, there is a **unique** algebra homomorphism $\bar{f}\colon C \to A$ with $\bar{f} \circ \iota = f$.

We write $\mathrm{Cl}(q)$, or $\mathrm{Cl}(V, q)$, for a Clifford algebra of (V, q), and $\mathrm{Cl}_{p,m}(\mathbb{R})$ when V is the space $\mathbb{R}^{p,m}$ of Definition 14.11.1, that is, $\mathbb{R}^{p+m}$ with the real form of signature (p, m).

Clause by clause. (C1) is the demand we started from: inside C every vector has a square, and that square is a *number* times the identity, not a vector. (C2) says C is the most economical algebra with that property. Any other algebra where the vectors square correctly receives a map from C, so C has no relations beyond those (C1) forces; and the map is unique, so C has no room to spare either. What is defined is the pair (C, ι), not C alone.

Remark. The subscripts in $\mathrm{Cl}_{p,m}(\mathbb{R})$ are usually written p, q, but q already names the quadratic form here, so the second index is m.

Existence is deferred. The definition says what a Clifford algebra would be; it does not produce one. The standard construction is not available yet: one takes the **tensor algebra** of V — the infinite-dimensional algebra whose elements are formal sums of strings of vectors, multiplied by concatenation — and divides out by the relations $v \otimes v - q(v)1$. Chapter 15 builds tensor algebras and carries this out, for every V and q at once. So everything below is of one of two kinds: a consequence of the definition, valid whenever a Clifford algebra exists, or an explicit algebra written down by hand, which proves existence in that case. We say which each time.

The first consequence is that "the" Clifford algebra is legitimate language.

Proposition 14.12.1. Uniqueness of the Clifford Algebra

Let (C, ι) and (C', ι') both be Clifford algebras of (V, q). Then there is a unique algebra isomorphism $\varphi\colon C \to C'$ with $\varphi \circ \iota = \iota'$.

Idea. Each one's universal property produces a map to the other. The two composites are self-maps that do nothing to ι, and the uniqueness clause says the only such self-map is the identity.

Proof. Since (C', ι') satisfies (C1), the universal property (C2) of C, applied with $A = C'$ and $f = \iota'$, gives an algebra homomorphism $\varphi\colon C \to C'$ with $\varphi\iota = \iota'$. Symmetrically, (C2) for C' gives $\psi\colon C' \to C$ with $\psi\iota' = \iota$. Then $\psi\varphi\colon C \to C$ is an algebra homomorphism with $\psi\varphi\iota = \psi\iota' = \iota$. The identity id_C is another one. By the uniqueness clause of (C2) for C, applied with $A = C$ and $f = \iota$, the two coincide: $\psi\varphi = \mathrm{id}_C$. Swapping the roles of C and C' gives $\varphi\psi = \mathrm{id}_{C'}$. Hence φ is an isomorphism, and it is unique because (C2) already delivers it uniquely. This proves the proposition.

■

Next, the identity that does all the computing.

Lemma 14.12.2. The Anticommutation Relation

Let (C, ι) be a Clifford algebra of (V, q). Then for all $u, v \in V$,

$$\iota(u)\iota(v) + \iota(v)\iota(u) = 2\beta(u, v)1_C.$$

In particular $\iota(u)$ and $\iota(v)$ anticommute whenever $\beta(u, v) = 0$.

Proof. By (C1) applied to $u + v$, and then bilinearity of the product together with linearity of ι,

$$q(u + v)1_C = \iota(u + v)^2 = \iota(u)^2 + \iota(u)\iota(v) + \iota(v)\iota(u) + \iota(v)^2.$$

By Theorem 14.3.5, $q(u+v) = q(u) + 2\beta(u, v) + q(v)$, and by (C1) again $\iota(u)^2 = q(u)1_C$ and $\iota(v)^2 = q(v)1_C$. Canceling those two terms from both sides leaves the stated identity. The last sentence is the case $\beta(u, v) = 0$.

■

So an **orthogonal** basis for β becomes a list of pairwise anticommuting elements of C, each with a scalar square. That is the whole computational content of a Clifford algebra, and Theorem 14.4.2 guarantees such a basis exists whenever $\operatorname{char} F \neq 2$ — which is where the characteristic hypothesis in Definition 14.12.2 earns its place, since Theorem 14.3.5 needs it too.

14.12.4 Three algebras, worked out

Example. The zero space

Let $V = \{0\}$, so $q = 0$ and $n = 0$. Show that (F, ι) with $\iota = 0$ is a Clifford algebra of (V, q), so that $\mathrm{Cl}(0) = F$.

Solution. (C1): the only vector is 0, and $\iota(0)^2 = 0 = q(0)1_F$.

(C2): let A be an F-algebra and $f\colon V \to A$ linear, necessarily $f = 0$, so the condition on f reads $0 = 0$. Define $\bar{f}\colon F \to A$ by $\bar{f}(a) = a1_A$: it is linear, sends 1 to 1_A, and is multiplicative because $\bar{f}(ab) = (ab)1_A = (a1_A)(b1_A)$ by bilinearity of the product and $1_A 1_A = 1_A$; and $\bar{f}\iota = 0 = f$. It is unique, since any algebra homomorphism $\varphi\colon F \to A$ is linear with $\varphi(1) = 1_A$, hence $\varphi(a) = a1_A$.

So $\mathrm{Cl}(0) = F$, of dimension $1 = 2^0$: the construction adds nothing when there is nothing to add.

Example. The complex numbers

Let $V = \mathbb{R}$ with $q(x) = -x^2$, the form of signature $(0,1)$. Show that $\mathbb{C}$, with $\iota\colon \mathbb{R} \to \mathbb{C}$ given by $\iota(x) = xi$, is a Clifford algebra of (V, q), so that $\mathrm{Cl}_{0,1}(\mathbb{R}) = \mathbb{C}$.

Solution. Here $\mathbb{C}$ is an algebra over $\mathbb{R}$ of dimension 2, with basis $(1, i)$.

(C1): $\iota(x)^2 = (xi)^2 = x^2 i^2 = -x^2 = q(x) \cdot 1$.

(C2): let A be a real algebra and $f\colon \mathbb{R} \to A$ linear with $f(x)^2 = -x^2 1_A$. Put $u = f(1)$, so $f(x) = xu$ by linearity and $u^2 = -1_A$. Define $\bar{f}\colon \mathbb{C} \to A$ by $\bar{f}(a + bi) = a1_A + bu$, which is $\mathbb{R}$-linear and sends 1 to 1_A. To see it is multiplicative, note that $(z, w) \mapsto \bar{f}(zw)$ and $(z, w) \mapsto \bar{f}(z)\bar{f}(w)$ are both $\mathbb{R}$-bilinear, so it suffices to compare them on the basis $(1, i)$: the pairs involving 1 are immediate, and $\bar{f}(i \cdot i) = \bar{f}(-1) = -1_A = u^2 = \bar{f}(i)\bar{f}(i)$. Finally $\bar{f}\iota(x) = \bar{f}(xi) = xu = f(x)$. For uniqueness, an algebra homomorphism φ with $\varphi\iota = f$ has $\varphi(1) = 1_A$ and $\varphi(i) = \varphi(\iota(1)) = f(1) = u$, and $(1, i)$ spans $\mathbb{C}$, so $\varphi = \bar{f}$.

So $\mathrm{Cl}_{0,1}(\mathbb{R}) = \mathbb{C}$, of dimension $2 = 2^1$. Read this backwards: $\mathbb{C}$ is not primarily "$\mathbb{R}$ with a square root of -1 attached", it is the smallest algebra in which the number x squares to $-x^2$.

For the two-dimensional case we need the quaternions, and we need them to be associative. The cheapest way to get associativity is to find them inside a matrix algebra, where it is free.

Definition 14.12.3. The Quaternions

Inside $M_2(\mathbb{C})$, regarded as an algebra over $\mathbb{R}$, put

$$i = \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix}, \quad j = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \quad k = \begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix}.$$

The **quaternions** $\mathbb{H}$ are the real span of $\mathbf{I}_2, i, j, k$ in $M_2(\mathbb{C})$.

The three matrices carry Hamilton's names and are written plain, as elements of an algebra are and as the complex unit is. The i in their entries is that complex unit, a scalar, which is a different object from the matrix i. The two look alike, and the place tells them apart: inside a matrix entry or a complex number such as $a + bi$ the letter is the scalar, and beside $\mathbf{I}_2, j, k$ it is the matrix. Multiplying them out gives

$$i^2 = j^2 = k^2 = -\mathbf{I}_2, \qquad ij = k = -ji,$$

and then $jk = i = -kj$ and $ki = j = -ik$. A real combination $a\mathbf{I}_2 + bi + cj + dk$ is the matrix with rows $(a + bi,\ c + di)$ and $(-c + di,\ a - bi)$, which vanishes only when $a = b = c = d = 0$; so $\dim_{\mathbb{R}} \mathbb{H} = 4$. The relations show the span is closed under multiplication, so $\mathbb{H}$ is a real subalgebra of $M_2(\mathbb{C})$ — and it is associative for free, because matrix multiplication is.

Example. The quaternions

Let $V = \mathbb{R}^2$ with $q(\mathbf{x}) = -(x_1^2 + x_2^2)$, the form of signature $(0,2)$. Show that $\mathbb{H}$, with $\iota(\mathbf{x}) = x_1 i + x_2 j$, is a Clifford algebra of (V, q), so that $\mathrm{Cl}_{0,2}(\mathbb{R}) = \mathbb{H}$.

Solution. (C1): using $ij + ji = 0$ and $i^2 = j^2 = -\mathbf{I}_2$,

$$\iota(\mathbf{x})^2 = x_1^2 i^2 + x_1 x_2 (ij + ji) + x_2^2 j^2 = -(x_1^2 + x_2^2)\mathbf{I}_2,$$

which is $q(\mathbf{x})1_{\mathbb{H}}$.

(C2): let A be a real algebra and $f\colon \mathbb{R}^2 \to A$ linear with $f(\mathbf{x})^2 = q(\mathbf{x})1_A$. Put $u = f(\mathbf{e}_1)$ and $w = f(\mathbf{e}_2)$, so $u^2 = w^2 = -1_A$, and by Lemma 14.12.2 — or directly, by expanding $f(\mathbf{e}_1 + \mathbf{e}_2)^2 = -2 \cdot 1_A$ — we get $uw + wu = 0$, that is, $wu = -uw$. Define the linear map

$$\bar{f}(a\mathbf{I}_2 + bi + cj + dk) = a1_A + bu + cw + d\,uw,$$

which is well defined because $(\mathbf{I}_2, i, j, k)$ is a basis of $\mathbb{H}$, and sends $1_{\mathbb{H}} = \mathbf{I}_2$ to 1_A. Both $(z, z') \mapsto \bar{f}(zz')$ and $(z, z') \mapsto \bar{f}(z)\bar{f}(z')$ are $\mathbb{R}$-bilinear, so multiplicativity need only be checked on the basis. The pairs involving $\mathbf{I}_2$ are immediate, and the remaining nine are

$$\begin{aligned}
ii &= -\mathbf{I}_2 \mapsto -1_A = u^2, & jj &= -\mathbf{I}_2 \mapsto -1_A = w^2,\\
ij &= k \mapsto uw, & ji &= -k \mapsto -uw = wu,\\
ik &= -j \mapsto -w = u(uw), & ki &= j \mapsto w = (uw)u,\\
jk &= i \mapsto u = w(uw), & kj &= -i \mapsto -u = (uw)w,\\
kk &= -\mathbf{I}_2 \mapsto -1_A = (uw)^2.
\end{aligned}$$

Each right-hand claim is one line of algebra in A, using associativity, $wu = -uw$, and $u^2 = w^2 = -1_A$. For instance $(uw)u = u(wu) = -u(uw) = -u^2 w = w$, and $(uw)^2 = u(wu)w = -u^2 w^2 = -1_A$.

Finally $\bar{f}\iota(\mathbf{x}) = x_1 u + x_2 w = f(\mathbf{x})$. For uniqueness, an algebra homomorphism φ with $\varphi\iota = f$ is forced on $\mathbf{I}_2$, on $i = \iota(\mathbf{e}_1)$, and on $j = \iota(\mathbf{e}_2)$, hence on $k = ij$, and these four span $\mathbb{H}$.

So $\mathrm{Cl}_{0,2}(\mathbb{R}) = \mathbb{H}$, of dimension $4 = 2^2$.

Warning. **A Clifford algebra is not commutative, and a product of two vectors is not a vector.** In $\mathrm{Cl}_{0,2}(\mathbb{R}) = \mathbb{H}$ the image $\iota(\mathbb{R}^2) = \mathrm{span}(i, j)$ is a plane, and $ij = k$ lies outside it; worse, $ij = -ji$, so the order matters. Two habits have to go: that multiplying two elements of V keeps you in V, and that $xy = yx$. Beyond the scalars $F \cdot 1_C$, nothing may be assumed to commute.

14.12.5 How big is it

The three examples all came out with $\dim = 2^n$. Half of that is a theorem we can prove.

Proposition 14.12.3. The Clifford Algebra Is Spanned by Sorted Products

Let (C, ι) be a Clifford algebra of (V, q), where $\dim V = n$ and $\mathrm{char}\, F \neq 2$, and let $(e_1, \dots, e_n)$ be an orthogonal basis of V for β. Then C is spanned over F by the 2^n products

$$\iota(e_{i_1})\iota(e_{i_2}) \cdots \iota(e_{i_k}), \qquad 1 \le i_1 < i_2 < \cdots < i_k \le n,$$

one for each subset $\{i_1, \dots, i_k\}$ of $\{1, \dots, n\}$, the empty product being 1_C. In particular

$\dim_F C \le 2^n$.

Idea. Two steps. ① Show that C is generated as an algebra by 1_C and $\iota(V)$ — this is not automatic, and it is where the *uniqueness* half of the universal property does its work. ② Sort: any product of basis vectors can be put in increasing order by swapping neighbors at the cost of a sign (Lemma 14.12.2), and a repeated neighbor collapses to a scalar (C1).

Proof. An orthogonal basis exists by Theorem 14.4.2, since $\operatorname{char} F \neq 2$. Write $d_i = q(e_i)$.

Step 1: C is generated by 1_C and $\iota(V)$. Let $C' \subseteq C$ be the set of all F-linear combinations of 1_C and of finite products $\iota(v_1)\cdots\iota(v_m)$ with $v_1, \dots, v_m \in V$. It is a subspace closed under multiplication and contains 1_C, hence is itself an F-algebra, and ι followed by the change of codomain gives a linear map $\iota' \colon V \to C'$ satisfying (C1). By the universal property of C, applied with $A = C'$ and $f = \iota'$, there is an algebra homomorphism $g \colon C \to C'$ with $g\iota = \iota'$. Let $j \colon C' \to C$ be the inclusion, an algebra homomorphism. Then $jg \colon C \to C$ is an algebra homomorphism with $jg\iota = \iota$, and so is id_C; by the uniqueness clause of (C2), applied with $A = C$ and $f = \iota$, we get $jg = \mathrm{id}_C$. Hence j is surjective, so $C' = C$.

Step 2: sorting. By Step 1 and linearity of ι, C is spanned by 1_C together with the products $\iota(e_{j_1})\cdots\iota(e_{j_m})$ over all finite index strings $(j_1, \dots, j_m)$. Fix such a product and induct on the pair $(m, \operatorname{inv}(j_1, \dots, j_m))$, ordered lexicographically, where inv counts the pairs $r < s$ with $j_r > j_s$. If the string is strictly increasing, the product is on the list and we are done. Otherwise some neighboring pair has $j_r \ge j_{r+1}$. If $j_r = j_{r+1}$, then (C1) replaces $\iota(e_{j_r})^2$ by the scalar d_{j_r}, producing a scalar multiple of a product of length $m - 2$. If $j_r > j_{r+1}$, then Lemma 14.12.2 with $\beta(e_{j_r}, e_{j_{r+1}}) = 0$ lets us swap the two factors at the cost of a sign, producing a product of the same length with one fewer inversion. Either way the induction applies, and the product is a scalar multiple of a listed one.

There are 2^n subsets of $\{1, \dots, n\}$, so the spanning list has 2^n members and $\dim_F C \le 2^n$. This proves the proposition.

■

Quick check. Where exactly does Proposition 14.12.3 use the word **unique** in (C2), and what would go wrong without it?

Solution. In Step 1, at the sentence "by the uniqueness clause of (C2) … we get $jg = \mathrm{id}_C$". Existence alone gives an algebra homomorphism $jg \colon C \to C$ fixing ι, but nothing then makes it the identity, and the argument cannot conclude that j is surjective. The conclusion really would be false. Take $V = \mathbb{R}$ with $q(x) = -x^2$, let $A = \mathbb{C} \times \mathbb{C}$ with componentwise multiplication, and set $\iota(x) = (xi, xi)$. Then $\iota(x)^2 = q(x)1_A$, so (C1) holds, and so does the existence half of (C2): the first projection, followed by the homomorphism $\mathbb{C} \to B$ built in Example (The complex numbers), does the job for every target B. But A is **not** spanned by 1_A and products of vectors, since every such product lies in the diagonal $\{(z, z)\}$, of dimension 2. What fails is uniqueness — the second projection is a second map — and that is the clause the proof needs.

The other half of the count is a different matter.

Theorem 14.12.4. Dimension of a Clifford Algebra

Let $\operatorname{char} F \neq 2$, let $\dim V = n$, and let (C, ι) be a Clifford algebra of (V, q). Then the 2^n sorted products of Proposition 14.12.3 form a **basis** of C, so $\dim_F C = 2^n$. In particular ι is injective.

What is proved here, and what is not. The inequality $\dim_F C \leq 2^n$ is Proposition 14.12.3, proved in full. The reverse inequality — that the 2^n sorted products are linearly independent — is **not proved in this section**, and it cannot be, because the definition alone does not even produce an algebra: independence and existence have to be established together, by constructing one algebra of dimension exactly 2^n that satisfies (C1) and (C2). Chapter 15 does that, from the tensor algebra. What this section does establish is the theorem in the three cases above: $n = 0$ with $\dim = 1$, $n = 1$ with $\dim = 2$, and the case $q = -x_1^2 - x_2^2$ of $n = 2$ with $\dim = 4$, each by writing the algebra down and checking (C1) and (C2) by hand. For $n \geq 3$, read the statement above as quoted from Chapter 15 and not as proved here.

Granting the theorem, the subsets of $\{1, \dots, n\}$ indexing the basis are the first sign that $\mathrm{Cl}(q)$ is assembled from the exterior powers $\Lambda^k V$, one for each k. Chapter 15 constructs both and makes that precise.

14.12.6 What Clifford algebras are for

One computation shows where the subject goes. Recall from Section 10 that the isometries of q form a group (Definition 14.10.1). Inside $\mathrm{Cl}(q)$, some of them are conjugations.

Proposition 14.12.5. Reflections Are Conjugations

Let (C, ι) be a Clifford algebra of (V, q) and let $u \in V$ with $q(u) \neq 0$. Then $\iota(u)$ is invertible in C, with inverse $q(u)^{-1}\iota(u)$, and for every $v \in V$,

$$-\iota(u)\iota(v)\iota(u)^{-1} = \iota\big(s_u(v)\big), \qquad s_u(v) \coloneqq v - \frac{2\beta(u,v)}{q(u)}\,u.$$

The map s_u fixes $\{v : \beta(u,v) = 0\}$ pointwise, sends u to $-u$, and satisfies $q(s_u(v)) = q(v)$.

Proof. By (C1), $\iota(u)^2 = q(u)1_C$ with $q(u) \neq 0$, so $\iota(u)\big(q(u)^{-1}\iota(u)\big) = 1_C$ and likewise on the other side. By Lemma 14.12.2, $\iota(u)\iota(v) = 2\beta(u,v)1_C - \iota(v)\iota(u)$, so multiplying on the right by $\iota(u)$,

$$\iota(u)\iota(v)\iota(u) = 2\beta(u,v)\iota(u) - q(u)\iota(v).$$

Dividing by $q(u)$ and negating gives $-\iota(u)\iota(v)\iota(u)^{-1} = \iota(v) - \frac{2\beta(u,v)}{q(u)}\iota(u)$, which is $\iota(s_u(v))$ by linearity of ι. If $\beta(u,v) = 0$ then $s_u(v) = v$; and $s_u(u) = u - 2u = -u$. Finally, writing $c = 2\beta(u,v)/q(u)$,

$$q(v - cu) = q(v) - 2c\beta(u,v) + c^2 q(u) = q(v),$$

since $c^2 q(u) = c \cdot 2\beta(u,v)$. This proves the proposition. ∎

So the reflection in a hyperplane, the basic isometry of a quadratic form, is carried out inside $\mathrm{Cl}(q)$ by conjugating with a single vector. Composing such conjugations reaches more isometries, and the invertible elements of $\mathrm{Cl}(q)$ acting on V this way form a group mapping onto the isometry group of q — two-to-one, since $\iota(u)$ and $-\iota(u)$ conjugate alike. That double covering is where **spinors** live: objects the covering group acts on and the isometry group does not, because going once around a rotation returns them with a sign. This book will not build them. What it will do is Chapter 15: construct tensor algebras, construct $\mathrm{Cl}(q)$ from one, and prove Theorem 14.12.4. Nothing later depends on this section, and the chapter ends here.

14.12.7 Exercises

14.12.7.1 A. Check your understanding

Exercise. A1

(a) State the definition of an F-algebra, and give one example of dimension 1 and one of dimension 4.

(b) State the universal property (C2), with every quantifier.

(c) What is $\mathrm{Cl}(0)$ when $V = \{0\}$, and what is its dimension?

(d) True or false: $\mathrm{Cl}_{0,2}(\mathbb{R})$ is commutative. Justify your answer.

(e) What does this book prove about $\dim \mathrm{Cl}(q)$ when $\dim V = 3$, and what does it only quote?

Solution.

(a) A vector space over F with a bilinear, associative, unital multiplication (Definition 14.12.1). Dimension 1: F itself. Dimension 4: $M_2(F)$, or $\mathbb{H}$ over $\mathbb{R}$.

(b) For every F-algebra A and every linear $f\colon V \to A$ such that $f(v)^2 = q(v)1_A$ for every $v \in V$, there exists a unique algebra homomorphism $\bar{f}\colon C \to A$ such that $\bar{f} \circ \iota = f$.

(c) $\mathrm{Cl}(0) = F$, of dimension $1 = 2^0$ (Example (The zero space)).

(d) False. It is $\mathbb{H}$ by Example (The quaternions), and $ij = k \neq -k = ji$.

(e) It proves $\dim \mathrm{Cl}(q) \leq 2^3 = 8$, by Proposition 14.12.3. The equality $\dim \mathrm{Cl}(q) = 8$, and indeed the existence of $\mathrm{Cl}(q)$ at all, is quoted from Chapter 15.

14.12.7.2 B. Practice

Exercise. B1: The split case

Let $V = \mathbb{R}$ with $q(x) = x^2$, the form of signature $(1, 0)$. Let $A = \mathbb{R} \times \mathbb{R}$ with componentwise addition and multiplication, and $\iota\colon \mathbb{R} \to A$, $\iota(x) = (x, -x)$. Prove that (A, ι) is a Clifford algebra of (V, q), and state its dimension.

Solution. A is a real algebra of dimension 2 with identity $1_A = (1, 1)$, and $\big(1_A, \iota(1)\big) = \big((1, 1), (1, -1)\big)$ is a basis, since $a(1, 1) + b(1, -1) = (a + b, a - b)$ vanishes only for $a = b = 0$.

(C1): $\iota(x)^2 = (x^2, x^2) = q(x)1_A$.

(C2): let B be a real algebra and $f\colon \mathbb{R} \to B$ linear with $f(x)^2 = x^2 1_B$. Put $u = f(1)$, so $f(x) = xu$ and $u^2 = 1_B$. Define $\bar{f}(a,b) = \frac{a+b}{2}1_B + \frac{a-b}{2}u$, which is linear and sends 1_A to 1_B and $\iota(1)$ to u. By bilinearity, multiplicativity need only be checked on that basis: the pairs involving 1_A are immediate, and $\iota(1)^2 = 1_A \mapsto 1_B = u^2$. Also $\bar{f}\iota(x) = xu = f(x)$, and any such homomorphism is forced on 1_A and $\iota(1)$, which span A.

So $\mathrm{Cl}_{1,0}(\mathbb{R}) = \mathbb{R}\times\mathbb{R}$, of dimension $2 = 2^1$. Comparing with Example (The complex numbers): the same one-dimensional V with the opposite sign of q gives $\mathbb{C}$ instead, so the Clifford algebra sees the sign of the form, not just its rank.

Exercise. B2

Work in $\mathrm{Cl}_{0,2}(\mathbb{R}) = \mathbb{H}$, with $\iota(\mathbf{x}) = x_1 i + x_2 j$. Let $\mathbf{u} = (1,2)$ and $\mathbf{v} = (3,-1)$.

(a) Compute $\iota(\mathbf{u})\iota(\mathbf{v})$ and $\iota(\mathbf{v})\iota(\mathbf{u})$, and verify Lemma 14.12.2 on this pair.

(b) Compute $q(\mathbf{u})$, verify $\iota(\mathbf{u})^2 = q(\mathbf{u})1_{\mathbb{H}}$, and write down $\iota(\mathbf{u})^{-1}$.

Solution.

(a) Using $i^2 = j^2 = -1$ and $ij = k = -ji$, and writing 1 for $\mathbf{I}_2$,

$$\iota(\mathbf{u})\iota(\mathbf{v}) = (i+2j)(3i-j) = -3 - k - 6k + 2 = -1 - 7k,$$
$$\iota(\mathbf{v})\iota(\mathbf{u}) = (3i-j)(i+2j) = -3 + 6k + k + 2 = -1 + 7k.$$

Their sum is -2. On the other side, $\beta(\mathbf{u},\mathbf{v}) = -(1\cdot 3 + 2\cdot(-1)) = -1$, so $2\beta(\mathbf{u},\mathbf{v})1_{\mathbb{H}} = -2$. The two agree.

(b) $q(\mathbf{u}) = -(1+4) = -5$. And $\iota(\mathbf{u})^2 = (i+2j)^2 = -1 + 2k - 2k - 4 = -5$, as required. Hence $\iota(\mathbf{u})^{-1} = q(\mathbf{u})^{-1}\iota(\mathbf{u}) = -\frac{1}{5}(i+2j)$ by Proposition 14.12.5.

Exercise. B3

In $\mathbb{R}^{0,2}$ let $\mathbf{u} = \mathbf{e}_1$ and $\mathbf{v} = \mathbf{e}_1 + \mathbf{e}_2$. Compute the reflection $s_{\mathbf{u}}(\mathbf{v})$ directly from the formula in Proposition 14.12.5, then compute $-\iota(\mathbf{u})\iota(\mathbf{v})\iota(\mathbf{u})^{-1}$ inside $\mathbb{H}$, and check that the two agree.

Solution. Here $\beta(\mathbf{x},\mathbf{y}) = -(x_1y_1 + x_2y_2)$, so $\beta(\mathbf{u},\mathbf{v}) = -1$ and $q(\mathbf{u}) = -1$. The formula gives

$$s_{\mathbf{u}}(\mathbf{v}) = \mathbf{v} - \frac{2(-1)}{-1}\mathbf{u} = \mathbf{v} - 2\mathbf{u} = -\mathbf{e}_1 + \mathbf{e}_2.$$

In $\mathbb{H}$: $\iota(\mathbf{u}) = i$, $\iota(\mathbf{v}) = i + j$, and $\iota(\mathbf{u})^{-1} = q(\mathbf{u})^{-1}i = -i$. Hence

$$-i(i+j)(-i) = i(i+j)i = (-1+k)i = -i + ki = -i + j,$$

using $ki = j$. That is $\iota(-\mathbf{e}_1 + \mathbf{e}_2)$, which agrees with $\iota(s_{\mathbf{u}}(\mathbf{v}))$. Geometrically $s_{\mathbf{u}}$ negates the $\mathbf{e}_1$ coordinate and fixes the $\mathbf{e}_2$ one, which is what the answer shows.

14.12.7.3 C. Going deeper

Exercise. C1: A Clifford algebra of matrices

In $M_2(\mathbb{R})$ put $\mathbf{E}_1 = \left(\begin{smallmatrix} 1 & 0 \\ 0 & -1 \end{smallmatrix}\right)$ and $\mathbf{E}_2 = \left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)$, and let $V = \mathbb{R}^2$ carry the form $q(\mathbf{x}) = x_1^2 - x_2^2$ of signature $(1,1)$, with $f(\mathbf{x}) = x_1\mathbf{E}_1 + x_2\mathbf{E}_2$.

(a) Prove that $f(\mathbf{x})^2 = q(\mathbf{x})\mathbf{I}_2$ for every $\mathbf{x}$, and that $(\mathbf{I}_2, \mathbf{E}_1, \mathbf{E}_2, \mathbf{E}_1\mathbf{E}_2)$ is a basis of $M_2(\mathbb{R})$.

(b) Assume a Clifford algebra (C, ι) of (V, q) exists. Deduce that $C \cong M_2(\mathbb{R})$, and hence that $\dim C = 4$.

Hint for (b): the homomorphism the universal property hands you has a surjective image; then count.

Solution.

(a) A direct multiplication gives $\mathbf{E}_1^2 = \mathbf{I}_2$, $\mathbf{E}_2^2 = -\mathbf{I}_2$ and

$$\mathbf{E}_1\mathbf{E}_2 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = -\mathbf{E}_2\mathbf{E}_1,$$

so $\mathbf{E}_1\mathbf{E}_2 + \mathbf{E}_2\mathbf{E}_1 = \mathbf{O}$ and

$$f(\mathbf{x})^2 = x_1^2\mathbf{E}_1^2 + x_1x_2(\mathbf{E}_1\mathbf{E}_2 + \mathbf{E}_2\mathbf{E}_1) + x_2^2\mathbf{E}_2^2 = (x_1^2 - x_2^2)\mathbf{I}_2.$$

For the basis: $a\mathbf{I}_2 + b\mathbf{E}_1 + c\mathbf{E}_2 + d\mathbf{E}_1\mathbf{E}_2$ is the matrix with rows $(a+b,\ c+d)$ and $(-c+d,\ a-b)$, and setting all four entries to zero forces $a = b = c = d = 0$. Four independent vectors in the 4-dimensional space $M_2(\mathbb{R})$ form a basis.

(b) By (a) and the universal property (C2) with $A = M_2(\mathbb{R})$, there is an algebra homomorphism $\bar{f}\colon C \to M_2(\mathbb{R})$ with $\bar{f}\iota = f$. Its image is a subalgebra containing $\mathbf{I}_2$, $\mathbf{E}_1 = f(\mathbf{e}_1)$, and $\mathbf{E}_2 = f(\mathbf{e}_2)$, hence also $\mathbf{E}_1\mathbf{E}_2$; by (a) those four span $M_2(\mathbb{R})$, so $\bar{f}$ is surjective and $\dim C \geq 4$. On the other hand $\dim C \leq 2^2 = 4$ by Proposition 14.12.3. Therefore $\dim C = 4$ and $\bar{f}$ is a surjective linear map between spaces of equal finite dimension, hence bijective, hence an algebra isomorphism. So $\mathrm{Cl}_{1,1}(\mathbb{R}) \cong M_2(\mathbb{R})$.

Exercise. C2: Why characteristic 2 is excluded

Explain why Definition 14.12.2 assumes $\mathrm{char}\, F \neq 2$, by naming the first statement of this section that fails without it and saying what replaces the missing input.

Solution. The first casualty is Lemma 14.12.2, whose proof cites Theorem 14.3.5 for $q(u+v) = q(u) + 2\beta(u,v) + q(v)$ — and that identity is how β is recovered from q, by a division by 2 that is unavailable in characteristic 2. Without it, the relation $\iota(u)\iota(v) + \iota(v)\iota(u) = 2\beta(u,v)1_C$ collapses to $\iota(u)\iota(v) + \iota(v)\iota(u) = 0$ and carries no information about the form at all. Proposition 14.12.3 then loses its sorting step in the form given, and Theorem 14.4.2, which supplied the orthogonal basis, has the same hypothesis.

What is missing is a definition. In characteristic 2 a quadratic form is not the same data as a symmetric bilinear form — Section 3 warned exactly this — so a quadratic form must be defined in its own right, as a function $q\colon V \to F$ with $q(av) = a^2q(v)$ whose associated $\beta(u,v) = q(u+v) - q(u) - q(v)$ is bilinear. Clifford algebras are defined in that setting too,

by the same (C1) and (C2), but this book does not define quadratic forms there and so does not go on.

Chapter 15

Tensors and Exterior Algebra

Three times already this book has built an object out of one construction and then walked past the construction without naming it. Chapter 7 characterized the determinant as the **unique** alternating n-linear form sending the standard basis to 1 (Theorem 7.4.2), and got a number out of a function of n vectors. Chapter 8 defined the Kronecker product $\mathbf{A} \otimes \mathbf{B}$ by writing $\mathbf{A}$ with its entries replaced by blocks, borrowed the symbol $\otimes$, and admitted that the symbol was borrowed from something it had not yet met. Chapter 14 defined a Clifford algebra by saying what maps out of it must look like, proved that at most one such algebra exists, and then could not produce one: the construction needed an algebra of "formal sums of strings of vectors", and no such algebra had been built. The missing object is the same in all three places. It is the **tensor product**, and this chapter builds it.

The problem it solves is stated in one line. A bilinear map $\beta\colon V \times W \to Z$ is not a linear map, so nothing from Chapters 2 to 5 applies to it: it has no matrix, no kernel, no rank, and $\beta(2v, 2w)$ is $4\beta(v, w)$ rather than $2\beta(v, w)$. Multilinearity is everywhere in the book — the determinant, the dot product, matrix multiplication, every bilinear form of Chapter 14 — and each time we have had to develop it by hand, because the general theory of linear maps was unavailable. The tensor product ends that. It manufactures a single space $V \otimes W$ and a single bilinear map $(v, w) \mapsto v \otimes w$ into it, through which **every** bilinear map out of $V \times W$ factors, uniquely, as a linear map. Bilinear maps on $V \times W$ become linear maps on $V \otimes W$. From there the whole chapter is bookkeeping: repeat the construction k times to get the tensor algebra, then divide it by the relation $v \otimes w = w \otimes v$ to get symmetric powers, or by the relation $v \otimes v = 0$ to get exterior powers, where the determinant reappears as an induced map and the Clifford algebra finally gets built.

The method of this chapter is new, and §02 is where it is learned. Until now every object in the book was defined by a formula or by a construction: the determinant by the Leibniz sum, the quotient by cosets, the Jordan form by an algorithm. From here on we define objects by what maps out of them look like — a **universal property** — and the order of business inverts. First we say what the object must do; then we prove that any two objects doing it are isomorphic by a *unique* isomorphism, so that "the" object is legitimate; and only then do we build one, to know that the definition is not empty. Chapter 14 §12 already ran this argument once, for Clifford algebras (Proposition 14.12.1), without saying that it was a pattern. It is a pattern. Sections 02, 06, 07, 08, and 10 each execute it in exactly the same three moves, on five different objects. Theorem 15.2.1 in §02 is the template, written out at full length; the four later repetitions are compressed, and they are compressed on purpose, so that the reader who has absorbed the template recognizes it rather than rereads it. It is worth learning here because it is the only way to define these objects that does not depend on a choice of basis, and because a proof from a universal property is usually three lines long where a proof from a construction

is a page.

The field is a general F throughout. Where a statement genuinely needs $\operatorname{char} F = 0$ or $\operatorname{char} F \neq 2$ — and some do, in §§07, 08, and 10 — the hypothesis is stated and its job is named, as Chapter 14 did. Everything else holds over every field.

What you need. From Chapter 2, the space of linear maps (Definition 2.4.1, Theorem 2.4.1) and rank with Rank–Nullity (Definition 2.3.2, Theorem 2.3.1). From Chapter 4, quotient spaces and their universal property, which is the engine of every construction here (Definition 4.4.3, Lemma 4.4.3, Theorem 4.4.4, Theorem 4.4.10, Theorem 4.4.5, Theorem 4.4.7), the space of linear maps seen as matrices (Theorem 4.1.4), the matrix of a map (Definition 4.1.1, Theorem 4.1.1), products of spaces (Definition 4.4.1, Theorem 4.4.1), and the trace (Definition 4.5.2, Theorem 4.2.5). From Chapter 1, bases in any dimension and the freedom to prescribe a linear map on one (Definition 1.5.1, Theorem 1.6.4, Theorem 1.8.2, Theorem 2.1.6, Theorem 1.2.1). From Chapter 5, duality, used constantly: Definition 5.1.2, Definition 5.1.1, Theorem 5.1.3, Theorem 5.1.5, Definition 5.4.1, Theorem 5.4.2, Theorem 5.4.6. From Chapter 7, multilinearity and alternation, which §01 extends rather than redefines (Definition 7.2.1, Definition 7.2.2, Definition 7.2.3, Theorem 7.2.1, Theorem 7.2.2), and the determinant with its Leibniz formula and multiplicativity (Definition 7.4.1, Theorem 7.4.3, Theorem 7.4.2, Theorem 7.5.4, Theorem 7.4.5, Theorem 7.6.1, Theorem 7.7.1). From Chapter 8, the Kronecker product, whose two open promises §03 pays (Definition 8.4.1, Theorem 8.4.1, Theorem 8.4.3, Definition 8.4.2, Theorem 8.4.4). From Chapter 13, the singular value decomposition, which §11 turns out to be reading in another language (Theorem 13.9.1), and, for the same section, from Chapter 11 the orthogonal apparatus that goes with it: the inner product and Pythagoras, Gram–Schmidt and orthonormal coordinates, and the orthogonal decomposition (Definition 11.1.1, Theorem 11.1.3, Theorem 11.2.2, Theorem 11.2.1, Theorem 11.3.2). From Chapter 14, algebras, forms, and the Clifford algebra whose existence §10 supplies (Definition 14.12.1, Definition 14.1.1, Definition 14.3.2, Definition 14.3.1, Definition 14.12.2, Proposition 14.12.1, Proposition 14.12.3, Theorem 14.12.4). From Chapter 6, the polynomial ring, which §07 identifies with a symmetric algebra (Definition 0.8.2).

15.0.1 Roadmap

- **Multilinear maps.** Chapter 7's multilinear forms, extended from $V^k \to F$ to $V_1 \times \cdots \times V_k \to W$; the space they form, its dimension $\dim W \cdot \prod_i \dim V_i$, and the problem that they are not linear maps.
- **The tensor product.** The universal property, uniqueness up to unique isomorphism, the construction as a free space modulo the bilinearity relations, the basis $(v_i \otimes w_j)$, and three warnings about what a tensor is not.
- **Tensor products of maps.** $S \otimes T$ built from the universal property; its matrix is exactly the Kronecker product of Chapter 8, which settles two debts; and the identification $V^* \otimes W \cong \mathcal{L}(V, W)$ that everything later uses.
- **Contraction.** Evaluation $V^* \otimes V \to F$, extended to mixed tensors, and the discovery that under the previous identification it is the trace — the basis-free definition of trace the book has wanted since Chapter 4.
- **Index notation.** Components, upper and lower indices, the summation convention, the transformation law derived rather than asserted, and the older definition of a tensor proved equivalent to ours.
- **The tensor algebra.** $T(V) = \bigoplus_k V^{\otimes k}$: graded, associative, noncommutative, infinite-dimensional, and free. Its universal property is what the next three sections quotient.
- **Symmetric powers.** $\operatorname{Sym}^k V$ by killing $v \otimes w - w \otimes v$; a basis indexed by nondecreasing tuples; the symmetric algebra is the polynomial ring of Chapter 6; and what goes wrong

in characteristic p.

- **Exterior powers.** $\Lambda^k V$ by killing $v \otimes v$ — the characteristic-free choice, exactly Chapter 14's distinction between alternating and skew. A basis indexed by increasing tuples, $\dim \Lambda^k V = \binom{n}{k}$, and the test that $v_1 \wedge \cdots \wedge v_k \neq 0$ exactly when the list is independent.
- **The determinant, again.** $\Lambda^n V$ is a line, and $\Lambda^n T$ is multiplication by $\det T$. Multiplicativity becomes a one-line consequence, Laplace expansion falls out of Λ^k, and the coordinates of a wedge are the minors.
- **Grassmann and Clifford.** The exterior algebra ΛV, of dimension 2^n; elements of $\Lambda^2 V$ that are not decomposable; and $\mathrm{Cl}(q) = T(V)/\langle v \otimes v - q(v)1 \rangle$, which constructs the Clifford algebra of Chapter 14 and proves its dimension is 2^n.
- **Tensors in quantum mechanics.** Composite systems as tensor products, entanglement as failure of simplicity, and the Schmidt decomposition, which is the singular value decomposition of Chapter 13 wearing different notation.

15.0.2 Named moves

- **Define by universal property, then build a model.** State what maps out of the object must look like. Prove that any two objects with that property are isomorphic by a unique isomorphism, using nothing but the property itself. Then construct one object, so that the definition has content. The uniqueness proof is always the same four lines: each object's property produces a map to the other, each composite is a self-map fixing the structure map, and the uniqueness clause says the identity is the only such self-map.
- **A quotient kills exactly the relations you want.** To impose relations on a space, take the free object, take the subspace spanned by the relations written as "left side minus right side", and divide. A map respecting the relations kills that subspace, so it factors through the quotient. This chapter does it three times: tensor, symmetric, exterior.
- **To define a map out of a quotient, define it upstairs first.** Never define a map on cosets, or on simple tensors, by a formula and hope it is well defined. Define it on the free object, where a map may be prescribed arbitrarily on a basis, check that it kills the relations, and let the universal property hand you the induced map.
- **Count a basis by counting index tuples.** Every dimension formula in this chapter comes from counting tuples: all k-tuples for $V^{\otimes k}$, giving n^k; nondecreasing ones for $\mathrm{Sym}^k V$, giving $\binom{n+k-1}{k}$; strictly increasing ones for $\Lambda^k V$, giving $\binom{n}{k}$. Spanning is a rewriting argument, independence is a supply of linear functionals, and the count is combinatorics.

15.1 Multilinear Maps

Chapter 7 needed functions of several vectors that were linear in each vector separately, and called them multilinear forms (Definition 7.2.1). Every one of them landed in the field F: the determinant, the signed area, a product of functionals. This chapter needs the same idea with the output allowed to be any vector space, because three of the most familiar operations in the book — matrix multiplication, composition of linear maps, the cross product — are linear in each argument and return something that is not a scalar. So we begin by extending Chapter 7's definition, count how many such maps there are, and then state the problem that the rest of the chapter exists to solve: a multilinear map is **not** a linear map, so none of the machinery of Chapters 2 to 5 is available to it.

Throughout, F is a field, $k \geq 1$ is an integer, and all vector spaces are over F. From the third subsection on, every space named is finite-dimensional.

15.1.1 From forms to maps

Here are three operations the book has used constantly.

- **Matrix multiplication.** For fixed sizes, $(\mathbf{A}, \mathbf{B}) \mapsto \mathbf{AB}$ satisfies $(\mathbf{A}+\mathbf{A}')\mathbf{B} = \mathbf{AB}+\mathbf{A}'\mathbf{B}$ and $\mathbf{A}(\mathbf{B}+\mathbf{B}') = \mathbf{AB}+\mathbf{AB}'$, and scalars pull out of either factor.
- **Composition of linear maps.** $(S, T) \mapsto ST$ satisfies the same four rules.
- **The cross product** on $\mathbb{R}^3$, which is linear in each factor and returns a vector.

Each is linear in one argument while the other is held fixed. Each therefore looks exactly like Definition 7.2.1 with $k = 2$ — except that the value is a matrix, a linear map, or a vector, not a scalar. Chapter 7 had no reason to allow that, because it was building the determinant. We do, so we widen the definition. Nothing about Chapter 7's statements changes; a multilinear **form** will simply be the special case where the target is F, and we will keep using Definition 7.2.1 by that name whenever the target really is F.

A multilinear map is a function of several vectors that is linear in each one while the others are held fixed, with values in a vector space rather than in the field.

Definition 15.1.1. Multilinear Map

Let $V_1, \dots, V_k$ and W be vector spaces over the **same** field F, with $k \geq 1$. A function

$$\beta \colon V_1 \times V_2 \times \cdots \times V_k \to W$$

is k**-linear**, or a **multilinear map**, if **for every** position $i \in \{1, \dots, k\}$ and **every** choice of fixed vectors $v_j \in V_j$ for the positions $j \neq i$, the function

$$V_i \to W, \qquad x \mapsto \beta(v_1, \dots, v_{i-1}, x, v_{i+1}, \dots, v_k),$$

is linear. A 2-linear map is called **bilinear**.

When $V_1 = \cdots = V_k = V$ we say β is k-linear **on** V; when in addition $W = F$, the definition is word for word Definition 7.2.1, and we call β a multilinear **form**.

In words, clause by clause. **Same field:** all $k+1$ spaces are over one F, since otherwise "linear" would mean different things in different slots. **For every position:** the condition is imposed k separate times, once per slot. **Every choice of fixed vectors:** the other $k-1$ arguments are frozen at arbitrary values, not at convenient ones. **Is linear:** written out, for all $x, y \in V_i$ and all $c \in F$,

$$\beta(\dots, x+y, \dots) = \beta(\dots, x, \dots) + \beta(\dots, y, \dots),$$
$$\beta(\dots, cx, \dots) = c\,\beta(\dots, x, \dots),$$

where the dots stand for the same frozen vectors on both sides. The spaces $V_1, \dots, V_k$ are allowed to be different from each other and from W; nothing forces k to be related to any dimension.

Two consequences are free. Putting $c = 0$ in slot i gives $\beta(\dots, 0, \dots) = 0$: a multilinear map kills any tuple with a zero entry. And by induction on the number of terms, expanding one slot at a time,

$$\beta\Big(\sum_{i_1} a_{i_1} x_{i_1},\ \dots,\ \sum_{i_k} a_{i_k} y_{i_k}\Big) = \sum_{i_1, \dots, i_k} a_{i_1} \cdots a_{i_k}\, \beta(x_{i_1}, \dots, y_{i_k}),$$

a sum over all tuples of indices, one index per slot. This identity is the whole of the section: it says that a multilinear map is determined by its values on tuples drawn from spanning sets, and that the number of such tuples is a product, not a sum.

Example. Seven functions of several vectors

Decide which of the following are multilinear, and with which k.

(a) Any linear map $T\colon V \to W$, viewed as a function of one vector.

(b) Scalar multiplication $F \times V \to V$, $(c, v) \mapsto cv$, with F regarded as a vector space over itself.

(c) Matrix multiplication $M_{m\times n}(F) \times M_{n\times p}(F) \to M_{m\times p}(F)$, $(\mathbf{A}, \mathbf{B}) \mapsto \mathbf{AB}$.

(d) Composition $\mathcal{L}(W, Z) \times \mathcal{L}(V, W) \to \mathcal{L}(V, Z)$, $(S, T) \mapsto ST$.

(e) The determinant $(F^n)^n \to F$ sending $(\mathbf{a}_1, \dots, \mathbf{a}_n)$ to $\det \mathbf{A}$, where $\mathbf{A}$ has columns $\mathbf{a}_1, \dots, \mathbf{a}_n$.

(f) The evaluation pairing $V^* \times V \to F$, $(\varphi, v) \mapsto \varphi(v)$.

(g) The cross product $\mathbb{R}^3 \times \mathbb{R}^3 \to \mathbb{R}^3$.

Solution.

(a) With $k = 1$ there is one position and no other argument to freeze, so the condition reads: $x \mapsto Tx$ is linear. That is the hypothesis. So a 1-linear map is **exactly** a linear map (Definition 2.1.1), and Definition 15.1.1 generalizes it.

(b) Fix v. Then $c \mapsto cv$ is linear by (VS6) and (VS7) of Definition 1.1.1, which give $(a+b)v = av+bv$ and $a(cv) = (ac)v$. Fix c. Then $v \mapsto cv$ is linear by (VS5) and (VS7), the second of which gives $c(av) = (ca)v = (ac)v = a(cv)$, using that F is commutative. So scalar multiplication is bilinear. It is the smallest interesting example, and it shows that the two input spaces need not be equal.

(c) Fix $\mathbf{B}$. Then $\mathbf{A} \mapsto \mathbf{AB}$ is linear, since $(\mathbf{A} + \mathbf{A}')\mathbf{B} = \mathbf{AB} + \mathbf{A}'\mathbf{B}$ and $(c\mathbf{A})\mathbf{B} = c(\mathbf{AB})$ by Theorem 0.9.4. Fixing $\mathbf{A}$ gives the other slot. So matrix multiplication is bilinear, with three different spaces in play when m, n, p differ.

(d) Same check, with Theorem 2.4.2 (a) supplying linearity of ST and Theorem 2.4.2 (d) supplying the two slots. Bilinear.

(e) The determinant is n-linear in the columns, and alternating, by Theorem 7.4.3. It is the standard example of a multilinear form with k equal to the dimension. Here $k = n$ is forced by the shape of a square matrix, not by the definition.

(f) Fix v. Then $\varphi \mapsto \varphi(v)$ is linear because the operations on V^* are pointwise (Definition 5.1.2). Fix φ. Then $v \mapsto \varphi(v)$ is linear because φ is a linear functional. Bilinear. This one-line example is the source of the contraction operation of a later section.

(g) The cross product of Chapter 7 has components $a_2b_3 - a_3b_2$ and its two cousins, each bilinear as in Example (Alternating Forms) (b). A vector of bilinear functions is bilinear, since linearity in a slot may be checked coordinate by coordinate. So it is bilinear, with values in $\mathbb{R}^3$ — an example that genuinely needs Definition 15.1.1 rather than Definition 7.2.1.

Nonexample, by minimal change. Keep $k = 2$, keep $V_1 = V_2 = W = F$, and change multiplication to addition: $\sigma(a, b) = a + b$. Freeze $b = 1$. The resulting function $a \mapsto a + 1$ sends 0 to 1, so it is not linear (Theorem 2.1.1), and the clause "is linear" fails in the first slot. What still works is that σ is linear as a function on the product space $F \times F$ of Definition 4.4.1. That is precisely the distinction the next subsection is about.

Why this definition. One could try to define a "map of several variables" by asking for linearity in all slots at once, that is, by asking for a linear map on $V_1 \times \cdots \times V_k$. The result would be a different and much poorer notion: it would exclude multiplication, composition, the determinant, and every inner product, and would include only sums of linear maps applied to the separate slots. Freezing all but one slot is what keeps products in the theory. The name says the same thing: *multi*, once per slot; *linear*, in that slot alone.

15.1.2 Multilinear is not linear

The product $V \times W$ is itself a vector space, with $(v, w) + (v', w') = (v + v', w + w')$ and $c(v, w) = (cv, cw)$ (Definition 4.4.1). So a function $\beta\colon V \times W \to Z$ has two different conditions it might satisfy, and they are stated on the same domain. They are not the same condition, and confusing them is the standard first error with tensors.

Side by side, for $\beta\colon V \times W \to Z$:

- **Linear on the product** means $\beta((v, w) + (v', w')) = \beta(v, w) + \beta(v', w')$ and $\beta(cv, cw) = c\,\beta(v, w)$. Both slots move together.
- **Bilinear** means $\beta(v + v', w) = \beta(v, w) + \beta(v', w)$, and the same in the second slot with the first frozen. One slot moves at a time.

Warning. **Do not add in both slots at once.** For a bilinear β, the correct expansion has **four** terms:

$$\beta(v + v', w + w') = \beta(v, w) + \beta(v, w') + \beta(v', w) + \beta(v', w').$$

Concretely, with $\beta(a, b) = ab$ on $F \times F$: $\beta(1+1, 1+1) = 4$, not $\beta(1, 1) + \beta(1, 1) = 2$. Scaling is worse: $\beta(cv, cw) = c^2\beta(v, w)$, so a bilinear map scales by the **square**. Pull out scalars one slot at a time, always.

Quick check. Is $\beta(\mathbf{x}, \mathbf{y}) = x_1y_1 + x_2$ a bilinear map $F^2 \times F^2 \to F$?

Solution. No. Freeze $\mathbf{y}$. Then $\mathbf{x} \mapsto y_1x_1 + x_2$ is a linear functional, so the first slot is fine. Freeze $\mathbf{x} = \mathbf{e}_2$. Then $\mathbf{y} \mapsto 0 \cdot y_1 + 1 = 1$ is the constant function 1, which sends $\mathbf{0}$ to 1 and so is not linear. The clause that fails is "is linear" at position $i = 2$. The extra summand x_2 does not involve $\mathbf{y}$ at all, and a bilinear map must vanish whenever an argument is $\mathbf{0}$.

15.1.3 Values on basis tuples

A linear map may be prescribed arbitrarily on a basis and is determined by those values (Theorem 2.1.6). The corresponding statement for multilinear maps replaces "basis vectors" by "tuples of basis vectors, one from each slot", and the proof is the expansion identity above read in both directions.

Theorem 15.1.1. Multilinear Maps from Values on Basis Tuples

Let $V_1, \ldots, V_k$ be finite-dimensional, with ordered bases $\mathcal{B}_i = (v_1^{(i)}, \ldots, v_{n_i}^{(i)})$ for $i = 1, \ldots, k$,

and let W be any vector space over F. For every family of vectors

$$w_{i_1 i_2 \dots i_k} \in W, \qquad 1 \le i_1 \le n_1, \ \dots, \ 1 \le i_k \le n_k,$$

there is **exactly one** k-linear map $\beta\colon V_1 \times \cdots \times V_k \to W$ with $\beta(v_{i_1}^{(1)}, \dots, v_{i_k}^{(k)}) = w_{i_1 \dots i_k}$ for every tuple of indices.

Idea. Uniqueness is the expansion identity: write each argument in its basis, expand one slot at a time, and every value of β is a combination of the prescribed ones. For existence we need a formula, and the coefficients in that expansion are exactly the coordinates, which Theorem 5.1.3 supplies as linear functionals. So multiply the right functionals together and use them as coefficients.

Proof. *Uniqueness.* Let β and β' be k-linear with the same values on all basis tuples, and let $x_i \in V_i$. Write $x_i = \sum_j a_j^{(i)} v_j^{(i)}$. Expanding each slot in turn by Definition 15.1.1,

$$\beta(x_1, \dots, x_k) = \sum_{i_1, \dots, i_k} a_{i_1}^{(1)} \cdots a_{i_k}^{(k)} \, w_{i_1 \dots i_k},$$

and the same computation applies to β'. Hence $\beta = \beta'$.

Existence. For each i, let $(\varphi_1^{(i)}, \dots, \varphi_{n_i}^{(i)})$ be the basis of V_i^* dual to $\mathcal{B}_i$, so that $\varphi_a^{(i)}(v_b^{(i)}) = \delta_{ab}$ and $\varphi_j^{(i)}(x)$ is the j-th coordinate of x (Theorem 5.1.3). Define

$$\beta(x_1, \dots, x_k) \coloneqq \sum_{i_1, \dots, i_k} \varphi_{i_1}^{(1)}(x_1) \cdots \varphi_{i_k}^{(k)}(x_k) \, w_{i_1 \dots i_k},$$

a finite sum of vectors of W, so a well-defined element of W.

It is k-linear: freeze every slot but the m-th. Each summand is then the vector $w_{i_1 \dots i_k}$ multiplied by a fixed scalar times $\varphi_{i_m}^{(m)}(x_m)$, and $\varphi_{i_m}^{(m)}$ is linear; a finite sum of linear functions of x_m, each scaled by a fixed vector, is linear in x_m. This holds for every m, which is the condition of Definition 15.1.1.

Its values are the prescribed ones: at $x_i = v_{j_i}^{(i)}$ the coefficient $\varphi_{i_1}^{(1)}(v_{j_1}^{(1)}) \cdots \varphi_{i_k}^{(k)}(v_{j_k}^{(k)})$ is 1 when $i_1 = j_1, \dots, i_k = j_k$ and 0 otherwise, so only the term $w_{j_1 \dots j_k}$ survives. This proves the theorem. ∎

So a multilinear map is "nothing but" a family of vectors of W indexed by tuples of basis indices, exactly as a linear map is nothing but a family indexed by single basis indices. The count of tuples is where the product in the next theorem comes from.

15.1.4 The space of multilinear maps

Two multilinear maps with the same slots and the same target can be added and scaled pointwise, and the results are again multilinear. That makes them a vector space, and we name it before counting it.

Definition 15.1.2. The Space of Multilinear Maps

Let $V_1, \dots, V_k$ and W be vector spaces over F. Write

$$\mathcal{M}(V_1, \dots, V_k; W)$$

for the set of all k-linear maps $V_1 \times \cdots \times V_k \to W$, with the **pointwise** operations

$$(\beta + \gamma)(x_1, \dots, x_k) \coloneqq \beta(x_1, \dots, x_k) + \gamma(x_1, \dots, x_k),$$

and $(c\beta)(x_1, \dots, x_k) \coloneqq c\,\beta(x_1, \dots, x_k)$. The semicolon separates the inputs from the output.

Well-definedness. The set of *all* functions $V_1 \times \cdots \times V_k \to W$ is a vector space under the pointwise operations, by the argument of Example (The Standard Vector Spaces) (d) with F replaced by W: two functions are equal when they agree at every point, so each axiom is checked one point at a time and reduces to the corresponding axiom in W. Inside it, $\mathcal{M}(V_1, \dots, V_k; W)$ passes Theorem 1.2.1. The zero function is k-linear, since freezing slots leaves the zero map. If β, γ are k-linear and $c \in F$, then freezing all slots but the m-th turns $\beta + \gamma$ and $c\beta$ into a sum and a scalar multiple of linear maps $V_m \to W$, which are linear by Theorem 2.4.1. So $\mathcal{M}(V_1, \dots, V_k; W)$ is a subspace, hence a vector space.

Two special cases are old friends. For $k = 1$, $\mathcal{M}(V; W) = \mathcal{L}(V, W)$ by Example (Seven functions of several vectors) (a). For $W = F$, the elements are the multilinear forms of Definition 7.2.1; Chapter 7 wrote the space of alternating n-linear forms on F^n and proved it one-dimensional (Corollary 7.4.4), which is a subspace of $\mathcal{M}(F^n, \dots, F^n; F)$.

Theorem 15.1.2. Dimension of the Space of Multilinear Maps

Let $V_1, \dots, V_k$ and W be finite-dimensional, with $\dim V_i = n_i$ and $\dim W = m$. Then $\mathcal{M}(V_1, \dots, V_k; W)$ is finite-dimensional, of dimension

$$\dim \mathcal{M}(V_1, \dots, V_k; W) = (\dim W) \prod_{i=1}^{k} \dim V_i = m\, n_1 n_2 \cdots n_k. \tag{15.1.1}$$

Explicitly, fix bases $\mathcal{B}_i$ of V_i with dual bases $(\varphi_j^{(i)})$, and a basis $(w_1, \dots, w_m)$ of W. For a tuple $I = (i_1, \dots, i_k)$ and an index r, let

$$E_I^r(x_1, \dots, x_k) \coloneqq \varphi_{i_1}^{(1)}(x_1) \cdots \varphi_{i_k}^{(k)}(x_k)\, w_r.$$

The $m\, n_1 \cdots n_k$ maps E_I^r form a basis of $\mathcal{M}(V_1, \dots, V_k; W)$.

Idea. Theorem 15.1.1 says a multilinear map is the same thing as a list of vectors of W, one per index tuple. So count the lists: there are $n_1 \cdots n_k$ tuples and each entry needs m coordinates. The maps E_I^r are the maps whose list has a single w_r in position I and zeros elsewhere, which is why they should be the basis.

Proof. Each E_I^r is k-linear by the existence half of Theorem 15.1.1, being the map prescribed to take the value w_r at the basis tuple $(v_{i_1}^{(1)}, \dots, v_{i_k}^{(k)})$ and 0 at every other basis tuple.

Spanning. Let $\beta \in \mathcal{M}(V_1, \dots, V_k; W)$. For each tuple I write its value on the corresponding basis tuple in the basis of W:

$$\beta(v_{i_1}^{(1)}, \dots, v_{i_k}^{(k)}) = \sum_{r=1}^{m} c_{I,r}\, w_r,$$

which is possible, and with unique coefficients, since $(w_1, \dots, w_m)$ is a basis. The map $\sum_{I,r} c_{I,r} E_I^r$ is k-linear and takes exactly these values on the basis tuples. By the uniqueness half of Theorem 15.1.1 it equals β.

Independence. Suppose $\sum_{I,r} c_{I,r} E_I^r = 0$, the zero map. Evaluate at a basis tuple indexed by J. By the displayed formula for E_I^r, the coefficient $\varphi_{i_1}^{(1)}(v_{j_1}^{(1)}) \cdots \varphi_{i_k}^{(k)}(v_{j_k}^{(k)})$ is 1 for $I = J$ and 0 otherwise, so the value is $\sum_r c_{J,r} w_r = 0$. Since (w_r) is independent, $c_{J,r} = 0$ for every r. As J was arbitrary, every coefficient vanishes.

Hence the E_I^r form a basis. There are $n_1 n_2 \cdots n_k$ index tuples I and m choices of r, so the count is $m\, n_1 \cdots n_k$, which is Eq. 15.1.1. This proves the theorem.

■

Aftermath. Two sanity checks. With $k = 1$, Eq. 15.1.1 reads $\dim \mathcal{L}(V, W) = (\dim W)(\dim V)$, which is Theorem 4.1.4. With $W = F$ and $k = 2$, it reads $\dim \mathcal{M}(V_1, V_2; F) = n_1 n_2$, which is the count of $n_1 \times n_2$ matrices — and indeed Chapter 14 recorded a bilinear form on a single space by its Gram matrix (Definition 14.1.2). The n_i enter as a **product**, once per slot, and that is the number to remember.

Example. Counting bilinear maps

(a) Find $\dim \mathcal{M}(F^2, F^3; F^2)$.

(b) Find $\dim \mathcal{M}(F^3, F^3, F^3; F)$, and the dimension of the subspace of alternating forms inside it when $F = \mathbb{R}$.

(c) Write the bilinear map $\beta\colon F^2 \times F^2 \to F^2$, $\beta(\mathbf{x}, \mathbf{y}) = (x_1 y_1,\ x_1 y_2 + x_2 y_1)$, in the basis of Theorem 15.1.2 for the standard bases.

Solution.

(a) $\dim W \cdot \dim V_1 \cdot \dim V_2 = 2 \cdot 2 \cdot 3 = 12$.

(b) $1 \cdot 3 \cdot 3 \cdot 3 = 27$. The alternating ones form a subspace of dimension 1 by Corollary 7.4.4, spanned by the determinant. So alternation is a very strong condition: it cuts 27 down to 1.

(c) Here $m = n_1 = n_2 = 2$, so there are $2 \cdot 2 \cdot 2 = 8$ basis maps $E_{(i,j)}^r$, and $E_{(i,j)}^r(\mathbf{x}, \mathbf{y}) = x_i y_j \mathbf{e}_r$. Evaluate β on the four basis pairs:

$$\beta(\mathbf{e}_1, \mathbf{e}_1) = (1, 0) = \mathbf{e}_1, \quad \beta(\mathbf{e}_1, \mathbf{e}_2) = (0, 1) = \mathbf{e}_2,$$
$$\beta(\mathbf{e}_2, \mathbf{e}_1) = (0, 1) = \mathbf{e}_2, \quad \beta(\mathbf{e}_2, \mathbf{e}_2) = (0, 0) = \mathbf{0}.$$

By the spanning argument, $\beta = E_{(1,1)}^1 + E_{(1,2)}^2 + E_{(2,1)}^2$, with the other five coefficients zero. Reading the formula back, $\beta(\mathbf{x}, \mathbf{y}) = x_1 y_1 \mathbf{e}_1 + x_1 y_2 \mathbf{e}_2 + x_2 y_1 \mathbf{e}_2$, which is the given map.

Quick check. What does Eq. 15.1.1 give for $\mathcal{M}(V_1, V_2; W)$ when $V_2 = \{0\}$? Does the answer make sense?

Solution. It gives $(\dim W) \cdot (\dim V_1) \cdot 0 = 0$, so the space is $\{0\}$. That is right: a bilinear map must satisfy $\beta(v, 0) = 0$, and 0 is the only vector of V_2, so the only bilinear map is the zero map. The degenerate case is worth checking because it is where a formula with a product behaves differently from one with a sum.

15.1.5 The problem this chapter solves

We now have a good supply of multilinear maps and we know how many there are. What we do not have is any way to *work* with them. Consider what Chapters 2 to 5 give a linear map $T\colon V \to W$: a matrix, a kernel, an image, a rank, Rank–Nullity, a dual map, a determinant when $V = W$, eigenvalues, a Jordan form. A bilinear map $\beta\colon V \times W \to Z$ has none of these. It has no matrix, because Definition 4.1.1 applies to linear maps. It has no kernel that is a subspace: the set where β vanishes contains $(v, 0)$ and $(0, w)$ for all v, w but usually not their sum. It is not an element of $\mathcal{L}(V \times W, Z)$, as the previous subsection insisted. Every fact we know about it has had to be proved by hand, one slot at a time.

The wish is therefore easy to state. **We want to replace multilinearity by linearity.** Concretely, for two spaces V and W we want a single vector space — call it X for now — and a single bilinear map $\tau\colon V \times W \to X$, so universal that every bilinear map out of $V \times W$ is τ followed by a **linear** map. If that works, then bilinear maps $V \times W \to Z$ correspond to linear maps $X \to Z$, and the entire toolkit of Chapters 2 to 5 applies to the latter.

Before building anything, we can predict the size of X, and the prediction is a good reason to believe the object exists. Suppose such an X and τ existed, and suppose the correspondence "β on $V \times W \leftrightarrow$ linear map on X" is a bijection for every target Z. Take $Z = F$. Then Eq. 15.1.1 gives

$$\dim \mathcal{M}(V, W; F) = 1 \cdot \dim V \cdot \dim W,$$

while Theorem 4.1.4 gives $\dim \mathcal{L}(X, F) = \dim X$. A bijection that respects addition and scalars forces these to agree, so

$$\dim X = (\dim V)(\dim W).$$

The product of the dimensions, not the sum. This already rules out the naive guess $X = V \times W$, whose dimension is $\dim V + \dim W$ by Theorem 4.4.1, and it tells us what to aim for. The next section constructs X, calls it $V \otimes W$, and proves that its dimension is exactly this product.

One more remark before we go. The prediction used only the case $Z = F$, but the same count works for every Z: Eq. 15.1.1 gives $\dim Z \cdot \dim V \cdot \dim W$ on one side and $\dim Z \cdot \dim X$ on the other, and the two agree precisely when $\dim X = (\dim V)(\dim W)$. When the next section produces a basis of $V \otimes W$ indexed by pairs (i, j) of basis indices, that is the same count seen a second time, and the agreement is not a coincidence — it is the universal property, counted.

15.1.6 Exercises

15.1.6.1 A. Check your understanding

Exercise. A1

(a) Define what it means for $\beta\colon V_1 \times \cdots \times V_k \to W$ to be k-linear, with every quantifier.

(b) State the difference between "β is bilinear" and "β is linear on the product space $V \times W$", and give a function that is one but not the other.

(c) Write out $\beta(u+u', v+v')$ for a bilinear β. How many terms are there?

(d) State $\dim \mathcal{M}(V_1, \dots, V_k; W)$ in terms of the dimensions of the spaces involved.

(e) True or false: every bilinear map $\beta\colon V \times W \to Z$ satisfies $\beta(v, 0) = 0$. Justify your answer.

(f) What is a 1-linear map? What is a k-linear map with $W = F$ called?

Solution.

(a) It is k-linear if for **every** position i and **every** choice of fixed $v_j \in V_j$ with $j \neq i$, the function $x \mapsto \beta(v_1, \dots, x, \dots, v_k)$ with x in position i is a linear map $V_i \to W$ (Definition 15.1.1). All spaces are over the same field.

(b) Bilinear means linear in each slot with the other **frozen**; linear on the product means linear when both slots move **together**. The multiplication $\beta(a, b) = ab$ on $F \times F$ is bilinear but not linear on the product, since $\beta(2 \cdot (1,1)) = 4 \neq 2\beta(1,1)$. The sum $\sigma(a, b) = a + b$ is linear on the product but not bilinear, since $a \mapsto a + 1$ is not linear.

(c) $\beta(u, v) + \beta(u, v') + \beta(u', v) + \beta(u', v')$: four terms.

(d) $(\dim W) \prod_{i=1}^{k} \dim V_i$ (Eq. 15.1.1).

(e) True. Freeze v; the map $y \mapsto \beta(v, y)$ is linear, and a linear map sends 0 to 0 by Theorem 2.1.1.

(f) A 1-linear map is exactly a linear map. A k-linear map into F is a multilinear form (Definition 7.2.1).

15.1.6.2 B. Practice

Exercise. B1: Determine which are multilinear

Determine which of the following are multilinear maps. For those that are, say what k is; for those that are not, name the clause that fails and give a concrete witness. Justify your answer.

(a) $\beta\colon M_n(F) \times M_n(F) \to M_n(F)$, $\beta(\mathbf{A}, \mathbf{B}) = \mathbf{AB} - \mathbf{BA}$.

(b) $\beta\colon \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$, $\beta(\mathbf{x}, \mathbf{y}) = x_1 y_2 - x_2 y_1 + 1$.

(c) $\beta\colon F^n \times F^n \to M_n(F)$, $\beta(\mathbf{x}, \mathbf{y}) = \mathbf{x}\mathbf{y}^{\top}$.

(d) $\beta\colon M_n(F) \times M_n(F) \to F$, $\beta(\mathbf{A}, \mathbf{B}) = \det(\mathbf{AB})$, for $n \geq 2$.

(e) $\beta\colon V^* \times V^* \times V \times V \to F$, $\beta(\varphi, \psi, u, v) = \varphi(u)\psi(v)$.

Solution.

(a) Bilinear, $k = 2$. This is the commutator $[\mathbf{A}, \mathbf{B}]$ of Definition 8.5.1. Freeze $\mathbf{B}$: then $\mathbf{A} \mapsto \mathbf{AB} - \mathbf{BA}$ is a difference of two linear maps of $\mathbf{A}$, hence linear (Theorem 2.4.1). Freezing $\mathbf{A}$ is the same computation.

(b) Not multilinear. Freeze $\mathbf{y} = \mathbf{0}$; then $\mathbf{x} \mapsto 1$ is constant, so it sends $\mathbf{0}$ to 1 and the clause "is linear" fails at position 1. Removing the $+1$ leaves the signed area, which is bilinear.

(c) Bilinear, $k = 2$. The (i, j) entry of $\mathbf{x}\mathbf{y}^\top$ is $x_i y_j$. Freeze $\mathbf{y}$: each entry is a linear functional of $\mathbf{x}$ and the entries are the coordinates of the output in the basis of matrix units, so $\mathbf{x} \mapsto \mathbf{x}\mathbf{y}^\top$ is linear. Symmetrically in $\mathbf{y}$. This map will reappear as the model of the tensor product of two coordinate spaces.

(d) Not multilinear for $n \geq 2$. By Theorem 7.5.4, $\beta(\mathbf{A}, \mathbf{B}) = \det \mathbf{A} \det \mathbf{B}$. Freeze $\mathbf{B} = \mathbf{I}_n$; then $\mathbf{A} \mapsto \det \mathbf{A}$ is not linear, since $\det(c\mathbf{A}) = c^n \det \mathbf{A}$ and $n \geq 2$. Concretely with $n = 2$, $\det(2\mathbf{I}_2) = 4 \neq 2 \det \mathbf{I}_2 = 2$. The failing clause is again "is linear".

(e) 4-linear. Freeze three of the four arguments; in each case what remains is one of $\varphi \mapsto \varphi(u) \cdot \psi(v)$, $\psi \mapsto \psi(v) \cdot \varphi(u)$, $u \mapsto \varphi(u) \cdot \psi(v)$, $v \mapsto \psi(v) \cdot \varphi(u)$, each a fixed scalar times a linear functional, hence linear. Note that the four slots do **not** all lie in the same space, which Definition 15.1.1 allows and Definition 7.2.1 does not.

Exercise. B2: Dimension counts

Compute the dimension of each space. The field is F throughout.

(a) $\mathcal{M}(F^4, F^5; F^3)$.

(b) $\mathcal{M}(F^2, F^2, F^2; F^2)$.

(c) $\mathcal{M}(M_2(F), F^3; F)$.

(d) $\mathcal{M}(F^n; W)$ for $\dim W = m$, and the space $\mathcal{L}(F^n, W)$. Compare.

Solution.

(a) $3 \cdot 4 \cdot 5 = 60$ by Eq. 15.1.1.

(b) $2 \cdot 2 \cdot 2 \cdot 2 = 16$.

(c) $\dim M_2(F) = 4$, so $1 \cdot 4 \cdot 3 = 12$.

(d) Both are mn. With $k = 1$ the two spaces are equal, by Example (Seven functions of several vectors) (a), and Eq. 15.1.1 reduces to Theorem 4.1.4. The formula is consistent with what Chapter 2 already proved.

Exercise. B3: Expanding in the basis

Let $\beta\colon F^2 \times F^3 \to F^2$ be given by

$$\beta(\mathbf{x}, \mathbf{y}) = (x_1 y_1 - 2x_2 y_3,\ x_1 y_2 + x_2 y_1).$$

(a) Verify that β is bilinear.

(b) Compute the six values $\beta(\mathbf{e}_i, \mathbf{f}_j)$, where $(\mathbf{e}_1, \mathbf{e}_2)$ and $(\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3)$ are the standard bases.

(c) Write β in the basis $\{E^r_{(i,j)}\}$ of Theorem 15.1.2, and say how many of the 12 coefficients are nonzero.

Solution.

(a) Freeze $\mathbf{y}$. Each output coordinate is a fixed linear combination of x_1 and x_2, hence a linear functional of $\mathbf{x}$ (Theorem 5.1.1), and a map into F^2 whose coordinates are linear is linear. Freeze $\mathbf{x}$: each output coordinate is a fixed linear combination of y_1, y_2, y_3, and the same argument applies. So β is bilinear.

(b) Substituting the standard basis vectors:

$$\begin{aligned}\beta(\mathbf{e}_1,\mathbf{f}_1)&=(1,0), \quad \beta(\mathbf{e}_1,\mathbf{f}_2)=(0,1), \quad \beta(\mathbf{e}_1,\mathbf{f}_3)=(0,0),\\ \beta(\mathbf{e}_2,\mathbf{f}_1)&=(0,1), \quad \beta(\mathbf{e}_2,\mathbf{f}_2)=(0,0), \quad \beta(\mathbf{e}_2,\mathbf{f}_3)=(-2,0).\end{aligned}$$

(c) Reading the coordinates off, $c_{(1,1),1}=1$, $c_{(1,2),2}=1$, $c_{(2,1),2}=1$, and $c_{(2,3),1}=-2$, all others zero. Hence

$$\beta = E^1_{(1,1)} + E^2_{(1,2)} + E^2_{(2,1)} - 2\,E^1_{(2,3)},$$

so 4 of the $2\cdot 2\cdot 3 = 12$ coefficients are nonzero.

15.1.6.3 C. Going deeper

Exercise. C1: Bilinear forms on coordinate spaces are matrices

Let $m, n \geq 1$.

(a) Prove that for every bilinear $\beta\colon F^m \times F^n \to F$ there is a **unique** $\mathbf{A} \in M_{m\times n}(F)$ with $\beta(\mathbf{x},\mathbf{y}) = \mathbf{x}^\top \mathbf{A}\mathbf{y}$ for all $\mathbf{x},\mathbf{y}$.

(b) Hence deduce that $\dim \mathcal{M}(F^m, F^n; F) = mn$, and check that this agrees with Eq. 15.1.1.

Hint for (a): the entries of $\mathbf{A}$ can only be the values of β on pairs of standard basis vectors.

Solution.

(a) *Existence.* Put $a_{ij} := \beta(\mathbf{e}_i, \mathbf{f}_j)$, where $(\mathbf{e}_i)$ and $(\mathbf{f}_j)$ are the standard bases of F^m and F^n, and let $\mathbf{A} = (a_{ij})$. For $\mathbf{x} = \sum_i x_i \mathbf{e}_i$ and $\mathbf{y} = \sum_j y_j \mathbf{f}_j$, expanding one slot at a time by Definition 15.1.1 gives

$$\beta(\mathbf{x},\mathbf{y}) = \sum_{i,j} x_i y_j\, \beta(\mathbf{e}_i,\mathbf{f}_j) = \sum_{i,j} x_i a_{ij} y_j = \mathbf{x}^\top \mathbf{A}\mathbf{y}.$$

Uniqueness. If $\mathbf{x}^\top \mathbf{A}\mathbf{y} = \mathbf{x}^\top \mathbf{A}'\mathbf{y}$ for all $\mathbf{x},\mathbf{y}$, take $\mathbf{x} = \mathbf{e}_i$ and $\mathbf{y} = \mathbf{f}_j$: the left side is a_{ij} and the right side is a'_{ij}. Hence $\mathbf{A} = \mathbf{A}'$.

(b) By (a) the map $\mathcal{M}(F^m, F^n; F) \to M_{m\times n}(F)$, $\beta \mapsto \mathbf{A}$, is a bijection, and it is linear because the entries $\beta(\mathbf{e}_i,\mathbf{f}_j)$ depend linearly on β by the definition of the pointwise operations. So it is an isomorphism, and $\dim \mathcal{M}(F^m, F^n; F) = \dim M_{m\times n}(F) = mn$ by Theorem 2.5.4. Eq. 15.1.1 gives $1\cdot m\cdot n = mn$, the same value.

Exercise. C2: Linear and bilinear at once

Let V, W, Z be vector spaces over F and let $\beta\colon V \times W \to Z$ be a function.

(a) Prove that if β is **both** bilinear and linear as a map on the product space $V \times W$, then β is the zero map.

(b) Give an example, with $V = W = Z = F$, of a nonzero bilinear map and a nonzero linear map on the product, and say which property each one fails.

(c) Does (a) still hold when $V = \{0\}$? Explain what the statement says in that case.

Solution.

(a) Let $v \in V$ and $w \in W$. In the product space, $(v, w) = (v, 0) + (0, w)$. Since β is linear on the product,

$$\beta(v, w) = \beta(v, 0) + \beta(0, w).$$

Since β is bilinear, freezing the first slot at v makes $y \mapsto \beta(v, y)$ linear, so $\beta(v, 0) = 0$ by Theorem 2.1.1; freezing the second slot at w gives $\beta(0, w) = 0$ likewise. Therefore $\beta(v, w) = 0$ for all v, w, that is, $\beta = 0$.

(b) The product $\mu(a, b) = ab$ is bilinear and fails linearity on the product, since $\mu(2\cdot(1, 1)) = 4 \neq 2\mu(1, 1)$. The sum $\sigma(a, b) = a + b$ is linear on the product and fails bilinearity, since $\sigma(0, 1) = 1 \neq 0$ while a bilinear map must vanish when a slot is 0.

(c) Yes, but it says nothing: if $V = \{0\}$ then $v = 0$ always, and the only function $\{0\}\times W \to Z$ that is bilinear is already the zero map, since $\beta(0, w) = 0$. The content of (a) appears only when both V and W are nonzero and $Z \neq \{0\}$.

Exercise. C3: Counting over a finite field

Work over $\mathbb{F}_2$, the field with two elements.

(a) How many bilinear maps $\mathbb{F}_2^2 \times \mathbb{F}_2^2 \to \mathbb{F}_2$ are there? How many trilinear maps $\mathbb{F}_2^2 \times \mathbb{F}_2^2 \times \mathbb{F}_2^2 \to \mathbb{F}_2$?

(b) How many of the bilinear maps in (a) are symmetric, that is, satisfy $\beta(\mathbf{x}, \mathbf{y}) = \beta(\mathbf{y}, \mathbf{x})$ for all $\mathbf{x}, \mathbf{y}$? How many are alternating in the sense of Definition 7.2.2?

(c) Is every alternating bilinear form on $\mathbb{F}_2^2$ symmetric? Is every symmetric one alternating?

Solution.

(a) By Eq. 15.1.1, $\dim \mathcal{M}(\mathbb{F}_2^2, \mathbb{F}_2^2; \mathbb{F}_2) = 1 \cdot 2 \cdot 2 = 4$. A vector space of dimension d over $\mathbb{F}_2$ has exactly 2^d elements, since each of the d coordinates is 0 or 1. So there are $2^4 = 16$ bilinear maps. For three slots the dimension is $2^3 = 8$, giving $2^8 = 256$ trilinear maps.

(b) By Exercise (C1: Bilinear forms on coordinate spaces are matrices) a bilinear form on $\mathbb{F}_2^2$ is $\mathbf{x}^\top \mathbf{A}\mathbf{y}$ for a unique $\mathbf{A} \in M_2(\mathbb{F}_2)$, and β is symmetric exactly when $\mathbf{A} = \mathbf{A}^\top$ (compare the values at $(\mathbf{e}_i, \mathbf{e}_j)$ and $(\mathbf{e}_j, \mathbf{e}_i)$). A symmetric 2×2 matrix is determined by a_{11}, a_{12}, a_{22}, so there are $2^3 = 8$ symmetric ones. Alternating means $\beta(\mathbf{x}, \mathbf{x}) = 0$ for all $\mathbf{x}$, which forces $a_{11} = \beta(\mathbf{e}_1, \mathbf{e}_1) = 0$, $a_{22} = 0$, and then $\beta(\mathbf{e}_1+\mathbf{e}_2, \mathbf{e}_1+\mathbf{e}_2) = a_{12}+a_{21} = 0$. Over $\mathbb{F}_2$ that says $a_{21} = a_{12}$, so $\mathbf{A} = \left(\begin{smallmatrix} 0 & a \\ a & 0 \end{smallmatrix}\right)$ and there are 2 alternating forms: the zero form and the signed area.

(c) Both alternating forms found in (b) have symmetric matrices, so yes, every alternating form here is symmetric — which is Theorem 14.3.3 in characteristic 2, where $-1 = 1$.

The converse fails: the dot product $\beta(\mathbf{x},\mathbf{y}) = x_1y_1 + x_2y_2$ is symmetric, and $\beta(\mathbf{e}_1,\mathbf{e}_1) = 1 \neq 0$, so it is not alternating. Of the 8 symmetric forms only 2 are alternating.

15.2 The Tensor Product

The previous section ended with a wish: a space through which every bilinear map out of $V \times W$ becomes a **linear** map. This section grants it. The object is called the tensor product, and building it is the main technical work of the chapter — but the way we go about it matters more than the object itself, because the same three moves recur four more times, for the tensor algebra, the symmetric powers, the exterior powers and the Clifford algebra.

The three moves are: **state what the object must do**, in the form of a property of the maps out of it; **prove that any two objects doing it are isomorphic by a unique isomorphism**, using nothing but that property; and only then **construct one**, to know that the definition is not empty. The middle move is what licenses the word "the". Theorem 15.2.1 below is written out at full length and is the template for the rest of the chapter, so it is worth reading twice.

Throughout, F is a field and all vector spaces are over F. No finite-dimensionality is assumed until we reach bases, where it is stated.

15.2.1 What the object has to do

Here is the goal, stated before its terms are defined. We want a space X and a bilinear map $\tau\colon V \times W \to X$ such that *every bilinear map β out of $V \times W$ can be pushed through τ, and what comes out the other side is linear.* Two demands hide in that sentence. There must be **at least one** way to push β through, or the property is useless; and **at most one**, or we cannot recover β from its linear shadow. So we ask for existence and uniqueness together.

A tensor product is the most efficient bilinear map out of $V \times W$: every other one is it, followed by a single linear map.

Definition 15.2.1. Tensor Product

Let V and W be vector spaces over F. A **tensor product of V and W** is a pair (X, τ), where X is a vector space over F and $\tau\colon V \times W \to X$ is a bilinear map, satisfying:

(T1) (**universal property**) **For every** vector space Z over F and **every** bilinear map $\beta\colon V \times W \to Z$, there is a **unique** linear map $\bar{\beta}\colon X \to Z$ with $\bar{\beta} \circ \tau = \beta$.

In words, clause by clause. **For every Z:** the property is a demand about *all* targets at once, not about one favorite target. **Every bilinear β:** the arbitrary bilinear map is the input; nothing is assumed about it beyond Definition 15.1.1 with $k = 2$. **There is a unique linear $\bar{\beta}$:** exactly one, so both existence and uniqueness are being asserted. **With $\bar{\beta} \circ \tau = \beta$:** this is an equality of functions on $V \times W$, namely $\bar{\beta}(\tau(v,w)) = \beta(v,w)$ for all v, w. We say β **factors through τ**, and picture it as a triangle: the two routes from $V \times W$ to Z, one direct and one through X, give the same answer.

Notice what the definition does **not** say. It says nothing about the elements of X, nothing about a formula, and nothing about a basis: it is a statement purely about maps. That is deliberate, since a definition naming elements would depend on a construction, and there are several.

Warning. **The definition does not say that every element of X has the form $\tau(v, w)$.** It says that the values $\tau(v, w)$ are enough to *determine* a linear map on X — which, as we prove below, means they span X, not that they exhaust it. Sums of them are generally not of that form. This is the single most common misreading of the tensor product, and we return to it with a proof at the end of the section.

15.2.2 Any two tensor products are the same

Before building anything, we settle the word "the". If two pairs both satisfy (T1), each one's property produces a map to the other, and the composites have no choice but to be identities. This is the argument Chapter 14 used for the Clifford algebra (Proposition 14.12.1), and it is the argument we will use in four more sections.

Theorem 15.2.1. Uniqueness of the Tensor Product

Let V and W be vector spaces over F, and let (X, τ) and (X', τ') both be tensor products of V and W. Then there is a **unique** isomorphism $\varphi\colon X \to X'$ with $\varphi \circ \tau = \tau'$.

Idea. Three steps. ① Each pair's universal property, applied with the *other* pair's bilinear map as input, produces a linear map; so we get $\varphi\colon X \to X'$ and $\psi\colon X' \to X$ at once, for free. ② Compose them. The composite $\psi\varphi$ is a linear self-map of X that does nothing to τ — and so is the identity. ③ The uniqueness half of (T1), applied with $Z = X$ and $\beta = \tau$, says there is only **one** linear self-map doing nothing to τ. So the two coincide. The uniqueness half is the engine; without it the argument produces two maps and no reason for them to be inverse.

Proof. **Step 1.** The map $\tau'\colon V \times W \to X'$ is bilinear. Apply (T1) for (X, τ) with $Z = X'$ and $\beta = \tau'$: there is a unique linear $\varphi\colon X \to X'$ with $\varphi \circ \tau = \tau'$. Symmetrically, applying (T1) for (X', τ') with $Z = X$ and $\beta = \tau$ gives a unique linear $\psi\colon X' \to X$ with $\psi \circ \tau' = \tau$.

Step 2. The composite $\psi\varphi\colon X \to X$ is linear (Theorem 2.4.2 (a)), and

$$(\psi\varphi) \circ \tau = \psi \circ (\varphi \circ \tau) = \psi \circ \tau' = \tau.$$

The identity id_X is also linear and satisfies $\mathrm{id}_X \circ \tau = \tau$.

Step 3. Apply (T1) for (X, τ) with $Z = X$ and $\beta = \tau$, which is bilinear by hypothesis. It says there is **exactly one** linear map $X \to X$ whose composite with τ is τ. Both $\psi\varphi$ and id_X are such maps. Hence $\psi\varphi = \mathrm{id}_X$. Swapping the roles of (X, τ) and (X', τ') gives $\varphi\psi = \mathrm{id}_{X'}$.

So φ has a two-sided inverse and is a bijection, hence an isomorphism (Theorem 2.5.1). It is the unique isomorphism with $\varphi \circ \tau = \tau'$, because (T1) already produced it as the unique **linear** map with that property, and an isomorphism is in particular linear. This proves the theorem.

■

Aftermath, and the template. Read the proof again with the names removed. It says: *an object defined by a universal property is determined up to a unique isomorphism compatible with its structure map, because each object's property hands you a comparison map, and the uniqueness clause forces the composites to be identities.* Nothing in it used bilinearity, or vector spaces, or two factors. Sections 06, 07, 08, and 10 define the tensor algebra, the symmetric power, the exterior power, and the Clifford algebra by universal properties of exactly this shape, and each time the uniqueness proof is these three steps with different letters. When you meet them, do not reread the argument; recognize it.

Because of Theorem 15.2.1 we may now speak of **the** tensor product and fix notation.

Definition 15.2.2. Notation for Tensors

When a tensor product of V and W exists, we write it $V \otimes W$, write $v \otimes w$ for $\tau(v, w)$, and call an element of the form $v \otimes w$ a **simple tensor** (also called a *pure* or *decomposable* tensor). The defining property then reads: every bilinear $\beta\colon V \times W \to Z$ has a unique linear $\bar{\beta}\colon V \otimes W \to Z$ with $\bar{\beta}(v \otimes w) = \beta(v, w)$.

Bilinearity of τ, rewritten in this notation, gives the rules we compute with:

$$\begin{aligned} (v + v') \otimes w &= v \otimes w + v' \otimes w, \\ v \otimes (w + w') &= v \otimes w + v \otimes w', \\ (cv) \otimes w = c(v \otimes w) &= v \otimes (cw). \end{aligned}$$

The third line is the one to remember: **a scalar moves freely across the $\otimes$ sign.** That is the whole reason $v \otimes w$ does not determine v and w.

15.2.3 Building one: a free space modulo relations

Uniqueness is worthless if nothing satisfies (T1), so we construct a tensor product. The strategy is the one named in the chapter introduction: *take a space in which the symbols $v \otimes w$ are completely unrelated, then divide out exactly the relations bilinearity demands.* Both halves need a tool.

The first tool is a space with a basis indexed by an arbitrary set. It is the space of finitely supported functions, which Chapter 1 met for $S = \mathbb{N}$ (Exercise (B1: Finitely Supported Sequences)).

Definition 15.2.3. Free Vector Space on a Set

Let S be a set and F a field. The **free vector space on S over F** is

$$F^{(S)} := \{\, f\colon S \to F \;:\; f(s) = 0 \text{ for all but finitely many } s \,\},$$

a subset of the space F^S of all functions $S \to F$ (Example (The Standard Vector Spaces) (d)), with the pointwise operations. For $s \in S$, let $\delta_s \in F^{(S)}$ be the function with $\delta_s(s) = 1$ and $\delta_s(t) = 0$ for $t \neq s$.

Lemma 15.2.2. Universal Property of the Free Vector Space

Let S be a set and F a field.

(a) $F^{(S)}$ is a subspace of F^S, and $\{\delta_s : s \in S\}$ is a basis of it.

(b) For **every** vector space Z over F and **every** function $g\colon S \to Z$ — no structure on g is assumed, since S is only a set — there is **exactly one** linear map $\bar{g}\colon F^{(S)} \to Z$ with $\bar{g}(\delta_s) = g(s)$ for all $s \in S$.

Proof.

(a) We use Theorem 1.2.1. The zero function vanishes everywhere, so it lies in $F^{(S)}$. If $f, h \in F^{(S)}$ vanish outside the finite sets A and B, then $f + h$ vanishes outside $A \cup B$, which is finite, and cf vanishes outside A. So $F^{(S)}$ is a subspace.
Each δ_s lies in $F^{(S)}$. *Spanning:* if $f \in F^{(S)}$ is nonzero exactly on the finite set $\{s_1, \dots, s_r\}$, then f and $\sum_i f(s_i)\delta_{s_i}$ take the same value at every point of S, hence are equal. *Independence:* if $\sum_i a_i \delta_{s_i} = 0$ for distinct $s_1, \dots, s_r$, evaluate at s_j to get $a_j = 0$. By Definition 1.5.1, $\{\delta_s\}$ is a basis, possibly infinite in the sense of Theorem 1.8.2.

(b) Immediate from (a) and Theorem 2.1.6, which prescribes a linear map arbitrarily on a basis, in one and only one way.

■

In words: **a function on a set becomes a linear map on the free space, for free.** That is the point of the construction, and the whole of it: $F^{(S)}$ imposes no relations among the δ_s, so nothing can obstruct a prescription.

The second tool is the quotient. Chapter 2 recorded exactly the statement we need: a linear map that kills a subspace U passes, in exactly one way, to a linear map on V/U (Theorem 4.4.10 (a)). We use it in the form "to define a map out of a quotient, define it upstairs and check it kills the relations".

Theorem 15.2.3. Existence of the Tensor Product

For every pair of vector spaces V, W over F, a tensor product (X, τ) of V and W exists.

Idea. Write down the symbols $v \otimes w$ as a basis of a huge space P, one basis vector for each **pair** of vectors — so at this stage $v \otimes w$ has no properties at all. The four bilinearity rules are then four families of elements of P that we want to be zero, so divide by the subspace they span. A bilinear β becomes a linear map on P by Lemma 15.2.2, kills each relation because it is bilinear, and therefore passes to the quotient by Theorem 4.4.10. Uniqueness of the induced map is spanning: the cosets of the basis vectors span the quotient.

Proof. Let $S = V \times W$, regarded as a **set**, and let $P = F^{(S)}$ be the free vector space on it (Definition 15.2.3), with basis $\{\delta_{(v,w)}\}$. Let $R \subseteq P$ be the span of all elements of the four forms

$$\delta_{(v+v',\,w)} - \delta_{(v,w)} - \delta_{(v',w)}, \qquad \delta_{(v,\,w+w')} - \delta_{(v,w)} - \delta_{(v,w')},$$
$$\delta_{(cv,\,w)} - c\,\delta_{(v,w)}, \qquad \delta_{(v,\,cw)} - c\,\delta_{(v,w)},$$

taken over all $v, v' \in V$, all $w, w' \in W$, and all $c \in F$. A span is a subspace, so R is one. Put

$$X := P/R, \qquad \tau(v,w) := \pi(\delta_{(v,w)}),$$

where $\pi\colon P \to P/R$ is the quotient map, which is linear with kernel R (Theorem 4.4.4).

τ is bilinear. The element $\delta_{(v+v',w)} - \delta_{(v,w)} - \delta_{(v',w)}$ lies in $R = \ker\pi$, so applying π and using its linearity gives $\tau(v+v',w) - \tau(v,w) - \tau(v',w) = 0$. The other three families give, in the same way, additivity in the second slot and the two scalar rules. These are the clauses of Definition 15.1.1 for $k = 2$.

(T1), existence. Let Z be a vector space and $\beta\colon V \times W \to Z$ bilinear. Regarding β as a **function** on the set S, Lemma 15.2.2 (b) gives a unique linear $b\colon P \to Z$ with $b(\delta_{(v,w)}) = \beta(v,w)$. It kills every generator of R: for instance

$$b\big(\delta_{(v+v',w)} - \delta_{(v,w)} - \delta_{(v',w)}\big) = \beta(v+v',w) - \beta(v,w) - \beta(v',w) = 0$$

because β is linear in its first slot, and the other three families vanish by the other three clauses of bilinearity. Since b is linear and vanishes on a spanning set of R, it vanishes on R (Theorem 2.1.4); that is, $R \subseteq \ker b$. By Theorem 4.4.10 (a) there is exactly one linear $\bar\beta\colon X \to Z$ with $\bar\beta \circ \pi = b$, and then

$$\bar\beta(\tau(v,w)) = \bar\beta(\pi(\delta_{(v,w)})) = b(\delta_{(v,w)}) = \beta(v,w),$$

so $\bar\beta \circ \tau = \beta$.

(T1), uniqueness. Suppose $L\colon X \to Z$ is linear with $L \circ \tau = \beta$. Since π is surjective and the $\delta_{(v,w)}$ span P, the elements $\tau(v,w) = \pi(\delta_{(v,w)})$ span X. On each of them, L and $\bar\beta$ agree, both giving $\beta(v,w)$. Two linear maps that agree on a spanning set agree everywhere, by Theorem 2.1.4. Hence $L = \bar\beta$.

So (X, τ) satisfies (T1), which proves the theorem.

■

Aftermath. The construction is ugly and the object is not. That is normal for a universal property: the model is a scaffolding, and after this proof we never open it again. Every later fact about $V \otimes W$ in this book is derived from (T1), never from cosets of formal sums — which is also why the identification $F^m \otimes F^n \cong M_{m\times n}(F)$ below is legitimate as an alternative model, and much more pleasant to compute in.

One property is worth extracting at once, because we will use it constantly and because it is the second appearance of the self-map trick.

Proposition 15.2.4. Simple Tensors Span

Let $(V \otimes W, \tau)$ be a tensor product. Then $V \otimes W = \operatorname{span}\{\, v \otimes w : v \in V,\ w \in W \,\}$.

Proof. Let $Y = \operatorname{span}\{v \otimes w\}$, a subspace of $V \otimes W$, and let $j\colon Y \to V \otimes W$ be the inclusion, which is linear. Restricting the codomain of τ gives a bilinear map $\tau_Y\colon V \times W \to Y$, $(v,w) \mapsto v \otimes w$; it is bilinear because τ is and because Y is a subspace, so the identities hold already in Y. By (T1) with $Z = Y$ there is a linear $L\colon V \otimes W \to Y$ with $L \circ \tau = \tau_Y$. Then jL is a linear self-map of $V \otimes W$ with $jL\tau = j\tau_Y = \tau$, and so is id. By the uniqueness clause

of (T1) with $Z = V \otimes W$ and $\beta = \tau$, we get $jL = \mathrm{id}$. Hence j is surjective, so $Y = V \otimes W$.

■

Remark. That proof used no basis, no construction, and no finite dimension — only (T1). It is the same three lines as Step 1 of Proposition 14.12.3's proof in Chapter 14, where the generated subalgebra was shown to be everything. Once you see "a subobject that receives the structure map must be the whole thing", you can reuse it verbatim.

15.2.4 A basis and a dimension

Now we can count. The prediction at the end of the previous section was $\dim(V \otimes W) = (\dim V)(\dim W)$, and here it is, with a basis attached.

Theorem 15.2.5. Basis of a Tensor Product

Let V and W be finite-dimensional, with bases $(v_1, \dots, v_n)$ of V and $(w_1, \dots, w_m)$ of W. Then the nm simple tensors

$$v_i \otimes w_j, \qquad 1 \le i \le n, \quad 1 \le j \le m,$$

form a basis of $V \otimes W$. In particular $\dim(V \otimes W) = (\dim V)(\dim W)$.

Idea. Spanning is bilinearity: expand both slots of a simple tensor and use Proposition 15.2.4. Independence needs linear functionals on $V \otimes W$ that pick out one coefficient — and (T1) manufactures functionals out of bilinear forms, while Theorem 5.1.3 supplies all the bilinear forms we could want, namely $(x, y) \mapsto \varphi_p(x)\psi_q(y)$.

Proof. Spanning. By Proposition 15.2.4 it suffices to write each simple tensor in terms of the listed ones. Let $v = \sum_i a_i v_i$ and $w = \sum_j b_j w_j$. Expanding one slot at a time by the rules following Definition 15.2.2,

$$v \otimes w = \Big(\sum_i a_i v_i\Big) \otimes \Big(\sum_j b_j w_j\Big) = \sum_{i,j} a_i b_j \, (v_i \otimes w_j).$$

Independence. Let $(\varphi_1, \dots, \varphi_n)$ and $(\psi_1, \dots, \psi_m)$ be the bases of V^* and W^* dual to the given ones, so $\varphi_p(v_i) = \delta_{pi}$ and $\psi_q(w_j) = \delta_{qj}$ (Theorem 5.1.3). Fix p and q. The function

$$\beta_{pq} \colon V \times W \to F, \qquad \beta_{pq}(x, y) = \varphi_p(x)\, \psi_q(y),$$

is bilinear: freezing y leaves $\psi_q(y)\varphi_p$, a scalar multiple of a linear functional, and symmetrically. By (T1) there is a linear $f_{pq} \colon V \otimes W \to F$ with $f_{pq}(x \otimes y) = \varphi_p(x)\psi_q(y)$; in particular $f_{pq}(v_i \otimes w_j) = \delta_{pi}\delta_{qj}$.

Now suppose $\sum_{i,j} c_{ij}\,(v_i \otimes w_j) = 0$. Apply f_{pq}. By linearity the left side becomes $\sum_{i,j} c_{ij}\delta_{pi}\delta_{qj} = c_{pq}$, and the right side is 0. Hence $c_{pq} = 0$. Since p and q were arbitrary, all coefficients vanish.

So the nm listed tensors form a basis, and $\dim(V \otimes W) = nm$. This proves the theorem.

■

Aftermath: the count agrees. We found $\dim \mathcal{M}(V, W; Z) = (\dim Z)(\dim V)(\dim W)$ in Theorem 15.1.2, and we argued that a universal object X would have to satisfy $\dim \mathcal{L}(X, Z) = \dim \mathcal{M}(V, W; Z)$, forcing $\dim X = (\dim V)(\dim W)$. Theorem 15.2.5 delivers exactly that number, and the agreement is not a coincidence — it is the universal property, counted. Here is the statement behind it.

Corollary 15.2.6. Bilinear Maps Are Linear Maps on the Tensor Product

Let V, W, Z be vector spaces over F. The map

$$\Xi \colon \mathcal{L}(V \otimes W, Z) \to \mathcal{M}(V, W; Z), \qquad L \mapsto L \circ \tau,$$

is an isomorphism of vector spaces. When V, W, and Z are finite-dimensional, both sides have dimension $(\dim Z)(\dim V)(\dim W)$.

Proof. Ξ lands where claimed: $L \circ \tau$ is bilinear, since freezing one slot of τ gives a linear map and composing with the linear L keeps it linear (Theorem 2.4.2 (a)). It is linear in L by Theorem 2.4.2 (d), which gives $(L_1 + L_2) \circ \tau = L_1\tau + L_2\tau$ and $(cL) \circ \tau = c(L\tau)$. It is surjective by the existence half of (T1) and injective by the uniqueness half: if $L\tau = 0$, then both L and the zero map are linear maps composing with τ to give the zero bilinear map, so $L = 0$, and Theorem 2.2.3 applies. The dimension count is Theorem 15.1.2 on one side and Theorem 4.1.4 together with Theorem 15.2.5 on the other.

■

Quick check. What is $V \otimes W$ when $W = F$, viewed as a one-dimensional space over itself? What is it when $W = \{0\}$?

Solution. For $W = F$: $\dim(V \otimes F) = (\dim V) \cdot 1 = \dim V$ by Theorem 15.2.5, and indeed $v \mapsto v \otimes 1$ is an isomorphism $V \to V \otimes F$, since it carries a basis (v_i) to the basis $(v_i \otimes 1)$. For $W = \{0\}$: the dimension is 0, so $V \otimes \{0\} = \{0\}$. Both are the degenerate cases one should be able to predict from the product formula, and both come out right.

15.2.5 Coordinates: the tensor product of coordinate spaces

An abstract object is easier to trust once it has a concrete model. For coordinate spaces there is a very good one: $F^m \otimes F^n$ is the space of $m \times n$ matrices, with the simple tensor $\mathbf{x} \otimes \mathbf{y}$ corresponding to the outer product $\mathbf{x}\mathbf{y}^\top$.

Theorem 15.2.7. $F^m \otimes F^n \cong M_{m\times n}(F)$

Let $m, n \geq 1$. There is a unique linear map

$$\Theta\colon F^m \otimes F^n \to M_{m\times n}(F), \qquad \Theta(\mathbf{x}\otimes\mathbf{y}) = \mathbf{x}\mathbf{y}^{\top},$$

and it is an isomorphism, carrying $\mathbf{e}_i \otimes \mathbf{f}_j$ to the matrix unit $\mathbf{E}_{ij}$. Under Θ, the simple tensors correspond exactly to the matrices of rank at most 1.

Proof. The map $(\mathbf{x},\mathbf{y}) \mapsto \mathbf{x}\mathbf{y}^{\top}$ is bilinear: its (i,j) entry is $x_i y_j$, and with $\mathbf{y}$ frozen, $(\mathbf{x}+\mathbf{x}')\mathbf{y}^{\top} = \mathbf{x}\mathbf{y}^{\top} + \mathbf{x}'\mathbf{y}^{\top}$ and $(c\mathbf{x})\mathbf{y}^{\top} = c(\mathbf{x}\mathbf{y}^{\top})$ by Theorem 0.9.4 (2) and (4); freezing $\mathbf{x}$ is the same computation (Exercise (B1: Determine which are multilinear) (c)). By (T1) there is a unique linear Θ with $\Theta(\mathbf{x}\otimes\mathbf{y}) = \mathbf{x}\mathbf{y}^{\top}$. On standard basis vectors, $\mathbf{e}_i\mathbf{f}_j^{\top}$ has a single 1 in position (i,j), so $\Theta(\mathbf{e}_i\otimes\mathbf{f}_j) = \mathbf{E}_{ij}$.

By Theorem 15.2.5 the mn tensors $\mathbf{e}_i \otimes \mathbf{f}_j$ form a basis of $F^m \otimes F^n$, and the mn matrix units form a basis of $M_{m\times n}(F)$ (Example (Standard Bases) (c)). By Theorem 2.1.5 there is a linear $\Psi\colon M_{m\times n}(F) \to F^m \otimes F^n$ with $\Psi(\mathbf{E}_{ij}) = \mathbf{e}_i \otimes \mathbf{f}_j$. Then $\Psi\Theta$ and $\Theta\Psi$ are linear maps fixing a basis, hence are the identities by the uniqueness half of Theorem 2.1.5. So Θ is an isomorphism.

For the last claim, a nonzero $\mathbf{x}\mathbf{y}^{\top}$ is a product of an $m\times 1$ and a $1\times n$ matrix, so it has rank 1 by Proposition 4.3.2; and $\mathbf{x}\mathbf{y}^{\top} = \mathbf{O}$ when $\mathbf{x} = \mathbf{0}$ or $\mathbf{y} = \mathbf{0}$, of rank 0. Conversely, a matrix $\mathbf{A}$ of rank 1 factors as $\mathbf{A} = \mathbf{x}\mathbf{y}^{\top}$ with $\mathbf{x} \in F^m$ and $\mathbf{y}^{\top} \in M_{1\times n}(F)$, again by Proposition 4.3.2, and the zero matrix is $\mathbf{0}\otimes\mathbf{0}$. This proves the theorem. ∎

Example. Computing in $F^2 \otimes F^3$

Work in $F^2 \otimes F^3$ with standard bases $(\mathbf{e}_1,\mathbf{e}_2)$ and $(\mathbf{f}_1,\mathbf{f}_2,\mathbf{f}_3)$.

(a) Expand $(\mathbf{e}_1+2\mathbf{e}_2)\otimes(3\mathbf{f}_1-\mathbf{f}_2)$ in the basis of Theorem 15.2.5, and write the corresponding matrix.

(b) Decide whether $t = \mathbf{e}_1\otimes\mathbf{f}_1 + 2\,\mathbf{e}_1\otimes\mathbf{f}_2 + 3\,\mathbf{e}_2\otimes\mathbf{f}_1 + 6\,\mathbf{e}_2\otimes\mathbf{f}_2$ is a simple tensor, and if so write it as one.

(c) Decide whether $u = \mathbf{e}_1\otimes\mathbf{f}_1 + \mathbf{e}_2\otimes\mathbf{f}_2$ is a simple tensor.

Solution.

(a) Expanding one slot at a time,

$$(\mathbf{e}_1+2\mathbf{e}_2)\otimes(3\mathbf{f}_1-\mathbf{f}_2) = 3\,\mathbf{e}_1\otimes\mathbf{f}_1 - \mathbf{e}_1\otimes\mathbf{f}_2 + 6\,\mathbf{e}_2\otimes\mathbf{f}_1 - 2\,\mathbf{e}_2\otimes\mathbf{f}_2.$$

The coefficient of $\mathbf{e}_i\otimes\mathbf{f}_j$ is the (i,j) entry of the matrix, so Θ sends it to

$$\begin{pmatrix} 3 & -1 & 0 \\ 6 & -2 & 0 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}\begin{pmatrix} 3 & -1 & 0 \end{pmatrix},$$

which is the outer product predicted by Theorem 15.2.7.

(b) The coefficients give $\Theta(t) = \left(\begin{smallmatrix} 1 & 2 & 0 \\ 3 & 6 & 0 \end{smallmatrix}\right)$. Its second row is 3 times its first, so every 2×2 minor vanishes and the rank is 1 (Theorem 7.7.1). By Theorem 15.2.7, t is simple. Factoring, $\Theta(t) = (1,3)^{\top}(1,2,0)$, so

$$t = (\mathbf{e}_1 + 3\mathbf{e}_2) \otimes (\mathbf{f}_1 + 2\mathbf{f}_2),$$

which expands back to t.

(c) Here $\Theta(u) = \left(\begin{smallmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{smallmatrix}\right)$, whose first two columns are independent, so its rank is 2. By Theorem 15.2.7, u is **not** simple. No amount of rewriting will turn it into a single $\mathbf{v} \otimes \mathbf{w}$.

15.2.6 Three things a tensor is not

Part (c) above is not a curiosity; it is the central negative fact about tensor products, and the first of three misreadings worth naming before they happen.

15.2.6.1 Not every tensor is simple

Example. A tensor that is not simple

In $F^2 \otimes F^2$, the element $u = \mathbf{e}_1 \otimes \mathbf{e}_1 + \mathbf{e}_2 \otimes \mathbf{e}_2$ is not a simple tensor, for any field F.

Idea. Two routes, and both are worth having. The quick one: $\Theta(u) = \mathbf{I}_2$, of rank 2, while simple tensors have rank at most 1. The portable one: suppose $u = \mathbf{v} \otimes \mathbf{w}$, hit it with the four coordinate functionals f_{pq} built in the proof of Theorem 15.2.5, and read off four scalar equations that cannot all hold.

Proof. Suppose $u = \mathbf{v} \otimes \mathbf{w}$ for some $\mathbf{v}, \mathbf{w} \in F^2$. Let (φ_1, φ_2) be the basis of $(F^2)^*$ dual to $(\mathbf{e}_1, \mathbf{e}_2)$, and for $p, q \in \{1,2\}$ let $f_{pq} \colon F^2 \otimes F^2 \to F$ be the linear map with $f_{pq}(\mathbf{x} \otimes \mathbf{y}) = \varphi_p(\mathbf{x})\varphi_q(\mathbf{y})$, supplied by (T1) as in the proof of Theorem 15.2.5. Applying f_{pq} to both sides of $u = \mathbf{v} \otimes \mathbf{w}$ gives

$$v_p w_q = f_{pq}(u) = \delta_{p1}\delta_{q1} + \delta_{p2}\delta_{q2},$$

that is, $v_1 w_1 = 1$, $v_2 w_2 = 1$, $v_1 w_2 = 0$, and $v_2 w_1 = 0$. From $v_1 w_1 = 1$ we get $v_1 \neq 0$, so $v_1 w_2 = 0$ forces $w_2 = 0$ — and then $v_2 w_2 = 0 \neq 1$. This contradiction shows that no such $\mathbf{v}, \mathbf{w}$ exist, so u is not simple. ∎

Warning. **A general element of $V \otimes W$ is a *sum* of simple tensors, not a simple tensor.** Proposition 15.2.4 says the simple tensors span; Example (A tensor that is not simple) says they do not exhaust. A statement proved only for simple tensors is therefore proved for nothing until you extend it by linearity — which is legitimate precisely when the statement is about a linear map, and illegitimate when it is about, say, ranks or products. Whenever you want to define something on $V \otimes W$, do not write "define it on $v \otimes w$ and extend": use (T1), which is that move done correctly.

The smallest number of simple tensors needed to write a given tensor is a genuine invariant, and for coordinate spaces Theorem 15.2.7 identifies it: it is the rank of the corresponding matrix. So "not simple" means "rank at least 2". A later section meets the same failure for elements of $\Lambda^2 V$ that are not wedges of two vectors, and the closing section meets it again under the name *entanglement*.

15.2.6.2 A simple tensor vanishes only for an obvious reason

Proposition 15.2.8. When a Simple Tensor Is Zero

Let V and W be finite-dimensional, $v \in V$ and $w \in W$. Then $v \otimes w = 0$ if and only if $v = 0$ or $w = 0$.

Proof. ($\Leftarrow$) If $v = 0$, then $v \otimes w = (0 \cdot 0) \otimes w = 0(0 \otimes w) = 0$, using that a scalar passes across $\otimes$. The case $w = 0$ is the same with the slots swapped.

($\Rightarrow$) Suppose $v \neq 0$ and $w \neq 0$. By Theorem 5.1.5 (a) there are $\varphi \in V^*$ and $\psi \in W^*$ with $\varphi(v) = 1$ and $\psi(w) = 1$. The map $(x, y) \mapsto \varphi(x)\psi(y)$ is bilinear, so (T1) gives a linear $f\colon V \otimes W \to F$ with $f(x \otimes y) = \varphi(x)\psi(y)$. Then $f(v \otimes w) = 1 \neq 0$, while $f(0) = 0$. Hence $v \otimes w \neq 0$. This proves the proposition.

■

Warning. **A product of nonzero things can be zero in many algebraic settings, but not here.** It is tempting to expect $V \otimes W$ to behave like a ring with zero divisors, so that some $v \otimes w$ vanishes accidentally. It does not: Proposition 15.2.8 says the only way to get 0 is to feed in 0. The zero divisors appear one level later — in the algebras of the following sections, where products of nonzero *sums* of simple tensors can vanish.

15.2.6.3 A simple tensor does not remember its factors

Proposition 15.2.9. When Two Simple Tensors Agree

Let V and W be finite-dimensional, and let $v, v' \in V$ and $w, w' \in W$ with $v \otimes w \neq 0$. Then

$$v \otimes w = v' \otimes w'$$

if and only if there is a scalar $c \neq 0$ with $v' = cv$ and $w' = c^{-1}w$.

Idea. The "if" half is the scalar rule. For "only if", the danger is that v' points somewhere new. Suppose it does; then (v, v') is independent, so there is a functional killing v and not v', and the corresponding f on $V \otimes W$ separates the two sides.

Proof. ($\Leftarrow$) If $v' = cv$ and $w' = c^{-1}w$, then $v' \otimes w' = (cv) \otimes (c^{-1}w) = cc^{-1}(v \otimes w) = v \otimes w$, moving each scalar across $\otimes$.

($\Rightarrow$) Suppose $v \otimes w = v' \otimes w' \neq 0$. By Proposition 15.2.8 all four vectors are nonzero. Suppose, for a contradiction, that $v' \notin \operatorname{span}(v)$. Then (v, v') is linearly independent, so

it extends to a basis of V (Theorem 1.6.5), and the dual basis (Theorem 5.1.3) provides $\varphi \in V^*$ with $\varphi(v) = 0$ and $\varphi(v') = 1$. Since $w' \neq 0$, Theorem 5.1.5 (a) provides $\psi \in W^*$ with $\psi(w') = 1$. By (T1) there is a linear $f\colon V \otimes W \to F$ with $f(x \otimes y) = \varphi(x)\psi(y)$. Evaluating both sides of $v \otimes w = v' \otimes w'$,

$$0 = \varphi(v)\psi(w) = f(v \otimes w) = f(v' \otimes w') = \varphi(v')\psi(w') = 1,$$

which is absurd. Hence $v' = cv$ for some $c \in F$, and $c \neq 0$ because $v' \neq 0$.

Then $v \otimes w = (cv) \otimes w' = v \otimes (cw')$, so $v \otimes (w - cw') = 0$ by bilinearity. Since $v \neq 0$, Proposition 15.2.8 forces $w - cw' = 0$, that is, $w' = c^{-1}w$. This proves the proposition.

■

Warning. $v \otimes w = v' \otimes w'$ **does not mean** $v = v'$ **and** $w = w'$**.** Already $2v \otimes w = v \otimes 2w$, with different factors on the two sides. So "the first factor of a simple tensor" is **not** a well-defined function: writing $t = v \otimes w$ and then speaking of "the v of t" is an error. What Proposition 15.2.9 salvages is that the *lines* $\operatorname{span}(v)$ and $\operatorname{span}(w)$ are determined, which is exactly the information carried by a rank-one matrix: its column space and its row space.

Quick check. In $\mathbb{R}^2 \otimes \mathbb{R}^2$, is $\mathbf{e}_1 \otimes \mathbf{e}_2 + \mathbf{e}_2 \otimes \mathbf{e}_1$ a simple tensor?

Solution. No. Its matrix under Θ is $\left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$, of determinant $-1 \neq 0$, hence of rank 2 (Theorem 7.5.2). By Theorem 15.2.7 a simple tensor would have rank at most 1. Note how little the symmetry of the expression helps: being a sum of two simple tensors is the generic situation, not a special one.

15.2.7 Exercises

15.2.7.1 A. Check your understanding

Exercise. A1

(a) State the universal property (T1) of the tensor product, with every quantifier.

(b) What does Theorem 15.2.1 assert, and which clause of (T1) makes the composites identities?

(c) Give $\dim(V \otimes W)$ in terms of $\dim V$ and $\dim W$, and a basis.

(d) True or false: every element of $V \otimes W$ is of the form $v \otimes w$. Justify your answer.

(e) True or false: if $v \otimes w = 0$ then $v = 0$. Justify your answer.

(f) Explain in one sentence why the tensor product is **not** defined by a formula for its elements.

Solution.

(a) A pair (X, τ) with $\tau\colon V \times W \to X$ bilinear such that for **every** vector space Z and **every** bilinear $\beta\colon V \times W \to Z$ there is a **unique** linear $\bar{\beta}\colon X \to Z$ with $\bar{\beta} \circ \tau = \beta$ (Definition 15.2.1).

(b) That any two tensor products of V and W are isomorphic by a unique isomorphism commuting with the two structure maps. The **uniqueness** clause: it says there is only one linear self-map $X \to X$ composing with τ to give τ, and since both $\psi\varphi$ and id_X are such maps, they are equal.

(c) $(\dim V)(\dim W)$, with basis $(v_i \otimes w_j)$ for bases (v_i) of V and (w_j) of W (Theorem 15.2.5).

(d) False. In $F^2 \otimes F^2$, $\mathbf{e}_1 \otimes \mathbf{e}_1 + \mathbf{e}_2 \otimes \mathbf{e}_2$ is not simple (Example (A tensor that is not simple)).

(e) False as stated: it says $v = 0$ **or** $w = 0$ (Proposition 15.2.8). For instance $v \otimes 0 = 0$ for every v.

(f) Because a formula would require choosing a construction, and the object is characterized by what maps out of it do, not by what its elements look like; different constructions give different-looking elements and the same space up to a unique isomorphism.

15.2.7.2 B. Practice

Exercise. B1: Simple or not

For each element of $F^3 \otimes F^3$ below, determine whether it is a simple tensor. If it is, write it as $\mathbf{v} \otimes \mathbf{w}$; if it is not, justify your answer. Use the standard basis $(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$ in both factors.

(a) $t_1 = \mathbf{e}_1 \otimes \mathbf{e}_2 - 2\,\mathbf{e}_1 \otimes \mathbf{e}_3 + 3\,\mathbf{e}_3 \otimes \mathbf{e}_2 - 6\,\mathbf{e}_3 \otimes \mathbf{e}_3$.

(b) $t_2 = \mathbf{e}_1 \otimes \mathbf{e}_1 + \mathbf{e}_2 \otimes \mathbf{e}_2 + \mathbf{e}_3 \otimes \mathbf{e}_3$.

(c) $t_3 = \mathbf{e}_1 \otimes \mathbf{e}_2 + \mathbf{e}_2 \otimes \mathbf{e}_1$.

(d) $t_4 = 0$.

Solution. Throughout we use Theorem 15.2.7: t is simple exactly when the matrix $\Theta(t)$, whose (i, j) entry is the coefficient of $\mathbf{e}_i \otimes \mathbf{e}_j$, has rank at most 1.

(a) $\Theta(t_1) = \begin{pmatrix} 0 & 1 & -2 \\ 0 & 0 & 0 \\ 0 & 3 & -6 \end{pmatrix}$. Row 3 is 3 times row 1 and row 2 is zero, so the rank is 1. Factoring, $\Theta(t_1) = (1, 0, 3)^{\top}(0, 1, -2)$, so

$$t_1 = (\mathbf{e}_1 + 3\mathbf{e}_3) \otimes (\mathbf{e}_2 - 2\mathbf{e}_3),$$

which expands to the four given terms.

(b) $\Theta(t_2) = \mathbf{I}_3$, of rank 3. Not simple. This is Example (A tensor that is not simple) one size larger.

(c) $\Theta(t_3)$ has 1 in positions $(1, 2)$ and $(2, 1)$ and zeros elsewhere. Its top-left 2×2 minor is $0 \cdot 0 - 1 \cdot 1 = -1 \neq 0$, so its rank is 2 (Theorem 7.7.1). Not simple.

(d) $\Theta(t_4) = \mathbf{O}$, of rank $0 \leq 1$. So t_4 is simple, and indeed $0 = \mathbf{0} \otimes \mathbf{0}$. The degenerate case is simple for a boring reason, and it is worth keeping in the statement of Theorem 15.2.7.

Exercise. B2: Inducing maps from bilinear maps

In each case, a bilinear map is given. Use (T1) to produce the induced linear map on the tensor product, and compute the requested value.

(a) $\beta\colon F^2 \times F^2 \to F$, $\beta(\mathbf{x},\mathbf{y}) = x_1y_1 + x_2y_2$. Compute $\bar{\beta}(\mathbf{e}_1 \otimes \mathbf{e}_1 + \mathbf{e}_2 \otimes \mathbf{e}_2)$.

(b) $\beta\colon F^n \times F^n \to M_n(F)$, $\beta(\mathbf{x},\mathbf{y}) = \mathbf{x}\mathbf{y}^{\top} - \mathbf{y}\mathbf{x}^{\top}$. Compute $\bar{\beta}(\mathbf{v} \otimes \mathbf{v})$ for any $\mathbf{v}$.

(c) $\beta\colon V^* \times V \to F$, $\beta(\varphi, v) = \varphi(v)$, for V finite-dimensional. Compute $\bar{\beta}(\varphi_1 \otimes v_1 + \varphi_2 \otimes v_2)$ for a basis (v_i) with dual basis (φ_i), assuming $\dim V \geq 2$.

Solution.

(a) β is the dot product, bilinear by Example (Six two-slot functions) (a). By (T1) there is a unique linear $\bar{\beta}$ with $\bar{\beta}(\mathbf{x} \otimes \mathbf{y}) = x_1y_1 + x_2y_2$. By linearity,

$$\bar{\beta}(\mathbf{e}_1 \otimes \mathbf{e}_1 + \mathbf{e}_2 \otimes \mathbf{e}_2) = \bar{\beta}(\mathbf{e}_1 \otimes \mathbf{e}_1) + \bar{\beta}(\mathbf{e}_2 \otimes \mathbf{e}_2) = 1 + 1 = 2.$$

(b) Each entry of $\mathbf{x}\mathbf{y}^{\top} - \mathbf{y}\mathbf{x}^{\top}$ is $x_iy_j - y_ix_j$, linear in each slot, so β is bilinear and (T1) applies. Then $\bar{\beta}(\mathbf{v} \otimes \mathbf{v}) = \mathbf{v}\mathbf{v}^{\top} - \mathbf{v}\mathbf{v}^{\top} = \mathbf{O}$ for every $\mathbf{v}$, so the induced map kills every "diagonal" simple tensor without being zero: $\bar{\beta}(\mathbf{e}_1 \otimes \mathbf{e}_2) = \mathbf{E}_{12} - \mathbf{E}_{21} \neq \mathbf{O}$ for $n \geq 2$.

(c) The evaluation pairing is bilinear by Example (Seven functions of several vectors) (f), so $\bar{\beta}\colon V^* \otimes V \to F$ exists with $\bar{\beta}(\varphi \otimes v) = \varphi(v)$. Hence

$$\bar{\beta}(\varphi_1 \otimes v_1 + \varphi_2 \otimes v_2) = \varphi_1(v_1) + \varphi_2(v_2) = 1 + 1 = 2.$$

This map is the contraction, and the answer 2 is a first glimpse of a later section: it is the trace of a rank-two projection.

Exercise. B3: Dimensions and bases

Let V, W, and U be finite-dimensional over F, with $\dim V = 3$, $\dim W = 4$, and $\dim U = 2$.

(a) Find $\dim(V \otimes W)$ and $\dim((V \otimes W) \otimes U)$.

(b) Find $\dim \mathcal{L}(V \otimes W, U)$ and $\dim \mathcal{M}(V, W; U)$, and explain why they agree.

(c) Let $V = F^3$ with basis $(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$ and $W = F^4$ with basis $(\mathbf{f}_1, \dots, \mathbf{f}_4)$. Write down a basis of $V \otimes W$ and say which basis vector corresponds to the matrix unit $\mathbf{E}_{23}$ under Θ.

Solution.

(a) $\dim(V \otimes W) = 3 \cdot 4 = 12$ by Theorem 15.2.5, and then $\dim((V \otimes W) \otimes U) = 12 \cdot 2 = 24$, applying the theorem again to the pair $(V \otimes W, U)$.

(b) $\dim \mathcal{L}(V \otimes W, U) = (\dim U)\dim(V \otimes W) = 2 \cdot 12 = 24$ by Theorem 4.1.4, and $\dim \mathcal{M}(V, W; U) = 2 \cdot 3 \cdot 4 = 24$ by Eq. 15.1.1. They agree because Corollary 15.2.6 exhibits an isomorphism between the two spaces; the equality of numbers is that isomorphism, counted.

(c) The 12 tensors $\mathbf{e}_i \otimes \mathbf{f}_j$ with $1 \leq i \leq 3$ and $1 \leq j \leq 4$. Since $\Theta(\mathbf{e}_i \otimes \mathbf{f}_j) = \mathbf{E}_{ij}$, the basis vector corresponding to $\mathbf{E}_{23}$ is $\mathbf{e}_2 \otimes \mathbf{f}_3$.

15.2.7.3 C. Going deeper

Exercise. C1: The tensor product is commutative, up to a unique isomorphism

Let V and W be vector spaces over F.

(a) Prove that there is a unique linear map $\sigma\colon V \otimes W \to W \otimes V$ with $\sigma(v \otimes w) = w \otimes v$ for all v, w, and that it is an isomorphism.

(b) Deduce that $\sigma^2 = \mathrm{id}$ when $V = W$, where σ^2 means σ applied twice to $V \otimes V$.

(c) Explain why the argument in (a) never mentions a basis, and why that matters.

Hint for (a): the map $(v, w) \mapsto w \otimes v$ is bilinear; then build the inverse the same way and use the uniqueness clause of (T1).

Solution.

(a) The map $V \times W \to W \otimes V$, $(v, w) \mapsto w \otimes v$, is bilinear, because $\otimes$ is bilinear and swapping the slots exchanges the two clauses of Definition 15.1.1. By (T1) for $V \otimes W$ there is a unique linear σ with $\sigma(v \otimes w) = w \otimes v$. Symmetrically, (T1) for $W \otimes V$ gives a unique linear $\sigma'\colon W \otimes V \to V \otimes W$ with $\sigma'(w \otimes v) = v \otimes w$. Then $\sigma'\sigma$ is a linear self-map of $V \otimes W$ with $(\sigma'\sigma)(v \otimes w) = v \otimes w$, and so is the identity; by the uniqueness clause of (T1) applied with $Z = V \otimes W$ and $\beta = \tau$, the two agree, so $\sigma'\sigma = \mathrm{id}$. Symmetrically $\sigma\sigma' = \mathrm{id}$. Hence σ is an isomorphism (Theorem 2.5.1).

(b) With $V = W$, part (a) gives $\sigma' = \sigma$, since both are the unique linear map sending $v \otimes w$ to $w \otimes v$. Hence $\sigma^2 = \sigma'\sigma = \mathrm{id}$.

(c) Every step used only (T1) and the bilinearity of $\otimes$. So the argument works for infinite-dimensional spaces, where Theorem 15.2.5 is unavailable, and it produces a **canonical** isomorphism, whereas a basis-dependent construction would give one isomorphism per basis and no reason to prefer any of them.

Exercise. C2: Recovering the factors from a simple tensor

Let V and W be finite-dimensional and let $t = v \otimes w \neq 0$.

(a) Prove that the subspace $\mathrm{span}(v) \subseteq V$ depends only on t, not on the chosen factorization.

(b) Prove the corresponding statement for $\mathrm{span}(w) \subseteq W$.

(c) For $V = F^m$ and $W = F^n$, identify these two subspaces in terms of the matrix $\Theta(t)$ of Theorem 15.2.7, and say why (a) and (b) are then statements you already knew.

Solution.

(a) Suppose $t = v \otimes w = v' \otimes w'$ with $t \neq 0$. By Proposition 15.2.9 there is $c \neq 0$ with $v' = cv$. A nonzero scalar multiple spans the same subspace, so $\mathrm{span}(v') = \mathrm{span}(v)$. Hence the line is determined by t alone.

(b) The same proposition gives $w' = c^{-1}w$ with $c^{-1} \neq 0$, so $\mathrm{span}(w') = \mathrm{span}(w)$.

(c) Under Θ, t corresponds to $\mathbf{A} = \mathbf{v}\mathbf{w}^{\top}$, a matrix of rank 1 (Theorem 15.2.7). Every column of $\mathbf{A}$ is a scalar multiple of $\mathbf{v}$, and $\mathbf{A}$ is nonzero, so $\mathrm{col}(\mathbf{A}) = \mathrm{span}(\mathbf{v})$; every row is a multiple of $\mathbf{w}^{\top}$, so $\mathrm{row}(\mathbf{A}) = \mathrm{span}(\mathbf{w}^{\top})$. Thus (a) and (b) say that a matrix has

a well-defined column space and row space, which Chapter 2 established long ago. The tensor language adds nothing here except the reason the statement looks surprising: it is surprising only if one expects $\mathbf{v}$ itself, rather than its line, to be recoverable.

15.3 Tensor Products of Maps

Section 02 built the space $V \otimes W$. A construction on spaces is worth little until it acts on maps as well: the philosophy of the whole book has been objects first, then the maps between them. So given $S\colon V \to V'$ and $T\colon W \to W'$, we want a single map $V \otimes W \to V' \otimes W'$ built from the two. This section builds it, computes its matrix, and discovers that the matrix is the Kronecker product of Chapter 8 — which pays both of the debts that Chapter 8 §04 left open. It ends with the identification $V^* \otimes W \cong \mathcal{L}(V, W)$, which every later section of the chapter uses.

15.3.1 A map on each factor gives a map on the product

The definition writes itself, and that is the danger. We want a map that sends $v \otimes w$ to $Sv \otimes Tw$, so the temptation is to write

> "Define $(S \otimes T)(v \otimes w) \coloneqq Sv \otimes Tw$, and extend linearly."

Nothing in that sentence is a definition. Simple tensors span $V \otimes W$ (Proposition 15.2.4), but they do not exhaust it, and worse, the expression of an element as a combination of them is wildly nonunique: $v \otimes w = (2v) \otimes (\frac{1}{2}w)$, and $v \otimes w + v \otimes w' = v \otimes (w + w')$. A formula prescribed on simple tensors may therefore assign two different values to one element, and then it defines nothing at all. Here is a formula that really does fail. Try to define $g\colon V \otimes W \to V$ by $g(v \otimes w) \coloneqq v$. Taking $w = 0$ gives $v \otimes 0 = 0$ (Proposition 15.2.8), so g would have to send 0 to every v at once.

To build a map out of $V \otimes W$, never write a formula on simple tensors: write down a bilinear map and let the universal property produce the linear one.

That is the discipline of this chapter, and applied here it costs three lines.

Theorem 15.3.1. Tensor Product of Two Linear Maps

Let V, V', W, W' be vector spaces over F, and let $S \in \mathcal{L}(V, V')$ and $T \in \mathcal{L}(W, W')$. Then there is **exactly one** linear map

$$f\colon V \otimes W \to V' \otimes W'$$

with $f(v \otimes w) = Sv \otimes Tw$ for **all** $v \in V$ and $w \in W$.

Idea. The assignment $(v, w) \mapsto Sv \otimes Tw$ is bilinear, because S and T are linear and $\otimes$ is bilinear in its two slots. A bilinear map out of $V \times W$ is exactly what the universal property consumes, and what it returns is a unique linear map out of $V \otimes W$. There is nothing else to check.

Proof. Define $\beta\colon V \times W \to V' \otimes W'$ by $\beta(v,w) := Sv \otimes Tw$. Fix $w \in W$. The map $v \mapsto Sv \otimes Tw$ is the composite of S, which is linear, with the map $u \mapsto u \otimes Tw$, which is linear because the map $\otimes\colon V' \times W' \to V' \otimes W'$ of Definition 15.2.1 is bilinear, hence linear in its first slot with the second slot held fixed. A composite of linear maps is linear (Theorem 2.4.2), so β is linear in its first argument. Fixing v instead and swapping the roles of the two slots gives linearity in the second. Hence β is bilinear.

By the universal property in Definition 15.2.1 there is a unique linear $f\colon V \otimes W \to V' \otimes W'$ with $f(v \otimes w) = \beta(v,w) = Sv \otimes Tw$ for all v, w. This proves both existence and uniqueness.

■

Uniqueness is what licenses a name.

Definition 15.3.1. Tensor product of linear maps

Let $S \in \mathcal{L}(V, V')$ and $T \in \mathcal{L}(W, W')$. The **tensor product** of S and T is the unique linear map

$$S \otimes T\colon V \otimes W \to V' \otimes W'$$

of Theorem 15.3.1, characterized by

$$(S \otimes T)(v \otimes w) = Sv \otimes Tw \quad (v \in V,\ w \in W).$$

The symbol $\otimes$ now has three jobs — a space, a vector, a map — and they agree: $S \otimes T$ applied to $v \otimes w$ is $Sv \otimes Tw$, which is the only thing the notation could mean.

Examples.

1. **The identity.** $\mathrm{id}_V \otimes \mathrm{id}_W$ sends $v \otimes w$ to $v \otimes w$, and so does $\mathrm{id}_{V \otimes W}$. Both are linear and both are the factorization of the same bilinear map, so $\mathrm{id}_V \otimes \mathrm{id}_W = \mathrm{id}_{V \otimes W}$ by the uniqueness clause. Degenerate, and used in almost every proof below.
2. **Acting on one slot.** $S \otimes \mathrm{id}_W$ sends $v \otimes w$ to $Sv \otimes w$: it does S in the first slot and leaves the second alone. If $V = W = F[x]$ and $S = D$ is differentiation, $D \otimes \mathrm{id}$ is "differentiate the first factor".
3. **Scalars.** For $c, d \in F$, $(c\,\mathrm{id}_V) \otimes (d\,\mathrm{id}_W)$ sends $v \otimes w$ to $cv \otimes dw = cd\,(v \otimes w)$, so it is $cd\,\mathrm{id}_{V \otimes W}$. In particular a scalar may be moved from one slot to the other, which is the map-level shadow of $cv \otimes w = v \otimes cw$.
4. **Zero.** If $S = 0$, then $(S \otimes T)(v \otimes w) = 0 \otimes Tw = 0$, so $S \otimes T = 0$. The converse also holds; see Exercise (C1: When a tensor product of maps vanishes).

Nonexample by minimal change. Keep S and T linear, but try to send $v \otimes w$ to $Sv \otimes Tw + v \otimes w$. This one *is* legitimate: the assignment $(v,w) \mapsto Sv \otimes Tw + v \otimes w$ is a sum of two bilinear maps, hence bilinear, and the universal property applies. What is not legitimate is $(v,w) \mapsto Sv \otimes Tw + u_0$ for a fixed nonzero $u_0 \in V' \otimes W'$: the added constant destroys linearity in each slot, so no linear map on $V \otimes W$ does this, and indeed the prescription contradicts itself at $w = 0$. The exact clause that fails is bilinearity, and it fails for the same reason as in the map g above.

We will lean repeatedly on the following consequence of the uniqueness clause. It replaces the phrase "and extend linearly", which we are not allowed to use.

Lemma 15.3.2. Simple Tensors Determine a Map

Let Z be a vector space over F and let $f, g\colon V \otimes W \to Z$ be linear. If $f(v \otimes w) = g(v \otimes w)$ for all $v \in V$ and $w \in W$, then $f = g$.

Proof. Put $\beta(v, w) := f(v \otimes w)$. Since $\otimes$ is bilinear and f is linear, β is bilinear. By hypothesis $\beta(v, w) = g(v \otimes w)$ as well, so f and g are both linear maps factoring β through $\otimes$. The universal property of Definition 15.2.1 says there is only one such map. Hence $f = g$. ∎

No finite-dimensionality is needed, and no basis is chosen. With it, the algebra of $\otimes$ on maps is a formality.

Proposition 15.3.3. Algebra of Tensor Products of Maps

Let $S, S_1, S_2 \in \mathcal{L}(V, V')$, $T, T_1, T_2 \in \mathcal{L}(W, W')$, and $c \in F$. Let also $R \in \mathcal{L}(V', V'')$ and $U \in \mathcal{L}(W', W'')$. Then:

(a) $(S_1 + S_2) \otimes T = S_1 \otimes T + S_2 \otimes T$ and $S \otimes (T_1 + T_2) = S \otimes T_1 + S \otimes T_2$;
(b) $(cS) \otimes T = c(S \otimes T) = S \otimes (cT)$;
(c) $(R \otimes U)(S \otimes T) = RS \otimes UT$;
(d) $\mathrm{id}_V \otimes \mathrm{id}_W = \mathrm{id}_{V \otimes W}$.

Proof. In each part both sides are linear maps out of $V \otimes W$, so by Lemma 15.3.2 it suffices to compare their values at a simple tensor $v \otimes w$.

(a) The left side gives $(S_1 + S_2)v \otimes Tw = (S_1 v + S_2 v) \otimes Tw$, which equals $S_1 v \otimes Tw + S_2 v \otimes Tw$ because $\otimes$ is additive in its first slot. That is the value of the right side. The second identity is the same argument in the other slot.
(b) $(cS)v \otimes Tw = (cSv) \otimes Tw = c(Sv \otimes Tw) = Sv \otimes (cTw)$, where both outer equalities use bilinearity of $\otimes$.
(c) $(R \otimes U)\big((S \otimes T)(v \otimes w)\big) = (R \otimes U)(Sv \otimes Tw) = RSv \otimes UTw$, which is the value of $RS \otimes UT$.
(d) Done in Example 1.

∎

Part (c) is the statement that $\otimes$ respects composition: it is *functoriality*, and we will see in a moment that it is the mixed product rule of Theorem 8.4.1 in disguise.

Two more structural facts are needed before §04 can speak of tensors with several slots. The first extends the universal property from two factors to any number, and the second says that the brackets do not matter.

For $k \geq 2$ define the iterated tensor product by

$$V_1 \otimes \cdots \otimes V_k := (V_1 \otimes \cdots \otimes V_{k-1}) \otimes V_k,$$

starting from V_1 itself when $k = 1$, and write $v_1 \otimes \cdots \otimes v_k$ for $(v_1 \otimes \cdots \otimes v_{k-1}) \otimes v_k$. Multilinear maps of k arguments were defined in Definition 15.1.1.

Lemma 15.3.4. Universal Property for k Factors

Let $V_1, \dots, V_k$ and Z be vector spaces over F, and let $\beta\colon V_1 \times \cdots \times V_k \to Z$ be multilinear. Then there is **exactly one** linear map

$$\bar{\beta}\colon V_1 \otimes \cdots \otimes V_k \to Z$$

with $\bar{\beta}(v_1 \otimes \cdots \otimes v_k) = \beta(v_1, \dots, v_k)$ for all $v_1, \dots, v_k$. Moreover the tensors $v_1 \otimes \cdots \otimes v_k$ span $V_1 \otimes \cdots \otimes V_k$.

Idea. Induction on k, freezing the last argument. With v_k held fixed, β is multilinear in the first $k-1$ arguments, so the inductive hypothesis converts it into a linear map on $X = V_1 \otimes \cdots \otimes V_{k-1}$. Letting v_k move again gives a bilinear map on $X \times V_k$, and (T1) finishes. The spanning claim is the same induction.

Proof. For $k = 1$ take $\bar{\beta} = \beta$, which is linear, and the claim is trivial; for $k = 2$ the statement is Definition 15.2.1 together with Proposition 15.2.4. Let $k \geq 3$ and assume the lemma for $k-1$. Put $X \coloneqq V_1 \otimes \cdots \otimes V_{k-1}$.

Fix $v_k \in V_k$. The map $(v_1, \dots, v_{k-1}) \mapsto \beta(v_1, \dots, v_k)$ is multilinear, so the inductive hypothesis gives a unique linear $g_{v_k}\colon X \to Z$ with $g_{v_k}(v_1 \otimes \cdots \otimes v_{k-1}) = \beta(v_1, \dots, v_k)$.

The map $X \times V_k \to Z$, $(x, v_k) \mapsto g_{v_k}(x)$, is bilinear. It is linear in x because each g_{v_k} is. For linearity in v_k, fix $v_k, v_k' \in V_k$ and $c \in F$: the linear maps $g_{v_k + c v_k'}$ and $g_{v_k} + c\, g_{v_k'}$ agree at every $v_1 \otimes \cdots \otimes v_{k-1}$, because β is linear in its last argument, and those tensors span X by the inductive hypothesis; so the two maps are equal.

By Definition 15.2.1 there is a unique linear $\bar{\beta}$ on $X \otimes V_k = V_1 \otimes \cdots \otimes V_k$ with $\bar{\beta}(x \otimes v_k) = g_{v_k}(x)$; taking $x = v_1 \otimes \cdots \otimes v_{k-1}$ gives $\bar{\beta}(v_1 \otimes \cdots \otimes v_k) = \beta(v_1, \dots, v_k)$.

Finally, $X \otimes V_k$ is spanned by the simple tensors $x \otimes v_k$ (Proposition 15.2.4), each x is a combination of $v_1 \otimes \cdots \otimes v_{k-1}$ by the inductive hypothesis, and $\otimes$ is linear in its first slot; so the k-fold tensors span. A linear map is determined by its values on a spanning set, which gives uniqueness of $\bar{\beta}$.

■

Proposition 15.3.5. Associativity of the Tensor Product

Let U, V, W be vector spaces over F. There is a unique isomorphism

$$\alpha\colon (U \otimes V) \otimes W \to U \otimes (V \otimes W)$$

with $\alpha\big((u \otimes v) \otimes w\big) = u \otimes (v \otimes w)$ for all u, v, w.

Proof. The map $(u, v, w) \mapsto u \otimes (v \otimes w)$ is trilinear, since $\otimes$ is bilinear in each of its two uses. By Lemma 15.3.4 with $k = 3$ there is a unique linear α on $(U \otimes V) \otimes W$ with the stated values. Symmetrically, freezing the *first* argument instead and running the two-factor property twice produces a linear $\alpha' \colon U \otimes (V \otimes W) \to (U \otimes V) \otimes W$ with $\alpha'(u \otimes (v \otimes w)) = (u \otimes v) \otimes w$. The composite $\alpha'\alpha$ fixes every $(u \otimes v) \otimes w$, and those span by Lemma 15.3.4, so $\alpha'\alpha = \mathrm{id}$; symmetrically $\alpha\alpha' = \mathrm{id}$. Hence α is an isomorphism. ∎

From now on we drop the brackets and write $U \otimes V \otimes W$, with $u \otimes v \otimes w$ for the element that both ways of bracketing name, and likewise for any number of factors.

15.3.2 The matrix of a tensor product of maps

Now we compute. A matrix needs **ordered** bases, and $V \otimes W$ has a basis indexed by *pairs* (i, k), so the whole content of the computation is which order we put the pairs in.

Order the basis $v_i \otimes w_k$ by dictionary order on the pair (i, k): the index of the first factor changes slowly, the index of the second factor runs fastest.

Definition 15.3.2. The product basis, ordered

Let $\mathcal{B} = (v_1, \dots, v_n)$ be an ordered basis of V and $\mathcal{C} = (w_1, \dots, w_q)$ an ordered basis of W. Write $\mathcal{B} \otimes \mathcal{C}$ for the basis $(v_i \otimes w_k)$ of $V \otimes W$ given by Theorem 15.2.5, **ordered** so that $v_i \otimes w_k$ occupies position

$$(i-1)q + k \qquad (1 \le i \le n,\ 1 \le k \le q),$$

that is, in the order

$$v_1 \otimes w_1,\ \dots,\ v_1 \otimes w_q,\ v_2 \otimes w_1,\ \dots,\ v_n \otimes w_q.$$

Division with remainder writes every position $1, \dots, nq$ as $(i-1)q + k$ for exactly one pair, so this really is an ordering of the whole basis. It is the dictionary order on pairs, and it is the same rule that Chapter 8 used to label the rows and columns of a Kronecker product in the formula $(*)$ after Definition 8.4.1. That is not a coincidence, and the next theorem says why.

Theorem 15.3.6. The Kronecker Product Is the Matrix of a Tensor Product of Maps

Let V, V', W, W' be finite-dimensional vector spaces over F with ordered bases

$$\mathcal{B} = (v_1, \dots, v_n) \text{ of } V, \qquad \mathcal{B}' = (v'_1, \dots, v'_m) \text{ of } V',$$
$$\mathcal{C} = (w_1, \dots, w_q) \text{ of } W, \qquad \mathcal{C}' = (w'_1, \dots, w'_p) \text{ of } W'.$$

Let $S \in \mathcal{L}(V, V')$ and $T \in \mathcal{L}(W, W')$. Then, with the orderings of Definition 15.3.2 on both sides,

$$[S \otimes T]^{\mathcal{B}' \otimes \mathcal{C}'}_{\mathcal{B} \otimes \mathcal{C}} = [S]^{\mathcal{B}'}_{\mathcal{B}} \otimes [T]^{\mathcal{C}'}_{\mathcal{C}},$$

the Kronecker product of Definition 8.4.1.

Idea. A matrix is read off column by column, and column (j,l) of the left side holds the coordinates of $(S\otimes T)(v_j\otimes w_l)=Sv_j\otimes Tw_l$. Expanding Sv_j and Tw_l in their bases and multiplying out by bilinearity produces the coefficient $a_{ij}b_{kl}$ on $v'_i\otimes w'_k$ — a product of one entry of each matrix. The entry formula $(*)$ of Chapter 8 says the Kronecker product puts exactly that number in exactly that place, provided the pairs are read in dictionary order. So the theorem is the entry formula and bilinearity, nothing more.

Proof. Write $\mathbf{A}=(a_{ij})=[S]_{\mathcal{B}}^{\mathcal{B}'}\in M_{m\times n}(F)$ and $\mathbf{B}=(b_{kl})=[T]_{\mathcal{C}}^{\mathcal{C}'}\in M_{p\times q}(F)$, so that by Definition 4.1.1

$$Sv_j=\sum_{i=1}^{m}a_{ij}v'_i,\qquad Tw_l=\sum_{k=1}^{p}b_{kl}w'_k.$$

Fix j and l. Using Definition 15.3.1 and then bilinearity of $\otimes$ in each slot,

$$\begin{aligned}(S\otimes T)(v_j\otimes w_l)&=Sv_j\otimes Tw_l\\&=\Big(\sum_i a_{ij}v'_i\Big)\otimes\Big(\sum_k b_{kl}w'_k\Big)\\&=\sum_{i=1}^{m}\sum_{k=1}^{p}a_{ij}b_{kl}\,(v'_i\otimes w'_k).\end{aligned}$$

By Definition 15.3.2, $v_j\otimes w_l$ is the basis vector in position $(j-1)q+l$ of $\mathcal{B}\otimes\mathcal{C}$, and $v'_i\otimes w'_k$ is the basis vector in position $(i-1)p+k$ of $\mathcal{B}'\otimes\mathcal{C}'$. So the displayed expansion says that the entry of $[S\otimes T]_{\mathcal{B}\otimes\mathcal{C}}^{\mathcal{B}'\otimes\mathcal{C}'}$ in row $(i-1)p+k$ and column $(j-1)q+l$ is $a_{ij}b_{kl}$.

By the entry formula $(*)$ accompanying Definition 8.4.1, that is exactly the entry of $\mathbf{A}\otimes\mathbf{B}$ in the same position. Every row index of an $mp\times nq$ matrix is $(i-1)p+k$ for exactly one pair (i,k), and every column index is $(j-1)q+l$ for exactly one pair (j,l), so the two matrices agree in every position. This proves the theorem.

■

The ordering is the whole content, so it is worth being exact about it. Suppose we had ordered the product basis the other way, putting $v_i\otimes w_k$ in position $(k-1)n+i$, so that the *second* index changes slowly. Running the proof again, the entry in row $(k-1)m+i$ and column $(l-1)n+j$ is still $a_{ij}b_{kl}$, which by $(*)$ is the entry of $\mathbf{B}\otimes\mathbf{A}$ in that position. So the other ordering gives

$$[T]_{\mathcal{C}}^{\mathcal{C}'}\otimes[S]_{\mathcal{B}}^{\mathcal{B}'},$$

the Kronecker product with the factors **swapped** — not the transpose, and not the same matrix. For square factors the two answers are related exactly as Proposition 8.4.2 says, by conjugating with the permutation matrix that relabels dictionary order into its reverse — which is precisely the change-of-coordinates matrix between the two orderings of the product basis. In the rectangular setting of Theorem 15.3.6 the relation is not a conjugation at all: the row and column relabelings are different permutations, so $\mathbf{B}\otimes\mathbf{A}=\mathbf{P}_\sigma^{\top}(\mathbf{A}\otimes\mathbf{B})\mathbf{P}_\tau$ with two matrices, as Chapter 8 notes just after that proposition's proof.

Warning. **Do not reverse the two factors of the basis ordering.** The theorem needs the index of the **first** tensor factor to be the slow one, matching Chapter 8's dictionary order in $(*)$. With the fast and slow indices exchanged the matrix of $S\otimes T$ is $[T]\otimes[S]$. For

$\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$ and $\mathbf{B} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ these two differ in the $(1,2)$ entry, which is $a_{11}b_{12} = 1$ in the first and $b_{11}a_{12} = 0$ in the second (the nonexample after Definition 8.4.1).

Chapter 8's two promises, paid. At Definition 8.4.1, Chapter 8 wrote that the symbol $\otimes$ "anticipates the tensor product of Chapter 15, of which this is the matrix version". Theorem 15.3.6 is that statement, made precise and proved: the Kronecker product of two matrices is the matrix of the tensor product of the two maps they represent, taken in the product bases. At the end of the same section, Chapter 8 wrote that the Kronecker product is "the coordinate shadow of a basis-free construction: the tensor product of linear maps, in Chapter 15, whose matrix in suitable bases is exactly $\mathbf{A} \otimes \mathbf{B}$", and promised that "the dictionary order in $(*)$ will reappear there as an ordering of a basis $v_i \otimes w_k$". The suitable bases are $\mathcal{B} \otimes \mathcal{C}$ and $\mathcal{B}' \otimes \mathcal{C}'$; the dictionary order has reappeared, as Definition 15.3.2; and the shadow is cast by $S \otimes T$. The second promise gets a sharper payment at the end of this section, where the vec identity turns out to be this same theorem applied to a space of matrices.

Example. Computing with a tensor product of maps

Let $S\colon F^2 \to F^2$ and $T\colon F^3 \to F^2$ have matrices, in the standard bases,

$$\mathbf{A} = \begin{pmatrix} 2 & 1 \\ 0 & 3 \end{pmatrix}, \qquad \mathbf{B} = \begin{pmatrix} 1 & 0 & -1 \\ 2 & 1 & 0 \end{pmatrix}.$$

Compute $(S \otimes T)(\mathbf{e}_1 \otimes \mathbf{f}_2 + \mathbf{e}_2 \otimes \mathbf{f}_3)$ directly from Definition 15.3.1, and again from the matrix of Theorem 15.3.6. Here $(\mathbf{e}_1, \mathbf{e}_2)$ is the standard basis of F^2 and $(\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3)$ that of F^3.

Solution. **Directly.** Columns of $\mathbf{A}$ and $\mathbf{B}$ give $S\mathbf{e}_1 = 2\mathbf{e}_1$, $S\mathbf{e}_2 = \mathbf{e}_1 + 3\mathbf{e}_2$, $T\mathbf{f}_2 = \mathbf{e}_2$, and $T\mathbf{f}_3 = -\mathbf{e}_1$. Hence

$$\begin{aligned}(S \otimes T)(\mathbf{e}_1 \otimes \mathbf{f}_2 + \mathbf{e}_2 \otimes \mathbf{f}_3) &= 2\mathbf{e}_1 \otimes \mathbf{e}_2 + (\mathbf{e}_1 + 3\mathbf{e}_2) \otimes (-\mathbf{e}_1) \\ &= -\,\mathbf{e}_1 \otimes \mathbf{e}_1 + 2\,\mathbf{e}_1 \otimes \mathbf{e}_2 - 3\,\mathbf{e}_2 \otimes \mathbf{e}_1,\end{aligned}$$

using bilinearity of $\otimes$ in both slots.

By matrix. The domain basis, ordered as in Definition 15.3.2, is $(\mathbf{e}_1 \otimes \mathbf{f}_1, \mathbf{e}_1 \otimes \mathbf{f}_2, \mathbf{e}_1 \otimes \mathbf{f}_3, \mathbf{e}_2 \otimes \mathbf{f}_1, \mathbf{e}_2 \otimes \mathbf{f}_2, \mathbf{e}_2 \otimes \mathbf{f}_3)$, so the input has coordinate column $(0,1,0,0,0,1)$. By Theorem 15.3.6 the matrix of $S \otimes T$ is

$$\mathbf{A} \otimes \mathbf{B} = \begin{pmatrix} 2 & 0 & -2 & 1 & 0 & -1 \\ 4 & 2 & 0 & 2 & 1 & 0 \\ 0 & 0 & 0 & 3 & 0 & -3 \\ 0 & 0 & 0 & 6 & 3 & 0 \end{pmatrix},$$

and multiplying it by that column gives $(-1, 2, -3, 0)$. In the codomain basis $(\mathbf{e}_1 \otimes \mathbf{e}_1, \mathbf{e}_1 \otimes \mathbf{e}_2, \mathbf{e}_2 \otimes \mathbf{e}_1, \mathbf{e}_2 \otimes \mathbf{e}_2)$ this is the same element as before.

15.3.3 What Chapter 8 was computing

Every property of the Kronecker product proved by entry-chasing in Chapter 8 is now a property of maps, and several of them become one-liners. We record the three that the rest of the chapter

uses.

The mixed product rule comes first, because it is the one Chapter 8 called "the engine of everything that follows".

Corollary 15.3.7. The Mixed Product Rule Is Functoriality

Let $\mathbf{A} \in M_{m\times n}(F)$, $\mathbf{C} \in M_{n\times t}(F)$, $\mathbf{B} \in M_{p\times q}(F)$, and $\mathbf{D} \in M_{q\times u}(F)$. Then

$$(\mathbf{A}\otimes\mathbf{B})(\mathbf{C}\otimes\mathbf{D}) = (\mathbf{AC})\otimes(\mathbf{BD}).$$

Proof. Let S, R, T, U be the linear maps with matrices $\mathbf{A}, \mathbf{C}, \mathbf{B}, \mathbf{D}$ in the standard bases of the relevant spaces F^k. By Theorem 15.3.6, with all product bases ordered as in Definition 15.3.2, $\mathbf{A}\otimes\mathbf{B}$ and $\mathbf{C}\otimes\mathbf{D}$ are the matrices of $S\otimes T$ and $R\otimes U$. By Theorem 4.1.2 the product of the matrices is the matrix of the composite $(S\otimes T)(R\otimes U)$, which is $SR\otimes TU$ by Proposition 15.3.3 (c). Applying Theorem 15.3.6 once more, and Theorem 4.1.2 again to each factor, the matrix of $SR\otimes TU$ is $(\mathbf{AC})\otimes(\mathbf{BD})$.

■

This is Theorem 8.4.1 (c), which Chapter 8 proved by partitioning both matrices into conformable blocks and multiplying blockwise. The proof above says what the identity *means*: doing two things one after the other in each slot separately is the same as doing them one after the other in both slots at once. In the same way, the bilinearity of Theorem 8.4.1 (a) is parts (a) and (b) of Proposition 15.3.3, its associativity (b) is Proposition 15.3.5, and its inverse rule (e) is parts (c) and (d) together with invertibility of a composite.

Next, the numerical invariants.

Corollary 15.3.8. Rank, Determinant, and Trace of a Tensor Product of Maps

Let V, V', W, W' be finite-dimensional over F, $S \in \mathcal{L}(V, V')$, $T \in \mathcal{L}(W, W')$. Then

$$\operatorname{rank}(S\otimes T) = \operatorname{rank} S \cdot \operatorname{rank} T.$$

If moreover $V = V'$ and $W = W'$, with $n = \dim V \geq 1$ and $m = \dim W \geq 1$, then

$$\det(S\otimes T) = (\det S)^m (\det T)^n, \qquad \operatorname{tr}(S\otimes T) = \operatorname{tr} S \cdot \operatorname{tr} T.$$

Idea. For the rank, choose bases adapted to the two kernels: then $S\otimes T$ kills most product basis vectors and sends the remaining $\operatorname{rank} S\cdot\operatorname{rank} T$ of them to independent vectors. For the determinant, split $S\otimes T = (S\otimes \mathrm{id})(\mathrm{id}\otimes T)$. The factor $\mathrm{id}_V\otimes T$ preserves each of the n subspaces $v_i\otimes W$ and acts on each as T; the factor $S\otimes \mathrm{id}_W$ preserves each of the m subspaces $V\otimes w_k$ and acts as S. So each factor is block diagonal with repeated blocks, and the exponents count the *other* space's dimension, which is where the exponents come from. The trace we take from the matrix.

Proof. **Rank.** Let $r = \operatorname{rank} S$ and $s = \operatorname{rank} T$. Choose a basis $(v_1, \dots, v_n)$ of V whose last $n - r$ vectors form a basis of $\ker S$; this is possible by Theorem 1.8.2 applied to a basis of

$\ker S$, after reordering. Then $Sv_1, \dots, Sv_r$ span $\operatorname{im} S$ by Theorem 2.2.4, and being $r = \operatorname{rank} S$ spanning vectors of an r-dimensional space they form a basis of $\operatorname{im} S$ (Theorem 1.6.4). Choose a basis $(w_1, \dots, w_m)$ of W the same way, so that $Tw_1, \dots, Tw_s$ is a basis of $\operatorname{im} T$ and $Tw_l = 0$ for $l > s$.

By Theorem 2.2.4, $\operatorname{im}(S \otimes T)$ is spanned by the vectors $Sv_j \otimes Tw_l$, and those with $j > r$ or $l > s$ are 0. Extend $(Sv_1, \dots, Sv_r)$ to a basis of V' and $(Tw_1, \dots, Tw_s)$ to a basis of W' (Theorem 1.8.2). By Theorem 15.2.5 the products of these basis vectors form a basis of $V' \otimes W'$; in particular the rs vectors $Sv_j \otimes Tw_l$ with $j \le r$, $l \le s$ are linearly independent. They therefore form a basis of $\operatorname{im}(S \otimes T)$, and $\operatorname{rank}(S \otimes T) = rs$.

Determinant. By Proposition 15.3.3 (c) and (d),

$$S \otimes T = (S \otimes \mathrm{id}_W)(\mathrm{id}_V \otimes T),$$

so $\det(S \otimes T) = \det(S \otimes \mathrm{id}_W) \det(\mathrm{id}_V \otimes T)$ by Theorem 7.5.8. Fix bases $\mathcal{B}$ of V and $\mathcal{C}$ of W and use the ordering of Definition 15.3.2. For each i, the subspace $v_i \otimes W \coloneqq \operatorname{span}(v_i \otimes w_1, \dots, v_i \otimes w_m)$ is carried into itself by $\mathrm{id}_V \otimes T$, which acts on it as T does on W; and $V \otimes W$ is the direct sum of these n subspaces, because their bases together form the basis $\mathcal{B} \otimes \mathcal{C}$. Consecutive blocks of the ordering are exactly these subspaces, so

$$[\mathrm{id}_V \otimes T]^{\mathcal{B}\otimes\mathcal{C}}_{\mathcal{B}\otimes\mathcal{C}} = \underbrace{[T]^{\mathcal{C}}_{\mathcal{C}} \oplus \cdots \oplus [T]^{\mathcal{C}}_{\mathcal{C}}}_{n \text{ blocks}},$$

whose determinant is $(\det T)^n$ by Theorem 8.1.4 (d) and induction on the number of blocks. The same argument for $S \otimes \mathrm{id}_W$, using instead the m subspaces $V \otimes w_k$, gives $\det(S \otimes \mathrm{id}_W) = (\det S)^m$. (This time the relevant basis vectors are not consecutive in the ordering of Definition 15.3.2, so reorder the basis to make them so; matrices of one operator in two bases are similar, by Theorem 4.2.3, hence have the same determinant, by Corollary 7.5.7.) Multiplying the two gives the formula.

Trace. By Theorem 15.3.6 the matrix of $S \otimes T$ in $\mathcal{B} \otimes \mathcal{C}$ is $[S]^{\mathcal{B}}_{\mathcal{B}} \otimes [T]^{\mathcal{C}}_{\mathcal{C}}$, so by Definition 4.5.2 and Theorem 8.4.3 (b), $\operatorname{tr}(S \otimes T) = \operatorname{tr} S \cdot \operatorname{tr} T$.

■

These are parts (a), (c), and (b) of Theorem 8.4.3, transported to operators. Two remarks on what has changed. First, the exponents in the determinant formula are no longer something to memorize: $\det S$ appears to the power $\dim W$ because $S \otimes \mathrm{id}_W$ runs S once on each of the $\dim W$ copies of V inside $V \otimes W$. Second, the proof of the trace formula above is still a computation in a basis. There is a genuinely basis-free reason, and §04 gives it. The determinant, too, has a basis-free explanation, which has to wait for the top exterior power in §09.

Warning. **Most operators on $V \otimes W$ are not of the form $S \otimes T$.** The operators that are of this form are heavily constrained: by Corollary 15.3.8 their rank is a product rs with $r \le \dim V$ and $s \le \dim W$. Take $V = W = F^2$, so $\dim(V \otimes W) = 4$, and let P be any operator on $V \otimes W$ of rank 3, for instance the projection onto the span of the first three basis vectors of Definition 15.3.2. Then $3 = rs$ with $r, s \in \{0, 1, 2\}$ is impossible, so $P \ne S \otimes T$ for every S and T. The maps $S \otimes T$ do **span** $\mathcal{L}(V \otimes W)$, which is a different and weaker statement; this is the same distinction as between simple tensors and general tensors in §02.

Quick check. Let $S \in \mathcal{L}(V)$ and $T \in \mathcal{L}(W)$ with $\dim V = 3$ and $\dim W = 2$, and suppose $\det S = 2$ and $\det T = -1$. What are $\det(S \otimes T)$ and $\det(T \otimes S)$, where the second is an operator on $W \otimes V$?

Solution. By Corollary 15.3.8, $\det(S \otimes T) = (\det S)^{\dim W}(\det T)^{\dim V} = 2^2(-1)^3 = -4$. For $T \otimes S$ the roles swap, giving $(\det T)^{\dim V}(\det S)^{\dim W} = (-1)^3 2^2 = -4$ as well. The two agree, as they must: a determinant is a product of scalars, and the exponent always belongs to the *other* factor's dimension. The tempting wrong answer, $(\det S)^3(\det T)^2 = 8$, puts each map's own dimension in its exponent.

15.3.4 Maps as tensors

We come to the identification that the rest of the chapter runs on. A functional $\varphi \in V^*$ and a vector $w \in W$ together determine a linear map $V \to W$: measure with φ, then scale w by the answer. That recipe is bilinear in (φ, w), which is now a reflex: it must come from a linear map on $V^* \otimes W$.

Theorem 15.3.9. Tensors Are Maps

Let V and W be **finite-dimensional** vector spaces over F. There is a unique linear map

$$\Theta\colon V^* \otimes W \to \mathcal{L}(V, W), \qquad \Theta(\varphi \otimes w)(v) = \varphi(v)\, w,$$

and it is an isomorphism. Moreover, for $R \in \mathcal{L}(V_0, V)$ and $U \in \mathcal{L}(W, W_0)$ with V_0, W_0 finite-dimensional,

$$\Theta((R' \otimes U)(\xi)) = U\, \Theta(\xi)\, R \qquad (\xi \in V^* \otimes W),$$

where $R'\colon V^* \to V_0^*$ is the dual map of Definition 5.4.1 and the Θ on the left is the one for the pair (V_0, W_0).

Idea. Existence is the universal property again. For bijectivity, count: both sides have dimension $\dim V \cdot \dim W$, so it is enough to see that Θ carries a basis onto a basis. It carries $\varphi_j \otimes w_i$ to the map whose matrix is the matrix unit $\mathbf{E}_{ij}$, and the matrix units are a basis. The last formula is the statement that Θ turns the tensor product of maps into composition on the two sides; it is checked on simple tensors.

Proof. **Existence and uniqueness.** For fixed φ and w, the assignment $v \mapsto \varphi(v)w$ is linear, so it is an element $\varphi \cdot w$ of $\mathcal{L}(V, W)$. The map $(\varphi, w) \mapsto \varphi \cdot w$ is bilinear: for every v,

$$\begin{aligned} (\varphi + \psi)(v)w &= \varphi(v)w + \psi(v)w, \\ \varphi(v)(w + w') &= \varphi(v)w + \varphi(v)w', \\ (c\varphi)(v)w &= c\,\varphi(v)w = \varphi(v)(cw), \end{aligned}$$

and two maps agreeing at every v are equal. By Definition 15.2.1 there is a unique linear Θ with $\Theta(\varphi \otimes w) = \varphi \cdot w$.

Bijectivity. Let $\mathcal{B} = (v_1, \dots, v_n)$ be a basis of V with dual basis $\mathcal{B}^* = (\varphi_1, \dots, \varphi_n)$ (Definition 5.1.3), and $\mathcal{C} = (w_1, \dots, w_m)$ a basis of W. By Theorem 15.2.5 the products $\varphi_j \otimes w_i$ form a basis of $V^* \otimes W$, of size nm. Now $\Theta(\varphi_j \otimes w_i)$ sends v_k to $\varphi_j(v_k) w_i = \delta_{jk} w_i$, so it sends v_j to w_i and every other basis vector to 0. By Definition 4.1.1 its matrix is the matrix unit $\mathbf{E}_{ij}$. The matrix units form a basis of $M_{m\times n}(F)$ (Example (Standard Bases)) and $T \mapsto [T]_{\mathcal{B}}^{\mathcal{C}}$ is an isomorphism (Theorem 4.1.4), so the nm maps $\Theta(\varphi_j \otimes w_i)$ form a basis of $\mathcal{L}(V, W)$.

Hence $\operatorname{im}\Theta$ contains a basis of $\mathcal{L}(V, W)$, so Θ is surjective (Theorem 2.2.4). Since $\dim(V^* \otimes W) = nm = \dim \mathcal{L}(V, W)$, Theorem 2.3.1 gives $\operatorname{nullity}\Theta = 0$, so Θ is injective (Theorem 2.2.3). Therefore Θ is an isomorphism.

Naturality. Both $\xi \mapsto \Theta((R' \otimes U)\xi)$ and $\xi \mapsto U\,\Theta(\xi)\,R$ are linear in ξ, so by Lemma 15.3.2 it suffices to check them at $\xi = \varphi \otimes w$. On the left, $(R' \otimes U)(\varphi \otimes w) = R'\varphi \otimes Uw$, and for $v \in V_0$,

$$\Theta(R'\varphi \otimes Uw)(v) = (R'\varphi)(v)\,Uw = \varphi(Rv)\,Uw,$$

using Definition 5.4.1. On the right, $(U\,\Theta(\varphi \otimes w)\,R)(v) = U(\varphi(Rv)w) = \varphi(Rv)\,Uw$ by linearity of U. The two agree.

■

Warning. **Finite dimension is a hypothesis, not decoration.** The map Θ is defined and injective for any V and W, but it is surjective only onto the maps of finite rank: $\Theta(\sum_{t=1}^{N} \varphi_t \otimes w_t)$ has image inside $\operatorname{span}(w_1, \dots, w_N)$, so its rank is at most N. On $V = W = F[x]$, the identity map has infinite rank and is therefore not $\Theta(\xi)$ for any $\xi \in V^* \otimes V$. Everything below uses the finite-dimensional statement.

The letter Θ is deliberate: taking $V = F^n$ and $W = F^m$ and using the dual basis to identify $(F^n)^*$ with F^n, this Θ is the isomorphism $F^m \otimes F^n \cong M_{m\times n}(F)$ of Theorem 15.2.7, with the two factors written in the order that makes the source a space of *maps* rather than of matrices. Section 02's model was the special case; this is the general statement.

Two uses of the theorem, immediately. First, it names the matrix units: under Θ, the map with matrix $\mathbf{E}_{ij}$ *is* the tensor $\varphi_j \otimes w_i$, so a matrix is a sum $\sum_{i,j} a_{ij}\,\varphi_j \otimes w_i$. Note the index order — the functional carries the column index, the vector the row index — and compare it with Chapter 8's vectorization, which stacks columns.

Second, the naturality formula is the vec identity.

Example. The vec identity, read as a tensor product of maps

Let $\mathbf{A} \in M_{k\times m}(F)$ and $\mathbf{B} \in M_{n\times l}(F)$, and let

$$\Psi\colon M_{m\times n}(F) \to M_{k\times l}(F), \qquad \mathbf{X} \mapsto \mathbf{AXB},$$

which is linear. Identify $M_{m\times n}(F)$ with $\mathcal{L}(F^n, F^m)$ and $M_{k\times l}(F)$ with $\mathcal{L}(F^l, F^k)$. Explain why Theorem 8.4.4, $\operatorname{vec}(\mathbf{AXB}) = (\mathbf{B}^{\top} \otimes \mathbf{A})\operatorname{vec}\mathbf{X}$, is Theorem 15.3.6 applied to Ψ.

Solution. Let $R\colon F^l \to F^n$ and $U\colon F^m \to F^k$ be the maps with matrices $\mathbf{B}$ and $\mathbf{A}$ in the standard bases. Under the identification, Ψ becomes the map $\mathcal{L}(F^n, F^m) \to \mathcal{L}(F^l, F^k)$

sending T to UTR, so by the naturality clause of Theorem 15.3.9 it corresponds, through Θ, to

$$R' \otimes U \colon (F^n)^* \otimes F^m \to (F^l)^* \otimes F^k.$$

Order the basis of $(F^n)^* \otimes F^m$ as in Definition 15.3.2, using the dual standard basis $(\varphi_1, \dots, \varphi_n)$ first and the standard basis $(\mathbf{e}_1, \dots, \mathbf{e}_m)$ second. Then $\varphi_j \otimes \mathbf{e}_i$ sits in position $(j-1)m + i$, and $\Theta(\varphi_j \otimes \mathbf{e}_i) = \mathbf{E}_{ij}$. By Definition 8.4.2, $\operatorname{vec} \mathbf{E}_{ij} = \mathbf{e}_{(j-1)m+i}$. So vec **is** the coordinate map of this ordered basis: the dictionary order with the *column* index slow is exactly the order in which vectorization stacks the columns.

Now apply Theorem 15.3.6. The matrix of R' in the dual standard bases is $\mathbf{B}^\top$, because the matrix of a dual map is the transpose (Theorem 5.4.2), and the matrix of U is $\mathbf{A}$. Hence the matrix of $R' \otimes U$ in the product bases above is $\mathbf{B}^\top \otimes \mathbf{A}$. Reading that equality of matrices through the coordinate map vec gives $\operatorname{vec}(\mathbf{AXB}) = (\mathbf{B}^\top \otimes \mathbf{A}) \operatorname{vec} \mathbf{X}$, which is Theorem 8.4.4.

For a numerical check, take $\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix}$ and $\mathbf{B} = \begin{pmatrix} 3 & 0 \\ 1 & -1 \end{pmatrix}$. Writing the four images $\mathbf{A}\mathbf{E}_{ij}\mathbf{B}$ in the order $\mathbf{E}_{11}, \mathbf{E}_{21}, \mathbf{E}_{12}, \mathbf{E}_{22}$ and stacking the four columns $\operatorname{vec}(\mathbf{A}\mathbf{E}_{ij}\mathbf{B})$ side by side gives

$$\begin{pmatrix} 3 & 6 & 1 & 2 \\ 0 & 3 & 0 & 1 \\ 0 & 0 & -1 & -2 \\ 0 & 0 & 0 & -1 \end{pmatrix},$$

which is $\mathbf{B}^\top \otimes \mathbf{A}$ written out.

So the transpose in the vec identity, which Chapter 8 had to warn readers not to get wrong, is not a bookkeeping accident: it is there because Ψ multiplies $\mathbf{X}$ on the right by $\mathbf{B}$, and right multiplication is *precomposition*, which reverses arrows. The dual map is what reverses them, and the matrix of a dual map is a transpose. That is the sharper form of Chapter 8's second promise, and it is why the wrong guesses $\mathbf{A} \otimes \mathbf{B}^\top$ and $\mathbf{B} \otimes \mathbf{A}$ are wrong.

15.3.5 Exercises

15.3.5.1 A. Check your understanding

Exercise. A1

(a) State the defining property of $S \otimes T$, and say why writing "$(S \otimes T)(v \otimes w) = Sv \otimes Tw$, extended linearly" is not by itself a definition.

(b) Let $\dim V = 3$, $\dim W = 4$, $\dim V' = 2$, $\dim W' = 5$. What are the sizes of $[S]_{\mathcal{B}}^{\mathcal{B}'}$, $[T]_{\mathcal{C}}^{\mathcal{C}'}$, and $[S \otimes T]_{\mathcal{B}\otimes\mathcal{C}}^{\mathcal{B}'\otimes\mathcal{C}'}$?

(c) In the ordering of Definition 15.3.2 with $\dim V = 3$ and $\dim W = 4$, which basis vector of $V \otimes W$ sits in position 7?

(d) True or false: every operator on $V \otimes W$ is of the form $S \otimes T$. Justify your answer.

(e) State Theorem 15.3.9, including its hypothesis, and say what $\Theta(\varphi \otimes w)$ does to a vector v.

Solution.

(a) $S \otimes T$ is the **unique** linear map $V \otimes W \to V' \otimes W'$ with $(S \otimes T)(v \otimes w) = Sv \otimes Tw$ for all v, w (Definition 15.3.1). The quoted phrase is not a definition because a general element of $V \otimes W$ is a sum of simple tensors in many different ways, so a prescription on simple tensors may be inconsistent; $v \otimes 0 = 0$ already forces consistency conditions. What makes it work is that $(v, w) \mapsto Sv \otimes Tw$ is bilinear, so Definition 15.2.1 supplies the map.

(b) 2×3, 5×4, and 10×12 (Theorem 15.3.6: an $m \times n$ and a $p \times q$ matrix give an $mp \times nq$ one).

(c) With $q = 4$, position $7 = (i-1) \cdot 4 + k$ forces $i = 2$, $k = 3$. So it is $v_2 \otimes w_3$.

(d) False. By Corollary 15.3.8, $\operatorname{rank}(S \otimes T) = \operatorname{rank} S \cdot \operatorname{rank} T$. With $\dim V = \dim W = 2$, an operator of rank 3 on the 4-dimensional space $V \otimes W$ cannot have its rank factored as a product of two integers each at most 2.

(e) For V, W **finite-dimensional**, there is a unique linear $\Theta\colon V^* \otimes W \to \mathcal{L}(V, W)$ with $\Theta(\varphi \otimes w)(v) = \varphi(v)w$, and it is an isomorphism. So $\Theta(\varphi \otimes w)$ measures v with φ and scales w by the result.

15.3.5.2 B. Practice

Exercise. B1: A tensor product of maps, two ways

Let $S\colon \mathbb{Q}^2 \to \mathbb{Q}^2$ and $T\colon \mathbb{Q}^2 \to \mathbb{Q}^3$ have matrices, in the standard bases,

$$\mathbf{A} = \begin{pmatrix} 1 & -1 \\ 2 & 0 \end{pmatrix}, \qquad \mathbf{B} = \begin{pmatrix} 1 & 0 \\ 0 & 2 \\ -1 & 1 \end{pmatrix}.$$

(a) Write down $[S \otimes T]^{\mathcal{B}' \otimes \mathcal{C}'}_{\mathcal{B} \otimes \mathcal{C}}$ for the standard bases, ordered as in Definition 15.3.2.

(b) Compute $(S \otimes T)(\mathbf{e}_1 \otimes \mathbf{e}_2 - \mathbf{e}_2 \otimes \mathbf{e}_1)$ directly from Definition 15.3.1.

(c) Check (b) against (a) by multiplying the matrix by the coordinate column of the input.

Solution.

(a) By Theorem 15.3.6 the matrix is $\mathbf{A} \otimes \mathbf{B}$, of size 6×4:

$$\mathbf{A} \otimes \mathbf{B} = \begin{pmatrix} 1 & 0 & -1 & 0 \\ 0 & 2 & 0 & -2 \\ -1 & 1 & 1 & -1 \\ 2 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ -2 & 2 & 0 & 0 \end{pmatrix}.$$

(b) The columns of $\mathbf{A}$ give $S\mathbf{e}_1 = \mathbf{e}_1 + 2\mathbf{e}_2$ and $S\mathbf{e}_2 = -\mathbf{e}_1$; the columns of $\mathbf{B}$ give $T\mathbf{e}_1 = \mathbf{f}_1 - \mathbf{f}_3$ and $T\mathbf{e}_2 = 2\mathbf{f}_2 + \mathbf{f}_3$, where $(\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3)$ is the standard basis of $\mathbb{Q}^3$. Hence

$$\begin{aligned} &(S \otimes T)(\mathbf{e}_1 \otimes \mathbf{e}_2 - \mathbf{e}_2 \otimes \mathbf{e}_1) \\ &\quad = (\mathbf{e}_1 + 2\mathbf{e}_2) \otimes (2\mathbf{f}_2 + \mathbf{f}_3) - (-\mathbf{e}_1) \otimes (\mathbf{f}_1 - \mathbf{f}_3) \\ &\quad = \mathbf{e}_1 \otimes \mathbf{f}_1 + 2\,\mathbf{e}_1 \otimes \mathbf{f}_2 + 4\,\mathbf{e}_2 \otimes \mathbf{f}_2 + 2\,\mathbf{e}_2 \otimes \mathbf{f}_3, \end{aligned}$$

where the $\mathbf{e}_1 \otimes \mathbf{f}_3$ terms cancel: $+1$ from the first product and -1 from the second.

(c) The input has coordinates $(0, 1, -1, 0)$ in the basis $(\mathbf{e}_1 \otimes \mathbf{e}_1, \mathbf{e}_1 \otimes \mathbf{e}_2, \mathbf{e}_2 \otimes \mathbf{e}_1, \mathbf{e}_2 \otimes \mathbf{e}_2)$. Multiplying gives $(1, 2, 0, 0, 4, 2)$, which in the basis $(\mathbf{e}_1 \otimes \mathbf{f}_1, \mathbf{e}_1 \otimes \mathbf{f}_2, \mathbf{e}_1 \otimes \mathbf{f}_3, \mathbf{e}_2 \otimes \mathbf{f}_1, \mathbf{e}_2 \otimes \mathbf{f}_2, \mathbf{e}_2 \otimes \mathbf{f}_3)$ is exactly the answer to (b).

Exercise. B2: Invariants without multiplying matrices

Let $S \in \mathcal{L}(V)$ and $T \in \mathcal{L}(W)$ over $\mathbb{R}$, with $\dim V = 4$, $\dim W = 3$, $\operatorname{rank} S = 2$, $\operatorname{rank} T = 3$, $\operatorname{tr} S = 5$, $\operatorname{tr} T = -2$, $\det T = 6$.

(a) Find $\operatorname{rank}(S \otimes T)$ and $\operatorname{nullity}(S \otimes T)$.

(b) Find $\operatorname{tr}(S \otimes T)$.

(c) Find $\det(S \otimes T)$. Justify the value from the ranks alone.

(d) Find $\operatorname{tr}(S \otimes \mathrm{id}_W)$ and $\operatorname{tr}(\mathrm{id}_V \otimes T)$, and check that their product is **not** $\operatorname{tr}(S \otimes T)$ in general by comparing with (b).

Solution.

(a) By Corollary 15.3.8, $\operatorname{rank}(S \otimes T) = 2 \cdot 3 = 6$. Since $\dim(V \otimes W) = 12$ by Theorem 15.2.5, Theorem 2.3.1 gives $\operatorname{nullity}(S \otimes T) = 12 - 6 = 6$.

(b) $\operatorname{tr}(S \otimes T) = 5 \cdot (-2) = -10$.

(c) $\operatorname{rank} S = 2 < 4 = \dim V$, so S is not invertible and $\det S = 0$ by Theorem 3.4.4 and Definition 7.5.1. Hence $\det(S \otimes T) = 0^3 \cdot 6^4 = 0$. Directly: $\operatorname{rank}(S \otimes T) = 6 < 12$, so $S \otimes T$ is not invertible.

(d) $\operatorname{tr}(S \otimes \mathrm{id}_W) = \operatorname{tr} S \cdot \operatorname{tr} \mathrm{id}_W = 5 \cdot 3 = 15$ and $\operatorname{tr}(\mathrm{id}_V \otimes T) = 4 \cdot (-2) = -8$, using $\operatorname{tr} \mathrm{id} = (\dim) \cdot 1$ from Theorem 4.5.3. Their product is $-120 \neq -10$. There is no contradiction: $S \otimes T$ is the *composite* of the two, and the trace of a composite is not the product of the traces.

Exercise. B3: Which statements hold?

Let V, W be finite-dimensional over F. Determine which of the following hold for **all** $S, S_1, S_2 \in \mathcal{L}(V)$ and $T \in \mathcal{L}(W)$. Justify your answer.

(a) $(S_1 S_2) \otimes T = (S_1 \otimes T)(S_2 \otimes T)$.

(b) $(S_1 S_2) \otimes T^2 = (S_1 \otimes T)(S_2 \otimes T)$.

(c) $S \otimes T = 0$ implies $S = 0$ or $T = 0$.

(d) $S \otimes \mathrm{id}_W$ and $\mathrm{id}_V \otimes T$ commute.

(e) If S and T are invertible, so is $S \otimes T$, with inverse $S^{-1} \otimes T^{-1}$.

Solution.

(a) Fails. By Proposition 15.3.3 (c) the right side is $S_1 S_2 \otimes T^2$, and with $S_1 = S_2 = \mathrm{id}_V$ and $T = 2\,\mathrm{id}_W$ over $\mathbb{Q}$ the two sides are 2 id and 4 id on $V \otimes W$, by Example 3 after Definition 15.3.1.

(b) Holds: it is exactly Proposition 15.3.3 (c).

(c) Holds. If $S \neq 0$ and $T \neq 0$, then $\operatorname{rank} S \geq 1$ and $\operatorname{rank} T \geq 1$, so $\operatorname{rank}(S \otimes T) \geq 1$ by Corollary 15.3.8, and $S \otimes T \neq 0$.

(d) Holds. By Proposition 15.3.3 (c) and (d), both composites equal $S \otimes T$.

(e) Holds. By Proposition 15.3.3 (c) and (d), $(S \otimes T)(S^{-1} \otimes T^{-1}) = \mathrm{id}_V \otimes \mathrm{id}_W = \mathrm{id}_{V \otimes W}$, and likewise in the other order.

15.3.5.3 C. Going deeper

Exercise. C1: When a tensor product of maps vanishes

Let V, V', W, W' be finite-dimensional over F, $S \in \mathcal{L}(V, V')$ and $T \in \mathcal{L}(W, W')$.

(a) Prove that $S \otimes T = 0$ if and only if $S = 0$ or $T = 0$.

(b) Prove that $\ker(S \otimes \mathrm{id}_W) = \ker S \otimes W$, meaning the span of all $v \otimes w$ with $v \in \ker S$.

(c) Deduce $\operatorname{nullity}(S \otimes \mathrm{id}_W) = \operatorname{nullity} S \cdot \dim W$, and check it against Corollary 15.3.8.

Solution.

(a) ($\Leftarrow$) If $S = 0$, then $(S \otimes T)(v \otimes w) = 0 \otimes Tw = 0$ for all v, w, so $S \otimes T = 0$ by Lemma 15.3.2. Likewise if $T = 0$. ($\Rightarrow$) Suppose $S \neq 0$ and $T \neq 0$. Then $\operatorname{rank} S \geq 1$ and $\operatorname{rank} T \geq 1$, so $\operatorname{rank}(S \otimes T) = \operatorname{rank} S \cdot \operatorname{rank} T \geq 1$ by Corollary 15.3.8, and $S \otimes T \neq 0$.

(b) ($\supseteq$) For $v \in \ker S$ and $w \in W$, $(S \otimes \mathrm{id}_W)(v \otimes w) = 0 \otimes w = 0$; a kernel is a subspace (Theorem 2.2.1), so it contains the span of these elements. ($\subseteq$) Choose a basis $(v_1, \dots, v_n)$ of V whose last $n - r$ vectors span $\ker S$, where $r = \operatorname{rank} S$, and such that $Sv_1, \dots, Sv_r$ are independent (as in the proof of Corollary 15.3.8). Let $(w_1, \dots, w_q)$ be a basis of W and let $z = \sum_{j,l} c_{jl}\, v_j \otimes w_l$ lie in the kernel. Then

$$0 = (S \otimes \mathrm{id}_W)(z) = \sum_{j \leq r} \sum_{l} c_{jl}\, Sv_j \otimes w_l,$$

and the vectors $Sv_j \otimes w_l$ with $j \leq r$ are linearly independent by Theorem 15.2.5, after extending $(Sv_1, \dots, Sv_r)$ to a basis of V'. Hence $c_{jl} = 0$ for all $j \leq r$, so z is a combination of the $v_j \otimes w_l$ with $j > r$, all of which lie in $\ker S \otimes W$.

(c) By (b) and Theorem 15.2.5 applied to $\ker S$ and W, $\operatorname{nullity}(S \otimes \mathrm{id}_W) = \dim \ker S \cdot \dim W = \operatorname{nullity} S \cdot \dim W$. Consistency: Corollary 15.3.8 gives $\operatorname{rank}(S \otimes \mathrm{id}_W) = \operatorname{rank} S \cdot \dim W$, and adding the two gives $(\operatorname{rank} S + \operatorname{nullity} S) \dim W = \dim V \cdot \dim W = \dim(V \otimes W)$, as Theorem 2.3.1 requires.

Exercise. C2: Rank-one maps and simple tensors

Let V, W be finite-dimensional over F and let Θ be the isomorphism of Theorem 15.3.9.

(a) Prove that $\Theta(\xi)$ has rank at most 1 if and only if ξ is a simple tensor $\varphi \otimes w$.

(b) Deduce that "not every tensor is simple" (§02) says exactly that not every linear map has rank at most 1, and identify the tensor corresponding to id_V in a basis.

(c) Let $\dim V = \dim W = 2$. Exhibit $\xi \in V^* \otimes W$ that is not simple, by exhibiting a map of rank 2.

Solution.

(a) ($\Leftarrow$) $\Theta(\varphi \otimes w)$ sends every v into $\operatorname{span}(w)$, so its image has dimension at most 1. ($\Rightarrow$) Suppose $R := \Theta(\xi)$ has rank at most 1. If $R = 0$, then $\xi = 0 = 0 \otimes 0$ since Θ is injective. Otherwise $\operatorname{im} R = \operatorname{span}(w)$ for some $w \neq 0$, so for each v there is a unique scalar $\varphi(v)$ with $Rv = \varphi(v)w$; uniqueness and linearity of R make φ linear, so $\varphi \in V^*$. Then $R = \Theta(\varphi \otimes w)$, and $\xi = \varphi \otimes w$ because Θ is injective.

(b) By (a) and bijectivity of Θ, the nonsimple tensors in $V^* \otimes W$ correspond exactly to the maps of rank at least 2. Taking $W = V$ with basis $\mathcal{B} = (v_1, \dots, v_n)$ and dual basis $(\varphi_1, \dots, \varphi_n)$, the matrix of id_V is $\mathbf{I}_n = \sum_i \mathbf{E}_{ii}$, so $\mathrm{id}_V = \Theta(\sum_{i=1}^{n} \varphi_i \otimes v_i)$. For $n \geq 2$ this has rank $n \geq 2$, hence is not simple. For $n = 2$ it is the witness $\varphi_1 \otimes v_1 + \varphi_2 \otimes v_2$, which is Example (A tensor that is not simple) read through Θ.

(c) Take $\xi = \varphi_1 \otimes w_1 + \varphi_2 \otimes w_2$ for bases (v_1, v_2) of V and (w_1, w_2) of W. Then $\Theta(\xi)$ sends $v_1 \mapsto w_1$ and $v_2 \mapsto w_2$, so it has rank 2, and by (a) ξ is not simple.

Exercise. C3: Reproving the swap

Let V and W be finite-dimensional over F with ordered bases $\mathcal{B}$ and $\mathcal{C}$.

(a) Prove that there is a unique isomorphism $\tau\colon V \otimes W \to W \otimes V$ with $\tau(v \otimes w) = w \otimes v$.

(b) Prove that $\tau(S \otimes T) = (T \otimes S)\tau$ for all $S \in \mathcal{L}(V)$, $T \in \mathcal{L}(W)$.

(c) Deduce Proposition 8.4.2: for $\mathbf{A} \in M_n(F)$ and $\mathbf{B} \in M_m(F)$ there is a permutation matrix $\mathbf{P}$ with $\mathbf{B} \otimes \mathbf{A} = \mathbf{P}^{-1}(\mathbf{A} \otimes \mathbf{B})\mathbf{P}$, and identify $\mathbf{P}$.

Hint for (c): think of τ as renaming basis vectors rather than as moving them.

Solution.

(a) The map $(v, w) \mapsto w \otimes v$ is bilinear, so Definition 15.2.1 gives a unique linear τ with $\tau(v \otimes w) = w \otimes v$. The same construction with the roles swapped gives $\sigma\colon W \otimes V \to V \otimes W$ with $\sigma(w \otimes v) = v \otimes w$. Then $\sigma\tau$ fixes every simple tensor, so $\sigma\tau = \mathrm{id}$ by Lemma 15.3.2, and likewise $\tau\sigma = \mathrm{id}$. Hence τ is an isomorphism.

(b) Both sides are linear maps out of $V \otimes W$, so by Lemma 15.3.2 compare at $v \otimes w$: the left gives $\tau(Sv \otimes Tw) = Tw \otimes Sv$, and the right gives $(T \otimes S)(w \otimes v) = Tw \otimes Sv$.

(c) Let S, T be the operators on F^n, F^m with matrices $\mathbf{A}, \mathbf{B}$. Both $\mathcal{B} \otimes \mathcal{C}$ and $\tau^{-1}(\mathcal{C} \otimes \mathcal{B})$ are orderings of the same basis of $F^n \otimes F^m$, differing by the permutation σ that sends position $(k-1)n + i$ to position $(i-1)m + k$ — the direction matters, since σ is not an involution unless $n = m$ — and let $\mathbf{P}$ be its matrix, which is exactly the $\mathbf{P}_\sigma$ built in the proof of Proposition 8.4.2.

15.4 Contraction and the Trace

Chapter 2 defined the trace of an operator by choosing a basis, writing down the matrix, and adding the diagonal entries; it then proved that the answer does not depend on the choice (Definition 4.5.2). That is a correct definition and an unsatisfying one. It says how to compute the number, not what the number is, and the independence of the basis arrives afterwards, as a theorem about similar matrices. Section 03 offers a way out. An operator on V is an element of $V^* \otimes V$ (Theorem 15.3.9), and on $V^* \otimes V$ there is one obvious scalar-valued map that uses no basis at all: apply the functional to the vector. This section builds that map, proves that it is the trace, and extends it to tensors with many slots.

15.4.1 Evaluation, made linear

The dual space was invented so that its elements could eat vectors. The eating is a function of two arguments,

$$V^* \times V \to F, \qquad (\varphi, v) \mapsto \varphi(v),$$

and it is bilinear: linear in v because φ is a linear map, and linear in φ because that is how V^* is added and scaled (Definition 5.1.2). A bilinear map is exactly what the tensor product converts into a linear one, and by now this should be a reflex.

Contraction is the evaluation pairing, turned into a single linear map by the tensor product.

Definition 15.4.1. Contraction

Let V be a vector space over F. The **contraction** of V is the unique linear map

$$C\colon V^* \otimes V \to F, \qquad C(\varphi \otimes v) = \varphi(v) \quad (\varphi \in V^*,\ v \in V).$$

Well-definedness. The pairing $(\varphi, v) \mapsto \varphi(v)$ is bilinear, so Definition 15.2.1 supplies exactly one linear map on $V^* \otimes V$ taking those values. Nothing else is chosen: no basis, no finite dimension, no extra structure on V.

Examples.

1. **In coordinates.** Let $\dim V = n$, let $\mathcal{B} = (v_1, \dots, v_n)$ be a basis and $\mathcal{B}^* = (\varphi^1, \dots, \varphi^n)$ its dual basis (Definition 5.1.3), so $\varphi^j(v_i) = \delta_i^j$. A general $\xi \in V^* \otimes V$ is $\xi = \sum_{i,j} a_{ij}\, \varphi^j \otimes v_i$ by Theorem 15.2.5, and
$$C(\xi) = \sum_{i,j} a_{ij}\, \varphi^j(v_i) = \sum_{i=1}^{n} a_{ii}.$$
The contraction sums the coefficients whose two indices agree. That is a diagonal sum, which is the first hint of what is coming.
2. **The smallest case.** If $\dim V = 1$, pick $v \neq 0$ and let φ be dual to it. Every element of $V^* \otimes V$ is $a\,\varphi \otimes v$, and $C(a\,\varphi \otimes v) = a$. So C is the isomorphism $V^* \otimes V \cong F$. Degenerate, and worth seeing: the contraction is the only thing that can happen here.
3. **A functional times a vector, in F^n.** With $V = F^n$, $\varphi = \mathbf{y}^\top$, and $\mathbf{v} = \mathbf{x}$, $C(\varphi \otimes \mathbf{v}) = \mathbf{y}^\top\mathbf{x}$, the dot product. Chapter 11's inner product on $\mathbb{R}^n$ is this contraction composed with the identification of $\mathbb{R}^n$ with its dual — which is extra structure, and is why the dot product is not a contraction of $V \otimes V$.

Nonexample by minimal change. Replace $\varphi(v)$ by $\varphi(v)^2$. The assignment $(\varphi, v) \mapsto \varphi(v)^2$ is not bilinear — doubling v multiplies it by 4 — so no linear map on $V^* \otimes V$ agrees with it,

and Definition 15.2.1 does not apply. The exact clause that fails is linearity in each slot.

Why one upper slot and one lower slot. Contraction needs a functional *and* a vector: the functional supplies the mouth. There is no analogous map $V \otimes V \to F$.

Warning. **There is no canonical contraction of two vectors.** A linear map $V \otimes V \to F$ is the same thing as a bilinear form on V (Definition 15.2.1 with $Z = F$), and Chapter 14 spent a chapter on the fact that a vector space carries *many* bilinear forms and prefers none of them. Each basis of V supplies one, $(v, w) \mapsto \sum_i \varphi^i(v)\varphi^i(w)$, and different bases supply different ones: on $\mathbb{R}^2$ the standard basis gives the dot product, while $((1,0),(1,1))$ gives a form under which $(1,0)$ and $(0,1)$ are not orthogonal. Only the pairing between V^* and V is free of charge.

15.4.2 Contraction is the trace

We can now say what the trace is.

Theorem 15.4.1. The Trace Is a Contraction

Let V be a finite-dimensional vector space over F, and let

$$\Theta \colon V^* \otimes V \to \mathcal{L}(V)$$

be the isomorphism of Theorem 15.3.9. Then

$$\operatorname{tr}\left(\Theta(\xi)\right) = C(\xi) \qquad \text{for every } \xi \in V^* \otimes V,$$

that is, $\operatorname{tr} = C \circ \Theta^{-1}$ as maps $\mathcal{L}(V) \to F$.

Idea. Both sides are linear in ξ, so by Lemma 15.3.2 it is enough to check simple tensors. And $\Theta(\varphi \otimes v)$ is the rank-at-most-one map $u \mapsto \varphi(u)v$, whose trace can be read off from a single well-chosen basis: put v first.

Proof. Both $\xi \mapsto \operatorname{tr}(\Theta(\xi))$ and $\xi \mapsto C(\xi)$ are linear, being composites of linear maps (Theorem 4.5.3 for the first). By Lemma 15.3.2 it suffices to prove $\operatorname{tr}(\Theta(\varphi \otimes v)) = \varphi(v)$ for all $\varphi \in V^*$, $v \in V$.

Write $R := \Theta(\varphi \otimes v)$, so $Ru = \varphi(u)v$ for every u. If $v = 0$, then $R = 0$ and $\varphi(v) = 0$, so both sides vanish. Otherwise (v) is a linearly independent list, so by Theorem 1.8.2 it extends to a basis $\mathcal{B} = (u_1, \dots, u_n)$ of V with $u_1 = v$. For each j,

$$Ru_j = \varphi(u_j)\, v = \varphi(u_j)\, u_1,$$

so column j of $[R]_{\mathcal{B}}^{\mathcal{B}}$ has $\varphi(u_j)$ in row 1 and zeros below it (Definition 4.1.1). Hence the only nonzero diagonal entry is the one in position $(1,1)$, equal to $\varphi(u_1) = \varphi(v)$. By Definition 4.5.2, $\operatorname{tr} R = \varphi(v)$, as required.

■

What has changed. Chapter 2's definition named a basis and then had to prove that the answer survived a change of basis, using Theorem 4.2.5. Here nothing was named. The map C was manufactured by the universal property out of the evaluation pairing, which exists before any basis does, and Θ was manufactured the same way. So basis-independence is not a theorem about $C \circ \Theta^{-1}$; there is no basis in the construction to be independent of. Chapter 4's theorem is still true, and it is what makes the two descriptions agree — but it is now an observation about a *computation* of the trace, not part of its definition. The single basis used in the proof above was a convenience for reading off one number, not a choice the answer could depend on.

Two consequences come free. First, the trace of a rank-one map.

Corollary 15.4.2. Trace of a Rank-One Map

Let V be finite-dimensional, $\varphi \in V^*$ and $v \in V$, and let $R \in \mathcal{L}(V)$ be $Ru = \varphi(u)v$. Then $\operatorname{tr} R = \varphi(v)$. In particular, with a basis $\mathcal{B}$ and its dual basis, $\mathrm{id}_V = \Theta(\sum_i \varphi^i \otimes v_i)$ and $\operatorname{tr} \mathrm{id}_V = \sum_i \varphi^i(v_i) = n \cdot 1$.

Proof. The first claim was proved inside Theorem 15.4.1. For the second, $\Theta(\sum_i \varphi^i \otimes v_i)$ sends v_j to $\sum_i \varphi^i(v_j)v_i = v_j$, so it is id_V (Theorem 2.1.5); applying C gives $\sum_i \varphi^i(v_i) = \sum_i 1 = n \cdot 1$. ∎

Second, the promise made in §03: the trace of a tensor product of maps multiplies, and now we can see why. The key is that a functional on $V \otimes W$ is built from one on each factor.

Proposition 15.4.3. The Dual of a Tensor Product

Let V and W be finite-dimensional over F. There is a unique linear map

$$\kappa \colon V^* \otimes W^* \to (V \otimes W)^*, \qquad \kappa(\varphi \otimes \psi)(v \otimes w) = \varphi(v)\,\psi(w),$$

and it is an isomorphism.

Proof. For fixed φ, ψ, the map $(v, w) \mapsto \varphi(v)\psi(w)$ is bilinear, so Definition 15.2.1 produces a unique functional $\varphi \cdot \psi \in (V \otimes W)^*$ with the stated values. The assignment $(\varphi, \psi) \mapsto \varphi \cdot \psi$ is itself bilinear, since two functionals agreeing on all simple tensors are equal (Lemma 15.3.2), so Definition 15.2.1 gives the unique linear κ.

Let $\mathcal{B} = (v_1, \dots, v_n)$ and $\mathcal{C} = (w_1, \dots, w_m)$ be bases with dual bases (φ^i) and (ψ^j). Then $\kappa(\varphi^i \otimes \psi^j)(v_k \otimes w_l) = \delta^i_k \delta^j_l$, so κ carries the basis $(\varphi^i \otimes \psi^j)$ of $V^* \otimes W^*$ (Theorem 15.2.5) to the dual basis of $\mathcal{B} \otimes \mathcal{C}$ (Theorem 5.1.3 (a), which characterizes it by these values). A linear map carrying a basis to a basis is an isomorphism, by Theorem 2.2.4 and Theorem 2.3.1. ∎

Corollary 15.4.4. Why the Trace Multiplies

Let V, W be finite-dimensional over F, $S \in \mathcal{L}(V)$ and $T \in \mathcal{L}(W)$. Then $\operatorname{tr}(S \otimes T) = \operatorname{tr} S \cdot \operatorname{tr} T$.

Proof. Both $(S,T) \mapsto \operatorname{tr}(S \otimes T)$ and $(S,T) \mapsto \operatorname{tr} S \cdot \operatorname{tr} T$ are bilinear on $\mathcal{L}(V) \times \mathcal{L}(W)$, by Proposition 15.3.3 (a), (b), and Theorem 4.5.3. The maps of rank at most one span $\mathcal{L}(V)$, since Θ is surjective and simple tensors span $V^* \otimes V$ (Proposition 15.2.4); likewise for $\mathcal{L}(W)$. So it suffices to check the identity when $S = \Theta(\varphi \otimes v)$ and $T = \Theta(\psi \otimes w)$.

For such S and T and any $a \otimes b \in V \otimes W$,

$$(S \otimes T)(a \otimes b) = \varphi(a)v \otimes \psi(b)w = \kappa(\varphi \otimes \psi)(a \otimes b)\,(v \otimes w),$$

so $S \otimes T = \Theta\big(\kappa(\varphi \otimes \psi) \otimes (v \otimes w)\big)$ by Lemma 15.3.2. By Theorem 15.4.1 and Proposition 15.4.3,

$$\operatorname{tr}(S \otimes T) = \kappa(\varphi \otimes \psi)(v \otimes w) = \varphi(v)\psi(w),$$

which is $\operatorname{tr} S \cdot \operatorname{tr} T$ by Corollary 15.4.2.

■

So the trace multiplies because *evaluation* multiplies: a functional on a tensor product is a product of functionals, and contraction is evaluation. Chapter 8 proved this by summing $a_{ii}b_{kk}$ over all pairs (Theorem 8.4.3 (b)), which is the same fact with the sum written out.

15.4.3 Tensors with many slots

Contraction is not confined to one upper and one lower slot; it can be applied to a chosen pair inside a tensor with any number of them. First we name those tensors. Write

$$V^{\otimes p} := \underbrace{V \otimes \cdots \otimes V}_{p}, \qquad V^{\otimes 0} := F,$$

which is unambiguous by Proposition 15.3.5.

Definition 15.4.2. Tensor of type (p,q)

Let V be a vector space over F and $p, q \geq 0$. A **tensor of type** (p,q) on V is an element of

$$T^p_q(V) := V^{\otimes p} \otimes (V^*)^{\otimes q}.$$

The p copies of V are the **contravariant** slots and the q copies of V^* the **covariant** slots. The number $p+q$ is the **rank** or **order** of the tensor.

If $\dim V = n$, then $\dim T^p_q(V) = n^{p+q}$ by Theorem 15.2.5 and Corollary 5.1.4. The small cases are all things we have met:

- $T^0_0(V) = F$: scalars.
- $T^1_0(V) = V$ and $T^0_1(V) = V^*$: vectors and functionals.
- $T^1_1(V) = V \otimes V^* \cong \mathcal{L}(V)$: operators, by Theorem 15.3.9 together with the swap of Exercise (C3: Reproving the swap), which sends $v \otimes \varphi$ to $\varphi \otimes v$.
- $T^0_2(V) \cong (V \otimes V)^*$ by Proposition 15.4.3, and a functional on $V \otimes V$ is a bilinear form on V by Definition 15.2.1. So a type $(0,2)$ tensor is exactly a bilinear form — the whole subject matter of Chapter 14.

Warning. **The word "rank" is overloaded here.** The rank of a tensor in the sense of Definition 15.4.2 is the number of slots, $p+q$; it has nothing to do with $\operatorname{rank} T$ for a linear map, nor with the smallest number of simple tensors needed to write an element. A type $(1,1)$ tensor always has rank 2 in this sense, whatever the rank of the corresponding operator. Where confusion is possible we say **order**.

Now the general contraction. Contracting slot a against slot b means feeding the b-th functional the a-th vector and keeping everything else.

Definition 15.4.3. Contraction on a pair of slots

Let V be a vector space over F, let $p, q \geq 1$, and fix $1 \leq a \leq p$ and $1 \leq b \leq q$. The **contraction on slots a and b** is the unique linear map

$$C^a_b \colon T^p_q(V) \to T^{p-1}_{q-1}(V)$$

with, for all $v_1, \dots, v_p \in V$ and $\varphi^1, \dots, \varphi^q \in V^*$,

$$\begin{aligned} &C^a_b(v_1 \otimes \cdots \otimes v_p \otimes \varphi^1 \otimes \cdots \otimes \varphi^q) \\ &\qquad = \varphi^b(v_a)\; v_1 \otimes \dots \widehat{v_a} \cdots \otimes v_p \otimes \varphi^1 \otimes \dots \widehat{\varphi^b} \cdots \otimes \varphi^q, \end{aligned}$$

where a hat means that the factor is omitted.

Well-definedness. The right-hand side is multilinear in the $p+q$ arguments: each argument appears exactly once, either inside the scalar $\varphi^b(v_a)$ or as one factor of a tensor product, and both are linear in it. By Lemma 15.3.4 there is exactly one linear map on $T^p_q(V)$ with these values. For $p = q = 1$ it lands in $T^0_0(V) = F$ and sends $v \otimes \varphi$ to $\varphi(v)$: it is Definition 15.4.1 read on $V \otimes V^*$ instead of $V^* \otimes V$, that is, C composed with the swap of Exercise (C3: Reproving the swap). We use that identification silently from here on, and Theorem 15.4.1 applies to either reading.

Warning. **Which slots you contract matters.** Let $\dim V = 2$ with basis (v_1, v_2) and dual basis (φ^1, φ^2), and take the type $(2,2)$ tensor

$$t = v_1 \otimes v_2 \otimes \varphi^1 \otimes \varphi^2.$$

Then $C^1_1(t) = \varphi^1(v_1)\, v_2 \otimes \varphi^2 = v_2 \otimes \varphi^2$, while $C^1_2(t) = \varphi^2(v_1)\, v_2 \otimes \varphi^1 = 0$. The two contractions of the same tensor differ, one of them being zero. There is no "the" contraction of a tensor of order ≥ 3: the slots must be named.

Quick check. Let $\dim V = 3$ with basis (v_1, v_2, v_3) and dual basis $(\varphi^1, \varphi^2, \varphi^3)$, and let

$$t = v_1 \otimes \varphi^1 + 2\, v_2 \otimes \varphi^2 - v_1 \otimes \varphi^3 \in T^1_1(V).$$

Compute $C^1_1(t)$, and say which operator t corresponds to.

Solution. $C_1^1(t) = \varphi^1(v_1) + 2\varphi^2(v_2) - \varphi^3(v_1) = 1 + 2 - 0 = 3$. The corresponding operator sends $v_1 \mapsto v_1$, $v_2 \mapsto 2v_2$, $v_3 \mapsto -v_1$, so its matrix in (v_1, v_2, v_3) is $\begin{pmatrix} 1 & 0 & -1 \\ 0 & 2 & 0 \\ 0 & 0 & 0 \end{pmatrix}$, of trace 3, in agreement with Theorem 15.4.1.

Index notation, in the next section, makes contraction visibly what it looks like in physics: set an upper index equal to a lower one and sum.

15.4.4 Exercises

15.4.4.1 A. Check your understanding

Exercise. A1

(a) Define the contraction $C\colon V^* \otimes V \to F$, and say what makes it well defined.

(b) State the relation between the trace and the contraction, with its hypothesis.

(c) Explain why there is no canonical linear map $V \otimes V \to F$.

(d) For $\dim V = 4$, what is $\dim T_3^2(V)$?

(e) True or false: for $t \in T_2^2(V)$, the four contractions $C_b^a(t)$ all agree. Justify your answer.

Solution.

(a) C is the unique linear map with $C(\varphi \otimes v) = \varphi(v)$ (Definition 15.4.1). It is well defined because $(\varphi, v) \mapsto \varphi(v)$ is bilinear, so Definition 15.2.1 produces exactly one such linear map.

(b) If V is **finite-dimensional**, then $\operatorname{tr}(\Theta(\xi)) = C(\xi)$ for all $\xi \in V^* \otimes V$, where Θ is the isomorphism of Theorem 15.3.9 (Theorem 15.4.1).

(c) Such a map is the same thing as a bilinear form on V, and a vector space has many bilinear forms with none distinguished; each basis produces a different one. Contraction works only because V^* comes with the evaluation pairing built in.

(d) $4^{2+3} = 4^5 = 1024$.

(e) False. Definition 15.4.3's warning gives $t = v_1 \otimes v_2 \otimes \varphi^1 \otimes \varphi^2$ with $C_1^1(t) = v_2 \otimes \varphi^2 \neq 0$ and $C_2^1(t) = 0$.

15.4.4.2 B. Practice

Exercise. B1: Contracting a mixed tensor

Let $\dim V = 3$ with basis $\mathcal{B} = (v_1, v_2, v_3)$ and dual basis $(\varphi^1, \varphi^2, \varphi^3)$. Let

$$t = 2\, v_1 \otimes \varphi^1 - v_1 \otimes \varphi^2 + 3\, v_2 \otimes \varphi^2 + v_3 \otimes \varphi^1 - 4\, v_3 \otimes \varphi^3.$$

(a) Write the matrix of the operator S corresponding to t, in the basis $\mathcal{B}$.

(b) Compute $C_1^1(t)$ from Definition 15.4.3, and check it against $\operatorname{tr} S$.

(c) Compute $\operatorname{tr}(S \otimes S)$ without writing down a 9×9 matrix.

Solution.

(a) The tensor $v_i \otimes \varphi^j$ corresponds to the operator $u \mapsto \varphi^j(u) v_i$, whose matrix is the matrix unit $\mathbf{E}_{ij}$ (Theorem 15.3.9). Adding up,

$$[S]_{\mathcal{B}}^{\mathcal{B}} = \begin{pmatrix} 2 & -1 & 0 \\ 0 & 3 & 0 \\ 1 & 0 & -4 \end{pmatrix}.$$

(b) Here

$$\begin{aligned} C_1^1(t) &= 2\varphi^1(v_1) - \varphi^2(v_1) + 3\varphi^2(v_2) + \varphi^1(v_3) - 4\varphi^3(v_3) \\ &= 2 - 0 + 3 + 0 - 4 = 1, \end{aligned}$$

and the diagonal of the matrix in (a) sums to $2 + 3 - 4 = 1$.

(c) By Corollary 15.4.4, $\operatorname{tr}(S \otimes S) = (\operatorname{tr} S)^2 = 1$.

Exercise. B2: Two contractions of one tensor

Let $\dim V = 2$ with basis (v_1, v_2) and dual basis (φ^1, φ^2), and let

$$t = v_1 \otimes \varphi^1 \otimes \varphi^2 + v_2 \otimes \varphi^2 \otimes \varphi^2 \in T_2^1(V).$$

Compute $C_1^1(t)$ and $C_2^1(t)$, and state which space each lies in.

Solution. Both lie in $T_1^0(V) = V^*$. Contracting the single upper slot against the **first** lower slot,

$$C_1^1(t) = \varphi^1(v_1)\,\varphi^2 + \varphi^2(v_2)\,\varphi^2 = \varphi^2 + \varphi^2 = 2\varphi^2.$$

Against the **second** lower slot, the remaining factor is the first functional:

$$C_2^1(t) = \varphi^2(v_1)\,\varphi^1 + \varphi^2(v_2)\,\varphi^2 = 0 + \varphi^2 = \varphi^2.$$

So $C_1^1(t) = 2\varphi^2 \neq \varphi^2 = C_2^1(t)$.

Exercise. B3: Which statements hold?

Let V, W be finite-dimensional over F. Determine which of the following hold. Justify your answer.

(a) $C(\xi) = 0$ implies $\xi = 0$.

(b) $\operatorname{tr}(\Theta(\varphi \otimes v)) = \varphi(v)$ for all $\varphi \in V^*$, $v \in V$.

(c) Every element of $T_2^0(V)$ is a bilinear form on V, and conversely.

(d) $\operatorname{tr}(S \otimes T) = \operatorname{tr}(T \otimes S)$ for $S \in \mathcal{L}(V)$, $T \in \mathcal{L}(W)$.

(e) $C\colon V^* \otimes V \to F$ is surjective whenever $V \neq \{0\}$.

Solution.

(a) Fails for $\dim V \geq 2$. With a basis and its dual, $\xi = \varphi^1 \otimes v_2 \neq 0$ (Proposition 15.2.8) but $C(\xi) = \varphi^1(v_2) = 0$. Contraction is a functional, not an injection; its kernel has dimension $n^2 - 1$.

(b) Holds: Corollary 15.4.2.

(c) Holds. By Proposition 15.4.3, $T_2^0(V) = V^* \otimes V^* \cong (V \otimes V)^*$, and by Definition 15.2.1 the linear maps $V \otimes V \to F$ correspond exactly to the bilinear maps $V \times V \to F$, which are the bilinear forms of Definition 14.1.1.

(d) Holds: both equal $\operatorname{tr} S \cdot \operatorname{tr} T$ by Corollary 15.4.4, the two products of scalars being equal.

(e) Holds. Take $v \neq 0$; by Theorem 5.1.5 there is φ with $\varphi(v) = 1$, so $C(c\,\varphi \otimes v) = c$ for every $c \in F$.

15.4.4.3 C. Going deeper

Exercise. C1: Composition, in tensor language

Let V be finite-dimensional over F and $\Theta\colon V^* \otimes V \to \mathcal{L}(V)$ the isomorphism of Theorem 15.3.9.

(a) Prove that $\Theta(\varphi \otimes v)\,\Theta(\psi \otimes w) = \varphi(w)\,\Theta(\psi \otimes v)$ for all $\varphi, \psi \in V^*$ and $v, w \in V$. Recall that ST means first T, then S.

(b) Hence compute $\operatorname{tr}\big(\Theta(\varphi \otimes v)\,\Theta(\psi \otimes w)\big)$, and observe that the answer is unchanged when (φ, v) and (ψ, w) are exchanged.

(c) Deduce that $\operatorname{tr}(ST) = \operatorname{tr}(TS)$ for all $S, T \in \mathcal{L}(V)$.

Solution.

(a) Both sides are operators on V, so evaluate at an arbitrary $u \in V$. Applying $\Theta(\psi \otimes w)$ first gives $\psi(u)w$, and then

$$\Theta(\varphi \otimes v)\big(\psi(u)w\big) = \psi(u)\,\varphi(w)\,v,$$

using linearity of $\Theta(\varphi \otimes v)$. The right side gives $\varphi(w)\,\psi(u)\,v$, the same scalar times the same vector. Hence the two operators agree at every u.

(b) By (a) and Corollary 15.4.2,

$$\operatorname{tr}\big(\Theta(\varphi \otimes v)\,\Theta(\psi \otimes w)\big) = \varphi(w)\,\psi(v).$$

Exchanging (φ, v) with (ψ, w) turns this into $\psi(v)\,\varphi(w)$, which is the same product of scalars.

(c) The maps $(S, T) \mapsto \operatorname{tr}(ST)$ and $(S, T) \mapsto \operatorname{tr}(TS)$ are bilinear on $\mathcal{L}(V) \times \mathcal{L}(V)$, and the rank-one maps $\Theta(\varphi \otimes v)$ span $\mathcal{L}(V)$ (as Θ is surjective and simple tensors span, by Proposition 15.2.4). By (b) the two maps agree on such pairs, hence everywhere.

Exercise. C2: The contraction of the identity, over any field

Let V be finite-dimensional over F, $n = \dim V \geq 1$, and let $\iota \in V^* \otimes V$ be the tensor with $\Theta(\iota) = \mathrm{id}_V$.

(a) Prove that $\iota = \sum_{i=1}^{n} \varphi^i \otimes v_i$ for **every** basis $\mathcal{B}$ of V with dual basis $\mathcal{B}^*$. (This tensor is basis-free, though each expression for it is not.)

(b) Prove that $C(\iota) = n \cdot 1$ in F.

(c) Over $\mathbb{F}_2$ with $n = 2$, find a tensor $\xi \neq \iota$ in $V^* \otimes V$ with $C(\xi) = C(\iota)$, and relate this to Chapter 4's observation that the trace does not determine the rank of a projection over $\mathbb{F}_2$.

Solution.

(a) By Corollary 15.4.2, $\Theta(\sum_i \varphi^i \otimes v_i) = \mathrm{id}_V$ for any basis and its dual basis. Since Θ is injective (Theorem 15.3.9), all these tensors are equal, and each equals ι.

(b) $C(\iota) = \sum_i \varphi^i(v_i) = \sum_i 1 = n \cdot 1$ by Definition 5.1.3.

(c) Take $\xi = 0$. Then $C(\xi) = 0 = 2 \cdot 1 = C(\iota)$ in $\mathbb{F}_2$, while $\xi \neq \iota$ because $\Theta(\iota) = \mathrm{id} \neq 0$. Under Θ these are the identity and the zero operator, both projections, of ranks 2 and 0; this is exactly Theorem 4.5.4's warning that over a field of characteristic 2 the integer rank P cannot be recovered from $\operatorname{tr} P$. The contraction is faithful to the tensor, but the field can be too small for the resulting scalar to distinguish them.

15.5 Index Notation and the Transformation Law

Outside mathematics, a tensor is usually not a vector in $T^p_q(V)$ at all: it is an array of numbers with some indices written upstairs and some downstairs, together with a rule saying how the array changes when the coordinates change. That older definition is still the working one in physics and in differential geometry, and a reader who meets it after this chapter deserves a dictionary. This section supplies one. We write the components of a tensor, derive — not assert — how they transform under a change of basis, and then prove that the classical definition and ours describe the same objects in finite dimensions. The section ends with the warning the classical definition exists to make: an array of numbers is not automatically a tensor.

15.5.1 Components, upstairs and downstairs

Fix a finite-dimensional V over F with $n = \dim V \geq 1$, an ordered basis $\mathcal{B} = (v_1, \dots, v_n)$, and its dual basis $\mathcal{B}^* = (\varphi^1, \dots, \varphi^n)$, characterized by $\varphi^i(v_j) = \delta^i_j$ (Definition 5.1.3). By Theorem 15.2.5, the n^{p+q} tensors

$$v_{i_1} \otimes \cdots \otimes v_{i_p} \otimes \varphi^{j_1} \otimes \cdots \otimes \varphi^{j_q}$$

form a basis of $T^p_q(V)$, so every tensor has unique coordinates in it.

Definition 15.5.1. Components of a tensor

Let $t \in T^p_q(V)$. The **components of t in the basis $\mathcal{B}$** are the unique scalars $T^{i_1 \dots i_p}_{j_1 \dots j_q} \in F$,

for $1 \le i_1, \dots, i_p, j_1, \dots, j_q \le n$, with

$$t = \sum T^{i_1 \dots i_p}_{j_1 \dots j_q}\, v_{i_1} \otimes \cdots \otimes v_{i_p} \otimes \varphi^{j_1} \otimes \cdots \otimes \varphi^{j_q},$$

the sum running over all $p+q$ indices from 1 to n. An index written **upstairs** belongs to a V-slot and is called **contravariant**; an index written **downstairs** belongs to a V^*-slot and is called **covariant**.

In words: the position of an index records which kind of slot it labels, so that the shape of a symbol already tells you the type of the tensor. Nothing here is new mathematics; it is the coordinate vector of Theorem 15.2.5, with the indices arranged to be readable.

Two pieces of shorthand come with it. The first is the placement rule: a vector has its coordinates upstairs, $v = \sum_i v^i v_i$, while a functional has them downstairs, $\xi = \sum_j \xi_j \varphi^j$ — which is Theorem 5.1.3 (c) written in the new notation. The second is the **summation convention**: an index appearing once upstairs and once downstairs in a product is summed over, and the $\sum$ is not written. With it, the displayed formula above reads

$$t = T^{i_1 \dots i_p}_{j_1 \dots j_q}\, v_{i_1} \otimes \cdots \otimes \varphi^{j_q},$$

and the evaluation of a functional on a vector is $\xi(v) = \xi_i v^i$. The rest of this book writes its sums out; this section is the one place where the convention is used, and it is used because the transformation law is unreadable without it.

Examples.

1. **Type $(1,1)$: an operator.** If $t \in T^1_1(V)$ corresponds to $S \in \mathcal{L}(V)$ under Theorem 15.3.9, then T^i_j is the (i,j) entry of $[S]^{\mathcal{B}}_{\mathcal{B}}$: the tensor $v_i \otimes \varphi^j$ is the operator $u \mapsto \varphi^j(u) v_i$, whose matrix is $\mathbf{E}_{ij}$. Upper index = row, lower index = column.
2. **Type $(0,2)$: a bilinear form.** If $t \in T^0_2(V)$ corresponds to the bilinear form β (Proposition 15.4.3 and Exercise (B3: Which statements hold?)), then $T_{ij} = \beta(v_i, v_j)$, which is the Gram matrix $[\beta]^{\mathcal{B}}_{\mathcal{B}}$ of Definition 14.1.2. Both indices downstairs.
3. **Contraction.** In components, Definition 15.4.3 is "set an upper index equal to a lower one and sum". For $t \in T^1_1(V)$,

$$C^1_1(t) = T^i_i \quad \Big(= \sum_{i=1}^{n} T^i_i \Big),$$

which is the trace, by Theorem 15.4.1. This is the whole reason the convention pairs an upper with a lower index: a repeated index is a contraction, and contraction is the only basis-free way to remove slots.

Remark. The names are historically backwards, and it is worth saying so once. A *contravariant* index sits upstairs, on a V-slot, and its components transform by the **inverse** of the matrix that transforms the basis vectors; a *covariant* index sits downstairs and transforms by the same matrix as the basis vectors. The names describe the components, not the slots. Theorem 15.5.2 below is what they refer to.

15.5.2 How the components change

Let $\mathcal{B} = (v_1, \dots, v_n)$ and $\mathcal{B}' = (v'_1, \dots, v'_n)$ be two ordered bases, and let

$$\mathbf{P} = [\mathrm{id}]^{\mathcal{B}}_{\mathcal{B}'} = (p^i_j)$$

be the change-of-coordinates matrix from $\mathcal{B}'$ to $\mathcal{B}$ (Definition 4.2.1), whose j-th column is $[v'_j]_{\mathcal{B}}$. That is,

$$v'_j = \sum_{i=1}^{n} p^i_j \, v_i,$$

with the row index upstairs and the column index downstairs. Write $\mathbf{Q} = \mathbf{P}^{-1} = (q^i_j)$, which exists because a change-of-coordinates matrix is invertible (Proposition 4.2.2).

The dual bases move by the inverse matrix. This is the one computation the whole section rests on, so we do it first.

Lemma 15.5.1. Dual Bases Move by the Inverse Matrix

With the notation above, $(\varphi')^k = \sum_{l=1}^{n} q^k_l \, \varphi^l$ for each k, where $(\varphi')^1, \dots, (\varphi')^n$ is the dual basis of $\mathcal{B}'$ supplied by Theorem 5.1.3.

Proof. Both sides are linear maps on V, so by the uniqueness clause of Theorem 2.1.5 it is enough to check that they agree at each vector of the basis $\mathcal{B}'$. By Definition 5.1.3 the left side gives $(\varphi')^k(v'_j) = \delta^k_j$. The right side gives

$$\sum_{l} q^k_l \, \varphi^l\Big(\sum_{i} p^i_j v_i\Big) = \sum_{l}\sum_{i} q^k_l \, p^i_j \, \delta^l_i = \sum_{i} q^k_i p^i_j,$$

which is the (k, j) entry of $\mathbf{QP} = \mathbf{I}_n$, that is δ^k_j. The two agree.

■

So the dual basis vectors carry the entries of $\mathbf{P}^{-1}$, while the basis vectors carry the entries of $\mathbf{P}$. Substituting both into the expansion of a tensor gives the transformation law.

Theorem 15.5.2. Transformation Law for Components

Let $t \in T^p_q(V)$ have components $T^{i_1 \dots i_p}_{j_1 \dots j_q}$ in $\mathcal{B}$ and $(T')^{k_1 \dots k_p}_{l_1 \dots l_q}$ in $\mathcal{B}'$, and let $\mathbf{P} = (p^i_j)$ and $\mathbf{Q} = \mathbf{P}^{-1} = (q^i_j)$ be as above. Then, with the summation convention,

$$(T')^{k_1 \dots k_p}_{l_1 \dots l_q} = q^{k_1}_{i_1} \cdots q^{k_p}_{i_p} \times p^{j_1}_{l_1} \cdots p^{j_q}_{l_q} \, T^{i_1 \dots i_p}_{j_1 \dots j_q}.$$

Each **upper** index is contracted with a factor of $\mathbf{P}^{-1}$, each **lower** index with a factor of $\mathbf{P}$.

Idea. There is only one move: write t in the new basis, replace every new basis vector and every new dual basis vector by its expression in the old ones, multiply out using multilinearity of $\otimes$, and compare with the expansion of t in the old basis. Coordinates in a basis are unique, so the comparison is an equality of coefficients. Doing it once for type $(1, 0)$ shows the whole pattern; the general case is the same computation with more indices.

Proof. Start with the expansion of t in the new basis:

$$t = (T')^{k_1\dots k_p}_{l_1\dots l_q}\; v'_{k_1}\otimes\cdots\otimes v'_{k_p}\otimes(\varphi')^{l_1}\otimes\cdots\otimes(\varphi')^{l_q}.$$

Substitute $v'_k = p^i_k v_i$ and, by Lemma 15.5.1, $(\varphi')^l = q^l_j\varphi^j$. Since $\otimes$ is linear in each of its slots, the scalars may be pulled out of all $p+q$ factors at once, giving

$$t = (T')^{k_1\dots k_p}_{l_1\dots l_q}\; p^{i_1}_{k_1}\cdots p^{i_p}_{k_p}\; q^{l_1}_{j_1}\cdots q^{l_q}_{j_q} \\ \times\; v_{i_1}\otimes\cdots\otimes v_{i_p}\otimes\varphi^{j_1}\otimes\cdots\otimes\varphi^{j_q}.$$

The tensors on the right form a basis of $T^p_q(V)$ (Theorem 15.2.5), so coordinates in them are unique (Theorem 1.5.1). Comparing with Definition 15.5.1,

$$T^{i_1\dots i_p}_{j_1\dots j_q} = p^{i_1}_{k_1}\cdots p^{i_p}_{k_p}\; q^{l_1}_{j_1}\cdots q^{l_q}_{j_q}\;(T')^{k_1\dots k_p}_{l_1\dots l_q}. \tag{$*$}$$

It remains to invert $(*)$. Contract both sides of $(*)$ with $q^{m_1}_{i_1}\cdots q^{m_p}_{i_p}\,p^{j_1}_{r_1}\cdots p^{j_q}_{r_q}$ and sum over $i_1,\dots,i_p,j_1,\dots,j_q$. On the right, each pair $q^m_i p^i_k$ sums to $(\mathbf{QP})^m_k = \delta^m_k$, and each pair $q^l_j p^j_r$ sums to δ^l_r; every delta then collapses one summation. What survives on the right is $(T')^{m_1\dots m_p}_{r_1\dots r_q}$, and what stands on the left is the asserted expression. Renaming m to k and r to l gives the theorem.

■

The two smallest cases are results the book already has, and they are worth reading off as a check.

- **Type $(1,1)$.** The law reads $(T')^k_l = q^k_i\,p^j_l\,T^i_j$, that is $\mathbf{T}' = \mathbf{P}^{-1}\mathbf{T}\mathbf{P}$. This is Theorem 4.2.3: matrices of an operator in two bases are **similar**.
- **Type $(0,2)$.** The law reads $(T')_{kl} = p^i_k\,p^j_l\,T_{ij}$, that is $\mathbf{T}' = \mathbf{P}^{\top}\mathbf{T}\mathbf{P}$. This is Theorem 14.2.1: Gram matrices of a bilinear form in two bases are **congruent**.

So similarity and congruence are not two unrelated relations that happen to look alike. They are the transformation laws of type $(1,1)$ and type $(0,2)$, and the difference between $\mathbf{P}^{-1}$ and $\mathbf{P}^{\top}$ is exactly the difference between an upper and a lower index.

15.5.3 The classical definition

The classical definition turns the theorem into a definition: a tensor *is* a rule producing arrays that obey the law.

Definition 15.5.2. Classical tensor

Let V be a vector space over F with $\dim V = n \geq 1$, and let $p, q \geq 0$. A **classical tensor of type** (p,q) on V is a rule $\mathcal{B} \mapsto A_{\mathcal{B}}$ assigning to **every** ordered basis $\mathcal{B}$ of V an array

$$A_{\mathcal{B}} = (A^{i_1\dots i_p}_{j_1\dots j_q})$$

of n^{p+q} scalars, such that for **any two** ordered bases $\mathcal{B}, \mathcal{B}'$, with $\mathbf{P} = [\mathrm{id}]_{\mathcal{B}'}^{\mathcal{B}}$ and $\mathbf{Q} = \mathbf{P}^{-1}$,

$$(A_{\mathcal{B}'})_{l_1 \dots l_q}^{k_1 \dots k_p} = q_{i_1}^{k_1} \cdots q_{i_p}^{k_p} \times p_{l_1}^{j_1} \cdots p_{l_q}^{j_q} \, (A_{\mathcal{B}})_{j_1 \dots j_q}^{i_1 \dots i_p}.$$

Theorem 15.5.3. The Two Definitions Agree

Let V be finite-dimensional over F with $\dim V = n \geq 1$, and let $p, q \geq 0$. The map sending $t \in T_q^p(V)$ to the rule

$$\mathcal{B} \longmapsto (\text{components of } t \text{ in } \mathcal{B})$$

is a bijection from $T_q^p(V)$ onto the set of classical tensors of type (p, q) on V. It is linear for the entrywise operations on arrays.

Idea. Well defined is Theorem 15.5.2. Injective because a tensor is determined by its coordinates in one basis. Surjective is the only step with content, and it is short: fix one basis $\mathcal{B}_0$, build the tensor whose components there are the given array, and then observe that its components in any other basis obey the same law as the rule's array does, starting from the same array — so the two agree.

Proof. **Well defined.** If $t \in T_q^p(V)$, its components in the various bases satisfy the required identity, by Theorem 15.5.2. So the rule attached to t is a classical tensor.

Injective. Fix a basis $\mathcal{B}_0$. If two tensors give the same rule, they have the same components in $\mathcal{B}_0$, hence are equal, since coordinates in the basis of Theorem 15.2.5 are unique (Theorem 1.5.1).

Surjective. Let $\mathcal{B} \mapsto A_{\mathcal{B}}$ be a classical tensor. Fix an ordered basis $\mathcal{B}_0 = (v_1, \dots, v_n)$ with dual basis $(\varphi^1, \dots, \varphi^n)$, and define

$$t \coloneqq (A_{\mathcal{B}_0})_{j_1 \dots j_q}^{i_1 \dots i_p} \, v_{i_1} \otimes \cdots \otimes \varphi^{j_q} \in T_q^p(V),$$

so that the components of t in $\mathcal{B}_0$ are exactly $A_{\mathcal{B}_0}$. Let $\mathcal{B}'$ be any other ordered basis and put $\mathbf{P} = [\mathrm{id}]_{\mathcal{B}'}^{\mathcal{B}_0}$. By Theorem 15.5.2 the components of t in $\mathcal{B}'$ are obtained from $A_{\mathcal{B}_0}$ by the law with this $\mathbf{P}$; by Definition 15.5.2, $A_{\mathcal{B}'}$ is obtained from $A_{\mathcal{B}_0}$ by the law with the same $\mathbf{P}$. Both are the value of the same expression, so they are equal. Hence the rule attached to t is the given one.

Linearity. The components of $t + ct'$ in each basis are $A + cA'$, because coordinate maps are linear (Corollary 2.5.5).

■

Why the older definition survives. It is tempting to file Definition 15.5.2 as a historical curiosity, and that would be a mistake. Three honest reasons keep it in use.

- **Components are what gets measured and computed.** A stress tensor, a metric, an inertia tensor: the physicist has numbers in a chosen frame, and the question that matters is what those numbers become in another frame. The transformation law is not a

consequence to be derived afterwards; it is the datum.

- **There is often no global basis to be abstract about.** On a curved space, the objects of interest are tensor *fields*, and a manifold generally admits no single coordinate system covering it. What it does admit is many local coordinate patches with prescribed rules for passing between them. In that setting "an array in each chart, transforming correctly on overlaps" is a definition that works, while "an element of $T^p_q(V)$" needs a V that varies from point to point, which is a later and heavier construction.
- **It is a usable test.** Given a recipe that produces an array from a basis, checking the transformation law is a finite computation, and it is the standard way to certify that a quantity is geometric rather than an artifact of the coordinates. The next warning is exactly such a test, failed.

15.5.4 An array is not a tensor

Warning. **A matrix is not automatically a tensor.** Consider the rule that assigns to **every** ordered basis of V the same array, written as a matrix,

$$\mathbf{A}_{\mathcal{B}} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \qquad \dim V = 2,$$

read as a type $(1,1)$ array. Take $\mathcal{B} = (v_1, v_2)$ and $\mathcal{B}' = (v_1, v_1 + v_2)$, so that

$$\mathbf{P} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \qquad \mathbf{P}^{-1} = \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix}.$$

The law demands $\mathbf{A}_{\mathcal{B}'} = \mathbf{P}^{-1}\mathbf{A}_{\mathcal{B}}\mathbf{P}$, which is $\begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}$, not $\mathbf{A}_{\mathcal{B}}$. So this rule is **not** a classical tensor of type $(1,1)$: it is a matrix, and nothing more. The operator it describes in $\mathcal{B}$ — projection onto $\operatorname{span}(v_1)$ along $\operatorname{span}(v_2)$ — is a perfectly good tensor, but then its array in $\mathcal{B}'$ is the second matrix, not the first. "Write down the same numbers in every basis" is a rule, and almost never a tensor.

Two refinements of the same point are worth having.

The same array can be two different tensors. Take the array $\mathbf{I}_2$. As a type $(1,1)$ rule it *is* a tensor: $\mathbf{P}^{-1}\mathbf{I}_2\mathbf{P} = \mathbf{I}_2$ for every $\mathbf{P}$, and the tensor is the identity operator, which is the ι of Exercise (C2: The contraction of the identity, over any field). As a type $(0,2)$ rule, the same constant array is **not** a tensor: the law demands $\mathbf{P}^{\top}\mathbf{I}_2\mathbf{P}$, and for the $\mathbf{P}$ above that is $\begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix} \neq \mathbf{I}_2$. Geometrically: "the identity operator" is basis-free, while "the bilinear form whose Gram matrix is the identity", that is, the choice of an inner product declaring the basis orthonormal, is not — it depends on the basis, and orthonormality, as Chapter 11 defines it, is relative to a chosen inner product.

Warning. **Which type an array has is not visible in the array.** The array $\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$ transforms as $\mathbf{P}^{-1}\mathbf{A}\mathbf{P}$ if it is the matrix of an operator and as $\mathbf{P}^{\top}\mathbf{A}\mathbf{P}$ if it is the Gram matrix of a form, and for the $\mathbf{P}$ above those give $\begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}$ and $\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$. Same numbers, same

basis change, different answers. This is why index position is not decoration: writing A^i_j rather than A_{ij} is a statement about what the object is.

Quick check. Let $\dim V = 2$. Is the rule assigning to every ordered basis the array $A_{ij} = \delta_{ij}$ of type $(0, 2)$ ever a classical tensor — say, over $\mathbb{F}_2$, or if we only allow bases obtained from one another by permutation? Decide, with a reason.

Solution. Over $\mathbb{F}_2$ it still fails: with $\mathbf{P} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$, the law demands $\mathbf{P}^\top \mathbf{I}_2 \mathbf{P} = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}$ (since $2 = 0$ in $\mathbb{F}_2$), which is not $\mathbf{I}_2$. If the bases are restricted to permutations of one fixed basis, then $\mathbf{P}$ is a permutation matrix, so $\mathbf{P}^\top = \mathbf{P}^{-1}$ and $\mathbf{P}^\top \mathbf{I}_2 \mathbf{P} = \mathbf{I}_2$: the rule survives. But Definition 15.5.2 quantifies over **every** ordered basis, and a rule that works only for a restricted family is exactly the kind of coordinate-dependent object the definition is designed to exclude. The lesson is that the law is a demand about all changes of basis, not about some.

15.5.5 Exercises

15.5.5.1 A. Check your understanding

Exercise. A1

(a) Define the components of a tensor of type (p, q) in an ordered basis, and say which indices go upstairs.

(b) State the summation convention, and write $\xi(v)$ with it.

(c) For a type $(1, 1)$ tensor, write out the transformation law and name the relation it expresses between the two matrices.

(d) Give the transformation law for type $(0, 2)$ and name the relation.

(e) True or false: any $n \times n$ array of scalars, attached to every basis, is a tensor of type $(1, 1)$. Justify your answer.

Solution.

(a) They are the unique coordinates of t in the basis of Theorem 15.2.5 built from $\mathcal{B}$ and $\mathcal{B}^*$ (Definition 15.5.1). An index belonging to a V-slot goes upstairs; one belonging to a V^*-slot goes downstairs.

(b) An index occurring once upstairs and once downstairs in a product is summed from 1 to n, with the summation sign suppressed. So $\xi(v) = \xi_i v^i$.

(c) $(T')^k_l = q^k_i p^j_l T^i_j$, that is $\mathbf{T}' = \mathbf{P}^{-1}\mathbf{T}\mathbf{P}$: the two matrices are **similar** (Theorem 4.2.3).

(d) $(T')_{kl} = p^i_k p^j_l T_{ij}$, that is $\mathbf{T}' = \mathbf{P}^\top \mathbf{T} \mathbf{P}$: the two matrices are **congruent** (Theorem 14.2.1, Definition 14.2.1).

(e) False. The rule must satisfy the transformation law; the constant rule $\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$ does

not, as the warning above shows.

15.5.5.2 B. Practice

Exercise. B1: Transforming components

Let V be a 2-dimensional space over $\mathbb{Q}$ with basis $\mathcal{B} = (v_1, v_2)$, and let $\mathcal{B}' = (v_1+v_2,\ v_1+2v_2)$.

(a) Write down $\mathbf{P} = [\mathrm{id}]^{\mathcal{B}}_{\mathcal{B}'}$ and $\mathbf{Q} = \mathbf{P}^{-1}$.

(b) Let $t \in T^1_1(V)$ have components $T^1_1 = 1$, $T^1_2 = 2$, $T^2_1 = 3$, $T^2_2 = 4$ in $\mathcal{B}$. Find its components in $\mathcal{B}'$.

(c) Let $s \in T^0_2(V)$ have the same array S_{ij} of components in $\mathcal{B}$. Find its components in $\mathcal{B}'$, and compare with (b).

(d) Compute $C^1_1(t)$ in both bases and check that the two answers agree.

Solution.

(a) The columns of $\mathbf{P}$ are the $\mathcal{B}$-coordinates of the vectors of $\mathcal{B}'$, so

$$\mathbf{P} = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}, \qquad \mathbf{Q} = \mathbf{P}^{-1} = \begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix},$$

using Theorem 0.9.8 with $\det \mathbf{P} = 1$.

(b) By Theorem 15.5.2, $\mathbf{T}' = \mathbf{P}^{-1}\mathbf{T}\mathbf{P}$. With $\mathbf{T} = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$,

$$\mathbf{P}^{-1}\mathbf{T} = \begin{pmatrix} -1 & 0 \\ 2 & 2 \end{pmatrix}, \qquad \mathbf{T}' = \begin{pmatrix} -1 & -1 \\ 4 & 6 \end{pmatrix}.$$

(c) Now the law gives $\mathbf{S}' = \mathbf{P}^{\top}\mathbf{S}\mathbf{P}$. With $\mathbf{P}^{\top} = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}$ (this $\mathbf{P}$ is symmetric),

$$\mathbf{P}^{\top}\mathbf{S} = \begin{pmatrix} 4 & 6 \\ 7 & 10 \end{pmatrix}, \qquad \mathbf{S}' = \begin{pmatrix} 10 & 16 \\ 17 & 27 \end{pmatrix}.$$

The same array in the same two bases produces completely different answers, because the two objects are of different types.

(d) In $\mathcal{B}$: $T^1_1 + T^2_2 = 1 + 4 = 5$. In $\mathcal{B}'$: $-1 + 6 = 5$. They agree, as they must, since Definition 15.4.3 is defined without reference to a basis. (By contrast, the "trace" of the type $(0,2)$ arrays is $1 + 4 = 5$ against $10 + 27 = 37$: there is no contraction of two lower indices, and summing S_{ii} is not a basis-free operation.)

Exercise. B2: Tensor or not?

Let $\dim V = 2$. For each rule below, determine whether it is a classical tensor of the stated type. Justify your answer, using $\mathcal{B} = (v_1, v_2)$ and $\mathcal{B}' = (v_2, v_1)$ or $\mathcal{B}' = (v_1, v_1 + v_2)$ where a counterexample is needed.

(a) Type $(1,1)$: $A_{\mathcal{B}} = \mathbf{I}_2$ for every $\mathcal{B}$.

(b) Type $(1,1)$: $A_{\mathcal{B}} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$ for every $\mathcal{B}$.

(c) Type $(0,2)$: $A_{\mathcal{B}} = \mathbf{O}$ for every $\mathcal{B}$.

(d) Type $(1,1)$: $A_{\mathcal{B}} = (\operatorname{tr} \mathbf{M}_{\mathcal{B}})\mathbf{I}_2$, where $\mathbf{M}_{\mathcal{B}}$ is the matrix in $\mathcal{B}$ of one fixed operator $M \in \mathcal{L}(V)$.

Solution.

(a) Yes. The law demands $\mathbf{P}^{-1}\mathbf{I}_2\mathbf{P} = \mathbf{I}_2$, which holds for every invertible $\mathbf{P}$. The tensor is id_V.

(b) No. With $\mathcal{B}' = (v_2, v_1)$ we get $\mathbf{P} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = \mathbf{P}^{-1}$, and $\mathbf{P}^{-1}A\mathbf{P} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} \neq A$. (This is the standard nondiagonalizable matrix of Chapter 9; here it is failing a different test.)

(c) Yes. The law reads $\mathbf{O} = \mathbf{P}^{\top}\mathbf{O}\mathbf{P}$, which holds. The tensor is the zero tensor, and the zero array is the one array that is a tensor of every type.

(d) Yes. By Theorem 4.2.5 the scalar $c = \operatorname{tr} \mathbf{M}_{\mathcal{B}}$ is the same for every $\mathcal{B}$, so the rule is the constant rule $c\,\mathbf{I}_2$, and $\mathbf{P}^{-1}(c\mathbf{I}_2)\mathbf{P} = c\mathbf{I}_2$. The tensor is $c\,\mathrm{id}_V$. The point of contrast with (b) is that c was itself built basis-freely.

15.5.5.3 C. Going deeper

Exercise. C1: Contraction in indices

Let $\dim V = n$ and let $t \in T^1_1(V)$ have components T^i_j in $\mathcal{B}$ and $(T')^k_l$ in $\mathcal{B}'$.

(a) Prove directly from Theorem 15.5.2 that $(T')^k_k = T^i_i$, without using Theorem 15.4.1.

(b) Let $t \in T^2_1(V)$ with components T^{ij}_k. Show that $U^j := T^{ij}_i$ are the components of a type $(1,0)$ tensor, and identify it in terms of Definition 15.4.3.

(c) Let $s \in T^2_0(V)$ have components S^{ij}. Show that the scalar $\sum_i S^{ii}$ is **not** basis-independent, and say which step of (b) fails for it.

Solution.

(a) By the law, $(T')^k_l = q^k_i p^j_l T^i_j$. Setting $l = k$ and summing,

$$(T')^k_k = \sum_k \sum_{i,j} q^k_i p^j_k T^i_j = \sum_{i,j} \Big(\sum_k p^j_k q^k_i \Big) T^i_j = \sum_{i,j} \delta^j_i T^i_j = \sum_i T^i_i,$$

where the inner sum is the (j,i) entry of $\mathbf{P}\mathbf{Q} = \mathbf{I}_n$.

(b) By Theorem 15.5.2 for type $(2,1)$, $(T')^{km}_l = q^k_i q^m_j p^r_l T^{ij}_r$. Setting $l = k$ and summing over k,

$$(U')^m = \sum_k (T')^{km}_k = q^m_j \Big(\sum_{k,i,r} q^k_i p^r_k T^{ij}_r \Big) = q^m_j\, \delta^r_i T^{ij}_r = q^m_j U^j,$$

which is the transformation law for a single upper index. So (U^j) is a classical tensor of type $(1,0)$, and by Theorem 15.5.3 it is a vector; it is $C^1_1(t)$, the contraction of the

first upper slot against the lower one (Definition 15.4.3).

(c) Take $n = 2$ and $s = v_1 \otimes v_1$, so $S^{11} = 1$ and the other components vanish; then $\sum_i S^{ii} = 1$. Let $\mathcal{B}' = (v_1 + v_2,\ v_2)$, so that $\mathbf{P} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$ and $v_1 = v'_1 - v'_2$. Expanding,

$$s = (v'_1 - v'_2) \otimes (v'_1 - v'_2),$$

so $(S')^{11} = (S')^{22} = 1$ and $(S')^{12} = (S')^{21} = -1$. Hence $\sum_k (S')^{kk} = 2 \neq 1$. The step that fails is the collapse in (b): setting the two indices equal pairs a factor q_i^k with another factor q_j^k, and $\sum_k q_i^k q_j^k$ is an entry of $\mathbf{Q}^\top \mathbf{Q}$, not of $\mathbf{P}\mathbf{Q} = \mathbf{I}_n$, so nothing cancels. Only an upper index repeated against a **lower** one is a contraction.

Exercise. C2: Raising an index needs a form

Let V be finite-dimensional over F and let β be a nondegenerate bilinear form on V, with components $g_{ij} = \beta(v_i, v_j)$ in $\mathcal{B}$.

(a) Prove that the matrix $\mathbf{G} = (g_{ij})$ is invertible, and let (g^{ij}) denote the entries of $\mathbf{G}^{-1}$.

(b) Let $t \in T^0_1(V)$ have components T_j. Show that $U^i := g^{ij} T_j$ satisfies the type $(1,0)$ transformation law, so that "raising the index with β" produces a genuine tensor.

(c) Explain in one or two sentences why this does **not** contradict the warning that there is no canonical map $V^* \to V$.

Solution.

(a) By Proposition 14.1.2, β is nondegenerate exactly when $\mathbf{G} = [\beta]^{\mathcal{B}}_{\mathcal{B}}$ is invertible.

(b) By Theorem 14.2.1, $\mathbf{G}' = \mathbf{P}^\top \mathbf{G} \mathbf{P}$, so $(\mathbf{G}')^{-1} = \mathbf{P}^{-1}\mathbf{G}^{-1}(\mathbf{P}^\top)^{-1} = \mathbf{Q}\mathbf{G}^{-1}\mathbf{Q}^\top$, that is $(g')^{kl} = q_i^k q_j^l g^{ij}$: the inverse array carries two upper indices. Also $(T')_l = p_l^r T_r$ by Theorem 15.5.2. Hence

$$\begin{aligned} (U')^k = (g')^{kl}(T')_l &= q_i^k\, g^{ij} \Big(\sum_l q_j^l p_l^r \Big) T_r \\ &= q_i^k\, g^{ij}\, \delta_j^r\, T_r = q_i^k\, g^{ij} T_j = q_i^k U^i, \end{aligned}$$

using $\sum_l p_l^r q_j^l = (\mathbf{P}\mathbf{Q})_j^r = \delta_j^r$. That is the type $(1,0)$ law.

(c) The map is canonical *given* β, not given V alone: a different nondegenerate form raises indices differently, and there is no distinguished choice. This is the same point as the warning after Definition 15.4.1: the isomorphism $V \cong V^*$ is extra structure, which Chapter 11 supplies with an inner product and Chapter 14 with a form.

15.6 The Tensor Algebra

Sections 03 and 04 already iterated the tensor product: $V^{\otimes k}$ is $V \otimes \cdots \otimes V$ with k factors, its brackets do not matter (Proposition 15.3.5), and k-linear maps out of V^k become linear maps on it (Lemma 15.3.4). Two things are still missing, and Chapter 14 needs both. The first is a **basis** of $V^{\otimes k}$, which the next three sections will cut down in two different ways. The second is a single algebra holding all the powers at once: Chapter 14 §12 described it as the algebra of

"formal sums of strings of vectors, multiplied by concatenation", said no such algebra had been built, and deferred the Clifford algebra until it was. This section builds it, calls it $\mathrm{T}(V)$, and proves the universal property that Sections 07, 08, and 10 will each quotient.

Throughout, F is a field and V is a finite-dimensional vector space over F, with $n = \dim V$. The letter k always denotes a nonnegative integer.

15.6.1 The k-th tensor power

Section 04 wrote $V^{\otimes k}$ for the k-fold tensor product and $V^{\otimes 0} = F$, and Section 03 proved what maps out of it look like. Before using that space three more times it is worth restating the property as a **characterization**, in the form the rest of the chapter will copy: *say what maps out of the object look like, and demand that the object be the most economical one with that property.* The definition below is Definition 15.2.1 with k slots in place of two, and nothing else changes.

The k-th tensor power is the space in which k-linear maps become linear.

Definition 15.6.1. k-th Tensor Power

Let V be a vector space over F and let $k \geq 1$. A **k-th tensor power of** V is a pair (P, μ), where P is a vector space over F and $\mu\colon V^k \to P$ is a k-linear map, such that

(T1) (**universal property**) for **every** vector space W over F and **every** k-linear map $f\colon V^k \to W$, there is a **unique** linear map $\bar{f}\colon P \to W$ with $\bar{f} \circ \mu = f$.

We write $V^{\otimes k}$ for P and

$$v_1 \otimes v_2 \otimes \cdots \otimes v_k \coloneqq \mu(v_1, \ldots, v_k)$$

for the value of μ, and we call such an element a **simple tensor**. By convention $V^{\otimes 0} = F$, with μ the empty product $1 \in F$.

Clause by clause. The pair is what is defined, not the space alone: without μ there is nothing for f to factor through, and a bare vector space of the right dimension satisfies nothing. **Every** vector space W is allowed as a target, including $W = F$ and including P itself; the property is a demand about all of them simultaneously. **Unique** is the clause that does the work later: existence of $\bar{f}$ says P is big enough to record f, and uniqueness says P is not bigger than that.

At $k = 1$ a 1-linear map is a linear map, and the pair (V, id_V) satisfies (T1) with $\bar{f} = f$. So $V^{\otimes 1} = V$, and the notation agrees with itself. At $k = 2$ the definition is Definition 15.2.1 with W replaced by V, so $V^{\otimes 2} = V \otimes V$. The convention $V^{\otimes 0} = F$ is a convention in the same sense as $\mathrm{span}(\emptyset) = \{0\}$ and $0! = 1$: it is the choice that keeps the formulas below true at the boundary, as the dimension count $\dim V^{\otimes k} = n^k$ and the multiplication rule will both confirm.

Uniqueness first, exactly as in Section 02.

Theorem 15.6.1. Uniqueness of the Tensor Power

Let (P, μ) and (P', μ') both be k-th tensor powers of V. Then there is a **unique** linear isomorphism $\varphi\colon P \to P'$ with $\varphi \circ \mu = \mu'$.

Proof. This is the proof of Theorem 15.2.1 with k slots in place of two, and we run it once more so that the pattern is visible. Since μ' is k-linear, (T1) for P gives a linear $\varphi\colon P \to P'$ with $\varphi\mu = \mu'$. Since μ is k-linear, (T1) for P' gives a linear $\psi\colon P' \to P$ with $\psi\mu' = \mu$. Then $\psi\varphi\colon P \to P$ is linear with $\psi\varphi\mu = \psi\mu' = \mu$, and id_P is another such map. By the uniqueness clause of (T1) for P, applied to the k-linear map μ itself, the two coincide: $\psi\varphi = \mathrm{id}_P$. Swapping the roles of P and P' gives $\varphi\psi = \mathrm{id}_{P'}$. Hence φ is an isomorphism, and (T1) delivered it uniquely. This proves the theorem. ∎

Four lines, and they used nothing but the property itself. That is the point of defining objects this way, and it is why the remaining uniqueness proofs in this chapter will be compressed to a citation of this one.

Now existence. Here the chapter can be lazy: a k-th tensor power does not need a new construction, because iterating the binary one already produces it.

Theorem 15.6.2. Existence of the Tensor Power

Let $k \geq 1$ and let $V^{\otimes k} = V \otimes \cdots \otimes V$ be the iterated tensor product of Section 03, with

$$\mu_k(v_1, \ldots, v_k) = v_1 \otimes \cdots \otimes v_k.$$

Then $(V^{\otimes k}, \mu_k)$ is a k-th tensor power of V, and it is spanned by its simple tensors.

Proof. The map μ_k is k-linear, being $\otimes$ applied $k-1$ times and $\otimes$ bilinear in each use. Property (T1) and the spanning claim are exactly Lemma 15.3.4, read with $V_1 = \cdots = V_k = V$. This proves the theorem. ∎

The proof is one sentence because Section 03 did the induction: freeze the last argument, apply the $(k-1)$-factor statement, let the last argument move again, and finish with the two-factor property. Two consequences are used in every argument below, so they are worth saying in words. $V^{\otimes k}$ **is spanned by its simple tensors**, so two linear maps out of $V^{\otimes k}$ are equal as soon as they agree on simple tensors; and **a k-linear recipe on V^k defines a linear map on $V^{\otimes k}$**, which is how every map out of a tensor power gets built, here and in Sections 07 and 08.

From now on $V^{\otimes k}$ denotes this space, and by Theorem 15.6.1 any other k-th tensor power is identified with it by a unique isomorphism respecting the simple tensors. So it is legitimate to speak of *the* k-th tensor power, and the model may be forgotten: every proof below argues from (T1).

15.6.2 A basis, and the dimension

The count is the one Section 01 predicted. A k-linear map on an n-dimensional space may be prescribed freely on the n^k tuples of basis vectors and is determined by those values (Theorem 15.1.1); so $V^{\otimes k}$ had better have dimension n^k, and it does.

Theorem 15.6.3. Basis of a Tensor Power

Let $\mathcal{B} = (v_1, \dots, v_n)$ be a basis of V and let $k \geq 1$. Then the list of simple tensors

$$v_{i_1} \otimes v_{i_2} \otimes \cdots \otimes v_{i_k}, \qquad (i_1, \dots, i_k) \in \{1, \dots, n\}^k,$$

is a basis of $V^{\otimes k}$. In particular $\dim V^{\otimes k} = n^k$.

Idea. Spanning is multilinear expansion: write each argument in the basis and expand one slot at a time. Independence needs, for each tuple J, a linear functional on $V^{\otimes k}$ that is 1 on the tensor indexed by J and 0 on the others. Build it upstairs, as a product of dual basis functionals, and let the universal property carry it down.

Proof. Spanning. By Theorem 15.6.2, $V^{\otimes k}$ is spanned by its simple tensors, so it suffices to expand one of them in the basis $\mathcal{B}$. Write $u_t = \sum_{i=1}^{n} c_{ti} v_i$ for $t = 1, \dots, k$. Expanding $u_1 \otimes \cdots \otimes u_k$ in slot 1, then in slot 2, and so on, using k-linearity of μ_k once per slot, gives

$$u_1 \otimes \cdots \otimes u_k = \sum_{i_1, \dots, i_k} c_{1 i_1} \cdots c_{k i_k} \, v_{i_1} \otimes \cdots \otimes v_{i_k},$$

a linear combination of the listed tensors.

Independence. Let $(\varphi_1, \dots, \varphi_n)$ be the dual basis of $\mathcal{B}$ (Theorem 5.1.3), so $\varphi_i(v_j) = \delta_{ij}$. Fix a tuple $J = (j_1, \dots, j_k)$ and define

$$f_J(u_1, \dots, u_k) = \varphi_{j_1}(u_1)\, \varphi_{j_2}(u_2) \cdots \varphi_{j_k}(u_k).$$

This is k-linear: in position t it is the functional φ_{j_t} times a fixed scalar. By (T1) it induces a linear $\bar{f}_J \colon V^{\otimes k} \to F$ with

$$\bar{f}_J(v_{i_1} \otimes \cdots \otimes v_{i_k}) = \delta_{j_1 i_1} \cdots \delta_{j_k i_k},$$

which is 1 if $I = J$ and 0 otherwise. Now suppose $\sum_I a_I \, v_{i_1} \otimes \cdots \otimes v_{i_k} = 0$, the sum over all n^k tuples I. Applying $\bar{f}_J$ to both sides leaves $a_J = 0$. Since J was arbitrary, every coefficient vanishes.

Hence the list is a basis. It has one member per tuple in $\{1, \dots, n\}^k$, and there are n^k of those, so $\dim V^{\otimes k} = n^k$. This proves the theorem.

■

The count also confirms the convention $V^{\otimes 0} = F$: the formula n^k at $k = 0$ reads $n^0 = 1 = \dim F$, and the only 0-tuple is the empty one.

Quick check. Let $\dim V = 3$. What is $\dim V^{\otimes 4}$, and how many of its basis tensors have all four indices distinct?

Solution. $\dim V^{\otimes 4} = 3^4 = 81$. None: an index tuple of length 4 drawn from $\{1,2,3\}$ must repeat a value, by the pigeonhole principle. This is the first sign that the tensors with a repeated index are not a negligible part of $V^{\otimes k}$ — Section 08 will divide by the span of the simple tensors $u_1 \otimes \cdots \otimes u_k$ that repeat a **vector**, a subspace containing every one of these.

Example. A tensor cube, in coordinates

Let $V = F^2$ with standard basis $(\mathbf{e}_1, \mathbf{e}_2)$. Write out a basis of $V^{\otimes 3}$, and expand $(\mathbf{e}_1 + \mathbf{e}_2) \otimes (\mathbf{e}_1 - \mathbf{e}_2) \otimes \mathbf{e}_1$ in it.

Solution. By Theorem 15.6.3 a basis is indexed by the $2^3 = 8$ triples of indices from $\{1,2\}$:

$$\begin{aligned} &\mathbf{e}_1 \otimes \mathbf{e}_1 \otimes \mathbf{e}_1, \quad \mathbf{e}_1 \otimes \mathbf{e}_1 \otimes \mathbf{e}_2, \quad \mathbf{e}_1 \otimes \mathbf{e}_2 \otimes \mathbf{e}_1, \quad \mathbf{e}_1 \otimes \mathbf{e}_2 \otimes \mathbf{e}_2, \\ &\mathbf{e}_2 \otimes \mathbf{e}_1 \otimes \mathbf{e}_1, \quad \mathbf{e}_2 \otimes \mathbf{e}_1 \otimes \mathbf{e}_2, \quad \mathbf{e}_2 \otimes \mathbf{e}_2 \otimes \mathbf{e}_1, \quad \mathbf{e}_2 \otimes \mathbf{e}_2 \otimes \mathbf{e}_2. \end{aligned}$$

Expanding slot by slot, first in slot 1 and then in slot 2, and noting that slot 3 is already a basis vector,

$$\begin{aligned} (\mathbf{e}_1 + \mathbf{e}_2) \otimes (\mathbf{e}_1 - \mathbf{e}_2) \otimes \mathbf{e}_1 = \mathbf{e}_1 \otimes \mathbf{e}_1 \otimes \mathbf{e}_1 - \mathbf{e}_1 \otimes \mathbf{e}_2 \otimes \mathbf{e}_1 \\ + \mathbf{e}_2 \otimes \mathbf{e}_1 \otimes \mathbf{e}_1 - \mathbf{e}_2 \otimes \mathbf{e}_2 \otimes \mathbf{e}_1. \end{aligned}$$

Four of the eight coordinates are nonzero and the other four are 0. Note also that $\mathbf{e}_1 \otimes \mathbf{e}_2 \otimes \mathbf{e}_1$ and $\mathbf{e}_2 \otimes \mathbf{e}_1 \otimes \mathbf{e}_1$ are **different** basis vectors: the order of the slots is part of the data.

15.6.3 Stacking the powers: the tensor algebra

We now have one space for each k. Chapter 14 wanted a single algebra containing all of them, in which $v_1 \otimes v_2$ and $v_3 \otimes v_4 \otimes v_5$ could be multiplied to give $v_1 \otimes v_2 \otimes v_3 \otimes v_4 \otimes v_5$. The multiplication is concatenation of strings; the only work is to define it without choosing a basis.

First the multiplication on a single pair of powers.

Proposition 15.6.4. Concatenation of Tensors

Let $k, l \geq 1$. There is a unique bilinear map

$$V^{\otimes k} \times V^{\otimes l} \to V^{\otimes (k+l)}, \qquad (x, y) \mapsto x \cdot y,$$

such that for all $u_1, \ldots, u_k, w_1, \ldots, w_l \in V$,

$$(u_1 \otimes \cdots \otimes u_k) \cdot (w_1 \otimes \cdots \otimes w_l) = u_1 \otimes \cdots \otimes u_k \otimes w_1 \otimes \cdots \otimes w_l.$$

Idea. The same peeling move as in Theorem 15.6.2, applied to the two blocks of slots in turn: first make the last l slots into a linear map, then make the first k slots into a linear map whose values are those linear maps.

Proof. Fix $u_1, \dots, u_k \in V$. The map $V^l \to V^{\otimes(k+l)}$ sending $(w_1, \dots, w_l)$ to $u_1 \otimes \cdots \otimes u_k \otimes w_1 \otimes \cdots \otimes w_l$ is l-linear, since μ_{k+l} is $(k+l)$-linear and the first k arguments are fixed. By (T1) for $V^{\otimes l}$ it induces a unique linear map $R(u_1, \dots, u_k)\colon V^{\otimes l} \to V^{\otimes(k+l)}$.

Next, $(u_1, \dots, u_k) \mapsto R(u_1, \dots, u_k)$ is k-linear as a map into $\mathcal{L}(V^{\otimes l}, V^{\otimes(k+l)})$. Indeed, two linear maps $V^{\otimes l} \to V^{\otimes(k+l)}$ are equal as soon as they agree on simple tensors, and on a simple tensor both sides of each linearity identity reduce to linearity of μ_{k+l} in one of its first k slots. By (T1) for $V^{\otimes k}$ we get a linear map

$$M\colon V^{\otimes k} \to \mathcal{L}\big(V^{\otimes l},\, V^{\otimes(k+l)}\big),$$

and we set $x \cdot y = M(x)(y)$. This is linear in x because M is linear, and linear in y because each $M(x)$ is linear, so it is bilinear, and on simple tensors it is the displayed concatenation.

Uniqueness: a bilinear map is determined by its values on pairs of spanning sets, and each factor is spanned by its simple tensors. This proves the proposition.

■

With $k = 0$ or $l = 0$ we let $\cdot$ be scalar multiplication, which is the only bilinear option on $F \times V^{\otimes l}$ agreeing with "prepend nothing". Now stack.

Definition 15.6.2. Tensor Algebra

Let V be a vector space over F. The **tensor algebra** of V is the vector space

$$\mathrm{T}(V) = \bigoplus_{k \geq 0} V^{\otimes k},$$

whose elements are the sequences $x = (x_0, x_1, x_2, \dots)$ with $x_k \in V^{\otimes k}$ and $x_k = 0$ for **all but finitely many** k, with componentwise addition and scalar multiplication, and whose multiplication is

$$(xy)_m = \sum_{k+l=m} x_k \cdot y_l,$$

with $\cdot$ as in Proposition 15.6.4. We identify $V^{\otimes k}$ with the set of sequences supported in position k, so that $F = V^{\otimes 0}$ and $V = V^{\otimes 1}$ are subspaces of $\mathrm{T}(V)$. An element of $V^{\otimes k} \subseteq \mathrm{T}(V)$ is called **homogeneous of degree** k.

The "all but finitely many" clause is the same one that makes $F[x]$ a space of polynomials rather than power series (Definition 0.8.1): an element of $\mathrm{T}(V)$ is a **finite** sum $x_0 + x_1 + \cdots + x_N$, not a series. The formula for $(xy)_m$ is then a finite sum, so the product is defined.

Theorem 15.6.5. The Tensor Algebra Is an Algebra

$\mathrm{T}(V)$ with the above operations is an associative unital F-algebra (Definition 14.12.1), with identity $1 \in F = V^{\otimes 0}$. Moreover $V^{\otimes k} \cdot V^{\otimes l} \subseteq V^{\otimes(k+l)}$ for all k, l.

Proof. Bilinearity of the multiplication is immediate from bilinearity of each $\cdot$ in Proposition 15.6.4, since the components of xy are finite sums of bilinear expressions in the components of x and y.

For associativity it suffices, by bilinearity in each of the three arguments and because the homogeneous components span, to check $(xy)z = x(yz)$ for homogeneous x, y, z, and then, because each power is spanned by its simple tensors, for simple tensors. Both sides are then the single simple tensor obtained by writing the three strings of vectors one after another, so they agree.

For the identity, $1 \in F = V^{\otimes 0}$ acts on each $V^{\otimes k}$ as scalar multiplication by 1, hence as the identity. The grading statement is the target of Proposition 15.6.4. This proves the theorem.

■

From here on we drop the dot and write xy, and for simple tensors we write $u_1 \otimes \cdots \otimes u_k$ for both the element and the product $u_1 u_2 \cdots u_k$ of k elements of V — they are the same thing, since concatenating k strings of length 1 gives the string of length k.

Warning. $\mathrm{T}(V)$ **is not commutative, and not finite-dimensional.** If $\dim V \geq 2$, pick basis vectors $v_1 \neq v_2$; then $v_1 \otimes v_2$ and $v_2 \otimes v_1$ are **distinct basis vectors** of $V^{\otimes 2}$ by Theorem 15.6.3, so $v_1 v_2 \neq v_2 v_1$. And $\dim \mathrm{T}(V) = \sum_k n^k$ is infinite as soon as $n \geq 1$, even though V is finite-dimensional. Every individual element still has only finitely many nonzero components.

Example. The tensor algebra of a line

Let $\dim V = 1$, say $V = Fv$. Identify $\mathrm{T}(V)$ with a familiar algebra.

Solution. By Theorem 15.6.3, $V^{\otimes k}$ has the single basis vector $v^{\otimes k} = v \otimes \cdots \otimes v$, so $\dim V^{\otimes k} = 1^k = 1$. An element of $\mathrm{T}(V)$ is therefore a finite sum $a_0 + a_1 v + a_2 v^{\otimes 2} + \dots$ with $a_k \in F$, and concatenation gives $v^{\otimes k} v^{\otimes l} = v^{\otimes (k+l)}$. Matching $v^{\otimes k}$ with x^k turns these into exactly the addition and multiplication rules of Definition 0.8.2, so $\mathrm{T}(V) \cong F[x]$ as F-algebras.

In particular $\mathrm{T}(V)$ *is* commutative when $\dim V = 1$: there is only one letter, so no two letters can fail to commute. The warning above needed $\dim V \geq 2$, and it needed it for a reason.

15.6.4 The universal property of the tensor algebra

The tensor powers were characterized by what k-linear maps do. The algebra assembled from them has its own characterization, and it is the one Sections 07, 08, and 10 will use: **to define an algebra homomorphism out of $\mathrm{T}(V)$, you only have to say what it does to the vectors.**

Theorem 15.6.6. Universal Property of the Tensor Algebra

Let V be a vector space over F, let A be an associative unital F-algebra, and let $f \colon V \to A$

be a **linear** map. Then there is a **unique** algebra homomorphism

$$\bar{f}\colon \mathrm{T}(V) \to A \qquad \text{with} \qquad \bar{f}|_V = f.$$

Explicitly, $\bar{f}(u_1 \otimes \cdots \otimes u_k) = f(u_1)\cdots f(u_k)$ and $\bar{f}(a) = a1_A$ for $a \in F$.

Idea. Degree by degree. In degree k, the recipe $(u_1, \dots, u_k) \mapsto f(u_1)\cdots f(u_k)$ is k-linear because the product of A is bilinear and f is linear, so (T1) converts it into a linear map on $V^{\otimes k}$. Assemble the pieces and check multiplicativity on simple tensors, which span.

Proof. Existence. For $k \geq 1$ the map $V^k \to A$ given by $(u_1, \dots, u_k) \mapsto f(u_1)f(u_2)\cdots f(u_k)$ is k-linear: with all arguments but the t-th fixed, it is f followed by multiplication on the left and on the right by fixed elements of A, and all three of those maps are linear, the last two by bilinearity of the product (A1). By (T1) there is a unique linear $\bar{f}_k\colon V^{\otimes k} \to A$ with $\bar{f}_k(u_1 \otimes \cdots \otimes u_k) = f(u_1)\cdots f(u_k)$. Put $\bar{f}_0\colon F \to A$, $a \mapsto a1_A$, and define $\bar{f}(x) = \sum_k \bar{f}_k(x_k)$, a finite sum. This is linear, and $\bar{f}(1) = 1_A$.

Multiplicativity: the maps $(x, y) \mapsto \bar{f}(xy)$ and $(x, y) \mapsto \bar{f}(x)\bar{f}(y)$ are both bilinear, so it suffices to compare them on a spanning set of each argument, and the simple tensors of all degrees span $\mathrm{T}(V)$. For two simple tensors,

$$\begin{aligned}\bar{f}((u_1 \otimes \cdots \otimes u_k)(w_1 \otimes \cdots \otimes w_l)) &= f(u_1)\cdots f(u_k)f(w_1)\cdots f(w_l)\\ &= \bar{f}(u_1 \otimes \cdots \otimes u_k)\,\bar{f}(w_1 \otimes \cdots \otimes w_l),\end{aligned}$$

the first equality by Proposition 15.6.4 and the definition of $\bar{f}_{k+l}$, the second by the definition of $\bar{f}_k$ and $\bar{f}_l$ together with associativity in A. The degree-0 cases are the identity $\bar{f}(ax) = a\bar{f}(x)$, which holds by linearity. Finally $\bar{f}|_V = \bar{f}_1 = f$.

Uniqueness. Let φ be an algebra homomorphism with $\varphi|_V = f$. Then $\varphi(1) = 1_A$, so $\varphi(a) = a1_A$ on $V^{\otimes 0}$, and for a simple tensor, multiplicativity gives

$$\varphi(u_1 \otimes \cdots \otimes u_k) = \varphi(u_1)\cdots\varphi(u_k) = f(u_1)\cdots f(u_k).$$

So φ agrees with $\bar{f}$ on the simple tensors, which span $\mathrm{T}(V)$, and both are linear. Hence $\varphi = \bar{f}$. This proves the theorem.

■

In words: $\mathrm{T}(V)$ **is the free associative unital algebra on** V**.** "Free" means that a linear map out of V extends to an algebra map in exactly one way, with no condition on f whatsoever — no multiplicativity is asked of f, because V carries no multiplication to preserve. Compare the corresponding statement one level down, Theorem 2.1.6: a *function* on a basis extends to a linear map in exactly one way. The tensor algebra plays the same role one level up.

Remark. The hypothesis on f is empty, and that is precisely what makes $\mathrm{T}(V)$ too big to be useful on its own. Every interesting algebra built from V is a quotient of $\mathrm{T}(V)$ by the relations one wants to impose, and Theorem 15.6.6 is what shows that the quotient has the expected universal property. Section 07 imposes $uw = wu$, Section 08 imposes $vv = 0$, and Section 10 imposes $vv = q(v)1$, which is the construction Chapter 14 §12 said was

unavailable.

Warning. **Do not confuse the degree with the dimension.** $V^{\otimes k}$ is the space of tensors of degree k; its dimension is n^k, not k. A general element of $V^{\otimes k}$ is a sum of simple tensors, and — exactly as in Section 02 — usually not a single one. A general element of $\mathrm{T}(V)$ is not even homogeneous: $1 + v_1 + v_1 \otimes v_2$ is a perfectly good element of degrees $0, 1, 2$ at once.

15.6.5 Exercises

15.6.5.1 A. Check your understanding

Exercise. A1

(a) State the universal property (T1) of the k-th tensor power, with every quantifier.

(b) Let $\dim V = 4$. Compute $\dim V^{\otimes 0}$, $\dim V^{\otimes 1}$, and $\dim V^{\otimes 3}$.

(c) Give a spanning set of $V^{\otimes k}$, and explain why it lets you check an identity between linear maps out of $V^{\otimes k}$ on simple tensors only.

(d) Determine whether $\mathrm{T}(V)$ is commutative. Justify your answer, distinguishing the possible values of $\dim V$.

(e) State what data is needed to define an algebra homomorphism out of $\mathrm{T}(V)$, and what must be checked about it.

Solution.

(a) For every vector space W over F and every k-linear map $f\colon V^k \to W$, there exists exactly one linear map $\bar{f}\colon V^{\otimes k} \to W$ such that $\bar{f} \circ \mu = f$, where $\mu(v_1, \dots, v_k) = v_1 \otimes \cdots \otimes v_k$.

(b) $\dim V^{\otimes 0} = \dim F = 1$; $\dim V^{\otimes 1} = 4$; $\dim V^{\otimes 3} = 4^3 = 64$, by Theorem 15.6.3.

(c) The simple tensors $u_1 \otimes \cdots \otimes u_k$ span $V^{\otimes k}$ (Theorem 15.6.2); so does the smaller list of Theorem 15.6.3. Two linear maps that agree on a spanning set agree on every linear combination of it, hence everywhere.

(d) It is commutative if $\dim V \leq 1$ and noncommutative if $\dim V \geq 2$. For $\dim V = 0$, $\mathrm{T}(V) = F$; for $\dim V = 1$, $\mathrm{T}(V) \cong F[x]$ by Example (The tensor algebra of a line). For $\dim V \geq 2$, $v_1 \otimes v_2$ and $v_2 \otimes v_1$ are distinct members of a basis by Theorem 15.6.3, so they are unequal.

(e) A **linear** map $f\colon V \to A$ into an associative unital algebra A, and nothing else: no condition on f needs checking, by Theorem 15.6.6. What must be checked is that A really is an associative unital algebra and that f really is linear.

15.6.5.2 B. Practice

Exercise. B1: Expanding a simple tensor

Let $V = F^2$. Expand each of the following in the basis of Theorem 15.6.3.

(a) $(2\mathbf{e}_1 - \mathbf{e}_2) \otimes (\mathbf{e}_1 + 3\mathbf{e}_2)$ in $V^{\otimes 2}$.
(b) $\mathbf{e}_1 \otimes (\mathbf{e}_1 - \mathbf{e}_2) \otimes (\mathbf{e}_1 + \mathbf{e}_2)$ in $V^{\otimes 3}$.

Solution.

(a) Expanding in slot 1, then slot 2,

$$(2\mathbf{e}_1 - \mathbf{e}_2) \otimes (\mathbf{e}_1 + 3\mathbf{e}_2) = 2\,\mathbf{e}_1 \otimes \mathbf{e}_1 + 6\,\mathbf{e}_1 \otimes \mathbf{e}_2 - \mathbf{e}_2 \otimes \mathbf{e}_1 - 3\,\mathbf{e}_2 \otimes \mathbf{e}_2.$$

(b) Slot 1 is already a basis vector, so expand slots 2 and 3:

$$\mathbf{e}_1 \otimes (\mathbf{e}_1 - \mathbf{e}_2) \otimes (\mathbf{e}_1 + \mathbf{e}_2) = \mathbf{e}_1 \otimes \mathbf{e}_1 \otimes \mathbf{e}_1 + \mathbf{e}_1 \otimes \mathbf{e}_1 \otimes \mathbf{e}_2 - \mathbf{e}_1 \otimes \mathbf{e}_2 \otimes \mathbf{e}_1 - \mathbf{e}_1 \otimes \mathbf{e}_2 \otimes \mathbf{e}_2.$$

The four basis tensors beginning with $\mathbf{e}_2$ have coefficient 0, as they must: every term of the expansion carries $\mathbf{e}_1$ in slot 1.

Exercise. B2: A product in the tensor algebra

Let $V = F^2$ and let $x = 1 + \mathbf{e}_1$ and $y = \mathbf{e}_2 - \mathbf{e}_1 \otimes \mathbf{e}_1$ in $\mathrm{T}(V)$. Compute xy and yx, sort the answers by degree, and say whether x and y commute.

Solution. Multiply out, using that 1 is the identity and that concatenation is Proposition 15.6.4:

$$xy = \mathbf{e}_2 - \mathbf{e}_1 \otimes \mathbf{e}_1 + \mathbf{e}_1 \otimes \mathbf{e}_2 - \mathbf{e}_1 \otimes \mathbf{e}_1 \otimes \mathbf{e}_1,$$
$$yx = \mathbf{e}_2 - \mathbf{e}_1 \otimes \mathbf{e}_1 + \mathbf{e}_2 \otimes \mathbf{e}_1 - \mathbf{e}_1 \otimes \mathbf{e}_1 \otimes \mathbf{e}_1.$$

Both have degree-1 part $\mathbf{e}_2$ and degree-3 part $-\mathbf{e}_1 \otimes \mathbf{e}_1 \otimes \mathbf{e}_1$. Their degree-2 parts differ: $-\mathbf{e}_1 \otimes \mathbf{e}_1 + \mathbf{e}_1 \otimes \mathbf{e}_2$ against $-\mathbf{e}_1 \otimes \mathbf{e}_1 + \mathbf{e}_2 \otimes \mathbf{e}_1$. Since $\mathbf{e}_1 \otimes \mathbf{e}_2 \neq \mathbf{e}_2 \otimes \mathbf{e}_1$ by Theorem 15.6.3, $xy \neq yx$: they do not commute.

Exercise. B3: A homomorphism built from the universal property

Let $V = F^2$ and let $A = M_2(F)$. Define $f\colon V \to A$ by

$$f(x_1\mathbf{e}_1 + x_2\mathbf{e}_2) = \begin{pmatrix} 0 & x_1 \\ x_2 & 0 \end{pmatrix}.$$

Prove that there is a unique algebra homomorphism $\bar{f}\colon \mathrm{T}(V) \to M_2(F)$ with $\bar{f}|_V = f$, and compute $\bar{f}(\mathbf{e}_1 \otimes \mathbf{e}_2)$, $\bar{f}(\mathbf{e}_2 \otimes \mathbf{e}_1)$, and $\bar{f}(\mathbf{e}_1 \otimes \mathbf{e}_1)$.

Solution. $M_2(F)$ is an associative unital F-algebra (Definition 14.12.1) and f is linear, since each entry of $f(\mathbf{x})$ is a linear function of $\mathbf{x}$. By Theorem 15.6.6 there is exactly one algebra homomorphism $\bar{f}$ with $\bar{f}|_V = f$.

On simple tensors $\bar{f}$ is the product of the images, so with $\mathbf{A} = f(\mathbf{e}_1)$ and $\mathbf{B} = f(\mathbf{e}_2)$,

$$\mathbf{A} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix},$$

$$\bar{f}(\mathbf{e}_1 \otimes \mathbf{e}_2) = \mathbf{AB} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad \bar{f}(\mathbf{e}_2 \otimes \mathbf{e}_1) = \mathbf{BA} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix},$$

and $\bar{f}(\mathbf{e}_1 \otimes \mathbf{e}_1) = \mathbf{A}^2 = \mathbf{O}$. The first two differ, which is another way to see that $\mathbf{e}_1 \otimes \mathbf{e}_2 \neq \mathbf{e}_2 \otimes \mathbf{e}_1$: a homomorphism cannot separate equal elements. The third shows $\bar{f}$ is far from injective — as it must be, since $\mathrm{T}(V)$ is infinite-dimensional and $M_2(F)$ is not.

15.6.5.3 C. Going deeper

Exercise. C1: The grading is a direct sum of subspaces

(a) Prove that if $x \in V^{\otimes k}$ and $y \in V^{\otimes l}$ with $k \neq l$ are both nonzero, then $x + y$ lies in no single $V^{\otimes m}$.

(b) Deduce that an algebra homomorphism $\varphi\colon \mathrm{T}(V) \to A$ is determined by its restrictions $\varphi|_{V^{\otimes k}}$, and that these are in turn determined by $\varphi|_V$.

Solution.

(a) Under the identification of Definition 15.6.2, x is the sequence with x in position k and 0 elsewhere, and similarly for y. Their sum has nonzero entries in both positions k and l. An element of $V^{\otimes m}$ has at most one nonzero entry, in position m. Since $k \neq l$, the sum has two, so it is in no $V^{\otimes m}$.

(b) Every element of $\mathrm{T}(V)$ is the finite sum of its homogeneous components, so a linear map is determined by its restrictions to the $V^{\otimes k}$. By multiplicativity, $\varphi(u_1 \otimes \cdots \otimes u_k) = \varphi(u_1) \cdots \varphi(u_k)$, which depends only on $\varphi|_V$; and $V^{\otimes k}$ is spanned by its simple tensors, so $\varphi|_{V^{\otimes k}}$ is determined by $\varphi|_V$. (This is the uniqueness half of Theorem 15.6.6, isolated.)

Exercise. C2: Extending an operator to the algebra

Let $T\colon V \to V$ be linear.

(a) Prove that there is a unique algebra homomorphism $\mathrm{T}(T)\colon \mathrm{T}(V) \to \mathrm{T}(V)$ with $\mathrm{T}(T)|_V = T$, and that it maps $V^{\otimes k}$ into $V^{\otimes k}$.

(b) Prove that $\mathrm{T}(ST) = \mathrm{T}(S)\,\mathrm{T}(T)$ for linear $S, T\colon V \to V$, and that $\mathrm{T}(\mathrm{id}_V) = \mathrm{id}$.

(c) Deduce that if T is invertible then so is $\mathrm{T}(T)$.

Solution.

(a) $\mathrm{T}(V)$ is an associative unital algebra and the composite $V \xrightarrow{T} V \hookrightarrow \mathrm{T}(V)$ is linear, so Theorem 15.6.6 gives a unique algebra homomorphism $\mathrm{T}(T)$ restricting to T on V. On a simple tensor it is $u_1 \otimes \cdots \otimes u_k \mapsto Tu_1 \otimes \cdots \otimes Tu_k$, which lies in $V^{\otimes k}$; those tensors span $V^{\otimes k}$, so $\mathrm{T}(T)(V^{\otimes k}) \subseteq V^{\otimes k}$.

(b) $\mathrm{T}(S)\,\mathrm{T}(T)$ is a composite of algebra homomorphisms, hence an algebra homomorphism, and on V it is ST. By the uniqueness clause of Theorem 15.6.6 applied to the linear map $ST\colon V \to \mathrm{T}(V)$, it equals $\mathrm{T}(ST)$. Likewise $\mathrm{id}_{\mathrm{T}(V)}$ is an algebra homomorphism restricting to id_V, so it equals $\mathrm{T}(\mathrm{id}_V)$.

(c) By (b), $\mathrm{T}(T)\,\mathrm{T}(T^{-1}) = \mathrm{T}(TT^{-1}) = \mathrm{T}(\mathrm{id}_V) = \mathrm{id}$, and the same with the factors swapped. Hence $\mathrm{T}(T)$ is invertible with inverse $\mathrm{T}(T^{-1})$.

Exercise. C3: Why the empty degree is F

Suppose we had instead set $V^{\otimes 0} = \{0\}$ and kept every other definition.

(a) Explain why Theorem 15.6.3 would then fail at $k = 0$.

(b) Explain why $\mathrm{T}(V)$ would have no identity element, and hence would not be an algebra in the sense of Definition 14.12.1.

(c) Give one further formula in this section that the convention $V^{\otimes 0} = F$ keeps true at $k = 0$.

Solution.

(a) The theorem counts one basis vector per k-tuple of indices. There is exactly one 0-tuple, the empty one, so the count gives $n^0 = 1$, the dimension of F, not of $\{0\}$.

(b) The identity of $\mathrm{T}(V)$ is $1 \in V^{\otimes 0}$. With $V^{\otimes 0} = \{0\}$ there is no candidate: a homogeneous element of degree $m \geq 1$ multiplies $V^{\otimes k}$ into $V^{\otimes(k+m)}$, so it cannot fix a nonzero element of $V^{\otimes k}$, and neither can a sum of such elements, by Exercise (C1: The grading is a direct sum of subspaces) (a). Hypothesis (A3) of Definition 14.12.1 would fail.

(c) The universal property Theorem 15.6.6: its formula $\bar{f}(a) = a1_A$ for $a \in F$ is the degree-0 case, and it is what makes $\bar{f}$ unital. Equally acceptable: $\dim \mathrm{T}(V) = \sum_{k\geq 0} n^k$, or the multiplication rule $V^{\otimes 0} \cdot V^{\otimes l} \subseteq V^{\otimes l}$, which reads "multiplying by a scalar does not change the degree".

15.7 Symmetric Powers

The tensor algebra is free, and being free it is useless on its own: it records the order of the slots even when nothing in the problem cares about the order. A symmetric bilinear form does not care (Definition 14.3.1); nor does a quadratic form; nor does the monomial x_1x_2, which is the same polynomial as x_2x_1. This section forces $\mathrm{T}(V)$ to forget the order, by dividing the tensor power $V^{\otimes k}$ by the span of the differences $u\otimes w - w\otimes u$. The result, the symmetric power $\mathrm{Sym}^k V$, turns out to have a basis indexed by nondecreasing index tuples, and the algebra assembled from all the $\mathrm{Sym}^k V$ turns out to be the polynomial algebra of Chapter 0, one variable per basis vector.

The construction repeats Section 02's three moves for the second time: universal property, uniqueness, then a model built as a quotient. If the shape of the argument below feels familiar, that is the intention.

Throughout, F is a field, V is a finite-dimensional vector space over F with $n = \dim V \geq 1$, and $k \geq 1$ unless stated otherwise. **The field is arbitrary**, and the single place where the characteristic matters is isolated in the last part of the section.

15.7.1 Symmetric multilinear maps

Before asking for a space, we name the maps it is supposed to linearize. Chapter 14 called a *bilinear form* symmetric when swapping its two arguments changed nothing (Definition 14.3.1). With k arguments there are more ways to rearrange them, and we demand invariance under all of them.

A symmetric map cannot tell in which order its arguments were handed to it.

Definition 15.7.1. Symmetric Multilinear Map

Let W be a vector space over F. A k-linear map $f\colon V^k \to W$ (Definition 15.1.1) is **symmetric** if

$$f(u_{\sigma(1)}, u_{\sigma(2)}, \dots, u_{\sigma(k)}) = f(u_1, u_2, \dots, u_k)$$

for every permutation $\sigma \in S_k$ and **for all** $u_1, \dots, u_k \in V$.

At $k = 2$ the only nonidentity permutation is the swap, so this is exactly Definition 14.3.1 with a general target W in place of F. At $k = 1$ the condition is empty and every linear map is symmetric.

Three examples and one nonexample.

- **A symmetric bilinear form**, $\beta\colon V \times V \to F$ with $\beta(u, w) = \beta(w, u)$. This is the case $k = 2$, $W = F$.
- **Products of one fixed functional.** For $\varphi \in V^*$, the map $(u_1, \dots, u_k) \mapsto \varphi(u_1) \cdots \varphi(u_k)$ is k-linear, by the third example after Definition 7.2.1, and symmetric because multiplication in F is commutative: rearranging the factors does not change the product.
- **Multiplication in a commutative algebra.** If A is a commutative F-algebra and $f\colon V \to A$ is linear, then $(u_1, \dots, u_k) \mapsto f(u_1) \cdots f(u_k)$ is symmetric, by commutativity of A.
- **A nonexample by minimal change.** On $V = F^2$, the map $(\mathbf{u}, \mathbf{w}) \mapsto u_1 w_2$ is bilinear but not symmetric: it sends $(\mathbf{e}_1, \mathbf{e}_2)$ to 1 and $(\mathbf{e}_2, \mathbf{e}_1)$ to 0. Only the symmetry clause fails; bilinearity survives untouched.

Note which permutations are allowed: **all** of S_k, not only the swaps of adjacent slots. Nothing is lost by asking for all of them, and asking for all of them spares us a lemma about generating S_k.

15.7.2 The symmetric power

Now the object. The demand is Definition 15.6.1's, restricted to the symmetric maps.

The k-th symmetric power is the space in which symmetric k-linear maps become linear.

Definition 15.7.2. k-th Symmetric Power

A **k-th symmetric power of** V is a pair (P, π), where P is a vector space over F and $\pi\colon V^k \to P$ is a **symmetric** k-linear map, such that

(S1) (**universal property**) for **every** vector space W over F and **every symmetric** k-linear map $f\colon V^k \to W$, there is a **unique** linear map $\bar{f}\colon P \to W$ with $\bar{f} \circ \pi = f$.

We write $\operatorname{Sym}^k V$ for P and denote the value $\pi(u_1, \dots, u_k)$ by juxtaposition,

$$u_1 u_2 \cdots u_k \coloneqq \pi(u_1, \dots, u_k),$$

calling it a **symmetric product**. By convention $\operatorname{Sym}^0 V = F$.

The one difference from (T1) is the word **symmetric**, in two places: the structure map π is required to be symmetric, and only symmetric maps f are required to factor. Both changes are needed. If π were not symmetric, the notation $uw = wu$ would be false; if nonsymmetric maps had to factor, no space could do the job, since $\bar{f} \circ \pi$ is symmetric whenever π is.

Uniqueness is the template of Theorem 15.2.1 for the third time, and we no longer write it out.

Theorem 15.7.1. Uniqueness of the Symmetric Power

Let (P, π) and (P', π') both be k-th symmetric powers of V. Then there is a unique linear isomorphism $\varphi\colon P \to P'$ with $\varphi \circ \pi = \pi'$.

Proof. Word for word the proof of Theorem 15.6.1, with "symmetric k-linear" in place of "k-linear" throughout: (S1) for P applied to the symmetric map π' gives φ, (S1) for P' applied to π gives ψ, the composite $\psi\varphi$ is a linear self-map of P with $\psi\varphi\pi = \pi$, and the uniqueness clause of (S1) applied to π forces $\psi\varphi = \mathrm{id}_P$. Symmetrically $\varphi\psi = \mathrm{id}_{P'}$. This proves the theorem. ∎

And now the model: the tensor power, divided by the relations we want to impose.

Theorem 15.7.2. Existence of the Symmetric Power

Let $C_k \subseteq V^{\otimes k}$ be the span of all the elements

$$u_1 \otimes \cdots \otimes u_k - u_{\sigma(1)} \otimes \cdots \otimes u_{\sigma(k)}, \qquad u_i \in V,\ \sigma \in S_k,$$

and let $q\colon V^{\otimes k} \to V^{\otimes k}/C_k$ be the quotient map (Definition 4.4.3). Then $V^{\otimes k}/C_k$, together with

$$\pi(u_1, \dots, u_k) = q(u_1 \otimes \cdots \otimes u_k),$$

is a k-th symmetric power of V. At $k = 2$ the subspace C_2 is the span of the elements $u \otimes w - w \otimes u$.

Idea. Two things to check, and both are one line each once the right tool is named. That π is symmetric is the definition of C_k: permuting the arguments changes the tensor by an element of C_k, which q kills. That symmetric maps factor is Theorem 4.4.10: turn f into a linear map on $V^{\otimes k}$ by (T1), observe that symmetry of f makes it vanish on the generators of C_k, and let the quotient's universal property do the rest. This is the move the chapter uses three times: **a quotient kills exactly the relations you want.**

Proof. C_k is a subspace, being a span. The map π is k-linear, being the k-linear μ_k of Theorem 15.6.2 followed by the linear map q. It is symmetric: for $\sigma \in S_k$, the difference

$$u_1 \otimes \cdots \otimes u_k - u_{\sigma(1)} \otimes \cdots \otimes u_{\sigma(k)}$$

lies in C_k by definition, so the two tensors have the same image under q: $\pi(u_{\sigma(1)}, \dots, u_{\sigma(k)}) = \pi(u_1, \dots, u_k)$.

Now let $f\colon V^k \to W$ be symmetric k-linear. By (T1) there is a unique linear $\tilde{f}\colon V^{\otimes k} \to W$ with $\tilde{f}(u_1 \otimes \cdots \otimes u_k) = f(u_1, \dots, u_k)$. On a generator of C_k,

$$\tilde{f}(u_1 \otimes \cdots \otimes u_k - u_{\sigma(1)} \otimes \cdots \otimes u_{\sigma(k)}) = f(u_1, \dots, u_k) - f(u_{\sigma(1)}, \dots, u_{\sigma(k)}),$$

which is 0 because f is symmetric. A linear map vanishing on a spanning set of C_k vanishes on C_k, so $C_k \subseteq \ker \tilde{f}$. By Theorem 4.4.10 (a) there is exactly one linear $\bar{f}\colon V^{\otimes k}/C_k \to W$ with $\bar{f} \circ q = \tilde{f}$, and then

$$\bar{f} \circ \pi = \bar{f} \circ q \circ \mu_k = \tilde{f} \circ \mu_k = f.$$

For uniqueness, suppose h is linear with $h \circ \pi = f$. Since q is surjective and $V^{\otimes k}$ is spanned by its simple tensors (Theorem 15.6.2), the elements $\pi(u_1, \dots, u_k)$ span $V^{\otimes k}/C_k$. So h and $\bar{f}$ agree on a spanning set, hence everywhere. This proves the theorem.

■

Warning. $\operatorname{Sym}^k V$ **is a quotient of $V^{\otimes k}$, not a subspace of it.** Its elements are cosets, and the identity $uw = wu$ in $\operatorname{Sym}^2 V$ says that $u \otimes w$ and $w \otimes u$ lie in the **same** coset. It does **not** say $u \otimes w = w \otimes u$ in $V^{\otimes 2}$, which is false for independent u, w by Theorem 15.6.3. Confusing the two is the commonest error here; the safe habit is never to write an element of $\operatorname{Sym}^k V$ as a tensor.

15.7.3 A basis, and the dimension

Sorting an index tuple is now legal, so the basis of Theorem 15.6.3 collapses: only the sorted tuples survive. Counting them is the substance of the section.

Theorem 15.7.3. Basis of a Symmetric Power

Let $\mathcal{B} = (v_1, \dots, v_n)$ be a basis of V and $k \geq 1$. Then the symmetric products

$$v_{i_1} v_{i_2} \cdots v_{i_k}, \qquad 1 \leq i_1 \leq i_2 \leq \cdots \leq i_k \leq n,$$

one for each **nondecreasing** tuple, form a basis of $\operatorname{Sym}^k V$. Consequently

$$\dim \operatorname{Sym}^k V = \binom{n+k-1}{k},$$

over **every** field F.

Idea. Spanning is sorting: the image of any index tuple equals the image of its sorted version, so the sorted ones already span. Independence needs one linear functional per nondecreasing tuple, and it must be built **without dividing by anything**, since we want the theorem over every field. The functional that works is a sum over all ways of dealing out the n dual basis functionals to the k slots with prescribed multiplicities: every coefficient in it is 1. Counting the tuples is then the stars-and-bars bijection, which turns a nondecreasing tuple into a strictly increasing one by adding $0, 1, 2, \dots$ to its entries.

Proof. Spanning. By Theorem 15.6.3 the tensors $v_{j_1} \otimes \cdots \otimes v_{j_k}$, over all tuples J, form a basis of $V^{\otimes k}$, and q is surjective, so their images $v_{j_1} \cdots v_{j_k}$ span $\operatorname{Sym}^k V$. Given any J, let $\sigma \in S_k$ sort it, so that $(j_{\sigma(1)}, \dots, j_{\sigma(k)})$ is nondecreasing. Since π is symmetric, $v_{j_1} \cdots v_{j_k} = v_{j_{\sigma(1)}} \cdots v_{j_{\sigma(k)}}$. So the nondecreasing tuples already give a spanning set.

Independence. Let $(\varphi_1, \dots, \varphi_n)$ be the dual basis of $\mathcal{B}$ (Theorem 5.1.3). For a nondecreasing tuple $I = (i_1, \dots, i_k)$, let

$$m_j = |\{\, t : i_t = j \,\}|, \qquad j = 1, \dots, n,$$

be its **multiplicity vector**, so $m_1 + \cdots + m_n = k$, and let $\Phi(I)$ be the set of functions $\tau \colon \{1, \dots, k\} \to \{1, \dots, n\}$ taking the value j exactly m_j times, for each j. Define

$$g_I(u_1, \dots, u_k) = \sum_{\tau \in \Phi(I)} \varphi_{\tau(1)}(u_1)\, \varphi_{\tau(2)}(u_2) \cdots \varphi_{\tau(k)}(u_k).$$

Each summand is k-linear, being a product of one functional per slot, so g_I is k-linear.

It is symmetric. Let $\rho \in S_k$. Substituting and then reindexing the product by $s = \rho(t)$,

$$g_I(u_{\rho(1)}, \dots, u_{\rho(k)}) = \sum_{\tau \in \Phi(I)} \prod_{s=1}^{k} \varphi_{(\tau\rho^{-1})(s)}(u_s).$$

The map $\tau \mapsto \tau\rho^{-1}$ is a bijection of $\Phi(I)$ onto itself, since composing with a permutation of the domain does not change how often a value is taken. So the sum is unchanged, and g_I is symmetric.

By (S1), g_I induces a linear $\bar{g}_I \colon \operatorname{Sym}^k V \to F$. Evaluate it on a basis candidate: for any tuple J,

$$\bar{g}_I(v_{j_1} \cdots v_{j_k}) = \sum_{\tau \in \Phi(I)} \prod_{t=1}^{k} \delta_{\tau(t)\, j_t}.$$

A summand is 1 exactly when $\tau(t) = j_t$ for every t, and 0 otherwise; so the whole sum is 1 if the function $t \mapsto j_t$ belongs to $\Phi(I)$, and 0 if it does not. For nondecreasing I and J, that function lies in $\Phi(I)$ exactly when J has the same multiplicity vector as I, which for nondecreasing tuples means $J = I$. Hence $\bar{g}_I(v_{j_1} \cdots v_{j_k}) = \delta_{IJ}$.

Now suppose $\sum_I a_I\, v_{i_1} \cdots v_{i_k} = 0$, the sum over nondecreasing tuples. Applying $\bar{g}_J$ leaves $a_J = 0$. Since J was arbitrary, every coefficient is 0, and the list is independent. So it is a basis.

The count. Send a nondecreasing tuple $(i_1, \dots, i_k)$ to

$$(i_1,\ i_2 + 1,\ i_3 + 2,\ \dots,\ i_k + k - 1).$$

Consecutive entries satisfy $i_t + t - 1 < i_{t+1} + t$ because $i_t \le i_{t+1}$, so the image is **strictly** increasing, and its entries lie between 1 and $n+k-1$. Conversely, subtracting $0, 1, \dots, k-1$ from a strictly increasing tuple in $\{1, \dots, n+k-1\}$ returns a nondecreasing tuple in $\{1, \dots, n\}$: the differences $j_{t+1} - j_t \ge 1$ become $i_{t+1} - i_t \ge 0$, and $1 \le j_1$, $j_k \le n + k - 1$ give $1 \le i_1$ and $i_k \le n$. The two assignments are mutually inverse, so the nondecreasing k-tuples from $\{1, \dots, n\}$ are in bijection with the k-element subsets of $\{1, \dots, n + k - 1\}$, and there are $\binom{n+k-1}{k}$ of those. This proves the theorem.

■

Nothing in the proof divided by an integer, which is why the conclusion holds over $\mathbb{F}_2$ as it does over $\mathbb{C}$. Keep that in view: the last part of this section shows that a *different* and very common description of $\operatorname{Sym}^k V$ does not survive over $\mathbb{F}_2$, and the difference between the two is exactly one division.

Quick check. Let $\dim V = 3$. Compute $\dim \operatorname{Sym}^2 V$ and $\dim V^{\otimes 2}$, and say which basis tensors of $V^{\otimes 2}$ become equal in $\operatorname{Sym}^2 V$.

Solution. $\dim \operatorname{Sym}^2 V = \binom{4}{2} = 6$ and $\dim V^{\otimes 2} = 3^2 = 9$. The six surviving basis elements are $v_1^2, v_1v_2, v_1v_3, v_2^2, v_2v_3, v_3^2$. The three pairs $v_i \otimes v_j$ and $v_j \otimes v_i$ with $i < j$ become equal, which removes 3 from the count: $9 - 3 = 6$.

Example. A symmetric square, in coordinates

Let $V = F^2$. Write down a basis of $\operatorname{Sym}^2 V$ and expand $(\mathbf{e}_1 + 2\mathbf{e}_2)(3\mathbf{e}_1 - \mathbf{e}_2)$ in it.

Solution. By Theorem 15.7.3 a basis is $(\mathbf{e}_1^2,\ \mathbf{e}_1\mathbf{e}_2,\ \mathbf{e}_2^2)$, of size $\binom{3}{2} = 3$. Expanding one slot at a time and then using $\mathbf{e}_2\mathbf{e}_1 = \mathbf{e}_1\mathbf{e}_2$,

$$\begin{aligned}(\mathbf{e}_1 + 2\mathbf{e}_2)(3\mathbf{e}_1 - \mathbf{e}_2) &= 3\mathbf{e}_1\mathbf{e}_1 - \mathbf{e}_1\mathbf{e}_2 + 6\mathbf{e}_2\mathbf{e}_1 - 2\mathbf{e}_2\mathbf{e}_2 \\ &= 3\mathbf{e}_1^2 + 5\mathbf{e}_1\mathbf{e}_2 - 2\mathbf{e}_2^2.\end{aligned}$$

Compare with expanding $(x + 2y)(3x - y) = 3x^2 + 5xy - 2y^2$ in ordinary algebra. That is not a coincidence, and the next part of this section says exactly what it is.

Warning. **Not every element of $\operatorname{Sym}^k V$ is a product of k vectors.** Take $F = \mathbb{R}$, $V = \mathbb{R}^2$, and $\omega = \mathbf{e}_1^2 + \mathbf{e}_2^2 \in \operatorname{Sym}^2 V$. If $\omega = \mathbf{uw}$ with $\mathbf{u} = a\mathbf{e}_1 + b\mathbf{e}_2$ and $\mathbf{w} = c\mathbf{e}_1 + d\mathbf{e}_2$, then comparing coordinates in the basis $(\mathbf{e}_1^2, \mathbf{e}_1\mathbf{e}_2, \mathbf{e}_2^2)$ gives $ac = 1$, $ad + bc = 0$, and $bd = 1$. The first and last force $a, b, c, d \neq 0$, and multiplying $ad = -bc$ by bc gives $-(bc)^2 = (ac)(bd) = 1$, impossible in $\mathbb{R}$. This is the same phenomenon as the nonsimple tensors of Section 02, and it is why $\operatorname{Sym}^k V$ is much larger than the set of products.

15.7.4 The symmetric algebra is a polynomial algebra

Stacking the symmetric powers gives an algebra, exactly as stacking the tensor powers did, and the multiplication descends from concatenation.

Proposition 15.7.4. The Symmetric Product of Two Symmetric Tensors

Let $k, l \geq 1$. There is a unique bilinear map $\operatorname{Sym}^k V \times \operatorname{Sym}^l V \to \operatorname{Sym}^{k+l} V$ with

$$(u_1 \cdots u_k)\,(w_1 \cdots w_l) = u_1 \cdots u_k w_1 \cdots w_l.$$

Proof. *Concatenation carries the relations into the relations.* Let $x \in C_k$ and $y \in V^{\otimes l}$. For a generator $x = u_1 \otimes \cdots \otimes u_k - u_{\sigma(1)} \otimes \cdots \otimes u_{\sigma(k)}$ and a simple $y = w_1 \otimes \cdots \otimes w_l$, the product $x \cdot y$ of Proposition 15.6.4 is again a generator of C_{k+l}, for the permutation acting as σ on the first k slots and fixing the last l. Since generators span C_k, since $V^{\otimes l}$ is spanned by its simple tensors, and since $\cdot$ is bilinear, $C_k \cdot V^{\otimes l} \subseteq C_{k+l}$. Swapping the roles of the two blocks gives $V^{\otimes k} \cdot C_l \subseteq C_{k+l}$.

Descending, one argument at a time. Let $q_m \colon V^{\otimes m} \to \operatorname{Sym}^m V$ denote the quotient maps. Fix $y \in V^{\otimes l}$. The linear map $x \mapsto q_{k+l}(x \cdot y)$ kills C_k by the first inclusion, so Theorem 4.4.10 gives a unique linear $\beta_y \colon \operatorname{Sym}^k V \to \operatorname{Sym}^{k+l} V$ with $\beta_y(q_k x) = q_{k+l}(x \cdot y)$. The assignment $y \mapsto \beta_y$ is linear, since $\beta_{ay+y'}$ and $a\beta_y + \beta_{y'}$ agree on the spanning set $q_k(V^{\otimes k}) = \operatorname{Sym}^k V$; and it kills C_l, since $y \in C_l$ gives $x \cdot y \in C_{k+l}$ and hence $\beta_y = 0$. By Theorem 4.4.10 again it descends to a linear map $\operatorname{Sym}^l V \to \mathcal{L}(\operatorname{Sym}^k V, \operatorname{Sym}^{k+l} V)$, and evaluating gives the required bilinear map. It is unique because the products of vectors span each factor. When $k = 0$ or $l = 0$ the product is scalar multiplication, as in Section 06. This proves the proposition. ∎

Definition 15.7.3. Symmetric Algebra

The **symmetric algebra** of V is the vector space

$$\operatorname{Sym}(V) = \bigoplus_{k \geq 0} \operatorname{Sym}^k V,$$

of sequences with finitely many nonzero components, with the multiplication assembled from Proposition 15.7.4 by $(xy)_m = \sum_{k+l=m} x_k y_l$.

Proposition 15.7.5. $\operatorname{Sym}(V)$ Is a Commutative Algebra

$\operatorname{Sym}(V)$ is an associative, **commutative**, unital F-algebra with identity $1 \in \operatorname{Sym}^0 V = F$, and $\operatorname{Sym}^k V \cdot \operatorname{Sym}^l V \subseteq \operatorname{Sym}^{k+l} V$.

Proof. Bilinearity, associativity, and the identity are proved exactly as in Theorem 15.6.5: all three are identities between multilinear expressions, so it suffices to check them on products of vectors, which span each $\operatorname{Sym}^k V$, and there all three read “concatenate the strings”.

Commutativity is the new point. Both $(u_1 \cdots u_k)(w_1 \cdots w_l)$ and $(w_1 \cdots w_l)(u_1 \cdots u_k)$ equal the symmetric product of the same $k+l$ vectors in two different orders, and π is symmetric, so they are equal. Two bilinear maps agreeing on spanning sets of both arguments agree everywhere, so $xy = yx$ for all x, y. This proves the proposition.

■

To name what $\mathrm{Sym}(V)$ is, we need polynomials in more than one variable. Chapter 0 built $F[x]$; the several-variable version is the same construction with monomials indexed by exponent vectors.

Definition 15.7.4. The Polynomial Algebra $F[x_1, \dots, x_n]$

Let $n \geq 1$. The **polynomial algebra in n variables** over F is the vector space $F[x_1, \dots, x_n]$ with basis the **monomials**

$$x^a = x_1^{a_1} x_2^{a_2} \cdots x_n^{a_n}, \qquad a = (a_1, \dots, a_n) \in \mathbb{N}^n,$$

with the unique bilinear multiplication satisfying $x^a x^b = x^{a+b}$. The **degree** of x^a is $|a| = a_1 + \cdots + a_n$, and $F[x_1, \dots, x_n]_k$ denotes the span of the monomials of degree k.

The bilinear extension exists and is unique because a bilinear map may be prescribed freely on a pair of bases, by the argument used twice in the proof of Proposition 15.6.4. The multiplication is associative, commutative, and unital, with identity the monomial all of whose exponents are 0, because exponent addition in $\mathbb{N}^n$ is associative, commutative, and has an identity. For $n = 1$ the monomials are $1, x, x^2, \dots$ and the rule is $x^a x^b = x^{a+b}$, so $F[x_1]$ is the ring $F[x]$ of Definition 0.8.2, whose basis of monomials is Example (A Basis of $F[x]$). Everything Chapter 6 proved — division, greatest common divisors, unique factorization — lives in that one-variable case.

Theorem 15.7.6. The Symmetric Algebra Is a Polynomial Algebra

Let $\mathcal{B} = (v_1, \dots, v_n)$ be a basis of V. Then there is an isomorphism of F-algebras

$$\Psi \colon \mathrm{Sym}(V) \xrightarrow{\sim} F[x_1, \dots, x_n], \qquad \Psi(v_i) = x_i,$$

and it carries $\mathrm{Sym}^k V$ onto $F[x_1, \dots, x_n]_k$. In particular

$$\dim F[x_1, \dots, x_n]_k = \binom{n+k-1}{k}.$$

Idea. Both sides have a basis indexed by the same combinatorial object: a nondecreasing k-tuple on the left, a degree-k monomial on the right, and each is determined by its multiplicity vector $m \in \mathbb{N}^n$ with $|m| = k$. So match the bases, which gives a linear isomorphism for free, and then check multiplicativity on basis elements, where both products are "add the exponent vectors".

Proof. Write $|m| = m_1 + \cdots + m_n$ for $m \in \mathbb{N}^n$. Given m with $|m| = k$, let $I(m)$ be the nondecreasing tuple taking the value j exactly m_j times, and write v^m for the corresponding

symmetric product $v_{i_1} \cdots v_{i_k}$. The assignment $m \mapsto I(m)$ is a bijection onto the nondecreasing k-tuples, since a nondecreasing tuple is recovered from its multiplicity vector and conversely. So by Theorem 15.7.3 the family $(v^m)_{|m|=k}$ is a basis of $\mathrm{Sym}^k V$, and hence $(v^m)_{m \in \mathbb{N}^n}$ is a basis of $\mathrm{Sym}(V)$.

Define Ψ to be the unique linear map with $\Psi(v^m) = x^m$, which exists by Theorem 2.1.6. It sends a basis of $\mathrm{Sym}(V)$ bijectively onto the basis of monomials, so it is a linear isomorphism, and it sends the degree-k basis onto the degree-k monomials, so $\Psi(\mathrm{Sym}^k V) = F[x_1, \ldots, x_n]_k$.

Multiplicativity: the maps $(y, z) \mapsto \Psi(yz)$ and $(y, z) \mapsto \Psi(y)\Psi(z)$ are bilinear, so it is enough to compare them on the basis. By Proposition 15.7.4, $v^m v^{m'}$ is the symmetric product of the $|m| + |m'|$ vectors of the two strings, which after sorting is $v^{m+m'}$; and $x^m x^{m'} = x^{m+m'}$ by Definition 15.7.4. Also Ψ sends the empty product 1 to the monomial with all exponents 0, which is 1. Hence Ψ is an algebra isomorphism. Finally v_i is the basis element v^m whose m has a single 1 in position i, so $\Psi(v_i) = x_i$.

The dimension formula follows from Theorem 15.7.3, since Ψ restricts to an isomorphism in each degree. This proves the theorem.

■

So the symmetric algebra is nothing exotic: *it is the polynomial algebra, with one variable per basis vector of V.* The isomorphism depends on the basis, exactly as the coordinate isomorphism $V \cong F^n$ does; $\mathrm{Sym}(V)$ is the basis-free version of "polynomials in n variables", and $\mathrm{Sym}^k V$ is the basis-free version of "homogeneous polynomials of degree k". Read backwards, the theorem also computes something Chapter 6 never did: the number of monomials of degree k in n variables is $\binom{n+k-1}{k}$.

Remark. The case $k = 2$ recovers a count from Chapter 14. A symmetric bilinear form on V is a symmetric 2-linear map into F, so by (S1) the symmetric forms on V are in bijection with the linear functionals on $\mathrm{Sym}^2 V$, and there are $\binom{n+1}{2} = n(n+1)/2$ dimensions of them — the number of entries on and above the diagonal of a symmetric matrix, which is what Proposition 14.3.1 would give by counting.

15.7.5 Symmetric powers versus symmetric tensors

Here is the trap, and it is worth the space. Many treatments define the k-th symmetric power as a **subspace** of $V^{\otimes k}$: the tensors left fixed by every permutation of the slots. That is a different construction from the quotient above, and over a field of characteristic 0 the two agree. Over $\mathbb{F}_2$ they do not, and the reason is a single division by $k!$.

We first need the permutation operators.

Lemma 15.7.7. Permuting the Slots of a Tensor

For each $\sigma \in S_k$ there is a unique linear operator P_σ on $V^{\otimes k}$ with

$$P_\sigma(u_1 \otimes \cdots \otimes u_k) = u_{\sigma(1)} \otimes \cdots \otimes u_{\sigma(k)},$$

and $P_\sigma P_\tau = P_{\tau\sigma}$ for all $\sigma, \tau \in S_k$, with $P_{\mathrm{id}} = \mathrm{id}$.

Proof. The map $(u_1, \dots, u_k) \mapsto u_{\sigma(1)} \otimes \cdots \otimes u_{\sigma(k)}$ is k-linear, since in slot t it is μ_k in slot $\sigma^{-1}(t)$. So (T1) gives a unique linear P_σ with the stated values, and uniqueness holds because those tensors span $V^{\otimes k}$. For the composition rule, put $w_t = u_{\tau(t)}$; then

$$P_\sigma P_\tau (u_1 \otimes \cdots \otimes u_k) = P_\sigma (w_1 \otimes \cdots \otimes w_k) = w_{\sigma(1)} \otimes \cdots \otimes w_{\sigma(k)},$$

and $w_{\sigma(t)} = u_{\tau(\sigma(t))}$, which is the t-th argument of $P_{\tau\sigma}$. Both sides are linear and agree on simple tensors, hence are equal. This proves the lemma. ∎

Definition 15.7.5. Symmetric Tensor

A tensor $x \in V^{\otimes k}$ is **symmetric** if $P_\sigma x = x$ for **every** $\sigma \in S_k$. The symmetric tensors form a subspace of $V^{\otimes k}$, which we write $(V^{\otimes k})^{S_k}$; it is the intersection of the kernels of the linear maps $P_\sigma - \mathrm{id}$.

Theorem 15.7.8. Symmetric Tensors and the Symmetric Power

Let $n = \dim V \geq 1$ and $k \geq 1$, and let $\iota \colon (V^{\otimes k})^{S_k} \to \operatorname{Sym}^k V$ be the restriction of the quotient map q.

(a) Over **every** field, $\dim (V^{\otimes k})^{S_k} = \binom{n+k-1}{k} = \dim \operatorname{Sym}^k V$.

(b) If $\operatorname{char} F = 0$ or $\operatorname{char} F > k$, then ι is an isomorphism.

(c) If $\operatorname{char} F = 2$, $k = 2$, and $n \geq 2$, then ι is **neither** injective nor surjective.

Idea. For (a), a symmetric tensor must have equal coordinates along each orbit of basis tensors, so the **orbit sums** are a basis; orbits correspond to multisets, which Theorem 15.7.3 already counted. For (b), average: $s = \sum_\sigma P_\sigma$ maps everything into the symmetric tensors and kills C_k, so $\frac{1}{k!} s$ is an inverse for ι — provided $k!$ is invertible, which is exactly the hypothesis. For (c), the tensor $v_1 \otimes v_2 + v_2 \otimes v_1$ is symmetric and nonzero, but its image is $2 v_1 v_2 = 0$.

Proof.

(a) Let $\mathcal{B} = (v_1, \dots, v_n)$ be a basis of V. For a tuple J write $v_J = v_{j_1} \otimes \cdots \otimes v_{j_k}$, so that $(v_J)_J$ is a basis of $V^{\otimes k}$ (Theorem 15.6.3), and note $P_\sigma v_J = v_{J\sigma}$, where $J\sigma$ is the tuple $t \mapsto j_{\sigma(t)}$. Two tuples lie in the same **orbit** when one is a rearrangement of the other, that is, when they have the same multiplicity vector. For each nondecreasing I let $s_I = \sum_{J \in O_I} v_J$, the sum over the orbit of I.
Each s_I is symmetric, because $J \mapsto J\sigma$ permutes O_I. The s_I are independent, because distinct orbits involve disjoint sets of basis vectors. They span: if $x = \sum_J a_J v_J$ satisfies $P_\sigma x = x$ for every σ, then comparing coordinates in $\sum_J a_J v_{J\sigma} = \sum_J a_J v_J$ gives $a_J = a_{J\sigma}$ for all J and σ, so $J \mapsto a_J$ is constant on orbits and $x = \sum_I a_I s_I$. Hence (s_I) is a basis, with one member per orbit, that is, per multiplicity vector of total k. By

the bijection in the proof of Theorem 15.7.3 there are $\binom{n+k-1}{k}$ of those. No division occurred, so this holds over every field.

(b) Let $s = \sum_{\sigma \in S_k} P_\sigma$, a linear operator on $V^{\otimes k}$. By Lemma 15.7.7, for any τ,

$$P_\tau s = \sum_\sigma P_\tau P_\sigma = \sum_\sigma P_{\sigma\tau} = s, \qquad s P_\tau = \sum_\sigma P_\sigma P_\tau = \sum_\sigma P_{\tau\sigma} = s,$$

since $\sigma \mapsto \sigma\tau$ and $\sigma \mapsto \tau\sigma$ are bijections of S_k. The first identity says $\operatorname{im} s \subseteq (V^{\otimes k})^{S_k}$. The second gives $s(\mathrm{id} - P_\tau) = 0$, and the generators of C_k are exactly the elements $(\mathrm{id} - P_\sigma)(u_1 \otimes \cdots \otimes u_k)$, so $C_k \subseteq \ker s$.
By hypothesis $k! \cdot 1 \neq 0$ in F. Indeed $k! \cdot 1 = (1 \cdot 1)(2 \cdot 1) \cdots (k \cdot 1)$ in F, a product of k factors; if $\operatorname{char} F = 0$ none of them is 0, and if $\operatorname{char} F = p > k$ then $i \cdot 1 \neq 0$ for $1 \leq i \leq k < p$ by Definition 0.7.2. A field has no zero divisors, so the product is nonzero, hence invertible. Since $C_k \subseteq \ker s$, Theorem 4.4.10 gives a linear map $\theta\colon \operatorname{Sym}^k V \to (V^{\otimes k})^{S_k}$ with $\theta(q(x)) = \frac{1}{k!} s(x)$.
Now compute both composites. For $x \in V^{\otimes k}$, each $P_\sigma x$ differs from x by an element of C_k, so $q(s(x)) = \sum_\sigma q(P_\sigma x) = k!\, q(x)$, whence

$$\iota\theta(q(x)) = \tfrac{1}{k!} q(s(x)) = q(x),$$

and q is surjective, so $\iota\theta = \mathrm{id}$. Conversely, if x is symmetric then $s(x) = \sum_\sigma P_\sigma x = k!\, x$, so $\theta\iota(x) = \theta(q(x)) = \frac{1}{k!} k!\, x = x$. Hence ι is an isomorphism with inverse θ.

(c) Let $\operatorname{char} F = 2$, $k = 2$, $n \geq 2$, and put $t = v_1 \otimes v_2 + v_2 \otimes v_1$. It is symmetric, since the only nonidentity element of S_2 swaps the two summands, and it is nonzero, being a sum of two distinct basis vectors of $V^{\otimes 2}$ (Theorem 15.6.3). But

$$\iota(t) = v_1 v_2 + v_2 v_1 = 2\, v_1 v_2 = 0$$

because $1 + 1 = 0$ in F. So $\ker \iota \neq \{0\}$ and ι is not injective. By (a) the two spaces have the same finite dimension, so a surjective ι would be injective by Theorem 2.3.1; hence ι is not surjective either. This proves the theorem.

■

Part (b) is where the division lives, and part (c) is what it costs. The operator s still exists over $\mathbb{F}_2$; what fails is the *rescaling* $\frac{1}{k!} s$, and with it the splitting of the quotient.

Warning. **In characteristic $p \leq k$, "symmetric power" and "symmetric tensors" are different objects.** They have the same dimension by Theorem 15.7.8 (a), so they are abstractly isomorphic, but the *natural* map between them is not an isomorphism, and no natural one is available. So: do not define $\operatorname{Sym}^k V$ as $(V^{\otimes k})^{S_k}$, do not write $\frac{1}{k!} \sum_\sigma P_\sigma$ without saying that $k!$ is invertible, and do not transport a statement proved for one to the other. This is Chapter 14's discipline again, where alternating and skew-symmetric forms coincided only away from characteristic 2 (Theorem 14.3.3): the same clause, the same division, and the same smallest witness, $\mathbb{F}_2^2$.

Everything else in this section is characteristic-free. In particular Theorem 15.7.3, Theorem 15.7.6, and the dimension $\binom{n+k-1}{k}$ hold over every field, because the quotient construction never divides. The next section takes the same care with the same reward: the exterior power is built from a relation chosen so that no division is ever needed.

15.7.6 Exercises

15.7.6.1 A. Check your understanding

Exercise. A1

(a) State the universal property (S1), with every quantifier, and say which maps are required to factor.

(b) Which subspace of $V^{\otimes k}$ is divided out to build $\operatorname{Sym}^k V$? Write down its generators for $k = 3$.

(c) Let $\dim V = 4$. Compute $\dim \operatorname{Sym}^3 V$ and $\dim V^{\otimes 3}$.

(d) Determine whether $\operatorname{Sym}^2 V$ is a subspace of $V^{\otimes 2}$. Justify your answer.

(e) What hypothesis on F makes the symmetric tensors of degree k naturally isomorphic to $\operatorname{Sym}^k V$, and which step of the proof needs it?

Solution.

(a) For every vector space W over F and every **symmetric** k-linear map $f\colon V^k \to W$, there exists exactly one linear $\bar{f}\colon \operatorname{Sym}^k V \to W$ with $\bar{f} \circ \pi = f$. Only the symmetric maps factor; a nonsymmetric one cannot, since $\bar{f} \circ \pi$ is symmetric for every linear $\bar{f}$.

(b) The subspace C_k of Theorem 15.7.2, spanned by the differences $u_1 \otimes \cdots \otimes u_k - u_{\sigma(1)} \otimes \cdots \otimes u_{\sigma(k)}$. For $k = 3$ the generators are the six differences obtained from the six $\sigma \in S_3$, for all $u_1, u_2, u_3 \in V$.

(c) $\dim \operatorname{Sym}^3 V = \binom{4+3-1}{3} = \binom{6}{3} = 20$ and $\dim V^{\otimes 3} = 4^3 = 64$.

(d) No. It is a quotient of $V^{\otimes 2}$, and its elements are cosets of C_2, not tensors. (Over a field of characteristic 0 there is a subspace of $V^{\otimes 2}$ naturally isomorphic to it, namely the symmetric tensors, but the isomorphism is not an equality, and it does not exist in characteristic 2.)

(e) $\operatorname{char} F = 0$ or $\operatorname{char} F > k$, so that $k!$ is invertible in F (Theorem 15.7.8 (b)). The step that needs it is the construction of the inverse map θ, which divides the symmetrization $s = \sum_\sigma P_\sigma$ by $k!$.

15.7.6.2 B. Practice

Exercise. B1: Expanding symmetric products

Let $V = F^2$.

(a) Expand $(2\mathbf{e}_1 - \mathbf{e}_2)(\mathbf{e}_1 + 3\mathbf{e}_2)$ in the basis $(\mathbf{e}_1^2, \mathbf{e}_1\mathbf{e}_2, \mathbf{e}_2^2)$ of $\operatorname{Sym}^2 V$.

(b) Expand $(\mathbf{e}_1 + \mathbf{e}_2)(\mathbf{e}_1 - \mathbf{e}_2)\mathbf{e}_1$ in the basis of $\operatorname{Sym}^3 V$ given by Theorem 15.7.3.

Solution.

(a) Expanding one slot at a time and then using $\mathbf{e}_2\mathbf{e}_1 = \mathbf{e}_1\mathbf{e}_2$,

$$(2\mathbf{e}_1 - \mathbf{e}_2)(\mathbf{e}_1 + 3\mathbf{e}_2) = 2\mathbf{e}_1^2 + 6\mathbf{e}_1\mathbf{e}_2 - \mathbf{e}_1\mathbf{e}_2 - 3\mathbf{e}_2^2 = 2\mathbf{e}_1^2 + 5\mathbf{e}_1\mathbf{e}_2 - 3\mathbf{e}_2^2.$$

(b) The basis is $(\mathbf{e}_1^3,\ \mathbf{e}_1^2\mathbf{e}_2,\ \mathbf{e}_1\mathbf{e}_2^2,\ \mathbf{e}_2^3)$, of size $\binom{4}{3} = 4$. Expanding the first two slots,

$$(\mathbf{e}_1 + \mathbf{e}_2)(\mathbf{e}_1 - \mathbf{e}_2)\mathbf{e}_1 = \mathbf{e}_1^2\mathbf{e}_1 - \mathbf{e}_1\mathbf{e}_2\mathbf{e}_1 + \mathbf{e}_2\mathbf{e}_1\mathbf{e}_1 - \mathbf{e}_2^2\mathbf{e}_1,$$

and the two middle terms are equal after sorting, so they cancel:

$$(\mathbf{e}_1 + \mathbf{e}_2)(\mathbf{e}_1 - \mathbf{e}_2)\mathbf{e}_1 = \mathbf{e}_1^3 - \mathbf{e}_1\mathbf{e}_2^2.$$

In the polynomial language of Theorem 15.7.6 this is $(x+y)(x-y)x = x^3 - xy^2$.

Exercise. B2: Counting

(a) Compute $\dim \operatorname{Sym}^k V$ for $\dim V = 3$ and $k = 0, 1, 2, 3, 4$.

(b) For $\dim V = n$, compute $\dim \operatorname{Sym}^2 V$ in closed form and check it against the number of entries on or above the diagonal of an $n \times n$ matrix.

(c) Determine all pairs (n, k) with $k \geq 1$ for which $\dim \operatorname{Sym}^k V = \dim V^{\otimes k}$. Justify your answer.

Solution.

(a) $\binom{k+2}{k}$ for $n = 3$, giving $1, 3, 6, 10, 15$ for $k = 0, 1, 2, 3, 4$.

(b) $\binom{n+1}{2} = n(n+1)/2$. An $n \times n$ matrix has n diagonal entries and $(n^2 - n)/2$ entries strictly above the diagonal, and $n + (n^2 - n)/2 = n(n+1)/2$. The agreement is the remark after Theorem 15.7.6: a symmetric bilinear form is the same data as a functional on $\operatorname{Sym}^2 V$, and also the same data as a symmetric matrix.

(c) Exactly those with $n = 1$, or with $k = 1$. If $k = 1$ both sides are n. If $n = 1$ both sides are 1. Otherwise $n \geq 2$ and $k \geq 2$, and then $v_1 \otimes v_2 \neq v_2 \otimes v_1$ are two basis tensors with the same image, so $C_k \neq \{0\}$ and $\dim \operatorname{Sym}^k V = n^k - \dim C_k < n^k$ by Theorem 4.4.5.

Exercise. B3: A form factored through the symmetric square

Let β be a symmetric bilinear form on V. Prove that there is a unique linear functional $\bar{\beta} \in (\operatorname{Sym}^2 V)^*$ with $\bar{\beta}(uw) = \beta(u, w)$ for all $u, w \in V$, and compute $\bar{\beta}$ on the basis of Theorem 15.7.3 when $V = F^2$ and $\beta(\mathbf{x}, \mathbf{y}) = x_1y_1 - x_2y_2$.

Solution. β is a symmetric 2-linear map $V^2 \to F$ (Definition 15.7.1 with $k = 2$, $W = F$), so (S1) gives exactly one linear $\bar{\beta}\colon \operatorname{Sym}^2 V \to F$ with $\bar{\beta} \circ \pi = \beta$, which is the displayed identity.

For the given β: $\bar{\beta}(\mathbf{e}_1^2) = \beta(\mathbf{e}_1, \mathbf{e}_1) = 1$, $\bar{\beta}(\mathbf{e}_1\mathbf{e}_2) = \beta(\mathbf{e}_1, \mathbf{e}_2) = 0$, and $\bar{\beta}(\mathbf{e}_2^2) = \beta(\mathbf{e}_2, \mathbf{e}_2) = -1$. So in the coordinates given by the basis $(\mathbf{e}_1^2, \mathbf{e}_1\mathbf{e}_2, \mathbf{e}_2^2)$ the functional reads $(1, 0, -1)$, and under Theorem 15.7.6 it is the rule "read off the coefficients of x^2 and y^2 and subtract".

15.7.6.3 C. Going deeper

Exercise. C1: The characteristic-two witness in full

Let $F = \mathbb{F}_2$ and $V = \mathbb{F}_2^2$, and take $k = 2$.

(a) Write down a basis of $(V^{\otimes 2})^{S_2}$ and verify that its dimension is 3.

(b) Compute $\dim C_2$ and $\dim \operatorname{Sym}^2 V$, and identify $\ker \iota$ and $\operatorname{im} \iota$ for the map ι of Theorem 15.7.8.

(c) Are $(V^{\otimes 2})^{S_2}$ and $\operatorname{Sym}^2 V$ isomorphic as vector spaces? Explain why this does not repair the situation.

Solution.

(a) Here $S_2 = \{\mathrm{id}, \sigma\}$ with σ the swap, and P_σ exchanges $\mathbf{e}_1 \otimes \mathbf{e}_2$ with $\mathbf{e}_2 \otimes \mathbf{e}_1$ and fixes $\mathbf{e}_1 \otimes \mathbf{e}_1$ and $\mathbf{e}_2 \otimes \mathbf{e}_2$. So $x = a\,\mathbf{e}_1 \otimes \mathbf{e}_1 + b\,\mathbf{e}_1 \otimes \mathbf{e}_2 + c\,\mathbf{e}_2 \otimes \mathbf{e}_1 + d\,\mathbf{e}_2 \otimes \mathbf{e}_2$ is symmetric if and only if $b = c$. A basis is

$$\mathbf{e}_1 \otimes \mathbf{e}_1, \qquad \mathbf{e}_2 \otimes \mathbf{e}_2, \qquad \mathbf{e}_1 \otimes \mathbf{e}_2 + \mathbf{e}_2 \otimes \mathbf{e}_1,$$

the orbit sums of Theorem 15.7.8 (a), so the dimension is $3 = \binom{3}{2}$.

(b) C_2 is spanned by the elements $\mathbf{u} \otimes \mathbf{w} - \mathbf{w} \otimes \mathbf{u}$. Writing $\mathbf{u} = a\mathbf{e}_1 + b\mathbf{e}_2$ and $\mathbf{w} = c\mathbf{e}_1 + d\mathbf{e}_2$ and expanding both slots, the $\mathbf{e}_1 \otimes \mathbf{e}_1$ and $\mathbf{e}_2 \otimes \mathbf{e}_2$ terms cancel and

$$\mathbf{u} \otimes \mathbf{w} - \mathbf{w} \otimes \mathbf{u} = (ad - bc)(\mathbf{e}_1 \otimes \mathbf{e}_2 - \mathbf{e}_2 \otimes \mathbf{e}_1).$$

So C_2 is the line spanned by $\mathbf{e}_1 \otimes \mathbf{e}_2 + \mathbf{e}_2 \otimes \mathbf{e}_1$, the sign being irrelevant over $\mathbb{F}_2$, and $\dim C_2 = 1$ and $\dim \operatorname{Sym}^2 V = 4 - 1 = 3$ by Theorem 4.4.5. Now $\ker \iota = (V^{\otimes 2})^{S_2} \cap C_2 = C_2$, of dimension 1, and $\operatorname{im} \iota$ is therefore 2-dimensional, spanned by $\mathbf{e}_1^2$ and $\mathbf{e}_2^2$; it misses $\mathbf{e}_1\mathbf{e}_2$.

(c) Yes, they are isomorphic: both have dimension 3 over $\mathbb{F}_2$, and Theorem 2.5.4 applies. It does not repair anything, because the isomorphism is not ι and does not respect the structure maps: an isomorphism chosen by hand would not satisfy $\varphi(\mathbf{u} \otimes \mathbf{w}) = \mathbf{uw}$, so it could not be used to transport statements about products of vectors. A construction is only as good as its universal property, and $(V^{\otimes 2})^{S_2}$ has none here.

Exercise. C2: $\operatorname{Sym}(V)$ is the free commutative algebra

(a) Prove that the products $u_1 \cdots u_k$ with $u_i \in V$ span $\operatorname{Sym}^k V$, and deduce that an algebra homomorphism out of $\operatorname{Sym}(V)$ is determined by its restriction to V.

(b) Prove that for every **commutative** associative unital F-algebra A and every linear map $f\colon V \to A$ there is a unique algebra homomorphism $\bar{f}\colon \operatorname{Sym}(V) \to A$ with $\bar{f}|_V = f$.

(c) Explain in one sentence why (b) fails if the word "commutative" is deleted, and name the algebra that has the corresponding property for all A.

Solution.

(a) $V^{\otimes k}$ is spanned by its simple tensors (Theorem 15.6.2) and q is surjective and linear, so their images, the products $u_1 \cdots u_k$, span $\operatorname{Sym}^k V$. If φ, ψ are algebra homomorphisms agreeing on V, then $\varphi(u_1 \cdots u_k) = \varphi(u_1) \cdots \varphi(u_k) = \psi(u_1 \cdots u_k)$, and both send 1 to 1_A; since these elements span $\operatorname{Sym}(V)$ and both maps are linear, $\varphi = \psi$.

(b) For $k \geq 1$ the map $(u_1, \ldots, u_k) \mapsto f(u_1) \cdots f(u_k)$ is k-linear, as in the proof of Theo-

rem 15.6.6, and it is **symmetric** because A is commutative (the third example after Definition 15.7.1). By (S1) it induces a linear $\bar{f}_k\colon \operatorname{Sym}^k V \to A$; set $\bar{f}_0(a) = a1_A$ and $\bar{f} = \sum_k \bar{f}_k$ on the direct sum. Multiplicativity is checked on products of vectors, which span, exactly as in Theorem 15.6.6, and $\bar{f}|_V = f$. Uniqueness is (a).

(c) It fails because $\bar{f}$ would have to satisfy $f(u)f(w) = \bar{f}(uw) = \bar{f}(wu) = f(w)f(u)$, so the images of V would be forced to commute, which need not happen in a noncommutative A. The algebra with the property for **all** A is the tensor algebra $\mathrm{T}(V)$, by Theorem 15.6.6.

Exercise. C3: One variable

Let $\dim V = 1$.

(a) Prove that $C_k = \{0\}$ for every k, and deduce that the quotient map $\mathrm{T}(V) \to \operatorname{Sym}(V)$ assembled from the q's is an isomorphism of algebras.

(b) Hence deduce that $\mathrm{T}(V) \cong \operatorname{Sym}(V) \cong F[x]$, and reconcile this with the fact that $\mathrm{T}(V)$ is noncommutative for $\dim V \geq 2$.

Solution.

(a) Write $V = Fv$. By Theorem 15.6.3, $V^{\otimes k}$ is spanned by $v^{\otimes k}$. A generator of C_k is $u_1 \otimes \cdots \otimes u_k - u_{\sigma(1)} \otimes \cdots \otimes u_{\sigma(k)}$; writing $u_t = c_t v$ and expanding, both terms equal $c_1 c_2 \cdots c_k\, v^{\otimes k}$, since the scalars multiply in the same way whatever the order. So every generator is 0 and $C_k = \{0\}$. Hence each $q\colon V^{\otimes k} \to \operatorname{Sym}^k V$ is an isomorphism (Theorem 4.4.5, or directly: its kernel is C_k), and the assembled map is a bijective algebra homomorphism, so an isomorphism.

(b) By Example (The tensor algebra of a line), $\mathrm{T}(V) \cong F[x]$, so (a) gives $\operatorname{Sym}(V) \cong F[x]$ as well; the same conclusion comes from Theorem 15.7.6 with $n = 1$. There is no tension with noncommutativity: $\mathrm{T}(V)$ is commutative precisely when $\dim V \leq 1$, because noncommutativity needs two independent vectors to put in the two slots (the warning after Theorem 15.6.5). With one variable there is nothing to reorder, so imposing the symmetry relations imposes nothing.

15.8 Exterior Powers

The symmetric power made the slots of a tensor interchangeable. This section makes them *hostile*: we divide $V^{\otimes k}$ by a subspace chosen so that a tensor dies as soon as two of its slots hold the same vector. The result, the exterior power $\Lambda^k V$, has a basis indexed by **strictly increasing** index tuples, so $\dim \Lambda^k V = \binom{n}{k}$ and $\Lambda^k V = \{0\}$ once k exceeds n. That collapse is the whole reason the determinant exists, and Section 09 will read the determinant off the top exterior power.

One choice in the construction is delicate, and it is the choice Chapter 14 prepared us for. There are two natural ways to say "the value flips when two slots are exchanged", and they are not the same condition: Theorem 14.3.3 showed that *alternating* implies *skew-symmetric* over every field, while the converse needs $\operatorname{char} F \neq 2$. We must therefore decide which relation to divide by, and the answer is the one that works over every field.

This is Section 02's method for the third time: universal property, uniqueness, then a model built as a quotient of the tensor power. Throughout, F is a field, V is a finite-dimensional vector space over F with $n = \dim V$, and $k \geq 1$ unless stated otherwise. **No hypothesis on the characteristic is made anywhere in this section**, and the construction is arranged so that none is needed.

15.8.1 Alternating multilinear maps

Chapter 7 defined alternating k-linear **forms**, valued in F (Definition 7.2.2), because that is all the determinant needed. The target plays no role in the definition, so we extend it, as Section 01 extended multilinearity.

An alternating map dies the moment two of its arguments coincide.

Definition 15.8.1. Alternating Multilinear Map

Let W be a vector space over F. A k-linear map $f\colon V^k \to W$ (Definition 15.1.1) is **alternating** if

$$f(u_1, \dots, u_k) = 0 \qquad \text{whenever } u_i = u_j \text{ for some pair of positions } i \neq j.$$

The condition is on **some** pair of positions, adjacent or not, and it says nothing about inputs whose entries are all distinct. For $W = F$ this is Definition 7.2.2 verbatim. The examples of Chapter 7 are all available: the determinant $\det(\mathbf{a}_1, \dots, \mathbf{a}_n)$ as a function of the columns; the signed area $(\mathbf{u}, \mathbf{w}) \mapsto u_1 w_2 - u_2 w_1$ on F^2; any linear functional, vacuously, at $k = 1$. A nonexample by minimal change: the dot product on F^2 is bilinear but $\mathbf{e}_1 \cdot \mathbf{e}_1 = 1 \neq 0$, so the alternating clause fails at that one input while bilinearity survives.

Everything Chapter 7 deduced from the definition used only the vector-space structure of the target, so it carries over.

Lemma 15.8.1. Properties of Alternating Maps

Let $f\colon V^k \to W$ be alternating k-linear, let $u_1, \dots, u_k \in V$, and let $\sigma \in S_k$.

(a) Exchanging two arguments multiplies the value by -1; in particular f is skew-symmetric.

(b) $f(u_{\sigma(1)}, \dots, u_{\sigma(k)}) = \operatorname{sgn}(\sigma)\, f(u_1, \dots, u_k)$.

(c) If $(u_1, \dots, u_k)$ is linearly dependent, then $f(u_1, \dots, u_k) = 0$.

Proof.

(a) Fix positions $i < j$, freeze the other arguments, and let $g(x, y) \in W$ be the value of f with x in position i and y in position j. Then g is bilinear and $g(x, x) = 0$, so expanding

$$0 = g(u_i + u_j, u_i + u_j) = g(u_i, u_j) + g(u_j, u_i)$$

gives the swap rule. This is the computation of Theorem 7.2.1 (a), which used nothing about F as a target beyond its being a vector space.

(b) For $k = 1$ the only permutation is the identity and there is nothing to prove. For $k \geq 2$, use Theorem 0.10.4 to write $\sigma = \tau_1 \cdots \tau_r$ as a product of transpositions. Apply-

ing (a) once per transposition multiplies the value by $(-1)^r$, and $(-1)^r = \operatorname{sgn}(\sigma)$ by Theorem 7.3.1 and Corollary 7.3.2. This is Lemma 7.4.1, with the same proof.

(c) By the Linear Dependence Lemma (Theorem 1.4.1) some u_j lies in $\operatorname{span}(u_1, \dots, u_{j-1})$. Expanding f in slot j writes the value as a combination of values in which some u_m with $m < j$ sits both in slot m and in slot j, and each such value is 0. This is Theorem 7.2.1 (d). If $j = 1$ then $u_1 = 0$ and linearity in slot 1 gives 0 at once.

■

15.8.2 Which relation to divide by

Now the delicate point, stated before the definition so that the definition looks inevitable. We want a quotient of $V^{\otimes k}$ in which the image of $u_1 \otimes \cdots \otimes u_k$ vanishes exactly when it should. There are two candidate subspaces to divide by:

$$A_k = \operatorname{span}\{\, u_1 \otimes \cdots \otimes u_k \;:\; u_i = u_j \text{ for some } i \neq j \,\},$$
$$N_k = \operatorname{span}\{\, x + P_\tau x \;:\; x \text{ a simple tensor},\ \tau \text{ a transposition} \,\},$$

where P_τ permutes the slots as in Lemma 15.7.7. Dividing by A_k imposes "repeat a vector and you get zero"; dividing by N_k imposes "swap two slots and the sign flips". The two are not the same subspace.

Proposition 15.8.2. The Repetition Relations Are the Stronger Ones

For every field F and every $k \geq 2$ we have $N_k \subseteq A_k$. If $\operatorname{char} F \neq 2$, then $N_k = A_k$. If $\operatorname{char} F = 2$, $k = 2$, and $n \geq 1$, the inclusion is **strict**.

Idea. For the inclusion, feed the sum $u_i + u_j$ into **both** slots i and j: the resulting tensor is in A_k, and expanding it produces the two "square" terms, also in A_k, plus the two cross terms, whose sum is therefore in A_k too. That is Theorem 14.3.3 (a) transcribed into tensors. For the reverse, a tensor with a repeat is its own swap, so $x + P_\tau x = 2x$, and recovering x is a division by 2 — the same division, in the same place, as in Theorem 14.3.3 (b).

Proof. ($\subseteq$) Let $x = u_1 \otimes \cdots \otimes u_k$ be simple and τ the transposition of positions $i < j$. Let y be the tensor obtained from x by putting $u_i + u_j$ in **both** positions i and j, which lies in A_k because those two slots now agree. Expanding y by multilinearity in those two slots gives four terms:

$$y = [u_i, u_i] + [u_i, u_j] + [u_j, u_i] + [u_j, u_j],$$

where $[p, q]$ denotes x with p in position i and q in position j. The first and last terms lie in A_k, again by a repeat, and $[u_i, u_j] = x$ while $[u_j, u_i] = P_\tau x$. Hence $x + P_\tau x = y - [u_i, u_i] - [u_j, u_j] \in A_k$. Generators of N_k lie in A_k, so $N_k \subseteq A_k$.

($\supseteq$), assuming $\operatorname{char} F \neq 2$. Let $x = u_1 \otimes \cdots \otimes u_k$ with $u_i = u_j$, $i \neq j$, and let τ swap i and j. Then $P_\tau x = x$, so $x + P_\tau x = (1 + 1)x$ lies in N_k. Since $\operatorname{char} F \neq 2$, the scalar $1 + 1$ is invertible (Definition 0.7.2), so $x \in N_k$. Generators of A_k lie in N_k, so $A_k \subseteq N_k$.

Strictness in characteristic 2, $k = 2$. Suppose $\operatorname{char} F = 2$ and $n \geq 1$, and fix a basis $(v_1, \dots, v_n)$. Expand a generator of N_2 in the basis of Theorem 15.6.3: writing $u = \sum_i a_i v_i$

and $w = \sum_j b_j v_j$,

$$u \otimes w + w \otimes u = \sum_{i,j} (a_i b_j + a_j b_i)\, v_i \otimes v_j,$$

and the coefficient of a **diagonal** basis tensor $v_i \otimes v_i$ is $2a_i b_i = 0$. Every element of N_2 therefore has coefficient 0 on $v_1 \otimes v_1$. But $v_1 \otimes v_1$ is itself a generator of A_2, and it is a basis vector of $V^{\otimes 2}$, hence nonzero. So $v_1 \otimes v_1 \in A_2 \setminus N_2$ and the inclusion is strict. This proves the proposition.

■

So the two relations differ, and we must choose. **We divide by A_k.** Three reasons, and they are the same three that made Chapter 7 build the determinant on alternation rather than skewness:

- The quotient by A_k is the object whose maps are the **alternating** ones, and alternating is the condition the determinant satisfies over every field.
- A_k is the larger subspace, so the quotient by it is the smaller space; and by Proposition 15.8.2 the two quotients agree whenever $\operatorname{char} F \neq 2$. So nothing is lost outside characteristic 2, and something is gained inside it.
- The dimension $\binom{n}{k}$ comes out right over every field. Dividing by N_2 instead would give, for $V = \mathbb{F}_2^2$, a space of dimension 3 where $\binom{2}{2} = 1$ is wanted; Exercise C1 carries out that count.

15.8.3 The exterior power

The k-th exterior power is the space in which alternating k-linear maps become linear.

Definition 15.8.2. k-th Exterior Power

A **k-th exterior power of** V is a pair $(P, \wedge)$, where P is a vector space over F and $\wedge\colon V^k \to P$ is an **alternating** k-linear map, such that

(E1) (**universal property**) for **every** vector space W over F and **every alternating** k-linear map $f\colon V^k \to W$, there is a **unique** linear map $\bar{f}\colon P \to W$ with $\bar{f} \circ \wedge = f$.

We write $\Lambda^k V$ for P and

$$v_1 \wedge v_2 \wedge \cdots \wedge v_k \coloneqq \wedge(v_1, \ldots, v_k),$$

calling it a **wedge** of the list. By convention $\Lambda^0 V = F$.

Clause by clause, and the small words matter. The structure map $\wedge$ must itself be **alternating**, which is what makes $v \wedge v = 0$ a fact rather than a wish. Only **alternating** maps are required to factor, and that is forced: $\bar{f} \circ \wedge$ is alternating for every linear $\bar{f}$, so no space could linearize a nonalternating map. **Unique** does the usual two jobs: existence of $\bar{f}$ says $\Lambda^k V$ is large enough, uniqueness says it is no larger.

At $k = 1$ the condition is vacuous and (V, id_V) works, so $\Lambda^1 V = V$. At $k = 0$ the convention $\Lambda^0 V = F$ is the one that makes $\binom{n}{k}$ correct at $k = 0$, where it reads $\binom{n}{0} = 1$.

Theorem 15.8.3. Uniqueness of the Exterior Power

Let $(P, \wedge)$ and $(P', \wedge')$ both be k-th exterior powers of V. Then there is a unique linear isomorphism $\varphi\colon P \to P'$ with $\varphi \circ \wedge = \wedge'$.

Proof. The proof of Theorem 15.6.1, with "alternating k-linear" in place of "k-linear": (E1) for P applied to $\wedge'$ gives φ, (E1) for P' applied to $\wedge$ gives ψ, and $\psi\varphi$ is a linear self-map of P with $\psi\varphi\wedge = \wedge$, so the uniqueness clause of (E1) applied to $\wedge$ forces $\psi\varphi = \mathrm{id}_P$; symmetrically $\varphi\psi = \mathrm{id}_{P'}$. This proves the theorem. ∎

Theorem 15.8.4. Existence of the Exterior Power

Let $A_k \subseteq V^{\otimes k}$ be as above and let $q\colon V^{\otimes k} \to V^{\otimes k}/A_k$ be the quotient map (Definition 4.4.3). Then $V^{\otimes k}/A_k$, together with

$$v_1 \wedge \cdots \wedge v_k = q(v_1 \otimes \cdots \otimes v_k),$$

is a k-th exterior power of V.

Proof. A_k is a subspace, being a span, and $\wedge$ is k-linear, being μ_k followed by the linear map q. It is alternating: if $u_i = u_j$ with $i \neq j$, then $u_1 \otimes \cdots \otimes u_k$ is by definition a generator of A_k, so its image under q is 0.

Let $f\colon V^k \to W$ be alternating k-linear. By (T1) there is a unique linear $\tilde{f}\colon V^{\otimes k} \to W$ with $\tilde{f}(u_1 \otimes \cdots \otimes u_k) = f(u_1, \ldots, u_k)$. On a generator of A_k, that is, on a simple tensor with $u_i = u_j$ for some $i \neq j$, the value is $f(u_1, \ldots, u_k) = 0$ because f is alternating. A linear map vanishing on a spanning set of A_k vanishes on A_k, so $A_k \subseteq \ker \tilde{f}$, and Theorem 4.4.10 (a) gives exactly one linear $\bar{f}$ on the quotient with $\bar{f} \circ q = \tilde{f}$, whence $\bar{f} \circ \wedge = f$.

Uniqueness: q is surjective and $V^{\otimes k}$ is spanned by its simple tensors (Theorem 15.6.2), so the wedges $u_1 \wedge \cdots \wedge u_k$ span $V^{\otimes k}/A_k$, and two linear maps agreeing on them agree everywhere. This proves the theorem. ∎

From Lemma 15.8.1 applied to the alternating map $\wedge$ itself we get the rules used in every computation below, for all $u_i \in V$ and $\sigma \in S_k$:

$$u_{\sigma(1)} \wedge \cdots \wedge u_{\sigma(k)} = \operatorname{sgn}(\sigma)\; u_1 \wedge \cdots \wedge u_k,$$
$$u_1 \wedge \cdots \wedge u_k = 0 \quad \text{if the list is dependent.}$$

Four instances, smallest first, and one look-alike that is not one.

- **One slot.** $\Lambda^1 V = V$, and the wedge of a single vector is the vector itself. The alternating condition asks about pairs of positions, and with one position there are none.
- **The plane.** For $V = F^2$, expanding both slots and canceling the two repeats gives $\mathbf{u} \wedge \mathbf{w} = (u_1 w_2 - u_2 w_1)\, \mathbf{e}_1 \wedge \mathbf{e}_2$. The coefficient is the signed area of Chapter 7 §01; the wedge is that number, moved out of F and into a line inside $\Lambda^2 V$.
- **Polynomials.** On $V = F[x]_{\leq 2}$ with basis $(1, x, x^2)$, the same expansion gives $(1 + x) \wedge (x + x^2) = 1 \wedge x + 1 \wedge x^2 + x \wedge x^2$, the term $x \wedge x$ having died.

- **A degenerate case.** If $V = \{0\}$ then every wedge of $k \geq 1$ vectors is 0, so $\Lambda^k V = \{0\}$. It matters because it is the base of the count below: $\binom{0}{k} = 0$ for $k \geq 1$.
- **Not an example: the symmetric product.** In $\mathrm{Sym}^2 V$ the product uw is bilinear and satisfies $uw = wu$, but $e_1 e_1 \neq 0$ by Theorem 15.7.3, so the symmetric product map is **not** alternating. Exactly one clause differs between the two constructions, and it is this one.

Warning. **The relation is $v \wedge v = 0$, not $u \wedge w = -w \wedge u$.** The second follows from the first (Lemma 15.8.1 (a)), but not conversely: over $\mathbb{F}_2$ the two say different things, since $-1 = 1$ there makes "$u \wedge w = -w \wedge u$" mean "$u \wedge w = w \wedge u$", which is no condition at all on a repeated slot. This is exactly Theorem 14.3.3, transcribed from forms to tensors by Proposition 15.8.2, and it is why the construction above divides by A_k. A reader who remembers that theorem should recognize the choice as the only characteristic-free one.

Quick check. In $\Lambda^2 V$, simplify $(u + w) \wedge (u - w)$.

Solution. Expand both slots: $u \wedge u - u \wedge w + w \wedge u - w \wedge w$. The first and last terms are 0 because $\wedge$ is alternating, and $w \wedge u = -u \wedge w$ by Lemma 15.8.1 (a). So the value is $-2\, u \wedge w$. Over $\mathbb{F}_2$ that is 0, which is correct there and not a sign of anything going wrong: over $\mathbb{F}_2$, $u + w$ and $u - w$ are the same vector, so the wedge of the pair must vanish.

15.8.4 A basis, and the dimension

Sorting is again the move, but now it costs a sign, and tuples with a repeat are simply gone. What survives is the strictly increasing tuples.

Theorem 15.8.5. Basis of an Exterior Power

Let $\mathcal{B} = (v_1, \ldots, v_n)$ be a basis of V and $k \geq 1$. Then the wedges

$$v_{i_1} \wedge v_{i_2} \wedge \cdots \wedge v_{i_k}, \qquad 1 \leq i_1 < i_2 < \cdots < i_k \leq n,$$

one for each **strictly increasing** tuple, form a basis of $\Lambda^k V$. Consequently

$$\dim \Lambda^k V = \binom{n}{k}, \qquad \text{and} \qquad \Lambda^k V = \{0\} \ \text{ for } k > n,$$

over **every** field F.

Idea. Spanning: start from the basis of Theorem 15.6.3, throw away every tuple with a repeated index — those wedges are 0 — and sort the rest, paying a sign. Independence: one linear functional per increasing tuple I, and the functional must be **alternating**, so the natural candidate is a determinant: read the k arguments through the k dual functionals $\varphi_{i_1}, \ldots, \varphi_{i_k}$, assemble the $k \times k$ matrix of values, and take its determinant. Determinants are alternating in their columns by construction, which is exactly what is needed, and the value at the basis wedge indexed by J is $\det \mathbf{I}_k = 1$ if $I = J$ and 0 otherwise, because some row is then zero.

Proof. Spanning. By Theorem 15.6.3 the tensors $v_{j_1} \otimes \cdots \otimes v_{j_k}$, over all tuples J, form a basis of $V^{\otimes k}$, and q is surjective, so the wedges $v_{j_1} \wedge \cdots \wedge v_{j_k}$ span $\Lambda^k V$. If J has two equal entries, the corresponding wedge is 0 because $\wedge$ is alternating. If the entries of J are distinct, let $\sigma \in S_k$ sort them into increasing order; then by Lemma 15.8.1 (b) the wedge equals $\pm$ the wedge of the sorted tuple. So the increasing tuples already span, and if $k > n$ there are no such tuples at all, whence $\Lambda^k V = \{0\}$.

Independence. Let $(\varphi_1, \dots, \varphi_n)$ be the dual basis of $\mathcal{B}$ (Theorem 5.1.3). For an increasing tuple $I = (i_1 < \cdots < i_k)$ define

$$g_I(u_1, \dots, u_k) = \det\Big(\varphi_{i_a}(u_b)\Big)_{a,b=1}^{k} \in F,$$

the determinant of the $k \times k$ matrix whose (a, b) entry is $\varphi_{i_a}(u_b)$. Column b of that matrix depends linearly on u_b, because each φ_{i_a} is linear, and det is linear in each column and vanishes when two columns coincide (Theorem 7.4.3). Hence g_I is k-linear, and alternating: if $u_a = u_b$ with $a \neq b$, two columns of the matrix are equal and the determinant is 0.

By (E1), g_I induces a linear $\bar{g}_I \colon \Lambda^k V \to F$. Evaluate it at the wedge of an increasing tuple J: the matrix has (a, b) entry $\varphi_{i_a}(v_{j_b}) = \delta_{i_a j_b}$. If $I = J$ this matrix is $\mathbf{I}_k$ and the determinant is 1. If $I \neq J$, then since I and J each list k distinct indices and are not the same increasing tuple, some i_a is not among $j_1, \dots, j_k$; row a is then zero, so the rows are dependent, and by Theorem 7.4.5 together with Theorem 7.2.1 (d) the determinant is 0. Hence $\bar{g}_I(v_{j_1} \wedge \cdots \wedge v_{j_k}) = \delta_{IJ}$.

Now let $\sum_I a_I\, v_{i_1} \wedge \cdots \wedge v_{i_k} = 0$, the sum over increasing tuples. Applying $\bar{g}_J$ leaves $a_J = 0$, and J was arbitrary. So the list is independent, hence a basis.

The count. An increasing k-tuple in $\{1, \dots, n\}$ is the same thing as a k-element subset of $\{1, \dots, n\}$, listed in its only increasing order. There are $\binom{n}{k}$ such subsets, and none at all when $k > n$. No step divided by an integer, so the conclusion holds over every field. This proves the theorem.

■

Corollary 15.8.6. The Top Exterior Power Is a Line

If $\dim V = n \geq 1$, then $\dim \Lambda^n V = 1$, with basis the single wedge $v_1 \wedge \cdots \wedge v_n$ of any basis of V. For $k > n$, $\Lambda^k V = \{0\}$.

Proof. There is exactly one increasing n-tuple in $\{1, \dots, n\}$, namely $(1, 2, \dots, n)$, so Theorem 15.8.5 gives a basis with one member, and $\binom{n}{n} = 1$. The second statement is the last clause of that theorem.

■

That one-dimensionality is the source of the determinant, and Section 09 is about it. For now, notice the shape of the numbers: $\dim \Lambda^k V$ rises from 1 at $k = 0$ to a maximum in the middle and back to 1 at $k = n$, and then stops. Compare $\dim \operatorname{Sym}^k V = \binom{n+k-1}{k}$, which grows forever. The symmetric power forgets the order; the exterior power forgets the order **and** forbids repetition, and forbidding repetition is what makes the tower finite.

Remark. The two counts fit together at $k = 2$: $\binom{n+1}{2} + \binom{n}{2} = n^2 = \dim V^{\otimes 2}$ over every field. In characteristic $\neq 2$ this reflects an actual decomposition of $V^{\otimes 2}$ into symmetric and alternating tensors, in the manner of Theorem 14.3.4. Over $\mathbb{F}_2$ the identity between the numbers survives while the decomposition does not, exactly as in Theorem 15.7.8.

Warning. **$\Lambda^k V$ is a quotient of $V^{\otimes k}$, not a subspace.** Its elements are cosets of A_k, and $u \wedge w = -w \wedge u$ is an identity between cosets, not between tensors. It is also a mistake to replace it by a subspace of $V^{\otimes k}$ — say "the tensors that change sign under every swap of two slots". Over $\mathbb{F}_2$ that subspace is the space of symmetric tensors of Definition 15.7.5, because $-1 = 1$ there; for $V = \mathbb{F}_2^2$ and $k = 2$ it has dimension $\binom{3}{2} = 3$ by Theorem 15.7.8 (a), while $\Lambda^2 V$ has dimension $\binom{2}{2} = 1$.

Example. A wedge in $\Lambda^2(F^3)$

Let $V = F^3$, $\mathbf{u} = \mathbf{e}_1 + 2\mathbf{e}_2 + 3\mathbf{e}_3$, and $\mathbf{w} = 4\mathbf{e}_1 + 5\mathbf{e}_2 + 6\mathbf{e}_3$. Write out a basis of $\Lambda^2 V$ and expand $\mathbf{u} \wedge \mathbf{w}$ in it.

Solution. By Theorem 15.8.5 a basis is $(\mathbf{e}_1 \wedge \mathbf{e}_2,\ \mathbf{e}_1 \wedge \mathbf{e}_3,\ \mathbf{e}_2 \wedge \mathbf{e}_3)$, of size $\binom{3}{2} = 3$. Expanding both slots gives nine terms; the three with a repeated index vanish, and each remaining pair combines by $\mathbf{e}_j \wedge \mathbf{e}_i = -\mathbf{e}_i \wedge \mathbf{e}_j$. The coefficient of $\mathbf{e}_i \wedge \mathbf{e}_j$ is therefore $u_i w_j - u_j w_i$:

$$\begin{aligned} u \wedge w &= (1 \cdot 5 - 2 \cdot 4)\, \mathbf{e}_1 \wedge \mathbf{e}_2 + (1 \cdot 6 - 3 \cdot 4)\, \mathbf{e}_1 \wedge \mathbf{e}_3 \\ &\quad + (2 \cdot 6 - 3 \cdot 5)\, \mathbf{e}_2 \wedge \mathbf{e}_3 \\ &= -3\, \mathbf{e}_1 \wedge \mathbf{e}_2 - 6\, \mathbf{e}_1 \wedge \mathbf{e}_3 - 3\, \mathbf{e}_2 \wedge \mathbf{e}_3. \end{aligned}$$

The three coefficients are the 2×2 minors of the 3×2 matrix with columns $\mathbf{u}$ and $\mathbf{w}$. That is not an accident, and Section 09 proves it in general.

15.8.5 Independence detects wedges

The last result of the section is small to state and is the tool everything after it uses: **a wedge is nonzero exactly when its factors are independent.** One direction is Lemma 15.8.1 (c); the other needs Theorem 15.8.5, which is why it comes last.

Theorem 15.8.7. Wedges Detect Independence

Let $u_1, \dots, u_k \in V$ with $1 \le k \le n$. Then

$$u_1 \wedge u_2 \wedge \cdots \wedge u_k \neq 0$$
$$\text{if and only if } (u_1, \dots, u_k) \text{ is independent.}$$

For $k > n$ the wedge is always 0.

Idea. ($\Rightarrow$) is the contrapositive of "alternating maps kill dependent lists". ($\Leftarrow$) is the reason Theorem 15.8.5 was worth proving: an independent list extends to a basis of V, and that theorem then says its wedge **is** one of the basis vectors of $\Lambda^k V$, hence nonzero.

Proof. ($\Rightarrow$) We prove the contrapositive. If $(u_1, \dots, u_k)$ is dependent, then $u_1 \wedge \cdots \wedge u_k = 0$ by Lemma 15.8.1 (c), applied to the alternating map $\wedge$.

($\Leftarrow$) Suppose the list is independent. By the Basis Extension Theorem (Theorem 1.6.5) it extends to a basis $\mathcal{B} = (u_1, \dots, u_k, u_{k+1}, \dots, u_n)$ of V, which is possible because $k \leq n$. Apply Theorem 15.8.5 to $\mathcal{B}$: the wedges indexed by increasing k-tuples form a basis of $\Lambda^k V$, and $u_1 \wedge \cdots \wedge u_k$ is the one indexed by $(1, 2, \dots, k)$. A member of a basis is nonzero, so $u_1 \wedge \cdots \wedge u_k \neq 0$.

For $k > n$, $\Lambda^k V = \{0\}$ by Theorem 15.8.5, so every wedge vanishes; consistently, any $k > n$ vectors in V are dependent. This proves the theorem.

■

Remark. The theorem turns a question about a list into a question about a single vector, and that is its value. Section 09 uses it to prove that $\Lambda^n T$ is multiplication by $\det T$ and that T is invertible exactly when that scalar is nonzero; Section 10 uses it to decide which elements of $\Lambda^2 V$ are wedges at all. Note the hypothesis it does **not** need: no assumption on F, because neither half divided by anything.

15.8.6 Exercises

15.8.6.1 A. Check your understanding

Exercise. A1

(a) State the universal property (E1), with every quantifier, and say which maps are required to factor.

(b) Which subspace of $V^{\otimes k}$ is divided out, and why is the other candidate N_k rejected?

(c) Let $\dim V = 5$. Compute $\dim \Lambda^k V$ for $k = 0, 1, 2, 3, 4, 5, 6$.

(d) Determine whether $u \wedge w \wedge u$ can be nonzero. Justify your answer.

(e) State the test for $u_1 \wedge \cdots \wedge u_k \neq 0$, and say which hypothesis on the field it needs.

Solution.

(a) For every vector space W over F and every **alternating** k-linear map $f\colon V^k \to W$, there exists exactly one linear $\bar{f}\colon \Lambda^k V \to W$ with $\bar{f} \circ \wedge = f$. Only alternating maps factor; a nonalternating one cannot, since $\bar{f} \circ \wedge$ is alternating for every linear $\bar{f}$.

(b) The subspace A_k spanned by the simple tensors with a repeated entry (Theorem 15.8.4). The candidate N_k, spanned by the elements $x + P_\tau x$, is rejected because $N_k \subsetneq A_k$ in characteristic 2 (Proposition 15.8.2), so the quotient by N_k would have the wrong

dimension there and its maps would be the skew-symmetric ones rather than the alternating ones.

(c) $\binom{5}{k}$, giving $1, 5, 10, 10, 5, 1, 0$ for $k = 0, \dots, 6$.

(d) No: the list (u, w, u) repeats u in positions 1 and 3, so the wedge is 0 because $\wedge$ is alternating. (Equivalently, the list is dependent, so Theorem 15.8.7 applies.)

(e) $u_1 \wedge \cdots \wedge u_k \neq 0$ if and only if the list is linearly independent (Theorem 15.8.7). It needs no hypothesis on F.

15.8.6.2 B. Practice

Exercise. B1: Computing wedges

Let $V = F^3$ with $F = \mathbb{Q}$.

(a) Expand $(2\mathbf{e}_1 - \mathbf{e}_2 + \mathbf{e}_3) \wedge (\mathbf{e}_1 + \mathbf{e}_2 - \mathbf{e}_3)$ in the basis $(\mathbf{e}_1 \wedge \mathbf{e}_2, \mathbf{e}_1 \wedge \mathbf{e}_3, \mathbf{e}_2 \wedge \mathbf{e}_3)$.

(b) Expand $\mathbf{e}_1 \wedge (\mathbf{e}_1 + \mathbf{e}_2) \wedge (\mathbf{e}_1 + \mathbf{e}_2 + \mathbf{e}_3)$ in the basis of $\Lambda^3 V$.

Solution.

(a) With $\mathbf{u} = (2, -1, 1)$ and $\mathbf{w} = (1, 1, -1)$, the coefficient of $\mathbf{e}_i \wedge \mathbf{e}_j$ is $u_i w_j - u_j w_i$, as in Example (A wedge in $\Lambda^2(F^3)$):

$$2 \cdot 1 - (-1) \cdot 1 = 3, \qquad 2 \cdot (-1) - 1 \cdot 1 = -3,$$
$$(-1) \cdot (-1) - 1 \cdot 1 = 0,$$

so the wedge is $3\,\mathbf{e}_1 \wedge \mathbf{e}_2 - 3\,\mathbf{e}_1 \wedge \mathbf{e}_3$.

(b) By Corollary 15.8.6 the basis is the single wedge $\mathbf{e}_1 \wedge \mathbf{e}_2 \wedge \mathbf{e}_3$. Expand from the left: in the second slot, $\mathbf{e}_1 \wedge \mathbf{e}_1 \wedge \cdots = 0$, so only $\mathbf{e}_2$ survives; then in the third slot, both $\mathbf{e}_1$ and $\mathbf{e}_2$ already occur, so only $\mathbf{e}_3$ survives. Hence the wedge is $\mathbf{e}_1 \wedge \mathbf{e}_2 \wedge \mathbf{e}_3$. (Equivalently, the three vectors are the columns of a unit upper triangular matrix, of determinant 1; Exercise C3 explains why the answer had to be that determinant.)

Exercise. B2: Independence by wedging

Determine whether each list is linearly independent, by computing a wedge. Justify your answer.

(a) $((1, 2), (3, 6))$ in $\mathbb{Q}^2$.

(b) $((1, 0, 1), (0, 1, 1), (1, 1, 2))$ in $\mathbb{Q}^3$.

(c) $((1, 0, 0), (0, 1, 0), (0, 0, 1), (1, 1, 1))$ in $\mathbb{Q}^3$.

Solution.

(a) $(1, 2) \wedge (3, 6) = (1 \cdot 6 - 2 \cdot 3)\,\mathbf{e}_1 \wedge \mathbf{e}_2 = 0$. By Theorem 15.8.7 the list is dependent, as it visibly is: the second vector is three times the first.

(b) The minor computation $(1 \cdot 1 - 0 \cdot 0,\ 1 \cdot 1 - 1 \cdot 0,\ 0 \cdot 1 - 1 \cdot 1) = (1, 1, -1)$ gives, for the first two vectors, $\mathbf{e}_1 \wedge \mathbf{e}_2 + \mathbf{e}_1 \wedge \mathbf{e}_3 - \mathbf{e}_2 \wedge \mathbf{e}_3$. Wedging with the third vector $(1, 1, 2)$

and keeping only the terms with all three indices distinct,

$$(\mathbf{e}_1 \wedge \mathbf{e}_2) \wedge 2\mathbf{e}_3 + (\mathbf{e}_1 \wedge \mathbf{e}_3) \wedge \mathbf{e}_2 - (\mathbf{e}_2 \wedge \mathbf{e}_3) \wedge \mathbf{e}_1 = (2 - 1 - 1)\,\mathbf{e}_1 \wedge \mathbf{e}_2 \wedge \mathbf{e}_3 = 0,$$

using $\mathbf{e}_1 \wedge \mathbf{e}_3 \wedge \mathbf{e}_2 = -\mathbf{e}_1 \wedge \mathbf{e}_2 \wedge \mathbf{e}_3$ and $\mathbf{e}_2 \wedge \mathbf{e}_3 \wedge \mathbf{e}_1 = \mathbf{e}_1 \wedge \mathbf{e}_2 \wedge \mathbf{e}_3$ from Lemma 15.8.1 (b). So the list is dependent; indeed the third vector is the sum of the first two.

(c) Here $k = 4 > 3 = n$, so $\Lambda^4(\mathbb{Q}^3) = \{0\}$ by Theorem 15.8.5 and the wedge is 0 without any computation. The list is dependent, as any 4 vectors in a 3-dimensional space must be.

Exercise. B3: A functional built from the universal property

Let $V = F^3$ and let $f\colon V^2 \to F$ be $f(\mathbf{u}, \mathbf{w}) = u_1 w_3 - u_3 w_1$.

(a) Prove that f is alternating and bilinear, and deduce that there is a unique linear $\bar{f}\colon \Lambda^2 V \to F$ with $\bar{f}(\mathbf{u} \wedge \mathbf{w}) = f(\mathbf{u}, \mathbf{w})$.

(b) Compute $\bar{f}$ on the basis $(\mathbf{e}_1 \wedge \mathbf{e}_2, \mathbf{e}_1 \wedge \mathbf{e}_3, \mathbf{e}_2 \wedge \mathbf{e}_3)$, and evaluate $\bar{f}$ on the answer to Example (A wedge in $\Lambda^2(F^3)$).

Solution.

(a) Each slot is a linear functional with the other fixed, so f is bilinear, and $f(\mathbf{u}, \mathbf{u}) = u_1 u_3 - u_3 u_1 = 0$, so f is alternating (Definition 15.8.1 with $k = 2$). By (E1) there is exactly one linear $\bar{f}\colon \Lambda^2 V \to F$ with $\bar{f} \circ \wedge = f$.

(b) $\bar{f}(\mathbf{e}_1 \wedge \mathbf{e}_2) = f(\mathbf{e}_1, \mathbf{e}_2) = 0$; $\bar{f}(\mathbf{e}_1 \wedge \mathbf{e}_3) = 1$; $\bar{f}(\mathbf{e}_2 \wedge \mathbf{e}_3) = 0$. So $\bar{f}$ reads off the $\mathbf{e}_1 \wedge \mathbf{e}_3$ coordinate. On $\mathbf{u} \wedge \mathbf{w} = -3\mathbf{e}_1 \wedge \mathbf{e}_2 - 6\mathbf{e}_1 \wedge \mathbf{e}_3 - 3\mathbf{e}_2 \wedge \mathbf{e}_3$ it gives -6, which agrees with $f(\mathbf{u}, \mathbf{w}) = 1 \cdot 6 - 3 \cdot 4 = -6$.

15.8.6.3 C. Going deeper

Exercise. C1: The characteristic-two witness

Let $F = \mathbb{F}_2$, $V = \mathbb{F}_2^2$, and $k = 2$.

(a) List the elements of A_2 and of N_2, and compute both dimensions.

(b) Compute $\dim \Lambda^2 V$ from (a) and check it against Theorem 15.8.5.

(c) Compute $\dim(V^{\otimes 2}/N_2)$ and explain, in one sentence, what would go wrong if the exterior power had been defined as that quotient.

Solution.

(a) By Theorem 15.6.3, $V^{\otimes 2}$ has basis $\mathbf{e}_1 \otimes \mathbf{e}_1, \mathbf{e}_1 \otimes \mathbf{e}_2, \mathbf{e}_2 \otimes \mathbf{e}_1, \mathbf{e}_2 \otimes \mathbf{e}_2$. Now A_2 is spanned by the tensors $\mathbf{u} \otimes \mathbf{u}$, and the three choices $\mathbf{u} = \mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_1 + \mathbf{e}_2$ give

$$\mathbf{e}_1 \otimes \mathbf{e}_1, \quad \mathbf{e}_2 \otimes \mathbf{e}_2, \quad \mathbf{e}_1 \otimes \mathbf{e}_1 + \mathbf{e}_1 \otimes \mathbf{e}_2 + \mathbf{e}_2 \otimes \mathbf{e}_1 + \mathbf{e}_2 \otimes \mathbf{e}_2,$$

whose span also contains $t = \mathbf{e}_1 \otimes \mathbf{e}_2 + \mathbf{e}_2 \otimes \mathbf{e}_1$. Those three vectors $\mathbf{e}_1 \otimes \mathbf{e}_1, \mathbf{e}_2 \otimes \mathbf{e}_2, t$ are

independent, being distinct disjoint sums of basis vectors, and they span all three listed generators, so $\dim A_2 = 3$; there are no further generators, since $\mathbf{0}$ is the only remaining vector of $\mathbb{F}_2^2$. For N_2, the coefficient formula in the proof of Proposition 15.8.2 gives

$$\mathbf{u} \otimes \mathbf{w} + \mathbf{w} \otimes \mathbf{u} = (a_1 b_2 + a_2 b_1)\, t,$$

since the diagonal coefficients vanish and the two off-diagonal ones are equal. Hence $N_2 = \operatorname{span}(t)$ and $\dim N_2 = 1$.

(b) $\dim \Lambda^2 V = 4 - 3 = 1$ by Theorem 4.4.5, matching $\binom{2}{2} = 1$.

(c) $\dim(V^{\otimes 2}/N_2) = 4 - 1 = 3$. Defining $\Lambda^2 V$ as that quotient would make the image of $\mathbf{e}_1 \otimes \mathbf{e}_1$ nonzero, so $\mathbf{v} \wedge \mathbf{v} \neq 0$ for some $\mathbf{v}$, the structure map would be skew-symmetric but not alternating, and the dimension would be 3 instead of $\binom{2}{2} = 1$.

Exercise. C2: Alternating forms, recounted

(a) Prove that the alternating k-linear forms on V are in bijection with the linear functionals on $\Lambda^k V$, and that the bijection is an isomorphism of vector spaces.

(b) Deduce that the space of alternating k-linear forms on V has dimension $\binom{n}{k}$.

(c) Hence recover Corollary 7.4.4: the alternating n-linear forms on F^n form a line.

Solution.

(a) Write $\mathcal{A}$ for the space of alternating k-linear forms on V, a vector space by Theorem 7.2.3. The map $\Theta\colon (\Lambda^k V)^* \to \mathcal{A}$, $\Theta(\lambda) = \lambda \circ \wedge$, is well defined, since a composite of the alternating $\wedge$ with a linear map is alternating and k-linear, and it is linear in λ because composition with a fixed map is. It is injective: if $\lambda \circ \wedge = 0$ then λ vanishes on the wedges, which span $\Lambda^k V$ (Theorem 15.8.4), so $\lambda = 0$. It is surjective: given $f \in \mathcal{A}$, (E1) supplies $\bar{f}$ with $\bar{f} \circ \wedge = f$. So Θ is an isomorphism.

(b) By (a) and Corollary 5.1.4, $\dim \mathcal{A} = \dim(\Lambda^k V)^* = \dim \Lambda^k V = \binom{n}{k}$, using Theorem 15.8.5.

(c) Take $V = F^n$ and $k = n$: the dimension is $\binom{n}{n} = 1$. Since det is a nonzero alternating n-linear form (Theorem 7.4.3, with $\det \mathbf{I}_n = 1$), it spans that line, which is Corollary 7.4.4.

Exercise. C3: A wedge sees only the subspace

Let $(u_1, \dots, u_k)$ and $(w_1, \dots, w_k)$ be two lists in V with $w_b = \sum_{a=1}^{k} p_{ab}\, u_a$ for a matrix $\mathbf{P} = (p_{ab}) \in M_k(F)$.

(a) Prove that $w_1 \wedge \cdots \wedge w_k = (\det \mathbf{P})\; u_1 \wedge \cdots \wedge u_k$.

(b) Deduce that if $(u_1, \dots, u_k)$ and $(w_1, \dots, w_k)$ are two bases of the **same** k-dimensional subspace $U \subseteq V$, then their wedges differ by a nonzero scalar factor.

(c) Deduce that the line $\operatorname{span}(u_1 \wedge \cdots \wedge u_k) \subseteq \Lambda^k V$ depends only on U, not on the basis chosen for it.

Hint: for (a), expand every slot and use Lemma 15.8.1 (b) together with Definition 7.4.1.

Solution.

(a) Expanding slot by slot,

$$w_1 \wedge \cdots \wedge w_k = \sum_{a_1,\ldots,a_k} p_{a_1 1} \cdots p_{a_k k}\, u_{a_1} \wedge \cdots \wedge u_{a_k},$$

the sum over all k-tuples $(a_1, \ldots, a_k)$ from $\{1, \ldots, k\}$. A term with two equal indices has a repeated argument, so its wedge is 0. A term with all indices distinct has $(a_1, \ldots, a_k) = (\sigma(1), \ldots, \sigma(k))$ for a unique $\sigma \in S_k$, and then $u_{\sigma(1)} \wedge \cdots \wedge u_{\sigma(k)} = \operatorname{sgn}(\sigma)\, u_1 \wedge \cdots \wedge u_k$ by Lemma 15.8.1 (b). Hence

$$w_1 \wedge \cdots \wedge w_k = \Big(\sum_{\sigma \in S_k} \operatorname{sgn}(\sigma)\, p_{\sigma(1)1} \cdots p_{\sigma(k)k} \Big) u_1 \wedge \cdots \wedge u_k,$$

and the bracket is $\det \mathbf{P}$ by Definition 7.4.1.

(b) Both lists are bases of U, so each w_b is a combination of the u_a, giving $\mathbf{P}$ as above, and $\mathbf{P}$ is the change-of-coordinates matrix between two bases of U, hence invertible (Proposition 4.2.2). By Theorem 7.5.2, $\det \mathbf{P} \neq 0$, and (a) gives the claim. Both wedges are nonzero, by Theorem 15.8.7.

(c) By (b) the two wedges are nonzero scalar multiples of each other, so they span the same one-dimensional subspace of $\Lambda^k V$. Any basis of U therefore determines the same line, which consequently depends on U alone. (This line is the object that Section 09's coordinates — the $k \times k$ minors — describe, and it is how a k-dimensional subspace is turned into a single point.)

15.9 The Determinant as a Top Exterior Power

Chapter 7 built the determinant twice. First it proved that an alternating n-linear form on F^n is determined by its value at the standard basis (Theorem 7.4.2), which says *at most one* normalized such form exists; then it wrote the Leibniz sum down and checked the three rules (Theorem 7.4.3), which says *at least one* does. The number $\det \mathbf{A}$ of Definition 7.4.1 is that form's value. This section builds the determinant a third time, and the honest way to describe what it adds is to say up front what it does not: no new value, no faster computation, and no theorem Chapter 7 was missing. What it adds is an **object**. The exterior powers of the previous section turn $\det T$ from a number attached to a matrix into the operator that T induces on a one-dimensional space, and from that description multiplicativity stops being a computation and becomes the statement that composing two induced maps induces the composite.

Throughout, F is a field, V and W are finite-dimensional vector spaces over F, and $\Lambda^k V$, the wedge $v_1 \wedge \cdots \wedge v_k$, and the universal property for alternating maps are as in Definition 15.8.2. We use freely that the wedges $v_{i_1} \wedge \cdots \wedge v_{i_k}$ over strictly increasing index tuples form a basis of $\Lambda^k V$, so that $\dim \Lambda^k V = \binom{n}{k}$ when $\dim V = n$ (Theorem 15.8.5), and that a wedge of vectors is nonzero exactly when the list is linearly independent (Theorem 15.8.7).

15.9.1 Wedging with a fixed vector

Two facts from the previous section will be used on every page, and one small tool is still missing.

The first fact is the **reordering rule**: the wedge map $(v_1, \dots, v_k) \mapsto v_1 \wedge \cdots \wedge v_k$ is alternating (Definition 15.8.2), and an alternating multilinear map satisfies

$$f(v_{\sigma(1)}, \dots, v_{\sigma(k)}) = \operatorname{sgn}(\sigma)\, f(v_1, \dots, v_k)$$

for every $\sigma \in S_k$, by Lemma 15.8.1 (b). Reordering a wedge costs the sign of the permutation, and nothing else. The second is that $\Lambda^n V$ is one-dimensional when $\dim V = n \geq 1$ (Corollary 15.8.6); that is the whole engine of this section.

The missing tool is the ability to wedge a fixed vector onto a whole exterior power, not just onto a list of vectors. Section 10 builds the product on all of ΛV; here one vector at a time is enough.

Lemma 15.9.1. Wedging with a Fixed Vector

Let $v \in V$ and $k \geq 1$. There is a **unique** linear map

$$\varepsilon_v \colon \Lambda^k V \to \Lambda^{k+1} V, \qquad \varepsilon_v(w_1 \wedge \cdots \wedge w_k) = v \wedge w_1 \wedge \cdots \wedge w_k.$$

Proof. The map $V^k \to \Lambda^{k+1} V$ sending $(w_1, \dots, w_k)$ to $v \wedge w_1 \wedge \cdots \wedge w_k$ is k-linear, because the wedge of $k+1$ vectors is linear in each slot and the first slot is frozen at v; and it is alternating, because if $w_p = w_r$ with $p \neq r$ then the wedge of $k+1$ vectors has two equal arguments and is 0. By the universal property in Definition 15.8.2 it factors through $\Lambda^k V$, giving a linear ε_v with the stated values. Uniqueness holds because the wedges $w_1 \wedge \cdots \wedge w_k$ span $\Lambda^k V$ (Theorem 15.8.5), and a linear map is determined on a spanning set. ∎

15.9.2 Induced maps on exterior powers

A linear map should act on everything built from its domain. A map $T \colon V \to W$ sends a list $(v_1, \dots, v_k)$ to the list $(Tv_1, \dots, Tv_k)$, so it ought to send the wedge $v_1 \wedge \cdots \wedge v_k$ to $Tv_1 \wedge \cdots \wedge Tv_k$. The only question is whether that rule defines a map at all: an element of $\Lambda^k V$ can be written as a wedge in many ways, and an assignment made on wedges must be checked. The universal property answers it in two lines, and that is the point of having a universal property.

The k-th exterior power of a map wedges its values together.

Definition 15.9.1. Exterior Power of a Linear Map

Let $T \colon V \to W$ be linear and let $k \geq 1$. The **k-th exterior power** of T is the unique linear map

$$\Lambda^k T \colon \Lambda^k V \to \Lambda^k W$$

satisfying $\Lambda^k T(v_1 \wedge \cdots \wedge v_k) = Tv_1 \wedge \cdots \wedge Tv_k$ for all $v_1, \dots, v_k \in V$.

Well-definedness. The map $V^k \to \Lambda^k W$ sending $(v_1, \dots, v_k)$ to $Tv_1 \wedge \cdots \wedge Tv_k$ is k-linear, since T is linear and the wedge in W is linear in each slot; and it is alternating, since $v_p = v_r$ forces $Tv_p = Tv_r$ and a wedge with two equal arguments is 0. So Definition 15.8.2 produces exactly one linear map on $\Lambda^k V$ with the stated values, and the wedges span $\Lambda^k V$ (Theorem 15.8.5), so no other linear map has them.

Examples.

- $k = 1$. $\Lambda^1 V = V$ and $\Lambda^1 T = T$: the definition says nothing new in degree one.
- **The identity.** $\Lambda^k(\mathrm{id}_V) = \mathrm{id}$, since both sides send $v_1 \wedge \cdots \wedge v_k$ to itself.
- **A scalar.** For $T = c\,\mathrm{id}_V$, $\Lambda^k T(v_1 \wedge \cdots \wedge v_k) = c^k\, v_1 \wedge \cdots \wedge v_k$, so $\Lambda^k T = c^k\,\mathrm{id}$. One factor of c comes out of each of the k slots.
- **A degenerate case.** If $k > \dim V$ then $\Lambda^k V = \{0\}$ (Theorem 15.8.5) and $\Lambda^k T$ is the zero map between zero spaces. Worth recording so that statements below need no size caveat.

Nonexample by minimal change. Replace "alternating" by nothing: the rule $v_1 \otimes \cdots \otimes v_k \mapsto Tv_1 \wedge \cdots \wedge Tv_k$ on $V^{\otimes k}$ is perfectly well defined, but it is **not** the map above, because $\Lambda^k V$ is a quotient of $V^{\otimes k}$ and not a subspace of it. What makes Definition 15.9.1 legal is precisely that the rule kills the relations that were divided out.

The reason for having the notation is the next sentence.

Theorem 15.9.2. Exterior Powers Respect Composition

Let $T\colon U \to V$ and $S\colon V \to W$ be linear, and let $k \geq 1$. Then

$$\Lambda^k(ST) = (\Lambda^k S)(\Lambda^k T), \qquad \Lambda^k(\mathrm{id}_V) = \mathrm{id}_{\Lambda^k V}.$$

Proof. Both $\Lambda^k(ST)$ and $(\Lambda^k S)(\Lambda^k T)$ are linear maps $\Lambda^k U \to \Lambda^k W$, so it suffices to compare them on the wedges, which span $\Lambda^k U$ (Theorem 15.8.5). By Definition 15.9.1 applied three times,

$$(\Lambda^k S)(\Lambda^k T)(u_1 \wedge \cdots \wedge u_k) = STu_1 \wedge \cdots \wedge STu_k = \Lambda^k(ST)(u_1 \wedge \cdots \wedge u_k).$$

The second identity was checked above. This proves the theorem. ∎

Example. The second exterior power of a 3×3 matrix

Let $T\colon \mathbb{Q}^3 \to \mathbb{Q}^3$ have matrix

$$\mathbf{A} = \begin{pmatrix} 2 & 1 & 3 \\ 0 & 4 & 1 \\ 5 & 2 & 1 \end{pmatrix}$$

in the standard basis. Find the matrix of $\Lambda^2 T$ in the basis $(\mathbf{e}_1 \wedge \mathbf{e}_2,\ \mathbf{e}_1 \wedge \mathbf{e}_3,\ \mathbf{e}_2 \wedge \mathbf{e}_3)$ of $\Lambda^2 \mathbb{Q}^3$, and compute its determinant.

Solution. By Definition 15.9.1, $\Lambda^2 T(\mathbf{e}_i \wedge \mathbf{e}_j) = \mathbf{a}_i \wedge \mathbf{a}_j$, where $\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3$ are the columns of $\mathbf{A}$. Expand one column:

$$\begin{aligned} \mathbf{a}_1 \wedge \mathbf{a}_2 &= (2\mathbf{e}_1 + 5\mathbf{e}_3) \wedge (\mathbf{e}_1 + 4\mathbf{e}_2 + 2\mathbf{e}_3) \\ &= 8\,\mathbf{e}_1 \wedge \mathbf{e}_2 + 4\,\mathbf{e}_1 \wedge \mathbf{e}_3 + 5\,\mathbf{e}_3 \wedge \mathbf{e}_1 + 20\,\mathbf{e}_3 \wedge \mathbf{e}_2 \\ &= 8\,\mathbf{e}_1 \wedge \mathbf{e}_2 - \mathbf{e}_1 \wedge \mathbf{e}_3 - 20\,\mathbf{e}_2 \wedge \mathbf{e}_3, \end{aligned}$$

where the terms $2\,\mathbf{e}_1 \wedge \mathbf{e}_1$ and $10\,\mathbf{e}_3 \wedge \mathbf{e}_3$ were dropped in the second line because a wedge with a repeated vector is 0, and the third line used $\mathbf{e}_3 \wedge \mathbf{e}_1 = -\mathbf{e}_1 \wedge \mathbf{e}_3$ and $\mathbf{e}_3 \wedge \mathbf{e}_2 = -\mathbf{e}_2 \wedge \mathbf{e}_3$ (Lemma 15.8.1 (b)). The coefficients $8, -1, -20$ are the three 2×2 minors of the first two

columns of $\mathbf{A}$, on rows $\{1,2\},\{1,3\},\{2,3\}$ — a pattern the last part of this section explains. Doing the same for the other two column pairs gives

$$[\Lambda^2 T] = \begin{pmatrix} 8 & 2 & -11 \\ -1 & -13 & -5 \\ -20 & -5 & 2 \end{pmatrix},$$

whose determinant is 2601. Since $\det \mathbf{A} = -51$, this is $(\det \mathbf{A})^2$.

Quick check. In the example, why is $\Lambda^2 T$ **not** multiplication by a scalar, while $\Lambda^3 T$ is?

Solution. $\dim \Lambda^2 \mathbb{Q}^3 = \binom{3}{2} = 3$ by Theorem 15.8.5, and the displayed matrix is not a scalar multiple of $\mathbf{I}_3$, so $\Lambda^2 T$ is an operator on a three-dimensional space with no reason to be scalar. But $\dim \Lambda^3 \mathbb{Q}^3 = \binom{3}{3} = 1$, and every linear operator on a one-dimensional space is multiplication by a scalar. Being a line is what makes the top power special.

15.9.3 The top exterior power

That last observation is the whole theorem. When k equals $\dim V$, the space $\Lambda^k V$ is a line, an operator on a line is a scalar, and the scalar can be read off by feeding in a basis.

Theorem 15.9.3. The Determinant Is the Top Exterior Power

Let V be a vector space over F with $\dim V = n \geq 1$, and let $T \in \mathcal{L}(V)$. Then $\dim \Lambda^n V = 1$, and $\Lambda^n T$ is multiplication by $\det T$:

$$\Lambda^n T(\omega) = (\det T)\,\omega \qquad \text{for every } \omega \in \Lambda^n V.$$

Equivalently, for all $v_1, \ldots, v_n \in V$,

$$Tv_1 \wedge \cdots \wedge Tv_n = (\det T)\, v_1 \wedge \cdots \wedge v_n. \tag{15.9.1}$$

Idea. Three moves. ① $\Lambda^n V$ is one-dimensional, so $\Lambda^n T$ is multiplication by one unknown scalar c. ② To find c, evaluate on a basis wedge $v_1 \wedge \cdots \wedge v_n$ and expand each Tv_j in the basis. ③ The expansion produces one term for every function from slots to basis indices; the noninjective ones die because a wedge with a repeat is zero, and what survives is a sum over permutations with signs — the Leibniz formula, arriving on its own.

Proof. Fix a basis $\mathcal{B} = (v_1, \ldots, v_n)$ of V. By Corollary 15.8.6, $\dim \Lambda^n V = 1$ with basis the single wedge $\omega_0 = v_1 \wedge \cdots \wedge v_n$. Since $\Lambda^n T$ is a linear operator on a one-dimensional space, there is a unique $c \in F$ with $\Lambda^n T(\omega_0) = c\,\omega_0$, and then $\Lambda^n T(a\omega_0) = ac\,\omega_0$ for every $a \in F$, so $\Lambda^n T = c\,\mathrm{id}$ on all of $\Lambda^n V$.

It remains to show $c = \det T$. Let $\mathbf{A} = [T]_{\mathcal{B}}^{\mathcal{B}}$ have entries a_{ij}, so that $Tv_j = \sum_{i=1}^{n} a_{ij} v_i$ (Definition 4.1.1). The wedge of n vectors is linear in each slot, so expanding slot by slot gives one term for each function $\varphi\colon \{1,\dots,n\} \to \{1,\dots,n\}$:

$$Tv_1 \wedge \cdots \wedge Tv_n = \sum_{\varphi} a_{\varphi(1)1} \cdots a_{\varphi(n)n}\, v_{\varphi(1)} \wedge \cdots \wedge v_{\varphi(n)}.$$

If φ is not injective, two slots hold the same vector and the wedge is 0. A noninjective φ is possible only because the domain and codomain are finite of the same size, so the surviving φ are exactly the bijections, that is, the permutations $\sigma \in S_n$. For those, Lemma 15.8.1 (b) gives $v_{\sigma(1)} \wedge \cdots \wedge v_{\sigma(n)} = \operatorname{sgn}(\sigma)\, \omega_0$. Therefore

$$\Lambda^n T(\omega_0) = \Big(\sum_{\sigma \in S_n} \operatorname{sgn}(\sigma)\, a_{\sigma(1)1} \cdots a_{\sigma(n)n} \Big) \omega_0,$$

and the bracket is $\det \mathbf{A}$ by Definition 7.4.1, which is $\det T$ by Definition 7.5.1. So $c = \det T$.

For the last display, note that both sides of Eq. 15.9.1 are n-linear in $(v_1,\dots,v_n)$ and agree when the list is the basis $\mathcal{B}$; but in fact no extra argument is needed, since $v_1 \wedge \cdots \wedge v_n \in \Lambda^n V$ for **any** list, and the first display applies to it. This proves the theorem.

■

Warning. **Only the top power is a line.** For $1 \le k < n$ the operator $\Lambda^k T$ is not multiplication by a scalar in general: Example (The second exterior power of a 3×3 matrix) exhibits a $\Lambda^2 T$ with three different diagonal entries. The determinant appears because $\binom{n}{n} = 1$, not because exterior powers are scalars.

Warning. **A map between different spaces still has no determinant.** If $T\colon V \to W$ is linear with $\dim V = \dim W = n$, then $\Lambda^n T$ is a linear map between two **different** one-dimensional spaces, and there is no scalar attached to it until bases of $\Lambda^n V$ and $\Lambda^n W$ are chosen. Changing the basis of $\Lambda^n W$ alone rescales the number arbitrarily. This is the same obstruction Chapter 7 recorded after Definition 7.5.1, now visible as the failure of a line to have a canonical generator.

Quick check. Let T be the operator on F^3 with $T\mathbf{e}_1 = \mathbf{e}_2$, $T\mathbf{e}_2 = \mathbf{e}_3$, and $T\mathbf{e}_3 = \mathbf{e}_1$. Read $\det T$ off Eq. 15.9.1, without writing down a matrix.

Solution. $T\mathbf{e}_1 \wedge T\mathbf{e}_2 \wedge T\mathbf{e}_3 = \mathbf{e}_2 \wedge \mathbf{e}_3 \wedge \mathbf{e}_1$. The permutation carrying $(1,2,3)$ to $(2,3,1)$ is a 3-cycle, of sign $+1$ (Proposition 7.3.4), so by Lemma 15.8.1 (b) this equals $\mathbf{e}_1 \wedge \mathbf{e}_2 \wedge \mathbf{e}_3$. Comparing with Eq. 15.9.1 on the nonzero element $\mathbf{e}_1 \wedge \mathbf{e}_2 \wedge \mathbf{e}_3$ gives $\det T = 1$.

15.9.4 What the third construction adds

It is worth saying plainly, because the temptation to oversell a slick construction is real.

It adds no new value. The scalar produced in the proof is the Leibniz sum, arrived at by the same expansion that proved Theorem 7.4.2 in Chapter 7: expand multilinearly, discard the terms with a repeated basis vector, sort the survivors, and collect signs. Anyone who has read Chapter 7 §04 has read that computation. The exterior power did not compute the determinant; it stored the same computation in a different place.

It adds no faster method. Nothing here helps evaluate a determinant. Elimination remains the way to do that (Example (A 4×4 Determinant by Elimination)).

What it does add is that det **is a map, not just a number.** Chapter 7 defined $\det \mathbf{A}$ by a formula in the entries and then had to *prove* that operators inherit it: the formula depends on a basis, so Corollary 7.5.7 was needed before Definition 7.5.1 could be written down. Here $\Lambda^n T$ is defined with no basis in sight, and the scalar exists because $\Lambda^n V$ is a line. Basis-independence is not a theorem about the determinant; it is the absence of any choice to be independent of. One honest qualification: the route is not self-contained. Theorem 15.8.5, which is what makes $\Lambda^n V$ a line in the first place, proves independence with a $k \times k$ determinant, and the identification of the scalar with $\det T$ goes through Definition 7.5.1. Chapter 7 is not being replaced here; it is being re-read.

And from that, multiplicativity is free.

Corollary 15.9.4. Multiplicativity and Invertibility, from Functoriality

Let $\dim V = n \geq 1$ and $S, T \in \mathcal{L}(V)$.

(a) $\det(ST) = \det S \det T$, and $\det \mathrm{id}_V = 1$.

(b) T is invertible if and only if $\det T \neq 0$.

Proof.

(a) By Theorem 15.9.2, $\Lambda^n(ST) = (\Lambda^n S)(\Lambda^n T)$. By Theorem 15.9.3 all three maps are multiplications by scalars on the line $\Lambda^n V$, namely by $\det(ST)$, $\det S$, and $\det T$; and multiplication by $\det T$ followed by multiplication by $\det S$ is multiplication by $\det S \det T$. Two scalar multiplications on a nonzero space are equal only if the scalars are, so $\det(ST) = \det S \det T$. Likewise $\Lambda^n(\mathrm{id}_V) = \mathrm{id}$, so $\det \mathrm{id}_V = 1$.

(b) ($\Rightarrow$) If T is invertible, apply (a) to $TT^{-1} = \mathrm{id}_V$: then $\det T \cdot \det(T^{-1}) = 1$, so $\det T \neq 0$. ($\Leftarrow$) Suppose T is not invertible. Then T is not injective, so $\ker T \neq \{0\}$; choose $v_1 \in \ker T$ with $v_1 \neq 0$ and extend to a basis $(v_1, \ldots, v_n)$ of V (Theorem 1.6.5). Since $Tv_1 = 0$ and the wedge is linear in its first slot, $Tv_1 \wedge \cdots \wedge Tv_n = 0$. By Eq. 15.9.1 this is $(\det T)\, v_1 \wedge \cdots \wedge v_n$, and the wedge of a basis is nonzero (Theorem 15.8.7), so $\det T = 0$.

■

Part (b) is the promise Theorem 15.8.7 made: the test "is this list independent?" and the test "is this determinant nonzero?" are the same test, once the determinant is read as an induced map.

Compare the work this replaces. Chapter 7 proved Theorem 7.5.4 by fixing $\mathbf{A}$, checking that $\mathbf{B} \mapsto \det(\mathbf{AB})$ is alternating and n-linear in the columns of $\mathbf{B}$, and applying uniqueness. That proof is short and correct; what the present one shows is that the content is not really about determinants at all. Multiplicativity is the statement that Λ^n turns composition into composition

— and Λ^n does that for the same reason that Λ^k does it for every k, where no scalar is available and no multiplicativity statement can even be phrased.

Remark. What is **not** improved: Theorem 7.4.5. The identity $\det \mathbf{A}^\top = \det \mathbf{A}$ is a statement about rows versus columns, and Λ^n sees only the map T, which has no rows. Chapter 7's proof, reindexing the Leibniz sum by σ^{-1}, remains the natural one.

15.9.5 The coordinates of a wedge

The example above produced the minors of a matrix as the coefficients of a wedge, apparently by accident. It was not an accident, and the general statement is the bridge between this chapter and Chapter 7 §07.

Fix $n \geq 1$, write $[n] = \{1, \dots, n\}$ as Chapter 7 §07 does, and recall the notation of Definition 7.7.1: for $\mathbf{A} \in M_{n\times k}(F)$ and a set $I = \{i_1 < \cdots < i_k\} \subseteq [n]$ of size k, the submatrix $\mathbf{A}_{I,[k]} \in M_k(F)$ keeps the rows in I and all k columns, and its determinant is a $k \times k$ minor of $\mathbf{A}$. Write

$$e_I \coloneqq \mathbf{e}_{i_1} \wedge \mathbf{e}_{i_2} \wedge \cdots \wedge \mathbf{e}_{i_k} \in \Lambda^k F^n,$$

so that the e_I, over all k-element $I \subseteq [n]$, form the basis of Theorem 15.8.5.

Theorem 15.9.5. The Coordinates of a Wedge Are the Minors

Let $1 \leq k \leq n$, let $\mathbf{v}_1, \dots, \mathbf{v}_k \in F^n$, and let $\mathbf{A} \in M_{n\times k}(F)$ be the matrix with these vectors as its columns. Then

$$\mathbf{v}_1 \wedge \mathbf{v}_2 \wedge \cdots \wedge \mathbf{v}_k = \sum_I (\det \mathbf{A}_{I,[k]})\, e_I,$$

the sum running over all subsets $I \subseteq [n]$ with $|I| = k$. That is, the coordinate of $\mathbf{v}_1 \wedge \cdots \wedge \mathbf{v}_k$ at e_I is the $k \times k$ minor of $\mathbf{A}$ on the rows I.

Idea. The same expansion as in Theorem 15.9.3, but now the index functions run into a bigger set. An injective $\varphi\colon [k] \to [n]$ is a choice of image set I followed by an ordering of it, so the terms group into one bunch per I, and each bunch is a signed sum over S_k: a determinant of size k.

Proof. Write $\mathbf{A} = (a_{ij})$, so $\mathbf{v}_j = \sum_{i=1}^n a_{ij}\mathbf{e}_i$. Expanding the wedge in each of its k slots gives one term per function $\varphi\colon [k] \to [n]$:

$$\mathbf{v}_1 \wedge \cdots \wedge \mathbf{v}_k = \sum_{\varphi} a_{\varphi(1)1} \cdots a_{\varphi(k)k}\, \mathbf{e}_{\varphi(1)} \wedge \cdots \wedge \mathbf{e}_{\varphi(k)}.$$

A noninjective φ repeats a basis vector, so its wedge is 0 and its term drops out. An injective φ is determined by its image $I = \{i_1 < \cdots < i_k\}$ together with the permutation $\sigma \in S_k$ defined by $\varphi(j) = i_{\sigma(j)}$; conversely every pair (I, σ) arises from exactly one injective φ. For such a φ, Lemma 15.8.1 (b) gives

$$\mathbf{e}_{\varphi(1)} \wedge \cdots \wedge \mathbf{e}_{\varphi(k)} = \mathbf{e}_{i_{\sigma(1)}} \wedge \cdots \wedge \mathbf{e}_{i_{\sigma(k)}} = \operatorname{sgn}(\sigma)\, e_I.$$

Grouping the terms by I therefore gives

$$\mathbf{v}_1 \wedge \cdots \wedge \mathbf{v}_k = \sum_I \Big(\sum_{\sigma \in S_k} \operatorname{sgn}(\sigma)\, a_{i_{\sigma(1)}1} \cdots a_{i_{\sigma(k)}k} \Big) e_I .$$

By Definition 7.7.1 the (p,q)-entry of $\mathbf{A}_{I,[k]}$ is $a_{i_p q}$, so the inner sum is exactly the Leibniz sum of Definition 7.4.1 for the $k \times k$ matrix $\mathbf{A}_{I,[k]}$, that is, $\det \mathbf{A}_{I,[k]}$. This proves the theorem.

■

Example. A wedge of two vectors in four-dimensional space

In $\mathbb{Q}^4$, let $\mathbf{v}_1 = (1,0,2,1)$ and $\mathbf{v}_2 = (0,1,1,3)$. Write $\mathbf{v}_1 \wedge \mathbf{v}_2$ in the basis of $\Lambda^2 \mathbb{Q}^4$.

Solution. Let $\mathbf{A} \in M_{4\times 2}(\mathbb{Q})$ have these two vectors as its columns, so that $\mathbf{A}_{\{1,2\},[2]} = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$ and the other five 2×2 submatrices read off the same way. Their determinants are

$$\begin{aligned} \det \mathbf{A}_{\{1,2\}} &= 1, \quad \det \mathbf{A}_{\{1,3\}} = 1, \quad \det \mathbf{A}_{\{1,4\}} = 3, \\ \det \mathbf{A}_{\{2,3\}} &= -2, \quad \det \mathbf{A}_{\{2,4\}} = -1, \quad \det \mathbf{A}_{\{3,4\}} = 5, \end{aligned}$$

abbreviating $\mathbf{A}_{I,[2]}$ to $\mathbf{A}_I$. By Theorem 15.9.5,

$$\begin{aligned} \mathbf{v}_1 \wedge \mathbf{v}_2 = {} & e_{12} + e_{13} + 3\, e_{14} \\ & - 2\, e_{23} - e_{24} + 5\, e_{34}. \end{aligned}$$

As a check, some coordinate is nonzero, so $\mathbf{v}_1 \wedge \mathbf{v}_2 \neq 0$, so $(\mathbf{v}_1, \mathbf{v}_2)$ is linearly independent by Theorem 15.8.7 — which is visible directly, since neither vector is a multiple of the other.

That check generalizes at once, and what comes out is a theorem of Chapter 7.

Corollary 15.9.6. Independence via Minors

Let $\mathbf{v}_1, \ldots, \mathbf{v}_k \in F^n$ and let $\mathbf{A} \in M_{n\times k}(F)$ have these vectors as its columns. Then $(\mathbf{v}_1, \ldots, \mathbf{v}_k)$ is linearly independent if and only if some $k \times k$ minor of $\mathbf{A}$ is nonzero.

Proof. By Theorem 15.8.7 the list is independent exactly when $\mathbf{v}_1 \wedge \cdots \wedge \mathbf{v}_k \neq 0$. By Theorem 15.9.5 the coordinates of that wedge in the basis (e_I) of Theorem 15.8.5 are the $k \times k$ minors of $\mathbf{A}$, and a vector is 0 exactly when all of its coordinates in a basis vanish. (For $k > n$ both sides fail: $\Lambda^k F^n = \{0\}$ and $\mathbf{A}$ has no $k \times k$ submatrix.)

■

This is Theorem 7.7.1 in the case of full column rank, and the general case follows from it: $\operatorname{rank} \mathbf{A}$ is the largest number of independent columns, hence the largest k for which some k columns carry a nonzero $k \times k$ minor. Chapter 7 proved it with two careful steps involving pivot columns and row rank. The wedge proves it by naming the coordinates.

15.9.6 Laplace expansion

Cofactor expansion also comes out of the coordinate theorem, once the wedge of $n-1$ vectors is available. The route is: pull the chosen column out of the wedge, recognize what is left as an element of $\Lambda^{n-1}F^n$ whose coordinates are the $(n-1)\times(n-1)$ minors, and wedge $\mathbf{e}_i$ back on.

Corollary 15.9.7. Cofactor Expansion, from a Wedge

Let $n \geq 2$, let $\mathbf{A} \in M_n(F)$ with minors M_{ij} as in Definition 7.6.1, and fix $j \in [n]$. Then

$$\det \mathbf{A} = \sum_{i=1}^{n} (-1)^{i+j} a_{ij} M_{ij}.$$

Idea. Write $\det \mathbf{A}$ as the wedge of the columns. Linearity in the j-th slot splits it into n pieces, one per i, with $\mathbf{e}_i$ sitting in that slot. Move $\mathbf{e}_i$ to the front at a cost of $(-1)^{j-1}$; the remaining $n-1$ columns form a wedge whose coordinates are the minors of the matrix with column j deleted, and wedging $\mathbf{e}_i$ on kills every coordinate except the one whose index set misses i.

Proof. Let $\mathbf{a}_1, \dots, \mathbf{a}_n$ be the columns of $\mathbf{A}$ and $\omega_0 = \mathbf{e}_1 \wedge \cdots \wedge \mathbf{e}_n$. By Theorem 15.9.3 applied to the operator $\mathbf{x} \mapsto \mathbf{A}\mathbf{x}$, whose matrix in the standard basis is $\mathbf{A}$,

$$\mathbf{a}_1 \wedge \cdots \wedge \mathbf{a}_n = (\det \mathbf{A})\, \omega_0.$$

Substituting $\mathbf{a}_j = \sum_{i=1}^{n} a_{ij} \mathbf{e}_i$ and using linearity in slot j,

$$(\det \mathbf{A})\, \omega_0 = \sum_{i=1}^{n} a_{ij}\, \mathbf{a}_1 \wedge \cdots \wedge \underset{\text{slot } j}{\mathbf{e}_i} \wedge \cdots \wedge \mathbf{a}_n.$$

Fix i. Moving $\mathbf{e}_i$ from slot j to the front means $j-1$ exchanges of neighbors, so by Lemma 15.8.1 (b) the i-th wedge above equals $(-1)^{j-1}\, \varepsilon_{\mathbf{e}_i}(\eta)$, where $\eta = \mathbf{a}_1 \wedge \cdots \wedge \widehat{\mathbf{a}_j} \wedge \cdots \wedge \mathbf{a}_n \in \Lambda^{n-1}F^n$ omits the j-th column, and $\varepsilon_{\mathbf{e}_i}$ is the map of Lemma 15.9.1.

Let $\mathbf{B} \in M_{n\times(n-1)}(F)$ be $\mathbf{A}$ with column j deleted. By Theorem 15.9.5,

$$\eta = \sum_{l=1}^{n} (\det \mathbf{B}_{[n]\setminus\{l\},\,[n-1]})\, e_{[n]\setminus\{l\}},$$

since the $(n-1)$-element subsets of $[n]$ are exactly the complements of single elements. Deleting row l from $\mathbf{B}$ deletes row l and column j from $\mathbf{A}$, so $\det \mathbf{B}_{[n]\setminus\{l\},[n-1]} = M_{lj}$ by Definition 7.6.1. Now apply $\varepsilon_{\mathbf{e}_i}$. For $l \neq i$ the index set $[n]\setminus\{l\}$ contains i, so $\mathbf{e}_i \wedge e_{[n]\setminus\{l\}}$ repeats $\mathbf{e}_i$ and vanishes; for $l = i$, moving $\mathbf{e}_i$ into its place costs $i-1$ exchanges, so $\mathbf{e}_i \wedge e_{[n]\setminus\{i\}} = (-1)^{i-1}\omega_0$ by Lemma 15.8.1 (b). Hence $\varepsilon_{\mathbf{e}_i}(\eta) = (-1)^{i-1} M_{ij}\, \omega_0$, and

$$(\det \mathbf{A})\, \omega_0 = \sum_{i=1}^{n} a_{ij} (-1)^{j-1} (-1)^{i-1} M_{ij}\, \omega_0.$$

Since $(-1)^{i+j-2} = (-1)^{i+j}$ and $\omega_0 \neq 0$ (Theorem 15.8.7), comparing coefficients gives the stated formula.

■

This is Theorem 7.6.1 (a). The row version follows from it by Theorem 7.4.5, exactly as in Chapter 7. The gain is again structural rather than practical: the sign $(-1)^{i+j}$, which in Chapter 7 arrived from a careful count of row and column moves, is here the price of two reversals, $j-1$ in one direction and $i-1$ in the other.

15.9.7 Exercises

15.9.7.1 A. Check your understanding

Exercise. A1

(a) State the defining property of $\Lambda^k T$, and say why a rule given on wedges needs an argument before it defines a map.

(b) State Theorem 15.9.3.

(c) Name one thing this section's construction of the determinant adds to Chapter 7's, and one thing it does not.

(d) True or false: for every $T \in \mathcal{L}(V)$ and every k, the map $\Lambda^k T$ is multiplication by a scalar. Justify your answer.

(e) What are the coordinates of $\mathbf{v}_1 \wedge \cdots \wedge \mathbf{v}_k$ in $\Lambda^k F^n$, in the basis (e_I)?

Solution.

(a) $\Lambda^k T$ is the unique linear map $\Lambda^k V \to \Lambda^k W$ with $\Lambda^k T(v_1 \wedge \cdots \wedge v_k) = Tv_1 \wedge \cdots \wedge Tv_k$ (Definition 15.9.1). An element of $\Lambda^k V$ has many expressions as a sum of wedges, so a rule stated on wedges might assign two different values to one element; the universal property of Definition 15.8.2 removes the danger by producing the map from an alternating multilinear map upstairs.

(b) If $\dim V = n \geq 1$ and $T \in \mathcal{L}(V)$, then $\dim \Lambda^n V = 1$ and $\Lambda^n T(\omega) = (\det T)\omega$ for every $\omega \in \Lambda^n V$.

(c) It adds that the determinant is the induced map on a line, so that no basis and no similarity-invariance argument is needed to define $\det T$, and multiplicativity becomes Theorem 15.9.2. It does not add a new value, a new method of computation, or any improvement to $\det \mathbf{A}^\top = \det \mathbf{A}$.

(d) False. It is true when $k = n = \dim V$, and when $k > n$ for the trivial reason that $\Lambda^k V = \{0\}$. For $1 \leq k < n$ it usually fails: see Example (The second exterior power of a 3×3 matrix), where $\Lambda^2 T$ has matrix with rows $(8, 2, -11)$, $(-1, -13, -5)$, $(-20, -5, 2)$, not a scalar matrix.

(e) The $k \times k$ minors of the matrix whose columns are $\mathbf{v}_1, \ldots, \mathbf{v}_k$: the coordinate at e_I is $\det \mathbf{A}_{I,[k]}$ (Theorem 15.9.5).

15.9.7.2 B. Practice

Exercise. B1: A second exterior power

Let $T\colon \mathbb{Q}^3 \to \mathbb{Q}^3$ have standard matrix

$$\mathbf{A} = \begin{pmatrix} 1 & 2 & 0 \\ 0 & 1 & 3 \\ 2 & 0 & 1 \end{pmatrix}.$$

Find the matrix of $\Lambda^2 T$ in the basis $(\mathbf{e}_1 \wedge \mathbf{e}_2, \mathbf{e}_1 \wedge \mathbf{e}_3, \mathbf{e}_2 \wedge \mathbf{e}_3)$. Hence verify that $\det(\Lambda^2 T) = (\det T)^2$.

Solution. By Definition 15.9.1 and Theorem 15.9.5, the entry of $[\Lambda^2 T]$ in row I and column J is $\det \mathbf{A}_{I,J}$, with rows and columns indexed by $\{1,2\}, \{1,3\}, \{2,3\}$ in that order. Computing the nine 2×2 minors,

$$[\Lambda^2 T] = \begin{pmatrix} 1 & 3 & 6 \\ -4 & 1 & 2 \\ -2 & -6 & 1 \end{pmatrix}.$$

Its determinant is 169. Expanding $\det \mathbf{A}$ along the first column (Theorem 7.6.1) gives $1 \cdot (1 - 0) + 2 \cdot (6 - 0) = 13$, and $13^2 = 169$.

Exercise. B2: Independence by minors

In $\mathbb{Q}^4$, let $\mathbf{v}_1 = (1, 2, 0, 1)$, $\mathbf{v}_2 = (0, 1, 1, 1)$, and $\mathbf{v}_3 = (1, 0, -2, -1)$. Compute the coordinates of $\mathbf{v}_1 \wedge \mathbf{v}_2 \wedge \mathbf{v}_3$ in $\Lambda^3 \mathbb{Q}^4$, and hence determine whether the list is linearly independent. Justify your answer.

Solution. Let $\mathbf{A} \in M_{4\times 3}(\mathbb{Q})$ have these columns. By Theorem 15.9.5 the four coordinates are the 3×3 minors on the row sets $\{1,2,3\}, \{1,2,4\}, \{1,3,4\}, \{2,3,4\}$. Computing them,

$$\det \mathbf{A}_{\{1,2,3\}} = \det \begin{pmatrix} 1 & 0 & 1 \\ 2 & 1 & 0 \\ 0 & 1 & -2 \end{pmatrix} = 0,$$

and the same for the other three: $\det \mathbf{A}_{\{1,2,4\}} = \det \mathbf{A}_{\{1,3,4\}} = \det \mathbf{A}_{\{2,3,4\}} = 0$. So $\mathbf{v}_1 \wedge \mathbf{v}_2 \wedge \mathbf{v}_3 = 0$, and by Theorem 15.8.7 the list is **dependent**. Indeed $\mathbf{v}_3 = \mathbf{v}_1 - 2\mathbf{v}_2$.

Exercise. B3: Determinant of an inverse

Let $\dim V = n \geq 1$ and let $T \in \mathcal{L}(V)$ be invertible. Using Theorem 15.9.2 and Theorem 15.9.3, and no formula for the determinant, prove that $\Lambda^n T$ is invertible and that $\det(T^{-1}) = (\det T)^{-1}$.

Solution. By Theorem 15.9.2, $(\Lambda^n T)(\Lambda^n T^{-1}) = \Lambda^n(TT^{-1}) = \Lambda^n(\mathrm{id}_V) = \mathrm{id}$, and symmetrically in the other order. So $\Lambda^n T$ is invertible with inverse $\Lambda^n T^{-1}$. By Theorem 15.9.3 these two maps are multiplication by $\det T$ and by $\det(T^{-1})$ on the line $\Lambda^n V$, and their com-

posite is multiplication by $\det T \cdot \det(T^{-1})$, which equals id. Since $\Lambda^n V \neq \{0\}$, this forces $\det T \cdot \det(T^{-1}) = 1$. In particular $\det T \neq 0$, and $\det(T^{-1}) = (\det T)^{-1}$.

15.9.7.3 C. Going deeper

Exercise. C1: The trace of an exterior power

Let $\mathbf{A} \in M_n(F)$ and $1 \le k \le n$, and let $T\colon F^n \to F^n$ be $\mathbf{x} \mapsto \mathbf{A}\mathbf{x}$.

(a) Prove that the matrix of $\Lambda^k T$ in the basis (e_I) has (I,J)-entry $\det \mathbf{A}_{I,J}$.

(b) Deduce that $\operatorname{tr}(\Lambda^k T)$ is the sum of the $\binom{n}{k}$ **principal** $k \times k$ minors of $\mathbf{A}$, those with $I = J$.

(c) Compute $\operatorname{tr}(\Lambda^2 T)$ for the matrix $\mathbf{A}$ of Example (The second exterior power of a 3×3 matrix), and check it against the matrix found there.

Solution.

(a) Column $J = \{j_1 < \cdots < j_k\}$ of the matrix is the coordinate vector of $\Lambda^k T(e_J) = \mathbf{a}_{j_1} \wedge \cdots \wedge \mathbf{a}_{j_k}$, where $\mathbf{a}_j = T\mathbf{e}_j$ is the j-th column of $\mathbf{A}$ (Definition 15.9.1). By Theorem 15.9.5 applied to the $n \times k$ matrix with columns $\mathbf{a}_{j_1}, \dots, \mathbf{a}_{j_k}$, the coordinate at e_I is the determinant of its submatrix on rows I, which is $\det \mathbf{A}_{I,J}$ (Definition 7.7.1).

(b) The trace is the sum of the diagonal entries, that is, of the entries with $I = J$, and by (a) those are the principal minors $\det \mathbf{A}_{I,I}$.

(c) The three principal 2×2 minors of $\mathbf{A}$ are $\det\left(\begin{smallmatrix} 2 & 1 \\ 0 & 4 \end{smallmatrix}\right) = 8$, $\det\left(\begin{smallmatrix} 2 & 3 \\ 5 & 1 \end{smallmatrix}\right) = -13$, and $\det\left(\begin{smallmatrix} 4 & 1 \\ 2 & 1 \end{smallmatrix}\right) = 2$. Their sum is -3, which is the sum of the diagonal entries $8, -13, 2$ of the matrix computed in Example (The second exterior power of a 3×3 matrix).

Exercise. C2: Cauchy–Binet from a wedge

Let $m \le n$, let $\mathbf{A} \in M_{m \times n}(F)$ and $\mathbf{B} \in M_{n \times m}(F)$, and let $T_{\mathbf{A}}\colon F^n \to F^m$ and $T_{\mathbf{B}}\colon F^m \to F^n$ be the associated maps.

(a) Show that $\Lambda^m T_{\mathbf{B}}(\mathbf{e}_1 \wedge \cdots \wedge \mathbf{e}_m) = \sum_J (\det \mathbf{B}_{J,[m]})\, e_J$, the sum over m-element subsets $J \subseteq [n]$.

(b) Show that $\Lambda^m T_{\mathbf{A}}(e_J) = (\det \mathbf{A}_{[m],J})\, \mathbf{e}_1 \wedge \cdots \wedge \mathbf{e}_m$ for each such J.

(c) Hence deduce the Cauchy–Binet formula $\det(\mathbf{A}\mathbf{B}) = \sum_J \det \mathbf{A}_{[m],J} \det \mathbf{B}_{J,[m]}$ (Theorem 7.7.2).

Solution.

(a) $\Lambda^m T_{\mathbf{B}}(\mathbf{e}_1 \wedge \cdots \wedge \mathbf{e}_m) = \mathbf{B}\mathbf{e}_1 \wedge \cdots \wedge \mathbf{B}\mathbf{e}_m$ by Definition 15.9.1, and the columns of $\mathbf{B}$ are $\mathbf{B}\mathbf{e}_1, \dots, \mathbf{B}\mathbf{e}_m$. Apply Theorem 15.9.5 in F^n with $k = m$.

(b) Writing $J = \{j_1 < \cdots < j_m\}$, we get $\Lambda^m T_{\mathbf{A}}(e_J) = \mathbf{A}\mathbf{e}_{j_1} \wedge \cdots \wedge \mathbf{A}\mathbf{e}_{j_m}$, a wedge of m vectors of F^m. By Theorem 15.9.5 with $k = m = n$ there is only one index set, $[m]$ itself, and the coefficient is the determinant of the matrix with those columns, which is $\mathbf{A}_{[m],J}$.

(c) By Theorem 15.9.2, $\Lambda^m(T_{\mathbf{A}} T_{\mathbf{B}}) = (\Lambda^m T_{\mathbf{A}})(\Lambda^m T_{\mathbf{B}})$. The map $T_{\mathbf{A}} T_{\mathbf{B}}$ is the opera-

tor on F^m with matrix $\mathbf{AB}$, so by Theorem 15.9.3 the left side sends $\mathbf{e}_1 \wedge \cdots \wedge \mathbf{e}_m$ to $\det(\mathbf{AB})\,\mathbf{e}_1 \wedge \cdots \wedge \mathbf{e}_m$. By (a) and then (b), the right side sends it to $\sum_J \det \mathbf{B}_{J,[m]} \det \mathbf{A}_{[m],J}\,\mathbf{e}_1 \wedge \cdots \wedge \mathbf{e}_m$. Comparing coefficients on the basis vector $\mathbf{e}_1 \wedge \cdots \wedge \mathbf{e}_m \neq 0$ gives the formula.

Exercise. C3: Exterior powers of an injection

Let $T\colon V \to W$ be an **injective** linear map between finite-dimensional spaces and let $1 \leq k \leq \dim V$. Prove that $\Lambda^k T$ is injective. Is the analogous statement true with "injective" replaced by "surjective"?

Hint: choose a basis of V and use Theorem 15.8.5 in both spaces.

Solution. Let $(v_1, \ldots, v_m)$ be a basis of V. Since T is injective, $(Tv_1, \ldots, Tv_m)$ is linearly independent in W, so it extends to a basis $(Tv_1, \ldots, Tv_m, w_{m+1}, \ldots)$ of W (Theorem 1.6.5). By Theorem 15.8.5, the wedges $v_{i_1} \wedge \cdots \wedge v_{i_k}$ over increasing tuples form a basis of $\Lambda^k V$, and $\Lambda^k T$ sends each of them to $Tv_{i_1} \wedge \cdots \wedge Tv_{i_k}$, which is one of the basis vectors of $\Lambda^k W$ provided by the extended basis. So $\Lambda^k T$ carries a basis to a linearly independent list, hence is injective.

Yes for surjective as well, and more cheaply: if T is surjective then every $w_1 \wedge \cdots \wedge w_k$ equals $Tv_1 \wedge \cdots \wedge Tv_k$ for preimages v_i, so the image of $\Lambda^k T$ contains a spanning set of $\Lambda^k W$ (Theorem 15.8.5) and is therefore all of it.

15.10 The Exterior Algebra, and Clifford at Last

The exterior powers have so far been used one degree at a time, and the wedge of two of them has had nowhere to live: $\Lambda^2 V$ and $\Lambda^3 V$ are different spaces, and until now nothing said what $(u \wedge w) \wedge x$ means as an operation. Stacking them fixes that, exactly as stacking the tensor powers produced the tensor algebra (Definition 15.6.2) and stacking the symmetric powers produced the symmetric algebra (Definition 15.7.3). The result is a 2^n-dimensional algebra, and it is the last object this chapter needs. With it we can finally pay the debt Chapter 14 recorded in writing: a Clifford algebra of (V, q) exists, and its dimension is 2^n.

Throughout, F is a field and V is a vector space over F with $\dim V = n$. From the subsection on decomposability on, $\operatorname{char} F \neq 2$, because Definition 14.12.2 assumes it. We use the exterior powers of Definition 15.8.2 with their universal property for alternating maps, their bases of increasing wedges (Theorem 15.8.5), the test Theorem 15.8.7, and the reordering rule Lemma 15.8.1 (b) of the previous section.

15.10.1 Stacking the exterior powers

The pieces are already in place; only the multiplication is missing. The pattern is the one Proposition 15.7.4 used for symmetric powers: a product of a k-fold wedge and an l-fold wedge ought to be the $(k+l)$-fold wedge of all the vectors, and the only thing to check is that this depends on the two factors and not on the way each was written.

Proposition 15.10.1. The Wedge of Two Exterior Powers

Let $k, l \geq 1$. There is a **unique** bilinear map

$$\Lambda^k V \times \Lambda^l V \to \Lambda^{k+l} V, \qquad (\omega, \eta) \mapsto \omega \wedge \eta,$$

such that for all $u_1, \ldots, u_k, w_1, \ldots, w_l \in V$,

$$(u_1 \wedge \cdots \wedge u_k) \wedge (w_1 \wedge \cdots \wedge w_l) = u_1 \wedge \cdots \wedge u_k \wedge w_1 \wedge \cdots \wedge w_l.$$

Idea. Freeze the second factor's vectors and use the universal property once, in the first k slots; that turns the left-hand factor into an element of $\Lambda^k V$. Then let the second factor's vectors vary: what we have built depends on them l-linearly, and vanishes as soon as two of them coincide, with values in a space of linear maps, so the universal property applies a second time.

Proof. Fix $w_1, \ldots, w_l \in V$. The map $V^k \to \Lambda^{k+l} V$ given by

$$(u_1, \ldots, u_k) \mapsto u_1 \wedge \cdots \wedge u_k \wedge w_1 \wedge \cdots \wedge w_l$$

is k-linear, since the wedge of $k + l$ vectors is linear in each slot and the last l are frozen, and alternating, since a repeat among the first k arguments is a repeat among all $k + l$. By Definition 15.8.2 it induces a unique linear map

$$R(w_1, \ldots, w_l) \colon \Lambda^k V \to \Lambda^{k+l} V$$

with $R(w_1, \ldots, w_l)(u_1 \wedge \cdots \wedge u_k)$ equal to the displayed wedge.

Now let the w's vary. The assignment $(w_1, \ldots, w_l) \mapsto R(w_1, \ldots, w_l)$, with values in the vector space $\mathcal{L}(\Lambda^k V, \Lambda^{k+l} V)$, is l-linear and alternating: two linear maps on $\Lambda^k V$ are equal as soon as they agree on the wedges $u_1 \wedge \cdots \wedge u_k$, which span $\Lambda^k V$ (Theorem 15.8.5), and on those the required identities are the linearity and the alternation of the wedge of $k + l$ vectors in its last l slots. So Definition 15.8.2 applies again and gives a unique linear map

$$\Lambda^l V \to \mathcal{L}(\Lambda^k V, \Lambda^{k+l} V), \qquad w_1 \wedge \cdots \wedge w_l \mapsto R(w_1, \ldots, w_l).$$

Reading a linear map into a space of linear maps as a bilinear map gives the required product, with the stated values. It is unique because the wedges span both factors and a bilinear map is determined on a pair of spanning sets. This proves the proposition.

■

With the product in hand, the algebra assembles itself.

The exterior algebra is all the exterior powers at once, multiplied by wedging.

Definition 15.10.1. Exterior Algebra

Let V be a vector space over F, with $\Lambda^0 V = F$ and $\Lambda^1 V = V$ as in Definition 15.8.2. The

exterior algebra, or **Grassmann algebra**, of V is the vector space

$$\Lambda V := \bigoplus_{k \geq 0} \Lambda^k V,$$

whose elements are the sequences $x = (x_0, x_1, \dots)$ with $x_k \in \Lambda^k V$ and only finitely many x_k nonzero, with componentwise operations, and whose multiplication is

$$(x \wedge y)_m = \sum_{k+l=m} x_k \wedge y_l,$$

where $\wedge$ on the right is Proposition 15.10.1 for $k, l \geq 1$, and is scalar multiplication when $k = 0$ or $l = 0$. We identify each $\Lambda^k V$ with the sequences supported in position k, so that F and V are subspaces of ΛV, and call an element of $\Lambda^k V \subseteq \Lambda V$ **homogeneous of degree** k.

In words: an element of ΛV is a finite sum $x_0 + x_1 + \cdots + x_N$ of pieces of different degrees, the scalars sit inside as the degree-0 piece, and **the product adds degrees**. The finiteness clause is automatic here, since $\Lambda^k V = \{0\}$ once $k > n$ (Theorem 15.8.5).

Theorem 15.10.2. Structure of the Exterior Algebra

Let $\dim V = n$ with basis $(e_1, \dots, e_n)$. For a subset $S = \{i_1 < \cdots < i_k\} \subseteq [n]$ write $e_S = e_{i_1} \wedge \cdots \wedge e_{i_k}$, with $e_\emptyset = 1$. Then

(a) ΛV is an associative unital F-algebra with identity $1 \in \Lambda^0 V = F$, and $\Lambda^k V \wedge \Lambda^l V \subseteq \Lambda^{k+l} V$.

(b) The 2^n elements e_S, one for each subset $S \subseteq [n]$, form a basis of ΛV. In particular $\dim_F \Lambda V = 2^n$.

(c) ΛV is generated as an algebra by 1 and V.

Proof.

(a) Bilinearity of the product holds componentwise, by Proposition 15.10.1. For associativity, both $(x \wedge y) \wedge z$ and $x \wedge (y \wedge z)$ are trilinear in (x, y, z), so it is enough to compare them on homogeneous elements, and then on the wedges of vectors, which span each $\Lambda^k V$ for $k \geq 1$ (Theorem 15.8.5) while $\Lambda^0 V$ is spanned by 1. On those both sides are the single wedge of all the listed vectors in order, by the defining identity of Proposition 15.10.1 applied twice. The scalar 1 acts on each $\Lambda^k V$ as multiplication by 1, hence as the identity. The grading statement is built into the definition of the product.

(b) By Theorem 15.8.5 the e_S with $|S| = k$ form a basis of $\Lambda^k V$, and $\Lambda^k V = \{0\}$ for $k > n$. A basis of a direct sum is the union of bases of the summands, so the e_S over all $S \subseteq [n]$ form a basis of ΛV. Their number is $\sum_{k=0}^{n} \binom{n}{k} = 2^n$, by the binomial theorem applied to $(1 + 1)^n$.

(c) Each e_S is a product of elements of V by the defining identity of the product, and the e_S span by (b).

■

The one structural fact that is not inherited from the tensor algebra is how the order of a product matters.

Theorem 15.10.3. Graded Commutativity

Let $\omega \in \Lambda^k V$ and $\eta \in \Lambda^l V$. Then

$$\omega \wedge \eta = (-1)^{kl}\, \eta \wedge \omega.$$

In particular $\omega \wedge \omega = 0$ whenever k is **odd**, over every field.

Idea. Count the swaps. To turn $w_1 \wedge \cdots \wedge w_l \wedge u_1 \wedge \cdots \wedge u_k$ into $u_1 \wedge \cdots \wedge u_k \wedge w_1 \wedge \cdots \wedge w_l$, each of the k vectors u_i must cross each of the l vectors w_j, and every crossing costs a sign.

Proof. Both sides are bilinear in (ω, η), so by Theorem 15.8.5 it suffices to take $\omega = u_1 \wedge \cdots \wedge u_k$ and $\eta = w_1 \wedge \cdots \wedge w_l$, and the cases $k = 0$ or $l = 0$ are scalar multiplication and are clear. The permutation of $\{1, \ldots, k+l\}$ carrying the list $(w_1, \ldots, w_l, u_1, \ldots, u_k)$ to $(u_1, \ldots, u_k, w_1, \ldots, w_l)$ has exactly kl inversions, one for each pair consisting of a u and a w, so its sign is $(-1)^{kl}$ (Definition 7.3.2). By Lemma 15.8.1 (b) the two wedges differ by that factor, which is the claim.

For the last sentence let k be odd and $\omega \in \Lambda^k V$. Write $\omega = \sum_S c_S\, e_S$ in the basis of Theorem 15.10.2, the sum over the S with $|S| = k$. Then

$$\omega \wedge \omega = \sum_S \sum_T c_S c_T\, e_S \wedge e_T.$$

A term with $S = T$ has a repeated vector, so $e_S \wedge e_S = 0$. A pair $S \neq T$ occurs twice, and the two occurrences are $c_S c_T\, e_S \wedge e_T$ and $c_T c_S\, e_T \wedge e_S = -c_S c_T\, e_S \wedge e_T$, since $(-1)^{kl} = (-1)^{k^2} = -1$ for odd k. So the two cancel, and nothing divides by 2 anywhere.

■

Warning. **Graded commutative is not commutative.** In ΛF^2 we have $\mathbf{e}_1 \wedge \mathbf{e}_2 = -\mathbf{e}_2 \wedge \mathbf{e}_1$, and these are different elements unless $\operatorname{char} F = 2$. What is true in all degrees is only the signed rule. And the sign is $+1$ as soon as one of the degrees is even, so an element of $\Lambda^2 V$ commutes with everything, while two elements of $\Lambda^1 V = V$ anticommute.

Warning. $\omega \wedge \omega$ **need not be zero.** The identity $v \wedge v = 0$ is a statement about degree 1. For $\omega = \mathbf{e}_1 \wedge \mathbf{e}_2 + \mathbf{e}_3 \wedge \mathbf{e}_4$ in ΛF^4, a direct expansion gives $\omega \wedge \omega = 2\, \mathbf{e}_1 \wedge \mathbf{e}_2 \wedge \mathbf{e}_3 \wedge \mathbf{e}_4$, which is nonzero whenever $\operatorname{char} F \neq 2$. Squaring to zero is the privilege of odd degree.

Example. The exterior algebra of a three-dimensional space

Describe ΛF^3 completely, and compute $(\mathbf{e}_1 + \mathbf{e}_2 \wedge \mathbf{e}_3) \wedge (\mathbf{e}_2 + \mathbf{e}_1 \wedge \mathbf{e}_3)$.

Solution. By Theorem 15.10.2 it has dimension $2^3 = 8$, with basis

$$1; \quad \mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3; \quad e_{12}, e_{13}, e_{23}; \quad e_{123},$$

writing e_{ij} for $\mathbf{e}_i \wedge \mathbf{e}_j$ and e_{123} for $\mathbf{e}_1 \wedge \mathbf{e}_2 \wedge \mathbf{e}_3$. The degrees have dimensions $1, 3, 3, 1$. Every product of two elements of degree ≥ 2 lands in degree ≥ 4, hence is 0.

For the computation, expand by bilinearity into four products:

$$\begin{aligned} \mathbf{e}_1 \wedge \mathbf{e}_2 &= e_{12}, \\ \mathbf{e}_1 \wedge (\mathbf{e}_1 \wedge \mathbf{e}_3) &= 0 \quad (\text{repeated } \mathbf{e}_1), \\ (\mathbf{e}_2 \wedge \mathbf{e}_3) \wedge \mathbf{e}_2 &= 0 \quad (\text{repeated } \mathbf{e}_2), \\ (\mathbf{e}_2 \wedge \mathbf{e}_3) \wedge (\mathbf{e}_1 \wedge \mathbf{e}_3) &= 0 \quad (\text{degree } 4). \end{aligned}$$

So the product is e_{12}, while the reversed product is $-e_{12}$ plus the same three zeros.

Quick check. In ΛF^4, what is $(\mathbf{e}_1 \wedge \mathbf{e}_2) \wedge (\mathbf{e}_2 \wedge \mathbf{e}_3)$, and what is $(\mathbf{e}_1 \wedge \mathbf{e}_2) \wedge (\mathbf{e}_3 \wedge \mathbf{e}_4)$?

Solution. The first is $\mathbf{e}_1 \wedge \mathbf{e}_2 \wedge \mathbf{e}_2 \wedge \mathbf{e}_3 = 0$, since $\mathbf{e}_2$ is repeated. The second is $\mathbf{e}_1 \wedge \mathbf{e}_2 \wedge \mathbf{e}_3 \wedge \mathbf{e}_4$, a basis vector of $\Lambda^4 F^4$, so it is nonzero. Wedging two 2-vectors gives zero exactly when the four vectors involved are dependent.

15.10.2 Decomposable elements

Section 2 warned that not every element of $V \otimes W$ is a simple tensor $v \otimes w$, and proved it with $\mathbf{e}_1 \otimes \mathbf{e}_1 + \mathbf{e}_2 \otimes \mathbf{e}_2$ (Example (A tensor that is not simple)). The same thing happens one floor up, and for the same reason: the simple objects form a low-dimensional cone inside a large space, and a sum of two of them usually escapes it.

Definition 15.10.2. Decomposable Element

Let $k \geq 1$. An element $\omega \in \Lambda^k V$ is **decomposable**, or a **k-blade**, if there exist $v_1, \ldots, v_k \in V$ with

$$\omega = v_1 \wedge v_2 \wedge \cdots \wedge v_k.$$

Examples. Every element of $\Lambda^1 V = V$ is decomposable, with $k = 1$; so is 0, being $v \wedge v \wedge v_3 \wedge \cdots \wedge v_k$; and so is every element of the line $\Lambda^n V$, since $c\,e_1 \wedge \cdots \wedge e_n = (ce_1) \wedge e_2 \wedge \cdots \wedge e_n$.

Why the word. A decomposable $\omega \neq 0$ carries a k-dimensional subspace: its factors are independent by Theorem 15.8.7, and they span one. Nondecomposable elements are exactly those that come from no single subspace, which is why the failure below is a phenomenon and not a notational accident.

Theorem 15.10.4. Not Every 2-Vector Is Decomposable

Let $\dim V = n \geq 4$, let $(e_1, \ldots, e_n)$ be a basis of V, and set

$$\omega = e_1 \wedge e_2 + e_3 \wedge e_4 \in \Lambda^2 V.$$

Then ω is **not** decomposable.

Idea. A decomposable 2-vector $u \wedge w$ squares to zero, because the four-fold wedge $u \wedge w \wedge u \wedge w$ repeats a vector. So compute $\omega \wedge \omega$ and find it nonzero. That test needs $\operatorname{char} F \neq 2$; the proof below takes a second route that does not, by asking which vectors annihilate ω.

Proof. Suppose $\omega = u \wedge w$ for some $u, w \in V$. Since $e_1 \wedge e_2$ and $e_3 \wedge e_4$ are distinct basis vectors of $\Lambda^2 V$ (Theorem 15.8.5), $\omega \neq 0$, so (u, w) is linearly independent by Theorem 15.8.7; in particular $u \neq 0$.

Consider $u \wedge \omega = u \wedge u \wedge w = 0$, the wedge having a repeated argument. We show this is impossible. Let $v = \sum_{i=1}^{n} x_i e_i$ be any vector and compute $v \wedge \omega$, using bilinearity and dropping every term with a repeated basis vector:

$$\begin{aligned} v \wedge \omega = {} & x_3\, e_3 \wedge e_1 \wedge e_2 + x_4\, e_4 \wedge e_1 \wedge e_2 \\ & + x_1\, e_1 \wedge e_3 \wedge e_4 + x_2\, e_2 \wedge e_3 \wedge e_4 \\ & + \sum_{i \geq 5} x_i\, (e_i \wedge e_1 \wedge e_2 + e_i \wedge e_3 \wedge e_4). \end{aligned}$$

Sorting each wedge by Lemma 15.8.1 (b) turns this into a combination of the distinct basis vectors

$$e_{123},\ e_{124},\ e_{134},\ e_{234},\ e_{12i},\ e_{34i}\ (i \geq 5)$$

of $\Lambda^3 V$, with coefficients x_3, x_4, x_1, x_2 and x_i, x_i respectively, up to signs. Since these basis vectors are distinct and the e_S are independent (Theorem 15.8.5), $v \wedge \omega = 0$ forces $x_1 = x_2 = \cdots = x_n = 0$, that is, $v = 0$. This contradicts $u \neq 0$ with $u \wedge \omega = 0$. Hence ω is not decomposable.

■

Remark. When $\operatorname{char} F \neq 2$ there is a one-line version: $\omega \wedge \omega = 2\, e_1 \wedge e_2 \wedge e_3 \wedge e_4 \neq 0$, whereas $(u \wedge w) \wedge (u \wedge w) = 0$ always. The proof above was written to avoid that division, because the test fails over $\mathbb{F}_2$ — see the warning next.

Warning. **The square test is blind in characteristic 2.** Over $\mathbb{F}_2$ every $\omega \in \Lambda^2 V$ satisfies $\omega \wedge \omega = 0$: in the expansion $\sum_{S,T} c_S c_T\, e_S \wedge e_T$ the diagonal terms vanish and each off-diagonal pair contributes $2 c_S c_T\, e_S \wedge e_T = 0$. So over $\mathbb{F}_2$ the vanishing of $\omega \wedge \omega$ says nothing at all, while $e_1 \wedge e_2 + e_3 \wedge e_4$ is still not decomposable, by the proof above. A criterion and a phenomenon are different things.

Quick check. Why does Theorem 15.10.4 need $n \geq 4$? Show that every element of $\Lambda^2 V$ is decomposable when $\dim V = 3$.

Solution. With $n \leq 3$ there is no room to write $e_1 \wedge e_2 + e_3 \wedge e_4$ at all. For $n = 3$, take $\omega = a\,e_{12} + b\,e_{13} + c\,e_{23}$. If $a = b = c = 0$ then $\omega = 0$, which is decomposable. If $a \neq 0$, then

$$\left(e_1 - \tfrac{c}{a} e_3\right) \wedge (a e_2 + b e_3) = a\,e_{12} + b\,e_{13} + c\,e_{23} = \omega,$$

using $e_3 \wedge e_2 = -e_{23}$. If $a = 0$, relabel the basis so that the surviving nonzero coefficient sits in the first slot and repeat. So the phenomenon starts exactly at $n = 4$.

Together with Example (A tensor that is not simple) this is the chapter's negative result, stated in its two forms: a general tensor is not a simple tensor, and a general k-vector is not a blade. The optimistic reading — that every element of a space built from V is built from finitely many vectors of V in the obvious way — is false in both places, and Section 11 will find the first form of it wearing a physical name.

15.10.3 Algebras modulo relations

One tool is still missing. The Clifford algebra is to be built by imposing relations on $\mathrm{T}(V)$, and imposing relations on an **algebra** means dividing by more than a subspace: the relations must stay relations after multiplication on either side. Chapter 6 met the commutative version of the idea for $F[x]$ (Definition 6.2.1).

Definition 15.10.3. Two-Sided Ideal, and the Ideal Generated by a Set

Let A be an F-algebra (Definition 14.12.1). A **two-sided ideal** of A is a subspace $\mathcal{I} \subseteq A$ such that $yx \in \mathcal{I}$ and $xy \in \mathcal{I}$ for **every** $x \in \mathcal{I}$ and **every** $y \in A$. For a subset $S \subseteq A$, the **ideal generated by** S is

$$\langle S \rangle \coloneqq \operatorname{span}\{\, ysz \;:\; y, z \in A,\ s \in S \,\}.$$

Examples. In $F[x]$ the ideals are the sets $\langle p \rangle$ of Definition 6.2.1; in any A, both $\{0\}$ and A are ideals. **Nonexample by minimal change.** The diagonal matrices in $M_2(F)$ form a subspace closed under multiplication, but not an ideal: it contains $\mathbf{E}_{11}$, while $\mathbf{E}_{21}\mathbf{E}_{11} = \mathbf{E}_{21}$ is not diagonal. Being closed among one's own elements is not the condition; absorbing products with the **whole** algebra is.

The span above is a subspace by construction, and it is a two-sided ideal because $a(ysz) = (ay)sz$ and $(ysz)a = ys(za)$ are again of the listed form. It contains S, taking $y = z = 1_A$, and every two-sided ideal containing S contains each ysz; so $\langle S \rangle$ is the smallest two-sided ideal containing S, which is what "generated" means.

Lemma 15.10.5. Quotient of an Algebra by an Ideal

Let A be an F-algebra and $\mathcal{I}$ a two-sided ideal of A.

(a) The quotient space $A/\mathcal{I}$ carries exactly one multiplication making the quotient map

$\pi\colon A \to A/\mathcal{I}$ an algebra homomorphism, namely $(x+\mathcal{I})(y+\mathcal{I}) = xy+\mathcal{I}$; with it $A/\mathcal{I}$ is an F-algebra with identity $1_A+\mathcal{I}$, and $\ker\pi = \mathcal{I}$.

(b) If $\varphi\colon A \to B$ is an algebra homomorphism with $\mathcal{I} \subseteq \ker\varphi$, then there is exactly one algebra homomorphism $\bar{\varphi}\colon A/\mathcal{I} \to B$ with $\bar{\varphi}\circ\pi = \varphi$.

(c) The kernel of any algebra homomorphism is a two-sided ideal.

Proof.

(a) The rule is well defined: if $x' = x+u$ and $y' = y+w$ with $u, w \in \mathcal{I}$, then

$$x'y' = xy + uy + xw + uw \in xy + \mathcal{I},$$

since $\mathcal{I}$ absorbs multiplication on both sides and is closed under addition. Bilinearity, associativity, and the identity are then inherited from A, because each is an identity among cosets that follows by applying π to the corresponding identity in A. Any multiplication making π multiplicative must satisfy the displayed rule, and π is surjective, so the multiplication is unique. Finally $\ker\pi = \mathcal{I}$ by Lemma 4.4.3.

(b) Since φ is linear with $\mathcal{I} \subseteq \ker\varphi$, Theorem 4.4.10 (a) gives exactly one **linear** $\bar{\varphi}$ with $\bar{\varphi}\pi = \varphi$. It is multiplicative: every element of $A/\mathcal{I}$ is $\pi(x)$ for some x, and

$$\bar{\varphi}(\pi(x)\pi(y)) = \bar{\varphi}(\pi(xy)) = \varphi(xy) = \varphi(x)\varphi(y),$$

which is $\bar{\varphi}(\pi(x))\,\bar{\varphi}(\pi(y))$. Also $\bar{\varphi}(\pi(1_A)) = \varphi(1_A) = 1_B$. Uniqueness is already contained in the uniqueness of the linear map.

(c) If $\varphi(x) = 0$ and $y \in A$, then $\varphi(xy) = \varphi(x)\varphi(y) = 0$ and $\varphi(yx) = 0$; and $\ker\varphi$ is a subspace because φ is linear.

■

15.10.4 The Clifford algebra, constructed

Chapter 14 §12 defined a Clifford algebra of (V, q) as a pair (C, ι) with $\iota\colon V \to C$ linear satisfying

- (C1) $\iota(v)^2 = q(v)1_C$ for every $v \in V$, and
- (C2) the universal property: every linear $f\colon V \to A$ into an F-algebra with $f(v)^2 = q(v)1_A$ factors as $f = \bar{f}\circ\iota$ for a unique algebra homomorphism $\bar{f}$,

and then said in as many words that it could not produce one: "The definition says what a Clifford algebra would be; it does not produce one." What it needed was an algebra of formal strings of vectors, which is now Definition 15.6.2, and a way to impose the relation $v\otimes v = q(v)1$, which is now Lemma 15.10.5.

Theorem 15.10.6. Existence of the Clifford Algebra

Let F be a field with $\operatorname{char} F \neq 2$, let V be a finite-dimensional F-vector space, let β be a symmetric bilinear form on V and $q(v) = \beta(v, v)$. Let

$$\mathcal{I}_q \coloneqq \langle\, v\otimes v - q(v)1 \;:\; v \in V \,\rangle \subseteq \mathrm{T}(V)$$

be the two-sided ideal generated by these elements, put $\mathrm{Cl}(q) \coloneqq \mathrm{T}(V)/\mathcal{I}_q$, and let $\iota\colon V \to$

$\mathrm{Cl}(q)$ be the restriction to $V \subseteq \mathrm{T}(V)$ of the quotient map π. Then $(\mathrm{Cl}(q), \iota)$ is a Clifford algebra of (V, q) in the sense of Definition 14.12.2.

In particular a Clifford algebra exists for every (V, q), and by Proposition 14.12.1 any two are isomorphic by a unique isomorphism compatible with the structure maps.

Idea. The tensor algebra is the free associative algebra on V: a linear map out of V extends uniquely to it (Theorem 15.6.6), with no relations in the way. Dividing by the smallest ideal containing $v \otimes v - q(v)1$ imposes exactly the relation (C1) and nothing else, and "nothing else" is precisely what the universal property (C2) asserts.

Proof. By Lemma 15.10.5 (a), $\mathrm{Cl}(q)$ is an F-algebra and π is a surjective algebra homomorphism with $\ker \pi = \mathcal{I}_q$. The map $\iota = \pi|_V$ is linear, being a restriction of a linear map.

(C1). Let $v \in V$. In $\mathrm{T}(V)$ the product of v with itself is $v \otimes v$ (Definition 15.6.2), and $v \otimes v - q(v)1$ is a generator of $\mathcal{I}_q = \ker \pi$. Since π is multiplicative and linear,

$$\iota(v)^2 = \pi(v)\pi(v) = \pi(v \otimes v) = \pi\big(q(v)1\big) = q(v)1_{\mathrm{Cl}(q)}.$$

(C2). Let A be an F-algebra and let $f \colon V \to A$ be linear with $f(v)^2 = q(v)1_A$ for every $v \in V$. By Theorem 15.6.6 there is a unique algebra homomorphism $\tilde{f} \colon \mathrm{T}(V) \to A$ with $\tilde{f}|_V = f$. For every $v \in V$,

$$\tilde{f}(v \otimes v - q(v)1) = f(v)f(v) - q(v)1_A = 0,$$

so $\ker \tilde{f}$ contains every generator of $\mathcal{I}_q$. By Lemma 15.10.5 (c), $\ker \tilde{f}$ is a two-sided ideal, and $\mathcal{I}_q$ is the **smallest** one containing those generators (Definition 15.10.3), so $\mathcal{I}_q \subseteq \ker \tilde{f}$. By Lemma 15.10.5 (b) there is exactly one algebra homomorphism $\bar{f} \colon \mathrm{Cl}(q) \to A$ with $\bar{f} \circ \pi = \tilde{f}$, and restricting to V gives $\bar{f} \circ \iota = \tilde{f}|_V = f$.

For the uniqueness clause of (C2), let $g \colon \mathrm{Cl}(q) \to A$ be any algebra homomorphism with $g \circ \iota = f$. Then $g \circ \pi$ is an algebra homomorphism $\mathrm{T}(V) \to A$ whose restriction to V is $g\iota = f$, so $g\pi = \tilde{f}$ by the uniqueness clause of Theorem 15.6.6. Hence $g\pi = \bar{f}\pi$, and π is surjective, so $g = \bar{f}$. This proves the theorem.

■

Remark. Neither $\mathrm{char}\, F \neq 2$ nor finite dimension was used in this proof; the construction and both checks work for any V and any quadratic form. The hypotheses are carried because Definition 14.12.2 states them, and because the dimension count below needs an orthogonal basis.

15.10.5 The dimension, at last

Chapter 14 proved half of the count. Its Proposition 14.12.3 showed that for an orthogonal basis $(e_1, \dots, e_n)$ of β the 2^n **sorted products**

$$\iota(e_{i_1})\iota(e_{i_2}) \cdots \iota(e_{i_k}), \qquad i_1 < i_2 < \cdots < i_k,$$

span C, so that $\dim_F C \le 2^n$. The other half — their independence — is what this section owes. The idea is to make ΛV, which has exactly the right dimension, into a place where the vectors of V act with square $q(v)$; the universal property then sends C into the operators on ΛV, and evaluating at 1 turns the sorted products into distinct basis vectors.

Two families of operators on ΛV do the work: wedging a basis vector on, and deleting it again.

Lemma 15.10.7. Creation and Deletion Operators

Let $(e_1, \dots, e_n)$ be a basis of V, and for $S \subseteq [n]$ let e_S be as in Theorem 15.10.2. For $i \in [n]$ write $\sigma(i, S) = |\{j \in S : j < i\}|$. Let $E_i \in \mathcal{L}(\Lambda V)$ be left multiplication by e_i, that is, $E_i(\omega) = e_i \wedge \omega$, and let $D_i \in \mathcal{L}(\Lambda V)$ be the unique linear map with

$$D_i(e_S) = \begin{cases} (-1)^{\sigma(i,S)}\, e_{S\setminus\{i\}}, & i \in S, \\ 0, & i \notin S, \end{cases}$$

which exists and is unique because the e_S form a basis (Theorem 15.10.2). Then for all $i, j \in [n]$,

$$E_iE_j + E_jE_i = 0,$$
$$D_iD_j + D_jD_i = 0,$$
$$D_iE_j + E_jD_i = \delta_{ij}\ \mathrm{id}_{\Lambda V}.$$

Idea. E_i inserts the index i into the set S and D_i removes it, each at the cost of the sign needed to move e_i past the smaller members of S. All three identities are then a comparison of two signs, and in the third the two cases "$i \in S$" and "$i \notin S$" are exactly the two ways the identity can be satisfied: one term is zero and the other is e_S.

Proof. First record what E_i does on the basis. If $i \in S$ then $e_i \wedge e_S$ repeats e_i and is 0. If $i \notin S$, then sorting $e_i \wedge e_S$ moves e_i past the $\sigma(i, S)$ members of S smaller than i, so by Lemma 15.8.1 (b)

$$E_i(e_S) = (-1)^{\sigma(i,S)}\, e_{S\cup\{i\}}.$$

First identity. For every $\omega \in \Lambda V$, $E_iE_j(\omega) = e_i \wedge e_j \wedge \omega$ and $E_jE_i(\omega) = e_j \wedge e_i \wedge \omega$, and $e_j \wedge e_i = -e_i \wedge e_j$ by Theorem 15.10.3. Adding gives 0. (For $i = j$ each term is separately 0.)

Third identity, case $i = j$. If $i \notin S$, then $D_i(e_S) = 0$ and $E_i(e_S) = (-1)^{\sigma(i,S)} e_{S\cup\{i\}}$; since $\sigma(i, S\cup\{i\}) = \sigma(i, S)$, applying D_i returns $(-1)^{2\sigma(i,S)} e_S = e_S$. If $i \in S$, then $E_i(e_S) = 0$ and $D_i(e_S) = (-1)^{\sigma(i,S)} e_{S\setminus\{i\}}$; since $\sigma(i, S \setminus \{i\}) = \sigma(i, S)$, applying E_i returns e_S. Either way the sum is e_S.

Third identity, case $i \neq j$. If $j \in S$ then $E_j(e_S) = 0$, and $D_i(e_S)$ is either 0 or a multiple of $e_{S\setminus\{i\}}$, which still contains j, so E_j kills it; both terms vanish. If $j \notin S$ and $i \notin S$ then $D_i(e_S) = 0$ and $D_i(e_{S\cup\{j\}}) = 0$ as well, since $i \neq j$; again both terms vanish. So let $j \notin S$ and $i \in S$. Then

$$D_iE_j(e_S) = (-1)^{\sigma(j,S)+\sigma(i,S\cup\{j\})}\, e_{(S\cup\{j\})\setminus\{i\}},$$
$$E_jD_i(e_S) = (-1)^{\sigma(i,S)+\sigma(j,S\setminus\{i\})}\, e_{(S\setminus\{i\})\cup\{j\}},$$

and the two index sets agree. Now $\sigma(i, S\cup\{j\}) = \sigma(i, S) + [\,j < i\,]$ and $\sigma(j, S\setminus\{i\}) = \sigma(j, S) - [\,i < j\,]$, writing $[\,P\,]$ for 1 if P holds and 0 otherwise. So the two exponents differ

by $[\,j<i\,]+[\,i<j\,]=1$, which is odd because $i\neq j$. The two terms therefore cancel.

Second identity. Both $D_iD_j(e_S)$ and $D_jD_i(e_S)$ vanish unless $i,j\in S$ and $i\neq j$, and in that case they are $(-1)^{\sigma(j,S)+\sigma(i,S\setminus\{j\})}$ and $(-1)^{\sigma(i,S)+\sigma(j,S\setminus\{i\})}$ times the same $e_{S\setminus\{i,j\}}$. The same two identities for σ show the exponents differ by an odd number, so the terms cancel. For $i=j$, $D_i^2(e_S)=0$ because $i\notin S\setminus\{i\}$. This proves the lemma.

■

Theorem 15.10.8. Dimension of a Clifford Algebra

Let F be a field with $\operatorname{char}F\neq 2$, let $\dim V=n$, let β be a symmetric bilinear form on V with quadratic form q, and let (C,ι) be a Clifford algebra of (V,q). Let $(e_1,\dots,e_n)$ be an orthogonal basis of V for β. Then the 2^n sorted products of Proposition 14.12.3 form a **basis** of C. Consequently

$$\dim_F C=2^n,$$

the map ι is injective, and C is isomorphic to ΛV **as a vector space**. This is Theorem 14.12.4.

Idea. Build a representation. ① Turn each basis vector e_i into the operator $c_i=E_i+d_iD_i$ on ΛV, where $d_i=q(e_i)$; Lemma 15.10.7 makes these anticommute and gives $c_i^2=d_i\,\mathrm{id}$, which is exactly (C1). ② The universal property then hands us an algebra homomorphism $C\to\mathcal{L}(\Lambda V)$. ③ Evaluate at the element $1\in\Lambda^0V$: the sorted product for the index set S is sent to e_S, because at each step the new index is smaller than all the ones already there, so the deletion half contributes nothing. Distinct basis vectors come out, so the sorted products were independent.

Proof. An orthogonal basis exists by Theorem 14.4.2, since $\operatorname{char}F\neq 2$. Write $d_i=q(e_i)$ and $c_i=E_i+d_iD_i\in\mathcal{L}(\Lambda V)$, with E_i,D_i as in Lemma 15.10.7. By Proposition 14.12.3 the sorted products span C and $\dim_F C\le 2^n$, so only independence is left.

Step 1: the c_i satisfy the Clifford relations. Expanding the product and collecting the four brackets,

$$\begin{aligned}c_ic_j+c_jc_i=&(E_iE_j+E_jE_i)+d_j(E_iD_j+D_jE_i)\\&+d_i(D_iE_j+E_jD_i)+d_id_j(D_iD_j+D_jD_i).\end{aligned}$$

By Lemma 15.10.7 the first and last brackets are 0 and each middle bracket is $\delta_{ij}\,\mathrm{id}$. So $c_ic_j=-c_jc_i$ for $i\neq j$. Setting $i=j$ in the display would give $2c_i^2$, so expand c_i^2 directly instead; with $E_i^2=D_i^2=0$, which were checked separately in the lemma,

$$c_i^2=E_i^2+d_i(E_iD_i+D_iE_i)+d_i^2D_i^2=d_i\,\mathrm{id},$$

with no division by 2 anywhere. Now $\mathcal{L}(\Lambda V)$ is an F-algebra under composition (Definition 14.12.1). Let $f\colon V\to\mathcal{L}(\Lambda V)$ be the linear map with $f(e_i)=c_i$, which exists and is unique by Theorem 2.1.6. For $v=\sum_i x_ie_i$,

$$f(v)^2=\sum_i x_i^2c_i^2+\sum_{i<j}x_ix_j(c_ic_j+c_jc_i)=\Big(\sum_i d_ix_i^2\Big)\,\mathrm{id},$$

the cross terms vanishing because $i \neq j$ there. Since the basis is orthogonal, $\beta(e_i, e_j) = 0$ for $i \neq j$, so $q(v) = \beta(v, v) = \sum_i d_i x_i^2$. Hence $f(v)^2 = q(v)\,\mathrm{id}$, which is (C1) for f.

Step 2: the induced homomorphism. By the universal property (C2) of Definition 14.12.2 there is an algebra homomorphism $\Psi\colon C \to \mathcal{L}(\Lambda V)$ with $\Psi\iota = f$. Define the linear map

$$\Phi\colon C \to \Lambda V, \qquad \Phi(x) = \Psi(x)(1),$$

where $1 \in \Lambda^0 V = F \subseteq \Lambda V$. It is linear because Ψ is linear and evaluation at a fixed element is linear.

Step 3: Φ sends sorted products to basis vectors. Let $S = \{i_1 < \cdots < i_k\} \subseteq [n]$. Since Ψ is multiplicative and $\Psi\iota = f$,

$$\Psi\big(\iota(e_{i_1}) \cdots \iota(e_{i_k})\big) = c_{i_1} c_{i_2} \cdots c_{i_k}.$$

We claim $c_{i_r} c_{i_{r+1}} \cdots c_{i_k}(1) = e_{T_r}$ with $T_r = \{i_r, \ldots, i_k\}$, by downward induction on r. For $r = k+1$ the product is empty and the value is $1 = e_\emptyset$. Assume the claim for $r+1$. Because $i_r < i_{r+1}$ and the indices increase, $i_r \notin T_{r+1}$, so $D_{i_r}(e_{T_{r+1}}) = 0$; and i_r is smaller than every element of T_{r+1}, so $\sigma(i_r, T_{r+1}) = 0$ and $E_{i_r}(e_{T_{r+1}}) = e_{T_r}$. Hence $c_{i_r}(e_{T_{r+1}}) = e_{T_r}$, completing the induction. Taking $r = 1$,

$$\Phi\big(\iota(e_{i_1}) \cdots \iota(e_{i_k})\big) = e_S,$$

and for $S = \emptyset$, $\Phi(1_C) = \mathrm{id}(1) = 1 = e_\emptyset$.

Step 4: conclusion. Suppose a linear combination of the sorted products is 0. Applying Φ gives the same combination of the e_S, which are linearly independent by Theorem 15.10.2 (b); so all coefficients vanish. The sorted products are therefore linearly independent, and being also a spanning list they form a basis of C, of size 2^n. Thus $\dim_F C = 2^n$.

Since 1_C and $\iota(e_1), \ldots, \iota(e_n)$ are among the basis elements just found, the list $(\iota(e_1), \ldots, \iota(e_n))$ is independent, so ι carries a basis of V to an independent list and is injective. Finally Φ carries the basis of C bijectively onto the basis (e_S) of ΛV, so Φ is a vector space isomorphism. This proves the theorem.

■

The debt, and how it was paid. Chapter 14 §12 stated Theorem 14.12.4 and then wrote, under the heading "What is proved here, and what is not", that the inequality $\dim_F C \leq 2^n$ was proved in full, while "the reverse inequality — that the 2^n sorted products are linearly independent — is **not proved in this section**", and that the two halves would have to be established together by exhibiting one algebra of dimension exactly 2^n satisfying (C1) and (C2). Both halves are now supplied. The missing one was independence, and Theorem 15.10.8 supplies it; the prior question, whether any Clifford algebra exists at all, is Theorem 15.10.6. Chapter 14 read the statement for $n \geq 3$ "as quoted from Chapter 15 and not as proved here"; it may now be read as proved.

The prediction was right in substance and slightly off in shape. The algebra of dimension 2^n that settled the matter is the exterior algebra, but it is not the Clifford algebra: it is only a space that $\mathrm{Cl}(q)$ acts on, and independence came from the action rather than from an identification. Chapter 14 also guessed that "$\mathrm{Cl}(q)$ is assembled from the exterior powers $\Lambda^k V$, one for each k", and Φ is that assembly, matching the sorted product on S with the wedge e_S.

Warning. $\mathrm{Cl}(q)$ **and ΛV are the same size, not the same algebra.** The map Φ above is an isomorphism of **vector spaces** only. It is not multiplicative, and no other map is either, unless $q = 0$: in ΛV the square of a vector is 0, while in $\mathrm{Cl}(q)$ it is $q(v)1$. Writing an element of $\mathrm{Cl}(q)$ as a sum of wedges is a choice of coordinates, not a description of the multiplication.

That warning deserves a proof rather than a slogan.

Proposition 15.10.9. Comparing the Two Algebras

Let $\operatorname{char} F \neq 2$ and $\dim V = n \geq 1$, and let q be the quadratic form of a symmetric bilinear form on V.

(a) If $q = 0$, then $\mathrm{Cl}(0)$ and ΛV are isomorphic as F-algebras.

(b) If $q \neq 0$, they are **not** isomorphic as F-algebras, although they are isomorphic as vector spaces.

Idea. For (b), find a property of ΛV that $\mathrm{Cl}(q)$ fails. In ΛV everything of positive degree is nilpotent, so an element whose square is a **nonzero scalar** can have no positive-degree part at all; but in $\mathrm{Cl}(q)$ a vector v with $q(v) \neq 0$ squares to a nonzero scalar and is not a scalar itself.

Proof.

(a) Let $q = 0$ and let $j\colon V \to \Lambda V$ be the inclusion $V = \Lambda^1 V \subseteq \Lambda V$. Then $j(v)^2 = v \wedge v = 0 = q(v)1$, so (C1) holds for j. By the universal property (C2) of $\mathrm{Cl}(0)$ there is an algebra homomorphism $\psi\colon \mathrm{Cl}(0) \to \Lambda V$ with $\psi\iota = j$. Its image is a subalgebra containing 1 and V, hence all of ΛV by Theorem 15.10.2 (c); so ψ is surjective. Both spaces have dimension 2^n, by Theorem 15.10.8 and Theorem 15.10.2 (b), so ψ is bijective and is an algebra isomorphism.

(b) *Claim.* If $x \in \Lambda V$ and $x \wedge x = a \cdot 1$ with $a \in F$, $a \neq 0$, then $x \in F \cdot 1$.

Proof of claim. Throughout, y^t means the t-fold wedge product of y with itself. Write $x = x_0 + m$ with $x_0 \in \Lambda^0 V = F$ and m the sum of the components of positive degree. Then

$$x \wedge x = x_0^2 + 2x_0 m + m \wedge m,$$

and the only degree-0 term is x_0^2, so $x_0^2 = a \neq 0$ and hence $x_0 \neq 0$. Canceling that term leaves

$$m \wedge (2x_0 \cdot 1 + m) = 0.$$

Every product of $n+1$ elements of positive degree lies in $\bigoplus_{k>n} \Lambda^k V = \{0\}$, so $m^{n+1} = 0$. Put $c = 2x_0$, which is nonzero because $\operatorname{char} F \neq 2$ and $x_0 \neq 0$, and $u = -c^{-1}m$, so that $u^{n+1} = 0$ and $c \cdot 1 + m = c(1 - u)$. Then

$$(1 - u) \wedge (1 + u + u^2 + \cdots + u^n) = 1 - u^{n+1} = 1,$$

and likewise in the other order, so $c \cdot 1 + m$ is invertible. Multiplying $m \wedge (c \cdot 1 + m) = 0$ by its inverse gives $m = 0$, so $x = x_0 \in F \cdot 1$, proving the claim.

Now suppose $q \neq 0$ and let $\varphi\colon \mathrm{Cl}(q) \to \Lambda V$ be an algebra isomorphism. Choose $v \in V$ with $q(v) \neq 0$ and put $x = \varphi(\iota(v))$. Then

$$x \wedge x = \varphi(\iota(v)^2) = \varphi(q(v)1) = q(v) \cdot 1,$$

a nonzero scalar, so $x \in F \cdot 1$ by the claim, say $x = c \cdot 1 = \varphi(c\,1_{\mathrm{Cl}(q)})$. Since φ is injective, $\iota(v) = c\,1_{\mathrm{Cl}(q)}$. But by Theorem 15.10.8 the elements $1_{\mathrm{Cl}(q)}, \iota(e_1), \dots, \iota(e_n)$ are part of a basis of $\mathrm{Cl}(q)$, and $\iota(v)$ is a combination of $\iota(e_1), \dots, \iota(e_n)$ alone; comparing coordinates forces $c = 0$ and $\iota(v) = 0$, contradicting $\iota(v)^2 = q(v)1 \neq 0$. So no such φ exists. The vector space isomorphism is Theorem 15.10.8.

■

The dimension count also makes Clifford algebras easy to recognize, which is how one identifies them in practice.

Corollary 15.10.10. Recognizing a Clifford Algebra

Let $\mathrm{char}\, F \neq 2$, $\dim V = n$, and let q be as above. Let A be an F-algebra with $\dim_F A = 2^n$, and let $\iota_A\colon V \to A$ be linear with $\iota_A(v)^2 = q(v)1_A$ for every $v \in V$. If A is generated as an algebra by 1_A and $\iota_A(V)$, then (A, ι_A) is a Clifford algebra of (V, q).

Proof. Let (C, ι) be the Clifford algebra of Theorem 15.10.6. By its universal property there is an algebra homomorphism $\psi\colon C \to A$ with $\psi\iota = \iota_A$. Its image is a subalgebra of A containing 1_A and $\iota_A(V)$, hence is all of A by the generation hypothesis. So ψ is surjective, and $\dim_F C = 2^n = \dim_F A$ by Theorem 15.10.8, so ψ is an isomorphism.

It remains to transport (C2). Let $f\colon V \to B$ be linear with $f(v)^2 = q(v)1_B$. Then $\bar{f}\psi^{-1}\colon A \to B$ is an algebra homomorphism with $\bar{f}\psi^{-1}\iota_A = \bar{f}\iota = f$, where $\bar{f}\colon C \to B$ comes from (C2) for C; and if $g\colon A \to B$ is another one, then $g\psi\colon C \to B$ satisfies $g\psi\iota = g\iota_A = f$, so $g\psi = \bar{f}$ by uniqueness for C, whence $g = \bar{f}\psi^{-1}$.

■

Example. The Clifford algebra of the Euclidean plane

Let $V = \mathbb{R}^2$ with $q(\mathbf{x}) = x_1^2 + x_2^2$, the form of signature $(2, 0)$. Show that $\mathrm{Cl}_{2,0}(\mathbb{R}) \cong M_2(\mathbb{R})$.

Solution. Put

$$\mathbf{A} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \qquad \mathbf{B} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

in $M_2(\mathbb{R})$, and let $\iota(\mathbf{x}) = x_1\mathbf{A} + x_2\mathbf{B}$, a linear map $\mathbb{R}^2 \to M_2(\mathbb{R})$. Multiplying out, $\mathbf{A}^2 = \mathbf{B}^2 = \mathbf{I}_2$ and $\mathbf{AB} = -\mathbf{BA}$, the products being $\left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)$ and its negative. Hence

$$\iota(\mathbf{x})^2 = x_1^2\mathbf{A}^2 + x_1x_2(\mathbf{AB} + \mathbf{BA}) + x_2^2\mathbf{B}^2 = q(\mathbf{x})\mathbf{I}_2,$$

which is (C1). Next, $\dim_{\mathbb{R}} M_2(\mathbb{R}) = 4 = 2^2$, and the four matrices $\mathbf{I}_2, \mathbf{A}, \mathbf{B}, \mathbf{AB}$ are linearly independent: the first two span the diagonal matrices and the last two span the off-diagonal ones, and the diagonal and off-diagonal matrices meet only in 0. So they are a basis, and

$M_2(\mathbb{R})$ is generated by $\mathbf{I}_2$ and $\iota(\mathbb{R}^2)$. By Corollary 15.10.10, $(M_2(\mathbb{R}), \iota)$ is a Clifford algebra of $(\mathbb{R}^2, q)$, so $\mathrm{Cl}_{2,0}(\mathbb{R}) \cong M_2(\mathbb{R})$.

Compare Chapter 14's $\mathrm{Cl}_{0,2}(\mathbb{R}) = \mathbb{H}$ (Example (The quaternions)). All three of $\mathbb{H}$, $M_2(\mathbb{R})$, and $\Lambda\mathbb{R}^2$ have dimension 4, and no two are isomorphic: $\mathbb{H}$ has no zero divisors, $M_2(\mathbb{R})$ has plenty, and $\Lambda\mathbb{R}^2$ differs from both by Proposition 15.10.9 (b). Changing a sign in q changes the algebra completely and never changes its size.

Quick check. Where in the proof of Theorem 15.10.8 is the hypothesis $\operatorname{char} F \neq 2$ used?

Solution. Twice, both times through the orthogonal basis. It is needed to produce one at all: Theorem 14.4.2 assumes $\operatorname{char} F \neq 2$. And it is needed for Chapter 14's Proposition 14.12.3, which supplies the spanning half, since that proof uses Lemma 14.12.2, which comes from polarization. The operator identities of Lemma 15.10.7 and the induction of Step 3 use no such hypothesis.

With that, Chapter 14's last open statement is closed and the construction side of this chapter is complete. The final section spends the machinery on one application, where the failure of decomposability proved above reappears under another name.

15.10.6 Exercises

15.10.6.1 A. Check your understanding

Exercise. A1

(a) Define the exterior algebra ΛV and state its dimension when $\dim V = n$.

(b) State the graded commutativity rule, and say what it gives when both degrees are odd.

(c) Define what it means for $\omega \in \Lambda^k V$ to be decomposable, and give one decomposable and one nondecomposable element of $\Lambda^2 F^4$.

(d) Which half of Theorem 14.12.4 was left unproved in Chapter 14, and which result of this section supplies it?

(e) True or false: $\mathrm{Cl}(q)$ and ΛV are isomorphic algebras. Justify your answer.

Solution.

(a) $\Lambda V = \bigoplus_{k \geq 0} \Lambda^k V$ with the product of Proposition 15.10.1 extended by bilinearity (Definition 15.10.1). Its dimension is $\sum_k \binom{n}{k} = 2^n$ (Theorem 15.10.2).

(b) $\omega \wedge \eta = (-1)^{kl} \eta \wedge \omega$ for $\omega \in \Lambda^k V$, $\eta \in \Lambda^l V$ (Theorem 15.10.3). Both odd gives $\omega \wedge \eta = -\eta \wedge \omega$: odd-degree elements anticommute.

(c) ω is decomposable if $\omega = v_1 \wedge \cdots \wedge v_k$ for some $v_i \in V$ (Definition 15.10.2). Decomposable: $e_1 \wedge e_2$. Nondecomposable: $e_1 \wedge e_2 + e_3 \wedge e_4$ (Theorem 15.10.4).

(d) The independence of the 2^n sorted products, that is, the inequality $\dim_F C \geq 2^n$. It is supplied by Theorem 15.10.8, which needs Theorem 15.10.6 first so that the object exists.

(e) False in general. They are isomorphic as vector spaces, and as algebras exactly when $q = 0$ (Proposition 15.10.9).

15.10.6.2 B. Practice

Exercise. B1: Multiplying in the exterior algebra

In ΛF^4 with $F = \mathbb{Q}$, put $x = \mathbf{e}_1 + \mathbf{e}_2 \wedge \mathbf{e}_3$ and $y = \mathbf{e}_2 + \mathbf{e}_1 \wedge \mathbf{e}_4$. Compute the following, expressed in the basis of Theorem 15.10.2.

(a) $x \wedge y$.

(b) $y \wedge x$.

(c) $x \wedge x$.

Solution. Write e_S for the sorted wedge on S.

(a) Expanding into four products:

$$\mathbf{e}_1 \wedge \mathbf{e}_2 = e_{12}, \qquad \mathbf{e}_1 \wedge (\mathbf{e}_1 \wedge \mathbf{e}_4) = 0,$$
$$(\mathbf{e}_2 \wedge \mathbf{e}_3) \wedge \mathbf{e}_2 = 0, \qquad (\mathbf{e}_2 \wedge \mathbf{e}_3) \wedge (\mathbf{e}_1 \wedge \mathbf{e}_4) = e_{1234}.$$

The last one sorts as $\mathbf{e}_2 \wedge \mathbf{e}_3 \wedge \mathbf{e}_1 \wedge \mathbf{e}_4 = e_{1234}$, since moving $\mathbf{e}_1$ two places to the left costs $(-1)^2$. So $x \wedge y = e_{12} + e_{1234}$.

(b) The surviving products are $\mathbf{e}_2 \wedge \mathbf{e}_1 = -e_{12}$ and $(\mathbf{e}_1 \wedge \mathbf{e}_4) \wedge (\mathbf{e}_2 \wedge \mathbf{e}_3) = e_{1234}$, the latter because $\mathbf{e}_1 \wedge \mathbf{e}_4 \wedge \mathbf{e}_2 \wedge \mathbf{e}_3$ sorts with two transpositions. So $y \wedge x = -e_{12} + e_{1234}$. This agrees with Theorem 15.10.3 applied degree by degree: the $(1,1)$ part changes sign, the $(1,2)$ and $(2,2)$ parts do not.

(c) $x \wedge x = \mathbf{e}_1 \wedge \mathbf{e}_1 + \mathbf{e}_1 \wedge \mathbf{e}_2 \wedge \mathbf{e}_3 + \mathbf{e}_2 \wedge \mathbf{e}_3 \wedge \mathbf{e}_1 + 0 = 2\,e_{123}$, since $\mathbf{e}_2 \wedge \mathbf{e}_3 \wedge \mathbf{e}_1 = e_{123}$. An inhomogeneous element need not square to zero.

Exercise. B2: Decomposable or not

In $\Lambda^2\mathbb{Q}^4$, writing $e_{ij} = \mathbf{e}_i \wedge \mathbf{e}_j$, determine whether each of the following is decomposable. Justify your answer.

$$\omega_1 = e_{12} + e_{13} + 3e_{14} - 2e_{23} - e_{24} + 5e_{34}, \qquad \omega_2 = e_{12} + e_{34} + e_{14}.$$

Solution. For $\omega = \sum_{i<j} c_{ij} e_{ij}$ in $\Lambda^2\mathbb{Q}^4$, expanding gives

$$\omega \wedge \omega = 2(c_{12}c_{34} - c_{13}c_{24} + c_{14}c_{23})\, e_{1234},$$

since the only pairs of index sets with empty intersection are $\{12\},\{34\}$ and $\{13\},\{24\}$ and $\{14\},\{23\}$, each occurring in two orders, and $e_{13} \wedge e_{24} = -e_{1234}$ while the other two products are $+e_{1234}$.

For ω_1: $1 \cdot 5 - 1 \cdot (-1) + 3 \cdot (-2) = 5 + 1 - 6 = 0$, so $\omega_1 \wedge \omega_1 = 0$ and the test is inconclusive. In fact ω_1 **is** decomposable: it is $\mathbf{v}_1 \wedge \mathbf{v}_2$ for $\mathbf{v}_1 = (1,0,2,1)$ and $\mathbf{v}_2 = (0,1,1,3)$, by Example (A wedge of two vectors in four-dimensional space).

For ω_2: $1 \cdot 1 - 0 \cdot 0 + 1 \cdot 0 = 1$, so $\omega_2 \wedge \omega_2 = 2\, e_{1234} \neq 0$. A decomposable $u \wedge w$ satisfies $(u \wedge w) \wedge (u \wedge w) = 0$, so ω_2 is **not** decomposable.

Exercise. B3: A Clifford algebra of a line

Let $V = \mathbb{R}$ with $q(x) = x^2$, the form of signature $(1, 0)$. Let $A = \mathbb{R} \times \mathbb{R}$ with componentwise operations, and $\iota_A(x) = (x, -x)$. Prove that (A, ι_A) is a Clifford algebra of (V, q), so that $\mathrm{Cl}_{1,0}(\mathbb{R}) \cong \mathbb{R} \times \mathbb{R}$. Contrast this with $\mathrm{Cl}_{0,1}(\mathbb{R})$.

Solution. A is an associative unital $\mathbb{R}$-algebra with identity $(1, 1)$ and $\dim_{\mathbb{R}} A = 2 = 2^1$. The map ι_A is linear, and

$$\iota_A(x)^2 = (x, -x)(x, -x) = (x^2, x^2) = q(x)(1, 1),$$

which is (C1). The subalgebra generated by $(1, 1)$ and $\iota_A(\mathbb{R}) = \mathrm{span}\{(1, -1)\}$ contains the independent pair $(1, 1), (1, -1)$, hence is all of A. By Corollary 15.10.10, (A, ι_A) is a Clifford algebra of (V, q).

The contrast: Chapter 14 showed $\mathrm{Cl}_{0,1}(\mathbb{R}) = \mathbb{C}$ (Example (The complex numbers)), where $q(x) = -x^2$. Both algebras have dimension 2, but $\mathbb{C}$ is a field and $\mathbb{R} \times \mathbb{R}$ has zero divisors, $(1, 0)(0, 1) = (0, 0)$. A single sign in q decides which.

15.10.6.3 C. Going deeper

Exercise. C1: The Plücker relation

Let $\dim V = 4$ with basis $(e_1, \ldots, e_4)$, let $\mathrm{char}\, F \neq 2$, and let $\omega = \sum_{i<j} c_{ij} e_{ij} \in \Lambda^2 V$.

(a) Prove that $\omega \wedge \omega = 2(c_{12}c_{34} - c_{13}c_{24} + c_{14}c_{23})\, e_{1234}$.

(b) Deduce that if $c_{12}c_{34} - c_{13}c_{24} + c_{14}c_{23} \neq 0$ then ω is not decomposable.

(c) Prove the converse: if the expression vanishes, then ω is decomposable. *Hint: if $\omega \neq 0$, some $c_{ij} \neq 0$; after relabeling the basis assume $c_{12} \neq 0$ and look for ω as a wedge of two vectors whose first two coordinates you may normalize.*

Solution.

(a) By bilinearity, $\omega \wedge \omega = \sum_{S,T} c_S c_T\, e_S \wedge e_T$ over 2-element sets S, T. A term vanishes unless $S \cap T = \emptyset$, so the surviving pairs are $(\{1, 2\}, \{3, 4\})$, $(\{1, 3\}, \{2, 4\})$, $(\{1, 4\}, \{2, 3\})$, and their reverses. By Theorem 15.10.3 with $k = l = 2$, reversing a pair does not change the product, so each contributes twice. Sorting, $e_{12} \wedge e_{34} = e_{1234}$; $e_{13} \wedge e_{24} = e_1 \wedge e_3 \wedge e_2 \wedge e_4 = -e_{1234}$, one transposition; and $e_{14} \wedge e_{23} = e_1 \wedge e_4 \wedge e_2 \wedge e_3 = e_{1234}$, two transpositions. Collecting gives the stated formula.

(b) If $\omega = u \wedge w$ then $\omega \wedge \omega = u \wedge w \wedge u \wedge w = 0$, a wedge with a repeated vector. Since $\mathrm{char}\, F \neq 2$, (a) then forces the bracket to vanish. Contrapositively, a nonzero bracket rules out decomposability.

(c) If $\omega = 0$ it is decomposable. Otherwise relabel the basis so that $c_{12} \neq 0$; this is legitimate because a permutation of the basis permutes the six coefficients up to sign

and changes the bracket by at most a sign, so both the hypothesis and the conclusion are preserved. Put

$$u = c_{12}e_1 - c_{23}e_3 - c_{24}e_4, \qquad w = e_2 + \tfrac{c_{13}}{c_{12}}e_3 + \tfrac{c_{14}}{c_{12}}e_4.$$

Expanding and sorting each wedge,

$$\begin{aligned} u \wedge w = {} & c_{12}e_{12} + c_{13}e_{13} + c_{14}e_{14} \\ & + c_{23}e_{23} + c_{24}e_{24} + \tfrac{1}{c_{12}}(c_{13}c_{24} - c_{14}c_{23})\, e_{34}. \end{aligned}$$

Each of the first five coefficients is read off a single product: for instance the e_{23} term comes from $(-c_{23}e_3)\wedge e_2 = c_{23}e_{23}$, and the e_{13} term from $c_{12}e_1 \wedge \frac{c_{13}}{c_{12}}e_3$. The last comes from

$$(-c_{23}e_3 - c_{24}e_4) \wedge \left(\tfrac{c_{13}}{c_{12}}e_3 + \tfrac{c_{14}}{c_{12}}e_4\right) = \tfrac{c_{13}c_{24} - c_{14}c_{23}}{c_{12}}\, e_{34}.$$

By hypothesis $c_{13}c_{24} - c_{14}c_{23} = c_{12}c_{34}$, so that coefficient is c_{34} and $u \wedge w = \omega$. Hence ω is decomposable.

Exercise. C2: Characteristic two

Let $F = \mathbb{F}_2$ and $\dim V = 4$ with basis $(e_1, \dots, e_4)$.

(a) Prove that $\omega \wedge \omega = 0$ for **every** $\omega \in \Lambda^2 V$.

(b) Explain why (a) does not make every element of $\Lambda^2 V$ decomposable, and name the element and the argument that settle it.

(c) Is the statement "$\omega \wedge \omega = 0$ for every ω of odd degree" also special to characteristic 2? Justify your answer.

Solution.

(a) By the computation of Exercise (C1: The Plücker relation) (a), $\omega \wedge \omega$ is 2 times an element of $\Lambda^4 V$, and $2 = 0$ in $\mathbb{F}_2$. (Equivalently: in $\sum_{S,T} c_S c_T\, e_S \wedge e_T$ the diagonal terms vanish and the off-diagonal ones pair up, each pair summing to $2c_Sc_T\, e_S \wedge e_T = 0$.)

(b) Because $\omega \wedge \omega = 0$ is only a **necessary** condition for decomposability, and over $\mathbb{F}_2$ everything satisfies it, so it carries no information at all. The element $\omega = e_1 \wedge e_2 + e_3 \wedge e_4$ is still not decomposable, by Theorem 15.10.4, whose proof computes $v \wedge \omega$ for a general v and never divides by 2.

(c) No, that one is characteristic-free: Theorem 15.10.3 proves $\omega \wedge \omega = 0$ for odd degree over every field, by pairing off the terms with $S \neq T$ rather than by dividing. Over $\mathbb{F}_2$ the statement is true but vacuous as a test, for the same reason as in (a).

Exercise. C3: The even part of a Clifford algebra

Let $\operatorname{char} F \neq 2$, $\dim V = n \geq 1$, and let (C, ι) be a Clifford algebra of (V, q), and let $(e_1, \dots, e_n)$ be an orthogonal basis of V for β. Let $C^+ \subseteq C$ be the span of the sorted products with k **even**, and C^- the span of those with k odd. Prove that $C = C^+ \oplus C^-$, that C^+ is a subalgebra of C, and that $\dim_F C^+ = \dim_F C^- = 2^{n-1}$. Is C^- a subalgebra?

Hint: what does the sorting procedure of Proposition 14.12.3 do to the parity of the length of a product?

Solution. The sorted products form a basis of C by Theorem 15.10.8, and they are partitioned by the parity of $|S|$, so $C = C^+ \oplus C^-$ and the two dimensions are $\sum_{k \text{ even}} \binom{n}{k}$ and $\sum_{k \text{ odd}} \binom{n}{k}$. These are equal and sum to 2^n, since $\sum_k (-1)^k \binom{n}{k} = (1-1)^n = 0$ for $n \geq 1$; so each is 2^{n-1}.

For the subalgebra claim, note that the two moves in the proof of Proposition 14.12.3 — swapping two neighboring distinct factors, and replacing $\iota(e_i)^2$ by the scalar d_i — change the length of a product by 0 and by 2 respectively. Both preserve parity. So the product of two sorted products of lengths k and l is a combination of sorted products of length congruent to $k + l$ modulo 2. Hence $C^+C^+ \subseteq C^+$, and since $1_C \in C^+$ (the empty product, of length 0), C^+ is a subalgebra.

C^- is not: $1_C \notin C^-$, and $C^-C^- \subseteq C^+$ by the same parity count, so C^- fails to be closed as soon as it holds an element with nonzero square, such as $\iota(e_i)$ with $q(e_i) \neq 0$.

15.11 Tensors in Quantum Mechanics

The chapter ends outside mathematics, and the arrangement is the one Chapter 14 used for Minkowski space and for the second derivative test: physics supplies a modeling assumption, the book quotes it and proves nothing about it, and everything after it is linear algebra. The assumption is that a composite quantum system is the tensor product of its parts. Two things follow. One is a phenomenon, **entanglement**, which turns out to be nothing but §02's observation that not every element of $V \otimes W$ is a simple tensor. The other is a theorem, the **Schmidt decomposition**, which turns out to be Chapter 13's singular value decomposition wearing different notation. The second is the real payoff, and it is why this section closes Part IV: it joins the two halves of the book's later chapters at a single point.

15.11.1 What is quoted and what is proved

Quantum mechanics, as a physical theory, is not derived here and cannot be. What we take from it is a short list of modeling conventions, stated once so that the reader can see exactly where the mathematics starts.

(Q1) The states of a quantum system are the **unit vectors** of a complex inner product space V, where two unit vectors differing by a scalar of modulus 1 describe the same state.

(Q2) If a system A has state space V and a system B has state space W, then the composite system "A together with B" has state space $V \otimes W$.

(Q3) If A is in the state v and B is in the state w, then the composite is in the state $v \otimes w$.

None of (Q1)–(Q3) is a theorem, and this book does not argue for any of them. They are quoted exactly as Taylor's theorem was quoted in Chapter 14 §06 — as hypotheses that give the mathematics below something to stand on. Everything proved in this section is a statement about finite-dimensional complex inner product spaces, true whether or not (Q1)–(Q3) describe the world.

Two remarks make the list less arbitrary. First, $\otimes$ rather than $\oplus$ is visible in a dimension count: a spin has a two-dimensional state space, a **qubit**, and two spins together have a four-dimensional one, not a three-dimensional one. Of the two standard constructions, only $\otimes$ produces the

observed count, since $\dim(V \otimes W) = \dim V \cdot \dim W$ by Theorem 15.2.5. Second, (Q3) very nearly forces (Q2): the assignment $(v, w) \mapsto$ (joint state) is required to be bilinear, and by Definition 15.2.1 every bilinear map out of $V \times W$ factors uniquely through $V \otimes W$. The universal property is the precise statement that the joint state depends on the parts bilinearly and in no other way.

Genuine quantum mechanics uses infinite-dimensional spaces with a completeness condition. Everything below is finite-dimensional, which is where this book lives, and where qubits live too.

Before states can be unit vectors, $V \otimes W$ needs an inner product. There is only one reasonable candidate, and it exists.

Proposition 15.11.1. The Inner Product on a Tensor Product

Let V and W be finite-dimensional inner product spaces over $F = \mathbb{C}$ (or over $\mathbb{R}$). Then there is **exactly one** inner product on $V \otimes W$ satisfying

$$\langle v \otimes w, v' \otimes w' \rangle = \langle v, v' \rangle \, \langle w, w' \rangle$$

for all $v, v' \in V$ and $w, w' \in W$. Moreover, if $\mathcal{E}$ and $\mathcal{F}$ are orthonormal bases of V and W, then $(e_i \otimes f_j)$ is an orthonormal basis of $V \otimes W$.

Proof. *Uniqueness.* The simple tensors span $V \otimes W$ (Proposition 15.2.4), and an inner product is additive in the first slot and conjugate-additive in the second, so its values on all pairs of vectors are determined by its values on pairs of simple tensors.

Existence. Fix orthonormal bases $\mathcal{E} = (e_1, \dots, e_m)$ of V and $\mathcal{F} = (f_1, \dots, f_n)$ of W, which exist by Theorem 11.2.2. By Theorem 15.2.5 the mn vectors $e_i \otimes f_j$ form a basis of $V \otimes W$, so the rule

$$\Big\langle \sum_{i,j} c_{ij}\, e_i \otimes f_j,\ \sum_{i,j} d_{ij}\, e_i \otimes f_j \Big\rangle := \sum_{i,j} c_{ij} \overline{d_{ij}}$$

is well defined; it is the standard inner product of coordinate vectors, hence satisfies (IP1)–(IP3) of Definition 11.1.1. To check the product formula, write $v = \sum_i a_i e_i$ and $w = \sum_j b_j f_j$. Bilinearity of $\otimes$ gives $v \otimes w = \sum_{i,j} a_i b_j \, e_i \otimes f_j$, and likewise for $v' \otimes w'$, so

$$\langle v \otimes w, v' \otimes w' \rangle = \sum_{i,j} a_i b_j \overline{a'_i b'_j} = \Big(\sum_i a_i \overline{a'_i}\Big)\Big(\sum_j b_j \overline{b'_j}\Big),$$

which is $\langle v, v' \rangle \langle w, w' \rangle$ by Theorem 11.2.1. Finally, the product formula gives $\langle e_i \otimes f_j, e_k \otimes f_l \rangle = \delta_{ik}\delta_{jl}$ for **any** orthonormal $\mathcal{E}$ and $\mathcal{F}$, so the displayed construction did not depend on the pair chosen. This proves the proposition. ∎

15.11.2 Entanglement is failure of simplicity

The whole phenomenon fits in the language the chapter already has. A state of the composite is a unit vector of $V \otimes W$, and a state of the form $v \otimes w$ is one in which each part has a state of its own. Section 02 proved that not every vector of a tensor product has that form; a composite state that does not describes a system whose parts have no separate states at all.

A state is entangled when it is a vector of $V \otimes W$ that is not a simple tensor.

Definition 15.11.1. Product State, Entangled State

Let V and W be finite-dimensional complex inner product spaces, and let $z \in V \otimes W$ be a unit vector. Then z is a **product state**, also called **separable**, if

$$z = v \otimes w \quad \text{for some } v \in V,\ w \in W,$$

and z is **entangled** otherwise. The same words are used for any nonzero $z \in V \otimes W$, the unit-length condition playing no part in the distinction.

Clause by clause. **Product state** is a synonym for "simple tensor", the term of §02; only the context is new. The definition is relative to the **splitting** $V \otimes W$, not to the vector alone: "entangled" is a statement about a vector *and* a choice of two factors. And v and w need not be unit vectors, though they can be taken so: $\|v \otimes w\| = \|v\|\|w\|$ by Proposition 15.11.1, so rescaling one factor up and the other down costs nothing.

Example. The Bell state

In $\mathbb{C}^2 \otimes \mathbb{C}^2$, write $\mathcal{E} = (e_1, e_2)$ for the standard orthonormal basis of the first factor and $\mathcal{F} = (f_1, f_2)$ for the same basis regarded as sitting in the second factor, the two names being there only to keep the factors apart. Let

$$z_0 \coloneqq \tfrac{1}{\sqrt{2}}(e_1 \otimes f_1 + e_2 \otimes f_2).$$

Show that z_0 is a unit vector, and that it is entangled.

Solution. By Proposition 15.11.1 the four vectors $e_i \otimes f_j$ are orthonormal, so $\|z_0\|^2 = \tfrac{1}{2}(1+1) = 1$.

For entanglement there is nothing to prove that §02 has not already proved. Once the two bases are recognized as the same one, $e_1 \otimes f_1 + e_2 \otimes f_2$ is exactly the tensor of Example (A tensor that is not simple), and z_0 is a nonzero multiple of it: if $z_0 = v \otimes w$ then $e_1 \otimes f_1 + e_2 \otimes f_2 = (\sqrt{2}\,v) \otimes w$, which Example (A tensor that is not simple) rules out. So z_0 is entangled.

The physical reading is one paragraph long, and the book stops there. Under (Q1)–(Q3), a composite in the state z_0 is a pair of qubits of which two things are true at once: the pair has a completely specified state, and neither member has a state at all, since z_0 is not of the form $v \otimes w$. Nothing about the individual qubits is being suppressed; the postulates simply do not assign them states. That is what is meant by calling the two qubits entangled, and it is why the word attaches to a piece of linear algebra rather than to a measurement. Whether the postulates are true of the world is a question this book does not address.

Warning. **A sum of simple tensors can still be simple.** "Entangled" does not mean "written as a sum of more than one term". In $\mathbb{C}^2 \otimes \mathbb{C}^2$,

$$\tfrac{1}{2}(e_1 \otimes f_1 + e_1 \otimes f_2 + e_2 \otimes f_1 + e_2 \otimes f_2) = \tfrac{e_1+e_2}{\sqrt{2}} \otimes \tfrac{f_1+f_2}{\sqrt{2}},$$

a four-term expression for a product state. Entanglement is a property of the **vector**, not

of an expression for it, and deciding it needs an invariant. Supplying one is the job of the rest of this section.

Quick check. Is $z = \frac{1}{\sqrt{3}}(e_1 \otimes f_1 + e_1 \otimes f_2 + e_2 \otimes f_1)$ in $\mathbb{C}^2 \otimes \mathbb{C}^2$ entangled?

Solution. Yes. Suppose $z = v \otimes w$ with $v = a_1 e_1 + a_2 e_2$ and $w = b_1 f_1 + b_2 f_2$. Expanding, $v \otimes w = \sum_{i,j} a_i b_j \, e_i \otimes f_j$, and the $e_i \otimes f_j$ are a basis (Theorem 15.2.5), so comparing coefficients gives

$$a_1 b_1 = a_1 b_2 = a_2 b_1 = \tfrac{1}{\sqrt{3}}, \qquad a_2 b_2 = 0.$$

Then $(a_1 b_1)(a_2 b_2) = 0$ while $(a_1 b_2)(a_2 b_1) = \frac{1}{3}$; but both products equal $a_1 a_2 b_1 b_2$. So no such v, w exist.

15.11.3 The Schmidt decomposition

The Quick check used a 2×2 determinant on an array of coefficients, and that is a hint about where to look. Section 03 already said what the array is: by Theorem 15.3.9 a tensor **is** a linear map, and a linear map between finite-dimensional spaces is a matrix. Chapter 13 knows exactly one thing about an arbitrary matrix, and it knows it completely.

Write out the identification for the case at hand. Fix bases $\mathcal{E} = (e_1, \dots, e_m)$ of V and $\mathcal{F} = (f_1, \dots, f_n)$ of W, and let $(\varphi_1, \dots, \varphi_n)$ be the basis of W^* dual to $\mathcal{F}$ (Theorem 5.1.3). The map

$$V \times W \to \mathcal{L}(W^*, V), \qquad (v, w) \mapsto (\varphi \mapsto \varphi(w)\, v),$$

is bilinear in (v, w), so by Definition 15.2.1 it induces a linear $\Theta \colon V \otimes W \to \mathcal{L}(W^*, V)$. It sends $e_i \otimes f_j$ to the map $\varphi_k \mapsto \delta_{jk} e_i$, whose matrix from $(\varphi_1, \dots, \varphi_n)$ to $\mathcal{E}$ is the matrix unit $\mathbf{E}_{ij}$. The matrix units are a basis of $M_{m \times n}(\mathbb{C})$ and taking the matrix of a map in fixed bases is an isomorphism (Theorem 4.1.4), so Θ carries a basis to a basis and is itself an isomorphism. It is Theorem 15.3.9 with the two factors in the other order and W identified with W^{**} (Theorem 5.4.6), and for coordinate spaces it is §02's matrix model Theorem 15.2.7; the letter is shared for that reason. What matters below is the consequence: the matrix of $\Theta(z)$ is the array of coefficients of z in the basis $(e_i \otimes f_j)$.

Theorem 15.11.2. Schmidt Decomposition

Let V and W be finite-dimensional complex inner product spaces of dimensions m and n, and let $z \in V \otimes W$ be nonzero. Then there are an integer r with $1 \le r \le \min(m, n)$, real numbers

$$\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_r > 0,$$

an orthonormal list $(x_1, \dots, x_r)$ in V and an orthonormal list $(y_1, \dots, y_r)$ in W such that

$$z = \sum_{k=1}^{r} \sigma_k \, x_k \otimes y_k. \tag{15.11.1}$$

Moreover r and the numbers $\sigma_1, \dots, \sigma_r$ are determined by z: they do not depend on the

lists chosen, nor on any basis. Finally $\|z\|^2 = \sum_k \sigma_k^2$, so z is a unit vector exactly when $\sum_k \sigma_k^2 = 1$.

Idea. The statement is the singular value decomposition of Theorem 13.9.1 with the words changed, so the plan is to change the words back. ① Fix orthonormal bases on both sides; Theorem 15.2.5 turns z into a matrix $\mathbf{C}$ of coefficients, which by the paragraph above is the matrix of the linear map $\Theta(z)$. ② Feed $\mathbf{C}$ to the compact singular value decomposition, Theorem 13.9.3, which over $\mathbb{C}$ still returns **real positive** numbers σ_k and two orthonormal lists of complex column vectors. ③ Read the column vectors back as vectors of V and W; orthonormal columns become orthonormal vectors because coordinates in an orthonormal basis compute inner products. For uniqueness, run ③ backwards: any expression of the shape Eq. 15.11.1 turns into a factorization $\mathbf{C} = \mathbf{X}\mathbf{D}\mathbf{Y}^{\top}$ with $\mathbf{X}, \mathbf{Y}$ having orthonormal columns, and then $\mathbf{C}\mathbf{C}^{*} = \mathbf{X}\mathbf{D}^2\mathbf{X}^{*}$ hands over the σ_k^2 as the nonzero eigenvalues of a matrix built from z alone.

Proof. **Step 1: turn z into a matrix.** Fix orthonormal bases $\mathcal{E} = (e_1, \dots, e_m)$ of V and $\mathcal{F} = (f_1, \dots, f_n)$ of W (Theorem 11.2.2). By Theorem 15.2.5 there are unique scalars $c_{ij} \in \mathbb{C}$ with $z = \sum_{i,j} c_{ij}\, e_i \otimes f_j$. Let $\mathbf{C} \in M_{m\times n}(\mathbb{C})$ be the matrix with entries c_{ij}; it is the matrix of the linear map $\Theta(z)$ described above, and it is nonzero because z is.

Step 2: decompose the matrix. Put $r = \operatorname{rank}\mathbf{C} \geq 1$. By Theorem 13.9.3 there are orthonormal lists $(\mathbf{u}_1, \dots, \mathbf{u}_r)$ in $\mathbb{C}^m$ and $(\mathbf{v}_1, \dots, \mathbf{v}_r)$ in $\mathbb{C}^n$, and the nonzero singular values $\sigma_1 \geq \cdots \geq \sigma_r > 0$ of $\mathbf{C}$ (Definition 13.9.1), with

$$\mathbf{C} = \sum_{k=1}^{r} \sigma_k\, \mathbf{u}_k \mathbf{v}_k^{*}, \qquad \text{that is} \qquad c_{ij} = \sum_{k=1}^{r} \sigma_k (\mathbf{u}_k)_i \overline{(\mathbf{v}_k)_j}.$$

The σ_k are real and positive even though $\mathbf{C}$ is complex, since they are square roots of eigenvalues of the positive semidefinite matrix $\mathbf{C}^{*}\mathbf{C}$. Also $r = \operatorname{rank}\mathbf{C} \leq \min(m, n)$.

Step 3: read the columns back. Set

$$x_k \coloneqq \sum_{i=1}^{m} (\mathbf{u}_k)_i\, e_i \ \in V,$$
$$y_k \coloneqq \sum_{j=1}^{n} \overline{(\mathbf{v}_k)_j}\, f_j \ \in W.$$

Substituting the formula for c_{ij} into $z = \sum_{i,j} c_{ij}\, e_i \otimes f_j$ and interchanging the three finite sums, bilinearity of $\otimes$ gives

$$z = \sum_{k=1}^{r} \sigma_k \Big(\sum_i (\mathbf{u}_k)_i e_i\Big) \otimes \Big(\sum_j \overline{(\mathbf{v}_k)_j} f_j\Big) = \sum_{k=1}^{r} \sigma_k\, x_k \otimes y_k.$$

Since $\mathcal{E}$ is orthonormal, Theorem 11.2.1 gives

$$\langle x_k, x_l \rangle = \sum_i (\mathbf{u}_k)_i \overline{(\mathbf{u}_l)_i} = \langle \mathbf{u}_k, \mathbf{u}_l \rangle = \delta_{kl},$$

and likewise, since $\mathcal{F}$ is orthonormal,

$$\langle y_k, y_l\rangle = \sum_j \overline{(\mathbf{v}_k)_j}(\mathbf{v}_l)_j = \langle \mathbf{v}_l, \mathbf{v}_k\rangle = \delta_{kl}.$$

So both lists are orthonormal, and Eq. 15.11.1 holds.

Step 4: uniqueness. Suppose $z = \sum_{k=1}^{s} \tau_k\, x'_k \otimes y'_k$ is **any** expression of the stated shape, with $\tau_1 \geq \cdots \geq \tau_s > 0$ and both lists orthonormal. Keep the bases $\mathcal{E}, \mathcal{F}$ and write $x'_k = \sum_i \xi_{ik} e_i$ and $y'_k = \sum_j \eta_{jk} f_j$. Let $\mathbf{X} \in M_{m\times s}(\mathbb{C})$ have entries ξ_{ik} and $\mathbf{Y} \in M_{n\times s}(\mathbb{C})$ have entries η_{jk}. Orthonormality of the two lists says, by Theorem 11.2.1 again, that the columns of $\mathbf{X}$ and of $\mathbf{Y}$ are orthonormal, that is

$$\mathbf{X}^*\mathbf{X} = \mathbf{I}_s, \qquad \mathbf{Y}^*\mathbf{Y} = \mathbf{I}_s.$$

Expanding the expression for z in the basis $(e_i \otimes f_j)$ and comparing with the unique coefficients c_{ij} of Step 1,

$$c_{ij} = \sum_{k=1}^{s} \tau_k\, \xi_{ik}\eta_{jk}, \qquad \text{i.e.} \qquad \mathbf{C} = \mathbf{X}\mathbf{D}\mathbf{Y}^{\top},$$

where $\mathbf{D} = \operatorname{diag}(\tau_1, \ldots, \tau_s)$. Taking conjugate transposes and using $\mathbf{Y}^{\top}\overline{\mathbf{Y}} = \overline{\mathbf{Y}^*\mathbf{Y}} = \mathbf{I}_s$,

$$\mathbf{C}\mathbf{C}^* = \mathbf{X}\mathbf{D}(\mathbf{Y}^{\top}\overline{\mathbf{Y}})\mathbf{D}\mathbf{X}^* = \mathbf{X}\mathbf{D}^2\mathbf{X}^*.$$

Now read off the eigenvalues of $\mathbf{C}\mathbf{C}^* \in M_m(\mathbb{C})$. For each k, the k-th column $\mathbf{X}\mathbf{e}_k$ satisfies

$$\mathbf{C}\mathbf{C}^*(\mathbf{X}\mathbf{e}_k) = \mathbf{X}\mathbf{D}^2\mathbf{X}^*\mathbf{X}\mathbf{e}_k = \mathbf{X}\mathbf{D}^2\mathbf{e}_k = \tau_k^2\,(\mathbf{X}\mathbf{e}_k),$$

and these s columns are orthonormal. If instead $\mathbf{z}' \in \mathbb{C}^m$ is orthogonal to every column of $\mathbf{X}$ then $\mathbf{X}^*\mathbf{z}' = \mathbf{0}$, so $\mathbf{C}\mathbf{C}^*\mathbf{z}' = \mathbf{0}$. Since $\mathbb{C}^m = \operatorname{col}(\mathbf{X}) \oplus \operatorname{col}(\mathbf{X})^{\perp}$ by Theorem 11.3.2, putting an orthonormal basis of $\operatorname{col}(\mathbf{X})^{\perp}$ beside the columns of $\mathbf{X}$ produces an orthonormal basis of $\mathbb{C}^m$ consisting of eigenvectors of $\mathbf{C}\mathbf{C}^*$. Hence the eigenvalue list of $\mathbf{C}\mathbf{C}^*$, with multiplicity, is $\tau_1^2, \ldots, \tau_s^2$ together with $m - s$ zeros.

Every τ_k^2 is positive, so s is the number of nonzero eigenvalues of $\mathbf{C}\mathbf{C}^*$, and $(\tau_1^2, \ldots, \tau_s^2)$ is the list of those eigenvalues in decreasing order. Both are determined by $\mathbf{C}$, which is determined by z once $\mathcal{E}$ and $\mathcal{F}$ are fixed. So any two expressions of the shape Eq. 15.11.1 have the same s and the same τ_k; in particular they agree with the r and σ_k of Steps 2 and 3, and since the statement "z admits Eq. 15.11.1 with these r, σ_k" mentions no basis, the numbers cannot depend on $\mathcal{E}$ and $\mathcal{F}$ either.

Step 5: the norm. By Proposition 15.11.1, $\langle x_k \otimes y_k, x_l \otimes y_l\rangle = \langle x_k, x_l\rangle\langle y_k, y_l\rangle = \delta_{kl}$, so the r vectors $x_k \otimes y_k$ are orthonormal in $V \otimes W$. By Theorem 11.1.3, $\|z\|^2 = \sum_k \sigma_k^2$. This proves the theorem.

■

Definition 15.11.2. Schmidt Coefficients, Schmidt Rank

Let $z \in V \otimes W$ be nonzero. The numbers $\sigma_1 \geq \cdots \geq \sigma_r > 0$ of Theorem 15.11.2 are the **Schmidt coefficients** of z, and the integer r is the **Schmidt rank** of z. By convention the Schmidt rank of 0 is 0.

Both are well defined precisely because of Step 4, and the proof identifies them: the Schmidt coefficients are the nonzero singular values of the coefficient matrix of z in any pair of orthonormal bases, and the Schmidt rank is the rank of that matrix. That is the invariant the warning above asked for, and it is also §02's: the smallest number of simple tensors summing to z is the rank of $\Theta(z)$, and Theorem 15.11.2 now exhibits a shortest such sum, with its terms orthogonal on both sides at once.

Corollary 15.11.3. Entanglement Is Schmidt Rank at Least Two

A nonzero $z \in V \otimes W$ is a product state if and only if its Schmidt rank is 1. Hence a state is entangled exactly when its Schmidt rank is at least 2.

Proof. ($\Leftarrow$) If $r = 1$ then $z = \sigma_1 x_1 \otimes y_1 = (\sigma_1 x_1) \otimes y_1$, a simple tensor.

($\Rightarrow$) Suppose $z = v \otimes w$. Neither factor is 0, since $z \neq 0$. Put $x_1 = v/\|v\|$ and $y_1 = w/\|w\|$, which is legitimate because $\|v\|, \|w\| > 0$. Then $z = \sigma_1 x_1 \otimes y_1$ with $\sigma_1 = \|v\|\|w\| > 0$, and the one-term lists (x_1), (y_1) are orthonormal. This is an expression of the shape Eq. 15.11.1 with one term, so the uniqueness clause of Theorem 15.11.2 gives $r = 1$.

■

Warning. **The Schmidt coefficients are determined; the vectors are not.** For the Bell state z_0 of Example (The Bell state) the coefficient matrix is $\frac{1}{\sqrt{2}}\mathbf{I}_2$, whose singular values are $\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}$. Repeated coefficients leave the lists free, exactly as a repeated eigenvalue of $\mathbf{A}^*\mathbf{A}$ leaves $\mathbf{V}$ free in Theorem 13.9.2: for **every** orthonormal basis (x_1, x_2) of $\mathbb{C}^2$ one has $z_0 = \frac{1}{\sqrt{2}}(x_1 \otimes y_1 + x_2 \otimes y_2)$ for a suitable orthonormal (y_1, y_2). Never speak of *the* Schmidt vectors; speak of *the* Schmidt coefficients.

Example. A Schmidt decomposition in two by three

In $\mathbb{C}^2 \otimes \mathbb{C}^3$, find the Schmidt decomposition, the Schmidt rank, and the Schmidt coefficients of

$$z = \tfrac{1}{2}(e_1 \otimes f_1 + e_1 \otimes f_3 + e_2 \otimes f_2 - e_2 \otimes f_3),$$

where $\mathcal{E}$ and $\mathcal{F}$ are the standard orthonormal bases.

Solution. The coefficient matrix is

$$\mathbf{C} = \tfrac{1}{2}\begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & -1 \end{pmatrix} \in M_{2\times 3}(\mathbb{R}).$$

This is half the matrix of Example (A singular value decomposition of a 2 by 3 matrix), so by Exercise (B4: Determine whether) (b) its singular values are half of those found there: $\sigma_1 = \frac{\sqrt{3}}{2}$ and $\sigma_2 = \frac{1}{2}$. Both are nonzero, so $\operatorname{rank}\mathbf{C} = 2$ and the Schmidt rank is 2; by Corollary 15.11.3, z is entangled. As a check, $\sigma_1^2 + \sigma_2^2 = \frac{3}{4} + \frac{1}{4} = 1$, and indeed $\|z\|^2 = 4 \cdot \frac{1}{4} = 1$.

For the vectors, Theorem 13.9.3 asks for an orthonormal eigenbasis of the 3×3 matrix $\mathbf{C}^*\mathbf{C}$; as in Example (A singular value decomposition of a 2 by 3 matrix) it is cheaper to work with

the smaller $\mathbf{C}\mathbf{C}^{\top} = \frac{1}{4}\begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}$, whose orthonormal eigenvectors are the left singular vectors $\mathbf{u}_k$. They are $\frac{1}{\sqrt{2}}(1,-1)$ for the eigenvalue $\frac{3}{4}$ and $\frac{1}{\sqrt{2}}(1,1)$ for $\frac{1}{4}$. So $\mathbf{u}_1 = \frac{1}{\sqrt{2}}(1,-1)$ and $\mathbf{u}_2 = \frac{1}{\sqrt{2}}(1,1)$, and the relation $\mathbf{C}^{*}\mathbf{u}_i = \sigma_i \mathbf{v}_i$ of Eq. 13.9.3 gives

$$\mathbf{v}_1 = \frac{\mathbf{C}^{\top}\mathbf{u}_1}{\sigma_1} = \frac{1}{\sqrt{6}}(1,-1,2), \qquad \mathbf{v}_2 = \frac{\mathbf{C}^{\top}\mathbf{u}_2}{\sigma_2} = \frac{1}{\sqrt{2}}(1,1,0).$$

All entries are real, so the conjugation in Step 3 of Theorem 15.11.2 does nothing here, and reading the columns back gives

$$\begin{aligned} x_1 &= \tfrac{1}{\sqrt{2}}(e_1 - e_2), \quad y_1 = \tfrac{1}{\sqrt{6}}(f_1 - f_2 + 2f_3), \\ x_2 &= \tfrac{1}{\sqrt{2}}(e_1 + e_2), \quad y_2 = \tfrac{1}{\sqrt{2}}(f_1 + f_2), \end{aligned}$$

so that $z = \frac{\sqrt{3}}{2}\, x_1 \otimes y_1 + \frac{1}{2}\, x_2 \otimes y_2$. Multiplying out recovers z: the two terms are $\frac{1}{4}(e_1 - e_2) \otimes (f_1 - f_2 + 2f_3)$ and $\frac{1}{4}(e_1 + e_2) \otimes (f_1 + f_2)$, whose coefficients at $e_1 \otimes f_2$ and at $e_2 \otimes f_1$ cancel, leaving $\frac{1}{2}$ at $e_1 \otimes f_1$, $e_1 \otimes f_3$, and $e_2 \otimes f_2$ and $-\frac{1}{2}$ at $e_2 \otimes f_3$.

Remark. Truncating Eq. 15.11.1 after k terms is truncating a singular value decomposition, so Theorem 13.11.3 applies verbatim: the closest state of Schmidt rank at most k is the truncated sum, at distance $(\sum_{i>k} \sigma_i^2)^{1/2}$. Exercise (C2: Best approximation by a state of small Schmidt rank) works this out. The remark also marks the limit of the theorem: everything here concerns **two** factors, because a matrix has two indices. For $V \otimes W \otimes U$ a general vector admits no expression as a sum of simple tensors built from three orthonormal families; the book records that without proof.

15.11.4 Looking back at Part IV

Part IV asked one question in two ways. Chapter 14 took a single bilinear form and asked what it is up to a change of basis; Chapter 15 asked what all multilinear maps have in common, and answered by building the space that turns every one of them into a linear map.

The two chapters answer each other. Chapter 14 worked one form at a time and classified: congruence rather than similarity, complete the square, split off a hyperbolic plane, count signs rather than values. Every invariant it produced — the inertia of Theorem 14.5.1, the Witt index of Definition 14.7.4, the discriminant of Definition 14.2.2 — was a number attached to a form that survives a change of basis, and every classification theorem said that a short list of such numbers is all there is. Chapter 15 worked with all forms at once and constructed: define by universal property and then build a model, impose relations by a quotient, define a map out of a quotient upstairs first, and count a basis by counting index tuples. Where Chapter 14 diagonalized a symmetric form by completing the square, Chapter 15 produced the space $\operatorname{Sym}^2 V$ on which that form is a single linear functional; where Chapter 14 proved that alternating is the characteristic-free strengthening of skew, Chapter 15 built $\Lambda^k V$ by killing $v \otimes v$ for exactly that reason; and where Chapter 14 could only say that a Clifford algebra is unique if it exists, Chapter 15 built one out of $T(V)$ and settled its dimension.

This last section adds a third connection, backwards rather than sideways. The Schmidt decomposition is not a new theorem; it is Theorem 13.9.3 with $V \otimes W$ written in place of $M_{m\times n}$. Part III spent a chapter on what a matrix does to two orthonormal bases, one in each of two spaces; Part IV spent a chapter on the space in which a pair of vectors, one from each of two

spaces, becomes a single object. Those are the same statement read in two directions, and Theorem 15.3.9 is the sentence that says so. Asked years from now what the tensor product is for, a reader may answer that it is where a bilinear map becomes linear; the shortest true answer is that $V \otimes W$ is where the two sides of a matrix live.

15.11.5 Exercises

15.11.5.1 A. Check your understanding

Exercise. A1

(a) State the three modeling assumptions (Q1)–(Q3) of this section, and say which of them this book proves.

(b) Define **product state** and **entangled state**, and say what data the words depend on besides the vector itself.

(c) State the Schmidt decomposition, with every hypothesis, and say what is unique in it and what is not.

(d) Define the Schmidt rank, and give the criterion for entanglement in terms of it.

(e) True or false: a state written as a sum of three simple tensors has Schmidt rank 3. Justify your answer.

(f) Which theorem of Chapter 13 is the Schmidt decomposition, and which theorem of §03 licenses the translation?

Solution.

(a) (Q1) states are unit vectors of a complex inner product space, up to a scalar of modulus 1; (Q2) the state space of a composite system is the tensor product of the state spaces of the parts; (Q3) if the parts are in states v and w, the composite is in the state $v \otimes w$. The book proves none of them: they are quoted from physics.

(b) A unit vector $z \in V \otimes W$ is a product state if $z = v \otimes w$ for some $v \in V$ and $w \in W$, and entangled otherwise (Definition 15.11.1). The words depend on the **splitting**: the pair of factors V and W, not the vector alone.

(c) Theorem 15.11.2: for V, W finite-dimensional complex inner product spaces and $z \in V \otimes W$ nonzero, there are $r \geq 1$, reals $\sigma_1 \geq \cdots \geq \sigma_r > 0$, and orthonormal lists (x_k) in V and (y_k) in W with $z = \sum_k \sigma_k x_k \otimes y_k$. The number r and the σ_k are unique; the lists are not.

(d) The Schmidt rank is that r (Definition 15.11.2). By Corollary 15.11.3, z is entangled exactly when $r \geq 2$.

(e) False. The Schmidt rank is at most 3, but need not equal 3: the warning after Definition 15.11.1 displays a four-term sum whose Schmidt rank is 1.

(f) It is Theorem 13.9.3, the compact form of the Singular Value Decomposition Theorem 13.9.1. The translation is licensed by Theorem 15.3.9, which identifies a tensor with a linear map and hence, in bases, with a matrix.

15.11.5.2 B. Practice

Exercise. B1: Determine which are entangled

Work in $\mathbb{C}^2 \otimes \mathbb{C}^2$ with the standard orthonormal bases. Determine which of the following unit vectors are entangled. Justify your answer, and for each product state exhibit the two factors.

(a) $\frac{1}{2}(e_1 \otimes f_1 + e_1 \otimes f_2 + e_2 \otimes f_1 + e_2 \otimes f_2)$

(b) $\frac{1}{\sqrt{2}}(e_1 \otimes f_2 - e_2 \otimes f_1)$

(c) $\frac{1}{\sqrt{3}}(e_1 \otimes f_1 + e_1 \otimes f_2 + e_2 \otimes f_2)$

(d) $\frac{1}{\sqrt{10}}(e_1 \otimes f_1 + 2e_1 \otimes f_2 - e_2 \otimes f_1 - 2e_2 \otimes f_2)$

Solution. By Corollary 15.11.3 and the identification of the Schmidt rank with the rank of the coefficient matrix, it suffices to compute $\operatorname{rank} \mathbf{C}$ in each case, and for a 2×2 matrix that is decided by the determinant.

(a) $\mathbf{C} = \frac{1}{2}\left(\begin{smallmatrix} 1 & 1 \\ 1 & 1 \end{smallmatrix}\right)$, with $\det \mathbf{C} = 0$ and $\mathbf{C} \neq \mathbf{O}$, so $\operatorname{rank} \mathbf{C} = 1$: a product state. Indeed it equals $\frac{1}{\sqrt{2}}(e_1 + e_2) \otimes \frac{1}{\sqrt{2}}(f_1 + f_2)$.

(b) $\mathbf{C} = \frac{1}{\sqrt{2}}\left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)$, with $\det \mathbf{C} = \frac{1}{2} \neq 0$, so $\operatorname{rank} \mathbf{C} = 2$: entangled.

(c) $\mathbf{C} = \frac{1}{\sqrt{3}}\left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$, with $\det \mathbf{C} = \frac{1}{3} \neq 0$, so $\operatorname{rank} \mathbf{C} = 2$: entangled.

(d) $\mathbf{C} = \frac{1}{\sqrt{10}}\left(\begin{smallmatrix} 1 & 2 \\ -1 & -2 \end{smallmatrix}\right)$, with $\det \mathbf{C} = 0$ and $\mathbf{C} \neq \mathbf{O}$, so $\operatorname{rank} \mathbf{C} = 1$: a product state, equal to $\frac{1}{\sqrt{2}}(e_1 - e_2) \otimes \frac{1}{\sqrt{5}}(f_1 + 2f_2)$.

Exercise. B2: A Schmidt decomposition

In $\mathbb{C}^2 \otimes \mathbb{C}^2$, let

$$z = \frac{1}{\sqrt{10}}(e_1 \otimes f_1 + 2e_1 \otimes f_2 + 2e_2 \otimes f_1 + e_2 \otimes f_2).$$

Find the Schmidt coefficients, the Schmidt rank, and a Schmidt decomposition of z, and verify that $\sum_k \sigma_k^2 = 1$.

Solution. The coefficient matrix is $\mathbf{C} = \frac{1}{\sqrt{10}}\left(\begin{smallmatrix} 1 & 2 \\ 2 & 1 \end{smallmatrix}\right)$, which is real symmetric. Then

$$\mathbf{C}\mathbf{C}^{\top} = \frac{1}{10}\begin{pmatrix} 5 & 4 \\ 4 & 5 \end{pmatrix},$$

with characteristic polynomial $x^2 - x + \frac{9}{100}$ and eigenvalues $\frac{9}{10}$ and $\frac{1}{10}$. So the Schmidt coefficients are $\sigma_1 = \frac{3}{\sqrt{10}}$ and $\sigma_2 = \frac{1}{\sqrt{10}}$, the Schmidt rank is 2, and $\sigma_1^2 + \sigma_2^2 = \frac{9}{10} + \frac{1}{10} = 1$.

Eigenvectors of $\mathbf{C}\mathbf{C}^{\top}$ are $\mathbf{u}_1 = \frac{1}{\sqrt{2}}(1, 1)$ for $\frac{9}{10}$ and $\mathbf{u}_2 = \frac{1}{\sqrt{2}}(1, -1)$ for $\frac{1}{10}$. Then

$$\mathbf{v}_1 = \frac{\mathbf{C}^{\top}\mathbf{u}_1}{\sigma_1} = \frac{1}{\sqrt{2}}(1, 1), \qquad \mathbf{v}_2 = \frac{\mathbf{C}^{\top}\mathbf{u}_2}{\sigma_2} = \frac{1}{\sqrt{2}}(-1, 1).$$

All entries are real, so reading the columns back as in Step 3 of Theorem 15.11.2,

$$x_1 = \tfrac{e_1+e_2}{\sqrt{2}}, \quad y_1 = \tfrac{f_1+f_2}{\sqrt{2}},$$
$$x_2 = \tfrac{e_1-e_2}{\sqrt{2}}, \quad y_2 = \tfrac{-f_1+f_2}{\sqrt{2}},$$

and $z = \tfrac{3}{\sqrt{10}}\,x_1 \otimes y_1 + \tfrac{1}{\sqrt{10}}\,x_2 \otimes y_2$. Expanding the two terms gives $\tfrac{1}{2\sqrt{10}}(3-1)e_1 \otimes f_1$ and similarly for the other three coefficients, recovering z.

Exercise. B3: How large can the Schmidt rank be?

Let V and W be complex inner product spaces with $\dim V = m$ and $\dim W = n$.

(a) Prove that every nonzero $z \in V \otimes W$ has Schmidt rank at most $\min(m, n)$.

(b) Give, for each r with $1 \le r \le \min(m, n)$, a unit vector of Schmidt rank exactly r.

Solution.

(a) Fix orthonormal bases and let $\mathbf{C} \in M_{m\times n}(\mathbb{C})$ be the coefficient matrix of z. By Theorem 15.11.2 the Schmidt rank is $\operatorname{rank}\mathbf{C}$, and the rank of an $m \times n$ matrix is at most $\min(m, n)$ because $\operatorname{rank}\mathbf{C} = \dim\operatorname{col}(\mathbf{C}) \le m$ and $\operatorname{rank}\mathbf{C} = \dim\operatorname{row}(\mathbf{C}) \le n$.

(b) Take $z_r = \tfrac{1}{\sqrt{r}}\sum_{k=1}^{r} e_k \otimes f_k$. The lists $(e_1, \dots, e_r)$ and $(f_1, \dots, f_r)$ are orthonormal and every coefficient $1/\sqrt{r}$ is positive, so this is already an expression of the shape Eq. 15.11.1; by the uniqueness clause the Schmidt rank is r. It is a unit vector since $r \cdot (1/\sqrt{r})^2 = 1$.

15.11.5.3 C. Going deeper

Exercise. C1: Acting on one part at a time

Let $S \in \mathcal{L}(V)$ and $T \in \mathcal{L}(W)$ be unitary, and let $z \in V \otimes W$ be nonzero.

(a) Prove that $(S \otimes T)(z)$ has the same Schmidt coefficients, and hence the same Schmidt rank, as z.

(b) Hence deduce that no operator of the form $S \otimes T$ with S, T unitary carries the Bell state z_0 of Example (The Bell state) to a product state.

Hint for (a): apply $S \otimes T$ to a Schmidt decomposition.

Solution.

(a) Write $z = \sum_{k=1}^{r} \sigma_k\, x_k \otimes y_k$ as in Theorem 15.11.2. By Definition 15.3.1, $(S \otimes T)(x_k \otimes y_k) = Sx_k \otimes Ty_k$, so by linearity

$$(S \otimes T)(z) = \sum_{k=1}^{r} \sigma_k\, (Sx_k) \otimes (Ty_k).$$

A unitary operator preserves inner products (Theorem 11.7.3), so $\langle Sx_k, Sx_l\rangle = \langle x_k, x_l\rangle = \delta_{kl}$ and likewise for T. Both lists are therefore orthonormal, the σ_k are

still positive and decreasing, and the display is an expression of the shape Eq. 15.11.1 for $(S \otimes T)(z)$. By the uniqueness clause of Theorem 15.11.2, its Schmidt coefficients are $\sigma_1, \dots, \sigma_r$ and its Schmidt rank is r.

(b) The Schmidt rank of z_0 is 2, as computed in the warning after Corollary 15.11.3. By (a), $(S \otimes T)(z_0)$ also has Schmidt rank 2, and by Corollary 15.11.3 a product state has Schmidt rank 1. So $(S \otimes T)(z_0)$ is entangled.

Exercise. C2: Best approximation by a state of small Schmidt rank

Let $z \in V \otimes W$ be nonzero with Schmidt decomposition $z = \sum_{k=1}^{r} \sigma_k\, x_k \otimes y_k$, and let $0 \le k \le r$. Put $z_{(k)} = \sum_{i=1}^{k} \sigma_i\, x_i \otimes y_i$.

(a) Fix orthonormal bases of V and W and let $\mathbf{C}$ be the coefficient matrix of z. Prove that $\|z\| = \|\mathbf{C}\|_F$, and that the Schmidt rank of a vector equals the rank of its coefficient matrix.

(b) Hence deduce that for every $z' \in V \otimes W$ of Schmidt rank at most k,

$$\|z - z'\| \;\ge\; \Big(\sum_{i>k} \sigma_i^2\Big)^{1/2},$$

with equality for $z' = z_{(k)}$.

Solution.

(a) By Proposition 15.11.1 the basis $(e_i \otimes f_j)$ of $V \otimes W$ is orthonormal, so $\|z\|^2 = \sum_{i,j} |c_{ij}|^2$ by Theorem 11.9.2, and that sum is $\|\mathbf{C}\|_F^2$ by the definition of the Frobenius norm. For the second claim, Step 2 of Theorem 15.11.2 produced the Schmidt coefficients as the nonzero singular values of $\mathbf{C}$, and Theorem 13.9.1 says there are $\operatorname{rank}\mathbf{C}$ of them.

(b) The map $z \mapsto \mathbf{C}$ is the coordinate isomorphism attached to the basis $(e_i \otimes f_j)$ (Corollary 2.5.5), with the mn coordinates arranged in a rectangle, so it is linear and bijective; and by (a) it carries $\|\cdot\|$ to $\|\cdot\|_F$ and Schmidt rank to matrix rank. Let $\mathbf{B}$ be the coefficient matrix of z', so $\operatorname{rank}\mathbf{B} \le k$. By Theorem 13.11.3,

$$\|z - z'\| = \|\mathbf{C} - \mathbf{B}\|_F \;\ge\; \Big(\sum_{i>k} \sigma_i^2\Big)^{1/2},$$

since the nonzero σ_i are the singular values of $\mathbf{C}$. Running Step 3 of Theorem 15.11.2 backwards turns $z_{(k)}$ into the rank-k truncation $\mathbf{C}_k$ (Definition 13.11.1) of a compact singular value decomposition of $\mathbf{C}$ (Theorem 13.9.3), and Theorem 13.11.3 attains equality there.

Exercise. C3: Does the field matter?

Let V and W be **real** inner product spaces of dimensions m and n, with orthonormal bases $\mathcal{E}$ and $\mathcal{F}$, and let $z \in V \otimes W$ have coefficient matrix $\mathbf{C} \in M_{m\times n}(\mathbb{R})$. Let $z_{\mathbb{C}} \in \mathbb{C}^m \otimes \mathbb{C}^n$ be the vector whose coefficient matrix in the standard orthonormal bases is the same $\mathbf{C}$.

(a) Prove that Theorem 15.11.2 holds verbatim over $\mathbb{R}$, with orthonormal lists of real

vectors.

(b) Prove that z and $z_{\mathbb{C}}$ have the same Schmidt coefficients and the same Schmidt rank.

(c) Is the same true of the Schmidt **vectors**? Justify your answer.

Solution.

(a) Every step of the proof runs over $F = \mathbb{R}$. Theorem 11.2.2, Theorem 15.2.5, Theorem 13.9.3, Theorem 11.2.1, and Theorem 11.3.2 are all stated for $F = \mathbb{R}$ or $\mathbb{C}$, and over $\mathbb{R}$ the conjugations in Step 3 and Step 4 are the identity, so $y_k = \sum_j (\mathbf{v}_k)_j f_j$ and $\mathbf{C}\mathbf{C}^{\top} = \mathbf{X}\mathbf{D}^2\mathbf{X}^{\top}$.

(b) By Step 4 of Theorem 15.11.2, read over each field in turn, the Schmidt coefficients of z and of $z_{\mathbb{C}}$ are both times the positive square roots of the nonzero eigenvalues of $\mathbf{C}\mathbf{C}^{\top}$, and both Schmidt ranks equal $\operatorname{rank}\mathbf{C}$. This is one and the same real symmetric matrix with one and the same characteristic polynomial, so the eigenvalues agree; and $\operatorname{rank}\mathbf{C}$ is unchanged because row reduction over $\mathbb{R}$ is row reduction over $\mathbb{C}$.

(c) No. The warning after Corollary 15.11.3 says why: a repeated Schmidt coefficient leaves the vectors free, and over $\mathbb{C}$ there is more freedom to use. For z_0 the orthonormal pair $x_1 = \frac{1}{\sqrt{2}}(e_1 + ie_2)$, $x_2 = \frac{1}{\sqrt{2}}(e_1 - ie_2)$ occurs in a complex Schmidt decomposition — with $y_1 = \frac{1}{\sqrt{2}}(f_1 - if_2)$ and $y_2 = \frac{1}{\sqrt{2}}(f_1 + if_2)$ — and has no real counterpart. What is invariant is the coefficient list, not the vectors.

Chapter 16

Norms and Matrix Analysis

Nine chapters have been writing promissory notes. Chapter 9 asked what happens to $\mathbf{A}^m$ and Chapter 10 §10 answered exactly, entry by entry, with no number attached to the answer: Theorem 10.10.2 says *whether* the powers die out, never *how fast*. Chapter 10 §09 built $e^{\mathbf{A}}$ out of a series (Definition 10.9.2) and promised that the estimates would come once there was something to estimate with. Chapter 11 solved the least-squares problem (Theorem 11.4.2) and then, at Theorem 11.8.1, recommended one algorithm over another for a reason it could not yet state. Chapter 12 §01 proved that every matrix has a diagonalizable matrix arbitrarily close to it (Corollary 12.1.6) and left the word "close" doing work no definition supported. Chapter 13 §09 met a number, the largest stretching factor of a matrix, and said that this chapter would give it a name. Every one of those debts is the same debt: the book has had no way to say **how big** a vector or a matrix is, except through one particular inner product.

A **norm** is that way. It keeps exactly what measurement needs — a length is zero only for the zero vector, it scales with the scalar, and a detour is never shorter than the direct route — and it throws away the inner product that Chapter 11 used to produce it. There are many norms, and the first surprise is that in finite dimension it does not matter which one you pick: any two of them are trapped between constant multiples of each other, so they agree about every limit, every bounded set, and every continuous function. The second surprise is that the constants cannot be made independent of the dimension, and that the whole theorem collapses in infinite dimension. From there the chapter measures matrices, and then uses the measurements: the spectral radius as the infimum of all the norms, a geometric series that inverts $\mathbf{I}-\mathbf{A}$, quantitative bounds for the exponential, the continuity of eigenvalues, and finally the condition number, which is the exact constant in the sentence "a small change in the data produces a small change in the answer".

The field in this chapter is $\mathbb{R}$ or $\mathbb{C}$. A norm returns a real number and its homogeneity clause uses $|c|$, so the scalars must have an absolute value; no other field in this book has one. This is the same restriction Chapters 11 to 13 worked under, and for the same reason.

16.0.1 The analysis this chapter imports

This chapter leans on analysis harder than any before it, and the book's practice is to name what it borrows rather than let it pass unnoticed — as Chapter 14 §06 did with Taylor's theorem, and Chapter 11 with the theorem of Abel and Ruffini. Here is the complete list. Each item is standard, each is quoted without proof, and **each is named again at the point where it is used**.

(A1) *(Completeness.)* Every nonempty set of real numbers that is bounded above has a least upper bound; equivalently, every Cauchy sequence of real or complex numbers converges.

One standard consequence is used as well: every real number is the limit of a sequence of rational numbers.

(A2) *(Monotone convergence.)* An increasing sequence of real numbers that is bounded above converges, and its limit is its least upper bound.

(A3) *(Compactness of the closed unit ball.)* In $\mathbb{R}^n$ and in $\mathbb{C}^n$ with the Euclidean norm, every sequence lying in a closed bounded set has a subsequence converging to a point **of that set**. In particular the closed unit ball and the unit sphere are compact.

(A4) *(Extreme value theorem.)* A continuous real-valued function on a nonempty compact set attains a maximum value and a minimum value; the bounds are not merely approached, they are reached.

(A5) *(Intermediate value theorem.)* A continuous real-valued function on an interval takes every value between any two of its values.

(A6) *(Mean value theorem.)* A real-valued function differentiable on an interval satisfies $f(b) - f(a) = f'(c)(b - a)$ for some c strictly between a and b. Two consequences are used: a function with negative derivative throughout an interval is strictly decreasing there, and a function whose derivative is positive to the left of a point and negative to its right attains its maximum at that point.

Alongside these, the **elementary algebra of limits** is used throughout without further comment: limits respect sums, products, and quotients with nonvanishing denominator; sums, products, and composites of continuous functions are continuous; and a nonstrict inequality survives a limit, so that $a_k \le M$ for every k forces $\lim a_k \le M$. Chapter 10 §09 quoted the same package as its fact (A3). What the list above records is everything beyond that.

Several of these speak of **continuity**, which the book has not needed until now, so fix the meaning here. A function f defined on a subset of a normed space is **continuous at** x when $\|x_k - x\| \to 0$ forces $f(x_k) \to f(x)$ for every sequence (x_k) in its domain, and **continuous** when it is continuous at every point of its domain. This is the sequential form, matching the way (A3) is stated. Every continuity claim in this chapter is established the same cheap way: by exhibiting a constant M with $|f(x) - f(y)| \le M\|x - y\|$, which gives the implication at once. §02 proves the first such bound.

Nothing else analytic enters silently. Where a section needs a fact outside this list — the differentiability of the entries of a matrix-valued function, or the continuous dependence of the roots of a polynomial on its coefficients — it is recorded as an assumption in the same manner, with a counterexample proved where one is available, and never used as a step in a proof that pretends to be self-contained.

16.0.2 What you need

The chapter is built on results the book already has. Gathered by source, so that a reader can go back to one place at a time:

- **Chapter 0 §06 and §09** — two small facts, each used once: the triangle inequality for complex numbers, which §07 needs before it can bound a sum of moduli, and the cyclic invariance of the trace, which §06 uses on the Frobenius inner product: Theorem 0.6.4, Theorem 0.9.6.
- **Chapter 2, with Chapter 1 for the words** — bases, dimension, and coordinates: Definition 1.5.1, Definition 1.6.1, Definition 1.5.2, Theorem 1.5.2, Corollary 2.5.5. What the proofs cite is the coordinate isomorphism; **basis** and **dimension** are vocabulary the chapter assumes rather than results it leans on.
- **Chapter 11 §01 and §07** — the inner product and the norm it induces, with their

three properties and the two inequalities that follow: Definition 11.1.1, Definition 11.1.3, Theorem 11.1.2, Theorem 11.1.5, Corollary 11.1.6, together with the two exercises that recover an inner product from its norm, Exercise (C1: The parallelogram law characterizes inner-product norms) and Exercise (C2: Recovering the inner product from the norm); and, from §07, the Frobenius norm and its unitary invariance, Lemma 11.7.5.

- **Chapter 11 §04 and §08** — least squares and the **QR** factorization whose stability claim was deferred: Theorem 11.4.2, Corollary 11.4.4, Theorem 11.8.1.
- **Chapter 10 §10** — the whole convergence apparatus that this chapter re-proves with numbers attached: Definition 10.10.1, Lemma 10.10.1, Theorem 10.10.2, Theorem 10.10.3, Theorem 10.10.4, Theorem 10.10.6, Corollary 10.10.5.
- **Chapter 10 §09 and §03** — the matrix exponential and the Jordan form: Definition 10.9.2, Theorem 10.9.3, Theorem 10.9.4, Theorem 10.3.1.
- **Chapter 7** — determinants, and only in §§06 and 07: the Leibniz formula, minors and cofactors, Laplace expansion, and the adjugate identity, which between them make the entries of an inverse polynomial in the entries of the matrix; and the characteristic polynomial: Theorem 7.4.3, Definition 7.6.1, Theorem 7.6.1, Definition 7.6.2, Theorem 7.6.2, Theorem 7.10.2, Theorem 7.10.3.
- **Chapter 9** — where the spectral radius comes from: the characterizations of an eigenvalue and the algebraic multiplicity, the eigenvalues on the diagonal of a triangular form, and the spectral mapping theorem: Theorem 9.2.1, Definition 9.3.1, Theorem 9.7.4, Theorem 9.10.1.
- **Chapter 12** — Schur triangularization with its density corollary, and the defect the Frobenius norm measures: Theorem 12.1.2, Corollary 12.1.6, Theorem 12.11.1, Definition 12.11.1.
- **Chapter 13** — the singular value decomposition and the two constructions built on it: Theorem 13.9.1, Definition 13.9.1, Definition 13.12.1.

16.0.3 Roadmap

- **Norms.** The three axioms, the 1-, 2-, and ∞-norms with their unit balls drawn, and the test that decides which norms come from an inner product: the parallelogram law. This pays off a promise made in Chapter 11 §01.
- **All norms agree.** On a finite-dimensional space any two norms are trapped between constant multiples of each other. Compactness enters the book here, and the payoff is that Chapter 10 §10's entrywise convergence is simply *the* convergence, whatever norm one prefers.
- **Measuring a matrix.** Submultiplicativity, and the operator norm $\|\mathbf{A}\| = \max\{\|\mathbf{Ax}\| : \|\mathbf{x}\| = 1\}$ induced by a vector norm. Its formulas for $p = 1, 2, \infty$, including $\|\mathbf{A}\|_2 = \sigma_1(\mathbf{A})$, which is the name Chapter 13 §09 asked for.
- **The spectral radius.** $\rho(\mathbf{A}) \le \|\mathbf{A}\|$ for every matrix norm, a norm that comes within ε of $\rho(\mathbf{A})$, and Gelfand's formula $\rho(\mathbf{A}) = \lim \|\mathbf{A}^k\|^{1/k}$.
- **Inverting a perturbation of the identity.** The Neumann series, the openness of the invertible matrices, and the first perturbation estimate.
- **The exponential, with bounds; and derivatives.** $\|e^{\mathbf{A}}\| \le e^{\|\mathbf{A}\|}$, decay when the spectrum sits in the left half plane, and the derivatives of the determinant and of the inverse.
- **Eigenvalues move continuously.** Continuously as a multiset, and no better: a Jordan block's eigenvalues move like $\varepsilon^{1/k}$.
- **How wrong can the answer be?** The condition number $\kappa(\mathbf{A}) = \|\mathbf{A}\|\|\mathbf{A}^{-1}\|$, what it predicts about solving $\mathbf{Ax} = \mathbf{b}$, and why the normal equations square it.

16.0.4 Named moves

- **Compactness turns "for all unit vectors" into a maximum.** A supremum over the unit sphere is hard to control; a maximum is attained at a vector you can name and feed back into the argument. Both the equivalence of norms and the very existence of the operator norm are this one move.
- **Ask for submultiplicativity.** A matrix is a vector, so a matrix norm is a vector norm; what makes it useful is the extra clause $\|\mathbf{AB}\| \le \|\mathbf{A}\|\|\mathbf{B}\|$, which is the only reason a bound on $\|\mathbf{A}\|$ says anything about $\mathbf{A}^k$. Test every candidate against it, and expect natural-looking candidates to fail.
- **Bound the tail of a series.** Neumann, the exponential, and Gelfand all reduce to estimating $\sum_{k\ge K}\|\mathbf{A}\|^k$ or $\sum_{k\ge K}\|\mathbf{A}\|^k/k!$ by something that visibly tends to zero.
- **Perturb and compare.** Replace $\mathbf{A}$ by $\mathbf{A}+\mathbf{E}$, subtract the two conclusions, and read the result as a bound on the relative change in the answer by a constant times the relative change in the data. The constant is the condition number.
- **Reduce to the 2-norm.** Equivalence of norms means any single norm may be used as the hub: prove the estimate for the Euclidean norm, where the inner product and the singular value decomposition are available, and transfer.

16.1 Norms

Chapter 11 measured vectors, and measured them in exactly one way: $\|v\| = \sqrt{\langle v, v\rangle}$. Every length in Chapters 11 to 13 came out of that formula. This section keeps the measuring and throws the inner product away. The reason is not economy. The quantities this part of the book owes — how fast a sequence of matrices settles down, how far a solution moves when the data moves — are often measured most naturally by a size that no inner product produces, and the arguments that use them turn out to need only three properties of a length.

The field here is $\mathbb{R}$ or $\mathbb{C}$, as it is throughout the chapter. A length is a real number, and the scaling rule below multiplies it by $|c|$; so the scalars must carry an absolute value, and these are the only fields in this book that do.

16.1.1 What measurement actually needs

Look back at how $\|\cdot\|$ was used after Chapter 11 defined it. Cauchy–Schwarz needed the inner product. So did Pythagoras, orthogonality, and the adjoint. But a long list of arguments did not: that a convergent sequence is bounded, that $\|\mathbf{Ax}\|$ is small when $\mathbf{x}$ is small, that a limit of unit vectors is a unit vector, that $\|u-w\| \le \|u-v\|+\|v-w\|$. Every one of those used only that $\|v\|$ vanishes just for $v=0$, that $\|cv\| = |c|\|v\|$, and the triangle inequality of Corollary 11.1.6. Three properties, used over and over. Following the book's habit, we give the three of them a name and see what else satisfies them.

A norm is a length: zero only at the zero vector, scaling with the scalar, and never rewarding a detour.

Definition 16.1.1. Norm

Let F be $\mathbb{R}$ or $\mathbb{C}$ and let V be a vector space **over** F. A **norm** on V is a function

$$\|\cdot\| : V \to \mathbb{R}$$

such that for **all** $u, v \in V$ and **all** $c \in F$:

(N1) $\|v\| \geq 0$, and $\|v\| = 0$ **only** for $v = 0$ (**positive definiteness**);
(N2) $\|cv\| = |c|\,\|v\|$ (**absolute homogeneity**);
(N3) $\|u + v\| \leq \|u\| + \|v\|$ (**the triangle inequality**).

A vector space carrying a norm is a **normed space**. A vector with $\|v\| = 1$ is a **unit vector**, and replacing a nonzero v by $v/\|v\|$ is **normalizing** it.

In words, clause by clause. (N1) says the output is a genuine nonnegative size, and that only the zero vector has size zero; the word **only** is the whole content, since the reverse implication is free (take $c = 0$ in (N2) to get $\|0\| = 0 \cdot \|0\| = 0$). (N2) says that stretching a vector by c stretches its length by $|c|$ — the **absolute value** of c, so that $\|-v\| = \|v\|$ and, over $\mathbb{C}$, $\|iv\| = \|v\|$. (N3) says that going from 0 to $u + v$ directly is at most as long as going by way of u. Note what the definition does not mention: no second vector is paired with the first, no angle appears, and there is no inner product anywhere.

Two remarks on the shape of the definition. First, the codomain is $\mathbb{R}$ even when $F = \mathbb{C}$: a length is never complex. Second, the three clauses are not independent of the arithmetic of $\mathbb{R}$ — (N3) is an inequality, and inequalities are why the field had to be restricted.

Example. The 1-, 2-, and ∞-norms on F^n

For $\mathbf{x} = (x_1, \ldots, x_n) \in F^n$ put

$$\begin{aligned} \|\mathbf{x}\|_1 &= |x_1| + \cdots + |x_n|, \\ \|\mathbf{x}\|_2 &= \left(|x_1|^2 + \cdots + |x_n|^2\right)^{1/2}, \\ \|\mathbf{x}\|_\infty &= \max\{|x_1|, \ldots, |x_n|\}. \end{aligned}$$

Check each of them against Definition 16.1.1.

Solution. (N1) holds for all three: each is a sum, a square root of a sum, or a maximum of the nonnegative numbers $|x_i|$, so each is ≥ 0; and each vanishes only when every $|x_i| = 0$, that is, only for $\mathbf{x} = \mathbf{0}$. (N2) holds because $|cx_i| = |c||x_i|$ for every i, and the common factor $|c|$ comes out of a sum, out of a square root of a sum of squares, and out of a maximum of nonnegative numbers alike.

For (N3), the 1-norm inherits it one coordinate at a time: $|x_i + y_i| \leq |x_i| + |y_i|$ for each i by the triangle inequality for the modulus (Theorem 0.6.4, which covers $F = \mathbb{R}$ as well), and summing over i gives $\|\mathbf{x} + \mathbf{y}\|_1 \leq \|\mathbf{x}\|_1 + \|\mathbf{y}\|_1$. The ∞-norm inherits it too, but from one coordinate rather than all: for each i,

$$|x_i + y_i| \leq |x_i| + |y_i| \leq \|\mathbf{x}\|_\infty + \|\mathbf{y}\|_\infty,$$

and a bound valid for every i bounds the maximum. The 2-norm is the norm induced by the standard inner product on F^n (Definition 11.1.3), so its triangle inequality is Corollary 11.1.6 and needs no fresh argument.

So one space carries at least three different norms. For $\mathbf{x} = (3, -4, 12) \in \mathbb{R}^3$ they give $\|\mathbf{x}\|_1 = 19$, $\|\mathbf{x}\|_2 = 13$, and $\|\mathbf{x}\|_\infty = 12$ — three honest answers to "how big is $\mathbf{x}$", and no reason yet to

prefer one.

Remark. There is a p-norm $\|\mathbf{x}\|_p = \left(\sum_i |x_i|^p\right)^{1/p}$ for every real $p \geq 1$, and $\|\mathbf{x}\|_\infty$ is its limit as $p \to \infty$. Its triangle inequality is Minkowski's inequality, which Chapter 18 proves from Hölder's inequality (and which is the first place real powers t^p and the logarithm are set up properly). This chapter uses only $p = 1, 2, \infty$, and each of those was verified above from scratch.

Example. Inner products, weights, and a function space

Three further families.

(a) **Every inner product gives a norm.** If V is an inner product space, then $\|v\| = \sqrt{\langle v, v\rangle}$ satisfies (N1) and (N2) by Theorem 11.1.2 (a), (b) and (N3) by Corollary 11.1.6. In particular the Frobenius norm $\|\mathbf{A}\|_F$ on $M_{m\times n}(F)$ and the norm $\left(\int_0^1 |f|^2\right)^{1/2}$ on $C[0,1]$ are norms.

(b) **Weighted norms.** Fix real weights $w_1, \ldots, w_n > 0$ and put $\|\mathbf{x}\|_w = \max_i w_i|x_i|$. All three clauses go through exactly as for $\|\cdot\|_\infty$, the weights riding along untouched; positivity of every w_i is what (N1) needs.

(c) **The sup norm.** On $C[0,1]$, the real continuous functions on $[0,1]$, put $\|f\|_\infty = \max_{t\in[0,1]} |f(t)|$.

For (c) the maximum has to exist before the formula defines anything, and that is the first place this chapter borrows: $[0,1]$ is compact and $|f|$ is continuous, so by **the extreme value theorem, fact (A4) of the chapter introduction**, the maximum is attained. Granting that, (N1) says a continuous function with $\max|f| = 0$ is the zero function, which is true pointwise; (N2) and (N3) follow from the corresponding facts about $|\cdot|$ at each point, exactly as for $\|\cdot\|_\infty$ on F^n. Note that $C[0,1]$ is infinite-dimensional — it contains all the polynomials — and it is kept in view deliberately, because §02's main theorem will fail on it.

A degenerate case is worth one line. On $V = F$ itself, $\|c\| = |c|$ is a norm, and (N2) forces every norm on F to be $\|c\| = |c|\|1\|$, a positive multiple of that one. On the zero space $V = \{0\}$ the only function to $\mathbb{R}$ is $\|0\| = 0$, and it is a norm. Both matter: they say that the theory is not empty at the bottom, and the first of them is the model for the general statement that any two norms differ by bounded factors.

Example. Three near misses

Change one feature of an example above and watch a single clause break.

(a) **Drop the square root.** On F^n put $N(\mathbf{x}) = |x_1|^2 + \cdots + |x_n|^2$. Then (N1) still holds. But $N(c\mathbf{x}) = |c|^2 N(\mathbf{x})$, so (N2) fails as soon as $|c| \neq 0, 1$: for $\mathbf{x} = \mathbf{e}_1$ and $c = 2$ it gives $N(2\mathbf{e}_1) = 4 \neq 2 = |2|N(\mathbf{e}_1)$. The square root in $\|\cdot\|_2$ is there for exactly this reason.

(b) **Forget a coordinate.** On F^2 put $N(\mathbf{x}) = |x_1|$. Then (N2) and (N3) both hold, inherited from $|\cdot|$ on F. What fails is the word **only** in (N1): $N(\mathbf{e}_2) = 0$ while $\mathbf{e}_2 \neq \mathbf{0}$. A function satisfying (N1) with "only" deleted, together with (N2) and (N3), is called a **seminorm**; it measures a quotient of V rather than V.

(c) **Bend the exponent the other way.** On $\mathbb{R}^2$ put $N(\mathbf{x}) = \left(|x_1|^{1/2} + |x_2|^{1/2}\right)^2$. Then (N1) holds, and (N2) holds because $|c|^{1/2}$ factors out of the inner sum and is squared

back to $|c|$. Only (N3) fails: $N(\mathbf{e}_1) = N(\mathbf{e}_2) = 1$ but $N(\mathbf{e}_1 + \mathbf{e}_2) = (1+1)^2 = 4 > 2$.

Item (c) is the one to remember, because it shows the three clauses are genuinely independent: two of them can hold exactly while the third fails, and it is the failing one that carries all the geometry.

Warning. **A norm is not determined by what it does to the standard basis.** Both $\|\cdot\|_1$ and $\|\cdot\|_\infty$ give every $\mathbf{e}_i$ length 1, and so does $\|\cdot\|_2$, yet the three disagree everywhere else: on $\mathbf{1} = (1, \dots, 1)$ they return n, 1, and $\sqrt{n}$. Knowing the lengths of a basis pins down nothing, and this is worth holding on to, because §02's theorem is sometimes misremembered as saying that norms which agree on a basis agree everywhere. It says no such thing.

One small consequence of (N3) will be used twice below and again in §02, so it gets a name here. It is Exercise (B4: A lower bound for a difference of norms) with the inner product removed from the proof.

Lemma 16.1.1. Reverse Triangle Inequality

Let $\|\cdot\|$ be a norm on V. Then for all $u, v \in V$,

$$\big|\, \|u\| - \|v\| \,\big| \le \|u - v\|.$$

Proof. By (N3) applied to $u = (u-v)+v$, we have $\|u\| \le \|u-v\| + \|v\|$, so $\|u\| - \|v\| \le \|u-v\|$. Swapping the roles of u and v gives $\|v\| - \|u\| \le \|v-u\|$, and $\|v-u\| = |-1|\|u-v\| = \|u-v\|$ by (N2). The number $|\|u\| - \|v\||$ is one of the two quantities just bounded, so it is at most $\|u - v\|$. This proves the inequality. ∎

Read with v replaced by v', it says that the function $\|\cdot\|$ changes by at most $\|u - v\|$ when its argument moves from u to v. That is a Lipschitz estimate, and it is the reason a norm is continuous with respect to itself — a fact §02 turns into the main theorem.

16.1.2 Unit balls

The quickest way to see the difference between two norms is to draw the set of vectors they call short.

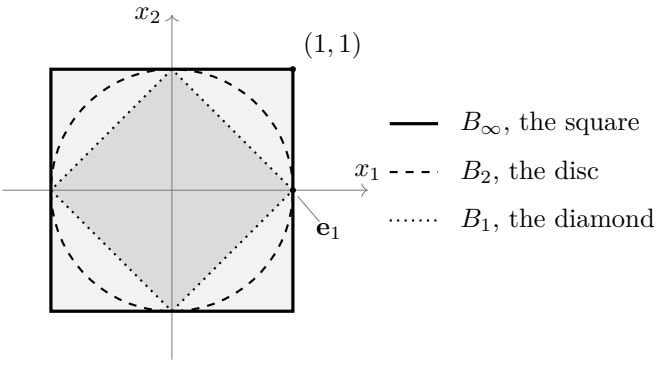
Definition 16.1.2. Closed Unit Ball

Let $\|\cdot\|$ be a norm on V. Its **closed unit ball** is

$$B_{\|\cdot\|} = \{v \in V : \|v\| \le 1\}.$$

In F^n we write B_1, B_2, and B_∞ for the balls of Example (The 1-, 2-, and ∞-norms on F^n).

In $\mathbb{R}^2$ the three are a diamond, a disc, and a square, and they are nested.



The closed unit balls of the three norms on $\mathbb{R}^2$. A smaller ball belongs to a larger norm: $B_1 \subseteq B_2 \subseteq B_\infty$ says exactly that $\|\mathbf{x}\|_\infty \le \|\mathbf{x}\|_2 \le \|\mathbf{x}\|_1$, and all three agree at $\mathbf{e}_1$

Two features of the picture are general. Each ball is **symmetric**, $v \in B$ if and only if $-v \in B$, which is (N2) with $c = -1$; and each is **convex**, since if $\|u\|, \|v\| \le 1$ and $0 \le t \le 1$ then

$$\|tu + (1-t)v\| \le t\|u\| + (1-t)\|v\| \le 1$$

by (N3) and (N2). The diamond has corners and the square has corners; convexity is what the triangle inequality buys, and it is all it buys. The disc is rounder than the other two for a reason that is the subject of the last part of this section.

The nesting in the picture is an inequality, and it holds in every dimension, with a matching chain running the other way.

Proposition 16.1.2. Comparing the Three Norms

For every $\mathbf{x} \in F^n$,

$$\|\mathbf{x}\|_\infty \le \|\mathbf{x}\|_2 \le \|\mathbf{x}\|_1$$

and

$$\|\mathbf{x}\|_1 \le \sqrt{n}\,\|\mathbf{x}\|_2, \qquad \|\mathbf{x}\|_2 \le \sqrt{n}\,\|\mathbf{x}\|_\infty.$$

Each of the four inequalities is attained: the first two by $\mathbf{x} = \mathbf{e}_1$, the last two by $\mathbf{x} = \mathbf{1}$.

Idea. Four inequalities, two moves. The ones **without** a $\sqrt{n}$ compare a sum against its largest term, or a sum of squares against the square of a sum, and need nothing but nonnegativity. The ones **with** a $\sqrt{n}$ are a trade of n terms against one, and the sharp constant comes from Cauchy–Schwarz against the all-ones vector $\mathbf{1}$ — which is exactly why $\mathbf{1}$ is the vector that attains them.

Proof. Write $a_i = |x_i| \ge 0$ and let $a_k = \max_i a_i$. Then $a_k^2 \le \sum_i a_i^2$, and taking nonnegative square roots gives $\|\mathbf{x}\|_\infty \le \|\mathbf{x}\|_2$. Next,

$$\|\mathbf{x}\|_1^2 = \Big(\sum_i a_i\Big)^2 = \sum_i a_i^2 + \sum_{i \ne j} a_i a_j \;\ge\; \|\mathbf{x}\|_2^2,$$

since every cross term $a_i a_j$ is ≥ 0; taking square roots gives $\|\mathbf{x}\|_2 \le \|\mathbf{x}\|_1$.

For the third inequality, apply Theorem 11.1.5 in $\mathbb{R}^n$ with the dot product to the vectors

$(a_1, \dots, a_n)$ and $\mathbf{1}$:

$$\|\mathbf{x}\|_1 = \sum_i a_i \cdot 1 \le \Big(\sum_i a_i^2\Big)^{1/2} \Big(\sum_i 1\Big)^{1/2} = \sqrt{n}\,\|\mathbf{x}\|_2.$$

For the fourth, $\sum_i a_i^2 \le n \max_i a_i^2$, and taking square roots gives $\|\mathbf{x}\|_2 \le \sqrt{n}\,\|\mathbf{x}\|_\infty$.

For the equality cases, $\mathbf{e}_1$ has $\|\mathbf{e}_1\|_1 = \|\mathbf{e}_1\|_2 = \|\mathbf{e}_1\|_\infty = 1$, and $\mathbf{1}$ has $\|\mathbf{1}\|_1 = n$, $\|\mathbf{1}\|_2 = \sqrt{n}$, and $\|\mathbf{1}\|_\infty = 1$, so $\|\mathbf{1}\|_1 = \sqrt{n}\,\|\mathbf{1}\|_2$ and $\|\mathbf{1}\|_2 = \sqrt{n}\,\|\mathbf{1}\|_\infty$. This proves the proposition.

■

So the three norms never disagree by a factor worse than $\sqrt{n}$, and the factor $\sqrt{n}$ really occurs. Hold on to both halves of that sentence: §02 proves the first half for *every* pair of norms, and the second half is why its constants cannot be made to forget the dimension.

Quick check. Sketch the set $\{\mathbf{x} \in \mathbb{R}^2 : \|\mathbf{x}\|_w \le 1\}$ for the weighted norm $\|\mathbf{x}\|_w = \max(|x_1|, 2|x_2|)$ of Example (Inner products, weights, and a function space) (b), and find the smallest C with $\|\mathbf{x}\|_\infty \le C\|\mathbf{x}\|_w$ for all $\mathbf{x}$.

Solution. The ball is the rectangle $|x_1| \le 1$, $|x_2| \le \frac{1}{2}$: the square B_∞ squashed vertically by a factor 2. Since $|x_1| \le \|\mathbf{x}\|_w$ and $|x_2| \le \frac{1}{2}\|\mathbf{x}\|_w$, we get $\|\mathbf{x}\|_\infty \le \|\mathbf{x}\|_w$, so $C = 1$ works, and it is smallest because $\mathbf{x} = \mathbf{e}_1$ gives $\|\mathbf{x}\|_\infty = \|\mathbf{x}\|_w = 1$. The other direction costs a factor: $\|\mathbf{x}\|_w \le 2\|\mathbf{x}\|_\infty$, attained at $\mathbf{e}_2$. Squashing a ball inward enlarges the norm, and the largest squash factor is the constant.

16.1.3 Which norms come from an inner product

Chapter 11 §01 closed with a promise. It observed that Theorem 11.1.2 (a), (b), and Corollary 11.1.6 make every induced norm a norm in the general sense, warned that the converse fails, and named the test: the **parallelogram law**

$$\|u + v\|^2 + \|u - v\|^2 = 2\|u\|^2 + 2\|v\|^2. \tag{$*$}$$

It then proved one direction outright — expand both sides with Theorem 11.1.2 (c) and add — and left the converse to Exercise (C1: The parallelogram law characterizes inner-product norms), which carries it out in $\mathbb{R}^2$ while **granting** one step, the homogeneity $\beta(u, tv) = t\beta(u, v)$ for real t. That exercise's own remark says where the gap is: additivity gives homogeneity for rational t by induction, and passing to real t needs a limit.

Here is what this section adds. The dimension 2 in Exercise (C1: The parallelogram law characterizes inner-product norms) was never used, so the computation runs in any real space; the granted step is supplied by a limit, using the completeness of $\mathbb{R}$; and the complex case, which Chapter 11 did not touch at all, is built from the real one. The result is the theorem the promise described, with no hypothesis on the dimension.

Theorem 16.1.3. The Parallelogram Law Characterizes Inner-Product Norms

Let F be $\mathbb{R}$ or $\mathbb{C}$, let V be a vector space over F, and let $\|\cdot\|$ be a norm on V. Then there

is an inner product on V with $\langle v, v\rangle = \|v\|^2$ for every v **if and only if** $\|\cdot\|$ satisfies the parallelogram law $(*)$ for all $u, v \in V$. When it exists, that inner product is unique.

Idea. The forward direction is Chapter 11's computation. For the converse there is only one candidate for the inner product, because Exercise (C2: Recovering the inner product from the norm) already shows that an inner product is determined by its norm through polarization; so write down the polarization formula and check the axioms. ① Over $\mathbb{R}$ the symmetry and the additivity are the computation of Exercise (C1: The parallelogram law characterizes inner-product norms), which never mentions the dimension. ② Homogeneity comes free for rational scalars and is pushed to real scalars by a limit, which is the one analytic step. ③ Over $\mathbb{C}$, apply ① and ② to V regarded as a real space, then glue the real form β and its twist $\beta(u, iv)$ into a complex inner product; the single identity $\beta(iu, iv) = \beta(u, v)$ does all the work.

Proof. $(\Rightarrow)$ Suppose $\langle v, v\rangle = \|v\|^2$ for an inner product on V. By Theorem 11.1.2 (c) and the same identity with $-v$ in place of v,

$$\|u + v\|^2 = \|u\|^2 + 2\,\mathrm{Re}\langle u, v\rangle + \|v\|^2,$$
$$\|u - v\|^2 = \|u\|^2 - 2\,\mathrm{Re}\langle u, v\rangle + \|v\|^2.$$

Adding them gives $(*)$.

$(\Leftarrow)$ Suppose $(*)$ holds. Regard V as a vector space over $\mathbb{R}$, by restricting which scalars are allowed. This changes nothing about $\|\cdot\|$: the vectors, the addition, and the values of the norm are the same, and (N2) for real c is a special case of (N2) for $c \in F$. Define

$$\beta(u, v) = \tfrac{1}{4}\big(\|u + v\|^2 - \|u - v\|^2\big).$$

Step 1. β is symmetric and additive in each slot. Exercise (C1: The parallelogram law characterizes inner-product norms) runs this computation in $\mathbb{R}^2$; rather than transport an exercise past the hypotheses it was stated under, here it is in full, for an arbitrary V.

Setting $v = u$ gives $\beta(u, u) = \frac{1}{4}\|2u\|^2 = \|u\|^2$ by (N2). Swapping the arguments replaces $\|u - v\|$ by $\|v - u\|$, equal by (N2) with $c = -1$, so $\beta(u, v) = \beta(v, u)$; and $\beta(u, 0) = 0$.

For additivity, apply $(*)$ twice. To the pair $u + \frac{1}{2}(v + w)$, $\frac{1}{2}(v - w)$, whose sum is $u + v$ and whose difference is $u + w$:

$$\|u + v\|^2 + \|u + w\|^2 = 2\|u + \tfrac{1}{2}(v + w)\|^2 + 2\|\tfrac{1}{2}(v - w)\|^2.$$

Then to the pair $u - \frac{1}{2}(v + w)$, $\frac{1}{2}(w - v)$, whose sum is $u - v$ and whose difference is $u - w$:

$$\|u - v\|^2 + \|u - w\|^2 = 2\|u - \tfrac{1}{2}(v + w)\|^2 + 2\|\tfrac{1}{2}(w - v)\|^2.$$

Subtract the second from the first. On the left, the two differences of squares are $4\beta(u, v)$ and $4\beta(u, w)$, by the definition of β. On the right, the last terms cancel, since $\|\frac{1}{2}(v - w)\| = \|\frac{1}{2}(w - v)\|$ by (N2) with $c = -1$, and the remaining difference of squares is $4\beta\big(u, \frac{1}{2}(v + w)\big)$. So $4\beta(u, v) + 4\beta(u, w) = 8\beta\big(u, \frac{1}{2}(v + w)\big)$, that is,

$$\beta(u, v) + \beta(u, w) = 2\,\beta\big(u, \tfrac{1}{2}(v + w)\big).$$

Taking $w = 0$ and using $\beta(u, 0) = 0$ turns the right-hand side into $2\beta(u, \frac{1}{2}v)$, so $\beta(u, v) = 2\beta(u, \frac{1}{2}v)$ for every v; substituting that back with $v + w$ in place of v gives

$$\beta(u, v) + \beta(u, w) = \beta(u, v + w)$$

for all $u, v, w \in V$. Nothing in this argument mentions a dimension. Additivity in the first slot follows from the symmetry.

Step 2. $\beta(u, tv) = t\,\beta(u, v)$ for every real t. This is the step Exercise (C1: The parallelogram law characterizes inner-product norms) granted. First, $\beta(u, -v) = -\beta(u, v)$, directly from the definition of β, since replacing v by $-v$ exchanges the two terms. Next, induction on $m \geq 0$ using Step 1 gives $\beta(u, mv) = m\beta(u, v)$, the case $m = 0$ being $\beta(u, 0) = 0$; combined with the previous sentence this covers every integer m. For a rational $q = m/n$ with $n \geq 1$,

$$n\,\beta(u, qv) = \beta(u, nqv) = \beta(u, mv) = m\,\beta(u, v),$$

so $\beta(u, qv) = q\beta(u, v)$.

Now fix a real t and put $M = \|u\| + (|t| + 1)\|v\|$. For any rational q with $|q - t| \leq 1$, both $\|u \pm tv\|$ and $\|u \pm qv\|$ are at most M by (N2) and (N3), while Lemma 16.1.1 and (N2) give

$$\big|\|u \pm tv\| - \|u \pm qv\|\big| \leq |t - q|\,\|v\|.$$

Since $|a^2 - b^2| = |a - b|(a + b)$ for $a, b \geq 0$, each of the two squared terms in β moves by at most $2M\|v\||t - q|$, and therefore

$$|\beta(u, tv) - \beta(u, qv)| \leq M\|v\|\,|t - q|.$$

By **the completeness of $\mathbb{R}$, fact (A1) of the chapter introduction**, there is a sequence of rationals $q_k \to t$, and we may assume $|q_k - t| \leq 1$. Then

$$\begin{aligned} &|\beta(u, tv) - t\beta(u, v)| \\ &\quad \leq |\beta(u, tv) - \beta(u, q_k v)| + |q_k - t|\,|\beta(u, v)| \\ &\quad \leq (M\|v\| + |\beta(u, v)|)|q_k - t|, \end{aligned}$$

where the middle term used $\beta(u, q_k v) = q_k \beta(u, v)$. The left-hand side does not depend on k, and the right-hand side tends to 0, so the left-hand side is 0.

If $F = \mathbb{R}$, we are done: β is symmetric, bilinear by Steps 1 and 2, and $\beta(v, v) = \|v\|^2 > 0$ for $v \neq 0$ by (N1), so (IP1), (IP2), and (IP3) of Definition 11.1.1 all hold and the induced norm is $\|\cdot\|$.

Step 3. The complex case. Suppose $F = \mathbb{C}$. Steps 1 and 2 apply to V as a real space, so β is a symmetric $\mathbb{R}$-bilinear form with $\beta(v, v) = \|v\|^2$. Define

$$\langle u, v\rangle = \beta(u, v) + i\,\beta(u, iv).$$

Everything follows from one identity: $\beta(iu, iv) = \beta(u, v)$, which holds because $\|iu \pm iv\| = \|i(u \pm v)\| = |i|\|u \pm v\| = \|u \pm v\|$ by (N2). Applying it twice, $\beta(iu, v) = \beta(i \cdot iu, iv) = -\beta(u, iv)$.

(IP3) and the norm: $\beta(u, iu) = \frac{1}{4}(\|u + iu\|^2 - \|u - iu\|^2)$, and (N2) turns this into $\frac{1}{4}(|1 + i|^2 - |1 - i|^2)\|u\|^2 = 0$. Hence $\langle u, u\rangle = \beta(u, u) = \|u\|^2$, which is > 0 for $u \neq 0$, and the induced norm is $\|\cdot\|$.

(IP1): additivity in the first slot and homogeneity for real scalars come from Steps 1 and 2 applied to both $\beta(\cdot, v)$ and $\beta(\cdot, iv)$. For the scalar i,

$$\begin{aligned}\langle iu, v\rangle &= \beta(iu, v) + i\,\beta(iu, iv)\\ &= -\beta(u, iv) + i\,\beta(u, v) = i\,\langle u, v\rangle.\end{aligned}$$

Every complex scalar is $a + bi$ with a, b real, so (IP1) follows.

(IP2): using $\beta(v, iu) = \beta(iv, i \cdot iu) = -\beta(u, iv)$ and the symmetry of β,

$$\langle v, u\rangle = \beta(u, v) - i\,\beta(u, iv) = \overline{\langle u, v\rangle}.$$

So $\langle\cdot,\cdot\rangle$ is an inner product inducing $\|\cdot\|$.

Uniqueness. Suppose $\langle\cdot,\cdot\rangle'$ is any inner product on V with $\langle v, v\rangle' = \|v\|^2$ for every v. The two displays of the ($\Rightarrow$) direction apply to it verbatim, and subtracting them instead of adding them gives

$$\|u + v\|^2 - \|u - v\|^2 = 4\,\mathrm{Re}\langle u, v\rangle',$$

that is, $\mathrm{Re}\langle u, v\rangle' = \beta(u, v)$: the real part of $\langle\cdot,\cdot\rangle'$ is forced to be the β built above. Over $\mathbb{R}$ an inner product is real-valued, so $\langle u, v\rangle' = \beta(u, v)$ and there is nothing left to choose. Over $\mathbb{C}$, apply the same identity to the pair u, iv. By (IP1) and (IP2) together the form is conjugate-linear in its second slot, so $\langle u, iv\rangle' = \bar{i}\,\langle u, v\rangle' = -i\,\langle u, v\rangle'$, whose real part is $\mathrm{Im}\langle u, v\rangle'$. Hence $\mathrm{Im}\langle u, v\rangle' = \beta(u, iv)$, and the two parts together give $\langle u, v\rangle' = \beta(u, v) + i\,\beta(u, iv)$, which is exactly the inner product constructed above. So it is the only one. (Exercise (C2: Recovering the inner product from the norm) set this computation as an exercise in Chapter 11; it is carried out here because the theorem rests on it.)

This proves the theorem.

■

The theorem says that the parallelogram law is not one property of induced norms among many — it is the whole difference. A norm either obeys it, and then it is secretly an inner product with an inner product's entire apparatus of angles, orthogonality, and adjoints; or it violates it once, and then no inner product will ever produce it.

Warning. **Neither $\|\cdot\|_1$ nor $\|\cdot\|_\infty$ comes from an inner product.** Test $(*)$ on $\mathbf{u} = \mathbf{e}_1$ and $\mathbf{v} = \mathbf{e}_2$ in F^2. Both $\mathbf{u} + \mathbf{v} = (1, 1)$ and $\mathbf{u} - \mathbf{v} = (1, -1)$ have $\|\cdot\|_1 = 2$, so the left side of $(*)$ is $4 + 4 = 8$, while the right side is $2 + 2 = 4$. For $\|\cdot\|_\infty$ both have norm 1, so the left side is 2 and the right side is again 4. One failing pair is enough, and there is no repair: by Theorem 16.1.3 no inner product whatsoever induces these norms, in any dimension ≥ 2. The same two vectors $f(t) = t$ and $g(t) = 1 - t$ in $C[0, 1]$ refute it for the sup norm, since $\|f\|_\infty = \|g\|_\infty = \|f + g\|_\infty = \|f - g\|_\infty = 1$.

Quick check. The function $N(\mathbf{x}) = \left(x_1^2 + 4x_2^2\right)^{1/2}$ is a norm on $\mathbb{R}^2$, and it is lopsided: $N(\mathbf{e}_1) = 1$ while $N(\mathbf{e}_2) = 2$. Does it come from an inner product?

Solution. Yes. It is the norm induced by $\langle \mathbf{x}, \mathbf{y} \rangle = x_1 y_1 + 4x_2 y_2$, which satisfies (IP1) and (IP2) by inspection and (IP3) because $x_1^2 + 4x_2^2 > 0$ for $\mathbf{x} \neq \mathbf{0}$. One may also check $(*)$ directly: both sides expand to $2x_1^2 + 8x_2^2 + 2y_1^2 + 8y_2^2$. The lesson is that a norm can be visibly lopsided — its unit ball here is an ellipse, not a circle — and still be an inner-product norm. What matters is not roundness but the parallelogram law.

So the norms of this book split in two. On one side sit $\|\cdot\|_2$, the Frobenius norm, and every weighted Euclidean norm, where Chapters 11 to 13 apply in full. On the other sit $\|\cdot\|_1$, $\|\cdot\|_\infty$, and the sup norm, where there is no orthogonality and no adjoint, and the only tools are the three axioms. The next section shows that for questions about convergence the split does not matter at all.

16.1.4 Exercises

16.1.4.1 A. Check your understanding

Exercise. A1

(a) State the three clauses of Definition 16.1.1, and say which one fails for $N(\mathbf{x}) = |x_1|$ on F^2.

(b) Deduce $\|0\| = 0$ from (N2) alone, and explain why (N1) is therefore stated with the word **only**.

(c) True or false: a function satisfying (N1) and (N3) but not (N2) cannot exist. Justify your answer.

(d) Name the test that decides whether a norm is induced by an inner product, and state what Exercise (C1: The parallelogram law characterizes inner-product norms) had already established before this section.

(e) Order $\|\mathbf{x}\|_1$, $\|\mathbf{x}\|_2$, $\|\mathbf{x}\|_\infty$ for a general $\mathbf{x} \in F^n$, and give a vector for which all three are equal.

Solution.

(a) (N1) positive definiteness, (N2) absolute homogeneity, (N3) the triangle inequality (Definition 16.1.1). For $N(\mathbf{x}) = |x_1|$ only (N1) fails, and only its "only" half: $N(\mathbf{e}_2) = 0$ with $\mathbf{e}_2 \neq \mathbf{0}$ (Example (Three near misses) (b)).

(b) Take $c = 0$ and any v: (N2) gives $\|0\| = \|0v\| = |0|\|v\| = 0$. So "$v = 0 \Rightarrow \|v\| = 0$" is automatic, and the only content left in (N1) is the converse, which is why it is the clause that is stated.

(c) False. Take $N(\mathbf{x}) = \min(\|\mathbf{x}\|_2, 1)$ on $\mathbb{R}^2$. It is ≥ 0 and vanishes only at $\mathbf{0}$, so (N1) holds. For (N3), note first that $\min(a+b, 1) \leq \min(a, 1) + \min(b, 1)$ for all $a, b \geq 0$: if either $a \geq 1$ or $b \geq 1$ the right side is at least 1, which bounds the left side; and otherwise the right side is $a + b$, which is at least $\min(a+b, 1)$. Applying this to $a = \|\mathbf{u}\|_2$, $b = \|\mathbf{v}\|_2$ and using that $s \mapsto \min(s, 1)$ is increasing gives (N3). But $N(2\mathbf{e}_1) = 1 \neq 2 = 2N(\mathbf{e}_1)$, so (N2) fails.

(d) The parallelogram law $(*)$: a norm is induced by an inner product exactly when $\|u+v\|^2 + \|u-v\|^2 = 2\|u\|^2 + 2\|v\|^2$ always (Theorem 16.1.3). Before this section, Exercise (C1: The parallelogram law characterizes inner-product norms) had proved the

converse direction on $\mathbb{R}^2$, and only modulo the granted homogeneity $\beta(u,tv) = t\beta(u,v)$ for real t; the general dimension, the real homogeneity step, and the complex case are what Theorem 16.1.3 adds.

(e) $\|\mathbf{x}\|_\infty \le \|\mathbf{x}\|_2 \le \|\mathbf{x}\|_1$ (Proposition 16.1.2). All three agree at $\mathbf{x} = \mathbf{e}_1$, and more generally at any $\mathbf{x}$ with at most one nonzero coordinate.

16.1.4.2 B. Practice

Exercise. B1: Three measurements

For $\mathbf{x} = (1,-1,2,-2) \in \mathbb{R}^4$, compute $\|\mathbf{x}\|_1$, $\|\mathbf{x}\|_2$, and $\|\mathbf{x}\|_\infty$, and verify all four inequalities of Proposition 16.1.2. Which of them is closest to being an equality?

Solution. $\|\mathbf{x}\|_1 = 1+1+2+2 = 6$; $\|\mathbf{x}\|_2 = (1+1+4+4)^{1/2} = \sqrt{10}$; $\|\mathbf{x}\|_\infty = 2$. Here $n = 4$, so $\sqrt{n} = 2$. The four inequalities read

$$2 \le \sqrt{10} \le 6, \qquad 6 \le 2\sqrt{10}, \qquad \sqrt{10} \le 4,$$

and $\sqrt{10} \approx 3.162$, so all four hold. The tightest is $\|\mathbf{x}\|_1 \le \sqrt{n}\|\mathbf{x}\|_2$, that is $6 \le 6.32$; this is as expected, since equality there needs all coordinates of equal modulus, and $\mathbf{x}$ has only two distinct moduli.

Exercise. B2: Determine which are norms

Determine which of the following define norms on $\mathbb{R}^2$. For those that do, verify all three clauses of Definition 16.1.1; for those that do not, name the exact clause that fails and give a witness. Justify your answer.

(a) $N(\mathbf{x}) = |x_1| + 2|x_2|$.

(b) $N(\mathbf{x}) = |x_1 - x_2|$.

(c) $N(\mathbf{x}) = \max(|x_1|,|x_2|) - \tfrac{1}{4}\min(|x_1|,|x_2|)$.

(d) $N(\mathbf{x}) = |x_1| + |x_2| - \tfrac{1}{2}\min(|x_1|,|x_2|)$.

Solution.

(a) A norm. It is the weighted norm of Example (Inner products, weights, and a function space) (b) in its 1-norm form: (N1) because both terms are ≥ 0 and both vanish only for $\mathbf{x} = \mathbf{0}$; (N2) because $|cx_i| = |c||x_i|$; (N3) coordinatewise, exactly as for $\|\cdot\|_1$.

(b) Not a norm. (N2) and (N3) hold, inherited from $|\cdot|$ on $\mathbb{R}$. **(N1) fails**: $N((1,1)) = 0$ with $(1,1) \neq \mathbf{0}$.

(c) Not a norm. (N1) holds, since $N(\mathbf{x}) \ge \tfrac{3}{4}\max(|x_1|,|x_2|)$, which is > 0 for $\mathbf{x} \neq \mathbf{0}$; and (N2) holds because max and min of the scaled moduli both pick up the factor $|c|$. **(N3) fails**: with $\mathbf{u} = (1,1)$ and $\mathbf{v} = (1,-1)$ we get $N(\mathbf{u}) = N(\mathbf{v}) = 1 - \tfrac{1}{4} = \tfrac{3}{4}$, while $\mathbf{u}+\mathbf{v} = (2,0)$ has $N(\mathbf{u}+\mathbf{v}) = 2 - 0 = 2 > \tfrac{3}{2}$. Subtracting from the largest coordinate is exactly what a cancellation in the sum can punish.

(d) A norm, despite the minus sign. Since $|x_1| + |x_2| = \max + \min$, the formula is $N(\mathbf{x}) =$

$\max + \frac{1}{2}\min$, which is

$$N(\mathbf{x}) = \tfrac{1}{2}\|\mathbf{x}\|_1 + \tfrac{1}{2}\|\mathbf{x}\|_\infty.$$

A sum of two norms, each scaled by a positive number, is a norm: (N1), (N2), and (N3) are each inherited term by term, and (N1) needs only one of the two summands to detect $\mathbf{x} \neq \mathbf{0}$. Compare (c): there the subtracted term was large enough to break (N3), here it is not, and the way to tell is to rewrite the formula rather than to guess.

Exercise. B3: Which of these come from an inner product?

For each norm, determine whether it is induced by an inner product. Justify your answer, using Theorem 16.1.3 in each case.

(a) $\|\mathbf{x}\| = (x_1^2 + x_1x_2 + x_2^2)^{1/2}$ on $\mathbb{R}^2$.

(b) $\|\mathbf{A}\|_F$ on $M_2(\mathbb{R})$.

(c) $N(\mathbf{A}) = \max_{i,j}|a_{ij}|$ on $M_2(\mathbb{R})$.

Solution.

(a) Yes. Put $\langle \mathbf{x}, \mathbf{y}\rangle = x_1y_1 + \frac{1}{2}(x_1y_2 + x_2y_1) + x_2y_2$. It is bilinear and symmetric by inspection, and $\langle \mathbf{x}, \mathbf{x}\rangle = x_1^2 + x_1x_2 + x_2^2 = (x_1 + \frac{1}{2}x_2)^2 + \frac{3}{4}x_2^2$, which is > 0 unless $x_2 = 0$ and $x_1 = 0$; so (IP3) holds and the induced norm is the given one. (Completing the square is also how one sees that the given function is a norm at all.)

(b) Yes. The Frobenius norm is induced by the Frobenius inner product $\langle \mathbf{A}, \mathbf{B}\rangle = \operatorname{tr}(\mathbf{B}^\top \mathbf{A})$ (Example (Inner products, weights, and a function space) (a)), so $(*)$ holds by the forward direction of Theorem 16.1.3.

(c) No. Take $\mathbf{A} = \mathbf{E}_{11}$ and $\mathbf{B} = \mathbf{E}_{12}$, the matrix units. Then $\mathbf{A} \pm \mathbf{B}$ has entries 1 and ± 1 in the first row and zeros elsewhere, so $N(\mathbf{A} + \mathbf{B}) = N(\mathbf{A} - \mathbf{B}) = 1$ and the left side of $(*)$ is 2, while the right side is $2 \cdot 1 + 2 \cdot 1 = 4$. By Theorem 16.1.3 no inner product induces N.

16.1.4.3 C. Going deeper

Exercise. C1: Sharpness of the comparison constants

(a) Prove that $\|\mathbf{x}\|_1 = \sqrt{n}\,\|\mathbf{x}\|_2$ holds if and only if all the numbers $|x_1|, \dots, |x_n|$ are equal. *Hint: the equality case of Theorem 11.1.5.*

(b) Hence deduce that no constant C independent of n satisfies $\|\mathbf{x}\|_1 \le C\|\mathbf{x}\|_2$ for all n and all $\mathbf{x} \in \mathbb{R}^n$.

Solution.

(a) With $\mathbf{a} = (|x_1|, \dots, |x_n|)$, the proof of Proposition 16.1.2 reads $\|\mathbf{x}\|_1 = \langle \mathbf{a}, \mathbf{1}\rangle \le \|\mathbf{a}\|_2\|\mathbf{1}\|_2 = \sqrt{n}\|\mathbf{x}\|_2$, an application of Theorem 11.1.5 to $\mathbf{a}$ and $\mathbf{1}$. Equality in Theorem 11.1.5 holds exactly when $(\mathbf{a}, \mathbf{1})$ is dependent. Since $\mathbf{1} \neq \mathbf{0}$, that means $\mathbf{a} = t\mathbf{1}$ for some real t, which is $t \ge 0$ because the entries of $\mathbf{a}$ are; and $\mathbf{a} = t\mathbf{1}$ says precisely that all the $|x_i|$ are equal.

(b) By (a), $\mathbf{x} = \mathbf{1} \in \mathbb{R}^n$ gives $\|\mathbf{1}\|_1 = n$ and $\|\mathbf{1}\|_2 = \sqrt{n}$, so any valid C satisfies $C \geq n/\sqrt{n} = \sqrt{n}$. As n is arbitrary, no single C works for all n. The comparison constants between two norms are genuinely functions of the dimension.

Exercise. C2: A real norm that is not a complex norm

Regard $\mathbb{C}$ as a one-dimensional complex vector space, and also as $\mathbb{R}^2$ by $z = a + bi \leftrightarrow (a, b)$. Define $N(z) = 2|a| + |b|$.

(a) Prove that N is a norm on $\mathbb{R}^2$.

(b) Prove that N is **not** a norm on $\mathbb{C}$ as a complex vector space, and name the exact clause that fails.

(c) Prove that if $\|\cdot\|$ is any norm on a complex vector space V, then it is also a norm on V regarded as a real vector space. Explain why (b) does not contradict this.

Solution.

(a) It is the weighted 1-norm $|x_1| \cdot 2 + |x_2| \cdot 1$ of Exercise (B2: Determine which are norms) (a) with the weights swapped, so all three clauses hold as there.

(b) **(N2) fails** for the complex scalar $c = i$. Indeed $N(1) = 2$ and $N(i) = 1$, while (N2) would demand $N(i \cdot 1) = |i|N(1) = 2$.

(c) Restricting the scalars from $\mathbb{C}$ to $\mathbb{R}$ leaves the vectors, the addition, and the zero vector unchanged, so (N1) and (N3) are the same statements. (N2) for real c is a special case of (N2) for complex c, the absolute value of a real number being the same in $\mathbb{R}$ and in $\mathbb{C}$. There is no contradiction with (b) because the implication runs one way only: (N2) over $\mathbb{C}$ is a **strictly stronger** demand than (N2) over $\mathbb{R}$, the extra content being $\|iv\| = \|v\|$, and N fails exactly that.

Exercise. C3: The ball determines the norm

Let $\|\cdot\|$ and N be two norms on a vector space V with the **same** closed unit ball, in the sense of Definition 16.1.2:

$$\{v : \|v\| \leq 1\} = \{v : N(v) \leq 1\}.$$

(a) Prove that for every $v \in V$ and every real $t > 0$, we have $\|v\| \leq t$ if and only if v/t lies in the closed unit ball of $\|\cdot\|$.

(b) Hence prove that $\|v\| = N(v)$ for every $v \in V$.

(c) Give two different norms on $\mathbb{R}^2$ whose closed unit balls have the same area, and say why (b) does not apply to them.

Solution.

(a) Since $t > 0$, (N2) gives $\|v/t\| = \|v\|/t$. So v/t lies in the ball exactly when $\|v\|/t \leq 1$, that is, exactly when $\|v\| \leq t$.

(b) For $v = 0$ both sides are 0 by (N1). Let $v \neq 0$ and put $t = \|v\| > 0$. By (a), v/t lies

in the ball of $\|\cdot\|$, hence in the ball of N, hence $N(v) \le t = \|v\|$ by (a) applied to N. Swapping the roles of $\|\cdot\|$ and N gives $\|v\| \le N(v)$. Therefore $\|v\| = N(v)$.

(c) Take $\|\cdot\|_1$, whose ball is the diamond with vertices $\pm\mathbf{e}_1$ and $\pm\mathbf{e}_2$, of area 2; and $N(\mathbf{x}) = 2|x_1| + \frac{1}{2}|x_2|$, whose ball is the diamond with vertices $\pm\frac{1}{2}\mathbf{e}_1$ and $\pm 2\mathbf{e}_2$, also of area 2. They differ at $\mathbf{e}_1$, where they give 1 and 2. Part (b) does not apply because equal **area** is far weaker than equal **set**: the recipe in (a) reads the norm off the ball as a set of vectors, and the area of that set retains almost none of it.

16.2 All Norms Agree

§01 left a worry. There are many norms on F^n, they give different answers, and every statement of the form "this quantity is small" now has to say small in which norm. This section removes the worry in finite dimension, in the strongest way available: any two norms on a finite-dimensional space are trapped between fixed multiples of each other, so they agree about every limit, every bounded set, and every continuous function. It also says what the theorem costs — constants that grow with the dimension — and where it stops being true.

16.2.1 Trapped between multiples

Look again at Proposition 16.1.2. Its four inequalities say that on F^n,

$$\tfrac{1}{\sqrt{n}}\,\|\mathbf{x}\|_1 \;\le\; \|\mathbf{x}\|_2 \;\le\; \|\mathbf{x}\|_1,$$

so $\|\cdot\|_2$ never differs from $\|\cdot\|_1$ by a factor worse than $\sqrt{n}$, in either direction. Whatever $\|\cdot\|_1$ calls small, $\|\cdot\|_2$ calls small too, at the cost of a fixed factor. That is a relation worth naming.

Two norms are equivalent when each is sandwiched between two fixed multiples of the other, so that neither can be small while the other is large.

Definition 16.2.1. Equivalent Norms

Let V be a vector space over $F = \mathbb{R}$ or $\mathbb{C}$, and let $\|\cdot\|_a$ and $\|\cdot\|_b$ be norms on V. They are **equivalent** if there exist real constants $c, C > 0$, **not depending on the vector**, such that

$$c\,\|v\|_b \;\le\; \|v\|_a \;\le\; C\,\|v\|_b \qquad \text{for every } v \in V.$$

In words. The two inequalities are read separately: the right one says $\|\cdot\|_a$ is *no larger than* a fixed multiple of $\|\cdot\|_b$, and the left one says it is *no smaller than* a fixed positive multiple. The phrase **not depending on the vector** is the whole content, since for each individual $v \neq 0$ the ratio $\|v\|_a/\|v\|_b$ is some positive number and one could always take that number as the constant. What is being asked is that one pair of constants works for all v at once. Note also that the definition says nothing about the two norms being close; c may be 10^{-6}.

Example. The three p-norms, with best constants

On F^n, Proposition 16.1.2 gives

$$\begin{aligned} \|\mathbf{x}\|_\infty \le \|\mathbf{x}\|_2 &\le \sqrt{n}\,\|\mathbf{x}\|_\infty, \\ \|\mathbf{x}\|_\infty \le \|\mathbf{x}\|_1 &\le n\,\|\mathbf{x}\|_\infty, \\ \tfrac{1}{\sqrt{n}}\|\mathbf{x}\|_1 \le \|\mathbf{x}\|_2 &\le \|\mathbf{x}\|_1, \end{aligned}$$

so the three are pairwise equivalent. The middle line's upper constant comes from $\|\mathbf{x}\|_1 \le \sqrt{n}\|\mathbf{x}\|_2 \le \sqrt{n}\cdot\sqrt{n}\,\|\mathbf{x}\|_\infty$. Every constant shown is attained, so none of them can be improved. On the first two lines the lower constant is attained at $\mathbf{e}_1$ and the upper one at $\mathbf{1}$. **The third line runs the other way**: at $\mathbf{1}$ the lower inequality reads $\frac{1}{\sqrt{n}}\cdot n = \sqrt{n} \le \sqrt{n}$, and at $\mathbf{e}_1$ the upper one reads $1 \le 1$.

Before the theorem, one small fact makes the proof strategy possible: it is enough to compare every norm with a single one of our choosing.

Proposition 16.2.1. Equivalence of Norms Is an Equivalence Relation

On a fixed vector space V over F, the relation of Definition 16.2.1 is reflexive, symmetric, and transitive.

Proof. Reflexive: take $c = C = 1$. Symmetric: if $c\|v\|_b \le \|v\|_a \le C\|v\|_b$ for all v, then dividing through by the positive constants gives $C^{-1}\|v\|_a \le \|v\|_b \le c^{-1}\|v\|_a$. Transitive: if in addition $c'\|v\|_d \le \|v\|_b \le C'\|v\|_d$, then multiplying the first chain's outer inequalities by the positive constants c and C gives

$$cc'\|v\|_d \le c\|v\|_b \le \|v\|_a \le C\|v\|_b \le CC'\|v\|_d,$$

and $cc', CC' > 0$. This proves all three properties. ∎

So to prove that *any two* norms on V are equivalent, it suffices to prove that *every* norm is equivalent to one fixed reference norm. We take the reference to be the Euclidean norm of a chosen basis, because that is where compactness lives.

16.2.2 Compactness enters the book

Theorem 16.2.2. Equivalence of Norms in Finite Dimension

Let V be a vector space over $F = \mathbb{R}$ or $\mathbb{C}$ with $\dim V = n < \infty$. Then **any two** norms on V are equivalent. Explicitly, if $\mathcal{B} = (v_1, \dots, v_n)$ is a basis of V, if $\|\cdot\|_2$ denotes the Euclidean norm of the coordinate vector,

$$\Big\|\sum_i x_i v_i\Big\|_2 := \left(|x_1|^2 + \cdots + |x_n|^2\right)^{1/2},$$

and if N is any norm on V, then there are constants $0 < m \le M$ with

$$m\,\|v\|_2 \;\le\; N(v) \;\le\; M\,\|v\|_2 \qquad \text{for every } v \in V.$$

Idea. The two inequalities are not symmetric, and it is worth seeing why before starting. The **upper** bound is cheap: write v in the basis, apply the triangle inequality once, and the constant M is built out of the numbers $N(v_i)$. The **lower** bound cannot be obtained the same way; there is no inequality running from $N(\sum x_i v_i)$ down to the individual coordinates, and the temptation to assume that the two norms are comparable on the basis and then "extend

by linearity" is exactly the error the theorem exists to forbid — Example (Agreeing on a basis is not enough) below shows two norms agreeing on a basis and differing by a factor of n elsewhere.

The route is therefore ① upper bound by the triangle inequality; ② use that upper bound to show N is continuous as a function of the coordinates; ③ restrict N to the Euclidean unit sphere, which is compact, so N attains a minimum m there; ④ the minimum is positive because N vanishes only at $\mathbf{0}$, which is not on the sphere; ⑤ scale back. Compactness is what turns the infimum over infinitely many unit vectors into a value attained at a particular unit vector, and a value attained at a vector can be tested for positivity one vector at a time.

Proof. By Proposition 16.2.1 it suffices to prove the displayed two-sided bound, since then any two norms N_1, N_2 are each equivalent to $\|\cdot\|_2$ and hence to each other. If $n = 0$ then $V = \{0\}$, both norms are the zero function and $m = M = 1$ works; assume $n \geq 1$. The coordinate map $\Phi\colon F^n \to V$, $\mathbf{x} \mapsto \sum_i x_i v_i$, is an isomorphism (Corollary 2.5.5), and by construction $\|\Phi(\mathbf{x})\|_2 = \|\mathbf{x}\|_2$ is the Euclidean norm of Example (The 1-, 2-, and ∞-norms on F^n). It is enough to prove the two bounds for $\mathbf{x} \in F^n$ with N replaced by $f(\mathbf{x}) \coloneqq N(\Phi(\mathbf{x}))$.

Step 1. The upper bound. Let $M = \left(\sum_{i=1}^{n} N(v_i)^2\right)^{1/2}$. For $\mathbf{x} \in F^n$, the triangle inequality (N3) applied $n-1$ times and then (N2) give

$$f(\mathbf{x}) = N\Big(\sum_i x_i v_i\Big) \leq \sum_i |x_i|\, N(v_i),$$

and Theorem 11.1.5 in $\mathbb{R}^n$, applied to the vectors $(|x_1|, \dots, |x_n|)$ and $(N(v_1), \dots, N(v_n))$, bounds the right-hand side by $M\|\mathbf{x}\|_2$. Hence $f(\mathbf{x}) \leq M\|\mathbf{x}\|_2$ for every $\mathbf{x}$. Also $M > 0$, since $N(v_1) > 0$ by (N1) and $v_1 \neq 0$.

Step 2. f is continuous on F^n for the Euclidean norm. By Lemma 16.1.1 applied to N, and then Step 1 applied to $\mathbf{x} - \mathbf{y}$,

$$|f(\mathbf{x}) - f(\mathbf{y})| \leq N\big(\Phi(\mathbf{x} - \mathbf{y})\big) \leq M\,\|\mathbf{x} - \mathbf{y}\|_2.$$

So f changes by at most M times the distance moved: it is continuous, and the continuity was **deduced** from Step 1 rather than assumed.

Step 3. The minimum on the unit sphere. Let

$$S = \{\mathbf{x} \in F^n : \|\mathbf{x}\|_2 = 1\},$$

which is nonempty because $n \geq 1$ (it contains $\mathbf{e}_1$). It is bounded, and it is closed: if $\|\mathbf{x}_k - \mathbf{x}\|_2 \to 0$ with every $\|\mathbf{x}_k\|_2 = 1$, then Lemma 16.1.1 gives $\big|\|\mathbf{x}\|_2 - 1\big| \leq \|\mathbf{x} - \mathbf{x}_k\|_2 \to 0$, so $\|\mathbf{x}\|_2 = 1$ and $\mathbf{x} \in S$. By **the compactness of closed bounded sets in F^n, fact (A3) of the chapter introduction**, S is compact; by **the extreme value theorem, fact (A4)**, the continuous function f of Step 2 attains a minimum value on S. Say $m = f(\mathbf{x}_0)$ with $\mathbf{x}_0 \in S$.

Step 4. The minimum is positive. Since $\mathbf{x}_0 \in S$ we have $\|\mathbf{x}_0\|_2 = 1 \neq 0$, so $\mathbf{x}_0 \neq \mathbf{0}$ and hence $\Phi(\mathbf{x}_0) \neq 0$, Φ being injective. By (N1) for N, $m = N(\Phi(\mathbf{x}_0)) > 0$. This is the step that needs a minimum rather than an infimum: an infimum of positive numbers can be 0, while a minimum is the value at an actual vector, and there (N1) can be applied.

Step 5. Scaling back. Let $\mathbf{x} \neq \mathbf{0}$. Then $\mathbf{x}/\|\mathbf{x}\|_2 \in S$, so $f(\mathbf{x}/\|\mathbf{x}\|_2) \geq m$, and (N2) for N gives $f(\mathbf{x}) \geq m\|\mathbf{x}\|_2$. For $\mathbf{x} = \mathbf{0}$ both sides are 0. Combining with Step 1,

$$m\,\|\mathbf{x}\|_2 \;\leq\; f(\mathbf{x}) \;\leq\; M\,\|\mathbf{x}\|_2 \qquad (\mathbf{x} \in F^n),$$

and transporting along Φ gives the statement for V. Since $m \leq f(\mathbf{x}_0) \leq M\|\mathbf{x}_0\|_2 = M$, the constants satisfy $0 < m \leq M$. This proves the theorem.

■

The proof used the finite dimension twice, and it is worth saying where: once in Step 1, where the triangle inequality was applied finitely many times to a finite basis expansion, and once in Step 3, where F^n is the space in which closed bounded sets are compact. Remove either and the argument collapses, as the last part of this section shows.

Example. Agreeing on a basis is not enough

On F^n, the norms $\|\cdot\|_1$ and $\|\cdot\|_\infty$ both give every standard basis vector the value 1. Yet $\|\mathbf{1}\|_1 = n$ and $\|\mathbf{1}\|_\infty = 1$. So the data "$N_1(v_i) = N_2(v_i)$ for every basis vector" implies no inequality between N_1 and N_2 beyond what Theorem 16.2.2 gives for any pair at all.

Quick check. In Step 1 the constant was $M = \left(\sum_i N(v_i)^2\right)^{1/2}$. Where would the proof break if we tried the symmetric guess $m = \min_i N(v_i)$ for the lower bound?

Solution. It breaks immediately, and Example (Agreeing on a basis is not enough) is the witness. Take $V = F^2$ with the basis $(\mathbf{e}_1, \mathbf{e}_2)$, $N = \|\cdot\|_\infty$, and $\|\cdot\|_2$ the Euclidean norm. Then $\min_i N(\mathbf{e}_i) = 1$, and the guessed inequality $\|\mathbf{x}\|_\infty \geq \|\mathbf{x}\|_2$ fails at $\mathbf{x} = \mathbf{1}$, where the two sides are 1 and $\sqrt{2}$. The reason is that the triangle inequality only ever bounds $N(\sum x_i v_i)$ from **above**; there is no lower bound in the axioms, because a sum of long vectors can be short. Getting a lower bound needs a genuinely different mechanism, and compactness is that mechanism.

16.2.3 What the theorem buys

The point of the theorem is that whole sentences may now drop the words "in which norm".

Corollary 16.2.3. Convergence and Boundedness Do Not Depend on the Norm

Let V be finite-dimensional over $\mathbb{R}$ or $\mathbb{C}$, let $\|\cdot\|_a$ and $\|\cdot\|_b$ be norms on V, and let (u_k) be a sequence in V with $u \in V$. Then:

(a) $\|u_k - u\|_a \to 0$ if and only if $\|u_k - u\|_b \to 0$;
(b) (u_k) is bounded for $\|\cdot\|_a$ if and only if it is bounded for $\|\cdot\|_b$;
(c) (u_k) is Cauchy for $\|\cdot\|_a$ if and only if it is Cauchy for $\|\cdot\|_b$.

Proof. By Theorem 16.2.2 there are $c, C > 0$ with $c\|v\|_b \le \|v\|_a \le C\|v\|_b$ for every $v \in V$. For (a), apply this to $v = u_k - u$: if $\|u_k - u\|_b \to 0$ then $0 \le \|u_k - u\|_a \le C\|u_k - u\|_b \to 0$, and conversely $\|u_k - u\|_b \le c^{-1}\|u_k - u\|_a \to 0$. For (b), a bound $\|u_k\|_b \le K$ for all k gives $\|u_k\|_a \le CK$, and conversely with c^{-1}. For (c), apply the same two estimates to $v = u_k - u_l$. This proves all three.

■

Now the connection this chapter was built for. Chapter 10 §10 defined convergence of matrices **entrywise** (Definition 10.10.1) and said, in its opening paragraph, that no notion of the size of a matrix was used and that Chapter 16 would supply one. Here is the reconciliation.

Corollary 16.2.4. Entrywise Convergence Is Convergence in Every Norm

Let $\mathbf{M}_1, \mathbf{M}_2, \ldots$ and $\mathbf{M}$ lie in $M_{p\times q}(\mathbb{C})$, and let $\|\cdot\|$ be **any** norm on $M_{p\times q}(\mathbb{C})$. Then $\mathbf{M}_k \to \mathbf{M}$ entrywise in the sense of Definition 10.10.1 if and only if $\|\mathbf{M}_k - \mathbf{M}\| \to 0$. Likewise $(\mathbf{M}_k)$ is bounded in the sense of Definition 10.10.1 if and only if $(\|\mathbf{M}_k\|)$ is a bounded sequence of real numbers.

Idea. Entrywise convergence is convergence in one particular norm, the entrywise maximum. Theorem 16.2.2 says every other norm is trapped between two multiples of it, and a quantity squeezed between two multiples of something tending to 0 tends to 0 as well. Both directions are that one sentence.

Proof. The space $M_{p\times q}(\mathbb{C})$ is a complex vector space of dimension pq, and

$$N_\infty(\mathbf{M}) = \max_{r,s} |m_{rs}|$$

is a norm on it: it is the ∞-norm of Example (The 1-, 2-, and ∞-norms on F^n) read on the list of entries, through the coordinate isomorphism of Corollary 2.5.5 that sends a matrix to its pq entries. By Definition 10.10.1 (a), $\mathbf{M}_k \to \mathbf{M}$ entrywise means $(\mathbf{M}_k)_{rs} \to \mathbf{M}_{rs}$ for each of the finitely many positions (r, s), and since a maximum of finitely many nonnegative sequences tends to 0 exactly when each of them does, this says precisely $N_\infty(\mathbf{M}_k - \mathbf{M}) \to 0$. By Theorem 16.2.2 the norms N_∞ and $\|\cdot\|$ are equivalent, so Corollary 16.2.3 (a) converts this into $\|\mathbf{M}_k - \mathbf{M}\| \to 0$. The statement about boundedness is Definition 10.10.1 (b) together with Corollary 16.2.3 (b), read the same way.

■

Example. Watching one sequence converge

In $M_2(\mathbb{R})$ let

$$\mathbf{M}_k = \begin{pmatrix} \frac{1}{k} & 2 \\ 0 & 1 - \frac{1}{k} \end{pmatrix}, \qquad k \ge 1.$$

Entry by entry, $\frac{1}{k} \to 0$, $2 \to 2$, $0 \to 0$, and $1 - \frac{1}{k} \to 1$, so $\mathbf{M}_k \to \mathbf{M} = \left(\begin{smallmatrix} 0 & 2 \\ 0 & 1 \end{smallmatrix}\right)$ in the sense of Definition 10.10.1. The corollary says this is convergence in *every* norm, and here one can

see it directly: $\mathbf{M}_k - \mathbf{M} = \operatorname{diag}(\frac{1}{k}, -\frac{1}{k})$, so

$$\|\mathbf{M}_k - \mathbf{M}\|_F = \sqrt{\tfrac{1}{k^2} + \tfrac{1}{k^2}} = \frac{\sqrt{2}}{k} \longrightarrow 0,$$

and $\|\mathbf{M}_k - \mathbf{M}\|_\infty$, the largest absolute row sum, is $\frac{1}{k}$. Different norms, different constants, the same verdict — which is the content of the corollary.

So Chapter 10 §10 lost nothing by working entrywise. Its Theorem 10.10.2, Theorem 10.10.3, Theorem 10.10.4, and Theorem 10.10.6 are statements about *the* convergence of matrices, and every norm this chapter introduces — the Frobenius norm, the operator norms of §03, anything at all — measures the same thing. What the norms add is not a new notion of limit but a **rate**: a single number $\|\mathbf{A}^k\|$ that can be bounded, compared, and summed, where an entrywise statement gives pq separate sequences and no way to combine them.

One more corollary, which §03 will use to know that the operator norm exists at all.

Corollary 16.2.5. Closed and Bounded Is Compact, in Every Norm

Let V be finite-dimensional over $\mathbb{R}$ or $\mathbb{C}$. Call a subset $K \subseteq V$ **bounded** for a norm if $K \subseteq \{v : \|v\| \le R\}$ for some real R, and **closed** for that norm if whenever $\|u_k - u\| \to 0$ with every $u_k \in K$, also $u \in K$. Then a subset that is closed and bounded for one norm is closed and bounded for every norm, and is **compact**: every sequence in it has a subsequence converging to a point of it. In particular, when $V \neq \{0\}$, the unit sphere $\{v : \|v\| = 1\}$ of any norm is nonempty and compact, and a continuous real-valued function on it attains a maximum and a minimum.

Idea. Both words in "closed and bounded" are defined by a norm, so the first thing to check is that changing the norm changes neither. That is Theorem 16.2.2 again. Once the words are norm-free, the statement is Theorem 16.2.2's own compactness step read in a coordinate-free way: carry K to F^n by coordinates, apply (A3), and carry the limit back.

Proof. Fix a basis and transport everything to F^n by the coordinate isomorphism (Corollary 2.5.5), as in the proof of Theorem 16.2.2. Let $\|\cdot\|_a$ and $\|\cdot\|_b$ be norms, with $c\|\cdot\|_b \le \|\cdot\|_a \le C\|\cdot\|_b$ from Theorem 16.2.2. A set bounded for one is bounded for the other by the same two estimates as in Corollary 16.2.3 (b). A set closed for one is closed for the other, because by Corollary 16.2.3 (a) the two norms have the same convergent sequences with the same limits, and the notion of closed defined in the statement asks exactly that limits of convergent sequences stay in the set.

For compactness, let $K \subseteq F^n$ be closed and bounded for some norm; by the previous paragraph it is closed and bounded for $\|\cdot\|_2$. By **the compactness of closed bounded sets in F^n, fact (A3) of the chapter introduction**, every sequence in K has a subsequence converging in $\|\cdot\|_2$ to a point of K, and by Corollary 16.2.3 (a) that subsequence converges in every norm to the same point. The unit sphere of a norm N is bounded, and it is closed because $|N(v) - N(u_k)| \le N(v - u_k)$ by Lemma 16.1.1, so a limit of vectors with $N = 1$ has $N = 1$; it is nonempty when $V \neq \{0\}$, since $N(v)^{-1}v$ lies on it for any $v \neq 0$. The last sentence is then **the extreme value theorem, fact (A4)**, applied to that nonempty compact set. This proves the corollary.

■

16.2.4 Two warnings, and where the theorem stops

The theorem is not permission to forget which norm is in use. Here is the first reason.

Warning. **The constants depend on the dimension, and grow.** Example (The three p-norms, with best constants) gives $\|\mathbf{x}\|_1 \le n\|\mathbf{x}\|_\infty$ on F^n, and Exercise (C1: Sharpness of the comparison constants) shows the constant $\sqrt{n}$ in $\|\mathbf{x}\|_1 \le \sqrt{n}\|\mathbf{x}\|_2$ is attained, hence best possible. So a statement such as "the error is at most 10^{-6} in the ∞-norm" translates into the 1-norm as "at most $n \cdot 10^{-6}$", which is a different statement when n is a million. Theorem 16.2.2 says the two norms cannot disagree about **whether** something tends to zero; it says nothing about **how fast**, and every quantitative estimate in the rest of this chapter names its norm for that reason.

The second reason is that the theorem is false without the finite dimension, and not by a narrow margin.

Example. Two inequivalent norms on a space of polynomials

Let $V = \mathbb{R}[x]$, the real polynomials — equivalently, the sequences of real coefficients with only finitely many nonzero terms. For $p = a_0 + a_1x + \cdots + a_dx^d$ put

$$\|p\|_1 = |a_0| + \cdots + |a_d|, \qquad \|p\|_\infty = \max_i |a_i|.$$

Both are norms, by the same coordinatewise checks as in Example (The 1-, 2-, and ∞-norms on F^n); the only difference is that the number of coordinates is unbounded. One inequality survives: $\|p\|_\infty \le \|p\|_1$ always. The other fails completely. Put

$$p_k = 1 + x + x^2 + \cdots + x^{k-1}.$$

Then $\|p_k\|_1 = k$ while $\|p_k\|_\infty = 1$, so a constant C with $\|p\|_1 \le C\|p\|_\infty$ for all p would have to satisfy $C \ge k$ for every k. No such C exists, and the two norms are not equivalent.

Look at what the sequence $(\frac{1}{k}p_k)$ does: it tends to 0 in $\|\cdot\|_\infty$, since $\|\frac{1}{k}p_k\|_\infty = 1/k$, while $\|\frac{1}{k}p_k\|_1 = 1$ for every k. So the two norms disagree about which sequences converge — not about the rate, about the fact. Everything in Corollary 16.2.3 and Corollary 16.2.5 fails here at once, and it fails because Step 3 of the proof of Theorem 16.2.2 has nothing to stand on. The general reason is recorded, and not proved, in the warning below.

Warning. **Infinite dimension is outside the scope of this book.** The closed unit ball of a normed space is compact **only** when the space is finite-dimensional. That is a genuine theorem, due to Riesz, and it is **recorded here, not proved**: it lies outside the four imported facts of the chapter introduction, this book never uses it as a step in an argument, and it is stated only so that Example (Two inequivalent norms on a space of polynomials) is not mistaken for an accident of a badly chosen pair. Every result of this chapter carries the hypothesis $\dim V < \infty$, and on $C[0,1]$ or $\mathbb{R}[x]$ none of them is available. The study of

norms on infinite-dimensional spaces is functional analysis, and there the choice of norm is the main content rather than an irrelevance.

Quick check. On $C[0,1]$, compare the sup norm $\|f\|_\infty$ of Example (Inner products, weights, and a function space) (c) with $\|f\|_1 = \int_0^1 |f|$, which is a norm because a continuous $f \geq 0$ with $\int_0^1 f = 0$ is the zero function — the one fact from analysis that Chapter 11 §01 recorded when it put an inner product on $C[0,1]$, quoted again here. Which of the two inequalities of Definition 16.2.1 holds, and which fails?

Solution. The inequality $\|f\|_1 \leq \|f\|_\infty$ holds, since $\int_0^1 |f| \leq \int_0^1 \|f\|_\infty = \|f\|_\infty$. The other fails. Take $f_k(t) = \max(0, 1 - kt)$ for $k \geq 1$, the spike that drops from 1 to 0 on $[0, 1/k]$. Then $\|f_k\|_\infty = f_k(0) = 1$, while

$$\|f_k\|_1 = \int_0^{1/k} (1 - kt)\,\mathrm{d}t = \frac{1}{2k},$$

the area of a triangle of base $1/k$ and height 1, so no integration theory is needed to evaluate it. A constant C with $\|f\|_\infty \leq C\|f\|_1$ would need $C \geq 2k$ for every k. A tall thin spike has small integral and large maximum, and in infinite dimension nothing stops it from being arbitrarily thin.

With the choice of norm made harmless in finite dimension, the next section can measure matrices without apologizing for which measurement it uses — and then, when the estimates start to matter, name the norm on purpose.

16.2.5 Exercises

16.2.5.1 A. Check your understanding

Exercise. A1

(a) State Definition 16.2.1, and explain why the phrase "not depending on the vector" cannot be dropped.

(b) State Theorem 16.2.2, including its hypothesis on V.

(c) Which two of the four imported analysis facts does the proof of Theorem 16.2.2 use, and at which step is each used?

(d) True or false: if two norms on $\mathbb{R}^n$ are equivalent, then some sequence converges in one and not the other. Justify your answer.

(e) True or false: equivalent norms give the same value to every vector. Justify your answer.

Solution.

(a) Two norms $\|\cdot\|_a, \|\cdot\|_b$ on V are equivalent if there are constants $c, C > 0$, independent of the vector, with $c\|v\|_b \leq \|v\|_a \leq C\|v\|_b$ for every $v \in V$ (Definition 16.2.1). Without the phrase the condition is vacuous: for each $v \neq 0$ the number $\|v\|_a/\|v\|_b$ is positive

and could be taken as both c and C at that one vector, so every pair of norms would qualify.

(b) On a finite-dimensional vector space over $\mathbb{R}$ or $\mathbb{C}$, any two norms are equivalent (Theorem 16.2.2). Finite dimension is essential — see Example (Two inequivalent norms on a space of polynomials).

(c) Fact (A3), the compactness of closed bounded sets in F^n, at Step 3, to know the Euclidean unit sphere is compact; and fact (A4), the extreme value theorem, also at Step 3, to know that the continuous function f attains a minimum there rather than merely approaching an infimum.

(d) False, and this is the content of Corollary 16.2.3 (a): equivalent norms have exactly the same convergent sequences with exactly the same limits.

(e) False. $\|\cdot\|_1$ and $\|\cdot\|_\infty$ on $\mathbb{R}^2$ are equivalent by Example (The three p-norms, with best constants), yet $\|\mathbf{1}\|_1 = 2 \neq 1 = \|\mathbf{1}\|_\infty$. Equivalence constrains ratios, not values.

16.2.5.2 B. Practice

Exercise. B1: Best constants for a weighted norm

On $\mathbb{R}^2$ let $N(\mathbf{x}) = \left(x_1^2 + 4x_2^2\right)^{1/2}$. Find the **best** constants $c, C > 0$ with

$$c\,\|\mathbf{x}\|_\infty \le N(\mathbf{x}) \le C\,\|\mathbf{x}\|_\infty \qquad (\mathbf{x} \in \mathbb{R}^2),$$

and give a vector attaining each. Hence deduce that N and $\|\cdot\|_\infty$ are equivalent.

Solution. Write $a = |x_1|$, $b = |x_2|$, and $M = \max(a, b) = \|\mathbf{x}\|_\infty$. Then $N(\mathbf{x})^2 = a^2 + 4b^2$.

Upper: $a^2 + 4b^2 \le M^2 + 4M^2 = 5M^2$, so $N(\mathbf{x}) \le \sqrt{5}\,\|\mathbf{x}\|_\infty$. Equality needs $a = b = M$, and $\mathbf{x} = (1, 1)$ gives $N(\mathbf{x}) = \sqrt{5}$ and $\|\mathbf{x}\|_\infty = 1$. So $C = \sqrt{5}$ and it cannot be lowered.

Lower: if $M = a$ then $a^2 + 4b^2 \ge a^2 = M^2$, and if $M = b$ then $a^2 + 4b^2 \ge 4b^2 \ge M^2$. Either way $N(\mathbf{x}) \ge \|\mathbf{x}\|_\infty$, so $c = 1$ works, and $\mathbf{x} = \mathbf{e}_1$ gives $N(\mathbf{x}) = \|\mathbf{x}\|_\infty = 1$, so it cannot be raised.

Since 1 and $\sqrt{5}$ are positive constants independent of $\mathbf{x}$, the two norms are equivalent by Definition 16.2.1.

Exercise. B2: Reading a convergence off the entries

In $M_2(\mathbb{R})$ let

$$\mathbf{M}_k = \begin{pmatrix} 1 + \frac{1}{k} & \frac{(-1)^k}{k^2} \\ \frac{1}{2^k} & 3 - \frac{5}{k} \end{pmatrix}.$$

Determine $\lim_k \mathbf{M}_k$ in the sense of Definition 10.10.1, and compute $\|\mathbf{M}_k - \mathbf{M}\|_F$ explicitly. Hence confirm Corollary 16.2.4 for this sequence and the Frobenius norm.

Solution. Each entry converges: $1 + 1/k \to 1$, $(-1)^k/k^2 \to 0$, $2^{-k} \to 0$, $3 - 5/k \to 3$. So

$\mathbf{M}_k \to \mathbf{M} = \begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix}$ entrywise. Then

$$\|\mathbf{M}_k - \mathbf{M}\|_F = \left(\tfrac{1}{k^2} + \tfrac{1}{k^4} + \tfrac{1}{4^k} + \tfrac{25}{k^2}\right)^{1/2},$$

which is at most $(28/k^2)^{1/2} = \sqrt{28}/k$ for $k \geq 1$, since $1/k^4 \leq 1/k^2$ and $4^{-k} \leq 1/k^2$ for $k \geq 1$. Hence $\|\mathbf{M}_k - \mathbf{M}\|_F \to 0$, as Corollary 16.2.4 predicts. Note what the norm added: a single sequence $\sqrt{28}/k$ bounding the whole matrix, where the entrywise statement gave four unrelated sequences.

Exercise. B3: An equivalence with the constants written down

Let $V = \mathbb{R}[x]_{\leq 2}$ with basis $\mathcal{B} = (1, x, x^2)$. For $p \in V$ let $\|p\|_{\mathcal{B}}$ be the Euclidean norm of the coefficient vector, and let $N(p) = |p(0)| + |p(1)| + |p(-1)|$.

(a) Prove that N is a norm on V.

(b) Find an explicit M with $N(p) \leq M\|p\|_{\mathcal{B}}$, following Step 1 of the proof of Theorem 16.2.2.

(c) Explain, without computing it, why a constant $m > 0$ with $N(p) \geq m\|p\|_{\mathcal{B}}$ exists.

Solution.

(a) (N2) and (N3) are inherited pointwise from $|\cdot|$ on $\mathbb{R}$, exactly as for $\|\cdot\|_1$ in Example (The 1-, 2-, and ∞-norms on F^n). For (N1): $N(p) = 0$ forces $p(0) = p(1) = p(-1) = 0$, so p has three distinct roots while $\deg p \leq 2$, hence $p = 0$. This is the only clause with any content, and it is where $\dim V = 3$ and the three sample points match up.

(b) Step 1 takes $M = \left(\sum_i N(v_i)^2\right)^{1/2}$ over the basis. Here $N(1) = 1 + 1 + 1 = 3$, $N(x) = 0 + 1 + 1 = 2$, and $N(x^2) = 0 + 1 + 1 = 2$, so

$$M = (9 + 4 + 4)^{1/2} = \sqrt{17}.$$

(c) V is finite-dimensional, so Theorem 16.2.2 applies to the pair N, $\|\cdot\|_{\mathcal{B}}$ and delivers both constants at once. Only the upper one is computable by hand from the basis; the lower one comes from compactness and the proof gives no formula for it.

16.2.5.3 C. Going deeper

Exercise. C1: Every linear map out of a finite-dimensional space is continuous

Let V be finite-dimensional over F with norm $\|\cdot\|_V$, let W be a normed space over F with norm $\|\cdot\|_W$, and let $T\colon V \to W$ be linear. Fix a basis $\mathcal{B} = (v_1, \dots, v_n)$ of V and write $\|\cdot\|_1$ for the 1-norm of the coordinate vector in the basis $\mathcal{B}$.

(a) Prove that $\|Tv\|_W \leq K\|v\|_1$ for every $v \in V$, where $K = \max_i \|Tv_i\|_W$.

(b) Hence prove that there is a constant L with $\|Tu - Tv\|_W \leq L\|u - v\|_V$ for all $u, v \in V$.

(c) Explain why (b) may fail when V is infinite-dimensional, using Example (Two inequivalent norms on a space of polynomials).

Solution.

(a) Write $v = \sum_i x_i v_i$. By linearity of T, then (N3) and (N2) in W,

$$\|Tv\|_W = \Big\|\sum_i x_i T v_i\Big\|_W \le \sum_i |x_i|\,\|Tv_i\|_W \le K\|v\|_1.$$

(b) The function $\|\cdot\|_1$ is a norm on V (it is $\|\cdot\|_1$ of Example (The 1-, 2-, and ∞-norms on F^n) transported by the coordinate isomorphism Corollary 2.5.5), and V is finite-dimensional, so by Theorem 16.2.2 there is $C > 0$ with $\|v\|_1 \le C\|v\|_V$ for all v. Put $L = KC$. By linearity, $Tu - Tv = T(u - v)$, so (a) gives

$$\|Tu - Tv\|_W \le K\|u - v\|_1 \le L\|u - v\|_V.$$

(c) Part (a) needs a **finite** basis expansion, and part (b) needs Theorem 16.2.2, which is false in infinite dimension. Concretely, take $V = \mathbb{R}[x]$ with the coefficient norm $\|\cdot\|_\infty$ of Example (Two inequivalent norms on a space of polynomials), $W = \mathbb{R}$ with $|\cdot|$, and the linear functional $T(p) = p(1)$, which adds up the coefficients. With $p_k = 1 + x + \cdots + x^{k-1}$ as there, the polynomials $q_k = \frac{1}{k}p_k$ satisfy $\|q_k\|_\infty = 1/k \to 0$ while $T(q_k) = 1$ for every k. So T sends a sequence tending to 0 to a sequence that does not, and no constant L can exist.

Exercise. C2: No dimension-free constants

For each $n \ge 1$ let $c_n, C_n > 0$ be the best constants with

$$c_n\|\mathbf{x}\|_\infty \le \|\mathbf{x}\|_1 \le C_n\|\mathbf{x}\|_\infty \qquad (\mathbf{x} \in F^n).$$

(a) Determine c_n and C_n, with a vector attaining each.

(b) Deduce that there is no pair of constants $c, C > 0$ working simultaneously for every n.

(c) A statement about $\mathbb{R}^n$ says "the error $\mathbf{e}$ satisfies $\|\mathbf{e}\|_\infty \le 10^{-6}$". What is the strongest bound on $\|\mathbf{e}\|_1$ that follows, and for which n does it stop being informative if the entries of $\mathbf{e}$ are themselves of size about 10^{-6}?

Solution.

(a) $c_n = 1$ and $C_n = n$. Indeed $\|\mathbf{x}\|_\infty = |x_k|$ for some k, and $|x_k| \le \sum_i |x_i| = \|\mathbf{x}\|_1$, with equality at $\mathbf{x} = \mathbf{e}_1$; and $\|\mathbf{x}\|_1 = \sum_i |x_i| \le n \max_i |x_i| = n\|\mathbf{x}\|_\infty$, with equality at $\mathbf{x} = \mathbf{1}$. Both constants are therefore attained and cannot be improved.

(b) A dimension-free C would need $C \ge C_n = n$ for every n, which is impossible. (The constant $c = 1$ does happen to work for all n; it is the upper constant that blows up, and one blown-up constant is enough.)

(c) The strongest bound is $\|\mathbf{e}\|_1 \le n \cdot 10^{-6}$, by (a), and it is attained when all entries have modulus 10^{-6} — which is exactly the situation described. So for $n = 10^6$ the bound reads $\|\mathbf{e}\|_1 \le 1$, which says nothing useful about a quantity that is itself of order 1. The two norms agree that the error tends to 0 as the computation improves; they disagree, by a factor of n, about whether it is currently small.

Exercise. C3: A finite-dimensional subspace is closed

Let W be a normed space over F with norm $\|\cdot\|$, and let $U \subseteq W$ be a **finite-dimensional** subspace. Let (u_k) be a sequence in U with $\|u_k - w\| \to 0$ for some $w \in W$.

(a) Prove that (u_k) is bounded in U for the restricted norm.

(b) Using Corollary 16.2.5, produce a subsequence (u_{k_j}) converging to some $u \in U$.

(c) Deduce that $w = u$, so that U contains the limit of every convergent sequence of its own elements. *Hint: the triangle inequality, through* u_{k_j}.

Solution.

(a) By Lemma 16.1.1, $|\|u_k\| - \|w\|| \le \|u_k - w\|$, and the right side tends to 0. So there is K with $\|u_k - w\| \le 1$ for $k \ge K$, whence $\|u_k\| \le \|w\| + 1$ for those k; the finitely many earlier terms are bounded by their maximum. Hence (u_k) is bounded.

(b) The restriction of $\|\cdot\|$ to U is a norm on U, and U is finite-dimensional. The set $K = \{u \in U : \|u\| \le R\}$, with R the bound from (a), is closed and bounded in U, hence compact by Corollary 16.2.5. The sequence (u_k) lies in K, so some subsequence (u_{k_j}) converges to a point $u \in K \subseteq U$.

(c) For every j, by (N3),

$$\|w - u\| \le \|w - u_{k_j}\| + \|u_{k_j} - u\|.$$

Both terms on the right tend to 0 as $j \to \infty$: the first because (u_k) converges to w and a subsequence of a convergent sequence has the same limit, the second by (b). The left side does not depend on j, so $\|w - u\| = 0$, and (N1) gives $w = u \in U$. Hence U contains every such limit, which is what it means for U to be closed in W. Note that W itself was never assumed finite-dimensional; only U was, and that is exactly where Corollary 16.2.5 was applied.

16.3 Measuring a Matrix

Chapter 12 §11 gave the Frobenius norm its own symbol and explained the subscript: a bare $\|\cdot\|$ already meant the norm of a *vector*, and a matrix, it warned, would shortly be asked to carry both kinds of measurement. This is that moment. A matrix is a list of mn numbers, and it is also a map; measuring the list is easy, measuring the map is what we usually want, and the two answers are not the same number. This section builds the second measurement, computes it in the three cases anyone computes, and says exactly where the first one fails.

Throughout, F is $\mathbb{R}$ or $\mathbb{C}$, all spaces are finite-dimensional, and $\|\cdot\|$ with no subscript is a norm in the sense of Definition 16.1.1.

16.3.1 What a matrix norm has to do

Chapter 10 §10 decided whether $\mathbf{A}^m \to \mathbf{O}$ by looking at one entry of one Jordan block at a time. A single number attached to $\mathbf{A}$ would be a far better instrument — but only if it interacts with matrix multiplication, because the questions we want to answer are questions about products: $\mathbf{A}^m$, $\sum_k \mathbf{A}^k$, $e^{\mathbf{A}}$. A norm that satisfies

$$\|\mathbf{AB}\| \le \|\mathbf{A}\|\,\|\mathbf{B}\|$$

immediately gives $\|\mathbf{A}^m\| \le \|\mathbf{A}\|^m$, and with that one line a geometric series is in reach. A norm that does not satisfy it gives nothing at all about powers. So the property is not a technical nicety; it is the entire reason for the definition.

A matrix norm is a norm on matrices that never lets a product be bigger than the product of the sizes.

Definition 16.3.1. Matrix Norm

Let $n \ge 1$. A **matrix norm** on $M_n(F)$ is a function $\|\cdot\|\colon M_n(F) \to \mathbb{R}$ which is a norm on $M_n(F)$ as a vector space (Definition 16.1.1) and which is in addition **submultiplicative**: for **all** $\mathbf{A}, \mathbf{B} \in M_n(F)$,

$$\|\mathbf{AB}\| \le \|\mathbf{A}\|\,\|\mathbf{B}\|.$$

In words: the first requirement is the ordinary one — clauses (N1), (N2), (N3) of Definition 16.1.1 — and it sees $M_n(F)$ only as the vector space F^{n^2} in disguise. The second requirement is the one that notices the multiplication. Note that submultiplicativity is an inequality, not an equality: it is easy for a norm to make $\|\mathbf{AB}\|$ far smaller than $\|\mathbf{A}\|\|\mathbf{B}\|$, and impossible for a matrix norm to make it larger.

The condition needs a square matrix only so that $\mathbf{AB}$ is defined for every pair; nothing else in the definition cares about the shape.

Example. The Frobenius norm is submultiplicative

Show that $\|\mathbf{AB}\|_F \le \|\mathbf{A}\|_F \|\mathbf{B}\|_F$ for all $\mathbf{A}, \mathbf{B} \in M_n(F)$, so that $\|\cdot\|_F$ is a matrix norm.

Solution. The entry $(\mathbf{AB})_{ij} = \sum_k a_{ik} b_{kj}$ is the inner product in F^n of the i-th row of $\mathbf{A}$ with the j-th column of $\mathbf{B}$ (up to a conjugation, which does not change moduli). By Theorem 11.1.5,

$$|(\mathbf{AB})_{ij}|^2 \le \Big(\sum_k |a_{ik}|^2\Big)\Big(\sum_k |b_{kj}|^2\Big).$$

Summing over all i and j, and noting that the two bracketed factors depend only on i and only on j respectively,

$$\|\mathbf{AB}\|_F^2 \le \Big(\sum_{i,k} |a_{ik}|^2\Big)\Big(\sum_{k,j} |b_{kj}|^2\Big) = \|\mathbf{A}\|_F^2\,\|\mathbf{B}\|_F^2.$$

Both sides are nonnegative, so taking square roots gives the claim. That $\|\cdot\|_F$ is a norm was settled in Chapter 11: it is the norm induced by the Frobenius inner product (Definition 11.1.3), so Theorem 11.1.2 and Corollary 11.1.6 apply.

Now the minimal change that breaks it. Keep the idea "read the matrix as a list of numbers" but measure the list with the maximum instead of the Euclidean length.

Example. The entrywise maximum is not submultiplicative

For $\mathbf{A} \in M_n(F)$ put $N(\mathbf{A}) = \max_{i,j} |a_{ij}|$. Show that N is a norm on $M_n(F)$ but not a matrix norm for $n \geq 2$.

Solution. N is the ∞-norm of the list of n^2 entries, so it is a norm (Definition 16.1.1): it vanishes only on the zero matrix, $N(c\mathbf{A}) = |c|N(\mathbf{A})$, and (N3) holds entry by entry.

Submultiplicativity fails already for $n = 2$. Take $\mathbf{A} = \mathbf{B} = \mathbf{J} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$. Then $N(\mathbf{J}) = 1$, while

$$\mathbf{J}^2 = \begin{pmatrix} 2 & 2 \\ 2 & 2 \end{pmatrix}, \qquad N(\mathbf{J}^2) = 2 > 1 = N(\mathbf{J})N(\mathbf{J}).$$

The mechanism is plain: an entry of a product is a sum of n terms, each as large as the largest entry, and a maximum taken entry by entry has no way of charging for the length of that sum. For $n \times n$ the same matrix gives $N(\mathbf{J}^2) = n$.

Warning. **Submultiplicativity is a property of the norm, not of matrices.** The failure above is not caused by $\mathbf{J}$ being special; it is caused by N being the wrong yardstick. Exercise (C2: Repairing the entrywise maximum) shows that the single rescaling $\mathbf{A} \mapsto nN(\mathbf{A})$ repairs it, which is a good way to see that nothing deep is being violated — only a normalization.

16.3.2 The norm a vector norm induces

The definition above is a filter, not a construction: it says which norms are usable, not where to find one. The construction comes from remembering what $\mathbf{A}$ is for. It sends vectors to vectors, and if both are already measured, we can ask how much $\mathbf{A}$ magnifies.

This is exactly the question Chapter 13 §09 asked of the singular value decomposition, and the question Chapter 13 §13 asked of an oblique projection, in each case computing $\max\{\|\mathbf{Ax}\| : \|\mathbf{x}\| = 1\}$ and then promising that the number would get a name here.

Before naming it, we must know that the maximum exists. "Take the largest value of a function over infinitely many vectors" is not something one may write without a reason, and the reason is compactness, already packaged for us in Corollary 16.2.5.

Lemma 16.3.1. The Largest Stretch Is Attained

Let $n, m \geq 1$, let $\|\cdot\|$ be a norm on F^n and also a norm on F^m, and let $\mathbf{A} \in M_{m\times n}(F)$. Then the function $\mathbf{x} \mapsto \|\mathbf{Ax}\|$ attains a maximum on the unit sphere $S = \{\mathbf{x} \in F^n : \|\mathbf{x}\| = 1\}$.

Idea. Two steps, and the second is already packaged for us. ① The function is Lipschitz, hence continuous: Lemma 16.1.1 converts a difference of norms into the norm of a difference, and a crude expansion in the standard basis bounds $\|\mathbf{Az}\|$ by a constant times $\|\mathbf{z}\|$. ② The unit sphere is closed and bounded, hence compact, and a continuous real function on it attains its bounds — which is exactly what Corollary 16.2.5 says.

Proof. **Step 1: a Lipschitz bound.** Write $\mathbf{z} = \sum_{j=1}^{n} z_j \mathbf{e}_j$. By subadditivity and homogeneity,

$$\|\mathbf{Az}\| \le \sum_{j=1}^{n} |z_j| \, \|\mathbf{Ae}_j\| \le \Big(\max_j \|\mathbf{Ae}_j\|\Big) \|\mathbf{z}\|_1.$$

By Theorem 16.2.2, applied on F^n to the norms $\|\cdot\|_1$ and $\|\cdot\|$, there is a constant $c > 0$ with $\|\mathbf{z}\|_1 \le c\|\mathbf{z}\|$ for every $\mathbf{z}$. Hence $\|\mathbf{Az}\| \le M\|\mathbf{z}\|$ with $M = c \max_j \|\mathbf{Ae}_j\|$, a constant depending on $\mathbf{A}$ but not on $\mathbf{z}$.

Step 2: continuity. Let $\mathbf{x}, \mathbf{y} \in F^n$. By Lemma 16.1.1, applied to the vectors $\mathbf{Ax}$ and $\mathbf{Ay}$ of F^m, and then by Step 1,

$$\big|\, \|\mathbf{Ax}\| - \|\mathbf{Ay}\| \,\big| \le \|\mathbf{A}(\mathbf{x} - \mathbf{y})\| \le M\|\mathbf{x} - \mathbf{y}\|.$$

So $\mathbf{x} \mapsto \|\mathbf{Ax}\|$ is Lipschitz with constant M, and in particular continuous.

Step 3: compactness. The set S is nonempty, since $n \ge 1$ and $\mathbf{e}_1/\|\mathbf{e}_1\| \in S$, the division being legal because $\mathbf{e}_1 \neq \mathbf{0}$. By Corollary 16.2.5 the unit sphere of any norm on a finite-dimensional space is compact, and a continuous real-valued function on it attains a maximum. That corollary is where this chapter's imported facts (A3), the compactness of closed bounded sets, and (A4), the extreme value theorem, are cashed in; here we only apply it, to the continuous function of Step 2. This proves the lemma. ∎

The operator norm of a matrix is the worst stretching it can do to a unit vector.

Definition 16.3.2. Operator Norm

Let $n, m \ge 1$ and let a norm $\|\cdot\|$ be given on F^n and on F^m; we write both with the same symbol, the argument saying which is meant. The **operator norm** of $\mathbf{A} \in M_{m\times n}(F)$ **induced by** those norms is

$$\|\mathbf{A}\| := \max\{\, \|\mathbf{Ax}\| : \mathbf{x} \in F^n,\ \|\mathbf{x}\| = 1 \,\},$$

which exists by Lemma 16.3.1. When the vector norm is $\|\cdot\|_p$ on both sides we write $\|\mathbf{A}\|_p$, and $\|\mathbf{A}\|_2$ is also called the **spectral norm**.

Clause by clause: the maximum is over **unit** vectors only, so a matrix that stretches one long

vector enormously is not thereby given a large norm — what counts is the *factor*. The norm on the source and the norm on the target may be different, and then $\|\mathbf{A}\|$ depends on both; the notation hides that, which is why the vector norm must always be named. And the word **induced** is doing real work: not every matrix norm arises this way, as the end of this section shows.

Three equivalent descriptions are worth having at once. For $\mathbf{A} \neq \mathbf{O}$,

$$\|\mathbf{A}\| = \max_{\|\mathbf{x}\|=1} \|\mathbf{A}\mathbf{x}\| = \max_{\mathbf{x}\neq\mathbf{0}} \frac{\|\mathbf{A}\mathbf{x}\|}{\|\mathbf{x}\|} = \max_{\|\mathbf{x}\|\leq 1} \|\mathbf{A}\mathbf{x}\|.$$

The middle expression equals the first because $\|\mathbf{A}\mathbf{x}\|/\|\mathbf{x}\| = \|\mathbf{A}(\mathbf{x}/\|\mathbf{x}\|)\|$ by homogeneity, and $\mathbf{x}/\|\mathbf{x}\|$ runs over S as $\mathbf{x}$ runs over the nonzero vectors. The third equals the first because scaling any $\mathbf{x}$ with $0 < \|\mathbf{x}\| \leq 1$ up to the sphere only increases $\|\mathbf{A}\mathbf{x}\|$, and $\mathbf{x} = \mathbf{0}$ contributes 0.

The smallest cases are worth seeing before any real example, because they pin the scale. For $\mathbf{A} = \mathbf{O}$ every $\mathbf{A}\mathbf{x}$ is $\mathbf{0}$, so $\|\mathbf{O}\| = 0$, as (N1) demands. For $m = n = 1$, a matrix is a scalar (a), the only unit vectors are the scalars of modulus 1, and $\|(a)\| = |a|$: the operator norm of a number is its absolute value, so the definition extends the one-dimensional case rather than replacing it. And for $\mathbf{A} = \mathbf{I}$, every unit vector satisfies $\|\mathbf{I}\mathbf{x}\| = \|\mathbf{x}\| = 1$, so $\|\mathbf{I}\| = 1$ **for every choice of vector norm** — a fact used so often below that it is worth fixing now.

Quick check. Why is the maximum in Definition 16.3.2 taken over $\|\mathbf{x}\| = 1$ rather than over all $\mathbf{x}$? What would the supremum over all $\mathbf{x} \in F^n$ be?

Solution. For a nonzero $\mathbf{A}$ it would be infinite: pick $\mathbf{x}_0$ with $\mathbf{A}\mathbf{x}_0 \neq \mathbf{0}$; then $\|\mathbf{A}(t\mathbf{x}_0)\| = t\|\mathbf{A}\mathbf{x}_0\| \to \infty$ as $t \to \infty$. Restricting to unit vectors measures the *ratio* of output size to input size, which is what "how much does $\mathbf{A}$ magnify?" means, and by homogeneity the ratio is already known on all of F^n once it is known on the sphere.

Theorem 16.3.2. Properties of the Operator Norm

Let $\|\cdot\|$ be a norm on each of F^n, F^m, F^ℓ, and let $\mathbf{A} \in M_{m\times n}(F)$. Then:

(a) $\|\mathbf{A}\mathbf{x}\| \leq \|\mathbf{A}\|\,\|\mathbf{x}\|$ for **every** $\mathbf{x} \in F^n$, and $\|\mathbf{A}\|$ is the smallest constant with this property;

(b) $\|\cdot\|$ is a norm on the vector space $M_{m\times n}(F)$;

(c) $\|\mathbf{I}_n\| = 1$;

(d) $\|\mathbf{A}\mathbf{B}\| \leq \|\mathbf{A}\|\,\|\mathbf{B}\|$ for every $\mathbf{B} \in M_{n\times \ell}(F)$. In particular the operator norm induced on $M_n(F)$ by a single norm on F^n is a matrix norm in the sense of Definition 16.3.1.

Idea. Everything is the definition plus homogeneity. The one move worth naming is in (a): a general $\mathbf{x} \neq \mathbf{0}$ is $\|\mathbf{x}\|$ times a unit vector, so the inequality on the sphere upgrades to an inequality everywhere; and once (a) is available, (d) is two applications of it.

Proof.

(a) For $\mathbf{x} = \mathbf{0}$ both sides are 0. For $\mathbf{x} \neq \mathbf{0}$, the vector $\mathbf{u} = \mathbf{x}/\|\mathbf{x}\|$ satisfies $\|\mathbf{u}\| = 1$, so $\|\mathbf{Au}\| \leq \|\mathbf{A}\|$ by Definition 16.3.2; multiplying by $\|\mathbf{x}\|$ and using homogeneity gives $\|\mathbf{Ax}\| \leq \|\mathbf{A}\|\|\mathbf{x}\|$. If K is any constant with $\|\mathbf{Ax}\| \leq K\|\mathbf{x}\|$ for all $\mathbf{x}$, then taking a unit $\mathbf{x}_0$ at which the maximum is attained (Lemma 16.3.1) gives $\|\mathbf{A}\| = \|\mathbf{Ax}_0\| \leq K$.

(b) Each value is a maximum of nonnegative numbers, so $\|\mathbf{A}\| \geq 0$. If $\|\mathbf{A}\| = 0$ then $\mathbf{Au} = \mathbf{0}$ for every unit $\mathbf{u}$, hence $\mathbf{Ax} = \mathbf{0}$ for every $\mathbf{x}$ by (a), hence $\mathbf{A} = \mathbf{O}$. For a scalar c, $\|c\mathbf{A}\| = \max_{\|\mathbf{x}\|=1} |c|\|\mathbf{Ax}\| = |c|\|\mathbf{A}\|$. For subadditivity, let $\mathbf{x}$ be a unit vector; then $\|(\mathbf{A}+\mathbf{B})\mathbf{x}\| \leq \|\mathbf{Ax}\| + \|\mathbf{Bx}\| \leq \|\mathbf{A}\| + \|\mathbf{B}\|$, and taking the maximum over $\mathbf{x}$ gives $\|\mathbf{A}+\mathbf{B}\| \leq \|\mathbf{A}\| + \|\mathbf{B}\|$.

(c) $\|\mathbf{I}_n\mathbf{x}\| = \|\mathbf{x}\| = 1$ for every unit $\mathbf{x}$, so the maximum is 1.

(d) Let $\mathbf{x} \in F^\ell$ be a unit vector. By (a) applied twice, first to $\mathbf{A}$ and then to $\mathbf{B}$,

$$\|\mathbf{ABx}\| \leq \|\mathbf{A}\|\,\|\mathbf{Bx}\| \leq \|\mathbf{A}\|\,\|\mathbf{B}\|\,\|\mathbf{x}\| = \|\mathbf{A}\|\,\|\mathbf{B}\|.$$

Taking the maximum over unit $\mathbf{x}$ gives $\|\mathbf{AB}\| \leq \|\mathbf{A}\|\|\mathbf{B}\|$. This proves the theorem.

■

Part (c) is small and will do a large amount of work: it is the single property that separates the operator norms from the rest of the matrix norms.

16.3.3 The three formulas

A definition by maximization is not yet a computation. For the three standard vector norms the maximization can be carried out completely, and in two of the three cases the answer is read straight off the entries.

Theorem 16.3.3. Computing the 1-, 2-, and Infinity-Norms

Let $\mathbf{A} = (a_{ij}) \in M_{m\times n}(F)$ with $m, n \geq 1$. Then

(a) $\|\mathbf{A}\|_1 = \max\limits_{1\leq j\leq n} \sum\limits_{i=1}^{m} |a_{ij}|$, the largest absolute column sum;

(b) $\|\mathbf{A}\|_\infty = \max\limits_{1\leq i\leq m} \sum\limits_{j=1}^{n} |a_{ij}|$, the largest absolute row sum;

(c) $\|\mathbf{A}\|_2 = \sigma_1(\mathbf{A})$, the largest singular value of $\mathbf{A}$ (Definition 13.9.1).

Idea. Each part is an upper bound and a witness. ① For $\|\cdot\|_1$ the upper bound is subadditivity of the modulus with the column sums pulled out, and the witness is a standard basis vector $\mathbf{e}_k$ selecting the heaviest column. ② For $\|\cdot\|_\infty$ the upper bound is the same estimate read along a row, and the witness is a vector of unimodular entries chosen to rotate every term of the heaviest row into the positive reals, so that no cancellation occurs. ③ For $\|\cdot\|_2$ the witness is the first right singular vector, and the upper bound is the singular value decomposition: unitary factors are invisible to the Euclidean norm, so only the diagonal matrix $\mathbf{\Sigma}$ is left, and a diagonal matrix stretches most along its largest entry, because it scales each

coordinate separately.

Proof. Write $C = \max_j \sum_i |a_{ij}|$ and $R = \max_i \sum_j |a_{ij}|$.

(a). Let $\|\mathbf{x}\|_1 = 1$. Then

$$\|\mathbf{A}\mathbf{x}\|_1 = \sum_i \Big|\sum_j a_{ij}x_j\Big| \le \sum_j |x_j| \sum_i |a_{ij}| \le C \sum_j |x_j| = C,$$

where the first inequality is subadditivity of the modulus together with an exchange of two finite sums. So $\|\mathbf{A}\|_1 \le C$. For the reverse, choose k with $\sum_i |a_{ik}| = C$. The vector $\mathbf{e}_k$ has $\|\mathbf{e}_k\|_1 = 1$ and $\mathbf{A}\mathbf{e}_k$ is the k-th column of $\mathbf{A}$, so $\|\mathbf{A}\mathbf{e}_k\|_1 = C$. Hence $\|\mathbf{A}\|_1 \ge C$.

(b). Let $\|\mathbf{x}\|_\infty = 1$. For each i,

$$\Big|\sum_j a_{ij}x_j\Big| \le \sum_j |a_{ij}|\,|x_j| \le \sum_j |a_{ij}| \le R,$$

so $\|\mathbf{A}\mathbf{x}\|_\infty \le R$ and $\|\mathbf{A}\|_\infty \le R$. For the reverse, choose k with $\sum_j |a_{kj}| = R$ and define $\mathbf{x}$ by $x_j = \overline{a_{kj}}/|a_{kj}|$ when $a_{kj} \neq 0$ and $x_j = 1$ otherwise. Every x_j has modulus 1, so $\|\mathbf{x}\|_\infty = 1$; and $a_{kj}x_j = |a_{kj}|$ for every j, the choice being legal exactly because we checked $a_{kj} \neq 0$ before dividing. Hence the k-th entry of $\mathbf{A}\mathbf{x}$ is $\sum_j |a_{kj}| = R$, giving $\|\mathbf{A}\mathbf{x}\|_\infty \ge R$.

(c). Let $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$ be a singular value decomposition (Theorem 13.9.1), with $\mathbf{U}, \mathbf{V}$ unitary and $\mathbf{\Sigma}$ carrying $\sigma_1 \ge \cdots \ge \sigma_p \ge 0$ on its diagonal, $p = \min(m, n)$. Let $\|\mathbf{x}\|_2 = 1$ and put $\mathbf{y} = \mathbf{V}^*\mathbf{x}$. A unitary matrix is an isometry for $\|\cdot\|_2$ (Theorem 11.7.3, (a) and (f)), so $\|\mathbf{y}\|_2 = 1$ and $\|\mathbf{A}\mathbf{x}\|_2 = \|\mathbf{U}\mathbf{\Sigma}\mathbf{y}\|_2 = \|\mathbf{\Sigma}\mathbf{y}\|_2$. The vector $\mathbf{\Sigma}\mathbf{y}$ has entries $\sigma_i y_i$ for $i \le p$ and zeros below, so

$$\|\mathbf{A}\mathbf{x}\|_2^2 = \sum_{i=1}^{p} \sigma_i^2 |y_i|^2 \le \sigma_1^2 \sum_{i=1}^{n} |y_i|^2 = \sigma_1^2.$$

Hence $\|\mathbf{A}\|_2 \le \sigma_1$. For the reverse, take $\mathbf{x} = \mathbf{v}_1$, the first column of $\mathbf{V}$; then $\mathbf{y} = \mathbf{V}^*\mathbf{v}_1 = \mathbf{e}_1$, so $\|\mathbf{A}\mathbf{v}_1\|_2 = \|\mathbf{\Sigma}\mathbf{e}_1\|_2 = \sigma_1$, and $\|\mathbf{v}_1\|_2 = 1$ because the columns of a unitary matrix are orthonormal. This proves the theorem.

■

Part (c) is the promise this section was written to keep. Chapter 13 §09 computed $\sigma_1 = \max\{\|\mathbf{A}\mathbf{x}\| : \|\mathbf{x}\| = 1\}$ and said that Chapter 16 would give that quantity a name, the operator norm $\|\mathbf{A}\|_2$, "and this computation is why $\|\mathbf{A}\|_2 = \sigma_1$". The name is Definition 16.3.2; the reason has now been separated from the name, so that the equality is a theorem about two independently defined numbers rather than a definition dressed up as one. Chapter 13 §13 measured an oblique projection by the same maximum and deferred the word twice; Example (The oblique projection, finally measured) below supplies it. And Chapter 12 §11, when it wrote $\|\mathbf{A}\|_F$ with an obligatory subscript, explained that the bare symbol was reserved because a matrix would soon be asked to be measured both as a list and as a map: parts (a)–(c) are that second measurement, and the rest of this section is the difference between the two.

Example. One matrix, four numbers

For $\mathbf{A} = \begin{pmatrix} 3 & 0 \\ 4 & 5 \end{pmatrix} \in M_2(\mathbb{R})$, compute $\|\mathbf{A}\|_1$, $\|\mathbf{A}\|_\infty$, $\|\mathbf{A}\|_F$, and $\|\mathbf{A}\|_2$.

Solution. The column sums of absolute values are $3+4=7$ and $0+5=5$, so $\|\mathbf{A}\|_1 = 7$ by Theorem 16.3.3 (a). The row sums are 3 and 9, so $\|\mathbf{A}\|_\infty = 9$ by (b). The Frobenius norm is $\sqrt{9+16+25} = 5\sqrt{2}$.

For the spectral norm, form

$$\mathbf{A}^{\top}\mathbf{A} = \begin{pmatrix} 25 & 20 \\ 20 & 25 \end{pmatrix}.$$

This matrix is $25\mathbf{I} + 20\left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$, whose eigenvalues are 25 ± 20, namely 45 and 5. So the singular values are $\sigma_1 = 3\sqrt{5}$ and $\sigma_2 = \sqrt{5}$, and $\|\mathbf{A}\|_2 = 3\sqrt{5} \approx 6.708$ by (c).

Two checks come for free. The product $\sigma_1\sigma_2 = 15$ equals $|\det \mathbf{A}| = 15$, and $\sigma_1^2 + \sigma_2^2 = 50 = \|\mathbf{A}\|_F^2$. The four numbers 7, 9, $5\sqrt{2} \approx 7.071$, and $3\sqrt{5} \approx 6.708$ are all different, which is the point: "the size of $\mathbf{A}$" is not a well-posed phrase until the norm is named.

Example. The oblique projection, finally measured

Chapter 13 §13 studied a projection P of F^2 onto a line, with matrix $\left(\begin{smallmatrix} 1 & x \\ 0 & 0 \end{smallmatrix}\right)$ in a suitable orthonormal basis, and found that the largest value of $\|P\mathbf{v}\|$ over unit $\mathbf{v}$ is $\sqrt{1+|x|^2}$, which over $\mathbb{R}$ is $1/\sin\theta$ for the angle θ between image and kernel. Confirm this with Theorem 16.3.3 (c) for $x = -1$, the matrix $\mathbf{P}' = \left(\begin{smallmatrix} 1 & -1 \\ 0 & 0 \end{smallmatrix}\right)$ of Example (Two Oblique Projections, Measured).

Solution. $(\mathbf{P}')^{\top}\mathbf{P}' = \left(\begin{smallmatrix} 1 & -1 \\ -1 & 1 \end{smallmatrix}\right)$, with eigenvalues 2 and 0; so $\sigma_1 = \sqrt{2}$ and $\|\mathbf{P}'\|_2 = \sqrt{2}$. This agrees with $\sqrt{1+|-1|^2} = \sqrt{2}$, and with $1/\sin\theta$ for $\theta = \pi/4$, the angle between the two lines. An orthogonal projection, by contrast, has $\|P_U\|_2 = 1$ whenever $U \neq \{\mathbf{0}\}$, since it fixes every unit vector of U and shortens the rest.

The two matrix norms now in play compare in a way worth recording, because the Frobenius norm is far cheaper to compute and the spectral norm is usually the one wanted.

Proposition 16.3.4. Frobenius Against Spectral

Let $\mathbf{A} \in M_{m\times n}(F)$ be nonzero, of rank r. Then

$$\|\mathbf{A}\|_2 \le \|\mathbf{A}\|_F \le \sqrt{r}\,\|\mathbf{A}\|_2,$$

and the left inequality is an equality if and only if $r = 1$.

Idea. Write both norms in terms of the singular values: $\|\mathbf{A}\|_2 = \sigma_1$ and $\|\mathbf{A}\|_F^2 = \sigma_1^2 + \cdots + \sigma_r^2$. The claim is then the statement that one term of a sum of r nonnegative numbers is at most the sum, which is at most r times the largest — and that the first is an equality exactly when there is only one term.

Proof. By Lemma 11.7.5 and $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$ (Theorem 13.9.1), $\|\mathbf{A}\|_F = \|\mathbf{\Sigma}\|_F$, so $\|\mathbf{A}\|_F^2 = \sum_{i=1}^{p} \sigma_i^2$. Exactly r of the σ_i are nonzero (Theorem 13.9.1), and each satisfies $\sigma_i \le \sigma_1$. Hence

$$\sigma_1^2 \le \sum_{i=1}^{p} \sigma_i^2 \le r\sigma_1^2,$$

which is the displayed chain after taking square roots and using $\|\mathbf{A}\|_2 = \sigma_1$ (Theorem 16.3.3 (c)). The left inequality is an equality exactly when $\sigma_i = 0$ for all $i \ge 2$, that is, when $r = 1$.

■

So the two norms never differ by a factor worse than $\sqrt{r} \le \sqrt{\min(m,n)}$, and they agree exactly on the rank-one matrices. What they do *not* agree on is more interesting.

16.3.4 What submultiplicativity does not buy

The Frobenius norm is a matrix norm: Example (The Frobenius norm is submultiplicative) proved it submultiplicative. It is tempting to conclude that it is therefore the operator norm induced by some vector norm — perhaps a cleverly chosen one. It is not, and the obstruction is a single line.

Warning. **Submultiplicative does not mean induced.** Every operator norm satisfies $\|\mathbf{I}_n\| = 1$ by Theorem 16.3.2 (c), whereas

$$\|\mathbf{I}_n\|_F = \sqrt{\underbrace{1 + 1 + \cdots + 1}_{n}} = \sqrt{n}.$$

So for $n \ge 2$ the Frobenius norm is **not** induced by any norm on F^n whatsoever — not by a bad choice, but by no choice at all. It is a matrix norm and not an operator norm, and the two phrases must be kept apart.

The identity is the cheapest witness, and it is decisive because $\|\mathbf{I}\| = 1$ is forced: the identity stretches nothing, so a norm that measures stretching must give it 1. A norm that assigns the identity the value $\sqrt{n}$ is measuring something else — the amount of material in the matrix, which for $\mathbf{I}_n$ grows with n.

Warning. **Two different objects are called $\|\mathbf{A}\|_\infty$ in the wild.** In this book $\|\mathbf{A}\|_\infty$ is the operator norm induced by $\|\cdot\|_\infty$ on vectors, that is, the largest row sum. The entrywise maximum $N(\mathbf{A}) = \max_{i,j}|a_{ij}|$ of Example (The entrywise maximum is not submultiplicative) is a different function, and it is not even submultiplicative. Whenever a source writes $\|\mathbf{A}\|_\infty$,

check which one is meant by evaluating it at $\mathbf{I}_2$: the row-sum norm gives 1, the entrywise maximum gives 1 as well — so evaluate instead at $\mathbf{J}$, where the row-sum norm gives 2 and the entrywise maximum gives 1.

Quick check. The Frobenius norm satisfies $\|\mathbf{Ax}\|_2 \le \|\mathbf{A}\|_F\|\mathbf{x}\|_2$ for every $\mathbf{x}$, by Proposition 16.3.4 and Theorem 16.3.2 (a). Does that make it an operator norm?

Solution. No. Being *compatible* with a vector norm — satisfying $\|\mathbf{Ax}\| \le \|\mathbf{A}\|\|\mathbf{x}\|$ — is weaker than being induced by it. The induced norm is the **smallest** constant that works (Theorem 16.3.2 (a)), and for $\mathbf{A} = \mathbf{I}_n$ with $n \ge 2$ the smallest constant is 1 while $\|\mathbf{I}_n\|_F = \sqrt{n}$. Compatibility says the number is big enough; being induced says it is exactly right.

Where this leaves us: we have a supply of matrix norms (Definition 16.3.1), among them a distinguished family, the induced ones (Definition 16.3.2), and inside that family three that are computable (Theorem 16.3.3). The next section asks which quantity all of these norms are simultaneously bounding from above, and finds that the answer is the spectral radius.

16.3.5 Exercises

16.3.5.1 A. Check your understanding

Exercise. A1

(a) Define a **matrix norm** on $M_n(F)$, and define the **operator norm** induced by a vector norm.

(b) State the three formulas of Theorem 16.3.3.

(c) Which analysis facts make Definition 16.3.2 legitimate, and where is each used?

(d) Determine whether the following statement is correct, and justify your answer: every matrix norm on $M_n(\mathbb{R})$ is the operator norm induced by some norm on $\mathbb{R}^n$.

(e) Determine whether the following statement is correct, and justify your answer: for every $\mathbf{A} \in M_n(F)$ and every operator norm, $\|\mathbf{A}^2\| = \|\mathbf{A}\|^2$.

Solution.

(a) A matrix norm is a norm on the vector space $M_n(F)$ that also satisfies $\|\mathbf{AB}\| \le \|\mathbf{A}\|\|\mathbf{B}\|$ for all $\mathbf{A}, \mathbf{B}$ (Definition 16.3.1). Given norms on F^n and F^m, the operator norm of $\mathbf{A} \in M_{m\times n}(F)$ is $\|\mathbf{A}\| = \max\{\|\mathbf{Ax}\| : \|\mathbf{x}\| = 1\}$ (Definition 16.3.2).

(b) $\|\mathbf{A}\|_1$ is the largest absolute column sum, $\|\mathbf{A}\|_\infty$ the largest absolute row sum, and $\|\mathbf{A}\|_2 = \sigma_1(\mathbf{A})$.

(c) Two: the compactness of closed bounded sets, fact (A3) of the chapter introduction, used for the unit sphere S; and the extreme value theorem, fact (A4), used for $\mathbf{x} \mapsto \|\mathbf{Ax}\|$. Both enter through Corollary 16.2.5, quoted in Step 3 of Lemma 16.3.1. The continuity needed for the second is proved, not imported, and rests on Theorem 16.2.2 and Lemma 16.1.1.

(d) Incorrect. The Frobenius norm is a matrix norm, and $\|\mathbf{I}_n\|_F = \sqrt{n} \neq 1$ for $n \geq 2$, whereas every operator norm has $\|\mathbf{I}_n\| = 1$ by Theorem 16.3.2 (c).

(e) Incorrect; only $\leq$ holds, by Theorem 16.3.2 (d). Take $\mathbf{A} = \left(\begin{smallmatrix} 0 & 1 \\ 0 & 0 \end{smallmatrix}\right)$, for which $\|\mathbf{A}\|_2 = 1$ but $\mathbf{A}^2 = \mathbf{O}$, so $\|\mathbf{A}^2\|_2 = 0$.

16.3.5.2 B. Practice

Exercise. B1: Four numbers for one matrix

Let $\mathbf{A} = \begin{pmatrix} 2 & -1 \\ 2 & 2 \end{pmatrix} \in M_2(\mathbb{R})$. Compute $\|\mathbf{A}\|_1$, $\|\mathbf{A}\|_\infty$, $\|\mathbf{A}\|_F$, and $\|\mathbf{A}\|_2$, and verify that $\|\mathbf{A}\|_2 \leq \|\mathbf{A}\|_F$.

Solution. Absolute column sums: $2+2 = 4$ and $1+2 = 3$, so $\|\mathbf{A}\|_1 = 4$. Absolute row sums: $2+1 = 3$ and $2+2 = 4$, so $\|\mathbf{A}\|_\infty = 4$. Also $\|\mathbf{A}\|_F = \sqrt{4+1+4+4} = \sqrt{13}$.

For the spectral norm,

$$\mathbf{A}^{\top}\mathbf{A} = \begin{pmatrix} 8 & 2 \\ 2 & 5 \end{pmatrix},$$

with trace 13 and determinant $40 - 4 = 36$, so its eigenvalues solve $\lambda^2 - 13\lambda + 36 = 0$, giving $\lambda = 9$ and $\lambda = 4$. Hence $\sigma_1 = 3$, $\sigma_2 = 2$, and $\|\mathbf{A}\|_2 = 3$ by Theorem 16.3.3 (c). Finally $3 \leq \sqrt{13}$, as Proposition 16.3.4 requires; the inequality is strict because $\operatorname{rank}\mathbf{A} = 2$.

Exercise. B2: A rank-one matrix

Let $\mathbf{u} \in F^m$ and $\mathbf{v} \in F^n$ be nonzero and put $\mathbf{A} = \mathbf{u}\mathbf{v}^*$.

(a) Prove that $\|\mathbf{A}\|_2 = \|\mathbf{u}\|_2\|\mathbf{v}\|_2$.

(b) Verify the formula for $\mathbf{u} = (1, 2)$ and $\mathbf{v} = (2, -1)$ in $\mathbb{R}^2$ by computing $\sigma_1(\mathbf{A})$ directly.

Solution.

(a) For $\mathbf{x} \in F^n$ we have $\mathbf{A}\mathbf{x} = (\mathbf{v}^*\mathbf{x})\mathbf{u}$, so $\|\mathbf{A}\mathbf{x}\|_2 = |\mathbf{v}^*\mathbf{x}|\,\|\mathbf{u}\|_2$. By Theorem 11.1.5, $|\mathbf{v}^*\mathbf{x}| = |\langle \mathbf{x}, \mathbf{v}\rangle| \leq \|\mathbf{x}\|_2\|\mathbf{v}\|_2$, with equality when $\mathbf{x} = \mathbf{v}/\|\mathbf{v}\|_2$. Hence for unit $\mathbf{x}$ the quantity $\|\mathbf{A}\mathbf{x}\|_2$ is at most $\|\mathbf{u}\|_2\|\mathbf{v}\|_2$ and attains that value, so $\|\mathbf{A}\|_2 = \|\mathbf{u}\|_2\|\mathbf{v}\|_2$.

(b) Here $\mathbf{A} = \left(\begin{smallmatrix} 2 & -1 \\ 4 & -2 \end{smallmatrix}\right)$ and $\mathbf{A}^{\top}\mathbf{A} = \left(\begin{smallmatrix} 20 & -10 \\ -10 & 5 \end{smallmatrix}\right)$, of trace 25 and determinant $100 - 100 = 0$. So the eigenvalues are 25 and 0, giving $\sigma_1 = 5$. And $\|\mathbf{u}\|_2\|\mathbf{v}\|_2 = \sqrt{5}\cdot\sqrt{5} = 5$, in agreement. Note also $\|\mathbf{A}\|_F = 5 = \|\mathbf{A}\|_2$, as Proposition 16.3.4 predicts for a rank-one matrix.

Exercise. B3: The star trick

Let $\mathbf{A} \in M_{m\times n}(F)$. Prove that $\|\mathbf{A}^*\mathbf{A}\|_2 = \|\mathbf{A}\|_2^2$. *Hint: the eigenvalues of* $\mathbf{A}^*\mathbf{A}$.

Solution. By Definition 13.9.1 the eigenvalues of $\mathbf{A}^*\mathbf{A}$ are $\sigma_1^2 \geq \cdots \geq \sigma_n^2 \geq 0$, where the σ_i are the singular values of $\mathbf{A}$ padded with zeros if $n > m$. The matrix $\mathbf{A}^*\mathbf{A}$ is positive semidefinite, hence Hermitian, so by Corollary 12.4.3 it is $\mathbf{W}\mathbf{D}\mathbf{W}^*$ with $\mathbf{W}$ unitary and $\mathbf{D} = \operatorname{diag}(\sigma_1^2, \dots, \sigma_n^2)$. Then $(\mathbf{A}^*\mathbf{A})^*(\mathbf{A}^*\mathbf{A}) = \mathbf{W}\mathbf{D}^2\mathbf{W}^*$, whose eigenvalues are σ_i^4, so the singular values of $\mathbf{A}^*\mathbf{A}$ are the σ_i^2 and its largest is σ_1^2. By Theorem 16.3.3 (c), applied to $\mathbf{A}^*\mathbf{A}$ and then to $\mathbf{A}$,

$$\|\mathbf{A}^*\mathbf{A}\|_2 = \sigma_1^2 = \|\mathbf{A}\|_2^2.$$

This proves the claim. Notice that it is an equality, whereas Theorem 16.3.2 (d) alone would give only $\|\mathbf{A}^*\mathbf{A}\|_2 \leq \|\mathbf{A}^*\|_2\|\mathbf{A}\|_2$.

16.3.5.3 C. Going deeper

Exercise. C1: The spectral norm between the other two

Let $\mathbf{A} \in M_{m\times n}(F)$.

(a) Prove that $\|\mathbf{A}\|_2 \leq \sqrt{\|\mathbf{A}\|_1\|\mathbf{A}\|_\infty}$.

(b) Check the inequality on the matrix of Example (One matrix, four numbers), and say how far it is from equality.

Hint for (a): apply Theorem 11.1.5 to the splitting $|a_{ij}| = |a_{ij}|^{1/2}\cdot|a_{ij}|^{1/2}|x_j|$ *inside one row.*

Solution.

(a) Write $C = \|\mathbf{A}\|_1$ and $R = \|\mathbf{A}\|_\infty$ for the largest column and row sums (Theorem 16.3.3). Let $\|\mathbf{x}\|_2 = 1$. For each i, Theorem 11.1.5 applied to the vectors with entries $|a_{ij}|^{1/2}$ and $|a_{ij}|^{1/2}|x_j|$ gives

$$\Big|\sum_j a_{ij}x_j\Big|^2 \leq \Big(\sum_j |a_{ij}|\Big)\Big(\sum_j |a_{ij}|\,|x_j|^2\Big) \leq R\sum_j |a_{ij}|\,|x_j|^2.$$

Summing over i and exchanging the two finite sums,

$$\|\mathbf{A}\mathbf{x}\|_2^2 \leq R\sum_j |x_j|^2 \sum_i |a_{ij}| \leq RC\sum_j |x_j|^2 = RC.$$

Taking the maximum over unit $\mathbf{x}$ gives $\|\mathbf{A}\|_2^2 \leq RC$, as claimed.

(b) There $\|\mathbf{A}\|_1 = 7$ and $\|\mathbf{A}\|_\infty = 9$, so the bound is $\sqrt{63} \approx 7.937$, while $\|\mathbf{A}\|_2 = 3\sqrt{5} \approx 6.708$. The bound overestimates by about 18 per cent. It is sharp in general: for $\mathbf{A} = \mathbf{I}_n$ all three numbers are 1.

Exercise. C2: Repairing the entrywise maximum

For $\mathbf{A} \in M_n(F)$ let $N(\mathbf{A}) = \max_{i,j}|a_{ij}|$, as in Example (The entrywise maximum is not submultiplicative).

(a) Prove that $N(\mathbf{AB}) \le nN(\mathbf{A})N(\mathbf{B})$, and deduce that $\mathbf{A} \mapsto nN(\mathbf{A})$ **is** a matrix norm.

(b) Deduce that $nN(\cdot)$ is not an operator norm for any $n \ge 2$.

(c) Explain why (a) and (b) together show that submultiplicativity and $\|\mathbf{I}\| = 1$ are independent requirements.

Solution.

(a) For all i, j,

$$|(\mathbf{AB})_{ij}| = \left|\sum_k a_{ik}b_{kj}\right| \le \sum_k |a_{ik}|\,|b_{kj}| \le n\,N(\mathbf{A})N(\mathbf{B}),$$

so $N(\mathbf{AB}) \le nN(\mathbf{A})N(\mathbf{B})$. Put $\|\mathbf{A}\| = nN(\mathbf{A})$. It is a norm, being a positive multiple of the norm N, and

$$\|\mathbf{AB}\| = nN(\mathbf{AB}) \le n^2 N(\mathbf{A})N(\mathbf{B}) = \|\mathbf{A}\|\,\|\mathbf{B}\|.$$

(b) $\|\mathbf{I}_n\| = nN(\mathbf{I}_n) = n \ne 1$ for $n \ge 2$, while every operator norm gives the identity the value 1 (Theorem 16.3.2 (c)).

(c) N itself is a norm with $N(\mathbf{I}_n) = 1$ that is not submultiplicative; nN is submultiplicative with $nN(\mathbf{I}_n) = n$. So neither property implies the other, and an operator norm is precisely a matrix norm that also passes the identity test — a necessary condition which, as $\|\cdot\|_F$ already showed, is still not sufficient on its own for n large.

Exercise. C3: When is a bound an equality?

Let $\mathbf{A} \in M_{m\times n}(F)$ be nonzero.

(a) Prove that $\|\mathbf{A}\|_2 = \|\mathbf{A}\|_F$ if and only if $\operatorname{rank}\mathbf{A} = 1$.

(b) Give an example with $m = n = 2$ where $\|\mathbf{A}\|_F = \sqrt{2}\,\|\mathbf{A}\|_2$, the worst case allowed by Proposition 16.3.4.

(c) Is there a nonzero $\mathbf{A}$ with $\|\mathbf{A}\|_F < \|\mathbf{A}\|_2$? Justify your answer.

Solution.

(a) This is the equality case of Proposition 16.3.4, proved there: $\|\mathbf{A}\|_F^2 = \sum_i \sigma_i^2$ and $\|\mathbf{A}\|_2^2 = \sigma_1^2$, so the two agree exactly when $\sigma_i = 0$ for every $i \ge 2$, which by Theorem 13.9.1 says $\operatorname{rank}\mathbf{A} = 1$.

(b) Take $\mathbf{A} = \mathbf{I}_2$. Its singular values are $1, 1$, so $\|\mathbf{A}\|_2 = 1$ and $\|\mathbf{A}\|_F = \sqrt{2}$. Any matrix with two equal nonzero singular values does the same, for instance any 2×2 unitary matrix.

(c) No. $\|\mathbf{A}\|_2 = \sigma_1 \le \left(\sum_i \sigma_i^2\right)^{1/2} = \|\mathbf{A}\|_F$ always, since every term of the sum is nonneg-

ative. The Frobenius norm counts all the singular values and the spectral norm only the largest, so the Frobenius norm can never be the smaller of the two.

16.4 The Spectral Radius

Every criterion in Chapter 10 §10 began by looking at the largest modulus of an eigenvalue, and that section gave the number a name in passing — the spectral radius — without a numbered definition and without a single norm available to compare it with. It also left a note: "Chapter 16 measures matrices by norms, proves that the spectral radius never exceeds the norm of $\mathbf{A}$ induced by any vector norm, and recovers $\rho(\mathbf{A})$ as a limit built from the norms of the powers $\mathbf{A}^m$." This section pays that note in full, and is careful about what is new when it does.

Throughout, matrices are complex. A real matrix is read inside $M_n(\mathbb{C})$, exactly as Chapter 10 §10 read it when it counted eigenvalues; over $\mathbb{R}$ a matrix may have no eigenvalue at all, and then there is nothing to take the largest modulus of.

16.4.1 The number Chapter 10 named

The spectral radius is the radius of the smallest disk about the origin that holds the whole spectrum.

Definition 16.4.1. Spectral Radius

Let $\mathbf{A} \in M_n(\mathbb{C})$ with $n \geq 1$. The **spectral radius** of $\mathbf{A}$ is

$$\rho(\mathbf{A}) := \max\{\, |\lambda| : \lambda \in \operatorname{spec}(\mathbf{A}) \,\}.$$

For an operator T on a nonzero finite-dimensional complex vector space, $\rho(T)$ is defined by the same formula with $\operatorname{spec}(T)$ in place of $\operatorname{spec}(\mathbf{A})$.

The maximum is over a **finite nonempty** set, so it exists with no analysis at all: $\operatorname{spec}(\mathbf{A})$ is the root set of the characteristic polynomial $p_{\mathbf{A}}$, which is nonempty by the Fundamental Theorem of Algebra (Theorem 6.5.1) and has at most n elements. It is the largest modulus of an eigenvalue, **not** the largest eigenvalue: over $\mathbb{C}$ there is no order to take a largest in, and even for a real spectrum $\rho(\operatorname{diag}(-5, 1)) = 5$, not 1. This is the number Chapter 10 §10 introduced in prose, now with a label so that later sections can cite it.

Some values, simplest first.

- **Diagonal or triangular.** The eigenvalues of a triangular matrix are its diagonal entries (Theorem 9.7.4 (b)), so ρ is the largest modulus on the diagonal. For $\operatorname{diag}(d_1, \dots, d_n)$ it is $\max_i |d_i|$.
- **Normal.** If $\mathbf{A}$ is normal then $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{U}^*$ with $\mathbf{D} = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$ (Corollary 12.4.3), so $\mathbf{A}^*\mathbf{A} = \mathbf{U}\operatorname{diag}(|\lambda_i|^2)\mathbf{U}^*$ and the singular values of $\mathbf{A}$ are the moduli $|\lambda_i|$. Hence $\rho(\mathbf{A}) = \sigma_1(\mathbf{A}) = \|\mathbf{A}\|_2$ by Theorem 16.3.3 (c). For a normal matrix the spectral radius **is** the spectral norm, which is where the second name comes from.
- **A rotation.** $\left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$ has eigenvalues $\pm i$, so $\rho = 1$, although no real vector is left in place.
- **Nilpotent.** $\mathbf{J}_k(0)$ is triangular with zero diagonal, so by the first bullet $\rho(\mathbf{J}_k(0)) = 0$. The matrix itself is not $\mathbf{O}$ when $k \geq 2$.

The last entry is the degenerate case, and it is the important one: it says at once that ρ is not a norm.

Warning. **ρ is not a norm, and it is neither subadditive nor submultiplicative.** Take

$$\mathbf{A} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \qquad \mathbf{B} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}.$$

Both are nilpotent, so $\rho(\mathbf{A}) = \rho(\mathbf{B}) = 0$ although neither is the zero matrix. Their sum $\left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$ has eigenvalues ± 1 and their product $\mathbf{AB} = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 0 \end{smallmatrix}\right)$ has eigenvalues 1 and 0. So $\rho(\mathbf{A}+\mathbf{B}) = 1 > 0 = \rho(\mathbf{A}) + \rho(\mathbf{B})$ and $\rho(\mathbf{AB}) = 1 > 0 = \rho(\mathbf{A})\rho(\mathbf{B})$. So positive definiteness and subadditivity both fail, and so does submultiplicativity, all on the same small pair. Only absolute homogeneity survives, since $\rho(c\mathbf{A}) = |c|\rho(\mathbf{A})$ for every scalar c.

So ρ is a genuinely different kind of measurement from the ones of the previous section. The relation between the two is the subject of the rest of this section, and it is one-sided.

16.4.2 Every norm dominates the spectral radius

Theorem 16.4.1. The Spectral Radius Is a Lower Bound for Every Norm

Let $\mathbf{A} \in M_n(\mathbb{C})$.

(a) If $\|\cdot\|$ is the operator norm on $M_n(\mathbb{C})$ induced by a norm on $\mathbb{C}^n$ (Definition 16.3.2), then $\rho(\mathbf{A}) \le \|\mathbf{A}\|$.

(b) More generally, $\rho(\mathbf{A}) \le \|\mathbf{A}\|$ for **every** matrix norm $\|\cdot\|$ on $M_n(\mathbb{C})$ (Definition 16.3.1).

Idea. For (a) there is nothing to do but write down the eigenvalue equation: an eigenvector is a vector on which $\mathbf{A}$ acts by multiplication by λ, so the stretching factor $|\lambda|$ is one of the factors the maximum in Definition 16.3.2 ranges over. For (b) there is no vector to feed the norm, only matrices — so manufacture one: fill an $n \times n$ matrix with copies of the eigenvector, and the eigenvalue equation becomes a matrix equation that submultiplicativity can be applied to.

Proof.

(a) Let $\lambda \in \operatorname{spec}(\mathbf{A})$ and let $\mathbf{x} \neq \mathbf{0}$ with $\mathbf{Ax} = \lambda\mathbf{x}$. Then $\|\mathbf{Ax}\| = |\lambda|\,\|\mathbf{x}\|$ by homogeneity, while $\|\mathbf{Ax}\| \le \|\mathbf{A}\|\|\mathbf{x}\|$ by Theorem 16.3.2 (a). Since $\mathbf{x} \neq \mathbf{0}$ we have $\|\mathbf{x}\| > 0$ and may divide, giving $|\lambda| \le \|\mathbf{A}\|$. This holds for every eigenvalue, so it holds for the largest modulus.

(b) Let λ and $\mathbf{x}$ be as above, and let $\mathbf{X} \in M_n(\mathbb{C})$ be the matrix all of whose n columns equal $\mathbf{x}$. Then $\mathbf{X} \neq \mathbf{O}$, so $\|\mathbf{X}\| > 0$. Computing column by column, $\mathbf{AX}$ has every column equal to $\mathbf{Ax} = \lambda\mathbf{x}$, that is,

$$\mathbf{AX} = \lambda\mathbf{X}.$$

Taking norms and using homogeneity on the left and submultiplicativity on the right,

$$|\lambda|\,\|\mathbf{X}\| = \|\mathbf{A}\mathbf{X}\| \le \|\mathbf{A}\|\,\|\mathbf{X}\|.$$

Dividing by $\|\mathbf{X}\| > 0$ gives $|\lambda| \le \|\mathbf{A}\|$, and again the largest modulus obeys the same bound. This proves the theorem.

■

This is the inequality Chapter 10 §10 promised: the spectral radius never exceeds the norm of $\mathbf{A}$ induced by any vector norm. Part (b) says more than was promised, and costs one extra line.

The immediate use is that every computable norm is a computable *upper bound* for every eigenvalue — and two of the three formulas of Theorem 16.3.3 need no arithmetic beyond adding up moduli.

Corollary 16.4.2. Cheap Bounds on the Spectrum

Let $\mathbf{A} \in M_n(\mathbb{C})$. Then every eigenvalue λ of $\mathbf{A}$ satisfies

$$|\lambda| \le \rho(\mathbf{A}) \le \min\{\|\mathbf{A}\|_1,\ \|\mathbf{A}\|_\infty,\ \|\mathbf{A}\|_F\},$$

that is, $\rho(\mathbf{A})$ is at most the largest absolute column sum, at most the largest absolute row sum, and at most $\|\mathbf{A}\|_F$.

Proof. The first two are operator norms (Theorem 16.3.3 (a), (b)), so Theorem 16.4.1 (a) applies; $\|\cdot\|_F$ is a matrix norm (Example (The Frobenius norm is submultiplicative)), so Theorem 16.4.1 (b) applies. A minimum of upper bounds is an upper bound.

■

Remark. The bound $\rho(\mathbf{A}) \le \|\mathbf{A}\|_2 = \sigma_1(\mathbf{A})$ is the same crude inequality $|\lambda| \le \sigma_1$ that Chapter 13 §09 obtained directly from $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$. It is now one instance of a general principle rather than an isolated computation, and Theorem 16.4.1 (a) is that computation with the norm left unnamed.

Example. Bounding a spectrum without finding it

Let $\mathbf{A} = \frac{1}{10}\begin{pmatrix} 2 & 5 & 1 \\ 3 & 1 & 4 \\ 1 & 2 & 5 \end{pmatrix}$. Show that $\rho(\mathbf{A}) \le \frac{4}{5}$, and hence that $\mathbf{A}^m \to \mathbf{O}$.

Solution. Every row of $\mathbf{A}$ sums to $\frac{8}{10} = \frac{4}{5}$, so $\|\mathbf{A}\|_\infty = \frac{4}{5}$ by Theorem 16.3.3 (b), and $\rho(\mathbf{A}) \le \frac{4}{5}$ by Corollary 16.4.2. Since $\frac{4}{5} < 1$, every eigenvalue has modulus less than 1 and $\mathbf{A}^m \to \mathbf{O}$ by Theorem 10.10.2. No characteristic polynomial was computed.

The column sums are $\frac{6}{10}$, $\frac{8}{10}$, and 1, so $\|\mathbf{A}\|_1 = 1$ and the column bound gives nothing. It is worth computing both and keeping the smaller. In fact $\rho(\mathbf{A}) = \frac{4}{5}$ exactly, because the constant row sums make $\mathbf{1}$ an eigenvector with eigenvalue $\frac{4}{5}$, so the row bound here is attained.

Quick check. Theorem 16.4.1 says $\rho(\mathbf{A}) \le \|\mathbf{A}\|$ for every matrix norm. Why does it not follow that ρ is the smallest matrix norm?

Solution. Because ρ is not a matrix norm at all: ρ vanishes on every nonzero nilpotent matrix, and it fails both subadditivity and submultiplicativity, as the warning above showed. The correct statement is that $\rho(\mathbf{A})$ is a lower bound for the *values* $\|\mathbf{A}\|$, one matrix at a time, and the next theorem shows it is the greatest such lower bound.

16.4.3 A norm as close to the spectral radius as we like

Theorem 16.4.1 leaves open how large the gap can be. It can be enormous for a fixed norm: $\mathbf{J}_2(0)$ has $\rho = 0$ and $\|\mathbf{J}_2(0)\|_2 = 1$, and multiplying the corner entry by 10^6 leaves $\rho = 0$ while the norm becomes 10^6. But the norm may be chosen after the matrix, and then the gap can be made as small as we please.

Theorem 16.4.3. Norms Close to the Spectral Radius

Let $\mathbf{A} \in M_n(\mathbb{C})$ and let $\varepsilon > 0$. Then there is a norm $\|\cdot\|_t$ on $\mathbb{C}^n$ whose induced operator norm satisfies

$$\rho(\mathbf{A}) \le \|\mathbf{A}\|_t \le \rho(\mathbf{A}) + \varepsilon.$$

Idea. Schur-triangularize and then squash. ① By Corollary 12.1.3, $\mathbf{A} = \mathbf{U}\mathbf{T}\mathbf{U}^*$ with $\mathbf{T}$ upper triangular carrying the eigenvalues on its diagonal; the row sums of $\mathbf{T}$ are the diagonal entry plus the off-diagonal junk. ② Conjugating $\mathbf{T}$ by $\mathbf{D}_t = \operatorname{diag}(1, t, \dots, t^{n-1})$ multiplies the entry in position (i, j) by t^{i-j}, which leaves the diagonal alone and, for large t, shrinks everything strictly above it, because $i - j < 0$ there. ③ Choose t so that every off-diagonal row sum is below ε; the largest row sum is then below $\rho(\mathbf{A}) + \varepsilon$, and the largest row sum is a norm we can compute, by Theorem 16.3.3 (b). ④ Transport that norm back to $\mathbb{C}^n$ along the invertible matrix $\mathbf{D}_t\mathbf{U}^*$.

Proof. By Corollary 12.1.3 there are a unitary $\mathbf{U}$ and an upper triangular $\mathbf{T} = (t_{ij})$ with $\mathbf{A} = \mathbf{U}\mathbf{T}\mathbf{U}^*$ and $t_{ii} = \lambda_i$, the eigenvalues of $\mathbf{A}$. For $t \ge 1$ put $\mathbf{D}_t = \operatorname{diag}(1, t, t^2, \dots, t^{n-1})$, an invertible matrix, and $\mathbf{T}_t = \mathbf{D}_t \mathbf{T} \mathbf{D}_t^{-1}$. Multiplying out,

$$(\mathbf{T}_t)_{ij} = t^{\,i-1}\, t_{ij}\, t^{-(j-1)} = t^{\,i-j} t_{ij}.$$

So $\mathbf{T}_t$ is again upper triangular with the same diagonal $\lambda_1, \dots, \lambda_n$, while for $i < j$ the entry is scaled by $t^{i-j} \le t^{-1}$, using $t \ge 1$ and $j - i \ge 1$. Let $M = \max_{i<j} |t_{ij}|$ if $n \ge 2$, and $M = 0$ if $n = 1$. Then the i-th absolute row sum of $\mathbf{T}_t$ is at most

$$|\lambda_i| + \sum_{j>i} t^{-1}|t_{ij}| \le \rho(\mathbf{A}) + \frac{(n-1)M}{t}.$$

Choose $t = \max(1,\ (n-1)M/\varepsilon)$. Then every absolute row sum of $\mathbf{T}_t$ is at most $\rho(\mathbf{A}) + \varepsilon$, so $\|\mathbf{T}_t\|_\infty \le \rho(\mathbf{A}) + \varepsilon$ by Theorem 16.3.3 (b).

Now define, for $\mathbf{x} \in \mathbb{C}^n$,

$$\|\mathbf{x}\|_t := \|\mathbf{D}_t \mathbf{U}^* \mathbf{x}\|_\infty.$$

This is a norm: it is nonnegative, homogeneous, and subadditive because $\|\cdot\|_\infty$ is and $\mathbf{x} \mapsto \mathbf{D}_t\mathbf{U}^*\mathbf{x}$ is linear, and it vanishes only at $\mathbf{x} = \mathbf{0}$ because $\mathbf{D}_t\mathbf{U}^*$ is invertible.

It remains to compute the induced operator norm. Put $\mathbf{Q} = \mathbf{D}_t\mathbf{U}^*$, so $\|\mathbf{x}\|_t = \|\mathbf{Q}\mathbf{x}\|_\infty$. As $\mathbf{x}$ runs over $\{\|\mathbf{x}\|_t = 1\}$, the vector $\mathbf{y} = \mathbf{Q}\mathbf{x}$ runs over $\{\|\mathbf{y}\|_\infty = 1\}$, since $\mathbf{Q}$ is a bijection. Moreover $\mathbf{Q}\mathbf{A}\mathbf{Q}^{-1} = \mathbf{D}_t\mathbf{U}^*\mathbf{A}\mathbf{U}\mathbf{D}_t^{-1} = \mathbf{D}_t\mathbf{T}\mathbf{D}_t^{-1} = \mathbf{T}_t$, so

$$\|\mathbf{A}\mathbf{x}\|_t = \|\mathbf{Q}\mathbf{A}\mathbf{x}\|_\infty = \|\mathbf{T}_t\,\mathbf{y}\|_\infty.$$

Taking maxima over the two matching sets gives $\|\mathbf{A}\|_t = \|\mathbf{T}_t\|_\infty \le \rho(\mathbf{A}) + \varepsilon$. The lower bound $\rho(\mathbf{A}) \le \|\mathbf{A}\|_t$ is Theorem 16.4.1 (a). This proves the theorem.

■

Read together, the last two theorems say

$$\rho(\mathbf{A}) = \inf\{\|\mathbf{A}\| : \|\cdot\| \text{ an induced matrix norm}\}.$$

Remark. The infimum need not be attained. If $\mathbf{A} \ne \mathbf{O}$ is nilpotent then $\rho(\mathbf{A}) = 0$ while $\|\mathbf{A}\| > 0$ for every norm, so no norm achieves the value. At the other extreme, for a normal $\mathbf{A}$ the spectral norm achieves it exactly: $\|\mathbf{A}\|_2 = \rho(\mathbf{A})$, as computed at the start of this section. The gap between ρ and the norms is a measure of how far $\mathbf{A}$ is from being normal, which is the theme Chapter 12 §11 opened with the departure from normality.

16.4.4 Gelfand's formula

The infimum above is over norms, and to use it we must choose one. There is a second description of $\rho(\mathbf{A})$ that needs no choice: fix any norm at all, and let the powers of $\mathbf{A}$ tell you the answer.

Theorem 16.4.4. Gelfand's Formula

Let $\mathbf{A} \in M_n(\mathbb{C})$ and let $\|\cdot\|$ be any matrix norm on $M_n(\mathbb{C})$. Then

$$\lim_{k\to\infty} \|\mathbf{A}^k\|^{1/k} = \rho(\mathbf{A}).$$

In particular the limit exists and does not depend on which matrix norm is used.

Idea. Two inequalities, one for each direction, and they come from opposite ends of the chapter. ① The lower bound $\rho(\mathbf{A}) \le \|\mathbf{A}^k\|^{1/k}$ is Theorem 16.4.1 applied to $\mathbf{A}^k$, once we know $\rho(\mathbf{A}^k) = \rho(\mathbf{A})^k$ — which is the Spectral Mapping Theorem. This holds for every k, with no limit taken. ② For the upper bound, scale $\mathbf{A}$ down by a whisker more than $\rho(\mathbf{A})$. The scaled matrix has spectral radius below 1, so Chapter 10 §10's criterion makes its powers tend to $\mathbf{O}$, and once they are below 1 in norm the inequality we want has been written down. Theorem 16.4.3 says the same decay is geometric, which is what the corollary below extracts.

Proof. **Lower bound.** Fix $k \ge 1$. By Theorem 9.10.1 applied to $q(x) = x^k$, $\operatorname{spec}(\mathbf{A}^k) = \{\lambda^k : \lambda \in \operatorname{spec}(\mathbf{A})\}$, so

$$\rho(\mathbf{A}^k) = \max_{\lambda\in\operatorname{spec}(\mathbf{A})} |\lambda|^k = \rho(\mathbf{A})^k.$$

By Theorem 16.4.1 (b) applied to $\mathbf{A}^k$, $\rho(\mathbf{A})^k = \rho(\mathbf{A}^k) \le \|\mathbf{A}^k\|$. Taking k-th roots of these nonnegative reals,

$$\rho(\mathbf{A}) \le \|\mathbf{A}^k\|^{1/k} \qquad \text{for every } k \ge 1.$$

Upper bound. Let $\varepsilon > 0$, put $s = \rho(\mathbf{A}) + \varepsilon > 0$ and $\mathbf{B} = s^{-1}\mathbf{A}$. Every eigenvalue of $\mathbf{B}$ is λ/s for some $\lambda \in \operatorname{spec}(\mathbf{A})$ (Theorem 9.10.1 with $q(x) = x/s$), so

$$\rho(\mathbf{B}) = \frac{\rho(\mathbf{A})}{\rho(\mathbf{A}) + \varepsilon} < 1.$$

By Theorem 10.10.2, $\mathbf{B}^k \to \mathbf{O}$ entrywise. Entrywise convergence of matrices is convergence in **every** norm on $M_n(\mathbb{C})$, by Corollary 16.2.4, so $\|\mathbf{B}^k\| \to 0$. A real sequence tending to 0 is eventually below 1: there is K with $\|\mathbf{B}^k\| \le 1$ for all $k \ge K$. Since $\mathbf{B}^k = s^{-k}\mathbf{A}^k$, homogeneity gives $\|\mathbf{A}^k\| \le s^k$, that is,

$$\|\mathbf{A}^k\|^{1/k} \le \rho(\mathbf{A}) + \varepsilon \qquad (k \ge K).$$

Conclusion. Combining, for every $\varepsilon > 0$ there is K with

$$\rho(\mathbf{A}) \le \|\mathbf{A}^k\|^{1/k} \le \rho(\mathbf{A}) + \varepsilon \qquad (k \ge K),$$

which is the definition of $\|\mathbf{A}^k\|^{1/k} \to \rho(\mathbf{A})$. The independence of the norm is automatic, since the limit equals $\rho(\mathbf{A})$ whichever matrix norm was fixed at the start. This proves the theorem.

■

Remark. **Which analysis was used.** The lower bound used none. The upper bound used Theorem 10.10.2, which is a theorem of this book, and Corollary 16.2.4, likewise; the only analysis is that a real sequence tending to 0 is eventually less than 1, and that a sequence trapped between $\rho(\mathbf{A})$ and $\rho(\mathbf{A})+\varepsilon$ for every $\varepsilon>0$ converges to $\rho(\mathbf{A})$. Neither is an extra import: both are the definition of a limit of a real sequence, read forwards and backwards. Nothing from the list (A1)–(A4) of the chapter introduction is needed here, and nothing outside it is used.

The formula is a genuine computation rule as well as a theorem. For a fixed k the number $\|\mathbf{A}^k\|^{1/k}$ is an *upper* estimate of $\rho(\mathbf{A})$ — that is the lower bound above, read the other way — and the estimates improve as k grows.

Example. Watching the limit converge

Let $\mathbf{A}=\begin{pmatrix}1/2 & 10\\ 0 & 1/3\end{pmatrix}$. Compute $\|\mathbf{A}^k\|_\infty$ exactly and confirm Gelfand's formula.

Solution. $\mathbf{A}$ is upper triangular, so $\rho(\mathbf{A})=\max(\tfrac12,\tfrac13)=\tfrac12$ and the powers are upper triangular with diagonal $2^{-k},3^{-k}$. For the corner entry, induction on k gives

$$(\mathbf{A}^k)_{12}=10\sum_{j=0}^{k-1}2^{-j}3^{-(k-1-j)}=60(2^{-k}-3^{-k}),$$

the last step by summing the geometric series with ratio $3/2$. The two absolute row sums are $2^{-k}+60(2^{-k}-3^{-k})$ and 3^{-k}, the first being the larger, so by Theorem 16.3.3 (b)

$$\|\mathbf{A}^k\|_\infty=61\cdot 2^{-k}-60\cdot 3^{-k}.$$

At $k=1$ this is 10.5, which overestimates $\rho(\mathbf{A})=\tfrac12$ by a factor of 21; the k-th roots run $10.5,\ 2.93,\ 1.75,\ 1.32,\ 1.11,\ 0.98,\ldots$ and settle toward $\tfrac12$, which is what Theorem 16.4.4 guarantees. The convergence is slow, and it has to be: the single number $\|\mathbf{A}\|_\infty$ cannot know that the entry 10 sits above the diagonal rather than on it.

16.4.5 What is new, and what is not

Corollary 16.4.5. Powers Tending to Zero, with a Rate

Let $\mathbf{A}\in M_n(\mathbb{C})$ and let $\|\cdot\|$ be any matrix norm on $M_n(\mathbb{C})$.

(a) $\mathbf{A}^k\to\mathbf{O}$ if and only if $\rho(\mathbf{A})<1$, and this holds if and only if $\|\mathbf{A}^k\|\to 0$.

(b) If $\rho(\mathbf{A})<1$, then for every q with $\rho(\mathbf{A})<q<1$ there is a constant $C\geq 1$, depending

on $\mathbf{A}$, on q, and on the norm, such that $\|\mathbf{A}^k\| \le C\,q^k$ for **every** $k \ge 0$.

Idea. Part (a) is Theorem 10.10.2 translated into norms by Corollary 16.2.4; nothing is proved that was not proved in Chapter 10. Part (b) is the new content, and it is Theorem 16.4.3 doing what it was built for: choose a norm in which $\mathbf{A}$ itself is already smaller than q, take powers there, and pay one fixed constant to return to the norm we were given.

Proof.

(a) The first equivalence is Theorem 10.10.2, since $\rho(\mathbf{A}) < 1$ says exactly that every eigenvalue has modulus less than 1. The second is Corollary 16.2.4: entrywise convergence of matrices and convergence in the norm $\|\cdot\|$ are the same thing.

(b) Let $\varepsilon = q - \rho(\mathbf{A}) > 0$. By Theorem 16.4.3 there is an induced matrix norm $\|\cdot\|_t$ on $M_n(\mathbb{C})$ with $\|\mathbf{A}\|_t \le \rho(\mathbf{A}) + \varepsilon = q$. Submultiplicativity (Theorem 16.3.2 (d)) gives $\|\mathbf{A}^k\|_t \le \|\mathbf{A}\|_t^k \le q^k$ for every $k \ge 0$. By Theorem 16.2.2, applied to the two norms $\|\cdot\|$ and $\|\cdot\|_t$ on the finite-dimensional space $M_n(\mathbb{C})$, there is $C_0 > 0$ with $\|\mathbf{M}\| \le C_0\|\mathbf{M}\|_t$ for every $\mathbf{M}$. Taking $C = \max(C_0, 1)$,

$$\|\mathbf{A}^k\| \le C_0\|\mathbf{A}^k\|_t \le C\,q^k.$$

This proves the corollary.

■

It is worth being exact about the accounting. Part (a) is **not** new. Chapter 10 §10 proved that $\mathbf{A}^m \to \mathbf{O}$ if and only if every eigenvalue has modulus less than 1, and it proved it with the Jordan form and no norms whatsoever; all that has happened here is that the conclusion has been restated in a language where "small" is a number. What is new is part (b): a **rate**. Chapter 10 could say that the powers die; it could not say how fast, because it had nothing to measure a matrix with. Now every decaying matrix comes with a geometric envelope Cq^k for every q above its spectral radius, and that envelope is what later sections integrate, sum, and differentiate.

Two honest limitations of the envelope deserve saying out loud.

Warning. $\rho(\mathbf{A}) < 1$ **promises eventual decay, not immediate decay.** The constant C in Corollary 16.4.5 (b) is generally larger than 1 and grows as q approaches $\rho(\mathbf{A})$. For

$$\mathbf{A} = \begin{pmatrix} 9/10 & 10 \\ 0 & 9/10 \end{pmatrix}, \qquad \mathbf{A}^k = \begin{pmatrix} (9/10)^k & 10k(9/10)^{k-1} \\ 0 & (9/10)^k \end{pmatrix},$$

we have $\rho(\mathbf{A}) = 9/10 < 1$, yet $\|\mathbf{A}^k\|_\infty = (9/10)^{k-1}(9/10 + 10k)$ **rises** from 10.9 at $k = 1$ to about 39.13 at $k = 9$ before beginning its decay, and does not fall below 1 until $k = 63$. A transient hump of this kind is invisible to the spectral radius and is exactly what the constant C is paying for.

Quick check. For which matrices can one take $C = 1$ and $q = \rho(\mathbf{A})$ in the bound of Corollary 16.4.5 (b), with the spectral norm?

Solution. For normal $\mathbf{A}$, among others. If $\mathbf{A}$ is normal then so is $\mathbf{A}^k$ (its adjoint is $(\mathbf{A}^*)^k$, and powers of commuting matrices commute), so $\|\mathbf{A}^k\|_2 = \rho(\mathbf{A}^k) = \rho(\mathbf{A})^k$ by the computation at the start of this section together with Theorem 9.10.1. There is then no transient at all: the norm of the powers is exactly geometric. The hump in the warning is possible only because that matrix is far from normal, with departure from normality 10 in the sense of Definition 12.11.1.

16.4.6 Exercises

16.4.6.1 A. Check your understanding

Exercise. A1

(a) Define the **spectral radius**, and say why the maximum in the definition exists.

(b) State Theorem 16.4.1 and Theorem 16.4.4.

(c) Determine whether the following statement is correct, and justify your answer: ρ is a matrix norm on $M_n(\mathbb{C})$.

(d) Determine whether the following statement is correct, and justify your answer: if $\rho(\mathbf{A}) < 1$ then $\|\mathbf{A}\|_2 < 1$.

(e) Chapter 10 §10 already proved that $\mathbf{A}^m \to \mathbf{O}$ exactly when $\rho(\mathbf{A}) < 1$. What does Corollary 16.4.5 add to that?

Solution.

(a) $\rho(\mathbf{A}) = \max\{|\lambda| : \lambda \in \operatorname{spec}(\mathbf{A})\}$ (Definition 16.4.1). The set $\operatorname{spec}(\mathbf{A})$ is finite and nonempty — it is the root set of $p_{\mathbf{A}}$, nonempty by Theorem 6.5.1 — so the maximum of the finitely many moduli exists.

(b) For every matrix norm on $M_n(\mathbb{C})$, $\rho(\mathbf{A}) \le \|\mathbf{A}\|$; and for every matrix norm, $\|\mathbf{A}^k\|^{1/k} \to \rho(\mathbf{A})$ as $k \to \infty$.

(c) Incorrect. ρ vanishes on nonzero nilpotent matrices, so it is not even a norm, and it is neither subadditive nor submultiplicative: see the warning after Definition 16.4.1.

(d) Incorrect. $\mathbf{A} = \left(\begin{smallmatrix} 0 & 10 \\ 0 & 0 \end{smallmatrix}\right)$ has $\rho(\mathbf{A}) = 0$ and $\|\mathbf{A}\|_2 = 10$. The correct statement is that *some* norm has $\|\mathbf{A}\| < 1$, by Theorem 16.4.3.

(e) Nothing about the criterion, which is Theorem 10.10.2 restated; what it adds is the quantitative envelope $\|\mathbf{A}^k\| \le Cq^k$ for every $q \in (\rho(\mathbf{A}), 1)$, a rate of decay rather than the bare fact of it.

16.4.6.2 B. Practice

Exercise. B1: Four numbers again

Let $\mathbf{A} = \begin{pmatrix} 0 & 2 \\ 1/4 & 0 \end{pmatrix}$. Compute $\rho(\mathbf{A})$, $\|\mathbf{A}\|_1$, $\|\mathbf{A}\|_\infty$, and $\|\mathbf{A}\|_2$, and verify Theorem 16.4.1 in each case.

Solution. $p_{\mathbf{A}}(x) = x^2 - \frac{1}{2}$, so $\operatorname{spec}(\mathbf{A}) = \{\pm 1/\sqrt{2}\}$ and $\rho(\mathbf{A}) = 1/\sqrt{2} \approx 0.707$.

The absolute column sums are $\frac{1}{4}$ and 2, so $\|\mathbf{A}\|_1 = 2$; the absolute row sums are 2 and $\frac{1}{4}$, so $\|\mathbf{A}\|_\infty = 2$. For the spectral norm, $\mathbf{A}^{\top}\mathbf{A} = \operatorname{diag}(\frac{1}{16}, 4)$, whose eigenvalues are $\frac{1}{16}$ and 4; hence $\sigma_1 = 2$ and $\|\mathbf{A}\|_2 = 2$ by Theorem 16.3.3 (c).

So $\rho(\mathbf{A}) = 0.707\cdots \le 2$ in all three cases, and the gap is a factor of nearly 3 even though $\mathbf{A}$ is a very small matrix. The reason is the mismatch between the two off-diagonal entries: $\mathbf{A}$ is far from normal.

Exercise. B2: Gelfand exactly

With $\mathbf{A}$ as in Exercise (B1: Four numbers again), compute $\mathbf{A}^k$ for every k and evaluate $\lim_k \|\mathbf{A}^k\|_\infty^{1/k}$ directly.

Solution. $\mathbf{A}^2 = \frac{1}{2}\mathbf{I}$, so by induction $\mathbf{A}^{2l} = 2^{-l}\mathbf{I}$ and $\mathbf{A}^{2l+1} = 2^{-l}\mathbf{A}$. Hence

$$\|\mathbf{A}^{2l}\|_\infty = 2^{-l}, \qquad \|\mathbf{A}^{2l+1}\|_\infty = 2^{-l}\cdot 2 = 2^{1-l}.$$

For even $k = 2l$, $\|\mathbf{A}^k\|_\infty^{1/k} = (2^{-l})^{1/(2l)} = 2^{-1/2}$, exactly $\rho(\mathbf{A})$ for every l. For odd $k = 2l+1$, $\|\mathbf{A}^k\|_\infty^{1/k} = 2^{(1-l)/(2l+1)}$, and $(1-l)/(2l+1) \to -\frac{1}{2}$, so this too tends to $2^{-1/2}$. Both subsequences converge to $1/\sqrt{2} = \rho(\mathbf{A})$, confirming Theorem 16.4.4.

Exercise. B3: Deciding convergence without eigenvalues

For each matrix, decide whether $\mathbf{A}^m \to \mathbf{O}$, using only Corollary 16.4.2 and Theorem 10.10.2.

(a) $\mathbf{A} = \frac{1}{4}\begin{pmatrix} 1 & 2 & 0 \\ 0 & 1 & 1 \\ 2 & 0 & 1 \end{pmatrix}$.

(b) $\mathbf{A} = \begin{pmatrix} 1/2 & 3 \\ 0 & 2 \end{pmatrix}$.

Solution.

(a) The absolute row sums are $\frac{3}{4}$, $\frac{1}{2}$, and $\frac{3}{4}$, so $\|\mathbf{A}\|_\infty = \frac{3}{4}$ and $\rho(\mathbf{A}) \le \frac{3}{4} < 1$. By Theorem 10.10.2, $\mathbf{A}^m \to \mathbf{O}$.

(b) Here the row sums are $\frac{7}{2}$ and 2 and the column sums are $\frac{1}{2}$ and 5, so the bounds give only $\rho(\mathbf{A}) \le \frac{7}{2}$ and decide nothing. But the bound is an upper bound, and here the matrix is triangular, so the eigenvalues are visible: $\frac{1}{2}$ and 2, giving $\rho(\mathbf{A}) = 2 > 1$. By Theorem 10.10.2 the powers do **not** tend to $\mathbf{O}$. The moral is that Corollary 16.4.2 can

only ever certify convergence, never rule it out.

16.4.6.3 C. Going deeper

Exercise. C1: A norm that sees the decay

Let $\mathbf{A} \in M_n(\mathbb{C})$.

(a) Prove that if $\|\mathbf{A}\| < 1$ for **some** matrix norm, then $\mathbf{A}^k \to \mathbf{O}$.

(b) Give a matrix with $\mathbf{A}^k \to \mathbf{O}$ and $\|\mathbf{A}\|_2 > 1$, so that the converse of (a) fails for a fixed norm.

(c) Prove that the converse of (a) nevertheless holds if the norm is allowed to depend on $\mathbf{A}$.

Solution.

(a) Submultiplicativity gives $\|\mathbf{A}^k\| \le \|\mathbf{A}\|^k$ by induction on k. With $r = \|\mathbf{A}\| < 1$ the right-hand side tends to 0, so $\|\mathbf{A}^k\| \to 0$, and by Corollary 16.2.4 this is the same as $\mathbf{A}^k \to \mathbf{O}$ entrywise.

(b) $\mathbf{A} = \left(\begin{smallmatrix} 0 & 10 \\ 0 & 0 \end{smallmatrix}\right)$ satisfies $\mathbf{A}^2 = \mathbf{O}$, so certainly $\mathbf{A}^k \to \mathbf{O}$; but $\|\mathbf{A}\|_2 = 10 > 1$. The same happens for $\left(\begin{smallmatrix} 1/2 & 10 \\ 0 & 1/2 \end{smallmatrix}\right)$, which is not even nilpotent.

(c) Suppose $\mathbf{A}^k \to \mathbf{O}$. By Theorem 10.10.2, $\rho(\mathbf{A}) < 1$. Apply Theorem 16.4.3 with $\varepsilon = \frac{1}{2}(1 - \rho(\mathbf{A})) > 0$: it produces an induced matrix norm $\|\cdot\|_t$ with

$$\|\mathbf{A}\|_t \le \rho(\mathbf{A}) + \varepsilon = \tfrac{1}{2}(1 + \rho(\mathbf{A})) < 1.$$

So a norm witnessing the decay always exists, though it must be chosen after $\mathbf{A}$.

Exercise. C2: Spectral radius zero

Let $\mathbf{A} \in M_n(\mathbb{C})$.

(a) Prove that $\rho(\mathbf{A}) = 0$ if and only if $\mathbf{A}$ is nilpotent.

(b) Deduce that for a nilpotent $\mathbf{A}$ and any matrix norm, $\|\mathbf{A}^k\|^{1/k} \to 0$, and verify this directly for $\mathbf{A} = \mathbf{J}_n(0)$.

(c) Explain why (a) shows that no inequality of the form $\|\mathbf{A}\| \le c\,\rho(\mathbf{A})$, with c depending only on n, can hold.

Solution.

(a) ($\Rightarrow$) If $\rho(\mathbf{A}) = 0$ then every eigenvalue is 0, so $p_{\mathbf{A}}(x) = x^n$ because $p_{\mathbf{A}}$ is monic of degree n and splits over $\mathbb{C}$ with all roots 0 (Theorem 6.5.1). By the Cayley–Hamilton theorem (Theorem 9.6.1), $\mathbf{A}^n = p_{\mathbf{A}}(\mathbf{A}) = \mathbf{O}$, so $\mathbf{A}$ is nilpotent (Definition 10.2.1).
($\Leftarrow$) If $\mathbf{A}^d = \mathbf{O}$ and $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$ with $\mathbf{x} \neq \mathbf{0}$, then $\mathbf{0} = \mathbf{A}^d\mathbf{x} = \lambda^d\mathbf{x}$, so $\lambda^d = 0$ and $\lambda = 0$. Hence $\operatorname{spec}(\mathbf{A}) = \{0\}$ and $\rho(\mathbf{A}) = 0$.

(b) By Theorem 16.4.4 the limit is $\rho(\mathbf{A}) = 0$. Directly for $\mathbf{A} = \mathbf{J}_n(0)$: $\mathbf{A}^k = \mathbf{O}$ for $k \ge n$,

so $\|\mathbf{A}^k\|^{1/k} = 0$ for all $k \geq n$, and the sequence is eventually constant at 0.

(c) Such an inequality would force $\|\mathbf{A}\| = 0$, hence $\mathbf{A} = \mathbf{O}$, for every nilpotent $\mathbf{A}$. But $\mathbf{J}_2(0) \neq \mathbf{O}$ is nilpotent. So the two quantities cannot be compared in that direction at all; Theorem 16.4.3 is the best available substitute, and it buys its conclusion by changing the norm.

Exercise. C3: When the infimum is attained

Call $\mathbf{A} \in M_n(\mathbb{C})$ **radius-attaining** if $\|\mathbf{A}\| = \rho(\mathbf{A})$ for some matrix norm.

(a) Prove that every normal $\mathbf{A}$ is radius-attaining.

(b) Prove that a nonzero nilpotent $\mathbf{A}$ is not.

(c) Deduce that $\left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$ is not radius-attaining, even though it is invertible.

Hint for (c): consider $\mathbf{A} - \mathbf{I}$.

Solution.

(a) Let $\mathbf{A}$ be normal, so $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{U}^*$ with $\mathbf{U}$ unitary and $\mathbf{D} = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$ (Corollary 12.4.3). Then $\mathbf{A}^*\mathbf{A} = \mathbf{U}\operatorname{diag}(|\lambda_i|^2)\mathbf{U}^*$, so the singular values of $\mathbf{A}$ are the $|\lambda_i|$ in decreasing order and $\|\mathbf{A}\|_2 = \sigma_1 = \max_i |\lambda_i| = \rho(\mathbf{A})$ by Theorem 16.3.3 (c).

(b) If $\mathbf{A} \neq \mathbf{O}$ is nilpotent then $\rho(\mathbf{A}) = 0$ by Exercise (C2: Spectral radius zero) (a), while $\|\mathbf{A}\| > 0$ for every norm, since a norm vanishes only at $\mathbf{O}$ (Definition 16.1.1). So no norm can give $\|\mathbf{A}\| = \rho(\mathbf{A})$.

(c) Put $\mathbf{A} = \left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$, whose only eigenvalue is 1, so $\rho(\mathbf{A}) = 1$. Suppose some matrix norm had $\|\mathbf{A}\| = 1$. Then $\|\mathbf{A}^k\| \leq \|\mathbf{A}\|^k = 1$ for every k, so the sequence $(\mathbf{A}^k)$ is bounded in that norm, hence bounded entrywise by Corollary 16.2.4, whose second sentence covers boundedness. But $\mathbf{A}^k = \left(\begin{smallmatrix} 1 & k \\ 0 & 1 \end{smallmatrix}\right)$, whose $(1,2)$-entry is unbounded. Contradiction. (The hint points at the same phenomenon: $\mathbf{A} - \mathbf{I}$ is a nonzero nilpotent, so $\mathbf{A}$ is a unipotent matrix, and it is the nilpotent part that the spectral radius cannot see.)

16.5 Inverting a Perturbation of the Identity

Chapter 10 §10 summed the series $\mathbf{I} + \mathbf{A} + \mathbf{A}^2 + \dots$ and identified its sum as $(\mathbf{I} - \mathbf{A})^{-1}$, under the exact hypothesis that every eigenvalue of $\mathbf{A}$ has modulus less than 1. What it could not say is how large that inverse is. This section reruns the argument with a norm in hand. The hypothesis becomes cruder — $\|\mathbf{A}\| < 1$ instead of $\rho(\mathbf{A}) < 1$ — and the conclusion becomes quantitative, which is the trade this chapter keeps making. The payoff is immediate: the invertible matrices turn out to form an open set, inversion turns out to be continuous, and a first estimate appears for how far the solution of $\mathbf{A}\mathbf{x} = \mathbf{b}$ moves when $\mathbf{A}$ and $\mathbf{b}$ are disturbed.

Throughout, $\|\cdot\|$ is the operator norm on $M_n(\mathbb{C})$ induced by a norm on $\mathbb{C}^n$ (Definition 16.3.2), so that $\|\mathbf{I}\| = 1$ and $\|\mathbf{X}\mathbf{Y}\| \leq \|\mathbf{X}\|\|\mathbf{Y}\|$ by Theorem 16.3.2 (c) and (d). Both properties are used in every proof below, and the first of them is not free: the closing warning shows what happens without it.

16.5.1 The geometric series, with a number attached

The scalar identity behind everything here is $1+r+r^2+\cdots = 1/(1-r)$ for $0 \le r < 1$. Chapter 10 §10 gave the matrix version of the *identity*. What a norm adds is the matrix version of the *inequality* $|1+r+\ldots| \le 1/(1-r)$: a ceiling on the inverse, expressed in the data.

Theorem 16.5.1. Neumann Series with a Bound

Let $\mathbf{A} \in M_n(\mathbb{C})$ with $\|\mathbf{A}\| < 1$, and write $a = \|\mathbf{A}\|$. Then:

(a) $\mathbf{I}-\mathbf{A}$ is invertible, and $\sum_{k\ge 0} \mathbf{A}^k = (\mathbf{I}-\mathbf{A})^{-1}$;

(b) $\|(\mathbf{I}-\mathbf{A})^{-1}\| \le \dfrac{1}{1-a}$;

(c) for every $K \ge 0$, the partial sum $\mathbf{S}_K = \sum_{k=0}^{K} \mathbf{A}^k$ satisfies

$$\|(\mathbf{I}-\mathbf{A})^{-1} - \mathbf{S}_K\| \le \frac{a^{K+1}}{1-a};$$

(d) $\|(\mathbf{I}-\mathbf{A})^{-1}\| \ge \dfrac{1}{1+a}$.

Idea. Part (a) is not new work: $\|\mathbf{A}\| < 1$ forces $\rho(\mathbf{A}) < 1$ by Theorem 16.4.1, which is exactly the hypothesis of Chapter 10's Theorem 10.10.6. Everything else is the tail-bounding move. Submultiplicativity turns $\|\mathbf{A}^k\|$ into a^k, so each partial sum is bounded by a scalar geometric sum; the bound then has to be carried across the limit, and what carries it is the fact that a norm is continuous. For (c), the tail of the series factors as $\mathbf{A}^{K+1}$ times the whole series again, so (b) bounds it. For (d), read $\mathbf{I} = (\mathbf{I}-\mathbf{A})(\mathbf{I}-\mathbf{A})^{-1}$ through submultiplicativity.

Proof.

(a) By Theorem 16.4.1 (a), $\rho(\mathbf{A}) \le \|\mathbf{A}\| = a < 1$, so every eigenvalue of $\mathbf{A}$ has modulus less than 1. That is the hypothesis of Theorem 10.10.6, which gives both the invertibility of $\mathbf{I}-\mathbf{A}$ and the value of the sum.

(b) Submultiplicativity (Theorem 16.3.2 (d)) gives $\|\mathbf{A}^k\| \le a^k$ for every $k \ge 1$, by induction on k; and $\|\mathbf{A}^0\| = \|\mathbf{I}\| = 1 = a^0$ by part (c) of the same theorem. Hence, by subadditivity applied K times,

$$\|\mathbf{S}_K\| \le \sum_{k=0}^{K} \|\mathbf{A}^k\| \le \sum_{k=0}^{K} a^k = \frac{1-a^{K+1}}{1-a} \le \frac{1}{1-a},$$

the last step because $a^{K+1} \ge 0$ and $1-a > 0$. By (a) the sequence $(\mathbf{S}_K)$ converges entrywise to $(\mathbf{I}-\mathbf{A})^{-1}$, so $\|\mathbf{S}_K - (\mathbf{I}-\mathbf{A})^{-1}\| \to 0$ by Corollary 16.2.4. Then Lemma 16.1.1 gives

$$\big|\, \|\mathbf{S}_K\| - \|(\mathbf{I}-\mathbf{A})^{-1}\| \,\big| \le \|\mathbf{S}_K - (\mathbf{I}-\mathbf{A})^{-1}\| \to 0,$$

so $\|\mathbf{S}_K\| \to \|(\mathbf{I}-\mathbf{A})^{-1}\|$. A nonstrict inequality survives a limit, so a limit of numbers each at most $1/(1-a)$ is at most $1/(1-a)$, which proves (b).

(c) Fix K. For every $M \geq 1$, multiplying out and canceling,

$$\mathbf{S}_{K+M} - \mathbf{S}_K = \sum_{k=K+1}^{K+M} \mathbf{A}^k = \mathbf{A}^{K+1}\,\mathbf{S}_{M-1}.$$

Let $M \to \infty$. The left-hand side tends to $(\mathbf{I}-\mathbf{A})^{-1} - \mathbf{S}_K$ by (a). The right-hand side tends to $\mathbf{A}^{K+1}(\mathbf{I}-\mathbf{A})^{-1}$, because

$$\|\mathbf{A}^{K+1}\mathbf{S}_{M-1} - \mathbf{A}^{K+1}(\mathbf{I}-\mathbf{A})^{-1}\| \leq \|\mathbf{A}^{K+1}\|\,\|\mathbf{S}_{M-1} - (\mathbf{I}-\mathbf{A})^{-1}\|,$$

which tends to 0 by (a) and Corollary 16.2.4 again. A sequence has at most one limit, so

$$(\mathbf{I}-\mathbf{A})^{-1} - \mathbf{S}_K = \mathbf{A}^{K+1}(\mathbf{I}-\mathbf{A})^{-1}.$$

Taking norms and using submultiplicativity together with (b),

$$\|(\mathbf{I}-\mathbf{A})^{-1} - \mathbf{S}_K\| \leq \|\mathbf{A}^{K+1}\|\,\|(\mathbf{I}-\mathbf{A})^{-1}\| \leq \frac{a^{K+1}}{1-a}.$$

(d) $1 = \|\mathbf{I}\| = \|(\mathbf{I}-\mathbf{A})(\mathbf{I}-\mathbf{A})^{-1}\| \leq \|\mathbf{I}-\mathbf{A}\|\,\|(\mathbf{I}-\mathbf{A})^{-1}\|$, and $\|\mathbf{I}-\mathbf{A}\| \leq \|\mathbf{I}\| + \|\mathbf{A}\| = 1+a$ by subadditivity. Since $1+a > 0$ we may divide, which proves (d). This proves the theorem.

■

Parts (b) and (d) together pin the inverse between two numbers computed from a alone:

$$\frac{1}{1+a} \leq \|(\mathbf{I}-\mathbf{A})^{-1}\| \leq \frac{1}{1-a}.$$

As a approaches 1 the upper bound escapes to infinity while the lower one stays near $\frac{1}{2}$, and the gap is real rather than an artifact of the proof: $\mathbf{A} = a\mathbf{I}$ attains the upper bound, and $\mathbf{A} = -a\mathbf{I}$ attains the lower one.

Example. The bound and the truth

Let

$$\mathbf{A} = \begin{pmatrix} 0 & 1/2 \\ 1/3 & 0 \end{pmatrix}.$$

Using the ∞-norm, check the hypothesis, compute $(\mathbf{I}-\mathbf{A})^{-1}$ exactly, and compare it with the bound of Theorem 16.5.1 (b). Then compare $\mathbf{S}_3$ with the true inverse and with the tail bound of (c).

Solution. The hypothesis. The absolute row sums are $\frac{1}{2}$ and $\frac{1}{3}$, so $\|\mathbf{A}\|_\infty = \frac{1}{2} < 1$ by Theorem 16.3.3 (b). So Theorem 16.5.1 applies with $a = \frac{1}{2}$.

The inverse. $\mathbf{I}-\mathbf{A} = \left(\begin{smallmatrix} 1 & -1/2 \\ -1/3 & 1 \end{smallmatrix}\right)$ has determinant $1 - \frac{1}{6} = \frac{5}{6}$, so by Theorem 0.9.8

$$(\mathbf{I}-\mathbf{A})^{-1} = \tfrac{6}{5}\begin{pmatrix} 1 & 1/2 \\ 1/3 & 1 \end{pmatrix} = \begin{pmatrix} 6/5 & 3/5 \\ 2/5 & 6/5 \end{pmatrix},$$

whose largest absolute row sum is $\frac{6}{5} + \frac{3}{5} = \frac{9}{5} = 1.8$. The bound is $1/(1-\frac{1}{2}) = 2$. So the

estimate is correct and loses only ten percent.

The partial sum. Here $\mathbf{A}^2 = \frac{1}{6}\mathbf{I}$, so $\mathbf{A}^3 = \frac{1}{6}\mathbf{A}$ and

$$\mathbf{S}_3 = (1 + \tfrac{1}{6})(\mathbf{I} + \mathbf{A}) = \tfrac{7}{6}\begin{pmatrix} 1 & 1/2 \\ 1/3 & 1 \end{pmatrix}.$$

Hence $(\mathbf{I} - \mathbf{A})^{-1} - \mathbf{S}_3 = (\frac{6}{5} - \frac{7}{6})(\mathbf{I} + \mathbf{A}) = \frac{1}{30}(\mathbf{I} + \mathbf{A})$, with ∞-norm $\frac{1}{30} \cdot \frac{3}{2} = \frac{1}{20} = 0.05$. The bound of (c) with $K = 3$ is $(\frac{1}{2})^4/(1 - \frac{1}{2}) = \frac{1}{8} = 0.125$. Again correct, and this time loose by a factor of 2.5 — the price of replacing four matrices by four copies of one number.

16.5.2 Which theorem is stronger

The two Neumann theorems must be compared honestly, because neither contains the other.

Chapter 10's Theorem 10.10.6 assumes $\rho(\mathbf{A}) < 1$ and concludes that the series converges, with sum $(\mathbf{I} - \mathbf{A})^{-1}$; and its hypothesis is **sharp**, since the statement is an "if and only if". Theorem 16.5.1 assumes $\|\mathbf{A}\| < 1$, which by Theorem 16.4.1 is a **strictly stronger** demand: $\rho(\mathbf{A}) \le \|\mathbf{A}\|$ always, and the inequality is often strict. For

$$\mathbf{A} = \begin{pmatrix} 0 & 2 \\ 0 & 0 \end{pmatrix}$$

we have $\rho(\mathbf{A}) = 0$, so Chapter 10's theorem applies and returns $\sum_k \mathbf{A}^k = \mathbf{I} + \mathbf{A}$; but $\|\mathbf{A}\| = 2$ in each of the 1-, 2-, and ∞-norms, so this section's theorem says nothing at all about it.

In the other direction, when this section's theorem does apply it says strictly more: a ceiling on the inverse, a rate at which the partial sums approach it, and a floor as well. Chapter 10 could not have stated any of the three, having no way to measure a matrix.

There is a third thing to say, and it dissolves most of the apparent conflict. If $\rho(\mathbf{A}) < 1$, then by Theorem 16.4.3 there is a norm $\|\cdot\|_t$ with $\|\mathbf{A}\|_t \le \rho(\mathbf{A}) + \varepsilon < 1$, and Theorem 16.5.1 applies **in that norm**. For the matrix above, taking $\mathbf{D} = \mathrm{diag}(1, 4)$ and $\|\mathbf{x}\|_t = \|\mathbf{D}\mathbf{x}\|_\infty$ gives $\|\mathbf{A}\|_t = \|\mathbf{D}\mathbf{A}\mathbf{D}^{-1}\|_\infty = \frac{1}{2}$, so the bound reads $\|(\mathbf{I} - \mathbf{A})^{-1}\|_t \le 2$; and indeed $\mathbf{D}(\mathbf{I} + \mathbf{A})\mathbf{D}^{-1} = \left(\begin{smallmatrix} 1 & 1/2 \\ 0 & 1 \end{smallmatrix}\right)$ has ∞-norm $\frac{3}{2} \le 2$. So the quantitative theorem can always be brought to bear on a matrix Chapter 10 accepts — at the cost of changing the norm, and the bound it then produces is a bound in that new norm, which may be a strange unit of measurement.

Quick check. Both theorems need a hypothesis. Which of the two hypotheses, $\rho(\mathbf{A}) < 1$ and $\|\mathbf{A}\| < 1$, can be checked by looking at the entries of $\mathbf{A}$ and doing no linear algebra?

Solution. The second, for a suitable norm: $\|\mathbf{A}\|_1$ and $\|\mathbf{A}\|_\infty$ are the largest absolute column and row sums (Theorem 16.3.3), read straight off the entries. Deciding $\rho(\mathbf{A}) < 1$ requires the eigenvalues, that is, the roots of a degree-n polynomial. This is the practical reason the cruder hypothesis is worth having: it is the one a reader can verify at a glance, and Corollary 16.4.2 is the same observation used in the other direction.

Warning. **The hypothesis $\|\mathbf{A}\| < 1$ is sufficient and never necessary.** A matrix can be as far from $\mathbf{I}$ as one likes and still have $\mathbf{I} - \mathbf{A}$ invertible: $\mathbf{A} = 5\mathbf{I}$ gives $(\mathbf{I} - \mathbf{A})^{-1} = -\frac{1}{4}\mathbf{I}$. What fails when $\|\mathbf{A}\| \geq 1$ is not the invertibility but the *argument*, and with it the bound. Read $\|\mathbf{A}\| < 1$ as "$\mathbf{A}$ is a small perturbation", not as a criterion.

16.5.3 The invertible matrices form an open set

The theorem above concerns perturbations of $\mathbf{I}$. Every invertible matrix can be moved to $\mathbf{I}$ by multiplication, so the general statement follows at once, and it is the statement several earlier arguments wanted: a matrix near an invertible matrix is invertible, and its inverse is near.

Corollary 16.5.2. Invertibility Is an Open Condition, and Inversion Is Continuous

Let $\mathbf{B} \in M_n(\mathbb{C})$ be invertible and let $\mathbf{E} \in M_n(\mathbb{C})$ satisfy

$$\|\mathbf{E}\| < \frac{1}{\|\mathbf{B}^{-1}\|}.$$

Write $r = \|\mathbf{B}^{-1}\|\|\mathbf{E}\| < 1$. Then:

(a) $\mathbf{B} + \mathbf{E}$ is invertible, and $\|(\mathbf{B} + \mathbf{E})^{-1}\| \leq \dfrac{\|\mathbf{B}^{-1}\|}{1 - r}$;

(b) $\|(\mathbf{B} + \mathbf{E})^{-1} - \mathbf{B}^{-1}\| \leq \dfrac{\|\mathbf{B}^{-1}\|^2\|\mathbf{E}\|}{1 - r}$.

Consequently $\mathrm{GL}_n(\mathbb{C})$ is an open subset of $M_n(\mathbb{C})$: around every invertible $\mathbf{B}$ the ball of radius $1/\|\mathbf{B}^{-1}\|$ consists of invertible matrices. And inversion is continuous on it: if $\mathbf{B}_m \to \mathbf{B}$ with $\mathbf{B}$ invertible, then $\mathbf{B}_m$ is invertible for all large m and $\mathbf{B}_m^{-1} \to \mathbf{B}^{-1}$.

Idea. Factor the perturbed matrix as $\mathbf{B} + \mathbf{E} = \mathbf{B}(\mathbf{I} + \mathbf{B}^{-1}\mathbf{E})$ — legitimate because $\mathbf{B}$ is invertible — and notice that $\|\mathbf{B}^{-1}\mathbf{E}\| \leq r < 1$, so the second factor is a perturbation of the identity of exactly the kind Theorem 16.5.1 handles, with $\mathbf{A} = -\mathbf{B}^{-1}\mathbf{E}$. For (b), the difference of two inverses always factors as $\mathbf{X}^{-1} - \mathbf{Y}^{-1} = \mathbf{X}^{-1}(\mathbf{Y} - \mathbf{X})\mathbf{Y}^{-1}$, and both inverses are now bounded.

Proof. Put $\mathbf{A} = -\mathbf{B}^{-1}\mathbf{E}$. By submultiplicativity, $\|\mathbf{A}\| \leq \|\mathbf{B}^{-1}\|\|\mathbf{E}\| = r < 1$.

(a) Since $\mathbf{B}$ is invertible, $\mathbf{B} + \mathbf{E} = \mathbf{B}(\mathbf{I} + \mathbf{B}^{-1}\mathbf{E}) = \mathbf{B}(\mathbf{I} - \mathbf{A})$. By Theorem 16.5.1 (a) the factor $\mathbf{I} - \mathbf{A}$ is invertible, so $\mathbf{B} + \mathbf{E}$ is a product of invertible matrices and is invertible, with

$$(\mathbf{B} + \mathbf{E})^{-1} = (\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}^{-1}.$$

By submultiplicativity and Theorem 16.5.1 (b),

$$\|(\mathbf{B} + \mathbf{E})^{-1}\| \leq \|(\mathbf{I} - \mathbf{A})^{-1}\|\,\|\mathbf{B}^{-1}\| \leq \frac{\|\mathbf{B}^{-1}\|}{1 - \|\mathbf{A}\|} \leq \frac{\|\mathbf{B}^{-1}\|}{1 - r},$$

the last step because $\|\mathbf{A}\| \leq r$ and $t \mapsto 1/(1 - t)$ increases on $[0, 1)$.

(b) Multiplying the identity $\mathbf{B} - (\mathbf{B} + \mathbf{E}) = -\mathbf{E}$ on the left by $(\mathbf{B} + \mathbf{E})^{-1}$ and on the right

by $\mathbf{B}^{-1}$ gives

$$(\mathbf{B}+\mathbf{E})^{-1}-\mathbf{B}^{-1}=-(\mathbf{B}+\mathbf{E})^{-1}\mathbf{E}\,\mathbf{B}^{-1}.$$

Taking norms and applying submultiplicativity twice and then (a),

$$\|(\mathbf{B}+\mathbf{E})^{-1}-\mathbf{B}^{-1}\|\le\frac{\|\mathbf{B}^{-1}\|}{1-r}\,\|\mathbf{E}\|\,\|\mathbf{B}^{-1}\|=\frac{\|\mathbf{B}^{-1}\|^2\|\mathbf{E}\|}{1-r}.$$

For the two consequences: openness is the statement of (a) with $\mathbf{E}$ ranging over the ball of radius $1/\|\mathbf{B}^{-1}\|$. For continuity, let $\mathbf{B}_m\to\mathbf{B}$ with $\mathbf{B}$ invertible, and put $\mathbf{E}_m=\mathbf{B}_m-\mathbf{B}$. By Corollary 16.2.4, $\|\mathbf{E}_m\|\to 0$, so $\|\mathbf{E}_m\|<1/(2\|\mathbf{B}^{-1}\|)$ for all large m; for those m, part (a) applies with $r\le\frac{1}{2}$ and $\mathbf{B}_m$ is invertible, and part (b) gives

$$\|\mathbf{B}_m^{-1}-\mathbf{B}^{-1}\|\le 2\|\mathbf{B}^{-1}\|^2\|\mathbf{E}_m\|\to 0.$$

This proves the corollary.

■

Example. How far an invertible matrix can be pushed

Take $\mathbf{B}=\left(\begin{smallmatrix}2&1\\0&2\end{smallmatrix}\right)$ and work with $\|\cdot\|_\infty$. Then $\mathbf{B}^{-1}=\left(\begin{smallmatrix}\frac{1}{2}&-\frac{1}{4}\\0&\frac{1}{2}\end{smallmatrix}\right)$, whose largest absolute row sum is $\frac{1}{2}+\frac{1}{4}=\frac{3}{4}$, so $\|\mathbf{B}^{-1}\|_\infty=\frac{3}{4}$. Corollary 16.5.2 therefore keeps $\mathbf{B}+\mathbf{E}$ invertible for every $\mathbf{E}$ with

$$\|\mathbf{E}\|_\infty<\frac{1}{\|\mathbf{B}^{-1}\|_\infty}=\frac{4}{3}.$$

For a perturbation of size $\|\mathbf{E}\|_\infty\le\frac{1}{2}$ we get $r=\frac{3}{4}\cdot\frac{1}{2}=\frac{3}{8}$, and part (a) bounds

$$\|(\mathbf{B}+\mathbf{E})^{-1}\|_\infty\;\le\;\frac{3/4}{1-3/8}\;=\;\frac{6}{5}.$$

The bound is honest rather than tight: pushing every entry of $\mathbf{E}$ to $\pm\frac{1}{4}$ gives at worst $\|(\mathbf{B}+\mathbf{E})^{-1}\|_\infty=\frac{12}{11}\approx 1.09$, against the promised 1.2. What the corollary buys is not the sharpest constant but a radius inside which no invertibility check is needed at all.

Part (a) has a reading worth recording, since the relative-error theorem of §08 is the same sentence with the constant named: **the distance from an invertible $\mathbf{B}$ to the nearest singular matrix is at least $1/\|\mathbf{B}^{-1}\|$.** A large $\|\mathbf{B}^{-1}\|$ is therefore not an abstract complaint; it says that $\mathbf{B}$ is close, in the literal sense of the norm, to a matrix that cannot be inverted at all.

Remark. Chapter 12 §01 proved that every matrix has diagonalizable matrices arbitrarily close to it (Corollary 12.1.6) and used the word "close" before this chapter existed. Corollary 16.5.2 is the first of the two facts that argument needs; the second is the continuity of the eigenvalues, and the two are put together in §07.

16.5.4 Perturbing a solve

The question the rest of the chapter circles is this. A linear system $\mathbf{A}\mathbf{x} = \mathbf{b}$ is solved; the data $\mathbf{A}$ and $\mathbf{b}$ were only known approximately; how wrong can $\mathbf{x}$ be? With Corollary 16.5.2 available the answer is three lines, and it already contains the quantity §08 will name.

Theorem 16.5.3. Relative Error in a Perturbed Solve

Let $\mathbf{A} \in M_n(\mathbb{C})$ be invertible, let $\mathbf{b} \neq \mathbf{0}$, and let $\mathbf{x}$ be the solution of $\mathbf{A}\mathbf{x} = \mathbf{b}$. Let $\mathbf{E} \in M_n(\mathbb{C})$ and $\mathbf{f} \in \mathbb{C}^n$ be perturbations with

$$r = \|\mathbf{A}^{-1}\|\|\mathbf{E}\| < 1,$$

and let $\tilde{\mathbf{x}}$ be the solution of $(\mathbf{A}+\mathbf{E})\tilde{\mathbf{x}} = \mathbf{b}+\mathbf{f}$, which exists and is unique by Corollary 16.5.2. Then

$$\frac{\|\tilde{\mathbf{x}} - \mathbf{x}\|}{\|\mathbf{x}\|} \leq \frac{\|\mathbf{A}\|\,\|\mathbf{A}^{-1}\|}{1-r}\left(\frac{\|\mathbf{E}\|}{\|\mathbf{A}\|} + \frac{\|\mathbf{f}\|}{\|\mathbf{b}\|}\right).$$

Idea. Subtract the two equations to isolate $\mathbf{A}(\tilde{\mathbf{x}} - \mathbf{x})$; what is left on the other side is $\mathbf{f} - \mathbf{E}\tilde{\mathbf{x}}$. Apply $\mathbf{A}^{-1}$ and take norms. The one awkward term is $\|\tilde{\mathbf{x}}\|$, which is not data; write $\tilde{\mathbf{x}} = \mathbf{x} + (\tilde{\mathbf{x}} - \mathbf{x})$ and move the resulting copy of $\|\tilde{\mathbf{x}} - \mathbf{x}\|$ to the left, which is where the factor $1/(1-r)$ comes from. Finally $\|\mathbf{b}\| \leq \|\mathbf{A}\|\|\mathbf{x}\|$ converts $1/\|\mathbf{x}\|$ into $\|\mathbf{A}\|/\|\mathbf{b}\|$, and the two perturbations appear only through their relative sizes.

Proof. Write $\mathbf{d} = \tilde{\mathbf{x}} - \mathbf{x}$. Subtracting $\mathbf{A}\mathbf{x} = \mathbf{b}$ from $(\mathbf{A}+\mathbf{E})\tilde{\mathbf{x}} = \mathbf{b}+\mathbf{f}$ gives $\mathbf{A}\tilde{\mathbf{x}} + \mathbf{E}\tilde{\mathbf{x}} - \mathbf{A}\mathbf{x} = \mathbf{f}$, that is, $\mathbf{A}\mathbf{d} = \mathbf{f} - \mathbf{E}\tilde{\mathbf{x}}$. Since $\mathbf{A}$ is invertible, $\mathbf{d} = \mathbf{A}^{-1}(\mathbf{f} - \mathbf{E}\tilde{\mathbf{x}})$, and Theorem 16.3.2 (a) with subadditivity gives

$$\|\mathbf{d}\| \leq \|\mathbf{A}^{-1}\|(\|\mathbf{f}\| + \|\mathbf{E}\|\,\|\tilde{\mathbf{x}}\|).$$

Now $\|\tilde{\mathbf{x}}\| \leq \|\mathbf{x}\| + \|\mathbf{d}\|$, so

$$\|\mathbf{d}\| \leq \|\mathbf{A}^{-1}\|\|\mathbf{f}\| + \|\mathbf{A}^{-1}\|\|\mathbf{E}\|\|\mathbf{x}\| + r\|\mathbf{d}\|,$$

using $\|\mathbf{A}^{-1}\|\|\mathbf{E}\| = r$ on the last term. Since $r < 1$ we may subtract $r\|\mathbf{d}\|$ and divide by $1 - r > 0$:

$$\|\mathbf{d}\| \leq \frac{\|\mathbf{A}^{-1}\|}{1-r}(\|\mathbf{f}\| + \|\mathbf{E}\|\|\mathbf{x}\|).$$

Divide by $\|\mathbf{x}\|$, which is nonzero because $\mathbf{b} \neq \mathbf{0}$:

$$\frac{\|\mathbf{d}\|}{\|\mathbf{x}\|} \leq \frac{\|\mathbf{A}^{-1}\|}{1-r}\left(\frac{\|\mathbf{f}\|}{\|\mathbf{x}\|} + \|\mathbf{E}\|\right).$$

Finally $\|\mathbf{b}\| = \|\mathbf{A}\mathbf{x}\| \leq \|\mathbf{A}\|\|\mathbf{x}\|$, so $1/\|\mathbf{x}\| \leq \|\mathbf{A}\|/\|\mathbf{b}\|$. Substituting this in the first term and writing $\|\mathbf{E}\| = \|\mathbf{A}\| \cdot \|\mathbf{E}\|/\|\mathbf{A}\|$ in the second gives the stated inequality. This proves the theorem.

■

The shape of the conclusion is the point. On the right, the two perturbations appear only as **relative** sizes, $\|\mathbf{E}\|/\|\mathbf{A}\|$ and $\|\mathbf{f}\|/\|\mathbf{b}\|$; on the left is the relative size of the error. Between them

stands one number built from $\mathbf{A}$ alone,

$$\|\mathbf{A}\|\,\|\mathbf{A}^{-1}\|,$$

inflated slightly by $1/(1-r)$, which is close to 1 whenever the perturbation of $\mathbf{A}$ is small. That number is the condition number of §08. Everything that section does is to give it a name, compute it for the 2-norm, and show that the bound is attained.

Example. A perturbation of the matrix

Let

$$\mathbf{A}=\begin{pmatrix}2&1\\1&1\end{pmatrix},\quad \mathbf{b}=\begin{pmatrix}3\\2\end{pmatrix},\quad \mathbf{E}=\begin{pmatrix}0&0\\0&1/10\end{pmatrix},\quad \mathbf{f}=\mathbf{0}.$$

Solve both systems in the ∞-norm, compute the relative error, and compare it with Theorem 16.5.3.

Solution. The two solutions. $\det\mathbf{A}=1$, so $\mathbf{A}^{-1}=\left(\begin{smallmatrix}1&-1\\-1&2\end{smallmatrix}\right)$ by Theorem 0.9.8, and $\mathbf{x}=\mathbf{A}^{-1}\mathbf{b}=(3-2,\ -3+4)=(1,1)$. For the perturbed matrix, $\mathbf{A}+\mathbf{E}=\left(\begin{smallmatrix}2&1\\1&11/10\end{smallmatrix}\right)$ has determinant $\frac{22}{10}-1=\frac{6}{5}$, so

$$\widetilde{\mathbf{x}}=\tfrac{5}{6}\begin{pmatrix}11/10&-1\\-1&2\end{pmatrix}\begin{pmatrix}3\\2\end{pmatrix}=\tfrac{5}{6}\begin{pmatrix}13/10\\1\end{pmatrix}=\begin{pmatrix}13/12\\5/6\end{pmatrix}.$$

The error. $\widetilde{\mathbf{x}}-\mathbf{x}=(\frac{1}{12},-\frac{1}{6})$, so $\|\widetilde{\mathbf{x}}-\mathbf{x}\|_\infty=\frac{1}{6}$ and, since $\|\mathbf{x}\|_\infty=1$, the relative error is $\frac{1}{6}\approx 0.167$.

The bound. The absolute row sums give $\|\mathbf{A}\|_\infty=3$ and $\|\mathbf{A}^{-1}\|_\infty=3$, so $\|\mathbf{A}\|_\infty\|\mathbf{A}^{-1}\|_\infty=9$. Also $\|\mathbf{E}\|_\infty=\frac{1}{10}$ and $r=3\cdot\frac{1}{10}=\frac{3}{10}<1$. With $\mathbf{f}=\mathbf{0}$ the bound is

$$\frac{9}{1-\frac{3}{10}}\cdot\frac{1/10}{3}=\frac{9}{7/10}\cdot\frac{1}{30}=\frac{3}{7}\approx 0.429.$$

So $0.167\le 0.429$: correct, and loose by a factor of about 2.6. Note the accounting. The matrix was disturbed by 3.3% of its size and the answer moved by 16.7%, an amplification of about 5; the theorem promised at most $9/(1-0.3)\approx 12.9$.

Warning. **All four bounds of this section need $\|\mathbf{I}\|=1$, which is a property of *induced* norms only.** The Frobenius norm is submultiplicative (Example (The Frobenius norm is submultiplicative)) but has $\|\mathbf{I}_n\|_F=\sqrt{n}$. Take $\mathbf{A}=\mathbf{O}$ in $M_2(\mathbb{C})$: then $\|\mathbf{A}\|_F=0<1$, while

$$\|(\mathbf{I}-\mathbf{A})^{-1}\|_F=\|\mathbf{I}_2\|_F=\sqrt{2}>1=\frac{1}{1-\|\mathbf{A}\|_F},$$

so Theorem 16.5.1 (b) is false for it. The repair is easy — Exercise (C3: Without $\|\mathbf{I}\|=1$) — but it is not the same inequality, and using the Frobenius norm here is the most common way to arrive at a wrong constant.

16.5.5 Exercises

16.5.5.1 A. Check your understanding

Exercise. A1

(a) State Theorem 16.5.1 in full, including which kind of norm it assumes.

(b) Chapter 10 §10 already proved that $\sum_k \mathbf{A}^k$ converges to $(\mathbf{I}-\mathbf{A})^{-1}$ when $\rho(\mathbf{A}) < 1$. Say precisely what this section adds, and what it gives up.

(c) Determine whether the following statement is correct, and justify your answer: if $\mathbf{I}-\mathbf{A}$ is invertible then $\|\mathbf{A}\| < 1$.

(d) Where in the proof of Theorem 16.5.1 is $\|\mathbf{I}\| = 1$ used?

(e) State what Corollary 16.5.2 says about the distance from an invertible matrix to the singular matrices.

Solution.

(a) If $\|\cdot\|$ is the operator norm induced by a vector norm and $\|\mathbf{A}\| = a < 1$, then $\mathbf{I}-\mathbf{A}$ is invertible with $\sum_{k\ge 0}\mathbf{A}^k = (\mathbf{I}-\mathbf{A})^{-1}$; moreover $\|(\mathbf{I}-\mathbf{A})^{-1}\| \le 1/(1-a)$, the partial sums satisfy $\|(\mathbf{I}-\mathbf{A})^{-1} - \mathbf{S}_K\| \le a^{K+1}/(1-a)$, and $\|(\mathbf{I}-\mathbf{A})^{-1}\| \ge 1/(1+a)$.

(b) It adds three numbers: a ceiling on the inverse, a geometric rate for the partial sums, and a floor. It gives up sharpness of the hypothesis: $\rho(\mathbf{A}) < 1$ is necessary and sufficient for convergence, whereas $\|\mathbf{A}\| < 1$ is only sufficient, and strictly stronger by Theorem 16.4.1.

(c) Incorrect. $\mathbf{A} = 5\mathbf{I}$ has $\mathbf{I}-\mathbf{A} = -4\mathbf{I}$, invertible, while $\|\mathbf{A}\| = 5$ in every induced norm. Invertibility of $\mathbf{I}-\mathbf{A}$ says only $1 \notin \operatorname{spec}(\mathbf{A})$.

(d) Twice, and both times to make a geometric sum start at 1: in (b), where $\|\mathbf{A}^0\| = \|\mathbf{I}\| = 1$ is the $k=0$ term of $\sum_k a^k$; and in (d), where $1 = \|\mathbf{I}\|$ is the left-hand side of the estimate.

(e) It is at least $1/\|\mathbf{B}^{-1}\|$: every $\mathbf{E}$ with $\|\mathbf{E}\| < 1/\|\mathbf{B}^{-1}\|$ leaves $\mathbf{B}+\mathbf{E}$ invertible.

16.5.5.2 B. Practice

Exercise. B1: Check, sum, compare

Let $\mathbf{A} = \begin{pmatrix} 1/4 & 1/4 \\ 0 & 1/2 \end{pmatrix}$.

(a) Verify the hypothesis of Theorem 16.5.1 for $\|\cdot\|_\infty$.

(b) Compute $(\mathbf{I}-\mathbf{A})^{-1}$ exactly and compare $\|(\mathbf{I}-\mathbf{A})^{-1}\|_\infty$ with the bound of (b) and with the lower bound of (d).

Solution.

(a) The absolute row sums are $\frac{1}{2}$ and $\frac{1}{2}$, so $\|\mathbf{A}\|_\infty = \frac{1}{2} < 1$ by Theorem 16.3.3 (b).

(b) $\mathbf{I}-\mathbf{A}=\left(\begin{smallmatrix}3/4 & -1/4\\ 0 & 1/2\end{smallmatrix}\right)$, upper triangular with determinant $\frac{3}{8}$, so by Theorem 0.9.8

$$(\mathbf{I}-\mathbf{A})^{-1}=\tfrac{8}{3}\begin{pmatrix}1/2 & 1/4\\ 0 & 3/4\end{pmatrix}=\begin{pmatrix}4/3 & 2/3\\ 0 & 2\end{pmatrix}.$$

The absolute row sums are 2 and 2, so $\|(\mathbf{I}-\mathbf{A})^{-1}\|_\infty = 2$. The upper bound is $1/(1-\frac{1}{2})=2$, attained exactly; the lower bound is $1/(1+\frac{1}{2})=\frac{2}{3}$, comfortably satisfied. The upper bound is attained because both rows of $\mathbf{A}$ have the same absolute sum and no cancellation occurs anywhere in the geometric series.

Exercise. B2: A perturbed right-hand side

Let $\mathbf{A}=\begin{pmatrix}3 & 1\\ 1 & 2\end{pmatrix}$, $\mathbf{b}=(4,3)$, and $\mathbf{f}=(1/10,0)$, with $\mathbf{E}=\mathbf{O}$. Solve both systems, compute the relative error in the ∞-norm, and compare with Theorem 16.5.3.

Solution. $\det\mathbf{A}=5$, so $\mathbf{A}^{-1}=\frac{1}{5}\left(\begin{smallmatrix}2 & -1\\ -1 & 3\end{smallmatrix}\right)$ by Theorem 0.9.8, and $\mathbf{x}=\mathbf{A}^{-1}\mathbf{b}=\frac{1}{5}(8-3,-4+9)=(1,1)$.

With $\mathbf{b}+\mathbf{f}=(\frac{41}{10},3)$ we get $\widetilde{\mathbf{x}}=\frac{1}{5}(\frac{82}{10}-3,-\frac{41}{10}+9)=(\frac{52}{50},\frac{49}{50})$. Hence $\widetilde{\mathbf{x}}-\mathbf{x}=(\frac{2}{50},-\frac{1}{50})$ and $\|\widetilde{\mathbf{x}}-\mathbf{x}\|_\infty=\frac{1}{25}=0.04$, a relative error of 0.04 since $\|\mathbf{x}\|_\infty=1$.

For the bound: $\|\mathbf{A}\|_\infty=4$, $\|\mathbf{A}^{-1}\|_\infty=\frac{4}{5}$, so $\|\mathbf{A}\|_\infty\|\mathbf{A}^{-1}\|_\infty=\frac{16}{5}$. Here $\mathbf{E}=\mathbf{O}$, so $r=0$, and $\|\mathbf{f}\|_\infty/\|\mathbf{b}\|_\infty=\frac{1/10}{4}=\frac{1}{40}$. The bound is $\frac{16}{5}\cdot\frac{1}{40}=\frac{2}{25}=0.08$. So $0.04\le 0.08$, loose by a factor of exactly 2.

Exercise. B3: Staying invertible

Let $\mathbf{B}=\begin{pmatrix}2 & 0\\ 0 & 1/2\end{pmatrix}$. Using $\|\cdot\|_\infty$, find a radius $\delta>0$ such that every $\mathbf{E}$ with $\|\mathbf{E}\|_\infty<\delta$ leaves $\mathbf{B}+\mathbf{E}$ invertible, and bound $\|(\mathbf{B}+\mathbf{E})^{-1}\|_\infty$ for $\|\mathbf{E}\|_\infty\le\frac{1}{4}$. Is your δ the largest possible?

Solution. $\mathbf{B}^{-1}=\operatorname{diag}(\frac{1}{2},2)$, so $\|\mathbf{B}^{-1}\|_\infty=2$ and Corollary 16.5.2 gives $\delta=1/\|\mathbf{B}^{-1}\|_\infty=\frac{1}{2}$.

For $\|\mathbf{E}\|_\infty\le\frac{1}{4}$ we get $r=2\cdot\frac{1}{4}=\frac{1}{2}$, so part (a) of the corollary gives

$$\|(\mathbf{B}+\mathbf{E})^{-1}\|_\infty\le\frac{2}{1-\frac{1}{2}}=4.$$

It is the largest possible here. Taking $\mathbf{E}=\operatorname{diag}(0,-\frac{1}{2})$ gives $\|\mathbf{E}\|_\infty=\frac{1}{2}$ and $\mathbf{B}+\mathbf{E}=\operatorname{diag}(2,0)$, which is singular. So no radius larger than $\frac{1}{2}$ works, and the estimate $1/\|\mathbf{B}^{-1}\|$ is attained by this $\mathbf{B}$. (It is not attained by every $\mathbf{B}$ in every norm; Exercise (C2: Distance to the singular matrices) returns to this.)

16.5.5.3 C. Going deeper

Exercise. C1: The inverse to first order

Let $\|\mathbf{A}\| < 1$.

(a) Prove that $\|(\mathbf{I}-\mathbf{A})^{-1} - \mathbf{I} - \mathbf{A}\| \le \dfrac{\|\mathbf{A}\|^2}{1-\|\mathbf{A}\|}$.

(b) Deduce that for a fixed $\mathbf{M} \in M_n(\mathbb{C})$ and small real t,

$$(\mathbf{I}-t\mathbf{M})^{-1} = \mathbf{I} + t\mathbf{M} + \mathbf{R}(t), \qquad \|\mathbf{R}(t)\|/|t| \to 0$$

as $t \to 0$.

Hint for (a): this is Theorem 16.5.1 (c) with a particular K.

Solution.

(a) Take $K = 1$ in Theorem 16.5.1 (c): $\mathbf{S}_1 = \mathbf{I} + \mathbf{A}$, so

$$\|(\mathbf{I}-\mathbf{A})^{-1} - \mathbf{I} - \mathbf{A}\| \le \frac{\|\mathbf{A}\|^{1+1}}{1-\|\mathbf{A}\|} = \frac{\|\mathbf{A}\|^2}{1-\|\mathbf{A}\|}.$$

(b) If $\mathbf{M} = \mathbf{O}$ the claim is trivial, so assume $\mathbf{M} \neq \mathbf{O}$ and let $|t| < 1/(2\|\mathbf{M}\|)$. Then $\|t\mathbf{M}\| = |t|\|\mathbf{M}\| < \frac{1}{2}$, so (a) applies with $\mathbf{A} = t\mathbf{M}$ and gives, with $\mathbf{R}(t) = (\mathbf{I}-t\mathbf{M})^{-1} - \mathbf{I} - t\mathbf{M}$,

$$\|\mathbf{R}(t)\| \le \frac{t^2\|\mathbf{M}\|^2}{1-|t|\|\mathbf{M}\|} \le 2t^2\|\mathbf{M}\|^2,$$

using $1 - |t|\|\mathbf{M}\| > \frac{1}{2}$. Hence $\|\mathbf{R}(t)\|/|t| \le 2|t|\|\mathbf{M}\|^2 \to 0$ as $t \to 0$. So $t \mapsto (\mathbf{I}-t\mathbf{M})^{-1}$ is differentiable at $t = 0$ with derivative $\mathbf{M}$, which is the case $\mathbf{A}(t) = \mathbf{I} - t\mathbf{M}$ of the general formula for differentiating an inverse, proved in §06.

Exercise. C2: Distance to the singular matrices

Let $\mathbf{B} \in M_n(\mathbb{C})$ be invertible and let $\|\cdot\|$ be an induced matrix norm.

(a) Prove that every singular $\mathbf{S} \in M_n(\mathbb{C})$ satisfies $\|\mathbf{B} - \mathbf{S}\| \ge 1/\|\mathbf{B}^{-1}\|$.

(b) For $\mathbf{B} = \operatorname{diag}(1, \varepsilon)$ with $0 < \varepsilon < 1$ and the 2-norm, exhibit a singular $\mathbf{S}$ attaining the bound.

(c) Hence say what $\|\mathbf{B}^{-1}\|_2$ measures.

Solution.

(a) Suppose $\|\mathbf{B} - \mathbf{S}\| < 1/\|\mathbf{B}^{-1}\|$. Apply Corollary 16.5.2 with $\mathbf{E} = \mathbf{S} - \mathbf{B}$: since $\|\mathbf{E}\| < 1/\|\mathbf{B}^{-1}\|$, the matrix $\mathbf{B} + \mathbf{E} = \mathbf{S}$ is invertible, contradicting the assumption that $\mathbf{S}$ is singular. Hence $\|\mathbf{B} - \mathbf{S}\| \ge 1/\|\mathbf{B}^{-1}\|$ for every singular $\mathbf{S}$.

(b) $\mathbf{B}^{-1} = \operatorname{diag}(1, 1/\varepsilon)$, a normal matrix, so its singular values are the moduli of its eigenvalues and $\|\mathbf{B}^{-1}\|_2 = 1/\varepsilon$ by Theorem 16.3.3 (c). The bound is therefore ε. Take $\mathbf{S} = \operatorname{diag}(1, 0)$, which is singular; then $\mathbf{B} - \mathbf{S} = \operatorname{diag}(0, \varepsilon)$ and $\|\mathbf{B} - \mathbf{S}\|_2 = \varepsilon$. So

the bound is attained.

(c) The reciprocal $1/\|\mathbf{B}^{-1}\|_2$ is a lower bound for the distance from $\mathbf{B}$ to the singular matrices, attained in the example. So $\|\mathbf{B}^{-1}\|_2$ is large exactly when $\mathbf{B}$ is close to being singular, which is the geometric content behind the condition number of §08.

Exercise. C3: Without $\|\mathbf{I}\| = 1$

Let $\|\cdot\|$ be any matrix norm on $M_n(\mathbb{C})$ in the sense of Definition 16.3.1, not necessarily induced, and let $\|\mathbf{A}\| < 1$.

(a) Prove that $\mathbf{I} - \mathbf{A}$ is invertible and that

$$\|(\mathbf{I} - \mathbf{A})^{-1}\| \leq \|\mathbf{I}\| + \frac{\|\mathbf{A}\|}{1 - \|\mathbf{A}\|}.$$

(b) Verify that this bound is correct for the Frobenius norm on $M_2(\mathbb{C})$ with $\mathbf{A} = \mathbf{O}$, where the bound of Theorem 16.5.1 (b) fails.

(c) Explain why the argument of (a) cannot be improved to $1/(1 - \|\mathbf{A}\|)$ for a general matrix norm.

Solution.

(a) By Theorem 16.4.1 (b), which holds for every matrix norm, $\rho(\mathbf{A}) \leq \|\mathbf{A}\| < 1$; so Theorem 10.10.6 applies and gives both the invertibility of $\mathbf{I} - \mathbf{A}$ and $\mathbf{S}_K \to (\mathbf{I} - \mathbf{A})^{-1}$. Submultiplicativity gives $\|\mathbf{A}^k\| \leq \|\mathbf{A}\|^k$ for $k \geq 1$, so separating the term $k = 0$,

$$\|\mathbf{S}_K\| \leq \|\mathbf{I}\| + \sum_{k=1}^{K} \|\mathbf{A}\|^k \leq \|\mathbf{I}\| + \frac{\|\mathbf{A}\|}{1 - \|\mathbf{A}\|}.$$

As in the proof of Theorem 16.5.1 (b), Corollary 16.2.4 and Lemma 16.1.1 let the bound pass to the limit.

(b) With $\mathbf{A} = \mathbf{O}$, the right-hand side is $\|\mathbf{I}_2\|_F + 0 = \sqrt{2}$, and the left-hand side is $\|\mathbf{I}_2\|_F = \sqrt{2}$. The bound holds, with equality; the bound $1/(1 - 0) = 1$ of Theorem 16.5.1 (b) does not.

(c) The case $\mathbf{A} = \mathbf{O}$ already decides it: the left-hand side is $\|\mathbf{I}\|$, so any valid bound must be at least $\|\mathbf{I}\|$, and $1/(1 - \|\mathbf{A}\|) = 1$ there. A norm with $\|\mathbf{I}\| > 1$ therefore falsifies the sharper form at once, and by Theorem 16.3.2 (c) an induced norm never has $\|\mathbf{I}\| > 1$. This is exactly the difference between a matrix norm and an induced one.

16.6 Bounds for the Exponential, and Two Derivatives

Chapter 10 §09 built $e^{\mathbf{A}}$ out of a series and closed with a promise: "Chapter 16 will return to $e^{\mathbf{A}}$ with norms available, which give quantitative bounds instead of exact formulas." That is what this section does. It does not rebuild the exponential; Definition 10.9.2 and Theorem 10.9.3 stand as they are, and the Jordan form is not touched. What is new is that the series now has

a size, so its sum has a ceiling, its truncation has an error bound, and its behavior as $t \to \infty$ has a rate.

The second half takes the other step Chapter 10 could not: it differentiates. The derivative of the determinant and the derivative of the inverse are both needed later in the book, and both require one fact from calculus that linear algebra does not supply. That fact is stated before it is used.

Throughout, $\|\cdot\|$ is the operator norm on $M_n(\mathbb{C})$ induced by a norm on $\mathbb{C}^n$ (Definition 16.3.2), so $\|\mathbf{I}\| = 1$ and $\|\mathbf{XY}\| \le \|\mathbf{X}\|\|\mathbf{Y}\|$ (Theorem 16.3.2 (c), (d)).

16.6.1 How big is the exponential

Chapter 10 §10's criteria answered *whether*; the norm answers *how much*. The bound to aim at is the one the scalar case suggests: $|e^z| \le e^{|z|}$, because the series has nonnegative terms once every term is replaced by its modulus. The matrix proof is the same sentence with submultiplicativity in place of $|zw| = |z||w|$.

One scalar fact is needed and is not on the chapter's list of imports. We name it here rather than let it in quietly.

Remark. **A quoted fact.** For every real $s \ge 0$ the series $\sum_{m\ge 0} s^m/m!$ converges, with sum e^s. This is fact (A1) of Chapter 10 §09, quoted there and quoted again here; this book does not prove it. Because its terms are nonnegative, its partial sums increase to e^s, so every partial sum is at most e^s — that last step is **(A2), monotone convergence**, from the list in this chapter's introduction.

Theorem 16.6.1. Bounds for the Exponential

Let $\mathbf{A} \in M_n(\mathbb{C})$ and write $a = \|\mathbf{A}\|$. Then

(a) $\|e^{\mathbf{A}}\| \le e^a$;

(b) $\|e^{\mathbf{A}} - \mathbf{I}\| \le e^a - 1$;

(c) $\|e^{\mathbf{A}} - \mathbf{I} - \mathbf{A}\| \le e^a - 1 - a \le \dfrac{a^2 e^a}{2}$.

In particular $\|e^{t\mathbf{A}}\| \le e^{|t|a}$ for every real t.

Idea. All three are the same computation stopped at a different place: bound the norm of a partial sum by the corresponding partial sum of $\sum s^m/m!$, then let the partial sums converge. The only thing that needs care is the passage to the limit, and Lemma 16.1.1 does it, exactly as in Theorem 16.5.1. For the second inequality in (c), compare $m!$ with $2\,(m-2)!$.

Proof. Write $\mathbf{S}_M = \sum_{m=0}^{M} \mathbf{A}^m/m!$. By Theorem 10.9.3 the sequence $(\mathbf{S}_M)$ converges entrywise to $e^{\mathbf{A}}$, hence $\|\mathbf{S}_M - e^{\mathbf{A}}\| \to 0$ by Corollary 16.2.4, hence $\|\mathbf{S}_M\| \to \|e^{\mathbf{A}}\|$ by Lemma 16.1.1. The same three steps apply to $\mathbf{S}_M - \mathbf{I}$ and to $\mathbf{S}_M - \mathbf{I} - \mathbf{A}$. So in each part it suffices to bound the partial sums, uniformly in M.

(a) Submultiplicativity gives $\|\mathbf{A}^m\| \le a^m$ for $m \ge 1$, by induction, and $\|\mathbf{A}^0\| = \|\mathbf{I}\| = 1$. By subadditivity and the quoted fact above,

$$\|\mathbf{S}_M\| \le \sum_{m=0}^{M} \frac{\|\mathbf{A}^m\|}{m!} \le \sum_{m=0}^{M} \frac{a^m}{m!} \le e^a.$$

Letting $M \to \infty$ gives $\|e^{\mathbf{A}}\| \le e^a$.

(b) The same estimate with the term $m = 0$ removed:

$$\|\mathbf{S}_M - \mathbf{I}\| \le \sum_{m=1}^{M} \frac{a^m}{m!} \le e^a - 1,$$

since the partial sums of $\sum_{m\ge1} a^m/m!$ increase to $e^a - 1$.

(c) Removing the terms $m = 0$ and $m = 1$ in the same way gives

$$\|e^{\mathbf{A}} - \mathbf{I} - \mathbf{A}\| \le \sum_{m\ge2} \frac{a^m}{m!} = e^a - 1 - a.$$

For the second inequality, note that $m! = m(m-1)\,(m-2)! \ge 2\,(m-2)!$ for every $m \ge 2$, so $a^m/m! \le \frac{1}{2}a^2 \cdot a^{m-2}/(m-2)!$. Summing over $m \ge 2$ and substituting $l = m - 2$,

$$\sum_{m\ge2} \frac{a^m}{m!} \le \frac{a^2}{2} \sum_{l\ge0} \frac{a^l}{l!} = \frac{a^2 e^a}{2}.$$

The final claim is (a) applied to $t\mathbf{A}$, whose norm is $|t|a$ by homogeneity. This proves the theorem.

■

Part (c) is the statement that $e^{\mathbf{A}}$ agrees with $\mathbf{I} + \mathbf{A}$ to second order in $\|\mathbf{A}\|$, with an explicit constant. It is the bound one uses when $\mathbf{A}$ is a small perturbation, and it is the counterpart, for the exponential, of Theorem 16.5.1 (c).

Example. Where the bound is exact, and where it is useless

Compare $\|e^{\mathbf{A}}\|_2$ with $e^{\|\mathbf{A}\|_2}$ for

(a) $\mathbf{A} = \operatorname{diag}(2, 1)$;

(b) $\mathbf{A} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$, and more generally $t\mathbf{A}$ for $t > 0$.

Solution.

(a) $\mathbf{A}$ is Hermitian, hence normal, so its singular values are the moduli of its eigenvalues and $\|\mathbf{A}\|_2 = 2$ by Theorem 16.3.3 (c). By Definition 10.9.1 (or directly from the series, entry by entry) $e^{\mathbf{A}} = \operatorname{diag}(e^2, e)$, again normal, so $\|e^{\mathbf{A}}\|_2 = e^2$. The bound $e^{\|\mathbf{A}\|_2} = e^2$ is attained exactly. The same happens for every $\mathbf{A} = c\mathbf{I}$ with $c > 0$, so no constant smaller than 1 can be inserted in Theorem 16.6.1 (a).

(b) $\mathbf{A}^2 = -\mathbf{I}$, so $\mathbf{A}^3 = -\mathbf{A}$, $\mathbf{A}^4 = \mathbf{I}$, and grouping the series by parity gives, using fact (A1)

of Chapter 10 §09 for the series of cos and sin and its (A3) to rearrange a convergent series into its even and odd parts,

$$e^{t\mathbf{A}} = (\cos t)\mathbf{I} + (\sin t)\mathbf{A} = \begin{pmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{pmatrix},$$

a real orthogonal matrix. Hence $\|e^{t\mathbf{A}}\|_2 = 1$ for every t, while $\|t\mathbf{A}\|_2 = t$ and the bound reads e^t. At $t = 10$ the bound is about 22026 and the truth is 1.

Warning. **The bound sees only the size of A, never the direction of its spectrum.** Part (b) above is the extreme case: $\|e^{t\mathbf{A}}\|$ stays at 1 while $e^{t\|\mathbf{A}\|}$ runs away. The reason is that Theorem 16.6.1 throws away every cancellation in the series at the first step, when $\|\mathbf{A}^m\|$ is replaced by $\|\mathbf{A}\|^m$. Any statement about $e^{t\mathbf{A}}$ for **large** t must come from the eigenvalues instead, which is the subject of the next subsection.

Quick check. Why does Theorem 16.6.1 (a) fail for the Frobenius norm, and what is the correct statement for it?

Solution. Because $\|\mathbf{I}_n\|_F = \sqrt{n}$, so already at $\mathbf{A} = \mathbf{O}$ the claim reads $\sqrt{n} \le 1$. Repeating the proof with the term $m = 0$ kept separate gives the correct version for any matrix norm: $\|e^{\mathbf{A}}\| \le \|\mathbf{I}\| + e^{\|\mathbf{A}\|} - 1$. This is the same correction as in Exercise (C3: Without $\|\mathbf{I}\| = 1$), and it comes from the same place: only an induced norm has $\|\mathbf{I}\| = 1$.

16.6.2 Decay in the stable case

The warning above says the bound $e^{t\|\mathbf{A}\|}$ is the wrong instrument for large t. The right one reads the eigenvalues, and the bridge to the discrete theory of §04 is the observation that the exponential turns the *real part* of an eigenvalue into a *modulus*.

Lemma 16.6.2. The Spectrum of the Exponential

Let $\mathbf{A} \in M_n(\mathbb{C})$. Then $\operatorname{spec}(e^{\mathbf{A}}) = \{e^{\lambda} : \lambda \in \operatorname{spec}(\mathbf{A})\}$, and

$$\rho(e^{\mathbf{A}}) = e^{\mu}, \qquad \mu = \max\{\operatorname{Re}\lambda : \lambda \in \operatorname{spec}(\mathbf{A})\}.$$

Idea. The eigenvalues of $e^{\mathbf{A}}$ are the numbers e^{λ} for $\lambda \in \operatorname{spec}(\mathbf{A})$, which the Jordan form makes visible. Taking moduli turns e^{λ} into $e^{\operatorname{Re}\lambda}$, and the largest of finitely many such numbers is the one with the largest exponent.

Proof. By Theorem 10.3.1 there are an invertible $\mathbf{P}$ and a Jordan matrix $\mathbf{J}$ with $\mathbf{A} = \mathbf{P}\mathbf{J}\mathbf{P}^{-1}$, and the diagonal of $\mathbf{J}$ lists the eigenvalues $\lambda_1, \dots, \lambda_n$ of $\mathbf{A}$ with multiplicity. By Theorem 10.9.4 (a), $e^{\mathbf{J}} = e^{\mathbf{P}^{-1}\mathbf{A}\mathbf{P}} = \mathbf{P}^{-1}e^{\mathbf{A}}\mathbf{P}$. By Theorem 10.9.3 each block $e^{\mathbf{J}_k(\lambda)}$ is

upper triangular with e^{λ} on its diagonal, so $e^{\mathbf{J}}$ is upper triangular with diagonal $e^{\lambda_1}, \dots, e^{\lambda_n}$; hence $\operatorname{spec}(e^{\mathbf{J}}) = \{e^{\lambda_i}\}$ by Theorem 9.7.4 (b). Similar matrices have the same characteristic polynomial (Theorem 7.10.3), so $\operatorname{spec}(e^{\mathbf{A}}) = \operatorname{spec}(e^{\mathbf{J}})$, which is the first claim.

For the second, $|e^{\lambda}| = e^{\operatorname{Re}\lambda}$ by fact (A1) of Chapter 10 §09, and $s \mapsto e^s$ is nondecreasing on $\mathbb{R}$ — for $h \geq 0$, $e^{s+h} = e^s e^h \geq e^s$ by fact (A1) of Chapter 10 §09 — so the largest of the finitely many numbers $e^{\operatorname{Re}\lambda_i}$ is e^{μ}. This proves the lemma.

■

Chapter 12 §10 called **A stable** when every $\lambda \in \operatorname{spec}(\mathbf{A})$ has $\operatorname{Re}\lambda < 0$ (Definition 12.10.2). In the language of the lemma, $\mathbf{A}$ is stable exactly when $\rho(e^{\mathbf{A}}) < 1$ — the continuous-time condition and Chapter 10 §10's discrete-time condition are the same condition, read through the exponential.

Theorem 16.6.3. Exponential Decay of a Stable System

Let $\mathbf{A} \in M_n(\mathbb{C})$, let $\mu = \max\{\operatorname{Re}\lambda : \lambda \in \operatorname{spec}(\mathbf{A})\}$, and let $\nu > \mu$ be real. Then there is a constant $C \geq 1$, depending on $\mathbf{A}$, on ν, and on the norm, with

$$\|e^{t\mathbf{A}}\| \leq C\, e^{\nu t} \qquad \text{for every } t \geq 0.$$

In particular, if $\mathbf{A}$ is **stable** then ν may be taken negative, and every solution of $\mathbf{x}' = \mathbf{A}\mathbf{x}$ satisfies

$$\|\mathbf{x}(t)\| \leq C e^{\nu t}\|\mathbf{x}(0)\| \xrightarrow[t\to\infty]{} 0.$$

Idea. Sample the flow at integer times and use §04. The matrix $\mathbf{M} = e^{-\nu}e^{\mathbf{A}}$ has $\rho(\mathbf{M}) = e^{\mu-\nu} < 1$ by Lemma 16.6.2, so Corollary 16.4.5 bounds all of its powers by one constant; unwinding the powers gives $\|e^{k\mathbf{A}}\| \leq C_0 e^{\nu k}$ at every integer k. Between the integers, $e^{t\mathbf{A}} = e^{(t-k)\mathbf{A}}e^{k\mathbf{A}}$ with $0 \leq t-k < 1$, and the crude bound Theorem 16.6.1 (a) is more than enough to cover a gap of length 1. The whole point is that the crude bound is applied only on a bounded interval, where crudeness costs a constant and not a rate.

Proof. Put $\mathbf{M} = e^{-\nu}e^{\mathbf{A}}$. Multiplying a matrix by a scalar c multiplies every eigenvalue by c, so by Lemma 16.6.2,

$$\rho(\mathbf{M}) = e^{-\nu}\rho(e^{\mathbf{A}}) = e^{-\nu}e^{\mu} = e^{\mu-\nu} < 1,$$

using $\mu - \nu < 0$. Fix any q with $\rho(\mathbf{M}) < q < 1$. By Corollary 16.4.5 (b) there is $C_0 \geq 1$ with $\|\mathbf{M}^k\| \leq C_0 q^k \leq C_0$ for every integer $k \geq 0$.

Since $\mathbf{A}$ commutes with itself, Theorem 10.9.4 (d) gives $e^{(s+u)\mathbf{A}} = e^{s\mathbf{A}}e^{u\mathbf{A}}$ for all real s, u; in particular $(e^{\mathbf{A}})^k = e^{k\mathbf{A}}$ by induction on k. Hence $\mathbf{M}^k = e^{-\nu k}e^{k\mathbf{A}}$, and

$$\|e^{k\mathbf{A}}\| = e^{\nu k}\|\mathbf{M}^k\| \leq C_0\, e^{\nu k} \qquad (k = 0, 1, 2, \dots).$$

Now let $t \geq 0$ and let k be the integer part of t, so that $k \geq 0$ and $0 \leq t - k < 1$. Then $e^{t\mathbf{A}} = e^{(t-k)\mathbf{A}}e^{k\mathbf{A}}$, so by submultiplicativity and Theorem 16.6.1 (a),

$$\|e^{t\mathbf{A}}\| \leq \|e^{(t-k)\mathbf{A}}\|\,\|e^{k\mathbf{A}}\| \leq e^{(t-k)\|\mathbf{A}\|}\, C_0 e^{\nu k} \leq C_0\, e^{\|\mathbf{A}\|}\, e^{\nu k},$$

using $0 \le t-k < 1$ in the last step. Finally $e^{\nu k} = e^{\nu t}e^{-\nu(t-k)} \le e^{\nu t}e^{|\nu|}$, again because $0 \le t-k < 1$. So the constant

$$C = C_0\, e^{\|\mathbf{A}\|+|\nu|} \ge 1$$

works for every $t \ge 0$.

For the last statement, suppose $\mathbf{A}$ is stable, so $\mu < 0$, and choose $\nu \in (\mu, 0)$. By Theorem 10.9.6 the unique solution with the given initial value is $\mathbf{x}(t) = e^{t\mathbf{A}}\mathbf{x}(0)$, so $\|\mathbf{x}(t)\| \le \|e^{t\mathbf{A}}\|\|\mathbf{x}(0)\| \le Ce^{\nu t}\|\mathbf{x}(0)\|$ by Theorem 16.3.2 (a). Since $\nu < 0$, $e^{\nu t} \to 0$ as $t \to \infty$ — this is fact (A5) of Chapter 10 §09 with $j = 0$, quoted there and quoted again here — so $\|\mathbf{x}(t)\| \to 0$, and by Corollary 16.2.4 the vectors $\mathbf{x}(t)$ tend to $\mathbf{0}$. This proves the theorem.

■

Example. Decay after a detour

Let $\mathbf{A} = \begin{pmatrix} -1 & 5 \\ 0 & -1 \end{pmatrix}$. Find μ, compute $\|e^{t\mathbf{A}}\|_\infty$ exactly, and find the smallest constant C that works in Theorem 16.6.3 for $\nu = -\frac{1}{2}$.

Solution. The eigenvalues. $\mathbf{A}$ is upper triangular with both diagonal entries -1, so $\operatorname{spec}(\mathbf{A}) = \{-1\}$ (Theorem 9.7.4 (b)) and $\mu = -1 < 0$: the matrix is stable.

The exponential. Write $\mathbf{A} = -\mathbf{I} + 5\mathbf{N}$ with $\mathbf{N} = \left(\begin{smallmatrix} 0 & 1 \\ 0 & 0 \end{smallmatrix}\right)$. The two summands commute and $\mathbf{N}^2 = \mathbf{O}$, so Corollary 10.9.5 gives

$$e^{t\mathbf{A}} = e^{-t}(\mathbf{I} + 5t\mathbf{N}) = e^{-t}\begin{pmatrix} 1 & 5t \\ 0 & 1 \end{pmatrix}.$$

The absolute row sums are $e^{-t}(1+5t)$ and e^{-t}, so $\|e^{t\mathbf{A}}\|_\infty = e^{-t}(1+5t)$ for $t \ge 0$ by Theorem 16.3.3 (b).

The transient. The function $t \mapsto e^{-t}(1+5t)$ has derivative $e^{-t}(4-5t)$, so it **rises** from 1 at $t = 0$ to $5e^{-4/5} \approx 2.247$ at $t = \frac{4}{5}$, and only then decays. A stable matrix does not decay from the start.

The constant. We need the smallest C with $e^{-t}(1+5t) \le Ce^{-t/2}$ for all $t \ge 0$, that is, $C = \max_{t\ge 0}(1+5t)e^{-t/2}$. The derivative of $(1+5t)e^{-t/2}$ is $e^{-t/2}(5 - \frac{1}{2}(1+5t))$, which is positive for $t < \frac{9}{5}$ and negative for $t > \frac{9}{5}$, so by fact (A6) of the chapter introduction the maximum is at $t = \frac{9}{5}$ and

$$C = 10\, e^{-9/10} \approx 4.066.$$

The rate $e^{-t/2}$ is honest and the constant pays for the hump.

Remark. **What this settles from Chapter 12 §10, and what it does not.** That section introduced the Lyapunov equation $\mathbf{AX}+\mathbf{XA}^* = -\mathbf{Q}$ and the word *stable* (Definition 12.10.2). It then stated, without proof, that a stable $\mathbf{A}$ and a positive definite $\mathbf{Q}$ produce a positive definite solution $\mathbf{X}$, and conversely; it assigned the positivity half to Chapter 13, and assigned to this chapter "the consequence for differential equations — that a stable $\mathbf{A}$ makes every solution of $\dot{\mathbf{x}} = \mathbf{Ax}$ decay, with $\mathbf{x}^*\mathbf{Xx}$ as the quantity that measures the decay".

Theorem 16.6.3 delivers the decay, with a rate, and does so without the Lyapunov equation at all. What it does not deliver is the existence of a positive definite $\mathbf{X}$ for a stable $\mathbf{A}$; that is Theorem 13.6.3, proved in Chapter 13 with no exponential anywhere in the argument, and nothing here depends on it. What can be said cheaply is the other half of the connection: **any** positive definite $\mathbf{X}$ satisfying $\mathbf{A}^*\mathbf{X} + \mathbf{X}\mathbf{A} = -\mathbf{Q}$ with $\mathbf{Q} \succ \mathbf{O}$ does measure the decay, in the sense that $\mathbf{x}^*\mathbf{X}\mathbf{x}$ is strictly decreasing along every nonzero solution of $\mathbf{x}' = \mathbf{A}\mathbf{x}$. That is Exercise (C3: A quantity that measures the decay), and it is three lines once the product rule below is available. Note the transpose: the equation the decay computation wants is Definition 12.10.2 for the pair $(\mathbf{A}^*, \mathbf{Q})$, and $\mathbf{A}$ is stable if and only if $\mathbf{A}^*$ is, since the eigenvalues of $\mathbf{A}^*$ are the conjugates of those of $\mathbf{A}$.

16.6.3 Differentiating a matrix-valued function

A matrix whose entries vary — a family $\mathbf{A}(t)$, a curve in $M_n(F)$ — is differentiated entrywise, exactly as Chapter 10 §09 differentiated $t \mapsto e^{t\mathbf{A}}$. The two derivatives this book needs later, of det and of inversion, are algebraic facts with one analytic ingredient, and the ingredient is named first.

Definition 16.6.1. Derivative of a Matrix-Valued Function

Let $I \subseteq \mathbb{R}$ be an open interval and $\mathbf{A}\colon I \to M_{m\times n}(\mathbb{C})$, written $\mathbf{A}(t) = (a_{ij}(t))$. Call $\mathbf{A}$ **differentiable** at $t_0 \in I$ if every entry function $a_{ij}\colon I \to \mathbb{C}$ is differentiable at t_0, and then set

$$\mathbf{A}'(t_0) \coloneqq (a_{ij}'(t_0)) \in M_{m\times n}(\mathbb{C}).$$

The prime here always means the derivative in t; no dual map (Definition 5.4.1) appears anywhere in this section, so the notation cannot collide. By Corollary 16.2.4 the definition can be restated without coordinates: $\mathbf{A}$ is differentiable at t_0 with derivative $\mathbf{B}$ exactly when

$$\Big\| \frac{\mathbf{A}(t) - \mathbf{A}(t_0)}{t - t_0} - \mathbf{B} \Big\| \longrightarrow 0 \qquad (t \to t_0),$$

in any norm at all. Both readings are used below, the entrywise one to import calculus and the norm one to take limits.

Remark. **The calculus this section quotes.** Let $f, g\colon I \to \mathbb{C}$ be differentiable at t_0.

(D1) *(Linearity and the product rule.)* $f + g$ and fg are differentiable at t_0, with $(f+g)' = f' + g'$ and $(fg)' = f'g + fg'$; a constant function has derivative 0.

(D2) *(Differentiable implies continuous.)* f is continuous at t_0.

Neither is proved here. (D1) is Chapter 10 §09's fact (A3), read for complex-valued functions of a real variable; (D2) is one more standard fact of the same kind, quoted here for the first time in this book. Both are of a piece with Chapter 14 §06's Taylor assumption: standard one-variable calculus, quoted so that the theorems below have hypotheses to stand on. Everything that follows them is linear algebra.

Lemma 16.6.4. Product Rule for Matrices

Let $\mathbf{A}\colon I \to M_{m\times n}(\mathbb{C})$ and $\mathbf{B}\colon I \to M_{n\times p}(\mathbb{C})$ be differentiable at t_0. Then so is $t \mapsto \mathbf{A}(t)\mathbf{B}(t)$,

with

$$(\mathbf{AB})'(t_0) = \mathbf{A}'(t_0)\mathbf{B}(t_0) + \mathbf{A}(t_0)\mathbf{B}'(t_0).$$

Proof. Fix a position (i,k). The entry function is $t \mapsto \sum_{j=1}^{n} a_{ij}(t)b_{jk}(t)$, a finite sum of products of functions differentiable at t_0. By (D1) it is differentiable at t_0 with derivative

$$\sum_{j=1}^{n} \big(a'_{ij}(t_0)b_{jk}(t_0) + a_{ij}(t_0)b'_{jk}(t_0)\big),$$

which is the (i,k)-entry of $\mathbf{A}'(t_0)\mathbf{B}(t_0) + \mathbf{A}(t_0)\mathbf{B}'(t_0)$. Since this holds at every position, Definition 16.6.1 gives the claim. Note that the order of the factors is preserved throughout: matrices do not commute, and the product rule must not be written $2\mathbf{A}\mathbf{A}'$ even for $\mathbf{B} = \mathbf{A}$. This proves the lemma.

■

16.6.4 The derivative of the inverse

Nothing so far says that $t \mapsto \mathbf{A}(t)^{-1}$ is even defined near t_0, let alone differentiable. Both come from §05.

Theorem 16.6.5. Derivative of the Inverse

Let $\mathbf{A}\colon I \to M_n(\mathbb{C})$ be differentiable at $t_0 \in I$, with $\mathbf{A}(t_0)$ invertible. Then $\mathbf{A}(t)$ is invertible for all t in some open interval around t_0, the function $t \mapsto \mathbf{A}(t)^{-1}$ is differentiable at t_0, and

$$\big(\mathbf{A}^{-1}\big)'(t_0) = -\,\mathbf{A}(t_0)^{-1}\,\mathbf{A}'(t_0)\,\mathbf{A}(t_0)^{-1}.$$

Idea. The formula is what one gets by differentiating $\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$ with Lemma 16.6.4 and solving — but that argument presupposes what has to be proved, namely that $\mathbf{A}^{-1}$ is differentiable. So run it in the honest order. The difference of two inverses factors as $\mathbf{X}^{-1} - \mathbf{Y}^{-1} = \mathbf{X}^{-1}(\mathbf{Y} - \mathbf{X})\mathbf{Y}^{-1}$; divide by $t - t_0$ and the middle factor becomes a difference quotient, which converges by hypothesis, while the outer factor converges because inversion is continuous (Corollary 16.5.2). The limit of the product is the product of the limits, and that is the formula.

Proof. Write $\mathbf{A}_0 = \mathbf{A}(t_0)$. By (D2) every entry of $\mathbf{A}$ is continuous at t_0, so $\mathbf{A}(t) \to \mathbf{A}_0$ entrywise as $t \to t_0$, hence $\|\mathbf{A}(t) - \mathbf{A}_0\| \to 0$ by Corollary 16.2.4. By Corollary 16.5.2 there is therefore an open interval $J \ni t_0$ on which $\|\mathbf{A}(t) - \mathbf{A}_0\| < 1/\|\mathbf{A}_0^{-1}\|$, so that $\mathbf{A}(t)$ is invertible for $t \in J$, and on which $\|\mathbf{A}(t)^{-1} - \mathbf{A}_0^{-1}\| \to 0$ as $t \to t_0$.

Let $t \in J$ with $t \neq t_0$. Multiplying $\mathbf{A}_0 - \mathbf{A}(t)$ on the left by $\mathbf{A}(t)^{-1}$ and on the right by $\mathbf{A}_0^{-1}$,

$$\mathbf{A}(t)^{-1} - \mathbf{A}_0^{-1} = -\,\mathbf{A}(t)^{-1}\big(\mathbf{A}(t) - \mathbf{A}_0\big)\mathbf{A}_0^{-1}.$$

Divide by $t - t_0$:

$$\frac{\mathbf{A}(t)^{-1} - \mathbf{A}_0^{-1}}{t - t_0} = -\,\mathbf{A}(t)^{-1}\,\frac{\mathbf{A}(t) - \mathbf{A}_0}{t - t_0}\,\mathbf{A}_0^{-1}.$$

As $t \to t_0$ the middle factor tends to $\mathbf{A}'(t_0)$, by Definition 16.6.1. The left factor tends to $\mathbf{A}_0^{-1}$, by the first paragraph. Products of convergent matrix families converge to the product of the limits: if $\|\mathbf{X}(t) - \mathbf{X}\| \to 0$ and $\|\mathbf{Y}(t) - \mathbf{Y}\| \to 0$ then

$$\|\mathbf{X}(t)\mathbf{Y}(t) - \mathbf{X}\mathbf{Y}\| \le \|\mathbf{X}(t) - \mathbf{X}\|\|\mathbf{Y}(t)\| + \|\mathbf{X}\|\|\mathbf{Y}(t) - \mathbf{Y}\|,$$

and $\|\mathbf{Y}(t)\|$ is bounded near t_0 because $\|\mathbf{Y}(t)\| \le \|\mathbf{Y}\| + \|\mathbf{Y}(t) - \mathbf{Y}\|$. Applying this twice, the right-hand side converges to $-\mathbf{A}_0^{-1}\mathbf{A}'(t_0)\mathbf{A}_0^{-1}$. Hence the difference quotient on the left converges, which by Definition 16.6.1 is exactly the differentiability of $t \mapsto \mathbf{A}(t)^{-1}$ at t_0, with the stated derivative. This proves the theorem.

■

Warning. **The two copies of $\mathbf{A}^{-1}$ sit on either side and cannot be combined.** The scalar formula $(1/a)' = -a'/a^2$ tempts one to write $-\mathbf{A}^{-2}\mathbf{A}'$, and that is wrong unless $\mathbf{A}'(t_0)$ commutes with $\mathbf{A}(t_0)$. For $\mathbf{A}(t) = \left(\begin{smallmatrix} 1 & t \\ 0 & 2 \end{smallmatrix}\right)$ at $t = 0$: we have $\mathbf{A}(0)^{-1} = \operatorname{diag}(1, \tfrac{1}{2})$ and $\mathbf{A}'(0) = \left(\begin{smallmatrix} 0 & 1 \\ 0 & 0 \end{smallmatrix}\right)$, so the correct answer is $-\operatorname{diag}(1, \tfrac{1}{2})\left(\begin{smallmatrix} 0 & 1 \\ 0 & 0 \end{smallmatrix}\right)\operatorname{diag}(1, \tfrac{1}{2}) = \left(\begin{smallmatrix} 0 & -1/2 \\ 0 & 0 \end{smallmatrix}\right)$, whereas $-\mathbf{A}(0)^{-2}\mathbf{A}'(0) = -\left(\begin{smallmatrix} 1 & 0 \\ 0 & 1/4 \end{smallmatrix}\right)\left(\begin{smallmatrix} 0 & 1 \\ 0 & 0 \end{smallmatrix}\right) = \left(\begin{smallmatrix} 0 & -1 \\ 0 & 0 \end{smallmatrix}\right)$.

16.6.5 Jacobi's formula

The determinant is a polynomial in the entries, so differentiating it needs no analysis beyond (D1); the work is in recognizing the answer.

Theorem 16.6.6. Jacobi's Formula

Let $\mathbf{A}\colon I \to M_n(\mathbb{C})$ be differentiable at $t_0 \in I$. Then $t \mapsto \det \mathbf{A}(t)$ is differentiable at t_0, with

$$\frac{d}{dt}\Big|_{t_0} \det \mathbf{A}(t) = \operatorname{tr}\big(\operatorname{adj} \mathbf{A}(t_0)\,\mathbf{A}'(t_0)\big).$$

If moreover $\mathbf{A}(t_0)$ is invertible, this equals $\det \mathbf{A}(t_0)\,\operatorname{tr}\big(\mathbf{A}(t_0)^{-1}\mathbf{A}'(t_0)\big)$.

Idea. Use the Leibniz formula, which writes det as a sum of $n!$ products of n entries, one from each column. The product rule differentiates each such product into n terms, the j-th of which has column j differentiated and the rest untouched. Collecting by j gives n determinants, the j-th being $\mathbf{A}$ with its j-th column replaced by $\mathbf{a}_j'$. Expanding that determinant along column j (Theorem 7.6.1) produces $\sum_i a_{ij}' C_{ij}$, and summing over j gives a double sum which is precisely a trace against the adjugate.

Proof. For $n = 1$ the statement reads $\frac{d}{dt}\big|_{t_0} a_{11}(t) = a_{11}'(t_0)$, since $\operatorname{adj} \mathbf{A} = (1)$ by Definition 7.6.2, so assume $n \ge 2$; that is what lets us speak of cofactors. Write $\mathbf{A}(t) = (a_{ij}(t))$

with columns $\mathbf{a}_1(t), \dots, \mathbf{a}_n(t)$, and write C_{ij} for the cofactors of $\mathbf{A}(t_0)$ (Definition 7.6.1). By Theorem 7.4.3,

$$\det \mathbf{A}(t) = \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma)\, a_{\sigma(1)1}(t) \cdots a_{\sigma(n)n}(t),$$

a finite sum of products of n functions each differentiable at t_0. By (D1), applied $n-1$ times to each product and then to the sum, $\det \mathbf{A}$ is differentiable at t_0 with

$$\frac{d}{dt}\Big|_{t_0} \det \mathbf{A} = \sum_{j=1}^{n} \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma)\, a'_{\sigma(j)j}(t_0) \prod_{l \neq j} a_{\sigma(l)l}(t_0),$$

where the outer sum collects, for each j, the terms in which the j-th factor was the one differentiated. By Theorem 7.4.3 read backwards, the inner sum is the determinant of the matrix $\mathbf{A}_j$ obtained from $\mathbf{A}(t_0)$ by replacing its j-th column with $\mathbf{a}'_j(t_0)$. Expanding $\det \mathbf{A}_j$ along that column with Theorem 7.6.1 (a), and using that the cofactors of column j do not involve column j and are therefore the cofactors C_{ij} of $\mathbf{A}(t_0)$ itself,

$$\det \mathbf{A}_j = \sum_{i=1}^{n} a'_{ij}(t_0)\, C_{ij}.$$

Summing over j,

$$\frac{d}{dt}\Big|_{t_0} \det \mathbf{A} = \sum_{j=1}^{n} \sum_{i=1}^{n} C_{ij}\, a'_{ij}(t_0).$$

It remains to recognize the right-hand side. By Definition 7.6.2, $(\operatorname{adj} \mathbf{A}(t_0))_{ji} = C_{ij}$, so

$$\operatorname{tr}\big(\operatorname{adj} \mathbf{A}(t_0)\, \mathbf{A}'(t_0)\big) = \sum_{j=1}^{n} \sum_{i=1}^{n} (\operatorname{adj} \mathbf{A}(t_0))_{ji}\, a'_{ij}(t_0),$$

which is the same double sum. For the last claim, if $\mathbf{A}(t_0)$ is invertible then $\operatorname{adj} \mathbf{A}(t_0) = \det \mathbf{A}(t_0)\, \mathbf{A}(t_0)^{-1}$ by Theorem 7.6.2, and the trace is linear (Theorem 0.9.6), so the scalar $\det \mathbf{A}(t_0)$ comes out in front. This proves the theorem.

■

Corollary 16.6.7. The Determinant's Derivative at the Identity Is the Trace

For every $\mathbf{B} \in M_n(\mathbb{C})$,

$$\frac{d}{dt}\Big|_{t=0} \det(\mathbf{I} + t\mathbf{B}) = \operatorname{tr} \mathbf{B}.$$

Proof. Apply Theorem 16.6.6 to $\mathbf{A}(t) = \mathbf{I} + t\mathbf{B}$, which is differentiable with $\mathbf{A}'(t) = \mathbf{B}$ and $\mathbf{A}(0) = \mathbf{I}$. Since $\mathbf{I}$ is invertible with $\mathbf{I}^{-1} = \mathbf{I}$ and $\det \mathbf{I} = 1$, the second form of the formula gives $\operatorname{tr}(\mathbf{IB}) = \operatorname{tr} \mathbf{B}$.

■

So the determinant, which is a product of n numbers, has the trace, a sum of n numbers, as its linearization at $\mathbf{I}$. That is the honest source of the identity $\det e^{\mathbf{A}} = e^{\operatorname{tr} \mathbf{A}}$ of Theorem 10.9.4 (e), and Exercise (C1: The determinant of the exponential, again) reproves that identity along these lines.

Example. Both derivatives on one family

Let $\mathbf{A}(t) = \begin{pmatrix} 1+t & t^2 \\ t & 1 \end{pmatrix}$.

(a) Compute $\frac{d}{dt}\det \mathbf{A}(t)$ directly and by Theorem 16.6.6.

(b) Compute $(\mathbf{A}^{-1})'(0)$ by Theorem 16.6.5, and check it against a direct expansion.

Solution.

(a) *Directly.* $\det \mathbf{A}(t) = (1+t)\cdot 1 - t^2 \cdot t = 1 + t - t^3$, so the derivative is $1 - 3t^2$.
By the formula. $\operatorname{adj} \mathbf{A}(t) = \left(\begin{smallmatrix} 1 & -t^2 \\ -t & 1+t \end{smallmatrix}\right)$ (Definition 7.6.2) and $\mathbf{A}'(t) = \left(\begin{smallmatrix} 1 & 2t \\ 1 & 0 \end{smallmatrix}\right)$, so

$$\operatorname{adj} \mathbf{A}(t)\, \mathbf{A}'(t) = \begin{pmatrix} 1 - t^2 & 2t \\ 1 & -2t^2 \end{pmatrix},$$

whose trace is $1 - t^2 - 2t^2 = 1 - 3t^2$. The two agree.

(b) $\mathbf{A}(0) = \mathbf{I}$ and $\mathbf{A}'(0) = \left(\begin{smallmatrix} 1 & 0 \\ 1 & 0 \end{smallmatrix}\right)$, so Theorem 16.6.5 gives

$$(\mathbf{A}^{-1})'(0) = -\mathbf{I}\begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}\mathbf{I} = \begin{pmatrix} -1 & 0 \\ -1 & 0 \end{pmatrix}.$$

Check. $\mathbf{A}(t)^{-1} = \dfrac{1}{1+t-t^3}\left(\begin{smallmatrix} 1 & -t^2 \\ -t & 1+t \end{smallmatrix}\right)$. Near $t = 0$ the scalar factor is $1 - t + (\text{terms in } t^2)$, so

$$\mathbf{A}(t)^{-1} = \begin{pmatrix} 1 - t & 0 \\ -t & 1 \end{pmatrix} + (\text{terms in } t^2),$$

whose derivative at 0 is $\left(\begin{smallmatrix} -1 & 0 \\ -1 & 0 \end{smallmatrix}\right)$, as predicted. The formula did the work that the quotient rule would otherwise have had to do four times.

Quick check. Theorem 16.6.6 holds with no invertibility hypothesis, yet the second form of the formula has one. Where exactly is invertibility used, and what does the formula say at a t_0 where $\det \mathbf{A}(t_0) = 0$?

Solution. Only in the last line, where $\operatorname{adj} \mathbf{A}(t_0)$ is rewritten as $\det \mathbf{A}(t_0)\mathbf{A}(t_0)^{-1}$. The adjugate itself is defined for every square matrix (Definition 7.6.2), so the first form survives. At a t_0 with $\det \mathbf{A}(t_0) = 0$ it still gives a number: for $\mathbf{A}(t) = \operatorname{diag}(t, t)$ at $t_0 = 0$ we get $\operatorname{adj} \mathbf{A}(0) = \mathbf{O}$, so the derivative of $\det \mathbf{A}(t) = t^2$ is 0 there, which is correct. The second form would read "0 times something undefined".

16.6.6 Exercises

16.6.6.1 A. Check your understanding

Exercise. A1

(a) State the three bounds of Theorem 16.6.1, and say which property of the norm each one needs.

(b) Determine whether the following statement is correct, and justify your answer: if $\|\mathbf{A}\|_2$ is large then $\|e^{\mathbf{A}}\|_2$ is large.

(c) State Theorem 16.6.3, and say what the constant C is paying for.

(d) State Jacobi's formula (Theorem 16.6.6), and name the fact from calculus its proof quotes.

(e) Write down the derivative of $t \mapsto \mathbf{A}(t)^{-1}$ and explain in one sentence why it is not $-\mathbf{A}^{-2}\mathbf{A}'$.

Solution.

(a) $\|e^{\mathbf{A}}\| \le e^{\|\mathbf{A}\|}$, $\|e^{\mathbf{A}} - \mathbf{I}\| \le e^{\|\mathbf{A}\|} - 1$, and $\|e^{\mathbf{A}} - \mathbf{I} - \mathbf{A}\| \le e^{\|\mathbf{A}\|} - 1 - \|\mathbf{A}\| \le \|\mathbf{A}\|^2 e^{\|\mathbf{A}\|}/2$. All three need subadditivity (to bound a sum term by term) and submultiplicativity (to get $\|\mathbf{A}^m\| \le \|\mathbf{A}\|^m$); the first also needs $\|\mathbf{I}\| = 1$, which is what makes the $m = 0$ term equal 1.

(b) Incorrect. For $\mathbf{A} = t\left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)$ we have $\|\mathbf{A}\|_2 = t$, arbitrarily large, while $e^{\mathbf{A}}$ is a rotation matrix and $\|e^{\mathbf{A}}\|_2 = 1$ (Example (Where the bound is exact, and where it is useless)).

(c) If μ is the largest real part of an eigenvalue of $\mathbf{A}$ and $\nu > \mu$, there is $C \ge 1$ with $\|e^{t\mathbf{A}}\| \le Ce^{\nu t}$ for all $t \ge 0$. The constant pays for the transient: $\|e^{t\mathbf{A}}\|$ may first grow, as in Example (Decay after a detour), and it also pays for the gap between ν and μ.

(d) $\frac{d}{dt} \det \mathbf{A}(t) = \operatorname{tr}(\operatorname{adj} \mathbf{A}(t)\, \mathbf{A}'(t))$, for a matrix-valued $\mathbf{A}$ differentiable at t. The quoted fact is (D1): the linearity of differentiation and the product rule for complex-valued functions of a real variable.

(e) $(\mathbf{A}^{-1})' = -\mathbf{A}^{-1}\mathbf{A}'\mathbf{A}^{-1}$. The two copies of $\mathbf{A}^{-1}$ sit on opposite sides of $\mathbf{A}'$, and matrices do not commute, so they may not be collected into $\mathbf{A}^{-2}$ unless $\mathbf{A}'$ commutes with $\mathbf{A}$.

16.6.6.2 B. Practice

Exercise. B1: Bound against truth

For each matrix, compute $\|e^{\mathbf{A}}\|_\infty$ exactly and compare with $e^{\|\mathbf{A}\|_\infty}$.

(a) $\mathbf{A} = \begin{pmatrix} 0 & 3 \\ 0 & 0 \end{pmatrix}$.

(b) $\mathbf{A} = \operatorname{diag}(1, -4)$.

Solution.

(a) $\mathbf{A}^2 = \mathbf{O}$, so the series stops: $e^{\mathbf{A}} = \mathbf{I} + \mathbf{A} = \left(\begin{smallmatrix} 1 & 3 \\ 0 & 1 \end{smallmatrix}\right)$, with $\|e^{\mathbf{A}}\|_\infty = 4$. Also $\|\mathbf{A}\|_\infty = 3$, so the bound is $e^3 \approx 20.09$. Correct, and loose by a factor of about 5.

(b) $e^{\mathbf{A}} = \operatorname{diag}(e, e^{-4})$, so $\|e^{\mathbf{A}}\|_\infty = e \approx 2.718$. Here $\|\mathbf{A}\|_\infty = 4$ and the bound is $e^4 \approx 54.6$. The bound is governed by the entry of largest modulus, -4, while the exponential is

governed by the largest **real part**, 1. This is the discrepancy Theorem 16.6.3 exists to remove.

Exercise. B2: A rate for a stable matrix

Let $\mathbf{A} = \begin{pmatrix} -2 & 4 \\ 0 & -2 \end{pmatrix}$.

(a) Compute μ and $e^{t\mathbf{A}}$, and give $\|e^{t\mathbf{A}}\|_\infty$ for $t \geq 0$.

(b) Find the largest value of $\|e^{t\mathbf{A}}\|_\infty$ on $t \geq 0$.

(c) Find the smallest C that works in Theorem 16.6.3 for $\nu = -1$.

Solution.

(a) $\mathbf{A}$ is upper triangular with diagonal $-2, -2$, so $\operatorname{spec}(\mathbf{A}) = \{-2\}$ (Theorem 9.7.4 (b)) and $\mu = -2$. Writing $\mathbf{A} = -2\mathbf{I} + 4\mathbf{N}$ with $\mathbf{N} = \left(\begin{smallmatrix} 0 & 1 \\ 0 & 0 \end{smallmatrix}\right)$, the summands commute and $\mathbf{N}^2 = \mathbf{O}$, so by Corollary 10.9.5

$$e^{t\mathbf{A}} = e^{-2t} \begin{pmatrix} 1 & 4t \\ 0 & 1 \end{pmatrix}, \qquad \|e^{t\mathbf{A}}\|_\infty = e^{-2t}(1 + 4t).$$

(b) The derivative of $e^{-2t}(1 + 4t)$ is $e^{-2t}(4 - 2 - 8t) = e^{-2t}(2 - 8t)$, which is positive for $t < \frac{1}{4}$ and negative after, so by fact (A6) the maximum is at $t = \frac{1}{4}$. The value there is $2e^{-1/2} \approx 1.213$.

(c) We need the smallest C with $e^{-2t}(1 + 4t) \leq Ce^{-t}$, that is, $C = \max_{t \geq 0}(1 + 4t)e^{-t}$. Its derivative is $e^{-t}(4 - 1 - 4t) = e^{-t}(3 - 4t)$, positive for $t < \frac{3}{4}$ and negative after, so by fact (A6) the maximum is at $t = \frac{3}{4}$ and $C = 4e^{-3/4} \approx 1.890$.

Exercise. B3: Two derivatives

Let $\mathbf{A}(t) = \begin{pmatrix} \cos t & -\sin t \\ \sin t & \cos t \end{pmatrix}$.

(a) Compute $\frac{d}{dt} \det \mathbf{A}(t)$ by Theorem 16.6.6 and check it against the exact value of $\det \mathbf{A}(t)$.

(b) Compute $(\mathbf{A}^{-1})'(t)$ by Theorem 16.6.5 and check it by differentiating $\mathbf{A}(t)^{-1}$ directly.

Solution.

(a) $\det \mathbf{A}(t) = \cos^2 t + \sin^2 t = 1$, so the derivative is 0. By the formula: $\operatorname{adj} \mathbf{A}(t) = \left(\begin{smallmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{smallmatrix}\right)$ and $\mathbf{A}'(t) = \left(\begin{smallmatrix} -\sin t & -\cos t \\ \cos t & -\sin t \end{smallmatrix}\right)$, so

$$\operatorname{adj} \mathbf{A}(t)\mathbf{A}'(t) = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix},$$

whose trace is 0. The two agree. (The diagonal entries are $-\cos t \sin t + \sin t \cos t$ and $\sin t \cos t - \cos t \sin t$, both zero.)

(b) Since $\det \mathbf{A}(t) = 1$, $\mathbf{A}(t)^{-1} = \operatorname{adj}\mathbf{A}(t) = \left(\begin{smallmatrix}\cos t & \sin t \\ -\sin t & \cos t\end{smallmatrix}\right)$, whose entrywise derivative is $\left(\begin{smallmatrix}-\sin t & \cos t \\ -\cos t & -\sin t\end{smallmatrix}\right)$. By the formula, using the product $\operatorname{adj}\mathbf{A}(t)\mathbf{A}'(t)$ computed in (a),

$$-\mathbf{A}^{-1}\mathbf{A}'\mathbf{A}^{-1} = -\begin{pmatrix}0 & -1 \\ 1 & 0\end{pmatrix}\begin{pmatrix}\cos t & \sin t \\ -\sin t & \cos t\end{pmatrix} = \begin{pmatrix}-\sin t & \cos t \\ -\cos t & -\sin t\end{pmatrix}.$$

The two agree.

16.6.6.3 C. Going deeper

Exercise. C1: The determinant of the exponential, again

Let $\mathbf{A} \in M_n(\mathbb{C})$ and put $\varphi(t) = \det e^{t\mathbf{A}}$.

(a) Prove that φ is differentiable with $\varphi'(t) = (\operatorname{tr}\mathbf{A})\varphi(t)$.

(b) Hence deduce $\det e^{\mathbf{A}} = e^{\operatorname{tr}\mathbf{A}}$, recovering Theorem 10.9.4 (e).

Hint for (b): consider $\psi(t) = e^{-t\operatorname{tr}\mathbf{A}}\varphi(t)$.

Solution.

(a) The function $t \mapsto e^{t\mathbf{A}}$ is differentiable with derivative $\mathbf{A}e^{t\mathbf{A}}$ (Theorem 10.9.4 (b)), and $e^{t\mathbf{A}}$ is invertible with inverse $e^{-t\mathbf{A}}$ (Theorem 10.9.4 (c)). So Theorem 16.6.6 applies in its second form:

$$\varphi'(t) = \det(e^{t\mathbf{A}})\,\operatorname{tr}(e^{-t\mathbf{A}}\mathbf{A}e^{t\mathbf{A}}).$$

By Theorem 0.9.6 (3) with the two factors $e^{-t\mathbf{A}}$ and $\mathbf{A}e^{t\mathbf{A}}$, we get $\operatorname{tr}(e^{-t\mathbf{A}}\mathbf{A}e^{t\mathbf{A}}) = \operatorname{tr}(\mathbf{A}e^{t\mathbf{A}}e^{-t\mathbf{A}}) = \operatorname{tr}\mathbf{A}$, giving $\varphi'(t) = (\operatorname{tr}\mathbf{A})\varphi(t)$.

(b) Put $c = \operatorname{tr}\mathbf{A}$ and $\psi(t) = e^{-ct}\varphi(t)$. By the product rule (D1) and (a),

$$\psi'(t) = -c\,e^{-ct}\varphi(t) + e^{-ct}c\,\varphi(t) = 0$$

for every t, where $\frac{d}{dt}e^{-ct} = -ce^{-ct}$ is Theorem 10.9.4 (b) in the case $n = 1$. A differentiable function with vanishing derivative is constant — fact (A4) of Chapter 10 §09 — so $\psi(t) = \psi(0) = \det(e^0) = \det\mathbf{I} = 1$. Hence $\varphi(t) = e^{ct}$, and $t = 1$ gives $\det e^{\mathbf{A}} = e^{\operatorname{tr}\mathbf{A}}$.

Exercise. C2: Comparing two exponentials

Let $\mathbf{A}, \mathbf{B} \in M_n(\mathbb{C})$ with $\mathbf{AB} = \mathbf{BA}$.

(a) Prove that $\|e^{\mathbf{A}} - e^{\mathbf{B}}\| \le e^{\|\mathbf{B}\|}\left(e^{\|\mathbf{A}-\mathbf{B}\|} - 1\right)$.

(b) Deduce that $\mathbf{M} \mapsto e^{\mathbf{M}}$ is continuous at every $\mathbf{M}$ that commutes with a neighborhood of itself, and identify those $\mathbf{M}$: they are exactly the scalar matrices $c\mathbf{I}$.

(c) Say exactly where the commuting hypothesis is used, and why the argument cannot simply be dropped.

Solution.

(a) Since $\mathbf{A}$ and $\mathbf{B}$ commute, so do $\mathbf{B}$ and $\mathbf{A}-\mathbf{B}$. By Theorem 10.9.4 (d), $e^{\mathbf{A}} = e^{\mathbf{B}+(\mathbf{A}-\mathbf{B})} = e^{\mathbf{B}}e^{\mathbf{A}-\mathbf{B}}$. Hence

$$e^{\mathbf{A}} - e^{\mathbf{B}} = e^{\mathbf{B}}(e^{\mathbf{A}-\mathbf{B}} - \mathbf{I}),$$

and submultiplicativity with Theorem 16.6.1 (a) and (b) gives

$$\|e^{\mathbf{A}} - e^{\mathbf{B}}\| \le \|e^{\mathbf{B}}\|\,\|e^{\mathbf{A}-\mathbf{B}} - \mathbf{I}\| \le e^{\|\mathbf{B}\|}(e^{\|\mathbf{A}-\mathbf{B}\|} - 1).$$

(b) Take $\mathbf{B} = c\mathbf{I}$, which commutes with every matrix. For $\mathbf{A}$ with $\|\mathbf{A} - c\mathbf{I}\| = \delta$, part (a) gives $\|e^{\mathbf{A}} - e^{c}\mathbf{I}\| \le e^{|c|}(e^{\delta} - 1)$, and $e^{\delta} - 1 \to 0$ as $\delta \to 0$, because $t \mapsto e^{t}$ is differentiable by fact (A2) of Chapter 10 §09 and so continuous at 0. So $e^{\mathbf{A}} \to e^{c\mathbf{I}}$ as $\mathbf{A} \to c\mathbf{I}$.

(c) Only in the first line, where $e^{\mathbf{B}+(\mathbf{A}-\mathbf{B})}$ is split as $e^{\mathbf{B}}e^{\mathbf{A}-\mathbf{B}}$; Theorem 10.9.4 (d) has commutation as a hypothesis, and Chapter 10 §09 exhibits matrices with $e^{\mathbf{A}+\mathbf{B}} \neq e^{\mathbf{A}}e^{\mathbf{B}}$. Without it there is no factorization to take norms of, and a different argument is needed. (The inequality itself is in fact true in general, but proving it needs an integral representation of $e^{\mathbf{A}} - e^{\mathbf{B}}$, which this book does not have.)

Exercise. C3: A quantity that measures the decay

Let $\mathbf{A}, \mathbf{X}, \mathbf{Q} \in M_n(\mathbb{C})$ with $\mathbf{X}$ and $\mathbf{Q}$ positive definite (Definition 13.1.1) and

$$\mathbf{A}^*\mathbf{X} + \mathbf{X}\mathbf{A} = -\mathbf{Q}.$$

Let $\mathbf{x}\colon \mathbb{R} \to \mathbb{C}^n$ be differentiable with $\mathbf{x}' = \mathbf{A}\mathbf{x}$, and put $V(t) = \mathbf{x}(t)^*\mathbf{X}\mathbf{x}(t)$.

(a) Prove that V is real-valued and differentiable, with $V'(t) = \mathbf{x}(t)^*(\mathbf{A}^*\mathbf{X} + \mathbf{X}\mathbf{A})\mathbf{x}(t)$.

(b) Deduce that if $\mathbf{x}(0) \neq \mathbf{0}$, then V is strictly decreasing.

(c) Explain how this relates to the Lyapunov equation of Definition 12.10.2, whose unknown sits between $\mathbf{A}$ and $\mathbf{A}^*$ the other way round.

Hint for (a): $\mathbf{x}^*$ *is a* $1 \times n$ *matrix-valued function, and* $\overline{f}' = \overline{f'}$ *for a differentiable* $f\colon \mathbb{R} \to \mathbb{C}$.

Solution.

(a) $V(t)$ is a 1×1 matrix, and $V(t)^* = \mathbf{x}^*\mathbf{X}^*\mathbf{x} = \mathbf{x}^*\mathbf{X}\mathbf{x} = V(t)$ since $\mathbf{X}$ is Hermitian, so $V(t)$ is real. Each entry of $t \mapsto \mathbf{x}(t)^*$ is the conjugate of an entry of $\mathbf{x}$, hence differentiable with $(\mathbf{x}^*)' = (\mathbf{x}')^* = (\mathbf{A}\mathbf{x})^* = \mathbf{x}^*\mathbf{A}^*$. By Lemma 16.6.4, applied twice to the product $\mathbf{x}^* \cdot \mathbf{X} \cdot \mathbf{x}$ and using $\mathbf{X}' = \mathbf{O}$,

$$V'(t) = \mathbf{x}^*\mathbf{A}^*\mathbf{X}\mathbf{x} + \mathbf{x}^*\mathbf{X}\mathbf{A}\mathbf{x} = \mathbf{x}^*(\mathbf{A}^*\mathbf{X} + \mathbf{X}\mathbf{A})\mathbf{x}.$$

(b) By hypothesis $V'(t) = -\mathbf{x}(t)^*\mathbf{Q}\mathbf{x}(t)$. If $\mathbf{x}(t_1) = \mathbf{0}$ for some t_1, then by Theorem 10.9.6 $\mathbf{x}(t) = e^{t\mathbf{A}}\mathbf{x}(0)$ for every t, and $e^{t_1\mathbf{A}}$ is invertible by Theorem 10.9.4 (c), so $\mathbf{x}(0) = \mathbf{0}$, contradicting the hypothesis. So $\mathbf{x}(t) \neq \mathbf{0}$ for every t, and $\mathbf{Q} \succ \mathbf{O}$ gives $\mathbf{x}(t)^*\mathbf{Q}\mathbf{x}(t) > 0$, hence $V'(t) < 0$ for every t. A differentiable real function with everywhere negative derivative is strictly decreasing; that last step is the mean value theorem, one more fact of one-variable calculus quoted here and not proved.

(c) Definition 12.10.2 writes the equation as $\mathbf{A}\mathbf{X} + \mathbf{X}\mathbf{A}^* = -\mathbf{Q}$; the equation used here is that one for the pair $(\mathbf{A}^*, \mathbf{Q})$. The two are interchanged by replacing $\mathbf{A}$ with $\mathbf{A}^*$, which changes nothing about stability, since the eigenvalues of $\mathbf{A}^*$ are the complex conjugates

of those of $\mathbf{A}$ and conjugation preserves real parts. So the decay of V is exactly the statement Chapter 12 §10 pointed forward to, for whichever of the two conventions one adopts, and such an $\mathbf{X}$ does exist: Theorem 13.6.3 applied to $\mathbf{A}^*$ produces one.

16.7 Eigenvalues Move Continuously

Chapter 12 §01 proved that every complex matrix has diagonalizable matrices arbitrarily close to it (Corollary 12.1.6), then admitted that the word "close" was doing work no definition supported, and handed the repair to this chapter. The repair is now in place: a norm on $M_n(\mathbb{C})$ says how close two matrices are, and Corollary 16.2.4 says that it does not matter which norm. This section asks the question that could not even be posed before that point. If the matrix moves a little, how far do its eigenvalues move?

The answer has two halves, and the second is the one worth remembering. The eigenvalues do move only a little — provided they are read as an unordered list, which is the only way they can be read. But "a little" is not proportional to the movement of the matrix, and for a matrix with a large Jordan block it is spectacularly larger.

Throughout, matrices are complex, as in Section 4, so that an $n \times n$ matrix always has exactly n eigenvalues counted with algebraic multiplicity (Corollary 6.5.5).

16.7.1 From matrices to polynomials

The eigenvalues of $\mathbf{A}$ are the roots of $p_{\mathbf{A}}$ (Theorem 9.2.1), with algebraic multiplicity meaning multiplicity as a root (Definition 9.3.1). So the question splits in two: do the coefficients of $p_{\mathbf{A}}$ move a little when $\mathbf{A}$ does, and do the roots of a polynomial move a little when its coefficients do? The first half is easy and we do it now; the whole difficulty is in the second.

The reason the first half is easy is that everything in sight is a polynomial in the entries. It is worth isolating that observation once, because it will be used four times in this section.

Lemma 16.7.1. Polynomial Maps of the Entries Are Continuous

Let $\Phi\colon M_{p\times q}(\mathbb{C}) \to M_{r\times s}(\mathbb{C})$ be a map each of whose output entries is given by a fixed polynomial, with complex coefficients, in the pq entries of the input. Fix any norm on the source and any norm on the target. Then Φ is **continuous**: whenever $\|\mathbf{A}_k - \mathbf{A}\| \to 0$ we have $\|\Phi(\mathbf{A}_k) - \Phi(\mathbf{A})\| \to 0$, and equivalently, for every $\mathbf{A}$ and every $\varepsilon > 0$ there is $\delta > 0$ such that $\|\mathbf{B} - \mathbf{A}\| < \delta$ forces $\|\Phi(\mathbf{B}) - \Phi(\mathbf{A})\| < \varepsilon$. The same holds for a map into $\mathbb{C}$.

Idea. Two reductions and one piece of arithmetic. Convergence in any norm is entrywise convergence, by Corollary 16.2.4, in the source and in the target alike; so the claim is about pq complex sequences producing rs complex sequences. That reduces it to the statement that a polynomial in finitely many complex variables respects limits, which is the algebra of limits.

Proof. Suppose $\|\mathbf{A}_k - \mathbf{A}\| \to 0$. By Corollary 16.2.4, each entry $(\mathbf{A}_k)_{ij} \to a_{ij}$. Sums and products of convergent complex sequences converge to the sum and the product of the limits:

for sums this is Theorem 0.6.4, and for products it follows from

$$|z_k w_k - z w| \le |z_k|\,|w_k - w| + |w|\,|z_k - z|$$

together with the fact that a convergent sequence is bounded. A polynomial is built from its variables and constants by finitely many sums and products, so each entry of $\Phi(\mathbf{A}_k)$ converges to the corresponding entry of $\Phi(\mathbf{A})$. Applying Corollary 16.2.4 in $M_{r\times s}(\mathbb{C})$ gives $\|\Phi(\mathbf{A}_k) - \Phi(\mathbf{A})\| \to 0$.

For the equivalence of the two formulations, suppose the ε–δ statement failed at some $\mathbf{A}$ for some $\varepsilon > 0$. Then for each $k \ge 1$ the value $\delta = 1/k$ is not good enough, so there is $\mathbf{B}_k$ with $\|\mathbf{B}_k - \mathbf{A}\| < 1/k$ and $\|\Phi(\mathbf{B}_k) - \Phi(\mathbf{A})\| \ge \varepsilon$. The sequence $(\mathbf{B}_k)$ contradicts the sequential statement just proved. This proves the lemma.

■

Now the first half of the question is answered. Write $p_{\mathbf{A}}(x) = x^n + c_{n-1}(\mathbf{A})x^{n-1} + \cdots + c_0(\mathbf{A})$, which is legitimate because $p_{\mathbf{A}}$ is monic of degree n (Theorem 7.10.2). Each c_j is a polynomial in the entries of $\mathbf{A}$: the Leibniz sum of Definition 7.10.1 expresses $\det(x\mathbf{I} - \mathbf{A})$ as a signed sum of products of the entries of $x\mathbf{I} - \mathbf{A}$, each of which is $x - a_{jj}$ or $-a_{ij}$, and collecting the powers of x leaves every coefficient a polynomial with integer coefficients in the a_{ij}. So $\mathbf{A} \mapsto (c_0(\mathbf{A}), \dots, c_{n-1}(\mathbf{A}))$ is continuous by Lemma 16.7.1.

That leaves the second half: roots against coefficients. Before stating what is true there, we settle what cannot be true.

16.7.2 Why the answer must be about the multiset

The statement a reader would write down first is *"each eigenvalue is a continuous function of the matrix"*. It is not merely unproved; it is meaningless as written, because the eigenvalues come with no numbering — and the natural repair, "there is a way of numbering them that makes each one continuous", is false. One two-by-two family shows it, and the failure is the ordinary square root.

Proposition 16.7.2. There Is No Continuous Choice of a Single Eigenvalue

For $t \in \mathbb{C}$ put

$$\mathbf{A}(t) = \begin{pmatrix} 0 & 1 \\ t & 0 \end{pmatrix},$$

so that $p_{\mathbf{A}(t)}(x) = x^2 - t$ and the eigenvalues of $\mathbf{A}(t)$ are the two square roots of t. Then there is **no** continuous function $\lambda\colon D \to \mathbb{C}$ on the closed unit disc $D = \{t \in \mathbb{C} : |t| \le 1\}$ such that $\lambda(t)$ is an eigenvalue of $\mathbf{A}(t)$ for every $t \in D$. Consequently there is no continuous $\Lambda\colon M_2(\mathbb{C}) \to \mathbb{C}$ with $\Lambda(\mathbf{B}) \in \operatorname{spec}(\mathbf{B})$ for every $\mathbf{B}$.

Idea. Walk t once around the unit circle. The two roots $\pm\sqrt{t}$ trade places on the way round, so a choice that is continuous along the walk comes home as the *other* root. The bookkeeping is done by dividing the chosen root by one fixed root, which turns the walk into a continuous function with only the two values 1 and -1; on an interval such a function cannot jump, and

it must.

Proof. The characteristic polynomial is $\det\left(\begin{smallmatrix} x & -1 \\ -t & x \end{smallmatrix}\right) = x^2 - t$, so μ is an eigenvalue of $\mathbf{A}(t)$ if and only if $\mu^2 = t$ (Theorem 9.2.1).

Suppose such a λ exists, so $\lambda(t)^2 = t$ for all $|t| \le 1$. Define $g\colon [0, 2\pi] \to \mathbb{C}$ by

$$g(\theta) = \lambda(e^{i\theta})\, e^{-i\theta/2}.$$

It is continuous, being a composite and a product of continuous functions — the elementary algebra of limits recorded in the chapter introduction, applied to $\theta \mapsto e^{i\theta}$, to λ, and to $\theta \mapsto e^{-i\theta/2}$. Squaring,

$$g(\theta)^2 = \lambda(e^{i\theta})^2 e^{-i\theta} = e^{i\theta} e^{-i\theta} = 1,$$

so $g(\theta) \in \{1, -1\}$ for every θ. In particular g is real-valued, and a continuous real-valued function on an interval taking only the values 1 and -1 is constant: if it took both, the intermediate value theorem would produce a θ with $g(\theta) = 0$. Hence $g(2\pi) = g(0)$.

But $e^{i\cdot 0} = e^{i\cdot 2\pi} = 1$, so $g(0) = \lambda(1)$ while $g(2\pi) = \lambda(1)e^{-i\pi} = -\lambda(1)$. Therefore $\lambda(1) = -\lambda(1)$, that is $\lambda(1) = 0$, contradicting $\lambda(1)^2 = 1$. So no such λ exists.

For the last sentence, if Λ were such a function then $t \mapsto \Lambda(\mathbf{A}(t))$ would be continuous, because $t \mapsto \mathbf{A}(t)$ is continuous by Lemma 16.7.1, and it would select an eigenvalue of $\mathbf{A}(t)$ for each t. This proves the proposition.

■

The intermediate value theorem is quoted here, and it is the only place in this chapter that uses it; it is fact (A5) of the chapter introduction, borrowed and not proved.

Two things are worth reading off the argument. First, the obstruction is not exotic: $\mathbf{A}(0) = \mathbf{J}_2(0)$ is the standard nilpotent matrix, and the trouble happens exactly at the parameter where the two eigenvalues collide. Second, nothing is wrong with the *pair* $\{\sqrt{t}, -\sqrt{t}\}$, which shrinks to $\{0, 0\}$ as $t \to 0$ in any reasonable sense. It is only the attempt to call one of them "the first" that fails. This is why the correct statement speaks of the roots as a **multiset**: a list with multiplicities and no order.

16.7.3 The imported fact

Here is what is true. It is the one statement in this chapter that the chapter does not prove.

Theorem 16.7.3. Continuous Dependence of the Roots on the Coefficients

Let $n \ge 1$ and let

$$p(x) = x^n + c_{n-1}x^{n-1} + \cdots + c_1 x + c_0 \in \mathbb{C}[x]$$

be monic, with roots $\mu_1, \ldots, \mu_n \in \mathbb{C}$ listed with multiplicity (Corollary 6.5.5). Let $\varepsilon > 0$. Then there is $\delta > 0$ with the following property. If

$$q(x) = x^n + b_{n-1}x^{n-1} + \cdots + b_1 x + b_0$$

is monic of the **same degree** n and $|b_j - c_j| < \delta$ for every j, then the roots $\nu_1, \dots, \nu_n$ of q, listed with multiplicity, can be **numbered** so that

$$|\nu_i - \mu_i| < \varepsilon \qquad (i = 1, \dots, n).$$

This chapter quotes this theorem, and Chapter 20 proves it. The classical proof uses complex analysis: one encircles each root of p by a small circle on which p does not vanish, counts the roots inside by a contour integral, and observes that an integer-valued quantity depending continuously on the coefficients cannot change. The book never develops that machinery, and it does not need to: compactness, fact (A3), together with unique factorization in $\mathbb{C}[x]$, is enough, and Chapter 20 §03 gives that proof. It is postponed there because it belongs with the quantitative bounds on how far roots move. Until then the statement is recorded in the manner this book has used before — Chapter 14 §06 with Taylor's theorem, Chapter 11 with the theorem of Abel and Ruffini: named at the point of use, never buried inside a proof that presents itself as self-contained, and never strengthened in passing. What *can* be proved with the tools at hand is that the numbering in the statement has to be chosen afresh for each q, and that was Proposition 16.7.2.

Two clauses of the statement carry weight. **Monic of the same degree** is essential: the roots of $\delta x^2 + x$ are 0 and $-1/\delta$, which runs off to infinity as $\delta \to 0$, while the limit polynomial x has the single root 0. Allowing the leading coefficient to vanish destroys everything, and monic polynomials of a fixed degree are exactly the family where it cannot. And **numbered so that** is the multiset formulation: the permutation matching up the two lists depends on q, and Proposition 16.7.2 says it must.

16.7.4 Eigenvalues, and the spectral radius

Now the two halves are assembled.

Corollary 16.7.4. Eigenvalues Depend Continuously on the Matrix

Let $\mathbf{A} \in M_n(\mathbb{C})$ with eigenvalues $\lambda_1, \dots, \lambda_n$ listed with algebraic multiplicity, let $\|\cdot\|$ be any norm on $M_n(\mathbb{C})$, and let $\varepsilon > 0$. Then there is $\delta > 0$ such that every $\mathbf{B} \in M_n(\mathbb{C})$ with $\|\mathbf{B} - \mathbf{A}\| < \delta$ has eigenvalues $\mu_1, \dots, \mu_n$, listed with algebraic multiplicity, which can be numbered so that $|\mu_i - \lambda_i| < \varepsilon$ for every i.

Idea. The characteristic polynomial's coefficients are polynomials in the entries, so they move continuously with $\mathbf{A}$; the quoted root theorem then moves the roots. The only work is composing the two estimates, choosing the tolerance for the coefficients small enough that the roots come out within ε.

Proof. By Theorem 7.10.2 the polynomials $p_{\mathbf{A}}$ and $p_{\mathbf{B}}$ are monic of degree n, and by Theorem 9.2.1 together with Definition 9.3.1 their root lists with multiplicity are the eigenvalue lists with algebraic multiplicity. Let $\delta' > 0$ be the number supplied by Theorem 16.7.3 for $p = p_{\mathbf{A}}$ and this ε. The map $\mathbf{B} \mapsto (c_0(\mathbf{B}), \dots, c_{n-1}(\mathbf{B}))$ is continuous by Lemma 16.7.1, as noted after that lemma, so there is $\delta > 0$ such that $\|\mathbf{B} - \mathbf{A}\| < \delta$ forces $|c_j(\mathbf{B}) - c_j(\mathbf{A})| < \delta'$ for every j. For such a $\mathbf{B}$, Theorem 16.7.3 applies to $q = p_{\mathbf{B}}$. This proves the corollary.

■

The corollary says exactly as much as Theorem 16.7.3 and no more: the numbering is not canonical and depends on $\mathbf{B}$. Any quantity that ignores the numbering, however, inherits honest continuity, and the spectral radius of Section 4 is the first example.

Quick check. Does Corollary 16.7.4 make $\rho\colon M_n(\mathbb{C}) \to \mathbb{R}$ continuous, given that no individual eigenvalue is a continuous function of $\mathbf{A}$?

Solution. Yes. Let $\varepsilon > 0$ and take δ from Corollary 16.7.4. If $\|\mathbf{B} - \mathbf{A}\| < \delta$, fix a numbering with $|\mu_i - \lambda_i| < \varepsilon$. Then $|\mu_i| \le |\lambda_i| + \varepsilon \le \rho(\mathbf{A}) + \varepsilon$ for each i by Definition 16.4.1, so $\rho(\mathbf{B}) \le \rho(\mathbf{A}) + \varepsilon$; the same numbering read the other way gives $|\lambda_i| \le |\mu_i| + \varepsilon \le \rho(\mathbf{B}) + \varepsilon$ for each i, hence $\rho(\mathbf{A}) \le \rho(\mathbf{B}) + \varepsilon$, so $|\rho(\mathbf{B}) - \rho(\mathbf{A})| \le \varepsilon$. The argument never names an individual eigenvalue, which is why it survives: ρ is a function of the multiset alone.

16.7.5 The limit argument, made routine

Chapter 12 §01 called Corollary 12.1.6 "the tool behind a standard strategy: *prove it for diagonalizable matrices, then take a limit*", and said that this chapter, "where matrices acquire norms and limits can be spoken of properly, runs arguments of exactly this shape". Here is the shape. It costs two corollaries, and afterwards the strategy is three lines every time.

Corollary 16.7.5. Diagonalizable Matrices Are Dense

Let $\|\cdot\|$ be any norm on $M_n(\mathbb{C})$. For every $\mathbf{A} \in M_n(\mathbb{C})$ and every $\varepsilon > 0$ there is a **diagonalizable** $\mathbf{A}' \in M_n(\mathbb{C})$ with $\|\mathbf{A} - \mathbf{A}'\| < \varepsilon$. Equivalently, every complex square matrix is the limit of a sequence of diagonalizable matrices, in every norm.

Idea. Chapter 12 produced a diagonalizable matrix near $\mathbf{A}$ with the *entries* close. "Close in every norm" is now the same thing, by Corollary 16.2.4, so the only task is to convert one tolerance into the other.

Proof. Let $N_\infty(\mathbf{M}) = \max_{i,j} |m_{ij}|$, which is a norm on $M_n(\mathbb{C})$: it is the ∞-norm of Example (The 1-, 2-, and ∞-norms on F^n) read on the list of n^2 entries. By Theorem 16.2.2 there is $C > 0$ with $\|\mathbf{M}\| \le C\, N_\infty(\mathbf{M})$ for every $\mathbf{M}$. Apply Corollary 12.1.6 with the tolerance ε/C: it supplies a diagonalizable $\mathbf{A}'$ with $|a_{ij} - a'_{ij}| < \varepsilon/C$ for all i, j, hence with $N_\infty(\mathbf{A} - \mathbf{A}') < \varepsilon/C$, a maximum of finitely many numbers each below that bound. Therefore $\|\mathbf{A} - \mathbf{A}'\| \le C\, N_\infty(\mathbf{A} - \mathbf{A}') < \varepsilon$. Taking $\varepsilon = 1/k$ gives a sequence $\mathbf{A}_k$ of diagonalizable matrices with $\|\mathbf{A} - \mathbf{A}_k\| \to 0$. This proves the corollary. ∎

Corollary 16.7.6. Proving an Identity by Density

Let $\Phi\colon M_n(\mathbb{C}) \to M_{r\times s}(\mathbb{C})$ be a map each of whose output entries is a polynomial in the entries of the input. If $\Phi(\mathbf{A}) = \mathbf{O}$ for every **diagonalizable** $\mathbf{A} \in M_n(\mathbb{C})$, then $\Phi(\mathbf{A}) = \mathbf{O}$

for every $\mathbf{A} \in M_n(\mathbb{C})$.

Idea. Both sides of the identity are continuous in $\mathbf{A}$, and they agree on the diagonalizable matrices, which are dense. A continuous function is determined by its values on a dense set, so they agree everywhere.

Proof. Fix $\mathbf{A}$ and a norm on each side. By Corollary 16.7.5 choose diagonalizable $\mathbf{A}_k$ with $\|\mathbf{A} - \mathbf{A}_k\| \to 0$. By Lemma 16.7.1, $\|\Phi(\mathbf{A}_k) - \Phi(\mathbf{A})\| \to 0$. Since $\Phi(\mathbf{A}_k) = \mathbf{O}$ by hypothesis, that says $\|\Phi(\mathbf{A})\| \to 0$; but $\|\Phi(\mathbf{A})\|$ does not depend on k, so $\|\Phi(\mathbf{A})\| = 0$ and $\Phi(\mathbf{A}) = \mathbf{O}$ by the positivity clause of Definition 16.1.1. This proves the corollary.

■

The hypothesis "each output entry is a polynomial in the input entries" is the one to check, and it is usually free: any expression built from $\mathbf{A}$ by addition, multiplication, transposition, conjugate transposition of a real matrix, determinants, adjugates, traces, and the coefficients of $p_{\mathbf{A}}$ qualifies. What does *not* qualify is anything built from an inverse, an eigenvector, or a choice of ordering. Here is the strategy in action, on the theorem that Chapter 9 proved the long way.

Example. Cayley–Hamilton, by a limit

Prove that $p_{\mathbf{A}}(\mathbf{A}) = \mathbf{O}$ for every $\mathbf{A} \in M_n(\mathbb{C})$, using density rather than the adjugate identity of Theorem 7.6.2.

Solution. Put $\Phi(\mathbf{A}) = p_{\mathbf{A}}(\mathbf{A})$. Each entry of $\Phi(\mathbf{A})$ is a polynomial in the entries of $\mathbf{A}$: the coefficients $c_j(\mathbf{A})$ are, as noted after Lemma 16.7.1, and $\Phi(\mathbf{A}) = \mathbf{A}^n + c_{n-1}(\mathbf{A})\mathbf{A}^{n-1} + \cdots + c_0(\mathbf{A})\mathbf{I}$ is then built from them by products and sums.

Let $\mathbf{A}$ be diagonalizable, say $\mathbf{A} = \mathbf{P}\mathbf{D}\mathbf{P}^{-1}$ with $\mathbf{D} = \operatorname{diag}(d_1, \dots, d_n)$. By Theorem 9.4.3, $p_{\mathbf{A}}(\mathbf{A}) = \mathbf{P}\operatorname{diag}(p_{\mathbf{A}}(d_1), \dots, p_{\mathbf{A}}(d_n))\,\mathbf{P}^{-1}$. Each d_i is an eigenvalue of $\mathbf{A}$, hence a root of $p_{\mathbf{A}}$ (Theorem 9.2.1), so every diagonal entry is 0 and $\Phi(\mathbf{A}) = \mathbf{O}$.

By Corollary 16.7.6, $\Phi(\mathbf{A}) = \mathbf{O}$ for every $\mathbf{A} \in M_n(\mathbb{C})$, which is Theorem 9.6.1 over $\mathbb{C}$.

That is the whole strategy: check the identity where the matrix is diagonal in some basis, where everything is a scalar computation, and let density carry it to the rest. It proves nothing new here — Theorem 9.6.1 holds over every field, and the density argument is confined to $\mathbb{C}$ — but the three lines at the end are the part to transfer. Exercise (C2: Density of the invertible matrices, and **AB** versus **BA**) runs the same argument with a different dense set.

16.7.6 The warning: continuous is not Lipschitz

Everything so far has been qualitative. The quantitative question is whether there is a constant L with

$$\text{(eigenvalue movement)} \;\le\; L\,\|\mathbf{B} - \mathbf{A}\|,$$

which is what "continuous" is usually imagined to mean and what a numerical analyst actually needs. There is not, and the failure is severe.

Example. A Jordan block's eigenvalues move like $\varepsilon^{1/k}$

Let $k \geq 2$, let $\varepsilon > 0$, and perturb the nilpotent Jordan block in its bottom left corner:

$$\mathbf{A} = \mathbf{J}_k(0), \qquad \mathbf{A}_\varepsilon = \mathbf{J}_k(0) + \varepsilon \mathbf{E}_{k1}.$$

Find the eigenvalues of $\mathbf{A}_\varepsilon$ and compare their distance from $\operatorname{spec}(\mathbf{A}) = \{0\}$ with $\|\mathbf{A}_\varepsilon - \mathbf{A}\|$.

Solution. The characteristic polynomial is $x^k - \varepsilon$. Write $\mathbf{M} = x\mathbf{I} - \mathbf{A}_\varepsilon$, so $\mathbf{M}$ has x on the diagonal, -1 on the superdiagonal, $-\varepsilon$ in position $(k, 1)$, and zeros elsewhere. Expand along the first column (Theorem 7.6.1 (a)), whose only nonzero entries are $m_{11} = x$ and $m_{k1} = -\varepsilon$.

Deleting row 1 and column 1 leaves the $(k-1) \times (k-1)$ upper triangular matrix with x on the diagonal, of determinant x^{k-1} (Theorem 7.4.6). Deleting row k and column 1 leaves the matrix with rows $1, \dots, k-1$ and columns $2, \dots, k$ of $\mathbf{M}$; its (i, j) entry is $m_{i,j+1}$, which is -1 when $j = i$ and x when $i = j + 1$, so it is lower triangular with -1 all along its diagonal and determinant $(-1)^{k-1}$. The cofactor signs are $(-1)^{1+1} = 1$ and $(-1)^{k+1}$, so

$$\det \mathbf{M} = x \cdot x^{k-1} + (-\varepsilon)(-1)^{k+1}(-1)^{k-1} = x^k - \varepsilon,$$

since $(-1)^{k+1}(-1)^{k-1} = (-1)^{2k} = 1$.

The eigenvalues. They are the k-th roots of ε, namely $\varepsilon^{1/k} e^{2\pi i j/k}$ for $j = 0, \dots, k-1$, each of modulus $\varepsilon^{1/k}$. For $k = 2$ this is the pair $\pm\sqrt{\varepsilon}$, and $\mathbf{A}_\varepsilon$ is the matrix $\mathbf{A}(\varepsilon)$ of Proposition 16.7.2.

The comparison. The perturbation $\mathbf{E} = \varepsilon \mathbf{E}_{k1}$ has a single nonzero entry, so its largest absolute row sum and largest absolute column sum are both ε, and $\mathbf{E}^*\mathbf{E} = \varepsilon^2 \mathbf{E}_{11}$ has largest eigenvalue ε^2; hence $\|\mathbf{E}\|_1 = \|\mathbf{E}\|_\infty = \|\mathbf{E}\|_2 = \varepsilon$ by Theorem 16.3.3. Every eigenvalue of $\mathbf{A}$ is 0, so the eigenvalues have moved a distance $\varepsilon^{1/k}$ while the matrix moved a distance ε. The ratio is

$$\frac{\varepsilon^{1/k}}{\varepsilon} = \varepsilon^{-(k-1)/k} \xrightarrow[\varepsilon \to 0^+]{} \infty.$$

A by-product worth keeping. The k roots $\varepsilon^{1/k} e^{2\pi i j/k}$ are **distinct** for every $\varepsilon > 0$, so $\mathbf{A}_\varepsilon$ has k distinct eigenvalues and is diagonalizable by Corollary 9.4.2. A single nilpotent Jordan block, which is as far from diagonalizable as a matrix gets, therefore has diagonalizable matrices arbitrarily close to it, and here they are written down. That is Corollary 16.7.5 made concrete for one family.

Put numbers on it. For $k = 2$ and $\varepsilon = 10^{-6}$ the matrix moves by one part in a million and the eigenvalues move to $\pm 10^{-3}$: an amplification of a thousand. For $k = 10$ and $\varepsilon = 10^{-10}$ the matrix moves by 10^{-10} and the eigenvalues land on the circle of radius 10^{-1}, an amplification of a billion. No constant L survives this, so the continuity of Corollary 16.7.4 cannot be upgraded to a Lipschitz bound, and the δ it produces genuinely depends on $\mathbf{A}$ and not only on n.

Warning. **Continuous does not mean stable.** Corollary 16.7.4 guarantees that a small enough perturbation moves the eigenvalues a little, and says nothing whatever about how small "small enough" is. For $\mathbf{J}_{10}(0)$ a perturbation of size 10^{-10} — the kind of change one would call negligible in data recorded to ten digits — scatters the eigenvalues over a circle of radius 0.1. Reading a printed eigenvalue of a matrix with a large Jordan block as a property

of the matrix, rather than of the matrix plus whatever noise reached it, is the mistake this example exists to prevent.

The opposite extreme is worth seeing beside it. Take the Hermitian $\mathbf{A} = \operatorname{diag}(1, 2)$ and the Hermitian perturbation $\mathbf{E} = \varepsilon(\mathbf{E}_{12} + \mathbf{E}_{21})$, of norm $\|\mathbf{E}\|_2 = \varepsilon$. The matrix $\mathbf{A} + \mathbf{E} = \left(\begin{smallmatrix} 1 & \varepsilon \\ \varepsilon & 2 \end{smallmatrix}\right)$ has trace 3 and determinant $2 - \varepsilon^2$, so its eigenvalues are

$$\frac{3 \pm \sqrt{1 + 4\varepsilon^2}}{2},$$

and since $\sqrt{1+u} \le 1 + u/2$ for $u \ge 0$, each of them lies within ε^2 of the corresponding eigenvalue of $\mathbf{A}$. Here the eigenvalues move *less* than the matrix does, by a whole order in ε. That is not an accident of the example: Chapter 17 proves that the eigenvalues of a Hermitian matrix move by at most $\|\mathbf{E}\|_2$ under a Hermitian perturbation, and Chapter 20 proves the theorem of Bauer and Fike, which bounds the movement for a diagonalizable matrix by $\|\mathbf{E}\|_2$ times a factor measuring how far its eigenvector basis is from orthonormal.

So the exponent $1/k$ in Example (A Jordan block's eigenvalues move like $\varepsilon^{1/k}$) is not about size; it is about shape. It is decided by the largest Jordan block at the eigenvalue (Theorem 10.3.1), and block size is precisely what Chapter 12 §11 was measuring when Example (The defect of a Jordan block) computed

$$\Delta(\mathbf{J}_k(\lambda)) = \sqrt{k - 1}$$

for the departure from normality of Definition 12.11.1. The two statements match at both ends: $\Delta = 0$ exactly for the normal matrices, whose eigenvalues are the stable ones, and Δ grows with the block that makes them fragile. They do not match in between, and it would be wrong to read Δ as a formula for the exponent — Δ is one number for the whole matrix, so it cannot report which eigenvalue carries which block, and Example (The defect of a Jordan block) records that it does not depend on λ at all. What $\Delta(\mathbf{A})$ does is tell you whether the question needs asking at all.

16.7.7 Exercises

16.7.7.1 A. Check your understanding

Exercise. A1

(a) State Theorem 16.7.3, paying attention to the two hypotheses on q.

(b) Explain in one sentence why the theorem is stated for the multiset of roots rather than for a list of n individual roots.

(c) Decide whether the following is correct, and justify your answer: "the function sending $\mathbf{A} \in M_2(\mathbb{C})$ to its eigenvalue of largest modulus is continuous".

(d) For the $k \times k$ Jordan block, how far do the eigenvalues move under a perturbation of norm ε, and what does that say about a Lipschitz bound?

Solution.

(a) See Theorem 16.7.3. The two hypotheses are that q is **monic** and of the **same degree** n as p; without them the roots can escape to infinity, as $\delta x^2 + x$ shows.

(b) Because no continuous choice of an individual root exists: Proposition 16.7.2 shows that following one square root of t once around the unit circle returns the other one,

so the matching numbering must be allowed to depend on q.

(c) Incorrect as stated, because a matrix can have two eigenvalues of equal largest modulus and then the phrase does not name a number; and incorrect even where it does name a number, since $\mathbf{A}(t)$ of Proposition 16.7.2 has the two eigenvalues $\pm\sqrt{t}$ of equal modulus for every $t \neq 0$. The modulus itself, that is ρ, is continuous — see the Quick check.

(d) They move a distance $\varepsilon^{1/k}$ (Example (A Jordan block's eigenvalues move like $\varepsilon^{1/k}$)), so the ratio of eigenvalue movement to matrix movement is $\varepsilon^{-(k-1)/k}$, which is unbounded as $\varepsilon \to 0^+$. No Lipschitz constant exists.

16.7.7.2 B. Practice

Exercise. B1: Two perturbations of the same size

Let $\varepsilon = 10^{-8}$. Compute the eigenvalues of

$$\mathbf{B} = \begin{pmatrix} 0 & 1 \\ \varepsilon & 0 \end{pmatrix} \qquad \text{and} \qquad \mathbf{C} = \begin{pmatrix} \varepsilon & 1 \\ 0 & 0 \end{pmatrix},$$

each of which differs from $\mathbf{J}_2(0)$ by a matrix of 2-norm ε. Hence explain why the constant in Corollary 16.7.4 cannot depend on $\|\mathbf{B} - \mathbf{A}\|$ alone.

Solution. For $\mathbf{B}$: $p_{\mathbf{B}}(x) = x^2 - \varepsilon$ by Example (A Jordan block's eigenvalues move like $\varepsilon^{1/k}$) with $k = 2$, so the eigenvalues are $\pm 10^{-4}$.

For $\mathbf{C}$: the matrix is upper triangular, so its eigenvalues are its diagonal entries $\varepsilon = 10^{-8}$ and 0 (Theorem 9.7.4 (b)).

Both perturbations have 2-norm $\varepsilon = 10^{-8}$, since each is ε times a single matrix unit. The eigenvalues of $\mathbf{B}$ have moved 10^{-4} from $\operatorname{spec}(\mathbf{J}_2(0)) = \{0\}$, those of $\mathbf{C}$ only 10^{-8}: a factor of 10^4 between two perturbations of exactly the same size. So the movement is not a function of $\|\mathbf{E}\|$, and the δ of Corollary 16.7.4 cannot be computed from ε and n alone.

Exercise. B2: A diagonalizable neighbor, explicitly

Let $\mathbf{A} = \begin{pmatrix} 2 & 1 \\ 0 & 2 \end{pmatrix}$ and let $\varepsilon > 0$.

(a) Show that $\mathbf{A}$ is not diagonalizable.

(b) Exhibit a diagonalizable $\mathbf{A}'$ with $\|\mathbf{A} - \mathbf{A}'\|_2 < \varepsilon$, and say which theorem certifies that your $\mathbf{A}'$ is diagonalizable.

(c) What are the eigenvalues of your $\mathbf{A}'$, and how far have they moved?

Solution.

(a) The only eigenvalue is 2, with algebraic multiplicity 2 (Theorem 9.7.4 (b)). But $\mathbf{A} - 2\mathbf{I} = \left(\begin{smallmatrix} 0 & 1 \\ 0 & 0 \end{smallmatrix}\right)$ has rank 1, so $g_{\mathbf{A}}(2) = 1 < 2 = a_{\mathbf{A}}(2)$ and $\mathbf{A}$ is not diagonalizable by Theorem 9.4.1 (e).

(b) Take δ with $0 < \delta < \varepsilon$ and $\mathbf{A}' = \left(\begin{smallmatrix} 2 & 1 \\ 0 & 2+\delta \end{smallmatrix}\right)$. It is upper triangular with distinct diagonal entries 2 and $2+\delta$, hence has two distinct eigenvalues (Theorem 9.7.4 (b)) and is diagonalizable by Corollary 9.4.2. And $\mathbf{A} - \mathbf{A}' = -\delta \mathbf{E}_{22}$, so $\|\mathbf{A} - \mathbf{A}'\|_2 = \delta < \varepsilon$ by Theorem 16.3.3 (c).

(c) They are 2 and $2+\delta$. One eigenvalue has not moved and the other has moved δ, which is exactly $\|\mathbf{A} - \mathbf{A}'\|_2$: here the movement is Lipschitz with constant 1. The failure in Example (A Jordan block's eigenvalues move like $\varepsilon^{1/k}$) needed a perturbation *below* the diagonal, which is the direction a triangular form cannot absorb.

Exercise. B3: The determinant and the trace, by density

Let $\mathbf{A}, \mathbf{B} \in M_n(\mathbb{C})$. Use Corollary 16.7.6 to prove that $\operatorname{tr}(\mathbf{AB}) = \operatorname{tr}(\mathbf{BA})$ — after first checking that the map in question satisfies the hypothesis of that corollary — and then explain why this is a bad way to prove it.

Solution. Fix $\mathbf{B}$ and put $\Phi(\mathbf{A}) = \operatorname{tr}(\mathbf{AB}) - \operatorname{tr}(\mathbf{BA})$, a 1×1 matrix whose single entry is $\sum_{i,j}(a_{ij}b_{ji} - b_{ij}a_{ji})$, a polynomial in the entries of $\mathbf{A}$; so the hypothesis of Corollary 16.7.6 holds.

Let $\mathbf{A} = \mathbf{PDP}^{-1}$ be diagonalizable, $\mathbf{D} = \operatorname{diag}(d_1, \dots, d_n)$, and put $\mathbf{X} = \mathbf{P}^{-1}\mathbf{BP}$. By Theorem 4.2.5, similar matrices have equal traces, and $\mathbf{AB} = \mathbf{P}(\mathbf{DX})\mathbf{P}^{-1}$ while $\mathbf{BA} = \mathbf{P}(\mathbf{XD})\mathbf{P}^{-1}$, so $\operatorname{tr}(\mathbf{AB}) = \operatorname{tr}(\mathbf{DX})$ and $\operatorname{tr}(\mathbf{BA}) = \operatorname{tr}(\mathbf{XD})$. For a diagonal $\mathbf{D}$ both of those are $\sum_i d_i x_{ii}$, by inspection of the diagonal entries of the two products. Hence $\Phi(\mathbf{A}) = \mathbf{O}$ for every diagonalizable $\mathbf{A}$, and Corollary 16.7.6 extends it to every $\mathbf{A}$.

It is a bad way to prove it because the identity is a two-line computation from the definition, $\operatorname{tr}(\mathbf{AB}) = \sum_{i,j} a_{ij}b_{ji} = \operatorname{tr}(\mathbf{BA})$, valid over every field; the density argument needs $\mathbb{C}$, an imported theorem, and two corollaries, and delivers less. A density argument earns its keep only when the diagonalizable case is genuinely easier, as in Example (Cayley–Hamilton, by a limit).

16.7.7.3 C. Going deeper

Exercise. C1: How close to singular a nearby eigenvalue forces you to be

Let $\mathbf{A} \in M_n(\mathbb{C})$, let $\|\cdot\|$ be an operator norm on $M_n(\mathbb{C})$ (Definition 16.3.2), and let $\mathbf{E} \in M_n(\mathbb{C})$ with $\mathbf{E} \neq \mathbf{O}$.

(a) Suppose μ is an eigenvalue of $\mathbf{A} + \mathbf{E}$ and $\mu \notin \operatorname{spec}(\mathbf{A})$. Prove that

$$\|(\mathbf{A} - \mu\mathbf{I})^{-1}\| \geq \frac{1}{\|\mathbf{E}\|}.$$

(b) Deduce that if $\|(\mathbf{A} - z\mathbf{I})^{-1}\| \leq M$ for every z with $|z - \lambda| = r$, then no eigenvalue of $\mathbf{A} + \mathbf{E}$ lies on that circle as soon as $\|\mathbf{E}\| < 1/M$.

Hint: for (a), factor $\mathbf{A} + \mathbf{E} - \mu\mathbf{I}$ *and use Theorem 16.5.1.*

Solution.

(a) Since $\mu \notin \operatorname{spec}(\mathbf{A})$, the matrix $\mathbf{A} - \mu\mathbf{I}$ is invertible (Theorem 9.2.1). Write

$$\mathbf{A} + \mathbf{E} - \mu\mathbf{I} = (\mathbf{A} - \mu\mathbf{I})(\mathbf{I} + (\mathbf{A} - \mu\mathbf{I})^{-1}\mathbf{E}).$$

The left side is singular, because μ is an eigenvalue of $\mathbf{A} + \mathbf{E}$. The first factor on the right is invertible, so the second factor is singular: a product with an invertible left factor is invertible as soon as the right factor is. Put $\mathbf{X} = -(\mathbf{A} - \mu\mathbf{I})^{-1}\mathbf{E}$, so $\mathbf{I} - \mathbf{X}$ is singular. By Theorem 16.5.1, $\|\mathbf{X}\| < 1$ would make $\mathbf{I} - \mathbf{X}$ invertible; hence $\|\mathbf{X}\| \geq 1$. Submultiplicativity (Theorem 16.3.2) gives

$$1 \leq \|\mathbf{X}\| \leq \|(\mathbf{A} - \mu\mathbf{I})^{-1}\| \, \|\mathbf{E}\|,$$

and dividing by $\|\mathbf{E}\| > 0$ gives the claim.

(b) Suppose μ is an eigenvalue of $\mathbf{A} + \mathbf{E}$ with $|\mu - \lambda| = r$. Then $\|(\mathbf{A} - \mu\mathbf{I})^{-1}\| \leq M$ by hypothesis, and in particular $\mathbf{A} - \mu\mathbf{I}$ is invertible, so $\mu \notin \operatorname{spec}(\mathbf{A})$ and part (a) applies: $M \geq 1/\|\mathbf{E}\|$, that is $\|\mathbf{E}\| \geq 1/M$. Contrapositively, $\|\mathbf{E}\| < 1/M$ leaves no eigenvalue of $\mathbf{A} + \mathbf{E}$ on the circle.

Exercise. C2: Density of the invertible matrices, and AB versus BA

(a) Prove that for every $\mathbf{A} \in M_n(\mathbb{C})$ and every $\varepsilon > 0$ there is an **invertible** $\mathbf{A}'$ with $\|\mathbf{A} - \mathbf{A}'\| < \varepsilon$, in any norm. *Hint: consider* $\mathbf{A} - t\mathbf{I}$.

(b) Prove a version of Corollary 16.7.6 with "invertible" in place of "diagonalizable".

(c) Hence give a second proof that $p_{\mathbf{AB}} = p_{\mathbf{BA}}$ for all $\mathbf{A}, \mathbf{B} \in M_n(\mathbb{C})$, a statement first proved in Theorem 9.10.4 (b).

Solution.

(a) The matrix $\mathbf{A} - t\mathbf{I}$ is singular exactly when $t \in \operatorname{spec}(\mathbf{A})$ (Theorem 9.2.1), and $\operatorname{spec}(\mathbf{A})$ has at most n elements. Let N_∞ and C be as in the proof of Corollary 16.7.5, so $\|\mathbf{M}\| \leq C\, N_\infty(\mathbf{M})$. Choose $t \neq 0$ with $|t| < \varepsilon/C$ and $t \notin \operatorname{spec}(\mathbf{A})$, possible because the disc of radius ε/C is infinite and only finitely many of its points are forbidden. Then $\mathbf{A}' = \mathbf{A} - t\mathbf{I}$ is invertible and $N_\infty(\mathbf{A} - \mathbf{A}') = |t|$, so $\|\mathbf{A} - \mathbf{A}'\| < \varepsilon$.

(b) Verbatim the proof of Corollary 16.7.6, with the sequence of diagonalizable matrices replaced by the sequence of invertible matrices supplied by (a) with $\varepsilon = 1/k$.

(c) Fix $\mathbf{B}$ and set $\Phi(\mathbf{A}) = (c_0(\mathbf{AB}) - c_0(\mathbf{BA}), \dots, c_{n-1}(\mathbf{AB}) - c_{n-1}(\mathbf{BA}))$, a row of n entries. Each entry of $\mathbf{AB}$ and of $\mathbf{BA}$ is a polynomial in the entries of $\mathbf{A}$, and each c_j is a polynomial in the entries of its argument, so Φ satisfies the hypothesis of (b). For invertible $\mathbf{A}$ we have $\mathbf{BA} = \mathbf{A}^{-1}(\mathbf{AB})\mathbf{A}$, so $\mathbf{AB}$ and $\mathbf{BA}$ are similar and $p_{\mathbf{AB}} = p_{\mathbf{BA}}$ by Theorem 7.10.3; that is exactly $\Phi(\mathbf{A}) = \mathbf{O}$, and it is Exercise (B3: **AB** and **BA**) (a). By (b), $\Phi(\mathbf{A}) = \mathbf{O}$ for every $\mathbf{A}$, and two monic polynomials of degree n with the same lower coefficients are equal.

Exercise. C3: Distinct eigenvalues is an open condition

Let $\mathbf{A} \in M_n(\mathbb{C})$ have n **distinct** eigenvalues $\lambda_1, \dots, \lambda_n$, and let $\|\cdot\|$ be a norm on $M_n(\mathbb{C})$.

(a) Prove that there is $\delta > 0$ such that every $\mathbf{B}$ with $\|\mathbf{B} - \mathbf{A}\| < \delta$ also has n distinct eigenvalues, and is therefore diagonalizable.

(b) Deduce that the set of matrices in $M_n(\mathbb{C})$ with a repeated eigenvalue contains no ball, that is, that its complement is dense.

(c) Is the set of diagonalizable matrices open? Justify your answer.

Solution.

(a) Put $2\varepsilon = \min_{i \neq j} |\lambda_i - \lambda_j|$, which is positive because the λ_i are distinct and there are finitely many pairs. Let δ be the number that Corollary 16.7.4 supplies for this ε. If $\|\mathbf{B} - \mathbf{A}\| < \delta$, number the eigenvalues of $\mathbf{B}$ so that $|\mu_i - \lambda_i| < \varepsilon$ for every i. For $i \neq j$, Theorem 0.6.4 gives

$$|\mu_i - \mu_j| \geq |\lambda_i - \lambda_j| - |\mu_i - \lambda_i| - |\mu_j - \lambda_j| > 2\varepsilon - \varepsilon - \varepsilon = 0,$$

so the μ_i are pairwise distinct. There are n of them, so $\mathbf{B}$ is diagonalizable by Corollary 9.4.2.

(b) Let $\mathbf{A}$ be arbitrary and $\varepsilon > 0$. By Exercise (C3: Distinct eigenvalues are dense, but not over the reals) (a) there is $\mathbf{A}'$ with n distinct eigenvalues and $|a_{ij} - a'_{ij}| < \varepsilon/C$ for all i, j, with C as in the proof of Corollary 16.7.5; hence $\|\mathbf{A} - \mathbf{A}'\| < \varepsilon$. So every ball meets the set of matrices with n distinct eigenvalues, and no ball is contained in its complement.

(c) No. The zero matrix is diagonalizable, and in $M_2(\mathbb{C})$ every ball around it contains $\varepsilon \mathbf{J}_2(0)$, whose only eigenvalue is 0, with eigenspace of dimension 1 and algebraic multiplicity 2; so it is not diagonalizable by Theorem 9.4.1 (e). Dense (Corollary 16.7.5) and open are independent properties, and the diagonalizable matrices are the first without being the second.

16.8 How Wrong Can the Answer Be

Chapter 3 solved $\mathbf{Ax} = \mathbf{b}$ and Chapter 11 solved it when it had no solution. Neither chapter could ask the question a user of either method asks first: the vector $\mathbf{b}$ is a measurement, so it is wrong in the last few digits, and the answer $\mathbf{x}$ is then wrong too — by how much? The question needs a way to compare the size of an error with the size of the thing it corrupts, and that is what the last seven sections have built.

The answer turns out to be one number attached to $\mathbf{A}$, independent of $\mathbf{b}$, and it is the last object this chapter defines. It also carries a warning, and the warning is more useful than the number.

Throughout, matrices are complex, as in Sections 4 to 7, and $\|\cdot\|$ is a norm on $\mathbb{C}^n$ together with the operator norm it induces on $M_n(\mathbb{C})$ (Definition 16.3.2) — the standing convention of Section 5, whose results are used here throughout. A real matrix is read inside $M_n(\mathbb{C})$, which changes none of the three norms we actually compute with, since Theorem 16.3.3 gives the same three formulas over either field.

16.8.1 A recurring product

Let $\mathbf{A} \in M_n(\mathbb{C})$ be invertible, let $\mathbf{b} \neq \mathbf{0}$, and let $\mathbf{x}$ solve $\mathbf{A}\mathbf{x} = \mathbf{b}$. Suppose the data arrives as $\mathbf{b} + \Delta\mathbf{b}$ instead, so the computed answer solves $\mathbf{A}\tilde{\mathbf{x}} = \mathbf{b} + \Delta\mathbf{b}$. Subtracting the two equations, $\mathbf{A}(\tilde{\mathbf{x}} - \mathbf{x}) = \Delta\mathbf{b}$, so

$$\tilde{\mathbf{x}} - \mathbf{x} = \mathbf{A}^{-1}\Delta\mathbf{b}, \qquad \|\tilde{\mathbf{x}} - \mathbf{x}\| \leq \|\mathbf{A}^{-1}\|\,\|\Delta\mathbf{b}\|$$

by Theorem 16.3.2 (a). That is an *absolute* bound, and absolute bounds are the wrong currency: whether an error of 10^{-3} is a catastrophe depends on whether the answer is of size 1 or of size 10^6. Divide by $\|\mathbf{x}\|$, and use $\|\mathbf{b}\| = \|\mathbf{A}\mathbf{x}\| \leq \|\mathbf{A}\|\|\mathbf{x}\|$, that is $1/\|\mathbf{x}\| \leq \|\mathbf{A}\|/\|\mathbf{b}\|$:

$$\frac{\|\tilde{\mathbf{x}} - \mathbf{x}\|}{\|\mathbf{x}\|} \;\leq\; \|\mathbf{A}^{-1}\|\,\|\Delta\mathbf{b}\|\,\frac{\|\mathbf{A}\|}{\|\mathbf{b}\|} \;=\; \|\mathbf{A}\|\|\mathbf{A}^{-1}\|\,\frac{\|\Delta\mathbf{b}\|}{\|\mathbf{b}\|}.$$

The relative error in the answer is at most the relative error in the data, multiplied by $\|\mathbf{A}\|\|\mathbf{A}^{-1}\|$. That product depends only on $\mathbf{A}$. It will reappear in every estimate in this section, so we name it.

The condition number is the worst factor by which $\mathbf{A}$ *can magnify a relative error.*

Definition 16.8.1. Condition Number

Let $\|\cdot\|$ be a norm on $\mathbb{C}^n$, and write $\|\cdot\|$ also for the operator norm it induces on $M_n(\mathbb{C})$. For $\mathbf{A} \in M_n(\mathbb{C})$, the **condition number** of $\mathbf{A}$ **with respect to that norm** is

$$\kappa(\mathbf{A}) := \begin{cases} \|\mathbf{A}\|\,\|\mathbf{A}^{-1}\| & \text{if } \mathbf{A} \text{ is invertible,} \\ \infty & \text{if } \mathbf{A} \text{ is singular.} \end{cases}$$

We write $\kappa_1, \kappa_2, \kappa_\infty$ for the condition numbers with respect to $\|\cdot\|_1$, $\|\cdot\|_2$, $\|\cdot\|_\infty$; κ_2 is also called the **spectral condition number**. For a rectangular $\mathbf{A} \in M_{m\times n}(\mathbb{C})$ with $m \geq n$ and $\operatorname{rank}\mathbf{A} = n$, we define

$$\kappa_2(\mathbf{A}) := \frac{\sigma_1(\mathbf{A})}{\sigma_n(\mathbf{A})},$$

the ratio of the largest to the smallest singular value (Definition 13.9.1).

Clause by clause. The number is attached to **A and to a norm**: $\kappa_1(\mathbf{A})$, $\kappa_2(\mathbf{A})$, and $\kappa_\infty(\mathbf{A})$ are three different numbers, and a statement like "$\kappa(\mathbf{A}) = 400$" is incomplete until the norm is named. They cannot disagree wildly — Theorem 16.2.2 traps them within dimension-dependent factors of each other, and Exercise (C1: Changing the norm changes κ, but not much) makes that precise — so the *order of magnitude* of κ, which is all anyone uses, does not depend on the choice. The value ∞ for a singular matrix is a convention, and the right one: a singular $\mathbf{A}$ either makes $\mathbf{A}\mathbf{x} = \mathbf{b}$ unsolvable or makes its solution set a whole coset $\mathbf{x} + \operatorname{nul}(\mathbf{A})$ with $\operatorname{nul}(\mathbf{A}) \neq \{\mathbf{0}\}$ (Theorem 3.1.3), so no finite bound on the error can hold. Finally, the square case and the rectangular case agree where they overlap: Proposition 16.8.2 (d) shows $\kappa_2(\mathbf{A}) = \sigma_1/\sigma_n$ for an invertible square $\mathbf{A}$ as well.

Before the examples, a small lemma that says what κ is measuring. It is the one geometric picture in this section, and it is worth carrying.

Lemma 16.8.1. The Inverse Measures the Smallest Stretch

Let $\mathbf{A} \in M_n(\mathbb{C})$ be invertible and let $\|\cdot\|$ be a norm on $\mathbb{C}^n$. Then

$$\|\mathbf{A}^{-1}\| = \frac{1}{\min\{\|\mathbf{Ax}\| : \|\mathbf{x}\| = 1\}}, \qquad \kappa(\mathbf{A}) = \frac{\max\{\|\mathbf{Ax}\| : \|\mathbf{x}\| = 1\}}{\min\{\|\mathbf{Ax}\| : \|\mathbf{x}\| = 1\}}.$$

Idea. Both maxima and the minimum exist by compactness, which is Lemma 16.3.1 and the same argument with a minimum. After that it is one substitution: as $\mathbf{x}$ runs over the nonzero vectors so does $\mathbf{y} = \mathbf{Ax}$, and the ratio $\|\mathbf{A}^{-1}\mathbf{y}\|/\|\mathbf{y}\|$ is the reciprocal of $\|\mathbf{Ax}\|/\|\mathbf{x}\|$.

Proof. The function $\mathbf{x} \mapsto \|\mathbf{Ax}\|$ is continuous on the unit sphere S, which is compact, so it attains a maximum and a minimum there (Lemma 16.3.1 for the maximum; the same proof, or the extreme value theorem (A4) applied to the same continuous function, for the minimum). Write $m = \min_{\mathbf{x}\in S}\|\mathbf{Ax}\|$. Since $\mathbf{A}$ is invertible, $\mathbf{Ax} \neq \mathbf{0}$ for $\mathbf{x} \in S$, so $m > 0$.

By Definition 16.3.2 and homogeneity, $\|\mathbf{A}^{-1}\| = \max\{\|\mathbf{A}^{-1}\mathbf{y}\|/\|\mathbf{y}\| : \mathbf{y} \neq \mathbf{0}\}$. The map $\mathbf{y} \mapsto \mathbf{A}^{-1}\mathbf{y}$ is a bijection of $\mathbb{C}^n \setminus \{\mathbf{0}\}$ onto itself, so substituting $\mathbf{y} = \mathbf{Ax}$ with $\mathbf{x} \neq \mathbf{0}$,

$$\|\mathbf{A}^{-1}\| = \max_{\mathbf{x}\neq\mathbf{0}} \frac{\|\mathbf{x}\|}{\|\mathbf{Ax}\|} = \frac{1}{\min\limits_{\mathbf{x}\neq\mathbf{0}} \dfrac{\|\mathbf{Ax}\|}{\|\mathbf{x}\|}} = \frac{1}{m},$$

the last step by homogeneity again, which lets the minimum over $\mathbf{x} \neq \mathbf{0}$ be taken over S. Multiplying by $\|\mathbf{A}\| = \max_{\mathbf{x}\in S}\|\mathbf{Ax}\|$ gives the second formula. This proves the lemma. ∎

So $\kappa(\mathbf{A})$ is the **distortion** of $\mathbf{A}$: feed $\mathbf{A}$ the unit sphere and it comes out as a lopsided surface, and κ is the ratio of its longest radius to its shortest. A matrix with $\kappa = 1$ carries the sphere to a sphere. A matrix with $\kappa = 10^6$ flattens it into something a million times longer than it is wide, and a direction of $\mathbf{b}$ that lands along the short axis has its errors magnified by a million on the way back.

Some values, simplest first.

- **The identity, and any unitary matrix.** $\|\mathbf{I}\| = 1 = \|\mathbf{I}^{-1}\|$, so $\kappa(\mathbf{I}) = 1$ in every norm. For unitary $\mathbf{U}$, $\|\mathbf{Ux}\|_2 = \|\mathbf{x}\|_2$ for every $\mathbf{x}$ (Theorem 11.7.3), so every stretch is 1 and $\kappa_2(\mathbf{U}) = 1$ by Lemma 16.8.1. These are the best-conditioned matrices there are.
- **A diagonal matrix.** For $\mathbf{D} = \operatorname{diag}(d_1, \dots, d_n)$ with all $d_i \neq 0$, $\|\mathbf{D}\|_2 = \max_i |d_i|$ and $\|\mathbf{D}^{-1}\|_2 = 1/\min_i |d_i|$, so $\kappa_2(\mathbf{D})$ is the ratio of the largest to the smallest modulus. Thus $\kappa_2(\operatorname{diag}(1, 10^{-6})) = 10^6$.
- **A small matrix that is perfectly conditioned.** $\mathbf{A} = 10^{-6}\mathbf{I}$ has $\|\mathbf{A}\|_2 = 10^{-6}$, a tiny norm, and $\kappa_2(\mathbf{A}) = 1$. Being small and being ill-conditioned are unrelated.
- **A singular matrix**, the degenerate case: $\kappa = \infty$ by convention, consistent with Lemma 16.8.1, where the minimum stretch is 0.

Example. Three condition numbers of one matrix

Compute $\kappa_\infty(\mathbf{A})$ and $\kappa_2(\mathbf{A})$ for $\mathbf{A} = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$.

Solution. $\det \mathbf{A} = 1$, so $\mathbf{A}^{-1} = \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix}$. The largest absolute row sums are 3 for $\mathbf{A}$ and 3 for $\mathbf{A}^{-1}$, so by Theorem 16.3.3 (b), $\kappa_\infty(\mathbf{A}) = 3 \cdot 3 = 9$.

For κ_2: $\mathbf{A}$ is real symmetric with $\operatorname{tr} \mathbf{A} = 3$ and $\det \mathbf{A} = 1$, so its eigenvalues are $(3 \pm \sqrt{5})/2$, both positive, and the singular values of a symmetric matrix with positive eigenvalues are those eigenvalues (its $\mathbf{A}^*\mathbf{A} = \mathbf{A}^2$ has eigenvalues their squares). Hence

$$\kappa_2(\mathbf{A}) = \frac{3 + \sqrt{5}}{3 - \sqrt{5}} = \frac{(3 + \sqrt{5})^2}{4} = \frac{7 + 3\sqrt{5}}{2} \approx 6.854.$$

Two different norms, two different numbers, the same order of magnitude — which is the general picture.

Warning. **A large κ is not a large determinant defect, and a small determinant is not ill-conditioning.** The matrix $10^{-6}\mathbf{I}_{10}$ has determinant 10^{-60} and $\kappa_2 = 1$; the matrix $\operatorname{diag}(10^6, 10^{-6})$ has determinant 1 and $\kappa_2 = 10^{12}$. The determinant scales with c^n under $\mathbf{A} \mapsto c\mathbf{A}$ while κ does not move at all (Proposition 16.8.2 (b)), so no threshold on $|\det \mathbf{A}|$ can detect ill-conditioning. This is the most common wrong test, and it fails in both directions.

The basic properties are all one line each, and the third is the reason κ behaves well under the factorizations of Chapters 11 and 13.

Proposition 16.8.2. Properties of the Condition Number

Let $\mathbf{A}, \mathbf{B} \in M_n(\mathbb{C})$ be invertible and let $c \in \mathbb{C} \setminus \{0\}$. With respect to any operator norm:

(a) $\kappa(\mathbf{A}) \geq 1$;
(b) $\kappa(c\mathbf{A}) = \kappa(\mathbf{A})$;
(c) $\kappa(\mathbf{AB}) \leq \kappa(\mathbf{A})\kappa(\mathbf{B})$;
(d) $\kappa_2(\mathbf{A}) = \sigma_1(\mathbf{A})/\sigma_n(\mathbf{A})$, and $\kappa_2(\mathbf{UAV}) = \kappa_2(\mathbf{A})$ for all unitary $\mathbf{U}, \mathbf{V}$.

Proof.

(a) $1 = \|\mathbf{I}\| = \|\mathbf{A}\mathbf{A}^{-1}\| \leq \|\mathbf{A}\|\|\mathbf{A}^{-1}\| = \kappa(\mathbf{A})$, using Theorem 16.3.2 (c) and (d).

(b) $(c\mathbf{A})^{-1} = c^{-1}\mathbf{A}^{-1}$, so by homogeneity $\|c\mathbf{A}\|\|(c\mathbf{A})^{-1}\| = |c|\,|c|^{-1}\|\mathbf{A}\|\|\mathbf{A}^{-1}\|$.

(c) $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$, so $\|\mathbf{AB}\|\|\mathbf{B}^{-1}\mathbf{A}^{-1}\| \leq \|\mathbf{A}\|\|\mathbf{B}\|\|\mathbf{B}^{-1}\|\|\mathbf{A}^{-1}\|$ by submultiplicativity (Theorem 16.3.2 (d)).

(d) By Theorem 16.3.3 (c), $\|\mathbf{A}\|_2 = \sigma_1(\mathbf{A})$. By Lemma 16.8.1, $\|\mathbf{A}^{-1}\|_2 = 1/m$ with $m = \min\{\|\mathbf{Ax}\|_2 : \|\mathbf{x}\|_2 = 1\}$, so it remains to see $m = \sigma_n(\mathbf{A})$. Write $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$ (Theorem 13.9.1) and $\mathbf{y} = \mathbf{V}^*\mathbf{x}$; unitary matrices preserve $\|\cdot\|_2$ (Theorem 11.7.3), so $\|\mathbf{x}\|_2 = \|\mathbf{y}\|_2$ and $\|\mathbf{Ax}\|_2 = \|\mathbf{\Sigma y}\|_2$. Then

$$\|\mathbf{\Sigma y}\|_2^2 = \sum_{i=1}^{n} \sigma_i^2 |y_i|^2 \geq \sigma_n^2 \sum_{i=1}^{n} |y_i|^2 = \sigma_n^2$$

for $\|\mathbf{y}\|_2 = 1$, with equality at $\mathbf{y} = \mathbf{e}_n$, that is at $\mathbf{x} = \mathbf{v}_n$, the last column of $\mathbf{V}$. So

$m = \sigma_n$. For the last claim, if $\mathbf{A} = \mathbf{U}_1\mathbf{\Sigma}\mathbf{V}_1^*$ then $\mathbf{UAV} = (\mathbf{UU}_1)\mathbf{\Sigma}(\mathbf{V}^*\mathbf{V}_1)^*$, and a product of unitary matrices is unitary, so this is a singular value decomposition of $\mathbf{UAV}$ with the same $\mathbf{\Sigma}$: unitary factors do not change the singular values. This proves the proposition.

■

Part (d) is the reason κ_2 is the condition number of choice in theory: it is invariant under the unitary changes of coordinates that Chapters 11 to 13 use throughout, so it is a property of the linear map and not of the basis it is written in.

16.8.2 What the condition number predicts

The computation that opened the section is now a theorem, together with the version where $\mathbf{A}$ itself, and not only $\mathbf{b}$, carries the error. Both halves matter in practice: the entries of $\mathbf{A}$ are measurements too. Section 5 already proved the two perturbations together, in Theorem 16.5.3, and its constant $\|\mathbf{A}\|\|\mathbf{A}^{-1}\|$ is the product we have just named; what is added here is that the first half is sharp on its own, and that it needs no smallness hypothesis at all.

Theorem 16.8.3. Relative Error Under Perturbation of the Data

Let $\mathbf{A} \in M_n(\mathbb{C})$ be invertible, let $\mathbf{b} \neq \mathbf{0}$, and let $\mathbf{Ax} = \mathbf{b}$.

(a) **(Perturbing b.)** If $\mathbf{A}\tilde{\mathbf{x}} = \mathbf{b} + \Delta\mathbf{b}$, then

$$\frac{\|\tilde{\mathbf{x}} - \mathbf{x}\|}{\|\mathbf{x}\|} \leq \kappa(\mathbf{A}) \frac{\|\Delta\mathbf{b}\|}{\|\mathbf{b}\|},$$

and there are $\mathbf{b}$ and $\Delta\mathbf{b}$ for which equality holds.

(b) **(Perturbing A.)** Let $\mathbf{E} \in M_n(\mathbb{C})$ with $r := \kappa(\mathbf{A})\|\mathbf{E}\|/\|\mathbf{A}\| < 1$. Then $\mathbf{A} + \mathbf{E}$ is invertible; and if $(\mathbf{A} + \mathbf{E})\tilde{\mathbf{x}} = \mathbf{b}$, then

$$\frac{\|\tilde{\mathbf{x}} - \mathbf{x}\|}{\|\tilde{\mathbf{x}}\|} \leq \kappa(\mathbf{A}) \frac{\|\mathbf{E}\|}{\|\mathbf{A}\|}, \qquad \frac{\|\tilde{\mathbf{x}} - \mathbf{x}\|}{\|\mathbf{x}\|} \leq \frac{r}{1 - r}.$$

Idea. For (a), the displayed computation at the head of the section, plus a choice of the two vectors that makes both inequalities in it equalities — and both are attained, because Lemma 16.3.1 says an operator norm is a maximum and not a supremum. For (b), factor $\mathbf{A} + \mathbf{E} = \mathbf{A}(\mathbf{I} + \mathbf{A}^{-1}\mathbf{E})$ so that Theorem 16.5.1 decides invertibility, then subtract the two equations. Subtracting gives $\tilde{\mathbf{x}} - \mathbf{x}$ in terms of $\tilde{\mathbf{x}}$, which is why the clean bound is relative to $\tilde{\mathbf{x}}$; subadditivity of the norm converts it to a bound relative to $\mathbf{x}$ at the cost of the factor $1/(1 - r)$, and that factor is where "to first order" hides.

Proof. (a). Subtracting $\mathbf{Ax} = \mathbf{b}$ from $\mathbf{A}\tilde{\mathbf{x}} = \mathbf{b} + \Delta\mathbf{b}$ gives $\mathbf{A}(\tilde{\mathbf{x}} - \mathbf{x}) = \Delta\mathbf{b}$, so $\tilde{\mathbf{x}} - \mathbf{x} = \mathbf{A}^{-1}\Delta\mathbf{b}$ and $\|\tilde{\mathbf{x}} - \mathbf{x}\| \leq \|\mathbf{A}^{-1}\|\|\Delta\mathbf{b}\|$ by Theorem 16.3.2 (a). Also $\|\mathbf{b}\| = \|\mathbf{Ax}\| \leq \|\mathbf{A}\|\|\mathbf{x}\|$, and $\mathbf{x} \neq \mathbf{0}$ because $\mathbf{b} \neq \mathbf{0}$, so $1/\|\mathbf{x}\| \leq \|\mathbf{A}\|/\|\mathbf{b}\|$. Multiplying the two estimates,

$$\frac{\|\tilde{\mathbf{x}} - \mathbf{x}\|}{\|\mathbf{x}\|} \leq \|\mathbf{A}^{-1}\|\|\Delta\mathbf{b}\| \cdot \frac{\|\mathbf{A}\|}{\|\mathbf{b}\|} = \kappa(\mathbf{A}) \frac{\|\Delta\mathbf{b}\|}{\|\mathbf{b}\|}.$$

For equality, use Lemma 16.3.1 twice. Choose a unit vector $\mathbf{x}$ with $\|\mathbf{A}\mathbf{x}\| = \|\mathbf{A}\|$ and put $\mathbf{b} = \mathbf{A}\mathbf{x}$, so that $\|\mathbf{b}\| = \|\mathbf{A}\|\|\mathbf{x}\|$ and the second estimate is an equality. Choose a unit vector $\mathbf{y}$ with $\|\mathbf{A}^{-1}\mathbf{y}\| = \|\mathbf{A}^{-1}\|$ and put $\Delta\mathbf{b} = \mathbf{y}$, so that the first estimate is an equality. Then

$$\frac{\|\widetilde{\mathbf{x}} - \mathbf{x}\|}{\|\mathbf{x}\|} \Big/ \frac{\|\Delta\mathbf{b}\|}{\|\mathbf{b}\|} = \frac{\|\mathbf{A}^{-1}\|}{1} \cdot \frac{\|\mathbf{A}\|}{1} = \kappa(\mathbf{A}).$$

(b). Write $\mathbf{A} + \mathbf{E} = \mathbf{A}(\mathbf{I} + \mathbf{A}^{-1}\mathbf{E})$. By Theorem 16.3.2 (d),

$$\|\mathbf{A}^{-1}\mathbf{E}\| \le \|\mathbf{A}^{-1}\|\|\mathbf{E}\| = \kappa(\mathbf{A})\frac{\|\mathbf{E}\|}{\|\mathbf{A}\|} = r < 1,$$

so $\mathbf{I} - (-\mathbf{A}^{-1}\mathbf{E})$ is invertible by Theorem 16.5.1, and $\mathbf{A} + \mathbf{E}$ is a product of two invertible matrices, hence invertible. (This is the openness of $\mathrm{GL}_n(F)$ recorded in Corollary 16.5.2, with the radius made explicit.)

Since $\mathbf{b} \neq \mathbf{0}$ we have $\widetilde{\mathbf{x}} = (\mathbf{A} + \mathbf{E})^{-1}\mathbf{b} \neq \mathbf{0}$. Subtracting $\mathbf{A}\mathbf{x} = \mathbf{b}$ from $(\mathbf{A} + \mathbf{E})\widetilde{\mathbf{x}} = \mathbf{b}$ gives $\mathbf{A}\widetilde{\mathbf{x}} - \mathbf{A}\mathbf{x} = -\mathbf{E}\widetilde{\mathbf{x}}$, hence

$$\widetilde{\mathbf{x}} - \mathbf{x} = -\mathbf{A}^{-1}\mathbf{E}\widetilde{\mathbf{x}}, \qquad \|\widetilde{\mathbf{x}} - \mathbf{x}\| \le \|\mathbf{A}^{-1}\|\|\mathbf{E}\|\,\|\widetilde{\mathbf{x}}\| = r\,\|\widetilde{\mathbf{x}}\|,$$

which is the first bound. For the second, $\|\widetilde{\mathbf{x}}\| \le \|\mathbf{x}\| + \|\widetilde{\mathbf{x}} - \mathbf{x}\|$ since $\|\cdot\|$ is subadditive (Definition 16.1.1), so $\|\widetilde{\mathbf{x}} - \mathbf{x}\| \le r\|\mathbf{x}\| + r\|\widetilde{\mathbf{x}} - \mathbf{x}\|$; as $r < 1$ we may divide by $1 - r > 0$ to get $\|\widetilde{\mathbf{x}} - \mathbf{x}\| \le \frac{r}{1-r}\|\mathbf{x}\|$. This proves the theorem.

■

The two bounds in (b) say the same thing differently. The first is exact and needs nothing but invertibility, but it measures the error against the *computed* answer. The second measures it against the true answer, at the cost of the factor $1/(1 - r)$ — which is $1 + r + r^2 + \cdots$, so for a small r it is 1 to first order and the two bounds agree. That is all "to first order in the perturbation of $\mathbf{A}$" ever means here, and it is now a stated inequality rather than a gesture. Adding the two halves, with the factor $1/(1-r)$ covering both, recovers Theorem 16.5.3 exactly.

Quick check. Where in the proof of (a) is the hypothesis $\mathbf{b} \neq \mathbf{0}$ used, and what goes wrong without it?

Solution. It is used to know $\mathbf{x} \neq \mathbf{0}$, so that $\|\mathbf{x}\|$ may be divided by, and to know $\|\mathbf{b}\| \neq 0$, so that the relative perturbation $\|\Delta\mathbf{b}\|/\|\mathbf{b}\|$ means anything. With $\mathbf{b} = \mathbf{0}$ the true solution is $\mathbf{x} = \mathbf{0}$ and *every* nonzero $\Delta\mathbf{b}$ produces an infinite relative error. Relative error is only defined against a nonzero reference.

Now a case where the bound of (a) is not merely an inequality. The point of computing the example exactly is to see that κ is not a pessimistic overestimate manufactured by the inequalities used to reach it: it is achieved.

Example. A system where the bound is attained

Let

$$\mathbf{A} = \begin{pmatrix} 1 & 1 \\ 1 & 1.01 \end{pmatrix}, \qquad \mathbf{b} = \begin{pmatrix} 2 \\ 2.01 \end{pmatrix},$$

whose solution is $\mathbf{x} = (1, 1)$. Compute $\kappa_\infty(\mathbf{A})$; then perturb $\mathbf{b}$ to $\mathbf{b} + \Delta\mathbf{b}$ with $\Delta\mathbf{b} = (0.01, -0.01)$, solve again, and compare the two relative errors.

Solution. Work with exact fractions: $1.01 = \frac{101}{100}$ and $\det \mathbf{A} = \frac{101}{100} - 1 = \frac{1}{100}$, so

$$\mathbf{A}^{-1} = 100 \begin{pmatrix} \frac{101}{100} & -1 \\ -1 & 1 \end{pmatrix} = \begin{pmatrix} 101 & -100 \\ -100 & 100 \end{pmatrix}.$$

The largest absolute row sums are $\frac{201}{100}$ for $\mathbf{A}$ and 201 for $\mathbf{A}^{-1}$, so by Theorem 16.3.3 (b),

$$\kappa_\infty(\mathbf{A}) = \tfrac{201}{100} \cdot 201 = \tfrac{40401}{100} = 404.01.$$

Check that $\mathbf{A}\mathbf{x} = \mathbf{b}$: the two rows give $1 + 1 = 2$ and $1 + \frac{101}{100} = \frac{201}{100}$. Now

$$\widetilde{\mathbf{x}} - \mathbf{x} = \mathbf{A}^{-1}\Delta\mathbf{b} = \tfrac{1}{100} \begin{pmatrix} 101 + 100 \\ -100 - 100 \end{pmatrix} = \begin{pmatrix} 2.01 \\ -2 \end{pmatrix},$$

so $\widetilde{\mathbf{x}} = (3.01, -1)$. Verify directly: $3.01 - 1 = 2.01$ and $3.01 - 1.01 = 2$, which is $\mathbf{b} + \Delta\mathbf{b} = (2.01, 2)$.

The two relative errors, in $\|\cdot\|_\infty$:

$$\frac{\|\Delta\mathbf{b}\|_\infty}{\|\mathbf{b}\|_\infty} = \frac{1/100}{201/100} = \frac{1}{201} \approx 0.50\%,$$
$$\frac{\|\widetilde{\mathbf{x}} - \mathbf{x}\|_\infty}{\|\mathbf{x}\|_\infty} = \frac{201/100}{1} = \frac{201}{100} = 201\%.$$

Their ratio is $\frac{201}{100} \cdot 201 = 404.01$, which is exactly $\kappa_\infty(\mathbf{A})$. A perturbation of half a percent in the data has turned the answer $(1, 1)$ into $(3.01, -1)$, and the bound of Theorem 16.8.3 (a) is attained, not approached.

Geometrically the example is Lemma 16.8.1 at work: the two rows of $\mathbf{A}$ are nearly parallel, so the two lines whose intersection is $\mathbf{x}$ cross at a very small angle, and moving either line slightly slides the crossing point a long way. Ill-conditioning is that picture, in any dimension.

Warning. **The condition number bounds the error; it does not produce it.** Theorem 16.8.3 is a worst case over all perturbations $\Delta\mathbf{b}$ of a given size. A particular $\Delta\mathbf{b}$ may do far less damage: in the example above, $\Delta\mathbf{b} = (0.01, 0.01)$ gives $\widetilde{\mathbf{x}} - \mathbf{x} = \mathbf{A}^{-1}\Delta\mathbf{b} = (0.01, 0)$, a relative error of 1% against a relative perturbation of 0.5% — a magnification of 2, not 404. A large κ says that *some* direction is dangerous, not that yours is.

16.8.3 Why the normal equations are worse

Chapter 11 §04 solved the least-squares problem through the normal equations $\mathbf{A}^*\mathbf{A}\mathbf{x} = \mathbf{A}^*\mathbf{b}$ (Theorem 11.4.2), and Chapter 11 §08 then recommended against actually forming that system. Its words were: "The deeper reason to prefer it is not the operation count: forming $\mathbf{A}^*\mathbf{A}$ squares the sensitivity of the problem to rounding, while (†) never forms that product at all. Chapter 24 makes 'sensitivity' precise and proves the comparison; here we record only the practice." The word "sensitivity" now has a meaning, and the word "squares" is now a theorem.

Theorem 16.8.4. The Normal Equations Square the Condition Number

Let $m \geq n \geq 1$ and let $\mathbf{A} \in M_{m\times n}(\mathbb{C})$ have rank n. Then $\mathbf{A}^*\mathbf{A} \in M_n(\mathbb{C})$ is invertible, and

$$\kappa_2(\mathbf{A}^*\mathbf{A}) = \kappa_2(\mathbf{A})^2.$$

Idea. An SVD of $\mathbf{A}$ is also, after one multiplication, an eigenvalue decomposition of $\mathbf{A}^*\mathbf{A}$: the singular values get squared and nothing else happens. Both condition numbers are then ratios of the same list of numbers, one list being the squares of the other.

Proof. Write $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$ with $\mathbf{U} \in M_m(\mathbb{C})$, $\mathbf{V} \in M_n(\mathbb{C})$ unitary and $\mathbf{\Sigma} \in M_{m\times n}(\mathbb{R})$ carrying $\sigma_1 \geq \cdots \geq \sigma_n \geq 0$ (Theorem 13.9.1). Since $\operatorname{rank}\mathbf{A} = n$, exactly n singular values are nonzero (Theorem 13.9.1), so $\sigma_n > 0$. Using $\mathbf{U}^*\mathbf{U} = \mathbf{I}_m$,

$$\mathbf{A}^*\mathbf{A} = \mathbf{V}\mathbf{\Sigma}^*\mathbf{U}^*\mathbf{U}\mathbf{\Sigma}\mathbf{V}^* = \mathbf{V}(\mathbf{\Sigma}^*\mathbf{\Sigma})\mathbf{V}^*, \qquad \mathbf{\Sigma}^*\mathbf{\Sigma} = \operatorname{diag}(\sigma_1^2, \dots, \sigma_n^2).$$

So $\mathbf{A}^*\mathbf{A}$ is unitarily similar to a diagonal matrix with the positive entries σ_i^2; it is therefore invertible, Hermitian, and positive definite, with eigenvalues $\sigma_1^2 \geq \cdots \geq \sigma_n^2 > 0$.

The singular values of $\mathbf{A}^*\mathbf{A}$ are those same numbers. Indeed the displayed identity *is* a singular value decomposition of $\mathbf{A}^*\mathbf{A}$, with both unitary factors equal to $\mathbf{V}$ and the diagonal entries already in decreasing order; alternatively, $(\mathbf{A}^*\mathbf{A})^*(\mathbf{A}^*\mathbf{A}) = \mathbf{V}\operatorname{diag}(\sigma_i^4)\mathbf{V}^*$ has eigenvalues σ_i^4, whose nonnegative square roots are the σ_i^2 (Definition 13.9.1). Hence by Definition 16.8.1 and Proposition 16.8.2 (d),

$$\kappa_2(\mathbf{A}^*\mathbf{A}) = \frac{\sigma_1^2}{\sigma_n^2} = \left(\frac{\sigma_1}{\sigma_n}\right)^2 = \kappa_2(\mathbf{A})^2.$$

This proves the theorem. ∎

Example. Squaring, on a 3×2 example

Let $\mathbf{A} = \begin{pmatrix} 1 & 1 \\ 0.1 & 0 \\ 0 & 0.1 \end{pmatrix}$. Compute $\kappa_2(\mathbf{A})$ and $\kappa_2(\mathbf{A}^\top\mathbf{A})$.

Solution. $\mathbf{A}^\top\mathbf{A} = \left(\begin{smallmatrix} 1.01 & 1 \\ 1 & 1.01 \end{smallmatrix}\right)$. It sends $(1,1)$ to $2.01\,(1,1)$ and $(1,-1)$ to $0.01\,(1,-1)$, so its eigenvalues are $\frac{201}{100}$ and $\frac{1}{100}$. Hence $\sigma_1(\mathbf{A}) = \sqrt{201}/10$ and $\sigma_2(\mathbf{A}) = 1/10$, so

$$\kappa_2(\mathbf{A}) = \sqrt{201} \approx 14.18, \qquad \kappa_2(\mathbf{A}^\top\mathbf{A}) = \frac{201/100}{1/100} = 201 = \left(\sqrt{201}\right)^2.$$

A matrix whose conditioning is unremarkable becomes, on forming $\mathbf{A}^\top\mathbf{A}$, a matrix whose conditioning is fourteen times worse.

What is proved here, and what is not. Theorem 16.8.4 is a statement about matrices: the matrix one *forms* when one writes down the normal equations has the square of the condition number of the matrix one started with. Combined with Theorem 16.8.3, it says that the linear system $\mathbf{A}^*\mathbf{A}\mathbf{x} = \mathbf{A}^*\mathbf{b}$, considered as a problem in its own data, is more sensitive than $\mathbf{A}$ itself by a factor $\kappa_2(\mathbf{A})$ — and the $\mathbf{QR}$ route $\mathbf{R}\mathbf{x} = \mathbf{Q}^*\mathbf{b}$ of Chapter 11 §08 never forms $\mathbf{A}^*\mathbf{A}$, so it never incurs that factor. Indeed the triangular matrix it does solve with is no worse conditioned than $\mathbf{A}$: from $\mathbf{A} = \mathbf{QR}$ with $\mathbf{Q}^*\mathbf{Q} = \mathbf{I}_n$ (Theorem 11.8.1) we get $\mathbf{A}^*\mathbf{A} = \mathbf{R}^*\mathbf{Q}^*\mathbf{Q}\mathbf{R} = \mathbf{R}^*\mathbf{R}$, so $\mathbf{A}$ and $\mathbf{R}$ have the same singular values (Definition 13.9.1) and $\kappa_2(\mathbf{R}) = \kappa_2(\mathbf{A})$ exactly.

Three things are *not* proved here, and all three belong to Chapter 24.

- **The conditioning of the least-squares problem itself** is not $\kappa_2(\mathbf{A})$. The map $\mathbf{b} \mapsto \mathbf{x}^+$ from data to least-squares solution has a sensitivity that depends on $\kappa_2(\mathbf{A})$, on the residual $\|\mathbf{A}\mathbf{x}^+ - \mathbf{b}\|$, and on the angle between $\mathbf{b}$ and $\operatorname{col}(\mathbf{A})$; when the residual is large, a term in $\kappa_2(\mathbf{A})^2$ appears even for the QR route. The bound above is about the *linear systems being solved*, not about the underlying fitting problem.
- **Nothing here is about arithmetic.** A condition number describes a problem, and an error bound of the form "relative error $\leq \kappa\cdot$ (relative perturbation)" only becomes a statement about a computation once one knows how large a perturbation the computation itself introduces. That requires a model of floating-point arithmetic and a backward error analysis, neither of which exists in this book yet.
- **Hence the comparison of the two algorithms** — that a stable implementation of the QR route loses about $\log_{10}\kappa_2(\mathbf{A})$ decimal digits while forming and solving the normal equations loses about $2\log_{10}\kappa_2(\mathbf{A})$ — is still Chapter 24's to prove. What this section supplies is the number that appears in both statements, and the reason a square appears in one of them.

16.8.4 Conditioning is a property of the problem

This is the last warning of the chapter, and it is the one readers most often need.

Warning. **Conditioning belongs to the problem; stability belongs to the algorithm.** $\kappa(\mathbf{A})$ is computed from $\mathbf{A}$ alone. It knows nothing about how anyone intends to solve $\mathbf{A}\mathbf{x} = \mathbf{b}$, and it does not change if the method changes. Two consequences, both regularly got wrong:

- **A large $\kappa(\mathbf{A})$ does not mean a computed answer is wrong.** It means that the answer is not determined to high relative accuracy by data of limited accuracy — a statement about the information available, not about the arithmetic performed. If the data are exact and the arithmetic is exact, the answer is exact however large κ is. And even in inexact arithmetic, a good algorithm returns the exact answer to a problem

whose data differ from yours by a rounding error, which is the most anyone can ask of it.

- **A small $\kappa(\mathbf{A})$ does not make an algorithm trustworthy.** An unstable method can wreck a perfectly conditioned problem all by itself, by introducing errors far larger than the data warrant; Exercise (C3: A well-conditioned problem and a bad method) gives an elementary example in which κ_∞ is below 5 and a computed coordinate comes out as 0 instead of 1.

Chapter 24 draws the distinction properly, defines backward stability, and proves which of the algorithms in this book have it. Until then, read κ as an upper bound on what any method can promise, and never as a verdict on a particular answer.

16.8.5 What this chapter has settled

Nine chapters wrote promissory notes that only a norm could pay. Here is the ledger, with the section that closed each entry.

- **Chapter 11 §01** said that not every norm comes from an inner product and that the parallelogram law is the test. Section 1 proved it, in Theorem 16.1.3, building on the $\mathbb{R}^2$ case of Exercise (C1: The parallelogram law characterizes inner-product norms).
- **Chapter 10 §10** promised that this chapter would prove $\rho(\mathbf{A}) \le \|\mathbf{A}\|$ for the norm induced by any vector norm, and would recover $\rho(\mathbf{A})$ as a limit built from the norms of the powers. Section 4 did both: Theorem 16.4.1 and Gelfand's formula Theorem 16.4.4. The same section also gave ρ the labeled definition Definition 16.4.1 that Chapter 10 used only in prose, and Corollary 16.4.5 restored Chapter 10's convergence criterion with a rate attached. And Chapter 10's entrywise convergence turned out to be *the* convergence, by Corollary 16.2.4.
- **Chapter 13 §09 and §13** asked, twice, for a name for $\max\{\|\mathbf{Ax}\| : \|\mathbf{x}\| = 1\}$. Section 3 supplied it — Definition 16.3.2 — and separated the name from the computation $\|\mathbf{A}\|_2 = \sigma_1(\mathbf{A})$, which is Theorem 16.3.3 (c); Example (The oblique projection, finally measured) measured Chapter 13 §13's oblique projection. The spectral norm that Chapter 13 §11 needed in order even to *state* the second half of its approximation theorem now exists; the theorem itself waits for Chapter 17, which proves it from the min–max description of the singular values. The version covering every unitarily invariant norm at once is Chapter 21's.
- **Chapter 10 §09** promised a return to $e^{\mathbf{A}}$ with norms available, giving quantitative bounds in place of exact formulas. Section 6 gave them, in Theorem 16.6.1, and its Theorem 16.6.3 is the statement Chapter 12 §10 pointed at when it said that the consequence of stability for $\dot{\mathbf{x}} = \mathbf{Ax}$ belonged to Chapter 16.
- **Chapter 12 §01** said that this chapter, "where matrices acquire norms and limits can be spoken of properly, runs arguments of exactly this shape". Section 7 ran one: Corollary 16.7.5 and Corollary 16.7.6 turn "prove it for diagonalizable matrices, then take a limit" into three lines, and Example (Cayley–Hamilton, by a limit) is the demonstration.
- **Chapter 11 §08** deferred its reason for preferring **QR** to the normal equations. Theorem 16.8.4 supplies the matrix half of it; the arithmetic half is Chapter 24's, as listed above.

What the chapter has *not* done is equally worth recording. It has not bounded how far an individual eigenvalue moves under a perturbation — Example (A Jordan block's eigenvalues move like $\varepsilon^{1/k}$) shows there is no bound of the obvious kind, and the bounds that do exist, Weyl's and Bauer–Fike's, are Chapters 17 and 20. It has not said which norms are invariant

under unitary multiplication, which is Chapter 21. And it has not touched arithmetic. The condition number is the bridge to that last subject: it is the exact point at which linear algebra stops and numerical analysis begins.

16.8.6 Exercises

16.8.6.1 A. Check your understanding

Exercise. A1

(a) Define $\kappa(\mathbf{A})$, naming what it depends on besides $\mathbf{A}$.

(b) Explain why $\kappa(\mathbf{A}) \geq 1$ for every invertible $\mathbf{A}$.

(c) Decide whether each statement is correct, and justify: (i) "if $\kappa(\mathbf{A}) = 10^8$ then any computed solution of $\mathbf{Ax} = \mathbf{b}$ has at most a few correct digits"; (ii) "if $\kappa(\mathbf{A})$ is small then any method of solving $\mathbf{Ax} = \mathbf{b}$ is reliable".

(d) State what $\kappa_2(\mathbf{A}^*\mathbf{A})$ is in terms of $\kappa_2(\mathbf{A})$, and say what that has to do with the **QR** factorization.

Solution.

(a) See Definition 16.8.1: $\kappa(\mathbf{A}) = \|\mathbf{A}\|\|\mathbf{A}^{-1}\|$, with $\kappa(\mathbf{A}) = \infty$ for singular $\mathbf{A}$. Besides $\mathbf{A}$ it depends on the **choice of vector norm**, through the operator norm it induces; κ_1, κ_2, and κ_∞ are three different numbers.

(b) $1 = \|\mathbf{I}\| = \|\mathbf{A}\mathbf{A}^{-1}\| \leq \|\mathbf{A}\|\|\mathbf{A}^{-1}\|$, by Theorem 16.3.2 (c) and (d). See Proposition 16.8.2 (a).

(c) (i) Incorrect. κ bounds the effect of a perturbation of the data; with exact data and exact arithmetic the answer is exact. It says that the answer *cannot be determined* to high relative accuracy from inaccurate data, which is a statement about the problem, not a verdict on a computed answer. (ii) Incorrect. Conditioning is a property of the problem and stability a property of the algorithm; a small κ leaves an unstable method free to introduce errors of its own.

(d) $\kappa_2(\mathbf{A}^*\mathbf{A}) = \kappa_2(\mathbf{A})^2$ for $\mathbf{A}$ of full column rank (Theorem 16.8.4). The normal equations form $\mathbf{A}^*\mathbf{A}$ and so solve a system with the squared condition number, while the route $\mathbf{Rx} = \mathbf{Q}^*\mathbf{b}$ of Chapter 11 §08 never forms that product and keeps $\kappa_2(\mathbf{R}) = \kappa_2(\mathbf{A})$.

16.8.6.2 B. Practice

Exercise. B1: Three condition numbers

For $\mathbf{A} = \begin{pmatrix} 3 & 0 \\ 0 & -\frac{1}{2} \end{pmatrix}$ and $\mathbf{B} = \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix}$, compute κ_∞ and κ_2 of each. *Hint:* $\mathbf{B}^*\mathbf{B}$ *has trace* 6 *and determinant* 1.

Solution. $\mathbf{A}^{-1} = \operatorname{diag}(\frac{1}{3}, -2)$. Largest absolute row sums: 3 for $\mathbf{A}$, 2 for $\mathbf{A}^{-1}$, so $\kappa_\infty(\mathbf{A}) = 6$ by Theorem 16.3.3 (b). For κ_2: $\mathbf{A}^*\mathbf{A} = \operatorname{diag}(9, \frac{1}{4})$, so $\sigma_1 = 3$, $\sigma_2 = \frac{1}{2}$, and $\kappa_2(\mathbf{A}) = 6$ by Proposition 16.8.2 (d). The two agree here because $\mathbf{A}$ is diagonal.

$\mathbf{B}^{-1} = \left(\begin{smallmatrix} 1 & -2 \\ 0 & 1 \end{smallmatrix}\right)$, and both have largest absolute row sum 3, so $\kappa_\infty(\mathbf{B}) = 9$. For κ_2: $\mathbf{B}^*\mathbf{B} =$

$\left(\begin{smallmatrix} \frac{1}{2} & \frac{2}{5} \end{smallmatrix}\right)$ has trace 6 and determinant 1, so its eigenvalues are $3 \pm 2\sqrt{2}$, giving

$$\kappa_2(\mathbf{B}) = \sqrt{\frac{3+2\sqrt{2}}{3-2\sqrt{2}}} = 3 + 2\sqrt{2} \approx 5.828,$$

using $(3-2\sqrt{2})(3+2\sqrt{2}) = 1$, so that the ratio is $(3+2\sqrt{2})^2$.

Exercise. B2: A perturbation, measured against the bound

Let $\mathbf{A} = \begin{pmatrix} 1 & 1 \\ 1 & 1.01 \end{pmatrix}$ and $\mathbf{b} = (2, 2.01)$ as in Example (A system where the bound is attained), and take $\Delta\mathbf{b} = (0, 0.01)$.

(a) Solve $\mathbf{A}\tilde{\mathbf{x}} = \mathbf{b} + \Delta\mathbf{b}$ exactly.

(b) Compute both relative errors in $\|\cdot\|_\infty$ and the factor between them, and compare with $\kappa_\infty(\mathbf{A}) = 404.01$.

Solution.

(a) With $\mathbf{A}^{-1} = \left(\begin{smallmatrix} 101 & -100 \\ -100 & 100 \end{smallmatrix}\right)$ from Example (A system where the bound is attained) and $\Delta\mathbf{b} = (0, \frac{1}{100})$,

$$\tilde{\mathbf{x}} - \mathbf{x} = \mathbf{A}^{-1}\Delta\mathbf{b} = \begin{pmatrix} -1 \\ 1 \end{pmatrix}, \qquad \tilde{\mathbf{x}} = \begin{pmatrix} 0 \\ 2 \end{pmatrix}.$$

Check: $0 + 2 = 2$ and $0 + 2 \cdot \frac{101}{100} = \frac{202}{100} = 2.02 = \mathbf{b} + \Delta\mathbf{b}$ in the second coordinate.

(b) $\|\Delta\mathbf{b}\|_\infty / \|\mathbf{b}\|_\infty = \frac{1/100}{201/100} = \frac{1}{201}$, and $\|\tilde{\mathbf{x}} - \mathbf{x}\|_\infty / \|\mathbf{x}\|_\infty = 1$. The factor is 201, which is below $\kappa_\infty(\mathbf{A}) = 404.01$ by a factor of about 2. The bound holds, as Theorem 16.8.3 (a) requires, and is not attained: this $\Delta\mathbf{b}$ is not the worst direction, which Example (A system where the bound is attained) found.

Exercise. B3: Squaring, again

Let $\mathbf{A} = \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{2} \\ 0 & 0 \end{pmatrix}$.

(a) Compute the singular values of $\mathbf{A}$ and $\kappa_2(\mathbf{A})$.

(b) Compute $\mathbf{A}^\top\mathbf{A}$ and $\kappa_2(\mathbf{A}^\top\mathbf{A})$ directly, and confirm Theorem 16.8.4.

(c) Does κ_∞ also square? Compute $\kappa_\infty(\mathbf{A}^\top\mathbf{A})$ and compare with κ_∞ of the square matrix $\left(\begin{smallmatrix} 1 & 0 \\ 0 & \frac{1}{2} \end{smallmatrix}\right)$.

Solution.

(a) $\mathbf{A}^\top\mathbf{A} = \mathrm{diag}(1, \frac{1}{4})$, whose eigenvalues are 1 and $\frac{1}{4}$; so $\sigma_1(\mathbf{A}) = 1$, $\sigma_2(\mathbf{A}) = \frac{1}{2}$, and $\kappa_2(\mathbf{A}) = 2$.

(b) $\mathbf{A}^{\top}\mathbf{A} = \operatorname{diag}(1, \frac{1}{4})$ is diagonal with positive entries, so its singular values are 1 and $\frac{1}{4}$ and $\kappa_2(\mathbf{A}^{\top}\mathbf{A}) = 4 = 2^2 = \kappa_2(\mathbf{A})^2$, as Theorem 16.8.4 says.

(c) For a diagonal matrix every κ_p is the ratio of the largest to the smallest modulus on the diagonal, so $\kappa_\infty(\mathbf{A}^{\top}\mathbf{A}) = 4$ while $\kappa_\infty(\operatorname{diag}(1, \frac{1}{2})) = 2$; here the squaring happens in κ_∞ too. It does not in general: Theorem 16.8.4 is stated for κ_2, and the proof uses singular values, which only $\|\cdot\|_2$ sees. Exercise (C1: Changing the norm changes κ, but not much) bounds how far the other condition numbers can stray.

16.8.6.3 C. Going deeper

Exercise. C1: Changing the norm changes κ, but not much

Let $\|\cdot\|_a$ and $\|\cdot\|_b$ be norms on $\mathbb{C}^n$ with $c\|\mathbf{v}\|_b \le \|\mathbf{v}\|_a \le C\|\mathbf{v}\|_b$ for all $\mathbf{v}$, as supplied by Theorem 16.2.2, and write κ_a, κ_b for the corresponding condition numbers.

(a) Prove that $\|\mathbf{A}\|_a \le (C/c)\|\mathbf{A}\|_b$ for every $\mathbf{A} \in M_n(\mathbb{C})$.

(b) Deduce that $\kappa_a(\mathbf{A}) \le (C/c)^2\kappa_b(\mathbf{A})$ and $\kappa_b(\mathbf{A}) \le (C/c)^2\kappa_a(\mathbf{A})$ for every invertible $\mathbf{A}$.

(c) Using $\|\mathbf{v}\|_\infty \le \|\mathbf{v}\|_2 \le \sqrt{n}\,\|\mathbf{v}\|_\infty$, bound κ_2 in terms of κ_∞ for $n = 2$, and check the bound on the matrix $\mathbf{B} = \left(\begin{smallmatrix} 1 & 2 \\ 0 & 1 \end{smallmatrix}\right)$ of Exercise (B1: Three condition numbers).

Solution.

(a) Let $\mathbf{x} \neq \mathbf{0}$. Then

$$\frac{\|\mathbf{A}\mathbf{x}\|_a}{\|\mathbf{x}\|_a} \le \frac{C\|\mathbf{A}\mathbf{x}\|_b}{c\|\mathbf{x}\|_b} \le \frac{C}{c}\,\|\mathbf{A}\|_b,$$

using the two inequalities and Theorem 16.3.2 (a) for $\|\cdot\|_b$. Taking the maximum over $\mathbf{x} \neq \mathbf{0}$, which is $\|\mathbf{A}\|_a$ by Definition 16.3.2 and homogeneity, gives the claim.

(b) Apply (a) to $\mathbf{A}$ and $\mathbf{A}^{-1}$ and multiply: $\kappa_a(\mathbf{A}) = \|\mathbf{A}\|_a\|\mathbf{A}^{-1}\|_a \le (C/c)^2\|\mathbf{A}\|_b\|\mathbf{A}^{-1}\|_b = (C/c)^2\kappa_b(\mathbf{A})$. The reverse follows by symmetry, since the hypothesis also reads $C^{-1}\|\mathbf{v}\|_a \le \|\mathbf{v}\|_b \le c^{-1}\|\mathbf{v}\|_a$, whose ratio of constants is again C/c.

(c) With $\|\cdot\|_a = \|\cdot\|_2$ and $\|\cdot\|_b = \|\cdot\|_\infty$, we have $c = 1$ and $C = \sqrt{n} = \sqrt{2}$, so $\kappa_2(\mathbf{A}) \le 2\kappa_\infty(\mathbf{A})$ and $\kappa_\infty(\mathbf{A}) \le 2\kappa_2(\mathbf{A})$. For $\mathbf{A} = \left(\begin{smallmatrix} 1 & 2 \\ 0 & 1 \end{smallmatrix}\right)$: $\kappa_2 = 3 + 2\sqrt{2} \approx 5.83$ and $\kappa_\infty = 9$, and indeed $5.83 \le 18$ and $9 \le 11.66$. Both bounds hold, with room to spare; the point is only that no choice of norm can turn a well-conditioned matrix into an ill-conditioned one.

Exercise. C2: The condition number is the reciprocal distance to singularity

Let $\mathbf{A} \in M_n(\mathbb{C})$ be invertible.

(a) Prove that every singular $\mathbf{B} \in M_n(\mathbb{C})$ satisfies

$$\frac{\|\mathbf{A} - \mathbf{B}\|}{\|\mathbf{A}\|} \ge \frac{1}{\kappa(\mathbf{A})}.$$

(b) For the 2-norm, prove that equality is achieved: exhibit a singular $\mathbf{B}$ with $\|\mathbf{A} - \mathbf{B}\|_2 = \sigma_n(\mathbf{A})$. *Hint: use a singular value decomposition and delete one term.*

(c) Hence interpret $1/\kappa_2(\mathbf{A})$ in one sentence, and say what it gives for the matrix of

Example (A system where the bound is attained).

Solution.

(a) Put $\mathbf{E} = \mathbf{B} - \mathbf{A}$, so $\mathbf{A} + \mathbf{E} = \mathbf{B}$ is singular. If $\|\mathbf{A}^{-1}\mathbf{E}\| < 1$ then $\mathbf{A} + \mathbf{E} = \mathbf{A}(\mathbf{I} + \mathbf{A}^{-1}\mathbf{E})$ would be invertible, by Theorem 16.5.1 and the fact that a product of invertible matrices is invertible; so $\|\mathbf{A}^{-1}\mathbf{E}\| \geq 1$. By Theorem 16.3.2 (d), $1 \leq \|\mathbf{A}^{-1}\|\|\mathbf{E}\|$, that is $\|\mathbf{E}\| \geq 1/\|\mathbf{A}^{-1}\|$. Dividing by $\|\mathbf{A}\|$,

$$\frac{\|\mathbf{A} - \mathbf{B}\|}{\|\mathbf{A}\|} = \frac{\|\mathbf{E}\|}{\|\mathbf{A}\|} \geq \frac{1}{\|\mathbf{A}\|\,\|\mathbf{A}^{-1}\|} = \frac{1}{\kappa(\mathbf{A})}.$$

(b) Let $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$ (Theorem 13.9.1) with $\mathbf{\Sigma} = \operatorname{diag}(\sigma_1, \ldots, \sigma_n)$, and put

$$\mathbf{B} = \mathbf{U}\,\operatorname{diag}(\sigma_1, \ldots, \sigma_{n-1}, 0)\,\mathbf{V}^*.$$

Then $\mathbf{B}$ is singular, since $\operatorname{diag}(\sigma_1, \ldots, \sigma_{n-1}, 0)$ is and $\mathbf{U}, \mathbf{V}$ are invertible; and $\mathbf{A} - \mathbf{B} = \mathbf{U}\,\operatorname{diag}(0, \ldots, 0, \sigma_n)\,\mathbf{V}^*$, whose singular values are those of $\operatorname{diag}(0, \ldots, 0, \sigma_n)$ — for any $\mathbf{M}$, $(\mathbf{U}\mathbf{M}\mathbf{V}^*)^*(\mathbf{U}\mathbf{M}\mathbf{V}^*) = \mathbf{V}\mathbf{M}^*\mathbf{M}\mathbf{V}^*$ is similar to $\mathbf{M}^*\mathbf{M}$, so unitary factors leave the singular values alone, invertible or not — the largest of them being σ_n; so $\|\mathbf{A} - \mathbf{B}\|_2 = \sigma_n$ by Theorem 16.3.3 (c). Since $\|\mathbf{A}\|_2 = \sigma_1$, the ratio is $\sigma_n/\sigma_1 = 1/\kappa_2(\mathbf{A})$, matching (a) with equality.

(c) $1/\kappa_2(\mathbf{A})$ is the relative distance, in the 2-norm, from $\mathbf{A}$ to the nearest singular matrix. For $\mathbf{A} = \left(\begin{smallmatrix} 1 & 1 \\ 1 & 1.01 \end{smallmatrix}\right)$, whose κ_2 is the ratio of the eigenvalues $(201 \pm \sqrt{40001})/200$ of the symmetric $\mathbf{A}$, one gets $\kappa_2(\mathbf{A}) \approx 402.0$, so a relative perturbation of about 0.25% suffices to make $\mathbf{A}$ singular. Ill-conditioned means *nearly singular*, in exactly this quantitative sense.

Exercise. C3: A well-conditioned problem and a bad method

Consider the two-by-two system

$$\begin{aligned} 10^{-15}x_1 + x_2 &= 1, \\ x_1 + x_2 &= 2, \end{aligned}$$

with coefficient matrix $\mathbf{A} = \left(\begin{smallmatrix} 10^{-15} & 1 \\ 1 & 1 \end{smallmatrix}\right)$.

(a) Show that $\kappa_\infty(\mathbf{A}) < 5$, so that the problem is well conditioned.

(b) Solve the system exactly.

(c) Eliminate x_1 using the **first** equation as the pivot row, rounding every intermediate quantity to 12 significant decimal digits, and compare the result with (b). Then do the same after exchanging the two equations, and say what the exchange achieved.

(d) What does this show about the relationship between $\kappa(\mathbf{A})$ and the accuracy of a computed answer?

Solution.

(a) Write $\varepsilon = 10^{-15}$. Then $\det \mathbf{A} = \varepsilon - 1 \neq 0$ and

$$\mathbf{A}^{-1} = \frac{1}{\varepsilon - 1}\begin{pmatrix} 1 & -1 \\ -1 & \varepsilon \end{pmatrix}.$$

The largest absolute row sum of $\mathbf{A}$ is 2, and that of $\mathbf{A}^{-1}$ is $2/(1-\varepsilon)$, so by Theorem 16.3.3 (b),

$$\kappa_\infty(\mathbf{A}) = \frac{4}{1-\varepsilon} < 5.$$

(b) Subtracting the first equation from the second gives $(1-\varepsilon)x_1 = 1$, so

$$x_1 = \frac{1}{1-\varepsilon} = 1 + 10^{-15} + \dots, \qquad x_2 = 1 - \varepsilon x_1 = \frac{1-2\varepsilon}{1-\varepsilon} = 1 - 10^{-15} + \dots.$$

Both coordinates are 1 to fourteen decimal places.

(c) *Without the exchange.* The multiplier is $1/\varepsilon = 10^{15}$, and the second equation becomes

$$(1 - 10^{15})x_2 = 2 - 10^{15}.$$

Both sides need fifteen significant digits to be recorded exactly, so at twelve digits both round to $-1.00000000000 \times 10^{15}$, and the computed x_2 is exactly 1 — accurate, as it happens. Back-substitution into the first equation then gives

$$x_1 = \frac{1 - x_2}{\varepsilon} = \frac{0}{10^{-15}} = 0,$$

against a true value of $1 + 10^{-15}$: every digit is wrong. The damage was done by subtracting two numbers that agree in all twelve retained digits, so that the difference retains none of them.

With the exchange. Pivoting on $x_1 + x_2 = 2$, the multiplier is $\varepsilon = 10^{-15}$ and the second equation becomes $(1-\varepsilon)x_2 = 1 - 2\varepsilon$. At twelve digits both coefficients round to 1.00000000000, giving $x_2 = 1.00000000000$, and back-substitution gives $x_1 = 2 - x_2 = 1.00000000000$. Both agree with (b) to all twelve digits. The exchange achieved one thing: it made the multiplier at most 1 in modulus, so nothing in the elimination was amplified, and no large quantity was ever created to be subtracted from a comparable one. This is why every practical elimination scheme chooses the largest available pivot.

(d) The conditioning of the problem is excellent, $\kappa_\infty(\mathbf{A}) < 5$, and it is the same for both computations, since it does not mention a method. Yet one method returns $(0, 1)$ and the other returns the right answer. Conditioning bounds what *any* method can promise; whether a given method keeps that promise is stability, and it is Chapter 24's subject.

Chapter 17

Variational Principles and Interlacing

Since Chapter 13 the book has written $\lambda_1(\mathbf{A}) \geq \cdots \geq \lambda_n(\mathbf{A})$ for the eigenvalues of a Hermitian matrix without ever saying what the middle ones *are*. The two at the ends have an answer. Chapter 11 observed that $\langle Tv, v\rangle$ is real for a self-adjoint operator, and Chapter 13 §05 turned that into Lemma 13.5.2: on the unit sphere, $\langle \mathbf{Ax}, \mathbf{x}\rangle$ runs from $\lambda_n(\mathbf{A})$ up to $\lambda_1(\mathbf{A})$, and both ends are attained at eigenvectors. Everything between the ends was left blank, twice and deliberately. Chapter 13 §05 said it in one sentence: the intermediate eigenvalues "have no description as a plain maximum over the whole unit sphere". Chapter 13 §13 said where the blank would be filled: "Chapter 17 characterizes the eigenvalues of a Hermitian matrix as maxima and minima of $\langle \mathbf{Ax}, \mathbf{x}\rangle$ — at which point the test becomes the theory."

This is that chapter, and the device that fills the blank is a **min–max**. One maximum over the unit sphere is not enough information to see λ_2; a maximum over a *subspace*, followed by a minimum over all subspaces of the right dimension, is exactly enough. That is the Courant–Fischer theorem, and its proof is one dimension count repeated: two subspaces of F^n whose dimensions add to more than n must share a nonzero vector. Once every eigenvalue is a maximum and a minimum of a quantity that does not mention an eigenvector, eigenvalues of *different* matrices become comparable, and the rest of the chapter is the harvest. How far the eigenvalues move when the matrix moves (Weyl). How they move when a row and its column are deleted, or a rank-one matrix is added (Cauchy interlacing, and the secular equation). What sums of the largest eigenvalues optimize (Poincaré, Ky Fan), and what arbitrary index sets do (Wielandt, Lidskii). What the diagonal of a Hermitian matrix can be, given its spectrum (Schur). The same theory for singular values, which finally proves the approximation theorem Chapter 13 §11 could only state. Inertia, re-proved without a basis and then made additive across a Schur complement. And a first look at which scalar functions respect the Loewner order.

Every matrix whose eigenvalues are indexed in this chapter is Hermitian, and the field is $\mathbb{R}$ or $\mathbb{C}$. This is not a convenience. The index i in $\lambda_i(\mathbf{A})$ presupposes that the eigenvalues are real and can be sorted, which is Theorem 12.2.1, and every eigenvalue inequality below is false without it — §04 exhibits a non-Hermitian matrix with a 1×1 principal submatrix lying outside the convex hull of the whole matrix's spectrum. Each section states the hypothesis once and keeps it.

Eigenvalues are indexed decreasingly, $\lambda_1(\mathbf{A}) \geq \lambda_2(\mathbf{A}) \geq \cdots \geq \lambda_n(\mathbf{A})$, repeated according to multiplicity, as Chapter 13 §05 fixed it. Many sources index the min–max theorem the other way, and the two statements differ by $i \leftrightarrow n + 1 - i$; §02 gives the increasing form once, as a

remark, so that a reader who meets it elsewhere is not lost.

17.0.1 The analysis this chapter imports

None that is new. Chapter 16 named everything it borrowed from analysis in a numbered list (A1)–(A6) in its introduction, and this chapter uses that list and nothing else, citing an item by the same name at the point of use — "by the extreme value theorem, fact (A4) of Chapter 16's introduction". Compactness of the unit sphere (A3) and the extreme value theorem (A4) are the two that a reader will expect to see, since a min–max is an optimization and optimizations need their extrema to exist.

It is worth saying in advance that the expectation is not always met. In §§01–02, where the maxima and minima are first written down, **no analytic fact is invoked**: the spectral theorem is already available, and it hands over a maximizer explicitly, so the extreme value theorem is never called on to prove that one exists. That is not the same as saying no analysis lies underneath. The spectral theorem rests on the fundamental theorem of algebra, and Chapter 6 §05 proves that with the extreme value theorem on a closed disc; the analysis was paid for there, once, and is not paid again here. Chapter 12 §02 avoided that theorem on purpose and said so; §§01–02 below return to the remark and settle the account. Where a later section does need an item of the list, it names it. Nothing is added to the list.

17.0.2 What you need

- **Chapter 0 §09** — the trace and its cyclic invariance, which §§05 and 06 compute with: Theorem 0.9.6.
- **Chapter 1 §07** — the dimension formula $\dim(U+W) = \dim U + \dim W - \dim(U \cap W)$, Theorem 1.7.3. It is the engine of every min–max proof in the chapter, and §02 packages it as a lemma.
- **Chapter 3 §04 and §05** — permutation matrices, which §08 uses to put a spectrum in order, and Rank–Nullity for matrices, which §09 counts singular values with: Definition 3.4.2, Lemma 3.4.7, Theorem 3.5.7.
- **Chapter 2** — the coordinate isomorphism and what it preserves, and Rank–Nullity, which §§07 and 10 count with: Corollary 2.5.5, Theorem 2.5.2, Theorem 2.3.1.
- **Chapter 4** — the matrix of a map read through coordinates, which is how §02 passes between an operator and its matrix: Theorem 4.1.1.
- **Chapter 6 §01** — one fact, in §04: a nonzero polynomial of degree at most n has at most n roots, which is how the bordered-determinant identity passes from all but finitely many scalars to every scalar: Corollary 6.1.4.
- **Chapter 7 and Chapter 8** — principal submatrices (Definition 7.7.1) and Schur complements: Definition 8.3.1, Theorem 8.3.2, Theorem 8.3.1, Corollary 8.3.6. §§04, 05, and 10 lean on them.
- **Chapter 9** — the trace and determinant as sum and product of the eigenvalues, Theorem 9.3.3.
- **Chapter 11** — inner products, orthonormal bases, and the adjoint: Definition 11.1.1, Definition 11.1.3, Theorem 11.1.5, Theorem 11.2.1, Definition 11.3.1, Theorem 11.3.2, Corollary 11.2.3, Definition 11.6.2, Theorem 11.6.3, Proposition 11.6.7.
- **Chapter 12** — the spectral theorems, in both the operator and the matrix form: Theorem 12.2.1, Theorem 12.4.2, Corollary 12.4.3, Theorem 12.5.1, Corollary 12.5.2, Corollary 12.5.4, and the functional calculus Definition 12.8.2, Theorem 12.8.3.
- **Chapter 13** — positivity, the Loewner order, and the singular value decomposition: Definition 13.1.1, Theorem 13.1.4, Theorem 13.1.5, Theorem 13.2.1, Definition 13.5.1, Proposition 13.5.1, Lemma 13.5.2, Theorem 13.5.3, Proposition 13.5.4, Definition 13.9.1,

Theorem 13.9.1, Theorem 13.9.3, Definition 13.11.1, Lemma 13.11.2, Theorem 13.11.3; from §04 the generalized eigenvalues (Definition 13.4.1, Theorem 13.4.2), and from §06 Hadamard's inequality and the block criterion (Theorem 13.7.3, Theorem 13.6.1), each of which gets a second proof here.

- **Chapter 14 §§02, 05** — congruence (Definition 14.2.1), inertia, and signature: Theorem 14.5.1, Definition 14.5.1, Theorem 14.5.3. This one is a connection rather than a prerequisite: §10 gives Sylvester's law a second proof and reworks its examples, and no proof in the chapter cites Chapter 14 — one worked example and two exercises of §10 do.
- **Chapter 16** — the way a matrix is measured, and the questions left open: Definition 16.3.2, Theorem 16.3.2, and Theorem 16.3.3, whose part (c) is the formula $\|\mathbf{A}\|_2 = \sigma_1(\mathbf{A})$; then Definition 16.4.1, Theorem 16.4.1, Definition 16.8.1, and the imported-analysis list of the Chapter 16 introduction.

17.0.3 Roadmap

- **The Rayleigh quotient.** $R_{\mathbf{A}}(\mathbf{x}) = \langle \mathbf{A}\mathbf{x}, \mathbf{x}\rangle / \langle \mathbf{x}, \mathbf{x}\rangle$ promoted from a test to an object: its range is exactly $[\lambda_n, \lambda_1]$, its critical points on the unit sphere are exactly the unit eigenvectors, and the first description of λ_k — maximize after deflating by the top eigenvectors — is correct but circular.
- **The min–max theorem.** Courant–Fischer: λ_k is a minimum over subspaces of a maximum over vectors, and equally a maximum over subspaces of a minimum. No eigenvector is named, so two matrices can now be compared eigenvalue by eigenvalue.
- **How far eigenvalues move.** Weyl's inequalities, and the Lipschitz bound $|\lambda_i(\mathbf{A}+\mathbf{E}) - \lambda_i(\mathbf{A})| \le \|\mathbf{E}\|_2$ that Chapter 16 §07 promised.
- **Deleting a row and a column.** Cauchy interlacing: the spectrum of a principal submatrix is caught between consecutive eigenvalues of the whole matrix. When the interlacing is strict, which settles a promise of Chapter 11: the zeros of consecutive orthogonal polynomials interlace strictly. And why none of it survives without symmetry.
- **Adding a rank-one matrix.** Interlacing again, the secular equation that locates the new eigenvalues between the old, and the shifts that add up to $\tau\|\mathbf{v}\|^2$.
- **Subspaces and sums.** Poincaré's separation theorem, and Ky Fan's: the sum of the k largest eigenvalues is a maximum of a trace over orthonormal k-frames, hence a convex function of the matrix. Chapter 13's capture bound turns out to be one case of it.
- **Arbitrary index sets.** The Wielandt min–max over a flag of subspaces, and Lidskii's inequality. It contains Ky Fan's subadditivity and the Weyl inequalities with one index at an end — but not all of Weyl's, which is worth seeing.
- **The diagonal and the spectrum.** Majorization; Schur's theorem and Horn's converse, which together say that the diagonals of the Hermitian matrices with a given spectrum are exactly the vectors that spectrum majorizes.
- **The same theory for singular values.** The min–max for σ_k, perturbation with no hypothesis on the perturbation at all, the spectral-norm form of Eckart–Young, interlacing when a row or column is deleted, and the distance from an invertible matrix to the nearest singular one — which is what Chapter 16's condition number was measuring all along.
- **Inertia, again and additively.** A second proof of Sylvester's law with no basis in it, now over $\mathbb{C}$ as well; how far a rank-r perturbation can move the inertia; and Haynsworth's additivity across a Schur complement.
- **Which functions respect the order.** Operator monotone and operator convex functions, why the square root behaves and the square does not, and a statement of the classification that Chapter 21 proves.

17.0.4 Named moves

- **Count dimensions to force an intersection.** Two subspaces of an n-dimensional space whose dimensions add to more than n meet in a nonzero vector. This one line proves Courant–Fischer and Weyl directly, and Cauchy interlacing and Poincaré through Courant–Fischer. It is the chapter's single technique, and everything else is bookkeeping around it.
- **Prove a min–max by two inequalities of different characters.** One direction exhibits a **witness** subspace — usually a span of eigenvectors — and computes on it. The other takes an **arbitrary** subspace and intersects it with a span of eigenvectors from the other end. Neither half resembles the other, and recognizing which half is which is most of the work in reading any proof below.
- **Apply the result to $-\mathbf{A}$.** The two halves of Courant–Fischer, the two families of Weyl inequalities, and the top- and bottom-k trace theorems are each other's images under $\mathbf{A} \mapsto -\mathbf{A}$, which reverses the order and sends λ_i to $-\lambda_{n+1-i}$. Deriving the second statement this way is a genuine argument, not a wave of the hand, and it is written out each time.
- **Compress, then interlace.** Restricting the quadratic form to a subspace produces a smaller Hermitian matrix whose eigenvalues are trapped by the original's. Deleting rows and columns, adding a rank-one matrix, and cutting a singular value decomposition short are all this move.
- **Perturb and compare, eigenvalue by eigenvalue.** Chapter 16's move, now applied one index at a time rather than to the spectrum as a whole.
- **Reduce singular values to eigenvalues.** Either through $\mathbf{A}^*\mathbf{A}$, which is shortest, or through the Hermitian dilation, which is symmetric in $\mathbf{A}$ and $\mathbf{A}^*$. Both appear, and each section says which it uses and why.
- **Majorize.** When the right comparison is between *sums* of the largest entries rather than individual ones, compare the partial sums of the decreasing rearrangements. The diagonal of a Hermitian matrix against its spectrum (§08), and the Ky Fan and Lidskii sums (§§06–07), are all of this type.

17.1 The Rayleigh Quotient

Chapter 13 used the number $\langle \mathbf{Ax}, \mathbf{x} \rangle$ as a *test*: positive for every nonzero $\mathbf{x}$ means positive definite, nonnegative means positive semidefinite, and that was the end of it. Chapter 13 §05 then noticed that the test carries more information than the verdict — normalized, it ranges over the interval between the smallest and largest eigenvalues — and stopped there. This section makes the normalized quantity an object with a name, finds its range exactly, finds the vectors where it is flat, and extracts a first description of every eigenvalue as a maximum. The description is correct and unsatisfying, and saying precisely why is the business of §02.

Throughout, F is $\mathbb{R}$ or $\mathbb{C}$, every matrix is Hermitian, and its eigenvalues are indexed decreasingly, $\lambda_1(\mathbf{A}) \geq \cdots \geq \lambda_n(\mathbf{A})$, repeated according to multiplicity. The indexing presupposes that the eigenvalues are real, which is Theorem 12.2.1. Inner products are the standard ones of F^n, linear in the first slot.

17.1.1 From a test to an object

Look at what Lemma 13.5.2 actually says. For a unit vector $\mathbf{x}$, the number $\langle \mathbf{Ax}, \mathbf{x} \rangle$ lies between $\lambda_n(\mathbf{A})$ and $\lambda_1(\mathbf{A})$, and it reaches both ends. The restriction to unit vectors is a nuisance: doubling $\mathbf{x}$ quadruples $\langle \mathbf{Ax}, \mathbf{x} \rangle$, so for $\mathbf{A} \neq \mathbf{O}$ the raw quantity has no bounded range at all and

every statement about it must carry the clause "for $\|\mathbf{x}\| = 1$" along. Dividing by $\|\mathbf{x}\|^2$ removes the nuisance permanently, and the resulting expression occurs often enough in what follows to deserve a name of its own.

The Rayleigh quotient measures how much a Hermitian matrix stretches a vector along its own direction, on a scale where the vector's length has been divided out.

Definition 17.1.1. Rayleigh Quotient

Let F be $\mathbb{R}$ or $\mathbb{C}$ and let $\mathbf{A} \in M_n(F)$ be **Hermitian**. The **Rayleigh quotient** of $\mathbf{A}$ is the function

$$R_{\mathbf{A}}(\mathbf{x}) := \frac{\langle \mathbf{A}\mathbf{x}, \mathbf{x} \rangle}{\langle \mathbf{x}, \mathbf{x} \rangle},$$

defined at **every** $\mathbf{x} \in F^n$ with $\mathbf{x} \neq \mathbf{0}$, and at no other vector.

For a self-adjoint operator T on a finite-dimensional inner product space V over F, the same formula $R_T(v) = \langle Tv, v \rangle / \langle v, v \rangle$, for $v \neq 0$, defines the Rayleigh quotient of T.

Clause by clause. The word **Hermitian** is what makes the values real, and so what makes it sensible to ask for a maximum: by Proposition 11.6.7 (a), $\langle \mathbf{A}\mathbf{x}, \mathbf{x} \rangle \in \mathbb{R}$ for every $\mathbf{x}$, even over $\mathbb{C}$. The condition $\mathbf{x} \neq \mathbf{0}$ is what makes the fraction mean anything: $\langle \mathbf{x}, \mathbf{x} \rangle = \|\mathbf{x}\|^2 > 0$ exactly for $\mathbf{x} \neq \mathbf{0}$, by positive definiteness of the inner product (Definition 11.1.1), and there is no value to assign at $\mathbf{0}$ — the numerator vanishes there too, and the quotient $0/0$ would have to be given a value by hand, which no theorem below would thank us for. So $R_{\mathbf{A}}$ is a real-valued function on $F^n \setminus \{\mathbf{0}\}$.

Well-definedness has a second half, and it is the useful one. For a scalar $c \neq 0$,

$$\langle \mathbf{A}(c\mathbf{x}), c\mathbf{x} \rangle = c\,\bar{c}\,\langle \mathbf{A}\mathbf{x}, \mathbf{x} \rangle = |c|^2 \langle \mathbf{A}\mathbf{x}, \mathbf{x} \rangle,$$
$$\langle c\mathbf{x}, c\mathbf{x} \rangle = |c|^2 \langle \mathbf{x}, \mathbf{x} \rangle,$$

so the factor $|c|^2$ cancels and

$$R_{\mathbf{A}}(c\mathbf{x}) = R_{\mathbf{A}}(\mathbf{x}) \qquad \text{for every } c \neq 0.$$

The Rayleigh quotient is therefore constant on each line through the origin with the origin removed. It is really a function on the set of lines, and the cheapest way to evaluate it is to normalize: for a **unit** vector $\mathbf{x}$,

$$R_{\mathbf{A}}(\mathbf{x}) = \langle \mathbf{A}\mathbf{x}, \mathbf{x} \rangle.$$

Every statement below may be read either way — over all nonzero vectors, or over the unit sphere — and we shall move between the two without comment.

Example. A diagonal matrix, and the whole chapter in miniature

Let $\mathbf{A} = \operatorname{diag}(d_1, \dots, d_n)$ with real d_i. Compute $R_{\mathbf{A}}$.

Solution. For $\mathbf{x} = (x_1, \dots, x_n) \neq \mathbf{0}$ we have $\mathbf{A}\mathbf{x} = (d_1 x_1, \dots, d_n x_n)$, so

$$R_{\mathbf{A}}(\mathbf{x}) = \frac{\sum_i d_i |x_i|^2}{\sum_i |x_i|^2} = \sum_i w_i d_i, \qquad w_i = \frac{|x_i|^2}{\|\mathbf{x}\|^2}.$$

The weights w_i are nonnegative and sum to 1. So $R_{\mathbf{A}}(\mathbf{x})$ is a **weighted average of the diagonal entries**, and $\mathbf{x}$ is nothing but the recipe for the weights.

That one line is the entire chapter in miniature, and it is worth pausing on. A weighted average of $d_1, \dots, d_n$ can be any number between the smallest and the largest, and nothing else; it equals the largest only when all the weight sits on the largest entries; and to force it down to the second largest one must **forbid** the weight from sitting on the first. Every theorem in this chapter is one of those three sentences, stated for a general Hermitian matrix, where the diagonal entries are replaced by the eigenvalues and the coordinates x_i by the coefficients of $\mathbf{x}$ in an orthonormal eigenbasis. The spectral theorem is what turns the general case into this one.

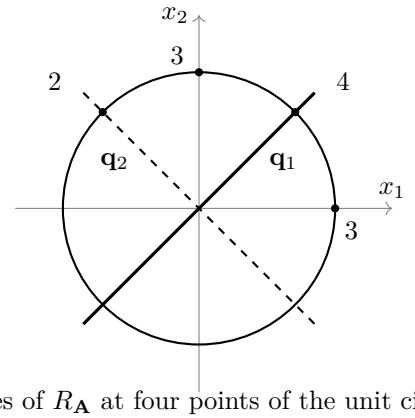
Example. A symmetric matrix on the unit circle

Let $\mathbf{A} = \begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix} \in M_2(\mathbb{R})$. Describe $R_{\mathbf{A}}$ on the unit circle, and find its largest and smallest values.

Solution. The eigenvalues are 4 and 2, with orthogonal eigenvectors $\mathbf{q}_1 = \frac{1}{\sqrt{2}}(1,1)$ and $\mathbf{q}_2 = \frac{1}{\sqrt{2}}(1,-1)$: indeed $\mathbf{A}(1,1) = (4,4)$ and $\mathbf{A}(1,-1) = (2,-2)$. Writing a unit vector as $\mathbf{x} = (\cos\theta, \sin\theta)$,

$$R_{\mathbf{A}}(\mathbf{x}) = \langle \mathbf{A}\mathbf{x}, \mathbf{x} \rangle = 3\cos^2\theta + 2\cos\theta\sin\theta + 3\sin^2\theta = 3 + \sin 2\theta.$$

So the values run over $[2, 4]$; the maximum 4 occurs at $\theta = \pi/4$ and $\theta = 5\pi/4$, that is along the line through $\mathbf{q}_1$, and the minimum 2 at $\theta = 3\pi/4$ and $7\pi/4$, along the line through $\mathbf{q}_2$. At $\theta = 0$ and $\theta = \pi/2$ the value is 3, the average of the two eigenvalues, which is also $a_{11} = a_{22}$.



The values of $R_{\mathbf{A}}$ at four points of the unit circle, for the symmetric matrix with diagonal entries 3 and off-diagonal entries 1. It is largest along the eigenline through $\mathbf{q}_1$, smallest along the one through $\mathbf{q}_2$, and takes every value in between

Example. Two matrices whose quotient is forced

Two further cases, at opposite extremes.

(a) **The identity.** $R_{\mathbf{I}}(\mathbf{x}) = \langle \mathbf{x}, \mathbf{x} \rangle / \langle \mathbf{x}, \mathbf{x} \rangle = 1$ for every $\mathbf{x} \neq \mathbf{0}$. The quotient is constant, and the constant is the only eigenvalue. More generally $R_{c\mathbf{I}} \equiv c$ for real c.

(b) **An orthogonal projection.** Let $\mathbf{P} \in M_n(F)$ be the matrix of the orthogonal projection onto a subspace U, so $\mathbf{P}^2 = \mathbf{P}$ and $\mathbf{P}^* = \mathbf{P}$ (Theorem 13.13.1). Then $\langle \mathbf{Px}, \mathbf{x} \rangle = \langle \mathbf{P}^2\mathbf{x}, \mathbf{x} \rangle = \langle \mathbf{Px}, \mathbf{Px} \rangle = \|\mathbf{Px}\|^2$, so

$$R_{\mathbf{P}}(\mathbf{x}) = \frac{\|\mathbf{Px}\|^2}{\|\mathbf{x}\|^2} \in [0, 1],$$

the fraction of $\mathbf{x}$ that survives the projection. Since $\|\mathbf{x}\|^2 = \|\mathbf{Px}\|^2 + \|\mathbf{x} - \mathbf{Px}\|^2$ by Theorem 11.1.3, it equals 1 exactly when $\mathbf{Px} = \mathbf{x}$, that is on $U \setminus \{\mathbf{0}\}$, and 0 exactly when $\mathbf{Px} = \mathbf{0}$, that is on $U^{\perp} \setminus \{\mathbf{0}\}$ — matching the eigenvalues 1 and 0 of $\mathbf{P}$.

Case (a) is the degenerate one, and it earns its place: when all the eigenvalues coincide the interval of values collapses to a point, every nonzero vector is an eigenvector, and every theorem below becomes an equality between two copies of the same number. A statement that fails to survive $\mathbf{A} = c\mathbf{I}$ has been stated wrongly.

Example. Change one word: drop Hermitian

Take $\mathbf{A} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \in M_2(\mathbb{C})$, which is not Hermitian, and form $\langle \mathbf{Ax}, \mathbf{x} \rangle / \langle \mathbf{x}, \mathbf{x} \rangle$ anyway. At $\mathbf{x} = (1, i)$ the numerator is $\langle (i, 0), (1, i) \rangle = i \cdot \overline{1} = i$ and the denominator is 2, so the quotient is $i/2$, which is not real. The clause that fails is the word **Hermitian** in Definition 17.1.1, and it fails at the first step: Proposition 11.6.7 (a) is what made the numerator real, and its hypothesis is gone.

Over $\mathbb{R}$ the failure is quieter and worse. For $\mathbf{A} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \in M_2(\mathbb{R})$ and real $\mathbf{x} = (x_1, x_2)$,

$$\langle \mathbf{Ax}, \mathbf{x} \rangle = -x_2 x_1 + x_1 x_2 = 0,$$

so the quotient is real, and identically 0. It is perfectly well behaved and tells us nothing: $\mathbf{A}$ has no real eigenvalue at all, and there is no interval $[\lambda_n, \lambda_1]$ for the quotient to describe.

Why this definition, then. Dropping "Hermitian" costs the reality of the values, hence any notion of largest; the real 2×2 example above shows that it also destroys the link with the spectrum even when the values happen to stay real. Dropping the denominator costs scale invariance, and with it the possibility of writing max without a side condition. Keeping the denominator but allowing $\mathbf{x} = \mathbf{0}$ costs nothing but sense.

Warning. **$R_{\mathbf{A}}(\mathbf{x})$ and $\langle \mathbf{Ax}, \mathbf{x} \rangle$ are the same number only for unit $\mathbf{x}$.** The bound $\langle \mathbf{Ax}, \mathbf{x} \rangle \le \lambda_1(\mathbf{A})$ of Lemma 13.5.2 is stated for unit vectors and is false without that hypothesis: for $\mathbf{A} = \mathbf{I}_2$ and $\mathbf{x} = (10, 0)$ it would read $100 \le 1$. The bound $R_{\mathbf{A}}(\mathbf{x}) \le \lambda_1(\mathbf{A})$, by contrast, holds for every $\mathbf{x} \neq \mathbf{0}$. Whenever a maximum over "all $\mathbf{x}$" appears below, it is a maximum of the **quotient**; the un-normalized form of the same statement always carries $\|\mathbf{x}\| = 1$ with it.

17.1.2 What values it takes

Lemma 13.5.2 gives the two ends of the range. The following proposition adds that nothing in between is missing, which is a smaller statement than it looks and is worth proving explicitly

rather than quoting a continuity argument.

Proposition 17.1.1. The Range of the Rayleigh Quotient

Let $\mathbf{A} \in M_n(F)$ be Hermitian, $n \geq 1$. Then

$$\{R_{\mathbf{A}}(\mathbf{x}) : \mathbf{x} \in F^n,\ \mathbf{x} \neq \mathbf{0}\} = [\lambda_n(\mathbf{A}),\ \lambda_1(\mathbf{A})],$$

the whole closed interval. In particular

$$\lambda_1(\mathbf{A}) = \max_{\mathbf{x}\neq\mathbf{0}} R_{\mathbf{A}}(\mathbf{x}), \qquad \lambda_n(\mathbf{A}) = \min_{\mathbf{x}\neq\mathbf{0}} R_{\mathbf{A}}(\mathbf{x}),$$

and both are attained, at unit eigenvectors for $\lambda_1(\mathbf{A})$ and $\lambda_n(\mathbf{A})$.

Idea. Two inclusions of quite different weights. ($\subseteq$) is Lemma 13.5.2 with the normalization divided out, and costs one line. ($\supseteq$) asks for a vector realizing each intermediate value; the natural instinct is to move continuously along the sphere from a top eigenvector to a bottom one and quote the intermediate value theorem, but there is no need to import anything. Mixing a top eigenvector and a bottom eigenvector in the proportion $t : 1-t$ puts weight t on λ_1 and $1-t$ on λ_n, and the weighted average of Example (A diagonal matrix, and the whole chapter in miniature) sweeps the interval as t does.

Proof. By the spectral theorem for matrices — Corollary 12.4.3 over $\mathbb{C}$, Corollary 12.5.2 over $\mathbb{R}$ — there is an orthonormal basis $(\mathbf{q}_1, \dots, \mathbf{q}_n)$ of F^n with $\mathbf{A}\mathbf{q}_i = \lambda_i(\mathbf{A})\mathbf{q}_i$ for every i.

($\subseteq$) Let $\mathbf{x} \neq \mathbf{0}$ and put $\mathbf{u} = \mathbf{x}/\|\mathbf{x}\|$, a unit vector. By the scale invariance established above, $R_{\mathbf{A}}(\mathbf{x}) = R_{\mathbf{A}}(\mathbf{u}) = \langle \mathbf{A}\mathbf{u}, \mathbf{u}\rangle$, which lies in $[\lambda_n(\mathbf{A}), \lambda_1(\mathbf{A})]$ by Lemma 13.5.2.

($\supseteq$) Let $\mu \in [\lambda_n(\mathbf{A}), \lambda_1(\mathbf{A})]$. If $\lambda_1(\mathbf{A}) = \lambda_n(\mathbf{A})$ the interval is the single point $\mu = \lambda_1(\mathbf{A}) = R_{\mathbf{A}}(\mathbf{q}_1)$, and we are done. Otherwise put

$$t = \frac{\mu - \lambda_n(\mathbf{A})}{\lambda_1(\mathbf{A}) - \lambda_n(\mathbf{A})} \in [0, 1], \qquad \mathbf{x} = \sqrt{t}\,\mathbf{q}_1 + \sqrt{1-t}\,\mathbf{q}_n,$$

which is legitimate because the denominator is nonzero and both square roots are of non-negative reals. Note $n \geq 2$ here, since $\lambda_1 \neq \lambda_n$, so $\mathbf{q}_1$ and $\mathbf{q}_n$ are distinct members of an orthonormal list. By Theorem 11.2.1 (c),

$$\|\mathbf{x}\|^2 = t + (1-t) = 1,$$

so in particular $\mathbf{x} \neq \mathbf{0}$. Since $\mathbf{A}\mathbf{x} = \sqrt{t}\,\lambda_1(\mathbf{A})\mathbf{q}_1 + \sqrt{1-t}\,\lambda_n(\mathbf{A})\mathbf{q}_n$, Theorem 11.2.1 (b) gives

$$R_{\mathbf{A}}(\mathbf{x}) = \langle \mathbf{A}\mathbf{x}, \mathbf{x}\rangle = t\,\lambda_1(\mathbf{A}) + (1-t)\,\lambda_n(\mathbf{A}) = \mu,$$

the last equality by the choice of t. So every point of the interval is a value.

For the attainment, $R_{\mathbf{A}}(\mathbf{q}_1) = \lambda_1(\mathbf{A})$ and $R_{\mathbf{A}}(\mathbf{q}_n) = \lambda_n(\mathbf{A})$. This proves the proposition. ∎

So the range is an **interval**, not a finite set, even though $\mathbf{A}$ has at most n eigenvalues. That

single fact is responsible for most of the confusion this chapter has to clear up, and the warning below spells out why.

Quick check. For $\mathbf{A} = \operatorname{diag}(6, 3, 0)$, write down two unit vectors with Rayleigh quotient 4, each having one zero coordinate, in different positions.

Solution. Mixing $\mathbf{e}_1$ and $\mathbf{e}_3$ as in the proof: $4 = 6t + 0(1-t)$ gives $t = 2/3$ and $\mathbf{x} = (\sqrt{2/3},\, 0,\, \sqrt{1/3})$, for which $R_{\mathbf{A}}(\mathbf{x}) = 6\cdot\frac{2}{3} + 0\cdot\frac{1}{3} = 4$. Mixing $\mathbf{e}_1$ and $\mathbf{e}_2$ instead: $4 = 6t + 3(1-t)$ gives $t = 1/3$ and $\mathbf{y} = (\sqrt{1/3},\, \sqrt{2/3},\, 0)$, for which $R_{\mathbf{A}}(\mathbf{y}) = 6\cdot\frac{1}{3} + 3\cdot\frac{2}{3} = 4$. A value of $R_{\mathbf{A}}$ is realized by many vectors, and they need not resemble each other.

17.1.3 Where the quotient is flat

The maximum and the minimum are attained at eigenvectors. There is a converse, and it identifies the eigenvectors among all vectors by a condition that does not mention eigenvalues: they are the places where $R_{\mathbf{A}}$ does not move to first order. Everything needed is one algebraic identity, exact for every t, so no calculus enters.

Proposition 17.1.2. Critical Points of the Rayleigh Quotient

Let $\mathbf{A} \in M_n(F)$ be Hermitian, let $\mathbf{x} \in F^n$ be a **unit** vector, and put $\mu = R_{\mathbf{A}}(\mathbf{x}) = \langle \mathbf{A}\mathbf{x}, \mathbf{x}\rangle$. For every $\mathbf{v} \in F^n$ and every $t \in \mathbb{R}$ with $\mathbf{x} + t\mathbf{v} \neq \mathbf{0}$,

$$R_{\mathbf{A}}(\mathbf{x}+t\mathbf{v}) - \mu = \frac{2t\,\operatorname{Re}\langle \mathbf{A}\mathbf{x} - \mu\mathbf{x}, \mathbf{v}\rangle + t^2\big(\langle \mathbf{A}\mathbf{v}, \mathbf{v}\rangle - \mu\|\mathbf{v}\|^2\big)}{\|\mathbf{x}+t\mathbf{v}\|^2}. \tag{17.1.1}$$

Moreover:

(a) $\operatorname{Re}\langle \mathbf{A}\mathbf{x} - \mu\mathbf{x}, \mathbf{v}\rangle = 0$ for every $\mathbf{v} \in F^n$ **if and only if** $\mathbf{A}\mathbf{x} = \mu\mathbf{x}$; that is, if and only if $\mathbf{x}$ is an eigenvector of $\mathbf{A}$, necessarily for the eigenvalue μ.

(b) If $R_{\mathbf{A}}(\mathbf{x}) \geq R_{\mathbf{A}}(\mathbf{y})$ for every $\mathbf{y} \neq \mathbf{0}$, then the condition in (a) holds. The same is true if $R_{\mathbf{A}}(\mathbf{x}) \leq R_{\mathbf{A}}(\mathbf{y})$ for every $\mathbf{y} \neq \mathbf{0}$.

Idea. Expand the numerator and the denominator of $R_{\mathbf{A}}(\mathbf{x}+t\mathbf{v})$ in powers of t and subtract μ times the denominator from the numerator. The constant terms cancel by the definition of μ, and what is left is exactly Eq. 17.1.1. The coefficient of t is the object of interest: (a) says it vanishes in every direction precisely at eigenvectors, and (b) says that at a maximum it must vanish, since otherwise a small step in the right direction would increase the quotient — the t^2 term cannot interfere, because for small t it is smaller than the t term.

Proof. The identity. Expanding by sesquilinearity,

$$\begin{aligned}\langle \mathbf{A}(\mathbf{x}+t\mathbf{v}), \mathbf{x}+t\mathbf{v}\rangle &= \langle \mathbf{A}\mathbf{x},\mathbf{x}\rangle + t\langle \mathbf{A}\mathbf{x},\mathbf{v}\rangle + t\langle \mathbf{A}\mathbf{v},\mathbf{x}\rangle + t^2\langle \mathbf{A}\mathbf{v},\mathbf{v}\rangle \\ &= \mu + 2t\,\mathrm{Re}\langle \mathbf{A}\mathbf{x},\mathbf{v}\rangle + t^2\langle \mathbf{A}\mathbf{v},\mathbf{v}\rangle,\end{aligned}$$

where the second line used $\langle \mathbf{A}\mathbf{v},\mathbf{x}\rangle = \langle \mathbf{v},\mathbf{A}\mathbf{x}\rangle = \overline{\langle \mathbf{A}\mathbf{x},\mathbf{v}\rangle}$, the first equality because $\mathbf{A}$ is Hermitian and the second by conjugate symmetry (Definition 11.1.1), together with $\langle \mathbf{A}\mathbf{x},\mathbf{x}\rangle = \mu$ for the unit vector $\mathbf{x}$. The same computation with $\mathbf{A}$ replaced by $\mathbf{I}$, where the constant term is $\|\mathbf{x}\|^2 = 1$, gives

$$\|\mathbf{x}+t\mathbf{v}\|^2 = 1 + 2t\,\mathrm{Re}\langle \mathbf{x},\mathbf{v}\rangle + t^2\|\mathbf{v}\|^2.$$

Since μ is real, $\mu\,\mathrm{Re}\langle \mathbf{x},\mathbf{v}\rangle = \mathrm{Re}\langle \mu\mathbf{x},\mathbf{v}\rangle$. Subtracting $\mu\|\mathbf{x}+t\mathbf{v}\|^2$ from the first display therefore leaves

$$2t\,\mathrm{Re}\langle \mathbf{A}\mathbf{x}-\mu\mathbf{x},\mathbf{v}\rangle + t^2\big(\langle \mathbf{A}\mathbf{v},\mathbf{v}\rangle - \mu\|\mathbf{v}\|^2\big),$$

and dividing by $\|\mathbf{x}+t\mathbf{v}\|^2 \neq 0$ gives Eq. 17.1.1.

(a) Put $\mathbf{w} = \mathbf{A}\mathbf{x} - \mu\mathbf{x}$. ($\Leftarrow$) If $\mathbf{A}\mathbf{x} = \mu\mathbf{x}$ then $\mathbf{w} = \mathbf{0}$ and every $\langle \mathbf{w},\mathbf{v}\rangle$ vanishes. ($\Rightarrow$) Take $\mathbf{v} = \mathbf{w}$: the hypothesis gives $\mathrm{Re}\langle \mathbf{w},\mathbf{w}\rangle = \|\mathbf{w}\|^2 = 0$, so $\mathbf{w} = \mathbf{0}$ by positive definiteness, that is $\mathbf{A}\mathbf{x} = \mu\mathbf{x}$. Since $\mathbf{x} \neq \mathbf{0}$, this exhibits $\mathbf{x}$ as an eigenvector for the eigenvalue μ.

(b) Suppose $\mathbf{x}$ maximizes $R_{\mathbf{A}}$ and, for a contradiction, that $c := 2\,\mathrm{Re}\langle \mathbf{A}\mathbf{x}-\mu\mathbf{x},\mathbf{v}\rangle \neq 0$ for some $\mathbf{v}$; then $\mathbf{v} \neq \mathbf{0}$. Put $d = \langle \mathbf{A}\mathbf{v},\mathbf{v}\rangle - \mu\|\mathbf{v}\|^2$, and let t be any real number with

$$0 < |t| < \frac{1}{2\|\mathbf{v}\|}, \qquad |t|\,|d| < \tfrac{1}{2}|c|, \qquad t \text{ of the same sign as } c,$$

which is possible because the two upper bounds are positive. The first bound gives $\|\mathbf{x}+t\mathbf{v}\| \geq \|\mathbf{x}\| - |t|\|\mathbf{v}\| > 1 - \tfrac{1}{2} > 0$, by the triangle inequality (Corollary 11.1.6), so $\mathbf{x}+t\mathbf{v} \neq \mathbf{0}$ and the identity applies. The second gives $|td| < |c|/2$, so $c + td$ has the same sign as c, and hence $t(c+td) > 0$. By Eq. 17.1.1,

$$R_{\mathbf{A}}(\mathbf{x}+t\mathbf{v}) - \mu = \frac{t\,(c+td)}{\|\mathbf{x}+t\mathbf{v}\|^2} > 0,$$

contradicting the maximality of $R_{\mathbf{A}}(\mathbf{x})$. So $c = 0$ for every $\mathbf{v}$. For a minimizer, apply what has just been proved to $-\mathbf{A}$, whose Rayleigh quotient is $-R_{\mathbf{A}}$, so that a minimizer of $R_{\mathbf{A}}$ is a maximizer of $R_{-\mathbf{A}}$; the condition in (a) for $-\mathbf{A}$ and $-\mu$ is the same condition. This proves the proposition. ∎

Call a unit vector $\mathbf{x}$ a **critical point** of $R_{\mathbf{A}}$ when the condition in (a) holds, that is, when the coefficient of t in Eq. 17.1.1 vanishes in every direction $\mathbf{v}$ — when $R_{\mathbf{A}}$ moves away from $R_{\mathbf{A}}(\mathbf{x})$ only at second order. Parts (a) and (b) then say: *the critical points of $R_{\mathbf{A}}$ on the unit sphere are exactly the unit eigenvectors of $\mathbf{A}$, the critical values are exactly the eigenvalues, and every maximizer and every minimizer is among them.* That the eigenvalues are all realized as critical values is a direct computation: $R_{\mathbf{A}}(\mathbf{q}_i) = \langle \lambda_i\mathbf{q}_i,\mathbf{q}_i\rangle = \lambda_i$, and each $\mathbf{q}_i$ is a unit eigenvector, hence a critical point by (b).

Remark. A reader who knows calculus will recognize Eq. 17.1.1 as saying that the directional derivative of $R_{\mathbf{A}}$ at a unit vector $\mathbf{x}$ in the direction $\mathbf{v}$ is $2\,\mathrm{Re}\langle \mathbf{A}\mathbf{x}-\mu\mathbf{x},\mathbf{v}\rangle$, so that "critical point" has its usual meaning of a vanishing gradient. Nothing above uses derivatives: the

identity Eq. 17.1.1 is exact for every t, not an approximation, and the argument in (b) compares two explicit numbers. This matters for the accounting in the next remark.

Remark. **The promise from Chapter 12 §02, and what it cost.** Chapter 12 §02 proved that a real self-adjoint operator has an eigenvalue by an algebraic route, and then described the alternative it was declining: "the function $v \mapsto \langle Tv, v\rangle$ is real-valued and continuous on the unit sphere of V, which is closed and bounded, so it attains a maximum, and a vector where it does so turns out to be an eigenvector for the largest eigenvalue. That argument is shorter to state but it is analysis, not algebra, and it needs the extreme value theorem; we keep to the algebraic route so that nothing in this chapter depends on it. The maximization picture returns in Chapter 17, where the Rayleigh quotient and the Courant–Fischer theorem describe every eigenvalue of a self-adjoint operator by optimization."

The picture has now returned, and the account can be settled exactly. The clause "a vector where it does so turns out to be an eigenvector" is Proposition 17.1.2 (b) and (a), and that the eigenvalue is the *largest* one is Proposition 17.1.1; neither invokes an analytic fact. What Chapter 12 would have had to import is only the *existence* of the maximizer, which is the extreme value theorem — fact (A4) of Chapter 16's introduction — applied to a continuous function on the unit sphere, compact by fact (A3). This chapter does not import it, because it does not have to: the spectral theorem is available here and it hands over the maximizer by name, namely $\mathbf{q}_1$, in Proposition 17.1.1. So the optimization description of eigenvalues in this chapter is a **consequence** of the spectral theorem, not an independent proof of it. One honest qualification: the spectral theorem itself rests on the fundamental theorem of algebra (Theorem 6.5.1), which Chapter 6 §05 proved with the extreme value theorem on a disc. So the analysis is not absent from the foundations — it was spent once, there, and nothing here spends it again. Section 2 completes the promise for the intermediate eigenvalues, on the same terms.

17.1.4 Why the second eigenvalue is not a second maximum

Warning. **$\lambda_2(\mathbf{A})$ is not "the second largest value of $R_{\mathbf{A}}$", and there is no sense in which it could be.** Take $\mathbf{A} = \operatorname{diag}(3, 1, -1)$, so $\lambda_1 = 3$, $\lambda_2 = 1$, $\lambda_3 = -1$. By Proposition 17.1.1 the set of values of $R_{\mathbf{A}}$ is the whole interval $[-1, 3]$. Three things go wrong at once.

- **There is no second largest value.** The values below 3 form the interval $[-1, 3)$, which has no largest element. So the phrase does not describe a number.
- **The value λ_2 is attained, but so is every other value, and not only at eigenvectors.** At $\mathbf{x} = \frac{1}{\sqrt{2}}(1, 0, 1)$ the quotient is $\frac{1}{2}(3) + \frac{1}{2}(-1) = 1 = \lambda_2(\mathbf{A})$, and $\mathbf{x}$ is not an eigenvector of $\mathbf{A}$: it mixes the eigenvectors $\mathbf{e}_1$ and $\mathbf{e}_3$ of the first and the *third* eigenvalues, and λ_2 plays no part in it.
- **Removing a direction does not help by itself.** At $\mathbf{y} = \frac{1}{\sqrt{2}}(1, 1, 0)$, which is orthogonal to the eigenvector $\mathbf{e}_3$ of λ_3, the quotient is $\frac{1}{2}(3) + \frac{1}{2}(1) = 2$, strictly between λ_1 and λ_2. Constraining $\mathbf{x}$ to a subspace changes the answer, and which subspace is everything.

Whatever describes λ_2 must be a maximum over something smaller than the sphere.

The last bullet says what to try. Cutting the sphere down by orthogonality to the top eigenvector does work, and the proof is Example (A diagonal matrix, and the whole chapter in miniature) again, with the top entry removed from the average.

Proposition 17.1.3. Eigenvalues by Deflation

Let $\mathbf{A} \in M_n(F)$ be Hermitian and let $(\mathbf{q}_1, \dots, \mathbf{q}_n)$ be an orthonormal basis of F^n with $\mathbf{A}\mathbf{q}_i = \lambda_i(\mathbf{A})\mathbf{q}_i$ for every i. Then for every k with $1 \le k \le n$,

$$\lambda_k(\mathbf{A}) = \max\{R_{\mathbf{A}}(\mathbf{x}) : \mathbf{x} \neq \mathbf{0},\ \mathbf{x} \perp \mathbf{q}_1, \dots, \mathbf{q}_{k-1}\},$$

the maximum being attained at $\mathbf{x} = \mathbf{q}_k$. For $k = 1$ the list of constraints is empty and the statement is Proposition 17.1.1.

Idea. The constraints do one thing: they delete the first $k-1$ coordinates in the eigenbasis. What is left is the weighted average of Example (A diagonal matrix, and the whole chapter in miniature) over the remaining eigenvalues $\lambda_k \ge \cdots \ge \lambda_n$, whose largest member is λ_k.

Proof. Let $W = \{\mathbf{x} \in F^n : \langle \mathbf{x}, \mathbf{q}_i \rangle = 0 \text{ for } 1 \le i \le k-1\}$. By Theorem 11.2.1 (a) every $\mathbf{x} \in F^n$ satisfies $\mathbf{x} = \sum_{i=1}^{n} \langle \mathbf{x}, \mathbf{q}_i \rangle \mathbf{q}_i$, so $\mathbf{x} \in W$ exactly when $\mathbf{x} \in \operatorname{span}(\mathbf{q}_k, \dots, \mathbf{q}_n)$.

Let $\mathbf{x} \in W$ with $\mathbf{x} \neq \mathbf{0}$, and write $c_i = \langle \mathbf{x}, \mathbf{q}_i \rangle$, so that $\mathbf{x} = \sum_{i \ge k} c_i \mathbf{q}_i$. Then $\mathbf{A}\mathbf{x} = \sum_{i \ge k} \lambda_i(\mathbf{A}) c_i \mathbf{q}_i$, and Theorem 11.2.1 (b), (c) give

$$\langle \mathbf{A}\mathbf{x}, \mathbf{x} \rangle = \sum_{i \ge k} \lambda_i(\mathbf{A}) |c_i|^2, \qquad \|\mathbf{x}\|^2 = \sum_{i \ge k} |c_i|^2 > 0.$$

Hence $R_{\mathbf{A}}(\mathbf{x})$ is a weighted average of $\lambda_k(\mathbf{A}), \dots, \lambda_n(\mathbf{A})$ with nonnegative weights $|c_i|^2 / \|\mathbf{x}\|^2$ summing to 1. The largest of those eigenvalues is $\lambda_k(\mathbf{A})$, because the list is indexed decreasingly, so $R_{\mathbf{A}}(\mathbf{x}) \le \lambda_k(\mathbf{A})$.

Finally $\mathbf{q}_k \in W$, since the $\mathbf{q}_i$ are orthonormal, and $R_{\mathbf{A}}(\mathbf{q}_k) = \langle \lambda_k(\mathbf{A})\mathbf{q}_k, \mathbf{q}_k \rangle = \lambda_k(\mathbf{A})$. So the bound is attained and the supremum is a maximum. This proves the proposition. ∎

Quick check. In Proposition 17.1.3, does the value of the maximum depend on which orthonormal eigenbasis $(\mathbf{q}_1, \dots, \mathbf{q}_n)$ is chosen? Does the *maximizing vector*?

Solution. The value does not: it is $\lambda_k(\mathbf{A})$, a number attached to $\mathbf{A}$ alone. The maximizing vector generally does, and so does the constraint set W. For $\mathbf{A} = \operatorname{diag}(1, 1, 0)$ and $k = 2$, one admissible eigenbasis is $(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$, giving $W = \operatorname{span}(\mathbf{e}_2, \mathbf{e}_3)$ and maximizer $\mathbf{e}_2$; another is $(\frac{1}{\sqrt{2}}(1,1,0), \frac{1}{\sqrt{2}}(1,-1,0), \mathbf{e}_3)$, giving a different W and the maximizer $\frac{1}{\sqrt{2}}(1,-1,0)$. Both maxima equal $\lambda_2(\mathbf{A}) = 1$.

17.1.5 What is still wrong

Proposition 17.1.3 is a complete answer to the question "what is $\lambda_k(\mathbf{A})$, as an optimization?" and it is nearly useless. The reason is visible in its statement: the constraint set is built from $\mathbf{q}_1, \dots, \mathbf{q}_{k-1}$, which are eigenvectors of $\mathbf{A}$. To use the formula one must already know the eigenvectors, and to know the eigenvectors one must already have solved the problem the formula claims to describe. It is a description of $\lambda_k(\mathbf{A})$ in terms of $\mathbf{A}$'s own answer.

The cost is not merely aesthetic. The whole point of turning an eigenvalue into an optimization is to compare the eigenvalues of **different** matrices: to say that $\mathbf{A} \succeq \mathbf{B}$ forces $\lambda_i(\mathbf{A}) \geq \lambda_i(\mathbf{B})$, or that perturbing $\mathbf{A}$ moves λ_i only a little. A comparison needs the two optimizations to run over the *same* set. Deflation gives $\lambda_i(\mathbf{A})$ as a maximum over the orthogonal complement of $\mathbf{A}$'s top eigenvectors, and $\lambda_i(\mathbf{B})$ as a maximum over the orthogonal complement of $\mathbf{B}$'s top eigenvectors, and those two subspaces have nothing to do with each other. Nothing can be concluded.

Chapter 13 §13 promised that "Chapter 17 characterizes the eigenvalues of a Hermitian matrix as maxima and minima of $\langle \mathbf{Ax}, \mathbf{x} \rangle$ — at which point the test becomes the theory". Proposition 17.1.1 and Proposition 17.1.3 are the first installment of that, and they show that the promise can be kept. The test has already become a statement about a single number: $\mathbf{A} \succ \mathbf{O}$ exactly when $\lambda_n(\mathbf{A}) = \min_{\mathbf{x} \neq \mathbf{0}} R_{\mathbf{A}}(\mathbf{x})$ is positive, and $\mathbf{A} \succeq \mathbf{O}$ exactly when it is nonnegative, which is the eigenvalue criterion of Theorem 13.1.5 and Theorem 13.1.4 read through Proposition 17.1.1. The rest of it is the removal of the eigenvectors from the statement, which §02 achieves by a device that looks extravagant and is exactly right: keep the maximum over a subspace, but let the subspace vary over **all** subspaces of the appropriate dimension, and take the minimum of what comes out.

17.1.6 Exercises

17.1.6.1 A. Check your understanding

Exercise. A1

(a) State Definition 17.1.1, naming both hypotheses, and say what each one is for.

(b) Prove that $R_{\mathbf{A}}(c\mathbf{x}) = R_{\mathbf{A}}(\mathbf{x})$ for every scalar $c \neq 0$, and explain why this makes $R_{\mathbf{A}}$ a function on lines.

(c) Determine whether the following is correct, and justify your answer: if $\mathbf{A} \in M_3(\mathbb{R})$ is symmetric with three distinct eigenvalues, then $R_{\mathbf{A}}$ takes exactly three values.

(d) State Proposition 17.1.3, and say in one sentence why it is not a satisfactory description of $\lambda_k(\mathbf{A})$.

(e) Give a Hermitian $\mathbf{A}$ and a vector $\mathbf{x}$ that is **not** an eigenvector of $\mathbf{A}$ but has $R_{\mathbf{A}}(\mathbf{x})$ equal to an eigenvalue of $\mathbf{A}$.

Solution.

(a) Let F be $\mathbb{R}$ or $\mathbb{C}$ and $\mathbf{A} \in M_n(F)$ Hermitian. For $\mathbf{x} \neq \mathbf{0}$, the Rayleigh quotient is $R_{\mathbf{A}}(\mathbf{x}) = \langle \mathbf{Ax}, \mathbf{x} \rangle / \langle \mathbf{x}, \mathbf{x} \rangle$. Hermitian makes the numerator real (Proposition 11.6.7 (a)) and ties the values to the spectrum; $\mathbf{x} \neq \mathbf{0}$ makes the denominator nonzero (Definition 11.1.1).

(b) $\langle \mathbf{A}(c\mathbf{x}), c\mathbf{x} \rangle = c\bar{c}\langle \mathbf{Ax}, \mathbf{x} \rangle = |c|^2 \langle \mathbf{Ax}, \mathbf{x} \rangle$ and likewise $\langle c\mathbf{x}, c\mathbf{x} \rangle = |c|^2 \langle \mathbf{x}, \mathbf{x} \rangle$; the factor $|c|^2 \neq 0$ cancels. So $R_{\mathbf{A}}$ is constant on $\{c\mathbf{x} : c \neq 0\}$, the punctured line through $\mathbf{x}$, and

its value depends only on which line $\mathbf{x}$ lies on.

(c) Incorrect. By Proposition 17.1.1 the set of values is the whole of the interval $[\lambda_3(\mathbf{A}), \lambda_1(\mathbf{A})]$, which is infinite whenever $\lambda_1(\mathbf{A}) \neq \lambda_3(\mathbf{A})$, and distinct eigenvalues guarantee that.

(d) With $(\mathbf{q}_1, \dots, \mathbf{q}_n)$ an orthonormal eigenbasis ordered so that $\mathbf{A}\mathbf{q}_i = \lambda_i(\mathbf{A})\mathbf{q}_i$, $\lambda_k(\mathbf{A})$ is the maximum of $R_{\mathbf{A}}$ over the nonzero vectors orthogonal to $\mathbf{q}_1, \dots, \mathbf{q}_{k-1}$. It is unsatisfactory because the constraint set is built from eigenvectors of $\mathbf{A}$, so the formula presupposes the answer and cannot compare two different matrices.

(e) $\mathbf{A} = \operatorname{diag}(3, 1, -1)$ and $\mathbf{x} = (1, 0, 1)$: $R_{\mathbf{A}}(\mathbf{x}) = (3 - 1)/2 = 1 = \lambda_2(\mathbf{A})$, while $\mathbf{A}\mathbf{x} = (3, 0, -1)$ is not a multiple of $\mathbf{x}$. (Any mixture of a top and a bottom eigenvector in the right proportion does the same job.)

17.1.6.2 B. Practice

Exercise. B1: Locating four values

Let

$$\mathbf{A} = \begin{pmatrix} 2 & 2 & 0 \\ 2 & 3 & 2 \\ 0 & 2 & 4 \end{pmatrix} \in M_3(\mathbb{R}).$$

Its eigenvalues are $6, 3, 0$, with eigenvectors $(1, 2, 2)$, $(2, 1, -2)$, and $(2, -2, 1)$ respectively; you may take this as given. Compute $R_{\mathbf{A}}(\mathbf{x})$ for $\mathbf{x} = \mathbf{e}_1$, $\mathbf{x} = (1, 1, 1)$, $\mathbf{x} = (1, 1, 0)$, and $\mathbf{x} = (1, 0, -1)$, and check each value against Proposition 17.1.1. Hence exhibit a noneigenvector at which $R_{\mathbf{A}}$ equals an eigenvalue.

Solution. The three given vectors are pairwise orthogonal, each of norm 3, so $\mathbf{q}_1 = \frac{1}{3}(1, 2, 2)$, $\mathbf{q}_2 = \frac{1}{3}(2, 1, -2)$, $\mathbf{q}_3 = \frac{1}{3}(2, -2, 1)$ is an orthonormal eigenbasis with $\lambda_1 = 6$, $\lambda_2 = 3$, $\lambda_3 = 0$. By Proposition 17.1.1 every value must lie in $[0, 6]$.

For $\mathbf{x} = \mathbf{e}_1$: $\mathbf{A}\mathbf{e}_1 = (2, 2, 0)$, so $\langle \mathbf{A}\mathbf{x}, \mathbf{x} \rangle = 2$ and $\|\mathbf{x}\|^2 = 1$, giving $R_{\mathbf{A}}(\mathbf{e}_1) = 2$. (This is the diagonal entry a_{11}, as it must be for a standard basis vector.)

For $\mathbf{x} = (1, 1, 1)$: $\mathbf{A}\mathbf{x} = (4, 7, 6)$ and $\langle \mathbf{A}\mathbf{x}, \mathbf{x} \rangle = 4 + 7 + 6 = 17$, while $\|\mathbf{x}\|^2 = 3$; so $R_{\mathbf{A}}(\mathbf{x}) = 17/3 \approx 5.67$.

For $\mathbf{x} = (1, 1, 0)$: $\mathbf{A}\mathbf{x} = (4, 5, 2)$ and $\langle \mathbf{A}\mathbf{x}, \mathbf{x} \rangle = 9$, $\|\mathbf{x}\|^2 = 2$; so $R_{\mathbf{A}}(\mathbf{x}) = 9/2$.

For $\mathbf{x} = (1, 0, -1)$: $\mathbf{A}\mathbf{x} = (2, 0, -4)$ and $\langle \mathbf{A}\mathbf{x}, \mathbf{x} \rangle = 2 + 4 = 6$, $\|\mathbf{x}\|^2 = 2$; so $R_{\mathbf{A}}(\mathbf{x}) = 3$.

All four values lie in $[0, 6]$. The last one equals $\lambda_2(\mathbf{A}) = 3$, yet $\mathbf{A}(1, 0, -1) = (2, 0, -4)$ is not a multiple of $(1, 0, -1)$, so $(1, 0, -1)$ is not an eigenvector. This is the trap of the warning above, in a concrete matrix.

Exercise. B2: A prescribed value

With $\mathbf{A}$ and the orthonormal eigenbasis of Exercise (B1: Locating four values), use the construction in the proof of Proposition 17.1.1 to produce a unit vector $\mathbf{x}$ with $R_{\mathbf{A}}(\mathbf{x}) = 2$.

Then find, by mixing $\mathbf{q}_1$ and $\mathbf{q}_2$ in equal proportion, a vector with integer entries whose Rayleigh quotient is $9/2$.

Solution. For the first, $\mu = 2$ with $\lambda_1 = 6$ and $\lambda_3 = 0$ gives $t = (2-0)/(6-0) = 1/3$, so

$$\mathbf{x} = \sqrt{\tfrac{1}{3}}\,\mathbf{q}_1 + \sqrt{\tfrac{2}{3}}\,\mathbf{q}_3,$$

a unit vector with $R_{\mathbf{A}}(\mathbf{x}) = \frac{1}{3}\cdot 6 + \frac{2}{3}\cdot 0 = 2$.

For the second, $\mathbf{y} = \mathbf{q}_1 + \mathbf{q}_2 = \frac{1}{3}((1,2,2) + (2,1,-2)) = (1,1,0)$, whose Rayleigh quotient is the average $\frac{1}{2}(6) + \frac{1}{2}(3) = 9/2$ — matching the direct computation in Exercise (B1: Locating four values). Equal proportions of two orthonormal eigenvectors give the plain average of their eigenvalues.

Exercise. B3: The quotient of a projection

Let $U = \operatorname{span}\{(1,1,0),(0,0,1)\} \subseteq \mathbb{R}^3$ and let $\mathbf{P}$ be the matrix of the orthogonal projection onto U.

(a) Compute $\mathbf{P}$, and list the eigenvalues of $\mathbf{P}$ with multiplicity.

(b) Compute $R_{\mathbf{P}}(\mathbf{x})$ for $\mathbf{x} = \mathbf{e}_1$, $\mathbf{x} = (1,-1,0)$, and $\mathbf{x} = (1,-1,1)$, and check each against Proposition 17.1.1.

Solution.

(a) The two spanning vectors are orthogonal, of norms $\sqrt{2}$ and 1, so $\mathbf{u}_1 = \frac{1}{\sqrt{2}}(1,1,0)$ and $\mathbf{u}_2 = (0,0,1)$ form an orthonormal basis of U and $\mathbf{P} = \mathbf{u}_1\mathbf{u}_1^{\top} + \mathbf{u}_2\mathbf{u}_2^{\top}$, that is

$$\mathbf{P} = \begin{pmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 1/2 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Its eigenvalues are 1 with multiplicity 2 (on U) and 0 with multiplicity 1 (on $U^{\perp} = \operatorname{span}\{(1,-1,0)\}$), so $\lambda_1 = \lambda_2 = 1$ and $\lambda_3 = 0$.

(b) By Example (Two matrices whose quotient is forced) (b), $R_{\mathbf{P}}(\mathbf{x}) = \|\mathbf{P}\mathbf{x}\|^2/\|\mathbf{x}\|^2$. For $\mathbf{x} = \mathbf{e}_1$: $\mathbf{P}\mathbf{e}_1 = (\frac{1}{2},\frac{1}{2},0)$, so $\|\mathbf{P}\mathbf{x}\|^2 = \frac{1}{2}$ and $R_{\mathbf{P}}(\mathbf{e}_1) = \frac{1}{2}$. For $\mathbf{x} = (1,-1,0)$: $\mathbf{P}\mathbf{x} = \mathbf{0}$, so $R_{\mathbf{P}}(\mathbf{x}) = 0$. For $\mathbf{x} = (1,-1,1)$: $\mathbf{P}\mathbf{x} = (0,0,1)$, so $\|\mathbf{P}\mathbf{x}\|^2 = 1$ and $\|\mathbf{x}\|^2 = 3$, giving $R_{\mathbf{P}}(\mathbf{x}) = 1/3$. All three lie in $[\lambda_3, \lambda_1] = [0,1]$, and the middle one attains the lower end, as it must, being a nonzero vector of $U^{\perp}$.

17.1.6.3 C. Going deeper

Exercise. C1: Why the quotient estimates eigenvalues so well

Let $\mathbf{A} \in M_n(F)$ be Hermitian with orthonormal eigenbasis $(\mathbf{q}_1, \dots, \mathbf{q}_n)$, $\mathbf{A}\mathbf{q}_i = \lambda_i(\mathbf{A})\mathbf{q}_i$.

(a) Prove that for every unit vector $\mathbf{x}$,

$$\lambda_1(\mathbf{A}) - R_{\mathbf{A}}(\mathbf{x}) \;\le\; (\lambda_1(\mathbf{A}) - \lambda_n(\mathbf{A}))\,\|\mathbf{x} - \mathbf{q}_1\|^2.$$

(b) Deduce that if $\|\mathbf{x} - \mathbf{q}_1\| \le \varepsilon$ then $R_{\mathbf{A}}(\mathbf{x})$ approximates $\lambda_1(\mathbf{A})$ with error at most $(\lambda_1 - \lambda_n)\varepsilon^2$: an error in the eigen*vector* of size ε costs only ε^2 in the eigen*value*.

(c) Show that the exponent 2 cannot be improved, by computing both sides for $\mathbf{A} = \operatorname{diag}(1, 0)$ and $\mathbf{x} = (\cos\varepsilon, \sin\varepsilon)$.

Hint for (a): expand $\mathbf{x}$ *in the eigenbasis and compare the two sums term by term.*

Solution.

(a) Write $c_i = \langle \mathbf{x}, \mathbf{q}_i \rangle$, so $\sum_i |c_i|^2 = 1$ by Theorem 11.2.1 (c). As in Proposition 17.1.3, $R_{\mathbf{A}}(\mathbf{x}) = \sum_i \lambda_i(\mathbf{A})|c_i|^2$, so

$$\lambda_1(\mathbf{A}) - R_{\mathbf{A}}(\mathbf{x}) = \sum_{i=1}^{n} (\lambda_1(\mathbf{A}) - \lambda_i(\mathbf{A}))|c_i|^2 \le (\lambda_1(\mathbf{A}) - \lambda_n(\mathbf{A})) \sum_{i \ge 2} |c_i|^2,$$

where the $i = 1$ term was dropped because it is 0 and each remaining factor $\lambda_1 - \lambda_i$ was enlarged to $\lambda_1 - \lambda_n$, legitimate since $|c_i|^2 \ge 0$ and the eigenvalues decrease. On the other side, again by Theorem 11.2.1 (c),

$$\|\mathbf{x} - \mathbf{q}_1\|^2 = |c_1 - 1|^2 + \sum_{i \ge 2} |c_i|^2 \;\ge\; \sum_{i \ge 2} |c_i|^2.$$

Combining the two displays gives the inequality.

(b) Immediate from (a), since the right-hand side is increasing in $\|\mathbf{x} - \mathbf{q}_1\|$.

(c) Here $\lambda_1 = 1$, $\lambda_2 = 0$, and $\mathbf{q}_1 = \mathbf{e}_1$. The vector $\mathbf{x} = (\cos\varepsilon, \sin\varepsilon)$ is a unit vector with

$$R_{\mathbf{A}}(\mathbf{x}) = \cos^2\varepsilon, \qquad \lambda_1 - R_{\mathbf{A}}(\mathbf{x}) = \sin^2\varepsilon,$$

while $\|\mathbf{x} - \mathbf{e}_1\|^2 = (\cos\varepsilon - 1)^2 + \sin^2\varepsilon = 2 - 2\cos\varepsilon$. The bound in (a) reads $\sin^2\varepsilon \le 2 - 2\cos\varepsilon$, which is true, and the ratio of the two sides is

$$\frac{\sin^2\varepsilon}{2 - 2\cos\varepsilon} = \frac{(1 - \cos\varepsilon)(1 + \cos\varepsilon)}{2(1 - \cos\varepsilon)} = \frac{1 + \cos\varepsilon}{2},$$

which tends to 1 as $\varepsilon \to 0$. So the error really is of order ε^2, not smaller, and the constant in (a) is sharp in the limit. This second-order accuracy is why the Rayleigh quotient is the standard way to read an eigenvalue off an approximate eigenvector, and Chapter 24 builds an iteration on it.

Exercise. C2: The equality case at the top

Let $\mathbf{A} \in M_n(F)$ be Hermitian and let $E_{\lambda_1}(\mathbf{A}) = \ker(\mathbf{A} - \lambda_1(\mathbf{A})\mathbf{I})$ be the eigenspace of the largest eigenvalue.

(a) Prove that if $\mathbf{x} \neq \mathbf{0}$ and $R_{\mathbf{A}}(\mathbf{x}) = \lambda_1(\mathbf{A})$, then $\mathbf{x} \in E_{\lambda_1}(\mathbf{A})$.

(b) Deduce that $\{\mathbf{x} \neq \mathbf{0} : R_{\mathbf{A}}(\mathbf{x}) = \lambda_1(\mathbf{A})\} \cup \{\mathbf{0}\} = E_{\lambda_1}(\mathbf{A})$.

(c) Show that the corresponding statement for an intermediate eigenvalue is **false**, by

naming a matrix, an index k with $1 < k < n$, and a vector.

Solution.

(a) We may assume $\|\mathbf{x}\| = 1$, by scale invariance and because $E_{\lambda_1}(\mathbf{A})$ is a subspace. Write $c_i = \langle \mathbf{x}, \mathbf{q}_i \rangle$ for an orthonormal eigenbasis as above, so $\sum_i |c_i|^2 = 1$ and $R_{\mathbf{A}}(\mathbf{x}) = \sum_i \lambda_i(\mathbf{A})|c_i|^2$. Then

$$0 = \lambda_1(\mathbf{A}) - R_{\mathbf{A}}(\mathbf{x}) = \sum_{i=1}^{n} (\lambda_1(\mathbf{A}) - \lambda_i(\mathbf{A}))|c_i|^2,$$

a sum of nonnegative terms. Every term is therefore 0, so $c_i = 0$ whenever $\lambda_i(\mathbf{A}) < \lambda_1(\mathbf{A})$. Hence $\mathbf{x} = \sum_{i:\lambda_i=\lambda_1} c_i \mathbf{q}_i$, a combination of eigenvectors for $\lambda_1(\mathbf{A})$, so $\mathbf{x} \in E_{\lambda_1}(\mathbf{A})$.

(b) ($\subseteq$) is (a), together with $\mathbf{0} \in E_{\lambda_1}(\mathbf{A})$. ($\supseteq$) If $\mathbf{0} \neq \mathbf{x} \in E_{\lambda_1}(\mathbf{A})$ then $\mathbf{A}\mathbf{x} = \lambda_1(\mathbf{A})\mathbf{x}$, so $R_{\mathbf{A}}(\mathbf{x}) = \lambda_1(\mathbf{A})\langle \mathbf{x}, \mathbf{x} \rangle / \langle \mathbf{x}, \mathbf{x} \rangle = \lambda_1(\mathbf{A})$.

(c) Take $\mathbf{A} = \operatorname{diag}(3, 1, -1)$, $k = 2$, and $\mathbf{x} = (1, 0, 1)$. Then $R_{\mathbf{A}}(\mathbf{x}) = 1 = \lambda_2(\mathbf{A})$, but $\mathbf{A}\mathbf{x} = (3, 0, -1)$ is not a multiple of $\mathbf{x}$, so $\mathbf{x} \notin E_{\lambda_2}(\mathbf{A}) = \operatorname{span}\{\mathbf{e}_2\}$. The argument in (a) breaks because the terms $(\lambda_2 - \lambda_i)|c_i|^2$ are no longer all of the same sign: the $i = 1$ term is negative and cancels against the $i = 3$ term instead of forcing both to vanish.

Exercise. C3: What the quotient cannot see

Let $\mathbf{M} \in M_n(F)$ be an arbitrary matrix, not assumed Hermitian, and put

$$\mathbf{A} = \tfrac{1}{2}(\mathbf{M} + \mathbf{M}^*), \qquad \mathbf{B} = \tfrac{1}{2}(\mathbf{M} - \mathbf{M}^*),$$

so that $\mathbf{M} = \mathbf{A} + \mathbf{B}$ with $\mathbf{A}^* = \mathbf{A}$ and $\mathbf{B}^* = -\mathbf{B}$.

(a) Prove that $\operatorname{Re}\langle \mathbf{M}\mathbf{x}, \mathbf{x} \rangle = \langle \mathbf{A}\mathbf{x}, \mathbf{x} \rangle$ for every $\mathbf{x} \in F^n$.

(b) Deduce that over $\mathbb{R}$, the function $\mathbf{x} \mapsto \langle \mathbf{M}\mathbf{x}, \mathbf{x} \rangle$ determines $\mathbf{A}$ and nothing whatever about $\mathbf{B}$. Give two distinct real 2×2 matrices with the same such function.

(c) Over $\mathbb{C}$, determine whether $\langle \mathbf{M}\mathbf{x}, \mathbf{x} \rangle = 0$ for all $\mathbf{x}$ forces $\mathbf{M} = \mathbf{O}$. Justify your answer.

Solution.

(a) By conjugate symmetry and the definition of the adjoint, $\langle \mathbf{M}^*\mathbf{x}, \mathbf{x} \rangle = \langle \mathbf{x}, \mathbf{M}\mathbf{x} \rangle = \overline{\langle \mathbf{M}\mathbf{x}, \mathbf{x} \rangle}$. Hence

$$\langle \mathbf{A}\mathbf{x}, \mathbf{x} \rangle = \tfrac{1}{2}\big(\langle \mathbf{M}\mathbf{x}, \mathbf{x} \rangle + \overline{\langle \mathbf{M}\mathbf{x}, \mathbf{x} \rangle}\big) = \operatorname{Re}\langle \mathbf{M}\mathbf{x}, \mathbf{x} \rangle.$$

(b) Over $\mathbb{R}$ every inner product value is real, so (a) reads $\langle \mathbf{M}\mathbf{x}, \mathbf{x} \rangle = \langle \mathbf{A}\mathbf{x}, \mathbf{x} \rangle$ for all $\mathbf{x}$: the function sees only $\mathbf{A}$. It determines $\mathbf{A}$ completely, since two symmetric matrices with the same quadratic form are equal (Theorem 12.2.5 applied to their difference). It sees nothing of $\mathbf{B}$: for

$$\mathbf{M}_1 = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, \qquad \mathbf{M}_2 = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix},$$

both functions are identically 0, since $\langle \mathbf{M}_2\mathbf{x}, \mathbf{x} \rangle = -x_2x_1 + x_1x_2 = 0$, yet $\mathbf{M}_1 \neq \mathbf{M}_2$.

This is Exercise (C3: What the real quadratic form does determine) in matrix language.

(c) Yes, over $\mathbb{C}$ it does force $\mathbf{M} = \mathbf{O}$. This is Theorem 12.2.6, and the reason the complex theory is the cleaner one: a complex quadratic form determines its matrix, so nothing is invisible. Concretely, (a) gives $\mathbf{A} = \mathbf{O}$, and applying the hypothesis to $i\mathbf{M}$ — or expanding $\langle \mathbf{M}(\mathbf{x}+\mathbf{y}), \mathbf{x}+\mathbf{y}\rangle$ and $\langle \mathbf{M}(\mathbf{x}+i\mathbf{y}), \mathbf{x}+i\mathbf{y}\rangle$ — kills $\mathbf{B}$ as well. The real 2×2 matrix $\mathbf{M}_2$ of (b) shows that the same statement over $\mathbb{R}$ is false, so this is genuinely a fact about the field.

17.2 The Min–Max Theorem

Section 1 ended with a description of $\lambda_k(\mathbf{A})$ that works and cannot be used: a maximum of the Rayleigh quotient over the orthogonal complement of $\mathbf{A}$'s own top eigenvectors. The defect is that the constraint set is built from the answer. This section repairs it by a move that looks like an extravagance and is exactly the right one: keep the maximum over a subspace, stop insisting that the subspace be the right one, and let it range over **all** subspaces of the appropriate dimension. The minimum of what comes out is again $\lambda_k(\mathbf{A})$, and no eigenvector appears anywhere in the statement. That is the Courant–Fischer theorem, and everything in the rest of the chapter is a consequence of it.

Throughout, F is $\mathbb{R}$ or $\mathbb{C}$ and every matrix is Hermitian. The eigenvalues of a Hermitian $\mathbf{A} \in M_n(F)$ are real (Theorem 12.2.1) and are **indexed decreasingly**,

$$\lambda_1(\mathbf{A}) \;\geq\; \lambda_2(\mathbf{A}) \;\geq\; \dots \;\geq\; \lambda_n(\mathbf{A}),$$

each repeated according to multiplicity, so that the list has exactly n entries. This is the book's convention, fixed in Chapter 13 §05, and it is used without exception below. Many sources index the theorem of this section the other way round; the remark after the theorem gives that form, once.

17.2.1 Two subspaces that have to meet

One dimension count does all the work in this chapter, and it is worth a label of its own, because four later proofs cite it.

Two subspaces of an n-dimensional space cannot both be large and disjoint: if their dimensions add to more than n, they share a nonzero vector.

Lemma 17.2.1. Large Subspaces Intersect

Let V be a vector space with $\dim V = n$ over any field, and let U and W be subspaces of V with

$$\dim U + \dim W > n.$$

Then $U \cap W \neq \{0\}$. More precisely, $\dim(U \cap W) \geq \dim U + \dim W - n$.

Proof. By Theorem 1.7.3,

$$\dim(U + W) = \dim U + \dim W - \dim(U \cap W).$$

Both U and W are finite-dimensional, being subspaces of the n-dimensional V (Theorem 1.6.7); $U + W$ is a subspace of V (Theorem 1.7.1), so the same theorem gives $\dim(U +$

$W) \le n$, and rearranging gives

$$\dim(U \cap W) \;\ge\; \dim U + \dim W - n,$$

which is > 0 by hypothesis. A subspace of positive dimension contains a nonzero vector. This proves the lemma.

■

Remark. The bound is sharp, and the hypothesis cannot be weakened to $\dim U + \dim W = n$: two distinct lines in $\mathbb{R}^2$ meet only at $\mathbf{0}$, and so do a line and a plane in $\mathbb{R}^3$ chosen in general position. What makes the count useful is that it needs nothing about the two subspaces beyond their dimensions — in every application below, one of them is a span of eigenvectors that we choose and the other is an arbitrary competitor we know nothing about.

17.2.2 The quotient on a subspace

Before quantifying over subspaces, we must know that the inner maximum and minimum exist. They do, and for an algebraic reason: restricting the quadratic form to a subspace produces a smaller Hermitian matrix, to which Proposition 17.1.1 applies verbatim.

Lemma 17.2.2. The Rayleigh Quotient on a Subspace

Let $\mathbf{A} \in M_n(F)$ be Hermitian and let $W \subseteq F^n$ be a subspace with $\dim W = m \ge 1$. Let $\mathbf{Q} \in M_{n\times m}(F)$ be a matrix whose columns form an orthonormal basis of W, and put $\mathbf{B} = \mathbf{Q}^*\mathbf{A}\mathbf{Q} \in M_m(F)$. Then $\mathbf{B}$ is Hermitian and

$$\{R_{\mathbf{A}}(\mathbf{x}) : \mathbf{0} \ne \mathbf{x} \in W\} = [\lambda_m(\mathbf{B}),\ \lambda_1(\mathbf{B})].$$

In particular $\max_{\mathbf{0}\ne\mathbf{x}\in W} R_{\mathbf{A}}(\mathbf{x})$ and $\min_{\mathbf{0}\ne\mathbf{x}\in W} R_{\mathbf{A}}(\mathbf{x})$ exist and are attained.

Idea. Coordinates on W. Writing a vector of W as $\mathbf{Q}\mathbf{c}$ turns $\langle \mathbf{A}\mathbf{x}, \mathbf{x}\rangle$ into $\langle \mathbf{B}\mathbf{c}, \mathbf{c}\rangle$ and, because the columns of $\mathbf{Q}$ are orthonormal, turns $\langle \mathbf{x}, \mathbf{x}\rangle$ into $\langle \mathbf{c}, \mathbf{c}\rangle$ with no distortion. So the Rayleigh quotient of $\mathbf{A}$ on W *is* the Rayleigh quotient of $\mathbf{B}$ on F^m, and Proposition 17.1.1 answers the question there.

Proof. Such a $\mathbf{Q}$ exists: W has an orthonormal basis by Theorem 11.2.2, and we take its members as columns. Orthonormality of the columns says exactly $\mathbf{Q}^*\mathbf{Q} = \mathbf{I}_m$. Then $\mathbf{B}^* = \mathbf{Q}^*\mathbf{A}^*\mathbf{Q} = \mathbf{Q}^*\mathbf{A}\mathbf{Q} = \mathbf{B}$, so $\mathbf{B}$ is Hermitian and Proposition 17.1.1 applies to it.

The map $\mathbf{c} \mapsto \mathbf{Q}\mathbf{c}$ sends F^m onto W, since the columns of $\mathbf{Q}$ span W, and it is injective, since $\mathbf{Q}\mathbf{c} = \mathbf{0}$ gives $\mathbf{c} = \mathbf{Q}^*\mathbf{Q}\mathbf{c} = \mathbf{0}$. So it is a bijection $F^m \to W$ carrying $\mathbf{0}$ to $\mathbf{0}$, hence a bijection from the nonzero vectors of F^m onto those of W. For $\mathbf{c} \ne \mathbf{0}$, writing the standard inner product as $\langle \mathbf{u}, \mathbf{v}\rangle = \mathbf{v}^*\mathbf{u}$,

$$\langle \mathbf{A}\mathbf{Q}\mathbf{c}, \mathbf{Q}\mathbf{c}\rangle = \mathbf{c}^*\mathbf{Q}^*\mathbf{A}\mathbf{Q}\mathbf{c} = \langle \mathbf{B}\mathbf{c}, \mathbf{c}\rangle,$$
$$\langle \mathbf{Q}\mathbf{c}, \mathbf{Q}\mathbf{c}\rangle = \mathbf{c}^*\mathbf{Q}^*\mathbf{Q}\mathbf{c} = \langle \mathbf{c}, \mathbf{c}\rangle,$$

so $R_{\mathbf{A}}(\mathbf{Qc}) = R_{\mathbf{B}}(\mathbf{c})$. Therefore the two value sets $\{R_{\mathbf{A}}(\mathbf{x}) : \mathbf{0} \neq \mathbf{x} \in W\}$ and $\{R_{\mathbf{B}}(\mathbf{c}) : \mathbf{0} \neq \mathbf{c} \in F^m\}$ are equal, and the second is $[\lambda_m(\mathbf{B}), \lambda_1(\mathbf{B})]$ with both ends attained, by Proposition 17.1.1. This proves the lemma.

■

Remark. **Which analysis was invoked: none.** The existence of a maximum of a function over an infinite set is normally an analytic fact — the set of unit vectors of W is compact, by fact (A3) of Chapter 16's introduction, and $R_{\mathbf{A}}$ is continuous on it, so the extreme value theorem, fact (A4), applies. That route is available and correct. It is not needed. The spectral theorem, already proved in Chapter 12, produces the maximizing vector by name, and Lemma 17.2.2 inherits that. Nothing in this section, or in §01, uses an item of Chapter 16's imported-analysis list. The analysis has not vanished, only moved: the spectral theorem rests on the fundamental theorem of algebra (Theorem 6.5.1), whose proof in Chapter 6 §05 uses the extreme value theorem on a closed disc.

17.2.3 Every eigenvalue is a min–max

Theorem 17.2.3. Courant–Fischer Min–Max Theorem

Let F be $\mathbb{R}$ or $\mathbb{C}$, let $\mathbf{A} \in M_n(F)$ be Hermitian, and let $1 \leq k \leq n$. Then

$$\begin{aligned} \lambda_k(\mathbf{A}) &= \min_{\dim W = n-k+1} \; \max_{\mathbf{0} \neq \mathbf{x} \in W} R_{\mathbf{A}}(\mathbf{x}) \\ &= \max_{\dim W = k} \; \min_{\mathbf{0} \neq \mathbf{x} \in W} R_{\mathbf{A}}(\mathbf{x}), \end{aligned}$$

where in each line W runs over **all** subspaces of F^n of the stated dimension. All four extrema are attained.

Idea. Read the second line first, because it is the one with a slogan: λ_k *is the best value you can guarantee on a* k*-dimensional subspace.* On a k-dimensional subspace the quotient has a smallest value, which is what that subspace guarantees; the span of the top k eigenvectors guarantees λ_k, and no other k-dimensional subspace guarantees more.

The proof is three steps, and the first two have entirely different characters. **Step 1** exhibits a **witness** subspace and computes on it, giving one inequality. **Step 2** takes an **arbitrary** competitor subspace and shows it cannot beat the witness; the only thing known about a competitor is its dimension, so the only available tool is Lemma 17.2.1, which forces the competitor to meet a span of eigenvectors from the other end of the list. On that shared nonzero vector the quotient is pinned, and the inequality follows. **Step 3** obtains the second line from the first by applying it to $-\mathbf{A}$, which reverses the order of the eigenvalues and turns every maximum into a minimum.

Proof. By Corollary 12.4.3 over $\mathbb{C}$, or Corollary 12.5.2 over $\mathbb{R}$, fix an orthonormal basis

$(\mathbf{q}_1, \dots, \mathbf{q}_n)$ of F^n with $\mathbf{A}\mathbf{q}_i = \lambda_i(\mathbf{A})\mathbf{q}_i$ for every i. Abbreviate $\lambda_i = \lambda_i(\mathbf{A})$ and put

$$U_j = \operatorname{span}(\mathbf{q}_1, \dots, \mathbf{q}_j), \qquad L_j = \operatorname{span}(\mathbf{q}_j, \dots, \mathbf{q}_n),$$

so that $\dim U_j = j$ and $\dim L_j = n-j+1$, the lists being orthonormal and hence independent. By Lemma 17.2.2 the inner extrema below all exist.

Claim. For every j: $R_{\mathbf{A}}(\mathbf{x}) \geq \lambda_j$ for all $\mathbf{0} \neq \mathbf{x} \in U_j$, and $R_{\mathbf{A}}(\mathbf{x}) \leq \lambda_j$ for all $\mathbf{0} \neq \mathbf{x} \in L_j$. Both bounds are attained at $\mathbf{q}_j$.

Proof. Let $\mathbf{0} \neq \mathbf{x} \in U_j$ and write $c_i = \langle \mathbf{x}, \mathbf{q}_i \rangle$, so $\mathbf{x} = \sum_{i \leq j} c_i \mathbf{q}_i$ by Theorem 11.2.1 (a) and the definition of U_j. Then $\mathbf{A}\mathbf{x} = \sum_{i \leq j} \lambda_i c_i \mathbf{q}_i$, and Theorem 11.2.1 (b), (c) give

$$R_{\mathbf{A}}(\mathbf{x}) = \frac{\sum_{i \leq j} \lambda_i |c_i|^2}{\sum_{i \leq j} |c_i|^2},$$

a weighted average of $\lambda_1, \dots, \lambda_j$ with nonnegative weights summing to 1. Since the list is indexed decreasingly, the smallest of those eigenvalues is λ_j, so the average is $\geq \lambda_j$. The statement for L_j is Proposition 17.1.3: L_j is precisely the set of vectors orthogonal to $\mathbf{q}_1, \dots, \mathbf{q}_{j-1}$, and the maximum of $R_{\mathbf{A}}$ there is λ_j. In both cases $R_{\mathbf{A}}(\mathbf{q}_j) = \lambda_j$.

■

Step 1. The witness for the first line. For each subspace W of dimension $n - k + 1$ write $f(W) = \max_{\mathbf{0} \neq \mathbf{x} \in W} R_{\mathbf{A}}(\mathbf{x})$, which exists by Lemma 17.2.2. The outer minimum of f over all such W is not yet known to exist — there are infinitely many W — so we show that λ_k is a value of f and a lower bound for all its values; it is then the least value, which is what the minimum means. For the first half take $W = L_k$, of dimension $n - k + 1$: by the claim, $f(L_k) = \lambda_k$.

Step 2. Every competitor is at least as large. Let $W \subseteq F^n$ be any subspace with $\dim W = n - k + 1$. Then

$$\dim W + \dim U_k = (n - k + 1) + k = n + 1 > n,$$

so by Lemma 17.2.1 there is a vector $\mathbf{x} \neq \mathbf{0}$ lying in $W \cap U_k$. Being in U_k, it satisfies $R_{\mathbf{A}}(\mathbf{x}) \geq \lambda_k$ by the claim; being in W, it is one of the vectors the inner maximum ranges over. Therefore

$$\max_{\mathbf{0} \neq \mathbf{y} \in W} R_{\mathbf{A}}(\mathbf{y}) \;\geq\; R_{\mathbf{A}}(\mathbf{x}) \;\geq\; \lambda_k.$$

This holds for every competitor W, so $f(W) \geq \lambda_k$ for all of them. With Step 1, λ_k is a value of f and a lower bound for every value, so it is the least value: the minimum exists, equals λ_k, and is attained at $W = L_k$. This proves the first line.

Step 3. The second line, from the first applied to $-\mathbf{A}$. The matrix $-\mathbf{A}$ is Hermitian, its eigenvalues are the numbers $-\lambda_i$, and sorting those decreasingly reverses the list:

$$\lambda_j(-\mathbf{A}) = -\lambda_{n+1-j}(\mathbf{A}) \qquad (1 \leq j \leq n).$$

Also $R_{-\mathbf{A}}(\mathbf{x}) = -R_{\mathbf{A}}(\mathbf{x})$ for every $\mathbf{x} \neq \mathbf{0}$, directly from the definition. Since negation turns a maximum of a set of reals into the minimum of the negated set, and a minimum into a maximum,

$$\max_{\mathbf{0} \neq \mathbf{x} \in W} R_{-\mathbf{A}}(\mathbf{x}) = -\min_{\mathbf{0} \neq \mathbf{x} \in W} R_{\mathbf{A}}(\mathbf{x}),$$

$$\min_{\dim W = m} \bigl(-f(W)\bigr) = -\max_{\dim W = m} f(W)$$

for any real-valued f on subspaces, where each side exists exactly when the other does. Apply the first line, already proved, to $-\mathbf{A}$ with index j:

$$\begin{aligned} -\lambda_{n+1-j}(\mathbf{A}) = \lambda_j(-\mathbf{A}) &= \min_{\dim W = n-j+1} \max_{\mathbf{0}\neq\mathbf{x}\in W} R_{-\mathbf{A}}(\mathbf{x}) \\ &= -\max_{\dim W = n-j+1} \min_{\mathbf{0}\neq\mathbf{x}\in W} R_{\mathbf{A}}(\mathbf{x}). \end{aligned}$$

Now substitute $j = n+1-k$, so that $n-j+1 = k$ and $n+1-j = k$. Multiplying by -1 gives

$$\lambda_k(\mathbf{A}) = \max_{\dim W = k} \min_{\mathbf{0}\neq\mathbf{x}\in W} R_{\mathbf{A}}(\mathbf{x}),$$

with the maximum attained, the attainment transported from Step 1 by the same negation; explicitly it occurs at $W = U_k$, which is the subspace L_{n+1-k} of the eigenbasis of $-\mathbf{A}$ taken in its own decreasing order. This proves the theorem.

■

The theorem removes the eigenvectors from the description of $\lambda_k(\mathbf{A})$. What is left mentions only $\mathbf{A}$, through $R_{\mathbf{A}}$, and the subspaces of F^n, which are the same for every matrix. Two Hermitian matrices of the same size can therefore be compared term by term, and that is the whole reason this chapter exists.

Remark. **The increasing-index form.** If the eigenvalues are listed *increasingly* as $\mu_1 \le \mu_2 \le \cdots \le \mu_n$, then $\mu_j = \lambda_{n+1-j}(\mathbf{A})$, and substituting $k = n+1-j$ in Theorem 17.2.3 turns it into

$$\begin{aligned} \mu_j &= \min_{\dim W = j} \max_{\mathbf{0}\neq\mathbf{x}\in W} R_{\mathbf{A}}(\mathbf{x}) \\ &= \max_{\dim W = n-j+1} \min_{\mathbf{0}\neq\mathbf{x}\in W} R_{\mathbf{A}}(\mathbf{x}). \end{aligned}$$

The pairing of min max with max min is unchanged; only the dimensions move. Many books state the theorem this way, so a reader comparing sources should check which convention is in force before comparing formulas. This book uses the decreasing one everywhere, and this is the only place the other appears.

Warning. **Both the order of the two extrema and the dimension attached to each are load-bearing.** Neither

$$\max_{\dim W = n-k+1} \min_{\mathbf{0}\neq\mathbf{x}\in W} R_{\mathbf{A}}(\mathbf{x}) \qquad \text{nor} \qquad \min_{\dim W = k} \max_{\mathbf{0}\neq\mathbf{x}\in W} R_{\mathbf{A}}(\mathbf{x})$$

equals $\lambda_k(\mathbf{A})$. Each of them is, by Theorem 17.2.3 itself, equal to $\lambda_{n+1-k}(\mathbf{A})$: the first is the second line of the theorem at index $n-k+1$, and the second is the first line at the index j with $n-j+1 = k$, namely $j = n-k+1$. So swapping either the order or the dimension returns the eigenvalue counted from the *other end* of the list.

For a witness take $\mathbf{A} = \operatorname{diag}(2, 0)$ and $k = 1$, so that $\lambda_1(\mathbf{A}) = 2$. The first false expression maximizes over the single subspace $W = F^2$ the minimum of $R_{\mathbf{A}}$, which is 0. The second minimizes over lines the maximum on the line, which is the value of $R_{\mathbf{A}}$ on that line, and the smallest such value is again 0. Both return $\lambda_2(\mathbf{A}) = 0$ instead of 2. The mistake is invisible when $n = 2k-1$, since then $n+1-k = k$, and this is exactly why it survives in the wild.

Quick check. In the first line of Theorem 17.2.3, why is the dimension $n-k+1$ rather than k? Check your reason against $k=1$.

Solution. The competitor W must be forced to meet $U_k = \operatorname{span}(\mathbf{q}_1,\dots,\mathbf{q}_k)$, which has dimension k; by Lemma 17.2.1 that needs $\dim W > n-k$, and $n-k+1$ is the smallest such dimension. It must not be larger either: competitors of dimension $n-k+2$ would be forced to meet U_{k-1} as well, and by the theorem itself the min–max over them is $\lambda_{k-1}(\mathbf{A})$, the wrong eigenvalue.

At $k=1$ the dimension is $n-1+1=n$, so the only competitor is $W=F^n$, and the statement reads $\lambda_1(\mathbf{A}) = \max_{\mathbf{x}\neq\mathbf{0}} R_{\mathbf{A}}(\mathbf{x})$, which is Proposition 17.1.1. With the dimension $k=1$ instead, we would be computing min over lines of the single value of $R_{\mathbf{A}}$ on the line, which is $\lambda_n(\mathbf{A})$ — the wrong end.

Example. Both forms on one matrix

Let

$$\mathbf{A} = \begin{pmatrix} 2 & 2 & 0 \\ 2 & 3 & 2 \\ 0 & 2 & 4 \end{pmatrix} \in M_3(\mathbb{R}),$$

whose eigenvalues are $6, 3, 0$, with orthonormal eigenvectors

$$\mathbf{q}_1 = \tfrac{1}{3}(1,2,2), \qquad \mathbf{q}_2 = \tfrac{1}{3}(2,1,-2), \qquad \mathbf{q}_3 = \tfrac{1}{3}(2,-2,1).$$

Verify both lines of Theorem 17.2.3 for $k=2$ by computing the witness and two competitors in each case.

Solution. Throughout we use Lemma 17.2.2: for a subspace W with orthonormal basis the columns of $\mathbf{Q}$, the values of $R_{\mathbf{A}}$ on W fill the interval between the smallest and largest eigenvalues of the 2×2 matrix $\mathbf{B}=\mathbf{Q}^{\top}\mathbf{A}\mathbf{Q}$.

Max–min form, $\dim W = k = 2$. The witness is $U_2 = \operatorname{span}(\mathbf{q}_1,\mathbf{q}_2)$; since $\mathbf{q}_1,\mathbf{q}_2$ are eigenvectors, $\mathbf{B} = \operatorname{diag}(6,3)$ and the minimum of $R_{\mathbf{A}}$ on U_2 is $3=\lambda_2(\mathbf{A})$. Two competitors:

- $W = \operatorname{span}(\mathbf{e}_1,\mathbf{e}_2)$, where $\mathbf{B}$ is the top-left 2×2 block $\left(\begin{smallmatrix} 2 & 2 \\ 2 & 3 \end{smallmatrix}\right)$, with eigenvalues $\frac{1}{2}(5\pm\sqrt{17})$. Its minimum is $\frac{1}{2}(5-\sqrt{17}) \approx 0.438$, well below 3.
- $W = \operatorname{span}(\mathbf{e}_1,\mathbf{e}_3)$, where $\mathbf{B} = \operatorname{diag}(2,4)$ because $a_{13}=0$. Its minimum is 2, again below 3.

Min–max form, $\dim W = n-k+1 = 2$ as well, since $n=3$. The witness is $L_2 = \operatorname{span}(\mathbf{q}_2,\mathbf{q}_3)$, where $\mathbf{B}=\operatorname{diag}(3,0)$ and the maximum is $3=\lambda_2(\mathbf{A})$. The same two competitors now go the other way:

- on $\operatorname{span}(\mathbf{e}_1,\mathbf{e}_2)$ the maximum is $\frac{1}{2}(5+\sqrt{17}) \approx 4.562 > 3$;
- on $\operatorname{span}(\mathbf{e}_1,\mathbf{e}_3)$ the maximum is $4>3$.

So each competitor overshoots in the min–max form and undershoots in the max–min form, and the two witnesses pin $\lambda_2(\mathbf{A})=3$ from both sides. Notice that the competitor $\operatorname{span}(\mathbf{e}_1,\mathbf{e}_2)$ is the coordinate plane obtained by deleting the third row and column, and its two eigenvalues

$0.438\ldots$ and $4.562\ldots$ are caught between consecutive eigenvalues $0 \le 0.438 \le 3 \le 4.562 \le 6$ of $\mathbf{A}$. That is no accident; it is the interlacing theorem of §04.

17.2.4 The two debts this pays at once

Corollary 17.2.4. The Loewner Order Compares Every Eigenvalue

Let $\mathbf{A}, \mathbf{B} \in M_n(F)$ be Hermitian with $\mathbf{A} \succeq \mathbf{B}$. Then

$$\lambda_i(\mathbf{A}) \;\ge\; \lambda_i(\mathbf{B}) \qquad \text{for every } i = 1, \ldots, n.$$

Idea. $\mathbf{A} \succeq \mathbf{B}$ says the Rayleigh quotient of $\mathbf{A}$ sits above that of $\mathbf{B}$ at every vector. Courant–Fischer builds each λ_i out of the Rayleigh quotient by a minimum and then a maximum, and both operations preserve "everywhere above". So the domination passes straight through to every eigenvalue.

Proof. By Definition 13.5.1, $\mathbf{A} \succeq \mathbf{B}$ means $\mathbf{A} - \mathbf{B} \succeq \mathbf{O}$, so $\langle (\mathbf{A}-\mathbf{B})\mathbf{x}, \mathbf{x}\rangle \ge 0$ for every $\mathbf{x}$. Dividing by $\langle \mathbf{x}, \mathbf{x}\rangle > 0$ for $\mathbf{x} \neq \mathbf{0}$,

$$R_{\mathbf{A}}(\mathbf{x}) \;\ge\; R_{\mathbf{B}}(\mathbf{x}) \qquad \text{for every } \mathbf{x} \neq \mathbf{0}.$$

Fix i and let W be any subspace of F^n with $\dim W = i$. Let $\mathbf{x}_1 \in W \setminus \{\mathbf{0}\}$ attain $\min_{\mathbf{0}\neq\mathbf{x}\in W} R_{\mathbf{A}}(\mathbf{x})$, which exists by Lemma 17.2.2. Then

$$\min_{\mathbf{0}\neq\mathbf{x}\in W} R_{\mathbf{A}}(\mathbf{x}) = R_{\mathbf{A}}(\mathbf{x}_1) \;\ge\; R_{\mathbf{B}}(\mathbf{x}_1) \;\ge\; \min_{\mathbf{0}\neq\mathbf{x}\in W} R_{\mathbf{B}}(\mathbf{x}).$$

So the function $W \mapsto \min_{\mathbf{0}\neq\mathbf{x}\in W} R_{\mathbf{A}}(\mathbf{x})$ dominates the corresponding function for $\mathbf{B}$ at every W, and therefore so do their maxima over the subspaces of dimension i. By Theorem 17.2.3 applied to $\mathbf{A}$ and to $\mathbf{B}$, those maxima are $\lambda_i(\mathbf{A})$ and $\lambda_i(\mathbf{B})$. This proves the corollary. ∎

Example. Checking the corollary on a pair

Let $\mathbf{A} = \left(\begin{smallmatrix} 4 & 1 \\ 1 & 3 \end{smallmatrix}\right)$ and $\mathbf{B} = \left(\begin{smallmatrix} 2 & 1 \\ 1 & 2 \end{smallmatrix}\right)$. Then $\mathbf{A} - \mathbf{B} = \operatorname{diag}(2, 1)$, which is positive semidefinite, so $\mathbf{A} \succeq \mathbf{B}$. The eigenvalues of $\mathbf{B}$ are 3 and 1, from trace 4 and determinant 3; those of $\mathbf{A}$ are $\frac{7 \pm \sqrt{5}}{2} \approx 4.618,\ 2.382$, from trace 7 and determinant 11. So $\lambda_1(\mathbf{A}) \approx 4.618 \ge 3 = \lambda_1(\mathbf{B})$ and $\lambda_2(\mathbf{A}) \approx 2.382 \ge 1 = \lambda_2(\mathbf{B})$, as the corollary demands. Note that the two gaps, $\frac{1+\sqrt{5}}{2} \approx 1.618$ and $\frac{5-\sqrt{5}}{2} \approx 1.382$, are different: the corollary orders the eigenvalues, and says nothing about moving them by the same amount.

This settles an account opened in Chapter 13 §05. Discussing Theorem 13.5.3 (b), which proved only $\lambda_1(\mathbf{A}) \ge \lambda_1(\mathbf{B})$ and $\lambda_n(\mathbf{A}) \ge \lambda_n(\mathbf{B})$, that section said: "The full statement is that $\lambda_i(\mathbf{A}) \ge \lambda_i(\mathbf{B})$ for **every** i, not only for $i = 1$ and $i = n$; but the intermediate eigenvalues have no description as a plain maximum over the whole unit sphere, and the min–max description

that does the job is the Courant–Fischer theorem of Chapter 17." Corollary 17.2.4 is that full statement, and the diagnosis was exact: what was missing was not an extra trick but a description of λ_i that two different matrices could be measured against. The proof above is short, and all of its weight is carried by Theorem 17.2.3.

Corollary 17.2.5. Courant–Fischer for a Self-Adjoint Operator

Let V be a finite-dimensional inner product space over $F = \mathbb{R}$ or $F = \mathbb{C}$, with $\dim V = n \geq 1$, and let $T \in \mathcal{L}(V)$ be self-adjoint, with eigenvalues $\lambda_1(T) \geq \cdots \geq \lambda_n(T)$ repeated according to multiplicity. Then for every k with $1 \leq k \leq n$,

$$\lambda_k(T) = \min_{\dim U = n-k+1} \max_{0 \neq v \in U} \frac{\langle Tv, v \rangle}{\langle v, v \rangle} = \max_{\dim U = k} \min_{0 \neq v \in U} \frac{\langle Tv, v \rangle}{\langle v, v \rangle},$$

where U runs over the subspaces of V of the stated dimension.

Idea. Transport the theorem along coordinates. An orthonormal basis of eigenvectors turns V into F^n without changing any inner product, turns T into a real diagonal matrix, and matches subspaces with subspaces of the same dimension. There is nothing to prove beyond checking that each of those three things is true.

Proof. By Proposition 11.6.7 (c) a self-adjoint operator is normal, so Theorem 12.4.2 over $\mathbb{C}$, and Theorem 12.5.1 over $\mathbb{R}$, give an orthonormal basis $\mathcal{E} = (e_1, \dots, e_n)$ of V consisting of eigenvectors of T; relabel it so that $Te_i = \lambda_i(T) e_i$ with the eigenvalues in decreasing order, which is possible because they are real by Theorem 12.2.1. In particular the eigenvalue list of T, repeated according to multiplicity, has exactly n entries.

Let $\Phi \colon V \to F^n$, $\Phi(v) = [v]_{\mathcal{E}}$, be the coordinate map, an isomorphism by Corollary 2.5.5. By Theorem 11.2.1 (b), $\langle \Phi u, \Phi v \rangle = \langle u, v \rangle$ for all $u, v \in V$, so Φ is an isometric isomorphism in the sense of Corollary 11.2.4. Put $\mathbf{A} = [T]_{\mathcal{E}}^{\mathcal{E}} = \operatorname{diag}(\lambda_1(T), \dots, \lambda_n(T))$, a real diagonal matrix, hence Hermitian, with $\lambda_i(\mathbf{A}) = \lambda_i(T)$ for every i. By Theorem 4.1.1, $\Phi(Tv) = \mathbf{A}\,\Phi(v)$.

Consequently, for $v \neq 0$,

$$\frac{\langle Tv, v \rangle}{\langle v, v \rangle} = \frac{\langle \mathbf{A}\Phi v, \Phi v \rangle}{\langle \Phi v, \Phi v \rangle} = R_{\mathbf{A}}(\Phi v),$$

and $\Phi v \neq \mathbf{0}$ because Φ is injective. Finally $U \mapsto \Phi(U)$ is a bijection from the subspaces of V to the subspaces of F^n preserving dimension, by Theorem 2.5.2, and it matches the nonzero vectors of U with those of $\Phi(U)$. So each of the two displayed expressions equals the corresponding expression for $\mathbf{A}$ in Theorem 17.2.3, which is $\lambda_k(\mathbf{A}) = \lambda_k(T)$. This proves the corollary. ∎

This is the statement Chapter 12 §02 promised when it wrote that "the maximization picture returns in Chapter 17, where the Rayleigh quotient and the Courant–Fischer theorem describe every eigenvalue of a self-adjoint operator by optimization". The picture has returned and the description is complete: every eigenvalue, not merely the extreme two, and stated for an operator

on an abstract inner product space, which is the setting Chapter 12 worked in. The remark in §01 records what it cost — nothing. Chapter 12 declined the maximization route because it would have needed the extreme value theorem to know that a maximum exists; here the spectral theorem supplies the maximizing vectors explicitly, so the min–max description is a *consequence* of Chapter 12 rather than an alternative to it, and no analysis has been imported.

What has been bought is comparison. Two Hermitian matrices are now comparable eigenvalue by eigenvalue whenever their Rayleigh quotients are comparable vector by vector, and Corollary 17.2.4 is the first instance. Section 3 takes the same idea and applies it to $\mathbf{A}$ and $\mathbf{A}+\mathbf{E}$, where the quotients differ by at most $\|\mathbf{E}\|_2$, and gets the perturbation bound that Chapter 16 could not state.

17.2.5 Exercises

17.2.5.1 A. Check your understanding

Exercise. A1

(a) State both lines of Theorem 17.2.3, being careful about which dimension goes with which line.

(b) State Lemma 17.2.1, and say at which point of the proof of Theorem 17.2.3 it is used and what it is used on.

(c) Determine whether the following is correct, and justify your answer:

$$\lambda_k(\mathbf{A}) = \min_{\dim W = k} \max_{\mathbf{0}\neq\mathbf{x}\in W} R_{\mathbf{A}}(\mathbf{x}).$$

(d) Write the first line of Theorem 17.2.3 in the increasing-index convention $\mu_1 \le \cdots \le \mu_n$.

(e) Explain in one or two sentences why Theorem 13.5.3 (b) could not be extended to intermediate eigenvalues by the argument Chapter 13 used, and what Theorem 17.2.3 supplies instead.

Solution.

(a) For Hermitian $\mathbf{A} \in M_n(F)$ and $1 \le k \le n$: $\lambda_k(\mathbf{A}) = \min$ max over subspaces of dimension $n-k+1$ of the Rayleigh quotient, and $= \max$ min over subspaces of dimension k. The large dimension goes with the minimum of maxima; the small one with the maximum of minima.

(b) If U, W are subspaces of an n-dimensional space with $\dim U + \dim W > n$, then $U \cap W \neq \{0\}$. It is used in Step 2, applied to an arbitrary competitor W of dimension $n-k+1$ and the span U_k of the top k eigenvectors, whose dimensions add to $n+1$.

(c) Incorrect. By Theorem 17.2.3 itself, that expression is the first line evaluated at the index j with $n-j+1=k$, so it equals $\lambda_{n-k+1}(\mathbf{A})$. For $\mathbf{A} = \operatorname{diag}(2,0)$ and $k=1$ it returns 0, not $\lambda_1(\mathbf{A}) = 2$.

(d) $\mu_j = \min_{\dim W = j} \max_{\mathbf{0}\neq\mathbf{x}\in W} R_{\mathbf{A}}(\mathbf{x})$, obtained by setting $k = n+1-j$.

(e) Chapter 13's argument evaluated the quadratic form at a vector that is extremal for the *other* matrix, which works only when the extremal vector is characterized by a plain maximum over the whole sphere — true for λ_1 and λ_n and for no other index, since the range of $R_{\mathbf{A}}$ is an interval (Proposition 17.1.1). Theorem 17.2.3 supplies a description of every λ_i as an optimization over a family of subspaces that does not depend on the

matrix, so the same subspace can be fed to both matrices; that is exactly what the proof of Corollary 17.2.4 does.

17.2.5.2 B. Practice

Exercise. B1: Both forms on a diagonal matrix

Let $\mathbf{A} = \operatorname{diag}(5, 2, -1) \in M_3(\mathbb{R})$ and $k = 2$.

(a) Give the witness subspace for the max–min form, compute the minimum of $R_{\mathbf{A}}$ on it, and check the value against Theorem 17.2.3.

(b) Do the same for the min–max form.

(c) Put $\mathbf{u} = \frac{1}{\sqrt{2}}(0, 1, 1)$. Compute the minimum of $R_{\mathbf{A}}$ over the nonzero vectors of $\operatorname{span}(\mathbf{e}_1, \mathbf{u})$, and over those of $\operatorname{span}(\mathbf{e}_2, \mathbf{e}_3)$, and verify that neither subspace beats the witness in (a).

Solution.

(a) The eigenvalues are $\lambda_1 = 5$, $\lambda_2 = 2$, $\lambda_3 = -1$ with eigenvectors $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$. The witness is $U_2 = \operatorname{span}(\mathbf{e}_1, \mathbf{e}_2)$, of dimension $k = 2$. There $\mathbf{B} = \operatorname{diag}(5, 2)$, so by Lemma 17.2.2 the values of $R_{\mathbf{A}}$ fill $[2, 5]$ and the minimum is $2 = \lambda_2(\mathbf{A})$.

(b) Here $n - k + 1 = 2$ as well, and the witness is $L_2 = \operatorname{span}(\mathbf{e}_2, \mathbf{e}_3)$, where $\mathbf{B} = \operatorname{diag}(2, -1)$; the values fill $[-1, 2]$ and the maximum is $2 = \lambda_2(\mathbf{A})$.

(c) For $W = \operatorname{span}(\mathbf{e}_1, \mathbf{u})$ the two spanning vectors are orthonormal, and

$$\mathbf{B} = \begin{pmatrix} 5 & 0 \\ 0 & \frac{1}{2}(2-1) \end{pmatrix} = \operatorname{diag}(5, \tfrac{1}{2}),$$

the off-diagonal entry vanishing because $\mathbf{A}\mathbf{e}_1 = 5\mathbf{e}_1$ is orthogonal to $(0, 1, 1)$. So the minimum is $\frac{1}{2} < 2$. For $W = \operatorname{span}(\mathbf{e}_2, \mathbf{e}_3)$ we get $\mathbf{B} = \operatorname{diag}(2, -1)$ and minimum $-1 < 2$. Neither competitor reaches $\lambda_2(\mathbf{A}) = 2$, as Theorem 17.2.3 requires.

Exercise. B2: Shifting by a multiple of the identity

Let $\mathbf{A} \in M_n(F)$ be Hermitian and let $c \in \mathbb{R}$. Prove that $\lambda_k(\mathbf{A} + c\mathbf{I}) = \lambda_k(\mathbf{A}) + c$ for every k, using the first line of Theorem 17.2.3 for both matrices.

Solution. The matrix $\mathbf{A} + c\mathbf{I}$ is Hermitian, since c is real. For every $\mathbf{x} \neq \mathbf{0}$,

$$R_{\mathbf{A}+c\mathbf{I}}(\mathbf{x}) = \frac{\langle \mathbf{A}\mathbf{x}, \mathbf{x} \rangle + c\langle \mathbf{x}, \mathbf{x} \rangle}{\langle \mathbf{x}, \mathbf{x} \rangle} = R_{\mathbf{A}}(\mathbf{x}) + c.$$

Adding a constant to a real-valued function adds it to every maximum and every minimum of that function over any nonempty set. So for each subspace W,

$$\max_{\mathbf{0} \neq \mathbf{x} \in W} R_{\mathbf{A}+c\mathbf{I}}(\mathbf{x}) = \Big(\max_{\mathbf{0} \neq \mathbf{x} \in W} R_{\mathbf{A}}(\mathbf{x}) \Big) + c,$$

and taking the minimum over all W of dimension $n-k+1$ adds c again. By Theorem 17.2.3 applied to $\mathbf{A}+c\mathbf{I}$ on the left and to $\mathbf{A}$ on the right, $\lambda_k(\mathbf{A}+c\mathbf{I}) = \lambda_k(\mathbf{A}) + c$.

Exercise. B3: Comparing two matrices

Let

$$\mathbf{A} = \begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix}, \qquad \mathbf{B} = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}.$$

Verify that $\mathbf{A} \succeq \mathbf{B}$, compute both eigenvalue lists, and check Corollary 17.2.4. Check whether $\lambda_1(\mathbf{A}) - \lambda_1(\mathbf{B}) \le \lambda_1(\mathbf{A}-\mathbf{B})$ holds for this pair. (Whether it holds for every pair is a question for Section 3.)

Solution. $\mathbf{A}-\mathbf{B} = \left(\begin{smallmatrix} 1 & 1 \\ 1 & 2 \end{smallmatrix}\right)$ is symmetric with leading principal minors $1 > 0$ and $\det = 2-1 = 1 > 0$, so $\mathbf{A}-\mathbf{B} \succ \mathbf{O}$ by Theorem 13.1.5; in particular $\mathbf{A} \succeq \mathbf{B}$.

The eigenvalues: $\mathbf{A}$ has $p_{\mathbf{A}}(x) = (x-3)^2 - 1$, so $\lambda_1(\mathbf{A}) = 4$, $\lambda_2(\mathbf{A}) = 2$; and $\mathbf{B}$ is diagonal with $\lambda_1(\mathbf{B}) = 2$, $\lambda_2(\mathbf{B}) = 1$. Indeed $4 \ge 2$ and $2 \ge 1$, as Corollary 17.2.4 requires.

For the last question, $\mathbf{A}-\mathbf{B}$ has trace 3 and determinant 1, so its eigenvalues are $\frac{1}{2}(3 \pm \sqrt{5})$ and $\lambda_1(\mathbf{A}-\mathbf{B}) = \frac{1}{2}(3+\sqrt{5}) \approx 2.618$. The inequality asks whether $4-2 \le 2.618$, which holds. (It is not an accident: it is the simplest of the perturbation inequalities of §03, which carry Weyl's name. Note that Corollary 17.2.4 alone gives no upper bound on the gap $\lambda_i(\mathbf{A}) - \lambda_i(\mathbf{B})$.)

17.2.5.3 C. Going deeper

Exercise. C1: The definite generalized eigenvalue problem

Let $\mathbf{A}, \mathbf{B} \in M_n(F)$ be Hermitian with $\mathbf{A} \succ \mathbf{O}$, and recall Definition 13.4.1: λ is a generalized eigenvalue of the pair $(\mathbf{A}, \mathbf{B})$ when $\mathbf{B}\mathbf{x} = \lambda\mathbf{A}\mathbf{x}$ for some $\mathbf{x} \neq \mathbf{0}$. Define the **generalized Rayleigh quotient**

$$R_{\mathbf{A},\mathbf{B}}(\mathbf{x}) = \frac{\langle \mathbf{B}\mathbf{x}, \mathbf{x}\rangle}{\langle \mathbf{A}\mathbf{x}, \mathbf{x}\rangle} \qquad (\mathbf{x} \neq \mathbf{0}).$$

(a) Prove that $R_{\mathbf{A},\mathbf{B}}$ is well defined and real-valued.

(b) Put $\mathbf{C} = \mathbf{A}^{-1/2}\mathbf{B}\mathbf{A}^{-1/2}$. Prove that $\mathbf{C}$ is Hermitian and that $\det(\mathbf{B} - \lambda\mathbf{A}) = \det(\mathbf{A})\det(\mathbf{C} - \lambda\mathbf{I})$, so that the generalized eigenvalues of $(\mathbf{A}, \mathbf{B})$, with multiplicity, are exactly the eigenvalues of $\mathbf{C}$. Write them $d_1 \ge \cdots \ge d_n$.

(c) Prove that for every k,

$$d_k = \min_{\dim W = n-k+1} \; \max_{\mathbf{0} \neq \mathbf{x} \in W} R_{\mathbf{A},\mathbf{B}}(\mathbf{x}).$$

Hint for (c): the substitution $\mathbf{x} = \mathbf{A}^{-1/2}\mathbf{y}$.

Solution.

(a) By Theorem 13.2.1 and Theorem 13.1.5, $\mathbf{A} \succ \mathbf{O}$ has a positive definite square root

$\mathbf{A}^{1/2}$, which is invertible; write $\mathbf{A}^{-1/2}$ for its inverse, again Hermitian and positive definite. For $\mathbf{x} \neq \mathbf{0}$, $\langle \mathbf{A}\mathbf{x}, \mathbf{x}\rangle > 0$ by definiteness, so the denominator is a nonzero real; the numerator is real by Proposition 11.6.7 (a). So the quotient is a well-defined real number.

(b) $\mathbf{C}^* = \mathbf{A}^{-1/2}\mathbf{B}^*\mathbf{A}^{-1/2} = \mathbf{C}$. Since $\mathbf{A}^{1/2}\mathbf{C}\mathbf{A}^{1/2} = \mathbf{B}$ and $\mathbf{A}^{1/2}\mathbf{A}^{1/2} = \mathbf{A}$,

$$\mathbf{B} - \lambda\mathbf{A} = \mathbf{A}^{1/2}(\mathbf{C} - \lambda\mathbf{I})\mathbf{A}^{1/2},$$

so taking determinants and using Theorem 7.5.4 gives

$$\det(\mathbf{B} - \lambda\mathbf{A}) = \det(\mathbf{A}^{1/2})^2 \det(\mathbf{C} - \lambda\mathbf{I}) = \det(\mathbf{A}) \det(\mathbf{C} - \lambda\mathbf{I}).$$

Since $\det \mathbf{A} \neq 0$, the two polynomials in λ have the same roots with the same multiplicities. By Theorem 13.4.2 (b) the generalized eigenvalues of $(\mathbf{A}, \mathbf{B})$ are the roots of $\det(\mathbf{B} - \lambda\mathbf{A})$, so they are the eigenvalues of $\mathbf{C}$, all real because $\mathbf{C}$ is Hermitian.

(c) Put $\mathbf{S} = \mathbf{A}^{-1/2}$, which is Hermitian and invertible. For $\mathbf{y} \neq \mathbf{0}$ set $\mathbf{x} = \mathbf{S}\mathbf{y} \neq \mathbf{0}$. Then

$$\langle \mathbf{B}\mathbf{x}, \mathbf{x}\rangle = \langle \mathbf{S}\mathbf{B}\mathbf{S}\mathbf{y}, \mathbf{y}\rangle = \langle \mathbf{C}\mathbf{y}, \mathbf{y}\rangle, \qquad \langle \mathbf{A}\mathbf{x}, \mathbf{x}\rangle = \langle \mathbf{S}\mathbf{A}\mathbf{S}\mathbf{y}, \mathbf{y}\rangle = \langle \mathbf{y}, \mathbf{y}\rangle,$$

using $\mathbf{S}\mathbf{A}\mathbf{S} = \mathbf{A}^{-1/2}\mathbf{A}\mathbf{A}^{-1/2} = \mathbf{I}$. Hence $R_{\mathbf{A},\mathbf{B}}(\mathbf{S}\mathbf{y}) = R_{\mathbf{C}}(\mathbf{y})$. The map $\mathbf{y} \mapsto \mathbf{S}\mathbf{y}$ is an isomorphism of F^n, so $W \mapsto \mathbf{S}^{-1}W$ is a bijection of the subspaces of dimension $n - k + 1$ onto themselves, matching nonzero vectors with nonzero vectors. Therefore

$$\max_{\mathbf{0}\neq\mathbf{x}\in W} R_{\mathbf{A},\mathbf{B}}(\mathbf{x}) = \max_{\mathbf{0}\neq\mathbf{y}\in\mathbf{S}^{-1}W} R_{\mathbf{C}}(\mathbf{y}),$$

and minimizing both sides over all W of dimension $n - k + 1$ gives, by Theorem 17.2.3 applied to $\mathbf{C}$, the value $\lambda_k(\mathbf{C}) = d_k$.

Exercise. C2: How much weaker the eigenvalue comparison is

(a) Give Hermitian $\mathbf{A}, \mathbf{B} \in M_2(\mathbb{R})$ with $\lambda_i(\mathbf{A}) \geq \lambda_i(\mathbf{B})$ for $i = 1, 2$ but $\mathbf{A} \not\succeq \mathbf{B}$. So the converse of Corollary 17.2.4 is false.

(b) Prove that if $\mathbf{A}$ and $\mathbf{B}$ are Hermitian with $\lambda_i(\mathbf{A}) \geq \lambda_i(\mathbf{B})$ for every i and $\operatorname{tr}\mathbf{A} = \operatorname{tr}\mathbf{B}$, then $\lambda_i(\mathbf{A}) = \lambda_i(\mathbf{B})$ for every i.

(c) Prove that if $\mathbf{A} \succeq \mathbf{B}$ and $\operatorname{tr}\mathbf{A} = \operatorname{tr}\mathbf{B}$, then $\mathbf{A} = \mathbf{B}$.

Solution.

(a) Take $\mathbf{A} = \operatorname{diag}(2, 0)$ and $\mathbf{B} = \operatorname{diag}(0, 1)$. Then $\lambda_1(\mathbf{A}) = 2 \geq 1 = \lambda_1(\mathbf{B})$ and $\lambda_2(\mathbf{A}) = 0 \geq 0 = \lambda_2(\mathbf{B})$. But $\mathbf{A} - \mathbf{B} = \operatorname{diag}(2, -1)$ has the eigenvalue $-1 < 0$, so $\mathbf{A} - \mathbf{B} \not\succeq \mathbf{O}$ by Theorem 13.1.4 and $\mathbf{A} \not\succeq \mathbf{B}$. Ordering the two spectra says nothing about how the eigen*vectors* are placed relative to each other, and the Loewner order sees both.

(b) By Theorem 9.3.3 the trace is the sum of the eigenvalues with multiplicity, so

$$0 = \operatorname{tr}\mathbf{A} - \operatorname{tr}\mathbf{B} = \sum_{i=1}^{n} (\lambda_i(\mathbf{A}) - \lambda_i(\mathbf{B})).$$

Every summand is ≥ 0 by hypothesis, and a sum of nonnegative reals is zero only if each is zero. Hence $\lambda_i(\mathbf{A}) = \lambda_i(\mathbf{B})$ for every i.

(c) Put $\mathbf{C} = \mathbf{A} - \mathbf{B}$, Hermitian with $\mathbf{C} \succeq \mathbf{O}$, so every eigenvalue of $\mathbf{C}$ is ≥ 0 by Theorem 13.1.4. Also $\operatorname{tr}\mathbf{C} = \operatorname{tr}\mathbf{A} - \operatorname{tr}\mathbf{B} = 0$, which by Theorem 9.3.3 is the sum of those eigenvalues; a sum of nonnegative reals vanishing forces each to vanish, so every eigenvalue of $\mathbf{C}$ is 0. By Corollary 12.4.3 (or Corollary 12.5.2), $\mathbf{C} = \mathbf{U}\mathbf{D}\mathbf{U}^*$ with $\mathbf{D}$ the diagonal matrix of eigenvalues, which is $\mathbf{O}$; hence $\mathbf{C} = \mathbf{O}$ and $\mathbf{A} = \mathbf{B}$. (One may also finish with (b) and Corollary 17.2.4, which give the equality of the two spectra but not yet of the matrices; the trace argument on $\mathbf{C}$ is what closes the gap.)

Exercise. C3: Does enlarging the field change the answer?

Let $\mathbf{A} \in M_n(\mathbb{R})$ be symmetric. It may be read as an element of $M_n(\mathbb{C})$, where it is Hermitian, and then Theorem 17.2.3 minimizes over a much larger collection of subspaces.

(a) Prove that the eigenvalue lists of $\mathbf{A}$ over $\mathbb{R}$ and over $\mathbb{C}$, with multiplicity, are the same.

(b) Deduce that for every k,

$$\min_{\substack{W \subseteq \mathbb{R}^n \\ \dim W = n-k+1}} \max_{\mathbf{0} \neq \mathbf{x} \in W} R_{\mathbf{A}}(\mathbf{x}) = \min_{\substack{W \subseteq \mathbb{C}^n \\ \dim W = n-k+1}} \max_{\mathbf{0} \neq \mathbf{x} \in W} R_{\mathbf{A}}(\mathbf{x}).$$

(c) Show that the second minimum really does range over strictly more subspaces, by exhibiting a one-dimensional subspace of $\mathbb{C}^2$ that contains no nonzero real vector.

Solution.

(a) By Corollary 12.5.2 there is an orthogonal $\mathbf{Q} \in M_n(\mathbb{R})$ and a real diagonal $\mathbf{D}$ with $\mathbf{A} = \mathbf{Q}\mathbf{D}\mathbf{Q}^{\top}$, and the diagonal of $\mathbf{D}$ is the eigenvalue list of $\mathbf{A}$ over $\mathbb{R}$ with multiplicity. Read over $\mathbb{C}$, the same $\mathbf{Q}$ is unitary, since $\mathbf{Q}^* = \overline{\mathbf{Q}}^{\top} = \mathbf{Q}^{\top} = \mathbf{Q}^{-1}$, and the same equation $\mathbf{A} = \mathbf{Q}\mathbf{D}\mathbf{Q}^*$ holds. So $\mathbf{A}$ is unitarily similar over $\mathbb{C}$ to the same $\mathbf{D}$, and its complex eigenvalue list with multiplicity is again the diagonal of $\mathbf{D}$.

(b) Apply Theorem 17.2.3 twice: once with $F = \mathbb{R}$, where the left-hand side equals $\lambda_k(\mathbf{A})$ computed over $\mathbb{R}$, and once with $F = \mathbb{C}$, where the right-hand side equals $\lambda_k(\mathbf{A})$ computed over $\mathbb{C}$. By (a) these are the same number.

(c) Take $W = \operatorname{span}_{\mathbb{C}}\{(1, i)\} \subseteq \mathbb{C}^2$. A nonzero element of W is $c(1, i) = (c, ci)$ with $c \neq 0$; for it to be real we would need $c \in \mathbb{R}$ and $ci \in \mathbb{R}$, and $c \neq 0$ makes the second impossible. So W is not the complex span of any real subspace.

The lesson is worth stating. Enlarging the field adds competitors to the minimization, which can only push the minimum **down**, and adds vectors to each inner maximization, which can only push those maxima **up**. The two effects cancel exactly, because both the witness subspace and the bound of Step 2 in the proof of Theorem 17.2.3 survive the enlargement unchanged: the witness is spanned by real eigenvectors, and the dimension count of Lemma 17.2.1 holds over any field.

17.3 How Far Eigenvalues Move

Chapter 16 §07 asked what happens to the eigenvalues of a matrix when the matrix is nudged, and came back with a discouraging answer: they move continuously, and nothing better. Example (A Jordan block's eigenvalues move like $\varepsilon^{1/k}$) exhibits a perturbation of size ε that moves

every eigenvalue a distance $\varepsilon^{1/k}$, so no Lipschitz constant survives. That section also promised that the Hermitian case is different, and left the proof here.

The reason the Hermitian case is different is now available. Theorem 17.2.3 describes $\lambda_k(\mathbf{A})$ as an optimization over subspaces, and an optimization over subspaces reacts to a change in $\mathbf{A}$ the way any optimization reacts to a change in the objective: by at most the change in the objective. Turning that sentence into inequalities is the business of this section, and the inequalities are sharper than the sentence suggests, because the index may be shifted as well.

Throughout this section the field is $\mathbb{R}$ or $\mathbb{C}$, and every matrix whose eigenvalues are written down is Hermitian, that is, $\mathbf{A}^* = \mathbf{A}$ with $\mathbf{A} \in M_n(F)$ and $n \geq 1$. Its eigenvalues are real and are always indexed **decreasingly**, $\lambda_1(\mathbf{A}) \geq \lambda_2(\mathbf{A}) \geq \cdots \geq \lambda_n(\mathbf{A})$, as they have been since Chapter 13. Sums of Hermitian matrices are Hermitian, so $\mathbf{A} + \mathbf{B}$ is again a matrix this section may write λ_i of.

17.3.1 Two small facts about eigenvalue lists

One small fact is used in every proof below, and it is the reason the index bookkeeping works out. If a vector lies in the span of the *bottom* eigenvectors of $\mathbf{A}$, its Rayleigh quotient cannot exceed the largest eigenvalue in that span; if it lies in the span of the *top* ones, the quotient cannot fall below the smallest eigenvalue there.

Lemma 17.3.1. The Quotient on a Span of Eigenvectors

Let $\mathbf{A} \in M_n(F)$ be Hermitian and let $(\mathbf{u}_1, \dots, \mathbf{u}_n)$ be an orthonormal basis of F^n with $\mathbf{A}\mathbf{u}_p = \lambda_p(\mathbf{A})\mathbf{u}_p$ for every p. Fix i with $1 \leq i \leq n$.

(a) If $\mathbf{x} \neq \mathbf{0}$ lies in $\operatorname{span}\{\mathbf{u}_i, \mathbf{u}_{i+1}, \dots, \mathbf{u}_n\}$, then $R_{\mathbf{A}}(\mathbf{x}) \leq \lambda_i(\mathbf{A})$.

(b) If $\mathbf{x} \neq \mathbf{0}$ lies in $\operatorname{span}\{\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_i\}$, then $R_{\mathbf{A}}(\mathbf{x}) \geq \lambda_i(\mathbf{A})$.

Proof. Such a basis exists by the spectral theorem: Corollary 12.4.3 over $\mathbb{C}$, where a Hermitian matrix is normal and its eigenvalues may be listed in any prescribed order, and Corollary 12.5.2 over $\mathbb{R}$, where Hermitian means symmetric.

(a) Write $\mathbf{x} = \sum_{p \geq i} c_p \mathbf{u}_p$, with the c_p not all zero. Since the $\mathbf{u}_p$ are orthonormal,

$$\langle \mathbf{A}\mathbf{x}, \mathbf{x} \rangle = \Big\langle \sum_{p \geq i} c_p \lambda_p(\mathbf{A}) \mathbf{u}_p,\ \sum_{q \geq i} c_q \mathbf{u}_q \Big\rangle = \sum_{p \geq i} \lambda_p(\mathbf{A}) |c_p|^2,$$
$$\|\mathbf{x}\|^2 = \sum_{p \geq i} |c_p|^2.$$

Every index in these sums satisfies $p \geq i$, so $\lambda_p(\mathbf{A}) \leq \lambda_i(\mathbf{A})$ by the decreasing indexing, and therefore $\langle \mathbf{A}\mathbf{x}, \mathbf{x} \rangle \leq \lambda_i(\mathbf{A}) \|\mathbf{x}\|^2$. Since $\mathbf{x} \neq \mathbf{0}$ we have $\|\mathbf{x}\|^2 > 0$ and may divide, which gives $R_{\mathbf{A}}(\mathbf{x}) \leq \lambda_i(\mathbf{A})$ by Definition 17.1.1.

(b) The same computation with $\mathbf{x} = \sum_{p \leq i} c_p \mathbf{u}_p$ gives $\langle \mathbf{A}\mathbf{x}, \mathbf{x} \rangle = \sum_{p \leq i} \lambda_p(\mathbf{A}) |c_p|^2$, and now every index satisfies $p \leq i$, so $\lambda_p(\mathbf{A}) \geq \lambda_i(\mathbf{A})$ and the inequality reverses. This proves the lemma.

■

The two halves are the two ways of using an eigenvector span: (a) caps a quotient from above,

(b) props it up from below. The proof of Weyl's inequalities needs only (a), applied to a single vector for two different matrices at once; (b) is the form in which the same fact enters the witness half of a min–max argument.

A second small fact converts statements about the top of an eigenvalue list into statements about the bottom. Negating a matrix negates its eigenvalues, and negation reverses the order of real numbers.

Lemma 17.3.2. Negation Reverses the List

Let $\mathbf{A} \in M_n(F)$ be Hermitian. Then $-\mathbf{A}$ is Hermitian and

$$\lambda_p(-\mathbf{A}) = -\lambda_{n+1-p}(\mathbf{A}) \qquad (1 \le p \le n).$$

Proof. $(-\mathbf{A})^* = -\mathbf{A}^* = -\mathbf{A}$, so $-\mathbf{A}$ is Hermitian. By the spectral theorem quoted in the proof of Lemma 17.3.1, $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{U}^*$ with $\mathbf{D} = \operatorname{diag}(\lambda_1(\mathbf{A}), \dots, \lambda_n(\mathbf{A}))$ and $\mathbf{U}$ unitary, whence $-\mathbf{A} = \mathbf{U}(-\mathbf{D})\mathbf{U}^*$. So the eigenvalue list of $-\mathbf{A}$, with multiplicity, is $-\lambda_1(\mathbf{A}), \dots, -\lambda_n(\mathbf{A})$. Multiplying by -1 reverses the order of real numbers, so listing these decreasingly gives

$$-\lambda_n(\mathbf{A}) \;\ge\; -\lambda_{n-1}(\mathbf{A}) \;\ge\; \dots \;\ge\; -\lambda_1(\mathbf{A}),$$

whose p-th entry is $-\lambda_{n+1-p}(\mathbf{A})$. This proves the lemma. ∎

17.3.2 The two families of Weyl inequalities

Here is the question. Given the eigenvalues of $\mathbf{A}$ and the eigenvalues of $\mathbf{B}$, separately, what can be said about the eigenvalues of $\mathbf{A}+\mathbf{B}$? Nothing exact: $\operatorname{diag}(1,0)$ and $\operatorname{diag}(0,1)$ have the same eigenvalue list as $\operatorname{diag}(1,0)$ and $\operatorname{diag}(1,0)$, while the two sums $\mathbf{I}$ and $\operatorname{diag}(2,0)$ do not. But a great deal inexactly, and the shape of the answer is dictated by the index arithmetic.

Theorem 17.3.3. Weyl's Inequalities

Let $\mathbf{A}, \mathbf{B} \in M_n(F)$ be Hermitian and let $1 \le i, j \le n$.

(a) If $i + j - 1 \le n$, then

$$\lambda_{i+j-1}(\mathbf{A}+\mathbf{B}) \;\le\; \lambda_i(\mathbf{A}) + \lambda_j(\mathbf{B}).$$

(b) If $i + j - n \ge 1$, then

$$\lambda_{i+j-n}(\mathbf{A}+\mathbf{B}) \;\ge\; \lambda_i(\mathbf{A}) + \lambda_j(\mathbf{B}).$$

Idea. For (a), read $\lambda_{i+j-1}(\mathbf{A}+\mathbf{B})$ through the **max–min** half of Theorem 17.2.3: it is the best guaranteed value of $R_{\mathbf{A}+\mathbf{B}}$ on a subspace of dimension $i+j-1$. So take an arbitrary such subspace W and find one bad vector inside it — a vector on which $R_{\mathbf{A}}$ is capped by $\lambda_i(\mathbf{A})$ and $R_{\mathbf{B}}$ is capped by $\lambda_j(\mathbf{B})$ simultaneously. The span of the bottom $n-i+1$ eigenvectors of $\mathbf{A}$ is a *subspace* on which $R_{\mathbf{A}}$ is capped by $\lambda_i(\mathbf{A})$, and the span of the bottom $n-j+1$ eigenvectors of $\mathbf{B}$ is one on which $R_{\mathbf{B}}$ is capped by $\lambda_j(\mathbf{B})$. (The set of *all* vectors capped

by $\lambda_i(\mathbf{A})$ is larger and is not a subspace; the spans are what the dimension count can use.) Three subspaces must meet, and the dimensions have been chosen so that they do — twice, by Lemma 17.2.1. Since $R_{\mathbf{A}+\mathbf{B}} = R_{\mathbf{A}} + R_{\mathbf{B}}$, that one vector finishes the argument.

Part (b) is not a new proof. Negating a Hermitian matrix reverses its eigenvalue list (Lemma 17.3.2), so (a) applied to $-\mathbf{A}$ and $-\mathbf{B}$ at the reflected indices $n+1-i$ and $n+1-j$ says exactly (b), once every index is translated back.

Proof. **(a)** Put $k = i + j - 1$, so $1 \le k \le n$. By the spectral theorem choose an orthonormal basis $(\mathbf{u}_1, \dots, \mathbf{u}_n)$ of eigenvectors of $\mathbf{A}$ with $\mathbf{A}\mathbf{u}_p = \lambda_p(\mathbf{A})\mathbf{u}_p$, and an orthonormal basis $(\mathbf{v}_1, \dots, \mathbf{v}_n)$ of eigenvectors of $\mathbf{B}$ with $\mathbf{B}\mathbf{v}_q = \lambda_q(\mathbf{B})\mathbf{v}_q$. Set

$$U = \operatorname{span}\{\mathbf{u}_i, \dots, \mathbf{u}_n\}, \qquad V = \operatorname{span}\{\mathbf{v}_j, \dots, \mathbf{v}_n\},$$

so that $\dim U = n - i + 1$ and $\dim V = n - j + 1$, the two lists being independent as parts of a basis.

Let $W \le F^n$ be **any** subspace with $\dim W = k$. Then

$$\dim W + \dim U = (i + j - 1) + (n - i + 1) = n + j > n,$$

since $j \ge 1$, so $W \cap U \neq \{\mathbf{0}\}$ by Lemma 17.2.1. More than that is needed, namely how large $W \cap U$ is, and the dimension formula Theorem 1.7.3 supplies it:

$$\dim(W \cap U) = \dim W + \dim U - \dim(W + U) \;\ge\; (n + j) - n = j,$$

because $W + U$ is a subspace of F^n. Therefore

$$\dim(W \cap U) + \dim V \;\ge\; j + (n - j + 1) = n + 1 > n,$$

and a second application of Lemma 17.2.1, to the subspaces $W \cap U$ and V, produces a vector

$$\mathbf{x} \in W \cap U \cap V, \qquad \mathbf{x} \neq \mathbf{0}.$$

Now evaluate. Since $\mathbf{x} \in U$, part (a) of Lemma 17.3.1 gives $R_{\mathbf{A}}(\mathbf{x}) \le \lambda_i(\mathbf{A})$; since $\mathbf{x} \in V$, the same part applied to $\mathbf{B}$ gives $R_{\mathbf{B}}(\mathbf{x}) \le \lambda_j(\mathbf{B})$. The inner product is additive in its first slot, so $\langle(\mathbf{A} + \mathbf{B})\mathbf{x}, \mathbf{x}\rangle = \langle \mathbf{A}\mathbf{x}, \mathbf{x}\rangle + \langle \mathbf{B}\mathbf{x}, \mathbf{x}\rangle$ and hence $R_{\mathbf{A}+\mathbf{B}}(\mathbf{x}) = R_{\mathbf{A}}(\mathbf{x}) + R_{\mathbf{B}}(\mathbf{x})$. Since $\mathbf{x}$ is a nonzero vector of W,

$$\min_{\mathbf{0} \neq \mathbf{y} \in W} R_{\mathbf{A}+\mathbf{B}}(\mathbf{y}) \;\le\; R_{\mathbf{A}+\mathbf{B}}(\mathbf{x}) \;\le\; \lambda_i(\mathbf{A}) + \lambda_j(\mathbf{B}).$$

The subspace W of dimension k was arbitrary, so the maximum of the left-hand side over all such W obeys the same bound; and that maximum is $\lambda_k(\mathbf{A} + \mathbf{B})$ by the max–min half of Theorem 17.2.3. This proves (a).

(b) Suppose $i + j - n \ge 1$ and set

$$i' = n + 1 - i, \qquad j' = n + 1 - j,$$

both of which lie in $\{1, \dots, n\}$. Then

$$i' + j' - 1 = 2n + 1 - i - j \le n$$

precisely because $i+j-n \geq 1$, so part (a) applies to the Hermitian matrices $-\mathbf{A}$ and $-\mathbf{B}$ at the indices i', j':

$$\lambda_{i'+j'-1}(-\mathbf{A}-\mathbf{B}) \;\leq\; \lambda_{i'}(-\mathbf{A}) + \lambda_{j'}(-\mathbf{B}).$$

Translate each term with Lemma 17.3.2. On the right, $\lambda_{i'}(-\mathbf{A}) = -\lambda_{n+1-i'}(\mathbf{A}) = -\lambda_i(\mathbf{A})$, and likewise $\lambda_{j'}(-\mathbf{B}) = -\lambda_j(\mathbf{B})$. On the left, $-\mathbf{A}-\mathbf{B} = -(\mathbf{A}+\mathbf{B})$ with $\mathbf{A}+\mathbf{B}$ Hermitian, and $i'+j'-1 = 2n+1-i-j$, so

$$\lambda_{i'+j'-1}(-(\mathbf{A}+\mathbf{B})) = -\lambda_{n+1-(2n+1-i-j)}(\mathbf{A}+\mathbf{B}) = -\lambda_{i+j-n}(\mathbf{A}+\mathbf{B}).$$

The displayed inequality therefore reads $-\lambda_{i+j-n}(\mathbf{A}+\mathbf{B}) \leq -\lambda_i(\mathbf{A}) - \lambda_j(\mathbf{B})$. Multiplying by -1 reverses it and gives (b). This proves the theorem.

■

Two remarks on the bookkeeping, because it is the whole content. First, the index on the left is *not* i or j but $i+j-1$, so a fixed eigenvalue $\lambda_k(\mathbf{A}+\mathbf{B})$ receives k upper bounds from (a), one for each way of splitting $k = i+j-1$, and the best of them is their minimum. Second, (a) only ever bounds from above and (b) only from below. To trap one eigenvalue $\lambda_i(\mathbf{A}+\mathbf{B})$ from both sides with the index of $\mathbf{A}$ unchanged, take $j=1$ in (a) and $j=n$ in (b); that is the next corollary.

Quick check. Take $i=j=1$ in (a) and $i=j=n$ in (b). What do the two inequalities say, and which earlier result are they?

Solution. They say $\lambda_1(\mathbf{A}+\mathbf{B}) \leq \lambda_1(\mathbf{A}) + \lambda_1(\mathbf{B})$ and $\lambda_n(\mathbf{A}+\mathbf{B}) \geq \lambda_n(\mathbf{A}) + \lambda_n(\mathbf{B})$. For $i=j=n$ the index on the left of (b) is $n+n-n=n$, as required. These are the statements that the largest eigenvalue is subadditive and the smallest superadditive, which one can also read straight off Lemma 13.5.2: $\langle(\mathbf{A}+\mathbf{B})\mathbf{x},\mathbf{x}\rangle = \langle\mathbf{A}\mathbf{x},\mathbf{x}\rangle + \langle\mathbf{B}\mathbf{x},\mathbf{x}\rangle \leq \lambda_1(\mathbf{A}) + \lambda_1(\mathbf{B})$ for every unit $\mathbf{x}$. Weyl's inequalities are what happens when this easy observation is run at every index instead of only at the two ends.

Warning. **The eigenvalues are not subadditive index by index.** It is tempting to read Weyl as $\lambda_i(\mathbf{A}+\mathbf{B}) \leq \lambda_i(\mathbf{A}) + \lambda_i(\mathbf{B})$, and that statement is false. Take $\mathbf{A} = \operatorname{diag}(1,0)$ and $\mathbf{B} = \operatorname{diag}(0,1)$, so that $\lambda(\mathbf{A}) = \lambda(\mathbf{B}) = (1,0)$ and $\mathbf{A}+\mathbf{B} = \mathbf{I}_2$ with $\lambda(\mathbf{A}+\mathbf{B}) = (1,1)$. At $i=2$ the proposed inequality reads $1 \leq 0+0$. The reverse reading $\lambda_i(\mathbf{A}+\mathbf{B}) \geq \lambda_i(\mathbf{A})+\lambda_i(\mathbf{B})$ fails on the same pair at $i=1$, where it reads $1 \geq 2$. Only the shifted indices $i+j-1$ and $i+j-n$ are correct.

Here is the theorem in full on one pair of matrices, so that the index arithmetic can be seen doing its work.

Example. Both families on one pair

Let

$$\mathbf{A} = \begin{pmatrix} 1 & 1 & 2 \\ 1 & 1 & 2 \\ 2 & 2 & 0 \end{pmatrix}, \qquad \mathbf{B} = \begin{pmatrix} 3 & 2 & 0 \\ 2 & 3 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Find the eigenvalues of $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{A}+\mathbf{B}$, and check every inequality of Theorem 17.3.3.

Solution. All three matrices are real symmetric. Their characteristic polynomials factor over $\mathbb{Z}$:

$$p_{\mathbf{A}}(x) = x(x-4)(x+2),$$
$$p_{\mathbf{B}}(x) = x(x-5)(x-1),$$
$$p_{\mathbf{A}+\mathbf{B}}(x) = (x-8)(x-1)(x+1),$$

so, in decreasing order,

$$\lambda(\mathbf{A}) = (4, 0, -2), \qquad \lambda(\mathbf{B}) = (5, 1, 0), \qquad \lambda(\mathbf{A}+\mathbf{B}) = (8, 1, -1).$$

For $\mathbf{A}$ one can see the list without computing: the rows $(1,1,2)$ and $(1,1,2)$ are equal, so $\operatorname{rank}\mathbf{A} = 2$ and 0 is an eigenvalue; $\operatorname{tr}\mathbf{A} = 2$ and $\operatorname{tr}(\mathbf{A}^2) = 20$ then force the other two to be 4 and -2.

With $n = 3$, family (a) applies when $i + j \le 4$ and family (b) when $i + j \ge 4$.

(i,j)	family (a): $\lambda_{i+j-1}(\mathbf{A}+\mathbf{B}) \le \lambda_i(\mathbf{A}) + \lambda_j(\mathbf{B})$
$(1,1)$	$8 \le 4 + 5 = 9$
$(1,2)$	$1 \le 4 + 1 = 5$
$(2,1)$	$1 \le 0 + 5 = 5$
$(1,3)$	$-1 \le 4 + 0 = 4$
$(2,2)$	$-1 \le 0 + 1 = 1$
$(3,1)$	$-1 \le -2 + 5 = 3$

(i,j)	family (b): $\lambda_{i+j-3}(\mathbf{A}+\mathbf{B}) \ge \lambda_i(\mathbf{A}) + \lambda_j(\mathbf{B})$
$(1,3)$	$8 \ge 4 + 0 = 4$
$(3,1)$	$8 \ge -2 + 5 = 3$
$(2,2)$	$8 \ge 0 + 1 = 1$
$(2,3)$	$1 \ge 0 + 0 = 0$
$(3,2)$	$1 \ge -2 + 1 = -1$
$(3,3)$	$-1 \ge -2 + 0 = -2$

Every inequality holds, and every one of them is strict. At $(1,1)$ one can see why. Equality there would mean a unit $\mathbf{x}$ with $R_{\mathbf{A}+\mathbf{B}}(\mathbf{x}) = \lambda_1(\mathbf{A}) + \lambda_1(\mathbf{B})$; since $R_{\mathbf{A}}(\mathbf{x}) \le \lambda_1(\mathbf{A})$ and $R_{\mathbf{B}}(\mathbf{x}) \le \lambda_1(\mathbf{B})$ by Lemma 17.3.1 (a) at $i = 1$, both would have to be equalities, and the computation in that proof shows that equality in $\sum_p \lambda_p |c_p|^2 \le \lambda_1 \sum_p |c_p|^2$ forces $c_p = 0$ at every p with $\lambda_p < \lambda_1$. So $\mathbf{x}$ would lie in the λ_1-eigenspace of $\mathbf{A}$ and in that of $\mathbf{B}$. Those are the lines through $(1,1,1)$ and through $(1,1,0)$ — as $\mathbf{A}(1,1,1) = 4(1,1,1)$ and $\mathbf{B}(1,1,0) = 5(1,1,0)$, and both eigenvalues are simple — and two distinct lines meet only at $\mathbf{0}$.

17.3.3 Perturbing one matrix

Setting $j = 1$ in family (a) and $j = n$ in family (b) leaves the index on the left equal to i. That is the case in which Weyl becomes a statement about *perturbation*: one matrix, moved.

Corollary 17.3.4. Perturbation, Two-Sided

Let $\mathbf{A}, \mathbf{B} \in M_n(F)$ be Hermitian. Then for every i,

$$\lambda_i(\mathbf{A}) + \lambda_n(\mathbf{B}) \;\le\; \lambda_i(\mathbf{A}+\mathbf{B}) \;\le\; \lambda_i(\mathbf{A}) + \lambda_1(\mathbf{B}).$$

Proof. For the upper bound take $j = 1$ in Theorem 17.3.3 (a): the hypothesis $i+1-1 = i \le n$ holds, and the conclusion is $\lambda_i(\mathbf{A}+\mathbf{B}) \le \lambda_i(\mathbf{A}) + \lambda_1(\mathbf{B})$. For the lower bound take $j = n$ in Theorem 17.3.3 (b): the hypothesis $i + n - n = i \ge 1$ holds, and the conclusion is $\lambda_i(\mathbf{A}+\mathbf{B}) \ge \lambda_i(\mathbf{A}) + \lambda_n(\mathbf{B})$. This proves the corollary. ∎

In words: adding $\mathbf{B}$ shifts each eigenvalue of $\mathbf{A}$, in order, by an amount trapped between the smallest and the largest eigenvalue of $\mathbf{B}$. Taking $\mathbf{B} \succeq \mathbf{O}$, so that $\lambda_n(\mathbf{B}) \ge 0$ by Theorem 13.1.4 (b), the lower bound becomes $\lambda_i(\mathbf{A}+\mathbf{B}) \ge \lambda_i(\mathbf{A})$ — which is Corollary 17.2.4 again, now as a special case of Weyl's inequalities.

To turn Corollary 17.3.4 into a bound of the kind Chapter 16 wanted, the two numbers $\lambda_1(\mathbf{B})$ and $\lambda_n(\mathbf{B})$ must be replaced by one, and the right one is a norm.

Lemma 17.3.5. The Spectral Norm of a Hermitian Matrix

Let $\mathbf{E} \in M_n(F)$ be Hermitian. Then

$$\|\mathbf{E}\|_2 = \max_{1 \le i \le n} |\lambda_i(\mathbf{E})| = \max\bigl(\lambda_1(\mathbf{E}),\; -\lambda_n(\mathbf{E})\bigr).$$

Proof. By the spectral theorem, $\mathbf{E} = \mathbf{U}\mathbf{D}\mathbf{U}^*$ with $\mathbf{U}$ unitary and $\mathbf{D} = \operatorname{diag}(\lambda_1(\mathbf{E}), \dots, \lambda_n(\mathbf{E}))$ real. Then

$$\mathbf{E}^*\mathbf{E} = \mathbf{E}^2 = \mathbf{U}\mathbf{D}^2\mathbf{U}^*,$$

so the eigenvalues of $\mathbf{E}^*\mathbf{E}$ are the numbers $\lambda_i(\mathbf{E})^2$, and the singular values of $\mathbf{E}$, which are their nonnegative square roots listed decreasingly (Definition 13.9.1), are the numbers $|\lambda_i(\mathbf{E})|$ listed decreasingly. Hence $\sigma_1(\mathbf{E}) = \max_i |\lambda_i(\mathbf{E})|$, and $\|\mathbf{E}\|_2 = \sigma_1(\mathbf{E})$ by Theorem 16.3.3 (c).

For the second equality, $\lambda_n(\mathbf{E}) \le \lambda_i(\mathbf{E}) \le \lambda_1(\mathbf{E})$ for all i, so $|\lambda_i(\mathbf{E})| \le \max(|\lambda_1(\mathbf{E})|, |\lambda_n(\mathbf{E})|)$, so the maximum is attained at one end of the list. Now $\lambda_1(\mathbf{E}) \le |\lambda_1(\mathbf{E})|$ and $-\lambda_n(\mathbf{E}) \le |\lambda_n(\mathbf{E})|$, which gives one inequality between the two expressions. Conversely, $|\lambda_1(\mathbf{E})|$ is either $\lambda_1(\mathbf{E})$ or $-\lambda_1(\mathbf{E}) \le -\lambda_n(\mathbf{E})$, and $|\lambda_n(\mathbf{E})|$ is either $-\lambda_n(\mathbf{E})$ or $\lambda_n(\mathbf{E}) \le \lambda_1(\mathbf{E})$; in every case it is at most $\max(\lambda_1(\mathbf{E}), -\lambda_n(\mathbf{E}))$. This proves the lemma. ∎

Corollary 17.3.6. Hermitian Eigenvalues Are 1-Lipschitz

Let $\mathbf{A}, \mathbf{E} \in M_n(F)$ be Hermitian. Then

$$|\lambda_i(\mathbf{A}+\mathbf{E}) - \lambda_i(\mathbf{A})| \;\le\; \|\mathbf{E}\|_2 \qquad \text{for every } i = 1, \dots, n.$$

Proof. By Corollary 17.3.4 applied with $\mathbf{B} = \mathbf{E}$,

$$\lambda_n(\mathbf{E}) \;\le\; \lambda_i(\mathbf{A}+\mathbf{E}) - \lambda_i(\mathbf{A}) \;\le\; \lambda_1(\mathbf{E}).$$

By Lemma 17.3.5, $\lambda_1(\mathbf{E}) \le \|\mathbf{E}\|_2$ and $-\lambda_n(\mathbf{E}) \le \|\mathbf{E}\|_2$, that is, $-\|\mathbf{E}\|_2 \le \lambda_n(\mathbf{E})$. So the difference lies in $[-\|\mathbf{E}\|_2, \|\mathbf{E}\|_2]$, which is the assertion. This proves the corollary.

■

This is the sentence Chapter 16 §07 promised: **the eigenvalues of a Hermitian matrix move by at most $\|\mathbf{E}\|_2$ under a Hermitian perturbation.** It is worth being precise about what it adds to Corollary 16.7.4, the continuity statement of that section, because the gain is not continuity — that was already had — but a constant.

- Corollary 16.7.4 says: for each $\mathbf{A}$ and each $\varepsilon > 0$ there is a $\delta > 0$ such that every matrix within δ of $\mathbf{A}$ has eigenvalues that can be numbered to lie within ε of those of $\mathbf{A}$. The δ depends on $\mathbf{A}$, and Example (A Jordan block's eigenvalues move like $\varepsilon^{1/k}$) shows that it must.
- Corollary 17.3.6 says, for Hermitian $\mathbf{A}$ and Hermitian perturbations measured in $\|\cdot\|_2$: $\delta = \varepsilon$ works, for every $\mathbf{A}$ at once, and the numbering is simply the decreasing order. The movement is bounded by the perturbation itself, with **Lipschitz constant** 1.

The constant does not depend on the size n of the matrices, it does not depend on how ill-conditioned $\mathbf{A}$ is, and it does not depend on how close together the eigenvalues of $\mathbf{A}$ are — three quantities that every reader of Chapter 16 has learned to expect in an error bound. It also cannot be improved: $\mathbf{E} = t\mathbf{I}$ with $t \in \mathbb{R}$ gives $\lambda_i(\mathbf{A} + t\mathbf{I}) = \lambda_i(\mathbf{A}) + t$ for every i, while $\|t\mathbf{I}\|_2 = |t|$, so the inequality is an equality at every index. Finally, the eigenvalues are matched **by index**: the i-th of $\mathbf{A}+\mathbf{E}$ is compared with the i-th of $\mathbf{A}$, not with whichever eigenvalue of $\mathbf{A}$ happens to be nearest, and no search for a good numbering is needed.

Corollary 17.3.7. The Hermitian Eigenvalue Problem Is Perfectly Conditioned in Absolute Terms

Let $\mathbf{A}, \mathbf{E} \in M_n(F)$ be Hermitian and write $\lambda(\mathbf{A}) = (\lambda_1(\mathbf{A}), \dots, \lambda_n(\mathbf{A}))$ for the decreasing eigenvalue list. Then

$$\|\lambda(\mathbf{A}+\mathbf{E}) - \lambda(\mathbf{A})\|_\infty \;\le\; \|\mathbf{E}\|_2,$$

and the constant 1 is the smallest one for which this holds.

Proof. The ∞-norm of a vector is the largest modulus of its entries, so the left-hand side is $\max_i |\lambda_i(\mathbf{A}+\mathbf{E}) - \lambda_i(\mathbf{A})|$, and the inequality is Corollary 17.3.6 read over all i at once. For the last clause, $\mathbf{E} = t\mathbf{I}$ with $t > 0$ makes both sides equal to t, so no constant below 1 can work. This proves the corollary.

■

Chapter 16 §08 measured a linear system by $\kappa(\mathbf{A}) = \|\mathbf{A}\|\|\mathbf{A}^{-1}\|$ (Definition 16.8.1), the worst factor by which $\mathbf{A}$ magnifies a relative error in the data. The corollary above is the corresponding statement for a different problem — not "solve $\mathbf{Ax} = \mathbf{b}$" but "report the eigenvalues of $\mathbf{A}$" — and the factor there is 1, in absolute rather than relative terms. The difference between absolute and relative matters here, and it is worth being exact. $\mathbf{A} = \operatorname{diag}(10^6, 10^{-6})$ is Hermitian with $\kappa_2(\mathbf{A}) = 10^{12}$, so solving with it is a delicate business, and yet its eigenvalues move by at most $\|\mathbf{E}\|_2$ under any Hermitian $\mathbf{E}$, exactly as those of every other Hermitian matrix do. **Ill-conditioning of the linear system says nothing about the *absolute* conditioning of the eigenvalues.** In *relative* terms the two numbers meet again. The Hermitian perturbation $\mathbf{E} = -10^{-6}\,\mathbf{e}_2\mathbf{e}_2^{\top}$ has relative size $\|\mathbf{E}\|_2/\|\mathbf{A}\|_2 = 10^{-12}$ and sends the eigenvalue 10^{-6} to 0, a relative change of 100%: an amplification of exactly $10^{12} = \kappa_2(\mathbf{A})$. In general, for an invertible Hermitian $\mathbf{A}$, Corollary 17.3.6 bounds the relative change of the eigenvalue of smallest modulus by $\|\mathbf{E}\|_2/\sigma_n(\mathbf{A})$, which is $\kappa_2(\mathbf{A})$ times the relative size of $\mathbf{E}$, and a rank-one $\mathbf{E}$ along its eigenvector attains it. So for Hermitian matrices, κ_2 *is* the relative condition number of the smallest eigenvalue.

The contrast that makes the corollary worth having is with Example (A Jordan block's eigenvalues move like $\varepsilon^{1/k}$). There $\mathbf{A} = \mathbf{J}_k(0)$ and $\mathbf{E} = \varepsilon\mathbf{E}_{k1}$, and the eigenvalues move a distance $\varepsilon^{1/k}$ while $\|\mathbf{E}\|_2 = \varepsilon$, an amplification of $\varepsilon^{-(k-1)/k}$ that is unbounded as $\varepsilon \to 0^+$. Neither matrix there is Hermitian, and Corollary 17.3.6 explains which hypothesis is doing the work: for $k = 2$ the perturbed matrix $\left(\begin{smallmatrix} 0 & 1 \\ \varepsilon & 0 \end{smallmatrix}\right)$ has the perfectly real eigenvalues $\pm\sqrt{\varepsilon}$, so the conclusion of Corollary 17.3.6 can be *stated* for it, and it is false — $\sqrt{\varepsilon} > \varepsilon$ for $0 < \varepsilon < 1$, by a factor of 1000 at $\varepsilon = 10^{-6}$.

Warning. **Both matrices must be Hermitian, and "small" is no substitute.** If $\mathbf{A}$ is not Hermitian, the corollary fails even when $\mathbf{E}$ is: take $\mathbf{A} = \mathbf{J}_2(0)$ and the Hermitian $\mathbf{E} = \varepsilon\left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$, with $\|\mathbf{E}\|_2 = \varepsilon$. Then $\mathbf{A} + \mathbf{E} = \left(\begin{smallmatrix} 0 & 1+\varepsilon \\ \varepsilon & 0 \end{smallmatrix}\right)$ has the real eigenvalues $\pm\sqrt{\varepsilon(1+\varepsilon)}$, which have moved from 0 by more than ε. If $\mathbf{E}$ is not Hermitian, the conclusion usually cannot even be stated: $\mathbf{A} = \mathbf{O}$ is Hermitian and $\mathbf{E} = \left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$ has $\|\mathbf{E}\|_2 = 1$, but $\mathbf{A} + \mathbf{E}$ has eigenvalues $\pm i$, so there is no decreasing list $\lambda_i(\mathbf{A} + \mathbf{E})$ of real numbers to compare with anything. A bound for a diagonalizable $\mathbf{A}$ and an arbitrary $\mathbf{E}$ does exist — the theorem of Bauer and Fike, in Chapter 20 — and it carries an extra factor that measures how far the eigenvector basis of $\mathbf{A}$ is from orthonormal.

Example. Locating eigenvalues without recomputing them

Let

$$\mathbf{A} = \begin{pmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{pmatrix}, \qquad \mathbf{E} = \tfrac{1}{10}(\mathbf{E}_{13} + \mathbf{E}_{31}).$$

Use Corollary 17.3.6 to locate the eigenvalues of $\mathbf{A} + \mathbf{E}$, then compute them exactly and compare.

Solution. The unperturbed spectrum. Expanding along the first row, $p_{\mathbf{A}}(x) = (x-2)((x-2)^2 - 2)$, so

$$\lambda(\mathbf{A}) = (2 + \sqrt{2},\ 2,\ 2 - \sqrt{2}) \approx (3.41421,\ 2,\ 0.58579).$$

The size of the perturbation. $\mathbf{E}$ is real symmetric with $\mathbf{E}^2 = \frac{1}{100}(\mathbf{E}_{11} + \mathbf{E}_{33})$, whose eigenvalues

are $\frac{1}{100}, \frac{1}{100}, 0$; so the singular values of $\mathbf{E}$ are $\frac{1}{10}, \frac{1}{10}, 0$ and $\|\mathbf{E}\|_2 = \frac{1}{10}$ by Theorem 16.3.3 (c). (One can also read this off Lemma 17.3.5: the eigenvalues of $\mathbf{E}$ are $\frac{1}{10}, 0, -\frac{1}{10}$.)

The prediction. By Corollary 17.3.6 each eigenvalue of $\mathbf{A}+\mathbf{E}$ lies within $\frac{1}{10}$ of the corresponding eigenvalue of $\mathbf{A}$:

$$\lambda_1 \in [3.31421,\ 3.51421], \qquad \lambda_2 \in [1.9,\ 2.1], \qquad \lambda_3 \in [0.48579,\ 0.68579].$$

In particular $\mathbf{A}+\mathbf{E}$ is positive definite, since $\lambda_3 > 0$, and its three eigenvalues are distinct, since the three intervals are disjoint.

The check. The characteristic polynomial of $\mathbf{A}+\mathbf{E}$ factors as

$$p_{\mathbf{A}+\mathbf{E}}(x) = \Big(x - \tfrac{19}{10}\Big)\Big(x^2 - \tfrac{41}{10}x + \tfrac{11}{5}\Big),$$

so

$$\begin{aligned}\lambda(\mathbf{A}+\mathbf{E}) &= \Big(\tfrac{41+3\sqrt{89}}{20},\ \tfrac{19}{10},\ \tfrac{41-3\sqrt{89}}{20}\Big)\\ &\approx (3.46510,\ 1.9,\ 0.63490).\end{aligned}$$

All three lie in the predicted intervals. The middle one lies at the very end of its interval: λ_2 has moved by exactly $\frac{1}{10} = \|\mathbf{E}\|_2$. That is not luck. The vector $(1, 0, -1)$ is an eigenvector of $\mathbf{A}$ for 2 and of $\mathbf{E}$ for $-\frac{1}{10}$, hence an eigenvector of $\mathbf{A}+\mathbf{E}$ for $2 - \frac{1}{10}$; a shared eigenvector is exactly what makes the bound tight.

17.3.4 Perturbations of small rank

The last corollary trades size for rank. A perturbation of rank r need not be small in norm at all, but it can only disturb r directions, and Weyl converts that into a shift of the index by r.

Corollary 17.3.8. A Rank-r Perturbation Shifts the Index by r

Let $\mathbf{A}, \mathbf{B} \in M_n(F)$ be Hermitian, let $r \geq 0$ be an integer, and suppose $\operatorname{rank}\mathbf{B} \leq r$. Then

$$\begin{aligned}\lambda_{i+r}(\mathbf{A}+\mathbf{B}) &\leq \lambda_i(\mathbf{A}) && (1 \leq i,\ i+r \leq n),\\ \lambda_i(\mathbf{A}) &\leq \lambda_{i-r}(\mathbf{A}+\mathbf{B}) && (i \leq n,\ i-r \geq 1).\end{aligned}$$

Idea. A Hermitian matrix of rank at most r has at most r nonzero eigenvalues, so once one walks $r+1$ places down its decreasing list one is certainly at a nonpositive number, and once one stops r places short of the bottom one is certainly at a nonnegative number. Feed those two facts into the two Weyl families at $j = r+1$ and $j = n-r$.

Proof. By the spectral theorem $\mathbf{B} = \mathbf{U}\mathbf{D}\mathbf{U}^*$ with $\mathbf{D} = \operatorname{diag}(\lambda_1(\mathbf{B}), \dots, \lambda_n(\mathbf{B}))$ and $\mathbf{U}$ unitary, so $\operatorname{rank}\mathbf{B} = \operatorname{rank}\mathbf{D}$ is the number of nonzero $\lambda_q(\mathbf{B})$, which is at most r by hypothesis.

Two facts about the list of $\mathbf{B}$. If $\lambda_{r+1}(\mathbf{B}) > 0$, then $\lambda_1(\mathbf{B}), \dots, \lambda_{r+1}(\mathbf{B})$ are all positive by the decreasing indexing, giving $r+1$ nonzero eigenvalues, a contradiction; hence

$$\lambda_{r+1}(\mathbf{B}) \leq 0 \qquad \text{whenever } r+1 \leq n.$$

If $\lambda_{n-r}(\mathbf{B}) < 0$, then $\lambda_{n-r}(\mathbf{B}), \dots, \lambda_n(\mathbf{B})$ are all negative, again $r+1$ nonzero eigenvalues and a contradiction; hence

$$\lambda_{n-r}(\mathbf{B}) \geq 0 \qquad \text{whenever } n - r \geq 1.$$

First inequality. Assume $i \geq 1$ and $i + r \leq n$; then $r + 1 \leq n$, so the index $j = r + 1$ is legitimate and $i + j - 1 = i + r \leq n$. By Theorem 17.3.3 (a) and the first fact,

$$\lambda_{i+r}(\mathbf{A} + \mathbf{B}) \leq \lambda_i(\mathbf{A}) + \lambda_{r+1}(\mathbf{B}) \leq \lambda_i(\mathbf{A}).$$

Second inequality. Assume $i \leq n$ and $i - r \geq 1$; then $n - r \geq n - i + 1 \geq 1$, so the index $j = n - r$ is legitimate, and $i + j - n = i - r \geq 1$. By Theorem 17.3.3 (b) and the second fact,

$$\lambda_{i-r}(\mathbf{A} + \mathbf{B}) \geq \lambda_i(\mathbf{A}) + \lambda_{n-r}(\mathbf{B}) \geq \lambda_i(\mathbf{A}).$$

This proves the corollary.

■

Read together, the two inequalities say that the two eigenvalue lists interlace with a lag of r:

$$\lambda_{i+r}(\mathbf{A} + \mathbf{B}) \;\leq\; \lambda_i(\mathbf{A}) \;\leq\; \lambda_{i-r}(\mathbf{A} + \mathbf{B}),$$

whenever all three indices make sense. For $r = 0$ this collapses to $\lambda_i(\mathbf{A} + \mathbf{B}) = \lambda_i(\mathbf{A})$, correctly, since a Hermitian matrix of rank 0 is $\mathbf{O}$. For $r = 1$ it is the statement that adding a Hermitian rank-one matrix moves each eigenvalue by at most one place in the list. That case is the subject of Section 5, and Section 9 returns to low-rank perturbations when it finishes the approximation theorem that Chapter 13 §11 could only state in the Frobenius norm.

Remark. The bound is about the *index*, not the *value*. A rank-one $\mathbf{B}$ can be enormous: $\mathbf{B} = 10^6\, \mathbf{e}_1 \mathbf{e}_1^*$ has rank 1, and $\lambda_1(\mathbf{A} + \mathbf{B})$ lies within $\|\mathbf{A}\|_2$ of 10^6, by Corollary 17.3.6 with the roles of the two matrices exchanged. What Corollary 17.3.8 guarantees is that the others are trapped all the same: $\lambda_{k+1}(\mathbf{A}) \leq \lambda_k(\mathbf{A} + \mathbf{B}) \leq \lambda_{k-1}(\mathbf{A})$ for $2 \leq k \leq n - 1$, however large the one that escapes.

17.3.5 Exercises

17.3.5.1 A. Check your understanding

Exercise. A1

(a) State both families of Theorem 17.3.3, with their index conditions, and say why the condition $i + j - 1 \leq n$ cannot be dropped from the first.

(b) Determine whether the following is correct, and justify your answer: for Hermitian $\mathbf{A}, \mathbf{B}$, $\lambda_2(\mathbf{A} + \mathbf{B}) \leq \lambda_2(\mathbf{A}) + \lambda_2(\mathbf{B})$.

(c) Corollary 16.7.4 already said that eigenvalues move continuously. What does Corollary 17.3.6 add?

(d) Explain why Corollary 17.3.6 cannot be applied to $\mathbf{A} = \mathbf{J}_2(0)$ and $\mathbf{E} = \varepsilon \mathbf{E}_{21}$, and what goes wrong if one applies it anyway.

(e) If $\mathbf{B}$ is Hermitian of rank 3 and $\mathbf{A}$ is Hermitian of size 10, which eigenvalue of $\mathbf{A}+\mathbf{B}$ is guaranteed to be at most $\lambda_4(\mathbf{A})$?

Solution.

(a) See Theorem 17.3.3. Family (a) needs $i+j-1 \le n$ and family (b) needs $i+j-n \ge 1$. Without the first condition the symbol $\lambda_{i+j-1}(\mathbf{A}+\mathbf{B})$ names nothing: a matrix of size n has only n eigenvalues, and there is no λ_{n+1}.

(b) Incorrect. With $\mathbf{A} = \operatorname{diag}(1,0)$ and $\mathbf{B} = \operatorname{diag}(0,1)$ it reads $\lambda_2(\mathbf{I}_2) = 1 \le 0+0$. The correct instance of Theorem 17.3.3 (a) at $i=j=2$ needs $i+j-1 = 3 \le n$, so it says nothing when $n=2$; when $n \ge 3$ it bounds λ_3, not λ_2.

(c) A constant. Corollary 16.7.4 gives, for each $\mathbf{A}$ and ε, some unspecified δ depending on $\mathbf{A}$; Corollary 17.3.6 gives $\delta = \varepsilon$ for every $\mathbf{A}$ simultaneously, that is, a Lipschitz constant of 1, independent of n, of the conditioning of $\mathbf{A}$, and of the gaps between its eigenvalues. It also matches the eigenvalues by index rather than as a multiset.

(d) Neither matrix is Hermitian, and the corollary's hypothesis is that both are. Applying it anyway predicts movement at most $\|\mathbf{E}\|_2 = \varepsilon$, whereas the eigenvalues are $\pm\sqrt{\varepsilon}$ and have moved a distance $\sqrt{\varepsilon}$, which exceeds ε for every $\varepsilon \in (0,1)$ — see Example (A Jordan block's eigenvalues move like $\varepsilon^{1/k}$).

(e) $\lambda_7(\mathbf{A}+\mathbf{B})$, and hence every later one, by Corollary 17.3.8 with $i=4$ and $r=3$.

17.3.5.2 B. Practice

Exercise. B1: A small table

Let $\mathbf{A} = \begin{pmatrix} 0 & 4 \\ 4 & 0 \end{pmatrix}$ and $\mathbf{B} = \begin{pmatrix} 3 & 0 \\ 0 & -3 \end{pmatrix}$. Compute the eigenvalues of $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{A}+\mathbf{B}$, write out every inequality of Theorem 17.3.3 for this pair, and verify Corollary 17.3.6 with $\mathbf{E} = \mathbf{B}$.

Solution. $p_{\mathbf{A}}(x) = x^2 - 16$, so $\lambda(\mathbf{A}) = (4,-4)$; $\mathbf{B}$ is diagonal, so $\lambda(\mathbf{B}) = (3,-3)$; and $\mathbf{A}+\mathbf{B} = \left(\begin{smallmatrix} 3 & 4 \\ 4 & -3 \end{smallmatrix}\right)$ has $\operatorname{tr} = 0$ and $\det = -25$, so $p_{\mathbf{A}+\mathbf{B}}(x) = x^2 - 25$ and $\lambda(\mathbf{A}+\mathbf{B}) = (5,-5)$.

Family (a), for $i+j \le 3$:

$$(1,1):\ 5 \le 4+3 = 7,$$
$$(1,2):\ -5 \le 4-3 = 1,$$
$$(2,1):\ -5 \le -4+3 = -1.$$

Family (b), for $i+j \ge 3$:

$$(1,2):\ 5 \ge 4-3 = 1,$$
$$(2,1):\ 5 \ge -4+3 = -1,$$
$$(2,2):\ -5 \ge -4-3 = -7.$$

All six hold. For the last part, $\mathbf{B}$ is Hermitian with eigenvalues ± 3, so $\|\mathbf{B}\|_2 = 3$ by Lemma 17.3.5, and

$$|\lambda_1(\mathbf{A}+\mathbf{B}) - \lambda_1(\mathbf{A})| = |5-4| = 1 \le 3,$$
$$|\lambda_2(\mathbf{A}+\mathbf{B}) - \lambda_2(\mathbf{A})| = |-5+4| = 1 \le 3,$$

as Corollary 17.3.6 requires.

Exercise. B2: An interval for each eigenvalue

Let $\mathbf{A} = \begin{pmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{pmatrix}$, whose eigenvalues are $2 \pm \sqrt{2}$ and 2, and let $\mathbf{E} = \frac{1}{100}\mathbf{J}$, where $\mathbf{J} \in M_3(\mathbb{R})$ is the all-ones matrix.

(a) Compute $\|\mathbf{E}\|_2$ and give an interval containing each eigenvalue of $\mathbf{A} + \mathbf{E}$.

(b) Deduce that $\mathbf{A} + \mathbf{E}$ is positive definite.

Solution.

(a) $\mathbf{J}$ is real symmetric with $\mathbf{J}^2 = 3\mathbf{J}$, so its eigenvalues satisfy $\lambda^2 = 3\lambda$ and are 3 and 0; since $\operatorname{rank} \mathbf{J} = 1$ the list is $(3, 0, 0)$. Hence $\mathbf{E}$ has eigenvalues $(\frac{3}{100}, 0, 0)$ and $\|\mathbf{E}\|_2 = \frac{3}{100}$ by Lemma 17.3.5. By Corollary 17.3.6,

$$\begin{aligned} \lambda_1(\mathbf{A} + \mathbf{E}) &\in [3.38421,\ 3.44421], \\ \lambda_2(\mathbf{A} + \mathbf{E}) &\in [1.97,\ 2.03], \\ \lambda_3(\mathbf{A} + \mathbf{E}) &\in [0.55579,\ 0.61579], \end{aligned}$$

the centers being $2 + \sqrt{2} \approx 3.41421$, 2, and $2 - \sqrt{2} \approx 0.58579$.

(b) Every interval lies strictly to the right of 0, so all three eigenvalues of the Hermitian matrix $\mathbf{A} + \mathbf{E}$ are positive and $\mathbf{A} + \mathbf{E} \succ \mathbf{O}$ by Theorem 13.1.5 (b). No characteristic polynomial was computed. (For the record, the exact eigenvalues are 2 and $\frac{403 \pm \sqrt{81609}}{200}$, that is 2, $3.44336\ldots$, and $0.58664\ldots$; the middle one is unmoved because $(1, 0, -1)$ is an eigenvector of both $\mathbf{A}$ and $\mathbf{J}$.)

Exercise. B3: What a low-rank update can do

Let $\mathbf{A} \in M_5(\mathbb{C})$ be Hermitian with $\lambda(\mathbf{A}) = (5, 4, 3, 2, 1)$, and let $\mathbf{B} \in M_5(\mathbb{C})$ be Hermitian with $\operatorname{rank} \mathbf{B} \le 2$. Find the best bounds that Corollary 17.3.8 gives for $\lambda_3(\mathbf{A} + \mathbf{B})$, and give examples showing that each is attained.

Solution. Take $r = 2$. The first inequality reads $\lambda_{i+2}(\mathbf{A} + \mathbf{B}) \le \lambda_i(\mathbf{A})$, and it mentions $\lambda_3(\mathbf{A} + \mathbf{B})$ only at $i = 1$, giving $\lambda_3(\mathbf{A} + \mathbf{B}) \le \lambda_1(\mathbf{A}) = 5$. The second reads $\lambda_i(\mathbf{A}) \le \lambda_{i-2}(\mathbf{A} + \mathbf{B})$, and it mentions $\lambda_3(\mathbf{A} + \mathbf{B})$ only at $i = 5$, giving $\lambda_3(\mathbf{A} + \mathbf{B}) \ge \lambda_5(\mathbf{A}) = 1$. So the corollary gives exactly $\lambda_3(\mathbf{A} + \mathbf{B}) \in [1, 5]$.

Both ends are attained. Write $\mathbf{A} = \mathbf{U} \operatorname{diag}(5, 4, 3, 2, 1)\mathbf{U}^*$ with $\mathbf{U}$ unitary (Corollary 12.4.3) and let $\mathbf{q}_1, \ldots, \mathbf{q}_5$ be its columns, an orthonormal eigenbasis.

For the upper bound take $\mathbf{B} = 4(\mathbf{q}_4\mathbf{q}_4^* + \mathbf{q}_5\mathbf{q}_5^*)$, which is Hermitian of rank 2. It adds 4 to the fourth and fifth eigenvalues and leaves the rest alone, so $\mathbf{A} + \mathbf{B} = \mathbf{U} \operatorname{diag}(5, 4, 3, 6, 5)\mathbf{U}^*$, whose eigenvalue list sorted decreasingly is $(6, 5, 5, 4, 3)$. Hence $\lambda_3(\mathbf{A} + \mathbf{B}) = 5 = \lambda_1(\mathbf{A})$.

For the lower bound take $\mathbf{B} = -4(\mathbf{q}_1\mathbf{q}_1^* + \mathbf{q}_2\mathbf{q}_2^*)$, again Hermitian of rank 2. Then $\mathbf{A} + \mathbf{B} = \mathbf{U} \operatorname{diag}(1, 0, 3, 2, 1)\mathbf{U}^*$, sorted $(3, 2, 1, 1, 0)$, so $\lambda_3(\mathbf{A} + \mathbf{B}) = 1 = \lambda_5(\mathbf{A})$.

17.3.5.3 C. Going deeper

Exercise. C1: The Loewner order through Weyl

Let $\mathbf{A}, \mathbf{B} \in M_n(F)$ be Hermitian.

(a) Prove that if $\mathbf{A} \succeq \mathbf{B}$ then $\lambda_i(\mathbf{A}) \geq \lambda_i(\mathbf{B})$ for every i, as a consequence of Corollary 17.3.4.

(b) Hence deduce that $|\lambda_i(\mathbf{A}) - \lambda_i(\mathbf{B})| \leq \|\mathbf{A} - \mathbf{B}\|_2$ for every i, without invoking Corollary 17.3.6.

Solution.

(a) Put $\mathbf{C} = \mathbf{A} - \mathbf{B}$, which is Hermitian, and $\mathbf{C} \succeq \mathbf{O}$ by hypothesis and Definition 13.5.1. By Theorem 13.1.4 (b) every eigenvalue of $\mathbf{C}$ is ≥ 0, in particular $\lambda_n(\mathbf{C}) \geq 0$. Applying Corollary 17.3.4 to $\mathbf{B}$ and $\mathbf{C}$,

$$\lambda_i(\mathbf{A}) = \lambda_i(\mathbf{B} + \mathbf{C}) \;\geq\; \lambda_i(\mathbf{B}) + \lambda_n(\mathbf{C}) \;\geq\; \lambda_i(\mathbf{B}).$$

This is Corollary 17.2.4.

(b) Put $\mathbf{C} = \mathbf{A} - \mathbf{B}$ again, now with no sign hypothesis, and let $t = \|\mathbf{C}\|_2$. By Lemma 17.3.5, $-t \leq \lambda_n(\mathbf{C})$ and $\lambda_1(\mathbf{C}) \leq t$, so $t\mathbf{I} - \mathbf{C} \succeq \mathbf{O}$ and $\mathbf{C} + t\mathbf{I} \succeq \mathbf{O}$ by Theorem 13.1.4 (b) applied to the Hermitian matrices $t\mathbf{I} - \mathbf{C}$ and $\mathbf{C} + t\mathbf{I}$, whose eigenvalues are $t - \lambda_q(\mathbf{C})$ and $\lambda_q(\mathbf{C}) + t$. Hence $\mathbf{B} + t\mathbf{I} \succeq \mathbf{A} \succeq \mathbf{B} - t\mathbf{I}$, and part (a) gives

$$\lambda_i(\mathbf{B}) + t = \lambda_i(\mathbf{B} + t\mathbf{I}) \;\geq\; \lambda_i(\mathbf{A}),$$
$$\lambda_i(\mathbf{A}) \geq \lambda_i(\mathbf{B} - t\mathbf{I}) = \lambda_i(\mathbf{B}) - t,$$

using $\lambda_i(\mathbf{M} + s\mathbf{I}) = \lambda_i(\mathbf{M}) + s$ for real s, which holds because $\mathbf{M}$ and $\mathbf{M} + s\mathbf{I}$ have the same eigenvectors with eigenvalues shifted by s, and shifting by a constant preserves the decreasing order. Rearranged, this is $|\lambda_i(\mathbf{A}) - \lambda_i(\mathbf{B})| \leq t$, as claimed.

Exercise. C2: The largest eigenvalue is a convex function

Let $\mathbf{A}, \mathbf{B} \in M_n(F)$ be Hermitian and let $t \in [0, 1]$.

(a) Prove that $\lambda_1(t\mathbf{A} + (1-t)\mathbf{B}) \leq t\lambda_1(\mathbf{A}) + (1-t)\lambda_1(\mathbf{B})$.

(b) Deduce that for every $c \in \mathbb{R}$ the set $\{\mathbf{M} \in M_n(F) : \mathbf{M}^* = \mathbf{M},\ \lambda_1(\mathbf{M}) \leq c\}$ is convex.

(c) Show that λ_2 is **not** convex in this sense, by producing two Hermitian 2×2 matrices and a value of t where the inequality fails.

Solution.

(a) For $s \geq 0$ and Hermitian $\mathbf{M}$ we have $\lambda_1(s\mathbf{M}) = s\lambda_1(\mathbf{M})$, since $s\mathbf{M} = \mathbf{U}(s\mathbf{D})\mathbf{U}^*$ and multiplying a decreasing list of reals by $s \geq 0$ keeps it decreasing. By Theorem 17.3.3 (a) at $i = j = 1$, applied to the Hermitian matrices $t\mathbf{A}$ and $(1-t)\mathbf{B}$,

$$\lambda_1\big(t\mathbf{A} + (1-t)\mathbf{B}\big) \leq \lambda_1(t\mathbf{A}) + \lambda_1\big((1-t)\mathbf{B}\big) = t\lambda_1(\mathbf{A}) + (1-t)\lambda_1(\mathbf{B}).$$

(b) Let $\mathbf{A}, \mathbf{B}$ lie in the set and let $t \in [0, 1]$. The matrix $t\mathbf{A} + (1-t)\mathbf{B}$ is Hermitian, and

by (a) its largest eigenvalue is at most $tc + (1-t)c = c$. So the set contains the whole segment between any two of its points.

(c) Take $\mathbf{A} = \operatorname{diag}(1,0)$, $\mathbf{B} = \operatorname{diag}(0,1)$, and $t = \frac{1}{2}$. Then $\lambda_2(\mathbf{A}) = \lambda_2(\mathbf{B}) = 0$, while $t\mathbf{A} + (1-t)\mathbf{B} = \frac{1}{2}\mathbf{I}_2$ has $\lambda_2 = \frac{1}{2}$. So $\frac{1}{2} \le \frac{1}{2}\cdot 0 + \frac{1}{2}\cdot 0 = 0$ is false. The reason is visible in Theorem 17.3.3: at $i = j = 2$ the index on the left is 3, not 2.

Exercise. C3: Sharpness of the whole family

Fix n and indices i, j with $i + j - 1 \le n$. Construct Hermitian $\mathbf{A}, \mathbf{B} \in M_n(\mathbb{R})$, both diagonal, for which

$$\lambda_{i+j-1}(\mathbf{A} + \mathbf{B}) = \lambda_i(\mathbf{A}) + \lambda_j(\mathbf{B}).$$

Hint: with diagonal matrices every quantity in sight is a sorting problem; arrange for the $(i+j-1)$-st largest entry of the sum to sit in a coordinate where $\mathbf{A}$ contributes its i-th largest and $\mathbf{B}$ its j-th largest.

Solution. Let $k = i + j - 1$ and let $\mathbf{A} = \operatorname{diag}(a_1, \dots, a_n)$ and $\mathbf{B} = \operatorname{diag}(b_1, \dots, b_n)$ with

$$a_p = \begin{cases} 1 & p \le i-1, \\ 0 & p \ge i, \end{cases} \qquad b_p = \begin{cases} 0 & p \le i-1, \\ 1 & i \le p \le k, \\ 0 & p > k. \end{cases}$$

Both are diagonal with entries 0 and 1, hence Hermitian, and their diagonals are already their eigenvalue lists up to sorting. The matrix $\mathbf{A}$ has $i-1$ entries equal to 1, so $\lambda_i(\mathbf{A}) = 0$. The matrix $\mathbf{B}$ has $k - i + 1 = j$ entries equal to 1, so $\lambda_j(\mathbf{B}) = 1$. The sum is $\mathbf{A} + \mathbf{B} = \operatorname{diag}(1, \dots, 1, 0, \dots, 0)$ with exactly k ones, since the ones of $\mathbf{A}$ occupy coordinates $1, \dots, i-1$ and those of $\mathbf{B}$ occupy $i, \dots, k$, and these two sets of coordinates do not overlap. Hence $\lambda_k(\mathbf{A} + \mathbf{B}) = 1 = \lambda_i(\mathbf{A}) + \lambda_j(\mathbf{B})$.

Every inequality of the first family is therefore attained, for every admissible pair (i, j), already by 0–1 diagonal matrices. Applying Lemma 17.3.2 to $-\mathbf{A}$ and $-\mathbf{B}$, as in the proof of Theorem 17.3.3 (b), turns this family of equalities into one showing that the second family is attained as well.

17.4 Deleting a Row and a Column

Take a Hermitian matrix, cross out the last row and the last column, and look at what is left. It is again Hermitian, one size smaller, and it has its own eigenvalues. Two lists of numbers are now in play, n of them and $n-1$ of them, and they cannot be independent of each other: the smaller matrix is the larger one restricted to a hyperplane, and Theorem 17.2.3 describes every eigenvalue by what happens on subspaces. Restricting the supply of subspaces can only make the description worse in one direction and better in the other, and the exact bookkeeping is this section's subject.

The answer is as tight as it could be. The $n-1$ eigenvalues of the corner fall into the $n-1$ gaps of the eigenvalues of the whole, one to a gap. Nothing stronger holds: every interlacing pattern actually occurs (Exercise C2 proves the case $n = 2$).

Throughout this section the field is $\mathbb{R}$ or $\mathbb{C}$, and every matrix whose eigenvalues are written down is Hermitian, with eigenvalues real and indexed **decreasingly**. For an index set $I \subseteq \{1, \dots, n\}$ we write $\mathbf{A}_{I,I}$ for the **principal submatrix** of $\mathbf{A}$ on the rows and columns indexed by I, as in Definition 7.7.1. A principal submatrix of a Hermitian matrix is Hermitian, since $(\mathbf{A}_{I,I})^* = (\mathbf{A}^*)_{I,I} = \mathbf{A}_{I,I}$.

17.4.1 A matrix restricted to a coordinate hyperplane

Deleting row m and column m is not an arbitrary mutilation of $\mathbf{A}$. It is what $\mathbf{A}$ looks like to vectors whose m-th coordinate is zero, and that is the only observation the proof needs.

Lemma 17.4.1. Deleting a Row and a Column Is Restricting to a Hyperplane

Let $\mathbf{A} \in M_n(F)$ be Hermitian with $n \geq 2$, fix m with $1 \leq m \leq n$, put $I = \{1, \dots, n\} \setminus \{m\}$ and $\mathbf{B} = \mathbf{A}_{I,I} \in M_{n-1}(F)$. Let

$$H = \{\mathbf{x} \in F^n : x_m = 0\},$$

and let $\varphi \colon H \to F^{n-1}$ delete the m-th coordinate. Then φ is a linear isomorphism, and for every $\mathbf{x} \in H$,

$$\langle \mathbf{A}\mathbf{x}, \mathbf{x} \rangle = \langle \mathbf{B}\varphi(\mathbf{x}), \varphi(\mathbf{x}) \rangle, \qquad \|\mathbf{x}\| = \|\varphi(\mathbf{x})\|.$$

In particular $R_{\mathbf{A}}(\mathbf{x}) = R_{\mathbf{B}}(\varphi(\mathbf{x}))$ for every $\mathbf{x} \in H$ with $\mathbf{x} \neq \mathbf{0}$.

Proof. The map φ is linear, being given by deleting a coordinate, and it is bijective from H to F^{n-1} because inserting 0 in position m inverts it. Write $I = \{i_1 < \cdots < i_{n-1}\}$, so that $\varphi(\mathbf{x})_r = x_{i_r}$ and $(\mathbf{B})_{rs} = a_{i_r i_s}$ by Definition 7.7.1.

Let $\mathbf{x} \in H$. Expanding the inner product of F^n, which is linear in the first slot and conjugate-linear in the second,

$$\langle \mathbf{A}\mathbf{x}, \mathbf{x} \rangle = \sum_{p=1}^{n} \sum_{q=1}^{n} a_{pq}\, x_q\, \overline{x_p}.$$

Every term with $p = m$ or $q = m$ vanishes, because $x_m = 0$. The remaining terms are indexed by $p, q \in I$, say $p = i_r$ and $q = i_s$, and there the summand is $a_{i_r i_s}\varphi(\mathbf{x})_s\overline{\varphi(\mathbf{x})_r} = (\mathbf{B})_{rs}\varphi(\mathbf{x})_s\overline{\varphi(\mathbf{x})_r}$. Summing over r, s gives $\langle \mathbf{B}\varphi(\mathbf{x}), \varphi(\mathbf{x}) \rangle$. The same deletion of a zero coordinate gives $\|\mathbf{x}\|^2 = \sum_{p \neq m} |x_p|^2 = \|\varphi(\mathbf{x})\|^2$. Dividing the first identity by the second, which is positive when $\mathbf{x} \neq \mathbf{0}$, gives the statement about the Rayleigh quotients (Definition 17.1.1). This proves the lemma. ∎

Because φ is an isomorphism, it matches the k-dimensional subspaces of H with the k-dimensional subspaces of F^{n-1}, for every k. So a statement about $R_{\mathbf{B}}$ on subspaces of F^{n-1} *is* a statement about $R_{\mathbf{A}}$ on subspaces of H, and the only thing lost in passing from $\mathbf{A}$ to $\mathbf{B}$ is the subspaces of F^n that are not inside H.

17.4.2 Interlacing

Theorem 17.4.2. Cauchy's Interlacing Theorem

Let $\mathbf{A} \in M_n(F)$ be Hermitian with $n \geq 2$, let $I \subseteq \{1, \ldots, n\}$ with $|I| = n - 1$, and let $\mathbf{B} = \mathbf{A}_{I,I}$. Then

$$\lambda_k(\mathbf{A}) \;\geq\; \lambda_k(\mathbf{B}) \;\geq\; \lambda_{k+1}(\mathbf{A}) \qquad (1 \leq k \leq n-1),$$

that is,

$$\lambda_1(\mathbf{A}) \geq \lambda_1(\mathbf{B}) \geq \lambda_2(\mathbf{A}) \geq \lambda_2(\mathbf{B}) \geq \cdots \geq \lambda_{n-1}(\mathbf{B}) \geq \lambda_n(\mathbf{A}).$$

Idea. Both halves of Theorem 17.2.3 are used, one for each inequality, and each time the move is the same: a subspace of H is also a subspace of F^n.

① For $\lambda_k(\mathbf{B}) \leq \lambda_k(\mathbf{A})$ use the **max–min** form. The best k-dimensional subspace for $\mathbf{B}$ lives inside H; transplanted to F^n it is one competitor among more, so the maximum over all of them is at least as large.

② For $\lambda_{k+1}(\mathbf{A}) \leq \lambda_k(\mathbf{B})$ use the **min–max** form, and count dimensions. The best subspace for $\mathbf{B}$ there has dimension $(n-1) - k + 1 = n - k$, and $n - k = n - (k+1) + 1$ is exactly the dimension the min–max form for $\lambda_{k+1}(\mathbf{A})$ asks for. So the same subspace, viewed in F^n, is admissible for $\mathbf{A}$ at the index $k+1$.

The dimension count that makes min–max work at all is Lemma 17.2.1, and it has already been spent inside Theorem 17.2.3; nothing more is needed here than the coincidence of the two numbers in ②.

Proof. Let m be the index not in I, and let H and φ be as in Lemma 17.4.1, so that $R_{\mathbf{A}}(\mathbf{x}) = R_{\mathbf{B}}(\varphi(\mathbf{x}))$ for every nonzero $\mathbf{x} \in H$. Both $\mathbf{A}$ and $\mathbf{B}$ are Hermitian, so Theorem 17.2.3 applies to each, to $\mathbf{A}$ with n and to $\mathbf{B}$ with $n-1$ in the role of the size. Fix k with $1 \leq k \leq n-1$.

$(\lambda_k(\mathbf{B}) \leq \lambda_k(\mathbf{A}))$ By the max–min form of Theorem 17.2.3 for $\mathbf{B}$, there is a subspace $W' \leq F^{n-1}$ with $\dim W' = k$ and

$$\lambda_k(\mathbf{B}) = \min_{\mathbf{0} \neq \mathbf{y} \in W'} R_{\mathbf{B}}(\mathbf{y}).$$

Put $W = \varphi^{-1}(W') \leq H \leq F^n$. Since φ is an isomorphism, $\dim W = k$, and φ carries the nonzero vectors of W onto the nonzero vectors of W', so

$$\min_{\mathbf{0} \neq \mathbf{x} \in W} R_{\mathbf{A}}(\mathbf{x}) = \min_{\mathbf{0} \neq \mathbf{y} \in W'} R_{\mathbf{B}}(\mathbf{y}) = \lambda_k(\mathbf{B}).$$

Now W is one of the k-dimensional subspaces of F^n over which the max–min form for $\mathbf{A}$ takes its maximum, so $\lambda_k(\mathbf{A}) \geq \lambda_k(\mathbf{B})$.

$(\lambda_{k+1}(\mathbf{A}) \leq \lambda_k(\mathbf{B}))$ By the min–max form of Theorem 17.2.3 for $\mathbf{B}$, whose matrices have size $n-1$, there is a subspace $W' \leq F^{n-1}$ with $\dim W' = (n-1) - k + 1 = n - k$ and

$$\lambda_k(\mathbf{B}) = \max_{\mathbf{0} \neq \mathbf{y} \in W'} R_{\mathbf{B}}(\mathbf{y}).$$

Put $W = \varphi^{-1}(W') \leq F^n$, of dimension $n - k$. Since $1 \leq k+1 \leq n$ and $n - (k+1) + 1 = n - k = \dim W$, the subspace W is admissible in the min–max form for $\lambda_{k+1}(\mathbf{A})$, and that

form takes the minimum over all such subspaces. Hence

$$\lambda_{k+1}(\mathbf{A}) \;\le\; \max_{\mathbf{0}\neq\mathbf{x}\in W} R_{\mathbf{A}}(\mathbf{x}) = \max_{\mathbf{0}\neq\mathbf{y}\in W'} R_{\mathbf{B}}(\mathbf{y}) = \lambda_k(\mathbf{B}).$$

This proves the theorem.

■

The picture to keep is one of nesting. The interval $[\lambda_n(\mathbf{A}), \lambda_1(\mathbf{A})]$ contains the interval $[\lambda_{n-1}(\mathbf{B}), \lambda_1(\mathbf{B})]$, and inside it the two lists alternate, an eigenvalue of $\mathbf{B}$ between each consecutive pair of eigenvalues of $\mathbf{A}$. Deleting a row and a column can only shrink the spectrum inward, never spread it out.

Quick check. Both inequalities in Theorem 17.4.2 can be equalities. Give a matrix $\mathbf{A}$ for which $\lambda_1(\mathbf{A}) = \lambda_1(\mathbf{B})$, and one for which $\lambda_1(\mathbf{B}) = \lambda_2(\mathbf{A})$.

Solution. For the first, $\mathbf{A} = \mathbf{I}_2$ with $I = \{1\}$: then $\mathbf{B} = (1)$ and $\lambda_1(\mathbf{A}) = \lambda_1(\mathbf{B}) = 1$. For the second, $\mathbf{A} = \operatorname{diag}(1, 0)$ with $I = \{2\}$: then $\mathbf{B} = \mathbf{O}_1$, and $\lambda_1(\mathbf{B}) = 0 = \lambda_2(\mathbf{A})$. Nothing in the theorem forces a gap; when a gap is forced is exactly the question settled later in this section.

Deleting one index at a time and iterating gives the general statement.

Corollary 17.4.3. Deleting m Rows and the Same m Columns

Let $\mathbf{A} \in M_n(F)$ be Hermitian, let $0 \le m \le n-1$, let $I \subseteq \{1, \dots, n\}$ with $|I| = n - m$, and let $\mathbf{B} = \mathbf{A}_{I,I}$. Then

$$\lambda_i(\mathbf{A}) \;\ge\; \lambda_i(\mathbf{B}) \;\ge\; \lambda_{i+m}(\mathbf{A}) \qquad (1 \le i \le n-m).$$

Proof. Induction on m. For $m = 0$ we have $\mathbf{B} = \mathbf{A}$ and both inequalities are equalities.

Let $m \ge 1$ and suppose the statement holds for $m-1$, for every Hermitian matrix. Choose an index $j \notin I$ and put $I' = I \cup \{j\}$, so $|I'| = n - m + 1$, and $\mathbf{B}' = \mathbf{A}_{I',I'} \in M_{n-m+1}(F)$, which is Hermitian. Since $I \subseteq I'$, taking the principal submatrix on I of $\mathbf{B}'$ means keeping the rows and columns of $\mathbf{A}$ indexed by I, so $(\mathbf{B}')_{I,I} = \mathbf{B}$ after relabeling the indices of I inside I'; and $|I| = |I'| - 1$.

Fix i with $1 \le i \le n - m$. By the inductive hypothesis applied to $\mathbf{A}$ and $\mathbf{B}'$,

$$\lambda_i(\mathbf{A}) \;\ge\; \lambda_i(\mathbf{B}') \;\ge\; \lambda_{i+m-1}(\mathbf{A}),$$

which is available because $i \le n - m < n - m + 1$. By Theorem 17.4.2 applied to $\mathbf{B}'$ and its principal submatrix $\mathbf{B}$, which is one size smaller,

$$\lambda_k(\mathbf{B}') \;\ge\; \lambda_k(\mathbf{B}) \;\ge\; \lambda_{k+1}(\mathbf{B}') \qquad (1 \le k \le n-m).$$

Combining, $\lambda_i(\mathbf{B}) \le \lambda_i(\mathbf{B}') \le \lambda_i(\mathbf{A})$, which is the first inequality. For the second, take $k = i$ in the display above and then use the inductive hypothesis at the index $i+1$, legitimate since

$i+1 \le n-m+1$:

$$\lambda_i(\mathbf{B}) \;\ge\; \lambda_{i+1}(\mathbf{B}') \;\ge\; \lambda_{(i+1)+m-1}(\mathbf{A}) = \lambda_{i+m}(\mathbf{A}).$$

This proves the corollary.

■

The extreme case $m = n-1$ is worth writing down, because it recovers a fact about diagonal entries that also follows from Chapter 13's description of the extreme eigenvalues.

Example. The diagonal lies inside the spectrum

Let $\mathbf{A} \in M_n(F)$ be Hermitian. Show that $\lambda_n(\mathbf{A}) \le a_{ii} \le \lambda_1(\mathbf{A})$ for every i.

Solution. Take $I = \{i\}$, so $m = n-1$ and $\mathbf{B} = \mathbf{A}_{I,I} = (a_{ii})$, a 1×1 Hermitian matrix whose single eigenvalue is a_{ii} — real, as a diagonal entry of a Hermitian matrix must be. Corollary 17.4.3 at $i = 1$ gives

$$\lambda_1(\mathbf{A}) \;\ge\; \lambda_1(\mathbf{B}) = a_{ii} \;\ge\; \lambda_{1+(n-1)}(\mathbf{A}) = \lambda_n(\mathbf{A}).$$

The same conclusion follows from Lemma 13.5.2 applied to the unit vector $\mathbf{e}_i$, since $\langle \mathbf{A}\mathbf{e}_i, \mathbf{e}_i\rangle = a_{ii}$. What interlacing adds is the version for larger index sets: any k diagonal entries are the diagonal of a $k \times k$ principal submatrix, whose i-th eigenvalue is trapped between $\lambda_{i+n-k}(\mathbf{A})$ and $\lambda_i(\mathbf{A})$. How much more can be said about the diagonal of a Hermitian matrix, given its spectrum, is the subject of Section 8.

17.4.3 When is the interlacing strict?

Theorem 17.4.2 allows equalities, and the Quick check produced them. Deciding when they occur is the delicate part of the subject, and it is easiest to see when the deleted index is the last one, so that $\mathbf{A}$ is presented as $\mathbf{B}$ with a border added.

Theorem 17.4.4. Bordering a Hermitian Matrix

Let $\mathbf{B} \in M_{n-1}(F)$ be Hermitian with $n \ge 2$, let $\mathbf{b} \in F^{n-1}$, let $c \in \mathbb{R}$, and let

$$\mathbf{A} = \begin{pmatrix} \mathbf{B} & \mathbf{b} \\ \mathbf{b}^* & c \end{pmatrix} \in M_n(F),$$

which is Hermitian and has $\mathbf{A}_{I,I} = \mathbf{B}$ for $I = \{1, \dots, n-1\}$. For an eigenvalue ν of $\mathbf{B}$, write P_ν for the orthogonal projection of F^{n-1} onto the eigenspace $E_\nu(\mathbf{B})$, and $m_\nu = \dim E_\nu(\mathbf{B})$.

(a) If $P_\nu \mathbf{b} \neq \mathbf{0}$, then ν is a root of $p_{\mathbf{A}}$ of multiplicity **exactly** $m_\nu - 1$; in particular, if $m_\nu = 1$ then ν is not an eigenvalue of $\mathbf{A}$.

(b) If $P_\nu \mathbf{b} = \mathbf{0}$, then ν is a root of $p_{\mathbf{A}}$ of multiplicity **at least** m_ν.

(c) The interlacing $\lambda_1(\mathbf{A}) \ge \lambda_1(\mathbf{B}) \ge \cdots \ge \lambda_{n-1}(\mathbf{B}) \ge \lambda_n(\mathbf{A})$ is strict at every one of its $2(n-1)$ inequalities if and only if $\mathbf{B}$ has $n-1$ **distinct** eigenvalues **and** $P_\nu \mathbf{b} \neq \mathbf{0}$ for every eigenvalue ν of $\mathbf{B}$.

Idea. Write the characteristic polynomial of $\mathbf{A}$ down. With $\mathbf{B}$ diagonalized, a Schur complement turns $\det(x\mathbf{I}-\mathbf{A})$ into

$$\Big(\prod_k (x-\mu_k)\Big)\Big((x-c)-\sum_k \frac{|\tilde b_k|^2}{x-\mu_k}\Big),$$

and clearing the denominators leaves a polynomial identity in which each μ_k appears with one factor fewer than one might expect — provided $P_\nu \mathbf{b} \neq \mathbf{0}$, that is, provided the $|\tilde b_k|^2$ over the group belonging to that eigenvalue do not all vanish (for a repeated eigenvalue the individual $\tilde b_k$ depend on the choice of basis, but their sum does not). Part (c) then costs almost nothing: strict interlacing means the combined list of $2n-1$ numbers is strictly decreasing, hence has no repetitions at all, and (a) and (b) say exactly which repetitions are possible.

Proof. **Step 1: diagonalize the corner.** By the spectral theorem (Corollary 12.4.3 over $\mathbb{C}$, Corollary 12.5.2 over $\mathbb{R}$) write $\mathbf{B} = \mathbf{U}\mathbf{D}\mathbf{U}^*$ with $\mathbf{U} \in M_{n-1}(F)$ unitary and $\mathbf{D} = \operatorname{diag}(\mu_1, \dots, \mu_{n-1})$ carrying the eigenvalues of $\mathbf{B}$ in decreasing order, the columns $\mathbf{u}_1, \dots, \mathbf{u}_{n-1}$ of $\mathbf{U}$ being an orthonormal basis of eigenvectors. Let $\mathbf{V} = \left(\begin{smallmatrix} \mathbf{U} & \mathbf{0} \\ \mathbf{0}^* & 1 \end{smallmatrix}\right) \in M_n(F)$, which is unitary because $\mathbf{V}^*\mathbf{V} = \mathbf{I}_n$. Multiplying in blocks (Theorem 8.1.1) gives

$$\mathbf{V}^*\mathbf{A}\mathbf{V} = \begin{pmatrix} \mathbf{D} & \tilde{\mathbf{b}} \\ \tilde{\mathbf{b}}^* & c \end{pmatrix}, \qquad \tilde{\mathbf{b}} = \mathbf{U}^*\mathbf{b}.$$

Similar matrices have the same characteristic polynomial (Theorem 7.10.3), so we may compute with the right-hand side. Moreover $\tilde b_k = (\mathbf{U}^*\mathbf{b})_k = \langle \mathbf{b}, \mathbf{u}_k \rangle$, and the $\mathbf{u}_k$ with $\mu_k = \nu$ form an orthonormal basis of $E_\nu(\mathbf{B})$: they lie in it, and a vector $\mathbf{x} = \sum_k c_k \mathbf{u}_k$ with $\mathbf{B}\mathbf{x} = \nu\mathbf{x}$ has $(\mu_k - \nu)c_k = 0$ for every k, so $c_k = 0$ unless $\mu_k = \nu$. Hence exactly m_ν of the μ_k equal ν, and

$$P_\nu \mathbf{b} = \sum_{k:\mu_k=\nu} \langle \mathbf{b}, \mathbf{u}_k \rangle \, \mathbf{u}_k \quad \text{and} \quad \|P_\nu \mathbf{b}\|^2 = \sum_{k:\mu_k=\nu} |\tilde b_k|^2.$$

In particular $P_\nu \mathbf{b} = \mathbf{0}$ exactly when $\tilde b_k = 0$ for every k with $\mu_k = \nu$.

Step 2: the characteristic polynomial. We claim that, as polynomials in x,

$$\begin{aligned} p_{\mathbf{A}}(x) = (x-c)\prod_{k=1}^{n-1}(x-\mu_k) \\ -\sum_{k=1}^{n-1} |\tilde b_k|^2 \prod_{l\neq k}(x-\mu_l). \end{aligned} \tag{17.4.1}$$

Claim. Eq. 17.4.1 holds.

Proof. Fix a scalar x_0 that is not an eigenvalue of $\mathbf{B}$. Then $x_0\mathbf{I}_{n-1} - \mathbf{D}$ is invertible, being diagonal with nonzero diagonal entries, so Theorem 8.3.2 (a), applied to

$$\mathbf{M} = x_0\mathbf{I}_n - \mathbf{V}^*\mathbf{A}\mathbf{V} = \begin{pmatrix} x_0\mathbf{I}_{n-1} - \mathbf{D} & -\tilde{\mathbf{b}} \\ -\tilde{\mathbf{b}}^* & x_0 - c \end{pmatrix}$$

with the invertible block in the top left, gives $\det \mathbf{M} = \det(x_0\mathbf{I} - \mathbf{D}) \cdot \det(\mathbf{M}/(x_0\mathbf{I} - \mathbf{D}))$,

where the Schur complement is the 1×1 matrix

$$(x_0 - c) - (-\tilde{\mathbf{b}}^*)(x_0\mathbf{I} - \mathbf{D})^{-1}(-\tilde{\mathbf{b}}) = (x_0 - c) - \sum_k \frac{|\tilde{b}_k|^2}{x_0 - \mu_k}.$$

Multiplying out $\det(x_0\mathbf{I} - \mathbf{D}) = \prod_k (x_0 - \mu_k)$ against this scalar, and canceling one factor $(x_0 - \mu_k)$ in the k-th summand, gives exactly the right-hand side of Eq. 17.4.1 evaluated at x_0. So the two sides of Eq. 17.4.1, which are polynomials of degree n, agree at every scalar outside the finite set $\mathrm{spec}(\mathbf{B})$. Their difference is therefore a polynomial of degree at most n with infinitely many roots in F, and a nonzero polynomial of degree at most n has at most n roots (Corollary 6.1.4); so the difference is zero.

■

Step 3: multiplicities. Let $\nu_1 > \cdots > \nu_p$ be the distinct eigenvalues of $\mathbf{B}$, with multiplicities $m_1, \ldots, m_p$ summing to $n - 1$, and put $\beta_t = \|P_{\nu_t}\mathbf{b}\|^2$. Grouping equal μ_k in Eq. 17.4.1 — the product $\prod_{l \neq k}(x - \mu_l)$ is the same for all k in one group, and the coefficients $|\tilde{b}_k|^2$ add up to β_t by Step 1 — turns it into

$$p_{\mathbf{A}}(x) = (x - c)\prod_{t=1}^{p}(x - \nu_t)^{m_t} - \sum_{t=1}^{p} \beta_t\,(x - \nu_t)^{m_t - 1}\prod_{s \neq t}(x - \nu_s)^{m_s}.$$

Fix t. Every term on the right is divisible by $(x - \nu_t)^{m_t - 1}$, so $p_{\mathbf{A}}(x) = (x - \nu_t)^{m_t - 1} g(x)$ with

$$g(x) = (x - c)(x - \nu_t)\prod_{s \neq t}(x - \nu_s)^{m_s} - \beta_t \prod_{s \neq t}(x - \nu_s)^{m_s} - \sum_{u \neq t} \beta_u\,(x - \nu_t)(x - \nu_u)^{m_u - 1}\prod_{s \neq u,t}(x - \nu_s)^{m_s}.$$

Evaluate at $x = \nu_t$. The first term vanishes because of its factor $(x - \nu_t)$, and so does every summand of the third, for the same reason. What is left is

$$g(\nu_t) = -\beta_t \prod_{s \neq t}(\nu_t - \nu_s)^{m_s},$$

and the product is nonzero because the ν_s are distinct. So $g(\nu_t) \neq 0$ exactly when $\beta_t \neq 0$, that is, exactly when $P_{\nu_t}\mathbf{b} \neq \mathbf{0}$. In that case ν_t is a root of $p_{\mathbf{A}}$ of multiplicity exactly $m_t - 1$, which proves (a); in the other case $(x - \nu_t)$ divides g, so the multiplicity is at least m_t, which proves (b).

Step 4: strictness. ($\Leftarrow$) Suppose $\mathbf{B}$ has $n - 1$ distinct eigenvalues, so every $m_t = 1$, and $P_\nu\mathbf{b} \neq \mathbf{0}$ for each of them. By (a), no eigenvalue of $\mathbf{B}$ is an eigenvalue of $\mathbf{A}$. Theorem 17.4.2 gives $\lambda_k(\mathbf{A}) \geq \lambda_k(\mathbf{B}) \geq \lambda_{k+1}(\mathbf{A})$ for each k; since $\lambda_k(\mathbf{B})$ is an eigenvalue of $\mathbf{B}$ and $\lambda_k(\mathbf{A})$, $\lambda_{k+1}(\mathbf{A})$ are eigenvalues of $\mathbf{A}$, neither inequality can be an equality. So all $2(n - 1)$ of them are strict.

($\Rightarrow$) Suppose all of them are strict, that is,

$$\lambda_1(\mathbf{A}) > \lambda_1(\mathbf{B}) > \lambda_2(\mathbf{A}) > \cdots > \lambda_{n-1}(\mathbf{B}) > \lambda_n(\mathbf{A}).$$

Reading two steps at a time gives $\lambda_k(\mathbf{B}) > \lambda_{k+1}(\mathbf{A}) > \lambda_{k+1}(\mathbf{B})$ for $1 \le k \le n-2$, so the eigenvalues of $\mathbf{B}$ are distinct. And the displayed chain is strictly decreasing, so its $2n-1$ entries are pairwise distinct; in particular no eigenvalue of $\mathbf{B}$ is an eigenvalue of $\mathbf{A}$. By (b), read as a contrapositive, $P_\nu \mathbf{b} \neq \mathbf{0}$ for every eigenvalue ν of $\mathbf{B}$. This proves the theorem.

■

Remark. Part (b) is deliberately one-sided. When $P_\nu \mathbf{b} = \mathbf{0}$ the multiplicity of ν in $p_{\mathbf{A}}$ may be m_ν or $m_\nu + 1$ — never more, since by Cauchy interlacing q equal eigenvalues ν of $\mathbf{A}$ squeeze $q-1$ eigenvalues of $\mathbf{B}$ to ν — and both occur: $\mathbf{B} = \operatorname{diag}(3,1)$, $\mathbf{b} = (0,1)$, $c = 0$ gives $p_{\mathbf{A}}(x) = (x-3)(x^2-x-1)$, where 3 keeps multiplicity $1 = m_3$; while $\mathbf{B} = (\mu)$, $\mathbf{b} = \mathbf{0}$, $c = \mu$ gives $\mathbf{A} = \mu \mathbf{I}_2$, where μ has multiplicity $2 = m_\mu + 1$. Deciding between the two needs the value of c as well, and Theorem 17.4.4 does not attempt it; what it does settle completely is the strictness question (c), which is what interlacing asks.

Warning. **Strictness needs both clauses, and a nonzero border is not enough.** It is tempting to say that the interlacing is strict as soon as $\mathbf{b} \neq \mathbf{0}$, or as soon as $\mathbf{b}$ meets every eigenspace of $\mathbf{B}$. The first is wrong even when the eigenvalues of $\mathbf{B}$ are distinct (Exercise B3(a)); the second fails exactly when $\mathbf{B}$ has a repeated eigenvalue. Take $\mathbf{B} = 2\mathbf{I}_2$, $\mathbf{b} = (1,1)$, and $c = 2$. Here $\mathbf{b}$ has a nonzero component in the one eigenspace $E_2(\mathbf{B}) = F^2$, and yet $\lambda_1(\mathbf{B}) = \lambda_2(\mathbf{B}) = 2$, so the middle of the interlacing chain reads $2 \ge \lambda_2(\mathbf{A}) \ge 2$ and $\lambda_2(\mathbf{A}) = 2$ with no freedom at all. A repeated eigenvalue of $\mathbf{B}$ is passed on to $\mathbf{A}$, by Theorem 17.4.4 (a) with $m_\nu = 2$.

To apply Theorem 17.4.4 when the deleted index is not the last one, permute. Let $I = \{i_1 < \cdots < i_{n-1}\} = \{1, \dots, n\} \setminus \{m\}$, and let $\mathbf{P}$ be the matrix with columns $\mathbf{e}_{i_1}, \dots, \mathbf{e}_{i_{n-1}}, \mathbf{e}_m$, a permutation matrix. Its columns are orthonormal, so $\mathbf{P}^\top \mathbf{P} = \mathbf{I}$ and $\mathbf{P}$ is unitary (Lemma 3.4.7 (b)); hence $\mathbf{P}^\top \mathbf{A} \mathbf{P}$ is Hermitian and similar to $\mathbf{A}$, with the same eigenvalues. Its (r,s) entry is $\mathbf{e}_{i_r}^\top \mathbf{A} \mathbf{e}_{i_s} = a_{i_r i_s}$ for $r, s \le n-1$, so its leading $(n-1) \times (n-1)$ corner is $\mathbf{A}_{I,I}$, and its border is the column $(a_{i_1 m}, \dots, a_{i_{n-1} m})$ with corner entry a_{mm}. So $\mathbf{P}^\top \mathbf{A} \mathbf{P}$ is a bordered matrix with corner $\mathbf{A}_{I,I}$, and Theorem 17.4.4 applies to it verbatim.

The strictness criterion pays a debt from Chapter 11. The solution of Exercise (C1: Interlacing roots) proved by hand that the roots of consecutive orthogonal polynomials interlace for degrees $1, 2, 3$, and then said: "The general case is not proved this way. The coefficients a_k and b_k assemble into a symmetric tridiagonal matrix whose characteristic polynomial is p_k, and the interlacing then follows from the interlacing theorem for the eigenvalues of a symmetric matrix and its leading submatrices, which is Chapter 17's business." Two things in that sentence were asserted, not proved: that the characteristic polynomial is p_k, and that the interlacing is **strict**, which is what "exactly one root strictly between" demands. Plain Theorem 17.4.2 gives only $\ge$. Strictness needs Theorem 17.4.4 (c), and (c) needs one fact about tridiagonal matrices.

Lemma 17.4.5. Eigenvectors of an Unreduced Tridiagonal Matrix

Let $\mathbf{T} \in M_k(\mathbb{R})$ be symmetric and tridiagonal, with every entry $t_{i,i+1} = t_{i+1,i}$ $(1 \le i \le k-1)$ **nonzero**. Then every eigenvector $\mathbf{u}$ of $\mathbf{T}$ has $u_k \neq 0$. Consequently every eigenspace of $\mathbf{T}$ has dimension 1, so $\mathbf{T}$ has k distinct eigenvalues.

Proof. Let $\mathbf{Tu} = \nu\mathbf{u}$ and suppose $u_k = 0$; we show $\mathbf{u} = \mathbf{0}$. If $k = 1$ this is immediate. Otherwise row k of $\mathbf{Tu} = \nu\mathbf{u}$ reads $t_{k,k-1}u_{k-1}+t_{kk}u_k = \nu u_k$, so $t_{k,k-1}u_{k-1} = 0$ and $u_{k-1} = 0$, since $t_{k,k-1} \neq 0$. Now suppose $u_i = u_{i+1} = 0$ for some $2 \leq i \leq k-1$. Row i reads $t_{i,i-1}u_{i-1} + t_{ii}u_i + t_{i,i+1}u_{i+1} = \nu u_i$, so $t_{i,i-1}u_{i-1} = 0$ and $u_{i-1} = 0$. Walking up from $i = k-1$ to $i = 2$ gives $u_{k-2} = \cdots = u_1 = 0$, so $\mathbf{u} = \mathbf{0}$, and $\mathbf{u}$ is not an eigenvector.

If some eigenspace $E_\nu(\mathbf{T})$ contained two independent vectors $\mathbf{u}, \mathbf{u}'$, then $u'_k\mathbf{u} - u_k\mathbf{u}'$ would be a nonzero vector of $E_\nu(\mathbf{T})$ — nonzero by independence, since $u_k \neq 0$ — with last coordinate 0, which we have just excluded. So each eigenspace is a line. Since $\mathbf{T}$ is real symmetric, its eigenspaces have dimensions adding up to k (Corollary 12.5.2), so there are k of them, that is, k distinct eigenvalues. This proves the lemma.

■

Corollary 17.4.6. Zeros of Orthogonal Polynomials Interlace

Optional: nothing later depends on this.

Let (p_k) be the monic orthogonal polynomial sequence of an inner product on $\mathbb{R}[x]$ to which Theorem 11.10.3 applies, with coefficients a_k and $b_k > 0$; for the node inner product on $\mathbb{R}[x]_{\leq n}$ read everything below for $k+1 \leq n$. For $k \geq 1$ let $\mathbf{T}_k$ (the **Jacobi matrix** of the sequence) be the $k \times k$ real symmetric tridiagonal matrix with diagonal $a_0, a_1, \ldots, a_{k-1}$ and entries $\sqrt{b_1}, \ldots, \sqrt{b_{k-1}}$ just above and just below it.

(a) $p_{\mathbf{T}_k} = p_k$; so the roots of p_k are the eigenvalues of $\mathbf{T}_k$, and they are real and distinct.

(b) If $r_1 > \cdots > r_{k+1}$ are the roots of p_{k+1} and $s_1 > \cdots > s_k$ those of p_k, then

$$r_1 > s_1 > r_2 > s_2 > \cdots > s_k > r_{k+1}.$$

Idea. The polynomials p_k are the characteristic polynomials of a nested family of symmetric tridiagonal matrices, each the leading corner of the next. So their zeros are eigenvalues of a matrix and of its corner, and Cauchy interlacing applies at once — nonstrictly. Strictness is the bordered-matrix criterion, and both of its clauses come from one fact about tridiagonal matrices with nonzero off-diagonal entries: an eigenvector can never end in a zero.

Proof.

(a) Put $D_k(x) = \det(x\mathbf{I}_k - \mathbf{T}_k)$ and $D_0 = 1$. For each real x_0, the matrix $x_0\mathbf{I}_k - \mathbf{T}_k$ is tridiagonal in the sense of Chapter 7, with diagonal entries $x_0 - a_{i-1}$ and off-diagonal entries $-\sqrt{b_i}$ in both positions $(i, i+1)$ and $(i+1, i)$, so Theorem 7.8.3 gives $D_1(x_0) = x_0 - a_0$ and

$$\begin{aligned} D_{k+1}(x_0) &= (x_0 - a_k)D_k(x_0) - (-\sqrt{b_k})^2 D_{k-1}(x_0) \\ &= (x_0 - a_k)D_k(x_0) - b_k D_{k-1}(x_0) \end{aligned}$$

for $k \geq 1$. Here $D_k(x_0) = \det(x_0\mathbf{I} - \mathbf{T}_k)$ is the characteristic polynomial evaluated at x_0 (Lemma 7.10.1), and two real polynomials that agree at every real number are equal (Theorem 6.1.5 (b)), so these identities hold as polynomials. They are the recurrence of Theorem 11.10.3 with the same starting values $p_0 = 1$, $p_1 = x - a_0$, so $D_k = p_k$ for every

k by induction. The roots of $p_{\mathbf{T}_k}$ are the eigenvalues of $\mathbf{T}_k$ (Theorem 9.2.1), which are real because $\mathbf{T}_k$ is real symmetric (Theorem 12.2.1) and distinct by Lemma 17.4.5, whose hypothesis holds because every $\sqrt{b_i} > 0$.

(b) Write $\mathbf{T}_{k+1}$ in bordered form:

$$\mathbf{T}_{k+1} = \begin{pmatrix} \mathbf{T}_k & \sqrt{b_k}\,\mathbf{e}_k \\ \sqrt{b_k}\,\mathbf{e}_k^{\top} & a_k \end{pmatrix}, \qquad \mathbf{e}_k \in \mathbb{R}^k.$$

The border of $\mathbf{T}_{k+1}$ is $\mathbf{w} = \sqrt{b_k}\,\mathbf{e}_k$, whose k-th entry is $\sqrt{b_k}$. We call it $\mathbf{w}$ rather than $\mathbf{b}$ because b_k already names a recurrence coefficient. We check the two clauses of Theorem 17.4.4 (c) for $\mathbf{B} = \mathbf{T}_k$, with $\mathbf{w}$ in place of $\mathbf{b}$. By Lemma 17.4.5, $\mathbf{T}_k$ has k distinct eigenvalues. For each eigenvalue ν, the eigenspace $E_\nu(\mathbf{T}_k)$ is spanned by one unit vector $\mathbf{u}$, so $P_\nu \mathbf{w} = \langle \mathbf{w}, \mathbf{u} \rangle \mathbf{u} = \sqrt{b_k}\, u_k\, \mathbf{u}$, which is nonzero because $b_k > 0$ and $u_k \neq 0$ by the same lemma. So Theorem 17.4.4 (c) applies, and the eigenvalues of $\mathbf{T}_{k+1}$ and of $\mathbf{T}_k$ interlace strictly. By (a) these are the roots of p_{k+1} and of p_k. This proves the corollary.

■

Part (a) gives a second proof, with no integrals, that the roots of p_k are real and simple — half of Theorem 11.10.5. The other half, that they lie inside the interval of the weight, is not visible in the matrix and still needs the argument of Chapter 11. For the monic Legendre polynomials of Example (The monic Legendre polynomials, continued), $a_k = 0$, $b_1 = \frac{1}{3}$, and $b_2 = \frac{4}{15}$, so

$$\mathbf{T}_3 = \begin{pmatrix} 0 & 1/\sqrt{3} & 0 \\ 1/\sqrt{3} & 0 & 2/\sqrt{15} \\ 0 & 2/\sqrt{15} & 0 \end{pmatrix},$$

and expanding along the first row,

$$p_{\mathbf{T}_3}(x) = x^3 - (\tfrac{1}{3} + \tfrac{4}{15})x = x^3 - \tfrac{3}{5}x = p_3.$$

Its leading corner $\mathbf{T}_2$ has characteristic polynomial $x^2 - \frac{1}{3} = p_2$ and unit eigenvectors $\frac{1}{\sqrt{2}}(1, \pm 1)$, both with nonzero last coordinate, so the border $\frac{2}{\sqrt{15}}\mathbf{e}_2$ meets both eigenspaces. The prediction is strict interlacing, and indeed

$$\sqrt{3/5} > 1/\sqrt{3} > 0 > -1/\sqrt{3} > -\sqrt{3/5},$$

that is $0.7746 > 0.5774 > 0 > -0.5774 > -0.7746$, where the strict comparisons $3/5 > 1/3$ and $1/\sqrt{3} > 0$ need no decimals.

17.4.4 A worked example

Example. One matrix and all three of its corners

Let

$$\mathbf{A} = \begin{pmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{pmatrix}.$$

Compute the eigenvalues of $\mathbf{A}$ and of each of its three 2×2 principal submatrices, exhibit

the interlacing in each case, and explain the pattern of strict and nonstrict inequalities using Theorem 17.4.4.

Solution. The eigenvalues of $\mathbf{A}$. Expanding $\det(x\mathbf{I}-\mathbf{A})$ along the first row,

$$\begin{aligned} p_{\mathbf{A}}(x) &= (x-2)\big[(x-2)^2-1\big]+1\cdot\big[-(x-2)\big] \\ &= (x-2)\big[(x-2)^2-2\big], \end{aligned}$$

so

$$\lambda(\mathbf{A}) = (2+\sqrt{2},\ 2,\ 2-\sqrt{2}) \approx (3.41421,\ 2,\ 0.58579).$$

The three corners. With I the index set kept,

$$\mathbf{A}_{\{1,2\},\{1,2\}} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}, \quad \mathbf{A}_{\{1,3\},\{1,3\}} = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}, \quad \mathbf{A}_{\{2,3\},\{2,3\}} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}.$$

A matrix $\left(\begin{smallmatrix} 2 & 1 \\ 1 & 2 \end{smallmatrix}\right)$ has trace 4 and determinant 3, so its eigenvalues are 3 and 1; and $2\mathbf{I}_2$ has the eigenvalue 2 twice.

The interlacing. In all three cases Theorem 17.4.2 is confirmed:

$$\begin{aligned} I=\{1,2\}: &\quad 3.41421 > 3 > 2 > 1 > 0.58579, \\ I=\{1,3\}: &\quad 3.41421 > 2 = 2 = 2 > 0.58579, \\ I=\{2,3\}: &\quad 3.41421 > 3 > 2 > 1 > 0.58579. \end{aligned}$$

The first and third are strict everywhere; the second has two equalities in the middle. The exact comparisons need no decimals: $2+\sqrt{2} > 3$ is $\sqrt{2} > 1$, and $1 > 2-\sqrt{2}$ is $\sqrt{2} > 1$ again.

Why. Read $\mathbf{A}$ as a bordered matrix in each case, using the permutation remark above.

For $I=\{1,2\}$ the deleted index is already the last one, so $\mathbf{B} = \left(\begin{smallmatrix} 2 & 1 \\ 1 & 2 \end{smallmatrix}\right)$, $\mathbf{b}=(0,1)$, and $c=2$. The eigenvalues 3 and 1 of $\mathbf{B}$ are distinct, with unit eigenvectors $\frac{1}{\sqrt{2}}(1,1)$ and $\frac{1}{\sqrt{2}}(1,-1)$. Both inner products with $\mathbf{b}$ are nonzero:

$$\langle \mathbf{b}, \tfrac{1}{\sqrt{2}}(1,1)\rangle = \tfrac{1}{\sqrt{2}}, \qquad \langle \mathbf{b}, \tfrac{1}{\sqrt{2}}(1,-1)\rangle = -\tfrac{1}{\sqrt{2}}.$$

So both clauses of Theorem 17.4.4 (c) hold and the interlacing is strict, as computed. The case $I=\{2,3\}$ is the same: moving index 1 to the end gives the corner $\left(\begin{smallmatrix} 2 & 1 \\ 1 & 2 \end{smallmatrix}\right)$ again, with border $(a_{21}, a_{31}) = (1,0)$, whose inner products with the two unit eigenvectors $\frac{1}{\sqrt{2}}(1,\pm 1)$ are both $\frac{1}{\sqrt{2}}$, nonzero.

For $I=\{1,3\}$ the deleted index is 2. Conjugating by the permutation matrix that sends the coordinate order $(1,2,3)$ to $(1,3,2)$ turns $\mathbf{A}$ into

$$\begin{pmatrix} 2 & 0 & 1 \\ 0 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix}, \qquad \mathbf{B} = 2\mathbf{I}_2, \quad \mathbf{b} = (1,1), \quad c = 2.$$

Here $\mathbf{b} \neq \mathbf{0}$ and $P_2\mathbf{b} = \mathbf{b} \neq \mathbf{0}$, so the first clause of Theorem 17.4.4 (c) is the one that fails: $\mathbf{B}$ has the repeated eigenvalue 2. Part (a) with $m_2 = 2$ then says that 2 is an eigenvalue of

$\mathbf{A}$ of multiplicity exactly 1, which is visible in $p_{\mathbf{A}}$, and the interlacing collapses at the middle to $2 = 2 = 2$. This is exactly the situation of the warning above, occurring inside a perfectly ordinary integer matrix.

17.4.5 Inertia under compression

Interlacing controls not only where the eigenvalues are but how many of them have each sign, and that is a statement about inertia.

Corollary 17.4.7. A Principal Submatrix Keeps Most of the Inertia

Let $\mathbf{A} \in M_n(F)$ be Hermitian, let $0 \le m \le n-1$, let $|I| = n - m$ and $\mathbf{B} = \mathbf{A}_{I,I}$. Write $n_+(\mathbf{M})$ and $n_-(\mathbf{M})$ for the number of positive and of negative eigenvalues of a Hermitian $\mathbf{M}$, counted with multiplicity. Then

$$n_+(\mathbf{B}) \le n_+(\mathbf{A}) \le n_+(\mathbf{B}) + m, \qquad n_-(\mathbf{B}) \le n_-(\mathbf{A}) \le n_-(\mathbf{B}) + m.$$

In particular $\mathbf{A} \succeq \mathbf{O}$ implies $\mathbf{B} \succeq \mathbf{O}$, and $\mathbf{A} \succ \mathbf{O}$ implies $\mathbf{B} \succ \mathbf{O}$.

Proof. Write $r = n_+(\mathbf{B})$ and $s = n_+(\mathbf{A})$, and abbreviate $\lambda_i = \lambda_i(\mathbf{A})$, $\mu_i = \lambda_i(\mathbf{B})$. By Corollary 17.4.3, $\lambda_i \ge \mu_i \ge \lambda_{i+m}$ for $1 \le i \le n - m$.

($n_+(\mathbf{B}) \le n_+(\mathbf{A})$) For $i \le r$ we have $\mu_i > 0$, hence $\lambda_i \ge \mu_i > 0$. So $\mathbf{A}$ has at least r positive eigenvalues.

($n_+(\mathbf{A}) \le n_+(\mathbf{B}) + m$) If $r + m \ge n$ there is nothing to prove, since $s \le n$. Otherwise $r + 1 \le n - m$, so the index $i = r + 1$ is admissible and $\mu_{r+1} \le 0$ by the definition of r; hence $\lambda_{r+1+m} \le \mu_{r+1} \le 0$, and the eigenvalues $\lambda_{r+m+1}, \dots, \lambda_n$ are all ≤ 0 by the decreasing indexing. So $s \le r + m$.

For the two statements about n_-, apply what has just been proved to $-\mathbf{A}$, whose principal submatrix on I is $-\mathbf{B}$. By Lemma 17.3.2 the eigenvalue lists are negated and reversed, so $n_+(-\mathbf{A}) = n_-(\mathbf{A})$ and $n_+(-\mathbf{B}) = n_-(\mathbf{B})$, and the two displayed chains are the same statement.

Finally, $\mathbf{A} \succeq \mathbf{O}$ means every eigenvalue of $\mathbf{A}$ is ≥ 0 (Theorem 13.1.4 (b)), that is, $n_-(\mathbf{A}) = 0$; then $n_-(\mathbf{B}) \le n_-(\mathbf{A}) = 0$, so $\mathbf{B} \succeq \mathbf{O}$. And $\mathbf{A} \succ \mathbf{O}$ means $n_+(\mathbf{A}) = n$ (Theorem 13.1.5 (b)); then $n_+(\mathbf{B}) \ge n_+(\mathbf{A}) - m = n - m$, which is the size of $\mathbf{B}$, so every eigenvalue of $\mathbf{B}$ is positive and $\mathbf{B} \succ \mathbf{O}$. This proves the corollary. ∎

The last two sentences are a second proof of something Chapter 13 already knew. That a principal submatrix of a positive semidefinite matrix is positive semidefinite is the step (a) ⇒ (e) in the proof of Theorem 13.1.4, where it comes from the observation that a principal submatrix is what the quadratic form does to vectors supported on the corresponding coordinates — one line, needing no interlacing. What is new here is the quantitative form. Sylvester's law of inertia (Theorem 14.5.1) says that the triple of counts is an invariant of a real symmetric matrix up to congruence, and Theorem 14.5.3 identifies Chapter 14's counts with the eigenvalue counts n_+ and n_- used here; the corollary says that deleting m rows and the matching m columns can destroy at most m positive and at most m negative eigenvalues, and create none of either.

Section 10 gives a second proof of Sylvester's law from Courant–Fischer, and bounds how far a low-rank perturbation can move the inertia.

17.4.6 Interlacing is a Hermitian phenomenon

Everything above rests on Theorem 17.2.3, which rests on the quadratic form being real. Without the Hermitian hypothesis, nothing survives — not a weakened interlacing, not even a bound.

Example. A corner outside the spectrum

Let

$$\mathbf{A} = \begin{pmatrix} 5 & 1 \\ -8 & -1 \end{pmatrix}.$$

Show that $\mathbf{A}$ has real eigenvalues, so that the conclusion of Theorem 17.4.2 can be stated for it, and that the conclusion is false.

Solution. $\operatorname{tr}\mathbf{A} = 4$ and $\det\mathbf{A} = 5(-1)-1(-8) = 3$, so $p_{\mathbf{A}}(x) = x^2-4x+3 = (x-3)(x-1)$ and the eigenvalues are 3 and 1, both real. The two 1×1 principal submatrices are $\mathbf{A}_{\{1\},\{1\}} = (5)$ and $\mathbf{A}_{\{2\},\{2\}} = (-1)$, with eigenvalues 5 and -1.

Interlacing would require $3 \geq 5 \geq 1$ in the first case and $3 \geq -1 \geq 1$ in the second. Both fail, and they fail on opposite sides: the diagonal entries 5 and -1 lie outside the interval $[1,3]$, which is the convex hull of the spectrum. For a Hermitian matrix that is impossible, by Example (The diagonal lies inside the spectrum).

Remark. The failure is total, not marginal. For any $a \in F$, the matrix

$$\mathbf{A}_a = \begin{pmatrix} a & 1 \\ -(a-1)(a-3) & 4-a \end{pmatrix}$$

has trace 4 and determinant $a(4-a)+(a-1)(a-3) = 4a-a^2+a^2-4a+3 = 3$, hence the same characteristic polynomial $(x-3)(x-1)$ as above. Its $(1,1)$ principal submatrix is (a). So a general matrix with spectrum $\{1,3\}$ can have any prescribed number whatsoever as its $(1,1)$ entry: the eigenvalue of a single principal submatrix is subject to no constraint at all. (The two diagonal entries together still add up to the trace, 4; that constraint holds for every matrix and has nothing to do with interlacing.) The choice $a = 5$ recovers $\mathbf{A}$, and $a = 0$ gives $\left(\begin{smallmatrix} 0 & 1 \\ -3 & 4 \end{smallmatrix}\right)$.

Warning. **Principal means the same index set on both sides.** Interlacing says nothing about a submatrix obtained by deleting row i and column j with $i \neq j$, even for a Hermitian $\mathbf{A}$. Take $\mathbf{A} = \left(\begin{smallmatrix} -10 & 1 \\ 1 & -10 \end{smallmatrix}\right)$, whose eigenvalues are -9 and -11. Deleting row 1 and column 2 leaves the 1×1 matrix (1), and 1 is nowhere near $[-11,-9]$. The submatrix in Theorem 17.4.2 is $\mathbf{A}_{I,I}$, and the two I's are not decoration.

17.4.7 Exercises

17.4.7.1 A. Check your understanding

Exercise. A1

(a) State Theorem 17.4.2, and say which hypothesis on $\mathbf{A}$ and which hypothesis on the submatrix are used.

(b) Let $\mathbf{A} \in M_6(\mathbb{C})$ be Hermitian and let $\mathbf{B} = \mathbf{A}_{I,I}$ with $|I| = 4$. Which eigenvalues of $\mathbf{A}$ bound $\lambda_3(\mathbf{B})$ above and below?

(c) Determine whether the following is correct, and justify your answer: if $\mathbf{b} \neq \mathbf{0}$ then the interlacing between $\left(\begin{smallmatrix} \mathbf{B} & \mathbf{b} \\ \mathbf{b}^* & c \end{smallmatrix}\right)$ and $\mathbf{B}$ is strict.

(d) Explain why a principal submatrix of a positive semidefinite matrix is positive semidefinite, in one sentence, without using interlacing.

(e) Give a 2×2 matrix with real eigenvalues whose $(1,1)$ entry is larger than both of them.

Solution.

(a) See Theorem 17.4.2. The matrix $\mathbf{A}$ must be **Hermitian**, so that Theorem 17.2.3 applies and the eigenvalues are real and can be put in order; and $\mathbf{B}$ must be a **principal** submatrix, $\mathbf{B} = \mathbf{A}_{I,I}$ with the same index set on rows and columns, so that Lemma 17.4.1 identifies $R_{\mathbf{B}}$ with $R_{\mathbf{A}}$ restricted to a coordinate hyperplane.

(b) Here $n = 6$ and $m = 2$, so Corollary 17.4.3 gives $\lambda_3(\mathbf{A}) \geq \lambda_3(\mathbf{B}) \geq \lambda_5(\mathbf{A})$.

(c) Incorrect. The warning after Theorem 17.4.4 gives $\mathbf{B} = 2\mathbf{I}_2$, $\mathbf{b} = (1,1)$, $c = 2$, where $\mathbf{b} \neq \mathbf{0}$ and the interlacing is not strict, because $\mathbf{B}$ has a repeated eigenvalue. The correct criterion is Theorem 17.4.4 (c), which needs distinct eigenvalues of $\mathbf{B}$ as well as $P_\nu \mathbf{b} \neq \mathbf{0}$ for each ν.

(d) Because $\langle \mathbf{B}\mathbf{y}, \mathbf{y} \rangle = \langle \mathbf{A}\mathbf{x}, \mathbf{x} \rangle \geq 0$, where $\mathbf{x} \in F^n$ is $\mathbf{y}$ placed in the coordinates of I and 0 elsewhere (Lemma 17.4.1); this is how Theorem 13.1.4 obtained its clause (e).

(e) $\left(\begin{smallmatrix} 5 & 1 \\ -8 & -1 \end{smallmatrix}\right)$, with eigenvalues 3 and 1 and $(1,1)$ entry 5 (Example (A corner outside the spectrum)). It is necessarily non-Hermitian, by Example (The diagonal lies inside the spectrum).

17.4.7.2 B. Practice

Exercise. B1: Interlacing in integers

Let

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 2 \\ 1 & 0 & 2 \\ 2 & 2 & 3 \end{pmatrix}.$$

Compute the eigenvalues of $\mathbf{A}$ and of all three of its 2×2 principal submatrices, and exhibit the interlacing in each case. Hence say, for each corner, whether the interlacing is strict.

Solution. The eigenvalues of $\mathbf{A}$. The vector $(1,-1,0)$ satisfies $\mathbf{A}(1,-1,0) = (-1,1,0)$, so

-1 is an eigenvalue. Dividing, $p_{\mathbf{A}}(x) = x^3 - 3x^2 - 9x - 5 = (x-5)(x+1)^2$, so

$$\lambda(\mathbf{A}) = (5,\ -1,\ -1).$$

(As a check, $\operatorname{tr}\mathbf{A} = 3 = 5 - 1 - 1$, and expanding along the first row $\det\mathbf{A} = 0 - 1(3-4) + 2(2-0) = 5 = 5\cdot(-1)\cdot(-1)$.)

The corners.

$$\mathbf{A}_{\{1,2\},\{1,2\}} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \qquad \mathbf{A}_{\{1,3\},\{1,3\}} = \mathbf{A}_{\{2,3\},\{2,3\}} = \begin{pmatrix} 0 & 2 \\ 2 & 3 \end{pmatrix}.$$

The first has trace 0 and determinant -1, so its eigenvalues are 1 and -1. The second has trace 3 and determinant -4, so its characteristic polynomial is $x^2 - 3x - 4 = (x-4)(x+1)$ and its eigenvalues are 4 and -1.

Interlacing.

$$\begin{aligned} I = \{1,2\}: &\quad 5 > 1 > -1 = -1 = -1, \\ I = \{1,3\}: &\quad 5 > 4 > -1 = -1 = -1, \\ I = \{2,3\}: &\quad 5 > 4 > -1 = -1 = -1. \end{aligned}$$

No corner gives strict interlacing, and the reason is on the side of $\mathbf{A}$: whichever index is deleted, $\lambda_2(\mathbf{A}) = \lambda_3(\mathbf{A}) = -1$, so the chain $\lambda_2(\mathbf{A}) \geq \lambda_2(\mathbf{B}) \geq \lambda_3(\mathbf{A})$ squeezes $\lambda_2(\mathbf{B})$ to -1. Strict interlacing would make the combined chain strictly decreasing, and so would force $\mathbf{A}$ to have three distinct eigenvalues. Theorem 17.4.4 sees the same thing from the side of the corner. For $I = \{1,2\}$ the border is $\mathbf{b} = (2,2)$, orthogonal to $(1,-1)$, which spans $E_{-1}(\mathbf{B})$; so $P_{-1}\mathbf{b} = \mathbf{0}$, and part (b) predicts that -1 is an eigenvalue of $\mathbf{A}$. For $I = \{1,3\}$ the border is $(a_{12}, a_{32}) = (1,2)$, orthogonal to $(2,-1)$, which spans E_{-1} of $\left(\begin{smallmatrix} 0 & 2 \\ 2 & 3 \end{smallmatrix}\right)$, with the same conclusion.

Exercise. B2: Bounds without a characteristic polynomial

Let $\mathbf{A} \in M_4(\mathbb{R})$ be symmetric with

$$\mathbf{A} = \begin{pmatrix} 7 & * & * & * \\ * & 2 & * & * \\ * & * & -3 & * \\ * & * & * & 1 \end{pmatrix},$$

the entries marked $*$ being unknown.

(a) What do Corollary 17.4.3 and Example (The diagonal lies inside the spectrum) say about $\lambda_1(\mathbf{A})$ and $\lambda_4(\mathbf{A})$?

(b) Suppose in addition that $\mathbf{A}_{\{1,2\},\{1,2\}} = \left(\begin{smallmatrix} 7 & 4 \\ 4 & 2 \end{smallmatrix}\right)$. Improve the lower bound on $\lambda_1(\mathbf{A})$, and give a lower bound on $\lambda_2(\mathbf{A})$.

Solution.

(a) Each diagonal entry is a 1×1 principal submatrix, so $\lambda_1(\mathbf{A}) \geq \max\{7,2,-3,1\} = 7$ and $\lambda_4(\mathbf{A}) \leq \min\{7,2,-3,1\} = -3$. In particular $\mathbf{A}$ is indefinite whatever the unknown entries are.

(b) The submatrix $\mathbf{B} = \left(\begin{smallmatrix} 7 & 4 \\ 4 & 2 \end{smallmatrix}\right)$ has trace 9 and determinant $14 - 16 = -2$, so its characteristic

polynomial is $x^2 - 9x - 2$ and its eigenvalues are

$$\lambda_1(\mathbf{B}) = \tfrac{9+\sqrt{89}}{2} \approx 9.2170, \qquad \lambda_2(\mathbf{B}) = \tfrac{9-\sqrt{89}}{2} \approx -0.2170.$$

Here $n = 4$ and $m = 2$, and the left-hand inequality $\lambda_i(\mathbf{A}) \geq \lambda_i(\mathbf{B})$ of Corollary 17.4.3 at $i = 1, 2$ gives

$$\lambda_1(\mathbf{A}) \;\geq\; \tfrac{9+\sqrt{89}}{2} \approx 9.2170, \qquad \lambda_2(\mathbf{A}) \;\geq\; \tfrac{9-\sqrt{89}}{2} \approx -0.2170.$$

The first improves the bound 7 from (a). The second is new: the diagonal entries alone say nothing about $\lambda_2(\mathbf{A})$. The right-hand inequalities $\lambda_i(\mathbf{B}) \geq \lambda_{i+2}(\mathbf{A})$ give $\lambda_3(\mathbf{A}) \leq 9.2170$, which is true but of little use, and $\lambda_4(\mathbf{A}) \leq -0.2170$, which is weaker than the bound $\lambda_4(\mathbf{A}) \leq -3$ of (a). Each principal submatrix contributes what it can, and one keeps the best of the bounds.

Exercise. B3: Deciding strictness

For each of the following bordered matrices $\mathbf{A} = \left(\begin{smallmatrix} \mathbf{B} & \mathbf{b} \\ \mathbf{b}^{\top} & c \end{smallmatrix}\right)$, decide without computing $p_{\mathbf{A}}$ whether the interlacing between $\mathbf{A}$ and $\mathbf{B}$ is strict, and justify with Theorem 17.4.4.

(a) $\mathbf{B} = \operatorname{diag}(3, 1)$, $\mathbf{b} = (0, 1)$, $c = 0$.

(b) $\mathbf{B} = \operatorname{diag}(3, 1)$, $\mathbf{b} = (2, -1)$, $c = 5$.

(c) $\mathbf{B} = \left(\begin{smallmatrix} 4 & 0 \\ 0 & 4 \end{smallmatrix}\right)$, $\mathbf{b} = (1, 2)$, $c = 0$.

Solution. In each case $\mathbf{B}$ is diagonal, so the standard basis vectors are an orthonormal eigenbasis and $P_\nu \mathbf{b}$ is just the part of $\mathbf{b}$ in the coordinates where the diagonal entry equals ν.

(a) Not strict. The eigenvalues $3, 1$ of $\mathbf{B}$ are distinct, but $P_3\mathbf{b} = (0, 0) = \mathbf{0}$. By Theorem 17.4.4 (b), 3 is an eigenvalue of $\mathbf{A}$, so one of the inequalities $\lambda_1(\mathbf{A}) \geq 3 \geq \lambda_2(\mathbf{A})$ is an equality. (Indeed $p_{\mathbf{A}}(x) = (x - 3)(x^2 - x - 1)$.)

(b) Strict. The eigenvalues $3, 1$ are distinct and $P_3\mathbf{b} = (2, 0) \neq \mathbf{0}$, $P_1\mathbf{b} = (0, -1) \neq \mathbf{0}$. Both clauses of Theorem 17.4.4 (c) hold. (Indeed $p_{\mathbf{A}}(x) = (x - 2)(x^2 - 7x + 4)$, with roots $\tfrac{7+\sqrt{33}}{2} \approx 6.372$, 2, and $\tfrac{7-\sqrt{33}}{2} \approx 0.628$, and $6.372 > 3 > 2 > 1 > 0.628$.)

(c) Not strict. $P_4\mathbf{b} = \mathbf{b} = (1, 2) \neq \mathbf{0}$, so the second clause holds, but $\mathbf{B}$ has the repeated eigenvalue 4 and the first clause fails. By Theorem 17.4.4 (a) with $m_4 = 2$, the number 4 is an eigenvalue of $\mathbf{A}$ of multiplicity exactly 1, and the middle of the chain reads $4 \geq \lambda_2(\mathbf{A}) \geq 4$. (Indeed $p_{\mathbf{A}}(x) = (x - 5)(x - 4)(x + 1)$.)

17.4.7.3 C. Going deeper

Exercise. C1: Rank under compression

Let $\mathbf{A} \in M_n(F)$ be Hermitian of rank r, and let $\mathbf{B} = \mathbf{A}_{I,I}$ with $|I| = n - m$.

(a) Prove that $\operatorname{rank} \mathbf{B} \geq r - 2m$.

(b) Show that the bound is attained, by exhibiting for each n and each $m \le n/2$ a Hermitian $\mathbf{A}$ of rank $2m$ with a principal submatrix of size $n-m$ and rank 0.

Solution.

(a) For a Hermitian matrix $\mathbf{M}$, diagonalizing by the spectral theorem gives $\operatorname{rank}\mathbf{M} = n_+(\mathbf{M}) + n_-(\mathbf{M})$, the total number of nonzero eigenvalues. By Corollary 17.4.7, $n_+(\mathbf{B}) \ge n_+(\mathbf{A}) - m$ and $n_-(\mathbf{B}) \ge n_-(\mathbf{A}) - m$. Adding,

$$\operatorname{rank}\mathbf{B} = n_+(\mathbf{B}) + n_-(\mathbf{B}) \;\ge\; (n_+(\mathbf{A}) + n_-(\mathbf{A})) - 2m = r - 2m.$$

(b) Let $m \le n/2$ and let $I = \{m+1, m+2, \dots, n\}$, of size $n-m$. Define $\mathbf{A}$ by $a_{i,m+i} = a_{m+i,i} = 1$ for $1 \le i \le m$, and all other entries 0. Then $\mathbf{A}$ is real symmetric. It is a direct sum of m copies of $\left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$, acting on the coordinate pairs $\{i, m+i\}$, together with zeros on the remaining coordinates, so its eigenvalues are ± 1 each m times and 0 otherwise; hence $\operatorname{rank}\mathbf{A} = 2m$ and $n_+(\mathbf{A}) = n_-(\mathbf{A}) = m$. Every nonzero entry of $\mathbf{A}$ has one index in $\{1, \dots, m\}$, which I omits, so $\mathbf{A}_{I,I} = \mathbf{O}$ and $\operatorname{rank}\mathbf{A}_{I,I} = 0 = \operatorname{rank}\mathbf{A} - 2m$. The bound in (a) is therefore sharp for every m.

Exercise. C2: The converse, in the smallest case

Theorem 17.4.2 says that the spectrum of a corner interlaces the spectrum of the whole. This exercise asks whether every interlacing pattern occurs.

(a) Let $\lambda_1 > \mu_1 > \lambda_2$ be real numbers. Find $b \in \mathbb{R}$ and $c \in \mathbb{R}$ such that $\mathbf{A} = \left(\begin{smallmatrix} \mu_1 & b \\ b & c \end{smallmatrix}\right)$ has eigenvalues λ_1 and λ_2.

(b) Deduce that for $n = 2$ the interlacing condition of Theorem 17.4.2 is not merely necessary but also sufficient.

(c) Explain why the strict hypothesis $\lambda_1 > \mu_1 > \lambda_2$ may be relaxed to $\lambda_1 \ge \mu_1 \ge \lambda_2$.

Hint: match the trace and the determinant.

Solution.

(a) A real symmetric 2×2 matrix with eigenvalues λ_1, λ_2 must have $\operatorname{tr}\mathbf{A} = \lambda_1 + \lambda_2$ and $\det\mathbf{A} = \lambda_1\lambda_2$. The first forces

$$c = \lambda_1 + \lambda_2 - \mu_1,$$

and the second then forces $\mu_1 c - b^2 = \lambda_1\lambda_2$, that is,

$$b^2 = \mu_1(\lambda_1 + \lambda_2 - \mu_1) - \lambda_1\lambda_2 = -(\mu_1 - \lambda_1)(\mu_1 - \lambda_2) = (\lambda_1 - \mu_1)(\mu_1 - \lambda_2),$$

where the middle equality is the expansion $-(\mu_1^2 - (\lambda_1+\lambda_2)\mu_1 + \lambda_1\lambda_2)$. The hypothesis $\lambda_1 > \mu_1 > \lambda_2$ makes both factors positive, so $b = \sqrt{(\lambda_1 - \mu_1)(\mu_1 - \lambda_2)}$ is a real number. With this b and c, the matrix $\mathbf{A}$ is real symmetric with the prescribed trace and determinant, hence with characteristic polynomial $x^2 - (\lambda_1+\lambda_2)x + \lambda_1\lambda_2 = (x - \lambda_1)(x - \lambda_2)$, hence with eigenvalues λ_1 and λ_2.
For example $\lambda_1 = 4$, $\mu_1 = 2$, $\lambda_2 = 1$ gives $c = 3$ and $b^2 = 2$, that is $\mathbf{A} = \left(\begin{smallmatrix} 2 & \sqrt{2} \\ \sqrt{2} & 3 \end{smallmatrix}\right)$, whose trace is 5 and determinant $6 - 2 = 4$, as required.

(b) Necessity is Theorem 17.4.2 with $n = 2$ and $I = \{1\}$. Sufficiency is part (a): given any $\lambda_1 > \mu_1 > \lambda_2$, a Hermitian $\mathbf{A}$ with spectrum $\{\lambda_1, \lambda_2\}$ and $\mathbf{A}_{\{1\},\{1\}} = (\mu_1)$ exists.

(c) If $\mu_1 = \lambda_1$ or $\mu_1 = \lambda_2$ then the formula gives $b = 0$, and $\mathbf{A} = \operatorname{diag}(\mu_1, \lambda_1 + \lambda_2 - \mu_1)$ is diagonal with the two entries λ_1 and λ_2 in one order or the other. So the construction still works, with a zero border; consistently with Theorem 17.4.4, a zero border is exactly what makes the interlacing nonstrict.

Exercise. C3: How far apart can the extremes be

Let $\mathbf{A} \in M_n(F)$ be Hermitian.

(a) Prove that $\lambda_1(\mathbf{A}) - \lambda_n(\mathbf{A}) \geq \max_{i,j} |a_{ii} - a_{jj}|$.

(b) Prove that $\lambda_1(\mathbf{A}) - \lambda_n(\mathbf{A}) \geq 2 \max_{i \neq j} |a_{ij}|$.

Hint for (b): look at a 2×2 principal submatrix.

Solution.

(a) By Example (The diagonal lies inside the spectrum), $\lambda_n(\mathbf{A}) \leq a_{kk} \leq \lambda_1(\mathbf{A})$ for every k. So for any i, j, both $a_{ii} - a_{jj}$ and $a_{jj} - a_{ii}$ are at most $\lambda_1(\mathbf{A}) - \lambda_n(\mathbf{A})$, which gives the bound after taking the maximum over i, j.

(b) Fix $i \neq j$ and let $I = \{i, j\}$, so that

$$\mathbf{B} = \mathbf{A}_{I,I} = \begin{pmatrix} a_{ii} & a_{ij} \\ \overline{a_{ij}} & a_{jj} \end{pmatrix}$$

is Hermitian of size 2. Its trace is $a_{ii} + a_{jj}$ and its determinant is $a_{ii}a_{jj} - |a_{ij}|^2$, so its eigenvalues are

$$\frac{a_{ii} + a_{jj}}{2} \pm \sqrt{\left(\frac{a_{ii} - a_{jj}}{2}\right)^2 + |a_{ij}|^2},$$

and their difference is $2\sqrt{((a_{ii} - a_{jj})/2)^2 + |a_{ij}|^2} \geq 2|a_{ij}|$. Now apply Corollary 17.4.3 with $m = n - 2$: it gives $\lambda_1(\mathbf{A}) \geq \lambda_1(\mathbf{B})$ and $\lambda_2(\mathbf{B}) \geq \lambda_{2+(n-2)}(\mathbf{A}) = \lambda_n(\mathbf{A})$, so

$$\lambda_1(\mathbf{A}) - \lambda_n(\mathbf{A}) \;\geq\; \lambda_1(\mathbf{B}) - \lambda_2(\mathbf{B}) \;\geq\; 2|a_{ij}|.$$

Taking the maximum over the pairs $i \neq j$ finishes the proof. (For $n = 2$ the step with $m = 0$ is the identity $\mathbf{B} = \mathbf{A}$, and the conclusion is the computation itself.)

17.5 Adding a Rank-One Matrix

Section 4 deleted a row and the matching column and found that the spectrum of the smaller matrix is pinned between the eigenvalues of the larger one. This section makes the opposite kind of change: it leaves the size alone and adds the simplest nonzero Hermitian matrix there is, a positive multiple of $\mathbf{v}\mathbf{v}^*$. The answer has the same shape — the new eigenvalues interlace the old ones — but here we can say more than that they interlace. Under one extra hypothesis the new eigenvalues are the roots of a single explicit scalar equation in one variable, and the graph of that function shows at a glance where every one of them sits.

Throughout this section $F = \mathbb{R}$ or $F = \mathbb{C}$, the matrix $\mathbf{A} \in M_n(F)$ is Hermitian, and its eigenvalues are indexed decreasingly, $\lambda_1(\mathbf{A}) \geq \cdots \geq \lambda_n(\mathbf{A})$, as everywhere in this chapter. We fix $\mathbf{v} \in F^n$ and a real number τ, and study

$$\widetilde{\mathbf{A}} \coloneqq \mathbf{A} + \tau \mathbf{v}\mathbf{v}^*.$$

The added term is Hermitian, since $(\mathbf{v}\mathbf{v}^*)^* = \mathbf{v}\mathbf{v}^*$ and τ is real, so $\widetilde{\mathbf{A}}$ is Hermitian too and its eigenvalues are indexed the same way. The added term has rank 1 when $\tau \neq 0$ and $\mathbf{v} \neq \mathbf{0}$, and is the zero matrix otherwise. For $\tau > 0$ it is positive semidefinite, because

$$\langle \tau \mathbf{v}\mathbf{v}^*\mathbf{x}, \mathbf{x} \rangle = \tau\, \mathbf{x}^*\mathbf{v}\, \mathbf{v}^*\mathbf{x} = \tau |\mathbf{v}^*\mathbf{x}|^2 \;\geq\; 0$$

for every $\mathbf{x}$, which is clause (P2) of Definition 13.1.1.

17.5.1 One bump, every eigenvalue up

Adding a positive semidefinite matrix can only raise the quadratic form, so no eigenvalue can go down. The content of the theorem is the ceiling: however large τ is, the i-th new eigenvalue cannot climb past the $(i-1)$-st old one. All the room created by the bump is spent at the top.

Theorem 17.5.1. Interlacing for a Rank-One Update

Let $\mathbf{A} \in M_n(F)$ be Hermitian, let $\mathbf{v} \in F^n$, let $\tau \geq 0$, and put $\widetilde{\mathbf{A}} = \mathbf{A} + \tau \mathbf{v}\mathbf{v}^*$. Then

$$\lambda_i(\widetilde{\mathbf{A}}) \;\geq\; \lambda_i(\mathbf{A}) \;\geq\; \lambda_{i+1}(\widetilde{\mathbf{A}}) \qquad (1 \leq i \leq n-1),$$

and $\lambda_n(\widetilde{\mathbf{A}}) \geq \lambda_n(\mathbf{A})$. Written as one chain,

$$\begin{aligned} \lambda_1(\widetilde{\mathbf{A}}) \;\geq\; \lambda_1(\mathbf{A}) \;\geq\; \lambda_2(\widetilde{\mathbf{A}}) \;\geq\; \lambda_2(\mathbf{A}) \\ \geq\; \cdots \;\geq\; \lambda_n(\widetilde{\mathbf{A}}) \;\geq\; \lambda_n(\mathbf{A}). \end{aligned}$$

Idea. The two halves have different characters and different sources. The left-hand inequalities say the eigenvalues go **up**, and that is a statement about the quadratic form: on every vector, $\widetilde{\mathbf{A}}$ scores at least what $\mathbf{A}$ scores, so every max–min it competes in can only improve. The right-hand inequalities say the eigenvalues go up **by at most one index**, and that has nothing to do with the sign of τ: it is the rank of the perturbation doing the work, and Corollary 17.3.8 has already isolated exactly that.

Proof. If $\tau = 0$ or $\mathbf{v} = \mathbf{0}$ then $\widetilde{\mathbf{A}} = \mathbf{A}$ and every inequality is an equality, so assume $\tau > 0$ and $\mathbf{v} \neq \mathbf{0}$.

For the inequalities $\lambda_i(\widetilde{\mathbf{A}}) \geq \lambda_i(\mathbf{A})$, fix i and let $\mathbf{x} \neq \mathbf{0}$. By the computation above,

$$R_{\widetilde{\mathbf{A}}}(\mathbf{x}) = R_{\mathbf{A}}(\mathbf{x}) + \frac{\tau |\mathbf{v}^*\mathbf{x}|^2}{\|\mathbf{x}\|^2} \;\geq\; R_{\mathbf{A}}(\mathbf{x}),$$

using Definition 17.1.1 and $\tau > 0$. Multiplying through by $\|\mathbf{x}\|^2$, this says $\langle \widetilde{\mathbf{A}}\mathbf{x}, \mathbf{x} \rangle \geq \langle \mathbf{A}\mathbf{x}, \mathbf{x} \rangle$ for every $\mathbf{x}$, that is, $\widetilde{\mathbf{A}} \succeq \mathbf{A}$ in the Loewner order of Definition 13.5.1. So Corollary 17.2.4 gives $\lambda_i(\widetilde{\mathbf{A}}) \geq \lambda_i(\mathbf{A})$ for every i.

For the inequalities $\lambda_i(\mathbf{A}) \geq \lambda_{i+1}(\widetilde{\mathbf{A}})$, note that $\operatorname{rank}(\tau \mathbf{v}\mathbf{v}^*) = 1$, because every column of

$\mathbf{v}\mathbf{v}^*$ is a multiple of $\mathbf{v} \neq \mathbf{0}$ and at least one of them is nonzero. So Corollary 17.3.8 applies with $\mathbf{B} = \tau\mathbf{v}\mathbf{v}^*$ and $r = 1$, and its first inequality reads

$$\lambda_{i+1}(\mathbf{A} + \tau\mathbf{v}\mathbf{v}^*) \;\leq\; \lambda_i(\mathbf{A})$$

for $1 \leq i \leq n-1$. This proves the theorem.

■

Two remarks on the proof. The first half used nothing about the rank: it is Corollary 17.2.4, the general fact that $\widetilde{\mathbf{A}} \succeq \mathbf{A}$ forces $\lambda_i(\widetilde{\mathbf{A}}) \geq \lambda_i(\mathbf{A})$ for every i, specialized to a rank-one increment. The second half used nothing about the sign: the eigenvalues of $\mathbf{A}+\mathbf{B}$ and of $\mathbf{A}$ are never more than $\operatorname{rank}\mathbf{B}$ indices apart. Interlacing is what happens when both restrictions apply at once.

The secular equation of the next subsection gives a sharper, **strict** version of the theorem — it locates each new eigenvalue exactly, strictly between consecutive old ones — but only under extra hypotheses (a diagonal $\mathbf{A}$ with distinct entries, and a $\mathbf{v}$ with no zero coordinate), so it is not a second proof of the general statement. The other short route to the general theorem is Cauchy interlacing (Theorem 17.4.2) applied to a suitable bordered matrix.

Example. A bump on a diagonal matrix

Let $\mathbf{A} = \operatorname{diag}(4, 1)$, $\mathbf{v} = (1, 2)$, and $\tau = 1$. Compute $\widetilde{\mathbf{A}}$ and its eigenvalues, and check Theorem 17.5.1.

Solution. Here $\mathbf{v}\mathbf{v}^{\top} = \left(\begin{smallmatrix} 1 & 2 \\ 2 & 4 \end{smallmatrix}\right)$, so

$$\widetilde{\mathbf{A}} = \begin{pmatrix} 4 & 0 \\ 0 & 1 \end{pmatrix} + \begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix} = \begin{pmatrix} 5 & 2 \\ 2 & 5 \end{pmatrix}.$$

Its trace is 10 and its determinant is $25 - 4 = 21$, so its characteristic polynomial is $x^2 - 10x + 21 = (x-7)(x-3)$ and $\lambda_1(\widetilde{\mathbf{A}}) = 7$, $\lambda_2(\widetilde{\mathbf{A}}) = 3$. The chain of the theorem reads

$$7 \;\geq\; 4 \;\geq\; 3 \;\geq\; 1,$$

and all four inequalities are strict. Note the accounting: the two eigenvalues rose by 3 and by 2, a total of $5 = \|\mathbf{v}\|^2$. That is no coincidence, and the last result of this section says why.

Quick check. The matrix $\mathbf{A} = \operatorname{diag}(9, 5, 2)$ is bumped by $\tau\mathbf{v}\mathbf{v}^*$ with $\tau > 0$. Without knowing $\mathbf{v}$ or τ, what is the largest possible value of $\lambda_3(\mathbf{A} + \tau\mathbf{v}\mathbf{v}^*)$, and what is the smallest?

Solution. By Theorem 17.5.1, $\lambda_2(\mathbf{A}) \geq \lambda_3(\widetilde{\mathbf{A}}) \geq \lambda_3(\mathbf{A})$, that is, $5 \geq \lambda_3(\widetilde{\mathbf{A}}) \geq 2$. Both ends are reached: $\mathbf{v} = \mathbf{e}_3$ with τ large gives $\widetilde{\mathbf{A}} = \operatorname{diag}(9, 5, 2+\tau)$, whose third eigenvalue is 5 as soon as $\tau \geq 3$; and $\mathbf{v} = \mathbf{e}_1$ gives $\widetilde{\mathbf{A}} = \operatorname{diag}(9+\tau, 5, 2)$, whose third eigenvalue is 2. So the smallest eigenvalue can be pushed up as far as the *second* smallest of $\mathbf{A}$ and no further, no matter how big τ is.

17.5.2 Where the new eigenvalues actually are

Interlacing brackets each new eigenvalue in an interval. To pin it down we need an equation, and there is a cheap one: an eigenvalue of $\widetilde{\mathbf{A}}$ is a number λ making $\widetilde{\mathbf{A}} - \lambda\mathbf{I}$ singular, and $\widetilde{\mathbf{A}} - \lambda\mathbf{I} = (\mathbf{A} - \lambda\mathbf{I}) + \tau\mathbf{v}\mathbf{v}^*$ is a rank-one update of a matrix whose determinant we know. Chapter 8 computed the determinant of exactly such an update.

Divide the characteristic polynomial of the bumped matrix by that of the original one; what is left is a sum of n simple fractions.

Definition 17.5.1. Secular function

Let $\mathbf{A} = \operatorname{diag}(d_1, \dots, d_n)$ with $d_1, \dots, d_n \in \mathbb{R}$, let $\mathbf{v} \in F^n$, and let $\tau \in \mathbb{R}$. The **secular function** of the update $\mathbf{A} + \tau\mathbf{v}\mathbf{v}^*$ is

$$f(\lambda) \coloneqq 1 + \tau \sum_{i=1}^{n} \frac{|v_i|^2}{d_i - \lambda},$$

defined for every real λ **other than** $d_1, \dots, d_n$. The equation $f(\lambda) = 0$ is the **secular equation**.

In words: f is the constant 1 plus one simple pole at each diagonal entry of $\mathbf{A}$, weighted by τ times the squared modulus of the matching coordinate of $\mathbf{v}$. The name is inherited from celestial mechanics, where equations of this shape governed slow ("secular") drifts in planetary orbits.

Theorem 17.5.2. Eigenvalues from the Secular Equation

Let $\mathbf{A} = \operatorname{diag}(d_1, \dots, d_n)$ with

$$d_1 > d_2 > \cdots > d_n$$

strictly decreasing and real, let $\tau > 0$, and let $\mathbf{v} \in F^n$ have **every** coordinate nonzero. Write $\widetilde{\mathbf{A}} = \mathbf{A} + \tau\mathbf{v}\mathbf{v}^*$ and let f be its secular function. Then:

(a) no d_i is an eigenvalue of $\widetilde{\mathbf{A}}$;

(b) a real number $\lambda \notin \{d_1, \dots, d_n\}$ is an eigenvalue of $\widetilde{\mathbf{A}}$ if and only if $f(\lambda) = 0$;

(c) f has exactly n roots $\mu_1 > \cdots > \mu_n$, one in each of the intervals

$$(d_1, \infty),\ (d_2, d_1),\ \dots,\ (d_n, d_{n-1}),$$

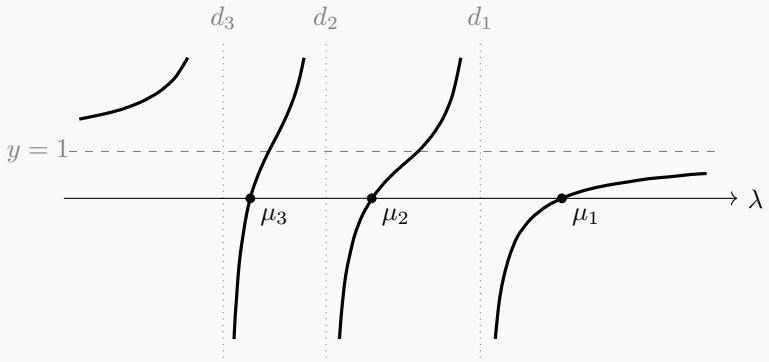
and none below d_n; these are the eigenvalues of $\widetilde{\mathbf{A}}$, each simple, and

$$\mu_1 > d_1 > \mu_2 > d_2 > \cdots > \mu_n > d_n;$$

(d) for each j, the vector $(\mathbf{A} - \mu_j\mathbf{I})^{-1}\mathbf{v}$ is an eigenvector of $\widetilde{\mathbf{A}}$ for μ_j.

Idea. Part (b) is one line of Chapter 8 once (a) has cleared the denominators out of the way. Part (c) is a picture. Each term $|v_i|^2/(d_i - \lambda)$ increases with λ on any interval that avoids d_i, so on each of the $n + 1$ open intervals into which the poles cut the line f climbs steadily from left to right. Just to the left of each pole f shoots up to $+\infty$; just to the right it starts again from $-\infty$; and far out in either direction it settles at 1. Draw that, and the roots count themselves. The proof turns the picture into a count that needs no limits at all: the n

eigenvalues turn out to be distinct, there are exactly n intervals where a root can live, and each holds at most one.



The dotted verticals are the poles, at the old eigenvalues $d_3 < d_2 < d_1$; the dashed horizontal is $y = 1$, the level f approaches at both ends. On the branch left of d_3 every term of the sum is positive, so the curve stays above 1 and never meets the axis. Each of the other three branches runs from $-\infty$ to $+\infty$ or from $-\infty$ to 1, increasing throughout, so each crosses the axis exactly once. Three branches, three roots, one strictly inside each gap and one above the top pole.

Proof. **(a)** Suppose $\widetilde{\mathbf{A}}\mathbf{x} = d_i\mathbf{x}$ for some $\mathbf{x} \neq \mathbf{0}$. Rearranging, $(\mathbf{A} - d_i\mathbf{I})\mathbf{x} = -\tau(\mathbf{v}^*\mathbf{x})\mathbf{v}$. The matrix $\mathbf{A} - d_i\mathbf{I}$ is diagonal with i-th diagonal entry 0, so the i-th coordinate of the left-hand side is 0; the i-th coordinate of the right-hand side is $-\tau(\mathbf{v}^*\mathbf{x})v_i$. Since $\tau \neq 0$ and $v_i \neq 0$, this forces $\mathbf{v}^*\mathbf{x} = 0$, and then $(\mathbf{A} - d_i\mathbf{I})\mathbf{x} = \mathbf{0}$. The diagonal entries d_j are distinct, so the only diagonal entry of $\mathbf{A} - d_i\mathbf{I}$ that vanishes is the i-th, and therefore $\mathbf{x} = c\mathbf{e}_i$ for some scalar c. But then $0 = \mathbf{v}^*\mathbf{x} = c\,\overline{v_i}$, and $v_i \neq 0$ gives $c = 0$, so $\mathbf{x} = \mathbf{0}$ — a contradiction.

(b) Let $\lambda \notin \{d_1, \dots, d_n\}$. Then $\mathbf{A} - \lambda\mathbf{I}$ is diagonal with nonzero diagonal entries, hence invertible, with

$$(\mathbf{A} - \lambda\mathbf{I})^{-1} = \operatorname{diag}\Big(\frac{1}{d_1 - \lambda}, \dots, \frac{1}{d_n - \lambda}\Big).$$

Apply the determinant statement of Corollary 8.3.6 — itself a 1×1 case of Theorem 8.3.2 — to the invertible matrix $\mathbf{A} - \lambda\mathbf{I}$, with the column $\tau\mathbf{v}$ and the row $\mathbf{v}^*$ — that is, with $\tau\mathbf{v}$ as its $\mathbf{u}$ and $\overline{\mathbf{v}}$ in place of its $\mathbf{v}$, so that its $\mathbf{v}^\top$ is our $\mathbf{v}^*$:

$$\begin{aligned}\det(\widetilde{\mathbf{A}} - \lambda\mathbf{I}) &= \det\big((\mathbf{A} - \lambda\mathbf{I}) + (\tau\mathbf{v})\mathbf{v}^*\big)\\ &= \det(\mathbf{A} - \lambda\mathbf{I})\,\big(1 + \tau\,\mathbf{v}^*(\mathbf{A} - \lambda\mathbf{I})^{-1}\mathbf{v}\big).\end{aligned}$$

The scalar in the second factor is $\sum_i |v_i|^2/(d_i - \lambda)$, since $\mathbf{v}^*\mathbf{D}\mathbf{v} = \sum_i \overline{v_i}\, d_{ii} v_i$ for diagonal $\mathbf{D}$. So

$$\det(\widetilde{\mathbf{A}} - \lambda\mathbf{I}) = \Big(\prod_{i=1}^{n}(d_i - \lambda)\Big) f(\lambda), \tag{$*$}$$

and the product is nonzero. Hence $\det(\widetilde{\mathbf{A}} - \lambda\mathbf{I}) = 0$ if and only if $f(\lambda) = 0$. Finally $\det(\lambda\mathbf{I} - \widetilde{\mathbf{A}}) = (-1)^n \det(\widetilde{\mathbf{A}} - \lambda\mathbf{I})$, since negating all n rows of a matrix multiplies its determinant by $(-1)^n$, by n applications of Theorem 7.5.1 (b); so the two determinants vanish together, and by Theorem 7.10.4 (a) this happens exactly when λ is an eigenvalue of $\widetilde{\mathbf{A}}$.

(c) Let J be one of the $n + 1$ open intervals into which $d_1, \dots, d_n$ cut the real line, and let $\lambda < \lambda'$ lie in J. For each i, the numbers $d_i - \lambda$ and $d_i - \lambda'$ are nonzero and have the same

sign, because $d_i \notin J$; hence

$$\frac{1}{d_i - \lambda'} - \frac{1}{d_i - \lambda} = \frac{\lambda' - \lambda}{(d_i - \lambda)(d_i - \lambda')} \;>\; 0.$$

Multiplying by $\tau |v_i|^2 > 0$ and summing over i gives $f(\lambda') > f(\lambda)$. So f is **strictly increasing** on each such interval, and in particular has at most one root in each.

Next, every eigenspace of $\widetilde{\mathbf{A}}$ is one-dimensional. Let $\widetilde{\mathbf{A}}\mathbf{x} = \mu\mathbf{x}$. By (a), $\mu \notin \{d_1, \dots, d_n\}$, so $\mathbf{A} - \mu\mathbf{I}$ is invertible, and rearranging as in (a) gives $(\mathbf{A} - \mu\mathbf{I})\mathbf{x} = -\tau(\mathbf{v}^*\mathbf{x})\mathbf{v}$, that is,

$$\mathbf{x} = -\tau(\mathbf{v}^*\mathbf{x})\,(\mathbf{A} - \mu\mathbf{I})^{-1}\mathbf{v}.$$

So $\ker(\widetilde{\mathbf{A}} - \mu\mathbf{I}) \subseteq \operatorname{span}\big((\mathbf{A} - \mu\mathbf{I})^{-1}\mathbf{v}\big)$, a space of dimension 1. Now write $\widetilde{\mathbf{A}} = \mathbf{U}\mathbf{D}\mathbf{U}^*$ with $\mathbf{U}$ unitary and $\mathbf{D}$ diagonal carrying the eigenvalue list of $\widetilde{\mathbf{A}}$ with multiplicity (Corollary 12.4.3, or Corollary 12.5.2 over $\mathbb{R}$). Since $\mathbf{U}$ is invertible, $\ker(\widetilde{\mathbf{A}} - \mu\mathbf{I}) = \mathbf{U}\ker(\mathbf{D} - \mu\mathbf{I})$, and $\dim\ker(\mathbf{D} - \mu\mathbf{I})$ is the number of diagonal entries of $\mathbf{D}$ equal to μ. That number is therefore at most 1: the n eigenvalues of $\widetilde{\mathbf{A}}$ are **distinct**, and each is simple.

Now count. Call the eigenvalues $\mu_1 > \cdots > \mu_n$. By (b) each is a root of f, and by (a) none equals a d_i, so each lies in one of the $n + 1$ open intervals. None lies in $(-\infty, d_n)$: there every $d_i - \lambda$ is positive, so every term of the sum is positive and $f(\lambda) > 1$. That leaves n intervals, each containing at most one root of f, to accommodate n distinct roots — so each of the intervals $(d_1, \infty), (d_2, d_1), \dots, (d_n, d_{n-1})$ contains exactly one of them, and since they are listed in decreasing order, $\mu_1 \in (d_1, \infty)$ and $\mu_{i+1} \in (d_{i+1}, d_i)$. Conversely every root of f is an eigenvalue by (b), so f has no roots besides these n. Reading the intervals in decreasing order gives the displayed chain.

(d) Fix j and put $\mathbf{x} = (\mathbf{A} - \mu_j\mathbf{I})^{-1}\mathbf{v}$, which is defined because $\mu_j \notin \{d_1, \dots, d_n\}$ and nonzero because $\mathbf{v} \neq \mathbf{0}$. As in (b),

$$\tau\,\mathbf{v}^*(\mathbf{A} - \mu_j\mathbf{I})^{-1}\mathbf{v} = \tau \sum_i \frac{|v_i|^2}{d_i - \mu_j} = f(\mu_j) - 1 = -1.$$

Therefore

$$\begin{aligned}(\widetilde{\mathbf{A}} - \mu_j\mathbf{I})\mathbf{x} &= (\mathbf{A} - \mu_j\mathbf{I})\mathbf{x} + \tau\mathbf{v}(\mathbf{v}^*\mathbf{x}) \\ &= \mathbf{v} + \mathbf{v}\,(\tau\,\mathbf{v}^*(\mathbf{A} - \mu_j\mathbf{I})^{-1}\mathbf{v}) = \mathbf{v} - \mathbf{v} = \mathbf{0}.\end{aligned}$$

This proves the theorem.

■

Identity $(*)$ is worth keeping on its own: it says that the characteristic polynomial of the bumped matrix is the characteristic polynomial of the original one multiplied by the secular function. The poles of f cancel the roots of the product exactly, and what survives is a polynomial of the same degree with its roots shifted one gap to the right.

Example. A three-by-three update, found from its secular equation

Let

$$\mathbf{A} = \operatorname{diag}(7, 1, -3), \qquad \mathbf{v} = (2, 1, 2), \qquad \tau = 1.$$

Write down $\widetilde{\mathbf{A}} = \mathbf{A} + \mathbf{v}\mathbf{v}^{\top}$ and its secular function, find all three eigenvalues from the secular equation, and confirm them independently.

Solution. The matrix. $\mathbf{v}\mathbf{v}^{\top}$ has (i,j) entry $v_i v_j$, so

$$\widetilde{\mathbf{A}} = \begin{pmatrix} 7 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -3 \end{pmatrix} + \begin{pmatrix} 4 & 2 & 4 \\ 2 & 1 & 2 \\ 4 & 2 & 4 \end{pmatrix} = \begin{pmatrix} 11 & 2 & 4 \\ 2 & 2 & 2 \\ 4 & 2 & 1 \end{pmatrix}.$$

The hypotheses of Theorem 17.5.2 hold: $7 > 1 > -3$ are distinct, $\tau = 1 > 0$, and no coordinate of $\mathbf{v}$ vanishes.

The secular function. With $|v_i|^2 = 4, 1, 4$,

$$f(\lambda) = 1 + \frac{4}{7-\lambda} + \frac{1}{1-\lambda} + \frac{4}{-3-\lambda}.$$

The roots. Multiply through by $(7-\lambda)(1-\lambda)(-3-\lambda)$, which by $(*)$ turns f into $\det(\widetilde{\mathbf{A}} - \lambda\mathbf{I})$:

$$\begin{aligned} &(7-\lambda)(1-\lambda)(-3-\lambda) + 4(1-\lambda)(-3-\lambda) \\ &\quad + (7-\lambda)(-3-\lambda) + 4(7-\lambda)(1-\lambda) \\ &= -(\lambda - 13)(\lambda - 2)(\lambda + 1). \end{aligned}$$

So the eigenvalues are $\mu_1 = 13$, $\mu_2 = 2$, $\mu_3 = -1$. They sit where the theorem promised:

$$13 > 7 > 2 > 1 > -1 > -3.$$

It is quicker to check the three roots than to expand the product. At $\lambda = 13$,

$$f(13) = 1 - \tfrac{4}{6} - \tfrac{1}{12} - \tfrac{4}{16} = \tfrac{12-8-1-3}{12} = 0;$$

at $\lambda = 2$, $f(2) = 1 + \tfrac{4}{5} - 1 - \tfrac{4}{5} = 0$; at $\lambda = -1$, $f(-1) = 1 + \tfrac{1}{2} + \tfrac{1}{2} - 2 = 0$.

An independent confirmation. Part (d) of the theorem produces eigenvectors without any further elimination: $(\mathbf{A} - \mu\mathbf{I})^{-1}\mathbf{v}$ is diagonal applied to $\mathbf{v}$. For $\mu = 13$, $\mathbf{A} - 13\mathbf{I} = \operatorname{diag}(-6, -12, -16)$ and

$$(\mathbf{A} - 13\mathbf{I})^{-1}\mathbf{v} = (-\tfrac{1}{3}, -\tfrac{1}{12}, -\tfrac{1}{8}) \sim (8, 2, 3)$$

after scaling by -24. Multiplying out, $\widetilde{\mathbf{A}}(8,2,3) = (88+4+12,\ 16+4+6,\ 32+4+3) = (104, 26, 39) = 13\,(8,2,3)$. The same recipe gives $(2,-5,-2)$ for $\mu = 2$ and $(1,2,-4)$ for $\mu = -1$, and both check by one matrix–vector product. Finally the trace: $11+2+1 = 14 = 13+2-1$, as Corollary 12.1.4 requires.

Warning. **Both extra hypotheses of Theorem 17.5.2 earn their place, and the sign of τ reverses the picture.**

A zero coordinate deflates. Take $\mathbf{A} = \operatorname{diag}(2,1,0)$ and $\mathbf{v} = (1,0,1)$, so $v_2 = 0$. Then $\mathbf{e}_2$ is an eigenvector of $\widetilde{\mathbf{A}} = \mathbf{A} + \mathbf{v}\mathbf{v}^{\top}$ with eigenvalue $1 = d_2$, because $\mathbf{v}^{*}\mathbf{e}_2 = 0$. So conclusion (a) fails. The secular function $1 + 1/(2-\lambda) - 1/\lambda$ has only two poles and only two roots, $2 \pm \sqrt{2}$; the third eigenvalue, 1, is invisible to it. The eigenvalues still interlace,

$2+\sqrt{2} \geq 2 \geq 1 \geq 1 \geq 2-\sqrt{2} \geq 0$, but not strictly.

A repeated d_i deflates too. If $d_1 = d_2$, any vector in $\operatorname{span}(\mathbf{e}_1, \mathbf{e}_2)$ orthogonal to $\mathbf{v}$ is an eigenvector of $\widetilde{\mathbf{A}}$ for d_1, for the same reason.

Negative τ reverses the interlacing. For $\tau < 0$ the matrix $\tau\mathbf{v}\mathbf{v}^*$ is negative semidefinite and every eigenvalue goes **down**: applying Theorem 17.5.1 to $\widetilde{\mathbf{A}}$ and $-\tau > 0$ gives $\lambda_i(\mathbf{A}) \geq \lambda_i(\widetilde{\mathbf{A}}) \geq \lambda_{i+1}(\mathbf{A})$. Reading the chain of the theorem in the wrong direction is the standard way to get this backwards.

17.5.3 The shifts add up

Interlacing says where each eigenvalue may go. One number says how far they all go together, and it costs a line.

Corollary 17.5.3. Total Eigenvalue Shift

Let $\mathbf{A} \in M_n(F)$ be Hermitian, $\mathbf{v} \in F^n$, $\tau \in \mathbb{R}$, and $\widetilde{\mathbf{A}} = \mathbf{A} + \tau\mathbf{v}\mathbf{v}^*$. Then

$$\sum_{i=1}^{n} (\lambda_i(\widetilde{\mathbf{A}}) - \lambda_i(\mathbf{A})) = \tau\|\mathbf{v}\|^2.$$

Proof. By Theorem 0.9.6 (3) applied to the $n \times 1$ matrix $\mathbf{v}$ and the $1 \times n$ matrix $\mathbf{v}^*$,

$$\operatorname{tr}(\mathbf{v}\mathbf{v}^*) = \operatorname{tr}(\mathbf{v}^*\mathbf{v}) = \|\mathbf{v}\|^2.$$

Hence $\operatorname{tr}\widetilde{\mathbf{A}} = \operatorname{tr}\mathbf{A} + \tau\|\mathbf{v}\|^2$, by part (1) of the same theorem. Both $\mathbf{A}$ and $\widetilde{\mathbf{A}}$ are Hermitian, so Corollary 12.1.4 expresses each trace as the sum of the eigenvalues with multiplicity, and subtracting the two expressions gives the claim. ∎

So the whole budget for the movement is $\tau\|\mathbf{v}\|^2$, and for $\tau \geq 0$ Theorem 17.5.1 says how it may be spent: each eigenvalue takes a nonnegative share, and no share except the top one is large enough to carry an eigenvalue past its old upper neighbor. In Example (A bump on a diagonal matrix) the budget 5 was split as $3 + 2$; in Example (A three-by-three update, found from its secular equation) the budget $\|\mathbf{v}\|^2 = 9$ was split as $6 + 1 + 2$.

Remark. The secular equation of Theorem 17.5.2 is not only a proof device. Because f is increasing between consecutive poles and the bracketing interval is known in advance, each root can be located by bisection on an interval that is guaranteed to contain exactly one of them — a far better-behaved problem than finding the roots of a characteristic polynomial. The secular equation is the engine of the divide-and-conquer eigenvalue algorithms for Hermitian matrices, which split the matrix into two halves joined by a rank-one term and solve one secular equation to merge their spectra. The deflation observed in the warning above, where a zero coordinate of $\mathbf{v}$ hands over an eigenvalue for free, is the same phenomenon that makes the algorithm fast in practice.

17.5.4 Exercises

17.5.4.1 A. Check your understanding

Exercise. A1

(a) State Theorem 17.5.1, including the hypothesis on τ.

(b) Write down the secular function of $\mathbf{A} + 2\mathbf{v}\mathbf{v}^{\top}$ for $\mathbf{A} = \operatorname{diag}(6, 2, -1)$ and $\mathbf{v} = (1, 3, 1)$.

(c) Explain in one sentence why no d_i is an eigenvalue of $\mathbf{A} + \tau\mathbf{v}\mathbf{v}^{*}$ under the hypotheses of Theorem 17.5.2.

(d) Determine whether the following is correct, and justify your answer: if $\tau < 0$ then $\lambda_i(\mathbf{A} + \tau\mathbf{v}\mathbf{v}^{*}) \geq \lambda_i(\mathbf{A})$ for every i.

(e) A Hermitian 4×4 matrix $\mathbf{A}$ has eigenvalues $8, 5, 5, 0$. What does Corollary 17.5.3 say about $\mathbf{A} + 3\mathbf{v}\mathbf{v}^{*}$ when $\|\mathbf{v}\| = 2$?

Solution.

(a) For Hermitian $\mathbf{A} \in M_n(F)$, $\mathbf{v} \in F^n$, and $\tau \geq 0$, the eigenvalues of $\widetilde{\mathbf{A}} = \mathbf{A} + \tau\mathbf{v}\mathbf{v}^{*}$ satisfy $\lambda_i(\widetilde{\mathbf{A}}) \geq \lambda_i(\mathbf{A}) \geq \lambda_{i+1}(\widetilde{\mathbf{A}})$ for $1 \leq i \leq n-1$, together with $\lambda_n(\widetilde{\mathbf{A}}) \geq \lambda_n(\mathbf{A})$. The hypothesis $\tau \geq 0$ is what makes the perturbation positive semidefinite; without it only the second family of inequalities survives.

(b) $f(\lambda) = 1 + 2\Big(\dfrac{1}{6-\lambda} + \dfrac{9}{2-\lambda} + \dfrac{1}{-1-\lambda}\Big)$.

(c) If $\widetilde{\mathbf{A}}\mathbf{x} = d_i\mathbf{x}$, comparing i-th coordinates in $(\mathbf{A} - d_i\mathbf{I})\mathbf{x} = -\tau(\mathbf{v}^{*}\mathbf{x})\mathbf{v}$ forces $\mathbf{v}^{*}\mathbf{x} = 0$ because $v_i \neq 0$; then $\mathbf{x}$ is a multiple of $\mathbf{e}_i$, and $\mathbf{v}^{*}\mathbf{x} = 0$ forces that multiple to be 0.

(d) Incorrect; the inequality goes the other way. For $\tau < 0$ the added matrix is negative semidefinite, and $\lambda_i(\mathbf{A}) \geq \lambda_i(\mathbf{A} + \tau\mathbf{v}\mathbf{v}^{*})$ for every i. Concretely, $\mathbf{A} = \operatorname{diag}(1, 0)$, $\tau = -1$, $\mathbf{v} = \mathbf{e}_1$ gives $\widetilde{\mathbf{A}} = \operatorname{diag}(0, 0)$ and λ_1 drops from 1 to 0.

(e) It says the four eigenvalues of $\mathbf{A} + 3\mathbf{v}\mathbf{v}^{*}$ sum to $8 + 5 + 5 + 0 + 3 \cdot 4 = 30$, that is, 12 more than those of $\mathbf{A}$. Combined with Theorem 17.5.1, which caps λ_2 at 8, λ_3 at 5, and λ_4 at 5, it forces $\lambda_1(\mathbf{A} + 3\mathbf{v}\mathbf{v}^{*}) \geq 30 - 18 = 12$.

17.5.4.2 B. Practice

Exercise. B1: A secular equation, end to end

Let $\mathbf{A} = \operatorname{diag}(7, 0, -3)$, $\mathbf{v} = (3, 2, 2)$, and $\tau = 1$.

(a) Write down $\widetilde{\mathbf{A}} = \mathbf{A} + \mathbf{v}\mathbf{v}^{\top}$ and the secular function f.

(b) Verify that 21, 2, and -2 are roots of f, and explain why these are all the eigenvalues of $\widetilde{\mathbf{A}}$.

(c) Exhibit the interlacing chain of Theorem 17.5.1 for this pair.

Solution.

(a) $\mathbf{v}\mathbf{v}^{\top}$ has entries $v_i v_j$, so

$$\widetilde{\mathbf{A}} = \begin{pmatrix} 7 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -3 \end{pmatrix} + \begin{pmatrix} 9 & 6 & 6 \\ 6 & 4 & 4 \\ 6 & 4 & 4 \end{pmatrix} = \begin{pmatrix} 16 & 6 & 6 \\ 6 & 4 & 4 \\ 6 & 4 & 1 \end{pmatrix},$$

and $f(\lambda) = 1 + \dfrac{9}{7-\lambda} + \dfrac{4}{0-\lambda} + \dfrac{4}{-3-\lambda}$.

(b) At $\lambda = 21$: $1 - \frac{9}{14} - \frac{4}{21} - \frac{4}{24} = 1 - \frac{9}{14} - \frac{4}{21} - \frac{1}{6}$. Over 42 this is $(42 - 27 - 8 - 7)/42 = 0$. At $\lambda = 2$: $1 + \frac{9}{5} - 2 - \frac{4}{5} = (5 + 9 - 10 - 4)/5 = 0$. At $\lambda = -2$: $1 + 1 + 2 - 4 = 0$. The hypotheses of Theorem 17.5.2 hold — $7 > 0 > -3$ are distinct, $\tau = 1 > 0$, and every coordinate of $\mathbf{v}$ is nonzero — so by part (c) the secular function has exactly three roots and they are the eigenvalues. We have found three distinct roots, so there are no others.

(c) $21 \geq 7 \geq 2 \geq 0 \geq -2 \geq -3$, all strict, as part (c) of Theorem 17.5.2 predicts.

Exercise. B2: Reading the chain backwards

Let $\mathbf{B} = \begin{pmatrix} 5 & 2 \\ 2 & 5 \end{pmatrix}$ and $\mathbf{v} = (1, 2)$.

(a) Compute the eigenvalues of $\mathbf{B}$ and of $\mathbf{B} - \mathbf{v}\mathbf{v}^{\top}$.

(b) State which chain of inequalities Theorem 17.5.1 gives here, being careful about which matrix plays the role of $\mathbf{A}$.

Solution.

(a) $\mathbf{B}$ has trace 10 and determinant 21, so its characteristic polynomial is $x^2 - 10x + 21 = (x - 7)(x - 3)$ and its eigenvalues are 7 and 3. Also

$$\mathbf{B} - \mathbf{v}\mathbf{v}^{\top} = \begin{pmatrix} 5 & 2 \\ 2 & 5 \end{pmatrix} - \begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix} = \begin{pmatrix} 4 & 0 \\ 0 & 1 \end{pmatrix},$$

with eigenvalues 4 and 1.

(b) Put $\mathbf{A} = \mathbf{B} - \mathbf{v}\mathbf{v}^{\top} = \operatorname{diag}(4, 1)$ and $\tau = 1$, so that $\mathbf{B} = \mathbf{A} + \tau\mathbf{v}\mathbf{v}^{\top}$ is the *bumped* matrix. The theorem then reads $\lambda_1(\mathbf{B}) \geq \lambda_1(\mathbf{A}) \geq \lambda_2(\mathbf{B}) \geq \lambda_2(\mathbf{A})$, that is, $7 \geq 4 \geq 3 \geq 1$. Taking $\mathbf{B}$ for $\mathbf{A}$ and $\tau = -1$ instead would give the reversed chain $\lambda_i(\mathbf{B}) \geq \lambda_i(\mathbf{B} - \mathbf{v}\mathbf{v}^{\top}) \geq \lambda_{i+1}(\mathbf{B})$, namely $7 \geq 4 \geq 3 \geq 1$ again: the same four numbers, read from the other side.

Exercise. B3: The missing eigenvalue

Let $\mathbf{A} = \operatorname{diag}(5, 3, 0)$, $\mathbf{v} = (2, 2, 2)$, and $\tau = 1$. It is given that two of the eigenvalues of $\widetilde{\mathbf{A}} = \mathbf{A} + \mathbf{v}\mathbf{v}^{\top}$ are 15 and 4. Find the third without computing a characteristic polynomial, and then confirm it from the secular equation.

Solution. By Corollary 17.5.3 the eigenvalues of $\widetilde{\mathbf{A}}$ sum to

$$(5+3+0)+\|\mathbf{v}\|^2 = 8+12 = 20.$$

Hence the third eigenvalue is $20-15-4=1$.

For the confirmation, $|v_i|^2 = 4$ for each i, so

$$f(\lambda) = 1 + \frac{4}{5-\lambda} + \frac{4}{3-\lambda} + \frac{4}{-\lambda},$$

and $f(1) = 1+1+2-4 = 0$. The hypotheses of Theorem 17.5.2 hold, and 1 lies in the interval $(0,3)$ where part (c) puts the third root. As a check on the other two, $15 > 5$ and $4 \in (3,5)$, exactly as the theorem requires.

17.5.4.3 C. Going deeper

Exercise. C1: What a zero coordinate does

Let $\mathbf{A} = \operatorname{diag}(d_1,\dots,d_n)$ with the d_i real, let $\tau \in \mathbb{R}$, and let $\mathbf{v} \in F^n$.

(a) Prove that if $v_i = 0$ then $\mathbf{e}_i$ is an eigenvector of $\mathbf{A}+\tau\mathbf{v}\mathbf{v}^*$ with eigenvalue d_i.

(b) Suppose $\tau > 0$, the d_i are strictly decreasing, and exactly one coordinate v_k vanishes. Show that the remaining $n-1$ eigenvalues of $\mathbf{A}+\tau\mathbf{v}\mathbf{v}^*$ are the roots of the function obtained from the secular function by deleting the k-th term. *Hint: restrict to the coordinate subspace spanned by the $\mathbf{e}_i$ with $i \neq k$.*

(c) Illustrate both parts with $\mathbf{A} = \operatorname{diag}(2,1,0)$, $\mathbf{v} = (1,0,1)$, and $\tau = 1$.

Solution.

(a) $\mathbf{v}^*\mathbf{e}_i = \overline{v_i} = 0$, so $(\mathbf{A}+\tau\mathbf{v}\mathbf{v}^*)\mathbf{e}_i = \mathbf{A}\mathbf{e}_i + \tau\mathbf{v}(\mathbf{v}^*\mathbf{e}_i) = d_i\mathbf{e}_i$.

(b) Let $U = \operatorname{span}(\mathbf{e}_i : i \neq k)$, a subspace of dimension $n-1$. Every column of $\mathbf{A}$ other than the k-th lies in U, and $\mathbf{v} \in U$ because $v_k = 0$; hence $(\mathbf{A}+\tau\mathbf{v}\mathbf{v}^*)U \subseteq U$. In the ordered basis $(\mathbf{e}_i)_{i\neq k}$ of U, the restriction of $\mathbf{A}+\tau\mathbf{v}\mathbf{v}^*$ to U has matrix $\mathbf{A}' + \tau\mathbf{v}'(\mathbf{v}')^*$, where $\mathbf{A}' = \operatorname{diag}(d_i : i \neq k)$ and $\mathbf{v}'$ is $\mathbf{v}$ with its k-th coordinate removed. Since $U \oplus \operatorname{span}(\mathbf{e}_k) = F^n$ and $\operatorname{span}(\mathbf{e}_k)$ is also invariant by (a), the eigenvalues of $\mathbf{A}+\tau\mathbf{v}\mathbf{v}^*$ are those of the restriction to U together with d_k. Now $\mathbf{A}'$ has strictly decreasing diagonal entries, $\tau > 0$, and $\mathbf{v}'$ has no zero coordinate, so Theorem 17.5.2 applies to $\mathbf{A}' + \tau\mathbf{v}'(\mathbf{v}')^*$: its $n-1$ eigenvalues are exactly the roots of its secular function. That function is the secular function of the original update with the k-th term deleted, because that term was $\tau|v_k|^2/(d_k-\lambda) = 0$ anyway.

(c) Here $v_2 = 0$, so $\mathbf{e}_2$ is an eigenvector with eigenvalue $d_2 = 1$ by (a). By (b) the other two eigenvalues are the roots of

$$1 + \frac{1}{2-\lambda} + \frac{1}{-\lambda} = 0,$$

that is, of $\lambda(2-\lambda) + \lambda - (2-\lambda) = 0$, i.e. $\lambda^2 - 4\lambda + 2 = 0$, with roots $2 \pm \sqrt{2}$. So the

spectrum is $\{2+\sqrt{2},\ 1,\ 2-\sqrt{2}\}$. Directly,

$$\mathbf{A}+\mathbf{v}\mathbf{v}^{\top}=\begin{pmatrix}3&0&1\\0&1&0\\1&0&1\end{pmatrix},$$

whose second row and column confirm the eigenvalue 1, and whose remaining 2×2 block $\left(\begin{smallmatrix}3&1\\1&1\end{smallmatrix}\right)$ has trace 4 and determinant 2.

Exercise. C2: Only one eigenvalue can escape

Let $\mathbf{A}\in M_n(F)$ be Hermitian, $\mathbf{v}\in F^n$, $\tau\in\mathbb{R}$, and $\widetilde{\mathbf{A}}=\mathbf{A}+\tau\mathbf{v}\mathbf{v}^{*}$.

(a) Prove that **at most one** eigenvalue of $\widetilde{\mathbf{A}}$ lies outside the closed interval $[\lambda_n(\mathbf{A}),\lambda_1(\mathbf{A})]$, and say which one it is and on which side, according to the sign of τ.

(b) Prove that the escaping eigenvalue moves by at most $|\tau|\|\mathbf{v}\|^2$, and show by examples with $n=2$ that it may move by exactly that amount, and that it may not escape at all.

Solution.

(a) Suppose first $\tau\geq 0$. By Theorem 17.5.1, $\lambda_n(\widetilde{\mathbf{A}})\geq\lambda_n(\mathbf{A})$, so no eigenvalue of $\widetilde{\mathbf{A}}$ falls below the interval; and for $i\geq 2$, $\lambda_i(\widetilde{\mathbf{A}})\leq\lambda_{i-1}(\mathbf{A})\leq\lambda_1(\mathbf{A})$, so only $\lambda_1(\widetilde{\mathbf{A}})$ can rise above it. So at most one eigenvalue escapes, namely the largest, and it escapes upwards.
For $\tau<0$, the warning above gives $\lambda_i(\mathbf{A})\geq\lambda_i(\widetilde{\mathbf{A}})\geq\lambda_{i+1}(\mathbf{A})$ for $i\leq n-1$ and $\lambda_n(\mathbf{A})\geq\lambda_n(\widetilde{\mathbf{A}})$. Hence $\lambda_1(\widetilde{\mathbf{A}})\leq\lambda_1(\mathbf{A})$, and $\lambda_i(\widetilde{\mathbf{A}})\geq\lambda_{i+1}(\mathbf{A})\geq\lambda_n(\mathbf{A})$ for $i\leq n-1$. So only $\lambda_n(\widetilde{\mathbf{A}})$ can escape, and it escapes downwards.

(b) Take $\tau>0$; the case $\tau<0$ is the mirror image. By Theorem 17.5.1 every shift $\lambda_i(\widetilde{\mathbf{A}})-\lambda_i(\mathbf{A})$ is ≥ 0, and by Corollary 17.5.3 they sum to $\tau\|\mathbf{v}\|^2$. So each shift, in particular that of λ_1, is at most $\tau\|\mathbf{v}\|^2$.

For equality, let $\mathbf{A}=\operatorname{diag}(2,1)$, $\mathbf{v}=\mathbf{e}_1$, $\tau=3$: then $\widetilde{\mathbf{A}}=\operatorname{diag}(5,1)$, and λ_1 moves from 2 to 5, by exactly $\tau\|\mathbf{v}\|^2=3$. For no escape, let $\mathbf{v}=\mathbf{e}_2$ and $\tau=1$: then $\widetilde{\mathbf{A}}=\operatorname{diag}(2,2)$, whose eigenvalues both lie in $[1,2]$. The whole budget went to λ_2, which Theorem 17.5.1 caps at $\lambda_1(\mathbf{A})=2$.

Exercise. C3: Bumping along an eigenvector

Let $\mathbf{A}\in M_n(F)$ be Hermitian with an orthonormal basis of eigenvectors $\mathbf{u}_1,\dots,\mathbf{u}_n$, where $\mathbf{A}\mathbf{u}_i=\lambda_i(\mathbf{A})\mathbf{u}_i$.

(a) Fix k and $\tau\in\mathbb{R}$. Prove that the eigenvalues of $\mathbf{A}+\tau\mathbf{u}_k\mathbf{u}_k^{*}$, listed with multiplicity but not necessarily in order, are $\lambda_k(\mathbf{A})+\tau$ together with $\lambda_i(\mathbf{A})$ for $i\neq k$.

(b) Deduce that $\mathbf{A}-\lambda_1(\mathbf{A})\mathbf{u}_1\mathbf{u}_1^{*}$ has eigenvalue list $0,\lambda_2(\mathbf{A}),\dots,\lambda_n(\mathbf{A})$. (This is the **deflation** step: it removes the top eigenvalue and leaves the rest untouched, which is

what makes it possible to find eigenvalues one at a time.)

(c) Explain why (a) does not contradict part (c) of Theorem 17.5.2.

Solution.

(a) For $i \neq k$, orthonormality gives $\mathbf{u}_k^* \mathbf{u}_i = 0$, so

$$(\mathbf{A} + \tau \mathbf{u}_k \mathbf{u}_k^*)\mathbf{u}_i = \mathbf{A}\mathbf{u}_i + \tau \mathbf{u}_k(\mathbf{u}_k^* \mathbf{u}_i) = \lambda_i(\mathbf{A})\mathbf{u}_i.$$

For $i = k$, $\mathbf{u}_k^* \mathbf{u}_k = 1$ gives $(\mathbf{A} + \tau \mathbf{u}_k \mathbf{u}_k^*)\mathbf{u}_k = (\lambda_k(\mathbf{A}) + \tau)\mathbf{u}_k$. So $(\mathbf{u}_1, \dots, \mathbf{u}_n)$ is a basis of F^n consisting of eigenvectors of $\mathbf{A} + \tau \mathbf{u}_k \mathbf{u}_k^*$, with the stated eigenvalues. A basis of eigenvectors exhibits the whole eigenvalue list with multiplicity.

(b) Take $k = 1$ and $\tau = -\lambda_1(\mathbf{A})$ in (a).

(c) There is no contradiction, because the hypothesis of Theorem 17.5.2 that fails here fails as badly as possible. Writing everything in the orthonormal eigenbasis, the update vector is $\mathbf{u}_k$, whose coordinate vector is $\mathbf{e}_k$: **all but one** of its coordinates are zero. Part (a) is the extreme case of Exercise (C1: What a zero coordinate does) (a), with $n - 1$ coordinates deflating at once, and only one eigenvalue is left to move.

17.6 Subspaces and Sums of Eigenvalues

Section 4 deleted a row and the matching column, which is a coordinate-dependent operation, and got a coordinate-free conclusion: the spectra interlace. That is a hint. This section replaces "delete some coordinates" by "restrict to a subspace", which is the same operation with the accident of a basis removed, and the interlacing survives in a stronger form. Then it asks a question Courant–Fischer cannot answer one eigenvalue at a time: not "what is λ_2?" but "what is $\lambda_1 + \lambda_2$?" That sum turns out to be a maximum over a set of subspaces, and the answer discharges a promise Chapter 13 made about singular values.

Throughout this section $F = \mathbb{R}$ or $F = \mathbb{C}$, every matrix whose eigenvalues are indexed is Hermitian, and the indexing is decreasing, $\lambda_1(\mathbf{A}) \geq \cdots \geq \lambda_n(\mathbf{A})$.

17.6.1 Restricting a matrix to a subspace

A principal submatrix of $\mathbf{A}$ records what the quadratic form $\mathbf{x} \mapsto \langle \mathbf{A}\mathbf{x}, \mathbf{x} \rangle$ does to vectors supported on a few coordinates. Nothing in that description needs the coordinates to be the standard ones: any subspace W and any orthonormal basis of it will do, and the matrix that results plays the same role.

Definition 17.6.1. Compression to a subspace

Let $\mathbf{A} \in M_n(F)$, let $1 \leq k \leq n$, and let $\mathbf{q}_1, \dots, \mathbf{q}_k$ be an **orthonormal** list in F^n, so that the matrix $\mathbf{Q} \in M_{n \times k}(F)$ with these columns satisfies $\mathbf{Q}^* \mathbf{Q} = \mathbf{I}_k$. The **compression** of $\mathbf{A}$ to $W = \operatorname{span}(\mathbf{q}_1, \dots, \mathbf{q}_k)$ with respect to this basis is the $k \times k$ matrix

$$\mathbf{Q}^* \mathbf{A} \mathbf{Q}, \qquad (\mathbf{Q}^* \mathbf{A} \mathbf{Q})_{ij} = \langle \mathbf{A}\mathbf{q}_j, \mathbf{q}_i \rangle.$$

The entry formula is just $(\mathbf{Q}^* \mathbf{A} \mathbf{Q})_{ij} = \mathbf{q}_i^* \mathbf{A} \mathbf{q}_j$, and $\mathbf{q}_i^* \mathbf{y} = \langle \mathbf{y}, \mathbf{q}_i \rangle$ for the standard inner product.

If $\mathbf{A}$ is Hermitian so is $\mathbf{Q}^*\mathbf{A}\mathbf{Q}$, since $(\mathbf{Q}^*\mathbf{A}\mathbf{Q})^* = \mathbf{Q}^*\mathbf{A}^*\mathbf{Q} = \mathbf{Q}^*\mathbf{A}\mathbf{Q}$.

The eigenvalues do not depend on the basis chosen. Let $(\mathbf{q}'_1, \dots, \mathbf{q}'_k)$ be another orthonormal basis of the same W, with matrix $\mathbf{Q}'$. Each $\mathbf{q}'_j$ lies in $W = \mathrm{col}(\mathbf{Q})$, so $\mathbf{Q}' = \mathbf{Q}\mathbf{U}$ for some $\mathbf{U} \in M_k(F)$, and then $\mathbf{I}_k = (\mathbf{Q}')^*\mathbf{Q}' = \mathbf{U}^*\mathbf{Q}^*\mathbf{Q}\mathbf{U} = \mathbf{U}^*\mathbf{U}$, so $\mathbf{U}$ is unitary. Hence

$$(\mathbf{Q}')^*\mathbf{A}\mathbf{Q}' = \mathbf{U}^*(\mathbf{Q}^*\mathbf{A}\mathbf{Q})\mathbf{U},$$

a unitary similarity, which leaves the eigenvalues alone. So it is legitimate to speak of *the* eigenvalues of the compression of $\mathbf{A}$ to W, and we do.

Three cases to keep in mind.

- **Coordinate subspaces.** If $\mathbf{q}_j = \mathbf{e}_{i_j}$ for indices $i_1 < \cdots < i_k$, then $(\mathbf{Q}^*\mathbf{A}\mathbf{Q})_{lj} = a_{i_l i_j}$: the compression is the principal submatrix $\mathbf{A}_{I,I}$ of Definition 7.7.1 on the index set $I = \{i_1, \dots, i_k\}$. Compression generalizes "delete rows and the matching columns".
- $k = n$**.** Then $\mathbf{Q}$ is a unitary matrix and $\mathbf{Q}^*\mathbf{A}\mathbf{Q}$ is unitarily similar to $\mathbf{A}$: nothing is lost and nothing is learned.
- $k = 1$**.** Then $\mathbf{Q}^*\mathbf{A}\mathbf{Q}$ is the 1×1 matrix $\langle \mathbf{A}\mathbf{q}_1, \mathbf{q}_1 \rangle = R_{\mathbf{A}}(\mathbf{q}_1)$, the Rayleigh quotient of Definition 17.1.1 at a unit vector. Everything below is a statement about k Rayleigh quotients at once.

Drop the orthonormality and the notion collapses. With $\mathbf{A} = \mathrm{diag}(1, 0)$ and the single column $\mathbf{Q} = (2, 0)^\top$, which has $\mathbf{Q}^*\mathbf{Q} = (4) \neq \mathbf{I}_1$, the "compression" is the 1×1 matrix (4), whose eigenvalue 4 is larger than every eigenvalue of $\mathbf{A}$. The clause that fails is $\mathbf{Q}^*\mathbf{Q} = \mathbf{I}_k$, and with it every conclusion of this section.

17.6.2 Interlacing without coordinates

Theorem 17.6.1. Poincaré Separation Theorem

Let $\mathbf{A} \in M_n(F)$ be Hermitian, let $1 \le k \le n$, and let $\mathbf{B} = \mathbf{Q}^*\mathbf{A}\mathbf{Q}$ be the compression of $\mathbf{A}$ to a k-dimensional subspace $W \subseteq F^n$, as in Definition 17.6.1. Then

$$\lambda_i(\mathbf{A}) \;\ge\; \lambda_i(\mathbf{B}) \;\ge\; \lambda_{i+n-k}(\mathbf{A}) \qquad (1 \le i \le k).$$

Idea. Both inequalities come from the same observation, used twice. The map $\mathbf{y} \mapsto \mathbf{Q}\mathbf{y}$ carries F^k isometrically onto W, and it carries the Rayleigh quotient of $\mathbf{B}$ to the Rayleigh quotient of $\mathbf{A}$. So each of the two Courant–Fischer descriptions of $\lambda_i(\mathbf{B})$ — a maximum of minima, and a minimum of maxima — is the corresponding description of $\lambda_i(\mathbf{A})$ with the competing subspaces restricted to those lying inside W. A maximum over fewer competitors is smaller; a minimum over fewer competitors is larger. That is the whole proof; the only care needed is in matching the dimensions, and it is there that $n - k$ appears.

Proof. **Step 1: transport.** Throughout, $R_{\mathbf{A}}$ and $R_{\mathbf{B}}$ are the Rayleigh quotients of Definition 17.1.1. For $\mathbf{0} \neq \mathbf{y} \in F^k$ we have $\mathbf{Q}\mathbf{y} \neq \mathbf{0}$, since $\|\mathbf{Q}\mathbf{y}\|^2 = \mathbf{y}^*\mathbf{Q}^*\mathbf{Q}\mathbf{y} = \|\mathbf{y}\|^2$ and $\mathbf{y} \neq \mathbf{0}$. The same computation with $\mathbf{A}$ inserted gives

$$\langle \mathbf{B}\mathbf{y}, \mathbf{y} \rangle = \mathbf{y}^*\mathbf{Q}^*\mathbf{A}\mathbf{Q}\mathbf{y} = \langle \mathbf{A}(\mathbf{Q}\mathbf{y}), \mathbf{Q}\mathbf{y} \rangle,$$

so $R_{\mathbf{B}}(\mathbf{y}) = R_{\mathbf{A}}(\mathbf{Q}\mathbf{y})$. Moreover $\mathbf{y} \mapsto \mathbf{Q}\mathbf{y}$ is injective and linear, so it maps the subspaces of

F^k of dimension j bijectively onto the subspaces of W of dimension j, for every j. Consequently, for each j,

$$\{\, R_{\mathbf{B}}(\mathbf{y}) : \mathbf{0} \neq \mathbf{y} \in S \,\} = \{\, R_{\mathbf{A}}(\mathbf{x}) : \mathbf{0} \neq \mathbf{x} \in \mathbf{Q}S \,\}$$

for every subspace $S \subseteq F^k$.

Step 2: the upper bound. By the max–min form of Theorem 17.2.3 applied to $\mathbf{B} \in M_k(F)$,

$$\begin{aligned} \lambda_i(\mathbf{B}) &= \max_{\substack{S \subseteq F^k \\ \dim S = i}} \; \min_{\mathbf{0} \neq \mathbf{y} \in S} R_{\mathbf{B}}(\mathbf{y}) \\ &= \max_{\substack{U \subseteq W \\ \dim U = i}} \; \min_{\mathbf{0} \neq \mathbf{x} \in U} R_{\mathbf{A}}(\mathbf{x}), \end{aligned}$$

the second equality by Step 1. The subspaces $U \subseteq W$ of dimension i form a subfamily of the subspaces $U \subseteq F^n$ of dimension i, and enlarging the family over which a maximum is taken cannot decrease it. So by the max–min form of Theorem 17.2.3 applied to $\mathbf{A}$, the right-hand side is at most $\lambda_i(\mathbf{A})$.

Step 3: the lower bound. By the min–max form of Theorem 17.2.3 applied to $\mathbf{B} \in M_k(F)$, with the complementary dimension $k - i + 1$,

$$\begin{aligned} \lambda_i(\mathbf{B}) &= \min_{\substack{S \subseteq F^k \\ \dim S = k-i+1}} \; \max_{\mathbf{0} \neq \mathbf{y} \in S} R_{\mathbf{B}}(\mathbf{y}) \\ &= \min_{\substack{U \subseteq W \\ \dim U = k-i+1}} \; \max_{\mathbf{0} \neq \mathbf{x} \in U} R_{\mathbf{A}}(\mathbf{x}), \end{aligned}$$

again by Step 1. Put $l = i + n - k$, so that $1 \leq l \leq n$ and $n - l + 1 = k - i + 1$. The min–max form of Theorem 17.2.3 applied to $\mathbf{A}$ therefore reads

$$\lambda_l(\mathbf{A}) = \min_{\substack{U \subseteq F^n \\ \dim U = k-i+1}} \; \max_{\mathbf{0} \neq \mathbf{x} \in U} R_{\mathbf{A}}(\mathbf{x}),$$

a minimum over a **larger** family than the one above, hence no larger. So $\lambda_i(\mathbf{B}) \geq \lambda_{i+n-k}(\mathbf{A})$. This proves the theorem.

■

Remark. Taking W to be the coordinate hyperplane $\operatorname{span}(\mathbf{e}_j : j \neq l)$ makes $\mathbf{B}$ the principal submatrix obtained by deleting row l and column l, and $k = n - 1$ turns the conclusion into $\lambda_i(\mathbf{A}) \geq \lambda_i(\mathbf{B}) \geq \lambda_{i+1}(\mathbf{A})$. That is Theorem 17.4.2. What the present theorem adds is that the coordinates played no part, and that deleting $n - k$ dimensions at once costs exactly $n - k$ indices.

Example. A compression of a diagonal matrix

Let $\mathbf{A} = \operatorname{diag}(5, 3, -2)$ and let

$$\mathbf{q}_1 = \tfrac{1}{3}(2, 2, 1), \qquad \mathbf{q}_2 = \tfrac{1}{3}(-2, 1, 2).$$

Check that $(\mathbf{q}_1, \mathbf{q}_2)$ is orthonormal, compute the compression $\mathbf{B}$ of $\mathbf{A}$ to $W = \operatorname{span}(\mathbf{q}_1, \mathbf{q}_2)$

and its eigenvalues, and verify Theorem 17.6.1.

Solution. Orthonormality. $\|\mathbf{q}_1\|^2 = (4+4+1)/9 = 1$ and $\|\mathbf{q}_2\|^2 = (4+1+4)/9 = 1$, while $9\langle \mathbf{q}_1, \mathbf{q}_2\rangle = -4+2+2 = 0$.

The compression. Multiplying by the diagonal matrix, $3\mathbf{A}\mathbf{q}_1 = (10, 6, -2)$ and $3\mathbf{A}\mathbf{q}_2 = (-10, 3, -4)$. Hence

$$\begin{aligned} b_{11} &= \tfrac{1}{9}(2,2,1)\cdot(10,6,-2) = \tfrac{20+12-2}{9} = \tfrac{10}{3}, \\ b_{21} &= \tfrac{1}{9}(-2,1,2)\cdot(10,6,-2) = \tfrac{-20+6-4}{9} = -2, \\ b_{22} &= \tfrac{1}{9}(-2,1,2)\cdot(-10,3,-4) = \tfrac{20+3-8}{9} = \tfrac{5}{3}, \end{aligned}$$

and $b_{12} = b_{21}$ because $\mathbf{B}$ is symmetric. So

$$\mathbf{B} = \begin{pmatrix} 10/3 & -2 \\ -2 & 5/3 \end{pmatrix}.$$

Its eigenvalues. $\operatorname{tr}\mathbf{B} = 5$ and $\det\mathbf{B} = \frac{50}{9} - 4 = \frac{14}{9}$, so the characteristic polynomial is $x^2 - 5x + \frac{14}{9}$, that is $9x^2 - 45x + 14$. Its discriminant is $2025 - 504 = 1521 = 39^2$, so the roots are $(45 \pm 39)/18$, namely $\lambda_1(\mathbf{B}) = \frac{14}{3}$ and $\lambda_2(\mathbf{B}) = \frac{1}{3}$.

The separation. Here $n = 3$ and $k = 2$, so the theorem says $\lambda_i(\mathbf{A}) \geq \lambda_i(\mathbf{B}) \geq \lambda_{i+1}(\mathbf{A})$ for $i = 1, 2$:

$$5 \;\geq\; \tfrac{14}{3} \;\geq\; 3 \;\geq\; \tfrac{1}{3} \;\geq\; -2,$$

and every inequality is strict. Note also $\operatorname{tr}\mathbf{B} = 5 \leq 5+3 = \lambda_1(\mathbf{A}) + \lambda_2(\mathbf{A})$; the next theorem says that this is no accident and that 8 is the best possible ceiling.

Quick check. Let $\mathbf{A}$ be Hermitian $n \times n$ and let $\mathbf{B}$ be its compression to a k-dimensional subspace. Which single inequality of Theorem 17.6.1 says that a compression of a positive semidefinite matrix is positive semidefinite?

Solution. The case $i = k$ of the lower bound: $\lambda_k(\mathbf{B}) \geq \lambda_{k+n-k}(\mathbf{A}) = \lambda_n(\mathbf{A})$. If $\mathbf{A} \succeq \mathbf{O}$ then $\lambda_n(\mathbf{A}) \geq 0$ by Theorem 13.1.4 (b), so every eigenvalue of $\mathbf{B}$ is ≥ 0, and $\mathbf{B}$ is Hermitian, so $\mathbf{B} \succeq \mathbf{O}$ by the same theorem. (A direct proof is also one line: $\langle \mathbf{B}\mathbf{y}, \mathbf{y}\rangle = \langle \mathbf{A}\mathbf{Q}\mathbf{y}, \mathbf{Q}\mathbf{y}\rangle \geq 0$. The point of the check is to locate the statement inside the interlacing.)

17.6.3 The top eigenvalues, summed

Chapter 13 described λ_1 as a maximum over the unit sphere (Lemma 13.5.2), and Section 1 pointed out that λ_2 is not the "second largest value" of anything. But $\lambda_1 + \lambda_2$ is a maximum again — over pairs of orthonormal vectors rather than single ones. The pattern continues for every k, and it is the sum, never the individual eigenvalue, that behaves.

Theorem 17.6.2. Ky Fan Maximum Principle

Let $\mathbf{A} \in M_n(F)$ be Hermitian and let $1 \le k \le n$. Then

$$\sum_{i=1}^{k} \lambda_i(\mathbf{A}) = \max\Big\{ \sum_{i=1}^{k} \langle \mathbf{A}\mathbf{q}_i, \mathbf{q}_i \rangle \;:\; \mathbf{q}_1, \dots, \mathbf{q}_k \text{ orthonormal} \Big\}$$
$$= \max\big\{ \operatorname{tr}(\mathbf{Q}^{*}\mathbf{A}\mathbf{Q}) \;:\; \mathbf{Q} \in M_{n\times k}(F),\ \mathbf{Q}^{*}\mathbf{Q} = \mathbf{I}_k \big\},$$

and the maximum is attained by $\mathbf{Q} = (\mathbf{u}_1 \ \dots \ \mathbf{u}_k)$ for any orthonormal basis $(\mathbf{u}_1, \dots, \mathbf{u}_n)$ of F^n with $\mathbf{A}\mathbf{u}_i = \lambda_i(\mathbf{A})\mathbf{u}_i$.

Idea. The two right-hand sides are the same number written twice, since the diagonal entries of $\mathbf{Q}^{*}\mathbf{A}\mathbf{Q}$ are exactly the k numbers $\langle \mathbf{A}\mathbf{q}_i, \mathbf{q}_i \rangle$. That reformulation is the whole idea: the quantity to be maximized is the **trace of a compression**, and Theorem 17.6.1 already bounds every eigenvalue of a compression by the matching eigenvalue of $\mathbf{A}$. Add k such bounds and read the left-hand side as a trace. The maximizing $\mathbf{Q}$ is the one that makes the compression diagonal with the top eigenvalues on it, and there is no cleverness in guessing it.

Proof. Write $\mathbf{B} = \mathbf{Q}^{*}\mathbf{A}\mathbf{Q}$ for a matrix $\mathbf{Q}$ with orthonormal columns $\mathbf{q}_1, \dots, \mathbf{q}_k$. Its diagonal entries are $b_{ii} = \langle \mathbf{A}\mathbf{q}_i, \mathbf{q}_i \rangle$ by Definition 17.6.1, so

$$\operatorname{tr} \mathbf{B} = \sum_{i=1}^{k} \langle \mathbf{A}\mathbf{q}_i, \mathbf{q}_i \rangle,$$

and the two sets over which the maxima are taken consist of the same numbers. It therefore suffices to prove the statement for $\operatorname{tr}(\mathbf{Q}^{*}\mathbf{A}\mathbf{Q})$.

The bound. Let $\mathbf{Q}^{*}\mathbf{Q} = \mathbf{I}_k$ and put $\mathbf{B} = \mathbf{Q}^{*}\mathbf{A}\mathbf{Q}$, a Hermitian matrix in $M_k(F)$. By Corollary 12.1.4, $\operatorname{tr} \mathbf{B} = \sum_{i=1}^{k} \lambda_i(\mathbf{B})$. By Theorem 17.6.1, $\lambda_i(\mathbf{B}) \le \lambda_i(\mathbf{A})$ for every $1 \le i \le k$. Summing these k inequalities,

$$\operatorname{tr}(\mathbf{Q}^{*}\mathbf{A}\mathbf{Q}) = \sum_{i=1}^{k} \lambda_i(\mathbf{B}) \;\le\; \sum_{i=1}^{k} \lambda_i(\mathbf{A}).$$

Attainment. Since $\mathbf{A}$ is Hermitian, Corollary 12.4.3 (for $F = \mathbb{C}$) or Corollary 12.5.2 (for $F = \mathbb{R}$) provides an orthonormal basis $(\mathbf{u}_1, \dots, \mathbf{u}_n)$ of F^n of eigenvectors, in any prescribed order; take the order with $\mathbf{A}\mathbf{u}_i = \lambda_i(\mathbf{A})\mathbf{u}_i$. Let $\mathbf{Q}$ have columns $\mathbf{u}_1, \dots, \mathbf{u}_k$; these are orthonormal, so $\mathbf{Q}^{*}\mathbf{Q} = \mathbf{I}_k$. Then

$$\langle \mathbf{A}\mathbf{u}_i, \mathbf{u}_i \rangle = \lambda_i(\mathbf{A}) \langle \mathbf{u}_i, \mathbf{u}_i \rangle = \lambda_i(\mathbf{A}),$$

so $\operatorname{tr}(\mathbf{Q}^{*}\mathbf{A}\mathbf{Q}) = \sum_{i \le k} \lambda_i(\mathbf{A})$.

The bound is therefore attained, so the supremum is a maximum and equals $\sum_{i \le k} \lambda_i(\mathbf{A})$. This proves the theorem.

■

For $k = 1$ this is Lemma 13.5.2 of Chapter 13: λ_1 is the largest value of $\langle \mathbf{A}\mathbf{x}, \mathbf{x} \rangle$ on the unit sphere. For $k = n$ it degenerates in an instructive way: every $\mathbf{Q}$ is unitary, every $\mathbf{Q}^{*}\mathbf{A}\mathbf{Q}$ is

unitarily similar to $\mathbf{A}$, and the "maximum" is attained everywhere, at the common value $\operatorname{tr}\mathbf{A}$. The content is in between.

Example. Which pairs of directions capture the most

Let

$$\mathbf{A} = \begin{pmatrix} 2 & 1 & 0 \\ 1 & 2 & 0 \\ 0 & 0 & -1 \end{pmatrix}.$$

Find $\max\{\operatorname{tr}(\mathbf{Q}^{\top}\mathbf{A}\mathbf{Q}) : \mathbf{Q}^{\top}\mathbf{Q} = \mathbf{I}_2\}$ and a maximizer, and compare the values at $\mathbf{Q} = (\mathbf{e}_1\ \mathbf{e}_2)$ and $\mathbf{Q} = (\mathbf{e}_1\ \mathbf{e}_3)$.

Solution. The eigenvalues: $\mathbf{e}_3$ is an eigenvector for -1, and the top-left block $\left(\begin{smallmatrix} 2 & 1 \\ 1 & 2 \end{smallmatrix}\right)$ has eigenvectors $(1,1)$ and $(1,-1)$ for 3 and 1. So $\lambda_1 = 3$, $\lambda_2 = 1$, $\lambda_3 = -1$, with orthonormal eigenvectors $\mathbf{u}_1 = \frac{1}{\sqrt{2}}(1,1,0)$, $\mathbf{u}_2 = \frac{1}{\sqrt{2}}(1,-1,0)$, $\mathbf{u}_3 = \mathbf{e}_3$. By Theorem 17.6.2 the maximum is $3 + 1 = 4$, attained at $\mathbf{Q} = (\mathbf{u}_1\ \mathbf{u}_2)$.

At $\mathbf{Q} = (\mathbf{e}_1\ \mathbf{e}_2)$ the trace is $a_{11} + a_{22} = 4$: this is also a maximizer. That is no accident — $\operatorname{span}(\mathbf{e}_1, \mathbf{e}_2) = \operatorname{span}(\mathbf{u}_1, \mathbf{u}_2)$, and the trace of a compression, like its eigenvalues, depends only on the subspace. At $\mathbf{Q} = (\mathbf{e}_1\ \mathbf{e}_3)$ the trace is $2 + (-1) = 1$, well below the ceiling, as Theorem 17.6.2 requires.

The simplest competitors are the standard basis vectors, and they give a statement about the diagonal of $\mathbf{A}$ for free.

Corollary 17.6.3. Partial Sums of the Diagonal

Let $\mathbf{A} \in M_n(F)$ be Hermitian with diagonal entries $a_{11}, \dots, a_{nn}$, and let $1 \le i_1 < \cdots < i_k \le n$. Then

$$\sum_{j=1}^{k} a_{i_j i_j} \;\le\; \sum_{i=1}^{k} \lambda_i(\mathbf{A}),$$

with equality for $k = n$.

Proof. The list $(\mathbf{e}_{i_1}, \dots, \mathbf{e}_{i_k})$ is orthonormal, and $\langle \mathbf{A}\mathbf{e}_i, \mathbf{e}_i \rangle = a_{ii}$. Apply Theorem 17.6.2. For $k = n$ both sides equal $\operatorname{tr}\mathbf{A}$, by Corollary 12.1.4. ∎

Quick check. Can a Hermitian 3×3 matrix with diagonal entries $5, 1, 0$ have eigenvalues $4, 2, 0$? Can it have eigenvalues $6, 0, 0$?

Solution. Not the first: the case $k = 1$ of Corollary 17.6.3 says every diagonal entry is at most λ_1, and $5 > 4$. The second passes every test — $5 \le 6$, $5 + 1 \le 6$, $6 = 6$ — and is realized: with $\mathbf{v} = (\sqrt{5}, 1, 0)$, the matrix $\mathbf{v}\mathbf{v}^{\top}$ has diagonal $5, 1, 0$; it sends $\mathbf{v}$ to $\|\mathbf{v}\|^2\mathbf{v} = 6\mathbf{v}$

and kills the plane orthogonal to $\mathbf{v}$, so its eigenvalues are $6, 0, 0$. Whether passing every test is always enough is the converse question, which Section 8 takes up.

17.6.4 The promise from Chapter 13

Chapter 13 §11 proved a bound on how much of a matrix an orthonormal list can capture (Lemma 13.11.2), and said of it: it "is the extreme case of a family of statements about sums of the largest eigenvalues of a Hermitian matrix; Chapter 17 develops the family through the Courant–Fischer min–max theorem". Theorem 17.6.2 is that family. Here is the lemma, recovered from it in four lines.

Corollary 17.6.4. The Capture Bound as a Case of Ky Fan

Let $\mathbf{A} \in M_{m\times n}(F)$ with singular values $\sigma_1 \geq \cdots \geq \sigma_p$, $p = \min(m, n)$, extended by $\sigma_i = 0$ for $p < i \leq m$, and let $\mathbf{q}_1, \ldots, \mathbf{q}_s$ be an orthonormal list in F^m. Then

$$\sum_{j=1}^{s} \|\mathbf{A}^*\mathbf{q}_j\|^2 \;\leq\; \sum_{i=1}^{s} \sigma_i^2.$$

Proof. If $s = 0$ both sides are empty sums, equal to 0; so assume $s \geq 1$. Put $\mathbf{M} = \mathbf{A}\mathbf{A}^* \in M_m(F)$, which is Hermitian. For each j,

$$\|\mathbf{A}^*\mathbf{q}_j\|^2 = (\mathbf{A}^*\mathbf{q}_j)^*(\mathbf{A}^*\mathbf{q}_j) = \mathbf{q}_j^*\mathbf{M}\mathbf{q}_j = \langle \mathbf{M}\mathbf{q}_j, \mathbf{q}_j\rangle.$$

Fix a singular value decomposition $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$ (Theorem 13.9.1). Since $\mathbf{V}^*\mathbf{V} = \mathbf{I}_n$,

$$\mathbf{M} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*\mathbf{V}\mathbf{\Sigma}^*\mathbf{U}^* = \mathbf{U}(\mathbf{\Sigma}\mathbf{\Sigma}^*)\mathbf{U}^*,$$

and $\mathbf{\Sigma}\mathbf{\Sigma}^*$ is the $m \times m$ diagonal matrix with diagonal $\sigma_1^2, \ldots, \sigma_p^2$ followed by $m - p$ zeros — that is, $\sigma_1^2 \geq \cdots \geq \sigma_m^2$ in the extended notation. So this display is a unitary diagonalization of $\mathbf{M}$, and Corollary 12.4.3 identifies its diagonal as the eigenvalue list: $\lambda_i(\mathbf{M}) = \sigma_i^2$ for $1 \leq i \leq m$. Now Theorem 17.6.2 applied to $\mathbf{M}$ with $k = s$ gives

$$\sum_{j=1}^{s} \langle \mathbf{M}\mathbf{q}_j, \mathbf{q}_j\rangle \;\leq\; \sum_{i=1}^{s} \lambda_i(\mathbf{M}) = \sum_{i=1}^{s} \sigma_i^2,$$

which is the claim. ∎

This is Lemma 13.11.2 exactly, hypotheses and conclusion, so the debt is paid. Two things are worth saying about the shape of the payment.

First, Chapter 13 had to prove the inequality by hand, weight by weight, because it had no min–max theorem; the argument there is a careful estimate on a weighted sum of a decreasing sequence. Here the same estimate is invisible: it has been absorbed into Theorem 17.2.3, whose proof is a dimension count, and passed on through Theorem 17.6.1.

Second, the words "extreme case". Strictly, the lemma is a *special* case of the family rather than an extreme one: it is the instance in which the Hermitian matrix is a product $\mathbf{A}\mathbf{A}^*$, so that

every eigenvalue is a squared singular value. Its crudest instance, $s = 1$, says $\|\mathbf{A}^*\mathbf{q}\|^2 \le \sigma_1^2$ for a unit $\mathbf{q}$, that is, $\|\mathbf{A}^*\|_2 = \sigma_1$. The same one line of Ky Fan, run with the other product — Theorem 17.6.2 with $k = 1$ applied to the Hermitian matrix $\mathbf{A}^*\mathbf{A} \in M_n(F)$ — whose largest eigenvalue is σ_1^2 by Definition 13.9.1 — gives

$$\max_{\|\mathbf{x}\|=1} \|\mathbf{A}\mathbf{x}\|^2 = \max_{\|\mathbf{x}\|=1} \langle \mathbf{A}^*\mathbf{A}\mathbf{x}, \mathbf{x}\rangle = \lambda_1(\mathbf{A}^*\mathbf{A}) = \sigma_1^2.$$

So the largest singular value is the largest stretching factor, which is Theorem 16.3.3 (c), $\|\mathbf{A}\|_2 = \sigma_1(\mathbf{A})$. Chapter 16 proved that from the singular value decomposition; here it is one line of Ky Fan.

17.6.5 Sums are convex; individual eigenvalues are not

Corollary 17.6.5. Subadditivity of the Top Partial Sums

Let $\mathbf{A}, \mathbf{B} \in M_n(F)$ be Hermitian and let $1 \le k \le n$. Then

$$\sum_{i=1}^{k} \lambda_i(\mathbf{A}+\mathbf{B}) \;\le\; \sum_{i=1}^{k} \lambda_i(\mathbf{A}) + \sum_{i=1}^{k} \lambda_i(\mathbf{B}).$$

Proof. By Theorem 17.6.2 applied to the Hermitian matrix $\mathbf{A}+\mathbf{B}$, there is a $\mathbf{Q} \in M_{n\times k}(F)$ with $\mathbf{Q}^*\mathbf{Q} = \mathbf{I}_k$ attaining the maximum, so that

$$\sum_{i=1}^{k} \lambda_i(\mathbf{A}+\mathbf{B}) = \operatorname{tr}(\mathbf{Q}^*(\mathbf{A}+\mathbf{B})\mathbf{Q}).$$

By Theorem 0.9.6 (1) the right-hand side is $\operatorname{tr}(\mathbf{Q}^*\mathbf{A}\mathbf{Q}) + \operatorname{tr}(\mathbf{Q}^*\mathbf{B}\mathbf{Q})$. This one $\mathbf{Q}$ is a competitor in each of the two maximization problems of Theorem 17.6.2, for $\mathbf{A}$ and for $\mathbf{B}$, so each summand is at most the corresponding maximum. Adding the two bounds proves the corollary. ∎

Write $s_k(\mathbf{A}) = \sum_{i\le k} \lambda_i(\mathbf{A})$ for the function just bounded. Two properties have now been proved: it is **subadditive**, $s_k(\mathbf{A}+\mathbf{B}) \le s_k(\mathbf{A}) + s_k(\mathbf{B})$, and it is **positively homogeneous**, $s_k(t\mathbf{A}) = t s_k(\mathbf{A})$ for $t \ge 0$, since multiplying a Hermitian matrix by $t \ge 0$ multiplies each eigenvalue by t without disturbing the order. Together these say that for $t \in [0,1]$,

$$s_k\big(t\mathbf{A} + (1-t)\mathbf{B}\big) \;\le\; t s_k(\mathbf{A}) + (1-t) s_k(\mathbf{B}) :$$

s_k is a **convex** function on the real vector space of Hermitian matrices. That is exactly what one expects of a maximum of linear functions, which is what Theorem 17.6.2 says s_k is: each fixed $\mathbf{Q}$ makes $\mathbf{A} \mapsto \operatorname{tr}(\mathbf{Q}^*\mathbf{A}\mathbf{Q})$ linear, and s_k is the pointwise maximum of this family. Chapter 18 studies convex functions in general, and the observation that a pointwise supremum of linear functions is convex is one of its basic facts.

Corollary 17.6.6. The Bottom Partial Sums as a Minimum

Let $\mathbf{A} \in M_n(F)$ be Hermitian and let $1 \le k \le n$. Then

$$\sum_{i=n-k+1}^{n} \lambda_i(\mathbf{A}) = \min\{\operatorname{tr}(\mathbf{Q}^*\mathbf{A}\mathbf{Q}) : \mathbf{Q}^*\mathbf{Q} = \mathbf{I}_k\},$$

the minimum being attained at the bottom k eigenvectors.

Proof. The eigenvalues of $-\mathbf{A}$, with multiplicity, are the negatives of those of $\mathbf{A}$, and negation reverses order, so

$$\lambda_i(-\mathbf{A}) = -\lambda_{n+1-i}(\mathbf{A}) \qquad (1 \le i \le n).$$

Apply Theorem 17.6.2 to $-\mathbf{A}$. Its left-hand side is

$$\sum_{i=1}^{k} \lambda_i(-\mathbf{A}) = -\sum_{i=1}^{k} \lambda_{n+1-i}(\mathbf{A}) = -\sum_{i=n-k+1}^{n} \lambda_i(\mathbf{A}),$$

the last step re-indexing the sum. Its right-hand side is

$$\max_{\mathbf{Q}^*\mathbf{Q}=\mathbf{I}_k} \operatorname{tr}(\mathbf{Q}^*(-\mathbf{A})\mathbf{Q}) = -\min_{\mathbf{Q}^*\mathbf{Q}=\mathbf{I}_k} \operatorname{tr}(\mathbf{Q}^*\mathbf{A}\mathbf{Q}),$$

using Theorem 0.9.6 (1) and the fact that maximizing $-g$ is minimizing g. Negating both sides gives the claim, and the maximizer for $-\mathbf{A}$ is a list of top eigenvectors of $-\mathbf{A}$, which are bottom eigenvectors of $\mathbf{A}$.

■

Warning. **Only the sums are convex. A single eigenvalue λ_k with $1 < k < n$ is neither convex nor concave in $\mathbf{A}$.** Both failures happen among 3×3 diagonal matrices, where λ_2 is just the middle entry; for general n and $1 < k < n$, take the direct sum of each witness with $2\mathbf{I}_{k-2}$ above and $-\mathbf{I}_{n-k-1}$ below, which moves the middle entry to position k without changing the computation.

Not convex. Take $\mathbf{A} = \operatorname{diag}(1,0,0)$ and $\mathbf{B} = \operatorname{diag}(0,1,0)$, so $\lambda_2(\mathbf{A}) = \lambda_2(\mathbf{B}) = 0$. Their midpoint is $\frac{1}{2}(\mathbf{A}+\mathbf{B}) = \operatorname{diag}(\frac{1}{2},\frac{1}{2},0)$, with $\lambda_2 = \frac{1}{2}$. Convexity would require $\frac{1}{2} \le 0$.

Not concave. Take $\mathbf{A} = \operatorname{diag}(1,1,0)$ and $\mathbf{B} = \operatorname{diag}(1,0,1)$, so $\lambda_2(\mathbf{A}) = \lambda_2(\mathbf{B}) = 1$. Their midpoint is $\operatorname{diag}(1,\frac{1}{2},\frac{1}{2})$, with $\lambda_2 = \frac{1}{2}$. Concavity would require $\frac{1}{2} \ge 1$.

The partial sum s_2 satisfies the **convexity** inequality on both pairs, as it must: in the first pair its values at the two matrices and their midpoint are $1, 1, 1$, and in the second $2, 2, \frac{3}{2}$. The moral is that $\lambda_k = s_k - s_{k-1}$ is a **difference** of two convex functions, and a difference of convex functions has, in general, no convexity property at all. This is why the sharpest statements in this chapter are about sums, and why a statement about a single intermediate eigenvalue usually has to be proved as a statement about two sums.

17.6.6 Exercises

17.6.6.1 A. Check your understanding

Exercise. A1

(a) Define the compression of $\mathbf{A}$ to a subspace, and say what has to be checked before the phrase "the eigenvalues of the compression" makes sense.

(b) State Theorem 17.6.2, including where the maximum is attained.

(c) Explain why the maximum in Theorem 17.6.2 does not need an appeal to compactness.

(d) Is $\mathbf{A} \mapsto \lambda_1(\mathbf{A})$ convex on the Hermitian matrices? Is $\mathbf{A} \mapsto \lambda_n(\mathbf{A})$? Justify each answer.

(e) State Theorem 17.6.1 and say which case of it is Theorem 17.4.2.

Solution.

(a) For an orthonormal list $\mathbf{q}_1, \dots, \mathbf{q}_k$ with matrix $\mathbf{Q}$, the compression of $\mathbf{A}$ to $W = \operatorname{span}(\mathbf{q}_1, \dots, \mathbf{q}_k)$ is $\mathbf{Q}^*\mathbf{A}\mathbf{Q}$, the $k \times k$ matrix with entries $\langle \mathbf{A}\mathbf{q}_j, \mathbf{q}_i \rangle$. What must be checked is that changing the orthonormal basis of W changes $\mathbf{Q}^*\mathbf{A}\mathbf{Q}$ only by a unitary similarity, so the eigenvalues depend on W alone.

(b) For Hermitian $\mathbf{A} \in M_n(F)$ and $1 \le k \le n$, $\sum_{i \le k} \lambda_i(\mathbf{A})$ is the maximum of $\operatorname{tr}(\mathbf{Q}^*\mathbf{A}\mathbf{Q})$ over all $\mathbf{Q} \in M_{n \times k}(F)$ with $\mathbf{Q}^*\mathbf{Q} = \mathbf{I}_k$, equivalently the maximum of $\sum_{i \le k} \langle \mathbf{A}\mathbf{q}_i, \mathbf{q}_i \rangle$ over orthonormal lists. It is attained at $\mathbf{Q} = (\mathbf{u}_1 \ \dots \ \mathbf{u}_k)$ for an orthonormal eigenbasis ordered with $\mathbf{A}\mathbf{u}_i = \lambda_i(\mathbf{A})\mathbf{u}_i$.

(c) Because the proof exhibits both a bound valid for every competitor and a competitor achieving it. A supremum that is achieved is a maximum, whatever the topology of the competing set.

(d) Incorrect. Take $\mathbf{A} = \operatorname{diag}(1,0,0)$ and $\mathbf{B} = \operatorname{diag}(0,1,0)$ as in the warning: $\lambda_2(\mathbf{A}) = \lambda_2(\mathbf{B}) = 0$ while $\mathbf{A}+\mathbf{B} = \operatorname{diag}(1,1,0)$ has $\lambda_2 = 1$. What is true is the statement about sums, Corollary 17.6.5: here $\lambda_1 + \lambda_2$ of the sum is 2, and $(1+0)+(1+0) = 2$.

(e) If $\mathbf{B}$ is the compression of the Hermitian $\mathbf{A} \in M_n(F)$ to a k-dimensional subspace, then $\lambda_i(\mathbf{A}) \ge \lambda_i(\mathbf{B}) \ge \lambda_{i+n-k}(\mathbf{A})$ for $1 \le i \le k$. Cauchy interlacing is the case $k = n-1$ with the subspace a coordinate hyperplane.

17.6.6.2 B. Practice

Exercise. B1: A compression, and two bounds

Let $\mathbf{A} = \operatorname{diag}(5, 0, -2)$ and let $\mathbf{q}_1 = \frac{1}{3}(1,2,2)$, $\mathbf{q}_2 = \frac{1}{3}(2,1,-2)$.

(a) Verify that $(\mathbf{q}_1, \mathbf{q}_2)$ is orthonormal and compute the compression $\mathbf{B}$ of $\mathbf{A}$ to $\operatorname{span}(\mathbf{q}_1, \mathbf{q}_2)$.

(b) Find the eigenvalues of $\mathbf{B}$ and verify Theorem 17.6.1.

(c) Compare $\operatorname{tr}\mathbf{B}$ with the bound of Theorem 17.6.2, and say how far this subspace is from optimal.

Solution.

(a) $\|\mathbf{q}_1\|^2 = (1+4+4)/9 = 1$, $\|\mathbf{q}_2\|^2 = (4+1+4)/9 = 1$, and $9\langle \mathbf{q}_1, \mathbf{q}_2 \rangle = 2+2-4 = 0$.

Next $3\mathbf{A}\mathbf{q}_1 = (5,0,-4)$ and $3\mathbf{A}\mathbf{q}_2 = (10,0,4)$, so

$$b_{11} = \tfrac{1}{9}(1,2,2)\cdot(5,0,-4) = \tfrac{5-8}{9} = -\tfrac{1}{3},$$
$$b_{12} = \tfrac{1}{9}(1,2,2)\cdot(10,0,4) = \tfrac{10+8}{9} = 2,$$
$$b_{22} = \tfrac{1}{9}(2,1,-2)\cdot(10,0,4) = \tfrac{20-8}{9} = \tfrac{4}{3},$$

so $\mathbf{B} = \left(\begin{smallmatrix} -1/3 & 2 \\ 2 & 4/3 \end{smallmatrix}\right)$.

(b) $\operatorname{tr}\mathbf{B} = 1$ and $\det\mathbf{B} = -\tfrac{4}{9} - 4 = -\tfrac{40}{9}$, so $9x^2 - 9x - 40 = 0$, with discriminant $81 + 1440 = 1521 = 39^2$ and roots $(9 \pm 39)/18$. Hence $\lambda_1(\mathbf{B}) = \tfrac{8}{3}$ and $\lambda_2(\mathbf{B}) = -\tfrac{5}{3}$. The separation $\lambda_i(\mathbf{A}) \geq \lambda_i(\mathbf{B}) \geq \lambda_{i+1}(\mathbf{A})$ reads

$$5 \;\geq\; \tfrac{8}{3} \;\geq\; 0 \;\geq\; -\tfrac{5}{3} \;\geq\; -2,$$

all strict.

(c) $\operatorname{tr}\mathbf{B} = 1$, while Theorem 17.6.2 gives $\lambda_1(\mathbf{A}) + \lambda_2(\mathbf{A}) = 5 + 0 = 5$ as the maximum, attained at $\operatorname{span}(\mathbf{e}_1, \mathbf{e}_2)$. So this subspace scores 1 out of a possible 5; it is far from optimal, and one can see why. Between them $\mathbf{q}_1$ and $\mathbf{q}_2$ put weight $\tfrac{4}{9} + \tfrac{4}{9} = \tfrac{8}{9}$ on the third coordinate, where $\mathbf{A}$ is -2, and only $\tfrac{1}{9} + \tfrac{4}{9} = \tfrac{5}{9}$ on the first, where it is 5: indeed $5\cdot\tfrac{5}{9} + 0 - 2\cdot\tfrac{8}{9} = 1$.

Exercise. B2: Reading the sums off the spectrum

A Hermitian $\mathbf{A} \in M_4(\mathbb{R})$ has eigenvalues $6, 4, 1, -3$.

(a) Compute $\max\{\operatorname{tr}(\mathbf{Q}^\top\mathbf{A}\mathbf{Q}) : \mathbf{Q}^\top\mathbf{Q} = \mathbf{I}_k\}$ for $k = 1, 2, 3, 4$.

(b) Compute the corresponding minima, using Corollary 17.6.6.

(c) For $\mathbf{A} = \operatorname{diag}(6,4,1,-3)$ and $\mathbf{q}_1 = \tfrac{1}{\sqrt{2}}(1,1,0,0)$, $\mathbf{q}_2 = \tfrac{1}{\sqrt{2}}(0,0,1,1)$, compute $\operatorname{tr}(\mathbf{Q}^\top\mathbf{A}\mathbf{Q})$ and check it against (a) and (b).

Solution.

(a) By Theorem 17.6.2 the maxima are the partial sums from the top: 6, 10, 11, 8.

(b) By Corollary 17.6.6 the minima are the partial sums from the bottom: -3, -2, 2, 8. The two agree at $k = 4$, as they must, both being $\operatorname{tr}\mathbf{A}$.

(c) $\langle\mathbf{A}\mathbf{q}_1, \mathbf{q}_1\rangle = \tfrac{1}{2}(6+4) = 5$ and $\langle\mathbf{A}\mathbf{q}_2, \mathbf{q}_2\rangle = \tfrac{1}{2}(1-3) = -1$, so $\operatorname{tr}(\mathbf{Q}^\top\mathbf{A}\mathbf{Q}) = 4$. Indeed $-2 \leq 4 \leq 10$, as (a) and (b) require for $k = 2$.

Exercise. B3: Subadditivity, and how loose it can be

Let

$$\mathbf{A} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}, \qquad \mathbf{B} = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}.$$

(a) Compute the eigenvalues of $\mathbf{A}$, of $\mathbf{B}$, and of $\mathbf{A} + \mathbf{B}$.

(b) Verify Corollary 17.6.5 for $k = 1$ and $k = 2$, and say for which k it is strict.

(c) Give a pair of Hermitian 2×2 matrices for which the $k = 1$ inequality is an equality.

Solution.

(a) $\mathbf{A}$ has trace 2 and determinant 0, so its eigenvalues are $2, 0$; the same for $\mathbf{B}$. And $\mathbf{A}+\mathbf{B} = 2\mathbf{I}_2$ has eigenvalues $2, 2$.

(b) For $k = 1$: $\lambda_1(\mathbf{A}+\mathbf{B}) = 2$ and $\lambda_1(\mathbf{A}) + \lambda_1(\mathbf{B}) = 4$, so $2 \le 4$, strictly. For $k = 2$: $4 \le 2 + 2 = 4$, an equality — as it must be, since for $k = n$ both sides are traces and Theorem 0.9.6 (1) makes the trace additive.

(c) Take $\mathbf{A} = \mathbf{B} = \operatorname{diag}(1, 0)$. Then $\mathbf{A}+\mathbf{B} = \operatorname{diag}(2, 0)$ and $\lambda_1(\mathbf{A}+\mathbf{B}) = 2 = 1+1$. More generally equality holds at $k = 1$ whenever $\mathbf{A}$ and $\mathbf{B}$ have a common top eigenvector, because then that one vector attains all three maxima at once.

17.6.6.3 C. Going deeper

Exercise. C1: The Loewner order and the partial sums

Let $\mathbf{A}, \mathbf{B} \in M_n(F)$ be Hermitian.

(a) Prove that $\mathbf{A} \succeq \mathbf{B}$ implies $\sum_{i\le k} \lambda_i(\mathbf{A}) \ge \sum_{i\le k} \lambda_i(\mathbf{B})$ for every k. *Hint: compare the two maximization problems of Theorem 17.6.2 competitor by competitor.*

(b) Show that the converse fails: give Hermitian $\mathbf{A}, \mathbf{B}$ satisfying all n of these inequalities with $\mathbf{A} \not\succeq \mathbf{B}$.

Solution.

(a) Let $\mathbf{Q}^*\mathbf{Q} = \mathbf{I}_k$ with columns $\mathbf{q}_1, \dots, \mathbf{q}_k$. Since $\mathbf{A} - \mathbf{B} \succeq \mathbf{O}$, clause (P2) of Definition 13.1.1 gives $\langle (\mathbf{A}-\mathbf{B})\mathbf{q}_i, \mathbf{q}_i\rangle \ge 0$ for each i, that is, $\langle \mathbf{A}\mathbf{q}_i, \mathbf{q}_i\rangle \ge \langle \mathbf{B}\mathbf{q}_i, \mathbf{q}_i\rangle$. Summing over i,

$$\operatorname{tr}(\mathbf{Q}^*\mathbf{A}\mathbf{Q}) \;\ge\; \operatorname{tr}(\mathbf{Q}^*\mathbf{B}\mathbf{Q}).$$

Now take $\mathbf{Q}$ to be a maximizer for $\mathbf{B}$ in Theorem 17.6.2. Then

$$\sum_{i\le k} \lambda_i(\mathbf{B}) = \operatorname{tr}(\mathbf{Q}^*\mathbf{B}\mathbf{Q}) \le \operatorname{tr}(\mathbf{Q}^*\mathbf{A}\mathbf{Q}) \le \sum_{i\le k} \lambda_i(\mathbf{A}),$$

the last step because $\mathbf{Q}$ is one competitor in the maximization for $\mathbf{A}$.

(b) Take $\mathbf{A} = \operatorname{diag}(2, 0)$ and $\mathbf{B} = \operatorname{diag}(1, 1)$. The partial sums are $2 \ge 1$ for $k = 1$ and $2 \ge 2$ for $k = 2$. But $\mathbf{A} - \mathbf{B} = \operatorname{diag}(1, -1)$ has a negative eigenvalue, so $\mathbf{A} - \mathbf{B} \not\succeq \mathbf{O}$ by Theorem 13.1.4 (b), and $\mathbf{A} \not\succeq \mathbf{B}$. Concretely, $\langle (\mathbf{A} - \mathbf{B})\mathbf{e}_2, \mathbf{e}_2\rangle = -1 < 0$.

Exercise. C2: Partial sums that behave like a norm

For positive semidefinite $\mathbf{A} \in M_n(F)$ and $1 \le k \le n$, consider $s_k(\mathbf{A}) = \sum_{i\le k} \lambda_i(\mathbf{A})$, the partial sum of this section, now restricted to the positive semidefinite matrices.

(a) Prove that $N_k(\mathbf{A}) \ge 0$, with equality if and only if $\mathbf{A} = \mathbf{O}$.

(b) Prove that $N_k(c\mathbf{A}) = cN_k(\mathbf{A})$ for every real $c \ge 0$, and that $N_k(\mathbf{A}+\mathbf{B}) \le N_k(\mathbf{A}) + N_k(\mathbf{B})$ for positive semidefinite $\mathbf{A}, \mathbf{B}$.

(c) The positive semidefinite matrices form a **cone**, not a vector space, so a norm cannot be defined on them. Deduce instead that s_k is positive, positively homogeneous, and

subadditive on this cone. Which step of your argument for (a) uses positive semidefiniteness, and what goes wrong without it?

Solution.

(a) By Theorem 13.1.4 (b) every eigenvalue of $\mathbf{A}$ is ≥ 0, so $N_k(\mathbf{A}) \geq \lambda_1(\mathbf{A}) \geq 0$ for $k \geq 1$. If $N_k(\mathbf{A}) = 0$ then in particular $\lambda_1(\mathbf{A}) = 0$, so every eigenvalue is 0; by Corollary 12.4.3 (or Corollary 12.5.2) $\mathbf{A} = \mathbf{U}\mathbf{O}\mathbf{U}^* = \mathbf{O}$. Conversely $N_k(\mathbf{O}) = 0$.

(b) Multiplying a Hermitian matrix by $c \geq 0$ multiplies each eigenvalue by c and preserves the decreasing order, so $\lambda_i(c\mathbf{A}) = c\lambda_i(\mathbf{A})$ and the first claim follows by summing. The second is Corollary 17.6.5, which needs no positivity at all.

(c) Positivity with $N_k(\mathbf{A}) = 0$ only for $\mathbf{A} = \mathbf{O}$ is (a); homogeneity under scalars $c \geq 0$ is (b); subadditivity is (b). So N_k behaves as a norm wherever it is defined. It cannot be extended to all Hermitian matrices by the same formula: for $\mathbf{A} = \operatorname{diag}(0, -1)$ and $k = 1$ we get $N_1(\mathbf{A}) = 0$ with $\mathbf{A} \neq \mathbf{O}$, so the positivity clause fails, and for $c = -1$ the value $N_1(-\mathbf{A}) = 1$ differs from $|c|N_1(\mathbf{A}) = 0$. The repair is to apply N_k to the singular values instead of the eigenvalues, which is what Chapter 21 does.

Exercise. C3: How much of a matrix a subspace holds

Let $\mathbf{A} \in M_{m\times n}(F)$ with singular values $\sigma_1 \geq \cdots \geq \sigma_p$ extended by zeros as in Corollary 17.6.4, let $S \subseteq F^m$ have dimension k, and let $\mathbf{Q}$ have as columns an orthonormal basis of S, so that $P_S = \mathbf{Q}\mathbf{Q}^*$ by Theorem 11.3.3 (a).

(a) Prove that $\|P_S\mathbf{A}\|_F^2 = \operatorname{tr}(\mathbf{Q}^*\mathbf{A}\mathbf{A}^*\mathbf{Q})$.

(b) Deduce from Theorem 17.6.2 that $\|P_S\mathbf{A}\|_F^2 \leq \sigma_1^2 + \cdots + \sigma_k^2$, with equality when S is spanned by the first k left singular vectors.

(c) Hence recover the value $\sum_{i>k}\sigma_i^2$ of the minimum in Corollary 13.11.4. *Hint:* $\mathbf{a} - P_S\mathbf{a}$ *is orthogonal to* $P_S\mathbf{a}$.

Solution.

(a) The Frobenius norm satisfies $\|\mathbf{X}\|_F^2 = \operatorname{tr}(\mathbf{X}^*\mathbf{X})$, because the (j, j) entry of $\mathbf{X}^*\mathbf{X}$ is the squared length of the j-th column of $\mathbf{X}$. With $P_S = \mathbf{Q}\mathbf{Q}^*$, which is Hermitian and satisfies $P_S^2 = \mathbf{Q}(\mathbf{Q}^*\mathbf{Q})\mathbf{Q}^* = P_S$,

$$\|P_S\mathbf{A}\|_F^2 = \operatorname{tr}(\mathbf{A}^*P_S^*P_S\mathbf{A}) = \operatorname{tr}(\mathbf{A}^*\mathbf{Q}\mathbf{Q}^*\mathbf{A}) = \operatorname{tr}(\mathbf{Q}^*\mathbf{A}\mathbf{A}^*\mathbf{Q}),$$

the last step by Theorem 0.9.6 (3) applied to $\mathbf{A}^*\mathbf{Q}$ and $\mathbf{Q}^*\mathbf{A}$.

(b) The matrix $\mathbf{M} = \mathbf{A}\mathbf{A}^*$ is Hermitian with $\lambda_i(\mathbf{M}) = \sigma_i^2$, as computed in the proof of Corollary 17.6.4. By Theorem 17.6.2 with k columns,

$$\operatorname{tr}(\mathbf{Q}^*\mathbf{M}\mathbf{Q}) \leq \sum_{i\leq k}\lambda_i(\mathbf{M}) = \sum_{i\leq k}\sigma_i^2,$$

with equality when the columns of $\mathbf{Q}$ are top eigenvectors of $\mathbf{M}$. From $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$ the columns $\mathbf{u}_1, \ldots, \mathbf{u}_m$ of $\mathbf{U}$ are exactly such an eigenbasis, so equality holds for $S = \operatorname{span}(\mathbf{u}_1, \ldots, \mathbf{u}_k)$.

(c) Write $\mathbf{a}_1, \dots, \mathbf{a}_n$ for the columns of $\mathbf{A}$. For each j the vectors $P_S\mathbf{a}_j$ and $\mathbf{a}_j - P_S\mathbf{a}_j$ are orthogonal, so Theorem 11.1.3 gives $\|\mathbf{a}_j\|^2 = \|P_S\mathbf{a}_j\|^2 + \|\mathbf{a}_j - P_S\mathbf{a}_j\|^2$. Summing over j,

$$\sum_j d(\mathbf{a}_j, S)^2 = \|\mathbf{A}\|_F^2 - \|P_S\mathbf{A}\|_F^2.$$

By Lemma 13.11.1, $\|\mathbf{A}\|_F^2 = \sum_i \sigma_i^2$. By (b) the subtracted term is at most $\sum_{i\le k}\sigma_i^2$, with equality for $S = \operatorname{span}(\mathbf{u}_1, \dots, \mathbf{u}_k)$. Hence the total squared distance is at least $\sum_{i>k}\sigma_i^2$, and equals it for that S; a subspace of smaller dimension $k' < k$ does no better, since the same argument bounds what it captures by $\sum_{i\le k'}\sigma_i^2 \le \sum_{i\le k}\sigma_i^2$ — which is Corollary 13.11.4, now read as a statement about a maximum of traces rather than a minimum of distances.

17.7 Arbitrary Index Sets

The Courant–Fischer theorem describes one eigenvalue at a time, and Ky Fan's theorem describes the sum of the top k. Between those two lies every other sum, for instance $\lambda_2 + \lambda_4$, and §06 ended by warning that the sharpest statements about eigenvalues tend to be statements about sums. This section describes the sum over an **arbitrary** index set as a max–min, which is Wielandt's theorem. It then uses that description to compare two Hermitian matrices across any index set at once, which is Lidskii's inequality. The price of the generality is that single subspaces are replaced by nested chains of them, and the proof takes some care.

Throughout, $F = \mathbb{R}$ or $F = \mathbb{C}$, every matrix whose eigenvalues are indexed is Hermitian, and the indexing is decreasing, $\lambda_1(\mathbf{A}) \ge \cdots \ge \lambda_n(\mathbf{A})$, with multiplicity. An **index set** of size k is a list of integers $1 \le i_1 < i_2 < \cdots < i_k \le n$.

17.7.1 What adding up the min–max gives

Here is the question that drives the section. Let $\mathbf{A}$ and $\mathbf{B}$ be Hermitian and fix an index set. How large can

$$\sum_{j=1}^{k} \lambda_{i_j}(\mathbf{A}) - \sum_{j=1}^{k} \lambda_{i_j}(\mathbf{B})$$

be, in terms of $\mathbf{A} - \mathbf{B}$? One answer is immediate. By Corollary 17.3.4, applied to $\mathbf{B}$ and the Hermitian matrix $\mathbf{A} - \mathbf{B}$, each single difference satisfies $\lambda_i(\mathbf{A}) - \lambda_i(\mathbf{B}) \le \lambda_1(\mathbf{A} - \mathbf{B})$. Adding k of these gives the bound $k\,\lambda_1(\mathbf{A} - \mathbf{B})$.

That bound is poor, and a rank-one example shows by how much. Let $\mathbf{A} = \mathbf{B} + \mathbf{u}\mathbf{u}^*$ with $\|\mathbf{u}\| = 1$. By Theorem 17.5.1 every shift $\lambda_i(\mathbf{A}) - \lambda_i(\mathbf{B})$ is ≥ 0, and by Corollary 17.5.3 the shifts add up to exactly $\|\mathbf{u}\|^2 = 1$. So **every** partial sum of shifts is at most 1, whatever the index set. The eigenvalues of $\mathbf{u}\mathbf{u}^*$ are $1, 0, \dots, 0$, so $1 = \lambda_1 + \cdots + \lambda_k$ of $\mathbf{A} - \mathbf{B}$ for every k. The summed Weyl bound says k, which is wrong by a factor of k.

The bound we would like is therefore

$$\sum_{j=1}^{k} \lambda_{i_j}(\mathbf{A}) - \sum_{j=1}^{k} \lambda_{i_j}(\mathbf{B}) \;\le\; \sum_{j=1}^{k} \lambda_j(\mathbf{A} - \mathbf{B}),$$

and the obstruction is visible in the proof of Weyl's inequality. That proof finds, for each index, **one** good vector. Summing k such bounds uses k unrelated vectors, and the only control over $\sum_j \langle(\mathbf{A} - \mathbf{B})\mathbf{x}_j, \mathbf{x}_j\rangle$ for unrelated unit vectors is $k\,\lambda_1(\mathbf{A} - \mathbf{B})$. If the $\mathbf{x}_j$ were **orthonormal**,

Theorem 17.6.2 would bound the same sum by $\lambda_1 + \cdots + \lambda_k$ of $\mathbf{A} - \mathbf{B}$. So we need a single orthonormal list $\mathbf{x}_1, \dots, \mathbf{x}_k$ that is simultaneously good for $\mathbf{A}$ and good for $\mathbf{B}$:

$$\sum_j \langle \mathbf{A}\mathbf{x}_j, \mathbf{x}_j \rangle \;\geq\; \sum_j \lambda_{i_j}(\mathbf{A}), \qquad \sum_j \langle \mathbf{B}\mathbf{x}_j, \mathbf{x}_j \rangle \;\leq\; \sum_j \lambda_{i_j}(\mathbf{B}).$$

The first requirement is easy to meet: it will hold whenever each $\mathbf{x}_j$ lies in the span W_j of the top i_j eigenvectors of $\mathbf{A}$. Those spans are nested, $W_1 \subset W_2 \subset \cdots \subset W_k$, and from $\mathbf{B}$'s point of view they are an arbitrary nested chain of subspaces. The second requirement is then a question about $\mathbf{B}$ alone: **inside any nested chain of subspaces of these dimensions, is there an orthonormal list whose sum for $\mathbf{B}$ is small?** Wielandt's theorem answers yes, and the chain needs a name.

17.7.2 Flags and adapted lists

Chapter 9 §07 called a chain $U_1 \subset U_2 \subset \cdots \subset U_n = V$ with $\dim U_k = k$ a **flag**, and a basis whose first k members span U_k a basis adapted to it. We need chains that skip dimensions, and lists that meet each member of the chain once.

A flag of type $(i_1, \dots, i_k)$ is a nested chain of subspaces with prescribed dimensions; an adapted list takes its j-th vector from the j-th subspace and is orthonormal.

Definition 17.7.1. Flags of a given type, and adapted lists

Let $F = \mathbb{R}$ or $\mathbb{C}$, and let $1 \leq i_1 < \cdots < i_k \leq n$ be an index set. A **flag of type** $(i_1, \dots, i_k)$ in F^n is a list $\mathcal{W} = (W_1, \dots, W_k)$ of subspaces of F^n with

$$W_1 \subseteq W_2 \subseteq \cdots \subseteq W_k \qquad \text{and} \qquad \dim W_j = i_j \;\text{ for every } j.$$

A list $(\mathbf{x}_1, \dots, \mathbf{x}_k)$ of vectors of F^n is **adapted to** $\mathcal{W}$ if it is **orthonormal** and $\mathbf{x}_j \in W_j$ for **every** j.

In words. The subspaces are nested, and the j-th one has dimension equal to the j-th **index** i_j, not to j; since the indices strictly increase, so do the dimensions, and each inclusion is proper. An adapted list has exactly one vector per subspace; the j-th vector may be taken anywhere in W_j, including in the smaller $W_1, \dots, W_{j-1}$, but the whole list must be orthonormal.

Every flag has an adapted list. Choose a basis of W_1, extend it to a basis of W_2 (Theorem 1.6.5), then to W_3, and so on up to W_k. This gives a basis $\mathbf{v}_1, \dots, \mathbf{v}_{i_k}$ of W_k whose first i_j members span W_j. Applying Theorem 11.2.2 yields an orthonormal list $\mathbf{e}_1, \dots, \mathbf{e}_{i_k}$ with $\operatorname{span}(\mathbf{e}_1, \dots, \mathbf{e}_{i_j}) = W_j$ for each j, and $(\mathbf{e}_{i_1}, \dots, \mathbf{e}_{i_k})$ is adapted to $\mathcal{W}$.

Four cases to keep in mind.

- **One index.** For $k = 1$ a flag is a single subspace W_1 of dimension i_1, and an adapted list is a single unit vector in it.
- **An initial segment.** For the index set $(1, 2, \dots, k)$, W_j has dimension j. An adapted list is then an orthonormal basis of W_k whose first j vectors span W_j, since j orthonormal vectors in the j-dimensional W_j span it. With $k = n$ this is Chapter 9's complete flag.
- **The eigenflag.** Let $(\mathbf{u}_1, \dots, \mathbf{u}_n)$ be an orthonormal basis of F^n with $\mathbf{A}\mathbf{u}_p = \lambda_p(\mathbf{A})\mathbf{u}_p$, and put $W_j = \operatorname{span}(\mathbf{u}_1, \dots, \mathbf{u}_{i_j})$. This is a flag of type $(i_1, \dots, i_k)$, and $(\mathbf{u}_{i_1}, \dots, \mathbf{u}_{i_k})$ is adapted to it, since $\mathbf{u}_{i_j}$ is among the vectors spanning W_j.
- **The degenerate case $k = n$.** Then $i_j = j$ is forced, $W_n = F^n$, and an adapted list is an orthonormal basis of F^n whose first j vectors span W_j.

A minimal change breaks adaptedness. For the eigenflag of type $(1,3)$ in F^3, namely $\operatorname{span}(\mathbf{u}_1) \subset F^3$, the list $(\mathbf{u}_1, \mathbf{u}_3)$ is adapted. Swap the two vectors to get $(\mathbf{u}_3, \mathbf{u}_1)$. It is still orthonormal and its members still lie in $W_2 = F^3$, but the clause $\mathbf{x}_1 \in W_1$ fails: $\mathbf{u}_3 \notin \operatorname{span}(\mathbf{u}_1)$.

Why this definition? Orthonormality is what makes $\sum_j \langle \mathbf{A}\mathbf{x}_j, \mathbf{x}_j\rangle$ the trace of a compression, which is what Theorem 17.6.2 controls. The nesting will let us orthonormalize a list by Gram–Schmidt without moving any vector out of its own subspace, since Gram–Schmidt keeps the j-th vector in the span of the first j.

Quick check. In $\mathbb{R}^3$, take the flag $W_1 = \operatorname{span}(\mathbf{e}_1) \subset W_2 = \operatorname{span}(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$, of type $(1,3)$. Which of these lists are adapted to it: (a) $(\mathbf{e}_1, \mathbf{e}_3)$; (b) $(\mathbf{e}_3, \mathbf{e}_1)$; (c) $(\mathbf{e}_1, \mathbf{e}_1 + \mathbf{e}_2)$; (d) $(-\mathbf{e}_1, \frac{1}{\sqrt{2}}(\mathbf{e}_2 - \mathbf{e}_3))$?

Solution.

(a) Yes: orthonormal, $\mathbf{e}_1 \in W_1$ and $\mathbf{e}_3 \in W_2$. (b) No: $\mathbf{x}_1 = \mathbf{e}_3$ is not in W_1. (c) No: the list is not orthonormal, since $\langle \mathbf{e}_1, \mathbf{e}_1 + \mathbf{e}_2\rangle = 1$ and $\|\mathbf{e}_1 + \mathbf{e}_2\| = \sqrt{2}$. (d) Yes: $-\mathbf{e}_1 \in W_1$, the second vector has length 1 and is orthogonal to $\mathbf{e}_1$, and every vector lies in $W_2 = \mathbb{R}^3$.

17.7.3 Wielandt's minimax principle

The eigenflag guarantees every adapted list a large sum, and no flag of the same type can guarantee more. That is the statement; the second half is the hard one.

Theorem 17.7.1. Wielandt's Minimax Principle

Let $\mathbf{A} \in M_n(F)$ be Hermitian, with $F = \mathbb{R}$ or $\mathbb{C}$, and let $1 \le i_1 < \cdots < i_k \le n$.

(a) For **every** flag $\mathcal{W}$ of type $(i_1, \dots, i_k)$ there is a list $(\mathbf{x}_1, \dots, \mathbf{x}_k)$ adapted to $\mathcal{W}$ with

$$\sum_{j=1}^{k} \langle \mathbf{A}\mathbf{x}_j, \mathbf{x}_j\rangle \;\le\; \sum_{j=1}^{k} \lambda_{i_j}(\mathbf{A}).$$

(b) Let $(\mathbf{u}_1, \dots, \mathbf{u}_n)$ be an orthonormal basis with $\mathbf{A}\mathbf{u}_p = \lambda_p(\mathbf{A})\mathbf{u}_p$ for every p, and let $\mathcal{W}^{\mathbf{A}}$ be the flag $W_j = \operatorname{span}(\mathbf{u}_1, \dots, \mathbf{u}_{i_j})$. Then **every** list $(\mathbf{x}_1, \dots, \mathbf{x}_k)$ adapted to $\mathcal{W}^{\mathbf{A}}$ satisfies

$$\sum_{j=1}^{k} \langle \mathbf{A}\mathbf{x}_j, \mathbf{x}_j\rangle \;\ge\; \sum_{j=1}^{k} \lambda_{i_j}(\mathbf{A}),$$

with equality for $\mathbf{x}_j = \mathbf{u}_{i_j}$.

(c) Consequently

$$\sum_{j=1}^{k} \lambda_{i_j}(\mathbf{A}) = \max_{\mathcal{W}} \; \min_{(\mathbf{x}_1, \dots, \mathbf{x}_k)} \; \sum_{j=1}^{k} \langle \mathbf{A}\mathbf{x}_j, \mathbf{x}_j\rangle,$$

where $\mathcal{W}$ runs over all flags of type $(i_1, \dots, i_k)$ in F^n, and for each $\mathcal{W}$ the list runs over all lists adapted to $\mathcal{W}$. The inner minimum exists for every $\mathcal{W}$, and the outer maximum is attained at $\mathcal{W}^{\mathbf{A}}$.

Idea. Two inequalities of different characters, as in Theorem 17.2.3. In both, the sum $\sum_j \langle \mathbf{A}\mathbf{x}_j, \mathbf{x}_j\rangle$ is the trace of the compression $\mathbf{B}$ of $\mathbf{A}$ to the span of the list, so it is $\sum_j \lambda_j(\mathbf{B})$. Both parts therefore come down to comparing $\lambda_j(\mathbf{B})$ with $\lambda_{i_j}(\mathbf{A})$, one j at a time, by Courant–Fischer applied to the small matrix $\mathbf{B}$.

For (b), the first j vectors of the list lie in $\operatorname{span}(\mathbf{u}_1, \dots, \mathbf{u}_{i_j})$, where $R_{\mathbf{A}} \geq \lambda_{i_j}(\mathbf{A})$, so $\lambda_j(\mathbf{B}) \geq \lambda_{i_j}(\mathbf{A})$.

For (a), we need a list whose span X meets each bottom span $L_{i_j} = \operatorname{span}(\mathbf{u}_{i_j}, \dots, \mathbf{u}_n)$ in dimension at least $k - j + 1$, since $R_{\mathbf{A}} \leq \lambda_{i_j}(\mathbf{A})$ there. For a **single** j, Lemma 17.2.1 gives such a meeting. For all j at once, with one X, it does not: the vectors chosen for different j can collide, and the example after the proof shows that they do. The remedy is to count in coordinates. Record, for each vector, the first eigenvector coordinate at which it is nonzero; vectors with different first coordinates are independent; and a greedy choice of these first coordinates achieves every count at once.

Step roadmap. A preliminary claim turns bounds on $R_{\mathbf{A}}$ over subspaces into bounds on the eigenvalues of a compression. **Step 1** proves (b). **Step 2** introduces leading indices and makes the greedy selection. **Step 3** orthonormalizes the selected vectors and proves (a). **Step 4** proves that the inner minimum exists, which is the only place analysis enters, and deduces (c).

Proof. Let $(\mathbf{u}_1, \dots, \mathbf{u}_n)$ be an orthonormal basis of F^n with $\mathbf{A}\mathbf{u}_p = \lambda_p(\mathbf{A})\mathbf{u}_p$ for every p; such a basis exists by Corollary 12.4.3 over $\mathbb{C}$ or Corollary 12.5.2 over $\mathbb{R}$. For part (b) take the basis named there; for part (a) any such basis will do. Abbreviate $\lambda_p = \lambda_p(\mathbf{A})$. For $1 \leq p \leq n$ put

$$L_p = \operatorname{span}(\mathbf{u}_p, \mathbf{u}_{p+1}, \dots, \mathbf{u}_n), \qquad L_{n+1} = \{\mathbf{0}\}.$$

By Lemma 17.3.1, $R_{\mathbf{A}}(\mathbf{x}) \leq \lambda_p$ for every nonzero $\mathbf{x} \in L_p$, and $R_{\mathbf{A}}(\mathbf{x}) \geq \lambda_p$ for every nonzero $\mathbf{x} \in \operatorname{span}(\mathbf{u}_1, \dots, \mathbf{u}_p)$.

Claim. **Claim 1 (compression bounds).** Let $(\mathbf{y}_1, \dots, \mathbf{y}_k)$ be orthonormal, let Y be its span, let $\mathbf{Q} \in M_{n\times k}(F)$ have columns $\mathbf{y}_1, \dots, \mathbf{y}_k$, and let $\mathbf{B} = \mathbf{Q}^*\mathbf{A}\mathbf{Q}$ be the compression of Definition 17.6.1. Let $1 \leq m \leq k$ and $c \in \mathbb{R}$.

(i) If $S \subseteq Y$ is a subspace with $\dim S = m$ and $R_{\mathbf{A}}(\mathbf{x}) \geq c$ for every nonzero $\mathbf{x} \in S$, then $\lambda_m(\mathbf{B}) \geq c$.

(ii) If $S \subseteq Y$ is a subspace with $\dim S = k - m + 1$ and $R_{\mathbf{A}}(\mathbf{x}) \leq c$ for every nonzero $\mathbf{x} \in S$, then $\lambda_m(\mathbf{B}) \leq c$.

Moreover $\sum_{j=1}^{k} \langle \mathbf{A}\mathbf{y}_j, \mathbf{y}_j\rangle = \operatorname{tr}\mathbf{B} = \sum_{j=1}^{k} \lambda_j(\mathbf{B})$.

Proof. The matrix $\mathbf{B}$ is Hermitian, and the map $\mathbf{c} \mapsto \mathbf{Q}\mathbf{c}$ is a linear bijection $F^k \to Y$ with $R_{\mathbf{B}}(\mathbf{c}) = R_{\mathbf{A}}(\mathbf{Q}\mathbf{c})$ for $\mathbf{c} \neq \mathbf{0}$, as computed in the proof of Lemma 17.2.2. Let $S' = \{\mathbf{c} \in F^k : \mathbf{Q}\mathbf{c} \in S\}$; it is a subspace, and $\mathbf{c} \mapsto \mathbf{Q}\mathbf{c}$ maps it bijectively onto S, so $\dim S' = \dim S$.

(i) By the max–min line of Theorem 17.2.3 for $\mathbf{B} \in M_k(F)$, $\lambda_m(\mathbf{B})$ is the largest of the numbers $\min_{\mathbf{0}\neq\mathbf{c}\in T} R_{\mathbf{B}}(\mathbf{c})$ over subspaces $T \subseteq F^k$ of dimension m. Taking $T = S'$,

$$\lambda_m(\mathbf{B}) \;\geq\; \min_{\mathbf{0}\neq\mathbf{c}\in S'} R_{\mathbf{B}}(\mathbf{c}) = \min_{\mathbf{0}\neq\mathbf{x}\in S} R_{\mathbf{A}}(\mathbf{x}) \;\geq\; c.$$

(ii) By the min–max line of Theorem 17.2.3 for $\mathbf{B}$, with $T = S'$ of dimension $k - m + 1$,

$$\lambda_m(\mathbf{B}) \;\le\; \max_{\mathbf{0}\neq\mathbf{c}\in S'} R_{\mathbf{B}}(\mathbf{c}) = \max_{\mathbf{0}\neq\mathbf{x}\in S} R_{\mathbf{A}}(\mathbf{x}) \;\le\; c.$$

Finally, the diagonal entries of $\mathbf{B}$ are $\langle \mathbf{A}\mathbf{y}_j, \mathbf{y}_j\rangle$ by Definition 17.6.1, and the trace of the Hermitian $\mathbf{B}$ is the sum of its eigenvalues by Corollary 12.1.4.

■

Step 1. The eigenflag guarantees the sum: part (b). Let $(\mathbf{x}_1, \dots, \mathbf{x}_k)$ be adapted to $\mathcal{W}^{\mathbf{A}}$, so $\mathbf{x}_j \in W_j = \operatorname{span}(\mathbf{u}_1, \dots, \mathbf{u}_{i_j})$. Fix j and put $S = \operatorname{span}(\mathbf{x}_1, \dots, \mathbf{x}_j)$. The list is orthonormal, hence independent, so $\dim S = j$. Each $\mathbf{x}_m$ with $m \le j$ lies in $W_m \subseteq W_j$, so $S \subseteq W_j$, where $R_{\mathbf{A}} \ge \lambda_{i_j}$. By Claim 1 (i) with $m = j$, $\lambda_j(\mathbf{B}) \ge \lambda_{i_j}$, where $\mathbf{B}$ is the compression to the span of the list. Summing over j and using the last line of Claim 1,

$$\sum_{j=1}^{k} \langle \mathbf{A}\mathbf{x}_j, \mathbf{x}_j\rangle = \sum_{j=1}^{k} \lambda_j(\mathbf{B}) \;\ge\; \sum_{j=1}^{k} \lambda_{i_j}.$$

For $\mathbf{x}_j = \mathbf{u}_{i_j}$ each term is $\langle \mathbf{A}\mathbf{u}_{i_j}, \mathbf{u}_{i_j}\rangle = \lambda_{i_j}$, so equality holds. This proves (b).

Step 2. Leading indices, and a selection. Let $\mathcal{W} = (W_1, \dots, W_k)$ be an arbitrary flag of type $(i_1, \dots, i_k)$. For $\mathbf{x} \in F^n$ write $c_p(\mathbf{x}) = \langle \mathbf{x}, \mathbf{u}_p\rangle$, so that $\mathbf{x} = \sum_p c_p(\mathbf{x})\mathbf{u}_p$ by Theorem 11.2.1 (a). Then $\mathbf{x} \in L_p$ exactly when $c_1(\mathbf{x}) = \cdots = c_{p-1}(\mathbf{x}) = 0$. The **leading index** $\ell(\mathbf{x})$ of a nonzero $\mathbf{x}$ is the smallest p with $c_p(\mathbf{x}) \neq 0$, so that

$$\mathbf{x} \in L_p \iff \ell(\mathbf{x}) \ge p \qquad (\mathbf{x} \neq \mathbf{0}).$$

For a subspace $V \subseteq F^n$, let $J(V)$ be the set of $p \in \{1, \dots, n\}$ with $V \cap L_{p+1} \neq V \cap L_p$. These are the indices at which the chain $V = V \cap L_1 \supseteq V \cap L_2 \supseteq \cdots \supseteq V \cap L_{n+1} = \{\mathbf{0}\}$ gets strictly smaller.

Claim. **Claim 2 (the jumps count the dimension).** For every subspace $V \subseteq F^n$ and every $1 \le p \le n + 1$,

$$\dim(V \cap L_p) = \bigl|J(V) \cap \{p, p+1, \dots, n\}\bigr|.$$

In particular $|J(V)| = \dim V$. Moreover, for each $p \in J(V)$ there is $\mathbf{x} \in V$ with $\ell(\mathbf{x}) = p$.

Proof. Fix $p \le n$. The map $\varphi_p \colon V \cap L_p \to F$, $\mathbf{x} \mapsto c_p(\mathbf{x})$, is linear. A vector of $V \cap L_p$ already has $c_1 = \cdots = c_{p-1} = 0$, so it lies in L_{p+1} exactly when also $c_p = 0$; hence $\ker \varphi_p = V \cap L_{p+1}$. By Theorem 2.3.1,

$$\dim(V \cap L_p) - \dim(V \cap L_{p+1}) = \operatorname{rank} \varphi_p,$$

and $\operatorname{rank} \varphi_p \le \dim F = 1$. The rank is 1 exactly when $\ker \varphi_p \neq V \cap L_p$, that is, when $p \in J(V)$. Adding these equations for $p, p+1, \dots, n$, the left side telescopes to $\dim(V \cap L_p) - \dim(V \cap L_{n+1}) = \dim(V \cap L_p)$, and the right side counts the elements of $J(V)$ that are $\ge p$. This proves the formula, which for $p = n + 1$ reads $0 = 0$. At $p = 1$, $L_1 = F^n$ gives $|J(V)| = \dim V$. Finally, if $p \in J(V)$, pick $\mathbf{x} \in (V \cap L_p) \setminus L_{p+1}$; then $c_1(\mathbf{x}) = \cdots = c_{p-1}(\mathbf{x}) = 0 \neq c_p(\mathbf{x})$, so $\ell(\mathbf{x}) = p$.

■

Claim. **Claim 3 (distinct leading indices force independence).** Nonzero vectors $\mathbf{x}_1, \dots, \mathbf{x}_r$ whose leading indices are pairwise distinct are linearly independent.

Proof. Suppose $\sum_m a_m \mathbf{x}_m = \mathbf{0}$ with not all a_m zero. Among the m with $a_m \neq 0$, let m_0 be the one with the smallest leading index, $q = \ell(\mathbf{x}_{m_0})$. Apply c_q. Every other $\mathbf{x}_m$ with $a_m \neq 0$ has $\ell(\mathbf{x}_m) > q$, so $c_q(\mathbf{x}_m) = 0$. Hence $0 = a_{m_0} c_q(\mathbf{x}_{m_0})$, a product of two nonzero numbers, which is a contradiction.

■

Claim. **Claim 4 (a greedy selection).** There are nonzero vectors $\mathbf{x}_1, \dots, \mathbf{x}_k$ with $\mathbf{x}_m \in W_m$, with pairwise distinct leading indices $q_m = \ell(\mathbf{x}_m)$, and such that for every j,

$$\#\{\, m : q_m \geq i_j \,\} \;\geq\; k - j + 1.$$

Proof. Choose $q_1, \dots, q_k$ in turn: q_m is the **largest** element of $J(W_m)$ not among $q_1, \dots, q_{m-1}$. This is possible, because $J(W_m)$ has $\dim W_m = i_m \geq m$ elements by Claim 2, so at most $m-1$ of them are used. By the last part of Claim 2, pick $\mathbf{x}_m \in W_m$ with $\ell(\mathbf{x}_m) = q_m$. The q_m are distinct by construction.

Fix j, and for $m \geq j-1$ let N_m be the number of $l \leq m$ with $q_l \geq i_j$. We show $N_m \geq m-j+1$ by induction on m. For $m = j-1$ this says $N_{j-1} \geq 0$, which is true. Let $m \geq j$ and assume $N_{m-1} \geq m-j$.

Case 1: $q_m \geq i_j$. Then $N_m = N_{m-1} + 1 \geq m - j + 1$.

Case 2: $q_m < i_j$. By Claim 2, $J(W_m)$ has i_m elements, of which at most $i_j - 1$ are smaller than i_j. So at least $i_m - i_j + 1$ of them are $\geq i_j$, and $i_m - i_j \geq m - j$ because the indices are strictly increasing integers. Every element of $J(W_m)$ that is $\geq i_j$ exceeds q_m, so by the choice of q_m as the largest unused element, each of them is already among $q_1, \dots, q_{m-1}$. These are at least $i_m - i_j + 1 \geq m - j + 1$ of the earlier q_l, all $\geq i_j$, so $N_{m-1} \geq m - j + 1$. Hence $N_m = N_{m-1} \geq m - j + 1$.

At $m = k$ this gives $N_k \geq k - j + 1$, which is the claim.

■

Step 3. An adapted list with a small sum: part (a). Take $\mathbf{x}_1, \dots, \mathbf{x}_k$ from Claim 4. They are linearly independent by Claim 3, so Theorem 11.2.2 applies and produces an orthonormal list $(\mathbf{y}_1, \dots, \mathbf{y}_k)$ with $\operatorname{span}(\mathbf{y}_1, \dots, \mathbf{y}_j) = \operatorname{span}(\mathbf{x}_1, \dots, \mathbf{x}_j)$ for each j. Since $\mathbf{x}_m \in W_m \subseteq W_j$ for $m \leq j$, the vector $\mathbf{y}_j$ lies in W_j. So $(\mathbf{y}_1, \dots, \mathbf{y}_k)$ is adapted to $\mathcal{W}$.

Let $X = \operatorname{span}(\mathbf{y}_1, \dots, \mathbf{y}_k) = \operatorname{span}(\mathbf{x}_1, \dots, \mathbf{x}_k)$, and let $\mathbf{B}$ be the compression of $\mathbf{A}$ to it with respect to the $\mathbf{y}_j$. Fix j. By Claim 4 at least $k - j + 1$ of the $\mathbf{x}_m$ have $q_m \geq i_j$; each of them lies in L_{i_j}, and any $k - j + 1$ of them are independent by Claim 3. Let S be the span of $k - j + 1$ of them, so $S \subseteq X \cap L_{i_j}$ and $\dim S = k - j + 1$. On $S \subseteq L_{i_j}$ we have $R_{\mathbf{A}} \leq \lambda_{i_j}$, so Claim 1 (ii) with $m = j$ gives $\lambda_j(\mathbf{B}) \leq \lambda_{i_j}$. Summing over j and using the last line of Claim 1,

$$\sum_{j=1}^{k} \langle \mathbf{A}\mathbf{y}_j, \mathbf{y}_j \rangle = \sum_{j=1}^{k} \lambda_j(\mathbf{B}) \;\leq\; \sum_{j=1}^{k} \lambda_{i_j}.$$

This proves (a).

Step 4. The minimum exists, and the max–min formula: part (c). Fix a flag $\mathcal{W}$. Identify a list $(\mathbf{x}_1, \dots, \mathbf{x}_k)$ with the column of F^{nk} obtained by stacking the $\mathbf{x}_j$; its Euclidean norm is $(\sum_j \|\mathbf{x}_j\|^2)^{1/2}$. Let $\mathcal{O}$ be the set of lists adapted to $\mathcal{W}$. It is nonempty, by the paragraph after Definition 17.7.1. It is bounded, since every element has norm $\sqrt{k}$. And it

contains the limit of each of its convergent sequences. Indeed, the conditions $\langle \mathbf{x}_i, \mathbf{x}_j \rangle = \delta_{ij}$ are equations between sums of products of the entries and their conjugates, which pass to limits. Each condition $\mathbf{x}_j \in W_j$ says $\mathbf{M}_j \mathbf{x}_j = \mathbf{0}$ for a matrix $\mathbf{M}_j$ whose null space is W_j, for instance the matrix whose rows are the conjugate transposes of a basis of $W_j^{\perp}$ (the zero matrix if $W_j = F^n$), and this passes to limits too. So by fact (A3) of Chapter 16's introduction, every sequence in $\mathcal{O}$ has a subsequence converging to a point of $\mathcal{O}$. The function $(\mathbf{x}_1, \dots, \mathbf{x}_k) \mapsto \sum_j \langle \mathbf{A}\mathbf{x}_j, \mathbf{x}_j \rangle$ is built from the real and imaginary parts of the entries by sums and products, so it is continuous, and it is real-valued since $\mathbf{A}$ is Hermitian. By fact (A4) it attains a minimum on $\mathcal{O}$.

Call that minimum $v(\mathcal{W})$. By (a), some adapted list has sum $\le \sum_j \lambda_{i_j}$, so $v(\mathcal{W}) \le \sum_j \lambda_{i_j}$ for every $\mathcal{W}$. By (b), every list adapted to $\mathcal{W}^{\mathbf{A}}$ has sum $\ge \sum_j \lambda_{i_j}$, with equality at $(\mathbf{u}_{i_1}, \dots, \mathbf{u}_{i_k})$, so $v(\mathcal{W}^{\mathbf{A}}) = \sum_j \lambda_{i_j}$. Hence the largest value of $v(\mathcal{W})$ over all flags of the type exists, is attained at $\mathcal{W}^{\mathbf{A}}$, and equals $\sum_j \lambda_{i_j}$. This proves (c), and the theorem.

■

Remark. **Where the analysis went.** Parts (a) and (b) are pure algebra: a dimension count, Gram–Schmidt, and Courant–Fischer for a small matrix. The only analytic input is in Step 4, facts (A3) and (A4) of Chapter 16's introduction. This is the first time in the chapter that an item of that list is used rather than merely mentioned, and it is used only so that the word "min" in (c) refers to a value that is attained. Every later *result* in this section uses (a) and (b) alone, so no later result depends on analysis; only the exercises that quote (c), such as A1(b), lean on the attained minimum.

Two index sets give back what we already have. For $k = 1$ a flag is a single subspace W of dimension i_1, an adapted list is a unit vector $\mathbf{x} \in W$, and $\langle \mathbf{A}\mathbf{x}, \mathbf{x} \rangle = R_{\mathbf{A}}(\mathbf{x})$. So (c) reads $\lambda_{i_1}(\mathbf{A}) = \max_{\dim W = i_1} \min_{\mathbf{0} \neq \mathbf{x} \in W} R_{\mathbf{A}}(\mathbf{x})$, the max–min line of Theorem 17.2.3. For the initial segment $(1, \dots, k)$, an adapted list is an orthonormal basis of the k-dimensional W_k, and conversely every orthonormal basis $(\mathbf{x}_1, \dots, \mathbf{x}_k)$ of a k-dimensional subspace is adapted to the flag $\operatorname{span}(\mathbf{x}_1) \subset \operatorname{span}(\mathbf{x}_1, \mathbf{x}_2) \subset \dots$. The sum $\sum_j \langle \mathbf{A}\mathbf{x}_j, \mathbf{x}_j \rangle$ is the trace of the compression to W_k. That trace does not depend on the orthonormal basis, because §06 showed that a change of orthonormal basis changes the compression by a unitary similarity. So the inner minimum is constant, and (c) says that $\sum_{j \le k} \lambda_j(\mathbf{A})$ is the largest trace of a compression to a k-dimensional subspace, which is Theorem 17.6.2. What is new is everything in between: index sets with gaps, such as $(2, 4)$, and index sets that do not start at 1.

Example. The selection at work, and why it is needed

Let $\mathbf{A} = \operatorname{diag}(4, 2, 1, -1) \in M_4(\mathbb{R})$, take the index set $(2, 4)$, and take the flag

$$W_1 = \operatorname{span}\big((1, 1, 0, 0),\ \mathbf{e}_4\big) \ \subset \ W_2 = \mathbb{R}^4,$$

of type $(2, 4)$. Show that the naive attempt, choosing $\mathbf{x}_j \in W_j \cap L_{i_j}$ for each j separately, cannot produce an adapted list. Then follow the proof of part (a) to produce one with sum at most $\lambda_2(\mathbf{A}) + \lambda_4(\mathbf{A})$.

Solution. Here the eigenvectors are $\mathbf{u}_p = \mathbf{e}_p$, so $L_p = \operatorname{span}(\mathbf{e}_p, \dots, \mathbf{e}_4)$ is the set of vectors whose first $p-1$ entries vanish, and the target is $\lambda_2 + \lambda_4 = 2 + (-1) = 1$.

The naive attempt. A vector $a(1,1,0,0) + b\mathbf{e}_4$ of W_1 lies in L_2 exactly when its first entry a is 0, so $W_1 \cap L_2 = \operatorname{span}(\mathbf{e}_4)$. Also $W_2 \cap L_4 = L_4 = \operatorname{span}(\mathbf{e}_4)$. Both vectors would have to be multiples of $\mathbf{e}_4$, and two such vectors are never orthonormal. Each of the two intersections is nonzero, as Lemma 17.2.1 promises, but the two choices collide.

The selection. The vectors of W_1 have the form $(a, a, 0, b)$. They lie in L_2, L_3, or L_4 exactly when $a = 0$, and in $L_5 = \{\mathbf{0}\}$ only when also $b = 0$. So the dimensions of $W_1 \cap L_p$ for $p = 1, \dots, 5$ are $2, 1, 1, 1, 0$, and the chain drops at $p = 1$ and $p = 4$: $J(W_1) = \{1, 4\}$. For $W_2 = \mathbb{R}^4$ the dimensions are $4, 3, 2, 1, 0$, so $J(W_2) = \{1, 2, 3, 4\}$. The greedy rule takes $q_1 = \max J(W_1) = 4$, with $\mathbf{x}_1 = \mathbf{e}_4 \in W_1$. It then takes $q_2 = \max(J(W_2) \setminus \{4\}) = 3$, with $\mathbf{x}_2 = \mathbf{e}_3$.

The counts of Claim 4. For $j = 1$, $i_1 = 2$, and both $q_1 = 4$ and $q_2 = 3$ are ≥ 2: the count is $2 \geq 2$. For $j = 2$, $i_2 = 4$, and only $q_1 = 4$ is ≥ 4: the count is $1 \geq 1$. Notice that the requirement for $j = 2$ is met by $\mathbf{x}_1$, not by $\mathbf{x}_2$; insisting that $\mathbf{x}_2$ itself lie in L_4 is exactly the naive attempt that failed.

The list and its sum. $(\mathbf{e}_4, \mathbf{e}_3)$ is already orthonormal, so Gram–Schmidt leaves it alone, and it is adapted: $\mathbf{e}_4 \in W_1$, $\mathbf{e}_3 \in W_2$. Its sum is $\langle \mathbf{A}\mathbf{e}_4, \mathbf{e}_4 \rangle + \langle \mathbf{A}\mathbf{e}_3, \mathbf{e}_3 \rangle = -1 + 1 = 0 \leq 1$. The compression to $X = \operatorname{span}(\mathbf{e}_4, \mathbf{e}_3)$ is $\operatorname{diag}(-1, 1)$, with eigenvalues $1 \leq \lambda_2(\mathbf{A}) = 2$ and $-1 \leq \lambda_4(\mathbf{A}) = -1$, as Step 3 predicts one index at a time.

Quick check. Suppose "orthonormal" is weakened to "each $\mathbf{x}_j$ is a unit vector" in the definition of an adapted list. Show that the formula (c) of Theorem 17.7.1 remains true. Why is this weaker version useless for comparing two matrices?

Solution. With the $\mathbf{x}_j$ chosen independently, the inner minimum is $\sum_j \min_{\mathbf{0} \neq \mathbf{x} \in W_j} R_{\mathbf{A}}(\mathbf{x})$, one term per subspace. By Theorem 17.2.3 each term is at most $\lambda_{i_j}(\mathbf{A})$, since $\dim W_j = i_j$, and the eigenflag makes every term equal to $\lambda_{i_j}(\mathbf{A})$ at once. So the max–min is still $\sum_j \lambda_{i_j}(\mathbf{A})$; it is just Courant–Fischer added up. But the lists it produces are not orthonormal, so Theorem 17.6.2 cannot be applied to them. Comparing $\mathbf{A}$ with $\mathbf{B}$ through such lists gives only the summed Weyl bound $k\,\lambda_1(\mathbf{A} - \mathbf{B})$ of the first subsection. The whole content of Wielandt's theorem is that orthonormality costs nothing.

17.7.4 Lidskii's inequality

The plan of the first subsection can now be carried out.

Theorem 17.7.2. Lidskii's Inequality

Let $\mathbf{A}, \mathbf{B} \in M_n(F)$ be Hermitian, with $F = \mathbb{R}$ or $\mathbb{C}$. Then for every index set $1 \leq i_1 < \cdots <$

$i_k \le n$,

$$\sum_{j=1}^{k} \lambda_{i_j}(\mathbf{A}) - \sum_{j=1}^{k} \lambda_{i_j}(\mathbf{B}) \;\le\; \sum_{j=1}^{k} \lambda_j(\mathbf{A}-\mathbf{B}).$$

Idea. Find one orthonormal list that is good for both matrices. Take the eigenflag of $\mathbf{A}$. Part (b) of Theorem 17.7.1 says that **every** list adapted to it is good for $\mathbf{A}$. Part (a), applied to $\mathbf{B}$ and this same flag, produces **one** adapted list that is good for $\mathbf{B}$. Subtract the two sums. What remains is $\sum_j \langle (\mathbf{A}-\mathbf{B})\mathbf{x}_j, \mathbf{x}_j\rangle$ over an orthonormal list, which Theorem 17.6.2 bounds by the top k eigenvalues of $\mathbf{A}-\mathbf{B}$.

Proof. Let $(\mathbf{u}_1, \dots, \mathbf{u}_n)$ be an orthonormal basis of F^n with $\mathbf{A}\mathbf{u}_p = \lambda_p(\mathbf{A})\mathbf{u}_p$, which exists by Corollary 12.4.3 or Corollary 12.5.2, and let $\mathcal{W}^{\mathbf{A}}$ be the flag $W_j = \operatorname{span}(\mathbf{u}_1, \dots, \mathbf{u}_{i_j})$, of type $(i_1, \dots, i_k)$. By Theorem 17.7.1 (a) applied to the Hermitian matrix $\mathbf{B}$ and the flag $\mathcal{W}^{\mathbf{A}}$, there is a list $(\mathbf{x}_1, \dots, \mathbf{x}_k)$ adapted to $\mathcal{W}^{\mathbf{A}}$ with

$$\sum_{j=1}^{k} \langle \mathbf{B}\mathbf{x}_j, \mathbf{x}_j\rangle \;\le\; \sum_{j=1}^{k} \lambda_{i_j}(\mathbf{B}).$$

By Theorem 17.7.1 (b) applied to $\mathbf{A}$, the same list satisfies

$$\sum_{j=1}^{k} \langle \mathbf{A}\mathbf{x}_j, \mathbf{x}_j\rangle \;\ge\; \sum_{j=1}^{k} \lambda_{i_j}(\mathbf{A}).$$

Subtracting the first inequality from the second, and using that the inner product is linear in its first slot,

$$\sum_{j=1}^{k} \lambda_{i_j}(\mathbf{A}) - \sum_{j=1}^{k} \lambda_{i_j}(\mathbf{B}) \;\le\; \sum_{j=1}^{k} \langle (\mathbf{A}-\mathbf{B})\mathbf{x}_j, \mathbf{x}_j\rangle.$$

The matrix $\mathbf{A}-\mathbf{B}$ is Hermitian and $\mathbf{x}_1, \dots, \mathbf{x}_k$ are orthonormal, so Theorem 17.6.2 bounds the right-hand side by $\sum_{j=1}^{k} \lambda_j(\mathbf{A}-\mathbf{B})$. This proves the theorem. ∎

For the rank-one example of the first subsection, the right-hand side is $\lambda_1 + \cdots + \lambda_k$ of $\mathbf{u}\mathbf{u}^*$, which is 1 for every k: the theorem gives exactly the bound that the interlacing argument found there, and it gives it for every Hermitian $\mathbf{A}-\mathbf{B}$, whatever its rank. One feature of the statement deserves a warning.

Warning. **The left side compares the two spectra at the same indices; the right side always uses the top k eigenvalues of $\mathbf{A}-\mathbf{B}$, whatever the index set.** The tempting "index-by-index" version

$$\lambda_i(\mathbf{A}) - \lambda_i(\mathbf{B}) \;\le\; \lambda_i(\mathbf{A}-\mathbf{B})$$

is false. Take $\mathbf{A} = \operatorname{diag}(1,1)$ and $\mathbf{B} = \operatorname{diag}(1,0)$, so $\mathbf{A}-\mathbf{B} = \operatorname{diag}(0,1)$ with $\lambda_1(\mathbf{A}-\mathbf{B}) = 1$ and $\lambda_2(\mathbf{A}-\mathbf{B}) = 0$. At $i = 2$ the left side is $1 - 0 = 1$, which exceeds $\lambda_2(\mathbf{A}-\mathbf{B}) = 0$. Lidskii's bound is $\lambda_1(\mathbf{A}-\mathbf{B}) = 1$, and it is attained.

Example. All seven index sets

Let

$$\mathbf{A} = \begin{pmatrix} 0 & 2 & 0 \\ 2 & 1 & 2 \\ 0 & 2 & 2 \end{pmatrix}, \qquad \mathbf{B} = \operatorname{diag}(0, 2, -1).$$

Find the three spectra, check Theorem 17.7.2 for every index set, and compare with the summed Weyl bound $k\,\lambda_1(\mathbf{A} - \mathbf{B})$.

Solution. The spectra. The vectors $(1, 2, 2)$, $(2, 1, -2)$, $(2, -2, 1)$ are eigenvectors of $\mathbf{A}$: for instance $\mathbf{A}(1, 2, 2) = (4, 2 + 2 + 4, 4 + 4) = 4(1, 2, 2)$, and the other two give the eigenvalues 1 and -2 in the same way. So $\boldsymbol{\lambda}(\mathbf{A}) = (4, 1, -2)$, writing $\boldsymbol{\lambda}(\mathbf{A})$ for the vector of eigenvalues in decreasing order. Sorting the diagonal of $\mathbf{B}$ gives $\boldsymbol{\lambda}(\mathbf{B}) = (2, 0, -1)$. Next,

$$\mathbf{A} - \mathbf{B} = \begin{pmatrix} 0 & 2 & 0 \\ 2 & -1 & 2 \\ 0 & 2 & 3 \end{pmatrix}$$

has eigenvectors $(1, 2, 4)$, $(-2, -1, 1)$, $(2, -3, 1)$. For example $(\mathbf{A} - \mathbf{B})(1, 2, 4) = (4,\ 2 - 2 + 8,\ 4 + 12) = 4(1, 2, 4)$. So $\boldsymbol{\lambda}(\mathbf{A} - \mathbf{B}) = (4, 1, -3)$, whose running totals from the top are $4, 5, 2$.

The inequalities. The differences at equal indices are $\lambda_i(\mathbf{A}) - \lambda_i(\mathbf{B}) = 2, 1, -1$ for $i = 1, 2, 3$.

index set	left side	Lidskii bound	summed Weyl bound
(1), (2), (3)	$2,\ 1,\ -1$	4	4
$(1,2)$, $(1,3)$, $(2,3)$	$3,\ 1,\ 0$	5	8
$(1,2,3)$	2	2	12

Every inequality holds. For $k = 3$ it is an equality, and must be: both sides are then $\operatorname{tr}\mathbf{A} - \operatorname{tr}\mathbf{B} = \operatorname{tr}(\mathbf{A} - \mathbf{B})$, by Corollary 12.1.4 and Theorem 0.9.6 (1). The summed Weyl bound grows like $k\,\lambda_1(\mathbf{A} - \mathbf{B})$, while Lidskii's bound can even decrease, because the eigenvalue -3 of $\mathbf{A} - \mathbf{B}$ enters the sum as soon as $k = 3$.

The index sets that give the results we already have are the two families from the discussion of Wielandt's theorem, now at the level of two matrices.

Corollary 17.7.3. Weyl and Ky Fan as Cases of Lidskii

Let $\mathbf{A}, \mathbf{B} \in M_n(F)$ be Hermitian.

(a) **(One index.)** For every i,

$$\lambda_n(\mathbf{A} - \mathbf{B}) \;\le\; \lambda_i(\mathbf{A}) - \lambda_i(\mathbf{B}) \;\le\; \lambda_1(\mathbf{A} - \mathbf{B}),$$

and in particular $|\lambda_i(\mathbf{A}) - \lambda_i(\mathbf{B})| \le \|\mathbf{A} - \mathbf{B}\|_2$.

(b) **(An initial segment.)** For every k,

$$\sum_{i=1}^{k} \lambda_i(\mathbf{A}) - \sum_{i=1}^{k} \lambda_i(\mathbf{B}) \;\le\; \sum_{i=1}^{k} \lambda_i(\mathbf{A}-\mathbf{B}).$$

Idea. Both are Theorem 17.7.2 with a particular index set, the lower bound in (a) coming from the same theorem with $\mathbf{A}$ and $\mathbf{B}$ exchanged.

Proof.

(a) The upper bound is Theorem 17.7.2 with $k = 1$ and the index set (i). Exchanging $\mathbf{A}$ and $\mathbf{B}$ in the same case gives $\lambda_i(\mathbf{B}) - \lambda_i(\mathbf{A}) \le \lambda_1(\mathbf{B}-\mathbf{A})$. By Lemma 17.3.2 with $p = 1$, $\lambda_1(\mathbf{B}-\mathbf{A}) = \lambda_1(-(\mathbf{A}-\mathbf{B})) = -\lambda_n(\mathbf{A}-\mathbf{B})$, and multiplying by -1 gives the lower bound. So $\lambda_i(\mathbf{A})-\lambda_i(\mathbf{B})$ lies between $\lambda_n(\mathbf{A}-\mathbf{B})$ and $\lambda_1(\mathbf{A}-\mathbf{B})$, and its absolute value is at most $\max(\lambda_1(\mathbf{A}-\mathbf{B}), -\lambda_n(\mathbf{A}-\mathbf{B}))$, which is $\|\mathbf{A}-\mathbf{B}\|_2$ by Lemma 17.3.5.

(b) This is Theorem 17.7.2 with the index set $(1, 2, \dots, k)$.

■

Read with $\mathbf{A} = \mathbf{X}+\mathbf{Y}$ and $\mathbf{B} = \mathbf{X}$, part (a) is Corollary 17.3.4 for the pair $\mathbf{X}, \mathbf{Y}$; read with $\mathbf{B} = \mathbf{Y}$ instead, it is the same corollary with $\mathbf{X}$ and $\mathbf{Y}$ exchanged. Together these are the cases of Theorem 17.3.3 in which one of the two indices on the right is 1 (in part (a) of that theorem) or n (in its part (b)); applied to negated pairs, the one-index case gives a few more, as the remark below explains. The last clause of part (a) is Corollary 17.3.6. Part (b), read the same way, is Corollary 17.6.5. Here is the hierarchy, one row per index set.

index set	Theorem 17.7.1 gives	Theorem 17.7.2 gives
(i), one index	Theorem 17.2.3, max–min line	Theorem 17.3.3 with one index 1 or n (and, through negation, with $i+j-1=n$); Corollary 17.3.6
$(1, \dots, k)$	Theorem 17.6.2	Corollary 17.6.5
any other, such as $(2, 4)$	new	new

Remark. **Not all of Weyl's inequalities are cases of Lidskii's.** To test this fairly, Theorem 17.7.2 must be allowed every pair it can see: all pairs of matrices among $\pm\mathbf{X}$, $\pm\mathbf{Y}$, $\pm(\mathbf{X}+\mathbf{Y})$ whose difference is another of the six. Negation matters. With it, the one-index case already reaches some Weyl inequalities with both indices inside: applied to $\mathbf{A} = -\mathbf{X}$, $\mathbf{B} = \mathbf{Y}$, whose difference is $-(\mathbf{X}+\mathbf{Y})$, and read through Lemma 17.3.2, it gives $\lambda_n(\mathbf{X}+\mathbf{Y}) \le \lambda_{n+1-i}(\mathbf{X}) + \lambda_i(\mathbf{Y})$ for every i, which is exactly the Weyl inequality (a) with $i+j-1=n$. So a small example that ignores negation can look like a counterexample and not be one.

Even with every signed pair allowed, the family does not contain all of Weyl's. Take $n = 4$

and the candidate spectra

$$\boldsymbol{\lambda}(\mathbf{X}) = (2,0,0,-1), \qquad \boldsymbol{\lambda}(\mathbf{Y}) = (2,0,0,0),$$
$$\boldsymbol{\lambda}(\mathbf{X}+\mathbf{Y}) = (2,1,1,-1).$$

The totals agree, $1+2=3$, and a direct check of every index set shows that all twelve signed pairs satisfy every inequality of Theorem 17.7.2. Yet Weyl's inequality with $i=j=2$ demands $\lambda_3(\mathbf{X}+\mathbf{Y}) \le \lambda_2(\mathbf{X}) + \lambda_2(\mathbf{Y}) = 0$, and here $\lambda_3(\mathbf{X}+\mathbf{Y}) = 1$. Since Theorem 17.3.3 is a theorem, no Hermitian $\mathbf{X}$, $\mathbf{Y}$ have these spectra: the lists are ruled out by Weyl and not by Lidskii.

Remark. **The majorization form.** Section 8 introduces the language that turns this family of inequalities into a single statement. We record the translation here, together with exactly what has been proved. For a real vector $\mathbf{x} \in \mathbb{R}^n$, the sum of its k largest entries is the largest value of $\sum_{i\in S} x_i$ over the sets S of k positions: sorting the entries shows that the k largest do best. Apply this to $\mathbf{x} = \boldsymbol{\lambda}(\mathbf{A}) - \boldsymbol{\lambda}(\mathbf{B})$, whose i-th entry is $\lambda_i(\mathbf{A}) - \lambda_i(\mathbf{B})$ and which need not be decreasing. Then Theorem 17.7.2, over all index sets of size k, says that the k largest entries of $\boldsymbol{\lambda}(\mathbf{A}) - \boldsymbol{\lambda}(\mathbf{B})$ add up to at most $\sum_{j\le k} \lambda_j(\mathbf{A}-\mathbf{B})$, for every k. The two vectors also have the same total, $\operatorname{tr}\mathbf{A} - \operatorname{tr}\mathbf{B} = \operatorname{tr}(\mathbf{A}-\mathbf{B})$. In the notation of Section 8 this is

$$\boldsymbol{\lambda}(\mathbf{A}) - \boldsymbol{\lambda}(\mathbf{B}) \;\prec\; \boldsymbol{\lambda}(\mathbf{A}-\mathbf{B}).$$

Proved here: every inequality in this majorization, namely Theorem 17.7.2 for every index set, and the equality of totals, which is the trace. **Not proved here, and left to Chapter 21:** the general consequences of a majorization $\mathbf{x} \prec \mathbf{y}$, namely $\sum_i \phi(x_i) \le \sum_i \phi(y_i)$ for every convex $\phi\colon \mathbb{R} \to \mathbb{R}$, together with the matching statement for the unitarily invariant norms of that chapter. Applied to the display above, they bound $\sum_i \phi(\lambda_i(\mathbf{A}) - \lambda_i(\mathbf{B}))$ by $\sum_i \phi(\lambda_i(\mathbf{A}-\mathbf{B}))$ for every convex ϕ. One case, $\phi(t) = |t|$, follows from this section directly, and exercise C2 proves it. Nothing in this chapter uses the general case.

17.7.5 What the flags bought

Courant–Fischer turned each eigenvalue into an optimization that two matrices can share, and the comparison theorems of §§02–05 followed one index at a time. Ky Fan did the same for the top partial sums. Wielandt's theorem does it for every sum, at the price of optimizing over flags instead of subspaces, and Lidskii's inequality is what the shared optimization yields: the eigenvalue shifts $\lambda_i(\mathbf{A}) - \lambda_i(\mathbf{B})$, summed over **any** k indices, never exceed the top k eigenvalues of $\mathbf{A}-\mathbf{B}$. The single idea behind it is the one the first subsection isolated. A comparison of sums needs one orthonormal list that serves both matrices, and a flag is exactly the structure that lets such a list be found by counting.

17.7.6 Exercises

17.7.6.1 A. Check your understanding

Exercise. A1

(a) Define a flag of type $(i_1, \dots, i_k)$ in F^n and a list adapted to it.

(b) State the max–min formula of Theorem 17.7.1, and name the index sets for which it reduces to Theorem 17.2.3 and to Theorem 17.6.2.

(c) Determine whether the following is true, and justify your answer: if $(\mathbf{x}_1, \dots, \mathbf{x}_k)$ is adapted to a flag $(W_1, \dots, W_k)$, then $\operatorname{span}(\mathbf{x}_1, \dots, \mathbf{x}_j) = W_j$ for every j.

(d) Determine whether the following is true for all Hermitian $\mathbf{A}, \mathbf{B} \in M_n(F)$, and justify your answer: $\lambda_2(\mathbf{A}) - \lambda_2(\mathbf{B}) \le \lambda_2(\mathbf{A} - \mathbf{B})$.

(e) In the proof of Theorem 17.7.2, which part of Theorem 17.7.1 is applied to $\mathbf{A}$, which to $\mathbf{B}$, and with which flag?

Solution.

(a) For an index set $1 \le i_1 < \cdots < i_k \le n$, a flag of that type is a list $(W_1, \dots, W_k)$ of subspaces of F^n with $W_1 \subseteq \cdots \subseteq W_k$ and $\dim W_j = i_j$. A list $(\mathbf{x}_1, \dots, \mathbf{x}_k)$ is adapted to it if it is orthonormal and $\mathbf{x}_j \in W_j$ for every j.

(b) $\sum_j \lambda_{i_j}(\mathbf{A})$ is the maximum over flags $\mathcal{W}$ of type $(i_1, \dots, i_k)$ of the minimum of $\sum_j \langle \mathbf{A}\mathbf{x}_j, \mathbf{x}_j \rangle$ over lists adapted to $\mathcal{W}$. For $k = 1$, with index set (i), it is the max–min line of Courant–Fischer for λ_i. For the initial segment $(1, \dots, k)$ it is Ky Fan's theorem for $\sum_{j \le k} \lambda_j$.

(c) False in general. The span of j orthonormal vectors has dimension j, while $\dim W_j = i_j$, which exceeds j unless $i_j = j$. For instance, with $k = 1$ and the flag $W_1 = \mathbb{R}^2$ of type (2), the adapted list $(\mathbf{e}_1)$ spans a line. What is true is $\operatorname{span}(\mathbf{x}_1, \dots, \mathbf{x}_j) \subseteq W_j$.

(d) False. With $\mathbf{A} = \operatorname{diag}(1, 1)$ and $\mathbf{B} = \operatorname{diag}(1, 0)$, the left side is $1 - 0 = 1$ and $\lambda_2(\mathbf{A} - \mathbf{B}) = \lambda_2(\operatorname{diag}(0, 1)) = 0$. The correct bound, from Theorem 17.7.2 with $k = 1$, is $\lambda_1(\mathbf{A} - \mathbf{B})$.

(e) The flag is the eigenflag $\mathcal{W}^{\mathbf{A}}$ of $\mathbf{A}$ for the given index set. Part (b) is applied to $\mathbf{A}$: every list adapted to $\mathcal{W}^{\mathbf{A}}$ has $\mathbf{A}$-sum at least $\sum_j \lambda_{i_j}(\mathbf{A})$. Part (a) is applied to $\mathbf{B}$ with the same flag: some adapted list has $\mathbf{B}$-sum at most $\sum_j \lambda_{i_j}(\mathbf{B})$.

17.7.6.2 B. Practice

Exercise. B1: Running the selection

Let $\mathbf{A} = \operatorname{diag}(4, 1, -1) \in M_3(\mathbb{R})$, take the index set $(1, 3)$, and let $\mathcal{W}$ be the flag $W_1 = \operatorname{span}((1, 1, 1)) \subset W_2 = \mathbb{R}^3$.

(a) Compute $\lambda_1(\mathbf{A}) + \lambda_3(\mathbf{A})$, and show directly that every list adapted to the eigenflag $\operatorname{span}(\mathbf{e}_1) \subset \mathbb{R}^3$ has sum at least this number.

(b) Follow the proof of Theorem 17.7.1 (a) for $\mathcal{W}$: find $J(W_1)$ and $J(W_2)$, select $\mathbf{x}_1, \mathbf{x}_2$, apply Gram–Schmidt, and compute $\sum_j \langle \mathbf{A}\mathbf{y}_j, \mathbf{y}_j \rangle$.

(c) Compute the compression $\mathbf{B}$ of $\mathbf{A}$ to $\operatorname{span}(\mathbf{y}_1, \mathbf{y}_2)$ and its eigenvalues, and check $\lambda_j(\mathbf{B}) \le \lambda_{i_j}(\mathbf{A})$ for $j = 1, 2$.

(d) Find the minimum of $\sum_j \langle \mathbf{A}\mathbf{x}_j, \mathbf{x}_j \rangle$ over all lists adapted to $\mathcal{W}$. Is it the value found in (b)?

Solution.

(a) The eigenvalues are $4, 1, -1$ with eigenvectors $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$, so $\lambda_1 + \lambda_3 = 3$, and $\operatorname{span}(\mathbf{e}_1) \subset \mathbb{R}^3$ is the eigenflag of type $(1, 3)$. An adapted list has $\mathbf{x}_1 = \pm \mathbf{e}_1$, the unit vectors of the

line, so $\langle \mathbf{A}\mathbf{x}_1, \mathbf{x}_1 \rangle = 4$. Its $\mathbf{x}_2$ is a unit vector orthogonal to $\mathbf{e}_1$, so $\mathbf{x}_2 = (0, b, c)$ with $b^2 + c^2 = 1$, and $\langle \mathbf{A}\mathbf{x}_2, \mathbf{x}_2 \rangle = b^2 - c^2 \geq -1$. The sum is at least $4 - 1 = 3$, with equality at $\mathbf{x}_2 = \pm\mathbf{e}_3$, as part (b) of the theorem says.

(b) Here $L_p = \operatorname{span}(\mathbf{e}_p, \dots, \mathbf{e}_3)$. The line W_1 is spanned by $\mathbf{w} = (1, 1, 1)$, whose first entry is nonzero, so $W_1 \cap L_2 = \{\mathbf{0}\}$ and $J(W_1) = \{1\}$. For $W_2 = \mathbb{R}^3$, $J(W_2) = \{1, 2, 3\}$. The greedy rule gives $q_1 = 1$ with $\mathbf{x}_1 = \mathbf{w}$, and $q_2 = \max\{2, 3\} = 3$ with $\mathbf{x}_2 = \mathbf{e}_3$. The counts of Claim 4 hold: for $j = 1$, both q's are $\geq i_1 = 1$; for $j = 2$, $q_2 = 3 \geq i_2 = 3$. Gram–Schmidt gives $\mathbf{y}_1 = \frac{1}{\sqrt{3}}(1, 1, 1)$ and

$$\mathbf{e}_3 - \langle \mathbf{e}_3, \mathbf{y}_1 \rangle \mathbf{y}_1 = \tfrac{1}{3}(-1, -1, 2), \qquad \mathbf{y}_2 = \tfrac{1}{\sqrt{6}}(-1, -1, 2).$$

Then $\langle \mathbf{A}\mathbf{y}_1, \mathbf{y}_1 \rangle = \frac{1}{3}(4 + 1 - 1) = \frac{4}{3}$ and $\langle \mathbf{A}\mathbf{y}_2, \mathbf{y}_2 \rangle = \frac{1}{6}(4 + 1 - 4) = \frac{1}{6}$, so the sum is $\frac{3}{2} \leq 3$, as the theorem requires.

(c) The off-diagonal entry is $\langle \mathbf{A}\mathbf{y}_2, \mathbf{y}_1 \rangle = \frac{1}{\sqrt{18}}(-4 - 1 - 2) = -\frac{7}{3\sqrt{2}}$, so

$$\mathbf{B} = \begin{pmatrix} 4/3 & -7/(3\sqrt{2}) \\ -7/(3\sqrt{2}) & 1/6 \end{pmatrix}, \qquad \operatorname{tr} \mathbf{B} = \tfrac{3}{2}, \quad \det \mathbf{B} = \tfrac{2}{9} - \tfrac{49}{18} = -\tfrac{5}{2}.$$

The characteristic polynomial is $t^2 - \frac{3}{2}t - \frac{5}{2}$, that is $\frac{1}{2}(2t - 5)(t + 1)$, so $\lambda_1(\mathbf{B}) = \frac{5}{2} \leq 4 = \lambda_1(\mathbf{A})$ and $\lambda_2(\mathbf{B}) = -1 \leq -1 = \lambda_3(\mathbf{A})$. The second is an equality, as it must be: Theorem 17.6.1 with $n = 3$, $k = 2$ gives the reverse bound $\lambda_2(\mathbf{B}) \geq \lambda_3(\mathbf{A})$.

(d) An adapted list has $\mathbf{x}_1 = \pm\mathbf{y}_1$, contributing $\frac{4}{3}$, and $\mathbf{x}_2$ is any unit vector of $\mathbf{w}^\perp$. By Lemma 17.2.2 the smallest value of $\langle \mathbf{A}\mathbf{x}_2, \mathbf{x}_2 \rangle$ there is the smallest eigenvalue of the compression of $\mathbf{A}$ to $\mathbf{w}^\perp$. With the orthonormal basis $\frac{1}{\sqrt{2}}(1, -1, 0)$, $\frac{1}{\sqrt{6}}(1, 1, -2)$ of $\mathbf{w}^\perp$, that compression is

$$\begin{pmatrix} 5/2 & \sqrt{3}/2 \\ \sqrt{3}/2 & 1/6 \end{pmatrix},$$

with trace $\frac{8}{3}$, determinant $\frac{5}{12} - \frac{3}{4} = -\frac{1}{3}$, and characteristic polynomial $\frac{1}{3}(3t^2 - 8t - 1)$. Its smaller eigenvalue is $\frac{1}{3}(4 - \sqrt{19})$. So the minimum over lists adapted to $\mathcal{W}$ is

$$\tfrac{4}{3} + \tfrac{1}{3}(4 - \sqrt{19}) = \tfrac{1}{3}(8 - \sqrt{19}) \approx 1.214.$$

This is smaller than $\frac{3}{2}$. The construction in the proof finds **an** adapted list below the bound 3, which is all part (a) asks; it need not find the best one. Both numbers are below 3, the value at the eigenflag, as part (c) requires.

Exercise. B2: A complex pair

Let

$$\mathbf{A} = \begin{pmatrix} 2 & 1 + i \\ 1 - i & 1 \end{pmatrix}, \qquad \mathbf{B} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \quad \text{in } M_2(\mathbb{C}).$$

(a) Check that $\mathbf{A}$ and $\mathbf{B}$ are Hermitian, and compute $\boldsymbol{\lambda}(\mathbf{A})$, $\boldsymbol{\lambda}(\mathbf{B})$, and $\boldsymbol{\lambda}(\mathbf{A} - \mathbf{B})$.

(b) Verify Theorem 17.7.2 for all three index sets, and say which inequality is an equality and why it had to be.

(c) Verify both bounds of Corollary 17.7.3 (a), and explain why here every $\lambda_i(\mathbf{A}) - \lambda_i(\mathbf{B})$ had to be nonnegative.

Solution.

(a) The diagonal of $\mathbf{A}$ is real and $a_{21} = 1 - i = \overline{a_{12}}$, so $\mathbf{A}^* = \mathbf{A}$; $\mathbf{B}$ is real diagonal. For $\mathbf{A}$, $\operatorname{tr}\mathbf{A} = 3$ and $\det\mathbf{A} = 2 - (1+i)(1-i) = 2 - 2 = 0$, so $\boldsymbol{\lambda}(\mathbf{A}) = (3, 0)$. Next, $\boldsymbol{\lambda}(\mathbf{B}) = (1, -1)$. Finally

$$\mathbf{A} - \mathbf{B} = \begin{pmatrix} 1 & 1+i \\ 1-i & 2 \end{pmatrix}$$

has trace 3 and determinant $2 - 2 = 0$, so $\boldsymbol{\lambda}(\mathbf{A} - \mathbf{B}) = (3, 0)$.

(b) The differences at equal indices are $3 - 1 = 2$ and $0 - (-1) = 1$. For (1): $2 \le 3$. For (2): $1 \le 3$. For $(1, 2)$: $3 \le 3 + 0 = 3$, an equality. It had to be one, since for the full index set both sides equal $\operatorname{tr}\mathbf{A} - \operatorname{tr}\mathbf{B} = \operatorname{tr}(\mathbf{A} - \mathbf{B})$.

(c) The bounds are $\lambda_2(\mathbf{A} - \mathbf{B}) = 0 \le 2 \le 3$ and $0 \le 1 \le 3$. The lower bound $\lambda_n(\mathbf{A} - \mathbf{B}) = 0$ is nonnegative because $\mathbf{A} - \mathbf{B} \succeq \mathbf{O}$: it is Hermitian with eigenvalues $3, 0 \ge 0$ (Theorem 13.1.4). So $\mathbf{A} \succeq \mathbf{B}$, and Corollary 17.2.4 already forces $\lambda_i(\mathbf{A}) \ge \lambda_i(\mathbf{B})$ for each i.

Exercise. B3: The lower bound

Let $\mathbf{A}, \mathbf{B} \in M_n(F)$ be Hermitian and $1 \le i_1 < \cdots < i_k \le n$. Prove that

$$\sum_{j=1}^{k} \lambda_{i_j}(\mathbf{A}) - \sum_{j=1}^{k} \lambda_{i_j}(\mathbf{B}) \;\ge\; \sum_{j=n-k+1}^{n} \lambda_j(\mathbf{A} - \mathbf{B}),$$

that is, the shifts over any k indices add up to at least the k **smallest** eigenvalues of $\mathbf{A} - \mathbf{B}$. Check the result on Example (All seven index sets).

Solution. Apply Theorem 17.7.2 with the roles of $\mathbf{A}$ and $\mathbf{B}$ exchanged:

$$\sum_{j=1}^{k} \lambda_{i_j}(\mathbf{B}) - \sum_{j=1}^{k} \lambda_{i_j}(\mathbf{A}) \;\le\; \sum_{j=1}^{k} \lambda_j(\mathbf{B} - \mathbf{A}).$$

Since $\mathbf{B} - \mathbf{A} = -(\mathbf{A} - \mathbf{B})$, Lemma 17.3.2 gives $\lambda_j(\mathbf{B} - \mathbf{A}) = -\lambda_{n+1-j}(\mathbf{A} - \mathbf{B})$. As j runs over $1, \dots, k$, the index $n + 1 - j$ runs over $n, \dots, n - k + 1$, so the right-hand side is $-\sum_{j=n-k+1}^{n} \lambda_j(\mathbf{A} - \mathbf{B})$. Multiplying the inequality by -1 gives the claim.

In Example (All seven index sets), $\boldsymbol{\lambda}(\mathbf{A} - \mathbf{B}) = (4, 1, -3)$, whose bottom partial sums are -3, -2, 2 for $k = 1, 2, 3$. The left sides are $2, 1, -1$ for $k = 1$, then $3, 1, 0$ for $k = 2$, then 2 for $k = 3$. Each is at least the corresponding bottom sum.

17.7.6.3 C. Going deeper

Exercise. C1: A second proof of Lidskii's inequality

Let $\mathbf{A}, \mathbf{B} \in M_n(F)$ be Hermitian, put $\mathbf{C} = \mathbf{A} - \mathbf{B}$, and fix an index set $1 \le i_1 < \cdots < i_k \le n$.

(a) Let $c \in \mathbb{R}$. Show that replacing $\mathbf{A}$ by $\mathbf{A} - c\mathbf{I}$, with $\mathbf{B}$ unchanged, lowers both sides of Theorem 17.7.2 by kc. Deduce that it suffices to prove the inequality when $\lambda_k(\mathbf{C}) = 0$.

(b) Assume $\lambda_k(\mathbf{C}) = 0$. Write $\mathbf{C} = \mathbf{U}\mathbf{D}\mathbf{U}^*$ with $\mathbf{U}$ unitary and $\mathbf{D} = \operatorname{diag}(\lambda_1(\mathbf{C}), \dots, \lambda_n(\mathbf{C}))$, and put $\mathbf{C}_+ = \mathbf{U}\mathbf{D}_+\mathbf{U}^*$, where $\mathbf{D}_+$ replaces each negative diagonal entry of $\mathbf{D}$ by 0. Prove that $\mathbf{C}_+ \succeq \mathbf{C}$, $\mathbf{C}_+ \succeq \mathbf{O}$, and $\operatorname{tr}\mathbf{C}_+ = \sum_{j\le k}\lambda_j(\mathbf{C})$.

(c) Prove that $\lambda_i(\mathbf{A}) \le \lambda_i(\mathbf{B}+\mathbf{C}_+)$ and $\lambda_i(\mathbf{B}+\mathbf{C}_+) \ge \lambda_i(\mathbf{B})$ for every i, and deduce Theorem 17.7.2.

(d) Which of the ingredients of this proof were available before this section? What does this proof not give that Theorem 17.7.1 does?

Hint: *Corollary 17.2.4*.

Solution.

(a) For every Hermitian $\mathbf{M}$ and real c, $\lambda_i(\mathbf{M} - c\mathbf{I}) = \lambda_i(\mathbf{M}) - c$ for all i. Indeed, if $\mathbf{M} = \mathbf{U}\mathbf{D}\mathbf{U}^*$ with $\mathbf{U}$ unitary and $\mathbf{D}$ real diagonal (Corollary 12.4.3, or Corollary 12.5.2 over $\mathbb{R}$), then $\mathbf{M} - c\mathbf{I} = \mathbf{U}(\mathbf{D} - c\mathbf{I})\mathbf{U}^*$, and subtracting c from every diagonal entry keeps the order. Replacing $\mathbf{A}$ by $\mathbf{A} - c\mathbf{I}$ lowers each $\lambda_{i_j}(\mathbf{A})$ by c, so the left side drops by kc. It also replaces $\mathbf{C}$ by $\mathbf{C} - c\mathbf{I}$, so each $\lambda_j(\mathbf{C})$ drops by c and the right side drops by kc. So the inequality for $(\mathbf{A} - c\mathbf{I}, \mathbf{B})$ is equivalent to the one for $(\mathbf{A}, \mathbf{B})$. With $c = \lambda_k(\mathbf{C})$, the new difference $\mathbf{C} - c\mathbf{I}$ has $\lambda_k = 0$.

(b) $\mathbf{C}_+ - \mathbf{C} = \mathbf{U}(\mathbf{D}_+ - \mathbf{D})\mathbf{U}^*$, and $\mathbf{D}_+ - \mathbf{D}$ is diagonal with entries $\max(-\lambda_i(\mathbf{C}), 0) \ge 0$. So $\mathbf{C}_+ - \mathbf{C}$ is Hermitian with nonnegative eigenvalues, hence $\mathbf{C}_+ - \mathbf{C} \succeq \mathbf{O}$ by Theorem 13.1.4, that is, $\mathbf{C}_+ \succeq \mathbf{C}$. In the same way $\mathbf{C}_+ \succeq \mathbf{O}$, since the entries of $\mathbf{D}_+$ are ≥ 0. For the trace, $\operatorname{tr}\mathbf{C}_+ = \operatorname{tr}\mathbf{D}_+ = \sum_i \max(\lambda_i(\mathbf{C}), 0)$ by Theorem 0.9.6 (3). For $i \le k$, $\lambda_i(\mathbf{C}) \ge \lambda_k(\mathbf{C}) = 0$, so the term is $\lambda_i(\mathbf{C})$. For $i > k$, $\lambda_i(\mathbf{C}) \le 0$, so the term is 0. Hence $\operatorname{tr}\mathbf{C}_+ = \sum_{j\le k}\lambda_j(\mathbf{C})$.

(c) $\mathbf{A} = \mathbf{B} + \mathbf{C} \preceq \mathbf{B} + \mathbf{C}_+$ by (b), and $\mathbf{B} + \mathbf{C}_+ \succeq \mathbf{B}$ because $\mathbf{C}_+ \succeq \mathbf{O}$. Corollary 17.2.4 turns both into the stated eigenvalue inequalities. Therefore

$$\begin{aligned}\sum_{j=1}^{k}(\lambda_{i_j}(\mathbf{A}) - \lambda_{i_j}(\mathbf{B})) &\le \sum_{j=1}^{k}(\lambda_{i_j}(\mathbf{B}+\mathbf{C}_+) - \lambda_{i_j}(\mathbf{B})) \\ &\le \sum_{i=1}^{n}(\lambda_i(\mathbf{B}+\mathbf{C}_+) - \lambda_i(\mathbf{B})),\end{aligned}$$

the second step adding the $n-k$ omitted terms, each ≥ 0. The last sum is $\operatorname{tr}(\mathbf{B}+\mathbf{C}_+) - \operatorname{tr}\mathbf{B} = \operatorname{tr}\mathbf{C}_+$, by Corollary 12.1.4 and Theorem 0.9.6 (1), and $\operatorname{tr}\mathbf{C}_+ = \sum_{j\le k}\lambda_j(\mathbf{C})$ by (b).

(d) Everything used here was available after §02: the spectral theorem, the characterization of positive semidefiniteness by eigenvalues, the trace, and Corollary 17.2.4. The proof never looks at a flag. What it does not give is the variational description itself. Theorem 17.7.1 expresses $\sum_j \lambda_{i_j}(\mathbf{A})$ as an optimization involving $\mathbf{A}$ alone, and like Courant–Fischer and Ky Fan before it, such a description can be reused for other comparisons. This proof is a clever route to one consequence of that description.

Exercise. C2: The total absolute shift

Let $\mathbf{A}, \mathbf{B} \in M_n(F)$ be Hermitian.

(a) Prove that
$$\sum_{i=1}^{n}|\lambda_i(\mathbf{A})-\lambda_i(\mathbf{B})| \;\le\; \sum_{i=1}^{n}|\lambda_i(\mathbf{A}-\mathbf{B})|.$$

(b) Show that equality holds when $\mathbf{A}\succeq\mathbf{B}$. Compute both sides for Example (All seven index sets), and compare with the bound $n\|\mathbf{A}-\mathbf{B}\|_2$ obtained by adding the bounds of Corollary 17.3.6.

Hint for (a): use the set of indices at which $\lambda_i(\mathbf{A})>\lambda_i(\mathbf{B})$.

Solution.

(a) Write $d_i=\lambda_i(\mathbf{A})-\lambda_i(\mathbf{B})$ and $\mu_j=\lambda_j(\mathbf{A}-\mathbf{B})$. Let $S=\{i:d_i>0\}$ and $k=|S|$. Then
$$\sum_{i=1}^{n}|d_i|=\sum_{i\in S}d_i-\sum_{i\notin S}d_i=2\sum_{i\in S}d_i-\sum_{i=1}^{n}d_i.$$
The total is $\sum_i d_i=\operatorname{tr}\mathbf{A}-\operatorname{tr}\mathbf{B}=\operatorname{tr}(\mathbf{A}-\mathbf{B})=\sum_j\mu_j$, by Corollary 12.1.4 and Theorem 0.9.6 (1). If $k\ge1$, Theorem 17.7.2 with the elements of S as index set gives $\sum_{i\in S}d_i\le\sum_{j\le k}\mu_j$; if $k=0$ both sides of this inequality are the empty sum 0. In either case
$$\sum_{i=1}^{n}|d_i| \;\le\; 2\sum_{j\le k}\mu_j-\sum_{j=1}^{n}\mu_j=\sum_{j\le k}\mu_j-\sum_{j>k}\mu_j \;\le\; \sum_{j=1}^{n}|\mu_j|,$$
the last step because $\pm\mu_j\le|\mu_j|$ for each j.

(b) If $\mathbf{A}\succeq\mathbf{B}$ then every $d_i\ge0$ by Corollary 17.2.4, and every $\mu_j\ge0$ by Theorem 13.1.4. Both sides then equal their plain sums, and those are equal, both being $\operatorname{tr}(\mathbf{A}-\mathbf{B})$.

In Example (All seven index sets), the d_i are $2,1,-1$, so the left side is 4; the μ_j are $4,1,-3$, so the right side is 8. By Lemma 17.3.5, $\|\mathbf{A}-\mathbf{B}\|_2=\max(4,3)=4$, and adding the three bounds of Corollary 17.3.6 gives $3\cdot4=12$. The bound of (a) is never worse than this one: each $|\mu_j|\le\|\mathbf{A}-\mathbf{B}\|_2$, so the right side of (a) is at most $n\|\mathbf{A}-\mathbf{B}\|_2$, and it is often much smaller.

17.8 The Diagonal and the Spectrum

A Hermitian matrix carries two lists of n real numbers: its diagonal entries, read off at a glance, and its eigenvalues, computed with effort. Since Chapter 9 we have known that the two lists have the same sum, the trace. This section asks what else the eigenvalues force on the diagonal, and answers the question completely. The answer needs a name, **majorization**. One half of it, Schur's theorem, is §06's bound on partial sums of the diagonal in the new language. The other half, Horn's theorem, says that nothing more is forced; it is proved by an induction driven by a single 2×2 rotation. The section ends with a short second proof of Hadamard's determinant inequality.

Throughout, the field is $\mathbb{R}$ or $\mathbb{C}$, and every matrix whose eigenvalues are indexed is Hermitian, with its eigenvalues real and listed **decreasingly** with multiplicity, $\lambda_1(\mathbf{A})\ge\cdots\ge$

$\lambda_n(\mathbf{A})$. Two vectors in $\mathbb{R}^n$ are attached to such an $\mathbf{A} \in M_n(F)$:

$$\mathbf{d}(\mathbf{A}) = (a_{11}, a_{22}, \ldots, a_{nn}),$$
$$\boldsymbol{\lambda}(\mathbf{A}) = (\lambda_1(\mathbf{A}), \lambda_2(\mathbf{A}), \ldots, \lambda_n(\mathbf{A})).$$

The first is the **diagonal vector**, in the order the entries sit on the diagonal; the second is the **eigenvalue vector**, always in decreasing order. The diagonal entries are real because $\mathbf{A}^* = \mathbf{A}$ gives $a_{ii} = \overline{a_{ii}}$, and the eigenvalues are real by Theorem 12.2.1. The bold $\mathbf{d}(\mathbf{A})$ is a vector, not the diagonal matrix $\operatorname{diag}(a_{11}, \ldots, a_{nn})$.

Two facts about a Hermitian $\mathbf{A}$ are used below. The spectral theorem (Corollary 12.4.3 over $\mathbb{C}$, Corollary 12.5.2 over $\mathbb{R}$) makes $\mathbf{A}$ similar to the real diagonal matrix of its eigenvalues, so its characteristic polynomial splits as $\prod_i (x - \lambda_i(\mathbf{A}))$. Then Theorem 9.3.3 gives

$$\operatorname{tr} \mathbf{A} = \sum_{i=1}^{n} \lambda_i(\mathbf{A}), \qquad \det \mathbf{A} = \prod_{i=1}^{n} \lambda_i(\mathbf{A}).$$

17.8.1 What the diagonal already knows

Start with the smallest interesting case. The matrix

$$\mathbf{A} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$$

has $\mathbf{d}(\mathbf{A}) = (2, 2)$ and eigenvalues 3 and 1, with eigenvectors $(1, 1)$ and $(1, -1)$, so $\boldsymbol{\lambda}(\mathbf{A}) = (3, 1)$. The totals agree, $2 + 2 = 4 = 3 + 1$. But the two lists are not alike: the diagonal is perfectly even, while the spectrum is spread out.

A larger example shows the pattern better. The tridiagonal matrix

$$\mathbf{A} = \begin{pmatrix} 3 & 2 & 0 \\ 2 & 4 & 2 \\ 0 & 2 & 5 \end{pmatrix}$$

has eigenvectors $(1, 2, 2)$, $(-2, -1, 2)$ and $(2, -2, 1)$, with eigenvalues 7, 4 and 1. To check the first, $\mathbf{A}(1, 2, 2) = (3 + 4,\ 2 + 8 + 4,\ 4 + 10) = (7, 14, 14)$; the other two are checked the same way. So $\boldsymbol{\lambda}(\mathbf{A}) = (7, 4, 1)$, while $\mathbf{d}(\mathbf{A}) = (3, 4, 5)$. Sort the diagonal decreasingly to $(5, 4, 3)$ and compare the running totals of the two lists:

k	1	2	3
largest k diagonal entries	5	9	12
largest k eigenvalues	7	11	12

Each running total of the eigenvalues is at least the matching running total of the diagonal, and at $k = n$ the two are equal. None of this is new. Corollary 17.6.3 says that any k diagonal entries add up to at most $\lambda_1(\mathbf{A}) + \cdots + \lambda_k(\mathbf{A})$, with equality at $k = n$, and the k largest diagonal entries are one such choice. What is new is the point of view: to ask which diagonals are **possible**, the corollary's list of inequalities has to be read as a single relation between two vectors.

Two features of this comparison deserve notice. First, the diagonal had to be **sorted**, and this is forced. For a permutation matrix $\mathbf{P}_\sigma$ (Definition 3.4.2), $\mathbf{P}_\sigma^\top = \mathbf{P}_\sigma^{-1}$ by Lemma 3.4.7 (b), so $\mathbf{P}_\sigma^\top \mathbf{A} \mathbf{P}_\sigma$ is Hermitian with the same eigenvalues as $\mathbf{A}$, and its (j, j) entry is $\mathbf{e}_{\sigma(j)}^\top \mathbf{A} \mathbf{e}_{\sigma(j)} = a_{\sigma(j)\sigma(j)}$. Relabeling the basis shuffles the diagonal and fixes the spectrum, so any true relation between $\mathbf{d}(\mathbf{A})$ and $\boldsymbol{\lambda}(\mathbf{A})$ must ignore the order of the diagonal. Second, the comparison is between **running totals from the top**, and it measures how unevenly a fixed total is spread.

17.8.2 Majorization

One vector is majorized by another when the two have the same total but the second piles more of it onto its largest entries: every "top k" sum of the second is at least that of the first.

The sorting step gets a notation first. For $\mathbf{x} = (x_1, \dots, x_n) \in \mathbb{R}^n$, the **decreasing rearrangement** $\mathbf{x}^{\downarrow} = (x_1^{\downarrow}, \dots, x_n^{\downarrow})$ is the vector with the same entries, repeated as often as in $\mathbf{x}$, listed so that $x_1^{\downarrow} \geq x_2^{\downarrow} \geq \cdots \geq x_n^{\downarrow}$. For example, $(0,3,1,3)^{\downarrow} = (3,3,1,0)$. In this notation, $\boldsymbol{\lambda}(\mathbf{A}) = \boldsymbol{\lambda}(\mathbf{A})^{\downarrow}$ by convention, while $\mathbf{d}(\mathbf{A})$ usually needs sorting.

Definition 17.8.1. Majorization

Let $n \geq 1$ and let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ be **real** vectors with the **same number** of entries. We say that $\mathbf{x}$ is **majorized by** $\mathbf{y}$, and write $\mathbf{x} \prec \mathbf{y}$, if

(M1) $\displaystyle\sum_{i=1}^{k} x_i^{\downarrow} \leq \sum_{i=1}^{k} y_i^{\downarrow}$ for **every** $k = 1, \dots, n-1$; and

(M2) $\displaystyle\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$ (**equal totals**).

We say that $\mathbf{x}$ is **weakly majorized by** $\mathbf{y}$, and write $\mathbf{x} \prec_w \mathbf{y}$, if the inequality in (M1) holds for **every** $k = 1, \dots, n$, including $k = n$, and (M2) is dropped.

In words, clause by clause. (M1) sorts both vectors and compares their **running totals from the top**: the largest entry of $\mathbf{x}$ is at most the largest entry of $\mathbf{y}$, the two largest of $\mathbf{x}$ add up to at most the two largest of $\mathbf{y}$, and so on up to $k = n-1$. The word **every** matters: one failing k refutes $\mathbf{x} \prec \mathbf{y}$. (M2) says the totals are equal; a sum does not care about order, so no sorting is needed there. Weak majorization keeps (M1), adds its case $k = n$, which reads $\sum_i x_i \leq \sum_i y_i$, and asks nothing more. Comparing the definitions: $\mathbf{x} \prec \mathbf{y}$ holds **exactly when** $\mathbf{x} \prec_w \mathbf{y}$ and the totals are equal.

There is an equivalent way to read (M1) from the bottom. Subtracting the k-th running total from the equal totals of (M2) gives

$$\sum_{i=k+1}^{n} x_i^{\downarrow} \geq \sum_{i=k+1}^{n} y_i^{\downarrow} \qquad (k = 1, \dots, n-1),$$

so under (M2), the $n-k$ **smallest** entries of $\mathbf{x}$ add up to at least the $n-k$ smallest entries of $\mathbf{y}$. The vector $\mathbf{y}$ is more extreme at both ends: its top is heavier and its bottom is lighter.

The definition makes sense because $\mathbf{x}^{\downarrow}$ is determined by $\mathbf{x}$: when two entries are equal, it does not matter which is written first. A consequence worth recording: **the relation sees only which numbers occur in each vector and how often, not where they sit.** If $\mathbf{x}'$ is a rearrangement of $\mathbf{x}$ and $\mathbf{y}'$ of $\mathbf{y}$, then $\mathbf{x} \prec \mathbf{y}$ if and only if $\mathbf{x}' \prec \mathbf{y}'$, because the rearrangements have the same decreasing rearrangements and the same totals.

Example. A chain, the two extremes, and size one

Check each claim against Definition 17.8.1.

(a) $(1,1,1) \prec (2,1,0) \prec (3,0,0)$.

(b) For every $\mathbf{x} \in \mathbb{R}^n$ with mean $m = \frac{1}{n}\sum_i x_i$, the **flat vector** $m\mathbf{1} = (m, \dots, m)$ satisfies $m\mathbf{1} \prec \mathbf{x}$.

(c) For every $\mathbf{x} \in \mathbb{R}^n$ with **nonnegative** entries and total $s = \sum_i x_i$, $\mathbf{x} \prec (s, 0, \ldots, 0)$.

(d) For $n = 1$, $\mathbf{x} \prec \mathbf{y}$ if and only if $\mathbf{x} = \mathbf{y}$.

Solution.

(a) All three vectors are already decreasing, and all have total 3, so (M2) holds for both pairs. Their running totals are

$$(1,1,1)\colon\ 1,\ 2,\ 3; \qquad (2,1,0)\colon\ 2,\ 3,\ 3; \qquad (3,0,0)\colon\ 3,\ 3,\ 3.$$

At $k = 1$ and $k = 2$ the first row is at most the second ($1 \le 2$, $2 \le 3$), and the second is at most the third ($2 \le 3$, $3 \le 3$). So (M1) holds for both pairs.

(b) The totals are $nm = \sum_i x_i$, so (M2) holds. For (M1), fix k and let $a = \frac{1}{k}\sum_{i \le k} x_i^{\downarrow}$ be the average of the k largest entries. Each of these is at least $x_k^{\downarrow}$, so $a \ge x_k^{\downarrow}$. Every remaining entry $x_j^{\downarrow}$ with $j > k$ is at most $x_k^{\downarrow}$, hence at most a. Therefore

$$nm = \sum_{i \le k} x_i^{\downarrow} + \sum_{j > k} x_j^{\downarrow} \le ka + (n-k)a = na,$$

so $m \le a$, and multiplying by k gives $km \le \sum_{i \le k} x_i^{\downarrow}$. The left side is the k-th running total of $m\mathbf{1}$. This is (M1).

(c) The totals are both s. Every running total of $(s, 0, \ldots, 0)$ equals s, and $\sum_{i \le k} x_i^{\downarrow} \le s$ because the omitted entries $x_{k+1}^{\downarrow}, \ldots, x_n^{\downarrow}$ are ≥ 0. This is (M1). Nonnegativity is really needed: $(1, -1)$ has total 0, but $(1, -1) \prec (0, 0)$ fails at $k = 1$, since $1 > 0$.

(d) For $n = 1$ there are no values of k in (M1), since k runs from 1 to 0. So only (M2) remains, and it says $x_1 = y_1$.

Parts (b) and (c) are the two ends of the order: among vectors with a given total, the flat vector is the most evenly spread, and among nonnegative ones, the vector that puts everything in one entry is the least. Part (d) is the degenerate case: with a single entry there is nothing to spread, and the relation collapses to equality.

The chain in (a) might suggest that any two vectors with the same total can be ranked this way. They cannot.

Example. Neither majorizes the other

Show that $\mathbf{x} = (4, 1, 1)$ and $\mathbf{y} = (3, 3, 0)$ have the same total, that neither $\mathbf{x} \prec \mathbf{y}$ nor $\mathbf{y} \prec \mathbf{x}$, and that both are majorized by $(4, 2, 0)$.

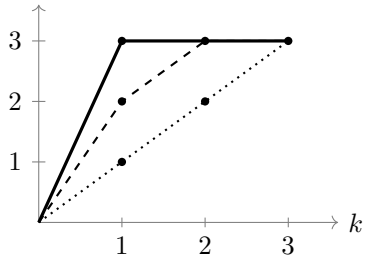
Solution. Both totals are 6, and both vectors are already decreasing. Their running totals are

$$\mathbf{x}\colon\ 4,\ 5,\ 6; \qquad \mathbf{y}\colon\ 3,\ 6,\ 6.$$

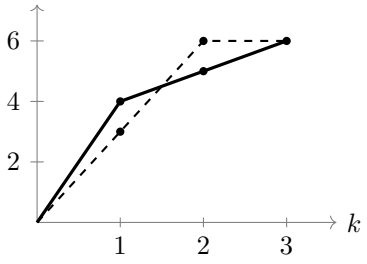
At $k = 1$, $4 > 3$, so (M1) fails for $\mathbf{x} \prec \mathbf{y}$. At $k = 2$, $6 > 5$, so (M1) fails for $\mathbf{y} \prec \mathbf{x}$. The vector $(4, 2, 0)$ has total 6 and running totals $4, 6, 6$, which are at least both $4, 5$ and $3, 6$ at $k = 1, 2$. So $\mathbf{x} \prec (4, 2, 0)$ and $\mathbf{y} \prec (4, 2, 0)$.

This is the nonexample by minimal change. Start from the comparable pair $(4,1,1) \prec (4,2,0)$, move one unit in the second vector from its top entry to its middle entry to get $(3,3,0)$, and the pair becomes incomparable. (M2) still holds; what fails is (M1), at a **different** k in each direction. Spreading is not a single number.

The picture makes this visible. Plot the running totals of a sorted vector against k, joined by segments, starting from $(0,0)$. Then $\mathbf{x} \prec \mathbf{y}$ says that the curve of $\mathbf{y}$ lies on or above the curve of $\mathbf{x}$ and ends at the same height. Two curves can also cross.



solid: $(3,0,0)$
dashed: $(2,1,0)$
dotted: $(1,1,1)$
nested, same endpoint



solid: $(4,1,1)$
dashed: $(3,3,0)$
the curves cross:
incomparable

Why this definition. Each clause earns its place, and changing any one of them changes the subject.

- *Sorting.* Relabeling a basis shuffles $\mathbf{d}(\mathbf{A})$ and fixes $\boldsymbol{\lambda}(\mathbf{A})$, so a relation comparing unsorted entries would answer a question the matrix does not pose.
- *Running totals from the top, with* $\leq$. These are the quantities that Theorem 17.6.2 and Corollary 17.6.3 control: the top k eigenvalues sum to a maximum. Totals from the bottom add nothing new, as shown above, because under (M2) they are the same information.
- *Equal totals.* This is the trace. Dropping it gives weak majorization, the right notion when there is no conservation law to enforce: $(1,0) \prec_w (2,0)$, but $(1,0) \not\prec (2,0)$, because the totals differ.

Warning. **$\mathbf{x} \prec \mathbf{y}$ does not mean $x_i \leq y_i$, not even after sorting both vectors.** In $(1,1,1) \prec (2,1,0)$, the third entries are 1 and 0, and $1 > 0$. This is no accident: subtracting (M1) at $k = n-1$ from (M2) gives $x_n^{\downarrow} \geq y_n^{\downarrow}$ whenever $\mathbf{x} \prec \mathbf{y}$, so at the bottom the inequality always runs the other way. Majorization compares **sums of the largest entries**, not entries one at a time.

Quick check. Arrange the five vectors $(2,1,1,0)$, $(4,0,0,0)$, $(1,1,1,1)$, $(3,1,0,0)$, and $(2,2,0,0)$ in a chain $\mathbf{x}^{(1)} \prec \mathbf{x}^{(2)} \prec \cdots \prec \mathbf{x}^{(5)}$.

Solution. All five are decreasing and all have total 4. Their running totals at $k = 1,2,3$ are

$$(1,1,1,1)\colon\ 1,\ 2,\ 3; \qquad (2,1,1,0)\colon\ 2,\ 3,\ 4;$$
$$(2,2,0,0)\colon\ 2,\ 4,\ 4; \qquad (3,1,0,0)\colon\ 3,\ 4,\ 4;$$
$$(4,0,0,0)\colon\ 4,\ 4,\ 4.$$

Read in this order, each triple is entrywise at most the next, so

$$(1,1,1,1) \prec (2,1,1,0) \prec (2,2,0,0) \prec (3,1,0,0) \prec (4,0,0,0).$$

The ends are the flat vector and the concentrated vector of Example (A chain, the two extremes, and size one) (b) and (c), as they must be. These five happen to be comparable in pairs, but Example (Neither majorizes the other) shows that this is not always so.

17.8.3 Majorization is an order

The first small result is that $\prec$ behaves like an order, and it says precisely how it falls short of being a partial order on $\mathbb{R}^n$.

Proposition 17.8.1. Basic Properties of Majorization

Let $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}^n$.

(a) $\mathbf{x} \prec \mathbf{x}$.

(b) If $\mathbf{x} \prec \mathbf{y}$ and $\mathbf{y} \prec \mathbf{z}$, then $\mathbf{x} \prec \mathbf{z}$.

(c) If $\mathbf{x} \prec \mathbf{y}$ and $\mathbf{y} \prec \mathbf{x}$, then $\mathbf{x}^{\downarrow} = \mathbf{y}^{\downarrow}$; that is, $\mathbf{y}$ is a rearrangement of $\mathbf{x}$.

Statements (a) and (b) hold with $\prec_w$ in place of $\prec$.

Proof. For $k = 1, \dots, n$, write $X_k = \sum_{i \le k} x_i^{\downarrow}$, and define Y_k and Z_k in the same way. Since $X_n = \sum_i x_i$, Definition 17.8.1 says that $\mathbf{x} \prec \mathbf{y}$ means $X_k \le Y_k$ for $k < n$ and $X_n = Y_n$.

(a) $X_k \le X_k$ for every k, and $X_n = X_n$.

(b) For $k < n$, $X_k \le Y_k \le Z_k$. At $k = n$, $X_n = Y_n = Z_n$. Hence $\mathbf{x} \prec \mathbf{z}$.

(c) The hypotheses give $X_k \le Y_k$ and $Y_k \le X_k$ for $k < n$, and $X_n = Y_n$, so $X_k = Y_k$ for every $k = 1, \dots, n$. Put $X_0 = Y_0 = 0$. Then for each k,

$$x_k^{\downarrow} = X_k - X_{k-1} = Y_k - Y_{k-1} = y_k^{\downarrow},$$

so $\mathbf{x}^{\downarrow} = \mathbf{y}^{\downarrow}$. Two vectors with the same decreasing rearrangement contain the same numbers, each the same number of times, so each is a rearrangement of the other.

For $\prec_w$, the arguments for (a) and (b) are the same with every $k \le n$ treated like $k < n$. This proves the proposition.

■

So $\prec$ is reflexive and transitive. A relation with these two properties is called a **preorder**. Part (c) is antisymmetry up to rearrangement, and no more: $(1,0) \prec (0,1) \prec (1,0)$, yet $(1,0) \neq (0,1)$. So $\prec$ is **not** a partial order on $\mathbb{R}^n$. Restricted to **decreasing** vectors, where a rearrangement is the vector itself, (c) is exact antisymmetry, and there $\prec$ is a partial order. By Example (Neither majorizes the other) it is **not total**, not even among nonnegative decreasing vectors with the same total. Like the Loewner order of Chapter 13 (Proposition 13.5.1), it leaves many pairs uncompared.

17.8.4 Schur's theorem

The running totals in the 3×3 example were not a coincidence. Restated in the language of majorization, §06's Corollary 17.6.3 says that for every Hermitian matrix the diagonal is more evenly spread than the spectrum:

Theorem 17.8.2. Schur's Majorization Theorem

Let $F = \mathbb{R}$ or $F = \mathbb{C}$, and let $\mathbf{A} \in M_n(F)$ be Hermitian. Then

$$\mathbf{d}(\mathbf{A}) \prec \boldsymbol{\lambda}(\mathbf{A}).$$

Explicitly, if $\mathbf{d}(\mathbf{A})^{\downarrow} = (a_1^{\downarrow}, \dots, a_n^{\downarrow})$ lists the diagonal entries of $\mathbf{A}$ decreasingly, then

$$\begin{aligned} a_1^{\downarrow} + \cdots + a_k^{\downarrow} &\le \lambda_1(\mathbf{A}) + \cdots + \lambda_k(\mathbf{A}) \qquad (1 \le k \le n-1), \\ a_1^{\downarrow} + \cdots + a_n^{\downarrow} &= \lambda_1(\mathbf{A}) + \cdots + \lambda_n(\mathbf{A}). \end{aligned}$$

Proof. Both vectors lie in $\mathbb{R}^n$, as noted at the start of the section. Fix k with $1 \le k \le n$. List the positions $1, \dots, n$ in an order in which the diagonal entries decrease, breaking ties arbitrarily, and let $i_1 < \cdots < i_k$ be the first k positions of that list, written in increasing order. Then $a_{i_1 i_1}, \dots, a_{i_k i_k}$ are $a_1^{\downarrow}, \dots, a_k^{\downarrow}$ in some order, so Corollary 17.6.3 gives

$$a_1^{\downarrow} + \cdots + a_k^{\downarrow} = \sum_{j=1}^{k} a_{i_j i_j} \le \lambda_1(\mathbf{A}) + \cdots + \lambda_k(\mathbf{A}),$$

with equality when $k = n$. For $k \le n-1$ this is (M1). For $k = n$ it is (M2), since a sum does not depend on the order of its terms. This proves the theorem. ∎

So Schur's theorem is Corollary 17.6.3 restated. What the restatement adds is a relation between two vectors, and a relation has a converse that can be asked about. Read from the bottom, (M1) says that the $n-k$ smallest diagonal entries add up to **at least** the $n-k$ smallest eigenvalues. At the two ends of Example (A chain, the two extremes, and size one), a **diagonal** matrix has $\mathbf{d}(\mathbf{A})$ a rearrangement of $\boldsymbol{\lambda}(\mathbf{A})$, the most spread a diagonal can be; a **constant** diagonal t is the flat vector $t\mathbf{1}$, where Schur's theorem says only that $nt = \lambda_1(\mathbf{A}) + \cdots + \lambda_n(\mathbf{A})$.

Example. Coordinate vectors against eigenvectors

For the matrix

$$\mathbf{A} = \begin{pmatrix} 3 & 2 & 0 \\ 2 & 4 & 2 \\ 0 & 2 & 5 \end{pmatrix}$$

from the start of the section, whose running totals were tabulated there, exhibit for $k = 2$ the coordinate pair that the proof of Corollary 17.6.3 feeds to Theorem 17.6.2 and an orthonormal pair that attains the maximum in Theorem 17.6.2. Account for the difference between the two sums.

Solution. We found $\boldsymbol{\lambda}(\mathbf{A}) = (7, 4, 1)$, with unit eigenvectors $\mathbf{u}_1 = \frac{1}{3}(1, 2, 2)$, $\mathbf{u}_2 = \frac{1}{3}(-2, -1, 2)$, and $\mathbf{u}_3 = \frac{1}{3}(2, -2, 1)$. They are pairwise orthogonal (for instance $\langle (1, 2, 2), (-2, -1, 2)\rangle = -2 - 2 + 4 = 0$), and each vector in parentheses has length 3.

For $k = 2$ the proof uses the two largest diagonal entries, $a_{33} = 5$ and $a_{22} = 4$, so the coordinate pair is $(\mathbf{e}_2, \mathbf{e}_3)$, and $\langle \mathbf{A}\mathbf{e}_2, \mathbf{e}_2\rangle + \langle \mathbf{A}\mathbf{e}_3, \mathbf{e}_3\rangle = 9$. The pair $(\mathbf{u}_1, \mathbf{u}_2)$ gives $\langle \mathbf{A}\mathbf{u}_1, \mathbf{u}_1\rangle + \langle \mathbf{A}\mathbf{u}_2, \mathbf{u}_2\rangle = 7 + 4 = 11$, since $\langle \mathbf{A}\mathbf{u}_i, \mathbf{u}_i\rangle = \lambda_i \|\mathbf{u}_i\|^2 = \lambda_i$. The coordinate pair falls short of the eigenvector pair by 2, and that shortfall is exactly the slack in Schur's theorem at $k = 2$. Ky Fan's maximum is attained by eigenvectors, and the diagonal only ever sees coordinate vectors.

Quick check. A Hermitian 3×3 matrix $\mathbf{A}$ has diagonal entries $4, -1, 2$. Without knowing anything else about $\mathbf{A}$, what does Theorem 17.8.2 say about $\lambda_1(\mathbf{A})$, about $\lambda_1(\mathbf{A}) + \lambda_2(\mathbf{A})$, and about $\lambda_3(\mathbf{A})$?

Solution. Sorted, $\mathbf{d}(\mathbf{A})^{\downarrow} = (4, 2, -1)$, with running totals $4, 6, 5$. So $\lambda_1(\mathbf{A}) \geq 4$, $\lambda_1(\mathbf{A}) + \lambda_2(\mathbf{A}) \geq 6$, and $\lambda_1(\mathbf{A}) + \lambda_2(\mathbf{A}) + \lambda_3(\mathbf{A}) = 5$. Subtracting the middle inequality from the last equality gives $\lambda_3(\mathbf{A}) \leq 5 - 6 = -1$. In particular $\mathbf{A}$ is indefinite, since $\lambda_1(\mathbf{A}) > 0 > \lambda_3(\mathbf{A})$, and that was visible without computing any eigenvalue.

17.8.5 Horn's converse

Schur's theorem gives a **necessary** condition on a diagonal. Is it also sufficient? Given $\mathbf{d} \prec \boldsymbol{\lambda}$, is there a Hermitian matrix with eigenvalues $\lambda_1, \dots, \lambda_n$ and diagonal entries $d_1, \dots, d_n$, in the given order? For $n = 2$ a rotation answers the question. Turning the eigenbasis of $\operatorname{diag}(\lambda_1, \lambda_2)$ moves its first diagonal entry through every value between λ_2 and λ_1, and the trace fixes the second entry.

Lemma 17.8.3. The Two-by-Two Rotation

Let $\lambda_1 \geq \lambda_2$ and t be real numbers with $\lambda_2 \leq t \leq \lambda_1$. Put $b = \sqrt{(\lambda_1 - t)(t - \lambda_2)}$. Then there is an orthogonal $\mathbf{R} \in \mathrm{O}(2)$ with

$$\mathbf{R} \operatorname{diag}(\lambda_1, \lambda_2) \mathbf{R}^{\top} = \begin{pmatrix} t & b \\ b & \lambda_1 + \lambda_2 - t \end{pmatrix}.$$

Idea. If the first row of $\mathbf{R}$ is $(c, -s)$ with $c^2 + s^2 = 1$, the $(1, 1)$ entry of $\mathbf{R} \operatorname{diag}(\lambda_1, \lambda_2) \mathbf{R}^{\top}$ is $c^2 \lambda_1 + s^2 \lambda_2$, an average of λ_1 and λ_2 with weights c^2 and s^2. Every t between them is such an average, and solving for the weights gives c and s.

Proof. If $\lambda_1 = \lambda_2$, then $t = \lambda_1 = \lambda_2$ and $b = 0$, and $\mathbf{R} = \mathbf{I}$ works. Otherwise $\lambda_1 - \lambda_2 > 0$,

and

$$c = \sqrt{\frac{t-\lambda_2}{\lambda_1-\lambda_2}}, \qquad s = \sqrt{\frac{\lambda_1-t}{\lambda_1-\lambda_2}}$$

are real numbers, because both radicands are ≥ 0 by the hypothesis on t. They satisfy $c^2+s^2=1$. Let $\mathbf{R} = \left(\begin{smallmatrix} c & -s \\ s & c \end{smallmatrix}\right)$. Then $\mathbf{R}^\top\mathbf{R} = (c^2+s^2)\mathbf{I} = \mathbf{I}$, so $\mathbf{R} \in \mathrm{O}(2)$. Multiplying out,

$$\mathbf{R}\,\mathrm{diag}(\lambda_1,\lambda_2)\mathbf{R}^\top = \begin{pmatrix} c^2\lambda_1 + s^2\lambda_2 & cs(\lambda_1-\lambda_2) \\ cs(\lambda_1-\lambda_2) & s^2\lambda_1 + c^2\lambda_2 \end{pmatrix}.$$

Since $s^2 = 1-c^2$, the $(1,1)$ entry is $\lambda_2 + c^2(\lambda_1-\lambda_2) = \lambda_2 + (t-\lambda_2) = t$. The two diagonal entries add up to $(c^2+s^2)(\lambda_1+\lambda_2) = \lambda_1+\lambda_2$, so the $(2,2)$ entry is $\lambda_1+\lambda_2-t$. Finally $cs(\lambda_1-\lambda_2) = \sqrt{(t-\lambda_2)(\lambda_1-t)} = b$. This proves the lemma.

■

This is the converse for $n=2$: if $\lambda_1 \geq \lambda_2$ and $(d_1,d_2) \prec (\lambda_1,\lambda_2)$, then $d_1, d_2 \leq d_1^{\downarrow} \leq \lambda_1$ by (M1), so by (M2) $d_1 = \lambda_1+\lambda_2-d_2 \geq \lambda_2$, so $t=d_1$ is allowed. For larger n, the lemma is used once, on a well-chosen pair of eigenvalues, to put d_1 in the corner, and induction does the rest.

Theorem 17.8.4. Horn's Theorem

Let $n \geq 1$, and let $\mathbf{d} = (d_1,\dots,d_n)$ and $\boldsymbol{\lambda} = (\lambda_1,\dots,\lambda_n)$ be vectors in $\mathbb{R}^n$ with $\mathbf{d} \prec \boldsymbol{\lambda}$. Then there is an orthogonal $\mathbf{Q} \in \mathrm{O}(n)$ such that the matrix

$$\mathbf{A} = \mathbf{Q}\,\mathrm{diag}(\lambda_1,\dots,\lambda_n)\mathbf{Q}^\top$$

has diagonal entries $d_1, d_2, \dots, d_n$, **in this order**. So $\mathbf{A}$ is a real symmetric matrix with $\mathbf{d}(\mathbf{A}) = \mathbf{d}$ and $\boldsymbol{\lambda}(\mathbf{A}) = \boldsymbol{\lambda}^{\downarrow}$.

Idea. Induction on n, one diagonal entry at a time.

① Permutation matrices are orthogonal and shuffle diagonal entries, so it is enough to treat $\mathbf{d}$ and $\boldsymbol{\lambda}$ both decreasing. ② The top and bottom clauses of majorization put d_1 between λ_n and λ_1, hence between two **consecutive** eigenvalues, $\lambda_k \geq d_1 \geq \lambda_{k+1}$. The lemma rotates that pair so that d_1 sits in the corner, with $\eta = \lambda_k + \lambda_{k+1} - d_1$ next to it. ③ The Claim: what is left over, the diagonal $d_2,\dots,d_n$ and the spectrum with λ_k, λ_{k+1} replaced by η, satisfies the hypothesis again, in size $n-1$. ④ Induction rotates the lower $(n-1)\times(n-1)$ block into place without touching the corner.

Step ③ is where the choice of k matters.

Proof. For every $\mathbf{Q} \in \mathrm{O}(n)$, the matrix $\mathbf{A} = \mathbf{Q}\,\mathrm{diag}(\lambda_1,\dots,\lambda_n)\mathbf{Q}^\top$ is real and symmetric, and by Corollary 12.5.2 its eigenvalues, with multiplicity, are $\lambda_1,\dots,\lambda_n$; listed decreasingly they form $\boldsymbol{\lambda}^{\downarrow}$. So only the diagonal needs work. We use two facts repeatedly. A product of orthogonal matrices is orthogonal, since $(\mathbf{XY})^\top\mathbf{XY} = \mathbf{Y}^\top\mathbf{X}^\top\mathbf{XY} = \mathbf{I}$. And for a permutation $\pi \in S_n$ and any $\mathbf{M} \in M_n(\mathbb{R})$, Definition 3.4.2 gives

$$(\mathbf{P}_\pi^\top \mathbf{M} \mathbf{P}_\pi)_{ij} = \mathbf{e}_{\pi(i)}^\top \mathbf{M} \mathbf{e}_{\pi(j)} = m_{\pi(i)\pi(j)}, \tag{†}$$

where $\mathbf{P}_\pi$ is orthogonal by Lemma 3.4.7 (b).

Step 1: reduction to decreasing vectors. Suppose the theorem holds whenever $\mathbf{d}$ and $\boldsymbol{\lambda}$ are both decreasing, and let $\mathbf{d} \prec \boldsymbol{\lambda}$ be arbitrary. Majorization sees only rearrangements (as noted after Definition 17.8.1), so $\mathbf{d}^{\downarrow} \prec \boldsymbol{\lambda}^{\downarrow}$, and there is $\mathbf{Q}_0 \in \mathrm{O}(n)$ such that $\mathbf{B} = \mathbf{Q}_0 \operatorname{diag}(\boldsymbol{\lambda}^{\downarrow})\mathbf{Q}_0^{\top}$ has diagonal $\mathbf{d}^{\downarrow}$. Here $\operatorname{diag}(\boldsymbol{\lambda}^{\downarrow})$ is the diagonal matrix with diagonal $\lambda_1^{\downarrow}, \dots, \lambda_n^{\downarrow}$. Since $\boldsymbol{\lambda}^{\downarrow}$ is a rearrangement of $\boldsymbol{\lambda}$, and $\mathbf{d}^{\downarrow}$ of $\mathbf{d}$, there are permutations $\sigma, \tau \in S_n$ with $\lambda_{\sigma(i)} = \lambda_i^{\downarrow}$ and $d_{\tau(j)}^{\downarrow} = d_j$ for all i, j. By (†), $\mathbf{P}_\sigma^{\top} \operatorname{diag}(\lambda_1, \dots, \lambda_n)\mathbf{P}_\sigma = \operatorname{diag}(\boldsymbol{\lambda}^{\downarrow})$, and the (j, j) entry of $\mathbf{P}_\tau^{\top}\mathbf{B}\mathbf{P}_\tau$ is $b_{\tau(j)\tau(j)} = d_{\tau(j)}^{\downarrow} = d_j$. So $\mathbf{Q} = \mathbf{P}_\tau^{\top}\mathbf{Q}_0\mathbf{P}_\sigma^{\top}$ is orthogonal and

$$\begin{aligned}\mathbf{Q} \operatorname{diag}(\lambda_1, \dots, \lambda_n)\mathbf{Q}^{\top} &= \mathbf{P}_\tau^{\top}\mathbf{Q}_0 \operatorname{diag}(\boldsymbol{\lambda}^{\downarrow})\mathbf{Q}_0^{\top}\mathbf{P}_\tau \\ &= \mathbf{P}_\tau^{\top}\mathbf{B}\mathbf{P}_\tau\end{aligned}$$

has diagonal $\mathbf{d}$. It therefore suffices to prove the theorem for decreasing $\mathbf{d}$ and $\boldsymbol{\lambda}$, and we do so by induction on n.

Step 2: the corner. For $n = 1$, $\mathbf{d} \prec \boldsymbol{\lambda}$ means $d_1 = \lambda_1$ (Example (A chain, the two extremes, and size one) (d)), and $\mathbf{Q} = (1)$ works. Let $n \geq 2$, assume the theorem for $n - 1$ (for vectors in any order, by Step 1), and let $\mathbf{d}$ and $\boldsymbol{\lambda}$ be decreasing with $\mathbf{d} \prec \boldsymbol{\lambda}$. By (M1) at $k = 1$, $d_1 \leq \lambda_1$. Subtracting (M1) at $k = n - 1$ from (M2) gives $d_n \geq \lambda_n$, so $d_1 \geq d_n \geq \lambda_n$. Let k be the smallest index in $\{1, \dots, n-1\}$ with $\lambda_{k+1} \leq d_1$; one exists, since $\lambda_n \leq d_1$. Then $\lambda_k \geq d_1$: for $k = 1$ this is $d_1 \leq \lambda_1$, and for $k \geq 2$ the minimality of k gives $\lambda_k > d_1$. So

$$\lambda_k \geq d_1 \geq \lambda_{k+1}.$$

By Lemma 17.8.3 with $t = d_1$, there are $\mathbf{R} \in \mathrm{O}(2)$ and a real b with

$$\mathbf{R} \operatorname{diag}(\lambda_k, \lambda_{k+1})\mathbf{R}^{\top} = \begin{pmatrix} d_1 & b \\ b & \eta \end{pmatrix}, \qquad \eta = \lambda_k + \lambda_{k+1} - d_1.$$

Moreover $\lambda_{k+1} \leq \eta \leq \lambda_k$, since $\eta - \lambda_{k+1} = \lambda_k - d_1 \geq 0$ and $\lambda_k - \eta = d_1 - \lambda_{k+1} \geq 0$.

Step 3: the leftover data. Put

$$\begin{aligned}\mathbf{d}' &= (d_2, \dots, d_n), \\ \boldsymbol{\lambda}' &= (\eta, \lambda_1, \dots, \lambda_{k-1}, \lambda_{k+2}, \dots, \lambda_n),\end{aligned}$$

two vectors in $\mathbb{R}^{n-1}$. So $\boldsymbol{\lambda}'$ is $\boldsymbol{\lambda}$ with the pair λ_k, λ_{k+1} replaced by the single number η, written first.

Claim. $\mathbf{d}' \prec \boldsymbol{\lambda}'$.

Proof. By Step 2, $\lambda_{k-1} \geq \lambda_k \geq \eta \geq \lambda_{k+1} \geq \lambda_{k+2}$, for those of these indices that exist. So

$$(\mu_1, \dots, \mu_{n-1}) = (\lambda_1, \dots, \lambda_{k-1}, \eta, \lambda_{k+2}, \dots, \lambda_n)$$

is $(\boldsymbol{\lambda}')^{\downarrow}$, and $\mathbf{d}'$ is already decreasing. For (M2), $\sum_i d_i' = \sum_i d_i - d_1 = \sum_i \lambda_i - d_1 = \sum_i \lambda_i'$, by (M2) for $\mathbf{d} \prec \boldsymbol{\lambda}$ and the definition of η. For (M1), fix m with $1 \leq m \leq n - 2$; when $n = 2$ there is nothing to check.

Case $m \leq k - 1$. Then $\mu_i = \lambda_i$ for $i \leq m$. For each such i, $d_{i+1} \leq d_1 \leq \lambda_k \leq \lambda_i$, because $\mathbf{d}$ and $\boldsymbol{\lambda}$ are decreasing and $i < k$. Summing over $i \leq m$ gives $d_2 + \cdots + d_{m+1} \leq \mu_1 + \cdots + \mu_m$.

Case $m \geq k$. The first m entries of $(\boldsymbol{\lambda}')^{\downarrow}$ are $\lambda_1, \dots, \lambda_{k-1}$, then η, then $\lambda_{k+2}, \dots, \lambda_{m+1}$. So

$$\begin{aligned} \mu_1 + \cdots + \mu_m &= (\lambda_1 + \cdots + \lambda_{m+1}) - \lambda_k - \lambda_{k+1} + \eta \\ &= (\lambda_1 + \cdots + \lambda_{m+1}) - d_1. \end{aligned}$$

Since $m+1 \leq n-1$, (M1) for $\mathbf{d} \prec \boldsymbol{\lambda}$ applies at $m+1$ and gives

$$\begin{aligned} d_2 + \cdots + d_{m+1} &= (d_1 + \cdots + d_{m+1}) - d_1 \\ &\leq (\lambda_1 + \cdots + \lambda_{m+1}) - d_1 = \mu_1 + \cdots + \mu_m. \end{aligned}$$

In both cases (M1) holds, which proves the claim.

■

Step 4: assembly. Let $\sigma \in S_n$ list the indices $k, k+1$ first and the others after them in increasing order: $\sigma(1) = k$, $\sigma(2) = k+1$, and $\sigma(3), \dots, \sigma(n)$ are $1, \dots, k-1, k+2, \dots, n$. By (†),

$$\mathbf{P}_\sigma^{\top} \operatorname{diag}(\lambda_1, \dots, \lambda_n) \mathbf{P}_\sigma = \begin{pmatrix} \operatorname{diag}(\lambda_k, \lambda_{k+1}) & \mathbf{O} \\ \mathbf{O} & \mathbf{D}_1 \end{pmatrix},$$

where $\mathbf{D}_1 = \operatorname{diag}(\lambda_1, \dots, \lambda_{k-1}, \lambda_{k+2}, \dots, \lambda_n) \in M_{n-2}(\mathbb{R})$; when $n = 2$ the blocks involving $\mathbf{D}_1$ are absent. Let $\mathbf{R}_1 = \left(\begin{smallmatrix} \mathbf{R} & \mathbf{O} \\ \mathbf{O} & \mathbf{I}_{n-2} \end{smallmatrix}\right)$. Multiplying blocks and using Step 2,

$$\mathbf{M} = \mathbf{R}_1 \mathbf{P}_\sigma^{\top} \operatorname{diag}(\lambda_1, \dots, \lambda_n) \mathbf{P}_\sigma \mathbf{R}_1^{\top} = \begin{pmatrix} d_1 & b & \mathbf{0}^{\top} \\ b & \eta & \mathbf{0}^{\top} \\ \mathbf{0} & \mathbf{0} & \mathbf{D}_1 \end{pmatrix} = \begin{pmatrix} d_1 & b\,\mathbf{e}_1^{\top} \\ b\,\mathbf{e}_1 & \mathbf{E} \end{pmatrix},$$

where $\mathbf{e}_1$ is the first standard basis vector of $\mathbb{R}^{n-1}$ and $\mathbf{E} = \operatorname{diag}(\eta, \lambda_1, \dots, \lambda_{k-1}, \lambda_{k+2}, \dots, \lambda_n)$ is the diagonal matrix of $\boldsymbol{\lambda}'$. By the Claim and the inductive hypothesis, there is $\mathbf{W} \in \mathrm{O}(n-1)$ such that $\mathbf{W}\mathbf{E}\mathbf{W}^{\top}$ has diagonal entries $d_2, \dots, d_n$. Let $\mathbf{W}_1 = \left(\begin{smallmatrix} 1 & \mathbf{0}^{\top} \\ \mathbf{0} & \mathbf{W} \end{smallmatrix}\right)$. Then

$$\mathbf{W}_1 \mathbf{M} \mathbf{W}_1^{\top} = \begin{pmatrix} d_1 & b\,(\mathbf{W}\mathbf{e}_1)^{\top} \\ b\,\mathbf{W}\mathbf{e}_1 & \mathbf{W}\mathbf{E}\mathbf{W}^{\top} \end{pmatrix},$$

whose diagonal entries are $d_1, d_2, \dots, d_n$. The block matrices $\mathbf{R}_1$ and $\mathbf{W}_1$ are orthogonal, since $\mathbf{R}_1^{\top}\mathbf{R}_1$ and $\mathbf{W}_1^{\top}\mathbf{W}_1$ are block diagonal with blocks $\mathbf{R}^{\top}\mathbf{R} = \mathbf{I}$, $\mathbf{I}$, and 1, $\mathbf{W}^{\top}\mathbf{W} = \mathbf{I}$. So $\mathbf{Q} = \mathbf{W}_1 \mathbf{R}_1 \mathbf{P}_\sigma^{\top}$ is orthogonal, and $\mathbf{Q} \operatorname{diag}(\lambda_1, \dots, \lambda_n) \mathbf{Q}^{\top} = \mathbf{W}_1 \mathbf{M} \mathbf{W}_1^{\top}$. This completes the induction and proves the theorem.

■

Over $\mathbb{C}$ nothing more is needed, because an orthogonal matrix is unitary and a real symmetric matrix is Hermitian. Together with Schur's theorem, this settles the question the section began with.

Theorem 17.8.5. Schur–Horn Theorem

Let $F = \mathbb{R}$ or $F = \mathbb{C}$, and let $\mathbf{d}, \boldsymbol{\lambda} \in \mathbb{R}^n$. There is a Hermitian $\mathbf{A} \in M_n(F)$ with $\boldsymbol{\lambda}(\mathbf{A}) = \boldsymbol{\lambda}^{\downarrow}$ and $\mathbf{d}(\mathbf{A}) = \mathbf{d}$ if and only if $\mathbf{d} \prec \boldsymbol{\lambda}$. In words: the diagonals of the Hermitian matrices with spectrum $\boldsymbol{\lambda}$ are **exactly** the vectors majorized by $\boldsymbol{\lambda}$, and real symmetric matrices already give all of them.

Proof. ($\Rightarrow$) By Theorem 17.8.2, $\mathbf{d} = \mathbf{d}(\mathbf{A}) \prec \boldsymbol{\lambda}(\mathbf{A}) = \boldsymbol{\lambda}^{\downarrow}$. Since majorization sees only rearrangements, $\mathbf{d} \prec \boldsymbol{\lambda}$.

($\Leftarrow$) Theorem 17.8.4 gives a real symmetric $\mathbf{A} \in M_n(\mathbb{R}) \subseteq M_n(F)$ with $\mathbf{d}(\mathbf{A}) = \mathbf{d}$ and $\boldsymbol{\lambda}(\mathbf{A}) = \boldsymbol{\lambda}^{\downarrow}$, and a real symmetric matrix is Hermitian.

■

Example. Building a matrix from its diagonal and spectrum

Find a real symmetric matrix with eigenvalues $6, 1, 0$ and diagonal $(4, 2, 1)$.

Solution. Both vectors are decreasing, and $(4,2,1) \prec (6,1,0)$: the running totals are $4, 6$ against $6, 7$, and both totals are 7. Follow the proof of Theorem 17.8.4.

The corner. $d_1 = 4$ lies between $\lambda_1 = 6$ and $\lambda_2 = 1$, so $k = 1$, $\eta = 6 + 1 - 4 = 3$, and Lemma 17.8.3 gives $b = \sqrt{(6-4)(4-1)} = \sqrt{6}$. After the first rotation the matrix is

$$\mathbf{M} = \begin{pmatrix} 4 & \sqrt{6} & 0 \\ \sqrt{6} & 3 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

The leftover. $\mathbf{d}' = (2, 1)$ and $\boldsymbol{\lambda}' = (3, 0)$, with $2 \le 3$ and $2 + 1 = 3 + 0$, so $\mathbf{d}' \prec \boldsymbol{\lambda}'$, as the Claim promises. The lemma with $t = 2$ gives $c = \sqrt{2/3}$, $s = \sqrt{1/3}$, and

$$\mathbf{W}\,\mathrm{diag}(3,0)\mathbf{W}^{\top} = \begin{pmatrix} 2 & \sqrt{2} \\ \sqrt{2} & 1 \end{pmatrix}, \qquad \mathbf{W}\mathbf{e}_1 = (c, s).$$

Assembly. The first row becomes $(4,\ \sqrt{6}\,c,\ \sqrt{6}\,s) = (4, 2, \sqrt{2})$, so

$$\mathbf{A} = \begin{pmatrix} 4 & 2 & \sqrt{2} \\ 2 & 2 & \sqrt{2} \\ \sqrt{2} & \sqrt{2} & 1 \end{pmatrix}.$$

Check directly: $\mathbf{A}(3, 2, \sqrt{2}) = (18, 12, 6\sqrt{2}) = 6\,(3, 2, \sqrt{2})$, $\mathbf{A}(-\sqrt{2}, \sqrt{2}, 1) = (-\sqrt{2}, \sqrt{2}, 1)$, and $\mathbf{A}(0, -1, \sqrt{2}) = (0, 0, 0)$. Three eigenvectors with distinct eigenvalues $6, 1, 0$ give the whole spectrum, and the diagonal is $(4, 2, 1)$.

17.8.6 Hadamard's inequality from the diagonal

Chapter 13 §07 proved Theorem 13.7.3: for a positive semidefinite $\mathbf{A}$, $\det \mathbf{A} \le a_{11}a_{22}\cdots a_{nn}$, with equality if and only if $\mathbf{A}$ is diagonal or some $a_{ii} = 0$. That proof inducts on the size through Fischer's inequality (Corollary 13.7.2). A much shorter route scales $\mathbf{A}$ so that its diagonal is flat, and then compares the product of the eigenvalues with their sum. That comparison is the inequality between the arithmetic and geometric means, which the book has not yet proved; it is pure algebra, and it comes first.

Lemma 17.8.6. The Arithmetic–Geometric Mean Inequality

Let $n \geq 1$ and let $t_1, \ldots, t_n$ be **nonnegative** real numbers. Then

$$t_1 t_2 \cdots t_n \leq \Big(\frac{t_1 + t_2 + \cdots + t_n}{n}\Big)^n,$$

with equality if and only if $t_1 = t_2 = \cdots = t_n$.

Idea. Scale so that the mean is 1. If the numbers are not all 1, one of them lies above 1 and another below. Merge that pair into the single number $t_j + t_k - 1$. This leaves $n - 1$ nonnegative numbers with mean 1, so induction applies to them. The merge can only increase the product, because

$$(t_j + t_k - 1) - t_j t_k = (t_j - 1)(1 - t_k),$$

and both factors are positive. So the original product is at most, and when it is nonzero strictly less than, a product that induction bounds by 1.

Proof. **Step 1: mean 1.** We prove by induction on $n \geq 1$: if $t_1, \ldots, t_n \geq 0$ and $t_1 + \cdots + t_n = n$, then $t_1 \cdots t_n \leq 1$, with equality only when every $t_i = 1$. For $n = 1$ the hypothesis says $t_1 = 1$, and the product is 1.

Let $n \geq 2$ and assume the claim for $n - 1$ numbers. If every $t_i = 1$, the product is 1. Otherwise some $t_j > 1$: if every $t_i \leq 1$ with at least one strictly smaller, the sum would be less than n. Similarly some $t_k < 1$, since if every $t_i \geq 1$ with at least one strictly larger, the sum would exceed n. Then $j \neq k$. Neither the sum nor the product depends on the order of the numbers, so we may relabel so that $j = n - 1$ and $k = n$. Put

$$s = t_{n-1} + t_n - 1.$$

Then $s > 0$, since $t_{n-1} > 1$ and $t_n \geq 0$. The $n - 1$ numbers $t_1, \ldots, t_{n-2}, s$ are nonnegative and add up to $n - 1$. So by the inductive hypothesis $p\, s \leq 1$, where $p = t_1 \cdots t_{n-2} \geq 0$ (with $p = 1$ when $n = 2$). Moreover

$$s - t_{n-1} t_n = (t_{n-1} - 1)(1 - t_n) > 0.$$

If $p = 0$, then $t_1 \cdots t_n = 0 < 1$. If $p > 0$, then $t_1 \cdots t_n = p\, t_{n-1} t_n < p\, s \leq 1$. In either case the product is strictly less than 1 when the t_i are not all 1. This completes the induction.

Step 2: the general case. Let $m = \frac{1}{n}(t_1 + \cdots + t_n) \geq 0$. If $m = 0$, then every $t_i = 0$, since nonnegative numbers with sum 0 are all 0. Both sides are then 0, and the t_i are all equal. If $m > 0$, the numbers t_i / m are nonnegative with sum n. By Step 1,

$$\frac{t_1}{m} \cdots \frac{t_n}{m} \leq 1, \qquad \text{that is,} \qquad t_1 \cdots t_n \leq m^n,$$

with equality if and only if every $t_i / m = 1$, that is, every $t_i = m$. Numbers that are all equal to some c have mean $m = c$, so "every $t_i = m$" is the same as "all the t_i are equal". This proves the lemma.

■

In the language of this section, the lemma says that among nonnegative vectors with a fixed total, the **flat** vector, which is the bottom of the majorization order (Example (A chain, the two extremes, and size one) (b)), has the largest product.

Corollary 17.8.7. Hadamard's Inequality, Second Proof

Let $F = \mathbb{R}$ or $F = \mathbb{C}$, and let $\mathbf{A} \in M_n(F)$ with $\mathbf{A} \succeq \mathbf{O}$. Then

$$\det \mathbf{A} \leq a_{11}a_{22}\cdots a_{nn},$$

with equality if and only if $\mathbf{A}$ is diagonal or some $a_{ii} = 0$.

This is Theorem 13.7.3 (a) again, with the same hypothesis and the same equality condition. What is new is the proof.

Idea. Suppose first that every diagonal entry equals 1. Then $\operatorname{tr}\mathbf{A} = n$, so the eigenvalues, which are nonnegative, add up to n, and Lemma 17.8.6 gives $\det\mathbf{A} = \prod_i \lambda_i(\mathbf{A}) \leq 1 = \prod_i a_{ii}$. A general diagonal is made flat by a diagonal congruence $\mathbf{A} \mapsto \mathbf{SAS}$ with $\mathbf{S} = \operatorname{diag}(a_{11}^{-1/2}, \ldots, a_{nn}^{-1/2})$. This congruence preserves positivity and divides both sides of the inequality by the same number $a_{11}\cdots a_{nn}$. Zero diagonal entries, where $\mathbf{S}$ does not exist, are handled first and separately.

Proof. Each $a_{ii} = \langle \mathbf{A}\mathbf{e}_i, \mathbf{e}_i\rangle \geq 0$ by Definition 13.1.1, and each $\lambda_i(\mathbf{A}) \geq 0$ by Theorem 13.1.4 (b).

Case 1: some $a_{ii} = 0$. The right-hand side is 0. Suppose $\det\mathbf{A} \neq 0$. By Theorem 9.3.3, $\det\mathbf{A} = \prod_j \lambda_j(\mathbf{A})$, so every $\lambda_j(\mathbf{A})$ is nonzero, hence positive. By Theorem 13.1.5 (b), $\mathbf{A} \succ \mathbf{O}$, so $a_{ii} = \langle \mathbf{A}\mathbf{e}_i, \mathbf{e}_i\rangle > 0$, a contradiction. Hence $\det\mathbf{A} = 0$, and the inequality holds with equality, as the statement says it should when some $a_{ii} = 0$.

Case 2: every $a_{ii} > 0$. Let $\mathbf{S} = \operatorname{diag}(a_{11}^{-1/2}, \ldots, a_{nn}^{-1/2})$, a real diagonal matrix with nonzero diagonal, so $\mathbf{S}^* = \mathbf{S}$ and $\mathbf{S}$ is invertible. Put $\mathbf{B} = \mathbf{S}^*\mathbf{A}\mathbf{S} = \mathbf{SAS}$. By Proposition 13.1.2 (a), $\mathbf{B} \succeq \mathbf{O}$. Its entries are $b_{ij} = a_{ij}/\sqrt{a_{ii}a_{jj}}$, so $b_{ii} = 1$ for every i, and $\operatorname{tr}\mathbf{B} = n$. By Theorem 9.3.3,

$$\lambda_1(\mathbf{B}) + \cdots + \lambda_n(\mathbf{B}) = n.$$

The $\lambda_i(\mathbf{B})$ are nonnegative by Theorem 13.1.4 (b). So Lemma 17.8.6 and Theorem 9.3.3 give

$$\det\mathbf{B} = \prod_{i=1}^{n} \lambda_i(\mathbf{B}) \leq \left(\frac{n}{n}\right)^n = 1,$$

with equality if and only if all $\lambda_i(\mathbf{B})$ are equal, hence all equal to 1. By Theorem 7.5.4, and since $\det\mathbf{S} = (a_{11}\cdots a_{nn})^{-1/2}$,

$$\det\mathbf{B} = (\det\mathbf{S})^2 \det\mathbf{A} = \frac{\det\mathbf{A}}{a_{11}a_{22}\cdots a_{nn}}.$$

Multiplying $\det\mathbf{B} \leq 1$ by the positive number $a_{11}\cdots a_{nn}$ gives $\det\mathbf{A} \leq a_{11}\cdots a_{nn}$.

For equality in Case 2: equality holds exactly when every $\lambda_i(\mathbf{B}) = 1$. If so, the spectral theorem (Corollary 12.4.3, or Corollary 12.5.2 over $\mathbb{R}$) gives $\mathbf{B} = \mathbf{UIU}^* = \mathbf{I}$ for some unitary

(or orthogonal) $\mathbf{U}$. Then $\mathbf{A} = \mathbf{S}^{-1}\mathbf{B}\mathbf{S}^{-1} = \mathbf{S}^{-2} = \operatorname{diag}(a_{11}, \dots, a_{nn})$ is diagonal. Conversely, if $\mathbf{A}$ is diagonal, then $\mathbf{B} = \mathbf{S}\mathbf{A}\mathbf{S} = \mathbf{I}$, and every $\lambda_i(\mathbf{B}) = 1$. So in Case 2, equality holds if and only if $\mathbf{A}$ is diagonal. Together with Case 1, this is the stated equality condition. This proves the corollary.

■

For the matrix of Example (Coordinate vectors against eigenvectors), which is positive definite since its eigenvalues are $7, 4, 1$, the inequality reads $\det \mathbf{A} = 7 \cdot 4 \cdot 1 = 28 \le 60 = 3 \cdot 4 \cdot 5$.

The two proofs show different things. The first proves more along the way: Fischer's inequality for blocks of any sizes. The second shows that once the diagonal is scaled flat, the trace and the arithmetic–geometric mean inequality suffice. Of $\mathbf{d}(\mathbf{B}) \prec \boldsymbol{\lambda}(\mathbf{B})$ it uses only clause (M2), which for a matrix is just the trace.

Remark. A proof by majorization in earnest would use all of $\mathbf{d}(\mathbf{A}) \prec \boldsymbol{\lambda}(\mathbf{A})$, without scaling, together with the fact that for vectors with **nonnegative** entries, $\mathbf{x} \prec \mathbf{y}$ implies $x_1 \cdots x_n \ge y_1 \cdots y_n$ (the product is **Schur-concave**). That fact is true but **not proved in this book yet**; Chapter 21 proves it, for the whole class of functions that reverse majorization.

Majorization is the language this chapter's eigenvalue inequalities were already using, and the Schur–Horn theorem shows that it fits the diagonal problem exactly. The same language compresses the inequality of Section 7 for arbitrary index sets, Theorem 17.7.2, into one line. The k largest entries of the vector $\boldsymbol{\lambda}(\mathbf{A}) - \boldsymbol{\lambda}(\mathbf{B})$ add up to the largest value of $\sum_j (\lambda_{i_j}(\mathbf{A}) - \lambda_{i_j}(\mathbf{B}))$ over index sets of size k, and the totals are $\operatorname{tr}\mathbf{A} - \operatorname{tr}\mathbf{B} = \operatorname{tr}(\mathbf{A} - \mathbf{B})$. So a bound of every such sum by $\sum_{j \le k} \lambda_j(\mathbf{A} - \mathbf{B})$ says exactly that

$$\boldsymbol{\lambda}(\mathbf{A}) - \boldsymbol{\lambda}(\mathbf{B}) \prec \boldsymbol{\lambda}(\mathbf{A} - \mathbf{B}).$$

The diagonal of a Hermitian matrix and the eigenvalues of a difference are both controlled by one ordering of vectors in $\mathbb{R}^n$.

17.8.7 Exercises

17.8.7.1 A. Check your understanding

Exercise. A1

(a) State the definitions of $\mathbf{x} \prec \mathbf{y}$ and $\mathbf{x} \prec_w \mathbf{y}$ for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$.

(b) Decide whether $(2, 1, 1) \prec (2, 2, 0)$, and whether $(2, 2, 0) \prec (2, 1, 1)$.

(c) True or false: if $\mathbf{x} \prec \mathbf{y}$, then $x_1 \le y_1$. Justify your answer.

(d) State Theorem 17.8.2. Which result of §06 is it a restatement of?

(e) State Theorem 17.8.5. Which half is proved by induction on n, and which lemma does each step of that induction use?

Solution.

(a) $\mathbf{x} \prec \mathbf{y}$ means that $\sum_{i \le k} x_i^{\downarrow} \le \sum_{i \le k} y_i^{\downarrow}$ for every $k = 1, \dots, n-1$, and $\sum_{i=1}^n x_i = \sum_{i=1}^n y_i$ (Definition 17.8.1). $\mathbf{x} \prec_w \mathbf{y}$ means that the same inequality holds for every $k = 1, \dots, n$,

with no condition on the totals.

(b) Both vectors are decreasing with total 4. Their running totals at $k = 1, 2$ are $2, 3$ and $2, 4$. Since $2 \le 2$ and $3 \le 4$, $(2, 1, 1) \prec (2, 2, 0)$. The reverse fails at $k = 2$, since $4 > 3$.

(c) False. The definition compares **sorted** vectors, so the unsorted first entries can go either way. Take $\mathbf{x} = (1, 0)$ and $\mathbf{y} = (0, 1)$. Both have decreasing rearrangement $(1, 0)$, so (M1) at $k = 1$ reads $1 \le 1$, and both totals are 1; hence $\mathbf{x} \prec \mathbf{y}$. But $x_1 = 1 > 0 = y_1$. What is true is $x_1^{\downarrow} \le y_1^{\downarrow}$, which is (M1) at $k = 1$.

(d) For a Hermitian $\mathbf{A} \in M_n(F)$, $F = \mathbb{R}$ or $\mathbb{C}$, $\mathbf{d}(\mathbf{A}) \prec \boldsymbol{\lambda}(\mathbf{A})$. It restates Corollary 17.6.3: applied to the positions of the k largest diagonal entries, the corollary gives (M1), and its equality at $k = n$ gives (M2).

(e) For $\mathbf{d}, \boldsymbol{\lambda} \in \mathbb{R}^n$, there is a Hermitian matrix over $\mathbb{R}$ or $\mathbb{C}$ with diagonal $\mathbf{d}$, in the given order, and eigenvalues $\lambda_1, \dots, \lambda_n$ if and only if $\mathbf{d} \prec \boldsymbol{\lambda}$. The half "$\mathbf{d} \prec \boldsymbol{\lambda}$ implies such a matrix exists" is Horn's theorem, Theorem 17.8.4, proved by induction on n. Each step uses Lemma 17.8.3, applied to a pair of consecutive eigenvalues $\lambda_k \ge d_1 \ge \lambda_{k+1}$, to put d_1 in the corner. The other half is Schur's theorem.

17.8.7.2 B. Practice

Exercise. B1: Determine which pairs are comparable

For each pair, determine whether $\mathbf{x} \prec \mathbf{y}$ and whether $\mathbf{x} \prec_w \mathbf{y}$. Where a relation fails, name the exact clause and the value of k at which it fails. Justify your answer.

(a) $\mathbf{x} = (2, 2, 1, 1)$, $\mathbf{y} = (3, 2, 1, 0)$.

(b) $\mathbf{x} = (0, 3, 1)$, $\mathbf{y} = (4, 0, 0)$.

(c) $\mathbf{x} = (1, 1, 0)$, $\mathbf{y} = (2, 1, 0)$.

(d) $\mathbf{x} = (-1, 2, 2)$, $\mathbf{y} = (3, 0, 0)$.

(e) $\mathbf{x} = (3, 1, 1, 1)$, $\mathbf{y} = (2, 2, 2, 0)$; decide also whether $\mathbf{y} \prec \mathbf{x}$.

Solution. In each case, sort first, then compare running totals.

(a) Both are decreasing with total 6. Running totals: $2, 4, 5, 6$ and $3, 5, 6, 6$. Every entry of the first list is at most the matching entry of the second, so $\mathbf{x} \prec \mathbf{y}$ and hence $\mathbf{x} \prec_w \mathbf{y}$.

(b) $\mathbf{x}^{\downarrow} = (3, 1, 0)$ with running totals $3, 4, 4$, and $\mathbf{y}$ has running totals $4, 4, 4$. The totals agree and $3 \le 4$, $4 \le 4$, so $\mathbf{x} \prec \mathbf{y}$ and $\mathbf{x} \prec_w \mathbf{y}$. This is Example (A chain, the two extremes, and size one) (c), applied after sorting.

(c) Running totals $1, 2, 2$ and $2, 3, 3$. Every inequality holds, including at $k = 3$, so $\mathbf{x} \prec_w \mathbf{y}$. But the totals are $2 \ne 3$, so **(M2) fails** and $\mathbf{x} \not\prec \mathbf{y}$.

(d) $\mathbf{x}^{\downarrow} = (2, 2, -1)$ with running totals $2, 4, 3$, and $\mathbf{y}$ has running totals $3, 3, 3$. The totals agree, but at $k = 2$, $4 > 3$, so **(M1) fails at** $k = 2$. Neither $\mathbf{x} \prec \mathbf{y}$ nor $\mathbf{x} \prec_w \mathbf{y}$ holds.

(e) Both are decreasing with total 6. Running totals: $3, 4, 5, 6$ and $2, 4, 6, 6$. At $k = 1$, $3 > 2$, so **(M1) fails at** $k = 1$ and $\mathbf{x} \not\prec \mathbf{y}$, and also $\mathbf{x} \not\prec_w \mathbf{y}$. For $\mathbf{y} \prec \mathbf{x}$, the inequalities $2 \le 3$ and $4 \le 4$ hold, but $6 > 5$ at $k = 3$, so **(M1) fails at** $k = 3$. The pair is incomparable, like Example (Neither majorizes the other).

Exercise. B2: Schur's theorem for one matrix

Let

$$\mathbf{A} = \begin{pmatrix} 3 & 2 & 2 \\ 2 & 4 & 0 \\ 2 & 0 & 7 \end{pmatrix}.$$

(a) Verify that $(2,1,4)$, $(-1,-2,1)$, and $(-3,2,1)$ are eigenvectors of $\mathbf{A}$, and find $\boldsymbol{\lambda}(\mathbf{A})$.

(b) Write down $\mathbf{d}(\mathbf{A})$ and verify every clause of $\mathbf{d}(\mathbf{A}) \prec \boldsymbol{\lambda}(\mathbf{A})$.

(c) Hence verify Hadamard's inequality (Theorem 13.7.3) for $\mathbf{A}$.

Solution.

(a) Multiply out:

$$\begin{aligned} \mathbf{A}(2,1,4) &= (6+2+8,\ 4+4,\ 4+28) = (16,8,32) = 8\,(2,1,4), \\ \mathbf{A}(-1,-2,1) &= (-3-4+2,\ -2-8,\ -2+7) = (-5,-10,5) = 5\,(-1,-2,1), \\ \mathbf{A}(-3,2,1) &= (-9+4+2,\ -6+8,\ -6+7) = (-3,2,1) = 1\,(-3,2,1). \end{aligned}$$

Three distinct eigenvalues of a 3×3 matrix account for the whole spectrum, so $\boldsymbol{\lambda}(\mathbf{A}) = (8,5,1)$.

(b) $\mathbf{d}(\mathbf{A}) = (3,4,7)$, and $\mathbf{d}(\mathbf{A})^{\downarrow} = (7,4,3)$. (M1): at $k=1$, $7 \le 8$; at $k=2$, $7+4 = 11 \le 13 = 8+5$. (M2): $3+4+7 = 14 = 8+5+1$. Hence $\mathbf{d}(\mathbf{A}) \prec \boldsymbol{\lambda}(\mathbf{A})$, as Theorem 17.8.2 requires.

(c) All eigenvalues are positive, so $\mathbf{A} \succ \mathbf{O}$ by Theorem 13.1.5 (b), and Corollary 17.8.7 applies. By Theorem 9.3.3, $\det \mathbf{A} = 8 \cdot 5 \cdot 1 = 40$, while $a_{11}a_{22}a_{33} = 3 \cdot 4 \cdot 7 = 84$. Hence $40 \le 84$. The inequality is strict, as it must be, since $\mathbf{A}$ is not diagonal and no diagonal entry is 0.

Exercise. B3: Which diagonals are ruled out?

A Hermitian 3×3 matrix has eigenvalues $5, 1, 0$. For each proposed list of diagonal entries, determine whether Theorem 17.8.2 rules it out. Justify your answer.

(a) $4, 2, 0$.

(b) $6, 0, 0$.

(c) $3, 3, 1$.

(d) $2, 2, 2$.

(e) $-1, 4, 3$.

For the lists that are not ruled out, what does Theorem 17.8.5 add?

Solution. Here $\boldsymbol{\lambda} = (5,1,0)$, with running totals $5, 6$ and total 6.

(a) Running totals $4, 6$, total 6. Since $4 \le 5$ and $6 \le 6$, the list is **not** ruled out.

(b) At $k=1$, $6 > 5$. **Ruled out.** No diagonal entry can exceed λ_1.

(c) The total is $7 \neq 6$. **Ruled out** by (M2): the trace would be wrong.

(d) This is the flat vector with total 6, which is majorized by every vector with total 6 (Example (A chain, the two extremes, and size one) (b)). **Not** ruled out.

(e) Sorted, it is $(4, 3, -1)$, with running totals $4, 7$ and total 6. At $k = 2$, $7 > 6$. **Ruled out.** Independently, $-1 < 0 = \lambda_3$ contradicts the bottom-end inequality of the warning.

Schur's theorem is only a necessary condition, so for (a) and (d) it shows only that nothing forbids these diagonals. Theorem 17.8.5 adds that they occur: for each of them there is a real symmetric matrix with eigenvalues $5, 1, 0$ and that diagonal.

17.8.7.3 C. Going deeper

Exercise. C1: Majorization squeezes the range

(a) Prove that if $\mathbf{x} \prec \mathbf{y}$, then $y_n^{\downarrow} \le x_n^{\downarrow} \le x_1^{\downarrow} \le y_1^{\downarrow}$.

(b) Deduce that every diagonal entry of a Hermitian $\mathbf{A}$ lies in $[\lambda_n(\mathbf{A}), \lambda_1(\mathbf{A})]$.

(c) Show that the conclusion of (b), together with equal totals, does **not** imply majorization. Prove that no Hermitian matrix has eigenvalues $4, 1, 1, 0$ and diagonal entries $3, 3, 0, 0$, although every one of these entries lies in $[0, 4]$ and the totals agree.

Solution.

(a) The inequality $x_1^{\downarrow} \le y_1^{\downarrow}$ is (M1) at $k = 1$ (for $n = 1$, $\mathbf{x} = \mathbf{y}$ by Example (A chain, the two extremes, and size one) (d), and there is nothing to prove). The middle inequality holds because $\mathbf{x}^{\downarrow}$ is decreasing. For the left inequality, subtract (M1) at $k = n-1$ from (M2):

$$\begin{aligned} x_n^{\downarrow} &= \sum_{i=1}^{n} x_i - \sum_{i=1}^{n-1} x_i^{\downarrow} \\ &\ge \sum_{i=1}^{n} y_i - \sum_{i=1}^{n-1} y_i^{\downarrow} = y_n^{\downarrow}. \end{aligned}$$

(b) By Theorem 17.8.2, $\mathbf{d}(\mathbf{A}) \prec \boldsymbol{\lambda}(\mathbf{A})$. Apply (a) with $\mathbf{x} = \mathbf{d}(\mathbf{A})$ and $\mathbf{y} = \boldsymbol{\lambda}(\mathbf{A})$, using $\boldsymbol{\lambda}(\mathbf{A})^{\downarrow} = \boldsymbol{\lambda}(\mathbf{A})$. Every diagonal entry lies between the smallest and the largest diagonal entry, which (a) places between $\lambda_n(\mathbf{A})$ and $\lambda_1(\mathbf{A})$. (This also follows from Lemma 13.5.2 at $\mathbf{x} = \mathbf{e}_i$. The point of (a) is that it is a property of majorization itself.)

(c) The vector $(3, 3, 0, 0)$ has running totals $3, 6, 6$, and $(4, 1, 1, 0)$ has $4, 5, 6$. At $k = 2$, $6 > 5$, so $(3, 3, 0, 0) \not\prec (4, 1, 1, 0)$, and by Theorem 17.8.2 no Hermitian matrix has this diagonal and this spectrum. Yet $0 \le 3, 3, 0, 0 \le 4$, and both totals are 6. The range condition controls only the running totals at $k = 1$ and $k = n - 1$, while majorization constrains the middle ones too, which is why the example needs $n \ge 4$.

Exercise. C2: A weaker determinant bound

Let $\mathbf{A} \in M_n(F)$ with $\mathbf{A} \succeq \mathbf{O}$.

(a) Prove that $\det \mathbf{A} \le \left(\frac{1}{n} \operatorname{tr} \mathbf{A}\right)^n$, with equality if and only if $\mathbf{A} = c\mathbf{I}$ for some real $c \ge 0$.

(b) Prove that $a_{11} \cdots a_{nn} \le \left(\frac{1}{n} \operatorname{tr} \mathbf{A}\right)^n$. Deduce that Hadamard's inequality (Theorem 13.7.3) always implies the bound in (a).

(c) For

$$\mathbf{A} = \begin{pmatrix} 2 & 0 & 2 \\ 0 & 5 & 2 \\ 2 & 2 & 6 \end{pmatrix},$$

whose eigenvectors are $(1,2,3)$, $(1,-2,1)$ and $(-4,-1,2)$, compute the three numbers $\det \mathbf{A}$, $a_{11}a_{22}a_{33}$ and $\left(\frac{1}{3}\operatorname{tr}\mathbf{A}\right)^3$.

Solution.

(a) The eigenvalues $\lambda_i(\mathbf{A})$ are nonnegative by Theorem 13.1.4 (b). By Theorem 9.3.3 and Lemma 17.8.6,

$$\det \mathbf{A} = \prod_i \lambda_i(\mathbf{A}) \le \Big(\frac{1}{n}\sum_i \lambda_i(\mathbf{A})\Big)^n = \Big(\frac{1}{n}\operatorname{tr}\mathbf{A}\Big)^n.$$

Equality holds if and only if all $\lambda_i(\mathbf{A})$ equal a common value $c \ge 0$. By the spectral theorem this happens if and only if $\mathbf{A} = \mathbf{U}(c\mathbf{I})\mathbf{U}^* = c\mathbf{I}$. Conversely, $\mathbf{A} = c\mathbf{I}$ gives equality directly.

(b) The a_{ii} are nonnegative (as in the proof of Corollary 17.8.7) and add up to $\operatorname{tr}\mathbf{A}$. So Lemma 17.8.6 gives $\prod_i a_{ii} \le (\frac{1}{n}\operatorname{tr}\mathbf{A})^n$. Combined with Corollary 17.8.7, this gives $\det\mathbf{A} \le \prod_i a_{ii} \le (\frac{1}{n}\operatorname{tr}\mathbf{A})^n$, so the Hadamard bound always sits between $\det\mathbf{A}$ and the bound of (a). The bound in (a) sees only the total of the diagonal, while Hadamard's bound sees each entry, and so it is never worse.

(c) Check the eigenvectors: $\mathbf{A}(1,2,3) = (2{+}6,\ 10{+}6,\ 2{+}4{+}18) = (8,16,24)$; $\mathbf{A}(1,-2,1) = (2{+}2,\ -10{+}2,\ 2{-}4{+}6) = (4,-8,4)$; $\mathbf{A}(-4,-1,2) = (-8{+}4,\ -5{+}4,\ -8{-}2{+}12) = (-4,-1,2)$. So $\boldsymbol{\lambda}(\mathbf{A}) = (8,4,1)$, and $\mathbf{A} \succ \mathbf{O}$. Then

$$\det\mathbf{A} = 8\cdot 4\cdot 1 = 32, \qquad a_{11}a_{22}a_{33} = 2\cdot 5\cdot 6 = 60,$$
$$\Big(\frac{\operatorname{tr}\mathbf{A}}{3}\Big)^3 = \Big(\frac{13}{3}\Big)^3 = \frac{2197}{27} \approx 81.4,$$

and $32 \le 60 \le 81.4$, with both inequalities strict.

Exercise. C3: The possible diagonals form a convex set

For $\mathbf{x} \in \mathbb{R}^n$ and $1 \le k \le n$, write $s_k(\mathbf{x}) = x_1^{\downarrow} + \cdots + x_k^{\downarrow}$.

(a) Prove that $s_k(\mathbf{x}) = \max\{\sum_{i\in I} x_i : I \subseteq \{1,\dots,n\},\ |I| = k\}$.

(b) Deduce that $s_k(\mathbf{x}+\mathbf{y}) \le s_k(\mathbf{x}) + s_k(\mathbf{y})$ for all $\mathbf{x},\mathbf{y} \in \mathbb{R}^n$.

(c) Let $\boldsymbol{\lambda} \in \mathbb{R}^n$. Prove that if $\mathbf{d} \prec \boldsymbol{\lambda}$ and $\mathbf{d}' \prec \boldsymbol{\lambda}$, then $t\mathbf{d} + (1-t)\mathbf{d}' \prec \boldsymbol{\lambda}$ for every $0 \le t \le 1$.

(d) Hence show that if $\mathbf{A}$ and $\mathbf{B}$ are Hermitian $n \times n$ matrices with the same eigenvalues, then for every $0 \le t \le 1$ there is a real symmetric matrix with those eigenvalues and diagonal $t\,\mathbf{d}(\mathbf{A}) + (1-t)\,\mathbf{d}(\mathbf{B})$.

Hint for (a): if $v_1 \ge \cdots \ge v_k$ are the entries x_i, $i \in I$, in decreasing order, count how many entries of $\mathbf{x}$ are at least v_j.

Solution.

(a) Let $|I| = k$, and list the entries x_i, $i \in I$, decreasingly as $v_1 \geq \cdots \geq v_k$. For each $j \leq k$, the entries $v_1, \ldots, v_j$ sit at j different positions of $\mathbf{x}$ and are all $\geq v_j$. If $x_j^{\downarrow} < v_j$, then, since $\mathbf{x}^{\downarrow}$ is decreasing, only $x_1^{\downarrow}, \ldots, x_{j-1}^{\downarrow}$ could be $\geq v_j$, which is fewer than j entries. Hence $v_j \leq x_j^{\downarrow}$ for every j, and summing gives $\sum_{i \in I} x_i \leq s_k(\mathbf{x})$. The bound is attained by the positions of the k largest entries, chosen as in the proof of Theorem 17.8.2. This proves (a).

(b) Let I be a set of k positions at which $\mathbf{x} + \mathbf{y}$ attains the maximum in (a). Applying (a) three times,

$$s_k(\mathbf{x} + \mathbf{y}) = \sum_{i \in I} x_i + \sum_{i \in I} y_i \leq s_k(\mathbf{x}) + s_k(\mathbf{y}).$$

(c) Put $\mathbf{z} = t\mathbf{d} + (1-t)\mathbf{d}'$. Its total is $t\sum_i \lambda_i + (1-t)\sum_i \lambda_i = \sum_i \lambda_i$, by (M2) for both hypotheses. For a scalar $c \geq 0$, multiplying by c does not change the order of the entries, so $s_k(c\mathbf{x}) = c\, s_k(\mathbf{x})$. For $1 \leq k \leq n-1$, (b) and (M1) for both hypotheses give

$$s_k(\mathbf{z}) \leq t\, s_k(\mathbf{d}) + (1-t)\, s_k(\mathbf{d}') \leq t\, s_k(\boldsymbol{\lambda}) + (1-t)\, s_k(\boldsymbol{\lambda}) = s_k(\boldsymbol{\lambda}).$$

So $\mathbf{z} \prec \boldsymbol{\lambda}$.

(d) Let $\boldsymbol{\lambda} = \boldsymbol{\lambda}(\mathbf{A}) = \boldsymbol{\lambda}(\mathbf{B})$. By Theorem 17.8.2, $\mathbf{d}(\mathbf{A}) \prec \boldsymbol{\lambda}$ and $\mathbf{d}(\mathbf{B}) \prec \boldsymbol{\lambda}$. By (c), $t\,\mathbf{d}(\mathbf{A}) + (1-t)\,\mathbf{d}(\mathbf{B}) \prec \boldsymbol{\lambda}$. By Theorem 17.8.4 there is a real symmetric matrix with eigenvalues $\boldsymbol{\lambda}$ and exactly this diagonal. So among Hermitian matrices with a fixed spectrum, the set of diagonals contains the segment between any two of its points: it is **convex**.

17.9 The Same Theory for Singular Values

Everything so far in this chapter has been about Hermitian matrices, and it had to be: a general square matrix need not have real eigenvalues to order, and a rectangular matrix has none at all. But every matrix has singular values, they are real and nonnegative, and Chapter 13 listed them decreasingly from the moment it defined them. This section carries the apparatus across — a min–max description, a perturbation bound, an interlacing theorem — and then spends the rest of its length paying debts that earlier chapters recorded and could not settle.

Throughout, $F = \mathbb{R}$ **or** $F = \mathbb{C}$, $\mathbf{A} \in M_{m \times n}(F)$, $p = \min(m, n)$, and $\sigma_1(\mathbf{A}) \geq \cdots \geq \sigma_p(\mathbf{A}) \geq 0$ are the singular values of $\mathbf{A}$ (Definition 13.9.1). Nothing here is assumed Hermitian and nothing is assumed square; where a matrix must be square, or invertible, the statement says so. Vector norms are Euclidean, and $\|\mathbf{A}\|_2$ is the spectral norm of Definition 16.3.2, which Theorem 16.3.3 (c) computes as $\sigma_1(\mathbf{A})$.

One convention, stated once and used everywhere below. We set

$$\sigma_i(\mathbf{A}) \coloneqq 0 \qquad \text{for } p < i \leq n.$$

This is not new: Theorem 13.9.2 (a) already makes it, and records what it buys. With it in force, if $\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^*$ is a singular value decomposition (Theorem 13.9.1) with columns $\mathbf{v}_1, \ldots, \mathbf{v}_n$ of $\mathbf{V}$, then

$$\mathbf{A}^*\mathbf{A}\,\mathbf{v}_i = \sigma_i(\mathbf{A})^2\,\mathbf{v}_i \qquad (1 \leq i \leq n), \tag{17.9.1}$$

so that $\mathbf{v}_1, \dots, \mathbf{v}_n$ is an orthonormal eigenbasis of $\mathbf{A}^*\mathbf{A}$ whose eigenvalues $\sigma_1^2 \geq \cdots \geq \sigma_n^2 \geq 0$ are already sorted. In the notation of this chapter,

$$\lambda_i(\mathbf{A}^*\mathbf{A}) = \sigma_i(\mathbf{A})^2 \qquad (1 \leq i \leq n).$$

Every statement in this section is that identity, pushed through a theorem about Hermitian matrices.

17.9.1 The stretch ratio is a Rayleigh quotient

Chapter 13 introduced the singular values by an algebraic route — take $\mathbf{A}^*\mathbf{A}$, diagonalize it, take square roots. Their geometric meaning is stretching: how much longer $\mathbf{Ax}$ is than $\mathbf{x}$. The bridge between the two descriptions is one line of algebra, and it is the only computation this section needs before the theorems start.

For $\mathbf{x} \neq \mathbf{0}$ in F^n, the matrix $\mathbf{A}^*\mathbf{A}$ is Hermitian, so its Rayleigh quotient (Definition 17.1.1) is defined, and

$$R_{\mathbf{A}^*\mathbf{A}}(\mathbf{x}) = \frac{\langle \mathbf{A}^*\mathbf{Ax}, \mathbf{x}\rangle}{\langle \mathbf{x}, \mathbf{x}\rangle} = \frac{\mathbf{x}^*\mathbf{A}^*\mathbf{Ax}}{\mathbf{x}^*\mathbf{x}} = \frac{(\mathbf{Ax})^*(\mathbf{Ax})}{\mathbf{x}^*\mathbf{x}} = \frac{\|\mathbf{Ax}\|^2}{\|\mathbf{x}\|^2}. \tag{17.9.2}$$

The Rayleigh quotient of $\mathbf{A}^*\mathbf{A}$ is the squared stretching factor of $\mathbf{A}$.

So a statement about the Rayleigh quotient of $\mathbf{A}^*\mathbf{A}$ is a statement about the function $\mathbf{x} \mapsto \|\mathbf{Ax}\|/\|\mathbf{x}\|$, and conversely. Nothing is lost in passing between them, because $t \mapsto \sqrt{t}$ is a strictly increasing bijection of $[0, \infty)$ onto itself: it carries maxima to maxima and minima to minima, over any set whatever.

Before the min–max theorem, it is worth writing down what the singular value decomposition says about this function in coordinates. The computation is three lines, and it is used four times below — for the witness subspaces in every proof of the section — so it gets a label.

Lemma 17.9.1. Stretching in the Right Singular Basis

Let $\mathbf{A} \in M_{m\times n}(F)$ have a singular value decomposition $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$ (Theorem 13.9.1), with $\mathbf{v}_1, \dots, \mathbf{v}_n$ the columns of $\mathbf{V}$, and write $\sigma_i = \sigma_i(\mathbf{A})$ with the convention above. Then for $\mathbf{x} = \sum_{i=1}^n c_i \mathbf{v}_i$,

$$\|\mathbf{Ax}\|^2 = \sum_{i=1}^n \sigma_i^2 \, |c_i|^2, \qquad \|\mathbf{x}\|^2 = \sum_{i=1}^n |c_i|^2.$$

Consequently, for each $1 \leq k \leq n$:

(a) on the **tail** subspace $T_k = \operatorname{span}(\mathbf{v}_k, \dots, \mathbf{v}_n)$, of dimension $n - k + 1$,

$$\max\left\{ \tfrac{\|\mathbf{Ax}\|}{\|\mathbf{x}\|} : \mathbf{0} \neq \mathbf{x} \in T_k \right\} = \sigma_k,$$

attained at $\mathbf{x} = \mathbf{v}_k$;

(b) on the **head** subspace $H_k = \operatorname{span}(\mathbf{v}_1, \dots, \mathbf{v}_k)$, of dimension k,

$$\min\left\{ \tfrac{\|\mathbf{Ax}\|}{\|\mathbf{x}\|} : \mathbf{0} \neq \mathbf{x} \in H_k \right\} = \sigma_k,$$

attained at $\mathbf{x} = \mathbf{v}_k$.

Idea. In the right singular basis $\mathbf{A}$ does nothing but scale: it sends $\mathbf{v}_i$ to $\sigma_i \mathbf{u}_i$, and the $\mathbf{u}_i$ are orthonormal. So measuring $\mathbf{Ax}$ is Pythagoras applied to the scaled coordinates, and every claim about how much $\mathbf{A}$ can stretch a subspace becomes a claim about a weighted sum of $|c_i|^2$.

Proof. By Eq. 17.9.1 the list $(\mathbf{v}_1, \dots, \mathbf{v}_n)$ is an orthonormal basis of F^n consisting of eigenvectors of $\mathbf{A}^*\mathbf{A}$, with $\mathbf{A}^*\mathbf{A}\mathbf{v}_i = \sigma_i^2 \mathbf{v}_i$. Hence, for $\mathbf{x} = \sum_i c_i \mathbf{v}_i$, using Eq. 17.9.2 in the form $\|\mathbf{Ax}\|^2 = \mathbf{x}^*\mathbf{A}^*\mathbf{Ax}$ and expanding with orthonormality,

$$\|\mathbf{Ax}\|^2 = \Big(\sum_i c_i \mathbf{v}_i\Big)^* \Big(\sum_j c_j \sigma_j^2 \mathbf{v}_j\Big) = \sum_{i=1}^{n} \sigma_i^2 |c_i|^2,$$

the cross terms vanishing because $\mathbf{v}_i^* \mathbf{v}_j = \delta_{ij}$. The second identity is the same expansion with $\mathbf{A}$ removed.

(a) For $\mathbf{0} \neq \mathbf{x} \in T_k$ we have $c_i = 0$ for $i < k$, so $\|\mathbf{Ax}\|^2 = \sum_{i \geq k} \sigma_i^2 |c_i|^2 \leq \sigma_k^2 \sum_{i \geq k} |c_i|^2 = \sigma_k^2 \|\mathbf{x}\|^2$, since $\sigma_i \leq \sigma_k$ for $i \geq k$ and the singular values are nonnegative. Taking square roots gives $\|\mathbf{Ax}\|/\|\mathbf{x}\| \leq \sigma_k$, and $\mathbf{x} = \mathbf{v}_k$ gives $\|\mathbf{Av}_k\| = \sigma_k = \sigma_k \|\mathbf{v}_k\|$, so the bound is attained.

(b) For $\mathbf{0} \neq \mathbf{x} \in H_k$ we have $c_i = 0$ for $i > k$, so $\|\mathbf{Ax}\|^2 = \sum_{i \leq k} \sigma_i^2 |c_i|^2 \geq \sigma_k^2 \|\mathbf{x}\|^2$, since now $\sigma_i \geq \sigma_k$ for $i \leq k$; and $\mathbf{x} = \mathbf{v}_k$ attains equality. This proves the lemma.

■

In words: the tail subspace is a place where $\mathbf{A}$ stretches **no more** than σ_k, and the head subspace is a place where it stretches **no less**. The min–max theorem says that these two subspaces are optimal.

17.9.2 Each singular value as an optimization

Theorem 17.9.2. Min–Max for Singular Values

Let $\mathbf{A} \in M_{m \times n}(F)$ and let $1 \leq k \leq n$, with $\sigma_k(\mathbf{A}) = 0$ read by convention when $k > p$. Then

$$\begin{aligned} \sigma_k(\mathbf{A}) &= \min_{\substack{W \leq F^n \\ \dim W = n-k+1}} \; \max_{\mathbf{0} \neq \mathbf{x} \in W} \frac{\|\mathbf{Ax}\|}{\|\mathbf{x}\|} \\ &= \max_{\substack{W \leq F^n \\ \dim W = k}} \; \min_{\mathbf{0} \neq \mathbf{x} \in W} \frac{\|\mathbf{Ax}\|}{\|\mathbf{x}\|}. \end{aligned}$$

The outer minimum is attained at $W = \operatorname{span}(\mathbf{v}_k, \dots, \mathbf{v}_n)$ and the outer maximum at $W = \operatorname{span}(\mathbf{v}_1, \dots, \mathbf{v}_k)$, for the right singular vectors of any singular value decomposition of $\mathbf{A}$ (Theorem 13.9.1).

Idea. There is nothing to discover: Eq. 17.9.2 turns the objective function into the Rayleigh quotient of the Hermitian matrix $\mathbf{A}^*\mathbf{A}$, Theorem 17.2.3 applies to that matrix, and the square root — increasing, hence order-preserving — moves through the minima and maxima

without changing which subspace attains them.

Proof. Put $\mathbf{B} = \mathbf{A}^*\mathbf{A} \in M_n(F)$, which is Hermitian, and fix $1 \le k \le n$. By Eq. 17.9.1, $\lambda_k(\mathbf{B}) = \sigma_k(\mathbf{A})^2$. By Theorem 17.2.3 applied to $\mathbf{B}$,

$$\sigma_k(\mathbf{A})^2 = \lambda_k(\mathbf{B}) = \min_{\dim W = n-k+1} \max_{\mathbf{0}\neq\mathbf{x}\in W} R_{\mathbf{B}}(\mathbf{x}),$$

and by Eq. 17.9.2, $R_{\mathbf{B}}(\mathbf{x}) = \left(\|\mathbf{A}\mathbf{x}\|/\|\mathbf{x}\|\right)^2$ for every $\mathbf{x} \neq \mathbf{0}$.

Fix a subspace W. The function $\mathbf{x} \mapsto \|\mathbf{A}\mathbf{x}\|/\|\mathbf{x}\|$ takes only nonnegative values, and $t \mapsto t^2$ is strictly increasing on $[0, \infty)$; so a vector of W maximizes $R_{\mathbf{B}}$ over W exactly when it maximizes $\|\mathbf{A}\mathbf{x}\|/\|\mathbf{x}\|$, and

$$\max_{\mathbf{0}\neq\mathbf{x}\in W} R_{\mathbf{B}}(\mathbf{x}) = \left(\max_{\mathbf{0}\neq\mathbf{x}\in W} \frac{\|\mathbf{A}\mathbf{x}\|}{\|\mathbf{x}\|}\right)^2.$$

(Both maxima exist: by homogeneity the quotient may be taken over the unit sphere of W, which is compact by fact (A3) of Chapter 16's introduction, and the function is continuous, so the extreme value theorem — fact (A4) of Chapter 16's introduction — applies. This is the same appeal Theorem 17.2.3 makes.) Taking square roots, and using once more that $t \mapsto \sqrt{t}$ is increasing, so that it commutes with a minimum over the family of subspaces,

$$\sigma_k(\mathbf{A}) = \sqrt{\lambda_k(\mathbf{B})} = \min_{\dim W = n-k+1} \max_{\mathbf{0}\neq\mathbf{x}\in W} \frac{\|\mathbf{A}\mathbf{x}\|}{\|\mathbf{x}\|}.$$

The max–min form follows from the second equality of Theorem 17.2.3 by the same two steps.

For the attainment claims, Lemma 17.9.1 (a) says that the tail subspace T_k, of dimension $n - k + 1$, has inner maximum exactly $\sigma_k(\mathbf{A})$, which is therefore the outer minimum; and part (b) says the head subspace H_k, of dimension k, has inner minimum exactly $\sigma_k(\mathbf{A})$. This proves the theorem.

■

Two things are worth noticing. First, the subspaces W live in F^n, the **domain**, whatever m is; the theorem compares two matrices with the same number of columns without caring how many rows they have. Second, the extremal subspaces are named, and they are spans of right singular vectors: the theorem is not only an identity but a recipe.

There is a second route to the same theorem, and it is worth recording because it converts a statement about a rectangular matrix into a statement about a Hermitian one, which is what the later chapters will want.

Proposition 17.9.3. The Hermitian Dilation

Let $\mathbf{A} \in M_{m\times n}(F)$ and put

$$\mathcal{H}(\mathbf{A}) \coloneqq \begin{pmatrix} \mathbf{O} & \mathbf{A} \\ \mathbf{A}^* & \mathbf{O} \end{pmatrix} \in M_{m+n}(F).$$

Then $\mathcal{H}(\mathbf{A})$ is Hermitian, and its eigenvalues, with multiplicity, are

$$\sigma_1(\mathbf{A}), \dots, \sigma_p(\mathbf{A}),\ -\sigma_1(\mathbf{A}), \dots, -\sigma_p(\mathbf{A}),\ \underbrace{0, \dots, 0}_{m+n-2p}.$$

Idea. Guess the eigenvectors from the SVD. The pair $(\mathbf{u}_i, \mathbf{v}_i)$, stacked, is sent by $\mathcal{H}(\mathbf{A})$ to $(\mathbf{A}\mathbf{v}_i, \mathbf{A}^*\mathbf{u}_i) = \sigma_i(\mathbf{u}_i, \mathbf{v}_i)$; flipping the sign of the second half flips the sign of the eigenvalue. That accounts for $2p$ orthonormal eigenvectors, and what is left over is a kernel whose dimension is a count.

Proof. Writing out the conjugate transpose block by block, $\mathcal{H}(\mathbf{A})^* = \left(\begin{smallmatrix} \mathbf{O} & (\mathbf{A}^*)^* \\ \mathbf{A}^* & \mathbf{O} \end{smallmatrix}\right) = \mathcal{H}(\mathbf{A})$, so $\mathcal{H}(\mathbf{A})$ is Hermitian.

Fix a singular value decomposition $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$ (Theorem 13.9.1), with columns $\mathbf{u}_1, \dots, \mathbf{u}_m$ of $\mathbf{U}$ and $\mathbf{v}_1, \dots, \mathbf{v}_n$ of $\mathbf{V}$. Since $\mathbf{V}^*\mathbf{v}_i = \mathbf{e}_i$ and $\mathbf{\Sigma}\mathbf{e}_i = \sigma_i\mathbf{e}_i$ for $i \le p$ while $\mathbf{\Sigma}\mathbf{e}_j = \mathbf{0}$ for $p < j \le n$, and symmetrically for $\mathbf{A}^* = \mathbf{V}\mathbf{\Sigma}^*\mathbf{U}^*$,

$$\mathbf{A}\mathbf{v}_i = \sigma_i\mathbf{u}_i, \quad \mathbf{A}^*\mathbf{u}_i = \sigma_i\mathbf{v}_i \ \ (i \le p), \qquad \mathbf{A}\mathbf{v}_j = \mathbf{0}, \quad \mathbf{A}^*\mathbf{u}_j = \mathbf{0} \ \ (j > p).$$

For $i \le p$ put

$$\mathbf{w}_i^{\pm} = \tfrac{1}{\sqrt{2}} \begin{pmatrix} \mathbf{u}_i \\ \pm\mathbf{v}_i \end{pmatrix} \in F^{m+n}.$$

Then $\mathcal{H}(\mathbf{A})\mathbf{w}_i^{\pm} = \frac{1}{\sqrt{2}}(\pm\mathbf{A}\mathbf{v}_i,\ \mathbf{A}^*\mathbf{u}_i) = \frac{1}{\sqrt{2}}(\pm\sigma_i\mathbf{u}_i,\ \sigma_i\mathbf{v}_i) = \pm\sigma_i\,\mathbf{w}_i^{\pm}$. Likewise $\mathcal{H}(\mathbf{A})(\mathbf{u}_j, \mathbf{0}) = (\mathbf{0}, \mathbf{A}^*\mathbf{u}_j) = \mathbf{0}$ for $j > p$, and $\mathcal{H}(\mathbf{A})(\mathbf{0}, \mathbf{v}_j) = (\mathbf{A}\mathbf{v}_j, \mathbf{0}) = \mathbf{0}$ for $j > p$.

These $2p + (m-p) + (n-p) = m+n$ vectors are orthonormal: the $\mathbf{w}_i^{\pm}$ are unit vectors because $\mathbf{u}_i$ and $\mathbf{v}_i$ are, and any two distinct vectors in the list pair to 0, since $(\mathbf{w}_i^{+})^*\mathbf{w}_i^{-} = \frac{1}{2}(1-1) = 0$ and all other pairs have orthogonal $\mathbf{u}$-parts or orthogonal $\mathbf{v}$-parts. So they form an orthonormal eigenbasis of F^{m+n} with the stated eigenvalues, which are therefore the eigenvalue list of $\mathcal{H}(\mathbf{A})$ with multiplicity. This proves the proposition.

■

Applying Theorem 17.2.3 to $\mathcal{H}(\mathbf{A})$ instead of $\mathbf{A}^*\mathbf{A}$ gives Theorem 17.9.2 again, with subspaces of F^{m+n} in place of F^n. We took the shorter road; the dilation is recorded because it turns *any* theorem about Hermitian eigenvalues into one about singular values, which Section 10's exercises and Chapter 20 both exploit.

17.9.3 Perturbation, with no hypothesis at all

For Hermitian matrices, Corollary 17.3.6 says that $|\lambda_i(\mathbf{A}+\mathbf{E}) - \lambda_i(\mathbf{A})| \le \|\mathbf{E}\|_2$ — provided $\mathbf{A}$ and $\mathbf{E}$ are Hermitian. That proviso is not decoration: Chapter 16's Example (A Jordan block's eigenvalues move like $\varepsilon^{1/k}$) shows the conclusion failing spectacularly without it. The corresponding statement for singular values needs no proviso whatsoever.

Theorem 17.9.4. Weyl's Inequalities for Singular Values

Let $\mathbf{A}, \mathbf{B} \in M_{m\times n}(F)$ and let $i, j \geq 1$ with $i + j - 1 \leq n$. Then

$$\sigma_{i+j-1}(\mathbf{A} + \mathbf{B}) \leq \sigma_i(\mathbf{A}) + \sigma_j(\mathbf{B}).$$

Idea. Put $k = i+j-1$. The min–max form of Theorem 17.9.2 bounds $\sigma_k(\mathbf{A}+\mathbf{B})$ from above as soon as we produce **one** subspace of dimension $n-k+1$ on which $\mathbf{A}+\mathbf{B}$ stretches by at most $\sigma_i(\mathbf{A})+\sigma_j(\mathbf{B})$. Each summand supplies a subspace where it is individually small — the tail subspaces of Lemma 17.9.1 — and the dimensions are chosen so that the two subspaces must overlap in something of the required size. On the overlap the triangle inequality finishes it.

Proof. Put $k = i + j - 1 \leq n$. Let $T = \operatorname{span}(\mathbf{v}_i, \dots, \mathbf{v}_n)$ for the right singular vectors of $\mathbf{A}$ and $T' = \operatorname{span}(\mathbf{v}'_j, \dots, \mathbf{v}'_n)$ for those of $\mathbf{B}$, so that $\dim T = n - i + 1$, $\dim T' = n - j + 1$, and by Lemma 17.9.1 (a),

$$\|\mathbf{A}\mathbf{x}\| \leq \sigma_i(\mathbf{A})\|\mathbf{x}\| \quad (\mathbf{x} \in T), \qquad \|\mathbf{B}\mathbf{x}\| \leq \sigma_j(\mathbf{B})\|\mathbf{x}\| \quad (\mathbf{x} \in T').$$

By the dimension formula (Theorem 1.7.3) and $\dim(T + T') \leq n$,

$$\begin{aligned} \dim(T \cap T') &\geq (n - i + 1) + (n - j + 1) - n \\ &= n - (i + j - 1) + 1 = n - k + 1, \end{aligned}$$

which is at least 1 because $k \leq n$. Choose a subspace $W \subseteq T \cap T'$ with $\dim W = n - k + 1$ exactly; this is possible because every subspace has a subspace of each smaller dimension, obtained by truncating a basis. For $\mathbf{0} \neq \mathbf{x} \in W$, the triangle inequality (Corollary 11.1.6) gives

$$\|(\mathbf{A} + \mathbf{B})\mathbf{x}\| \leq \|\mathbf{A}\mathbf{x}\| + \|\mathbf{B}\mathbf{x}\| \leq (\sigma_i(\mathbf{A}) + \sigma_j(\mathbf{B}))\|\mathbf{x}\|.$$

Hence the inner maximum over W is at most $\sigma_i(\mathbf{A}) + \sigma_j(\mathbf{B})$, and since $\sigma_k(\mathbf{A} + \mathbf{B})$ is the **minimum** of that inner maximum over all subspaces of dimension $n-k+1$ (Theorem 17.9.2), we get $\sigma_k(\mathbf{A} + \mathbf{B}) \leq \sigma_i(\mathbf{A}) + \sigma_j(\mathbf{B})$. This proves the theorem.

■

Corollary 17.9.5. Singular Values Are Perfectly Conditioned

Let $\mathbf{A}, \mathbf{E} \in M_{m\times n}(F)$. Then

$$|\sigma_i(\mathbf{A} + \mathbf{E}) - \sigma_i(\mathbf{A})| \leq \|\mathbf{E}\|_2 \qquad \text{for every } 1 \leq i \leq n.$$

No hypothesis is placed on $\mathbf{A}$ or on $\mathbf{E}$.

Idea. Put $j = 1$ in the Weyl inequality for singular values: one index passes through unchanged, and the other matrix contributes only its largest singular value, which is $\|\mathbf{E}\|_2$. That bounds $\sigma_i(\mathbf{A}+\mathbf{E})$ from above; writing $\mathbf{A} = (\mathbf{A}+\mathbf{E})+(-\mathbf{E})$ and doing it again bounds it from below.

Proof. Take $j = 1$ in Theorem 17.9.4, so that $i + j - 1 = i \le n$:

$$\sigma_i(\mathbf{A} + \mathbf{E}) \le \sigma_i(\mathbf{A}) + \sigma_1(\mathbf{E}) = \sigma_i(\mathbf{A}) + \|\mathbf{E}\|_2,$$

using Theorem 16.3.3 (c) for the last equality. Now apply the same inequality with $\mathbf{A} + \mathbf{E}$ in place of $\mathbf{A}$ and $-\mathbf{E}$ in place of $\mathbf{E}$, noting $\|-\mathbf{E}\|_2 = \|\mathbf{E}\|_2$ by homogeneity (Theorem 16.3.2 (b)):

$$\sigma_i(\mathbf{A}) = \sigma_i((\mathbf{A} + \mathbf{E}) + (-\mathbf{E})) \le \sigma_i(\mathbf{A} + \mathbf{E}) + \|\mathbf{E}\|_2.$$

The two inequalities together are the claim.

■

This is a Lipschitz bound with constant 1, exactly as Corollary 17.3.6 is for Hermitian eigenvalues, and with no dependence on n, on the conditioning of $\mathbf{A}$, or on the gaps between the singular values. What is new is the absence of hypotheses. It is worth seeing the contrast on the very matrix Chapter 16 used to show that eigenvalues have no such bound.

Example. The Jordan block again, measured by its singular values

Let $k \ge 2$, let $0 < \varepsilon < 1$, and let

$$\mathbf{A} = \mathbf{J}_k(0), \qquad \mathbf{A}_\varepsilon = \mathbf{J}_k(0) + \varepsilon \mathbf{E}_{k1}$$

be the matrices of Example (A Jordan block's eigenvalues move like $\varepsilon^{1/k}$). Compute the singular values of both, and compare how far they move with how far the eigenvalues move.

Solution. The columns of $\mathbf{A}_\varepsilon$ are $\varepsilon \mathbf{e}_k$ in position 1, and $\mathbf{e}_{j-1}$ in position j for $2 \le j \le k$. These k columns are pairwise orthogonal, with norms $\varepsilon, 1, \dots, 1$, so

$$\mathbf{A}_\varepsilon^* \mathbf{A}_\varepsilon = \operatorname{diag}(\varepsilon^2, 1, \dots, 1),$$

whose eigenvalues in decreasing order are 1 with multiplicity $k - 1$ and then ε^2, since $\varepsilon < 1$. By Definition 13.9.1 the singular values of $\mathbf{A}_\varepsilon$ are

$$\sigma_1 = \cdots = \sigma_{k-1} = 1, \qquad \sigma_k = \varepsilon.$$

Setting $\varepsilon = 0$ in the same computation, the singular values of $\mathbf{A} = \mathbf{J}_k(0)$ are $1, \dots, 1, 0$.

So every singular value moves by at most ε, and exactly one of them — the last — moves by exactly ε. The perturbation has $\|\varepsilon \mathbf{E}_{k1}\|_2 = \varepsilon$, computed in Example (A Jordan block's eigenvalues move like $\varepsilon^{1/k}$), so Corollary 17.9.5 is attained here and not merely satisfied.

Meanwhile Example (A Jordan block's eigenvalues move like $\varepsilon^{1/k}$) computed that the **eigenvalues** of the same two matrices are 0 (with multiplicity k) and the k-th roots of ε, all of modulus $\varepsilon^{1/k}$. For $k = 10$ and $\varepsilon = 10^{-10}$: the singular values move by 10^{-10}, and the eigenvalues move by 10^{-1}, a factor of a billion more. The same matrix, the same perturbation, two spectra — one of them stable and one of them not.

This is not a paradox. The singular values of $\mathbf{A}$ are the square roots of the eigenvalues of the Hermitian matrix $\mathbf{A}^*\mathbf{A}$, and Hermitian eigenvalues are stable; the eigenvalues of $\mathbf{A}$ itself are

roots of a polynomial, and a repeated root is fragile. The instability lives in the passage from $\mathbf{A}$ to its eigenvalues, not in $\mathbf{A}$ — which is why a question that can be phrased in singular values should be.

Warning. **Weyl's inequality does not say $\sigma_k(\mathbf{A}+\mathbf{B}) \le \sigma_k(\mathbf{A}) + \sigma_k(\mathbf{B})$.** The index bookkeeping $i+j-1$ is the content, not a technicality. Take

$$\mathbf{A} = \operatorname{diag}(1,0), \qquad \mathbf{B} = \operatorname{diag}(0,1).$$

Then $\sigma(\mathbf{A}) = \sigma(\mathbf{B}) = (1,0)$, while $\mathbf{A}+\mathbf{B} = \mathbf{I}_2$ has $\sigma(\mathbf{I}_2) = (1,1)$. So $\sigma_2(\mathbf{A}+\mathbf{B}) = 1$ while $\sigma_2(\mathbf{A}) + \sigma_2(\mathbf{B}) = 0$. What Theorem 17.9.4 does give here, with $i = j = 1$, is $\sigma_1(\mathbf{A}+\mathbf{B}) \le \sigma_1(\mathbf{A}) + \sigma_1(\mathbf{B}) = 2$, which holds. Only $i = 1$ or $j = 1$ lets an index pass through unchanged, which is why Corollary 17.9.5 is the case $j = 1$.

17.9.4 The best approximation of low rank

Chapter 13 §11 proved that the truncated singular value decomposition $\mathbf{A}_k$ of Definition 13.11.1 is the closest matrix of rank at most k to $\mathbf{A}$ in the Frobenius norm (Theorem 13.11.3), and then wrote, in the remark that closes its main proof:

> The same matrix $\mathbf{A}_k$ is also optimal in the **spectral norm** $\|\cdot\|_2$, where the minimum value is σ_{k+1} rather than $\left(\sum_{i>k}\sigma_i^2\right)^{1/2}$. We cannot state that here, because the spectral norm is defined only in Chapter 16.

Chapter 16 supplied the norm and then, in its closing ledger, handed the theorem to this chapter: "the theorem itself waits for Chapter 17, which proves it from the min–max description of the singular values." That description now exists. Here is the theorem.

Theorem 17.9.6. Eckart–Young Theorem, Spectral Case

Let $\mathbf{A} \in M_{m\times n}(F)$ and let $0 \le k < n$. Then for **every** $\mathbf{B} \in M_{m\times n}(F)$ with $\operatorname{rank}\mathbf{B} \le k$,

$$\|\mathbf{A}-\mathbf{B}\|_2 \;\ge\; \sigma_{k+1}(\mathbf{A}),$$

and equality holds for the truncation $\mathbf{B} = \mathbf{A}_k$ of Definition 13.11.1. In words: $\mathbf{A}_k$ minimizes $\|\mathbf{A}-\mathbf{B}\|_2$ over all matrices of rank at most k, and the minimum value is $\sigma_{k+1}(\mathbf{A})$.

Idea. A matrix $\mathbf{B}$ of rank at most k has a large null space — dimension at least $n-k$ — and on that null space $\mathbf{A}-\mathbf{B}$ acts exactly as $\mathbf{A}$ does. The head subspace $H_{k+1} = \operatorname{span}(\mathbf{v}_1,\dots,\mathbf{v}_{k+1})$ is a place where $\mathbf{A}$ stretches by at least σ_{k+1}, and it has dimension $k+1$. Two subspaces of F^n of dimensions $n-k$ and $k+1$ cannot avoid each other, by Lemma 17.2.1, and a vector in both settles the matter in one line.

Proof. The lower bound. Let $\operatorname{rank}\mathbf{B} \le k$. By Rank–Nullity for matrices (Theorem 3.5.7), $\dim\operatorname{nul}(\mathbf{B}) = n - \operatorname{rank}\mathbf{B} \ge n-k$. Let $H = \operatorname{span}(\mathbf{v}_1,\dots,\mathbf{v}_{k+1})$, where $\mathbf{v}_1,\dots,\mathbf{v}_n$ are the right singular vectors of a fixed singular value decomposition of $\mathbf{A}$ (Theorem 13.9.1); then $\dim H = k+1$. Since

$$\dim\operatorname{nul}(\mathbf{B}) + \dim H \;\ge\; (n-k)+(k+1) = n+1 > n,$$

Lemma 17.2.1 gives a vector $\mathbf{x} \neq \mathbf{0}$ with $\mathbf{x} \in \operatorname{nul}(\mathbf{B}) \cap H$. For that $\mathbf{x}$, $\mathbf{B}\mathbf{x} = \mathbf{0}$, so $(\mathbf{A}-\mathbf{B})\mathbf{x} = \mathbf{A}\mathbf{x}$, and by Lemma 17.9.1 (b) applied with $k+1$ in place of k,

$$\|(\mathbf{A}-\mathbf{B})\mathbf{x}\| = \|\mathbf{A}\mathbf{x}\| \;\geq\; \sigma_{k+1}(\mathbf{A})\,\|\mathbf{x}\|.$$

Since $\mathbf{x} \neq \mathbf{0}$, dividing by $\|\mathbf{x}\|$ and using Theorem 16.3.2 (a), which says $\|\mathbf{A}-\mathbf{B}\|_2$ is the largest such ratio, gives $\|\mathbf{A}-\mathbf{B}\|_2 \geq \sigma_{k+1}(\mathbf{A})$.

Equality at the truncation. If $k \geq p$, then $\mathbf{A}_k = \mathbf{A}$ by Definition 13.11.1 and $\sigma_{k+1}(\mathbf{A}) = 0$ by our convention, so equality holds trivially; assume $k < p$. By Definition 13.11.1, $\mathbf{A}_k = \sum_{i \leq k} \sigma_i \mathbf{u}_i \mathbf{v}_i^*$ has rank $\min(k, \operatorname{rank}\mathbf{A}) \leq k$, and

$$\mathbf{A} - \mathbf{A}_k = \sum_{i>k} \sigma_i \mathbf{u}_i \mathbf{v}_i^* = \mathbf{U}\boldsymbol{\Sigma}^{(k)}\mathbf{V}^*,$$

where $\boldsymbol{\Sigma}^{(k)}$ is $\boldsymbol{\Sigma}$ with its first k diagonal entries replaced by 0. Since $\mathbf{U}$ is unitary, $(\mathbf{A}-\mathbf{A}_k)^*(\mathbf{A}-\mathbf{A}_k) = \mathbf{V}\,(\boldsymbol{\Sigma}^{(k)})^*\boldsymbol{\Sigma}^{(k)}\,\mathbf{V}^*$, which is similar to the diagonal matrix $(\boldsymbol{\Sigma}^{(k)})^*\boldsymbol{\Sigma}^{(k)}$; its eigenvalues are therefore $\sigma_{k+1}^2, \ldots, \sigma_p^2$ together with zeros. By Definition 13.9.1 the singular values of $\mathbf{A}-\mathbf{A}_k$ are $\sigma_{k+1} \geq \sigma_{k+2} \geq \ldots$ followed by zeros, the largest being σ_{k+1}. Hence $\|\mathbf{A}-\mathbf{A}_k\|_2 = \sigma_{k+1}(\mathbf{A})$ by Theorem 16.3.3 (c). This proves the theorem.

■

So the *minimizer* is the same matrix in both norms, and only the *minimum value* changes: σ_{k+1} for the spectral norm, $\left(\sum_{i>k}\sigma_i^2\right)^{1/2}$ for the Frobenius norm. That the same $\mathbf{A}_k$ works for both is not a coincidence — it is optimal in every unitarily invariant norm at once, which is Chapter 21's theorem and needs machinery neither this section nor Chapter 13 has.

Example. One matrix, two norms

Let

$$\mathbf{A} = \begin{pmatrix} 3 & 1 & 0 \\ 1 & 3 & 0 \\ 0 & 0 & 1 \end{pmatrix} \in M_3(\mathbb{R}).$$

Find the best approximations of rank 1 and rank 2, and compute the error in both the spectral and the Frobenius norm.

Solution. $\mathbf{A}$ is symmetric and block diagonal. The 2×2 block $\left(\begin{smallmatrix} 3 & 1 \\ 1 & 3 \end{smallmatrix}\right)$ has trace 6 and determinant 8, hence eigenvalues 4 and 2, with orthonormal eigenvectors $\frac{1}{\sqrt{2}}(1,1)$ and $\frac{1}{\sqrt{2}}(1,-1)$; the third eigenvalue is 1, with eigenvector $\mathbf{e}_3$. All three eigenvalues are positive, so $\mathbf{A} \succ \mathbf{O}$ and its singular values are those eigenvalues (Theorem 13.1.5; $\mathbf{A}^*\mathbf{A} = \mathbf{A}^2$ has eigenvalues their squares):

$$\sigma_1 = 4, \quad \sigma_2 = 2, \quad \sigma_3 = 1,$$

with $\mathbf{u}_i = \mathbf{v}_i$ the unit eigenvectors $\frac{1}{\sqrt{2}}(1,1,0)$, $\frac{1}{\sqrt{2}}(1,-1,0)$, $\mathbf{e}_3$.

Truncating,

$$\mathbf{A}_1 = 4\mathbf{v}_1\mathbf{v}_1^{\top} = \begin{pmatrix} 2 & 2 & 0 \\ 2 & 2 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \qquad \mathbf{A}_2 = \begin{pmatrix} 3 & 1 & 0 \\ 1 & 3 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

the second because $4\mathbf{v}_1\mathbf{v}_1^\top + 2\mathbf{v}_2\mathbf{v}_2^\top$ restores the 2×2 block exactly. The error matrices are

$$\mathbf{A}-\mathbf{A}_1 = \begin{pmatrix} 1 & -1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \qquad \mathbf{A}-\mathbf{A}_2 = \operatorname{diag}(0,0,1).$$

For $k=1$: $\|\mathbf{A}-\mathbf{A}_1\|_2 = \sigma_2 = 2$ and $\|\mathbf{A}-\mathbf{A}_1\|_F = \sqrt{\sigma_2^2+\sigma_3^2} = \sqrt{5} \approx 2.236$. Check the first directly: $\mathbf{A}-\mathbf{A}_1$ is symmetric with the 2×2 block $\left(\begin{smallmatrix} 1 & -1 \\ -1 & 1 \end{smallmatrix}\right)$, of eigenvalues 2 and 0, and the entry 1, so its largest eigenvalue in modulus is 2. Check the second directly: the sum of the squares of the five nonzero entries is $4\cdot 1 + 1 = 5$.

For $k=2$: both norms of $\operatorname{diag}(0,0,1)$ equal $1=\sigma_3$, and the two formulas agree because only one singular value is discarded.

Quick check. A matrix has singular values $6,3,2,1$. What is the smallest spectral-norm error of a rank-2 approximation, and how does it compare with the Frobenius-norm error found in Exercise (B1: Errors from the singular values alone)?

Solution. By Theorem 17.9.6 with $k=2$, the minimum is $\sigma_3 = 2$. By Theorem 13.11.3 the Frobenius minimum is $\sqrt{\sigma_3^2+\sigma_4^2} = \sqrt{5} \approx 2.236$. The Frobenius error is the larger of the two, as it must be: it counts every discarded singular value, while the spectral error counts only the largest one.

17.9.5 How far an invertible matrix is from being singular

Take $k = n-1$ in Theorem 17.9.6, with $\mathbf{A}$ square. A matrix of rank at most $n-1$ is precisely a singular matrix, so the theorem computes the distance from $\mathbf{A}$ to the set of singular matrices. That is the geometric content of Chapter 16's condition number, and it is the best payoff in this chapter.

Corollary 17.9.7. The Distance to Singularity

Let $\mathbf{A} \in M_n(F)$ be invertible. Then

$$\min\{\|\mathbf{A}-\mathbf{B}\|_2 : \mathbf{B} \in M_n(F) \text{ singular}\} = \sigma_n(\mathbf{A}) = \frac{1}{\|\mathbf{A}^{-1}\|_2},$$

the minimum being attained at $\mathbf{B} = \mathbf{A}_{n-1}$. Consequently the **relative** distance from $\mathbf{A}$ to the nearest singular matrix is

$$\frac{\min\{\|\mathbf{A}-\mathbf{B}\|_2 : \mathbf{B} \text{ singular}\}}{\|\mathbf{A}\|_2} = \frac{\sigma_n(\mathbf{A})}{\sigma_1(\mathbf{A})} = \frac{1}{\kappa_2(\mathbf{A})},$$

the reciprocal of the spectral condition number of Definition 16.8.1.

Idea. A singular matrix has rank at most $n-1$, so this is the spectral Eckart–Young theorem with $k = n-1$, and the answer is the next singular value, σ_n. The formula $1/\|\mathbf{A}^{-1}\|_2$ is the same number read through the inverse, whose largest singular value is $1/\sigma_n$.

Proof. A matrix $\mathbf{B} \in M_n(F)$ is singular if and only if $\operatorname{rank}\mathbf{B} \leq n-1$ (Theorem 7.5.3 (d)). So the set over which the minimum is taken is exactly the set appearing in Theorem 17.9.6 with $k = n-1$, and that theorem gives the value $\sigma_{(n-1)+1}(\mathbf{A}) = \sigma_n(\mathbf{A})$, attained at $\mathbf{A}_{n-1}$. The matrix $\mathbf{A}_{n-1}$ is singular because its rank is $\min(n-1, \operatorname{rank}\mathbf{A}) = n-1 < n$.

For the second expression: since $\operatorname{rank}\mathbf{A} = n$, exactly n singular values of $\mathbf{A}$ are nonzero (Theorem 13.9.1), so $\sigma_n(\mathbf{A}) > 0$. By Definition 16.8.1 and Proposition 16.8.2 (d), $\|\mathbf{A}\|_2\|\mathbf{A}^{-1}\|_2 = \kappa_2(\mathbf{A}) = \sigma_1(\mathbf{A})/\sigma_n(\mathbf{A})$, and $\|\mathbf{A}\|_2 = \sigma_1(\mathbf{A})$ by Theorem 16.3.3 (c); dividing, $\|\mathbf{A}^{-1}\|_2 = 1/\sigma_n(\mathbf{A})$. Dividing the minimum $\sigma_n(\mathbf{A})$ by $\|\mathbf{A}\|_2 = \sigma_1(\mathbf{A})$ gives $\sigma_n/\sigma_1 = 1/\kappa_2(\mathbf{A})$, the last display. This proves the corollary.

■

What this settles. Chapter 16 defined $\kappa(\mathbf{A}) = \|\mathbf{A}\|\|\mathbf{A}^{-1}\|$ as the worst factor by which $\mathbf{A}$ magnifies a relative error, proved that the factor is attained, and described it geometrically through Lemma 16.8.1 as the ratio of the largest to the smallest stretch $\|\mathbf{Ax}\|$ over unit vectors $\mathbf{x}$. That is a statement about what $\mathbf{A}$ *does*. Corollary 17.9.7 is a statement about where $\mathbf{A}$ *is*: it says that $1/\kappa_2(\mathbf{A})$ measures the distance from $\mathbf{A}$ to the boundary of the invertible matrices, relative to the size of $\mathbf{A}$. A matrix with $\kappa_2 = 10^6$ is not merely a matrix that magnifies errors by a million; it is a matrix that a relative perturbation of one part in a million can make singular. "Ill-conditioned" and "nearly singular" are the same property, and this is the theorem that identifies them.

Two earlier exercises reached for this. Exercise (C1: How far is a matrix from being singular?) found the same distance in the Frobenius norm — also σ_n, since only one singular value is discarded when $k = n-1$. Exercise (C2: The condition number is the reciprocal distance to singularity) obtained the spectral statement by hand, with a Neumann-series lower bound and the truncation as witness. Nothing above rests on either. What is new is that the distance to singularity is now the case $k = n-1$ of one theorem covering every rank, proved by a dimension count and stated as an attained minimum.

Example. The nearest singular matrix, exactly and approximately

(a) For the matrix $\mathbf{A}$ of Example (One matrix, two norms), find the nearest singular matrix and the relative distance to it.

(b) Do the same, to three significant figures, for $\mathbf{C} = \left(\begin{smallmatrix} 1 & 1 \\ 1 & 1.01 \end{smallmatrix}\right)$, the ill-conditioned matrix of Example (A system where the bound is attained).

Solution.

(a) The singular values are $4, 2, 1$, so by Corollary 17.9.7 the nearest singular matrix is at

distance $\sigma_3 = 1$, and it is

$$\mathbf{A}_2 = \begin{pmatrix} 3 & 1 & 0 \\ 1 & 3 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

computed in Example (One matrix, two norms); it is singular because its last row is zero. Here $\kappa_2(\mathbf{A}) = 4/1 = 4$, so the relative distance is $\frac{1}{4}$: a perturbation of a quarter of the size of $\mathbf{A}$ is needed, and $\mathbf{A}$ is very well conditioned.

(b) $\mathbf{C}$ is symmetric with $\operatorname{tr}\mathbf{C} = \frac{201}{100}$ and $\det\mathbf{C} = \frac{1}{100}$, so its eigenvalues are

$$\frac{201 \pm \sqrt{201^2 - 400}}{200} = \frac{201 \pm \sqrt{40001}}{200},$$

both positive, hence equal to the singular values: $\sigma_1 \approx 2.00501$ and $\sigma_2 \approx 0.0049875$. So $\kappa_2(\mathbf{C}) \approx 402.0$ and the nearest singular matrix is at distance $\sigma_2 \approx 0.00499$, a **relative** distance of $1/\kappa_2(\mathbf{C}) \approx 0.00249$, about a quarter of one percent. The nearest singular matrix itself is $\mathbf{C}_1 = \sigma_1 \mathbf{v}_1 \mathbf{v}_1^{\top}$, which to five decimals is

$$\mathbf{C}_1 \approx \begin{pmatrix} 0.99749 & 1.00249 \\ 1.00249 & 1.00752 \end{pmatrix};$$

it is singular, and each entry differs from the corresponding entry of $\mathbf{C}$ by about 0.0025. Changing $\mathbf{C}$ in its third decimal place destroys its invertibility. That is what $\kappa_2 \approx 402$ means.

Warning. **The distance to singularity is not measured by the determinant.** Chapter 16 §08 warned that no threshold on $|\det\mathbf{A}|$ detects ill-conditioning, and Corollary 17.9.7 says exactly why: the quantity that decides nearness to singularity is σ_n, and $|\det\mathbf{A}| = \sigma_1\sigma_2\cdots\sigma_n$ is a product in which a tiny σ_n can be hidden by large partners, or a healthy σ_n buried by many small ones. For $\mathbf{A} = 10^{-6}\mathbf{I}_{10}$ all singular values are 10^{-6}, so $|\det\mathbf{A}| = 10^{-60}$ while the relative distance to singularity is $1/\kappa_2 = 1$, the largest it can ever be. For $\mathbf{A} = \operatorname{diag}(10^6, 10^{-6})$, $|\det\mathbf{A}| = 1$ while the relative distance is 10^{-12}. Only the *smallest* singular value knows.

17.9.6 Deleting a row

The last transfer is Cauchy interlacing. For a Hermitian matrix the operation was deleting a row **and** the matching column; for a general matrix there is no matching, and deleting a single row is enough.

Theorem 17.9.8. Interlacing Under Deletion of a Row or a Column

Let $\mathbf{A} \in M_{m\times n}(F)$ with $m \geq 2$, and let $\mathbf{B} \in M_{(m-1)\times n}(F)$ be obtained from $\mathbf{A}$ by deleting one row. Then

$$\sigma_i(\mathbf{B}) \leq \sigma_i(\mathbf{A}) \quad (1 \leq i \leq n), \qquad \sigma_{i+1}(\mathbf{A}) \leq \sigma_i(\mathbf{B}) \quad (1 \leq i \leq n-1),$$

singular values beyond the p-th being read as 0. If instead $n \geq 2$ and $\mathbf{B} \in M_{m\times(n-1)}(F)$ is obtained from $\mathbf{A}$ by deleting one **column**, the same two chains hold, both over $1 \leq i \leq n-1$;

written out, $\sigma_1(\mathbf{A}) \geq \sigma_1(\mathbf{B}) \geq \sigma_2(\mathbf{A}) \geq \cdots \geq \sigma_{n-1}(\mathbf{B}) \geq \sigma_n(\mathbf{A})$.

Idea. Deleting a row removes one term from the sum $\|\mathbf{Ax}\|^2 = \sum_{\text{rows}} |\text{row} \cdot \mathbf{x}|^2$. So $\|\mathbf{Bx}\| \leq \|\mathbf{Ax}\|$ for every $\mathbf{x}$, which gives the first chain from the min–max formula at once, because both matrices are compared over the same subspaces of F^n. For the second chain, the two norms **agree** on the hyperplane where the deleted row vanishes; intersecting a good subspace for $\mathbf{B}$ with that hyperplane costs one dimension, and one dimension is exactly one index.

Proof. Say row r of $\mathbf{A}$ is deleted, and write $\mathbf{r}^* \in M_{1\times n}(F)$ for that row. Summing squares of entries coordinate by coordinate,

$$\|\mathbf{Ax}\|^2 = \|\mathbf{Bx}\|^2 + |\mathbf{r}^*\mathbf{x}|^2 \qquad (\mathbf{x} \in F^n), \tag{17.9.3}$$

since the entries of $\mathbf{Ax}$ are those of $\mathbf{Bx}$ together with $\mathbf{r}^*\mathbf{x}$.

First chain. By Eq. 17.9.3, $\|\mathbf{Bx}\|/\|\mathbf{x}\| \leq \|\mathbf{Ax}\|/\|\mathbf{x}\|$ for every $\mathbf{x} \neq \mathbf{0}$. Both $\mathbf{A}$ and $\mathbf{B}$ have n columns, so Theorem 17.9.2 describes $\sigma_i(\mathbf{A})$ and $\sigma_i(\mathbf{B})$ as a minimum over the *same* family of subspaces of F^n, of a maximum of these two objective functions. A pointwise smaller function has a smaller maximum on each subspace and hence a smaller minimum over the family, so $\sigma_i(\mathbf{B}) \leq \sigma_i(\mathbf{A})$ for every $i \leq n$.

Second chain. Fix $1 \leq i \leq n-1$. Let $\mathbf{w}_1, \ldots, \mathbf{w}_n$ be the right singular vectors of a singular value decomposition of $\mathbf{B}$ (Theorem 13.9.1), and put $T = \text{span}(\mathbf{w}_i, \ldots, \mathbf{w}_n)$, of dimension $n - i + 1$; by Lemma 17.9.1 (a) applied to $\mathbf{B}$,

$$\|\mathbf{Bx}\| \leq \sigma_i(\mathbf{B})\|\mathbf{x}\| \qquad (\mathbf{x} \in T).$$

Let $H = \{\mathbf{x} \in F^n : \mathbf{r}^*\mathbf{x} = 0\} = \text{nul}(\mathbf{r}^*)$. Since $\mathbf{r}^*$ has at most one nonzero row, $\operatorname{rank} \mathbf{r}^* \leq 1$, so $\dim H \geq n - 1$ by Rank–Nullity for matrices (Theorem 3.5.7). By the dimension formula (Theorem 1.7.3),

$$\dim(T \cap H) \;\geq\; (n - i + 1) + (n - 1) - n = n - i \;\geq\; 1.$$

Choose $W \subseteq T \cap H$ with $\dim W = n - i = n - (i+1) + 1$. For $\mathbf{0} \neq \mathbf{x} \in W$ we have $\mathbf{r}^*\mathbf{x} = 0$, so Eq. 17.9.3 gives $\|\mathbf{Ax}\| = \|\mathbf{Bx}\| \leq \sigma_i(\mathbf{B})\|\mathbf{x}\|$. Hence the inner maximum over W is at most $\sigma_i(\mathbf{B})$, and $\sigma_{i+1}(\mathbf{A})$, being the minimum of that inner maximum over all subspaces of dimension $n - (i+1) + 1$ (Theorem 17.9.2), is at most $\sigma_i(\mathbf{B})$.

Columns. Deleting column c of $\mathbf{A}$ is deleting row c of $\mathbf{A}^*$. Taking conjugate transposes in $\mathbf{A} = \mathbf{U\Sigma V}^*$ gives $\mathbf{A}^* = \mathbf{V\Sigma}^*\mathbf{U}^*$, in which $\mathbf{\Sigma}^*$ has the same diagonal $\sigma_1, \ldots, \sigma_p$; so by Theorem 13.9.2, $\sigma_i(\mathbf{A}^*) = \sigma_i(\mathbf{A})$ for every $i \leq p$, and the same holds for $\mathbf{B}$ and $\mathbf{B}^*$. Both $\mathbf{A}^*$ and $\mathbf{B}^*$ have m columns, so the row statement applied to them gives $\sigma_i(\mathbf{B}) \leq \sigma_i(\mathbf{A})$ for $i \leq m$ and $\sigma_{i+1}(\mathbf{A}) \leq \sigma_i(\mathbf{B})$ for $i \leq m - 1$. The remaining indices are free: if $i > m$ then i exceeds $\min(m, n-1)$, so $\sigma_i(\mathbf{B}) = 0$ and the first inequality is trivial; and if $i \geq m$ then $i + 1 > p$, so $\sigma_{i+1}(\mathbf{A}) = 0$ and the second is trivial. Hence both chains hold on the stated ranges. This proves the theorem.

■

Example. Interlacing on a 3 × 2 matrix

Let $\mathbf{A} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{pmatrix}$, the matrix of Example (The best rank-one approximation of a 3 × 2 matrix). Verify Theorem 17.9.8 for each of the three rows that could be deleted.

Solution. Example (The best rank-one approximation of a 3 × 2 matrix) computed $\mathbf{A}^\top\mathbf{A} = \left(\begin{smallmatrix} 2 & 1 \\ 1 & 2 \end{smallmatrix}\right)$, with eigenvalues 3 and 1, so $\sigma_1(\mathbf{A}) = \sqrt{3} \approx 1.732$ and $\sigma_2(\mathbf{A}) = 1$.

Deleting row 1 or row 2. The remaining matrix is $\left(\begin{smallmatrix} 0 & 1 \\ 1 & 1 \end{smallmatrix}\right)$ or $\left(\begin{smallmatrix} 1 & 0 \\ 1 & 1 \end{smallmatrix}\right)$. In both cases $\mathbf{B}^\top\mathbf{B} = \left(\begin{smallmatrix} 1 & 1 \\ 1 & 2 \end{smallmatrix}\right)$, with trace 3 and determinant 1, hence eigenvalues $(3 \pm \sqrt{5})/2$. Their square roots are $(1+\sqrt{5})/2 \approx 1.618$ and $(\sqrt{5}-1)/2 \approx 0.618$ — the two are reciprocal, since $(3+\sqrt{5})/2 \cdot (3-\sqrt{5})/2 = 1$. The interlacing reads

$$1.732 \;\geq\; 1.618 \;\geq\; 1 \;\geq\; 0.618,$$

that is $\sigma_1(\mathbf{A}) \geq \sigma_1(\mathbf{B}) \geq \sigma_2(\mathbf{A}) \geq \sigma_2(\mathbf{B})$, as the theorem requires.

Deleting row 3. The remaining matrix is $\mathbf{I}_2$, with $\sigma_1 = \sigma_2 = 1$, and the chain reads $1.732 \geq 1 \geq 1 \geq 1$. Here the second inequality of the first chain and the inequality of the second chain are both equalities: deleting the row that carried all the interaction between the two columns leaves a perfectly conditioned matrix.

Everything in this section came from one move: a singular value is the square root of an eigenvalue of a Hermitian matrix, so the min–max theory transfers wholesale. What the transfer bought is worth listing. The spectral half of the Eckart–Young theorem, which Chapter 13 could state only in the Frobenius norm, is now proved. The condition number of Chapter 16 has a geometric meaning: $1/\kappa_2(\mathbf{A})$ is how far, relatively, $\mathbf{A}$ sits from the nearest singular matrix. And singular values turned out to be perfectly conditioned with no hypothesis at all, which the Jordan block showed eigenvalues are not. The Hermitian dilation $\mathcal{H}(\mathbf{A})$ is the other bridge from singular values to Hermitian eigenvalues; Section 10's exercises use it, and Chapter 20 leans on it for perturbation bounds.

17.9.7 Exercises

17.9.7.1 A. Check your understanding

Exercise. A1

(a) State the min–max description of $\sigma_k(\mathbf{A})$, naming the dimension of the subspaces in each of the two forms.

(b) Explain in one sentence why the identity $\|\mathbf{A}\mathbf{x}\|^2/\|\mathbf{x}\|^2 = R_{\mathbf{A}^*\mathbf{A}}(\mathbf{x})$ is what makes the whole section possible.

(c) State the bound of Corollary 17.9.5 and say precisely which hypotheses it needs, comparing with Corollary 17.3.6.

(d) What is $\min\{\|\mathbf{A}-\mathbf{B}\|_2 : \operatorname{rank}\mathbf{B} \leq k\}$, and what is the corresponding Frobenius-norm minimum? Which of the two is larger, and why?

(e) Decide whether the statement is correct, and justify: "a square matrix with a very small determinant is close to a singular matrix".

Solution.

(a) See Theorem 17.9.2: $\sigma_k(\mathbf{A})$ is the minimum, over subspaces $W \leq F^n$ of dimension $n-k+1$, of $\max\{\|\mathbf{Ax}\|/\|\mathbf{x}\| : \mathbf{0} \neq \mathbf{x} \in W\}$; equivalently the maximum, over subspaces of dimension k, of the corresponding minimum.

(b) It converts a statement about the rectangular matrix $\mathbf{A}$, which has no eigenvalues, into a statement about the Hermitian matrix $\mathbf{A}^*\mathbf{A}$, which has the whole of Sections 1 to 8 available; and the conversion loses nothing, because the square root is increasing.

(c) $|\sigma_i(\mathbf{A}+\mathbf{E}) - \sigma_i(\mathbf{A})| \leq \|\mathbf{E}\|_2$, for **arbitrary** $\mathbf{A}, \mathbf{E} \in M_{m\times n}(F)$ — no symmetry, no smallness, no squareness. Corollary 17.3.6 is the same bound for eigenvalues, but it requires both $\mathbf{A}$ and $\mathbf{E}$ to be Hermitian, and Example (A Jordan block's eigenvalues move like $\varepsilon^{1/k}$) shows the bound is false without that.

(d) $\sigma_{k+1}(\mathbf{A})$ in the spectral norm (Theorem 17.9.6) and $\left(\sum_{i>k}\sigma_i^2\right)^{1/2}$ in the Frobenius norm (Theorem 13.11.3). The Frobenius value is the larger, since $\sum_{i>k}\sigma_i^2 \geq \sigma_{k+1}^2$; it charges for every discarded singular value, the spectral norm only for the biggest.

(e) Incorrect. Nearness to singularity is measured by σ_n, not by det (Corollary 17.9.7). The matrix $10^{-6}\mathbf{I}_{10}$ has determinant 10^{-60} and relative distance 1 to the singular matrices, the largest possible value. The determinant is the product of all the singular values and cannot report the smallest one.

17.9.7.2 B. Practice

Exercise. B1: Two norms, one list of singular values

A matrix $\mathbf{A} \in M_{5\times 4}(\mathbb{R})$ has singular values $8, 4, 4, 1$.

(a) Compute $\min\{\|\mathbf{A}-\mathbf{B}\|_2 : \operatorname{rank}\mathbf{B} \leq k\}$ for $k = 0, 1, 2, 3$.

(b) Compute the corresponding Frobenius minima and say for which k the two agree.

(c) Compute $\kappa_2(\mathbf{A})$ and the relative distance from $\mathbf{A}$ to the set of matrices of rank at most 3.

Solution.

(a) By Theorem 17.9.6 the minimum is σ_{k+1}, so the four values are $8, 4, 4, 1$ for $k = 0, 1, 2, 3$.

(b) By Theorem 13.11.3 the minimum is $\left(\sum_{i>k}\sigma_i^2\right)^{1/2}$:

$$\sqrt{64+16+16+1} = \sqrt{97}, \quad \sqrt{33}, \quad \sqrt{17}, \quad 1$$

for $k = 0, 1, 2, 3$, approximately 9.849, 5.745, 4.123, 1. They agree only at $k = 3$, where exactly one singular value is discarded.

(c) $\mathbf{A}$ has rank 4 and full column rank, so Definition 16.8.1 gives $\kappa_2(\mathbf{A}) = \sigma_1/\sigma_4 = 8$. The distance to the matrices of rank at most 3 is $\sigma_4 = 1$ and $\|\mathbf{A}\|_2 = 8$, so the relative distance is $\frac{1}{8} = 1/\kappa_2(\mathbf{A})$ — the rectangular analogue of Corollary 17.9.7, and for the same reason.

Exercise. B2: The nearest singular matrix

Let $\mathbf{A} = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$, whose condition number $\kappa_2(\mathbf{A}) = \frac{7+3\sqrt{5}}{2} \approx 6.854$ was computed in Example (Three condition numbers of one matrix).

(a) Find the singular values of $\mathbf{A}$ and the distance from $\mathbf{A}$ to the nearest singular matrix.

(b) Write down a singular matrix at exactly that distance, and check that it is singular.

(c) Compute the relative distance and compare it with $1/\kappa_2(\mathbf{A})$.

Solution.

(a) $\mathbf{A}$ is symmetric with $\operatorname{tr}\mathbf{A} = 3$ and $\det\mathbf{A} = 1$, so its eigenvalues are $(3 \pm \sqrt{5})/2$, both positive; hence these are also its singular values,

$$\sigma_1 = \tfrac{3+\sqrt{5}}{2} \approx 2.618, \qquad \sigma_2 = \tfrac{3-\sqrt{5}}{2} \approx 0.382.$$

By Corollary 17.9.7 the distance to the nearest singular matrix is $\sigma_2 = (3-\sqrt{5})/2$.

(b) The witness is $\mathbf{A}_1 = \sigma_1 \mathbf{v}_1 \mathbf{v}_1^{\top}$, where $\mathbf{v}_1$ is a unit eigenvector for σ_1. From $(2-\sigma_1)x_1 + x_2 = 0$ we may take $\mathbf{v}_1$ parallel to $(2, \sqrt{5}-1)$, whose squared length is $4 + (6 - 2\sqrt{5}) = 10 - 2\sqrt{5}$. Hence

$$\mathbf{A}_1 = \frac{3+\sqrt{5}}{2} \cdot \frac{1}{10-2\sqrt{5}} \begin{pmatrix} 4 & 2(\sqrt{5}-1) \\ 2(\sqrt{5}-1) & 6-2\sqrt{5} \end{pmatrix} = \begin{pmatrix} 1 + \frac{2\sqrt{5}}{5} & \frac{1}{2} + \frac{3\sqrt{5}}{10} \\ \frac{1}{2} + \frac{3\sqrt{5}}{10} & \frac{1}{2} + \frac{\sqrt{5}}{10} \end{pmatrix},$$

approximately $\left(\begin{smallmatrix} 1.894 & 1.171 \\ 1.171 & 0.724 \end{smallmatrix}\right)$. It is singular: it is a scalar multiple of $\mathbf{v}_1\mathbf{v}_1^{\top}$, which has rank 1, so its determinant is 0. (Directly: $1.894 \times 0.724 - 1.171^2 = 0$ to three decimals.)

(c) $\|\mathbf{A}\|_2 = \sigma_1$, so the relative distance is

$$\frac{\sigma_2}{\sigma_1} = \frac{3-\sqrt{5}}{3+\sqrt{5}} = \frac{(3-\sqrt{5})^2}{4} = \frac{7-3\sqrt{5}}{2} \approx 0.1459,$$

using $(3-\sqrt{5})(3+\sqrt{5}) = 4$. And $1/\kappa_2(\mathbf{A}) = 2/(7+3\sqrt{5}) = (7-3\sqrt{5})/2$, since $(7+3\sqrt{5})(7-3\sqrt{5}) = 49 - 45 = 4$. The two agree, as Corollary 17.9.7 says.

Exercise. B3: Interlacing by hand

Let $\mathbf{A} = \begin{pmatrix} 2 & 0 \\ 0 & 1 \\ 0 & 2 \end{pmatrix}$.

(a) Compute the singular values of $\mathbf{A}$.

(b) Compute the singular values of each of the three matrices obtained by deleting one row, and check Theorem 17.9.8 in each case.

Solution.

(a) $\mathbf{A}^{\top}\mathbf{A} = \left(\begin{smallmatrix} 4 & 0 \\ 0 & 5 \end{smallmatrix}\right)$, so the eigenvalues are 5 and 4 and $\sigma_1(\mathbf{A}) = \sqrt{5} \approx 2.236$, $\sigma_2(\mathbf{A}) = 2$.

(b) Deleting row 1 leaves $\left(\begin{smallmatrix}0&1\\0&2\end{smallmatrix}\right)$, with $\mathbf{B}^{\top}\mathbf{B} = \operatorname{diag}(0,5)$, so $\sigma_1 = \sqrt{5}$, $\sigma_2 = 0$; the chain reads $\sqrt{5} \geq \sqrt{5} \geq 2 \geq 0$.

Deleting row 2 leaves $\left(\begin{smallmatrix}2&0\\0&2\end{smallmatrix}\right)$, with singular values $2, 2$; the chain reads $\sqrt{5} \geq 2 \geq 2 \geq 2$.

Deleting row 3 leaves $\left(\begin{smallmatrix}2&0\\0&1\end{smallmatrix}\right)$, with singular values $2, 1$; the chain reads $\sqrt{5} \geq 2 \geq 2 \geq 1$.

All three satisfy $\sigma_1(\mathbf{A}) \geq \sigma_1(\mathbf{B}) \geq \sigma_2(\mathbf{A}) \geq \sigma_2(\mathbf{B})$, as Theorem 17.9.8 requires.

17.9.7.3 C. Going deeper

Exercise. C1: Rank cannot drop suddenly

Let $\mathbf{A} \in M_{m\times n}(F)$ have rank $r \geq 1$.

(a) Prove that every $\mathbf{B} \in M_{m\times n}(F)$ with $\|\mathbf{A} - \mathbf{B}\|_2 < \sigma_r(\mathbf{A})$ has $\operatorname{rank}\mathbf{B} \geq r$.

(b) Deduce that for every $\mathbf{A}$ there is an $\varepsilon > 0$ such that every $\mathbf{B}$ within ε of $\mathbf{A}$ in the spectral norm has $\operatorname{rank}\mathbf{B} \geq \operatorname{rank}\mathbf{A}$.

(c) Show that the reverse inequality can fail however small ε is: give matrices $\mathbf{A}$ and $\mathbf{B}$ with $\|\mathbf{A} - \mathbf{B}\|_2$ as small as we like and $\operatorname{rank}\mathbf{B} > \operatorname{rank}\mathbf{A}$.

Hint: for (a), argue by contradiction with Theorem 17.9.6.

Solution.

(a) Suppose $\operatorname{rank}\mathbf{B} \leq r - 1$. By Theorem 17.9.6 with $k = r - 1$,

$$\|\mathbf{A} - \mathbf{B}\|_2 \;\geq\; \sigma_r(\mathbf{A}),$$

contradicting the hypothesis. Hence $\operatorname{rank}\mathbf{B} \geq r$.

(b) If $\mathbf{A} = \mathbf{O}$ any $\varepsilon > 0$ works, since then $r = 0$ and every $\mathbf{B}$ has rank at least 0. Otherwise take $\varepsilon = \sigma_r(\mathbf{A})$, which is > 0 because $\mathbf{A}$ has exactly $r = \operatorname{rank}\mathbf{A}$ nonzero singular values (Theorem 13.9.3). Then (a) applies.

(c) Take $\mathbf{A} = \operatorname{diag}(1,0)$, of rank 1, and $\mathbf{B} = \operatorname{diag}(1,\delta)$ with $\delta > 0$. Then $\mathbf{A} - \mathbf{B} = \operatorname{diag}(0,-\delta)$, so $\|\mathbf{A} - \mathbf{B}\|_2 = \delta$, as small as we like, while $\operatorname{rank}\mathbf{B} = 2 > 1 = \operatorname{rank}\mathbf{A}$. Rank can jump **up** under an arbitrarily small perturbation, and never jumps down under a small one: that asymmetry is exactly (a), and it is why numerical rank is defined by a threshold on the singular values rather than by exact vanishing.

Exercise. C2: The dilation, and a second proof of the perturbation bound

Let $\mathbf{A}, \mathbf{E} \in M_{m\times n}(F)$ and let $\mathcal{H}(\cdot)$ be the Hermitian dilation of Proposition 17.9.3.

(a) Prove that $\mathcal{H}(\mathbf{A} + \mathbf{E}) = \mathcal{H}(\mathbf{A}) + \mathcal{H}(\mathbf{E})$ and that $\mathcal{H}(\mathbf{E})$ is Hermitian.

(b) Prove that $\|\mathcal{H}(\mathbf{E})\|_2 = \|\mathbf{E}\|_2$.

(c) Hence give a second proof of Corollary 17.9.5 for $1 \leq i \leq p$, using Corollary 17.3.6, and say in one sentence why this route explains the absence of hypotheses on $\mathbf{E}$.

Hint: for (b), use Proposition 17.9.3 and Theorem 16.3.3 (c).

Solution.

(a) Both claims are immediate from the block form: $\mathcal{H}$ is built entrywise from $\mathbf{A}$ and $\mathbf{A}^*$, and both $\mathbf{A} \mapsto \mathbf{A}$ and $\mathbf{A} \mapsto \mathbf{A}^*$ are additive, so

$$\mathcal{H}(\mathbf{A}+\mathbf{E}) = \begin{pmatrix} \mathbf{O} & \mathbf{A}+\mathbf{E} \\ \mathbf{A}^*+\mathbf{E}^* & \mathbf{O} \end{pmatrix} = \mathcal{H}(\mathbf{A}) + \mathcal{H}(\mathbf{E}).$$

That $\mathcal{H}(\mathbf{E})$ is Hermitian was proved in Proposition 17.9.3, which places no hypothesis on its argument.

(b) Write $\mathbf{M} = \mathcal{H}(\mathbf{E})$, which is Hermitian by (a). Then $\mathbf{M}^*\mathbf{M} = \mathbf{M}^2$, whose eigenvalues are the squares of those of $\mathbf{M}$ (apply Corollary 12.4.3 to $\mathbf{M}$ and square the diagonal). By Proposition 17.9.3 the eigenvalues of $\mathbf{M}$ are $\pm\sigma_i(\mathbf{E})$ for $i \le p$ together with zeros, so the eigenvalues of $\mathbf{M}^*\mathbf{M}$ are the numbers $\sigma_i(\mathbf{E})^2$, each twice, together with zeros. The largest is $\sigma_1(\mathbf{E})^2$, so $\sigma_1(\mathbf{M}) = \sigma_1(\mathbf{E})$ by Definition 13.9.1, and Theorem 16.3.3 (c) turns this into $\|\mathcal{H}(\mathbf{E})\|_2 = \|\mathbf{E}\|_2$.

(c) Let $1 \le i \le p$. By Proposition 17.9.3 the largest p eigenvalues of $\mathcal{H}(\mathbf{A})$ are $\sigma_1(\mathbf{A}) \ge \cdots \ge \sigma_p(\mathbf{A})$, all of the remaining ones being ≤ 0; so $\lambda_i(\mathcal{H}(\mathbf{A})) = \sigma_i(\mathbf{A})$, and likewise $\lambda_i(\mathcal{H}(\mathbf{A}+\mathbf{E})) = \sigma_i(\mathbf{A}+\mathbf{E})$. By (a), $\mathcal{H}(\mathbf{A}+\mathbf{E}) = \mathcal{H}(\mathbf{A}) + \mathcal{H}(\mathbf{E})$ with $\mathcal{H}(\mathbf{A})$ and $\mathcal{H}(\mathbf{E})$ both Hermitian, so Corollary 17.3.6 applies and gives

$$|\sigma_i(\mathbf{A}+\mathbf{E}) - \sigma_i(\mathbf{A})| = |\lambda_i(\mathcal{H}(\mathbf{A}+\mathbf{E})) - \lambda_i(\mathcal{H}(\mathbf{A}))| \le \|\mathcal{H}(\mathbf{E})\|_2 = \|\mathbf{E}\|_2.$$

The route explains the missing hypotheses: Corollary 17.3.6 demands that the perturbation be Hermitian, and $\mathcal{H}(\mathbf{E})$ **is** Hermitian for every $\mathbf{E}$ whatsoever. Passing to the dilation manufactures the hypothesis instead of assuming it.

17.10 Inertia, Again and Additively

Chapter 14 attached three numbers to a real symmetric matrix — how many positive, how many negative, how many zero entries appear on the diagonal after a congruence — and proved that the three do not depend on which congruence was used. The proof there was elementary and self-contained: an orthogonal basis for the form, two subspaces, a dimension count. This section proves the same theorem a second time, from the min–max description of the eigenvalues, and then shows what the second proof buys that the first could not: the three numbers become **additive across a Schur complement**.

Throughout, $F = \mathbb{R}$ or $F = \mathbb{C}$, and every matrix whose eigenvalues or inertia are named is Hermitian — real symmetric when $F = \mathbb{R}$. Its eigenvalues are real and are indexed decreasingly, $\lambda_1(\mathbf{A}) \ge \cdots \ge \lambda_n(\mathbf{A})$, as everywhere in this chapter. **Congruence** means the operation $\mathbf{A} \mapsto \mathbf{P}^*\mathbf{A}\mathbf{P}$ with $\mathbf{P}$ invertible, which sends Hermitian matrices to Hermitian matrices; over $\mathbb{R}$ the star is the plain transpose and this is exactly the relation $\mathbf{A} \simeq \mathbf{B}$ of Definition 14.2.1, and Chapter 13 already used the complex version in Proposition 13.1.2.

17.10.1 A symbol for the triple

Chapter 14 §05 named the three counts and proved them well defined, and Definition 14.5.1 records the words: the triple (n_+, n_-, n_0) is the **inertia** and the pair (n_+, n_-) the **signature**. What it did not do is give the triple a symbol, because it never needed to write an equation

between two triples. This section writes several, so the symbol is worth the line it costs.

Definition 17.10.1. The Inertia Triple

Let $\mathbf{A} \in M_n(F)$ be Hermitian, with eigenvalues $\lambda_1(\mathbf{A}) \geq \cdots \geq \lambda_n(\mathbf{A})$ listed **with multiplicity**. Write $n_+(\mathbf{A})$, $n_-(\mathbf{A})$ and $n_0(\mathbf{A})$ for the numbers of indices i with $\lambda_i(\mathbf{A}) > 0$, with $\lambda_i(\mathbf{A}) < 0$ and with $\lambda_i(\mathbf{A}) = 0$, and set

$$\operatorname{In}(\mathbf{A}) := \big(n_+(\mathbf{A}),\ n_-(\mathbf{A}),\ n_0(\mathbf{A})\big) \in \mathbb{N}^3.$$

Triples are added **entrywise**.

Clause by clause. Every eigenvalue is counted, so $n_+ + n_- + n_0 = n$ always. The counts are *with multiplicity*: $\operatorname{In}(\mathbf{I}_n) = (n, 0, 0)$, not $(1, 0, 0)$. And $n_0(\mathbf{A}) = \dim \operatorname{nul}(\mathbf{A})$, because a Hermitian matrix is unitarily diagonalizable (Corollary 12.4.3, or Corollary 12.5.2 when $F = \mathbb{R}$), so the eigenspace for 0 has dimension equal to the multiplicity of 0 in the list; equivalently $n_+ + n_- = \operatorname{rank} \mathbf{A}$.

For $F = \mathbb{R}$ this is not a new object. Theorem 14.5.3 says precisely that the numbers of positive, negative and zero *eigenvalues* of a real symmetric $\mathbf{A}$ are the numbers of positive, negative and zero *diagonal entries* after any congruence to diagonal form — that is, the inertia of Definition 14.5.1. So $\operatorname{In}(\mathbf{A})$ is the old triple, freshly abbreviated.

A few values, and the dictionary with definiteness. $\operatorname{In}(\mathbf{O}_n) = (0, 0, n)$, and $\operatorname{In}(\operatorname{diag}(1, -1)) = (1, 1, 0)$. By Theorem 13.1.4, $\mathbf{A} \succeq \mathbf{O}$ exactly when every eigenvalue is ≥ 0, that is when $n_-(\mathbf{A}) = 0$; by Theorem 13.1.5, $\mathbf{A} \succ \mathbf{O}$ exactly when $\operatorname{In}(\mathbf{A}) = (n, 0, 0)$. Definiteness is the extreme case of inertia, and everything below degenerates to a known statement about definiteness when the triple is extreme.

17.10.2 Sylvester's law, proved again

The description of n_+ that this chapter can supply is not a count of diagonal entries at all. It is a maximum over subspaces, and it mentions no basis, no diagonalization, and no matrix.

Theorem 17.10.1. Sylvester's Law of Inertia, Second Proof

Let $\mathbf{A} \in M_n(F)$ be Hermitian. Call a subspace $W \leq F^n$ **$\mathbf{A}$-positive** if $\langle \mathbf{Ax}, \mathbf{x} \rangle > 0$ for every $\mathbf{x} \in W$ with $\mathbf{x} \neq \mathbf{0}$. Then:

(a) $n_+(\mathbf{A}) = \max\{\dim W : W \text{ is } \mathbf{A}\text{-positive}\}$, and the maximum is attained at the span of an orthonormal family of eigenvectors for the positive eigenvalues of $\mathbf{A}$;

(b) for every invertible $\mathbf{P} \in M_n(F)$, the matrix $\mathbf{P}^*\mathbf{AP}$ is Hermitian and

$$\operatorname{In}(\mathbf{P}^*\mathbf{AP}) = \operatorname{In}(\mathbf{A}).$$

Over $\mathbb{R}$, part (b) is Sylvester's Law of Inertia, Theorem 14.5.1.

Idea. For (a), read Theorem 17.2.3 through a sign. Its max–min form says $\lambda_k(\mathbf{A})$ is the largest value that $R_{\mathbf{A}}$ can be *guaranteed* to exceed on a k-dimensional subspace; and "W is $\mathbf{A}$-positive" says exactly that $R_{\mathbf{A}}$ is guaranteed to be positive on W. So $\mathbf{A}$-positive subspaces of dimension k exist precisely when $\lambda_k(\mathbf{A}) > 0$, which is precisely when $k \leq n_+$. For (b), the

right-hand side of (a) is defined from the quadratic form alone, and a congruence is a change of variables inside that form: $\mathbf{x} \mapsto \mathbf{P}\mathbf{x}$ matches up the two families of positive subspaces, dimension for dimension.

Proof. Throughout, note that $\langle \mathbf{A}\mathbf{x}, \mathbf{x} \rangle = \mathbf{x}^*\mathbf{A}\mathbf{x}$ is real for Hermitian $\mathbf{A}$ (Proposition 11.6.7 (a), or directly: $\overline{\mathbf{x}^*\mathbf{A}\mathbf{x}} = \mathbf{x}^*\mathbf{A}^*\mathbf{x} = \mathbf{x}^*\mathbf{A}\mathbf{x}$), and that for $\mathbf{x} \neq \mathbf{0}$ the sign of $\langle \mathbf{A}\mathbf{x}, \mathbf{x} \rangle$ is the sign of the Rayleigh quotient $R_{\mathbf{A}}(\mathbf{x}) = \langle \mathbf{A}\mathbf{x}, \mathbf{x} \rangle / \langle \mathbf{x}, \mathbf{x} \rangle$ of Definition 17.1.1, since $\langle \mathbf{x}, \mathbf{x} \rangle = \|\mathbf{x}\|^2 > 0$. So W is $\mathbf{A}$-positive if and only if $R_{\mathbf{A}}(\mathbf{x}) > 0$ for every $\mathbf{x} \in W \setminus \{\mathbf{0}\}$.

(a), the lower bound. Put $k = n_+(\mathbf{A})$. If $k = 0$ the zero subspace is $\mathbf{A}$-positive and there is nothing to prove, so assume $k \geq 1$. By the spectral theorem (Corollary 12.4.3, or Corollary 12.5.2 when $F = \mathbb{R}$) there is an orthonormal basis $\mathbf{q}_1, \dots, \mathbf{q}_n$ of F^n with $\mathbf{A}\mathbf{q}_i = \lambda_i(\mathbf{A})\mathbf{q}_i$, the eigenvalues in decreasing order. Put $W_0 = \operatorname{span}(\mathbf{q}_1, \dots, \mathbf{q}_k)$, of dimension k. For $\mathbf{0} \neq \mathbf{x} = \sum_{i \leq k} c_i \mathbf{q}_i$, expanding by orthonormality,

$$\langle \mathbf{A}\mathbf{x}, \mathbf{x} \rangle = \sum_{i=1}^{k} \lambda_i(\mathbf{A}) |c_i|^2 \;\geq\; \lambda_k(\mathbf{A}) \sum_{i=1}^{k} |c_i|^2 = \lambda_k(\mathbf{A}) \|\mathbf{x}\|^2 > 0,$$

because $\lambda_i(\mathbf{A}) \geq \lambda_k(\mathbf{A})$ for $i \leq k$ and $\lambda_k(\mathbf{A}) > 0$, the latter since $k = n_+(\mathbf{A})$ counts the positive eigenvalues of a decreasing list. So W_0 is $\mathbf{A}$-positive of dimension $n_+(\mathbf{A})$.

(a), the upper bound. Let W be $\mathbf{A}$-positive with $d = \dim W \geq 1$. The unit sphere $S_W = \{\mathbf{x} \in W : \|\mathbf{x}\| = 1\}$ is a closed bounded subset of F^n, hence compact by fact (A3) of Chapter 16's introduction, and it is nonempty because $d \geq 1$. The function $\mathbf{x} \mapsto R_{\mathbf{A}}(\mathbf{x})$ is continuous on it, so by the extreme value theorem — fact (A4) of Chapter 16's introduction — it attains a minimum $m = R_{\mathbf{A}}(\mathbf{x}_0)$ there, and $m > 0$ because $\mathbf{x}_0 \in W$ is nonzero. By homogeneity of $R_{\mathbf{A}}$, the same m is the minimum over all nonzero $\mathbf{x} \in W$. Now W is one competitor in the max–min form of Theorem 17.2.3, so

$$\lambda_d(\mathbf{A}) = \max_{\dim W' = d} \; \min_{\mathbf{0} \neq \mathbf{x} \in W'} R_{\mathbf{A}}(\mathbf{x}) \;\geq\; m \;>\; 0.$$

A decreasing list with $\lambda_d > 0$ has at least d positive entries, so $d \leq n_+(\mathbf{A})$. Together with the lower bound this proves (a).

(b). Put $\mathbf{B} = \mathbf{P}^*\mathbf{A}\mathbf{P}$. Then $\mathbf{B}^* = \mathbf{P}^*\mathbf{A}^*\mathbf{P} = \mathbf{B}$, so $\mathbf{B}$ is Hermitian. For every $\mathbf{x} \in F^n$,

$$\langle \mathbf{B}\mathbf{x}, \mathbf{x} \rangle = \mathbf{x}^*\mathbf{P}^*\mathbf{A}\mathbf{P}\mathbf{x} = (\mathbf{P}\mathbf{x})^*\mathbf{A}(\mathbf{P}\mathbf{x}) = \langle \mathbf{A}(\mathbf{P}\mathbf{x}), \mathbf{P}\mathbf{x} \rangle.$$

Since $\mathbf{P}$ is invertible, $\mathbf{x} \mapsto \mathbf{P}\mathbf{x}$ is an isomorphism of F^n; its restriction to a subspace W is an isomorphism $W \to \mathbf{P}W$, which carries a basis to a basis (Theorem 2.5.2), so $\dim \mathbf{P}W = \dim W$. And $W \mapsto \mathbf{P}W$ is a bijection on subspaces, with inverse $W' \mapsto \mathbf{P}^{-1}W'$. Moreover W is $\mathbf{B}$-positive if and only if $\mathbf{P}W$ is $\mathbf{A}$-positive: the nonzero vectors of $\mathbf{P}W$ are exactly the vectors $\mathbf{P}\mathbf{x}$ with $\mathbf{0} \neq \mathbf{x} \in W$, and the display says the two quadratic forms agree at matched vectors. Hence the two sets of dimensions appearing in (a) coincide, and (a) gives $n_+(\mathbf{B}) = n_+(\mathbf{A})$.

Applying this to $-\mathbf{A}$, whose eigenvalues are the negatives of those of $\mathbf{A}$ and whose congruence by $\mathbf{P}$ is $-\mathbf{B}$, gives $n_-(\mathbf{B}) = n_+(-\mathbf{B}) = n_+(-\mathbf{A}) = n_-(\mathbf{A})$. Finally $n_0 = n - n_+ - n_-$ for both. This proves the theorem.

■

What each proof gives. Over $\mathbb{R}$ the two theorems say the same thing, since a congruence is exactly a change of basis for the underlying form (Theorem 14.2.1). They are not interchangeable.

Chapter 14's proof uses only the existence of an orthogonal basis for the form, the invariance of rank under congruence, and the dimension formula. It mentions no inner product, no eigenvalue, and no limit — in Chapter 14 the eigenvalues enter only afterwards, in Theorem 14.5.3 — and it is the proof to give for a symmetric bilinear form on an abstract real vector space, where no inner product has been chosen and the words "Hermitian matrix" have no meaning until one is.

The proof above costs the spectral theorem and one appeal to compactness. In exchange it gives three things. First, a **description of n_+ with no basis in it at all** — a maximum over subspaces — together with an explicit optimal subspace, which is what Exercise (C1: An intrinsic description of n_+) asked the reader to establish by hand from an orthogonal basis; here it falls out of the min–max, with the eigenvalues attached. Second, it works **over** $\mathbb{C}$, for Hermitian matrices and the congruence $\mathbf{A} \mapsto \mathbf{P}^*\mathbf{A}\mathbf{P}$, which Theorem 14.5.1 does not cover. Third, it keeps the eigenvalues in view, so the perturbation theory of Section 3 applies to the inertia directly; that is the next theorem. Haynsworth's theorem at the end of the section leans on the second gain: for complex Hermitian matrices, only this proof supplies the law of inertia it needs.

The companions of part (a) are worth stating, because one of them is a trap.

Corollary 17.10.2. What Each Count Is a Maximum Of

Let $\mathbf{A} \in M_n(F)$ be Hermitian.

(a) $n_-(\mathbf{A}) = \max\{\dim W : \langle \mathbf{A}\mathbf{x}, \mathbf{x}\rangle < 0 \text{ for every } \mathbf{0} \neq \mathbf{x} \in W\}$;

(b) $n_+(\mathbf{A}) + n_0(\mathbf{A}) = \max\{\dim W : \langle \mathbf{A}\mathbf{x}, \mathbf{x}\rangle \geq 0 \text{ for every } \mathbf{x} \in W\}$, and symmetrically $n_-(\mathbf{A}) + n_0(\mathbf{A})$ is the largest dimension of a subspace on which $\langle \mathbf{A}\mathbf{x}, \mathbf{x}\rangle \leq 0$;

(c) $n_0(\mathbf{A}) = \dim \operatorname{nul}(\mathbf{A}) = n - n_+(\mathbf{A}) - n_-(\mathbf{A})$.

Idea. Each claim has a witness and a competitor, as in Courant–Fischer. The witness is the span of the eigenvectors with the right sign; any competing subspace that is too large must meet the span of the eigenvectors of the wrong sign, and on that intersection the form has the wrong sign. For (c), no maximum is involved at all — the kernel is read off the eigenvalues.

Proof.

(a) Apply Theorem 17.10.1 (a) to $-\mathbf{A}$: its eigenvalues are the negatives of those of $\mathbf{A}$, so $n_+(-\mathbf{A}) = n_-(\mathbf{A})$, and a subspace is $(-\mathbf{A})$-positive exactly when $\langle \mathbf{A}\mathbf{x}, \mathbf{x}\rangle < 0$ on its nonzero vectors.

(b) Fix, as in the previous proof, an orthonormal eigenbasis $\mathbf{q}_1, \dots, \mathbf{q}_n$ of $\mathbf{A}$ with $\mathbf{A}\mathbf{q}_i = \lambda_i(\mathbf{A})\mathbf{q}_i$ (Corollary 12.4.3). Let $t = n_+(\mathbf{A}) + n_0(\mathbf{A})$, so that $\lambda_i(\mathbf{A}) \geq 0$ exactly for $i \leq t$. On $W_1 = \operatorname{span}(\mathbf{q}_1, \dots, \mathbf{q}_t)$ the same expansion gives $\langle \mathbf{A}\mathbf{x}, \mathbf{x}\rangle = \sum_{i \leq t} \lambda_i(\mathbf{A})|c_i|^2 \geq 0$, so the maximum is at least t. Conversely let W satisfy $\langle \mathbf{A}\mathbf{x}, \mathbf{x}\rangle \geq 0$ throughout and suppose $\dim W > t$. Put $W_2 = \operatorname{span}(\mathbf{q}_{t+1}, \dots, \mathbf{q}_n)$, of dimension $n - t = n_-(\mathbf{A})$, on whose nonzero vectors $\langle \mathbf{A}\mathbf{x}, \mathbf{x}\rangle = \sum_{i > t} \lambda_i(\mathbf{A})|c_i|^2 < 0$, every $\lambda_i(\mathbf{A})$ with $i > t$ being negative. Since $\dim W + \dim W_2 > t + (n - t) = n$, Lemma 17.2.1 produces

$\mathbf{0} \neq \mathbf{x} \in W \cap W_2$, for which $\langle \mathbf{A}\mathbf{x}, \mathbf{x} \rangle$ is both ≥ 0 and < 0. This is impossible, so $\dim W \leq t$. The statement with $\leq$ follows by applying this one to $-\mathbf{A}$.

(c) The first equality was recorded after Definition 17.10.1, and the second is $n_+ + n_- + n_0 = n$.

■

Warning. n_0 **is not the largest dimension of a subspace on which the form vanishes.** Parts (a) and (b) describe n_-, n_+, and their sums with n_0 as maxima over subspaces; n_0 itself has no such description, and guessing one is the standard error here. Take $\mathbf{A} = \operatorname{diag}(1, -1)$, so $\operatorname{In}(\mathbf{A}) = (1, 1, 0)$ and $n_0(\mathbf{A}) = 0$. On the line $W = \operatorname{span}((1, 1))$,

$$\langle \mathbf{A}\mathbf{x}, \mathbf{x} \rangle = x_1^2 - x_2^2 = 0 \qquad \text{for every } \mathbf{x} \in W,$$

so a subspace of dimension 1 carries the identically zero form while $n_0(\mathbf{A}) = 0$. What n_0 *is* is $\dim \operatorname{nul}(\mathbf{A})$, by (c) — a kernel, not a maximum. The largest dimension of a subspace on which the form vanishes identically is a different invariant. For a real symmetric **invertible** $\mathbf{A}$ it is the Witt index of Chapter 14 §07, which that section defines only for nondegenerate forms and bounds by $n/2$. For a singular $\mathbf{A}$ the bound fails — for $\mathbf{A} = \mathbf{O}$ the whole space qualifies — and in general the answer is $n_0 + \min(n_+, n_-)$, which Exercise (C1: How large can a subspace be on which the form vanishes?) computes.

Quick check. A Hermitian $\mathbf{A} \in M_5(F)$ has $\operatorname{In}(\mathbf{A}) = (2, 2, 1)$. What is the largest dimension of a subspace on which $\langle \mathbf{A}\mathbf{x}, \mathbf{x} \rangle > 0$ for all nonzero $\mathbf{x}$, and the largest on which $\langle \mathbf{A}\mathbf{x}, \mathbf{x} \rangle \geq 0$ for all $\mathbf{x}$?

Solution. 2 and 3. The first is $n_+(\mathbf{A}) = 2$ by Theorem 17.10.1 (a); the second is $n_+(\mathbf{A}) + n_0(\mathbf{A}) = 2 + 1 = 3$ by Corollary 17.10.2 (b). The kernel direction may be added to a positive subspace without breaking ≥ 0, but it does break > 0 — which is the whole difference between the two numbers.

17.10.3 Inertia under a perturbation of small rank

A congruence leaves the inertia alone and does nothing else: it can move the eigenvalues anywhere compatible with their signs. Chapter 14 §05's warning displayed the pair

$$\mathbf{A} = \operatorname{diag}(1, -1), \qquad \mathbf{P}^{\top}\mathbf{A}\mathbf{P} = \operatorname{diag}(4, -9) \quad (\mathbf{P} = \operatorname{diag}(2, 3)),$$

where λ_1 rose from 1 to 4 and λ_2 fell from -1 to -9. There is no interlacing here: $\lambda_1(\mathbf{A}) \geq \lambda_1(\mathbf{P}^{\top}\mathbf{A}\mathbf{P})$ already fails. A congruence *is* a perturbation, $\mathbf{P}^{*}\mathbf{A}\mathbf{P} = \mathbf{A} + \mathbf{E}$ with $\mathbf{E} = \mathbf{P}^{*}\mathbf{A}\mathbf{P} - \mathbf{A}$, but $\mathbf{E}$ can have full rank, and then the rank bound of Section 3 says nothing. When the perturbation has small rank, it does.

Theorem 17.10.3. Inertia Moves by at Most the Rank

Let $\mathbf{A}, \mathbf{E} \in M_n(F)$ be Hermitian with $\operatorname{rank} \mathbf{E} \le r$. Then

$$|n_+(\mathbf{A}+\mathbf{E}) - n_+(\mathbf{A})| \le r, \qquad |n_-(\mathbf{A}+\mathbf{E}) - n_-(\mathbf{A})| \le r,$$

and $|n_0(\mathbf{A}+\mathbf{E}) - n_0(\mathbf{A})| \le r$ as well.

Idea. Corollary 17.3.8 says that adding a matrix of rank r shifts each eigenvalue by at most r places in the ordered list. A count of positive eigenvalues is a count of places, so it too can move by at most r.

Proof. Put $s = n_+(\mathbf{A}+\mathbf{E})$. If $s \le r$ then certainly $n_+(\mathbf{A}) \ge 0 \ge s - r$. Otherwise $i \coloneqq s - r$ satisfies $1 \le i$ and $i + r = s \le n$, so Corollary 17.3.8 applies with this i and gives

$$\lambda_s(\mathbf{A}+\mathbf{E}) = \lambda_{i+r}(\mathbf{A}+\mathbf{E}) \;\le\; \lambda_i(\mathbf{A}) = \lambda_{s-r}(\mathbf{A}).$$

Since $s = n_+(\mathbf{A}+\mathbf{E})$ we have $\lambda_s(\mathbf{A}+\mathbf{E}) > 0$, hence $\lambda_{s-r}(\mathbf{A}) > 0$, hence $n_+(\mathbf{A}) \ge s - r$. In either case

$$n_+(\mathbf{A}+\mathbf{E}) - n_+(\mathbf{A}) \le r.$$

For the opposite inequality, write $\mathbf{A} = (\mathbf{A}+\mathbf{E}) + (-\mathbf{E})$, where $-\mathbf{E}$ is Hermitian with $\operatorname{rank}(-\mathbf{E}) = \operatorname{rank}\mathbf{E} \le r$, and apply what was just proved with $\mathbf{A}+\mathbf{E}$ in the role of $\mathbf{A}$: $n_+(\mathbf{A}) - n_+(\mathbf{A}+\mathbf{E}) \le r$. The two together give the first claim.

The second claim is the first applied to the Hermitian pair $-\mathbf{A}, -\mathbf{E}$, using $n_-(\mathbf{M}) = n_+(-\mathbf{M})$. For the third, adding the first two would only give $2r$; the sharp bound comes from kernels instead. By Corollary 17.10.2 (c), $n_0(\mathbf{M}) = \dim\operatorname{nul}(\mathbf{M})$ for every Hermitian $\mathbf{M}$. A vector killed by both $\mathbf{A}$ and $\mathbf{E}$ is killed by $\mathbf{A}+\mathbf{E}$, so $\operatorname{nul}(\mathbf{A}) \cap \operatorname{nul}(\mathbf{E}) \subseteq \operatorname{nul}(\mathbf{A}+\mathbf{E})$; and $\dim\operatorname{nul}(\mathbf{E}) = n - \operatorname{rank}\mathbf{E} \ge n - r$ by Theorem 2.3.1. The sharper form of Lemma 17.2.1 then gives

$$n_0(\mathbf{A}+\mathbf{E}) \;\ge\; \dim\bigl(\operatorname{nul}(\mathbf{A}) \cap \operatorname{nul}(\mathbf{E})\bigr) \;\ge\; n_0(\mathbf{A}) + (n-r) - n \;=\; n_0(\mathbf{A}) - r.$$

Applied to $\mathbf{A} = (\mathbf{A}+\mathbf{E}) + (-\mathbf{E})$ the same argument gives $n_0(\mathbf{A}) \ge n_0(\mathbf{A}+\mathbf{E}) - r$, and the two together are the claim. The bound is attained: $\mathbf{A} = \mathbf{I}_r \oplus \mathbf{O}$ and $\mathbf{E} = -\mathbf{I}_r \oplus \mathbf{O}$ give $n_0(\mathbf{A}) = n - r$ and $n_0(\mathbf{A}+\mathbf{E}) = n$. This proves the theorem. ∎

Example. A rank-one perturbation, three ways

Let $\mathbf{A} = \operatorname{diag}(3, -1, -2)$ and $\mathbf{v} = (1, 1, 0)$, and let $\mathbf{E}_t = t\,\mathbf{v}\mathbf{v}^{\top}$, which has rank 1 for $t \neq 0$. Find $\operatorname{In}(\mathbf{A}+\mathbf{E}_t)$ for $t = 2$, $t = \tfrac{3}{2}$, and $t = -5$, and check Theorem 17.10.3.

Solution. $\operatorname{In}(\mathbf{A}) = (1, 2, 0)$, read off the diagonal. Since $\mathbf{v}$ has third entry 0,

$$\mathbf{A} + \mathbf{E}_t = \begin{pmatrix} 3+t & t & 0 \\ t & t-1 & 0 \\ 0 & 0 & -2 \end{pmatrix} = \begin{pmatrix} 3+t & t \\ t & t-1 \end{pmatrix} \oplus (-2),$$

and the eigenvalue list of a block diagonal matrix is the two lists together (its characteristic

polynomial is the product of the two, by Theorem 7.8.4). The entry -2 contributes one negative eigenvalue for every t. The 2×2 block has

$$\det = (3+t)(t-1) - t^2 = 2t - 3, \qquad \operatorname{tr} = 2t + 2,$$

and since the determinant of a real symmetric 2×2 matrix is the product of its two real eigenvalues and the trace is their sum, the eigenvalues both have the sign of the trace when the determinant is positive, have opposite signs when it is negative, and are 0 and tr when it is zero.

- $t = 2$: $\det = 1 > 0$, $\operatorname{tr} = 6 > 0$, so the block contributes $(2,0,0)$ and $\operatorname{In}(\mathbf{A} + \mathbf{E}_2) = (2,1,0)$.
- $t = \frac{3}{2}$: $\det = 0$, $\operatorname{tr} = 5$, so the block contributes $(1,0,1)$ and $\operatorname{In}(\mathbf{A} + \mathbf{E}_{3/2}) = (1,1,1)$.
- $t = -5$: $\det = -13 < 0$, so the block contributes $(1,1,0)$ and $\operatorname{In}(\mathbf{A} + \mathbf{E}_{-5}) = (1,2,0)$, unchanged.

Against $\operatorname{In}(\mathbf{A}) = (1,2,0)$: at $t = 2$ one negative direction has become positive, so n_+ and n_- each move by $1 = r$ and the bound of Theorem 17.10.3 is attained; at $t = \frac{3}{2}$, exactly on the way between, the direction passes through 0, so n_- and n_0 each move by 1. No rank-one perturbation of this $\mathbf{A}$ can reach $(3,0,0)$, which would need n_+ to move by 2.

17.10.4 Additivity across a Schur complement

Chapter 8 built the Schur complement out of block elimination and proved two things about it: $\det \mathbf{M} = \det \mathbf{A} \cdot \det(\mathbf{M}/\mathbf{A})$ (Theorem 8.3.2) and $\operatorname{rank} \mathbf{M} = k + \operatorname{rank}(\mathbf{M}/\mathbf{A})$ (Theorem 8.3.3). Chapter 13 then used the same elimination on Hermitian matrices — in §01, observing that it is the block ŁDU factorization written as a congruence, and in §06, to prove Theorem 13.6.1 — and each time read off a statement about positive definiteness only. Haynsworth's theorem is what that congruence is worth in general: the rank does not merely add, the *signs* add.

Theorem 17.10.4. Haynsworth's Inertia Additivity Formula

Let $1 \le k \le n-1$ and let

$$\mathbf{M} = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^* & \mathbf{D} \end{pmatrix} \in M_n(F)$$

be Hermitian, with $\mathbf{A} \in M_k(F)$ **invertible**, $\mathbf{B} \in M_{k\times(n-k)}(F)$ and $\mathbf{D} \in M_{n-k}(F)$. Then $\mathbf{A}$ and the Schur complement $\mathbf{M}/\mathbf{A} = \mathbf{D} - \mathbf{B}^*\mathbf{A}^{-1}\mathbf{B}$ (Definition 8.3.1) are Hermitian, and

$$\operatorname{In}(\mathbf{M}) = \operatorname{In}(\mathbf{A}) + \operatorname{In}(\mathbf{M}/\mathbf{A}).$$

Idea. Block elimination factors $\mathbf{M}$ as $Ł\,(\mathbf{A} \oplus \mathbf{M}/\mathbf{A})\,\mathbf{U}$ with Ł lower and $\mathbf{U}$ upper unit block triangular (Theorem 8.3.1). For a **Hermitian** $\mathbf{M}$ with $\mathbf{A}$ Hermitian, the two outer factors are conjugate transposes of each other, so the factorization is a congruence and not merely a factorization. Theorem 17.10.1 (b) then says the inertia of $\mathbf{M}$ is the inertia of $\mathbf{A} \oplus \mathbf{M}/\mathbf{A}$, and the inertia of a block diagonal matrix is the sum of the inertias of its blocks, because its characteristic polynomial is the product.

Proof. Since $\mathbf{M}^* = \mathbf{M}$, comparing blocks gives $\mathbf{A}^* = \mathbf{A}$ and $\mathbf{D}^* = \mathbf{D}$. As $\mathbf{A}$ is Hermitian and invertible, $(\mathbf{A}^{-1})^* = (\mathbf{A}^*)^{-1} = \mathbf{A}^{-1}$, so $\mathbf{A}^{-1}$ is Hermitian and

$$(\mathbf{M}/\mathbf{A})^* = \mathbf{D}^* - \mathbf{B}^*(\mathbf{A}^{-1})^*\mathbf{B} = \mathbf{D} - \mathbf{B}^*\mathbf{A}^{-1}\mathbf{B} = \mathbf{M}/\mathbf{A}.$$

Step 1: the elimination is a congruence. By Theorem 8.3.1 (a), applied with $\mathbf{C} = \mathbf{B}^*$,

$$\mathbf{M} = Ł\,(\mathbf{A} \oplus (\mathbf{M}/\mathbf{A}))\,\mathbf{U}, \qquad Ł = \begin{pmatrix} \mathbf{I}_k & \mathbf{O} \\ \mathbf{B}^*\mathbf{A}^{-1} & \mathbf{I}_{n-k} \end{pmatrix}, \quad \mathbf{U} = \begin{pmatrix} \mathbf{I}_k & \mathbf{A}^{-1}\mathbf{B} \\ \mathbf{O} & \mathbf{I}_{n-k} \end{pmatrix}.$$

Now $\mathbf{U}^*$ has blocks $\mathbf{I}_k, \mathbf{O}$ on top and $(\mathbf{A}^{-1}\mathbf{B})^* = \mathbf{B}^*\mathbf{A}^{-1}$, $\mathbf{I}_{n-k}$ below, using that $\mathbf{A}^{-1}$ is Hermitian; that is, $\mathbf{U}^* = Ł$. So

$$\mathbf{M} = \mathbf{U}^*(\mathbf{A} \oplus (\mathbf{M}/\mathbf{A}))\mathbf{U}.$$

The matrix $\mathbf{U}$ is invertible, with inverse $\mathbf{P} := \left(\begin{smallmatrix} \mathbf{I}_k & -\mathbf{A}^{-1}\mathbf{B} \\ \mathbf{O} & \mathbf{I}_{n-k} \end{smallmatrix}\right)$, as multiplying out confirms. Multiplying the display by $\mathbf{P}^*$ on the left and $\mathbf{P}$ on the right, and using $\mathbf{P}^*\mathbf{U}^* = (\mathbf{U}\mathbf{P})^* = \mathbf{I}_n$,

$$\mathbf{P}^*\mathbf{M}\mathbf{P} = \mathbf{A} \oplus (\mathbf{M}/\mathbf{A}). \tag{17.10.1}$$

Step 2: apply the law of inertia. $\mathbf{P}$ is invertible, so Theorem 17.10.1 (b) gives

$$\operatorname{In}(\mathbf{M}) = \operatorname{In}(\mathbf{P}^*\mathbf{M}\mathbf{P}) = \operatorname{In}(\mathbf{A} \oplus (\mathbf{M}/\mathbf{A})).$$

Step 3: inertia is additive over a block diagonal sum. Let $\mathbf{X} \in M_k(F)$ and $\mathbf{Y} \in M_{n-k}(F)$ be Hermitian. The matrix $x\mathbf{I}_n - (\mathbf{X} \oplus \mathbf{Y})$ is block diagonal with blocks $x\mathbf{I}_k - \mathbf{X}$ and $x\mathbf{I}_{n-k} - \mathbf{Y}$, so by Theorem 7.8.4 its determinant is $\det(x\mathbf{I}_k - \mathbf{X})\det(x\mathbf{I}_{n-k} - \mathbf{Y})$; that is, $p_{\mathbf{X}\oplus\mathbf{Y}} = p_{\mathbf{X}}\,p_{\mathbf{Y}}$. The eigenvalue list of a Hermitian matrix with multiplicity is the list of roots of its characteristic polynomial with multiplicity, so the list for $\mathbf{X} \oplus \mathbf{Y}$ is the two lists concatenated. Counting positive, negative, and zero entries gives $\operatorname{In}(\mathbf{X}\oplus\mathbf{Y}) = \operatorname{In}(\mathbf{X}) + \operatorname{In}(\mathbf{Y})$.

Applying Step 3 with $\mathbf{X} = \mathbf{A}$ and $\mathbf{Y} = \mathbf{M}/\mathbf{A}$ to the conclusion of Step 2 proves the theorem.

■

Three remarks. First, Eq. 17.10.1 is the cleanest illustration in the book of what a Schur complement *is*: the block elimination of Chapter 8, performed with the same matrix on both sides, is a congruence, and $\mathbf{M}/\mathbf{A}$ is simply what the congruence leaves in the lower corner. For Hermitian $\mathbf{M}$, Chapter 8's determinant and rank formulas are both shadows of this one identity. Second, taking ranks in Eq. 17.10.1 recovers Theorem 8.3.3 for Hermitian $\mathbf{M}$, since $n_+ + n_-$ = rank; Haynsworth refines the rank formula by splitting each rank into its two signs. Third, the theorem needs $\mathbf{A}$ invertible and nothing else — no definiteness anywhere, which is what separates it from Chapter 13 §06's test.

Example. The Chapter 14 example, through a Schur complement

Find $\mathrm{In}(\mathbf{M})$ for

$$\mathbf{M} = \begin{pmatrix} 1 & 1 & 2 \\ 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix},$$

the matrix of Example (A signature by symmetric elimination), using Theorem 17.10.4 with the leading 2×2 block.

Solution. Take $\mathbf{A} = \left(\begin{smallmatrix} 1 & 1 \\ 1 & 2 \end{smallmatrix}\right)$, $\mathbf{B} = (2,3)^{\top}$, and $\mathbf{D} = (1)$. Then $\det \mathbf{A} = 1 \neq 0$, so $\mathbf{A}$ is invertible with $\mathbf{A}^{-1} = \left(\begin{smallmatrix} 2 & -1 \\ -1 & 1 \end{smallmatrix}\right)$. Its leading principal minors are 1 and 1, both positive, so $\mathbf{A} \succ \mathbf{O}$ by Theorem 13.1.5 and $\mathrm{In}(\mathbf{A}) = (2,0,0)$.

For the Schur complement, $\mathbf{B}^{\top}\mathbf{A}^{-1} = (2,3)\left(\begin{smallmatrix} 2 & -1 \\ -1 & 1 \end{smallmatrix}\right) = (1,1)$, so $\mathbf{B}^{\top}\mathbf{A}^{-1}\mathbf{B} = (1,1)\cdot(2,3) = 5$ and

$$\mathbf{M}/\mathbf{A} = 1 - 5 = -4, \qquad \mathrm{In}(\mathbf{M}/\mathbf{A}) = (0,1,0).$$

Hence $\mathrm{In}(\mathbf{M}) = (2,0,0) + (0,1,0) = (2,1,0)$.

Two checks. The congruence of Eq. 17.10.1 is given by $\mathbf{A}^{-1}\mathbf{B} = (1,1)^{\top}$, hence

$$\mathbf{P} = \begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \\ 0 & 0 & 1 \end{pmatrix}, \qquad \mathbf{P}^{\top}\mathbf{M}\mathbf{P} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 2 & 0 \\ 0 & 0 & -4 \end{pmatrix},$$

which is $\mathbf{A} \oplus (-4)$ as the theorem says; and Example (A signature by symmetric elimination) reached $\mathrm{diag}(1,1,-4)$ by paired row and column operations, finding the same -4. The Schur complement of the leading block *is* the last pivot of the symmetric elimination. The eigenvalues were never needed: they are the roots of $x^3 - 4x^2 - 9x + 4$, which Example (A signature by symmetric elimination) already noted are irrational.

Example. A 4×4 inertia in two lines

Find $\mathrm{In}(\mathbf{N})$ for

$$\mathbf{N} = \begin{pmatrix} 1 & 0 & 1 & 2 \\ 0 & 1 & 1 & -1 \\ 1 & 1 & 1 & 0 \\ 2 & -1 & 0 & 3 \end{pmatrix}.$$

Solution. The leading 2×2 block is $\mathbf{A} = \mathbf{I}_2$, which is invertible with $\mathrm{In}(\mathbf{A}) = (2,0,0)$ and $\mathbf{A}^{-1} = \mathbf{I}_2$. With

$$\mathbf{B} = \begin{pmatrix} 1 & 2 \\ 1 & -1 \end{pmatrix}, \qquad \mathbf{D} = \begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix},$$

the Schur complement needs no inversion at all:

$$\mathbf{N}/\mathbf{A} = \mathbf{D} - \mathbf{B}^{\top}\mathbf{B} = \begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix} - \begin{pmatrix} 2 & 1 \\ 1 & 5 \end{pmatrix} = \begin{pmatrix} -1 & -1 \\ -1 & -2 \end{pmatrix}.$$

This 2×2 matrix has determinant $2 - 1 = 1 > 0$ and trace $-3 < 0$, so both its eigenvalues

are negative and $\mathrm{In}(\mathbf{N}/\mathbf{A}) = (0,2,0)$. By Theorem 17.10.4,

$$\mathrm{In}(\mathbf{N}) = (2,0,0) + (0,2,0) = (2,2,0).$$

In particular $\mathbf{N}$ is invertible and indefinite. The characteristic polynomial of $\mathbf{N}$ is $x^4 - 6x^3 + 5x^2 + 8x + 1$, which has no rational root; the four eigenvalues were never computed, and the signature came from one 2×2 determinant and one trace.

Warning. **A must be invertible, and the off-diagonal block is not a spectator.** Take $\mathbf{M} = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$, with $k = 1$. Neither diagonal block is invertible, so no Schur complement exists for this partition and Theorem 17.10.4 does not apply — and its conclusion would be false if one tried to read the blocks directly: $\mathrm{In}(\mathbf{M}) = (1,1,0)$, since the eigenvalues are ± 1, while the two diagonal blocks contribute $(0,0,1) + (0,0,1) = (0,0,2)$. Even when $\mathbf{A}$ *is* invertible, the diagonal blocks alone decide nothing: for $\mathbf{M}_b = \left(\begin{smallmatrix} 1 & b \\ b & 1 \end{smallmatrix}\right)$ with b real, the Schur complement is $1 - b^2$, so $\mathrm{In}(\mathbf{M}_b)$ is $(2,0,0)$, $(1,0,1)$ or $(1,1,0)$ according as $|b| < 1$, $= 1$ or > 1. All the information is in the coupling.

The positive definite case of Haynsworth is a theorem Chapter 13 proved by hand, and it is now one line.

Corollary 17.10.5. Block Definiteness, a Second Proof

Let $1 \le k \le n-1$ and let $\mathbf{M} = \left(\begin{smallmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^* & \mathbf{D} \end{smallmatrix}\right) \in M_n(F)$ be Hermitian with $\mathbf{A} \in M_k(F)$ and $\mathbf{A} \succ \mathbf{O}$. Then

$$\mathbf{M} \succ \mathbf{O} \iff \mathbf{M}/\mathbf{A} \succ \mathbf{O}, \qquad \mathbf{M} \succeq \mathbf{O} \iff \mathbf{M}/\mathbf{A} \succeq \mathbf{O}.$$

Proof. $\mathbf{A} \succ \mathbf{O}$ is invertible (Theorem 13.1.5) with $\mathrm{In}(\mathbf{A}) = (k,0,0)$, so Theorem 17.10.4 gives

$$\mathrm{In}(\mathbf{M}) = (k,0,0) + \mathrm{In}(\mathbf{M}/\mathbf{A}).$$

Comparing entries, $n_+(\mathbf{M}) = k + n_+(\mathbf{M}/\mathbf{A})$ and $n_-(\mathbf{M}) = n_-(\mathbf{M}/\mathbf{A})$, while $\mathbf{M}/\mathbf{A} \in M_{n-k}(F)$.

Now $\mathbf{M} \succ \mathbf{O}$ means $n_+(\mathbf{M}) = n$ (Theorem 13.1.5), that is $n_+(\mathbf{M}/\mathbf{A}) = n - k$, that is $\mathbf{M}/\mathbf{A} \succ \mathbf{O}$. And $\mathbf{M} \succeq \mathbf{O}$ means $n_-(\mathbf{M}) = 0$ (Theorem 13.1.4), that is $n_-(\mathbf{M}/\mathbf{A}) = 0$, that is $\mathbf{M}/\mathbf{A} \succeq \mathbf{O}$. This proves the corollary. ∎

This is Theorem 13.6.1 (b). Chapter 13 §06 proved it with the very same congruence, and then read positivity directly off the split quadratic form $\mathbf{y}_1^*\mathbf{A}\mathbf{y}_1 + \mathbf{y}_2^*(\mathbf{M}/\mathbf{A})\mathbf{y}_2$, with no eigenvalue in sight. The two proofs share their first step and differ in the second: Chapter 13's is shorter and needs nothing beyond Proposition 13.1.2, while the one above reads the verdict off a count, and so shows the definite case as the extreme instance of an equality that holds for every Hermitian $\mathbf{M}$ with $\mathbf{A}$ invertible, definite or not.

17.10.5 Exercises

17.10.5.1 A. Check your understanding

Exercise. A1

(a) Define $\mathrm{In}(\mathbf{A})$ for a Hermitian $\mathbf{A}$, and say what $n_+ + n_-$ equals.

(b) State the subspace description of $n_+(\mathbf{A})$ and name the subspace that attains it.

(c) Determine whether the following is correct, and justify: "$n_0(\mathbf{A})$ is the largest dimension of a subspace on which $\langle \mathbf{Ax}, \mathbf{x}\rangle$ vanishes identically."

(d) State Theorem 17.10.4, with all hypotheses.

(e) A congruence preserves the inertia of a Hermitian matrix. Does it preserve the eigenvalues, the trace, or the rank? Justify each answer.

Solution.

(a) $\mathrm{In}(\mathbf{A}) = (n_+, n_-, n_0)$, the numbers of positive, negative, and zero eigenvalues of $\mathbf{A}$ counted with multiplicity (Definition 17.10.1). Their sum is n, and $n_+ + n_- = \operatorname{rank}\mathbf{A}$, since $n_0 = \dim\operatorname{nul}(\mathbf{A})$ by Corollary 17.10.2 (c).

(b) $n_+(\mathbf{A})$ is the largest dimension of a subspace W with $\langle \mathbf{Ax}, \mathbf{x}\rangle > 0$ for every nonzero $\mathbf{x} \in W$ (Theorem 17.10.1 (a)), attained at the span of orthonormal eigenvectors for the positive eigenvalues, that is at $\operatorname{span}(\mathbf{q}_1, \dots, \mathbf{q}_{n_+})$ in an eigenbasis ordered decreasingly.

(c) Incorrect. For $\mathbf{A} = \operatorname{diag}(1, -1)$ we have $n_0(\mathbf{A}) = 0$, yet $\langle \mathbf{Ax}, \mathbf{x}\rangle = x_1^2 - x_2^2$ vanishes identically on the line spanned by $(1, 1)$. The correct description is $n_0(\mathbf{A}) = \dim\operatorname{nul}(\mathbf{A})$.

(d) If $\mathbf{M} = \left(\begin{smallmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^* & \mathbf{D} \end{smallmatrix}\right) \in M_n(F)$ is Hermitian with $\mathbf{A} \in M_k(F)$ invertible and $1 \le k \le n-1$, then $\mathrm{In}(\mathbf{M}) = \mathrm{In}(\mathbf{A}) + \mathrm{In}(\mathbf{M}/\mathbf{A})$, where $\mathbf{M}/\mathbf{A} = \mathbf{D} - \mathbf{B}^*\mathbf{A}^{-1}\mathbf{B}$.

(e) Rank: yes, since rank $= n_+ + n_-$ is determined by the inertia (and independently, congruence by an invertible matrix preserves rank by Theorem 3.5.9). Eigenvalues: no — $\operatorname{diag}(1, -1)$ and $\operatorname{diag}(4, -9)$ are congruent by $\mathbf{P} = \operatorname{diag}(2, 3)$. Trace: no, by the same pair, whose traces are 0 and -5.

17.10.5.2 B. Practice

Exercise. B1: Two inertias by Schur complement

Find the inertia of each of the following, using Theorem 17.10.4 with the leading 2×2 block, and compute no eigenvalues.

(a) $\mathbf{M} = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & -1 \end{pmatrix}$.

(b) $\mathbf{N} = \begin{pmatrix} 1 & 0 & 1 & 1 \\ 0 & -1 & 1 & 0 \\ 1 & 1 & 2 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix}$.

Solution.

(a) $\mathbf{A} = \left(\begin{smallmatrix} 2 & 1 \\ 1 & 1 \end{smallmatrix}\right)$ has leading principal minors 2 and 1, both positive, so $\mathbf{A} \succ \mathbf{O}$ (Theorem 13.1.5) and $\operatorname{In}(\mathbf{A}) = (2, 0, 0)$; also $\mathbf{A}^{-1} = \left(\begin{smallmatrix} 1 & -1 \\ -1 & 2 \end{smallmatrix}\right)$. With $\mathbf{B} = (1, 0)^{\top}$ and $\mathbf{D} = (-1)$,

$$\mathbf{B}^{\top}\mathbf{A}^{-1}\mathbf{B} = (1, 0)\begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix}\begin{pmatrix} 1 \\ 0 \end{pmatrix} = 1,$$

so $\mathbf{M}/\mathbf{A} = -1 - 1 = -2$ and $\operatorname{In}(\mathbf{M}/\mathbf{A}) = (0, 1, 0)$. Hence $\operatorname{In}(\mathbf{M}) = (2, 1, 0)$. (Consistency check: $\det \mathbf{M} = \det \mathbf{A} \cdot (\mathbf{M}/\mathbf{A}) = 1 \cdot (-2) = -2$ by Theorem 8.3.2, and a 3×3 matrix with inertia $(2, 1, 0)$ has negative determinant.)

(b) $\mathbf{A} = \operatorname{diag}(1, -1)$ is invertible with $\mathbf{A}^{-1} = \mathbf{A}$ and $\operatorname{In}(\mathbf{A}) = (1, 1, 0)$. With

$$\mathbf{B} = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}, \qquad \mathbf{D} = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix},$$

we get $\mathbf{A}^{-1}\mathbf{B} = \left(\begin{smallmatrix} 1 & 1 \\ -1 & 0 \end{smallmatrix}\right)$ and

$$\mathbf{B}^{\top}\mathbf{A}^{-1}\mathbf{B} = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 1 & 1 \\ -1 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix},$$

so

$$\mathbf{N}/\mathbf{A} = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix} - \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 2 & -1 \\ -1 & 0 \end{pmatrix},$$

whose determinant is $-1 < 0$, so its two eigenvalues have opposite signs and $\operatorname{In}(\mathbf{N}/\mathbf{A}) = (1, 1, 0)$. Hence $\operatorname{In}(\mathbf{N}) = (1, 1, 0) + (1, 1, 0) = (2, 2, 0)$.

Exercise. B2: How much can a rank-two perturbation do?

Let $\mathbf{A} \in M_6(\mathbb{R})$ be symmetric with $\operatorname{In}(\mathbf{A}) = (4, 1, 1)$, and let $\mathbf{E}$ be symmetric with $\operatorname{rank} \mathbf{E} = 2$.

(a) Which values of $n_+(\mathbf{A} + \mathbf{E})$ does Theorem 17.10.3 allow?

(b) Show that the two extreme values are both achieved, for suitable $\mathbf{A}$ and $\mathbf{E}$.

(c) Can $\mathbf{A} + \mathbf{E}$ be negative definite? Justify.

Solution.

(a) By Theorem 17.10.3 with $r = 2$, $|n_+(\mathbf{A}+\mathbf{E}) - 4| \le 2$, and $n_+ \le 6$ always, so $n_+(\mathbf{A}+\mathbf{E}) \in \{2, 3, 4, 5, 6\}$.

(b) Take $\mathbf{A} = \operatorname{diag}(1, 1, 1, 1, -1, 0)$, which has inertia $(4, 1, 1)$. With $\mathbf{E} = \operatorname{diag}(0, 0, 0, 0, 2, 1)$, of rank 2, we get $\mathbf{A} + \mathbf{E} = \mathbf{I}_6$, so $n_+ = 6$. With $\mathbf{E} = \operatorname{diag}(-2, -2, 0, 0, 0, 0)$, of rank 2, we get $\mathbf{A} + \mathbf{E} = \operatorname{diag}(-1, -1, 1, 1, -1, 0)$, so $n_+ = 2$.

(c) No. Negative definiteness would mean $n_+(\mathbf{A} + \mathbf{E}) = 0$, and $|0 - 4| = 4 > 2$ contradicts Theorem 17.10.3.

Exercise. B3: A bordered matrix

Let $\mathbf{D} \in M_n(\mathbb{R})$ be symmetric and invertible, let $\mathbf{b} \in \mathbb{R}^n$ and $c \in \mathbb{R}$, and put $\mathbf{M} = \left(\begin{smallmatrix} \mathbf{D} & \mathbf{b} \\ \mathbf{b}^\top & c \end{smallmatrix}\right)$.

(a) Express $\mathrm{In}(\mathbf{M})$ in terms of $\mathrm{In}(\mathbf{D})$ and the scalar $s = c - \mathbf{b}^\top \mathbf{D}^{-1}\mathbf{b}$.

(b) Take $\mathbf{D} = \mathrm{diag}(1, -1)$ and $\mathbf{b} = (1, 1)$. Compute $\mathrm{In}(\mathbf{M})$ for $c = 0$, $c = 1$, and $c = -1$.

(c) For which c is $\mathbf{M}$ singular?

Solution.

(a) By Theorem 17.10.4 with $\mathbf{A} = \mathbf{D}$, and $\mathbf{M}/\mathbf{D} = s$ a 1×1 matrix,

$$\mathrm{In}(\mathbf{M}) = \mathrm{In}(\mathbf{D}) + \begin{cases} (1,0,0) & s > 0, \\ (0,1,0) & s < 0, \\ (0,0,1) & s = 0. \end{cases}$$

(b) Here $\mathbf{D}^{-1} = \mathrm{diag}(1, -1) = \mathbf{D}$, so $\mathbf{b}^\top \mathbf{D}^{-1}\mathbf{b} = 1 - 1 = 0$ and $s = c$. Also $\mathrm{In}(\mathbf{D}) = (1, 1, 0)$. Hence $\mathrm{In}(\mathbf{M})$ is $(1, 1, 1)$ for $c = 0$, $(2, 1, 0)$ for $c = 1$, and $(1, 2, 0)$ for $c = -1$. The case $c = 0$ recovers Exercise (C3: Inertia of a block matrix) (c), which found the same triple by exhibiting the congruence directly.

(c) $\mathbf{M}$ is singular exactly when $n_0(\mathbf{M}) \geq 1$, which by (a) happens exactly when $s = 0$, since $\mathbf{D}$ is invertible and so contributes $n_0(\mathbf{D}) = 0$. Here that is $c = 0$.

17.10.5.3 C. Going deeper

Exercise. C1: How large can a subspace be on which the form vanishes?

Let $\mathbf{A} \in M_n(F)$ be Hermitian with $\mathrm{In}(\mathbf{A}) = (n_+, n_-, n_0)$. Call a subspace $W \leq F^n$ **totally isotropic** for $\mathbf{A}$ if $\langle \mathbf{A}\mathbf{x}, \mathbf{x} \rangle = 0$ for every $\mathbf{x} \in W$; for real symmetric $\mathbf{A}$ this is the notion of Chapter 14 §07, in the form that section derives from polarization.

(a) Prove that every totally isotropic W satisfies $\dim W \leq n_0 + \min(n_+, n_-)$.

(b) Construct a totally isotropic subspace of dimension exactly $n_0 + \min(n_+, n_-)$.

(c) Deduce that the largest dimension of a totally isotropic subspace equals n_0 if and only if $\mathbf{A}$ is positive semidefinite or negative semidefinite.

Hint: for (a), a totally isotropic subspace satisfies both inequalities of Corollary 17.10.2 (b). For (b), pair up a positive eigenvector with a negative one.

Solution.

(a) A totally isotropic W satisfies $\langle \mathbf{A}\mathbf{x}, \mathbf{x} \rangle \geq 0$ throughout and also $\langle \mathbf{A}\mathbf{x}, \mathbf{x} \rangle \leq 0$ throughout. By Corollary 17.10.2 (b), the first gives $\dim W \leq n_+ + n_0$ and the second $\dim W \leq n_- + n_0$. Hence $\dim W \leq n_0 + \min(n_+, n_-)$.

(b) Fix an orthonormal eigenbasis $\mathbf{q}_1, \dots, \mathbf{q}_n$ with $\mathbf{A}\mathbf{q}_i = \lambda_i \mathbf{q}_i$ and $\lambda_1 \geq \cdots \geq \lambda_n$ (Corollary 12.4.3, or Corollary 12.5.2 when $F = \mathbb{R}$), so that $\lambda_i > 0$ for $i \leq n_+$, $\lambda_i = 0$ for $n_+ < i \leq n_+ + n_0$, and $\lambda_i < 0$ for $i > n_+ + n_0$. Put $t = \min(n_+, n_-)$ and, for $1 \leq j \leq t$, let $j' = n + 1 - j$. Then $j \leq n_+$ and $j' \geq n + 1 - n_- > n_+ + n_0$, so $\lambda_j > 0 > \lambda_{j'}$ and the

indices $1,\dots,t,\ 1',\dots,t'$ are $2t$ distinct indices, none of them in the zero block. Put

$$\mathbf{x}_j = \frac{\mathbf{q}_j}{\sqrt{\lambda_j}} + \frac{\mathbf{q}_{j'}}{\sqrt{-\lambda_{j'}}} \qquad (1 \le j \le t),$$

and let W be spanned by $\mathbf{x}_1,\dots,\mathbf{x}_t$ together with the n_0 vectors $\mathbf{q}_i$ with $\lambda_i = 0$. These $t+n_0$ vectors are independent, since they are nonzero combinations of pairwise disjoint sets of vectors from a basis.
For $\mathbf{x} \in W$ write $\mathbf{x} = \sum_j c_j \mathbf{x}_j + \mathbf{z}$ with $\mathbf{A}\mathbf{z} = \mathbf{0}$. Then $\mathbf{A}\mathbf{x} = \sum_j c_j \mathbf{A}\mathbf{x}_j$ with $\mathbf{A}\mathbf{x}_j = \sqrt{\lambda_j}\,\mathbf{q}_j - \sqrt{-\lambda_{j'}}\,\mathbf{q}_{j'}$, and

$$\langle \mathbf{A}\mathbf{x}, \mathbf{x}\rangle = \mathbf{x}^*\mathbf{A}\mathbf{x} = \sum_{j,k} \overline{c_k}\, c_j\, \mathbf{x}_k^*\mathbf{A}\mathbf{x}_j + \sum_j c_j\, \mathbf{z}^*\mathbf{A}\mathbf{x}_j.$$

Each $\mathbf{z}^*\mathbf{A}\mathbf{x}_j = 0$, because $\mathbf{z}$ lies in the span of eigenvectors orthogonal to $\mathbf{q}_j$ and $\mathbf{q}_{j'}$. For $k \ne j$, $\mathbf{x}_k^*\mathbf{A}\mathbf{x}_j = 0$, because $\mathbf{x}_k$ and $\mathbf{A}\mathbf{x}_j$ are combinations of disjoint sets of orthonormal vectors. And

$$\mathbf{x}_j^*\mathbf{A}\mathbf{x}_j = \frac{\sqrt{\lambda_j}}{\sqrt{\lambda_j}} - \frac{\sqrt{-\lambda_{j'}}}{\sqrt{-\lambda_{j'}}} = 1 - 1 = 0.$$

Hence $\langle \mathbf{A}\mathbf{x}, \mathbf{x}\rangle = 0$ for every $\mathbf{x} \in W$, and W is totally isotropic of dimension $n_0 + \min(n_+, n_-)$.

(c) By (a) and (b) the largest totally isotropic dimension is $n_0 + \min(n_+, n_-)$, which equals n_0 exactly when $\min(n_+, n_-) = 0$, that is when $n_- = 0$ or $n_+ = 0$, that is when $\mathbf{A} \succeq \mathbf{O}$ or $-\mathbf{A} \succeq \mathbf{O}$ (Theorem 13.1.4).

Exercise. C2: The inertia of a Hermitian dilation

Let $\mathbf{A} \in M_{m\times n}(F)$ have rank r, and let $\mathcal{H}(\mathbf{A}) = \left(\begin{smallmatrix} \mathbf{O} & \mathbf{A} \\ \mathbf{A}^* & \mathbf{O} \end{smallmatrix}\right) \in M_{m+n}(F)$ be the Hermitian dilation of Proposition 17.9.3.

(a) Prove that $\operatorname{In}(\mathcal{H}(\mathbf{A})) = (r,\ r,\ m+n-2r)$.

(b) Deduce a formula for $\operatorname{rank}\mathbf{A}$ in terms of the inertia of $\mathcal{H}(\mathbf{A})$, and check that it gives $\operatorname{rank}\mathbf{O} = 0$.

(c) Verify (a) for $\mathbf{A} = \left(\begin{smallmatrix} 1 & 0 & 2 \\ 0 & 1 & 1 \end{smallmatrix}\right)$.

Solution.

(a) By Proposition 17.9.3 the eigenvalue list of $\mathcal{H}(\mathbf{A})$ with multiplicity consists of $\pm\sigma_i(\mathbf{A})$ for $1 \le i \le p = \min(m,n)$, together with $m+n-2p$ zeros. Exactly r of the $\sigma_i(\mathbf{A})$ are nonzero (Theorem 13.9.3), so the positive entries are the r numbers $\sigma_1,\dots,\sigma_r$ and the negative entries are their negatives, giving $n_+ = n_- = r$. The rest are zero: $(m+n)-2r$ of them. Hence $\operatorname{In}(\mathcal{H}(\mathbf{A})) = (r, r, m+n-2r)$.

(b) $\operatorname{rank}\mathbf{A} = n_+(\mathcal{H}(\mathbf{A})) = \frac{1}{2}(\operatorname{rank}\mathcal{H}(\mathbf{A}))$, using $n_+ + n_- = \operatorname{rank}$. For $\mathbf{A} = \mathbf{O}$ the dilation is the zero matrix of size $m+n$, whose inertia is $(0, 0, m+n)$, so the formula returns 0.

(c) Here $m = 2$, $n = 3$, and $\operatorname{rank}\mathbf{A} = 2$, since the first two columns are $\mathbf{e}_1, \mathbf{e}_2$. Also

$$\mathbf{A}\mathbf{A}^{\top} = \begin{pmatrix} 5 & 2 \\ 2 & 2 \end{pmatrix},$$

of trace 7 and determinant 6, hence eigenvalues 6 and 1; so $\sigma_1(\mathbf{A}) = \sqrt{6}$ and $\sigma_2(\mathbf{A}) = 1$. By Proposition 17.9.3 the eigenvalues of the 5×5 matrix $\mathcal{H}(\mathbf{A})$ are $\pm\sqrt{6}, \pm 1$ and one 0, so $\mathrm{In}(\mathcal{H}(\mathbf{A})) = (2, 2, 1)$, which is $(r, r, m + n - 2r) = (2, 2, 5 - 4)$.

Exercise. C3: Inertia from the leading principal minors

Let $\mathbf{M} \in M_n(F)$ be Hermitian and suppose that every leading principal submatrix $\mathbf{M}_k \in M_k(F)$ (rows and columns $1, \dots, k$) is **invertible**, for $k = 1, \dots, n$. Write $d_k = \det \mathbf{M}_k$ and $d_0 = 1$.

(a) Prove that for $2 \le k \le n$ the Schur complement $\mathbf{M}_k/\mathbf{M}_{k-1}$ is the scalar d_k/d_{k-1}.

(b) Deduce that $\mathrm{In}(\mathbf{M}) = (n_+, n_-, 0)$, where n_- is the number of indices $k \in \{1, \dots, n\}$ with $d_k/d_{k-1} < 0$ — that is, the number of **sign changes** in the sequence $d_0, d_1, \dots, d_n$.

(c) Apply this to the matrix $\mathbf{M}$ of Example (The Chapter 14 example, through a Schur complement) and compare.

Hint: for (b), induct on k using Theorem 17.10.4.

Solution.

(a) Write $\mathbf{M}_k = \left(\begin{smallmatrix} \mathbf{M}_{k-1} & \mathbf{b} \\ \mathbf{b}^* & c \end{smallmatrix}\right)$. Since $\mathbf{M}_{k-1}$ is invertible, $\mathbf{M}_k/\mathbf{M}_{k-1}$ is defined and is a 1×1 matrix, and Theorem 8.3.2 (a) gives $d_k = \det \mathbf{M}_k = \det \mathbf{M}_{k-1} \cdot (\mathbf{M}_k/\mathbf{M}_{k-1}) = d_{k-1}(\mathbf{M}_k/\mathbf{M}_{k-1})$. As $d_{k-1} \neq 0$, dividing gives $\mathbf{M}_k/\mathbf{M}_{k-1} = d_k/d_{k-1}$.

(b) Induct on k. For $k = 1$, $\mathbf{M}_1 = (d_1)$ is a nonzero real scalar (real because $\mathbf{M}$ is Hermitian), and $\mathrm{In}(\mathbf{M}_1)$ is $(1, 0, 0)$ or $(0, 1, 0)$ according to the sign of $d_1/d_0 = d_1$. For the step, $\mathbf{M}_{k-1}$ is invertible, so Theorem 17.10.4 applies to $\mathbf{M}_k$ with $\mathbf{A} = \mathbf{M}_{k-1}$:

$$\mathrm{In}(\mathbf{M}_k) = \mathrm{In}(\mathbf{M}_{k-1}) + \mathrm{In}(d_k/d_{k-1}),$$

and the last term is $(1, 0, 0)$ or $(0, 1, 0)$ according to the sign of the ratio, which is nonzero. Summing from $k = 1$ to n,

$$\mathrm{In}(\mathbf{M}) = \mathrm{In}(\mathbf{M}_n) = (\#\{k : d_k/d_{k-1} > 0\},\ \#\{k : d_k/d_{k-1} < 0\},\ 0),$$

and $d_k/d_{k-1} < 0$ says exactly that d_{k-1} and d_k have opposite signs. (Taking $n_- = 0$ here recovers the leading-minor criterion of Theorem 13.1.5: all $d_k > 0$ if and only if $\mathbf{M} \succ \mathbf{O}$.)

(c) For $\mathbf{M} = \left(\begin{smallmatrix} 1 & 1 & 2 \\ 1 & 2 & 3 \\ 2 & 3 & 1 \end{smallmatrix}\right)$ the leading minors are $d_1 = 1$, $d_2 = 2 - 1 = 1$, and $d_3 = -4$, all nonzero, so (b) applies. The sequence $d_0, d_1, d_2, d_3 = 1, 1, 1, -4$ has exactly one sign change, at the last step. Hence $n_- = 1$, $n_+ = 2$, and $\mathrm{In}(\mathbf{M}) = (2, 1, 0)$, agreeing with Example (The Chapter 14 example, through a Schur complement) and with Example (A signature by symmetric elimination).

17.11 Which Functions Respect the Order

Chapter 13 §05 ended on two facts that sit awkwardly together. If $\mathbf{A} \succeq \mathbf{B} \succeq \mathbf{O}$, then $\mathbf{A}^{1/2} \succeq \mathbf{B}^{1/2}$ (Proposition 13.5.4); but $\mathbf{A}^2 \succeq \mathbf{B}^2$ can fail (Example (Squaring does not preserve the Loewner

order)). For numbers, both $t \mapsto t^{1/2}$ and $t \mapsto t^2$ are increasing on $[0, \infty)$, and both preserve inequalities. This section names the property that separates them and proves it for every function the book can reach. It then compares that property with what §02's min–max theorem already gives for *every* increasing function, and states without proof the classification that settles the question. Since this is the last section of the chapter, it ends by summing up the chapter's moves and listing the promises the chapter has paid.

Throughout, $F = \mathbb{R}$ or $F = \mathbb{C}$, and every matrix whose eigenvalues are indexed is Hermitian, with eigenvalues $\lambda_1(\mathbf{A}) \geq \cdots \geq \lambda_n(\mathbf{A})$ indexed decreasingly and repeated according to multiplicity. The letter I denotes an **interval** of real numbers (open, closed, or half-open, bounded or not). It is plain italic, and it is not the bold identity matrix $\mathbf{I}$. A function $f\colon I \to \mathbb{R}$ is **increasing** when $s \leq t$ implies $f(s) \leq f(t)$; strictness is not required. Every result below holds for either choice of F, with the same proof. No analysis is used anywhere in this section except in the last part, which states a theorem without proving it.

17.11.1 Functions of a Hermitian matrix

Chapter 12 §08 already applies a function to a matrix, and nothing needs rebuilding. Let $\mathbf{A} \in M_n(F)$ be Hermitian. The operator $\mathbf{x} \mapsto \mathbf{A}\mathbf{x}$ on F^n has an orthonormal basis of eigenvectors (Corollary 12.4.3, Corollary 12.5.2), so it has a spectral resolution

$$\mathbf{A} = \mu_1 \mathbf{P}_1 + \cdots + \mu_k \mathbf{P}_k$$

with $\mu_1, \ldots, \mu_k$ the **distinct** eigenvalues and $\mathbf{P}_j$ the orthogonal projection onto the eigenspace of μ_j (Definition 12.8.1, Theorem 12.8.1). For any function f defined on $\operatorname{spec}(\mathbf{A})$, Definition 12.8.2 sets

$$f(\mathbf{A}) \coloneqq f(\mu_1)\mathbf{P}_1 + \cdots + f(\mu_k)\mathbf{P}_k.$$

In this section f is always defined on an interval I containing $\operatorname{spec}(\mathbf{A})$, and $f(\mathbf{A})$ is formed from the restriction of f to the spectrum: the values of f off the spectrum play no part. Four facts from Theorem 12.8.3 are used constantly.

- **Polynomials give the expected answer** (part (a)): $t \mapsto \alpha + \beta t$ gives $\alpha\mathbf{I} + \beta\mathbf{A}$, and $t \mapsto t^2$ gives $\mathbf{A}^2$.
- **Sums and products go through** (part (b)): $(f+g)(\mathbf{A}) = f(\mathbf{A}) + g(\mathbf{A})$, $(cf)(\mathbf{A}) = c\,f(\mathbf{A})$, and $(fg)(\mathbf{A}) = f(\mathbf{A})g(\mathbf{A})$.
- **Real-valued f gives a Hermitian $f(\mathbf{A})$** (part (c)). This is what lets $f(\mathbf{A})$ be compared with anything in the Loewner order.
- **The eigenvectors do not change** (part (d)): $f(\mathbf{A})$ has the same orthonormal eigenbasis as $\mathbf{A}$.

Two instances have names already. If $0 \notin \operatorname{spec}(\mathbf{A})$, then $t \mapsto 1/t$ gives $\mathbf{A}^{-1}$, as Chapter 12 §08 records right after Definition 12.8.2. If $\mathbf{A} \succeq \mathbf{O}$, then $t \mapsto t^{1/2}$ gives the positive square root $\mathbf{A}^{1/2}$ of Theorem 13.2.1, which Chapter 13 §02 constructed in exactly this way.

17.11.2 Operator monotone and operator convex functions

Chapter 13 §05 introduced the Loewner order $\mathbf{A} \succeq \mathbf{B}$, meaning $\mathbf{A} - \mathbf{B} \succeq \mathbf{O}$ (Definition 13.5.1). With the functional calculus available, the natural question is which functions f carry $\mathbf{A} \succeq \mathbf{B}$ to $f(\mathbf{A}) \succeq f(\mathbf{B})$.

The obvious guess is "the increasing ones", since that is the answer for numbers. It is wrong. The function $t \mapsto t^2$ is increasing on $[0, \infty)$, and Example (Squaring does not preserve the Loewner

order) gives Hermitian 2×2 matrices $\mathbf{A} \succeq \mathbf{B} \succeq \mathbf{O}$ with $\mathbf{A}^2 \not\succeq \mathbf{B}^2$. So the property we want is stronger than being increasing, and it needs its own name.

A function is operator monotone when applying it to Hermitian matrices never breaks the Loewner order, at every size at once.

Definition 17.11.1. Operator Monotone and Operator Convex Functions

Let $I \subseteq \mathbb{R}$ be an interval and let $f : I \to \mathbb{R}$.

(a) f is **operator monotone on** I if, **for every size** $n \geq 1$ and every pair of Hermitian matrices $\mathbf{A}, \mathbf{B} \in M_n(F)$ whose spectra lie in I,

$$\mathbf{A} \succeq \mathbf{B} \quad \Longrightarrow \quad f(\mathbf{A}) \succeq f(\mathbf{B}).$$

(b) f is **operator convex on** I if, **for every size** $n \geq 1$, every pair of Hermitian matrices $\mathbf{A}, \mathbf{B} \in M_n(F)$ whose spectra lie in I, and every $\theta \in [0, 1]$,

$$f(\theta\mathbf{A} + (1-\theta)\mathbf{B}) \preceq \theta f(\mathbf{A}) + (1-\theta) f(\mathbf{B}).$$

It is **operator concave on** I if $-f$ is operator convex on I.

Here $f(\mathbf{A})$ is the matrix of Definition 12.8.2.

In words: the first condition takes any two Hermitian matrices of the same size with spectra inside the domain of f. If the first dominates the second, applying f must preserve that. The second condition is the matrix version of "the chord lies above the graph". Both are infinitely many conditions, one for each size n. A function that passes the test for 2×2 matrices has not yet passed it for 3×3 ones. The clause "for every n" is why the notion is hard, and also why it is interesting.

Everything in the definition makes sense. Since $\operatorname{spec}(\mathbf{A}) \subseteq I$, the matrix $f(\mathbf{A})$ is defined, and it is Hermitian because f is real-valued (Theorem 12.8.3 (c)). So $f(\mathbf{A}) \succeq f(\mathbf{B})$ is a statement about two Hermitian matrices, as Definition 13.5.1 requires. In (b) one more thing needs checking: the spectrum of $\mathbf{C} = \theta\mathbf{A} + (1-\theta)\mathbf{B}$ must lie in I, or $f(\mathbf{C})$ would not be defined. The matrix $\mathbf{C}$ is Hermitian. Let $\mathbf{x}$ be a unit eigenvector of $\mathbf{C}$ for the eigenvalue μ. Then

$$\mu = \langle \mathbf{C}\mathbf{x}, \mathbf{x} \rangle = \theta \langle \mathbf{A}\mathbf{x}, \mathbf{x} \rangle + (1-\theta) \langle \mathbf{B}\mathbf{x}, \mathbf{x} \rangle.$$

By Lemma 13.5.2, $\langle \mathbf{A}\mathbf{x}, \mathbf{x} \rangle$ lies in $[\lambda_n(\mathbf{A}), \lambda_1(\mathbf{A})]$. That interval is contained in I, because its endpoints are eigenvalues of $\mathbf{A}$ and I is an interval. Likewise $\langle \mathbf{B}\mathbf{x}, \mathbf{x} \rangle \in I$. So μ is a convex combination of two points of the interval I, and it therefore lies in I.

The smallest size already says something, and the proof is three lines.

Proposition 17.11.1. Operator Monotone Functions Are Increasing

If f is operator monotone on I, then f is increasing on I.

Proof. Let $s \leq t$ in I, and take $n = 1$, $\mathbf{A} = (t)$, $\mathbf{B} = (s)$. These are Hermitian, their spectra $\{t\}$ and $\{s\}$ lie in I, and $\mathbf{A} - \mathbf{B} = (t - s) \succeq \mathbf{O}$. Hence $f(\mathbf{A}) \succeq f(\mathbf{B})$, and since $f(\mathbf{A}) = (f(t))$ and $f(\mathbf{B}) = (f(s))$, this says $f(t) - f(s) \geq 0$. This shows that f is increasing.

■

The converse is false, and the failure is the reason this section exists.

Warning. **Increasing is not enough.** "Increasing" is exactly the case $n = 1$ of the definition, and sizes $n \geq 2$ are where all the new content lies. The function $t \mapsto t^2$ is increasing on $[0, \infty)$ and is **not** operator monotone there. The 2×2 matrices $\mathbf{A} = \left(\begin{smallmatrix} 2 & 1 \\ 1 & 1 \end{smallmatrix}\right)$ and $\mathbf{B} = \left(\begin{smallmatrix} 1 & 1 \\ 1 & 1 \end{smallmatrix}\right)$ of Example (Squaring does not preserve the Loewner order) satisfy $\mathbf{A} \succeq \mathbf{B} \succeq \mathbf{O}$, yet $\mathbf{A}^2 - \mathbf{B}^2$ has negative determinant. Chapter 13 §05 traced the failure to noncommutativity: the number proof factors $a^2 - b^2$ as $(a-b)(a+b)$, and for matrices that factorization needs $\mathbf{AB} = \mathbf{BA}$.

Quick check. Is $t \mapsto t^3$ operator monotone on $\mathbb{R}$ when only $n = 1$ is tested? What does the answer tell you about $n = 2$?

Solution. For $n = 1$ the condition says only that f is increasing, as the proof of Proposition 17.11.1 shows, and $t \mapsto t^3$ is increasing on $\mathbb{R}$. So it passes at $n = 1$. That tells us nothing about $n = 2$, where the test is genuinely new. In fact $t \mapsto t^3$ fails at $n = 2$ (exercise C3).

17.11.3 The examples we can prove

Four families are within reach, and the first is the simplest. It includes a degenerate case, the constant functions, which respect every order because they flatten everything to one matrix.

Proposition 17.11.2. Increasing Affine Functions

Let $\alpha, \beta \in \mathbb{R}$ with $\beta \geq 0$. Then $f(t) = \alpha + \beta t$ is operator monotone on $\mathbb{R}$. Moreover, every affine function $t \mapsto \alpha + \beta t$, with β of either sign, is both operator convex and operator concave on $\mathbb{R}$.

Proof. By Theorem 12.8.3 (a), $f(\mathbf{A}) = \alpha\mathbf{I} + \beta\mathbf{A}$ for every Hermitian $\mathbf{A}$. Let $\mathbf{A} \succeq \mathbf{B}$ be Hermitian of the same size. Then $f(\mathbf{A}) - f(\mathbf{B}) = \beta(\mathbf{A} - \mathbf{B})$, and for every $\mathbf{x}$,

$$\langle \beta(\mathbf{A} - \mathbf{B})\mathbf{x}, \mathbf{x} \rangle = \beta \langle (\mathbf{A} - \mathbf{B})\mathbf{x}, \mathbf{x} \rangle \;\geq\; 0,$$

a product of two nonnegative numbers. Hence $f(\mathbf{A}) \succeq f(\mathbf{B})$. For the second statement, with $\mathbf{C} = \theta\mathbf{A} + (1-\theta)\mathbf{B}$,

$$\begin{aligned} f(\mathbf{C}) &= \alpha\mathbf{I} + \beta\theta\mathbf{A} + \beta(1-\theta)\mathbf{B} \\ &= \theta(\alpha\mathbf{I} + \beta\mathbf{A}) + (1-\theta)(\alpha\mathbf{I} + \beta\mathbf{B}) = \theta f(\mathbf{A}) + (1-\theta) f(\mathbf{B}), \end{aligned}$$

so the inequality in Definition 17.11.1 (b) holds with equality, for f and for $-f$ alike.

■

The second example is the one that behaves like the square root: $t \mapsto -1/t$. For numbers, $a \geq b > 0$ gives $-1/a \geq -1/b$ because one may divide by the product ab. For matrices that division is unavailable, since $\mathbf{A}$ and $\mathbf{B}$ need not commute. Chapter 13 §05 already found the way round.

Proposition 17.11.3. The Negative Reciprocal Is Operator Monotone

The function $t \mapsto -1/t$ is operator monotone on $(0, \infty)$. Explicitly: if $\mathbf{A}, \mathbf{B} \in M_n(F)$ are Hermitian with $\mathbf{A} \succeq \mathbf{B} \succ \mathbf{O}$, then $\mathbf{A} \succ \mathbf{O}$ and

$$-\mathbf{A}^{-1} \succeq -\mathbf{B}^{-1}, \qquad \text{that is,} \qquad \mathbf{B}^{-1} \succeq \mathbf{A}^{-1}.$$

Idea. A congruence removes the noncommutativity. Conjugating by $\mathbf{B}^{-1/2}$ turns $\mathbf{B}$ into $\mathbf{I}$ and $\mathbf{A}$ into $\mathbf{C} = \mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2} \succeq \mathbf{I}$. Now one of the two matrices is the identity, which commutes with everything, so the argument for numbers applies to the eigenvalues of $\mathbf{C}$: they are all ≥ 1, so those of $\mathbf{C}^{-1}$ are all ≤ 1. Undoing the congruence carries $\mathbf{I} \succeq \mathbf{C}^{-1}$ back to $\mathbf{B}^{-1} \succeq \mathbf{A}^{-1}$. That is exactly the proof of Theorem 13.5.3 (d), so the proposition is that clause read as a statement about the function $t \mapsto -1/t$.

Proof. Let $\mathbf{A} \succeq \mathbf{B}$ be Hermitian with spectra in $(0, \infty)$. By Theorem 13.1.5 (b), a Hermitian matrix whose eigenvalues are all positive is positive definite, so $\mathbf{B} \succ \mathbf{O}$. By Theorem 13.5.3 (d), $\mathbf{A} \succ \mathbf{O}$ and $\mathbf{B}^{-1} \succeq \mathbf{A}^{-1}$. Put $f(t) = -1/t$. By Definition 12.8.2 and the instance $t \mapsto 1/t$ recalled above, $f(\mathbf{A}) = -\mathbf{A}^{-1}$ and $f(\mathbf{B}) = -\mathbf{B}^{-1}$. Hence

$$f(\mathbf{A}) - f(\mathbf{B}) = \mathbf{B}^{-1} - \mathbf{A}^{-1} \succeq \mathbf{O},$$

which is the assertion. The size n was arbitrary, so f is operator monotone on $(0, \infty)$. ∎

Two things make the whitening work, and they are worth separating. First, once one of the two matrices is $\mathbf{I}$, the comparison $\mathbf{C} \succeq \mathbf{I}$ can be settled eigenvalue by eigenvalue, because $\mathbf{I}$ commutes with everything. Second, inversion gets along with congruence: $(\mathbf{S}^{*}\mathbf{M}\mathbf{S})^{-1} = \mathbf{S}^{-1}\mathbf{M}^{-1}(\mathbf{S}^{-1})^{*}$ is again a congruence, and that is what carries the conclusion back from $\mathbf{C}$ to $\mathbf{A}$. For a general f there is no such identity, since $f(\mathbf{S}^{*}\mathbf{M}\mathbf{S})$ is usually not a congruence of $f(\mathbf{M})$, so the argument does not simply transfer to other functions.

To build more examples out of these, one needs to know which ways of combining functions preserve the property.

Proposition 17.11.4. Combining Operator Monotone Functions

Let f and g be operator monotone on I.

(a) For every $c \geq 0$ and every $\alpha \in \mathbb{R}$, the functions $f + g$ and $\alpha + cf$ are operator monotone on I.

(b) For every $s \in \mathbb{R}$, the translate $t \mapsto f(t + s)$ is operator monotone on the interval $I - s = \{t - s : t \in I\}$.

Proof.

(a) Let $\mathbf{A} \succeq \mathbf{B}$ be Hermitian of the same size with spectra in I. By Theorem 12.8.3 (b), $(f+g)(\mathbf{A}) - (f+g)(\mathbf{B}) = (f(\mathbf{A}) - f(\mathbf{B})) + (g(\mathbf{A}) - g(\mathbf{B}))$. Both brackets are $\succeq \mathbf{O}$ by hypothesis, and for every $\mathbf{x}$ the quadratic form of the sum is the sum of the two nonnegative quadratic forms. So the sum is $\succeq \mathbf{O}$ by Definition 13.1.1. Likewise, using (b) and (a) of the same theorem, $(\alpha + cf)(\mathbf{A}) - (\alpha + cf)(\mathbf{B}) = c(f(\mathbf{A}) - f(\mathbf{B}))$, whose quadratic form is $c \geq 0$ times a nonnegative one.

(b) Let $g(t) = f(t+s)$ on $I - s$, and let $\mathbf{A}$ be Hermitian with spectrum in $I - s$, with spectral resolution $\mathbf{A} = \sum_j \mu_j \mathbf{P}_j$. By Theorem 12.8.1 (c) and (d),

$$\mathbf{A} + s\mathbf{I} = \sum_j (\mu_j + s)\mathbf{P}_j,$$

and the numbers $\mu_j + s$ are distinct. The $\mathbf{P}_j$ are nonzero self-adjoint idempotents with $\mathbf{P}_i\mathbf{P}_j = \mathbf{O}$ for $i \neq j$ and $\sum_j \mathbf{P}_j = \mathbf{I}$ (Theorem 12.8.1 (b), (c)). So by Theorem 12.8.2 this is the spectral resolution of $\mathbf{A} + s\mathbf{I}$, whose spectrum $\{\mu_j + s\}$ lies in I. By Definition 12.8.2, applied to each side,

$$f(\mathbf{A} + s\mathbf{I}) = \sum_j f(\mu_j + s)\mathbf{P}_j = g(\mathbf{A}).$$

Now if $\mathbf{A} \succeq \mathbf{B}$ with spectra in $I - s$, then $(\mathbf{A} + s\mathbf{I}) - (\mathbf{B} + s\mathbf{I}) = \mathbf{A} - \mathbf{B} \succeq \mathbf{O}$, and both shifted matrices have spectra in I. Hence $g(\mathbf{A}) = f(\mathbf{A} + s\mathbf{I}) \succeq f(\mathbf{B} + s\mathbf{I}) = g(\mathbf{B})$.

■

Combining the last two propositions produces a whole one-parameter family. It will reappear in the final part of the section as the building block of every operator monotone function.

Corollary 17.11.5. Shifted Reciprocals

For every $s \geq 0$, the function $t \mapsto -1/(t+s)$ is operator monotone on $(0, \infty)$. For every $s > 0$, so is

$$t \longmapsto \frac{t}{t+s} = 1 - \frac{s}{t+s}.$$

Proof. By Proposition 17.11.3 and Proposition 17.11.4 (b), $t \mapsto -1/(t+s)$ is operator monotone on $(0, \infty) - s = (-s, \infty)$. Every Hermitian matrix with spectrum in $(0, \infty)$ has spectrum in $(-s, \infty)$, so it is in particular operator monotone on $(0, \infty)$. For the second function, $t/(t+s) = 1 + s \cdot (-1/(t+s))$ with $s > 0$, so Proposition 17.11.4 (a) applies.

■

The fourth example is the one that started the section, and it is already proved. By Proposition 13.5.4, $\mathbf{A} \succeq \mathbf{B} \succeq \mathbf{O}$ implies $\mathbf{A}^{1/2} \succeq \mathbf{B}^{1/2}$. A Hermitian matrix has spectrum in $[0, \infty)$ exactly when it is $\succeq \mathbf{O}$ (Theorem 13.1.4 (b)), and $\mathbf{A}^{1/2}$ is $f(\mathbf{A})$ for $f(t) = t^{1/2}$. So Chapter 13 §05's proposition says precisely that $t \mapsto t^{1/2}$ **is operator monotone on** $[0, \infty)$. It is not re-proved here. Its proof was a different kind of argument, an eigenvector of $\mathbf{A}^{1/2} - \mathbf{B}^{1/2}$ pushed to a contradiction, and it too never needs $\mathbf{A}$ and $\mathbf{B}$ to commute.

A nonexample by minimal change. The function $t \mapsto t = t^1$ is operator monotone on $[0, \infty)$ by Proposition 17.11.2. Raise the exponent from 1 to 2. The function is still increasing, still a polynomial, and still satisfies the case $n = 1$. What fails is the clause "for every n", already at $n = 2$, with the witness of Example (Squaring does not preserve the Loewner order). Lowering the exponent from 1 to $\frac{1}{2}$ instead keeps the property. Among the powers, then, the dividing line lies somewhere around the exponent 1. Chapter 13 §05 said where it lies exactly, and the last part of this section returns to that.

Why "for every n". Each size imposes its own condition, and the conditions differ. At $n = 1$, $t \mapsto t^2$ passes on $[0, \infty)$ and at $n = 2$ it fails. A definition at one fixed size would admit functions that break at the next size, and it would not say what the functional calculus does to matrices in general. Requiring every size at once makes the class small enough to classify completely, which is what Loewner's theorem at the end of this section does. The name comes from the same place as the name of the order: both are Loewner's.

17.11.4 Eigenvalues move monotonically for every increasing function

Operator monotonicity is a strong requirement, and a weaker one comes for free. §02 proved that the whole ordered list of eigenvalues respects the Loewner order: by Corollary 17.2.4, $\mathbf{A} \succeq \mathbf{B}$ implies $\lambda_i(\mathbf{A}) \geq \lambda_i(\mathbf{B})$ for **every** i. That paid the debt Chapter 13 §05 recorded after Theorem 13.5.3, and it holds with no hypothesis on commuting. Composing with an increasing f keeps it, once one knows what f does to the ordered list of eigenvalues.

Lemma 17.11.6. The Eigenvalue List of an Increasing Function

Let $\mathbf{A} \in M_n(F)$ be Hermitian with $\operatorname{spec}(\mathbf{A}) \subseteq I$, and let $f\colon I \to \mathbb{R}$ be increasing. Then $f(\mathbf{A})$ is Hermitian and

$$\lambda_i(f(\mathbf{A})) = f(\lambda_i(\mathbf{A})) \qquad (i = 1, \dots, n).$$

Proof. The matrix $f(\mathbf{A})$ is Hermitian by Theorem 12.8.3 (c). By Corollary 12.4.3 (over $\mathbb{C}$) or Corollary 12.5.2 (over $\mathbb{R}$), there is an orthonormal basis $(\mathbf{q}_1, \dots, \mathbf{q}_n)$ of F^n with $\mathbf{A}\mathbf{q}_i = \lambda_i(\mathbf{A})\mathbf{q}_i$. Let $\mathbf{A} = \sum_j \mu_j \mathbf{P}_j$ be the spectral resolution, and fix i. Exactly one of the distinct eigenvalues equals $\lambda_i(\mathbf{A})$; call it μ_{j_0}. Then $\mathbf{q}_i$ lies in the eigenspace of μ_{j_0}, which is the image of $\mathbf{P}_{j_0}$, so $\mathbf{P}_{j_0}\mathbf{q}_i = \mathbf{q}_i$. For $j \neq j_0$, the eigenspace of μ_{j_0} is orthogonal to that of μ_j (Theorem 12.8.1 (a)), so $\mathbf{P}_j\mathbf{q}_i = \mathbf{0}$. By Definition 12.8.2,

$$f(\mathbf{A})\mathbf{q}_i = \sum_j f(\mu_j)\mathbf{P}_j\mathbf{q}_i = f(\mu_{j_0})\mathbf{q}_i = f(\lambda_i(\mathbf{A}))\mathbf{q}_i.$$

Thus $(\mathbf{q}_1, \dots, \mathbf{q}_n)$ is an orthonormal basis of eigenvectors of $f(\mathbf{A})$, and $f(\lambda_1(\mathbf{A})), \dots, f(\lambda_n(\mathbf{A}))$ is its list of eigenvalues with multiplicity. Since $\lambda_1(\mathbf{A}) \geq \cdots \geq \lambda_n(\mathbf{A})$ and f is increasing, this list is already in decreasing order. A decreasing list of all the eigenvalues with multiplicity is the list $\lambda_1(f(\mathbf{A})) \geq \cdots \geq \lambda_n(f(\mathbf{A}))$, so the two agree term by term.

■

The lemma needs f increasing only for the order of the list. For a general f, the numbers $f(\lambda_i(\mathbf{A}))$ are still the eigenvalues of $f(\mathbf{A})$, but they may have to be re-sorted. Two lines now

give the eigenvalue-level statement.

Proposition 17.11.7. Every Increasing Function Is Monotone on Eigenvalues

Let $f\colon I \to \mathbb{R}$ be increasing, and let $\mathbf{A}, \mathbf{B} \in M_n(F)$ be Hermitian with spectra in I and $\mathbf{A} \succeq \mathbf{B}$. Then

$$\lambda_i(f(\mathbf{A})) \;\geq\; \lambda_i(f(\mathbf{B})) \qquad \text{for every } i = 1, \dots, n.$$

Proof. By Corollary 17.2.4, $\lambda_i(\mathbf{A}) \geq \lambda_i(\mathbf{B})$ for every i. Since f is increasing, $f(\lambda_i(\mathbf{A})) \geq f(\lambda_i(\mathbf{B}))$. By Lemma 17.11.6, applied to $\mathbf{A}$ and to $\mathbf{B}$, the two sides are $\lambda_i(f(\mathbf{A}))$ and $\lambda_i(f(\mathbf{B}))$.

■

So the squaring map, which is not operator monotone, is nevertheless monotone on eigenvalues. The Chapter 13 witness shows both at once.

Example. The Squares Are Not Ordered, but Their Eigenvalues Are

Let $\mathbf{A} = \left(\begin{smallmatrix} 2 & 1 \\ 1 & 1 \end{smallmatrix}\right)$ and $\mathbf{B} = \left(\begin{smallmatrix} 1 & 1 \\ 1 & 1 \end{smallmatrix}\right)$, the pair of Example (Squaring does not preserve the Loewner order), for which $\mathbf{A} \succeq \mathbf{B} \succeq \mathbf{O}$ and $\mathbf{A}^2 \not\succeq \mathbf{B}^2$. Compute the eigenvalues of $\mathbf{A}^2$ and $\mathbf{B}^2$ and compare them index by index.

Solution. From Example (Squaring does not preserve the Loewner order), $\mathbf{A}^2 = \left(\begin{smallmatrix} 5 & 3 \\ 3 & 2 \end{smallmatrix}\right)$ and $\mathbf{B}^2 = \left(\begin{smallmatrix} 2 & 2 \\ 2 & 2 \end{smallmatrix}\right)$. The first has trace 7 and determinant $10 - 9 = 1$, so its eigenvalues are the roots of $x^2 - 7x + 1$:

$$\lambda_1(\mathbf{A}^2) = \frac{7 + 3\sqrt{5}}{2} \approx 6.854, \qquad \lambda_2(\mathbf{A}^2) = \frac{7 - 3\sqrt{5}}{2} \approx 0.146.$$

The second has trace 4 and determinant 0, so $\lambda_1(\mathbf{B}^2) = 4$ and $\lambda_2(\mathbf{B}^2) = 0$. Index by index: $\frac{7+3\sqrt{5}}{2} \geq 4$ because $3\sqrt{5} \geq 1$, and $\frac{7-3\sqrt{5}}{2} \geq 0$ because $49 \geq 45$. So $\lambda_i(\mathbf{A}^2) \geq \lambda_i(\mathbf{B}^2)$ for $i = 1, 2$, as Proposition 17.11.7 predicts for the increasing function $t \mapsto t^2$ on $[0, \infty)$. Nevertheless $\mathbf{A}^2 - \mathbf{B}^2$ is not $\succeq \mathbf{O}$.

So for every increasing f, the eigenvalues of $f(\mathbf{A})$ dominate those of $f(\mathbf{B})$ one index at a time. What operator monotonicity adds is that the *matrices* are ordered, and the example shows that this is strictly more.

Warning. **Eigenvalues in order do not make matrices in order.** It is tempting to read Proposition 17.11.7 as "almost" operator monotonicity. It is not, because $\lambda_i(\mathbf{M}) \geq \lambda_i(\mathbf{N})$ for every i does **not** imply $\mathbf{M} \succeq \mathbf{N}$. Take $\mathbf{M} = \operatorname{diag}(1, 0)$ and $\mathbf{N} = \operatorname{diag}(0, 1)$: they have the same eigenvalues $1 \geq 0$, yet $\mathbf{M} - \mathbf{N} = \operatorname{diag}(1, -1)$ is indefinite. The Loewner order sees the eigenvectors as well as the eigenvalues, and the eigenvalue list forgets them.

17.11.5 Operator convexity, and where the two notions part

Operator convexity is the second notion of Definition 17.11.1, and one might expect it to behave like operator monotonicity. It does not: the function that fails to be operator monotone turns out to be operator convex.

Proposition 17.11.8. The Square Is Operator Convex

The function $t \mapsto t^2$ is operator convex on $\mathbb{R}$. Explicitly: for Hermitian $\mathbf{A}, \mathbf{B} \in M_n(F)$ and $\theta \in [0,1]$,

$$\theta \mathbf{A}^2 + (1-\theta)\mathbf{B}^2 - \big(\theta \mathbf{A} + (1-\theta)\mathbf{B}\big)^2 = \theta(1-\theta)(\mathbf{A}-\mathbf{B})^2 \succeq \mathbf{O}.$$

Idea. For numbers, $\theta a^2 + (1-\theta)b^2 - (\theta a + (1-\theta)b)^2 = \theta(1-\theta)(a-b)^2$, which is a variance. The matrix version of the expansion produces $\mathbf{AB} + \mathbf{BA}$ where the number version has $2ab$, and that is exactly the cross term of $(\mathbf{A}-\mathbf{B})^2 = \mathbf{A}^2 - \mathbf{AB} - \mathbf{BA} + \mathbf{B}^2$. So this time noncommutativity cancels itself out and costs nothing. The square of a Hermitian matrix is $\succeq \mathbf{O}$, which finishes the proof.

Proof. By Theorem 12.8.3 (a), the function $t \mapsto t^2$ sends a Hermitian matrix to its square, so the displayed identity is what Definition 17.11.1 (b) requires. Expanding, and keeping the order of every product,

$$\big(\theta \mathbf{A} + (1-\theta)\mathbf{B}\big)^2 = \theta^2 \mathbf{A}^2 + \theta(1-\theta)(\mathbf{AB} + \mathbf{BA}) + (1-\theta)^2 \mathbf{B}^2.$$

Subtracting this from $\theta \mathbf{A}^2 + (1-\theta)\mathbf{B}^2$, and using $\theta - \theta^2 = \theta(1-\theta) = (1-\theta) - (1-\theta)^2$,

$$\theta \mathbf{A}^2 + (1-\theta)\mathbf{B}^2 - \big(\theta \mathbf{A} + (1-\theta)\mathbf{B}\big)^2 = \theta(1-\theta)(\mathbf{A}^2 - \mathbf{AB} - \mathbf{BA} + \mathbf{B}^2) = \theta(1-\theta)(\mathbf{A}-\mathbf{B})^2.$$

The matrix $\mathbf{D} = \mathbf{A} - \mathbf{B}$ is Hermitian, so $\mathbf{D}^2 = \mathbf{D}^* \mathbf{D} \succeq \mathbf{O}$ by Theorem 13.1.4 (c). Since $\theta(1-\theta) \geq 0$, the quadratic form of $\theta(1-\theta)\mathbf{D}^2$ is nonnegative, which proves the inequality. ∎

The contrast is now complete, and it runs both ways.

- $t \mapsto t^2$ is **operator convex but not operator monotone** on $[0, \infty)$: Proposition 17.11.8 and Example (Squaring does not preserve the Loewner order).
- $t \mapsto -1/t$ is **operator monotone but not operator convex** on $(0, \infty)$: Proposition 17.11.3, and at $n = 1$ the scalar function $-1/t$ is not even convex, since with $a = 1$, $b = 3$, and $\theta = \frac{1}{2}$ the value at the midpoint is $-\frac{1}{2}$, above the chord's $-\frac{2}{3}$.

So neither notion implies the other. Each implies its scalar version (increasing, convex) by taking $n = 1$, and they strengthen those versions in different directions. Exercise C2 adds that $t \mapsto 1/t$

is operator convex on $(0,\infty)$, so $-1/t$ is operator monotone *and* operator concave there.

17.11.6 What is not proved here

The question the section opened with has a complete answer, and this chapter does not prove it. Here is the statement, so that the reader knows what the answer looks like and which of the facts above are pieces of it.

Theorem 17.11.9. Loewner's Theorem, Statement

Let $f\colon (0,\infty)\to\mathbb{R}$. The following are equivalent.

(i) f is operator monotone on $(0,\infty)$.

(ii) There are $\alpha\in\mathbb{R}$, $\beta\geq 0$, and a positive measure ν on $[0,\infty)$ with $\int_{[0,\infty)}\frac{\mathrm{d}\nu(s)}{1+s^2}<\infty$, such that for every $t>0$

$$f(t)=\alpha+\beta t + \int_{[0,\infty)}\Big(\frac{s}{1+s^2}-\frac{1}{t+s}\Big)\,\mathrm{d}\nu(s).$$

(iii) f is the restriction to $(0,\infty)$ of a function that is analytic on $\mathbb{C}\backslash(-\infty,0]$ and sends every point of the upper half-plane $\{z : \operatorname{Im} z>0\}$ to a point with nonnegative imaginary part.

We do not prove this theorem here. Its proof needs complex analysis and some measure theory, which lie outside Chapter 16's list (A1)–(A6); Chapter 21 proves it. Nothing in this chapter depends on it.

Read (ii) against what was proved above. The functions it superposes are $t\mapsto -1/(t+s)$ for $s\geq 0$, the translates of $-1/t$ that Corollary 17.11.5 showed to be operator monotone. The constants $s/(1+s^2)$ do not depend on t and are there only to make the integral converge. The term $\alpha+\beta t$ is Proposition 17.11.2. So the theorem says that **the examples of this section are, in a precise sense, all the examples there are**: every operator monotone function on $(0,\infty)$ is an increasing affine function plus a nonnegative superposition of shifted reciprocals. The easy direction, (ii) $\Rightarrow$ (i), is essentially Proposition 17.11.4 applied to an integral rather than a finite sum. What this book lacks is the analysis that carries the Loewner order through an integral, which is not on the list (A1)–(A6) in Chapter 16's introduction. The hard direction, (i) $\Rightarrow$ (ii), is the classification itself.

Condition (iii) answers the opening question in one line. For z in the upper half-plane, $z^{1/2}$ (the branch that is positive on $(0,\infty)$) halves the argument of z, so its argument stays between 0 and $\pi/2$ and $z^{1/2}$ stays in the upper half-plane. The square doubles the argument, which can pass π. **The square root keeps the upper half-plane, and the square folds it over.**

Quick check. Find a point z of the upper half-plane whose square has negative imaginary part.

Solution. Take $z=-1+i$, with $\operatorname{Im} z=1>0$. Then $z^2=1-2i+i^2=-2i$, with imaginary

part $-2 < 0$. So $z \mapsto z^2$ fails condition (iii) of Theorem 17.11.9, consistently with Example (Squaring does not preserve the Loewner order).

The same test settles the powers $t \mapsto t^p$. For $0 < p < 1$, the function z^p multiplies the argument by p and keeps the upper half-plane, and indeed for $t > 0$ (a standard integral, not proved here)

$$t^p = \frac{\sin p\pi}{\pi} \int_0^\infty s^{p-1} \frac{t}{t+s}\, \mathrm{d}s,$$

a positive superposition of the functions $t/(t+s)$ of Corollary 17.11.5. For $p > 1$, the argument of z^p passes π for some z in the upper half-plane, and the test fails. This is the dividing line at exponent 1 that Chapter 13 §05 announced. Both halves, that t^p is operator monotone on $(0, \infty)$ for every $p \in (0, 1)$ and that it is not for any $p > 1$, are proved in Chapter 21, with Loewner's theorem. What is proved so far is the case $p = \frac{1}{2}$ (Proposition 13.5.4) and, by applying it repeatedly, the cases $p = 1/2^k$ (exercise C1).

17.11.7 Summary and transfer

The chapter began with a list of eigenvalues that the book had been able to *order* since Chapter 13 without being able to *describe*. It ends with each eigenvalue described as an optimization, and with a set of comparison theorems that follow from that description. Five moves did the work, and each transfers beyond this chapter.

- **The dimension count.** Two subspaces of an n-dimensional space whose dimensions add to more than n share a nonzero vector. This one line drives Courant–Fischer (Theorem 17.2.3) and Weyl's inequalities (§03) directly, and through Courant–Fischer everything after them. *Transfer:* whenever an extremal problem over vectors has to be compared with a fixed subspace, count dimensions before computing anything.
- **Min–max.** Describe a quantity as a minimum over subspaces of a maximum over vectors, and prove it with two inequalities of different characters: one exhibits a witness subspace and computes, and the other takes an arbitrary subspace and intersects it with a span of eigenvectors. Applying the result to $-\mathbf{A}$ gives the dual form for free. *Transfer:* a description that names no eigenvector is what makes two different matrices comparable. The same shape describes singular values (§09) and inertia (§10).
- **Compress and interlace.** Restricting a Hermitian form to a subspace gives a smaller Hermitian matrix whose eigenvalues are trapped by the original's: compressing to a k-dimensional subspace puts λ_i of the piece between λ_i and λ_{i+n-k} of the whole, which are consecutive when $k = n - 1$. Deleting a row and column (§04), compressing to an orthonormal frame (§06) and deleting a row or a column of a rectangular matrix (§09) are all this move; adding a rank-one term (§05) reaches the same interlaced picture by a different route, the Loewner order together with Weyl's rank bound. *Transfer:* to bound the spectrum of a piece, look for the whole of which it is a compression.
- **Perturb in order.** Compare $\lambda_i(\mathbf{A}+\mathbf{E})$ with $\lambda_i(\mathbf{A})$ index by index, never "each eigenvalue with its nearest neighbor". For Hermitian perturbations the Lipschitz constant is 1 in the spectral norm, independent of n, of the gaps, and of the eigenvectors (§03). For singular values it needs no hypothesis on $\mathbf{E}$ at all (§09). *Transfer:* where the Jordan block of Chapter 16 §07 showed that eigenvalues of general matrices can move like $\varepsilon^{1/k}$, the Hermitian and singular-value problems are perfectly conditioned. Reduce a question to one of them whenever possible.
- **Majorize.** When the right comparison is between *sums* of the largest entries rather than individual ones, compare partial sums of decreasing rearrangements. The diagonal of a Hermitian matrix is majorized by its spectrum, and by the Schur–Horn theorem

majorization describes the possible diagonals *exactly* (§08); and the Ky Fan and Lidskii sums (§§06–07) are of the same type. *Transfer:* a family of partial-sum inequalities indexed by k, together with equality of the totals, is a majorization in disguise, and writing it as one lets the whole family be used at once.

Two supporting moves recur under these five: **apply the result to $-\mathbf{A}$**, which turns every max–min into a min–max and every top-k statement into a bottom-k one, and **reduce singular values to eigenvalues**, through $\mathbf{A}^*\mathbf{A}$ or the Hermitian dilation. This section also reused a move from Chapter 13 §05, **whiten, then compare**: use a congruence to make one of the two matrices $\mathbf{I}$. It works for $t \mapsto -1/t$ because inversion turns a congruence into a congruence, and not for a general f.

The chapter also paid a series of promises made in earlier chapters, some of them long ago.

The promise	Made in	Paid by
The maximization route to eigenvalues "returns in Chapter 17, where the Rayleigh quotient and the Courant–Fischer theorem describe every eigenvalue of a self-adjoint operator by optimization"	Chapter 12 §02	§01 (The Rayleigh quotient) and §02 (The min–max theorem), Theorem 17.2.3
"Chapter 17 characterizes the eigenvalues of a Hermitian matrix as maxima and minima of $\langle \mathbf{Ax}, \mathbf{x} \rangle$ — at which point the test becomes the theory"	Chapter 13 §13	§01 and §02, Theorem 17.2.3
$\mathbf{A} \succeq \mathbf{B}$ gives $\lambda_i(\mathbf{A}) \geq \lambda_i(\mathbf{B})$ for **every** i, not only $i = 1$ and $i = n$ as in Theorem 13.5.3 (b)	Chapter 13 §05	§02, Corollary 17.2.4
"The eigenvalues of a Hermitian matrix move by at most $\|\mathbf{E}\|_2$ under a Hermitian perturbation"	Chapter 16 §07	§03 (How far eigenvalues move), Weyl's perturbation bound
Lemma 13.11.2 is "the extreme case of a family of statements about sums of the largest eigenvalues"	Chapter 13 §11	§06 (Subspaces and Sums of Eigenvalues), Ky Fan's theorem, of which it is a special case
The second half of the approximation theorem, in the spectral norm	Chapter 13 §11 and Chapter 16 §08	§09 (The same theory for singular values), Eckart–Young in the spectral norm
Interlacing of the zeros of consecutive orthogonal polynomials, "Chapter 17's business"	Chapter 11 §10	§04 (Deleting a row and a column): strict interlacing, through the bordered-matrix criterion

The promise	Made in	Paid by
The operator monotone functions named in Chapter 13 §05	Chapter 13 §05	§11: definition and first examples; t^s, $\log t$, and Loewner's theorem remain Chapter 21's, as that section said

Four earlier theorems also received second proofs from the variational side, each of which shows something the first proof did not. The positivity of principal submatrices of a positive semidefinite matrix now follows from interlacing (§04, Deleting a row and a column). Hadamard's inequality Theorem 13.7.3 follows from AM–GM after a diagonal scaling, with Schur's theorem supplying the equal totals (§08, The diagonal and the spectrum). Sylvester's law of inertia Theorem 14.5.1 is now a statement about the dimensions of subspaces, with no basis in it (§10, Inertia, again and additively). And the block test Theorem 13.6.1 is now a consequence of inertia additivity (§10).

Some things the chapter did not do. Bounds for the eigenvalues of non-Hermitian matrices under perturbation, beginning with Bauer–Fike, are Chapter 20's. Chapter 21 has Loewner's theorem with the powers t^p, Schur-concave functions, the strongest form of Lidskii's inequality, and Eckart–Young in every unitarily invariant norm. The variational principle itself, though, is now a tool the book can use without comment: an eigenvalue of a Hermitian matrix is a min–max, and comparison theorems follow from that.

17.11.8 Exercises

17.11.8.1 A. Check your understanding

Exercise. A1

(a) Define "f is operator monotone on I", and say what the clause "for every n" ranges over.

(b) Explain why an operator monotone function is increasing, and name an increasing function that is not operator monotone.

(c) Decide whether the following is correct, and justify your answer: if $f\colon I \to \mathbb{R}$ is increasing and $\mathbf{A} \succeq \mathbf{B}$ are Hermitian with spectra in I, then $\lambda_i(f(\mathbf{A})) \geq \lambda_i(f(\mathbf{B}))$ for every i.

(d) Decide whether the following is correct, and justify your answer: if $\mathbf{A}, \mathbf{B}$ are Hermitian of the same size and $\lambda_i(\mathbf{A}) \geq \lambda_i(\mathbf{B})$ for every i, then $\mathbf{A} \succeq \mathbf{B}$.

(e) Which of $t \mapsto t^2$ and $t \mapsto -1/t$ is operator monotone on $(0, \infty)$, and which is operator convex there?

(f) What does this section prove about Loewner's theorem?

Solution.

(a) For every $n \geq 1$ and all Hermitian $\mathbf{A}, \mathbf{B} \in M_n(F)$ with spectra in I, $\mathbf{A} \succeq \mathbf{B}$ implies $f(\mathbf{A}) \succeq f(\mathbf{B})$, where $f(\mathbf{A})$ is formed by Definition 12.8.2 (Definition 17.11.1 (a)). The clause ranges over all matrix sizes: the implication must hold for 1×1, 2×2, 3×3, ... matrices simultaneously.

(b) Taking $n = 1$ gives $f(t) \geq f(s)$ whenever $t \geq s$ in I (Proposition 17.11.1). The function $t \mapsto t^2$ on $[0, \infty)$ is increasing and is not operator monotone (Example (Squaring does not preserve the Loewner order)).

(c) Correct. This is Proposition 17.11.7: Corollary 17.2.4 orders the eigenvalues of $\mathbf{A}$ and $\mathbf{B}$ index by index, f preserves each inequality, and Lemma 17.11.6 identifies $f(\lambda_i(\mathbf{A}))$ with $\lambda_i(f(\mathbf{A}))$.

(d) Incorrect. $\operatorname{diag}(1, 0)$ and $\operatorname{diag}(0, 1)$ have the same eigenvalues, but their difference $\operatorname{diag}(1, -1)$ is indefinite, so neither dominates the other.

(e) $t \mapsto -1/t$ is operator monotone on $(0, \infty)$ (Proposition 17.11.3) and is not operator convex, since already the scalar function is not convex: at 1, 3, and their midpoint 2, $-\frac{1}{2} > \frac{1}{2}(-1 - \frac{1}{3}) = -\frac{2}{3}$. $t \mapsto t^2$ is operator convex (Proposition 17.11.8) and not operator monotone on $(0, \infty)$. The witness of Example (Squaring does not preserve the Loewner order) has 0 in the spectrum of $\mathbf{B}$, but the pair $\mathbf{A} = \left(\begin{smallmatrix} 3 & 1 \\ 1 & 1 \end{smallmatrix}\right) \succeq \mathbf{B} = \left(\begin{smallmatrix} 2 & 1 \\ 1 & 1 \end{smallmatrix}\right) \succ \mathbf{O}$ works: $\mathbf{A}^2 - \mathbf{B}^2 = \left(\begin{smallmatrix} 5 & 1 \\ 1 & 0 \end{smallmatrix}\right)$ has determinant -1.

(f) Only the statement (Theorem 17.11.9) and the fact that its building blocks $t \mapsto -1/(t + s)$ are operator monotone (Corollary 17.11.5). The theorem itself is proved in Chapter 21.

17.11.8.2 B. Practice

Exercise. B1: Which are operator monotone?

Determine which of the following functions are operator monotone on the interval given. Justify your answer.

(i) $f(t) = 5 + 4t$ on $\mathbb{R}$.

(ii) $f(t) = 3 - 2t$ on $\mathbb{R}$.

(iii) $f(t) = 2 - \dfrac{3}{t+1}$ on $(0, \infty)$.

(iv) $f(t) = \dfrac{t}{t+2}$ on $(0, \infty)$.

(v) $f(t) = t^2$ on $[0, \infty)$.

Solution.

(i) Operator monotone, by Proposition 17.11.2 with $\beta = 4 \geq 0$.

(ii) Not operator monotone. It is not increasing, since $f(1) = 1 < 3 = f(0)$, so it fails at $n = 1$ with $\mathbf{A} = (1) \succeq \mathbf{B} = \mathbf{O}_1$; compare Proposition 17.11.1.

(iii) Operator monotone. By Corollary 17.11.5 with $s = 1$, $t \mapsto -1/(t + 1)$ is operator monotone on $(0, \infty)$, and $f = 2 + 3 \cdot (-1/(t + 1))$ with $3 \geq 0$, so Proposition 17.11.4 (a) applies.

(iv) Operator monotone, by Corollary 17.11.5 with $s = 2 > 0$.

(v) Not operator monotone: Example (Squaring does not preserve the Loewner order) gives $\mathbf{A} \succeq \mathbf{B} \succeq \mathbf{O}$, with spectra $\{(3 \pm \sqrt{5})/2\}$ and $\{0, 2\}$ in $[0, \infty)$, and $\mathbf{A}^2 \not\succeq \mathbf{B}^2$.

Exercise. B2: One pair, three functions

Let

$$\mathbf{A} = \begin{pmatrix} 3 & 1 \\ 1 & 1 \end{pmatrix}, \qquad \mathbf{B} = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}.$$

(a) Show that $\mathbf{A} \succeq \mathbf{B} \succ \mathbf{O}$.

(b) Compute $\mathbf{A}^{-1}$ and $\mathbf{B}^{-1}$, and verify directly that $\mathbf{B}^{-1} \succeq \mathbf{A}^{-1}$, as Proposition 17.11.3 predicts.

(c) Show that $\mathbf{A}^2 \not\succeq \mathbf{B}^2$ by exhibiting a vector $\mathbf{x}$ with $\mathbf{x}^\top(\mathbf{A}^2 - \mathbf{B}^2)\mathbf{x} < 0$.

(d) Compute the eigenvalues of $\mathbf{A}^2$ and $\mathbf{B}^2$, and verify that $\lambda_i(\mathbf{A}^2) \geq \lambda_i(\mathbf{B}^2)$ for $i = 1, 2$. Hence explain which result of this section (c) and (d) illustrate together.

Solution.

(a) $\mathbf{A} - \mathbf{B} = \operatorname{diag}(1, 0)$, a diagonal matrix with nonnegative entries, so $\mathbf{A} - \mathbf{B} \succeq \mathbf{O}$ and $\mathbf{A} \succeq \mathbf{B}$. The leading principal minors of $\mathbf{B}$ are 2 and $2 - 1 = 1$, both positive, so $\mathbf{B} \succ \mathbf{O}$ by Sylvester's criterion (Theorem 13.1.5 (d)).

(b) $\det \mathbf{A} = 2$ and $\det \mathbf{B} = 1$, so

$$\mathbf{A}^{-1} = \frac{1}{2}\begin{pmatrix} 1 & -1 \\ -1 & 3 \end{pmatrix}, \qquad \mathbf{B}^{-1} = \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix}, \qquad \mathbf{B}^{-1} - \mathbf{A}^{-1} = \frac{1}{2}\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}.$$

For every $\mathbf{x} \in \mathbb{R}^2$, $\mathbf{x}^\top(\mathbf{B}^{-1} - \mathbf{A}^{-1})\mathbf{x} = \frac{1}{2}(x_1 - x_2)^2 \geq 0$, so $\mathbf{B}^{-1} \succeq \mathbf{A}^{-1}$.

(c) $\mathbf{A}^2 = \left(\begin{smallmatrix} 10 & 4 \\ 4 & 2 \end{smallmatrix}\right)$ and $\mathbf{B}^2 = \left(\begin{smallmatrix} 5 & 3 \\ 3 & 2 \end{smallmatrix}\right)$, so $\mathbf{A}^2 - \mathbf{B}^2 = \left(\begin{smallmatrix} 5 & 1 \\ 1 & 0 \end{smallmatrix}\right)$, with determinant $-1 < 0$. With $\mathbf{x} = (1, -3)$,

$$\mathbf{x}^\top(\mathbf{A}^2 - \mathbf{B}^2)\mathbf{x} = 5 + 2 \cdot (1)(-3) + 0 = -1 < 0.$$

(d) $\mathbf{A}^2$ has trace 12 and determinant $(\det \mathbf{A})^2 = 4$, so its eigenvalues are $6 \pm \sqrt{32} = 6 \pm 4\sqrt{2}$, approximately 11.657 and 0.343. $\mathbf{B}^2$ has trace 7 and determinant 1, so its eigenvalues are $(7 \pm 3\sqrt{5})/2$, approximately 6.854 and 0.146. For $i = 1$, $12 + 8\sqrt{2} \geq 7 + 3\sqrt{5}$ because $8\sqrt{2} > 3\sqrt{5}$. For $i = 2$, $12 - 8\sqrt{2} \geq 7 - 3\sqrt{5}$ is $5 + 3\sqrt{5} \geq 8\sqrt{2}$. Both sides are positive, and squaring gives $70 + 30\sqrt{5} \geq 128$, that is $\sqrt{5} \geq \frac{29}{15}$, which holds because $5 \geq \frac{841}{225}$. Hence (c) and (d) together illustrate Proposition 17.11.7 and the warning after it: the increasing function $t \mapsto t^2$ orders the eigenvalues but not the matrices.

Exercise. B3: The convexity identity in numbers

With $\mathbf{A}$ and $\mathbf{B}$ as in B2 and $\theta = \frac{1}{2}$, compute

$$\mathbf{M} = \tfrac{1}{2}\mathbf{A}^2 + \tfrac{1}{2}\mathbf{B}^2 - \left(\tfrac{1}{2}\mathbf{A} + \tfrac{1}{2}\mathbf{B}\right)^2$$

directly, confirm that $\mathbf{M} = \frac{1}{4}(\mathbf{A} - \mathbf{B})^2$, and hence conclude that $\left(\frac{1}{2}(\mathbf{A} + \mathbf{B})\right)^2 \preceq \frac{1}{2}(\mathbf{A}^2 + \mathbf{B}^2)$.

Solution. From B2, $\frac{1}{2}\mathbf{A}^2 + \frac{1}{2}\mathbf{B}^2 = \frac{1}{2}\left(\begin{smallmatrix} 15 & 7 \\ 7 & 4 \end{smallmatrix}\right) = \begin{pmatrix} 15/2 & 7/2 \\ 7/2 & 2 \end{pmatrix}$. Next $\frac{1}{2}(\mathbf{A} + \mathbf{B}) = \begin{pmatrix} 5/2 & 1 \\ 1 & 1 \end{pmatrix}$, whose

square is

$$\begin{pmatrix} \frac{25}{4}+1 & \frac{5}{2}+1 \\ \frac{5}{2}+1 & 1+1 \end{pmatrix} = \begin{pmatrix} \frac{29}{4} & \frac{7}{2} \\ \frac{7}{2} & 2 \end{pmatrix}.$$

Hence $\mathbf{M} = \begin{pmatrix} 15/2-29/4 & 0 \\ 0 & 0 \end{pmatrix} = \operatorname{diag}(\frac{1}{4}, 0)$. On the other side, $\mathbf{A}-\mathbf{B} = \operatorname{diag}(1, 0)$, so $\frac{1}{4}(\mathbf{A}-\mathbf{B})^2 = \operatorname{diag}(\frac{1}{4}, 0) = \mathbf{M}$, as Proposition 17.11.8 says with $\theta(1-\theta) = \frac{1}{4}$. Since $\operatorname{diag}(\frac{1}{4}, 0)$ is diagonal with nonnegative entries, $\mathbf{M} \succeq \mathbf{O}$, which is the displayed inequality.

17.11.8.3 C. Going deeper

Exercise. C1: Iterated square roots

(a) Let $\mathbf{A} \succeq \mathbf{B} \succeq \mathbf{O}$. Prove that $\mathbf{A}^{1/4} \succeq \mathbf{B}^{1/4}$, where $\mathbf{A}^{1/4}$ is $f(\mathbf{A})$ for $f(t) = t^{1/4}$ on $[0, \infty)$.

(b) Deduce that $t \mapsto t^{1/2^k}$ is operator monotone on $[0, \infty)$ for every integer $k \geq 0$.

(c) Explain why the argument of (b) cannot reach $t \mapsto t^{3/4}$, and say where the operator monotonicity of t^p for every $p \in (0, 1)$ is proved.

Hint for (a): the positive square root is unique.

Solution.

(a) Let $g(t) = t^{1/4}$ on $[0, \infty)$. Since $\mathbf{A} \succeq \mathbf{O}$, its spectrum lies in $[0, \infty)$ (Theorem 13.1.4 (b)), and $g(\mathbf{A}) = \sum_j \mu_j^{1/4} \mathbf{P}_j$ is Hermitian with nonnegative eigenvalues by Theorem 12.8.3 (c) and (d), so $g(\mathbf{A}) \succeq \mathbf{O}$. By Theorem 12.8.3 (b), $g(\mathbf{A})^2 = (g^2)(\mathbf{A})$, and $g^2(t) = t^{1/2}$, so $g(\mathbf{A})^2 = \mathbf{A}^{1/2}$. By the uniqueness in Theorem 13.1.4 (d), applied to $\mathbf{A}^{1/2} \succeq \mathbf{O}$, $g(\mathbf{A})$ is the positive square root of $\mathbf{A}^{1/2}$: $\mathbf{A}^{1/4} = (\mathbf{A}^{1/2})^{1/2}$. The same holds for $\mathbf{B}$. Now Proposition 13.5.4 applied to $\mathbf{A} \succeq \mathbf{B} \succeq \mathbf{O}$ gives $\mathbf{A}^{1/2} \succeq \mathbf{B}^{1/2}$, and $\mathbf{B}^{1/2} \succeq \mathbf{O}$ by Theorem 13.2.1. Applying Proposition 13.5.4 again, to $\mathbf{A}^{1/2} \succeq \mathbf{B}^{1/2} \succeq \mathbf{O}$, gives $(\mathbf{A}^{1/2})^{1/2} \succeq (\mathbf{B}^{1/2})^{1/2}$, that is $\mathbf{A}^{1/4} \succeq \mathbf{B}^{1/4}$.

(b) Induction on k. For $k = 0$ the function is $t \mapsto t$, operator monotone by Proposition 17.11.2. Suppose $\mathbf{A}^{1/2^k} \succeq \mathbf{B}^{1/2^k}$ whenever $\mathbf{A} \succeq \mathbf{B} \succeq \mathbf{O}$. The argument of (a), with $t^{1/2^{k+1}}$ in place of $t^{1/4}$ and $t^{1/2^k}$ in place of $t^{1/2}$, shows that $\mathbf{A}^{1/2^{k+1}} = (\mathbf{A}^{1/2^k})^{1/2}$, and similarly for $\mathbf{B}$. Both $\mathbf{A}^{1/2^k}$ and $\mathbf{B}^{1/2^k}$ are $\succeq \mathbf{O}$, so Proposition 13.5.4 applied to $\mathbf{A}^{1/2^k} \succeq \mathbf{B}^{1/2^k} \succeq \mathbf{O}$ gives $\mathbf{A}^{1/2^{k+1}} \succeq \mathbf{B}^{1/2^{k+1}}$.

(c) Each step of (b) takes a square root, which halves the exponent. Starting from 1, the exponents reached are exactly $1, \frac{1}{2}, \frac{1}{4}, \ldots$, and $\frac{3}{4}$ is not among them. The operator monotonicity of t^p for every $p \in (0, 1)$ needs a new idea, an integral representation of the kind in Theorem 17.11.9, and it is proved in Chapter 21.

Exercise. C2: The reciprocal is operator convex

(a) Let $\mathbf{G} \in M_n(F)$ with $\mathbf{G} \succ \mathbf{O}$. Prove that the $2n \times 2n$ matrix $\begin{pmatrix} \mathbf{G} & \mathbf{I}_n \\ \mathbf{I}_n & \mathbf{G}^{-1} \end{pmatrix}$ is Hermitian and $\succeq \mathbf{O}$.

(b) Let $\mathbf{A}, \mathbf{B} \succ \mathbf{O}$ be of the same size and $\theta \in [0,1]$. Prove that

$$(\theta\mathbf{A} + (1-\theta)\mathbf{B})^{-1} \preceq \theta\mathbf{A}^{-1} + (1-\theta)\mathbf{B}^{-1},$$

that is, $t \mapsto 1/t$ is operator convex on $(0,\infty)$.

(c) Deduce that $t \mapsto -1/t$ is both operator monotone and operator concave on $(0,\infty)$.

Hint: Schur complements.

Solution.

(a) $\mathbf{G}$ is Hermitian, and so is $\mathbf{G}^{-1}$, since $(\mathbf{G}^{-1})^* = (\mathbf{G}^*)^{-1} = \mathbf{G}^{-1}$. The off-diagonal blocks are $\mathbf{I}_n$ and $\mathbf{I}_n^* = \mathbf{I}_n$, so the block matrix $\mathbf{M}_\mathbf{G}$ is Hermitian. Its corner $\mathbf{G}$ is $\succ \mathbf{O}$, and its Schur complement is

$$\mathbf{M}_\mathbf{G}/\mathbf{G} = \mathbf{G}^{-1} - \mathbf{I}_n^*\mathbf{G}^{-1}\mathbf{I}_n = \mathbf{O} \succeq \mathbf{O}.$$

By Theorem 13.6.1 (b), $\mathbf{M}_\mathbf{G} \succeq \mathbf{O}$.

(b) Put $\mathbf{C} = \theta\mathbf{A} + (1-\theta)\mathbf{B}$. For $\mathbf{x} \neq \mathbf{0}$, $\langle \mathbf{C}\mathbf{x}, \mathbf{x}\rangle = \theta\langle \mathbf{A}\mathbf{x}, \mathbf{x}\rangle + (1-\theta)\langle \mathbf{B}\mathbf{x}, \mathbf{x}\rangle$. Both inner products are > 0, and θ and $1-\theta$ are ≥ 0 and not both zero, so the sum is > 0. Thus $\mathbf{C} \succ \mathbf{O}$. By (a), $\mathbf{M}_\mathbf{A} \succeq \mathbf{O}$ and $\mathbf{M}_\mathbf{B} \succeq \mathbf{O}$, so for every $\mathbf{y} \in F^{2n}$ the quadratic form of

$$\theta\mathbf{M}_\mathbf{A} + (1-\theta)\mathbf{M}_\mathbf{B} = \begin{pmatrix} \mathbf{C} & \mathbf{I}_n \\ \mathbf{I}_n & \theta\mathbf{A}^{-1} + (1-\theta)\mathbf{B}^{-1} \end{pmatrix}$$

is $\theta\langle \mathbf{M}_\mathbf{A}\mathbf{y}, \mathbf{y}\rangle + (1-\theta)\langle \mathbf{M}_\mathbf{B}\mathbf{y}, \mathbf{y}\rangle \geq 0$, and this block matrix is $\succeq \mathbf{O}$. Its corner $\mathbf{C}$ is $\succ \mathbf{O}$, so Theorem 13.6.1 (b) gives

$$\theta\mathbf{A}^{-1} + (1-\theta)\mathbf{B}^{-1} - \mathbf{C}^{-1} \succeq \mathbf{O},$$

which is the claim. Since $t \mapsto 1/t$ sends a positive definite matrix to its inverse (Definition 12.8.2), and a Hermitian matrix has spectrum in $(0,\infty)$ exactly when it is $\succ \mathbf{O}$ (Theorem 13.1.5 (b)), this is Definition 17.11.1 (b) for $f(t) = 1/t$.

(c) Operator monotone by Proposition 17.11.3. Operator concave by definition, because its negative $t \mapsto 1/t$ is operator convex by (b).

Exercise. C3: The cube needs its own witness

(a) With $\mathbf{A}$ and $\mathbf{B}$ as in B2, show that $\mathbf{A}^3 \not\succeq \mathbf{B}^3$. Conclude that $t \mapsto t^3$ is not operator monotone on $[0,\infty)$.

(b) One might hope to deduce (a) from the failure of $t \mapsto t^2$. Suppose $t \mapsto t^3$ were operator monotone on $[0,\infty)$. Applying it and $t \mapsto t^{1/2}$ (Proposition 13.5.4) to both sides of $\mathbf{A} \succeq \mathbf{B} \succeq \mathbf{O}$, repeatedly and in any order, yields inequalities $\mathbf{A}^q \succeq \mathbf{B}^q$ for certain exponents q. Show that every such q has the form $3^j/2^k$ with integers $j, k \geq 0$, and that 2 never has this form. Explain why this means (a) needed a witness of its own.

Solution.

(a) From B2, $\mathbf{A} \succeq \mathbf{B} \succeq \mathbf{O}$, with spectra in $(0,\infty)$. Multiplying the squares of B2 once

more,

$$\mathbf{A}^3 = \begin{pmatrix} 34 & 14 \\ 14 & 6 \end{pmatrix}, \qquad \mathbf{B}^3 = \begin{pmatrix} 13 & 8 \\ 8 & 5 \end{pmatrix}, \qquad \mathbf{A}^3 - \mathbf{B}^3 = \begin{pmatrix} 21 & 6 \\ 6 & 1 \end{pmatrix}.$$

The determinant is $21 - 36 = -15 < 0$, so the difference is indefinite. Explicitly, with $\mathbf{x} = (1, -5)$,

$$\mathbf{x}^\top (\mathbf{A}^3 - \mathbf{B}^3)\mathbf{x} = 21 + 2 \cdot 6 \cdot (-5) + 25 = -14 < 0.$$

By Theorem 12.8.3 (a), $t \mapsto t^3$ sends $\mathbf{A}$ to $\mathbf{A}^3$, so it is not operator monotone on $[0, \infty)$.

(b) Start from $\mathbf{A} = \mathbf{A}^{3^0/2^0}$. Taking the square root of $\mathbf{A}^q$ gives $\mathbf{A}^{q/2}$, and cubing it gives $\mathbf{A}^{3q}$, by Theorem 12.8.3 (b) and the uniqueness of the positive square root, as in C1 (a); likewise for $\mathbf{B}$. If $q = 3^j/2^k$, the results are $3^j/2^{k+1}$ and $3^{j+1}/2^k$, of the same form. By induction on the number of steps, every exponent reached has the form $3^j/2^k$. If $2 = 3^j/2^k$, then $3^j = 2^{k+1}$, and the left side is odd while the right side is even, since $k + 1 \geq 1$. So no exponent reached equals 2. The hypothetical operator monotonicity of the cube therefore cannot be converted, by these means, into that of the square. The failure of the square does not formally imply the failure of the cube, and (a) had to find a pair of matrices of its own.

Chapter 18

Convexity

Three chapters in a row have met convexity and moved on. Chapter 16 §01 drew the unit balls of three norms and remarked that each is convex, because the triangle inequality forces $\|tu + (1-t)v\| \leq 1$ whenever $\|u\|, \|v\| \leq 1$ (Definition 16.1.2). Chapter 13 §01 set as an exercise that the positive semidefinite matrices are closed under sums and nonnegative multiples (Exercise (C2: The cone of positive matrices)). Chapter 17 §06 proved that the sum of the k largest eigenvalues is subadditive, and so a convex function of the matrix (Corollary 17.6.5), and said that this was only to be expected of a maximum of linear functions. And Chapter 17 §08 showed, again in an exercise, that the possible diagonals of the Hermitian matrices with a given spectrum contain the segment between any two of their points (Exercise (C3: The possible diagonals form a convex set)). Each time, convexity was checked on the spot and not studied. One debt is older still: Chapter 16's plan listed **dual norms**, and Chapter 16 never defined one.

This chapter studies convexity for its own sake. A set is convex when it contains the segment between any two of its points, and the theory built on that single condition has two faces. One face is **combinatorial**. Every point of a convex hull in dimension n is an average of at most $n+1$ of the original points, and from that count follow the theorems of Radon and Helly on how convex sets intersect. The other face is **geometric**, and it carries most of the weight. A point outside a closed convex set can be separated from it by a hyperplane, and the hyperplane comes from the nearest point of the set, generalizing the orthogonal projection of Chapter 11. From separation come the rest: a convex set described by the half-spaces that contain it, the dual norm as the support function of the unit ball, Farkas's lemma and the duality of linear programming, the minimax theorem for matrix games, and the fact that a compact convex set is the hull of its extreme points. The chapter ends with convex functions, which is where the Ky Fan observation of Chapter 17 belongs.

The field is $\mathbb{R}$ throughout. Convexity asks for weights t with $0 \leq t \leq 1$, and "between 0 and 1" needs an order on the scalars, which $\mathbb{C}$ does not have. A complex space is treated as a real one when it is needed, by allowing only real scalars: the Hermitian $n \times n$ matrices, for instance, form a real vector space and not a complex one, and that is how positive semidefinite matrices and eigenvalue functions enter. Spaces are finite-dimensional, and each section says whether a result needs only the vector space, a norm, or an inner product. Most of the chapter works in $\mathbb{R}^n$ with the standard inner product.

18.0.1 The analysis this chapter imports

The list is Chapter 16's, facts (A1)–(A6) of that chapter's introduction, cited by the same name at the point of use: "by the extreme value theorem, fact (A4) of Chapter 16's introduction". This chapter needs chiefly two of them. **(A3)**, the compactness of closed bounded sets, and

(A4), the extreme value theorem, are what make a nearest point, a maximum over a hull, or an extreme point exist rather than merely be approached. §02 uses (A3) to prove that the hull of a compact set is compact, and the separation and extreme-point theorems of §§03 and 08 rest on both. **(A1)**, completeness, gives the suprema that define gauges and support functions. For the second-derivative test of convexity in §10, the one-variable case uses **(A6)**, the mean value theorem, and the several-variable case uses the symmetry and Taylor assumptions that Chapter 14 §06 recorded for its own second-derivative test. §11 builds the logarithm from the exponential of Chapter 10 §09 (that section's facts (A1)–(A3)) and needs **(A5)**, the intermediate value theorem, to show every positive number has one.

The words **closed** and **open** are defined in §01, in the sequential form in which Chapter 16's introduction defines continuity, and a closedness argument uses nothing but that definition and the elementary algebra of limits. A section that needs any analytic fact outside Chapter 16's list says so where it uses it.

18.0.2 What you need

- **Chapter 0 §06 and §09** — the triangle inequality for complex numbers, which §11 uses twice, and the trace, which §10 uses once: Theorem 0.6.4, Theorem 0.9.6.
- **Chapter 1** — independence and dimension, used to count points: Definition 1.4.1, Theorem 1.6.3, Theorem 1.3.1, Theorem 1.6.5, and the dimension formula Theorem 1.7.3. Zorn's lemma (Theorem 1.8.1) is mentioned once, to mark where the infinite-dimensional Hahn–Banach theorem would begin.
- **Chapter 2** — Rank–Nullity: Theorem 2.3.1.
- **Chapter 4** — cosets, which are the affine subspaces of this chapter: Definition 4.4.2, Lemma 4.4.3.
- **Chapter 3 §04 and §05** — the Invertible Matrix Theorem and the rank facts §09 counts the vertices of a polyhedron with: Theorem 3.4.4, Theorem 3.5.7, Theorem 3.5.5.
- **Chapter 5** — linear functionals and hyperplanes, the raw material of every separation argument: Definition 5.1.1, Theorem 5.1.1, Definition 5.5.2, Theorem 5.5.3, Proposition 5.6.4. This chapter is named for its language and not as a step in an argument: §§01, 03, 04, 05, and 09 cite it in the text, to say which abstract object a half-space or an affine hyperplane is, and two examples use it, but no proof here rests on it. Once $V = \mathbb{R}^n$, what is worked with is the concrete $\mathbf{x} \mapsto \mathbf{a}^\top \mathbf{x}$.
- **Chapter 10 §09** — the exponential of a number, with that section's imported facts (A1)–(A3), from which §11 builds the logarithm and real powers: Theorem 10.9.4. The reference sits in the text of §11 that sets the construction up, not inside a proof.
- **Chapter 11** — the inner product and the orthogonal projection, which §03 generalizes from subspaces to closed convex sets: Definition 11.1.1, Theorem 11.1.5, Theorem 11.3.2, Definition 11.3.2, Theorem 11.3.3, Theorem 11.3.5.
- **Chapter 12 §05** — the real spectral theorem, which is how §08 finds the extreme points of the density matrices and of the interval $\mathbf{O} \preceq \mathbf{X} \preceq \mathbf{I}$: Corollary 12.5.2.
- **Chapter 13** — positivity: Definition 13.1.1, Theorem 13.1.4, Definition 13.5.1.
- **Chapter 14 §06** — the Hessian and the second-derivative test, Definition 14.6.2 and Theorem 14.6.4, with that section's Taylor assumption. As with Chapter 10, §10 names these in the text, to say what it is generalizing and which imported facts it borrows; the proofs cite the assumptions, not the theorem.
- **Chapter 16** — norms, their unit balls and their equivalence, compactness, and the operator norm: Definition 16.1.1, Definition 16.1.2, Proposition 16.1.2, Lemma 16.1.1, Theorem 16.2.2, Corollary 16.2.3, Corollary 16.2.5, Definition 16.3.2, Theorem 16.3.3.
- **Chapter 17** — the places where convexity was met: Theorem 17.6.2 and Corollary 17.6.5, Definition 17.8.1 and Theorem 17.8.5, Lemma 17.8.6, Definition 17.11.1 and Proposi-

tion 17.11.8.

18.0.3 Roadmap

- **Convex sets and hulls.** The definition, the first examples (half-spaces, the balls of every norm, the standard simplex, the positive semidefinite matrices, the vectors majorized by a fixed vector), the operations that preserve convexity, and the convex hull as the set of all convex combinations. Affine hulls, which are Chapter 4's cosets. Closed and open sets.
- **Carathéodory's theorem.** In dimension n, every point of a hull is a convex combination of at most $n+1$ of the points, and no fewer will do in general. Hence the hull of a compact set is compact. Radon's theorem, and Helly's theorem on intersections.
- **Separating hyperplanes.** The nearest point of a closed convex set, characterized by an obtuse-angle condition that becomes Chapter 11's orthogonality when the set is a subspace. Strict separation of a point from a closed convex set, supporting hyperplanes, and the finite-dimensional Hahn–Banach theorem. Why two disjoint closed convex sets need not be strictly separated.
- **Gauges, support functions, dual norms.** Every bounded, closed, symmetric convex set that contains a ball around $\mathbf{0}$ is the unit ball of a norm, which is the converse of Chapter 16's convexity remark. The support function, the polar set, and the dual norm, with the proof that the dual of the dual norm is the original norm. This supplies the dual norms that Chapter 16 left out.
- **Cones and Farkas's lemma.** Convex cones and their duals; a finitely generated cone is closed, which is the subtle step; the alternative theorem of Farkas; and a closed cone with a nonclosed linear image, showing why finite generation matters.
- **Linear programming.** Weak and strong duality from Farkas's lemma, complementary slackness, and a worked problem solved together with its dual. This answers the remark in Chapter 3 §07 that flows with capacities are a problem about inequalities.
- **The minimax theorem.** Matrix games, mixed strategies over the simplex, and von Neumann's theorem, contrasted with the Courant–Fischer min–max of Chapter 17, whose proof is a dimension count rather than convexity.
- **Extreme points.** A compact convex set is the convex hull of its extreme points (Minkowski's theorem). The extreme points of the unit balls of the three norms, of the density matrices, and of the doubly stochastic matrices; for the last, Birkhoff's theorem that the permutation matrices are all of them is stated and left to Chapter 19. A compact convex set whose extreme points do not form a closed set.
- **Polytopes.** Hulls of finitely many points against bounded intersections of finitely many half-spaces, and the Minkowski–Weyl theorem that they are the same sets. Vertices as the basic feasible points of linear programming.
- **Convex functions.** The epigraph, Jensen's inequality, the second-derivative test, and continuity. A pointwise supremum of affine functions is convex, which accounts for Chapter 17's Ky Fan sums.
- **Three classical inequalities.** The logarithm and real powers, built from Chapter 10's exponential; the weighted arithmetic–geometric mean inequality as Jensen's inequality for $-\log$; Young's and Hölder's inequalities; Minkowski's inequality, so that every $\|\cdot\|_p$ is a norm on F^n; and the dual of every p-norm, which completes §04. The chapter's summary closes this section.

18.0.4 Named moves

- **Count, then drop a point.** Too many points in too few dimensions are dependent, affinely or linearly. Move the weights along the dependence until one of them reaches

zero, and drop that point without leaving the hull. Carathéodory's theorem, its version for cones, and Radon's theorem are all this move.

- **Nearest point, then the normal direction.** To separate a point from a closed convex set, take the nearest point of the set and use the vector from it to the outside point as the normal of the hyperplane. The separation theorems are the geometry of this one projection.
- **Dualize.** A closed convex set is determined by the half-spaces containing it, and a closed convex cone by its dual cone. Every statement about the one side has a statement about the other: Farkas's lemma is that move made into a theorem, and linear-programming duality is Farkas applied to one combined system.
- **A supremum of linear functions is convex.** Support functions, dual norms, the largest eigenvalue, and the Ky Fan sums are all suprema of linear functions, and each inherits its convexity from that alone.
- **Lift, then eliminate.** To see what a set looks like from fewer coordinates, eliminate one variable at a time: Fourier–Motzkin elimination pairs each upper bound with each lower bound, and is what proves the Minkowski–Weyl theorem of §09.
- **Compactness turns a supremum into a maximum.** Chapter 16's move, used again: once a hull or a set of weights is known to be compact, a maximum or a nearest point is attained at a point that can be named and fed back into the argument.

18.1 Convex Sets and Hulls

Convexity has turned up in this book four times without being asked for: in the unit balls of Chapter 16, the positive semidefinite matrices of Chapter 13, and the Ky Fan sums and the possible diagonals of Chapter 17 (the chapter introduction lists the places). Each time the property was checked by hand and set aside. This section gives it a name, finds the smallest convex set containing a given set, and fixes the words **closed** and **open** that the rest of the chapter needs.

The field is $\mathbb{R}$, throughout the chapter. The definition below asks for a scalar t with $0 \le t \le 1$, which needs an order on the scalars; $\mathbb{C}$ has none. A complex space is treated as a real one by allowing only real scalars. That is how the Hermitian matrices enter below: they form a vector space over $\mathbb{R}$, and not over $\mathbb{C}$, since $i\mathbf{A}$ is not Hermitian when $\mathbf{A} \neq \mathbf{O}$ is. Unless a statement says otherwise, V is a real vector space; where a result needs a finite dimension, a norm, or an inner product, it says so.

18.1.1 Segments, and sets that contain them

Chapter 16 §01 checked the convexity of a unit ball with one line: if $\|u\|, \|v\| \le 1$ and $0 \le t \le 1$, then $\|tu + (1-t)v\| \le 1$. The expression $tx + (1-t)y$, with t running over $[0, 1]$, keeps recurring, so it deserves a name, and so does the property of containing all of it.

Picture it first. At $t = 1$ the expression is x, at $t = 0$ it is y, and at $t = \frac{1}{2}$ it is the midpoint. Writing it as $y + t(x - y)$ shows the rest: we start at y and travel the fraction t of the way to x. As t runs over $[0, 1]$, the point runs along the straight **segment** from y to x, and nowhere else.

A convex set is one that contains the whole straight segment between any two of its points.

Definition 18.1.1. Convex Set

Let V be a vector space **over** $\mathbb{R}$. A subset $C \subseteq V$ is **convex** if, **for all** $x, y \in C$ and **every**

real t with $0 \le t \le 1$,

$$tx + (1-t)y \in C.$$

The set $\{tx + (1-t)y : 0 \le t \le 1\}$ is the **segment** joining x and y, written $[x, y]$.

In words: pick any two points of C, possibly equal; the segment joining them must lie inside C. The quantifier "for all x, y" is what makes this a property of the set rather than of a pair; one segment escaping C is enough to make it nonconvex. The restriction $0 \le t \le 1$ is the other half of the content, and the end of this subsection shows what happens without it. There is nothing to check for well-definedness: a subset either has the property or it does not.

Example. The first convex sets

Check each of the following against Definition 18.1.1.

(a) The empty set, and a single point $\{p\}$.

(b) A subspace U of V, and a coset $p + U$ of it (Definition 4.4.2).

(c) A **closed half-space** $\{x \in V : \varphi(x) \le c\}$, where φ is a nonzero linear functional on V (Definition 5.1.1) and $c \in \mathbb{R}$.

(d) A closed ball $\{x \in V : \|x - p\| \le r\}$ of any norm on V, with center p and radius $r \ge 0$.

(e) The set of positive semidefinite matrices (Definition 13.1.1) in the real vector space of $n \times n$ Hermitian matrices.

Solution.

(a) The empty set has no two points, so the condition holds vacuously; for $\{p\}$, $tp + (1-t)p = p$. These degenerate cases matter because an intersection of convex sets, which we meet shortly, is often empty or a single point, and the statements below must allow for both.

(b) Let $x = p + u$ and $y = p + w$ with $u, w \in U$. Then

$$tx + (1-t)y = p + \big(tu + (1-t)w\big),$$

and $tu + (1-t)w \in U$ because a subspace is closed under linear combinations. Taking $p = 0$ gives the subspace itself. Here the restriction on t was not used at all; we return to this below.

(c) Let $\varphi(x) \le c$ and $\varphi(y) \le c$. By linearity,

$$\varphi\big(tx + (1-t)y\big) = t\varphi(x) + (1-t)\varphi(y) \le tc + (1-t)c = c,$$

where the inequality multiplies the first hypothesis by t and the second by $1 - t$. Multiplying an inequality by a number preserves it only when the number is ≥ 0, so both $t \ge 0$ and $1 - t \ge 0$ were used. In $\mathbb{R}^n$ every linear functional is $\mathbf{x} \mapsto \mathbf{a}^{\top}\mathbf{x}$ for some $\mathbf{a}$ (Theorem 5.1.1), so the closed half-spaces of $\mathbb{R}^n$ are the sets $\{\mathbf{x} : \mathbf{a}^{\top}\mathbf{x} \le c\}$ with $\mathbf{a} \neq \mathbf{0}$. The same computation shows that the **open half-space** $\{x : \varphi(x) < c\}$ is convex, and that the **affine hyperplane** $\{x : \varphi(x) = c\}$ is convex.

(d) Let $\|x - p\| \le r$ and $\|y - p\| \le r$. Since $tx + (1-t)y - p = t(x - p) + (1-t)(y - p)$, (N3) and (N2) of Definition 16.1.1 give

$$\|tx + (1-t)y - p\| \le t\|x - p\| + (1-t)\|y - p\| \le r,$$

where homogeneity used $|t| = t$ and $|1-t| = 1-t$. With $p = 0$ and $r = 1$ this is the closed unit ball of Definition 16.1.2, and the computation is the one Chapter 16 §01 made there.

(e) The Hermitian matrices form a real vector space: a real linear combination of Hermitian matrices is Hermitian. Let $\mathbf{A} \succeq \mathbf{O}$ and $\mathbf{B} \succeq \mathbf{O}$. Then $t\mathbf{A} + (1-t)\mathbf{B}$ is Hermitian, and for every $\mathbf{x} \in \mathbb{C}^n$,

$$\langle (t\mathbf{A} + (1-t)\mathbf{B})\mathbf{x}, \mathbf{x}\rangle = t\langle \mathbf{A}\mathbf{x}, \mathbf{x}\rangle + (1-t)\langle \mathbf{B}\mathbf{x}, \mathbf{x}\rangle \geq 0,$$

since both inner products on the right are ≥ 0 by (P2) and both coefficients are ≥ 0. So (P1) and (P2) hold, and $t\mathbf{A} + (1-t)\mathbf{B} \succeq \mathbf{O}$. The same computation with any $s, t \geq 0$ in place of t and $1-t$ shows $s\mathbf{A} + t\mathbf{B} \succeq \mathbf{O}$, so the positive semidefinite matrices form a convex cone, in the sense of §05. (This is what Exercise (C2: The cone of positive matrices) (a) asks.)

The standard example of a convex set in coordinates gets its own name, because the rest of this section keeps returning to it.

Definition 18.1.2. Standard Simplex

For $n \geq 1$, the **standard simplex** in $\mathbb{R}^n$ is

$$\Delta_n := \Big\{ \mathbf{t} = (t_1, \ldots, t_n) \in \mathbb{R}^n : t_i \geq 0 \text{ for every } i, \quad t_1 + \cdots + t_n = 1 \Big\}.$$

$\Delta_1 = \{1\}$ is a point, Δ_2 is the segment from $(1,0)$ to $(0,1)$, and Δ_3 is the triangle with corners $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$. It is convex: it is the intersection of the n closed half-spaces $\{t_i \geq 0\}$ with the affine hyperplane $\{t_1 + \cdots + t_n = 1\}$, and each of these is convex by Example (The first convex sets) (c). (That an intersection of convex sets is convex is the first item of Proposition 18.1.1 below; it is also immediate here, because a point that lies in each set lies in their intersection.) An element of Δ_n is a list of n **weights**: nonnegative, and adding up to 1.

convex not convex: the dashed segment leaves the set

A nonexample by minimal change. Keep the closed unit ball B_2 of $\mathbb{R}^2$ but take only its edge, the unit circle $\{\mathbf{x} : \|\mathbf{x}\|_2 = 1\}$. Much survives: the circle is symmetric under $\mathbf{x} \mapsto -\mathbf{x}$, bounded, and as round as the disc. But the clause "for all $\mathbf{x}, \mathbf{y}$" fails at $\mathbf{x} = \mathbf{e}_1$, $\mathbf{y} = -\mathbf{e}_1$, $t = \frac{1}{2}$: the midpoint is $\mathbf{0}$, and $\|\mathbf{0}\|_2 = 0 \neq 1$. What was lost is exactly the inequality ≤ 1 that absorbed the triangle inequality in Example (The first convex sets) (d); with an equality there is nothing to absorb it.

Why this definition. Two variations are worth trying.

- *Let t range over all of $\mathbb{R}$.* Then C must contain the whole **line** through any two of its points. The subspaces and cosets of Example (The first convex sets) meet this demand, which is why (b) never used $0 \leq t \leq 1$; the half-spaces and balls do not. Exercise (C3: Sets that contain lines) shows that, apart from $\emptyset$, the sets meeting it are exactly the cosets.
- *Ask only for $t = \frac{1}{2}$.* A set containing the midpoint of any two of its points is **midpoint**

convex, which is genuinely weaker: $\mathbb{Q}^2$ is midpoint convex, yet it misses $\frac{1}{\sqrt{2}}\mathbf{e}_1 \in [\mathbf{0}, \mathbf{e}_1]$. Exercise (C1: Midpoint convexity) shows that the gap disappears for closed sets, defined at the end of this section.

The name is the everyday word: a convex lens has no dents.

Warning. **A union of convex sets need not be convex.** The two coordinate axes in $\mathbb{R}^2$ are subspaces, hence convex, but their union contains $\mathbf{e}_1$ and $\mathbf{e}_2$ and not their midpoint $(\frac{1}{2}, \frac{1}{2})$, which lies on neither axis. Even two points fail: $\{\mathbf{0}\} \cup \{\mathbf{e}_1\}$ is a union of two convex sets and misses $\frac{1}{2}\mathbf{e}_1$. The operation that does respect convexity is intersection (Proposition 18.1.1), and the convex replacement for a union is the hull of the union, defined below.

Quick check. Is $\{(x, y) \in \mathbb{R}^2 : y \geq |x|\}$ convex? Is $\{(x, y) \in \mathbb{R}^2 : y \leq |x|\}$ convex?

Solution. The first is convex. It is the set where both $y - x \geq 0$ and $y + x \geq 0$, an intersection of two closed half-spaces, and a point on a segment between two points of both half-spaces lies in both by Example (The first convex sets) (c). The second is not: $(1, 1)$ and $(-1, 1)$ belong to it, since $1 \leq 1$, but their midpoint $(0, 1)$ does not, since $1 \leq 0$ is false. The two sets differ only in the direction of the inequality, and the direction decides everything; a convex function such as $|x|$ has a convex region **above** its graph (§10).

Here is a payoff. Chapter 17 §08 met a set of vectors that it could describe exactly, but whose convexity it proved only in an exercise. Everything needed to prove it is now in place.

Example. The vectors majorized by a fixed vector

Let $\boldsymbol{\lambda} \in \mathbb{R}^n$. Prove that the set $\{\mathbf{d} \in \mathbb{R}^n : \mathbf{d} \prec \boldsymbol{\lambda}\}$ of vectors majorized by $\boldsymbol{\lambda}$ (Definition 17.8.1) is convex, and deduce that the diagonals of the Hermitian matrices with spectrum $\boldsymbol{\lambda}$ form a convex set.

Solution. For $\mathbf{x} \in \mathbb{R}^n$ and $1 \leq k \leq n$, write $s_k(\mathbf{x}) = x_1^{\downarrow} + \cdots + x_k^{\downarrow}$, the sum of the k largest entries.

Step 1: any k entries add up to at most $s_k(\mathbf{x})$. List the entries x_i at a set I of k positions decreasingly as $v_1 \geq \cdots \geq v_k$. Since $\mathbf{x}$ has j entries $v_1, \ldots, v_j \geq v_j$, we get $x_j^{\downarrow} \geq v_j$ for each j. Summing gives $\sum_{i \in I} x_i \leq s_k(\mathbf{x})$, with equality when I holds the positions of the k largest entries.

Step 2: s_k respects averages. Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ and $0 \leq t \leq 1$, and let I hold the positions of the k largest entries of $\mathbf{z} = t\mathbf{x} + (1 - t)\mathbf{y}$. By Step 1, used once with equality and twice as an inequality,

$$s_k(\mathbf{z}) = t \sum_{i \in I} x_i + (1 - t) \sum_{i \in I} y_i \leq t\, s_k(\mathbf{x}) + (1 - t)\, s_k(\mathbf{y}),$$

where the inequality again needs $t \geq 0$ and $1 - t \geq 0$.

Step 3: convexity. Let $\mathbf{d} \prec \boldsymbol{\lambda}$ and $\mathbf{d}' \prec \boldsymbol{\lambda}$, and put $\mathbf{z} = t\mathbf{d} + (1 - t)\mathbf{d}'$. For $k \leq n - 1$, Step

2 and (M1) for both vectors give $s_k(\mathbf{z}) \le t\, s_k(\boldsymbol{\lambda}) + (1-t)\, s_k(\boldsymbol{\lambda}) = s_k(\boldsymbol{\lambda})$. The total of $\mathbf{z}$ is $t\sum_i \lambda_i + (1-t)\sum_i \lambda_i = \sum_i \lambda_i$ by (M2) for both. So $\mathbf{z} \prec \boldsymbol{\lambda}$, and the set is convex.

For the deduction, the Schur–Horn Theorem (Theorem 17.8.5) says that the diagonals of the Hermitian matrices with eigenvalues $\boldsymbol{\lambda}$ are **exactly** the vectors $\mathbf{d} \prec \boldsymbol{\lambda}$. So the set of those diagonals is the set just shown to be convex.

This is the statement of Exercise (C3: The possible diagonals form a convex set), now with a proof in the text, and Step 2 is a first glimpse of §10's convex functions: s_k satisfies $s_k(t\mathbf{x}+(1-t)\mathbf{y}) \le t\, s_k(\mathbf{x}) + (1-t)\, s_k(\mathbf{y})$ because it is a maximum of linear functions.

18.1.2 Operations that preserve convexity

Checking the definition directly is rarely the fastest route. Most convex sets are built from simpler ones by a few operations, and the following proposition lists those that preserve convexity. It needs one construction. For subsets $A, B \subseteq V$ and a real c, the **Minkowski sum** and the **scalar multiple** are

$$A + B \coloneqq \{a + b : a \in A,\ b \in B\}, \qquad cA \coloneqq \{ca : a \in A\}.$$

A **translate** $p + A$ is the Minkowski sum $\{p\} + A$; a coset (Definition 4.4.2) is a translate of a subspace.

Proposition 18.1.1. Operations Preserving Convexity

Let V and W be real vector spaces.

(a) The intersection of **any** family of convex subsets of V is convex.

(b) If $T\colon V \to W$ is linear, $C \subseteq V$ is convex and $D \subseteq W$ is convex, then the image $T(C)$ and the preimage $T^{-1}(D) = \{v \in V : Tv \in D\}$ are convex.

(c) If $A, B \subseteq V$ are convex and $c \in \mathbb{R}$, then $A + B$, cA, and every translate $p + A$ are convex.

Idea. Rewrite the point of the segment so that convexity of the given sets applies: linearity moves the weights through T, and a Minkowski sum regroups term by term.

Proof. Throughout, let $0 \le t \le 1$.

(a) Let (C_α) be a family of convex sets and $x, y \in \bigcap_\alpha C_\alpha$. For each α, both points lie in C_α, so $tx + (1-t)y \in C_\alpha$ by convexity of C_α. Hence the point lies in every C_α, that is, in the intersection. (The intersection of the empty family is V by convention, and V is convex.)

(b) Let $Tx, Ty \in T(C)$ with $x, y \in C$. By linearity, $tTx + (1-t)Ty = T\big(tx + (1-t)y\big)$, and $tx + (1-t)y \in C$ by convexity; so the point lies in $T(C)$. For the preimage, let $Tx, Ty \in D$. Then $T\big(tx + (1-t)y\big) = tTx + (1-t)Ty \in D$ by convexity of D, so $tx + (1-t)y \in T^{-1}(D)$.

(c) Let $a + b$ and $a' + b'$ lie in $A + B$. Rearranging,

$$t(a+b) + (1-t)(a'+b') = \big(ta + (1-t)a'\big) + \big(tb + (1-t)b'\big),$$

and the two brackets lie in A and in B by convexity. So the point lies in $A + B$. The scalar multiple cA is the image of A under the linear map $v \mapsto cv$, and is convex by (b). A translate $p + A = \{p\} + A$ is convex because a single point is convex (Example (The first convex sets) (a)). This proves the proposition.

■

The proposition turns many checks into one line. The solution set of a system of linear inequalities,

$$\{\mathbf{x} \in \mathbb{R}^n : \mathbf{A}\mathbf{x} \le \mathbf{b}\} = \bigcap_{i=1}^{m} \{\mathbf{x} : \mathbf{a}_i^{\top}\mathbf{x} \le b_i\},$$

where $\mathbf{a}_i^{\top}$ is the i-th row of $\mathbf{A} \in M_{m\times n}(\mathbb{R})$ and $\mathbf{A}\mathbf{x} \le \mathbf{b}$ means $\le$ in every entry, is an intersection of closed half-spaces (a zero row gives $\mathbb{R}^n$ or $\emptyset$), hence convex by (a). This is the kind of set that the closing remark of Chapter 3 §07 points to, where nonnegativity and capacities are added to a flow problem, and §06 of this chapter optimizes over it. The set of **density matrices** $\{\mathbf{A} : \mathbf{A} \succeq \mathbf{O},\ \operatorname{tr}\mathbf{A} = 1\}$ is convex for the same reason: it is the intersection of the positive semidefinite set of Example (The first convex sets) (e) with the preimage of $\{1\}$ under the linear map tr.

18.1.3 Convex combinations and the convex hull

A convex set contains the segment between any two of its points. Apply that twice: it contains x_1, x_2, and hence every point of $[x_1, x_2]$, and then every point on a segment from such a point to x_3. The points reached are the averages of x_1, x_2, x_3 with nonnegative weights, and the pattern deserves a name.

Definition 18.1.3. Convex Combination

Let V be a real vector space and $x_1, \dots, x_k \in V$, with $k \ge 1$. A **convex combination** of $x_1, \dots, x_k$ is a vector

$$t_1x_1 + t_2x_2 + \cdots + t_kx_k \qquad \text{with } (t_1, \dots, t_k) \in \Delta_k,$$

that is, with **every** $t_i \ge 0$ and $t_1 + \cdots + t_k = 1$.

It is a linear combination with two restrictions on the coefficients, and each restriction has a job. Dropping "$t_i \ge 0$" gives the affine combinations of the next subsection; dropping "sum to 1" as well gives all linear combinations, and the span. A convex combination of two points is a point of the segment joining them, with $(t, 1-t)$ as the weights. The weights may be read as masses placed at the points, and the combination is then their center of mass. The case $k = 1$ is included and is $1 \cdot x_1 = x_1$.

Lemma 18.1.2. Convex Sets Contain Their Convex Combinations

Let C be a convex subset of a real vector space. Then every convex combination of points of C lies in C.

Idea. Peel off the last point. A convex combination of k points is a combination of two points, one of which is a convex combination of the first $k-1$ points, rescaled so that its weights again add up to 1. Induction on k does the rest; the only care needed is not to divide

by zero while rescaling.

Proof. We use induction on the number k of points. For $k = 1$, the combination is $x_1 \in C$. Let $k \geq 2$, suppose the statement holds for $k-1$ points, and let $x = \sum_{i=1}^{k} t_i x_i$ with $x_i \in C$ and $(t_1, \dots, t_k) \in \Delta_k$.

Case 1: $t_k = 1$. Then the other weights are nonnegative and add up to 0, so all of them are 0, and $x = x_k \in C$.

Case 2: $t_k < 1$. Put $s = 1 - t_k > 0$ and $y = \sum_{i=1}^{k-1} (t_i/s) x_i$. The division is legal because $s \neq 0$. The coefficients t_i/s are ≥ 0 and add up to $(1-t_k)/s = 1$, so y is a convex combination of $k-1$ points of C, and $y \in C$ by the inductive hypothesis. Then

$$x = sy + t_k x_k = (1 - t_k) y + t_k x_k,$$

which lies in C by Definition 18.1.1, applied to the points x_k and y of C with $t = t_k \in [0, 1]$. This completes the induction.

■

A set that is not convex can be enlarged to one that is. We want the most economical enlargement: nothing added that convexity does not force.

The convex hull of a set is the smallest convex set containing it.

Definition 18.1.4. Convex Hull

Let S be a subset of a real vector space V. The **convex hull** of S is the intersection of **all** convex subsets of V that contain S:

$$\operatorname{conv} S := \bigcap \{C \subseteq V : C \text{ is convex and } S \subseteq C\}.$$

The family being intersected is never empty, because V itself is convex and contains S. By Proposition 18.1.1 (a), $\operatorname{conv} S$ is convex; it contains S, since each set in the family does; and it lies inside every convex set containing S, since it is the intersection of all of them. So "smallest" in the slogan is literal: $\operatorname{conv} S$ is a convex set containing S and contained in every other one. In particular, $\operatorname{conv} C = C$ **exactly when** C is convex. This description from outside is clean but gives no way to decide whether a given point lies in the hull. The next theorem gives the description from inside.

Theorem 18.1.3. The Hull Is the Set of Convex Combinations

Let S be a subset of a real vector space V. Then $\operatorname{conv} S$ is the set of all convex combinations of points of S:

$$\operatorname{conv} S = \Big\{ \sum_{i=1}^{k} t_i x_i : k \geq 1,\ x_1, \dots, x_k \in S,\ (t_1, \dots, t_k) \in \Delta_k \Big\}.$$

Idea. Call the right-hand side D. One inclusion is Lemma 18.1.2 applied to the convex set $\operatorname{conv} S$. For the other, $\operatorname{conv} S$ is the smallest convex set containing S, so it suffices to show that D is **some** convex set containing S. An average of two averages is again an average, of the combined list.

Proof. Let D denote the right-hand side.

($\supseteq$) The set $\operatorname{conv} S$ is convex and contains S. By Lemma 18.1.2 it contains every convex combination of its own points, in particular of points of S. Hence $D \subseteq \operatorname{conv} S$.

($\subseteq$) Each $x \in S$ is the convex combination $1 \cdot x$, so $S \subseteq D$. To see that D is convex, let $x = \sum_{i=1}^{k} a_i x_i$ and $y = \sum_{j=1}^{l} b_j y_j$ be points of D, with all $x_i, y_j \in S$, and let $0 \le t \le 1$. Then

$$tx + (1-t)y = \sum_{i=1}^{k} (t a_i) x_i + \sum_{j=1}^{l} ((1-t) b_j) y_j,$$

a combination of the $k + l$ points $x_1, \dots, x_k, y_1, \dots, y_l$ of S, with coefficients $t a_i \ge 0$ and $(1-t) b_j \ge 0$ adding up to $t \cdot 1 + (1-t) \cdot 1 = 1$. So $tx + (1-t)y \in D$, and D is convex. Since $\operatorname{conv} S$ is contained in every convex set containing S, we get $\operatorname{conv} S \subseteq D$.

This proves $\operatorname{conv} S = D$.

■

A combination needs a point, so $\operatorname{conv} \emptyset = \emptyset$; and, padding with zero weights and merging repeated points, $\operatorname{conv}\{x_1, \dots, x_m\}$ is the set of all $\sum_i t_i x_i$ with $\mathbf{t} \in \Delta_m$; since $\sum_i t_i \mathbf{e}_i = (t_1, \dots, t_n)$, **the standard simplex is the hull of the standard basis**.

The theorem does not say how many points a combination needs. Every k is allowed, and for an infinite S nothing in it bounds k. That question is the business of §02.

Example. Deciding membership in a triangle

Let $\mathbf{p}_1 = (1, 0)$, $\mathbf{p}_2 = (3, 1)$, $\mathbf{p}_3 = (0, 2)$ in $\mathbb{R}^2$. Decide whether $\mathbf{x} = (\frac{5}{4}, \frac{3}{4})$ and $\mathbf{y} = (2, 2)$ lie in $\operatorname{conv}\{\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3\}$.

Solution. By Theorem 18.1.3 and the remark after it, a point $\mathbf{z}$ lies in the hull exactly when $\mathbf{z} = t_1\mathbf{p}_1 + t_2\mathbf{p}_2 + t_3\mathbf{p}_3$ for some $\mathbf{t} \in \Delta_3$. The two coordinates and the condition $t_1 + t_2 + t_3 = 1$ give three linear equations in t_1, t_2, t_3, with coefficient matrix

$$\mathbf{M} = \begin{pmatrix} 1 & 3 & 0 \\ 0 & 1 & 2 \\ 1 & 1 & 1 \end{pmatrix}, \qquad \det \mathbf{M} = 5 \neq 0.$$

So each $\mathbf{z}$ has exactly one solution $\mathbf{t}$, and $\mathbf{z}$ is in the hull if and only if that solution has no negative entry.

For $\mathbf{x}$, the solution is $\mathbf{t} = (\frac{1}{2}, \frac{1}{4}, \frac{1}{4})$. Check: $\frac{1}{2}(1,0) + \frac{1}{4}(3,1) + \frac{1}{4}(0,2) = (\frac{1}{2} + \frac{3}{4}, \frac{1}{4} + \frac{1}{2}) = (\frac{5}{4}, \frac{3}{4})$, and the weights add up to 1. All three are ≥ 0, so $\mathbf{x}$ lies in the hull.

For $\mathbf{y}$, the solution is $\mathbf{t} = (-\frac{2}{5}, \frac{4}{5}, \frac{3}{5})$. Check: $-\frac{2}{5}(1,0) + \frac{4}{5}(3,1) + \frac{3}{5}(0,2) = (-\frac{2}{5} + \frac{12}{5}, \frac{4}{5} + \frac{6}{5}) = (2,2)$, and $-\frac{2}{5} + \frac{4}{5} + \frac{3}{5} = 1$. The only solution has $t_1 < 0$, so **no** convex combination equals $\mathbf{y}$, and $\mathbf{y} \notin \operatorname{conv}\{\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3\}$. The negative weight even says where $\mathbf{y}$ is: on the far side of the line through $\mathbf{p}_2$ and $\mathbf{p}_3$ from $\mathbf{p}_1$.

Quick check. Is $(\frac{1}{2}, \frac{1}{2}, \frac{1}{2})$ in the convex hull of $\mathbf{0},\ \mathbf{e}_1 + \mathbf{e}_2,\ \mathbf{e}_2 + \mathbf{e}_3,\ \mathbf{e}_1 + \mathbf{e}_3$ in $\mathbb{R}^3$?

Solution. Yes. Give each of the four points weight $\frac{1}{4}$: the weights are ≥ 0 and add up to 1, and

$$\tfrac{1}{4}(\mathbf{0} + (\mathbf{e}_1 + \mathbf{e}_2) + (\mathbf{e}_2 + \mathbf{e}_3) + (\mathbf{e}_1 + \mathbf{e}_3)) = \tfrac{1}{4}(2,2,2) = (\tfrac{1}{2}, \tfrac{1}{2}, \tfrac{1}{2}).$$

By Theorem 18.1.3, the point lies in the hull. One convex combination is a complete certificate of membership; proving **non**-membership, as for $\mathbf{y}$ above, needs an argument about all combinations at once.

18.1.4 Affine combinations and the affine hull

Dropping the sign condition from a convex combination gives the other natural kind of average. It describes the flat pieces of V: lines, planes, and their analogues, not necessarily through 0. Chapter 4 §04 already has these objects, as cosets $p + U$ of subspaces, and the point of this subsection is to connect them with combinations.

Definition 18.1.5. Affine Combination, Affine Hull

Let V be a real vector space.

(a) An **affine combination** of $x_1, \dots, x_k \in V$, $k \geq 1$, is a vector $\sum_{i=1}^{k} t_i x_i$ with real t_i satisfying $t_1 + \cdots + t_k = 1$. No sign condition is imposed.

(b) An **affine subspace** of V is a coset $p + U$ of a subspace U (Definition 4.4.2). Its **dimension** is $\dim U$.

(c) The **affine hull** of $S \subseteq V$, written $\operatorname{aff} S$, is the set of all affine combinations of points of S.

The dimension in (b) is well defined, because the coset determines its subspace: $U = \{a - b : a, b \in p + U\}$. Indeed each such difference is $(p + u) - (p + w) = u - w \in U$, and each $u \in U$ is the difference $(p + u) - p$. So a line in $\mathbb{R}^3$ not through $\mathbf{0}$ is an affine subspace of dimension 1, and a single point $\{p\} = p + \{0\}$ is one of dimension 0.

Proposition 18.1.4. The Affine Hull Is a Coset

Let S be a **nonempty** subset of a real vector space V, let $x_0 \in S$, and put $U = \operatorname{span}\{x - x_0 : x \in S\}$. Then

$$\operatorname{aff} S = x_0 + U,$$

and $\operatorname{aff} S$ is contained in every affine subspace of V that contains S. In particular U does not depend on the choice of $x_0 \in S$.

Idea. Since the weights add up to 1, an affine combination is x_0 plus a linear combination of the differences $x_i - x_0$; that identity, read in both directions, gives the equality, and minimality follows because an affine subspace through x_0 is x_0 plus a subspace.

Proof. ($\subseteq$) Let $x = \sum_i t_i x_i$ with $x_i \in S$ and $\sum_i t_i = 1$. Since the t_i add up to 1, $x_0 = \sum_i t_i x_0$, and therefore

$$x = x_0 + \sum_i t_i(x_i - x_0) \in x_0 + U.$$

($\supseteq$) An element of x_0+U is $x_0+\sum_j c_j(y_j-x_0)$ for some $y_j \in S$ and real c_j, by Definition 1.3.2. It equals

$$\Big(1 - \sum_j c_j\Big)x_0 + \sum_j c_j y_j,$$

a combination of points of S whose coefficients add up to 1, hence an affine combination.

Now let $p+W$ be an affine subspace containing S. Since $x_0 \in p+W$, Lemma 4.4.3 (b) gives $p+W = x_0+W$. Each $x \in S$ also lies in x_0+W, so $x - x_0 \in W$. Thus W is a subspace containing every $x - x_0$, and $U \subseteq W$ by Theorem 1.3.1. Hence $\operatorname{aff} S = x_0 + U \subseteq x_0 + W$. Finally, U is the subspace determined by the coset $\operatorname{aff} S$, by the remark after Definition 18.1.5, and that coset does not mention x_0. This proves the proposition.

■

So the affine hull is to affine subspaces what the span is to subspaces, and what the convex hull is to convex sets: the smallest one containing S. It also gives every convex set a dimension. The **dimension** of a nonempty convex set C is $\dim \operatorname{aff} C$. A segment in $\mathbb{R}^3$ has dimension 1 and a triangle dimension 2. For Δ_n, take $\mathbf{x}_0 = \mathbf{e}_1$ in Proposition 18.1.4: each $\mathbf{t} - \mathbf{e}_1$ has coordinates adding up to 0, so U lies in the kernel H of $\mathbf{t} \mapsto t_1 + \cdots + t_n$, which has dimension $n-1$ by Theorem 2.3.1 (the map is surjective onto $\mathbb{R}$); and U contains the $n-1$ vectors $\mathbf{e}_i - \mathbf{e}_1$ ($i \geq 2$), independent because the i-th coordinate of $\sum_{j\geq 2} c_j(\mathbf{e}_j - \mathbf{e}_1)$ is c_i, so $U = H$ by Theorem 1.6.8. Hence Δ_n has dimension $n-1$, and it is flat inside $\mathbb{R}^n$: no ball of positive radius fits inside it.

18.1.5 Closed and open sets

The later sections take limits: §03 needs a nearest point of a convex set to exist, and §02 needs the hull of a compact set to be compact. Chapter 16 §02 used the right notion in passing, inside Corollary 16.2.5; here it is given a name, together with its partner.

Throughout this subsection, V is a **finite-dimensional** real vector space with a norm $\|\cdot\|$. By Example (Inner products, weights, and a function space) (a), an inner product space is a special case, with the induced norm. A sequence (x_k) **converges** to x, written $x_k \to x$, if $\|x_k - x\| \to 0$. By Corollary 16.2.3 (a), whether $x_k \to x$ does not depend on which norm is used, so neither does anything defined from it below.

A closed set keeps the limits of its sequences. An open set is one that every sequence converging to one of its points must eventually enter.

Definition 18.1.6. Closed Set, Open Set

Let V be a finite-dimensional real vector space with a norm, and let $K, U \subseteq V$.

(a) K is **closed** if, **whenever** (x_k) is a sequence of points **of** K with $x_k \to x$ for some $x \in V$, the limit x lies in K.

(b) U is **open** if, **whenever** $x_k \to x$ with $x \in U$, there is an index N such that $x_k \in U$ for **every** $k \geq N$.

By Corollary 16.2.3 (a), both properties are the same for every norm on V. Closedness in (a) is the notion used in Corollary 16.2.5.

In words: (a) demands that no convergent sequence inside K has its limit outside, and (b) that every sequence converging to a point of U is eventually inside U. This is the sequential form, matching the way Chapter 16's introduction defines continuity; it is the form taken as the definition, and Proposition 18.1.6 (a) gives the equivalent ball form, which §03 and §10 use.

Continuity will be needed for linear maps, and in finite dimension it costs nothing.

Lemma 18.1.5. Linear Maps Are Lipschitz

Let V and W be finite-dimensional real vector spaces with norms, and let $T\colon V \to W$ be linear. There is a constant $M \geq 0$ such that

$$\|Tx - Ty\| \leq M\|x - y\| \qquad \text{for all } x, y \in V.$$

In particular $x_k \to x$ implies $Tx_k \to Tx$; and every linear functional $\varphi\colon V \to \mathbb{R}$, with $|\cdot|$ as the norm on $\mathbb{R}$, is continuous.

Idea. In a basis, $\|Tx\|$ is at most a constant times the 1-norm of the coordinates of x, and norm equivalence trades that norm for the given one.

Proof. If $V = \{0\}$, then $T = 0$ and $M = 0$ works. Otherwise let $(v_1, \dots, v_n)$ be a basis of V, and write $x = \sum_i x_i v_i$. The function $N(x) = \sum_i |x_i|$ is a norm on V: it is the 1-norm of the coordinate vector (Example (The 1-, 2-, and ∞-norms on F^n)), and the coordinate map is linear (Theorem 1.5.2) and injective, so (N1), (N2) and (N3) transfer. By Theorem 16.2.2 there is $c > 0$ with $N(x) \leq c\|x\|$ for every x. By (N3) and (N2) of Definition 16.1.1 in W,

$$\|Tx\| = \Big\|\sum_i x_i T v_i\Big\| \leq \sum_i |x_i|\,\|Tv_i\| \leq \Big(\max_i \|Tv_i\|\Big) N(x) \leq M\|x\|,$$

with $M = c \max_i \|Tv_i\|$. Applying this to $x - y$ and using $Tx - Ty = T(x - y)$ gives the inequality. If $\|x_k - x\| \to 0$, then $\|Tx_k - Tx\| \leq M\|x_k - x\| \to 0$. The statement about φ is the case $W = \mathbb{R}$.

■

This is Step 1 of Lemma 16.3.1, with a basis of V in place of the standard basis of F^n.

Proposition 18.1.6. Closed and Open Sets: First Facts

Let V be a finite-dimensional real vector space with a norm.

(a) A subset $U \subseteq V$ is open **if and only if** for every $x \in U$ there is a real $r > 0$ with $\{y \in V : \|y - x\| < r\} \subseteq U$.

(b) The intersection of any family of closed sets is closed, and the union of finitely many closed sets is closed.

(c) For a linear functional φ on V and $c \in \mathbb{R}$, the sets $\{\varphi \le c\}$, $\{\varphi \ge c\}$, and $\{\varphi = c\}$ are closed. Every closed ball $\{x : \|x - p\| \le r\}$ is closed, and every subspace and every coset of a subspace is closed.

Idea. Each part unwinds the definitions. For the converse in (a), radii $1/k$ build a sequence breaking openness; for unions in (b), some set catches infinitely many terms.

Proof.

(a) ($\Leftarrow$) Let $x_k \to x \in U$, and take r as in the hypothesis. Since $\|x_k - x\| \to 0$, there is N with $\|x_k - x\| < r$ for $k \ge N$, and then $x_k \in U$.
($\Rightarrow$) Suppose some $x \in U$ has no such r. Then for each $k \ge 1$ the radius $r = 1/k$ fails, so there is $x_k \notin U$ with $\|x_k - x\| < 1/k$. Hence $x_k \to x \in U$ while no x_k lies in U, and U is not open.

(b) Let (K_α) be closed, and $x_k \to x$ with every $x_k \in \bigcap_\alpha K_\alpha$. For each α, the sequence lies in K_α, so $x \in K_\alpha$. Hence x lies in the intersection. For a union $K_1 \cup \cdots \cup K_m$: if $x_k \to x$ with every x_k in the union, then some K_i contains x_k for infinitely many k, because there are only m sets. Those terms form a subsequence of points of K_i converging to x, so $x \in K_i$ by closedness of K_i.

(c) Let $\varphi(x_k) \le c$ for every k, and $x_k \to x$. By Lemma 18.1.5, $\varphi(x_k) \to \varphi(x)$, and a nonstrict inequality survives a limit, so $\varphi(x) \le c$. The same argument handles $\ge$, and $\{\varphi = c\}$ is the intersection of the two, closed by (b). For the ball, let $\|x_k - p\| \le r$ and $x_k \to x$. By Lemma 16.1.1, $\big|\|x_k - p\| - \|x - p\|\big| \le \|x_k - x\| \to 0$, so $\|x - p\|$ is the limit of numbers $\le r$, and is $\le r$.

Finally let U be a subspace, with basis $(u_1, \dots, u_m)$. By Theorem 1.6.5 extend it to a basis $(u_1, \dots, u_m, v_{m+1}, \dots, v_n)$ of V, and let φ_j, for $m < j \le n$, send a vector to its coefficient on v_j; each φ_j is linear by Theorem 1.5.2. A vector lies in U exactly when these coefficients all vanish, since the basis expansion is unique (Theorem 1.5.1). So $U = \bigcap_{j>m} \{\varphi_j = 0\}$, which is closed by the previous paragraph and (b); when $m = n$, $U = V$, which is closed because every limit lies in V. A coset $p + U$ is closed because $x_k \to x$ with $x_k \in p + U$ gives $x_k - p \to x - p$ with $x_k - p \in U$, hence $x - p \in U$. This proves the proposition.

■

By (b) and (c), the solution set $\{\mathbf{x} : \mathbf{A}\mathbf{x} \le \mathbf{b}\}$ of a system of linear inequalities is closed as well as convex, and so is Δ_n. An open ball $\{x : \|x - p\| < r\}$ is open by (a): for x in it, the radius $r - \|x - p\| > 0$ works, by the triangle inequality.

Warning. **"Closed" is not the opposite of "open", and it is not "closed under an operation".** The interval $[0, 1) \subseteq \mathbb{R}$ is neither: $1 - 1/k$ stays inside and converges to $1 \notin [0, 1)$, and $-1/k$ converges to $0 \in [0, 1)$ without ever entering it. At the other extreme, $\emptyset$ and V are both closed and open. And $\mathbb{Q}^2$ is closed under addition and midpoints, yet not a

closed subset of $\mathbb{R}^2$: by fact (A1) of Chapter 16's introduction, $\sqrt{2}$ is a limit of rationals q_k, and $(q_k, 0) \to (\sqrt{2}, 0)$.

Quick check. Is the open half-plane $\{(x, y) \in \mathbb{R}^2 : x > 0\}$ closed? Is Δ_n open in $\mathbb{R}^n$?

Solution. Neither. The points $(1/k, 0)$ lie in the half-plane and converge to $(0, 0)$, which does not, so the half-plane is not closed. For Δ_n, take $\mathbf{x} = \mathbf{e}_1 \in \Delta_n$ and $\mathbf{x}_k = (1 + 1/k)\mathbf{e}_1 \to \mathbf{e}_1$; no $\mathbf{x}_k$ lies in Δ_n, because its coordinates add up to $1 + 1/k \neq 1$. So Δ_n is not open, which is the flatness of the simplex seen through sequences. (It is closed, by (b) and (c) of Proposition 18.1.6.)

With convex sets, hulls, and closed sets in hand, the question left open by Theorem 18.1.3 can be asked precisely: how many points does a convex combination need? The answer, in §02, depends only on the dimension, and it is what makes the hull of a compact set compact.

18.1.6 Exercises

18.1.6.1 A. Check your understanding

Exercise. A1

(a) State Definition 18.1.1 and Definition 18.1.3.

(b) True or false: the union of two convex sets is convex. Justify your answer.

(c) True or false: the intersection of two convex sets is convex. Justify your answer.

(d) Explain why the sphere $\{\mathbf{x} \in \mathbb{R}^3 : \|\mathbf{x}\|_2 = 1\}$ is **not** convex, naming the clause that fails.

(e) Describe $\operatorname{conv} S$ and $\operatorname{aff} S$ for $S = \{\mathbf{e}_1, \mathbf{e}_2\} \subseteq \mathbb{R}^2$.

(f) Give a subset of $\mathbb{R}$ that is neither open nor closed, and one that is both.

Solution.

(a) A subset C of a real vector space is convex if $tx + (1 - t)y \in C$ for all $x, y \in C$ and every $t \in [0, 1]$. A convex combination of $x_1, \dots, x_k$ is $\sum_i t_i x_i$ with every $t_i \geq 0$ and $\sum_i t_i = 1$.

(b) False. The two coordinate axes of $\mathbb{R}^2$ are convex, and their union contains $\mathbf{e}_1, \mathbf{e}_2$ but not $\frac{1}{2}(\mathbf{e}_1 + \mathbf{e}_2)$.

(c) True, by Proposition 18.1.1 (a): a point of a segment between two points of both sets lies in each set, by convexity of each.

(d) The clause "for all $\mathbf{x}, \mathbf{y} \in C$" fails for $\mathbf{x} = \mathbf{e}_1$, $\mathbf{y} = -\mathbf{e}_1$, $t = \frac{1}{2}$: both points lie on the sphere, and the midpoint $\mathbf{0}$ does not.

(e) By Theorem 18.1.3, $\operatorname{conv} S = \{t\mathbf{e}_1 + (1 - t)\mathbf{e}_2 : 0 \leq t \leq 1\}$, the segment from $\mathbf{e}_1$ to $\mathbf{e}_2$, which is Δ_2. By Definition 18.1.5, $\operatorname{aff} S = \{t\mathbf{e}_1 + (1 - t)\mathbf{e}_2 : t \in \mathbb{R}\}$, the whole line $x_1 + x_2 = 1$.

(f) $[0,1)$ is neither (see the warning above). $\mathbb{R}$ is both: every limit of real numbers lies in $\mathbb{R}$, so it is closed, and every sequence lies in $\mathbb{R}$, so it is open; the same holds for $\emptyset$ vacuously.

18.1.6.2 B. Practice

Exercise. B1: Determine which are convex

Determine which of the following sets are convex. Justify your answer.

(a) $\{(x,y) \in \mathbb{R}^2 : y \geq x^2\}$.
(b) $\{(x,y) \in \mathbb{R}^2 : xy \geq 1\}$.
(c) $\{\mathbf{A} \in M_2(\mathbb{R}) : \mathbf{A}^\top = \mathbf{A},\ \det \mathbf{A} \geq 0\}$.
(d) $\{\mathbf{x} \in \mathbb{R}^3 : x_1 + 2x_2 - x_3 = 4\}$.

Solution. Throughout, let $0 \leq t \leq 1$.

(a) Convex. Let $y \geq x^2$ and $y' \geq x'^2$, and put $x_t = tx + (1-t)x'$. Expanding,

$$tx^2 + (1-t)x'^2 - x_t^2 = t(1-t)(x-x')^2 \geq 0,$$

so $ty + (1-t)y' \geq tx^2 + (1-t)x'^2 \geq x_t^2$.

(b) Not convex. The points $(1,1)$ and $(-1,-1)$ satisfy $xy = 1$, but their midpoint $(0,0)$ has $xy = 0 < 1$.

(c) Not convex. $\mathbf{A} = \operatorname{diag}(1,0)$ and $\mathbf{B} = \operatorname{diag}(0,-1)$ are symmetric with $\det \mathbf{A} = \det \mathbf{B} = 0$, but $\frac{1}{2}(\mathbf{A}+\mathbf{B}) = \operatorname{diag}(\frac{1}{2}, -\frac{1}{2})$ has determinant $-\frac{1}{4} < 0$.

(d) Convex. It is the affine hyperplane $\{\varphi = 4\}$ for the linear functional $\varphi(\mathbf{x}) = x_1 + 2x_2 - x_3$, convex by Example (The first convex sets) (c). It is also a coset of $\ker \varphi$, convex by Example (The first convex sets) (b).

Exercise. B2: Membership in a hull

Let $\mathbf{p}_1 = (0,0)$, $\mathbf{p}_2 = (4,1)$, $\mathbf{p}_3 = (1,3)$ in $\mathbb{R}^2$. Decide whether $(2,2)$ and $(3,3)$ lie in $\operatorname{conv}\{\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3\}$. Justify your answer.

Solution. As in Example (Deciding membership in a triangle), $\mathbf{z}$ lies in the hull exactly when the system $t_1\mathbf{p}_1 + t_2\mathbf{p}_2 + t_3\mathbf{p}_3 = \mathbf{z}$, $t_1 + t_2 + t_3 = 1$ has a solution with every $t_i \geq 0$. Its coefficient matrix

$$\begin{pmatrix} 0 & 4 & 1 \\ 0 & 1 & 3 \\ 1 & 1 & 1 \end{pmatrix}$$

has determinant $11 \neq 0$ (expand along the first column: $1 \cdot (4 \cdot 3 - 1 \cdot 1) = 11$), so the solution is unique.

For $\mathbf{z} = (2,2)$: $\mathbf{t} = (\frac{1}{11}, \frac{4}{11}, \frac{6}{11})$. Check: $\frac{4}{11}(4,1) + \frac{6}{11}(1,3) = (\frac{16+6}{11}, \frac{4+18}{11}) = (2,2)$, and $\frac{1+4+6}{11} = 1$. All weights are ≥ 0, so $(2,2)$ is in the hull.

For $\mathbf{z} = (3,3)$: $\mathbf{t} = (-\frac{4}{11}, \frac{6}{11}, \frac{9}{11})$. Check: $\frac{6}{11}(4,1) + \frac{9}{11}(1,3) = (\frac{24+9}{11}, \frac{6+27}{11}) = (3,3)$, and $\frac{-4+6+9}{11} = 1$. The unique solution has $t_1 < 0$, so no convex combination equals $(3,3)$, and it is not in the hull.

18.1.6.3 C. Going deeper

Exercise. C1: Midpoint convexity

Let V be a finite-dimensional real vector space with a norm, and let $C \subseteq V$ be **midpoint convex**: $\frac{1}{2}(x+y) \in C$ whenever $x, y \in C$.

(a) Prove by induction on m that $tx + (1-t)y \in C$ for all $x, y \in C$ and every **dyadic** $t = j/2^m$ with $0 \le j \le 2^m$.

(b) Prove that if C is also closed, then C is convex.

(c) Give a midpoint convex subset of $\mathbb{R}$ that is not convex, and say which hypothesis of (b) it violates.

Hint for (b): approximate t by $\lfloor 2^m t \rfloor / 2^m$ (the floor exists and $2^{-m} \to 0$ by the completeness of $\mathbb{R}$, fact (A1) of Chapter 16's introduction).

Solution.

(a) For $m = 0$, $t \in \{0, 1\}$ and the point is y or x. Suppose the claim holds for m, and let $t = j/2^{m+1}$. If j is even, $t = (j/2)/2^m$ and the inductive hypothesis applies. If j is odd, then $t = \frac{1}{2}(t_- + t_+)$ with $t_\pm = (j \pm 1)/2^{m+1} = \frac{(j\pm1)/2}{2^m}$, which are dyadic of level m and lie in $[0,1]$. By the inductive hypothesis $z_\pm = t_\pm x + (1-t_\pm)y \in C$, and

$$\tfrac{1}{2}(z_- + z_+) = \tfrac{1}{2}(t_- + t_+)x + (1 - \tfrac{1}{2}(t_- + t_+))y = tx + (1-t)y,$$

which lies in C by midpoint convexity.

(b) Let $x, y \in C$ and $t \in [0,1]$. Put $t_m = \lfloor 2^m t \rfloor / 2^m$, a dyadic number in $[0,1]$ with $0 \le t - t_m < 2^{-m}$ (the floor exists and $2^{-m} \to 0$ by the completeness of $\mathbb{R}$, fact (A1) of Chapter 16's introduction). By (a), $z_m = t_m x + (1-t_m)y \in C$. Then

$$\|z_m - (tx + (1-t)y)\| = |t_m - t| \, \|x - y\| < 2^{-m}\|x - y\| \to 0,$$

so $z_m \to tx + (1-t)y$, and closedness of C puts the limit in C. Hence C is convex.

(c) $\mathbb{Q} \subseteq \mathbb{R}$. The midpoint of two rationals is rational, but $0, 1 \in \mathbb{Q}$ while $t \cdot 1 + (1-t)\cdot 0 = t \notin \mathbb{Q}$ for $t = 1/\sqrt{2}$. The set is not closed: the rationals $\lfloor 10^k/\sqrt{2} \rfloor / 10^k$ converge to $1/\sqrt{2}$ (the floor exists and $10^{-k} \to 0$ by the completeness of $\mathbb{R}$, fact (A1) of Chapter 16's introduction).

Exercise. C3: Sets that contain lines

Let V be a real vector space and $A \subseteq V$ a **nonempty** set containing $tx + (1-t)y$ for all $x, y \in A$ and **every** real t.

(a) Fix $a \in A$ and put $U = \{x - a : x \in A\}$. Prove that U is a subspace of V, so that $A = a + U$ is an affine subspace.

(b) Hence prove that a nonempty $S \subseteq V$ is an affine subspace if and only if $S = \operatorname{aff} S$.

Hint for (a): for closure under addition, use a midpoint and then a scaling.

Solution.

(a) $0 = a - a \in U$. *Scalars:* let $u = x - a \in U$ and $c \in \mathbb{R}$. The point $cx + (1-c)a$ lies in A by hypothesis, and it equals $a + cu$; so $cu \in U$. *Sums:* let $u = x - a$ and $w = y - a$ with $x, y \in A$. The midpoint $m = \frac{1}{2}x + \frac{1}{2}y \in A$, so $\frac{1}{2}(u + w) = m - a \in U$, and by the scalar step with $c = 2$, $u + w \in U$. By Theorem 1.2.1, U is a subspace, and $A = a + U$ by the definition of U.

(b) ($\Leftarrow$) If $S = \operatorname{aff} S$, then S is a coset by Proposition 18.1.4. ($\Rightarrow$) If $S = p + W$ is an affine subspace, then $\operatorname{aff} S \subseteq S$ by the minimality clause of Proposition 18.1.4, and $S \subseteq \operatorname{aff} S$ because each $x \in S$ is the affine combination $1 \cdot x$. So $S = \operatorname{aff} S$.

18.2 Carathéodory's Theorem

By Theorem 18.1.3, the convex hull of a set S consists of the convex combinations of points of S, with any number of points allowed. For a hundred points in the plane, that description permits a single point of the hull to be an average of all hundred at once. A picture says that this is wasteful: the hundred points spread out into a polygon, the polygon can be cut into triangles with corners among the points, and every point of the polygon lies in one of those triangles, so three points always suffice. This section proves the general statement. In a space of dimension n, every point of a convex hull is a convex combination of at most $n + 1$ points of the set, and the bound cannot be lowered. Two consequences follow. The hull of a compact set is compact, which §08 uses. And from the same counting argument come Radon's theorem and Helly's theorem, which describe how convex sets in $\mathbb{R}^n$ can intersect.

As in §01, all spaces are real. Carathéodory's theorem needs a **finite dimension** and nothing else; the compactness corollary needs, in addition, a norm, and any norm will do.

18.2.1 When there are too many points

The proof we want has a simple plan. If a point is written as a convex combination of **too many** points, find a way to shift the weights so that the point stays where it is while one weight falls to zero. That point can then be dropped. Two things are needed: a precise meaning of "too many", and a way to shift the weights without moving the point.

Both come from one observation. If $\sum_i \lambda_i x_i = 0$ and $\sum_i \lambda_i = 0$, then for every real s,

$$\sum_i (t_i - s\lambda_i) x_i = \sum_i t_i x_i \quad \text{and} \quad \sum_i (t_i - s\lambda_i) = \sum_i t_i.$$

So the weights $t_i - s\lambda_i$ describe the same point, and they still add up to 1. A relation of this shape is what we need a word for.

A list of points is affinely dependent when some combination with coefficients adding to zero, not all of them zero, cancels out.

Definition 18.2.1. Affine Independence

Let V be a real vector space and let $(x_0, x_1, \dots, x_m)$ be a list in V, $m \geq 0$. The list is **affinely independent** if, **for all** real $\lambda_0, \dots, \lambda_m$,

$$\sum_{i=0}^{m} \lambda_i x_i = 0 \quad \text{and} \quad \sum_{i=0}^{m} \lambda_i = 0 \quad \Longrightarrow \quad \lambda_0 = \lambda_1 = \cdots = \lambda_m = 0.$$

It is **affinely dependent** otherwise. A relation $\sum_i \lambda_i x_i = 0$ with $\sum_i \lambda_i = 0$ and the λ_i **not all zero** is an **affine dependence**.

In words: this is linear independence (Definition 1.4.1) with one extra hypothesis, that the coefficients add up to zero. The extra hypothesis makes the condition easier to satisfy, and it makes the notion blind to the choice of origin. Translating every point by p changes $\sum_i \lambda_i x_i$ by $(\sum_i \lambda_i)p = 0$, so a translated list is affinely dependent exactly when the original is.

Some small cases:

- A single point (x_0) is always affinely independent: $\lambda_0 = 0$ is forced by the sum condition alone.
- Two points (x_0, x_1) are affinely dependent exactly when they are equal: the condition forces $\lambda_1 = -\lambda_0$, and then $\lambda_0(x_0 - x_1) = 0$ has a solution $\lambda_0 \neq 0$ if and only if $x_0 = x_1$.
- Three collinear points are affinely dependent: $(0,0), (1,1), (2,2)$ satisfy $1\cdot(0,0) - 2\cdot(1,1) + 1\cdot(2,2) = \mathbf{0}$, with $1 - 2 + 1 = 0$.

A nonexample by minimal change. The list $(\mathbf{0}, \mathbf{e}_1, \mathbf{e}_2)$ in $\mathbb{R}^2$ is affinely independent: the $\mathbf{e}_1$- and $\mathbf{e}_2$-coordinates of $\lambda_0\mathbf{0} + \lambda_1\mathbf{e}_1 + \lambda_2\mathbf{e}_2 = \mathbf{0}$ give $\lambda_1 = \lambda_2 = 0$, and then the sum condition gives $\lambda_0 = 0$. Move the third point to $2\mathbf{e}_1$. The list $(\mathbf{0}, \mathbf{e}_1, 2\mathbf{e}_1)$ still spans the same line as $(\mathbf{0}, \mathbf{e}_1)$, but now $1\cdot\mathbf{0} - 2\mathbf{e}_1 + 1\cdot(2\mathbf{e}_1) = \mathbf{0}$ with $1 - 2 + 1 = 0$: the three points have fallen onto one line, and the implication in the definition fails.

Warning. **Affine independence is not linear independence.** The list $(\mathbf{0}, \mathbf{e}_1, \mathbf{e}_2)$ in $\mathbb{R}^2$ is linearly dependent, since it contains $\mathbf{0}$, yet it was just shown to be affinely independent. A triangle is an affinely independent list of three points in a two-dimensional space, which linear independence would never allow. That is why the bound in this section is $n + 1$, and not n. The implication does run one way: a linearly independent list is affinely independent, since an affine dependence is in particular a linear relation with coefficients not all zero.

The next lemma converts affine dependence into linear dependence of differences, which is what allows a dimension count.

Lemma 18.2.1. Affine Dependence by Differences

Let V be a real vector space and $x_0, \dots, x_m \in V$.

(a) The list $(x_0, \dots, x_m)$ is affinely independent **if and only if** the list $(x_1 - x_0, \dots, x_m - x_0)$ is linearly independent.

(b) If $\dim V = n$, every affinely independent list in V has at most $n+1$ entries. Equivalently, every list of $n + 2$ or more points of V is affinely dependent.

Idea. The sum condition determines λ_0 from the other coefficients, and eliminating it turns an affine relation among the points into a linear relation among the differences, and back.

Proof.

(a) Suppose $\sum_{i=0}^m \lambda_i x_i = 0$ and $\sum_{i=0}^m \lambda_i = 0$. Then $\lambda_0 = -\sum_{i\geq 1} \lambda_i$, and substituting,

$$\sum_{i=1}^m \lambda_i (x_i - x_0) = \sum_{i=0}^m \lambda_i x_i = 0.$$

Conversely, if $\sum_{i=1}^m \lambda_i (x_i - x_0) = 0$, then setting $\lambda_0 = -\sum_{i\geq 1} \lambda_i$ gives an affine relation $\sum_{i=0}^m \lambda_i x_i = 0$ with $\sum_{i=0}^m \lambda_i = 0$. These two passages are inverse to each other, and $(\lambda_0, \dots, \lambda_m)$ is not all zero exactly when $(\lambda_1, \dots, \lambda_m)$ is not all zero, because λ_0 is determined by the others. So affine dependences of the points correspond to linear relations, not all zero, among the differences, and one exists exactly when the other does.

(b) If $(x_0, \dots, x_m)$ is affinely independent, then by (a) the m differences $x_i - x_0$ are linearly independent, so $m \leq n$ by Theorem 1.6.3 (a). The list has $m + 1 \leq n + 1$ entries.

■

18.2.2 The theorem

Now the plan can be carried out. The statement is sharper than the count: the points can be taken affinely independent, and the count is a consequence.

Theorem 18.2.2. Carathéodory's Theorem

Let V be a real vector space with $\dim V = n$, and let $S \subseteq V$. Every point of $\operatorname{conv} S$ is a convex combination of an **affinely independent** list of points of S. In particular, every point of $\operatorname{conv} S$ is a convex combination of **at most** $n + 1$ points of S.

Idea. **Count, then drop a point.** Write x with as few points as possible, so that every weight is positive. If the points were affinely dependent, the weights $t_i - s\lambda_i$ would describe the same x for every s. Start at $s = 0$ and increase s. The weights with $\lambda_i > 0$ decrease, and there is at least one of these, because the λ_i add up to zero without all being zero. Stop at the first moment a weight reaches 0: all the weights are still ≥ 0, and one point has become unnecessary. That contradicts minimality.

Proof. Let $x \in \operatorname{conv} S$. By Theorem 18.1.3, x is a convex combination of finitely many points of S. Among all ways of writing it so, choose one,

$$x = \sum_{i=0}^m t_i x_i, \qquad x_i \in S, \quad (t_0, \dots, t_m) \in \Delta_{m+1},$$

with the number $m+1$ of points as small as possible. Then every $t_i > 0$: a zero weight could

be deleted, and the remaining weights would still be nonnegative and add up to 1, giving a shorter expression.

Suppose, for a contradiction, that $(x_0, \dots, x_m)$ is affinely dependent, and let $\sum_i \lambda_i x_i = 0$, $\sum_i \lambda_i = 0$ be an affine dependence. The λ_i are not all zero and add up to 0, so at least one of them is positive. Put

$$s = \min\Big\{ \frac{t_i}{\lambda_i} : \lambda_i > 0 \Big\},$$

a minimum over a nonempty finite set, and let j be an index where it is attained; $s > 0$ since each $t_i > 0$. Define $t_i' = t_i - s\lambda_i$. Then:

- $t_i' \geq 0$ for every i. If $\lambda_i \leq 0$, then $t_i' \geq t_i > 0$, since $s > 0$. If $\lambda_i > 0$, then $s \leq t_i/\lambda_i$ by the choice of s, so $s\lambda_i \leq t_i$.
- $t_j' = t_j - (t_j/\lambda_j)\lambda_j = 0$.
- $\sum_i t_i' = \sum_i t_i - s\sum_i \lambda_i = 1$ and $\sum_i t_i' x_i = \sum_i t_i x_i - s\sum_i \lambda_i x_i = x$.

So x is a convex combination of the m points x_i, $i \neq j$, contradicting the minimality of $m+1$. Hence $(x_0, \dots, x_m)$ is affinely independent. By Lemma 18.2.1 (b) it has at most $n+1$ entries. This proves the theorem.

■

In the plane, then, every point of the hull of any set S lies in a triangle whose corners are points of S, possibly a degenerate one (a segment or a single point). In $\mathbb{R}^3$ it lies in a tetrahedron with corners in S. The proof is constructive: given any convex combination, it says how to shorten it, one point at a time.

Example. Dropping a point in the plane

In $\mathbb{R}^2$ let $\mathbf{x}_0 = (0,0)$, $\mathbf{x}_1 = (3,0)$, $\mathbf{x}_2 = (0,3)$, $\mathbf{x}_3 = (2,2)$, and let $\mathbf{x} = \frac{1}{4}(\mathbf{x}_0 + \mathbf{x}_1 + \mathbf{x}_2 + \mathbf{x}_3) = (\frac{5}{4}, \frac{5}{4})$. Write $\mathbf{x}$ as a convex combination of three of these points by the method of the proof.

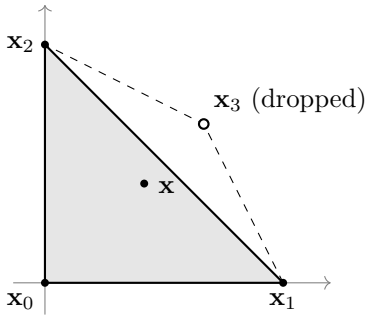
Solution. Four points in $\mathbb{R}^2$ are affinely dependent by Lemma 18.2.1 (b). To find a dependence, solve $\sum_i \lambda_i \mathbf{x}_i = \mathbf{0}$ and $\sum_i \lambda_i = 0$: the two coordinate equations read $3\lambda_1 + 2\lambda_3 = 0$ and $3\lambda_2 + 2\lambda_3 = 0$. Taking $\lambda_3 = 3$ gives $\lambda_1 = \lambda_2 = -2$, and then $\lambda_0 = -(\lambda_1 + \lambda_2 + \lambda_3) = 1$. Check: $1\cdot(0,0) - 2(3,0) - 2(0,3) + 3(2,2) = (-6+6, -6+6) = \mathbf{0}$, and $1 - 2 - 2 + 3 = 0$.

The positive coefficients are $\lambda_0 = 1$ and $\lambda_3 = 3$, with ratios $t_0/\lambda_0 = \frac{1}{4}$ and $t_3/\lambda_3 = \frac{1}{12}$. So $s = \frac{1}{12}$, attained at $j = 3$, and the new weights are

$$\mathbf{t}' = (\tfrac{1}{4} - \tfrac{1}{12},\ \tfrac{1}{4} + \tfrac{2}{12},\ \tfrac{1}{4} + \tfrac{2}{12},\ \tfrac{1}{4} - \tfrac{3}{12}) = (\tfrac{1}{6},\ \tfrac{5}{12},\ \tfrac{5}{12},\ 0).$$

Check: $\frac{5}{12}(3,0) + \frac{5}{12}(0,3) = (\frac{5}{4}, \frac{5}{4}) = \mathbf{x}$, and $\frac{2}{12} + \frac{5}{12} + \frac{5}{12} = 1$. So $\mathbf{x} = \frac{1}{6}\mathbf{x}_0 + \frac{5}{12}\mathbf{x}_1 + \frac{5}{12}\mathbf{x}_2$ lies in the triangle with corners $\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2$, and $\mathbf{x}_3$ has been dropped.

The minimum matters. Using the other positive coefficient, $s = \frac{1}{4}$, would make the weight of $\mathbf{x}_3$ equal to $\frac{1}{4} - \frac{3}{4} = -\frac{1}{2} < 0$: the point would still be described correctly, but not as a convex combination.



$\mathbf{x}$ is an average of all four points,
and already lies in the shaded triangle of three

The bound $n+1$ cannot be improved, and the standard simplex shows it.

Example. The bound is attained

In $\mathbb{R}^n$, let $S = \{\mathbf{0}, \mathbf{e}_1, \dots, \mathbf{e}_n\}$ and let $\mathbf{c} = \frac{1}{n+1}(\mathbf{0} + \mathbf{e}_1 + \cdots + \mathbf{e}_n) = \frac{1}{n+1}\mathbf{1}$. Show that $\mathbf{c} \in \operatorname{conv} S$ is not a convex combination of any n points of S.

Solution. The point $\mathbf{c}$ is a convex combination of the $n+1$ points of S, each with weight $\frac{1}{n+1}$. A convex combination of n of the points leaves one out. If it leaves out $\mathbf{e}_j$, the combination is $t_0\mathbf{0} + \sum_{i \neq j} t_i \mathbf{e}_i$, whose j-th coordinate is 0, while $c_j = \frac{1}{n+1} \neq 0$. If it leaves out $\mathbf{0}$, the combination is $\sum_i t_i \mathbf{e}_i = (t_1, \dots, t_n)$, whose coordinates add up to 1, while those of $\mathbf{c}$ add up to $\frac{n}{n+1} \neq 1$. Either way it is not $\mathbf{c}$.

The bound can be far from attained for a particular set, however, and it is worth seeing one.

Example. Two points suffice for a circle

Let $S = \{\mathbf{x} \in \mathbb{R}^2 : \|\mathbf{x}\|_2 = 1\}$ be the unit circle. Show that $\operatorname{conv} S$ is the closed unit disc B_2, and that every point of B_2 is a convex combination of **two** points of S.

Solution. B_2 is convex (Example (The first convex sets) (d)) and contains S, so $\operatorname{conv} S \subseteq B_2$ by Definition 18.1.4. Conversely, let $(a, b) \in B_2$, so $a^2 + b^2 \leq 1$, and put $c = \sqrt{1 - a^2}$, which is real since $|a| \leq 1$. The points (a, c) and $(a, -c)$ lie on S, and $|b| \leq c$ because $b^2 \leq 1 - a^2$. With $t = \frac{c+b}{2c}$ when $c > 0$,

$$t(a, c) + (1 - t)(a, -c) = (a,\ (2t - 1)c) = (a, b),$$

and $0 \leq t \leq 1$ because $-c \leq b \leq c$. When $c = 0$, we have $a = \pm 1$ and $b = 0$, and (a, b) is itself a point of S. So every point of the disc lies on a vertical chord of the circle, and $B_2 \subseteq \operatorname{conv} S$ by Theorem 18.1.3.

Quick check. A point of $\mathbb{R}^3$ is given as a convex combination of seven points of a set S. What is the smallest number of points of S that Theorem 18.2.2 guarantees will suffice? Can one always manage with three?

Solution. Four, since $n+1=4$ for $n=3$. Three do not always suffice: by Example (The bound is attained) with $n=3$, the point $\frac{1}{4}(1,1,1)$ is a convex combination of $\mathbf{0}, \mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$ and of no three of them.

Remark. The only place the proof used $\dim V = n$ was the final count, through Lemma 18.2.1 (b), and all that count needs is a space containing the differences of the points. If S lies in an affine subspace of dimension d, then $n+1$ can be replaced by $d+1$; Exercise (C1: Carathéodory in the affine hull) carries this out. A set of points on a line in $\mathbb{R}^{100}$ needs only two at a time.

18.2.3 The hull of a compact set

The first use of the theorem is a statement about limits. Recall from Corollary 16.2.5 that a subset K of a finite-dimensional real space with a norm is **compact** if every sequence in K has a subsequence converging to a point **of** K. Compactness is what lets later sections take a maximum, or a nearest point, over a hull, and it is not obvious that a hull keeps it. A convex combination may use arbitrarily many points, so a sequence of combinations could use more and more of them, and there would be no single compact set of parameters from which to extract a convergent subsequence. Carathéodory's bound removes that possibility: every point of the hull is described by exactly $n+1$ points of K and $n+1$ weights.

Corollary 18.2.3. The Hull of a Compact Set Is Compact

Let V be a finite-dimensional real vector space with a norm, and let $K \subseteq V$ be compact.

(a) $\operatorname{conv} K$ is compact.

(b) If K is nonempty, then every linear functional φ on V attains a maximum on $\operatorname{conv} K$, and

$$\max_{x \in \operatorname{conv} K} \varphi(x) = \max_{x \in K} \varphi(x).$$

Idea. For (a), write each term of a sequence in $\operatorname{conv} K$ as a combination of exactly $n+1$ points of K, using Theorem 18.2.2 and padding with zero weights. That turns one sequence into $n+2$ sequences: one of weight vectors in Δ_{n+1}, and $n+1$ of points of K. Extract a convergent subsequence from each in turn, a subsequence of the previous one each time, and pass to the limit in the combination. For (b), a linear functional cannot do better on an average than on the best of the points averaged.

Proof. Let $n = \dim V$.

(a) If $K = \emptyset$, then $\operatorname{conv} K = \emptyset$, which is compact vacuously. Otherwise let (y_k) be a sequence in $\operatorname{conv} K$. By Theorem 18.2.2, each y_k is a convex combination of at most $n+1$ points of K; repeating one of those points with weight 0 if necessary, we may write

$$y_k = \sum_{i=0}^{n} t_{k,i}\, x_{k,i}, \qquad x_{k,i} \in K, \quad \mathbf{t}_k = (t_{k,0}, \dots, t_{k,n}) \in \Delta_{n+1}.$$

Step 1: the weights. Δ_{n+1} is closed in $\mathbb{R}^{n+1}$ by Proposition 18.1.6 (b), (c), and bounded, since each coordinate of a point of it lies in $[0,1]$. By **the compactness of closed bounded sets in $\mathbb{R}^{n+1}$, fact (A3) of Chapter 16's introduction**, some subsequence of $(\mathbf{t}_k)$ converges to a point $\mathbf{t} = (t_0, \dots, t_n) \in \Delta_{n+1}$. In particular $t_{k,i} \to t_i$ along it for each i, since $|t_{k,i} - t_i| \le \|\mathbf{t}_k - \mathbf{t}\|_2$.
Step 2: the points. Along that subsequence, the sequence $(x_{k,0})$ lies in K, so by compactness of K a further subsequence has $x_{k,0} \to x_0 \in K$. Along it, a further subsequence has $x_{k,1} \to x_1 \in K$, and so on through $i = n$. A subsequence of a convergent sequence converges to the same limit, so along the final subsequence, which we index by k again, $\mathbf{t}_k \to \mathbf{t}$ and $x_{k,i} \to x_i$ for every i.
Step 3: the limit. Put $y = \sum_{i=0}^n t_i x_i$, which lies in $\operatorname{conv} K$ by Theorem 18.1.3. By (N3) and (N2) of Definition 16.1.1, along the final subsequence,

$$\begin{aligned} \|y_k - y\| &\le \sum_{i=0}^n \|t_{k,i}x_{k,i} - t_i x_i\| \\ &\le \sum_{i=0}^n \Big(|t_{k,i} - t_i| \, \|x_{k,i}\| + t_i \|x_{k,i} - x_i\| \Big), \end{aligned}$$

where the second line writes $t_{k,i}x_{k,i} - t_i x_i = (t_{k,i} - t_i)x_{k,i} + t_i(x_{k,i} - x_i)$. Each $\|x_{k,i}\| \le \|x_i\| + \|x_{k,i} - x_i\|$ is bounded, since the last term tends to 0. So every term on the right tends to 0, and $y_k \to y \in \operatorname{conv} K$ along the subsequence. Hence $\operatorname{conv} K$ is compact.

(b) By Lemma 18.1.5, φ is continuous. By **the extreme value theorem, fact (A4) of Chapter 16's introduction**, applied to φ on the nonempty compact set K, there is $x^* \in K$ with $\varphi(x) \le M = \varphi(x^*)$ for every $x \in K$. Now let $y = \sum_i t_i x_i$ be any point of $\operatorname{conv} K$, with $x_i \in K$ and $\mathbf{t} \in \Delta_m$ (Theorem 18.1.3). By linearity, and since every $t_i \ge 0$,

$$\varphi(y) = \sum_i t_i \varphi(x_i) \le \sum_i t_i M = M.$$

So M bounds φ on $\operatorname{conv} K$, and it is attained there, at $x^* \in K \subseteq \operatorname{conv} K$. This proves the corollary.

■

For instance, $\varphi(x,y) = x + 2y$ takes the values $0, 3, 6$ at the corners $(0,0), (3,0), (0,3)$ of the triangle in Example (Dropping a point in the plane); as in the proof of (b), at a point $\sum_i t_i \mathbf{x}_i$ of the triangle it takes the average $3t_1 + 6t_2 \le 6$ of these values, so its maximum over the triangle is 6, attained at $(0,3)$. At the interior point $\mathbf{x} = (\frac{5}{4}, \frac{5}{4})$ it takes $\frac{15}{4} = \frac{1}{6} \cdot 0 + \frac{5}{12} \cdot 3 + \frac{5}{12} \cdot 6$, an average of the corner values and so no larger than the largest.

To maximize a linear functional over the hull of finitely many points it is enough to evaluate it at those points (Exercise (B3: Linear functionals on a polygon)). §08 extends this to every compact convex set, and §09 shows that a bounded feasible region of §06 is such a hull.

Warning. **The hull of a closed set need not be closed.** In $\mathbb{R}^2$, let $S = \{(0,1)\} \cup \{(x,0) : x \in \mathbb{R}\}$, a point together with a line. S is closed, by Proposition 18.1.6 (b), (c). Its hull is

$$\operatorname{conv} S = \{(x,y) : 0 \le y < 1\} \cup \{(0,1)\}.$$

Indeed a convex combination of points of S with total weight y on $(0,1)$ has second coordinate y, and when $y < 1$ any first coordinate x is reached, as $y(0,1) + (1-y)(x/(1-y), 0) = (x,y)$;

when $y = 1$ all the weight is on $(0, 1)$. The points $(1, 1 - 1/k)$ lie in $\operatorname{conv} S$ and converge to $(1, 1)$, which does not. So compactness in Corollary 18.2.3 cannot be weakened to closedness: here S is closed but unbounded, and the hull loses the line $y = 1$, except for the one point that belongs to S.

18.2.4 Radon's theorem and Helly's theorem

The dimension count behind Carathéodory's theorem says something about intersections of convex sets as well. Its first form is a statement about $n + 2$ points.

Theorem 18.2.4. Radon's Theorem

Let V be a real vector space with $\dim V = n$, and let $x_1, \dots, x_m \in V$ with $m \geq n + 2$. Then the index set $\{1, \dots, m\}$ can be split into two **disjoint** nonempty sets I and J with

$$\operatorname{conv}\{x_i : i \in I\} \cap \operatorname{conv}\{x_j : j \in J\} \neq \emptyset.$$

Idea. Take an affine dependence and separate its positive coefficients from the rest. Moving the negative terms across the equation gives an equation between a positive combination of some points and a positive combination of the others, with the **same total weight** on each side, because the coefficients add up to zero. Divide by that total, and each side becomes a convex combination.

Proof. By Lemma 18.2.1 (b), the list $(x_1, \dots, x_m)$ is affinely dependent: there are $\lambda_1, \dots, \lambda_m$, not all zero, with $\sum_i \lambda_i x_i = 0$ and $\sum_i \lambda_i = 0$. Let $I = \{i : \lambda_i > 0\}$ and $J = \{j : \lambda_j \leq 0\}$, which are disjoint with union $\{1, \dots, m\}$. Since the λ_i add up to 0 and are not all zero, some λ_i is positive and some is negative, so neither I nor J is empty. Put $\sigma = \sum_{i \in I} \lambda_i > 0$; then $\sum_{j \in J} (-\lambda_j) = \sigma$ as well, since the total is 0. Moving the terms indexed by J across and dividing by σ,

$$y := \sum_{i \in I} \frac{\lambda_i}{\sigma} x_i = \sum_{j \in J} \frac{-\lambda_j}{\sigma} x_j.$$

On the left the coefficients are positive and add up to 1; on the right they are ≥ 0 and add up to 1. So y lies in both hulls, by Theorem 18.1.3. This proves the theorem.

■

A partition as in the theorem is a **Radon partition**. In the plane, four points in general position (no three on a line) either have one inside the triangle of the other three, or they form a quadrilateral whose two diagonals cross; Exercise (B2: Radon partitions in the plane) finds the partition in each case. The count $n + 2$ cannot be lowered: the $n + 1$ points $\mathbf{0}, \mathbf{e}_1, \dots, \mathbf{e}_n$ have no Radon partition (Exercise (C3: The counts are sharp)).

Helly's theorem turns the counting around. Instead of asking how many points a hull needs, it asks how many sets of a family must be checked to know that all of them have a common point. Checking pairs is not enough, since three lines in the plane can meet pairwise in three different points; checking triples in the plane is.

Theorem 18.2.5. Helly's Theorem

Let V be a real vector space with $\dim V = n$, and let $C_1, \dots, C_m$ be convex subsets of V, with $m \geq n+1$. If **every** $n+1$ of the sets have a point in common, then all m of them have a point in common.

Idea. Induction on the number of sets. Given $m \geq n+2$ sets, the inductive hypothesis provides, for each i, a point x_i lying in every set **except possibly** C_i. That is $m \geq n+2$ points, so Radon's theorem splits them into two groups whose hulls meet at some y. Any one set C_k misses at most one point, x_k, and that point lies in only one of the two groups; the other group lies entirely in C_k, and so does its hull, which contains y.

Proof. We use induction on m. For $m = n+1$, the hypothesis applied to all m sets is the conclusion. Let $m \geq n+2$, and suppose the theorem holds for families of $m-1$ convex sets. For each $i = 1, \dots, m$, the family $\{C_k : k \neq i\}$ has $m-1 \geq n+1$ members, and every $n+1$ of them have a common point by hypothesis. By the inductive hypothesis, there is a point x_i with

$$x_i \in C_k \quad \text{for every } k \neq i.$$

Since $m \geq n+2$, Theorem 18.2.4 splits $\{1, \dots, m\}$ into disjoint nonempty sets I, J and gives a point $y \in \operatorname{conv}\{x_i : i \in I\} \cap \operatorname{conv}\{x_j : j \in J\}$.

Fix $k \in \{1, \dots, m\}$. If $k \in I$, then $k \notin J$, so every x_j with $j \in J$ lies in C_k. Since C_k is convex, Lemma 18.1.2 puts every convex combination of these points in C_k, so $\operatorname{conv}\{x_j : j \in J\} \subseteq C_k$ by Theorem 18.1.3, and in particular $y \in C_k$. If $k \in J$, the same argument with I in place of J gives $y \in C_k$. Hence y lies in every C_k. This completes the induction.

■

Example. Helly's proof run on four half-planes

In $\mathbb{R}^2$, let

$$C_1 = \{y \geq 0\}, \qquad C_2 = \{x \geq 0\},$$
$$C_3 = \{x + y \leq 2\}, \quad C_4 = \{x - y \leq 1\}.$$

Find points $\mathbf{x}_i$ in all the sets except C_i, and follow the proof of Theorem 18.2.5 to a common point of all four.

Solution. Here $n = 2$ and $m = 4 = n+2$, so the proof applies Radon's theorem (Theorem 18.2.4) to four points. The following points lie in every set except their own, as substitution into the four inequalities shows:

$$\mathbf{x}_1 = (0, -1), \quad \mathbf{x}_2 = (-1, 0), \quad \mathbf{x}_3 = (2, 2), \quad \mathbf{x}_4 = (2, 0).$$

For instance, $\mathbf{x}_1$ has $x = 0 \geq 0$, $x + y = -1 \leq 2$, $x - y = 1 \leq 1$, and $y = -1 < 0$ violates C_1. And $\mathbf{x}_4$ has $y = 0$, $x = 2$, $x + y = 2$, and $x - y = 2 > 1$ violates C_4.

An affine dependence: the equations $\sum_i \lambda_i \mathbf{x}_i = \mathbf{0}$, $\sum_i \lambda_i = 0$ are satisfied by $\boldsymbol{\lambda} = (-6, 4, -3, 5)$. Check: the first coordinates give $-4 - 6 + 10 = 0$, the second $6 - 6 + 0 = 0$,

and $-6+4-3+5=0$. The positive coefficients are at $I=\{2,4\}$, with $\sigma=9$, so

$$\mathbf{y}=\tfrac{4}{9}\mathbf{x}_2+\tfrac{5}{9}\mathbf{x}_4=(\tfrac{-4+10}{9},0)=(\tfrac{2}{3},0),$$

and from the other side $\frac{6}{9}\mathbf{x}_1+\frac{3}{9}\mathbf{x}_3=(\frac{6}{9},\frac{-6+6}{9})=(\frac{2}{3},0)$, as it must be. Finally $\mathbf{y}=(\frac{2}{3},0)$ satisfies $y=0\geq 0$, $x=\frac{2}{3}\geq 0$, $x+y=\frac{2}{3}\leq 2$ and $x-y=\frac{2}{3}\leq 1$: it lies in all four half-planes. The proof explains why, without any checking: C_2 and C_4 contain $\mathbf{x}_1$ and $\mathbf{x}_3$, hence the segment between them, which contains $\mathbf{y}$; and C_1 and C_3 contain $\mathbf{x}_2$ and $\mathbf{x}_4$, hence the other segment through $\mathbf{y}$.

Remark. Helly's theorem is about **finite** families, and the finiteness cannot simply be dropped. The closed half-lines $[k,\infty)\subseteq\mathbb{R}$, for $k=1,2,3,\ldots$, are convex, and any finitely many of them meet (at their largest k), but no real number lies in all of them. For compact convex sets the infinite version does hold; Exercise (C2: Helly's theorem for a sequence of compact sets) proves it for a sequence of sets.

With these results the chapter has its combinatorial foundation. §03 turns to geometry: a point outside a closed convex set can be separated from it by a hyperplane.

18.2.5 Exercises

18.2.5.1 A. Check your understanding

Exercise. A1

(a) State Theorem 18.2.2.

(b) True or false: every point in the convex hull of ten points of $\mathbb{R}^2$ is a convex combination of three of them. Justify your answer.

(c) True or false: the convex hull of a closed subset of $\mathbb{R}^2$ is closed. Justify your answer.

(d) True or false: any four points of $\mathbb{R}^3$ have a Radon partition. Justify your answer.

(e) Is the list $(\mathbf{e}_1,\mathbf{e}_2,\mathbf{e}_3)$ in $\mathbb{R}^3$ affinely independent? Is $(\mathbf{e}_1,\mathbf{e}_2,\mathbf{e}_1+\mathbf{e}_2-\mathbf{e}_3,\mathbf{e}_3)$?

(f) Which facts from Chapter 16's list of imported analysis does Corollary 18.2.3 use, and where?

Solution.

(a) If $\dim V=n$ and $S\subseteq V$, every point of $\operatorname{conv}S$ is a convex combination of an affinely independent list of points of S, and so of at most $n+1$ points of S.

(b) True, by Theorem 18.2.2 with $n=2$: at most three are needed, and if fewer are used, a point can be added with weight 0.

(c) False. The warning in this section gives $S=\{(0,1)\}\cup\{(x,0):x\in\mathbb{R}\}$, closed, with $(1,1)$ a limit of points of $\operatorname{conv}S$ not in $\operatorname{conv}S$.

(d) False. Radon's theorem (Theorem 18.2.4) needs $n+2=5$ points in $\mathbb{R}^3$, and four can fail. The points $\mathbf{0},\mathbf{e}_1,\mathbf{e}_2,\mathbf{e}_3$ are affinely independent, since $\mathbf{e}_1,\mathbf{e}_2,\mathbf{e}_3$ are linearly independent (Lemma 18.2.1 (a)). A point common to the hulls of two disjoint groups of them would give two convex combinations with equal value, and subtracting them

gives a relation with coefficients adding up to 0 and not all zero, which is an affine dependence. So there is no Radon partition; Exercise (C3: The counts are sharp) carries this out for $n+1$ affinely independent points in general.

(e) Yes: it is linearly independent, and a linearly independent list is affinely independent. The second list is affinely dependent: $\mathbf{e}_1+\mathbf{e}_2-(\mathbf{e}_1+\mathbf{e}_2-\mathbf{e}_3)-\mathbf{e}_3=\mathbf{0}$ with coefficients $1,1,-1,-1$, which add up to 0.

(f) (A3), the compactness of closed bounded sets, for the weights in part (a); and (A4), the extreme value theorem, for the maximum over K in part (b). Part (a) also uses the hypothesis that K is compact, once for each of the $n+1$ points.

18.2.5.2 B. Practice

Exercise. B1: Shortening a combination

In $\mathbb{R}^2$, let $\mathbf{x}_1=(0,0)$, $\mathbf{x}_2=(4,0)$, $\mathbf{x}_3=(4,4)$, $\mathbf{x}_4=(0,4)$, and

$$\mathbf{x}=\tfrac{1}{10}\mathbf{x}_1+\tfrac{2}{10}\mathbf{x}_2+\tfrac{3}{10}\mathbf{x}_3+\tfrac{4}{10}\mathbf{x}_4.$$

(a) Compute $\mathbf{x}$, and find an affine dependence among $\mathbf{x}_1,\dots,\mathbf{x}_4$.

(b) Use the method of Theorem 18.2.2 to write $\mathbf{x}$ as a convex combination of three of the points.

(c) Apply the method with the dependence multiplied by -1. Which point is dropped now? Hence $\mathbf{x}$ lies in two different triangles with corners among the $\mathbf{x}_i$.

Solution.

(a) $\mathbf{x}=\frac{2}{10}(4,0)+\frac{3}{10}(4,4)+\frac{4}{10}(0,4)=(\frac{8+12}{10},\frac{12+16}{10})=(2,\frac{14}{5})$. The corners of a square satisfy $\mathbf{x}_1-\mathbf{x}_2+\mathbf{x}_3-\mathbf{x}_4=(0-4+4-0,0-0+4-4)=\mathbf{0}$, with $1-1+1-1=0$. So $\boldsymbol{\lambda}=(1,-1,1,-1)$.

(b) The positive coefficients are at $\mathbf{x}_1$ and $\mathbf{x}_3$, with ratios $\frac{1/10}{1}=\frac{1}{10}$ and $\frac{3/10}{1}=\frac{3}{10}$. So $s=\frac{1}{10}$, attained at $\mathbf{x}_1$, and the new weights are $\mathbf{t}-s\boldsymbol{\lambda}=(0,\frac{3}{10},\frac{2}{10},\frac{5}{10})$. Check: $\frac{3}{10}(4,0)+\frac{2}{10}(4,4)+\frac{5}{10}(0,4)=(\frac{12+8}{10},\frac{8+20}{10})=(2,\frac{14}{5})$, and the weights add up to 1. So $\mathbf{x}\in\operatorname{conv}\{\mathbf{x}_2,\mathbf{x}_3,\mathbf{x}_4\}$.

(c) With $-\boldsymbol{\lambda}=(-1,1,-1,1)$, the positive coefficients are at $\mathbf{x}_2$ and $\mathbf{x}_4$, with ratios $\frac{2}{10}$ and $\frac{4}{10}$. So $s=\frac{2}{10}$, attained at $\mathbf{x}_2$, and the weights become $\mathbf{t}+s\boldsymbol{\lambda}=(\frac{3}{10},0,\frac{5}{10},\frac{2}{10})$. Check: $\frac{5}{10}(4,4)+\frac{2}{10}(0,4)=(2,\frac{20+8}{10})=(2,\frac{14}{5})$. Now $\mathbf{x}_2$ is dropped, and $\mathbf{x}\in\operatorname{conv}\{\mathbf{x}_1,\mathbf{x}_3,\mathbf{x}_4\}$. The two answers are consistent: the triangle with corners $\mathbf{x}_2,\mathbf{x}_3,\mathbf{x}_4$ is the part of the square where $x+y\geq 4$, the triangle with corners $\mathbf{x}_1,\mathbf{x}_3,\mathbf{x}_4$ is the part where $y\geq x$, and $(2,\frac{14}{5})$ satisfies both. The shorter combination produced by Carathéodory's theorem is not unique; it depends on the dependence used and on its sign.

Exercise. B2: Radon partitions in the plane

Find a Radon partition, and a point common to the two hulls, for each list of four points in $\mathbb{R}^2$. Justify your answer.

(a) $(0,0),(4,0),(0,4),(1,1)$.

(b) $(0,0),(3,1),(1,3),(4,4)$.

Solution. Label the points $\mathbf{x}_1,\dots,\mathbf{x}_4$ in the order given, and follow the proof of Theorem 18.2.4.

(a) The equations $\sum_i \lambda_i \mathbf{x}_i = \mathbf{0}$, $\sum_i \lambda_i = 0$ read $4\lambda_2+\lambda_4 = 0$, $4\lambda_3+\lambda_4 = 0$, $\lambda_1+\lambda_2+\lambda_3+\lambda_4 = 0$. Taking $\lambda_4 = 4$ gives $\lambda_2 = \lambda_3 = -1$ and $\lambda_1 = -2$. So $I = \{4\}$, $J = \{1,2,3\}$, $\sigma = 4$, and

$$\mathbf{y} = \mathbf{x}_4 = (1,1) = \tfrac{2}{4}(0,0) + \tfrac{1}{4}(4,0) + \tfrac{1}{4}(0,4).$$

The fourth point lies inside the triangle of the other three.

(b) The equations read $3\lambda_2+\lambda_3+4\lambda_4 = 0$, $\lambda_2+3\lambda_3+4\lambda_4 = 0$, and $\sum_i \lambda_i = 0$. Subtracting the first two gives $\lambda_2 = \lambda_3$, and then $4\lambda_2+4\lambda_4 = 0$, so $\lambda_4 = -\lambda_2$ and $\lambda_1 = -\lambda_2-\lambda_3-\lambda_4 = -\lambda_2$. Taking $\lambda_2 = -1$ gives $\boldsymbol{\lambda} = (1,-1,-1,1)$. So $I = \{1,4\}$, $J = \{2,3\}$, $\sigma = 2$, and

$$\mathbf{y} = \tfrac{1}{2}(0,0) + \tfrac{1}{2}(4,4) = (2,2) = \tfrac{1}{2}(3,1) + \tfrac{1}{2}(1,3).$$

The two diagonals of the quadrilateral cross at $(2,2)$.

Exercise. B3: Linear functionals on a polygon

Let V be a finite-dimensional real vector space with a norm, and $x_1,\dots,x_m \in V$.

(a) Prove that $\{x_1,\dots,x_m\}$ is compact, and deduce that $P = \operatorname{conv}\{x_1,\dots,x_m\}$ is compact.

(b) Prove that every linear functional φ on V attains its maximum over P at one of the points x_i.

(c) Maximize and minimize $\varphi(x,y) = 2x - y$ on the triangle $\operatorname{conv}\{(1,0),(3,1),(0,2)\}$ of Example (Deciding membership in a triangle).

Solution.

(a) Let (y_k) be a sequence in the finite set. Some x_i occurs as y_k for infinitely many k, since there are only m candidates, and those terms form a constant subsequence, which converges to x_i, a point of the set. So the set is compact, and P is compact by Corollary 18.2.3 (a).

(b) By Corollary 18.2.3 (b), the maximum of φ on P equals its maximum on $\{x_1,\dots,x_m\}$, which is $\max_i \varphi(x_i)$, attained at an x_i with the largest value.

(c) The values at the corners are $\varphi(1,0) = 2$, $\varphi(3,1) = 5$, $\varphi(0,2) = -2$. By (b), the maximum over the triangle is 5, at $(3,1)$. Applying (b) to $-\varphi$, the minimum is -2, at $(0,2)$.

18.2.5.3 C. Going deeper

Exercise. C1: Carathéodory in the affine hull

Let V be a real vector space, $S \subseteq V$ nonempty, and suppose $\operatorname{aff} S$ has dimension d (Definition 18.1.5).

(a) Prove that every affinely independent list of points of S has at most $d+1$ entries.

(b) Hence prove that every point of $\operatorname{conv} S$ is a convex combination of at most $d+1$ points of S.

(c) Let $S = \{(t,t,t) : 0 \le t \le 1\} \cup \{(2,2,2)\} \subseteq \mathbb{R}^3$. What bound does Theorem 18.2.2 give, and what bound does (b) give?

Solution.

(a) By Proposition 18.1.4, $\operatorname{aff} S = x_0 + U$ with $\dim U = d$. If $(y_0, \dots, y_m)$ is an affinely independent list in S, then each y_i lies in $x_0 + U$, so each difference $y_i - y_0 = (y_i - x_0) - (y_0 - x_0)$ lies in U. By Lemma 18.2.1 (a) the m differences are linearly independent, and they lie in the d-dimensional space U, so $m \le d$ by Theorem 1.6.3 (a).

(b) By Theorem 18.2.2, each point of $\operatorname{conv} S$ is a convex combination of an affinely independent list of points of S, which has at most $d+1$ entries by (a).

(c) Here $n = 3$, so Theorem 18.2.2 gives 4. All of S lies on the line $\operatorname{span}(\mathbf{1})$, and S contains two distinct points, so $\operatorname{aff} S$ is that line and $d = 1$; (b) gives 2. Indeed $\operatorname{conv} S$ is the segment from $\mathbf{0}$ to $(2,2,2)$, and each of its points is a convex combination of the two endpoints.

Exercise. C2: Helly's theorem for a sequence of compact sets

Let V be a real vector space of dimension n with a norm, and let $C_1, C_2, C_3, \dots$ be a sequence of **compact** convex subsets of V such that every $n+1$ of them have a common point.

(a) Prove that a compact set K is closed.

(b) Prove that for each $m \ge 1$ there is a point $y_m \in C_1 \cap \cdots \cap C_m$.

(c) Prove that $\bigcap_{k \ge 1} C_k \neq \emptyset$.

(d) Explain where the argument fails for the closed half-lines $[k, \infty)$ of the remark after Theorem 18.2.5.

Hint for (c): all the y_m lie in C_1.

Solution.

(a) Let $x_k \to x$ with $x_k \in K$. By compactness, a subsequence converges to some $x' \in K$. It also converges to x, as a subsequence of a convergent sequence, and a limit is unique: $\|x - x'\| \le \|x - x_{k_j}\| + \|x_{k_j} - x'\| \to 0$, so $x = x'$ by (N1). Hence $x \in K$.

(b) For $m \le n$, the sets $C_1, \dots, C_m$, together with $C_{m+1}, \dots, C_{n+1}$, are $n+1$ of the sets, and a common point of all of them lies in $C_1 \cap \cdots \cap C_m$. For $m \ge n+1$, Theorem 18.2.5 applies to the convex sets $C_1, \dots, C_m$.

(c) Every y_m lies in the compact set C_1, so a subsequence y_{m_j} converges to some y. Fix k. For $m_j \ge k$, $y_{m_j} \in C_k$, and C_k is closed by (a); the terms with $m_j \ge k$ form a sequence in C_k converging to y, so $y \in C_k$. As k was arbitrary, $y \in \bigcap_k C_k$.

(d) Steps (a) and (b) survive for $C_k = [k, \infty)$ with $n = 1$: the sets are closed, and $\mathbf{y}_m = m$ lies in the first m of them. Step (c) fails, because $C_1 = [1, \infty)$ is not compact: the

sequence $\mathbf{y}_m = m$ has no convergent subsequence, and the points escape to infinity instead of accumulating at a common point.

Exercise. C3: The counts are sharp

Let V be a real vector space of dimension n, and let $(x_0, \dots, x_n)$ be affinely independent.

(a) Prove that each point of $\operatorname{conv}\{x_0, \dots, x_n\}$ has **exactly one** expression $\sum_i t_i x_i$ with $\mathbf{t} \in \Delta_{n+1}$.

(b) Deduce that $\frac{1}{n+1}\sum_i x_i$ is not a convex combination of any n of the points. (This recovers Example (The bound is attained).)

(c) Prove that $x_0, \dots, x_n$ have **no** Radon partition, so that $n+2$ in Theorem 18.2.4 cannot be lowered.

Solution.

(a) If $\sum_i t_i x_i = \sum_i t'_i x_i$ with $\mathbf{t}, \mathbf{t}' \in \Delta_{n+1}$, then $\sum_i (t_i - t'_i) x_i = 0$ and $\sum_i (t_i - t'_i) = 1 - 1 = 0$. By affine independence every $t_i - t'_i = 0$.

(b) The point has the expression with all weights $\frac{1}{n+1}$. A convex combination of n of the points is an expression of the form in (a) with the omitted point given weight 0. By the uniqueness in (a), it would have to coincide with the all-$\frac{1}{n+1}$ expression, which has no zero weight.

(c) Suppose I, J are disjoint, nonempty, and $\sum_{i\in I} a_i x_i = \sum_{j\in J} b_j x_j$ with nonnegative weights, each family adding up to 1. Then $\sum_{i\in I} a_i x_i - \sum_{j\in J} b_j x_j = 0$ is a relation among the x_i whose coefficients add up to $1 - 1 = 0$; an index lying in neither set gets coefficient 0, and no index lies in both, since $I \cap J = \emptyset$. By affine independence, all these coefficients are 0, so every $a_i = 0$, contradicting $\sum_{i\in I} a_i = 1$. Hence no Radon partition exists.

18.3 Separating Hyperplanes

Sections 01 and 02 looked at a convex set from the inside: which combinations of its points it contains, and how few points a combination needs. This section looks at it from the outside. A point that is not in a closed convex set can be cut off from it by a hyperplane, a set can be touched at its edge by a hyperplane that leaves the whole set on one side, and two disjoint convex sets can be pulled apart by one. All three facts come from a single construction, the point of the set nearest to a given point, and that construction is Chapter 11's orthogonal projection with the subspace replaced by a convex set.

Throughout, V is a finite-dimensional real inner product space, with inner product $\langle \cdot, \cdot \rangle$ (Definition 11.1.1), induced norm $\|\cdot\|$ (Definition 11.1.3), and $n = \dim V$. A sequence **converges**, $x_k \to x$, when $\|x_k - x\| \to 0$; **closed** means closed in the sense of Definition 18.1.6, so that a closed set contains the limit of every convergent sequence of its points. A set is **compact** if every sequence in it has a subsequence converging to a point of the set; this is the word of Corollary 16.2.5, which says that every closed bounded subset of V is compact. The field is $\mathbb{R}$ because the whole section is about inequalities.

18.3.1 The nearest point of a closed convex set

Chapter 11 answered a question about subspaces: given v and a finite-dimensional subspace U, which vector of U is closest to v? Theorem 11.3.5 said there is exactly one, the orthogonal projection $P_U v$, and that $v - P_U v$ is perpendicular to U. Now replace U by a set with less structure, such as a disc, a box, or a half-plane, and ask the same question. Two small experiments in $\mathbb{R}^2$ show which hypotheses the answer needs.

- The **open** disc $\{\mathbf{y} : \|\mathbf{y}\| < 1\}$ and the point $\mathbf{x} = (2, 0)$. The points $(1 - 1/k, 0)$ of the disc come closer and closer to $\mathbf{x}$, but the distance 1 is never reached, since the only candidate $(1, 0)$ is missing. There is no nearest point. The set is convex but not closed.
- The **circle** $\{\mathbf{y} : \|\mathbf{y}\| = 1\}$ and the point $\mathbf{x} = \mathbf{0}$. Every point of the circle is at distance 1, so every point is a nearest point. The set is closed but not convex.

So we ask for both properties, and then everything works.

Theorem 18.3.1. Nearest Point Theorem

Let C be a **nonempty, closed, convex** subset of V, and let $x \in V$.

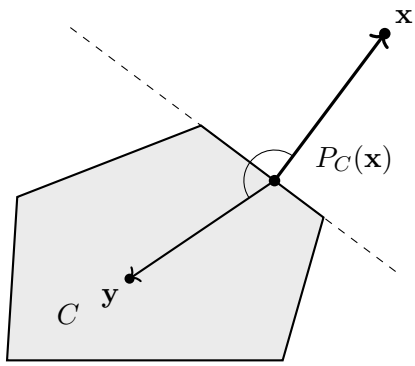
(a) There is **exactly one** point $p \in C$ with $\|x - p\| \le \|x - y\|$ for every $y \in C$.

(b) A point $p \in C$ is this nearest point **if and only if**

$$\langle x - p, y - p \rangle \le 0 \qquad \text{for every } y \in C. \tag{18.3.1}$$

The nearest point is written $P_C(x)$, and the map $P_C \colon V \to C$ is the **projection onto** C. The number $d(x, C) = \|x - P_C(x)\|$ is the **distance** from x to C.

Idea. Three jobs, three tools. ① Existence: a nearest point is a minimum of the continuous function $y \mapsto \|x - y\|$, and a minimum exists on a compact set; C itself may be unbounded, so cut it down to the part of C within a fixed distance of x. ② The inequality Eq. 18.3.1 is the first-order condition: slide from p a little way towards another point y of C, which convexity allows, and ask that the distance not decrease. ③ Read backwards, the inequality says that the angle at p between $x - p$ and $y - p$ is at least a right angle, and expanding $\|x - y\|^2$ then gives $\|x - y\|^2 \ge \|x - p\|^2 + \|y - p\|^2$. That inequality proves both that p is nearest and that nothing else is.



the angle at $P_C(\mathbf{x})$ between $\mathbf{x} - P_C(\mathbf{x})$ and $\mathbf{y} - P_C(\mathbf{x})$ is obtuse;
the dashed line through $P_C(\mathbf{x})$ is perpendicular to $\mathbf{x} - P_C(\mathbf{x})$

Proof. Existence. Since $C \neq \emptyset$, pick $y_0 \in C$ and put $R = \|x - y_0\|$. Let

$$K = \{y \in C : \|x - y\| \leq R\}.$$

Then $y_0 \in K$, and K is bounded, since $\|y\| \leq \|x\| + \|y - x\| \leq \|x\| + R$ for $y \in K$ by Corollary 11.1.6. It is closed: if $y_k \in K$ and $y_k \to y$, then $y \in C$ because C is closed, and

$$\|x - y\| \leq \|x - y_k\| + \|y_k - y\| \leq R + \|y_k - y\|$$

for every k; the right-hand side tends to R, and a nonstrict inequality survives a limit, so $\|x - y\| \leq R$. By Corollary 16.2.5, the form in which Chapter 16 carried **the compactness of closed bounded sets, fact (A3) of Chapter 16's introduction**, over to every finite-dimensional space, K is compact. The function $f(y) = \|x - y\|$ is continuous, since $|f(y) - f(y')| \leq \|y - y'\|$ by Lemma 16.1.1. By **the extreme value theorem, fact (A4) of Chapter 16's introduction**, f attains a minimum on K at some $p \in K$. Every $y \in C \setminus K$ has $f(y) > R \geq f(p)$. So $\|x - p\| \leq \|x - y\|$ for every $y \in C$.

(b), ($\Rightarrow$). Let p be a nearest point and let $y \in C$. For $0 < t \leq 1$, the point $p + t(y - p) = (1 - t)p + ty$ lies in C by Definition 18.1.1. Hence, expanding with Theorem 11.1.2 (c) over $\mathbb{R}$,

$$\|x - p\|^2 \leq \|(x - p) - t(y - p)\|^2 = \|x - p\|^2 - 2t\langle x - p, y - p\rangle + t^2\|y - p\|^2.$$

Cancel $\|x - p\|^2$ and divide by $2t > 0$: $\langle x - p, y - p\rangle \leq \frac{t}{2}\|y - p\|^2$. This holds for every $t = 1/k$, and letting $k \to \infty$ gives Eq. 18.3.1.

(b), ($\Leftarrow$), and uniqueness. Let $p \in C$ satisfy Eq. 18.3.1, and let $y \in C$. The same expansion with $t = 1$ gives

$$\begin{aligned} \|x - y\|^2 &= \|x - p\|^2 - 2\langle x - p, y - p\rangle + \|y - p\|^2 \\ &\geq \|x - p\|^2 + \|y - p\|^2. \end{aligned} \tag{$*$}$$

So $\|x - y\| \geq \|x - p\|$, and p is a nearest point. Finally, the nearest point p found in the existence step satisfies Eq. 18.3.1 by ($\Rightarrow$), so ($*$) applies to it; for $y \in C$ with $y \neq p$ it gives $\|x - y\|^2 \geq \|x - p\|^2 + \|y - p\|^2 > \|x - p\|^2$, and y is not a nearest point. This proves the theorem.

■

The inequality ($*$) is Pythagoras with the equality relaxed to an inequality. That is exactly the relation with Chapter 11, and it deserves to be said precisely.

This theorem generalizes Theorem 11.3.5. Let $C = U$ be a subspace of V. It is convex, and closed by Proposition 18.1.6 (c). For $p \in U$, the vector $y - p$ runs over **all** of U as y does, and with each $u \in U$ it also takes the value $-u$. So Eq. 18.3.1 says $\langle x - p, u\rangle \leq 0$ and $-\langle x - p, u\rangle \leq 0$, that is,

$$\langle x - p, u\rangle = 0 \qquad \text{for every } u \in U.$$

For a subspace the variational inequality becomes an equality, the statement $x - p \in U^\perp$. Then $x = p + (x - p)$ is the splitting of Theorem 11.3.2 (a), so $P_C(x) = P_U x$ by Definition 11.3.2, and ($*$) becomes the equation of Theorem 11.1.3 used in the proof of Theorem 11.3.5. The projection onto a closed convex set is the orthogonal projection with "perpendicular" weakened to "at an obtuse angle".

Example. Three projections

Find $P_C(x)$ in each case, and check Eq. 18.3.1.

(a) C is the closed unit ball $\{y : \|y\| \le 1\}$ of V, and $\|x\| > 1$.

(b) C is the box $[0,1]^3 = \{\mathbf{y} \in \mathbb{R}^3 : 0 \le y_i \le 1 \text{ for each } i\}$ with the dot product, and $\mathbf{x} = (3, -1, 2)$.

(c) C is the half-space $\{y : \langle a, y\rangle \le c\}$, where $a \neq 0$, and $\langle a, x\rangle > c$.

Solution.

(a) $P_C(x) = x/\|x\|$. With $p = x/\|x\|$ we have $\|p\| = 1$ and $x - p = (\|x\| - 1)p$, so for $y \in C$

$$\langle x - p, y - p\rangle = (\|x\| - 1)(\langle p, y\rangle - 1) \le 0,$$

because $\|x\| - 1 > 0$ and $\langle p, y\rangle \le \|p\|\|y\| \le 1$ by Theorem 11.1.5. By Theorem 18.3.1 (b), p is the nearest point, and $d(x, C) = \|x\| - 1$.

(b) Clamp each coordinate into $[0,1]$: $\mathbf{p} = (1, 0, 1)$. Then $\mathbf{x} - \mathbf{p} = (2, -1, 1)$ and, for $\mathbf{y} \in C$,

$$\langle \mathbf{x} - \mathbf{p}, \mathbf{y} - \mathbf{p}\rangle = 2(y_1 - 1) - y_2 + (y_3 - 1) \le 0,$$

since $y_1 \le 1$, $y_2 \ge 0$, and $y_3 \le 1$. So $P_C(\mathbf{x}) = (1, 0, 1)$ and $d(\mathbf{x}, C) = \sqrt{4 + 1 + 1} = \sqrt{6}$. The vector $\mathbf{a} = \mathbf{x} - \mathbf{p} = (2, -1, 1)$ also cuts $\mathbf{x}$ off: $\langle \mathbf{a}, \mathbf{x}\rangle = 9$ and $\|\mathbf{a}\|^2 = 6$, so $c = \langle \mathbf{a}, \mathbf{x}\rangle - \frac{1}{2}\|\mathbf{a}\|^2 = 6$, and the plane $2y_1 - y_2 + y_3 = 6$ strictly separates C from $\mathbf{x}$, since its left side is at most $2 + 0 + 1 = 3$ on C and equals 9 at $\mathbf{x}$. Theorem 18.3.2 below turns this recipe into a theorem.

(c) Put $\lambda = (\langle a, x\rangle - c)/\|a\|^2$, which is positive, and $p = x - \lambda a$. Then $\langle a, p\rangle = \langle a, x\rangle - \lambda\|a\|^2 = c$, so $p \in C$, and for $y \in C$

$$\langle x - p, y - p\rangle = \lambda(\langle a, y\rangle - c) \le 0.$$

So $P_C(x) = x - \lambda a$: drop the perpendicular to the boundary hyperplane, as Chapter 11 would have done. The distance is $\lambda\|a\| = (\langle a, x\rangle - c)/\|a\|$.

Two degenerate cases fix the scale. If $C = \{c\}$ is a single point, then $P_C(x) = c$ for every x, and Eq. 18.3.1 holds because $y - p = 0$. If $C = V$, then P_C is the identity. In every case $P_C(x) = x$ exactly when $x \in C$.

Quick check. Where did the proof of Theorem 18.3.1 use that C is closed, and where that it is convex? Match each use with one of the two failing experiments before the theorem.

Solution. Closedness was used in the existence step, to make K closed and hence compact; the open disc fails there, since the minimizing candidates converge to a point outside it. Convexity was used in ($\Rightarrow$), to know that $(1 - t)p + ty \in C$, and uniqueness rests on that step. The circle fails there: with $\mathbf{x} = \mathbf{0}$, $\mathbf{p} = \mathbf{e}_1$, and $\mathbf{y} = -\mathbf{e}_1$, $\langle \mathbf{x} - \mathbf{p}, \mathbf{y} - \mathbf{p}\rangle = 2 > 0$.

Warning. **The proof needs an inner product, and uniqueness can fail without one.** In $\mathbb{R}^2$ with $\|\cdot\|_\infty$, let C be the x_1-axis, which is closed and convex, and $\mathbf{x} = (0,1)$. Then $\|\mathbf{x} - (t,0)\|_\infty = \max(|t|, 1) = 1$ for every $t \in [-1,1]$, so there is a whole segment of nearest points. Existence survives, since its proof used only (N3) and compactness. Uniqueness does not, because it came from the expansion of $\|\mathbf{x} - \mathbf{y}\|^2$, which is available only for induced norms.

18.3.2 Hyperplanes and half-spaces

A hyperplane in the sense of Chapter 5 passes through 0: it is a subspace of codimension 1 (Definition 5.5.2). To separate sets we need translates of hyperplanes, and Chapter 5 §06 called a coset $p + H$ of a hyperplane H (Definition 4.4.2) an **affine hyperplane**. In an inner product space these have a concrete form. By Theorem 5.5.3, $H = \ker \varphi$ for a nonzero $\varphi \in V^*$; by Theorem 11.5.1, $\varphi = \langle \cdot, a \rangle$ for a nonzero $a \in V$. Hence

$$p + H = \{y \in V : \langle a, y \rangle = c\}, \qquad c = \langle a, p \rangle,$$

and conversely every such set with $a \neq 0$ is an affine hyperplane, with a as a **normal vector**. It cuts V into the two **closed half-spaces** $\{\langle a, \cdot \rangle \leq c\}$ and $\{\langle a, \cdot \rangle \geq c\}$. Both are convex, because $\langle a, \cdot \rangle$ is linear, and closed by Proposition 18.1.6 (c).

Definition 18.3.1. Separation by a Hyperplane

Let $A, B \subseteq V$ be nonempty. A **nonzero** vector $a \in V$ and a number $c \in \mathbb{R}$ **separate** A and B if

$$\langle a, x \rangle \leq c \leq \langle a, y \rangle \qquad \text{for all } x \in A \text{ and all } y \in B.$$

They **strictly separate** A and B if in addition there is a real $\varepsilon > 0$ with $\langle a, x \rangle \leq c - \varepsilon$ and $\langle a, y \rangle \geq c + \varepsilon$ for all $x \in A$, $y \in B$.

In words: A lies in one closed half-space of $\{\langle a, \cdot \rangle = c\}$ and B in the other, and strict separation puts a slab of width $2\varepsilon/\|a\|$ between them. Replacing (a, c) by $(-a, -c)$ swaps the roles of A and B. With $a = 0$ and $c = 0$ any two sets would be "separated".

18.3.3 Cutting off a point

The first consequence of Theorem 18.3.1 is the one the chapter uses most. If x is not in C, the vector from $P_C(x)$ to x is a normal vector for a hyperplane that cuts x off.

Theorem 18.3.2. Separating a Point from a Closed Convex Set

Let $C \subseteq V$ be nonempty, closed, and convex, and let $x \notin C$. Put $a = x - P_C(x)$. Then $a \neq 0$ and

$$\langle a, y \rangle \leq \langle a, x \rangle - \|a\|^2 \qquad \text{for every } y \in C.$$

In particular a and $c = \langle a, x \rangle - \frac{1}{2}\|a\|^2$ strictly separate C and $\{x\}$.

Proof. Write $p = P_C(x)$. Since $p \in C$ and $x \notin C$, $a = x - p \neq 0$. For $y \in C$, Eq. 18.3.1 gives

$\langle a, y - p\rangle \le 0$, hence

$$\langle a, y\rangle \le \langle a, p\rangle = \langle a, x\rangle - \langle a, x - p\rangle = \langle a, x\rangle - \|a\|^2.$$

With $\varepsilon = \frac{1}{2}\|a\|^2 > 0$ this reads $\langle a, y\rangle \le c - \varepsilon$, while $\langle a, x\rangle = c + \varepsilon$. This proves the theorem.

■

In the picture after Theorem 18.3.1, the hyperplane of the theorem is the line parallel to the dashed one, halfway between $P_C(x)$ and x.

The theorem turns into a description of a closed convex set from the outside. Sections 01 and 02 built convex sets up from points; this builds them down from half-spaces.

Corollary 18.3.3. A Closed Convex Set Is an Intersection of Half-Spaces

Every nonempty closed convex subset C of V is the intersection of all the closed half-spaces $\{y : \langle a, y\rangle \le c\}$, with $a \neq 0$, that contain C.

Proof. Each of these half-spaces contains C, so C lies in their intersection. Conversely, let $x \notin C$. By Theorem 18.3.2, with $a = x - P_C(x) \neq 0$, the half-space $\{y : \langle a, y\rangle \le \langle a, x\rangle - \|a\|^2\}$ contains C but not x. So x is not in the intersection. (If $C = V$, the family of half-spaces is empty and its intersection is V by convention.)

■

The corollary has a converse that is immediate: an intersection of closed half-spaces is closed and convex, because each half-space is and both properties pass to intersections. So the closed convex sets are **exactly** the intersections of closed half-spaces. §04 turns this into a formula, the support function, and §09 studies the case of finitely many half-spaces.

18.3.4 Supporting hyperplanes

A point on the edge of C cannot be cut off, because it belongs to C. The best one can hope for is a hyperplane through the point with all of C on one side, touching C without cutting it. To say "on the edge" we need two more words.

Definition 18.3.2. Interior Point, Closure, Boundary Point

Let $S \subseteq V$. A point $x \in V$ is an **interior point** of S if there is a real $r > 0$ such that **every** y with $\|y - x\| < r$ lies in S; the set of interior points is written $\operatorname{int}(S)$. The **closure** $\overline{S}$ of S is the set of limits of all convergent sequences of points of S. A **boundary point** of S is a point of $\overline{S}$ that is not an interior point of S.

Constant sequences show $S \subseteq \overline{S}$, and by Definition 18.1.6, S is closed exactly when $\overline{S} = S$. By Theorem 16.2.2 neither notion depends on the norm used: a ball of radius r for one norm contains a ball of positive radius for any other. For the closed unit disc of $\mathbb{R}^2$, the interior points are those with $\|\mathbf{x}\| < 1$, and the boundary points form the unit circle. A segment in $\mathbb{R}^2$ has no interior points at all, so every one of its points is a boundary point; that is the case the next lemma isolates.

Lemma 18.3.4. A Convex Set Is Solid or Flat

Let $V \neq \{0\}$ and let $C \subseteq V$ be nonempty and convex. Then **either** C has an interior point, **or** C lies in an affine hyperplane: there are $a \neq 0$ and $c \in \mathbb{R}$ with $\langle a, y\rangle = c$ for every $y \in C$.

Idea. Fix $c_0 \in C$ and look at the directions $y - c_0$ available inside C. If they span a proper subspace, a normal vector to that subspace is the a we want. If they span V, then C contains $n+1$ points in general position, hence the solid simplex they span, and the center of a solid simplex is an interior point: a point of the simplex is $\sum_i t_i v_i$ with $t_i \geq 0$ summing to 1, the t_i are determined by the point and depend continuously on it, and at the center they are all equal and positive, so every small enough move keeps them positive.

Proof. Fix $c_0 \in C$ and let $W = \operatorname{span}\{y - c_0 : y \in C\}$.

Case 1: $W \neq V$. Then $\dim W < n$, so $\dim W^{\perp} = n - \dim W \geq 1$ by Theorem 11.3.2 (c), and there is $a \neq 0$ in $W^{\perp}$. For $y \in C$, $\langle a, y - c_0\rangle = 0$, that is, $\langle a, y\rangle = \langle a, c_0\rangle$. Take $c = \langle a, c_0\rangle$.

Case 2: $W = V$. Choose a linearly independent list $(y_1 - c_0, \dots, y_m - c_0)$ with every $y_i \in C$ and m as large as possible; $m \leq n$ by Theorem 1.6.3. Every vector $y - c_0$ with $y \in C$ lies in the span of this list, since otherwise Lemma 1.5.3 would lengthen it. So the span contains every generator of W, hence equals V, and the list is a basis; thus $m = n$. Let $\varphi_i(v)$ be the i-th coordinate of v in this basis, a linear functional by Theorem 1.5.2, and write $\varphi_i = \langle \cdot, w_i\rangle$ by Theorem 11.5.1. Put $\mu_i(v) = \varphi_i(v - c_0)$. Then for every $v \in V$,

$$v = \Big(1 - \sum_{i=1}^{n} \mu_i(v)\Big) c_0 + \sum_{i=1}^{n} \mu_i(v)\, y_i,$$

and when $\mu_i(v) \geq 0$ for every i and $\sum_i \mu_i(v) \leq 1$, this is a convex combination of points of C, hence a point of C: by Theorem 18.1.3 it lies in $\operatorname{conv} C$, and $\operatorname{conv} C = C$ by Definition 18.1.4, since C is itself a convex set containing C.

Let $z = c_0 + \frac{1}{n+1}\sum_i (y_i - c_0)$, so that $\mu_i(z) = \frac{1}{n+1}$ for every i. Each $w_i \neq 0$, because $\varphi_i(y_i - c_0) = 1$; put $L = \max_i \|w_i\| > 0$ and $r = 1/(n(n+1)L)$. If $\|v - z\| < r$, then by Theorem 11.1.5

$$|\mu_i(v) - \mu_i(z)| = |\langle v - z, w_i\rangle| \leq L\|v - z\| < \frac{1}{n(n+1)},$$

so $\mu_i(v) > \frac{1}{n+1} - \frac{1}{n(n+1)} \geq 0$ and $\sum_i \mu_i(v) < \frac{n}{n+1} + \frac{1}{n+1} = 1$. Hence $v \in C$, and z is an interior point of C.

■

The second ingredient says that from an interior point one can see the whole closure.

Lemma 18.3.5. Closure and Interior of a Convex Set

Let $C \subseteq V$ be convex.

(a) $\overline{C}$ is closed and convex.

(b) If $z \in \operatorname{int}(C)$, $w \in \overline{C}$, and $0 < t \leq 1$, then $(1-t)w + tz \in \operatorname{int}(C)$.

Idea. Part (a) pushes limits through: a limit of limits is a limit, and a convex combination of limits is the limit of convex combinations. For (b), picture a small ball around z inside C, and the cone from w over it. The cone contains a smaller ball around $(1-t)w + tz$; since w is only a limit, replace it by a nearby point w' of C, which shrinks the ball by half.

Proof.

(a) Let $x_k \in \overline{C}$ with $x_k \to x$. Each x_k is a limit of points of C, so there is $c_k \in C$ with $\|c_k - x_k\| < 1/k$. Then $\|c_k - x\| \leq 1/k + \|x_k - x\| \to 0$, so $x \in \overline{C}$. For convexity, let $x = \lim c_k$ and $y = \lim d_k$ with $c_k, d_k \in C$, and $0 \leq s \leq 1$. The points $sc_k + (1-s)d_k$ lie in C and converge to $sx + (1-s)y$, by Corollary 11.1.6, so that point lies in $\overline{C}$.

(b) Take $\rho > 0$ with every point within distance ρ of z in C, and put $p = (1-t)w + tz$. If $t = 1$ then $p = z$ and there is nothing to prove, so let $t < 1$. Choose $w' \in C$ with $(1-t)\|w - w'\| < t\rho/2$, possible because w is a limit of points of C. Let $\|q - p\| < t\rho/2$, and put $z' = (q - (1-t)w')/t$, so that $q = (1-t)w' + tz'$. Since $p = (1-t)w + tz$,

$$z' - z = \frac{(q-p) + (1-t)(w-w')}{t}, \qquad \|z' - z\| < \frac{t\rho/2 + t\rho/2}{t} = \rho.$$

Hence $z' \in C$, and $q \in C$ by Definition 18.1.1. So every point within $t\rho/2$ of p lies in C, and $p \in \operatorname{int}(C)$.

■

Now the theorem. It asks only that x_0 not be an interior point, which covers the boundary points of C and also the points outside $\overline{C}$.

Theorem 18.3.6. Supporting Hyperplane Theorem

Let $V \neq \{0\}$, let $C \subseteq V$ be convex, and let $x_0 \in V$ be a point that is **not** an interior point of C. Then there is a **nonzero** $a \in V$ with

$$\langle a, y \rangle \leq \langle a, x_0 \rangle \qquad \text{for every } y \in C.$$

If moreover $x_0 \in C$, the affine hyperplane $\{y : \langle a, y \rangle = \langle a, x_0 \rangle\}$ passes through x_0 and has C in one of its closed half-spaces; it is called a **supporting hyperplane** of C at x_0.

Idea. Approach x_0 from outside $\overline{C}$ and cut off each approximating point with Theorem 18.3.2. The normal vectors, scaled to length 1, live on the compact unit sphere, so a subsequence converges, and in the limit the strict inequalities become the nonstrict one we want. The only thing that could go wrong is that there are no points outside $\overline{C}$ near x_0. Lemma 18.3.4 and Lemma 18.3.5 rule that out: if C is flat, its own hyperplane serves; if C is solid, pushing x_0 directly away from an interior point leaves $\overline{C}$ at once.

Proof. If $C = \emptyset$, any nonzero a works, and one exists because $V \neq \{0\}$. So let $C \neq \emptyset$.

Case 1: C has no interior point. By Lemma 18.3.4 there are $a \neq 0$ and c with $\langle a, y\rangle = c$ for all $y \in C$. If $\langle a, x_0\rangle \geq c$, then a works. Otherwise $-a$ works, since $\langle -a, y\rangle = -c < -\langle a, x_0\rangle$.

Case 2: C has an interior point z. For $k \geq 1$ let $x_k = x_0 + \frac{1}{k}(x_0 - z)$. We claim $x_k \notin \overline{C}$. Indeed, with $t = \frac{1}{k+1}$,

$$(1-t)x_k + tz = \tfrac{k}{k+1}x_0 + \tfrac{1}{k+1}(x_0 - z) + \tfrac{1}{k+1}z = x_0,$$

so if x_k were in $\overline{C}$, then x_0 would be an interior point of C by Lemma 18.3.5 (b), contrary to the hypothesis. By Lemma 18.3.5 (a), $\overline{C}$ is nonempty, closed, and convex, so Theorem 18.3.2, applied to $\overline{C}$ and x_k, gives $b_k \neq 0$ with $\langle b_k, y\rangle \leq \langle b_k, x_k\rangle$ for all $y \in \overline{C}$, in particular for all $y \in C$. Put $a_k = b_k/\|b_k\|$, a unit vector satisfying the same inequality.

The unit sphere of V is compact by Corollary 16.2.5, which rests on **fact (A3) of Chapter 16's introduction**. So a subsequence (a_{k_j}) converges to some a with $\|a\| = 1$; in particular $a \neq 0$. Fix $y \in C$. Then $\langle a_{k_j}, y - x_{k_j}\rangle \leq 0$ for every j, and $x_k \to x_0$ because $\|x_k - x_0\| = \|x_0 - z\|/k$. By Theorem 11.1.5,

$$\begin{aligned}\left|\langle a_{k_j}, y - x_{k_j}\rangle - \langle a, y - x_0\rangle\right| &\leq \left|\langle a_{k_j} - a, y - x_{k_j}\rangle\right| + \left|\langle a, x_0 - x_{k_j}\rangle\right| \\ &\leq \|a_{k_j} - a\|(\|y - x_0\| + \|x_0 - z\|) + \|x_0 - x_{k_j}\|,\end{aligned}$$

which tends to 0. A nonstrict inequality survives the limit, so $\langle a, y - x_0\rangle \leq 0$. This proves the theorem.

■

A supporting hyperplane need not be unique. At the corner $(1, 1)$ of the square $[-1, 1]^2$, every $\mathbf{a} = (a_1, a_2)$ with $a_1, a_2 \geq 0$, not both zero, works, since $a_1 y_1 + a_2 y_2 \leq a_1 + a_2$ on the square. At a point of the unit circle, the boundary of the disc, only the positive multiples of the point itself work. Corners admit many supporting hyperplanes and smooth points one; §08 builds its induction on dimension on these hyperplanes, and §04 uses them to prove that a norm is recovered from its dual.

18.3.5 Separating two convex sets

To separate two sets, separate one point from one set: the point 0 from the set of differences.

Theorem 18.3.7. Separating Hyperplane Theorem

Let $C_1, C_2 \subseteq V$ be **nonempty, disjoint**, and convex.

(a) If C_1 is **compact** and C_2 is **closed**, then C_1 and C_2 are strictly separated: there are $a \neq 0$ and c with $\langle a, x\rangle < c < \langle a, y\rangle$ for all $x \in C_1$, $y \in C_2$, and indeed with a margin $\varepsilon > 0$ as in Definition 18.3.1.

(b) In general, C_1 and C_2 are separated: there are $a \neq 0$ and c with $\langle a, x\rangle \leq c \leq \langle a, y\rangle$ for all $x \in C_1$, $y \in C_2$.

Idea. The difference set $D = \{y - x : x \in C_1, y \in C_2\}$ is convex and misses 0, and a normal a with D on its positive side separates C_1 from C_2. For (a), compactness of C_1 makes D closed, and Theorem 18.3.2 cuts 0 off; for (b), 0 is not an interior point of D, and Theorem 18.3.6 applies.

Proof. Let $D = \{y - x : x \in C_1, y \in C_2\}$. It is convex: for $0 \le t \le 1$,

$$t(y - x) + (1 - t)(y' - x') = (ty + (1 - t)y') - (tx + (1 - t)x'),$$

a difference of a point of C_2 and a point of C_1. And $0 \notin D$, since $C_1 \cap C_2 = \emptyset$.

(a) We show D is closed. Let $y_k - x_k \to d$ with $x_k \in C_1$, $y_k \in C_2$. Since C_1 is compact, a subsequence x_{k_j} converges to some $x \in C_1$. Then $y_{k_j} = (y_{k_j} - x_{k_j}) + x_{k_j} \to d + x$, since a subsequence of a convergent sequence has the same limit, part of the elementary algebra of limits that Chapter 16's introduction imports; so $d + x \in C_2$ because C_2 is closed, and $d = (d + x) - x \in D$. Now Theorem 18.3.2, applied to D and the point $0 \notin D$, gives $b \neq 0$ with $\langle b, d\rangle \le -\|b\|^2$ for all $d \in D$. With $a = -b$ this says

$$\langle a, x\rangle + \|a\|^2 \le \langle a, y\rangle \qquad \text{for all } x \in C_1,\ y \in C_2.$$

Fix $y \in C_2$. The numbers $\langle a, x\rangle$, $x \in C_1$, are bounded above by $\langle a, y\rangle - \|a\|^2$, so by **the completeness of $\mathbb{R}$, fact (A1) of Chapter 16's introduction**, they have a least upper bound s, and $s + \|a\|^2 \le \langle a, y\rangle$ for every $y \in C_2$. Take $\varepsilon = \frac{1}{2}\|a\|^2$ and $c = s + \varepsilon$. Then $\langle a, x\rangle \le s = c - \varepsilon$ and $\langle a, y\rangle \ge c + \varepsilon$.

(b) If $V = \{0\}$, two nonempty subsets both contain 0 and cannot be disjoint, so $V \neq \{0\}$. The point 0 is not in D, so it is not an interior point of D. By Theorem 18.3.6 there is $b \neq 0$ with $\langle b, d\rangle \le \langle b, 0\rangle = 0$ for all $d \in D$. With $a = -b$, this says $\langle a, x\rangle \le \langle a, y\rangle$ for all $x \in C_1$, $y \in C_2$. As in (a), the least upper bound c of the numbers $\langle a, x\rangle$, $x \in C_1$, exists by (A1) and satisfies $\langle a, x\rangle \le c \le \langle a, y\rangle$. This proves the theorem. ∎

18.3.6 Where strict separation fails

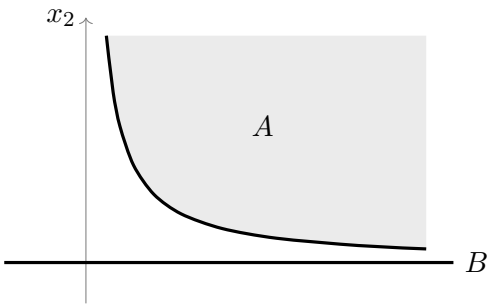
Part (a) asked one of the two sets to be compact. That hypothesis cannot be weakened to "closed", and the standard witness is a hyperbola and its asymptote.

Example. A hyperbola and its asymptote

In $\mathbb{R}^2$ with the dot product, let

$$A = \{\mathbf{x} : x_1 > 0,\ x_2 > 0,\ x_1 x_2 \ge 1\}, \qquad B = \{\mathbf{x} : x_2 = 0\}.$$

Show that A and B are nonempty, disjoint, closed, and convex, that the x_1-axis is the **only** hyperplane separating them, and that they are **not** strictly separated.



A and B are closed, convex and disjoint,
but the gap between them shrinks to zero

Solution. Nonempty and disjoint. $(1,1) \in A$ and $\mathbf{0} \in B$; every point of A has $x_2 > 0$, and every point of B has $x_2 = 0$.

B is closed and convex. It is a subspace, so it is convex, and closed by Proposition 18.1.6 (c).

A is convex. Let $\mathbf{x}, \mathbf{z} \in A$ and $0 \le t \le 1$, and let $\mathbf{w} = t\mathbf{x} + (1-t)\mathbf{z}$. Its coordinates are positive. Expanding,

$$w_1 w_2 = t^2 x_1 x_2 + (1-t)^2 z_1 z_2 + t(1-t)(x_1 z_2 + x_2 z_1).$$

By $\left(\sqrt{x_1 z_2} - \sqrt{x_2 z_1}\right)^2 \ge 0$, we get $x_1 z_2 + x_2 z_1 \ge 2\sqrt{x_1 x_2 z_1 z_2} \ge 2$, using $x_1 x_2 \ge 1$ and $z_1 z_2 \ge 1$. Hence $w_1 w_2 \ge t^2 + (1-t)^2 + 2t(1-t) = 1$, and $\mathbf{w} \in A$.

A is closed. If $\mathbf{x}_k \to \mathbf{x}$ with $\mathbf{x}_k \in A$, then the coordinates converge, so $x_1, x_2 \ge 0$, and $x_{k,1} x_{k,2} \to x_1 x_2$ by the algebra of limits, so $x_1 x_2 \ge 1$. A product that is at least 1 has no zero factor, so $x_1, x_2 > 0$, and $\mathbf{x} \in A$.

The only separating hyperplane. Let $\mathbf{a} \neq \mathbf{0}$ and c separate A and B, in either order. On B, $\langle \mathbf{a}, (t,0) \rangle = a_1 t$ for all real t, and these values lie on one side of c; that forces $a_1 = 0$. So $a_2 \neq 0$, and after replacing $(\mathbf{a}, c)$ by $(-\mathbf{a}, -c)$ we may take $a_2 > 0$. On B the value is 0. On A the values are $a_2 x_2$, and since $(1/s, s) \in A$ for every $s > 0$, they fill the interval $(0, \infty)$. These are unbounded above, so A is on the side $\ge c$ and B on the side $\le c$: $0 \le c \le a_2 s$ for every $s > 0$, which forces $c = 0$. The hyperplane $\{a_2 x_2 = 0\}$ is the x_1-axis, and it separates, since $a_2 x_2 > 0$ on A and $= 0$ on B.

No strict separation. Strict separation would need the only separating hyperplane to miss both sets, and B **is** that hyperplane.

The points $(t, 1/t) \in A$ and $(t, 0) \in B$ are $1/t$ apart, and Exercise (C2: Strict separation and distance) shows that this vanishing distance is exactly what rules out strict separation.

Warning. **Disjoint closed convex sets need not be strictly separated.** The hyperbola region and its asymptote above are the example to remember: both closed, both convex, disjoint, and every hyperplane that separates them contains one of them. "Disjoint" gives only the weak separation of Theorem 18.3.7 (b). Strict separation needs more, such as compactness of one set in (a).

18.3.7 Extending linear functionals

Separation has an analytic twin. Instead of a set and a point, take a linear functional defined on a subspace and bounded there by a convex function; can it be extended to the whole space without breaking the bound? The bounds allowed are the following.

Definition 18.3.3. Sublinear Functional

Let V be a real vector space. A function $p\colon V \to \mathbb{R}$ is **sublinear** if, for **all** $x, y \in V$ and **all** real $t \geq 0$,

$$p(x+y) \leq p(x) + p(y) \qquad \text{and} \qquad p(tx) = t\,p(x).$$

Every norm is sublinear: (N3) is the first condition, and (N2) with $t \geq 0$ the second. Every linear functional is sublinear, with equality in the first condition. So is $p(\mathbf{x}) = \max(x_1, x_2)$ on $\mathbb{R}^2$, which is neither a norm nor linear: $p(-\mathbf{e}_1) = 0$. Taking $t = 0$ gives $p(0) = 0$. A nonexample by minimal change is $p(x) = \|x\|^2$: the first condition fails at $x = y \neq 0$ (it would give $4\|x\|^2 \leq 2\|x\|^2$), and so does the second, since $p(tx) = t^2 p(x)$.

It turns out that sublinearity is exactly what makes one-step extensions possible.

Theorem 18.3.8. Hahn–Banach Theorem, Finite-Dimensional

Let V be a **finite-dimensional** real vector space, let p be a sublinear functional on V, let U be a subspace of V, and let $f \in U^*$ satisfy $f(u) \leq p(u)$ for every $u \in U$. Then there is $F \in V^*$ with

$$F(u) = f(u) \text{ for every } u \in U \qquad \text{and} \qquad F(v) \leq p(v) \text{ for every } v \in V.$$

Idea. Add one dimension at a time. On $U + \operatorname{span}(v)$, a linear extension is fixed by one number $c = F(v)$. The requirement $F \leq p$ at the vectors $u + tv$ with $t > 0$ gives upper bounds on c, and at those with $t < 0$ it gives lower bounds. Sublinearity of p, together with linearity of f, is exactly what makes every lower bound at most every upper bound, so some c fits. Then induct on $\dim V - \dim U$. No inner product is needed.

Proof. We induct on $k = \dim V - \dim U$. If $k = 0$, then $U = V$ and $F = f$ works.

Let $k \geq 1$ and assume the theorem whenever the difference of dimensions is $k - 1$. Pick $v \in V \setminus U$ and let $W = U + \operatorname{span}(v)$. Each $w \in W$ can be written as $u + tv$ with $u \in U$ and $t \in \mathbb{R}$, and only in one way: if $u + tv = u' + t'v$ with $t \neq t'$, then $v = (u' - u)/(t - t') \in U$, which is false; so $t = t'$ and $u = u'$. Appending v to a basis of U gives a basis of W by Lemma 1.5.3, so $\dim V - \dim W = k - 1$. For $c \in \mathbb{R}$, the formula $g(u + tv) = f(u) + tc$ therefore defines a linear functional on W that extends f.

For $u, u' \in U$, linearity of f, the hypothesis, and sublinearity of p give

$$f(u) + f(u') = f(u + u') \leq p(u + u') \leq p(u - v) + p(u' + v),$$

that is, $f(u) - p(u - v) \leq p(u' + v) - f(u')$. So every number in $L = \{f(u) - p(u - v) : u \in U\}$ is at most every number in $R = \{p(u' + v) - f(u') : u' \in U\}$. By **fact (A1) of Chapter 16's introduction**, the nonempty set L, bounded above by any element of R, has a least

upper bound c, and $c \le r$ for every $r \in R$. With this c, we check $g(u+tv) \le p(u+tv)$ in three cases.

- $t = 0$: this is the hypothesis $f(u) \le p(u)$.
- $t > 0$: $c \le p(u/t + v) - f(u/t)$, the element of R at $u' = u/t$. Multiply by t and use $p(t\,\cdot) = t\,p(\cdot)$: $tc \le p(u+tv) - f(u)$.
- $t = -s < 0$: $c \ge f(u/s) - p(u/s - v)$, since this is an element of L. Multiply by s: $sc \ge f(u) - p(u - sv)$, that is, $f(u) - sc \le p(u+tv)$.

So $g \le p$ on W. By the induction hypothesis applied to W and g, there is $F \in V^*$ extending g, hence f, with $F \le p$ on V. This proves the theorem.

■

Beyond finite dimension. The induction above stops after $\dim V - \dim U$ steps. In an infinite-dimensional space there is no last step, and the one-step extension has to be combined with Zorn's lemma (Theorem 1.8.1), exactly as Chapter 1 §08 combined "append one vector" with Zorn's lemma to prove Theorem 1.8.3. The resulting theorem, the Hahn–Banach theorem of functional analysis, is outside the scope of this book, and nothing here depends on it.

The special case that matters most is a norm.

Corollary 18.3.9. Norm-Preserving Extension

Let V be a finite-dimensional real vector space with a norm $\|\cdot\|$, let U be a subspace, and let $f \in U^*$ satisfy $|f(u)| \le \|u\|$ for every $u \in U$. Then there is $F \in V^*$ extending f with $|F(v)| \le \|v\|$ for every $v \in V$.

Proof. A norm is sublinear, so Theorem 18.3.8 with $p = \|\cdot\|$ gives F extending f with $F(v) \le \|v\|$ for all v. Applying this to $-v$ gives $-F(v) = F(-v) \le \|-v\| = \|v\|$ by (N2). Hence $|F(v)| \le \|v\|$.

■

In particular, for any $x \neq 0$, taking $U = \operatorname{span}(x)$ and $f(tx) = t\|x\|$ produces a functional F with $F(x) = \|x\|$ and $|F| \le \|\cdot\|$ everywhere: a linear functional that "sees" the full length of x without ever exceeding the norm. §04 meets this functional again, as the vector at which a dual norm is attained.

Example. Many extensions

On $\mathbb{R}^2$ with $\|\cdot\|_\infty$, let $U = \operatorname{span}((1,1))$ and $f(s(1,1)) = s$. Find every F as in Corollary 18.3.9, and compare with the interval produced by the proof of Theorem 18.3.8 for $\mathbf{v} = \mathbf{e}_1$.

Solution. First, $|f(s(1,1))| = |s| = \|s(1,1)\|_\infty$, so the hypothesis holds. An extension has $F(\mathbf{e}_1) = c$ for some c, and then $F(\mathbf{e}_2) = F\big((1,1)\big) - c = 1 - c$, so $F(\mathbf{x}) = cx_1 + (1-c)x_2$.

If $0 \le c \le 1$, then $|F(\mathbf{x})| \le c|x_1| + (1-c)|x_2| \le \|\mathbf{x}\|_\infty$. If $c > 1$, then $\mathbf{x} = (1,-1)$ gives $F(\mathbf{x}) = 2c - 1 > 1 = \|\mathbf{x}\|_\infty$. If $c < 0$, then $\mathbf{x} = (-1,1)$ gives $F(\mathbf{x}) = 1 - 2c > 1$. So the extensions are exactly $F(\mathbf{x}) = cx_1 + (1-c)x_2$ with $c \in [0,1]$: infinitely many.

In the proof, with $p = \|\cdot\|_\infty$, $\mathbf{u} = s(1,1)$ and $\mathbf{v} = \mathbf{e}_1$, the elements of L are $s - \max(|s-1|, |s|)$, which is 0 for $s \geq \frac{1}{2}$ and $2s - 1 < 0$ for $s < \frac{1}{2}$; so $\sup L = 0$. The elements of R are $\max(|s+1|, |s|) - s$, which is 1 for $s \geq -\frac{1}{2}$ and $-2s > 1$ for $s < -\frac{1}{2}$; so $\inf R = 1$. Every c between $\sup L$ and $\inf R$ works, and that is the interval $[0, 1]$ found directly. The proof picked its left end.

The corners of the square B_∞ are why the extension is not unique: U meets the unit sphere at the corner $(1, 1)$, and every supporting line of the square at that corner gives an extension.

18.3.8 Exercises

18.3.8.1 A. Check your understanding

Exercise. A1

(a) State the characterization Eq. 18.3.1 of the nearest point, and say what it becomes when C is a subspace.

(b) Which two facts of Chapter 16's introduction does the existence of $P_C(x)$ use, and to which set are they applied?

(c) Determine whether the following statement is correct, and justify your answer: every nonempty closed subset of $\mathbb{R}^2$ has a unique point nearest to $\mathbf{0}$.

(d) Determine whether the following statement is correct, and justify your answer: two disjoint nonempty closed convex subsets of $\mathbb{R}^2$ can always be strictly separated.

Solution.

(a) A point $p \in C$ is $P_C(x)$ exactly when $\langle x - p, y - p\rangle \leq 0$ for every $y \in C$ (Theorem 18.3.1 (b)). For a subspace U it becomes the equality $\langle x - p, u\rangle = 0$ for every $u \in U$, so $P_C(x) = P_U x$ and Theorem 11.3.5 is recovered.

(b) The compactness of closed bounded sets (A3), through Corollary 16.2.5, and the extreme value theorem (A4). They are applied to $K = \{y \in C : \|x - y\| \leq R\}$, the part of C within the distance R of x attained at one chosen point of C, and to the function $y \mapsto \|x - y\|$.

(c) Incorrect. The unit circle is closed and nonempty, and every one of its points is at distance 1 from $\mathbf{0}$. Uniqueness needs convexity.

(d) Incorrect. The region A above the hyperbola $x_1 x_2 = 1$ in the positive quadrant and its asymptote B, the x_1-axis, are disjoint, closed, and convex, but the only hyperplane separating them is B itself (Example (A hyperbola and its asymptote)).

18.3.8.2 B. Practice

Exercise. B1: Projection onto a half-plane

In $\mathbb{R}^2$ with the dot product, let $C = \{\mathbf{y} : y_1 + y_2 \leq 1\}$ and $\mathbf{x} = (2, 3)$. Find $P_C(\mathbf{x})$ and $d(\mathbf{x}, C)$, and verify Eq. 18.3.1.

Solution. Here $\mathbf{a} = (1, 1)$, $c = 1$, and $\langle \mathbf{a}, \mathbf{x} \rangle = 5 > 1$. By Example (Three projections) (c), $\lambda = (5 - 1)/2 = 2$ and $P_C(\mathbf{x}) = (2, 3) - 2(1, 1) = (0, 1)$. Check: $0 + 1 = 1$, so $\mathbf{p} = (0, 1) \in C$, and $\mathbf{x} - \mathbf{p} = (2, 2)$, so $\langle \mathbf{x} - \mathbf{p}, \mathbf{y} - \mathbf{p} \rangle = 2y_1 + 2(y_2 - 1) = 2(y_1 + y_2 - 1) \leq 0$ for $\mathbf{y} \in C$. By Theorem 18.3.1 (b) this is the nearest point, and $d(\mathbf{x}, C) = \|(2, 2)\| = 2\sqrt{2}$.

Exercise. B2: Cutting off a point from a diamond

Let $C = B_1 = \{\mathbf{y} \in \mathbb{R}^2 : |y_1| + |y_2| \leq 1\}$, with the dot product, and $\mathbf{x} = (2, 2)$.

(a) Show that $P_C(\mathbf{x}) = (\frac{1}{2}, \frac{1}{2})$.

(b) Hence find the strictly separating hyperplane of Theorem 18.3.2, and check it directly.

(c) Find every supporting hyperplane of C at $(\frac{1}{2}, \frac{1}{2})$.

Solution.

(a) Let $\mathbf{p} = (\frac{1}{2}, \frac{1}{2})$, which lies in C. Then $\mathbf{x} - \mathbf{p} = (\frac{3}{2}, \frac{3}{2})$ and, for $\mathbf{y} \in C$,

$$\langle \mathbf{x} - \mathbf{p}, \mathbf{y} - \mathbf{p} \rangle = \tfrac{3}{2}(y_1 + y_2 - 1) \leq 0,$$

because $y_1 + y_2 \leq |y_1| + |y_2| \leq 1$. By Theorem 18.3.1 (b), $P_C(\mathbf{x}) = \mathbf{p}$.

(b) With $\mathbf{a} = (\frac{3}{2}, \frac{3}{2})$, we have $\|\mathbf{a}\|^2 = \frac{9}{2}$ and $\langle \mathbf{a}, \mathbf{x} \rangle = 6$, so $c = 6 - \frac{9}{4} = \frac{15}{4}$. The hyperplane $\frac{3}{2}(y_1 + y_2) = \frac{15}{4}$ is the line $y_1 + y_2 = \frac{5}{2}$. Directly: on C, $y_1 + y_2 \leq 1 < \frac{5}{2}$, while $x_1 + x_2 = 4 > \frac{5}{2}$.

(c) Let $\mathbf{a} \neq \mathbf{0}$ satisfy $\langle \mathbf{a}, \mathbf{y} \rangle \leq \langle \mathbf{a}, \mathbf{p} \rangle = \frac{1}{2}(a_1 + a_2)$ for all $\mathbf{y} \in C$. Taking $\mathbf{y} = \mathbf{e}_1$ and $\mathbf{y} = \mathbf{e}_2$ gives $a_1 \leq \frac{1}{2}(a_1 + a_2)$ and $a_2 \leq \frac{1}{2}(a_1 + a_2)$, that is, $a_1 \leq a_2$ and $a_2 \leq a_1$; so $a_1 = a_2$. Taking $\mathbf{y} = -\mathbf{e}_1$ gives $-a_1 \leq a_1$, so $a_1 \geq 0$, and $a_1 \neq 0$ since $\mathbf{a} \neq \mathbf{0}$. Conversely $\mathbf{a} = (a, a)$ with $a > 0$ works, by the inequality in (a). So the only supporting hyperplane at $\mathbf{p}$ is the line $y_1 + y_2 = 1$ through the edge containing $\mathbf{p}$.

Exercise. B3: A unique extension

On $\mathbb{R}^2$ with $\|\cdot\|_1$, let $U = \operatorname{span}((1, 2))$ and $f(t(1, 2)) = 3t$. Show that f satisfies the hypothesis of Corollary 18.3.9, and find every extension F that the corollary allows.

Solution. $|f(t(1, 2))| = 3|t| = |t| + 2|t| = \|t(1, 2)\|_1$, so the hypothesis holds. Every $F \in (\mathbb{R}^2)^*$ is $F(\mathbf{x}) = c_1 x_1 + c_2 x_2$ (Theorem 5.1.1), and it extends f exactly when $F((1, 2)) = 3$, that is, $c_1 + 2c_2 = 3$.

Next, $|F(\mathbf{x})| \leq \|\mathbf{x}\|_1$ for all $\mathbf{x}$ holds exactly when $|c_1| \leq 1$ and $|c_2| \leq 1$. Indeed, if both hold then $|c_1 x_1 + c_2 x_2| \leq |x_1| + |x_2|$; and if, say, $|c_1| > 1$, then $\mathbf{x} = \mathbf{e}_1$ gives $|F(\mathbf{e}_1)| > 1 = \|\mathbf{e}_1\|_1$.

So we need $c_1 = 3 - 2c_2$ with $|3 - 2c_2| \leq 1$ and $|c_2| \leq 1$. The first gives $1 \leq c_2 \leq 2$, the second $c_2 \leq 1$; hence $c_2 = 1$ and $c_1 = 1$. The only extension is $F(\mathbf{x}) = x_1 + x_2$.

18.3.8.3 C. Going deeper

Exercise. C1: The projection does not stretch

Let $C \subseteq V$ be nonempty, closed, and convex. Prove that $\|P_C(x) - P_C(y)\| \le \|x - y\|$ for all $x, y \in V$.

Hint: write Eq. 18.3.1 once at x, tested with the point $P_C(y)$, and once at y, tested with $P_C(x)$, and add.

Solution. Put $p = P_C(x)$ and $q = P_C(y)$. By Theorem 18.3.1 (b), applied at x with the point $q \in C$ and at y with the point $p \in C$,

$$\langle x - p, q - p\rangle \le 0, \qquad \langle y - q, p - q\rangle \le 0.$$

The second is $\langle q - y, q - p\rangle \le 0$. Adding, $\langle (x - y) + (q - p), q - p\rangle \le 0$, that is,

$$\|q - p\|^2 \le \langle y - x, q - p\rangle \le \|y - x\|\,\|q - p\|$$

by Theorem 11.1.5. If $q = p$ there is nothing to prove; otherwise divide by $\|q - p\| > 0$. This proves $\|p - q\| \le \|x - y\|$.

Exercise. C2: Strict separation and distance

Let $C_1, C_2 \subseteq V$ be nonempty, closed, and convex, and put $\delta = \inf\{\|y - x\| : x \in C_1, y \in C_2\}$.

(a) Prove that if C_1 and C_2 are strictly separated, then $\delta > 0$.

(b) Prove that if $\delta > 0$, then C_1 and C_2 are strictly separated.

(c) Hence deduce, without the computation of Example (A hyperbola and its asymptote), that the sets A and B there are not strictly separated.

Hint for (b): replace the difference set D of Theorem 18.3.7 by its closure.

Solution.

(a) Let $a \neq 0$, c, and $\varepsilon > 0$ strictly separate the sets, with C_1 on the side $\le c - \varepsilon$. For $x \in C_1$ and $y \in C_2$, Theorem 11.1.5 gives

$$2\varepsilon \le \langle a, y\rangle - \langle a, x\rangle = \langle a, y - x\rangle \le \|a\|\,\|y - x\|,$$

so $\|y - x\| \ge 2\varepsilon/\|a\|$. Hence $\delta \ge 2\varepsilon/\|a\| > 0$.

(b) Let $D = \{y - x : x \in C_1, y \in C_2\}$, which is convex by the computation in the proof of Theorem 18.3.7. Every $d \in D$ has $\|d\| \ge \delta$, and so does every point of $\overline{D}$: if $d_k \to d$ with $d_k \in D$, then $\|d\| \ge \|d_k\| - \|d_k - d\| \ge \delta - \|d_k - d\|$ for every k, and the right side tends to δ. So $0 \notin \overline{D}$, and $\overline{D}$ is nonempty, closed, and convex by Lemma 18.3.5 (a). By Theorem 18.3.2 there is $b \neq 0$ with $\langle b, d\rangle \le -\|b\|^2$ on $\overline{D} \supseteq D$. With $a = -b$, this is $\langle a, x\rangle + \|a\|^2 \le \langle a, y\rangle$ for all $x \in C_1$, $y \in C_2$, and the last paragraph of the proof of Theorem 18.3.7 (a) turns it into strict separation with $\varepsilon = \frac{1}{2}\|a\|^2$.

(c) The points $(t, 1/t) \in A$ and $(t, 0) \in B$ are at distance $1/t$, so $\delta \le 1/t$ for every $t > 0$, and $\delta = 0$. By (a), A and B are not strictly separated.

18.4 Gauges, Support Functions, and Dual Norms

Chapter 16 §01 observed that the closed unit ball of every norm is convex and symmetric, and left the converse alone: does every convex, symmetric set arise as the unit ball of some norm? This section answers yes, by reading a norm off a convex set from the inside, the **gauge**. It then reads a function off a convex set from the outside, the **support function**, which turns §03's description of a closed convex set by half-spaces into a formula. Applied to a unit ball, the support function is the **dual norm**, the measurement Chapter 16 never gave to a linear functional, and the last part of the section proves that taking the dual twice returns the norm we started with.

Throughout, the field is $\mathbb{R}$, and we work in $\mathbb{R}^n$, $n \geq 1$, with the dot product $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^\top \mathbf{y}$ and its norm $\|\cdot\|_2$. So everything proved in §03 is available, with $V = \mathbb{R}^n$. A symbol $\|\cdot\|$ without a subscript is an **arbitrary** norm on $\mathbb{R}^n$ in the sense of Definition 16.1.1, and B is its closed unit ball (Definition 16.1.2). Interior points are meant in the sense of Definition 18.3.2, which does not depend on the norm. For a set K and a real t, we write $tK = \{t\mathbf{x} : \mathbf{x} \in K\}$.

18.4.1 Gauges

Every norm can be read off its unit ball. By (N2), for $t > 0$ we have $\|\mathbf{x}\| \leq t$ exactly when $\|\mathbf{x}/t\| \leq 1$, that is, when $\mathbf{x} \in tB$. Hence

$$\|\mathbf{x}\| = \inf\{t > 0 : \mathbf{x} \in tB\},$$

the inflation factor by which B has to be blown up before it reaches $\mathbf{x}$. The right-hand side makes sense for sets that are not unit balls, and this is the formula to study.

The gauge of a convex set measures a vector by how far the set must be inflated to reach it.

Definition 18.4.1. Gauge

Let $K \subseteq \mathbb{R}^n$ be convex, with $\mathbf{0}$ an **interior point** of K. The **gauge** (or **Minkowski functional**) of K is the function $p_K \colon \mathbb{R}^n \to \mathbb{R}$ given by

$$p_K(\mathbf{x}) = \inf\{t > 0 : \mathbf{x} \in tK\}.$$

The infimum exists. Choose $r > 0$ with every $\mathbf{y}$ satisfying $\|\mathbf{y}\|_2 < r$ in K. If $t > \|\mathbf{x}\|_2/r$, then $\|\mathbf{x}/t\|_2 < r$, so $\mathbf{x}/t \in K$ and $\mathbf{x} \in tK$. So the set of admissible t is nonempty, and it is bounded below by 0; by **the completeness of $\mathbb{R}$, fact (A1) of Chapter 16's introduction**, it has a greatest lower bound, and

$$0 \leq p_K(\mathbf{x}) \leq \|\mathbf{x}\|_2/r \qquad \text{for every } \mathbf{x} \in \mathbb{R}^n.$$

The interior point is what makes p_K finite. For the segment $[-\mathbf{e}_1, \mathbf{e}_1] \subseteq \mathbb{R}^2$, which has no interior points, the vector $\mathbf{e}_2$ lies in no multiple tK, and the formula would give the infimum of the empty set.

Proposition 18.4.1. Properties of the Gauge

Let $K \subseteq \mathbb{R}^n$ be convex with $\mathbf{0}$ an interior point, and let $\mathbf{x} \in \mathbb{R}^n$.

(a) $\mathbf{x} \in tK$ for **every** $t > p_K(\mathbf{x})$.
(b) p_K is sublinear (Definition 18.3.3).
(c) $\{\mathbf{x} : p_K(\mathbf{x}) < 1\} \subseteq K \subseteq \{\mathbf{x} : p_K(\mathbf{x}) \le 1\}$.
(d) If K is closed, then $K = \{\mathbf{x} : p_K(\mathbf{x}) \le 1\}$.

Idea. Everything comes from one observation, (a): since $\mathbf{0} \in K$ and K is convex, shrinking a point of K towards $\mathbf{0}$ keeps it in K, so the admissible t form a whole half-line. Subadditivity is then convexity applied to $\mathbf{x}/s$ and $\mathbf{y}/t$ with weights proportional to s and t.

Proof.

(a) Let $t > p_K(\mathbf{x})$. By the definition of an infimum there is s with $p_K(\mathbf{x}) \le s < t$ and $\mathbf{x} \in sK$, so $\mathbf{x}/s \in K$. Since $\mathbf{0} \in K$ and K is convex,

$$\frac{\mathbf{x}}{t} = \frac{s}{t} \cdot \frac{\mathbf{x}}{s} + \Big(1 - \frac{s}{t}\Big)\mathbf{0} \in K,$$

that is, $\mathbf{x} \in tK$.

(b) Let $\lambda > 0$. Since $\lambda\mathbf{x} \in tK$ exactly when $\mathbf{x} \in (t/\lambda)K$, the admissible t for $\lambda\mathbf{x}$ are λ times those for $\mathbf{x}$, and $p_K(\lambda\mathbf{x}) = \lambda p_K(\mathbf{x})$. For $\lambda = 0$, $\mathbf{0} \in tK$ for every $t > 0$, so $p_K(\mathbf{0}) = 0$. For subadditivity, let $\varepsilon > 0$, $s = p_K(\mathbf{x}) + \varepsilon$ and $t = p_K(\mathbf{y}) + \varepsilon$. By (a), $\mathbf{x}/s$ and $\mathbf{y}/t$ lie in K, so by convexity

$$\frac{\mathbf{x}+\mathbf{y}}{s+t} = \frac{s}{s+t} \cdot \frac{\mathbf{x}}{s} + \frac{t}{s+t} \cdot \frac{\mathbf{y}}{t} \in K.$$

Hence $p_K(\mathbf{x}+\mathbf{y}) \le s + t = p_K(\mathbf{x}) + p_K(\mathbf{y}) + 2\varepsilon$, for every $\varepsilon > 0$, and so $p_K(\mathbf{x}+\mathbf{y}) \le p_K(\mathbf{x}) + p_K(\mathbf{y})$.

(c) If $p_K(\mathbf{x}) < 1$, then $\mathbf{x} \in 1 \cdot K$ by (a). If $\mathbf{x} \in K$, then $t = 1$ is admissible, so $p_K(\mathbf{x}) \le 1$.

(d) Let $p_K(\mathbf{x}) \le 1$. By (a), $\mathbf{x} \in (1+1/k)K$ for every $k \ge 1$, so the points $\mathbf{x}/(1+1/k)$ lie in K. They converge to $\mathbf{x}$, and K is closed, so $\mathbf{x} \in K$. With (c) this gives $K = \{p_K \le 1\}$.

■

The gauge of a unit ball is its norm, by the computation that opened this section. A second example: for the rectangle $K = [-2, 2] \times [-1, 1] \subseteq \mathbb{R}^2$, a vector lies in tK exactly when $|x_1| \le 2t$ and $|x_2| \le t$, so

$$p_K(\mathbf{x}) = \max\big(\tfrac{1}{2}|x_1|,\ |x_2|\big),$$

a weighted ∞-norm of the kind in Example (Inner products, weights, and a function space) (b). At the degenerate extreme, $K = \mathbb{R}^n$ lies in every tK, and p_K is identically 0: sublinear, but no norm.

The next theorem says exactly when a gauge is a norm. Its part (b) records the facts about unit balls that make the match perfect.

Theorem 18.4.2. Convex Bodies and Norms

(a) Let $K \subseteq \mathbb{R}^n$ be **closed, bounded, convex**, and **symmetric** ($-K = K$), with $\mathbf{0}$ an

interior point. Then p_K is a norm on $\mathbb{R}^n$, and its closed unit ball is K.

(b) Conversely, the closed unit ball B of any norm $\|\cdot\|$ on $\mathbb{R}^n$ is closed, bounded, convex, and symmetric, with $\mathbf{0}$ an interior point, and $p_B = \|\cdot\|$.

So the norms on $\mathbb{R}^n$ correspond one to one with the sets described in (a).

Idea. Sublinearity is two thirds of a norm, and Proposition 18.4.1 already has it. The missing third is absolute homogeneity for a scalar $c < 0$, which is what symmetry buys, and positive definiteness, which is what boundedness buys: a nonzero vector cannot fit inside every shrunken copy tK of a bounded set. Closedness makes the unit ball of p_K equal to K rather than squeezed between the two sets of Proposition 18.4.1 (c).

Proof.

(a) By Proposition 18.4.1 (b), p_K is sublinear, which gives (N3) and (N2) for $c \geq 0$. For $c < 0$, symmetry gives $\mathbf{x} \in tK$ if and only if $-\mathbf{x} \in t(-K) = tK$, so $p_K(-\mathbf{x}) = p_K(\mathbf{x})$ and $p_K(c\mathbf{x}) = |c|p_K(-\mathbf{x}) = |c|p_K(\mathbf{x})$. For (N1), $p_K \geq 0$ by definition. Let R be such that $\|\mathbf{y}\|_2 \leq R$ for all $\mathbf{y} \in K$. If $p_K(\mathbf{x}) = 0$, then $\mathbf{x} \in tK$ for every $t > 0$ by Proposition 18.4.1 (a), so $\|\mathbf{x}\|_2 \leq tR$ for every $t > 0$, and $\mathbf{x} = \mathbf{0}$. So p_K is a norm, and its closed unit ball $\{p_K \leq 1\}$ is K by Proposition 18.4.1 (d).

(b) Chapter 16 §01 showed, right after Definition 16.1.2, that B is convex and symmetric. By Theorem 16.2.2 there are constants $0 < m \leq M$ with $m\|\mathbf{x}\|_2 \leq \|\mathbf{x}\| \leq M\|\mathbf{x}\|_2$. If $\|\mathbf{x}\| \leq 1$, then $\|\mathbf{x}\|_2 \leq 1/m$, so B is bounded. If $\|\mathbf{x}\|_2 < 1/M$, then $\|\mathbf{x}\| < 1$, so $\mathbf{0}$ is an interior point. If $\mathbf{x}_k \in B$ and $\mathbf{x}_k \to \mathbf{x}$, then $|\|\mathbf{x}_k\| - \|\mathbf{x}\|| \leq \|\mathbf{x}_k - \mathbf{x}\| \leq M\|\mathbf{x}_k - \mathbf{x}\|_2 \to 0$ by Lemma 16.1.1, and $\|\mathbf{x}\| \leq 1$ because a nonstrict inequality survives the limit; so B is closed. Finally, $\mathbf{x} \in tB$ exactly when $t \geq \|\mathbf{x}\|$, for $t > 0$, so $p_B(\mathbf{x}) = \|\mathbf{x}\|$ (for $\mathbf{x} = \mathbf{0}$, the infimum of all $t > 0$ is 0). This proves the theorem.

■

This is the converse of Chapter 16's "every unit ball is convex". There, a norm produced a convex symmetric set, the ball of Definition 16.1.2, and convexity was what the triangle inequality bought. Here a convex symmetric set produces a norm, and the triangle inequality is what convexity buys: the subadditivity step in Proposition 18.4.1 (b) is the convexity of K, used once. Norms and closed bounded symmetric convex sets with $\mathbf{0}$ inside are the same objects seen two ways, one analytic, one geometric.

Each hypothesis in (a) carries one clause of Definition 16.1.1, and dropping one breaks exactly that clause.

- **Drop symmetry.** The triangle $T = \{\mathbf{x} : x_1 \geq -1,\ x_2 \geq -1,\ x_1 + x_2 \leq 1\}$, with vertices $(-1,-1)$, $(2,-1)$, $(-1,2)$, is closed, bounded, and convex, and contains the disc of radius $\frac{1}{2}$ about $\mathbf{0}$. A vector lies in tT exactly when $-x_1 \leq t$, $-x_2 \leq t$, and $x_1 + x_2 \leq t$, so

$$p_T(\mathbf{x}) = \max(-x_1,\ -x_2,\ x_1 + x_2),$$

  which is never negative, since the three numbers add up to 0. This is sublinear and satisfies (N1), but $p_T((1,1)) = 2$ while $p_T((-1,-1)) = 1$, so (N2) fails at $c = -1$.
- **Drop boundedness.** The strip $S = \{\mathbf{x} \in \mathbb{R}^2 : |x_2| \leq 1\}$ is closed, convex, and symmetric, and $p_S(\mathbf{x}) = |x_2|$. Now (N2) and (N3) hold but (N1) fails: $p_S(\mathbf{e}_1) = 0$. This is the

seminorm of Example (Three near misses) (b).

Quick check. Let K be the open unit disc $\{\mathbf{x} \in \mathbb{R}^2 : \|\mathbf{x}\|_2 < 1\}$, which is bounded, symmetric, and convex (if $\|\mathbf{x}\|_2, \|\mathbf{y}\|_2 < 1$ and $0 \le t \le 1$, then $\|(1-t)\mathbf{x}+t\mathbf{y}\|_2 \le (1-t)\|\mathbf{x}\|_2+t\|\mathbf{y}\|_2 < 1$) but not closed. What is p_K, and what is its closed unit ball?

Solution. For $t > 0$, $\mathbf{x} \in tK$ exactly when $\|\mathbf{x}\|_2 < t$, so the admissible t form the interval $(\|\mathbf{x}\|_2, \infty)$ and $p_K(\mathbf{x}) = \|\mathbf{x}\|_2$. So p_K is a norm after all, but its closed unit ball is the closed disc, not K. Closedness is used only for the last claim of Theorem 18.4.2 (a), through Proposition 18.4.1 (d).

18.4.2 Support functions

§03 described a closed convex set from the outside: by Corollary 18.3.3 it is the intersection of the closed half-spaces containing it. For a fixed normal vector $\mathbf{y}$, the half-spaces $\{\mathbf{x} : \langle \mathbf{x}, \mathbf{y}\rangle \le c\}$ containing K are those with $c \ge \langle \mathbf{x}, \mathbf{y}\rangle$ for every $\mathbf{x} \in K$, and the tightest of them has $c = \sup_{\mathbf{x}\in K}\langle \mathbf{x}, \mathbf{y}\rangle$. That number, as a function of the direction $\mathbf{y}$, deserves a name.

The support function of a set records, for each direction, how far the set reaches in that direction.

Definition 18.4.2. Support Function

Let $K \subseteq \mathbb{R}^n$ be **nonempty**. The **support function** of K is

$$h_K(\mathbf{y}) = \sup_{\mathbf{x}\in K} \langle \mathbf{x}, \mathbf{y}\rangle, \qquad \mathbf{y} \in \mathbb{R}^n,$$

with the value $+\infty$ when the numbers $\langle \mathbf{x}, \mathbf{y}\rangle$, $\mathbf{x} \in K$, are unbounded above.

When the numbers are bounded above, the supremum exists by (A1). If K is **bounded**, say $\|\mathbf{x}\|_2 \le R$ on K, then $\langle \mathbf{x}, \mathbf{y}\rangle \le R\|\mathbf{y}\|_2$ by Theorem 11.1.5, so h_K is real-valued. If K is **compact**, the supremum is a maximum by **the extreme value theorem, fact (A4) of Chapter 16's introduction**, since $\mathbf{x} \mapsto \langle \mathbf{x}, \mathbf{y}\rangle$ is continuous: $|\langle \mathbf{x}, \mathbf{y}\rangle - \langle \mathbf{x}', \mathbf{y}\rangle| \le \|\mathbf{x} - \mathbf{x}'\|_2\|\mathbf{y}\|_2$. For a unit vector $\mathbf{y}$, the hyperplane $\{\langle \cdot, \mathbf{y}\rangle = h_K(\mathbf{y})\}$ is the one with normal $\mathbf{y}$ that touches K from outside, and $h_K(\mathbf{y})$ is its signed distance from $\mathbf{0}$.

Some examples, each computed from the definition.

- **A point.** $h_{\{\mathbf{x}_0\}}(\mathbf{y}) = \langle \mathbf{x}_0, \mathbf{y}\rangle$, a linear function.
- **The Euclidean ball.** $h_{B_2}(\mathbf{y}) = \|\mathbf{y}\|_2$: at most that by Theorem 11.1.5, and attained at $\mathbf{x} = \mathbf{y}/\|\mathbf{y}\|_2$ when $\mathbf{y} \neq \mathbf{0}$.
- **The standard simplex** $\operatorname{conv}\{\mathbf{e}_1, \dots, \mathbf{e}_n\}$. By Theorem 18.1.3 its points are $\sum_i t_i\mathbf{e}_i$ with $t_i \ge 0$ and $\sum_i t_i = 1$, and $\sum_i t_i y_i \le \max_i y_i$, with equality at a vertex $\mathbf{e}_k$ where y_k is largest. So $h(\mathbf{y}) = \max_i y_i$.
- **A segment.** $h_{[-\mathbf{e}_1, \mathbf{e}_1]}(\mathbf{y}) = |y_1|$ in $\mathbb{R}^2$.
- **An unbounded set.** For the half-line $\{t\mathbf{e}_1 : t \ge 0\}$, $h(\mathbf{y}) = 0$ if $y_1 \le 0$ and $h(\mathbf{y}) = +\infty$ if $y_1 > 0$.

Proposition 18.4.3. Properties of the Support Function

Let $K, L \subseteq \mathbb{R}^n$ be nonempty and bounded.

(a) h_K is sublinear.
(b) If $K \subseteq L$, then $h_K \le h_L$.
(c) $h_{\operatorname{conv} K} = h_K$.

Idea. A supremum of linear functions of $\mathbf{y}$ inherits whatever inequalities each of them satisfies, which gives (a) and (b). For (c), a linear function takes on a convex combination a weighted average of its values, and an average never exceeds the largest value.

Proof.

(a) For $\mathbf{x} \in K$, $\langle \mathbf{x}, \mathbf{y}+\mathbf{z}\rangle = \langle \mathbf{x}, \mathbf{y}\rangle + \langle \mathbf{x}, \mathbf{z}\rangle \le h_K(\mathbf{y}) + h_K(\mathbf{z})$, and taking the supremum over $\mathbf{x}$ gives $h_K(\mathbf{y}+\mathbf{z}) \le h_K(\mathbf{y}) + h_K(\mathbf{z})$. For $t > 0$, $\langle \mathbf{x}, t\mathbf{y}\rangle = t\langle \mathbf{x}, \mathbf{y}\rangle$, and a positive factor comes out of a supremum; for $t = 0$ both sides are 0, as $K \neq \emptyset$.

(b) A supremum over a larger set is at least as large.

(c) Since $K \subseteq \operatorname{conv} K$, (b) gives $h_K \le h_{\operatorname{conv} K}$, once we know $\operatorname{conv} K$ is bounded: it lies in the ball $\{\|\mathbf{x}\|_2 \le R\}$, which is convex and contains K (Definition 18.1.4). Conversely, by Theorem 18.1.3 a point of $\operatorname{conv} K$ is $\mathbf{x} = \sum_i t_i \mathbf{x}_i$ with $\mathbf{x}_i \in K$, $t_i \ge 0$, $\sum_i t_i = 1$, and then

$$\langle \mathbf{x}, \mathbf{y}\rangle = \sum_i t_i \langle \mathbf{x}_i, \mathbf{y}\rangle \le \sum_i t_i \, h_K(\mathbf{y}) = h_K(\mathbf{y}).$$

So $h_{\operatorname{conv} K} \le h_K$.

■

Warning. **A support function cannot see gaps.** The two-point set $\{\mathbf{e}_1, -\mathbf{e}_1\}$ and the segment $[-\mathbf{e}_1, \mathbf{e}_1]$ in $\mathbb{R}^2$ have the same support function $|y_1|$, by Proposition 18.4.3 (c). So h_K does not determine K in general; it determines only the convex hull, and the next theorem shows that among closed convex sets it determines the set.

Here is the promised formula. It is Corollary 18.3.3 with the family of half-spaces indexed by their normals.

Theorem 18.4.4. The Support Function Determines a Closed Convex Set

Let $K, L \subseteq \mathbb{R}^n$ be nonempty, closed, and convex. Then

(a) $K = \{\mathbf{x} \in \mathbb{R}^n : \langle \mathbf{x}, \mathbf{y}\rangle \le h_K(\mathbf{y}) \text{ for every } \mathbf{y} \in \mathbb{R}^n\}$;
(b) $K \subseteq L$ if and only if $h_K(\mathbf{y}) \le h_L(\mathbf{y})$ for every $\mathbf{y}$; in particular $K = L$ if and only if $h_K = h_L$.

Here h_K and h_L may take the value $+\infty$, with $a \le +\infty$ for every a.

Idea. Part (a) is Corollary 18.3.3, with the half-space for the normal $\mathbf{y}$ chosen as tight as possible, at level $h_K(\mathbf{y})$. Part (b) then compares the two families of half-spaces normal by normal.

Proof.

(a) Every $\mathbf{x} \in K$ satisfies $\langle \mathbf{x}, \mathbf{y} \rangle \le h_K(\mathbf{y})$ by definition. Conversely let $\mathbf{x} \notin K$. By Theorem 18.3.2 there is $\mathbf{a} \neq \mathbf{0}$ with $\langle \mathbf{a}, \mathbf{z} \rangle \le \langle \mathbf{a}, \mathbf{x} \rangle - \|\mathbf{a}\|_2^2$ for all $\mathbf{z} \in K$. Hence $h_K(\mathbf{a}) \le \langle \mathbf{a}, \mathbf{x} \rangle - \|\mathbf{a}\|_2^2 < \langle \mathbf{x}, \mathbf{a} \rangle$, and $\mathbf{x}$ violates the condition at $\mathbf{y} = \mathbf{a}$.

(b) If $K \subseteq L$, then $h_K \le h_L$ because a supremum over a larger set is at least as large. Conversely, suppose $h_K \le h_L$ and let $\mathbf{x} \in K$. For every $\mathbf{y}$, $\langle \mathbf{x}, \mathbf{y} \rangle \le h_K(\mathbf{y}) \le h_L(\mathbf{y})$, so $\mathbf{x} \in L$ by (a) applied to L. The last statement follows by applying this in both directions.

■

18.4.3 Polar sets

The condition $h_K(\mathbf{y}) \le 1$ cuts out a set of directions, and that set turns out to be a mirror image of K.

Definition 18.4.3. Polar Set

The **polar** of a set $K \subseteq \mathbb{R}^n$ is

$$K^\circ = \{\mathbf{y} \in \mathbb{R}^n : \langle \mathbf{x}, \mathbf{y} \rangle \le 1 \text{ for every } \mathbf{x} \in K\}.$$

For nonempty K, $K^\circ = \{\mathbf{y} : h_K(\mathbf{y}) \le 1\}$. The polar is always closed, convex, and contains $\mathbf{0}$, being an intersection of closed half-spaces $\{\mathbf{y} : \langle \mathbf{x}, \mathbf{y} \rangle \le 1\}$, one for each $\mathbf{x} \in K$. Two examples:

- $(B_2)^\circ = B_2$, because $h_{B_2}(\mathbf{y}) = \|\mathbf{y}\|_2$.
- For the single point $K = \{\mathbf{e}_1\}$ in $\mathbb{R}^2$, $K^\circ = \{\mathbf{y} : y_1 \le 1\}$, a half-plane. Its polar consists of the $\mathbf{x}$ with $x_1 y_1 + x_2 y_2 \le 1$ whenever $y_1 \le 1$. Letting y_2 run over $\mathbb{R}$ forces $x_2 = 0$, letting $y_1 \to -\infty$ forces $x_1 \ge 0$, and $y_1 = 1$ forces $x_1 \le 1$. So $K^{\circ\circ}$ is the segment $[\mathbf{0}, \mathbf{e}_1]$, which is not K.

The second example shows what the theorem must assume.

Theorem 18.4.5. Bipolar Theorem

Let $K \subseteq \mathbb{R}^n$ be closed and convex with $\mathbf{0} \in K$. Then $K^{\circ\circ} = K$.

Idea. One inclusion is the definition read twice. For the other, cut a point $\mathbf{x} \notin K$ off with Theorem 18.3.2, and rescale the normal so that K lies in $\{\langle \cdot, \mathbf{b} \rangle \le 1\}$ while $\mathbf{x}$ does not. That rescaled normal lies in K° and witnesses $\mathbf{x} \notin K^{\circ\circ}$; the rescaling needs $\mathbf{0} \in K$.

Proof. ($\supseteq$) Let $\mathbf{x} \in K$. For every $\mathbf{y} \in K^\circ$, $\langle \mathbf{x}, \mathbf{y} \rangle \le 1$ by the definition of K°. Hence $\mathbf{x} \in K^{\circ\circ}$.

($\subseteq$) Let $\mathbf{x} \notin K$. By Theorem 18.3.2 there is $\mathbf{a} \neq \mathbf{0}$ with $\langle \mathbf{a}, \mathbf{z} \rangle \le \langle \mathbf{a}, \mathbf{x} \rangle - \|\mathbf{a}\|_2^2$ for all $\mathbf{z} \in K$. Put $s = \sup_{\mathbf{z} \in K} \langle \mathbf{a}, \mathbf{z} \rangle$, which exists by (A1); since $\mathbf{0} \in K$, $s \ge 0$, and $s < \langle \mathbf{a}, \mathbf{x} \rangle$. If $s > 0$, let $\mathbf{b} = \mathbf{a}/s$. Then $\langle \mathbf{z}, \mathbf{b} \rangle \le 1$ for all $\mathbf{z} \in K$, so $\mathbf{b} \in K^\circ$, while $\langle \mathbf{x}, \mathbf{b} \rangle > 1$. If $s = 0$, then $\langle \mathbf{a}, \mathbf{x} \rangle > 0$; let $\mathbf{b} = 2\mathbf{a}/\langle \mathbf{a}, \mathbf{x} \rangle$. Then $\langle \mathbf{z}, \mathbf{b} \rangle \le 0 \le 1$ on K, so $\mathbf{b} \in K^\circ$, while $\langle \mathbf{x}, \mathbf{b} \rangle = 2 > 1$. In both cases $\mathbf{x} \notin K^{\circ\circ}$.

■

The hypothesis $\mathbf{0} \in K$ was used to make $s \ge 0$, so that dividing by s keeps the direction of the inequality; the counterexample $\{\mathbf{e}_1\}$ above is exactly a closed convex set without it.

18.4.4 Dual norms

Chapter 16 measured vectors (Definition 16.1.1) and matrices (Definition 16.3.2). A third kind of object sits between them and was never measured: a linear functional on $\mathbb{R}^n$, which by Theorem 5.1.1 is $\mathbf{x} \mapsto \mathbf{y}^\top \mathbf{x} = \langle \mathbf{x}, \mathbf{y} \rangle$ for exactly one $\mathbf{y}$. How big is it?

When $\mathbf{x}$ is measured by $\|\cdot\|_2$, Theorem 11.1.5 answers: $|\langle \mathbf{x}, \mathbf{y} \rangle| \le \|\mathbf{x}\|_2 \|\mathbf{y}\|_2$, and the constant $\|\mathbf{y}\|_2$ is the best possible. When $\mathbf{x}$ is measured by another norm, the question is the same — what is the smallest c with $|\langle \mathbf{x}, \mathbf{y} \rangle| \le c\|\mathbf{x}\|$ for every $\mathbf{x}$? — and the answer is the largest value of $\langle \mathbf{x}, \mathbf{y} \rangle$ over the unit ball. That is the support function of B, and it is also the operator norm of the $1 \times n$ matrix $\mathbf{y}^\top$. An expression that turns up three ways deserves a name.

The dual norm of $\mathbf{y}$ is the most that the functional $\langle \cdot, \mathbf{y} \rangle$ can extract from a vector of norm at most one.

Definition 18.4.4. Dual Norm

Let $\|\cdot\|$ be a norm on $\mathbb{R}^n$, with closed unit ball B. The **dual norm** of $\|\cdot\|$ is the function $\|\cdot\|_* \colon \mathbb{R}^n \to \mathbb{R}$ given by

$$\|\mathbf{y}\|_* = \max\{\langle \mathbf{x}, \mathbf{y} \rangle : \mathbf{x} \in \mathbb{R}^n,\ \|\mathbf{x}\| \le 1\} = h_B(\mathbf{y}).$$

The maximum exists. By Theorem 18.4.2 (b), B is nonempty, closed, and bounded, hence compact by Corollary 16.2.5, which rests on **fact (A3) of Chapter 16's introduction**; and $\mathbf{x} \mapsto \langle \mathbf{x}, \mathbf{y} \rangle$ is continuous, so by **the extreme value theorem, fact (A4)**, it attains its maximum on B.

In words: fix $\mathbf{y}$, let $\mathbf{x}$ range over the vectors that the **given** norm calls short, and record the largest value of $\langle \mathbf{x}, \mathbf{y} \rangle$. The inner product is part of the definition, and so is the norm; changing either changes $\|\cdot\|_*$. No absolute value is needed in the maximum: B is symmetric, so replacing $\mathbf{x}$ by $-\mathbf{x}$ shows that the largest value of $\langle \mathbf{x}, \mathbf{y} \rangle$ on B is also the largest value of $|\langle \mathbf{x}, \mathbf{y} \rangle|$.

Proposition 18.4.6. Properties of the Dual Norm

Let $\|\cdot\|$ be a norm on $\mathbb{R}^n$, with closed unit ball B.

(a) $\|\cdot\|_*$ is a norm on $\mathbb{R}^n$.

(b) $|\langle \mathbf{x}, \mathbf{y} \rangle| \le \|\mathbf{x}\|\, \|\mathbf{y}\|_*$ for all $\mathbf{x}, \mathbf{y}$, and $\|\mathbf{y}\|_*$ is the smallest constant c with $|\langle \mathbf{x}, \mathbf{y} \rangle| \le c\|\mathbf{x}\|$

for every $\mathbf{x}$.

(c) $\|\mathbf{y}\|_*$ is the operator norm (Definition 16.3.2) of the $1 \times n$ matrix $\mathbf{y}^\top$, with $\|\cdot\|$ on $\mathbb{R}^n$ and $|\cdot|$ on $\mathbb{R}$.

(d) The closed unit ball of $\|\cdot\|_*$ is the polar B°.

Idea. Each norm axiom for $\|\cdot\|_*$ is checked one $\mathbf{x} \in B$ at a time and then survives the maximum; (N1) needs a good test vector, $\mathbf{y}$ itself rescaled into B. For (b), rescale any $\mathbf{x} \neq \mathbf{0}$ into B. Parts (c) and (d) only compare definitions.

Proof.

(a) (N1): $\mathbf{x} = \mathbf{0} \in B$ gives $\|\mathbf{y}\|_* \geq 0$. If $\mathbf{y} \neq \mathbf{0}$, then $\mathbf{x} = \mathbf{y}/\|\mathbf{y}\| \in B$ gives $\|\mathbf{y}\|_* \geq \|\mathbf{y}\|_2^2/\|\mathbf{y}\| > 0$. (N2): for $c \geq 0$, $\max_B \langle \mathbf{x}, c\mathbf{y}\rangle = c \max_B \langle \mathbf{x}, \mathbf{y}\rangle$; for $c < 0$, $\langle \mathbf{x}, c\mathbf{y}\rangle = |c|\langle -\mathbf{x}, \mathbf{y}\rangle$, and $-\mathbf{x}$ runs over B as $\mathbf{x}$ does, by (N2) for $\|\cdot\|$. (N3): for $\mathbf{x} \in B$, $\langle \mathbf{x}, \mathbf{y}+\mathbf{z}\rangle = \langle \mathbf{x}, \mathbf{y}\rangle + \langle \mathbf{x}, \mathbf{z}\rangle \leq \|\mathbf{y}\|_* + \|\mathbf{z}\|_*$; take the maximum over $\mathbf{x}$.

(b) For $\mathbf{x} = \mathbf{0}$ there is nothing to prove. For $\mathbf{x} \neq \mathbf{0}$, $\mathbf{x}/\|\mathbf{x}\| \in B$, so $\pm\langle \mathbf{x}, \mathbf{y}\rangle/\|\mathbf{x}\| = \langle \pm\mathbf{x}/\|\mathbf{x}\|, \mathbf{y}\rangle \leq \|\mathbf{y}\|_*$, using that B is symmetric. If c is any constant with $|\langle \mathbf{x}, \mathbf{y}\rangle| \leq c\|\mathbf{x}\|$ for all $\mathbf{x}$, then $\langle \mathbf{x}, \mathbf{y}\rangle \leq c$ on B, so $\|\mathbf{y}\|_* \leq c$.

(c) By Definition 16.3.2 and the equivalent descriptions that follow it in Chapter 16 §03, the operator norm of $\mathbf{y}^\top$ is $\max\{|\mathbf{y}^\top \mathbf{x}| : \|\mathbf{x}\| \leq 1\}$, which is $\|\mathbf{y}\|_*$ by the remark after Definition 18.4.4.

(d) $\|\mathbf{y}\|_* \leq 1$ exactly when $\langle \mathbf{x}, \mathbf{y}\rangle \leq 1$ for every $\mathbf{x} \in B$, which is the definition of $\mathbf{y} \in B^\circ$ (Definition 18.4.3).

■

Part (b) is Cauchy–Schwarz for an arbitrary norm, with the price that the two factors are measured differently. Now the three norms of Chapter 16.

Proposition 18.4.7. Duals of the 1-, 2-, and Infinity-Norms

On $\mathbb{R}^n$ with the dot product,

(a) the dual of $\|\cdot\|_1$ is $\|\cdot\|_\infty$;

(b) the dual of $\|\cdot\|_\infty$ is $\|\cdot\|_1$;

(c) the dual of $\|\cdot\|_2$ is $\|\cdot\|_2$.

Idea. Each part is an upper bound and a witness, the same pattern as Theorem 16.3.3. For the 1-norm, all of the budget $\|\mathbf{x}\|_1 \leq 1$ goes on the coordinate where $|y_k|$ is largest. For the ∞-norm, every coordinate can be spent at once, each with the sign of y_i, so nothing cancels. For the 2-norm, the witness is $\mathbf{y}$ itself, rescaled.

Proof.

(a) For $\|\mathbf{x}\|_1 \leq 1$, $\langle \mathbf{x}, \mathbf{y}\rangle \leq \sum_i |x_i||y_i| \leq \|\mathbf{y}\|_\infty \sum_i |x_i| \leq \|\mathbf{y}\|_\infty$. Choose k with $|y_k| = \|\mathbf{y}\|_\infty$,

and let $\mathbf{x} = \pm\mathbf{e}_k$ with the sign of y_k (either sign if $y_k = 0$). Then $\|\mathbf{x}\|_1 = 1$ and $\langle \mathbf{x}, \mathbf{y} \rangle = |y_k| = \|\mathbf{y}\|_\infty$.

(b) For $\|\mathbf{x}\|_\infty \le 1$, $\langle \mathbf{x}, \mathbf{y} \rangle \le \sum_i |x_i||y_i| \le \sum_i |y_i| = \|\mathbf{y}\|_1$. Let $x_i = 1$ if $y_i \ge 0$ and $x_i = -1$ if $y_i < 0$. Then $\|\mathbf{x}\|_\infty = 1$ and $\langle \mathbf{x}, \mathbf{y} \rangle = \sum_i |y_i| = \|\mathbf{y}\|_1$.

(c) For $\|\mathbf{x}\|_2 \le 1$, $\langle \mathbf{x}, \mathbf{y} \rangle \le \|\mathbf{x}\|_2 \|\mathbf{y}\|_2 \le \|\mathbf{y}\|_2$ by Theorem 11.1.5. For $\mathbf{y} \neq \mathbf{0}$, $\mathbf{x} = \mathbf{y}/\|\mathbf{y}\|_2$ gives equality; for $\mathbf{y} = \mathbf{0}$ both sides are 0.

■

So the 1- and ∞-norms are each other's duals, and the Euclidean norm is its own. In the language of polars, (a) and (b) say $(B_1)^\circ = B_\infty$ and $(B_\infty)^\circ = B_1$: the diamond and the square of Chapter 16's picture are polar to each other, and the disc is polar to itself. Part (b) of Proposition 18.4.6 now reads $|\langle \mathbf{x}, \mathbf{y} \rangle| \le \|\mathbf{x}\|_1 \|\mathbf{y}\|_\infty$, the simplest of a family of inequalities. The general member pairs $\|\cdot\|_p$ with $\|\cdot\|_q$, $\frac{1}{p} + \frac{1}{q} = 1$; it is **Hölder's inequality**, and §11 of this chapter proves it and deduces that the dual of $\|\cdot\|_p$ is $\|\cdot\|_q$ for every $1 < p < \infty$. The degenerate case $n = 1$ is worth one line: every norm on $\mathbb{R}$ is $\|x\| = a|x|$ with $a > 0$, its ball is $[-1/a, 1/a]$, and $\|y\|_* = |y|/a$.

A nonexample by minimal change. Keep the formula $\max_K \langle \mathbf{x}, \mathbf{y} \rangle$, and replace the symmetric ball by the triangle $K = \operatorname{conv}\{(-1,-1), (2,-1), (-1,2)\}$, which is compact and convex and has $\mathbf{0}$ inside. By Proposition 18.4.3 (c), $h_K(\mathbf{y})$ is the largest of $-y_1 - y_2$, $2y_1 - y_2$, and $-y_1 + 2y_2$. It is sublinear by Proposition 18.4.3 (a), and positive at every $\mathbf{y} \neq \mathbf{0}$, since the three numbers add up to 0 and are not all zero. But $h_K(\mathbf{e}_1) = 2$ and $h_K(-\mathbf{e}_1) = 1$, so (N2) fails at $c = -1$. Symmetry of the ball is exactly what makes a dual norm a norm.

Why this definition, and not another? The measurement of a functional must be compatible with the measurement of its inputs, in the sense that $|\langle \mathbf{x}, \mathbf{y} \rangle| \le c\|\mathbf{x}\|$ for every $\mathbf{x}$, and Proposition 18.4.6 (b) says that $\|\mathbf{y}\|_*$ is the **smallest** compatible c. Any larger choice would be a bound, not a measurement. Abstractly, the dual norm lives on the dual space: for $\varphi \in (\mathbb{R}^n)^*$ one would put $\|\varphi\| = \max_{\|\mathbf{x}\| \le 1} \varphi(\mathbf{x})$, and the dot product is used only to name φ by a vector, through Theorem 5.1.1. That is why a different inner product on $\mathbb{R}^n$ gives a different formula for "the same" dual norm. The name is standard: the dual space carries the dual norm.

Warning. **A larger norm has a smaller dual.** If $\|\mathbf{x}\|_a \le \|\mathbf{x}\|_b$ for all $\mathbf{x}$, then the ball of $\|\cdot\|_b$ lies inside the ball of $\|\cdot\|_a$, and a maximum over a smaller set is smaller: $\|\mathbf{y}\|_{b*} \le \|\mathbf{y}\|_{a*}$. For instance $N(\mathbf{x}) = 2\|\mathbf{x}\|_2$ has ball of radius $\frac{1}{2}$ and dual $N_*(\mathbf{y}) = \frac{1}{2}\|\mathbf{y}\|_2$, not $2\|\mathbf{y}\|_2$. Scaling a norm up scales its dual down.

Chapter 16 measured vectors and matrices but never a linear functional; Definition 18.4.4 fills that gap, and Proposition 18.4.6 (c) shows it is exactly an operator norm of Chapter 16, the one for $1 \times n$ matrices. The one question it leaves open is whether the construction loses information.

18.4.5 The dual of the dual

Apply the construction twice. Since $\|\cdot\|_*$ is a norm, it has a dual $\|\cdot\|_{**}$, and the natural guess is that we are back where we started. For the three norms of Proposition 18.4.7 this is visible: $1 \to \infty \to 1$ and $2 \to 2 \to 2$. In general it is a theorem, and its proof is a supporting hyperplane.

Theorem 18.4.8. The Dual of the Dual Norm

For every norm $\|\cdot\|$ on $\mathbb{R}^n$, $\|\cdot\|_{**} = \|\cdot\|$.

Idea. One inequality is Cauchy–Schwarz for dual norms. For the other, take $\mathbf{x}$ on the unit sphere of $\|\cdot\|$. It is on the edge of the ball B, so §03 provides a supporting hyperplane of B at $\mathbf{x}$. Its normal $\mathbf{a}$ is a direction in which B reaches no further than $\mathbf{x}$ does, so the maximum defining $\|\mathbf{a}\|_*$ is attained **at** $\mathbf{x}$; rescaling $\mathbf{a}$ produces a functional of dual norm 1 that takes the value 1 at $\mathbf{x}$.

Proof. Let $\mathbf{x} \in \mathbb{R}^n$. For every $\mathbf{y}$ with $\|\mathbf{y}\|_* \le 1$, Proposition 18.4.6 (b) gives $\langle \mathbf{x}, \mathbf{y} \rangle \le \|\mathbf{x}\|\|\mathbf{y}\|_* \le \|\mathbf{x}\|$. By Definition 18.4.4 applied to $\|\cdot\|_*$, whose unit ball is $\{\mathbf{y} : \|\mathbf{y}\|_* \le 1\}$, this says $\|\mathbf{x}\|_{**} \le \|\mathbf{x}\|$.

For the reverse inequality, both sides vanish at $\mathbf{x} = \mathbf{0}$, and both are absolutely homogeneous, so it suffices to take $\|\mathbf{x}\| = 1$. Then $\mathbf{x} \in B$, but $\mathbf{x}$ is not an interior point of B: the vectors $(1+1/k)\mathbf{x}$ have norm $1+1/k > 1$ and $\|(1+1/k)\mathbf{x} - \mathbf{x}\|_2 = \|\mathbf{x}\|_2/k \to 0$. By Theorem 18.3.6, applied to the convex set B and the point $\mathbf{x}$, there is $\mathbf{a} \neq \mathbf{0}$ with $\langle \mathbf{z}, \mathbf{a} \rangle \le \langle \mathbf{x}, \mathbf{a} \rangle$ for every $\mathbf{z} \in B$. Since $\mathbf{x} \in B$, the maximum in Definition 18.4.4 is attained at $\mathbf{x}$:

$$\|\mathbf{a}\|_* = \max_{\mathbf{z} \in B} \langle \mathbf{z}, \mathbf{a} \rangle = \langle \mathbf{x}, \mathbf{a} \rangle.$$

By Proposition 18.4.6 (a), $\|\mathbf{a}\|_* > 0$. Let $\mathbf{y} = \mathbf{a}/\|\mathbf{a}\|_*$. Then $\|\mathbf{y}\|_* = 1$ and $\langle \mathbf{x}, \mathbf{y} \rangle = 1$. Hence $\|\mathbf{x}\|_{**} \ge 1 = \|\mathbf{x}\|$. This proves the theorem.

■

Alternatively, the vector $\mathbf{y}$ can be obtained from Hahn–Banach. Corollary 18.3.9, applied to $U = \operatorname{span}(\mathbf{x})$ and $f(t\mathbf{x}) = t\|\mathbf{x}\|$, gives a functional F with $F(\mathbf{x}) = \|\mathbf{x}\|$ and $|F| \le \|\cdot\|$; writing $F = \langle \cdot, \mathbf{y} \rangle$ by Theorem 5.1.1, the second property says $\|\mathbf{y}\|_* \le 1$ by Proposition 18.4.6 (b), so $\|\mathbf{x}\|_{**} \ge \langle \mathbf{x}, \mathbf{y} \rangle = \|\mathbf{x}\|$. So the analytic and the geometric forms of §03 each prove the theorem, which is one more sign that they are the same fact.

Read with the definition of $\|\cdot\|_{**}$ unwound, the theorem is a formula for the original norm.

Corollary 18.4.9. A Norm Is a Maximum of Linear Functions

For every norm $\|\cdot\|$ on $\mathbb{R}^n$ and every $\mathbf{x} \in \mathbb{R}^n$,

$$\|\mathbf{x}\| = \max\{\langle \mathbf{x}, \mathbf{y} \rangle : \|\mathbf{y}\|_* \le 1\},$$

and when $\mathbf{x} \neq \mathbf{0}$ the maximum is attained at some $\mathbf{y}$ with $\|\mathbf{y}\|_* = 1$.

Proof. The right-hand side is $\|\mathbf{x}\|_{**}$ by Definition 18.4.4 applied to $\|\cdot\|_*$, and it equals $\|\mathbf{x}\|$ by Theorem 18.4.8. For $\mathbf{x} \neq \mathbf{0}$, the proof of Theorem 18.4.8, applied to $\mathbf{x}/\|\mathbf{x}\|$, produced $\mathbf{y}$ with $\|\mathbf{y}\|_* = 1$ and $\langle \mathbf{x}/\|\mathbf{x}\|, \mathbf{y} \rangle = 1$, so $\langle \mathbf{x}, \mathbf{y} \rangle = \|\mathbf{x}\|$.

■

So every norm is the support function of the unit ball of its dual, $\|\cdot\| = h_{\{\mathbf{y}:\|\mathbf{y}\|_*\le 1\}}$: a maximum of the linear functions $\mathbf{x} \mapsto \langle \mathbf{x}, \mathbf{y}\rangle$, one for each $\mathbf{y}$ in a compact set. §10 of this chapter shows that a pointwise maximum of linear functions is always convex, and norms (their convexity was already the triangle inequality) are among its first examples.

Quick check. For $\mathbf{x} = (3, -4)$, find a $\mathbf{y}$ with $\|\mathbf{y}\|_\infty = 1$ and $\langle \mathbf{x}, \mathbf{y}\rangle = \|\mathbf{x}\|_1$, and a $\mathbf{y}$ with $\|\mathbf{y}\|_1 = 1$ and $\langle \mathbf{x}, \mathbf{y}\rangle = \|\mathbf{x}\|_\infty$.

Solution. The dual of $\|\cdot\|_1$ is $\|\cdot\|_\infty$ (Proposition 18.4.7 (a)), so the first $\mathbf{y}$ is a sign vector: $\mathbf{y} = (1, -1)$ gives $3 + 4 = 7 = \|\mathbf{x}\|_1$. The dual of $\|\cdot\|_\infty$ is $\|\cdot\|_1$, so the second $\mathbf{y}$ puts all its weight on the largest coordinate: $\mathbf{y} = (0, -1)$ gives $4 = \|\mathbf{x}\|_\infty$. Each is the witness from the proof of Proposition 18.4.7, with the roles of $\mathbf{x}$ and $\mathbf{y}$ exchanged.

18.4.6 Operator norms as maxima of bilinear values

The corollary also rewrites Chapter 16's operator norm. Fix a norm $\|\cdot\|_\alpha$ on $\mathbb{R}^n$ and a norm $\|\cdot\|_\beta$ on $\mathbb{R}^m$, and write $\|\mathbf{A}\|_{\alpha\to\beta}$ for the operator norm of $\mathbf{A} \in M_{m\times n}(\mathbb{R})$ that they induce (Definition 16.3.2). By the equivalent descriptions after that definition, $\|\mathbf{A}\|_{\alpha\to\beta} = \max\{\|\mathbf{A}\mathbf{x}\|_\beta : \|\mathbf{x}\|_\alpha \le 1\}$.

Proposition 18.4.10. Operator Norms through Dual Norms

Let $\mathbf{A} \in M_{m\times n}(\mathbb{R})$, and let $\|\cdot\|_\alpha$, $\|\cdot\|_\beta$ be norms on $\mathbb{R}^n$ and $\mathbb{R}^m$, with duals $\|\cdot\|_{\alpha*}$ and $\|\cdot\|_{\beta*}$. Then

(a) $\|\mathbf{A}\|_{\alpha\to\beta} = \max\{\mathbf{y}^\top \mathbf{A}\mathbf{x} : \|\mathbf{x}\|_\alpha \le 1,\ \|\mathbf{y}\|_{\beta*} \le 1\}$;
(b) $\|\mathbf{A}^\top\|_{\beta*\to\alpha*} = \|\mathbf{A}\|_{\alpha\to\beta}$.

Idea. Write the outer norm $\|\mathbf{A}\mathbf{x}\|_\beta$ as a maximum over its dual ball by Corollary 18.4.9, and the operator norm becomes one maximum over pairs $(\mathbf{x}, \mathbf{y})$. Since $\mathbf{y}^\top \mathbf{A}\mathbf{x} = \mathbf{x}^\top \mathbf{A}^\top \mathbf{y}$, the same double maximum, taken with $\mathbf{x}$ innermost, is the operator norm of $\mathbf{A}^\top$, which gives (b).

Proof.

(a) For $\|\mathbf{x}\|_\alpha \le 1$, Corollary 18.4.9 in $\mathbb{R}^m$ gives $\|\mathbf{A}\mathbf{x}\|_\beta = \max\{\mathbf{y}^\top \mathbf{A}\mathbf{x} : \|\mathbf{y}\|_{\beta*} \le 1\}$. So every value $\mathbf{y}^\top \mathbf{A}\mathbf{x}$ in the set is at most $\|\mathbf{A}\mathbf{x}\|_\beta \le \|\mathbf{A}\|_{\alpha\to\beta}$. Conversely, choose $\mathbf{x}_0$ with $\|\mathbf{x}_0\|_\alpha \le 1$ and $\|\mathbf{A}\mathbf{x}_0\|_\beta = \|\mathbf{A}\|_{\alpha\to\beta}$, which exists by Lemma 16.3.1, and then $\mathbf{y}_0$ attaining the maximum for $\|\mathbf{A}\mathbf{x}_0\|_\beta$. The value $\mathbf{y}_0^\top \mathbf{A}\mathbf{x}_0$ equals $\|\mathbf{A}\|_{\alpha\to\beta}$, so the maximum exists and has this value.

(b) By the equivalent descriptions after Definition 16.3.2, and then Definition 18.4.4 for $\|\cdot\|_{\alpha*}$,

$$\|\mathbf{A}^\top\|_{\beta*\to\alpha*} = \max_{\|\mathbf{y}\|_{\beta*}\le 1} \|\mathbf{A}^\top \mathbf{y}\|_{\alpha*} = \max_{\|\mathbf{y}\|_{\beta*}\le 1} \max_{\|\mathbf{x}\|_\alpha \le 1} \langle \mathbf{x}, \mathbf{A}^\top \mathbf{y}\rangle.$$

Since $\langle \mathbf{x}, \mathbf{A}^\top \mathbf{y} \rangle = \mathbf{x}^\top \mathbf{A}^\top \mathbf{y} = \mathbf{y}^\top \mathbf{A}\mathbf{x}$, this is the maximum in (a).

■

With $\alpha = \beta = \infty$, part (b) and Proposition 18.4.7 say $\|\mathbf{A}^\top\|_1 = \|\mathbf{A}\|_\infty$: the largest column sum of $\mathbf{A}^\top$ is the largest row sum of $\mathbf{A}$, which is also what Theorem 16.3.3 gives directly. With $\alpha = \beta = 2$ it says $\|\mathbf{A}^\top\|_2 = \|\mathbf{A}\|_2$, that is, $\sigma_1(\mathbf{A}^\top) = \sigma_1(\mathbf{A})$. For the matrix $\mathbf{A} = \left(\begin{smallmatrix} 3 & 0 \\ 4 & 5 \end{smallmatrix}\right)$ of Example (One matrix, four numbers), $\|\mathbf{A}\|_1 = 7$ and $\|\mathbf{A}\|_\infty = 9$. Its transpose $\left(\begin{smallmatrix} 3 & 4 \\ 0 & 5 \end{smallmatrix}\right)$ has column sums 3 and 9 and row sums 7 and 5, so $\|\mathbf{A}^\top\|_1 = 9 = \|\mathbf{A}\|_\infty$ and $\|\mathbf{A}^\top\|_\infty = 7 = \|\mathbf{A}\|_1$, as (b) predicts.

18.4.7 Exercises

18.4.7.1 A. Check your understanding

Exercise. A1

(a) Define the gauge p_K, the support function h_K, and the dual norm $\|\cdot\|_*$, stating what each needs of the set or norm involved.

(b) Determine whether the following statement is correct, and justify your answer: the support function of every nonempty bounded subset of $\mathbb{R}^2$ is a norm.

(c) What is the dual of $\|\cdot\|_\infty$ on $\mathbb{R}^n$? What is the dual of that?

(d) Which result of §03 does the proof of Theorem 18.4.8 use, and at which point?

(e) Determine whether the following statement is correct, and justify your answer: $K^{\circ\circ} = K$ for every nonempty closed convex $K \subseteq \mathbb{R}^n$.

Solution.

(a) For K convex with $\mathbf{0}$ an interior point, $p_K(\mathbf{x}) = \inf\{t > 0 : \mathbf{x} \in tK\}$ (Definition 18.4.1). For K nonempty, $h_K(\mathbf{y}) = \sup_{\mathbf{x} \in K} \langle \mathbf{x}, \mathbf{y} \rangle$, possibly $+\infty$ (Definition 18.4.2). For a norm $\|\cdot\|$ with ball B, $\|\mathbf{y}\|_* = \max_{\mathbf{x} \in B} \langle \mathbf{x}, \mathbf{y} \rangle = h_B(\mathbf{y})$ (Definition 18.4.4).

(b) Incorrect. For $K = \{\mathbf{0}\}$, $h_K = 0$, which violates (N1). For the segment $[-\mathbf{e}_1, \mathbf{e}_1]$, $h_K(\mathbf{y}) = |y_1|$ vanishes at $\mathbf{e}_2$. Support functions are sublinear (Proposition 18.4.3 (a)) but need not be norms.

(c) $\|\cdot\|_1$, by Proposition 18.4.7 (b); and the dual of $\|\cdot\|_1$ is $\|\cdot\|_\infty$ again, by (a) of the same proposition, as Theorem 18.4.8 requires.

(d) Theorem 18.3.6, applied to the unit ball B at a point $\mathbf{x}$ with $\|\mathbf{x}\| = 1$, to produce a normal $\mathbf{a}$ at which the maximum defining $\|\mathbf{a}\|_*$ is attained at $\mathbf{x}$.

(e) Incorrect. For $K = \{\mathbf{e}_1\}$ in $\mathbb{R}^2$, $K^{\circ\circ}$ is the segment $[\mathbf{0}, \mathbf{e}_1]$. Theorem 18.4.5 needs $\mathbf{0} \in K$.

18.4.7.2 B. Practice

Exercise. B1: A weighted norm and its dual

On $\mathbb{R}^2$ let $N(\mathbf{x}) = \max(|x_1|, 2|x_2|)$.

(a) Prove that $N_*(\mathbf{y}) = |y_1| + \frac{1}{2}|y_2|$.

(b) Verify Theorem 18.4.8 for N by computing N_{**} directly.

Solution.

(a) The ball of N is the rectangle $|x_1| \le 1$, $|x_2| \le \frac{1}{2}$. On it, $\langle \mathbf{x}, \mathbf{y} \rangle \le |x_1||y_1| + |x_2||y_2| \le |y_1| + \frac{1}{2}|y_2|$. The point $\mathbf{x} = (s_1, \frac{1}{2}s_2)$, where $s_i = 1$ if $y_i \ge 0$ and $s_i = -1$ otherwise, lies in the ball and gives equality. So $N_*(\mathbf{y}) = |y_1| + \frac{1}{2}|y_2|$.

(b) The ball of N_* is $\{|y_1| + \frac{1}{2}|y_2| \le 1\}$. On it,

$$\langle \mathbf{x}, \mathbf{y} \rangle \le |x_1||y_1| + 2|x_2| \cdot \tfrac{1}{2}|y_2| \le N(\mathbf{x})(|y_1| + \tfrac{1}{2}|y_2|) \le N(\mathbf{x}).$$

If $N(\mathbf{x}) = |x_1|$, the point $\mathbf{y} = \pm \mathbf{e}_1$ (sign of x_1) lies in the ball and gives $\langle \mathbf{x}, \mathbf{y} \rangle = |x_1|$. If $N(\mathbf{x}) = 2|x_2|$, the point $\mathbf{y} = \pm 2\mathbf{e}_2$ (sign of x_2) lies in the ball, since $\frac{1}{2} \cdot 2 = 1$, and gives $\langle \mathbf{x}, \mathbf{y} \rangle = 2|x_2|$. So $N_{**} = N$.

Exercise. B2: A hexagonal norm

Let $K = \{\mathbf{x} \in \mathbb{R}^2 : |x_1| \le 1,\ |x_2| \le 1,\ |x_1 - x_2| \le 1\}$.

(a) Check that K satisfies the hypotheses of Theorem 18.4.2 (a).

(b) Show that $p_K(\mathbf{x}) = \max(|x_1|, |x_2|, |x_1 - x_2|)$, and compute $p_K((2, -1))$.

(c) Determine whether p_K is induced by an inner product. Justify your answer.

Solution.

(a) K is the intersection of six closed half-planes, such as $\{x_1 \le 1\}$ and $\{x_1 - x_2 \ge -1\}$, each closed and convex (§03, before Definition 18.3.1); so K is closed and convex. It is bounded, since $\|\mathbf{x}\|_2 \le \sqrt{2}$ when $|x_1|, |x_2| \le 1$. It is symmetric, since each condition is unchanged by $\mathbf{x} \mapsto -\mathbf{x}$. And if $\|\mathbf{x}\|_2 < \frac{1}{2}$, then $|x_1|, |x_2| < \frac{1}{2}$ and $|x_1 - x_2| < 1$, so $\mathbf{x} \in K$ and $\mathbf{0}$ is an interior point.

(b) For $t > 0$, $\mathbf{x} \in tK$ exactly when $\mathbf{x}/t \in K$, that is, when $|x_1| \le t$, $|x_2| \le t$, and $|x_1 - x_2| \le t$. The infimum of such t is $M = \max(|x_1|, |x_2|, |x_1 - x_2|)$ when $\mathbf{x} \neq \mathbf{0}$, and $0 = M$ when $\mathbf{x} = \mathbf{0}$. So $p_K((2, -1)) = \max(2, 1, 3) = 3$.

(c) No. With $\mathbf{u} = \mathbf{e}_1$ and $\mathbf{v} = \mathbf{e}_2$, $p_K(\mathbf{u} + \mathbf{v}) = \max(1, 1, 0) = 1$ and $p_K(\mathbf{u} - \mathbf{v}) = \max(1, 1, 2) = 2$, so the left side of the parallelogram law is $1 + 4 = 5$, while the right side is $2 \cdot 1 + 2 \cdot 1 = 4$. By Theorem 16.1.3, no inner product induces p_K.

Exercise. B3: The polar of a triangle

Let $K = \operatorname{conv}\{(-1, -1), (2, -1), (-1, 2)\} \subseteq \mathbb{R}^2$.

(a) Compute $h_K(\mathbf{e}_1)$, $h_K(-\mathbf{e}_1)$, and $h_K((1, 1))$.

(b) Describe K° by three linear inequalities.

(c) Show that $(-1, 0)$, $(0, -1)$, and $(1, 1)$ lie in K°, and that $(1, 0)$ does not.

Solution.

(a) By Proposition 18.4.3 (c), $h_K(\mathbf{y})$ is the largest of $\langle \mathbf{v}, \mathbf{y} \rangle$ over the three vertices $\mathbf{v}$, that is, of $-y_1 - y_2$, $2y_1 - y_2$, $-y_1 + 2y_2$. So $h_K(\mathbf{e}_1) = \max(-1, 2, -1) = 2$, $h_K(-\mathbf{e}_1) = \max(1, -2, 1) = 1$, and $h_K((1,1)) = \max(-2, 1, 1) = 1$.

(b) $\mathbf{y} \in K^\circ$ exactly when $h_K(\mathbf{y}) \le 1$, that is,

$$-y_1 - y_2 \le 1, \qquad 2y_1 - y_2 \le 1, \qquad -y_1 + 2y_2 \le 1.$$

(c) At $(-1, 0)$ the three left sides are $1, -2, 1$; at $(0, -1)$ they are $1, 1, -2$; at $(1, 1)$ they are $-2, 1, 1$. All are ≤ 1. At $(1, 0)$ the second is $2 > 1$, so $(1, 0) \notin K^\circ$, in agreement with $h_K(\mathbf{e}_1) = 2$.

18.4.7.3 C. Going deeper

Exercise. C1: The norm from 1 to infinity

Let $\mathbf{A} \in M_{m \times n}(\mathbb{R})$, and let $\|\mathbf{A}\|_{1 \to \infty}$ be the operator norm with $\|\cdot\|_1$ on $\mathbb{R}^n$ and $\|\cdot\|_\infty$ on $\mathbb{R}^m$.

(a) Using Proposition 18.4.10 (a), prove that $\|\mathbf{A}\|_{1 \to \infty} = \max_{i,j} |a_{ij}|$.

(b) Deduce from Proposition 18.4.10 (b) that the operator norm of $\mathbf{A}^\top$ with $\|\cdot\|_1$ on $\mathbb{R}^m$ and $\|\cdot\|_\infty$ on $\mathbb{R}^n$ is the same number.

(c) Explain why this does not contradict the warning of Chapter 16 §03 that the entrywise maximum is not a matrix norm.

Solution.

(a) By Proposition 18.4.7 (b), the dual of $\|\cdot\|_\infty$ is $\|\cdot\|_1$, so Proposition 18.4.10 (a) gives $\|\mathbf{A}\|_{1 \to \infty} = \max\{\mathbf{y}^\top \mathbf{A} \mathbf{x} : \|\mathbf{x}\|_1 \le 1,\ \|\mathbf{y}\|_1 \le 1\}$. Write $M = \max_{i,j} |a_{ij}|$. For such $\mathbf{x}, \mathbf{y}$,

$$\mathbf{y}^\top \mathbf{A} \mathbf{x} = \sum_{i,j} y_i a_{ij} x_j \le M \sum_i |y_i| \sum_j |x_j| \le M.$$

If $|a_{kl}| = M$, then $\mathbf{x} = \mathbf{e}_l$ and $\mathbf{y} = \pm \mathbf{e}_k$, with the sign of a_{kl}, give $\mathbf{y}^\top \mathbf{A} \mathbf{x} = M$. So $\|\mathbf{A}\|_{1 \to \infty} = M$.

(b) By Proposition 18.4.10 (b) with $\alpha = 1$ and $\beta = \infty$, $\|\mathbf{A}^\top\|_{\beta* \to \alpha*} = \|\mathbf{A}^\top\|_{1 \to \infty}$ equals $\|\mathbf{A}\|_{1 \to \infty}$, using $\beta* = 1$ and $\alpha* = \infty$ from Proposition 18.4.7. This is consistent with (a), since $\mathbf{A}$ and $\mathbf{A}^\top$ have the same entries.

(c) $\|\cdot\|_{1 \to \infty}$ uses **different** norms on the source and the target, so the product $\mathbf{AB}$ of two square matrices is measured by a norm that does not compose: submultiplicativity in Theorem 16.3.2 (d) needs the target norm of $\mathbf{B}$ to be the source norm of $\mathbf{A}$. With one norm on $M_n(\mathbb{R})$ the entrywise maximum is not submultiplicative (Example (The entrywise maximum is not submultiplicative)), and nothing here says otherwise.

Exercise. C2: Equivalence constants dualize

Let N_1, N_2 be norms on $\mathbb{R}^n$ and $0 < m \le M$ constants with $m N_1(\mathbf{x}) \le N_2(\mathbf{x}) \le M N_1(\mathbf{x})$ for every $\mathbf{x}$.

(a) Prove that $\frac{1}{M}N_{1*}(\mathbf{y}) \le N_{2*}(\mathbf{y}) \le \frac{1}{m}N_{1*}(\mathbf{y})$ for every $\mathbf{y}$.

(b) Apply (a) to $\|\mathbf{x}\|_2 \le \|\mathbf{x}\|_1 \le \sqrt{n}\,\|\mathbf{x}\|_2$ from Proposition 16.1.2, and compare the result with that proposition.

Solution.

(a) If $N_2(\mathbf{x}) \le 1$, then $N_1(\mathbf{x}) \le 1/m$, so $\mathbf{x} = \frac{1}{m}\mathbf{x}'$ with $N_1(\mathbf{x}') \le 1$, and $\langle \mathbf{x}, \mathbf{y}\rangle = \frac{1}{m}\langle \mathbf{x}', \mathbf{y}\rangle \le \frac{1}{m}N_{1*}(\mathbf{y})$. Taking the maximum over the ball of N_2 gives $N_{2*}(\mathbf{y}) \le \frac{1}{m}N_{1*}(\mathbf{y})$. If $N_1(\mathbf{x}) \le 1$, then $N_2(\mathbf{x}) \le M$, and Proposition 18.4.6 (b) gives $\langle \mathbf{x}, \mathbf{y}\rangle \le N_2(\mathbf{x})N_{2*}(\mathbf{y}) \le MN_{2*}(\mathbf{y})$; so $N_{1*}(\mathbf{y}) \le MN_{2*}(\mathbf{y})$.

(b) Take $N_1 = \|\cdot\|_2$, $N_2 = \|\cdot\|_1$, $m = 1$, $M = \sqrt{n}$. The duals are $\|\cdot\|_2$ and $\|\cdot\|_\infty$ (Proposition 18.4.7), so (a) gives $\frac{1}{\sqrt{n}}\|\mathbf{y}\|_2 \le \|\mathbf{y}\|_\infty \le \|\mathbf{y}\|_2$. These are exactly the other two inequalities $\|\mathbf{y}\|_\infty \le \|\mathbf{y}\|_2 \le \sqrt{n}\,\|\mathbf{y}\|_\infty$ of Proposition 16.1.2: that proposition is closed under duality, with the constants carried over unchanged.

Exercise. C3: The gauge is the support function of the polar

Let $K \subseteq \mathbb{R}^n$ be closed and convex, with $\mathbf{0}$ an interior point.

(a) Prove that $p_K(\mathbf{x}) = h_{K^\circ}(\mathbf{x})$ for every $\mathbf{x}$. *Hint: for $t > 0$, rewrite "$\mathbf{x} \in tK$" using Theorem 18.4.5.*

(b) Deduce Theorem 18.4.8 again from (a), Theorem 18.4.2 (b) and Proposition 18.4.6 (d).

Solution.

(a) The polar K° contains $\mathbf{0}$, so $h_{K^\circ}(\mathbf{x}) \ge \langle \mathbf{x}, \mathbf{0}\rangle = 0$. Let $t > 0$. By Theorem 18.4.5, $K = K^{\circ\circ}$, so

$$\mathbf{x} \in tK \iff \mathbf{x}/t \in K^{\circ\circ} \iff \langle \mathbf{x}, \mathbf{y}\rangle \le t \text{ for every } \mathbf{y} \in K^\circ \iff h_{K^\circ}(\mathbf{x}) \le t.$$

So the admissible t in Definition 18.4.1 are the $t > 0$ with $t \ge h_{K^\circ}(\mathbf{x})$, and their infimum is $h_{K^\circ}(\mathbf{x})$ (if $h_{K^\circ}(\mathbf{x}) = 0$, the admissible t are all $t > 0$, with infimum 0). In particular $h_{K^\circ}(\mathbf{x})$ is finite.

(b) Let B be the ball of $\|\cdot\|$. By Theorem 18.4.2 (b), B is closed and convex with $\mathbf{0}$ interior, and $\|\cdot\| = p_B$. By (a), $p_B = h_{B^\circ}$, and B° is the ball of $\|\cdot\|_*$ by Proposition 18.4.6 (d). So $\|\mathbf{x}\| = \max\{\langle \mathbf{x}, \mathbf{y}\rangle : \|\mathbf{y}\|_* \le 1\} = \|\mathbf{x}\|_{**}$, the maximum existing by Definition 18.4.4 applied to $\|\cdot\|_*$.

18.5 Cones and Farkas's Lemma

Chapter 3 answered one question about a linear system completely: $\mathbf{Ax} = \mathbf{b}$ has a solution exactly when $\mathbf{b}$ lies in the column space of $\mathbf{A}$ (Theorem 3.1.1). Its applications section then hit a wall. Traffic flows cannot be negative, and once the unknowns are required to satisfy $\mathbf{x} \ge \mathbf{0}$, the column space is the wrong set to look in. This section finds the right set, which is a **cone**, and answers the new question with the separation theorems of §03: either $\mathbf{Ax} = \mathbf{b}$ has a nonnegative solution, or a single vector $\mathbf{y}$ certifies that it has none. That is Farkas's lemma,

and the next two sections run on it.

One step is harder than it looks. Separation needs a **closed** set, and nothing proved so far says that the relevant cone is closed. Most of the work goes into that step, and the last part of the section shows, with an explicit cone, that it can really fail.

Throughout, the field is $\mathbb{R}$. Vectors in $\mathbb{R}^n$ are columns, $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{y}^\top \mathbf{x}$ is the standard inner product, and $\|\cdot\|$ is the Euclidean norm it induces, the norm that Theorem 18.3.1 uses. Convergence $\mathbf{y}_k \to \mathbf{y}$ means $\|\mathbf{y}_k - \mathbf{y}\| \to 0$, and "closed" is meant in the sequential sense of Definition 18.1.6: a set S is closed when every limit of a convergent sequence of points of S again lies in S. Convergence in this norm is entrywise convergence, by the elementary bound

$$|y_i| \;\le\; \|\mathbf{y}\| \;\le\; |y_1| + \cdots + |y_n| \qquad \text{for } \mathbf{y} \in \mathbb{R}^n \text{ and } 1 \le i \le n. \tag{18.5.1}$$

The left inequality holds because $y_i^2 \le y_1^2 + \cdots + y_n^2$, and the right one is the triangle inequality for $\mathbf{y} = y_1\mathbf{e}_1 + \cdots + y_n\mathbf{e}_n$. Applied to $\mathbf{y}_k - \mathbf{y}$, with the algebra of limits of Chapter 16's introduction on the right, it shows that $\|\mathbf{y}_k - \mathbf{y}\| \to 0$ exactly when every entry of $\mathbf{y}_k - \mathbf{y}$ tends to 0.

18.5.1 Convex cones

For a vector with real entries we write $\mathbf{x} \ge \mathbf{0}$ when **every entry** of $\mathbf{x}$ is ≥ 0, and $\mathbf{x} > \mathbf{0}$ when every entry is > 0. The set

$$\mathbb{R}^m_{\ge 0} = \{\mathbf{x} \in \mathbb{R}^m : \mathbf{x} \ge \mathbf{0}\}$$

is the **nonnegative orthant**. The relation $\mathbf{x} \ge \mathbf{0}$ is an entrywise statement about a vector, and it has nothing to do with the Loewner order $\mathbf{A} \succeq \mathbf{O}$ of Chapter 13, which is a statement about a matrix.

Here is the question. Let $\mathbf{A} \in M_{m\times n}(\mathbb{R})$ have columns $\mathbf{a}_1, \ldots, \mathbf{a}_n$. By Theorem 0.9.3, $\mathbf{Ax} = x_1\mathbf{a}_1 + \cdots + x_n\mathbf{a}_n$, so

$$\{\mathbf{b} \in \mathbb{R}^m : \mathbf{Ax} = \mathbf{b} \text{ for some } \mathbf{x} \ge \mathbf{0}\} = \{x_1\mathbf{a}_1 + \cdots + x_n\mathbf{a}_n : \text{every } x_j \ge 0\}.$$

Without the sign condition this set would be the span of the columns, a subspace. With it, the set is closed under addition and under multiplication by **nonnegative** scalars, but not under multiplication by -1. Such sets turn up whenever quantities cannot be negative, so they get a name.

A cone is a set you may stretch and add in, but never reflect.

Definition 18.5.1. Convex Cone

Let V be a vector space over $\mathbb{R}$. A subset $K \subseteq V$ is a **convex cone** if K is **nonempty** and

$$au + bv \in K \qquad \text{for all } u, v \in K \text{ and all real } a, b \ge 0.$$

In words: a convex cone contains every **nonnegative** combination of two of its members, and hence, by induction on the number of terms, every nonnegative combination of finitely many of its members. Three special cases unpack the single condition. Taking $a = b = 0$ shows $0 \in K$, so every convex cone contains the origin. Taking $b = 0$ shows that K is closed under multiplication by nonnegative scalars. Taking $a = b = 1$ shows that it is closed under addition, and conversely a nonempty set with these two closure properties is a convex cone.

A convex cone is a convex set in the sense of Definition 18.1.1: take $a = 1 - t$ and $b = t$ with $0 \le t \le 1$. The converse fails, and the next examples show how.

Examples. Each is checked against the one condition of the definition.

- **The orthant.** If $\mathbf{u}, \mathbf{v} \ge \mathbf{0}$ and $a, b \ge 0$, then every entry of $a\mathbf{u} + b\mathbf{v}$ is a sum of products of nonnegative numbers, so $a\mathbf{u} + b\mathbf{v} \ge \mathbf{0}$. Hence $\mathbb{R}^m_{\ge 0}$ is a convex cone.
- **A half-space through the origin.** Fix $\mathbf{a} \in \mathbb{R}^n$ and let $H = \{\mathbf{x} : \langle \mathbf{x}, \mathbf{a} \rangle \ge 0\}$. If $\mathbf{u}, \mathbf{v} \in H$ and $a, b \ge 0$, then $\langle a\mathbf{u} + b\mathbf{v}, \mathbf{a} \rangle = a\langle \mathbf{u}, \mathbf{a} \rangle + b\langle \mathbf{v}, \mathbf{a} \rangle \ge 0$. So H is a convex cone. When $\mathbf{a} \ne \mathbf{0}$ it is one of the two closed half-spaces bounded by the hyperplane $\mathbf{a}^\perp$.
- **The ice-cream cone.** In $\mathbb{R}^{n+1}$, write vectors as $(\mathbf{x}, t)$ with $\mathbf{x} \in \mathbb{R}^n$ and $t \in \mathbb{R}$, and let $L = \{(\mathbf{x}, t) : \|\mathbf{x}\|_2 \le t\}$. If $(\mathbf{x}, t), (\mathbf{y}, s) \in L$ and $a, b \ge 0$, the triangle inequality and homogeneity of the norm (Definition 16.1.1) give $\|a\mathbf{x} + b\mathbf{y}\|_2 \le a\|\mathbf{x}\|_2 + b\|\mathbf{y}\|_2 \le at + bs$. So L is a convex cone. In $\mathbb{R}^3$ it is the round solid cone that gives the family its name.
- **Positive semidefinite matrices.** In the real vector space of real symmetric $n \times n$ matrices, the set of $\mathbf{A} \succeq \mathbf{O}$ is a convex cone. This is part (a) of Exercise (C2: The cone of positive matrices). Chapter 13 used the phrase there, and this is the definition it was using.
- **Degenerate cases.** $\{0\}$ and V are convex cones. More generally, **every subspace is a convex cone**, since it is closed under all linear combinations and in particular the nonnegative ones. A subspace is the extreme case of a cone that happens to be closed under reflection as well.

Nonexamples by minimal change. In $\mathbb{R}^2$, take the union of two opposite quadrants, $Q = \{\mathbf{x} : x_1 x_2 \ge 0\}$. It is closed under nonnegative scaling, since $(ax_1)(ax_2) = a^2 x_1 x_2$. The clause that fails is addition: $(1, 0)$ and $(0, -1)$ lie in Q, but their sum $(1, -1)$ does not. Or keep convexity and lose the scaling: the closed unit disc is convex, but $2\mathbf{e}_1$ is not in it. Or translate the orthant to $\mathbf{1} + \mathbb{R}^2_{\ge 0}$: it is convex and closed under addition, but it does not contain $\mathbf{0}$.

Why this definition. The scalars must be nonnegative. With all real scalars allowed, the definition becomes the definition of a subspace, and the sign information that started this section would be lost. The requirement $0 \in K$ is a convention that some authors drop: they call $\{\mathbf{x} : \mathbf{x} > \mathbf{0}\}$ a cone as well. Including the origin costs nothing here and keeps the statements below free of exceptions.

Warning. **A convex cone need not be closed.** In $\mathbb{R}^2$ let

$$C = \{(x, y) : x > 0\} \cup \{(0, 0)\}.$$

It is a convex cone. Let $\mathbf{u}, \mathbf{v} \in C$ and $a, b \ge 0$. The first entry of $a\mathbf{u} + b\mathbf{v}$ is $au_1 + bv_1 \ge 0$. If it is positive, the vector lies in C. If it is zero, then $au_1 = 0$ and $bv_1 = 0$; since the only member of C with first entry 0 is $\mathbf{0}$, this forces $a\mathbf{u} = \mathbf{0}$ and $b\mathbf{v} = \mathbf{0}$, and the sum is $\mathbf{0} \in C$. But $(1/k, 1) \in C$ for every $k \ge 1$, while the limit $(0, 1)$ is not in C. Separating a point from a cone needs the cone to be closed, so closedness must be proved. This particular cone comes back at the end of the section as the image of a closed cone.

The cones in the question at the start of this section are the simplest kind.

Definition 18.5.2. Finitely Generated Cone

Let $a_1, \dots, a_m$ be vectors in a real vector space V. The **cone generated by** them is

$$\operatorname{cone}(a_1, \dots, a_m) = \left\{ \sum_{i=1}^{m} x_i a_i : x_1, \dots, x_m \geq 0 \right\}.$$

A cone of this form is **finitely generated**. For the empty list we set $\operatorname{cone}() = \{0\}$, the value of the empty sum. When $V = \mathbb{R}^n$ and $\mathbf{A}$ is the matrix with columns $\mathbf{a}_1, \dots, \mathbf{a}_m$, this is $\{\mathbf{Ax} : \mathbf{x} \in \mathbb{R}^m_{\geq 0}\}$, and we also write it $\operatorname{cone}(\mathbf{A})$.

It is a convex cone: a nonnegative combination of two nonnegative combinations of the a_i is again one, with coefficients $ax_i + bx'_i \geq 0$. It is also the **smallest** convex cone containing $a_1, \dots, a_m$, since any convex cone containing the a_i contains all their nonnegative combinations, by the remark after Definition 18.5.1. The orthant is $\operatorname{cone}(\mathbf{e}_1, \dots, \mathbf{e}_m)$. A subspace with basis $v_1, \dots, v_k$ is $\operatorname{cone}(v_1, -v_1, \dots, v_k, -v_k)$, since cv_i is cv_i or $|c|(-v_i)$ according to the sign of c. The ice-cream cone is **not** finitely generated, a fact we shall neither need nor prove.

Quick check. Is $\operatorname{cone}((1,0),(1,1),(0,1))$ the same as $\operatorname{cone}((1,0),(0,1))$? Is $\operatorname{cone}((1,0),(-1,0))$ a line or a half-line?

Solution. Both cones are the quadrant $\mathbb{R}^2_{\geq 0}$. The second list's cone is the quadrant because $x_1\mathbf{e}_1 + x_2\mathbf{e}_2 = (x_1, x_2)$. The first contains the second, and every generator of the first lies in the quadrant, which is a convex cone, so the first is contained in the quadrant too. A redundant generator changes nothing. The cone of $(1,0)$ and $(-1,0)$ is the whole x-axis $\{(t,0) : t \in \mathbb{R}\}$, a line: for $t \geq 0$ use $t(1,0)$, and for $t < 0$ use $|t|(-1,0)$.

18.5.2 The dual cone

A certificate that $\mathbf{Ax} = \mathbf{b}$ has no nonnegative solution will be a vector $\mathbf{y}$ that makes a nonnegative inner product with every column of $\mathbf{A}$ but a negative one with $\mathbf{b}$. Every nonnegative combination of the columns then also has a nonnegative inner product with $\mathbf{y}$, so $\mathbf{b}$ cannot be such a combination. The set of vectors $\mathbf{y}$ that pass the first test deserves a name.

The dual cone of K is the set of directions that make an angle of at most 90° with everything in K.

Definition 18.5.3. Dual Cone

Let V be a finite-dimensional real inner product space and let $S \subseteq V$ be **any** subset. The **dual cone** of S is

$$S^* = \{y \in V : \langle x, y \rangle \geq 0 \text{ for every } x \in S\}.$$

We write S^{**} for $(S^*)^*$.

In words: $y \in S^*$ when the linear functional $\langle \cdot, y \rangle$ is nonnegative on the whole of S, that is, when S lies in the closed half-space $\{x : \langle x, y \rangle \geq 0\}$. The inner product is what places S^* inside V rather than in the dual space of Chapter 5 (Theorem 5.1.1 is the case $V = \mathbb{R}^n$).

Warning. S^* **is not the dual space.** The star here produces a subset of V itself, not a space of functionals, and it is defined for any subset, not only for a subspace. For a subspace

U the dual cone is the orthogonal complement: if $\langle u, y\rangle \geq 0$ for every $u \in U$, then applying this to $-u \in U$ gives $\langle u, y\rangle \leq 0$ as well, so $\langle u, y\rangle = 0$, and the converse is immediate. So for a subspace we always write $U^{\perp}$, never U^*.

The basic properties take one line each, and they explain why every dual cone is a good set to separate from.

Proposition 18.5.1. Basic Properties of the Dual Cone

Let V be a finite-dimensional real inner product space and let $S, T \subseteq V$.

(a) S^* is a closed convex cone.

(b) If $S \subseteq T$, then $T^* \subseteq S^*$.

(c) $S \subseteq S^{**}$.

(d) $\operatorname{cone}(a_1, \ldots, a_m)^* = \{y : \langle a_i, y\rangle \geq 0 \text{ for } i = 1, \ldots, m\}$. In $\mathbb{R}^n$, with $\mathbf{A}$ the matrix of columns $\mathbf{a}_i$, this is $\{\mathbf{y} : \mathbf{A}^{\top}\mathbf{y} \geq \mathbf{0}\}$.

Idea. S^* is the intersection of the sets $\{y : \langle x, y\rangle \geq 0\}$, one for each $x \in S$. Parts (a), (b) and (d) read off this description (more members of S mean more conditions, and for a cone the generators suffice), and (c) is the symmetry of the inner product.

Proof.

(a) $0 \in S^*$, so S^* is nonempty. If $y, z \in S^*$ and $a, b \geq 0$, then $\langle x, ay + bz\rangle = a\langle x, y\rangle + b\langle x, z\rangle \geq 0$ for every $x \in S$, using linearity in the second slot, which holds because the field is $\mathbb{R}$. Finally $S^* = \bigcap_{x \in S}\{y : \langle x, y\rangle \geq 0\}$ is an intersection of closed half-spaces (all of V when $x = 0$), closed by Proposition 18.1.6 (c) and (b), since $\langle x, \cdot\rangle$ is a linear functional; if $S = \emptyset$, then $S^* = V$, which is closed.

(b) A vector that is nonnegative against every member of T is in particular nonnegative against every member of S.

(c) If $x \in S$, then $\langle x, y\rangle \geq 0$ for every $y \in S^*$, by the definition of S^*. Since the inner product is symmetric over $\mathbb{R}$, this says $\langle y, x\rangle \geq 0$ for every $y \in S^*$, which is $x \in S^{**}$.

(d) The generators lie in the cone, so (b) gives $\subseteq$. Conversely, if $\langle a_i, y\rangle \geq 0$ for every i, then $\langle \sum x_i a_i, y\rangle = \sum x_i \langle a_i, y\rangle \geq 0$ whenever every $x_i \geq 0$. The matrix form follows because the i-th entry of $\mathbf{A}^{\top}\mathbf{y}$ is $\mathbf{a}_i^{\top}\mathbf{y} = \langle \mathbf{a}_i, \mathbf{y}\rangle$.

■

Part (a) holds for **every** set S, including one that is neither convex nor closed. Part (c) is only an inclusion, and the question of when it is an equality is really the question of this section.

Examples.

- **The orthant is its own dual.** By (d), $(\mathbb{R}^m_{\geq 0})^* = \operatorname{cone}(\mathbf{e}_1, \ldots, \mathbf{e}_m)^* = \{\mathbf{y} : \langle \mathbf{e}_i, \mathbf{y}\rangle = y_i \geq 0 \text{ for all } i\} = \mathbb{R}^m_{\geq 0}$.
- **A single generator.** $\operatorname{cone}(\mathbf{a})^* = \{\mathbf{y} : \langle \mathbf{a}, \mathbf{y}\rangle \geq 0\}$, a closed half-space when $\mathbf{a} \neq \mathbf{0}$. A thin cone has a fat dual.

- **The degenerate cases.** The dual cone of $\{0\}$ is all of V, and the dual cone of V is $\{0\}$, because a y in it must satisfy $\langle -y, y\rangle \geq 0$, that is $\|y\|^2 \leq 0$. As subspaces, by the warning, we write these as $\{0\}^\perp = V$ and $V^\perp = \{0\}$.
- **A wedge in the plane.** For $K = \operatorname{cone}((1,0),(1,1))$, part (d) gives $K^* = \{\mathbf{y} : y_1 \geq 0,\ y_1 + y_2 \geq 0\}$. This is the wedge between the directions $(0,1)$ and $(1,-1)$, each of which is perpendicular to one edge of K. Exercise (B3: A dual cone and its dual) finishes this computation.

18.5.3 Finitely generated cones are closed

Now the key step. We want to separate a point $\mathbf{b}$ from $\operatorname{cone}(\mathbf{A})$, and §03 separates points only from **closed** convex sets. So we must show that $\operatorname{cone}(\mathbf{A})$ is closed.

Here is why that is not obvious. Suppose $\mathbf{A}\mathbf{x}_k \to \mathbf{b}$ with every $\mathbf{x}_k \geq \mathbf{0}$. The obvious move is to take a limit of the $\mathbf{x}_k$, but they need not converge, or even be bounded. If the columns of $\mathbf{A}$ are dependent, then $\mathbf{A}\mathbf{x}_k$ can stay put while $\mathbf{x}_k$ runs off along the null space. The cure has two parts. First, throw away the dependence, generator by generator, which is the "count, then drop a point" move of Theorem 18.2.2 adapted to cones. Second, once the columns are independent, prove that $\mathbf{x}_k$ is recovered from $\mathbf{A}\mathbf{x}_k$ by a fixed linear map, so that the $\mathbf{x}_k$ cannot escape.

Theorem 18.5.2. Carathéodory's Theorem for Cones

Let V be a real vector space and let $a_1, \dots, a_m \in V$. Every $b \in \operatorname{cone}(a_1, \dots, a_m)$ lies in $\operatorname{cone}(a_i : i \in S)$ for some set of indices $S \subseteq \{1, \dots, m\}$ such that the list $(a_i)_{i \in S}$ is **linearly independent**. In particular, if $\dim V = n$, then b is a nonnegative combination of at most n of the a_i.

Idea. Take a representation of b that uses as few generators as possible. If those generators were dependent, a dependence relation could be added to the representation, with a scalar multiple chosen so that one coefficient reaches zero and none goes negative. That removes a generator, which contradicts minimality. It is the move of Theorem 18.2.2 with a *linear* dependence in place of an affine one, and that difference is why the bound is n rather than $n+1$.

Proof. By Definition 18.5.2, $b = \sum_{i=1}^m x_i a_i$ with every $x_i \geq 0$. Discarding the terms with $x_i = 0$, we can write $b = \sum_{i \in S} x_i a_i$ with $x_i > 0$ for every $i \in S$, for some $S \subseteq \{1, \dots, m\}$. The empty set is allowed, when $b = 0$. Among all such representations choose one with $|S|$ as small as possible. We claim that $(a_i)_{i \in S}$ is linearly independent.

Suppose not. Then there are reals c_i, for $i \in S$, not all zero, with $\sum_{i \in S} c_i a_i = 0$. Replacing every c_i by $-c_i$ if necessary, we may assume that $c_i > 0$ for at least one i. Put

$$t = \min\Big\{ \frac{x_i}{c_i} : i \in S,\ c_i > 0 \Big\},$$

a minimum over a nonempty finite set, so it exists. It is > 0, and it is attained at some $i_0 \in S$. Then

$$b = \sum_{i \in S} x_i a_i - t \sum_{i \in S} c_i a_i = \sum_{i \in S} (x_i - t c_i)\, a_i.$$

Every new coefficient is ≥ 0. If $c_i \leq 0$, then $x_i - t c_i \geq x_i > 0$. If $c_i > 0$, then $t \leq x_i / c_i$, so

$x_i - tc_i \geq 0$. Also $x_{i_0} - tc_{i_0} = 0$. Discarding the zero coefficients leaves a representation of b with positive coefficients on a subset of $S \setminus \{i_0\}$, which contradicts the minimality of $|S|$. Hence $(a_i)_{i \in S}$ is linearly independent, and $b \in \operatorname{cone}(a_i : i \in S)$.

If $\dim V = n$, a linearly independent list in V has length at most n by Theorem 1.6.3 (a), so $|S| \leq n$. This proves the theorem.

■

The second part is a statement about one cone with independent generators.

Lemma 18.5.3. Injective Images of the Orthant Are Closed

Let $\mathbf{B} \in M_{n \times k}(\mathbb{R})$ have **linearly independent columns**. Then $\operatorname{cone}(\mathbf{B}) = \{\mathbf{Bx} : \mathbf{x} \in \mathbb{R}^k_{\geq 0}\}$ is a closed subset of $\mathbb{R}^n$.

Idea. Independent columns make $\mathbf{x} \mapsto \mathbf{Bx}$ injective, and an injective matrix has a left inverse Ł with Ł$\mathbf{B} = \mathbf{I}_k$. So if $\mathbf{Bx}_j$ converges, then $\mathbf{x}_j = $ Ł$(\mathbf{Bx}_j)$ converges too, because Ł is a fixed matrix. The limit is still $\geq \mathbf{0}$, and it maps to the limit of the $\mathbf{Bx}_j$. The left inverse is what keeps the $\mathbf{x}_j$ from running away, and injectivity is exactly what provides it.

Proof. If $k = 0$, the cone is $\{\mathbf{0}\}$, which is closed. Let $k \geq 1$. Since the columns of $\mathbf{B}$ are linearly independent, Corollary 11.4.4 ((b) ⇒ (c), with $F = \mathbb{R}$, where $\mathbf{B}^* = \mathbf{B}^\top$) shows that $\mathbf{B}^\top\mathbf{B} \in M_k(\mathbb{R})$ is invertible. Put Ł $= (\mathbf{B}^\top\mathbf{B})^{-1}\mathbf{B}^\top \in M_{k \times n}(\mathbb{R})$, so that Ł$\mathbf{B} = \mathbf{I}_k$.

Let $\mathbf{y}_j \in \operatorname{cone}(\mathbf{B})$ with $\mathbf{y}_j \to \mathbf{y}$. Write $\mathbf{y}_j = \mathbf{Bx}_j$ with $\mathbf{x}_j \geq \mathbf{0}$. Then $\mathbf{x}_j =$ Ł$\mathbf{Bx}_j =$ Ł$\mathbf{y}_j$. By Eq. 18.5.1, $\mathbf{y}_j \to \mathbf{y}$ entrywise. Each entry of Ł$\mathbf{y}_j$ is a fixed linear combination of the entries of $\mathbf{y}_j$, so by the algebra of limits $\mathbf{x}_j =$ Ł$\mathbf{y}_j \to$ Ł$\mathbf{y}$ entrywise. Put $\mathbf{x} =$ Ł$\mathbf{y}$.

Each entry of $\mathbf{x}_j$ is ≥ 0, and a nonstrict inequality survives a limit, so $\mathbf{x} \geq \mathbf{0}$. By the same reasoning with the fixed matrix $\mathbf{B}$, we have $\mathbf{y}_j = \mathbf{Bx}_j \to \mathbf{Bx}$ entrywise. A convergent sequence of real numbers has only one limit, so $\mathbf{Bx} = \mathbf{y}$. Hence $\mathbf{y} = \mathbf{Bx}$ with $\mathbf{x} \geq \mathbf{0}$, that is, $\mathbf{y} \in \operatorname{cone}(\mathbf{B})$. This proves that the cone is closed.

■

The two pieces now fit together.

Theorem 18.5.4. Finitely Generated Cones Are Closed

Let $\mathbf{a}_1, \ldots, \mathbf{a}_m \in \mathbb{R}^n$. Then $\operatorname{cone}(\mathbf{a}_1, \ldots, \mathbf{a}_m)$ is a closed convex cone.

Idea. By Theorem 18.5.2 the cone is the union of the cones of its independent sublists. There are finitely many of them, and each is closed by Lemma 18.5.3. A finite union of closed sets is closed, by Proposition 18.1.6 (b).

Proof. Write $K = \operatorname{cone}(\mathbf{a}_1, \dots, \mathbf{a}_m)$, a convex cone by the remark after Definition 18.5.2. Let $\mathcal{S}$ be the set of those $S \subseteq \{1, \dots, m\}$ for which $(\mathbf{a}_i)_{i \in S}$ is linearly independent, and for $S \in \mathcal{S}$ put $K_S = \operatorname{cone}(\mathbf{a}_i : i \in S)$. The empty set is in $\mathcal{S}$, and $\mathcal{S}$ is finite, with at most 2^m members.

First, $K = \bigcup_{S \in \mathcal{S}} K_S$. Each $K_S \subseteq K$, because a nonnegative combination of some of the $\mathbf{a}_i$ is a nonnegative combination of all of them with the remaining coefficients zero. And $K \subseteq \bigcup_S K_S$ is Theorem 18.5.2.

Next, each K_S is closed. If $S = \emptyset$, then $K_S = \{\mathbf{0}\}$. Otherwise $K_S = \operatorname{cone}(\mathbf{B}_S)$, where $\mathbf{B}_S$ is the matrix whose columns are the $\mathbf{a}_i$ with $i \in S$. Those columns are linearly independent, so Lemma 18.5.3 applies.

Hence K is a finite union of closed sets, closed by Proposition 18.1.6 (b).

■

The proof uses no analysis beyond the algebra of limits of Chapter 16's introduction, through Eq. 18.5.1 and the subsequence argument of Proposition 18.1.6 (b). In particular it never uses compactness, fact (A3) (the only compactness in the background would be the equivalence of norms, which fixing the Euclidean norm avoids); a finitely generated cone other than $\{\mathbf{0}\}$ is unbounded, so compactness had nothing to offer anyway. What the proof uses is **finite generation**, twice: for finitely many pieces, and for writing each piece as the image of an orthant under a matrix with a left inverse. The last part of the section shows that without it the conclusion can fail. In Farkas's lemma below, compactness enters only through the existence half of Theorem 18.3.1.

18.5.4 Farkas's lemma

We can now separate. Theorem 18.3.2 would give some hyperplane between $\mathbf{b}$ and the cone, but for a cone we want one **through the origin**, with a normal vector whose sign on the cone we control. Both come most cleanly from the nearest point of Theorem 18.3.1.

Lemma 18.5.5. Separating a Point from a Closed Cone

Let V be a finite-dimensional real inner product space, let $K \subseteq V$ be a **closed** convex cone, and let $b \in V$ with $b \notin K$. Then there is $y \in K^*$ with $\langle b, y \rangle < 0$.

Idea. Let p be the point of K nearest to b, and use $y = p - b$, the vector from b back to the cone. The inequality of Theorem 18.3.1 says that every point of K is seen from p at an angle of at least 90° from b. Because K is a cone, we may test it against 0 and $2p$, which shows that the separating hyperplane passes through the origin.

Proof. K is nonempty, closed, and convex, so by Theorem 18.3.1 (a) there is a point $p \in K$ nearest to b, and by part (b) of the same theorem

$$\langle b - p, z - p \rangle \le 0 \qquad \text{for every } z \in K. \tag{18.5.2}$$

Put $y = p - b$, which is nonzero since $p \in K$ and $b \notin K$. The points $z = 0$ and $z = 2p$ lie in

K, since K is a cone. Substituting them into Eq. 18.5.2 gives $\langle b-p, p\rangle \geq 0$ and $\langle b-p, p\rangle \leq 0$, so $\langle y, p\rangle = 0$. Now Eq. 18.5.2 reads $\langle b-p, z\rangle \leq 0$, that is $\langle z, y\rangle \geq 0$, for every $z \in K$. So $y \in K^*$. Finally,

$$\langle b, y\rangle = \langle p - y, y\rangle = \langle p, y\rangle - \langle y, y\rangle = -\|y\|^2 < 0,$$

since $y \neq 0$. This proves the lemma.

■

With $K = \text{cone}(\mathbf{A})$, the lemma becomes the theorem this section is named after. It is stated for a matrix, since that is how it will be used.

Theorem 18.5.6. Farkas's Lemma

Let $\mathbf{A} \in M_{m\times n}(\mathbb{R})$ and $\mathbf{b} \in \mathbb{R}^m$. **Exactly one** of the following has a solution:

(F1) $\mathbf{Ax} = \mathbf{b}$ with $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{x} \geq \mathbf{0}$;
(F2) $\mathbf{A}^\top\mathbf{y} \geq \mathbf{0}$ and $\mathbf{b}^\top\mathbf{y} < 0$, with $\mathbf{y} \in \mathbb{R}^m$.

Idea. The two systems cannot both be solvable, which is one line of algebra: pair $\mathbf{Ax} = \mathbf{b}$ with $\mathbf{y}$. That at least one is solvable is geometry. If (F1) fails, then $\mathbf{b}$ lies outside the cone of the columns, the cone is closed by Theorem 18.5.4, and Lemma 18.5.5 produces $\mathbf{y}$.

Proof. Not both. Suppose $\mathbf{Ax} = \mathbf{b}$ with $\mathbf{x} \geq \mathbf{0}$, and $\mathbf{A}^\top\mathbf{y} \geq \mathbf{0}$. Then

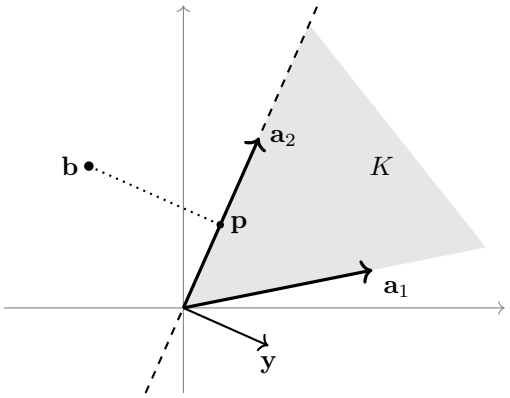
$$\mathbf{b}^\top\mathbf{y} = (\mathbf{Ax})^\top\mathbf{y} = \mathbf{x}^\top(\mathbf{A}^\top\mathbf{y}) \geq 0,$$

since it is a sum of products of the nonnegative entries of $\mathbf{x}$ and $\mathbf{A}^\top\mathbf{y}$. So $\mathbf{b}^\top\mathbf{y} < 0$ is impossible.

At least one. Suppose (F1) has no solution. Let $K = \text{cone}(\mathbf{A})$. By Definition 18.5.2, (F1) says exactly that $\mathbf{b} \in K$, so $\mathbf{b} \notin K$. By Theorem 18.5.4, K is a closed convex cone, so Lemma 18.5.5 gives $\mathbf{y} \in K^*$ with $\langle \mathbf{b}, \mathbf{y}\rangle = \mathbf{b}^\top\mathbf{y} < 0$. By Proposition 18.5.1 (d), $\mathbf{y} \in K^*$ means $\mathbf{A}^\top\mathbf{y} \geq \mathbf{0}$. So $\mathbf{y}$ solves (F2). This proves the theorem.

■

The geometric reading. Either $\mathbf{b}$ lies in the cone spanned by the columns of $\mathbf{A}$, or there is a hyperplane through the origin, $\{\mathbf{z} : \langle \mathbf{z}, \mathbf{y}\rangle = 0\}$, with every column on one side ($\langle \mathbf{a}_j, \mathbf{y}\rangle \geq 0$) and $\mathbf{b}$ strictly on the other ($\langle \mathbf{b}, \mathbf{y}\rangle < 0$). The vector $\mathbf{y}$ is a **certificate of infeasibility**: anyone can check (F2) by multiplying out, and the "not both" computation then proves that (F1) has no solution.



$\mathbf{b}$ lies outside $K = \operatorname{cone}(\mathbf{a}_1, \mathbf{a}_2)$, and $\mathbf{p}$ is the point of K nearest to it. The vector $\mathbf{y}$ points along $\mathbf{p} - \mathbf{b}$. The dashed line $\langle \mathbf{z}, \mathbf{y} \rangle = 0$ passes through $\mathbf{0}$, with K on the side $\langle \mathbf{z}, \mathbf{y} \rangle \geq 0$ and $\mathbf{b}$ strictly on the other

Example. Deciding a system both ways

Let

$$\mathbf{A} = \begin{pmatrix} 1 & 2 & 0 \\ 0 & 1 & 1 \end{pmatrix}, \qquad \mathbf{b} = \begin{pmatrix} 3 \\ 2 \end{pmatrix}, \qquad \mathbf{b}' = \begin{pmatrix} -1 \\ 1 \end{pmatrix}.$$

Decide, for $\mathbf{b}$ and for $\mathbf{b}'$, which alternative of Theorem 18.5.6 holds, and exhibit the solution.

Solution. The columns are $\mathbf{a}_1 = \mathbf{e}_1$, $\mathbf{a}_2 = (2, 1)$, and $\mathbf{a}_3 = \mathbf{e}_2$, all in the closed first quadrant. So $\operatorname{cone}(\mathbf{A}) \subseteq \mathbb{R}^2_{\geq 0}$, and in fact the two are equal, since $\mathbf{e}_1$ and $\mathbf{e}_2$ are among the generators.

For $\mathbf{b} = (3, 2)$ we expect (F1). The system $x_1 + 2x_2 = 3$, $x_2 + x_3 = 2$ has, for example, the solution $\mathbf{x} = (3, 0, 2) \geq \mathbf{0}$, and also $\mathbf{x} = (1, 1, 1)$. Nonnegative solutions need not be unique.

For $\mathbf{b}' = (-1, 1)$ we expect (F2), since $\mathbf{b}'$ has a negative first entry. Try $\mathbf{y} = \mathbf{e}_1$: then $\mathbf{A}^\top \mathbf{y} = (1, 2, 0) \geq \mathbf{0}$ and $\mathbf{b}'^\top \mathbf{y} = -1 < 0$. So $\mathbf{y} = \mathbf{e}_1$ is a certificate, and by the "not both" half of Theorem 18.5.6 the system $\mathbf{Ax} = \mathbf{b}'$, $\mathbf{x} \geq \mathbf{0}$ has no solution. The certificate encodes a one-line argument: the first equation reads $x_1 + 2x_2 = -1$, and its left side is ≥ 0 whenever $\mathbf{x} \geq \mathbf{0}$.

Linear programming, in the next section, meets systems of **inequalities**, $\mathbf{Ax} \leq \mathbf{b}$ with $\mathbf{x} \geq \mathbf{0}$. Farkas's lemma handles them after one standard move: add a **slack** variable to turn each inequality into an equation.

Corollary 18.5.7. Farkas's Lemma for Inequalities

Let $\mathbf{M} \in M_{m \times n}(\mathbb{R})$ and $\mathbf{d} \in \mathbb{R}^m$. Exactly one of the following has a solution:

(i) $\mathbf{Mz} \leq \mathbf{d}$ with $\mathbf{z} \in \mathbb{R}^n$, $\mathbf{z} \geq \mathbf{0}$;
(ii) $\mathbf{w} \geq \mathbf{0}$, $\mathbf{M}^\top \mathbf{w} \geq \mathbf{0}$, and $\mathbf{d}^\top \mathbf{w} < 0$, with $\mathbf{w} \in \mathbb{R}^m$.

Idea. A slack vector $\mathbf{s} = \mathbf{d} - \mathbf{Mz} \geq \mathbf{0}$ turns the inequalities into equations with matrix $\begin{pmatrix} \mathbf{M} & \mathbf{I}_m \end{pmatrix}$, and in Theorem 18.5.6 the identity block becomes the extra condition $\mathbf{w} \geq \mathbf{0}$.

Proof. Here $\mathbf{Mz} \le \mathbf{d}$ means $\mathbf{d} - \mathbf{Mz} \ge \mathbf{0}$. So (i) has a solution if and only if there are $\mathbf{z} \ge \mathbf{0}$ and $\mathbf{s} \ge \mathbf{0}$ in $\mathbb{R}^m$ with $\mathbf{Mz} + \mathbf{s} = \mathbf{d}$: given $\mathbf{z}$, take $\mathbf{s} = \mathbf{d} - \mathbf{Mz}$, and given $\mathbf{z}$ and $\mathbf{s}$, drop $\mathbf{s}$. In block form this reads

$$\begin{pmatrix} \mathbf{M} & \mathbf{I}_m \end{pmatrix} \begin{pmatrix} \mathbf{z} \\ \mathbf{s} \end{pmatrix} = \mathbf{d}, \qquad \begin{pmatrix} \mathbf{z} \\ \mathbf{s} \end{pmatrix} \ge \mathbf{0}.$$

Apply Theorem 18.5.6 to the $m \times (n+m)$ matrix $\begin{pmatrix} \mathbf{M} & \mathbf{I}_m \end{pmatrix}$ and to $\mathbf{d}$. Exactly one of the system above and the following is solvable: $\begin{pmatrix} \mathbf{M} & \mathbf{I}_m \end{pmatrix}^{\top} \mathbf{w} \ge \mathbf{0}$ and $\mathbf{d}^{\top}\mathbf{w} < 0$. The transpose has the blocks $\mathbf{M}^{\top}$ and $\mathbf{I}_m$ stacked, so the first condition says $\mathbf{M}^{\top}\mathbf{w} \ge \mathbf{0}$ and $\mathbf{w} \ge \mathbf{0}$. That is (ii).

■

18.5.5 The dual of the dual

Farkas's lemma has a coordinate-free form, which says that nothing is lost by passing to the dual cone twice, provided the cone is closed.

Theorem 18.5.8. The Bidual of a Closed Cone

Let V be a finite-dimensional real inner product space and let $K \subseteq V$ be a **closed** convex cone. Then $K^{**} = K$.

Proof. $K \subseteq K^{**}$ is Proposition 18.5.1 (c). Conversely, let $b \notin K$. By Lemma 18.5.5 there is $y \in K^*$ with $\langle b, y \rangle < 0$. That single y shows $b \notin K^{**}$. Hence $K^{**} \subseteq K$, and the two are equal.

■

For $K = \operatorname{cone}(\mathbf{A})$, which is closed by Theorem 18.5.4, the theorem **is** Farkas's lemma. By Proposition 18.5.1 (d), $K^* = \{\mathbf{y} : \mathbf{A}^{\top}\mathbf{y} \ge \mathbf{0}\}$, so "$\mathbf{b} \in K^{**}$" says that every $\mathbf{y}$ with $\mathbf{A}^{\top}\mathbf{y} \ge \mathbf{0}$ has $\mathbf{b}^{\top}\mathbf{y} \ge 0$, that is, (F2) has no solution. And "$\mathbf{b} \in K$" is (F1). This is the "dualize" move in its purest form: a question about a cone (is $\mathbf{b}$ in it?) is traded for one about its dual (does every member of the dual approve of $\mathbf{b}$?), and the trade is exact when the cone is closed.

Remark. For the wedge $K = \operatorname{cone}((1,0),(1,1))$, whose dual is $\operatorname{cone}((0,1),(1,-1))$ by Exercise (B3: A dual cone and its dual), the theorem and Proposition 18.5.1 (d) say that $K = \{\mathbf{x} : \langle \mathbf{x}, (0,1) \rangle \ge 0,\ \langle \mathbf{x}, (1,-1) \rangle \ge 0\}$, one inequality for each generator of K^*. In general the theorem describes a closed convex cone from outside, as an intersection of half-spaces through the origin; for a finitely generated cone finitely many suffice, which is one half of the Minkowski–Weyl theorem of Section 9.

18.5.6 When the cone is not finitely generated

The system "$\mathbf{Ax} = \mathbf{b}$, $\mathbf{x} \ge \mathbf{0}$" asks whether $\mathbf{b}$ lies in the image of the orthant under $\mathbf{x} \mapsto \mathbf{Ax}$. Replace the orthant by a closed convex cone $K \subseteq \mathbb{R}^k$ and $\mathbf{A}$ by a matrix $\mathbf{P} \in M_{m \times k}(\mathbb{R})$, and

ask whether $\mathbf{b} \in \mathbf{P}(K)$. The "not both" half of the proof of Theorem 18.5.6 survives unchanged. The "at least one" half needs $\mathbf{P}(K)$ to be closed, and the next example shows that this can fail.

Example. A closed cone with a nonclosed shadow

In $\mathbb{R}^3$, with coordinates (x, y, z), let

$$K = \{(x, y, z) : xs^2 + 2y\,st + zt^2 \geq 0 \text{ for all } s, t \in \mathbb{R}\},$$

the set of triples for which the symmetric matrix $\left(\begin{smallmatrix} x & y \\ y & z \end{smallmatrix}\right)$ is positive semidefinite (Definition 13.1.1). Let $\mathbf{P} = \left(\begin{smallmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{smallmatrix}\right)$, so that $\mathbf{P}(x, y, z) = (x, y)$ is the projection onto the first two coordinates, and write $\mathbf{P}(K) = \{\mathbf{Pk} : \mathbf{k} \in K\}$. Show that K is a closed convex cone, but that

$$\mathbf{P}(K) = \{(x, y) : x > 0\} \cup \{(0, 0)\},$$

the cone of the earlier warning, which is not closed. Then show that for $\mathbf{b} = (0, 1)$ **neither** of the two alternatives holds:

(i) $\mathbf{Pk} = \mathbf{b}$ for some $\mathbf{k} \in K$;

(ii) $\langle \mathbf{Pk}, \mathbf{y} \rangle \geq 0$ for every $\mathbf{k} \in K$, and $\langle \mathbf{b}, \mathbf{y} \rangle < 0$, for some $\mathbf{y} \in \mathbb{R}^2$.

Solution. K is a closed convex cone. Write $q_{\mathbf{k}}(s, t) = xs^2 + 2yst + zt^2$ for $\mathbf{k} = (x, y, z)$. It is linear in $\mathbf{k}$: $q_{a\mathbf{k}+b\mathbf{k}'} = aq_{\mathbf{k}} + bq_{\mathbf{k}'}$. So if $\mathbf{k}, \mathbf{k}' \in K$ and $a, b \geq 0$, then $q_{a\mathbf{k}+b\mathbf{k}'}(s, t) \geq 0$ for all s, t, and K is a convex cone that contains $\mathbf{0}$. If $\mathbf{k}_j \to \mathbf{k}$ with $\mathbf{k}_j \in K$, then for fixed s, t the number $q_{\mathbf{k}_j}(s, t)$ is a fixed linear combination of the entries of $\mathbf{k}_j$. So it converges to $q_{\mathbf{k}}(s, t)$, and $q_{\mathbf{k}}(s, t) \geq 0$ because a nonstrict inequality survives a limit. Hence K is closed.

Membership. By Theorem 13.1.4 (a) $\Leftrightarrow$ (e), $(x, y, z) \in K$ exactly when the principal minors x, z, and $xz - y^2$ are all ≥ 0.

The image. $(0, 0) = \mathbf{P}(\mathbf{0}) \in \mathbf{P}(K)$. If $x > 0$ and y is arbitrary, put $z = y^2/x$. Then $x \geq 0$, $z \geq 0$, and $xz - y^2 = 0$, so $(x, y, z) \in K$ and $(x, y) \in \mathbf{P}(K)$. Conversely, if $(x, y, z) \in K$, then $x \geq 0$, and if $x = 0$ then $0 = xz \geq y^2$ forces $y = 0$. So $\mathbf{P}(K)$ is exactly the displayed set. It is not closed: $(1/k, 1) \in \mathbf{P}(K)$ converges to $(0, 1) \notin \mathbf{P}(K)$.

Neither alternative holds for $\mathbf{b} = (0, 1)$. Alternative (i) fails because $\mathbf{b} \notin \mathbf{P}(K)$. For (ii), suppose $\mathbf{y} = (y_1, y_2)$ has $\langle \mathbf{Pk}, \mathbf{y} \rangle \geq 0$ for every $\mathbf{k} \in K$. Since $(1, t) \in \mathbf{P}(K)$ for every real t, we get $y_1 + ty_2 \geq 0$ for all t. A nonconstant affine function of t takes negative values, so $y_2 = 0$. But then $\langle \mathbf{b}, \mathbf{y} \rangle = y_2 = 0$, which is not < 0. So (ii) fails too.

It is worth seeing exactly where the proof of Theorem 18.5.6 breaks. The vectors $\mathbf{k}_j = (1/j, 1, j)$ lie in K, their images $(1/j, 1)$ converge to $\mathbf{b}$, and the $\mathbf{k}_j$ themselves run off to infinity in the third coordinate. This is the escape that the left inverse in Lemma 18.5.3 prevented. There is no nearest point of $\mathbf{P}(K)$ to $\mathbf{b}$: the distance is 0, but $\mathbf{b}$ is not in the set, so Lemma 18.5.5 has nothing to start from. What fails is **closedness**, which is why the finitely generated case needed its own theorem.

The dual cone cannot see the difference either: $\mathbf{P}(K)^{**}$ is the closed half-plane $\{x \geq 0\}$, strictly larger than $\mathbf{P}(K)$, as Exercise (C3: The bidual sees only the closure) shows.

18.5.7 Exercises

18.5.7.1 A. Check your understanding

Exercise. A1

(a) Define a convex cone and the dual cone of a subset S of $\mathbb{R}^n$.

(b) State Farkas's lemma, and say in one sentence what the vector $\mathbf{y}$ in the second alternative certifies.

(c) Determine whether the following is correct, and justify your answer: every convex cone in $\mathbb{R}^2$ is closed.

(d) Determine whether the following is correct, and justify your answer: the union of two convex cones in $\mathbb{R}^2$ is a convex cone.

(e) In Theorem 18.5.2 the bound is n generators, while Theorem 18.2.2 allows $n+1$ points. Which feature of the proof accounts for the difference?

Solution.

(a) A convex cone is a nonempty $K \subseteq \mathbb{R}^n$ with $a\mathbf{u} + b\mathbf{v} \in K$ for all $\mathbf{u}, \mathbf{v} \in K$ and all real $a, b \geq 0$. The dual cone of S is $S^* = \{\mathbf{y} : \langle \mathbf{x}, \mathbf{y} \rangle \geq 0 \text{ for every } \mathbf{x} \in S\}$.

(b) For $\mathbf{A} \in M_{m\times n}(\mathbb{R})$ and $\mathbf{b} \in \mathbb{R}^m$, exactly one of the following is solvable: $\mathbf{Ax} = \mathbf{b}$ with $\mathbf{x} \geq \mathbf{0}$, or $\mathbf{A}^\top \mathbf{y} \geq \mathbf{0}$ with $\mathbf{b}^\top \mathbf{y} < 0$. Such a $\mathbf{y}$ certifies that the first system has no solution, since any solution $\mathbf{x}$ would give $0 > \mathbf{b}^\top \mathbf{y} = \mathbf{x}^\top \mathbf{A}^\top \mathbf{y} \geq 0$.

(c) Incorrect. The cone $\{(x, y) : x > 0\} \cup \{\mathbf{0}\}$ of the warning in this section is a convex cone, but $(1/k, 1) \to (0, 1)$ leaves it.

(d) Incorrect. The two quadrants $\mathbb{R}^2_{\geq 0}$ and $-\mathbb{R}^2_{\geq 0}$ are convex cones, and their union contains $(1, 0)$ and $(0, -1)$ but not their sum $(1, -1)$.

(e) For cones the dependence to be removed is a **linear** dependence $\sum c_i a_i = 0$, and a linearly independent list in an n-dimensional space has at most n members. For convex hulls it is an **affine** dependence (coefficients also summing to 0), and an affinely independent list can have $n + 1$ members.

18.5.7.2 B. Practice

Exercise. B1: Recognizing cones

Determine which of the following subsets of $\mathbb{R}^2$ or $\mathbb{R}^3$ are convex cones. Justify your answers.

(a) $\{\mathbf{x} \in \mathbb{R}^2 : x_1 \geq |x_2|\}$.

(b) $\{\mathbf{x} \in \mathbb{R}^3 : x_1 + x_2 + x_3 = 0\}$.

(c) $\{\mathbf{x} \in \mathbb{R}^2 : x_1 \geq 1\}$.

(d) $\{\mathbf{x} \in \mathbb{R}^2 : x_2 \geq x_1^2\}$.

(e) $\{\mathbf{x} \in \mathbb{R}^3 : x_1 \geq 0,\ x_2 \geq 0,\ x_3 \leq x_1 + x_2\}$.

Solution.

(a) A convex cone. It is nonempty, containing $\mathbf{0}$. If $x_1 \geq |x_2|$, $y_1 \geq |y_2|$, and $a, b \geq 0$, then

by the triangle inequality for real numbers (Theorem 0.6.4)

$$|ax_2 + by_2| \le a|x_2| + b|y_2| \le ax_1 + by_1.$$

It is $\operatorname{cone}((1,1),(1,-1))$, a quarter-plane turned by 45°.

(b) A convex cone. It is a subspace, being the null space of $\begin{pmatrix}1 & 1 & 1\end{pmatrix}$, and every subspace is a convex cone.

(c) Not a convex cone: it does not contain $\mathbf{0}$, which taking $a = b = 0$ in the definition would force into it. It is convex and closed under addition; what fails is scaling by $0 \le a < 1$, for instance $\frac{1}{2}(1,0) = (\frac{1}{2},0)$ is not in it.

(d) Not a convex cone. It is convex, but it is not closed under nonnegative scaling: $(1,1)$ is in it, and $2(1,1) = (2,2)$ is not, since $2 < 4$.

(e) A convex cone. Each of the three conditions is a homogeneous linear inequality $\langle \mathbf{x}, \mathbf{a} \rangle \ge 0$, with $\mathbf{a} = \mathbf{e}_1$, $\mathbf{e}_2$, and $(1,1,-1)$, so the set is an intersection of three half-spaces through the origin. Each is a convex cone by the examples of this section, and an intersection of convex cones is a convex cone: it contains $\mathbf{0}$, and each member cone contains every nonnegative combination of points of the intersection.

Exercise. B2: A certificate either way

Let $\mathbf{A} = \begin{pmatrix}1 & 1 & -1\\ 0 & 1 & 2\end{pmatrix}$. For each of $\mathbf{b} = (0,3)$ and $\mathbf{b} = (-1,1)$, decide which alternative of Theorem 18.5.6 holds, and exhibit a solution of that alternative.

Solution. The columns are $\mathbf{a}_1 = (1,0)$, $\mathbf{a}_2 = (1,1)$, and $\mathbf{a}_3 = (-1,2)$.

For $\mathbf{b} = (0,3)$, try $\mathbf{x} = (0, x_2, x_3)$: the equations $x_2 - x_3 = 0$ and $x_2 + 2x_3 = 3$ give $x_2 = x_3 = 1$. So $\mathbf{x} = (0,1,1) \ge \mathbf{0}$ solves $\mathbf{Ax} = \mathbf{b}$, and (F1) holds.

For $\mathbf{b} = (-1,1)$, look for $\mathbf{y}$ perpendicular to the "outermost" generator $\mathbf{a}_3$ and nonnegative on the others. The vector $\mathbf{y} = (2,1)$ gives $\mathbf{A}^\top \mathbf{y} = (2,3,0) \ge \mathbf{0}$ and $\mathbf{b}^\top \mathbf{y} = -2 + 1 = -1 < 0$. So (F2) holds, and by Theorem 18.5.6 the system $\mathbf{Ax} = (-1,1)$, $\mathbf{x} \ge \mathbf{0}$ has no solution. As a direct check, adding twice the first equation to the second gives $2x_1 + 3x_2 = -1$, whose left side is ≥ 0 for $\mathbf{x} \ge \mathbf{0}$.

Exercise. B3: A dual cone and its dual

Let $K = \operatorname{cone}((1,0),(1,1)) \subseteq \mathbb{R}^2$.

(a) Show that $K^* = \operatorname{cone}((0,1),(1,-1))$.

(b) Compute K^{**} and check that it equals K, as Theorem 18.5.8 predicts.

Solution.

(a) By Proposition 18.5.1 (d), $K^* = \{\mathbf{y} : y_1 \ge 0,\ y_1 + y_2 \ge 0\}$. Given such a $\mathbf{y}$, write $\mathbf{y} = (y_1 + y_2)(0,1) + y_1(1,-1)$: indeed the right side is $(y_1, y_1 + y_2 - y_1) = (y_1, y_2)$,

and both coefficients are ≥ 0. So $\mathbf{y} \in \operatorname{cone}((0,1),(1,-1))$. Conversely, both generators satisfy the two inequalities, and the set defined by the inequalities is a convex cone, so it contains the cone they generate.

(b) Applying Proposition 18.5.1 (d) to the generators found in (a), $K^{**} = \{\mathbf{x} : x_2 \geq 0,\ x_1 - x_2 \geq 0\}$. The same decomposition argument, now with $\mathbf{x} = (x_1 - x_2)(1,0) + x_2(1,1)$, shows this set is $\operatorname{cone}((1,0),(1,1)) = K$.

18.5.7.3 C. Going deeper

Exercise. C1: Gordan's alternative

Let $\mathbf{A} \in M_{m\times n}(\mathbb{R})$.

(a) Prove that exactly one of the following has a solution: $\mathbf{Ax} = \mathbf{0}$ with $\mathbf{x} \geq \mathbf{0}$ and $\mathbf{x} \neq \mathbf{0}$; or $\mathbf{A}^\top \mathbf{y} > \mathbf{0}$ with $\mathbf{y} \in \mathbb{R}^m$.

(b) Deduce that if the columns of $\mathbf{A}$ are $(1,0)$, $(0,1)$, and $(-1,-1)$, then some nontrivial nonnegative combination of them is $\mathbf{0}$, and find one.

Hint for (a): scale $\mathbf{x}$ *so that its entries sum to* 1*, and apply Theorem 18.5.6 to* $\mathbf{A}$ *with an extra row of ones.*

Solution.

(a) *Not both.* If $\mathbf{Ax} = \mathbf{0}$ with $\mathbf{x} \geq \mathbf{0}$, $\mathbf{x} \neq \mathbf{0}$, and $\mathbf{A}^\top \mathbf{y} > \mathbf{0}$, then $0 = \mathbf{y}^\top \mathbf{Ax} = \sum_j x_j (\mathbf{A}^\top \mathbf{y})_j$. Every term is ≥ 0, and the term for an index with $x_j > 0$ is > 0. That is a contradiction.
At least one. Suppose the first system has no solution. If $\mathbf{Ax} = \mathbf{0}$, $\mathbf{x} \geq \mathbf{0}$, and $\mathbf{1}^\top \mathbf{x} = 1$, then $\mathbf{x} \neq \mathbf{0}$ would solve it. So the system

$$\begin{pmatrix} \mathbf{A} \\ \mathbf{1}^\top \end{pmatrix} \mathbf{x} = \begin{pmatrix} \mathbf{0} \\ 1 \end{pmatrix}, \qquad \mathbf{x} \geq \mathbf{0},$$

has no solution. By Theorem 18.5.6 there is $(\mathbf{y}, s) \in \mathbb{R}^m \times \mathbb{R}$ with $\mathbf{A}^\top \mathbf{y} + s\mathbf{1} \geq \mathbf{0}$ and $\mathbf{0}^\top \mathbf{y} + 1 \cdot s < 0$, that is, $s < 0$. Then $\mathbf{A}^\top \mathbf{y} \geq -s\mathbf{1} > \mathbf{0}$, so $\mathbf{y}$ solves the second system.

(b) Here $\mathbf{A} = \left(\begin{smallmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \end{smallmatrix}\right)$. The second system is impossible: $\mathbf{A}^\top \mathbf{y} = (y_1, y_2, -y_1 - y_2)$, and its entries sum to 0, so they cannot all be positive. By (a) the first system is solvable. Indeed $\mathbf{x} = (1,1,1)$ gives $\mathbf{Ax} = (1-1, 1-1) = \mathbf{0}$.

Exercise. C2: The ice-cream cone is its own dual

Let $L = \{(\mathbf{x}, t) \in \mathbb{R}^n \times \mathbb{R} : \|\mathbf{x}\|_2 \leq t\}$, with the inner product $\langle (\mathbf{x}, t), (\mathbf{y}, s) \rangle = \mathbf{y}^\top \mathbf{x} + ts$ on $\mathbb{R}^{n+1}$.

(a) Prove that $L \subseteq L^*$.

(b) Prove that $L^* \subseteq L$.

(c) Deduce that L is closed.

Hint for (b): test $(\mathbf{y}, s) \in L^*$ *against the member* $(-\mathbf{y}, \|\mathbf{y}\|_2)$ *of* L.

Solution.

(a) Let $(\mathbf{x}, t), (\mathbf{y}, s) \in L$. By Theorem 11.1.5, $|\mathbf{y}^\top \mathbf{x}| \le \|\mathbf{x}\|_2 \|\mathbf{y}\|_2 \le ts$, where the last step multiplies the inequalities $0 \le \|\mathbf{x}\|_2 \le t$ and $0 \le \|\mathbf{y}\|_2 \le s$. Hence $\mathbf{y}^\top \mathbf{x} + ts \ge ts - |\mathbf{y}^\top \mathbf{x}| \ge 0$. Fix $(\mathbf{y}, s) \in L$: we have shown that its inner product with every $(\mathbf{x}, t) \in L$ is ≥ 0, so $(\mathbf{y}, s) \in L^*$. Hence $L \subseteq L^*$.

(b) Let $(\mathbf{y}, s) \in L^*$. The point $(-\mathbf{y}, \|\mathbf{y}\|_2)$ lies in L, so

$$0 \le \langle (-\mathbf{y}, \|\mathbf{y}\|_2), (\mathbf{y}, s) \rangle = -\|\mathbf{y}\|_2^2 + s\|\mathbf{y}\|_2 = \|\mathbf{y}\|_2 (s - \|\mathbf{y}\|_2).$$

If $\mathbf{y} \neq \mathbf{0}$, dividing by $\|\mathbf{y}\|_2 > 0$ gives $s \ge \|\mathbf{y}\|_2$. If $\mathbf{y} = \mathbf{0}$, test against $(\mathbf{0}, 1) \in L$ instead: $s \ge 0 = \|\mathbf{y}\|_2$. Either way $(\mathbf{y}, s) \in L$.

(c) By (a) and (b), $L = L^*$, and every dual cone is closed by Proposition 18.5.1 (a).

Exercise. C3: The bidual sees only the closure

Let $C = \{(x, y) : x > 0\} \cup \{(0, 0)\} \subseteq \mathbb{R}^2$, the image cone of Example (A closed cone with a nonclosed shadow).

(a) Prove that $C^* = \{(y_1, 0) : y_1 \ge 0\}$.

(b) Prove that $C^{**} = \{(x, y) : x \ge 0\}$.

(c) Explain why this does not contradict Theorem 18.5.8, and show that C^{**} is the set of all limits of convergent sequences of points of C.

Solution.

(a) If $(y_1, y_2) \in C^*$, then testing against $(1, t) \in C$ gives $y_1 + t y_2 \ge 0$ for every real t. This forces $y_2 = 0$, as in the example, and then $t = 0$ gives $y_1 \ge 0$. Conversely, if $y_2 = 0$ and $y_1 \ge 0$, then $\langle (x, y), (y_1, 0) \rangle = x y_1 \ge 0$ for every $(x, y) \in C$, since $x \ge 0$ on C.

(b) By (a), $C^* = \operatorname{cone}((1, 0))$, so by Proposition 18.5.1 (d),

$$C^{**} = \{(x, y) : \langle (x, y), (1, 0) \rangle \ge 0\} = \{(x, y) : x \ge 0\}.$$

(c) Theorem 18.5.8 assumes the cone is closed, and C is not. Every point of C^{**} is such a limit: if $x > 0$ or $(x, y) = \mathbf{0}$, use the constant sequence; if $x = 0$, use $(1/k, y) \in C$, which tends to $(0, y)$. Conversely, a limit of points of $C \subseteq C^{**}$ lies in C^{**}, because C^{**} is closed by Proposition 18.5.1 (a). So C^{**} is exactly the set of limits of sequences in C.

18.6 Linear Programming Duality

Chapter 3 ended its traffic example (Example (Traffic in a town center)) by admitting what linear algebra alone cannot do: "Nonnegativity and street capacities turn the problem into one about inequalities, which is the subject of linear programming." This section takes up that subject. A **linear program** asks for the largest value of a linear function on a region cut out by finitely many linear inequalities. The theory is almost entirely one idea. Every such maximum comes with a companion minimum, the **dual** program, and the two optimal values are equal.

The proof of that equality is Farkas's lemma of §05, applied to one combined system.

The equality has a practical meaning that goes beyond the theorem. A feasible point proves a lower bound on the maximum, a dual feasible point proves an upper bound, and when the two bounds meet, the answer has been **certified**: anyone can check it by multiplying out, with no need to trust how it was found.

Throughout, the field is $\mathbb{R}$, vectors are columns, and $\mathbf{x} \geq \mathbf{0}$, $\mathbf{u} \leq \mathbf{v}$ are the entrywise relations of §05.

18.6.1 Linear programs

Here is a small instance of the kind of question we want to answer. A workshop makes two products. Each unit of the first uses 2 hours of machine time and 1 hour of labor, each unit of the second uses 1 hour of machine time and 3 hours of labor, and at most 3 units of the first can be sold. Machine time is limited to 7 hours, labor to 9, and the profits are 4 and 3 per unit. With x_1, x_2 units made, the question is:

$$\begin{aligned} \text{maximize} \quad & 4x_1 + 3x_2 \\ \text{subject to} \quad & 2x_1 + x_2 \leq 7, \quad x_1 + 3x_2 \leq 9, \quad x_1 \leq 3, \\ & x_1 \geq 0, \quad x_2 \geq 0. \end{aligned}$$

The objective and every constraint are linear, and there are finitely many constraints. In matrix form it reads "maximize $\mathbf{c}^{\top}\mathbf{x}$ subject to $\mathbf{A}\mathbf{x} \leq \mathbf{b}$, $\mathbf{x} \geq \mathbf{0}$". That shape deserves a name.

A linear program is the search for the best value of a linear function over a region defined by finitely many linear inequalities.

Definition 18.6.1. Linear Program

Let $\mathbf{A} \in M_{m\times n}(\mathbb{R})$, $\mathbf{b} \in \mathbb{R}^m$, and $\mathbf{c} \in \mathbb{R}^n$. The **linear program** with these data, in **inequality form**, is the problem

$$\text{(P)} \qquad \text{maximize } \mathbf{c}^{\top}\mathbf{x} \quad \text{subject to} \quad \mathbf{A}\mathbf{x} \leq \mathbf{b}, \quad \mathbf{x} \geq \mathbf{0}.$$

A vector $\mathbf{x} \in \mathbb{R}^n$ with $\mathbf{A}\mathbf{x} \leq \mathbf{b}$ and $\mathbf{x} \geq \mathbf{0}$ is a **feasible point**, and the set of all of them is the **feasible region**. The program is **feasible** if it has a feasible point and **infeasible** otherwise. It is **unbounded** if it is feasible and $\mathbf{c}^{\top}\mathbf{x}$ takes **arbitrarily large** values on the feasible region. A feasible $\mathbf{x}^{\star}$ is an **optimal solution** if $\mathbf{c}^{\top}\mathbf{x}^{\star} \geq \mathbf{c}^{\top}\mathbf{x}$ for **every** feasible $\mathbf{x}$, and then $\mathbf{c}^{\top}\mathbf{x}^{\star}$ is the **optimal value**.

Clause by clause. The function $\mathbf{x} \mapsto \mathbf{c}^{\top}\mathbf{x}$ is the **objective**. The m rows of $\mathbf{A}\mathbf{x} \leq \mathbf{b}$ are the **constraints**, and $\mathbf{x} \geq \mathbf{0}$ adds n more. An optimal solution is a feasible point at which the objective is **as large as anywhere** on the feasible region, and it need not be unique. The optimal value is unique when it exists. A program can fail to have an optimal solution in two ways: it can have no feasible point, or its objective can be unbounded above. Whether those are the **only** two ways is the question the section's second theorem settles.

Examples. The workshop problem is (P) with

$$\mathbf{A} = \begin{pmatrix} 2 & 1 \\ 1 & 3 \\ 1 & 0 \end{pmatrix}, \qquad \mathbf{b} = \begin{pmatrix} 7 \\ 9 \\ 3 \end{pmatrix}, \qquad \mathbf{c} = \begin{pmatrix} 4 \\ 3 \end{pmatrix}.$$

- **One variable.** With $n = m = 1$, "maximize x subject to $x \le 2$, $x \ge 0$" has the optimal solution $x = 2$. "Maximize x subject to $-x \le -1$, $x \ge 0$" asks for the largest $x \ge 1$, and it is unbounded. "Maximize x subject to $x \le -1$, $x \ge 0$" is infeasible.
- **Degenerate case.** If $\mathbf{c} = \mathbf{0}$, every feasible point is optimal, with value 0. The program then asks only whether the system $\mathbf{Ax} \le \mathbf{b}$, $\mathbf{x} \ge \mathbf{0}$ has a solution. So the feasibility question of Corollary 18.5.7 is itself a linear program, the one with no objective.

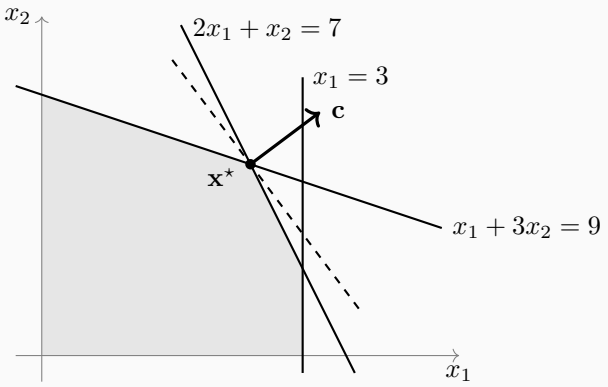
Nonexample by minimal change. Make one inequality strict: "maximize x subject to $x < 2$, $x \ge 0$". Everything is still linear, but the supremum 2 is not attained, and no feasible point is optimal. The definition allows only nonstrict inequalities. They define **closed** regions, and closedness will again be what the proofs need.

Why this form. Every problem with a linear objective and finitely many linear constraints can be brought to the shape (P). To minimize $\mathbf{c}^\top\mathbf{x}$, maximize $(-\mathbf{c})^\top\mathbf{x}$. A constraint $\mathbf{a}^\top\mathbf{x} \ge \beta$ is $(-\mathbf{a})^\top\mathbf{x} \le -\beta$. An equation $\mathbf{a}^\top\mathbf{x} = \beta$ is the pair $\mathbf{a}^\top\mathbf{x} \le \beta$, $-\mathbf{a}^\top\mathbf{x} \le -\beta$. A variable x_j without a sign constraint can be written $x_j = u_j - v_j$ with $u_j, v_j \ge 0$. Exercise (C2: Free variables) carries out the last conversion. The form (P), with its sign constraints on $\mathbf{x}$, is chosen because its dual, below, has the same shape, which makes the theory symmetric.

Example. The workshop problem, drawn

Find the optimal solution of the workshop problem, and explain why it is optimal.

Solution. The feasible region is the pentagon with corners $(0,0)$, $(3,0)$, $(3,1)$, $(\frac{12}{5}, \frac{11}{5})$, and $(0,3)$. For example, $(3,1)$ is where $x_1 = 3$ meets $2x_1 + x_2 = 7$, and $(\frac{12}{5}, \frac{11}{5})$ solves $2x_1 + x_2 = 7$, $x_1 + 3x_2 = 9$: subtracting twice the second equation from the first gives $-5x_2 = -11$.



The feasible region (shaded). The dashed line is the level set $4x_1 + 3x_2 = 81/5$. Moving it in the direction $\mathbf{c} = (4,3)$ leaves the region, which it last touches at $\mathbf{x}^\star = (12/5, 11/5)$

The level sets $4x_1 + 3x_2 = \text{const}$ are parallel lines, and the value grows in the direction $\mathbf{c} = (4,3)$. Sliding the line in that direction, the last point of the region it meets appears to be $\mathbf{x}^\star = (\frac{12}{5}, \frac{11}{5})$, with value

$$4 \cdot \tfrac{12}{5} + 3 \cdot \tfrac{11}{5} = \tfrac{48+33}{5} = \tfrac{81}{5}.$$

Comparing the corners supports this: they give 0, 12, 15, $\frac{81}{5} = 16.2$, and 9.

A picture is not a proof, and neither is a list of five corners, since we have not proved that an optimum sits at a corner. Here is a proof. Multiply the first constraint by $\frac{9}{5}$ and the second

by $\frac{2}{5}$, both nonnegative, and add:

$$\tfrac{9}{5}(2x_1 + x_2) + \tfrac{2}{5}(x_1 + 3x_2) \le \tfrac{9}{5} \cdot 7 + \tfrac{2}{5} \cdot 9 = \tfrac{81}{5}.$$

The left side is $(\frac{18}{5}+\frac{2}{5})x_1+(\frac{9}{5}+\frac{6}{5})x_2 = 4x_1+3x_2$. So **every** feasible point has $4x_1+3x_2 \le \frac{81}{5}$, and $\mathbf{x}^\star$, which is feasible ($\frac{12}{5} \le 3$ as well), reaches this bound. It is optimal, and the optimal value is $\frac{81}{5}$.

The multipliers $\frac{9}{5}$ and $\frac{2}{5}$ are the heart of the matter. They were chosen so that the combined constraint had exactly the objective as its left side. The best upper bound obtainable this way is itself the answer to a linear program.

18.6.2 The dual program

Multiply the i-th constraint of (P) by a number $y_i \ge 0$ and add. For every feasible $\mathbf{x}$ this gives $\mathbf{y}^\top \mathbf{A}\mathbf{x} \le \mathbf{y}^\top \mathbf{b}$. If the combination dominates the objective coefficient by coefficient, $\mathbf{A}^\top \mathbf{y} \ge \mathbf{c}$, then, since $\mathbf{x} \ge \mathbf{0}$,

$$\mathbf{c}^\top \mathbf{x} \le (\mathbf{A}^\top \mathbf{y})^\top \mathbf{x} = \mathbf{y}^\top \mathbf{A}\mathbf{x} \le \mathbf{y}^\top \mathbf{b}.$$

Each such $\mathbf{y}$ proves an upper bound $\mathbf{b}^\top \mathbf{y}$ on the objective, and the natural question is how small that bound can be made.

The dual program searches for the best upper bound that nonnegative combinations of the constraints can prove.

Definition 18.6.2. Dual Program

The **dual** of the linear program (P) of Definition 18.6.1 is the problem

$$\text{(D)} \qquad \text{minimize } \mathbf{b}^\top \mathbf{y} \quad \text{subject to} \quad \mathbf{A}^\top \mathbf{y} \ge \mathbf{c}, \quad \mathbf{y} \ge \mathbf{0},$$

with $\mathbf{y} \in \mathbb{R}^m$. The words **feasible**, **infeasible**, and **optimal** are used as for (P). (D) is **unbounded** if it is feasible and $\mathbf{b}^\top \mathbf{y}$ takes **arbitrarily negative** values on its feasible region. In this context (P) is called the **primal** program.

The dual has one variable for each constraint of the primal and one constraint for each variable of the primal. The matrix is transposed, the roles of $\mathbf{b}$ and $\mathbf{c}$ are exchanged, and maximization becomes minimization. The dual of (D), written back in inequality form, is (P) again. Exercise (C1: The dual of the dual) checks this, and it explains why a statement about (P) always has a mirror statement about (D).

For the workshop problem the dual is

$$\begin{aligned} \text{minimize} \quad & 7y_1 + 9y_2 + 3y_3 \\ \text{subject to} \quad & 2y_1 + y_2 + y_3 \ge 4, \quad y_1 + 3y_2 \ge 3, \\ & y_1, y_2, y_3 \ge 0, \end{aligned}$$

and the multipliers of Example (The workshop problem, drawn) form the dual feasible point $\mathbf{y}^\star = (\frac{9}{5}, \frac{2}{5}, 0)$, with $\mathbf{b}^\top \mathbf{y}^\star = \frac{81}{5}$. The dual has an economic reading too. y_i is a price for one unit of the i-th resource, and the constraints say that at these prices the resources used by one unit of each product are worth at least the profit it brings.

18.6.3 Weak duality

The computation that motivated (D) is the first theorem, and it is one line long.

Theorem 18.6.1. Weak Duality

If $\mathbf{x}$ is feasible for (P) and $\mathbf{y}$ is feasible for (D), then

$$\mathbf{c}^\top \mathbf{x} \le \mathbf{b}^\top \mathbf{y}.$$

Proof. Since $\mathbf{A}^\top \mathbf{y} - \mathbf{c} \ge \mathbf{0}$ and $\mathbf{x} \ge \mathbf{0}$, the number $(\mathbf{A}^\top \mathbf{y} - \mathbf{c})^\top \mathbf{x}$ is a sum of products of nonnegative numbers, so $\mathbf{c}^\top \mathbf{x} \le (\mathbf{A}^\top \mathbf{y})^\top \mathbf{x} = \mathbf{y}^\top (\mathbf{A}\mathbf{x})$. Since $\mathbf{b} - \mathbf{A}\mathbf{x} \ge \mathbf{0}$ and $\mathbf{y} \ge \mathbf{0}$, in the same way $\mathbf{y}^\top (\mathbf{A}\mathbf{x}) \le \mathbf{y}^\top \mathbf{b} = \mathbf{b}^\top \mathbf{y}$. Chaining the two inequalities proves the theorem. ■

Corollary 18.6.2. Certificates from Weak Duality

(a) If $\mathbf{x}$ is feasible for (P), $\mathbf{y}$ is feasible for (D), and $\mathbf{c}^\top \mathbf{x} = \mathbf{b}^\top \mathbf{y}$, then $\mathbf{x}$ is optimal for (P) and $\mathbf{y}$ is optimal for (D).

(b) If (P) is unbounded, then (D) is infeasible. If (D) is unbounded, then (P) is infeasible.

Idea. Weak duality makes every dual value an upper bound for every primal value. A primal value that meets one of these bounds cannot be beaten, and neither can the bound; an unbounded program leaves no room for any bound at all.

Proof.

(a) For every feasible $\mathbf{x}'$, Theorem 18.6.1 gives $\mathbf{c}^\top \mathbf{x}' \le \mathbf{b}^\top \mathbf{y} = \mathbf{c}^\top \mathbf{x}$, so $\mathbf{x}$ is optimal. For every dual feasible $\mathbf{y}'$, it gives $\mathbf{b}^\top \mathbf{y}' \ge \mathbf{c}^\top \mathbf{x} = \mathbf{b}^\top \mathbf{y}$, so $\mathbf{y}$ is optimal.

(b) If (D) had a feasible point $\mathbf{y}$, Theorem 18.6.1 would bound $\mathbf{c}^\top \mathbf{x}$ above by $\mathbf{b}^\top \mathbf{y}$ for every feasible $\mathbf{x}$, so (P) would not be unbounded. The second statement is the same argument with the roles exchanged. ■

Part (a) is exactly what Example (The workshop problem, drawn) did: $\mathbf{c}^\top \mathbf{x}^\star = \frac{81}{5} = \mathbf{b}^\top \mathbf{y}^\star$, so both points are optimal. It gives a certificate of optimality that can be checked by arithmetic alone. The question is whether such a certificate **always** exists. Weak duality allows the best lower bound to stay strictly below the best upper bound.

Quick check. Without solving anything, give an upper bound for $4x_1 + 3x_2$ on the workshop's feasible region using only the second constraint $x_1 + 3x_2 \le 9$ and $\mathbf{x} \ge \mathbf{0}$. Which dual feasible point is that?

Solution. We need $y_2 \geq 0$ with $y_2 \geq 4$ and $3y_2 \geq 3$, so $y_2 = 4$. Then $4x_1 + 3x_2 \leq 4x_1 + 12x_2 = 4(x_1 + 3x_2) \leq 36$, using $x_2 \geq 0$. This is the dual feasible point $\mathbf{y} = (0, 4, 0)$, with $\mathbf{b}^\top \mathbf{y} = 36$. It is a valid bound, but a poor one compared with the optimal $\frac{81}{5}$.

18.6.4 Strong duality

The main theorem says that the gap allowed by weak duality never occurs, and it gives the complete list of what can happen.

Theorem 18.6.3. Strong Duality for Linear Programs

Let $\mathbf{A} \in M_{m\times n}(\mathbb{R})$, $\mathbf{b} \in \mathbb{R}^m$, and $\mathbf{c} \in \mathbb{R}^n$, and let (P) and (D) be as in Definition 18.6.1 and Definition 18.6.2.

(a) If (P) and (D) are both feasible, then both have optimal solutions, and their optimal values are **equal**: there are feasible $\mathbf{x}^\star$ and $\mathbf{y}^\star$ with $\mathbf{c}^\top \mathbf{x}^\star = \mathbf{b}^\top \mathbf{y}^\star$.

(b) If (P) is feasible and (D) is infeasible, then (P) is unbounded.

(c) If (D) is feasible and (P) is infeasible, then (D) is unbounded.

Idea. Parts (b) and (c) are each one application of Corollary 18.5.7. The infeasibility of one program yields a vector that the other program can move along forever, with its objective improving without bound.

For (a) we want a single pair $(\mathbf{x}, \mathbf{y})$ that is feasible for both programs and has $\mathbf{c}^\top \mathbf{x} \geq \mathbf{b}^\top \mathbf{y}$. Weak duality then forces equality, and Corollary 18.6.2 forces optimality. The pair is a solution of one big system of inequalities in $(\mathbf{x}, \mathbf{y})$, so we ask Farkas whether it has one. If it had none, Farkas would hand us a certificate $(\mathbf{u}, \mathbf{v}, t)$. The number t is the multiplier on the row $\mathbf{b}^\top \mathbf{y} \leq \mathbf{c}^\top \mathbf{x}$. If $t > 0$, rescaling the certificate by $1/t$ produces a primal and a dual feasible point that violate weak duality. If $t = 0$, the certificate is a pair of directions along which one of the two programs is unbounded, which contradicts weak duality against a feasible point of the other.

Proof. **(b)** Let $\mathbf{x}_0$ be feasible for (P). That (D) is infeasible says that the system $(-\mathbf{A}^\top)\mathbf{y} \leq -\mathbf{c}$, $\mathbf{y} \geq \mathbf{0}$ has no solution. By Corollary 18.5.7, applied with $\mathbf{M} = -\mathbf{A}^\top$ and $\mathbf{d} = -\mathbf{c}$, there is $\mathbf{w} \in \mathbb{R}^n$ with $\mathbf{w} \geq \mathbf{0}$, $(-\mathbf{A}^\top)^\top \mathbf{w} = -\mathbf{A}\mathbf{w} \geq \mathbf{0}$, and $-\mathbf{c}^\top \mathbf{w} < 0$. That is, $\mathbf{A}\mathbf{w} \leq \mathbf{0}$ and $\mathbf{c}^\top \mathbf{w} > 0$. For every $s \geq 0$ the point $\mathbf{x}_0 + s\mathbf{w}$ is feasible, since $\mathbf{A}(\mathbf{x}_0 + s\mathbf{w}) = \mathbf{A}\mathbf{x}_0 + s\mathbf{A}\mathbf{w} \leq \mathbf{b}$ and $\mathbf{x}_0 + s\mathbf{w} \geq \mathbf{0}$. Its objective value is $\mathbf{c}^\top \mathbf{x}_0 + s\,\mathbf{c}^\top \mathbf{w}$, which is larger than any given number once s is large enough, because $\mathbf{c}^\top \mathbf{w} > 0$. So (P) is unbounded.

(c) Let $\mathbf{y}_0$ be feasible for (D). That (P) is infeasible says that $\mathbf{A}\mathbf{x} \leq \mathbf{b}$, $\mathbf{x} \geq \mathbf{0}$ has no solution. By Corollary 18.5.7 with $\mathbf{M} = \mathbf{A}$ and $\mathbf{d} = \mathbf{b}$, there is $\mathbf{w} \in \mathbb{R}^m$ with $\mathbf{w} \geq \mathbf{0}$, $\mathbf{A}^\top \mathbf{w} \geq \mathbf{0}$, and $\mathbf{b}^\top \mathbf{w} < 0$. For $s \geq 0$ the point $\mathbf{y}_0 + s\mathbf{w}$ is feasible for (D), since $\mathbf{A}^\top(\mathbf{y}_0 + s\mathbf{w}) \geq \mathbf{A}^\top \mathbf{y}_0 \geq \mathbf{c}$ and $\mathbf{y}_0 + s\mathbf{w} \geq \mathbf{0}$. Its value $\mathbf{b}^\top \mathbf{y}_0 + s\,\mathbf{b}^\top \mathbf{w}$ is below any given number for large s. So (D) is unbounded.

(a) Let $\mathbf{x}_0$ be feasible for (P) and $\mathbf{y}_0$ feasible for (D). Consider the $(m + n + 1) \times (n + m)$

matrix and the vector

$$\mathbf{M} = \begin{pmatrix} \mathbf{A} & \mathbf{O} \\ \mathbf{O} & -\mathbf{A}^{\top} \\ -\mathbf{c}^{\top} & \mathbf{b}^{\top} \end{pmatrix}, \qquad \mathbf{d} = \begin{pmatrix} \mathbf{b} \\ -\mathbf{c} \\ 0 \end{pmatrix}.$$

For $\mathbf{z} = (\mathbf{x}, \mathbf{y})$ with $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{y} \in \mathbb{R}^m$, the system $\mathbf{M}\mathbf{z} \le \mathbf{d}$, $\mathbf{z} \ge \mathbf{0}$ says exactly that

$$\mathbf{A}\mathbf{x} \le \mathbf{b}, \quad \mathbf{A}^{\top}\mathbf{y} \ge \mathbf{c}, \quad \mathbf{b}^{\top}\mathbf{y} \le \mathbf{c}^{\top}\mathbf{x}, \quad \mathbf{x} \ge \mathbf{0}, \quad \mathbf{y} \ge \mathbf{0}. \tag{18.6.1}$$

Suppose Eq. 18.6.1 has a solution $(\mathbf{x}^{\star}, \mathbf{y}^{\star})$. Then $\mathbf{x}^{\star}$ is feasible for (P) and $\mathbf{y}^{\star}$ is feasible for (D), so Theorem 18.6.1 gives $\mathbf{c}^{\top}\mathbf{x}^{\star} \le \mathbf{b}^{\top}\mathbf{y}^{\star}$. Together with $\mathbf{b}^{\top}\mathbf{y}^{\star} \le \mathbf{c}^{\top}\mathbf{x}^{\star}$, this gives equality, and by Corollary 18.6.2 (a) both points are optimal. That is (a).

It remains to show that Eq. 18.6.1 has a solution. Suppose not. By Corollary 18.5.7 there is $\mathbf{w} \ge \mathbf{0}$ with $\mathbf{M}^{\top}\mathbf{w} \ge \mathbf{0}$ and $\mathbf{d}^{\top}\mathbf{w} < 0$. Write $\mathbf{w} = (\mathbf{u}, \mathbf{v}, t)$ with $\mathbf{u} \in \mathbb{R}^m$, $\mathbf{v} \in \mathbb{R}^n$, and $t \in \mathbb{R}$; then $\mathbf{u} \ge \mathbf{0}$, $\mathbf{v} \ge \mathbf{0}$, and $t \ge 0$. Since

$$\mathbf{M}^{\top} = \begin{pmatrix} \mathbf{A}^{\top} & \mathbf{O} & -\mathbf{c} \\ \mathbf{O} & -\mathbf{A} & \mathbf{b} \end{pmatrix},$$

the conditions read

$$\mathbf{A}^{\top}\mathbf{u} \ge t\mathbf{c}, \qquad \mathbf{A}\mathbf{v} \le t\mathbf{b}, \qquad \mathbf{b}^{\top}\mathbf{u} < \mathbf{c}^{\top}\mathbf{v}. \tag{18.6.2}$$

Case 1: $t > 0$. Dividing Eq. 18.6.2 by t shows that $\mathbf{v}/t \ge \mathbf{0}$ is feasible for (P), that $\mathbf{u}/t \ge \mathbf{0}$ is feasible for (D), and that $\mathbf{b}^{\top}(\mathbf{u}/t) < \mathbf{c}^{\top}(\mathbf{v}/t)$. This contradicts Theorem 18.6.1.

Case 2: $t = 0$. Now $\mathbf{A}^{\top}\mathbf{u} \ge \mathbf{0}$, $\mathbf{A}\mathbf{v} \le \mathbf{0}$, and $\mathbf{b}^{\top}\mathbf{u} < \mathbf{c}^{\top}\mathbf{v}$, with $\mathbf{u}, \mathbf{v} \ge \mathbf{0}$. The strict inequality forces $\mathbf{c}^{\top}\mathbf{v} > 0$ or $\mathbf{b}^{\top}\mathbf{u} < 0$, since otherwise $\mathbf{b}^{\top}\mathbf{u} \ge 0 \ge \mathbf{c}^{\top}\mathbf{v}$. If $\mathbf{c}^{\top}\mathbf{v} > 0$, then, exactly as in the proof of (b), the points $\mathbf{x}_0 + s\mathbf{v}$ with $s \ge 0$ are feasible for (P) with objective values tending to $+\infty$. That contradicts the bound $\mathbf{c}^{\top}\mathbf{x} \le \mathbf{b}^{\top}\mathbf{y}_0$ that Theorem 18.6.1 gives for every feasible $\mathbf{x}$. If $\mathbf{b}^{\top}\mathbf{u} < 0$, then, as in the proof of (c), the points $\mathbf{y}_0 + s\mathbf{u}$ are feasible for (D) with values tending to $-\infty$, which contradicts $\mathbf{b}^{\top}\mathbf{y} \ge \mathbf{c}^{\top}\mathbf{x}_0$.

Both cases are impossible, so Eq. 18.6.1 has a solution, and (a) is proved. This proves the theorem.

■

Together with Corollary 18.6.2 (b), the theorem sorts every pair (P), (D) into exactly one of four cases:

(P)	(D)	what happens
feasible	feasible	both optimal, equal values (a)
feasible	infeasible	(P) unbounded (b)
infeasible	feasible	(D) unbounded (c)
infeasible	infeasible	nothing more to say

All four occur. The workshop problem is the first. The one-variable program “maximize x

subject to $-x \le -1$, $x \ge 0$" is the second: its dual, "minimize $-y$ subject to $-y \ge 1$, $y \ge 0$", is infeasible. "Maximize $0 \cdot x$ subject to $x \le -1$, $x \ge 0$" is the third: its dual, "minimize $-y$ subject to $y \ge 0$", is unbounded. For the fourth, take

$$\mathbf{A} = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}, \qquad \mathbf{b} = \begin{pmatrix} -1 \\ -1 \end{pmatrix}, \qquad \mathbf{c} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

Adding the two primal constraints gives $0 \le -2$, and adding the two dual constraints gives $0 \ge 2$, so neither program is feasible.

The table has a consequence that is easy to state and not at all obvious.

Corollary 18.6.4. Bounded Linear Programs Attain Their Optimum

If (P) is feasible and its objective is bounded above on the feasible region, that is, $\mathbf{c}^{\top}\mathbf{x} \le N$ for some N and every feasible $\mathbf{x}$, then (P) has an optimal solution. The same holds for (D) with "bounded below".

Proof. Since (P) is feasible and not unbounded, part (b) of Theorem 18.6.3 shows that (D) is not infeasible. So both programs are feasible, and part (a) gives an optimal solution of (P). The statement for (D) follows in the same way from part (c). ∎

Why is this not obvious? The feasible region is closed, but it may be unbounded, so the extreme value theorem (A4) of Chapter 16's introduction does not apply to it. And a linear function that is bounded above on a closed convex set need not attain its supremum. On the region $R = \{(x_1, x_2) : x_1 > 0,\ x_1 x_2 \ge 1\}$, bounded by a branch of a hyperbola, the linear function $-x_2$ is negative everywhere, and it takes values arbitrarily close to 0 at the points $(k, 1/k)$, so its supremum is 0. That supremum is never reached. The region R is the set A of Example (A hyperbola and its asymptote), since $x_1 > 0$ and $x_1 x_2 \ge 1$ force $x_2 > 0$, and that example shows it is closed and convex. But it is not cut out by **finitely many** linear inequalities. That finiteness is what made the cones in the proof finitely generated, and hence closed, by Theorem 18.5.4. The corollary is where finite generation pays off.

Warning. **Strong duality is a theorem about linear programs, not about optimization in general.** For programs whose objective or constraints are convex but not linear, one can still write down a dual problem, and weak duality still holds, but the two optimal values may differ: a **duality gap**. Extra hypotheses are needed to close it, of which the best known is a strictly feasible point (Slater's condition). That theory belongs to convex optimization and is not developed in this book. For linear programs no hypothesis beyond feasibility is needed, as Theorem 18.6.3 shows.

18.6.5 Complementary slackness

Strong duality says that at the optimum the two chains of inequalities in the proof of Theorem 18.6.1 are equalities. Reading that off term by term gives a test that locates the dual solution once the primal one is known.

Theorem 18.6.5. Complementary Slackness

Let $\mathbf{x}$ be feasible for (P) and $\mathbf{y}$ feasible for (D). Then $\mathbf{x}$ and $\mathbf{y}$ are **both** optimal if and only if

(i) for every i, $y_i > 0$ implies $(\mathbf{A}\mathbf{x})_i = b_i$; and
(ii) for every j, $x_j > 0$ implies $(\mathbf{A}^\top \mathbf{y})_j = c_j$.

Idea. The gap $\mathbf{b}^\top \mathbf{y} - \mathbf{c}^\top \mathbf{x}$ is a sum of products of nonnegative numbers, one product for each constraint of either program. Optimality means the gap is zero, which means every product is zero.

Proof. Expanding and using $\mathbf{y}^\top \mathbf{A}\mathbf{x} = \mathbf{x}^\top \mathbf{A}^\top \mathbf{y}$,

$$\mathbf{b}^\top \mathbf{y} - \mathbf{c}^\top \mathbf{x} = \mathbf{y}^\top (\mathbf{b} - \mathbf{A}\mathbf{x}) + \mathbf{x}^\top (\mathbf{A}^\top \mathbf{y} - \mathbf{c}). \tag{18.6.3}$$

The first term on the right is $\sum_i y_i (\mathbf{b} - \mathbf{A}\mathbf{x})_i$, a sum of products of nonnegative numbers, and so is the second. Condition (i) says that every product in the first sum is zero, and (ii) says the same for the second. So (i) and (ii) together hold exactly when the right side of Eq. 18.6.3 is 0, because a sum of nonnegative numbers is zero only when every term is.

($\Leftarrow$) If (i) and (ii) hold, then $\mathbf{c}^\top \mathbf{x} = \mathbf{b}^\top \mathbf{y}$, and both points are optimal by Corollary 18.6.2 (a).

($\Rightarrow$) If both are optimal, then both programs are feasible, and by Theorem 18.6.3 (a) their optimal values are equal. So $\mathbf{c}^\top \mathbf{x} = \mathbf{b}^\top \mathbf{y}$, the right side of Eq. 18.6.3 is 0, and (i) and (ii) hold.

■

In words: a constraint with a positive price must be tight, and a product that is actually made must exactly pay for its resources. For the workshop problem, $\mathbf{x}^\star = (\frac{12}{5}, \frac{11}{5})$ has both entries positive, so (ii) requires both dual constraints to be tight:

$$2y_1 + y_2 + y_3 = 4, \qquad y_1 + 3y_2 = 3.$$

The third primal constraint is slack at $\mathbf{x}^\star$, since $\frac{12}{5} < 3$, so (i) forces $y_3 = 0$. The two equations then give $y_1 = \frac{9}{5}$ and $y_2 = \frac{2}{5}$, both ≥ 0, which recovers the multipliers of Example (The workshop problem, drawn). So the multipliers were not a lucky guess: complementary slackness **finds** them from the primal solution.

18.6.6 Traffic, with inequalities

We return to Chapter 3's town center, to answer the remark made there. Recall from Example (Traffic in a town center) the four intersections, the five one-way streets $1 \to 2$, $2 \to 3$, $3 \to 4$, $4 \to 1$, $1 \to 3$ with flows $f_1, \dots, f_5$, and the incidence matrix $\mathbf{N}$. Conservation of cars is $\mathbf{N}\mathbf{f} = \mathbf{s}$ with $\mathbf{s} = (40, 20, -35, -25)$, and the flows that conserve cars are

$$\mathbf{f} = (40 - \beta + \alpha,\ 60 - \beta + \alpha,\ 25 + \alpha,\ \alpha,\ \beta), \qquad \alpha, \beta \in \mathbb{R}.$$

Chapter 3 asked, with street $4 \to 1$ closed, for the largest possible flow on the diagonal $1 \to 3$, and found 40 by inspecting the one remaining parameter. That was already a linear program: maximize f_5 subject to $\mathbf{Nf} = \mathbf{s}$, $\mathbf{f} \geq \mathbf{0}$, $f_4 = 0$. Its answer has a one-line certificate. The first row of $\mathbf{Nf} = \mathbf{s}$, the balance at intersection 1, reads $f_1 - f_4 + f_5 = 40$, so $f_5 = 40 - f_1 + f_4 = 40 - f_1 \leq 40$ because $f_1 \geq 0$.

Example. Capacities on every street

Suppose street $4 \to 1$ is open again, but every street can carry at most 35 cars per minute. What is the **smallest** flow the diagonal street $1 \to 3$ can carry, and how would you convince someone that it cannot be smaller?

Solution. The linear program is: minimize f_5 subject to $\mathbf{Nf} = \mathbf{s}$ and $\mathbf{0} \leq \mathbf{f} \leq 35 \cdot \mathbf{1}$. In the parameters, the capacity of street $2 \to 3$ reads $60 - \beta + \alpha \leq 35$, so $\beta \geq 25 + \alpha \geq 25$, since $\alpha = f_4 \geq 0$. The value $\beta = 25$ is achieved: with $\alpha = 0$ the flow is

$$\mathbf{f} = (15, 35, 25, 0, 25),$$

every entry of which lies between 0 and 35. So the minimum is 25.

The certificate, in the spirit of the dual, is a nonnegative combination of the constraints. Add the balance equations at intersections 1 and 2, $f_1 - f_4 + f_5 = 40$ and $-f_1 + f_2 = 20$:

$$f_2 - f_4 + f_5 = 60, \qquad \text{so} \qquad f_5 = 60 - f_2 + f_4 \;\geq\; 60 - 35 + 0 = 25,$$

using the capacity $f_2 \leq 35$ and the sign constraint $f_4 \geq 0$. In words: at least sixty cars per minute arrive at the pair of intersections $\{1, 2\}$ (sixty from outside, plus whatever street $4 \to 1$ brings), and only two streets leave that pair, $2 \to 3$ and the diagonal. The first can take at most 35, so the diagonal must take at least 25. This is a **cut** argument, and it is exactly what the dual program of a flow problem computes. The balance equations carry multipliers of either sign, the "potentials" of the intersections, and here they are 1 at intersections 1 and 2 and 0 at the others.

This is the answer to Chapter 3's remark. Linear algebra describes the conserving flows as a coset of the null space of $\mathbf{N}$. Linear programming chooses among them, and duality explains each choice with a certificate that can be read as a statement about the network.

Remark. Computing optimal solutions of large linear programs is a subject in its own right. The best known method, the simplex method, moves from corner to corner of the feasible region, each time improving the objective. This book does not develop algorithms for linear programming. Section 9 returns to the corners themselves, as the vertices of a polyhedron.

18.6.7 Exercises

18.6.7.1 A. Check your understanding

Exercise. A1

(a) Write down the dual of: maximize $2x_1 + x_2$ subject to $x_1 + x_2 \leq 3$, $x_1 - x_2 \leq 1$, $\mathbf{x} \geq \mathbf{0}$.

(b) State weak duality and strong duality.

(c) Determine whether the following is correct, and justify your answer: if (P) is infeasible, then (D) is unbounded.

(d) Determine whether the following is correct, and justify your answer: if (P) is feasible and $\mathbf{c}^\top \mathbf{x} \le 100$ for every feasible $\mathbf{x}$, then (P) has an optimal solution.

(e) At an optimal pair $\mathbf{x}, \mathbf{y}$, suppose $x_2 > 0$. What does complementary slackness say about the second constraint of (D)?

Solution.

(a) Minimize $3y_1 + y_2$ subject to $y_1 + y_2 \ge 2$, $y_1 - y_2 \ge 1$, $\mathbf{y} \ge \mathbf{0}$. The columns of $\mathbf{A}$ become the rows of $\mathbf{A}^\top$, and $\mathbf{b} = (3, 1)$ and $\mathbf{c} = (2, 1)$ exchange roles.

(b) Weak duality (Theorem 18.6.1): if $\mathbf{x}$ is feasible for (P) and $\mathbf{y}$ for (D), then $\mathbf{c}^\top \mathbf{x} \le \mathbf{b}^\top \mathbf{y}$. Strong duality (Theorem 18.6.3): if both are feasible, both have optimal solutions with equal values. If exactly one is feasible, that one is unbounded.

(c) Incorrect. Both programs can be infeasible, as for $\mathbf{A} = \left(\begin{smallmatrix} 1 & -1 \\ -1 & 1 \end{smallmatrix}\right)$, $\mathbf{b} = (-1, -1)$, $\mathbf{c} = (1, 1)$ in this section. What is true, by Theorem 18.6.3 (c), is that if (P) is infeasible **and** (D) is feasible, then (D) is unbounded.

(d) Correct, by Corollary 18.6.4. The bound shows that (P) is not unbounded, so by Theorem 18.6.3 (b) the dual is feasible, and then (a) gives an optimal solution.

(e) By Theorem 18.6.5 (ii), the second dual constraint is tight: $(\mathbf{A}^\top \mathbf{y})_2 = c_2$.

18.6.7.2 B. Practice

Exercise. B1: Solve and certify

Consider: maximize $x_1 + 2x_2$ subject to

$$x_1 + x_2 \le 5, \qquad -x_1 + x_2 \le 1, \qquad 2x_1 - x_2 \le 4, \qquad \mathbf{x} \ge \mathbf{0}.$$

(a) Sketch the feasible region and guess an optimal solution.

(b) Write down the dual, and use complementary slackness to find a candidate dual solution.

(c) Prove that both candidates are optimal, and state the optimal value.

Solution.

(a) The region is the pentagon with corners $(0,0)$, $(2,0)$, $(3,2)$, $(2,3)$, and $(0,1)$, where the objective takes the values $0, 2, 7, 8, 2$. The guess is $\mathbf{x}^\star = (2, 3)$, where $x_1 + x_2 = 5$ meets $-x_1 + x_2 = 1$.

(b) The dual is: minimize $5y_1 + y_2 + 4y_3$ subject to $y_1 - y_2 + 2y_3 \ge 1$, $y_1 + y_2 - y_3 \ge 2$, $\mathbf{y} \ge \mathbf{0}$. At $\mathbf{x}^\star$ the third constraint is slack, since $2 \cdot 2 - 3 = 1 < 4$, so (i) of Theorem 18.6.5 requires $y_3 = 0$. Both $x_j^\star > 0$, so (ii) requires both dual constraints to be tight: $y_1 - y_2 = 1$ and $y_1 + y_2 = 2$. Hence $\mathbf{y}^\star = (\tfrac{3}{2}, \tfrac{1}{2}, 0)$.

(c) $\mathbf{x}^\star$ is feasible: $5 \le 5$, $1 \le 1$, $1 \le 4$, and $\mathbf{x}^\star \ge \mathbf{0}$. $\mathbf{y}^\star$ is feasible: $\mathbf{y}^\star \ge \mathbf{0}$, $\tfrac{3}{2} - \tfrac{1}{2} = 1 \ge 1$, and $\tfrac{3}{2} + \tfrac{1}{2} = 2 \ge 2$. The objective values are $\mathbf{c}^\top \mathbf{x}^\star = 2 + 6 = 8$ and $\mathbf{b}^\top \mathbf{y}^\star = \tfrac{15}{2} + \tfrac{1}{2} = 8$. They are equal, so both points are optimal by Corollary 18.6.2 (a), and the optimal

value is 8.

Exercise. B2: Which case?

For each program, write down its dual, and determine which of the four cases of Theorem 18.6.3 occurs. Justify your answer, giving the optimal values where they exist.

(a) Maximize $x_1 + x_2$ subject to $x_1 - x_2 \le 1$, $\mathbf{x} \ge \mathbf{0}$.

(b) Maximize $-x_1$ subject to $x_1 \le 2$, $-x_1 \le -1$, $x_1 \ge 0$.

(c) Maximize x_1 subject to $x_1 + x_2 \le -1$, $\mathbf{x} \ge \mathbf{0}$.

Solution.

(a) The dual is: minimize y subject to $y \ge 1$, $-y \ge 1$, $y \ge 0$. The last two constraints conflict, so the dual is infeasible. The primal is feasible, with $\mathbf{0}$ a feasible point, so by Theorem 18.6.3 (b) it is unbounded. Directly, $(0, s)$ is feasible for every $s \ge 0$, with value s.

(b) The dual is: minimize $2y_1 - y_2$ subject to $y_1 - y_2 \ge -1$, $\mathbf{y} \ge \mathbf{0}$. Both programs are feasible: $x_1 = 1$ and $\mathbf{y} = (0, 1)$. Their values are -1 and $2 \cdot 0 - 1 = -1$, which are equal, so both are optimal by Corollary 18.6.2 (a), with optimal value -1. This is the first case.

(c) The primal is infeasible, since $x_1 + x_2 \ge 0 > -1$ for $\mathbf{x} \ge \mathbf{0}$. The dual is: minimize $-y$ subject to $y \ge 1$, $y \ge 0$, $y \ge 0$. It is feasible, and $-y$ takes arbitrarily negative values, so it is unbounded, as Theorem 18.6.3 (c) predicts.

Exercise. B3: A network with one flow

In the traffic network of this section, close street $4 \to 1$ and give every other street a capacity of 30 cars per minute. Show that there is exactly one feasible flow. For its diagonal flow $f_5 = 30$, give a certificate, in the style of Example (Capacities on every street), that $f_5 \ge 30$ for every feasible flow, and one that $f_5 \le 30$.

Solution. With $f_4 = \alpha = 0$ the conserving flows are $\mathbf{f} = (40 - \beta, 60 - \beta, 25, 0, \beta)$. The capacity of street $2 \to 3$ gives $60 - \beta \le 30$, so $\beta \ge 30$. The capacity of the diagonal gives $\beta \le 30$. So $\beta = 30$, and the only candidate is $\mathbf{f} = (10, 30, 25, 0, 30)$, which satisfies $0 \le f_i \le 30$ for every street. It is the unique feasible flow.

Lower bound: adding the balance equations at intersections 1 and 2 gives $f_2 - f_4 + f_5 = 60$, so $f_5 = 60 - f_2 + f_4 = 60 - f_2 \ge 30$ by the capacity $f_2 \le 30$, using $f_4 = 0$. Upper bound: $f_5 \le 30$ is itself a capacity constraint. Each certificate is a nonnegative combination of constraints, with free multipliers on the balance equations, as in the example.

18.6.7.3 C. Going deeper

Exercise. C1: The dual of the dual

(a) Rewrite (D) as a program in the inequality form of Definition 18.6.1, with data $(-\mathbf{A}^\top, -\mathbf{c}, -\mathbf{b})$. Show that the dual of that program, rewritten as a maximization, is (P).

(b) Deduce part (c) of Theorem 18.6.3 from part (b).

Solution.

(a) Minimizing $\mathbf{b}^\top\mathbf{y}$ is maximizing $(-\mathbf{b})^\top\mathbf{y}$, and $\mathbf{A}^\top\mathbf{y} \geq \mathbf{c}$ is $(-\mathbf{A}^\top)\mathbf{y} \leq -\mathbf{c}$. So (D) is the inequality-form program (D′): maximize $(-\mathbf{b})^\top\mathbf{y}$ subject to $(-\mathbf{A}^\top)\mathbf{y} \leq -\mathbf{c}$, $\mathbf{y} \geq \mathbf{0}$, whose data are $\mathbf{A}' = -\mathbf{A}^\top$, $\mathbf{b}' = -\mathbf{c}$, $\mathbf{c}' = -\mathbf{b}$. By Definition 18.6.2 its dual is: minimize $\mathbf{b}'^\top\mathbf{x} = (-\mathbf{c})^\top\mathbf{x}$ subject to $\mathbf{A}'^\top\mathbf{x} = -\mathbf{A}\mathbf{x} \geq \mathbf{c}' = -\mathbf{b}$, $\mathbf{x} \geq \mathbf{0}$. That is: maximize $\mathbf{c}^\top\mathbf{x}$ subject to $\mathbf{A}\mathbf{x} \leq \mathbf{b}$, $\mathbf{x} \geq \mathbf{0}$, which is (P). Its optimal value is the negative of the minimum above.

(b) Suppose (D) is feasible and (P) is infeasible. Then (D′) is feasible, with the same feasible points, and its dual, which by (a) has the same feasible points as (P), is infeasible. By Theorem 18.6.3 (b) applied to (D′), the objective $(-\mathbf{b})^\top\mathbf{y}$ takes arbitrarily large values on the feasible region. So $\mathbf{b}^\top\mathbf{y}$ takes arbitrarily negative values, and (D) is unbounded.

Exercise. C2: Free variables

Let $\mathbf{A} \in M_{m\times n}(\mathbb{R})$, $\mathbf{b} \in \mathbb{R}^m$, $\mathbf{c} \in \mathbb{R}^n$, and consider the program with **no sign constraint** on $\mathbf{x}$:

$$(\mathrm{P}_{\mathrm{free}}) \qquad \text{maximize } \mathbf{c}^\top\mathbf{x} \text{ subject to } \mathbf{A}\mathbf{x} \leq \mathbf{b}.$$

(a) Prove weak duality between $(\mathrm{P}_{\mathrm{free}})$ and $(\mathrm{D}_{\mathrm{eq}})$: minimize $\mathbf{b}^\top\mathbf{y}$ subject to $\mathbf{A}^\top\mathbf{y} = \mathbf{c}$, $\mathbf{y} \geq \mathbf{0}$.

(b) Writing $\mathbf{x} = \mathbf{u} - \mathbf{v}$ with $\mathbf{u}, \mathbf{v} \geq \mathbf{0}$, show that $(\mathrm{P}_{\mathrm{free}})$ has the same values as an inequality-form program whose dual is $(\mathrm{D}_{\mathrm{eq}})$. Deduce that if both are feasible, both have optimal solutions with equal values.

Solution.

(a) If $\mathbf{A}\mathbf{x} \leq \mathbf{b}$ and $\mathbf{A}^\top\mathbf{y} = \mathbf{c}$ with $\mathbf{y} \geq \mathbf{0}$, then $\mathbf{c}^\top\mathbf{x} = (\mathbf{A}^\top\mathbf{y})^\top\mathbf{x} = \mathbf{y}^\top\mathbf{A}\mathbf{x} \leq \mathbf{y}^\top\mathbf{b}$, the last step because $\mathbf{y} \geq \mathbf{0}$ and $\mathbf{b} - \mathbf{A}\mathbf{x} \geq \mathbf{0}$. No sign condition on $\mathbf{x}$ is needed, because the first step is an equality.

(b) Every $\mathbf{x} \in \mathbb{R}^n$ can be written as $\mathbf{u} - \mathbf{v}$ with $\mathbf{u}, \mathbf{v} \geq \mathbf{0}$: take $u_j = \max(x_j, 0)$ and $v_j = \max(-x_j, 0)$. So $(\mathrm{P}_{\mathrm{free}})$ has the same feasible objective values as the inequality-form program

$$\text{maximize } \begin{pmatrix}\mathbf{c}\\ -\mathbf{c}\end{pmatrix}^\top \begin{pmatrix}\mathbf{u}\\ \mathbf{v}\end{pmatrix} \text{ subject to } \begin{pmatrix}\mathbf{A} & -\mathbf{A}\end{pmatrix}\begin{pmatrix}\mathbf{u}\\ \mathbf{v}\end{pmatrix} \leq \mathbf{b}, \quad \begin{pmatrix}\mathbf{u}\\ \mathbf{v}\end{pmatrix} \geq \mathbf{0},$$

with an optimal $(\mathbf{u}, \mathbf{v})$ giving an optimal $\mathbf{x} = \mathbf{u} - \mathbf{v}$, and conversely. By Definition 18.6.2 its dual is: minimize $\mathbf{b}^\top\mathbf{y}$ subject to $\begin{pmatrix}\mathbf{A} & -\mathbf{A}\end{pmatrix}^\top\mathbf{y} \geq (\mathbf{c}, -\mathbf{c})$, $\mathbf{y} \geq \mathbf{0}$. The constraint reads $\mathbf{A}^\top\mathbf{y} \geq \mathbf{c}$ and $-\mathbf{A}^\top\mathbf{y} \geq -\mathbf{c}$, that is, $\mathbf{A}^\top\mathbf{y} = \mathbf{c}$. So the dual is $(\mathrm{D}_{\mathrm{eq}})$. If both programs are feasible, Theorem 18.6.3 (a) gives optimal $(\mathbf{u}^\star, \mathbf{v}^\star)$ and $\mathbf{y}^\star$ with equal values, and $\mathbf{x}^\star = \mathbf{u}^\star - \mathbf{v}^\star$ is optimal for $(\mathrm{P}_{\mathrm{free}})$.

18.7 Matrix Games and the Minimax Theorem

Chapter 17 was built on a min–max formula: the Courant–Fischer theorem describes each eigenvalue of a Hermitian matrix as a minimum of maxima over subspaces. This section proves a second min–max theorem that looks similar and works for a completely different reason. It concerns a game in which two players choose simultaneously, and it says that if the players are allowed to **randomize**, it does not matter which of them has to reveal a plan first. The theorem is due to von Neumann, and with linear programming duality in hand (§06) it takes a page to prove.

The two theorems are easy to confuse, and a reader who has just finished Chapter 17 will be tempted to. So the section also says precisely how they differ. Courant–Fischer is a statement about **subspaces**, and its equality comes from a **dimension count**. The minimax theorem is a statement about **simplices**, and its equality comes from **convexity**.

Throughout, the field is $\mathbb{R}$, $\mathbf{1}$ is the all-ones vector of whatever length the context requires, and $\mathbf{J} = \mathbf{1}\mathbf{1}^{\top}$ is the all-ones matrix. Inequalities between vectors are entrywise, as in §05.

18.7.1 Matrix games

Two players, Row and Column, play a game described by a real matrix $\mathbf{A} \in M_{m\times n}(\mathbb{R})$, the **payoff matrix**. Row secretly chooses a row i, Column secretly chooses a column j, both reveal their choices, and Column pays Row the amount a_{ij}. A negative entry means that Row pays Column. Whatever one player wins, the other loses, so the game is called **zero-sum**. Row wants a_{ij} large and Column wants it small. A choice of a row, or of a column, is a **pure strategy**.

Suppose Row must announce a row in advance. Column will then answer with the smallest entry in that row, so the best Row can guarantee is

$$\max_{i} \min_{j} a_{ij}.$$

If instead Column must announce first, the best Column can guarantee is to pay at most $\min_j \max_i a_{ij}$. Announcing first is never an advantage, and that holds for any payoff function whatsoever.

Lemma 18.7.1. Max–Min Never Exceeds Min–Max

Let X and Y be sets and $f\colon X \times Y \to \mathbb{R}$ a function for which the extrema below exist. Then

$$\max_{x\in X} \min_{y\in Y} f(x,y) \;\le\; \min_{y\in Y} \max_{x\in X} f(x,y).$$

Proof. Let $x_0 \in X$ and $y_0 \in Y$. Then

$$\min_{y\in Y} f(x_0,y) \;\le\; f(x_0,y_0) \;\le\; \max_{x\in X} f(x,y_0).$$

The left side does not depend on y_0 and the right side does not depend on x_0. Taking the maximum over x_0 on the left and then the minimum over y_0 on the right proves the lemma. ∎

The inequality can be strict. In the game with

$$\mathbf{A} = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix},$$

Row wins when the two choices match and loses when they differ. Every row contains a -1, so $\max_i \min_j a_{ij} = -1$. Every column contains a 1, so $\min_j \max_i a_{ij} = 1$. Whoever announces first loses a unit, and there is a gap of 2 between what the players can guarantee.

Some games have no gap. In

$$\mathbf{A} = \begin{pmatrix} 3 & 1 & 4 \\ 2 & 0 & 1 \end{pmatrix}$$

the row minima are 1 and 0, whose maximum is 1, and the column maxima are $3, 1, 4$, whose minimum is also 1. The entry $a_{12} = 1$ is the smallest in its row and the largest in its column. Such an entry is a **saddle point**. If Row plays row 1 and Column plays column 2, neither can gain by changing unilaterally, and the game is settled in pure strategies. The first game has no saddle point, and in that game the natural advice is: do not be predictable.

18.7.2 Mixed strategies

The way to be unpredictable is to randomize. Row chooses row i with probability x_i, and Column chooses column j with probability y_j, independently. The probabilities of the m rows form a vector with nonnegative entries summing to 1, that is, a point of the standard simplex of Definition 18.1.2,

$$\Delta_m = \{\mathbf{x} \in \mathbb{R}^m : \mathbf{x} \geq \mathbf{0} \text{ and } x_1 + \cdots + x_m = 1\}.$$

A mixed strategy is a probability distribution over the pure strategies, and the simplex is the set of all of them.

Definition 18.7.1. Mixed Strategy

In the game with payoff matrix $\mathbf{A} \in M_{m\times n}(\mathbb{R})$, a **mixed strategy** for Row is a vector $\mathbf{x} \in \Delta_m$, a mixed strategy for Column is a vector $\mathbf{y} \in \Delta_n$, and the **expected payoff** is

$$\mathbf{x}^\top \mathbf{A}\mathbf{y} = \sum_{i=1}^{m}\sum_{j=1}^{n} x_i\, a_{ij}\, y_j.$$

In words: a mixed strategy assigns to each pure strategy a weight that is **nonnegative**, and the weights **sum to exactly** 1. The payoff a_{ij} occurs with probability $x_i y_j$, since the two choices are independent, so $\mathbf{x}^\top \mathbf{A}\mathbf{y}$ is the average payoff over many plays. The pure strategy "row i" is the mixed strategy $\mathbf{e}_i$, a vertex of the simplex, and $\mathbf{e}_i^\top \mathbf{A}\mathbf{e}_j = a_{ij}$. So mixed strategies extend pure ones, and the expected payoff extends the payoff.

Examples. Δ_2 is the segment from $(1, 0)$ to $(0, 1)$ in the plane, whose points $(p, 1-p)$ with $0 \leq p \leq 1$ mean "row 1 with probability p". Δ_3 is the triangle with vertices $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$ in $\mathbb{R}^3$, and its center $\frac{1}{3}\mathbf{1}$ is the uniform strategy. The degenerate case is $\Delta_1 = \{1\}$: a player with only one pure strategy has no choice, and the game is a single row or a single column.

Nonexamples by minimal change. $(\frac{1}{2}, \frac{1}{2}, \frac{1}{2})$ has nonnegative entries, but they sum to $\frac{3}{2}$, so it is not a probability distribution. $(\frac{3}{2}, -\frac{1}{2})$ sums to 1, but it has a negative entry. It lies on the line through Δ_2, outside the segment.

Why this definition. Δ_n is exactly the set of convex combinations of $\mathbf{e}_1, \dots, \mathbf{e}_n$, since $\mathbf{x} = \sum x_j \mathbf{e}_j$. So it is the convex hull of the pure strategies, by Theorem 18.1.3, and in particular it is a convex set (Definition 18.1.1). That convexity is what makes the theorem below true. The convention of Definition 18.1.2, n weights for n pure strategies, is the one a game needs. The payoff $\mathbf{x}^\top \mathbf{A}\mathbf{y}$ is linear in $\mathbf{y}$ for fixed $\mathbf{x}$, and linear in $\mathbf{x}$ for fixed $\mathbf{y}$, so a mixture of strategies earns the same mixture of payoffs.

Warning. **A mixed strategy is not a compromise move.** Playing $(\frac{1}{2}, \frac{1}{2})$ in the matching game does not mean playing "half of row 1 and half of row 2" in a single round. Each round Row plays one actual row, chosen at random, and $\mathbf{x}^\top \mathbf{A}\mathbf{y}$ is an average over rounds. In the matching game with $\mathbf{x} = \mathbf{y} = (\frac{1}{2}, \frac{1}{2})$, the expected payoff is 0, yet every single round pays $+1$ or -1, never 0.

Against a fixed mixed strategy, the opponent never needs to randomize. That observation turns every "min over a simplex" into a min over finitely many numbers.

Lemma 18.7.2. A Pure Best Response Exists

Let $\mathbf{A} \in M_{m\times n}(\mathbb{R})$. For every $\mathbf{x} \in \Delta_m$,

$$\min_{\mathbf{y}\in\Delta_n} \mathbf{x}^\top \mathbf{A}\mathbf{y} = \min_{1\le j\le n} (\mathbf{A}^\top \mathbf{x})_j,$$

and the minimum on the left is attained at $\mathbf{y} = \mathbf{e}_j$ for any j attaining the minimum on the right. Likewise, for every $\mathbf{y} \in \Delta_n$, $\max_{\mathbf{x}\in\Delta_m} \mathbf{x}^\top \mathbf{A}\mathbf{y} = \max_{1\le i\le m} (\mathbf{A}\mathbf{y})_i$, attained at a vertex $\mathbf{e}_i$.

Idea. For fixed $\mathbf{x}$, the payoff $\mathbf{x}^\top \mathbf{A}\mathbf{y}$ is an average of the entries of $\mathbf{A}^\top \mathbf{x}$ with weights y_j, and an average is never below the smallest entry, which the vertex $\mathbf{e}_j$ attains.

Proof. Write $\mathbf{w} = \mathbf{A}^\top \mathbf{x} \in \mathbb{R}^n$ and $\mu = \min_j w_j$, a minimum of finitely many numbers. For $\mathbf{y} \in \Delta_n$,

$$\mathbf{x}^\top \mathbf{A}\mathbf{y} = \mathbf{w}^\top \mathbf{y} = \sum_{j=1}^{n} y_j w_j \;\ge\; \sum_{j=1}^{n} y_j \mu = \mu,$$

because every $y_j \ge 0$ and $\sum_j y_j = 1$. If $w_{j_0} = \mu$, then $\mathbf{e}_{j_0} \in \Delta_n$ gives $\mathbf{x}^\top \mathbf{A}\mathbf{e}_{j_0} = w_{j_0} = \mu$. So μ is a value of $\mathbf{y} \mapsto \mathbf{x}^\top \mathbf{A}\mathbf{y}$ on Δ_n and a lower bound for all its values, hence its minimum. The second statement is the same argument applied to $\mathbf{A}\mathbf{y}$ with the inequality reversed. ∎

So once Row fixes $\mathbf{x}$, the amount Row can **guarantee** is

$$g(\mathbf{x}) = \min_{\mathbf{y}\in\Delta_n} \mathbf{x}^\top \mathbf{A}\mathbf{y} = \min_j (\mathbf{A}^\top \mathbf{x})_j,$$

the smallest entry of $\mathbf{A}^\top \mathbf{x}$. Column, in the same way, guarantees by the second half of the lemma to pay at most $h(\mathbf{y}) = \max_{\mathbf{x}\in\Delta_m} \mathbf{x}^\top \mathbf{A}\mathbf{y} = \max_i (\mathbf{A}\mathbf{y})_i$. In the matching game, $\mathbf{x} = (\frac{1}{2}, \frac{1}{2})$ gives $\mathbf{A}^\top \mathbf{x} = (0, 0)$ and $g(\mathbf{x}) = 0$, and the same holds for Column's h. With mixing, both players

can guarantee 0, and the gap of 2 has closed. The theorem says this always happens.

18.7.3 The minimax theorem

Theorem 18.7.3. Von Neumann's Minimax Theorem

Let $\mathbf{A} \in M_{m\times n}(\mathbb{R})$. There are $\mathbf{x}^\star \in \Delta_m$, $\mathbf{y}^\star \in \Delta_n$, and $v \in \mathbb{R}$ such that

$$\mathbf{x}^\top \mathbf{A}\mathbf{y}^\star \;\le\; v \;\le\; \mathbf{x}^{\star\top}\mathbf{A}\mathbf{y} \qquad \text{for all } \mathbf{x} \in \Delta_m,\ \mathbf{y} \in \Delta_n. \tag{18.7.1}$$

Consequently

$$\max_{\mathbf{x}\in\Delta_m}\ \min_{\mathbf{y}\in\Delta_n} \mathbf{x}^\top\mathbf{A}\mathbf{y} \;=\; v \;=\; \min_{\mathbf{y}\in\Delta_n}\ \max_{\mathbf{x}\in\Delta_m} \mathbf{x}^\top\mathbf{A}\mathbf{y},$$

where the outer maximum is attained at $\mathbf{x}^\star$ and the outer minimum at $\mathbf{y}^\star$.

Idea. Row wants the $\mathbf{x} \in \Delta_m$ that maximizes the smallest entry of $\mathbf{A}^\top\mathbf{x}$. If every entry of $\mathbf{A}$ is positive, this can be rescaled into a linear program. Put $\mathbf{p} = \mathbf{x}/v$, where v is the guaranteed amount. Then "$\mathbf{A}^\top\mathbf{x} \ge v\mathbf{1}$ with $\mathbf{1}^\top\mathbf{x} = 1$, v as large as possible" becomes "$\mathbf{A}^\top\mathbf{p} \ge \mathbf{1}$, $\mathbf{p} \ge \mathbf{0}$, $\mathbf{1}^\top\mathbf{p}$ as small as possible". Column's problem becomes the dual of that program, by the same substitution. Strong duality (§06) says the two optimal values agree, and that common value is $1/v$. The plan:

① shift the entries by a constant to make them positive, which changes every payoff by the same constant; ② write down the two linear programs, check that both are feasible, and apply Theorem 18.6.3; ③ rescale the optimal solutions into strategies and read off Eq. 18.7.1.

Proof. **Step 1. We may assume every $a_{ij} > 0$.** Choose $c \in \mathbb{R}$ with $a_{ij} + c > 0$ for all i, j, and put $\mathbf{B} = \mathbf{A} + c\mathbf{J}$. For $\mathbf{x} \in \Delta_m$ and $\mathbf{y} \in \Delta_n$,

$$\mathbf{x}^\top\mathbf{J}\mathbf{y} = (\mathbf{x}^\top\mathbf{1})(\mathbf{1}^\top\mathbf{y}) = 1, \qquad \text{so} \qquad \mathbf{x}^\top\mathbf{B}\mathbf{y} = \mathbf{x}^\top\mathbf{A}\mathbf{y} + c.$$

Hence if $\mathbf{x}^\star, \mathbf{y}^\star, w$ satisfy Eq. 18.7.1 for $\mathbf{B}$, then $\mathbf{x}^\star, \mathbf{y}^\star, w - c$ satisfy it for $\mathbf{A}$. So it suffices to prove Eq. 18.7.1 when every entry of $\mathbf{A}$ is positive, and we assume this from now on. Let $\mu = \min_{i,j} a_{ij} > 0$.

Step 2. Two linear programs. Consider

$$\begin{aligned} &\text{(P)} \quad \text{maximize } \mathbf{1}^\top\mathbf{q} \ \text{ subject to } \ \mathbf{A}\mathbf{q} \le \mathbf{1},\ \mathbf{q} \ge \mathbf{0} \quad (\mathbf{q} \in \mathbb{R}^n),\\ &\text{(D)} \quad \text{minimize } \mathbf{1}^\top\mathbf{p} \ \text{ subject to } \ \mathbf{A}^\top\mathbf{p} \ge \mathbf{1},\ \mathbf{p} \ge \mathbf{0} \quad (\mathbf{p} \in \mathbb{R}^m). \end{aligned}$$

(P) is a linear program in the inequality form of Definition 18.6.1, with data $(\mathbf{A}, \mathbf{1}, \mathbf{1})$, and (D) is its dual (Definition 18.6.2). (P) is feasible, with $\mathbf{q} = \mathbf{0}$. (D) is feasible, with $\mathbf{p} = \mu^{-1}\mathbf{1}$: the j-th entry of $\mathbf{A}^\top\mathbf{p}$ is $\mu^{-1}\sum_i a_{ij} \ge \mu^{-1}a_{1j} \ge 1$. By Theorem 18.6.3 (a), there are optimal $\mathbf{q}^\star$ and $\mathbf{p}^\star$ with equal values; call the common value ω, so that

$$\mathbf{1}^\top\mathbf{q}^\star = \mathbf{1}^\top\mathbf{p}^\star = \omega.$$

Moreover $\omega > 0$. Since $\mathbf{A}^\top\mathbf{p}^\star \ge \mathbf{1}$, we have $\mathbf{p}^\star \ne \mathbf{0}$, and a nonzero vector with nonnegative entries has a positive entry sum.

Step 3. Strategies from the programs. Put

$$\mathbf{x}^\star = \omega^{-1}\mathbf{p}^\star, \qquad \mathbf{y}^\star = \omega^{-1}\mathbf{q}^\star, \qquad v = \omega^{-1}.$$

Then $\mathbf{x}^\star \geq \mathbf{0}$ with $\mathbf{1}^\top\mathbf{x}^\star = \omega^{-1}\omega = 1$, so $\mathbf{x}^\star \in \Delta_m$, and likewise $\mathbf{y}^\star \in \Delta_n$. The constraints of the two programs give

$$\mathbf{A}^\top\mathbf{x}^\star = \omega^{-1}\mathbf{A}^\top\mathbf{p}^\star \geq v\mathbf{1}, \qquad \mathbf{A}\mathbf{y}^\star = \omega^{-1}\mathbf{A}\mathbf{q}^\star \leq v\mathbf{1}.$$

For every $\mathbf{y} \in \Delta_n$, therefore, $\mathbf{x}^{\star\top}\mathbf{A}\mathbf{y} = \sum_j y_j(\mathbf{A}^\top\mathbf{x}^\star)_j \geq v\sum_j y_j = v$. For every $\mathbf{x} \in \Delta_m$, $\mathbf{x}^\top\mathbf{A}\mathbf{y}^\star = \sum_i x_i(\mathbf{A}\mathbf{y}^\star)_i \leq v$. This is Eq. 18.7.1.

Step 4. The two extrema. For $\mathbf{x} \in \Delta_m$ let $g(\mathbf{x}) = \min_{\mathbf{y}\in\Delta_n} \mathbf{x}^\top\mathbf{A}\mathbf{y}$, which exists by Lemma 18.7.2. By Eq. 18.7.1, $g(\mathbf{x}^\star) \geq v$, since every value of $\mathbf{y} \mapsto \mathbf{x}^{\star\top}\mathbf{A}\mathbf{y}$ is at least v. And $g(\mathbf{x}) \leq \mathbf{x}^\top\mathbf{A}\mathbf{y}^\star \leq v$ for every $\mathbf{x}$. So v is a value of g and an upper bound for all its values: $\max_{\mathbf{x}} g(\mathbf{x})$ exists, equals v, and is attained at $\mathbf{x}^\star$. In the same way $h(\mathbf{y}) = \max_{\mathbf{x}} \mathbf{x}^\top\mathbf{A}\mathbf{y}$ satisfies $h(\mathbf{y}^\star) \leq v \leq \mathbf{x}^{\star\top}\mathbf{A}\mathbf{y} \leq h(\mathbf{y})$ for every $\mathbf{y}$, so $\min_{\mathbf{y}} h(\mathbf{y}) = v$, attained at $\mathbf{y}^\star$. This proves the theorem.

■

Definition 18.7.2. Value and Optimal Strategies

The number v of Theorem 18.7.3 is the **value** of the game with payoff matrix $\mathbf{A}$, written $v(\mathbf{A})$. A strategy $\mathbf{x} \in \Delta_m$ is **optimal for Row** if it guarantees the value, $\min_{\mathbf{y}} \mathbf{x}^\top\mathbf{A}\mathbf{y} = v(\mathbf{A})$, and $\mathbf{y} \in \Delta_n$ is **optimal for Column** if $\max_{\mathbf{x}} \mathbf{x}^\top\mathbf{A}\mathbf{y} = v(\mathbf{A})$.

The value is well defined: it equals the max–min, and that depends only on $\mathbf{A}$. A strategy $\mathbf{x}$ is optimal for Row exactly when every entry of $\mathbf{A}^\top\mathbf{x}$ is at least $v(\mathbf{A})$, and $\mathbf{y}$ is optimal for Column exactly when every entry of $\mathbf{A}\mathbf{y}$ is at most $v(\mathbf{A})$; both statements are Lemma 18.7.2. These are linear conditions, so each player's set of optimal strategies is a convex subset of the simplex. Inequality Eq. 18.7.1 says that $(\mathbf{x}^\star, \mathbf{y}^\star)$ is a **saddle point in mixed strategies**. If Row plays $\mathbf{x}^\star$, Column cannot push the expected payment below v. If Column plays $\mathbf{y}^\star$, Row cannot push it above v. Neither player gains by deviating alone, and announcing an optimal strategy in advance costs nothing.

Alternatively, the theorem can be proved with no linear programming, by separation. The key fact is an alternative in the style of Farkas: either some $\mathbf{y} \in \Delta_n$ has $\mathbf{A}\mathbf{y} \leq \mathbf{0}$, or some $\mathbf{x} \in \Delta_m$ has $\mathbf{A}^\top\mathbf{x} > \mathbf{0}$. To prove it, separate the origin from the closed convex set $\{\mathbf{A}\mathbf{y} + \mathbf{z} : \mathbf{y} \in \Delta_n,\ \mathbf{z} \geq \mathbf{0}\}$ using §03, and apply the result to $\mathbf{A} - t\mathbf{J}$ for every real t. That route shows most directly that the theorem is about convexity. It also needs the compactness of Δ_n, (A3) of Chapter 16's introduction, to know that the set being separated is closed. The route taken above hides all of the analysis inside Farkas's lemma.

18.7.4 Two min–max theorems

Theorem 18.7.3 and the Courant–Fischer theorem of Chapter 17 (Theorem 17.2.3) both have "min" and "max" in them, and both are called min–max theorems. They are different theorems, with different proofs. Put side by side, for a Hermitian $\mathbf{A} \in M_n(F)$ ($F = \mathbb{R}$ or $\mathbb{C}$) and a real

$\mathbf{B} \in M_{m\times n}(\mathbb{R})$:

$$\begin{aligned}\lambda_k(\mathbf{A}) &= \min_{\dim W = n-k+1} \max_{\mathbf{0}\neq\mathbf{x}\in W} R_{\mathbf{A}}(\mathbf{x}) \\ &= \max_{\dim W = k} \min_{\mathbf{0}\neq\mathbf{x}\in W} R_{\mathbf{A}}(\mathbf{x}),\end{aligned}$$

$$\begin{aligned}v(\mathbf{B}) &= \min_{\mathbf{y}\in\Delta_n} \max_{\mathbf{x}\in\Delta_m} \mathbf{x}^\top\mathbf{B}\mathbf{y} \\ &= \max_{\mathbf{x}\in\Delta_m} \min_{\mathbf{y}\in\Delta_n} \mathbf{x}^\top\mathbf{B}\mathbf{y}.\end{aligned}$$

What is being chosen. In Courant–Fischer the outer variable is a **subspace** W, and the inner variable is a vector **inside that subspace**. The inner set depends on the outer choice, so the formula is a nested optimization, not a game between two independent players. In von Neumann's theorem the two variables range over **fixed** sets, Δ_m and Δ_n, neither of which depends on the other.

What the theorem asserts. Von Neumann's theorem is an **exchange of order**: the same function, with the max and the min taken in both orders, gives the same number. Courant–Fischer exchanges nothing. Its two forms use different dimensions, $n-k+1$ and k, and swapping the order while keeping the dimension gives $\lambda_{n+1-k}(\mathbf{A})$, which differs from $\lambda_k(\mathbf{A})$ unless $n = 2k-1$ or the two eigenvalues happen to be equal (the warning after Theorem 17.2.3). The sharpest contrast is in the inequality between the two forms. In von Neumann's theorem, "max–min $\leq$ min–max" is Lemma 18.7.1, and it is free for every function. In Courant–Fischer the corresponding inequality

$$\max_{\dim W = k} \min_{\mathbf{0}\neq\mathbf{x}\in W} R_{\mathbf{A}}(\mathbf{x}) \;\leq\; \min_{\dim W' = n-k+1} \max_{\mathbf{0}\neq\mathbf{x}\in W'} R_{\mathbf{A}}(\mathbf{x})$$

is **not** free: it is exactly the dimension count. Since $k + (n-k+1) > n$, any W of dimension k meets any W' of dimension $n-k+1$ in a nonzero vector $\mathbf{x}$ (Lemma 17.2.1), and then $\min_W R_{\mathbf{A}} \leq R_{\mathbf{A}}(\mathbf{x}) \leq \max_{W'} R_{\mathbf{A}}$.

Where the equality comes from. In Courant–Fischer it is a **dimension count**. Any subspace of dimension $n-k+1$ must meet the span of the top k eigenvectors, by Lemma 17.2.1, and on the shared vector the Rayleigh quotient is pinned. Convexity plays no role: the subspaces of dimension $n-k+1$ do not form a convex set in any sense used here, and $R_{\mathbf{A}}$ is not linear. In von Neumann's theorem it is **convexity**. The strategy sets are simplices, the payoff is linear in each variable separately, and the equality is linear programming duality, which is Farkas's lemma, which is a separating hyperplane. No dimension is counted anywhere.

What they need. Courant–Fischer needs a Hermitian matrix, since the spectral theorem supplies the eigenvectors that are its witnesses. Von Neumann's theorem holds for **every** real matrix, square or not.

How each fails when its key ingredient is removed. Replace the simplices by the finite sets of pure strategies, which are not convex, and von Neumann's equality fails, as in the matching game above. Replace the dimension $n-k+1$ by $n-k$, which is possible when $k < n$, so that the dimension count no longer forces an intersection, and Courant–Fischer's formula returns $\lambda_{k+1}(\mathbf{A})$ instead of $\lambda_k(\mathbf{A})$.

18.7.5 Solving a game

Example. A game with an unused column

Find the value and optimal strategies for the game with payoff matrix

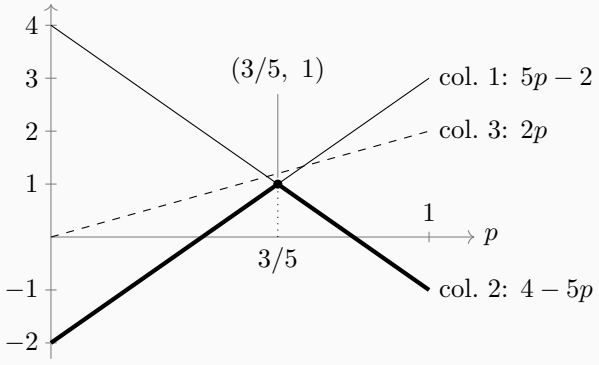
$$\mathbf{A} = \begin{pmatrix} 3 & -1 & 2 \\ -2 & 4 & 0 \end{pmatrix}.$$

Solution. No saddle point. The row minima are -1 and -2, so $\max_i \min_j a_{ij} = -1$. The column maxima are $3, 4, 2$, so $\min_j \max_i a_{ij} = 2$. The two differ, so both players must mix.

Row's side. Write $\mathbf{x} = (p, 1-p)$ with $0 \le p \le 1$. Then

$$\mathbf{A}^\top \mathbf{x} = (5p - 2,\ 4 - 5p,\ 2p),$$

and by Lemma 18.7.2 Row guarantees $g(p) = \min\{5p - 2,\ 4 - 5p,\ 2p\}$. The graph of g is the lower envelope of three lines.



Row's guarantee $g(p)$ is the lower envelope (thick). It is highest at $p = 3/5$, where columns 1 and 2 cross at height 1. The dashed line never touches it

For $p \le \frac{3}{5}$ the first line is the lowest: $5p - 2 \le 4 - 5p$ exactly when $p \le \frac{3}{5}$, and $5p - 2 \le 2p$ exactly when $p \le \frac{2}{3}$. For $p \ge \frac{3}{5}$ the second line is the lowest: $4 - 5p \le 2p$ exactly when $p \ge \frac{4}{7}$, which holds since $\frac{3}{5} > \frac{4}{7}$. So $g(p) = 5p - 2$ on $[0, \frac{3}{5}]$ and $g(p) = 4 - 5p$ on $[\frac{3}{5}, 1]$. With slopes 5 and -5, g **strictly** increases on the first interval and **strictly** decreases on the second, so its maximum $g(\frac{3}{5}) = 1$ is attained at $p = \frac{3}{5}$ and nowhere else. The only optimal strategy for Row is $\mathbf{x}^\star = (\frac{3}{5}, \frac{2}{5})$, and $\mathbf{A}^\top \mathbf{x}^\star = (1, 1, \frac{6}{5})$.

Column's side. Against $\mathbf{x}^\star$, column 3 costs Column $\frac{6}{5}$, which is more than the value, so we expect Column never to use it. Try $\mathbf{y} = (q, 1-q, 0)$: then $\mathbf{A}\mathbf{y} = (4q - 1,\ 4 - 6q)$, and the two entries are equal when $q = \frac{1}{2}$. So $\mathbf{y}^\star = (\frac{1}{2}, \frac{1}{2}, 0)$, with $\mathbf{A}\mathbf{y}^\star = (1, 1)$.

Verification. $\mathbf{A}^\top \mathbf{x}^\star = (1, 1, \frac{6}{5}) \ge 1 \cdot \mathbf{1}$ and $\mathbf{A}\mathbf{y}^\star = (1, 1) \le 1 \cdot \mathbf{1}$. As in Step 3 of the proof, these two inequalities give Eq. 18.7.1 with $v = 1$. So the value is $v(\mathbf{A}) = 1$, Row's optimal strategy is $(\frac{3}{5}, \frac{2}{5})$, and $(\frac{1}{2}, \frac{1}{2}, 0)$ is optimal for Column.

Column's optimal strategy is unique too. If $\mathbf{y}$ is optimal for Column, then $\mathbf{A}\mathbf{y} \le \mathbf{1}$, so $\mathbf{x}^{\star\top} \mathbf{A}\mathbf{y} \le 1$. But $\mathbf{x}^{\star\top} \mathbf{A}\mathbf{y} = y_1 + y_2 + \frac{6}{5} y_3 = 1 + \frac{1}{5} y_3$, so $y_3 = 0$. With $\mathbf{y} = (q, 1-q, 0)$, the conditions $4q - 1 \le 1$ and $4 - 6q \le 1$ give $q \le \frac{1}{2}$ and $q \ge \frac{1}{2}$. This is complementary slackness (Theorem 18.6.5) in the language of games: a column whose payoff against an optimal $\mathbf{x}$ exceeds the value is never used.

Quick check. In the game of Example (A game with an unused column), is $\mathbf{x} = (\frac{1}{2}, \frac{1}{2})$ an optimal strategy for Row? What does Column do against it?

Solution. No. $\mathbf{A}^\top \mathbf{x} = (\frac{1}{2}, \frac{3}{2}, 1)$, so $g(\mathbf{x}) = \frac{1}{2} < 1 = v(\mathbf{A})$. Column answers with column 1 and pays only $\frac{1}{2}$ on average. The symmetric-looking strategy is not the safe one.

Warning. **"Make the opponent indifferent between all columns" is not a method.** In Example (A game with an unused column), no $\mathbf{x} \in \Delta_2$ makes all three entries of $\mathbf{A}^\top \mathbf{x}$ equal: the first two are equal only at $p = \frac{3}{5}$, where the third is $\frac{6}{5}$, not 1. The optimal $\mathbf{x}$ equalizes only the columns that Column actually uses. Deciding **which** columns those are is the real content of solving the game, and it is what the linear program decides. Equalizing the wrong set of columns produces a strategy that is not optimal, or a "solution" with negative probabilities.

Finally, optimal strategies need not be unique, although in the example both were. Exercise (B3: Many optimal strategies) gives a game in which Row has a whole segment of optimal strategies, which is what the convexity of the optimal set allows.

18.7.6 Exercises

18.7.6.1 A. Check your understanding

Exercise. A1

(a) Define the standard simplex Δ_n, a mixed strategy, and the value of a matrix game.

(b) State von Neumann's minimax theorem.

(c) Determine whether the following is correct, and justify your answer: every matrix game has a saddle point in pure strategies.

(d) Explain in two sentences why $\max_x \min_y f(x,y) \le \min_y \max_x f(x,y)$ for every function f for which the extrema exist.

(e) Name the mechanism that produces the equality in Theorem 17.2.3 and the one that produces it in Theorem 18.7.3.

Solution.

(a) $\Delta_n = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{x} \ge \mathbf{0},\ \sum_i x_i = 1\}$. In a game with payoff matrix $\mathbf{A} \in M_{m\times n}(\mathbb{R})$, a mixed strategy for Row is an $\mathbf{x} \in \Delta_m$ and one for Column is a $\mathbf{y} \in \Delta_n$. The value is $v(\mathbf{A}) = \max_{\mathbf{x}\in\Delta_m} \min_{\mathbf{y}\in\Delta_n} \mathbf{x}^\top \mathbf{A}\mathbf{y}$.

(b) For every $\mathbf{A} \in M_{m\times n}(\mathbb{R})$, $\max_{\mathbf{x}\in\Delta_m} \min_{\mathbf{y}\in\Delta_n} \mathbf{x}^\top \mathbf{A}\mathbf{y} = \min_{\mathbf{y}\in\Delta_n} \max_{\mathbf{x}\in\Delta_m} \mathbf{x}^\top \mathbf{A}\mathbf{y}$, both extrema being attained. Equivalently, there are $\mathbf{x}^\star, \mathbf{y}^\star$ and v with $\mathbf{x}^\top \mathbf{A}\mathbf{y}^\star \le v \le \mathbf{x}^{\star\top} \mathbf{A}\mathbf{y}$ for all $\mathbf{x}, \mathbf{y}$.

(c) Incorrect. In the matching game with $\mathbf{A} = \left(\begin{smallmatrix} 1 & -1 \\ -1 & 1 \end{smallmatrix}\right)$, $\max_i \min_j a_{ij} = -1 < 1 = \min_j \max_i a_{ij}$. An entry that is smallest in its row and largest in its column would make these equal.

(d) For any x_0, y_0, $\min_y f(x_0, y) \le f(x_0, y_0) \le \max_x f(x, y_0)$. The left side does not involve y_0 and the right side does not involve x_0, so we may maximize the left over x_0 and minimize the right over y_0.

(e) For Courant–Fischer, a dimension count: Lemma 17.2.1 forces every competing subspace to meet a span of eigenvectors. For von Neumann, convexity: the strategy sets are simplices and the payoff is linear in each variable, and the equality is linear programming duality, which rests on separation.

18.7.6.2 B. Practice

Exercise. B1: A two-by-two game

Let $\mathbf{A} = \begin{pmatrix} 2 & -3 \\ -1 & 1 \end{pmatrix}$.

(a) Show that the game has no saddle point in pure strategies.

(b) Find the value and optimal strategies for both players, and verify the saddle inequality Eq. 18.7.1.

Solution.

(a) The row minima are -3 and -1, so $\max_i \min_j a_{ij} = -1$. The column maxima are 2 and 1, so $\min_j \max_i a_{ij} = 1$. These differ, so no entry is both smallest in its row and largest in its column.

(b) For $\mathbf{x} = (p, 1-p)$, $\mathbf{A}^\top \mathbf{x} = (3p - 1,\ 1 - 4p)$. The first entry increases and the second decreases in p, so their minimum is largest where they are equal: $3p - 1 = 1 - 4p$, that is, $p = \frac{2}{7}$, with common value $-\frac{1}{7}$. For $\mathbf{y} = (q, 1-q)$, $\mathbf{A}\mathbf{y} = (5q - 3,\ 1 - 2q)$, and the entries are equal when $q = \frac{4}{7}$, with common value $-\frac{1}{7}$. So take $\mathbf{x}^\star = (\frac{2}{7}, \frac{5}{7})$, $\mathbf{y}^\star = (\frac{4}{7}, \frac{3}{7})$, and $v = -\frac{1}{7}$. Then $\mathbf{A}^\top \mathbf{x}^\star = -\frac{1}{7}\mathbf{1}$ and $\mathbf{A}\mathbf{y}^\star = -\frac{1}{7}\mathbf{1}$. For every $\mathbf{y} \in \Delta_2$, $\mathbf{x}^{\star\top}\mathbf{A}\mathbf{y} = -\frac{1}{7}(y_1 + y_2) = -\frac{1}{7}$, and, since $\mathbf{A}\mathbf{y}^\star = -\frac{1}{7}\mathbf{1}$, for every $\mathbf{x} \in \Delta_2$ we have $\mathbf{x}^\top \mathbf{A}\mathbf{y}^\star = -\frac{1}{7}(x_1 + x_2) = -\frac{1}{7}$. So Eq. 18.7.1 holds with equality throughout, and $v(\mathbf{A}) = -\frac{1}{7}$. The game slightly favors Column.

Exercise. B2: A saddle point

Let $\mathbf{A} = \begin{pmatrix} 3 & 1 & 4 \\ 2 & 0 & 1 \end{pmatrix}$, the second game of this section. Show that $\mathbf{x}^\star = \mathbf{e}_1$, $\mathbf{y}^\star = \mathbf{e}_2$ satisfy Eq. 18.7.1 with $v = 1$, and deduce the value of the game.

Solution. For $\mathbf{x} \in \Delta_2$, $\mathbf{x}^\top \mathbf{A}\mathbf{e}_2 = x_1 \cdot 1 + x_2 \cdot 0 = x_1 \le 1$, since $x_1 \le x_1 + x_2 = 1$. For $\mathbf{y} \in \Delta_3$, $\mathbf{e}_1^\top \mathbf{A}\mathbf{y} = 3y_1 + y_2 + 4y_3 \ge y_1 + y_2 + y_3 = 1$. So $\mathbf{x}^\top \mathbf{A}\mathbf{e}_2 \le 1 \le \mathbf{e}_1^\top \mathbf{A}\mathbf{y}$ for all $\mathbf{x}, \mathbf{y}$, which is Eq. 18.7.1 with $v = 1$. By Step 4 of the proof of Theorem 18.7.3 (whose argument uses only Eq. 18.7.1), the max–min and the min–max both equal 1, so $v(\mathbf{A}) = 1$. A pure saddle point is a special case of a mixed one.

Exercise. B3: Many optimal strategies

Let $\mathbf{A} = \begin{pmatrix} 4 & 0 & 1 \\ -1 & 3 & 1 \end{pmatrix}$.

(a) Show that $v(\mathbf{A}) = 1$.

(b) Find **all** optimal strategies for Row.

(c) Show that Column has exactly one optimal strategy.

Solution.

(a) Column can play $\mathbf{e}_3$, which gives $\mathbf{x}^\top \mathbf{A}\mathbf{e}_3 = x_1 + x_2 = 1$ for every $\mathbf{x}$. Hence

$$\min_{\mathbf{y}} \max_{\mathbf{x}} \mathbf{x}^\top \mathbf{A}\mathbf{y} \le \max_{\mathbf{x}} \mathbf{x}^\top \mathbf{A}\mathbf{e}_3 = 1.$$

Row can play $\mathbf{x} = (\frac{1}{2}, \frac{1}{2})$, with $\mathbf{A}^\top \mathbf{x} = (\frac{3}{2}, \frac{3}{2}, 1) \ge \mathbf{1}$, so $\max_{\mathbf{x}} \min_{\mathbf{y}} \mathbf{x}^\top \mathbf{A}\mathbf{y} \ge 1$ by Lemma 18.7.2. By Theorem 18.7.3 both equal $v(\mathbf{A})$, which is therefore 1.

(b) By the remark after Definition 18.7.2, $\mathbf{x} = (p, 1-p)$ is optimal exactly when $\mathbf{A}^\top \mathbf{x} = (5p - 1,\ 3 - 3p,\ 1) \ge \mathbf{1}$, that is, $p \ge \frac{2}{5}$ and $p \le \frac{2}{3}$. So the optimal strategies for Row are the whole segment $\{(p, 1-p) : \frac{2}{5} \le p \le \frac{2}{3}\}$.

(c) $\mathbf{y} = (a, b, c) \in \Delta_3$ is optimal exactly when $\mathbf{A}\mathbf{y} \le \mathbf{1}$: $4a + c \le 1$ and $-a + 3b + c \le 1$. Substituting $c = 1 - a - b$, these become $3a \le b$ and $2b \le 2a$. Hence $3a \le b \le a$, so $a \le 0$, and with $a \ge 0$ we get $a = 0$ and then $b = 0$. So $\mathbf{y} = \mathbf{e}_3$ is the unique optimal strategy for Column.

18.7.6.3 C. Going deeper

Exercise. C1: Symmetric games

Let $\mathbf{A} \in M_n(\mathbb{R})$ be skew-symmetric, $\mathbf{A}^\top = -\mathbf{A}$. Such a game is **symmetric**: the two players have the same options, and swapping their roles swaps the payoff's sign.

(a) Prove that $\mathbf{x}^\top \mathbf{A}\mathbf{x} = 0$ for every $\mathbf{x} \in \mathbb{R}^n$, and deduce that $v(\mathbf{A}) = 0$.

(b) Prove that $\mathbf{x}$ is optimal for Row if and only if it is optimal for Column.

(c) In rock–paper–scissors, with the pure strategies in that order and payoff 1 for a win, -1 for a loss, and 0 for a tie,

$$\mathbf{A} = \begin{pmatrix} 0 & -1 & 1 \\ 1 & 0 & -1 \\ -1 & 1 & 0 \end{pmatrix}.$$

Prove that the uniform strategy $\frac{1}{3}\mathbf{1}$ is the **only** optimal strategy.

Solution.

(a) $\mathbf{x}^\top \mathbf{A}\mathbf{x}$ is a 1×1 matrix, so it equals its transpose: $\mathbf{x}^\top \mathbf{A}\mathbf{x} = \mathbf{x}^\top \mathbf{A}^\top \mathbf{x} = -\mathbf{x}^\top \mathbf{A}\mathbf{x}$. Hence $\mathbf{x}^\top \mathbf{A}\mathbf{x} = 0$. Let $\mathbf{x}^\star, \mathbf{y}^\star, v$ be as in Theorem 18.7.3. Taking $\mathbf{y} = \mathbf{x}^\star$ in the right half of Eq. 18.7.1 gives $v \le \mathbf{x}^{\star\top} \mathbf{A}\mathbf{x}^\star = 0$. Taking $\mathbf{x} = \mathbf{y}^\star$ in the left half gives $0 = \mathbf{y}^{\star\top} \mathbf{A}\mathbf{y}^\star \le v$.

So $v(\mathbf{A}) = 0$.

(b) By the remark after Definition 18.7.2 and (a), $\mathbf{x}$ is optimal for Row exactly when $\mathbf{A}^{\top}\mathbf{x} \geq \mathbf{0}$, and optimal for Column exactly when $\mathbf{A}\mathbf{x} \leq \mathbf{0}$. Since $\mathbf{A}^{\top}\mathbf{x} = -\mathbf{A}\mathbf{x}$, the two conditions are the same.

(c) With $\mathbf{x} = (x_1, x_2, x_3) \in \Delta_3$, the condition $\mathbf{A}^{\top}\mathbf{x} \geq \mathbf{0}$ reads

$$x_2 - x_3 \geq 0, \qquad x_3 - x_1 \geq 0, \qquad x_1 - x_2 \geq 0.$$

So $x_2 \geq x_3 \geq x_1 \geq x_2$, all three are equal, and since they sum to 1, each is $\frac{1}{3}$. By (a) and (b), $\frac{1}{3}\mathbf{1}$ is the unique optimal strategy for either player, and the value is 0.

Exercise. C2: The value depends continuously on the payoffs

For $\mathbf{A}, \mathbf{B} \in M_{m\times n}(\mathbb{R})$, put $\varepsilon = \max_{i,j}|a_{ij} - b_{ij}|$.

(a) Prove that $|v(\mathbf{A}) - v(\mathbf{B})| \leq \varepsilon$.

(b) Prove also that if $a_{ij} \geq b_{ij}$ for all i, j, then $v(\mathbf{A}) \geq v(\mathbf{B})$, and that $v(\mathbf{A}+c\mathbf{J}) = v(\mathbf{A})+c$ for every real c.

Solution.

(a) For $\mathbf{x} \in \Delta_m$ and $\mathbf{y} \in \Delta_n$,

$$|\mathbf{x}^{\top}\mathbf{A}\mathbf{y} - \mathbf{x}^{\top}\mathbf{B}\mathbf{y}| \leq \sum_{i,j} x_i y_j |a_{ij} - b_{ij}| \leq \varepsilon \sum_{i,j} x_i y_j = \varepsilon,$$

since $\sum_{i,j} x_i y_j = (\sum_i x_i)(\sum_j y_j) = 1$. So $\mathbf{x}^{\top}\mathbf{A}\mathbf{y} \leq \mathbf{x}^{\top}\mathbf{B}\mathbf{y} + \varepsilon$ for all $\mathbf{x}, \mathbf{y}$. Taking the minimum over $\mathbf{y}$ preserves this inequality between two functions of $\mathbf{y}$, and so does taking the maximum over $\mathbf{x}$. By Theorem 18.7.3 the resulting max–mins are the values, so $v(\mathbf{A}) \leq v(\mathbf{B}) + \varepsilon$. Exchanging $\mathbf{A}$ and $\mathbf{B}$ gives $v(\mathbf{B}) \leq v(\mathbf{A}) + \varepsilon$.

(b) If $a_{ij} \geq b_{ij}$, then $\mathbf{x}^{\top}\mathbf{A}\mathbf{y} \geq \mathbf{x}^{\top}\mathbf{B}\mathbf{y}$ for all strategies, since every $x_i y_j \geq 0$, and the same two steps give $v(\mathbf{A}) \geq v(\mathbf{B})$. For the shift, $\mathbf{x}^{\top}(\mathbf{A} + c\mathbf{J})\mathbf{y} = \mathbf{x}^{\top}\mathbf{A}\mathbf{y} + c$ for all strategies, as in Step 1 of the proof of Theorem 18.7.3, and adding a constant to a function adds it to every minimum and maximum. So $v(\mathbf{A} + c\mathbf{J}) = v(\mathbf{A}) + c$.

18.8 Extreme Points

Everything so far has gone from points to convex sets: a set of points has a convex hull, and Carathéodory's theorem says how few points each element of the hull needs. This section asks the reverse question: given a convex set, which of its points cannot be manufactured from the others? A triangle is the hull of its three corners, and no smaller set will do. The corners are the **extreme points**, and the main theorem, due to Minkowski, says that a compact convex set is always the hull of its extreme points. We then find the extreme points of three unit balls and of the density matrices, show that the permutation matrices are extreme among the doubly stochastic matrices, and end with a compact convex set whose extreme points do not form a closed set.

Throughout, the field is $\mathbb{R}$, and V is a finite-dimensional real inner product space: $\mathbb{R}^n$ with

the dot product, or $M_n(\mathbb{R})$ with the Frobenius inner product $\langle \mathbf{A}, \mathbf{B} \rangle = \operatorname{tr}(\mathbf{B}^\top \mathbf{A})$ of Chapter 11 §01. **Compact** is meant as in §03, and closed bounded sets are compact by Corollary 16.2.5. Conversely a compact set is closed, as a subsequence has the limit of the whole sequence, and bounded, as a sequence with $\|x_k\| > k$ has no convergent subsequence, convergent sequences being bounded; both facts belong to the elementary algebra of limits that Chapter 16's introduction imports.

18.8.1 Points that are not in between

A convex set contains the segment between any two of its points (Definition 18.1.1). In a filled triangle, a point in the middle of an edge sits between the two ends of that edge, and an interior point sits between many pairs. A corner sits between no pair: any segment of the triangle through a corner has the corner as an end. So the corners are the points you cannot do without: delete one from the list of points whose hull you take, and the hull shrinks.

An extreme point of a convex set is a point that is not strictly between two other points of the set.

Definition 18.8.1. Extreme Point

Let $C \subseteq V$ be a convex set. A point $x \in C$ is an **extreme point** of C if, **whenever** $x = ty + (1-t)z$ with $y, z \in C$ and $0 < t < 1$ **strictly**, then $y = z = x$. The set of extreme points of C is written $\operatorname{ext} C$.

In words: the only way to put x strictly inside a segment of C is the useless way, with both ends equal to x. The word **strictly** matters. With $t = 1$ allowed, every point would be $1 \cdot x + 0 \cdot z$ for any $z \in C$, and no point of a set with two elements or more would be extreme.

To test the definition, midpoints are enough.

Lemma 18.8.1. The Midpoint Test

Let $C \subseteq V$ be convex and $x \in C$. Then x is an extreme point of C if and only if, whenever $x = \frac{1}{2}(y + z)$ with $y, z \in C$, we have $y = z$.

Idea. If x sits at parameter $t \le \frac{1}{2}$ on a segment $[z, y]$, shorten the segment to $[z, y']$ with y' at parameter $2t$; then x is its midpoint.

Proof. ($\Rightarrow$) Take $t = \frac{1}{2}$ in Definition 18.8.1.

($\Leftarrow$) Let $x = ty + (1-t)z$ with $y, z \in C$ and $0 < t < 1$. Swapping the names of y and z if necessary, we may assume $t \le \frac{1}{2}$. Put $y' = 2ty + (1-2t)z$. Since $0 < 2t \le 1$, y' lies on the segment from z to y, so $y' \in C$ by convexity. Moreover

$$\tfrac{1}{2}(y' + z) = ty + (1-t)z = x,$$

so the hypothesis gives $y' = z$, that is, $2t(y - z) = 0$. As $t \neq 0$, $y = z$, and then $x = ty + (1-t)y = y$. So $y = z = x$, as claimed.

■

Now some examples, simplest first.

1. **A single point.** $C = \{x\}$ is convex, and x is extreme: any $y, z \in C$ equal x.
2. **A closed segment** $C = \{ta + (1-t)b : 0 \le t \le 1\}$ with $a \neq b$ has exactly two extreme points, a and b. A point x with parameter $0 < s < 1$ is not extreme, since $x = sa + (1-s)b$ with $a \neq b$. The end a is extreme: if $a = \frac{1}{2}(y + z)$ with y, z at parameters $r, r' \in [0, 1]$, then comparing coefficients along the line gives $\frac{1}{2}(r + r') = 1$, which forces $r = r' = 1$ and $y = z = a$. (The coefficient comparison is legitimate because $t \mapsto ta + (1-t)b = b + t(a-b)$ is injective, as $a - b \neq 0$.)
3. **The closed half-plane** $H = \{(x_1, x_2) \in \mathbb{R}^2 : x_2 \ge 0\}$ has **no** extreme points. Every $\mathbf{x} \in H$ is the midpoint of $\mathbf{x} + \mathbf{e}_1$ and $\mathbf{x} - \mathbf{e}_1$, both in H, and they differ.
4. **The empty set** is convex and compact with no extreme points. It satisfies the main theorem below only because $\operatorname{conv} \emptyset = \emptyset$, which is why the theorem's last sentence says "nonempty".

The closed disk $B_2 \subseteq \mathbb{R}^2$ is a richer example: its extreme points are the whole unit circle. The section on unit balls below proves this for every n.

A nonexample by minimal change. Take the closed triangle $T = \operatorname{conv}\{(0,0), (2,0), (0,2)\}$ and slide from the corner $(2,0)$ along the edge to its midpoint $(1,1)$. The point $(1,1)$ is still in T, still on its boundary, and still on a line with all of T on one side ($x_1 + x_2 \le 2$ on T). What fails is the one clause of Definition 18.8.1: $(1,1) = \frac{1}{2}((2,0) + (0,2))$ with $(2,0) \neq (0,2)$.

Warning. **Extreme is not the same as "on the boundary".** The point $(1,1)$ just used lies on the boundary of the triangle T, and every open disk around it contains points outside T; yet it is not extreme. The other direction does hold, for every convex set: an extreme point is never an interior point, since an interior point $\mathbf{x}$ is the midpoint of $\mathbf{x} \pm \varepsilon \mathbf{e}_1$ for small $\varepsilon > 0$.

Quick check. Let $S = \{(0,0), (2,0), (0,2), (1,1), (1,0)\}$. Which points of S are extreme points of $\operatorname{conv} S$?

Solution. Only $(0,0)$, $(2,0)$, and $(0,2)$. The hull is the triangle $T = \{x_1 \ge 0,\ x_2 \ge 0,\ x_1 + x_2 \le 2\}$ above, because $(1,1)$ and $(1,0) = \frac{1}{2}((0,0) + (2,0))$ already lie in T. Each of those two points is the midpoint of two distinct points of T, so neither is extreme. The corners are extreme. For $(2,0)$: if $(2,0) = \frac{1}{2}(\mathbf{y} + \mathbf{z})$ with $\mathbf{y}, \mathbf{z} \in T$, then $y_1, z_1 \le 2$ and $\frac{1}{2}(y_1 + z_1) = 2$ force $y_1 = z_1 = 2$, and the only point of T with first coordinate 2 is $(2,0)$, because $x_2 \ge 0$ and $x_1 + x_2 \le 2$. The other two corners are handled the same way, with x_2 and with $-x_1 - x_2$ in place of x_1.

Why this definition. It does not ask x to be far from the rest of C, or on a corner; on the disk, extreme points have extreme neighbors arbitrarily close. It asks only that x not be an average of other points of C, which is what "cannot be manufactured from the others" means for a convex hull. Exercise C1 shows that x is extreme exactly when $C \setminus \{x\}$ is still convex.

The first result makes the indispensability precise: if a convex set is the hull of S, then S contains every extreme point.

Proposition 18.8.2. Extreme Points of a Hull Come From the Generators

Let $S \subseteq V$ be any set. Then $\operatorname{ext}(\operatorname{conv} S) \subseteq S$.

Idea. Split off one generator from a convex combination: what remains is again a point of the hull, and x lies strictly between the two.

Proof. Let x be an extreme point of $\operatorname{conv} S$. By Theorem 18.1.3, $x = \sum_{i=1}^{k} t_i s_i$ with $s_i \in S$, every $t_i > 0$ (drop the zero terms) and $\sum_i t_i = 1$. If $k = 1$ then $x = s_1 \in S$. If $k \geq 2$, then $0 < t_1 < 1$, and

$$x = t_1 s_1 + (1 - t_1) r, \qquad r = \sum_{i=2}^{k} \frac{t_i}{1 - t_1} s_i.$$

The coefficients of r are positive and add up to 1, so $r \in \operatorname{conv} S$ by Theorem 18.1.3. Since x is extreme, $s_1 = r = x$. In either case $x \in S$.

■

In particular a **polytope**, the hull of finitely many points, has finitely many extreme points, all among the points one started from. Section 9 is built on this.

18.8.2 Faces

A single extreme point is too small an object for an induction on dimension. The right object is a piece of the set that behaves like an extreme point as a whole.

Definition 18.8.2. Face

Let $C \subseteq V$ be convex. A **face** of C is a convex subset $F \subseteq C$ such that, **whenever** $ty + (1 - t)z \in F$ with $y, z \in C$ and $0 < t < 1$, both y and z lie in F.

So a face swallows every segment of C that passes through it at an inner point. The empty set and C itself are faces, and a one-point set $\{x\}$ is a face exactly when x is an extreme point: that is Definition 18.8.1 word for word. The edges of a triangle are faces. The segment joining the midpoints of two edges is convex but is not a face: if the triangle has corners a, b, c and the segment joins $\frac{1}{2}(a + b)$ to $\frac{1}{2}(a + c)$, its midpoint is $\frac{1}{2}a + \frac{1}{2} \cdot \frac{1}{2}(b + c)$, an inner point of a segment of the triangle whose ends a and $\frac{1}{2}(b + c)$ do not lie on it.

What makes faces useful for an induction is that they pass extremeness upwards.

Lemma 18.8.3. Extreme Points of a Face

Let $C \subseteq V$ be convex and F a face of C. Then $\operatorname{ext} F = F \cap \operatorname{ext} C$. In particular an extreme point of a face is an extreme point of C.

Proof. Let $x \in \operatorname{ext} F$, and let $x = ty + (1-t)z$ with $y, z \in C$ and $0 < t < 1$. Since F is a face, $y, z \in F$, and since x is extreme in F, $y = z = x$. So $x \in \operatorname{ext} C$. Conversely a point of F that is extreme in C is extreme in the smaller set F, because the definition then quantifies over fewer pairs y, z.

■

The faces we can actually produce are the ones cut off by a hyperplane that has the whole set on one side. This is the situation of Theorem 18.3.6.

Proposition 18.8.4. Supporting Slices Are Faces

Let $C \subseteq V$ be convex, let $w \in V$ and $\alpha \in \mathbb{R}$, and suppose $\langle w, x\rangle \le \alpha$ for every $x \in C$. Then:

(a) $F = \{x \in C : \langle w, x\rangle = \alpha\}$ is a face of C;

(b) if F is a single point $\{c\}$, then c is an extreme point of C.

Idea. A weighted average of numbers that are all at most α equals α only if each of them does.

Proof.

(a) F is convex, since if $\langle w, y\rangle = \langle w, z\rangle = \alpha$, the same holds for $ty + (1-t)z$ by linearity of the inner product in its second slot. Let $y, z \in C$, $0 < t < 1$, and $ty + (1-t)z \in F$. Then

$$\alpha = t\langle w, y\rangle + (1-t)\langle w, z\rangle,$$

with $\langle w, y\rangle \le \alpha$ and $\langle w, z\rangle \le \alpha$. If either inequality were strict, the right side would be strictly less than $t\alpha + (1-t)\alpha = \alpha$, because t and $1-t$ are positive. So both are equalities, and $y, z \in F$.

(b) By (a), $\{c\}$ is a face, and its only point is extreme in it. By Lemma 18.8.3, $c \in \operatorname{ext} C$.

■

A point c as in (b), the **unique** maximizer over C of some linear function $x \mapsto \langle w, x\rangle$, is called an **exposed point**. Part (b) says that exposed points are extreme; it is the most convenient test for extremeness.

18.8.3 Minkowski's theorem

The main theorem is proved by induction on dimension. Recall from §01 that a nonempty convex set C has dimension $\dim C = \dim \operatorname{aff} C$; by Proposition 18.1.4, $\operatorname{aff} C = x_0 + W_C$ with $W_C = \operatorname{span}\{y - x_0 : y \in C\}$, independent of $x_0 \in C$. So $\dim C = \dim W_C$ (Definition 18.1.5 (b)). A convex set has dimension 0 exactly when it is a single point, and for every $y \in C$ the difference set $C - y = \{x - y : x \in C\}$ lies in W_C.

Theorem 18.8.5. Minkowski's Theorem

Let V be a finite-dimensional real inner product space and let $K \subseteq V$ be a compact convex set. Then

$$K = \operatorname{conv}(\operatorname{ext} K).$$

In particular, a nonempty compact convex set has at least one extreme point.

Idea. The inclusion $\supseteq$ is free, since K is convex and contains its extreme points. For $\subseteq$, induct on $\dim K$. Take $x \in K$ and a line through x in a direction of W_K. Compactness cuts the line down to a segment $[y, z]$ of K with x on it. An end y is not an interior point, and Theorem 18.3.6, applied **inside** W_K to the translate $K - y$, gives a hyperplane through y with K on one side. The slice of K on that hyperplane is a face of smaller dimension, so by induction y is a convex combination of its extreme points, and those are extreme in K by Lemma 18.8.3. The same goes for z, and x lies between y and z. The space matters: in all of V, a hyperplane containing all of K would qualify, and would give nothing.

Proof. ($\supseteq$) K is convex and contains $\operatorname{ext} K$, so it contains $\operatorname{conv}(\operatorname{ext} K)$, the smallest convex set containing $\operatorname{ext} K$ (Definition 18.1.4).

($\subseteq$) If $K = \emptyset$ there is nothing to prove. For nonempty K we prove, by induction on $d \geq 0$, the statement: *every nonempty compact convex set of dimension d, in any finite-dimensional real inner product space, is contained in the convex hull of its extreme points.*

If $d = 0$, then $K = \{x_0\}$, and x_0 is extreme, so $K = \operatorname{conv}(\operatorname{ext} K)$.

Let $d \geq 1$, assume the statement for all dimensions less than d, and let $\dim K = d$ and $x \in K$. Write $W = W_K$, and choose $u \in W$ with $u \neq 0$, which is possible because $\dim W = d \geq 1$.

Step 1: a segment through x. Let $T = \{t \in \mathbb{R} : x + tu \in K\}$, which contains 0. It is bounded: if $\|k\| \leq R$ for every $k \in K$, then for $t \in T$, $|t|\|u\| = \|(x + tu) - x\| \leq 2R$. It is closed: if $t_j \in T$ and $t_j \to t$, then $x + t_j u \to x + tu$, which lies in K because K is closed. So T is a nonempty closed bounded subset of $\mathbb{R}$, which is compact by fact (A3) of Chapter 16's introduction. The function $t \mapsto t$ is continuous, so by fact (A4) (the extreme value theorem) it attains a minimum a and a maximum b on T, and $a \leq 0 \leq b$. Put $y = x + au$ and $z = x + bu$, both in K. If $a < b$ then

$$x = \frac{b}{b-a}\, y + \frac{-a}{b-a}\, z,$$

a convex combination, since both coefficients are ≥ 0 and they add up to 1; if $a = b$ then $a = b = 0$ and $x = y = z$. Either way $x \in \operatorname{conv}\{y, z\}$.

Step 2: a supporting hyperplane at y, inside W. With the inner product of V restricted to it, $W \neq \{0\}$ is a finite-dimensional real inner product space. The set $K - y$ lies in W, is convex by Proposition 18.1.1 (c), and contains 0. But 0 is not an interior point of $K - y$ in W (Definition 18.3.2): for every $\varepsilon > 0$, the vector $-\varepsilon u \in W$, of norm $\varepsilon\|u\|$, is **not** in $K - y$, because $y - \varepsilon u = x + (a - \varepsilon)u \notin K$ by the choice of a. By Theorem 18.3.6, applied in W to $K - y$ and 0, there is a nonzero $w \in W$ with

$$\langle w, k - y \rangle \leq \langle w, 0 \rangle = 0 \qquad \text{for every } k \in K.$$

Step 3: the face through y has smaller dimension. So $\langle w, k\rangle \le \alpha := \langle w, y\rangle$ on K, and by Proposition 18.8.4 (a), $F = \{k \in K : \langle w, k\rangle = \alpha\}$ is a face of K. It is nonempty, as $y \in F$, and convex. It is bounded, being inside K, and closed: if $k_j \in F$ and $k_j \to k$, then $k \in K$, and $|\langle w, k_j\rangle - \langle w, k\rangle| \le \|w\|\|k_j - k\| \to 0$ by Theorem 11.1.5, so $\langle w, k\rangle = \alpha$. So F is compact. Using $y \in F$ as base point, W_F is spanned by vectors $k - y$ with $k \in F$, each of which lies in

$$W' = \{v \in W : \langle w, v\rangle = 0\},$$

the kernel of the linear functional $v \mapsto \langle w, v\rangle$ on W. That functional is nonzero, since it sends w to $\|w\|^2 > 0$, so its image is $\mathbb{R}$ and Theorem 2.3.1 gives $\dim W' = d - 1$. Hence $\dim F = \dim W_F \le d - 1$ by Theorem 1.6.7.

Step 4: induction. By the induction hypothesis, $y \in \operatorname{conv}(\operatorname{ext} F)$, and $\operatorname{ext} F \subseteq \operatorname{ext} K$ by Lemma 18.8.3. So $y \in \operatorname{conv}(\operatorname{ext} K)$. Running Steps 2 to 4 with $-u$ in place of u, which exchanges the roles of a and b, gives $z \in \operatorname{conv}(\operatorname{ext} K)$. Since $\operatorname{conv}(\operatorname{ext} K)$ is convex and contains y and z, Step 1 gives $x \in \operatorname{conv}(\operatorname{ext} K)$. As $x \in K$ was arbitrary, this proves the inclusion and the theorem. The last sentence of the statement follows, since $\operatorname{conv} \emptyset = \emptyset$.

■

Warning. **Compactness cannot be dropped.** The quarter-plane $Q = \{x_1 \ge 0,\ x_2 \ge 0\}$ is closed and convex with exactly one extreme point, the origin, and $\operatorname{conv}\{\mathbf{0}\} = \{\mathbf{0}\} \ne Q$. The open disk, bounded but not closed, has no extreme points at all, since each of its points $\mathbf{x}$ is the midpoint of $\mathbf{x} \pm \varepsilon \mathbf{e}_1$ for small $\varepsilon > 0$.

Combined with Carathéodory's theorem (Theorem 18.2.2), Minkowski's theorem says more: every point of a compact convex K of dimension d is a convex combination of at most $d + 1$ extreme points. To see this, pick $x_0 \in K$ and apply Theorem 18.2.2 to the translate $\operatorname{ext} K - x_0$, which lies in the d-dimensional space W_K; translating back preserves convex combinations, because their coefficients add up to 1.

The consequence used most often concerns linear functions. By the Riesz representation theorem (Theorem 11.5.1), every linear functional on V is $x \mapsto \langle x, a\rangle$ for some $a \in V$, and it is continuous by Theorem 11.1.5, since $|\langle x, a\rangle - \langle x', a\rangle| \le \|a\|\|x - x'\|$.

Corollary 18.8.6. Linear Functions Peak at Extreme Points

Let $K \subseteq V$ be a nonempty compact convex set and $a \in V$. Then the maximum of $x \mapsto \langle x, a\rangle$ over K exists and is attained at an extreme point of K. The same holds for the minimum.

Idea. Minkowski's theorem writes a maximizer as an average of extreme points, and an average of values at most M equals M only if every value does.

Proof. By fact (A4) of Chapter 16's introduction, the continuous function $x \mapsto \langle x, a\rangle$ attains a maximum M on K, at some $x^* \in K$. By Theorem 18.8.5 and Theorem 18.1.3, $x^* = \sum_{i=1}^{k} t_i v_i$

with $v_i \in \operatorname{ext} K$, $t_i > 0$, and $\sum_i t_i = 1$. Then

$$M = \langle x^*, a \rangle = \sum_{i=1}^{k} t_i \langle v_i, a \rangle,$$

while each $\langle v_i, a \rangle \le M$. If some $\langle v_i, a \rangle < M$, the weighted average would be strictly less than M, because every $t_i > 0$. So $\langle v_i, a \rangle = M$ for every i, and v_1 is an extreme point at which the maximum is attained. For the minimum, apply this to $-a$.

■

This is why linear programs are solved at corners: when one of Section 6's feasible regions is bounded and nonempty, the corollary puts an optimum at an extreme point, and Section 9 shows that there are finitely many. The same corollary gives a new view of Chapter 17's variational formulas, as we will see after the density matrices.

18.8.4 Three unit balls

The unit balls of the three norms of Chapter 16 (Definition 16.1.2, Example (The 1-, 2-, and ∞-norms on F^n)) are compact and convex. They are convex by the remark after Definition 16.1.2. They are bounded, and closed, because $\|\mathbf{x}_k\| \le 1$ and $\mathbf{x}_k \to \mathbf{x}$ give $\|\mathbf{x}\| \le \|\mathbf{x}_k\| + \|\mathbf{x} - \mathbf{x}_k\| \le 1 + \|\mathbf{x} - \mathbf{x}_k\|$ for every k, hence $\|\mathbf{x}\| \le 1$; so they are compact by Corollary 16.2.5.

Example. The diamond, the cube, and the sphere

In $\mathbb{R}^n$ with $n \ge 1$, show that:

(a) $\operatorname{ext} B_1 = \{\pm\mathbf{e}_1, \dots, \pm\mathbf{e}_n\}$, $2n$ points;
(b) $\operatorname{ext} B_\infty = \{\mathbf{s} : \text{every } s_i = \pm 1\}$, the 2^n sign vectors;
(c) $\operatorname{ext} B_2 = \{\mathbf{x} : \|\mathbf{x}\|_2 = 1\}$, the whole unit sphere.

Solution. In each case we show that the listed points are extreme and that no other point is. The second half uses the same device each time: a point $\mathbf{x}$ of the ball is not extreme as soon as we find $\mathbf{v} \neq \mathbf{0}$ with $\mathbf{x} \pm \mathbf{v}$ both in the ball, because $\mathbf{x}$ is their midpoint.

(a) *The points $\pm\mathbf{e}_i$ are extreme.* Let $\mathbf{e}_i = \frac{1}{2}(\mathbf{y} + \mathbf{z})$ with $\|\mathbf{y}\|_1, \|\mathbf{z}\|_1 \le 1$. Then

$$1 = \tfrac{1}{2}(y_i + z_i) \le \tfrac{1}{2}(|y_i| + |z_i|) \le \tfrac{1}{2}(\|\mathbf{y}\|_1 + \|\mathbf{z}\|_1) \le 1,$$

so every inequality is an equality. The first forces $y_i = |y_i|$ and $z_i = |z_i|$. The second, since $|y_i| \le \|\mathbf{y}\|_1$ and $|z_i| \le \|\mathbf{z}\|_1$ separately, forces $|y_i| = \|\mathbf{y}\|_1$, so all other entries of $\mathbf{y}$ vanish, and the third forces $\|\mathbf{y}\|_1 = 1$. Hence $\mathbf{y} = \mathbf{e}_i$, and likewise $\mathbf{z} = \mathbf{e}_i$. By Lemma 18.8.1, $\mathbf{e}_i$ is extreme; for $-\mathbf{e}_i$ apply this to $-\mathbf{y}, -\mathbf{z}$.
No other point is extreme. Let $\mathbf{x} \in B_1$. If $\|\mathbf{x}\|_1 < 1$, put $\varepsilon = 1 - \|\mathbf{x}\|_1 > 0$ and $\mathbf{v} = \varepsilon\mathbf{e}_1$; then $\|\mathbf{x} \pm \mathbf{v}\|_1 \le \|\mathbf{x}\|_1 + \varepsilon = 1$. If $\|\mathbf{x}\|_1 = 1$ and $\mathbf{x}$ is not some $\pm\mathbf{e}_i$, then $\mathbf{x}$ has two nonzero entries x_i, x_j with $i \neq j$. Let $\sigma_i, \sigma_j \in \{\pm 1\}$ be their signs, $\varepsilon = \min(|x_i|, |x_j|) > 0$, and $\mathbf{v} = \varepsilon(\sigma_i\mathbf{e}_i - \sigma_j\mathbf{e}_j)$. Adding $\mathbf{v}$ moves $|x_i|$ up by ε and $|x_j|$ down by ε, and subtracting $\mathbf{v}$ does the opposite; neither entry changes sign, because $\varepsilon \le |x_i|, |x_j|$. So $\|\mathbf{x} \pm \mathbf{v}\|_1 = \|\mathbf{x}\|_1 = 1$.

(b) *Sign vectors are extreme.* Let $\mathbf{s} = \frac{1}{2}(\mathbf{y} + \mathbf{z})$ with $\|\mathbf{y}\|_\infty, \|\mathbf{z}\|_\infty \le 1$. For each i, $s_i = \pm 1$

is the average of y_i and z_i, both in $[-1,1]$. If $s_i = 1$, then $y_i + z_i = 2$ with $y_i, z_i \le 1$ forces $y_i = z_i = 1$; if $s_i = -1$, then $y_i + z_i = -2$ with $y_i, z_i \ge -1$ forces $y_i = z_i = -1$. So $\mathbf{y} = \mathbf{z} = \mathbf{s}$.

No other point is extreme. If $\mathbf{x} \in B_\infty$ is not a sign vector, some $|x_i| < 1$. Put $\mathbf{v} = (1 - |x_i|)\mathbf{e}_i \neq \mathbf{0}$. Then the i-th entries of $\mathbf{x} \pm \mathbf{v}$ are $x_i \pm (1 - |x_i|)$, both in $[-1,1]$, and the other entries are unchanged.

(c) *Unit vectors are extreme.* Let $\|\mathbf{x}\|_2 = 1$ and $\mathbf{x} = \frac{1}{2}(\mathbf{y} + \mathbf{z})$ with $\|\mathbf{y}\|_2, \|\mathbf{z}\|_2 \le 1$. The parallelogram law of Chapter 11 §01, which comes from expanding with Theorem 11.1.2 (c), gives

$$\|\mathbf{y} - \mathbf{z}\|_2^2 = 2\|\mathbf{y}\|_2^2 + 2\|\mathbf{z}\|_2^2 - \|\mathbf{y} + \mathbf{z}\|_2^2 \le 2 + 2 - \|2\mathbf{x}\|_2^2 = 0,$$

so $\mathbf{y} = \mathbf{z}$.

No other point is extreme. If $\|\mathbf{x}\|_2 < 1$, put $\mathbf{v} = (1 - \|\mathbf{x}\|_2)\mathbf{e}_1$; then $\|\mathbf{x} \pm \mathbf{v}\|_2 \le \|\mathbf{x}\|_2 + (1 - \|\mathbf{x}\|_2) = 1$.

In (a) and (b), Minkowski's theorem (Theorem 18.8.5) now makes B_1 the hull of the $2n$ points $\pm\mathbf{e}_i$, and B_∞ the hull of the 2^n sign vectors. In (c) it says only that the ball is the hull of the sphere, and the infinitely many extreme points are why B_2 is not a polytope.

18.8.5 Density matrices

Sums and nonnegative multiples of positive semidefinite matrices are positive semidefinite, because $\mathbf{x}^\top(s\mathbf{A} + t\mathbf{B})\mathbf{x} = s\,\mathbf{x}^\top\mathbf{A}\mathbf{x} + t\,\mathbf{x}^\top\mathbf{B}\mathbf{x}$; Chapter 13 §01 recorded this in Exercise (C2: The cone of positive matrices) (a). Cutting that cone with a single linear condition gives a compact convex set whose extreme points we can find exactly.

Definition 18.8.3. Density Matrix

A **density matrix** of size n is a real symmetric matrix $\mathbf{X} \in M_n(\mathbb{R})$ with $\mathbf{X} \succeq \mathbf{O}$ and $\operatorname{tr}\mathbf{X} = 1$. We write $\mathcal{D}_n$ for the set of them.

The name comes from quantum mechanics, which uses the complex Hermitian version; the proofs below carry over, with $\mathbf{x}^*$ for $\mathbf{x}^\top$, $|\mathbf{x}^*\mathbf{y}|^2$ for $(\mathbf{x}^\top\mathbf{y})^2$, and Corollary 12.4.3 for Corollary 12.5.2. The set $\mathcal{D}_n$ is convex: if $\mathbf{X}, \mathbf{Y} \in \mathcal{D}_n$ and $0 \le t \le 1$, then $t\mathbf{X} + (1-t)\mathbf{Y}$ is symmetric, $\mathbf{x}^\top(t\mathbf{X} + (1-t)\mathbf{Y})\mathbf{x} = t\,\mathbf{x}^\top\mathbf{X}\mathbf{x} + (1-t)\,\mathbf{x}^\top\mathbf{Y}\mathbf{x} \ge 0$, and the trace is $t + (1-t) = 1$. It is compact in $M_n(\mathbb{R})$: the three conditions are preserved by entrywise limits, since symmetry and trace are linear and each inequality $\mathbf{x}^\top\mathbf{X}_j\mathbf{x} \ge 0$ survives the limit; and it is bounded, since by Corollary 12.5.2

$$\|\mathbf{X}\|_F^2 = \operatorname{tr}(\mathbf{X}^2) = \sum_{i=1}^{n} \lambda_i(\mathbf{X})^2 \le \Big(\sum_{i=1}^{n} \lambda_i(\mathbf{X})\Big)^2 = 1,$$

where the inequality uses $\lambda_i(\mathbf{X}) \ge 0$ (Theorem 13.1.4).

For a unit vector $\mathbf{x} \in \mathbb{R}^n$, the matrix $\mathbf{x}\mathbf{x}^\top$ is in $\mathcal{D}_n$: it is symmetric, $\mathbf{y}^\top\mathbf{x}\mathbf{x}^\top\mathbf{y} = (\mathbf{x}^\top\mathbf{y})^2 \ge 0$, and $\operatorname{tr}(\mathbf{x}\mathbf{x}^\top) = \mathbf{x}^\top\mathbf{x} = 1$. It is the orthogonal projection onto the line $\operatorname{span}\{\mathbf{x}\}$. One more fact about positive semidefinite matrices is needed: a zero value of the quadratic form pins down a null vector.

Lemma 18.8.7. A Zero of the Form Is a Null Vector

Let $\mathbf{B} \in M_n(\mathbb{R})$ be symmetric with $\mathbf{B} \succeq \mathbf{O}$, and let $\mathbf{y} \in \mathbb{R}^n$ satisfy $\mathbf{y}^\top \mathbf{B}\mathbf{y} = 0$. Then $\mathbf{B}\mathbf{y} = \mathbf{0}$.

Proof. By Corollary 12.5.2, $\mathbf{B} = \sum_{i=1}^n \mu_i \mathbf{q}_i \mathbf{q}_i^\top$ for an orthonormal basis $(\mathbf{q}_1, \dots, \mathbf{q}_n)$ of eigenvectors, with eigenvalues $\mu_i \geq 0$ by Theorem 13.1.4. Then $0 = \mathbf{y}^\top \mathbf{B}\mathbf{y} = \sum_i \mu_i (\mathbf{q}_i^\top \mathbf{y})^2$, a sum of nonnegative terms, so $\mu_i (\mathbf{q}_i^\top \mathbf{y})^2 = 0$ for every i. Hence $\mu_i (\mathbf{q}_i^\top \mathbf{y}) = 0$ for every i, and $\mathbf{B}\mathbf{y} = \sum_i \mu_i (\mathbf{q}_i^\top \mathbf{y}) \mathbf{q}_i = \mathbf{0}$.

■

Theorem 18.8.8. Extreme Points of the Density Matrices

For $n \geq 1$,

$$\operatorname{ext} \mathcal{D}_n = \{\mathbf{x}\mathbf{x}^\top : \mathbf{x} \in \mathbb{R}^n,\ \|\mathbf{x}\| = 1\},$$

the orthogonal projections of rank one. Every $\mathbf{X} \in \mathcal{D}_n$ is a convex combination of at most n of them.

Idea. The spectral theorem writes $\mathbf{X}$ as $\sum_i \lambda_i \mathbf{q}_i \mathbf{q}_i^\top$, with weights $\lambda_i \geq 0$ adding up to $\operatorname{tr} \mathbf{X} = 1$. That is already a convex combination of rank-one projections, which gives the last sentence and shows at once that a matrix with two positive eigenvalues is not extreme. For the converse, suppose $\mathbf{x}\mathbf{x}^\top$ is an average of two density matrices. On every vector orthogonal to $\mathbf{x}$, the form of $\mathbf{x}\mathbf{x}^\top$ is zero, so the forms of both pieces are zero there too, and Lemma 18.8.7 turns that into "both pieces kill $\mathbf{x}^\perp$". A symmetric matrix that kills $\mathbf{x}^\perp$ is a multiple of $\mathbf{x}\mathbf{x}^\top$, and the trace fixes the multiple.

Proof. Let $\mathbf{X} \in \mathcal{D}_n$. By Corollary 12.5.2, $\mathbf{X} = \sum_{i=1}^n \lambda_i \mathbf{q}_i \mathbf{q}_i^\top$ with $(\mathbf{q}_1, \dots, \mathbf{q}_n)$ orthonormal and $\lambda_i = \lambda_i(\mathbf{X})$. Each $\lambda_i \geq 0$ by Theorem 13.1.4, and $\sum_i \lambda_i = \operatorname{tr} \mathbf{X} = 1$, the trace being the sum of the eigenvalues. Dropping the zero terms, this writes $\mathbf{X}$ as a convex combination of at most n matrices $\mathbf{q}_i \mathbf{q}_i^\top$, which proves the last sentence.

Only rank-one projections are extreme. At least one eigenvalue of $\mathbf{X}$ is positive, since the trace is 1. Suppose first that $\mathbf{X}$ has at least two positive eigenvalues. As they are listed decreasingly, $\lambda_1 \geq \lambda_2 > 0$, and then $0 < \lambda_1 < 1$, because the total is 1 and λ_2 is part of it. Put

$$\mathbf{Y} = \frac{1}{1 - \lambda_1} \sum_{i=2}^n \lambda_i \mathbf{q}_i \mathbf{q}_i^\top.$$

Its coefficients are nonnegative and add up to 1, so $\mathbf{Y} \in \mathcal{D}_n$ by the convexity of $\mathcal{D}_n$, and $\mathbf{X} = \lambda_1 \mathbf{q}_1 \mathbf{q}_1^\top + (1 - \lambda_1)\mathbf{Y}$. The two ends differ, since $\mathbf{q}_1 \mathbf{q}_1^\top \mathbf{q}_1 = \mathbf{q}_1$ while $\mathbf{Y}\mathbf{q}_1 = \mathbf{0}$ by orthonormality. So $\mathbf{X}$ is not extreme. If instead $\mathbf{X}$ has exactly one positive eigenvalue, that eigenvalue is the whole trace, 1, and $\mathbf{X} = \mathbf{q}_1 \mathbf{q}_1^\top$. So every extreme point has the form $\mathbf{x}\mathbf{x}^\top$ with $\|\mathbf{x}\| = 1$.

Every rank-one projection is extreme. Let $\|\mathbf{x}\| = 1$ and $\mathbf{x}\mathbf{x}^\top = \frac{1}{2}(\mathbf{B} + \mathbf{C})$ with

$\mathbf{B}, \mathbf{C} \in \mathcal{D}_n$. For every $\mathbf{y}$ with $\mathbf{x}^{\top}\mathbf{y} = 0$,

$$0 = (\mathbf{x}^{\top}\mathbf{y})^2 = \mathbf{y}^{\top}\mathbf{x}\mathbf{x}^{\top}\mathbf{y} = \tfrac{1}{2}\,\mathbf{y}^{\top}\mathbf{B}\mathbf{y} + \tfrac{1}{2}\,\mathbf{y}^{\top}\mathbf{C}\mathbf{y},$$

a sum of two nonnegative numbers, so $\mathbf{y}^{\top}\mathbf{B}\mathbf{y} = 0$, and $\mathbf{B}\mathbf{y} = \mathbf{0}$ by Lemma 18.8.7. Now let $\mathbf{v} \in \mathbb{R}^n$ be arbitrary, and write $\mathbf{v} = (\mathbf{x}^{\top}\mathbf{v})\mathbf{x} + \mathbf{y}$ with $\mathbf{y} = \mathbf{v} - (\mathbf{x}^{\top}\mathbf{v})\mathbf{x}$, so that $\mathbf{x}^{\top}\mathbf{y} = \mathbf{x}^{\top}\mathbf{v} - \mathbf{x}^{\top}\mathbf{v} = 0$ because $\mathbf{x}^{\top}\mathbf{x} = 1$. Then $\mathbf{B}\mathbf{v} = (\mathbf{x}^{\top}\mathbf{v})\mathbf{B}\mathbf{x}$ for every $\mathbf{v}$, that is, $\mathbf{B} = (\mathbf{B}\mathbf{x})\mathbf{x}^{\top}$. Since $\mathbf{B}$ is symmetric, also $\mathbf{B} = \mathbf{B}^{\top} = \mathbf{x}(\mathbf{B}\mathbf{x})^{\top}$, and applying both expressions to $\mathbf{x}$ gives $\mathbf{B}\mathbf{x} = (\mathbf{x}^{\top}\mathbf{B}\mathbf{x})\,\mathbf{x}$. Hence $\mathbf{B} = c\,\mathbf{x}\mathbf{x}^{\top}$ with $c = \mathbf{x}^{\top}\mathbf{B}\mathbf{x}$, and $1 = \operatorname{tr}\mathbf{B} = c\,\operatorname{tr}(\mathbf{x}\mathbf{x}^{\top}) = c$. So $\mathbf{B} = \mathbf{x}\mathbf{x}^{\top}$, and by symmetry of the argument $\mathbf{C} = \mathbf{x}\mathbf{x}^{\top}$. By Lemma 18.8.1, $\mathbf{x}\mathbf{x}^{\top}$ is extreme. This proves the theorem.

■

For instance, $\mathbf{X} = \frac{1}{15}\begin{pmatrix} 9 & 2 \\ 2 & 6 \end{pmatrix}$ is symmetric with trace 1, and $\mathbf{X}(2,1) = \frac{2}{3}(2,1)$, $\mathbf{X}(1,-2) = \frac{1}{3}(1,-2)$. With $\mathbf{q}_1 = \frac{1}{\sqrt{5}}(2,1)$ and $\mathbf{q}_2 = \frac{1}{\sqrt{5}}(1,-2)$, the proof's splitting is

$$\begin{aligned} \mathbf{X} &= \frac{2}{3}\,\mathbf{q}_1\mathbf{q}_1^{\top} + \frac{1}{3}\,\mathbf{q}_2\mathbf{q}_2^{\top} \\ &= \frac{2}{3}\cdot\frac{1}{5}\begin{pmatrix} 4 & 2 \\ 2 & 1 \end{pmatrix} + \frac{1}{3}\cdot\frac{1}{5}\begin{pmatrix} 1 & -2 \\ -2 & 4 \end{pmatrix}, \end{aligned}$$

and the entries check: $\frac{8+1}{15} = \frac{9}{15}$, $\frac{4-2}{15} = \frac{2}{15}$, $\frac{2+4}{15} = \frac{6}{15}$.

Here the spectral theorem handed us the decomposition promised by Minkowski's theorem, with at most n terms. Carathéodory's theorem, in the coset of symmetric matrices of trace 1, of dimension $n(n+1)/2 - 1$, would only have promised $n(n+1)/2$. Structure beats counting.

Now the promised instance of Corollary 18.8.6. For a real symmetric $\mathbf{A}$, the function $\mathbf{X} \mapsto \operatorname{tr}(\mathbf{A}\mathbf{X}) = \langle \mathbf{X}, \mathbf{A} \rangle$ is linear on $M_n(\mathbb{R})$, with $\langle \cdot, \cdot \rangle$ the Frobenius inner product.

Example. The largest eigenvalue as a linear maximum

Let $\mathbf{A} \in M_n(\mathbb{R})$ be symmetric. Show that $\lambda_1(\mathbf{A}) = \max\{\operatorname{tr}(\mathbf{A}\mathbf{X}) : \mathbf{X} \in \mathcal{D}_n\}$, and find the maximum for $\mathbf{A} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$.

Solution. The set $\mathcal{D}_n$ is nonempty, compact, and convex, so by Corollary 18.8.6 the maximum exists and is attained at an extreme point, which by Theorem 18.8.8 is $\mathbf{x}\mathbf{x}^{\top}$ with $\|\mathbf{x}\| = 1$. There,

$$\operatorname{tr}(\mathbf{A}\mathbf{x}\mathbf{x}^{\top}) = \operatorname{tr}(\mathbf{x}^{\top}\mathbf{A}\mathbf{x}) = \mathbf{x}^{\top}\mathbf{A}\mathbf{x} \le \lambda_1(\mathbf{A})$$

by the cyclic property of the trace and Lemma 13.5.2. The bound is attained when $\mathbf{x}$ is a unit eigenvector for $\lambda_1(\mathbf{A})$, by the same lemma, so the maximum is $\lambda_1(\mathbf{A})$.

For the given $\mathbf{A}$, $\mathbf{A}(1,1) = (3,3)$ and $\mathbf{A}(1,-1) = (1,-1)$, so $\lambda_1(\mathbf{A}) = 3$, and the maximizer

is $\mathbf{x}\mathbf{x}^\top$ with $\mathbf{x} = \frac{1}{\sqrt{2}}(1,1)$:

$$\mathbf{X} = \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}, \qquad \mathbf{AX} = \frac{1}{2}\begin{pmatrix} 3 & 3 \\ 3 & 3 \end{pmatrix}, \qquad \operatorname{tr}(\mathbf{AX}) = 3.$$

This is the case $k = 1$ of Ky Fan's maximum principle (Theorem 17.6.2); for $k > 1$, $\operatorname{tr}(\mathbf{Q}^\top\mathbf{A}\mathbf{Q}) = \operatorname{tr}(\mathbf{A}\mathbf{Q}\mathbf{Q}^\top)$ is again linear in the projection $\mathbf{Q}\mathbf{Q}^\top$, but this section does not show that the rank-k projections are the extreme points of any convex set.

18.8.6 Doubly stochastic matrices

Chapter 9 met **stochastic** matrices, whose entries are nonnegative and whose columns add up to 1 (Definition 9.11.1). Asking the rows to add up to 1 as well gives a set with remarkable corners.

Definition 18.8.4. Doubly Stochastic Matrix

A matrix $\mathbf{M} \in M_n(\mathbb{R})$ is **doubly stochastic** if all its entries are nonnegative and **every** row and **every** column adds up to 1. We write Ω_n for the set of doubly stochastic $n \times n$ matrices.

Every permutation matrix $\mathbf{P}_\sigma$ is doubly stochastic, since it has exactly one 1 in each row and each column and zeros elsewhere (the remark after Definition 3.4.2). So is $\frac{1}{n}\mathbf{J}$, the matrix with every entry $\frac{1}{n}$. The set Ω_n is convex, since nonnegativity and the $2n$ linear conditions on sums survive convex combinations. It is compact in $M_n(\mathbb{R})$: its entries lie in $[0,1]$, so it is bounded, and each defining condition, being a nonstrict inequality or a linear equation in the entries, survives limits. By Minkowski's theorem, Ω_n is the hull of its extreme points, and the natural candidates are the permutation matrices.

Proposition 18.8.9. Permutation Matrices Are Extreme

Every permutation matrix is an extreme point of Ω_n.

Idea. Nonnegative entries cannot cancel, so a zero of $\mathbf{P}$ forces zeros in both pieces, and the row sums then leave no freedom.

Proof. Let $\mathbf{P} = \mathbf{P}_\sigma$ and $\mathbf{P} = t\mathbf{A} + (1-t)\mathbf{B}$ with $\mathbf{A}, \mathbf{B} \in \Omega_n$ and $0 < t < 1$. Wherever $\mathbf{P}$ has a zero entry, $0 = t\,a_{ij} + (1-t)\,b_{ij}$ with both terms nonnegative, so $a_{ij} = b_{ij} = 0$. Each row of $\mathbf{A}$ therefore has at most one nonzero entry, in the position where $\mathbf{P}$ has its 1, and since the row adds up to 1, that entry is 1. So $\mathbf{A} = \mathbf{P}$, and likewise $\mathbf{B} = \mathbf{P}$. ∎

The proof never used the column sums. The converse is much deeper.

Birkhoff's theorem. *The extreme points of Ω_n are exactly the $n!$ permutation matrices. Equivalently, every doubly stochastic matrix is a convex combination of permutation matrices.*

The two forms are equivalent by what we have. If the permutation matrices are all the extreme points, Minkowski's theorem (Theorem 18.8.5) makes Ω_n their hull. Conversely, if Ω_n is the hull of the permutation matrices, Proposition 18.8.2 puts every extreme point among them, and Proposition 18.8.9 shows that each of them is one. We do not prove Birkhoff's theorem here; Chapter 19 does. For $n = 2$ it can be checked by hand, as the next Quick check shows.

Quick check. Show that every $\mathbf{M} \in \Omega_2$ has the form $\begin{pmatrix} t & 1-t \\ 1-t & t \end{pmatrix}$ with $0 \le t \le 1$, and deduce Birkhoff's theorem for $n = 2$.

Solution. Let $t = m_{11}$. The first row sum gives $m_{12} = 1 - t$, the first column sum gives $m_{21} = 1 - t$, and the second row sum gives $m_{22} = 1 - (1 - t) = t$. Nonnegativity gives $0 \le t \le 1$. So

$$\mathbf{M} = t \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + (1-t) \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

a convex combination of the two 2×2 permutation matrices. Hence Ω_2 is the segment between them, and its extreme points are its two ends.

Birkhoff's theorem would also give a second route to Schur's theorem, Theorem 17.8.2, which Chapter 17 §08 proved without it. If $\mathbf{A} = \mathbf{Q}\mathbf{D}\mathbf{Q}^{\top}$ with $\mathbf{Q}$ orthogonal and $\mathbf{D} = \operatorname{diag}(\lambda_1(\mathbf{A}), \dots, \lambda_n(\mathbf{A}))$ (Corollary 12.5.2, after reordering), then $\mathbf{d}(\mathbf{A}) = \mathbf{S}\boldsymbol{\lambda}(\mathbf{A})$ with $\mathbf{S} = (q_{ij}^2)$, doubly stochastic because the rows and columns of $\mathbf{Q}$ are unit vectors, so $\mathbf{d}(\mathbf{A})$ would be an average of rearrangements of $\boldsymbol{\lambda}(\mathbf{A})$, each majorized by $\boldsymbol{\lambda}(\mathbf{A})$, hence majorized by $\boldsymbol{\lambda}(\mathbf{A})$ itself by Example (The vectors majorized by a fixed vector).

18.8.7 When the extreme points are not closed

In every example so far, the extreme points have formed a closed set: finitely many points, a sphere, or the rank-one projections. That is not always so.

Example. A double cone over a circle

In $\mathbb{R}^3$, let $\Gamma = \{(1,0,0) + (\mathbf{u}, 0) : \mathbf{u} \in \mathbb{R}^2,\ \|\mathbf{u}\| = 1\}$ be the circle of radius 1 centered at $(1,0,0)$ in the plane $x_3 = 0$, where $(\mathbf{u}, 0) = (u_1, u_2, 0)$; it passes through the origin, at $\mathbf{u} = (-1, 0)$. Let $\mathbf{p} = (0,0,1)$, $\mathbf{q} = (0,0,-1)$, $S = \Gamma \cup \{\mathbf{p}, \mathbf{q}\}$, and $K = \operatorname{conv} S$. Show that K is compact and convex, that

$$\operatorname{ext} K = (\Gamma \setminus \{\mathbf{0}\}) \cup \{\mathbf{p}, \mathbf{q}\},$$

and that this set is not closed.

Solution. Compact and convex. K is convex as a convex hull. The set S is bounded and closed: a limit of points of Γ satisfies the closed conditions $(x_1 - 1)^2 + x_2^2 = 1$ and $x_3 = 0$, which describe Γ, and adding two points keeps it closed. So S is compact, and K is compact by Corollary 18.2.3.

The candidates. By Proposition 18.8.2, $\operatorname{ext} K \subseteq S$.

The origin is not extreme. $\mathbf{0} \in \Gamma \subseteq K$, and $\mathbf{0} = \frac{1}{2}(\mathbf{p} + \mathbf{q})$ with $\mathbf{p} \neq \mathbf{q}$.

The points $\mathbf{p}$ *and* $\mathbf{q}$ *are extreme.* Let $\varphi(\mathbf{x}) = x_3$. On S, $\varphi(\mathbf{p}) = 1$, $\varphi(\mathbf{q}) = -1$, and $\varphi = 0$ on Γ. A point $\mathbf{x} = \sum_i t_i \mathbf{s}_i$ of K, written as a convex combination with every $t_i > 0$ (Theorem 18.1.3), has $\varphi(\mathbf{x}) = \sum_i t_i \varphi(\mathbf{s}_i) \leq 1$, with equality only if every $\mathbf{s}_i = \mathbf{p}$, that is, only if $\mathbf{x} = \mathbf{p}$. So $\varphi \leq 1$ on K with equality exactly at $\mathbf{p}$, and $\mathbf{p}$ is extreme by Proposition 18.8.4 (b). The same argument with $-\varphi$ handles $\mathbf{q}$.

The other points of Γ *are extreme.* Fix $\mathbf{c} = (1, 0, 0) + (\mathbf{u}, 0)$ with $\mathbf{c} \neq \mathbf{0}$, that is, $\mathbf{u} \neq (-1, 0)$; then $u_1 > -1$, because a unit vector with $u_1 = -1$ is $(-1, 0)$. Let $\varphi(\mathbf{x}) = \langle \mathbf{u}, (x_1, x_2) \rangle = u_1 x_1 + u_2 x_2$, so that $\varphi(\mathbf{c}) = u_1 + \|\mathbf{u}\|^2 = u_1 + 1$. For a point $\mathbf{s} = (1, 0, 0) + (\mathbf{u}', 0)$ of Γ,

$$\varphi(\mathbf{s}) = u_1 + \langle \mathbf{u}, \mathbf{u}' \rangle \leq u_1 + \|\mathbf{u}\| \|\mathbf{u}'\| = u_1 + 1$$

by Theorem 11.1.5. Equality forces $(\mathbf{u}, \mathbf{u}')$ to be linearly dependent, so $\mathbf{u}' = \pm \mathbf{u}$, and $\mathbf{u}' = -\mathbf{u}$ would give $\langle \mathbf{u}, \mathbf{u}' \rangle = -1$; so equality holds only at $\mathbf{u}' = \mathbf{u}$, that is, at $\mathbf{s} = \mathbf{c}$. Also $\varphi(\mathbf{p}) = \varphi(\mathbf{q}) = 0 < 1 + u_1 = \varphi(\mathbf{c})$. So $\varphi(\mathbf{s}) < \varphi(\mathbf{c})$ for every $\mathbf{s} \in S$ other than $\mathbf{c}$. As before, a convex combination $\mathbf{x} = \sum_i t_i \mathbf{s}_i$ with all $t_i > 0$ has $\varphi(\mathbf{x}) \leq \varphi(\mathbf{c})$, with equality only if every $\mathbf{s}_i = \mathbf{c}$. By Proposition 18.8.4 (b), $\mathbf{c}$ is extreme.

Not closed. For $k \geq 1$ let $\mathbf{c}_k = (1 - \sqrt{1 - 1/k^2},\ 1/k,\ 0)$. It is $(1, 0, 0) + (\mathbf{u}_k, 0)$ with $\mathbf{u}_k = (-\sqrt{1 - 1/k^2}, 1/k)$, a unit vector, so $\mathbf{c}_k \in \Gamma$; and $\mathbf{c}_k \neq \mathbf{0}$, as its second entry is $1/k$. So every $\mathbf{c}_k$ is extreme. For $0 \leq h \leq 1$ we have $(1 - h)^2 \leq 1 - h$, hence $1 - h \leq \sqrt{1 - h} \leq 1$ and $0 \leq 1 - \sqrt{1 - h} \leq h$. With $h = 1/k^2$, the first entry of $\mathbf{c}_k$ lies in $[0, 1/k^2]$, so $\mathbf{c}_k \to \mathbf{0}$, which is not extreme. So $\operatorname{ext} K$ contains a sequence converging to a point outside it, and it is not closed (Definition 18.1.6).

Geometrically, K is a double cone over the disk bounded by Γ, with apexes $\mathbf{p}$ and $\mathbf{q}$. The segment from $\mathbf{p}$ to $\mathbf{q}$ touches the rim of the disk at the origin, which robs the origin of its extremeness while leaving its neighbors on the circle intact. So the set of extreme points of a compact convex set need not be closed, and hence need not be compact.

18.8.8 Exercises

18.8.8.1 A. Check your understanding

Exercise. A1

(a) State the definition of an extreme point of a convex set $C \subseteq V$.

(b) True or false: every boundary point of a compact convex set in $\mathbb{R}^2$ is an extreme point. Justify your answer.

(c) True or false: there is a nonempty closed convex subset of $\mathbb{R}^2$ with no extreme points. Justify your answer.

(d) State Theorem 18.8.5, and name the two facts from Chapter 16's introduction that its proof uses.

(e) True or false: if $K = \operatorname{conv} S$ with S finite, then every point of S is an extreme point of K. Justify your answer.

Solution.

(a) A point $x \in C$ is extreme if $x = ty + (1-t)z$ with $y, z \in C$ and $0 < t < 1$ forces $y = z = x$.

(b) False. In the square $[-1,1]^2$, the point $(1,0)$ is on the boundary, since $(1+\varepsilon, 0)$ lies outside for every $\varepsilon > 0$, but it is the midpoint of $(1,1)$ and $(1,-1)$, both in the square.

(c) True. The closed half-plane $\{x_2 \geq 0\}$ is one: each of its points $\mathbf{x}$ is the midpoint of $\mathbf{x} \pm \mathbf{e}_1$, which lie in it. A line is another.

(d) A compact convex subset K of a finite-dimensional real inner product space is the convex hull of its extreme points. The proof uses (A3), to see that the set of t with $x + tu \in K$ is compact, and (A4), the extreme value theorem, to find its smallest and largest elements.

(e) False. Take $S = \{0, 1, 2\} \subseteq \mathbb{R}$. Then $K = [0, 2]$, whose extreme points are 0 and 2; the point $1 = \frac{1}{2}(0+2)$ of S is not extreme. What is true is the reverse inclusion $\operatorname{ext} K \subseteq S$ (Proposition 18.8.2).

18.8.8.2 B. Practice

Exercise. B1: Finding extreme points

Determine the extreme points of each of the following convex subsets of $\mathbb{R}^2$. Justify your answer.

(a) $\operatorname{conv}\{(0,0), (4,0), (0,2), (2,1), (1,1)\}$.

(b) $P = \{(x, y) : x^2 \leq y \leq 1\}$.

(c) The strip $\{(x, y) : 0 \leq y \leq 1\}$.

Hint: for (b), use Proposition 18.8.4 (b) with a tangent line.

Solution.

(a) By Proposition 18.8.2, every extreme point is one of the five listed points. The point $(2,1) = \frac{1}{2}((4,0) + (0,2))$ is not extreme. Neither is $(1,1) = \frac{1}{2}((2,1) + (0,1))$, since $(0,1) = \frac{1}{2}((0,0) + (0,2))$ lies in the hull, and $(2,1) \neq (0,1)$. The three corners are extreme: each is the unique maximizer over the five points of a linear function, namely $-x - y$, x, and y respectively, hence the unique maximizer over the hull by the convex-combination argument of Example (A double cone over a circle), and Proposition 18.8.4 (b) applies. So the extreme points are $(0,0)$, $(4,0)$, and $(0,2)$.

(b) *Points (a, a^2) with $-1 \leq a \leq 1$ are extreme.* Let $\varphi(x, y) = 2ax - y$. For $(x, y) \in P$,

$$\varphi(x, y) \leq 2ax - x^2 = a^2 - (x - a)^2 \leq a^2,$$

with equality only if $x = a$ and $y = x^2 = a^2$. So (a, a^2) is the unique maximizer of φ over P, and it lies in P because $a^2 \leq 1$. By Proposition 18.8.4 (b) it is extreme.
No other point is extreme. Let $(x, y) \in P$ with $y \neq x^2$, so $x^2 < y \leq 1$. If $y < 1$, let $\varepsilon = \min(y - x^2, 1 - y) > 0$; then $(x, y \pm \varepsilon) \in P$ and (x, y) is their midpoint. If $y = 1$, then $|x| < 1$; let $\varepsilon = 1 - |x| > 0$; then $(x \pm \varepsilon)^2 \leq (|x| + \varepsilon)^2 = 1$, so $(x \pm \varepsilon, 1) \in P$, and $(x, 1)$ is their midpoint. So $\operatorname{ext} P = \{(a, a^2) : -1 \leq a \leq 1\}$.

(c) None: every point $\mathbf{x}$ of the strip is the midpoint of $\mathbf{x} \pm \mathbf{e}_1$, which lie in the strip. The strip is closed and convex but not bounded, so Minkowski's theorem (Theorem 18.8.5) does not apply.

Exercise. B2: A density matrix split into pure pieces

Let

$$\mathbf{X} = \begin{pmatrix} 1/2 & 1/4 \\ 1/4 & 1/2 \end{pmatrix}, \qquad \mathbf{Y} = \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}, \qquad \mathbf{Z} = \begin{pmatrix} 1/2 & 1 \\ 1 & 1/2 \end{pmatrix}.$$

(a) For each of the three matrices, decide whether it lies in $\mathcal{D}_2$, and if so whether it is an extreme point.

(b) Write $\mathbf{X}$ as a convex combination of two extreme points of $\mathcal{D}_2$.

Solution.

(a) All three are symmetric with trace 1. For $\mathbf{X}$: $\mathbf{X}(1,1) = \frac{3}{4}(1,1)$ and $\mathbf{X}(1,-1) = \frac{1}{4}(1,-1)$, so its eigenvalues are $\frac{3}{4}$ and $\frac{1}{4}$, both nonnegative. Hence $\mathbf{X} \in \mathcal{D}_2$ (Theorem 13.1.4), and it has two positive eigenvalues, so it is **not** extreme by Theorem 18.8.8. For $\mathbf{Y}$: $\mathbf{Y} = \mathbf{x}\mathbf{x}^\top$ with $\mathbf{x} = \frac{1}{\sqrt{2}}(1,1)$, a unit vector, so $\mathbf{Y} \in \mathcal{D}_2$ and it **is** extreme. For $\mathbf{Z}$: $\det \mathbf{Z} = \frac{1}{4} - 1 = -\frac{3}{4} < 0$, so the product of its eigenvalues is negative and one of them is negative. Hence $\mathbf{Z}$ is not positive semidefinite, and $\mathbf{Z} \notin \mathcal{D}_2$.

(b) With the unit eigenvectors $\mathbf{q}_1 = \frac{1}{\sqrt{2}}(1,1)$ and $\mathbf{q}_2 = \frac{1}{\sqrt{2}}(1,-1)$,

$$\begin{aligned} \mathbf{X} &= \frac{3}{4}\,\mathbf{q}_1\mathbf{q}_1^\top + \frac{1}{4}\,\mathbf{q}_2\mathbf{q}_2^\top \\ &= \frac{3}{4}\cdot\frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} + \frac{1}{4}\cdot\frac{1}{2}\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}. \end{aligned}$$

The entries check: the diagonal is $\frac{3}{8} + \frac{1}{8} = \frac{1}{2}$ and the off-diagonal is $\frac{3}{8} - \frac{1}{8} = \frac{1}{4}$. The weights $\frac{3}{4}, \frac{1}{4}$ are nonnegative and add up to 1, and both matrices are extreme points by Theorem 18.8.8.

Exercise. B3: A doubly stochastic matrix from permutations

Let

$$\mathbf{M} = \frac{1}{6}\begin{pmatrix} 3 & 2 & 1 \\ 1 & 3 & 2 \\ 2 & 1 & 3 \end{pmatrix}.$$

(a) Check that $\mathbf{M} \in \Omega_3$.

(b) Write $\mathbf{M}$ as a convex combination of three permutation matrices. Hence explain why $\mathbf{M}$ is not an extreme point of Ω_3.

Solution.

(a) The entries are nonnegative. Each row of $6\mathbf{M}$ is a rearrangement of $3, 2, 1$, and so is each column (the columns are $(3,1,2)$, $(2,3,1)$, $(1,2,3)$), so every row and column of $\mathbf{M}$ adds up to $6/6 = 1$.

(b) The entries equal to 3 sit on the diagonal, the entries equal to 2 in positions $(1,2), (2,3), (3,1)$, and the entries equal to 1 in positions $(1,3), (2,1), (3,2)$. Each of these three patterns has exactly one position in each row and each column, so each

is a permutation matrix:

$$\mathbf{P} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}, \qquad \mathbf{P}^2 = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix},$$

and $\mathbf{I}$. Hence

$$\mathbf{M} = \frac{1}{2}\mathbf{I} + \frac{1}{3}\mathbf{P} + \frac{1}{6}\mathbf{P}^2,$$

with weights nonnegative and adding up to 1. Since $\mathbf{M} = \frac{1}{2}\mathbf{I} + \frac{1}{2}\left(\frac{2}{3}\mathbf{P} + \frac{1}{3}\mathbf{P}^2\right)$, with $\frac{2}{3}\mathbf{P} + \frac{1}{3}\mathbf{P}^2 \in \Omega_3$ by convexity and different from $\mathbf{I}$ (its diagonal is zero), $\mathbf{M}$ is the midpoint of two distinct points of Ω_3 and is not extreme.

18.8.8.3 C. Going deeper

Exercise. C1: Deleting an extreme point

Let $C \subseteq V$ be convex and $x \in C$.

(a) Prove that x is an extreme point of C if and only if $C \setminus \{x\}$ is convex.

(b) Deduce that x is an extreme point of C if and only if $x \notin \operatorname{conv}(C \setminus \{x\})$.

Solution.

(a) ($\Rightarrow$) Let $y, z \in C \setminus \{x\}$ and $0 \le t \le 1$. The point $w = ty + (1-t)z$ lies in C by convexity. If $t = 0$ or $t = 1$, w is z or y, which is not x. If $0 < t < 1$ and $w = x$, extremality would give $y = x$, which is false. So $w \neq x$, and $w \in C \setminus \{x\}$.
($\Leftarrow$) Suppose $x = ty + (1-t)z$ with $y, z \in C$, $0 < t < 1$, and not both equal to x. If exactly one of them, say z, equals x, then $x = ty + (1-t)x$, so $t(y-x) = 0$ and $y = x$, a contradiction. So neither equals x, both lie in $C \setminus \{x\}$, and their combination x does not: $C \setminus \{x\}$ is not convex.

(b) If x is extreme, $C \setminus \{x\}$ is convex by (a), so it is its own hull (Definition 18.1.4), which does not contain x. If x is not extreme, (a) gives $y, z \in C \setminus \{x\}$ and $t \in [0,1]$ with $ty + (1-t)z \notin C \setminus \{x\}$; this point lies in C, so it is x, and $x \in \operatorname{conv}(C \setminus \{x\})$.

Exercise. C2: Every contraction is an average of orthogonal matrices

Let $\mathcal{B} = \{\mathbf{A} \in M_n(\mathbb{R}) : \|\mathbf{A}\|_2 \le 1\}$, the unit ball of the operator 2-norm (Definition 16.3.2), so that $\mathbf{A} \in \mathcal{B}$ exactly when $\|\mathbf{A}\mathbf{x}\| \le \|\mathbf{x}\|$ for every $\mathbf{x} \in \mathbb{R}^n$.

(a) Prove that every orthogonal matrix is an extreme point of $\mathcal{B}$.

(b) Prove that every $\mathbf{A} \in \mathcal{B}$ that is not orthogonal is **not** an extreme point of $\mathcal{B}$.

(c) Deduce that every $\mathbf{A} \in \mathcal{B}$ is a convex combination of orthogonal matrices.

Hint for (b): use the singular value decomposition (Theorem 13.9.1) and change only the smallest singular value.

Solution.

(a) Let $\mathbf{Q}$ be orthogonal and $\mathbf{Q} = \frac{1}{2}(\mathbf{B}+\mathbf{C})$ with $\mathbf{B},\mathbf{C} \in \mathcal{B}$. Let $\mathbf{x}$ be a unit vector. Then $\mathbf{Qx}$ is a unit vector, $\mathbf{Bx}$ and $\mathbf{Cx}$ lie in the Euclidean unit ball B_2, and $\mathbf{Qx} = \frac{1}{2}(\mathbf{Bx}+\mathbf{Cx})$. By Example (The diamond, the cube, and the sphere) (c), $\mathbf{Qx}$ is an extreme point of B_2, so $\mathbf{Bx} = \mathbf{Cx}$ by Lemma 18.8.1. This holds for every unit $\mathbf{x}$, hence for every $\mathbf{x}$ by scaling, so $\mathbf{B} = \mathbf{C}$. By Lemma 18.8.1 again, $\mathbf{Q}$ is extreme.

(b) By Theorem 13.9.1, $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\top}$ with $\mathbf{U},\mathbf{V}$ orthogonal and $\mathbf{\Sigma} = \operatorname{diag}(\sigma_1,\dots,\sigma_n)$, $\sigma_1 \geq \cdots \geq \sigma_n \geq 0$, and $\sigma_1 = \|\mathbf{A}\|_2 \leq 1$ by Theorem 16.3.3 (c). If $\sigma_n = 1$, then every $\sigma_i = 1$ and $\mathbf{A} = \mathbf{U}\mathbf{V}^{\top}$ is orthogonal. So $\sigma_n < 1$. Put $\varepsilon = 1 - \sigma_n > 0$, let $\mathbf{u}_n, \mathbf{v}_n$ be the last columns of $\mathbf{U},\mathbf{V}$, and $\mathbf{A}_{\pm} = \mathbf{A} \pm \varepsilon\,\mathbf{u}_n\mathbf{v}_n^{\top} = \mathbf{U}\mathbf{D}_{\pm}\mathbf{V}^{\top}$ with $\mathbf{D}_{\pm} = \operatorname{diag}(\sigma_1,\dots,\sigma_{n-1},\sigma_n \pm \varepsilon)$. The last diagonal entries are 1 and $2\sigma_n - 1 \in [-1,1)$, so every diagonal entry d_i of $\mathbf{D}_{\pm}$ has $|d_i| \leq 1$. Since $\mathbf{U}$ and $\mathbf{V}^{\top}$ preserve Euclidean length,

$$\|\mathbf{A}_{\pm}\mathbf{x}\| = \|\mathbf{D}_{\pm}\mathbf{V}^{\top}\mathbf{x}\| \leq \|\mathbf{V}^{\top}\mathbf{x}\| = \|\mathbf{x}\|,$$

the inequality because $\mathbf{D}_{\pm}$ multiplies each coordinate by a number of absolute value at most 1. So $\mathbf{A}_{\pm} \in \mathcal{B}$, $\mathbf{A} = \frac{1}{2}(\mathbf{A}_+ + \mathbf{A}_-)$, and $\mathbf{A}_+ \neq \mathbf{A}_-$ because $\mathbf{u}_n\mathbf{v}_n^{\top} \neq \mathbf{O}$. Hence $\mathbf{A}$ is not extreme.

(c) $\mathcal{B}$ is convex, being the unit ball of a norm (the remark after Definition 16.1.2). It is closed and bounded for $\|\cdot\|_2$: if $\|\mathbf{A}_j - \mathbf{A}\|_2 \to 0$ with $\|\mathbf{A}_j\|_2 \leq 1$, then $\|\mathbf{A}\|_2 \leq \|\mathbf{A}_j\|_2 + \|\mathbf{A} - \mathbf{A}_j\|_2 \leq 1 + \|\mathbf{A} - \mathbf{A}_j\|_2$ for every j, and letting $j \to \infty$ gives $\|\mathbf{A}\|_2 \leq 1$. By Corollary 16.2.5 it is compact for the Frobenius norm as well. By (a) and (b), $\operatorname{ext}\mathcal{B}$ is the set of orthogonal matrices, and Theorem 18.8.5 gives $\mathcal{B} = \operatorname{conv}(\mathrm{O}(n))$.

Exercise. C3: The corners of the Loewner interval

Let $\mathcal{E} = \{\mathbf{X} \in M_n(\mathbb{R}) : \mathbf{X}^{\top} = \mathbf{X},\ \mathbf{O} \preceq \mathbf{X} \preceq \mathbf{I}\}$, where $\mathbf{X} \preceq \mathbf{I}$ means $\mathbf{I} - \mathbf{X} \succeq \mathbf{O}$ (Definition 13.5.1).

(a) Prove that every symmetric $\mathbf{P}$ with $\mathbf{P}^2 = \mathbf{P}$, that is, every orthogonal projection, is an extreme point of $\mathcal{E}$.

(b) Prove that no other point of $\mathcal{E}$ is extreme.

(c) Deduce that every $\mathbf{X} \in \mathcal{E}$ is a convex combination of orthogonal projections, and give such a combination for $\mathbf{X} = \operatorname{diag}(\frac{1}{2}, \frac{1}{4})$.

Hint for (a): use Lemma 18.8.7 twice, once on $\ker\mathbf{P}$ *and once on* $\operatorname{im}\mathbf{P}$.

Solution.

(a) First, $\mathbf{P} \in \mathcal{E}$: its eigenvalues satisfy $\lambda^2 = \lambda$, so each is 0 or 1, and $\mathbf{P}$ and $\mathbf{I} - \mathbf{P}$ have eigenvalues in $\{0,1\}$, hence are positive semidefinite (Theorem 13.1.4). Let $\mathbf{P} = \frac{1}{2}(\mathbf{A} + \mathbf{B})$ with $\mathbf{A},\mathbf{B} \in \mathcal{E}$. If $\mathbf{Px} = \mathbf{0}$, then $0 = \mathbf{x}^{\top}\mathbf{Px} = \frac{1}{2}\mathbf{x}^{\top}\mathbf{Ax} + \frac{1}{2}\mathbf{x}^{\top}\mathbf{Bx}$ with both terms ≥ 0, so $\mathbf{x}^{\top}\mathbf{Ax} = 0$ and $\mathbf{Ax} = \mathbf{0}$ by Lemma 18.8.7. If $\mathbf{Px} = \mathbf{x}$, then $0 = \mathbf{x}^{\top}(\mathbf{I} - \mathbf{P})\mathbf{x} = \frac{1}{2}\mathbf{x}^{\top}(\mathbf{I} - \mathbf{A})\mathbf{x} + \frac{1}{2}\mathbf{x}^{\top}(\mathbf{I} - \mathbf{B})\mathbf{x}$, so $\mathbf{x}^{\top}(\mathbf{I} - \mathbf{A})\mathbf{x} = 0$ and, since $\mathbf{I} - \mathbf{A} \succeq \mathbf{O}$, $(\mathbf{I} - \mathbf{A})\mathbf{x} = \mathbf{0}$, that is, $\mathbf{Ax} = \mathbf{x}$. Every $\mathbf{v} \in \mathbb{R}^n$ is $(\mathbf{v} - \mathbf{Pv}) + \mathbf{Pv}$, where $\mathbf{P}(\mathbf{v} - \mathbf{Pv}) = \mathbf{0}$ and $\mathbf{P}(\mathbf{Pv}) = \mathbf{Pv}$. So $\mathbf{Av} = \mathbf{0} + \mathbf{Pv} = \mathbf{Pv}$ for every $\mathbf{v}$, and $\mathbf{A} = \mathbf{P}$; then $\mathbf{B} = 2\mathbf{P} - \mathbf{A} = \mathbf{P}$. By Lemma 18.8.1, $\mathbf{P}$ is extreme.

(b) Let $\mathbf{X} \in \mathcal{E}$ not be idempotent. By Corollary 12.5.2, $\mathbf{X} = \sum_i \lambda_i \mathbf{q}_i\mathbf{q}_i^{\top}$ with $\mathbf{q}_i$ orthonor-

mal, and every $\lambda_i \in [0,1]$ because $\mathbf{X} \succeq \mathbf{O}$ and $\mathbf{I} - \mathbf{X} \succeq \mathbf{O}$ have eigenvalues λ_i and $1 - \lambda_i$. If every λ_i were 0 or 1, $\mathbf{X}^2 = \sum_i \lambda_i^2 \mathbf{q}_i \mathbf{q}_i^\top = \mathbf{X}$; so some λ_1, say, lies strictly between 0 and 1. Put $\varepsilon = \min(\lambda_1, 1 - \lambda_1) > 0$ and $\mathbf{X}_\pm = \mathbf{X} \pm \varepsilon \mathbf{q}_1 \mathbf{q}_1^\top$. These have the same eigenvectors, with λ_1 replaced by $\lambda_1 \pm \varepsilon \in [0,1]$, so $\mathbf{X}_\pm \in \mathcal{E}$; they differ, and $\mathbf{X}$ is their midpoint.

(c) $\mathcal{E}$ is convex (both conditions survive convex combinations, as for $\mathcal{D}_n$), closed (each condition survives limits), and bounded (its eigenvalues lie in $[0,1]$, so $\|\mathbf{X}\|_F^2 = \sum_i \lambda_i^2 \le n$). By (a), (b) and Theorem 18.8.5, $\mathcal{E}$ is the convex hull of the orthogonal projections. For $\mathbf{X} = \operatorname{diag}(\frac{1}{2}, \frac{1}{4})$, reading off the eigenvalues in layers,

$$\mathbf{X} = \frac{1}{4}\mathbf{I} + \frac{1}{4}\operatorname{diag}(1,0) + \frac{1}{2}\mathbf{O},$$

a convex combination of the projections $\mathbf{I}$, $\operatorname{diag}(1,0)$, and $\mathbf{O}$ with weights $\frac{1}{4}, \frac{1}{4}, \frac{1}{2}$. The diagonal entries check: $\frac{1}{4} + \frac{1}{4} = \frac{1}{2}$ and $\frac{1}{4}$.

18.9 Polytopes

There are two ways to pin down a convex set with finitely many numbers. One lists finitely many points and takes their convex hull; the other lists finitely many linear inequalities and takes their common solutions. A triangle can be given either way, by its three corners or by its three sides. This section proves that for bounded sets the two descriptions always describe the same sets, which is the Minkowski–Weyl theorem. On the way it locates the optima of Section 6's linear programs over bounded regions: they occur at corners, and a corner of $\{\mathbf{x} : \mathbf{A}\mathbf{x} \le \mathbf{b}\}$ can be recognized by a rank condition on the constraints that hold with equality there. The other direction of the theorem comes from Fourier–Motzkin elimination, an algorithm for projecting a system of inequalities that works the way Gaussian elimination does for equations.

Throughout, the field is $\mathbb{R}$ and the space is $\mathbb{R}^n$ with $n \ge 1$ and the dot product. An inequality between vectors is meant **entrywise**: $\mathbf{A}\mathbf{x} \le \mathbf{b}$ says that $(\mathbf{A}\mathbf{x})_i \le b_i$ for every i. For $\mathbf{A} \in M_{m\times n}(\mathbb{R})$ we write $\mathbf{a}_1^\top, \dots, \mathbf{a}_m^\top$ for its rows, so that $\mathbf{A}\mathbf{x} \le \mathbf{b}$ is the list of m inequalities $\mathbf{a}_i^\top \mathbf{x} \le b_i$.

18.9.1 Two finite descriptions

The first description already appeared in Section 8, just after Proposition 18.8.2.

A polytope is the convex hull of finitely many points; a polyhedron is the solution set of finitely many linear inequalities.

Definition 18.9.1. Polytope

A **polytope** in $\mathbb{R}^n$ is a set of the form $\operatorname{conv}\{\mathbf{v}_1, \dots, \mathbf{v}_k\}$ for **finitely many** points $\mathbf{v}_1, \dots, \mathbf{v}_k \in \mathbb{R}^n$, $k \ge 0$.

Definition 18.9.2. Polyhedron

A **polyhedron** in $\mathbb{R}^n$ is a set of the form

$$P = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{A}\mathbf{x} \le \mathbf{b}\}$$

for some $m \geq 0$, $\mathbf{A} \in M_{m\times n}(\mathbb{R})$, and $\mathbf{b} \in \mathbb{R}^m$; that is, the intersection of **finitely many** sets $\{\mathbf{x} : \mathbf{a}_i^\top \mathbf{x} \leq b_i\}$. An extreme point of a polyhedron or of a polytope (Definition 18.8.1) is called a **vertex**.

In words: a polytope is built from the inside, by mixing finitely many given points; a polyhedron is carved from the outside, by finitely many cuts. When $\mathbf{a}_i \neq \mathbf{0}$, the set $\{\mathbf{x} : \mathbf{a}_i^\top \mathbf{x} \leq b_i\}$ is a closed half-space, bounded by the hyperplane $\mathbf{a}_i^\top \mathbf{x} = b_i$ (Definition 5.5.2). A zero row $\mathbf{a}_i = \mathbf{0}$ is allowed and harmless: the condition $0 \leq b_i$ either holds for every $\mathbf{x}$ or for none. An **equation** $\mathbf{a}^\top \mathbf{x} = \beta$ is allowed too, as the two inequalities $\mathbf{a}^\top \mathbf{x} \leq \beta$ and $-\mathbf{a}^\top \mathbf{x} \leq -\beta$. By Theorem 18.1.3, the polytope $\operatorname{conv}\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ is the set of all $\sum_j t_j \mathbf{v}_j$ with every $t_j \geq 0$ and $\sum_j t_j = 1$.

Some examples, simplest first.

1. **The empty set and a point** are both. $\emptyset = \operatorname{conv} \emptyset$, and $\emptyset = \{\mathbf{x} : 0 \leq -1\}$ with the single zero row. A point $\mathbf{v} = \operatorname{conv}\{\mathbf{v}\}$ is the solution set of the $2n$ inequalities $x_i \leq v_i$ and $-x_i \leq -v_i$.
2. **All of** $\mathbb{R}^n$ is a polyhedron, with $m = 0$ (or with the one row $0 \leq 0$). So is a closed half-space. Neither is a polytope, because a polytope is bounded: if $\|\mathbf{v}_j\| \leq R$ for every j, then $\|\sum_j t_j \mathbf{v}_j\| \leq \sum_j t_j \|\mathbf{v}_j\| \leq R$.
3. **The triangle** $T \subseteq \mathbb{R}^2$ with corners $(0,0), (3,0), (0,3)$ is $\operatorname{conv}\{(0,0), (3,0), (0,3)\}$, and it is the polyhedron $\{x_1 \geq 0,\ x_2 \geq 0,\ x_1 + x_2 \leq 3\}$. That the two descriptions agree is not a definition but a fact; the worked example in the section on Fourier–Motzkin elimination below derives each from the other.
4. **The cube** $B_\infty = [-1, 1]^n$ is the polyhedron given by the $2n$ inequalities $\pm x_i \leq 1$, and by Section 8 it is the hull of its 2^n sign vectors (Example (The diamond, the cube, and the sphere) and Theorem 18.8.5). So it is both.

A nonexample by minimal change. Replace the cube's square $B_\infty \subseteq \mathbb{R}^2$ by the disk B_2. It is still compact and convex, still the unit ball of a norm, and still the solution set of inequalities $\mathbf{a}^\top \mathbf{x} \leq 1$, one for **every** unit vector $\mathbf{a}$: by Theorem 11.1.5, $\mathbf{a}^\top \mathbf{x} \leq \|\mathbf{x}\| \leq 1$ on the disk, and a point with $\|\mathbf{x}\| > 1$ violates the inequality for $\mathbf{a} = \mathbf{x}/\|\mathbf{x}\|$. What fails is the word **finitely**. The disk is not a polytope: by Proposition 18.8.2 a polytope has finitely many extreme points, while the disk has a whole circle of them (Example (The diamond, the cube, and the sphere)). We will see after the Minkowski–Weyl theorem that it is not a polyhedron either.

Why these definitions. Both allow only finite data, and that is the point: finitely many inequalities can be checked, finitely many points can be listed, and the whole theory of this section is about moving between two finite lists. An infinite family of inequalities describes every closed convex set, by the separation theorem (Theorem 18.3.2), so dropping "finitely many" would make the second definition say nothing new.

Section 1 already showed that every polyhedron is convex and closed; we record it for citation.

Proposition 18.9.1. Polyhedra Are Closed and Convex

Every polyhedron $P = \{\mathbf{x} : \mathbf{A}\mathbf{x} \leq \mathbf{b}\}$ is convex and closed.

Proof. With the linear functional $\varphi_i(\mathbf{x}) = \mathbf{a}_i^\top \mathbf{x}$, each set $\{\mathbf{x} : \mathbf{a}_i^\top \mathbf{x} \leq b_i\} = \{\varphi_i \leq b_i\}$ is closed by Proposition 18.1.6 (c), and convex by Proposition 18.1.1 (b), as the preimage of the interval $(-\infty, b_i]$ under φ_i. So their intersection P is closed by Proposition 18.1.6 (b) and convex by Proposition 18.1.1 (a).

■

Warning. **A polyhedron need not have any vertices, even when it is not empty.** The strip $\{\mathbf{x} \in \mathbb{R}^2 : -1 \le x_1 + x_2 \le 1\}$ is a closed convex set with no extreme points at all: every point $\mathbf{x}$ of it is the midpoint of $\mathbf{x} \pm (1, -1)$, which lie in it. Minkowski's theorem (Theorem 18.8.5) does not apply because the strip is unbounded, and in the language of the next theorem, its constraint rows $\pm(1, 1)$ have rank $1 < 2$.

18.9.2 Vertices are basic feasible points

Section 8 showed that a linear function on a nonempty compact convex set attains its maximum at an extreme point (Corollary 18.8.6). For a linear program over a bounded region $\{\mathbf{x} : \mathbf{A}\mathbf{x} \le \mathbf{b}\}$, that shrinks the search to the vertices. To make it a finite search we need a way to recognize a vertex from the data $\mathbf{A}, \mathbf{b}$, and a proof that there are finitely many.

Stand at a point $\mathbf{x}$ of P and look at the inequalities that are tight there. Each tight inequality $\mathbf{a}_i^\top \mathbf{x} = b_i$ confines $\mathbf{x}$ to a hyperplane. If the tight rows are rich enough to pin $\mathbf{x}$ down, there is no room to move in any direction, and $\mathbf{x}$ should be a corner. If not, there should be a direction along which one can move both ways without leaving P.

Definition 18.9.3. Active Set, Basic Feasible Point

Let $P = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{A}\mathbf{x} \le \mathbf{b}\}$ and $\mathbf{x} \in P$. The **active set** of $\mathbf{x}$ is $I(\mathbf{x}) = \{i : \mathbf{a}_i^\top \mathbf{x} = b_i\}$, the indices of the inequalities that hold with equality at $\mathbf{x}$. The point $\mathbf{x}$ is a **basic feasible point** of the system if among the active rows $\{\mathbf{a}_i : i \in I(\mathbf{x})\}$ there are n **linearly independent** ones.

Equivalently, the matrix $\mathbf{A}_{I(\mathbf{x})}$ formed by the active rows has a row space of dimension n, that is, rank n (Theorem 3.5.5); by Theorem 3.5.7 this says that its null space is $\{\mathbf{0}\}$, which means that $\mathbf{x}$ is the **only** solution of the equations $\mathbf{a}_i^\top \mathbf{y} = b_i$, $i \in I(\mathbf{x})$. The word "feasible" records that $\mathbf{x}$ lies in P; a solution of n independent equations $\mathbf{a}_i^\top \mathbf{y} = b_i$ that violates some other inequality is a **basic point** but not a basic feasible one. Note that the definition depends on the system $\mathbf{A}\mathbf{x} \le \mathbf{b}$, not only on the set P; the theorem shows that the resulting notion depends only on P.

Theorem 18.9.2. Vertices Are Basic Feasible Points

Let $P = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{A}\mathbf{x} \le \mathbf{b}\}$ and $\mathbf{x} \in P$. Then $\mathbf{x}$ is a vertex of P if and only if $\mathbf{x}$ is a basic feasible point.

Idea. Both directions are about the directions $\mathbf{u}$ with $\mathbf{a}_i^\top \mathbf{u} = 0$ for every active i, the directions along which every tight constraint stays tight. If the active rows have rank n, the only such direction is $\mathbf{0}$, and an averaging argument pins down any segment through $\mathbf{x}$. If their rank is less than n, a nonzero such $\mathbf{u}$ exists by rank–nullity, and a short step $\pm\varepsilon\mathbf{u}$ keeps the tight constraints tight and, because there are finitely many, keeps the slack ones slack.

Proof. Write $I = I(\mathbf{x})$ and let $\mathbf{A}_I$ be the matrix of active rows (empty if $I = \emptyset$).

($\Leftarrow$) Suppose $\mathbf{A}_I$ has rank n, and let $\mathbf{x} = \frac{1}{2}(\mathbf{y} + \mathbf{z})$ with $\mathbf{y}, \mathbf{z} \in P$. For $i \in I$,

$$b_i = \mathbf{a}_i^\top \mathbf{x} = \tfrac{1}{2}\mathbf{a}_i^\top \mathbf{y} + \tfrac{1}{2}\mathbf{a}_i^\top \mathbf{z},$$

with $\mathbf{a}_i^\top \mathbf{y} \le b_i$ and $\mathbf{a}_i^\top \mathbf{z} \le b_i$; if either were strict, the average would be less than b_i. So $\mathbf{a}_i^\top \mathbf{y} = b_i = \mathbf{a}_i^\top \mathbf{x}$, that is, $\mathbf{A}_I(\mathbf{y} - \mathbf{x}) = \mathbf{0}$. Since $\mathbf{A}_I$ has n columns and rank n, its null space is $\{\mathbf{0}\}$ by Theorem 3.5.7, so $\mathbf{y} = \mathbf{x}$, and then $\mathbf{z} = 2\mathbf{x} - \mathbf{y} = \mathbf{x}$. By Lemma 18.8.1, $\mathbf{x}$ is a vertex.

($\Rightarrow$) Suppose $\mathbf{A}_I$ has rank less than n (this includes $I = \emptyset$). By Theorem 3.5.7 its null space is nonzero; choose $\mathbf{u} \neq \mathbf{0}$ with $\mathbf{a}_i^\top \mathbf{u} = 0$ for every $i \in I$. For $i \notin I$, the slack $b_i - \mathbf{a}_i^\top \mathbf{x}$ is positive. Let $\varepsilon > 0$ be smaller than $(b_i - \mathbf{a}_i^\top \mathbf{x})/|\mathbf{a}_i^\top \mathbf{u}|$ for every $i \notin I$ with $\mathbf{a}_i^\top \mathbf{u} \neq 0$; there are finitely many such i, so such an ε exists, and if there are none, take $\varepsilon = 1$. Then for $i \in I$, $\mathbf{a}_i^\top(\mathbf{x} \pm \varepsilon\mathbf{u}) = b_i$, and for $i \notin I$,

$$\mathbf{a}_i^\top(\mathbf{x} \pm \varepsilon\mathbf{u}) \le \mathbf{a}_i^\top \mathbf{x} + \varepsilon|\mathbf{a}_i^\top \mathbf{u}| \le b_i.$$

So $\mathbf{x} \pm \varepsilon\mathbf{u} \in P$, they differ because $\mathbf{u} \neq \mathbf{0}$, and $\mathbf{x}$ is their midpoint. Hence $\mathbf{x}$ is not a vertex. This proves the theorem.

■

The payoff is finiteness.

Corollary 18.9.3. A Polyhedron Has Finitely Many Vertices

Let $P = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{A}\mathbf{x} \le \mathbf{b}\}$ with $\mathbf{A} \in M_{m\times n}(\mathbb{R})$. Every vertex of P is the unique solution of $\mathbf{A}_J\mathbf{x} = \mathbf{b}_J$ for some set J of n row indices with $\mathbf{A}_J$ invertible. In particular P has at most $\binom{m}{n}$ vertices.

Idea. By the theorem, n independent active rows pin a vertex down, so a vertex is determined by which n rows those are.

Proof. Let $\mathbf{x}$ be a vertex. By Theorem 18.9.2, some n active rows, with index set $J \subseteq I(\mathbf{x})$, are linearly independent. The $n \times n$ matrix $\mathbf{A}_J$ then has independent rows, so its rank is n (Theorem 3.5.5), its null space is $\{\mathbf{0}\}$ (Theorem 3.5.7), and it is invertible (Theorem 3.4.4 (b)). As $J \subseteq I(\mathbf{x})$, $\mathbf{A}_J\mathbf{x} = \mathbf{b}_J$, so $\mathbf{x} = \mathbf{A}_J^{-1}\mathbf{b}_J$ is determined by J. Different vertices therefore come from different sets J, and there are $\binom{m}{n}$ sets of n indices.

■

This gives the finite recipe for a vertex list: for each choice of n rows with $\mathbf{A}_J$ invertible, solve $\mathbf{A}_J\mathbf{x} = \mathbf{b}_J$ and keep the solution if it satisfies the remaining inequalities.

Example. The vertices of a quadrilateral

Find the vertices of

$$P = \{\mathbf{x} \in \mathbb{R}^2 : x_1 \geq 0,\ x_2 \geq 0,\ x_1 + 2x_2 \leq 4,\ 3x_1 + x_2 \leq 6\},$$

and the maximum of $\varphi(\mathbf{x}) = x_1 + x_2$ over P.

Solution. Number the constraints (1) $-x_1 \leq 0$, (2) $-x_2 \leq 0$, (3) $x_1+2x_2 \leq 4$, (4) $3x_1+x_2 \leq 6$. Any two of the four rows are linearly independent, so each of the $\binom{4}{2} = 6$ pairs gives one basic point. We solve each pair as equations and test the other two inequalities.

pair	basic point	other constraints
(1), (2)	$(0,0)$	$0 \leq 4,\ 0 \leq 6$: feasible
(1), (3)	$(0,2)$	$-2 \leq 0,\ 2 \leq 6$: feasible
(1), (4)	$(0,6)$	$12 \leq 4$ fails
(2), (3)	$(4,0)$	$12 \leq 6$ fails
(2), (4)	$(2,0)$	$-2 \leq 0,\ 2 \leq 4$: feasible
(3), (4)	$(\frac{8}{5}, \frac{6}{5})$	$-\frac{8}{5} \leq 0,\ -\frac{6}{5} \leq 0$: feasible

For the last pair, subtracting (3) from twice (4) gives $5x_1 = 8$, so $x_1 = \frac{8}{5}$ and $x_2 = 6-3x_1 = \frac{6}{5}$; check $\frac{8}{5} + \frac{12}{5} = 4$. By Theorem 18.9.2 the vertices are the four feasible basic points $(0,0)$, $(0,2)$, $(2,0)$, and $(\frac{8}{5}, \frac{6}{5})$, and P is a quadrilateral. The two infeasible basic points are where two of the lines cross outside P.

P is bounded, since $0 \leq x_1 \leq 2$ and $0 \leq x_2 \leq 2$ on it, so it is compact (Proposition 18.9.1, Corollary 16.2.5) and nonempty. By Corollary 18.8.6 the maximum of φ is attained at a vertex. The values are $0, 2, 2, \frac{14}{5}$, so the maximum is $\frac{14}{5}$, at $(\frac{8}{5}, \frac{6}{5})$.

Quick check. A polyhedron in $\mathbb{R}^2$ is given by $m = 5$ inequalities. What is the largest number of vertices the corollary allows? Could a polyhedron in $\mathbb{R}^3$ given by $m = 2$ inequalities have a vertex?

Solution. At most $\binom{5}{2} = 10$. In $\mathbb{R}^3$ a vertex needs 3 linearly independent active rows, and there are only 2 rows in all, so no: the rank of the active rows is at most $2 < 3$, and by Theorem 18.9.2 such a polyhedron has no vertices. Geometrically, the intersection of two half-spaces in $\mathbb{R}^3$ always contains a whole line through each of its points.

The recipe is honest but slow: $\binom{m}{n}$ grows very fast. The **simplex method**, the classical algorithm for linear programs, uses the same theorem more cleverly. It moves from a vertex to a neighboring vertex, one sharing $n-1$ of its independent active constraints, along an edge on which the objective improves, and stops at a vertex where no edge improves it.

18.9.3 Fourier–Motzkin elimination

Now the other direction. Given a polytope $\operatorname{conv}\{\mathbf{v}_1, \dots, \mathbf{v}_k\}$, how do we find inequalities for it? The hull is the set of $\mathbf{x}$ for which **there exist** weights $\mathbf{t} \geq \mathbf{0}$ with $\sum_j t_j = 1$ and $\mathbf{x} = \sum_j t_j \mathbf{v}_j$. The conditions on the pair $(\mathbf{x}, \mathbf{t})$ are linear inequalities, so the set of pairs is a polyhedron in

$\mathbb{R}^{n+k}$, and the polytope is its image under the projection that forgets $\mathbf{t}$. So we need to know that a projection of a polyhedron is a polyhedron, and a way to compute it. For equations, eliminating a variable is Gaussian elimination. For inequalities the device is to pair every upper bound on the variable with every lower bound.

Write a point of $\mathbb{R}^{n+1}$ as $(\mathbf{x}, s)$ with $\mathbf{x} \in \mathbb{R}^n$ and $s \in \mathbb{R}$, and let $\pi(\mathbf{x}, s) = \mathbf{x}$.

Lemma 18.9.4. Fourier–Motzkin Elimination

Let $P = \{(\mathbf{x}, s) \in \mathbb{R}^n \times \mathbb{R} : \mathbf{a}_i^\top \mathbf{x} + c_i s \le b_i \text{ for } i = 1, \dots, m\}$. Split the indices by the sign of the coefficient of s: $I_+ = \{i : c_i > 0\}$, $I_- = \{i : c_i < 0\}$, $I_0 = \{i : c_i = 0\}$. Then $\pi(P)$ is the set of $\mathbf{x} \in \mathbb{R}^n$ satisfying

$$\begin{aligned} \mathbf{a}_i^\top \mathbf{x} &\le b_i && \text{for } i \in I_0, \\ (c_i \mathbf{a}_j - c_j \mathbf{a}_i)^\top \mathbf{x} &\le c_i b_j - c_j b_i && \text{for } i \in I_+,\ j \in I_-. \end{aligned}$$

In particular $\pi(P)$ is a polyhedron, and if every $b_i = 0$, every right-hand side above is 0.

Idea. For fixed $\mathbf{x}$, each row with $c_i > 0$ is an upper bound on s, each row with $c_i < 0$ is a lower bound, and a row with $c_i = 0$ does not mention s. A number s satisfying all of them exists exactly when every lower bound is at most every upper bound, and those comparisons are the new rows. Each new row is the combination $c_i \cdot (\text{row } j) + (-c_j) \cdot (\text{row } i)$, with both multipliers positive and chosen to cancel s.

Proof. Let Q be the set described by the new rows. For $i \in I_+$ and $j \in I_-$, put

$$U_i(\mathbf{x}) = \frac{b_i - \mathbf{a}_i^\top \mathbf{x}}{c_i}, \qquad L_j(\mathbf{x}) = \frac{b_j - \mathbf{a}_j^\top \mathbf{x}}{c_j}.$$

Dividing row i by $c_i > 0$ shows that it says $s \le U_i(\mathbf{x})$, and dividing row j by $c_j < 0$, which reverses the inequality, shows that it says $s \ge L_j(\mathbf{x})$. Multiplying $L_j(\mathbf{x}) \le U_i(\mathbf{x})$ by the positive number $-c_i c_j$ and rearranging gives exactly the new row for the pair (i, j):

$$L_j(\mathbf{x}) \le U_i(\mathbf{x}) \iff (c_i \mathbf{a}_j - c_j \mathbf{a}_i)^\top \mathbf{x} \le c_i b_j - c_j b_i.$$

($\subseteq$) If $(\mathbf{x}, s) \in P$, then $L_j(\mathbf{x}) \le s \le U_i(\mathbf{x})$ for every pair, and the rows in I_0 hold since they do not involve s. So $\mathbf{x} \in Q$.

($\supseteq$) Let $\mathbf{x} \in Q$. If $I_- \neq \emptyset$, let $s = \max_{j \in I_-} L_j(\mathbf{x})$; if $I_- = \emptyset$ and $I_+ \neq \emptyset$, let $s = \min_{i \in I_+} U_i(\mathbf{x})$; if both are empty, let $s = 0$. In each case $L_j(\mathbf{x}) \le s \le U_i(\mathbf{x})$ for all $j \in I_-$ and $i \in I_+$: in the first case because every $L_j(\mathbf{x}) \le s$ by the choice of s, and $s = L_{j_0}(\mathbf{x}) \le U_i(\mathbf{x})$ for every i since $\mathbf{x} \in Q$; the second case is the same with the roles exchanged. The rows in I_0 hold because $\mathbf{x} \in Q$. So $(\mathbf{x}, s) \in P$ and $\mathbf{x} \in \pi(P)$.

The last sentence is visible: the new right-hand sides are b_i and $c_i b_j - c_j b_i$.

■

As an algorithm:

Algorithm. **Fourier–Motzkin elimination of one variable.** Input: a system of m inequalities $\mathbf{a}_i^{\top}\mathbf{x} + c_i s \le b_i$ in the unknowns $\mathbf{x} \in \mathbb{R}^n$ and $s \in \mathbb{R}$.

1. Sort the rows into I_+ ($c_i > 0$), I_- ($c_i < 0$) and I_0 ($c_i = 0$).
2. Copy every row of I_0, without its s term.
3. For every pair $i \in I_+$, $j \in I_-$, write down the row $c_i \cdot (\text{row } j) + (-c_j) \cdot (\text{row } i)$, which is $(c_i\mathbf{a}_j - c_j\mathbf{a}_i)^{\top}\mathbf{x} \le c_i b_j - c_j b_i$.
4. (Optional.) Delete every new row of the form $0 \le \beta$ with $\beta \ge 0$, which every $\mathbf{x}$ satisfies. A row $0 \le \beta$ with $\beta < 0$ means the projection is empty.

Output: $|I_0| + |I_+|\,|I_-|$ inequalities in $\mathbf{x}$ alone, at most, whose solution set is the projection of the input's solution set. To eliminate several variables, repeat.

When the variable to be eliminated appears in an **equation** with a nonzero coefficient, there is a shortcut: solve the equation for it and substitute into the other rows. The projection is unchanged, since the equation determines the eliminated variable from the others, and substitution is exactly Gaussian elimination.

Example. A triangle, both ways

Let $Q = \operatorname{conv}\{(0,0),(3,0),(0,3),(1,1)\} \subseteq \mathbb{R}^2$. Find a system of inequalities describing Q, and then recover the vertices of Q from it.

Solution. From points to inequalities. By Theorem 18.1.3, $\mathbf{x} \in Q$ if and only if there are $t_1, \dots, t_4 \ge 0$ with $\sum_j t_j = 1$ and

$$\mathbf{x} = t_1(0,0) + t_2(3,0) + t_3(0,3) + t_4(1,1),$$

that is,

$$x_1 = 3t_2 + t_4, \qquad x_2 = 3t_3 + t_4.$$

Three equations involve the weights, so we substitute three of them away and keep $s = t_4$: $t_2 = \tfrac{1}{3}(x_1 - s)$, $t_3 = \tfrac{1}{3}(x_2 - s)$, and

$$t_1 = 1 - t_2 - t_3 - s = 1 - \tfrac{1}{3}(x_1 + x_2 + s).$$

The four conditions $t_j \ge 0$, each multiplied by 3 where needed, become a system in (x_1, x_2, s):

$$\begin{aligned} &(1)\quad x_1 + x_2 + s \le 3, && (2)\quad -x_1 + s \le 0,\\ &(3)\quad -x_2 + s \le 0, && (4)\quad -s \le 0. \end{aligned}$$

Run the algorithm on s. The coefficients of s are $1, 1, 1, -1$, so $I_+ = \{1,2,3\}$, $I_- = \{4\}$, and $I_0 = \emptyset$. Each pair $(i,4)$ has $c_i = 1$ and $-c_4 = 1$, so the new row is row i plus row 4, which just deletes the s:

$$x_1 + x_2 \le 3, \qquad -x_1 \le 0, \qquad -x_2 \le 0.$$

By Lemma 18.9.4, $Q = \{x_1 \ge 0,\ x_2 \ge 0,\ x_1 + x_2 \le 3\}$, the triangle T of the examples above. The generator $(1,1)$ left no trace, as it should: it lies inside the triangle.

From inequalities to vertices. The three rows $(-1,0)$, $(0,-1)$, $(1,1)$ are pairwise independent, so each of the $\binom{3}{2} = 3$ pairs gives a basic point: $(0,0)$, $(0,3)$, and $(3,0)$. Each satisfies the

third inequality, so all three are vertices by Theorem 18.9.2. At $(1,1)$ no constraint is active, so $I = \emptyset$ and it is not a vertex, which matches Proposition 18.8.2: the extreme points of the hull are among the four generators, and $(1,1)$ is not one of them.

Warning. **Fourier–Motzkin does not return a minimal description.** Eliminating s from $-x_1 + s \le 0$, $s \le 1$, $-s \le 0$, and $x_1 - s \le 1$ pairs two upper bounds with two lower bounds, and gives $-x_1 \le 0$, $x_1 \le 2$, and twice the useless row $0 \le 1$. Even without trivial rows, the output can contain inequalities implied by the others. A single step can turn m rows into as many as $(m/2)^2$, and eliminating many variables compounds this, so the method is a proof device and a hand tool for small systems, not an efficient algorithm.

18.9.4 The Minkowski–Weyl theorem

Cones get the same treatment. Section 5 wrote $\mathrm{cone}(\mathbf{v}_1, \dots, \mathbf{v}_k)$ for the set of all nonnegative combinations $\sum_j c_j \mathbf{v}_j$, $c_j \ge 0$, a convex cone (Definition 18.5.1), and called such cones **finitely generated**; we allow $k = 0$, with $\mathrm{cone}() = \{\mathbf{0}\}$. The inequality side is a homogeneous system.

Definition 18.9.4. Polyhedral Cone

A **polyhedral cone** in $\mathbb{R}^n$ is a set of the form $C = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{A}\mathbf{x} \le \mathbf{0}\}$ for some $m \ge 0$ and $\mathbf{A} \in M_{m \times n}(\mathbb{R})$.

A polyhedral cone is a polyhedron, and it is a convex cone: it contains $\mathbf{0}$, and if $\mathbf{A}\mathbf{u} \le \mathbf{0}$, $\mathbf{A}\mathbf{v} \le \mathbf{0}$, and $a, b \ge 0$, then $\mathbf{A}(a\mathbf{u} + b\mathbf{v}) = a\mathbf{A}\mathbf{u} + b\mathbf{A}\mathbf{v} \le \mathbf{0}$. By Proposition 18.9.1 it is closed, while for finitely generated cones closedness took the work of Theorem 18.5.4.

Theorem 18.9.5. Minkowski–Weyl Theorem

(a) A subset of $\mathbb{R}^n$ is a polytope if and only if it is a bounded polyhedron.

(b) A subset of $\mathbb{R}^n$ is a finitely generated cone if and only if it is a polyhedral cone.

Idea. Four implications, two tools.

① *Bounded polyhedron ⇒ polytope.* This is Section 8 plus the vertex theorem: a bounded polyhedron is compact, so it is the hull of its vertices by Minkowski's theorem, and it has finitely many vertices.

② *Polyhedral cone ⇒ finitely generated.* Cut the cone with the cube $[-1,1]^n$. The result is a bounded polyhedron, hence by ① the hull of finitely many points, and scaling those points generates the cone.

③ *Polytope ⇒ polyhedron*, and ④ *finitely generated ⇒ polyhedral.* Write the set as the projection of a polyhedron in a bigger space, with the weights as extra coordinates, and eliminate the weights one at a time by Fourier–Motzkin. For a cone all right-hand sides are 0 and stay 0.

Proof. ① **A bounded polyhedron is a polytope.** Let $P = \{\mathbf{x} : \mathbf{A}\mathbf{x} \le \mathbf{b}\}$ be bounded. If $P = \emptyset$, then $P = \operatorname{conv}\emptyset$. Otherwise P is closed and convex by Proposition 18.9.1, and bounded, so it is compact by Corollary 16.2.5. By Theorem 18.8.5, $P = \operatorname{conv}(\operatorname{ext} P)$, and $\operatorname{ext} P$ is finite by Corollary 18.9.3. So P is a polytope.

② **A polyhedral cone is finitely generated.** Let $C = \{\mathbf{x} : \mathbf{A}\mathbf{x} \le \mathbf{0}\}$, and let $P = C \cap [-1,1]^n$, the polyhedron given by the rows of $\mathbf{A}$ together with the $2n$ rows $\pm x_i \le 1$. It is bounded and contains $\mathbf{0}$, so by ① $P = \operatorname{conv}\{\mathbf{v}_1, \dots, \mathbf{v}_k\}$, where $\mathbf{v}_1, \dots, \mathbf{v}_k$ are its vertices; in particular every $\mathbf{v}_j \in P \subseteq C$. We claim $C = \operatorname{cone}(\mathbf{v}_1, \dots, \mathbf{v}_k)$. If $\mathbf{c} \ge \mathbf{0}$, then $\mathbf{A}\sum_j c_j\mathbf{v}_j = \sum_j c_j\mathbf{A}\mathbf{v}_j \le \mathbf{0}$, since each $\mathbf{A}\mathbf{v}_j \le \mathbf{0}$ and each $c_j \ge 0$; so every such combination lies in C. Conversely let $\mathbf{x} \in C$. If $\mathbf{x} = \mathbf{0}$, take all $c_j = 0$. Otherwise put $r = \|\mathbf{x}\|_\infty > 0$. Then $\mathbf{A}(\mathbf{x}/r) = \frac{1}{r}\mathbf{A}\mathbf{x} \le \mathbf{0}$ and every entry of $\mathbf{x}/r$ lies in $[-1,1]$, so $\mathbf{x}/r \in P$, say $\mathbf{x}/r = \sum_j t_j\mathbf{v}_j$ with $t_j \ge 0$. Then $\mathbf{x} = \sum_j (rt_j)\mathbf{v}_j$ with $rt_j \ge 0$.

③ **A polytope is a bounded polyhedron.** Let $Q = \operatorname{conv}\{\mathbf{v}_1, \dots, \mathbf{v}_k\}$, and let $\mathbf{V}$ have columns $\mathbf{v}_j$. It is bounded, as in example 2 after Definition 18.9.2. By Theorem 18.1.3, Q is the image, under the map $(\mathbf{x}, \mathbf{t}) \mapsto \mathbf{x}$ that forgets the last k coordinates, of

$$\widehat{P} = \{(\mathbf{x}, \mathbf{t}) \in \mathbb{R}^n \times \mathbb{R}^k : \mathbf{x} - \mathbf{V}\mathbf{t} \le \mathbf{0}, \quad -\mathbf{x} + \mathbf{V}\mathbf{t} \le \mathbf{0}, \\ -\mathbf{t} \le \mathbf{0}, \quad \mathbf{1}^\top\mathbf{t} \le 1, \quad -\mathbf{1}^\top\mathbf{t} \le -1\},$$

a polyhedron in $\mathbb{R}^{n+k}$. Forgetting the last k coordinates is forgetting one coordinate k times in a row, and by Lemma 18.9.4 each step takes a polyhedron to a polyhedron. So Q is a polyhedron.

④ **A finitely generated cone is polyhedral.** Let $C = \operatorname{cone}(\mathbf{v}_1, \dots, \mathbf{v}_k) = \{\mathbf{V}\mathbf{c} : \mathbf{c} \ge \mathbf{0}\}$, where $\mathbf{V}$ has columns $\mathbf{v}_j$. It is the image under the same map of $\{(\mathbf{x}, \mathbf{c}) : \mathbf{x} - \mathbf{V}\mathbf{c} \le \mathbf{0}, \ -\mathbf{x} + \mathbf{V}\mathbf{c} \le \mathbf{0}, \ -\mathbf{c} \le \mathbf{0}\}$, whose right-hand sides are all 0. By the last sentence of Lemma 18.9.4, eliminating the k coordinates of $\mathbf{c}$ one at a time keeps every right-hand side 0, and the rows copied from I_0 keep theirs as well. So $C = \{\mathbf{x} : \mathbf{B}\mathbf{x} \le \mathbf{0}\}$ for the matrix $\mathbf{B}$ of the final rows (if no rows remain, $C = \mathbb{R}^n = \{\mathbf{x} : 0 \le 0\}$). This completes the proof. ■

Two consequences show what the theorem is for. First, **the intersection of two polytopes is a polytope**. By (a) each is a bounded polyhedron; putting the two systems of inequalities together describes the intersection, which is therefore a bounded polyhedron, hence a polytope by (a) again. From the point description alone this is not obvious at all: the corners of the intersection are, in general, new points found nowhere in the two lists. Second, (b) gives a new proof of Theorem 18.5.4, since every polyhedral cone is closed. In that proof the difficulty of Section 5 has moved into Fourier–Motzkin elimination. The same elimination shows that a linear image $\mathbf{M}(C)$ of a polyhedral cone $C = \{\mathbf{x} : \mathbf{A}\mathbf{x} \le \mathbf{0}\}$ is polyhedral: it is the image, under the map forgetting $\mathbf{x}$, of $\{(\mathbf{y}, \mathbf{x}) : \mathbf{y} - \mathbf{M}\mathbf{x} \le \mathbf{0}, \ -\mathbf{y} + \mathbf{M}\mathbf{x} \le \mathbf{0}, \ \mathbf{A}\mathbf{x} \le \mathbf{0}\}$, and the argument of ④ applies. So a linear image of a polyhedral cone is closed, and a closed convex cone with a nonclosed linear image, like the one in Section 5, can never be polyhedral.

The disk B_2 is not a polyhedron. If it were, it would be a bounded polyhedron, hence a polytope by (a), and it is not one.

18.9.5 The cube and the cross-polytope

The two unit balls with corners from Section 8 are the standard example of the theorem, and they illustrate a duality between the two descriptions.

Example. The cube and the cross-polytope

For $n \geq 1$, describe B_∞ and B_1 both as polytopes and as polyhedra, and find their vertices from the inequalities by Theorem 18.9.2.

Solution. The cube. B_∞ is the polyhedron given by the $2n$ rows $\mathbf{e}_i^\top \mathbf{x} \leq 1$ and $-\mathbf{e}_i^\top \mathbf{x} \leq 1$. At a point $\mathbf{x}$ of it, the active rows are $\mathbf{e}_i$ for each i with $x_i = 1$ and $-\mathbf{e}_i$ for each i with $x_i = -1$; they are distinct standard vectors up to sign, so they are linearly independent, and their number is the rank. The rank is n exactly when every $|x_i| = 1$. So the vertices are the 2^n sign vectors, as Example (The diamond, the cube, and the sphere) found directly, and B_∞ is their hull.

The cross-polytope. By Example (The diamond, the cube, and the sphere) and Theorem 18.8.5, $B_1 = \operatorname{conv}\{\pm\mathbf{e}_1, \dots, \pm\mathbf{e}_n\}$. For its inequalities, let $\mathbf{s}$ run over the 2^n sign vectors. For every $\mathbf{s}$, $\mathbf{s}^\top \mathbf{x} = \sum_i s_i x_i \leq \sum_i |x_i| = \|\mathbf{x}\|_1$, with equality when $s_i = 1$ for $x_i \geq 0$ and $s_i = -1$ for $x_i < 0$. So $\|\mathbf{x}\|_1 = \max_{\mathbf{s}} \mathbf{s}^\top \mathbf{x}$, and

$$B_1 = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{s}^\top \mathbf{x} \leq 1 \text{ for every sign vector } \mathbf{s}\},$$

a polyhedron with 2^n rows. At $\mathbf{e}_1$, the active rows are the $\mathbf{s}$ with $s_1 = 1$. They have rank n: they include $\mathbf{1}$, and for each $j \geq 2$ the vector $\mathbf{1} - 2\mathbf{e}_j$, so their span contains $\frac{1}{2}(\mathbf{1} - (\mathbf{1} - 2\mathbf{e}_j)) = \mathbf{e}_j$ for $j \geq 2$ and then $\mathbf{e}_1 = \mathbf{1} - \sum_{j \geq 2} \mathbf{e}_j$. The same holds at every $\pm\mathbf{e}_i$. Every other point $\mathbf{x}$ of B_1 has active rows of rank less than n. If $\|\mathbf{x}\|_1 < 1$, no row is active. If $\|\mathbf{x}\|_1 = 1$, then $\mathbf{x}$ has two nonzero entries x_i, x_j with $i \neq j$, since a point with one nonzero entry and $\|\mathbf{x}\|_1 = 1$ is some $\pm\mathbf{e}_i$; let $\sigma_i, \sigma_j \in \{\pm 1\}$ be their signs. An active $\mathbf{s}$ has $\sum_k s_k x_k = \sum_k |x_k|$ with each $s_k x_k \leq |x_k|$, so $s_k x_k = |x_k|$ for every k, and in particular $s_i = \sigma_i$, $s_j = \sigma_j$. Hence $\mathbf{u} = \sigma_i \mathbf{e}_i - \sigma_j \mathbf{e}_j \neq \mathbf{0}$ has $\mathbf{s}^\top \mathbf{u} = \sigma_i^2 - \sigma_j^2 = 0$ for every active $\mathbf{s}$, and the matrix of active rows has rank less than n by Theorem 3.5.7. By Theorem 18.9.2, the vertices are exactly the $2n$ points $\pm\mathbf{e}_i$.

Duality. The rows of the cube are $\pm\mathbf{e}_i$, which are the vertices of the cross-polytope; the rows of the cross-polytope are the sign vectors, which are the vertices of the cube. In $\mathbb{R}^3$, the cube has 8 vertices and 6 inequalities, the octahedron 6 vertices and 8 inequalities. This exchange is the polytope form of the fact that $\|\cdot\|_1$ and $\|\cdot\|_\infty$ are dual norms (Definition 18.4.4): the maximum of $\mathbf{y}^\top \mathbf{x}$ over $\mathbf{x} \in B_1$ is $\|\mathbf{y}\|_\infty$, and over $\mathbf{x} \in B_\infty$ it is $\|\mathbf{y}\|_1$.

18.9.6 Exercises

18.9.6.1 A. Check your understanding

Exercise. A1

(a) State the definitions of a polytope and of a polyhedron in $\mathbb{R}^n$.

(b) True or false: every polyhedron is a polytope. Justify your answer.

(c) True or false: the intersection of two polytopes in $\mathbb{R}^n$ is a polytope. Justify your answer.

(d) State the criterion of Theorem 18.9.2 for a point of $\{\mathbf{x} : \mathbf{A}\mathbf{x} \le \mathbf{b}\}$ to be a vertex.

(e) True or false: every inequality produced by Fourier–Motzkin elimination (Lemma 18.9.4) is needed to describe the projection. Justify your answer.

Solution.

(a) A polytope is the convex hull of finitely many points of $\mathbb{R}^n$. A polyhedron is a set $\{\mathbf{x} \in \mathbb{R}^n : \mathbf{A}\mathbf{x} \le \mathbf{b}\}$ for some $\mathbf{A} \in M_{m\times n}(\mathbb{R})$ and $\mathbf{b} \in \mathbb{R}^m$, that is, an intersection of finitely many sets $\{\mathbf{x} : \mathbf{a}_i^{\top}\mathbf{x} \le b_i\}$.

(b) False. The half-plane $\{x_2 \ge 0\} \subseteq \mathbb{R}^2$ is a polyhedron, with one row, but it is unbounded, while every polytope is bounded.

(c) True. By Theorem 18.9.5 (a), each is a bounded polyhedron. Listing both systems of inequalities together describes the intersection, which is therefore a polyhedron, and it is bounded because it lies inside one of the two. By Theorem 18.9.5 (a) again, it is a polytope.

(d) A point $\mathbf{x}$ with $\mathbf{A}\mathbf{x} \le \mathbf{b}$ is a vertex if and only if the rows $\mathbf{a}_i$ with $\mathbf{a}_i^{\top}\mathbf{x} = b_i$ include n linearly independent ones.

(e) False. In the warning after Example (A triangle, both ways), two of the four output rows are $0 \le 1$, which every point satisfies.

18.9.6.2 B. Practice

Exercise. B1: Vertices of a pentagon

Find all vertices of

$$P = \{\mathbf{x} \in \mathbb{R}^2 : x_1 \ge 0,\ x_2 \ge 0,\ x_1 + x_2 \le 4,\ x_1 - x_2 \le 2,\ -x_1 + 2x_2 \le 5\},$$

and the maximum of $2x_1 + 3x_2$ over P. Hence write P as a polytope.

Solution. Number the rows (1) $-x_1 \le 0$, (2) $-x_2 \le 0$, (3) $x_1 + x_2 \le 4$, (4) $x_1 - x_2 \le 2$, (5) $-x_1 + 2x_2 \le 5$. Any two rows are independent (no two are multiples of each other), so each of the $\binom{5}{2} = 10$ pairs gives one basic point.

pair	point	verdict
(1), (2)	$(0, 0)$	feasible
(1), (3)	$(0, 4)$	(5): $8 \le 5$ fails
(1), (4)	$(0, -2)$	(2) fails
(1), (5)	$(0, \frac{5}{2})$	feasible
(2), (3)	$(4, 0)$	(4): $4 \le 2$ fails
(2), (4)	$(2, 0)$	feasible
(2), (5)	$(-5, 0)$	(1) fails
(3), (4)	$(3, 1)$	feasible
(3), (5)	$(1, 3)$	feasible
(4), (5)	$(9, 7)$	(3): $16 \le 4$ fails

For instance, (3) and (4) added give $2x_1 = 6$, so $\mathbf{x} = (3,1)$; (3) and (5) added give $3x_2 = 9$, so $\mathbf{x} = (1,3)$; (4) and (5) added give $x_2 = 7$, then $x_1 = 9$. For the feasible points, the checks are routine: at $(0, \frac{5}{2})$, $0+\frac{5}{2} \le 4$ and $-\frac{5}{2} \le 2$; at $(2,0)$, $2 \le 4$ and $-2 \le 5$; at $(3,1)$, $-3+2 \le 5$; at $(1,3)$, $1-3 \le 2$. By Theorem 18.9.2 the vertices are $(0,0)$, $(2,0)$, $(3,1)$, $(1,3)$, $(0,\frac{5}{2})$.

P is bounded, since $0 \le x_1, x_2 \le 4$ on it by rows (1) to (3), and nonempty. So by Corollary 18.8.6 the maximum of $2x_1 + 3x_2$ is attained at a vertex. The values are $0, 4, 9, 11, \frac{15}{2}$, so the maximum is 11, at $(1,3)$. By the proof of Theorem 18.9.5 (a), $P = \operatorname{conv}\{(0,0),(2,0),(3,1),(1,3),(0,\frac{5}{2})\}$.

Exercise. B2: Inequalities for a quadrilateral

Use Fourier–Motzkin elimination to find a system of inequalities describing the quadrilateral

$$Q = \operatorname{conv}\{(0,0),(2,0),(2,1),(0,2)\}.$$

Hence find the vertices of Q.

Hint: substitute away t_1, t_2, t_4 and keep $s = t_3$, the weight of $(2,1)$.

Solution. $\mathbf{x} \in Q$ if and only if there are $t_1, \dots, t_4 \ge 0$ with $\sum t_j = 1$, $x_1 = 2t_2 + 2t_3$, and $x_2 = t_3 + 2t_4$. With $s = t_3$: $t_2 = \frac{1}{2}(x_1 - 2s)$, $t_4 = \frac{1}{2}(x_2 - s)$, and

$$t_1 = 1 - t_2 - s - t_4 = 1 - \tfrac{1}{2}x_1 - \tfrac{1}{2}x_2 + \tfrac{1}{2}s.$$

The conditions $2t_1 \ge 0$, $2t_2 \ge 0$, $2t_4 \ge 0$, $t_3 \ge 0$ read

$$\begin{aligned} &(1)\quad x_1 + x_2 - s \le 2, && (2)\quad -x_1 + 2s \le 0,\\ &(3)\quad -x_2 + s \le 0, && (4)\quad -s \le 0. \end{aligned}$$

The coefficients of s are $-1, 2, 1, -1$, so $I_+ = \{2,3\}$ and $I_- = \{1,4\}$. The four pairs give:

$$\begin{aligned} (2),(1):&\ 2\cdot(1) + 1\cdot(2) = x_1 + 2x_2 \le 4,\\ (2),(4):&\ 2\cdot(4) + 1\cdot(2) = -x_1 \le 0,\\ (3),(1):&\ 1\cdot(1) + 1\cdot(3) = x_1 \le 2,\\ (3),(4):&\ 1\cdot(4) + 1\cdot(3) = -x_2 \le 0. \end{aligned}$$

So $Q = \{0 \le x_1 \le 2,\ x_2 \ge 0,\ x_1 + 2x_2 \le 4\}$ by Lemma 18.9.4. Its rows are $(-1,0),(1,0),(0,-1),(1,2)$. The pair $\pm(1,0)$ is dependent; the other five pairs give the basic points $(0,0)$, $(0,2)$, $(2,0)$, $(2,1)$, and $(4,0)$ (from $x_2 = 0$, $x_1 + 2x_2 = 4$), and the last violates $x_1 \le 2$. So by Theorem 18.9.2 the vertices are the four given points, and none of them was redundant.

Exercise. B3: Which of these are vertices?

Let $P = \{\mathbf{x} \in \mathbb{R}^3 : \mathbf{x} \ge \mathbf{0},\ x_1 + x_2 + x_3 \le 3,\ x_1 + x_2 \le 2\}$. Determine which of the following points are vertices of P. Justify your answer.

(a) $(0,2,1)$

(b) $(1,1,1)$
(c) $(0,0,3)$
(d) $(3,0,0)$
(e) $(1,0,0)$

Solution. We use Theorem 18.9.2: check that the point lies in P, list its active rows, and compute their rank. The rows are $-\mathbf{e}_1, -\mathbf{e}_2, -\mathbf{e}_3$, $(1,1,1)$, and $(1,1,0)$.

(a) $(0,2,1) \in P$: the entries are ≥ 0, $0+2+1=3\leq 3$, $0+2\leq 2$. Active: $-\mathbf{e}_1$, $(1,1,1)$, $(1,1,0)$. Their span contains $(1,1,1)-(1,1,0)=\mathbf{e}_3$, $\mathbf{e}_1$, and $(1,1,0)-\mathbf{e}_1=\mathbf{e}_2$, so it is $\mathbb{R}^3$. Rank 3: **a vertex**.

(b) $(1,1,1)\in P$, with active rows $(1,1,1)$ and $(1,1,0)$ only. Rank $2<3$: **not a vertex**. Indeed it is the midpoint of $(0,2,1)$ and $(2,0,1)$, both in P.

(c) $(0,0,3)\in P$, with active rows $-\mathbf{e}_1$, $-\mathbf{e}_2$, $(1,1,1)$, which are independent (the third has a nonzero third entry, the first two do not). Rank 3: **a vertex**.

(d) $(3,0,0)\notin P$, since $3+0=3>2$. It is a basic point, the solution of $x_2=0$, $x_3=0$, $x_1+x_2+x_3=3$, but not a feasible one: **not a vertex**.

(e) $(1,0,0)\in P$, with active rows $-\mathbf{e}_2$ and $-\mathbf{e}_3$ only. Rank $2<3$: **not a vertex**. It is the midpoint of $(0,0,0)$ and $(2,0,0)$.

18.9.6.3 C. Going deeper

Exercise. C1: Unbounded polyhedra

Let $P=\{\mathbf{x}\in\mathbb{R}^n : \mathbf{A}\mathbf{x}\leq\mathbf{b}\}$ be a **nonempty** polyhedron, and let $C=\{(\mathbf{x},t)\in\mathbb{R}^n\times\mathbb{R} : \mathbf{A}\mathbf{x}-t\mathbf{b}\leq\mathbf{0},\ -t\leq 0\}$.

(a) Prove that $\mathbf{x}\in P$ if and only if $(\mathbf{x},1)\in C$.

(b) By Theorem 18.9.5 (b), C is generated by finitely many vectors $(\mathbf{w}_j,t_j)$, $j=1,\dots,k$. Prove that every $t_j\geq 0$, and that

$$P=\operatorname{conv}\{\mathbf{w}_j/t_j : t_j>0\}+\Big\{\sum_{t_j=0} c_j\mathbf{w}_j : c_j\geq 0\Big\},$$

the set of sums of a point of the polytope and a point of the cone. So every nonempty polyhedron is a polytope plus a finitely generated cone.

(c) Check the statement for $P=\{\mathbf{x}\in\mathbb{R}^2 : x_1\geq 0,\ 0\leq x_2\leq 1\}$, with the polytope $\operatorname{conv}\{(0,0),(0,1)\}$ and the cone generated by $\mathbf{e}_1$.

Solution.

(a) $(\mathbf{x},1)\in C$ says $\mathbf{A}\mathbf{x}-\mathbf{b}\leq\mathbf{0}$ and $-1\leq 0$, which is $\mathbf{x}\in P$.

(b) Each generator lies in C, so $-t_j\leq 0$. Split the indices into $J_+=\{j : t_j>0\}$ and $J_0=\{j : t_j=0\}$.
($\subseteq$) Let $\mathbf{x}\in P$. By (a), $(\mathbf{x},1)=\sum_j c_j(\mathbf{w}_j,t_j)$ with $c_j\geq 0$. The last coordinate gives

$1 = \sum_{j\in J_+} c_j t_j$, so $J_+ \neq \emptyset$, and

$$\mathbf{x} = \sum_{j\in J_+} (c_j t_j)\,\frac{\mathbf{w}_j}{t_j} + \sum_{j\in J_0} c_j \mathbf{w}_j,$$

where the weights $c_j t_j$ are nonnegative and add up to 1. So $\mathbf{x}$ lies in the right-hand side.

($\supseteq$) Let $\mathbf{x} = \sum_{j\in J_+} s_j \mathbf{w}_j/t_j + \sum_{j\in J_0} c_j \mathbf{w}_j$ with $s_j \geq 0$, $\sum s_j = 1$, and $c_j \geq 0$. Then

$$(\mathbf{x}, 1) = \sum_{j\in J_+} \frac{s_j}{t_j}(\mathbf{w}_j, t_j) + \sum_{j\in J_0} c_j(\mathbf{w}_j, 0),$$

a nonnegative combination of generators, so $(\mathbf{x}, 1) \in C$ and $\mathbf{x} \in P$ by (a).

(c) A point $(0, u) + c\mathbf{e}_1$ with $0 \leq u \leq 1$ and $c \geq 0$ is (c, u), which satisfies $x_1 \geq 0$ and $0 \leq x_2 \leq 1$. Conversely $(x_1, x_2) \in P$ is $(0, x_2) + x_1\mathbf{e}_1$ with $(0, x_2) \in \operatorname{conv}\{(0,0),(0,1)\}$ and $x_1 \geq 0$. So $P = \operatorname{conv}\{(0,0),(0,1)\} + \{c\mathbf{e}_1 : c \geq 0\}$.

Exercise. C2: Linear images of polyhedra

Let $\mathbf{M} \in M_{p\times n}(\mathbb{R})$.

(a) Prove that if $P \subseteq \mathbb{R}^n$ is a polyhedron, then $\mathbf{M}(P) = \{\mathbf{Mx} : \mathbf{x} \in P\}$ is a polyhedron in $\mathbb{R}^p$, and that if P is a polyhedral cone, so is $\mathbf{M}(P)$.

(b) Prove that if Q is a polytope, then $\mathbf{M}(Q)$ is a polytope, directly from Definition 18.9.1.

(c) Deduce that a linear image of a polyhedral cone is closed. Why does this not contradict the existence of a closed convex cone with a nonclosed linear image?

Solution.

(a) Let $P = \{\mathbf{x} : \mathbf{Ax} \leq \mathbf{b}\}$. Then $\mathbf{y} \in \mathbf{M}(P)$ if and only if there is $\mathbf{x}$ with $(\mathbf{y}, \mathbf{x})$ in

$$\widehat{P} = \{(\mathbf{y}, \mathbf{x}) \in \mathbb{R}^p \times \mathbb{R}^n : \mathbf{y} - \mathbf{Mx} \leq \mathbf{0},\ -\mathbf{y} + \mathbf{Mx} \leq \mathbf{0},\ \mathbf{Ax} \leq \mathbf{b}\},$$

a polyhedron in $\mathbb{R}^{p+n}$. So $\mathbf{M}(P)$ is the image of $\widehat{P}$ under the map forgetting the last n coordinates, and n applications of Lemma 18.9.4 show that it is a polyhedron. If $\mathbf{b} = \mathbf{0}$, every right-hand side in $\widehat{P}$ is 0, and stays 0 under elimination, so $\mathbf{M}(P)$ is a polyhedral cone.

(b) If $Q = \operatorname{conv}\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$, then by Theorem 18.1.3 and linearity,

$$\mathbf{M}\Big(\sum_j t_j \mathbf{v}_j\Big) = \sum_j t_j \mathbf{M}\mathbf{v}_j,$$

so $\mathbf{M}(Q) = \operatorname{conv}\{\mathbf{Mv}_1, \ldots, \mathbf{Mv}_k\}$, a polytope.

(c) By (a), the image of a polyhedral cone is a polyhedral cone, which is closed by Proposition 18.9.1. So a closed convex cone whose linear image is not closed cannot be polyhedral; by Theorem 18.9.5 (b), it cannot be finitely generated either. This is consistent with Section 5, where closedness was proved only for finitely generated cones.

Exercise. C3: Vertices in standard form

Let $\mathbf{A} \in M_{m\times n}(\mathbb{R})$ have columns $\mathbf{c}_1, \dots, \mathbf{c}_n$, let $\mathbf{b} \in \mathbb{R}^m$, and let $P = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{A}\mathbf{x} = \mathbf{b},\ \mathbf{x} \geq \mathbf{0}\}$, the feasible region of a linear program in standard form.

(a) Prove that $\mathbf{x} \in P$ is a vertex of P if and only if the columns $\mathbf{c}_j$ with $x_j > 0$ are linearly independent.

(b) Deduce that a vertex of P has at most rank $\mathbf{A}$ positive entries.

(c) Find the vertices of $P = \{\mathbf{x} \in \mathbb{R}^3 : x_1 + x_2 + x_3 = 1,\ \mathbf{x} \geq \mathbf{0}\}$, the standard triangle.

Solution.

(a) Write P as a polyhedron with the rows $\mathbf{A}\mathbf{x} \leq \mathbf{b}$, $-\mathbf{A}\mathbf{x} \leq -\mathbf{b}$, and $-\mathbf{e}_j^\top \mathbf{x} \leq 0$. At $\mathbf{x} \in P$, every row of $\pm\mathbf{A}$ is active, and $-\mathbf{e}_j$ is active exactly when $x_j = 0$. Let $S = \{j : x_j > 0\}$. By Theorem 18.9.2, $\mathbf{x}$ is a vertex if and only if the active rows have rank n, that is (Theorem 3.5.7), if and only if the only $\mathbf{u}$ with $\mathbf{A}\mathbf{u} = \mathbf{0}$ and $u_j = 0$ for all $j \notin S$ is $\mathbf{u} = \mathbf{0}$. For such $\mathbf{u}$, $\mathbf{A}\mathbf{u} = \sum_{j\in S} u_j \mathbf{c}_j$. So the condition says that $\sum_{j\in S} u_j \mathbf{c}_j = \mathbf{0}$ forces every $u_j = 0$, which is the linear independence of $\{\mathbf{c}_j : j \in S\}$.

(b) The columns $\mathbf{c}_j$ lie in $\mathrm{col}(\mathbf{A})$, of dimension $\operatorname{rank}\mathbf{A}$, and an independent list in it has at most $\operatorname{rank}\mathbf{A}$ vectors by Theorem 1.6.3 (a). By (a), a vertex has at most that many positive entries.

(c) Here $\mathbf{A} = (1\ 1\ 1)$, of rank 1. By (b) a vertex has at most one positive entry, and since the entries add up to 1, exactly one, equal to 1. Conversely $\mathbf{e}_j \in P$, and its one positive entry corresponds to the column $(1) \neq \mathbf{0}$, an independent list. So the vertices are $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$.

18.10 Convex Functions

The chapter so far has studied convex **sets**. Convex **functions** have been in the book for longer, as the first subsection recalls, and every finite support function of §04 is one too. This section defines the notion, ties it to convex sets through the epigraph, and proves its basic theorems: Jensen's inequality, the convexity of a supremum of affine functions, which pays a promise of Chapter 17 §06, a second-derivative test, and continuity. §11 then spends them on three classical inequalities.

Throughout, the field is $\mathbb{R}$. V is a real vector space, and a set $C \subseteq V$ called convex is convex in the sense of Definition 18.1.1. From the second-derivative test onward we work in $\mathbb{R}^n$, with the standard inner product and its norm $\|\cdot\| = \|\cdot\|_2$. A function g on an interval of $\mathbb{R}$ is **increasing** when $s \leq t$ implies $g(s) \leq g(t)$, as in Chapter 17 §11, and **strictly increasing** when $s < t$ implies $g(s) < g(t)$.

18.10.1 Convex functions and their epigraphs

Look again at three inequalities the book has already proved. For a norm, (N2) and (N3) of Definition 16.1.1 give $\|tx + (1-t)y\| \leq t\|x\| + (1-t)\|y\|$ for $0 \leq t \leq 1$. For the top partial sum $s_k(\mathbf{A}) = \sum_{i\leq k} \lambda_i(\mathbf{A})$ of a Hermitian matrix, Chapter 17 §06 derived $s_k(t\mathbf{A} + (1-t)\mathbf{B}) \leq t s_k(\mathbf{A}) + (1-t)s_k(\mathbf{B})$ from Corollary 17.6.5. And the case $n = 1$ of operator convexity in Definition 17.11.1 (b) reads $f(ta + (1-t)b) \leq tf(a) + (1-t)f(b)$ for numbers a, b. It is the same

inequality three times, and it deserves a name.

A function is convex when, on every segment of its domain, its graph lies on or below the chord.

Definition 18.10.1. Convex Function

Let C be a **nonempty convex** subset of a real vector space V, and let $f\colon C \to \mathbb{R}$. Then f is **convex** if

$$f(tx + (1-t)y) \;\le\; tf(x) + (1-t)f(y)$$

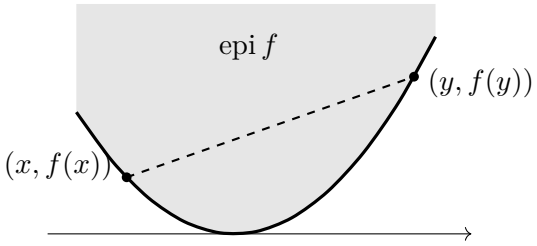
for **all** $x, y \in C$ and **all** real t with $0 \le t \le 1$. It is **strictly convex** if the inequality is **strict** whenever $x \neq y$ and $0 < t < 1$. It is **concave**, or **strictly concave**, if $-f$ is convex, or strictly convex.

The **epigraph** of f is the subset of the real vector space $V \times \mathbb{R}$, with componentwise operations,

$$\operatorname{epi} f = \{(x, s) \in V \times \mathbb{R} : x \in C,\ s \ge f(x)\}.$$

In words: take any two points x and y of the domain and any point $tx + (1-t)y$ between them. The left side is the value of f there. The right side is the value, at the same place, of the straight line through $(x, f(x))$ and $(y, f(y))$. So the condition says the chord is never below the graph. Strict convexity forbids the graph to touch the chord anywhere strictly between two distinct points. The epigraph is everything on or above the graph.

The domain must be convex, and this is the only thing to check. The point $tx+(1-t)y$ must lie in C, or f is not defined there. That is exactly what Definition 18.1.1 guarantees.



The picture shows $f(x) = 0.35x^2$ on $\mathbb{R}$: the dashed chord lies above the graph between its endpoints, and the shaded region is the epigraph.

Example. First convex functions

Check each against Definition 18.10.1.

(a) An **affine function** $a(x) = \varphi(x) + c$, with φ a linear functional on V (Definition 5.1.1) and $c \in \mathbb{R}$.

(b) A norm $\|\cdot\|$ on $\mathbb{R}^n$.

(c) $f(x) = x^2$ on $\mathbb{R}$, and more generally $q(\mathbf{x}) = \mathbf{x}^\top \mathbf{A}\mathbf{x}$ on $\mathbb{R}^n$ for a symmetric $\mathbf{A} \succeq \mathbf{O}$.

Solution.

(a) Since φ is linear and $c = tc + (1-t)c$,

$$a(tx + (1-t)y) = t\varphi(x) + (1-t)\varphi(y) + tc + (1-t)c = ta(x) + (1-t)a(y).$$

Equality holds, so a is convex, and so is $-a$: an affine function is both convex and concave, and neither strictly. The constant functions are the degenerate case $\varphi = 0$. They matter because they show that convexity by itself forces no growth at all.

(b) By (N3) and then (N2) of Definition 16.1.1, with $|t| = t$ and $|1-t| = 1-t$ for $0 \le t \le 1$,

$$\|t\mathbf{x} + (1-t)\mathbf{y}\| \le \|t\mathbf{x}\| + \|(1-t)\mathbf{y}\| = t\|\mathbf{x}\| + (1-t)\|\mathbf{y}\|.$$

A norm is not strictly convex: on the ray through a nonzero $\mathbf{x}$, $\|s\mathbf{x}\| = s\|\mathbf{x}\|$ is linear in $s \ge 0$, so the chord between $\mathbf{x}$ and $2\mathbf{x}$ lies on the graph.

(c) Expanding and collecting terms,

$$tx^2 + (1-t)y^2 - \big(tx + (1-t)y\big)^2 = t(1-t)(x-y)^2 \;\ge\; 0,$$

and the right side is strictly positive when $x \ne y$ and $0 < t < 1$. So x^2 is strictly convex. The same expansion with $\mathbf{A}$ symmetric gives

$$tq(\mathbf{x}) + (1-t)q(\mathbf{y}) - q\big(t\mathbf{x} + (1-t)\mathbf{y}\big) = t(1-t)\, q(\mathbf{x}-\mathbf{y}),$$

and $q(\mathbf{x}-\mathbf{y}) = \langle \mathbf{A}(\mathbf{x}-\mathbf{y}), \mathbf{x}-\mathbf{y}\rangle \ge 0$ by (P2) of Definition 13.1.1. So q is convex. The case $n = 1$, $\mathbf{A} = (1)$, is x^2.

A nonexample by minimal change. Replace x^2 by x^3 on $\mathbb{R}$. With $x = -2$, $y = 0$, and $t = \frac{1}{2}$, the left side is $(-1)^3 = -1$ and the right side is $\frac{1}{2}(-8) + \frac{1}{2}\cdot 0 = -4$, and $-1 \le -4$ is false. What fails is the inequality for this pair, on the negative half-line, where the graph bends downward. On the convex domain $(0, \infty)$ the same function is convex, as the one-variable test below shows. So convexity depends on the domain.

Why this definition. The inequality is nonstrict so that affine functions and norms, the most important examples, qualify. Requiring it for every $t \in [0,1]$, not just $t = \frac{1}{2}$, is what makes Jensen's inequality below immediate. The name comes from the epigraph: a function is convex exactly when the set above its graph is convex.

Proposition 18.10.1. Convex Function, Convex Epigraph

Let $C \subseteq V$ be nonempty and convex, and let $f\colon C \to \mathbb{R}$. Then f is convex if and only if $\operatorname{epi} f$ is a convex subset of $V \times \mathbb{R}$.

Idea. A point (x, s) lies in the epigraph exactly when $s \ge f(x)$. So the chord inequality for f and the statement that segments between points of the epigraph stay in it are one inequality, read in two directions.

Proof. ($\Rightarrow$) Let $(x,s), (y,u) \in \operatorname{epi} f$ and $0 \le t \le 1$. The point $tx + (1-t)y$ lies in C, since C is convex, and

$$f\big(tx + (1-t)y\big) \le tf(x) + (1-t)f(y) \le ts + (1-t)u,$$

the second step because $f(x) \le s$, $f(y) \le u$, and $t, 1-t \ge 0$. So $t(x,s) + (1-t)(y,u) \in \operatorname{epi} f$.

($\Leftarrow$) Let $x, y \in C$ and $0 \le t \le 1$. The points $(x, f(x))$ and $(y, f(y))$ lie in $\operatorname{epi} f$, so by convexity so does $\big(tx + (1-t)y,\ tf(x) + (1-t)f(y)\big)$. By the definition of the epigraph this

says $f(tx + (1-t)y) \le tf(x) + (1-t)f(y)$. This proves the proposition.

■

The same one-line argument shows that every **sublevel set** $\{x \in C : f(x) \le c\}$ of a convex f is convex: if $f(x) \le c$ and $f(y) \le c$, then $f(tx + (1-t)y) \le tc + (1-t)c = c$.

Quick check. Every sublevel set of a convex function is convex. Is the converse true? Is a function $f : \mathbb{R} \to \mathbb{R}$ whose sublevel sets are all convex necessarily convex?

Solution. No. Take $f(x) = x^3$. It is increasing: if $a < b$ then $b^3 - a^3 = (b-a)(a^2+ab+b^2) > 0$, since $a^2 + ab + b^2 = (a + \frac{b}{2})^2 + \frac{3}{4}b^2$ vanishes only at $a = b = 0$. So if $x \le z \le y$ and $x^3, y^3 \le c$, then $z^3 \le y^3 \le c$. Each sublevel set is therefore an interval, which is convex. But x^3 is not convex on $\mathbb{R}$, by the nonexample above.

Warning. **A convex function need not be differentiable.** The function $|x|$ is a norm on $\mathbb{R}$, hence convex, and it has a corner at 0: its difference quotients $|h|/h$ are 1 for $h > 0$ and -1 for $h < 0$, so they have no limit. The same corner appears in linear algebra. The map $\mathbf{A} \mapsto \lambda_1(\mathbf{A})$ on Hermitian matrices is convex, as the next two subsections prove, and on the diagonal matrices $\operatorname{diag}(t, -t)$ it equals $|t|$. The corner sits at $t = 0$, where the top eigenvalue is repeated.

18.10.2 Jensen's inequality

The definition speaks of two points. Induction extends it to any finite convex combination (Definition 18.1.3), and that extension is the form in which convexity is most often used.

Theorem 18.10.2. Jensen's Inequality

Let $C \subseteq V$ be nonempty and convex, and let $f : C \to \mathbb{R}$ be convex. Let $x_1, \dots, x_m \in C$, and let $w_1, \dots, w_m \ge 0$ with $w_1 + \cdots + w_m = 1$. Then $\sum_i w_i x_i \in C$ and

$$f\Big(\sum_{i=1}^{m} w_i x_i\Big) \;\le\; \sum_{i=1}^{m} w_i f(x_i).$$

If f is **strictly** convex and every $w_i > 0$, equality holds only when $x_1 = x_2 = \cdots = x_m$.

Idea. Split off the last point. The remaining weights add up to $s = 1 - w_m$, and after dividing by s they are the weights of a convex combination y of $m-1$ points, to which induction applies. The whole combination is then $sy + w_m x_m$, a combination of **two** points, which is where the definition applies.

Proof. We prove both statements by induction on $m \geq 1$. For $m = 1$ the weight is $w_1 = 1$, both sides equal $f(x_1)$, and the equality clause is empty.

Let $m \geq 2$ and assume both statements for $m - 1$ points. If $w_m = 1$, every other weight is 0, both sides equal $f(x_m)$, and the equality clause does not apply since $w_1 = 0$. Otherwise $s = 1 - w_m = w_1 + \cdots + w_{m-1} > 0$. The numbers w_i/s for $i < m$ are nonnegative with sum 1, so by the inductive hypothesis $y = \sum_{i<m}(w_i/s)x_i$ lies in C and

$$f(y) \leq \sum_{i<m} \frac{w_i}{s} f(x_i). \tag{1}$$

Now $\sum_{i=1}^{m} w_i x_i = sy + w_m x_m$ with $s + w_m = 1$, so this point lies in C because C is convex, and Definition 18.10.1 gives

$$\begin{aligned} f(sy + w_m x_m) &\leq s f(y) + w_m f(x_m) \\ &\leq \sum_{i<m} w_i f(x_i) + w_m f(x_m), \end{aligned} \tag{2}$$

where the second step multiplies (1) by $s > 0$. This proves the inequality.

For the equality clause, let f be strictly convex and every $w_i > 0$, so that $0 < w_m < 1$ and $0 < s < 1$. If equality holds in the theorem, both inequalities in (2) are equalities. The second is s times (1), so (1) is an equality, and the inductive hypothesis gives $x_1 = \cdots = x_{m-1}$; then y equals this common point. The first is the two-point inequality with weights strictly between 0 and 1, so strict convexity forces $y = x_m$. Hence all m points are equal, as claimed.

■

In the language of §01, the points $(x_i, f(x_i))$ lie in the convex set $\operatorname{epi} f$, and Jensen's inequality says that their convex combinations stay there. That is Theorem 18.1.3 read inside the epigraph.

18.10.3 Suprema of affine functions

Chapter 17 §06 showed that $s_k(\mathbf{A}) = \sum_{i \leq k} \lambda_i(\mathbf{A})$ is the largest of the linear functions $\mathbf{A} \mapsto \operatorname{tr}(\mathbf{Q}^{*}\mathbf{A}\mathbf{Q})$, deduced its convexity, and ended with a promise: "Chapter 18 studies convex functions in general, and the observation that a pointwise supremum of linear functions is convex is one of its basic facts." Here is that fact, stated for affine functions, which include the linear ones.

Proposition 18.10.3. A Supremum of Affine Functions Is Convex

Let $C \subseteq V$ be nonempty and convex, and let $(a_i)_{i \in I}$ be a nonempty family of affine functions on V. Suppose that for every $x \in C$ the set $\{a_i(x) : i \in I\}$ is bounded above, and put

$$g(x) = \sup_{i \in I} a_i(x) \qquad (x \in C).$$

Then $g \colon C \to \mathbb{R}$ is convex.

Idea. Each a_i meets its own chord inequality with equality, and each lies below g. So at a point between x and y, every a_i is at most the chord of g, and therefore so is their supremum.

Proof. The supremum exists for each $x \in C$ by completeness, fact (A1) of Chapter 16's introduction. Let $x, y \in C$ and $0 \le t \le 1$. For each $i \in I$, by Example (First convex functions) (a) and then $a_i \le g$ at x and at y,

$$a_i(tx + (1-t)y) = ta_i(x) + (1-t)a_i(y) \le tg(x) + (1-t)g(y),$$

where the inequality uses $t \ge 0$ and $1 - t \ge 0$. So the right side is an upper bound for the set $\{a_i(tx + (1-t)y) : i \in I\}$, and it is therefore at least that set's least upper bound, $g(tx + (1-t)y)$. This proves the proposition.

■

The proof never used that the a_i are affine beyond the fact that each is convex. So the same argument shows that **a pointwise supremum of convex functions is convex**, whenever it is finite. In particular the maximum of two convex functions is convex. The minimum need not be: $\min(x, -x) = -|x|$ is concave and not convex.

Instances. Many convex functions of this chapter are suprema of linear functions, and this proposition proves their convexity at a stroke.

- $|x| = \max(x, -x)$; $\|\mathbf{x}\|_\infty = \max_i \max(x_i, -x_i)$; and $\|\mathbf{x}\|_1 = \max\{\langle \boldsymbol{\epsilon}, \mathbf{x}\rangle : \boldsymbol{\epsilon} \in \{\pm 1\}^n\}$, the maximum being attained at the vector of signs of $\mathbf{x}$.
- The support function $h_K(\mathbf{y})$ of Definition 18.4.2 is the supremum over $\mathbf{x} \in K$ of the linear functions $\mathbf{y} \mapsto \langle \mathbf{x}, \mathbf{y}\rangle$. So it is convex wherever it is finite, and in particular the dual norm of Definition 18.4.4 is convex.
- Every norm is the maximum of $\mathbf{x} \mapsto \langle \mathbf{x}, \mathbf{y}\rangle$ over $\|\mathbf{y}\|_* \le 1$ (Corollary 18.4.9). This describes the norm; it is not a new proof of its convexity, since §04 derived the corollary from the triangle inequality.
- The largest eigenvalue: by Lemma 13.5.2, $\lambda_1(\mathbf{A}) = \max_{\|\mathbf{x}\|=1} \langle \mathbf{A}\mathbf{x}, \mathbf{x}\rangle$, and each $\mathbf{A} \mapsto \langle \mathbf{A}\mathbf{x}, \mathbf{x}\rangle$ is linear.
- The Ky Fan sums s_k, the case that prompted the promise. They are the subject of the next subsection.

18.10.4 Convex functions of a Hermitian matrix

Let $F = \mathbb{R}$ or $F = \mathbb{C}$. The Hermitian $n \times n$ matrices over F form a **real** vector space: if $\mathbf{A}^* = \mathbf{A}$, $\mathbf{B}^* = \mathbf{B}$, and $s, t \in \mathbb{R}$, then $(s\mathbf{A} + t\mathbf{B})^* = s\mathbf{A} + t\mathbf{B}$. They do not form a complex one, since $(i\mathbf{I})^* = -i\mathbf{I}$. This is the space on which Chapter 17 §06 called s_k convex, and it is the setting here. When $F = \mathbb{C}$ it is an instance of the chapter's convention that a complex space is treated as a real one.

Corollary 18.10.4. The Top Eigenvalue Sums Are Convex

Let $1 \le k \le n$. On the real vector space of Hermitian $n \times n$ matrices over F, the function

$$s_k(\mathbf{A}) = \lambda_1(\mathbf{A}) + \cdots + \lambda_k(\mathbf{A})$$

is convex. In particular $\mathbf{A} \mapsto \lambda_1(\mathbf{A})$ is convex, and $\mathbf{A} \mapsto \lambda_n(\mathbf{A})$ is concave.

Idea. Ky Fan's theorem writes s_k as a maximum of the functions $\mathbf{A} \mapsto \operatorname{tr}(\mathbf{Q}^*\mathbf{A}\mathbf{Q})$, each linear on the real space of Hermitian matrices, so Proposition 18.10.3 applies. For λ_n, change the

sign of $\mathbf{A}$.

Proof. By Theorem 17.6.2, $s_k(\mathbf{A}) = \max\{\operatorname{tr}(\mathbf{Q}^*\mathbf{A}\mathbf{Q}) : \mathbf{Q} \in M_{n\times k}(F),\ \mathbf{Q}^*\mathbf{Q} = \mathbf{I}_k\}$. Fix such a $\mathbf{Q}$. The matrix $\mathbf{Q}^*\mathbf{A}\mathbf{Q}$ is Hermitian, so its diagonal entries, and hence its trace, are real. By Theorem 0.9.6 (1), $\mathbf{A} \mapsto \operatorname{tr}(\mathbf{Q}^*\mathbf{A}\mathbf{Q})$ respects sums and real multiples. So it is a linear functional on the real space of Hermitian matrices, and s_k is the maximum of this family. By Proposition 18.10.3, s_k is convex, and $k = 1$ gives λ_1.

For λ_n, the eigenvalues of $-\mathbf{A}$ are the negatives of those of $\mathbf{A}$, so $\lambda_n(\mathbf{A}) = -\lambda_1(-\mathbf{A})$, as in the proof of Corollary 17.6.6. The map $\mathbf{A} \mapsto \lambda_1(-\mathbf{A})$ is convex, because it is the convex λ_1 composed with the linear map $\mathbf{A} \mapsto -\mathbf{A}$, and a convex function composed with a linear map is convex, by substituting into Definition 18.10.1. So λ_n is the negative of a convex function, that is, concave.

■

This is the fact Chapter 17 §06 anticipated, and its promise is now paid. Chapter 17 reached convexity by a different path: Corollary 17.6.5 gives $s_k(\mathbf{A}+\mathbf{B}) \le s_k(\mathbf{A}) + s_k(\mathbf{B})$, and positive homogeneity $s_k(t\mathbf{A}) = t s_k(\mathbf{A})$, $t \ge 0$, turns subadditivity into convexity. Exercise C1 shows that the two paths are the same, and the warning at the end of Chapter 17 §06 still stands: a single λ_k with $1 < k < n$, the difference $s_k - s_{k-1}$ of two convex functions, is neither convex nor concave.

What is not proved here. For a convex f on an interval, $\mathbf{A} \mapsto \operatorname{tr} f(\mathbf{A}) = \sum_i f(\lambda_i(\mathbf{A}))$ is convex on the Hermitian matrices with spectrum in that interval, a statement about sums over the spectrum and far weaker than the operator convexity of Chapter 17 §11. It is proved in Chapter 21, and nothing in this chapter uses it.

18.10.5 A second-derivative test

For a function on an interval, convexity says the slope never decreases. The mean value theorem turns an increasing derivative into a chord inequality.

Lemma 18.10.5. Increasing Derivative Gives Convexity

Let $I \subseteq \mathbb{R}$ be an open interval and let $g\colon I \to \mathbb{R}$ be differentiable. If g' is increasing on I, then g is convex; if g' is strictly increasing, then g is strictly convex. In particular, if g is twice differentiable with $g'' \ge 0$ on I, then g is convex, and if $g'' > 0$ on I, then g is strictly convex.

Idea. Split the chord at the intermediate point z into two sub-chords, over $[x, z]$ and over $[z, y]$. By the mean value theorem, each slope is a value of g' inside its interval, so an increasing g' makes the left slope at most the right one. A graph whose chord slopes increase in this way lies below the long chord.

Proof. Let $x, y \in I$ and $0 < t < 1$; the cases $t = 0$, $t = 1$, and $x = y$ are equalities. Exchanging x with y and t with $1 - t$ if necessary, we may assume $x < y$. Put $z = tx + (1-t)y$, so that

$x < z < y$, with

$$z - x = (1-t)(y-x), \qquad y - z = t(y-x).$$

By the mean value theorem, fact (A6) of Chapter 16's introduction, applied on $[x, z]$ and on $[z, y]$, there are $c_1 \in (x, z)$ and $c_2 \in (z, y)$ with

$$g(z) - g(x) = (1-t)(y-x)\, g'(c_1), \qquad g(y) - g(z) = t(y-x)\, g'(c_2).$$

Since $c_1 < c_2$ and g' is increasing, $g'(c_1) \le g'(c_2)$. Multiplying by $t(1-t)(y-x) > 0$,

$$\begin{aligned} t\big(g(z) - g(x)\big) &= t(1-t)(y-x)\, g'(c_1) \\ &\le t(1-t)(y-x)\, g'(c_2) = (1-t)\big(g(y) - g(z)\big). \end{aligned}$$

Rearranged, this is $g(z) \le tg(x) + (1-t)g(y)$. If g' is strictly increasing, then $g'(c_1) < g'(c_2)$, and every inequality is strict.

For the last sentence, let $a < b$ in I. Fact (A6) applied to g' gives $g'(b) - g'(a) = g''(c)(b-a)$ for some $c \in (a, b)$. This is ≥ 0 when $g'' \ge 0$, and > 0 when $g'' > 0$. So g' is increasing, or strictly increasing, and the first part applies.

■

Derivatives of polynomials, here and below, come from Chapter 10 §09: a polynomial is a power series with finitely many nonzero coefficients, so it converges for every real argument and that section's fact (A2) differentiates it term by term, while its fact (A3), the linearity of differentiation, handles sums and constant multiples; a partial derivative is such a derivative with the other variables fixed. For instance, $g(x) = x^4$ has $g'(x) = 4x^3$, which is strictly increasing (the Quick check after Proposition 18.10.1), so x^4 is strictly convex on $\mathbb{R}$. Likewise x^3, with $g'' = 6x \ge 0$ on $(0, \infty)$, is convex there, which is the claim made after the nonexample.

In several variables the second derivative is the Hessian $\mathbf{H}_f(\mathbf{a})$ of Definition 14.6.2, and "$g'' \ge 0$" becomes "positive semidefinite". Chapter 14 §06 read the Hessian at a single critical point, where its definiteness decides a local extremum (Theorem 14.6.4). Here the Hessian is asked to be positive semidefinite at **every** point, and what it decides is global. Restricting f to lines would need a chain rule the book has not imported, so the proof below uses only Chapter 14 §06's (A1), that $\mathbf{H}_f(\mathbf{a})$ is symmetric, and its (A2), Taylor's theorem with a remainder $r(\mathbf{h})$ satisfying $r(\mathbf{h})/\|\mathbf{h}\|^2 \to 0$. It works with **second differences**, which (A2) controls directly.

Theorem 18.10.6. Convexity from the Hessian

Let $U \subseteq \mathbb{R}^n$ be a nonempty open convex set, and let $f\colon U \to \mathbb{R}$ be twice continuously differentiable.

(a) f is convex on U if and only if $\mathbf{H}_f(\mathbf{a}) \succeq \mathbf{O}$ for every $\mathbf{a} \in U$.

(b) If $\mathbf{H}_f(\mathbf{a}) \succ \mathbf{O}$ for every $\mathbf{a} \in U$, then f is strictly convex on U.

Idea. Taylor's theorem at $\mathbf{z}$, applied at $\mathbf{z} + s\mathbf{v}$ and at $\mathbf{z} - s\mathbf{v}$ and added, kills the gradient term:

$$\frac{f(\mathbf{z} + s\mathbf{v}) + f(\mathbf{z} - s\mathbf{v}) - 2f(\mathbf{z})}{s^2} \longrightarrow \mathbf{v}^{\top}\mathbf{H}_f(\mathbf{z})\mathbf{v}.$$

For ($\Rightarrow$), convexity makes the left side ≥ 0, so the limit is too. For the other direction, restrict f to a segment and subtract the chord. If the difference were positive somewhere, it

would have an interior maximum, and at a maximum every second difference is ≤ 0. That contradicts a **strictly** positive limit. So the argument handles $\mathbf{H}_f \succ \mathbf{O}$ first, and reaches $\mathbf{H}_f \succeq \mathbf{O}$ by adding $\varepsilon\|\mathbf{x}\|^2$ and letting $\varepsilon \to 0$.

Proof. Write $\mathbf{H}(\mathbf{z}) = \mathbf{H}_f(\mathbf{z})$. For $\mathbf{z} \in U$, $\mathbf{v} \in \mathbb{R}^n$, and real $s \neq 0$ small enough that $\mathbf{z} \pm s\mathbf{v} \in U$, put

$$\Delta(\mathbf{z}, \mathbf{v}, s) = f(\mathbf{z} + s\mathbf{v}) + f(\mathbf{z} - s\mathbf{v}) - 2f(\mathbf{z}).$$

Step 1: the second difference. We claim $\Delta(\mathbf{z}, \mathbf{v}, s)/s^2 \to \mathbf{v}^\top \mathbf{H}(\mathbf{z})\mathbf{v}$ as $s \to 0$. If $\mathbf{v} = \mathbf{0}$ both sides are 0. Otherwise Chapter 14 §06's (A2) at $\mathbf{z}$, with $\mathbf{h} = s\mathbf{v}$ and with $\mathbf{h} = -s\mathbf{v}$, gives

$$f(\mathbf{z} \pm s\mathbf{v}) = f(\mathbf{z}) \pm s\,\nabla f(\mathbf{z}) \cdot \mathbf{v} + \tfrac{1}{2}s^2\,\mathbf{v}^\top \mathbf{H}(\mathbf{z})\mathbf{v} + r(\pm s\mathbf{v}).$$

Adding the two and subtracting $2f(\mathbf{z})$,

$$\frac{\Delta(\mathbf{z}, \mathbf{v}, s)}{s^2} = \mathbf{v}^\top \mathbf{H}(\mathbf{z})\mathbf{v} + \|\mathbf{v}\|^2 \Big(\frac{r(s\mathbf{v})}{\|s\mathbf{v}\|^2} + \frac{r(-s\mathbf{v})}{\|s\mathbf{v}\|^2} \Big),$$

and the bracket tends to 0 because $\|\pm s\mathbf{v}\| \to 0$. The same expansion shows that f is continuous: $f(\mathbf{z} + \mathbf{h}) - f(\mathbf{z}) = \nabla f(\mathbf{z}) \cdot \mathbf{h} + \tfrac{1}{2}\mathbf{h}^\top \mathbf{H}(\mathbf{z})\mathbf{h} + \|\mathbf{h}\|^2 \cdot r(\mathbf{h})/\|\mathbf{h}\|^2$, and each term tends to 0 as $\mathbf{h} \to \mathbf{0}$.

Step 2: (a), ($\Rightarrow$). Let f be convex, $\mathbf{a} \in U$, and $\mathbf{v} \in \mathbb{R}^n$. For small $s \neq 0$, $\mathbf{a} = \frac{1}{2}(\mathbf{a} + s\mathbf{v}) + \frac{1}{2}(\mathbf{a} - s\mathbf{v})$, so $f(\mathbf{a}) \le \frac{1}{2}f(\mathbf{a} + s\mathbf{v}) + \frac{1}{2}f(\mathbf{a} - s\mathbf{v})$, that is, $\Delta(\mathbf{a}, \mathbf{v}, s) \ge 0$. Dividing by $s^2 > 0$ and letting $s \to 0$, Step 1 gives $\mathbf{v}^\top \mathbf{H}(\mathbf{a})\mathbf{v} \ge 0$, since a nonstrict inequality survives a limit. The matrix $\mathbf{H}(\mathbf{a})$ is symmetric by Chapter 14 §06's (A1), so $\mathbf{H}(\mathbf{a}) \succeq \mathbf{O}$ by Definition 13.1.1.

Step 3: a maximum principle. Let $\mathbf{x} \neq \mathbf{y}$ in U, put $\mathbf{v} = \mathbf{y} - \mathbf{x}$, and let $\phi\colon U \to \mathbb{R}$ be continuous with the following property: for every point $\mathbf{z} = \mathbf{x} + t\mathbf{v}$ with $0 < t < 1$, the limit $L(\mathbf{z}) = \lim_{s \to 0} (\phi(\mathbf{z} + s\mathbf{v}) + \phi(\mathbf{z} - s\mathbf{v}) - 2\phi(\mathbf{z}))/s^2$ exists and is **strictly positive**. We claim

$$\phi(\mathbf{x} + t\mathbf{v}) < (1 - t)\phi(\mathbf{x}) + t\phi(\mathbf{y}) \qquad (0 < t < 1).$$

Define $h(t) = \phi(\mathbf{x} + t\mathbf{v}) - (1 - t)\phi(\mathbf{x}) - t\phi(\mathbf{y})$ on $[0, 1]$. Every point of the segment lies in U, since U is convex. So h is continuous, and $h(0) = h(1) = 0$. The subtracted part is affine in t, so its second differences vanish, and $h(t+s) + h(t-s) - 2h(t) = \phi(\mathbf{z} + s\mathbf{v}) + \phi(\mathbf{z} - s\mathbf{v}) - 2\phi(\mathbf{z})$ with $\mathbf{z} = \mathbf{x} + t\mathbf{v}$.

Suppose $h(t_1) \ge 0$ for some $t_1 \in (0, 1)$. The interval $[0, 1]$ is closed and bounded, hence compact by fact (A3) of Chapter 16's introduction, so by the extreme value theorem, fact (A4), h attains a maximum value M at some point. Then $M \ge h(t_1) \ge 0$. If the maximum is attained only at 0 or 1, then $M = 0 \le h(t_1)$, so t_1 attains it too. Either way, $M = h(t^*)$ for some $t^* \in (0, 1)$. For $0 < s < \min(t^*, 1 - t^*)$, both $h(t^* \pm s) \le h(t^*)$, so the second difference of h at t^* is ≤ 0, and so is its limit $L(\mathbf{x} + t^*\mathbf{v})$. This contradicts $L > 0$. Hence $h < 0$ on $(0, 1)$, which is the claim.

Step 4: (b). Let $\mathbf{H}(\mathbf{z}) \succ \mathbf{O}$ for every $\mathbf{z} \in U$. By Step 1, $\phi = f$ has $L(\mathbf{z}) = \mathbf{v}^\top \mathbf{H}(\mathbf{z})\mathbf{v} > 0$ whenever $\mathbf{v} \neq \mathbf{0}$, and it is continuous. So Step 3 gives $f((1 - t)\mathbf{x} + t\mathbf{y}) < (1 - t)f(\mathbf{x}) + tf(\mathbf{y})$ for all $\mathbf{x} \neq \mathbf{y}$ in U and $0 < t < 1$. Renaming t as $1 - t$, this is strict convexity.

Step 5: (a), ($\Leftarrow$). Let $\mathbf{H}(\mathbf{z}) \succeq \mathbf{O}$ for every $\mathbf{z} \in U$. Let $\mathbf{x} \neq \mathbf{y}$ in U, $\mathbf{v} = \mathbf{y} - \mathbf{x}$, $0 < t < 1$, and $\varepsilon > 0$. Put $\phi(\mathbf{z}) = f(\mathbf{z}) + \varepsilon\|\mathbf{z}\|^2$, which is continuous. Expanding the norms,

$$\|\mathbf{z} + s\mathbf{v}\|^2 + \|\mathbf{z} - s\mathbf{v}\|^2 - 2\|\mathbf{z}\|^2 = 2s^2\|\mathbf{v}\|^2,$$

so by Step 1 the limit for ϕ is $L(\mathbf{z}) = \mathbf{v}^{\top}\mathbf{H}(\mathbf{z})\mathbf{v} + 2\varepsilon\|\mathbf{v}\|^2 \geq 2\varepsilon\|\mathbf{v}\|^2 > 0$. Step 3 applies. With $\mathbf{w} = (1-t)\mathbf{x} + t\mathbf{y}$, it gives

$$\begin{aligned} f(\mathbf{w}) + \varepsilon\|\mathbf{w}\|^2 < (1-t)\big(f(\mathbf{x}) + \varepsilon\|\mathbf{x}\|^2\big) \\ + t\big(f(\mathbf{y}) + \varepsilon\|\mathbf{y}\|^2\big). \end{aligned}$$

Let $\varepsilon \to 0$ through $\varepsilon = 1/k$. The inequality survives as a nonstrict one, and gives $f(\mathbf{w}) \leq (1-t)f(\mathbf{x}) + tf(\mathbf{y})$. The cases $\mathbf{x} = \mathbf{y}$, $t = 0$, and $t = 1$ are equalities. So f is convex. This proves the theorem.

■

For $n = 1$, part (a) says that a twice continuously differentiable g on an open interval is convex exactly when $g'' \geq 0$, and its "if" half is Lemma 18.10.5 again.

Quick check. The function $f(x) = x^4$ is strictly convex on $\mathbb{R}$. Is its Hessian, the 1×1 matrix $(f''(x))$, positive definite at every point? What does this say about part (b)?

Solution. No. $f''(x) = 12x^2$ vanishes at $x = 0$, so the 1×1 matrix $\mathbf{O}_1$ is positive semidefinite but not positive definite. Yet f is strictly convex, as shown after Lemma 18.10.5. So the converse of part (b) fails: strict convexity does not force a positive definite Hessian.

Convexity is tested segment by segment, so f is convex exactly when its restriction $t \mapsto f(\mathbf{x}+t\mathbf{v})$ to every line is convex. The phrase *every line* matters.

Warning. **Convex along every line through one point is not enough.** Let $f(x, y) = x^2y^2$ on $\mathbb{R}^2$. On each line through the origin, $f(t\mathbf{v}) = t^4\, v_1^2 v_2^2$, a nonnegative multiple of the convex t^4, so each such restriction is convex. But f is not convex. On the segment from $(1, 0)$ to $(0, 1)$, a line that misses the origin, $f(1-t, t) = t^2(1-t)^2$ is 0 at both ends and $\frac{1}{16}$ at the midpoint. So the value at the midpoint exceeds the average 0 of the end values.

The Hessian test sees the same failure without guessing the segment. Here

$$\mathbf{H}_f(x, y) = \begin{pmatrix} 2y^2 & 4xy \\ 4xy & 2x^2 \end{pmatrix}, \qquad \det \mathbf{H}_f(x, y) = -12x^2y^2,$$

so at $(1, 1)$ the determinant is $-12 < 0$. The two eigenvalues of this symmetric matrix have product -12, so one is negative, and $\mathbf{H}_f(1, 1) \not\succeq \mathbf{O}$. Part (a) confirms that f is not convex.

18.10.6 Continuity

The warning about $|x|$ shows that convexity does not buy differentiability. It does buy continuity on an open set, and in finite dimension the proof is short: a bound from above near a point is

enough, and Jensen's inequality supplies one.

Theorem 18.10.7. Convex Functions Are Continuous

Let $U \subseteq \mathbb{R}^n$ be a nonempty open convex set and let $f\colon U \to \mathbb{R}$ be convex. Then f is continuous on U. More precisely, for every $\mathbf{a} \in U$ there are $\delta > 0$ and $L \geq 0$ with

$$|f(\mathbf{a}+\mathbf{h}) - f(\mathbf{a})| \leq L\|\mathbf{h}\| \qquad \text{whenever } \|\mathbf{h}\| \leq \delta.$$

Idea. ① Bound f above on a small ball around $\mathbf{a}$. The ball sits inside the "diamond" spanned by the $2n$ points $\mathbf{a} \pm \rho\mathbf{e}_i$, and Jensen bounds f on the diamond by its largest value at those points. ② Turn an upper bound into a two-sided linear bound. Walk from $\mathbf{a}$ toward $\mathbf{a}+\mathbf{h}$ and on to the edge of the ball: convexity makes the rise over the short step at most a fraction of the rise over the whole way. Then walk the other way for the lower bound.

Proof. Fix $\mathbf{a} \in U$. Since U is open, Proposition 18.1.6 (a) gives $r > 0$ such that every point within distance r of $\mathbf{a}$ lies in U. Put $\rho = r/2$. The $2n+1$ points $\mathbf{a}$ and $\mathbf{a} \pm \rho\mathbf{e}_i$ lie in U, so let M be the largest value of f at them.

Step 1: an upper bound. Let $\mathbf{h} = \sum_i h_i\mathbf{e}_i$ with $\|\mathbf{h}\|_1 \leq \rho$, and let $\sigma_i = 1$ if $h_i \geq 0$ and $\sigma_i = -1$ otherwise, so that $|h_i|\sigma_i = h_i$. Then

$$\mathbf{a}+\mathbf{h} = \sum_{i=1}^{n} \frac{|h_i|}{\rho}\,(\mathbf{a} + \rho\sigma_i\mathbf{e}_i) + \Big(1 - \frac{\|\mathbf{h}\|_1}{\rho}\Big)\mathbf{a},$$

a convex combination: the weights are nonnegative and add up to 1. By Theorem 18.10.2, $f(\mathbf{a}+\mathbf{h}) \leq M$. Put $\delta = \rho/\sqrt{n}$. If $\|\mathbf{h}\| \leq \delta$, then $\|\mathbf{h}\|_1 \leq \sqrt{n}\|\mathbf{h}\| \leq \rho$ by Proposition 16.1.2, so $f(\mathbf{a}+\mathbf{h}) \leq M$.

Step 2: the linear bound. Put $L = (M - f(\mathbf{a}))/\delta$, which is ≥ 0 because $\mathbf{a}$ is one of the points defining M. Let $0 < \|\mathbf{h}\| \leq \delta$; for $\mathbf{h} = \mathbf{0}$ there is nothing to prove. Put $s = \|\mathbf{h}\|/\delta \in (0,1]$ and $\mathbf{u} = \mathbf{h}/s$, so that $\|\mathbf{u}\| = \delta$ and $s\mathbf{u} = \mathbf{h}$. First,

$$\mathbf{a}+\mathbf{h} = (1-s)\mathbf{a} + s(\mathbf{a}+\mathbf{u}),$$

so convexity gives $f(\mathbf{a}+\mathbf{h}) \leq (1-s)f(\mathbf{a}) + sf(\mathbf{a}+\mathbf{u})$, that is,

$$f(\mathbf{a}+\mathbf{h}) - f(\mathbf{a}) \leq s(f(\mathbf{a}+\mathbf{u}) - f(\mathbf{a})) \leq s(M - f(\mathbf{a})),$$

using Step 1 at $\mathbf{u}$. Second,

$$\mathbf{a} = \frac{1}{1+s}(\mathbf{a}+\mathbf{h}) + \frac{s}{1+s}(\mathbf{a}-\mathbf{u}),$$

as one checks by multiplying out and using $s\mathbf{u} = \mathbf{h}$. Here $\mathbf{a}-\mathbf{u}$ lies in U, and convexity gives $(1+s)f(\mathbf{a}) \leq f(\mathbf{a}+\mathbf{h}) + sf(\mathbf{a}-\mathbf{u})$, that is,

$$f(\mathbf{a}) - f(\mathbf{a}+\mathbf{h}) \leq s(f(\mathbf{a}-\mathbf{u}) - f(\mathbf{a})) \leq s(M - f(\mathbf{a})),$$

using Step 1 at $-\mathbf{u}$. Together, $|f(\mathbf{a}+\mathbf{h}) - f(\mathbf{a})| \leq s(M - f(\mathbf{a})) = L\|\mathbf{h}\|$.

Step 3: continuity. If $\mathbf{x}_k \to \mathbf{a}$, then $\|\mathbf{x}_k - \mathbf{a}\| \le \delta$ for all large k, and for those k Step 2 gives $|f(\mathbf{x}_k) - f(\mathbf{a})| \le L\|\mathbf{x}_k - \mathbf{a}\| \to 0$. So f is continuous at $\mathbf{a}$, in the sequential sense of Chapter 16's introduction. As $\mathbf{a}$ was arbitrary, this proves the theorem.

■

By Theorem 16.2.2 the choice of the Euclidean norm is immaterial: every norm on $\mathbb{R}^n$ gives the same continuous functions. Both hypotheses of the theorem are needed.

Warning. **The domain must be open.** On the closed interval $[0,1]$, let $f = 0$ on $[0,1)$ and $f(1) = 1$. It is convex: the right side $tf(x) + (1-t)f(y)$ is never negative, and the left side is 0 unless $tx + (1-t)y = 1$, which for $0 < t < 1$ forces $x = y = 1$, where both sides equal 1. But $f(1 - 1/k) = 0 \not\to 1 = f(1)$. A convex function can jump at a boundary point of its domain.

Remark. **Finite dimension is needed too.** On the polynomials in $C[0,1]$, with the sup norm $\|p\|_\infty = \max_{t\in[0,1]}|p(t)|$ of Chapter 16 §01 restricted to them, the functional $\varphi(p) = p'(1)$ is linear, hence convex, on the whole space. The polynomials $p_k = x^{k^2}/k$ have $\|p_k\|_\infty = 1/k \to 0$ and $\varphi(p_k) = k \to \infty$, so φ is not continuous at 0. Step 1 is what fails: there is no finite set of points whose convex hull contains a ball.

The basic theory is now in place. A convex function is one with a convex epigraph; Jensen's inequality extends its defining inequality to every finite average; every supremum of affine functions is one, which accounts for the Ky Fan sums; the Hessian tests for it; and on an open set it is continuous. What it has not yet done is prove an inequality the book did not already know. §11 does that: with a logarithm constructed from the exponential, Jensen's inequality for $-\log$ gives the weighted arithmetic–geometric mean inequality, and from it come Young's and Hölder's inequalities, the triangle inequality for every p-norm, and the dual norms that §04 left open.

18.10.7 Exercises

18.10.7.1 A. Check your understanding

Exercise. A1

(a) Define "f is convex on C", and explain why C is required to be convex.

(b) Decide whether the following is correct, and justify your answer: if f and g are convex on C, then so are $f + g$ and $\max(f, g)$.

(c) Decide whether the following is correct, and justify your answer: if f and g are convex on $\mathbb{R}$, then so is their product fg.

(d) Decide whether the following is correct, and justify your answer: a convex function on $\mathbb{R}$ is differentiable.

(e) Decide whether the following is correct, and justify your answer: if $f\colon \mathbb{R}^2 \to \mathbb{R}$ is convex along every line through the origin, then f is convex.

(f) Name the facts of analysis that the proof of Theorem 18.10.6 uses.

Solution.

(a) Let C be a nonempty convex subset of a real vector space. Then $f\colon C \to \mathbb{R}$ is convex if $f(tx+(1-t)y) \le tf(x)+(1-t)f(y)$ for all $x, y \in C$ and all $t \in [0,1]$ (Definition 18.10.1). The left side evaluates f at $tx + (1-t)y$, and this point is in the domain for every choice exactly when C is convex.

(b) Correct. Adding the inequalities for f and for g gives the inequality for $f+g$. For the maximum, the remark after Proposition 18.10.3 applies: its proof works for any family of convex functions, and here the family is $\{f, g\}$.

(c) Incorrect. $f(x) = x$ is affine and $g(x) = x^2$ is convex (Example (First convex functions)), but $fg = x^3$ is not convex on $\mathbb{R}$: at -2, 0, and their midpoint -1 the nonexample after Example (First convex functions) gives $-1 > -4$.

(d) Incorrect. $|x|$ is convex and not differentiable at 0, as the first warning of the section shows.

(e) Incorrect. $f(x,y) = x^2y^2$ is convex along every line through the origin but not convex, since $f(\frac{1}{2}, \frac{1}{2}) = \frac{1}{16} > 0 = \frac{1}{2}f(1,0) + \frac{1}{2}f(0,1)$.

(f) Chapter 14 §06's assumptions (A1), the symmetry of the Hessian, and (A2), Taylor's theorem with a second-order remainder; facts (A3) and (A4) of Chapter 16's introduction, compactness of $[0,1]$ and the extreme value theorem, in the maximum principle; and the algebra of limits, in the form "a nonstrict inequality survives a limit".

18.10.7.2 B. Practice

Exercise. B1: Which are convex?

Determine which of the following functions are convex on the given domain. Justify your answer.

(a) $f(x,y) = 2x^2 - 2xy + y^2$ on $\mathbb{R}^2$.

(b) $f(x,y) = x^2 + 4xy + y^2$ on $\mathbb{R}^2$.

(c) $f(x) = \max(2x+1,\ -x,\ 3)$ on $\mathbb{R}$.

(d) $g(x) = 1/x$ on $(0, \infty)$.

(e) $g(x) = 1/x$ on $\{x \in \mathbb{R} : x \neq 0\}$.

Solution.

(a) Convex, indeed strictly convex. f is a polynomial, so twice continuously differentiable, with constant Hessian $\mathbf{H} = \left(\begin{smallmatrix} 4 & -2 \\ -2 & 2 \end{smallmatrix}\right)$, which is symmetric. For $\mathbf{v} = (v_1, v_2)$,

$$\mathbf{v}^{\top}\mathbf{H}\mathbf{v} = 4v_1^2 - 4v_1v_2 + 2v_2^2 = 2\big(v_1^2 + (v_1 - v_2)^2\big),$$

which is positive unless $v_1 = 0$ and $v_1 = v_2$, that is, unless $\mathbf{v} = \mathbf{0}$. So $\mathbf{H} \succ \mathbf{O}$ at every point, and Theorem 18.10.6 (b) applies.

(b) Not convex. Take $\mathbf{x} = (1,-1)$ and $\mathbf{y} = (-1,1)$, with midpoint $\mathbf{0}$. Then $f(\mathbf{x}) = 1 - 4 + 1 = -2 = f(\mathbf{y})$, while $f(\mathbf{0}) = 0 > \frac{1}{2}(-2) + \frac{1}{2}(-2) = -2$. The inequality of Definition 18.10.1 fails with $t = \frac{1}{2}$.

(c) Convex. The three functions $2x+1$, $-x$, and the constant 3 are affine, and f is their

maximum, so Proposition 18.10.3 applies.

(d) Strictly convex. For $x > 0$ and small $h \neq 0$,

$$\frac{1}{h}\Big(\frac{1}{x+h} - \frac{1}{x}\Big) = -\frac{1}{x(x+h)} \longrightarrow -\frac{1}{x^2},$$

so $g'(x) = -1/x^2$. If $0 < a < b$ then $1/a^2 > 1/b^2$, so $-1/a^2 < -1/b^2$: g' is strictly increasing on $(0, \infty)$, and Lemma 18.10.5 gives strict convexity.

(e) The question has no answer: Definition 18.10.1 requires a convex domain, and this one is not. It contains -1 and 1 but not their midpoint 0, so the defining inequality is undefined for $x = -1$, $y = 1$, $t = \frac{1}{2}$.

18.10.7.3 C. Going deeper

Exercise. C1: Subadditive and homogeneous

Let $g\colon V \to \mathbb{R}$ be **positively homogeneous**: $g(tx) = tg(x)$ for every $x \in V$ and every real $t \geq 0$.

(a) Prove that g is convex if and only if $g(x+y) \leq g(x) + g(y)$ for all $x, y \in V$.

(b) Deduce that every norm on $\mathbb{R}^n$ is convex, and that Corollary 17.6.5 and Corollary 18.10.4 are equivalent statements about s_k.

Solution.

(a) ($\Rightarrow$) Let g be convex. By homogeneity with $t = 2$ and then convexity with $t = \frac{1}{2}$,

$$g(x+y) = 2g\big(\tfrac{1}{2}x + \tfrac{1}{2}y\big) \leq 2\big(\tfrac{1}{2}g(x) + \tfrac{1}{2}g(y)\big) = g(x) + g(y).$$

($\Leftarrow$) Let g be subadditive, and let $0 \leq t \leq 1$. By subadditivity and then homogeneity with the nonnegative numbers t and $1-t$,

$$g\big(tx + (1-t)y\big) \leq g(tx) + g\big((1-t)y\big) = tg(x) + (1-t)g(y).$$

(b) A norm is positively homogeneous by (N2) of Definition 16.1.1, since $|t| = t$ for $t \geq 0$, and subadditive by (N3); so it is convex by (a). This is a second proof of Example (First convex functions) (b). The function s_k is positively homogeneous on the Hermitian matrices, since multiplying by $t \geq 0$ multiplies each eigenvalue by t and keeps their order. So by (a), subadditivity of s_k, which is Corollary 17.6.5, holds exactly when s_k is convex, which is Corollary 18.10.4.

Exercise. C2: Tangent lines

Let $I \subseteq \mathbb{R}$ be an open interval and let $g\colon I \to \mathbb{R}$ be differentiable and convex.

(a) Prove that $g(y) \geq g(x) + g'(x)(y-x)$ for all $x, y \in I$.

(b) Deduce that $g(y) = \max_{x \in I}\big(g(x) + g'(x)(y-x)\big)$ for every $y \in I$: a differentiable convex function is the maximum of its tangent lines.

Hint: in (a), apply the definition to the points x and y with weight s on y, and let $s \to 0$.

Solution.

(a) If $y = x$ both sides are equal. Let $y \neq x$ and $0 < s \leq 1$. Convexity with weight s on y gives $g(x + s(y - x)) \leq (1 - s)g(x) + sg(y)$, that is,

$$(y - x) \cdot \frac{g(x + s(y - x)) - g(x)}{s(y - x)} \;\leq\; g(y) - g(x).$$

The fraction is a difference quotient of g at x with step $s(y - x) \neq 0$. Take $s = 1/k$; the step tends to 0, so the fraction tends to $g'(x)$. A nonstrict inequality survives the limit, so $(y - x)g'(x) \leq g(y) - g(x)$, as claimed.

(b) For fixed y, each $\ell_x(y) = g(x) + g'(x)(y - x)$ is at most $g(y)$ by (a), and $\ell_y(y) = g(y)$. So the maximum over x exists, is attained at $x = y$, and equals $g(y)$. Each ℓ_x is an affine function of y, so this is the converse of Proposition 18.10.3, for differentiable functions.

Exercise. C3: Where a convex function is largest

(a) Let $f\colon \mathbb{R}^n \to \mathbb{R}$ be convex, and let P be the convex hull $\operatorname{conv}\{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$. Prove that $\max_{\mathbf{x} \in P} f(\mathbf{x}) = \max_i f(\mathbf{x}_i)$.

(b) Let $f(x, y) = x^2 - xy + y^2$ and let P be the triangle $\operatorname{conv}\{(2, 0), (0, 2), (-2, -2)\}$. Find the maximum and the minimum of f on P, and say where each is attained.

Solution.

(a) Each $\mathbf{x}_i$ lies in P, so the maximum over P, if it exists, is at least $\max_i f(\mathbf{x}_i)$. Conversely, let $\mathbf{x} \in P$. By Theorem 18.1.3, $\mathbf{x} = \sum_i w_i \mathbf{x}_i$ for some weights $w_i \geq 0$ with sum 1, and Theorem 18.10.2 gives $f(\mathbf{x}) \leq \sum_i w_i f(\mathbf{x}_i) \leq \max_i f(\mathbf{x}_i)$. So $\max_i f(\mathbf{x}_i)$ is an upper bound for f on P that is attained, hence the maximum. Compare Corollary 18.8.6, which says the same for a linear function on any compact convex set, with extreme points in place of the $\mathbf{x}_i$.

(b) The Hessian is the constant symmetric matrix $\left(\begin{smallmatrix} 2 & -1 \\ -1 & 2 \end{smallmatrix}\right)$, and $2v_1^2 - 2v_1v_2 + 2v_2^2 = v_1^2 + v_2^2 + (v_1 - v_2)^2 \geq 0$, so it is positive semidefinite and f is convex by Theorem 18.10.6 (a). By (a) the maximum is the largest vertex value: $f(2, 0) = 4$, $f(0, 2) = 4$, and $f(-2, -2) = 4 - 4 + 4 = 4$. So the maximum is 4, attained at all three vertices. For the minimum, $f(x, y) = \frac{1}{2}(x^2 + y^2 + (x - y)^2) \geq 0$, with equality only at $(0, 0)$. And $(0, 0) = \frac{1}{3}(2, 0) + \frac{1}{3}(0, 2) + \frac{1}{3}(-2, -2)$ lies in P. So the minimum is 0, attained only at the interior point $(0, 0)$. The maximum of a convex function sits at vertices; its minimum need not.

18.11 Three Classical Inequalities

§10 built the basic theory of convex functions. This section spends it on three classical inequalities, each obtained from the one before. Jensen's inequality for the logarithm is the weighted arithmetic–geometric mean inequality; with two weights it is Young's inequality; and Young's

inequality, applied coordinate by coordinate, is Hölder's. Hölder's inequality then settles two questions the book left open: the triangle inequality for the p-norm, which Chapter 16 §01 recorded without proof, and the dual of the p-norm, which §04 computed only for $p = 1, 2, \infty$.

Throughout, the field is $\mathbb{R}$, and "increasing" and "strictly increasing" are meant as in §10.

18.11.1 The logarithm and real powers

The first inequality needs a logarithm, and the book has none. It has the exponential: by fact (A1) of Chapter 10 §09, for real s the series $\sum_i s^i/i!$ converges to a real number e^s, with $e^{s+u} = e^s e^u$, $e^0 = 1$, and $e^s \geq 1$ for $s \geq 0$; and $\frac{d}{dt}e^t = e^t$ is the 1×1 case of Theorem 10.9.4 (b). We construct the logarithm as its inverse.

Lemma 18.11.1. Exponential and Logarithm

(a) $e^s > 0$ for every real s, and $s \mapsto e^s$ is strictly increasing and strictly convex on $\mathbb{R}$.

(b) For every real $y > 0$ there is exactly one real s with $e^s = y$; write $s = \log y$.

(c) $\log\colon (0, \infty) \to \mathbb{R}$ is strictly increasing and **strictly concave**, with $\log(xy) = \log x + \log y$ and $\log 1 = 0$.

Idea. The exponential is positive and is its own derivative, so it is strictly increasing with a strictly increasing derivative, hence strictly convex. The intermediate value theorem makes it take every positive value, so it has an inverse. Applying that increasing inverse to the convexity inequality of exp turns it into concavity of log.

Proof.

(a) If $s \geq 0$ then $e^s \geq 1 > 0$. If $s < 0$ then $e^{-s} \geq 1$, and $e^s e^{-s} = e^0 = 1$, so $e^s = 1/e^{-s} > 0$. The derivative e^s is therefore positive everywhere, and by the linearity of differentiation, fact (A3) of Chapter 10 §09, $-\exp$ has the negative derivative $-\exp$. By fact (A6) of Chapter 16's introduction, $-\exp$ is strictly decreasing, so exp is strictly increasing. So $g = \exp$ has a strictly increasing derivative, and Lemma 18.10.5 makes it strictly convex.

(b) Uniqueness holds because exp is strictly increasing. For existence, note first that exp is continuous: $e^{t+h} - e^t = e^t\, h \cdot (e^h - 1)/h$, and $(e^h - 1)/h \to 1$ is the derivative at 0, so the right side tends to 0. Let $y \geq 1$. For $s \geq 0$ every partial sum of the series with at least two terms is at least $1 + s$, so $e^s \geq 1 + s$, since a nonstrict inequality survives a limit. Hence $e^y \geq 1 + y > y$, while $e^0 = 1 \leq y$. The intermediate value theorem, fact (A5) of Chapter 16's introduction, gives $s \in [0, y]$ with $e^s = y$. If $0 < y < 1$, apply this to $1/y > 1$ to get $e^{s'} = 1/y$; then $e^{-s'} = y$.

(c) If $0 < x < y$ and $\log x \geq \log y$, then $x = e^{\log x} \geq e^{\log y} = y$, since exp is increasing, which is false. So log is strictly increasing. Next, $e^{\log x + \log y} = e^{\log x} e^{\log y} = xy$, so $\log(xy) = \log x + \log y$ by uniqueness in (b), and $e^0 = 1$ gives $\log 1 = 0$. For concavity, let $x, y > 0$ and $0 \leq t \leq 1$, and put $a = \log x$, $b = \log y$. By (a),

$$e^{ta + (1-t)b} \leq te^a + (1-t)e^b = tx + (1-t)y.$$

Applying the increasing function log to both sides gives $t \log x + (1-t)\log y \leq \log(tx + (1-t)y)$. If $x \neq y$ and $0 < t < 1$, then $a \neq b$, the first inequality is strict by strict

convexity, and log keeps it strict. So $-\log$ is strictly convex.

■

Real powers. For $t > 0$ and real r put $t^r = e^{r \log t}$, and $0^r = 0$ for $r > 0$. For a positive integer r this is the usual power, since $e^{r \log t} = (e^{\log t})^r = t^r$ by $e^{s+u} = e^s e^u$. For $t, a, b > 0$ and real r, u, the rules $e^{s+u} = e^s e^u$ and $\log(ab) = \log a + \log b$, with $\log(t^r) = r \log t$ from the uniqueness in (b), give

$$t^r t^u = t^{r+u}, \qquad (t^r)^u = t^{ru}, \qquad (ab)^r = a^r b^r, \qquad (a/b)^r = a^r / b^r.$$

The first three extend to $t, a, b \geq 0$, and the last to $a \geq 0$, when the exponents r and u are positive: then a zero base makes both sides 0. They do not extend further, since 0^r is not defined for $r \leq 0$. Finally, for $r > 0$ the function $t \mapsto t^r$ is strictly increasing on $[0, \infty)$: on $(0, \infty)$ it is exp applied to the strictly increasing $r \log t$, and $0^r = 0 < t^r$ for $t > 0$. This is the meaning of $|x_i|^p$ and of the exponent $1/p$ in the p-norm $\|\mathbf{x}\|_p = (\sum_i |x_i|^p)^{1/p}$ of Chapter 16 §01.

18.11.2 Arithmetic and geometric means

Chapter 17 §08 proved Lemma 17.8.6, the arithmetic–geometric mean inequality, by pure algebra: an induction that merges a number above the mean with one below it. Concavity of the logarithm gives a second proof, and a stronger statement, with arbitrary weights.

Theorem 18.11.2. Weighted Arithmetic–Geometric Mean Inequality

Let $t_1, \ldots, t_m \geq 0$, and let $w_1, \ldots, w_m > 0$ with $w_1 + \cdots + w_m = 1$. Then

$$t_1^{w_1} t_2^{w_2} \cdots t_m^{w_m} \;\leq\; w_1 t_1 + w_2 t_2 + \cdots + w_m t_m,$$

with equality if and only if $t_1 = t_2 = \cdots = t_m$.

Idea. Take logarithms. The product becomes the weighted average $\sum_i w_i \log t_i$, and the inequality becomes Jensen's inequality for the concave log. Then exponentiate back.

Proof. If some $t_j = 0$, the left side is 0, since $0^{w_j} = 0$, and the right side is ≥ 0. Equality then means $\sum_i w_i t_i = 0$, which with every $w_i > 0$ and $t_i \geq 0$ forces every $t_i = 0$; conversely, if every $t_i = 0$, both sides are 0.

So let every $t_i > 0$. By Lemma 18.11.1 (c), $-\log$ is strictly convex on $(0, \infty)$, so Theorem 18.10.2 gives

$$\sum_{i=1}^{m} w_i \log t_i \;\leq\; \log\Big(\sum_{i=1}^{m} w_i t_i\Big),$$

with equality only when all t_i are equal. Apply the strictly increasing function exp. The right side becomes $\sum_i w_i t_i$, and the left side becomes $\prod_i e^{w_i \log t_i} = \prod_i t_i^{w_i}$, by $e^{s+u} = e^s e^u$ and the definition of real powers. A strictly increasing function preserves both the inequality and its strictness, so equality holds exactly when all t_i are equal.

■

The case of equal weights is Lemma 17.8.6. With $w_i = 1/n$, the rules for powers turn

the theorem into $(t_1 \cdots t_n)^{1/n} \le \frac{1}{n}\sum_i t_i$. Since $u \mapsto u^n$ is strictly increasing on $[0,\infty)$, raising both sides to the n-th power gives $t_1 \cdots t_n \le (\frac{1}{n}\sum_i t_i)^n$, with the same equality case. That is Lemma 17.8.6, now with arbitrary weights and seen as Jensen's inequality for the logarithm.

With two weights, the theorem becomes an inequality about products that is exactly what Hölder's inequality needs.

Lemma 18.11.3. Young's Inequality

Let $p, q > 1$ with $\frac{1}{p} + \frac{1}{q} = 1$. Then for all real $a, b \ge 0$,

$$ab \;\le\; \frac{a^p}{p} + \frac{b^q}{q},$$

with equality if and only if $a^p = b^q$.

Proof. Apply Theorem 18.11.2 with $m = 2$, $t_1 = a^p$, $t_2 = b^q$, $w_1 = 1/p$, and $w_2 = 1/q$. The weights are positive and add up to 1, and $(a^p)^{1/p} = a$, $(b^q)^{1/q} = b$ by the rules for powers. So the left side of the theorem is ab and the right side is $a^p/p + b^q/q$, with equality exactly when $a^p = b^q$. ∎

18.11.3 Hölder's inequality and the dual p-norms

Two exponents $p, q \in (1,\infty)$ with $\frac{1}{p} + \frac{1}{q} = 1$ are called **conjugate**; equivalently $q = p/(p-1)$. The pair $p = q = 2$ is conjugate to itself. Hölder's inequality is the Cauchy–Schwarz inequality with the two factors measured in two different norms.

Theorem 18.11.4. Hölder's Inequality

Let $p, q \in (1,\infty)$ be conjugate exponents. For all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$,

$$|\langle \mathbf{x}, \mathbf{y} \rangle| \;\le\; \sum_{i=1}^{n} |x_i y_i| \;\le\; \|\mathbf{x}\|_p \|\mathbf{y}\|_q.$$

Idea. Both sides of the second inequality scale the same way when $\mathbf{x}$ or $\mathbf{y}$ is multiplied by a positive number. So normalize to $\|\mathbf{x}\|_p = \|\mathbf{y}\|_q = 1$. Then apply Young's inequality coordinate by coordinate and add: the right-hand sides add up to $\frac{1}{p} + \frac{1}{q} = 1$.

Proof. The first inequality is the triangle inequality for numbers (Theorem 0.6.4) applied to $\sum_i x_i y_i$. For the second, if $\mathbf{x} = \mathbf{0}$ or $\mathbf{y} = \mathbf{0}$ both sides are 0. Otherwise $\|\mathbf{x}\|_p > 0$ and $\|\mathbf{y}\|_q > 0$; put $a_i = |x_i|/\|\mathbf{x}\|_p$ and $b_i = |y_i|/\|\mathbf{y}\|_q$. Then $\sum_i a_i^p = \sum_i |x_i|^p / \|\mathbf{x}\|_p^p = 1$, because $\|\mathbf{x}\|_p^p = \sum_i |x_i|^p$ by the rules for powers, and likewise $\sum_i b_i^q = 1$. By Lemma 18.11.3 for each i, and adding,

$$\sum_{i=1}^{n} a_i b_i \;\le\; \frac{1}{p}\sum_{i=1}^{n} a_i^p + \frac{1}{q}\sum_{i=1}^{n} b_i^q = \frac{1}{p} + \frac{1}{q} = 1.$$

Multiplying by $\|\mathbf{x}\|_p\|\mathbf{y}\|_q$ gives $\sum_i|x_iy_i| \le \|\mathbf{x}\|_p\|\mathbf{y}\|_q$, as claimed.

■

For $p = q = 2$ this is Theorem 11.1.5 for the standard inner product on $\mathbb{R}^n$, reached without the inner product's geometry. The limiting pair $p = 1$, $q = \infty$ holds too, and more simply: $\sum_i|x_iy_i| \le \sum_i|x_i|\|\mathbf{y}\|_\infty$.

Chapter 16 §01 recorded the p-norm for every $p \ge 1$, but verified only $p = 1, 2, \infty$, because the triangle inequality for general p had not been proved. Hölder's inequality proves it.

Corollary 18.11.5. Minkowski's Inequality

Let $1 < p < \infty$. For all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, $\|\mathbf{x}+\mathbf{y}\|_p \le \|\mathbf{x}\|_p + \|\mathbf{y}\|_p$. Consequently $\|\cdot\|_p$ is a norm on $\mathbb{R}^n$.

Idea. Split $|z_i|^p = |z_i| \cdot |z_i|^{p-1}$ and use the triangle inequality on the first factor. That leaves two sums, and Hölder's inequality bounds each, with the same second factor $(|z_i|^{p-1})_i$. Dividing by that common factor's norm leaves the triangle inequality.

Proof. Let $q = p/(p-1)$ be the conjugate exponent, and put $\mathbf{z} = \mathbf{x} + \mathbf{y}$. If $\mathbf{z} = \mathbf{0}$ there is nothing to prove. Otherwise, by the triangle inequality for numbers in each coordinate (Theorem 0.6.4),

$$\begin{aligned}\sum_i |z_i|^p &= \sum_i |z_i|\,|z_i|^{p-1} \\ &\le \sum_i |x_i|\,|z_i|^{p-1} + \sum_i |y_i|\,|z_i|^{p-1}.\end{aligned}$$

Apply Theorem 18.11.4 to each sum, with the vector $\mathbf{w} = (|z_1|^{p-1}, \dots, |z_n|^{p-1})$ as the second factor. Since $(p-1)q = p$,

$$\|\mathbf{w}\|_q = \Big(\sum_i |z_i|^p\Big)^{1/q} = \|\mathbf{z}\|_p^{p/q}.$$

So $\|\mathbf{z}\|_p^p \le (\|\mathbf{x}\|_p + \|\mathbf{y}\|_p)\|\mathbf{z}\|_p^{p/q}$. Divide by $\|\mathbf{z}\|_p^{p/q} > 0$, and use $p - p/q = p(1-1/q) = 1$, to get $\|\mathbf{z}\|_p \le \|\mathbf{x}\|_p + \|\mathbf{y}\|_p$. This is (N3) of Definition 16.1.1. (N1) holds because $\|\mathbf{x}\|_p = 0$ forces every $|x_i|^p = 0$, and (N2) because $\big(\sum_i |c|^p|x_i|^p\big)^{1/p} = |c|\|\mathbf{x}\|_p$ by the rules for powers.

■

The proof uses the coordinates only through their moduli: the triangle inequality $|x_i + y_i| \le |x_i| + |y_i|$, which Theorem 0.6.4 proves for complex numbers as well, and Hölder's inequality for the real vectors $(|x_i|)_i$, $(|y_i|)_i$, and $\mathbf{w}$, whose p-norms are those of $\mathbf{x}$ and $\mathbf{y}$. So it works verbatim on $\mathbb{C}^n$, and $\|\cdot\|_p$ is a norm on F^n for $F = \mathbb{R}$ and for $F = \mathbb{C}$, as Chapter 16 §01 recorded.

Now the dual norm. Recall from Definition 18.4.4 that the dual norm of $\|\cdot\|$ at $\mathbf{y}$ is the largest value of $\langle \mathbf{x}, \mathbf{y}\rangle$ over the unit ball $\|\mathbf{x}\| \le 1$. §04 computed it for $p = 1, 2, \infty$: the 1-norm and the ∞-norm are dual to each other, and the 2-norm is dual to itself. Hölder's inequality supplies an upper bound for every other p, and a matching vector shows it is attained.

Theorem 18.11.6. The Dual of the p-Norm

Let $p, q \in (1, \infty)$ be conjugate exponents. On $\mathbb{R}^n$ with the standard inner product, the dual norm of $\|\cdot\|_p$ is $\|\cdot\|_q$:

$$\max\{\langle \mathbf{x}, \mathbf{y}\rangle : \|\mathbf{x}\|_p \le 1\} = \|\mathbf{y}\|_q \qquad (\mathbf{y} \in \mathbb{R}^n).$$

For $\mathbf{y} \neq \mathbf{0}$ the maximum is attained at $\mathbf{x}$ with $x_i = \sigma_i |y_i|^{q-1} / \|\mathbf{y}\|_q^{q-1}$, where $\sigma_i = 1$ if $y_i \ge 0$ and $\sigma_i = -1$ otherwise.

Idea. Hölder's inequality bounds $\langle \mathbf{x}, \mathbf{y}\rangle$ by $\|\mathbf{y}\|_q$ on the whole unit ball. To show the bound is attained, give x_i the sign of y_i and make $|x_i|^p$ proportional to $|y_i|^q$, which is when Young's inequality (Lemma 18.11.3) is an equality in each coordinate. That is the stated witness, and a direct computation checks it.

Proof. If $\|\mathbf{x}\|_p \le 1$, then $\langle \mathbf{x}, \mathbf{y}\rangle \le \|\mathbf{x}\|_p \|\mathbf{y}\|_q \le \|\mathbf{y}\|_q$ by Theorem 18.11.4. For $\mathbf{y} = \mathbf{0}$ the bound 0 is attained at $\mathbf{x} = \mathbf{0}$. For $\mathbf{y} \neq \mathbf{0}$, take $\mathbf{x}$ as in the statement. Since $(q-1)p = q$,

$$\sum_i |x_i|^p = \frac{\sum_i |y_i|^q}{\|\mathbf{y}\|_q^q} = 1,$$

so $\|\mathbf{x}\|_p = 1$; and since $\sigma_i y_i = |y_i|$,

$$\langle \mathbf{x}, \mathbf{y}\rangle = \frac{\sum_i |y_i|^q}{\|\mathbf{y}\|_q^{q-1}} = \frac{\|\mathbf{y}\|_q^q}{\|\mathbf{y}\|_q^{q-1}} = \|\mathbf{y}\|_q.$$

So the bound is attained, and the largest value is $\|\mathbf{y}\|_q$.

■

Example. A dual norm by hand

Let $\mathbf{y} = (3, -4, 5) \in \mathbb{R}^3$, and let $p = \frac{3}{2}$ and $q = 3$. Compute the dual norm of $\mathbf{y}$ with respect to $\|\cdot\|_{3/2}$, find a vector $\mathbf{x}$ with $\|\mathbf{x}\|_{3/2} = 1$ at which it is attained, and compare with the value at the unit vector $\mathbf{e}_3$.

Solution. The exponents are conjugate, since $\frac{1}{p} + \frac{1}{q} = \frac{2}{3} + \frac{1}{3} = 1$. Next, $|3|^3 + |-4|^3 + |5|^3 = 27 + 64 + 125 = 216 = 6^3$, so $\|\mathbf{y}\|_3 = 216^{1/3} = 6$, using $(6^3)^{1/3} = 6$. By Theorem 18.11.6 the dual norm is 6.

For the maximizer, $q - 1 = 2$, so $x_i = \sigma_i |y_i|^2 / 6^2$ with σ_i the sign of y_i:

$$\mathbf{x} = \left(\tfrac{9}{36}, -\tfrac{16}{36}, \tfrac{25}{36}\right) = \left(\tfrac{1}{4}, -\tfrac{4}{9}, \tfrac{25}{36}\right).$$

Check: $|x_i|^{3/2} = (|y_i|^2/36)^{3/2} = |y_i|^3/216$, giving $\frac{1}{8} + \frac{8}{27} + \frac{125}{216} = \frac{27+64+125}{216} = 1$. So $\|\mathbf{x}\|_{3/2} = 1$. And $\langle \mathbf{x}, \mathbf{y}\rangle = \frac{3}{4} + \frac{16}{9} + \frac{125}{36} = \frac{27+64+125}{36} = 6$.

The unit vector $\mathbf{e}_3$ has $\|\mathbf{e}_3\|_{3/2} = 1$ and gives $\langle \mathbf{e}_3, \mathbf{y}\rangle = 5 < 6$. Putting all the weight on the

largest coordinate of $\mathbf{y}$ is not optimal; the maximizer spreads it in proportion to $|y_i|^2$.

Exchanging the roles of p and q shows that the dual of $\|\cdot\|_q$ is $\|\cdot\|_p$, just as Theorem 18.4.8 requires. Together with §04, the dual norm of $\|\cdot\|_p$ is now known for every $p \in [1, \infty]$: it is $\|\cdot\|_q$ with $\frac{1}{p} + \frac{1}{q} = 1$, reading $\frac{1}{\infty}$ as 0. This completes §04's computation. Chapter 16 measured vectors and matrices but never a linear functional; §04's dual norm filled that gap, and now it is computed for every p-norm.

Look back at the chain. One convexity fact, that $-\log$ is strictly convex, went through Jensen's inequality to the weighted arithmetic–geometric mean inequality, through two weights to Young's inequality, and through normalization to Hölder's. Hölder's inequality then did two jobs that earlier chapters could not: it proved that every $\|\cdot\|_p$ is a norm, so that its unit ball is one of the convex sets this chapter began with, and it computed the dual of every such norm.

18.11.4 Summary and transfer

The chapter began with a word the book had been using without studying it. Convexity turned up in the positive semidefinite cone of Chapter 13, in the unit balls of Chapter 16, and in the Ky Fan sums of Chapter 17. The chapter ends with convexity as a subject: convex sets, the hyperplanes that separate them, the duality between a cone and its dual, extreme points and polytopes, and convex functions. Six moves did the work, and each transfers beyond this chapter.

- **Nearest point, then the normal direction.** Project onto a closed convex set; the vector from the projection to the point is the normal of a separating hyperplane (§03, Separating Hyperplanes). This is Chapter 11's best approximation, with an inequality where the equation was. *Transfer:* to prove that a point lies outside a closed convex set, and to certify it, look for the nearest point.
- **Count, then drop a point.** Too many points in n dimensions are affinely dependent, and moving the weights along the dependence removes one of them without leaving the hull (§02, Carathéodory's Theorem). Among too many generators of a cone, a linear dependence does the same (§05, Cones and Farkas's Lemma). *Transfer:* when a representation uses more pieces than the dimension allows, the surplus is a dependence, and a dependence can be spent.
- **Dualize.** A closed convex cone is determined by its dual cone, and every statement about the one has a statement about the other. Farkas's lemma is this move made into a theorem, linear programming duality (§06, Linear Programming Duality) is Farkas applied to one combined system, and the minimax theorem (§07, Matrix Games and the Minimax Theorem) is a duality statement of the same kind. *Transfer:* when a feasibility question will not yield, ask for the certificate that it fails.
- **Lift, then eliminate.** A hull of finitely many points is the projection of a polyhedron in a bigger space, with the weights as extra coordinates, and Fourier–Motzkin elimination removes those coordinates one at a time while keeping a system of inequalities (§09, Polytopes, Theorem 18.9.5). *Transfer:* to describe the image of a set given by inequalities, add the hidden variables as coordinates and eliminate them.
- **A supremum of linear functions is convex.** Support functions, dual norms, λ_1, and the Ky Fan sums are convex for this one reason (§04, Gauges, Support Functions, and Dual Norms, and §10, Convex Functions, Proposition 18.10.3). Its companion: a linear function on a compact convex set peaks at an extreme point (§08, Extreme Points). *Transfer:* to prove a function convex, write it as a maximum of affine functions.
- **Compactness turns a supremum into a maximum.** Once a hull, a ball, or a set of

weights is known to be compact, facts (A3) and (A4) of Chapter 16's introduction produce a nearest point, a maximizer, or an extreme point that can be named and used. *Transfer:* before optimizing, check that the set is compact; afterwards, feed the maximizer back into the argument.

§§10–11 added two moves of their own. **Second differences**: on a segment, convexity is read off from $f(\mathbf{z}+s\mathbf{v})+f(\mathbf{z}-s\mathbf{v})-2f(\mathbf{z})$ (§10, Theorem 18.10.6). **Normalize, then use a two-point inequality**: Hölder's inequality is Young's inequality after scaling both vectors to norm 1 (§11, Theorem 18.11.4). *Transfer:* if multiplying any one variable by $c>0$ multiplies both sides by the same power of c, check the cases with a zero variable directly, and then assume every variable has norm 1.

The chapter also paid promises made in earlier chapters.

The promise	Made in	Paid by
"Chapter 18 studies convex functions in general, and the observation that a pointwise supremum of linear functions is convex is one of its basic facts"	Chapter 17 §06	§10 (Convex Functions), Proposition 18.10.3, with λ_1 and the Ky Fan sums as its instances (Corollary 18.10.4)
Nonnegativity and capacities turn traffic flow "into one about inequalities, which is the subject of linear programming"	Chapter 3 §07	§06 (Linear Programming Duality)
A way to measure a linear functional: Chapter 16 measured vectors and matrices, but never a functional	Chapter 16	§04 (Gauges, Support Functions, and Dual Norms): the definition, the double dual, and $p=1,2,\infty$; §11 (Three Classical Inequalities): every $1<p<\infty$, Theorem 18.11.6
The triangle inequality for $\|\cdot\|_p$, recorded without proof, "which Chapter 18 proves from Hölder's inequality"	Chapter 16 §01	§11 (Three Classical Inequalities), Corollary 18.11.5

Some things the chapter did not do. It stated Birkhoff's theorem, that the extreme points of the doubly stochastic matrices are exactly the permutation matrices, and left its proof to Chapter 19. The convexity of $\mathbf{A}\mapsto \operatorname{tr} f(\mathbf{A})$ for convex f, and the consequences of a majorization for every convex function, belong to Chapter 21. Convex optimization beyond linear programs, where a duality gap can occur, lies outside this chapter.

18.11.5 Exercises

18.11.5.1 A. Check your understanding

Exercise. A1

(a) State Young's inequality, with its hypotheses and its equality case.

(b) Decide whether the following is correct, and justify your answer: the exponent conjugate to $p = 4$ is $q = \frac{4}{3}$.

(c) Decide whether the following is correct, and justify your answer: $(t^r)^u = t^{ru}$ for every real $t \geq 0$ and all real r, u.

(d) Decide whether the following is correct, and justify your answer: if p, q are conjugate and $\langle \mathbf{x}, \mathbf{y} \rangle = \|\mathbf{x}\|_p \|\mathbf{y}\|_q$ for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, then one of $\mathbf{x}, \mathbf{y}$ is a scalar multiple of the other.

(e) Name the convex function to which the proof of Theorem 18.11.2 applies Jensen's inequality, and say why the equality case needs it to be **strictly** convex.

Solution.

(a) Let $p, q > 1$ with $\frac{1}{p} + \frac{1}{q} = 1$. Then $ab \leq a^p/p + b^q/q$ for all real $a, b \geq 0$, with equality if and only if $a^p = b^q$ (Lemma 18.11.3).

(b) Correct. $\frac{1}{4} + \frac{3}{4} = 1$; equivalently $q = p/(p-1) = 4/3$.

(c) Incorrect. Take $t = 0$, $r = 1$, and $u = -1$. Then $t^r = 0$, and 0^{-1} is not defined, since 0^r is defined only for $r > 0$; neither side has a value. The rule holds for $t > 0$ and all real r, u, and for $t = 0$ when $r, u > 0$.

(d) Incorrect. Take $p = 3$, $q = \frac{3}{2}$, $\mathbf{x} = (1, 2)$, and $\mathbf{y} = (1, 4)$. Then $\langle \mathbf{x}, \mathbf{y} \rangle = 1 + 8 = 9$, $\|\mathbf{x}\|_3 = (1+8)^{1/3} = 9^{1/3}$, and $\|\mathbf{y}\|_{3/2} = (1 + 4^{3/2})^{2/3} = 9^{2/3}$, since $4^{3/2} = 8$; so $\|\mathbf{x}\|_3 \|\mathbf{y}\|_{3/2} = 9^{1/3+2/3} = 9$. But neither vector is a multiple of the other, since $1 \cdot 4 - 2 \cdot 1 \neq 0$. Equality asks for $|x_i|^p$ to be proportional to $|y_i|^q$, not $\mathbf{x}$ to $\mathbf{y}$ (exercise C2).

(e) It is $-\log$ on $(0, \infty)$, which is strictly convex by Lemma 18.11.1 (c). Theorem 18.10.2 proves its equality clause only for strictly convex functions, and the hypothesis is needed: an affine function makes Jensen's inequality an equality for every choice of points.

18.11.5.2 B. Practice

Exercise. B1: A dual norm by hand

Let $\mathbf{y} = (1, -6, 8) \in \mathbb{R}^3$, and let $p = \frac{3}{2}$ and $q = 3$.

(a) Check that p and q are conjugate, and compute $\|\mathbf{y}\|_3$.

(b) Find $\mathbf{x} \in \mathbb{R}^3$ with $\|\mathbf{x}\|_{3/2} = 1$ and $\langle \mathbf{x}, \mathbf{y} \rangle = \|\mathbf{y}\|_3$.

(c) Hence determine the dual norm of $\mathbf{y}$ with respect to $\|\cdot\|_{3/2}$, and compare it with the value $\langle \mathbf{e}_3, \mathbf{y} \rangle$ at the unit vector $\mathbf{e}_3$.

Solution.

(a) $\frac{1}{p} + \frac{1}{q} = \frac{2}{3} + \frac{1}{3} = 1$. Next, $|1|^3 + |-6|^3 + |8|^3 = 1 + 216 + 512 = 729 = 9^3$, so $\|\mathbf{y}\|_3 = 729^{1/3} = 9$, using $(9^3)^{1/3} = 9$.

(b) Follow Theorem 18.11.6: $x_i = \sigma_i |y_i|^2 / 9^2$ with σ_i the sign of y_i, since $q - 1 = 2$. So

$$\mathbf{x} = \left(\tfrac{1}{81}, -\tfrac{36}{81}, \tfrac{64}{81}\right) = \left(\tfrac{1}{81}, -\tfrac{4}{9}, \tfrac{64}{81}\right).$$

Check: $|x_i|^{3/2} = (|y_i|^2/81)^{3/2} = |y_i|^3/729$, giving $\frac{1}{729} + \frac{216}{729} + \frac{512}{729} = 1$. So $\|\mathbf{x}\|_{3/2} = 1$. And $\langle \mathbf{x}, \mathbf{y} \rangle = \frac{1}{81} + \frac{216}{81} + \frac{512}{81} = \frac{729}{81} = 9$.

(c) By Theorem 18.11.6 the dual norm is $\|\mathbf{y}\|_3 = 9$, attained at the $\mathbf{x}$ of (b). The unit vector $\mathbf{e}_3$ has $\|\mathbf{e}_3\|_{3/2} = 1$ and gives $\langle \mathbf{e}_3, \mathbf{y} \rangle = 8 < 9$: even when one coordinate of $\mathbf{y}$ dominates, putting all the weight on it is not optimal.

Exercise. B2: Harmonic and arithmetic means

Let $x_1, \dots, x_n > 0$. Prove that

$$\frac{n}{\frac{1}{x_1} + \cdots + \frac{1}{x_n}} \leq \frac{x_1 + \cdots + x_n}{n},$$

with equality if and only if $x_1 = \cdots = x_n$.

Solution. By Exercise (B1: Which are convex?) (d), $g(t) = 1/t$ is strictly convex on $(0, \infty)$. Let $m = \frac{1}{n}(x_1 + \cdots + x_n) > 0$. Theorem 18.10.2 with the weights $w_i = 1/n > 0$ gives

$$\frac{1}{m} = g\Big(\sum_{i=1}^n \frac{1}{n} x_i\Big) \leq \sum_{i=1}^n \frac{1}{n}\, g(x_i) = \frac{1}{n}\Big(\frac{1}{x_1} + \cdots + \frac{1}{x_n}\Big),$$

with equality only when all x_i are equal. Both sides are positive, and taking reciprocals reverses the inequality: $m \geq n/(\sum_i 1/x_i)$, which is the claim. If all x_i equal c, both sides equal c. So equality holds exactly when the x_i are all equal.

18.11.5.3 C. Going deeper

Exercise. C1: The exponential above its tangent

(a) Deduce from Exercise (C2: Tangent lines) (a) that $e^y \geq 1 + y$ for every real y.

(b) Deduce that $\log x \leq x - 1$ for every real $x > 0$.

Solution.

(a) By Lemma 18.11.1 (a), the function exp is convex and differentiable on the open interval $\mathbb{R}$, with derivative exp. So Exercise (C2: Tangent lines) (a) applies, and at $x = 0$ it gives $e^y \geq e^0 + e^0(y - 0) = 1 + y$. For $y \geq 0$ this was visible from the series, as in the proof of Lemma 18.11.1 (b); for $y < 0$ it is new.

(b) Let $x > 0$ and put $y = \log x$, so that $e^y = x$ by Lemma 18.11.1 (b). Part (a) gives $x \geq 1 + \log x$, that is, $\log x \leq x - 1$.

Exercise. C2: When Hölder is an equality

Let $p, q \in (1, \infty)$ be conjugate exponents, and let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ be nonzero.

(a) Prove that $\sum_i |x_i y_i| = \|\mathbf{x}\|_p \|\mathbf{y}\|_q$ if and only if $|x_i|^p / \|\mathbf{x}\|_p^p = |y_i|^q / \|\mathbf{y}\|_q^q$ for every i.

(b) Deduce that the maximizer in Theorem 18.11.6 is unique: if $\mathbf{y} \neq \mathbf{0}$, $\|\mathbf{x}\|_p \leq 1$, and $\langle \mathbf{x}, \mathbf{y} \rangle = \|\mathbf{y}\|_q$, then $\mathbf{x}$ is the vector given there.

Hint: in (a), look at where the proof of Theorem 18.11.4 adds up the inequalities of Lemma 18.11.3.

Solution.

(a) Put $a_i = |x_i|/\|\mathbf{x}\|_p$ and $b_i = |y_i|/\|\mathbf{y}\|_q$, as in the proof of Theorem 18.11.4, so that $\sum_i a_i^p = \sum_i b_i^q = 1$. Dividing by $\|\mathbf{x}\|_p \|\mathbf{y}\|_q > 0$, the equality $\sum_i |x_i y_i| = \|\mathbf{x}\|_p \|\mathbf{y}\|_q$ says $\sum_i a_i b_i = 1 = \sum_i (a_i^p/p + b_i^q/q)$. By Lemma 18.11.3, $a_i b_i \leq a_i^p/p + b_i^q/q$ for each i. A sum of nonnegative differences $(a_i^p/p + b_i^q/q) - a_i b_i$ is 0 exactly when each difference is 0, so equality holds exactly when $a_i b_i = a_i^p/p + b_i^q/q$ for every i, that is, by the equality case of Lemma 18.11.3, when $a_i^p = b_i^q$ for every i. By the rule $(a/b)^p = a^p/b^p$ for real powers, $a_i^p = |x_i|^p/\|\mathbf{x}\|_p^p$ and $b_i^q = |y_i|^q/\|\mathbf{y}\|_q^q$.

(b) Let $\mathbf{y} \neq \mathbf{0}$, $\|\mathbf{x}\|_p \leq 1$, and $\langle \mathbf{x}, \mathbf{y} \rangle = \|\mathbf{y}\|_q$. By Theorem 18.11.4,

$$\|\mathbf{y}\|_q = \langle \mathbf{x}, \mathbf{y} \rangle \leq \sum_i |x_i y_i| \leq \|\mathbf{x}\|_p \|\mathbf{y}\|_q \leq \|\mathbf{y}\|_q,$$

so every inequality is an equality. First, $\|\mathbf{x}\|_p = 1$, since $\|\mathbf{y}\|_q > 0$; in particular $\mathbf{x} \neq \mathbf{0}$. Second, $\sum_i x_i y_i = \sum_i |x_i y_i|$, and since $x_i y_i \leq |x_i y_i|$ for each i, every $x_i y_i \geq 0$. Third, by (a), $|x_i|^p = |y_i|^q/\|\mathbf{y}\|_q^q$ for every i. Raising to the power $1/p$ and using $q/p = q - 1$, which follows from $\frac{1}{p} + \frac{1}{q} = 1$, gives $|x_i| = |y_i|^{q-1}/\|\mathbf{y}\|_q^{q-1}$. If $y_i = 0$, then $x_i = 0$, as in the formula. If $y_i \neq 0$, then $x_i \neq 0$ and $x_i y_i \geq 0$, so x_i has the sign σ_i of y_i. Hence $x_i = \sigma_i |y_i|^{q-1}/\|\mathbf{y}\|_q^{q-1}$ for every i, which is the vector of Theorem 18.11.6.

Chapter 19

Nonnegative Matrices

Two chapters left debts that this one pays. Chapter 9 §11 found the long-run behavior of a Markov chain by diagonalizing its transition matrix, proved that the chain settles down under three hypotheses, that the transition matrix is diagonalizable, that 1 is a simple eigenvalue, and that every other eigenvalue has modulus less than 1 (Theorem 9.11.4), and promised that "Chapter 19 proves the Perron–Frobenius theorem, which gives the conclusions of Theorem 9.11.4 for every stochastic matrix with some power having all entries positive, without assuming diagonalizability and without computing eigenvalues." Chapter 18 §08 stated Birkhoff's theorem, that the extreme points of the doubly stochastic matrices are exactly the permutation matrices, proved its two forms equivalent, and left the theorem itself to this chapter.

Both debts are about matrices whose entries are all ≥ 0, and the theory of such matrices is unlike anything earlier in the book. It depends on the **standard basis**: the hypothesis $\mathbf{A} \geq \mathbf{O}$ is destroyed by almost every change of basis, so the similarity invariants that organized Chapters 9 and 10 are joined by a new kind of information, the pattern of zero entries. What that pattern buys is striking. A matrix with every entry positive has a positive eigenvalue that is simple, strictly larger in modulus than every other eigenvalue, and equal to the spectral radius, with an eigenvector whose entries are all positive; this is Perron's theorem. When some entries are zero, most of this survives when the zeros do not split the indices into a part that the rest cannot reach, a condition read off a directed graph. From those two facts follow the convergence of Markov chains, the ranking of web pages, Birkhoff's theorem, and the theory of the matrices $s\mathbf{I} - \mathbf{B}$ with $\mathbf{B} \geq \mathbf{O}$ that arise in economics and in the discretization of differential equations.

The field is $\mathbb{R}$. Matrices and vectors have real entries, and a real matrix is read in $M_n(\mathbb{C})$ whenever its eigenvalues are wanted, as in Chapter 16 §04, so that its spectral radius $\rho(\mathbf{A})$ is defined (Definition 16.4.1). Two conventions hold throughout and are never switched.

- **Stochastic means column-stochastic.** As in Definition 9.11.1, a stochastic matrix has nonnegative entries and every **column** adding up to 1, and a Markov chain is $\mathbf{x}_{k+1} = \mathbf{A}\mathbf{x}_k$. Many books use rows and multiply on the left; their stochastic matrices are the transposes of ours.
- **The order is entrywise.** $\mathbf{A} \geq \mathbf{O}$ means every entry of $\mathbf{A}$ is ≥ 0, and $\mathbf{A} > \mathbf{O}$ means every entry is > 0, as for vectors in Chapter 18 §05; §01 defines the matrix version (Definition 19.1.1). This is **not** the Loewner order of Chapter 13, which the book always writes $\mathbf{A} \succeq \mathbf{O}$. A positive matrix need not be positive definite, and a positive definite matrix need not be positive: §01 gives a 2×2 example of each.
- **The absolute value is entrywise.** In this chapter $|\mathbf{M}|$ is the matrix of the moduli $|m_{ij}|$ of the entries of $\mathbf{M}$ (Definition 19.1.1 (c)), **not** the absolute value $(\mathbf{A}^*\mathbf{A})^{1/2}$ of Chapter 13.

19.0.1 The analysis this chapter imports

The list is Chapter 16's, facts (A1)–(A6) of that chapter's introduction, cited by the same name at the point of use. Three of them carry the weight, together with the algebra of limits from the same introduction. In §01, **(A3)**, compactness of closed bounded sets, and **(A4)**, the extreme value theorem, produce the eigenvalue and positive eigenvector of Perron's theorem as a maximum that is attained rather than merely approached. The passage from positive to general nonnegative matrices in §03 is a limit, and it uses **(A2)**, monotone convergence, together with (A3) again. Nothing else analytic is used, and a section that needs anything more says so where it uses it.

19.0.2 What you need

- **Chapter 0 §06** — the modulus of a complex number and the triangle inequality: Theorem 0.6.2, Theorem 0.6.4.
- **Chapter 1 §§06 and 07** — two counting facts, both in §07's proof of Birkhoff's theorem: the size bound on an independent list, and the dimension formula for a sum of subspaces: Theorem 1.6.3, Theorem 1.7.3.
- **Chapter 3 §04** — permutation matrices, which relabel coordinates: Definition 3.4.2, Lemma 3.4.7.
- **Chapter 2 §03** — Rank–Nullity, used once, in §06, to give the kernel of $\mathbf{x} \mapsto \mathbf{1}^\top \mathbf{x}$ its dimension: Theorem 2.3.1.
- **Chapter 6 §05** — that a polynomial over $\mathbb{C}$ splits, which is how §§01 and 03 count the eigenvalues of a nonnegative matrix: Corollary 6.5.5.
- **Chapters 7 and 8** — the characteristic polynomial and block triangular matrices: Theorem 7.10.3, Theorem 7.8.4, Theorem 8.1.1.
- **Chapter 9** — eigenvalues and their multiplicities, invariant subspaces, left eigenvectors, the spectral mapping theorem, and the Markov chains this chapter returns to: Definition 9.3.1, Theorem 9.1.3, Definition 9.10.1, Proposition 9.10.5, Theorem 9.10.6, Theorem 9.10.1, Definition 9.11.1, Definition 9.11.2, Theorem 9.11.4.
- **Chapter 10** — generalized eigenspaces, and the convergence of matrix powers and of the Neumann series: Theorem 10.1.3, Theorem 10.10.4, Corollary 10.10.5, Theorem 10.10.6.
- **Chapter 13** — positive semidefinite and positive definite matrices, which the entrywise order must not be confused with: Definition 13.1.1.
- **Chapter 16** — entrywise convergence as the convergence of matrices (§02), the operator norms and their formulas (§03), and the spectral radius, Gelfand's formula, and the Neumann series: Corollary 16.2.4, Theorem 16.3.3, Theorem 16.3.2, Definition 16.4.1, Theorem 16.4.4, Corollary 16.4.2, Theorem 16.5.1.
- **Chapter 17 §08** — majorization, Schur's theorem, which Birkhoff's theorem reproves, and the Schur–Horn theorem: Definition 17.8.1, Theorem 17.8.2, Theorem 17.8.5.
- **Chapter 18** — the simplex, closed sets, Carathéodory's theorem, the doubly stochastic matrices, their permutation-matrix extreme points, and the vertices of polyhedra: Definition 18.1.2, Definition 18.1.6, Theorem 18.2.2, Definition 18.8.4, Proposition 18.8.9, Theorem 18.8.5, Theorem 18.9.2.

19.0.3 Roadmap

- **Positive matrices and Perron's theorem.** The entrywise order, the monotonicity of the spectral radius, and Perron's theorem for matrices with every entry positive, proved by maximizing over the simplex.
- **Irreducibility and graphs.** The directed graph of a matrix, strong connectivity, irre-

ducible matrices and the test $(\mathbf{I}+\mathbf{A})^{n-1} > \mathbf{O}$, and the block triangular normal form of a reducible matrix.

- **The Perron–Frobenius theorem.** Every nonnegative matrix has its spectral radius as an eigenvalue with a nonnegative eigenvector, and for an irreducible matrix that eigenvalue is simple with a positive eigenvector. The examples that show irreducibility cannot be dropped.
- **The Collatz–Wielandt formula.** The spectral radius of an irreducible matrix as a max–min and a min–max over test vectors, and the row-sum bounds that follow.
- **Primitive matrices.** Irreducible matrices with a positive power, the eigenvalues on the circle of radius ρ, and the convergence of $(\mathbf{A}/\rho)^k$ without any diagonalizability.
- **Markov chains.** Chapter 9's promise, paid in its own words; chains that are irreducible but not primitive, reducible chains, and absorbing chains.
- **Birkhoff's theorem.** Every extreme point of the doubly stochastic matrices is a permutation matrix, and a second proof of Schur's theorem.
- **PageRank.** The Google matrix is positive, so Perron's theorem ranks the pages, and the damping factor bounds the rate of convergence.
- **M-matrices.** The matrices $s\mathbf{I}-\mathbf{B}$ with $\mathbf{B} \geq \mathbf{O}$, their many equivalent characterizations, and two applications. The chapter's summary closes this section.

19.0.4 Named moves

- **Push up a slack.** If $\mathbf{Au} \geq r\mathbf{u}$ with r maximal, the slack $\mathbf{Au}-r\mathbf{u}$ must vanish: one more application of a positive matrix would make it strictly positive and leave room to increase r. Perron's proof makes this move once, in its push-up claim, and then uses the claim three times.
- **Pair a slack with a positive left eigenvector.** A nonnegative vector whose product with a positive vector is zero is itself zero. This is how a positive left eigenvector rules out a second Jordan block, and a nonnegative eigenvector for a smaller eigenvalue.
- **Read the powers as walks.** The nonzero entries of $\mathbf{A}^k$ record the walks of length k in a graph, so questions about positivity of powers become questions about reachability, answered by following arrows.
- **Relabel and rescale, nothing more.** The only changes of basis that respect the entrywise order are the relabelings of the coordinates, possibly combined with positive rescalings, and every normal form in the chapter uses only these.
- **Perturb to positive, then take the limit.** To prove something for a nonnegative matrix, prove it for $\mathbf{A}+\varepsilon\mathbf{J}$, which is positive, and let $\varepsilon \to 0$; what survives the limit is exactly the nonstrict part of the conclusion.

19.1 Positive Matrices and Perron's Theorem

Chapter 9 §11 found the long-run behavior of a Markov chain by diagonalizing its transition matrix, and warned that nothing guaranteed the method would work. Chapter 16 §04 gave every square matrix a spectral radius, the largest modulus of an eigenvalue, but said nothing about whether that modulus is itself an eigenvalue. For a general matrix it need not be: the swap $\left(\begin{smallmatrix}0&1\\1&0\end{smallmatrix}\right)$ has spectral radius 1 and the eigenvalue -1 sitting on the circle beside it, and the rotation by a right angle has no real eigenvalue at all. This section shows that a single hypothesis on the **entries** removes every one of these difficulties: if every entry of $\mathbf{A}$ is positive, then $\rho(\mathbf{A})$ is an eigenvalue, it is simple, it strictly dominates every other eigenvalue in modulus, and it has an

eigenvector with positive entries. That is Perron's theorem, and the rest of the chapter extends it.

Throughout, matrices and vectors have real entries, except where complex eigenvectors appear. The eigenvalues of a real matrix are those of the same matrix read in $M_n(\mathbb{C})$, exactly as in Chapter 16 §04, so that $\rho(\mathbf{A})$ is defined (Definition 16.4.1) and eigenvectors may have complex entries.

19.1.1 The entrywise order

Chapter 18 §05 wrote $\mathbf{x} \geq \mathbf{0}$ for a real vector with every entry ≥ 0. The same comparison, made one entry at a time, makes sense for matrices of any shape, and it is the only order this chapter uses. Its hook is a recurring expression: every probability argument of Chapter 9 §11 began "the entries are nonnegative, so each entry of the product is a sum of nonnegative terms". We name the hypothesis once and prove its consequences once.

The entrywise order compares two matrices of the same shape position by position.

Definition 19.1.1. Entrywise Order; Nonnegative and Positive Matrices

Let $\mathbf{A} = (a_{ij})$ and $\mathbf{B} = (b_{ij})$ be real $m \times n$ matrices.

(a) $\mathbf{A} \geq \mathbf{B}$ means $a_{ij} \geq b_{ij}$ for **every** i, j, and $\mathbf{A} > \mathbf{B}$ means $a_{ij} > b_{ij}$ for **every** i, j. We write $\mathbf{B} \leq \mathbf{A}$ and $\mathbf{B} < \mathbf{A}$ for the same statements.

(b) $\mathbf{A}$ is **nonnegative** if $\mathbf{A} \geq \mathbf{O}$, and **positive** if $\mathbf{A} > \mathbf{O}$, where $\mathbf{O}$ is the zero matrix of the same shape.

(c) For a complex $m \times n$ matrix $\mathbf{M} = (m_{ij})$, the **entrywise absolute value** is the real matrix $|\mathbf{M}| := (|m_{ij}|)$.

Vectors are the case $n = 1$, so $\mathbf{x} \geq \mathbf{0}$, $\mathbf{x} > \mathbf{0}$ and $|\mathbf{x}| = (|x_1|, \dots, |x_n|)$ have the same meaning as in Chapter 18 §05.

In words: (a) compares every pair of corresponding entries, and asks for a strict inequality in **every** position in the strict version. So $\mathbf{A} > \mathbf{B}$ is much stronger than "$\mathbf{A} \geq \mathbf{B}$ and $\mathbf{A} \neq \mathbf{B}$", and the two must not be confused. Clause (b) is the special case $\mathbf{B} = \mathbf{O}$. Clause (c) takes the modulus of each entry separately; it has nothing to do with the absolute value $(\mathbf{A}^*\mathbf{A})^{1/2}$ of Chapter 13, and in this chapter $|\mathbf{M}|$ always means the entrywise one.

Some examples, simplest first.

- $\begin{pmatrix} 1 & 2 \\ 3 & 2 \end{pmatrix} > \mathbf{O}$: every entry is positive. This is the matrix of Example (Eigenvalues and Eigenspaces of a 2×2 Matrix).
- $\mathbf{I}_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \geq \mathbf{O}$, but $\mathbf{I}_2$ is **not** positive, since its off-diagonal entries are 0. Every stochastic matrix (Definition 9.11.1) is nonnegative, and so is every permutation matrix.
- **The degenerate cases.** The zero matrix is nonnegative and not positive. A 1×1 matrix (a) is positive exactly when $a > 0$, and then everything this section proves reduces to the statement that a is an eigenvalue of (a) with eigenvector (1). It is worth keeping in mind that every theorem below must survive $n = 1$.
- For $\mathbf{M} = \begin{pmatrix} 3 & -4i \\ 0 & -1 \end{pmatrix}$, $|\mathbf{M}| = \begin{pmatrix} 3 & 4 \\ 0 & 1 \end{pmatrix}$.

Nonexample by minimal change. Take the positive matrix $\begin{pmatrix} 1 & 2 \\ 3 & 2 \end{pmatrix}$ and change its $(1,2)$ entry to 0. The result $\begin{pmatrix} 1 & 0 \\ 3 & 2 \end{pmatrix}$ is still nonnegative and still $\geq \mathbf{I}_2$, but it is not positive: the clause "**every** entry > 0" fails in one position. That one zero is enough to lose a conclusion of this section. The new matrix is triangular, with eigenvalues 1 and 2. Its eigenvectors for 2 are the multiples of $\mathbf{e}_2$, and those for 1 the multiples of $(1, -3)$. So the dominant eigenvalue 2 has a nonnegative eigenvector, but **no** positive one. Perron's theorem below says this cannot happen for a positive matrix.

Warning. **Positive is not positive definite, and positive definite is not positive.** The two words sound alike and measure different things. The entrywise order looks at entries; positive definiteness (Definition 13.1.1) looks at the quadratic form. The matrix $\begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}$ is positive, and symmetric, but its eigenvalues are 3 and -1, so it is not positive definite: $\mathbf{x} = (1, -1)$ gives $\mathbf{x}^\top \mathbf{A}\mathbf{x} = -2$. The matrix $\begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}$ is positive definite, with eigenvalues 1 and 3, but it has negative entries, so it is not even nonnegative. In this book $\mathbf{A} \geq \mathbf{O}$ and $\mathbf{A} > \mathbf{O}$ always mean the entrywise order, and the Loewner order is always written $\mathbf{A} \succeq \mathbf{O}$, $\mathbf{A} \succ \mathbf{O}$.

Why this definition. The order is entrywise because the only operations the theory needs are sums and products of entries, and a sum or product of nonnegative numbers is nonnegative. The hypothesis is not coordinate-free: a change of basis destroys it at once. That is also why this chapter, unlike most of the book, cares about the **standard** basis: the hypothesis $\mathbf{A} \geq \mathbf{O}$ is a statement about the matrix, not about the operator $T_{\mathbf{A}}$. The only changes of basis that preserve it are the ones that rearrange coordinates and rescale them by positive numbers, and §02 uses the first kind throughout.

The first facts are the ones used in every argument below. They are the three rules of arithmetic with nonnegative numbers, carried to matrices.

Lemma 19.1.1. Rules of the Entrywise Order

Let $\mathbf{A}, \mathbf{B} \in M_{m\times n}(\mathbb{R})$, $\mathbf{C}, \mathbf{D} \in M_{n\times p}(\mathbb{R})$ and $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$.

(a) If $\mathbf{O} \leq \mathbf{A} \leq \mathbf{B}$ and $\mathbf{0} \leq \mathbf{x} \leq \mathbf{y}$, then $\mathbf{0} \leq \mathbf{A}\mathbf{x} \leq \mathbf{B}\mathbf{y}$.

(b) If $\mathbf{A} > \mathbf{O}$, $\mathbf{x} \geq \mathbf{0}$ and $\mathbf{x} \neq \mathbf{0}$, then $\mathbf{A}\mathbf{x} > \mathbf{0}$.

(c) If $\mathbf{O} \leq \mathbf{A} \leq \mathbf{B}$ and $\mathbf{O} \leq \mathbf{C} \leq \mathbf{D}$, then $\mathbf{O} \leq \mathbf{A}\mathbf{C} \leq \mathbf{B}\mathbf{D}$. In particular, for square $\mathbf{O} \leq \mathbf{A} \leq \mathbf{B}$, $\mathbf{O} \leq \mathbf{A}^k \leq \mathbf{B}^k$ for every $k \geq 0$.

Idea. Each entry of a product is a sum of products of entries, so every rule comes down to the arithmetic of nonnegative numbers. For the comparison in (a), write $\mathbf{B}\mathbf{y} - \mathbf{A}\mathbf{x}$ as a sum of two products of nonnegative factors; for (b), one positive term in a sum of nonnegative terms is enough; and (c) is (a) applied one column at a time.

Proof.

(a) Each entry $(\mathbf{Ax})_i = \sum_j a_{ij}x_j$ is a sum of products of nonnegative numbers, so $\mathbf{Ax} \geq \mathbf{0}$. Next, $\mathbf{By} - \mathbf{Ax} = \mathbf{B}(\mathbf{y}-\mathbf{x}) + (\mathbf{B}-\mathbf{A})\mathbf{x}$, and both terms are nonnegative by what was just shown, because $\mathbf{B} \geq \mathbf{O}$, $\mathbf{y}-\mathbf{x} \geq \mathbf{0}$, $\mathbf{B}-\mathbf{A} \geq \mathbf{O}$ and $\mathbf{x} \geq \mathbf{0}$.

(b) Since $\mathbf{x} \neq \mathbf{0}$ and $\mathbf{x} \geq \mathbf{0}$, some entry x_k is positive. For every i, $(\mathbf{Ax})_i = \sum_j a_{ij}x_j \geq a_{ik}x_k > 0$, because the other terms are ≥ 0 and $a_{ik} > 0$.

(c) By Theorem 0.9.2, column l of $\mathbf{AC}$ is $\mathbf{Ac}_l$, where $\mathbf{c}_l$ is column l of $\mathbf{C}$, and likewise for $\mathbf{BD}$. Since $\mathbf{0} \leq \mathbf{c}_l \leq \mathbf{d}_l$, part (a) gives $\mathbf{0} \leq \mathbf{Ac}_l \leq \mathbf{Bd}_l$ for every l. The statement about powers follows by induction on k: $\mathbf{A}^0 = \mathbf{B}^0 = \mathbf{I}$, and $\mathbf{A}^{k+1} = \mathbf{A}^k\mathbf{A} \leq \mathbf{B}^k\mathbf{B} = \mathbf{B}^{k+1}$ by the case just proved.

■

The absolute value interacts with products through the triangle inequality, which is the second tool.

Lemma 19.1.2. Absolute Value of a Product

Let $\mathbf{A} \in M_{m\times n}(\mathbb{C})$, $\mathbf{B} \in M_{n\times p}(\mathbb{C})$ and $\mathbf{x} \in \mathbb{C}^n$. Then

(a) $|\mathbf{Ax}| \leq |\mathbf{A}|\,|\mathbf{x}|$;

(b) $|\mathbf{AB}| \leq |\mathbf{A}|\,|\mathbf{B}|$;

(c) if $m = n$, then $|\mathbf{A}^k| \leq |\mathbf{A}|^k$ for every $k \geq 0$.

Idea. Each entry of $\mathbf{Ax}$ is a sum of products, and the triangle inequality bounds its modulus by the sum of the moduli of the products: that is (a). Then (b) is (a) one column at a time, and (c) follows from (b) by induction on k, with the order rules to replace $|\mathbf{A}^k|$ by the larger $|\mathbf{A}|^k$.

Proof.

(a) For each i, the triangle inequality (Theorem 0.6.4, extended to finitely many terms by induction) and the multiplicativity of the modulus (Theorem 0.6.2 (d)) give

$$|(\mathbf{Ax})_i| = \left|\sum_j a_{ij}x_j\right| \leq \sum_j |a_{ij}|\,|x_j| = \big(|\mathbf{A}|\,|\mathbf{x}|\big)_i.$$

(b) Apply (a) to each column of $\mathbf{B}$, using Theorem 0.9.2 as in the previous proof.

(c) Induction on k. For $k = 0$, $|\mathbf{I}| = \mathbf{I}$. If $|\mathbf{A}^k| \leq |\mathbf{A}|^k$, then by (b) and Lemma 19.1.1 (c),

$$|\mathbf{A}^{k+1}| = |\mathbf{A}^k\mathbf{A}| \leq |\mathbf{A}^k|\,|\mathbf{A}| \leq |\mathbf{A}|^k|\mathbf{A}| = |\mathbf{A}|^{k+1}.$$

This proves the lemma.

■

For a nonnegative $\mathbf{A}$ we have $|\mathbf{A}| = \mathbf{A}$, and (a) reads $|\mathbf{Ax}| \leq \mathbf{A}|\mathbf{x}|$. That is the form used below:

whatever cancellation happens inside $\mathbf{Ax}$ can only make it smaller.

The proof of Perron's theorem also needs to know when the triangle inequality is an equality, a case Chapter 0 did not record. The answer is the expected one: when all the terms point in the same direction.

Lemma 19.1.3. Equality in the Triangle Inequality

Let $z_1, \dots, z_m \in \mathbb{C}$, put $s = z_1 + \cdots + z_m$, and suppose $|s| = |z_1| + \cdots + |z_m|$ and $s \neq 0$. Then $z_j = |z_j|\, c$ for every j, where $c = s/|s|$ has $|c| = 1$.

Idea. Rotate the sum onto the positive real axis: multiplying by $d = \overline{s}/|s|$ turns s into the real number $|s|$. Then $|s|$ is the real part of $\sum_j dz_j$, which is at most $\sum_j |z_j|$ term by term. Equality of the totals forces equality in every term, and a complex number whose real part equals its modulus is a real number ≥ 0.

Proof. Recall from the proof of Theorem 0.6.4 that $\operatorname{Re} u = x \leq |x| \leq \sqrt{x^2 + y^2} = |u|$ for $u = x + yi$. Equality $\operatorname{Re} u = |u|$ forces equality in both steps, that is $y^2 = 0$ and $x = |x|$, so u is a real number ≥ 0 and $u = |u|$.

Put $d = \overline{s}/|s|$. Then $|d| = 1$ and $ds = |s|^2/|s| = |s|$ by Theorem 0.6.2 (c), (d). Hence

$$\sum_j |z_j| = |s| = ds = \operatorname{Re} \sum_j dz_j = \sum_j \operatorname{Re}(dz_j) \leq \sum_j |dz_j| = \sum_j |z_j|.$$

So the inequality in the middle is an equality. It is a sum of the inequalities $\operatorname{Re}(dz_j) \leq |dz_j|$, so each of them is an equality, and by the first paragraph $dz_j = |dz_j| = |z_j|$. Multiplying by $c = s/|s|$, which satisfies $cd = s\overline{s}/|s|^2 = 1$, gives $z_j = |z_j|\, c$.

■

19.1.2 The spectral radius is monotone

A first payoff joins the entrywise order to Chapter 16. Enlarging the entries of a nonnegative matrix should not shrink its eigenvalues, and for the spectral radius this is true.

Proposition 19.1.4. Monotonicity of the Spectral Radius

(a) If $\mathbf{A}, \mathbf{B} \in M_n(\mathbb{R})$ and $\mathbf{O} \leq \mathbf{A} \leq \mathbf{B}$, then $\rho(\mathbf{A}) \leq \rho(\mathbf{B})$.

(b) More generally, if $\mathbf{A} \in M_n(\mathbb{C})$, $\mathbf{B} \in M_n(\mathbb{R})$ and $|\mathbf{A}| \leq \mathbf{B}$, then $\rho(\mathbf{A}) \leq \rho(|\mathbf{A}|) \leq \rho(\mathbf{B})$.

Idea. Eigenvalues do not see the entries directly, but Gelfand's formula (Theorem 16.4.4) expresses ρ through the norms of the powers, and the largest absolute row sum $\|\cdot\|_\infty$ is a norm that does see the entries, monotonically. So compare the powers entrywise, take norms, take k-th roots, and pass to the limit.

Proof.

(b) First, $|\mathbf{A}^k| \le |\mathbf{A}|^k \le \mathbf{B}^k$ for every $k \ge 1$: the first inequality is Lemma 19.1.2 (c), and the second is Lemma 19.1.1 (c) applied to $\mathbf{O} \le |\mathbf{A}| \le \mathbf{B}$. By Theorem 16.3.3 (b), $\|\mathbf{M}\|_\infty$ is the largest absolute row sum of $\mathbf{M}$, so it depends only on $|\mathbf{M}|$ and does not decrease when the entries of $|\mathbf{M}|$ increase. Hence

$$\|\mathbf{A}^k\|_\infty \le \||\mathbf{A}|^k\|_\infty \le \|\mathbf{B}^k\|_\infty \qquad (k \ge 1).$$

The k-th root is increasing on the nonnegative reals, so the same inequalities hold between the k-th roots. The operator norm $\|\cdot\|_\infty$ is a matrix norm (Theorem 16.3.2 (d)), so by Theorem 16.4.4 the three sequences of k-th roots converge to $\rho(\mathbf{A})$, $\rho(|\mathbf{A}|)$ and $\rho(\mathbf{B})$. A nonstrict inequality survives a limit, which gives $\rho(\mathbf{A}) \le \rho(|\mathbf{A}|) \le \rho(\mathbf{B})$.

(c) If $\mathbf{O} \le \mathbf{A} \le \mathbf{B}$, then $|\mathbf{A}| = \mathbf{A}$, and (b) gives $\rho(\mathbf{A}) \le \rho(\mathbf{B})$.

■

The inequality need not be strict, even when $\mathbf{A} \ne \mathbf{B}$: $\operatorname{diag}(1,0) \le \mathbf{I}_2$ and both have spectral radius 1. For positive matrices it is strict, which is Exercise C3. The hypothesis $\mathbf{O} \le \mathbf{A}$ cannot be dropped from (a): $\operatorname{diag}(-3,0) \le \mathbf{O}$ entrywise, yet $\rho(\operatorname{diag}(-3,0)) = 3 > 0 = \rho(\mathbf{O})$.

Example. Enlarging one entry

Let $\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 1 & 0 \end{pmatrix}$ and $\mathbf{B} = \begin{pmatrix} 1 & 2 \\ 1 & 1 \end{pmatrix}$. Compute both spectral radii and compare with Proposition 19.1.4.

Solution. $\mathbf{O} \le \mathbf{A} \le \mathbf{B}$, with a difference only in the $(2,2)$ entry. By the 2×2 formula $p(x) = x^2 - (\operatorname{tr})x + \det$, $p_{\mathbf{A}}(x) = x^2 - x - 2 = (x-2)(x+1)$, so $\rho(\mathbf{A}) = 2$, and $p_{\mathbf{B}}(x) = x^2 - 2x - 1$, with roots $1 \pm \sqrt{2}$, so $\rho(\mathbf{B}) = 1 + \sqrt{2} \approx 2.414$. Indeed $\rho(\mathbf{A}) \le \rho(\mathbf{B})$. In both cases the eigenvalue of largest modulus is positive, which Perron's theorem below (Theorem 19.1.5) guarantees for the positive $\mathbf{B}$.

19.1.3 Perron's theorem

Here is the theorem this section is built around.

Theorem 19.1.5. Perron's Theorem

Let $n \ge 1$ and let $\mathbf{A} \in M_n(\mathbb{R})$ be **positive**, $\mathbf{A} > \mathbf{O}$. Write $r = \rho(\mathbf{A})$. Then:

(a) $r > 0$, and r is an eigenvalue of $\mathbf{A}$ with an eigenvector $\mathbf{v} > \mathbf{0}$;

(b) r is **algebraically simple**: $a_{\mathbf{A}}(r) = 1$. In particular $E_r(\mathbf{A}) = \operatorname{span}(\mathbf{v})$;

(c) every eigenvalue $\lambda \in \mathbb{C}$ of $\mathbf{A}$ with $\lambda \ne r$ satisfies $|\lambda| < r$;

(d) every nonnegative eigenvector of $\mathbf{A}$ is a positive multiple of $\mathbf{v}$. In particular, no eigenvector of $\mathbf{A}$ for an eigenvalue other than r is nonnegative;

(e) $\mathbf{A}^\top$ is positive with $\rho(\mathbf{A}^\top) = r$, so (a)–(d) hold for $\mathbf{A}^\top$: there is $\mathbf{w} > \mathbf{0}$ with $\mathbf{w}^\top \mathbf{A} = r\mathbf{w}^\top$.

The route. There are several proofs of Perron's theorem. We take the one that finds r as the answer to an optimization problem: **how large can t be if $\mathbf{Ax} \geq t\mathbf{x}$ for some probability vector $\mathbf{x}$?** Its advantage is that it produces the positive eigenvector and the eigenvalue together, and that compactness is used exactly once. The quotient form of this question, maximizing $\min_i (\mathbf{Ax})_i / x_i$ over $\mathbf{x} \in \Delta_n$, is the Collatz–Wielandt formula, which §04 develops for all irreducible matrices.

The quotient itself is awkward on the boundary of the simplex. Where some $x_i = 0$, the i-th ratio is undefined, and the natural repair, taking the minimum over the indices with $x_i > 0$ only, gives a function that need not be continuous. For $\mathbf{D} = \operatorname{diag}(2, 1)$ that function takes the value 2 at the corner $\mathbf{e}_1$, where only the first ratio counts, but the value $\min(2, 1) = 1$ at every point $(1 - t, t)$ with $0 < t < 1$. So the extreme value theorem cannot be applied to it as it stands. (For positive $\mathbf{A}$ the repaired function is in fact continuous, but proving that costs more than the route below; for the nonnegative matrices of §04 it can genuinely fail.) We avoid the quotient: instead of maximizing a function on Δ_n, we maximize the coordinate t over the set of pairs $(\mathbf{x}, t)$ with $\mathbf{Ax} \geq t\mathbf{x}$, which is closed, and on which t is continuous.

Idea. **Step roadmap.** ① The set $S = \{(\mathbf{x}, t) : \mathbf{x} \in \Delta_n,\ t \geq 0,\ \mathbf{Ax} \geq t\mathbf{x}\}$ is compact, so t attains a maximum r on it: facts (A3) and (A4) of Chapter 16's introduction. ② **Push-up claim:** if $\mathbf{u} \geq \mathbf{0}$, $\mathbf{u} \neq \mathbf{0}$, and $\mathbf{Au} \geq r\mathbf{u}$, then $\mathbf{Au} = r\mathbf{u}$ and $\mathbf{u} > \mathbf{0}$. Otherwise one more application of the positive matrix $\mathbf{A}$ turns the slack $\mathbf{Au} - r\mathbf{u}$ into a **strict** inequality, and a strict inequality leaves room to increase r, contradicting maximality. The maximizer therefore gives the positive eigenvector. ③ For any eigenvalue λ with eigenvector $\mathbf{x}$, $|\lambda||\mathbf{x}| \leq \mathbf{A}|\mathbf{x}|$, so $|\lambda| \leq r$ by maximality; hence $r = \rho(\mathbf{A})$. ④ If $|\lambda| = r$, the push-up claim forces equality in the triangle inequality, which aligns all entries of $\mathbf{x}$ and gives $\lambda = r$. ⑤ Subtract a multiple of $\mathbf{v}$ until an entry vanishes: the push-up claim leaves nothing, so $E_r(\mathbf{A})$ is a line. ⑥ Steps ①–⑤ for $\mathbf{A}^\top$ give a positive $\mathbf{w}$ with $\mathbf{A}^\top \mathbf{w} = \rho(\mathbf{A}^\top)\mathbf{w}$, and pairing with $\mathbf{v}$ shows $\rho(\mathbf{A}^\top) = r$. A second vector in the generalized eigenspace would give $(\mathbf{A} - r\mathbf{I})\mathbf{y} = \mathbf{v}$, and pairing with $\mathbf{w}$ yields $0 = \mathbf{w}^\top \mathbf{v} > 0$. ⑦ A nonnegative eigenvector for $\lambda \neq r$ would be orthogonal to $\mathbf{w} > \mathbf{0}$, which is impossible.

Here is the proof, following the roadmap.

Proof. Let $\mathbf{A} > \mathbf{O}$ be $n \times n$. By Definition 18.1.2, Δ_n is the set of probability vectors in $\mathbb{R}^n$. Define

$$S = \{(\mathbf{x}, t) \in \mathbb{R}^n \times \mathbb{R} : \mathbf{x} \in \Delta_n,\ t \geq 0,\ \mathbf{Ax} \geq t\mathbf{x}\}.$$

Step 1: a maximal t. S is nonempty: it contains $(\mathbf{x}, 0)$ for every $\mathbf{x} \in \Delta_n$, by Lemma 19.1.1 (a).

S is bounded for the Euclidean norm on $\mathbb{R}^{n+1}$, the norm with which (A3) is stated. Let M be the largest entry of $\mathbf{A}$. If $(\mathbf{x}, t) \in S$, then $0 \leq x_i \leq 1$ for every i, and since the n entries of $\mathbf{x}$ add up to 1, some $x_k \geq 1/n$. Then

$$\frac{t}{n} \leq t\, x_k \leq (\mathbf{Ax})_k = \sum_j a_{kj} x_j \leq M \sum_j x_j = M,$$

so $0 \le t \le nM$. Hence $\|(\mathbf{x},t)\|_2^2 = \sum_i x_i^2 + t^2 \le n + (nM)^2$, and $\|(\mathbf{x},t)\|_2 \le \sqrt{n} + nM$.

S is closed in $\mathbb{R}^{n+1}$ (Definition 18.1.6). Let $(\mathbf{x}_k, t_k) \in S$ converge to $(\mathbf{x},t)$ in the Euclidean norm. Then each coordinate converges, since $|u_i| \le \|\mathbf{u}\|_2$ for every $\mathbf{u} \in \mathbb{R}^{n+1}$. Every condition defining S is a finite list of nonstrict inequalities and one equation between expressions built from the entries by sums and products: $x_i \ge 0$, $\sum_i x_i = 1$, $t \ge 0$, and $(\mathbf{Ax})_i - t\,x_i \ge 0$. Each holds for every $(\mathbf{x}_k, t_k)$, and by the algebra of limits of Chapter 16's introduction each survives the limit. So $(\mathbf{x},t) \in S$.

By fact (A3) of Chapter 16's introduction, the closed bounded set $S \subseteq \mathbb{R}^{n+1}$ is compact, and the function $(\mathbf{x},t) \mapsto t$ is continuous on it. By fact (A4), the extreme value theorem, it attains a maximum: there are $\mathbf{z} \in \Delta_n$ and a largest number r such that

$$\mathbf{Az} \ge r\mathbf{z}, \qquad \text{and } t \le r \text{ for every } (\mathbf{x},t) \in S.$$

Moreover $r > 0$. Indeed, let s be the smallest row sum of $\mathbf{A}$, which is positive as $\mathbf{A} > \mathbf{O}$. The vector $\mathbf{x} = \frac{1}{n}\mathbf{1} \in \Delta_n$ satisfies $(\mathbf{Ax})_i = \frac{1}{n}\sum_j a_{ij} \ge \frac{s}{n} = s\,x_i$, so $(\frac{1}{n}\mathbf{1}, s) \in S$ and $r \ge s > 0$.

Step 2: the push-up claim.

Claim. If $\mathbf{u} \in \mathbb{R}^n$, $\mathbf{u} \ge \mathbf{0}$, $\mathbf{u} \ne \mathbf{0}$, and $\mathbf{Au} \ge r\mathbf{u}$, then $\mathbf{Au} = r\mathbf{u}$ and $\mathbf{u} > \mathbf{0}$.

Proof. Suppose $\mathbf{Au} \ne r\mathbf{u}$. Then $\mathbf{y} = \mathbf{Au} - r\mathbf{u}$ satisfies $\mathbf{y} \ge \mathbf{0}$ and $\mathbf{y} \ne \mathbf{0}$, so $\mathbf{Ay} > \mathbf{0}$ by Lemma 19.1.1 (b). Put $\mathbf{q} = \mathbf{Au}$; by the same rule $\mathbf{q} > \mathbf{0}$, and

$$\mathbf{Aq} - r\mathbf{q} = \mathbf{A}(\mathbf{Au} - r\mathbf{u}) = \mathbf{Ay} > \mathbf{0}.$$

Every $q_i > 0$, so $\varepsilon = \min_i (\mathbf{Ay})_i / q_i$ is defined and positive, and $(\mathbf{Aq})_i - r\,q_i = (\mathbf{Ay})_i \ge \varepsilon q_i$ for every i, that is, $\mathbf{Aq} \ge (r+\varepsilon)\mathbf{q}$. Let $\sigma = q_1 + \cdots + q_n > 0$ and $\mathbf{x} = \mathbf{q}/\sigma$. Then $\mathbf{x} \in \Delta_n$, and dividing by $\sigma > 0$ preserves the inequality, so $\mathbf{Ax} \ge (r+\varepsilon)\mathbf{x}$. Hence $(\mathbf{x}, r+\varepsilon) \in S$, which contradicts the maximality of r. Therefore $\mathbf{Au} = r\mathbf{u}$. Finally $\mathbf{Au} > \mathbf{0}$ by Lemma 19.1.1 (b), and $r > 0$, so $\mathbf{u} = r^{-1}\mathbf{Au} > \mathbf{0}$.

■

Apply the claim to $\mathbf{u} = \mathbf{z}$, which is nonnegative and nonzero because its entries add up to 1. It gives $\mathbf{Az} = r\mathbf{z}$ with $\mathbf{z} > \mathbf{0}$. Put $\mathbf{v} = \mathbf{z}$. So r is an eigenvalue of $\mathbf{A}$ with a positive eigenvector.

Step 3: $r = \rho(\mathbf{A})$. Let $\lambda \in \mathbb{C}$ be an eigenvalue of $\mathbf{A}$ and $\mathbf{x} \in \mathbb{C}^n$ an eigenvector for it. By Theorem 0.6.2 (d) and Lemma 19.1.2 (a), with $|\mathbf{A}| = \mathbf{A}$,

$$|\lambda|\,|\mathbf{x}| = |\lambda\mathbf{x}| = |\mathbf{Ax}| \le \mathbf{A}|\mathbf{x}|.$$

The vector $|\mathbf{x}|$ is nonnegative and nonzero, since $\mathbf{x} \ne \mathbf{0}$. Dividing by the sum $\sigma > 0$ of its entries, $\mathbf{x}' = |\mathbf{x}|/\sigma \in \Delta_n$ satisfies $\mathbf{Ax}' \ge |\lambda|\mathbf{x}'$, so $(\mathbf{x}', |\lambda|) \in S$ and $|\lambda| \le r$. Since r is itself an eigenvalue, $r = \max\{|\lambda| : \lambda \in \operatorname{spec}(\mathbf{A})\} = \rho(\mathbf{A})$ (Definition 16.4.1). This proves (a).

Step 4: eigenvalues of modulus r. Let λ be an eigenvalue with $|\lambda| = r$, and $\mathbf{x}$ an eigenvector. The display of Step 3 reads $\mathbf{A}|\mathbf{x}| \ge r|\mathbf{x}|$, so the push-up claim applies to $|\mathbf{x}|$ and gives

$$\mathbf{A}|\mathbf{x}| = r|\mathbf{x}| \quad \text{and} \quad |\mathbf{x}| > \mathbf{0}.$$

Hence $|\mathbf{Ax}| = |\lambda||\mathbf{x}| = r|\mathbf{x}| = \mathbf{A}|\mathbf{x}|$. Read the first entry: with $z_j = a_{1j}x_j$,

$$\Big|\sum_j z_j\Big| = |(\mathbf{Ax})_1| = (\mathbf{A}|\mathbf{x}|)_1 = \sum_j a_{1j}|x_j| = \sum_j |z_j|,$$

and $\sum_j z_j = (\mathbf{Ax})_1 = \lambda x_1 \neq 0$, because $|\lambda x_1| = r|x_1| > 0$. By Lemma 19.1.3 there is $c \in \mathbb{C}$ with $|c| = 1$ and $a_{1j}x_j = a_{1j}|x_j|\,c$ for every j. Since $a_{1j} > 0$, we may divide: $x_j = c|x_j|$, that is, $\mathbf{x} = c|\mathbf{x}|$. Therefore

$$\lambda\mathbf{x} = \mathbf{Ax} = c\,\mathbf{A}|\mathbf{x}| = c\,r|\mathbf{x}| = r\mathbf{x},$$

and as $\mathbf{x} \neq \mathbf{0}$, $\lambda = r$. So every eigenvalue other than r has modulus $< r$, by Step 3. This proves (c). We record what the step showed for $\lambda = r$ itself: **every eigenvector $\mathbf{x}$ for r has the form $c|\mathbf{x}|$ with $|c| = 1$, $|\mathbf{x}| > \mathbf{0}$ and $\mathbf{A}|\mathbf{x}| = r|\mathbf{x}|$.**

Step 5: $E_r(\mathbf{A}) = \operatorname{span}(\mathbf{v})$. Let $\mathbf{x}$ be an eigenvector for r, and write $\mathbf{x} = c\,\mathbf{u}$ with $\mathbf{u} = |\mathbf{x}| > \mathbf{0}$ and $\mathbf{Au} = r\mathbf{u}$, as just recorded. Let $\theta = \min_i u_i/v_i$, which is positive, and put $\mathbf{y} = \mathbf{u} - \theta\mathbf{v}$. Then $\mathbf{y}$ is real, $\mathbf{y} \geq \mathbf{0}$ by the choice of θ, $y_i = 0$ for an index i attaining the minimum, and $\mathbf{Ay} = r\mathbf{y}$. If $\mathbf{y} \neq \mathbf{0}$, the push-up claim would give $\mathbf{y} > \mathbf{0}$, contradicting $y_i = 0$. So $\mathbf{y} = \mathbf{0}$, and $\mathbf{x} = c\theta\,\mathbf{v} \in \operatorname{span}(\mathbf{v})$. Hence $E_r(\mathbf{A}) = \operatorname{span}(\mathbf{v})$ and $g_{\mathbf{A}}(r) = 1$.

Step 6: $a_{\mathbf{A}}(r) = 1$. First a positive left eigenvector. $\mathbf{A}^\top$ is positive, since it has the same entries, and Steps 1–5 used nothing about $\mathbf{A}$ but its positivity. Applied to $\mathbf{A}^\top$, they give $\mathbf{w} > \mathbf{0}$ with $\mathbf{A}^\top\mathbf{w} = r'\mathbf{w}$, where $r' = \rho(\mathbf{A}^\top)$. By Theorem 0.9.5, $(\mathbf{A}^\top\mathbf{w})^\top = \mathbf{w}^\top\mathbf{A}$, so

$$r'\,\mathbf{w}^\top\mathbf{v} = (\mathbf{A}^\top\mathbf{w})^\top\mathbf{v} = \mathbf{w}^\top\mathbf{Av} = r\,\mathbf{w}^\top\mathbf{v}.$$

Here $\mathbf{w}^\top\mathbf{v} = \sum_i w_i v_i > 0$, because every term is positive, so $r' = r$. Hence $\rho(\mathbf{A}^\top) = r$ and $\mathbf{w}^\top\mathbf{A} = r\mathbf{w}^\top$.

Now suppose $a_{\mathbf{A}}(r) \geq 2$. The characteristic polynomial splits over $\mathbb{C}$ (Corollary 6.5.5), so Theorem 10.1.3 (c) applies to $\mathbf{A} \in M_n(\mathbb{C})$ and gives $\dim G_r(\mathbf{A}) = a_{\mathbf{A}}(r) \geq 2 > 1 = \dim E_r(\mathbf{A})$. Choose $\mathbf{x} \in G_r(\mathbf{A})$ with $\mathbf{x} \notin E_r(\mathbf{A})$, put $\mathbf{N} = \mathbf{A} - r\mathbf{I}$, and let $k \geq 1$ be least with $\mathbf{N}^k\mathbf{x} = \mathbf{0}$, which exists by Definition 10.1.2 and Theorem 0.2.11. Since $\mathbf{Nx} \neq \mathbf{0}$, $k \geq 2$. Put $\mathbf{y} = \mathbf{N}^{k-2}\mathbf{x}$. Then $\mathbf{Ny} = \mathbf{N}^{k-1}\mathbf{x} \neq \mathbf{0}$ by the minimality of k, and $\mathbf{N}(\mathbf{Ny}) = \mathbf{N}^k\mathbf{x} = \mathbf{0}$. So $\mathbf{Ny}$ is an eigenvector for r, and by Step 5, $\mathbf{Ny} = b\mathbf{v}$ with $b \neq 0$. But $\mathbf{w}^\top\mathbf{N} = \mathbf{w}^\top\mathbf{A} - r\mathbf{w}^\top = \mathbf{0}^\top$, so

$$0 = (\mathbf{w}^\top\mathbf{N})\mathbf{y} = \mathbf{w}^\top(\mathbf{Ny}) = b\,\mathbf{w}^\top\mathbf{v}.$$

Since $\mathbf{w}^\top\mathbf{v} > 0$ and $b \neq 0$, this is a contradiction. Hence $a_{\mathbf{A}}(r) \leq 1$, and $a_{\mathbf{A}}(r) \geq 1$ because r is an eigenvalue. Together with Step 5 this proves (b).

Step 7: nonnegative eigenvectors. Let $\mathbf{x} \geq \mathbf{0}$, $\mathbf{x} \neq \mathbf{0}$, be an eigenvector for an eigenvalue λ. If $\lambda \neq r$, then Theorem 9.10.6, applied over $\mathbb{C}$ to the left eigenvector $\mathbf{w}$ for r and the right eigenvector $\mathbf{x}$ for λ, gives $\mathbf{w}^\top\mathbf{x} = 0$. But $\mathbf{w}^\top\mathbf{x} > 0$ by Lemma 19.1.1 (b), applied to the positive $1 \times n$ matrix $\mathbf{w}^\top$. So $\lambda = r$, and by Step 5, $\mathbf{x} = \theta\mathbf{v}$ for a scalar θ. Comparing any entry, $\theta = x_i/v_i$ is real and ≥ 0, and it is not 0 because $\mathbf{x} \neq \mathbf{0}$. This proves (d).

Finally, (e): $\mathbf{A}^\top$ is positive, and Step 6 showed $\rho(\mathbf{A}^\top) = r$ by pairing with $\mathbf{v}$, so (a)–(d) hold for $\mathbf{A}^\top$ by what has been proved, with the positive eigenvector $\mathbf{w}$ of Step 6. This completes the proof.

■

Remark. **Which analysis was used.** Exactly two facts from Chapter 16's introduction, both in Step 1: (A3), that the closed bounded set S is compact, and (A4), that the continuous function t attains its maximum on it, together with the algebra of limits from the same introduction, which showed that S is closed. Everything else is algebra: the order rules, the triangle inequality with its equality case, and the generalized eigenspaces of Chapter 10.

Two examples show what the theorem says.

Example. Perron's theorem for Chapter 9's first example

Check each part of Theorem 19.1.5 for $\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 3 & 2 \end{pmatrix}$.

Solution. Example (Eigenvalues and Eigenspaces of a 2×2 Matrix) found $p_{\mathbf{A}}(x) = (x-4)(x+1)$, $E_4(\mathbf{A}) = \operatorname{span}((2,3))$ and $E_{-1}(\mathbf{A}) = \operatorname{span}((1,-1))$. So $r = \rho(\mathbf{A}) = 4 > 0$ is an eigenvalue, with the positive eigenvector $\mathbf{v} = (2,3)$: part (a). It is a simple root of $p_{\mathbf{A}}$: part (b). The other eigenvalue has $|-1| = 1 < 4$: part (c). Every eigenvector for -1 is $t(1,-1)$ with $t \neq 0$, which has entries of opposite signs, so none is nonnegative: part (d). For (e), the columns of $\mathbf{A}$ both sum to 4, so $\mathbf{1}^{\top}\mathbf{A} = 4\,\mathbf{1}^{\top}$ and $\mathbf{w} = (1,1)$ is a positive left eigenvector for 4, as the text after Definition 9.10.1 already observed.

Example. A positive 3×3 matrix

Let $\mathbf{A} = \begin{pmatrix} 1 & 3 & 1 \\ 2 & 2 & 1 \\ 4 & 2 & 3 \end{pmatrix}$. Find its eigenvalues and eigenvectors, and its positive left eigenvector.

Solution. Try the three vectors $(1,1,2)$, $(-1,-1,3)$ and $(-5,2,4)$:

$$\mathbf{A}\begin{pmatrix} 1 \\ 1 \\ 2 \end{pmatrix} = \begin{pmatrix} 6 \\ 6 \\ 12 \end{pmatrix}, \qquad \mathbf{A}\begin{pmatrix} -1 \\ -1 \\ 3 \end{pmatrix} = \begin{pmatrix} -1 \\ -1 \\ 3 \end{pmatrix}, \qquad \mathbf{A}\begin{pmatrix} -5 \\ 2 \\ 4 \end{pmatrix} = \begin{pmatrix} 5 \\ -2 \\ -4 \end{pmatrix}.$$

So 6, 1 and -1 are eigenvalues. They are distinct, and a 3×3 matrix has at most three eigenvalues (Corollary 9.2.3), so these are all, each with algebraic multiplicity 1 because the multiplicities add up to 3 (Corollary 6.5.5 (a)). As checks, $\operatorname{tr}\mathbf{A} = 6 = 6 + 1 - 1$ and $\det\mathbf{A} = -6 = 6 \cdot 1 \cdot (-1)$. Hence $\rho(\mathbf{A}) = 6$, with the positive eigenvector $\mathbf{v} = (1,1,2)$, and the other two eigenvalues have modulus $1 < 6$. The eigenvectors for 1 and -1 are the nonzero multiples of $(-1,-1,3)$ and $(-5,2,4)$, and each has entries of both signs, as Theorem 19.1.5 (d) requires.

For the left eigenvector, solve $(\mathbf{A}^{\top} - 6\mathbf{I})\mathbf{w} = \mathbf{0}$. Row reduction gives $\mathbf{w} = (10,11,7)$, and indeed

$$\mathbf{w}^{\top}\mathbf{A} = \begin{pmatrix} 10+22+28 & 30+22+14 & 10+11+21 \end{pmatrix} = \begin{pmatrix} 60 & 66 & 42 \end{pmatrix} = 6\,\mathbf{w}^{\top}.$$

The left and right eigenvectors for $\rho(\mathbf{A})$ are unrelated in general; here $\mathbf{w}^\top\mathbf{v} = 10+11+14 = 35$ for $\mathbf{v} = (1,1,2)$. Definition 19.1.2 below scales them so that this pairing is 1.

Quick check. The swap $\mathbf{S} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ is nonnegative. Which conclusions of Theorem 19.1.5 hold for $\mathbf{S}$, and which fail?

Solution. $\mathbf{S}$ has eigenvalues 1 and -1, with eigenvectors $(1,1)$ and $(1,-1)$. So $\rho(\mathbf{S}) = 1$ is an eigenvalue with a positive eigenvector, it is simple, and the only nonnegative eigenvectors are the positive multiples of $(1,1)$: (a), (b), (d) and (e) hold. Part (c) fails, since $-1 \neq 1$ and $|-1| = 1 = \rho(\mathbf{S})$. The hypothesis that fails is $\mathbf{S} > \mathbf{O}$: two entries are 0. This is the matrix whose powers do not converge in the warning of Chapter 9 §11, and (c) is exactly what was missing there.

The positive eigenvector is so useful that it gets a name.

Definition 19.1.2. Perron Root, Perron Vector

Let $\mathbf{A} \in M_n(\mathbb{R})$ be positive. The number $\rho(\mathbf{A})$ is the **Perron root** of $\mathbf{A}$. The **Perron vector** of $\mathbf{A}$ is the unique $\mathbf{v} > \mathbf{0}$ with $\mathbf{A}\mathbf{v} = \rho(\mathbf{A})\mathbf{v}$ and $v_1 + \cdots + v_n = 1$. The **left Perron vector** of $\mathbf{A}$ is the unique $\mathbf{w} > \mathbf{0}$ with $\mathbf{w}^\top\mathbf{A} = \rho(\mathbf{A})\mathbf{w}^\top$ and $\mathbf{w}^\top\mathbf{v} = 1$, where $\mathbf{v}$ is the Perron vector.

Existence and uniqueness are Theorem 19.1.5 (a) and (b): the eigenspace for $\rho(\mathbf{A})$ is a line through a positive vector, and it meets the hyperplane $\{\mathbf{1}^\top\mathbf{x} = 1\}$ in exactly one point, the vector $\mathbf{v}/(\mathbf{1}^\top\mathbf{v})$. By (d), the Perron vector is also the **only** eigenvector of $\mathbf{A}$ in Δ_n. Likewise, by (e) and (b) applied to $\mathbf{A}^\top$, the positive left eigenvectors for $\rho(\mathbf{A})$ are the positive multiples of one $\mathbf{w}_0 > \mathbf{0}$, and since $\mathbf{w}_0^\top\mathbf{v} > 0$, exactly one of them, $\mathbf{w}_0/(\mathbf{w}_0^\top\mathbf{v})$, pairs with the Perron vector to 1. The two normalizations differ on purpose: $\mathbf{v}$ is a probability vector, and $\mathbf{w}^\top\mathbf{v} = 1$ makes $\mathbf{v}\mathbf{w}^\top$ a projection, the limit of $(\mathbf{A}/\rho(\mathbf{A}))^k$ that §05 finds. For Example (Perron's theorem for Chapter 9's first example) the Perron vector is $\frac{1}{5}(2,3)$ and the left Perron vector is $(1,1)$, since $\mathbf{1}^\top\mathbf{v} = 1$. For Example (A positive 3×3 matrix) the Perron vector is $\frac{1}{4}(1,1,2)$, and the left Perron vector is $\frac{4}{35}(10,11,7)$, since $(10,11,7)$ pairs with $\frac{1}{4}(1,1,2)$ to $\frac{35}{4}$.

Warning. **Perron's theorem says nothing about the other eigenvalues except their size.** They may be negative, as -1 is in Example (Perron's theorem for Chapter 9's first example), or nonreal: the positive matrix $\left(\begin{smallmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \\ 2 & 3 & 1 \end{smallmatrix}\right)$ has row sums 6, hence the eigenvector $\mathbf{1}$ for 6, and its other two eigenvalues are the roots of $x^2 + 3x + 3$, namely $\frac{-3 \pm i\sqrt{3}}{2}$, of modulus $\sqrt{3} < 6$. They may also be repeated, as the eigenvalue 1 of $\left(\begin{smallmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{smallmatrix}\right)$ is. Simplicity is claimed for $\rho(\mathbf{A})$ alone.

19.1.4 What positivity buys, and what it costs

Each hypothesis of Theorem 19.1.5 has a job, and it is instructive to see where it went. **Positivity** of $\mathbf{A}$ was used in the push-up claim, to turn a nonnegative, nonzero slack into a strictly positive one; in Step 4, to divide by a_{1j}; and in Step 1, to make $r > 0$. For a matrix that is only nonnegative, each of these uses fails, and so can each conclusion: Exercise C2 gives a nonnegative 2×2 matrix for each of (a)–(d), and the swap of the Quick check is the one for (c).

§03 shows that one conclusion survives in general, that $\rho(\mathbf{A})$ is an eigenvalue with a nonnegative eigenvector, and that most of the rest returns under a hypothesis much weaker than positivity. That hypothesis, **irreducibility**, is a condition on where the zero entries are, and §02 introduces it.

Here is the payoff for Chapter 9. A stochastic matrix has $\mathbf{1}^{\top}\mathbf{A} = \mathbf{1}^{\top}$, so $\mathbf{1}$ is a positive left eigenvector for 1. If $\mathbf{A}$ is stochastic and positive, Theorem 19.1.5 (d) applied to $\mathbf{A}^{\top}$ says that $\mathbf{1}$ is a left eigenvector for $\rho(\mathbf{A})$ and for no other eigenvalue, so $\rho(\mathbf{A}) = 1$. Then (b) and (c) say that 1 is simple and every other eigenvalue has modulus less than 1: two of the three hypotheses of Theorem 9.11.4 hold automatically. The third, diagonalizability, is the one §06 removes.

19.1.5 Exercises

19.1.5.1 A. Check your understanding

Exercise. A1

(a) Define what it means for a real matrix to be **nonnegative** and to be **positive**, and define $|\mathbf{M}|$ for a complex matrix $\mathbf{M}$.

(b) State Theorem 19.1.5.

(c) True or false: a positive symmetric matrix is positive definite. Justify your answer.

(d) True or false: a positive matrix has no negative eigenvalue. Justify your answer.

(e) True or false: if $\mathbf{O} \le \mathbf{A} \le \mathbf{B}$ and $\mathbf{A} \neq \mathbf{B}$, then $\rho(\mathbf{A}) < \rho(\mathbf{B})$. Justify your answer.

(f) In the proof of Theorem 19.1.5, which step uses compactness, and why is the quotient $\min_i(\mathbf{Ax})_i/x_i$ not maximized directly?

Solution.

(a) $\mathbf{A}$ is nonnegative if every entry is ≥ 0, and positive if every entry is > 0 (Definition 19.1.1). $|\mathbf{M}|$ is the real matrix whose (i,j) entry is $|m_{ij}|$.

(b) If $\mathbf{A} \in M_n(\mathbb{R})$ and $\mathbf{A} > \mathbf{O}$, then $r = \rho(\mathbf{A}) > 0$ is an eigenvalue of $\mathbf{A}$ with a positive eigenvector $\mathbf{v}$; it is algebraically simple; every other eigenvalue has modulus less than r; every nonnegative eigenvector is a positive multiple of $\mathbf{v}$; and the same holds for $\mathbf{A}^{\top}$, which gives a positive left eigenvector.

(c) False. $\begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}$ is positive and symmetric with the eigenvalue -1; the warning after Definition 19.1.1 computes $\mathbf{x}^{\top}\mathbf{Ax} = -2$ for $\mathbf{x} = (1, -1)$.

(d) False. $\begin{pmatrix} 1 & 2 \\ 3 & 2 \end{pmatrix}$ has the eigenvalue -1 (Example (Perron's theorem for Chapter 9's first example)). Perron's theorem controls only the modulus of the other eigenvalues.

(e) False. $\operatorname{diag}(1, 0) \le \mathbf{I}_2$, the two differ, and both have spectral radius 1.

(f) Step 1: the set S is closed and bounded, so by (A3) and (A4) the coordinate t attains a maximum on it. The quotient is undefined where some $x_i = 0$. For a nonnegative matrix its natural repair can be discontinuous on the boundary of Δ_n (for $\mathrm{diag}(2,1)$ it jumps at $\mathbf{e}_1$), so the extreme value theorem does not apply to it directly. For positive $\mathbf{A}$ the repair is in fact continuous, but proving that costs more than maximizing t over S.

19.1.5.2 B. Practice

Exercise. B1: A Perron vector by hand

Let $\mathbf{A} = \begin{pmatrix} 2 & 1 & 1 \\ 3 & 4 & 3 \\ 3 & 1 & 4 \end{pmatrix}$. Find the eigenvalues of $\mathbf{A}$, its Perron root, Perron vector, and left Perron vector. Check that no eigenvector for the other eigenvalues is nonnegative.

Solution. Expanding $\det(x\mathbf{I} - \mathbf{A})$ along the first row,

$$\begin{aligned} p_{\mathbf{A}}(x) &= (x-2)((x-4)^2 - 3) + (-3(x-4) - 9) - (3 + 3(x-4)) \\ &= x^3 - 10x^2 + 23x - 14 = (x-1)(x-2)(x-7). \end{aligned}$$

(Check: $\operatorname{tr}\mathbf{A} = 10$ and $\det\mathbf{A} = 14$.) So the eigenvalues are $7, 2, 1$, and $\rho(\mathbf{A}) = 7$. Solving $(\mathbf{A} - 7\mathbf{I})\mathbf{x} = \mathbf{0}$, whose first two rows are $-5x_1 + x_2 + x_3 = 0$ and $3x_1 - 3x_2 + 3x_3 = 0$, gives $\mathbf{x} = t(1,3,2)$; indeed $\mathbf{A}(1,3,2) = (7,21,14)$. The entries of $(1,3,2)$ add up to 6, so the Perron vector is $\frac{1}{6}(1,3,2)$.

For the left eigenvector, $(\mathbf{A}^\top - 7\mathbf{I})\mathbf{w} = \mathbf{0}$ gives the multiples of $(3,2,3)$, and $(3,2,3)\mathbf{A} = (21,14,21) = 7\,(3,2,3)$. Against the Perron vector, $(3,2,3)\cdot\frac{1}{6}(1,3,2) = \frac{15}{6} = \frac{5}{2}$, so by Definition 19.1.2 the left Perron vector is $\frac{2}{5}(3,2,3) = (\frac{6}{5}, \frac{4}{5}, \frac{6}{5})$.

For 2: $(\mathbf{A} - 2\mathbf{I})\mathbf{x} = \mathbf{0}$ gives $\mathbf{x} = t(1,3,-3)$, since $\mathbf{A}(1,3,-3) = (2,6,-6)$. For 1: $\mathbf{x} = t(1,0,-1)$, since $\mathbf{A}(1,0,-1) = (1,0,-1)$. For $t \neq 0$ each has entries of both signs, so neither is nonnegative, as Theorem 19.1.5 (d) predicts.

Exercise. B2: Which order?

For each matrix, decide whether it is nonnegative, whether it is positive, and whether it is positive definite. Justify your answers, and say whether Theorem 19.1.5 applies.

(a) $\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$

(b) $\begin{pmatrix} 1 & 3 \\ 3 & 1 \end{pmatrix}$

(c) $\begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}$

(d) $\begin{pmatrix} 1 & 0 \\ 2 & 3 \end{pmatrix}$

(e) $\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$

Solution. For a real symmetric 2×2 matrix, positive definite means both eigenvalues are positive (Definition 13.1.1, with the spectral theorem).

(a) Positive, hence nonnegative. Eigenvalues $3, 1$, both positive, so positive definite. Perron applies: $\rho = 3$ with eigenvector $(1, 1)$.

(b) Positive, hence nonnegative. Eigenvalues $4, -2$, so **not** positive definite. Perron applies: $\rho = 4$ with eigenvector $(1, 1)$, and $|-2| < 4$.

(c) Not nonnegative, since the off-diagonal entries are -1; so not positive, and Perron does not apply. It is positive definite, with eigenvalues $1, 3$.

(d) Nonnegative, not positive (the $(1, 2)$ entry is 0), and not symmetric, so not positive definite by clause (P1). Perron does not apply, although here $\rho = 3$ is an eigenvalue with the eigenvector $\mathbf{e}_2 \geq \mathbf{0}$, which is not positive.

(e) Positive. Eigenvalues 2 and 0, so positive semidefinite but **not** positive definite, as $\mathbf{x} = (1, -1)$ gives $\mathbf{x}^\top \mathbf{A}\mathbf{x} = 0$. Perron applies: $\rho = 2$, eigenvector $(1, 1)$, and the eigenvalue 0 has modulus less than 2.

Exercise. B3: Bounds from the smallest and largest entry

(a) Show that $\rho(c\mathbf{J}) = nc$ for every real $c \geq 0$, where $\mathbf{J}$ is the $n \times n$ all-ones matrix.

(b) Let $\mathbf{A} \in M_n(\mathbb{R})$ with $a \leq a_{ij} \leq b$ for all i, j, where $0 \leq a \leq b$. Prove that $na \leq \rho(\mathbf{A}) \leq nb$.

(c) Hence bound $\rho(\mathbf{A})$ for $\mathbf{A} = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \\ 3 & 1 & 2 \end{pmatrix}$, and compare with its exact value.

Solution.

(a) $\mathbf{J}\mathbf{1} = n\mathbf{1}$, so n is an eigenvalue of $\mathbf{J}$. Also $\mathbf{J}^2 = n\mathbf{J}$, since every entry of $\mathbf{J}^2$ is a sum of n ones. If $\mathbf{J}\mathbf{x} = \lambda\mathbf{x}$ with $\mathbf{x} \neq \mathbf{0}$, then $\lambda^2\mathbf{x} = \mathbf{J}^2\mathbf{x} = n\mathbf{J}\mathbf{x} = n\lambda\mathbf{x}$, so $\lambda^2 = n\lambda$ and $\lambda \in \{0, n\}$. Hence $\rho(\mathbf{J}) = n$, and $\rho(c\mathbf{J}) = c\,\rho(\mathbf{J}) = nc$ because the eigenvalues of $c\mathbf{J}$ are c times those of $\mathbf{J}$.

(b) The hypothesis says $\mathbf{O} \leq a\mathbf{J} \leq \mathbf{A} \leq b\mathbf{J}$. By Proposition 19.1.4 (a), applied twice, $\rho(a\mathbf{J}) \leq \rho(\mathbf{A}) \leq \rho(b\mathbf{J})$, and by (a) this reads $na \leq \rho(\mathbf{A}) \leq nb$.

(c) The entries lie between 1 and 3 and $n = 3$, so $3 \leq \rho(\mathbf{A}) \leq 9$. In fact every row sums to 6, so $\mathbf{A}\mathbf{1} = 6\,\mathbf{1}$, and Theorem 19.1.5 (d) says the positive eigenvector $\mathbf{1}$ belongs to $\rho(\mathbf{A})$. Hence $\rho(\mathbf{A}) = 6$.

19.1.5.3 C. Going deeper

Exercise. C1: Symmetry passes to the Perron vector

Let $\mathbf{A} > \mathbf{O}$ be $n \times n$ with Perron vector $\mathbf{v}$, and let $\mathbf{P}$ be a permutation matrix with $\mathbf{P}^\top\mathbf{A}\mathbf{P} =$

$\mathbf{A}$.

(a) Prove that $\mathbf{Pv} = \mathbf{v}$.

(b) Deduce that the Perron vector of $\mathbf{A} = \begin{pmatrix} 1 & 1 & 3 \\ 1 & 1 & 3 \\ 1 & 1 & 1 \end{pmatrix}$ has equal first and second entries, and find it.

(c) Deduce that if $\mathbf{A} > \mathbf{O}$ commutes with the cyclic permutation matrix of $1 \to 2 \to \cdots \to n \to 1$, then its Perron vector is $\frac{1}{n}\mathbf{1}$.

Hint: for (a), $\mathbf{Pv}$ is a positive vector in Δ_n.

Solution.

(a) By Lemma 3.4.7 (b), $\mathbf{PP}^\top = \mathbf{I}$, so $\mathbf{AP} = \mathbf{PP}^\top\mathbf{AP} = \mathbf{PA}$. Hence $\mathbf{A}(\mathbf{Pv}) = \mathbf{PAv} = \rho(\mathbf{A})\,\mathbf{Pv}$. The entries of $\mathbf{Pv}$ are those of $\mathbf{v}$ in another order ($\mathbf{P} = \mathbf{P}_\sigma$ sends $\mathbf{e}_j$ to $\mathbf{e}_{\sigma(j)}$), so $\mathbf{Pv} > \mathbf{0}$ and its entries add up to 1. By Definition 19.1.2 and the uniqueness proved after it, $\mathbf{Pv} = \mathbf{v}$.

(b) Let $\mathbf{P}$ be the permutation matrix that swaps $\mathbf{e}_1$ and $\mathbf{e}_2$. Then $\mathbf{P}^\top\mathbf{AP}$ is $\mathbf{A}$ with rows $1, 2$ swapped and columns $1, 2$ swapped, which is $\mathbf{A}$ again, because rows 1 and 2 of $\mathbf{A}$ are equal and so are columns 1 and 2. By (a), $\mathbf{Pv} = \mathbf{v}$, that is, $v_1 = v_2$. Now try $\mathbf{v} = (1, 1, c)$: $\mathbf{Av} = (2 + 3c, 2 + 3c, 2 + c)$, and $\mathbf{Av} = \lambda\mathbf{v}$ requires $\lambda = 2 + 3c$ and $2 + c = (2 + 3c)c$, that is, $3c^2 + c - 2 = 0$, so $c = \frac{2}{3}$ or $c = -1$. The positive choice $c = \frac{2}{3}$ gives the eigenvector $(3, 3, 2)$ with eigenvalue 4, and by Theorem 19.1.5 (d) this is the Perron eigenvector, so $\rho(\mathbf{A}) = 4$ and the Perron vector is $\frac{1}{8}(3, 3, 2)$.

(c) Let $\mathbf{C}$ be the cyclic permutation matrix, $\mathbf{Ce}_j = \mathbf{e}_{j+1}$ for $j < n$ and $\mathbf{Ce}_n = \mathbf{e}_1$. If $\mathbf{AC} = \mathbf{CA}$, then $\mathbf{C}^\top\mathbf{AC} = \mathbf{C}^\top\mathbf{CA} = \mathbf{A}$ by Lemma 3.4.7 (b). By (a), $\mathbf{Cv} = \mathbf{v}$, which says $v_n = v_1$ and $v_j = v_{j+1}$ for $j < n$, so all entries are equal. As they add up to 1, $\mathbf{v} = \frac{1}{n}\mathbf{1}$.

Exercise. C2: Every conclusion fails for some nonnegative matrix

For each part of Theorem 19.1.5, (a) to (d), give a nonnegative 2×2 matrix for which that part fails, and name the step of the proof that breaks. Then give a nonnegative matrix for which (a), (b) and (c) fail at the same time.

Solution.

(a) $\mathbf{J}_2(0) = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$ has $\rho = 0$, and its eigenvectors are the nonzero multiples of $\mathbf{e}_1$, none positive. In the proof, $r > 0$ came from the smallest row sum, which here is 0; and the push-up claim fails, since $\mathbf{u} = \mathbf{e}_1$ satisfies $\mathbf{Au} = 0 \cdot \mathbf{u}$ without being positive.

(b) $\mathbf{I}_2$: $\rho = 1$ is a double eigenvalue and $E_1 = \mathbb{R}^2$. Step 5 fails: $\mathbf{u} - \theta\mathbf{v}$ can be nonzero and nonnegative with a zero entry, since $\mathbf{Ay} > \mathbf{0}$ is no longer forced.

(c) $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$: the eigenvalue -1 has modulus $\rho = 1$. Step 4 fails: equality in the triangle inequality (Lemma 19.1.3) is automatic when each row has a single nonzero term, so it aligns nothing.

(d) $\operatorname{diag}(2,1)$: the eigenvector $\mathbf{e}_2 \geq \mathbf{0}$ belongs to $1 \neq 2 = \rho$. Step 7 fails, because the left eigenvector for 2 is $\mathbf{e}_1$, which is not positive, and $\mathbf{e}_1^{\top}\mathbf{e}_2 = 0$ is no contradiction.

For all three at once, take

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

It is block upper triangular with diagonal blocks $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ and (1), so $p_{\mathbf{A}}(x) = (x^2-1)(x-1) = (x-1)^2(x+1)$ by Theorem 7.8.4, applied to $x\mathbf{I} - \mathbf{A}$ over $F[x]$, as the paragraph after it notes. Hence $\rho(\mathbf{A}) = 1$ and $a_{\mathbf{A}}(1) = 2$, so (b) fails, and $|-1| = \rho(\mathbf{A})$, so (c) fails. The system $(\mathbf{A}-\mathbf{I})\mathbf{x} = \mathbf{0}$ reads $-x_1 + x_2 + x_3 = 0$ and $x_1 - x_2 = 0$, which forces $x_3 = 0$; so every eigenvector for 1 is a multiple of $(1,1,0)$, none of them positive, and (a) fails.

Exercise. C3: The Perron root is strictly monotone

Let $\mathbf{A}, \mathbf{B} \in M_n(\mathbb{R})$ with $\mathbf{O} < \mathbf{A} \leq \mathbf{B}$ and $\mathbf{A} \neq \mathbf{B}$.

(a) Prove that $\rho(\mathbf{A}) < \rho(\mathbf{B})$.

(b) Show by an example that the conclusion fails if $\mathbf{A} > \mathbf{O}$ is weakened to $\mathbf{A} \geq \mathbf{O}$.

Hint: pair $\mathbf{Bv}$ *with a positive left eigenvector of* $\mathbf{B}$*, where* $\mathbf{v}$ *is the Perron vector of* $\mathbf{A}$*.*

Solution.

(a) $\mathbf{B} \geq \mathbf{A} > \mathbf{O}$, so $\mathbf{B}$ is positive too. Let $\mathbf{v} > \mathbf{0}$ be the Perron vector of $\mathbf{A}$, and by Theorem 19.1.5 (e) let $\mathbf{w} > \mathbf{0}$ satisfy $\mathbf{w}^{\top}\mathbf{B} = \rho(\mathbf{B})\mathbf{w}^{\top}$. Then

$$\rho(\mathbf{B})\,\mathbf{w}^{\top}\mathbf{v} = \mathbf{w}^{\top}\mathbf{B}\mathbf{v} = \mathbf{w}^{\top}\mathbf{A}\mathbf{v} + \mathbf{w}^{\top}(\mathbf{B}-\mathbf{A})\mathbf{v} = \rho(\mathbf{A})\,\mathbf{w}^{\top}\mathbf{v} + \mathbf{w}^{\top}(\mathbf{B}-\mathbf{A})\mathbf{v}.$$

The matrix $\mathbf{B}-\mathbf{A}$ is nonnegative and nonzero, say $b_{kl} - a_{kl} > 0$, so $\mathbf{w}^{\top}(\mathbf{B}-\mathbf{A})\mathbf{v} \geq w_k(b_{kl}-a_{kl})v_l > 0$, every other term being ≥ 0. Since $\mathbf{w}^{\top}\mathbf{v} > 0$, dividing gives $\rho(\mathbf{B}) > \rho(\mathbf{A})$.

(b) $\mathbf{A} = \operatorname{diag}(1,0)$ and $\mathbf{B} = \mathbf{I}_2$ satisfy $\mathbf{O} \leq \mathbf{A} \leq \mathbf{B}$ and $\mathbf{A} \neq \mathbf{B}$, with $\rho(\mathbf{A}) = \rho(\mathbf{B}) = 1$. In the argument of (a), the Perron vector of $\mathbf{A}$ is replaced by $\mathbf{e}_1$, and $(\mathbf{B}-\mathbf{A})\mathbf{e}_1 = \mathbf{0}$: the extra mass sits where the eigenvector does not look.

19.2 Irreducibility and Graphs

Perron's theorem (Theorem 19.1.5) asks for every entry to be positive, and most matrices that arise in practice do not oblige. The bicycle chain of Chapter 9 §11 has a zero entry, since a bicycle at station 3 never reaches station 1 in one night; the Leslie matrix of the same section is mostly zeros; the swap has only two nonzero entries. Some zeros are harmless and some are fatal: the swap keeps $\rho = 1$ as a simple eigenvalue with a positive eigenvector, while for $\mathbf{I}_2$ the eigenvalue 1 is double and its positive eigenvectors are no longer all multiples of one. This section finds the dividing line. It depends only on **where** the zero entries are, and the right way to read that pattern is as a map of one-step moves between the indices: a directed graph. The

matrices on the good side of the line are the **irreducible** ones, and §03 proves that they enjoy nearly all of Perron's conclusions.

Throughout, F is $\mathbb{R}$ or $\mathbb{C}$. The definitions of this section use only whether an entry is zero, so they make sense for any matrix. The statements about powers and positivity need $\mathbf{A} \geq \mathbf{O}$, and say so.

19.2.1 The graph of a matrix

The question that recurs is this: which entries of $\mathbf{A}^k$ are nonzero? For a nonnegative matrix, $(\mathbf{A}^2)_{ij} = \sum_l a_{il}a_{lj}$ is positive exactly when some a_{il} and a_{lj} are both positive, that is, when one can go from j to some l and from l to i along nonzero entries. In a Markov chain (Definition 9.11.2), a_{lj} is the probability of moving from state j to state l in one step, and the sentence says that state i can be reached from j in two steps. So we draw an arrow from j to i for each nonzero entry a_{ij} and count routes.

The graph of a matrix has a vertex for each index and an arrow from j to i whenever the entry a_{ij} is not zero.

Definition 19.2.1. Directed Graph of a Matrix; Walks and Paths

Let $n \geq 1$ and $\mathbf{A} = (a_{ij}) \in M_n(F)$.

(a) The **directed graph** $G(\mathbf{A})$ of $\mathbf{A}$ has **vertices** $1, 2, \ldots, n$, and an **edge from j to i**, written $j \to i$, for **each** pair (i, j) with $a_{ij} \neq 0$. An edge $j \to j$, from a nonzero diagonal entry, is a **loop**.

(b) A **walk of length $k \geq 0$ from j to i** is a list of vertices $j = i_0, i_1, \ldots, i_k = i$ such that $i_{t-1} \to i_t$ is an edge for **every** $t = 1, \ldots, k$.

(c) A **path** is a walk whose vertices $i_0, \ldots, i_k$ are distinct.

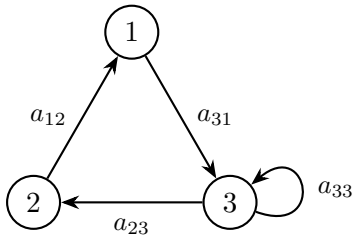
In words: (a) turns each nonzero entry into an arrow, and the arrow for a_{ij} points **from the column index to the row index**. That is the direction in which $\mathbf{A}$ moves things: in $\mathbf{x} \mapsto \mathbf{Ax}$, entry j of $\mathbf{x}$ contributes $a_{ij}x_j$ to entry i of the image, and in a column-stochastic chain a_{ij} is the probability of the move from j to i. Clause (b) allows $k = 0$: the one-vertex list (j) is a walk of length 0 from j to itself, whatever the matrix. A walk may revisit vertices and use an edge several times; a path may not.

Remark. Many books draw the arrow for $a_{ij} \neq 0$ from i to j. The two graphs differ by reversing every edge, and nothing below depends on the choice except the wording "from j to i". We take the direction that matches $\mathbf{x}_{k+1} = \mathbf{Ax}_k$ and the column-stochastic convention of Chapter 9 §11.

Examples, simplest first.

- $n = 1$. $G((a))$ has one vertex, with a loop if $a \neq 0$ and no edge if $a = 0$. In either case there is the walk of length 0 from 1 to 1.
- **The zero matrix $\mathbf{O} \in M_n(F)$** has n vertices and no edges at all. **A positive matrix** has every possible edge, loops included.
- **A cycle.** The permutation matrix $\mathbf{C} = \left(\begin{smallmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{smallmatrix}\right)$ has $c_{21} = c_{32} = c_{13} = 1$, so its edges are $1 \to 2$, $2 \to 3$ and $3 \to 1$.

- **A path.** $\left(\begin{smallmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{smallmatrix}\right)$ has the edges $1 \to 2$ and $2 \to 3$ only. There is a walk from 1 to 3, and none from 3 to anywhere but 3 itself.
- $\mathbf{A} = \left(\begin{smallmatrix} 0 & 2 & 0 \\ 0 & 0 & 1 \\ 3 & 0 & 1 \end{smallmatrix}\right)$ has the edges $2 \to 1$, $3 \to 2$, $1 \to 3$ and the loop $3 \to 3$, drawn below. The values of the entries play no part in the graph.



Nonexample by minimal change. The graph records whether an entry is zero, not how large it is. Change the entry 3 of the last matrix to 10^{-9}: the graph is unchanged. Change it to 0: the edge $1 \to 3$ disappears, and with it every walk out of vertex 1 other than the one of length 0.

The graph was built to count routes, and it does, for nonnegative matrices. The same lemma handles $\mathbf{I} + \mathbf{A}$, whose graph is that of $\mathbf{A}$ with a loop added at every vertex; a loop lets a walk wait, so walks of length **at most** k become walks of length exactly k.

Lemma 19.2.1. Powers Count Walks

Let $\mathbf{A} \in M_n(\mathbb{R})$ be nonnegative, and let $k \geq 0$ and $i, j \in \{1, \ldots, n\}$.

(a) $(\mathbf{A}^k)_{ij} > 0$ if and only if $G(\mathbf{A})$ has a walk of length k from j to i.

(b) $\left((\mathbf{I} + \mathbf{A})^k\right)_{ij} > 0$ if and only if $G(\mathbf{A})$ has a walk of length **at most** k from j to i.

Idea. Induction on k, peeling off the first step of a walk. Expanding $(\mathbf{A}^{k+1})_{ij} = \sum_l (\mathbf{A}^k)_{il} a_{lj}$ gives a sum of nonnegative terms, so it is positive exactly when one term is: a first step $j \to l$ followed by a walk of length k from l to i. For $\mathbf{I} + \mathbf{A}$ the diagonal 1 adds the option of waiting at j instead of stepping.

Proof.

(a) Induction on k. For $k = 0$, $(\mathbf{I})_{ij} > 0$ exactly when $i = j$, which is exactly when there is a walk of length 0 from j to i. Suppose (a) holds for k. By Lemma 19.1.1 (c), $\mathbf{A}^k \geq \mathbf{O}$, so

$$(\mathbf{A}^{k+1})_{ij} = \sum_{l=1}^{n} (\mathbf{A}^k)_{il}\, a_{lj}$$

is a sum of nonnegative terms. It is positive if and only if some term is, that is, if and only if there is an l with $a_{lj} > 0$ and $(\mathbf{A}^k)_{il} > 0$. By the definition of $G(\mathbf{A})$ and the induction hypothesis, this says: there is an edge $j \to l$ and a walk of length k from l to i. Putting j in front of that walk gives a walk of length $k+1$ from j to i, and every such walk arises this way, with l its second vertex.

(b) Induction on k, with the same base case, since $(\mathbf{I} + \mathbf{A})^0 = \mathbf{I}$. Suppose (b) holds for k.

As $(\mathbf{I}+\mathbf{A})^k \geq \mathbf{O}$ and $\mathbf{I}+\mathbf{A} \geq \mathbf{O}$,

$$\left((\mathbf{I}+\mathbf{A})^{k+1}\right)_{ij} = \sum_{l=1}^{n} \left((\mathbf{I}+\mathbf{A})^k\right)_{il} (\delta_{lj} + a_{lj})$$

is positive if and only if there is an l with $((\mathbf{I}+\mathbf{A})^k)_{il} > 0$ and either $l = j$ or $a_{lj} > 0$. By the induction hypothesis this says: there is a walk of length at most k from l to i, where either $l = j$ or $j \to l$ is an edge. In the first case we have a walk of length at most k from j to i; in the second, putting j in front gives one of length at most $k+1$. Conversely, let a walk of length $m \leq k+1$ from j to i be given. If $m \leq k$, take $l = j$. If $m = k+1 \geq 1$, take l to be its second vertex, so that $j \to l$ is an edge and the rest of the walk has length k. This proves the lemma.

■

For the matrix drawn above, $\mathbf{A}^2 = \left(\begin{smallmatrix} 0 & 0 & 2 \\ 3 & 0 & 1 \\ 3 & 6 & 1 \end{smallmatrix}\right)$. Its $(1,1)$ entry is 0, and indeed the shortest walk from 1 back to 1 is $1 \to 3 \to 2 \to 1$, of length 3. Its $(1,3)$ entry is 2, from the walk $3 \to 2 \to 1$.

A long walk can always be shortened to a path, and a path is short.

Lemma 19.2.2. Shortening a Walk to a Path

Let $\mathbf{A} \in M_n(F)$. If $G(\mathbf{A})$ has a walk from j to i, then it has a path from j to i, and every path has length at most $n-1$.

Idea. Take a shortest walk. If it visited some vertex twice, the loop between the two visits could be cut out, leaving a shorter walk; so a shortest walk is a path. A path visits distinct vertices, and there are only n of them.

Proof. Among the walks from j to i, choose one of least length k (Theorem 0.2.11), say $j = i_0, i_1, \ldots, i_k = i$. Suppose two of its vertices coincide, $i_s = i_t$ with $s < t$. Deleting $i_{s+1}, \ldots, i_t$ leaves the list $i_0, \ldots, i_s, i_{t+1}, \ldots, i_k$, which is still a walk, because the edge $i_t \to i_{t+1}$ (if $t < k$) is the edge $i_s \to i_{t+1}$. It has length $k - (t-s) < k$, contradicting the choice of k. So the vertices are distinct, and the walk is a path. A path of length k has $k+1$ distinct vertices among the n vertices of $G(\mathbf{A})$, so $k+1 \leq n$.

■

19.2.2 Strong connectivity

Positivity of $\mathbf{A}$ says every vertex has an edge to every vertex. The useful weakening is to ask only for a **walk**.

A graph is strongly connected when every vertex can be reached from every vertex.

Definition 19.2.2. Strongly Connected

Let $\mathbf{A} \in M_n(F)$. The graph $G(\mathbf{A})$ is **strongly connected** if, for **every** ordered pair (j, i) of vertices, there is a walk from j to i.

In words: whichever vertex we start from, every vertex can be reached, and the pair is **ordered**, so a walk from j to i and a walk from i back to j are both required. By Lemma 19.2.2, the walks may be taken to be paths of length at most $n-1$. For $n = 1$ the walk of length 0 suffices, so the graph of every 1×1 matrix, $\mathbf{O}_1$ included, is strongly connected.

The cycle $\mathbf{C}$ above is strongly connected: from any vertex, follow the arrows round. So is the graph drawn above, since it contains the cycle $1 \to 3 \to 2 \to 1$, and so is the graph of every positive matrix. The path is not: there is no walk from 3 to 1. Removing one edge from the cycle, by changing $c_{13} = 1$ to 0, gives exactly the path, so the clause that fails is the one for the pair $(3, 1)$; the walks from 1 to 2, from 1 to 3, and from 2 to 3 survive.

19.2.3 Irreducible matrices

Now the notion this chapter is built on. We want to single out the matrices whose eigenvalue problem cannot be split into two smaller ones. For a matrix with a zero block in the lower left, it can:

$$\mathbf{A} = \begin{pmatrix} \mathbf{B} & \mathbf{C} \\ \mathbf{O} & \mathbf{D} \end{pmatrix} \quad \Longrightarrow \quad p_{\mathbf{A}} = p_{\mathbf{B}}\, p_{\mathbf{D}},$$

by Theorem 7.8.4 applied to $x\mathbf{I} - \mathbf{A}$, and the subspace spanned by the first few standard basis vectors is mapped into itself. For such a matrix Perron's conclusions fail in general: $\mathbf{I}_2$ and $\operatorname{diag}(2, 1)$ are of this shape. In this chapter a change of basis must preserve nonnegativity, and the ones that do so while keeping the entries themselves are the relabelings of the coordinates. So the definition asks whether a relabeling produces the zero block.

A matrix is irreducible if no relabeling of the coordinates puts it in block upper triangular form.

Definition 19.2.3. Irreducible Matrix

Let $\mathbf{A} \in M_n(F)$.

(a) For $n \geq 2$, $\mathbf{A}$ is **reducible** if there are a permutation matrix $\mathbf{P}$ and an integer k with $1 \leq k \leq n-1$ such that

$$\mathbf{P}^{\top}\mathbf{A}\mathbf{P} = \begin{pmatrix} \mathbf{B} & \mathbf{C} \\ \mathbf{O} & \mathbf{D} \end{pmatrix}, \qquad \mathbf{B} \in M_k(F),\ \mathbf{D} \in M_{n-k}(F),$$

with $\mathbf{O}$ the $(n-k) \times k$ zero matrix. It is **irreducible** if it is **not** reducible.

(b) For $n = 1$, **every** matrix $(a) \in M_1(F)$ is irreducible, including $\mathbf{O}_1$.

In words: by Lemma 3.4.7 (b), $\mathbf{P}^{\top} = \mathbf{P}^{-1}$, so $\mathbf{P}^{\top}\mathbf{A}\mathbf{P}$ is similar to $\mathbf{A}$. It is the matrix of $T_{\mathbf{A}}$ in the standard basis taken in another order, and in (a) we ask whether some order has a zero block below the diagonal with **both** diagonal blocks nonempty, which is what $1 \leq k \leq n-1$ says. Irreducible means that **no** order does. Clause (b) is a convention, justified below.

To see what relabeling does to the entries, we compute it once.

Lemma 19.2.3. Relabeling the Coordinates

Let $\sigma \in S_n$ and $\mathbf{A} \in M_n(F)$. Then $(\mathbf{P}_{\sigma}^{\top}\mathbf{A}\mathbf{P}_{\sigma})_{st} = a_{\sigma(s)\sigma(t)}$ for all s, t. Consequently $t \to s$ is

an edge of $G(\mathbf{P}_\sigma^\top \mathbf{A}\mathbf{P}_\sigma)$ if and only if $\sigma(t) \to \sigma(s)$ is an edge of $G(\mathbf{A})$.

Proof. By Definition 3.4.2, $\mathbf{P}_\sigma \mathbf{e}_t = \mathbf{e}_{\sigma(t)}$. The (s,t) entry of a matrix $\mathbf{M}$ is $\mathbf{e}_s^\top \mathbf{M} \mathbf{e}_t$, so by Theorem 0.9.5,

$$(\mathbf{P}_\sigma^\top \mathbf{A}\mathbf{P}_\sigma)_{st} = \mathbf{e}_s^\top \mathbf{P}_\sigma^\top \mathbf{A}\mathbf{P}_\sigma \mathbf{e}_t = (\mathbf{P}_\sigma \mathbf{e}_s)^\top \mathbf{A}(\mathbf{P}_\sigma \mathbf{e}_t) = \mathbf{e}_{\sigma(s)}^\top \mathbf{A}\, \mathbf{e}_{\sigma(t)} = a_{\sigma(s)\sigma(t)}.$$

The statement about edges is the statement that one of these entries is nonzero exactly when the other is.

■

So the graph of $\mathbf{P}_\sigma^\top \mathbf{A}\mathbf{P}_\sigma$ is $G(\mathbf{A})$ with the vertex $\sigma(s)$ renamed s. In particular, Definition 19.2.2 holds for one exactly when it holds for the other. The zero block of Definition 19.2.3 becomes a statement about a set of vertices that no edge leaves.

Lemma 19.2.4. Reducible Means a Set of Vertices With No Exit

Let $n \geq 2$ and $\mathbf{A} \in M_n(F)$. Then $\mathbf{A}$ is reducible if and only if there is a set $S \subseteq \{1, \ldots, n\}$ with $S \neq \emptyset$ and $S \neq \{1, \ldots, n\}$ such that

$$a_{ij} = 0 \qquad \text{whenever } j \in S \text{ and } i \notin S,$$

that is, such that no edge of $G(\mathbf{A})$ goes from a vertex in S to a vertex outside S.

Idea. The zero block of Definition 19.2.3 sits in the columns of the first k relabeled vertices and the rows of the others. So take S to be the first k vertices in the new order; conversely, given S, choose the order that lists S first. Lemma 19.2.3 translates between the entries and the vertices.

Proof. ($\Leftarrow$) Let k be the number of elements of S, so $1 \leq k \leq n-1$, and choose $\sigma \in S_n$ listing the elements of S first: $\sigma(1), \ldots, \sigma(k)$ are the elements of S and $\sigma(k+1), \ldots, \sigma(n)$ the others. For $t \leq k < s$ we have $\sigma(t) \in S$ and $\sigma(s) \notin S$, so by Lemma 19.2.3, $(\mathbf{P}_\sigma^\top \mathbf{A}\mathbf{P}_\sigma)_{st} = a_{\sigma(s)\sigma(t)} = 0$. That is, the lower-left $(n-k) \times k$ block of $\mathbf{P}_\sigma^\top \mathbf{A}\mathbf{P}_\sigma$ is zero, and $\mathbf{A}$ is reducible.

($\Rightarrow$) Let $\mathbf{P}$ and k be as in Definition 19.2.3 (a). By Definition 3.4.2, $\mathbf{P} = \mathbf{P}_\sigma$ for some $\sigma \in S_n$. Put $S = \{\sigma(1), \ldots, \sigma(k)\}$, which has k elements, so it is neither empty nor everything. If $j \in S$ and $i \notin S$, then $j = \sigma(t)$ with $t \leq k$ and $i = \sigma(s)$ with $s > k$, and $a_{ij} = (\mathbf{P}_\sigma^\top \mathbf{A}\mathbf{P}_\sigma)_{st} = 0$ because (s,t) lies in the zero block.

■

The condition says that the coordinate subspace $\operatorname{span}\{\mathbf{e}_j : j \in S\}$ is mapped into itself by $\mathbf{A}$, since column j of $\mathbf{A}$ is $\mathbf{A}\mathbf{e}_j$ and its entries outside S vanish. So a reducible matrix is one with an invariant subspace spanned by some of the standard basis vectors, other than $\{\mathbf{0}\}$ and F^n, and Theorem 9.1.3 is the source of the block form.

Now some examples, simplest first.

- $n = 1$**.** Irreducible by (b).
- **Positive matrices** are irreducible: no entry is 0, so no set S as in Lemma 19.2.4 exists.
- **The swap** $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ is irreducible. For $n = 2$, the candidates are $S = \{1\}$, which needs $a_{21} = 0$, and $S = \{2\}$, which needs $a_{12} = 0$; both entries are 1. So a matrix may have zeros, even mostly zeros, and be irreducible.
- **Triangular matrices** with $n \geq 2$ are reducible. An upper triangular one already has the block form with $\mathbf{P} = \mathbf{I}$ and $k = 1$. A lower triangular one, such as the path matrix, is reducible too, by Lemma 19.2.4 with $S = \{n\}$: column n has its only possible nonzero entry on the diagonal. Reversing the order of the coordinates, $\sigma(s) = n + 1 - s$, turns the path matrix into $\mathbf{J}_3(0)$, which is upper triangular.
- **The zero matrix** of size $n \geq 2$ is reducible, and so is $\mathbf{I}_n$: any S will do.

Nonexample by minimal change. The cycle $\mathbf{C}$ is irreducible, as the next theorem will show. Change its entry c_{13} from 1 to 0: the result is the path matrix, and $S = \{3\}$ now has no exit, because the only nonzero entry of column 3 outside row 3, the entry c_{13}, has become 0. That single clause of Lemma 19.2.4 is what changed.

Example. A reducible matrix that does not look it

Decide whether $\mathbf{A} = \begin{pmatrix} 1 & 0 & 2 \\ 3 & 4 & 5 \\ 6 & 0 & 7 \end{pmatrix}$ is irreducible, and if not, put it in block triangular form.

Solution. Column 2 has its only nonzero entry in row 2, so $S = \{2\}$ has no exit: $a_{12} = a_{32} = 0$. By Lemma 19.2.4, $\mathbf{A}$ is reducible. Following the proof, list 2 first: $\sigma(1) = 2$, $\sigma(2) = 1$, $\sigma(3) = 3$. By Lemma 19.2.3, the (s, t) entry of $\mathbf{P}_\sigma^\top \mathbf{A} \mathbf{P}_\sigma$ is $a_{\sigma(s)\sigma(t)}$, so

$$\mathbf{P}_\sigma^\top \mathbf{A} \mathbf{P}_\sigma = \begin{pmatrix} a_{22} & a_{21} & a_{23} \\ a_{12} & a_{11} & a_{13} \\ a_{32} & a_{31} & a_{33} \end{pmatrix} = \begin{pmatrix} 4 & 3 & 5 \\ 0 & 1 & 2 \\ 0 & 6 & 7 \end{pmatrix}.$$

Hence $p_{\mathbf{A}}(x) = (x - 4)(x^2 - 8x - 5)$, and the eigenvalues are 4 and $4 \pm \sqrt{21}$. In the graph, vertex 2 has only a loop leaving it: a chain that reaches state 2 stays there.

Why this definition, and why every 1×1 matrix is irreducible. The block form is the matrix picture of an invariant coordinate subspace, and it is exactly what splits $p_{\mathbf{A}}$ into two factors and lets the conclusions of Perron's theorem fail, as $\mathbf{I}_2$ shows. For $n = 1$ there is no room for two nonempty blocks, and a convention is needed. Some books call $\mathbf{O}_1$ reducible. We call every 1×1 matrix irreducible because that is the choice that makes the two theorems below true with no exception: the graph of a 1×1 matrix is always strongly connected, through the walk of length 0, and $(\mathbf{I} + \mathbf{A})^{1-1} = \mathbf{I} = (1)$ is always positive. The price is paid in §03: the irreducible matrix $\mathbf{O}_1$ has spectral radius 0, so the statement there that an irreducible nonnegative matrix has $\rho(\mathbf{A}) > 0$ carries the hypothesis $n \geq 2$.

Warning. **Irreducibility depends only on which entries are zero, never on their size.** For every $t \neq 0$, however small, $\mathbf{A}_t = \begin{pmatrix} 1 & t \\ 1 & 1 \end{pmatrix}$ is irreducible, and for $t > 0$ its eigenvalues

$1 \pm \sqrt{t}$ are distinct, with the positive eigenvector $(\sqrt{t}, 1)$ for $1 + \sqrt{t}$. At $t = 0$ it is lower triangular, hence reducible, and 1 becomes a double eigenvalue with the eigenvector $\mathbf{e}_2$ only. An entry of 10^{-9} and an entry of 10^{9} count the same; an entry of 0 is different in kind. The property is not preserved by powers either: the swap is irreducible and its square is $\mathbf{I}_2$, which is not.

The first theorem says that irreducibility, defined through relabelings, is exactly strong connectivity of the graph.

Theorem 19.2.5. Irreducible If and Only If Strongly Connected

Let $n \geq 1$ and $\mathbf{A} \in M_n(F)$. Then $\mathbf{A}$ is irreducible if and only if $G(\mathbf{A})$ is strongly connected.

Idea. Both sides are about a set of vertices with no exit. If some vertex i cannot be reached from j, the set of vertices that **can** be reached from j contains j, misses i, and has no exit, since an exit would reach one more vertex. Conversely, a set with no exit traps every walk that starts inside it.

Proof. If $n = 1$, both sides hold, by Definition 19.2.3 (b) and the walk of length 0. Let $n \geq 2$. By Lemma 19.2.4 it suffices to show: $G(\mathbf{A})$ fails to be strongly connected if and only if there is a set S, neither empty nor all of $\{1, \ldots, n\}$, from which no edge leaves.

($\Rightarrow$) Suppose there are j and i with no walk from j to i. Let S be the set of vertices l such that there is a walk from j to l. Then $j \in S$, by the walk of length 0, and $i \notin S$. If $l \in S$ and $l \to m$ is an edge, then appending m to a walk from j to l gives a walk from j to m, so $m \in S$. Hence no edge leaves S.

($\Leftarrow$) Suppose no edge leaves S, and choose $j \in S$ and $i \notin S$. Let $j = i_0, i_1, \ldots, i_k$ be any walk from j. Then every i_t lies in S, by induction on t: $i_0 = j \in S$, and if $i_{t-1} \in S$, the edge $i_{t-1} \to i_t$ does not leave S, so $i_t \in S$. Hence no walk from j ends at i, and $G(\mathbf{A})$ is not strongly connected. This proves the theorem. ∎

So the cycle $\mathbf{C}$ is irreducible, and so is the matrix drawn earlier: their graphs are strongly connected. Deciding irreducibility from the definition means looking at $n!$ relabelings; deciding it from the graph means following arrows.

Quick check. Is $\mathbf{M} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}$ irreducible? Every row and every column of $\mathbf{M}$ contains a nonzero entry off the diagonal.

Solution. No. The edges of $G(\mathbf{M})$ are $1 \to 2$, $2 \to 1$, $3 \to 4$ and $4 \to 3$, so there is no walk from 1 to 3, and $\mathbf{M}$ is reducible by Theorem 19.2.5. Directly, $S = \{1, 2\}$ has no exit, and indeed $\mathbf{M} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \oplus \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ is already block diagonal, with $\mathbf{P} = \mathbf{I}$ and $k = 2$. Having

an off-diagonal entry in every row and column is necessary for irreducibility when $n \geq 2$ (Exercise C1), but it is not sufficient.

19.2.4 A test by one matrix power

For a nonnegative matrix, Lemma 19.2.1 turns strong connectivity into positivity of a single matrix, and that is the bridge back to Perron's theorem.

Theorem 19.2.6. Irreducibility Through a Power of $\mathbf{I}+\mathbf{A}$

Let $n \geq 1$ and let $\mathbf{A} \in M_n(\mathbb{R})$ be nonnegative. Then $\mathbf{A}$ is irreducible if and only if

$$(\mathbf{I}+\mathbf{A})^{n-1} > \mathbf{O}.$$

Idea. Read both sides as statements about walks. Positivity of $(\mathbf{I}+\mathbf{A})^{n-1}$ asks for a walk of length **at most** $n-1$ between every ordered pair of vertices, and strong connectivity asks for a walk of any length. The two agree because a walk can always be shortened to a path, and a path has at most $n-1$ edges.

Proof. By Lemma 19.2.1 (b), $(\mathbf{I}+\mathbf{A})^{n-1} > \mathbf{O}$ if and only if, for every ordered pair (j,i), there is a walk of length at most $n-1$ from j to i. By Lemma 19.2.2, if there is any walk from j to i then there is one of length at most $n-1$, namely a path; so the condition is equivalent to asking for a walk of any length from j to i, for every ordered pair. That is Definition 19.2.2, and by Theorem 19.2.5 it is equivalent to irreducibility. ∎

Both sides always hold for $n = 1$, as the convention intends. The hypothesis $\mathbf{A} \geq \mathbf{O}$ cannot be dropped: $\begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix}$ is irreducible, but adding $\mathbf{I}$ cancels its diagonal and leaves the swap, which is not positive. Cancellation is exactly what nonnegativity rules out, since a sum of nonnegative terms is zero only when every term is.

Example. The cycle, by one power

For the cycle $\mathbf{C} = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$, compute $(\mathbf{I}+\mathbf{C})^2$, and compare with the powers $\mathbf{C}^k$.

Solution. $\mathbf{C}$ is the permutation matrix of $1 \mapsto 2 \mapsto 3 \mapsto 1$, so $\mathbf{C}^2 = \left(\begin{smallmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{smallmatrix}\right)$ and $\mathbf{C}^3 = \mathbf{I}$ by Lemma 3.4.7 (a). As $\mathbf{I}$ and $\mathbf{C}$ commute,

$$(\mathbf{I}+\mathbf{C})^2 = \mathbf{I} + 2\mathbf{C} + \mathbf{C}^2 = \begin{pmatrix} 1 & 1 & 2 \\ 2 & 1 & 1 \\ 1 & 2 & 1 \end{pmatrix} > \mathbf{O},$$

so $\mathbf{C}$ is irreducible by Theorem 19.2.6, as the graph already showed. On the other hand every power $\mathbf{C}^k$ is $\mathbf{I}$, $\mathbf{C}$ or $\mathbf{C}^2$, each with six zero entries, so **no** power of $\mathbf{C}$ is positive. In the graph, the walks from j have lengths forced modulo 3 by where they end. The loops of $\mathbf{I}+\mathbf{C}$ let a walk wait, which is what $\mathbf{C}$ alone cannot do. §05 returns to this difference; it is the difference between irreducible and primitive.

The theorem is the reason irreducibility is enough for most of Perron's conclusions. If $\mathbf{A} \geq \mathbf{O}$ is irreducible, then $\mathbf{B} = (\mathbf{I}+\mathbf{A})^{n-1}$ is positive, so Theorem 19.1.5 applies to $\mathbf{B}$, and $\mathbf{B}$ is a polynomial in $\mathbf{A}$, so every eigenvector of $\mathbf{A}$ is an eigenvector of $\mathbf{B}$. §03 turns this observation into the Perron–Frobenius theorem.

19.2.5 The reducible normal form

A reducible matrix splits into two diagonal blocks, and each block may be reducible again. Splitting until nothing splits gives a normal form.

Theorem 19.2.7. Reducible Normal Form

Let $n \geq 1$ and $\mathbf{A} \in M_n(F)$.

(a) There are a permutation matrix $\mathbf{P}$ and an integer $m \geq 1$ such that

$$\mathbf{P}^{\top}\mathbf{A}\mathbf{P} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} & \cdots & \mathbf{A}_{1m} \\ \mathbf{O} & \mathbf{A}_{22} & \cdots & \mathbf{A}_{2m} \\ \vdots & & \ddots & \vdots \\ \mathbf{O} & \cdots & \mathbf{O} & \mathbf{A}_{mm} \end{pmatrix},$$

block upper triangular, with every diagonal block $\mathbf{A}_{ii}$ square and **irreducible**.

(b) In that case $p_{\mathbf{A}} = p_{\mathbf{A}_{11}}\, p_{\mathbf{A}_{22}} \cdots p_{\mathbf{A}_{mm}}$, so the eigenvalues of $\mathbf{A}$, with their algebraic multiplicities, are those of the diagonal blocks taken together.

(c) $\rho(\mathbf{A}) = \max_i \rho(\mathbf{A}_{ii})$; if $\mathbf{A} \geq \mathbf{O}$, every $\mathbf{A}_{ii} \geq \mathbf{O}$.

Idea. Strong induction on n: if $\mathbf{A}$ is irreducible there is nothing to do, and otherwise one relabeling gives two smaller diagonal blocks, which the induction hypothesis puts in normal form. The only work is to check that the two relabelings of the blocks combine into one relabeling of $\mathbf{A}$, and block multiplication does that.

Proof.

(a) Strong induction on n (Theorem 0.2.10). If $\mathbf{A}$ is irreducible, which includes the case $n = 1$, take $\mathbf{P} = \mathbf{I}$ and $m = 1$. Otherwise $n \geq 2$, and there are a permutation matrix $\mathbf{P}_0$ and $1 \leq k \leq n-1$ with

$$\mathbf{P}_0^{\top}\mathbf{A}\mathbf{P}_0 = \begin{pmatrix} \mathbf{B} & \mathbf{C} \\ \mathbf{O} & \mathbf{D} \end{pmatrix}, \qquad \mathbf{B} \in M_k(F),\ \mathbf{D} \in M_{n-k}(F).$$

Both sizes are less than n, so by the induction hypothesis there are permutation matrices $\mathbf{Q}_1 = \mathbf{P}_\alpha$ ($\alpha \in S_k$) and $\mathbf{Q}_2 = \mathbf{P}_\beta$ ($\beta \in S_{n-k}$) such that $\mathbf{Q}_1^{\top}\mathbf{B}\mathbf{Q}_1$ and $\mathbf{Q}_2^{\top}\mathbf{D}\mathbf{Q}_2$

are block upper triangular with square irreducible diagonal blocks. The direct sum $\mathbf{Q} = \mathbf{Q}_1 \oplus \mathbf{Q}_2$ (Definition 8.1.2) is the permutation matrix $\mathbf{P}_\gamma$ of the permutation γ with $\gamma(t) = \alpha(t)$ for $t \le k$ and $\gamma(k+q) = k + \beta(q)$ for $q \le n-k$, since its columns are $\mathbf{e}_{\gamma(1)}, \dots, \mathbf{e}_{\gamma(n)}$. By Proposition 8.1.2, $\mathbf{Q}^\top = \mathbf{Q}_1^\top \oplus \mathbf{Q}_2^\top$, and by Theorem 8.1.1,

$$\mathbf{Q}^\top(\mathbf{P}_0^\top \mathbf{A}\mathbf{P}_0)\mathbf{Q} = \begin{pmatrix} \mathbf{Q}_1^\top \mathbf{B}\mathbf{Q}_1 & \mathbf{Q}_1^\top \mathbf{C}\mathbf{Q}_2 \\ \mathbf{O} & \mathbf{Q}_2^\top \mathbf{D}\mathbf{Q}_2 \end{pmatrix}.$$

This is block upper triangular, and its diagonal blocks are those of $\mathbf{Q}_1^\top \mathbf{B}\mathbf{Q}_1$ followed by those of $\mathbf{Q}_2^\top \mathbf{D}\mathbf{Q}_2$, all square and irreducible. With $\mathbf{P} = \mathbf{P}_0\mathbf{Q}$, which is a permutation matrix by Lemma 3.4.7 (a), the left side is $\mathbf{P}^\top \mathbf{A}\mathbf{P}$ by Theorem 0.9.5. This proves (a).

(b) $\mathbf{P}^\top = \mathbf{P}^{-1}$ by Lemma 3.4.7 (b), so $\mathbf{P}^\top \mathbf{A}\mathbf{P}$ is similar to $\mathbf{A}$, and $p_{\mathbf{A}} = p_{\mathbf{P}^\top \mathbf{A}\mathbf{P}}$ by Theorem 7.10.3. The matrix $x\mathbf{I} - \mathbf{P}^\top \mathbf{A}\mathbf{P}$ is block upper triangular with diagonal blocks $x\mathbf{I} - \mathbf{A}_{ii}$. Applying Theorem 7.8.4 over $F[x]$, as the paragraph after it notes, to the first diagonal block and the rest, and then to the rest by induction on m, gives $p_{\mathbf{A}} = \prod_i p_{\mathbf{A}_{ii}}$. The roots of a product, with multiplicities, are those of the factors together.

(c) By (b), $\operatorname{spec}(\mathbf{A}) = \bigcup_i \operatorname{spec}(\mathbf{A}_{ii})$, and the largest modulus over a union is the largest of the largest moduli. If $\mathbf{A} \ge \mathbf{O}$, then by Lemma 19.2.3 the entries of $\mathbf{P}^\top \mathbf{A}\mathbf{P}$ are entries of $\mathbf{A}$, so they are ≥ 0, and so are those of each $\mathbf{A}_{ii}$.

■

The diagonal blocks of the normal form correspond to groups of vertices of $G(\mathbf{A})$ that can all reach one another. The first block is a group with no exit, and each later block can send edges only to the blocks before it. In a Markov chain the first block is therefore a set of states that the chain, once inside, never leaves; §06 uses exactly this. The blocks are unique up to their order and up to relabeling within each block, but we shall not need this and do not prove it.

Example. A normal form with three blocks

Put $\mathbf{A} = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 1 & 2 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 3 \end{pmatrix}$ in reducible normal form, and find $\rho(\mathbf{A})$.

Solution. Read the edges column by column. Column 1 gives $1 \to 2$ and $1 \to 3$; column 2 gives the loop $2 \to 2$; column 3 gives $3 \to 1$ and $3 \to 4$; column 4 gives $4 \to 2$ and the loop $4 \to 4$. So 1 and 3 reach each other, and from them one can go to 4 and then to 2, but never back. The groups are $\{2\}$, $\{4\}$ and $\{1,3\}$. Vertex 2 has no exit, so it comes first; the edges from 4 go only to 2 and 4, so 4 comes next; then $1, 3$. With $\sigma = (2,4,1,3)$, that is $\sigma(1) = 2$, $\sigma(2) = 4$, $\sigma(3) = 1$, $\sigma(4) = 3$, Lemma 19.2.3 gives

$$\mathbf{P}_\sigma^\top \mathbf{A}\mathbf{P}_\sigma = \begin{pmatrix} 2 & 1 & 1 & 0 \\ 0 & 3 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix},$$

with diagonal blocks (2), (3) and the swap $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$, each irreducible. By Theorem 19.2.7 (b), $p_{\mathbf{A}}(x) = (x-2)(x-3)(x^2-1)$, so the eigenvalues are $2, 3, 1, -1$, and $\rho(\mathbf{A}) = 3$, the spectral radius of the second block.

19.2.6 Exercises

19.2.6.1 A. Check your understanding

Exercise. A1

(a) Define the directed graph $G(\mathbf{A})$, a walk, and strong connectivity. Which way does the edge for $a_{ij} \neq 0$ point?

(b) Define **irreducible**, including the case $n = 1$, and state Theorem 19.2.6.

(c) True or false: a matrix with a zero entry is reducible. Justify your answer.

(d) True or false: if $\mathbf{A} \geq \mathbf{O}$ is irreducible, then so is $\mathbf{A}^2$. Justify your answer.

(e) True or false: $\mathbf{A}$ is irreducible if and only if $\mathbf{A}^\top$ is. Justify your answer.

(f) True or false: the 2×2 zero matrix is irreducible. Justify your answer.

Solution.

(a) $G(\mathbf{A})$ has vertices $1, \dots, n$ and an edge $j \to i$ for each $a_{ij} \neq 0$, pointing from the column index to the row index. A walk of length k from j to i is a list $j = i_0, \dots, i_k = i$ with each $i_{t-1} \to i_t$ an edge. $G(\mathbf{A})$ is strongly connected if there is a walk from j to i for every ordered pair (j, i).

(b) For $n \geq 2$, $\mathbf{A}$ is reducible if $\mathbf{P}^\top \mathbf{A} \mathbf{P} = \left(\begin{smallmatrix} \mathbf{B} & \mathbf{C} \\ \mathbf{O} & \mathbf{D} \end{smallmatrix}\right)$ for some permutation matrix $\mathbf{P}$ and square blocks $\mathbf{B}$, $\mathbf{D}$ of sizes k and $n-k$, $1 \leq k \leq n-1$; irreducible means not reducible. Every 1×1 matrix is irreducible. The theorem: a nonnegative $\mathbf{A} \in M_n(\mathbb{R})$ is irreducible if and only if $(\mathbf{I} + \mathbf{A})^{n-1} > \mathbf{O}$.

(c) False. The swap $\left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$ has two zero entries and is irreducible, since its graph $1 \to 2 \to 1$ is strongly connected (Theorem 19.2.5).

(d) False. The swap is irreducible and its square $\mathbf{I}_2$ is reducible, being triangular.

(e) True. $G(\mathbf{A}^\top)$ is $G(\mathbf{A})$ with every edge reversed, since $(\mathbf{A}^\top)_{ij} = a_{ji}$. Reversing a walk from j to i in $G(\mathbf{A})$ gives a walk from i to j in $G(\mathbf{A}^\top)$, so one graph is strongly connected exactly when the other is, and Theorem 19.2.5 finishes.

(f) False. With $n = 2$, the zero matrix is already upper triangular, so it is reducible with $\mathbf{P} = \mathbf{I}$, $k = 1$. Only the 1×1 zero matrix is irreducible, and that by convention.

19.2.6.2 B. Practice

Exercise. B1: Deciding irreducibility

Determine which of the following matrices are irreducible. Justify your answers, and for each reducible one give a permutation matrix that puts it in block upper triangular form.

(a) $\begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$

(b) $\begin{pmatrix} 1 & 2 & 0 \\ 0 & 3 & 0 \\ 4 & 5 & 6 \end{pmatrix}$

(c) $\begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \end{pmatrix}$

(d) $\begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix}$

Solution.

(a) The nonzero entries a_{12}, a_{23}, a_{31} give the edges $2 \to 1$, $3 \to 2$, $1 \to 3$: the cycle $1 \to 3 \to 2 \to 1$, which is strongly connected. Irreducible, by Theorem 19.2.5.

(b) Column 3 has its only nonzero entry in row 3, so $S = \{3\}$ has no exit, and the matrix is reducible by Lemma 19.2.4. With $\sigma(1) = 3$, $\sigma(2) = 1$, $\sigma(3) = 2$, Lemma 19.2.3 gives $\mathbf{P}_\sigma^\top \mathbf{A} \mathbf{P}_\sigma = \left(\begin{smallmatrix} 6 & 4 & 5 \\ 0 & 1 & 2 \\ 0 & 0 & 3 \end{smallmatrix}\right)$, which is even triangular.

(c) The edges are $1 \to 4$ (from a_{41}), $2 \to 1$ and $2 \to 4$ (from a_{12}, a_{42}), $3 \to 2$ (from a_{23}) and $4 \to 3$ (from a_{34}). The walk $1 \to 4 \to 3 \to 2 \to 1$ passes through every vertex and returns, so any vertex reaches any other by following it. Irreducible.

(d) Columns 3 and 4 have zeros in rows 1 and 2, so $S = \{3, 4\}$ has no exit: reducible. With $\sigma = (3, 4, 1, 2)$, $\mathbf{P}_\sigma^\top \mathbf{A} \mathbf{P}_\sigma = \left(\begin{smallmatrix} 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{smallmatrix}\right)$. The edge $2 \to 3$ goes into S, which is allowed; only edges out of S are forbidden.

Exercise. B2: One power decides

For each matrix, compute $(\mathbf{I} + \mathbf{A})^2$ and decide whether $\mathbf{A}$ is irreducible. For the reducible one, read off from $(\mathbf{I} + \mathbf{A})^2$ a set of vertices with no exit.

(a) $\mathbf{A} = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$

(b) $\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$

Solution.

(a) $\mathbf{I}+\mathbf{A} = \left(\begin{smallmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{smallmatrix}\right)$, and multiplying out,

$$(\mathbf{I}+\mathbf{A})^2 = \begin{pmatrix} 2 & 3 & 2 \\ 2 & 2 & 1 \\ 1 & 2 & 1 \end{pmatrix} > \mathbf{O}.$$

Hence $\mathbf{A}$ is irreducible by Theorem 19.2.6.

(b) $\mathbf{I}+\mathbf{A} = \left(\begin{smallmatrix} 2 & 0 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \end{smallmatrix}\right)$, and

$$(\mathbf{I}+\mathbf{A})^2 = \begin{pmatrix} 4 & 0 & 0 \\ 3 & 2 & 2 \\ 1 & 2 & 2 \end{pmatrix},$$

which has zero entries, so $\mathbf{A}$ is reducible. By Lemma 19.2.1 (b), column j of $(\mathbf{I}+\mathbf{A})^2$ is positive exactly in the rows i reachable from j. Column 2 is positive in rows $2, 3$ only, so the vertices reachable from 2 form $S = \{2, 3\}$, and by the proof of Theorem 19.2.5 no edge leaves it. Indeed $a_{12} = a_{13} = 0$.

Exercise. B3: Eigenvalues from the normal form

Let $\mathbf{A} = \begin{pmatrix} 1 & 0 & 2 \\ 1 & 3 & 0 \\ 4 & 0 & 5 \end{pmatrix}$. Put $\mathbf{A}$ in reducible normal form, and hence find its eigenvalues and $\rho(\mathbf{A})$.

Solution. The edges are $1 \to 1$, $1 \to 2$, $1 \to 3$ (column 1), the loop $2 \to 2$ (column 2), and $3 \to 1$, $3 \to 3$ (column 3). Vertex 2 has no exit, and 1, 3 reach each other. List 2 first: $\sigma(1) = 2$, $\sigma(2) = 1$, $\sigma(3) = 3$. By Lemma 19.2.3,

$$\mathbf{P}_\sigma^{\top}\mathbf{A}\mathbf{P}_\sigma = \begin{pmatrix} 3 & 1 & 0 \\ 0 & 1 & 2 \\ 0 & 4 & 5 \end{pmatrix}.$$

The blocks (3) and $\left(\begin{smallmatrix} 1 & 2 \\ 4 & 5 \end{smallmatrix}\right)$ are irreducible, the second because both its off-diagonal entries are nonzero. By Theorem 19.2.7 (b), $p_{\mathbf{A}}(x) = (x-3)(x^2-6x-3)$, so the eigenvalues are 3 and $3 \pm 2\sqrt{3}$, and by (c), $\rho(\mathbf{A}) = 3 + 2\sqrt{3} \approx 6.46$, coming from the second block.

19.2.6.3 C. Going deeper

Exercise. C1: A necessary condition that is not sufficient

(a) Prove that if $n \geq 2$ and $\mathbf{A} \in M_n(F)$ is irreducible, then every row and every column of $\mathbf{A}$ contains a nonzero entry off the diagonal.

(b) Show that the converse fails for every even $n \geq 4$.

Solution.

(a) Suppose column j has no nonzero entry off the diagonal: $a_{ij} = 0$ for all $i \neq j$. Then $S = \{j\}$ has no exit, and S is neither empty nor everything because $n \geq 2$, so $\mathbf{A}$ is reducible by Lemma 19.2.4. Suppose instead row i has $a_{ij} = 0$ for all $j \neq i$. Then $S = \{1, \dots, n\} \setminus \{i\}$ has no exit, since the only entries $a_{i'j}$ with $j \in S$ and $i' \notin S$ are the a_{ij} with $j \neq i$; again $\mathbf{A}$ is reducible.

(b) Let $n = 2m$ with $m \geq 2$, and let $\mathbf{M}$ be the direct sum of m copies of the swap $\left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$. Every row and column has a 1 off the diagonal. But $S = \{1, 2\}$ has no exit, since column 1 and column 2 are zero outside rows $1, 2$, so $\mathbf{M}$ is reducible. The Quick check is the case $m = 2$.

Exercise. C2: The exponent $n - 1$ is sharp

Let $n \geq 2$ and let $\mathbf{C}_n$ be the cyclic permutation matrix with $\mathbf{C}_n \mathbf{e}_j = \mathbf{e}_{j+1}$ for $j < n$ and $\mathbf{C}_n \mathbf{e}_n = \mathbf{e}_1$.

(a) Prove that $\mathbf{C}_n$ is irreducible, but that $(\mathbf{I} + \mathbf{C}_n)^{n-2}$ is not positive.

(b) Prove that if $\mathbf{A} \geq \mathbf{O}$ is irreducible, then $(\mathbf{I} + \mathbf{A})^k > \mathbf{O}$ for every $k \geq n - 1$.

Solution.

(a) The nonzero entries of $\mathbf{C}_n$ are $(\mathbf{C}_n)_{j+1,j} = 1$ for $j < n$ and $(\mathbf{C}_n)_{1n} = 1$, so the edges of $G(\mathbf{C}_n)$ are $j \to j + 1$ and $n \to 1$, and there are no others. Following them from any vertex visits every vertex, so $G(\mathbf{C}_n)$ is strongly connected and $\mathbf{C}_n$ is irreducible (Theorem 19.2.5). Each vertex has exactly one outgoing edge, so the only walk from 1 of length $m \leq n - 1$ is $1 \to 2 \to \cdots \to m + 1$. In particular the shortest walk from 1 to n has length $n - 1$, and there is no walk of length at most $n - 2$ from 1 to n. By Lemma 19.2.1 (b), $\left((\mathbf{I} + \mathbf{C}_n)^{n-2}\right)_{n1} = 0$.

(b) By Lemma 19.2.1 (b), $((\mathbf{I} + \mathbf{A})^k)_{ij} > 0$ exactly when there is a walk of length at most k from j to i. For $k \geq n - 1$, a walk of length at most $n - 1$ is one of length at most k, and such walks exist for every pair by the proof of Theorem 19.2.6. Hence $(\mathbf{I} + \mathbf{A})^k > \mathbf{O}$.

Exercise. C3: Irreducibility as growth of supports

Let $n \geq 2$ and $\mathbf{A} \in M_n(\mathbb{R})$ with $\mathbf{A} \geq \mathbf{O}$. For $\mathbf{x} \geq \mathbf{0}$, call $\{i : x_i > 0\}$ the **support** of $\mathbf{x}$.

(a) Prove that the support of $(\mathbf{I} + \mathbf{A})\mathbf{x}$ is the support S of $\mathbf{x}$ together with every vertex i for which there is an edge $j \to i$ with $j \in S$.

(b) Prove that $\mathbf{A}$ is irreducible if and only if, for every $\mathbf{x} \geq \mathbf{0}$ with at least one zero entry and at least one positive entry, $(\mathbf{I} + \mathbf{A})\mathbf{x}$ has more positive entries than $\mathbf{x}$.

(c) Deduce, without using walks, that $(\mathbf{I} + \mathbf{A})^{n-1}\mathbf{x} > \mathbf{0}$ for every nonzero $\mathbf{x} \geq \mathbf{0}$ when $\mathbf{A}$ is irreducible, and hence that $(\mathbf{I} + \mathbf{A})^{n-1} > \mathbf{O}$.

Hint for (b): use Lemma 19.2.4.

Solution.

(a) $((\mathbf{I} + \mathbf{A})\mathbf{x})_i = x_i + \sum_j a_{ij} x_j$ is a sum of nonnegative terms, so it is positive if and only

if $x_i > 0$ or some $a_{ij}x_j > 0$, that is, $i \in S$ or there is $j \in S$ with $a_{ij} > 0$, which is an edge $j \to i$.

(b) Let $\mathbf{x}$ be as stated, with support S, which is neither empty nor everything. By (a), $(\mathbf{I}+\mathbf{A})\mathbf{x}$ has support containing S, and strictly larger exactly when some edge goes from S to a vertex outside S. ($\Rightarrow$) If $\mathbf{A}$ is irreducible, Lemma 19.2.4 says every such S has an exit, so the support grows. ($\Leftarrow$) If $\mathbf{A}$ were reducible, take S with no exit from Lemma 19.2.4, and let $\mathbf{x}$ have entry 1 on S and 0 elsewhere. Then $(\mathbf{I}+\mathbf{A})\mathbf{x}$ has support exactly S, with the same number of positive entries, contradicting the hypothesis.

(c) Let $\mathbf{x} \geq \mathbf{0}$, $\mathbf{x} \neq \mathbf{0}$, and put $\mathbf{x}_k = (\mathbf{I}+\mathbf{A})^k\mathbf{x} \geq \mathbf{0}$. By (a) the supports never shrink, and by (b) the support grows by at least one element at each step until it is everything. It starts with at least one element, so after at most $n-1$ steps it has n, and then stays so: $\mathbf{x}_{n-1} > \mathbf{0}$. Applying this to $\mathbf{x} = \mathbf{e}_j$ shows that column j of $(\mathbf{I}+\mathbf{A})^{n-1}$ is positive, for every j.

19.3 The Perron–Frobenius Theorem

Section 1 proved Perron's theorem (Theorem 19.1.5): a matrix with **every** entry positive has its spectral radius as a simple eigenvalue, with an eigenvector whose entries are all positive. That hypothesis is rarely met in practice. A Markov chain whose states can only step to their neighbors has a transition matrix full of zeros, and so does the matrix of almost any network. This section asks which of Perron's conclusions survive when zeros are allowed. There are two answers. For **every** nonnegative matrix, the spectral radius is still an eigenvalue with a nonnegative eigenvector, and of Perron's conclusions nothing more survives. For an **irreducible** matrix, in the sense of Section 2, every conclusion survives except one: other eigenvalues may reach the spectral radius in modulus. The positivity of the spectral radius also needs $n \geq 2$, because the 1×1 zero matrix counts as irreducible.

Throughout, matrices have real entries, and $\mathbf{A} \geq \mathbf{O}$, $\mathbf{A} > \mathbf{O}$, $\mathbf{x} \geq \mathbf{0}$ and $\mathbf{x} > \mathbf{0}$ are the entrywise order of Section 1. The spectrum is read in $\mathbb{C}$, as in Definition 16.4.1, so $\rho(\mathbf{A})$ is the largest modulus of a complex eigenvalue. We write $\mathbf{J}$ for the $n \times n$ matrix with every entry 1, and $\mathbf{1}$ for the all-ones vector.

19.3.1 What survives for every nonnegative matrix

Start with the weakest hypothesis, $\mathbf{A} \geq \mathbf{O}$, and see what can be rescued. Three small matrices show at once that most of Perron's theorem is lost.

Example. Three nonnegative matrices that break Perron's theorem

For each matrix, find ρ, its algebraic multiplicity, and all the nonnegative eigenvectors, and say which conclusion of Theorem 19.1.5 fails:

$$\mathbf{N} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \qquad \mathbf{A}_c = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & c \end{pmatrix} \quad \text{with } c = 2 \text{ and with } c = 1.$$

Solution. The nilpotent matrix. $\mathbf{N}$ is upper triangular with zero diagonal, so its only eigenvalue is 0, with algebraic multiplicity 2 (Theorem 9.7.4 (b)), and $\rho(\mathbf{N}) = 0$. The eigenvectors are the nonzero solutions of $\mathbf{Nx} = \mathbf{0}$, that is, of $x_2 = 0$: the nonzero multiples of $\mathbf{e}_1$. So $\rho = 0$ is not positive, it is not simple, and no eigenvector is positive. What does survive is that $\rho(\mathbf{N}) = 0$ is an eigenvalue with the nonnegative eigenvector $\mathbf{e}_1$.

The block matrix. $\mathbf{A}_c$ is block diagonal, with the 2×2 block $\left(\begin{smallmatrix}1&1\\1&1\end{smallmatrix}\right)$ and the 1×1 block (c). The 2×2 block has trace 2 and determinant 0, so its eigenvalues are 2 and 0, with eigenvectors $(1, 1)$ and $(1, -1)$. Hence $p_{\mathbf{A}_c}(t) = t(t - 2)(t - c)$, and the eigenvectors of $\mathbf{A}_c$ are built from those of the blocks: $(1, 1, 0)$ for 2, $(1, -1, 0)$ for 0, and $\mathbf{e}_3$ for c.

For $c = 2$: $\rho = 2$ has algebraic multiplicity 2. Its eigenspace is spanned by $(1, 1, 0)$ and $\mathbf{e}_3$, which are nonnegative and not proportional, so the nonnegative eigenvector is **not unique up to scale**. (It does contain the positive vector $(1, 1, 1)$, their sum.)

For $c = 1$: $\rho = 2$ is simple, but its eigenvectors are the multiples of $(1, 1, 0)$, none of them positive. And $\mathbf{e}_3 \geq \mathbf{0}$ is an eigenvector for the eigenvalue $1 < \rho$: a **nonnegative eigenvector for a smaller eigenvalue**, which Theorem 19.1.5 rules out for positive matrices.

In each case ρ was still an eigenvalue with a nonnegative eigenvector. That is the part that survives in general, and of Perron's conclusions it is all that survives. The proof is the first real use in the book of a *limiting auxiliary matrix* with a nonnegative matrix in the limit. Perron's theorem applies to $\mathbf{A} + \varepsilon\mathbf{J}$, which is positive for every $\varepsilon > 0$; let $\varepsilon \to 0^+$ and keep whatever passes to the limit.

Theorem 19.3.1. The Spectral Radius of a Nonnegative Matrix Is an Eigenvalue

Let $\mathbf{A} \in M_n(\mathbb{R})$ with $n \geq 1$ and $\mathbf{A} \geq \mathbf{O}$. Then $\rho(\mathbf{A})$ is an eigenvalue of $\mathbf{A}$, and there is a vector $\mathbf{x} \in \mathbb{R}^n$ with $\mathbf{x} \geq \mathbf{0}$, $\mathbf{x} \neq \mathbf{0}$ and $\mathbf{Ax} = \rho(\mathbf{A})\mathbf{x}$.

Idea. ① Perturb: $\mathbf{A}_k = \mathbf{A} + \frac{1}{k}\mathbf{J}$ is positive, so Theorem 19.1.5 gives it an eigenvalue $\rho(\mathbf{A}_k)$ with a positive eigenvector $\mathbf{v}_k$, which we scale into the simplex Δ_n. ② The matrices $\mathbf{A}_k$ decrease as k grows, so by monotonicity the numbers $\rho(\mathbf{A}_k)$ decrease and stay $\geq \rho(\mathbf{A})$; by monotone convergence they have a limit $r \geq \rho(\mathbf{A})$. ③ The $\mathbf{v}_k$ need not converge, but they live in a closed bounded set, so a subsequence does. ④ Pass to the limit in $\mathbf{A}_k\mathbf{v}_k = \rho(\mathbf{A}_k)\mathbf{v}_k$: the limit equation makes r an eigenvalue of $\mathbf{A}$, so $r \leq \rho(\mathbf{A})$, and $r = \rho(\mathbf{A})$. The scaling into Δ_n is what stops the limit vector from being $\mathbf{0}$: its entries still add up to 1.

Proof. Let $\rho = \rho(\mathbf{A})$, and for each integer $k \geq 1$ put $\mathbf{A}_k = \mathbf{A} + \frac{1}{k}\mathbf{J}$. Every entry of $\mathbf{A}_k$ is at least $\frac{1}{k} > 0$, so $\mathbf{A}_k > \mathbf{O}$. By Theorem 19.1.5, $\rho_k := \rho(\mathbf{A}_k)$ is an eigenvalue of $\mathbf{A}_k$ with an eigenvector $\mathbf{v}_k > \mathbf{0}$. Dividing $\mathbf{v}_k$ by the sum of its entries, which is positive, we may assume $\mathbf{v}_k \in \Delta_n$ (Definition 18.1.2).

Step 1: ρ_k converges to some $r \geq \rho$. Since $\frac{1}{k+1} < \frac{1}{k}$ and $\mathbf{J} \geq \mathbf{O}$, we have $\mathbf{O} \leq \mathbf{A} \leq \mathbf{A}_{k+1} \leq \mathbf{A}_k$ entrywise for every k. By Proposition 19.1.4 (a), applied to each of these two inequalities, $\rho \leq \rho_{k+1} \leq \rho_k$. So the sequence $(-\rho_k)$ is increasing and bounded above by $-\rho$. By monotone convergence, fact (A2) of Chapter 16's introduction, it converges to its least upper bound, which is $\leq -\rho$ because $-\rho$ is an upper bound. Hence $\rho_k \to r$ for some real

number $r \geq \rho$.

Step 2: a convergent subsequence. The set Δ_n is bounded, since every entry of a point of Δ_n lies in $[0, 1]$. It is closed: if points of Δ_n converge to $\mathbf{v}$, then each entry of $\mathbf{v}$ is a limit of numbers ≥ 0, hence ≥ 0, and the entries of $\mathbf{v}$ add up to a limit of sums equal to 1. By the compactness of closed bounded sets, fact (A3) of Chapter 16's introduction, the sequence $(\mathbf{v}_k)$ has a subsequence $(\mathbf{v}_{k_j})$ converging to some $\mathbf{v} \in \Delta_n$. In particular $\mathbf{v} \geq \mathbf{0}$, and $\mathbf{v} \neq \mathbf{0}$ because its entries add up to 1.

Step 3: the limit. For every j we have $\mathbf{A}_{k_j}\mathbf{v}_{k_j} = \rho_{k_j}\mathbf{v}_{k_j}$. Convergence in the Euclidean norm is entrywise convergence (Corollary 16.2.4), and $\mathbf{A}_{k_j} \to \mathbf{A}$ entrywise. Each entry of $\mathbf{A}_{k_j}\mathbf{v}_{k_j}$ is a sum of products of entries of $\mathbf{A}_{k_j}$ and of $\mathbf{v}_{k_j}$, so by the algebra of limits it converges to the corresponding entry of $\mathbf{Av}$; likewise $\rho_{k_j} \to r$ by Step 1, since a subsequence of a convergent sequence has the same limit, and so $\rho_{k_j}\mathbf{v}_{k_j} \to r\mathbf{v}$. Limits are unique, so $\mathbf{Av} = r\mathbf{v}$ with $\mathbf{v} \neq \mathbf{0}$; r is an eigenvalue of $\mathbf{A}$, hence $r \leq \rho$ by Definition 16.4.1; with $r \geq \rho$, $r = \rho$. So $\mathbf{Av} = \rho\mathbf{v}$, and $\mathbf{v} \geq \mathbf{0}$ is the required eigenvector. This proves the theorem.

■

Two imported facts were used, both named where they act: (A2) in Step 1 and (A3) in Step 2. No statement about the roots of polynomials is needed.

Remark. Alternatively, $\rho_k \to \rho$ also follows from the continuity of eigenvalues proved in Chapter 16 §07, but that rests on a theorem about the roots of polynomials which Chapter 16 quotes without proof; the route above needs nothing of the kind.

The theorem is sharp, in the sense that each of Perron's further conclusions fails for some nonnegative matrix. Example (Three nonnegative matrices that break Perron's theorem) shows that ρ can be 0, it can be a repeated eigenvalue, the nonnegative eigenvector can fail to be positive or unique, and another eigenvalue can have a nonnegative eigenvector; and the swap matrix below shows that another eigenvalue can reach $|\lambda| = \rho$. The positive conclusions of Theorem 19.1.5 are *strict* statements — an entry is > 0, a multiplicity is 1, an inequality $|\lambda| < \rho$ is strict — and strict inequalities do not survive limits. Positive numbers can tend to 0; two simple eigenvalues can merge.

Warning. **A nonnegative eigenvector does not identify ρ for a general nonnegative matrix.** In $\mathbf{A}_1$ of Example (Three nonnegative matrices that break Perron's theorem) the vector $\mathbf{e}_3 \geq \mathbf{0}$ is an eigenvector, but its eigenvalue is 1, while $\rho(\mathbf{A}_1) = 2$. To find ρ from an eigenvector, you need either irreducibility (next subsection) or a **positive** eigenvector, which Exercise C1 at the end of this section shows is enough.

19.3.2 The theorem for irreducible matrices

What goes wrong in Example (Three nonnegative matrices that break Perron's theorem) is visible in the zero pattern: $\mathbf{A}_c$ splits into two blocks that do not talk to each other, and $\mathbf{N}$ passes mass from one coordinate to the other but never back. Section 2 made "every coordinate eventually talks to every other" precise. A nonnegative matrix is **irreducible** (Definition 19.2.3) when no simultaneous permutation of its rows and columns brings it to block upper triangular

form, and by Theorem 19.2.6 this happens exactly when $(\mathbf{I}+\mathbf{A})^{n-1} > \mathbf{O}$. That criterion is the bridge we need: it turns an irreducible matrix into a positive one, to which Theorem 19.1.5 applies.

Theorem 19.3.2. Perron–Frobenius Theorem

Let $\mathbf{A} \in M_n(\mathbb{R})$ with $n \geq 1$ be nonnegative and irreducible, and put $\rho = \rho(\mathbf{A})$. Then:

(a) ρ is an eigenvalue of $\mathbf{A}$ with an eigenvector $\mathbf{v} > \mathbf{0}$;

(b) ρ is **algebraically simple**: $a_{\mathbf{A}}(\rho) = 1$;

(c) every nonnegative eigenvector of $\mathbf{A}$, for **any** eigenvalue, is a positive multiple of $\mathbf{v}$; in particular its eigenvalue is ρ, and the nonnegative eigenvector is unique up to scale;

(d) if $n \geq 2$, then $\rho > 0$.

Idea. There are two ways to reach a positive matrix, and we take the one that keeps strict information. The perturbation $\mathbf{A} + \varepsilon\mathbf{J}$ of the last proof needs a limit, and a limit cannot certify strict conclusions; Exercise C2 at the end of this section shows them lost when the limit matrix is reducible. So instead we use $\mathbf{B} = (\mathbf{I}+\mathbf{A})^{n-1}$, which is positive by Theorem 19.2.6 and needs no limit. Step roadmap:

① Theorem 19.3.1 gives $\mathbf{Ax} = \rho\mathbf{x}$ with $\mathbf{x} \geq \mathbf{0}$, $\mathbf{x} \neq \mathbf{0}$.

② $\mathbf{Bx} = (1+\rho)^{n-1}\mathbf{x}$, and a positive matrix times a nonzero nonnegative vector is positive; so $\mathbf{x} > \mathbf{0}$.

③ Every eigenvector of $\mathbf{A}$ is an eigenvector of $\mathbf{B} = q(\mathbf{A})$, and the Spectral Mapping Theorem transports multiplicities. So a nonnegative eigenvector of $\mathbf{A}$, or a repeated ρ, would be a nonnegative eigenvector of $\mathbf{B}$, or a repeated $\rho(\mathbf{B})$ — both of which Theorem 19.1.5 controls.

④ If $\rho = 0$, then $\mathbf{Ax} = \mathbf{0}$ with $\mathbf{x} > \mathbf{0}$ forces $\mathbf{A} = \mathbf{O}$, which is not irreducible for $n \geq 2$.

Proof. Let $q(t) = (1+t)^{n-1}$ and $\mathbf{B} = q(\mathbf{A}) = (\mathbf{I}+\mathbf{A})^{n-1}$. Since $\mathbf{A}$ is irreducible, $\mathbf{B} > \mathbf{O}$ by Theorem 19.2.6.

Step 1: an eigenvector $\mathbf{v} > \mathbf{0}$ for ρ. By Theorem 19.3.1 there is $\mathbf{x} \geq \mathbf{0}$, $\mathbf{x} \neq \mathbf{0}$, with $\mathbf{Ax} = \rho\mathbf{x}$. Then $(\mathbf{I}+\mathbf{A})\mathbf{x} = (1+\rho)\mathbf{x}$, and applying this $n-1$ times gives $\mathbf{Bx} = (1+\rho)^{n-1}\mathbf{x}$. Choose j_0 with $x_{j_0} > 0$, possible because $\mathbf{x} \geq \mathbf{0}$ and $\mathbf{x} \neq \mathbf{0}$. For every i,

$$(\mathbf{Bx})_i = \sum_{j=1}^{n} b_{ij}x_j \geq b_{ij_0}x_{j_0} > 0,$$

since every term is ≥ 0 and $b_{ij_0} > 0$. So $\mathbf{Bx} > \mathbf{0}$. As $1+\rho \geq 1$, we may divide: $\mathbf{x} = (1+\rho)^{-(n-1)}\mathbf{Bx} > \mathbf{0}$. Put $\mathbf{v} = \mathbf{x}$. This proves (a).

Step 2: eigenvectors of A are eigenvectors of B. If $\mathbf{Ay} = \mu\mathbf{y}$ with $\mathbf{y} \neq \mathbf{0}$, then, exactly as in Step 1, $\mathbf{By} = q(\mu)\mathbf{y}$. In particular $\mathbf{Bv} = q(\rho)\mathbf{v}$ with $\mathbf{v} > \mathbf{0}$. By Theorem 19.1.5 (d), a nonnegative eigenvector of the positive matrix $\mathbf{B}$ belongs to the eigenvalue $\rho(\mathbf{B})$; hence

$$q(\rho) = (1+\rho)^{n-1} = \rho(\mathbf{B}).$$

Step 3: simplicity. Over $\mathbb{C}$, write $p_{\mathbf{A}}(t) = (t-\lambda_1)\cdots(t-\lambda_n)$ with each eigenvalue repeated according to its algebraic multiplicity (Corollary 6.5.5). By the Spectral Mapping Theorem (Theorem 9.10.1),

$$p_{\mathbf{B}}(t) = (t - q(\lambda_1))\cdots(t - q(\lambda_n)).$$

Suppose $a_{\mathbf{A}}(\rho) \geq 2$, say $\lambda_1 = \lambda_2 = \rho$. Then $q(\rho) = \rho(\mathbf{B})$ occurs at least twice in this factorization (other eigenvalues mapping to $\rho(\mathbf{B})$ could only increase this count), so $a_{\mathbf{B}}(\rho(\mathbf{B})) \geq 2$, contradicting the simplicity of $\rho(\mathbf{B})$ in Theorem 19.1.5. Hence $a_{\mathbf{A}}(\rho) = 1$, which is (b).

Step 4: nonnegative eigenvectors. Let $\mathbf{y} \geq \mathbf{0}$, $\mathbf{y} \neq \mathbf{0}$, with $\mathbf{A}\mathbf{y} = \mu\mathbf{y}$ for some $\mu \in \mathbb{C}$. By Step 2, $\mathbf{y}$ is a nonnegative eigenvector of $\mathbf{B}$, so by Theorem 19.1.5 (d) it belongs to $\rho(\mathbf{B})$. So $\mathbf{y}$ and $\mathbf{v}$ both lie in the eigenspace $E_{\rho(\mathbf{B})}(\mathbf{B})$. Since $\rho(\mathbf{B})$ is algebraically simple, Theorem 9.3.1 gives $\dim E_{\rho(\mathbf{B})}(\mathbf{B}) = 1$, so $\mathbf{y} = c\mathbf{v}$ for some scalar $c \neq 0$. Choosing i with $y_i > 0$ gives $c = y_i/v_i > 0$. Finally $\mathbf{A}\mathbf{y} = c\mathbf{A}\mathbf{v} = \rho\, c\mathbf{v} = \rho\mathbf{y}$, so $\mu = \rho$. This proves (c).

Step 5: $\rho > 0$ when $n \geq 2$. Suppose $\rho = 0$. Then $\mathbf{A}\mathbf{v} = \mathbf{0}$, so for each i the sum $\sum_j a_{ij}v_j$ of nonnegative terms is 0, which forces $a_{ij}v_j = 0$ and hence $a_{ij} = 0$ for every j, as $v_j > 0$. So $\mathbf{A} = \mathbf{O}$ and $\mathbf{B} = \mathbf{I}_n$. For $n \geq 2$ the matrix $\mathbf{I}_n$ has a zero entry off the diagonal, so it is not positive, contradicting $\mathbf{B} > \mathbf{O}$. This proves (d), and the theorem.

■

The first sentence of the proof is where irreducibility did its work, and irreducibility was used only there. Note also what the theorem does **not** claim. For positive matrices, Theorem 19.1.5 adds that every other eigenvalue has modulus strictly less than ρ. For irreducible matrices this is false.

Warning. **Irreducible does not make ρ strictly dominant.** The swap $\mathbf{P} = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$ is irreducible, since $\mathbf{I} + \mathbf{P} = \mathbf{J}$ is positive (Theorem 19.2.6 with $n = 2$). Its eigenvalues are 1 and -1, with eigenvectors $(1, 1)$ and $(1, -1)$. The theorem holds as stated: $\rho = 1$ is simple with the positive eigenvector $(1, 1)$, and $(1, -1)$ is not nonnegative. But $|-1| = \rho$. Section 5 identifies the irreducible matrices for which ρ is strictly dominant.

The case $n = 1$ is worth one sentence. A 1×1 matrix (a) with $a \geq 0$ has $\rho = a$, simple, with the positive eigenvector (1), so clauses (a)–(c) hold for it outright; clause (d) must exclude $n = 1$, since $\mathbf{O}_1$ has $\rho = 0$. For $n \geq 2$ the zero matrix is the one thing (d) had to rule out, and irreducibility does.

Example. The theorem on a matrix with zeros

Let

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 0 \\ 2 & 0 & 1 \\ 2 & 1 & 0 \end{pmatrix}.$$

Show that $\mathbf{A}$ is irreducible, find $\rho(\mathbf{A})$ and a positive eigenvector, and check clause (c) of Theorem 19.3.2 against the other eigenvalue.

Solution. Irreducible. We have $\mathbf{I}+\mathbf{A} = \left(\begin{smallmatrix} 1 & 1 & 0 \\ 2 & 1 & 1 \\ 2 & 1 & 1 \end{smallmatrix}\right)$ and, multiplying out,

$$(\mathbf{I}+\mathbf{A})^2 = \begin{pmatrix} 3 & 2 & 1 \\ 6 & 4 & 2 \\ 6 & 4 & 2 \end{pmatrix} > \mathbf{O},$$

so $\mathbf{A}$ is irreducible by Theorem 19.2.6 with $n = 3$. Note that $\mathbf{A}$ itself has four zero entries and is not positive, so Theorem 19.1.5 does not apply to it directly.

The spectrum. Expanding $\det(t\mathbf{I} - \mathbf{A})$ along the first row,

$$\begin{aligned} p_{\mathbf{A}}(t) &= t\,(t^2 - 1) + (-2t - 2) \\ &= t^3 - 3t - 2 = (t-2)(t+1)^2. \end{aligned}$$

So the eigenvalues are 2 and -1, the latter with algebraic multiplicity 2, and $\rho(\mathbf{A}) = 2$, algebraically simple as clause (b) says. A repeated eigenvalue elsewhere in the spectrum is no contradiction: the theorem speaks only about ρ.

The eigenvector. $(\mathbf{A} - 2\mathbf{I})\mathbf{x} = \mathbf{0}$ reads $-2x_1 + x_2 = 0$, $2x_1 - 2x_2 + x_3 = 0$, $2x_1 + x_2 - 2x_3 = 0$. The first gives $x_2 = 2x_1$, the second then $x_3 = 2x_1$, and the third is satisfied: $2 + 2 - 4 = 0$. So $\mathbf{v} = (1, 2, 2) > \mathbf{0}$ and $\mathbf{A}\mathbf{v} = (2, 4, 4) = 2\mathbf{v}$.

The other eigenvalue. $\mathbf{A}+\mathbf{I} = \left(\begin{smallmatrix} 1 & 1 & 0 \\ 2 & 1 & 1 \\ 2 & 1 & 1 \end{smallmatrix}\right)$ has rank 2, and its null space is spanned by $(1, -1, -1)$: the first row gives $x_2 = -x_1$, the second $x_3 = -2x_1 - x_2 = -x_1$. Every eigenvector for -1 is a multiple of $(1, -1, -1)$ and has entries of both signs, as clause (c) predicts.

Quick check. For which values of $c \geq 0$ is the matrix $\mathbf{A}_c$ of Example (Three nonnegative matrices that break Perron's theorem) irreducible? Is the answer consistent with what the example found?

Solution. For none. Whatever c is, $\mathbf{I} + \mathbf{A}_c$ is block diagonal with blocks of sizes 2 and 1, so every power of it is block diagonal too, and $(\mathbf{I}+\mathbf{A}_c)^2$ has zeros in positions $(1,3)$, $(2,3)$, $(3,1)$, $(3,2)$. By Theorem 19.2.6, $\mathbf{A}_c$ is reducible. This is consistent: the example found conclusions of Theorem 19.3.2 failing for $c = 1$ and $c = 2$, which is possible only for a reducible matrix.

19.3.3 The left Perron vector

Everything so far concerned the equation $\mathbf{A}\mathbf{v} = \rho\mathbf{v}$. A nonnegative matrix acts on rows as well: in a Markov chain, the row $\mathbf{1}^\top$ is a left eigenvector of every stochastic matrix (Proposition 9.11.3 (c)). Left eigenvectors of $\mathbf{A}$ are eigenvectors of $\mathbf{A}^\top$ (Proposition 9.10.5 (a)), and irreducibility passes to the transpose, so the theorem applies to rows at no extra cost.

Corollary 19.3.3. The Left Perron Vector

Let $\mathbf{A} \in M_n(\mathbb{R})$ be nonnegative and irreducible, with $\rho = \rho(\mathbf{A})$ and $\mathbf{v} > \mathbf{0}$ as in Theorem 19.3.2. Then:

(a) $\mathbf{A}^\top$ is nonnegative and irreducible, and $\rho(\mathbf{A}^\top) = \rho$;

(b) there is $\mathbf{w} > \mathbf{0}$ with $\mathbf{w}^\top \mathbf{A} = \rho\, \mathbf{w}^\top$, and every nonnegative left eigenvector of $\mathbf{A}$ is a positive multiple of $\mathbf{w}$;

(c) $\mathbf{w}^\top \mathbf{v} > 0$, so $\mathbf{w}$ can be scaled to make $\mathbf{w}^\top \mathbf{v} = 1$.

Idea. Apply Theorem 19.3.2 to $\mathbf{A}^\top$ and read the answer back as rows. Two things must be checked first: that $\mathbf{A}^\top$ meets the hypotheses, which holds because the test $(\mathbf{I}+\mathbf{A})^{n-1} > \mathbf{O}$ survives transposition, and that $\mathbf{A}^\top$ has the same spectral radius, which holds because it has the same characteristic polynomial. Positivity of both vectors then makes $\mathbf{w}^\top \mathbf{v}$ a sum of positive terms.

Proof.

(a) The entries of $\mathbf{A}^\top$ are those of $\mathbf{A}$, so $\mathbf{A}^\top \geq \mathbf{O}$. Transposition reverses products and fixes $\mathbf{I}$ (Theorem 0.9.5), so $(\mathbf{I}+\mathbf{A}^\top)^{n-1} = \big((\mathbf{I}+\mathbf{A})^{n-1}\big)^\top$, which is positive because $(\mathbf{I}+\mathbf{A})^{n-1}$ is; hence $\mathbf{A}^\top$ is irreducible by Theorem 19.2.6. By Proposition 9.10.5 (b), $p_{\mathbf{A}^\top} = p_{\mathbf{A}}$, so $\mathbf{A}$ and $\mathbf{A}^\top$ have the same eigenvalues and $\rho(\mathbf{A}^\top) = \rho$.

(b) By (a) and Theorem 19.3.2 (a) applied to $\mathbf{A}^\top$, there is $\mathbf{w} > \mathbf{0}$ with $\mathbf{A}^\top \mathbf{w} = \rho \mathbf{w}$, which by Proposition 9.10.5 (a) says $\mathbf{w}^\top \mathbf{A} = \rho \mathbf{w}^\top$. A nonnegative left eigenvector of $\mathbf{A}$ is a nonnegative eigenvector of $\mathbf{A}^\top$, hence a positive multiple of $\mathbf{w}$ by Theorem 19.3.2 (c) applied to $\mathbf{A}^\top$.

(c) $\mathbf{w}^\top \mathbf{v} = \sum_i w_i v_i$ is a sum of positive numbers, so it is positive, and $\mathbf{w}/(\mathbf{w}^\top \mathbf{v})$ has the stated property.

■

This makes the following names well defined, since each vector is pinned down by (c) of the theorem or (b) of the corollary together with its normalization.

Definition 19.3.1. Perron Root, Perron Vectors

This definition extends Definition 19.1.2 from positive matrices to irreducible ones. Let $\mathbf{A} \in M_n(\mathbb{R})$ be nonnegative and irreducible. Its **Perron root** is $\rho(\mathbf{A})$. Its **(right) Perron vector** is the unique $\mathbf{v} > \mathbf{0}$ with $\mathbf{A}\mathbf{v} = \rho(\mathbf{A})\mathbf{v}$ and $\mathbf{1}^\top \mathbf{v} = 1$. Its **left Perron vector** is the unique $\mathbf{w} > \mathbf{0}$ with $\mathbf{w}^\top \mathbf{A} = \rho(\mathbf{A})\mathbf{w}^\top$ and $\mathbf{w}^\top \mathbf{v} = 1$.

The two normalizations are chosen for what they do. $\mathbf{1}^\top \mathbf{v} = 1$ makes $\mathbf{v}$ a probability vector, which is what a Markov chain wants; $\mathbf{w}^\top \mathbf{v} = 1$ makes the rank-one matrix $\mathbf{v}\mathbf{w}^\top$ idempotent, since

$$(\mathbf{v}\mathbf{w}^\top)(\mathbf{v}\mathbf{w}^\top) = \mathbf{v}\,(\mathbf{w}^\top \mathbf{v})\,\mathbf{w}^\top = \mathbf{v}\mathbf{w}^\top.$$

Section 5 shows that, for a large class of irreducible matrices, the powers $(\mathbf{A}/\rho)^k$ converge to exactly this matrix.

A positive matrix is irreducible, since then $(\mathbf{I}+\mathbf{A})^{n-1} \geq \mathbf{I}+\mathbf{A} > \mathbf{O}$ for $n \geq 2$: expand $(\mathbf{I}+\mathbf{A})^{n-1}$ by the binomial theorem, which applies because $\mathbf{I}$ and $\mathbf{A}$ commute, and note that every term is $\geq \mathbf{O}$ and two of them are $\mathbf{I}$ and $(n-1)\mathbf{A} \geq \mathbf{A}$. So this definition covers the positive matrices

of Section 1. For them the Perron root and both Perron vectors are those of Definition 19.1.2.

For the matrix of Example (The theorem on a matrix with zeros), the left eigenvector is $(3,2,1)$: indeed $(3,2,1)\mathbf{A} = (0+4+2,\ 3+0+1,\ 0+2+0) = (6,4,2)$. The right Perron vector is $\mathbf{v} = \frac{1}{5}(1,2,2)$, and $(3,2,1)\cdot\frac{1}{5}(1,2,2) = \frac{9}{5}$, so the left Perron vector is $\mathbf{w} = \frac{5}{9}(3,2,1)$.

Example. An irreducible stochastic matrix

Let

$$\mathbf{A} = \begin{pmatrix} 0 & 1/2 & 1 \\ 1 & 0 & 0 \\ 0 & 1/2 & 0 \end{pmatrix}.$$

Show that $\mathbf{A}$ is stochastic and irreducible, and find its Perron root and its right and left Perron vectors.

Solution. Each column is a probability vector, so $\mathbf{A}$ is stochastic (Definition 9.11.1). Multiplying out,

$$(\mathbf{I}+\mathbf{A})^2 = \begin{pmatrix} 3/2 & 3/2 & 2 \\ 2 & 3/2 & 1 \\ 1/2 & 1 & 1 \end{pmatrix} > \mathbf{O},$$

so $\mathbf{A}$ is irreducible (Theorem 19.2.6).

Perron root. By Proposition 9.11.3 (c), 1 is an eigenvalue. Every column sum is 1, so $\|\mathbf{A}\|_1 = 1$ and $\rho(\mathbf{A}) \le 1$ by Corollary 16.4.2. Hence $\rho(\mathbf{A}) = 1$.

Left Perron vector. $\mathbf{1}^{\top}\mathbf{A} = \mathbf{1}^{\top}$ with $\mathbf{1} > \mathbf{0}$, so by Corollary 19.3.3 (b) the left Perron vector is a multiple of $\mathbf{1}$.

Right Perron vector. $\mathbf{A}\mathbf{x} = \mathbf{x}$ reads $\frac{1}{2}x_2 + x_3 = x_1$, $x_1 = x_2$, $\frac{1}{2}x_2 = x_3$. With $x_3 = 1$ this gives $x_2 = 2$, $x_1 = 2$, and the first equation holds: $1 + 1 = 2$. So $\mathbf{v} = \frac{1}{5}(2,2,1)$. Then $\mathbf{1}^{\top}\mathbf{v} = 1$, and the left Perron vector is $\mathbf{w} = \mathbf{1}$ itself.

By Theorem 19.3.2 (c), $\mathbf{v}$ is the **only** probability vector with $\mathbf{A}\mathbf{v} = \mathbf{v}$: this chain has exactly one steady state, and every state has positive probability in it. The other two eigenvalues are the roots of $2t^2 + 2t + 1$, namely $\frac{-1\pm i}{2}$, of modulus $1/\sqrt{2} < 1$; the factorization is $p_{\mathbf{A}}(t) = (t-1)(t^2 + t + \frac{1}{2})$, which the reader can confirm by expanding $\det(t\mathbf{I} - \mathbf{A})$ along the first column.

For every irreducible stochastic matrix the same three lines apply, and they show that the left Perron vector is always $\mathbf{1}$ and the steady state always unique and positive. Section 6 draws the consequences for Markov chains.

19.3.4 Exercises

19.3.4.1 A. Check your understanding

Exercise. A1

(a) State the Perron–Frobenius theorem for irreducible nonnegative matrices.

(b) True or false: if $\mathbf{A} \ge \mathbf{O}$, then $\rho(\mathbf{A}) > 0$. Justify your answer.

(c) True or false: if $\mathbf{A} \geq \mathbf{O}$ is irreducible, then every eigenvalue $\lambda \neq \rho(\mathbf{A})$ has $|\lambda| < \rho(\mathbf{A})$. Justify your answer.

(d) True or false: if $\mathbf{A} \geq \mathbf{O}$ and $\mathbf{Ay} = \mu\mathbf{y}$ with $\mathbf{y} \geq \mathbf{0}$, $\mathbf{y} \neq \mathbf{0}$, then $\mu = \rho(\mathbf{A})$. Justify your answer.

(e) Which positive matrix does the proof of Theorem 19.3.2 apply Theorem 19.1.5 to, and why not to $\mathbf{A} + \varepsilon\mathbf{J}$?

Solution.

(a) See Theorem 19.3.2: for $\mathbf{A} \geq \mathbf{O}$ irreducible, $\rho(\mathbf{A})$ is an algebraically simple eigenvalue with a positive eigenvector, every nonnegative eigenvector is a positive multiple of that one, and $\rho(\mathbf{A}) > 0$ when $n \geq 2$.

(b) False. $\mathbf{N} = \left(\begin{smallmatrix} 0 & 1 \\ 0 & 0 \end{smallmatrix}\right) \geq \mathbf{O}$ has $\rho(\mathbf{N}) = 0$ (Example (Three nonnegative matrices that break Perron's theorem)). It is reducible, so this does not contradict Theorem 19.3.2 (d).

(c) False. The swap $\left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$ is irreducible with eigenvalues $1 = \rho$ and -1, and $|-1| = \rho$.

(d) False in general: $\mathbf{A}_1$ of Example (Three nonnegative matrices that break Perron's theorem) has $\mathbf{A}_1\mathbf{e}_3 = \mathbf{e}_3$ but $\rho(\mathbf{A}_1) = 2$. It is true when $\mathbf{A}$ is irreducible, by Theorem 19.3.2 (c).

(e) To $\mathbf{B} = (\mathbf{I} + \mathbf{A})^{n-1}$, which is positive by Theorem 19.2.6. The perturbation $\mathbf{A} + \varepsilon\mathbf{J}$ requires a limit $\varepsilon \to 0^+$, and a limit cannot certify strict conclusions — positivity of the eigenvector and simplicity of ρ — which are exactly what the theorem asserts; Exercise C2 shows them lost when the limit matrix is reducible.

19.3.4.2 B. Practice

Exercise. B1: Does the theorem apply?

For each matrix, determine whether Theorem 19.3.2 applies. Where it does, find ρ and a positive eigenvector; where it does not, say which hypothesis fails and whether the conclusion fails too. Justify your answers.

(a) $\begin{pmatrix} 0 & 2 \\ 8 & 0 \end{pmatrix}$ (b) $\begin{pmatrix} 1 & 2 \\ 0 & 3 \end{pmatrix}$ (c) $\begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$ (d) $\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$

Solution.

(a) Nonnegative, and $\mathbf{I} + \mathbf{A} = \left(\begin{smallmatrix} 1 & 2 \\ 8 & 1 \end{smallmatrix}\right) > \mathbf{O}$, so irreducible by Theorem 19.2.6 with $n = 2$. The theorem applies. $p_{\mathbf{A}}(t) = t^2 - 16$, so the eigenvalues are ± 4 and $\rho = 4$. From $-4x_1 + 2x_2 = 0$, $\mathbf{v} = (1, 2) > \mathbf{0}$; check: $\mathbf{Av} = (4, 8) = 4\mathbf{v}$. As with the swap, -4 has the same modulus as ρ.

(b) Nonnegative, but $(\mathbf{I} + \mathbf{A})^1 = \left(\begin{smallmatrix} 2 & 2 \\ 0 & 4 \end{smallmatrix}\right)$ has a zero entry, so $\mathbf{A}$ is reducible and the theorem does not apply. The eigenvalues are 1 and 3 (Theorem 9.7.4 (b)), so $\rho = 3$. Here $(\mathbf{A} - 3\mathbf{I})\mathbf{x} = \mathbf{0}$ reads $-2x_1 + 2x_2 = 0$, so $(1, 1) > \mathbf{0}$ is an eigenvector for ρ: conclusions (a) and (b) happen to hold. Conclusion (c) fails: $\mathbf{e}_1 \geq \mathbf{0}$ is an eigenvector for $1 \neq \rho$.

(c) Nonnegative. With $\mathbf{C}$ this matrix, $\mathbf{Ce}_1 = \mathbf{e}_3$, $\mathbf{Ce}_3 = \mathbf{e}_2$, $\mathbf{Ce}_2 = \mathbf{e}_1$, and $\mathbf{I} + \mathbf{C} + \mathbf{C}^2 = \mathbf{J}$; so $(\mathbf{I} + \mathbf{C})^2 = \mathbf{I} + 2\mathbf{C} + \mathbf{C}^2 \geq \mathbf{I} + \mathbf{C} + \mathbf{C}^2 = \mathbf{J} > \mathbf{O}$, and $\mathbf{C}$ is irreducible. The theorem

applies. $\mathbf{C}^3 = \mathbf{I}$, and $p_{\mathbf{C}}(t) = t^3 - 1$ (expand along the first column), so the eigenvalues are the three cube roots of unity, all of modulus 1, and $\rho = 1$ with $\mathbf{v} = \mathbf{1}$, since every row of $\mathbf{C}$ has a single 1.

(d) Not nonnegative, so the theorem does not apply, and its conclusion fails: $p(t) = t^2 + 1$, the eigenvalues are $\pm i$, and $\rho = 1$ is not an eigenvalue at all.

Exercise. B2: Both Perron vectors

Let

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 0 \\ 2 & 0 & 2 \\ 1 & 0 & 1 \end{pmatrix}.$$

Show that $\mathbf{A}$ is irreducible, and find its Perron root, its right Perron vector, and its left Perron vector (Definition 19.3.1). Hence write down the matrix $\mathbf{v}\mathbf{w}^\top$ and check that it is idempotent.

Solution. Irreducible. $\mathbf{I} + \mathbf{A} = \left(\begin{smallmatrix} 1 & 1 & 0 \\ 2 & 1 & 2 \\ 1 & 0 & 2 \end{smallmatrix}\right)$, and

$$(\mathbf{I} + \mathbf{A})^2 = \begin{pmatrix} 3 & 2 & 2 \\ 6 & 3 & 6 \\ 3 & 1 & 4 \end{pmatrix} > \mathbf{O},$$

so $\mathbf{A}$ is irreducible by Theorem 19.2.6.

Perron root. Expanding $\det(t\mathbf{I} - \mathbf{A})$ along the first row,

$$\begin{aligned} p_{\mathbf{A}}(t) &= t\big(t(t-1) - 0\big) + \big(-2(t-1) - 2\big) \\ &= t^3 - t^2 - 2t = t(t-2)(t+1), \end{aligned}$$

so the eigenvalues are $2, 0, -1$ and $\rho = 2$.

Right. $(\mathbf{A} - 2\mathbf{I})\mathbf{x} = \mathbf{0}$: the first row gives $x_2 = 2x_1$, the third $x_1 - x_3 = 0$. So $\mathbf{x} = (1, 2, 1)$; check the second row: $2 - 4 + 2 = 0$. Normalizing, $\mathbf{v} = \frac{1}{4}(1, 2, 1)$.

Left. $(\mathbf{A}^\top - 2\mathbf{I})\mathbf{y} = \mathbf{0}$, with $\mathbf{A}^\top = \left(\begin{smallmatrix} 0 & 2 & 1 \\ 1 & 0 & 0 \\ 0 & 2 & 1 \end{smallmatrix}\right)$: the second row gives $y_1 = 2y_2$, the third $2y_2 - y_3 = 0$. So $\mathbf{y} = (2, 1, 2)$; check the first row: $-4+2+2 = 0$. Then $\mathbf{y}^\top\mathbf{v} = \frac{1}{4}(2+2+2) = \frac{3}{2}$, so $\mathbf{w} = \frac{2}{3}(2, 1, 2)$.

The product.

$$\mathbf{v}\mathbf{w}^\top = \frac{1}{4} \cdot \frac{2}{3} \begin{pmatrix} 2 & 1 & 2 \\ 4 & 2 & 4 \\ 2 & 1 & 2 \end{pmatrix} = \frac{1}{6} \begin{pmatrix} 2 & 1 & 2 \\ 4 & 2 & 4 \\ 2 & 1 & 2 \end{pmatrix}.$$

It is idempotent because $(\mathbf{v}\mathbf{w}^\top)^2 = \mathbf{v}(\mathbf{w}^\top\mathbf{v})\mathbf{w}^\top = \mathbf{v}\mathbf{w}^\top$, as $\mathbf{w}^\top\mathbf{v} = 1$. Directly: call the displayed integer matrix $\mathbf{M}$. Row i of $\mathbf{M}^2$ is row i of $\mathbf{M}$ times $\mathbf{M}$; the first row gives $2(2, 1, 2) + 1(4, 2, 4) + 2(2, 1, 2) = 6(2, 1, 2)$, and the other rows are multiples of the first, so $\mathbf{M}^2 = 6\mathbf{M}$. Hence $\left(\frac{1}{6}\mathbf{M}\right)^2 = \frac{1}{36} \cdot 6\mathbf{M} = \frac{1}{6}\mathbf{M}$.

Exercise. B3: A reducible matrix, clause by clause

Let

$$\mathbf{A} = \begin{pmatrix} 2 & 0 & 0 \\ 1 & 3 & 0 \\ 0 & 1 & 2 \end{pmatrix}.$$

Show that $\mathbf{A}$ is reducible. Find $\rho(\mathbf{A})$ and every nonnegative eigenvector of $\mathbf{A}$. Hence decide which of the four conclusions of Theorem 19.3.2 hold for $\mathbf{A}$ and which fail.

Solution. Reducible. $\mathbf{I} + \mathbf{A}$ is lower triangular, so $(\mathbf{I} + \mathbf{A})^2$ is lower triangular and has zero entries above the diagonal; by Theorem 19.2.6, $\mathbf{A}$ is reducible.

Spectrum. $\mathbf{A}$ is lower triangular, so its eigenvalues are $2, 3, 2$ (Theorem 9.7.4 (b), applied to $\mathbf{A}^\top$, which has the same characteristic polynomial by Proposition 9.10.5 (b)). So $\rho = 3$, with $a_{\mathbf{A}}(3) = 1$.

Eigenvectors for 3. $(\mathbf{A} - 3\mathbf{I})\mathbf{x} = \mathbf{0}$ reads $-x_1 = 0$, $x_1 = 0$, $x_2 - x_3 = 0$; the solutions are the multiples of $(0, 1, 1)$. *Eigenvectors for 2.* $(\mathbf{A} - 2\mathbf{I})\mathbf{x} = \mathbf{0}$ reads $x_1 + x_2 = 0$, $x_2 = 0$; the solutions are the multiples of $(0, 0, 1)$. So the nonnegative eigenvectors are the positive multiples of $(0, 1, 1)$ and of $\mathbf{e}_3$.

Verdict. Conclusion (b) holds and (d) holds ($\rho = 3 > 0$). Conclusion (a) fails in its strong form: ρ is an eigenvalue, as Theorem 19.3.1 guarantees, but no eigenvector for it is positive. Conclusion (c) fails: $\mathbf{e}_3 \geq \mathbf{0}$ is an eigenvector for $2 \neq \rho$.

19.3.4.3 C. Going deeper

Exercise. C1: A positive eigenvector always belongs to the spectral radius

Let $\mathbf{A} \in M_n(\mathbb{R})$ with $\mathbf{A} \geq \mathbf{O}$, **not** assumed irreducible.

(a) Prove that if $\mathbf{A}\mathbf{x} = \mu\mathbf{x}$ with $\mathbf{x} > \mathbf{0}$, then $\mu = \rho(\mathbf{A})$.

(b) Deduce that if every row of $\mathbf{A}$ has the same sum s, then $\rho(\mathbf{A}) = s$.

(c) Show by an example that (a) fails if $\mathbf{x} > \mathbf{0}$ is weakened to $\mathbf{x} \geq \mathbf{0}$, $\mathbf{x} \neq \mathbf{0}$.

Hint: apply Theorem 19.3.1 to $\mathbf{A}^\top$.

Solution.

(a) Since $\mathbf{A}$ and $\mathbf{x}$ are real and $\mathbf{x} \neq \mathbf{0}$, $\mu = (\mathbf{A}\mathbf{x})_i / x_i$ for any i is real, and it is ≥ 0 because $(\mathbf{A}\mathbf{x})_i \geq 0$ and $x_i > 0$. By Proposition 9.10.5 (b), $\rho(\mathbf{A}^\top) = \rho(\mathbf{A}) = \rho$, and $\mathbf{A}^\top \geq \mathbf{O}$, so Theorem 19.3.1 gives $\mathbf{y} \geq \mathbf{0}$, $\mathbf{y} \neq \mathbf{0}$, with $\mathbf{A}^\top\mathbf{y} = \rho\mathbf{y}$, that is, $\mathbf{y}^\top\mathbf{A} = \rho\mathbf{y}^\top$. Compute $\mathbf{y}^\top\mathbf{A}\mathbf{x}$ two ways:

$$\rho\,\mathbf{y}^\top\mathbf{x} = (\mathbf{y}^\top\mathbf{A})\mathbf{x} = \mathbf{y}^\top(\mathbf{A}\mathbf{x}) = \mu\,\mathbf{y}^\top\mathbf{x}.$$

Now $\mathbf{y}^\top\mathbf{x} = \sum_i y_i x_i > 0$, as every term is ≥ 0 and the terms with $y_i > 0$ are positive. Dividing, $\mu = \rho$.

(b) Equal row sums s say $\mathbf{A}\mathbf{1} = s\mathbf{1}$, and $\mathbf{1} > \mathbf{0}$, so $s = \rho(\mathbf{A})$ by (a).

(c) $\mathbf{A}_1$ of Example (Three nonnegative matrices that break Perron's theorem) has $\mathbf{A}_1\mathbf{e}_3 = \mathbf{e}_3$ with $\mathbf{e}_3 \geq \mathbf{0}$, $\mathbf{e}_3 \neq \mathbf{0}$, but $\rho(\mathbf{A}_1) = 2 \neq 1$.

Exercise. C2: What the limit loses

This exercise shows why the proof of Theorem 19.3.2 does not take the limit $\mathbf{A} + \varepsilon\mathbf{J} \to \mathbf{A}$.

(a) For $\varepsilon > 0$, find the eigenvalues of $\mathbf{I}_2 + \varepsilon\mathbf{J}$. Show that ρ is simple for every $\varepsilon > 0$, but not at $\varepsilon = 0$.

(b) For $\varepsilon > 0$, let $\mathbf{N}_\varepsilon = \left(\begin{smallmatrix} \varepsilon & 1+\varepsilon \\ \varepsilon & \varepsilon \end{smallmatrix}\right)$, which is $\mathbf{N} + \varepsilon\mathbf{J}$ for the nilpotent $\mathbf{N}$ of Example (Three nonnegative matrices that break Perron's theorem). Show that $\rho(\mathbf{N}_\varepsilon) = \varepsilon + \sqrt{\varepsilon(1+\varepsilon)}$ with eigenvector $(\sqrt{1+\varepsilon},\ \sqrt{\varepsilon}) > \mathbf{0}$, and find the limit of this eigenvector as $\varepsilon \to 0^+$.

(c) Which conclusions of Theorem 19.1.5 survive the limit in these two examples, and which do not? Explain in one sentence why this is to be expected.

Solution.

(a) $\mathbf{J}\mathbf{1} = 2 \cdot \mathbf{1}$ and $\mathbf{J}(1,-1) = \mathbf{0}$, so $\mathbf{I}_2 + \varepsilon\mathbf{J}$ has the eigenvalue $1 + 2\varepsilon$ on $(1,1)$ and 1 on $(1,-1)$. For $\varepsilon > 0$ these are distinct, so $\rho = 1 + 2\varepsilon$ is simple. At $\varepsilon = 0$ the matrix is $\mathbf{I}_2$ and $\rho = 1$ has algebraic multiplicity 2.

(b) $\mathbf{N}_\varepsilon - \varepsilon\mathbf{I} = \left(\begin{smallmatrix} 0 & 1+\varepsilon \\ \varepsilon & 0 \end{smallmatrix}\right)$ has characteristic polynomial $t^2 - \varepsilon(1+\varepsilon)$, so the eigenvalues of $\mathbf{N}_\varepsilon$ are $\varepsilon \pm \sqrt{\varepsilon(1+\varepsilon)}$, and $\rho(\mathbf{N}_\varepsilon) = \varepsilon + \sqrt{\varepsilon(1+\varepsilon)}$, which exceeds the modulus of the other one since $\varepsilon > 0$. With $s = \sqrt{\varepsilon(1+\varepsilon)}$ and $\mathbf{u} = (\sqrt{1+\varepsilon}, \sqrt{\varepsilon})$,

$$(\mathbf{N}_\varepsilon - \varepsilon\mathbf{I})\mathbf{u} = \big((1+\varepsilon)\sqrt{\varepsilon},\ \varepsilon\sqrt{1+\varepsilon}\big) = s\,\mathbf{u},$$

since $(1+\varepsilon)\sqrt{\varepsilon} = s\sqrt{1+\varepsilon}$ and $\varepsilon\sqrt{1+\varepsilon} = s\sqrt{\varepsilon}$. So $\mathbf{N}_\varepsilon\mathbf{u} = (\varepsilon + s)\mathbf{u}$ with $\mathbf{u} > \mathbf{0}$. As $\varepsilon \to 0^+$, $\mathbf{u} \to (1,0) = \mathbf{e}_1$, which is nonnegative but not positive.

(c) What survives is that ρ is an eigenvalue with a nonnegative eigenvector, which is Theorem 19.3.1. Simplicity is lost in (a), and positivity of the eigenvector in (b). This is expected because those conclusions are strict — a multiplicity equal to 1, entries > 0 — and a limit of positive numbers need only be ≥ 0, while two distinct eigenvalues can converge to one.

Exercise. C3: Uniqueness through the left Perron vector

Let $\mathbf{A} \geq \mathbf{O}$ be irreducible, with left Perron vector $\mathbf{w}$ (Definition 19.3.1, Corollary 19.3.3).

(a) Using Theorem 9.10.6, give a second proof that if $\mathbf{A}\mathbf{y} = \mu\mathbf{y}$ with $\mathbf{y} \geq \mathbf{0}$, $\mathbf{y} \neq \mathbf{0}$, then $\mu = \rho(\mathbf{A})$.

(b) Deduce that an irreducible stochastic matrix has exactly one steady state (Definition 9.11.2), and that it is positive.

Solution.

(a) Suppose $\mu \neq \rho(\mathbf{A})$. The vector $\mathbf{w}$ is a left eigenvector for $\rho(\mathbf{A})$ and $\mathbf{y}$ a right eigenvector for μ, so Theorem 9.10.6 gives $\mathbf{w}^{\top}\mathbf{y} = 0$. But $\mathbf{w} > \mathbf{0}$, $\mathbf{y} \geq \mathbf{0}$ and $\mathbf{y} \neq \mathbf{0}$, so $\mathbf{w}^{\top}\mathbf{y} =$

$\sum_i w_i y_i \geq w_i y_i > 0$ for an index i with $y_i > 0$. This contradiction shows $\mu = \rho(\mathbf{A})$.

(b) Let $\mathbf{A}$ be stochastic and irreducible. As in Example (An irreducible stochastic matrix), $\rho(\mathbf{A}) = 1$: 1 is an eigenvalue by Proposition 9.11.3 (c), and $\rho(\mathbf{A}) \leq \|\mathbf{A}\|_1 = 1$ by Corollary 16.4.2. A steady state is a probability vector $\mathbf{x}$ with $\mathbf{A}\mathbf{x} = \mathbf{x}$: a nonnegative eigenvector for $1 = \rho(\mathbf{A})$, normalized by $\mathbf{1}^\top \mathbf{x} = 1$. By Theorem 19.3.2 (c), every such $\mathbf{x}$ is a positive multiple of the Perron vector $\mathbf{v}$, and the normalization forces $\mathbf{x} = \mathbf{v}$. So there is exactly one steady state, namely $\mathbf{v} > \mathbf{0}$.

19.4 The Collatz–Wielandt Formula

Section 3 proved that the spectral radius of a nonnegative matrix is an eigenvalue, and that for an irreducible matrix it comes with a positive eigenvector. It did not say how to find the number. Finding it through the characteristic polynomial means finding the largest root of a polynomial of degree n, which in general has no formula in radicals once $n \geq 5$. This section gives a way to trap $\rho(\mathbf{A})$ between two numbers read off from a single test vector, with no eigenvalue computed. Each test vector gives a lower and an upper bound, and the bounds meet exactly at the Perron vector. That turns $\rho(\mathbf{A})$ into a maximum and a minimum over vectors, much as Chapter 17 turned the eigenvalues of a Hermitian matrix into extreme values of the Rayleigh quotient (Definition 17.1.1).

Throughout, matrices have real entries, and $\geq$, $>$ between matrices or vectors are the entrywise order of Section 1. We use one elementary fact repeatedly. If $\mathbf{A} \geq \mathbf{O}$ and $\mathbf{u} \geq \mathbf{u}'$, then $\mathbf{A}\mathbf{u} \geq \mathbf{A}\mathbf{u}'$, because each entry of $\mathbf{A}(\mathbf{u} - \mathbf{u}')$ is a sum of products of nonnegative numbers. In words, a nonnegative matrix preserves the order.

19.4.1 Bounds from a single test vector

If $\mathbf{v} > \mathbf{0}$ is an eigenvector for ρ, then the ratios $(\mathbf{A}\mathbf{v})_i / v_i$ are all equal to ρ. For a vector that is not an eigenvector, the ratios spread out. The idea of this section is that ρ always lies somewhere in that spread.

Take $\mathbf{x} = \mathbf{1}$. Then $(\mathbf{A}\mathbf{1})_i$ is the i-th row sum, so the ratios are the row sums. For

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 1 \\ 0 & 1 & 2 \\ 1 & 0 & 3 \end{pmatrix}$$

they are $2, 3, 4$. The characteristic polynomial is $t^3 - 4t^2 + 2t - 1$, and $\rho(\mathbf{A}) \approx 3.51$, which indeed lies between 2 and 4. Two propositions make "lies in the spread" precise, one for each end. The upper end is the classical one.

Proposition 19.4.1. Subinvariance

Let $\mathbf{A} \in M_n(\mathbb{R})$ with $\mathbf{A} \geq \mathbf{O}$, and let $s \in \mathbb{R}$.

(a) If $\mathbf{x} > \mathbf{0}$ and $\mathbf{A}\mathbf{x} \leq s\mathbf{x}$, then $\rho(\mathbf{A}) \leq s$.

(b) If $\mathbf{A}$ is irreducible, $\mathbf{x} \geq \mathbf{0}$, $\mathbf{x} \neq \mathbf{0}$ and $\mathbf{A}\mathbf{x} \leq s\mathbf{x}$, then $\rho(\mathbf{A}) \leq s$. If moreover $\mathbf{A}\mathbf{x} \neq s\mathbf{x}$, then $\rho(\mathbf{A}) < s$.

Idea. For (a), rescale the coordinates by the entries of $\mathbf{x}$. In the new coordinates, $\mathbf{x}$ becomes $\mathbf{1}$, the hypothesis says that every row sum is at most s, and Chapter 16's row-sum bound finishes the job. For (b), $\mathbf{x}$ may have zero entries, so no rescaling is possible. Instead, pair the inequality with the left Perron vector $\mathbf{w} > \mathbf{0}$, which turns $\mathbf{w}^{\top}\mathbf{A}\mathbf{x}$ into $\rho\,\mathbf{w}^{\top}\mathbf{x}$.

Proof.

(a) Let $\mathbf{D} = \operatorname{diag}(x_1, \dots, x_n)$, which is invertible because every $x_i > 0$, and put $\mathbf{C} = \mathbf{D}^{-1}\mathbf{A}\mathbf{D}$. Its entries are $c_{ij} = a_{ij}x_j/x_i \geq 0$, and its i-th row sum is

$$\sum_{j=1}^{n} \frac{a_{ij}x_j}{x_i} = \frac{(\mathbf{A}\mathbf{x})_i}{x_i} \leq s,$$

where the inequality divides $(\mathbf{A}\mathbf{x})_i \leq s x_i$ by $x_i > 0$. Since the entries of $\mathbf{C}$ are nonnegative, its absolute row sums are its row sums, so $\|\mathbf{C}\|_\infty \leq s$ by Theorem 16.3.3 (b). The matrices $\mathbf{A}$ and $\mathbf{C}$ are similar, so they have the same characteristic polynomial (Theorem 7.10.3) and the same spectral radius. Hence $\rho(\mathbf{A}) = \rho(\mathbf{C}) \leq \|\mathbf{C}\|_\infty \leq s$ by Corollary 16.4.2.

(b) Let $\rho = \rho(\mathbf{A})$ and let $\mathbf{w} > \mathbf{0}$ be a left eigenvector for ρ (Corollary 19.3.3). Since $\mathbf{w} \geq \mathbf{0}$ and $s\mathbf{x} - \mathbf{A}\mathbf{x} \geq \mathbf{0}$, the number $\mathbf{w}^{\top}(s\mathbf{x} - \mathbf{A}\mathbf{x})$ is ≥ 0. But $\mathbf{w}^{\top}\mathbf{A}\mathbf{x} = \rho\,\mathbf{w}^{\top}\mathbf{x}$, so

$$\mathbf{w}^{\top}(s\mathbf{x} - \mathbf{A}\mathbf{x}) = (s - \rho)\,\mathbf{w}^{\top}\mathbf{x}.$$

Choose i with $x_i > 0$. Then $\mathbf{w}^{\top}\mathbf{x} \geq w_i x_i > 0$, so $s - \rho \geq 0$. If moreover $\mathbf{A}\mathbf{x} \neq s\mathbf{x}$, then $s\mathbf{x} - \mathbf{A}\mathbf{x}$ has some entry > 0, and pairing with $\mathbf{w} > \mathbf{0}$ gives $\mathbf{w}^{\top}(s\mathbf{x} - \mathbf{A}\mathbf{x}) > 0$, hence $s - \rho > 0$. This proves the proposition.

■

The name comes from the inequality $\mathbf{A}\mathbf{x} \leq s\mathbf{x}$: the vector $\mathbf{x}$ is carried into something no bigger than a multiple of itself, "sub-invariant" up to the factor s.

Warning. **In part (a), the test vector must be positive unless A is irreducible.** For $\mathbf{A} = \operatorname{diag}(1, 5)$ and $\mathbf{x} = \mathbf{e}_1$, we have $\mathbf{A}\mathbf{x} = \mathbf{e}_1 \leq 1 \cdot \mathbf{x}$, yet $\rho(\mathbf{A}) = 5$. The zero entry of $\mathbf{x}$ hides the second coordinate from the test entirely. Part (b) allows zero entries only because irreducibility guarantees that every coordinate is seen through $\mathbf{w} > \mathbf{0}$.

The lower end needs no positivity at all, and it holds for every nonnegative matrix.

Proposition 19.4.2. Superinvariance

Let $\mathbf{A} \in M_n(\mathbb{R})$ with $\mathbf{A} \geq \mathbf{O}$, let $s \in \mathbb{R}$, and let $\mathbf{x} \geq \mathbf{0}$, $\mathbf{x} \neq \mathbf{0}$, with $\mathbf{A}\mathbf{x} \geq s\mathbf{x}$.

(a) Then $\rho(\mathbf{A}) \geq s$.

(b) If $\mathbf{A}$ is irreducible and $\mathbf{A}\mathbf{x} \neq s\mathbf{x}$, then $\rho(\mathbf{A}) > s$.

Idea. For (a), the proof cannot use a similarity, because $\mathbf{x}$ may have zero entries. Instead it iterates the inequality to get $\mathbf{A}^k\mathbf{x} \geq s^k\mathbf{x}$, so the powers of $\mathbf{A}$ grow at least like s^k, and Gelfand's formula reads $\rho(\mathbf{A})$ off that growth. For (b), pair with the left Perron vector exactly as in Proposition 19.4.1 (b).

Proof.

(a) If $s \leq 0$, then $\rho(\mathbf{A}) \geq 0 \geq s$. Let $s > 0$. We claim that $\mathbf{A}^k\mathbf{x} \geq s^k\mathbf{x}$ for every $k \geq 0$. The case $k = 0$ is an equality. If it holds for k, then

$$\mathbf{A}^{k+1}\mathbf{x} = \mathbf{A}(\mathbf{A}^k\mathbf{x}) \geq \mathbf{A}(s^k\mathbf{x}) = s^k\mathbf{A}\mathbf{x} \geq s^{k+1}\mathbf{x},$$

where the first inequality holds because $\mathbf{A} \geq \mathbf{O}$ preserves the order, and the second multiplies $\mathbf{A}\mathbf{x} \geq s\mathbf{x}$ by $s^k > 0$. This proves the claim.
Choose i with $x_i = \|\mathbf{x}\|_\infty$, which is positive since $\mathbf{x} \geq \mathbf{0}$ and $\mathbf{x} \neq \mathbf{0}$. By the claim, and Theorem 16.3.2 (a),

$$s^k\|\mathbf{x}\|_\infty = s^k x_i \leq (\mathbf{A}^k\mathbf{x})_i \leq \|\mathbf{A}^k\mathbf{x}\|_\infty \leq \|\mathbf{A}^k\|_\infty\|\mathbf{x}\|_\infty.$$

Dividing by $\|\mathbf{x}\|_\infty > 0$ and taking k-th roots gives $\|\mathbf{A}^k\|_\infty^{1/k} \geq s$ for every $k \geq 1$. The operator norm $\|\cdot\|_\infty$ is a matrix norm (Theorem 16.3.2 (d)), so by Gelfand's formula (Theorem 16.4.4) the left side tends to $\rho(\mathbf{A})$ as $k \to \infty$. A nonstrict inequality survives the limit, so $\rho(\mathbf{A}) \geq s$.

(b) Let $\mathbf{w} > \mathbf{0}$ be a left eigenvector of $\mathbf{A}$ for $\rho = \rho(\mathbf{A})$ (Corollary 19.3.3). As in the proof of Proposition 19.4.1 (b), $\mathbf{w}^\top(\mathbf{A}\mathbf{x} - s\mathbf{x}) = (\rho - s)\,\mathbf{w}^\top\mathbf{x}$ with $\mathbf{w}^\top\mathbf{x} > 0$. Now $\mathbf{A}\mathbf{x} - s\mathbf{x} \geq \mathbf{0}$ has a positive entry, so the left side is > 0 and $\rho > s$.

■

Remark. For an irreducible $\mathbf{A}$, part (a) also follows from the one-line pairing argument of part (b), exactly as in Proposition 19.4.1 (b). The detour through Gelfand's formula is what makes (a) hold for **every** nonnegative matrix, where $\mathbf{w}$ may have zero entries and $\mathbf{w}^\top\mathbf{x}$ may vanish.

Quick check. Let $\mathbf{A} = \begin{pmatrix} 2 & 3 \\ 1 & 0 \end{pmatrix}$. Use $\mathbf{x} = \mathbf{1}$ and $\mathbf{x} = (3, 1)$ to bound $\rho(\mathbf{A})$ above and below. What do you conclude?

Solution. $\mathbf{A}\mathbf{1} = (5, 1)$, so $\mathbf{A}\mathbf{1} \geq 1 \cdot \mathbf{1}$ and $\mathbf{A}\mathbf{1} \leq 5 \cdot \mathbf{1}$; the two propositions give $1 \leq \rho(\mathbf{A}) \leq 5$. Next, $\mathbf{A}(3, 1) = (9, 3) = 3 \cdot (3, 1)$, so $3 \leq \rho(\mathbf{A}) \leq 3$, that is, $\rho(\mathbf{A}) = 3$. Indeed $p_{\mathbf{A}}(t) = t^2 - 2t - 3 = (t - 3)(t + 1)$. A test vector that happens to be an eigenvector closes the gap completely.

19.4.2 The two Collatz–Wielandt functions

The propositions ask for a number s with $\mathbf{Ax} \le s\mathbf{x}$ or $\mathbf{Ax} \ge s\mathbf{x}$. The best such numbers for a given $\mathbf{x}$ deserve names, since they are the bounds a test vector actually delivers.

Definition 19.4.1. Collatz–Wielandt Functions

Let $\mathbf{A} \in M_n(\mathbb{R})$ with $\mathbf{A} \ge \mathbf{O}$.

(a) For $\mathbf{x} \ge \mathbf{0}$ with $\mathbf{x} \ne \mathbf{0}$, the **lower Collatz–Wielandt function** is

$$\underline{r}_{\mathbf{A}}(\mathbf{x}) := \min\Big\{ \frac{(\mathbf{Ax})_i}{x_i} : 1 \le i \le n,\ x_i > 0 \Big\}.$$

(b) For $\mathbf{x} > \mathbf{0}$, the **upper Collatz–Wielandt function** is

$$\overline{r}_{\mathbf{A}}(\mathbf{x}) := \max\Big\{ \frac{(\mathbf{Ax})_i}{x_i} : 1 \le i \le n \Big\}.$$

In words: $\underline{r}_{\mathbf{A}}(\mathbf{x})$ is the smallest of the ratios, taken only over the coordinates where $\mathbf{x}$ is positive, and $\overline{r}_{\mathbf{A}}(\mathbf{x})$ is the largest of the ratios, for a vector with no zero coordinates. Both are well defined: the first set is nonempty because $\mathbf{x} \ne \mathbf{0}$, and both are finite sets of real numbers.

The definition is built so that

$$\mathbf{Ax} \ge \underline{r}_{\mathbf{A}}(\mathbf{x})\,\mathbf{x} \quad (\mathbf{x} \ge \mathbf{0},\ \mathbf{x} \ne \mathbf{0}), \qquad \mathbf{Ax} \le \overline{r}_{\mathbf{A}}(\mathbf{x})\,\mathbf{x} \quad (\mathbf{x} > \mathbf{0}).$$

In the first, the coordinates with $x_i > 0$ satisfy $(\mathbf{Ax})_i \ge \underline{r}_{\mathbf{A}}(\mathbf{x})x_i$ by the choice of the minimum, and the coordinates with $x_i = 0$ satisfy $(\mathbf{Ax})_i \ge 0 = \underline{r}_{\mathbf{A}}(\mathbf{x})x_i$ because $\mathbf{Ax} \ge \mathbf{0}$. And $\underline{r}_{\mathbf{A}}(\mathbf{x})$ is the **largest** s with $\mathbf{Ax} \ge s\mathbf{x}$, since any such s is at most each ratio; likewise $\overline{r}_{\mathbf{A}}(\mathbf{x})$ is the **smallest** s with $\mathbf{Ax} \le s\mathbf{x}$.

Some examples, simplest first.

- **The all-ones vector.** $\underline{r}_{\mathbf{A}}(\mathbf{1})$ and $\overline{r}_{\mathbf{A}}(\mathbf{1})$ are the smallest and the largest row sum of $\mathbf{A}$.
- **An eigenvector.** If $\mathbf{x} > \mathbf{0}$ and $\mathbf{Ax} = \mu\mathbf{x}$, every ratio equals μ, so $\underline{r}_{\mathbf{A}}(\mathbf{x}) = \overline{r}_{\mathbf{A}}(\mathbf{x}) = \mu$.
- **A standard basis vector.** $\underline{r}_{\mathbf{A}}(\mathbf{e}_j) = a_{jj}$: only the coordinate j counts, and there the ratio is $(\mathbf{Ae}_j)_j/1 = a_{jj}$.
- **The 1×1 case.** For $\mathbf{A} = (a)$, both functions are the constant $a = \rho(\mathbf{A})$.

Nonexample by minimal change. Try to define $\overline{r}_{\mathbf{A}}$ on vectors with zero entries in the same way as $\underline{r}_{\mathbf{A}}$, as the maximum over the coordinates with $x_i > 0$. For the swap $\mathbf{A} = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$ and $\mathbf{x} = \mathbf{e}_1$, we get $\mathbf{Ax} = \mathbf{e}_2$, and the only ratio is $0/1 = 0$. But $\rho(\mathbf{A}) = 1$, so this "upper bound" is $0 < \rho$. What fails is the inequality $\mathbf{Ax} \le 0 \cdot \mathbf{x}$: in the ignored coordinate, $(\mathbf{Ax})_2 = 1 > 0$. The coordinates where $\mathbf{x}$ vanishes are harmless for a lower bound, where they contribute $(\mathbf{Ax})_i \ge 0$, and fatal for an upper bound.

Warning. **The two functions are not mirror images.** The lower function accepts every nonzero $\mathbf{x} \ge \mathbf{0}$; the upper one accepts only $\mathbf{x} > \mathbf{0}$. The formula of the next subsection is therefore a maximum over the closed orthant and a minimum over the open one, and letting the minimum range over the closed orthant too would make it false, as the nonexample shows.

19.4.3 The formula

With the two functions in hand, the propositions of the first subsection say that $\underline{r}_{\mathbf{A}}(\mathbf{x}) \le \rho(\mathbf{A}) \le \overline{r}_{\mathbf{A}}(\mathbf{x})$, and the second example shows both bounds attained at a positive eigenvector. For an irreducible matrix, such an eigenvector exists (Theorem 19.3.2), and more is true: nothing else attains either bound.

Theorem 19.4.3. Collatz–Wielandt Formula

Let $\mathbf{A} \in M_n(\mathbb{R})$ be nonnegative and irreducible, with Perron vector $\mathbf{v}$ (Definition 19.3.1). Then

$$\rho(\mathbf{A}) = \max_{\substack{\mathbf{x} \ge \mathbf{0} \\ \mathbf{x} \ne \mathbf{0}}} \underline{r}_{\mathbf{A}}(\mathbf{x}) = \min_{\mathbf{x} > \mathbf{0}} \overline{r}_{\mathbf{A}}(\mathbf{x}),$$

both extreme values are attained at $\mathbf{x} = \mathbf{v}$, and they are attained **only** at the positive multiples of $\mathbf{v}$: if $\mathbf{x} \ge \mathbf{0}$, $\mathbf{x} \ne \mathbf{0}$, and $\underline{r}_{\mathbf{A}}(\mathbf{x}) = \rho(\mathbf{A})$, or if $\mathbf{x} > \mathbf{0}$ and $\overline{r}_{\mathbf{A}}(\mathbf{x}) = \rho(\mathbf{A})$, then $\mathbf{x}$ is a positive multiple of $\mathbf{v}$.

Idea. The inequalities are the two propositions, applied with the best constants $\underline{r}_{\mathbf{A}}(\mathbf{x})$ and $\overline{r}_{\mathbf{A}}(\mathbf{x})$. Attainment is the eigenvector example. For the "only" clause, use the strict halves of the propositions: if $\mathbf{x}$ attains a bound without being an eigenvector, the strict inequality would push ρ strictly past itself. The usual proofs of this formula start from the optimization problem and need a compactness argument, which is delicate for a matrix with zero entries: then $\underline{r}_{\mathbf{A}}$ can jump where a coordinate of $\mathbf{x}$ becomes 0, whereas for a positive matrix it is continuous. Here the compactness was spent once, in Sections 1 and 3, and it has already produced the maximizer.

Proof. Let $\rho = \rho(\mathbf{A})$.

The inequalities. Let $\mathbf{x} \ge \mathbf{0}$, $\mathbf{x} \ne \mathbf{0}$. Since $\mathbf{A}\mathbf{x} \ge \underline{r}_{\mathbf{A}}(\mathbf{x})\,\mathbf{x}$, Proposition 19.4.2 (a) gives $\underline{r}_{\mathbf{A}}(\mathbf{x}) \le \rho$. Let $\mathbf{x} > \mathbf{0}$. Since $\mathbf{A}\mathbf{x} \le \overline{r}_{\mathbf{A}}(\mathbf{x})\,\mathbf{x}$, Proposition 19.4.1 (a) gives $\rho \le \overline{r}_{\mathbf{A}}(\mathbf{x})$.

Attainment. By Theorem 19.3.2 (a) and Definition 19.3.1, $\mathbf{v} > \mathbf{0}$ and $\mathbf{A}\mathbf{v} = \rho\mathbf{v}$, so every ratio $(\mathbf{A}\mathbf{v})_i / v_i$ equals ρ, and $\underline{r}_{\mathbf{A}}(\mathbf{v}) = \overline{r}_{\mathbf{A}}(\mathbf{v}) = \rho$. With the inequalities, this shows that ρ is the maximum of $\underline{r}_{\mathbf{A}}$ and the minimum of $\overline{r}_{\mathbf{A}}$.

Only at multiples of $\mathbf{v}$. Let $\mathbf{x} \ge \mathbf{0}$, $\mathbf{x} \ne \mathbf{0}$, with $\underline{r}_{\mathbf{A}}(\mathbf{x}) = \rho$, so that $\mathbf{A}\mathbf{x} \ge \rho\mathbf{x}$. If $\mathbf{A}\mathbf{x} \ne \rho\mathbf{x}$, then Proposition 19.4.2 (b) with $s = \rho$ gives $\rho > \rho$, which is absurd. Hence $\mathbf{A}\mathbf{x} = \rho\mathbf{x}$, and $\mathbf{x}$ is a positive multiple of $\mathbf{v}$ by Theorem 19.3.2 (c). If instead $\mathbf{x} > \mathbf{0}$ and $\overline{r}_{\mathbf{A}}(\mathbf{x}) = \rho$, then $\mathbf{A}\mathbf{x} \le \rho\mathbf{x}$, and Proposition 19.4.1 (b) rules out $\mathbf{A}\mathbf{x} \ne \rho\mathbf{x}$ in the same way. Conversely, $\underline{r}_{\mathbf{A}}(c\mathbf{v}) = \overline{r}_{\mathbf{A}}(c\mathbf{v}) = \rho$ for $c > 0$, since the ratios do not change when $\mathbf{x}$ is scaled. This proves the theorem.

■

So $\rho(\mathbf{A})$ is characterized without eigenvalues: it is the largest growth factor that $\mathbf{A}$ guarantees in every coordinate of some nonnegative vector, and the smallest growth factor that bounds every coordinate of some positive vector.

Remark. Irreducibility is used only for the attainment of the minimum and for the "only" clause. For **every** $\mathbf{A} \geq \mathbf{O}$, the inequalities hold by the same two propositions, and the maximum is still attained, at the nonnegative eigenvector $\mathbf{x}$ of Theorem 19.3.1: every ratio over a coordinate with $x_i > 0$ equals ρ. The minimum can fail to be attained; Exercise C3 at the end of this section works out an example.

19.4.4 Row sums and column sums

The simplest test vector is $\mathbf{1}$, and the transpose supplies a second family of bounds for free. For a nonnegative matrix, write $r_i = \sum_j a_{ij}$ for the i-th row sum and $c_j = \sum_i a_{ij}$ for the j-th column sum.

Corollary 19.4.4. Row and Column Sums Bound the Spectral Radius

Let $\mathbf{A} \in M_n(\mathbb{R})$ with $\mathbf{A} \geq \mathbf{O}$.

(a) $\min_i r_i \leq \rho(\mathbf{A}) \leq \max_i r_i$ and $\min_j c_j \leq \rho(\mathbf{A}) \leq \max_j c_j$.

(b) If $\mathbf{A}$ is irreducible, then $\rho(\mathbf{A}) = \min_i r_i$ if and only if $\rho(\mathbf{A}) = \max_i r_i$, if and only if all the row sums are equal. The same holds for the column sums.

Idea. Use the test vector $\mathbf{1}$, whose ratios are the row sums: the two propositions give (a) for rows, and the "only" clause of Theorem 19.4.3 gives (b), since equality at $\mathbf{1}$ makes $\mathbf{1}$ an eigenvector. For columns, run the same argument on $\mathbf{A}^\top$, which has the same spectral radius.

Proof.

(a) $\mathbf{A}\mathbf{1}$ is the vector of row sums, so $\mathbf{A}\mathbf{1} \geq (\min_i r_i)\mathbf{1}$ and $\mathbf{A}\mathbf{1} \leq (\max_i r_i)\mathbf{1}$, with $\mathbf{1} > \mathbf{0}$. Proposition 19.4.2 (a) and Proposition 19.4.1 (a) give the row bounds. The column sums of $\mathbf{A}$ are the row sums of $\mathbf{A}^\top \geq \mathbf{O}$, and $\rho(\mathbf{A}^\top) = \rho(\mathbf{A})$ because $p_{\mathbf{A}^\top} = p_{\mathbf{A}}$ (Proposition 9.10.5 (b)); so the row bounds for $\mathbf{A}^\top$ are the column bounds for $\mathbf{A}$.

(b) If all row sums equal r, then $\min_i r_i = \max_i r_i = r$, and (a) gives $\rho(\mathbf{A}) = r$. Conversely, suppose $\rho(\mathbf{A}) = \min_i r_i = \underline{r}_{\mathbf{A}}(\mathbf{1})$. By the "only" clause of Theorem 19.4.3, $\mathbf{1}$ is a positive multiple of the Perron vector, so $\mathbf{A}\mathbf{1} = \rho(\mathbf{A})\mathbf{1}$, and every row sum equals $\rho(\mathbf{A})$. The same argument with $\overline{r}_{\mathbf{A}}(\mathbf{1}) = \max_i r_i$ handles the maximum. For the columns, apply this to $\mathbf{A}^\top$, which is irreducible with $\rho(\mathbf{A}^\top) = \rho(\mathbf{A})$ by Corollary 19.3.3 (a).

■

Compare this with Chapter 16. Corollary 16.4.2 gives $\rho(\mathbf{A}) \leq \|\mathbf{A}\|_\infty$ and $\rho(\mathbf{A}) \leq \|\mathbf{A}\|_1$ for **every** complex matrix, and for $\mathbf{A} \geq \mathbf{O}$ these are exactly the upper bounds in (a), since the absolute row and column sums are then the row and column sums. What nonnegativity adds is the **lower** bound. For a matrix with entries of both signs there is none:

$$\mathbf{M} = \begin{pmatrix} 1 & -1 \\ 1 & -1 \end{pmatrix}$$

has absolute row sums $2, 2$, but $\mathbf{M}^2 = \mathbf{O}$, so every eigenvalue λ satisfies $\lambda^2 = 0$ and $\rho(\mathbf{M}) = 0$.

Nor do the signed row sums help:

$$\mathbf{M}' = \begin{pmatrix} 7 & -3 \\ 19 & -7 \end{pmatrix}$$

has row sums 4 and 12, both larger than $\rho(\mathbf{M}') = 2\sqrt{2}$, since $p_{\mathbf{M}'}(t) = t^2 + 8$.

The equality clause (b) genuinely needs irreducibility. The reducible matrix $\operatorname{diag}(1, 2)$ has $\rho = 2 = \max_i r_i$ with unequal row sums, and $\left(\begin{smallmatrix} 1 & 0 \\ 1 & 1 \end{smallmatrix}\right)$ has $\rho = 1 = \min_i r_i$ with row sums 1 and 2.

Example. Four bounds, and which one to keep

Bound $\rho(\mathbf{A})$ above and below using row and column sums, for

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 1 \\ 0 & 1 & 2 \\ 1 & 0 & 3 \end{pmatrix}.$$

Solution. The row sums are $2, 3, 4$ and the column sums are $1, 2, 6$. By Corollary 19.4.4 (a), $2 \le \rho(\mathbf{A}) \le 4$ from the rows and $1 \le \rho(\mathbf{A}) \le 6$ from the columns. Keep the larger lower bound and the smaller upper bound: $2 \le \rho(\mathbf{A}) \le 4$. Chapter 16's Corollary 16.4.2 gives only the upper bound 4.

The matrix is irreducible: $(\mathbf{I} + \mathbf{A})^2 = \left(\begin{smallmatrix} 2 & 3 & 7 \\ 2 & 4 & 12 \\ 5 & 1 & 17 \end{smallmatrix}\right) > \mathbf{O}$ (Theorem 19.2.6). Its row sums are not all equal, so by Corollary 19.4.4 (b) both bounds are strict: $2 < \rho(\mathbf{A}) < 4$.

19.4.5 Estimating the spectral radius

A bracket of width 2 is crude. The Collatz–Wielandt formula says that a better test vector gives a better bracket, and a test vector close to the Perron vector gives a very narrow one. There is a cheap way to improve a test vector: multiply it by $\mathbf{A}$.

The improvement is guaranteed. Let $\mathbf{x} > \mathbf{0}$, and suppose $\mathbf{Ax} > \mathbf{0}$ too. Since $\mathbf{A}$ preserves the order,

$$\mathbf{A}(\mathbf{Ax}) \ge \mathbf{A}(\underline{r}_{\mathbf{A}}(\mathbf{x})\,\mathbf{x}) = \underline{r}_{\mathbf{A}}(\mathbf{x})\,\mathbf{Ax},$$

so $\underline{r}_{\mathbf{A}}(\mathbf{Ax}) \ge \underline{r}_{\mathbf{A}}(\mathbf{x})$, because $\underline{r}_{\mathbf{A}}(\mathbf{Ax})$ is the largest s with $\mathbf{A}(\mathbf{Ax}) \ge s\,\mathbf{Ax}$. In the same way $\overline{r}_{\mathbf{A}}(\mathbf{Ax}) \le \overline{r}_{\mathbf{A}}(\mathbf{x})$. So the brackets from $\mathbf{x}, \mathbf{Ax}, \mathbf{A}^2\mathbf{x}, \dots$ are nested, as long as each $\mathbf{A}^k\mathbf{x} > \mathbf{0}$: each lower bound is at least the previous one, and each upper bound is at most the previous one. The condition holds whenever $\mathbf{x} > \mathbf{0}$ and $\mathbf{A}$ is irreducible with $n \ge 2$, since no row of $\mathbf{A}$ is then zero (a zero row, moved to the bottom by a permutation, would give the block form of Definition 19.2.3 with $k = n - 1$), so every entry of $\mathbf{Ay}$ is positive when $\mathbf{y} > \mathbf{0}$.

Example. Trapping the spectral radius

For the matrix $\mathbf{A}$ of Example (Four bounds, and which one to keep), compute the brackets $[\underline{r}_{\mathbf{A}}(\mathbf{x}), \overline{r}_{\mathbf{A}}(\mathbf{x})]$ for $\mathbf{x} = \mathbf{1}, \mathbf{A1}, \mathbf{A}^2\mathbf{1}, \mathbf{A}^3\mathbf{1}, \mathbf{A}^4\mathbf{1}$. Hence give $\rho(\mathbf{A})$ to two decimal places, and check the answer against the characteristic polynomial.

Solution. Each step multiplies the previous vector by $\mathbf{A}$, whose rows are $(0,1,1)$, $(0,1,2)$, $(1,0,3)$. The ratios are taken coordinate by coordinate.

$\mathbf{x}$	$\mathbf{Ax}$	ratios $(\mathbf{Ax})_i/x_i$	bracket
$(1,1,1)$	$(2,3,4)$	$2,\ 3,\ 4$	$[2,\ 4]$
$(2,3,4)$	$(7,11,14)$	$\frac{7}{2},\ \frac{11}{3},\ \frac{7}{2}$	$[\frac{7}{2},\ \frac{11}{3}]$
$(7,11,14)$	$(25,39,49)$	$\frac{25}{7},\ \frac{39}{11},\ \frac{7}{2}$	$[\frac{7}{2},\ \frac{25}{7}]$
$(25,39,49)$	$(88,137,172)$	$\frac{88}{25},\ \frac{137}{39},\ \frac{172}{49}$	$[\frac{172}{49},\ \frac{88}{25}]$
$(88,137,172)$	$(309,481,604)$	$\frac{309}{88},\ \frac{481}{137},\ \frac{151}{43}$	$[\frac{481}{137},\ \frac{151}{43}]$

For instance, in the last row $\mathbf{A}(88,137,172) = (137+172,\ 137+344,\ 88+516) = (309,481,604)$, and $\frac{604}{172} = \frac{151}{43}$. All five test vectors are positive, so each bracket is valid by Theorem 19.4.3, and they are nested, as the argument before the example predicts. The five brackets have widths 2, $\frac{1}{6}$, $\frac{1}{14}$, $\frac{12}{1225}$ and $\frac{4}{5891}$. Since $\frac{481}{137} = 3.51094\ldots$ and $\frac{151}{43} = 3.51162\ldots$,

$$3.5109 < \rho(\mathbf{A}) < 3.5117,$$

and therefore $\rho(\mathbf{A}) = 3.51$ to two decimal places.

Check. Expanding $\det(t\mathbf{I} - \mathbf{A})$ along the first column gives

$$\begin{aligned} p_{\mathbf{A}}(t) &= t((t-1)(t-3)) - ((-1)(-2) - (-1)(t-1)) \\ &= t^3 - 4t^2 + 2t - 1. \end{aligned}$$

Exactly, $p_{\mathbf{A}}(3.51) = -0.016849 < 0$ and $p_{\mathbf{A}}(3.512) = 0.004937728 > 0$. A real polynomial that changes sign on an interval has a root there, by the intermediate value theorem, fact (A5) of Chapter 16's introduction. So $p_{\mathbf{A}}$ has a real root between 3.51 and 3.512, consistent with the bracket.

Two things about this computation deserve a comment. First, every number in it is a certified bound, not an approximation whose error is unknown: $\rho(\mathbf{A}) \geq \frac{172}{49}$ is a theorem about this matrix, proved by one multiplication. Second, the iteration $\mathbf{x} \mapsto \mathbf{Ax}$ is the power method, and Chapter 24 studies how fast it converges. Nesting alone does not make the brackets shrink to a point, as the next warning shows.

Warning. **The brackets need not close up.** For the swap $\mathbf{A} = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$, which is irreducible with $\rho = 1$, start from $\mathbf{x} = (1,2)$. Then $\mathbf{Ax} = (2,1)$, with ratios 2 and $\frac{1}{2}$, and $\mathbf{A}^2\mathbf{x} = \mathbf{x}$. The bracket is $[\frac{1}{2}, 2]$ at every step and never improves. The iteration only swaps the two coordinates, because the eigenvalue -1 has the same modulus as ρ. Section 5 studies the irreducible matrices whose other eigenvalues all have modulus strictly less than ρ.

19.4.6 Exercises

19.4.6.1 A. Check your understanding

Exercise. A1

(a) Define $\underline{r}_{\mathbf{A}}(\mathbf{x})$ and $\overline{r}_{\mathbf{A}}(\mathbf{x})$, and state on which vectors each is defined.

(b) For $\mathbf{A} = \left(\begin{smallmatrix} 1 & 2 \\ 3 & 0 \end{smallmatrix}\right)$, compute $\underline{r}_{\mathbf{A}}(\mathbf{1})$ and $\overline{r}_{\mathbf{A}}(\mathbf{1})$. What do you conclude about $\rho(\mathbf{A})$?

(c) True or false: if $\mathbf{A} \geq \mathbf{O}$, $\mathbf{x} \geq \mathbf{0}$, $\mathbf{x} \neq \mathbf{0}$ and $\mathbf{Ax} \leq 2\mathbf{x}$, then $\rho(\mathbf{A}) \leq 2$. Justify your answer.

(d) True or false: for every $\mathbf{A} \geq \mathbf{O}$, $\rho(\mathbf{A})$ is at least the smallest row sum of $\mathbf{A}$. Justify your answer.

(e) Name the result from Chapter 16 that gives the upper bound in Corollary 19.4.4 (a), and say what that result does not give.

Solution.

(a) See Definition 19.4.1: $\underline{r}_{\mathbf{A}}(\mathbf{x})$ is the minimum of $(\mathbf{Ax})_i/x_i$ over the indices with $x_i > 0$, defined for $\mathbf{x} \geq \mathbf{0}$, $\mathbf{x} \neq \mathbf{0}$; $\overline{r}_{\mathbf{A}}(\mathbf{x})$ is the maximum of $(\mathbf{Ax})_i/x_i$ over all indices, defined only for $\mathbf{x} > \mathbf{0}$.

(b) $\mathbf{A1} = (3, 3)$, so both equal 3, and $\rho(\mathbf{A}) = 3$ by Proposition 19.4.2 (a) and Proposition 19.4.1 (a). Indeed $p_{\mathbf{A}}(t) = t^2 - t - 6 = (t - 3)(t + 2)$.

(c) False: $\mathbf{A} = \operatorname{diag}(1, 5)$ and $\mathbf{x} = \mathbf{e}_1$ give $\mathbf{Ax} = \mathbf{x} \leq 2\mathbf{x}$, but $\rho(\mathbf{A}) = 5$. It becomes true if $\mathbf{x} > \mathbf{0}$ (Proposition 19.4.1 (a)) or if $\mathbf{A}$ is irreducible (Proposition 19.4.1 (b)).

(d) True, by Corollary 19.4.4 (a), which rests on Proposition 19.4.2 (a) with $\mathbf{x} = \mathbf{1}$. No irreducibility is needed.

(e) Corollary 16.4.2, which bounds ρ above by the largest absolute row sum and the largest absolute column sum of any complex matrix. It gives no lower bound, and none exists for matrices with entries of both signs: $\left(\begin{smallmatrix} 1 & -1 \\ 1 & -1 \end{smallmatrix}\right)$ has $\rho = 0$.

19.4.6.2 B. Practice

Exercise. B1: A certified bracket

Let

$$\mathbf{A} = \begin{pmatrix} 1 & 2 & 0 \\ 0 & 1 & 3 \\ 1 & 0 & 2 \end{pmatrix}.$$

Show that $\mathbf{A}$ is irreducible. Compute the brackets $[\underline{r}_{\mathbf{A}}(\mathbf{x}), \overline{r}_{\mathbf{A}}(\mathbf{x})]$ for $\mathbf{x} = \mathbf{1}$ and for $\mathbf{x} = (10, 11, 8)$. Hence show that $3.18 < \rho(\mathbf{A}) \leq 3.25$, and check this against $p_{\mathbf{A}}$.

Solution. Irreducible. $\mathbf{I} + \mathbf{A} = \left(\begin{smallmatrix} 2 & 2 & 0 \\ 0 & 2 & 3 \\ 1 & 0 & 3 \end{smallmatrix}\right)$, and $(\mathbf{I} + \mathbf{A})^2 = \left(\begin{smallmatrix} 4 & 8 & 6 \\ 3 & 4 & 15 \\ 5 & 2 & 9 \end{smallmatrix}\right) > \mathbf{O}$, so $\mathbf{A}$ is irreducible by Theorem 19.2.6.

Brackets. $\mathbf{A1} = (3, 4, 3)$, giving $[3, 4]$. Next,

$$\mathbf{A}(10, 11, 8) = (10 + 22,\ 11 + 24,\ 10 + 16) = (32, 35, 26),$$

with ratios $\frac{32}{10} = 3.2$, $\frac{35}{11} = 3.1818\ldots$, $\frac{26}{8} = 3.25$. The bracket is $[\frac{35}{11}, \frac{13}{4}]$. By Theorem 19.4.3, $\frac{35}{11} \leq \rho(\mathbf{A}) \leq \frac{13}{4}$. Since $\frac{35}{11} > 3.18$, this gives $3.18 < \rho(\mathbf{A}) \leq 3.25$.

Check. Expanding along the first row,

$$\begin{aligned} p_{\mathbf{A}}(t) &= (t-1)\big((t-1)(t-2)\big) + 2(0-3) \\ &= t^3 - 4t^2 + 5t - 8. \end{aligned}$$

Then $p_{\mathbf{A}}(\frac{35}{11}) = -\frac{498}{1331} < 0$ and $p_{\mathbf{A}}(\frac{13}{4}) = \frac{21}{64} > 0$, so by the intermediate value theorem (fact (A5) of Chapter 16's introduction) $p_{\mathbf{A}}$ has a root in the bracket, consistent with the estimate. (In fact $\rho(\mathbf{A}) \approx 3.219$.)

Exercise. B2: When a sum bound is exact

For each matrix, give the bounds of Corollary 19.4.4 (a) from rows and from columns, and decide whether they determine ρ. Justify your answers.

$$\text{(a)}\ \begin{pmatrix} 1 & 2 & 1 \\ 3 & 0 & 1 \\ 0 & 2 & 2 \end{pmatrix} \qquad \text{(b)}\ \begin{pmatrix} 1 & 3 \\ 2 & 0 \end{pmatrix} \qquad \text{(c)}\ \begin{pmatrix} 2 & 0 \\ 5 & 3 \end{pmatrix}$$

Solution.

(a) Row sums $4, 4, 4$; column sums $4, 4, 4$. Both bracket ρ in $[4, 4]$, so $\rho = 4$, with eigenvector $\mathbf{1}$.

(b) Row sums $4, 2$, giving $[2, 4]$; column sums $3, 3$, giving $[3, 3]$. So $\rho = 3$, with $\mathbf{1}$ as a left eigenvector: $\mathbf{1}^{\top}\mathbf{A} = (3, 3)$. Check: $p(t) = t^2 - t - 6 = (t-3)(t+2)$. The matrix is irreducible ($\mathbf{I} + \mathbf{A} > \mathbf{O}$), and the row sums are unequal, so by Corollary 19.4.4 (b) the row bracket is strict, $2 < 3 < 4$, as it is.

(c) Row sums $2, 8$, giving $[2, 8]$; column sums $7, 3$, giving $[3, 7]$. Together, $3 \le \rho \le 7$. The matrix is lower triangular, so its eigenvalues are 2 and 3 and $\rho = 3$, the smallest column sum, although the column sums are unequal. This does not contradict Corollary 19.4.4 (b): the matrix is reducible, since $\mathbf{I}+\mathbf{A}$ has a zero entry in position $(1, 2)$ and $n-1 = 1$.

Exercise. B3: A weighted test vector

Let

$$\mathbf{A} = \frac{1}{10}\begin{pmatrix} 5 & 20 \\ 1 & 2 \end{pmatrix}.$$

Show that neither the row sums nor the column sums prove $\rho(\mathbf{A}) < 1$. Then use the test vector $\mathbf{x} = (5, 1)$ to prove $\rho(\mathbf{A}) \le \frac{9}{10}$. Hence deduce that $\mathbf{A}^k \to \mathbf{O}$ as $k \to \infty$.

Solution. The row sums are $\frac{25}{10}$ and $\frac{3}{10}$, and the column sums $\frac{6}{10}$ and $\frac{22}{10}$. Each upper bound in Corollary 19.4.4 (a) exceeds 1, so neither proves $\rho(\mathbf{A}) < 1$.

With $\mathbf{x} = (5, 1) > \mathbf{0}$,

$$\mathbf{A}\mathbf{x} = \tfrac{1}{10}(25 + 20,\ 5 + 2) = (\tfrac{45}{10},\ \tfrac{7}{10}),$$

and the ratios are $\frac{45}{50} = \frac{9}{10}$ and $\frac{7}{10}$. So $\mathbf{A}\mathbf{x} \le \frac{9}{10}\mathbf{x}$, and Proposition 19.4.1 (a) gives $\rho(\mathbf{A}) \le$

$\frac{9}{10} < 1$. Hence $\mathbf{A}^k \to \mathbf{O}$ by Corollary 16.4.5 (a).

19.4.6.3 C. Going deeper

Exercise. C1: Lower bounds from single entries

Let $\mathbf{A} \in M_n(\mathbb{R})$ with $\mathbf{A} \geq \mathbf{O}$, not necessarily irreducible.

(a) Prove that $\rho(\mathbf{A}) \geq a_{ii}$ for every i.

(b) Prove that $\rho(\mathbf{A}) \geq \sqrt{a_{ij}a_{ji}}$ for every $i \neq j$.

(c) Hence show that $\left(\begin{smallmatrix} 1 & 4 & 0 \\ 9 & 2 & 0 \\ 0 & 0 & 5 \end{smallmatrix}\right)$ has $\rho \geq 6$, and check this exactly.

Hint: choose a test vector supported on $\{i\}$ or on $\{i, j\}$.

Solution.

(a) Take $\mathbf{x} = \mathbf{e}_i$. Then $(\mathbf{A}\mathbf{e}_i)_i = a_{ii}$ and every other coordinate of $\mathbf{A}\mathbf{e}_i$ is ≥ 0, so $\mathbf{A}\mathbf{e}_i \geq a_{ii}\mathbf{e}_i$, and Proposition 19.4.2 (a) gives $\rho(\mathbf{A}) \geq a_{ii}$.

(b) If $a_{ij}a_{ji} = 0$, the claim is $\rho(\mathbf{A}) \geq 0$, which holds. Otherwise put $s = \sqrt{a_{ij}a_{ji}} > 0$ and $\mathbf{x} = \sqrt{a_{ij}}\,\mathbf{e}_i + \sqrt{a_{ji}}\,\mathbf{e}_j \geq \mathbf{0}$, $\mathbf{x} \neq \mathbf{0}$. Dropping nonnegative terms,

$$(\mathbf{A}\mathbf{x})_i \geq a_{ij}\sqrt{a_{ji}} = s\sqrt{a_{ij}} = sx_i, \qquad (\mathbf{A}\mathbf{x})_j \geq a_{ji}\sqrt{a_{ij}} = s\sqrt{a_{ji}} = sx_j,$$

and the other coordinates satisfy $(\mathbf{A}\mathbf{x})_k \geq 0 = sx_k$. So $\mathbf{A}\mathbf{x} \geq s\mathbf{x}$, and Proposition 19.4.2 (a) gives $\rho(\mathbf{A}) \geq s$.

(c) With $i = 1$, $j = 2$: $\sqrt{4 \cdot 9} = 6$, so $\rho \geq 6$. The diagonal bound of (a) gives only 5. Exactly: the matrix is block diagonal, the 2×2 block has $p(t) = t^2 - 3t - 34$ with roots $\frac{3 \pm \sqrt{145}}{2}$, and the 1×1 block contributes 5. So $\rho = \frac{3+\sqrt{145}}{2} \approx 7.52 \geq 6$.

Exercise. C2: Strict monotonicity

Let $\mathbf{A}, \mathbf{B} \in M_n(\mathbb{R})$ with $\mathbf{O} \leq \mathbf{A} \leq \mathbf{B}$ and $\mathbf{A} \neq \mathbf{B}$.

(a) Prove that if $\mathbf{A}$ is irreducible, then so is $\mathbf{B}$, and $\rho(\mathbf{A}) < \rho(\mathbf{B})$.

(b) Show by an example that the strict inequality can fail if $\mathbf{A}$ is reducible.

Solution.

(a) Since $\mathbf{I} + \mathbf{B} \geq \mathbf{I} + \mathbf{A} \geq \mathbf{O}$ and a nonnegative matrix preserves the order, induction on k gives $(\mathbf{I}+\mathbf{B})^k \geq (\mathbf{I}+\mathbf{A})^k$: indeed $(\mathbf{I}+\mathbf{B})^{k+1} = (\mathbf{I}+\mathbf{B})(\mathbf{I}+\mathbf{B})^k \geq (\mathbf{I}+\mathbf{B})(\mathbf{I}+\mathbf{A})^k \geq (\mathbf{I}+\mathbf{A})^{k+1}$, column by column. With $k = n-1$, $(\mathbf{I}+\mathbf{B})^{n-1} \geq (\mathbf{I}+\mathbf{A})^{n-1} > \mathbf{O}$, so $\mathbf{B}$ is irreducible by Theorem 19.2.6.
Let $\mathbf{v} > \mathbf{0}$ be the Perron vector of $\mathbf{A}$. Then $\mathbf{B}\mathbf{v} = \mathbf{A}\mathbf{v} + (\mathbf{B}-\mathbf{A})\mathbf{v} \geq \mathbf{A}\mathbf{v} = \rho(\mathbf{A})\mathbf{v}$. Moreover $\mathbf{B} - \mathbf{A} \geq \mathbf{O}$ has some entry $b_{ij} - a_{ij} > 0$, and then $\big((\mathbf{B}-\mathbf{A})\mathbf{v}\big)_i \geq (b_{ij} - a_{ij})v_j > 0$, so $\mathbf{B}\mathbf{v} \neq \rho(\mathbf{A})\mathbf{v}$. By Proposition 19.4.2 (b) applied to the irreducible $\mathbf{B}$, $\rho(\mathbf{B}) > \rho(\mathbf{A})$.

(b) $\mathbf{A} = \operatorname{diag}(1, 0)$ and $\mathbf{B} = \operatorname{diag}(1, \frac{1}{2})$ satisfy $\mathbf{O} \leq \mathbf{A} \leq \mathbf{B}$, $\mathbf{A} \neq \mathbf{B}$, and $\rho(\mathbf{A}) = \rho(\mathbf{B}) = 1$.

Exercise. C3: The minimum need not be attained

Let $\mathbf{A} = \left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$.

(a) Show that $\mathbf{A}$ is reducible and $\rho(\mathbf{A}) = 1$.

(b) Show that $\overline{r}_{\mathbf{A}}(\mathbf{x}) > 1$ for every $\mathbf{x} > \mathbf{0}$, so that $\min_{\mathbf{x}>\mathbf{0}} \overline{r}_{\mathbf{A}}(\mathbf{x})$ does not exist.

(c) Show that $\overline{r}_{\mathbf{A}}(\mathbf{x})$ nevertheless comes arbitrarily close to 1, and that $\max_{\mathbf{x}\geq\mathbf{0},\mathbf{x}\neq\mathbf{0}} \underline{r}_{\mathbf{A}}(\mathbf{x}) = 1$ is attained.

Solution.

(a) $\mathbf{I} + \mathbf{A} = \left(\begin{smallmatrix} 2 & 1 \\ 0 & 2 \end{smallmatrix}\right)$ has a zero entry, so $\mathbf{A}$ is reducible (Theorem 19.2.6 with $n = 2$). It is triangular with diagonal $1, 1$, so $\rho(\mathbf{A}) = 1$ (Theorem 9.7.4 (b)).

(b) For $\mathbf{x} = (x_1, x_2) > \mathbf{0}$, $\mathbf{A}\mathbf{x} = (x_1 + x_2,\ x_2)$, with ratios $1 + x_2/x_1$ and 1. Hence $\overline{r}_{\mathbf{A}}(\mathbf{x}) = 1 + x_2/x_1 > 1$. Suppose some $\mathbf{x} > \mathbf{0}$ gave the smallest value $m = 1 + x_2/x_1$. Then $\mathbf{x}' = (x_1, x_2/2) > \mathbf{0}$ gives $\overline{r}_{\mathbf{A}}(\mathbf{x}') = 1 + x_2/(2x_1) < m$, a contradiction. So there is no minimum.

(c) For $\mathbf{x} = (1, t)$ with $t > 0$, $\overline{r}_{\mathbf{A}}(\mathbf{x}) = 1 + t$, which is as close to 1 as we like. For the lower function, $\mathbf{x} = \mathbf{e}_1$ gives $\mathbf{A}\mathbf{e}_1 = \mathbf{e}_1$, so $\underline{r}_{\mathbf{A}}(\mathbf{e}_1) = 1 = \rho(\mathbf{A})$; since every value of $\underline{r}_{\mathbf{A}}$ is $\leq \rho(\mathbf{A})$ by Proposition 19.4.2 (a), the maximum is attained at $\mathbf{e}_1$, which is the nonnegative eigenvector of Theorem 19.3.1.

19.5 Primitive Matrices

The Perron–Frobenius theorem tells us what an irreducible nonnegative matrix looks like at its spectral radius: $\rho(\mathbf{A})$ is an algebraically simple eigenvalue with a positive eigenvector. It says nothing about what the powers $\mathbf{A}^m$ do, and Chapter 9 §11 already met the obstacle. The swap $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ is irreducible, has a positive eigenvector for $\rho = 1$, and its powers alternate forever. This section finds the extra hypothesis that makes the powers settle down, proves that under it $(\mathbf{A}/\rho)^m$ converges to an explicit rank-one matrix, and explains what goes wrong without it: the eigenvalues of modulus ρ are then spread evenly around a circle, and their number is read off the zero pattern of $\mathbf{A}$.

Throughout, matrices are real and $\mathbf{A} \geq \mathbf{O}$ is meant entrywise, as are $\mathbf{A} > \mathbf{O}$, $\mathbf{x} \geq \mathbf{0}$ and $\mathbf{x} > \mathbf{0}$. The spectrum is taken over $\mathbb{C}$, and $\rho(\mathbf{A})$ is the spectral radius (Definition 16.4.1).

19.5.1 Walks in a nonnegative matrix

Whether a power $\mathbf{A}^m$ has a positive entry at position (i, j) depends only on which entries of $\mathbf{A}$ are zero. The bookkeeping device is the directed graph $G(\mathbf{A})$ of §02 (Definition 19.2.1): it has an edge $j \to i$ for each nonzero entry a_{ij}, and a walk of length m from j to i is a list $j = i_0, i_1, \ldots, i_m = i$ of vertices in which each $i_{t-1} \to i_t$ is an edge. The edge for a_{ij} points from the column index to the row index, the direction in which a column-stochastic chain moves. A walk is **closed** if $i_0 = i_m$ and $m \geq 1$.

By §02's count of walks (Lemma 19.2.1 (a)), for $\mathbf{A} \geq \mathbf{O}$ and $m \geq 0$, $(\mathbf{A}^m)_{ij} > 0$ if and only if $G(\mathbf{A})$ has a walk of length m from j to i. Everything in this section about zero patterns rests on that fact.

19.5.2 Primitive matrices

Here is the property the swap lacks. Its powers are $\mathbf{I}$ and the swap itself, and each of them has zero entries; no power ever mixes every state with every other.

A primitive matrix is one that, after enough steps, connects every index to every index at once.

Definition 19.5.1. Primitive Matrix

A matrix $\mathbf{A} \in M_n(\mathbb{R})$, $n \geq 1$, with $\mathbf{A} \geq \mathbf{O}$ is **primitive** if $\mathbf{A}^k > \mathbf{O}$ for **some** integer $k \geq 1$, that is, if **every** entry of some power of $\mathbf{A}$ is **strictly** positive.

In words: we are allowed to wait, but only for a finite number k of steps, and after those k steps all n^2 entries must be positive together. By Lemma 19.2.1 (a), $\mathbf{A}$ is primitive exactly when there is one length k such that every vertex of $G(\mathbf{A})$ has a walk of length k to every vertex, including to itself.

Examples.

- **Positive matrices.** If $\mathbf{A} > \mathbf{O}$, take $k = 1$. Perron's theorem (Theorem 19.1.5) is about these.
- **A matrix with zeros.** $\mathbf{A} = \begin{pmatrix} 1 & 1 \\ 2 & 0 \end{pmatrix}$ has a zero entry, but $\mathbf{A}^2 = \begin{pmatrix} 3 & 1 \\ 2 & 2 \end{pmatrix} > \mathbf{O}$. So $\mathbf{A}$ is primitive with $k = 2$. In the graph: the edges are $1 \to 1$, $1 \to 2$ and $2 \to 1$, and $1 \to 1 \to 1$, $1 \to 1 \to 2$, $2 \to 1 \to 1$, $2 \to 1 \to 2$ are walks of length 2 between every ordered pair.
- **A slower one.** $\mathbf{W} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}$. Its edges are $2 \to 1$, $3 \to 2$, $1 \to 3$, and $2 \to 3$, so its closed walks include $1 \to 3 \to 2 \to 1$ of length 3 and $2 \to 3 \to 2$ of length 2. Multiplying out, $\mathbf{W}^4 = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 2 & 1 \end{pmatrix}$ still has a zero, and $\mathbf{W}^5 = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 2 & 2 \end{pmatrix} > \mathbf{O}$. So $\mathbf{W}$ is primitive, and the smallest k is 5.
- **Degenerate case.** A 1×1 matrix (a) is primitive if and only if $a > 0$. The zero 1×1 matrix is not, which is why the statements below exclude it.

Nonexample by minimal change. Change the $(1,1)$-entry of $\begin{pmatrix} 1 & 1 \\ 2 & 0 \end{pmatrix}$ to 0. The result $\mathbf{A}' = \begin{pmatrix} 0 & 1 \\ 2 & 0 \end{pmatrix}$ is still nonnegative, and it is still irreducible, since $(\mathbf{I}+\mathbf{A}')^1 > \mathbf{O}$ (Theorem 19.2.6). But $\mathbf{A}'^2 = 2\mathbf{I}$, so the even powers of $\mathbf{A}'$ are multiples of $\mathbf{I}$ and the odd powers multiples of $\mathbf{A}'$, and every power has two zero entries: $\mathbf{A}'$ is not primitive. The clause that fails is "some power is positive"; what was lost is the loop $1 \to 1$; without it every edge joins 1 and 2, so every closed walk has even length. More generally, the powers of a permutation matrix are permutation matrices (Lemma 3.4.7), and for $n \geq 2$ a permutation matrix has zeros; so no permutation matrix of size $n \geq 2$ is primitive.

Why this definition. Asking for $\mathbf{A} > \mathbf{O}$ itself would be too strong: it excludes the matrices above, whose powers behave exactly like those of a positive matrix. Asking only that each position (i, j) be positive in **some** power, possibly a different power for each position, is too weak: that is irreducibility, and the swap satisfies it. Primitivity sits between the two. It asks for a **single** power that is positive everywhere.

Warning. **Irreducible does not mean primitive.** Irreducibility asks that each j be reachable from each i by a walk of **some** length; primitivity asks for walks of **one common** length. For the swap, every walk from 1 to 2 has odd length and every walk from 1 to 1 has even length, so no single power is positive at both $(1,1)$ and $(1,2)$.

Quick check. Is $\mathbf{A} = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}$ primitive? Is $\mathbf{B} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$?

Solution. $\mathbf{A}^2 = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix} > \mathbf{O}$, so $\mathbf{A}$ is primitive. For $\mathbf{B}$, every power is upper triangular, $\mathbf{B}^m = \begin{pmatrix} 1 & m \\ 0 & 1 \end{pmatrix}$, so the $(2,1)$-entry is always 0 and $\mathbf{B}$ is not primitive: $G(\mathbf{B})$ has no walk from 1 to 2 at all.

The first consequences of the definition come straight from the graph and from Perron's theorem applied to the positive power.

Proposition 19.5.1. First Properties of Primitive Matrices

Let $\mathbf{A} \in M_n(\mathbb{R})$ be primitive, with $\mathbf{A}^k > \mathbf{O}$. Then:

(a) $\mathbf{A}^m > \mathbf{O}$ for **every** $m \geq k$;

(b) $\mathbf{A}$ is irreducible;

(c) $\rho(\mathbf{A}) > 0$, and $\rho(\mathbf{A})$ is the **only** eigenvalue of $\mathbf{A}$ of modulus $\rho(\mathbf{A})$.

Idea.

(a) Multiplying a positive matrix by a nonnegative matrix with no zero column keeps it positive. (b) A positive power gives walks between all vertices, which is strong connectivity. (c) The eigenvalues of $\mathbf{A}^k$ are the k-th powers of those of $\mathbf{A}$. A second eigenvalue of $\mathbf{A}$ on the circle of radius ρ would land, after raising to the k-th power, on the circle where Perron's theorem allows only the simple eigenvalue $\rho(\mathbf{A}^k)$.

Proof.

(a) Column j of $\mathbf{A}^k$ is $\mathbf{A}^{k-1}$ times column j of $\mathbf{A}$ (Theorem 0.9.2; for $k = 1$ read $\mathbf{A}^0 = \mathbf{I}$). Since $\mathbf{A}^k > \mathbf{O}$, no column of $\mathbf{A}$ is zero. Now suppose $\mathbf{A}^m > \mathbf{O}$ for some $m \geq k$. Then $(\mathbf{A}^{m+1})_{ij} = \sum_l (\mathbf{A}^m)_{il} a_{lj}$ is a sum of nonnegative terms, and the term for an l with $a_{lj} > 0$ is positive. So $\mathbf{A}^{m+1} > \mathbf{O}$, and (a) follows by induction on m.

(b) For every ordered pair (j,i), $(\mathbf{A}^k)_{ij} > 0$, so Lemma 19.2.1 (a) gives a walk from j to i in $G(\mathbf{A})$. So $G(\mathbf{A})$ is strongly connected (Definition 19.2.2), and $\mathbf{A}$ is irreducible by Theorem 19.2.5.

(c) Over $\mathbb{C}$, write $p_{\mathbf{A}} = (x - \lambda_1) \cdots (x - \lambda_n)$, with each eigenvalue repeated according to

its algebraic multiplicity. By the Spectral Mapping Theorem (Theorem 9.10.1) with $q = x^k$,

$$p_{\mathbf{A}^k} = (x - \lambda_1^k) \cdots (x - \lambda_n^k).$$

So $\rho(\mathbf{A}^k) = \max_i |\lambda_i|^k = \rho(\mathbf{A})^k$. By Theorem 19.1.5 applied to $\mathbf{A}^k > \mathbf{O}$, $\rho(\mathbf{A}^k) > 0$ is an algebraically simple eigenvalue of $\mathbf{A}^k$, and every other eigenvalue of $\mathbf{A}^k$ has modulus less than $\rho(\mathbf{A}^k)$. In particular $\rho(\mathbf{A})^k > 0$, so $\rho(\mathbf{A}) > 0$.

By (b) and Theorem 19.3.2, $\rho := \rho(\mathbf{A})$ is an eigenvalue of $\mathbf{A}$; say $\lambda_1 = \rho$. Suppose some λ_s with $s \geq 2$ had $|\lambda_s| = \rho$ but $\lambda_s \neq \rho$. Then λ_s^k is an eigenvalue of $\mathbf{A}^k$ of modulus $\rho^k = \rho(\mathbf{A}^k)$, so by Perron's theorem it is not an "other" eigenvalue: $\lambda_s^k = \rho^k$. Then ρ^k occurs at least twice in the list $\lambda_1^k, \ldots, \lambda_n^k$, at positions 1 and s, so $a_{\mathbf{A}^k}(\rho^k) \geq 2$, contradicting simplicity. Hence ρ is the only eigenvalue of modulus ρ.

■

Part (c) is the whole spectral content of primitivity, and the rest of the section shows that it is also enough. Part (a) says that once the powers become positive they stay positive, so the least such k, the **exponent** of $\mathbf{A}$, is well defined and marks a permanent change.

19.5.3 The limit theorem

For the swap, $\rho = 1$ and the powers do not converge. We now show that primitivity is exactly what makes $(\mathbf{A}/\rho)^m$ converge, and we identify the limit. The answer involves eigenvectors on both sides.

If $\mathbf{A} \geq \mathbf{O}$ is irreducible, then Theorem 19.3.2 (a) and Corollary 19.3.3 (a)–(c) provide $\mathbf{v} > \mathbf{0}$ and $\mathbf{w} > \mathbf{0}$ with $\mathbf{A}\mathbf{v} = \rho(\mathbf{A})\mathbf{v}$, $\mathbf{w}^\top \mathbf{A} = \rho(\mathbf{A})\mathbf{w}^\top$, and $\mathbf{w}^\top \mathbf{v} > 0$, and the right and left Perron vectors of Definition 19.3.1 are the pair scaled so that $\mathbf{1}^\top \mathbf{v} = 1$ and $\mathbf{w}^\top \mathbf{v} = 1$. Each of $\mathbf{v}, \mathbf{w}$ is determined up to a positive factor, because $\rho(\mathbf{A})$ is algebraically simple (Theorem 19.3.2 (b)), so its eigenspaces for $\mathbf{A}$ and $\mathbf{A}^\top$ are lines (Theorem 9.3.1, Proposition 9.10.5 (b)).

Theorem 19.5.2. Limit Theorem for Primitive Matrices

Let $\mathbf{A} \in M_n(\mathbb{R})$ be primitive, $\rho = \rho(\mathbf{A})$, and let $\mathbf{v} > \mathbf{0}$ and $\mathbf{w} > \mathbf{0}$ satisfy $\mathbf{A}\mathbf{v} = \rho\mathbf{v}$ and $\mathbf{w}^\top \mathbf{A} = \rho\mathbf{w}^\top$, for instance the right and left Perron vectors of $\mathbf{A}$. Then, entrywise,

$$\left(\frac{1}{\rho}\mathbf{A}\right)^m \longrightarrow \frac{\mathbf{v}\mathbf{w}^\top}{\mathbf{w}^\top \mathbf{v}} \qquad (m \to \infty).$$

The limit is a positive matrix of rank one, and it does not depend on how $\mathbf{v}$ and $\mathbf{w}$ are scaled. The same conclusion holds for every irreducible $\mathbf{A} \geq \mathbf{O}$ with $\rho(\mathbf{A}) > 0$ whose only eigenvalue of modulus $\rho(\mathbf{A})$ is $\rho(\mathbf{A})$ itself.

Idea. Divide by ρ, so that the Perron eigenvalue becomes 1 and every other eigenvalue lies strictly inside the unit circle. Chapter 10 §10 then says the powers converge, with no diagonalizability needed, because it works with the Jordan form. It remains to name the limit Ł. Three identities pin it down: $\mathbf{B}Ł = Ł$ puts every column of Ł on the line through $\mathbf{v}$; $Ł\mathbf{B} = Ł$ puts every row on the line through $\mathbf{w}^\top$; and $Ł\mathbf{v} = \mathbf{v}$ fixes the scale.

Proof. It suffices to prove the last sentence, since a primitive matrix is irreducible with $\rho(\mathbf{A}) > 0$ and has $\rho(\mathbf{A})$ as its only eigenvalue of modulus $\rho(\mathbf{A})$, by Proposition 19.5.1 (b) and (c).

So let $\mathbf{A} \geq \mathbf{O}$ be irreducible with $\rho > 0$ the only eigenvalue of modulus ρ, and put $\mathbf{B} = \rho^{-1}\mathbf{A}$. By Theorem 9.10.1 with $q = x/\rho$, the eigenvalues of $\mathbf{B}$ are the numbers λ/ρ, $\lambda \in \operatorname{spec}(\mathbf{A})$, with the same algebraic multiplicities. So 1 is an eigenvalue of $\mathbf{B}$ with $a_{\mathbf{B}}(1) = a_{\mathbf{A}}(\rho) = 1$ by Theorem 19.3.2, and every other eigenvalue of $\mathbf{B}$ has modulus less than 1. Since the 1-blocks in a Jordan form of $\mathbf{B}$ have total size $a_{\mathbf{B}}(1) = 1$ (Corollary 10.3.2 (a)), there is a single 1-block, of size 1. By Theorem 10.10.4, $\mathbf{B}^m$ converges to some $Ł \in M_n(\mathbb{C})$.

Three identities. By the algebra of limits, (B3) of Chapter 10 §10, applied entry by entry, $\mathbf{B}^{m+1} = \mathbf{B}\,\mathbf{B}^m \to \mathbf{B}Ł$ and $\mathbf{B}^{m+1} = \mathbf{B}^m\mathbf{B} \to Ł\mathbf{B}$; but $(\mathbf{B}^{m+1})$ is a tail of $(\mathbf{B}^m)$, so it also tends to Ł. Hence $\mathbf{B}Ł = Ł = Ł\mathbf{B}$. Also $\mathbf{B}\mathbf{v} = \mathbf{v}$, so $\mathbf{B}^m\mathbf{v} = \mathbf{v}$ for every m, and letting $m \to \infty$ gives $Ł\mathbf{v} = \mathbf{v}$.

The columns. $\mathbf{B}Ł = Ł$ says that every column of Ł lies in $E_1(\mathbf{B}) = E_\rho(\mathbf{A})$, which is the line $\operatorname{span}(\mathbf{v})$, as noted before the theorem. So column j of Ł is $u_j\mathbf{v}$ for some $u_j \in \mathbb{C}$, that is, $Ł = \mathbf{v}\mathbf{u}^\top$ with $\mathbf{u} = (u_1, \dots, u_n)$.

The rows. $Ł\mathbf{B} = Ł$ reads $\mathbf{v}(\mathbf{u}^\top\mathbf{B}) = \mathbf{v}\mathbf{u}^\top$. Choose r with $v_r \neq 0$ and compare row r: $\mathbf{u}^\top\mathbf{B} = \mathbf{u}^\top$, so $\mathbf{B}^\top\mathbf{u} = \mathbf{u}$, and $\mathbf{u}$ lies in the eigenspace of $\mathbf{A}^\top$ for ρ. That eigenspace is the line through $\mathbf{w}$, so $\mathbf{u} = c\mathbf{w}$ for some $c \in \mathbb{C}$.

The scale. Now $\mathbf{v} = Ł\mathbf{v} = c\,\mathbf{v}(\mathbf{w}^\top\mathbf{v})$, and comparing entry r gives $c\,\mathbf{w}^\top\mathbf{v} = 1$. Since $\mathbf{w}^\top\mathbf{v} > 0$, $c = 1/(\mathbf{w}^\top\mathbf{v})$, and $Ł = \mathbf{v}\mathbf{w}^\top/(\mathbf{w}^\top\mathbf{v})$.

Finally, $\mathbf{v}\mathbf{w}^\top$ has (i,j)-entry $v_i w_j > 0$ and every column a multiple of $\mathbf{v}$, so the limit is positive of rank one; replacing $\mathbf{v}, \mathbf{w}$ by $s\mathbf{v}, t\mathbf{w}$ with $s, t > 0$ multiplies numerator and denominator by st. This proves the theorem.

■

Note what the proof did **not** use: no eigenvector basis, no computation of the other eigenvalues. Chapter 10's Jordan-form criterion did the analysis once and for all; Perron–Frobenius supplied the simple eigenvalue and the two positive eigenvectors. With the normalization $\mathbf{w}^\top\mathbf{v} = 1$ of §03, the limit is simply the idempotent $\mathbf{v}\mathbf{w}^\top$. §06 turns the theorem into the long-run theorem for Markov chains.

Example. A Limit Computed Two Ways

Let $\mathbf{A} = \begin{pmatrix} 1 & 1 \\ 2 & 0 \end{pmatrix}$, which is primitive since $\mathbf{A}^2 > \mathbf{O}$. Find $\lim_m (\mathbf{A}/\rho)^m$ from Theorem 19.5.2, and confirm it with an exact formula for $\mathbf{A}^m$.

Solution. The theorem. $p_{\mathbf{A}} = x^2 - x - 2 = (x-2)(x+1)$, so $\rho = 2$. A positive right eigenvector solves $(\mathbf{A} - 2\mathbf{I})\mathbf{v} = \mathbf{0}$, that is, $-v_1 + v_2 = 0$: take $\mathbf{v} = (1, 1)$. A positive left eigenvector solves $(\mathbf{A}^\top - 2\mathbf{I})\mathbf{w} = \mathbf{0}$, with $\mathbf{A}^\top - 2\mathbf{I} = \begin{pmatrix} -1 & 2 \\ 1 & -2 \end{pmatrix}$: take $\mathbf{w} = (2, 1)$. Then

$\mathbf{w}^\top\mathbf{v} = 3$, and

$$\left(\tfrac{1}{2}\mathbf{A}\right)^m \longrightarrow Ł = \frac{1}{3}\begin{pmatrix}1\\1\end{pmatrix}\begin{pmatrix}2 & 1\end{pmatrix} = \frac{1}{3}\begin{pmatrix}2 & 1\\2 & 1\end{pmatrix}.$$

The formula. $Ł^2 = Ł$, because $\mathbf{w}^\top\mathbf{v}/3 = 1$, and $Ł(\mathbf{I} - Ł) = \mathbf{O}$. A direct check gives $3Ł - \mathbf{I} = \begin{pmatrix}1 & 1\\2 & 0\end{pmatrix} = \mathbf{A}$, so $\mathbf{A} = 2Ł - (\mathbf{I} - Ł)$. Since $Ł$ and $\mathbf{I} - Ł$ are complementary projections, induction on m gives

$$\mathbf{A}^m = 2^m Ł + (-1)^m(\mathbf{I} - Ł),$$
$$\left(\tfrac{1}{2}\mathbf{A}\right)^m = Ł + \left(-\tfrac{1}{2}\right)^m \frac{1}{3}\begin{pmatrix}1 & -1\\-2 & 2\end{pmatrix}.$$

For $m = 2$ this is $\frac{1}{3}\begin{pmatrix}2 & 1\\2 & 1\end{pmatrix} + \frac{1}{12}\begin{pmatrix}1 & -1\\-2 & 2\end{pmatrix} = \frac{1}{4}\begin{pmatrix}3 & 1\\2 & 2\end{pmatrix}$, which is $\mathbf{A}^2/4$. The error term is $(-\frac{1}{2})^m$ times a fixed matrix: it decays at the rate $|\lambda_2|/\rho = \frac{1}{2}$ and changes sign at every step.

Warning. **The limit uses the left eigenvector, not the transpose of the right one.** For the matrix above, $\mathbf{v}\mathbf{v}^\top/(\mathbf{v}^\top\mathbf{v}) = \frac{1}{2}\begin{pmatrix}1 & 1\\1 & 1\end{pmatrix}$, which is not the limit. The two agree when $\mathbf{A}$ is symmetric, since then $\mathbf{w} = \mathbf{v}$; in general $\mathbf{w}$ carries the information about how much each starting index contributes in the long run.

The limit theorem also gives the converse of Proposition 19.5.1 (c).

Theorem 19.5.3. Primitivity from the Spectrum

Let $\mathbf{A} \in M_n(\mathbb{R})$, $\mathbf{A} \geq \mathbf{O}$, be irreducible with $\rho(\mathbf{A}) > 0$. Then $\mathbf{A}$ is primitive if and only if $\rho(\mathbf{A})$ is the only eigenvalue of $\mathbf{A}$ of modulus $\rho(\mathbf{A})$.

Idea. One direction is already proved. For the other, the limit theorem gives a **positive** limit, and a sequence that converges to a positive number is eventually positive. There are only n^2 entries to wait for, so some single power is positive everywhere.

Proof. ($\Rightarrow$) This is Proposition 19.5.1 (c).

($\Leftarrow$) By the last sentence of Theorem 19.5.2, $(\mathbf{A}/\rho)^m \to Ł$ with $Ł > \mathbf{O}$. Each of the n^2 entries of $(\mathbf{A}/\rho)^m$ converges to a positive number c_{ij}, so it is positive for all m beyond some m_{ij}: this is the definition of the limit with $\varepsilon = c_{ij}/2$. For m larger than the largest of these finitely many m_{ij}, $(\mathbf{A}/\rho)^m > \mathbf{O}$, and hence $\mathbf{A}^m = \rho^m(\mathbf{A}/\rho)^m > \mathbf{O}$. So $\mathbf{A}$ is primitive.

■

The hypothesis $\rho(\mathbf{A}) > 0$ excludes only the 1×1 zero matrix: for $n \geq 2$ it holds for every irreducible $\mathbf{A} \geq \mathbf{O}$ by Theorem 19.3.2 (d), and for $n = 1$, $\rho((a)) = a$.

19.5.4 The peripheral spectrum and the period

What does an irreducible, nonprimitive matrix look like? The eigenvalues of modulus $\rho(\mathbf{A})$ are called the **peripheral eigenvalues** of $\mathbf{A}$, and the theorem just proved says that a nonprimitive irreducible matrix, other than the 1×1 zero matrix, has more than one of them. For the swap they are ± 1. For the 3×3 cyclic permutation they are the three cube roots of unity, as we shall compute. The pattern is general, and it is controlled by a number read off the walks.

The period is the common step size of all round trips.

Definition 19.5.2. Period

Let $\mathbf{A} \in M_n(\mathbb{R})$, $\mathbf{A} \geq \mathbf{O}$, be irreducible and not the 1×1 zero matrix. The **period** of $\mathbf{A}$ is the greatest common divisor of the lengths of all closed walks in $\mathbf{A}$.

There is at least one closed walk, so the definition makes sense. If $n = 1$, then $\mathbf{A} = (a)$ with $a > 0$, and the loop $1 \to 1$ is a closed walk of length 1. If $n \geq 2$, then $G(\mathbf{A})$ is strongly connected (Theorem 19.2.5), so there are walks from 1 to 2 and from 2 to 1, and one followed by the other is a closed walk. For the swap, every edge joins 1 and 2, so closed walks have even length, and $1 \to 2 \to 1$ has length 2: the period is 2. For $\mathbf{W}$ above, closed walks of lengths 3 and 2 give period 1. A single positive diagonal entry a_{ii} is a loop $i \to i$, a closed walk of length 1, and forces period 1.

The main theorem of this part says that the period counts the peripheral eigenvalues, and that they sit at the corners of a regular polygon. As in §01, $|\mathbf{z}| = (|z_1|, \dots, |z_n|)$ is the entrywise absolute value of $\mathbf{z} \in \mathbb{C}^n$.

Theorem 19.5.4. The Peripheral Spectrum

Let $\mathbf{A} \in M_n(\mathbb{R})$, $\mathbf{A} \geq \mathbf{O}$, be irreducible and not the 1×1 zero matrix. Put $\rho = \rho(\mathbf{A})$, let h be the number of **distinct** eigenvalues of $\mathbf{A}$ of modulus ρ, and let $\omega = e^{2\pi i/h}$. Then:

(a) the eigenvalues of $\mathbf{A}$ of modulus ρ are exactly $\rho, \rho\omega, \rho\omega^2, \dots, \rho\omega^{h-1}$, and each is algebraically simple;

(b) $\mathbf{A}$ is similar to $\omega\mathbf{A}$; in particular $\operatorname{spec}(\mathbf{A})$ is unchanged by multiplication by ω, with algebraic multiplicities;

(c) h equals the period of $\mathbf{A}$.

Idea. **Step roadmap.** ① Take an eigenvector $\mathbf{z}$ for a peripheral eigenvalue $\rho e^{i\theta}$. The triangle inequality (Theorem 0.6.4) gives $\mathbf{A}|\mathbf{z}| \geq \rho|\mathbf{z}|$, and superinvariance (Proposition 19.4.2) forces equality, so $|\mathbf{z}|$ is a positive multiple of $\mathbf{v}$. Equality in the triangle inequality then says all terms of each row sum point in the same direction, and that is a similarity $\mathbf{D}^{-1}\mathbf{A}\mathbf{D} = e^{i\theta}\mathbf{A}$ with $\mathbf{D}$ diagonal. ② So the spectrum is invariant under rotation by every peripheral angle, the peripheral points form a finite group of rotations, and a finite group of h points on the unit circle must be the h-th roots of unity. ③ The diagonal matrix $\mathbf{D}$ records a phase at each vertex, and each edge advances the phase by one step; going round a closed walk returns the phase, which ties h to walk lengths.

Proof. Recall that $\rho > 0$ (the remark after Theorem 19.5.3), and let $\mathbf{v} > \mathbf{0}$ be an eigenvector of $\mathbf{A}$ for ρ (Theorem 19.3.2 (a)).

Step 1. *Let $\lambda = \rho e^{i\theta}$ be an eigenvalue. Then there are real numbers $\varphi_1, \dots, \varphi_n$ such that $e^{i\varphi_j} = e^{i\theta} e^{i\varphi_i}$ whenever $a_{ij} > 0$; consequently, with $\mathbf{D} = \operatorname{diag}(e^{i\varphi_1}, \dots, e^{i\varphi_n})$, we have $\mathbf{D}^{-1}\mathbf{A}\mathbf{D} = e^{i\theta}\mathbf{A}$.*

Let $\mathbf{A}\mathbf{z} = \lambda\mathbf{z}$ with $\mathbf{z} \neq \mathbf{0}$. By Lemma 19.1.2 (a), with $|\mathbf{A}| = \mathbf{A}$ since $\mathbf{A} \geq \mathbf{O}$, in each row

$$\rho\,|z_i| = |\lambda z_i| = \Big|\sum_j a_{ij} z_j\Big| \leq \sum_j a_{ij}|z_j| = (\mathbf{A}|\mathbf{z}|)_i.$$

So $\mathbf{A}|\mathbf{z}| \geq \rho|\mathbf{z}|$, where $|\mathbf{z}| \geq \mathbf{0}$ and $|\mathbf{z}| \neq \mathbf{0}$. If $\mathbf{A}|\mathbf{z}| \neq \rho|\mathbf{z}|$, then Proposition 19.4.2 (b) with $s = \rho$ would give $\rho > \rho$; hence $\mathbf{A}|\mathbf{z}| = \rho|\mathbf{z}|$. So $|\mathbf{z}|$ is a nonnegative eigenvector of $\mathbf{A}$, and by Theorem 19.3.2 (c) it is a positive multiple of $\mathbf{v}$. Therefore $|\mathbf{z}| > \mathbf{0}$, and we may write $z_j = |z_j| e^{i\varphi_j}$ in polar form.

Now every inequality in the display is an equality. Fix i. The number $S = \sum_j a_{ij} z_j = \lambda z_i = \rho|z_i| e^{i(\theta+\varphi_i)}$ (Theorem 0.6.5) is nonzero, since $\rho > 0$ and $|z_i| > 0$, and $|S| = \sum_j |a_{ij} z_j|$. So Lemma 19.1.3, applied to the n numbers $a_{ij} z_j$, with $S/|S| = e^{i(\theta+\varphi_i)}$, gives $a_{ij} z_j = a_{ij}|z_j| e^{i(\theta+\varphi_i)}$ for every j. If $a_{ij} > 0$, dividing by $a_{ij}|z_j| > 0$ gives $e^{i\varphi_j} = e^{i(\theta+\varphi_i)} = e^{i\theta} e^{i\varphi_i}$. This is the first statement. For the second, the (i,j)-entry of $\mathbf{D}^{-1}\mathbf{A}\mathbf{D}$ is $e^{-i\varphi_i} a_{ij} e^{i\varphi_j}$, which is $e^{i\theta} a_{ij}$ when $a_{ij} > 0$ and $0 = e^{i\theta} a_{ij}$ when $a_{ij} = 0$.

Step 2. *Proof of (a) and (b).* Write $p_{\mathbf{A}} = (x - \lambda_1) \cdots (x - \lambda_n)$ over $\mathbb{C}$. For every peripheral $\lambda = \rho e^{i\theta}$, Step 1 and Theorem 7.10.3 give $p_{\mathbf{A}} = p_{e^{i\theta}\mathbf{A}}$, and by Theorem 9.10.1 with $q = e^{i\theta} x$,

$$\begin{aligned}(x - \lambda_1) \cdots (x - \lambda_n) = p_{\mathbf{A}} &= p_{e^{i\theta}\mathbf{A}} \\ &= (x - e^{i\theta}\lambda_1) \cdots (x - e^{i\theta}\lambda_n).\end{aligned}$$

So multiplication by $e^{i\theta}$ permutes the eigenvalues listed with multiplicity:

$$a_{\mathbf{A}}(e^{i\theta}\mu) = a_{\mathbf{A}}(\mu) \qquad \text{for every } \mu \in \mathbb{C}.$$

Let $G = \{\lambda/\rho : \lambda \in \operatorname{spec}(\mathbf{A}),\ |\lambda| = \rho\}$, a set of h complex numbers of modulus 1 containing 1, because $\rho \in \operatorname{spec}(\mathbf{A})$ by Theorem 19.3.2. If $g = e^{i\theta}$ and g' lie in G, then $\rho g' \in \operatorname{spec}(\mathbf{A})$, so $g\rho g' \in \operatorname{spec}(\mathbf{A})$ by the invariance just proved, and it has modulus ρ: thus $gg' \in G$. For fixed $g \in G$, the map $G \to G$, $x \mapsto gx$, is injective, since $g \neq 0$; as G is finite it is a bijection. Hence the product Π of all elements of G satisfies $\Pi = \prod_{x \in G} gx = g^h \Pi$, and $\Pi \neq 0$, so $g^h = 1$. So G is contained in the set of h-th roots of unity, which has exactly h elements $1, \omega, \dots, \omega^{h-1}$ (Example (Roots of Unity)). Both sets have h elements, so they are equal. This proves the first half of (a).

Each $\rho\omega^j$ has $a_{\mathbf{A}}(\rho\omega^j) = a_{\mathbf{A}}(\rho) = 1$, by the invariance with $e^{i\theta} = \omega^j$ and Theorem 19.3.2. This proves (a). Taking $\lambda = \rho\omega$ in Step 1 gives (b).

Step 3. *Proof of (c).* Let d be the period.

h divides d. Apply Step 1 to $\lambda = \rho\omega$, so $e^{i\theta} = \omega$. Along a closed walk $i_0 \to i_1 \to \cdots \to i_L = i_0$ each entry $a_{i_t i_{t-1}}$ is positive, so Step 1 gives $e^{i\varphi_{i_{t-1}}} = \omega e^{i\varphi_{i_t}}$ for each t, and chaining these L equations gives $e^{i\varphi_{i_0}} = \omega^L e^{i\varphi_{i_L}} = \omega^L e^{i\varphi_{i_0}}$. So $\omega^L = 1$, and h divides L (Example (Roots of Unity)). As this holds for every closed walk, h divides d.

d divides h. If $n = 1$ then $d = 1$. Let $n \geq 2$. Since $G(\mathbf{A})$ is strongly connected (Theorem 19.2.5), every vertex j is joined to vertex 1 by walks in both directions; for $j = 1$ this includes the walk of length 0. Choose for each j the length ℓ_j of some walk from 1 to j, with $\ell_1 = 0$. Any two walks from 1 to j, of lengths ℓ and ℓ', have lengths congruent modulo d: following either by a fixed walk from j back to 1, of length s, gives closed walks of lengths $\ell + s$ and $\ell' + s$ (or length 0, trivially a multiple of d), both multiples of d. Now if $a_{ij} > 0$, a walk from 1 to j followed by the edge $j \to i$ is a walk from 1 to i of length $\ell_j + 1$, so $\ell_i \equiv \ell_j + 1 \pmod{d}$. Put $\zeta = e^{2\pi i/d}$ and $\mathbf{E} = \operatorname{diag}(\zeta^{-\ell_1}, \dots, \zeta^{-\ell_n})$. Then the (i,j)-entry of $\mathbf{E}^{-1}\mathbf{A}\mathbf{E}$ is $\zeta^{\ell_i - \ell_j} a_{ij} = \zeta a_{ij}$ when $a_{ij} > 0$, since $\zeta^d = 1$, and 0 otherwise. So $\mathbf{E}^{-1}\mathbf{A}\mathbf{E} = \zeta\mathbf{A}$. As in Step 2, $\operatorname{spec}(\mathbf{A}) = \zeta \operatorname{spec}(\mathbf{A})$, so $\zeta\rho \in \operatorname{spec}(\mathbf{A})$, with modulus ρ. By (a), $\zeta = \omega^j$ for some j, so $\zeta^h = (\omega^h)^j = 1$, and d divides h: $e^{2\pi i h/d} = 1 = e^{2\pi i \cdot 0/d}$, and two integers k, l give the same number $e^{2\pi i k/d} = e^{2\pi i l/d}$ only when d divides $k - l$ (Example (Roots of Unity)).

Both h and d are positive, so $h = d$. This proves the theorem.

■

The two halves of Step 3 are mirror images. The diagonal similarity attached to a peripheral eigenvalue is a consistent labeling of the indices by h phases, with every positive entry advancing the phase by one; the walks of $\mathbf{A}$ build such a labeling with d phases.

Combining (c) with Theorem 19.5.3 gives a test that needs no eigenvalues at all.

Corollary 19.5.5. Primitivity from Walks

Let $\mathbf{A} \in M_n(\mathbb{R})$, $\mathbf{A} \geq \mathbf{O}$, be irreducible and not the 1×1 zero matrix. Then $\mathbf{A}$ is primitive if and only if its period is 1. In particular, an irreducible $\mathbf{A} \geq \mathbf{O}$ with at least one positive diagonal entry is primitive.

Proof. Since $\mathbf{A}$ is not the 1×1 zero matrix, $\rho(\mathbf{A}) > 0$ by the remark after Theorem 19.5.3. So by that theorem, $\mathbf{A}$ is primitive if and only if $h = 1$, and h is the period by Theorem 19.5.4 (c). A positive diagonal entry is a closed walk of length 1, so the period divides 1.

■

19.5.5 The cyclic permutation

The standard irreducible matrix that is not primitive is the cyclic shift. It generalizes Chapter 9's swap, and it shows every part of Theorem 19.5.4 at once.

Example. The Cyclic Permutation

Let $n \geq 2$ and let $\sigma \in S_n$ send $j \mapsto j + 1$ for $j < n$ and $n \mapsto 1$. Its permutation matrix $\mathbf{P} = \mathbf{P}_\sigma$ (Definition 3.4.2) has columns $\mathbf{e}_2, \mathbf{e}_3, \dots, \mathbf{e}_n, \mathbf{e}_1$. Show that $\mathbf{P}$ is irreducible with period n, find its peripheral eigenvalues, and show that $\mathbf{P}^m$ does not converge.

Solution. Walks. The nonzero entries of $\mathbf{P}$ are $p_{j+1,\,j}$ for $j < n$ and p_{1n}, so the edges of $G(\mathbf{P})$ are $j \to j + 1$ for $j < n$ and $n \to 1$: the graph is the cycle $1 \to 2 \to \cdots \to n \to 1$. Every walk runs round it, and a closed walk goes round a whole number of times, so its length is a

multiple of n; the walk $1 \to 2 \to \cdots \to n \to 1$ has length n. So the period is n. Following the cycle, every vertex reaches every vertex, so $G(\mathbf{P})$ is strongly connected and $\mathbf{P}$ is irreducible (Theorem 19.2.5).

Eigenvalues. Since $\mathbf{P}\mathbf{e}_j = \mathbf{e}_{\sigma(j)}$ and σ^n is the identity, $\mathbf{P}^n\mathbf{e}_j = \mathbf{e}_j$ for every j, so $\mathbf{P}^n = \mathbf{I}$. If $\mathbf{P}\mathbf{z} = \lambda\mathbf{z}$ with $\mathbf{z} \neq \mathbf{0}$, then $\mathbf{z} = \mathbf{P}^n\mathbf{z} = \lambda^n\mathbf{z}$, so $\lambda^n = 1$ and $|\lambda| = 1$. Hence $\rho(\mathbf{P}) = 1$ and **every** eigenvalue is peripheral. By Theorem 19.5.4, $h = n$, so the peripheral eigenvalues are the n numbers $1, \omega, \dots, \omega^{n-1}$, $\omega = e^{2\pi i/n}$, each algebraically simple. That accounts for all n eigenvalues, and $p_{\mathbf{P}} = (x-1)(x-\omega)\cdots(x-\omega^{n-1}) = x^n - 1$, the last equality because both sides are monic of degree n with the same n distinct roots.

Powers. $\mathbf{P}^n = \mathbf{I}$, so the sequence $\mathbf{P}, \mathbf{P}^2, \dots$ repeats with period n, and $\mathbf{P}^m$ runs through n distinct permutation matrices. A sequence that returns to each of two different matrices infinitely often has no limit, so $\mathbf{P}^m$ does not converge. Starting a chain at $\mathbf{e}_1$, $\mathbf{P}^m\mathbf{e}_1$ runs round $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n, \mathbf{e}_1, \dots$ forever, although $\mathbf{P}$ has the steady state $\frac{1}{n}\mathbf{1}$.

For $n = 2$ this is exactly the warning in Chapter 9 §11: the swap has the eigenvalue $-1 = e^{2\pi i/2}$, and Chapter 10 §10 identified an eigenvalue of modulus 1 other than 1 as the only obstruction to convergence (Corollary 10.10.5). We now know where that eigenvalue comes from. It is forced by the period: every closed walk has even length. §06 shows that the chain still settles down **on average**.

19.5.6 How large a power is needed

The definition of primitivity asks for some $\mathbf{A}^k > \mathbf{O}$ but gives no bound on k, so a direct check might seem to require computing powers without end. By Lemma 19.2.1 (a), the zero pattern of $\mathbf{A}^m$ depends only on the zero pattern of $\mathbf{A}$, and there are finitely many patterns, so in principle the question is finite. The sharp answer is classical.

Theorem 19.5.6. Wielandt's Bound

If $\mathbf{A} \in M_n(\mathbb{R})$ is primitive, then $\mathbf{A}^{(n-1)^2+1} > \mathbf{O}$. The bound cannot be improved: for $n \geq 2$ the matrix $\mathbf{W}_n$ with ones at the positions $(i, i+1)$ for $i < n$ and at $(n, 1)$ and $(n, 2)$, and zeros elsewhere, is primitive and $\mathbf{W}_n^{(n-1)^2}$ has a zero entry.

This theorem is due to H. Wielandt (1950). **We state it without proof**; the proof is a careful count of walk lengths, and nothing in this book depends on it. The matrix $\mathbf{W}_3$ is the matrix $\mathbf{W}$ of the examples, where we saw that $\mathbf{W}^4$ has a zero at position $(1,1)$ and $\mathbf{W}^5 > \mathbf{O}$; here $(n-1)^2 + 1 = 5$. In the walks: a closed walk at 1 must leave by $1 \to 3$ and return by $2 \to 1$, so it is made of 3-cycles $1 \to 3 \to 2 \to 1$ and detours round $3 \to 2 \to 3$, and its length $3 + 3a + 2b$, with $a, b \geq 0$, is never 4. For $n = 4$ the bound is 10, and $\mathbf{W}_4^9$ is not positive while $\mathbf{W}_4^{10}$ is.

Remark. The bound is quadratic in n, not $n-1$ as for $(\mathbf{I}+\mathbf{A})^{n-1}$ in Theorem 19.2.6. Adding $\mathbf{I}$ puts a loop at every vertex, which lets walks wait; a primitive matrix may have no such loops, and then the walks must make up lengths from long cycles, as $\mathbf{W}_n$ does with its cycles of lengths n and $n-1$.

19.5.7 Exercises

19.5.7.1 A. Check your understanding

Exercise. A1

(a) Define a primitive matrix, and state the limit theorem for primitive matrices.

(b) True or false: every primitive matrix is irreducible. Justify your answer.

(c) True or false: every irreducible nonnegative matrix is primitive. Justify your answer.

(d) True or false: if $\mathbf{A} \geq \mathbf{O}$ is irreducible and $a_{11} > 0$, then $\mathbf{A}$ is primitive. Justify your answer.

(e) An irreducible $\mathbf{A} \geq \mathbf{O}$ has $\rho(\mathbf{A}) = 2$ and period 4. Which of its eigenvalues can you name without further information?

(f) Name the two ingredients from earlier chapters that the proof of Theorem 19.5.2 combines.

Solution.

(a) $\mathbf{A} \in M_n(\mathbb{R})$, $\mathbf{A} \geq \mathbf{O}$, is primitive if $\mathbf{A}^k > \mathbf{O}$ for some $k \geq 1$ (Definition 19.5.1). Theorem 19.5.2: if $\mathbf{A}$ is primitive with $\rho = \rho(\mathbf{A})$, and $\mathbf{v} > \mathbf{0}$, $\mathbf{w} > \mathbf{0}$ satisfy $\mathbf{A}\mathbf{v} = \rho\mathbf{v}$ and $\mathbf{w}^{\top}\mathbf{A} = \rho\mathbf{w}^{\top}$, then $(\mathbf{A}/\rho)^m \to \mathbf{v}\mathbf{w}^{\top}/(\mathbf{w}^{\top}\mathbf{v})$ entrywise.

(b) True, by Proposition 19.5.1 (b).

(c) False. The swap $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ is irreducible, but each of its powers is $\mathbf{I}$ or the swap, and both have zero entries.

(d) True, by Corollary 19.5.5: the loop $1 \to 1$ is a closed walk of length 1, so the period is 1. (If $n = 1$, then $\mathbf{A} = (a_{11})$ with $a_{11} > 0$ is primitive directly.)

(e) By Theorem 19.5.4, $h = 4$, so the eigenvalues of modulus 2 are $2, 2i, -2, -2i$, each algebraically simple. The others cannot be named, but they come in orbits of multiplication by i, by part (b) of that theorem.

(f) The Jordan-form convergence criterion of Chapter 10 §10 (Theorem 10.10.4), which gives convergence of $(\mathbf{A}/\rho)^m$; and the Perron–Frobenius theorem (Theorem 19.3.2), which gives the simple eigenvalue ρ and the positive vectors $\mathbf{v}, \mathbf{w}$ that identify the limit.

19.5.7.2 B. Practice

Exercise. B1: Which are primitive?

Determine which of the following matrices are primitive. Justify your answer.

(a) $\begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}$.

(b) $\begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$.

(c) $\begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$.

(d) $\begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}$.

Solution.

(a) Primitive: its square is $\begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix} > \mathbf{O}$.

(b) Not primitive. It is the cyclic permutation matrix of Example (The Cyclic Permutation) with $n = 3$ (columns $\mathbf{e}_3, \mathbf{e}_1, \mathbf{e}_2$, the same shift run the other way), and every power of a permutation matrix is a permutation matrix (Lemma 3.4.7), which has zero entries.

(c) Not primitive. The positive entries are at $(1,2), (2,1), (2,3), (3,2)$, so the edges of $G(\mathbf{A})$ are $1 \to 2$, $2 \to 1$, $2 \to 3$ and $3 \to 2$, and every edge joins vertex 2 to the set $\{1,3\}$. A closed walk therefore has even length, and $1 \to 2 \to 1$ has length 2: the matrix has period 2. It is irreducible, since every vertex reaches every other through 2, so $G(\mathbf{A})$ is strongly connected (Theorem 19.2.5). By Corollary 19.5.5 it is not primitive. (Directly: $\mathbf{A}^m$ has zero $(1,2)$-entry for even m and zero $(1,1)$-entry for odd m.)

(d) Not primitive. The positive entries sit at positions (i,j) with $i \le j$, so every edge $j \to i$ has $i \le j$, vertex numbers never increase along a walk, and there is no walk from 1 to 2. By Lemma 19.2.1 (a), the $(2,1)$-entry of every power is 0.

Exercise. B2: A limit without diagonalizing

Let $\mathbf{A} = \begin{pmatrix} 0 & 0 & 1 \\ 2 & 0 & 1 \\ 2 & 1 & 0 \end{pmatrix}$.

(a) Show that $\mathbf{A}$ is primitive.

(b) Show that $p_{\mathbf{A}} = (x-2)(x+1)^2$, and that $\mathbf{A}$ is **not** diagonalizable.

(c) Find positive right and left eigenvectors for $\rho(\mathbf{A})$, and hence $\lim_m (\mathbf{A}/2)^m$.

Solution.

(a) Compute

$$\mathbf{A}^2 = \begin{pmatrix} 2 & 1 & 0 \\ 2 & 1 & 2 \\ 2 & 0 & 3 \end{pmatrix}, \qquad \mathbf{A}^4 = (\mathbf{A}^2)^2 = \begin{pmatrix} 6 & 3 & 2 \\ 10 & 3 & 8 \\ 10 & 2 & 9 \end{pmatrix} > \mathbf{O}.$$

So $\mathbf{A}$ is primitive.

(b) Here $x\mathbf{I}-\mathbf{A}=\begin{pmatrix} x & 0 & -1 \\ -2 & x & -1 \\ -2 & -1 & x \end{pmatrix}$, and expansion along the first row gives

$$\begin{aligned} p_{\mathbf{A}} &= x(x^2-1)+(-1)\big((-2)(-1)-x(-2)\big) \\ &= x^3-x-2-2x = x^3-3x-2. \end{aligned}$$

Since $(x-2)(x+1)^2=(x-2)(x^2+2x+1)=x^3-3x-2$, this is the claimed factorization. For $\lambda=-1$, $\mathbf{A}+\mathbf{I}=\begin{pmatrix} 1 & 0 & 1 \\ 2 & 1 & 1 \\ 2 & 1 & 1 \end{pmatrix}$ has rank 2 (rows 2 and 3 agree, rows 1 and 2 are independent), so $g_{\mathbf{A}}(-1)=1<2=a_{\mathbf{A}}(-1)$, and $\mathbf{A}$ is not diagonalizable.

(c) $\mathbf{A}-2\mathbf{I}=\begin{pmatrix} -2 & 0 & 1 \\ 2 & -2 & 1 \\ 2 & 1 & -2 \end{pmatrix}$: the first row gives $v_3=2v_1$, and then the second gives $2v_1-2v_2+2v_1=0$, so $v_2=2v_1$. Take $\mathbf{v}=(1,2,2)$; check: $\mathbf{A}\mathbf{v}=(2,4,4)$. For the left vector, $\mathbf{A}^{\top}-2\mathbf{I}=\begin{pmatrix} -2 & 2 & 2 \\ 0 & -2 & 1 \\ 1 & 1 & -2 \end{pmatrix}$: the second row gives $w_3=2w_2$, and the first gives $w_1=w_2+w_3=3w_2$. Take $\mathbf{w}=(3,1,2)$; check: $\mathbf{w}^{\top}\mathbf{A}=(0+2+4,\ 0+0+2,\ 3+1+0)=(6,2,4)=2\mathbf{w}^{\top}$. Then $\mathbf{w}^{\top}\mathbf{v}=3+2+4=9$, and by Theorem 19.5.2

$$\left(\tfrac{1}{2}\mathbf{A}\right)^m \longrightarrow \frac{1}{9}\begin{pmatrix}1\\2\\2\end{pmatrix}\begin{pmatrix}3 & 1 & 2\end{pmatrix}=\frac{1}{9}\begin{pmatrix}3 & 1 & 2\\6 & 2 & 4\\6 & 2 & 4\end{pmatrix}.$$

Part (b) shows that Chapter 9's diagonalization method is unavailable here; the theorem does not need it.

Exercise. B3: Period and peripheral eigenvalues

Let $\mathbf{A}=\begin{pmatrix} 0 & 1 & 0 \\ 2 & 0 & 2 \\ 0 & 1 & 0 \end{pmatrix}$.

(a) Show that $\mathbf{A}$ is irreducible with period 2.

(b) Use Theorem 19.5.4 to predict the eigenvalues of modulus $\rho(\mathbf{A})$ in terms of $\rho(\mathbf{A})$, then compute $p_{\mathbf{A}}$ and confirm.

(c) Hence explain why $(\mathbf{A}/\rho(\mathbf{A}))^m$ does not converge.

Solution.

(a) The positive entries are at $(1,2),(2,1),(2,3),(3,2)$, the same pattern as in B1(c), so the argument there applies: $\mathbf{A}$ is irreducible with period 2.

(b) With $h=2$ and $\omega=-1$, the peripheral eigenvalues are ρ and $-\rho$, and the spectrum is

symmetric under $\lambda \mapsto -\lambda$. Expanding along the first row,

$$p_{\mathbf{A}} = \det \begin{pmatrix} x & -1 & 0 \\ -2 & x & -2 \\ 0 & -1 & x \end{pmatrix} = x(x^2 - 2) + 1 \cdot (-2x) = x(x-2)(x+2).$$

So $\rho = 2$, the peripheral eigenvalues are ± 2, and the remaining eigenvalue 0 is its own negative, as predicted.

(c) The vector $\mathbf{z} = (1, -2, 1)$ satisfies $\mathbf{Az} = (-2, 4, -2) = -2\mathbf{z}$, so $\mathbf{A}/2$ has the eigenvalue -1. By Theorem 10.10.4, the powers of a matrix with an eigenvalue of modulus 1 other than 1 do not converge. Concretely, $(\mathbf{A}/2)^m \mathbf{z} = (-1)^m \mathbf{z}$ alternates in sign.

19.5.7.3 C. Going deeper

Exercise. C1: Shifting by the identity

Let $\mathbf{A} \in M_n(\mathbb{R})$, $\mathbf{A} \geq \mathbf{O}$, be irreducible, with right and left Perron vectors $\mathbf{v}, \mathbf{w}$, and let $t > 0$.

(a) Prove that $\mathbf{A} + t\mathbf{I}$ is primitive, with $\rho(\mathbf{A} + t\mathbf{I}) = \rho(\mathbf{A}) + t$ and the same Perron vectors $\mathbf{v}, \mathbf{w}$.

(b) Deduce that $\left((\mathbf{A} + t\mathbf{I})/(\rho(\mathbf{A}) + t)\right)^m \to \mathbf{v}\mathbf{w}^\top/(\mathbf{w}^\top \mathbf{v})$, and apply this to the swap with $t = 1$.

Solution.

(a) $\mathbf{A} + t\mathbf{I} \geq \mathbf{O}$ has the same off-diagonal pattern as $\mathbf{A}$, so $(\mathbf{I} + \mathbf{A} + t\mathbf{I})^{n-1} \geq (\mathbf{I} + \mathbf{A})^{n-1} > \mathbf{O}$ entrywise, since each product of nonnegative matrices can only grow when an entry grows; so $\mathbf{A} + t\mathbf{I}$ is irreducible (Theorem 19.2.6). Its diagonal entries are $\geq t > 0$, so it is primitive by Corollary 19.5.5. By Theorem 9.10.1 with $q = x + t$, its eigenvalues are $\lambda + t$, $\lambda \in \operatorname{spec}(\mathbf{A})$, and $|\lambda + t| \leq |\lambda| + t \leq \rho(\mathbf{A}) + t$ by Theorem 0.6.4, with $\rho(\mathbf{A}) + t$ itself an eigenvalue, since $\rho(\mathbf{A}) \in \operatorname{spec}(\mathbf{A})$. So $\rho(\mathbf{A} + t\mathbf{I}) = \rho(\mathbf{A}) + t$. Finally $(\mathbf{A} + t\mathbf{I})\mathbf{v} = (\rho(\mathbf{A}) + t)\mathbf{v}$ and $\mathbf{w}^\top(\mathbf{A} + t\mathbf{I}) = (\rho(\mathbf{A}) + t)\mathbf{w}^\top$, with $\mathbf{v}, \mathbf{w} > \mathbf{0}$.

(b) By (a) and Theorem 19.5.2 applied to $\mathbf{A} + t\mathbf{I}$. For the swap $\mathbf{S}$, $\rho = 1$, and $\mathbf{1}$ is a positive right and left eigenvector, so the limit is $\mathbf{1}\mathbf{1}^\top/(\mathbf{1}^\top\mathbf{1}) = \mathbf{1}\mathbf{1}^\top/2$. Directly, $(\mathbf{S} + \mathbf{I})/2 = \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$, whose powers are all equal to itself, in agreement.

Exercise. C2: Traces detect the period

Let $\mathbf{A} \geq \mathbf{O}$ be irreducible, not the 1×1 zero matrix, with period h.

(a) Prove that $\operatorname{tr}(\mathbf{A}^m) = 0$ for every $m \geq 1$ that is not a multiple of h, in two ways: from Theorem 19.5.4 (b), and from Lemma 19.2.1 (a).

(b) Deduce that if $\operatorname{tr}(\mathbf{A}^2) > 0$ and $\operatorname{tr}(\mathbf{A}^3) > 0$, then $\mathbf{A}$ is primitive.

(c) Check (b) on the matrix $\mathbf{W}$ of the examples.

Solution.

(a) *Spectrum.* Let $\omega = e^{2\pi i/h}$ and $p_{\mathbf{A}} = \prod_i (x - \lambda_i)$. By Theorem 19.5.4 (c), the period h is also the number of peripheral eigenvalues, so part (b) of that theorem applies with this ω. By Corollary 9.10.2, applied over $\mathbb{C}$, $\operatorname{tr}(\mathbf{A}^m) = \sum_i \lambda_i^m$. By Theorem 19.5.4 (b), the list $\omega\lambda_1, \dots, \omega\lambda_n$ is a rearrangement of $\lambda_1, \dots, \lambda_n$, so $\sum_i \lambda_i^m = \sum_i (\omega\lambda_i)^m = \omega^m \sum_i \lambda_i^m$. If $h \nmid m$, then $\omega^m \neq 1$ (Example (Roots of Unity)), so $\operatorname{tr}(\mathbf{A}^m) = 0$.
Walks. $\operatorname{tr}(\mathbf{A}^m) = \sum_i (\mathbf{A}^m)_{ii}$ is a sum of nonnegative terms, positive exactly when some $(\mathbf{A}^m)_{ii} > 0$, that is, by Lemma 19.2.1 (a), when there is a closed walk of length m. Every closed walk has length a multiple of the period, so for $h \nmid m$ the trace is 0.

(b) By (a), $h \mid 2$ and $h \mid 3$, so $h = 1$, and $\mathbf{A}$ is primitive by Corollary 19.5.5.

(c) $\mathbf{W}^2 = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix}$ has trace 2, and $\mathbf{W}^3 = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}$ has trace 3; so $\mathbf{W}$ is primitive, as found directly.

Exercise. C3: A finite test for primitivity

Let $\mathbf{A} \in M_n(\mathbb{R})$, $\mathbf{A} \geq \mathbf{O}$, and let $\mathbf{Z}_m$ be the zero pattern of $\mathbf{A}^m$: the 0/1 matrix with a 1 exactly where $\mathbf{A}^m$ is positive.

(a) Prove that $\mathbf{Z}_{m+1}$ is determined by $\mathbf{Z}_m$ and $\mathbf{Z}_1$.

(b) Deduce that if $\mathbf{Z}_a = \mathbf{Z}_b$ with $a < b$, then $\mathbf{Z}_{a+j} = \mathbf{Z}_{b+j}$ for every $j \geq 0$.

(c) Prove that if $\mathbf{A}$ is primitive, then $\mathbf{A}^k > \mathbf{O}$ for some $k \leq 2^{n^2}$. Compare with Theorem 19.5.6.

Hint: there are only 2^{n^2} zero patterns; use Proposition 19.5.1 (a).

Solution.

(a) $(\mathbf{A}^{m+1})_{ij} = \sum_l (\mathbf{A}^m)_{il} a_{lj}$ is a sum of nonnegative terms, positive exactly when some l has $(\mathbf{A}^m)_{il} > 0$ and $a_{lj} > 0$. So $(\mathbf{Z}_{m+1})_{ij} = 1$ if and only if $(\mathbf{Z}_m)_{il} = (\mathbf{Z}_1)_{lj} = 1$ for some l, which depends only on $\mathbf{Z}_m$ and $\mathbf{Z}_1$.

(b) By induction on j: the case $j = 0$ is the hypothesis, and if $\mathbf{Z}_{a+j} = \mathbf{Z}_{b+j}$, then (a) applied to both gives $\mathbf{Z}_{a+j+1} = \mathbf{Z}_{b+j+1}$.

(c) Among $\mathbf{Z}_1, \dots, \mathbf{Z}_{N+1}$, $N = 2^{n^2}$, two agree, say $\mathbf{Z}_a = \mathbf{Z}_b$ with $a < b \leq N + 1$, so $a \leq N$. By (b), $\mathbf{Z}_{a+q(b-a)} = \mathbf{Z}_a$ for every $q \geq 0$, by induction on q. If $\mathbf{A}^k > \mathbf{O}$, then by Proposition 19.5.1 (a) $\mathbf{A}^m > \mathbf{O}$ for all $m \geq k$; choosing q with $a + q(b-a) \geq k$ shows that $\mathbf{Z}_a$ is the all-ones pattern, so $\mathbf{A}^a > \mathbf{O}$ with $a \leq 2^{n^2}$. Wielandt's bound $(n-1)^2 + 1$ is enormously smaller: for $n = 3$ it is 5 against 512.

19.6 Markov Chains

Chapter 9 §11 proved that a Markov chain settles down to a unique steady state, but only under three hypotheses on the eigenvalues of its transition matrix, one of them diagonalizability. It closed with a promise: the Perron–Frobenius theorem would give the same conclusions "for every

stochastic matrix with some power having all entries positive, without assuming diagonalizability and without computing eigenvalues". This section keeps that promise, using the limit theorem of §05. It then asks what happens when no power is positive. Irreducible chains still have a unique steady state, and they approach it on average even when they oscillate. Reducible chains can have many steady states, and among them the absorbing chains, which end in a trap, have an exact theory built on the Neumann series.

Conventions. Stochastic means **column**-stochastic, as in Definition 9.11.1: $\mathbf{A} \geq \mathbf{O}$ and every column sums to 1, that is, $\mathbf{1}^{\top}\mathbf{A} = \mathbf{1}^{\top}$. The entry a_{ij} is the probability of moving **to** state i **from** state j, and a Markov chain is $\mathbf{x}_{k+1} = \mathbf{A}\mathbf{x}_k$ (Definition 9.11.2). Many books use rows and the transposed matrix; we never do. The order $\geq$ is entrywise.

19.6.1 Steady states and the Perron root

We first place stochastic matrices inside the theory of this chapter. Since $\mathbf{x}_m = \mathbf{A}^m\mathbf{x}_0$, the entry $(\mathbf{A}^m)_{ij}$ is the probability of being in state i after m steps when starting from state j. By Lemma 19.2.1 (a) it is positive exactly when the graph $G(\mathbf{A})$ has a walk of length m from j to i: the edges of the graph are the moves the chain can make in one step, and the walks are the routes it can take.

Proposition 19.6.1. The Perron Root of a Stochastic Matrix

Let $\mathbf{A} \in M_n(\mathbb{R})$ be stochastic. Then:

(a) $\rho(\mathbf{A}) = 1$, and $\mathbf{1}$ is a positive left eigenvector for it;

(b) the steady states of $\mathbf{A}$ are exactly the nonnegative eigenvectors for $\rho(\mathbf{A})$ whose entries sum to 1;

(c) if $\mathbf{A}$ is irreducible, then $\mathbf{A}$ has **exactly one** steady state $\mathbf{v}$, and $\mathbf{v} > \mathbf{0}$.

Idea. Example (An irreducible stochastic matrix) already found these facts for one matrix, and the argument there is general. The column sums make $\mathbf{1}^{\top}$ a left eigenvector for 1, and no eigenvalue is larger in modulus; so a steady state is a nonnegative eigenvector for the Perron root; when $\mathbf{A}$ is irreducible, Perron–Frobenius with the normalization $\mathbf{1}^{\top}\mathbf{v} = 1$ leaves room for only one.

Proof.

(a) By Proposition 9.11.3 (c), $1 \in \operatorname{spec}(\mathbf{A})$ and $\mathbf{1}^{\top}\mathbf{A} = \mathbf{1}^{\top}$. By Corollary 10.10.5 (a), every eigenvalue has modulus at most 1. So $\rho(\mathbf{A}) = 1$.

(b) This is Definition 9.11.2 read with (a): a steady state is a probability vector $\mathbf{v}$ with $\mathbf{A}\mathbf{v} = \mathbf{v} = \rho(\mathbf{A})\mathbf{v}$.

(c) By Theorem 19.3.2 (a), there is $\mathbf{u} > \mathbf{0}$ with $\mathbf{A}\mathbf{u} = \mathbf{u}$; then $\mathbf{v} = \mathbf{u}/(\mathbf{1}^{\top}\mathbf{u})$ is a steady state with $\mathbf{v} > \mathbf{0}$. If $\mathbf{v}'$ is any steady state, it is a nonnegative eigenvector, so $\mathbf{v}' = c\mathbf{v}$ with $c > 0$ by Theorem 19.3.2 (c), and $1 = \mathbf{1}^{\top}\mathbf{v}' = c\,\mathbf{1}^{\top}\mathbf{v} = c$. Hence $\mathbf{v}' = \mathbf{v}$.

■

In the notation of §03, the Perron root of a stochastic matrix is 1, the all-ones vector is a

positive left eigenvector, and for an irreducible stochastic matrix the steady state is the right Perron vector.

Warning. **A unique positive steady state does not make the chain converge.** The swap $\mathbf{S} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ is stochastic and irreducible, so by (c) it has exactly one steady state, $(\frac{1}{2}, \frac{1}{2})$, and it is positive. Yet the chain started at $\mathbf{e}_1$ alternates $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_1, \dots$ and never approaches it. Existence and uniqueness of the steady state come from irreducibility; convergence needs more.

19.6.2 The long-run theorem

The "more" is primitivity. Here are the three conclusions of Theorem 9.11.4, in the same words, under the hypothesis that Chapter 9 §11 promised.

Theorem 19.6.2. Long-Run Behavior of a Markov Chain

Let $\mathbf{A} \in M_n(\mathbb{R})$ be stochastic, and suppose that $\mathbf{A}^k > \mathbf{O}$ for some $k \geq 1$. Then:

(a) $\mathbf{A}$ has exactly one steady state $\mathbf{v}$, and $\mathbf{v} > \mathbf{0}$;

(b) $\mathbf{A}^m \to \mathbf{v}\mathbf{1}^\top$ entrywise as $m \to \infty$;

(c) for every probability vector $\mathbf{x}_0$, the Markov chain $\mathbf{x}_m = \mathbf{A}^m \mathbf{x}_0$ converges entrywise to $\mathbf{v}$.

Idea. For a stochastic matrix, Theorem 19.5.2 has $\rho = 1$, and the left eigenvector can be taken to be $\mathbf{1}$, so the limit $\mathbf{v}\mathbf{w}^\top/(\mathbf{w}^\top\mathbf{v})$ becomes $\mathbf{v}\mathbf{1}^\top$ once the steady state is normalized by $\mathbf{1}^\top\mathbf{v} = 1$. The chain inherits the limit column by column.

Proof. By hypothesis $\mathbf{A}$ is primitive (Definition 19.5.1), hence irreducible (Proposition 19.5.1 (b)). So (a) is Proposition 19.6.1 (c).

(b) By Proposition 19.6.1 (a), $\rho(\mathbf{A}) = 1$ and $\mathbf{w} = \mathbf{1}$ is positive with $\mathbf{w}^\top\mathbf{A} = \mathbf{w}^\top$; and $\mathbf{v} > \mathbf{0}$ satisfies $\mathbf{A}\mathbf{v} = \mathbf{v}$. By Theorem 19.5.2,

$$\mathbf{A}^m = \Big(\frac{1}{\rho(\mathbf{A})}\mathbf{A}\Big)^m \longrightarrow \frac{\mathbf{v}\mathbf{1}^\top}{\mathbf{1}^\top\mathbf{v}} = \mathbf{v}\mathbf{1}^\top,$$

since $\mathbf{1}^\top\mathbf{v} = 1$.

(c) Entry r of $\mathbf{A}^m\mathbf{x}_0$ is $\sum_s (\mathbf{A}^m)_{rs}(\mathbf{x}_0)_s$, which by (b) converges to $\sum_s v_r(\mathbf{x}_0)_s = v_r\,(\mathbf{1}^\top\mathbf{x}_0) = v_r$.

■

Compare the hypotheses with those of Theorem 9.11.4. That theorem asked for three facts about the eigenvalues: $\mathbf{A}$ diagonalizable over $\mathbb{C}$, $a_{\mathbf{A}}(1) = 1$, and $|\lambda| < 1$ for every other eigenvalue. Chapter 10 §10 (Corollary 10.10.5) removed the first, but still asked to locate every eigenvalue on the unit circle. The new hypothesis is a statement about the **zero pattern** of one power of

$\mathbf{A}$, checked by multiplying matrices; no eigenvalue is computed, and none needs to be. The last two eigenvalue conditions become conclusions, by Proposition 19.5.1 (c) and Theorem 19.3.2 (b), and diagonalizability is simply irrelevant. By Corollary 19.5.5 the hypothesis can often be checked without multiplying at all: an irreducible stochastic matrix with one positive diagonal entry, a state that can stay put, qualifies. In the vocabulary of probability, a stochastic matrix with a positive power is called **regular**.

Example. A Chain that Cannot Be Diagonalized

A token moves among three positions. From position 1 it always moves to 3; from 2 it moves to 1 or 3 with probability $\frac{1}{2}$ each; from 3 it moves to 1 or 2 with probability $\frac{1}{2}$ each. So

$$\mathbf{A} = \begin{pmatrix} 0 & 1/2 & 1/2 \\ 0 & 0 & 1/2 \\ 1 & 1/2 & 0 \end{pmatrix}.$$

Show that Theorem 9.11.4 does not apply but Theorem 19.6.2 does, and find the long-run distribution.

Solution. Diagonalization fails. Expanding $\det(x\mathbf{I} - \mathbf{A})$ along the first column,

$$\begin{aligned} p_{\mathbf{A}} &= x\Big(x^2 - \frac{1}{4}\Big) - \Big(\frac{1}{4} + \frac{x}{2}\Big) \\ &= x^3 - \frac{3}{4}x - \frac{1}{4} = (x-1)\Big(x + \frac{1}{2}\Big)^2. \end{aligned}$$

For the double eigenvalue, $\mathbf{A} + \frac{1}{2}\mathbf{I} = \begin{pmatrix} 1/2 & 1/2 & 1/2 \\ 0 & 1/2 & 1/2 \\ 1 & 1/2 & 1/2 \end{pmatrix}$ has rank 2: its first two rows are independent, and the third is twice the first minus the second. So $g_{\mathbf{A}}(-\frac{1}{2}) = 1 < 2 = a_{\mathbf{A}}(-\frac{1}{2})$, and $\mathbf{A}$ is not diagonalizable (Theorem 9.4.1). Theorem 9.11.4 does not apply.

A positive power. Squaring,

$$\mathbf{A}^2 = \begin{pmatrix} 1/2 & 1/4 & 1/4 \\ 1/2 & 1/4 & 0 \\ 0 & 1/2 & 3/4 \end{pmatrix}, \qquad \mathbf{A}^4 = (\mathbf{A}^2)^2 = \begin{pmatrix} 3/8 & 5/16 & 5/16 \\ 3/8 & 3/16 & 1/8 \\ 1/4 & 1/2 & 9/16 \end{pmatrix} > \mathbf{O}.$$

So Theorem 19.6.2 applies with $k = 4$. ($\mathbf{A}^3$ still has a zero at position $(2,1)$: three steps cannot take the token from 1 to 2.)

The steady state. $(\mathbf{A} - \mathbf{I})\mathbf{v} = \mathbf{0}$ reads $-v_1 + \frac{1}{2}v_2 + \frac{1}{2}v_3 = 0$ and $-v_2 + \frac{1}{2}v_3 = 0$. So $v_3 = 2v_2$, then $v_1 = \frac{1}{2}v_2 + v_2 = \frac{3}{2}v_2$, and $\mathbf{v}$ is a multiple of $(3, 2, 4)$. Scaling to sum 1,

$$\mathbf{v} = \Big(\frac{1}{3}, \frac{2}{9}, \frac{4}{9}\Big), \qquad \mathbf{A}^m \longrightarrow \mathbf{v}\mathbf{1}^{\top}.$$

Check: $\mathbf{A}\mathbf{v} = (\frac{1}{9} + \frac{2}{9},\ \frac{2}{9},\ \frac{1}{3} + \frac{1}{9}) = (\frac{1}{3}, \frac{2}{9}, \frac{4}{9})$. In the long run the token is at position 3 four times out of nine, wherever it starts.

What the Jordan block does. Solving the recurrence exactly (for instance from the Jordan

form, as in Chapter 10 §10) gives

$$(\mathbf{A}^m)_{11} = \frac{1}{3}\Big(1 + 2\Big(-\frac{1}{2}\Big)^m\Big),$$
$$(\mathbf{A}^m)_{21} = \frac{2}{9}\Big(1 + (3m-1)\Big(-\frac{1}{2}\Big)^m\Big),$$

which one can confirm for $m = 0, 1, 2$ against $\mathbf{I}, \mathbf{A}, \mathbf{A}^2$. The factor m in the second formula comes from the 2×2 Jordan block for $-\frac{1}{2}$, whose powers carry the entry $m\lambda^{m-1}$ (Chapter 10 §10): the error decays like $m\,2^{-m}$ rather than 2^{-m}. It still decays, and the theorem never needed to know.

Quick check. Does Theorem 19.6.2 apply to $\mathbf{A} = \begin{pmatrix} 1/2 & 1 \\ 1/2 & 0 \end{pmatrix}$? If so, what is $\lim_m \mathbf{A}^m$?

Solution. Yes: $\mathbf{A}$ is stochastic and $\mathbf{A}^2 = \begin{pmatrix} 3/4 & 1/2 \\ 1/4 & 1/2 \end{pmatrix} > \mathbf{O}$. The steady state solves $-\frac{1}{2}v_1 + v_2 = 0$, so $\mathbf{v} = (\frac{2}{3}, \frac{1}{3})$, and $\mathbf{A}^m \to \mathbf{v}\mathbf{1}^\top = \begin{pmatrix} 2/3 & 2/3 \\ 1/3 & 1/3 \end{pmatrix}$.

The hypothesis of Theorem 19.6.2 is also necessary for the theorem as stated: if $\mathbf{A}^m \to \mathbf{v}\mathbf{1}^\top$ with $\mathbf{v} > \mathbf{0}$, then $\mathbf{A}$ is primitive (exercise C2). Without the positivity of $\mathbf{v}$ it is not needed for convergence: $\begin{pmatrix} 1 & 1/2 \\ 0 & 1/2 \end{pmatrix}$ has no positive power, yet its powers converge to $\mathbf{e}_1\mathbf{1}^\top$, whose steady state has a zero entry.

19.6.3 Irreducible chains: convergence on average

Now drop primitivity but keep irreducibility. By Proposition 19.6.1 (c) the steady state $\mathbf{v}$ is still unique and positive, but by Theorem 19.5.3 and Theorem 19.5.4 the chain has $h \geq 2$ eigenvalues on the unit circle, and its powers do not converge (Corollary 10.10.5 (b)). What survives is convergence of **averages**. The average $\frac{1}{m}(\mathbf{x}_0 + \mathbf{x}_1 + \cdots + \mathbf{x}_{m-1})$ is, entry by entry, the average over the first m steps of the probability of being in each state; read through linearity of expectation (see the end of this section), it is the expected fraction of those steps spent there. It is this that settles down.

Theorem 19.6.3. Averages of an Irreducible Chain

Let $\mathbf{A} \in M_n(\mathbb{R})$ be stochastic and irreducible, with steady state $\mathbf{v}$. Then:

(a) $\frac{1}{m}(\mathbf{I} + \mathbf{A} + \mathbf{A}^2 + \cdots + \mathbf{A}^{m-1}) \to \mathbf{v}\mathbf{1}^\top$ as $m \to \infty$;

(b) for every probability vector $\mathbf{x}_0$, the averages $\frac{1}{m}(\mathbf{x}_0 + \mathbf{x}_1 + \cdots + \mathbf{x}_{m-1})$ of the chain converge to $\mathbf{v}$;

(c) $\mathbf{A}^m$ converges if and only if $\mathbf{A}$ is primitive.

Idea. Split every vector into its component along $\mathbf{v}$ and a remainder whose entries sum to 0. The remainder lies in the image of $\mathbf{I}-\mathbf{A}$, by a dimension count, and on that image the sum $\sum_{k<m}\mathbf{A}^k$ telescopes: $\sum_{k<m}\mathbf{A}^k(\mathbf{I}-\mathbf{A})\mathbf{y}=\mathbf{y}-\mathbf{A}^m\mathbf{y}$, which stays bounded, so dividing by m kills it. The component along $\mathbf{v}$ is fixed by every power.

Proof.

(a) Let $U=\{\mathbf{x}\in\mathbb{R}^n:\mathbf{1}^\top\mathbf{x}=0\}$, the kernel of the nonzero linear functional $\mathbf{x}\mapsto\mathbf{1}^\top\mathbf{x}$, so $\dim U=n-1$ by Rank–Nullity (Theorem 2.3.1). Since $\mathbf{1}^\top(\mathbf{I}-\mathbf{A})=\mathbf{1}^\top-\mathbf{1}^\top=\mathbf{0}^\top$, the image of $\mathbf{I}-\mathbf{A}$ lies in U. The kernel of $\mathbf{I}-\mathbf{A}$ on $\mathbb{R}^n$ is the real eigenspace of $\mathbf{A}$ for 1; it contains $\mathbf{v}\neq\mathbf{0}$, and it has dimension at most $g_{\mathbf{A}}(1)\le a_{\mathbf{A}}(1)=1$ by Theorem 9.3.1 and Theorem 19.3.2 (b). (The real and the complex kernel of the real matrix $\mathbf{I}-\mathbf{A}$ have the same dimension, n minus its rank, since row reduction uses only real arithmetic.) So $\operatorname{rank}(\mathbf{I}-\mathbf{A})=n-1$ by Rank–Nullity, and an inclusion of subspaces of equal dimension is an equality: $\operatorname{im}(\mathbf{I}-\mathbf{A})=U$.
Now fix $\mathbf{x}\in\mathbb{R}^n$. Since $\mathbf{1}^\top\mathbf{v}=1$, the vector $\mathbf{x}-(\mathbf{1}^\top\mathbf{x})\mathbf{v}$ lies in U, so $\mathbf{x}=(\mathbf{1}^\top\mathbf{x})\mathbf{v}+(\mathbf{I}-\mathbf{A})\mathbf{y}$ for some $\mathbf{y}\in\mathbb{R}^n$. As $\mathbf{A}^k\mathbf{v}=\mathbf{v}$ for every k,

$$\begin{aligned}\sum_{k=0}^{m-1}\mathbf{A}^k\mathbf{x}&=m\,(\mathbf{1}^\top\mathbf{x})\mathbf{v}+\sum_{k=0}^{m-1}(\mathbf{A}^k\mathbf{y}-\mathbf{A}^{k+1}\mathbf{y})\\&=m\,(\mathbf{1}^\top\mathbf{x})\mathbf{v}+\mathbf{y}-\mathbf{A}^m\mathbf{y}.\end{aligned}$$

Every $\mathbf{A}^m$ is stochastic (Proposition 9.11.3 (b)), so its entries lie in $[0,1]$ and $|(\mathbf{A}^m\mathbf{y})_i|\le\sum_j|y_j|$. Hence each entry of $\frac{1}{m}(\mathbf{y}-\mathbf{A}^m\mathbf{y})$ has modulus at most $\frac{2}{m}\sum_j|y_j|$, which tends to 0 because $\frac{1}{m}\to0$, by the Archimedean property, a consequence of completeness (fact (A1) in the introduction to Chapter 16). Dividing the display by m, $\frac{1}{m}\sum_{k<m}\mathbf{A}^k\mathbf{x}\to(\mathbf{1}^\top\mathbf{x})\mathbf{v}=\mathbf{v}\mathbf{1}^\top\mathbf{x}$. Taking $\mathbf{x}=\mathbf{e}_j$ gives the convergence of column j, for each j.

(b) $\frac{1}{m}\sum_{k<m}\mathbf{x}_k=(\frac{1}{m}\sum_{k<m}\mathbf{A}^k)\mathbf{x}_0\to\mathbf{v}\mathbf{1}^\top\mathbf{x}_0=\mathbf{v}$ by (a), entry by entry as in the proof of Theorem 19.6.2 (c).

(c) ($\Leftarrow$) is Theorem 19.6.2 (b). ($\Rightarrow$) Suppose $\mathbf{A}^m\to Ł$. Then $\mathbf{A}^{m+1}=\mathbf{A}\mathbf{A}^m$ tends both to Ł and to $\mathbf{A}Ł$, so $\mathbf{A}Ł=Ł$, and every column of Ł lies in the real eigenspace for 1, which is $\operatorname{span}(\mathbf{v})$ by the proof of (a). Each column of $\mathbf{A}^m$ has entries ≥0 summing to 1, and both properties survive the limit, by the elementary algebra of limits recalled in the introduction to Chapter 16 (limits respect sums, and a nonstrict inequality survives a limit), so each column of Ł is a probability vector on the line $\operatorname{span}(\mathbf{v})$, namely $\mathbf{v}$ itself. Thus $Ł=\mathbf{v}\mathbf{1}^\top>\mathbf{O}$. As in the proof of Theorem 19.5.3, the n^2 entries of $\mathbf{A}^m$ are then positive for all large m, so $\mathbf{A}$ is primitive.

■

The proof of (a) used irreducibility only to know that the eigenspace for 1 is a line through a probability vector $\mathbf{v}$; nothing about the period or the other eigenvalues entered.

Example. A Walk that Oscillates

A walker on three sites $1,2,3$ in a row steps to a neighboring site at every step, choosing

each neighbor with probability $\frac{1}{2}$ when there are two. So

$$\mathbf{A} = \begin{pmatrix} 0 & 1/2 & 0 \\ 1 & 0 & 1 \\ 0 & 1/2 & 0 \end{pmatrix}.$$

Show that $\mathbf{A}^m$ does not converge, and compute the averages of Theorem 19.6.3 exactly.

Solution. The edges of $G(\mathbf{A})$ are $1 \to 2$, $2 \to 1$, $2 \to 3$ and $3 \to 2$. Every site reaches every other through site 2, so $\mathbf{A}$ is irreducible (Theorem 19.2.5); and every edge joins site 2 to $\{1, 3\}$, so closed walks have even length and the period is 2. The steady state solves $v_1 = \frac{1}{2}v_2$, $v_3 = \frac{1}{2}v_2$, so $\mathbf{v} = (\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$. Multiplying out,

$$\mathbf{A}^2 = \begin{pmatrix} 1/2 & 0 & 1/2 \\ 0 & 1 & 0 \\ 1/2 & 0 & 1/2 \end{pmatrix}, \qquad \mathbf{A}^3 = \mathbf{A}\mathbf{A}^2 = \mathbf{A}.$$

So $\mathbf{A}^m = \mathbf{A}$ for odd m and $\mathbf{A}^m = \mathbf{A}^2$ for even $m \geq 2$, and since $\mathbf{A} \neq \mathbf{A}^2$ the powers do not converge. A walker who starts at site 2 is at site 2 at every even step and never at an odd one.

For the averages, note that $\frac{1}{2}(\mathbf{A} + \mathbf{A}^2) = \begin{pmatrix} 1/4 & 1/4 & 1/4 \\ 1/2 & 1/2 & 1/2 \\ 1/4 & 1/4 & 1/4 \end{pmatrix} = \mathbf{v}\mathbf{1}^\top$. For $m = 2p$, the sum $\sum_{k<2p} \mathbf{A}^k$ contains $\mathbf{I}$ once, $\mathbf{A}$ p times, and $\mathbf{A}^2$ $p-1$ times, so

$$\begin{aligned} \frac{1}{2p}\sum_{k=0}^{2p-1} \mathbf{A}^k &= \frac{\mathbf{I} - \mathbf{A}^2 + p(\mathbf{A} + \mathbf{A}^2)}{2p} \\ &= \mathbf{v}\mathbf{1}^\top + \frac{1}{2p}(\mathbf{I} - \mathbf{A}^2) \longrightarrow \mathbf{v}\mathbf{1}^\top. \end{aligned}$$

For $m = 2p+1$ there is one more term, $\mathbf{A}^{2p} = \mathbf{A}^2$, which changes the average by at most $\frac{1}{m}$ in each entry. So the walker spends, in the long run, half its time at the middle site and a quarter at each end, although at no single time is its distribution close to $\mathbf{v}$ if it starts at a site.

The oscillation can also be removed rather than averaged. The **lazy** chain $\frac{1}{2}(\mathbf{I} + \mathbf{A})$, which at each step stays put with probability $\frac{1}{2}$ and otherwise moves by $\mathbf{A}$, is stochastic, irreducible, and has positive diagonal, so it is primitive by Corollary 19.5.5, and it has the same steady state; exercise C1 works this out.

19.6.4 Reducible chains: many steady states

Without irreducibility, uniqueness fails too. The chain can split into parts that never communicate, and each can carry its own equilibrium.

Example. Two Traps

States 1 and 2 exchange with each other and are never left; state 3 is never left either; state

4 stays with probability $\frac{1}{4}$ and otherwise moves to 1, 2 or 3 with probability $\frac{1}{4}$ each:

$$\mathbf{A} = \begin{pmatrix} 1/2 & 1/4 & 0 & 1/4 \\ 1/2 & 3/4 & 0 & 1/4 \\ 0 & 0 & 1 & 1/4 \\ 0 & 0 & 0 & 1/4 \end{pmatrix}.$$

Find all steady states, and the limit of the chain from each starting state.

Solution. Steady states. Let $\mathbf{Au} = \mathbf{u}$ with $\mathbf{u}$ a probability vector. The last row gives $\frac{1}{4}u_4 = u_4$, so $u_4 = 0$. The first row then gives $\frac{1}{2}u_1 + \frac{1}{4}u_2 = u_1$, so $u_2 = 2u_1$; the second and third rows then hold automatically. So $\mathbf{u} = (s, 2s, t, 0)$ with $s, t \geq 0$ and $3s + t = 1$. Writing $\mathbf{v}_1 = (\frac{1}{3}, \frac{2}{3}, 0, 0)$ and $\mathbf{v}_2 = \mathbf{e}_3$, the steady states are exactly

$$\mathbf{u} = c\,\mathbf{v}_1 + (1 - c)\,\mathbf{v}_2, \qquad 0 \leq c \leq 1,$$

the whole segment $[\mathbf{v}_2, \mathbf{v}_1]$ of Chapter 18. There are infinitely many, and $E_1(\mathbf{A}) = \operatorname{span}(\mathbf{v}_1, \mathbf{v}_2)$ is a plane: $a_{\mathbf{A}}(1) = 2$, since $p_{\mathbf{A}} = (x - 1)^2(x - \frac{1}{4})^2$, the first factor from the block of states $1, 2$ and from state 3, the second from the block's other eigenvalue $\frac{1}{4}$ and from state 4. In particular $\mathbf{A}$ is not irreducible, by Proposition 19.6.1 (c).

Limits. The only eigenvalue of modulus 1 is 1, so $\mathbf{A}^m$ converges by Corollary 10.10.5 (b). Its limit $\mathbf{\Pi}$ satisfies $\mathbf{A\Pi} = \mathbf{\Pi}$, as in the proof of Theorem 19.6.3 (c), and its columns are limits of probability vectors, hence probability vectors; so each column of $\mathbf{\Pi}$ is a steady state. Columns 1 and 2 of every $\mathbf{A}^m$ have zeros in rows 3 and 4, by induction on m: $\mathbf{A}$ maps a vector with zeros in those rows to a combination of its own first two columns, which again has zeros there. So columns 1 and 2 of $\mathbf{\Pi}$ are steady states with $u_3 = 0$, that is, $\mathbf{v}_1$. Column 3 of $\mathbf{A}$ is $\mathbf{e}_3$, so column 3 of every power, and of $\mathbf{\Pi}$, is $\mathbf{e}_3 = \mathbf{v}_2$. For column 4, write it as $c\mathbf{v}_1 + (1 - c)\mathbf{v}_2$ and find c from a conserved quantity: the row vector $\mathbf{y}^\top = (1, 1, 0, \frac{2}{3})$ satisfies $\mathbf{y}^\top\mathbf{A} = \mathbf{y}^\top$, as a column-by-column check shows (column 4: $\frac{1}{4} + \frac{1}{4} + \frac{2}{3} \cdot \frac{1}{4} = \frac{2}{3}$). Then $\mathbf{y}^\top\mathbf{A}^m = \mathbf{y}^\top$ for all m, hence $\mathbf{y}^\top\mathbf{\Pi} = \mathbf{y}^\top$, and entry 4 gives $\mathbf{y}^\top(c\mathbf{v}_1 + (1 - c)\mathbf{v}_2) = c = \frac{2}{3}$. Thus

$$\mathbf{\Pi} = \begin{pmatrix} 1/3 & 1/3 & 0 & 2/9 \\ 2/3 & 2/3 & 0 & 4/9 \\ 0 & 0 & 1 & 1/3 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

with last column $\frac{2}{3}\mathbf{v}_1 + \frac{1}{3}\mathbf{v}_2$. The long run depends on where the chain starts: from state 4, the chain ends up in the pair $\{1, 2\}$ with probability $\frac{2}{3}$ and in state 3 with probability $\frac{1}{3}$, the ratio $\frac{1}{2} : \frac{1}{4}$ of the one-step probabilities of leaving 4 for each.

In general the steady states of a stochastic matrix form a convex set, since a convex combination of probability vectors fixed by $\mathbf{A}$ is again one; the example shows it can be a segment rather than a point. One steady state for each part of the chain that, once entered, is never left, and all convex combinations of these: that is the general picture, which we do not prove here.

19.6.5 Absorbing chains

The simplest traps are single states. A state j is **absorbing** if $a_{jj} = 1$, so that column j of $\mathbf{A}$ is $\mathbf{e}_j$: once there, the chain stays. Absorbing states model the end of a process: a game won or

lost, a customer who has left, a molecule that has reacted.

Definition 19.6.1. Absorbing Chain

A stochastic matrix $\mathbf{A} \in M_n(\mathbb{R})$ is **absorbing** if it has at least one absorbing state, and **from every state some absorbing state can be reached**: for every j there are an absorbing state i and an $m \geq 0$ with $(\mathbf{A}^m)_{ij} > 0$. The states that are not absorbing are called **transient**.

Number the states so that the $t \geq 1$ transient states come first and the $r \geq 1$ absorbing states last. Column j for a transient j splits into the probabilities of moving to transient states and to absorbing states, and the column of an absorbing state is a standard basis vector. So $\mathbf{A}$ has the block form

$$\mathbf{A} = \begin{pmatrix} \mathbf{Q} & \mathbf{O} \\ \mathbf{R} & \mathbf{I}_r \end{pmatrix}, \qquad \mathbf{Q} \in M_t(\mathbb{R}), \quad \mathbf{R} \in M_{r \times t}(\mathbb{R}),$$

with $\mathbf{Q}, \mathbf{R} \geq \mathbf{O}$ and $\mathbf{1}^\top \mathbf{Q} + \mathbf{1}^\top \mathbf{R} = \mathbf{1}^\top$. Multiplying blocks and inducting on m,

$$\mathbf{A}^m = \begin{pmatrix} \mathbf{Q}^m & \mathbf{O} \\ \mathbf{R}_m & \mathbf{I}_r \end{pmatrix}, \qquad \mathbf{R}_m = \mathbf{R}(\mathbf{I} + \mathbf{Q} + \cdots + \mathbf{Q}^{m-1}),$$

since $\mathbf{A}^{m+1} = \mathbf{A}\mathbf{A}^m$ has lower-left block $\mathbf{R}\mathbf{Q}^m + \mathbf{R}_m = \mathbf{R}_{m+1}$. The entry $(\mathbf{Q}^m)_{ij}$ is the probability of being at transient state i after m steps from transient state j; the column sums of $\mathbf{Q}^m$ are the probabilities of not yet having been absorbed.

Lemma 19.6.4. Transient Mass Dies Out

If $\mathbf{A}$ is absorbing, with $\mathbf{Q}$ as above, then every eigenvalue of $\mathbf{Q}$ has modulus less than 1, and $\mathbf{Q}^m \to \mathbf{O}$.

Idea. Every column of $\mathbf{Q}$ sums to at most 1, which only gives $\rho(\mathbf{Q}) \leq 1$. But after enough steps every transient state has leaked some mass into an absorbing state, so every column of a suitable power $\mathbf{Q}^M$ sums to **strictly** less than 1, and the column-sum bound of Chapter 16 gives $\rho(\mathbf{Q}^M) < 1$.

Proof. For absorbing i and any j, $(\mathbf{A}^{m+1})_{ij} = \sum_l a_{il}(\mathbf{A}^m)_{lj} \geq a_{ii}(\mathbf{A}^m)_{ij} = (\mathbf{A}^m)_{ij}$, since all terms are nonnegative; so the probability of being at i never decreases. Fix a transient j. By Definition 19.6.1 there are an absorbing i and m_j with $(\mathbf{A}^{m_j})_{ij} > 0$, and then $(\mathbf{A}^m)_{ij} > 0$ for every $m \geq m_j$. Let M be the largest of the m_j over the transient j. Column j of $\mathbf{A}^M$ sums to 1 (Proposition 9.11.3 (b)) and has a positive entry $(\mathbf{A}^M)_{ij}$ in the absorbing rows, so column j of $\mathbf{Q}^M$ sums to less than 1. Since $\mathbf{Q}^M \geq \mathbf{O}$, its largest absolute column sum is some $c < 1$, and by Corollary 16.4.2, $\rho(\mathbf{Q}^M) \leq c < 1$. By Theorem 9.10.1 with $q = x^M$, the eigenvalues of $\mathbf{Q}^M$ are the M-th powers of those of $\mathbf{Q}$, so $\rho(\mathbf{Q})^M = \rho(\mathbf{Q}^M) < 1$ and $\rho(\mathbf{Q}) < 1$. Finally $\mathbf{Q}^m \to \mathbf{O}$ by Theorem 10.10.2.

■

The Neumann series now does all the work.

Theorem 19.6.5. Absorbing Chains

Let $\mathbf{A} = \begin{pmatrix} \mathbf{Q} & \mathbf{O} \\ \mathbf{R} & \mathbf{I}_r \end{pmatrix}$ be absorbing, in the block form above. Then:

(a) $\mathbf{I} - \mathbf{Q}$ is invertible, and the **fundamental matrix** $\mathbf{N} := (\mathbf{I} - \mathbf{Q})^{-1} = \sum_{k \geq 0} \mathbf{Q}^k$ satisfies $\mathbf{N} \geq \mathbf{O}$;

(b) $\mathbf{A}^m \to \begin{pmatrix} \mathbf{O} & \mathbf{O} \\ \mathbf{RN} & \mathbf{I}_r \end{pmatrix}$ entrywise;

(c) every column of the $r \times t$ matrix $\mathbf{B} := \mathbf{RN}$ is a probability vector.

Idea. The lemma gives $\rho(\mathbf{Q}) < 1$, which is exactly what the Neumann series needs. In the block formula for $\mathbf{A}^m$, the lower-left block is $\mathbf{R}$ times a partial sum of that series, so it converges to $\mathbf{RN}$; and column sums equal to 1 pass to the limit.

Proof.

(a) By Lemma 19.6.4 and Theorem 10.10.6, $\sum_k \mathbf{Q}^k$ converges and equals $(\mathbf{I} - \mathbf{Q})^{-1}$. Each partial sum is $\geq \mathbf{O}$, and a limit of nonnegative numbers is nonnegative, since a nonstrict inequality survives a limit (the elementary algebra of limits, recalled in the introduction to Chapter 16); so $\mathbf{N} \geq \mathbf{O}$.

(b) By the block formula, the upper-left block of $\mathbf{A}^m$ is $\mathbf{Q}^m \to \mathbf{O}$ (Lemma 19.6.4), and the lower-left block is $\mathbf{R}_m = \mathbf{R} \sum_{k<m} \mathbf{Q}^k$. Each entry of $\mathbf{R}_m$ is a fixed linear combination of entries of the partial sum, so $\mathbf{R}_m \to \mathbf{RN}$ by the algebra of limits, (B3) of Chapter 10 §10. The other two blocks do not depend on m.

(c) Every column of $\mathbf{A}^m$ sums to 1, and so does every column of the limit in (b). For a transient j, column j of the limit is $(\mathbf{0}, \mathbf{Be}_j)$, so the entries of $\mathbf{Be}_j$ sum to 1; they are nonnegative because $\mathbf{R}, \mathbf{N} \geq \mathbf{O}$.

■

Reading the answer. The theorem is pure algebra; its meaning comes from reading entries as probabilities. The (i, j)-entry of the limit in (b), for absorbing i and transient j, is the limit of the nondecreasing probabilities of being at i after m steps from j: the **probability of being absorbed at** i. So $\mathbf{B} = \mathbf{RN}$ is the matrix of absorption probabilities. Two further readings use a fact from probability that this book does not develop, **linearity of expectation**: the expected number of steps at which an event occurs is the sum over the steps of the probabilities of that event. With it, since $(\mathbf{Q}^k)_{ij}$ is the probability of being at transient i at step k from j,

$$\begin{aligned} N_{ij} &= \sum_{k \geq 0} (\mathbf{Q}^k)_{ij} \\ &= \text{expected number of visits to } i \text{ starting from } j, \end{aligned}$$

counting the visit at step 0 when $i = j$; and summing over i, the column sums $\tau_j = (\mathbf{1}^\top \mathbf{N})_j$ are the **expected numbers of steps before absorption**. Nothing in the proofs depends on these readings.

Example. Gambler's Ruin

A gambler holds 1, 2, or 3 coins and bets one coin at a time on a fair toss, winning or losing a coin with probability $\frac{1}{2}$ each. Play stops when the gambler holds 0 coins (ruin) or 4 coins (target). Find the fundamental matrix, the expected length of play, and the probability of reaching the target, from each starting fortune.

Solution. The transient states are the fortunes $1, 2, 3$, in that order, and the absorbing ones are 0 and 4, in that order. From fortune j the gambler moves to $j-1$ or $j+1$ with probability $\frac{1}{2}$ each, so

$$\mathbf{Q} = \begin{pmatrix} 0 & 1/2 & 0 \\ 1/2 & 0 & 1/2 \\ 0 & 1/2 & 0 \end{pmatrix}, \qquad \mathbf{R} = \begin{pmatrix} 1/2 & 0 & 0 \\ 0 & 0 & 1/2 \end{pmatrix},$$

the first row of $\mathbf{R}$ recording ruin from fortune 1 and the second reaching the target from fortune 3. Each column of $\begin{pmatrix} \mathbf{Q} \\ \mathbf{R} \end{pmatrix}$ sums to 1, and every fortune reaches 0 by losing, so the chain is absorbing.

The fundamental matrix. Solve $(\mathbf{I} - \mathbf{Q})\mathbf{N} = \mathbf{I}$ with $\mathbf{I} - \mathbf{Q} = \begin{pmatrix} 1 & -1/2 & 0 \\ -1/2 & 1 & -1/2 \\ 0 & -1/2 & 1 \end{pmatrix}$. The result is

$$\mathbf{N} = \begin{pmatrix} 3/2 & 1 & 1/2 \\ 1 & 2 & 1 \\ 1/2 & 1 & 3/2 \end{pmatrix};$$

check the first column: $(\mathbf{I} - \mathbf{Q})(\frac{3}{2}, 1, \frac{1}{2}) = (\frac{3}{2} - \frac{1}{2},\ -\frac{3}{4} + 1 - \frac{1}{4},\ -\frac{1}{2} + \frac{1}{2}) = (1, 0, 0)$, and the other columns likewise. A gambler starting with 2 coins holds exactly 2 coins at 2 steps on average, counting the start, and holds 1 or 3 coins at one step each on average.

Length of play. $\mathbf{1}^{\top}\mathbf{N} = (3, 4, 3)$: from fortunes $1, 2, 3$ the game lasts 3, 4, 3 tosses on average.

Absorption.

$$\mathbf{B} = \mathbf{R}\mathbf{N} = \begin{pmatrix} 3/4 & 1/2 & 1/4 \\ 1/4 & 1/2 & 3/4 \end{pmatrix}.$$

From fortune j the gambler reaches the target with probability $j/4$ and is ruined with probability $1 - j/4$. Each column sums to 1, as Theorem 19.6.5 (c) requires: the game ends with probability 1. Both answers match the classical formulas for a fair game with target $T = 4$: success probability j/T and expected length $j(T - j)$, which gives $3, 4, 3$.

The matrix $\mathbf{Q}$ of this example has eigenvalues 0 and $\pm\frac{1}{\sqrt{2}}$, of modulus less than 1 as Lemma 19.6.4 requires; but the lemma was proved without finding them, which is the pattern of this whole section.

19.6.6 Exercises

19.6.6.1 A. Check your understanding

Exercise. A1

(a) State the long-run theorem Theorem 19.6.2, and name the three hypotheses of Theorem 9.11.4 that it no longer needs.

(b) True or false: every irreducible stochastic matrix has exactly one steady state. Justify your answer.

(c) True or false: for every irreducible stochastic matrix, $\mathbf{A}^m$ converges. Justify your answer.

(d) True or false: if a stochastic matrix has two different steady states, it has infinitely many. Justify your answer.

(e) In an absorbing chain, what do the column sums of the fundamental matrix measure, and why do the columns of $\mathbf{RN}$ sum to 1?

Solution.

(a) If $\mathbf{A}$ is stochastic and $\mathbf{A}^k > \mathbf{O}$ for some k, then $\mathbf{A}$ has exactly one steady state $\mathbf{v}$, it is positive, $\mathbf{A}^m \to \mathbf{v}\mathbf{1}^\top$, and every chain $\mathbf{A}^m\mathbf{x}_0$ converges to $\mathbf{v}$. It no longer assumes that $\mathbf{A}$ is diagonalizable, that $a_{\mathbf{A}}(1) = 1$, or that every other eigenvalue has modulus less than 1.

(b) True, by Proposition 19.6.1 (c); moreover the steady state is positive.

(c) False. The swap $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ is irreducible and stochastic, and its powers alternate between $\mathbf{I}$ and itself. By Theorem 19.6.3 (c), the powers converge exactly when the matrix is primitive.

(d) True. If $\mathbf{u} \neq \mathbf{u}'$ are steady states, then $c\mathbf{u} + (1-c)\mathbf{u}'$ for $0 \leq c \leq 1$ is nonnegative, has entries summing to 1, and is fixed by $\mathbf{A}$; different c give different vectors, since $\mathbf{u} - \mathbf{u}' \neq \mathbf{0}$.

(e) The column sum $(\mathbf{1}^\top\mathbf{N})_j$ is the expected number of steps before absorption from transient state j. The columns of $\mathbf{RN}$ sum to 1 because the columns of $\mathbf{A}^m$ do and $\mathbf{Q}^m \to \mathbf{O}$ (Theorem 19.6.5 (c)): absorption is certain.

19.6.6.2 B. Practice

Exercise. B1: When does the long-run theorem apply?

For each stochastic matrix, determine whether Theorem 19.6.2 applies. Justify your answer, and when it applies, find the limit of $\mathbf{A}^m$.

(a) $\begin{pmatrix} 0 & 0 & 1/2 \\ 1 & 0 & 0 \\ 0 & 1 & 1/2 \end{pmatrix}$.

(b) $\begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$.

(c) $\begin{pmatrix} 1/2 & 0 \\ 1/2 & 1 \end{pmatrix}$.

Solution.

(a) It applies. The edges of $G(\mathbf{A})$ are $1 \to 2$, $2 \to 3$, $3 \to 1$ and the loop $3 \to 3$. The cycle $1 \to 2 \to 3 \to 1$ reaches every state from every state, so $\mathbf{A}$ is irreducible (Theorem 19.2.5). It has the positive diagonal entry a_{33}, so it is primitive by Corollary 19.5.5. The steady state solves $v_1 = \frac{1}{2}v_3$, $v_2 = v_1$, so $\mathbf{v} = (1,1,2)/4$, and $\mathbf{A}^m \to \mathbf{v}\mathbf{1}^\top$, every column equal to $(\frac{1}{4}, \frac{1}{4}, \frac{1}{2})$. Check: $\mathbf{A}\mathbf{v} = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4} + \frac{1}{4}) = \mathbf{v}$.

(b) It does not apply. This is a permutation matrix, so every power is a permutation matrix (Lemma 3.4.7) and has zero entries. Its powers cycle with period 3 (Example (The Cyclic Permutation)) and do not converge.

(c) It does not apply. Every power is lower triangular with $(1,2)$-entry 0, as one checks by induction: $\mathbf{A}^m = \begin{pmatrix} 2^{-m} & 0 \\ 1 - 2^{-m} & 1 \end{pmatrix}$. The powers converge nonetheless, to $\begin{pmatrix} 0 & 0 \\ 1 & 1 \end{pmatrix} = \mathbf{e}_2\mathbf{1}^\top$: state 2 is absorbing, and $\mathbf{e}_2$ is the unique steady state, with a zero entry.

Exercise. B2: A three-state chain

Let $\mathbf{A} = \begin{pmatrix} 1/2 & 1/4 & 0 \\ 1/2 & 1/2 & 1/2 \\ 0 & 1/4 & 1/2 \end{pmatrix}$.

(a) Show that $\mathbf{A}$ is stochastic and that $\mathbf{A}^2 > \mathbf{O}$.

(b) Find the steady state, and hence $\lim_m \mathbf{A}^m$ and the long-run distribution of every chain with this transition matrix.

Solution.

(a) The entries are nonnegative and the columns sum to 1. Multiplying out,

$$\mathbf{A}^2 = \begin{pmatrix} 3/8 & 1/4 & 1/8 \\ 1/2 & 1/2 & 1/2 \\ 1/8 & 1/4 & 3/8 \end{pmatrix} > \mathbf{O}.$$

(b) $(\mathbf{A} - \mathbf{I})\mathbf{v} = \mathbf{0}$: the first row gives $-\frac{1}{2}v_1 + \frac{1}{4}v_2 = 0$, so $v_2 = 2v_1$, and the third gives $\frac{1}{4}v_2 - \frac{1}{2}v_3 = 0$, so $v_3 = \frac{1}{2}v_2 = v_1$. So $\mathbf{v}$ is a multiple of $(1,2,1)$, and scaling to sum 1 gives $\mathbf{v} = (\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$. By Theorem 19.6.2, $\mathbf{A}^m \to \mathbf{v}\mathbf{1}^\top$, every column equal to $\mathbf{v}$, and every chain converges to $\mathbf{v}$ whatever its initial distribution.

Exercise. B3: An unfair game

A gambler holds 1 or 2 coins and wins each toss with probability $\frac{1}{3}$, losing with probability $\frac{2}{3}$. Play stops at 0 or 3 coins. Find the fundamental matrix, the expected length of play, and the probability of reaching 3 coins from each starting fortune.

Solution. Order the states as $1, 2$ (transient) and $0, 3$ (absorbing). From 1: to 2 with

probability $\frac{1}{3}$, to 0 with $\frac{2}{3}$. From 2: to 3 with $\frac{1}{3}$, to 1 with $\frac{2}{3}$. So

$$\mathbf{Q} = \begin{pmatrix} 0 & 2/3 \\ 1/3 & 0 \end{pmatrix}, \qquad \mathbf{R} = \begin{pmatrix} 2/3 & 0 \\ 0 & 1/3 \end{pmatrix}.$$

The chain is absorbing, since from either fortune repeated losses reach 0. Then $\mathbf{I} - \mathbf{Q} = \begin{pmatrix} 1 & -2/3 \\ -1/3 & 1 \end{pmatrix}$ has determinant $1 - \frac{2}{9} = \frac{7}{9}$, so by Theorem 0.9.8

$$\mathbf{N} = \frac{9}{7}\begin{pmatrix} 1 & 2/3 \\ 1/3 & 1 \end{pmatrix} = \begin{pmatrix} 9/7 & 6/7 \\ 3/7 & 9/7 \end{pmatrix}.$$

The expected lengths of play are $\mathbf{1}^\top \mathbf{N} = (\frac{12}{7}, \frac{15}{7})$, and

$$\mathbf{RN} = \begin{pmatrix} 6/7 & 4/7 \\ 1/7 & 3/7 \end{pmatrix}.$$

So the probability of reaching 3 coins is $\frac{1}{7}$ from 1 coin and $\frac{3}{7}$ from 2 coins, far below the fair values $\frac{1}{3}$ and $\frac{2}{3}$. The columns of $\mathbf{RN}$ sum to 1, as they must.

19.6.6.3 C. Going deeper

Exercise. C1: The lazy chain

Let $\mathbf{A} \in M_n(\mathbb{R})$ be stochastic and irreducible, with steady state $\mathbf{v}$, and let $\mathbf{M} = \frac{1}{2}(\mathbf{I} + \mathbf{A})$.

(a) Prove that $\mathbf{M}$ is stochastic and primitive, and that $\mathbf{v}$ is its unique steady state. Deduce that $\mathbf{M}^m \to \mathbf{v}\mathbf{1}^\top$.

(b) Let $\mathbf{A}$ be the walk of Example (A Walk that Oscillates). Show that $\mathbf{M}$ is the matrix of exercise B2, and that $\mathbf{M}^m = \mathbf{v}\mathbf{1}^\top + 2^{-m}\mathbf{F}$ for every $m \geq 1$, where $\mathbf{F} = \frac{1}{2}\begin{pmatrix} 1 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 1 \end{pmatrix}$.

Hint: write $\mathbf{M}$ *in terms of* $\mathbf{v}\mathbf{1}^\top$ *and* $\mathbf{F}$.

Solution.

(a) $\mathbf{M} \geq \mathbf{O}$, and $\mathbf{1}^\top\mathbf{M} = \frac{1}{2}(\mathbf{1}^\top + \mathbf{1}^\top\mathbf{A}) = \mathbf{1}^\top$, so $\mathbf{M}$ is stochastic. By Exercise (C1: Shifting by the identity) (a) with $t = 1$, $\mathbf{A}+\mathbf{I}$ is primitive, say $(\mathbf{A}+\mathbf{I})^k > \mathbf{O}$; then $\mathbf{M}^k = 2^{-k}(\mathbf{A}+\mathbf{I})^k > \mathbf{O}$, so $\mathbf{M}$ is primitive. Also $\mathbf{M}\mathbf{v} = \frac{1}{2}(\mathbf{v}+\mathbf{v}) = \mathbf{v}$. By Theorem 19.6.2 (a), $\mathbf{M}$ has exactly one steady state, so it is $\mathbf{v}$, and by part (b) of that theorem $\mathbf{M}^m \to \mathbf{v}\mathbf{1}^\top$.

(b) With $\mathbf{A}$ as in Example (A Walk that Oscillates),

$$\mathbf{M} = \frac{1}{2}\begin{pmatrix} 1 & 1/2 & 0 \\ 1 & 1 & 1 \\ 0 & 1/2 & 1 \end{pmatrix} = \begin{pmatrix} 1/2 & 1/4 & 0 \\ 1/2 & 1/2 & 1/2 \\ 0 & 1/4 & 1/2 \end{pmatrix},$$

the matrix of Exercise (B2: A three-state chain), and $\mathbf{v} = (\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$. Put $Ł = \mathbf{v}\mathbf{1}^\top$, every column equal to $\mathbf{v}$. Then $Ł^2 = \mathbf{v}(\mathbf{1}^\top\mathbf{v})\mathbf{1}^\top = Ł$; $\mathbf{1}^\top\mathbf{F} = \mathbf{0}^\top$, since each column of $\mathbf{F}$ sums to 0, so $Ł\mathbf{F} = \mathbf{v}(\mathbf{1}^\top\mathbf{F}) = \mathbf{O}$; and $\mathbf{F}\mathbf{v} = \frac{1}{2}(\frac{1}{4} - \frac{1}{4},\ 0,\ -\frac{1}{4} + \frac{1}{4}) = \mathbf{0}$, so $\mathbf{F}Ł = (\mathbf{F}\mathbf{v})\mathbf{1}^\top = \mathbf{O}$.

Also $\mathbf{F}^2 = \frac{1}{4}\begin{pmatrix} 2 & 0 & -2 \\ 0 & 0 & 0 \\ -2 & 0 & 2 \end{pmatrix} = \mathbf{F}$. Subtracting, $\mathbf{M} - Ł = \frac{1}{4}\begin{pmatrix} 1 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 1 \end{pmatrix} = \frac{1}{2}\mathbf{F}$, so $\mathbf{M} = Ł + \frac{1}{2}\mathbf{F}$. If $\mathbf{M}^m = Ł + 2^{-m}\mathbf{F}$, then

$$\begin{aligned} \mathbf{M}^{m+1} &= \left(Ł + \frac{1}{2}\mathbf{F}\right)(Ł + 2^{-m}\mathbf{F}) \\ &= Ł^2 + 2^{-m}Ł\mathbf{F} + \frac{1}{2}\mathbf{F}Ł + 2^{-m-1}\mathbf{F}^2 = Ł + 2^{-(m+1)}\mathbf{F}, \end{aligned}$$

so the formula holds for every $m \geq 1$ by induction, starting from $\mathbf{M} = Ł + \frac{1}{2}\mathbf{F}$. The lazy walk converges although the walk itself oscillates, and its distance to the limit halves at every step.

Exercise. C2: The converse of the long-run theorem

Let $\mathbf{A} \in M_n(\mathbb{R})$ be stochastic.

(a) Prove that if $\mathbf{A}^m \to \mathbf{v}\mathbf{1}^\top$ for some $\mathbf{v} > \mathbf{0}$, then $\mathbf{A}$ is primitive.

(b) Give a stochastic $\mathbf{A}$ with $\mathbf{A}^m \to \mathbf{v}\mathbf{1}^\top$ for a probability vector $\mathbf{v}$, but no power of $\mathbf{A}$ positive. Which entry of $\mathbf{v}$ is responsible?

Solution.

(a) The limit $\mathbf{v}\mathbf{1}^\top$ has (i,j)-entry $v_i > 0$. Each of the n^2 entries of $\mathbf{A}^m$ converges to a positive number, so it is positive for all m beyond some m_{ij}; for m larger than all the m_{ij}, $\mathbf{A}^m > \mathbf{O}$. So $\mathbf{A}$ is primitive.

(b) $\mathbf{A} = \begin{pmatrix} 1 & 1/2 \\ 0 & 1/2 \end{pmatrix}$. By induction $\mathbf{A}^m = \begin{pmatrix} 1 & 1-2^{-m} \\ 0 & 2^{-m} \end{pmatrix}$: indeed $\mathbf{A}\mathbf{A}^m$ has first row $(1,\ 1 - 2^{-m} + 2^{-m-1}) = (1,\ 1 - 2^{-(m+1)})$ and second row $(0,\ 2^{-(m+1)})$. So $\mathbf{A}^m \to \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix} = \mathbf{e}_1\mathbf{1}^\top$, while every power has $(2,1)$-entry 0. The responsible entry is $v_2 = 0$: (a) fails exactly because $\mathbf{v}$ is not positive.

Exercise. C3: First-step analysis

Let $\mathbf{A}$ be absorbing, with fundamental matrix $\mathbf{N}$ and $\boldsymbol{\tau}^\top = \mathbf{1}^\top\mathbf{N}$.

(a) Prove that $\mathbf{N} = \mathbf{I} + \mathbf{N}\mathbf{Q}$, and deduce that $\tau_j = 1 + \sum_i q_{ij}\tau_i$ for every transient j. Interpret this equation.

(b) Prove, by multiplying by $\mathbf{R}$, that $\mathbf{B} = \mathbf{R}\mathbf{N}$ satisfies $\mathbf{B} = \mathbf{R} + \mathbf{B}\mathbf{Q}$, and interpret it.

(c) Use (a) to recompute the expected lengths of play $(3,4,3)$ of Example (Gambler's Ruin) without inverting a matrix.

Solution.

(a) $\mathbf{N}(\mathbf{I} - \mathbf{Q}) = \mathbf{I}$ because $\mathbf{N}$ is the inverse of $\mathbf{I} - \mathbf{Q}$, so $\mathbf{N} = \mathbf{I} + \mathbf{N}\mathbf{Q}$. Multiplying on the

left by $\mathbf{1}^\top$, $\boldsymbol{\tau}^\top = \mathbf{1}^\top + \boldsymbol{\tau}^\top\mathbf{Q}$, whose j-th entry is $\tau_j = 1 + \sum_i \tau_i q_{ij}$. In words: from j, one step is taken, the chain lands at transient i with probability q_{ij}, and from there the expected remaining time is τ_i; landing at an absorbing state contributes nothing more.

(b) Multiply $\mathbf{N} = \mathbf{I} + \mathbf{NQ}$ on the left by $\mathbf{R}$: $\mathbf{B} = \mathbf{R} + \mathbf{BQ}$, that is, $b_{ij} = r_{ij} + \sum_l b_{il} q_{lj}$. The probability of absorption at i from j is the probability of stepping into i at once, plus the probability of stepping to a transient l and being absorbed at i from there.

(c) With $\mathbf{Q}$ as in Example (Gambler's Ruin), (a) reads $\tau_1 = 1 + \frac{1}{2}\tau_2$, $\tau_2 = 1 + \frac{1}{2}\tau_1 + \frac{1}{2}\tau_3$, $\tau_3 = 1 + \frac{1}{2}\tau_2$. The first and third equations give $\tau_1 = \tau_3 = 1 + \frac{1}{2}\tau_2$, so $\tau_2 = 1 + \tau_1 = 1 + 1 + \frac{1}{2}\tau_2$, giving $\tau_2 = 4$ and $\tau_1 = \tau_3 = 3$. (The system has only one solution, since its matrix is $(\mathbf{I} - \mathbf{Q})^\top$, which is invertible by Theorem 19.6.5 (a).)

19.7 Birkhoff's Theorem

Chapter 18 §08 met the doubly stochastic matrices, the nonnegative matrices whose rows and columns all add up to 1. It showed that every permutation matrix is an extreme point of the set Ω_n they form, stated Birkhoff's theorem that there are no other extreme points, proved that this is equivalent to "every doubly stochastic matrix is an average of permutation matrices", and left the proof to this chapter. This section supplies it. The proof is a count: a corner of Ω_n has so many zero entries that some row holds a single 1, and deleting that row and its column leaves a corner of Ω_{n-1}. We then bound how many permutation matrices an average needs, turn the theorem into a procedure for finding them, and use it to give the second proof of Schur's majorization theorem that Chapter 18 §08 sketched.

Throughout, the field is $\mathbb{R}$, and inequalities between matrices or vectors are entrywise, as in §01. Recall Definition 18.8.4: $\mathbf{M} \in M_n(\mathbb{R})$ is **doubly stochastic** if $m_{ij} \geq 0$ for all i, j, and every row and every column adds up to 1; in terms of the all-ones vector, $\mathbf{M1} = \mathbf{1}$ and $\mathbf{1}^\top\mathbf{M} = \mathbf{1}^\top$. So $\mathbf{M}$ is stochastic in the book's column sense (Definition 9.11.1), and so is $\mathbf{M}^\top$.

19.7.1 What is left to prove

Here is where Chapter 18 §08 stopped. The set Ω_n is convex and compact. The $n!$ permutation matrices $\mathbf{P}_\sigma$, $\sigma \in S_n$ (Definition 3.4.2), lie in it, and each is an extreme point (Proposition 18.8.9). The two forms of Birkhoff's theorem, "the extreme points are the permutation matrices" and "Ω_n is the convex hull of the permutation matrices", were shown to be equivalent there, using Minkowski's theorem (Theorem 18.8.5) in one direction and Proposition 18.8.2 in the other. For $n = 2$, the Quick check after the statement found Ω_2 by hand: it is the segment from $\mathbf{I}$ to the swap matrix.

So one implication is missing: **every extreme point of Ω_n is a permutation matrix.** For $n \geq 3$ nothing as simple as the 2×2 computation is available, since Ω_3 already has four free parameters. What does the work instead is a result of Chapter 18 §09 that recognizes a corner of a polyhedron from its tight constraints.

19.7.2 The doubly stochastic matrices form a polyhedron

Recall Definition 18.9.2 and Theorem 18.9.2. A polyhedron is a set $\{\mathbf{x} \in \mathbb{R}^N : \mathbf{Ax} \leq \mathbf{b}\}$; its extreme points are called vertices; and a point $\mathbf{x}$ of it is a vertex exactly when the rows $\mathbf{a}_i^\top$ of

the constraints that hold with equality at $\mathbf{x}$, the **active** rows, include N linearly independent ones. An equation is allowed, as a pair of inequalities $\mathbf{a}^\top\mathbf{x} \le \beta$ and $-\mathbf{a}^\top\mathbf{x} \le -\beta$; both are active at every point of the polyhedron.

To put Ω_n in this form, list the n^2 entries of a matrix $\mathbf{X}$ as one column, the vectorization $\operatorname{vec}\mathbf{X} \in \mathbb{R}^{n^2}$ of Chapter 8. A linear expression $\sum_{i,j} b_{ij}x_{ij}$ in the entries is then the dot product of $\operatorname{vec}\mathbf{X}$ with $\operatorname{vec}\mathbf{B}$, and we may treat the coefficient matrix $\mathbf{B} \in M_n(\mathbb{R})$ itself as the row of the constraint. Three families of coefficient matrices appear. With $\mathbf{E}_{ij}$ the matrix unit (a 1 in position (i,j), zeros elsewhere), let

$$\mathbf{R}_i = \mathbf{e}_i\mathbf{1}^\top, \qquad \mathbf{C}_j = \mathbf{1}\mathbf{e}_j^\top \qquad (1 \le i, j \le n),$$

the matrices whose i-th row, respectively j-th column, consists of ones and whose other entries are zero. The letter $\mathbf{C}_j$ is local to this section; elsewhere $\mathbf{C}(p)$ is a companion matrix and $\mathbf{C}_k(\lambda)$ a real Jordan block. Then $\sum_{k,l}(\mathbf{R}_i)_{kl}x_{kl}$ is the i-th row sum of $\mathbf{X}$, and $\mathbf{C}_j$ gives the j-th column sum. So $\Omega_n \subseteq M_n(\mathbb{R}) \cong \mathbb{R}^{n^2}$ is the polyhedron cut out by

(i) the n^2 inequalities $-x_{ij} \le 0$, with rows $-\mathbf{E}_{ij}$;

(ii) the n row equations "row sum i equals 1", with rows $\pm\mathbf{R}_i$;

(iii) the n column equations "column sum j equals 1", with rows $\pm\mathbf{C}_j$.

At a point $\mathbf{M} \in \Omega_n$, every row $\pm\mathbf{R}_i$ and $\pm\mathbf{C}_j$ is active, and $-\mathbf{E}_{ij}$ is active exactly when $m_{ij} = 0$. The whole argument turns on how much the $2n$ equations are worth. They are not independent: adding all row sums and adding all column sums both give the sum of all entries, that is,

$$\mathbf{R}_1 + \cdots + \mathbf{R}_n = \mathbf{J} = \mathbf{C}_1 + \cdots + \mathbf{C}_n,$$

with $\mathbf{J}$ the all-ones matrix. The next lemma says that this is the only relation.

Lemma 19.7.1. The Row and Column Sum Conditions

Let $n \ge 1$, and let

$$W_n = \{\mathbf{X} \in M_n(\mathbb{R}) : \mathbf{X}\mathbf{1} = \mathbf{0},\ \mathbf{1}^\top\mathbf{X} = \mathbf{0}^\top\}$$

be the space of real matrices whose rows and columns all add up to 0. Then:

(a) $\dim W_n = (n-1)^2$;

(b) the $2n$ matrices $\mathbf{R}_1, \dots, \mathbf{R}_n, \mathbf{C}_1, \dots, \mathbf{C}_n$ span a subspace of $M_n(\mathbb{R})$ of dimension $2n-1$.

Idea. A matrix in W_n is determined by its top-left $(n-1) \times (n-1)$ block, which can be anything: the last column and the last row are then forced by the zero sums, and the corner comes out consistent. That gives (a). For (b), W_n is exactly the set of matrices killed by the $2n$ sum conditions, so rank–nullity converts one count into the other.

Proof.

(a) If $n = 1$, $W_1 = \{\mathbf{O}\}$ and $(n-1)^2 = 0$. Let $n \ge 2$, and let $\varphi\colon W_n \to M_{n-1}(\mathbb{R})$ delete the last row and the last column; it is linear. It is injective: if $\mathbf{X} \in W_n$ has top-left block zero, then for $i < n$ the i-th row sum is $x_{in} = 0$, for $j < n$ the j-th column sum is $x_{nj} = 0$, and then the last row sum is $x_{nn} = 0$. So $\mathbf{X} = \mathbf{O}$. It is surjective: given $\mathbf{B} \in M_{n-1}(\mathbb{R})$, with row sums r_i, column sums c_j, and total $s = \sum_i r_i = \sum_j c_j$, let $\mathbf{X}$ have top-left block $\mathbf{B}$, last column $(-r_1, \dots, -r_{n-1}, s)$, and last row $(-c_1, \dots, -c_{n-1}, s)$.

The i-th row sum for $i < n$ is $r_i - r_i = 0$, the j-th column sum for $j < n$ is $c_j - c_j = 0$, and the last row and last column both sum to $-s + s = 0$. So $\mathbf{X} \in W_n$ and $\varphi(\mathbf{X}) = \mathbf{B}$. Hence φ is an isomorphism and $\dim W_n = \dim M_{n-1}(\mathbb{R}) = (n-1)^2$.

(b) Let $\mathbf{K}$ be the $2n \times n^2$ matrix whose rows are the transposed vectorizations $(\operatorname{vec} \mathbf{R}_i)^\top$ and $(\operatorname{vec} \mathbf{C}_j)^\top$, for $1 \le i, j \le n$. Then $\mathbf{K} \operatorname{vec} \mathbf{X}$ lists the row sums and the column sums of $\mathbf{X}$, so the null space of $\mathbf{K}$ is $\operatorname{vec}(W_n)$, of dimension $(n-1)^2$ by (a), as vec is an isomorphism. By Theorem 3.5.7, $\operatorname{rank} \mathbf{K} = n^2 - (n-1)^2 = 2n-1$, and by Theorem 3.5.5 this is the dimension of the span of the rows of $\mathbf{K}$. Since vec is an isomorphism, the $2n$ matrices $\mathbf{R}_i, \mathbf{C}_j$ span a subspace of dimension $2n-1$, as claimed.

■

So of the $n^2 + 4n$ constraints defining Ω_n, the $4n$ that come from equations contribute only $2n-1$ dimensions to the active rows, wherever we stand. The rest must come from zero entries.

Quick check. A vertex of Ω_3 needs 9 linearly independent active rows. How many zero entries must it therefore have, at least? Does the identity matrix $\mathbf{I}_3$ meet this requirement?

Solution. The row and column equations supply at most $2 \cdot 3 - 1 = 5$ independent rows, by Lemma 19.7.1 (b). Each zero entry adds one row $-\mathbf{E}_{ij}$, so at least $9 - 5 = 4$ zero entries are needed; that is, at most 5 nonzero entries. The identity has 6 zero entries and 3 nonzero ones, so it meets the requirement with room to spare.

19.7.3 Birkhoff's theorem

Now the main theorem. We state both forms, since both are used, but prove only the missing implication.

Theorem 19.7.2. Birkhoff's Theorem

Let $n \ge 1$. The extreme points of Ω_n are exactly the $n!$ permutation matrices of size n. Equivalently, every doubly stochastic matrix is a convex combination of permutation matrices:

$$\Omega_n = \operatorname{conv}\{\mathbf{P}_\sigma : \sigma \in S_n\}.$$

Idea. **Step roadmap.** Let $\mathbf{M}$ be an extreme point of Ω_n and induct on n.

① **Count.** $\mathbf{M}$ is a vertex of the polyhedron above, so it has n^2 independent active rows. At most $2n-1$ of them come from the equations, so at least $n^2 - (2n-1)$ entries of $\mathbf{M}$ are zero, and at most $2n-1$ are nonzero.

② **Pigeonhole.** Every row has a nonzero entry, since it adds up to 1. With n rows and fewer than $2n$ nonzero entries, some row has exactly one, and that entry is 1. Its column then has no room for anything else.

③ **Strip.** Delete that row and column. What is left is doubly stochastic of size $n-1$, and, the key point, it is again **extreme**: any way of splitting it would split $\mathbf{M}$.

④ **Induct.** The smaller matrix is a permutation matrix, and putting the 1 back gives one of size n.

Proof. By Proposition 18.8.9, every permutation matrix is an extreme point of Ω_n. Distinct permutations give distinct matrices, since column j of $\mathbf{P}_\sigma$ is $\mathbf{e}_{\sigma(j)}$, so there are $n!$ of them. The equivalence of the two forms was proved in Chapter 18 §08, from Theorem 18.8.5 and Proposition 18.8.2. It remains to show that every extreme point of Ω_n is a permutation matrix. We prove this by induction on $n \geq 1$.

For $n = 1$, $\Omega_1 = \{(1)\}$, and $(1) = \mathbf{P}_{\mathrm{id}}$.

Let $n \geq 2$, assume the statement for $n-1$, and let $\mathbf{M}$ be an extreme point of Ω_n.

Step 1: at most $2n-1$ nonzero entries. As shown above, Ω_n is a polyhedron in $M_n(\mathbb{R}) \cong \mathbb{R}^{n^2}$, and $\mathbf{M}$ is one of its vertices (Definition 18.9.2). By Theorem 18.9.2, the active rows at $\mathbf{M}$ include n^2 linearly independent ones. Let z be the number of zero entries of $\mathbf{M}$. The active rows are $\pm\mathbf{R}_i$, $\pm\mathbf{C}_j$, and $-\mathbf{E}_{ij}$ for the z positions with $m_{ij} = 0$. All of them lie in the subspace

$$\operatorname{span}\{\mathbf{R}_1, \dots, \mathbf{R}_n, \mathbf{C}_1, \dots, \mathbf{C}_n\} + \operatorname{span}\{\mathbf{E}_{ij} : m_{ij} = 0\},$$

whose dimension is at most $(2n-1)+z$, by Lemma 19.7.1 (b) and the dimension formula for sums (Theorem 1.7.3). A subspace containing n^2 linearly independent vectors has dimension at least n^2, by Theorem 1.6.3 (a) applied in that subspace. Hence $n^2 \leq 2n-1+z$, and the number of nonzero entries of $\mathbf{M}$ is $n^2 - z \leq 2n-1$.

Step 2: a row with a single 1. Each row of $\mathbf{M}$ adds up to 1, so it has at least one nonzero entry. If every row had at least two, $\mathbf{M}$ would have at least $2n$ nonzero entries, contradicting Step 1. So some row i has exactly one nonzero entry, say m_{ij}, and the row sum gives $m_{ij} = 1$. Column j consists of nonnegative entries adding up to 1, one of which is $m_{ij} = 1$; so $m_{kj} = 0$ for every $k \neq i$.

Step 3: the reduced matrix is extreme. Let $\mathbf{M}' \in M_{n-1}(\mathbb{R})$ be $\mathbf{M}$ with row i and column j deleted. Its entries are nonnegative. Each of its rows is a row $k \neq i$ of $\mathbf{M}$ with the entry $m_{kj} = 0$ removed, so it still adds up to 1; likewise each of its columns is a column $l \neq j$ of $\mathbf{M}$ with $m_{il} = 0$ removed. So $\mathbf{M}' \in \Omega_{n-1}$. For $\mathbf{Y} \in M_{n-1}(\mathbb{R})$, let $E(\mathbf{Y}) \in M_n(\mathbb{R})$ be the matrix obtained by inserting a new row at position i and a new column at position j, with a 1 where they cross and zeros elsewhere along them. Then $E(\mathbf{M}') = \mathbf{M}$; $E(\mathbf{Y}) \in \Omega_n$ whenever $\mathbf{Y} \in \Omega_{n-1}$, since the inserted row and column add up to 1 and add 0 to every old row and column; $E(\mathbf{Y}) = E(\mathbf{Z})$ only if $\mathbf{Y} = \mathbf{Z}$; and $E\big(\tfrac{1}{2}(\mathbf{Y}+\mathbf{Z})\big) = \tfrac{1}{2}(E(\mathbf{Y})+E(\mathbf{Z}))$, because the inserted entries are the same for $\mathbf{Y}$, $\mathbf{Z}$ and their midpoint. Now suppose $\mathbf{M}' = \tfrac{1}{2}(\mathbf{Y}+\mathbf{Z})$ with $\mathbf{Y}, \mathbf{Z} \in \Omega_{n-1}$. Then $\mathbf{M} = \tfrac{1}{2}(E(\mathbf{Y}) + E(\mathbf{Z}))$ with $E(\mathbf{Y}), E(\mathbf{Z}) \in \Omega_n$, and since $\mathbf{M}$ is extreme, $E(\mathbf{Y}) = E(\mathbf{Z})$, so $\mathbf{Y} = \mathbf{Z}$. By the midpoint test (Lemma 18.8.1), $\mathbf{M}'$ is an extreme point of Ω_{n-1}.

Step 4: induction. By the induction hypothesis, $\mathbf{M}'$ is a permutation matrix: it has exactly one 1 in each row and each column and zeros elsewhere. Then so does $\mathbf{M} = E(\mathbf{M}')$: row i and column j contain only the inserted 1, and every other row and column is one of $\mathbf{M}'$ with a 0 inserted. By the remark after Definition 3.4.2, $\mathbf{M}$ is a permutation matrix. This completes the induction and proves the theorem.

■

This pays Chapter 18 §08's promise. With Minkowski's theorem it describes Ω_n completely: a compact convex set with exactly $n!$ corners, each a permutation matrix.

Step 3 is not decoration. It is tempting to finish from Step 1 alone, by an induction on the number of nonzero entries, and that induction does not close.

Warning. **Few nonzero entries do not by themselves make a permutation matrix.** The matrix

$$\mathbf{N} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & \frac{1}{2} & \frac{1}{2} \end{pmatrix} \in \Omega_3$$

has $5 = 2{\cdot}3{-}1$ nonzero entries, the most Step 1 allows, and it is not a permutation matrix; it is the midpoint of $\mathbf{I}_3$ and the permutation matrix that swaps the last two coordinates. Deleting its first row and column leaves $\frac{1}{2}\mathbf{J}_2$, with 4 nonzero entries, more than the $2 \cdot 2 - 1 = 3$ that a vertex of Ω_2 may have. The count must be applied afresh at each size, and that needs the smaller matrix to be extreme, which is what Step 3 proves.

19.7.4 How many permutation matrices

Birkhoff's theorem writes every doubly stochastic matrix as an average of permutation matrices, but a priori with up to $n!$ terms. Carathéodory's theorem (Theorem 18.2.2) cuts this down to one more than the dimension, and the dimension is the one computed in Lemma 19.7.1.

Corollary 19.7.3. Birkhoff's Theorem with a Count

Let $n \geq 1$. Every $\mathbf{M} \in \Omega_n$ is a convex combination of at most $(n-1)^2 + 1$ permutation matrices.

Idea. All of Ω_n lies in the coset of matrices with row and column sums 1, which is $\frac{1}{n}\mathbf{J} + W_n$, of dimension $(n-1)^2$. Translate to W_n and apply Carathéodory's theorem there.

Proof. Let $\mathbf{M} \in \Omega_n$. By Theorem 19.7.2 and Theorem 18.1.3, $\mathbf{M} = \sum_{k=1}^{m} t_k \mathbf{P}_k$ for some permutation matrices $\mathbf{P}_k$ and weights $(t_1, \ldots, t_m) \in \Delta_m$. Put $\mathbf{X}_0 = \frac{1}{n}\mathbf{J} \in \Omega_n$. Every doubly stochastic matrix $\mathbf{X}$ satisfies $(\mathbf{X}-\mathbf{X}_0)\mathbf{1} = \mathbf{1}-\mathbf{1} = \mathbf{0}$ and $\mathbf{1}^\top(\mathbf{X}-\mathbf{X}_0) = \mathbf{0}^\top$, so $\mathbf{X}-\mathbf{X}_0 \in W_n$. As the weights add up to 1,

$$\mathbf{M} - \mathbf{X}_0 = \sum_{k=1}^{m} t_k(\mathbf{P}_k - \mathbf{X}_0),$$

so $\mathbf{M} - \mathbf{X}_0$ lies in the convex hull of $S = \{\mathbf{P}_\sigma - \mathbf{X}_0 : \sigma \in S_n\}$, a subset of the real vector space W_n, which has dimension $(n-1)^2$ by Lemma 19.7.1 (a). By Theorem 18.2.2, $\mathbf{M} - \mathbf{X}_0 = \sum_{k=1}^{r} s_k(\mathbf{P}'_k - \mathbf{X}_0)$ for some $r \leq (n-1)^2 + 1$, permutation matrices $\mathbf{P}'_k$, and weights $(s_1, \ldots, s_r) \in \Delta_r$. Adding $\mathbf{X}_0 = \sum_k s_k \mathbf{X}_0$ back gives $\mathbf{M} = \sum_{k=1}^{r} s_k \mathbf{P}'_k$, as claimed.

■

For $n = 2$ the bound is 2, and every point strictly inside the segment Ω_2 needs both ends. For $n = 3$, Exercise C1 shows that 5 can be needed. That $(n-1)^2 + 1$ is best possible for every $n \geq 2$ is a known result which we state without proof. Compare the density matrices of Chapter 18 §08, where the spectral theorem did much better than Carathéodory's count. Here there is no such structure to exploit.

19.7.5 Finding the permutations

The proof of Birkhoff's theorem says that the permutation matrices suffice. It does not say how to find them. One consequence of the theorem does, because it only concerns which entries are zero.

Corollary 19.7.4. A Positive Diagonal

Let $\mathbf{M} \in \Omega_n$. Then there is a permutation $\sigma \in S_n$ with $m_{\sigma(j)j} > 0$ for **every** $j = 1, \dots, n$.

Proof. By Theorem 19.7.2, $\mathbf{M} = \sum_{\tau \in S_n} t_\tau \mathbf{P}_\tau$ with every $t_\tau \geq 0$ and $\sum_\tau t_\tau = 1$, so some $t_\sigma > 0$. The $(\sigma(j), j)$ entry of $\mathbf{P}_\sigma$ is 1, and every term of the sum has nonnegative entries. Hence $m_{\sigma(j)j} \geq t_\sigma > 0$ for every j. ∎

In words, one can pick one positive entry in each column, all in different rows. This gives a procedure, often called **peeling**. Given $\mathbf{M} \in \Omega_n$, choose σ as in the corollary and let $t = \min_j m_{\sigma(j)j}$, so $0 < t \leq 1$.

- If $t = 1$, then each column j has its whole sum 1 at position $\sigma(j)$, and $\mathbf{M} = \mathbf{P}_\sigma$.
- If $t < 1$, then $\mathbf{M} - t\mathbf{P}_\sigma$ has nonnegative entries and all row and column sums $1 - t$, so $\mathbf{M}_1 = (\mathbf{M} - t\mathbf{P}_\sigma)/(1-t) \in \Omega_n$ and $\mathbf{M} = t\mathbf{P}_\sigma + (1-t)\mathbf{M}_1$. Every zero of $\mathbf{M}$ is a zero of $\mathbf{M}_1$, since $\mathbf{P}_\sigma$ has its 1's at positive entries of $\mathbf{M}$, and the entry where the minimum t was attained becomes a new zero.

Repeat with $\mathbf{M}_1$, producing $\mathbf{M}_2, \mathbf{M}_3, \dots$, and write $\mathbf{M}_0 = \mathbf{M}$. A doubly stochastic matrix has at least n nonzero entries, one per column, and each round removes at least one, so the procedure stops after at most $n^2 - n + 1$ rounds, with a permutation matrix. Unwinding the rounds writes $\mathbf{M}$ as a convex combination of the permutation matrices found. In fact peeling never needs more than the $(n-1)^2 + 1$ of Corollary 19.7.3. Let U_k be the subspace of those $\mathbf{X} \in W_n$ with $x_{ij} = 0$ wherever $\mathbf{M}_k$ has a zero; since zeros of $\mathbf{M}_k$ stay zeros, $U_{k+1} \subseteq U_k$, and $\mathbf{M}_k - \mathbf{M}_{k+1}$ lies in U_k but not in U_{k+1}, being positive at the new zero. So each round with $t < 1$ lowers $\dim U_k$, which starts at most $\dim W_n = (n-1)^2$ by Lemma 19.7.1 (a) and never drops below 0; there are at most $(n-1)^2$ such rounds, and one final round. Still, peeling finds **a** decomposition, not necessarily the shortest one.

Example. Peeling a 4 × 4 matrix

Write

$$\mathbf{M} = \frac{1}{8}\begin{pmatrix} 4 & 2 & 2 & 0 \\ 1 & 3 & 0 & 4 \\ 3 & 0 & 4 & 1 \\ 0 & 3 & 2 & 3 \end{pmatrix}$$

as a convex combination of permutation matrices.

Solution. The entries are nonnegative, each row of $8\mathbf{M}$ adds up to 8, and so does each column: $4+1+3+0$, $2+3+0+3$, $2+0+4+2$, $0+4+1+3$. So $\mathbf{M} \in \Omega_4$. We peel $\mathbf{N} = 8\mathbf{M}$, whose row and column sums start at 8; subtracting w times a permutation matrix lowers them all by w. A permutation is recorded by the positions $(\sigma(1), 1), \ldots, (\sigma(4), 4)$ of its 1's, one per column.

Round 1. The diagonal of $\mathbf{N}$ is $4, 3, 4, 3$, all positive, so take $\sigma_1 = \mathrm{id}$ with weight $w_1 = 3$. Subtracting $3\mathbf{I}$ leaves

$$\begin{pmatrix} 1 & 2 & 2 & 0 \\ 1 & 0 & 0 & 4 \\ 3 & 0 & 1 & 1 \\ 0 & 3 & 2 & 0 \end{pmatrix} \quad \text{(sums 5).}$$

Round 2. Column 2 is now nonzero only in rows 1 and 4. Try row 4 there: the positions $(1,1)$, $(4,2)$, $(3,3)$, $(2,4)$ carry $1, 3, 1, 4$, so take this σ_2 with weight $w_2 = 1$:

$$\begin{pmatrix} 0 & 2 & 2 & 0 \\ 1 & 0 & 0 & 3 \\ 3 & 0 & 0 & 1 \\ 0 & 2 & 2 & 0 \end{pmatrix} \quad \text{(sums 4).}$$

Round 3. The positions $(3,1), (1,2), (4,3), (2,4)$ carry $3, 2, 2, 3$; weight $w_3 = 2$:

$$\begin{pmatrix} 0 & 0 & 2 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 2 & 0 & 0 \end{pmatrix} \quad \text{(sums 2).}$$

Round 4. The positions $(2,1), (4,2), (1,3), (3,4)$ carry $1, 2, 2, 1$; weight $w_4 = 1$. What remains has a single 1 in positions $(3,1), (4,2), (1,3), (2,4)$, a permutation matrix $\mathbf{P}_{\sigma_5}$, with weight $w_5 = 1$.

The weights $3 + 1 + 2 + 1 + 1 = 8$ account for the whole sum, and dividing by 8,

$$\mathbf{M} = \tfrac{3}{8}\mathbf{P}_{\sigma_1} + \tfrac{1}{8}\mathbf{P}_{\sigma_2} + \tfrac{2}{8}\mathbf{P}_{\sigma_3} + \tfrac{1}{8}\mathbf{P}_{\sigma_4} + \tfrac{1}{8}\mathbf{P}_{\sigma_5},$$

where $\sigma_1 = \mathrm{id}$, and $\sigma_2, \sigma_3, \sigma_4, \sigma_5$ send $1, 2, 3, 4$ to $1, 4, 3, 2$, to $3, 1, 4, 2$, to $2, 4, 1, 3$, and to $3, 4, 1, 2$ respectively. Five terms, within the bound $(4-1)^2 + 1 = 10$. To check, add up any entry: position $(1, 3)$ is covered by σ_4 and σ_5, giving $\frac{1}{8} + \frac{1}{8} = \frac{2}{8}$, which is m_{13}.

Remark. The route can be run the other way. Corollary 19.7.4 is a statement about the zero pattern of $\mathbf{M}$ alone, and it can be proved directly by combinatorics: by Hall's marriage theorem, or by the Frobenius–König theorem, which says that every $n \times n$ matrix in which each choice of one entry per column, in distinct rows, meets a zero has an $r \times s$ block of zeros with $r + s = n + 1$. With the corollary in hand first, peeling proves Birkhoff's theorem constructively. Both routes are standard; the one taken here uses only what Chapter 18 had

already built.

19.7.6 Schur's theorem again

Chapter 17 §08 proved Schur's majorization theorem, Theorem 17.8.2: for a Hermitian $\mathbf{A}$, the diagonal vector $\mathbf{d}(\mathbf{A})$ is majorized by the eigenvalue vector $\boldsymbol{\lambda}(\mathbf{A})$. The proof there went through Ky Fan's maximum principle. Chapter 18 §08 noted that Birkhoff's theorem would give a second proof, and sketched it. Here it is in full. The first ingredient says that the diagonal is a doubly stochastic image of the spectrum.

Lemma 19.7.5. The Diagonal Is a Doubly Stochastic Image of the Spectrum

Let $F = \mathbb{R}$ or $\mathbb{C}$, let $\mathbf{A} \in M_n(F)$ be Hermitian, and let $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{U}^*$ with $\mathbf{U}$ unitary and $\mathbf{D} = \operatorname{diag}(\lambda_1(\mathbf{A}), \dots, \lambda_n(\mathbf{A}))$. Then the matrix $\mathbf{S} = (|u_{ij}|^2)$ is doubly stochastic, and

$$\mathbf{d}(\mathbf{A}) = \mathbf{S}\,\boldsymbol{\lambda}(\mathbf{A}).$$

Such a $\mathbf{U}$ exists by Corollary 12.4.3, after reordering the columns of $\mathbf{U}$, since a Hermitian matrix is normal: the diagonal entries of the $\mathbf{D}$ it gives are the eigenvalues of $\mathbf{A}$ with multiplicity, because similar matrices have the same characteristic polynomial $p_{\mathbf{A}}$ (Theorem 7.10.3), and they are real (Theorem 12.2.1), so they can be put in decreasing order.

Proof. Write $\lambda_j = \lambda_j(\mathbf{A})$. By the definition of the matrix product, $a_{ii} = \sum_j u_{ij}\lambda_j\overline{u_{ij}} = \sum_j |u_{ij}|^2\lambda_j$, which is the i-th entry of $\mathbf{S}\boldsymbol{\lambda}(\mathbf{A})$. The entries of $\mathbf{S}$ are nonnegative. Row i of $\mathbf{S}$ adds up to $\sum_j |u_{ij}|^2 = (\mathbf{U}\mathbf{U}^*)_{ii} = 1$, and column j adds up to $\sum_i |u_{ij}|^2 = (\mathbf{U}^*\mathbf{U})_{jj} = 1$, since $\mathbf{U}$ is unitary. So $\mathbf{S} \in \Omega_n$. ∎

A doubly stochastic matrix of the form $(|u_{ij}|^2)$ is called **unistochastic**. Not every doubly stochastic matrix is one (Exercise C3 asks for a witness), but the lemma needs only that $\mathbf{S} \in \Omega_n$. The second ingredient is Birkhoff's theorem, which makes $\mathbf{S}\boldsymbol{\lambda}$ an average of rearrangements of $\boldsymbol{\lambda}$.

Proposition 19.7.6. Doubly Stochastic Maps Even Out

Let $\mathbf{S} \in \Omega_n$ and $\mathbf{y} \in \mathbb{R}^n$. Then $\mathbf{S}\mathbf{y} \prec \mathbf{y}$.

Idea. A permutation matrix only rearranges $\mathbf{y}$, and rearrangements are invisible to majorization. Birkhoff's theorem writes $\mathbf{S}\mathbf{y}$ as an average of rearrangements, and the vectors majorized by $\mathbf{y}$ form a convex set.

Proof. By Theorem 19.7.2, $\mathbf{S} = \sum_{k=1}^{m} t_k \mathbf{P}_{\sigma_k}$ with $(t_1, \dots, t_m) \in \Delta_m$, so $\mathbf{S}\mathbf{y} = \sum_k t_k \mathbf{P}_{\sigma_k}\mathbf{y}$. Each $\mathbf{P}_\sigma\mathbf{y} = \sum_j y_j \mathbf{e}_{\sigma(j)}$ has entry y_j in position $\sigma(j)$, so it is a rearrangement of $\mathbf{y}$. As observed after Definition 17.8.1, majorization sees only which numbers occur and how often, so $\mathbf{y} \prec \mathbf{y}$ (Proposition 17.8.1 (a)) gives $\mathbf{P}_\sigma\mathbf{y} \prec \mathbf{y}$. By Example (The vectors majorized

by a fixed vector) the set $\{\mathbf{x} : \mathbf{x} \prec \mathbf{y}\}$ is convex, and a convex set contains every convex combination of its points (Lemma 18.1.2). Hence $\mathbf{Sy} \prec \mathbf{y}$.

■

Second proof of Theorem 17.8.2. Let $\mathbf{A}$ be Hermitian. By Lemma 19.7.5, $\mathbf{d}(\mathbf{A}) = \mathbf{S}\boldsymbol{\lambda}(\mathbf{A})$ with $\mathbf{S} \in \Omega_n$, and by Proposition 19.7.6, $\mathbf{S}\boldsymbol{\lambda}(\mathbf{A}) \prec \boldsymbol{\lambda}(\mathbf{A})$. So $\mathbf{d}(\mathbf{A}) \prec \boldsymbol{\lambda}(\mathbf{A})$. □

The first proof compared running totals using a variational principle; this one never compares running totals directly. It sees Schur's theorem as a statement about averaging: the diagonal entries are weighted averages of the eigenvalues, with weights that form a doubly stochastic matrix, and averaging can only even out a vector.

Example. The averaging made visible

For

$$\mathbf{A} = \begin{pmatrix} 3 & 2 & 0 \\ 2 & 4 & 2 \\ 0 & 2 & 5 \end{pmatrix}$$

of Chapter 17 §08, whose eigenvalues are $\boldsymbol{\lambda}(\mathbf{A}) = (7, 4, 1)$ with unit eigenvectors $\mathbf{u}_1 = \frac{1}{3}(1, 2, 2)$, $\mathbf{u}_2 = \frac{1}{3}(-2, -1, 2)$, $\mathbf{u}_3 = \frac{1}{3}(2, -2, 1)$ (Example (Coordinate vectors against eigenvectors)), compute $\mathbf{S}$, check $\mathbf{d}(\mathbf{A}) = \mathbf{S}\boldsymbol{\lambda}(\mathbf{A})$, and write $\mathbf{d}(\mathbf{A})$ as an average of rearrangements of $\boldsymbol{\lambda}(\mathbf{A})$.

Solution. Take $\mathbf{U}$ with columns $\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3$; it is orthogonal and $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{U}^{\top}$. Squaring its entries,

$$\mathbf{U} = \frac{1}{3}\begin{pmatrix} 1 & -2 & 2 \\ 2 & -1 & -2 \\ 2 & 2 & 1 \end{pmatrix}, \qquad \mathbf{S} = \frac{1}{9}\begin{pmatrix} 1 & 4 & 4 \\ 4 & 1 & 4 \\ 4 & 4 & 1 \end{pmatrix}.$$

Every row and column of $9\mathbf{S}$ adds up to 9. Then

$$\mathbf{S}\boldsymbol{\lambda}(\mathbf{A}) = \frac{1}{9}\begin{pmatrix} 7+16+4 \\ 28+4+4 \\ 28+16+1 \end{pmatrix} = \begin{pmatrix} 3 \\ 4 \\ 5 \end{pmatrix} = \mathbf{d}(\mathbf{A}).$$

To decompose $\mathbf{S}$, let $\mathbf{P}$ be the permutation matrix of $\sigma = (1\ 2\ 3)$, $\sigma(1) = 2$, $\sigma(2) = 3$, $\sigma(3) = 1$. Then $\mathbf{P} + \mathbf{P}^2 = \mathbf{J} - \mathbf{I}$, since $\mathbf{P}$ and $\mathbf{P}^2$ have their 1's in the off-diagonal positions and never in the same place, so

$$\mathbf{S} = \tfrac{1}{9}\,\mathbf{I} + \tfrac{4}{9}\,\mathbf{P} + \tfrac{4}{9}\,\mathbf{P}^2.$$

With $\mathbf{P}\boldsymbol{\lambda} = (1, 7, 4)$ and $\mathbf{P}^2\boldsymbol{\lambda} = (4, 1, 7)$, the cyclic shifts of $(7, 4, 1)$,

$$\mathbf{d}(\mathbf{A}) = \tfrac{1}{9}(7, 4, 1) + \tfrac{4}{9}(1, 7, 4) + \tfrac{4}{9}(4, 1, 7) = (3, 4, 5).$$

Each diagonal entry is an average of all three eigenvalues, with weight $\frac{1}{9}$ on one of them and $\frac{4}{9}$ on the other two. That is why the diagonal $(3, 4, 5)$ is so much flatter than $(7, 4, 1)$.

Together with Theorem 17.8.5, the lemma and the proposition already show that the diagonals

of the Hermitian matrices with spectrum $\boldsymbol{\lambda}$ are exactly the vectors $\mathbf{S}\boldsymbol{\lambda}$ with $\mathbf{S} \in \Omega_n$. The lemma puts every such diagonal among the $\mathbf{S}\boldsymbol{\lambda}$; the proposition puts every $\mathbf{S}\boldsymbol{\lambda}$ among the $\mathbf{x} \prec \boldsymbol{\lambda}$; and Schur–Horn realizes every $\mathbf{x} \prec \boldsymbol{\lambda}$ as a diagonal. So all three sets coincide, and in particular every $\mathbf{x} \prec \boldsymbol{\lambda}$ is $\mathbf{S}\boldsymbol{\lambda}$ for some $\mathbf{S} \in \Omega_n$. This converse of the proposition is part of the theorem of Hardy, Littlewood, and Pólya, which Chapter 21 takes up.

19.7.7 Exercises

19.7.7.1 A. Check your understanding

Exercise. A1

(a) State Birkhoff's theorem in both forms.

(b) Name the results of Chapter 18 that the proof of Theorem 19.7.2 uses, and say what each contributes.

(c) What upper bound does Step 1 of the proof give for the number of nonzero entries of a vertex of Ω_4? How many does a vertex actually have?

(d) True or false: every matrix in Ω_n with at most $2n-1$ nonzero entries is a permutation matrix. Justify your answer.

(e) True or false: the product of two doubly stochastic matrices is doubly stochastic. Justify your answer.

(f) True or false: every $\mathbf{M} \in \Omega_3$ is a convex combination of at most five permutation matrices. Justify your answer.

Solution.

(a) The extreme points of Ω_n are exactly the $n!$ permutation matrices; equivalently, Ω_n is the convex hull of the permutation matrices, that is, every doubly stochastic matrix is a convex combination of them.

(b) Theorem 18.9.2, which says that a vertex has n^2 independent active constraints and so forces many zero entries; Lemma 18.8.1, the midpoint test, used to show that the stripped matrix is again extreme; and Proposition 18.8.9 with the equivalence of the two forms, which rests on Theorem 18.8.5 and Proposition 18.8.2.

(c) At most $2\cdot 4-1=7$. By Theorem 19.7.2 a vertex is a permutation matrix, with exactly 4 nonzero entries. The bound of Step 1 is far from sharp; it is only the first step of an induction.

(d) False. The matrix $\mathbf{N}$ in the warning after Theorem 19.7.2 lies in Ω_3 and has $5 = 2\cdot 3-1$ nonzero entries, but it is not a permutation matrix.

(e) True. If $\mathbf{A}, \mathbf{B} \in \Omega_n$, the entries of $\mathbf{AB}$ are sums of products of nonnegative numbers, and $\mathbf{AB1} = \mathbf{A1} = \mathbf{1}$, $\mathbf{1}^{\top}\mathbf{AB} = \mathbf{1}^{\top}\mathbf{B} = \mathbf{1}^{\top}$.

(f) True, by Corollary 19.7.3 with $n = 3$: $(3-1)^2 + 1 = 5$.

19.7.7.2 B. Practice

Exercise. B1: Which are corners?

For each matrix, determine whether it lies in Ω_3 and, if so, whether it is an extreme point

of Ω_3. Justify your answers.

$$\mathbf{M}_1 = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}, \qquad \mathbf{M}_2 = \frac{1}{2}\begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{pmatrix},$$

$$\mathbf{M}_3 = \frac{1}{2}\begin{pmatrix} 2 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}, \qquad \mathbf{M}_4 = \begin{pmatrix} 1 & 1 & -1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}.$$

Solution. $\mathbf{M}_1$ has exactly one 1 in each row and each column and zeros elsewhere, so it is a permutation matrix, in Ω_3 and extreme by Proposition 18.8.9.

$\mathbf{M}_2$ has nonnegative entries and every row and column of $2\mathbf{M}_2$ adds up to 2, so $\mathbf{M}_2 \in \Omega_3$. It is not a permutation matrix, so by Theorem 19.7.2 it is not extreme. Explicitly, $\mathbf{M}_2 = \frac{1}{2}\mathbf{I} + \frac{1}{2}\mathbf{P}$ with $\mathbf{P}$ the permutation matrix with 1's at $(1,2), (2,3), (3,1)$, and $\mathbf{I} \neq \mathbf{P}$.

$\mathbf{M}_3$ is not in Ω_3: its rows add up to $1, 1, \frac{1}{2}$.

$\mathbf{M}_4$ is not in Ω_3: its rows and columns add up to 1 (rows $1,1,1$; columns $1,1,1$), but the entry -1 is negative. It is in the coset $\frac{1}{3}\mathbf{J} + W_3$ of the proof of Corollary 19.7.3, which shows that the nonnegativity conditions matter.

Exercise. B2: Peeling a 3×3 matrix

Let

$$\mathbf{M} = \frac{1}{6}\begin{pmatrix} 3 & 3 & 0 \\ 1 & 2 & 3 \\ 2 & 1 & 3 \end{pmatrix}.$$

Check that $\mathbf{M} \in \Omega_3$, and write $\mathbf{M}$ as a convex combination of permutation matrices by peeling. Hence confirm the bound of Corollary 19.7.3 for this $\mathbf{M}$.

Solution. The entries are nonnegative; the rows of $6\mathbf{M}$ add up to $3+3+0$, $1+2+3$, $2+1+3$, and the columns to $3+1+2$, $3+2+1$, $0+3+3$, all 6. So $\mathbf{M} \in \Omega_3$. Peel $\mathbf{N} = 6\mathbf{M}$.

Round 1. The diagonal $3, 2, 3$ is positive; subtract $2\mathbf{I}$:

$$\begin{pmatrix} 1 & 3 & 0 \\ 1 & 0 & 3 \\ 2 & 1 & 1 \end{pmatrix} \quad \text{(sums 4)}.$$

Round 2. The positions $(3,1), (1,2), (2,3)$ carry $2, 3, 3$; subtract $2\mathbf{P}_1$, where $\mathbf{P}_1$ has its 1's there:

$$\begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix} \quad \text{(sums 2)}.$$

Round 3. The positions $(1,1), (3,2), (2,3)$ carry $1, 1, 1$; subtract $\mathbf{P}_2$, with 1's there. What remains has 1's at $(2,1), (1,2), (3,3)$: a permutation matrix $\mathbf{P}_3$, with weight 1.

So $6\mathbf{M} = 2\mathbf{I} + 2\mathbf{P}_1 + \mathbf{P}_2 + \mathbf{P}_3$ and

$$\mathbf{M} = \tfrac{1}{3}\,\mathbf{I} + \tfrac{1}{3}\,\mathbf{P}_1 + \tfrac{1}{6}\,\mathbf{P}_2 + \tfrac{1}{6}\,\mathbf{P}_3,$$

with weights nonnegative and adding up to 1. Hence $\mathbf{M}$ uses four permutation matrices, within the bound $(3-1)^2 + 1 = 5$.

Exercise. B3: Schur's theorem by averaging

Let $\mathbf{A} = \begin{pmatrix} 4 & 2 \\ 2 & 1 \end{pmatrix}$.

(a) Find $\boldsymbol{\lambda}(\mathbf{A})$ and an orthogonal $\mathbf{U}$ with $\mathbf{A} = \mathbf{U}\,\mathrm{diag}(\boldsymbol{\lambda}(\mathbf{A}))\mathbf{U}^\top$.

(b) Compute $\mathbf{S} = (u_{ij}^2)$ and check that $\mathbf{d}(\mathbf{A}) = \mathbf{S}\boldsymbol{\lambda}(\mathbf{A})$.

(c) Write $\mathbf{S}$ as a convex combination of permutation matrices, and hence $\mathbf{d}(\mathbf{A})$ as an average of rearrangements of $\boldsymbol{\lambda}(\mathbf{A})$. Confirm $\mathbf{d}(\mathbf{A}) \prec \boldsymbol{\lambda}(\mathbf{A})$ from Definition 17.8.1.

Solution.

(a) $\mathbf{A}(2,1) = (10,5) = 5(2,1)$ and $\mathbf{A}(1,-2) = (0,0)$, so $\boldsymbol{\lambda}(\mathbf{A}) = (5,0)$. With the unit eigenvectors $\frac{1}{\sqrt{5}}(2,1)$ and $\frac{1}{\sqrt{5}}(1,-2)$ as columns,

$$\mathbf{U} = \frac{1}{\sqrt{5}}\begin{pmatrix} 2 & 1 \\ 1 & -2 \end{pmatrix},$$

which is orthogonal, and $\mathbf{A} = \mathbf{U}\,\mathrm{diag}(5,0)\mathbf{U}^\top$ by Corollary 12.5.2.

(b) $\mathbf{S} = \frac{1}{5}\begin{pmatrix} 4 & 1 \\ 1 & 4 \end{pmatrix}$, and $\mathbf{S}(5,0) = (4,1) = \mathbf{d}(\mathbf{A})$.

(c) $\mathbf{S} = \frac{4}{5}\mathbf{I} + \frac{1}{5}\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$, so $\mathbf{d}(\mathbf{A}) = \frac{4}{5}(5,0) + \frac{1}{5}(0,5) = (4,1)$. Directly: both vectors are decreasing, the totals are $5 = 5$, and at $k = 1$, $4 \le 5$. So $\mathbf{d}(\mathbf{A}) \prec \boldsymbol{\lambda}(\mathbf{A})$.

19.7.7.3 C. Going deeper

Exercise. C1: Five permutation matrices can be needed

Let $\mathbf{P}$ be the permutation matrix of the 3-cycle $1 \mapsto 2 \mapsto 3 \mapsto 1$, and let $\mathbf{T}_{12}, \mathbf{T}_{13}, \mathbf{T}_{23}$ be those of the transpositions $(1\ 2), (1\ 3), (2\ 3)$. The six permutation matrices of size 3 are $\mathbf{I}, \mathbf{P}, \mathbf{P}^2$ (the **even** ones) and $\mathbf{T}_{12}, \mathbf{T}_{13}, \mathbf{T}_{23}$ (the **odd** ones).

(a) Show that each of the nine positions (i,j) holds a 1 in exactly one even and exactly one odd permutation matrix, and that each even and each odd permutation matrix share exactly one position of a 1.

(b) Deduce that if $\sum_\sigma c_\sigma \mathbf{P}_\sigma = \mathbf{O}$ for real numbers c_σ, then $c_\sigma = c$ for the even σ and $c_\sigma = -c$ for the odd σ, for some $c \in \mathbb{R}$.

(c) Let $\mathbf{M} = \frac{1}{15}(\mathbf{I} + 2\mathbf{P} + 3\mathbf{P}^2 + 4\mathbf{T}_{12} + 5\mathbf{T}_{13})$. Show that every way of writing $\mathbf{M}$ as a convex combination of permutation matrices uses at least five of them with positive

weight.

Hint: in (a), a permutation of $\{1,2,3\}$ is determined by where it sends any two points.

Solution.

(a) A permutation matrix $\mathbf{P}_\sigma$ has a 1 in position (i,j) exactly when $\sigma(j) = i$. The permutations of $\{1,2,3\}$ with $\sigma(j) = i$ are determined by where they send the other two points, which must go to the other two values in one of two ways, so there are exactly two. If σ is one of them, the other is $\tau\sigma$, with τ the transposition exchanging the two remaining values, so their signs are opposite (Definition 7.3.2, Theorem 7.3.1): one is even and one is odd. For the second claim, let σ be even and τ odd. They cannot agree at all three points, since they differ; they cannot agree at exactly two, since two agreements force the third; and if they agreed nowhere, $\tau^{-1}\sigma$ would fix no point and so be a 3-cycle, which is even, while $\tau^{-1}\sigma$ is odd. So they agree at exactly one point j, that is, they share exactly one position $(\sigma(j), j)$.

(b) Look at position (i,j): by (a) exactly one even σ and one odd τ have a 1 there, so the equation gives $c_\sigma + c_\tau = 0$. By (a) again, every pair of an even σ and an odd τ arises in this way at their shared position, so $c_\sigma = -c_\tau$ for every such pair. Fixing one odd τ, all even c_σ equal $c = -c_\tau$, and then every odd $c_{\tau'} = -c$.

(c) The given weights are $\frac{1}{15}(1,2,3)$ on $\mathbf{I}, \mathbf{P}, \mathbf{P}^2$ and $\frac{1}{15}(4,5,0)$ on $\mathbf{T}_{12}, \mathbf{T}_{13}, \mathbf{T}_{23}$; they are nonnegative with sum 1. If $\mathbf{M} = \sum_\sigma t_\sigma \mathbf{P}_\sigma$ is any other convex combination, the difference of the two weight lists satisfies (b), so the even weights are $\frac{1}{15}(1,2,3) + s$ and the odd ones $\frac{1}{15}(4,5,0) - s$ for some real s. Nonnegativity needs $s \geq -\frac{1}{15}$ (from $\mathbf{I}$) and $s \leq 0$ (from $\mathbf{T}_{23}$). For $-\frac{1}{15} < s < 0$, all six weights are positive. At $s = 0$ exactly one vanishes, that of $\mathbf{T}_{23}$; at $s = -\frac{1}{15}$ exactly one vanishes, that of $\mathbf{I}$, the others being $\frac{1}{15}(1,2)$ and $\frac{1}{15}(5,6,1)$. So every representation has at least five positive weights, and the bound $(3-1)^2 + 1 = 5$ of Corollary 19.7.3 is attained.

Exercise. C2: Integer matrices with equal line sums

Let $k \geq 1$, and let $\mathbf{N} \in M_n(\mathbb{R})$ have nonnegative **integer** entries, with every row and every column adding up to k.

(a) Prove that $\mathbf{N}$ is a sum of k permutation matrices, not necessarily distinct.

(b) Carry this out for $\mathbf{N} = \begin{pmatrix} 2 & 1 & 0 \\ 0 & 1 & 2 \\ 1 & 1 & 1 \end{pmatrix}$.

Hint: apply Corollary 19.7.4 to $\frac{1}{k}\mathbf{N}$.

Solution.

(a) By induction on k. The matrix $\frac{1}{k}\mathbf{N}$ is doubly stochastic, so by Corollary 19.7.4 there is σ with $n_{\sigma(j)j} > 0$ for every j. These are positive integers, so each is at least 1, and $\mathbf{N} - \mathbf{P}_\sigma$ has nonnegative integer entries with every row and column adding up to $k-1$. If $k = 1$, then $\mathbf{N} - \mathbf{P}_\sigma$ has line sums 0 and nonnegative entries, so it is $\mathbf{O}$ and $\mathbf{N} = \mathbf{P}_\sigma$. If $k \geq 2$, the induction hypothesis writes $\mathbf{N} - \mathbf{P}_\sigma$ as a sum of $k-1$

permutation matrices, and adding $\mathbf{P}_\sigma$ gives k.

(b) Every row and column adds up to 3. The diagonal $2, 1, 1$ is positive; subtracting $\mathbf{I}$ leaves

$$\begin{pmatrix} 1 & 1 & 0 \\ 0 & 0 & 2 \\ 1 & 1 & 0 \end{pmatrix}.$$

The positions $(1,1), (3,2), (2,3)$ are positive; subtracting that permutation matrix leaves the permutation matrix with 1's at $(3,1), (1,2), (2,3)$. So $\mathbf{N}$ is $\mathbf{I}$ plus these two permutation matrices.

Exercise. C3: The dimension, and a matrix that is not unistochastic

Let $n \geq 2$.

(a) Prove that if $\mathbf{X} \in W_n$ has every $|x_{ij}| \leq \frac{1}{n}$, then $\frac{1}{n}\mathbf{J} + \mathbf{X} \in \Omega_n$. Deduce that $\operatorname{aff} \Omega_n = \frac{1}{n}\mathbf{J} + W_n$ and $\dim \Omega_n = (n-1)^2$.

(b) Show that $\mathbf{M} = \frac{1}{2}\begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{pmatrix}$ is doubly stochastic but is not of the form $(|u_{ij}|^2)$ for any unitary $\mathbf{U} \in M_3(\mathbb{C})$.

Hint for (b): two columns of $\mathbf{U}$ *must be orthogonal; look at which rows both can be nonzero in.*

Solution.

(a) The entries of $\frac{1}{n}\mathbf{J} + \mathbf{X}$ are $\frac{1}{n} + x_{ij} \geq \frac{1}{n} - \frac{1}{n} = 0$, and its rows and columns add up to $1 + 0 = 1$, since $\mathbf{X} \in W_n$. So it lies in Ω_n. The proof of Corollary 19.7.3 showed $\Omega_n \subseteq \frac{1}{n}\mathbf{J} + W_n$, a coset, which is an affine set containing Ω_n, so $\operatorname{aff} \Omega_n \subseteq \frac{1}{n}\mathbf{J} + W_n$. Conversely, let $\mathbf{X} \in W_n$, $\mathbf{X} \neq \mathbf{O}$, and put $c = n \max_{i,j} |x_{ij}| > 0$. Then $\frac{1}{n}\mathbf{J}$ and $\frac{1}{n}\mathbf{J} + \frac{1}{c}\mathbf{X}$ lie in Ω_n, and the affine combination

$$\tfrac{1}{n}\mathbf{J} + \mathbf{X} = (1-c)\,\tfrac{1}{n}\mathbf{J} + c\,(\tfrac{1}{n}\mathbf{J} + \tfrac{1}{c}\mathbf{X})$$

lies in $\operatorname{aff} \Omega_n$. So $\operatorname{aff} \Omega_n = \frac{1}{n}\mathbf{J} + W_n$, and $\dim \Omega_n = \dim W_n = (n-1)^2$ by Lemma 19.7.1 (a).

(b) The entries are nonnegative and every row and column adds up to 1. Suppose $\mathbf{M} = (|u_{ij}|^2)$ with $\mathbf{U}$ unitary. The zero entries $m_{13} = m_{21} = m_{32} = 0$ force $u_{13} = u_{21} = u_{32} = 0$. Columns 1 and 2 of $\mathbf{U}$ are $(u_{11}, 0, u_{31})$ and $(u_{12}, u_{22}, 0)$, and they are orthogonal because $\mathbf{U}$ is unitary:

$$0 = u_{11}\overline{u_{12}} + 0 \cdot \overline{u_{22}} + u_{31} \cdot 0 = u_{11}\overline{u_{12}}.$$

But $|u_{11}|^2 = |u_{12}|^2 = \frac{1}{2}$, so $u_{11}\overline{u_{12}} \neq 0$, a contradiction. Hence $\mathbf{M}$ is not unistochastic.

19.8 PageRank

A search engine that finds thousands of pages containing a word must decide which to list first. One answer, which made a search engine famous, is to let the pages vote. Each page spreads its vote over the pages it links to, a vote counts for more when it comes from a page that is itself important, and "important" is whatever this circular description settles on. Written out, the circle is an eigenvector equation for a stochastic matrix. This section sets up that matrix, shows why the raw web does not give a well-defined ranking, repairs it with a single parameter, and uses Perron's theorem (Theorem 19.1.5) to prove that the repaired matrix has exactly one ranking, with every entry positive. We then prove how fast the ranking can be computed, and work a small web exactly.

Throughout, the field is $\mathbb{R}$, inequalities between vectors and matrices are entrywise as in §01, and **stochastic** means column-stochastic (Definition 9.11.1): nonnegative entries, every column adding up to 1. Many accounts of PageRank use rows instead; their matrices are the transposes of ours. We write $\mathbf{1}$ for the all-ones vector and $\mathbf{J} = \mathbf{1}\mathbf{1}^{\top}$ for the all-ones matrix.

19.8.1 The web as a Markov chain

Number the pages $1, \dots, n$. Page j **links to** page i if it contains a link to it; we ignore links from a page to itself, and several links from j to i count as one. Let d_j be the number of pages that page j links to, its **out-degree**. A page with $d_j = 0$ is called **dangling**: a document with no links out, such as an image or a file.

Imagine a surfer who, on page j, clicks one of its d_j links, each with probability $1/d_j$. If $\mathbf{x}_k$ is the probability vector of the surfer's position after k clicks, then, exactly as in Chapter 9 §11, $\mathbf{x}_{k+1} = \mathbf{H}\mathbf{x}_k$, where $h_{ij} = 1/d_j$ if j links to i and $h_{ij} = 0$ otherwise. This hyperlink matrix $\mathbf{H}$ is local to this section; elsewhere in the book $\mathbf{H}$ is a Hessian. The long-run share of time the surfer spends on page i is the candidate for page i's importance. It captures the circular description: page i receives, from each page j linking to it, the share $1/d_j$ of j's own weight.

At a dangling page the surfer has nowhere to go, and column j of $\mathbf{H}$ is zero. The standard patch sends the surfer to a page chosen uniformly at random.

Definition 19.8.1. Link Matrix

Let a web have pages $1, \dots, n$, with $n \geq 1$, and out-degrees $d_1, \dots, d_n$. Its **link matrix** is the $n \times n$ matrix $\mathbf{P}$ whose column j is:

(a) if $d_j \geq 1$: the vector with entry $1/d_j$ in each position i such that page j links to page i, and 0 elsewhere;

(b) if $d_j = 0$ (page j is **dangling**): the vector $\frac{1}{n}\mathbf{1}$.

Each column is a probability vector: in case (a) it has d_j entries equal to $1/d_j$, and in case (b) n entries equal to $1/n$. So $\mathbf{P}$ is stochastic. For instance, the web on three pages in which page 1 links to pages 2 and 3, page 2 links to page 3, and page 3 links nowhere, has

$$\mathbf{P} = \begin{pmatrix} 0 & 0 & \frac{1}{3} \\ \frac{1}{2} & 0 & \frac{1}{3} \\ \frac{1}{2} & 1 & \frac{1}{3} \end{pmatrix}.$$

A ranking should then be a steady state of $\mathbf{P}$ (Definition 9.11.2): a probability vector $\mathbf{x}$ with

$\mathbf{Px} = \mathbf{x}$. When $\mathbf{P}$ is irreducible (Definition 19.2.3), there is exactly one, with every entry positive, by Proposition 19.6.1 (c); part (a) of the same proposition gives $\rho(\mathbf{P}) = 1$ for every stochastic $\mathbf{P}$. The real web is very far from irreducible, and the three failures below are the typical ones.

Example. Three ways a web can fail to rank

For each web, describe what goes wrong with the surfer's chain.

(a) Two pages; page 1 links to page 2, and page 2 is dangling, but we use $\mathbf{H}$, without the patch.

(b) Four pages; pages 1 and 2 link to each other, and so do pages 3 and 4.

(c) Two pages that link to each other.

Solution.

(a) $\mathbf{H} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}$ has a zero column, so it is not stochastic, and $\mathbf{H}^2 = \mathbf{O}$. From any start the surfer is gone after two clicks: $\mathbf{H}^2\mathbf{x}_0 = \mathbf{0}$ is not a probability vector, and $\mathbf{H}$ has no steady state at all, since $\mathbf{Hx} = \mathbf{x}$ forces $x_1 = 0$ and then $x_2 = x_1 = 0$. The patched $\mathbf{P}$ has second column $(\frac{1}{2}, \frac{1}{2})$ and the steady state $(\frac{1}{3}, \frac{2}{3})$, since $\mathbf{P}(1, 2) = (0 \cdot 1 + \frac{1}{2} \cdot 2,\ 1 \cdot 1 + \frac{1}{2} \cdot 2) = (1, 2)$.

(b) $\mathbf{P}$ is block diagonal with two copies of the swap $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$. Both $(\frac{1}{2}, \frac{1}{2}, 0, 0)$ and $(0, 0, \frac{1}{2}, \frac{1}{2})$ are steady states, and so is every average of the two. The chain cannot say whether page 1 matters more than page 3: the ranking is not unique.

(c) $\mathbf{P}$ is the swap. Its steady state $(\frac{1}{2}, \frac{1}{2})$ is unique, but from $\mathbf{x}_0 = \mathbf{e}_1$ the chain alternates $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_1, \ldots$ and never settles. The eigenvalue -1 of the swap has the same modulus as 1.

In (b) the matrix is reducible, and in (c) it is irreducible but not primitive (Definition 19.5.1): the powers of the swap are $\mathbf{I}$ and the swap, and neither is positive. Real webs have all three features, on a vast scale.

19.8.2 The Google matrix

The dangling patch of Definition 19.8.1 already repairs failure (a). One modification repairs the remaining two at once. Let the surfer, at every step, toss a biased coin: with probability α follow a link as before, and with probability $1 - \alpha$ get bored and jump to a page chosen uniformly at random from all n. The jump is called **teleportation**, and α the **damping factor**; a value around 0.85 is customary.

The Google matrix mixes following links with jumping to a random page; its fixed probability vector is the ranking.

Definition 19.8.2. Google Matrix, PageRank Vector

Let $\mathbf{P} \in M_n(\mathbb{R})$ be the link matrix of a web with $n \geq 1$ pages, and let $0 < \alpha < 1$. The matrix

$$\mathbf{G} = \alpha \mathbf{P} + \frac{1-\alpha}{n}\mathbf{J}$$

is called the **Google matrix** in the literature, with damping factor α. A **PageRank vector** of the web is a probability vector $\boldsymbol{\pi} \in \mathbb{R}^n$ with $\mathbf{G}\boldsymbol{\pi} = \boldsymbol{\pi}$. Page i is ranked above page k when $\pi_i > \pi_k$.

In words: column j of $\mathbf{G}$ is α times column j of $\mathbf{P}$, the link step, plus $1-\alpha$ times the uniform vector $\frac{1}{n}\mathbf{1}$, the jump. A PageRank vector is a steady state of the surfer's new chain. The letter $\mathbf{G}$ is local to this section; elsewhere in the book $\mathbf{G}$ is a Gram matrix. The theorem below shows that there is exactly one, which justifies calling it **the** PageRank vector.

Two properties of $\mathbf{G}$ do all the work.

- **$\mathbf{G}$ is stochastic.** Its entries are nonnegative, and column j adds up to $\alpha \cdot 1 + (1-\alpha)\cdot\frac{1}{n}\cdot n = 1$.
- **$\mathbf{G}$ is positive.** Every entry is at least $\frac{1-\alpha}{n} > 0$, because $\alpha\mathbf{P} \geq \mathbf{O}$ and $\alpha < 1$ **strictly**.

Two degenerate values of α are excluded, and each shows why. With $\alpha = 1$, $\mathbf{G} = \mathbf{P}$, and the failures (b) and (c) of Example (Three ways a web can fail to rank) come back. With $\alpha = 0$, $\mathbf{G} = \frac{1}{n}\mathbf{J}$ for every web, the PageRank vector is $\frac{1}{n}\mathbf{1}$, and the links are ignored: every page ties. The damping factor trades faithfulness to the links against the good behavior of the chain, and we will see that the speed of the computation is governed by α alone.

For the three-page web above, with $\alpha = \frac{1}{2}$,

$$\mathbf{G} = \frac{1}{2}\mathbf{P} + \frac{1}{6}\mathbf{J} = \begin{pmatrix} \frac{1}{6} & \frac{1}{6} & \frac{1}{3} \\ \frac{5}{12} & \frac{1}{6} & \frac{1}{3} \\ \frac{5}{12} & \frac{2}{3} & \frac{1}{3} \end{pmatrix},$$

with columns adding up to 1 and every entry positive.

A nonexample by minimal change. Replace the uniform jump by a jump to page 1 only: $\mathbf{G}' = \alpha\mathbf{P} + (1-\alpha)\mathbf{e}_1\mathbf{1}^\top$. It is still stochastic, since each column of $\mathbf{e}_1\mathbf{1}^\top$ is $\mathbf{e}_1$. What fails is positivity: in the web of Example (Three ways a web can fail to rank) (b), $\mathbf{G}'$ has zeros in rows 3 and 4 of columns 1 and 2, so Perron's theorem no longer applies. (Here the ranking is still unique, but it gives pages 3 and 4 weight zero, since no surfer ever returns to them from pages 1 and 2.) Exercise B3 shows that any jump distribution with all entries positive works as well as the uniform one.

Warning. **PageRank is not a count of incoming links.** A link is worth the rank of the page it comes from, divided by that page's out-degree. In the four-page example at the end of this section, page 3 has more incoming links than any other page, yet page 1, with a single incoming link, ranks above it, because that link comes from page 3 and is page 3's only link.

19.8.3 A unique positive ranking

Perron's theorem, applied to the positive matrix $\mathbf{G}$, settles existence, uniqueness, and positivity at once. It also turns the eigenvector problem into a linear system.

Theorem 19.8.1. The PageRank Vector

Let $\mathbf{P} \in M_n(\mathbb{R})$ be stochastic, $0 < \alpha < 1$, and $\mathbf{G} = \alpha\mathbf{P} + \frac{1-\alpha}{n}\mathbf{J}$. Then:

(a) $\rho(\mathbf{G}) = 1$, and 1 is an eigenvalue of $\mathbf{G}$ of algebraic multiplicity one;

(b) there is **exactly one** probability vector $\boldsymbol{\pi}$ with $\mathbf{G}\boldsymbol{\pi} = \boldsymbol{\pi}$, and every entry of $\boldsymbol{\pi}$ is **positive**;

(c) $\mathbf{I} - \alpha\mathbf{P}$ is invertible, and $\boldsymbol{\pi}$ is the unique solution of

$$(\mathbf{I} - \alpha\mathbf{P})\,\boldsymbol{\pi} = \frac{1-\alpha}{n}\,\mathbf{1}.$$

Idea. Perron's theorem hands us a simple eigenvalue $\rho(\mathbf{G})$ with a positive eigenvector, so the only question is the value of $\rho(\mathbf{G})$; the column sums pin it at 1. For (c), on probability vectors the rank-one part $\frac{1-\alpha}{n}\mathbf{J}\mathbf{x}$ is the constant vector $\frac{1-\alpha}{n}\mathbf{1}$, so the eigenvector equation becomes linear with a right-hand side.

Proof.

(a) By Theorem 16.3.3 (a), $\|\mathbf{G}\|_1$ is the largest column sum of $|\mathbf{G}|$, which is 1 since $\mathbf{G}$ is stochastic; so $\rho(\mathbf{G}) \le 1$ by Theorem 16.4.1 (a). By Proposition 9.11.3 (c), 1 is an eigenvalue of $\mathbf{G}$. Hence $\rho(\mathbf{G}) = 1$. Since every entry of $\mathbf{G}$ is at least $\frac{1-\alpha}{n} > 0$, Perron's theorem (Theorem 19.1.5) applies: by its parts (a) and (b), $\rho(\mathbf{G}) = 1$ is an eigenvalue of algebraic multiplicity one, with an eigenvector $\mathbf{v} > \mathbf{0}$ spanning $E_1(\mathbf{G})$.

(b) Put $\boldsymbol{\pi} = \mathbf{v}/(\mathbf{1}^\top\mathbf{v})$, which is defined because $\mathbf{1}^\top\mathbf{v} > 0$. It is a probability vector with positive entries and $\mathbf{G}\boldsymbol{\pi} = \boldsymbol{\pi}$. If $\boldsymbol{\pi}'$ is another probability vector with $\mathbf{G}\boldsymbol{\pi}' = \boldsymbol{\pi}'$, then $\boldsymbol{\pi}'$ lies in $E_1(\mathbf{G}) = \operatorname{span}(\mathbf{v})$, so $\boldsymbol{\pi}' = c\boldsymbol{\pi}$ for some c, and comparing entry sums gives $1 = c$. Hence $\boldsymbol{\pi}' = \boldsymbol{\pi}$. In the language of §01, $\boldsymbol{\pi}$ is the Perron vector of $\mathbf{G}$ (Definition 19.1.2).

(c) If $(\mathbf{I} - \alpha\mathbf{P})\mathbf{z} = \mathbf{0}$ with $\mathbf{z} \neq \mathbf{0}$, then $\mathbf{P}\mathbf{z} = \alpha^{-1}\mathbf{z}$, so $\alpha^{-1} > 1$ is an eigenvalue of $\mathbf{P}$; but $\rho(\mathbf{P}) \le \|\mathbf{P}\|_1 = 1$ as in (a). So 0 is not an eigenvalue of $\mathbf{I} - \alpha\mathbf{P}$, and $\mathbf{I} - \alpha\mathbf{P}$ is invertible by Theorem 9.2.6. For a probability vector $\mathbf{x}$, $\mathbf{J}\mathbf{x} = \mathbf{1}(\mathbf{1}^\top\mathbf{x}) = \mathbf{1}$, so

$$\mathbf{G}\mathbf{x} = \alpha\mathbf{P}\mathbf{x} + \frac{1-\alpha}{n}\,\mathbf{1},$$

and $\mathbf{G}\boldsymbol{\pi} = \boldsymbol{\pi}$ is the equation $(\mathbf{I} - \alpha\mathbf{P})\boldsymbol{\pi} = \frac{1-\alpha}{n}\mathbf{1}$. So $\boldsymbol{\pi}$ solves the system, and by invertibility it is the only solution.

■

So the ranking exists, is unique, and gives every page a positive score: no page is ignored, however obscure. Part (c) is how one computes it for a small web: solve an $n \times n$ linear system, with no eigenvalue in sight. A useful bonus is that the solution of (c) is automatically a probability vector. Multiplying the system on the left by $\mathbf{1}^\top$, and using $\mathbf{1}^\top\mathbf{P} = \mathbf{1}^\top$,

$$(1-\alpha)\,\mathbf{1}^\top\boldsymbol{\pi} = \frac{1-\alpha}{n}\,\mathbf{1}^\top\mathbf{1} = 1 - \alpha,$$

so $\mathbf{1}^\top\boldsymbol{\pi} = 1$ without being imposed.

Quick check. Two pages link to each other, and $\alpha = \frac{1}{2}$. Write down $\mathbf{G}$, find $\boldsymbol{\pi}$, and find the eigenvalues of $\mathbf{G}$.

Solution. $\mathbf{P}$ is the swap, so

$$\mathbf{G} = \frac{1}{2}\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} + \frac{1}{4}\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} \frac{1}{4} & \frac{3}{4} \\ \frac{3}{4} & \frac{1}{4} \end{pmatrix}.$$

Then $\mathbf{G}(1,1) = (1,1)$ and $\mathbf{G}(1,-1) = (-\frac{1}{2}, \frac{1}{2}) = -\frac{1}{2}(1,-1)$. So $\boldsymbol{\pi} = (\frac{1}{2}, \frac{1}{2})$, and the eigenvalues are 1 and $-\frac{1}{2}$. The swap has eigenvalues 1 and -1: the second one has been multiplied by α. The next theorem says that this always happens.

19.8.4 The second eigenvalue

The PageRank vector of a real web is computed by iterating $\mathbf{x}_{k+1} = \mathbf{G}\mathbf{x}_k$, since the web has billions of pages and no one solves a system of that size exactly. How fast the iterates settle depends on the eigenvalues of $\mathbf{G}$ other than 1. For a general positive matrix, Perron's theorem says only that they are strictly smaller than ρ in modulus, with no bound on how much smaller. For the Google matrix there is an exact description.

Theorem 19.8.2. The Eigenvalues of the Google Matrix

Let $n \geq 2$, let $\mathbf{P} \in M_n(\mathbb{R})$ be stochastic, let $0 \leq \alpha \leq 1$, and let $\mathbf{G} = \alpha\mathbf{P} + \frac{1-\alpha}{n}\mathbf{J}$. Since 1 is an eigenvalue of $\mathbf{P}$ (Proposition 9.11.3) and $p_{\mathbf{P}}$ splits over $\mathbb{C}$ (Theorem 6.5.1), it factors as $p_{\mathbf{P}}(x) = (x-1)(x-\mu_2)\cdots(x-\mu_n)$ with $\mu_2, \ldots, \mu_n \in \mathbb{C}$. Then

$$p_{\mathbf{G}}(x) = (x-1)(x-\alpha\mu_2)\cdots(x-\alpha\mu_n).$$

In particular, if $\alpha < 1$, every eigenvalue $\lambda \neq 1$ of $\mathbf{G}$ satisfies $|\lambda| \leq \alpha$.

Idea. Both $\mathbf{P}$ and $\mathbf{G}$ have $\mathbf{1}^\top$ as a left eigenvector for 1. Change basis so that $\mathbf{1}^\top$ becomes the first coordinate functional $\mathbf{e}_1^\top$; then both matrices become block lower triangular with a 1 in the corner. The teleportation term $\frac{1}{n}\mathbf{J} = \frac{1}{n}\mathbf{1}\mathbf{1}^\top$ only touches the first column in the new basis, so the lower-right blocks are $\mathbf{P}_1$ and $\alpha\mathbf{P}_1$ for the **same** $\mathbf{P}_1$. The eigenvalues of $\mathbf{P}$ other than the 1 in the corner are those of $\mathbf{P}_1$; those of $\mathbf{G}$ are α times them.

Proof. Let $\mathbf{N} = \mathbf{e}_1(\mathbf{1} - \mathbf{e}_1)^\top$, the matrix whose first row is $(0, 1, \ldots, 1)$ and whose other rows are zero, and let $\mathbf{S} = \mathbf{I} + \mathbf{N}$. Its first row is $\mathbf{1}^\top$ and its other rows are those of $\mathbf{I}$, so $\mathbf{e}_1^\top\mathbf{S} = \mathbf{1}^\top$. Since $(\mathbf{1} - \mathbf{e}_1)^\top\mathbf{e}_1 = 0$, $\mathbf{N}^2 = \mathbf{e}_1((\mathbf{1} - \mathbf{e}_1)^\top\mathbf{e}_1)(\mathbf{1} - \mathbf{e}_1)^\top = \mathbf{O}$, and therefore $(\mathbf{I} + \mathbf{N})(\mathbf{I} - \mathbf{N}) = \mathbf{I} - \mathbf{N}^2 = \mathbf{I}$. So $\mathbf{S}$ is invertible with $\mathbf{S}^{-1} = \mathbf{I} - \mathbf{N}$ (Theorem 3.4.5).

Claim. **Claim.** If $\mathbf{M} \in M_n(\mathbb{R})$ satisfies $\mathbf{1}^\top\mathbf{M} = \mathbf{1}^\top$, then the first row of $\mathbf{S}\mathbf{M}\mathbf{S}^{-1}$ is $\mathbf{e}_1^\top$.

Proof. The first row is $\mathbf{e}_1^\top\mathbf{S}\mathbf{M}\mathbf{S}^{-1} = \mathbf{1}^\top\mathbf{M}\mathbf{S}^{-1} = \mathbf{1}^\top\mathbf{S}^{-1} = \mathbf{e}_1^\top\mathbf{S}\mathbf{S}^{-1} = \mathbf{e}_1^\top$.

■

Since $\mathbf{P}$ is stochastic, $\mathbf{1}^{\top}\mathbf{P} = \mathbf{1}^{\top}$, and the claim gives

$$\mathbf{S}\mathbf{P}\mathbf{S}^{-1} = \begin{pmatrix} 1 & \mathbf{0}^{\top} \\ \mathbf{b} & \mathbf{P}_1 \end{pmatrix}$$

for some $\mathbf{b} \in \mathbb{R}^{n-1}$ and $\mathbf{P}_1 \in M_{n-1}(\mathbb{R})$. For the teleportation term,

$$\mathbf{S}\left(\tfrac{1}{n}\mathbf{1}\mathbf{1}^{\top}\right)\mathbf{S}^{-1} = \tfrac{1}{n}(\mathbf{S}\mathbf{1})(\mathbf{1}^{\top}\mathbf{S}^{-1}) = \tfrac{1}{n}(\mathbf{S}\mathbf{1})\,\mathbf{e}_1^{\top},$$

using $\mathbf{1}^{\top}\mathbf{S}^{-1} = \mathbf{e}_1^{\top}$ from the proof of the claim. Its only nonzero column is the first, $\tfrac{1}{n}\mathbf{S}\mathbf{1}$, whose first entry is $\tfrac{1}{n}\mathbf{1}^{\top}\mathbf{1} = 1$. So it equals $\begin{pmatrix} 1 & \mathbf{0}^{\top} \\ \mathbf{c} & \mathbf{O} \end{pmatrix}$ for some $\mathbf{c} \in \mathbb{R}^{n-1}$, and

$$\begin{aligned} \mathbf{S}\mathbf{G}\mathbf{S}^{-1} &= \alpha\,\mathbf{S}\mathbf{P}\mathbf{S}^{-1} + (1-\alpha)\,\mathbf{S}\left(\tfrac{1}{n}\mathbf{J}\right)\mathbf{S}^{-1} \\ &= \begin{pmatrix} 1 & \mathbf{0}^{\top} \\ \alpha\mathbf{b} + (1-\alpha)\mathbf{c} & \alpha\mathbf{P}_1 \end{pmatrix}. \end{aligned}$$

Similar matrices have the same characteristic polynomial (Theorem 7.10.3), and $x\mathbf{I}$ minus a block lower triangular matrix is block lower triangular, so by Theorem 7.8.4

$$p_{\mathbf{P}}(x) = (x-1)\,p_{\mathbf{P}_1}(x), \qquad p_{\mathbf{G}}(x) = (x-1)\,p_{\alpha\mathbf{P}_1}(x).$$

Comparing the first with $p_{\mathbf{P}}(x) = (x-1)(x-\mu_2)\cdots(x-\mu_n)$ and canceling the nonzero factor $x-1$ (in $\mathbb{C}[x]$, a product of nonzero polynomials is nonzero) gives $p_{\mathbf{P}_1}(x) = (x-\mu_2)\cdots(x-\mu_n)$. If $\alpha > 0$, then

$$\begin{aligned} p_{\alpha\mathbf{P}_1}(x) &= \det(x\mathbf{I} - \alpha\mathbf{P}_1) = \alpha^{n-1}\det\left(\tfrac{x}{\alpha}\mathbf{I} - \mathbf{P}_1\right) \\ &= \alpha^{n-1}\prod_{j=2}^{n}\left(\frac{x}{\alpha} - \mu_j\right) = \prod_{j=2}^{n}(x - \alpha\mu_j), \end{aligned}$$

where the second step takes the factor α out of each of the $n-1$ rows. If $\alpha = 0$, then $p_{\alpha\mathbf{P}_1}(x) = \det(x\mathbf{I}) = x^{n-1} = \prod_j (x - 0\cdot\mu_j)$. Either way the formula for $p_{\mathbf{G}}$ follows.

For the last sentence, each μ_j is an eigenvalue of $\mathbf{P}$, so $|\mu_j| \le \rho(\mathbf{P}) \le \|\mathbf{P}\|_1 = 1$ by Theorem 16.4.1 (a) and Theorem 16.3.3 (a). The eigenvalues of $\mathbf{G}$ are the roots of $p_{\mathbf{G}}$: the number 1 and the $\alpha\mu_j$, each of modulus at most α. If $\alpha < 1$, none of the $\alpha\mu_j$ equals 1, so every eigenvalue $\lambda \ne 1$ is some $\alpha\mu_j$, and $|\lambda| \le \alpha$. This proves the theorem.

■

The theorem gives more than a bound. It shows again that 1 is a simple eigenvalue of $\mathbf{G}$ when $\alpha < 1$, without Perron's theorem, and it says that the rest of the spectrum of $\mathbf{G}$ is the rest of the spectrum of $\mathbf{P}$, shrunk by the factor α. The bound is attained whenever $\mathbf{P}$ itself has 1 as a repeated eigenvalue, as for the two separate webs of Example (Three ways a web can fail to rank) (b) (Exercise C2), and whenever $\mathbf{P}$ has -1 as an eigenvalue, as in the Quick check. So for a web with two or more pieces that no link leaves, as large webs typically have, the second eigenvalue of $\mathbf{G}$ has modulus exactly α, and no better bound is possible.

19.8.5 How fast the iteration converges

The eigenvalue bound predicts that iterates approach $\boldsymbol{\pi}$ at a rate like α^k. For this particular matrix the prediction can be proved directly, with an explicit constant, by measuring errors in the 1-norm of Chapter 16.

Proposition 19.8.3. Convergence at the Damping Rate

Let $\mathbf{P}$, α, $\mathbf{G}$ and $\boldsymbol{\pi}$ be as in Theorem 19.8.1, let $\mathbf{x}_0$ be any probability vector, and let $\mathbf{x}_k = \mathbf{G}^k\mathbf{x}_0$. Then for every $k \geq 0$,

$$\|\mathbf{x}_k - \boldsymbol{\pi}\|_1 \leq \alpha^k \|\mathbf{x}_0 - \boldsymbol{\pi}\|_1 \leq 2\alpha^k.$$

In particular, $\mathbf{x}_k \to \boldsymbol{\pi}$ entrywise.

Idea. The error $\boldsymbol{\epsilon}_k = \mathbf{x}_k - \boldsymbol{\pi}$ is a difference of probability vectors, so its entries add up to 0, and the teleportation term kills it. What remains is $\alpha\mathbf{P}$, and a stochastic matrix never increases the 1-norm.

Proof. Each $\mathbf{x}_k$ is a probability vector by Proposition 9.11.3 (a), since $\mathbf{G}$ is stochastic. Put $\boldsymbol{\epsilon}_k = \mathbf{x}_k - \boldsymbol{\pi}$, so $\mathbf{1}^\top\boldsymbol{\epsilon}_k = 1 - 1 = 0$. Since $\mathbf{G}\boldsymbol{\pi} = \boldsymbol{\pi}$,

$$\boldsymbol{\epsilon}_{k+1} = \mathbf{G}\mathbf{x}_k - \mathbf{G}\boldsymbol{\pi} = \mathbf{G}\boldsymbol{\epsilon}_k = \alpha\mathbf{P}\boldsymbol{\epsilon}_k + \frac{1-\alpha}{n}\mathbf{1}(\mathbf{1}^\top\boldsymbol{\epsilon}_k) = \alpha\mathbf{P}\boldsymbol{\epsilon}_k.$$

By Theorem 16.3.3 (a), $\|\mathbf{P}\|_1 = 1$, the largest column sum of the nonnegative matrix $\mathbf{P}$, so $\|\mathbf{P}\boldsymbol{\epsilon}_k\|_1 \leq \|\boldsymbol{\epsilon}_k\|_1$ (Definition 16.3.2). Hence $\|\boldsymbol{\epsilon}_{k+1}\|_1 \leq \alpha\|\boldsymbol{\epsilon}_k\|_1$, and by induction on k, $\|\boldsymbol{\epsilon}_k\|_1 \leq \alpha^k\|\boldsymbol{\epsilon}_0\|_1$. Finally $\|\boldsymbol{\epsilon}_0\|_1 \leq \|\mathbf{x}_0\|_1 + \|\boldsymbol{\pi}\|_1 = 2$, as both are probability vectors. Each entry of $\boldsymbol{\epsilon}_k$ has modulus at most $\|\boldsymbol{\epsilon}_k\|_1 \leq 2\alpha^k$, and $\alpha^k \to 0$ because $0 < \alpha < 1$, one of the calculus facts recorded in Chapter 9 §11. A sequence whose terms are at most $2\alpha^k$ in absolute value, with $2\alpha^k \to 0$, tends to 0 straight from the definition of a limit: once $2\alpha^k < \varepsilon$, so is every such term. So $\mathbf{x}_k \to \boldsymbol{\pi}$ entrywise.

■

The bound does not depend on the web at all: not on the number of pages, not on how they link. With $\alpha = 0.85$, fifty steps guarantee $\|\mathbf{x}_{50} - \boldsymbol{\pi}\|_1 \leq 2(0.85)^{50} < 6 \cdot 10^{-4}$, and ninety guarantee an error below 10^{-6}, for a web of any size. Each step is cheap, because $\mathbf{G}$ is never formed: by the proof of Theorem 19.8.1 (c), $\mathbf{G}\mathbf{x} = \alpha\mathbf{P}\mathbf{x} + \frac{1-\alpha}{n}\mathbf{1}$ for a probability vector $\mathbf{x}$, and $\mathbf{P}\mathbf{x}$ needs one operation per link, plus one shared correction for the dangling pages.

In the language of §05, $\mathbf{G}$ is primitive (Definition 19.5.1), since $\mathbf{G}^1$ is already positive, and Theorem 19.5.2 gives $\mathbf{G}^k \to \boldsymbol{\pi}\mathbf{1}^\top$, with right Perron vector $\boldsymbol{\pi}$ and left Perron vector $\mathbf{1}$, normalized by $\mathbf{1}^\top\boldsymbol{\pi} = 1$. The proposition adds the rate, and the rate is what makes the method practical. The iteration $\mathbf{x}_{k+1} = \mathbf{G}\mathbf{x}_k$ is the **power method**; Chapter 24 studies it for general matrices, where the rate is governed by the ratio of the two largest eigenvalue moduli, here at most α.

19.8.6 A small web, exactly

Example. Four pages, one dangling

A web has four pages. Page 1 links to pages 2 and 3; page 2 links to pages 3 and 4; page 3 links to page 1; page 4 is dangling. With $\alpha = \frac{4}{5}$, find the Google matrix, the PageRank vector, and the ranking, and compare $|\lambda_2|$ with α.

Solution. The link matrix. Pages 1 and 2 have out-degree 2, page 3 has out-degree 1, and page 4 is dangling, so by Definition 19.8.1

$$\mathbf{P} = \begin{pmatrix} 0 & 0 & 1 & \frac{1}{4} \\ \frac{1}{2} & 0 & 0 & \frac{1}{4} \\ \frac{1}{2} & \frac{1}{2} & 0 & \frac{1}{4} \\ 0 & \frac{1}{2} & 0 & \frac{1}{4} \end{pmatrix}, \qquad 20\,\mathbf{G} = 16\,\mathbf{P} + \mathbf{J} = \begin{pmatrix} 1 & 1 & 17 & 5 \\ 9 & 1 & 1 & 5 \\ 9 & 9 & 1 & 5 \\ 1 & 9 & 1 & 5 \end{pmatrix},$$

since $20\alpha = 16$ and $20 \cdot \frac{1-\alpha}{4} = 1$. Every column of $20\mathbf{G}$ adds up to 20, and every entry is positive.

The linear system. By Theorem 19.8.1 (c), $(\mathbf{I} - \frac{4}{5}\mathbf{P})\boldsymbol{\pi} = \frac{1}{20}\mathbf{1}$. Multiplying by 5 gives $(5\mathbf{I} - 4\mathbf{P})\boldsymbol{\pi} = \frac{1}{4}\mathbf{1}$, that is,

$$\begin{aligned} 5\pi_1 - 4\pi_3 - \pi_4 &= \tfrac{1}{4}, & -2\pi_1 + 5\pi_2 - \pi_4 &= \tfrac{1}{4}, \\ -2\pi_1 - 2\pi_2 + 5\pi_3 - \pi_4 &= \tfrac{1}{4}, & -2\pi_2 + 4\pi_4 &= \tfrac{1}{4}. \end{aligned}$$

Express everything through π_4. The fourth equation gives $\pi_2 = 2\pi_4 - \frac{1}{8}$. The second then gives $2\pi_1 = 5\pi_2 - \pi_4 - \frac{1}{4} = 9\pi_4 - \frac{7}{8}$, so $\pi_1 = \frac{9}{2}\pi_4 - \frac{7}{16}$. The first gives $4\pi_3 = 5\pi_1 - \pi_4 - \frac{1}{4} = \frac{43}{2}\pi_4 - \frac{39}{16}$, so $\pi_3 = \frac{43}{8}\pi_4 - \frac{39}{64}$. Substituting into the third,

$$\Big(-9 - 4 + \tfrac{215}{8} - 1\Big)\pi_4 + \Big(\tfrac{7}{8} + \tfrac{1}{4} - \tfrac{195}{64}\Big) = \tfrac{103}{8}\pi_4 - \tfrac{123}{64} = \tfrac{1}{4},$$

so $\frac{103}{8}\pi_4 = \frac{139}{64}$ and $\pi_4 = \frac{139}{824}$. Back-substituting,

$$\boldsymbol{\pi} = \frac{1}{824}\,(265,\ 175,\ 245,\ 139).$$

Check. The entries add up to $824/824 = 1$, as they must. With $\mathbf{y} = (265, 175, 245, 139)$, the first row of $20\mathbf{G}$ gives $265 + 175 + 17 \cdot 245 + 5 \cdot 139 = 5300 = 20 \cdot 265$, and the other rows give $3500 = 20 \cdot 175$, $4900 = 20 \cdot 245$ and $2780 = 20 \cdot 139$. So $\mathbf{G}\boldsymbol{\pi} = \boldsymbol{\pi}$.

The ranking. Page 1 first (≈ 0.322), then page 3 (≈ 0.297), page 2 (≈ 0.212) and page 4 (≈ 0.169). Page 3 is linked from pages 1 and 2, while the dangling page 4 feeds every page equally; page 1 is linked only from page 3. Yet page 1 wins, because page 3 passes its whole weight to it, while pages 1, 2 and 4 split theirs.

The second eigenvalue. Expanding the determinant gives

$$\begin{aligned} p_{\mathbf{P}}(x) &= (x-1)\Big(x^3 + \tfrac{3}{4}x^2 + \tfrac{1}{8}x - \tfrac{1}{16}\Big), \\ p_{\mathbf{G}}(x) &= (x-1)\Big(x^3 + \tfrac{3}{5}x^2 + \tfrac{2}{25}x - \tfrac{4}{125}\Big), \end{aligned}$$

and the second cubic is α^3 times the first evaluated at x/α, as Theorem 19.8.2 predicts: $\frac{3}{4} \cdot \frac{4}{5} = \frac{3}{5}$, $\frac{1}{8} \cdot \frac{16}{25} = \frac{2}{25}$, $\frac{1}{16} \cdot \frac{64}{125} = \frac{4}{125}$. The cubic for $\mathbf{P}$ has one real root near 0.199 and two complex ones of modulus about 0.560. So $|\lambda_2(\mathbf{G})| \approx 0.8 \times 0.560 \approx 0.448$, comfortably below $\alpha = 0.8$: for this web the iteration converges faster than Proposition 19.8.3 guarantees.

19.8.7 Exercises

19.8.7.1 A. Check your understanding

Exercise. A1

(a) Define the Google matrix of a web and its PageRank vector.

(b) Why is every entry of $\mathbf{G}$ positive, and which three conclusions does Theorem 19.1.5 then give?

(c) True or false: with $\alpha = 1$, every web has exactly one PageRank vector. Justify your answer.

(d) True or false: the page with the most incoming links always has the highest PageRank. Justify your answer.

(e) For $\alpha = 0.85$, what does Theorem 19.8.2 say about the eigenvalues of $\mathbf{G}$ other than 1? Can the bound be improved for every web?

(f) True or false: after 10 steps of $\mathbf{x}_{k+1} = \mathbf{G}\mathbf{x}_k$ from any probability vector, $\|\mathbf{x}_{10} - \boldsymbol{\pi}\|_1 \leq 2\alpha^{10}$. Justify your answer.

Solution.

(a) For a web with link matrix $\mathbf{P}$ (Definition 19.8.1) and $0 < \alpha < 1$, $\mathbf{G} = \alpha\mathbf{P} + \frac{1-\alpha}{n}\mathbf{J}$. A PageRank vector is a probability vector $\boldsymbol{\pi}$ with $\mathbf{G}\boldsymbol{\pi} = \boldsymbol{\pi}$.

(b) $\alpha\mathbf{P}$ has nonnegative entries and $\frac{1-\alpha}{n}\mathbf{J}$ has every entry equal to $\frac{1-\alpha}{n} > 0$, since $\alpha < 1$. Perron's theorem then gives: $\rho(\mathbf{G})$ is an eigenvalue of algebraic multiplicity one; it has an eigenvector with all entries positive; every other eigenvalue has modulus strictly less than $\rho(\mathbf{G})$. Here $\rho(\mathbf{G}) = 1$.

(c) False. With $\alpha = 1$, $\mathbf{G} = \mathbf{P}$, and for two pairs of pages linking only to each other, Example (Three ways a web can fail to rank) (b) gives two different steady states.

(d) False. In Example (Four pages, one dangling), page 3 has the most incoming links, but page 1 ranks first.

(e) Every eigenvalue $\lambda \neq 1$ has $|\lambda| \leq 0.85$. The bound cannot be improved for every web: when the link matrix has -1 or a repeated 1 among its eigenvalues, $\mathbf{G}$ has an eigenvalue of modulus exactly 0.85 (the Quick check, and Exercise C2).

(f) True, by Proposition 19.8.3 with $k = 10$.

19.8.7.2 B. Practice

Exercise. B1: A three-page web

For the three-page web after Definition 19.8.1 (page 1 links to pages 2 and 3, page 2 to page 3, page 3 is dangling), with $\alpha = \frac{1}{2}$, find the PageRank vector by solving the system of Theorem 19.8.1 (c). Hence rank the pages.

Solution. With $\mathbf{P}$ as in the text, $(\mathbf{I}-\frac{1}{2}\mathbf{P})\boldsymbol{\pi}=\frac{1}{6}\mathbf{1}$. Multiplying by 12 gives $(12\mathbf{I}-6\mathbf{P})\boldsymbol{\pi}=2\,\mathbf{1}$:

$$\begin{aligned} 12\pi_1 - 2\pi_3 &= 2,\\ -3\pi_1 + 12\pi_2 - 2\pi_3 &= 2,\\ -3\pi_1 - 6\pi_2 + 10\pi_3 &= 2. \end{aligned}$$

The first gives $\pi_3 = 6\pi_1 - 1$. The second becomes $-3\pi_1 + 12\pi_2 - 12\pi_1 + 2 = 2$, so $\pi_2 = \frac{5}{4}\pi_1$. The third becomes $-3\pi_1 - \frac{15}{2}\pi_1 + 60\pi_1 - 10 = 2$, so $\frac{99}{2}\pi_1 = 12$ and $\pi_1 = \frac{8}{33}$. Then $\pi_2 = \frac{10}{33}$ and $\pi_3 = \frac{48}{33} - 1 = \frac{15}{33}$. So

$$\boldsymbol{\pi} = \frac{1}{33}(8,\ 10,\ 15),$$

whose entries add up to 1. Check against the original system: $12\cdot 8 - 2\cdot 15 = 66 = 2\cdot 33$, $-24 + 120 - 30 = 66$, $-24 - 60 + 150 = 66$. Hence page 3 ranks first, then page 2, then page 1.

Exercise. B2: A cycle of three pages

Three pages link in a cycle: 1 to 2, 2 to 3, 3 to 1. Let $0 < \alpha < 1$.

(a) Write down $\mathbf{P}$ and show that $p_{\mathbf{P}}(x) = x^3 - 1$.

(b) Use Theorem 19.8.2 to find the eigenvalues of $\mathbf{G}$, and check your answer against $\operatorname{tr}\mathbf{G}$.

(c) Find $\boldsymbol{\pi}$, and explain why the bound of Theorem 19.8.2 is attained.

Solution.

(a) Column j of $\mathbf{P}$ is $\mathbf{e}_{j+1}$, with $\mathbf{e}_4$ read as $\mathbf{e}_1$:

$$\mathbf{P} = \begin{pmatrix} 0 & 0 & 1\\ 1 & 0 & 0\\ 0 & 1 & 0 \end{pmatrix}, \qquad \det(x\mathbf{I}-\mathbf{P}) = \det\begin{pmatrix} x & 0 & -1\\ -1 & x & 0\\ 0 & -1 & x \end{pmatrix}$$
$$= x\cdot x^2 + (-1)\cdot\big((-1)(-1) - x\cdot 0\big) = x^3 - 1,$$

expanding along the first row.

(b) $x^3 - 1 = (x-1)(x-\omega)(x-\overline{\omega})$ with $\omega = -\frac{1}{2} + \frac{\sqrt{3}}{2}i$, so $\mu_2 = \omega$, $\mu_3 = \overline{\omega}$, and the eigenvalues of $\mathbf{G}$ are $1, \alpha\omega, \alpha\overline{\omega}$. Their sum is $1 + \alpha(\omega + \overline{\omega}) = 1 - \alpha$. Directly, $\operatorname{tr}\mathbf{G} = \alpha\operatorname{tr}\mathbf{P} + \frac{1-\alpha}{3}\operatorname{tr}\mathbf{J} = 0 + (1-\alpha)$, which agrees (Theorem 9.3.3).

(c) $\mathbf{P}\mathbf{1} = \mathbf{1}$ and $\mathbf{J}\mathbf{1} = 3\,\mathbf{1}$, so $\mathbf{G}\mathbf{1} = \alpha\mathbf{1} + (1-\alpha)\mathbf{1} = \mathbf{1}$, and $\boldsymbol{\pi} = \frac{1}{3}\mathbf{1}$ by the uniqueness in Theorem 19.8.1. The eigenvalues $\alpha\omega$ and $\alpha\overline{\omega}$ have modulus $\alpha|\omega| = \alpha$, so the bound $|\lambda| \le \alpha$ holds with equality.

Exercise. B3: Personalized jumps

Let $\mathbf{P} \in M_n(\mathbb{R})$ be stochastic, $0 < \alpha < 1$, and let $\mathbf{v}$ be a probability vector with **every** $v_i > 0$. Put $\mathbf{G}_{\mathbf{v}} = \alpha\mathbf{P} + (1-\alpha)\mathbf{v}\mathbf{1}^{\top}$: the surfer now jumps to page i with probability v_i.

(a) Prove that $\mathbf{G}_{\mathbf{v}}$ is stochastic and positive, and deduce that it has exactly one probability vector $\boldsymbol{\pi}$ with $\mathbf{G}_{\mathbf{v}}\boldsymbol{\pi} = \boldsymbol{\pi}$, and that $\boldsymbol{\pi} > \mathbf{0}$.

(b) Prove that $\boldsymbol{\pi}$ is the unique solution of $(\mathbf{I} - \alpha\mathbf{P})\boldsymbol{\pi} = (1-\alpha)\mathbf{v}$.

(c) For two pages linking to each other, $\alpha = \frac{1}{2}$ and $\mathbf{v} = (\frac{3}{4}, \frac{1}{4})$, find $\boldsymbol{\pi}$.

Solution.

(a) The entries of $\mathbf{G}_\mathbf{v}$ are $\alpha p_{ij} + (1-\alpha)v_i \geq (1-\alpha)v_i > 0$. Column j adds up to $\alpha + (1-\alpha)\mathbf{1}^\top\mathbf{v} = \alpha + 1 - \alpha = 1$. So $\mathbf{G}_\mathbf{v}$ is stochastic and positive. The proof of Theorem 19.8.1 (a) and (b) used only these two facts, so it applies word for word: $\rho(\mathbf{G}_\mathbf{v}) = 1$ is an eigenvalue of algebraic multiplicity one with a positive eigenvector, and its normalization $\boldsymbol{\pi}$ is the only probability vector fixed by $\mathbf{G}_\mathbf{v}$.

(b) For a probability vector $\mathbf{x}$, $\mathbf{G}_\mathbf{v}\mathbf{x} = \alpha\mathbf{P}\mathbf{x} + (1-\alpha)\mathbf{v}(\mathbf{1}^\top\mathbf{x}) = \alpha\mathbf{P}\mathbf{x} + (1-\alpha)\mathbf{v}$, so $\mathbf{G}_\mathbf{v}\boldsymbol{\pi} = \boldsymbol{\pi}$ says $(\mathbf{I} - \alpha\mathbf{P})\boldsymbol{\pi} = (1-\alpha)\mathbf{v}$. The matrix $\mathbf{I} - \alpha\mathbf{P}$ is invertible, as in the proof of Theorem 19.8.1 (c), so the solution is unique.

(c) $\mathbf{P}$ is the swap, and the system is $\pi_1 - \frac{1}{2}\pi_2 = \frac{3}{8}$, $-\frac{1}{2}\pi_1 + \pi_2 = \frac{1}{8}$. The first gives $\pi_1 = \frac{3}{8} + \frac{1}{2}\pi_2$, and the second becomes $-\frac{3}{16} + \frac{3}{4}\pi_2 = \frac{1}{8}$, so $\pi_2 = \frac{5}{12}$ and $\pi_1 = \frac{3}{8} + \frac{5}{24} = \frac{7}{12}$. So $\boldsymbol{\pi} = (\frac{7}{12}, \frac{5}{12})$: the preferred page 1 gains, but less than its jump weight $\frac{3}{4}$, because half the time the surfer follows the link to page 2.

19.8.7.3 C. Going deeper

Exercise. C1: PageRank as a sum over paths

Let $\mathbf{P}$, α, $\mathbf{G}$ and $\boldsymbol{\pi}$ be as in Theorem 19.8.1.

(a) Prove that $\boldsymbol{\pi} = \dfrac{1-\alpha}{n}\displaystyle\sum_{k=0}^{\infty} \alpha^k \mathbf{P}^k \mathbf{1}$.

(b) Deduce that $\pi_i \geq \frac{1-\alpha}{n}$ for every i.

(c) Suppose no page is dangling and no page links to page i. Prove that $\pi_i = \frac{1-\alpha}{n}$ exactly.

Hint for (a): Theorem 16.5.1 with the 1-norm.

Solution.

(a) By Theorem 16.3.3 (a), $\|\alpha\mathbf{P}\|_1 = \alpha\|\mathbf{P}\|_1 = \alpha < 1$, the largest column sum of $\alpha\mathbf{P}$. By Theorem 16.5.1 (a), $\mathbf{I} - \alpha\mathbf{P}$ is invertible and $(\mathbf{I} - \alpha\mathbf{P})^{-1} = \sum_{k\geq 0} \alpha^k\mathbf{P}^k$. By Theorem 19.8.1 (c), $\boldsymbol{\pi} = \frac{1-\alpha}{n}(\mathbf{I} - \alpha\mathbf{P})^{-1}\mathbf{1}$, which is the stated series.

(b) Every term $\alpha^k\mathbf{P}^k\mathbf{1}$ has nonnegative entries, and the term $k = 0$ is $\mathbf{1}$. So each entry of the sum is at least 1, since a limit of partial sums that are all at least 1 is at least 1, and $\pi_i \geq \frac{1-\alpha}{n}$.

(c) From $\mathbf{G}\boldsymbol{\pi} = \boldsymbol{\pi}$ and the proof of Theorem 19.8.1 (c), $\boldsymbol{\pi} = \alpha\mathbf{P}\boldsymbol{\pi} + \frac{1-\alpha}{n}\mathbf{1}$. Row i of $\mathbf{P}$ is zero: $p_{ij} \neq 0$ only if page j links to page i or page j is dangling, and neither happens. So $(\mathbf{P}\boldsymbol{\pi})_i = 0$ and $\pi_i = \frac{1-\alpha}{n}$, the smallest possible value by (b).

Exercise. C2: When the bound is attained

Let $\mathbf{P} \in M_n(\mathbb{R})$ be stochastic, $0 < \alpha < 1$, and $\mathbf{G} = \alpha\mathbf{P} + \frac{1-\alpha}{n}\mathbf{J}$.

(a) Suppose $\mathbf{P}$ has two **different** steady states $\mathbf{u} \neq \mathbf{w}$. Prove that $\mathbf{G}(\mathbf{u}-\mathbf{w}) = \alpha(\mathbf{u}-\mathbf{w})$, so that α is an eigenvalue of $\mathbf{G}$.

(b) For the web of Example (Three ways a web can fail to rank) (b), find all eigenvalues of $\mathbf{G}$, with multiplicity, and check your answer against (a).

Solution.

(a) Put $\mathbf{z} = \mathbf{u} - \mathbf{w} \neq \mathbf{0}$. Both are probability vectors, so $\mathbf{1}^\top\mathbf{z} = 0$ and $\mathbf{J}\mathbf{z} = \mathbf{1}(\mathbf{1}^\top\mathbf{z}) = \mathbf{0}$. Hence $\mathbf{G}\mathbf{z} = \alpha\mathbf{P}\mathbf{u} - \alpha\mathbf{P}\mathbf{w} + \mathbf{0} = \alpha(\mathbf{u}-\mathbf{w}) = \alpha\mathbf{z}$, and α is an eigenvalue of $\mathbf{G}$ with eigenvector $\mathbf{z}$.

(b) $\mathbf{P}$ is block diagonal with two blocks $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$, each with characteristic polynomial x^2-1, so $p_{\mathbf{P}}(x) = (x^2-1)^2 = (x-1)(x-1)(x+1)(x+1)$ by Theorem 7.8.4. By Theorem 19.8.2, with $\mu_2 = 1$ and $\mu_3 = \mu_4 = -1$, the eigenvalues of $\mathbf{G}$ are $1, \alpha, -\alpha, -\alpha$. The eigenvalue α is the one found in (a), with $\mathbf{u} = (\frac{1}{2}, \frac{1}{2}, 0, 0)$, $\mathbf{w} = (0, 0, \frac{1}{2}, \frac{1}{2})$ and eigenvector $\mathbf{z} = \frac{1}{2}(1, 1, -1, -1)$; indeed $\mathbf{P}\mathbf{z} = \mathbf{z}$ and $\mathbf{J}\mathbf{z} = \mathbf{0}$, so $\mathbf{G}\mathbf{z} = \alpha\mathbf{z}$.

19.9 M-Matrices

Every section of this chapter so far has studied a nonnegative matrix $\mathbf{B}$ through its spectral radius. This last section looks at the matrices $s\mathbf{I} - \mathbf{B}$ built from one, which turn up wherever a quantity is balanced against what flows into it from its neighbors: a discretized differential equation, an economy that consumes part of its own output, the fundamental matrix of an absorbing Markov chain. The question is always the same. When is $s\mathbf{I} - \mathbf{B}$ invertible with a **nonnegative** inverse, so that a nonnegative right-hand side is guaranteed a nonnegative solution? The answer is a single inequality, $s > \rho(\mathbf{B})$, and the section's main theorem shows that six other conditions, several of which can be checked by hand, say the same thing. The chapter's summary closes the section.

Throughout, matrices have real entries, the spectrum $\operatorname{spec}(\mathbf{A})$ of $\mathbf{A} \in M_n(\mathbb{R})$ is taken in $\mathbb{C}$, as in the definition of the spectral radius (Definition 16.4.1), and $\geq$, $>$ between matrices or vectors are entrywise.

19.9.1 Z-matrices and M-matrices

A matrix of the form $s\mathbf{I} - \mathbf{B}$ with $\mathbf{B} \geq \mathbf{O}$ has an unmistakable sign pattern: the diagonal entries $s - b_{ii}$ can be anything, and every entry off the diagonal is $-b_{ij} \leq 0$. The pattern comes first, because it is what one sees on the page.

Definition 19.9.1. Z-Matrix

A matrix $\mathbf{A} = (a_{ij}) \in M_n(\mathbb{R})$, $n \geq 1$, is a **Z-matrix** if $a_{ij} \leq 0$ for **every** pair $i \neq j$.

The diagonal is unrestricted. The identity is a Z-matrix, and so is $-\mathbf{I}$; every 1×1 matrix is one, since there is no off-diagonal entry to check. The matrix $\left(\begin{smallmatrix} 2 & -1 \\ -1 & 2 \end{smallmatrix}\right)$ is one; $\left(\begin{smallmatrix} 2 & 1 \\ -1 & 2 \end{smallmatrix}\right)$ is not, because of the single entry $a_{12} = 1 > 0$.

Lemma 19.9.1. Z-Matrices Are Shifts of Nonnegative Matrices

A matrix $\mathbf{A} \in M_n(\mathbb{R})$ is a Z-matrix if and only if $\mathbf{A} = s\mathbf{I} - \mathbf{B}$ for some real s and some $\mathbf{B} \in M_n(\mathbb{R})$ with $\mathbf{B} \geq \mathbf{O}$. For a Z-matrix, **every** $s \geq \max_i a_{ii}$ gives such a representation, with $\mathbf{B} = s\mathbf{I} - \mathbf{A}$.

Proof. If $\mathbf{A} = s\mathbf{I} - \mathbf{B}$ with $\mathbf{B} \geq \mathbf{O}$, then $a_{ij} = -b_{ij} \leq 0$ for $i \neq j$. Conversely, let $\mathbf{A}$ be a Z-matrix, let $s \geq \max_i a_{ii}$ and put $\mathbf{B} = s\mathbf{I} - \mathbf{A}$. Its diagonal entries are $s - a_{ii} \geq 0$ by the choice of s, and its off-diagonal entries are $-a_{ij} \geq 0$ because $\mathbf{A}$ is a Z-matrix. So $\mathbf{B} \geq \mathbf{O}$ and $\mathbf{A} = s\mathbf{I} - \mathbf{B}$.

■

The representation is far from unique: raising s by t raises every diagonal entry of $\mathbf{B}$ by t. A definition built on it must therefore not depend on which representation is used, and that is the first thing to check below.

Now the goal. For a Z-matrix $\mathbf{A} = s\mathbf{I} - \mathbf{B}$ with $s > 0$, write $\mathbf{A} = s(\mathbf{I} - s^{-1}\mathbf{B})$. If $\rho(\mathbf{B}) < s$, the Neumann series of Chapter 10 (Theorem 10.10.6) inverts the second factor as $\sum_k (s^{-1}\mathbf{B})^k$, a sum of nonnegative matrices. So a comparison between s and $\rho(\mathbf{B})$ is exactly what decides whether the inverse is nonnegative, and it deserves a name.

An M-matrix is a nonnegative matrix subtracted from a multiple of the identity large enough to dominate its spectral radius.

Definition 19.9.2. M-Matrix

A matrix $\mathbf{A} \in M_n(\mathbb{R})$, $n \geq 1$, is an **M-matrix** if it can be written as

$$\mathbf{A} = s\mathbf{I} - \mathbf{B} \qquad \text{with } s \in \mathbb{R},\ \mathbf{B} \in M_n(\mathbb{R}),\ \mathbf{B} \geq \mathbf{O} \text{ and } s \geq \rho(\mathbf{B}).$$

An M-matrix that is **invertible** is a **nonsingular M-matrix**.

In words: the first clause asks that $\mathbf{A}$ be a Z-matrix (by Lemma 19.9.1, that is what "$s\mathbf{I} - \mathbf{B}$ with $\mathbf{B} \geq \mathbf{O}$" means), and the second asks that the shift s be **at least** the spectral radius of what is subtracted. The definition says "can be written", so a single good representation suffices. The next lemma shows that one representation is as good as any other, and also identifies the singular case.

Lemma 19.9.2. The Smallest Real Eigenvalue of a Z-Matrix

Let $\mathbf{A} \in M_n(\mathbb{R})$ be a Z-matrix, and let $\mathbf{A} = s\mathbf{I} - \mathbf{B}$ with $\mathbf{B} \geq \mathbf{O}$ be **any** representation as in Lemma 19.9.1.

(a) $s - \rho(\mathbf{B})$ is an eigenvalue of $\mathbf{A}$, and every $\lambda \in \operatorname{spec}(\mathbf{A})$ satisfies $\operatorname{Re}\lambda \geq s - \rho(\mathbf{B})$. In particular $s - \rho(\mathbf{B})$ is the **smallest real eigenvalue** of $\mathbf{A}$, and so it is the same number for every representation.

(b) $\mathbf{A}$ is an M-matrix if and only if $s \geq \rho(\mathbf{B})$, and a nonsingular M-matrix if and only if $s > \rho(\mathbf{B})$.

Idea. Shifting by $s\mathbf{I}$ moves the spectrum rigidly: the eigenvalues of $\mathbf{A}$ are $s-\mu$ for the eigenvalues μ of $\mathbf{B}$. All of those μ lie in the disc of radius $\rho(\mathbf{B})$, and the point of that disc farthest to the right, $\rho(\mathbf{B})$ itself, is an eigenvalue because $\mathbf{B} \geq \mathbf{O}$. Reflected through s, that point becomes the leftmost point of the spectrum of $\mathbf{A}$, and it is real.

Proof.

(a) For $\mathbf{v} \in \mathbb{C}^n$ and $\lambda \in \mathbb{C}$, $\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$ if and only if $\mathbf{B}\mathbf{v} = (s-\lambda)\mathbf{v}$. Hence $\operatorname{spec}(\mathbf{A}) = \{s-\mu : \mu \in \operatorname{spec}(\mathbf{B})\}$. By Theorem 19.3.1, $\rho(\mathbf{B}) \in \operatorname{spec}(\mathbf{B})$, so $s-\rho(\mathbf{B}) \in \operatorname{spec}(\mathbf{A})$. Now let $\lambda \in \operatorname{spec}(\mathbf{A})$ and put $\mu = s-\lambda \in \operatorname{spec}(\mathbf{B})$. By Definition 16.4.1, $|\mu| \leq \rho(\mathbf{B})$, so

$$\operatorname{Re}\lambda = s - \operatorname{Re}\mu \geq s - |\mu| \geq s - \rho(\mathbf{B}).$$

For real λ this reads $\lambda \geq s-\rho(\mathbf{B})$. So $s-\rho(\mathbf{B})$ is the smallest real eigenvalue of $\mathbf{A}$, a number determined by $\mathbf{A}$ alone.

(b) If $s \geq \rho(\mathbf{B})$, then $\mathbf{A}$ is an M-matrix by Definition 19.9.2. Conversely, if $\mathbf{A}$ is an M-matrix, there is a representation $\mathbf{A} = s'\mathbf{I} - \mathbf{B}'$ with $\mathbf{B}' \geq \mathbf{O}$ and $s' \geq \rho(\mathbf{B}')$, and by (a) $s-\rho(\mathbf{B}) = s'-\rho(\mathbf{B}') \geq 0$. For the second statement, $\mathbf{A}$ is invertible if and only if $0 \notin \operatorname{spec}(\mathbf{A})$ (Theorem 9.2.6). If $s > \rho(\mathbf{B})$, every eigenvalue has real part at least $s-\rho(\mathbf{B}) > 0$ by (a), so $0 \notin \operatorname{spec}(\mathbf{A})$. If $s = \rho(\mathbf{B})$, then $0 = s-\rho(\mathbf{B}) \in \operatorname{spec}(\mathbf{A})$ by (a), and $\mathbf{A}$ is singular. This proves the lemma.

■

So the M-matrix test can be run with the most convenient representation, usually the smallest admissible shift $s = \max_i a_{ii}$, and the verdict does not change.

Examples.

- *Size one.* The 1×1 matrix (a) is $s-(s-a)$ for any $s \geq a$, and $\rho(s-a) = s-a$. So (a) is an M-matrix exactly when $s \geq s-a$, that is, $a \geq 0$, and nonsingular exactly when $a > 0$. This degenerate case is worth keeping in mind: an M-matrix is the matrix version of a **nonnegative number**, and a nonsingular one the version of a **positive number**, whose reciprocal is again positive.
- *A standard example.* $\mathbf{A} = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix} = 2\mathbf{I} - \mathbf{B}$ with $\mathbf{B} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$. Here $p_{\mathbf{B}}(x) = x^2 - 1$, so $\operatorname{spec}(\mathbf{B}) = \{1, -1\}$ and $\rho(\mathbf{B}) = 1 < 2$. By Lemma 19.9.2, $\mathbf{A}$ is a nonsingular M-matrix, and indeed $\mathbf{A}^{-1} = \frac{1}{3}\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \geq \mathbf{O}$.

Nonexample by minimal change. Double the off-diagonal entries of the standard example: $\mathbf{A}' = \begin{pmatrix} 2 & -2 \\ -2 & 2 \end{pmatrix} = 2\mathbf{I} - 2\mathbf{B}$, with $\rho(2\mathbf{B}) = 2$. It is still a Z-matrix and still has the smallest shift $s = 2$, but now $s = \rho(2\mathbf{B})$: an M-matrix, and a singular one ($\mathbf{A}'\mathbf{1} = \mathbf{0}$). Go one step further, to $\mathbf{A}'' = \begin{pmatrix} 1 & -2 \\ -2 & 1 \end{pmatrix} = \mathbf{I} - 2\mathbf{B}$. The clause that fails is the comparison of the shift with the spectral radius: here $s = 1 < 2 = \rho(2\mathbf{B})$. By Lemma 19.9.2 its smallest real eigenvalue is $1-2 = -1$, and its inverse, $-\frac{1}{3}\begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}$, is entrywise negative.

Why this definition. Drop the clause $s \geq \rho(\mathbf{B})$ and the definition describes every Z-matrix, including $\mathbf{A}''$, whose inverse has the wrong sign everywhere; the whole point is lost. Drop the sign pattern instead, and ask only that some shift of $\mathbf{A}$ have small spectral radius, and the nonnegativity of the terms $(s^{-1}\mathbf{B})^k$ in the Neumann series disappears with it. The strict inequality $s > \rho(\mathbf{B})$ is what the series needs to converge, and the lemma shows that the boundary

case $s = \rho(\mathbf{B})$ is exactly the singular one, which is why the definition admits it but names the nonsingular case separately.

Warning. **Positive diagonal and nonpositive off-diagonal entries do not make an M-matrix.** The matrix $\mathbf{A}'' = \left(\begin{smallmatrix} 1 & -2 \\ -2 & 1 \end{smallmatrix}\right)$ has both, yet $\det \mathbf{A}'' = -3 < 0$ and its inverse has no nonnegative entry at all. The sign pattern is only half the definition; the size of the off-diagonal entries relative to the diagonal is the other half, and it is measured by $\rho(\mathbf{B})$, not entry by entry.

Quick check. Is $\left(\begin{smallmatrix} 1 & -1 \\ -1 & 1 \end{smallmatrix}\right)$ an M-matrix? Is it a nonsingular one?

Solution. Take $s = 1$ and $\mathbf{B} = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right) \geq \mathbf{O}$, whose spectrum is $\{1, -1\}$, so $\rho(\mathbf{B}) = 1 = s$. By Lemma 19.9.2 (b) the matrix is an M-matrix but not a nonsingular one. Directly: it sends $\mathbf{1}$ to $\mathbf{0}$.

The payoff of the lemma is already visible: for a Z-matrix, the leftmost point of the spectrum is a real eigenvalue, and being a (nonsingular) M-matrix means precisely that this eigenvalue is nonnegative (positive). The theorem below turns that into conditions that can be checked without finding any eigenvalue.

19.9.2 The characterization theorem

Deciding $s > \rho(\mathbf{B})$ directly means locating the eigenvalues of $\mathbf{B}$, which is the one computation we would like to avoid. The theorem gives three kinds of substitute: a certificate (one positive vector), a sign condition on the inverse, and a finite list of determinants. Its proof needs one small observation about certificates.

Lemma 19.9.3. Test Vectors Survive Deletion

Let $\mathbf{A} \in M_n(\mathbb{R})$ be a Z-matrix, and let $\mathbf{x} \in \mathbb{R}^n$ satisfy $\mathbf{x} > \mathbf{0}$ and $\mathbf{Ax} > \mathbf{0}$. Let $S \subseteq [n]$ be nonempty, let $\mathbf{A}_{S,S}$ be the principal submatrix on rows and columns S (Definition 7.7.1), and let $\mathbf{x}_S$ be the vector of the entries x_j, $j \in S$, in increasing order of j. Then $\mathbf{A}_{S,S}$ is a Z-matrix, $\mathbf{x}_S > \mathbf{0}$ and $\mathbf{A}_{S,S}\mathbf{x}_S > \mathbf{0}$.

Idea. Deleting columns throws away only nonpositive contributions to each row, so the surviving rows can only get more positive. That is the whole reason principal minors enter the story.

Proof. The off-diagonal entries of $\mathbf{A}_{S,S}$ are off-diagonal entries of $\mathbf{A}$, so they are ≤ 0, and $\mathbf{x}_S > \mathbf{0}$ because $\mathbf{x} > \mathbf{0}$. Let $i \in S$. The entry of $\mathbf{A}_{S,S}\mathbf{x}_S$ belonging to i is

$$\sum_{j \in S} a_{ij} x_j = (\mathbf{Ax})_i - \sum_{j \notin S} a_{ij} x_j \geq (\mathbf{Ax})_i > 0,$$

because each $j \notin S$ differs from $i \in S$, so $a_{ij} \leq 0$, while $x_j > 0$; every subtracted term is therefore ≤ 0. This proves the lemma.

■

Write $\mathbf{A}_k$ for the leading principal submatrix $\mathbf{A}_{[k],[k]}$, the top-left $k \times k$ corner, as in Sylvester's criterion (Theorem 13.1.5).

Theorem 19.9.4. Characterizations of Nonsingular M-Matrices

Let $\mathbf{A} \in M_n(\mathbb{R})$, $n \geq 1$, be a Z-matrix. The following are equivalent.

(a) $\mathbf{A}$ is a nonsingular M-matrix.
(b) $\mathbf{A}$ is invertible and $\mathbf{A}^{-1} \geq \mathbf{O}$.
(c) There is $\mathbf{x} \in \mathbb{R}^n$ with $\mathbf{x} > \mathbf{0}$ and $\mathbf{A}\mathbf{x} > \mathbf{0}$.
(d) Every eigenvalue of $\mathbf{A}$ in $\mathbb{C}$ has positive real part.
(e) Every real eigenvalue of $\mathbf{A}$ is positive.
(f) Every principal minor of $\mathbf{A}$ is positive: $\det \mathbf{A}_{S,S} > 0$ for every nonempty $S \subseteq [n]$.
(g) Every leading principal minor of $\mathbf{A}$ is positive: $\det \mathbf{A}_k > 0$ for $k = 1, \ldots, n$.

Idea. Three loops, all passing through (a), (b) or (c).

① (a) ⇒ (b) ⇒ (c) ⇒ (a). The Neumann series writes $\mathbf{A}^{-1}$ as a sum of nonnegative matrices. The vector $\mathbf{A}^{-1}\mathbf{1}$ is then a certificate for (c). And a certificate $\mathbf{x}$ bounds $\rho(\mathbf{B})$: rescaling the coordinates by $\mathbf{x}$ turns $\mathbf{B}\mathbf{x} < s\mathbf{x}$ into "every row sum is less than s", and Chapter 16 bounds a spectral radius by row sums.

② (a) ⇒ (d) ⇒ (e) ⇒ (a). This is Lemma 19.9.2 read three ways.

③ (c) ⇒ (f) ⇒ (g) ⇒ (b). A certificate survives deletion (Lemma 19.9.3), so every principal submatrix has a nonnegative inverse by ①, and one cofactor of that inverse compares two nested minors. The way back from (g) is the induction of Sylvester's criterion: peel off the last row and column, and read the sign of every block of the inverse from Chapter 8's block inverse formula.

Proof. **(a) ⇒ (b).** By Lemma 19.9.2 (b), $\mathbf{A} = s\mathbf{I} - \mathbf{B}$ with $\mathbf{B} \geq \mathbf{O}$ and $s > \rho(\mathbf{B}) \geq 0$; in particular $s > 0$. Put $\mathbf{C} = s^{-1}\mathbf{B} \geq \mathbf{O}$. Its eigenvalues are the numbers μ/s, $\mu \in \operatorname{spec}(\mathbf{B})$, so $\rho(\mathbf{C}) = \rho(\mathbf{B})/s < 1$. By Theorem 10.10.6, $\mathbf{I} - \mathbf{C}$ is invertible and $(\mathbf{I} - \mathbf{C})^{-1} = \sum_{k \geq 0} \mathbf{C}^k$. Every partial sum $\sum_{k=0}^{K} \mathbf{C}^k$ is $\geq \mathbf{O}$, since sums and products of nonnegative matrices are nonnegative; a nonstrict inequality survives a limit (the algebra of limits quoted in Chapter 16's introduction), so $(\mathbf{I} - \mathbf{C})^{-1} \geq \mathbf{O}$. Hence $\mathbf{A} = s(\mathbf{I} - \mathbf{C})$ is invertible with $\mathbf{A}^{-1} = s^{-1}(\mathbf{I} - \mathbf{C})^{-1} \geq \mathbf{O}$.

(b) ⇒ (c). Put $\mathbf{x} = \mathbf{A}^{-1}\mathbf{1}$, so that $\mathbf{A}\mathbf{x} = \mathbf{1} > \mathbf{0}$. Each $x_i = \sum_j (\mathbf{A}^{-1})_{ij}$ is a sum of nonnegative numbers, and it is 0 only if row i of $\mathbf{A}^{-1}$ is zero. That cannot happen: row i of $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$ is $\mathbf{e}_i^{\top} \neq \mathbf{0}^{\top}$. Hence $\mathbf{x} > \mathbf{0}$.

(c) ⇒ (a). Let $\mathbf{x} > \mathbf{0}$ with $\mathbf{A}\mathbf{x} > \mathbf{0}$. Choose $s \geq \max_i a_{ii}$, so that $\mathbf{B} = s\mathbf{I} - \mathbf{A} \geq \mathbf{O}$ by Lemma 19.9.1. Then $\mathbf{B}\mathbf{x} = s\mathbf{x} - \mathbf{A}\mathbf{x}$, so $(\mathbf{B}\mathbf{x})_i < sx_i$ for every i. Let $\mathbf{D} = \operatorname{diag}(x_1, \ldots, x_n)$,

invertible because every $x_i > 0$. The matrix $\mathbf{D}^{-1}\mathbf{B}\mathbf{D}$ has entries $b_{ij}x_j/x_i \geq 0$, and its i-th row sum is $(\mathbf{B}\mathbf{x})_i/x_i < s$. By Corollary 16.4.2, $\rho(\mathbf{D}^{-1}\mathbf{B}\mathbf{D})$ is at most the largest of these row sums, so $\rho(\mathbf{D}^{-1}\mathbf{B}\mathbf{D}) < s$. Similar matrices have the same characteristic polynomial (Theorem 7.10.3), hence the same eigenvalues, so $\rho(\mathbf{B}) < s$. By Lemma 19.9.2 (b), $\mathbf{A}$ is a nonsingular M-matrix.

(a) $\Rightarrow$ (d). Write $\mathbf{A} = s\mathbf{I} - \mathbf{B}$ with $\mathbf{B} \geq \mathbf{O}$ and $s > \rho(\mathbf{B})$, by Lemma 19.9.2 (b). By part (a) of the same lemma, every $\lambda \in \operatorname{spec}(\mathbf{A})$ has $\operatorname{Re}\lambda \geq s - \rho(\mathbf{B}) > 0$.

(d) $\Rightarrow$ (e). A real eigenvalue is its own real part.

(e) $\Rightarrow$ (a). Write $\mathbf{A} = s\mathbf{I} - \mathbf{B}$ with $\mathbf{B} \geq \mathbf{O}$ (Lemma 19.9.1). By Lemma 19.9.2 (a), $s - \rho(\mathbf{B})$ is a real eigenvalue of $\mathbf{A}$, so it is positive by (e). Hence $s > \rho(\mathbf{B})$, and $\mathbf{A}$ is a nonsingular M-matrix by part (b) of the lemma.

(c) $\Rightarrow$ (f). Let $\mathbf{x} > \mathbf{0}$ with $\mathbf{A}\mathbf{x} > \mathbf{0}$. We show $\det \mathbf{A}_{S,S} > 0$ by induction on $k = |S|$. If $k = 1$, say $S = \{i\}$, then $a_{ii}x_i > 0$ by Lemma 19.9.3, so $\det \mathbf{A}_{S,S} = a_{ii} > 0$. Let $k \geq 2$, and suppose every principal minor of size $k-1$ is positive. Put $\mathbf{N} = \mathbf{A}_{S,S} \in M_k(\mathbb{R})$. By Lemma 19.9.3, $\mathbf{N}$ is a Z-matrix satisfying (c), so the implications (c) $\Rightarrow$ (a) $\Rightarrow$ (b), already proved for Z-matrices of every size, show that $\mathbf{N}$ is invertible with $\mathbf{N}^{-1} \geq \mathbf{O}$. In particular $\det \mathbf{N} \neq 0$ (Theorem 7.5.2). Deleting the last row and column of $\mathbf{N}$ leaves $\mathbf{A}_{S',S'}$, where S' is S without its largest element; so the (k,k) cofactor of $\mathbf{N}$ is $\det \mathbf{A}_{S',S'}$, which is positive by the induction hypothesis. By Corollary 7.6.3, $\mathbf{N}^{-1} = (\det \mathbf{N})^{-1} \operatorname{adj} \mathbf{N}$, whose (k,k)-entry is therefore

$$(\mathbf{N}^{-1})_{kk} = \frac{\det \mathbf{A}_{S',S'}}{\det \mathbf{N}}.$$

This number is ≥ 0 because $\mathbf{N}^{-1} \geq \mathbf{O}$, and nonzero because its numerator is. So it is positive, and $\det \mathbf{N} = \det \mathbf{A}_{S',S'}/(\mathbf{N}^{-1})_{kk} > 0$.

(f) $\Rightarrow$ (g). A leading principal minor is the principal minor with $S = [k]$.

(g) $\Rightarrow$ (b). By induction on n. If $n = 1$, then $\mathbf{A} = (a_{11})$ with $a_{11} = \det \mathbf{A}_1 > 0$, and $\mathbf{A}^{-1} = (1/a_{11}) \geq \mathbf{O}$. Let $n \geq 2$, and assume the implication for Z-matrices of size $n-1$. Partition

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{n-1} & \mathbf{b} \\ \mathbf{c}^{\top} & a_{nn} \end{pmatrix}, \qquad \mathbf{b}, \mathbf{c} \in \mathbb{R}^{n-1}.$$

The corner $\mathbf{A}_{n-1}$ is a Z-matrix whose leading principal minors $\det \mathbf{A}_1, \dots, \det \mathbf{A}_{n-1}$ are positive, so by the induction hypothesis it is invertible with $\mathbf{A}_{n-1}^{-1} \geq \mathbf{O}$. The entries of $\mathbf{b}$ and $\mathbf{c}$ are off-diagonal entries of $\mathbf{A}$, so $\mathbf{b} \leq \mathbf{0}$ and $\mathbf{c} \leq \mathbf{0}$. By Theorem 8.3.2 (a), $\det \mathbf{A} = \det \mathbf{A}_{n-1} \cdot \sigma$, where $\sigma = a_{nn} - \mathbf{c}^{\top}\mathbf{A}_{n-1}^{-1}\mathbf{b}$ is the Schur complement of $\mathbf{A}_{n-1}$; since $\det \mathbf{A} > 0$ and $\det \mathbf{A}_{n-1} > 0$, $\sigma > 0$. By Theorem 8.3.4 (a), $\mathbf{A}$ is invertible and

$$\mathbf{A}^{-1} = \begin{pmatrix} \mathbf{A}_{n-1}^{-1} + \sigma^{-1}\mathbf{u}\mathbf{w}^{\top} & -\sigma^{-1}\mathbf{u} \\ -\sigma^{-1}\mathbf{w}^{\top} & \sigma^{-1} \end{pmatrix}, \qquad \mathbf{u} = \mathbf{A}_{n-1}^{-1}\mathbf{b}, \quad \mathbf{w}^{\top} = \mathbf{c}^{\top}\mathbf{A}_{n-1}^{-1}.$$

Now $\mathbf{u} \leq \mathbf{0}$, being a nonnegative matrix times a nonpositive vector, and likewise $\mathbf{w} \leq \mathbf{0}$. Hence $\mathbf{u}\mathbf{w}^{\top} \geq \mathbf{O}$, each of its entries being a product of two nonpositive numbers, and $-\sigma^{-1}\mathbf{u} \geq \mathbf{0}$, $-\sigma^{-1}\mathbf{w} \geq \mathbf{0}$, $\sigma^{-1} > 0$. Every block of $\mathbf{A}^{-1}$ is therefore nonnegative, so $\mathbf{A}^{-1} \geq \mathbf{O}$.

The three loops connect all seven conditions, which proves the theorem.

■

Each condition has its use. Condition (c) is a **certificate**: to prove that a given Z-matrix is a nonsingular M-matrix, exhibit one positive vector and multiply. Condition (g) is a **finite test** of n determinants, the counterpart for Z-matrices of Sylvester's criterion for Hermitian matrices, and condition (b) is the **payoff**: $\mathbf{Ax} = \mathbf{b}$ with $\mathbf{b} \geq \mathbf{0}$ has exactly one solution, and it is $\geq \mathbf{0}$. Condition (d) links the theorem to differential equations: $-\mathbf{A}$ is stable, so the solutions of $\mathbf{x}' = -\mathbf{Ax}$ decay (exercise C3).

Remark. The two determinant conditions need the Z-hypothesis, and they need **strict** positivity. The matrix $\left(\begin{smallmatrix} 1 & 2 \\ 0 & 1 \end{smallmatrix}\right)$ has all three principal minors equal to 1, but it is not a Z-matrix and its inverse $\left(\begin{smallmatrix} 1 & -2 \\ 0 & 1 \end{smallmatrix}\right)$ has a negative entry. And the Z-matrix $\left(\begin{smallmatrix} 0 & -1 \\ 0 & -1 \end{smallmatrix}\right)$ has leading principal minors 0 and 0, both nonnegative, but its eigenvalue -1 shows it is not an M-matrix of any kind (exercise C2).

Remark. **Stated, not proved.** There is a companion statement for M-matrices that may be singular: *a Z-matrix is an M-matrix if and only if every principal minor is nonnegative.* Exercise C2 proves the forward direction by a limiting argument; the converse is not proved in this book and is used nowhere. Longer lists of equivalent conditions exist in the literature; the seven above are the ones proved here.

Warning. **A nonnegative inverse does not make a matrix a Z-matrix, and so does not make it an M-matrix.** Condition (b) characterizes nonsingular M-matrices only **among Z-matrices**. Take

$$\mathbf{W} = \begin{pmatrix} 1 & -1 & 1 \\ 1 & 1 & -1 \\ -1 & 1 & 1 \end{pmatrix}, \qquad \mathbf{W}^{-1} = \frac{1}{2}\begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{pmatrix} \geq \mathbf{O}.$$

Multiplying out confirms $\mathbf{W}\mathbf{W}^{-1} = \mathbf{I}$: for instance row 1 of $\mathbf{W}$ against the columns of $2\mathbf{W}^{-1}$ gives $1 - 0 + 1 = 2$, $1 - 1 + 0 = 0$, and $0 - 1 + 1 = 0$. The diagonal of $\mathbf{W}$ is positive, its principal minors are $1, 1, 1, 2, 2, 2, 4$, $\mathbf{W1} = \mathbf{1} > \mathbf{0}$, and its eigenvalues 1 and $1 \pm i\sqrt{3}$ have positive real parts. So $\mathbf{W}$ satisfies every one of (b)–(g), yet $w_{13} = 1 > 0$: it is not a Z-matrix, and not an M-matrix.

Quick check. Let $\mathbf{A} = \left(\begin{smallmatrix} a & -b \\ -c & d \end{smallmatrix}\right)$ with $b, c \geq 0$. Use the theorem to write down exactly when $\mathbf{A}$ is a nonsingular M-matrix, and explain why the condition $d > 0$ need not be listed.

Solution. $\mathbf{A}$ is a Z-matrix, so by (g) it is a nonsingular M-matrix if and only if $a > 0$ and $ad - bc > 0$. Then $ad > bc \geq 0$ and $a > 0$ force $d > 0$; so condition (f), which also asks for $d > 0$, adds nothing, as the theorem says it cannot.

A symmetric Z-matrix has one more characterization, because condition (g) is Sylvester's criterion word for word.

Corollary 19.9.5. Symmetric M-Matrices Are Positive Definite

A symmetric Z-matrix $\mathbf{A} \in M_n(\mathbb{R})$ is a nonsingular M-matrix if and only if $\mathbf{A} \succ \mathbf{O}$.

Proof. By Theorem 19.9.4, $\mathbf{A}$ is a nonsingular M-matrix if and only if $\det \mathbf{A}_k > 0$ for every k. By Sylvester's criterion, Theorem 13.1.5 (d), a real symmetric matrix satisfies this if and only if $\mathbf{A} \succ \mathbf{O}$. ∎

The Z-hypothesis cannot be dropped here either: $\left(\begin{smallmatrix} 2 & 1 \\ 1 & 2 \end{smallmatrix}\right) \succ \mathbf{O}$, but its inverse $\frac{1}{3}\left(\begin{smallmatrix} 2 & -1 \\ -1 & 2 \end{smallmatrix}\right)$ has negative entries. Entrywise order and the Loewner order answer different questions, and the corollary is a place where they meet.

19.9.3 The second-difference matrix

The first application is the one that gave M-matrices their importance. To solve $-u''(t) = f(t)$ on $[0,1]$ with $u(0) = u(1) = 0$, choose n, put $h = 1/(n+1)$ and $t_i = ih$, and replace $u(t_i)$ by an unknown u_i and $-u''(t_i)$ by the second difference $(2u_i - u_{i-1} - u_{i+1})/h^2$, with $u_0 = u_{n+1} = 0$. The equations become

$$\frac{1}{h^2}\mathbf{K}_n\mathbf{u} = \mathbf{f}, \qquad \mathbf{K}_n = \begin{pmatrix} 2 & -1 & & \\ -1 & 2 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{pmatrix} \in M_n(\mathbb{R}),$$

where $\mathbf{f} = (f(t_1), \dots, f(t_n))$. Physically u is the sag of a string under the load f, and a load pushing everywhere in one direction should never make any part of the string move the other way. In matrix language: $\mathbf{K}_n^{-1}$ should be nonnegative.

$\mathbf{K}_n$ is a Z-matrix, but the obvious certificate fails. Its rows sum to $1, 0, \dots, 0, 1$, so $\mathbf{K}_n\mathbf{1}$ has zero entries once $n \geq 3$: the matrix is diagonally dominant, but only weakly. The continuous problem suggests a better test vector. The function $u(t) = t(1-t)$ solves $-u'' = 2$ with zero boundary values, so try its samples, scaled to be integers.

Proposition 19.9.6. The Second-Difference Matrix

Let $n \geq 1$. Then $\mathbf{K}_n$ is a nonsingular M-matrix, and its inverse is the matrix with entries

$$(\mathbf{K}_n^{-1})_{ij} = \frac{\min(i,j)\,(n+1-\max(i,j))}{n+1}, \qquad 1 \leq i, j \leq n.$$

In particular every entry of $\mathbf{K}_n^{-1}$ is positive.

Idea. For the M-matrix property, feed $\mathbf{K}_n$ the vector $x_i = i(n+1-i)$, a sampled parabola: second differences of a quadratic are constant, so $\mathbf{K}_n\mathbf{x} = 2\mathbf{1}$. For the inverse, guess column j from the continuous picture: the response to a point load at t_j is a tent, linear on each side

of j and zero at both ends. The second difference of a function that is linear on each side of j vanishes except at the kink, and there it equals the change in slope.

Proof. The certificate. Put $x_i = i(n+1-i)$ for $0 \le i \le n+1$, so that $x_0 = x_{n+1} = 0$ and $x_i > 0$ for $1 \le i \le n$. For $1 \le i \le n$, the i-th entry of $\mathbf{K}_n\mathbf{x}$ is $2x_i - x_{i-1} - x_{i+1}$, where the terms x_0 and x_{n+1}, which do not occur in rows 1 and n, are zero anyway. Writing $x_i = (n+1)i - i^2$, the linear part contributes $(n+1)(2i - (i-1) - (i+1)) = 0$, and the quadratic part contributes

$$-2i^2 + (i-1)^2 + (i+1)^2 = 2.$$

So $\mathbf{K}_n\mathbf{x} = 2\mathbf{1} > \mathbf{0}$ with $\mathbf{x} > \mathbf{0}$, and Theorem 19.9.4, (c) $\Rightarrow$ (a), shows that $\mathbf{K}_n$ is a nonsingular M-matrix.

The inverse. Let $\mathbf{G}$ be the matrix given by the formula, fix a column j, and put $g_i = \min(i,j)(n+1-\max(i,j))/(n+1)$ for $0 \le i \le n+1$. Then

$$g_i = \begin{cases} i\,\dfrac{n+1-j}{n+1} & (0 \le i \le j), \\[2mm] j\,\dfrac{n+1-i}{n+1} & (j \le i \le n+1), \end{cases}$$

the two lines agreeing at $i = j$. So $g_0 = g_{n+1} = 0$, and (g_i) is an affine function of i on $0 \le i \le j$, with slope $(n+1-j)/(n+1)$, and on $j \le i \le n+1$, with slope $-j/(n+1)$. For $1 \le i \le n$, entry i of $\mathbf{K}_n$ times column j of $\mathbf{G}$ is $2g_i - g_{i-1} - g_{i+1} = (g_i - g_{i-1}) - (g_{i+1} - g_i)$. If $i < j$ or $i > j$, the three points $i-1, i, i+1$ lie on one side of j, both differences are the same slope, and the entry is 0. If $i = j$, it is

$$\frac{n+1-j}{n+1} - \left(-\frac{j}{n+1}\right) = 1.$$

Hence $\mathbf{K}_n\mathbf{G} = \mathbf{I}$, so $\mathbf{G} = \mathbf{K}_n^{-1}$. Finally $\min(i,j) \ge 1$ and $n+1-\max(i,j) \ge 1$, so every entry is at least $1/(n+1) > 0$.

■

The matrix $\mathbf{K}_n^{-1}$ is a **discrete Green's function**: column j is the sag produced by a unit load at the single point t_j, and positivity of every entry says that a load at one point moves every point of the string, in the same direction. The nonnegativity of $\mathbf{K}_n^{-1}$ is the discrete **maximum principle**: if $\mathbf{K}_n\mathbf{u} \ge \mathbf{0}$ then $\mathbf{u} \ge \mathbf{0}$. Since $\mathbf{K}_n$ is symmetric, Corollary 19.9.5 also shows $\mathbf{K}_n \succ \mathbf{O}$, which Chapter 7 foreshadowed by computing $\det \mathbf{K}_n = n+1$ in Exercise (B2: The second-difference matrix).

Example. The Case n = 4

Write down $\mathbf{K}_4^{-1}$, check one column against $\mathbf{K}_4$, and recover the certificate $\mathbf{x}$ from it.

Solution. With $n+1=5$, Proposition 19.9.6 gives

$$\mathbf{K}_4^{-1} = \frac{1}{5}\begin{pmatrix} 4 & 3 & 2 & 1 \\ 3 & 6 & 4 & 2 \\ 2 & 4 & 6 & 3 \\ 1 & 2 & 3 & 4 \end{pmatrix}.$$

Check of column 2. $\mathbf{K}_4$ applied to $(3,6,4,2)$ gives $(6-6,\ -3+12-4,\ -6+8-2,\ -4+4) = (0,5,0,0)$, and dividing by 5 gives $\mathbf{e}_2$, as it must. Each column is a tent: column 2 rises by 3 per step up to row 2 and then falls by 2 per step towards the zero at the fictitious row 5.

The certificate. The row sums of $5\mathbf{K}_4^{-1}$ are $10, 15, 15, 10$, so $\mathbf{K}_4^{-1}\mathbf{1} = (2,3,3,2)$, which is $\frac{1}{2}(1\cdot 4, 2\cdot 3, 3\cdot 2, 4\cdot 1) = \frac{1}{2}\mathbf{x}$. This is the proof of (b) $\Rightarrow$ (c) run backwards: the vector $\mathbf{A}^{-1}\mathbf{1}$ is always a certificate, and here it is the sampled parabola.

19.9.4 The Leontief input–output model

The second application comes from economics. An economy has n sectors, each producing one good. To produce one unit of good j, sector j consumes $c_{ij} \geq 0$ units of good i, so column j of the **consumption matrix** $\mathbf{C} = (c_{ij}) \geq \mathbf{O}$ is sector j's recipe. A production plan $\mathbf{x} \geq \mathbf{0}$ uses up $\mathbf{Cx}$ as inputs and leaves the net output $\mathbf{x} - \mathbf{Cx}$ for outside use. Given a final demand $\mathbf{d} \geq \mathbf{0}$, the planner needs

$$(\mathbf{I} - \mathbf{C})\mathbf{x} = \mathbf{d}, \qquad \mathbf{x} \geq \mathbf{0}.$$

The economy is **productive** if this has a solution for **every** $\mathbf{d} \geq \mathbf{0}$. The matrix $\mathbf{I} - \mathbf{C}$ is a Z-matrix, and the theorem answers the question completely.

Corollary 19.9.7. Productive Economies

Let $\mathbf{C} \in M_n(\mathbb{R})$ with $\mathbf{C} \geq \mathbf{O}$. The following are equivalent.

(i) For every $\mathbf{d} \geq \mathbf{0}$ there is $\mathbf{x} \geq \mathbf{0}$ with $(\mathbf{I} - \mathbf{C})\mathbf{x} = \mathbf{d}$.

(ii) There is $\mathbf{x} \geq \mathbf{0}$ with $\mathbf{x} - \mathbf{Cx} > \mathbf{0}$: some plan leaves a surplus of every good.

(iii) $\rho(\mathbf{C}) < 1$.

(iv) $\mathbf{I} - \mathbf{C}$ is invertible and $(\mathbf{I} - \mathbf{C})^{-1} \geq \mathbf{O}$.

(v) Every leading principal minor of $\mathbf{I} - \mathbf{C}$ is positive.

When they hold, the solution in (i) is unique, and $(\mathbf{I} - \mathbf{C})^{-1} = \mathbf{I} + \mathbf{C} + \mathbf{C}^2 + \cdots$. They hold in particular when every column sum of $\mathbf{C}$ is less than 1.

Idea. The matrix $\mathbf{I} - \mathbf{C}$ is a Z-matrix, so this is the main theorem read in economic terms: (iii) is (a) with $s = 1$; (ii), (iv) and (v) are (c), (b) and (g), where in (ii) a plan $\mathbf{x} \geq \mathbf{0}$ with a surplus is automatically positive, since $\mathbf{x} > \mathbf{Cx} \geq \mathbf{0}$; and (i) sits between (iv) and (ii).

Proof. $\mathbf{I}-\mathbf{C}$ is a Z-matrix with the representation $s = 1$, $\mathbf{B} = \mathbf{C}$, so (iii) says, by Lemma 19.9.2 (b), that $\mathbf{I} - \mathbf{C}$ is a nonsingular M-matrix. By Theorem 19.9.4, (iii) is therefore equivalent to (iv) (condition (b) there) and to (v) (condition (g)). It is also equivalent to (ii): if $\mathbf{x} \geq \mathbf{0}$ and

$\mathbf{x} - \mathbf{C}\mathbf{x} > \mathbf{0}$, then $\mathbf{x} > \mathbf{C}\mathbf{x} \geq \mathbf{0}$, so $\mathbf{x} > \mathbf{0}$ and condition (c) of the theorem holds; conversely (c) supplies such an $\mathbf{x}$.

(iv) $\Rightarrow$ (i): given $\mathbf{d} \geq \mathbf{0}$, the vector $\mathbf{x} = (\mathbf{I} - \mathbf{C})^{-1}\mathbf{d}$ is a product of nonnegative factors, hence $\geq \mathbf{0}$, and it solves the system. (i) $\Rightarrow$ (ii): take $\mathbf{d} = \mathbf{1}$.

When they hold, $\mathbf{I} - \mathbf{C}$ is invertible, so the solution is unique, and $\rho(\mathbf{C}) < 1$ gives the series by Theorem 10.10.6. Finally, if every column sum of $\mathbf{C}$ is less than 1, then $\rho(\mathbf{C}) \leq \|\mathbf{C}\|_1 < 1$ by Corollary 16.4.2, which is (iii).

■

Condition (v) is known to economists as the Hawkins–Simon condition, and the series has a direct reading. To deliver $\mathbf{d}$, the economy must first produce $\mathbf{d}$; that production consumes $\mathbf{C}\mathbf{d}$, which must also be produced; that consumes $\mathbf{C}^2\mathbf{d}$; and so on. The Leontief inverse $(\mathbf{I} - \mathbf{C})^{-1}$ adds up all the rounds, and its (i, j)-entry is the total amount of good i that the economy must produce, directly and indirectly, to deliver one unit of good j.

Example. Three Sectors

An economy has three sectors, agriculture, manufacturing, and services, with consumption matrix

$$\mathbf{C} = \frac{1}{5}\begin{pmatrix} 1 & 1 & 1 \\ 2 & 0 & 1 \\ 0 & 2 & 1 \end{pmatrix}.$$

Show that it is productive, find $\rho(\mathbf{C})$ exactly, compute the Leontief inverse, and find the production plan meeting the demand $\mathbf{d} = (12, 6, 6)$.

Solution. Productive. Every column of $5\mathbf{C}$ sums to 3, so every column sum of $\mathbf{C}$ is $\frac{3}{5} < 1$, and Corollary 19.9.7 applies.

The spectral radius. Expanding $\det(x\mathbf{I} - \mathbf{C})$ along the first row gives

$$p_{\mathbf{C}}(x) = x^3 - \tfrac{2}{5}x^2 - \tfrac{3}{25}x = x(x - \tfrac{3}{5})(x + \tfrac{1}{5}),$$

so $\operatorname{spec}(\mathbf{C}) = \{0, \frac{3}{5}, -\frac{1}{5}\}$ and $\rho(\mathbf{C}) = \frac{3}{5}$, equal to the column-sum bound $\|\mathbf{C}\|_1$ of Corollary 16.4.2. The bound is attained because all column sums are equal: $\mathbf{C}^{\top}\mathbf{1} = \frac{3}{5}\mathbf{1}$.

The Leontief inverse. Here $\mathbf{I} - \mathbf{C} = \frac{1}{5}\begin{pmatrix} 4 & -1 & -1 \\ -2 & 5 & -1 \\ 0 & -2 & 4 \end{pmatrix}$, and

$$(\mathbf{I} - \mathbf{C})^{-1} = \frac{1}{6}\begin{pmatrix} 9 & 3 & 3 \\ 4 & 8 & 3 \\ 2 & 4 & 9 \end{pmatrix}.$$

To check, multiply $\begin{pmatrix} 4 & -1 & -1 \\ -2 & 5 & -1 \\ 0 & -2 & 4 \end{pmatrix}$ by $\begin{pmatrix} 9 & 3 & 3 \\ 4 & 8 & 3 \\ 2 & 4 & 9 \end{pmatrix}$: the first row gives $36 - 4 - 2 = 30$, $12 - 8 - 4 = 0$, $12 - 3 - 9 = 0$; the second $-18 + 20 - 2 = 0$, $-6 + 40 - 4 = 30$, $-6 + 15 - 9 = 0$; the third $-8 + 8 = 0$, $-16 + 16 = 0$, $-6 + 36 = 30$. The product is $30\mathbf{I}$, and $\frac{1}{5} \cdot \frac{1}{6} \cdot 30 = 1$. Every entry of the inverse is positive: extra demand for any one good raises the output of every

sector. For instance, one extra unit of services requires $\frac{1}{2}$ unit of agricultural output in total, although services consume only $\frac{1}{5}$ unit of it directly.

The plan.

$$\mathbf{x} = (\mathbf{I} - \mathbf{C})^{-1}\mathbf{d} = \frac{1}{6}\begin{pmatrix} 108 + 18 + 18 \\ 48 + 48 + 18 \\ 24 + 24 + 54 \end{pmatrix} = \begin{pmatrix} 24 \\ 19 \\ 17 \end{pmatrix}.$$

Check: $\mathbf{Cx} = \frac{1}{5}(24 + 19 + 17,\ 48 + 17,\ 38 + 17) = (12, 13, 11)$, and $\mathbf{x} - \mathbf{Cx} = (12, 6, 6) = \mathbf{d}$. Delivering this demand takes 24 units of agricultural output, twice the 12 that leave the economy.

Quick check. A two-sector economy has $\mathbf{C} = \left(\begin{smallmatrix} 1/2 & 1 \\ 1/2 & 0 \end{smallmatrix}\right)$. Is it productive?

Solution. No. The leading principal minors of $\mathbf{I} - \mathbf{C} = \left(\begin{smallmatrix} 1/2 & -1 \\ -1/2 & 1 \end{smallmatrix}\right)$ are $\frac{1}{2}$ and $\frac{1}{2} - \frac{1}{2} = 0$, so (v) of Corollary 19.9.7 fails. Indeed $\mathbf{C}\left(\begin{smallmatrix} 2 \\ 1 \end{smallmatrix}\right) = \left(\begin{smallmatrix} 2 \\ 1 \end{smallmatrix}\right)$: running the sectors in the ratio $2 : 1$ exactly reproduces its own inputs and leaves nothing over, and $\rho(\mathbf{C}) = 1$.

19.9.5 Summary and transfer

The chapter set out to explain what the sign of the entries alone forces on the spectrum. For a positive matrix the answer is Perron's theorem (§01, Positive matrices and Perron's theorem): the spectral radius is a simple eigenvalue with a positive eigenvector, and it strictly dominates every other eigenvalue in modulus. Irreducibility, a property of the zero pattern read off a directed graph (§02, Irreducibility and graphs), is what keeps most of that conclusion alive when zeros are allowed (§03, The Perron–Frobenius theorem), and without it one still has $\rho(\mathbf{B})$ as an eigenvalue with a nonnegative eigenvector. From there the chapter drew the long-run behavior of powers (§05, Primitive matrices), Markov chains (§06, Markov chains), the extreme points of the doubly stochastic matrices (§07, Birkhoff's theorem), a ranking of web pages (§08, PageRank), and in this section the matrices $s\mathbf{I} - \mathbf{B}$. Six moves did the work.

- **Compare entrywise, then take the spectral radius.** Nonnegativity makes the spectral radius monotone and lets $|\mathbf{Ax}| \le |\mathbf{A}||\mathbf{x}|$ replace cancellation by an inequality (§01, Positive matrices and Perron's theorem). *Transfer:* to bound the spectral radius of a matrix you do not understand, dominate it entrywise by one you do.
- **A positive test vector bounds ρ from both sides.** The ratios $(\mathbf{Ax})_i/x_i$ for a single $\mathbf{x} > \mathbf{0}$ trap $\rho(\mathbf{A})$ between their minimum and maximum (§04, The Collatz–Wielandt formula), and in this section one $\mathbf{x} > \mathbf{0}$ with $\mathbf{Ax} > \mathbf{0}$ certified a nonsingular M-matrix. The mechanism used here is the rescaling $\mathbf{D}^{-1}\mathbf{BD}$ with $\mathbf{D} = \operatorname{diag}(\mathbf{x})$, which turns $\mathbf{Bx}$ into row sums. *Transfer:* guess a positive vector from the structure of the problem, the parabola for $\mathbf{K}_n$, and verify it by one multiplication.
- **Read the zero pattern as a graph.** Which entries of $\mathbf{A}^k$ are positive depends only on which entries of $\mathbf{A}$ are, and that is a question about paths (§02, Irreducibility and graphs; §05, Primitive matrices). *Transfer:* when only signs matter, draw the graph and count paths.
- **Prove it for positive matrices, then reach the boundary.** Perron's theorem is proved

where every entry is positive, and the nonnegative and irreducible cases are reached from it through positive matrices built from $\mathbf{A}$ (§03, The Perron–Frobenius theorem). *Transfer:* when a hypothesis is an inequality that may hold with equality, prove the strict case and ask what survives the passage to the boundary; $s > \rho(\mathbf{B})$ versus $s \geq \rho(\mathbf{B})$ in this section is the same division.

- **Divide by the Perron root, then use Chapter 10.** Once $\rho(\mathbf{A})$ is known to be a simple eigenvalue dominating the others, $\mathbf{A}/\rho(\mathbf{A})$ meets the hypotheses of the convergence theorems of Chapter 10 §10, and no diagonalizability is ever needed (§05, Primitive matrices; §06, Markov chains). *Transfer:* normalize the dominant eigenvalue to 1 before asking about limits.
- **Invert with a series of nonnegative terms.** If $\mathbf{B} \geq \mathbf{O}$ and $\rho(\mathbf{B}) < 1$, then $(\mathbf{I}-\mathbf{B})^{-1} = \sum_k \mathbf{B}^k \geq \mathbf{O}$. This is the fundamental matrix of an absorbing chain (§06, Markov chains), the Leontief inverse, and the nonnegativity of the inverse of every nonsingular M-matrix. *Transfer:* to prove that an inverse is nonnegative, or to interpret its entries, expand it as the sum of the rounds of a process.

The chapter also paid promises made in earlier chapters.

The promise	Made in	Paid by
"Chapter 19 proves the Perron–Frobenius theorem, which gives the conclusions of Theorem 9.11.4 for every stochastic matrix with some power having all entries positive, without assuming diagonalizability and without computing eigenvalues"	Chapter 9 §11	§06 (Stochastic matrices and Markov chains), resting on §03 (The Perron–Frobenius theorem) and §05 (Primitive matrices and the limit theorem)
"We do not prove Birkhoff's theorem here; Chapter 19 does"	Chapter 18 §08, and again in its summary, §11	§07 (Doubly stochastic matrices and Birkhoff's theorem)

Some things the chapter did not do. The characterization of possibly singular M-matrices by nonnegative principal minors was stated here and proved in one direction only. The theorem of Hardy, Littlewood, and Pólya, that $\mathbf{x} \prec \mathbf{y}$ exactly when $\mathbf{x} = \mathbf{D}\mathbf{y}$ for a doubly stochastic $\mathbf{D}$, belongs to Chapter 21. And the power method, the algorithm that computes a Perron vector at the size of the web, belongs to Chapter 24.

19.9.6 Exercises

19.9.6.1 A. Check your understanding

Exercise. A1

(a) Define a Z-matrix and a nonsingular M-matrix.

(b) True or false: every Z-matrix with positive diagonal entries is a nonsingular M-matrix. Justify your answer.

(c) True or false: every principal submatrix of a nonsingular M-matrix is a nonsingular M-matrix. Justify your answer.

(d) True or false: if $\mathbf{A}$ is invertible and $\mathbf{A}^{-1} \geq \mathbf{O}$, then $\mathbf{A}$ is an M-matrix. Justify your answer.

(e) True or false: the inverse of a nonsingular M-matrix has every entry positive. Justify your answer.

(f) Name a vector $\mathbf{x} > \mathbf{0}$ with $\mathbf{K}_n\mathbf{x} > \mathbf{0}$, and explain why $\mathbf{x} = \mathbf{1}$ does not work when $n \geq 3$.

Solution.

(a) $\mathbf{A} \in M_n(\mathbb{R})$ is a Z-matrix if $a_{ij} \leq 0$ for all $i \neq j$ (Definition 19.9.1). It is a nonsingular M-matrix if $\mathbf{A} = s\mathbf{I} - \mathbf{B}$ with $\mathbf{B} \geq \mathbf{O}$ and $s \geq \rho(\mathbf{B})$, and $\mathbf{A}$ is invertible; equivalently, $s > \rho(\mathbf{B})$ (Definition 19.9.2, Lemma 19.9.2 (b)).

(b) False. $\left(\begin{smallmatrix} 1 & -2 \\ -2 & 1 \end{smallmatrix}\right)$ is a Z-matrix with positive diagonal, but its determinant is $1 - 4 = -3$, so condition (g) of Theorem 19.9.4 fails.

(c) True. Let $\mathbf{A}$ be a nonsingular M-matrix. By Theorem 19.9.4 there is $\mathbf{x} > \mathbf{0}$ with $\mathbf{A}\mathbf{x} > \mathbf{0}$. By Lemma 19.9.3, $\mathbf{A}_{S,S}$ is a Z-matrix with $\mathbf{A}_{S,S}\mathbf{x}_S > \mathbf{0}$ and $\mathbf{x}_S > \mathbf{0}$, so it is a nonsingular M-matrix by (c) $\Rightarrow$ (a) of the theorem.

(d) False. The swap $\mathbf{P} = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$ satisfies $\mathbf{P}^{-1} = \mathbf{P} \geq \mathbf{O}$, but $p_{12} = 1 > 0$, so $\mathbf{P}$ is not a Z-matrix, and every M-matrix is one.

(e) False. $\mathbf{I}_2$ is a nonsingular M-matrix, and $\mathbf{I}_2^{-1} = \mathbf{I}_2$ has zero off-diagonal entries.

(f) $x_i = i(n + 1 - i)$, for which $\mathbf{K}_n\mathbf{x} = 2\mathbf{1}$ (Proposition 19.9.6). For $\mathbf{x} = \mathbf{1}$, row i of $\mathbf{K}_n\mathbf{1}$ is $2 - 1 - 1 = 0$ for every $2 \leq i \leq n - 1$, and such rows exist when $n \geq 3$.

19.9.6.2 B. Practice

Exercise. B1: Which are nonsingular M-matrices?

Determine which of the following are nonsingular M-matrices. Justify your answer.

(a) $\begin{pmatrix} 3 & -1 & -1 \\ -1 & 3 & -1 \\ -1 & -1 & 3 \end{pmatrix}$

(b) $\begin{pmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{pmatrix}$

(c) $\begin{pmatrix} 2 & -3 \\ -1 & 2 \end{pmatrix}$

(d) $\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$

(e) $\begin{pmatrix} 1 & -2 & 0 \\ 0 & 1 & -2 \\ -2 & 0 & 1 \end{pmatrix}$

Solution.

(a) Yes. It is a Z-matrix, and every row sums to 1, so $\mathbf{x} = \mathbf{1} > \mathbf{0}$ gives $\mathbf{Ax} = \mathbf{1} > \mathbf{0}$. By Theorem 19.9.4, (c) $\Rightarrow$ (a), it is a nonsingular M-matrix. (Its inverse is $\frac{1}{4}\begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix} \geq \mathbf{O}$, as (b) predicts.)

(b) No. Every row sums to 0, so $\mathbf{A1} = \mathbf{0}$ and $\mathbf{A}$ is singular. (It is a singular M-matrix: with $s = 2$, $\mathbf{B} = 2\mathbf{I} - \mathbf{A}$ has every row sum equal to 2, so $\rho(\mathbf{B}) \leq 2$ by Corollary 16.4.2, while $\mathbf{B1} = 2\mathbf{1}$ gives $\rho(\mathbf{B}) \geq 2$.)

(c) Yes. It is a Z-matrix with leading principal minors $2 > 0$ and $4 - 3 = 1 > 0$, so (g) $\Rightarrow$ (a) applies. Check: $\mathbf{A}^{-1} = \left(\begin{smallmatrix} 2 & 3 \\ 1 & 2 \end{smallmatrix}\right) \geq \mathbf{O}$.

(d) No. The off-diagonal entries are $1 > 0$, so it is not a Z-matrix. (It is positive definite, and its inverse $\frac{1}{3}\begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}$ has negative entries.)

(e) No. It is a Z-matrix, but expanding along the first row,

$$\det \mathbf{A} = 1 \cdot (1 - 0) - (-2) \cdot (0 - 4) + 0 = 1 - 8 = -7 < 0,$$

so condition (f) of Theorem 19.9.4 fails.

Exercise. B2: A string with a free end

Let

$$\mathbf{A} = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{pmatrix}.$$

(a) Show that $\mathbf{A1}$ is not positive, and find $\mathbf{x} > \mathbf{0}$ with $\mathbf{Ax} > \mathbf{0}$.

(b) Compute the leading principal minors of $\mathbf{A}$, and check that they agree with (a).

(c) Compute $\mathbf{A}^{-1}$ and check that $\mathbf{A}^{-1} \geq \mathbf{O}$. Hence solve $\mathbf{Au} = \mathbf{e}_3$, and say why the answer had to be nonnegative.

Hint: try an increasing $\mathbf{x}$ *whose increments decrease.*

Solution.

(a) $\mathbf{A1} = (1, 0, 0)$, which is not positive. Take $\mathbf{x} = (3, 5, 6)$: then $\mathbf{Ax} = (6 - 5,\ -3 + 10 - 6,\ -5 + 6) = (1, 1, 1) > \mathbf{0}$. Since $\mathbf{A}$ is a Z-matrix, it is a nonsingular M-matrix by Theorem 19.9.4, (c) $\Rightarrow$ (a).

(b) $\det \mathbf{A}_1 = 2$, $\det \mathbf{A}_2 = 4 - 1 = 3$, and, expanding along the first row, $\det \mathbf{A} = 2(2 - 1) - (-1)(-1 - 0) = 2 - 1 = 1$. All three are positive, which is condition (g), as (a) and the theorem require.

(c)

$$\mathbf{A}^{-1} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 2 \\ 1 & 2 & 3 \end{pmatrix},$$

the matrix with entries $\min(i, j)$. Check: row 1 of $\mathbf{A}$ against the columns gives $2 - 1 = 1$, $2 - 2 = 0$, $2 - 2 = 0$; row 2 gives $-1 + 2 - 1 = 0$, $-1 + 4 - 2 = 1$, $-1 + 4 - 3 = 0$; row 3 gives $-1 + 1 = 0$, $-2 + 2 = 0$, $-2 + 3 = 1$. So $\mathbf{AA}^{-1} = \mathbf{I}$, and $\mathbf{A}^{-1} \geq \mathbf{O}$. Hence

$\mathbf{u} = \mathbf{A}^{-1}\mathbf{e}_3 = (1,2,3)$. It had to be $\geq \mathbf{0}$ because $\mathbf{e}_3 \geq \mathbf{0}$ and $\mathbf{A}^{-1} \geq \mathbf{O}$. Note that $\mathbf{A}^{-1}\mathbf{1} = (3,5,6)$ is the certificate of (a).

Exercise. B3: A two-sector economy

Let

$$\mathbf{C} = \begin{pmatrix} 1/5 & 2/5 \\ 3/10 & 1/10 \end{pmatrix}.$$

(a) Show that the economy with consumption matrix $\mathbf{C}$ is productive.

(b) Compute the Leontief inverse $(\mathbf{I}-\mathbf{C})^{-1}$.

(c) Find the production plan meeting the demand $\mathbf{d} = (30, 60)$, and check it.

(d) Find $\rho(\mathbf{C})$ exactly.

Solution.

(a) The column sums are $\frac{1}{5}+\frac{3}{10} = \frac{1}{2}$ and $\frac{2}{5}+\frac{1}{10} = \frac{1}{2}$, both less than 1, so the economy is productive by Corollary 19.9.7.

(b) $\mathbf{I}-\mathbf{C} = \left(\begin{smallmatrix} 4/5 & -2/5 \\ -3/10 & 9/10 \end{smallmatrix}\right)$, with determinant $\frac{36}{50} - \frac{6}{50} = \frac{3}{5}$. By Theorem 0.9.8,

$$(\mathbf{I}-\mathbf{C})^{-1} = \frac{5}{3}\begin{pmatrix} 9/10 & 2/5 \\ 3/10 & 4/5 \end{pmatrix} = \begin{pmatrix} 3/2 & 2/3 \\ 1/2 & 4/3 \end{pmatrix} \geq \mathbf{O}.$$

(c) $\mathbf{x} = (\frac{3}{2}\cdot 30 + \frac{2}{3}\cdot 60,\ \frac{1}{2}\cdot 30 + \frac{4}{3}\cdot 60) = (85, 95)$. Check: $\mathbf{C}\mathbf{x} = (17+38,\ \frac{51}{2}+\frac{19}{2}) = (55, 35)$, and $\mathbf{x} - \mathbf{C}\mathbf{x} = (30, 60)$.

(d) $p_{\mathbf{C}}(x) = x^2 - \frac{3}{10}x + (\frac{1}{50} - \frac{12}{100}) = x^2 - \frac{3}{10}x - \frac{1}{10} = (x-\frac{1}{2})(x+\frac{1}{5})$. So $\operatorname{spec}(\mathbf{C}) = \{\frac{1}{2}, -\frac{1}{5}\}$ and $\rho(\mathbf{C}) = \frac{1}{2}$. Here the column-sum bound $\|\mathbf{C}\|_1 = \frac{1}{2}$ is attained, as it must be when all column sums are equal.

19.9.6.3 C. Going deeper

Exercise. C1: Raising the diagonal

(a) Let $\mathbf{A}$ be a nonsingular M-matrix and $\mathbf{M}$ a Z-matrix of the same size with $\mathbf{M} \geq \mathbf{A}$ entrywise. Prove that $\mathbf{M}$ is a nonsingular M-matrix and that $\mathbf{O} \leq \mathbf{M}^{-1} \leq \mathbf{A}^{-1}$.

(b) Deduce that for every diagonal $\mathbf{D} \geq \mathbf{O}$, $\mathbf{K}_n + \mathbf{D}$ is a nonsingular M-matrix with $(\mathbf{K}_n + \mathbf{D})^{-1} \leq \mathbf{K}_n^{-1}$. Check this for $n = 2$ and $\mathbf{D} = \operatorname{diag}(1, 0)$.

Hint: for the inequality, factor $\mathbf{A}^{-1} - \mathbf{M}^{-1}$.

Solution.

(a) By Theorem 19.9.4 there is $\mathbf{x} > \mathbf{0}$ with $\mathbf{A}\mathbf{x} > \mathbf{0}$. Since $\mathbf{M} - \mathbf{A} \geq \mathbf{O}$ and $\mathbf{x} > \mathbf{0}$, $(\mathbf{M}-\mathbf{A})\mathbf{x} \geq \mathbf{0}$, so $\mathbf{M}\mathbf{x} = \mathbf{A}\mathbf{x} + (\mathbf{M}-\mathbf{A})\mathbf{x} > \mathbf{0}$. As $\mathbf{M}$ is a Z-matrix, (c) $\Rightarrow$ (a) of the theorem makes it a nonsingular M-matrix, and (a) $\Rightarrow$ (b) gives $\mathbf{M}^{-1} \geq \mathbf{O}$. Next,

$$\mathbf{A}^{-1} - \mathbf{M}^{-1} = \mathbf{A}^{-1}\mathbf{M}\mathbf{M}^{-1} - \mathbf{A}^{-1}\mathbf{A}\mathbf{M}^{-1} = \mathbf{A}^{-1}(\mathbf{M}-\mathbf{A})\mathbf{M}^{-1},$$

a product of three nonnegative matrices, hence $\geq \mathbf{O}$. This proves $\mathbf{O} \leq \mathbf{M}^{-1} \leq \mathbf{A}^{-1}$.

(b) $\mathbf{K}_n$ is a nonsingular M-matrix by Proposition 19.9.6, and $\mathbf{K}_n + \mathbf{D}$ is a Z-matrix with $\mathbf{K}_n + \mathbf{D} \geq \mathbf{K}_n$, since $\mathbf{D}$ only changes the diagonal and does so upwards. Part (a) gives the claim. For $n = 2$: $\mathbf{K}_2^{-1} = \frac{1}{3}\left(\begin{smallmatrix} 2 & 1 \\ 1 & 2 \end{smallmatrix}\right)$, and $\mathbf{K}_2 + \mathbf{D} = \left(\begin{smallmatrix} 3 & -1 \\ -1 & 2 \end{smallmatrix}\right)$ has determinant 5 and inverse $\frac{1}{5}\left(\begin{smallmatrix} 2 & 1 \\ 1 & 3 \end{smallmatrix}\right)$. The difference is

$$\frac{1}{3}\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} - \frac{1}{5}\begin{pmatrix} 2 & 1 \\ 1 & 3 \end{pmatrix} = \frac{1}{15}\begin{pmatrix} 4 & 2 \\ 2 & 1 \end{pmatrix} \geq \mathbf{O}.$$

Exercise. C2: The singular case

(a) Show that the Z-matrix $\mathbf{Z} = \left(\begin{smallmatrix} 0 & -1 \\ 0 & -1 \end{smallmatrix}\right)$ has nonnegative leading principal minors but is not an M-matrix.

(b) Let $\mathbf{A}$ be an M-matrix. Prove that $\mathbf{A} + \varepsilon\mathbf{I}$ is a nonsingular M-matrix for every real $\varepsilon > 0$.

(c) Deduce that every principal minor of an M-matrix is nonnegative.

(d) Check (c) for $\left(\begin{smallmatrix} 1 & -1 \\ -1 & 1 \end{smallmatrix}\right)$, and explain why (a) does not contradict the stated converse in the remark after Theorem 19.9.4.

Solution.

(a) $\det \mathbf{Z}_1 = 0$ and $\det \mathbf{Z} = 0 \cdot (-1) - (-1) \cdot 0 = 0$, both nonnegative. $\mathbf{Z}$ is upper triangular, so its eigenvalues are its diagonal entries 0 and -1 (Theorem 9.7.4). Its smallest real eigenvalue is $-1 < 0$, so by Lemma 19.9.2 it is not an M-matrix: for any representation $\mathbf{Z} = s\mathbf{I} - \mathbf{B}$, $s - \rho(\mathbf{B}) = -1$.

(b) Write $\mathbf{A} = s\mathbf{I} - \mathbf{B}$ with $\mathbf{B} \geq \mathbf{O}$ and $s \geq \rho(\mathbf{B})$. Then $\mathbf{A} + \varepsilon\mathbf{I} = (s + \varepsilon)\mathbf{I} - \mathbf{B}$ with $s + \varepsilon > \rho(\mathbf{B})$, so it is a nonsingular M-matrix by Lemma 19.9.2 (b).

(c) Fix a nonempty $S \subseteq [n]$ with k elements. Then $(\mathbf{A} + \varepsilon\mathbf{I})_{S,S} = \mathbf{A}_{S,S} + \varepsilon\mathbf{I}_k$. By (b) and condition (f) of Theorem 19.9.4, $q(\varepsilon) = \det(\mathbf{A}_{S,S} + \varepsilon\mathbf{I}_k) > 0$ for every $\varepsilon > 0$. By the Leibniz formula, q is a polynomial in ε, hence continuous by the algebra of limits quoted in Chapter 16's introduction, so $q(1/m) \to q(0)$ as $m \to \infty$. A nonstrict inequality survives a limit, so $\det \mathbf{A}_{S,S} = q(0) \geq 0$.

(d) It is an M-matrix (see the Quick check after Lemma 19.9.2). Its principal minors are 1, 1 and $1 - 1 = 0$, all nonnegative. The converse in the remark asks for **all** principal minors to be nonnegative, and $\mathbf{Z}$ of (a) has the principal minor $\det \mathbf{Z}_{\{2\},\{2\}} = -1 < 0$. Only its **leading** ones are nonnegative, and for the nonstrict inequality leading minors are not enough.

Exercise. C3: Nonnegative dynamics

Let $\mathbf{A} \in M_n(\mathbb{R})$ be a Z-matrix, written as $\mathbf{A} = s\mathbf{I} - \mathbf{B}$ with $\mathbf{B} \geq \mathbf{O}$.

(a) Prove that $e^{-t\mathbf{A}} \geq \mathbf{O}$ for every real $t \geq 0$.

(b) Suppose $\mathbf{A}$ is a nonsingular M-matrix, and let $\mathbf{x}$ be the solution of $\mathbf{x}' = -\mathbf{A}\mathbf{x}$ with $\mathbf{x}(0) = \mathbf{x}_0 \geq \mathbf{0}$. Prove that $\mathbf{x}(t) \geq \mathbf{0}$ for every $t \geq 0$ and that $\mathbf{x}(t) \to \mathbf{0}$ as $t \to \infty$.

Hint: split $-t\mathbf{A}$ into two commuting matrices.

Solution.

(a) $-t\mathbf{A} = t\mathbf{B} + (-ts)\mathbf{I}$, and the two terms commute because every matrix commutes with a multiple of $\mathbf{I}$. By Theorem 10.9.4 (d), $e^{-t\mathbf{A}} = e^{t\mathbf{B}}e^{-ts\mathbf{I}}$. Every partial sum of the series for $e^{-ts\mathbf{I}}$ is $(\sum_{m=0}^{M}(-ts)^m/m!)\mathbf{I}$, so $e^{-ts\mathbf{I}} = e^{-ts}\mathbf{I}$. The series for $e^{t\mathbf{B}}$ converges by Theorem 10.9.3, and its partial sums $\sum_{m=0}^{M} t^m\mathbf{B}^m/m!$ are $\geq \mathbf{O}$ because $t \geq 0$ and $\mathbf{B} \geq \mathbf{O}$; a nonstrict inequality survives a limit, so $e^{t\mathbf{B}} \geq \mathbf{O}$. Hence $e^{-t\mathbf{A}} = e^{-ts}e^{t\mathbf{B}} \geq \mathbf{O}$.

(b) By Theorem 10.9.6, $\mathbf{x}(t) = e^{-t\mathbf{A}}\mathbf{x}_0$, which is $\geq \mathbf{0}$ for $t \geq 0$ by (a). By Theorem 19.9.4 (d), every eigenvalue of $\mathbf{A}$ has positive real part, so every eigenvalue of $-\mathbf{A}$ has negative real part: $-\mathbf{A}$ is stable. By Theorem 16.6.3, $\mathbf{x}(t) \to \mathbf{0}$. Both conclusions matter in models of concentrations or populations, which must stay nonnegative and, when $-\mathbf{A}$ is stable, die out.

Chapter 20

Perturbation Theory

Every matrix that comes from a measurement or a computation is slightly wrong, so the eigenvalues one actually knows are those of a nearby matrix $\mathbf{A} + \mathbf{E}$. Two earlier chapters began to say what that costs. Chapter 16 showed that the eigenvalues move continuously with the matrix (Corollary 16.7.4), and that "continuously" can be very bad: a perturbation of size ε moves the eigenvalue of a $k \times k$ Jordan block by $\varepsilon^{1/k}$ (Example (A Jordan block's eigenvalues move like $\varepsilon^{1/k}$)). Chapter 17 showed that Hermitian matrices are as good as one could hope: each eigenvalue moves by at most $\|\mathbf{E}\|_2$ (Corollary 17.3.6). Most matrices are neither Jordan blocks nor Hermitian, and this chapter handles them. It asks four questions.

- **Where are the eigenvalues,** before any of them is computed? Discs and ovals read off the entries, and the numerical range, locate them.
- **How far do they move** when $\mathbf{A}$ is not Hermitian? For every matrix there is a bound with an n-th root, and the Jordan block shows that the root cannot be removed. For a diagonalizable matrix, the Bauer–Fike theorem gives a bound proportional to $\|\mathbf{E}\|$, with a factor that measures how far the eigenvectors are from orthonormal.
- **Which eigenvalue goes where?** A bound saying that each new eigenvalue is near **some** old one does not say that they can be paired off. For normal matrices they can, and the pairing comes from Birkhoff's theorem of Chapter 19.
- **What happens to the rest?** Eigenvectors and invariant subspaces, singular values, and the factors of the polar decomposition all move, and the chapter bounds each of them. It ends with the first-order sensitivity of a single simple eigenvalue.

One result settles an old account. Chapter 16 quoted the theorem that the roots of a polynomial depend continuously on its coefficients (Theorem 16.7.3) and left the proof to this chapter. §03 gives it from compactness and unique factorization alone, so Corollary 16.7.4, and everything Chapter 16 built on it, now rests on a proved theorem.

The field is $\mathbb{C}$ throughout, unless a statement says $\mathbb{R}$. A real matrix is read in $M_n(\mathbb{C})$ when its eigenvalues are wanted. Three conventions hold in every section.

- **The perturbation is E,** and the perturbed matrix is $\mathbf{A} + \mathbf{E}$, as in Chapters 16 and 17. A different letter, $\widetilde{\mathbf{A}}$, is used only where the perturbed matrix is not naturally written as a sum.
- **Every bound names its norm.** $\|\cdot\|_2$ is the spectral norm and $\|\cdot\|_F$ the Frobenius norm. A norm without a subscript means an arbitrary operator norm, and it is written that way only where the statement holds for all of them. There is one other use, announced where it starts: in §§09–10 a norm without a subscript stands for either the spectral or the Frobenius norm, the same one throughout the statement, and there it ranges over just

those two.

- **Near some eigenvalue is not the same as matched.** "Every eigenvalue of $\mathbf{A}+\mathbf{E}$ lies within r of some eigenvalue of $\mathbf{A}$" is a one-sided statement. "The eigenvalues can be **numbered** so that each pair is within r" is a matching, and it is stronger. §§03–05 are largely about the difference, and the chapter never blurs it.

20.0.1 The analysis this chapter imports

The list is Chapter 16's, facts (A1)–(A6) of that chapter's introduction, together with the algebra of limits, and each fact is named again where it is used. Two carry most of the weight. **(A3)**, compactness of closed bounded sets, is the engine of §03's proof of the continuity of roots, and it is also what makes the numerical range of §02 compact. **(A5)**, the intermediate value theorem, gives the convexity of the numerical range in §02. **(A4)**, the extreme value theorem, is used once, also in §02, to know that the numerical radius — the largest modulus on the numerical range — is attained. §03's counting theorem needs one more fact, that an integer-valued function on an interval that is locally constant is constant, and it derives this from (A5) as well. Nothing else analytic is used. In particular, the first-order formula of §11 for the derivative of a simple eigenvalue is proved without the implicit function theorem.

20.0.2 What you need

- **Chapter 0 §06 and §09** — the triangle inequality for complex numbers, which is the first move of nearly every estimate in the chapter (§§01, 02, 03, 09, and 11), and the trace, which §08 computes with: Theorem 0.6.4, Theorem 0.9.6.
- **Chapter 1 §06** — basis extension, once, in §11, to turn an eigenvector into the first column of an invertible matrix: Theorem 1.6.5.
- **Chapter 3 §05 and §07** — Rank–Nullity for matrices and the factor theorem, both in §11: Theorem 3.5.7, Lemma 3.7.1.
- **Chapter 4 §05** — that a projection's image is the subspace it projects onto, which is what lets §08 compare two subspaces by comparing their projections: Theorem 4.5.1.
- **Chapter 6** — polynomials over $\mathbb{C}$: Corollary 6.5.5, Theorem 6.3.5, Theorem 6.4.5.
- **Chapter 7** — determinants, the adjugate, and minors: Definition 7.6.2, Theorem 7.6.2, Theorem 7.7.1, Theorem 7.10.2, Theorem 7.10.4, Theorem 7.8.4.
- **Chapter 8** — block matrices and the Kronecker and vec identities: Theorem 8.1.1, Theorem 8.4.4.
- **Chapter 9** — eigenvalues, multiplicities, diagonalization, and left eigenvectors: Theorem 9.2.1, Definition 9.3.1, Theorem 9.3.1, Definition 9.4.1, Theorem 9.4.1, Definition 9.10.1, Proposition 9.10.5, Proposition 9.10.7.
- **Chapter 10** — the Jordan form, and its warning that the form is unstable: Theorem 10.3.1, Definition 10.3.1. Here the Jordan form is the thing being measured rather than a tool: §§04 and 11 discuss it in the text and pay off Chapter 10's promise, and no proof in the chapter cites it.
- **Chapter 11** — inner products, orthogonal projections, unitary maps, and the unitary invariance of the Frobenius norm: Theorem 11.1.5, Definition 11.3.2, Theorem 11.3.2, Corollary 11.2.3, Theorem 11.7.3, Lemma 11.7.5.
- **Chapter 12** — Schur's theorem, the spectral theorem, the Sylvester equation, and the departure from normality: Theorem 12.1.2, Corollary 12.4.3, Theorem 12.3.3, Definition 12.10.1, Theorem 12.10.2, Corollary 12.10.3, Definition 12.11.1, Example (A perturbation that moves the eigenvalues a thousand times as far).
- **Chapter 13** — the singular value and polar decompositions, and Hadamard's inequality: Theorem 13.9.1, Definition 13.9.1, Theorem 13.10.1, Theorem 13.7.3.

- **Chapter 16** — operator norms, the Neumann series, continuity of eigenvalues, and condition numbers: Definition 16.3.2, Theorem 16.3.2, Theorem 16.3.3, Theorem 16.5.1, Lemma 16.7.1, Theorem 16.7.3, Corollary 16.7.4, Example (A Jordan block's eigenvalues move like $\varepsilon^{1/k}$), Definition 16.8.1.
- **Chapter 17** — the Rayleigh quotient, Weyl's inequalities, Lidskii's inequality, majorization, the Hermitian dilation, and §10's law of inertia for Hermitian matrices: Definition 17.1.1, Theorem 17.2.3, Theorem 17.3.3, Corollary 17.3.6, Theorem 17.7.2, Definition 17.8.1, Proposition 17.9.3, Corollary 17.9.5, Definition 17.10.1, Theorem 17.10.1.
- **Chapter 18** — extreme points, and linear functions on compact convex sets: Definition 18.8.1, Corollary 18.8.6.
- **Chapter 19** — irreducible matrices, and Birkhoff's theorem: Definition 19.2.3, Theorem 19.2.5, Theorem 19.7.2.

20.0.3 Roadmap

- **Gershgorin discs.** Every eigenvalue lies in a union of discs read off the rows, a diagonal rescaling moves the discs, and an irreducible matrix can have an eigenvalue on the boundary only if it lies on every circle.
- **Other inclusion regions.** Brauer's ovals, which are never worse than the discs, and the numerical range: it contains the spectrum, it is convex by the Toeplitz–Hausdorff theorem, and it equals the convex hull of the eigenvalues for a normal matrix.
- **Continuity of roots.** Chapter 16's quoted theorem, proved by compactness. Elsner's bound, with its exponent $1/n$ shown to be sharp by a Jordan block, and its matching version. The counting theorem: a group of Gershgorin discs separated from the others contains exactly as many eigenvalues as discs.
- **The Bauer–Fike theorem.** For a diagonalizable matrix, each perturbed eigenvalue lies within $\kappa_2(\mathbf{X})\|\mathbf{E}\|_2$ of an old one, with factor 1 for a normal matrix. It comes with a residual bound for computed eigenpairs and a version with an m-th root for defective matrices.
- **The Hoffman–Wielandt theorem.** The eigenvalues of two normal matrices can be matched within their Frobenius distance, by Birkhoff's theorem. The inequality fails for nonnormal matrices.
- **Hermitian perturbations.** Residual bounds, the Kato–Temple bound, which is quadratic in the residual, and Ostrowski's theorem, which says that a congruence changes each eigenvalue by a bounded relative factor.
- **The Sylvester equation, quantitatively.** The separation of two matrices, which measures how badly conditioned $\mathbf{AX} - \mathbf{XB} = \mathbf{Y}$ is. For normal matrices it is the distance between the spectra, and in general it can be much smaller. Bounds for separated Hermitian spectra follow.
- **Principal angles.** The angles between two subspaces, their sines, and the distance $\|P_U - P_W\|_2$ between the orthogonal projections.
- **The Davis–Kahan theorem.** An invariant subspace of a Hermitian matrix moves by at most the perturbation divided by the spectral gap. Single eigenvectors inside a cluster may not be stable, but the subspace of the whole cluster is.
- **Singular values and polar factors.** Mirsky's Frobenius bound for singular values, and bounds for the positive and the unitary factors of the polar decomposition of an invertible matrix.
- **Condition numbers of eigenvalues.** The derivative of a simple eigenvalue, from a right and a left eigenvector, and its worst size, which is the condition number. The chapter's summary closes this section.

20.0.4 Named moves

- **Dominant coordinate plus the triangle inequality.** Read the eigen-equation at the coordinate where the eigenvector is largest.
- **A determinant bounded two ways.** A product of distances to the eigenvalues is a determinant, and Hadamard's inequality bounds it by column lengths.
- **Compactness plus uniqueness.** Bound the roots, extract a convergent subsequence, and let unique factorization identify the limit.
- **An integer-valued continuous function is constant.** Count eigenvalues along a path on which none of them can escape the region.
- **Factor out the resolvent.** Write $\mathbf{A}+\mathbf{E}-\mu\mathbf{I} = (\mathbf{A}-\mu\mathbf{I})(\mathbf{I}+(\mathbf{A}-\mu\mathbf{I})^{-1}\mathbf{E})$ and invert the second factor by the Neumann series.
- **Birkhoff turns a doubly stochastic weight into a permutation.** Minimize a linear function over the doubly stochastic matrices, and the minimum sits at a permutation.
- **The perturbation of a subspace is a Sylvester equation.** The block coupling an invariant subspace to the new complement solves one, so the spectral gap is the conditioning.
- **The singular-pair trick.** Test a matrix equation against the top singular pair of the unknown.
- **The adjugate as a polynomial eigenvector.** A column of $\operatorname{adj}(\mu\mathbf{I}-\mathbf{B})$ is an eigenvector whose entries are polynomials, so it moves continuously with the data.

20.1 Gershgorin Discs

Chapter 16 bounded the whole spectrum with one number: every eigenvalue lies in the disc $|z| \le \|\mathbf{A}\|_\infty$ (Corollary 16.4.2). That disc is centered at 0, and it knows nothing about where in it the eigenvalues sit. For a matrix that is nearly diagonal the answer ought to be visible at a glance: the eigenvalues should sit near the diagonal entries, and the off-diagonal entries should say how near. This section makes that precise. It puts one disc around each diagonal entry, with radius measured from the rest of its row, and proves that the discs together hold every eigenvalue. It then sharpens the picture twice: by rescaling the coordinates, and, for irreducible matrices, by looking at the boundary.

Throughout, matrices are complex. A real matrix is read inside $M_n(\mathbb{C})$, as in Chapter 16 §04, because its eigenvalues may be nonreal, and they are what we want to locate.

20.1.1 Discs around the diagonal

Here is the situation to keep in mind. Take

$$\mathbf{A} = \begin{pmatrix} 6 & 2 & 5 \\ 1 & -1 & 1 \\ 1 & 1 & -1 \end{pmatrix}.$$

The row-sum bound gives $|\lambda| \le \|\mathbf{A}\|_\infty = 13$ for every eigenvalue λ, a disc of radius 13 about 0. But if the off-diagonal entries were 0, the eigenvalues would be exactly $6, -1, -1$. The off-diagonal entries are a perturbation of the diagonal matrix, and the natural guess is that each eigenvalue stays within reach of some diagonal entry, where the reach of row i is measured by the size of the entries of row i off the diagonal. We give that reach a name.

Around each diagonal entry, draw the disc whose radius is the total size of the other entries in its row.

Definition 20.1.1. Gershgorin Discs

Let $n \geq 1$ and $\mathbf{A} = (a_{ij}) \in M_n(\mathbb{C})$. For $i = 1, \ldots, n$, the ***i*-th Gershgorin radius** of $\mathbf{A}$ is the real number

$$r_i(\mathbf{A}) \coloneqq \sum_{j \neq i} |a_{ij}| \ \geq 0,$$

the sum over **all** $j \in \{1, \ldots, n\}$ **other than** i, and the ***i*-th Gershgorin disc** is the **closed** disc

$$D_i(\mathbf{A}) \coloneqq \{\, z \in \mathbb{C} : |z - a_{ii}| \leq r_i(\mathbf{A}) \,\}.$$

The discs $D_i(\mathbf{A}^\top)$, whose radii $r_i(\mathbf{A}^\top) = \sum_{j \neq i} |a_{ji}|$ are taken from the **columns** of $\mathbf{A}$, are the **column discs**.

In words: the center of $D_i(\mathbf{A})$ is the diagonal entry a_{ii}, and the radius adds up the moduli of the entries of row i that are **not** on the diagonal. The disc includes its boundary circle. A radius may be 0, and then the disc is the single point a_{ii}. Since $\mathbf{A}^\top$ has the same diagonal as $\mathbf{A}$, the column discs have the same centers as the row discs and, in general, different radii. The letter r is lowercase on purpose, since $R_{\mathbf{A}}$ is the Rayleigh quotient, and the plain D_i is a set of complex numbers, not the bold diagonal matrix $\mathbf{D}$.

Some examples, simplest first.

- **Diagonal matrices and $n = 1$.** For $\mathbf{A} = \operatorname{diag}(d_1, \ldots, d_n)$ every radius is 0, because every off-diagonal entry is 0, and $D_i(\mathbf{A}) = \{d_i\}$. The discs are the eigenvalues. A 1×1 matrix (a) is the case $n = 1$: the sum defining r_1 is empty, so $D_1 = \{a\}$. These degenerate cases are the ones any theorem about the discs must get exactly right.
- **$\mathbf{A}$ 2×2 matrix.** For $\mathbf{B} = \left(\begin{smallmatrix} 0 & 1 \\ 4 & 0 \end{smallmatrix}\right)$, $r_1(\mathbf{B}) = 1$ and $r_2(\mathbf{B}) = 4$, so the row discs are $|z| \leq 1$ and $|z| \leq 4$. The column discs swap the radii: $|z| \leq 4$ and $|z| \leq 1$.
- **The matrix above.** $r_1(\mathbf{A}) = 2 + 5 = 7$, $r_2(\mathbf{A}) = 1 + 1 = 2$, and $r_3(\mathbf{A}) = 1 + 1 = 2$, so $D_1(\mathbf{A}) = \{|z - 6| \leq 7\}$ and $D_2(\mathbf{A}) = D_3(\mathbf{A}) = \{|z + 1| \leq 2\}$. The column radii are $1 + 1 = 2$, $2 + 1 = 3$, and $5 + 1 = 6$.

Nonexample by minimal change. Keep the radius and move the center to 0: replace $D_i(\mathbf{A})$ by $\{|z| \leq r_i(\mathbf{A})\}$. For $\operatorname{diag}(1, 2)$ both radii are 0 and the modified "discs" are both $\{0\}$, which misses both eigenvalues. The clause that changed is the center, and the diagonal example shows it is the part that carries the information: the radius only measures how far the off-diagonal part can pull an eigenvalue away from it.

20.1.2 The theorem

It turns out that the guess is right, with no loss at all in the constant:

Theorem 20.1.1. Gershgorin's Theorem

Let $n \geq 1$, $\mathbf{A} \in M_n(\mathbb{C})$, and let λ be an eigenvalue of $\mathbf{A}$ with eigenvector $\mathbf{x} = (x_1, \ldots, x_n)$. If k is an index with $|x_k| = \max_j |x_j|$, then $\lambda \in D_k(\mathbf{A})$. Consequently:

(a) every eigenvalue of $\mathbf{A}$ lies in the union $D_1(\mathbf{A}) \cup \cdots \cup D_n(\mathbf{A})$ of the row discs;

(b) every eigenvalue of $\mathbf{A}$ lies in the union $D_1(\mathbf{A}^\top) \cup \cdots \cup D_n(\mathbf{A}^\top)$ of the column discs;

(c) hence every eigenvalue lies in the intersection of the two unions.

Idea. We want $|\lambda - a_{kk}|$ to be small for one well-chosen k, so look at row k of $\mathbf{Ax} = \lambda\mathbf{x}$ and move the diagonal term across: $(\lambda - a_{kk})x_k = \sum_{j\neq k} a_{kj}x_j$. The triangle inequality bounds the right side by $\sum_{j\neq k}|a_{kj}||x_j|$, and this becomes $r_k(\mathbf{A})|x_k|$ exactly when every $|x_j| \le |x_k|$. So choose k where the eigenvector is largest in modulus. Then divide by $|x_k|$, which is positive because $\mathbf{x} \neq \mathbf{0}$.

Proof. Let $\mathbf{Ax} = \lambda\mathbf{x}$ with $\mathbf{x} \neq \mathbf{0}$, and let k be as in the statement; such a k exists because a finite set of real numbers has a largest element. Since $\mathbf{x} \neq \mathbf{0}$, some $x_j \neq 0$, so $|x_k| > 0$. Row k of $\mathbf{Ax} = \lambda\mathbf{x}$ reads $\sum_j a_{kj}x_j = \lambda x_k$, that is,

$$(\lambda - a_{kk})\, x_k = \sum_{j\neq k} a_{kj}x_j.$$

Taking moduli and using the triangle inequality (Theorem 0.6.4), then $|x_j| \le |x_k|$ for every j,

$$|\lambda - a_{kk}|\,|x_k| \le \sum_{j\neq k}|a_{kj}|\,|x_j| \le \Big(\sum_{j\neq k}|a_{kj}|\Big)|x_k| = r_k(\mathbf{A})\,|x_k|.$$

Dividing by $|x_k| > 0$ gives $|\lambda - a_{kk}| \le r_k(\mathbf{A})$, that is, $\lambda \in D_k(\mathbf{A})$.

(a) Every eigenvalue has an eigenvector, and the disc just found is one of the $D_i(\mathbf{A})$.

(b) By Proposition 9.10.5 (b), $\mathbf{A}^\top$ has the same eigenvalues as $\mathbf{A}$. Applying (a) to $\mathbf{A}^\top$ puts each of them in $D_1(\mathbf{A}^\top) \cup \cdots \cup D_n(\mathbf{A}^\top)$.

(c) A point lying in each of two sets lies in their intersection. This proves the theorem.

■

The move in this proof is worth naming, since the chapter uses it again: **choose the dominant coordinate of an eigenvector, then apply the triangle inequality to its row.** The choice of the largest coordinate is exactly what lets every $|x_j|$ be replaced by $|x_k|$ without looking at the others.

The theorem recovers Chapter 16's bound and improves on it. If $\lambda \in D_k(\mathbf{A})$, then

$$\begin{aligned}|\lambda| &\le |a_{kk}| + |\lambda - a_{kk}| \le |a_{kk}| + r_k(\mathbf{A}) \\ &= \sum_j |a_{kj}| \le \|\mathbf{A}\|_\infty,\end{aligned}$$

so (a) implies $\rho(\mathbf{A}) \le \|\mathbf{A}\|_\infty$, and (b) likewise implies $\rho(\mathbf{A}) \le \|\mathbf{A}\|_1$. The union of the discs is never larger than the row-sum disc, and it is usually much smaller, because it keeps the centers.

Example. Discs, eigenvectors, and where each eigenvalue sits

Let $\mathbf{A} = \begin{pmatrix} 6 & 2 & 5 \\ 1 & -1 & 1 \\ 1 & 1 & -1 \end{pmatrix}$. Describe the regions given by Theorem 20.1.1 (a) and (b). Then check that $(7,1,1)$, $(-1,1,1)$, and $(1,1,-2)$ are eigenvectors, and locate each eigenvalue in the disc the theorem assigns to it.

Solution. From the radii computed after Definition 20.1.1, the row discs give

$$\{|z-6|\le 7\}\cup\{|z+1|\le 2\},$$

the second disc counted twice, and the column discs give $\{|z-6|\le 2\}\cup\{|z+1|\le 3\}\cup\{|z+1|\le 6\}$, where the middle disc lies inside the last. So every eigenvalue lies in $\{|z-6|\le 2\}\cup\{|z+1|\le 6\}$, and also in $\{|z-6|\le 7\}\cup\{|z+1|\le 2\}$. The column discs have separated the two groups of centers; the row discs have not.

Multiplying out,

$$\mathbf{A}\begin{pmatrix}7\\1\\1\end{pmatrix}=\begin{pmatrix}49\\7\\7\end{pmatrix},\quad \mathbf{A}\begin{pmatrix}-1\\1\\1\end{pmatrix}=\begin{pmatrix}1\\-1\\-1\end{pmatrix},\quad \mathbf{A}\begin{pmatrix}1\\1\\-2\end{pmatrix}=\begin{pmatrix}-2\\-2\\4\end{pmatrix},$$

so the eigenvalues include 7, -1, and -2. These are three distinct eigenvalues of a 3×3 matrix, hence three distinct roots of the degree-3 polynomial $p_{\mathbf{A}}$ (Theorem 9.2.1), so by Corollary 6.1.4 there are no others.

- The eigenvector $(7,1,1)$ peaks at $k=1$, and indeed $|7-6|=1\le 7$: $7\in D_1(\mathbf{A})$.
- The eigenvector $(1,1,-2)$ peaks at $k=3$, and $|-2+1|=1\le 2$: $-2\in D_3(\mathbf{A})$.
- The eigenvector $(-1,1,1)$ has all three coordinates of modulus 1, so every k is a peak and -1 must lie in every row disc. It does: $|-1-6|=7=r_1(\mathbf{A})$, so it sits on the boundary circle of $D_1(\mathbf{A})$, and it is the center of the other two.

Both unions pass the test, and the eigenvector tells us which disc to look in. The row-sum disc $|z|\le 13$ says only that $|\lambda|\le 13$.

Warning. **The union is the claim, not each disc.** A single Gershgorin disc need not contain any eigenvalue. For $\mathbf{B}=\left(\begin{smallmatrix}0&1\\4&0\end{smallmatrix}\right)$, $p_{\mathbf{B}}(x)=x^2-4$, so the eigenvalues are ± 2, while the first row disc is $|z|\le 1$: it contains neither. Both eigenvalues lie in the second disc $|z|\le 4$. Nor may the row disc and the column disc of the same index be intersected: $D_1(\mathbf{B})\cap D_1(\mathbf{B}^{\top})$ and $D_2(\mathbf{B})\cap D_2(\mathbf{B}^{\top})$ are both $|z|\le 1$, and ± 2 lies in neither. Part (c) of the theorem intersects the two **unions**.

Quick check. Without computing any eigenvalue, show that every eigenvalue of $\mathbf{M}=\begin{pmatrix}5&2&-1\\1&6&2\\-1&0&3\end{pmatrix}$ has real part at least 2.

Solution. The row discs are $|z-5|\le 3$, $|z-6|\le 3$, and $|z-3|\le 1$. A point z of the first has $\operatorname{Re} z\ge 5-|z-5|\ge 2$, of the second $\operatorname{Re} z\ge 3$, of the third $\operatorname{Re} z\ge 2$. By Theorem 20.1.1 (a) every eigenvalue lies in one of them, so its real part is at least 2. In particular 0 is not an eigenvalue, and $\mathbf{M}$ is invertible.

20.1.3 Diagonal dominance

The last sentence of the Quick check is a pattern. If 0 lies outside every disc, then 0 is not an eigenvalue, and the matrix is invertible. The point 0 lies outside $D_i(\mathbf{A})$ exactly when $|0 - a_{ii}| > r_i(\mathbf{A})$, which is a condition on row i alone.

Corollary 20.1.2. Strict Diagonal Dominance Implies Invertibility

Let $n \geq 1$ and $\mathbf{A} \in M_n(\mathbb{C})$. Call **A strictly diagonally dominant** (by rows) if

$$|a_{ii}| > r_i(\mathbf{A}) = \sum_{j \neq i} |a_{ij}| \qquad \text{for **every**} \ i = 1, \ldots, n.$$

A strictly diagonally dominant matrix is invertible. The same holds if $\mathbf{A}$ is strictly diagonally dominant **by columns**, that is, if $\mathbf{A}^\top$ is strictly diagonally dominant.

Proof. Suppose $\mathbf{A}$ is strictly diagonally dominant. For every i, $|0 - a_{ii}| > r_i(\mathbf{A})$, so $0 \notin D_i(\mathbf{A})$. Hence 0 lies in no row disc, and by Theorem 20.1.1 (a) it is not an eigenvalue. By Theorem 9.2.1, $0\mathbf{I} - \mathbf{A} = -\mathbf{A}$ is invertible, and so is $\mathbf{A}$. If instead $\mathbf{A}^\top$ is strictly diagonally dominant, then 0 lies in no column disc, and Theorem 20.1.1 (b) finishes in the same way. ∎

This is the **Lévy–Desplanques theorem**. The word **strictly** carries the weight, and the definition asks it of **every** row. Examples: $\left(\begin{smallmatrix} 3 & 1 & 1 \\ 0 & 2 & 1 \\ 1 & -1 & 4 \end{smallmatrix}\right)$ is strictly dominant, since $3 > 2$, $2 > 1$, $4 > 2$; every diagonal matrix with nonzero diagonal is, since every radius is 0; and a 1×1 matrix (a) is strictly dominant exactly when $a \neq 0$. A minimal change breaks it. $\left(\begin{smallmatrix} 1 & 1 \\ 1 & 1 \end{smallmatrix}\right)$ has $|a_{ii}| = 1 = r_i$ in both rows: it is dominant only weakly, and it is singular. The converse of the corollary is false: $\left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$ is invertible and has zero diagonal. Diagonal dominance is a certificate of invertibility, cheap to check and often available, not a characterization.

20.1.4 Hermitian matrices, and intervals

For a Hermitian $\mathbf{A}$ the eigenvalues are real (Theorem 12.2.1), and so are the diagonal entries, since $a_{ii} = \overline{a_{ii}}$. So only the part of each disc on the real line matters, which is the interval $[a_{ii} - r_i(\mathbf{A}),\ a_{ii} + r_i(\mathbf{A})]$. Theorem 20.1.1 (a) then says

$$\min_i \big(a_{ii} - r_i(\mathbf{A})\big) \leq \lambda_n(\mathbf{A}) \leq \lambda_1(\mathbf{A}) \leq \max_i \big(a_{ii} + r_i(\mathbf{A})\big).$$

For $\mathbf{A} = \left(\begin{smallmatrix} 4 & 1 & 0 \\ 1 & 6 & 2 \\ 0 & 2 & 9 \end{smallmatrix}\right)$ the intervals are $[3, 5]$, $[3, 9]$, and $[7, 11]$, so every eigenvalue lies in $[3, 11]$, and $\mathbf{A}$ is positive definite by Theorem 13.1.5, since its eigenvalues are positive; Exercise C1 turns this into a general criterion. For a Hermitian matrix the row and column discs coincide, since $|a_{ji}| = |\overline{a_{ij}}|$, so (b) adds nothing.

20.1.5 Scaling the discs

The discs depend on the matrix, and the eigenvalues depend only on its similarity class. So any similarity that is easy to compute gives new discs for the same eigenvalues. The cheapest one rescales the coordinates.

Proposition 20.1.3. Rescaling the Gershgorin Discs

Let $\mathbf{A} \in M_n(\mathbb{C})$, let $d_1, \dots, d_n$ be **positive** real numbers, and $\mathbf{D} = \operatorname{diag}(d_1, \dots, d_n)$. Then:

(a) $\mathbf{D}^{-1}\mathbf{A}\mathbf{D}$ has entries $a_{ij}d_j/d_i$. It has the same diagonal as $\mathbf{A}$, and the same eigenvalues with the same algebraic multiplicities.

(b) $r_i(\mathbf{D}^{-1}\mathbf{A}\mathbf{D}) = \dfrac{1}{d_i}\sum_{j\neq i}|a_{ij}|\,d_j$ for each i.

(c) Every eigenvalue of $\mathbf{A}$ lies in

$$\bigcup_{i=1}^{n}\Big\{\, z : |z - a_{ii}| \le \frac{1}{d_i}\sum_{j\neq i}|a_{ij}|\,d_j \,\Big\}.$$

Proof.

(a) $\mathbf{D}$ is invertible with $\mathbf{D}^{-1} = \operatorname{diag}(1/d_1, \dots, 1/d_n)$, because every $d_i \neq 0$. Multiplying by a diagonal matrix on the left scales rows and on the right scales columns, so $(\mathbf{D}^{-1}\mathbf{A}\mathbf{D})_{ij} = d_i^{-1}a_{ij}d_j$, and for $i = j$ this is a_{ii}. The matrices $\mathbf{A}$ and $\mathbf{D}^{-1}\mathbf{A}\mathbf{D}$ are similar, so they have the same characteristic polynomial (Theorem 7.10.3), hence the same eigenvalues with the same multiplicities.

(b) By (a) and $d_i, d_j > 0$, $|a_{ij}d_j/d_i| = |a_{ij}|d_j/d_i$; sum over $j \neq i$.

(c) Apply Theorem 20.1.1 (a) to $\mathbf{D}^{-1}\mathbf{A}\mathbf{D}$ and use (a) and (b).

■

Taking $d_1 > 1$ and the other $d_j = 1$ divides the first radius by d_1 and multiplies the entry $|a_{i1}|$ inside every other radius by d_1. So scaling shrinks some discs at the expense of others, and the art is to shrink the disc we care about while the others can afford to grow. The same move appeared in Chapter 19 §09, where $\mathbf{D} = \operatorname{diag}(\mathbf{x})$ turned the entries of $\mathbf{B}\mathbf{x}$ into row sums.

Example. Separating a disc by scaling

For $\mathbf{A} = \begin{pmatrix} 6 & 2 & 5 \\ 1 & -1 & 1 \\ 1 & 1 & -1 \end{pmatrix}$ of Example (Discs, eigenvectors, and where each eigenvalue sits), the row disc $D_1(\mathbf{A}) = \{|z - 6| \le 7\}$ overlaps $\{|z + 1| \le 2\}$. Find a scaling $\mathbf{D} = \operatorname{diag}(d, 1, 1)$ whose row discs split into two disjoint groups, and find the range of d that does it.

Solution. By Proposition 20.1.3 (b) with $d_1 = d$ and $d_2 = d_3 = 1$, the scaled radii are

$$r_1 = \frac{2 + 5}{d} = \frac{7}{d}, \qquad r_2 = 1 \cdot d + 1 = d + 1, \qquad r_3 = d + 1,$$

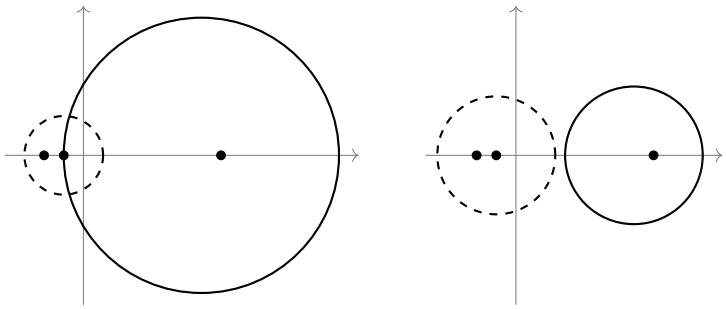
with centers $6, -1, -1$. The first disc is disjoint from the other two exactly when the distance 7 between the centers exceeds the sum of the radii:

$$\frac{7}{d} + d + 1 < 7 \iff d^2 - 6d + 7 < 0 \iff 3 - \sqrt{2} < d < 3 + \sqrt{2},$$

where we multiplied by $d > 0$. For instance $d = 2$ gives $\mathbf{D}^{-1}\mathbf{A}\mathbf{D} = \begin{pmatrix} 6 & 1 & 5/2 \\ 2 & -1 & 1 \\ 2 & 1 & -1 \end{pmatrix}$ and the discs

$$\{|z-6| \le \tfrac{7}{2}\} \quad \text{and} \quad \{|z+1| \le 3\},$$

the second counted twice, which are disjoint since $\frac{7}{2} + 3 < 7$. The eigenvalues 7, -1, -2 are distributed one in the first disc and two in the second, as the figure shows.



unscaled scaled by $\mathrm{diag}(2, 1, 1)$

Row discs of the matrix of the example before and after scaling: the solid disc is centered at 6, the dashed disc at -1 (rows 2 and 3), and the dots are the eigenvalues 7, -1, -2

Scaling has bought a picture in which the eigenvalues fall into two groups, one near 6 and two near -1. What it has **not** yet bought is the statement that the first disc contains exactly one eigenvalue. Theorem 20.1.1 says only that the union holds all three, and nothing so far forbids all three from lying in the second disc. The counting statement is true, but it needs the eigenvalues to move continuously as the off-diagonal part is switched on, and §03 proves it. For a real matrix like this one, §03 will also show that an isolated disc centered on the real axis holds a **real** eigenvalue.

20.1.6 Eigenvalues on the boundary

Diagonal dominance fails for a matrix that matters. The second-difference matrix $\mathbf{K}_n$ of Chapter 19 §09 has 2 on the diagonal and -1 beside it, so for $n \ge 3$ its interior rows have $|a_{ii}| = 2 = r_i(\mathbf{K}_n)$. The discs $|z-2| \le 2$ pass through 0, and Corollary 20.1.2 says nothing, although $\mathbf{K}_n$ is invertible. Only the first and last rows are strict. What rescues it is that the rows are **linked**: the dominant coordinate of an eigenvector cannot stay in the equality rows without spreading to the strict ones. The link is irreducibility, in the sense of Definition 19.2.3, and by Theorem 19.2.5 it means that the graph $G(\mathbf{A})$, with an edge $j \to i$ whenever $a_{ij} \neq 0$, is strongly connected.

Theorem 20.1.4. Taussky's Theorem

Let $n \ge 1$, let $\mathbf{A} \in M_n(\mathbb{C})$ be **irreducible**, and let λ be an eigenvalue of $\mathbf{A}$ with

$$|\lambda - a_{ii}| \ge r_i(\mathbf{A}) \qquad \text{for **every** } i,$$

that is, λ lies in no open disc $\{|z - a_{ii}| < r_i(\mathbf{A})\}$. Then $|\lambda - a_{ii}| = r_i(\mathbf{A})$ for every i: the eigenvalue lies on the boundary circle of **every** Gershgorin disc. Moreover, every eigenvector for λ has all its coordinates of the same modulus. The hypothesis holds in particular when λ is a boundary point (Definition 18.3.2) of $D_1(\mathbf{A}) \cup \cdots \cup D_n(\mathbf{A})$.

Idea. Run the proof of Theorem 20.1.1 at a peak index k. The hypothesis $|\lambda - a_{kk}| \geq r_k$ makes the two ends of its chain of inequalities equal, so every inequality in the chain is an equality. The second one, $\sum_{j\neq k}|a_{kj}||x_j| \leq r_k|x_k|$, is an equality only if $|x_j| = |x_k|$ whenever $a_{kj} \neq 0$. So the set of peak indices is closed under following edges of $G(\mathbf{A})$ backwards. Strong connectivity then drags every index into it.

Proof. Let $\mathbf{Ax} = \lambda\mathbf{x}$ with $\mathbf{x} \neq \mathbf{0}$, put $m = \max_j|x_j| > 0$ and $S = \{k : |x_k| = m\}$, which is nonempty.

Claim. If $k \in S$, then $|\lambda - a_{kk}| = r_k(\mathbf{A})$, and every $j \neq k$ with $a_{kj} \neq 0$ lies in S.

Proof. By the hypothesis and the chain of inequalities in the proof of Theorem 20.1.1, with $|x_k| = m$,

$$r_k(\mathbf{A})\, m \leq |\lambda - a_{kk}|\, m \leq \sum_{j\neq k}|a_{kj}|\,|x_j| \leq r_k(\mathbf{A})\, m.$$

The two ends agree, so every inequality is an equality. The first gives $|\lambda - a_{kk}| = r_k(\mathbf{A})$, after dividing by $m > 0$. The last says $\sum_{j\neq k}|a_{kj}|\,(m - |x_j|) = 0$. Each term is nonnegative, so each is 0, and $a_{kj} \neq 0$ forces $|x_j| = m$, that is, $j \in S$.

■

Now let i be any index and fix $k \in S$. By Theorem 19.2.5, $G(\mathbf{A})$ has a walk $i = i_0, i_1, \ldots, i_t = k$. We show $i_s \in S$ for $s = t, t-1, \ldots, 0$. For $s = t$ this is $k \in S$. If $i_s \in S$ with $s \geq 1$, then $i_{s-1} \to i_s$ is an edge, so $a_{i_s i_{s-1}} \neq 0$; either $i_{s-1} = i_s \in S$, or the Claim applied to i_s puts i_{s-1} in S. Hence $i \in S$. So $S = \{1, \ldots, n\}$: every coordinate of $\mathbf{x}$ has modulus m, and the Claim gives $|\lambda - a_{ii}| = r_i(\mathbf{A})$ for every i.

For the last sentence, suppose $|\lambda - a_{ii}| < r_i(\mathbf{A})$ for some i, and put $\delta = r_i(\mathbf{A}) - |\lambda - a_{ii}| > 0$. Every z with $|z - \lambda| < \delta$ satisfies $|z - a_{ii}| \leq |z - \lambda| + |\lambda - a_{ii}| < r_i(\mathbf{A})$, so it lies in $D_i(\mathbf{A})$ and in the union. Then λ is an interior point of the union, not a boundary point. This proves the theorem.

■

For $n = 1$ the theorem is empty but true: $r_1 = 0$ and the only eigenvalue is a_{11}. For the all-ones matrix $\mathbf{J} \in M_n(\mathbb{C})$, $n \geq 2$, which is irreducible because every entry is nonzero, every disc is $|z - 1| \leq n - 1$, and the eigenvalue n, with eigenvector $\mathbf{1}$, satisfies $|n - 1| = n - 1$: it lies on all of the boundary circles at once, and its eigenvector has equal coordinates, exactly as the theorem predicts.

The corollary is the one $\mathbf{K}_n$ needed: equality in most rows is allowed, provided one row is strict and the rows are linked.

Corollary 20.1.5. Irreducible Diagonal Dominance Implies Invertibility

Let $n \geq 1$ and let $\mathbf{A} \in M_n(\mathbb{C})$ be **irreducible**, with

$$|a_{ii}| \geq r_i(\mathbf{A}) \quad \text{for every } i, \qquad |a_{ii}| > r_i(\mathbf{A}) \quad \text{for **at least one** } i.$$

Then $\mathbf{A}$ is invertible.

Proof. Suppose $\mathbf{A}$ is not invertible. By Theorem 9.2.1, 0 is an eigenvalue. The first hypothesis says $|0 - a_{ii}| \geq r_i(\mathbf{A})$ for every i, so Theorem 20.1.4 applies to $\lambda = 0$ and gives $|a_{ii}| = r_i(\mathbf{A})$ for every i. This contradicts the second hypothesis. Hence $\mathbf{A}$ is invertible.

■

Both hypotheses are needed. $\left(\begin{smallmatrix} 1 & 1 \\ 1 & 1 \end{smallmatrix}\right)$ is irreducible and weakly dominant, with no strict row, and it is singular: its eigenvalue 0 lies on both circles $|z - 1| = 1$. Exercise C2 shows that irreducibility cannot be dropped either.

Example. The second-difference matrix

Show that $\mathbf{K}_n$ is invertible for every $n \geq 1$, and that its eigenvalues lie in the open interval $(0, 4)$.

Solution. For $n = 1$, $\mathbf{K}_1 = (2)$, with eigenvalue 2. Let $n \geq 2$. (Chapter 19 §09 proved invertibility by exhibiting the inverse, Proposition 19.9.6; here the entries alone suffice.) The only nonzero entries off the diagonal are $a_{i,i+1} = a_{i+1,i} = -1$, so $G(\mathbf{K}_n)$ has the edges $i \to i+1$ and $i+1 \to i$, and walking along them reaches every vertex from every vertex. So $\mathbf{K}_n$ is irreducible by Theorem 19.2.5. Its rows have $r_1 = r_n = 1 < 2$ and, when $n \geq 3$, $r_i = 2$ for $1 < i < n$. So $|a_{ii}| \geq r_i$ everywhere, strictly in row 1, and $\mathbf{K}_n$ is invertible by Corollary 20.1.5.

$\mathbf{K}_n$ is real symmetric, so its eigenvalues are real, and by Theorem 20.1.1 (a) they lie in $\bigcup_i [2 - r_i, 2 + r_i] \subseteq [0, 4]$. The value 0 is excluded by invertibility. The value 4 satisfies $|4 - 2| = 2 \geq r_i$ for every i, so if it were an eigenvalue, Theorem 20.1.4 would give $2 = r_1 = 1$, which is false. Hence every eigenvalue lies in $(0, 4)$. In particular $\mathbf{K}_n \succ \mathbf{O}$, which Chapter 19 §09 obtained from its M-matrix structure.

Quick check. The all-ones matrix $\mathbf{J} \in M_3(\mathbb{C})$ is irreducible, and 0 is an eigenvalue of it, since $\mathbf{J}$ has rank 1. But 0 does not lie on the circle $|z - 1| = 2$. Does this contradict Theorem 20.1.4?

Solution. No. The theorem applies only to an eigenvalue with $|\lambda - a_{ii}| \geq r_i(\mathbf{J})$ for every i, and $|0 - 1| = 1 < 2 = r_i(\mathbf{J})$: the eigenvalue 0 lies inside every open disc, so the hypothesis fails and the theorem says nothing about it. The other eigenvalue, 3, satisfies $|3 - 1| = 2$, and it does lie on every circle.

The picture to take away is this. The union of the discs holds the spectrum. An eigenvalue sits in the disc of the row where its eigenvector peaks. Scaling moves radius from one disc to another. And for an irreducible matrix, an eigenvalue can reach the outer edge of the union only by reaching the edge of every disc at once. §02 replaces the discs by other regions built from the same entries, and §03 counts the eigenvalues inside a group of discs.

20.1.7 Exercises

20.1.7.1 A. Check your understanding

Exercise. A1

(a) Define the Gershgorin radius $r_i(\mathbf{A})$ and the disc $D_i(\mathbf{A})$, and state Theorem 20.1.1, including the statement about the peak coordinate of an eigenvector.

(b) True or false: every Gershgorin disc of $\mathbf{A}$ contains an eigenvalue of $\mathbf{A}$. Justify your answer.

(c) True or false: every eigenvalue of $\mathbf{A}$ lies in $\bigcup_i (D_i(\mathbf{A}) \cap D_i(\mathbf{A}^\top))$. Justify your answer.

(d) True or false: an invertible matrix is strictly diagonally dominant by rows or by columns. Justify your answer.

(e) True or false: an irreducible matrix with $|a_{ii}| \geq r_i(\mathbf{A})$ for every i is invertible. Justify your answer.

(f) Name the method of proof of Theorem 20.1.1 in one line.

Solution.

(a) $r_i(\mathbf{A}) = \sum_{j \neq i} |a_{ij}|$ and $D_i(\mathbf{A}) = \{z \in \mathbb{C} : |z - a_{ii}| \leq r_i(\mathbf{A})\}$. If $\mathbf{Ax} = \lambda \mathbf{x}$ with $\mathbf{x} \neq \mathbf{0}$ and $|x_k| = \max_j |x_j|$, then $\lambda \in D_k(\mathbf{A})$; hence every eigenvalue lies in $\bigcup_i D_i(\mathbf{A})$, in $\bigcup_i D_i(\mathbf{A}^\top)$, and in the intersection of these two unions.

(b) False. For $\left(\begin{smallmatrix} 0 & 1 \\ 4 & 0 \end{smallmatrix}\right)$ the eigenvalues are ± 2 and the first disc is $|z| \leq 1$.

(c) False. For the same matrix both intersections are $|z| \leq 1$, which misses ± 2.

(d) False. $\left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$ is invertible, and $|a_{11}| = 0 < 1$ both for rows and for columns.

(e) False. $\left(\begin{smallmatrix} 1 & 1 \\ 1 & 1 \end{smallmatrix}\right)$ is irreducible, with $|a_{ii}| = 1 = r_i$, and singular. Corollary 20.1.5 needs one strict row.

(f) Choose the coordinate of the eigenvector of largest modulus, and apply the triangle inequality (Theorem 0.6.4) to that row of $\mathbf{Ax} = \lambda \mathbf{x}$.

20.1.7.2 B. Practice

Exercise. B1: Deciding invertibility

Determine which of the following matrices are invertible. Justify your answers, using the results of this section where they apply and a direct computation where they do not.

(a) $\begin{pmatrix} 5 & 2 & -2 \\ 1 & -4 & 2 \\ 0 & 3 & 4i \end{pmatrix}$

(b) $\begin{pmatrix} 3 & 1 & 2 \\ 1 & 3 & 1 \\ 1 & 1 & 3 \end{pmatrix}$

(c) $\begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix}$

(d) $\begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 1 & 2 \end{pmatrix}$

Solution.

(a) The radii are $4, 3, 3$ and the diagonal moduli $5, 4, |4i| = 4$, so every row is strictly dominant. The matrix is invertible by Corollary 20.1.2.

(b) Row 1 has $3 = 1 + 2$, and rows 2 and 3 have $3 > 2$. Every off-diagonal entry is nonzero, so every edge $j \to i$ with $i \neq j$ is present in the graph, which is strongly connected, and the matrix is irreducible (Theorem 19.2.5). It is invertible by Corollary 20.1.5. (Indeed its determinant is 18.)

(c) Every row has $2 = 1 + 1$: no row is strict, so neither corollary applies. Directly, the matrix is $\mathbf{I} + \mathbf{J}$, and $\mathbf{J1} = 3 \cdot \mathbf{1}$, while $\mathbf{Jv} = \mathbf{0}$ for $\mathbf{v} = (1, -1, 0)$ and $\mathbf{v} = (1, 0, -1)$. So the eigenvalues of $\mathbf{I} + \mathbf{J}$ are $4, 1, 1$, none zero, and the matrix is invertible, with determinant 4. The corollaries give sufficient conditions only.

(d) Rows 1 and 2 have equality and row 3 is strict, but the matrix is singular: its first two rows are equal. The corollary does not apply because the matrix is reducible: column 3 is zero outside row 3, so no edge leaves vertex 3, and there is no walk from 3 to 1.

Exercise. B2: An interval for a symmetric matrix

Let $\mathbf{A} = \begin{pmatrix} 5 & 1 & -1 \\ 1 & 3 & 1 \\ -1 & 1 & 6 \end{pmatrix}$.

(a) Use Theorem 20.1.1 to find an interval containing every eigenvalue of $\mathbf{A}$, and compare it with the interval $[-\|\mathbf{A}\|_\infty, \|\mathbf{A}\|_\infty]$ given by Corollary 16.4.2.

(b) Hence show that $\mathbf{A}$ is positive definite, and that $\det \mathbf{A}$ lies between 1 and 8^3.

Solution.

(a) $\mathbf{A}$ is real symmetric, so its eigenvalues are real. The radii are $2, 2, 2$, so the discs meet $\mathbb{R}$ in $[3, 7]$, $[1, 5]$, $[4, 8]$, whose union is $[1, 8]$. By Theorem 20.1.1 (a), every eigenvalue lies in $[1, 8]$. The largest absolute row sum is $\|\mathbf{A}\|_\infty = 8$, so Chapter 16's bound gives only $[-8, 8]$; the upper ends agree, but only the discs see that the eigenvalues are positive.

(b) Every eigenvalue is at least $1 > 0$, so $\mathbf{A} \succ \mathbf{O}$ by Theorem 13.1.5. The determinant is the product of the three eigenvalues, each in $[1, 8]$, so $1 \leq \det \mathbf{A} \leq 8^3 = 512$. (In fact $\det \mathbf{A} = 74$.)

Exercise. B3: Choosing the scaling

Let $\mathbf{A} = \begin{pmatrix} 7 & 4 & 4 \\ 1 & 0 & -1 \\ -1 & -2 & -1 \end{pmatrix}$.

(a) Show that the row disc $D_1(\mathbf{A})$ meets the other two row discs.

(b) For $\mathbf{D} = \operatorname{diag}(d, 1, 1)$ with $d > 0$, find the scaled radii, and find all d for which the first scaled disc is disjoint from the other two.

(c) Check that $(-6, -1, 1)$ is an eigenvector of $\mathbf{A}$, and that its eigenvalue lies in the first scaled disc for $d = 3$.

Solution.

(a) The discs are $|z - 7| \le 8$, $|z| \le 2$, and $|z + 1| \le 3$. The point 0 lies in the first two, and -1 in the first and third.

(b) By Proposition 20.1.3 (b), the radii of $\mathbf{D}^{-1}\mathbf{A}\mathbf{D}$ are $r_1 = 8/d$, $r_2 = 1 \cdot d + 1 = d + 1$, and $r_3 = 1 \cdot d + 2 = d + 2$, with centers $7, 0, -1$. The first disc misses the second when $8/d + d + 1 < 7$, and misses the third when $8/d + d + 2 < 8$: both say $d + 8/d < 6$. Multiplying by $d > 0$, this is $d^2 - 6d + 8 < 0$, that is, $(d - 2)(d - 4) < 0$, so $2 < d < 4$. At $d = 2$ the discs touch.

(c) Here

$$\begin{aligned}\mathbf{A}(-6, -1, 1) &= (-42 - 4 + 4,\ -6 + 0 - 1,\ 6 + 2 - 1) \\ &= (-42, -7, 7) = 7\,(-6, -1, 1),\end{aligned}$$

so 7 is an eigenvalue. For $d = 3$ the first scaled disc is $|z - 7| \le \frac{8}{3}$, which contains 7. The eigenvector peaks at the first coordinate, so Theorem 20.1.1 already placed 7 in D_1; the scaled theorem places it in the smaller disc.

20.1.7.3 C. Going deeper

Exercise. C1: Dominance and definiteness

(a) Prove that a Hermitian matrix $\mathbf{A}$ with $a_{ii} > r_i(\mathbf{A})$ for every i is positive definite.

(b) Prove that a Hermitian **irreducible** matrix with $a_{ii} \ge r_i(\mathbf{A})$ for every i, strictly for at least one i, is positive definite.

(c) Deduce that for every $n \ge 1$ and every real $t \ge 0$, the matrix $\mathbf{K}_n + t\,\mathbf{e}_1\mathbf{e}_1^{\top}$ is positive definite.

Hint for (b): apply (a)'s argument to find where the eigenvalues can be, and then use Theorem 20.1.4.

Solution.

(a) The diagonal entries of a Hermitian matrix are real, and so are its eigenvalues (Theorem 12.2.1). By Theorem 20.1.1 (a), each eigenvalue λ lies in some $D_k(\mathbf{A})$, so $\lambda \ge a_{kk} - |\lambda - a_{kk}| \ge a_{kk} - r_k(\mathbf{A}) > 0$. A Hermitian matrix with only positive eigenvalues is positive definite (Theorem 13.1.5).

(b) As in (a), every eigenvalue satisfies $\lambda \ge a_{kk} - r_k(\mathbf{A}) \ge 0$ for some k, so it is enough to show that 0 is not an eigenvalue. The hypothesis gives $|0 - a_{ii}| = a_{ii} \ge r_i(\mathbf{A})$ for every i, since $a_{ii} \ge r_i(\mathbf{A}) \ge 0$. If 0 were an eigenvalue, Theorem 20.1.4 would give $a_{ii} = r_i(\mathbf{A})$ for every i, contradicting the strict row. Hence all eigenvalues are positive and $\mathbf{A} \succ \mathbf{O}$. (This is also Corollary 20.1.5, together with the sign information from (a).)

(c) The matrix is real symmetric, and it differs from $\mathbf{K}_n$ only in the $(1,1)$ entry, which is $2+t$. The graph is unchanged, so for $n \geq 2$ it is irreducible, as shown in Example (The second-difference matrix). Its rows satisfy $a_{ii} \geq r_i$, with $2+t > 1$ in row 1. By (b), it is positive definite. For $n = 1$ the matrix is $(2+t)$ with $2+t > 0$.

Exercise. C2: Irreducibility cannot be dropped

(a) Give a 3×3 reducible matrix with $|a_{ii}| \geq r_i(\mathbf{A})$ for every i, strictly in at least one row, that is singular. Which step of the proof of Theorem 20.1.4 fails for it?

(b) Let $\mathbf{A}$ be irreducible and let λ be an eigenvalue with $|\lambda - a_{ii}| \geq r_i(\mathbf{A})$ for every i. Prove that the eigenspace $E_\lambda(\mathbf{A})$ has dimension 1.

Hint for (b): given two independent eigenvectors, form a combination with a zero coordinate.

Solution.

(a) Take $\mathbf{A} = \left(\begin{smallmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 1 & 2 \end{smallmatrix}\right)$ from Exercise (B1: Deciding invertibility) (d): rows 1 and 2 have equality, row 3 is strict, and the first two rows are equal, so $\mathbf{A}$ is singular. The eigenvector $\mathbf{x} = (1, -1, \frac{1}{2})$ for 0 has $\mathbf{A}\mathbf{x} = (0, 0, -1+1) = \mathbf{0}$. Its peak set is $S = \{1, 2\}$, and the Claim of the proof holds: the rows in S have equality and their off-diagonal nonzero entries point into S. What fails is the spreading step. There is no walk from 3 to 1 in $G(\mathbf{A})$, so nothing forces $3 \in S$, and the strict row 3 is never reached.

(b) Suppose $\mathbf{x}$ and $\mathbf{y}$ are linearly independent eigenvectors for λ. Since $\mathbf{y} \neq \mathbf{0}$, choose i with $y_i \neq 0$, and put $\mathbf{z} = \mathbf{x} - (x_i/y_i)\mathbf{y}$. Then $\mathbf{z} \neq \mathbf{0}$, by independence, and $\mathbf{A}\mathbf{z} = \lambda\mathbf{z}$, so $\mathbf{z}$ is an eigenvector for λ with $z_i = 0$. But by Theorem 20.1.4 every eigenvector for λ has all coordinates of the same modulus, which is positive because the vector is nonzero. This contradiction shows $\dim E_\lambda(\mathbf{A}) \leq 1$, and $\dim E_\lambda(\mathbf{A}) \geq 1$ because λ is an eigenvalue.

20.2 Other Inclusion Regions

The Gershgorin discs of §01 come from one row of $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$, read at the largest coordinate of $\mathbf{x}$. That leaves two kinds of information unused. The first is the second largest coordinate, which the triangle inequality simply replaced by the largest; using it gives a region of ovals that is never worse than the discs. The second is the geometry of the inner product, which the discs ignore entirely: the number $\mathbf{x}^*\mathbf{A}\mathbf{x}$ over unit vectors $\mathbf{x}$ fills a region of the plane, the numerical range, that contains every eigenvalue. This section builds both regions. Its main theorem is that the numerical range is always convex. It ends with a rectangle for the spectrum whose sides are eigenvalues of Hermitian matrices, and with a warning about what these regions cannot do.

Throughout, matrices are complex and $n \geq 1$, and $\|\cdot\|$ on $\mathbb{C}^n$ is the Euclidean norm. As in §01, $r_i(\mathbf{A}) = \sum_{j \neq i} |a_{ij}|$ is the i-th Gershgorin radius and $D_i(\mathbf{A})$ the i-th disc (Definition 20.1.1). The plane $\mathbb{C}$ is treated as a real vector space when convexity is discussed, as Chapter 18 §01 does.

20.2.1 Using two coordinates: Brauer's ovals

In the proof of Theorem 20.1.1 every $|x_j|$ with $j \neq k$ was bounded by the largest one, $|x_k|$. That is wasteful when the second largest coordinate is much smaller. So keep two indices: p, where $|x_p|$ is largest, and $q \neq p$, where it is largest among the rest. Row p bounds $|\lambda - a_{pp}||x_p|$ by $r_p|x_q|$, since every other coordinate is at most $|x_q|$. Row q bounds $|\lambda - a_{qq}||x_q|$ by $r_q|x_p|$. Multiplying the two bounds cancels the unknown coordinates. The result is a region defined by a product of distances.

Theorem 20.2.1. Brauer's Ovals of Cassini

Let $n \geq 2$ and $\mathbf{A} \in M_n(\mathbb{C})$. For $i \neq j$ put

$$O_{ij}(\mathbf{A}) := \{\, z \in \mathbb{C} : |z - a_{ii}|\,|z - a_{jj}| \leq r_i(\mathbf{A})\, r_j(\mathbf{A}) \,\}.$$

(a) Every eigenvalue of $\mathbf{A}$ lies in $\bigcup_{i \neq j} O_{ij}(\mathbf{A})$.

(b) $O_{ij}(\mathbf{A}) \subseteq D_i(\mathbf{A}) \cup D_j(\mathbf{A})$ for every $i \neq j$. Hence $\bigcup_{i \neq j} O_{ij}(\mathbf{A}) \subseteq \bigcup_i D_i(\mathbf{A})$.

Idea. For (a), take the two largest coordinates of an eigenvector as above and multiply the two row estimates. One case needs separate care: if the second largest coordinate is 0, there is nothing to divide by, but then $\mathbf{x}$ is a multiple of $\mathbf{e}_p$ and $\lambda = a_{pp}$, which lies in every oval through a_{pp}. For (b), a point outside both discs is farther than r_i from a_{ii} and farther than r_j from a_{jj}, so the product of its distances exceeds $r_i r_j$.

Proof.

(a) Let $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$ with $\mathbf{x} \neq \mathbf{0}$. Choose p with $|x_p| = \max_j |x_j|$, which is positive, and then $q \neq p$ with $|x_q| = \max_{j \neq p} |x_j|$; this is possible because $n \geq 2$.
Case 1: $x_q = 0$. Then $x_j = 0$ for every $j \neq p$, and row p of $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$ reads $a_{pp}x_p = \lambda x_p$. As $x_p \neq 0$, $\lambda = a_{pp}$. The left side of the inequality defining $O_{pq}(\mathbf{A})$ is then 0, and the right side is ≥ 0, so $\lambda \in O_{pq}(\mathbf{A})$.
Case 2: $x_q \neq 0$. Row p gives $(\lambda - a_{pp})x_p = \sum_{j \neq p} a_{pj}x_j$. Every $j \neq p$ has $|x_j| \leq |x_q|$, so by the triangle inequality (Theorem 0.6.4)

$$|\lambda - a_{pp}|\,|x_p| \leq r_p(\mathbf{A})\,|x_q|.$$

Row q gives $(\lambda - a_{qq})x_q = \sum_{j \neq q} a_{qj}x_j$. Every j has $|x_j| \leq |x_p|$, so

$$|\lambda - a_{qq}|\,|x_q| \leq r_q(\mathbf{A})\,|x_p|.$$

Both sides of both inequalities are nonnegative, so they may be multiplied:

$$|\lambda - a_{pp}|\,|\lambda - a_{qq}|\,|x_p|\,|x_q| \leq r_p(\mathbf{A})\, r_q(\mathbf{A})\,|x_p|\,|x_q|.$$

Dividing by $|x_p||x_q| > 0$ gives $\lambda \in O_{pq}(\mathbf{A})$.

(b) Let $z \notin D_i(\mathbf{A}) \cup D_j(\mathbf{A})$, so $a := |z - a_{ii}| > r_i(\mathbf{A}) \geq 0$ and $c := |z - a_{jj}| > r_j(\mathbf{A}) \geq 0$. Since $c > 0$ and $r_i(\mathbf{A}) \geq 0$,

$$ac > r_i(\mathbf{A})\, c \geq r_i(\mathbf{A})\, r_j(\mathbf{A}),$$

so $z \notin O_{ij}(\mathbf{A})$. Taking complements, $O_{ij}(\mathbf{A}) \subseteq D_i(\mathbf{A}) \cup D_j(\mathbf{A})$, and taking the union over all pairs gives the last claim. This proves the theorem.

■

The sets $O_{ij}(\mathbf{A})$ are the **ovals of Cassini**: the points whose distances to two fixed foci have product at most a constant. Since $O_{ij}(\mathbf{A}) = O_{ji}(\mathbf{A})$, there are $n(n-1)/2$ of them to draw, against n discs. The hypothesis $n \geq 2$ is there because an oval needs two distinct indices; for $n = 1$ there are none, and the eigenvalue a_{11} must be found some other way. For $n = 2$ the single oval is as good as a region can be. The eigenvalues satisfy $(\lambda - a_{11})(\lambda - a_{22}) = a_{12}a_{21}$, so taking moduli, $|\lambda - a_{11}||\lambda - a_{22}| = r_1 r_2$: both eigenvalues lie on the boundary curve of O_{12}.

Example. Ovals where discs fail

Let $\mathbf{A} = \begin{pmatrix} 4 & 1 & 0 \\ 0 & 4 & 1 \\ 1 & 1 & 1 \end{pmatrix}$. Show that the Gershgorin discs cannot prove $\mathbf{A}$ invertible, and that the ovals can.

Solution. The radii are $r_1 = 1$, $r_2 = 1$, $r_3 = 2$, and $D_3(\mathbf{A}) = \{|z - 1| \leq 2\}$ contains 0. So Theorem 20.1.1 (a) allows the eigenvalue 0. The column radii are $1, 2, 1$, and the second column disc $|z - 4| \leq 2$ misses 0, but the third, $|z - 1| \leq 1$, contains it. So the discs cannot decide.

Now test $z = 0$ against the three ovals:

$$\begin{aligned} O_{12}: &\ |0-4|\,|0-4| = 16 > 1 = r_1 r_2, \\ O_{13}: &\ |0-4|\,|0-1| = 4 > 2 = r_1 r_3, \\ O_{23}: &\ |0-4|\,|0-1| = 4 > 2 = r_2 r_3. \end{aligned}$$

So 0 lies in no oval, it is not an eigenvalue by Theorem 20.2.1 (a), and $\mathbf{A}$ is invertible by Theorem 9.2.1. Indeed, expanding along the first row, $\det \mathbf{A} = 4(4-1) - 1(0-1) = 13$. The ovals also exclude the point -1, which lies on the boundary of $D_3(\mathbf{A})$: there $|-1-4|\,|-1-4| = 25 > 1$ rules out O_{12}, and $|-1-4|\,|-1-1| = 10 > 2$ rules out O_{13} and O_{23}.

The same test gives a criterion in the spirit of Corollary 20.1.2: if $n \geq 2$, as in Theorem 20.2.1, and $|a_{ii}|\,|a_{jj}| > r_i(\mathbf{A})\, r_j(\mathbf{A})$ for **every** pair $i \neq j$, then 0 lies in no oval and $\mathbf{A}$ is invertible. One row may fail to dominate, as row 3 did above, provided every other row dominates enough to compensate.

Quick check. Why does the proof of Theorem 20.2.1 (a) choose q to maximize $|x_q|$ among the indices other than p, rather than taking any $q \neq p$?

Solution. The estimate from row p bounds every $|x_j|$, $j \neq p$, by $|x_q|$. That is true only if $|x_q|$ is the largest of them. With an arbitrary q the row-p estimate could fail, and the product of the two estimates would no longer cancel down to $r_p r_q$.

20.2.2 The numerical range

Now a different source of information. For a Hermitian $\mathbf{A}$, Chapter 17 §01 studied the Rayleigh quotient, and Proposition 17.1.1 showed that its values $\mathbf{x}^*\mathbf{A}\mathbf{x}$ on unit vectors fill exactly the interval $[\lambda_n(\mathbf{A}), \lambda_1(\mathbf{A})]$, which contains every eigenvalue. The number $\mathbf{x}^*\mathbf{A}\mathbf{x} = \langle \mathbf{A}\mathbf{x}, \mathbf{x}\rangle$ makes sense for every square matrix. Without the Hermitian hypothesis it is complex, and its values fill a region of the plane rather than an interval.

The numerical range is the set of all values $\mathbf{x}^*\mathbf{A}\mathbf{x}$ as $\mathbf{x}$ runs over the unit vectors.

Definition 20.2.1. Numerical Range

Let $\mathbf{A} \in M_n(\mathbb{C})$. The **numerical range** of $\mathbf{A}$ is the subset of $\mathbb{C}$

$$W(\mathbf{A}) \coloneqq \{\, \mathbf{x}^*\mathbf{A}\mathbf{x} : \mathbf{x} \in \mathbb{C}^n,\ \|\mathbf{x}\| = 1 \,\}.$$

In words: feed $\mathbf{A}$ a vector of length **exactly** 1, take the inner product of the output with the input, and collect all the complex numbers that arise. The vectors range over **complex** unit vectors, even when $\mathbf{A}$ is real, and the conjugate in $\mathbf{x}^*$ is part of the definition. $W(\mathbf{A})$ is always nonempty, since $\mathbf{e}_1$ is a unit vector, and $\mathbf{e}_i^*\mathbf{A}\mathbf{e}_i = a_{ii}$, so every diagonal entry lies in $W(\mathbf{A})$. The notation always carries its argument: $W(\mathbf{A})$, never W alone, since W also names subspaces.

Examples, simplest first.

- $n = 1$**, and scalar matrices.** For $\mathbf{A} = (a)$ the unit vectors are the scalars x with $|x| = 1$, and $\overline{x}ax = a|x|^2 = a$. So $W((a)) = \{a\}$. Likewise $W(c\mathbf{I}) = \{c\}$, since $\mathbf{x}^*(c\mathbf{I})\mathbf{x} = c\|\mathbf{x}\|^2 = c$.
- **Diagonal matrices.** For $\mathbf{A} = \operatorname{diag}(d_1, \dots, d_n)$, $\mathbf{x}^*\mathbf{A}\mathbf{x} = \sum_i d_i|x_i|^2$. The weights $t_i = |x_i|^2$ are nonnegative with sum $\|\mathbf{x}\|^2 = 1$, and every such list of weights arises, from $x_i = \sqrt{t_i}$. So $W(\mathbf{A})$ is the set of all weighted averages of $d_1, \dots, d_n$, which is $\operatorname{conv}\{d_1, \dots, d_n\}$ by Theorem 18.1.3. For $\operatorname{diag}(1, i, -1)$ it is the filled triangle with these three vertices.
- **Hermitian matrices.** By Proposition 17.1.1, $W(\mathbf{A}) = [\lambda_n(\mathbf{A}), \lambda_1(\mathbf{A})]$, a segment of the real axis.
- **A Jordan block.** The next example computes $W(\mathbf{J}_2(0))$, and it is a whole disc, although the only eigenvalue is 0.

Example. The numerical range of a Jordan block

Show that $W(\mathbf{J}_2(0)) = \{z \in \mathbb{C} : |z| \le \tfrac{1}{2}\}$.

Solution. Write $\mathbf{x} = (x_1, x_2)$ with $|x_1|^2 + |x_2|^2 = 1$. Then $\mathbf{J}_2(0)\mathbf{x} = (x_2, 0)$, so $\mathbf{x}^*\mathbf{J}_2(0)\mathbf{x} = \overline{x_1}\, x_2$.

($\subseteq$) Since $0 \le (|x_1| - |x_2|)^2 = 1 - 2|x_1||x_2|$, we get $|\overline{x_1}x_2| = |x_1||x_2| \le \tfrac{1}{2}$.

($\supseteq$) Let $z = \rho\,\omega$ with $0 \le \rho \le \tfrac{1}{2}$ and $|\omega| = 1$. Since $1 - 4\rho^2 \ge 0$, we may put

$$a = \sqrt{\tfrac{1+\sqrt{1-4\rho^2}}{2}}, \qquad b = \sqrt{\tfrac{1-\sqrt{1-4\rho^2}}{2}},$$

both real and nonnegative. Then $a^2 + b^2 = 1$ and $a^2b^2 = \tfrac{1}{4}(1 - (1 - 4\rho^2)) = \rho^2$, so $ab = \rho$.

The unit vector $\mathbf{x} = (a, b\,\omega)$ gives $\mathbf{x}^*\mathbf{J}_2(0)\mathbf{x} = a\,b\,\omega = z$. Hence $W(\mathbf{J}_2(0))$ is the closed disc of radius $\frac{1}{2}$ about 0.

Nonexample by minimal change. Drop the conjugate, and collect the numbers $\mathbf{x}^\top\mathbf{A}\mathbf{x}$ for unit $\mathbf{x}$ instead. For $\mathbf{A} = \mathbf{I}_2$ and the unit vector $\mathbf{x} = \frac{1}{\sqrt{2}}(1, i)$, $\mathbf{x}^\top\mathbf{I}_2\mathbf{x} = \frac{1}{2}(1 + i^2) = 0$, while $\mathbf{x}^*\mathbf{I}_2\mathbf{x} = 1$. The modified set for $\mathbf{I}_2$ contains 0, which is not an eigenvalue, while $W(\mathbf{I}_2) = \{1\}$ is exactly the spectrum. What breaks is that $\mathbf{x}^\top\mathbf{x}$ is not $\|\mathbf{x}\|^2$ over $\mathbb{C}$. The conjugate is what makes $\mathbf{x}^*(c\mathbf{I})\mathbf{x} = c$ on the unit sphere, and every property below leans on that.

Why this definition. The unit length plays the role that division by $\|\mathbf{x}\|^2$ played for the Rayleigh quotient. Without it the values $\mathbf{x}^*\mathbf{A}\mathbf{x}$ over all $\mathbf{x}$ form a set closed under multiplication by positive reals, which is all of $\mathbb{C}$ or a cone, and records nothing about the size of $\mathbf{A}$. With it, the set is bounded, and the eigenvalues are in it.

Proposition 20.2.2. Basic Properties of the Numerical Range

Let $\mathbf{A} \in M_n(\mathbb{C})$.

(a) $\operatorname{spec}(\mathbf{A}) \subseteq W(\mathbf{A})$.

(b) $W(\alpha\mathbf{A} + \beta\mathbf{I}) = \alpha W(\mathbf{A}) + \beta \coloneqq \{\alpha w + \beta : w \in W(\mathbf{A})\}$ for all $\alpha, \beta \in \mathbb{C}$.

(c) $W(\mathbf{U}^*\mathbf{A}\mathbf{U}) = W(\mathbf{A})$ for every unitary $\mathbf{U} \in M_n(\mathbb{C})$.

(d) $W(\mathbf{A})$ is compact: every sequence in $W(\mathbf{A})$ has a subsequence converging to a point of $W(\mathbf{A})$. In particular $W(\mathbf{A})$ is closed.

(e) $W(\mathbf{A}) \subseteq \{z : |z| \le \|\mathbf{A}\|_2\}$.

Idea. Four of the five parts are one line of algebra with a unit vector: evaluate at a unit eigenvector for (a), push the scalars or the unitary through $\mathbf{x}^*(\cdot)\mathbf{x}$ for (b) and (c), and apply Cauchy–Schwarz for (e). Only (d) needs analysis, and there the point is that $W(\mathbf{A})$ is the image of the unit sphere under a continuous map, so the compactness of the sphere passes to it.

Proof.

(a) Let $\lambda \in \operatorname{spec}(\mathbf{A})$, with eigenvector $\mathbf{v}$. Then $\mathbf{x} = \mathbf{v}/\|\mathbf{v}\|$ is a unit eigenvector, and $\mathbf{x}^*\mathbf{A}\mathbf{x} = \lambda\,\mathbf{x}^*\mathbf{x} = \lambda$.

(b) For a unit $\mathbf{x}$, $\mathbf{x}^*(\alpha\mathbf{A} + \beta\mathbf{I})\mathbf{x} = \alpha\,\mathbf{x}^*\mathbf{A}\mathbf{x} + \beta\,\mathbf{x}^*\mathbf{x} = \alpha\,\mathbf{x}^*\mathbf{A}\mathbf{x} + \beta$. As $\mathbf{x}$ runs over the unit vectors, the left side runs over $W(\alpha\mathbf{A} + \beta\mathbf{I})$ and the right side over $\alpha W(\mathbf{A}) + \beta$.

(c) For a unit $\mathbf{x}$, $\mathbf{x}^*\mathbf{U}^*\mathbf{A}\mathbf{U}\mathbf{x} = (\mathbf{U}\mathbf{x})^*\mathbf{A}(\mathbf{U}\mathbf{x})$, and $\mathbf{U}\mathbf{x}$ is a unit vector because $\mathbf{U}$ is an isometry (Theorem 11.7.3). So $W(\mathbf{U}^*\mathbf{A}\mathbf{U}) \subseteq W(\mathbf{A})$. Conversely, every unit $\mathbf{y}$ equals $\mathbf{U}\mathbf{x}$ for the unit vector $\mathbf{x} = \mathbf{U}^*\mathbf{y}$, so every value $\mathbf{y}^*\mathbf{A}\mathbf{y}$ arises, and the two sets are equal.

(d) First, $f(\mathbf{x}) = \mathbf{x}^*\mathbf{A}\mathbf{x}$ is continuous on the unit sphere. For unit vectors $\mathbf{x}, \mathbf{y}$,

$$\begin{aligned} |f(\mathbf{x}) - f(\mathbf{y})| &= |\mathbf{x}^*\mathbf{A}(\mathbf{x} - \mathbf{y}) + (\mathbf{x} - \mathbf{y})^*\mathbf{A}\mathbf{y}| \\ &\le 2\,\|\mathbf{A}\|_2\,\|\mathbf{x} - \mathbf{y}\|, \end{aligned}$$

by the triangle inequality (Theorem 0.6.4), the Cauchy–Schwarz inequality (Theo-

rem 11.1.5), and Theorem 16.3.2 (a). This is a bound of the kind fixed as the book's continuity test in Chapter 16's introduction. Now let (w_k) be a sequence in $W(\mathbf{A})$, with $w_k = f(\mathbf{x}_k)$ and $\|\mathbf{x}_k\| = 1$. By **the compactness of the unit sphere, fact (A3) of Chapter 16's introduction**, a subsequence $(\mathbf{x}_{k_l})$ converges to a unit vector $\mathbf{x}$. By continuity, $w_{k_l} = f(\mathbf{x}_{k_l}) \to f(\mathbf{x}) \in W(\mathbf{A})$. If (w_k) itself converges to w, its subsequence converges to w too, and limits are unique, so $w = f(\mathbf{x}) \in W(\mathbf{A})$: the set is closed.

(e) For a unit $\mathbf{x}$, by the Cauchy–Schwarz inequality and Theorem 16.3.2 (a), $|\mathbf{x}^*\mathbf{A}\mathbf{x}| = |\langle \mathbf{A}\mathbf{x}, \mathbf{x}\rangle| \le \|\mathbf{A}\mathbf{x}\|\,\|\mathbf{x}\| \le \|\mathbf{A}\|_2$.

■

Part (a) makes $W(\mathbf{A})$ an inclusion region for the spectrum, and (e) puts it inside the disc of radius $\|\mathbf{A}\|_2$, which upgrades $\rho(\mathbf{A}) \le \|\mathbf{A}\|_2$ from the spectrum to the whole numerical range. Part (b) is the tool that lets us move any two points of $W(\mathbf{A})$ to 0 and 1, and the next proof depends on it. Part (c) says $W(\mathbf{A})$ depends only on the operator $T_{\mathbf{A}}$ and the inner product, not on the orthonormal basis used to write it down.

20.2.3 Convexity: the Toeplitz–Hausdorff theorem

The examples so far were all convex: a point, a segment, a triangle, a disc. That is not a coincidence. To prove it, split $\mathbf{A}$ into two Hermitian pieces, the way a complex number splits into its real and imaginary parts.

Lemma 20.2.3. Real and Imaginary Parts of a Matrix

Let $\mathbf{A} \in M_n(\mathbb{C})$, and put

$$\mathbf{H} = \frac{\mathbf{A} + \mathbf{A}^*}{2}, \qquad \mathbf{K} = \frac{\mathbf{A} - \mathbf{A}^*}{2i}.$$

Then $\mathbf{H}$ and $\mathbf{K}$ are Hermitian, $\mathbf{A} = \mathbf{H} + i\mathbf{K}$, and for every $\mathbf{x} \in \mathbb{C}^n$ the numbers $\mathbf{x}^*\mathbf{H}\mathbf{x}$ and $\mathbf{x}^*\mathbf{K}\mathbf{x}$ are real, with

$$\operatorname{Re}(\mathbf{x}^*\mathbf{A}\mathbf{x}) = \mathbf{x}^*\mathbf{H}\mathbf{x}, \qquad \operatorname{Im}(\mathbf{x}^*\mathbf{A}\mathbf{x}) = \mathbf{x}^*\mathbf{K}\mathbf{x}.$$

Proof. By Definition 11.1.2, $(\mathbf{A}^*)^* = \mathbf{A}$, $(\mathbf{B}+\mathbf{C})^* = \mathbf{B}^* + \mathbf{C}^*$, and $(c\mathbf{B})^* = \bar{c}\mathbf{B}^*$. Hence $\mathbf{H}^* = \frac{1}{2}(\mathbf{A}^* + \mathbf{A}) = \mathbf{H}$ and $\mathbf{K}^* = \frac{1}{-2i}(\mathbf{A}^* - \mathbf{A}) = \frac{1}{2i}(\mathbf{A} - \mathbf{A}^*) = \mathbf{K}$. Also $\mathbf{H} + i\mathbf{K} = \frac{1}{2}(\mathbf{A} + \mathbf{A}^*) + \frac{1}{2}(\mathbf{A} - \mathbf{A}^*) = \mathbf{A}$. For $\mathbf{x} \in \mathbb{C}^n$, $\mathbf{x}^*\mathbf{H}\mathbf{x} = \langle \mathbf{H}\mathbf{x}, \mathbf{x}\rangle$ is real by Proposition 11.6.7 (a), applied to the self-adjoint operator $T_{\mathbf{H}}$, and likewise $\mathbf{x}^*\mathbf{K}\mathbf{x}$. So $\mathbf{x}^*\mathbf{A}\mathbf{x} = \mathbf{x}^*\mathbf{H}\mathbf{x} + i\,\mathbf{x}^*\mathbf{K}\mathbf{x}$ is the splitting of a complex number into real and imaginary parts.

■

For example $\mathbf{J}_2(0)$ has $\mathbf{H} = \frac{1}{2}\left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$ and $\mathbf{K} = \frac{1}{2i}\left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right) = \frac{1}{2}\left(\begin{smallmatrix} 0 & -i \\ i & 0 \end{smallmatrix}\right)$, both with eigenvalues $\pm\frac{1}{2}$. A Hermitian $\mathbf{A}$ has $\mathbf{K} = \mathbf{O}$, and then $W(\mathbf{A})$ lies on the real axis.

Theorem 20.2.4. Toeplitz–Hausdorff Theorem

For every $\mathbf{A} \in M_n(\mathbb{C})$, the numerical range $W(\mathbf{A})$ is a convex subset of $\mathbb{C}$.

Idea. **Step roadmap.** ① By Proposition 20.2.2 (b), an affine change $z \mapsto \alpha z + \beta$ carries any two points of $W(\mathbf{A})$ to 0 and 1, and carries segments to segments; so it suffices to show that $[0,1] \subseteq W(\mathbf{B})$ whenever $0, 1 \in W(\mathbf{B})$. ② Take unit vectors $\mathbf{x}$, $\mathbf{y}$ with $\mathbf{x}^*\mathbf{B}\mathbf{x} = 0$ and $\mathbf{y}^*\mathbf{B}\mathbf{y} = 1$, and walk from $\mathbf{x}$ to $\mathbf{y}$ along the straight path $(1-t)\mathbf{x} + t\mathbf{y}$. Along the way the value of $\mathbf{B}$, normalized, moves continuously from 0 to 1, but it may wander off the real axis. ③ Its imaginary part along the path is $\mathbf{K}$ evaluated there, and expanding shows that only the cross term $\mathbf{x}^*\mathbf{K}\mathbf{y} + \mathbf{y}^*\mathbf{K}\mathbf{x} = 2\,\mathrm{Re}(\mathbf{x}^*\mathbf{K}\mathbf{y})$ can be nonzero. Multiplying $\mathbf{y}$ by a unimodular scalar rotates $\mathbf{x}^*\mathbf{K}\mathbf{y}$ freely without changing $\mathbf{y}^*\mathbf{B}\mathbf{y}$, so we rotate it onto the imaginary axis, where its real part is 0. ④ Then the path stays on the real axis, and the intermediate value theorem fills in $[0,1]$.

Proof. **Step 1: reduction.** Let $w_0, w_1 \in W(\mathbf{A})$ and $0 \le s \le 1$; we must show $(1-s)w_0 + sw_1 \in W(\mathbf{A})$. If $w_0 = w_1$ there is nothing to prove. Otherwise put $\alpha = 1/(w_1 - w_0)$, $\beta = -\alpha w_0$, and $\mathbf{B} = \alpha\mathbf{A} + \beta\mathbf{I}$. By Proposition 20.2.2 (b), $W(\mathbf{B}) = \alpha W(\mathbf{A}) + \beta$, which contains $\alpha w_0 + \beta = 0$ and $\alpha w_1 + \beta = 1$. Suppose we show $s \in W(\mathbf{B})$. Then $s = \alpha w + \beta$ for some $w \in W(\mathbf{A})$, and solving, $w = w_0 + s(w_1 - w_0) = (1-s)w_0 + sw_1$. So it suffices to prove: **if** $0, 1 \in W(\mathbf{B})$**, then** $[0,1] \subseteq W(\mathbf{B})$**.**

Step 2: choosing the vectors. Let $\mathbf{B} = \mathbf{H} + i\mathbf{K}$ as in Lemma 20.2.3, and choose unit vectors $\mathbf{x}$, $\mathbf{y}_0$ with $\mathbf{x}^*\mathbf{B}\mathbf{x} = 0$ and $\mathbf{y}_0^*\mathbf{B}\mathbf{y}_0 = 1$. By Lemma 20.2.3, $\mathbf{x}^*\mathbf{H}\mathbf{x} = \mathbf{x}^*\mathbf{K}\mathbf{x} = 0$, $\mathbf{y}_0^*\mathbf{H}\mathbf{y}_0 = 1$, and $\mathbf{y}_0^*\mathbf{K}\mathbf{y}_0 = 0$. Let $c = \mathbf{x}^*\mathbf{K}\mathbf{y}_0$, and put $\gamma = 1$ if $c = 0$, and $\gamma = i\bar{c}/|c|$ otherwise. In both cases $|\gamma| = 1$ and $\gamma c = i|c|$ is purely imaginary. Set $\mathbf{y} = \gamma\mathbf{y}_0$. Then $\|\mathbf{y}\| = 1$ and $\mathbf{y}^*\mathbf{B}\mathbf{y} = |\gamma|^2\,\mathbf{y}_0^*\mathbf{B}\mathbf{y}_0 = 1$, so still $\mathbf{y}^*\mathbf{H}\mathbf{y} = 1$ and $\mathbf{y}^*\mathbf{K}\mathbf{y} = 0$. Since $\mathbf{K}$ is Hermitian, $\mathbf{y}^*\mathbf{K}\mathbf{x} = (\mathbf{x}^*\mathbf{K}\mathbf{y})^* = \overline{\mathbf{x}^*\mathbf{K}\mathbf{y}}$, and so

$$\mathbf{x}^*\mathbf{K}\mathbf{y} + \mathbf{y}^*\mathbf{K}\mathbf{x} = 2\,\mathrm{Re}(\mathbf{x}^*\mathbf{K}\mathbf{y}) = 2\,\mathrm{Re}(\gamma c) = 0.$$

Finally, $\mathbf{x}$ and $\mathbf{y}$ are linearly independent. Otherwise, both being nonzero, $\mathbf{y} = \mu\mathbf{x}$ for a scalar μ, and then $1 = \mathbf{y}^*\mathbf{B}\mathbf{y} = |\mu|^2\,\mathbf{x}^*\mathbf{B}\mathbf{x} = 0$, which is absurd.

Step 3: the path stays real. For $t \in [0,1]$ put $\mathbf{v}(t) = (1-t)\mathbf{x} + t\mathbf{y}$. The coefficients $1-t$ and t are not both zero, so $\mathbf{v}(t) \neq \mathbf{0}$ by independence. Expanding, and using $\mathbf{x}^*\mathbf{K}\mathbf{x} = \mathbf{y}^*\mathbf{K}\mathbf{y} = 0$ and Step 2,

$$\mathbf{v}(t)^*\mathbf{K}\mathbf{v}(t) = (1-t)^2\,\mathbf{x}^*\mathbf{K}\mathbf{x} + t^2\,\mathbf{y}^*\mathbf{K}\mathbf{y} + t(1-t)(\mathbf{x}^*\mathbf{K}\mathbf{y} + \mathbf{y}^*\mathbf{K}\mathbf{x}) = 0.$$

So $\mathbf{v}(t)^*\mathbf{B}\mathbf{v}(t) = \mathbf{v}(t)^*\mathbf{H}\mathbf{v}(t)$ is real for every t. The same expansion with $\mathbf{H}$ in place of $\mathbf{K}$, and with $\|\mathbf{v}(t)\|^2 = \mathbf{v}(t)^*\mathbf{v}(t)$, gives

$$\begin{aligned} \mathbf{v}(t)^*\mathbf{H}\mathbf{v}(t) &= t^2 + 2t(1-t)\,\mathrm{Re}(\mathbf{x}^*\mathbf{H}\mathbf{y}), \\ \|\mathbf{v}(t)\|^2 &= (1-t)^2 + t^2 + 2t(1-t)\,\mathrm{Re}(\mathbf{x}^*\mathbf{y}), \end{aligned}$$

two polynomials in t with real coefficients, the second positive on $[0,1]$.

Step 4: filling the segment. The function

$$f(t) = \frac{\mathbf{v}(t)^*\mathbf{H}\mathbf{v}(t)}{\|\mathbf{v}(t)\|^2}, \qquad 0 \le t \le 1,$$

is a quotient of polynomials with a denominator that never vanishes, hence continuous by the algebra of limits, and $f(0) = \mathbf{x}^*\mathbf{H}\mathbf{x} = 0$, $f(1) = \mathbf{y}^*\mathbf{H}\mathbf{y} = 1$. Let $s \in [0,1]$. By **the intermediate value theorem, fact (A5) of Chapter 16's introduction**, there is $t_s \in [0,1]$ with $f(t_s) = s$. The vector $\mathbf{u} = \mathbf{v}(t_s)/\|\mathbf{v}(t_s)\|$ is a unit vector, and by Step 3,

$$\mathbf{u}^*\mathbf{B}\mathbf{u} = \frac{\mathbf{v}(t_s)^*\mathbf{B}\mathbf{v}(t_s)}{\|\mathbf{v}(t_s)\|^2} = \frac{\mathbf{v}(t_s)^*\mathbf{H}\mathbf{v}(t_s)}{\|\mathbf{v}(t_s)\|^2} = f(t_s) = s.$$

So $s \in W(\mathbf{B})$. By Step 1, $W(\mathbf{A})$ contains the segment between any two of its points, which is Definition 18.1.1. This proves the theorem.

■

The phase choice in Step 2 is the heart of the proof. Without it the imaginary part along the path is $2t(1-t)\,\mathrm{Re}(\mathbf{x}^*\mathbf{K}\mathbf{y}_0)/\|\mathbf{v}(t)\|^2$, which may be nonzero, and the path would leave the real axis between 0 and 1. The rotated vector $\gamma\mathbf{y}_0$ spans the same line as $\mathbf{y}_0$ and gives the same value of $\mathbf{B}$, so the rotation costs nothing. Only the path between $\mathbf{x}$ and $\mathbf{y}$ changes.

With convexity, $W(\mathbf{A})$ contains the convex hull of $\mathrm{spec}(\mathbf{A})$, by (a) and Definition 18.1.4, and it contains the diagonal entries, so it contains $\mathrm{conv}\{a_{11}, \dots, a_{nn}\}$ as well.

20.2.4 Normal matrices

For a diagonal matrix the numerical range is the convex hull of the diagonal entries, and those are the eigenvalues. Unitary similarity does not change W, and a normal matrix is unitarily similar to a diagonal one. So for a normal matrix $W(\mathbf{A})$ is as small as (a) and convexity allow.

Proposition 20.2.5. The Numerical Range of a Normal Matrix

Let $\mathbf{A} \in M_n(\mathbb{C})$ be normal, with eigenvalues $\lambda_1, \dots, \lambda_n$, listed with multiplicity. Then

$$W(\mathbf{A}) = \mathrm{conv}\{\lambda_1, \dots, \lambda_n\}.$$

Idea. Unitary similarity leaves W alone, by Proposition 20.2.2 (c), so we may replace $\mathbf{A}$ by a diagonal matrix of its eigenvalues. Then $\mathbf{x}^*\mathbf{A}\mathbf{x}$ is the average of the λ_i with weights $|x_i|^2$, and the two inclusions say exactly that the lists of admissible weights and of convex coefficients are the same.

Proof. By Corollary 12.4.3, $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{U}^*$ with $\mathbf{U}$ unitary and $\mathbf{D} = \mathrm{diag}(\lambda_1, \dots, \lambda_n)$. So $\mathbf{U}^*\mathbf{A}\mathbf{U} = \mathbf{D}$, and $W(\mathbf{A}) = W(\mathbf{D})$ by Proposition 20.2.2 (c). For a unit $\mathbf{x}$, $\mathbf{x}^*\mathbf{D}\mathbf{x} = \sum_i |x_i|^2 \lambda_i$.

($\subseteq$) The weights $|x_i|^2$ are nonnegative with sum 1, so $\mathbf{x}^*\mathbf{D}\mathbf{x}$ is a convex combination of $\lambda_1, \dots, \lambda_n$, and lies in their convex hull by Theorem 18.1.3.

($\supseteq$) By Theorem 18.1.3, a point of the hull is $\sum_{k=1}^{m} t_k \mu_k$ with $t_k \ge 0$, $\sum_k t_k = 1$, and each μ_k one of the λ_i. Assign each k to one index i with $\lambda_i = \mu_k$, and let s_i be the sum of the

t_k assigned to i. Then $s_i \geq 0$, $\sum_i s_i = 1$, and the point is $\sum_i s_i \lambda_i$. The unit vector with $x_i = \sqrt{s_i}$ gives $\mathbf{x}^* \mathbf{D} \mathbf{x} = \sum_i s_i \lambda_i$.

■

For a Hermitian matrix this recovers Proposition 17.1.1. For a unitary matrix, whose eigenvalues lie on the unit circle, $W(\mathbf{A})$ is the convex hull of those eigenvalues: a polygon inscribed in the circle when three or more of them are distinct, and a chord or a single point when there are two or one. For a nonnormal matrix $W(\mathbf{A})$ can be much larger than $\operatorname{conv}\operatorname{spec}(\mathbf{A})$: by Example (The numerical range of a Jordan block), $W(\mathbf{J}_2(0))$ is a disc of radius $\frac{1}{2}$ around the single eigenvalue 0. The extra room is a measure of nonnormality, and Exercise C3 turns it into a comparison with $\|\mathbf{A}\|_2$.

Quick check. Find $W(\mathbf{A})$ for the rotation $\mathbf{A} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$, regarded as a complex matrix.

Solution. $\mathbf{A}^{\top}\mathbf{A} = \mathbf{I}$, so $\mathbf{A}$ is real orthogonal, hence unitary and normal. Its characteristic polynomial is $x^2 + 1$, with roots $\pm i$. By Proposition 20.2.5, $W(\mathbf{A})$ is the segment from $-i$ to i on the imaginary axis. For real unit vectors $\mathbf{x}$ the value $\mathbf{x}^*\mathbf{A}\mathbf{x}$ is 0, which is why the real Rayleigh-type quotient of Chapter 17 §01 saw nothing; the complex unit vectors fill in the rest.

20.2.5 Bendixson's rectangle

By Lemma 20.2.3, the real part of $\mathbf{x}^*\mathbf{A}\mathbf{x}$ is a value of the Hermitian form of $\mathbf{H}$, and Proposition 17.1.1 confines those values to $[\lambda_n(\mathbf{H}), \lambda_1(\mathbf{H})]$. The same holds for the imaginary part and $\mathbf{K}$. Together they box in the whole numerical range, hence the spectrum.

Theorem 20.2.6. Bendixson's Theorem

Let $\mathbf{A} \in M_n(\mathbb{C})$, with $\mathbf{H} = (\mathbf{A} + \mathbf{A}^*)/2$ and $\mathbf{K} = (\mathbf{A} - \mathbf{A}^*)/(2i)$. Every $w \in W(\mathbf{A})$ satisfies

$$\lambda_n(\mathbf{H}) \leq \operatorname{Re} w \leq \lambda_1(\mathbf{H}),$$
$$\lambda_n(\mathbf{K}) \leq \operatorname{Im} w \leq \lambda_1(\mathbf{K}).$$

In particular every eigenvalue λ of $\mathbf{A}$ satisfies the same inequalities with λ in place of w.

Proof. Let $w = \mathbf{x}^*\mathbf{A}\mathbf{x}$ with $\|\mathbf{x}\| = 1$. By Lemma 20.2.3, $\operatorname{Re} w = \mathbf{x}^*\mathbf{H}\mathbf{x} = \langle \mathbf{H}\mathbf{x}, \mathbf{x} \rangle$ and $\operatorname{Im} w = \mathbf{x}^*\mathbf{K}\mathbf{x}$, and $\mathbf{H}$, $\mathbf{K}$ are Hermitian. By Lemma 13.5.2, $\lambda_n(\mathbf{H}) \leq \langle \mathbf{H}\mathbf{x}, \mathbf{x} \rangle \leq \lambda_1(\mathbf{H})$, and likewise for $\mathbf{K}$. Every eigenvalue lies in $W(\mathbf{A})$ by Proposition 20.2.2 (a).

■

The most used consequence concerns the sign of the real parts. If $\mathbf{A} + \mathbf{A}^* \succ \mathbf{O}$, then $\lambda_n(\mathbf{H}) > 0$, so every eigenvalue of $\mathbf{A}$ has positive real part, and in particular $\mathbf{A}$ is invertible. This needs nothing about the size of the off-diagonal entries, only about their symmetric part.

Example. A rectangle beats the discs

Let $\mathbf{A} = \begin{pmatrix} 3 & 2 \\ -4 & 1 \end{pmatrix}$. Show that every eigenvalue of $\mathbf{A}$ has positive real part, using Theorem 20.2.6, and check that the Gershgorin discs cannot show it. Compare with the eigenvalues.

Solution. Here $\mathbf{A}^* = \mathbf{A}^\top$, so

$$\mathbf{H} = \begin{pmatrix} 3 & -1 \\ -1 & 1 \end{pmatrix}, \qquad \mathbf{K} = \frac{1}{2i}\begin{pmatrix} 0 & 6 \\ -6 & 0 \end{pmatrix} = \begin{pmatrix} 0 & -3i \\ 3i & 0 \end{pmatrix}.$$

$\mathbf{H}$ has trace 4 and determinant 2, so $p_{\mathbf{H}}(x) = x^2 - 4x + 2$, with roots $2 \pm \sqrt{2}$. $\mathbf{K}$ has trace 0 and determinant $0 - (-3i)(3i) = -9$, so its eigenvalues are ± 3. By Theorem 20.2.6, every eigenvalue lies in the rectangle

$$2 - \sqrt{2} \le \operatorname{Re}\lambda \le 2 + \sqrt{2}, \qquad -3 \le \operatorname{Im}\lambda \le 3,$$

and in particular $\operatorname{Re}\lambda \ge 2 - \sqrt{2} > 0$.

The row discs are $|z - 3| \le 2$ and $|z - 1| \le 4$, and the second contains -3. The column discs are $|z - 3| \le 4$ and $|z - 1| \le 2$, and the first contains -1. So neither union rules out an eigenvalue with negative real part.

Directly, $p_{\mathbf{A}}(x) = x^2 - 4x + 11$, so $\lambda = 2 \pm i\sqrt{7}$. The real part 2 lies in $[2 - \sqrt{2}, 2 + \sqrt{2}]$, and the imaginary parts $\pm\sqrt{7} \approx \pm 2.65$ lie in $[-3, 3]$.

20.2.6 What these regions do not do

Every region in this section and the last is an **inclusion region**: it contains the spectrum. None of them is a perturbation bound in the sense of the rest of this chapter, which asks how far the eigenvalues move when $\mathbf{A}$ changes by a small $\mathbf{E}$. The numerical range does give a statement of that kind. For a unit $\mathbf{x}$, $\mathbf{x}^*(\mathbf{A}+\mathbf{E})\mathbf{x} = \mathbf{x}^*\mathbf{A}\mathbf{x} + \mathbf{x}^*\mathbf{E}\mathbf{x}$, and $|\mathbf{x}^*\mathbf{E}\mathbf{x}| \le \|\mathbf{E}\|_2$ by Proposition 20.2.2 (e). Hence

$$W(\mathbf{A}+\mathbf{E}) \subseteq W(\mathbf{A}) + W(\mathbf{E}) \subseteq \{\, w + e : w \in W(\mathbf{A}),\ |e| \le \|\mathbf{E}\|_2 \,\},$$

and so every eigenvalue of $\mathbf{A} + \mathbf{E}$ lies within $\|\mathbf{E}\|_2$ of $W(\mathbf{A})$. This is only as good as $W(\mathbf{A})$ is small.

Warning. **$W(\mathbf{A})$ can be far larger than $\operatorname{spec}(\mathbf{A})$, so "within $\|\mathbf{E}\|_2$ of $W(\mathbf{A})$" is not "within $\|\mathbf{E}\|_2$ of an eigenvalue".** For $\mathbf{A} = M\mathbf{J}_2(0)$ with $M > 0$, Proposition 20.2.2 (b) and Example (The numerical range of a Jordan block) give $W(\mathbf{A}) = \{|z| \le M/2\}$, while $\operatorname{spec}(\mathbf{A}) = \{0\}$. So the bound above allows the eigenvalues of $\mathbf{A} + \mathbf{E}$ to lie anywhere in $|z| \le M/2 + \|\mathbf{E}\|_2$, and this does not shrink as $\mathbf{E} \to \mathbf{O}$. The eigenvalues do move a long way: $\mathbf{A} + \varepsilon\mathbf{e}_2\mathbf{e}_1^\top$ has characteristic polynomial $x^2 - M\varepsilon$, so for $\varepsilon > 0$ its eigenvalues $\pm\sqrt{M\varepsilon}$ have moved by $\sqrt{M\varepsilon}$, far more than ε. For a normal $\mathbf{A}$, $W(\mathbf{A}) = \operatorname{conv}\operatorname{spec}(\mathbf{A})$ is small, but it still contains points far from every eigenvalue, such as the midpoint of two distant ones.

The genuine perturbation bounds need a different mechanism: §03 proves that eigenvalues move

continuously and at what rate in general, and §04 gives Bauer–Fike's bound, in which the distance to the nearest **eigenvalue** is controlled.

20.2.7 Exercises

20.2.7.1 A. Check your understanding

Exercise. A1

(a) Define the Brauer oval $O_{ij}(\mathbf{A})$ and the numerical range $W(\mathbf{A})$, and state the Toeplitz–Hausdorff theorem.

(b) True or false: $W(\mathbf{A}) = \operatorname{conv}\operatorname{spec}(\mathbf{A})$ for every $\mathbf{A} \in M_n(\mathbb{C})$. Justify your answer.

(c) True or false: $W(\mathbf{A})$ is contained in the union of the Gershgorin discs of $\mathbf{A}$. Justify your answer.

(d) True or false: if $\mathbf{A} + \mathbf{A}^* \succ \mathbf{O}$, then $\mathbf{A}$ is invertible. Justify your answer.

(e) Why does Theorem 20.2.1 require $n \geq 2$?

Solution.

(a) For $i \neq j$, $O_{ij}(\mathbf{A}) = \{z : |z - a_{ii}||z - a_{jj}| \leq r_i(\mathbf{A})r_j(\mathbf{A})\}$, and $W(\mathbf{A}) = \{\mathbf{x}^*\mathbf{A}\mathbf{x} : \|\mathbf{x}\| = 1\}$. Theorem 20.2.4: for every square complex $\mathbf{A}$, $W(\mathbf{A})$ is convex.

(b) False. $W(\mathbf{J}_2(0))$ is the disc $|z| \leq \frac{1}{2}$ (Example (The numerical range of a Jordan block)), while $\operatorname{conv}\operatorname{spec}(\mathbf{J}_2(0)) = \{0\}$. Equality holds for normal matrices (Proposition 20.2.5).

(c) False. For $\mathbf{A} = \operatorname{diag}(1, -1)$ the discs are the two points $\{1\}$ and $\{-1\}$, while $W(\mathbf{A}) = [-1, 1]$ contains 0. The discs contain the spectrum, not the numerical range.

(d) True. By Theorem 20.2.6 every eigenvalue has $\operatorname{Re}\lambda \geq \lambda_n(\mathbf{H}) > 0$, with $\mathbf{H} = (\mathbf{A} + \mathbf{A}^*)/2 \succ \mathbf{O}$. So 0 is not an eigenvalue, and $\mathbf{A}$ is invertible (Theorem 9.2.1).

(e) An oval is built from two distinct indices $i \neq j$, and the proof needs a second index $q \neq p$. For $n = 1$ there are no ovals, and the union in (a) would be empty, although the matrix has an eigenvalue.

20.2.7.2 B. Practice

Exercise. B1: Invertibility from ovals

Let $\mathbf{A} = \begin{pmatrix} 5 & 1 & 1 \\ 1 & 5 & 1 \\ 2 & 2 & 2 \end{pmatrix}$.

(a) Show that neither the row discs nor the column discs of $\mathbf{A}$ exclude 0.

(b) Use Theorem 20.2.1 to show that $\mathbf{A}$ is invertible.

(c) Check that $(1, -1, 0)$ is an eigenvector, and verify that its eigenvalue lies in $O_{12}(\mathbf{A})$.

Solution.

(a) The row radii are $2, 2, 4$, and $D_3(\mathbf{A}) = \{|z - 2| \leq 4\}$ contains 0. The column radii are $3, 3, 2$, and the third column disc $|z - 2| \leq 2$ contains 0.

(b) At $z = 0$: for O_{12}, $5 \cdot 5 = 25 > 4 = r_1 r_2$; for O_{13} and O_{23}, $5 \cdot 2 = 10 > 8 = r_1 r_3 = r_2 r_3$. So 0 lies in no oval, is not an eigenvalue by Theorem 20.2.1 (a), and $\mathbf{A}$ is invertible. (Its determinant is 32.)

(c) $\mathbf{A}(1, -1, 0) = (5 - 1, 1 - 5, 2 - 2) = (4, -4, 0)$, so the eigenvalue is 4. For O_{12}: $|4 - 5||4 - 5| = 1 \le 4$.

Exercise. B2: A Bendixson rectangle

Let $\mathbf{A} = \begin{pmatrix} 1 & 3 \\ -1 & 1 \end{pmatrix}$. Find the rectangle of Theorem 20.2.6, and hence show that every eigenvalue of $\mathbf{A}$ satisfies $0 \le \operatorname{Re}\lambda \le 2$ and $|\operatorname{Im}\lambda| \le 2$. Compare with the eigenvalues.

Solution. $\mathbf{H} = \frac{1}{2}(\mathbf{A} + \mathbf{A}^\top) = \left(\begin{smallmatrix} 1 & 1 \\ 1 & 1 \end{smallmatrix}\right)$, with $p_{\mathbf{H}}(x) = x^2 - 2x$, so its eigenvalues are 0 and 2. $\mathbf{K} = \frac{1}{2i}(\mathbf{A} - \mathbf{A}^\top) = \frac{1}{2i}\left(\begin{smallmatrix} 0 & 4 \\ -4 & 0 \end{smallmatrix}\right) = \left(\begin{smallmatrix} 0 & -2i \\ 2i & 0 \end{smallmatrix}\right)$, with trace 0 and determinant -4, so its eigenvalues are ± 2. By Theorem 20.2.6 every eigenvalue has $0 \le \operatorname{Re}\lambda \le 2$ and $-2 \le \operatorname{Im}\lambda \le 2$. Directly, $p_{\mathbf{A}}(x) = x^2 - 2x + 4$, so $\lambda = 1 \pm i\sqrt{3}$: the real part 1 lies in $[0, 2]$, and $\sqrt{3} \approx 1.73 \le 2$.

Exercise. B3: Computing numerical ranges

Determine $W(\mathbf{A})$ for each matrix. Justify your answers.

(a) $\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix}$

(b) $\mathbf{A} = \operatorname{diag}(1, i, -1, -i)$

(c) $\mathbf{A} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$

Solution.

(a) $\mathbf{A} = 2\mathbf{J}_2(0) + \mathbf{I}$, so by Proposition 20.2.2 (b) and Example (The numerical range of a Jordan block), $W(\mathbf{A}) = 2W(\mathbf{J}_2(0)) + 1 = \{z : |z - 1| \le 1\}$, the closed disc of radius 1 about 1.

(b) A diagonal matrix is normal, so by Proposition 20.2.5 $W(\mathbf{A}) = \operatorname{conv}\{1, i, -1, -i\}$, the filled square with these four vertices.

(c) $\mathbf{A}$ is real symmetric, with eigenvalues 3 and 1 (eigenvectors $(1, 1)$ and $(1, -1)$). By Proposition 17.1.1, or Proposition 20.2.5, $W(\mathbf{A}) = [1, 3]$.

20.2.7.3 C. Going deeper

Exercise. C1: Direct sums

(a) Let $\mathbf{B} \in M_k(\mathbb{C})$ and $\mathbf{C} \in M_m(\mathbb{C})$. Prove that $W(\mathbf{B} \oplus \mathbf{C}) = \operatorname{conv}(W(\mathbf{B}) \cup W(\mathbf{C}))$.

(b) Deduce that $\mathbf{A} = \operatorname{diag}(2, -2, 2i, -2i) \oplus \mathbf{J}_2(0)$ is not normal, but satisfies $W(\mathbf{A}) = \operatorname{conv}\operatorname{spec}(\mathbf{A})$.

Hint for (a): one inclusion needs Theorem 20.2.4.

Solution.

(a) ($\subseteq$) Write a unit vector of $\mathbb{C}^{k+m}$ as $\mathbf{x} = (\mathbf{y}, \mathbf{z})$, with $\|\mathbf{y}\|^2 + \|\mathbf{z}\|^2 = 1$. Then $\mathbf{x}^*(\mathbf{B} \oplus \mathbf{C})\mathbf{x} = \mathbf{y}^*\mathbf{B}\mathbf{y} + \mathbf{z}^*\mathbf{C}\mathbf{z}$. If $\mathbf{z} = \mathbf{0}$ this is a point of $W(\mathbf{B})$, and if $\mathbf{y} = \mathbf{0}$ a point of $W(\mathbf{C})$. Otherwise, with $s = \|\mathbf{y}\|^2 \in (0, 1)$,

$$\mathbf{x}^*(\mathbf{B} \oplus \mathbf{C})\mathbf{x} = s\,\frac{\mathbf{y}^*\mathbf{B}\mathbf{y}}{\|\mathbf{y}\|^2} + (1 - s)\,\frac{\mathbf{z}^*\mathbf{C}\mathbf{z}}{\|\mathbf{z}\|^2},$$

a convex combination of a point of $W(\mathbf{B})$, at the unit vector $\mathbf{y}/\|\mathbf{y}\|$, and a point of $W(\mathbf{C})$. In every case the value lies in $\operatorname{conv}(W(\mathbf{B}) \cup W(\mathbf{C}))$, by Theorem 18.1.3.
($\supseteq$) Taking $\mathbf{x} = (\mathbf{y}, \mathbf{0})$ with $\mathbf{y}$ a unit vector shows $W(\mathbf{B}) \subseteq W(\mathbf{B} \oplus \mathbf{C})$, and likewise $W(\mathbf{C}) \subseteq W(\mathbf{B} \oplus \mathbf{C})$. By Theorem 20.2.4, $W(\mathbf{B} \oplus \mathbf{C})$ is convex, so it contains the smallest convex set containing $W(\mathbf{B}) \cup W(\mathbf{C})$ (Definition 18.1.4).

(b) $\mathbf{A}$ is not normal: its lower-right block is $\mathbf{J}_2(0)$, and $\mathbf{A}^*\mathbf{A}$ and $\mathbf{A}\mathbf{A}^*$ differ in that block, since $\mathbf{J}_2(0)^*\mathbf{J}_2(0) = \operatorname{diag}(0, 1)$ and $\mathbf{J}_2(0)\mathbf{J}_2(0)^* = \operatorname{diag}(1, 0)$. By (a), Proposition 20.2.5, and Example (The numerical range of a Jordan block), $W(\mathbf{A})$ is the convex hull of the square $Q = \operatorname{conv}\{\pm 2, \pm 2i\}$ and the disc $\{|z| \le \frac{1}{2}\}$. The square contains the disc: $Q = \{z : |\operatorname{Re} z| + |\operatorname{Im} z| \le 2\}$, and $|\operatorname{Re} z| + |\operatorname{Im} z| \le 2|z| \le 1$ on the disc. So $W(\mathbf{A}) = Q$. The spectrum is $\{\pm 2, \pm 2i, 0\}$, whose hull is also Q, since $0 \in Q$. Hence $W(\mathbf{A}) = \operatorname{conv}\operatorname{spec}(\mathbf{A})$, although $\mathbf{A}$ is not normal: the converse of Proposition 20.2.5 fails.

Exercise. C2: When the numerical range is real, or a point

Let $\mathbf{A} \in M_n(\mathbb{C})$.

(a) Prove that $W(\mathbf{A}) \subseteq \mathbb{R}$ if and only if $\mathbf{A}$ is Hermitian.

(b) Deduce that $W(\mathbf{A}) = \{c\}$ for a single $c \in \mathbb{C}$ if and only if $\mathbf{A} = c\mathbf{I}$.

Solution.

(a) ($\Leftarrow$) If $\mathbf{A}$ is Hermitian, $\mathbf{x}^*\mathbf{A}\mathbf{x} = \langle \mathbf{A}\mathbf{x}, \mathbf{x}\rangle$ is real by Proposition 11.6.7 (a). ($\Rightarrow$) Write $\mathbf{A} = \mathbf{H} + i\mathbf{K}$ as in Lemma 20.2.3. If $W(\mathbf{A}) \subseteq \mathbb{R}$, then $\mathbf{x}^*\mathbf{K}\mathbf{x} = \operatorname{Im}(\mathbf{x}^*\mathbf{A}\mathbf{x}) = 0$ for every unit $\mathbf{x}$. By Proposition 17.1.1, $\lambda_1(\mathbf{K})$ and $\lambda_n(\mathbf{K})$ are values of $\mathbf{x}^*\mathbf{K}\mathbf{x}$ on unit vectors, so both are 0, and every eigenvalue of $\mathbf{K}$ is 0. By Corollary 12.4.3, $\mathbf{K} = \mathbf{U}\,\mathbf{O}\,\mathbf{U}^* = \mathbf{O}$. Hence $\mathbf{A} = \mathbf{H}$ is Hermitian.

(b) ($\Leftarrow$) is the example $W(c\mathbf{I}) = \{c\}$. ($\Rightarrow$) By Proposition 20.2.2 (b), $W(\mathbf{A} - c\mathbf{I}) = W(\mathbf{A}) - c = \{0\} \subseteq \mathbb{R}$, so $\mathbf{A} - c\mathbf{I}$ is Hermitian by (a). By Proposition 17.1.1 its extreme eigenvalues are values of the form on unit vectors, hence both 0, and as in (a) the Hermitian matrix $\mathbf{A} - c\mathbf{I}$ with all eigenvalues 0 is $\mathbf{O}$.

Exercise. C3: The numerical radius

For $\mathbf{A} \in M_n(\mathbb{C})$, the **numerical radius** is $w(\mathbf{A}) = \max\{|z| : z \in W(\mathbf{A})\}$.

(a) Explain why the maximum exists, and prove that $\rho(\mathbf{A}) \le w(\mathbf{A}) \le \|\mathbf{A}\|_2$.

(b) Prove that $\|\mathbf{A}\|_2 \le 2\, w(\mathbf{A})$.

(c) Show that both constants are attained: $w(\mathbf{A}) = \|\mathbf{A}\|_2$ for every normal $\mathbf{A}$, and $\|\mathbf{A}\|_2 = 2\, w(\mathbf{A})$ for $\mathbf{A} = \mathbf{J}_2(0)$.

Hint for (b): bound $\|\mathbf{H}\|_2$ and $\|\mathbf{K}\|_2$ separately.

Solution.

(a) $W(\mathbf{A})$ is nonempty and compact (Proposition 20.2.2 (d)), and $z \mapsto |z|$ is continuous, so the maximum is attained by **the extreme value theorem, fact (A4) of Chapter 16's introduction**. Every eigenvalue lies in $W(\mathbf{A})$, by Proposition 20.2.2 (a), so $\rho(\mathbf{A}) \le w(\mathbf{A})$; and $w(\mathbf{A}) \le \|\mathbf{A}\|_2$ by Proposition 20.2.2 (e).

(b) Let $\mathbf{A} = \mathbf{H} + i\mathbf{K}$ as in Lemma 20.2.3. By Lemma 17.3.5, $\|\mathbf{H}\|_2$ is the larger of $\lambda_1(\mathbf{H})$ and $-\lambda_n(\mathbf{H})$. By Proposition 17.1.1 there is a unit $\mathbf{x}$ with

$$\lambda_1(\mathbf{H}) = \mathbf{x}^*\mathbf{H}\mathbf{x} = \operatorname{Re}(\mathbf{x}^*\mathbf{A}\mathbf{x}) \le |\mathbf{x}^*\mathbf{A}\mathbf{x}| \le w(\mathbf{A}),$$

and a unit $\mathbf{y}$ with $-\lambda_n(\mathbf{H}) = -\operatorname{Re}(\mathbf{y}^*\mathbf{A}\mathbf{y}) \le w(\mathbf{A})$. So $\|\mathbf{H}\|_2 \le w(\mathbf{A})$, and the same argument with imaginary parts gives $\|\mathbf{K}\|_2 \le w(\mathbf{A})$. As the operator norm is a norm (Theorem 16.3.2 (b)), $\|\mathbf{A}\|_2 \le \|\mathbf{H}\|_2 + \|i\mathbf{K}\|_2 \le 2\, w(\mathbf{A})$.

(c) For normal $\mathbf{A}$, $\rho(\mathbf{A}) = \|\mathbf{A}\|_2$, as Chapter 16 §04 showed after Definition 16.4.1, so the chain in (a) collapses to equalities. For $\mathbf{J}_2(0)$, Example (The numerical range of a Jordan block) gives $w(\mathbf{J}_2(0)) = \frac{1}{2}$, and $\|\mathbf{J}_2(0)\|_2 = 1$, since $\|\mathbf{J}_2(0)\mathbf{x}\| = |x_2| \le \|\mathbf{x}\|$ with equality at $\mathbf{x} = \mathbf{e}_2$. So $\|\mathbf{J}_2(0)\|_2 = 2\, w(\mathbf{J}_2(0))$.

20.3 Continuity of Roots

Chapter 16 §07 proved that the eigenvalues of a matrix move continuously with the matrix, but only after quoting a theorem it did not prove: the roots of a monic polynomial move continuously with its coefficients (Theorem 16.7.3). This section proves that theorem, with nothing more than the compactness of closed bounded sets. It then asks two quantitative questions the qualitative theorem cannot answer. How far can an eigenvalue move when the matrix moves by a given amount, with no hypothesis on the matrix at all? And how many eigenvalues does a region of the plane hold? The second question finishes the story of Gershgorin's discs begun in Section 1: an isolated disc holds **exactly one** eigenvalue.

Throughout, matrices are complex and eigenvalues are listed with algebraic multiplicity, so that $\mathbf{A} \in M_n(\mathbb{C})$ has exactly n of them (Corollary 6.5.5). Two imported facts of analysis are used, and each is named where it acts: compactness, fact (A3) of Chapter 16's introduction, and the intermediate value theorem, fact (A5).

20.3.1 A bound for the roots

The proof of continuity will take a sequence of polynomials and follow their roots. For that to work, the roots must not escape to infinity, and the first step is a bound on the roots in terms of the coefficients.

Lemma 20.3.1. Cauchy's Bound for the Roots

Let $n \geq 1$ and let $p(x) = x^n + c_{n-1}x^{n-1} + \cdots + c_1 x + c_0 \in \mathbb{C}[x]$ be monic. Put $M = \max_{0 \leq j \leq n-1} |c_j|$. Then every root $z \in \mathbb{C}$ of p satisfies

$$|z| \leq 1 + M.$$

Idea. A root satisfies $z^n = -(c_{n-1}z^{n-1} + \cdots + c_0)$. If $|z|$ is large, the left side is much bigger than every single term on the right, and there are only n terms; summing the geometric series on the right makes this precise.

Proof. If $|z| \leq 1$ there is nothing to prove, so suppose $|z| > 1$. Since $p(z) = 0$, we have $z^n = -\sum_{j=0}^{n-1} c_j z^j$, and the triangle inequality (Theorem 0.6.4) gives

$$|z|^n \leq M \sum_{j=0}^{n-1} |z|^j = M \, \frac{|z|^n - 1}{|z| - 1} \leq M \, \frac{|z|^n}{|z| - 1},$$

where the equality sums a geometric series with ratio $|z| \neq 1$. Dividing by $|z|^n > 0$ and multiplying by $|z| - 1 > 0$ gives $|z| - 1 \leq M$, as claimed.

■

The bound depends only on the size of the coefficients, and that is what the next proof needs. Polynomials whose coefficients stay bounded have roots that stay bounded.

Quick check. Show that the 1 in the bound cannot be dropped: find a monic polynomial with a root of modulus larger than $M = \max_j |c_j|$.

Solution. Take $p(x) = x^2 - x - 1$, with $M = 1$. Its roots are $(1 \pm \sqrt{5})/2$, and $(1 + \sqrt{5})/2 \approx 1.618 > 1$. The bound $1 + M = 2$ does hold.

20.3.2 The roots move continuously

Here is the theorem of Chapter 16 (Theorem 16.7.3), restated in full under its own label, since from now on it is a proved result.

Theorem 20.3.2. Continuous Dependence of the Roots on the Coefficients

Let $n \geq 1$ and let

$$p(x) = x^n + c_{n-1}x^{n-1} + \cdots + c_1 x + c_0 \in \mathbb{C}[x]$$

be monic, with roots $\mu_1, \ldots, \mu_n \in \mathbb{C}$ listed with multiplicity. Let $\varepsilon > 0$. Then there is $\delta > 0$ with the following property. If

$$q(x) = x^n + b_{n-1}x^{n-1} + \cdots + b_1 x + b_0$$

is monic of the **same degree** n and $|b_j - c_j| < \delta$ for every j, then the roots $\nu_1, \dots, \nu_n$ of q, listed with multiplicity, can be **numbered** so that $|\nu_i - \mu_i| < \varepsilon$ for $i = 1, \dots, n$.

Idea. The hard direction, from coefficients to roots, has no formula. The easy direction does: expanding $(x-\nu_1)\cdots(x-\nu_n)$ writes each coefficient as a polynomial in the roots. So argue by contradiction and let compactness supply the missing formula.

① A bad sequence $q_k \to p$ has root vectors that stay bounded, by Cauchy's bound. ② By compactness, a subsequence of root vectors converges to some $\boldsymbol{\nu} \in \mathbb{C}^n$. ③ Coefficients are polynomials in roots, so $(x-\nu_1)\cdots(x-\nu_n)$ has the limiting coefficients, which are those of p. ④ Unique factorization says $\boldsymbol{\nu}$ is a rearrangement of $(\mu_1, \dots, \mu_n)$. ⑤ That rearrangement is a good numbering for the late q_k, which were supposed to have none.

Proof. Suppose the conclusion fails for some $\varepsilon > 0$. Then no δ works; in particular $\delta = 1/k$ fails for every integer $k \geq 1$. So there is a monic $q_k(x) = x^n + \sum_{j<n} b_j^{(k)} x^j$ with $|b_j^{(k)} - c_j| < 1/k$ for every j, whose roots admit **no** numbering within ε of $\mu_1, \dots, \mu_n$. By Corollary 6.5.5 (a), $q_k = (x-\nu_1^{(k)})\cdots(x-\nu_n^{(k)})$ for some $\boldsymbol{\nu}^{(k)} = (\nu_1^{(k)}, \dots, \nu_n^{(k)}) \in \mathbb{C}^n$, which lists the roots of q_k with multiplicity.

The root vectors are bounded. Put $M = \max_j |c_j|$. Then $|b_j^{(k)}| < M + 1$, so by Lemma 20.3.1 every $|\nu_i^{(k)}| \leq M + 2$, and $\|\boldsymbol{\nu}^{(k)}\|_2 \leq \sqrt{n}\,(M+2)$. By compactness, fact (A3), the sequence $(\boldsymbol{\nu}^{(k)})$ has a subsequence $(\boldsymbol{\nu}^{(k_l)})_{l \geq 1}$ converging in $\mathbb{C}^n$ to some $\boldsymbol{\nu} = (\nu_1, \dots, \nu_n)$. Since $|\nu_i^{(k_l)} - \nu_i| \leq \|\boldsymbol{\nu}^{(k_l)} - \boldsymbol{\nu}\|_2$, each coordinate converges: $\nu_i^{(k_l)} \to \nu_i$.

The limit is a root vector of p. Expanding the product $(x-z_1)\cdots(x-z_n)$ gives $x^n + \sum_{j<n} s_j(z_1, \dots, z_n) x^j$, where each s_j is a fixed polynomial with integer coefficients in $z_1, \dots, z_n$ (it is $\pm$ the sum of all products of $n-j$ of the z_i with distinct indices). Hence $b_j^{(k)} = s_j(\boldsymbol{\nu}^{(k)})$. Along the subsequence, the algebra of limits gives $s_j(\boldsymbol{\nu}^{(k_l)}) \to s_j(\boldsymbol{\nu})$, while $b_j^{(k_l)} \to c_j$ by the choice of q_k. Limits are unique, so $s_j(\boldsymbol{\nu}) = c_j$ for every j, that is,

$$(x-\nu_1)(x-\nu_2)\cdots(x-\nu_n) = p(x) = (x-\mu_1)(x-\mu_2)\cdots(x-\mu_n).$$

Each factor $x - a$ is monic of degree 1, hence irreducible, so both sides are factorizations of p into monic irreducibles. By the uniqueness in Theorem 6.3.5 (b), there is a permutation π of $\{1, \dots, n\}$ with $\nu_{\pi(i)} = \mu_i$ for every i.

The contradiction. For each i, $\nu_{\pi(i)}^{(k_l)} \to \nu_{\pi(i)} = \mu_i$, so there is L_i with $|\nu_{\pi(i)}^{(k_l)} - \mu_i| < \varepsilon$ for all $l \geq L_i$. Take $l \geq \max_i L_i$. Then the list $\nu_{\pi(1)}^{(k_l)}, \dots, \nu_{\pi(n)}^{(k_l)}$ is a numbering of the roots of q_{k_l}, with multiplicity, within ε of $\mu_1, \dots, \mu_n$. This contradicts the choice of q_{k_l}, and proves the theorem.

■

Chapter 16 quoted this statement as Theorem 16.7.3 and left its proof to this section. The proof needed nothing beyond compactness and unique factorization, and Corollary 16.7.4, with everything Chapter 16 built on it, now rests on a proved theorem.

Each hypothesis of the theorem did a job in the proof. **Monic of the same degree** gave both

the bound on the roots and the factorization of q_k with exactly n linear factors. **Numbered so that** is where the permutation π went. It was produced by unique factorization and depends on the subsequence, which is exactly the freedom that Proposition 16.7.2 says cannot be removed.

Remark. The proof is an instance of a move worth naming: **compactness plus uniqueness**. To show that x depends continuously on y when only the inverse relation $y = F(x)$ is explicit, take a sequence, extract a convergent subsequence by compactness, and use the uniqueness of the solution of $y = F(x)$ to identify its limit. Here F sends a root vector to the coefficients, and "unique" means unique up to order.

20.3.3 How far: Elsner's bound

The theorem says that a small enough perturbation moves the roots a little, and it gives no rate. Chapter 16's example Example (A Jordan block's eigenvalues move like $\varepsilon^{1/k}$) showed that no bound of the form $L\|\mathbf{E}\|$ can hold. What *does* hold, for every matrix, is a bound by the n-th root of $\|\mathbf{E}\|$. The proof bounds one determinant in two ways: from below by the eigenvalues, and from above by Hadamard's inequality.

Theorem 20.3.3. Elsner's Spectral Variation Bound

Let $\mathbf{A}, \mathbf{B} \in M_n(\mathbb{C})$, and let $\lambda_1, \dots, \lambda_n$ be the eigenvalues of $\mathbf{A}$. Then every eigenvalue μ of $\mathbf{B}$ satisfies

$$\min_{1 \le i \le n} |\lambda_i - \mu| \le \left(\|\mathbf{A}\|_2 + \|\mathbf{B}\|_2\right)^{1-1/n} \|\mathbf{A} - \mathbf{B}\|_2^{1/n}.$$

(For $n = 1$ the first factor is 1.)

Idea. The quantity $\prod_i |\lambda_i - \mu|$ is $|\det(\mathbf{A} - \mu\mathbf{I})|$, and it is at least the n-th power of the smallest factor. Bound the determinant above by Hadamard's inequality, after rotating to an orthonormal basis whose first vector is an eigenvector $\mathbf{x}$ of $\mathbf{B}$. The first column is then $(\mathbf{A} - \mu\mathbf{I})\mathbf{x} = (\mathbf{A} - \mathbf{B})\mathbf{x}$, which is small. The other $n - 1$ columns are merely bounded. So one small factor meets $n - 1$ bounded ones, and taking n-th roots produces the exponent $1/n$.

Proof. Let $\mathbf{x}$ be an eigenvector of $\mathbf{B}$ for μ with $\|\mathbf{x}\|_2 = 1$. By Corollary 11.2.3, extend $(\mathbf{x})$ to an orthonormal basis $(\mathbf{u}_1, \dots, \mathbf{u}_n)$ of $\mathbb{C}^n$ with $\mathbf{u}_1 = \mathbf{x}$, and let $\mathbf{U}$ be the unitary matrix with these columns. Then $\mathbf{U}^*\mathbf{U} = \mathbf{I}$, so $|\det \mathbf{U}|^2 = \det \mathbf{U}^* \det \mathbf{U} = 1$ by Theorem 7.5.4 and Theorem 7.4.5.

By Corollary 12.1.5 applied to $\mathbf{A} - \mu\mathbf{I}$, whose eigenvalues are the $\lambda_i - \mu$, listed with the same multiplicities (Theorem 9.10.1, with $q(x) = x - \mu$), and by Theorem 7.5.4,

$$\prod_{i=1}^{n} |\lambda_i - \mu| = |\det(\mathbf{A} - \mu\mathbf{I})| = |\det((\mathbf{A} - \mu\mathbf{I})\mathbf{U})|.$$

The j-th column of $(\mathbf{A} - \mu\mathbf{I})\mathbf{U}$ is $(\mathbf{A} - \mu\mathbf{I})\mathbf{u}_j$. Hadamard's inequality Theorem 13.7.3 (b) bounds the determinant by the product of the column lengths.

For $j = 1$, $(\mathbf{A} - \mu\mathbf{I})\mathbf{x} = \mathbf{A}\mathbf{x} - \mathbf{B}\mathbf{x} = (\mathbf{A} - \mathbf{B})\mathbf{x}$, so $\|(\mathbf{A} - \mu\mathbf{I})\mathbf{u}_1\|_2 \le \|\mathbf{A} - \mathbf{B}\|_2$ by Theorem 16.3.2 (a). For $j \ge 2$, the triangle inequality (Corollary 11.1.6) gives $\|(\mathbf{A} - \mu\mathbf{I})\mathbf{u}_j\|_2 \le \|\mathbf{A}\|_2 + |\mu|$,

and $|\mu| \le \rho(\mathbf{B}) \le \|\mathbf{B}\|_2$ by Theorem 16.4.1. Writing $\delta = \min_i |\lambda_i - \mu|$, each factor on the left is at least δ, so

$$\delta^n \le \prod_{i=1}^{n} |\lambda_i - \mu| \le \|\mathbf{A} - \mathbf{B}\|_2 \left(\|\mathbf{A}\|_2 + \|\mathbf{B}\|_2\right)^{n-1}.$$

Taking nonnegative n-th roots, which preserves inequalities between nonnegative reals, gives the claim.

■

The bound needs no hypothesis on $\mathbf{A}$: no diagonalizability, no normality. It is a **Hölder modulus of continuity with exponent** $1/n$. The next example shows that the exponent cannot be improved, and that is the quantitative form of the instability that Chapter 10 §03 warned about.

Example. The Jordan block meets Elsner's bound

Let $n \ge 2$ and $0 < \varepsilon \le 1$, and put $\mathbf{A} = \mathbf{J}_n(0)$ and $\mathbf{B} = \mathbf{J}_n(0) + \varepsilon\, \mathbf{e}_n \mathbf{e}_1^{\top}$. Compare the movement of the eigenvalues with Elsner's bound, and show that no bound of the form $C(\|\mathbf{A}\|_2 + \|\mathbf{B}\|_2)^{1-\alpha}\|\mathbf{A} - \mathbf{B}\|_2^{\alpha}$ with $\alpha > 1/n$ can hold for all $\mathbf{A}, \mathbf{B} \in M_n(\mathbb{C})$.

Solution. The eigenvalues. The matrix $\mathbf{B}$ is the one of Example (A Jordan block's eigenvalues move like $\varepsilon^{1/k}$) with $k = n$, where $p_{\mathbf{B}}(x) = x^n - \varepsilon$ was computed. So the eigenvalues of $\mathbf{B}$ are the n complex n-th roots of ε, each of modulus $\varepsilon^{1/n}$, while $\mathbf{A}$ has the single eigenvalue 0. Every eigenvalue has moved exactly $\varepsilon^{1/n}$.

The norms. $\mathbf{A} - \mathbf{B} = -\varepsilon\, \mathbf{e}_n \mathbf{e}_1^{\top}$ has the single nonzero entry $-\varepsilon$, so $\|\mathbf{A} - \mathbf{B}\|_2 = \varepsilon$ (Theorem 16.3.3 (c)). Next, $\mathbf{B}\mathbf{e}_1 = \varepsilon \mathbf{e}_n$ and $\mathbf{B}\mathbf{e}_j = \mathbf{e}_{j-1}$ for $j \ge 2$, so for $\mathbf{x} = (x_1, \dots, x_n)$

$$\|\mathbf{B}\mathbf{x}\|_2^2 = \varepsilon^2 |x_1|^2 + \sum_{j=2}^{n} |x_j|^2 \le \|\mathbf{x}\|_2^2,$$

with equality at $\mathbf{x} = \mathbf{e}_2$. Hence $\|\mathbf{B}\|_2 = 1$, and the same computation with $\varepsilon = 0$ gives $\|\mathbf{A}\|_2 = 1$.

The comparison. Elsner's bound is $2^{1-1/n}\varepsilon^{1/n}$, and the actual movement is $\varepsilon^{1/n}$. The bound is attained up to a factor $2^{1-1/n} < 2$.

The exponent is optimal. If a bound $C\, 2^{1-\alpha}\varepsilon^{\alpha}$ held for these pairs with $\alpha > 1/n$, then $\varepsilon^{1/n} \le C\, 2^{1-\alpha}\varepsilon^{\alpha}$, that is, $\varepsilon^{1/n-\alpha} \le 2^{1-\alpha} C$, for every $\varepsilon \in (0, 1]$. But $1/n - \alpha < 0$, so the left side tends to ∞ as $\varepsilon \to 0^+$. No such C exists.

Put numbers on it. For $n = 16$ and $\varepsilon = 10^{-16}$, a change in the sixteenth decimal place of one entry, the eigenvalues of $\mathbf{J}_{16}(0)$ move to the circle of radius $(10^{-16})^{1/16} = 10^{-1}$. This pays the debt of Chapter 10 §03, whose warning said that the Jordan form is not stable under small changes of the matrix and that Chapter 20 returns to this point. Elsner's theorem says how unstable the eigenvalues can be: never worse than $\|\mathbf{E}\|^{1/n}$. The Jordan block says that this worst case happens. Arithmetic that keeps about sixteen significant digits commits errors of about this relative size at every step. So what such a computation returns are, at best, the eigenvalues of some matrix within about 10^{-16} of $\mathbf{J}_{16}(0)$, and these may be spread around a circle of radius 0.1. The computation cannot tell a nilpotent Jordan block from a diagonalizable

matrix with sixteen distinct eigenvalues. This is the numerical reason the Jordan form is not computed in floating-point arithmetic.

20.3.4 Counting eigenvalues in a region

Section 1 showed that every eigenvalue lies in the union of the Gershgorin discs (Theorem 20.1.1), and left open how many lie in each part of that union. The tool for counting is a principle in the spirit of one the proof of Proposition 16.7.2 already used: **an integer-valued function that varies continuously cannot change**. We apply it to the number of eigenvalues in a region, while the matrix moves along a segment.

Lemma 20.3.4. Eigenvalue Counts Are Constant Along a Path

Let $\mathbf{A}_0, \mathbf{A}_1 \in M_n(\mathbb{C})$, and put $\mathbf{A}(t) = (1-t)\mathbf{A}_0 + t\mathbf{A}_1$ for $t \in [0,1]$. Let $K, L \subseteq \mathbb{C}$ be sets at **positive distance**: there is $d > 0$ with $|z - w| \geq d$ for all $z \in K$ and $w \in L$. Suppose that for **every** $t \in [0,1]$, every eigenvalue of $\mathbf{A}(t)$ lies in $K \cup L$. Then the number $N(t)$ of eigenvalues of $\mathbf{A}(t)$ in K, counted with algebraic multiplicity, does not depend on t. In particular $N(0) = N(1)$.

Idea. The eigenvalues cannot jump across the gap between K and L, because they move continuously and are never allowed inside the gap. So near each t_0 the count is unchanged, and a function that is locally constant on an interval is constant. The last step is the intermediate value theorem: a continuous function with integer values cannot pass from one integer to another without taking the values in between.

Proof. Fix $t_0 \in [0,1]$. By Corollary 16.7.4, applied to $\mathbf{A}(t_0)$ with the tolerance d and the norm $\|\cdot\|_2$, there is $\delta > 0$ such that every $\mathbf{B}$ with $\|\mathbf{B} - \mathbf{A}(t_0)\|_2 < \delta$ has its eigenvalues μ_i numbered so that $|\mu_i - \lambda_i| < d$, where $\lambda_1, \dots, \lambda_n$ are the eigenvalues of $\mathbf{A}(t_0)$. Now $\mathbf{A}(t) - \mathbf{A}(t_0) = (t - t_0)(\mathbf{A}_1 - \mathbf{A}_0)$, so $\|\mathbf{A}(t) - \mathbf{A}(t_0)\|_2 < \delta$ whenever $|t - t_0|\,\|\mathbf{A}_1 - \mathbf{A}_0\|_2 < \delta$, which holds for all t close enough to t_0.

For such t, number the eigenvalues μ_i of $\mathbf{A}(t)$ in this way. Both λ_i and μ_i lie in $K \cup L$, and a point of K is at distance at least d from every point of L. Since $|\mu_i - \lambda_i| < d$, the two lie in the same one of the two sets: $\mu_i \in K$ if and only if $\lambda_i \in K$. Counting indices, $N(t) = N(t_0)$.

So N is constant on a neighborhood of each point of $[0,1]$. In particular N is continuous: if $t_k \to t_0$, then $N(t_k) = N(t_0)$ for all large k. Suppose $N(s) \neq N(0)$ for some $s \in (0,1]$, say $N(s) > N(0)$; the other case is symmetric. By the intermediate value theorem, fact (A5), applied to N on $[0,s]$, there is t with $N(t) = N(0) + \frac{1}{2}$, which is not an integer. This contradiction shows $N(s) = N(0)$ for every $s \in [0,1]$, and proves the lemma. ∎

Now the counting theorem. The trick is to shrink the off-diagonal part of $\mathbf{A}$ to zero. At the start of the path the matrix is diagonal and its eigenvalues are the disc centers, which we can count by looking.

Theorem 20.3.5. Gershgorin's Counting Theorem

Let $\mathbf{A} \in M_n(\mathbb{C})$, let $S \subseteq \{1, \dots, n\}$ have k elements with $1 \le k \le n-1$, and put

$$U = \bigcup_{i \in S} D_i(\mathbf{A}), \qquad V = \bigcup_{i \notin S} D_i(\mathbf{A}).$$

If $U \cap V = \emptyset$, then U contains exactly k eigenvalues of $\mathbf{A}$ and V contains exactly $n-k$, counted with algebraic multiplicity.

Idea. Let $\mathbf{D} = \operatorname{diag}(a_{11}, \dots, a_{nn})$ and slide from $\mathbf{D}$ to $\mathbf{A}$ along $\mathbf{A}(t) = (1-t)\mathbf{D} + t\mathbf{A}$. The discs of $\mathbf{A}(t)$ have the same centers and radii $t\, r_i(\mathbf{A})$, so they grow from points into the discs of $\mathbf{A}$ and never leave $U \cup V$. By Theorem 20.1.1 the eigenvalues never leave either, and Lemma 20.3.4 carries the count from $t = 0$, where it is k, to $t = 1$.

Proof. The two unions are at positive distance. Let $i \in S$ and $j \notin S$. The discs $D_i(\mathbf{A})$ and $D_j(\mathbf{A})$ are disjoint, so $|a_{ii} - a_{jj}| > r_i(\mathbf{A}) + r_j(\mathbf{A})$; otherwise the point on the segment from a_{ii} to a_{jj} at distance $\min\{r_i(\mathbf{A}), |a_{ii} - a_{jj}|\}$ from a_{ii} would lie in both. For $z \in D_i(\mathbf{A})$ and $w \in D_j(\mathbf{A})$, the triangle inequality (Theorem 0.6.4) gives

$$|z - w| \ge |a_{ii} - a_{jj}| - r_i(\mathbf{A}) - r_j(\mathbf{A}) > 0.$$

Call the middle quantity d_{ij}. The minimum d of the finitely many d_{ij} is positive, and every point of U is at distance at least d from every point of V.

The path. Let $\mathbf{D} = \operatorname{diag}(a_{11}, \dots, a_{nn})$ and $\mathbf{A}(t) = (1-t)\mathbf{D} + t\mathbf{A}$ for $t \in [0, 1]$. The diagonal entries of $\mathbf{A}(t)$ are the a_{ii}, and its off-diagonal entries are $t\, a_{ij}$, so $r_i(\mathbf{A}(t)) = t\, r_i(\mathbf{A}) \le r_i(\mathbf{A})$ and $D_i(\mathbf{A}(t)) \subseteq D_i(\mathbf{A})$. By Theorem 20.1.1 applied to $\mathbf{A}(t)$, every eigenvalue of $\mathbf{A}(t)$ lies in $\bigcup_i D_i(\mathbf{A}(t)) \subseteq U \cup V$. By Lemma 20.3.4 with $K = U$ and $L = V$, the number of eigenvalues of $\mathbf{A}(t)$ in U is the same at $t = 0$ and at $t = 1$.

The count at $t = 0$. $\mathbf{A}(0) = \mathbf{D}$ is diagonal, so $p_{\mathbf{D}}(x) = \prod_i (x - a_{ii})$ by Theorem 7.4.6, and its eigenvalue list is $a_{11}, \dots, a_{nn}$. For $i \in S$, a_{ii} is the center of $D_i(\mathbf{A})$, so $a_{ii} \in U$. For $i \notin S$, $a_{ii} \in V$, hence $a_{ii} \notin U$. So exactly k entries of the list lie in U.

Therefore $\mathbf{A}(1) = \mathbf{A}$ has exactly k eigenvalues in U. All n eigenvalues lie in $U \cup V$, and $U \cap V = \emptyset$, so the other $n-k$ lie in V. This proves the theorem.

■

The case $k = 1$ is the one used most, and for a real matrix it says something about reality as well.

Corollary 20.3.6. An Isolated Disc Holds One Eigenvalue

Let $\mathbf{A} \in M_n(\mathbb{C})$ with $n \ge 2$, and suppose $D_i(\mathbf{A})$ is disjoint from $D_j(\mathbf{A})$ for every $j \ne i$.

(a) $D_i(\mathbf{A})$ contains exactly one eigenvalue of $\mathbf{A}$, and it is **algebraically simple**.

(b) If $\mathbf{A}$ is **real**, that eigenvalue is real, and it lies in the interval $[a_{ii} - r_i(\mathbf{A}), a_{ii} + r_i(\mathbf{A})]$.

Proof.

(a) Take $S = \{i\}$. Then $U = D_i(\mathbf{A})$ is disjoint from the union V of the other discs, because it is disjoint from each of them. By Theorem 20.3.5, $D_i(\mathbf{A})$ contains exactly one eigenvalue counted with algebraic multiplicity, so that eigenvalue λ has algebraic multiplicity 1.

(b) Suppose $\lambda \notin \mathbb{R}$. By Theorem 9.2.5, $\overline{\lambda} \neq \lambda$ is also an eigenvalue of $\mathbf{A}$. The center a_{ii} is real, so $|\overline{\lambda} - a_{ii}| = |\overline{\lambda - a_{ii}}| = |\lambda - a_{ii}| \leq r_i(\mathbf{A})$, and $\overline{\lambda} \in D_i(\mathbf{A})$ too. Then $D_i(\mathbf{A})$ contains two distinct eigenvalues, contradicting (a). So λ is real, and a real number within $r_i(\mathbf{A})$ of a_{ii} lies in the stated interval.

■

Scaling (Proposition 20.1.3) and counting work together. Scaling moves the radii around without changing the eigenvalues, and a disc that scaling manages to isolate then holds exactly one eigenvalue.

Example. Isolating a real eigenvalue by scaling

Let

$$\mathbf{A} = \begin{pmatrix} 7 & 3 & -1 \\ 0 & -1 & -3 \\ 1 & 1 & 1 \end{pmatrix}.$$

Show that $\mathbf{A}$ has exactly one eigenvalue λ with $|\lambda - 7| \leq 2$, and that it is real.

Solution. The unscaled discs do not decide it. The row radii are $r_1 = 4$, $r_2 = 3$, $r_3 = 2$, so the discs are $|z - 7| \leq 4$, $|z + 1| \leq 3$, and $|z - 1| \leq 2$. The first and the third meet at $z = 3$, since $7 - 1 = 4 + 2$. So no disc is isolated, and Theorem 20.3.5 does not apply.

Scale. Take $\mathbf{D} = \operatorname{diag}(2, 1, 1)$. By Proposition 20.1.3, $\mathbf{D}^{-1}\mathbf{A}\mathbf{D}$ has the same eigenvalues as $\mathbf{A}$, and its (i, j) entry is $a_{ij}d_j/d_i$:

$$\mathbf{D}^{-1}\mathbf{A}\mathbf{D} = \begin{pmatrix} 7 & 3/2 & -1/2 \\ 0 & -1 & -3 \\ 2 & 1 & 1 \end{pmatrix}.$$

Its discs are $|z - 7| \leq 2$, $|z + 1| \leq 3$, and $|z - 1| \leq 3$. The first is disjoint from the other two, because $7 - (-1) = 8 > 2 + 3$ and $7 - 1 = 6 > 2 + 3$.

Count. By Corollary 20.3.6, applied to the real matrix $\mathbf{D}^{-1}\mathbf{A}\mathbf{D}$, the disc $|z - 7| \leq 2$ contains exactly one eigenvalue, it is simple and real, and it lies in $[5, 9]$. The other two eigenvalues lie in $|z + 1| \leq 3$ or $|z - 1| \leq 3$, outside $[5, 9]$.

A check by hand. Expanding along the first column, $p_{\mathbf{A}}(x) = x^3 - 7x^2 + 3x - 4$. Then $p_{\mathbf{A}}(5) = -39 < 0$ and $p_{\mathbf{A}}(9) = 185 > 0$, consistent with a real root in $(5, 9)$. (It is about 6.64. The other two roots are the nonreal pair $0.18 \pm 0.75i$, approximately, so here Gershgorin has isolated the only real eigenvalue.)

Quick check. A real 3×3 matrix has three pairwise disjoint Gershgorin discs. What can you say about its eigenvalues, and is it diagonalizable?

Solution. By Corollary 20.3.6, each disc contains exactly one eigenvalue, which is real. The discs are disjoint, so the three eigenvalues are distinct and real, and the matrix is diagonalizable over $\mathbb{R}$ by Corollary 9.4.2.

Warning. **The count is for the union, not for each disc in it.** Take $\mathbf{A} = \left(\begin{smallmatrix} 0 & 1 & 0 \\ 4 & 0 & 0 \\ 0 & 0 & 10 \end{smallmatrix}\right)$. Its discs are $|z| \le 1$, $|z| \le 4$, and the single point $\{10\}$. The union of the first two is disjoint from the third, and it does contain two eigenvalues, ± 2. But both lie in the second disc, and the first disc contains none. Inside a cluster of overlapping discs, the counting theorem says nothing about the individual discs.

20.3.5 From one-sided to matched

Elsner's bound is **one-sided**. It says that every eigenvalue of $\mathbf{B}$ is near **some** eigenvalue of $\mathbf{A}$, and it does not forbid two eigenvalues of $\mathbf{B}$ from crowding near the same eigenvalue of $\mathbf{A}$ while another eigenvalue of $\mathbf{A}$ is left alone. The counting lemma turns it into a statement about a **numbering**. The price is a factor that grows with n.

Theorem 20.3.7. Elsner's Matching Bound

Let $\mathbf{A}, \mathbf{B} \in M_n(\mathbb{C})$ have eigenvalues $\lambda_1, \dots, \lambda_n$ and $\mu_1, \dots, \mu_n$, and put

$$R = \left(\|\mathbf{A}\|_2 + \|\mathbf{B}\|_2\right)^{1-1/n} \|\mathbf{A} - \mathbf{B}\|_2^{1/n}.$$

Then the μ_i can be numbered so that $|\lambda_i - \mu_i| \le (2n-1)R$ for every i.

Idea. Slide from $\mathbf{A}$ to $\mathbf{B}$ along a segment. By Theorem 20.3.3, every matrix on the way has its eigenvalues in the discs of radius R about the λ_i. Group those discs into clusters that do not touch one another. By Lemma 20.3.4 a cluster built from m of the λ_i holds exactly m eigenvalues of $\mathbf{B}$, so we can pair within clusters. A cluster of m discs has diameter at most $2mR$, which gives the factor $2n-1$.

Proof. The conclusion is symmetric in $\mathbf{A}$ and $\mathbf{B}$, and so is R, so we may assume $\|\mathbf{A}\|_2 \le \|\mathbf{B}\|_2$. For $t \in [0,1]$ put $\mathbf{A}(t) = (1-t)\mathbf{A} + t\mathbf{B}$. Then $\|\mathbf{A}(t)\|_2 \le (1-t)\|\mathbf{A}\|_2 + t\|\mathbf{B}\|_2 \le \|\mathbf{B}\|_2$ and $\|\mathbf{A} - \mathbf{A}(t)\|_2 = t\|\mathbf{A} - \mathbf{B}\|_2 \le \|\mathbf{A} - \mathbf{B}\|_2$. By Theorem 20.3.3 applied to $\mathbf{A}$ and $\mathbf{A}(t)$, and since the exponents $1 - 1/n$ and $1/n$ are nonnegative, every eigenvalue of $\mathbf{A}(t)$ lies within R of some λ_i.

Clusters. Call indices i and j **linked** if $|\lambda_i - \lambda_j| \le 2R$, and put two indices in the same **cluster** if a chain of links joins them. The clusters partition $\{1, \dots, n\}$. For a cluster C let K_C be the set of z with $|z - \lambda_i| \le R$ for some $i \in C$, and let L_C be the union of the $K_{C'}$ for

the other clusters C'. If $i \in C$ and $j \notin C$, then i and j are not linked, so for z within R of λ_i and w within R of λ_j,

$$|z - w| \geq |\lambda_i - \lambda_j| - 2R > 0.$$

Taking the minimum over the finitely many such pairs, K_C and L_C are at positive distance; and if C is the only cluster, L_C is empty, there is no pair to check, and the requirement holds with any $d > 0$. Every eigenvalue of every $\mathbf{A}(t)$ lies in $K_C \cup L_C$.

Counting. At $t = 0$ the eigenvalues are the λ_j. Those with $j \in C$ lie in K_C, and the others lie in L_C, hence not in K_C; so $\mathbf{A}(0)$ has exactly $|C|$ eigenvalues in K_C. By Lemma 20.3.4, so does $\mathbf{A}(1) = \mathbf{B}$. Each μ lies in exactly one K_C, since the K_C are pairwise disjoint and cover every eigenvalue. So we may number the μ's so that, for each cluster C, those in K_C receive the indices in C.

Distances. Let $i \in C$, with $m = |C|$. Since $\mu_i \in K_C$, there is $j \in C$ with $|\mu_i - \lambda_j| \leq R$. A chain of links from j to i inside C can be chosen without repeating an index, so it has at most $m - 1$ links, each of length at most $2R$, and $|\lambda_j - \lambda_i| \leq 2(m-1)R$. Hence

$$|\mu_i - \lambda_i| \leq R + 2(m-1)R = (2m-1)R \leq (2n-1)R.$$

This proves the theorem.

■

The factor $2n-1$ is not the best possible, and bounds with a factor independent of n are known; we will not need them. What matters is the shape: a matching distance of order $\|\mathbf{A} - \mathbf{B}\|^{1/n}$, with no hypothesis on either matrix. Sections 4 and 5 replace the exponent $1/n$ by 1, and the price is a hypothesis on $\mathbf{A}$: diagonalizable in Section 4, and normal in Section 5.

20.3.6 Exercises

20.3.6.1 A. Check your understanding

Exercise. A1

(a) State Theorem 20.3.2, including both hypotheses on q.

(b) Which imported fact of analysis does the proof of Theorem 20.3.2 use, and at which step? Which one does Lemma 20.3.4 use?

(c) True or false: if $\|\mathbf{B} - \mathbf{A}\|_2 \leq 10^{-8}$, then every eigenvalue of $\mathbf{B}$ lies within 10^{-8} of an eigenvalue of $\mathbf{A}$. Justify your answer.

(d) True or false: if $\mathbf{A} \in M_3(\mathbb{C})$ and $D_1(\mathbf{A}) \cup D_2(\mathbf{A})$ is disjoint from $D_3(\mathbf{A})$, then $D_1(\mathbf{A})$ contains exactly one eigenvalue. Justify your answer.

(e) What is the difference between the conclusions of Theorem 20.3.3 and Theorem 20.3.7?

Solution.

(a) See Theorem 20.3.2. The polynomial q must be **monic** and of the **same degree** n as p.

(b) Compactness, fact (A3), supplies a convergent subsequence of the bounded root vectors $\boldsymbol{\nu}^{(k)}$. The lemma uses the intermediate value theorem, fact (A5), to show that a locally constant integer-valued function on $[0, 1]$ is constant.

(c) False. Take $\mathbf{A} = \mathbf{J}_2(0)$ and $\mathbf{B} = \mathbf{J}_2(0) + 10^{-8}\mathbf{e}_2\mathbf{e}_1^\top$, so $\|\mathbf{B} - \mathbf{A}\|_2 = 10^{-8}$. By Example (The Jordan block meets Elsner's bound) with $n = 2$, the eigenvalues of $\mathbf{B}$ are $\pm 10^{-4}$, at distance 10^{-4} from the only eigenvalue 0 of $\mathbf{A}$.

(d) False. In the warning after Corollary 20.3.6, $D_1 \cup D_2$ is disjoint from $D_3 = \{10\}$, yet $D_1 = \{|z| \le 1\}$ contains neither eigenvalue ± 2. The counting theorem counts in the union.

(e) The first is one-sided: every eigenvalue of $\mathbf{B}$ is within R of **some** eigenvalue of $\mathbf{A}$, and several may be near the same one. The second gives a **numbering** pairing each μ_i with its own λ_i, within $(2n-1)R$.

20.3.6.2 B. Practice

Exercise. B1: Cauchy's bound

Let $p(x) = x^3 - 3x^2 + 4x - 12$.

(a) Use Lemma 20.3.1 to bound the moduli of the roots of p.

(b) Find the roots exactly, and compare.

Solution.

(a) Here $M = \max\{3, 4, 12\} = 12$, so every root has $|z| \le 13$.

(b) $p(x) = x^2(x-3) + 4(x-3) = (x-3)(x^2+4)$, so the roots are 3 and $\pm 2i$, of moduli 3 and 2. The bound 13 is true but crude, because the constant term 12 dominates M.

Exercise. B2: Counting with discs

Let

$$\mathbf{A} = \begin{pmatrix} 10 & 1 & -1 \\ 1 & 2 & 2 \\ 0 & -2 & 1 \end{pmatrix}.$$

(a) Show that $\mathbf{A}$ has exactly one eigenvalue λ_1 with $|\lambda_1 - 10| \le 2$, and that it is real.

(b) Explain why the Gershgorin discs cannot decide whether the other two eigenvalues are real.

(c) Use the trace to locate the sum of the other two eigenvalues in an interval.

Solution.

(a) The row radii are $r_1 = 2$, $r_2 = 3$, $r_3 = 2$, so the discs are $|z - 10| \le 2$, $|z - 2| \le 3$, and $|z - 1| \le 2$. The first is disjoint from the other two, since $10 - 2 = 8 > 2 + 3$ and $10 - 1 = 9 > 2 + 2$. By Corollary 20.3.6, it contains exactly one eigenvalue, which is simple, real, and in $[8, 12]$.

(b) The other two eigenvalues lie in $\{|z-2| \le 3\} \cup \{|z-1| \le 2\}$, a union of two overlapping discs that contains both real and nonreal points. The counting theorem only says that this union holds two eigenvalues. A conjugate pair and two real numbers are both consistent with that. (In fact they are the nonreal pair $1.43 \pm 1.99i$, approximately.)

(c) By Corollary 12.1.4, $\lambda_1+\lambda_2+\lambda_3=\operatorname{tr}\mathbf{A}=13$. Since $8\le\lambda_1\le 12$, the sum $\lambda_2+\lambda_3=13-\lambda_1$ is real and lies in $[1,5]$.

20.3.6.3 C. Going deeper

Exercise. C1: Stability is an open condition

(a) Let p be monic of degree n with every root in the open disc $\{|z|<1\}$. Prove that there is $\delta>0$ such that every monic q of degree n whose coefficients are within δ of those of p also has every root in $\{|z|<1\}$.

(b) Deduce that the set of $\mathbf{A}\in M_n(\mathbb{C})$ with $\rho(\mathbf{A})<1$ is **open**: for each such $\mathbf{A}$ there is $\eta>0$ with $\rho(\mathbf{B})<1$ whenever $\|\mathbf{B}-\mathbf{A}\|_2<\eta$.

Solution.

(a) Let $\mu_1,\dots,\mu_n$ be the roots of p and put $\varepsilon=1-\max_i|\mu_i|$, which is positive since there are finitely many roots, each of modulus less than 1. Let δ be given by Theorem 20.3.2 for this ε. If q is within δ, number its roots so that $|\nu_i-\mu_i|<\varepsilon$. Then $|\nu_i|<|\mu_i|+\varepsilon\le 1$ for every i.

(b) Let $\rho(\mathbf{A})<1$, so every root of $p_{\mathbf{A}}$ lies in $\{|z|<1\}$ (Definition 16.4.1). Take δ from (a) for $p=p_{\mathbf{A}}$. The coefficients of $p_{\mathbf{B}}$ are polynomials in the entries of $\mathbf{B}$, as noted after Lemma 16.7.1, so that lemma gives $\eta>0$ such that $\|\mathbf{B}-\mathbf{A}\|_2<\eta$ makes every coefficient of $p_{\mathbf{B}}$ within δ of the corresponding coefficient of $p_{\mathbf{A}}$. For such $\mathbf{B}$, (a) puts every eigenvalue of $\mathbf{B}$ in the open unit disc, and $\rho(\mathbf{B})<1$ because there are finitely many of them.

Exercise. C2: Why one-sided bounds do not give a matching

For lists $(\lambda_1,\dots,\lambda_n)$ and $(\mu_1,\dots,\mu_n)$ of complex numbers and $r\ge 0$, say that the list of μ's is **r-close to** the list of λ's if every μ_j is within r of some λ_i.

(a) Let $n=2$. Prove that if each list is r-close to the other, then the μ_j can be numbered so that $|\lambda_i-\mu_i|\le r$ for $i=1,2$.

(b) Show that (a) fails for $n=3$: give two lists of length 3, each 0-close to the other, such that every numbering has $\max_i|\lambda_i-\mu_i|=2$.

(c) Explain why the proof of Theorem 20.3.7 needs Lemma 20.3.4, and cannot simply apply Theorem 20.3.3 twice, once in each direction.

Solution.

(a) By hypothesis, λ_2 is within r of some μ_j; number the μ's so that this one is μ_2, so $|\lambda_2-\mu_2|\le r$. The other one, μ_1, is within r of λ_1 or of λ_2. If of λ_1, we are done. If not, then $|\mu_1-\lambda_2|\le r$ and $|\mu_1-\lambda_1|>r$. Now λ_1 is within r of some μ_j, and not of μ_1, so $|\lambda_1-\mu_2|\le r$. Renumber by swapping μ_1 and μ_2. Then $|\lambda_1-\mu_1|\le r$ and $|\lambda_2-\mu_2|\le r$.

(b) Take $(\lambda_1,\lambda_2,\lambda_3)=(0,2,2)$ and $(\mu_1,\mu_2,\mu_3)=(0,0,2)$. Every entry of each list occurs in the other, so each is 0-close to the other. But the μ's contain 0 twice and the λ's only

once, so every numbering pairs some $\mu_i = 0$ with some $\lambda_i = 2$, and $\max_i|\lambda_i - \mu_i| = 2$.

(c) Applying Theorem 20.3.3 to $(\mathbf{A}, \mathbf{B})$ and to $(\mathbf{B}, \mathbf{A})$ gives exactly the two-sided closeness of (a) with $r = R$, since R is symmetric in $\mathbf{A}$ and $\mathbf{B}$. By (b), two-sided closeness does not produce a numbering once $n \geq 3$, because it ignores multiplicities. The counting lemma supplies the missing information: each cluster of discs holds as many eigenvalues of $\mathbf{B}$, counted with multiplicity, as it holds eigenvalues of $\mathbf{A}$.

Exercise. C3: Sharpening Cauchy's bound by rescaling

Let $n \geq 1$, let $p(x) = x^n + c_{n-1}x^{n-1} + \cdots + c_1 x + c_0 \in \mathbb{C}[x]$ be monic, and let $s > 0$.

(a) Prove that $q(y) := p(sy)/s^n$ is monic of degree n, with coefficients c_j/s^{n-j}, and that z is a root of p if and only if z/s is a root of q. Hence deduce that every root z of p satisfies

$$|z| \leq s\Big(1 + \max_{0 \leq j \leq n-1} \frac{|c_j|}{s^{n-j}}\Big).$$

(b) Apply (a) with $s = 3$ to $p(x) = x^3 - 3x^2 + 4x - 12$, and compare the bound with the one Lemma 20.3.1 gives directly and with the roots found in Exercise (B1: Cauchy's bound).

Solution.

(a) Substituting $x = sy$ and dividing by $s^n > 0$,

$$q(y) = \frac{p(sy)}{s^n} = \frac{s^n y^n + \sum_{j<n} c_j s^j y^j}{s^n} = y^n + \sum_{j<n} \frac{c_j}{s^{n-j}}\, y^j,$$

which is monic of degree n with the stated coefficients. Since $s^n \neq 0$, the identity $q(z/s) = p(z)/s^n$ shows that $q(z/s) = 0$ if and only if $p(z) = 0$. Now let z be a root of p. Then z/s is a root of q, so by Lemma 20.3.1 applied to q, whose coefficient maximum is $M_s = \max_j |c_j|/s^{n-j}$, we get $|z|/s = |z/s| \leq 1 + M_s$. Multiplying by $s > 0$ gives the bound.

(b) For $s = 3$ the coefficients of q are $-3/3 = -1$, $4/9$, and $-12/27 = -4/9$, so $M_3 = 1$ and every root of p has $|z| \leq 3(1 + 1) = 6$. Applied to p itself, Lemma 20.3.1 gives only $|z| \leq 13$, as in Exercise (B1: Cauchy's bound) (a). The true moduli are 3 and 2, so the rescaled bound is closer, and the trade-off is visible in the formula: enlarging s shrinks every $|c_j|/s^{n-j}$ but multiplies the whole bound by s.

20.4 The Bauer–Fike Theorem

Section 3 bounded the movement of the eigenvalues of an arbitrary matrix by a constant times $\|\mathbf{E}\|_2^{1/n}$, and the Jordan block showed that nothing better holds in general. The culprit was the Jordan block itself. This section removes it by assuming that $\mathbf{A}$ is diagonalizable, and the exponent $1/n$ becomes 1: every eigenvalue of $\mathbf{A} + \mathbf{E}$ lies within $\kappa_2(\mathbf{X})\|\mathbf{E}\|_2$ of some eigenvalue of $\mathbf{A}$, where $\mathbf{X}$ is a matrix of eigenvectors.

This is the theorem several earlier chapters pointed to. Chapter 12 §11 described it as the

statement "that for a diagonalizable $\mathbf{A}$ every eigenvalue of a perturbed $\mathbf{A}+\mathbf{E}$ lies within $\kappa\|\mathbf{E}\|$ of some eigenvalue of $\mathbf{A}$, where κ measures how far from unitary the diagonalizing matrix is, and that $\kappa = 1$ exactly in the normal case", and said that it "needs matrix norms and a little analysis, and Chapter 20 develops both and proves it". Chapter 16 §07 and Chapter 17 §03 promised a bound "for a diagonalizable matrix" carrying "an extra factor that measures how far the eigenvector basis of $\mathbf{A}$ is from orthonormal". Chapter 17 §11 left the bounds for non-Hermitian matrices, "beginning with Bauer–Fike", to this chapter, and Chapter 16 §08 recorded that "the bounds that do exist, Weyl's and Bauer–Fike's, are Chapters 17 and 20". We prove the theorem, identify the factor, prove the "exactly", and then return to the defective case. There the m-th root comes back, where m is the size of the largest Jordan block.

Throughout, $\mathbf{A} \in M_n(\mathbb{C})$, the perturbation is $\mathbf{E} \in M_n(\mathbb{C})$, and $\kappa_p(\mathbf{X}) = \|\mathbf{X}\|_p\|\mathbf{X}^{-1}\|_p$ is the condition number of Definition 16.8.1. No fact of analysis is imported in this section beyond those already used by the results it cites.

20.4.1 Factoring out the inverse

The whole section runs on one estimate. If μ is an eigenvalue of $\mathbf{A}+\mathbf{E}$ but not of $\mathbf{A}$, then $\mathbf{A}-\mu\mathbf{I}$ is invertible while $\mathbf{A}+\mathbf{E}-\mu\mathbf{I}$ is not. A small $\mathbf{E}$ cannot destroy invertibility (Corollary 16.5.2), so $\mathbf{E}$ must be large compared with the inverse of $\mathbf{A}-\mu\mathbf{I}$. The matrix $(\mathbf{A}-\mu\mathbf{I})^{-1}$, a function of μ, is called the **resolvent** of $\mathbf{A}$, up to a sign convention. Chapter 16 asked for this estimate in Exercise (C1: How close to singular a nearby eigenvalue forces you to be). Since the theorems below rest on it, we prove it here in full.

Lemma 20.4.1. A Perturbed Eigenvalue Forces a Large Resolvent

Let $\mathbf{A}, \mathbf{E} \in M_n(\mathbb{C})$, let $\|\cdot\|$ be an operator norm on $M_n(\mathbb{C})$ (Definition 16.3.2), and let μ be an eigenvalue of $\mathbf{A}+\mathbf{E}$ that is **not** an eigenvalue of $\mathbf{A}$. Then

$$1 \le \|(\mathbf{A}-\mu\mathbf{I})^{-1}\mathbf{E}\| \le \|(\mathbf{A}-\mu\mathbf{I})^{-1}\|\,\|\mathbf{E}\|.$$

Proof. Since $\mu \notin \operatorname{spec}(\mathbf{A})$, the matrix $\mathbf{A}-\mu\mathbf{I}$ is invertible (Theorem 9.2.1), and

$$\mathbf{A}+\mathbf{E}-\mu\mathbf{I} = (\mathbf{A}-\mu\mathbf{I})(\mathbf{I}+(\mathbf{A}-\mu\mathbf{I})^{-1}\mathbf{E}).$$

The left side is singular, because μ is an eigenvalue of $\mathbf{A}+\mathbf{E}$. If the second factor on the right were invertible, the product of two invertible matrices would be invertible. So $\mathbf{I}-\mathbf{F}$ is singular, where $\mathbf{F} = -(\mathbf{A}-\mu\mathbf{I})^{-1}\mathbf{E}$. By Theorem 16.5.1 (a), $\|\mathbf{F}\| < 1$ would make $\mathbf{I}-\mathbf{F}$ invertible, hence $\|\mathbf{F}\| \ge 1$. Since $\|\mathbf{F}\| = \|(\mathbf{A}-\mu\mathbf{I})^{-1}\mathbf{E}\|$, this is the first inequality, and the second is submultiplicativity (Theorem 16.3.2 (d)).

■

The lemma holds for every matrix and needs no eigenvectors. To turn it into a bound on $|\lambda_i - \mu|$, we need to compute $\|(\mathbf{A}-\mu\mathbf{I})^{-1}\|$ in terms of the distances from μ to the eigenvalues. For a diagonal matrix that is immediate, and diagonalizable matrices are diagonal in the right coordinates.

20.4.2 The theorem

We need one fact about diagonal matrices. Let $\mathbf{D} = \operatorname{diag}(d_1, \dots, d_n)$ and $p \in \{1, 2, \infty\}$. For every $\mathbf{x}$, each coordinate of $\mathbf{Dx}$ is $d_i x_i$, with $|d_i x_i| \le (\max_j |d_j|)|x_i|$, so $\|\mathbf{Dx}\|_p \le (\max_j |d_j|)\|\mathbf{x}\|_p$. Equality holds at $\mathbf{x} = \mathbf{e}_k$, where k is an index with $|d_k| = \max_j |d_j|$. Hence

$$\|\operatorname{diag}(d_1, \dots, d_n)\|_p = \max_j |d_j| \qquad (p = 1, 2, \infty). \tag{20.4.1}$$

Theorem 20.4.2. Bauer–Fike Theorem

Let $\mathbf{A} \in M_n(\mathbb{C})$ be diagonalizable. Write $\mathbf{A} = \mathbf{X}\mathbf{\Lambda}\mathbf{X}^{-1}$, where $\mathbf{X}$ is invertible and $\mathbf{\Lambda}$ is the diagonal matrix with diagonal entries $\lambda_1, \dots, \lambda_n$. Let $\mathbf{E} \in M_n(\mathbb{C})$ be arbitrary, and let $p \in \{1, 2, \infty\}$. Then every eigenvalue μ of $\mathbf{A} + \mathbf{E}$ satisfies

$$\min_{1 \le i \le n} |\lambda_i - \mu| \le \kappa_p(\mathbf{X})\, \|\mathbf{E}\|_p.$$

Idea. Change coordinates so that $\mathbf{A}$ becomes $\mathbf{\Lambda}$. Then $\mathbf{A} + \mathbf{E}$ becomes $\mathbf{\Lambda} + \mathbf{X}^{-1}\mathbf{E}\mathbf{X}$, with the same eigenvalues, and the perturbation has grown by at most the factor $\kappa_p(\mathbf{X})$. For the diagonal matrix $\mathbf{\Lambda}$, the resolvent $(\mathbf{\Lambda} - \mu\mathbf{I})^{-1}$ has norm exactly $1/\min_i |\lambda_i - \mu|$, and Lemma 20.4.1 finishes the proof.

Proof. If $\mu = \lambda_i$ for some i, the left side is 0 and there is nothing to prove. So suppose $\mu \neq \lambda_i$ for every i, and put $\delta = \min_i |\lambda_i - \mu| > 0$.

Since $\mathbf{X}^{-1}(\mathbf{A} + \mathbf{E})\mathbf{X} = \mathbf{\Lambda} + \mathbf{F}$ with $\mathbf{F} = \mathbf{X}^{-1}\mathbf{E}\mathbf{X}$, the matrices $\mathbf{A} + \mathbf{E}$ and $\mathbf{\Lambda} + \mathbf{F}$ are similar, so μ is an eigenvalue of $\mathbf{\Lambda} + \mathbf{F}$ (Theorem 7.10.3). The eigenvalues of $\mathbf{\Lambda}$ are the λ_i, so $\mu \notin \operatorname{spec}(\mathbf{\Lambda})$. By Lemma 20.4.1 with the operator norm $\|\cdot\|_p$,

$$1 \le \|(\mathbf{\Lambda} - \mu\mathbf{I})^{-1}\|_p\, \|\mathbf{F}\|_p.$$

Now $(\mathbf{\Lambda} - \mu\mathbf{I})^{-1} = \operatorname{diag}(1/(\lambda_1 - \mu), \dots, 1/(\lambda_n - \mu))$, whose norm is $\max_i 1/|\lambda_i - \mu| = 1/\delta$ by Eq. 20.4.1. By submultiplicativity (Theorem 16.3.2 (d)), $\|\mathbf{F}\|_p \le \|\mathbf{X}^{-1}\|_p \|\mathbf{E}\|_p \|\mathbf{X}\|_p = \kappa_p(\mathbf{X})\|\mathbf{E}\|_p$. Therefore $1 \le \kappa_p(\mathbf{X})\|\mathbf{E}\|_p/\delta$, that is, $\delta \le \kappa_p(\mathbf{X})\|\mathbf{E}\|_p$. This proves the theorem. ∎

The proof used only two properties of $\|\cdot\|_p$: it is an operator norm, and it gives a diagonal matrix the norm $\max_j |d_j|$. So the theorem holds for every operator norm with that second property. The spectral norm is the one used in practice, and $\kappa_2(\mathbf{X})$ is the headline constant.

Read the theorem as a picture. Draw the closed disc of radius $\kappa_p(\mathbf{X})\|\mathbf{E}\|_p$ about each λ_i; every eigenvalue of $\mathbf{A} + \mathbf{E}$ lies in the union. The bound is **one-sided**, like Elsner's (Theorem 20.3.3). It does not say that each disc holds one perturbed eigenvalue. When the discs are disjoint, Lemma 20.3.4 adds that, and Exercise C1 does it.

Quick check. Let $\mathbf{A} = \operatorname{diag}(1, 2)$. Which matrices $\mathbf{X}$ satisfy $\mathbf{X}^{-1}\mathbf{A}\mathbf{X}$ diagonal, and what does Theorem 20.4.2 give for each of them?

Solution. The eigenvalues $1, 2$ are distinct and each eigenspace is a line: $E_1(\mathbf{A}) = \operatorname{span}(\mathbf{e}_1)$ and $E_2(\mathbf{A}) = \operatorname{span}(\mathbf{e}_2)$. The columns of $\mathbf{X}$ must be eigenvectors (Theorem 9.4.1), so $\mathbf{X} = \operatorname{diag}(a, b)$ or $\mathbf{X} = \left(\begin{smallmatrix} 0 & a \\ b & 0 \end{smallmatrix}\right)$ with $a, b \neq 0$. In both cases the singular values of $\mathbf{X}$ are $|a|$ and $|b|$, so $\kappa_2(\mathbf{X}) = \max(|a|, |b|)/\min(|a|, |b|)$ (Proposition 16.8.2 (d)). The theorem gives $\min_i |\lambda_i - \mu| \le \kappa_2(\mathbf{X})\|\mathbf{E}\|_2$. This is $\|\mathbf{E}\|_2$ when $|a| = |b|$, and arbitrarily weak as $|a|/|b|$ grows. The theorem is true for every choice, and one uses the best.

20.4.3 Normal matrices

For a normal matrix the eigenvectors can be chosen orthonormal, and then the factor is 1. The second half of the next result proves the converse, in the form Chapter 12 §11 stated it: $\kappa = 1$ **exactly** in the normal case.

Corollary 20.4.3. Bauer–Fike for Normal Matrices

(a) Let $\mathbf{A} \in M_n(\mathbb{C})$ be normal with eigenvalues $\lambda_1, \dots, \lambda_n$, and let $\mathbf{E} \in M_n(\mathbb{C})$. Then every eigenvalue μ of $\mathbf{A} + \mathbf{E}$ satisfies $\min_i |\lambda_i - \mu| \le \|\mathbf{E}\|_2$.

(b) Let $\mathbf{A} \in M_n(\mathbb{C})$ be diagonalizable. Then

$$\inf\{\kappa_2(\mathbf{X}) : \mathbf{X}^{-1}\mathbf{A}\mathbf{X} \text{ is diagonal}\} = 1,$$

the infimum over invertible $\mathbf{X}$, if and only if $\mathbf{A}$ is normal.

Idea. Part (a) is the theorem with a unitary $\mathbf{X}$. For (b), write $\mathbf{X} = \mathbf{W}\mathbf{\Sigma}\mathbf{V}^*$ by the singular value decomposition. Then $\mathbf{A}$ is $\mathbf{W}\mathbf{\Sigma}\mathbf{M}\mathbf{\Sigma}^{-1}\mathbf{W}^*$ with $\mathbf{M}$ normal. If all the singular values of $\mathbf{X}$ are nearly equal, $\mathbf{\Sigma}\mathbf{M}\mathbf{\Sigma}^{-1}$ is nearly $\mathbf{M}$, so $\mathbf{A}$ is nearly the normal matrix $\mathbf{W}\mathbf{M}\mathbf{W}^*$, and its commutator $\mathbf{A}^*\mathbf{A} - \mathbf{A}\mathbf{A}^*$ is nearly $\mathbf{O}$. A commutator that is arbitrarily small is zero.

Proof.

(a) By Corollary 12.4.3, $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^*$ with $\mathbf{U}$ unitary and $\mathbf{\Lambda} = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$, so $\mathbf{U}^{-1} = \mathbf{U}^*$. A unitary matrix preserves $\|\cdot\|_2$ (Theorem 11.7.3), so $\|\mathbf{U}\|_2 = \|\mathbf{U}^*\|_2 = 1$ and $\kappa_2(\mathbf{U}) = 1$. Now apply Theorem 20.4.2 with $\mathbf{X} = \mathbf{U}$ and $p = 2$.

(b) ($\Leftarrow$) If $\mathbf{A}$ is normal, the $\mathbf{U}$ of (a) diagonalizes $\mathbf{A}$ with $\kappa_2(\mathbf{U}) = 1$, and every $\kappa_2(\mathbf{X}) \ge 1$ by Proposition 16.8.2 (a). So the infimum is 1.

($\Rightarrow$) First we prove that every invertible $\mathbf{X}$ with $\mathbf{X}^{-1}\mathbf{A}\mathbf{X} = \mathbf{\Lambda}$ diagonal satisfies

$$\|\mathbf{A}^*\mathbf{A} - \mathbf{A}\mathbf{A}^*\|_F \le 2(\kappa_2(\mathbf{X}) - 1)\|\mathbf{\Lambda}\|_F \cdot (\|\mathbf{A}\|_F + \|\mathbf{\Lambda}\|_F). \tag{20.4.2}$$

Write $\mathbf{X} = \mathbf{W}\mathbf{\Sigma}\mathbf{V}^*$ (Theorem 13.9.1), with $\mathbf{W}, \mathbf{V}$ unitary and $\mathbf{\Sigma} = \operatorname{diag}(\sigma_1, \dots, \sigma_n)$. Here

$\sigma_n > 0$, because $\mathbf{X}$ is invertible and so has rank n. Then $\mathbf{X}^{-1} = \mathbf{V}\mathbf{\Sigma}^{-1}\mathbf{W}^*$, and

$$\mathbf{A} = \mathbf{X}\mathbf{\Lambda}\mathbf{X}^{-1} = \mathbf{W}\mathbf{\Sigma}\mathbf{M}\mathbf{\Sigma}^{-1}\mathbf{W}^*, \qquad \mathbf{M} = \mathbf{V}^*\mathbf{\Lambda}\mathbf{V}.$$

Put $\mathbf{N} = \mathbf{W}\mathbf{M}\mathbf{W}^* = (\mathbf{W}\mathbf{V}^*)\mathbf{\Lambda}(\mathbf{W}\mathbf{V}^*)^*$. It is unitarily diagonalizable, hence normal by Corollary 12.4.3. The (i,j) entry of $\mathbf{\Sigma}\mathbf{M}\mathbf{\Sigma}^{-1} - \mathbf{M}$ is $m_{ij}(\sigma_i/\sigma_j - 1)$. Write $\kappa = \kappa_2(\mathbf{X}) = \sigma_1/\sigma_n$ (Proposition 16.8.2 (d)). If $\sigma_i \geq \sigma_j$, then $0 \leq \sigma_i/\sigma_j - 1 \leq \kappa - 1$. If $\sigma_i < \sigma_j$, then $0 < 1 - \sigma_i/\sigma_j \leq 1 - 1/\kappa \leq \kappa - 1$, the last step because $(\kappa - 1) - (1 - 1/\kappa) = (\kappa - 1)^2/\kappa \geq 0$. So every entry is at most $(\kappa - 1)|m_{ij}|$ in modulus. In the display below, the first equality is Lemma 11.7.5 applied to $\mathbf{A} - \mathbf{N} = \mathbf{W}(\mathbf{\Sigma}\mathbf{M}\mathbf{\Sigma}^{-1} - \mathbf{M})\mathbf{W}^*$; the inequality is the entrywise bound just proved, summed over the entries; and the last equality is the same lemma again, applied to $\mathbf{M} = \mathbf{V}^*\mathbf{\Lambda}\mathbf{V}$. So

$$\|\mathbf{A} - \mathbf{N}\|_F = \|\mathbf{\Sigma}\mathbf{M}\mathbf{\Sigma}^{-1} - \mathbf{M}\|_F \leq (\kappa - 1)\|\mathbf{M}\|_F = (\kappa - 1)\|\mathbf{\Lambda}\|_F.$$

Now $\mathbf{N}^*\mathbf{N} = \mathbf{N}\mathbf{N}^*$, and a direct expansion gives

$$\begin{aligned}\mathbf{A}^*\mathbf{A} - \mathbf{A}\mathbf{A}^* &= (\mathbf{A}^*\mathbf{A} - \mathbf{N}^*\mathbf{N}) - (\mathbf{A}\mathbf{A}^* - \mathbf{N}\mathbf{N}^*) \\ &= \mathbf{A}^*(\mathbf{A} - \mathbf{N}) + (\mathbf{A} - \mathbf{N})^*\mathbf{N} \\ &\quad - \mathbf{A}(\mathbf{A} - \mathbf{N})^* - (\mathbf{A} - \mathbf{N})\mathbf{N}^*.\end{aligned}$$

The Frobenius norm is submultiplicative (Example (The Frobenius norm is submultiplicative)), $\|\mathbf{Y}^*\|_F = \|\mathbf{Y}\|_F$ for every $\mathbf{Y}$, since the entries of $\mathbf{Y}^*$ are the conjugates of those of $\mathbf{Y}$ and conjugation preserves the modulus, and $\|\mathbf{N}\|_F = \|\mathbf{\Lambda}\|_F$ by Lemma 11.7.5. So each of the four terms has norm at most $\|\mathbf{A} - \mathbf{N}\|_F$ times $\|\mathbf{A}\|_F$ or $\|\mathbf{\Lambda}\|_F$, and

$$\|\mathbf{A}^*\mathbf{A} - \mathbf{A}\mathbf{A}^*\|_F \leq 2\|\mathbf{A} - \mathbf{N}\|_F(\|\mathbf{A}\|_F + \|\mathbf{\Lambda}\|_F),$$

which with the previous display gives Eq. 20.4.2.

Now suppose the infimum is 1. Every diagonalizing $\mathbf{X}$ produces the same $\|\mathbf{\Lambda}\|_F = (\sum_i |\lambda_i|^2)^{1/2}$, since $\mathbf{\Lambda}$ lists the eigenvalues of $\mathbf{A}$ in some order. So the right side of Eq. 20.4.2 can be made smaller than any positive number, while the left side does not depend on $\mathbf{X}$. Hence $\|\mathbf{A}^*\mathbf{A} - \mathbf{A}\mathbf{A}^*\|_F = 0$, so $\mathbf{A}^*\mathbf{A} = \mathbf{A}\mathbf{A}^*$ and $\mathbf{A}$ is normal. This proves the corollary.

■

Part (b) says more than the bare equivalence. By Eq. 20.4.2, a diagonalizable matrix whose eigenvector matrix is **nearly** unitary is **nearly** normal. It also shows that $\kappa_2(\mathbf{X}) = 1$ forces $\mathbf{A}$ to be normal. In fact $\kappa_2(\mathbf{X}) = 1$ says that $\sigma_1 = \sigma_n$, so $\mathbf{\Sigma} = \sigma_1\mathbf{I}$ and $\mathbf{X} = \sigma_1\mathbf{W}\mathbf{V}^*$, a nonzero multiple of a unitary matrix.

Part (a) is sharp, and the example that shows it is the one Chapter 17 §03 used to warn that Weyl's inequality needs Hermitian matrices. Take $\mathbf{A} = \mathbf{O}$, which is normal with the single eigenvalue 0, and $\mathbf{E} = \left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$, with $\|\mathbf{E}\|_2 = 1$ because $\mathbf{E}$ is unitary. The eigenvalues of $\mathbf{A} + \mathbf{E}$ are $\pm i$, since $p_{\mathbf{E}}(x) = x^2 + 1$. Weyl's comparison of the i-th eigenvalues cannot even be stated, because $\pm i$ are not real. The Bauer–Fike bound can: each of $\pm i$ is within $\|\mathbf{E}\|_2 = 1$ of 0, and the distance is exactly 1. So the constant 1 in (a) cannot be lowered.

20.4.4 Residuals

In practice one rarely knows $\mathbf{E}$. What one has is an approximate eigenpair $(\mu, \widehat{\mathbf{x}})$, computed somehow, and the **residual** $\mathbf{r} = \mathbf{A}\widehat{\mathbf{x}} - \mu\widehat{\mathbf{x}}$, which can be evaluated. The residual turns into a

perturbation: a small residual means that $(\mu, \widehat{\mathbf{x}})$ is an exact eigenpair of a nearby matrix.

Theorem 20.4.4. Residual Bound

Let $\mathbf{A} = \mathbf{X}\mathbf{\Lambda}\mathbf{X}^{-1} \in M_n(\mathbb{C})$ be diagonalizable, with $\mathbf{\Lambda} = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$. Let $\mu \in \mathbb{C}$ and $\widehat{\mathbf{x}} \in \mathbb{C}^n$ with $\widehat{\mathbf{x}} \neq \mathbf{0}$, and let $\eta \geq 0$ satisfy

$$\|\mathbf{A}\widehat{\mathbf{x}} - \mu\widehat{\mathbf{x}}\|_2 \leq \eta \, \|\widehat{\mathbf{x}}\|_2.$$

Then some eigenvalue λ_i of $\mathbf{A}$ satisfies $|\lambda_i - \mu| \leq \kappa_2(\mathbf{X})\, \eta$. If $\mathbf{A}$ is normal, some eigenvalue satisfies $|\lambda_i - \mu| \leq \eta$.

Proof. Put $\mathbf{r} = \mathbf{A}\widehat{\mathbf{x}} - \mu\widehat{\mathbf{x}}$ and $\mathbf{E} = -\mathbf{r}\widehat{\mathbf{x}}^{*}/\|\widehat{\mathbf{x}}\|_2^2$. Then

$$(\mathbf{A} + \mathbf{E})\widehat{\mathbf{x}} = \mathbf{A}\widehat{\mathbf{x}} - \mathbf{r}\,\frac{\widehat{\mathbf{x}}^{*}\widehat{\mathbf{x}}}{\|\widehat{\mathbf{x}}\|_2^2} = \mathbf{A}\widehat{\mathbf{x}} - \mathbf{r} = \mu\widehat{\mathbf{x}},$$

so μ is an eigenvalue of $\mathbf{A}+\mathbf{E}$ with eigenvector $\widehat{\mathbf{x}}$. For every $\mathbf{z}$, the Cauchy–Schwarz inequality (Theorem 11.1.5) gives $|\widehat{\mathbf{x}}^{*}\mathbf{z}| \leq \|\widehat{\mathbf{x}}\|_2\|\mathbf{z}\|_2$, hence

$$\|\mathbf{E}\mathbf{z}\|_2 = \frac{|\widehat{\mathbf{x}}^{*}\mathbf{z}|\,\|\mathbf{r}\|_2}{\|\widehat{\mathbf{x}}\|_2^2} \leq \frac{\|\mathbf{r}\|_2}{\|\widehat{\mathbf{x}}\|_2}\,\|\mathbf{z}\|_2,$$

so $\|\mathbf{E}\|_2 \leq \|\mathbf{r}\|_2/\|\widehat{\mathbf{x}}\|_2 \leq \eta$. (The first inequality is an equality, since $\mathbf{E}\widehat{\mathbf{x}} = -\mathbf{r}$.) By Theorem 20.4.2 with $p = 2$, some λ_i satisfies $|\lambda_i - \mu| \leq \kappa_2(\mathbf{X})\|\mathbf{E}\|_2 \leq \kappa_2(\mathbf{X})\eta$. If $\mathbf{A}$ is normal, use Corollary 20.4.3 (a) instead.

■

The theorem reverses the usual question. Instead of asking how far the eigenvalues of a given perturbed matrix have moved, it asks: of which nearby matrix is my computed answer exactly right? The answer is $\mathbf{A} + \mathbf{E}$ with $\|\mathbf{E}\|_2 = \|\mathbf{r}\|_2/\|\widehat{\mathbf{x}}\|_2$. An error analysis of this kind, which measures an answer by a change in the data that makes it exact, is a **backward error analysis**. Chapter 24 builds on it. For a normal matrix, the scaled residual norm $\|\mathbf{r}\|_2/\|\widehat{\mathbf{x}}\|_2$ is itself a certified error bar for μ.

Example. An error bar from one product

Let $\mathbf{A} = \begin{pmatrix} 2 & 1 \\ 1 & 3 \end{pmatrix}$, $\widehat{\mathbf{x}} = (1, 2)$, and $\mu = \frac{7}{2}$. Compute the residual, and say what interval Theorem 20.4.4 certifies around μ. Then compare it with the truth.

Solution. $\mathbf{A}\widehat{\mathbf{x}} = (4, 7)$ and $\mu\widehat{\mathbf{x}} = (\frac{7}{2}, 7)$, so $\mathbf{r} = (\frac{1}{2}, 0)$, with $\|\mathbf{r}\|_2 = \frac{1}{2}$ and $\|\widehat{\mathbf{x}}\|_2 = \sqrt{5}$. Take $\eta = \|\mathbf{r}\|_2/\|\widehat{\mathbf{x}}\|_2 = \sqrt{5}/10 \approx 0.2236$. The matrix $\mathbf{A}$ is real symmetric, hence normal, so Theorem 20.4.4 gives an eigenvalue of $\mathbf{A}$ within η of $\frac{7}{2}$, and that eigenvalue is real (Theorem 12.2.1). So $\mathbf{A}$ has an eigenvalue in $[\frac{7}{2} - \eta, \frac{7}{2} + \eta] \approx [3.276, 3.724]$, and no eigenvalue was computed to say so: one matrix–vector product did it.

The truth. $p_{\mathbf{A}}(x) = (x - 2)(x - 3) - 1 = x^2 - 5x + 5$, with roots $(5 \pm \sqrt{5})/2$. The larger is ≈ 3.618, which does lie in the interval, at distance ≈ 0.118 from $\frac{7}{2}$ — about half of η, so the

certificate costs a factor of 2 in sharpness.

20.4.5 A worked example

Chapter 12 §11 perturbed a corner of $\left(\begin{smallmatrix}1 & M\\ 0 & 1\end{smallmatrix}\right)$, with a double eigenvalue, and found the eigenvalues moving by $\sqrt{M\varepsilon}$ (Example (A perturbation that moves the eigenvalues a thousand times as far)). Its closing remark said that "a matrix with well-separated eigenvalues and a large defect behaves less dramatically". Bauer–Fike says how much less.

Example. A large corner, separated eigenvalues

Let $M > 0$, $\mathbf{A} = \begin{pmatrix}1 & M\\ 0 & 2\end{pmatrix}$, and $\mathbf{E} = \varepsilon\,\mathbf{e}_2\mathbf{e}_1^{\top}$ with $0 < \varepsilon \le 1$. Diagonalize $\mathbf{A}$ in two ways, compute κ_2 of each eigenvector matrix, and compare the Bauer–Fike bound with the exact movement of the eigenvalues.

Solution. Eigenvectors. $\mathbf{A}$ is upper triangular with distinct diagonal entries $1, 2$, which are its eigenvalues. For 1, $\mathbf{A} - \mathbf{I} = \left(\begin{smallmatrix}0 & M\\ 0 & 1\end{smallmatrix}\right)$ has kernel $\operatorname{span}(\mathbf{e}_1)$. For 2, $\mathbf{A} - 2\mathbf{I} = \left(\begin{smallmatrix}-1 & M\\ 0 & 0\end{smallmatrix}\right)$ has kernel spanned by $(M, 1)$. Two natural choices are

$$\mathbf{X}_1 = \begin{pmatrix}1 & M\\ 0 & 1\end{pmatrix}, \qquad \mathbf{X}_2 = \begin{pmatrix}1 & M/s\\ 0 & 1/s\end{pmatrix}, \quad s = \sqrt{1+M^2}.$$

The second has columns of length 1. Both satisfy $\mathbf{X}^{-1}\mathbf{A}\mathbf{X} = \operatorname{diag}(1, 2)$.

$\kappa_2(\mathbf{X}_1)$. $\mathbf{X}_1^{*}\mathbf{X}_1 = \left(\begin{smallmatrix}1 & M\\ M & 1+M^2\end{smallmatrix}\right)$ has trace $2 + M^2$ and determinant 1. Its eigenvalues $\sigma_1^2 \ge \sigma_2^2$ therefore have product 1, and $\kappa_2(\mathbf{X}_1) = \sigma_1/\sigma_2 = \sigma_1^2$. Solving the quadratic,

$$\kappa_2(\mathbf{X}_1) = \frac{2 + M^2 + M\sqrt{M^2+4}}{2},$$

which exceeds M^2.

$\kappa_2(\mathbf{X}_2)$. $\mathbf{X}_2^{*}\mathbf{X}_2 = \left(\begin{smallmatrix}1 & q\\ q & 1\end{smallmatrix}\right)$ with $q = M/s$, since the columns are unit vectors whose inner product is M/s. Its eigenvalues are $1 \pm q$, so

$$\kappa_2(\mathbf{X}_2)^2 = \frac{1+q}{1-q} = \frac{s+M}{s-M} = (s+M)^2,$$

using $(s+M)(s-M) = s^2 - M^2 = 1$. Hence $\kappa_2(\mathbf{X}_2) = M + \sqrt{1+M^2}$, which lies between $2M$ and $2M+1$.

The bound. $\|\mathbf{E}\|_2 = \varepsilon$, since $\mathbf{E}$ has a single nonzero entry. With $\mathbf{X}_2$, Theorem 20.4.2 gives movement at most $(M + \sqrt{1+M^2})\varepsilon$, about $2M\varepsilon$. With $\mathbf{X}_1$ it gives about $M^2\varepsilon$.

The truth. $p_{\mathbf{A}+\mathbf{E}}(x) = (x-1)(x-2) - M\varepsilon = x^2 - 3x + 2 - M\varepsilon$, with roots $\left(3 \pm \sqrt{1+4M\varepsilon}\right)/2$. Each eigenvalue has moved by

$$\frac{\sqrt{1+4M\varepsilon} - 1}{2} = \frac{2M\varepsilon}{1 + \sqrt{1+4M\varepsilon}},$$

multiplying top and bottom by $\sqrt{1+4M\varepsilon}+1$. The denominator is at least 2, and at most $2+2M\varepsilon$ because $\sqrt{1+4M\varepsilon} \le 1+2M\varepsilon$. So the movement lies between $M\varepsilon/(1+M\varepsilon)$ and $M\varepsilon$.

Numbers. For $M = 100$ and $\varepsilon = 10^{-6}$, the eigenvalues move by about 0.9999×10^{-4}, and the bound with $\mathbf{X}_2$ is about 2.0×10^{-4}. The bound with $\mathbf{X}_1$ is about 10^{-2}. By comparison, in Example (A perturbation that moves the eigenvalues a thousand times as far), the same corner and the same ε moved a **double** eigenvalue by $\sqrt{M\varepsilon} = 10^{-2}$. Separating the eigenvalues turns a square-root amplification into a linear one, with a constant of about M. The constant is still large, and $\kappa_2(\mathbf{X}_2)$ predicts it correctly to within a factor of about 2.

Warning. **$\kappa_2(\mathbf{X})$ belongs to the choice of $\mathbf{X}$, not to $\mathbf{A}$.** Eigenvectors can be rescaled, and rescaling changes κ_2: in Example (A large corner, separated eigenvalues) the same matrix gave $\kappa_2(\mathbf{X}_1) \approx 10^4$ and $\kappa_2(\mathbf{X}_2) \approx 200$ at $M = 100$. The theorem holds for every choice, so use the best one you can find; normalizing the columns is a reasonable first step. And even the best $\mathbf{X}$ measures all the eigenvalues at once. Section 11 attaches a separate condition number to each simple eigenvalue, and one eigenvalue can be well conditioned while $\mathbf{X}$ is badly conditioned.

20.4.6 Defective matrices

When $\mathbf{A}$ is not diagonalizable, write $\mathbf{A} = \mathbf{X}\mathbf{J}\mathbf{X}^{-1}$ with $\mathbf{J}$ a Jordan form (Theorem 10.3.1). The letter $\mathbf{J}$ is local to this subsection; elsewhere in the book $\mathbf{J}$ is the all-ones matrix, which does not appear here. The proof of Theorem 20.4.2 goes through verbatim, except that $(\mathbf{J}-\mu\mathbf{I})^{-1}$ is no longer diagonal. Its norm is no longer $1/\delta$; it grows like $1/\delta^m$ as $\delta \to 0$, where m is the size of the largest Jordan block. Taking m-th roots brings back the exponent of Section 3.

Theorem 20.4.5. Bauer–Fike for a Jordan Form

Let $\mathbf{A} = \mathbf{X}\mathbf{J}\mathbf{X}^{-1} \in M_n(\mathbb{C})$, where $\mathbf{X}$ is invertible and $\mathbf{J}$ is a Jordan form of $\mathbf{A}$ whose largest block has size m. Let $\lambda_1, \dots, \lambda_n$ be the eigenvalues of $\mathbf{A}$, let $\mathbf{E} \in M_n(\mathbb{C})$, and put $\theta = m\,\kappa_2(\mathbf{X})\|\mathbf{E}\|_2$. Then every eigenvalue μ of $\mathbf{A}+\mathbf{E}$ satisfies

$$\min_{1 \le i \le n} |\lambda_i - \mu| \le \max(\theta, \theta^{1/m}).$$

Idea. Run the proof of Theorem 20.4.2 with $\mathbf{J}$ in place of $\mathbf{\Lambda}$. The only new work is to bound $\|(\mathbf{J}_k(\lambda)-\mu\mathbf{I})^{-1}\|_2$ for one block. The block is $(\lambda-\mu)\mathbf{I}+\mathbf{N}$, with $\mathbf{N}$ nilpotent of norm at most 1, so its inverse is a **finite** Neumann series $\sum_{j<k}(-1)^j\mathbf{N}^j/(\lambda-\mu)^{j+1}$ with at most m terms. Bound it by $m\delta^{-1}$ when $\delta \ge 1$ and by $m\delta^{-m}$ when $\delta < 1$.

Proof. Put $\delta = \min_i|\lambda_i-\mu|$. If $\delta = 0$ there is nothing to prove, so suppose $\delta > 0$. The upper triangular $\mathbf{J}$ has its diagonal entries as eigenvalues (Theorem 9.7.4 (b)), and these are the λ_i because $\mathbf{J}$ and $\mathbf{A}$ are similar (Theorem 7.10.3); so $\mu \notin \operatorname{spec}(\mathbf{J})$. As in the proof of Theorem 20.4.2, μ is an eigenvalue of $\mathbf{J}+\mathbf{X}^{-1}\mathbf{E}\mathbf{X}$, and Lemma 20.4.1 with submultiplicativity

gives

$$1 \le \|(\mathbf{J}-\mu\mathbf{I})^{-1}\|_2 \, \kappa_2(\mathbf{X}) \, \|\mathbf{E}\|_2.$$

One block. Let $\mathbf{J}_k(\lambda)$ be a block of $\mathbf{J}$, so $k \le m$ and $|\lambda-\mu| \ge \delta$, and write $\mathbf{J}_k(\lambda)-\mu\mathbf{I} = a\mathbf{I}+\mathbf{N}$ with $a = \lambda-\mu \ne 0$ and $\mathbf{N} = \mathbf{J}_k(0)$. The matrix $\mathbf{N}$ sends $\mathbf{e}_1$ to $\mathbf{0}$ and $\mathbf{e}_j$ to $\mathbf{e}_{j-1}$ for $j \ge 2$, so $\mathbf{N}^k$ sends every $\mathbf{e}_j$ to $\mathbf{0}$ and $\mathbf{N}^k = \mathbf{O}$. Then

$$(a\mathbf{I}+\mathbf{N})\sum_{j=0}^{k-1}\frac{(-1)^j\mathbf{N}^j}{a^{j+1}} = \sum_{j=0}^{k-1}\frac{(-1)^j\mathbf{N}^j}{a^j} - \sum_{j=1}^{k}\frac{(-1)^j\mathbf{N}^j}{a^j} = \mathbf{I} - \frac{(-1)^k\mathbf{N}^k}{a^k} = \mathbf{I},$$

so the sum is the inverse of $a\mathbf{I}+\mathbf{N}$. The same description of $\mathbf{N}$ gives $\|\mathbf{N}\mathbf{x}\|_2^2 = \sum_{j\ge 2}|x_j|^2 \le \|\mathbf{x}\|_2^2$, so $\|\mathbf{N}\|_2 \le 1$. Hence $\|\mathbf{N}^j\|_2 \le 1$ for every $j \ge 0$ by submultiplicativity, and the triangle inequality (Theorem 16.3.2 (b)) gives

$$\|(\mathbf{J}_k(\lambda)-\mu\mathbf{I})^{-1}\|_2 \le \sum_{j=1}^{k}|a|^{-j} \le \sum_{j=1}^{m}\delta^{-j},$$

using $|a| \ge \delta$ and $k \le m$.

All blocks. $(\mathbf{J}-\mu\mathbf{I})^{-1}$ is block diagonal, with the inverses of the blocks $\mathbf{J}_k(\lambda)-\mu\mathbf{I}$ on its diagonal. Split $\mathbf{x}$ into the matching pieces $\mathbf{x}_1,\dots,\mathbf{x}_r$ and let $\mathbf{C}_1,\dots,\mathbf{C}_r$ be the diagonal blocks. Then $\|(\mathbf{J}-\mu\mathbf{I})^{-1}\mathbf{x}\|_2^2 = \sum_l \|\mathbf{C}_l\mathbf{x}_l\|_2^2 \le (\max_l\|\mathbf{C}_l\|_2)^2\|\mathbf{x}\|_2^2$. So $\|(\mathbf{J}-\mu\mathbf{I})^{-1}\|_2 \le \sum_{j=1}^m \delta^{-j}$, and

$$1 \le \frac{\theta}{m}\sum_{j=1}^{m}\delta^{-j}.$$

Two cases. If $\delta \ge 1$, each $\delta^{-j} \le \delta^{-1}$, so $1 \le \theta/\delta$ and $\delta \le \theta$. If $\delta < 1$, each $\delta^{-j} \le \delta^{-m}$, so $1 \le \theta\delta^{-m}$, that is $\delta^m \le \theta$ and $\delta \le \theta^{1/m}$. In both cases $\delta \le \max(\theta, \theta^{1/m})$. This proves the theorem.

■

For $m = 1$, $\mathbf{J}$ is diagonal, $\theta = \kappa_2(\mathbf{X})\|\mathbf{E}\|_2$, and the theorem is Theorem 20.4.2 with $p = 2$. For $m \ge 2$ and small $\mathbf{E}$, the term $\theta^{1/m}$ dominates, and the eigenvalues move like the m-th root of the perturbation. That exponent is the Jordan block's signature, and it is the answer to Chapter 10 §03's warning that the Jordan form is not stable. Section 3's Example (The Jordan block meets Elsner's bound) shows it cannot be improved. There $\mathbf{A} = \mathbf{J}_n(0)$, $\mathbf{X} = \mathbf{I}$, $m = n$, and $\|\mathbf{E}\|_2 = \varepsilon \le 1/n$, so $\theta = n\varepsilon \le 1$ and the theorem gives $\max(\theta, \theta^{1/n}) = (n\varepsilon)^{1/n}$ while the eigenvalues actually move $\varepsilon^{1/n}$. The two differ by the factor $n^{1/n}$, which never exceeds $3^{1/3} < 1.45$.

Only the size of the largest block enters, and a diagonalizable part of $\mathbf{A}$ is charged the same m-th root as the defective part. Section 11 separates the eigenvalues. There a simple eigenvalue gets its own first-order condition number, and that number becomes infinite exactly when two eigenvalues merge into a Jordan block.

20.4.7 Exercises

20.4.7.1 A. Check your understanding

Exercise. A1

(a) State Theorem 20.4.2 with all its hypotheses.

(b) True or false: the constant $\kappa_2(\mathbf{X})$ in Theorem 20.4.2 is determined by $\mathbf{A}$. Justify your answer.

(c) True or false: for every $\mathbf{A}, \mathbf{E} \in M_n(\mathbb{C})$, every eigenvalue of $\mathbf{A}+\mathbf{E}$ lies within $\|\mathbf{E}\|_2$ of an eigenvalue of $\mathbf{A}$. Justify your answer.

(d) What does Corollary 20.4.3 (a) give when $\mathbf{A}$ is normal, and which example shows that its constant cannot be lowered?

(e) In Theorem 20.4.5, what exponent appears for small $\|\mathbf{E}\|_2$, and what decides it?

Solution.

(a) See Theorem 20.4.2: $\mathbf{A} = \mathbf{X}\mathbf{\Lambda}\mathbf{X}^{-1}$ diagonalizable, $\mathbf{E}$ arbitrary, $p \in \{1, 2, \infty\}$; every eigenvalue μ of $\mathbf{A}+\mathbf{E}$ is within $\kappa_p(\mathbf{X})\|\mathbf{E}\|_p$ of some λ_i.

(b) False. The columns of $\mathbf{X}$ can be rescaled, and κ_2 changes. The Quick check gave $\mathbf{X} = \operatorname{diag}(a, b)$ for $\mathbf{A} = \operatorname{diag}(1, 2)$, with $\kappa_2(\mathbf{X}) = \max(|a|, |b|)/\min(|a|, |b|)$, any number ≥ 1.

(c) False. By Example (The Jordan block meets Elsner's bound) with $n = 2$, $\mathbf{A} = \mathbf{J}_2(0)$ and $\mathbf{E} = \varepsilon \mathbf{e}_2\mathbf{e}_1^{\top}$ with $0 < \varepsilon < 1$ give eigenvalues $\pm\sqrt{\varepsilon}$ of $\mathbf{A}+\mathbf{E}$, at distance $\sqrt{\varepsilon} > \varepsilon = \|\mathbf{E}\|_2$ from 0. The hypothesis that fails is diagonalizability.

(d) Every eigenvalue of $\mathbf{A}+\mathbf{E}$ is within $\|\mathbf{E}\|_2$ of an eigenvalue of $\mathbf{A}$. The example $\mathbf{A} = \mathbf{O}$, $\mathbf{E} = \left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$ has perturbed eigenvalues $\pm i$ at distance exactly $1 = \|\mathbf{E}\|_2$.

(e) The exponent $1/m$, where m is the size of the largest Jordan block of $\mathbf{A}$.

20.4.7.2 B. Practice

Exercise. B1: A Bauer–Fike region

Let $\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 0 & 3 \end{pmatrix}$.

(a) Find an eigenvector matrix $\mathbf{X}$ with columns of length 1, and compute $\kappa_2(\mathbf{X})$.

(b) Hence find r such that every eigenvalue of $\mathbf{A}+\mathbf{E}$ lies within r of 1 or of 3 whenever $\|\mathbf{E}\|_2 \leq 0.01$.

(c) Compute the eigenvalues of $\mathbf{A} + 0.01\,\mathbf{e}_2\mathbf{e}_1^{\top}$ exactly, and compare their movement with r.

Solution.

(a) The eigenvalues are 1 and 3, with eigenvectors $\mathbf{e}_1$ and $(1, 1)$, since $(\mathbf{A} - 3\mathbf{I})(1, 1) = (-2 + 2, 0) = \mathbf{0}$. So $\mathbf{X} = \left(\begin{smallmatrix} 1 & 1/\sqrt{2} \\ 0 & 1/\sqrt{2} \end{smallmatrix}\right)$. Then $\mathbf{X}^*\mathbf{X} = \left(\begin{smallmatrix} 1 & q \\ q & 1 \end{smallmatrix}\right)$ with $q = 1/\sqrt{2}$, with eigenvalues $1 \pm q$, so

$$\kappa_2(\mathbf{X})^2 = \frac{1+q}{1-q} = \frac{\sqrt{2}+1}{\sqrt{2}-1} = (\sqrt{2}+1)^2,$$

and $\kappa_2(\mathbf{X}) = 1 + \sqrt{2} \approx 2.414$.

(b) By Theorem 20.4.2 with $p = 2$, $r = (1 + \sqrt{2})(0.01) \approx 0.0241$ works.

(c) $p(x) = (x - 1)(x - 3) - 0.02 = x^2 - 4x + 2.98$, with roots $2 \pm \sqrt{1.02}$. Each eigenvalue has moved by $\sqrt{1.02} - 1 \approx 0.00995$, well within $r \approx 0.0241$.

Exercise. B2: A residual certificate

Let $\mathbf{A} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$, $\widehat{\mathbf{x}} = (10, 9)$, and $\mu = 3$.

(a) Compute the residual $\mathbf{r} = \mathbf{A}\widehat{\mathbf{x}} - \mu\widehat{\mathbf{x}}$ and $\eta = \|\mathbf{r}\|_2 / \|\widehat{\mathbf{x}}\|_2$.

(b) Hence show that $\mathbf{A}$ has an eigenvalue in the interval $[3 - \eta, 3 + \eta]$, without computing the eigenvalues of $\mathbf{A}$.

Solution.

(a) $\mathbf{A}\widehat{\mathbf{x}} = (29, 28)$ and $3\widehat{\mathbf{x}} = (30, 27)$, so $\mathbf{r} = (-1, 1)$, $\|\mathbf{r}\|_2 = \sqrt{2}$, $\|\widehat{\mathbf{x}}\|_2 = \sqrt{181}$, and $\eta = \sqrt{2/181} \approx 0.105$.

(b) $\mathbf{A}$ is real symmetric, hence normal. By Theorem 20.4.4, some eigenvalue λ satisfies $|\lambda - 3| \le \eta$. The eigenvalues of a real symmetric matrix are real (Theorem 12.2.1), so $\lambda \in [3 - \eta, 3 + \eta]$. (In fact $\lambda = 3$, with eigenvector $(1, 1)$.)

Exercise. B3: A defective bound

Let $\mathbf{A} = \mathbf{J}_2(3)$.

(a) Explain why Theorem 20.4.2 does not apply to $\mathbf{A}$.

(b) Use Theorem 20.4.5 to find r such that every eigenvalue of $\mathbf{A} + \mathbf{E}$ lies within r of 3 whenever $\|\mathbf{E}\|_2 \le 10^{-4}$.

(c) Compute the eigenvalues of $\mathbf{A} + 10^{-4}\mathbf{e}_2\mathbf{e}_1^{\top}$, and compare.

Solution.

(a) $\mathbf{A}$ has the single eigenvalue 3 with algebraic multiplicity 2, and $\mathbf{A} - 3\mathbf{I} = \mathbf{J}_2(0)$ has rank 1, so the eigenspace has dimension 1. Hence $\mathbf{A}$ is not diagonalizable (Theorem 9.4.1), and there is no $\mathbf{X}$ to put in Theorem 20.4.2.

(b) $\mathbf{A}$ is its own Jordan form, so take $\mathbf{X} = \mathbf{I}$, with $\kappa_2(\mathbf{I}) = 1$, and $m = 2$. Then $\theta = 2 \cdot 1 \cdot 10^{-4} = 2 \times 10^{-4}$, and $\max(\theta, \theta^{1/2}) = \sqrt{2} \times 10^{-2} \approx 0.0141$. So $r = \sqrt{2} \times 10^{-2}$ works.

(c) $p(x) = (x - 3)^2 - 10^{-4}$, with roots 3 ± 10^{-2}. The movement 0.01 is below $r \approx 0.0141$, and it is a hundred times the size 10^{-4} of the perturbation: the square root at work.

20.4.7.3 C. Going deeper

Exercise. C1: One eigenvalue in each disc

Let $\mathbf{A} = \mathbf{X}\mathbf{\Lambda}\mathbf{X}^{-1}$ have n distinct eigenvalues $\lambda_1, \dots, \lambda_n$, put $g = \min_{i \neq j} |\lambda_i - \lambda_j|$, and let $\mathbf{E}$ satisfy $r = \kappa_2(\mathbf{X})\|\mathbf{E}\|_2 < g/2$.

(a) Prove that for each i, $\mathbf{A} + \mathbf{E}$ has exactly one eigenvalue in the disc $\{z : |z - \lambda_i| \leq r\}$, and that $\mathbf{A} + \mathbf{E}$ is diagonalizable.

(b) Deduce that if $\mathbf{A}$ and $\mathbf{E}$ are real and every λ_i is real, then every eigenvalue of $\mathbf{A} + \mathbf{E}$ is real.

Hint: apply Lemma 20.3.4 to the path $\mathbf{A} + t\mathbf{E}$.

Solution.

(a) Let $Q_i = \{z : |z - \lambda_i| \leq r\}$. For $i \neq j$, $z \in Q_i$, and $w \in Q_j$, the triangle inequality (Theorem 0.6.4) gives $|z - w| \geq |\lambda_i - \lambda_j| - 2r \geq g - 2r > 0$. So Q_i and the union L_i of the other Q_j are at distance at least $g - 2r > 0$. For $t \in [0, 1]$, Theorem 20.4.2 applied to $\mathbf{A}$ and $t\mathbf{E}$ puts every eigenvalue of $\mathbf{A} + t\mathbf{E}$ within $\kappa_2(\mathbf{X})\|t\mathbf{E}\|_2 \leq r$ of some λ_j, that is, in $Q_i \cup L_i$. The path $\mathbf{A} + t\mathbf{E} = (1 - t)\mathbf{A} + t(\mathbf{A} + \mathbf{E})$ is of the form in Lemma 20.3.4, so the number of eigenvalues in Q_i is the same at $t = 0$ and $t = 1$. At $t = 0$ it is 1: $\lambda_i \in Q_i$, and $\lambda_j \in Q_j \subseteq L_i$ is not in Q_i for $j \neq i$. So $\mathbf{A} + \mathbf{E}$ has exactly one eigenvalue in each Q_i. The Q_i are pairwise disjoint, so these n eigenvalues are distinct, and $\mathbf{A} + \mathbf{E}$ is diagonalizable by Corollary 9.4.2.

(b) Let μ be the eigenvalue of $\mathbf{A} + \mathbf{E}$ in Q_i. Since λ_i is real, $|\overline{\mu} - \lambda_i| = |\mu - \lambda_i| \leq r$, so $\overline{\mu} \in Q_i$. Since $\mathbf{A} + \mathbf{E}$ is real, $\overline{\mu}$ is also an eigenvalue (Theorem 9.2.5). By (a), Q_i holds only one eigenvalue, so $\overline{\mu} = \mu$, and μ is real. Every eigenvalue of $\mathbf{A} + \mathbf{E}$ lies in some Q_i, so all of them are real.

Exercise. C2: The best shift for a residual

Let $\mathbf{A} \in M_n(\mathbb{C})$ and $\widehat{\mathbf{x}} \in \mathbb{C}^n$ with $\widehat{\mathbf{x}} \neq \mathbf{0}$, and put $\rho = \widehat{\mathbf{x}}^* \mathbf{A} \widehat{\mathbf{x}} / \widehat{\mathbf{x}}^* \widehat{\mathbf{x}}$.

(a) Prove that $\|\mathbf{A}\widehat{\mathbf{x}} - \mu\widehat{\mathbf{x}}\|_2 \geq \|\mathbf{A}\widehat{\mathbf{x}} - \rho\widehat{\mathbf{x}}\|_2$ for every $\mu \in \mathbb{C}$, with equality only for $\mu = \rho$.

(b) Deduce that if $\mathbf{A}$ is normal, then $\mathbf{A}$ has an eigenvalue λ with $|\lambda - \rho| \leq \|\mathbf{A}\widehat{\mathbf{x}} - \mu\widehat{\mathbf{x}}\|_2 / \|\widehat{\mathbf{x}}\|_2$ for **every** $\mu \in \mathbb{C}$.

Solution.

(a) Put $\mathbf{w} = \mathbf{A}\widehat{\mathbf{x}} - \rho\widehat{\mathbf{x}}$. Then $\langle \mathbf{w}, \widehat{\mathbf{x}} \rangle = \widehat{\mathbf{x}}^* \mathbf{w} = \widehat{\mathbf{x}}^* \mathbf{A} \widehat{\mathbf{x}} - \rho\, \widehat{\mathbf{x}}^* \widehat{\mathbf{x}} = 0$ by the choice of ρ. For any μ, $\mathbf{A}\widehat{\mathbf{x}} - \mu\widehat{\mathbf{x}} = \mathbf{w} + (\rho - \mu)\widehat{\mathbf{x}}$ is a sum of two orthogonal vectors, so by the Pythagorean theorem (Theorem 11.1.3)

$$\|\mathbf{A}\widehat{\mathbf{x}} - \mu\widehat{\mathbf{x}}\|_2^2 = \|\mathbf{w}\|_2^2 + |\rho - \mu|^2 \|\widehat{\mathbf{x}}\|_2^2.$$

This is at least $\|\mathbf{w}\|_2^2$, with equality exactly when $|\rho - \mu|\, \|\widehat{\mathbf{x}}\|_2 = 0$, that is, when $\mu = \rho$, since $\widehat{\mathbf{x}} \neq \mathbf{0}$.

(b) Apply Theorem 20.4.4 for normal $\mathbf{A}$ with the shift ρ and $\eta = \|\mathbf{A}\widehat{\mathbf{x}} - \rho\widehat{\mathbf{x}}\|_2 / \|\widehat{\mathbf{x}}\|_2$. It gives an eigenvalue λ with $|\lambda - \rho| \leq \eta$, and $\eta \leq \|\mathbf{A}\widehat{\mathbf{x}} - \mu\widehat{\mathbf{x}}\|_2 / \|\widehat{\mathbf{x}}\|_2$ for every μ by (a).

20.5 The Hoffman–Wielandt Theorem

Section 04 showed that for a normal $\mathbf{A}$, every eigenvalue of $\mathbf{A}+\mathbf{E}$ lies within $\|\mathbf{E}\|_2$ of **some** eigenvalue of $\mathbf{A}$. That answers "how far", one perturbed eigenvalue at a time. It does not answer "which goes where": it never says that the eigenvalues of the two matrices can be **paired off**, each with a partner of its own. This section proves that for two normal matrices they can, with the total squared distance of the pairs bounded by $\|\mathbf{A}-\mathbf{B}\|_F^2$. The proof is short once the right tool is in hand, and the tool is Birkhoff's theorem from Chapter 19 §07. We then sharpen the result for Hermitian matrices, where the pairing is simply the decreasing order, and show that without normality the inequality fails for every pairing.

Throughout, the field is $\mathbb{C}$, and eigenvalues are listed **with algebraic multiplicity**. For a Hermitian matrix, $\lambda_1(\mathbf{A}) \geq \cdots \geq \lambda_n(\mathbf{A})$ is the decreasing list, as in Chapter 17.

20.5.1 One-sided bounds do not pair eigenvalues

Recall Corollary 20.4.3 (a): if $\mathbf{A} \in M_n(\mathbb{C})$ is normal and $\mathbf{E} \in M_n(\mathbb{C})$ is arbitrary, then every eigenvalue μ of $\mathbf{A}+\mathbf{E}$ satisfies $\min_i|\lambda_i-\mu| \leq \|\mathbf{E}\|_2$, where $\lambda_1,\dots,\lambda_n$ are the eigenvalues of $\mathbf{A}$. The statement starts from a perturbed eigenvalue and finds a partner for it. Nothing in it stops two perturbed eigenvalues from choosing the **same** partner.

Here is the smallest case. Let

$$\mathbf{A} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}, \qquad \mathbf{A}+\mathbf{E} = \begin{pmatrix} \frac{1}{5} & 0 \\ 0 & \frac{2}{5} \end{pmatrix}, \qquad \mathbf{E} = \begin{pmatrix} \frac{1}{5} & 0 \\ 0 & -\frac{3}{5} \end{pmatrix}.$$

Then $\|\mathbf{E}\|_2 = \frac{3}{5}$, and both perturbed eigenvalues are within $\frac{3}{5}$ of an eigenvalue of $\mathbf{A}$, as the corollary promises. But if each one goes to its **nearest** eigenvalue of $\mathbf{A}$, both go to 0: the distances are $\frac{1}{5}$ and $\frac{2}{5}$ to 0, against $\frac{4}{5}$ and $\frac{3}{5}$ to 1. The nearest-eigenvalue map is not a bijection, and the eigenvalue 1 is left without a partner. A pairing does exist here, $0 \leftrightarrow \frac{1}{5}$ and $1 \leftrightarrow \frac{2}{5}$, with distances $\frac{1}{5}$ and $\frac{3}{5}$, but finding it was not the corollary's doing.

Why ask for a pairing at all? Because it is what "the eigenvalues move a little" should mean. Chapter 16's Corollary 16.7.4 is a pairing statement: the eigenvalues of a nearby matrix **can be numbered** so that each is close to its counterpart. It gives no rate. For Hermitian matrices, Weyl's Corollary 17.3.6 gives a pairing with a rate, and the pairing is the decreasing order. For normal matrices with complex eigenvalues there is no decreasing order, and the pairing has to be found.

To say this precisely, let $\lambda_1,\dots,\lambda_n$ and $\mu_1,\dots,\mu_n$ be two lists of complex numbers. A **pairing** of them is a permutation $\sigma \in S_n$, which matches λ_i with $\mu_{\sigma(i)}$. Its cost, in the sense of this section, is the sum of squared distances

$$\sum_{i=1}^{n} |\lambda_i - \mu_{\sigma(i)}|^2.$$

A one-sided statement, "every μ_j is within r of some λ_i", is a statement about sets. A pairing is a statement about lists with multiplicity, and it is strictly stronger.

Quick check. Let $\mathbf{A} = \operatorname{diag}(0,0,1)$ and $\mathbf{B} = \operatorname{diag}(0,1,1)$. Show that every eigenvalue of $\mathbf{B}$ is an eigenvalue of $\mathbf{A}$ and conversely, so each one-sided distance is 0. What is the smallest possible largest distance $\max_i|\lambda_i-\mu_{\sigma(i)}|$ over all pairings σ?

Solution. Both matrices have the eigenvalue set $\{0, 1\}$, so every eigenvalue of either one is at distance 0 from an eigenvalue of the other. As lists, $\mathbf{A}$ has $(0, 0, 1)$ and $\mathbf{B}$ has $(0, 1, 1)$. Any pairing must send one of the two 0's of $\mathbf{A}$ to a 1 of $\mathbf{B}$, since $\mathbf{B}$ has only one 0. So every pairing has largest distance at least 1, and the identity pairing attains 1. Sets that agree can still be far apart as lists.

20.5.2 Birkhoff turns weights into a permutation

The proof below will produce a sum of the form $\sum_{i,j} s_{ij} c_{ij}$, in which $c_{ij} = |\lambda_i - \mu_j|^2$ is the cost of pairing λ_i with μ_j, and the weights s_{ij} form a doubly stochastic matrix. Such a sum is a **fractional pairing**: every λ_i spreads a total weight 1 over the μ_j, and every μ_j receives a total weight 1. A genuine pairing is the special case where each weight is 0 or 1, that is, where (s_{ij}) is a permutation matrix. The lemma says that fractional pairings never beat the best genuine one.

Recall Definition 18.8.4: $\mathbf{S} \in M_n(\mathbb{R})$ is doubly stochastic if its entries are nonnegative and every row and every column adds up to 1; the set of such matrices is Ω_n. Recall also Definition 3.4.2: $\mathbf{P}_\sigma$ is the matrix whose j-th column is $\mathbf{e}_{\sigma(j)}$, so its entry in position $(\sigma(j), j)$ is 1 for each j, and all its other entries are 0.

Lemma 20.5.1. A Fractional Pairing Is Never Cheaper

Let $\mathbf{C} = (c_{ij}) \in M_n(\mathbb{R})$ and $\mathbf{S} = (s_{ij}) \in \Omega_n$. Then there is a permutation $\sigma \in S_n$ with

$$\sum_{j=1}^{n} c_{\sigma(j)j} \;\le\; \sum_{i,j=1}^{n} s_{ij} c_{ij}.$$

Idea. Write $f(\mathbf{X}) = \sum_{i,j} x_{ij} c_{ij}$, a linear function of $\mathbf{X}$. Birkhoff's theorem writes $\mathbf{S}$ as an average of permutation matrices, so $f(\mathbf{S})$ is the same average of the numbers $f(\mathbf{P}_\sigma)$, and an average is at least its smallest term.

Proof. Let $f(\mathbf{X}) = \sum_{i,j} x_{ij} c_{ij}$ for $\mathbf{X} \in M_n(\mathbb{R})$; it is linear in $\mathbf{X}$. By Theorem 19.7.2, $\mathbf{S} = \sum_{k=1}^{m} t_k \mathbf{P}_{\sigma_k}$ for some permutations $\sigma_1, \dots, \sigma_m$ and weights $t_k \ge 0$ with $\sum_k t_k = 1$. By linearity,

$$f(\mathbf{S}) = \sum_{k=1}^{m} t_k \, f(\mathbf{P}_{\sigma_k}) \;\ge\; \Big(\sum_{k=1}^{m} t_k\Big) \min_k f(\mathbf{P}_{\sigma_k}) = \min_k f(\mathbf{P}_{\sigma_k}),$$

where the inequality uses $t_k \ge 0$. Let σ be a σ_k attaining the minimum. The only nonzero entries of $\mathbf{P}_\sigma$ are the 1's in positions $(\sigma(j), j)$, so $f(\mathbf{P}_\sigma) = \sum_j c_{\sigma(j)j}$. Hence $\sum_j c_{\sigma(j)j} \le f(\mathbf{S})$, as claimed.

■

In the language of Chapter 18, this is Corollary 18.8.6 applied to Ω_n. With the Frobenius inner product $\langle \mathbf{X}, \mathbf{C} \rangle = \operatorname{tr}(\mathbf{C}^{\top}\mathbf{X}) = \sum_{i,j} x_{ij} c_{ij}$, the function f is $\mathbf{X} \mapsto \langle \mathbf{X}, \mathbf{C} \rangle$. The set Ω_n is nonempty, compact, and convex (Chapter 18 §08), so the minimum of f over it is attained at an extreme point, and by Theorem 19.7.2 the extreme points are the permutation matrices. The

proof above uses the second form of Birkhoff's theorem instead, so no compactness argument appears in it.

20.5.3 The theorem

We can now prove that two normal matrices have a pairing of their eigenvalues whose cost is at most the squared Frobenius distance between the matrices.

Theorem 20.5.2. Hoffman–Wielandt Theorem

Let $\mathbf{A}, \mathbf{B} \in M_n(\mathbb{C})$ be **normal**, with eigenvalues $\lambda_1, \dots, \lambda_n$ and $\mu_1, \dots, \mu_n$ respectively, each listed with algebraic multiplicity in any order. Then there is a permutation $\sigma \in S_n$ with

$$\sum_{i=1}^{n} |\lambda_i - \mu_{\sigma(i)}|^2 \;\le\; \|\mathbf{A} - \mathbf{B}\|_F^2.$$

Idea. **Step roadmap.**

① **Diagonalize both.** The spectral theorem gives $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^*$ and $\mathbf{B} = \mathbf{V}\mathbf{M}\mathbf{V}^*$ with $\mathbf{U}, \mathbf{V}$ unitary and $\mathbf{\Lambda}, \mathbf{M}$ diagonal. The two orthonormal bases of eigenvectors are different, and the unitary $\mathbf{W} = \mathbf{U}^*\mathbf{V}$ records how they sit relative to each other.

② **Compute the distance.** The Frobenius norm does not see unitary factors, so $\|\mathbf{A} - \mathbf{B}\|_F = \|\mathbf{\Lambda}\mathbf{W} - \mathbf{W}\mathbf{M}\|_F$. The (i, j) entry of $\mathbf{\Lambda}\mathbf{W} - \mathbf{W}\mathbf{M}$ is $(\lambda_i - \mu_j)w_{ij}$, so $\|\mathbf{A} - \mathbf{B}\|_F^2 = \sum_{i,j} |w_{ij}|^2 |\lambda_i - \mu_j|^2$. This is a fractional pairing with weights $|w_{ij}|^2$.

③ **Round it.** The weights form a doubly stochastic matrix, because the rows and the columns of a unitary matrix are unit vectors. The lemma replaces the fractional pairing by a genuine one that costs no more.

Proof. By Corollary 12.4.3, since $\mathbf{A}$ and $\mathbf{B}$ are normal, there are unitary $\mathbf{U}, \mathbf{V} \in \mathrm{U}(n)$ with

$$\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^*, \qquad \mathbf{B} = \mathbf{V}\mathbf{M}\mathbf{V}^*,$$

where $\mathbf{\Lambda} = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$ and $\mathbf{M} = \operatorname{diag}(\mu_1, \dots, \mu_n)$ in the given orders. Put $\mathbf{W} = \mathbf{U}^*\mathbf{V}$, which is unitary as a product of unitary matrices. Since $\mathbf{U}^*\mathbf{U} = \mathbf{I}$ and $\mathbf{V}^*\mathbf{V} = \mathbf{I}$,

$$\mathbf{U}^*(\mathbf{A} - \mathbf{B})\mathbf{V} = \mathbf{\Lambda}\mathbf{U}^*\mathbf{V} - \mathbf{U}^*\mathbf{V}\mathbf{M} = \mathbf{\Lambda}\mathbf{W} - \mathbf{W}\mathbf{M}.$$

By Lemma 11.7.5, applied with the unitary factors $\mathbf{U}^*$ on the left and $\mathbf{V}$ on the right, $\|\mathbf{A} - \mathbf{B}\|_F = \|\mathbf{\Lambda}\mathbf{W} - \mathbf{W}\mathbf{M}\|_F$. The (i, j) entry of $\mathbf{\Lambda}\mathbf{W}$ is $\lambda_i w_{ij}$ and that of $\mathbf{W}\mathbf{M}$ is $w_{ij}\mu_j$, since multiplying by a diagonal matrix on the left scales rows and on the right scales columns. Hence

$$\|\mathbf{A} - \mathbf{B}\|_F^2 = \sum_{i,j=1}^{n} |w_{ij}|^2 \, |\lambda_i - \mu_j|^2. \tag{20.5.1}$$

Let $\mathbf{S} = (s_{ij})$ with $s_{ij} = |w_{ij}|^2$. Its entries are nonnegative. Row i adds up to $\sum_j |w_{ij}|^2 = (\mathbf{W}\mathbf{W}^*)_{ii} = 1$, and column j adds up to $\sum_i |w_{ij}|^2 = (\mathbf{W}^*\mathbf{W})_{jj} = 1$, since $\mathbf{W}$ is unitary. So

$\mathbf{S} \in \Omega_n$. Apply Lemma 20.5.1 to $\mathbf{S}$ and the real matrix $\mathbf{C}$ with $c_{ij} = |\lambda_i - \mu_j|^2$. It gives $\tau \in S_n$ with

$$\sum_{j=1}^{n} |\lambda_{\tau(j)} - \mu_j|^2 \;\le\; \sum_{i,j} s_{ij} c_{ij} = \|\mathbf{A} - \mathbf{B}\|_F^2,$$

the equality being Eq. 20.5.1. Put $\sigma = \tau^{-1}$. As j runs over $1, \dots, n$, so does $i = \tau(j)$, and $j = \sigma(i)$; so the left-hand side is $\sum_i |\lambda_i - \mu_{\sigma(i)}|^2$. This proves the theorem.

■

Both hypotheses were used in the first line and nowhere else: normality is what makes the diagonalizing matrices unitary. The rest of the proof is bookkeeping and Birkhoff's theorem, which is where Chapter 19 §07's theorem earns its keep.

In words: *the eigenvalues of a normal matrix, as a list with multiplicity, are a* 1*-Lipschitz function of the matrix, when matrices are measured in the Frobenius norm and lists by their best pairing.* In particular every pair in the best pairing is within $\|\mathbf{A} - \mathbf{B}\|_F$, so each eigenvalue of $\mathbf{B}$ has its own partner among those of $\mathbf{A}$ within that distance. This is a matching bound, with a constant, valid for all normal matrices at once. Compare Corollary 16.7.4, which gives a matching with no constant for all matrices, and Corollary 20.4.3, which gives a constant with no matching.

The theorem asserts that **some** pairing is cheap. The proof does not say which, beyond what the lemma gives, but the cheapest of the $n!$ pairings certainly works, since it costs no more than any other. The next example shows that a pairing chosen by any other rule can fail.

Example. A Hermitian and a diagonal matrix

Let

$$\mathbf{A} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}, \qquad \mathbf{B} = \begin{pmatrix} i & 0 \\ 0 & 2+i \end{pmatrix}.$$

Both are normal. Compute $\|\mathbf{A} - \mathbf{B}\|_F^2$, the cost of both pairings of the eigenvalues, and the matrix $\mathbf{S}$ of the proof. Which pairings satisfy the conclusion of Theorem 20.5.2?

Solution. $\mathbf{A}$ is real symmetric, hence Hermitian and normal, and $\mathbf{B}$ is diagonal, hence normal. The eigenvalues of $\mathbf{A}$ are $\lambda_1 = 3$ and $\lambda_2 = 1$, with unit eigenvectors $\frac{1}{\sqrt{2}}(1,1)$ and $\frac{1}{\sqrt{2}}(1,-1)$: indeed $\mathbf{A}(1,1) = (3,3)$ and $\mathbf{A}(1,-1) = (1,-1)$. Those of $\mathbf{B}$ are $\mu_1 = i$ and $\mu_2 = 2+i$, with eigenvectors $\mathbf{e}_1, \mathbf{e}_2$.

The distance. $\mathbf{A} - \mathbf{B} = \left(\begin{smallmatrix} 2-i & 1 \\ 1 & -i \end{smallmatrix}\right)$, so

$$\|\mathbf{A} - \mathbf{B}\|_F^2 = |2-i|^2 + 1 + 1 + |-i|^2 = 5 + 1 + 1 + 1 = 8.$$

The costs. The four squared distances $c_{ij} = |\lambda_i - \mu_j|^2$ are

$$c_{11} = |3-i|^2 = 10, \qquad c_{12} = |1-i|^2 = 2,$$
$$c_{21} = |1-i|^2 = 2, \qquad c_{22} = |-1-i|^2 = 2.$$

The identity pairing $3 \leftrightarrow i$, $1 \leftrightarrow 2+i$ costs $c_{11} + c_{22} = 12$. The swap $3 \leftrightarrow 2+i$, $1 \leftrightarrow i$ costs $c_{12} + c_{21} = 4$.

The weights. With $\mathbf{U} = \frac{1}{\sqrt{2}}\left(\begin{smallmatrix} 1 & 1 \\ 1 & -1 \end{smallmatrix}\right)$ and $\mathbf{V} = \mathbf{I}$, $\mathbf{W} = \mathbf{U}^*$, and every entry of $\mathbf{W}$ has modulus $\frac{1}{\sqrt{2}}$. So $\mathbf{S} = \frac{1}{2}\mathbf{J}$, and Eq. 20.5.1 reads

$$\|\mathbf{A} - \mathbf{B}\|_F^2 = \tfrac{1}{2}(10 + 2 + 2 + 2) = 8,$$

the average of the two costs, $\frac{1}{2}(12 + 4)$. An average lies between its terms, so the cheaper pairing, the swap with cost 4, satisfies $4 \le 8$. The identity pairing, with cost $12 > 8$, does not.

Warning. **The theorem does not say that the given numbering works.** In the example, the lists $(3, 1)$ and $(i, 2 + i)$, paired in the order written, cost $12 > 8 = \|\mathbf{A} - \mathbf{B}\|_F^2$. Complex numbers have no natural order to sort by, so, unlike the Hermitian case below, there is no rule fixed in advance that always picks a good pairing. The safe reading is: the **cheapest** pairing satisfies the inequality.

20.5.4 Normality cannot be dropped

The proof used normality once, in its first line, to make the diagonalizing matrices unitary. For diagonalizable matrices $\mathbf{A} = \mathbf{X}\mathbf{\Lambda}\mathbf{X}^{-1}$ and $\mathbf{B} = \mathbf{Y}\mathbf{M}\mathbf{Y}^{-1}$ with $\mathbf{X}$, $\mathbf{Y}$ not unitary, Lemma 11.7.5 no longer applies, and $\|\mathbf{A} - \mathbf{B}\|_F$ is no longer the norm of $\mathbf{\Lambda}\mathbf{W} - \mathbf{W}\mathbf{M}$ with $\mathbf{W} = \mathbf{X}^{-1}\mathbf{Y}$. The failure is real, not an artifact of the proof.

Example. Every pairing fails

Let

$$\mathbf{A} = \begin{pmatrix} 1 & 3 \\ 0 & -1 \end{pmatrix}, \qquad \mathbf{B} = \begin{pmatrix} 1 & 3 \\ 1 & -1 \end{pmatrix}.$$

Show that $\mathbf{A}$ and $\mathbf{B}$ are diagonalizable but not normal, and that the inequality of Theorem 20.5.2 fails for **every** pairing of their eigenvalues.

Solution. $\mathbf{A}$ is triangular with eigenvalues 1 and -1. The matrix $\mathbf{B}$ has trace 0 and determinant $-1 - 3 = -4$, so its characteristic polynomial is $x^2 - 4$ and its eigenvalues are 2 and -2. Each has two distinct eigenvalues, so each is diagonalizable (Theorem 9.4.1). Neither is normal:

$$\mathbf{A}^*\mathbf{A} = \begin{pmatrix} 1 & 3 \\ 3 & 10 \end{pmatrix} \neq \begin{pmatrix} 10 & -3 \\ -3 & 1 \end{pmatrix} = \mathbf{A}\mathbf{A}^*,$$

$$\mathbf{B}^*\mathbf{B} = \begin{pmatrix} 2 & 2 \\ 2 & 10 \end{pmatrix} \neq \begin{pmatrix} 10 & -2 \\ -2 & 2 \end{pmatrix} = \mathbf{B}\mathbf{B}^*.$$

The difference $\mathbf{A} - \mathbf{B}$ has a single nonzero entry, -1 in position $(2, 1)$, so $\|\mathbf{A} - \mathbf{B}\|_F^2 = 1$.

There are two pairings. The one that matches signs, $1 \leftrightarrow 2$ and $-1 \leftrightarrow -2$, costs $1^2 + 1^2 = 2$. The other, $1 \leftrightarrow -2$ and $-1 \leftrightarrow 2$, costs $3^2 + 3^2 = 18$. Both exceed 1, so no pairing satisfies the inequality.

The entry 3 is what makes this work, and the failure is not confined to a perturbation as large

as 1. Put $\delta > 0$ in the corner instead: $\mathbf{B}_\delta = \left(\begin{smallmatrix} 1 & 3 \\ \delta & -1 \end{smallmatrix}\right)$ has determinant $-1 - 3\delta$ and eigenvalues $\pm\sqrt{1+3\delta}$, each of which moves by $\sqrt{1+3\delta} - 1$, about $\frac{3}{2}\delta$ for small δ. So the cheaper pairing costs about $2\left(\frac{3}{2}\delta\right)^2 = \frac{9}{2}\delta^2$, well above $\|\mathbf{A} - \mathbf{B}_\delta\|_F^2 = \delta^2$. With an entry m in place of 3 the cost is about $\frac{1}{2}m^2\delta^2$, so the inequality fails for small δ as soon as $m > \sqrt{2}$. A large entry above the diagonal is exactly what Chapter 12 §11 measured as departure from normality, and it is what makes the eigenvalues sensitive.

A Jordan block shows that no constant can repair the inequality. Let $\mathbf{A} = \mathbf{J}_2(0)$ and, for $\varepsilon > 0$, $\mathbf{B}_\varepsilon = \left(\begin{smallmatrix} 0 & 1 \\ \varepsilon & 0 \end{smallmatrix}\right)$. Then $\mathbf{A}$ has the eigenvalue 0 twice, $\mathbf{B}_\varepsilon$ has characteristic polynomial $x^2 - \varepsilon$ and eigenvalues $\pm\sqrt{\varepsilon}$, and $\|\mathbf{A} - \mathbf{B}_\varepsilon\|_F^2 = \varepsilon^2$. Both pairings cost the same, $(\sqrt{\varepsilon})^2 + (\sqrt{\varepsilon})^2 = 2\varepsilon$, and

$$\frac{2\varepsilon}{\varepsilon^2} = \frac{2}{\varepsilon} \longrightarrow \infty \qquad (\varepsilon \to 0^+).$$

So there is no constant K with $\min_\sigma \sum_i |\lambda_i - \mu_{\sigma(i)}|^2 \le K\|\mathbf{A} - \mathbf{B}\|_F^2$ for all pairs of matrices. This is Example (A Jordan block's eigenvalues move like $\varepsilon^{1/k}$) of Chapter 16 read in the language of pairings: the eigenvalues move like $\sqrt{\varepsilon}$ while the matrix moves by ε.

Warning. **Diagonalizable is not enough, and one normal matrix is not enough.** In Example (Every pairing fails) both matrices are diagonalizable, and every pairing fails. Nor does it help to make just one of the two matrices normal: for $\mathbf{A} = \mathbf{J}_2(0)$ and the Hermitian $\mathbf{B} = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$, with eigenvalues ± 1, both pairings cost $1 + 1 = 2$, while $\|\mathbf{A} - \mathbf{B}\|_F^2 = 1$. The theorem needs **both** matrices normal.

20.5.5 Hermitian matrices: the pairing is the order

When both matrices are Hermitian, the eigenvalues are real, and there is a natural guess for the best pairing: largest with largest, second with second, and so on. The guess is right, and the reason is a fact about real numbers alone.

Lemma 20.5.3. The Sorted Pairing Is Cheapest

Let $a_1 \ge a_2 \ge \cdots \ge a_n$ and $b_1 \ge b_2 \ge \cdots \ge b_n$ be real numbers. Then for every $\sigma \in S_n$,

$$\sum_{i=1}^{n} (a_i - b_i)^2 \;\le\; \sum_{i=1}^{n} (a_i - b_{\sigma(i)})^2.$$

Idea. Expanding the squares, $\sum_i a_i^2$ and $\sum_i b_{\sigma(i)}^2 = \sum_j b_j^2$ do not depend on σ, so the claim is that the cross term $\sum_i a_i b_{\sigma(i)}$ is largest at the identity. If σ puts a smaller b before a larger one in two neighboring positions, swapping those two values can only increase the cross term, and it removes one inversion. Repeating the swap sorts σ into the identity.

Proof. For every σ, since σ is a bijection, $\sum_i b_{\sigma(i)}^2 = \sum_j b_j^2$, and so

$$\sum_{i=1}^{n}(a_i - b_{\sigma(i)})^2 = \sum_i a_i^2 + \sum_j b_j^2 - 2g(\sigma),$$
$$g(\sigma) = \sum_{i=1}^{n} a_i b_{\sigma(i)}.$$

It therefore suffices to show $g(\sigma) \le g(\mathrm{id})$ for every σ. The set S_n is finite, so g attains a maximum on it; among the permutations where it does, choose one, σ, with the fewest inversions (Definition 7.3.1). Suppose $\sigma \neq \mathrm{id}$. Then $\sigma(i) > \sigma(i+1)$ for some i, since a permutation with $\sigma(1) < \sigma(2) < \cdots < \sigma(n)$ is the identity. Let σ' agree with σ except that $\sigma'(i) = \sigma(i+1)$ and $\sigma'(i+1) = \sigma(i)$. Then

$$g(\sigma') - g(\sigma) = (a_i - a_{i+1})(b_{\sigma(i+1)} - b_{\sigma(i)}) \;\ge\; 0,$$

because $a_i \ge a_{i+1}$, and $\sigma(i+1) < \sigma(i)$ gives $b_{\sigma(i+1)} \ge b_{\sigma(i)}$. So σ' also maximizes g. Now count inversions. The pair of positions $(i, i+1)$ is an inversion of σ and not of σ'. A pair involving neither i nor $i+1$ sees the same values under both. For a position $p \notin \{i, i+1\}$, the two pairs joining p to i and to $i+1$ compare $\sigma(p)$ with the same two values $\sigma(i), \sigma(i+1)$ under σ and under σ', from the same side, so together they contain the same number of inversions. Hence σ' has one inversion fewer than σ, contradicting the choice of σ. So $\sigma = \mathrm{id}$ maximizes g, and $g(\sigma) \le g(\mathrm{id})$ for every σ. This proves the lemma.

■

The move is an **exchange argument**: to show that an arrangement is optimal, show that repairing one local defect never makes things worse. The inequality $\sum_i a_i b_{\sigma(i)} \le \sum_i a_i b_i$ is known as the rearrangement inequality.

Example. Three numbers, six pairings

Let $\mathbf{a} = (4, 2, 1)$ and $\mathbf{b} = (3, 0, -1)$, both decreasing. Compute $\sum_i (a_i - b_{\sigma(i)})^2$ for all six $\sigma \in S_3$, and find the cheapest and the dearest.

Solution. The nine squared distances $(a_i - b_j)^2$, with i the row and j the column, are

$$\begin{pmatrix} 1 & 16 & 25 \\ 1 & 4 & 9 \\ 4 & 1 & 4 \end{pmatrix}.$$

A pairing picks one entry from each row and each column. Writing σ by its values $(\sigma(1), \sigma(2), \sigma(3))$,

$$(1,2,3) : 1 + 4 + 4 = 9, \qquad (1,3,2) : 1 + 9 + 1 = 11,$$
$$(2,1,3) : 16 + 1 + 4 = 21, \quad (2,3,1) : 16 + 9 + 4 = 29,$$
$$(3,1,2) : 25 + 1 + 1 = 27, \quad (3,2,1) : 25 + 4 + 4 = 33.$$

The identity, which pairs the two sorted lists in order, costs 9, the least of the six, as Lemma 20.5.3 says. The dearest is the reversed pairing $(3, 2, 1)$, at 33.

One exchange step of the proof is visible here. The permutation $(1, 3, 2)$ has the single inversion $\sigma(2) > \sigma(3)$; undoing it gives the identity and lowers the cost by 2, which is $2(a_2 - a_3)(b_2 - b_3) = 2 \cdot 1 \cdot 1$.

Corollary 20.5.4. Hoffman–Wielandt for Hermitian Matrices

Let $\mathbf{A}, \mathbf{B} \in M_n(\mathbb{C})$ be Hermitian. Then

$$\sum_{i=1}^{n} \big(\lambda_i(\mathbf{A}) - \lambda_i(\mathbf{B})\big)^2 \;\le\; \|\mathbf{A} - \mathbf{B}\|_F^2.$$

Proof. Hermitian matrices are normal (Proposition 11.6.7 (c)) and have real eigenvalues. By Theorem 20.5.2 applied to the lists $\lambda_1(\mathbf{A}), \dots, \lambda_n(\mathbf{A})$ and $\lambda_1(\mathbf{B}), \dots, \lambda_n(\mathbf{B})$, there is $\sigma \in S_n$ with $\sum_i \big(\lambda_i(\mathbf{A}) - \lambda_{\sigma(i)}(\mathbf{B})\big)^2 \le \|\mathbf{A} - \mathbf{B}\|_F^2$, where $|t|^2 = t^2$ for real t. Both lists are decreasing, so by Lemma 20.5.3 the sorted pairing costs no more than σ. This proves the corollary. ∎

This is the Frobenius-norm companion of Weyl's bound. Put $\mathbf{E} = \mathbf{A} - \mathbf{B}$, Hermitian. By the spectral theorem $\mathbf{E} = \mathbf{U}\mathbf{D}\mathbf{U}^*$ with $\mathbf{U}$ unitary and $\mathbf{D} = \operatorname{diag}(\lambda_1(\mathbf{E}), \dots, \lambda_n(\mathbf{E}))$, so Lemma 11.7.5 gives $\|\mathbf{E}\|_F^2 = \|\mathbf{D}\|_F^2 = \sum_i \lambda_i(\mathbf{E})^2$, while $\|\mathbf{E}\|_2 = \max_i |\lambda_i(\mathbf{E})|$ by Lemma 17.3.5. So Weyl's Corollary 17.3.6 says

$$\max_i \big|\lambda_i(\mathbf{A}) - \lambda_i(\mathbf{B})\big| \;\le\; \max_i |\lambda_i(\mathbf{E})|,$$

and Corollary 20.5.4 says

$$\sum_i \big(\lambda_i(\mathbf{A}) - \lambda_i(\mathbf{B})\big)^2 \;\le\; \sum_i \lambda_i(\mathbf{E})^2.$$

In both, a norm of the vector of shifts $\boldsymbol{\lambda}(\mathbf{A}) - \boldsymbol{\lambda}(\mathbf{B})$ is bounded by the same norm of $\boldsymbol{\lambda}(\mathbf{E})$: the ∞-norm for Weyl, the 2-norm for Hoffman–Wielandt.

Neither implies the other. For $\mathbf{E} = t\mathbf{I}$ with $t > 0$, every eigenvalue shifts by t; Weyl's bound t is exact for each shift, while Hoffman–Wielandt alone allows a single shift as large as $\|\mathbf{E}\|_F = \sqrt{n}\, t$. For a rank-one $\mathbf{E} = t\,\mathbf{x}\mathbf{x}^*$ with $\|\mathbf{x}\|_2 = 1$, $\|\mathbf{E}\|_F = \|\mathbf{E}\|_2 = t$; Weyl lets **every** eigenvalue move by up to t, while Hoffman–Wielandt says that the squared shifts together use up at most t^2, so if one eigenvalue moves by t, none of the others moves at all.

Remark. Lidskii's inequality contains both. Chapter 17 §07 proved $\boldsymbol{\lambda}(\mathbf{A}) - \boldsymbol{\lambda}(\mathbf{B}) \prec \boldsymbol{\lambda}(\mathbf{A} - \mathbf{B})$ (the remark after Theorem 17.7.2), and Weyl's bound follows from it there, as Corollary 17.7.3 (a). Chapter 21 shows that a majorization $\mathbf{x} \prec \mathbf{y}$ gives $\sum_i \phi(x_i) \le \sum_i \phi(y_i)$ for every convex ϕ; with $\phi(t) = t^2$ this is Corollary 20.5.4 again, and Chapter 21 turns the same majorization into one bound for every unitarily invariant norm. Nothing here depends on that route; the proof above goes through Birkhoff's theorem, Theorem 19.7.2.

Quick check. Let $\mathbf{A} = \operatorname{diag}(3, 1)$ and $\mathbf{B} = \left(\begin{smallmatrix} 2 & 1 \\ 1 & 2 \end{smallmatrix}\right)$. Compute $\|\mathbf{A} - \mathbf{B}\|_F^2$ and both sides of Corollary 20.5.4. What does the reversed pairing $\lambda_1(\mathbf{A}) \leftrightarrow \lambda_2(\mathbf{B})$, $\lambda_2(\mathbf{A}) \leftrightarrow \lambda_1(\mathbf{B})$ cost?

Solution. $\mathbf{A}-\mathbf{B}=\left(\begin{smallmatrix}1 & -1\\ -1 & -1\end{smallmatrix}\right)$, so $\|\mathbf{A}-\mathbf{B}\|_F^2=4$. The eigenvalues of $\mathbf{B}$ are 3 and 1 (as in Example (A Hermitian and a diagonal matrix)), the same as those of $\mathbf{A}$. So the left side of the corollary is $0\le 4$: two matrices at Frobenius distance 2 can have identical spectra, and the inequality is only an upper bound. The reversed pairing costs $(3-1)^2+(1-3)^2=8>4$. Even for Hermitian matrices, the pairing matters.

20.5.6 Exercises

20.5.6.1 A. Check your understanding

Exercise. A1

(a) State the Hoffman–Wielandt theorem, with its hypotheses.

(b) True or false: if $\mathbf{A}$ and $\mathbf{B}$ are normal, the eigenvalues of $\mathbf{B}$ listed in any order satisfy $\sum_i|\lambda_i-\mu_i|^2\le\|\mathbf{A}-\mathbf{B}\|_F^2$. Justify your answer.

(c) True or false: the Hoffman–Wielandt inequality holds for every pair of diagonalizable matrices. Justify your answer.

(d) For normal $\mathbf{A}$ and $\mathbf{B}$, what bound does the theorem give on the distance between an eigenvalue of $\mathbf{B}$ and its partner?

(e) Where exactly does Birkhoff's theorem (Theorem 19.7.2) enter the proof, and what would be missing without it?

(f) For Hermitian $\mathbf{A}$, $\mathbf{B}$ with $\mathbf{E}=\mathbf{A}-\mathbf{B}$ of rank one, explain why at most one eigenvalue can move by the full amount $\|\mathbf{E}\|_2$.

Solution.

(a) If $\mathbf{A},\mathbf{B}\in M_n(\mathbb{C})$ are normal, with eigenvalues $\lambda_1,\dots,\lambda_n$ and $\mu_1,\dots,\mu_n$ listed with multiplicity, then there is a permutation σ with $\sum_i|\lambda_i-\mu_{\sigma(i)}|^2\le\|\mathbf{A}-\mathbf{B}\|_F^2$.

(b) False. In Example (A Hermitian and a diagonal matrix), the order $(3,1)$ and $(i,2+i)$ costs $12>8$. Only a suitable pairing, for instance the cheapest one, is guaranteed.

(c) False. In Example (Every pairing fails) both matrices have two distinct eigenvalues, hence are diagonalizable, and both pairings cost more than $\|\mathbf{A}-\mathbf{B}\|_F^2=1$.

(d) Each summand is at most the whole sum, so $|\lambda_i-\mu_{\sigma(i)}|\le\|\mathbf{A}-\mathbf{B}\|_F$ for every i, for the pairing σ of the theorem.

(e) The proof writes $\|\mathbf{A}-\mathbf{B}\|_F^2$ as $\sum_{i,j}s_{ij}|\lambda_i-\mu_j|^2$ with (s_{ij}) doubly stochastic, a fractional pairing. Birkhoff's theorem, through Lemma 20.5.1, replaces it by a genuine pairing that costs no more. Without it, we would know only that a weighted average of the costs $|\lambda_i-\mu_j|^2$ is small, which does not produce a permutation.

(f) For Hermitian $\mathbf{E}$ of rank one, $\mathbf{E}$ has at most one nonzero eigenvalue, so $\|\mathbf{E}\|_F^2=\sum_i\lambda_i(\mathbf{E})^2=\|\mathbf{E}\|_2^2$. By Corollary 20.5.4 the squared shifts add up to at most $\|\mathbf{E}\|_2^2$. If one shift equals $\|\mathbf{E}\|_2$ in absolute value, the others are 0.

20.5.6.2 B. Practice

Exercise. B1: A pair of normal matrices

Let

$$\mathbf{A} = \begin{pmatrix} 0 & 2 \\ 2 & 0 \end{pmatrix}, \qquad \mathbf{B} = \begin{pmatrix} 1 & 0 \\ 0 & -1+2i \end{pmatrix}.$$

Check that both are normal, compute $\|\mathbf{A}-\mathbf{B}\|_F^2$ and the costs of both pairings of the eigenvalues, and compute the matrix $\mathbf{S}$ of the proof of Theorem 20.5.2. Hence decide which pairings satisfy the Hoffman–Wielandt inequality.

Solution. $\mathbf{A}$ is real symmetric, hence normal, and $\mathbf{B}$ is diagonal, hence normal. $\mathbf{A}(1,1) = (2,2)$ and $\mathbf{A}(1,-1) = (-2,2)$, so $\mathbf{A}$ has eigenvalues $\lambda_1 = 2$, $\lambda_2 = -2$ with unit eigenvectors $\frac{1}{\sqrt{2}}(1,\pm 1)$. $\mathbf{B}$ has eigenvalues $\mu_1 = 1$, $\mu_2 = -1+2i$.

$\mathbf{A}-\mathbf{B} = \left(\begin{smallmatrix} -1 & 2 \\ 2 & 1-2i \end{smallmatrix}\right)$, so $\|\mathbf{A}-\mathbf{B}\|_F^2 = 1+4+4+5 = 14$.

The squared distances are $c_{11} = |2-1|^2 = 1$, $c_{12} = |3-2i|^2 = 13$, $c_{21} = |-3|^2 = 9$, $c_{22} = |-1-2i|^2 = 5$. The identity pairing costs $1+5 = 6$, and the swap costs $13+9 = 22$.

With $\mathbf{U} = \frac{1}{\sqrt{2}}\left(\begin{smallmatrix} 1 & 1 \\ 1 & -1 \end{smallmatrix}\right)$ and $\mathbf{V} = \mathbf{I}$, $\mathbf{W} = \mathbf{U}^*$ has all entries of modulus $\frac{1}{\sqrt{2}}$, so $\mathbf{S} = \frac{1}{2}\mathbf{J}$, and indeed $\frac{1}{2}(1+13+9+5) = 14$. Hence the identity pairing satisfies the inequality, $6 \le 14$, and the swap does not, $22 > 14$.

Exercise. B2: Hermitian, compared with Weyl

Let

$$\mathbf{A} = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \qquad \mathbf{B} = \begin{pmatrix} 2 & 1 & 0 \\ 1 & 2 & 0 \\ 0 & 0 & -1 \end{pmatrix}.$$

Compute $\boldsymbol{\lambda}(\mathbf{A})$, $\boldsymbol{\lambda}(\mathbf{B})$, $\|\mathbf{A}-\mathbf{B}\|_F$, and $\|\mathbf{A}-\mathbf{B}\|_2$. Verify Corollary 20.5.4 and Corollary 17.3.6 for this pair. Hence say which of the two bounds is closer to the truth here.

Solution. $\boldsymbol{\lambda}(\mathbf{A}) = (2,1,0)$. The top-left block of $\mathbf{B}$ has eigenvalues 3 and 1 (with eigenvectors $(1,1)$ and $(1,-1)$), and the last entry is -1, so $\boldsymbol{\lambda}(\mathbf{B}) = (3,1,-1)$. The shifts are $\lambda_i(\mathbf{B}) - \lambda_i(\mathbf{A}) = 1, 0, -1$.

$\mathbf{E} = \mathbf{B}-\mathbf{A} = \left(\begin{smallmatrix} 0 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & -1 \end{smallmatrix}\right)$, so $\|\mathbf{E}\|_F^2 = 1+1+1+1 = 4$ and $\|\mathbf{E}\|_F = 2$. Its top-left block has characteristic polynomial $x^2 - x - 1$, with roots $\frac{1\pm\sqrt{5}}{2}$, and its last eigenvalue is -1. By Lemma 17.3.5, $\|\mathbf{E}\|_2 = \frac{1+\sqrt{5}}{2} \approx 1.618$.

Hoffman–Wielandt: $1^2 + 0^2 + (-1)^2 = 2 \le 4$. Weyl: each $|\text{shift}| \le 1 \le 1.618$. Hence both hold. Weyl overestimates the largest shift by a factor of about 1.6; Hoffman–Wielandt overestimates the sum of squares by a factor of 2. Neither is sharp here, and neither dominates.

Exercise. B3: Sorting

Let $\mathbf{a} = (5, 2, -1)$ and $\mathbf{b} = (4, 3, 0)$. Compute $\sum_i (a_i - b_{\sigma(i)})^2$ for all six $\sigma \in S_3$. Hence confirm Lemma 20.5.3, and say which pairing is the most expensive.

Solution. Write σ by its values $(\sigma(1), \sigma(2), \sigma(3))$. Then

$$\begin{aligned} &(1,2,3) : 1 + 1 + 1 = 3, && (1,3,2) : 1 + 4 + 16 = 21, \\ &(2,1,3) : 4 + 4 + 1 = 9, && (2,3,1) : 4 + 4 + 25 = 33, \\ &(3,1,2) : 25 + 4 + 16 = 45, && (3,2,1) : 25 + 1 + 25 = 51. \end{aligned}$$

The identity, which pairs the sorted lists in order, costs 3, the least of the six, as the lemma says. The most expensive is $(3, 2, 1)$, which pairs the largest a with the smallest b and so on: the reversed order. Exercise C1 (b) shows this is no accident.

20.5.6.3 C. Going deeper

Exercise. C1: The most expensive pairing

(a) Let $\mathbf{A}, \mathbf{B} \in M_n(\mathbb{C})$ be normal, with eigenvalues λ_i and μ_i. Prove that there is also a permutation σ with $\sum_i |\lambda_i - \mu_{\sigma(i)}|^2 \geq \|\mathbf{A} - \mathbf{B}\|_F^2$.

(b) Deduce that for Hermitian $\mathbf{A}, \mathbf{B}$,

$$\sum_{i=1}^{n} \big(\lambda_i(\mathbf{A}) - \lambda_{n+1-i}(\mathbf{B})\big)^2 \;\geq\; \|\mathbf{A} - \mathbf{B}\|_F^2.$$

Hint for (b): apply Lemma 20.5.3 to the list $-\lambda_{n+1-i}(\mathbf{B})$.

Solution.

(a) Apply Lemma 20.5.1 to the same $\mathbf{S} \in \Omega_n$ as in the proof of Theorem 20.5.2 and to the matrix $-\mathbf{C}$, where $c_{ij} = |\lambda_i - \mu_j|^2$. It gives τ with $-\sum_j c_{\tau(j)j} \leq -\sum_{i,j} s_{ij} c_{ij} = -\|\mathbf{A} - \mathbf{B}\|_F^2$, by Eq. 20.5.1. Multiplying by -1 and putting $\sigma = \tau^{-1}$, as in that proof, gives $\sum_i |\lambda_i - \mu_{\sigma(i)}|^2 \geq \|\mathbf{A} - \mathbf{B}\|_F^2$.

(b) Write $a_i = \lambda_i(\mathbf{A})$ and $c_i = \lambda_i(\mathbf{B})$, and let $b_i = -c_{n+1-i}$. Both (a_i) and (b_i) are decreasing, the second because c_{n+1-i} increases with i. Let $\pi \in S_n$, and let $\sigma(i) = n + 1 - \pi(i)$, which is again a permutation, with $b_{\sigma(i)} = -c_{\pi(i)}$. By Lemma 20.5.3, $\sum_i (a_i - b_i)^2 \leq \sum_i (a_i - b_{\sigma(i)})^2$, that is,

$$\sum_i (a_i + c_{n+1-i})^2 \;\leq\; \sum_i (a_i + c_{\pi(i)})^2.$$

Expanding both sides, the squares $\sum_i a_i^2$ and $\sum_j c_j^2$ cancel, and what remains is $\sum_i a_i c_{n+1-i} \leq \sum_i a_i c_{\pi(i)}$. Hence

$$\sum_i (a_i - c_{\pi(i)})^2 = \sum_i a_i^2 + \sum_j c_j^2 - 2\sum_i a_i c_{\pi(i)} \;\leq\; \sum_i (a_i - c_{n+1-i})^2.$$

So the reversed pairing $i \mapsto n + 1 - i$ is the most expensive of all. By (a) some pairing costs at least $\|\mathbf{A} - \mathbf{B}\|_F^2$, so the most expensive one does too, which is the claim.

Exercise. C2: Each eigenvalue keeps its own disc

Let $\mathbf{A} \in M_n(\mathbb{C})$ be normal with **distinct** eigenvalues $\lambda_1, \dots, \lambda_n$, and let $\delta = \min_{i \neq j} |\lambda_i - \lambda_j|$. Let $\mathbf{B}$ be normal with $\|\mathbf{A} - \mathbf{B}\|_F < \delta/2$. Prove that each open disc $\{z : |z - \lambda_i| < \delta/2\}$ contains **exactly one** eigenvalue of $\mathbf{B}$, counted with multiplicity. Explain why Corollary 20.4.3 alone does not give this.

Solution. The discs are pairwise disjoint: if $|z - \lambda_i| < \delta/2$ and $|z - \lambda_j| < \delta/2$ with $i \neq j$, the triangle inequality (Theorem 0.6.4) gives $|\lambda_i - \lambda_j| < \delta$, contradicting the definition of δ. By Theorem 20.5.2 there is σ with $\sum_i |\lambda_i - \mu_{\sigma(i)}|^2 \leq \|\mathbf{A} - \mathbf{B}\|_F^2 < \delta^2/4$, so $|\lambda_i - \mu_{\sigma(i)}| < \delta/2$ for every i. Thus the entry $\mu_{\sigma(i)}$ of the list lies in the i-th disc, and, the discs being disjoint, in no other. Since σ is a bijection, every entry of the list $\mu_1, \dots, \mu_n$ is $\mu_{\sigma(i)}$ for exactly one i. So the i-th disc contains exactly one entry of the list, namely $\mu_{\sigma(i)}$, which is the claim.

Corollary 20.4.3 (a), with $\|\mathbf{A} - \mathbf{B}\|_2 \leq \|\mathbf{A} - \mathbf{B}\|_F < \delta/2$ (Proposition 16.3.4, or both norms being 0 if $\mathbf{A} = \mathbf{B}$, the case the proposition excludes), puts every eigenvalue of $\mathbf{B}$ in the union of the discs, but, as in the first example of this section, it allows two of them to fall in the same disc and leave another disc empty.

Exercise. C3: A Jordan block of any size

Let $n \geq 2$, $\varepsilon > 0$, $\mathbf{A} = \mathbf{J}_n(0)$, and $\mathbf{B} = \mathbf{J}_n(0) + \varepsilon \mathbf{E}_{n1}$, where $\mathbf{E}_{n1}$ is the matrix unit.

(a) Show that the characteristic polynomial of $\mathbf{B}$ is $x^n - \varepsilon$, and hence that every pairing of the eigenvalues of $\mathbf{A}$ and $\mathbf{B}$ costs $n\varepsilon^{2/n}$.

(b) Deduce that the Hoffman–Wielandt inequality fails for every pairing whenever $0 < \varepsilon < n^{n/(2n-2)}$, and that the ratio of the cost to $\|\mathbf{A} - \mathbf{B}\|_F^2$ tends to ∞ as $\varepsilon \to 0^+$.

Hint for (a): expand $\det(x\mathbf{I} - \mathbf{B})$ *along the first column.*

Solution.

(a) $x\mathbf{I} - \mathbf{B}$ has x on the diagonal, -1 on the superdiagonal, $-\varepsilon$ in position $(n, 1)$, and zeros elsewhere. Expanding along the first column, the $(1, 1)$ entry x multiplies the determinant of an upper triangular matrix with diagonal x, which is x^{n-1}. The $(n, 1)$ entry $-\varepsilon$ carries the sign $(-1)^{n+1}$ and multiplies the determinant of the matrix left after deleting row n and column 1, which is lower triangular with diagonal entries -1, hence has determinant $(-1)^{n-1}$. So

$$\det(x\mathbf{I} - \mathbf{B}) = x^n + (-1)^{n+1}(-\varepsilon)(-1)^{n-1} = x^n - \varepsilon.$$

Its roots are the n complex numbers z with $z^n = \varepsilon$, each of modulus $\varepsilon^{1/n}$ (these are Chapter 16's computation in Example (A Jordan block's eigenvalues move like $\varepsilon^{1/k}$)). Every eigenvalue of $\mathbf{A}$ is 0, so every pairing costs $\sum_i |0 - \mu_{\sigma(i)}|^2 = n\,\varepsilon^{2/n}$.

(b) $\mathbf{A} - \mathbf{B} = -\varepsilon \mathbf{E}_{n1}$, so $\|\mathbf{A} - \mathbf{B}\|_F^2 = \varepsilon^2$. The inequality fails for every pairing exactly when $n\varepsilon^{2/n} > \varepsilon^2$, that is, $\varepsilon^{2-2/n} < n$, that is, $\varepsilon < n^{1/(2-2/n)} = n^{n/(2n-2)}$; for $n = 2$ this is $\varepsilon < 2$, as in the text. The ratio is $n\varepsilon^{2/n}/\varepsilon^2 = n\varepsilon^{-(2n-2)/n}$, and the exponent $-(2n-2)/n$ is negative, so the ratio tends to ∞ as $\varepsilon \to 0^+$.

20.6 Hermitian Perturbations

For Hermitian matrices, Chapter 17 already gave the basic perturbation bound: each eigenvalue moves by at most $\|\mathbf{E}\|_2$, and no smaller constant is possible. This section asks what that bound does **not** tell us, and answers two such questions. The first is practical. A computation rarely hands us $\mathbf{E}$; it hands us an approximate eigenvector, and the question is how good the eigenvalue estimate it yields is. The answer is a residual bound, and when the eigenvalue is separated from the others it improves from linear to **quadratic** in the residual; that is the Kato–Temple theorem. The second concerns small eigenvalues. An absolute bound of size $\|\mathbf{E}\|_2$ says nothing about an eigenvalue much smaller than $\|\mathbf{E}\|_2$, and for perturbations of the form $\mathbf{A} \mapsto \mathbf{SAS}^*$ we prove a **relative** bound instead, Ostrowski's theorem, which is Sylvester's law of inertia made quantitative.

Throughout, $F = \mathbb{C}$, matrices whose eigenvalues are ordered are Hermitian, and $\lambda_1(\mathbf{A}) \geq \cdots \geq \lambda_n(\mathbf{A})$ are the eigenvalues of $\mathbf{A}$ in decreasing order, with multiplicity. Vectors are measured in $\|\cdot\|_2$, written $\|\cdot\|$ when no other norm is in sight, and $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{y}^*\mathbf{x}$.

20.6.1 What Weyl's bound leaves open

Recall Corollary 17.3.6: if $\mathbf{A}$ and $\mathbf{E}$ are Hermitian, then $|\lambda_i(\mathbf{A} + \mathbf{E}) - \lambda_i(\mathbf{A})| \leq \|\mathbf{E}\|_2$ for every i. By Corollary 17.3.7 the constant 1 cannot be lowered, since $\mathbf{E} = t\mathbf{I}$ moves every eigenvalue by exactly $t = \|\mathbf{E}\|_2$. As a bound on **absolute** error, for an **unknown** Hermitian perturbation of known size, there is nothing more to be said.

The two assumptions in bold are where there is room. First, in practice $\mathbf{E}$ is not known at all. An iterative method stops with a unit vector $\mathbf{x}$ and a number μ that it believes to be an eigenpair, and what can be measured is the **residual** $\mathbf{r} = \mathbf{Ax} - \mu\mathbf{x}$. Second, an absolute error bound is only informative for eigenvalues larger than the error. For the matrix $\operatorname{diag}(1, 10^{-6})$ and a Hermitian perturbation of size 10^{-2}, Weyl's bound allows the eigenvalue 10^{-6} to end up anywhere in an interval of radius 10^{-2}, including negative numbers. Whether that is a fair description depends on the kind of perturbation, and for an important kind it is not.

20.6.2 Residual bounds

Suppose $\mathbf{x}$ is a unit vector and μ a real number, and $\mathbf{r} = \mathbf{Ax} - \mu\mathbf{x}$ is small. If $\mathbf{r} = \mathbf{0}$, then μ is an eigenvalue. The next theorem says that if $\mathbf{r}$ is merely small, μ is **near** an eigenvalue, with the size of $\mathbf{r}$ as the distance.

Theorem 20.6.1. Residual Bound for Hermitian Matrices

Let $\mathbf{A} \in M_n(\mathbb{C})$ be Hermitian, let $\mathbf{x} \in \mathbb{C}^n$ with $\|\mathbf{x}\| = 1$, let $\mu \in \mathbb{R}$, and put $\mathbf{r} = \mathbf{Ax} - \mu\mathbf{x}$. Then some eigenvalue λ of $\mathbf{A}$ satisfies $|\lambda - \mu| \leq \|\mathbf{r}\|$; that is, the interval $[\mu - \|\mathbf{r}\|, \mu + \|\mathbf{r}\|]$ contains an eigenvalue of $\mathbf{A}$.

This is the normal case of §04's Theorem 20.4.4, since a Hermitian matrix is normal. We give the direct proof from the spectral theorem instead, because the expansion it uses is the one Kato–Temple needs.

Proof. By Corollary 12.4.3 there is an orthonormal basis $(\mathbf{q}_1, \ldots, \mathbf{q}_n)$ of $\mathbb{C}^n$ with $\mathbf{Aq}_i = \lambda_i \mathbf{q}_i$, where $\lambda_i = \lambda_i(\mathbf{A})$. Write $\mathbf{x} = \sum_i c_i \mathbf{q}_i$, with $\sum_i |c_i|^2 = \|\mathbf{x}\|^2 = 1$ by Theorem 11.2.1 (c). Then

$\mathbf{r} = \sum_i c_i(\lambda_i - \mu)\mathbf{q}_i$, and by the same result

$$\|\mathbf{r}\|^2 = \sum_{i=1}^{n} |c_i|^2(\lambda_i - \mu)^2 \;\geq\; \Big(\min_i (\lambda_i - \mu)^2\Big) \sum_{i=1}^{n} |c_i|^2 = \min_i (\lambda_i - \mu)^2.$$

Taking square roots, $\min_i |\lambda_i - \mu| \leq \|\mathbf{r}\|$, as claimed.

■

The bound holds for every real μ, and one choice of μ makes the residual smallest: the **Rayleigh quotient** $R_{\mathbf{A}}(\mathbf{x}) = \mathbf{x}^*\mathbf{A}\mathbf{x}$ of Definition 17.1.1, which for a unit vector needs no denominator. With that choice the residual is orthogonal to $\mathbf{x}$, since

$$\langle \mathbf{r}, \mathbf{x} \rangle = \mathbf{x}^*\mathbf{A}\mathbf{x} - \mu\, \mathbf{x}^*\mathbf{x} = \mu - \mu = 0,$$

and Pythagoras (Theorem 11.1.3) applied to $\mathbf{A}\mathbf{x} = \mu\mathbf{x} + \mathbf{r}$ gives

$$\|\mathbf{r}\|^2 = \|\mathbf{A}\mathbf{x}\|^2 - \mu^2 \qquad (\mu = R_{\mathbf{A}}(\mathbf{x}),\ \|\mathbf{x}\| = 1). \tag{20.6.1}$$

Exercise B3 shows that any other μ gives a larger residual. From now on, μ is the Rayleigh quotient and $\varepsilon = \|\mathbf{A}\mathbf{x} - \mu\mathbf{x}\|$ its residual.

Quick check. Let $\mathbf{A} = \operatorname{diag}(1, 3)$ and $\mathbf{x} = \frac{1}{\sqrt{2}}(1, 1)$. Compute $\mu = R_{\mathbf{A}}(\mathbf{x})$, the residual $\mathbf{r}$, and $\varepsilon = \|\mathbf{r}\|$. Which eigenvalues does the interval $[\mu - \varepsilon, \mu + \varepsilon]$ contain?

Solution. $\mathbf{A}\mathbf{x} = \frac{1}{\sqrt{2}}(1, 3)$, so $\mu = \mathbf{x}^*\mathbf{A}\mathbf{x} = \frac{1}{2}(1 + 3) = 2$ and $\mathbf{r} = \frac{1}{\sqrt{2}}(1 - 2, 3 - 2) = \frac{1}{\sqrt{2}}(-1, 1)$, with $\varepsilon = 1$. As a check, Eq. 20.6.1 gives $\|\mathbf{A}\mathbf{x}\|^2 - \mu^2 = 5 - 4 = 1$. The interval is $[1, 3]$, and it contains both eigenvalues, at its two ends. The bound is attained: a vector halfway between two eigenvectors is as bad an approximation as it can be.

20.6.3 The Kato–Temple bound

The residual bound is linear in ε, and in general no better is possible: in the Quick check, $|\lambda - \mu| = \varepsilon$ for both eigenvalues. But that example had $\mathbf{x}$ equally far from both eigenvectors. When $\mathbf{x}$ is close to one eigenvector, the Rayleigh quotient is a much better estimate than ε suggests. Exercise (C1: Why the quotient estimates eigenvalues so well) of Chapter 17 found that an eigenvector error of size δ costs only about δ^2 in the Rayleigh quotient. The trouble is that the eigenvector error cannot be measured, while the residual can. The next theorem restates the δ^2 phenomenon in terms of the residual, and the price is a hypothesis on how far the other eigenvalues are.

Here is where the gap comes in. If no eigenvalue of $\mathbf{A}$ lies strictly between two real numbers λ and β, then $(t - \lambda)(t - \beta) \geq 0$ at every eigenvalue t, so the matrix $(\mathbf{A} - \lambda\mathbf{I})(\mathbf{A} - \beta\mathbf{I})$ is positive semidefinite. Evaluating its quadratic form at $\mathbf{x}$ gives an inequality that involves only μ, ε, λ, and β.

Theorem 20.6.2. Kato–Temple Theorem

Let $\mathbf{A} \in M_n(\mathbb{C})$ be Hermitian, let $\mathbf{x} \in \mathbb{C}^n$ with $\|\mathbf{x}\| = 1$, let $\mu = \mathbf{x}^*\mathbf{A}\mathbf{x}$, and let $\varepsilon = \|\mathbf{A}\mathbf{x} - \mu\mathbf{x}\|$.

(a) Let $\beta > \mu$, and let $\lambda < \beta$ be an eigenvalue of $\mathbf{A}$ such that **no** eigenvalue of $\mathbf{A}$ lies in the open interval (λ, β). Then

$$\lambda \;\geq\; \mu - \frac{\varepsilon^2}{\beta - \mu}.$$

(b) Let $\alpha < \mu$, and let $\lambda > \alpha$ be an eigenvalue of $\mathbf{A}$ such that **no** eigenvalue of $\mathbf{A}$ lies in the open interval (α, λ). Then

$$\lambda \;\leq\; \mu + \frac{\varepsilon^2}{\mu - \alpha}.$$

In particular, if $\alpha < \mu < \beta$ and λ is the **only** eigenvalue of $\mathbf{A}$ in (α, β), then

$$\mu - \frac{\varepsilon^2}{\beta - \mu} \;\leq\; \lambda \;\leq\; \mu + \frac{\varepsilon^2}{\mu - \alpha}.$$

Idea. Compute the number $\mathbf{x}^*(\mathbf{A} - \lambda\mathbf{I})(\mathbf{A} - \beta\mathbf{I})\mathbf{x}$ two ways. ① In the eigenbasis it is $\sum_i |c_i|^2(\lambda_i - \lambda)(\lambda_i - \beta)$, and the gap hypothesis makes every term nonnegative: each eigenvalue lies on or outside the interval between λ and β. ② Expanding the product, it is $\|\mathbf{A}\mathbf{x}\|^2 - (\lambda + \beta)\mu + \lambda\beta$, and Eq. 20.6.1 turns this into $\varepsilon^2 + (\mu - \lambda)(\mu - \beta)$. So $(\mu - \lambda)(\beta - \mu) \leq \varepsilon^2$; divide by the positive number $\beta - \mu$. Part (b) is the same with α in place of β.

Proof.

(a) As in the proof of Theorem 20.6.1, write $\mathbf{x} = \sum_i c_i \mathbf{q}_i$ in an orthonormal eigenbasis, $\mathbf{A}\mathbf{q}_i = \lambda_i \mathbf{q}_i$, $\sum_i |c_i|^2 = 1$. Then $(\mathbf{A} - \lambda\mathbf{I})(\mathbf{A} - \beta\mathbf{I})\mathbf{x} = \sum_i c_i(\lambda_i - \lambda)(\lambda_i - \beta)\mathbf{q}_i$, and by Theorem 11.2.1 (b),

$$\mathbf{x}^*(\mathbf{A} - \lambda\mathbf{I})(\mathbf{A} - \beta\mathbf{I})\mathbf{x} = \sum_{i=1}^{n} |c_i|^2(\lambda_i - \lambda)(\lambda_i - \beta).$$

Each eigenvalue λ_i satisfies $\lambda_i \leq \lambda$ or $\lambda_i \geq \beta$, because none lies in (λ, β). In the first case both factors are ≤ 0, in the second both are ≥ 0, since $\lambda < \beta$. So every term is ≥ 0, and the left-hand side is ≥ 0.
On the other hand, expanding the product and using $\mathbf{x}^*\mathbf{x} = 1$, $\mathbf{x}^*\mathbf{A}\mathbf{x} = \mu$, and $\mathbf{x}^*\mathbf{A}^2\mathbf{x} = (\mathbf{A}\mathbf{x})^*(\mathbf{A}\mathbf{x}) = \|\mathbf{A}\mathbf{x}\|^2$, where $\mathbf{A}^* = \mathbf{A}$,

$$\begin{aligned}\mathbf{x}^*(\mathbf{A} - \lambda\mathbf{I})(\mathbf{A} - \beta\mathbf{I})\mathbf{x} &= \|\mathbf{A}\mathbf{x}\|^2 - (\lambda + \beta)\mu + \lambda\beta \\ &= \varepsilon^2 + \mu^2 - (\lambda + \beta)\mu + \lambda\beta \\ &= \varepsilon^2 + (\mu - \lambda)(\mu - \beta),\end{aligned}$$

where the second line is Eq. 20.6.1. Hence $\varepsilon^2 + (\mu - \lambda)(\mu - \beta) \geq 0$, that is, $(\mu - \lambda)(\beta - \mu) \leq \varepsilon^2$. Since $\beta - \mu > 0$, dividing gives $\mu - \lambda \leq \varepsilon^2/(\beta - \mu)$, which is (a).

(b) Apply (a) to $-\mathbf{A}$, the unit vector $\mathbf{x}$, the eigenvalue $-\lambda$, and the number $-\alpha$. The Rayleigh quotient of $-\mathbf{A}$ at $\mathbf{x}$ is $-\mu$, its residual is $-(\mathbf{A}\mathbf{x} - \mu\mathbf{x})$, of the same norm ε,

and $-\alpha > -\mu$. The eigenvalues of $-\mathbf{A}$ are the negatives of those of $\mathbf{A}$, so $-\lambda < -\alpha$ and none lies in $(-\lambda, -\alpha)$. Part (a) gives $-\lambda \geq -\mu - \varepsilon^2/(-\alpha + \mu)$, which is (b).

The final statement follows: if λ is the only eigenvalue in (α, β), then no eigenvalue lies in (λ, β) or in (α, λ), so (a) and (b) both apply. This proves the theorem.

■

Compare the two bounds on the error $|\lambda - \mu|$. The residual bound gives ε. Kato–Temple gives ε^2 divided by the distance from μ to the neighboring part of the spectrum. When ε is small compared with that gap, the square wins by a wide margin: a residual of 10^{-4} and a gap of 1 give an error of at most 10^{-8}. *The Rayleigh quotient's error is quadratic in the residual, provided there is a gap.* That is the reason the Rayleigh quotient is the eigenvalue estimate of choice, and Chapter 24 builds an iteration on it.

The theorem needs a number α or β that is known to lie beyond the neighboring eigenvalue, and in practice that information has to come from somewhere. The next example gets it from interlacing.

Example. Bracketing the top eigenvalue

Let

$$\mathbf{A} = \begin{pmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{pmatrix}, \qquad \mathbf{x} = \tfrac{1}{\sqrt{3}}(1, 1, 1).$$

Estimate $\lambda_1(\mathbf{A})$ from $\mathbf{x}$: compute μ and ε, the interval given by Theorem 20.6.1, and the interval given by Theorem 20.6.2, using only information that can be read off $\mathbf{A}$ without computing its eigenvalues. Compare with the true value $\lambda_1(\mathbf{A}) = 2 + \sqrt{2}$.

Solution. Rayleigh quotient and residual. $\mathbf{A}(1, 1, 1) = (3, 4, 3)$, so $\mathbf{Ax} = \frac{1}{\sqrt{3}}(3, 4, 3)$ and

$$\mu = \mathbf{x}^*\mathbf{Ax} = \tfrac{1}{3}(3 + 4 + 3) = \tfrac{10}{3}, \qquad \|\mathbf{Ax}\|^2 = \tfrac{1}{3}(9 + 16 + 9) = \tfrac{34}{3}.$$

By Eq. 20.6.1, $\varepsilon^2 = \frac{34}{3} - \frac{100}{9} = \frac{2}{9}$, so $\varepsilon = \frac{\sqrt{2}}{3} \approx 0.471$. Directly: $\mathbf{r} = \frac{1}{\sqrt{3}}(3 - \frac{10}{3}, 4 - \frac{10}{3}, 3 - \frac{10}{3}) = \frac{1}{3\sqrt{3}}(-1, 2, -1)$, of squared norm $\frac{6}{27} = \frac{2}{9}$.

Residual bound. Some eigenvalue lies in $[\frac{10}{3} - 0.471, \frac{10}{3} + 0.471] \approx [2.862, 3.805]$.

Kato–Temple. By Proposition 17.1.1, $\lambda_1(\mathbf{A}) \geq \mu = \frac{10}{3}$. For an upper bound we need $\alpha < \mu$ with no eigenvalue in (α, λ_1), which says $\lambda_2(\mathbf{A}) \leq \alpha$. Deleting the second row and column leaves $\mathbf{B} = \operatorname{diag}(2, 2)$, and Cauchy's interlacing theorem (Theorem 17.4.2, with $I = \{1, 3\}$) gives $\lambda_2(\mathbf{A}) \leq \lambda_1(\mathbf{B}) = 2$. So $\alpha = 2$ is allowed: $\alpha < \mu$, and no eigenvalue lies in $(2, \lambda_1)$, because $\lambda_2 \leq 2$ and $\lambda_1 \geq \frac{10}{3} > 2$. Part (b) gives

$$\lambda_1(\mathbf{A}) \;\leq\; \frac{10}{3} + \frac{2/9}{10/3 - 2} = \frac{10}{3} + \frac{2}{9} \cdot \frac{3}{4} = \frac{10}{3} + \frac{1}{6} = \frac{7}{2}.$$

So $\frac{10}{3} \leq \lambda_1(\mathbf{A}) \leq \frac{7}{2}$, an interval of width $\frac{1}{6} \approx 0.167$, against width $2\varepsilon \approx 0.943$ for the residual bound.

Check. The true value is $2 + \sqrt{2} \approx 3.414$, which lies in both intervals; the other eigenvalues

are 2 and $2-\sqrt{2}$. The actual error of the Rayleigh quotient is $2+\sqrt{2}-\frac{10}{3}\approx 0.081$, about a sixth of ε.

Kato–Temple cannot be improved in general: for $\mathbf{A}=\operatorname{diag}(\lambda_1,\lambda_2)$ with $\lambda_1>\lambda_2$, and $\alpha=\lambda_2$ the other eigenvalue, the upper bound in (b) is an **equality** at every real unit $\mathbf{x}$ that is not an eigenvector for λ_2 (Exercise C3). At such an eigenvector the hypothesis $\alpha<\mu$ fails, since there $\mu=\lambda_2$.

Warning. **The gap hypothesis is not decoration.** Take $\mathbf{A}=\operatorname{diag}(1,0)$ and $\mathbf{x}=\frac{1}{\sqrt{2}}(1,1)$, so $\mu=\frac{1}{2}$ and $\mathbf{r}=\frac{1}{\sqrt{2}}(\frac{1}{2},-\frac{1}{2})$, $\varepsilon^2=\frac{1}{4}$. With $\lambda=1$ and $\alpha=0$, part (b) gives $1\le\frac{1}{2}+\frac{1/4}{1/2}=1$, true with equality. With $\alpha=-\frac{1}{2}$ instead, the eigenvalue 0 lies in (α,λ), the hypothesis fails, and the formula would claim $1\le\frac{1}{2}+\frac{1/4}{1}=\frac{3}{4}$, which is false. A number α chosen too far from μ, past an unseen eigenvalue, gives a bound that is simply wrong.

20.6.4 Relative perturbations and congruence

Now the second question: small eigenvalues. Weyl's bound is absolute, and an absolute bound is useless for an eigenvalue smaller than itself. Some perturbations deserve better. In a change of variables $\mathbf{x}=\mathbf{S}^*\mathbf{y}$ with $\mathbf{S}$ invertible, the quadratic form $\mathbf{x}^*\mathbf{A}\mathbf{x}$ becomes $\mathbf{y}^*(\mathbf{S}\mathbf{A}\mathbf{S}^*)\mathbf{y}$; the matrix undergoes a **congruence**, as in Chapter 14 and in Chapter 17 §10. When $\mathbf{S}$ is close to $\mathbf{I}$, for instance when it comes from a slightly inaccurate elimination step or from rounding in a factor, $\mathbf{S}\mathbf{A}\mathbf{S}^*$ is a perturbation of $\mathbf{A}$, but a very structured one. Its effect on the eigenvalues is **multiplicative**.

Two facts are known already. By Theorem 17.10.1 (b), applied with $\mathbf{P}=\mathbf{S}^*$, a congruence preserves the numbers of positive, negative, and zero eigenvalues; it cannot change a sign. And by Chapter 14 §05's warning, it can still move eigenvalues a long way: with $\mathbf{P}=\operatorname{diag}(2,3)$, $\mathbf{P}^{\top}\operatorname{diag}(1,-1)\mathbf{P}=\operatorname{diag}(4,-9)$. Ostrowski's theorem measures how far.

Theorem 20.6.3. Ostrowski's Theorem

Let $\mathbf{A}\in M_n(\mathbb{C})$ be Hermitian and let $\mathbf{S}\in M_n(\mathbb{C})$ be invertible. Then $\mathbf{S}\mathbf{S}^*$ is positive definite, and for each $k=1,\dots,n$ there is a real number θ_k with

$$\lambda_k(\mathbf{S}\mathbf{A}\mathbf{S}^*)=\theta_k\,\lambda_k(\mathbf{A}),\qquad \lambda_n(\mathbf{S}\mathbf{S}^*)\ \le\ \theta_k\ \le\ \lambda_1(\mathbf{S}\mathbf{S}^*).$$

In particular $\theta_k>0$, and

$$\bigl|\lambda_k(\mathbf{S}\mathbf{A}\mathbf{S}^*)-\lambda_k(\mathbf{A})\bigr|\ \le\ \|\mathbf{S}\mathbf{S}^*-\mathbf{I}\|_2\,|\lambda_k(\mathbf{A})|.$$

Idea. **Step roadmap.** Fix k and put $\mu=\lambda_k(\mathbf{A})$.

① **Shift, so the k-th eigenvalue is 0.** $\mathbf{A}-\mu\mathbf{I}$ has at most $k-1$ positive eigenvalues and at least k nonnegative ones. A congruence preserves both counts (Sylvester's law), so $\mathbf{S}(\mathbf{A}-\mu\mathbf{I})\mathbf{S}^*$ has the same property, and that forces its k-th eigenvalue to be exactly 0.

② **Split.** $\mathbf{S}\mathbf{A}\mathbf{S}^*=\mathbf{S}(\mathbf{A}-\mu\mathbf{I})\mathbf{S}^*+\mu\mathbf{S}\mathbf{S}^*$.

③ **Weyl.** Adding $\mu\mathbf{S}\mathbf{S}^*$ moves the k-th eigenvalue from 0 to somewhere between the small-

est and the largest eigenvalue of $\mu\mathbf{S}\mathbf{S}^*$, which are $\mu\lambda_n(\mathbf{S}\mathbf{S}^*)$ and $\mu\lambda_1(\mathbf{S}\mathbf{S}^*)$ in some order, depending on the sign of μ.

Proof. For $\mathbf{x} \neq \mathbf{0}$, $\mathbf{x}^*\mathbf{S}\mathbf{S}^*\mathbf{x} = \|\mathbf{S}^*\mathbf{x}\|^2 > 0$, because $\mathbf{S}^*$ is invertible. So $\mathbf{S}\mathbf{S}^*$, which is Hermitian, is positive definite, and its eigenvalues are positive by Theorem 13.1.5; write $m = \lambda_n(\mathbf{S}\mathbf{S}^*) > 0$ and $M = \lambda_1(\mathbf{S}\mathbf{S}^*)$.

Fix k and put $\mu = \lambda_k(\mathbf{A})$. The eigenvalues of the Hermitian matrix $\mathbf{A} - \mu\mathbf{I}$ are $\lambda_i(\mathbf{A}) - \mu$, in the same decreasing order. Those with $i \geq k$ are ≤ 0, so at most $k-1$ of them are positive; those with $i \leq k$ are ≥ 0, so at least k are nonnegative. In the notation of Definition 17.10.1, $n_+(\mathbf{A} - \mu\mathbf{I}) \leq k-1$ and $n_+(\mathbf{A} - \mu\mathbf{I}) + n_0(\mathbf{A} - \mu\mathbf{I}) \geq k$. Put $\mathbf{Y} = \mathbf{S}(\mathbf{A} - \mu\mathbf{I})\mathbf{S}^*$. This is $\mathbf{P}^*(\mathbf{A} - \mu\mathbf{I})\mathbf{P}$ with $\mathbf{P} = \mathbf{S}^*$ invertible, so by Theorem 17.10.1 (b), $\mathbf{Y}$ is Hermitian with the same inertia triple. Hence $\mathbf{Y}$ has at most $k-1$ positive eigenvalues, so $\lambda_k(\mathbf{Y}) \leq 0$; and at least k nonnegative ones, so $\lambda_k(\mathbf{Y}) \geq 0$. Thus $\lambda_k(\mathbf{Y}) = 0$.

Now $\mathbf{S}\mathbf{A}\mathbf{S}^* = \mathbf{Y} + \mu\mathbf{S}\mathbf{S}^*$, a sum of two Hermitian matrices. By Corollary 17.3.4,

$$\lambda_n(\mu\mathbf{S}\mathbf{S}^*) \;\leq\; \lambda_k(\mathbf{S}\mathbf{A}\mathbf{S}^*) - \lambda_k(\mathbf{Y}) \;\leq\; \lambda_1(\mu\mathbf{S}\mathbf{S}^*),$$

and $\lambda_k(\mathbf{Y}) = 0$. If $\mu \geq 0$, multiplying by μ keeps the order of the eigenvalues, so $\lambda_n(\mu\mathbf{S}\mathbf{S}^*) = \mu m$ and $\lambda_1(\mu\mathbf{S}\mathbf{S}^*) = \mu M$. If $\mu < 0$, then $\mu\mathbf{S}\mathbf{S}^* = -(|\mu|\mathbf{S}\mathbf{S}^*)$, and by Lemma 17.3.2 its extreme eigenvalues are $\lambda_n = -|\mu|M = \mu M$ and $\lambda_1 = -|\mu|m = \mu m$. In both cases $\lambda_k(\mathbf{S}\mathbf{A}\mathbf{S}^*)$ lies between μm and μM. If $\mu \neq 0$, put $\theta_k = \lambda_k(\mathbf{S}\mathbf{A}\mathbf{S}^*)/\mu$; dividing by μ (and reversing the inequalities when $\mu < 0$) gives $m \leq \theta_k \leq M$. If $\mu = 0$, both ends are 0, so $\lambda_k(\mathbf{S}\mathbf{A}\mathbf{S}^*) = 0$, and $\theta_k = m$ will do. In either case $\theta_k \geq m > 0$.

For the last inequality, $|\lambda_k(\mathbf{S}\mathbf{A}\mathbf{S}^*) - \lambda_k(\mathbf{A})| = |\theta_k - 1|\,|\mu|$, and $\theta_k - 1$ lies in $[m-1, M-1]$. The numbers $m-1$ and $M-1$ are the smallest and largest eigenvalues of the Hermitian matrix $\mathbf{S}\mathbf{S}^* - \mathbf{I}$, so by Lemma 17.3.5 both have absolute value at most $\|\mathbf{S}\mathbf{S}^* - \mathbf{I}\|_2$, and so does everything between them. This proves the theorem.

■

This is a quantitative Sylvester's law of inertia. Sylvester's law says that a congruence cannot change the sign of an eigenvalue; Ostrowski's theorem says that it multiplies the k-th eigenvalue by a factor θ_k that is always positive and always between the extreme eigenvalues of $\mathbf{S}\mathbf{S}^*$. These are $\lambda_1(\mathbf{S}\mathbf{S}^*) = \max_{\|\mathbf{y}\|=1}\|\mathbf{S}^*\mathbf{y}\|^2$ and $\lambda_n(\mathbf{S}\mathbf{S}^*) = \min_{\|\mathbf{y}\|=1}\|\mathbf{S}^*\mathbf{y}\|^2$, by Proposition 17.1.1, since $\mathbf{y}^*\mathbf{S}\mathbf{S}^*\mathbf{y} = \|\mathbf{S}^*\mathbf{y}\|^2$: the most and the least that the change of variables stretches a vector, squared. When $\mathbf{S}$ is close to a unitary matrix, both are close to 1, and **every** eigenvalue, however small, keeps most of its significant digits.

Quick check. Chapter 14 §05's warning used $\mathbf{A} = \operatorname{diag}(1, -1)$ and $\mathbf{P} = \operatorname{diag}(2, 3)$, with $\mathbf{P}^\top\mathbf{A}\mathbf{P} = \operatorname{diag}(4, -9)$. Take $\mathbf{S} = \mathbf{P}^\top = \mathbf{P}$. Find θ_1, θ_2, and the interval $[\lambda_n(\mathbf{S}\mathbf{S}^*), \lambda_1(\mathbf{S}\mathbf{S}^*)]$. Can the interval in Theorem 20.6.3 be shortened?

Solution. $\mathbf{S}\mathbf{S}^* = \operatorname{diag}(4, 9)$, so the interval is $[4, 9]$. The eigenvalues of $\mathbf{S}\mathbf{A}\mathbf{S}^* = \operatorname{diag}(4, -9)$ in decreasing order are 4 and -9, and those of $\mathbf{A}$ are 1 and -1. So $\theta_1 = 4/1 = 4$ and $\theta_2 = -9/(-1) = 9$. They sit at the two ends of the interval, so neither end can be moved

inwards: the interval is sharp.

The next example shows the difference between the absolute and the relative view on a matrix whose eigenvalues differ greatly in size.

Example. A tiny eigenvalue under a change of variables

Let $\mathbf{A} = \operatorname{diag}(1, 10^{-6})$ and $\mathbf{S} = \left(\begin{smallmatrix} 1 & 0 \\ t & 1 \end{smallmatrix}\right)$ with $t = 10^{-2}$, the change of variables that adds t times the first coordinate to the second. Compute $\mathbf{SAS}^*$ and $\mathbf{E} = \mathbf{SAS}^* - \mathbf{A}$. What do Corollary 17.3.4 and Theorem 20.6.3 say about $\lambda_2(\mathbf{SAS}^*)$, and what is its true value?

Solution. Write $\eta = 10^{-6}$. Multiplying out,

$$\mathbf{SAS}^* = \begin{pmatrix} 1 & t \\ t & t^2 + \eta \end{pmatrix}, \qquad \mathbf{E} = \begin{pmatrix} 0 & t \\ t & t^2 \end{pmatrix},$$

with $t^2 + \eta = 1.01 \times 10^{-4}$.

Weyl. $\mathbf{E}$ has trace t^2 and determinant $-t^2$, so its eigenvalues are $\frac{1}{2}(t^2 \pm t\sqrt{t^2+4})$, approximately 0.01005 and -0.00995. By Corollary 17.3.4,

$$10^{-6} - 0.00995 \;\le\; \lambda_2(\mathbf{SAS}^*) \;\le\; 10^{-6} + 0.01005.$$

This interval is about 10^4 times wider than the eigenvalue it is meant to locate, and it contains negative numbers.

Ostrowski. $\mathbf{SS}^* = \left(\begin{smallmatrix} 1 & t \\ t & 1+t^2 \end{smallmatrix}\right)$ has trace $2 + t^2$ and determinant 1, so its eigenvalues are $\frac{1}{2}(2 + t^2 \pm t\sqrt{4+t^2})$, approximately 1.01005 and 0.99005. By Theorem 20.6.3,

$$0.99005 \times 10^{-6} \;\le\; \lambda_2(\mathbf{SAS}^*) \;\le\; 1.01005 \times 10^{-6}\,:$$

the eigenvalue is positive and known to within about 1%.

The truth. $\det(\mathbf{SAS}^*) = \det(\mathbf{S})^2 \det \mathbf{A} = \eta$, since $\det \mathbf{S} = 1$, and the trace is $1 + t^2 + \eta$. So the two eigenvalues multiply to η, and the larger is $1.0001\ldots$; hence $\lambda_2(\mathbf{SAS}^*) = \eta/\lambda_1(\mathbf{SAS}^*) \approx 0.99990 \times 10^{-6}$. The relative change is about 10^{-4}, well inside Ostrowski's 1%; Weyl's bound, by contrast, cannot even decide the sign.

Warning. **Weyl's bound is sharp and still useless for small eigenvalues.** It is the best possible bound on **absolute** error for **arbitrary** Hermitian $\mathbf{E}$ of a given norm, and for exactly that reason it cannot distinguish an eigenvalue of size 10^{-6} from 0 once $\|\mathbf{E}\|_2 \ge 10^{-6}$. A general Hermitian $\mathbf{E}$ really can destroy a small eigenvalue: $\mathbf{E} = -10^{-6}\mathbf{e}_2\mathbf{e}_2^{\top}$ sends the eigenvalue 10^{-6} of $\operatorname{diag}(1, 10^{-6})$ to 0. Relative accuracy needs a structured perturbation, such as a congruence $\mathbf{SAS}^*$ with $\mathbf{S}$ near $\mathbf{I}$.

20.6.5 Exercises

20.6.5.1 A. Check your understanding

Exercise. A1

(a) State the Kato–Temple theorem, with its hypotheses.

(b) True or false: for Hermitian $\mathbf{A}$, a unit vector $\mathbf{x}$, and real μ, the eigenvalue of $\mathbf{A}$ **closest** to μ is within $\|\mathbf{A}\mathbf{x} - \mu\mathbf{x}\|$ of μ. Justify your answer.

(c) True or false: for Hermitian $\mathbf{A}$ and invertible $\mathbf{S}$, $\lambda_k(\mathbf{S}\mathbf{A}\mathbf{S}^*)$ and $\lambda_k(\mathbf{A})$ have the same sign for every k. Justify your answer.

(d) Which of the three bounds of this section need information that is not contained in $\mathbf{A}$, $\mathbf{x}$, and μ? What information?

(e) Why can the constant 1 in Corollary 17.3.6 not be lowered, and why does that not contradict Ostrowski's theorem?

Solution.

(a) Let $\mathbf{A}$ be Hermitian, $\mathbf{x}$ a unit vector, $\mu = \mathbf{x}^*\mathbf{A}\mathbf{x}$, $\varepsilon = \|\mathbf{A}\mathbf{x} - \mu\mathbf{x}\|$. If $\beta > \mu$ and $\lambda < \beta$ is an eigenvalue with no eigenvalue in (λ, β), then $\lambda \geq \mu - \varepsilon^2/(\beta - \mu)$. If $\alpha < \mu$ and $\lambda > \alpha$ is an eigenvalue with no eigenvalue in (α, λ), then $\lambda \leq \mu + \varepsilon^2/(\mu - \alpha)$.

(b) True. Theorem 20.6.1 gives some eigenvalue within $\|\mathbf{A}\mathbf{x} - \mu\mathbf{x}\|$ of μ, and the closest one is at least as close.

(c) True. By Theorem 20.6.3, $\lambda_k(\mathbf{S}\mathbf{A}\mathbf{S}^*) = \theta_k\lambda_k(\mathbf{A})$ with $\theta_k > 0$; a positive multiple has the same sign, and 0 goes to 0.

(d) The residual bound needs nothing more. Kato–Temple needs a number α or β beyond the neighboring eigenvalue, that is, a lower bound on the gap; in Example (Bracketing the top eigenvalue) it came from interlacing. Ostrowski's theorem is not about an approximate eigenvector, but it needs $\mathbf{S}$, or at least the extreme eigenvalues of $\mathbf{S}\mathbf{S}^*$.

(e) $\mathbf{E} = t\mathbf{I}$ moves every eigenvalue by exactly $t = \|\mathbf{E}\|_2$. Ostrowski's theorem is about the perturbations $\mathbf{S}\mathbf{A}\mathbf{S}^* - \mathbf{A}$, which form a special class; $\mathbf{A} + t\mathbf{I}$ is not of the form $\mathbf{S}\mathbf{A}\mathbf{S}^*$ in general (for $\mathbf{A} = \mathbf{O}$ it never is, when $t \neq 0$). A better bound for a smaller class contradicts nothing.

20.6.5.2 B. Practice

Exercise. B1: Three bounds for one estimate

Let $\mathbf{A} = \left(\begin{smallmatrix} 5 & 2 \\ 2 & 2 \end{smallmatrix}\right)$ and $\mathbf{x} = \frac{1}{\sqrt{10}}(3, 1)$.

(a) Compute $\mu = \mathbf{x}^*\mathbf{A}\mathbf{x}$ and $\varepsilon = \|\mathbf{A}\mathbf{x} - \mu\mathbf{x}\|$, and the interval given by Theorem 20.6.1.

(b) Using Theorem 17.4.2 to bound $\lambda_2(\mathbf{A})$, apply Theorem 20.6.2 to bracket $\lambda_1(\mathbf{A})$.

(c) Find the eigenvalues of $\mathbf{A}$ exactly. Hence compare the actual error of μ with ε and with the Kato–Temple bound.

Solution.

(a) $\mathbf{A}(3,1) = (17,8)$, so $\mu = \frac{1}{10}(3 \cdot 17 + 1 \cdot 8) = \frac{59}{10} = 5.9$. The residual is $\frac{1}{\sqrt{10}}((17,8) - 5.9(3,1)) = \frac{1}{\sqrt{10}}(-0.7, 2.1)$, so $\varepsilon^2 = \frac{1}{10}(0.49+4.41) = 0.49$ and $\varepsilon = 0.7$. Some eigenvalue lies in $[5.2, 6.6]$.

(b) Deleting the first row and column leaves (2), so Theorem 17.4.2 gives $\lambda_2(\mathbf{A}) \le 2$. Also $\lambda_1(\mathbf{A}) \ge \mu = 5.9$ by Proposition 17.1.1. So with $\alpha = 2 < \mu$, no eigenvalue lies in $(2, \lambda_1)$, and part (b) of Theorem 20.6.2 gives

$$\lambda_1(\mathbf{A}) \;\le\; 5.9 + \frac{0.49}{5.9 - 2} = 5.9 + \frac{0.49}{3.9} \approx 6.026.$$

Hence $5.9 \le \lambda_1(\mathbf{A}) \le 6.026$.

(c) $\mathbf{A}$ has trace 7 and determinant 6, so its eigenvalues are 6 and 1; indeed $\mathbf{A}(2,1) = (12,6)$. The actual error of μ is $6 - 5.9 = 0.1$. The residual bound allows 0.7, seven times too much; Kato–Temple allows 0.126, within about 26% of the truth.

Exercise. B2: Ostrowski's factors

Let $\mathbf{A} = \operatorname{diag}(2, -1)$ and $\mathbf{S} = \left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$. Compute $\mathbf{SAS}^*$ and its eigenvalues, the eigenvalues of $\mathbf{SS}^*$, and the factors θ_1, θ_2 of Theorem 20.6.3. Hence verify the theorem for this pair.

Solution. $\mathbf{AS}^* = \operatorname{diag}(2,-1)\left(\begin{smallmatrix} 1 & 0 \\ 1 & 1 \end{smallmatrix}\right) = \left(\begin{smallmatrix} 2 & 0 \\ -1 & -1 \end{smallmatrix}\right)$, and

$$\mathbf{SAS}^* = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 2 & 0 \\ -1 & -1 \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ -1 & -1 \end{pmatrix},$$

with trace 0 and determinant -2, so eigenvalues $\pm\sqrt{2}$. Next $\mathbf{SS}^* = \left(\begin{smallmatrix} 2 & 1 \\ 1 & 1 \end{smallmatrix}\right)$, with trace 3 and determinant 1, so eigenvalues $\frac{3 \pm \sqrt{5}}{2} \approx 2.618$ and 0.382.

Then $\theta_1 = \lambda_1(\mathbf{SAS}^*)/\lambda_1(\mathbf{A}) = \sqrt{2}/2 \approx 0.707$ and $\theta_2 = (-\sqrt{2})/(-1) = \sqrt{2} \approx 1.414$. Both lie in $[0.382, 2.618]$, as the theorem says, and both are positive, so the signs $+, -$ are preserved.

Exercise. B3: The Rayleigh quotient minimizes the residual

Let $\mathbf{A}$ be Hermitian and $\mathbf{x}$ a unit vector, with $\rho = \mathbf{x}^*\mathbf{A}\mathbf{x}$. Prove that for every real ν,

$$\|\mathbf{A}\mathbf{x} - \nu\mathbf{x}\|^2 = \|\mathbf{A}\mathbf{x} - \rho\mathbf{x}\|^2 + (\nu - \rho)^2.$$

Hence deduce that among all real ν, the residual bound of Theorem 20.6.1 is sharpest at $\nu = \rho$.

Solution. Write $\mathbf{A}\mathbf{x} - \nu\mathbf{x} = (\mathbf{A}\mathbf{x} - \rho\mathbf{x}) + (\rho - \nu)\mathbf{x}$. The first vector is orthogonal to $\mathbf{x}$, since $\langle \mathbf{A}\mathbf{x} - \rho\mathbf{x}, \mathbf{x}\rangle = \mathbf{x}^*\mathbf{A}\mathbf{x} - \rho = 0$. By Theorem 11.1.3,

$$\|\mathbf{A}\mathbf{x} - \nu\mathbf{x}\|^2 = \|\mathbf{A}\mathbf{x} - \rho\mathbf{x}\|^2 + (\rho - \nu)^2\|\mathbf{x}\|^2 = \|\mathbf{A}\mathbf{x} - \rho\mathbf{x}\|^2 + (\nu - \rho)^2.$$

Hence $\|\mathbf{A}\mathbf{x} - \nu\mathbf{x}\| \ge \|\mathbf{A}\mathbf{x} - \rho\mathbf{x}\|$, with equality only for $\nu = \rho$. The interval $[\nu - \|\mathbf{A}\mathbf{x} - \nu\mathbf{x}\|, \nu + \|\mathbf{A}\mathbf{x} - \nu\mathbf{x}\|]$ of Theorem 20.6.1 therefore has its smallest radius at $\nu = \rho$.

20.6.5.3 C. Going deeper

Exercise. C1: A residual that guarantees an eigenvalue

Let $\mathbf{A}$ be Hermitian, $\mathbf{x}$ a unit vector, $\mu = \mathbf{x}^*\mathbf{A}\mathbf{x}$, and $\varepsilon = \|\mathbf{A}\mathbf{x} - \mu\mathbf{x}\|$.

(a) Let $\alpha < \mu < \beta$. Prove that if $\varepsilon^2 < (\mu - \alpha)(\beta - \mu)$, then $\mathbf{A}$ has an eigenvalue in (α, β).

(b) Deduce that $\mathbf{A}$ has an eigenvalue in $[\mu - \varepsilon, \mu + \varepsilon]$, which is Theorem 20.6.1 again, for this choice of μ.

Hint for (a): if there is none, consider $\mathbf{x}^*(\mathbf{A} - \alpha\mathbf{I})(\mathbf{A} - \beta\mathbf{I})\mathbf{x}$.

Solution.

(a) Suppose no eigenvalue lies in (α, β). Then every eigenvalue t satisfies $t \le \alpha$ or $t \ge \beta$, so $(t - \alpha)(t - \beta) \ge 0$. As in the proof of Theorem 20.6.2, with α in place of λ,

$$0 \;\le\; \mathbf{x}^*(\mathbf{A} - \alpha\mathbf{I})(\mathbf{A} - \beta\mathbf{I})\mathbf{x} = \varepsilon^2 + (\mu - \alpha)(\mu - \beta),$$

so $\varepsilon^2 \ge (\mu - \alpha)(\beta - \mu)$, contradicting the hypothesis. Hence some eigenvalue lies in (α, β).

(b) Let d be the smallest distance from μ to an eigenvalue of $\mathbf{A}$, and suppose $d > \varepsilon$. Choose s with $\varepsilon < s < d$, and apply (a) with $\alpha = \mu - s$ and $\beta = \mu + s$: the hypothesis holds, since $(\mu - \alpha)(\beta - \mu) = s^2 > \varepsilon^2$. So some eigenvalue lies in $(\mu - s, \mu + s)$, at distance less than $s < d$ from μ, contradicting the choice of d. Hence $d \le \varepsilon$, which is the claim.

Exercise. C2: Relative bounds need structured perturbations

(a) Let $\mathbf{A}$ be Hermitian and $\mathbf{S} = \mathbf{I} + \mathbf{F}$ with $\mathbf{F} \in M_n(\mathbb{C})$. Prove that

$$|\lambda_k(\mathbf{S}\mathbf{A}\mathbf{S}^*) - \lambda_k(\mathbf{A})| \;\le\; (2\|\mathbf{F}\|_2 + \|\mathbf{F}\|_2^2)\,|\lambda_k(\mathbf{A})|$$

for every k, whenever $\mathbf{S}$ is invertible.

(b) Show that there is **no** constant c such that $|\lambda_k(\mathbf{A} + \mathbf{E}) - \lambda_k(\mathbf{A})| \le c\|\mathbf{E}\|_2|\lambda_k(\mathbf{A})|$ for all Hermitian $\mathbf{A}$, $\mathbf{E}$ and all k.

Hint for (a): expand $\mathbf{S}\mathbf{S}^* - \mathbf{I}$, *and show first that* $\|\mathbf{F}^*\|_2 = \|\mathbf{F}\|_2$.

Solution.

(a) First, $\|\mathbf{F}^*\|_2 \le \|\mathbf{F}\|_2$: for every $\mathbf{x}$, by Theorem 11.1.5 and Theorem 16.3.2 (a),

$$\|\mathbf{F}^*\mathbf{x}\|^2 = \langle \mathbf{F}\mathbf{F}^*\mathbf{x}, \mathbf{x}\rangle \le \|\mathbf{F}\mathbf{F}^*\mathbf{x}\|\,\|\mathbf{x}\| \le \|\mathbf{F}\|_2\|\mathbf{F}^*\mathbf{x}\|\,\|\mathbf{x}\|,$$

so $\|\mathbf{F}^*\mathbf{x}\| \le \|\mathbf{F}\|_2\|\mathbf{x}\|$ (trivially if $\mathbf{F}^*\mathbf{x} = \mathbf{0}$), and the reverse inequality follows by applying this to $\mathbf{F}^*$, since $(\mathbf{F}^*)^* = \mathbf{F}$. Now $\mathbf{S}\mathbf{S}^* - \mathbf{I} = \mathbf{F} + \mathbf{F}^* + \mathbf{F}\mathbf{F}^*$, so by the triangle inequality for $\|\cdot\|_2$, which is a norm by Theorem 16.3.2 (b), and by part (d) of the same theorem,

$$\|\mathbf{S}\mathbf{S}^* - \mathbf{I}\|_2 \le \|\mathbf{F}\|_2 + \|\mathbf{F}^*\|_2 + \|\mathbf{F}\|_2\|\mathbf{F}^*\|_2 = 2\|\mathbf{F}\|_2 + \|\mathbf{F}\|_2^2.$$

The last inequality of Theorem 20.6.3 then gives the claim.

(b) Take $\mathbf{A} = \operatorname{diag}(1, 0)$ and $\mathbf{E} = \operatorname{diag}(0, t)$ with $t > 0$. Then $\lambda_2(\mathbf{A}) = 0$ and $\lambda_2(\mathbf{A} + \mathbf{E}) = \min(1, t) > 0$, so the left side is positive while the right side is $c\,t \cdot 0 = 0$. No constant c works. By (a), such an $\mathbf{E}$ is not of the form $\mathbf{SAS}^* - \mathbf{A}$: congruence can never move the eigenvalue 0.

Exercise. C3: Kato–Temple is sharp

Let $\lambda_1 > \lambda_2$ be real, $\mathbf{A} = \operatorname{diag}(\lambda_1, \lambda_2)$, and $\mathbf{x} = (\cos\phi, \sin\phi)$ with $0 < \phi < \pi/2$.

(a) Compute $\mu = \mathbf{x}^*\mathbf{A}\mathbf{x}$ and show that $\varepsilon^2 = (\lambda_1 - \lambda_2)^2 \cos^2\phi \sin^2\phi$.

(b) Show that Theorem 20.6.2 (b), with $\lambda = \lambda_1$ and $\alpha = \lambda_2$, holds with **equality**. Hence the constant in the Kato–Temple bound cannot be improved.

Solution. Write $c = \cos\phi$, $s = \sin\phi$, and $d = \lambda_1 - \lambda_2 > 0$.

(a) $\mathbf{A}\mathbf{x} = (\lambda_1 c, \lambda_2 s)$, so $\mu = \lambda_1 c^2 + \lambda_2 s^2$. Using $c^2 + s^2 = 1$, $\lambda_1 - \mu = ds^2$ and $\lambda_2 - \mu = -dc^2$. The residual is $\mathbf{A}\mathbf{x} - \mu\mathbf{x} = ((\lambda_1 - \mu)c, (\lambda_2 - \mu)s) = (ds^2 c, -dc^2 s)$, so

$$\varepsilon^2 = d^2 s^4 c^2 + d^2 c^4 s^2 = d^2 c^2 s^2 (s^2 + c^2) = d^2 c^2 s^2.$$

(b) Here $\alpha = \lambda_2 < \mu$, since $\mu - \lambda_2 = dc^2 > 0$, and no eigenvalue lies in (λ_2, λ_1). The bound reads

$$\mu + \frac{\varepsilon^2}{\mu - \lambda_2} = \mu + \frac{d^2 c^2 s^2}{dc^2} = \mu + ds^2 = \lambda_1.$$

So equality holds for every $\phi \in (0, \pi/2)$, and no smaller multiple of $\varepsilon^2/(\mu - \alpha)$ would give a valid bound.

20.7 The Sylvester Equation Revisited

Chapter 12 settled when the equation $\mathbf{AX} - \mathbf{XB} = \mathbf{C}$ can be solved: over $\mathbb{C}$ it has exactly one solution for every $\mathbf{C}$ if and only if $\mathbf{A}$ and $\mathbf{B}$ have no common eigenvalue (Theorem 12.10.2). It then set a question aside. "The quantitative version — *how badly* conditioned the equation is when the spectra are close but disjoint — is a separate matter, and belongs with perturbation theory in Chapter 20." Chapter 8, which first wrote the equation as a linear system, said the same thing more briefly: "its quantitative form appears in Chapter 20." This section pays both debts. It attaches to each pair $(\mathbf{A}, \mathbf{B})$ a number that says how large a solution can be, computes it exactly for normal pairs, shows how badly it can behave otherwise, and proves a bound in the spectral norm for Hermitian pairs whose spectra sit on opposite sides of a gap.

The reason to want these bounds now is that Sections 9 and 10 need them. In Section 9 the error in a computed invariant subspace turns out to satisfy a Sylvester equation, and in Section 10 so does the error in a square root or a polar factor. Each of those three errors is measured by one of the bounds proved here, applied once.

Throughout, the field is $\mathbb{C}$, $\mathbf{A} \in M_m(\mathbb{C})$ and $\mathbf{B} \in M_n(\mathbb{C})$, and $\mathbf{X}, \mathbf{C} \in M_{m\times n}(\mathbb{C})$. We write $\mathcal{S} = \mathcal{S}_{\mathbf{A},\mathbf{B}}$ for the Sylvester operator $\mathbf{X} \mapsto \mathbf{AX} - \mathbf{XB}$ of Definition 12.10.1, and $\mathbf{K} = \mathbf{I}_n \otimes \mathbf{A} - \mathbf{B}^{\top} \otimes \mathbf{I}_m$ for its matrix, so that $\operatorname{vec}(\mathcal{S}(\mathbf{X})) = \mathbf{K}\operatorname{vec}\mathbf{X}$ by Eq. 12.10.1. The eigenvalues of $\mathbf{A}$ are $\lambda_1, \dots, \lambda_m$ and those of $\mathbf{B}$ are $\mu_1, \dots, \mu_n$, with multiplicity. Every bound names its norm: $\|\cdot\|_F$ is the

Frobenius norm and $\|\cdot\|_2$ the spectral norm.

20.7.1 How large can the solution be?

Start with the smallest case, $m = n = 1$. The equation is $(a - b)x = c$, and when $a \neq b$ its solution is $x = c/(a - b)$. So the size of the answer is governed by one number, the distance $|a - b|$ between the two "spectra": the closer they are, the larger x can be for a given c.

The natural guess for larger matrices is that the distance between the two spectra, $\min_{i,j}|\lambda_i - \mu_j|$, plays the same role, so that $\|\mathbf{X}\|_F \leq \|\mathbf{C}\|_F / \min_{i,j}|\lambda_i - \mu_j|$. We will see that this guess is right for normal pairs and can be wrong by any factor otherwise. The number that really controls the solution is the smallest amount by which $\mathcal{S}$ can stretch a matrix, and it deserves a name.

The separation of $\mathbf{A}$ *and* $\mathbf{B}$ *is the least that the Sylvester operator can shrink a matrix of unit size.*

Definition 20.7.1. Separation

Let $\mathbf{A} \in M_m(\mathbb{C})$ and $\mathbf{B} \in M_n(\mathbb{C})$. The **separation** of $\mathbf{A}$ and $\mathbf{B}$ is

$$\operatorname{sep}_F(\mathbf{A}, \mathbf{B}) \coloneqq \min\{\|\mathbf{AX} - \mathbf{XB}\|_F : \mathbf{X} \in M_{m\times n}(\mathbb{C}),\ \|\mathbf{X}\|_F = 1\}.$$

In words: we run over **every** matrix $\mathbf{X}$ of the right shape whose Frobenius norm is exactly 1, apply the Sylvester operator, measure the result in the Frobenius norm, and take the **smallest** value. By homogeneity, $\|\mathbf{AX} - \mathbf{XB}\|_F \geq \operatorname{sep}_F(\mathbf{A}, \mathbf{B})\|\mathbf{X}\|_F$ for **every** $\mathbf{X}$, and $\operatorname{sep}_F(\mathbf{A}, \mathbf{B})$ is the largest constant with this property. The order matters: $\operatorname{sep}_F(\mathbf{A}, \mathbf{B})$ belongs to $\mathbf{X} \mapsto \mathbf{AX} - \mathbf{XB}$, and $\operatorname{sep}_F(\mathbf{B}, \mathbf{A})$ to a different operator on a different space. The subscript F records the norm, as in $\|\cdot\|_F$.

Well-definedness. A minimum over an infinite set needs to exist, and the first part of Proposition 20.7.1 below shows that it does, without any appeal to compactness: it is the smallest singular value of $\mathbf{K}$.

Examples.

1. $m = n = 1$**.** Then $\mathbf{X} = (x)$ with $|x| = 1$, and $|ax - xb| = |a - b|$. So $\operatorname{sep}_F(a, b) = |a - b|$, the one-dimensional picture above.
2. **Equal matrices.** For square $\mathbf{A}$, $\operatorname{sep}_F(\mathbf{A}, \mathbf{A}) = 0$: the matrix $\mathbf{X} = \mathbf{I}_m/\sqrt{m}$ has $\|\mathbf{X}\|_F = 1$ and $\mathbf{AX} - \mathbf{XA} = \mathbf{O}$. This degenerate case matters because it is the extreme of a shared eigenvalue, and the first property below says that any shared eigenvalue forces the separation to 0.
3. **Diagonal matrices.** If $\mathbf{A} = \operatorname{diag}(\lambda_1, \ldots, \lambda_m)$ and $\mathbf{B} = \operatorname{diag}(\mu_1, \ldots, \mu_n)$, then $(\mathbf{AX} - \mathbf{XB})_{ij} = (\lambda_i - \mu_j)x_{ij}$, so $\mathcal{S}$ multiplies each entry by its own factor. The smallest stretch is the smallest factor in modulus, and $\operatorname{sep}_F(\mathbf{A}, \mathbf{B}) = \min_{i,j}|\lambda_i - \mu_j|$. The proof of Theorem 20.7.2 below writes this out.

Nonexample by minimal change. Drop the normalization and take $\min\{\|\mathbf{AX} - \mathbf{XB}\|_F : \mathbf{X} \neq \mathbf{O}\}$ instead. The set of values is still nonempty and bounded below, but its infimum is 0 for every pair, and when the spectra are disjoint it is never attained: replacing $\mathbf{X}$ by $t\mathbf{X}$ multiplies the value by t, and t can be as small as we like, while the value 0 is taken at some $\mathbf{X} \neq \mathbf{O}$ exactly when the two spectra meet (Theorem 12.10.2), as they do in the second example above, where $\mathbf{A} = \mathbf{B}$ and $\mathbf{X} = \mathbf{I}_m$. It is exactly the clause $\|\mathbf{X}\|_F = 1$ that fails, and it is what turns a

meaningless infimum into a measurement of the operator.

Why the Frobenius norm. Vectorization lists the entries of $\mathbf{X}$ in one column, so $\|\operatorname{vec}\mathbf{X}\|_2 = \|\mathbf{X}\|_F$: vec is an isometry from $(M_{m\times n}(\mathbb{C}), \|\cdot\|_F)$ to $(\mathbb{C}^{mn}, \|\cdot\|_2)$. That turns sep_F into a question about the ordinary matrix $\mathbf{K}$, which the singular value decomposition answers. With $\|\cdot\|_2$ on both sides the minimum is also a sensible quantity, but it has no such formula, and we will not need it: the spectral-norm results below go through a different argument.

Warning. **The separation is not the distance between the spectra.** It is at most that distance, by part (c) of the proposition below, but it can be far smaller. For $\mathbf{A} = \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix}$ and the 1×1 matrix $\mathbf{B} = \mathbf{O}_1$, the spectra $\{1\}$ and $\{0\}$ are at distance 1 for every t, yet $\operatorname{sep}_F(\mathbf{A}, \mathbf{B}) < 1/t$ for $t > 0$ (Example (Separation Far Below the Gap)). Whenever $\mathbf{A}$ or $\mathbf{B}$ is not normal, compute the separation; do not read it off the eigenvalues.

Everything we need about the separation is collected in one proposition.

Proposition 20.7.1. Properties of the Separation

Let $\mathbf{A} \in M_m(\mathbb{C})$, $\mathbf{B} \in M_n(\mathbb{C})$, and $\mathbf{K} = \mathbf{I}_n \otimes \mathbf{A} - \mathbf{B}^{\top} \otimes \mathbf{I}_m$.

(a) The minimum defining $\operatorname{sep}_F(\mathbf{A}, \mathbf{B})$ exists, and $\operatorname{sep}_F(\mathbf{A}, \mathbf{B}) = \sigma_{mn}(\mathbf{K})$, the smallest singular value of $\mathbf{K}$.

(b) $\operatorname{sep}_F(\mathbf{A}, \mathbf{B}) > 0$ if and only if $\operatorname{spec}(\mathbf{A}) \cap \operatorname{spec}(\mathbf{B}) = \varnothing$.

(c) $\operatorname{sep}_F(\mathbf{A}, \mathbf{B}) \le \min_{i,j} |\lambda_i - \mu_j|$.

(d) If $\operatorname{spec}(\mathbf{A}) \cap \operatorname{spec}(\mathbf{B}) = \varnothing$, then for every $\mathbf{C}$ the unique solution of $\mathbf{AX} - \mathbf{XB} = \mathbf{C}$ satisfies

$$\|\mathbf{X}\|_F \le \frac{\|\mathbf{C}\|_F}{\operatorname{sep}_F(\mathbf{A}, \mathbf{B})},$$

and equality holds for some $\mathbf{C} \neq \mathbf{O}$.

(e) For unitary $\mathbf{U} \in \mathrm{U}(m)$ and $\mathbf{V} \in \mathrm{U}(n)$, $\operatorname{sep}_F(\mathbf{U}^*\mathbf{AU}, \mathbf{V}^*\mathbf{BV}) = \operatorname{sep}_F(\mathbf{A}, \mathbf{B})$.

(f) If $n = 1$ and $\mathbf{B} = (\mu)$, then $\operatorname{sep}_F(\mathbf{A}, \mathbf{B}) = \sigma_m(\mathbf{A} - \mu\mathbf{I}_m)$, the smallest singular value of $\mathbf{A} - \mu\mathbf{I}_m$.

Idea. Part (a) is the isometry vec plus the fact that the smallest stretch of a square matrix is its smallest singular value. Part (b) is Sylvester's theorem, read through (a). Part (c) needs one good test matrix: an eigenvector of $\mathbf{A}$ times a left eigenvector of $\mathbf{B}$ is multiplied by $\lambda - \mu$ under $\mathcal{S}$, exactly as in the 1×1 case. Part (d) is the definition read backwards. Part (e) holds because unitary factors change neither the set of unit matrices nor the Frobenius norm. Part (f) is the case $n = 1$ of (a), where $\mathbf{K}$ is nothing but $\mathbf{A} - \mu\mathbf{I}_m$.

Proof.

(a) Let $\mathbf{X} \in M_{m\times n}(\mathbb{C})$ and $\mathbf{x} = \operatorname{vec}\mathbf{X}$. The entries of $\mathbf{x}$ are those of $\mathbf{X}$, so $\|\mathbf{x}\|_2 = \|\mathbf{X}\|_F$, and likewise $\|\mathbf{AX} - \mathbf{XB}\|_F = \|\mathbf{Kx}\|_2$ by Eq. 12.10.1. Since vec is a bijection from $M_{m\times n}(\mathbb{C})$ onto $\mathbb{C}^{mn}$ (Definition 8.4.2), the set whose minimum defines $\operatorname{sep}_F(\mathbf{A}, \mathbf{B})$ equals $\{\|\mathbf{Kx}\|_2 : \|\mathbf{x}\|_2 = 1\}$. By Lemma 17.9.1 (b) with k equal to the size mn of $\mathbf{K}$,

the head subspace is all of $\mathbb{C}^{mn}$, and the minimum of $\|\mathbf{Kx}\|/\|\mathbf{x}\|$ over $\mathbf{x} \neq \mathbf{0}$ exists and equals $\sigma_{mn}(\mathbf{K})$. Restricting to unit vectors changes nothing, since the quotient is unchanged when $\mathbf{x}$ is scaled.

(b) By Theorem 13.9.1, the number of nonzero singular values of $\mathbf{K}$ is $\operatorname{rank}\mathbf{K}$. Since $\sigma_{mn}(\mathbf{K})$ is the smallest, it is nonzero exactly when all mn of them are, that is, when $\operatorname{rank}\mathbf{K} = mn$, which says that $\mathbf{K}$ is invertible (Theorem 7.5.3 (d)). By Eq. 12.10.1, $\mathbf{K}$ is invertible exactly when $\mathcal{S}$ is, and by Theorem 12.10.2 ((a) $\Leftrightarrow$ (b)) that happens exactly when $\operatorname{spec}(\mathbf{A}) \cap \operatorname{spec}(\mathbf{B}) = \varnothing$. Combining with (a) proves (b).

(c) Fix i, j. Choose an eigenvector $\mathbf{x}$ of $\mathbf{A}$ for λ_i, and, by Proposition 9.10.5 (b), a left eigenvector $\mathbf{y}$ of $\mathbf{B}$ for μ_j, so that $\mathbf{y}^{\top}\mathbf{B} = \mu_j\mathbf{y}^{\top}$ (Definition 9.10.1). Scale both to length 1, and put $\mathbf{X} = \mathbf{x}\mathbf{y}^{\top}$. Its (p, q) entry is $x_p y_q$, so

$$\|\mathbf{X}\|_F^2 = \sum_{p,q} |x_p|^2 |y_q|^2 = \|\mathbf{x}\|^2 \|\mathbf{y}\|^2 = 1.$$

By associativity, $\mathbf{AX} - \mathbf{XB} = (\mathbf{Ax})\mathbf{y}^{\top} - \mathbf{x}(\mathbf{y}^{\top}\mathbf{B}) = (\lambda_i - \mu_j)\mathbf{X}$. Hence $\operatorname{sep}_F(\mathbf{A}, \mathbf{B}) \leq \|(\lambda_i - \mu_j)\mathbf{X}\|_F = |\lambda_i - \mu_j|$. This holds for every pair (i, j), so it holds for the minimum.

(d) By Theorem 12.10.2 the solution $\mathbf{X}$ exists and is unique. If $\mathbf{X} = \mathbf{O}$ there is nothing to prove. Otherwise $\mathbf{X}/\|\mathbf{X}\|_F$ has Frobenius norm 1, so by Definition 20.7.1

$$\operatorname{sep}_F(\mathbf{A}, \mathbf{B}) \leq \frac{\|\mathbf{AX} - \mathbf{XB}\|_F}{\|\mathbf{X}\|_F} = \frac{\|\mathbf{C}\|_F}{\|\mathbf{X}\|_F},$$

and $\operatorname{sep}_F(\mathbf{A}, \mathbf{B}) > 0$ by (b), so we may rearrange. For equality, take $\mathbf{X}_0$ with $\|\mathbf{X}_0\|_F = 1$ attaining the minimum, which exists by (a), and put $\mathbf{C} = \mathbf{AX}_0 - \mathbf{X}_0\mathbf{B}$. Then $\mathbf{C} \neq \mathbf{O}$, because $\|\mathbf{C}\|_F = \operatorname{sep}_F(\mathbf{A}, \mathbf{B}) > 0$, and the unique solution is $\mathbf{X}_0$, for which the two sides are both 1.

(e) Put $\mathbf{A}' = \mathbf{U}^*\mathbf{AU}$ and $\mathbf{B}' = \mathbf{V}^*\mathbf{BV}$. For every $\mathbf{X}$,

$$\mathbf{A}'(\mathbf{U}^*\mathbf{XV}) - (\mathbf{U}^*\mathbf{XV})\mathbf{B}' = \mathbf{U}^*(\mathbf{AX} - \mathbf{XB})\mathbf{V},$$

because $\mathbf{UU}^* = \mathbf{I}_m$ and $\mathbf{VV}^* = \mathbf{I}_n$. By Lemma 11.7.5, $\|\mathbf{U}^*\mathbf{XV}\|_F = \|\mathbf{X}\|_F$ and the right side has Frobenius norm $\|\mathbf{AX} - \mathbf{XB}\|_F$. The map $\mathbf{X} \mapsto \mathbf{U}^*\mathbf{XV}$ is a bijection of $M_{m\times n}(\mathbb{C})$, with inverse $\mathbf{Y} \mapsto \mathbf{UYV}^*$, so it carries the matrices of Frobenius norm 1 onto themselves. Hence the two minima are taken over the same set of values, and they are equal.

(f) With $n = 1$ and $\mathbf{B} = (\mu)$, the Kronecker matrix is $\mathbf{K} = \mathbf{I}_1 \otimes \mathbf{A} - (\mu) \otimes \mathbf{I}_m = \mathbf{A} - \mu\mathbf{I}_m$, of size $mn = m$. So (a) gives $\operatorname{sep}_F(\mathbf{A}, \mathbf{B}) = \sigma_m(\mathbf{A} - \mu\mathbf{I}_m)$.

■

Part (d) is the quantitative form that Chapter 8 and Chapter 12 promised, and it says more than an inequality. The number $1/\operatorname{sep}_F(\mathbf{A}, \mathbf{B})$ is **exactly** the largest factor by which solving the equation can magnify the data, measured in the Frobenius norm. Since $\mathcal{S}$ is linear, it is also the factor for errors: if $\mathbf{C}$ is replaced by $\mathbf{C} + \Delta\mathbf{C}$, the solution moves by $\mathcal{S}^{-1}(\Delta\mathbf{C})$, whose norm is at most $\|\Delta\mathbf{C}\|_F/\operatorname{sep}_F(\mathbf{A}, \mathbf{B})$, and some $\Delta\mathbf{C}$ achieves this. A small separation is the precise meaning of "badly conditioned". Part (c) says that close spectra always make the separation small. The rest of the section asks whether the converse holds, that is, whether well-separated spectra make it large.

Quick check. Part (f) of Proposition 20.7.1 was read off the Kronecker matrix. Prove it instead straight from Definition 20.7.1, without vectorizing: for $\mathbf{B} = (\mu)$, unwind what the matrices $\mathbf{X}$ of Frobenius norm 1 are.

Solution. Here $n = 1$, so $\mathbf{X} = \mathbf{x}$ is a column in $\mathbb{C}^m$ with $\|\mathbf{x}\|_F = \|\mathbf{x}\|$, and $\mathbf{A}\mathbf{x} - \mathbf{x}\mu = (\mathbf{A} - \mu\mathbf{I}_m)\mathbf{x}$. So $\operatorname{sep}_F(\mathbf{A}, \mathbf{B}) = \min_{\|\mathbf{x}\|=1} \|(\mathbf{A} - \mu\mathbf{I}_m)\mathbf{x}\|$, which is $\sigma_m(\mathbf{A} - \mu\mathbf{I}_m)$ by Lemma 17.9.1 (b) with $k = m$.

20.7.2 Normal pairs: the separation is the gap

For normal matrices the naive guess is right, and the proof is the diagonal example written out.

Theorem 20.7.2. Separation of Normal Matrices

Let $\mathbf{A} \in M_m(\mathbb{C})$ and $\mathbf{B} \in M_n(\mathbb{C})$ be **normal**, with eigenvalues $\lambda_1, \dots, \lambda_m$ and $\mu_1, \dots, \mu_n$. Then

$$\operatorname{sep}_F(\mathbf{A}, \mathbf{B}) = \min_{i,j} |\lambda_i - \mu_j|.$$

Consequently, if $\delta = \min_{i,j} |\lambda_i - \mu_j| > 0$, the unique solution of $\mathbf{A}\mathbf{X} - \mathbf{X}\mathbf{B} = \mathbf{C}$ satisfies $\|\mathbf{X}\|_F \le \|\mathbf{C}\|_F / \delta$.

Idea. Unitary diagonalization costs nothing in the Frobenius norm, by Proposition 20.7.1 (e). Once both matrices are diagonal, the operator scales the (i, j) entry by $\lambda_i - \mu_j$, so the total shrinkage is a weighted average of the $|\lambda_i - \mu_j|^2$, and a weighted average is at least the smallest weight.

Proof. By Corollary 12.4.3 there are unitary $\mathbf{U} \in \mathrm{U}(m)$ and $\mathbf{V} \in \mathrm{U}(n)$ with $\mathbf{U}^*\mathbf{A}\mathbf{U} = \mathbf{\Lambda} := \operatorname{diag}(\lambda_1, \dots, \lambda_m)$ and $\mathbf{V}^*\mathbf{B}\mathbf{V} = \mathbf{M} := \operatorname{diag}(\mu_1, \dots, \mu_n)$. By Proposition 20.7.1 (e), $\operatorname{sep}_F(\mathbf{A}, \mathbf{B}) = \operatorname{sep}_F(\mathbf{\Lambda}, \mathbf{M})$.

Let $\mathbf{X} = (x_{ij})$ have $\|\mathbf{X}\|_F = 1$, and put $\delta = \min_{i,j} |\lambda_i - \mu_j|$. The (i, j) entry of $\mathbf{\Lambda}\mathbf{X} - \mathbf{X}\mathbf{M}$ is $\lambda_i x_{ij} - x_{ij}\mu_j = (\lambda_i - \mu_j) x_{ij}$, since multiplying by a diagonal matrix on the left scales rows and on the right scales columns. Hence

$$\begin{aligned} \|\mathbf{\Lambda}\mathbf{X} - \mathbf{X}\mathbf{M}\|_F^2 &= \sum_{i,j} |\lambda_i - \mu_j|^2 |x_{ij}|^2 \\ &\ge \delta^2 \sum_{i,j} |x_{ij}|^2 = \delta^2. \end{aligned}$$

Therefore $\operatorname{sep}_F(\mathbf{\Lambda}, \mathbf{M}) \ge \delta$. The reverse inequality is Proposition 20.7.1 (c). This proves the equality, and the bound on $\mathbf{X}$ is then Proposition 20.7.1 (d).

■

The theorem answers the question of the opening completely in the normal case: there, how badly conditioned the equation is depends on the spectra and on nothing else. In particular it

applies to Hermitian pairs, which is the case Section 9 needs.

Example. A Normal Pair

Let

$$\mathbf{A} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}, \qquad \mathbf{B} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}.$$

Compute $\operatorname{sep}_F(\mathbf{A}, \mathbf{B})$ in two ways: from Theorem 20.7.2, and from the singular values of $\mathbf{K}$.

Solution. The matrix $\mathbf{A}$ is real symmetric, hence normal, with $p_{\mathbf{A}}(x) = (x-2)^2 - 1 = (x-1)(x-3)$, so its eigenvalues are $1, 3$. The matrix $\mathbf{B}$ is real with $\mathbf{B}^\top\mathbf{B} = \mathbf{I}_2 = \mathbf{B}\mathbf{B}^\top$, hence normal, and $p_{\mathbf{B}}(x) = x^2 + 1$, so its eigenvalues are $\pm i$. The four distances are

$$|1 \mp i| = \sqrt{2}, \qquad |3 \mp i| = \sqrt{10},$$

so Theorem 20.7.2 gives $\operatorname{sep}_F(\mathbf{A}, \mathbf{B}) = \sqrt{2}$.

For the second computation, $\mathbf{B}^\top = \left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)$, and by Definition 8.4.1, in 2×2 blocks,

$$\mathbf{K} = \mathbf{I}_2 \otimes \mathbf{A} - \mathbf{B}^\top \otimes \mathbf{I}_2 = \begin{pmatrix} \mathbf{A} & -\mathbf{I}_2 \\ \mathbf{I}_2 & \mathbf{A} \end{pmatrix}.$$

Since $\mathbf{K}$ is real, $\mathbf{K}^* = \mathbf{K}^\top$, so the singular values of $\mathbf{K}$ are the nonnegative square roots of the eigenvalues of $\mathbf{K}^\top\mathbf{K}$ (Definition 13.9.1). Since $\mathbf{A}$ is real symmetric, block multiplication (Theorem 8.1.1) gives

$$\begin{aligned} \mathbf{K}^\top\mathbf{K} &= \begin{pmatrix} \mathbf{A} & \mathbf{I}_2 \\ -\mathbf{I}_2 & \mathbf{A} \end{pmatrix} \begin{pmatrix} \mathbf{A} & -\mathbf{I}_2 \\ \mathbf{I}_2 & \mathbf{A} \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{A}^2 + \mathbf{I}_2 & \mathbf{O} \\ \mathbf{O} & \mathbf{A}^2 + \mathbf{I}_2 \end{pmatrix}. \end{aligned}$$

The eigenvalues of $\mathbf{A}^2 + \mathbf{I}_2$ are $1^2 + 1 = 2$ and $3^2 + 1 = 10$ (Theorem 9.10.1), so the singular values of $\mathbf{K}$ are $\sqrt{10}, \sqrt{10}, \sqrt{2}, \sqrt{2}$. The smallest is $\sqrt{2}$, in agreement with Proposition 20.7.1 (a).

20.7.3 Nonnormal pairs: the separation can collapse

For a nonnormal pair the inequality of Proposition 20.7.1 (c) can be strict, and by as large a factor as we please. The smallest example uses a 1×1 matrix $\mathbf{B}$, so that Proposition 20.7.1 (f) computes the separation.

Example. Separation Far Below the Gap

For a real number $t \geq 0$, let

$$\mathbf{A} = \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix}, \qquad \mathbf{B} = \mathbf{O}_1.$$

Compute $\operatorname{sep}_F(\mathbf{A},\mathbf{B})$ and compare it with the distance between the spectra. Then, for $t = 99/10$, solve $\mathbf{AX} - \mathbf{XB} = \mathbf{C}$ with $\mathbf{C} = \mathbf{e}_2$ and compare the size of the solution with the guess $\|\mathbf{C}\|_F / \min|\lambda_i - \mu_j|$.

Solution. The spectra are $\operatorname{spec}(\mathbf{A}) = \{1\}$ (triangular, Theorem 9.7.4) and $\operatorname{spec}(\mathbf{B}) = \{0\}$, at distance 1 for every t. By Proposition 20.7.1 (f) with $\mu = 0$, $\operatorname{sep}_F(\mathbf{A},\mathbf{B}) = \sigma_2(\mathbf{A})$, the square root of the smaller eigenvalue of

$$\mathbf{A}^{\top}\mathbf{A} = \begin{pmatrix} 1 & t \\ t & 1 + t^2 \end{pmatrix}.$$

Its characteristic polynomial is $(z-1)(z-1-t^2) - t^2 = z^2 - (2+t^2)z + 1$. Both roots are positive and their product is 1, so the smaller is the reciprocal of the larger, $\sigma_2(\mathbf{A}) = 1/\sigma_1(\mathbf{A})$. Solving the quadratic,

$$\sigma_2(\mathbf{A})^2 = \frac{2 + t^2 - t\sqrt{t^2+4}}{2} = \Big(\frac{\sqrt{t^2+4} - t}{2}\Big)^2,$$

as expanding the square confirms. Hence

$$\operatorname{sep}_F(\mathbf{A},\mathbf{B}) = \frac{\sqrt{t^2+4} - t}{2} = \frac{2}{\sqrt{t^2+4} + t},$$

where the second form multiplies top and bottom by $\sqrt{t^2+4} + t$. The second form is less than $2/(2t) = 1/t$ for $t > 0$, because $\sqrt{t^2+4} > t$. The spectra stay at distance 1, and the separation tends to 0 as $t \to \infty$.

For $t = 99/10$ we have $t^2 + 4 = 10201/100$, whose square root is $101/10$, so

$$\operatorname{sep}_F(\mathbf{A},\mathbf{B}) = \frac{1}{2}\Big(\frac{101}{10} - \frac{99}{10}\Big) = \frac{1}{10}.$$

The equation is $\mathbf{Ax} = \mathbf{e}_2$ for the column $\mathbf{x} = \mathbf{X}$, so $\mathbf{x} = \mathbf{A}^{-1}\mathbf{e}_2 = (-t, 1)$, since $\mathbf{A}^{-1} = \left(\begin{smallmatrix} 1 & -t \\ 0 & 1 \end{smallmatrix}\right)$. Its length is $\sqrt{t^2+1} = \sqrt{9901}/10 \approx 9.95$. The guess from the eigenvalues was $\|\mathbf{C}\|_F/1 = 1$, wrong by a factor of almost 10. The true bound $\|\mathbf{C}\|_F / \operatorname{sep}_F(\mathbf{A},\mathbf{B}) = 10$ from Proposition 20.7.1 (d) is almost attained, and it is attained when $\mathbf{C}$ is a unit left singular vector of $\mathbf{A}$ for σ_2.

The example has a second reading. The eigenvalue 1 of $\mathbf{A}$ is far from 0, yet a perturbation of $\mathbf{A}$ of size $\operatorname{sep}_F(\mathbf{A},\mathbf{B}) = 1/10$ in the spectral norm makes 0 an eigenvalue, because that is how far $\mathbf{A}$ is from the singular matrices (Corollary 17.9.7). Exercise C1 below shows that, for every pair, no perturbation of spectral norm less than $\operatorname{sep}_F(\mathbf{A},\mathbf{B})$ can create a common eigenvalue, and that when $\mathbf{B}$ is 1×1 one of exactly that size can. For a nonnormal pair, the distance between the spectra measures the wrong thing, and the separation measures the right one.

20.7.4 A one-sided bound in the spectral norm

Section 9 will measure the angle between two invariant subspaces, and the largest such angle is a spectral norm, not a Frobenius norm. For Hermitian pairs, Theorem 20.7.2 already gives $\|\mathbf{X}\|_F \le \|\mathbf{C}\|_F/\delta$. The spectral-norm analogue with the same constant is **false** in general: Exercise C2 below exhibits real diagonal $\mathbf{A}$ and $\mathbf{B}$ whose spectra interlace at distance 1, yet

$\|\mathbf{X}\|_2 > \|\mathbf{C}\|_2$. What rescues the spectral norm is a stronger hypothesis on where the spectra lie: all of $\operatorname{spec}(\mathbf{A})$ on one side of a gap and all of $\operatorname{spec}(\mathbf{B})$ on the other.

Theorem 20.7.3. One-Sided Sylvester Bound

Let $\mathbf{A} \in M_m(\mathbb{C})$ and $\mathbf{B} \in M_n(\mathbb{C})$ be Hermitian, and let $\alpha > \beta$ be real numbers with

$$\mathbf{A} \succeq \alpha \mathbf{I}_m, \qquad \mathbf{B} \preceq \beta \mathbf{I}_n.$$

Then for every $\mathbf{C} \in M_{m\times n}(\mathbb{C})$ the equation $\mathbf{AX} - \mathbf{XB} = \mathbf{C}$ has exactly one solution, and it satisfies

$$\|\mathbf{X}\|_2 \le \frac{\|\mathbf{C}\|_2}{\alpha - \beta}, \qquad \|\mathbf{X}\|_F \le \frac{\|\mathbf{C}\|_F}{\alpha - \beta}.$$

Idea. The **singular-pair trick.** A single pair of vectors sees the whole spectral norm of $\mathbf{X}$: by the singular value decomposition there are unit vectors $\mathbf{u}, \mathbf{v}$ with $\mathbf{Xv} = s\mathbf{u}$ and $\mathbf{X}^*\mathbf{u} = s\mathbf{v}$, where $s = \|\mathbf{X}\|_2$. Sandwich the equation between them. The left side becomes $\mathbf{u}^*\mathbf{AXv} - \mathbf{u}^*\mathbf{XBv} = s(\mathbf{u}^*\mathbf{Au} - \mathbf{v}^*\mathbf{Bv})$, a real number at least $s(\alpha - \beta)$ because of the two one-sided hypotheses. The right side becomes $\mathbf{u}^*\mathbf{Cv}$, which is at most $\|\mathbf{C}\|_2$ in modulus. The one-sided hypotheses are used exactly once, to make both quadratic forms pull in the same direction.

Proof. **Existence and uniqueness.** Let $\lambda \in \operatorname{spec}(\mathbf{A})$ with unit eigenvector $\mathbf{x}$. Then $\lambda = \langle \mathbf{Ax}, \mathbf{x}\rangle \ge \alpha$, because $\langle(\mathbf{A}-\alpha\mathbf{I}_m)\mathbf{x}, \mathbf{x}\rangle \ge 0$ by Definition 13.1.1 and Definition 13.5.1. Likewise every $\mu \in \operatorname{spec}(\mathbf{B})$ satisfies $\mu \le \beta < \alpha$. So $\operatorname{spec}(\mathbf{A}) \cap \operatorname{spec}(\mathbf{B}) = \varnothing$, and Theorem 12.10.2 gives exactly one solution $\mathbf{X}$.

The spectral norm. Put $s = \|\mathbf{X}\|_2$, which equals $\sigma_1(\mathbf{X})$ by Theorem 16.3.3 (c). If $s = 0$ there is nothing to prove. Otherwise $\mathbf{X}$ has rank at least 1, and Theorem 13.9.1 with Eq. 13.9.3 (for $i = 1$) gives unit vectors $\mathbf{u} \in \mathbb{C}^m$ and $\mathbf{v} \in \mathbb{C}^n$ with

$$\mathbf{Xv} = s\mathbf{u}, \qquad \mathbf{X}^*\mathbf{u} = s\mathbf{v}. \tag{20.7.1}$$

Taking adjoints in the second, $\mathbf{u}^*\mathbf{X} = s\mathbf{v}^*$. Multiply $\mathbf{AX} - \mathbf{XB} = \mathbf{C}$ by $\mathbf{u}^*$ on the left and $\mathbf{v}$ on the right:

$$\begin{aligned}\mathbf{u}^*\mathbf{Cv} &= \mathbf{u}^*\mathbf{A}(\mathbf{Xv}) - (\mathbf{u}^*\mathbf{X})\mathbf{Bv} \\ &= s\,\mathbf{u}^*\mathbf{Au} - s\,\mathbf{v}^*\mathbf{Bv}.\end{aligned}$$

Since $\mathbf{A} - \alpha\mathbf{I}_m \succeq \mathbf{O}$ and $\mathbf{u}$ is a unit vector, $\mathbf{u}^*\mathbf{Au} - \alpha = \langle(\mathbf{A} - \alpha\mathbf{I}_m)\mathbf{u}, \mathbf{u}\rangle \ge 0$. Since $\beta\mathbf{I}_n - \mathbf{B} \succeq \mathbf{O}$, likewise $\mathbf{v}^*\mathbf{Bv} \le \beta$. Hence $\mathbf{u}^*\mathbf{Cv}$ is real and

$$\mathbf{u}^*\mathbf{Cv} \ge s(\alpha - \beta).$$

On the other hand, by the Cauchy–Schwarz inequality (Theorem 11.1.5) and Theorem 16.3.2 (a),

$$|\mathbf{u}^*\mathbf{Cv}| \le \|\mathbf{u}\|\,\|\mathbf{Cv}\| \le \|\mathbf{C}\|_2.$$

Combining, $s(\alpha - \beta) \le \|\mathbf{C}\|_2$, and dividing by $\alpha - \beta > 0$ gives the first bound.

The Frobenius norm. Hermitian matrices are normal. Every $\lambda_i \geq \alpha$ and every $\mu_j \leq \beta$, as shown above, so $\lambda_i - \mu_j \geq \alpha - \beta$ for all i, j. By Theorem 20.7.2, $\operatorname{sep}_F(\mathbf{A}, \mathbf{B}) = \min_{i,j}(\lambda_i - \mu_j) \geq \alpha - \beta$, and Proposition 20.7.1 (d) gives $\|\mathbf{X}\|_F \leq \|\mathbf{C}\|_F / \operatorname{sep}_F(\mathbf{A}, \mathbf{B}) \leq \|\mathbf{C}\|_F / (\alpha - \beta)$. This proves the theorem.

■

The trick deserves its name because it is portable. Whenever a matrix $\mathbf{X}$ satisfies a linear equation and one wants $\|\mathbf{X}\|_2$, take a top singular pair of $\mathbf{X}$ itself and test the equation against it. The spectral norm is then attained on the test, and the rest is a pair of scalar inequalities. Notice what was **not** used: the size of the gap between individual eigenvalues. Only the two half-lines $[\alpha, \infty)$ and $(-\infty, \beta]$ matter, which is why the bound survives when $\mathbf{A}$ and $\mathbf{B}$ have many eigenvalues each.

Remark. The one-sided hypothesis does real work in the spectral-norm bound, as Exercise C2 shows. When the spectra of two Hermitian matrices are merely at distance δ, in any interlaced pattern, the spectral-norm bound does hold with a larger constant: $\|\mathbf{X}\|_2 \leq (\pi/2)\|\mathbf{C}\|_2/\delta$, and $\pi/2$ cannot be improved. We do not prove this, and nothing in the book depends on it.

The form of the theorem used most often has a plus sign and two positive definite matrices. It is a direct translation.

Corollary 20.7.4. Sylvester Bound for Positive Matrices

Let $\mathbf{S} \in M_m(\mathbb{C})$ and $\mathbf{T} \in M_n(\mathbb{C})$ be Hermitian with $\mathbf{S} \succeq \alpha \mathbf{I}_m$ and $\mathbf{T} \succeq \beta \mathbf{I}_n$, where $\alpha + \beta > 0$. Then for every $\mathbf{C} \in M_{m \times n}(\mathbb{C})$ the equation

$$\mathbf{S}\mathbf{X} + \mathbf{X}\mathbf{T} = \mathbf{C}$$

has exactly one solution, and it satisfies

$$\|\mathbf{X}\|_2 \leq \frac{\|\mathbf{C}\|_2}{\alpha + \beta}, \qquad \|\mathbf{X}\|_F \leq \frac{\|\mathbf{C}\|_F}{\alpha + \beta}.$$

Proof. The equation reads $\mathbf{S}\mathbf{X} - \mathbf{X}(-\mathbf{T}) = \mathbf{C}$. The matrix $-\mathbf{T}$ is Hermitian, and $-\beta\mathbf{I}_n - (-\mathbf{T}) = \mathbf{T} - \beta\mathbf{I}_n \succeq \mathbf{O}$, so $-\mathbf{T} \preceq -\beta\mathbf{I}_n$. Since $\alpha > -\beta$, Theorem 20.7.3 applies to the pair $(\mathbf{S}, -\mathbf{T})$ with the numbers α and $-\beta$, and $\alpha - (-\beta) = \alpha + \beta$.

■

The hypothesis is weaker than it looks: α or β may be negative, provided their sum is positive. The typical case is $\mathbf{S}, \mathbf{T} \succ \mathbf{O}$ with $\alpha = \lambda_m(\mathbf{S})$ and $\beta = \lambda_n(\mathbf{T})$, the smallest eigenvalues, which are the best constants because $\mathbf{S} \succeq \lambda_m(\mathbf{S})\mathbf{I}_m$ (Lemma 13.5.2). Section 10 applies the corollary twice in this form, once with $\mathbf{S}$ and $\mathbf{T}$ the square roots of two positive definite matrices and once with them the positive factors of two polar decompositions.

Example. The Bound Can Be Attained

Let $\mathbf{S} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$, $\mathbf{T} = (1)$, and $\mathbf{C} = \begin{pmatrix} 4 \\ -4 \end{pmatrix}$. Solve $\mathbf{SX} + \mathbf{XT} = \mathbf{C}$ and compare with Corollary 20.7.4.

Solution. The eigenvalues of $\mathbf{S}$ are 1 and 3 (as for $\mathbf{A}$ in Example (A Normal Pair)), so $\mathbf{S} \succeq 1 \cdot \mathbf{I}_2$ by Lemma 13.5.2, and $\mathbf{T} \succeq 1$. The corollary applies with $\alpha = \beta = 1$ and predicts $\|\mathbf{X}\|_2 \le \|\mathbf{C}\|_2/2$.

Here $\mathbf{X} = \mathbf{x}$ is a column and the equation is $(\mathbf{S} + \mathbf{I}_2)\mathbf{x} = \mathbf{C}$, with

$$\mathbf{S} + \mathbf{I}_2 = \begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix}, \qquad (\mathbf{S} + \mathbf{I}_2)^{-1} = \frac{1}{8}\begin{pmatrix} 3 & -1 \\ -1 & 3 \end{pmatrix}.$$

So $\mathbf{x} = \frac{1}{8}(12 + 4, -4 - 12) = (2, -2)$. For a column, the spectral and Frobenius norms are both the Euclidean length, so $\|\mathbf{X}\|_2 = 2\sqrt{2}$, while $\|\mathbf{C}\|_2/2 = 4\sqrt{2}/2 = 2\sqrt{2}$. The bound is an equality.

The reason is visible: $\mathbf{C}$ is an eigenvector of $\mathbf{S}$ for its smallest eigenvalue 1, so $(\mathbf{S}+\mathbf{I}_2)\mathbf{C} = 2\mathbf{C}$, and the solution is $\mathbf{x} = \mathbf{C}/2$, of norm exactly $\|\mathbf{C}\|_2/(\alpha + \beta)$. The data sits where both one-sided hypotheses are tight.

20.7.5 Exercises

20.7.5.1 A. Check your understanding

Exercise. A1

(a) Define $\mathrm{sep}_F(\mathbf{A}, \mathbf{B})$ for $\mathbf{A} \in M_m(\mathbb{C})$ and $\mathbf{B} \in M_n(\mathbb{C})$, and state how it is computed from the Kronecker matrix $\mathbf{K}$.

(b) True or false: $\mathrm{sep}_F(\mathbf{A}, \mathbf{B}) = \min_{i,j}|\lambda_i - \mu_j|$ for all square $\mathbf{A}, \mathbf{B}$. Justify your answer.

(c) True or false: $\mathrm{sep}_F(\mathbf{A} + c\mathbf{I}_m, \mathbf{B} + c\mathbf{I}_n) = \mathrm{sep}_F(\mathbf{A}, \mathbf{B})$ for every $c \in \mathbb{C}$. Justify your answer.

(d) State the bound on the solution of $\mathbf{AX} - \mathbf{XB} = \mathbf{C}$ in terms of the separation, including its hypothesis.

(e) In the proof of Theorem 20.7.3, which vectors are the equation multiplied by, and where are the hypotheses $\mathbf{A} \succeq \alpha\mathbf{I}_m$ and $\mathbf{B} \preceq \beta\mathbf{I}_n$ used?

Solution.

(a) $\mathrm{sep}_F(\mathbf{A}, \mathbf{B}) = \min\{\|\mathbf{AX} - \mathbf{XB}\|_F : \|\mathbf{X}\|_F = 1\}$, the minimum over $\mathbf{X} \in M_{m\times n}(\mathbb{C})$ (Definition 20.7.1). It equals $\sigma_{mn}(\mathbf{K})$, the smallest singular value of $\mathbf{K} = \mathbf{I}_n \otimes \mathbf{A} - \mathbf{B}^\top \otimes \mathbf{I}_m$ (Proposition 20.7.1 (a)).

(b) False. It is true for normal pairs (Theorem 20.7.2), but Example (Separation Far Below the Gap) has spectra at distance 1 and separation $1/10$ when $t = 99/10$. What is always true is the inequality $\le$ of Proposition 20.7.1 (c).

(c) True. For every $\mathbf{X}$, $(\mathbf{A} + c\mathbf{I}_m)\mathbf{X} - \mathbf{X}(\mathbf{B} + c\mathbf{I}_n) = \mathbf{AX} + c\mathbf{X} - \mathbf{XB} - c\mathbf{X} = \mathbf{AX} - \mathbf{XB}$, so the two minima are taken over the same set of values.

(d) If $\operatorname{spec}(\mathbf{A}) \cap \operatorname{spec}(\mathbf{B}) = \varnothing$, the unique solution satisfies $\|\mathbf{X}\|_F \le \|\mathbf{C}\|_F / \operatorname{sep}_F(\mathbf{A}, \mathbf{B})$, and equality holds for some $\mathbf{C} \neq \mathbf{O}$ (Proposition 20.7.1 (d)).

(e) By $\mathbf{u}^*$ on the left and $\mathbf{v}$ on the right, where $\mathbf{u}, \mathbf{v}$ are unit vectors with $\mathbf{X}\mathbf{v} = s\mathbf{u}$ and $\mathbf{X}^*\mathbf{u} = s\mathbf{v}$, $s = \|\mathbf{X}\|_2$. The hypotheses give $\mathbf{u}^*\mathbf{A}\mathbf{u} \ge \alpha$ and $\mathbf{v}^*\mathbf{B}\mathbf{v} \le \beta$, which make $\mathbf{u}^*\mathbf{C}\mathbf{v} = s(\mathbf{u}^*\mathbf{A}\mathbf{u} - \mathbf{v}^*\mathbf{B}\mathbf{v})$ at least $s(\alpha - \beta)$.

20.7.5.2 B. Practice

Exercise. B1: A positive Sylvester equation

Let

$$\mathbf{S} = \begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix}, \quad \mathbf{T} = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}, \quad \mathbf{C} = \begin{pmatrix} 3 & 4 \\ -3 & -4 \end{pmatrix}.$$

(a) Find the largest α and β with $\mathbf{S} \succeq \alpha\mathbf{I}_2$ and $\mathbf{T} \succeq \beta\mathbf{I}_2$, and state the bound that Corollary 20.7.4 gives for $\|\mathbf{X}\|_2$.

(b) Solve $\mathbf{S}\mathbf{X} + \mathbf{X}\mathbf{T} = \mathbf{C}$, column by column.

(c) Compute $\|\mathbf{X}\|_2$ and $\|\mathbf{C}\|_2$, and check the bound. Hence decide whether the bound is attained here.

Solution.

(a) $p_{\mathbf{S}}(x) = (x-3)^2 - 1 = (x-2)(x-4)$, so $\lambda_2(\mathbf{S}) = 2$, and $\lambda_2(\mathbf{T}) = 1$. By Lemma 13.5.2, $\langle \mathbf{S}\mathbf{x}, \mathbf{x} \rangle \ge 2$ for unit $\mathbf{x}$, with equality at an eigenvector, so $\alpha = 2$ is the largest constant with $\mathbf{S} \succeq \alpha\mathbf{I}_2$; likewise $\beta = 1$. The corollary gives $\|\mathbf{X}\|_2 \le \|\mathbf{C}\|_2/3$.

(b) Since $\mathbf{T}$ is diagonal, column j of $\mathbf{X}\mathbf{T}$ is t_{jj} times column j of $\mathbf{X}$. So column j of the equation is $(\mathbf{S} + t_{jj}\mathbf{I}_2)\mathbf{x}_j = \mathbf{c}_j$. For $j = 1$: $\left(\begin{smallmatrix} 4 & 1 \\ 1 & 4 \end{smallmatrix}\right)\mathbf{x}_1 = (3, -3)$, and $\mathbf{x}_1 = (1, -1)$ works, since $4 - 1 = 3$ and $1 - 4 = -3$. For $j = 2$: $\left(\begin{smallmatrix} 5 & 1 \\ 1 & 5 \end{smallmatrix}\right)\mathbf{x}_2 = (4, -4)$, and $\mathbf{x}_2 = (1, -1)$ works. The solution is unique by Corollary 20.7.4, so

$$\mathbf{X} = \begin{pmatrix} 1 & 1 \\ -1 & -1 \end{pmatrix}.$$

(c) Both matrices have rank one: $\mathbf{X} = \mathbf{a}\mathbf{b}^{\top}$ with $\mathbf{a} = (1, -1)$, $\mathbf{b} = (1, 1)$, and $\mathbf{C} = \mathbf{a}\mathbf{c}^{\top}$ with $\mathbf{c} = (3, 4)$. For a rank-one matrix $\mathbf{a}\mathbf{b}^{\top}$, the spectral norm equals the Frobenius norm (Proposition 16.3.4), which is $\|\mathbf{a}\|\,\|\mathbf{b}\|$ by the computation in Proposition 20.7.1 (c). So $\|\mathbf{X}\|_2 = \sqrt{2} \cdot \sqrt{2} = 2$ and $\|\mathbf{C}\|_2 = \sqrt{2} \cdot 5 = 5\sqrt{2}$. The bound reads $2 \le 5\sqrt{2}/3 \approx 2.357$, which holds. It is not attained. Each column of $\mathbf{C}$ lies along the eigenvector $(1, -1)$ of $\mathbf{S}$ for $\alpha = 2$, but the second column is paired with $t_{22} = 2$ rather than $\beta = 1$, so the second column of $\mathbf{X}$ is $\mathbf{c}_2/4$, not $\mathbf{c}_2/3$.

Exercise. B2: Separations of four pairs

For each pair, compute $\operatorname{sep}_F(\mathbf{A}, \mathbf{B})$ and decide whether it equals $\min_{i,j} |\lambda_i - \mu_j|$. Justify your answer.

(a) $\mathbf{A} = \operatorname{diag}(1, 3)$, $\mathbf{B} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$.

(b) $\mathbf{A} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$, $\mathbf{B} = (-1)$.

(c) $\mathbf{A} = \begin{pmatrix} 0 & -2 \\ 2 & 0 \end{pmatrix}$, $\mathbf{B} = \operatorname{diag}(1, -1)$.

(d) $\mathbf{A} = \begin{pmatrix} 1 & 4 \\ 0 & 1 \end{pmatrix}$, $\mathbf{B} = \mathbf{O}_1$.

Solution.

(a) $\mathbf{B}$ is real symmetric with eigenvalues ± 1, and $1 \in \operatorname{spec}(\mathbf{A}) \cap \operatorname{spec}(\mathbf{B})$. So $\operatorname{sep}_F(\mathbf{A}, \mathbf{B}) = 0$ by Proposition 20.7.1 (b), which equals the minimum distance $|1 - 1| = 0$.

(b) Both matrices are real symmetric, hence normal. The eigenvalues of $\mathbf{A}$ are $1, 3$ and that of $\mathbf{B}$ is -1, so Theorem 20.7.2 gives $\operatorname{sep}_F(\mathbf{A}, \mathbf{B}) = \min(2, 4) = 2$. It equals the minimum distance.

(c) $\mathbf{A}^\top \mathbf{A} = 4\mathbf{I}_2 = \mathbf{A}\mathbf{A}^\top$, so $\mathbf{A}$ is normal, with $p_{\mathbf{A}}(x) = x^2 + 4$ and eigenvalues $\pm 2i$. $\mathbf{B}$ is diagonal. The distances $|\pm 2i \mp 1|$ all equal $\sqrt{5}$, so $\operatorname{sep}_F(\mathbf{A}, \mathbf{B}) = \sqrt{5}$ by Theorem 20.7.2, equal to the minimum distance.

(d) This is Example (Separation Far Below the Gap) with $t = 4$: $\operatorname{sep}_F(\mathbf{A}, \mathbf{B}) = (\sqrt{20} - 4)/2 = \sqrt{5} - 2 \approx 0.236$, while the spectra $\{1\}$ and $\{0\}$ are at distance 1. They are not equal. The theorem does not apply because $\mathbf{A}$ is not normal: $\mathbf{A}^\top \mathbf{A} = \left(\begin{smallmatrix} 1 & 4 \\ 4 & 17 \end{smallmatrix}\right) \neq \left(\begin{smallmatrix} 17 & 4 \\ 4 & 1 \end{smallmatrix}\right) = \mathbf{A}\mathbf{A}^\top$.

Exercise. B3: A symmetric Sylvester equation

Let $\mathbf{A} \in M_n(\mathbb{C})$ be Hermitian with $\mathbf{A} \succeq 2\mathbf{I}_n$, and let $\mathbf{C} \in M_n(\mathbb{C})$.

(a) Prove that $\mathbf{A}\mathbf{X} + \mathbf{X}\mathbf{A} = \mathbf{C}$ has exactly one solution, and that it satisfies $\|\mathbf{X}\|_2 \leq \|\mathbf{C}\|_2/4$.

(b) Prove that if $\mathbf{C}$ is Hermitian, then so is $\mathbf{X}$.

Solution.

(a) Apply Corollary 20.7.4 with $\mathbf{S} = \mathbf{T} = \mathbf{A}$ and $\alpha = \beta = 2$. Since $\alpha + \beta = 4 > 0$, the equation has exactly one solution, and $\|\mathbf{X}\|_2 \leq \|\mathbf{C}\|_2/4$.

(b) Take the adjoint of $\mathbf{A}\mathbf{X} + \mathbf{X}\mathbf{A} = \mathbf{C}$. Since $\mathbf{A}^* = \mathbf{A}$ and $\mathbf{C}^* = \mathbf{C}$, we get $\mathbf{X}^*\mathbf{A} + \mathbf{A}\mathbf{X}^* = \mathbf{C}$, so $\mathbf{X}^*$ is also a solution. By the uniqueness in (a), $\mathbf{X}^* = \mathbf{X}$.

20.7.5.3 C. Going deeper

Exercise. C1: The separation is stable, and it measures a distance

Let $\mathbf{A} \in M_m(\mathbb{C})$, $\mathbf{B} \in M_n(\mathbb{C})$, $\mathbf{E} \in M_m(\mathbb{C})$, and $\mathbf{F} \in M_n(\mathbb{C})$.

(a) Prove that $\|\mathbf{E}\mathbf{X}\|_F \leq \|\mathbf{E}\|_2 \|\mathbf{X}\|_F$ and $\|\mathbf{X}\mathbf{F}\|_F \leq \|\mathbf{F}\|_2 \|\mathbf{X}\|_F$ for every $\mathbf{X} \in M_{m \times n}(\mathbb{C})$.

(b) Prove that $|\operatorname{sep}_F(\mathbf{A} + \mathbf{E}, \mathbf{B} + \mathbf{F}) - \operatorname{sep}_F(\mathbf{A}, \mathbf{B})| \leq \|\mathbf{E}\|_2 + \|\mathbf{F}\|_2$.

(c) Now let $n = 1$ and $\mathbf{B} = (\mu)$. Prove that there is $\mathbf{E}$ with $\|\mathbf{E}\|_2 = \operatorname{sep}_F(\mathbf{A}, \mathbf{B})$ such that μ is an eigenvalue of $\mathbf{A} + \mathbf{E}$.

(d) Hence, for $t = 99/10$ in Example (Separation Far Below the Gap), exhibit the size of a perturbation of $\mathbf{A}$ that makes 0 an eigenvalue, and explain why the eigenvalue gap 1 is a poor guide to how close the pair is to having a common eigenvalue.

Hint for (c): use a singular value decomposition of $\mathbf{A} - \mu\mathbf{I}_m$.

Solution.

(a) Let $\mathbf{x}_1, \dots, \mathbf{x}_n$ be the columns of $\mathbf{X}$. The columns of $\mathbf{EX}$ are $\mathbf{E}\mathbf{x}_q$, so by Theorem 16.3.2 (a)

$$\|\mathbf{EX}\|_F^2 = \sum_q \|\mathbf{E}\mathbf{x}_q\|^2 \le \|\mathbf{E}\|_2^2 \sum_q \|\mathbf{x}_q\|^2 = \|\mathbf{E}\|_2^2 \|\mathbf{X}\|_F^2.$$

For the second, $\|\mathbf{XF}\|_F = \|(\mathbf{XF})^*\|_F = \|\mathbf{F}^*\mathbf{X}^*\|_F$, since conjugate transposition does not change the moduli of the entries. By the first part this is at most $\|\mathbf{F}^*\|_2 \|\mathbf{X}^*\|_F = \|\mathbf{F}^*\|_2 \|\mathbf{X}\|_F$. Finally $\|\mathbf{F}^*\|_2 = \|\mathbf{F}\|_2$: if $\mathbf{F} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$ is a singular value decomposition, then $\mathbf{F}^* = \mathbf{V}\mathbf{\Sigma}^\top\mathbf{U}^*$ is one for $\mathbf{F}^*$, with the same diagonal, so $\sigma_1(\mathbf{F}^*) = \sigma_1(\mathbf{F})$ by Theorem 13.9.2 and Theorem 16.3.3 (c).

(b) Let $\|\mathbf{X}\|_F = 1$. By the triangle inequality (Corollary 11.1.6, the Frobenius norm being induced by an inner product) and (a),

$$\begin{aligned}\|(\mathbf{A}+\mathbf{E})\mathbf{X} - \mathbf{X}(\mathbf{B}+\mathbf{F})\|_F &\ge \|\mathbf{AX} - \mathbf{XB}\|_F - \|\mathbf{EX}\|_F - \|\mathbf{XF}\|_F \\ &\ge \operatorname{sep}_F(\mathbf{A}, \mathbf{B}) - \|\mathbf{E}\|_2 - \|\mathbf{F}\|_2.\end{aligned}$$

Taking the minimum over $\mathbf{X}$ gives $\operatorname{sep}_F(\mathbf{A}+\mathbf{E}, \mathbf{B}+\mathbf{F}) \ge \operatorname{sep}_F(\mathbf{A}, \mathbf{B}) - \|\mathbf{E}\|_2 - \|\mathbf{F}\|_2$. Applying this to the pair $(\mathbf{A}+\mathbf{E}, \mathbf{B}+\mathbf{F})$ with the perturbations $-\mathbf{E}, -\mathbf{F}$ gives the reverse inequality, and together they are the claim.

(c) By Proposition 20.7.1 (f), $s := \operatorname{sep}_F(\mathbf{A}, \mathbf{B}) = \sigma_m(\mathbf{A} - \mu\mathbf{I}_m)$. Let $\mathbf{A} - \mu\mathbf{I}_m = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$ be a singular value decomposition (Theorem 13.9.1), with last columns $\mathbf{u}_m$ of $\mathbf{U}$ and $\mathbf{v}_m$ of $\mathbf{V}$. Then $(\mathbf{A} - \mu\mathbf{I}_m)\mathbf{v}_m = s\mathbf{u}_m$, comparing column m of $(\mathbf{A} - \mu\mathbf{I}_m)\mathbf{V} = \mathbf{U}\mathbf{\Sigma}$. Put $\mathbf{E} = -s\,\mathbf{u}_m\mathbf{v}_m^*$. Then $(\mathbf{A} + \mathbf{E} - \mu\mathbf{I}_m)\mathbf{v}_m = s\mathbf{u}_m - s\mathbf{u}_m(\mathbf{v}_m^*\mathbf{v}_m) = \mathbf{0}$ with $\mathbf{v}_m \neq \mathbf{0}$, so μ is an eigenvalue of $\mathbf{A} + \mathbf{E}$. Also $\|\mathbf{Ex}\| = s|\mathbf{v}_m^*\mathbf{x}| \le s\|\mathbf{x}\|$ by Cauchy–Schwarz, with equality at $\mathbf{x} = \mathbf{v}_m$, so $\|\mathbf{E}\|_2 = s$.

(d) With $\mu = 0$ and $\operatorname{sep}_F(\mathbf{A}, \mathbf{O}_1) = 1/10$, part (c) gives a perturbation of spectral norm $1/10$ after which 0 is an eigenvalue of $\mathbf{A}+\mathbf{E}$, so $\mathbf{A}+\mathbf{E}$ and $\mathbf{B}$ share an eigenvalue. The eigenvalues 1 and 0 are at distance 1, but a perturbation ten times smaller than that distance destroys the solvability of the equation. By (b) with $\mathbf{F} = \mathbf{O}$, no perturbation of spectral norm less than $1/10$ can do it, since it leaves the separation positive. So the separation, not the eigenvalue gap, is the distance to a shared eigenvalue here.

Exercise. C2: Interlaced spectra break the spectral-norm bound

Let $\mathbf{A} = \operatorname{diag}(0, 2)$, $\mathbf{B} = \operatorname{diag}(1, 3)$, and $\mathbf{C} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$.

(a) Show that every eigenvalue of $\mathbf{A}$ is at distance at least 1 from every eigenvalue of $\mathbf{B}$, and explain why Theorem 20.7.3 does not apply with $\alpha - \beta = 1$.

(b) Solve $\mathbf{AX} - \mathbf{XB} = \mathbf{C}$.

(c) Show that $\|\mathbf{X}\|_2 > \|\mathbf{C}\|_2$.

(d) Verify directly that $\|\mathbf{X}\|_F \le \|\mathbf{C}\|_F$, as Theorem 20.7.2 requires.

Hint for (c): compute $\mathbf{C}^\top\mathbf{C}$*, and test* $\mathbf{X}$ *on the vector* $(1, 1)$.

Solution.

(a) The eigenvalues are $0, 2$ and $1, 3$, and the four differences $\lambda_i - \mu_j$ are $-1, -3, 1, -1$, all of modulus at least 1. The theorem needs $\mathbf{A} \succeq \alpha\mathbf{I}$ and $\mathbf{B} \preceq \beta\mathbf{I}$ with $\alpha > \beta$, that is, every eigenvalue of $\mathbf{A}$ above every eigenvalue of $\mathbf{B}$, or, applying it to $(-\mathbf{A}, -\mathbf{B})$, every one below. Neither holds, since $0 < 1 < 2 < 3$ interlace. Applied honestly, the best one-sided constants are $\alpha = 0$ and $\beta = 3$, and $\alpha > \beta$ fails.

(b) As in the diagonal example after Definition 20.7.1, $(\mathbf{AX} - \mathbf{XB})_{ij} = (a_{ii} - b_{jj})x_{ij}$, so $x_{ij} = c_{ij}/(a_{ii} - b_{jj})$:

$$\mathbf{X} = \begin{pmatrix} 1/(0-1) & 1/(0-3) \\ 1/(2-1) & -1/(2-3) \end{pmatrix} = \begin{pmatrix} -1 & -1/3 \\ 1 & 1 \end{pmatrix}.$$

(c) $\mathbf{C}^\top\mathbf{C} = 2\mathbf{I}_2$, so both singular values of $\mathbf{C}$ are $\sqrt{2}$ and $\|\mathbf{C}\|_2 = \sqrt{2}$ (Theorem 16.3.3 (c)). For $\mathbf{x} = (1, 1)/\sqrt{2}$, a unit vector, $\mathbf{Xx} = (-4/3, 2)/\sqrt{2}$, so

$$\|\mathbf{Xx}\|^2 = \frac{1}{2}\Big(\frac{16}{9} + 4\Big) = \frac{26}{9} > 2.$$

Hence $\|\mathbf{X}\|_2 \ge \sqrt{26}/3 > \sqrt{2} = \|\mathbf{C}\|_2$.

(d) $\|\mathbf{X}\|_F^2 = 1 + \frac{1}{9} + 1 + 1 = \frac{28}{9}$ and $\|\mathbf{C}\|_F^2 = 4$, so $\|\mathbf{X}\|_F \le \|\mathbf{C}\|_F$, in agreement with Theorem 20.7.2, which gives $\operatorname{sep}_F(\mathbf{A}, \mathbf{B}) = 1$. So the Frobenius bound with constant $1/\delta$ survives interlacing, while the spectral-norm bound with the same constant does not.

20.8 Principal Angles

The sections so far measured how far eigenvalues move. The next section asks how far **eigenvectors** move, and more generally invariant subspaces, and before it can answer it needs a way to say how far apart two subspaces are. For two lines there is an obvious answer, the angle between them. This section shows that for subspaces of dimension $k \ge 2$ one angle is not enough, and that the right replacement is a list of k angles read off from a singular value decomposition. It then proves the two facts Section 9 will use: the sines of these angles are the singular values of a matrix built from the orthogonal complement, and the largest sine is the distance between the two orthogonal projections.

Throughout, F is $\mathbb{R}$ or $\mathbb{C}$, and F^n carries its standard inner product $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{y}^*\mathbf{x}$, with $\mathbf{x}^* = \mathbf{x}^\top$ when $F = \mathbb{R}$. For a subspace U of F^n, P_U is the orthogonal projection onto U (Definition 11.3.2), identified with its matrix in the standard basis. Singular values of a matrix with k columns follow the convention of Chapter 17 §09: $\sigma_i := 0$ for $\min(\text{rows}, k) < i \le k$, so that σ_i^2 is the i-th largest eigenvalue of $\mathbf{A}^*\mathbf{A}$ for every $i \le k$ (Eq. 17.9.1).

20.8.1 One angle is not enough

For two lines $\operatorname{span}(\mathbf{u})$ and $\operatorname{span}(\mathbf{w})$ there is one number to compute, the angle between them. For planes the first idea is to copy it: call the angle between U and W the **smallest** angle between a line in U and a line in W.

That idea fails at once in $\mathbb{R}^3$. Two distinct planes $U, W \subseteq \mathbb{R}^3$ always meet in a line, since

$$\dim(U \cap W) = 2 + 2 - \dim(U + W) \geq 4 - 3 = 1$$

by Theorem 1.7.3. So the smallest angle is 0 for **every** pair of planes: it cannot tell a plane tilted by one degree from a plane at right angles.

Taking the **largest** such angle instead does no better in $\mathbb{R}^4$. Let $U = \operatorname{span}(\mathbf{e}_1, \mathbf{e}_2)$ and compare

$$W_1 = \operatorname{span}(\mathbf{e}_1, \mathbf{e}_3), \qquad W_2 = \operatorname{span}(\mathbf{e}_3, \mathbf{e}_4).$$

Each contains a vector perpendicular to all of U (namely $\mathbf{e}_3$), so each has a line at angle $\pi/2$ from U. But W_1 shares the whole line $\operatorname{span}(\mathbf{e}_1)$ with U, while W_2 is perpendicular to U in every direction. One number cannot record both facts. A list can: $(0, \pi/2)$ for W_1 and $(\pi/2, \pi/2)$ for W_2. The rest of this section explains where such a list comes from in general, when the subspaces are not lined up with the coordinate axes.

Two subspaces can be given orthonormal bases that face each other in pairs; the principal angles are the angles between the partners.

20.8.2 Principal angles

Everything will be computed from orthonormal bases written as the columns of a matrix, so we first record how such matrices behave.

Let $U \subseteq F^n$ be a subspace with $\dim U = k \geq 1$. An **orthonormal basis matrix** of U is a matrix $\mathbf{Q} \in M_{n\times k}(F)$ whose columns form an orthonormal basis of U. One exists: apply Gram–Schmidt (Theorem 11.2.2) to any basis of U.

Lemma 20.8.1. Orthonormal Basis Matrices

Let $U \subseteq F^n$ have dimension $k \geq 1$ and let $\mathbf{Q} \in M_{n\times k}(F)$ be an orthonormal basis matrix of U.

(a) $\mathbf{Q}^*\mathbf{Q} = \mathbf{I}_k$ and $P_U = \mathbf{Q}\mathbf{Q}^*$.

(b) $\|\mathbf{Q}\mathbf{y}\| = \|\mathbf{y}\|$ for every $\mathbf{y} \in F^k$. For every matrix $\mathbf{Y}$ with k rows and every matrix $\mathbf{Z}$ with k columns,

$$\|\mathbf{Q}\mathbf{Y}\|_2 = \|\mathbf{Y}\|_2, \quad \|\mathbf{Q}\mathbf{Y}\|_F = \|\mathbf{Y}\|_F,$$
$$\|\mathbf{Z}\mathbf{Q}^*\|_2 = \|\mathbf{Z}\|_2, \quad \|\mathbf{Z}\mathbf{Q}^*\|_F = \|\mathbf{Z}\|_F.$$

(c) Every orthonormal basis matrix of U is $\mathbf{Q}\mathbf{R}$ for some $\mathbf{R} \in \mathrm{U}(k)$ (orthogonal when $F = \mathbb{R}$).

(d) $\mathbf{I}_n - P_U = P_{U^\perp}$. If moreover $k < n$ and $\mathbf{Q}_\perp \in M_{n\times(n-k)}(F)$ is an orthonormal basis matrix of $U^\perp$, then the square matrix $(\mathbf{Q} \mid \mathbf{Q}_\perp)$ is unitary, and

$$\mathbf{Q}\mathbf{Q}^* + \mathbf{Q}_\perp\mathbf{Q}_\perp^* = \mathbf{I}_n.$$

Proof.

(a) The (i,j) entry of $\mathbf{Q}^*\mathbf{Q}$ is $\mathbf{q}_i^*\mathbf{q}_j = \langle \mathbf{q}_j, \mathbf{q}_i\rangle = \delta_{ij}$, because the columns $\mathbf{q}_1, \dots, \mathbf{q}_k$ are orthonormal. By Theorem 11.3.3 (a), for every $\mathbf{v} \in F^n$,

$$P_U\mathbf{v} = \sum_{i=1}^{k} \langle \mathbf{v}, \mathbf{q}_i\rangle \mathbf{q}_i = \sum_{i=1}^{k} \mathbf{q}_i(\mathbf{q}_i^*\mathbf{v}) = \mathbf{Q}\mathbf{Q}^*\mathbf{v},$$

the last step being column-times-row expansion of $\mathbf{Q}\mathbf{Q}^*$.

(b) By (a), $\|\mathbf{Q}\mathbf{y}\|^2 = \mathbf{y}^*\mathbf{Q}^*\mathbf{Q}\mathbf{y} = \mathbf{y}^*\mathbf{y} = \|\mathbf{y}\|^2$. Hence $\|\mathbf{Q}\mathbf{Y}\mathbf{x}\| = \|\mathbf{Y}\mathbf{x}\|$ for every $\mathbf{x}$, and taking the maximum over unit $\mathbf{x}$ (Definition 16.3.2) gives $\|\mathbf{Q}\mathbf{Y}\|_2 = \|\mathbf{Y}\|_2$. For the Frobenius norm, $\|\mathbf{Q}\mathbf{Y}\|_F^2 = \operatorname{tr}(\mathbf{Y}^*\mathbf{Q}^*\mathbf{Q}\mathbf{Y}) = \operatorname{tr}(\mathbf{Y}^*\mathbf{Y}) = \|\mathbf{Y}\|_F^2$. Next, $\|\mathbf{Z}\mathbf{Q}^*\|_F^2 = \operatorname{tr}(\mathbf{Q}\mathbf{Z}^*\mathbf{Z}\mathbf{Q}^*) = \operatorname{tr}(\mathbf{Z}^*\mathbf{Z}\mathbf{Q}^*\mathbf{Q}) = \|\mathbf{Z}\|_F^2$, by Theorem 0.9.6 (3). For the spectral norm, let $\mathbf{x} \in F^n$. By the first sentence and (a), $\|\mathbf{Q}^*\mathbf{x}\| = \|\mathbf{Q}\mathbf{Q}^*\mathbf{x}\| = \|P_U\mathbf{x}\| \le \|\mathbf{x}\|$ (Theorem 11.3.3 (c)). So $\|\mathbf{Z}\mathbf{Q}^*\mathbf{x}\| \le \|\mathbf{Z}\|_2\|\mathbf{Q}^*\mathbf{x}\| \le \|\mathbf{Z}\|_2\|\mathbf{x}\|$ by Theorem 16.3.2 (a), and $\|\mathbf{Z}\mathbf{Q}^*\|_2 \le \|\mathbf{Z}\|_2$. Conversely, choose a unit $\mathbf{y} \in F^k$ with $\|\mathbf{Z}\mathbf{y}\| = \|\mathbf{Z}\|_2$ (Lemma 16.3.1). Then $\mathbf{x} = \mathbf{Q}\mathbf{y}$ is a unit vector with $\mathbf{Z}\mathbf{Q}^*\mathbf{x} = \mathbf{Z}\mathbf{y}$, so $\|\mathbf{Z}\mathbf{Q}^*\|_2 \ge \|\mathbf{Z}\|_2$.

(c) Let $\mathbf{Q}'$ be another orthonormal basis matrix of U, and put $\mathbf{R} = \mathbf{Q}^*\mathbf{Q}' \in M_k(F)$. Each column of $\mathbf{Q}'$ lies in U, where P_U is the identity, so $\mathbf{Q}' = P_U\mathbf{Q}' = \mathbf{Q}\mathbf{Q}^*\mathbf{Q}' = \mathbf{Q}\mathbf{R}$ by (a). Then $\mathbf{R}^*\mathbf{R} = \mathbf{Q}'^*\mathbf{Q}\mathbf{Q}^*\mathbf{Q}' = \mathbf{Q}'^*P_U\mathbf{Q}' = \mathbf{Q}'^*\mathbf{Q}' = \mathbf{I}_k$, so the square matrix $\mathbf{R}$ is unitary (Definition 11.7.2).

(d) For the first identity, let $\mathbf{v} \in F^n$ and write $\mathbf{v} = \mathbf{u}+\mathbf{w}$ with $\mathbf{u} \in U$ and $\mathbf{w} \in U^\perp$, which is possible in exactly one way by Theorem 11.3.2 (a). Then $P_U\mathbf{v} = \mathbf{u}$ by Definition 11.3.2. Since $(U^\perp)^\perp = U$ by Theorem 11.3.2 (b), the splitting $\mathbf{v} = \mathbf{w} + \mathbf{u}$ is also the one that defines $P_{U^\perp}$, so $P_{U^\perp}\mathbf{v} = \mathbf{w} = \mathbf{v} - P_U\mathbf{v}$. As $\mathbf{v}$ was arbitrary, $\mathbf{I}_n - P_U = P_{U^\perp}$. Nothing here restricts k: when $k = n$, $U^\perp = \{\mathbf{0}\}$ and both sides are $\mathbf{O}$.

Now let $k < n$. The columns of $\mathbf{Q}$ and of $\mathbf{Q}_\perp$ are orthonormal, and each column of $\mathbf{Q}$ is orthogonal to each column of $\mathbf{Q}_\perp$ because $U \perp U^\perp$. Their number is $k + (n-k) = n$, by Theorem 11.3.2 (c). So the square matrix $\mathbf{W} = (\mathbf{Q} \mid \mathbf{Q}_\perp) \in M_n(F)$ satisfies $\mathbf{W}^*\mathbf{W} = \mathbf{I}_n$, hence is unitary, and then $\mathbf{W}\mathbf{W}^* = \mathbf{I}_n$ as well, since $\mathbf{W}^* = \mathbf{W}^{-1}$. By block multiplication (Theorem 8.1.1), $\mathbf{W}\mathbf{W}^* = \mathbf{Q}\mathbf{Q}^* + \mathbf{Q}_\perp\mathbf{Q}_\perp^*$. This also re-proves the first identity here, since $\mathbf{Q}_\perp\mathbf{Q}_\perp^* = P_{U^\perp}$ by (a) applied to $U^\perp$.

■

Now the definition. The question is how much of U survives when it is projected into W. By parts (a) and (b) of the lemma, for a unit vector $\mathbf{u} = \mathbf{Q}_U\mathbf{y}$ of U the surviving length is $\|P_W\mathbf{u}\| = \|\mathbf{Q}_W\mathbf{Q}_W^*\mathbf{u}\| = \|\mathbf{Q}_W^*\mathbf{u}\| = \|(\mathbf{Q}_W^*\mathbf{Q}_U)\mathbf{y}\|$. So the matrix $\mathbf{Q}_W^*\mathbf{Q}_U$ records the answer for every direction of U at once, and its singular values are the stretch factors along the best directions.

Definition 20.8.1. Principal Angles

Let $U, W \subseteq F^n$ be subspaces with

$$1 \le k = \dim U \le l = \dim W,$$

and let $\mathbf{Q}_U \in M_{n\times k}(F)$ and $\mathbf{Q}_W \in M_{n\times l}(F)$ be orthonormal basis matrices of U and W.

The **principal angles** between U and W are the k numbers

$$0 \le \theta_1 \le \theta_2 \le \cdots \le \theta_k \le \frac{\pi}{2}$$

determined by

$$\cos\theta_i = \sigma_i(\mathbf{Q}_W^* \mathbf{Q}_U) \qquad (i = 1, \dots, k).$$

We write $\Theta(U, W) = \operatorname{diag}(\theta_1, \dots, \theta_k)$ and $\sin\Theta(U, W) = \operatorname{diag}(\sin\theta_1, \dots, \sin\theta_k)$.

In words: fix orthonormal bases of the two subspaces, form the $l \times k$ matrix whose (i, j) entry is the inner product of the j-th basis vector of U with the i-th basis vector of W, and take its k singular values. These lie in $[0, 1]$, as we check next, so each is the cosine of exactly one angle in $[0, \pi/2]$. The largest cosine gives the **smallest** angle, so the decreasing singular values produce increasing angles. The number of angles is the **smaller** dimension, k, and the definition asks for $k \le l$; for a pair with $\dim U > \dim W$, swap the names. The two notations at the end are for use in norms: $\|\sin\Theta(U, W)\|_2 = \sin\theta_k$, the largest entry of a nonnegative diagonal matrix (Lemma 17.3.5), and $\|\sin\Theta(U, W)\|_F = \left(\sum_i \sin^2\theta_i\right)^{1/2}$.

Well-definedness. Three things need checking. The answer must not depend on the orthonormal bases chosen. The singular values must be at most 1, or there is no angle. And when $k = l$ the definition treats U and W asymmetrically, so we should check that swapping them changes nothing.

Proposition 20.8.2. Principal Angles Are Well Defined

In the setting of Definition 20.8.1, put $\mathbf{M} = \mathbf{Q}_W^* \mathbf{Q}_U \in M_{l\times k}(F)$.

(a) The singular values of $\mathbf{M}$ do not depend on the choice of the orthonormal basis matrices $\mathbf{Q}_U$ and $\mathbf{Q}_W$.

(b) $1 \ge \sigma_1(\mathbf{M}) \ge \cdots \ge \sigma_k(\mathbf{M}) \ge 0$, so $\theta_1, \dots, \theta_k$ exist, are unique, and are nondecreasing.

(c) $\sigma_i(\mathbf{Q}_U^* \mathbf{Q}_W) = \sigma_i(\mathbf{M})$ for $i = 1, \dots, k$. In particular, if $k = l$, the principal angles between W and U are those between U and W.

(d) If $k = 1$ and $U = \operatorname{span}(\mathbf{u})$ with $\|\mathbf{u}\| = 1$, then $\cos\theta_1 = \|P_W \mathbf{u}\|$. If also $l = 1$ and $W = \operatorname{span}(\mathbf{w})$ with $\|\mathbf{w}\| = 1$, then $\cos\theta_1 = |\langle \mathbf{u}, \mathbf{w}\rangle|$.

Proof.

(a) By Lemma 20.8.1 (c), any other choices are $\mathbf{Q}_U\mathbf{R}$ and $\mathbf{Q}_W\mathbf{R}'$ with $\mathbf{R} \in \mathrm{U}(k)$ and $\mathbf{R}' \in \mathrm{U}(l)$. The new matrix is $\mathbf{M}' = \mathbf{R}'^* \mathbf{M} \mathbf{R}$, and

$$\mathbf{M}'^* \mathbf{M}' = \mathbf{R}^* \mathbf{M}^* \mathbf{R}' \mathbf{R}'^* \mathbf{M} \mathbf{R} = \mathbf{R}^* (\mathbf{M}^* \mathbf{M}) \mathbf{R},$$

using $\mathbf{R}'\mathbf{R}'^* = \mathbf{I}_l$. Since $\mathbf{R}^* = \mathbf{R}^{-1}$, this is similar to $\mathbf{M}^*\mathbf{M}$, so the two have the same characteristic polynomial (Theorem 7.10.3) and the same eigenvalues with multiplicity. By Definition 13.9.1, $\mathbf{M}'$ and $\mathbf{M}$ have the same singular values.

(b) By Theorem 16.3.2 (d) and Lemma 20.8.1 (b), $\|\mathbf{M}\|_2 \le \|\mathbf{Q}_W^*\|_2 \|\mathbf{Q}_U\|_2$, and $\|\mathbf{Q}_U\|_2 = \|\mathbf{Q}_U \mathbf{I}_k\|_2 = \|\mathbf{I}_k\|_2 = 1$ by Theorem 16.3.2 (c). Also $\|\mathbf{Q}_W^* \mathbf{x}\| = \|P_W \mathbf{x}\| \le \|\mathbf{x}\|$, as in the proof of that lemma, so $\|\mathbf{Q}_W^*\|_2 \le 1$. Hence $\sigma_1(\mathbf{M}) = \|\mathbf{M}\|_2 \le 1$ by Theorem 16.3.3 (c), and all singular values lie in $[0, 1]$. The cosine is a strictly decreasing bijection from $[0, \pi/2]$ onto $[0, 1]$, so each $\sigma_i(\mathbf{M})$ is $\cos\theta_i$ for exactly one $\theta_i \in [0, \pi/2]$, and $\sigma_1 \ge \cdots \ge \sigma_k$ gives $\theta_1 \le \cdots \le \theta_k$.

(c) Let $\mathbf{M} = \mathbf{Y}\mathbf{\Sigma}\mathbf{Z}^*$ be a singular value decomposition (Theorem 13.9.1). Then $\mathbf{M}^* = \mathbf{Z}\mathbf{\Sigma}^\top\mathbf{Y}^*$, where $\mathbf{Z}, \mathbf{Y}$ are unitary and $\mathbf{\Sigma}^\top \in M_{k\times l}(\mathbb{R})$ carries the same diagonal entries $\sigma_1(\mathbf{M}) \geq \cdots \geq \sigma_k(\mathbf{M})$ and zeros elsewhere. So this is a factorization of $\mathbf{M}^*$ of the shape in Theorem 13.9.1, and Theorem 13.9.2 gives $\sigma_i(\mathbf{M}^*) = \sigma_i(\mathbf{M})$. Since $\mathbf{M}^* = \mathbf{Q}_U^*\mathbf{Q}_W$, this is the first claim. When $k = l$, the principal angles between W and U are defined by $\sigma_i(\mathbf{Q}_U^*\mathbf{Q}_W)$, so they coincide with those between U and W.

(d) With $\mathbf{Q}_U = \mathbf{u}$, the matrix $\mathbf{M} = \mathbf{Q}_W^*\mathbf{u}$ is a single column, and its one singular value is $(\mathbf{M}^*\mathbf{M})^{1/2} = \|\mathbf{Q}_W^*\mathbf{u}\|$. By Lemma 20.8.1 (b) and (a), $\|\mathbf{Q}_W^*\mathbf{u}\| = \|\mathbf{Q}_W\mathbf{Q}_W^*\mathbf{u}\| = \|P_W\mathbf{u}\|$. If $l = 1$ and $\mathbf{Q}_W = \mathbf{w}$, then $\mathbf{M} = \mathbf{w}^*\mathbf{u} = \langle \mathbf{u}, \mathbf{w}\rangle$, whose singular value is $|\langle \mathbf{u}, \mathbf{w}\rangle|$.

■

Examples.

1. **Two lines in $\mathbb{R}^2$.** For $\mathbf{u} = \mathbf{e}_1$ and $\mathbf{w} = (1, 1)/\sqrt{2}$, part (d) gives $\cos\theta_1 = 1/\sqrt{2}$, so $\theta_1 = \pi/4$. Replacing $\mathbf{w}$ by $-\mathbf{w}$ changes neither the line nor the answer.
2. **Nested subspaces.** If $U \subseteq W$, then every column of $\mathbf{Q}_U$ is fixed by P_W, so $\mathbf{M}^*\mathbf{M} = \mathbf{Q}_U^*\mathbf{Q}_W\mathbf{Q}_W^*\mathbf{Q}_U = \mathbf{Q}_U^*P_W\mathbf{Q}_U = \mathbf{Q}_U^*\mathbf{Q}_U = \mathbf{I}_k$. All singular values are 1 and all principal angles are 0. The degenerate case $U = W$ is included, and it is the reason the angles start at 0.
3. **Perpendicular subspaces.** If $U \subseteq W^\perp$, every entry of $\mathbf{M}$ is the inner product of a vector of U with a vector of W, hence 0, so all principal angles are $\pi/2$.
4. **The planes of the opening.** For $U = \operatorname{span}(\mathbf{e}_1, \mathbf{e}_2)$ and $W_1 = \operatorname{span}(\mathbf{e}_1, \mathbf{e}_3)$, take the standard vectors as orthonormal bases. Then $\mathbf{M} = \mathbf{Q}_{W_1}^\top\mathbf{Q}_U = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 0 \end{smallmatrix}\right)$, with singular values $1, 0$, so the principal angles are $(0, \pi/2)$. For $W_2 = \operatorname{span}(\mathbf{e}_3, \mathbf{e}_4)$ we are in Example 3, with angles $(\pi/2, \pi/2)$. These are the two lists that the single-number attempts could not tell apart.

Nonexample by minimal change. In Example 4, replace the orthonormal basis $(\mathbf{e}_1, \mathbf{e}_3)$ of W_1 by the basis $(\mathbf{e}_1, (\mathbf{e}_1+\mathbf{e}_3)/\sqrt{2})$, whose vectors still have length 1 but are no longer orthogonal. The matrix of inner products becomes $\left(\begin{smallmatrix} 1 & 0 \\ 1/\sqrt{2} & 0 \end{smallmatrix}\right)$, with largest singular value $\sqrt{3/2} > 1$, and no angle has that cosine. What fails is the orthogonality of the columns of $\mathbf{Q}_W$: without it, $\mathbf{Q}_W\mathbf{Q}_W^*$ is not the projection P_W, and the matrix no longer measures how much of U survives in W.

Why this definition: the principal vectors. The singular value decomposition of $\mathbf{M}$ does more than produce numbers. It produces bases of U and W in which the whole configuration is visible, and this is the precise meaning of the slogan.

Proposition 20.8.3. Principal Vectors

In the setting of Definition 20.8.1, let $\mathbf{Q}_W^*\mathbf{Q}_U = \mathbf{Y}\mathbf{\Sigma}\mathbf{Z}^*$ be a singular value decomposition (Theorem 13.9.1), with $\mathbf{Y} \in M_l(F)$ and $\mathbf{Z} \in M_k(F)$ unitary, columns $\mathbf{y}_1, \ldots, \mathbf{y}_l$ and $\mathbf{z}_1, \ldots, \mathbf{z}_k$. Put

$$\mathbf{u}_i = \mathbf{Q}_U\mathbf{z}_i \ (1 \leq i \leq k), \qquad \mathbf{w}_j = \mathbf{Q}_W\mathbf{y}_j \ (1 \leq j \leq l).$$

Then $(\mathbf{u}_1, \ldots, \mathbf{u}_k)$ is an orthonormal basis of U, $(\mathbf{w}_1, \ldots, \mathbf{w}_l)$ is an orthonormal basis of W, and

$$\langle \mathbf{u}_i, \mathbf{w}_j\rangle = \begin{cases} \cos\theta_i & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases}$$

Proof. The matrix $\mathbf{Q}_U\mathbf{Z}$ has columns $\mathbf{u}_1, \dots, \mathbf{u}_k$, all in U, and $(\mathbf{Q}_U\mathbf{Z})^*(\mathbf{Q}_U\mathbf{Z}) = \mathbf{Z}^*\mathbf{Z} = \mathbf{I}_k$ by Lemma 20.8.1 (a). So the $\mathbf{u}_i$ are k orthonormal vectors of U. They are linearly independent (Theorem 11.1.4), and $\dim U = k$, so they form a basis. The same argument applies to the $\mathbf{w}_j$ in W. Finally,

$$\langle \mathbf{u}_i, \mathbf{w}_j \rangle = \mathbf{w}_j^* \mathbf{u}_i = \mathbf{y}_j^* (\mathbf{Q}_W^* \mathbf{Q}_U) \mathbf{z}_i = \mathbf{y}_j^* \mathbf{Y} \mathbf{\Sigma} \mathbf{Z}^* \mathbf{z}_i = \mathbf{e}_j^* \mathbf{\Sigma} \mathbf{e}_i,$$

which is the (j, i) entry of $\mathbf{\Sigma}$: it equals $\sigma_i = \cos\theta_i$ when $j = i$, and 0 otherwise.

■

So U and W have orthonormal bases that face each other in pairs. Each $\mathbf{u}_i$ makes the angle θ_i with its partner $\mathbf{w}_i$, in the sense of part (d) of the proposition, and is perpendicular to every other $\mathbf{w}_j$. The last $l - k$ vectors of W have no partner, and they are perpendicular to all of U. The vectors $\mathbf{u}_i, \mathbf{w}_i$ are called **principal vectors**. Two subspaces, however they sit in F^n, look like a list of k independent pairs of lines, and the principal angles are the angles within the pairs.

The principal vectors are usually **not** the basis vectors one started with. For $U = W = \mathbb{R}^2$ with bases $(\mathbf{e}_1, \mathbf{e}_2)$ and $((\mathbf{e}_1 + \mathbf{e}_2)/\sqrt{2}, (\mathbf{e}_1 - \mathbf{e}_2)/\sqrt{2})$, each given basis vector of U is at angle $\pi/4$ from each given basis vector of W. The principal angles are nevertheless $0, 0$, by Example 2, since the subspaces are equal. Pairing the given columns one by one measures the bases, not the subspaces.

Warning. **Lines have no orientation, so principal angles use $|\langle \mathbf{u}, \mathbf{w} \rangle|$ and lie in $[0, \pi/2]$.** The angle of Definition 11.1.5 is between **vectors** in a **real** space, and it can reach π. For $\mathbf{u} = \mathbf{e}_1$ and $\mathbf{w} = -\mathbf{e}_1$ in $\mathbb{R}^2$ that angle is π, yet $\operatorname{span}(\mathbf{u}) = \operatorname{span}(\mathbf{w})$ and the principal angle is 0. Over $\mathbb{C}$ the vector angle is not even defined: $\mathbf{u} = \mathbf{e}_1$ and $\mathbf{w} = i\mathbf{e}_1$ span the same complex line, $\langle \mathbf{u}, \mathbf{w} \rangle = \bar{i} = -i$ is not real, and the principal angle is again 0 because $|-i| = 1$.

Quick check. Find the principal angle between the line $U = \operatorname{span}((1, 2, 2))$ and the plane $W = \{\mathbf{x} \in \mathbb{R}^3 : x_3 = 0\}$.

Solution. Here $k = 1 \le 2 = l$. The unit vector $\mathbf{u} = \frac{1}{3}(1, 2, 2)$ spans U, and $P_W\mathbf{u} = \frac{1}{3}(1, 2, 0)$, since $(\mathbf{e}_1, \mathbf{e}_2)$ is an orthonormal basis of W (Theorem 11.3.3 (a)). By Proposition 20.8.2 (d), $\cos\theta_1 = \|P_W\mathbf{u}\| = \sqrt{5}/3$. So $\theta_1 = \arccos(\sqrt{5}/3) \approx 41.8°$, and $\sin\theta_1 = 2/3$, which is the distance from $\mathbf{u}$ to the plane.

20.8.3 Sines and the orthogonal complement

The cosines measure how much of U survives projection into W. What Section 9 can actually bound is the part that does **not** survive: the component of U in $W^\perp$. Pythagoras says the two parts are complementary, and the theorem makes this exact.

Theorem 20.8.4. Sines of Principal Angles

Let $U, W \subseteq F^n$ with $1 \le k = \dim U \le l = \dim W$, principal angles $\theta_1 \le \cdots \le \theta_k$, and orthonormal basis matrices $\mathbf{Q}_U, \mathbf{Q}_W$.

(a) Suppose $l < n$, and let $\mathbf{Q}_{W^\perp} \in M_{n\times(n-l)}(F)$ be an orthonormal basis matrix of $W^\perp$. Put $\mathbf{M} = \mathbf{Q}_W^* \mathbf{Q}_U$ and $\mathbf{N} = \mathbf{Q}_{W^\perp}^* \mathbf{Q}_U \in M_{(n-l)\times k}(F)$. Then

$$\mathbf{M}^*\mathbf{M} + \mathbf{N}^*\mathbf{N} = \mathbf{I}_k,$$

and the eigenvalues of $\mathbf{N}^*\mathbf{N}$, with multiplicity, are $\sin^2\theta_1, \ldots, \sin^2\theta_k$. Equivalently, $\sigma_i(\mathbf{N}) = \sin\theta_{k+1-i}$ for $i = 1, \ldots, k$, with the convention for σ_i stated at the start of the section.

(b) For every $l \le n$,

$$\|(\mathbf{I} - P_W)P_U\|_2 = \|(\mathbf{I} - P_W)\mathbf{Q}_U\|_2 = \sin\theta_k = \|\sin\Theta(U, W)\|_2,$$

and

$$\|(\mathbf{I} - P_W)P_U\|_F = \|(\mathbf{I} - P_W)\mathbf{Q}_U\|_F = \|\sin\Theta(U, W)\|_F.$$

Idea. Split the identity of F^n along $W \oplus W^\perp$ and sandwich it between $\mathbf{Q}_U^*$ and $\mathbf{Q}_U$. That gives two Hermitian $k \times k$ matrices adding up to $\mathbf{I}_k$, and such matrices share their eigenvectors, with eigenvalues c and $1 - c$. The first has eigenvalues $\cos^2\theta_i$ by definition, so the second has eigenvalues $\sin^2\theta_i$. Part (b) then removes the orthonormal factors, which change no norm.

Proof.

(a) By Lemma 20.8.1 (d) applied to W, $\mathbf{I}_n = \mathbf{Q}_W\mathbf{Q}_W^* + \mathbf{Q}_{W^\perp}\mathbf{Q}_{W^\perp}^*$. Multiplying on the left by $\mathbf{Q}_U^*$ and on the right by $\mathbf{Q}_U$, and using $\mathbf{Q}_U^*\mathbf{Q}_U = \mathbf{I}_k$ (part (a) of the lemma), gives $\mathbf{I}_k = \mathbf{M}^*\mathbf{M} + \mathbf{N}^*\mathbf{N}$.
The matrix $\mathbf{M}^*\mathbf{M} \in M_k(F)$ is Hermitian (real symmetric when $F = \mathbb{R}$). By Corollary 12.4.3 (or Corollary 12.5.2 when $F = \mathbb{R}$) there is an orthonormal basis $\mathbf{z}_1, \ldots, \mathbf{z}_k$ of F^k consisting of eigenvectors of $\mathbf{M}^*\mathbf{M}$. Since $\mathbf{M}$ has $l \ge k$ rows, all k eigenvalues of $\mathbf{M}^*\mathbf{M}$ are squared singular values (Definition 13.9.1), so we may number the $\mathbf{z}_i$ so that $\mathbf{M}^*\mathbf{M}\mathbf{z}_i = \sigma_i(\mathbf{M})^2\mathbf{z}_i = \cos^2\theta_i\,\mathbf{z}_i$. Then

$$\mathbf{N}^*\mathbf{N}\mathbf{z}_i = (\mathbf{I}_k - \mathbf{M}^*\mathbf{M})\mathbf{z}_i = (1 - \cos^2\theta_i)\mathbf{z}_i = \sin^2\theta_i\,\mathbf{z}_i.$$

So the unitary matrix with columns $\mathbf{z}_1, \ldots, \mathbf{z}_k$ diagonalizes $\mathbf{N}^*\mathbf{N}$ with diagonal $\sin^2\theta_1, \ldots, \sin^2\theta_k$, and these are its eigenvalues with multiplicity (Theorem 7.10.3). The sine is increasing on $[0, \pi/2]$, so in decreasing order they read $\sin^2\theta_k \ge \cdots \ge \sin^2\theta_1 \ge 0$. By Eq. 17.9.1, the i-th largest eigenvalue of $\mathbf{N}^*\mathbf{N}$ is $\sigma_i(\mathbf{N})^2$ for every $i \le k$, with the stated convention. Taking nonnegative square roots gives $\sigma_i(\mathbf{N}) = \sin\theta_{k+1-i}$.

(b) If $l = n$, then $W = F^n$, so $U \subseteq W$ and every $\theta_i = 0$ (Example 2 above), while $\mathbf{I} - P_W = \mathbf{O}$; all four norms are 0. Suppose $l < n$. By Lemma 20.8.1 (d) and (a),

$$(\mathbf{I} - P_W)\mathbf{Q}_U = \mathbf{Q}_{W^\perp}\mathbf{Q}_{W^\perp}^*\mathbf{Q}_U = \mathbf{Q}_{W^\perp}\mathbf{N},$$

and $(\mathbf{I} - P_W)P_U = \mathbf{Q}_{W^\perp}\mathbf{N}\mathbf{Q}_U^*$. By part (b) of the lemma, applied to $\mathbf{Q}_{W^\perp}$ and then to $\mathbf{Q}_U$, both matrices have the spectral and Frobenius norms of $\mathbf{N}$. Now $\|\mathbf{N}\|_2 = \sigma_1(\mathbf{N}) =$

$\sin\theta_k$ by Theorem 16.3.3 (c) and part (a). Also $\|\mathbf{N}\|_F^2 = \operatorname{tr}(\mathbf{N}^*\mathbf{N})$, which is the sum of the eigenvalues of $\mathbf{N}^*\mathbf{N}$ (Theorem 9.3.3), that is $\sum_i \sin^2\theta_i = \|\sin\Theta(U,W)\|_F^2$. This proves the theorem.

■

Part (b) has a plain geometric reading. For a unit vector $\mathbf{u} \in U$, the vector $(\mathbf{I} - P_W)\mathbf{u}$ is the error in approximating $\mathbf{u}$ from W, and its length is the distance from $\mathbf{u}$ to W (Theorem 11.3.5). So $\sin\theta_k$ is the **largest distance from a unit vector of U to W**, and θ_k is the worst angle any direction of U makes with W. That is the number to bound when one asks whether a computed subspace is close to the true one, and part (a) turns it into the norm of the concrete matrix $\mathbf{Q}_{W^\perp}^*\mathbf{Q}_U$. In Section 9, U will be an invariant subspace of $\mathbf{A}$ and W the corresponding one for $\mathbf{A}+\mathbf{E}$. The columns of $\mathbf{Q}_{W^\perp}$ will be the remaining eigenvectors of $\mathbf{A}+\mathbf{E}$, and $\mathbf{N}$ will turn out to satisfy a Sylvester equation, so the bounds of Section 7 will apply to it.

20.8.4 The distance between two subspaces

A subspace is determined by its orthogonal projection, since $U = \operatorname{im} P_U$ by Theorem 4.5.1 (b), applied to $F^n = U \oplus U^\perp$. So a second natural way to compare U and W is the size of $P_U - P_W$. When the dimensions agree, this measures exactly the largest principal angle.

Theorem 20.8.5. The Distance Between Projections

Let $U, W \subseteq F^n$ with $\dim U = \dim W = k \geq 1$, and principal angles $\theta_1 \leq \cdots \leq \theta_k$. Then

(a) $\|P_U - P_W\|_2 = \sin\theta_k$;
(b) $\|P_U - P_W\|_F = \sqrt{2}\,\|\sin\Theta(U,W)\|_F$;
(c) $\|P_U - P_W\|_2 \leq 1$, with equality if and only if $U \cap W^\perp \neq \{\mathbf{0}\}$.

Idea. For (b), the **squared** Frobenius norm of the Hermitian matrix $\mathbf{D} = P_U - P_W$ is $\operatorname{tr}(\mathbf{D}^2)$, which expands into traces of projections: $k + k - 2\operatorname{tr}(P_U P_W)$, and $\operatorname{tr}(P_U P_W) = \sum \cos^2\theta_i$. For (a), split $\mathbf{x} = \mathbf{x}_1 + \mathbf{x}_2$ along $U \oplus U^\perp$. Then $\mathbf{D}$ sends $\mathbf{x}_1$ to $(\mathbf{I} - P_W)\mathbf{x}_1 \in W^\perp$ and $\mathbf{x}_2$ to $-P_W\mathbf{x}_2 \in W$, so Pythagoras separates the two contributions. The first is controlled by the sines for (U,W), the second by the sines for (W,U), and equal dimensions make these the same list.

Proof. Put $\mathbf{D} = P_U - P_W$, and let $\mathbf{Q}_U, \mathbf{Q}_W$ be orthonormal basis matrices and $\mathbf{M} = \mathbf{Q}_W^*\mathbf{Q}_U \in M_k(F)$.

(b) By Lemma 20.8.1 (a), $P_U = \mathbf{Q}_U\mathbf{Q}_U^*$ and $P_W = \mathbf{Q}_W\mathbf{Q}_W^*$ are Hermitian and satisfy $P_U^2 = P_U$, $P_W^2 = P_W$. Hence $\mathbf{D}$ is Hermitian and

$$\begin{aligned}\|\mathbf{D}\|_F^2 &= \operatorname{tr}(\mathbf{D}^*\mathbf{D}) = \operatorname{tr}(\mathbf{D}^2)\\ &= \operatorname{tr}(P_U) + \operatorname{tr}(P_W) - \operatorname{tr}(P_U P_W) - \operatorname{tr}(P_W P_U).\end{aligned}$$

By Theorem 0.9.6 (3), $\operatorname{tr}(P_U) = \operatorname{tr}(\mathbf{Q}_U^*\mathbf{Q}_U) = \operatorname{tr}(\mathbf{I}_k) = k$, likewise $\operatorname{tr}(P_W) = k$, and

$\operatorname{tr}(P_W P_U) = \operatorname{tr}(P_U P_W)$. Also

$$\begin{aligned}\operatorname{tr}(P_U P_W) &= \operatorname{tr}(\mathbf{Q}_U \mathbf{Q}_U^* \mathbf{Q}_W \mathbf{Q}_W^*) \\ &= \operatorname{tr}(\mathbf{Q}_W^* \mathbf{Q}_U \mathbf{Q}_U^* \mathbf{Q}_W) = \operatorname{tr}(\mathbf{M}\mathbf{M}^*),\end{aligned}$$

and $\operatorname{tr}(\mathbf{M}\mathbf{M}^*) = \|\mathbf{M}^*\|_F^2 = \|\mathbf{M}\|_F^2 = \operatorname{tr}(\mathbf{M}^*\mathbf{M}) = \sum_i \cos^2\theta_i$, since the eigenvalues of $\mathbf{M}^*\mathbf{M}$ are $\cos^2\theta_1, \ldots, \cos^2\theta_k$ (Definition 13.9.1, Theorem 9.3.3). Therefore

$$\|\mathbf{D}\|_F^2 = 2k - 2\sum_{i=1}^{k} \cos^2\theta_i = 2\sum_{i=1}^{k} \sin^2\theta_i,$$

which is (b).

(c) Put $s = \sin\theta_k$. By Theorem 20.8.4 (b), $\|(\mathbf{I} - P_W)P_U\|_2 = s$. The same theorem with the roles of U and W exchanged, which is allowed because $\dim W = \dim U$, gives $\|(\mathbf{I} - P_U)P_W\|_2 = s$ too, because by Proposition 20.8.2 (c) the principal angles between W and U are those between U and W.

Let $\mathbf{x} \in F^n$, and write $\mathbf{x}_1 = P_U\mathbf{x} \in U$ and $\mathbf{x}_2 = \mathbf{x} - \mathbf{x}_1 \in U^\perp$ (Definition 11.3.2). Then $P_U\mathbf{x} = \mathbf{x}_1$ and $P_W\mathbf{x} = P_W\mathbf{x}_1 + P_W\mathbf{x}_2$, so

$$\mathbf{D}\mathbf{x} = (\mathbf{I} - P_W)\mathbf{x}_1 - P_W\mathbf{x}_2.$$

The first term lies in $W^\perp$ and the second in W, by Theorem 11.3.2, so they are orthogonal, and by Theorem 11.1.3

$$\|\mathbf{D}\mathbf{x}\|^2 = \|(\mathbf{I} - P_W)\mathbf{x}_1\|^2 + \|P_W\mathbf{x}_2\|^2. \tag{20.8.1}$$

We bound the two terms. Since $\mathbf{x}_1 = P_U\mathbf{x}_1$, $\|(\mathbf{I} - P_W)\mathbf{x}_1\| = \|(\mathbf{I} - P_W)P_U\mathbf{x}_1\| \le s\|\mathbf{x}_1\|$ by Theorem 16.3.2 (a). For the second term, P_W is Hermitian and idempotent and $\mathbf{x}_2 = (\mathbf{I} - P_U)\mathbf{x}_2$, so

$$\begin{aligned}\|P_W\mathbf{x}_2\|^2 &= \langle P_W\mathbf{x}_2, \mathbf{x}_2\rangle = \langle P_W\mathbf{x}_2, (\mathbf{I} - P_U)\mathbf{x}_2\rangle \\ &= \langle (\mathbf{I} - P_U)P_W(P_W\mathbf{x}_2), \mathbf{x}_2\rangle \le s\,\|P_W\mathbf{x}_2\|\,\|\mathbf{x}_2\|,\end{aligned}$$

using Theorem 11.3.3 (b) for $\mathbf{I} - P_U = P_{U^\perp}$ (Lemma 20.8.1 (d)), then $P_W = P_W^2$, then Cauchy–Schwarz (Theorem 11.1.5) with $\|(\mathbf{I} - P_U)P_W\|_2 = s$. Hence $\|P_W\mathbf{x}_2\| \le s\|\mathbf{x}_2\|$, the case $P_W\mathbf{x}_2 = \mathbf{0}$ being trivial. Adding, and using $\|\mathbf{x}_1\|^2 + \|\mathbf{x}_2\|^2 = \|\mathbf{x}\|^2$ (Theorem 11.1.3),

$$\|\mathbf{D}\mathbf{x}\|^2 \le s^2(\|\mathbf{x}_1\|^2 + \|\mathbf{x}_2\|^2) = s^2\|\mathbf{x}\|^2.$$

So $\|\mathbf{D}\|_2 \le s$. For the reverse, Theorem 20.8.4 (b) gives $\|(\mathbf{I} - P_W)\mathbf{Q}_U\|_2 = s$, and by Lemma 16.3.1 there is a unit $\mathbf{y} \in F^k$ with $\|(\mathbf{I} - P_W)\mathbf{Q}_U\mathbf{y}\| = s$. The vector $\mathbf{x} = \mathbf{Q}_U\mathbf{y} \in U$ is a unit vector (Lemma 20.8.1 (b)) with $P_U\mathbf{x} = \mathbf{x}$, so $\mathbf{D}\mathbf{x} = (\mathbf{I} - P_W)\mathbf{x}$ and $\|\mathbf{D}\mathbf{x}\| = s$. Hence $\|\mathbf{D}\|_2 = s$.

(c) By (a) and Proposition 20.8.2 (b), $\|\mathbf{D}\|_2 = \sin\theta_k \le 1$, with equality if and only if $\cos\theta_k = \sigma_k(\mathbf{M}) = 0$. The square matrix $\mathbf{M}$ has $\sigma_k(\mathbf{M}) = 0$ if and only if $\operatorname{rank}\mathbf{M} < k$ (Theorem 13.9.1), that is, if and only if $\mathbf{M}\mathbf{y} = \mathbf{0}$ for some $\mathbf{y} \ne \mathbf{0}$. Put $\mathbf{u} = \mathbf{Q}_U\mathbf{y}$, which is a nonzero vector of U exactly when $\mathbf{y} \ne \mathbf{0}$, by Lemma 20.8.1 (b). Then $\mathbf{M}\mathbf{y} = \mathbf{Q}_W^*\mathbf{u}$, and $\mathbf{Q}_W^*\mathbf{u} = \mathbf{0}$ if and only if $P_W\mathbf{u} = \mathbf{Q}_W\mathbf{Q}_W^*\mathbf{u} = \mathbf{0}$, again by parts (b) and (a) of the lemma. Finally $P_W\mathbf{u} = \mathbf{0}$ means that $\mathbf{u} \in \ker P_W = W^\perp$, by Theorem 4.5.1 (b) applied to $F^n = W \oplus W^\perp$. So equality holds if and only if U contains a nonzero vector of $W^\perp$.

■

The factor $\sqrt{2}$ in (b) has a clean explanation. When $k = l$, the principal vectors of Proposition 20.8.3 give $P_U = \sum_i \mathbf{u}_i\mathbf{u}_i^*$ and $P_W = \sum_i \mathbf{w}_i\mathbf{w}_i^*$ (Lemma 20.8.1 (a)), so $P_U - P_W$ is a sum of the k terms $\mathbf{u}_i\mathbf{u}_i^* - \mathbf{w}_i\mathbf{w}_i^*$. These act in the mutually orthogonal planes $\operatorname{span}(\mathbf{u}_i, \mathbf{w}_i)$. Each term is Hermitian with trace 0 and rank at most 2, so its nonzero eigenvalues are a pair $\pm t_i$, and being Hermitian it is normal, so $\|\mathbf{u}_i\mathbf{u}_i^* - \mathbf{w}_i\mathbf{w}_i^*\|_F^2 = t_i^2 + t_i^2$ by Theorem 12.11.1. On the other hand, expanding the trace with $\mathbf{u}_i^*\mathbf{u}_i = \mathbf{w}_i^*\mathbf{w}_i = 1$ and $|\langle \mathbf{u}_i, \mathbf{w}_i\rangle| = \cos\theta_i$ (Proposition 20.8.3),

$$\|\mathbf{u}_i\mathbf{u}_i^* - \mathbf{w}_i\mathbf{w}_i^*\|_F^2 = 2 - 2|\langle \mathbf{u}_i, \mathbf{w}_i\rangle|^2 = 2\sin^2\theta_i,$$

so $t_i = \sin\theta_i$. Each angle is therefore counted twice in the Frobenius norm and once in the spectral norm. The worked example below shows the $\pm$ pairs explicitly.

Example. Two Planes in Four Dimensions

Let $U = \operatorname{span}(\mathbf{e}_1, \mathbf{e}_2) \subseteq \mathbb{R}^4$ and $W = \operatorname{span}(\mathbf{a}, \mathbf{b})$, where

$$\mathbf{a} = (3, 4, 4, 3), \qquad \mathbf{b} = (3, -4, 4, -3).$$

Find the principal angles, a pair of principal vectors for each, the sines from Theorem 20.8.4, and $\|P_U - P_W\|$ in both norms. Then check the last two numbers directly.

Solution. **Orthonormal bases.** $\langle \mathbf{a}, \mathbf{b}\rangle = 9 - 16 + 16 - 9 = 0$ and $\|\mathbf{a}\|^2 = \|\mathbf{b}\|^2 = 50$, so $\mathbf{Q}_W = \frac{1}{5\sqrt{2}}(\mathbf{a} \mid \mathbf{b})$ is an orthonormal basis matrix of W, and $\mathbf{Q}_U = (\mathbf{e}_1 \mid \mathbf{e}_2)$. The matrix $\mathbf{Q}_W^\top\mathbf{Q}_U$ holds the first two entries of $\mathbf{a}$ and $\mathbf{b}$:

$$\mathbf{M} = \mathbf{Q}_W^\top\mathbf{Q}_U = \frac{1}{5\sqrt{2}}\begin{pmatrix} 3 & 4 \\ 3 & -4 \end{pmatrix}, \qquad \mathbf{M}^\top\mathbf{M} = \begin{pmatrix} 9/25 & 0 \\ 0 & 16/25 \end{pmatrix}.$$

So the singular values are $4/5$ and $3/5$, and

$$\theta_1 = \arccos\tfrac{4}{5} \approx 36.87^\circ, \qquad \theta_2 = \arccos\tfrac{3}{5} \approx 53.13^\circ.$$

Note $\sin\theta_1 = 3/5 = \cos\theta_2$, so $\theta_1 + \theta_2 = \pi/2$ here.

Principal vectors. The eigenvector of $\mathbf{M}^\top\mathbf{M}$ for $16/25$ is $\mathbf{z}_1 = \mathbf{e}_2$, and $\mathbf{y}_1 = \mathbf{M}\mathbf{z}_1/\sigma_1 = \frac{1}{5\sqrt{2}}(4, -4) \cdot \frac{5}{4} = \frac{1}{\sqrt{2}}(1, -1)$. By Proposition 20.8.3,

$$\mathbf{u}_1 = \mathbf{e}_2, \qquad \mathbf{w}_1 = \mathbf{Q}_W\mathbf{y}_1 = \tfrac{1}{10}(\mathbf{a} - \mathbf{b}) = (0, \tfrac{4}{5}, 0, \tfrac{3}{5}),$$

and $\langle \mathbf{u}_1, \mathbf{w}_1\rangle = 4/5$. Likewise $\mathbf{z}_2 = \mathbf{e}_1$, $\mathbf{y}_2 = \frac{1}{\sqrt{2}}(1, 1)$, $\mathbf{u}_2 = \mathbf{e}_1$, and $\mathbf{w}_2 = \frac{1}{10}(\mathbf{a} + \mathbf{b}) = (\frac{3}{5}, 0, \frac{4}{5}, 0)$, with $\langle \mathbf{u}_2, \mathbf{w}_2\rangle = 3/5$, while $\langle \mathbf{u}_1, \mathbf{w}_2\rangle = \langle \mathbf{u}_2, \mathbf{w}_1\rangle = 0$. So W is obtained from U by tilting $\mathbf{e}_2$ towards $\mathbf{e}_4$ through θ_1 and $\mathbf{e}_1$ towards $\mathbf{e}_3$ through θ_2, in two perpendicular planes.

Sines. The vectors $\mathbf{c} = (4, -3, -3, 4)$ and $\mathbf{d} = (4, 3, -3, -4)$ are orthogonal to $\mathbf{a}$, to $\mathbf{b}$, and to each other:

$$\langle \mathbf{c}, \mathbf{a}\rangle = 12 - 12 - 12 + 12 = 0, \quad \langle \mathbf{c}, \mathbf{b}\rangle = 12 + 12 - 12 - 12 = 0,$$
$$\langle \mathbf{d}, \mathbf{a}\rangle = 12 + 12 - 12 - 12 = 0, \quad \langle \mathbf{d}, \mathbf{b}\rangle = 12 - 12 - 12 + 12 = 0,$$

and $\langle \mathbf{c}, \mathbf{d}\rangle = 16 - 9 + 9 - 16 = 0$. Also $\|\mathbf{c}\|^2 = \|\mathbf{d}\|^2 = 50$. Since $\dim W^\perp = 2$ (Theorem 11.3.2

(c)), $\mathbf{Q}_{W^\perp} = \frac{1}{5\sqrt{2}}(\mathbf{c} \mid \mathbf{d})$ is an orthonormal basis matrix of $W^\perp$, and

$$\mathbf{N} = \mathbf{Q}_{W^\perp}^\top \mathbf{Q}_U = \frac{1}{5\sqrt{2}} \begin{pmatrix} 4 & -3 \\ 4 & 3 \end{pmatrix}, \qquad \mathbf{N}^\top \mathbf{N} = \begin{pmatrix} 16/25 & 0 \\ 0 & 9/25 \end{pmatrix}.$$

Its eigenvalues are $16/25 = \sin^2\theta_2$ and $9/25 = \sin^2\theta_1$, and $\mathbf{M}^\top\mathbf{M} + \mathbf{N}^\top\mathbf{N} = \mathbf{I}_2$, as Theorem 20.8.4 (a) says.

Distances. By Theorem 20.8.5, $\|P_U - P_W\|_2 = \sin\theta_2 = 4/5$ and $\|P_U - P_W\|_F = \sqrt{2}(\frac{9}{25} + \frac{16}{25})^{1/2} = \sqrt{2}$.

Direct check. By Lemma 20.8.1 (a), $P_W = \mathbf{Q}_W\mathbf{Q}_W^\top = \frac{1}{50}(\mathbf{a}\mathbf{a}^\top + \mathbf{b}\mathbf{b}^\top)$, and

$$P_W = \frac{1}{50}\begin{pmatrix} 18 & 0 & 24 & 0 \\ 0 & 32 & 0 & 24 \\ 24 & 0 & 32 & 0 \\ 0 & 24 & 0 & 18 \end{pmatrix}, \qquad P_U = \operatorname{diag}(1,1,0,0).$$

In $\mathbf{D} = P_U - P_W$, coordinates $1,3$ do not interact with coordinates $2,4$, and the two 2×2 blocks are

$$\frac{1}{50}\begin{pmatrix} 32 & -24 \\ -24 & -32 \end{pmatrix}, \qquad \frac{1}{50}\begin{pmatrix} 18 & -24 \\ -24 & -18 \end{pmatrix}.$$

Each is symmetric with trace 0, so its eigenvalues are $\pm\sqrt{-\det}$: $\pm\frac{1}{50}\sqrt{32^2+24^2} = \pm\frac{4}{5}$ and $\pm\frac{1}{50}\sqrt{18^2+24^2} = \pm\frac{3}{5}$. So $\mathbf{D}$ has eigenvalues $\pm\frac{4}{5}, \pm\frac{3}{5}$. By Lemma 17.3.5, $\|\mathbf{D}\|_2 = 4/5$, and $\|\mathbf{D}\|_F^2 = \operatorname{tr}(\mathbf{D}^2) = 2(\frac{16}{25} + \frac{9}{25}) = 2$. Both agree with the theorem, and the eigenvalues come in the $\pm\sin\theta_i$ pairs described before the example.

Remark. **A metric, and why the dimensions must match.** On the set of k-dimensional subspaces of F^n, the number $d(U,W) = \|P_U - P_W\|_2 = \sin\theta_k$ is a metric. It is symmetric. It is zero only when $P_U = P_W$, that is, when $U = \operatorname{im} P_U = \operatorname{im} P_W = W$. And it satisfies the triangle inequality because $\|\cdot\|_2$ is a norm (Theorem 16.3.2 (b)). Between subspaces of **different** dimensions it is useless. If $k < l$, then $\dim(W \cap U^\perp) \geq l + (n-k) - n > 0$ by Theorem 1.7.3, so some unit $\mathbf{w} \in W$ is perpendicular to U. For it $(P_U - P_W)\mathbf{w} = -\mathbf{w}$, so $\|P_U - P_W\|_2 \geq 1$. On the other hand, Eq. 20.8.1, which did not use the dimensions, together with $\|P\mathbf{v}\| \leq \|\mathbf{v}\|$ for orthogonal projections (Theorem 11.3.3 (c)), gives $\|P_U - P_W\|_2 \leq 1$. So the distance is exactly 1, whatever the principal angles are. This is why Section 9 always compares invariant subspaces of the same dimension.

20.8.5 Exercises

20.8.5.1 A. Check your understanding

Exercise. A1

(a) Define the principal angles between subspaces $U, W \subseteq F^n$ with $1 \leq \dim U \leq \dim W$.

(b) True or false: if $U \subseteq W$, then every principal angle between U and W is 0. Justify your answer.

(c) True or false: the principal angle between $\operatorname{span}(\mathbf{u})$ and $\operatorname{span}(-\mathbf{u})$, for $\mathbf{u} \neq \mathbf{0}$, is π. Justify your answer.

(d) True or false: if $U \cap W \neq \{\mathbf{0}\}$, then $\theta_1 = 0$. Justify your answer.

(e) State the formula for $\|P_U - P_W\|_2$ in terms of principal angles, with its hypothesis.

(f) True or false: for two distinct lines U, W in $\mathbb{R}^2$, $\|P_U - P_W\|_2 < 1$. Justify your answer.

Solution.

(a) Choose orthonormal basis matrices $\mathbf{Q}_U \in M_{n\times k}(F)$ and $\mathbf{Q}_W \in M_{n\times l}(F)$. The principal angles $0 \le \theta_1 \le \cdots \le \theta_k \le \pi/2$ are defined by $\cos\theta_i = \sigma_i(\mathbf{Q}_W^*\mathbf{Q}_U)$, $i = 1, \dots, k$ (Definition 20.8.1). They do not depend on the bases (Proposition 20.8.2).

(b) True. As in Example 2 after Proposition 20.8.2, $\mathbf{M}^*\mathbf{M} = \mathbf{Q}_U^* P_W \mathbf{Q}_U = \mathbf{Q}_U^*\mathbf{Q}_U = \mathbf{I}_k$, so every $\cos\theta_i = 1$.

(c) False. The two lines are equal, so the angle is 0: with $\mathbf{u}_0 = \mathbf{u}/\|\mathbf{u}\|$, $|\langle \mathbf{u}_0, -\mathbf{u}_0\rangle| = 1$ (Proposition 20.8.2 (d)). The value π is the angle between the vectors $\mathbf{u}$ and $-\mathbf{u}$ of Definition 11.1.5, which is a different notion.

(d) True. Take a unit $\mathbf{u} \in U \cap W$ and write $\mathbf{u} = \mathbf{Q}_U\mathbf{y}$ with $\|\mathbf{y}\| = 1$ (Lemma 20.8.1 (b)). Then $\|\mathbf{M}\mathbf{y}\| = \|\mathbf{Q}_W^*\mathbf{u}\| = \|P_W\mathbf{u}\| = \|\mathbf{u}\| = 1$, so $\sigma_1(\mathbf{M}) = \|\mathbf{M}\|_2 \ge 1$. With Proposition 20.8.2 (b), $\cos\theta_1 = 1$.

(e) If $\dim U = \dim W = k \ge 1$, then $\|P_U - P_W\|_2 = \sin\theta_k$, the sine of the largest principal angle (Theorem 20.8.5 (a)).

(f) False. The two axes $U = \operatorname{span}(\mathbf{e}_1)$ and $W = \operatorname{span}(\mathbf{e}_2)$ have $U \cap W^\perp = U \neq \{\mathbf{0}\}$, so $\|P_U - P_W\|_2 = 1$ by Theorem 20.8.5 (c). Directly, $P_U - P_W = \operatorname{diag}(1, -1)$.

20.8.5.2 B. Practice

Exercise. B1: Two planes in three dimensions

Let $U = \operatorname{span}(\mathbf{e}_1, \mathbf{e}_2)$ and $W = \operatorname{span}((1,0,0),(1,3,4))$ in $\mathbb{R}^3$.

(a) Find an orthonormal basis matrix of W.

(b) Compute the principal angles between U and W.

(c) Hence compute $\|P_U - P_W\|_2$ and $\|P_U - P_W\|_F$.

(d) Explain why $\theta_1 = 0$ could have been predicted without computation.

Solution.

(a) Gram–Schmidt: $\mathbf{q}_1 = \mathbf{e}_1$, and $(1,3,4) - \langle (1,3,4), \mathbf{e}_1\rangle \mathbf{e}_1 = (0,3,4)$, of length 5, so $\mathbf{q}_2 = (0, \tfrac{3}{5}, \tfrac{4}{5})$ and $\mathbf{Q}_W = (\mathbf{q}_1 \mid \mathbf{q}_2)$.

(b) With $\mathbf{Q}_U = (\mathbf{e}_1 \mid \mathbf{e}_2)$, the matrix $\mathbf{M} = \mathbf{Q}_W^\top \mathbf{Q}_U$ holds the first two entries of $\mathbf{q}_1$ and $\mathbf{q}_2$:

$$\mathbf{M} = \begin{pmatrix} 1 & 0 \\ 0 & 3/5 \end{pmatrix}.$$

It is diagonal with nonnegative entries, so $\mathbf{M}^\top\mathbf{M} = \operatorname{diag}(1, \tfrac{9}{25})$ and the singular values are 1 and $\tfrac{3}{5}$. So $\theta_1 = 0$ and $\theta_2 = \arccos\tfrac{3}{5} \approx 53.13°$, with $\sin\theta_2 = \tfrac{4}{5}$.

(c) By Theorem 20.8.5, $\|P_U - P_W\|_2 = \sin\theta_2 = \frac{4}{5}$ and $\|P_U - P_W\|_F = \sqrt{2}\,(0+\frac{16}{25})^{1/2} = \frac{4\sqrt{2}}{5}$.

(d) Two planes in $\mathbb{R}^3$ meet in a line (Theorem 1.7.3, as in the opening of the section), and a common unit vector forces $\cos\theta_1 = 1$ (Exercise A1 (d)). Here the common line is $\operatorname{span}(\mathbf{e}_1)$.

Exercise. B2: Which projections are at distance one?

Determine, for each pair, whether $\|P_U - P_W\|_2 = 1$. Justify your answer.

(a) $U = \operatorname{span}(\mathbf{e}_1, \mathbf{e}_2)$ and $W = \operatorname{span}(\mathbf{e}_2, \mathbf{e}_3)$ in $\mathbb{R}^3$.

(b) $U = \operatorname{span}((1,1,0))$ and $W = \operatorname{span}((1,-1,1))$ in $\mathbb{R}^3$.

(c) $U = \operatorname{span}(\mathbf{e}_1, \mathbf{e}_2)$ and $W = \operatorname{span}((1,0,1),(0,1,1))$ in $\mathbb{R}^3$.

(d) $U = \operatorname{span}((1,2,3))$ and $W = \operatorname{span}((1,0,0),(0,1,0))$ in $\mathbb{R}^3$.

Solution.

(a) Yes. $W^\perp = \operatorname{span}(\mathbf{e}_1) \subseteq U$, so $U \cap W^\perp \neq \{\mathbf{0}\}$, and Theorem 20.8.5 (c) applies since both dimensions are 2.

(b) Yes. $\langle (1,1,0), (1,-1,1)\rangle = 0$, so $U \subseteq W^\perp$, and (c) of the theorem applies with $k = 1$.

(c) No. Both are planes, and $W^\perp = \operatorname{span}((1,1,-1))$, since $(1,1,-1)$ is orthogonal to both spanning vectors of W and $\dim W^\perp = 1$. This vector has nonzero third entry, so it is not in U, and $U \cap W^\perp = \{\mathbf{0}\}$. By the theorem, $\|P_U - P_W\|_2 < 1$. (In fact the principal angles are 0 and $\arccos(1/\sqrt{3})$, so the distance is $\sqrt{2/3}$.)

(d) Yes, for a different reason: the dimensions are 1 and 2, and the remark after Example (Two Planes in Four Dimensions) shows that $\|P_U - P_W\|_2 = 1$ whenever the dimensions differ. The theorem itself does not apply. Its principal angle $\theta_1 = \arccos(\sqrt{5}/\sqrt{14})$ is far from $\pi/2$, which illustrates why the projection distance is only used for equal dimensions.

Exercise. B3: Two lines

Let $\mathbf{u}, \mathbf{w} \in F^n$ be unit vectors and $c = |\langle \mathbf{u}, \mathbf{w}\rangle|$. Prove that

$$\|\mathbf{u}\mathbf{u}^* - \mathbf{w}\mathbf{w}^*\|_2 = \sqrt{1-c^2}, \qquad \|\mathbf{u}\mathbf{u}^* - \mathbf{w}\mathbf{w}^*\|_F = \sqrt{2}\,\sqrt{1-c^2}.$$

Solution. Let $U = \operatorname{span}(\mathbf{u})$ and $W = \operatorname{span}(\mathbf{w})$, both of dimension 1. By Lemma 20.8.1 (a) with $\mathbf{Q}_U = \mathbf{u}$ and $\mathbf{Q}_W = \mathbf{w}$, $P_U = \mathbf{u}\mathbf{u}^*$ and $P_W = \mathbf{w}\mathbf{w}^*$. By Proposition 20.8.2 (d), the single principal angle satisfies $\cos\theta_1 = c$, so $\sin\theta_1 = \sqrt{1-c^2}$, since $\sin\theta_1 \geq 0$ on $[0, \pi/2]$. Now Theorem 20.8.5 (a) and (b) with $k = 1$ give $\|P_U - P_W\|_2 = \sin\theta_1$ and $\|P_U - P_W\|_F = \sqrt{2}\sin\theta_1$, which are the two claims.

20.8.5.3 C. Going deeper

Exercise. C1: Counting zero and right angles

Let $U, W \subseteq F^n$ with $1 \le k = \dim U \le l = \dim W$, orthonormal basis matrices $\mathbf{Q}_U, \mathbf{Q}_W$, and $\mathbf{M} = \mathbf{Q}_W^* \mathbf{Q}_U$.

(a) Prove that $\mathbf{y}^*(\mathbf{I}_k - \mathbf{M}^*\mathbf{M})\mathbf{y} = \|(\mathbf{I} - P_W)\mathbf{Q}_U\mathbf{y}\|^2$ for every $\mathbf{y} \in F^k$.

(b) Deduce that the number of principal angles equal to 0 is $\dim(U \cap W)$.

(c) Prove that the number of principal angles equal to $\pi/2$ is $\dim(U \cap W^\perp)$.

Hint for (b): a positive semidefinite matrix $\mathbf{B}$ *with* $\mathbf{y}^*\mathbf{B}\mathbf{y} = 0$ *has* $\mathbf{B}\mathbf{y} = \mathbf{0}$.

Solution.

(a) Let $\mathbf{u} = \mathbf{Q}_U\mathbf{y}$. By Lemma 20.8.1 (a) and (b), $\mathbf{y}^*\mathbf{y} = \|\mathbf{u}\|^2$ and $\mathbf{y}^*\mathbf{M}^*\mathbf{M}\mathbf{y} = \|\mathbf{Q}_W^*\mathbf{u}\|^2 = \|P_W\mathbf{u}\|^2$. Since $\mathbf{u} = P_W\mathbf{u} + (\mathbf{I} - P_W)\mathbf{u}$ with the two parts orthogonal (Theorem 11.3.2), Theorem 11.1.3 gives $\|\mathbf{u}\|^2 - \|P_W\mathbf{u}\|^2 = \|(\mathbf{I} - P_W)\mathbf{u}\|^2$.

(b) Put $\mathbf{B} = \mathbf{I}_k - \mathbf{M}^*\mathbf{M}$. It is Hermitian, and by (a) it is positive semidefinite (Definition 13.1.1). Its eigenvalues are $1 - \cos^2\theta_i = \sin^2\theta_i$, with an orthonormal eigenbasis $\mathbf{z}_1, \dots, \mathbf{z}_k$ (as in the proof of Theorem 20.8.4). So the number of i with $\theta_i = 0$ is the number of zero eigenvalues, which is $\dim\ker\mathbf{B}$.
We claim $\ker\mathbf{B} = \{\mathbf{y} : \mathbf{Q}_U\mathbf{y} \in W\}$. If $\mathbf{B}\mathbf{y} = \mathbf{0}$, then $\mathbf{y}^*\mathbf{B}\mathbf{y} = 0$, so $(\mathbf{I} - P_W)\mathbf{Q}_U\mathbf{y} = \mathbf{0}$ by (a), and $\mathbf{Q}_U\mathbf{y} = P_W\mathbf{Q}_U\mathbf{y} \in W$. Conversely, if $\mathbf{Q}_U\mathbf{y} \in W$, then $\mathbf{y}^*\mathbf{B}\mathbf{y} = 0$ by (a). Writing $\mathbf{y} = \sum_i c_i\mathbf{z}_i$, this says $\sum_i \sin^2\theta_i |c_i|^2 = 0$. So $c_i = 0$ whenever $\sin\theta_i \neq 0$, and then $\mathbf{B}\mathbf{y} = \sum_i c_i \sin^2\theta_i \mathbf{z}_i = \mathbf{0}$. This proves the claim; the second half of it is the fact in the hint, proved here for this $\mathbf{B}$.
The map $\mathbf{y} \mapsto \mathbf{Q}_U\mathbf{y}$ is linear and injective (Lemma 20.8.1 (b)). It carries $\ker\mathbf{B}$ onto $\{\mathbf{u} \in U : \mathbf{u} \in W\} = U \cap W$, because every vector of U is $\mathbf{Q}_U\mathbf{y}$ for some $\mathbf{y}$. So $\dim\ker\mathbf{B} = \dim(U \cap W)$.

(c) $\theta_i = \pi/2$ exactly when $\sigma_i(\mathbf{M}) = 0$. The number of nonzero singular values of $\mathbf{M}$ is $\operatorname{rank}\mathbf{M}$ (Theorem 13.9.1), so the number of zero ones among the k is $k - \operatorname{rank}\mathbf{M} = \dim\ker\mathbf{M}$, by the Rank–Nullity Theorem (Theorem 2.3.1) for $\mathbf{y} \mapsto \mathbf{M}\mathbf{y}$ on F^k. Next, $\mathbf{M}\mathbf{y} = \mathbf{Q}_W^*\mathbf{Q}_U\mathbf{y} = \mathbf{0}$ if and only if $P_W\mathbf{Q}_U\mathbf{y} = \mathbf{Q}_W(\mathbf{Q}_W^*\mathbf{Q}_U\mathbf{y}) = \mathbf{0}$, because $\mathbf{Q}_W$ is injective. That holds if and only if $\mathbf{Q}_U\mathbf{y} \in W^\perp$. As in (b), $\mathbf{y} \mapsto \mathbf{Q}_U\mathbf{y}$ carries $\ker\mathbf{M}$ injectively onto $U \cap W^\perp$, so the count is $\dim(U \cap W^\perp)$.

Exercise. C2: The smallest angle as a maximum

Let $U, W \subseteq F^n$ with $1 \le k = \dim U \le l = \dim W$ and principal angles $\theta_1 \le \cdots \le \theta_k$.

(a) Prove that

$$\cos\theta_1 = \max\{|\langle \mathbf{u}, \mathbf{w}\rangle| : \mathbf{u} \in U,\ \mathbf{w} \in W,\ \|\mathbf{u}\| = \|\mathbf{w}\| = 1\},$$

and that the maximum is attained at the principal vectors $\mathbf{u}_1, \mathbf{w}_1$.

(b) Deduce that θ_1 is the smallest principal angle between a line of U and a line of W, which is the notion tried and rejected at the start of the section.

(c) Hence explain, in one sentence, why θ_1 alone cannot distinguish the pairs (U, W_1) and (U, U) of the opening, but the full list θ_1, θ_2 can.

Solution.

(a) Let $\mathbf{u} \in U$ and $\mathbf{w} \in W$ be unit vectors. By Lemma 20.8.1 (a) and (b), $\mathbf{u} = \mathbf{Q}_U\mathbf{y}$ and $\mathbf{w} = \mathbf{Q}_W\mathbf{z}$ with $\mathbf{y} = \mathbf{Q}_U^*\mathbf{u}$ and $\mathbf{z} = \mathbf{Q}_W^*\mathbf{w}$ unit vectors. Then $\langle \mathbf{u}, \mathbf{w}\rangle = \mathbf{z}^*\mathbf{Q}_W^*\mathbf{Q}_U\mathbf{y} = \mathbf{z}^*\mathbf{M}\mathbf{y}$, and by Cauchy–Schwarz and Theorem 16.3.2 (a),

$$|\langle \mathbf{u}, \mathbf{w}\rangle| \le \|\mathbf{z}\|\,\|\mathbf{M}\mathbf{y}\| \le \|\mathbf{M}\|_2 = \sigma_1(\mathbf{M}) = \cos\theta_1,$$

using Theorem 16.3.3 (c). By Proposition 20.8.3, the unit vectors $\mathbf{u}_1 \in U$ and $\mathbf{w}_1 \in W$ satisfy $\langle \mathbf{u}_1, \mathbf{w}_1\rangle = \cos\theta_1$, so the maximum exists and equals $\cos\theta_1$.

(b) By Proposition 20.8.2 (d), the principal angle between the lines $\operatorname{span}(\mathbf{u})$ and $\operatorname{span}(\mathbf{w})$ is $\arccos|\langle \mathbf{u}, \mathbf{w}\rangle|$. Since arccos is decreasing, the smallest such angle corresponds to the largest $|\langle \mathbf{u}, \mathbf{w}\rangle|$, which by (a) is $\cos\theta_1$.

(c) Both pairs share a line, so both have $\theta_1 = 0$, and the first number cannot separate them. The second angle is $\pi/2$ for (U, W_1) and 0 for (U, U), by Examples 4 and 2 after Proposition 20.8.2.

20.9 The Davis–Kahan Theorem

Weyl's inequality settled the eigenvalues of a Hermitian matrix: under a Hermitian perturbation $\mathbf{E}$, each of them moves by at most $\|\mathbf{E}\|_2$, whatever the matrix and however its eigenvalues are spaced. Eigenvectors are a different matter, and this section is about them. Section 8 supplied the yardstick, the principal angles between two subspaces, and Section 7 supplied the tool, bounds for the equation $\mathbf{AX} - \mathbf{XB} = \mathbf{Y}$. The theorem of this section is what happens when the two meet: the perturbation of an invariant subspace satisfies a Sylvester equation, and the bounds of Section 7 then measure it.

Throughout, $\mathbf{A}$ and $\widetilde{\mathbf{A}} = \mathbf{A} + \mathbf{E}$ are Hermitian matrices in $M_n(\mathbb{C})$, so $\mathbf{E} = \widetilde{\mathbf{A}} - \mathbf{A}$ is Hermitian too, and $1 \le k \le n-1$. As everywhere in this chapter, $\|\cdot\|_2$ is the spectral norm and $\|\cdot\|_F$ the Frobenius norm, and every bound names the norm it holds in. For a subspace we write $\operatorname{col}(\mathbf{Q})$ for the column space of a matrix $\mathbf{Q}$ whose columns span it.

20.9.1 Eigenvectors can move a long way

Before any theorem, here is what goes wrong. The example is as small as it can be, and every number in it is exact.

Example. A tiny perturbation that turns the eigenvectors through 45°

Let $0 < \eta < 1$ and

$$\mathbf{A} = \begin{pmatrix} 1+\eta & 0 \\ 0 & 1-\eta \end{pmatrix}, \qquad \mathbf{E} = \eta \begin{pmatrix} -1 & 1 \\ 1 & 1 \end{pmatrix}.$$

Compare the eigenvalues and the eigenvectors of $\mathbf{A}$ and of $\widetilde{\mathbf{A}} = \mathbf{A} + \mathbf{E}$, and compute $\|\mathbf{E}\|_2$.

Solution. Adding, $\widetilde{\mathbf{A}} = \left(\begin{smallmatrix} 1 & \eta \\ \eta & 1 \end{smallmatrix}\right)$. Since

$$\widetilde{\mathbf{A}} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = (1+\eta) \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \qquad \widetilde{\mathbf{A}} \begin{pmatrix} 1 \\ -1 \end{pmatrix} = (1-\eta) \begin{pmatrix} 1 \\ -1 \end{pmatrix},$$

the eigenvalues of $\widetilde{\mathbf{A}}$ are $1+\eta$ and $1-\eta$: **exactly** those of $\mathbf{A}$. The eigenvectors are another story. For the eigenvalue $1+\eta$, the matrix $\mathbf{A}$ has the unit eigenvector $\mathbf{e}_1$ and $\widetilde{\mathbf{A}}$ has $\frac{1}{\sqrt{2}}(1,1)$; the two lines they span meet at the angle $\pi/4$, because $|\langle \mathbf{e}_1, \frac{1}{\sqrt{2}}(1,1)\rangle| = \frac{1}{\sqrt{2}} = \cos\frac{\pi}{4}$. The same happens for $1-\eta$, with $\mathbf{e}_2$ and $\frac{1}{\sqrt{2}}(1,-1)$.

The matrix $\mathbf{E}/\eta$ is Hermitian and squares to $2\mathbf{I}_2$, so its eigenvalues are $\pm\sqrt{2}$ (it has trace 0), and by Lemma 17.3.5

$$\|\mathbf{E}\|_2 = \sqrt{2}\,\eta.$$

So as $\eta \to 0^+$ the perturbation vanishes, the eigenvalues do not move at all, and yet the eigenvectors of $\widetilde{\mathbf{A}}$ stay a full 45° away from those of $\mathbf{A}$.

The rotation does not shrink with η. What does shrink with η is the distance 2η between the two eigenvalues of $\mathbf{A}$, and that is the whole story: the two eigenvalues are so close that a perturbation of their own size can mix their eigenvectors freely. An eigenvector bound must therefore divide by a gap, and the natural guess is "angle $\lesssim$ perturbation / gap". Here that ratio is $\sqrt{2}\eta/2\eta = 1/\sqrt{2} = \sin(\pi/4)$, which is exactly the sine of the angle. The theorem below turns the guess into an inequality, and this example will turn out to attain it.

Warning. **Well-conditioned eigenvalues do not make well-conditioned eigenvectors.** Corollary 17.3.6 guarantees that every eigenvalue of a Hermitian matrix moves by at most $\|\mathbf{E}\|_2$, with no mention of gaps, and it is tempting to expect the same of eigenvectors. Example (A tiny perturbation that turns the eigenvectors through 45°) refutes this: a perturbation of size $\sqrt{2}\,\eta$ leaves every eigenvalue fixed and turns every eigenvector through $\pi/4$, however small η is. No inequality of the form $\sin\theta \le C\|\mathbf{E}\|_2$, with C depending only on n, can hold.

20.9.2 Two norms, three facts

The proofs of this section and the next shuffle unitary factors and matrices with orthonormal columns in and out of norms, sometimes the spectral one and sometimes the Frobenius one. The facts needed are elementary, but they are needed so often that they are worth collecting once. Part (c) is mostly quoted: its norm identities for a matrix with orthonormal columns are Lemma 20.8.1 (b) of Section 8, read with $U = \operatorname{col}(\mathbf{X})$, and for the Frobenius norm the unitary invariance is Lemma 11.7.5. The rest — (a), (b), and the inequality $\|\mathbf{X}^*\mathbf{N}\| \le \|\mathbf{N}\|$ — is proved below, and what the lemma is really for is to have all of it, in both norms, in one place.

Lemma 20.9.1. Moving Factors in and out of the Two Norms

Let $\|\cdot\|$ stand for either $\|\cdot\|_2$ or $\|\cdot\|_F$, the same one throughout each statement.

(a) $\|\mathbf{M}^*\| = \|\mathbf{M}\|$ for every $\mathbf{M} \in M_{m\times n}(\mathbb{C})$.

(b) $\|\mathbf{XMY}\| \le \|\mathbf{X}\|_2\,\|\mathbf{M}\|\,\|\mathbf{Y}\|_2$ whenever the product is defined.

(c) If $\mathbf{X} \in M_{m\times k}(\mathbb{C})$ has orthonormal columns, that is $\mathbf{X}^*\mathbf{X} = \mathbf{I}_k$, then $\|\mathbf{X}\|_2 = 1$, $\|\mathbf{XM}\| = \|\mathbf{M}\|$, and $\|\mathbf{X}^*\mathbf{N}\| \le \|\mathbf{N}\|$ for all $\mathbf{M} \in M_{k\times n}(\mathbb{C})$ and $\mathbf{N} \in M_{m\times n}(\mathbb{C})$. In particular $\|\mathbf{UMV}\| = \|\mathbf{M}\|$ for unitary $\mathbf{U}$ and $\mathbf{V}$.

Proof.

(a) For $\|\cdot\|_F$ this holds because $\mathbf{M}^*$ has the same entries as $\mathbf{M}$ up to position and conjugation. For $\|\cdot\|_2$, let $\mathbf{x} \in \mathbb{C}^n$. By Theorem 11.1.5 and Theorem 16.3.2 (a), applied twice,

$$\|\mathbf{Mx}\|^2 = \langle \mathbf{M}^*\mathbf{Mx}, \mathbf{x}\rangle \le \|\mathbf{M}^*\mathbf{Mx}\|\,\|\mathbf{x}\| \le \|\mathbf{M}^*\|_2\|\mathbf{M}\|_2\|\mathbf{x}\|^2.$$

Taking the maximum over unit $\mathbf{x}$ gives $\|\mathbf{M}\|_2^2 \le \|\mathbf{M}^*\|_2\|\mathbf{M}\|_2$, hence $\|\mathbf{M}\|_2 \le \|\mathbf{M}^*\|_2$ when $\mathbf{M} \neq \mathbf{O}$ (and trivially when $\mathbf{M} = \mathbf{O}$). Applying this to $\mathbf{M}^*$, whose conjugate transpose is $\mathbf{M}$, gives the reverse inequality.

(b) For $\|\cdot\|_2$ this is Theorem 16.3.2 (d) used twice. For $\|\cdot\|_F$, let $\mathbf{m}_1, \dots, \mathbf{m}_n$ be the columns of $\mathbf{M}$. The columns of $\mathbf{XM}$ are the $\mathbf{Xm}_j$, and the squared Frobenius norm is the sum of the squared lengths of the columns, so

$$\|\mathbf{XM}\|_F^2 = \sum_j \|\mathbf{Xm}_j\|^2 \le \|\mathbf{X}\|_2^2 \sum_j \|\mathbf{m}_j\|^2 = \|\mathbf{X}\|_2^2\|\mathbf{M}\|_F^2.$$

For a factor on the right, (a) and the left-hand case give $\|\mathbf{MY}\|_F = \|\mathbf{Y}^*\mathbf{M}^*\|_F \le \|\mathbf{Y}^*\|_2\|\mathbf{M}^*\|_F = \|\mathbf{Y}\|_2\|\mathbf{M}\|_F$. Combining the two cases gives (b).

(c) For every $\mathbf{v} \in \mathbb{C}^k$, $\|\mathbf{Xv}\|^2 = \mathbf{v}^*\mathbf{X}^*\mathbf{Xv} = \|\mathbf{v}\|^2$. Hence $\|\mathbf{X}\|_2 = 1$, and multiplying by $\mathbf{X}$ changes the length of no column, so $\|\mathbf{XM}\|_F = \|\mathbf{M}\|_F$ and $\|\mathbf{XMv}\| = \|\mathbf{Mv}\|$ for every $\mathbf{v}$, which gives $\|\mathbf{XM}\|_2 = \|\mathbf{M}\|_2$. Next, $\|\mathbf{X}^*\|_2 = \|\mathbf{X}\|_2 = 1$ by (a), so $\|\mathbf{X}^*\mathbf{N}\| \le \|\mathbf{N}\|$ by (b). Finally, for unitary $\mathbf{U}$ and $\mathbf{V}$ both $\mathbf{U}$ and $\mathbf{V}^*$ have orthonormal columns, so $\|\mathbf{UMV}\| = \|\mathbf{MV}\| = \|\mathbf{V}^*\mathbf{M}^*\| = \|\mathbf{M}^*\| = \|\mathbf{M}\|$, using (a) twice.

■

Part (c) is the one used most. Its inequality $\|\mathbf{X}^*\mathbf{N}\| \le \|\mathbf{N}\|$ says that reading $\mathbf{N}$ in only some of the coordinates of an orthonormal basis can only lose mass, never gain it.

20.9.3 The perturbation of a subspace is a Sylvester equation

Fix the data. A Hermitian matrix is normal, so by Corollary 12.4.3 there is a unitary $\mathbf{Q} \in M_n(\mathbb{C})$ with $\mathbf{Q}^*\mathbf{AQ}$ diagonal, the eigenvalues appearing in any order we like, and they are real by Theorem 12.2.1. Choose k of them, list them first, and split the columns accordingly:

$$\mathbf{Q} = \begin{pmatrix}\mathbf{Q}_1 & \mathbf{Q}_2\end{pmatrix}, \qquad \mathbf{Q}^*\mathbf{AQ} = \begin{pmatrix}\mathbf{\Lambda}_1 & \mathbf{O} \\ \mathbf{O} & \mathbf{\Lambda}_2\end{pmatrix},$$

with $\mathbf{Q}_1 \in M_{n\times k}(\mathbb{C})$ and $\mathbf{\Lambda}_1 \in M_k(\mathbb{R})$, $\mathbf{\Lambda}_2 \in M_{n-k}(\mathbb{R})$ diagonal. Do the same for $\widetilde{\mathbf{A}}$, producing $\widetilde{\mathbf{Q}} = \begin{pmatrix}\widetilde{\mathbf{Q}}_1 & \widetilde{\mathbf{Q}}_2\end{pmatrix}$, $\widetilde{\mathbf{\Lambda}}_1$, and $\widetilde{\mathbf{\Lambda}}_2$. Reading $\mathbf{AQ} = \mathbf{Q}(\mathbf{\Lambda}_1 \oplus \mathbf{\Lambda}_2)$ column by column,

$$\mathbf{AQ}_1 = \mathbf{Q}_1\mathbf{\Lambda}_1, \qquad \widetilde{\mathbf{A}}\widetilde{\mathbf{Q}}_2 = \widetilde{\mathbf{Q}}_2\widetilde{\mathbf{\Lambda}}_2,$$

so the columns of $\mathbf{Q}_1$ are orthonormal eigenvectors of $\mathbf{A}$, and $U = \operatorname{col}(\mathbf{Q}_1)$ is a k-dimensional subspace invariant under $\mathbf{A}$. Likewise $W = \operatorname{col}(\widetilde{\mathbf{Q}}_1)$ is a k-dimensional subspace invariant under

$\widetilde{\mathbf{A}}$. The question is how far apart U and W are, and Section 8 answers "measure the principal angles $\theta_1 \le \cdots \le \theta_k$ of Definition 20.8.1".

Put

$$\mathbf{X} \coloneqq \widetilde{\mathbf{Q}}_2^{*}\mathbf{Q}_1 \in M_{(n-k)\times k}(\mathbb{C}).$$

It measures the angles, and it satisfies an equation.

The unperturbed subspace, read in the perturbed complement, solves a Sylvester equation whose coefficients are the two spectra and whose right-hand side is the perturbation.

Lemma 20.9.2. The Subspace Perturbation Equation

In the notation above,

$$\widetilde{\mathbf{\Lambda}}_2\mathbf{X} - \mathbf{X}\mathbf{\Lambda}_1 = \widetilde{\mathbf{Q}}_2^{*}\mathbf{E}\mathbf{Q}_1, \tag{20.9.1}$$

and the nonzero singular values of $\mathbf{X}$ are the nonzero sines $\sin\theta_i$ of the principal angles between $U = \operatorname{col}(\mathbf{Q}_1)$ and $W = \operatorname{col}(\widetilde{\mathbf{Q}}_1)$, with multiplicity. In particular

$$\|\mathbf{X}\|_F = \|\sin\Theta(U,W)\|_F, \qquad \|\mathbf{X}\|_2 = \sin\theta_k = \|\sin\Theta(U,W)\|_2,$$

where $\sin\Theta(U,W) = \operatorname{diag}(\sin\theta_1,\dots,\sin\theta_k)$.

Idea. Compute $\widetilde{\mathbf{Q}}_2^{*}\widetilde{\mathbf{A}}\mathbf{Q}_1$ twice. Letting $\widetilde{\mathbf{A}}$ act to the left, on the perturbed eigenvectors, produces $\widetilde{\mathbf{\Lambda}}_2$. Writing $\widetilde{\mathbf{A}} = \mathbf{A} + \mathbf{E}$ and letting $\mathbf{A}$ act to the right, on the unperturbed ones, produces $\mathbf{\Lambda}_1$, plus an error term that carries $\mathbf{E}$. The difference of the two computations is the identity.

Proof. Taking conjugate transposes in $\widetilde{\mathbf{A}}\widetilde{\mathbf{Q}}_2 = \widetilde{\mathbf{Q}}_2\widetilde{\mathbf{\Lambda}}_2$, and using $\widetilde{\mathbf{A}}^{*} = \widetilde{\mathbf{A}}$ and $\widetilde{\mathbf{\Lambda}}_2^{*} = \widetilde{\mathbf{\Lambda}}_2$ (a real diagonal matrix), gives $\widetilde{\mathbf{Q}}_2^{*}\widetilde{\mathbf{A}} = \widetilde{\mathbf{\Lambda}}_2\widetilde{\mathbf{Q}}_2^{*}$. Hence

$$\widetilde{\mathbf{Q}}_2^{*}\widetilde{\mathbf{A}}\mathbf{Q}_1 = \widetilde{\mathbf{\Lambda}}_2\widetilde{\mathbf{Q}}_2^{*}\mathbf{Q}_1 = \widetilde{\mathbf{\Lambda}}_2\mathbf{X}.$$

On the other hand $\widetilde{\mathbf{A}} = \mathbf{A} + \mathbf{E}$ and $\mathbf{A}\mathbf{Q}_1 = \mathbf{Q}_1\mathbf{\Lambda}_1$, so

$$\begin{aligned}\widetilde{\mathbf{Q}}_2^{*}\widetilde{\mathbf{A}}\mathbf{Q}_1 &= \widetilde{\mathbf{Q}}_2^{*}\mathbf{Q}_1\mathbf{\Lambda}_1 + \widetilde{\mathbf{Q}}_2^{*}\mathbf{E}\mathbf{Q}_1\\ &= \mathbf{X}\mathbf{\Lambda}_1 + \widetilde{\mathbf{Q}}_2^{*}\mathbf{E}\mathbf{Q}_1.\end{aligned}$$

Subtracting the two expressions gives Eq. 20.9.1.

For the angles: the columns of the unitary $\widetilde{\mathbf{Q}}$ form an orthonormal basis of $\mathbb{C}^n$ (Theorem 11.7.3), so the $n-k$ columns of $\widetilde{\mathbf{Q}}_2$ are orthonormal and orthogonal to every column of $\widetilde{\mathbf{Q}}_1$. They therefore span a subspace of $W^{\perp}$ of dimension $n-k$, which is all of $W^{\perp}$ because $\dim W^{\perp} = n-k$ by Theorem 11.3.2 (c). So $\widetilde{\mathbf{Q}}_2$ is a matrix whose columns are an orthonormal basis of $W^{\perp}$, and Theorem 20.8.4 (a) says that the singular values of $\widetilde{\mathbf{Q}}_2^{*}\mathbf{Q}_1 = \mathbf{X}$ are the $\sin\theta_i$, as a multiset completed by zeros; zeros change neither a sum of squares nor a maximum of nonnegative numbers. For the two norms, $\mathbf{I} - P_W = P_{W^{\perp}} = \widetilde{\mathbf{Q}}_2\widetilde{\mathbf{Q}}_2^{*}$ by Lemma 20.8.1

(d) and (a), so $(\mathbf{I} - P_W)\mathbf{Q}_1 = \widetilde{\mathbf{Q}}_2\mathbf{X}$, and Lemma 20.9.1 (c) removes the factor $\widetilde{\mathbf{Q}}_2$:

$$\|\mathbf{X}\|_F = \|(\mathbf{I} - P_W)\mathbf{Q}_1\|_F, \qquad \|\mathbf{X}\|_2 = \|(\mathbf{I} - P_W)\mathbf{Q}_1\|_2.$$

By Theorem 20.8.4 (b), applied with U and W, these are $\|\sin\Theta(U,W)\|_F$ and $\sin\theta_k = \|\sin\Theta(U,W)\|_2$. This proves the lemma.

■

Quick check. Take $\mathbf{A} = \widetilde{\mathbf{A}} = \mathbf{I}_2$, so $\mathbf{E} = \mathbf{O}$, with $k = 1$, $\mathbf{Q}_1 = \mathbf{e}_1$, and $\widetilde{\mathbf{Q}}_1 = \frac{1}{\sqrt{2}}(1,1)$; both are legitimate choices of unit eigenvector. The subspaces U and W are 45° apart, although the matrices are equal. What does Eq. 20.9.1 say here, and why is there no contradiction?

Solution. Here $\mathbf{\Lambda}_1 = (1)$ and $\widetilde{\mathbf{\Lambda}}_2 = (1)$, so the equation reads $1 \cdot \mathbf{X} - \mathbf{X} \cdot 1 = \mathbf{O}$, which **every** 1×1 matrix satisfies. The equation holds, as it must, but it does not determine $\mathbf{X}$, and $\mathbf{X} = \widetilde{\mathbf{Q}}_2^*\mathbf{Q}_1 = \frac{1}{\sqrt{2}}(1,-1)\cdot\mathbf{e}_1 = \frac{1}{\sqrt{2}} = \sin\frac{\pi}{4}$ is one of its many solutions. The trouble is that the spectra of $\widetilde{\mathbf{\Lambda}}_2$ and $\mathbf{\Lambda}_1$ overlap. The theorems below assume that they are separated, and then the equation has exactly one solution (Theorem 12.10.2), which is small when the right-hand side is.

20.9.4 The $\sin\Theta$ theorem in the Frobenius norm

With the identity in hand, the theorem is a matter of dividing by the gap. The gap that matters compares **perturbed** eigenvalues outside the subspace with **unperturbed** eigenvalues inside it, because those are the two coefficient matrices of the equation.

Theorem 20.9.3. The Davis–Kahan $\sin\Theta$ Theorem

Let $\mathbf{A}, \widetilde{\mathbf{A}} = \mathbf{A} + \mathbf{E} \in M_n(\mathbb{C})$ be Hermitian, let $1 \le k \le n-1$, and let $\mathbf{Q} = \begin{pmatrix}\mathbf{Q}_1 & \mathbf{Q}_2\end{pmatrix}$ and $\widetilde{\mathbf{Q}} = \begin{pmatrix}\widetilde{\mathbf{Q}}_1 & \widetilde{\mathbf{Q}}_2\end{pmatrix}$ be unitary, with $\mathbf{Q}_1, \widetilde{\mathbf{Q}}_1 \in M_{n\times k}(\mathbb{C})$, such that

$$\mathbf{Q}^*\mathbf{A}\mathbf{Q} = \mathbf{\Lambda}_1 \oplus \mathbf{\Lambda}_2, \qquad \widetilde{\mathbf{Q}}^*\widetilde{\mathbf{A}}\widetilde{\mathbf{Q}} = \widetilde{\mathbf{\Lambda}}_1 \oplus \widetilde{\mathbf{\Lambda}}_2$$

are real diagonal. Suppose there is a $\delta > 0$ such that $|\mu - \lambda| \ge \delta$ for **every** diagonal entry μ of $\widetilde{\mathbf{\Lambda}}_2$ and **every** diagonal entry λ of $\mathbf{\Lambda}_1$. Then the principal angles $\Theta = \Theta(\mathrm{col}(\mathbf{Q}_1), \mathrm{col}(\widetilde{\mathbf{Q}}_1))$ satisfy

$$\|\sin\Theta\|_F \;\le\; \frac{\|\mathbf{E}\mathbf{Q}_1\|_F}{\delta} \;\le\; \frac{\sqrt{k}\,\|\mathbf{E}\|_2}{\delta}.$$

Idea. Eq. 20.9.1 is a Sylvester equation whose two coefficient matrices are diagonal, hence normal, and for a normal pair the separation of Section 7 is exactly the distance between the two spectra. The hypothesis says that this distance is at least δ, so Section 7's bound divides the right-hand side by δ. What is left is Lemma 20.9.2, which reads the size of $\mathbf{X}$ as the size of $\sin\Theta$.

Proof. Write $\mu_1, \dots, \mu_{n-k}$ for the diagonal entries of $\widetilde{\mathbf{\Lambda}}_2$, $\lambda_1, \dots, \lambda_k$ for those of $\mathbf{\Lambda}_1$, and $\mathbf{Y} = \widetilde{\mathbf{Q}}_2^{*}\mathbf{E}\mathbf{Q}_1$, so that $\widetilde{\mathbf{\Lambda}}_2\mathbf{X} - \mathbf{X}\mathbf{\Lambda}_1 = \mathbf{Y}$ by Lemma 20.9.2. A diagonal matrix is normal, and its eigenvalues are its diagonal entries, so Theorem 20.7.2 gives

$$\operatorname{sep}_F(\widetilde{\mathbf{\Lambda}}_2, \mathbf{\Lambda}_1) = \min_{i,j} |\mu_i - \lambda_j| \;\geq\; \delta > 0.$$

In particular the two spectra are disjoint, so $\mathbf{X}$ is the **unique** solution of that equation (Theorem 12.10.2), and Proposition 20.7.1 (d) applies to it:

$$\|\mathbf{X}\|_F \;\leq\; \frac{\|\mathbf{Y}\|_F}{\operatorname{sep}_F(\widetilde{\mathbf{\Lambda}}_2, \mathbf{\Lambda}_1)} \;\leq\; \frac{\|\mathbf{Y}\|_F}{\delta}.$$

The columns of $\widetilde{\mathbf{Q}}_2$ are orthonormal, so Lemma 20.9.1 (c) gives $\|\mathbf{Y}\|_F = \|\widetilde{\mathbf{Q}}_2^{*}(\mathbf{E}\mathbf{Q}_1)\|_F \leq \|\mathbf{E}\mathbf{Q}_1\|_F$. By Lemma 20.9.1 (b), $\|\mathbf{E}\mathbf{Q}_1\|_F \leq \|\mathbf{E}\|_2 \|\mathbf{Q}_1\|_F$, and $\|\mathbf{Q}_1\|_F = \sqrt{k}$ because $\mathbf{Q}_1$ has k unit columns. Finally $\|\mathbf{X}\|_F = \|\sin\Theta\|_F$ by Lemma 20.9.2. Chaining the three inequalities proves the theorem.

■

The right-hand side is a Lipschitz bound with constant $1/\delta$, and it is sharp. In Example (A tiny perturbation that turns the eigenvectors through 45°), take $k = 1$, $\mathbf{Q}_1 = \mathbf{e}_1$ with $\mathbf{\Lambda}_1 = (1 + \eta)$, and $\widetilde{\mathbf{Q}}_2 = \frac{1}{\sqrt{2}}(1, -1)$ with $\widetilde{\mathbf{\Lambda}}_2 = (1 - \eta)$. Then $\delta = 2\eta$ and $\mathbf{E}\mathbf{e}_1 = \eta(-1, 1)$ has length $\sqrt{2}\,\eta$, so the theorem gives

$$\sin\theta_1 \leq \frac{\sqrt{2}\,\eta}{2\eta} = \frac{1}{\sqrt{2}},$$

and the true angle is $\pi/4$, whose sine is exactly $1/\sqrt{2}$. The large rotation there was not a failure of the theorem. It is what the theorem predicts when the gap is no bigger than the perturbation.

Remark. The roles of $\mathbf{A}$ and $\widetilde{\mathbf{A}}$ may be exchanged, since $\mathbf{A} = \widetilde{\mathbf{A}} + (-\mathbf{E})$ and the principal angles between U and W are those between W and U (a matrix and its conjugate transpose have the same singular values). This gives the companion bound $\|\sin\Theta\|_F \leq \|\mathbf{E}\widetilde{\mathbf{Q}}_1\|_F/\delta'$, where δ' separates the diagonal of $\mathbf{\Lambda}_2$ from that of $\widetilde{\mathbf{\Lambda}}_1$. Use whichever one's data are at hand.

20.9.5 A one-sided version in the spectral norm

The Frobenius theorem controls all k angles at once. Often only the largest angle matters, $\|\sin\Theta\|_2 = \sin\theta_k$, and the Frobenius theorem bounds it too, since $\|\sin\Theta\|_2 \leq \|\sin\Theta\|_F$. But its right-hand side $\|\mathbf{E}\mathbf{Q}_1\|_F$ can be $\sqrt{k}$ times $\|\mathbf{E}\mathbf{Q}_1\|_2$, and for large k that factor is a real loss. When the two spectra are separated by a single point, one entirely above it and the other entirely below, the one-sided bound of Section 7 removes it.

Theorem 20.9.4. The $\sin\Theta$ Theorem, One-Sided Case

In the setting of Theorem 20.9.3, suppose instead that there are $a \in \mathbb{R}$ and $\delta > 0$ such that either

(i) every eigenvalue in $\mathbf{\Lambda}_1$ is $\geq a$ and every eigenvalue in $\widetilde{\mathbf{\Lambda}}_2$ is $\leq a - \delta$, or

(ii) every eigenvalue in $\mathbf{\Lambda}_1$ is $\le a$ and every eigenvalue in $\widetilde{\mathbf{\Lambda}}_2$ is $\ge a+\delta$.

Then

$$\|\sin\Theta\|_2 = \sin\theta_k \;\le\; \frac{\|\mathbf{E}\mathbf{Q}_1\|_2}{\delta} \;\le\; \frac{\|\mathbf{E}\|_2}{\delta}.$$

Proof. Let $\mathbf{X} = \widetilde{\mathbf{Q}}_2^*\mathbf{Q}_1$ and $\mathbf{Y} = \widetilde{\mathbf{Q}}_2^*\mathbf{E}\mathbf{Q}_1$, so that $\widetilde{\mathbf{\Lambda}}_2\mathbf{X} - \mathbf{X}\mathbf{\Lambda}_1 = \mathbf{Y}$ by Lemma 20.9.2. For a real diagonal matrix, $\mathbf{\Lambda} \succeq \alpha\mathbf{I}$ means that every diagonal entry is at least α (Theorem 13.1.4 (b), applied to $\mathbf{\Lambda} - \alpha\mathbf{I}$).

Case (ii). Here $\widetilde{\mathbf{\Lambda}}_2 \succeq (a+\delta)\mathbf{I}$ and $\mathbf{\Lambda}_1 \preceq a\mathbf{I}$, with $a+\delta > a$. By Theorem 20.7.3, applied with $\widetilde{\mathbf{\Lambda}}_2$ and $\mathbf{\Lambda}_1$ in the roles of its two Hermitian coefficients, the solution of $\widetilde{\mathbf{\Lambda}}_2\mathbf{X} - \mathbf{X}\mathbf{\Lambda}_1 = \mathbf{Y}$, which is unique because the two spectra are disjoint, satisfies $\|\mathbf{X}\|_2 \le \|\mathbf{Y}\|_2/\delta$.

Case (i). Multiply the equation by -1: $(-\widetilde{\mathbf{\Lambda}}_2)\mathbf{X} - \mathbf{X}(-\mathbf{\Lambda}_1) = -\mathbf{Y}$, where now $-\widetilde{\mathbf{\Lambda}}_2 \succeq (\delta-a)\mathbf{I}$ and $-\mathbf{\Lambda}_1 \preceq -a\mathbf{I}$, and $(\delta - a) - (-a) = \delta > 0$. So case (ii)'s argument applies and gives $\|\mathbf{X}\|_2 \le \|-\mathbf{Y}\|_2/\delta = \|\mathbf{Y}\|_2/\delta$.

In both cases, Lemma 20.9.1 (c) and (b) give $\|\mathbf{Y}\|_2 \le \|\mathbf{E}\mathbf{Q}_1\|_2 \le \|\mathbf{E}\|_2\|\mathbf{Q}_1\|_2 = \|\mathbf{E}\|_2$, and $\|\mathbf{X}\|_2 = \sin\theta_k$ by Lemma 20.9.2. This proves the theorem.

■

The typical use is the top of the spectrum: $\mathbf{\Lambda}_1$ holds the k largest eigenvalues of $\mathbf{A}$, and $\widetilde{\mathbf{\Lambda}}_2$ the $n-k$ smallest of $\widetilde{\mathbf{A}}$. Then (i) holds with $a = \lambda_k(\mathbf{A})$ and $\delta = \lambda_k(\mathbf{A}) - \lambda_{k+1}(\widetilde{\mathbf{A}})$, provided that number is positive.

Remark. Davis and Kahan proved more. If the eigenvalues in $\mathbf{\Lambda}_1$ lie in an interval $[a,b]$ and every eigenvalue in $\widetilde{\mathbf{\Lambda}}_2$ lies outside $(a-\delta, b+\delta)$, on either side, then the spectral bound $\sin\theta_k \le \|\mathbf{E}\mathbf{Q}_1\|_2/\delta$ still holds. The proof needs more than the one-sided Sylvester bound, and we do not give it. Nothing in this book depends on it. When the two spectra merely avoid each other by δ, interleaved in any pattern, what we have proved is the Frobenius theorem, and through $\|\sin\Theta\|_2 \le \|\sin\Theta\|_F$ the spectral consequence $\sin\theta_k \le \|\mathbf{E}\mathbf{Q}_1\|_F/\delta$.

20.9.6 Single eigenvectors

The most common question is about one eigenvector. A simple eigenvalue has a one-dimensional eigenspace, so its unit eigenvector is determined up to a unimodular factor, and the right measure of how far it moves is the angle between the two **lines**. For unit vectors $\mathbf{x}$ and $\widetilde{\mathbf{x}}$ we write $\theta(\mathbf{x},\widetilde{\mathbf{x}}) \in [0,\pi/2]$ for that angle, so that $\cos\theta(\mathbf{x},\widetilde{\mathbf{x}}) = |\langle\widetilde{\mathbf{x}},\mathbf{x}\rangle|$. This is the single principal angle between $\operatorname{span}(\mathbf{x})$ and $\operatorname{span}(\widetilde{\mathbf{x}})$, since the 1×1 matrix $\mathbf{x}^*\widetilde{\mathbf{x}}$ has the singular value $|\mathbf{x}^*\widetilde{\mathbf{x}}|$.

The theorem needs a separation between the perturbed eigenvalues of $\widetilde{\mathbf{A}}$ and the unperturbed one of $\mathbf{A}$. In practice only $\mathbf{A}$ is known, and Weyl's inequality converts a gap in the spectrum of $\mathbf{A}$ into the separation the theorem wants.

Corollary 20.9.5. Perturbation of a Single Eigenvector

Let $\mathbf{A}, \widetilde{\mathbf{A}} = \mathbf{A} + \mathbf{E} \in M_n(\mathbb{C})$ be Hermitian, let $\lambda_i(\mathbf{A})$ be a **simple** eigenvalue of $\mathbf{A}$ with unit

eigenvector $\mathbf{x}$, and put

$$\gamma \coloneqq \min_{j \neq i} |\lambda_i(\mathbf{A}) - \lambda_j(\mathbf{A})| > 0.$$

If $\|\mathbf{E}\|_2 < \gamma/2$, then $\lambda_i(\widetilde{\mathbf{A}})$ is a simple eigenvalue of $\widetilde{\mathbf{A}}$, and every unit eigenvector $\widetilde{\mathbf{x}}$ of $\widetilde{\mathbf{A}}$ for it satisfies

$$\sin\theta(\mathbf{x}, \widetilde{\mathbf{x}}) \;\leq\; \frac{\|\mathbf{E}\|_2}{\gamma - \|\mathbf{E}\|_2} \;\leq\; \frac{2\|\mathbf{E}\|_2}{\gamma}.$$

Idea. Apply the theorem with $k = 1$. The only thing to supply is δ: every eigenvalue $\lambda_j(\widetilde{\mathbf{A}})$ with $j \neq i$ sits within $\|\mathbf{E}\|_2$ of $\lambda_j(\mathbf{A})$ by Weyl, which is at least γ away from $\lambda_i(\mathbf{A})$. So the separation is at least $\gamma - \|\mathbf{E}\|_2$, by the triangle inequality. The hypothesis $\|\mathbf{E}\|_2 < \gamma/2$ keeps $\lambda_i(\widetilde{\mathbf{A}})$ away from its neighbors too, so that "the" eigenvector makes sense.

Proof. Put $\varepsilon = \|\mathbf{E}\|_2$. By Corollary 17.3.6, $|\lambda_j(\widetilde{\mathbf{A}}) - \lambda_j(\mathbf{A})| \leq \varepsilon$ for every j. Hence for $j \neq i$, by the triangle inequality in $\mathbb{R}$ (Theorem 0.6.4),

$$\begin{aligned} |\lambda_j(\widetilde{\mathbf{A}}) - \lambda_i(\mathbf{A})| &\geq |\lambda_j(\mathbf{A}) - \lambda_i(\mathbf{A})| - |\lambda_j(\widetilde{\mathbf{A}}) - \lambda_j(\mathbf{A})| \\ &\geq \gamma - \varepsilon, \end{aligned}$$

and similarly $|\lambda_j(\widetilde{\mathbf{A}}) - \lambda_i(\widetilde{\mathbf{A}})| \geq \gamma - 2\varepsilon > 0$. So $\lambda_i(\widetilde{\mathbf{A}})$ differs from every other $\lambda_j(\widetilde{\mathbf{A}})$, and it is a simple eigenvalue of $\widetilde{\mathbf{A}}$.

By Corollary 12.4.3 there is a unitary matrix whose columns are eigenvectors of $\widetilde{\mathbf{A}}$, with the eigenvalue $\lambda_i(\widetilde{\mathbf{A}})$ first. That eigenspace is one-dimensional, so the first column is $c\widetilde{\mathbf{x}}$ with $|c| = 1$, and replacing it by $\widetilde{\mathbf{x}}$ keeps the columns orthonormal eigenvectors. Call the result $\widetilde{\mathbf{Q}} = \begin{pmatrix} \widetilde{\mathbf{x}} & \widetilde{\mathbf{Q}}_2 \end{pmatrix}$. Then $\widetilde{\mathbf{\Lambda}}_2$ is diagonal with entries $\lambda_j(\widetilde{\mathbf{A}})$, $j \neq i$. In the same way build $\mathbf{Q} = \begin{pmatrix} \mathbf{x} & \mathbf{Q}_2 \end{pmatrix}$ for $\mathbf{A}$, with $\mathbf{\Lambda}_1 = (\lambda_i(\mathbf{A}))$. By the display, the hypothesis of Theorem 20.9.3 holds with $k = 1$ and $\delta = \gamma - \varepsilon > 0$, so

$$\sin\theta(\mathbf{x}, \widetilde{\mathbf{x}}) = \|\sin\Theta\|_F \leq \frac{\|\mathbf{E}\mathbf{x}\|}{\gamma - \varepsilon} \leq \frac{\varepsilon}{\gamma - \varepsilon},$$

using $\|\mathbf{E}\mathbf{x}\| \leq \|\mathbf{E}\|_2\|\mathbf{x}\| = \varepsilon$. Finally $\varepsilon < \gamma/2$ gives $\gamma - \varepsilon > \gamma/2$, so $\varepsilon/(\gamma - \varepsilon) \leq 2\varepsilon/\gamma$. This proves the corollary.

■

To first order in $\|\mathbf{E}\|_2$ the bound reads $\sin\theta \lesssim \|\mathbf{E}\|_2/\gamma$: *an eigenvector is as well conditioned as its eigenvalue is isolated.* The constant cannot be improved, and Exercise (C2: The constant in the eigenvector bound is sharp) shows it. The factor 2 in the cruder form is the price of the hypothesis $\|\mathbf{E}\|_2 < \gamma/2$, nothing more. In Example (A tiny perturbation that turns the eigenvectors through 45°), $\gamma = 2\eta$ and $\|\mathbf{E}\|_2 = \sqrt{2}\,\eta > \gamma/2$, so the corollary does not apply there.

Example. The theorem and the corollary, side by side

Let $\mathbf{A} = \operatorname{diag}(2, 0)$ and $\mathbf{E} = \frac{3}{4}\left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$. Find the exact angle between the eigenvectors of $\mathbf{A}$ and of $\widetilde{\mathbf{A}} = \mathbf{A} + \mathbf{E}$ for the larger eigenvalue, and compare it with the bounds of Theorem 20.9.3

and Corollary 20.9.5.

Solution. Exact. $\widetilde{\mathbf{A}} = \left(\begin{smallmatrix} 2 & 3/4 \\ 3/4 & 0 \end{smallmatrix}\right)$ has trace 2 and determinant $-\frac{9}{16}$, so its eigenvalues are $1 \pm \sqrt{1 + 9/16} = 1 \pm \frac{5}{4}$, that is $\frac{9}{4}$ and $-\frac{1}{4}$. For $\frac{9}{4}$, the first row of $\widetilde{\mathbf{A}} - \frac{9}{4}\mathbf{I}$ is $(-\frac{1}{4}, \frac{3}{4})$, which is orthogonal to $(3, 1)$, and the second row $(\frac{3}{4}, -\frac{9}{4})$ is a multiple of the first; so $\widetilde{\mathbf{x}} = \frac{1}{\sqrt{10}}(3, 1)$. With $\mathbf{x} = \mathbf{e}_1$,

$$\cos\theta = \frac{3}{\sqrt{10}}, \qquad \sin\theta = \frac{1}{\sqrt{10}} \approx 0.316.$$

The theorem. Take $\mathbf{\Lambda}_1 = (2)$ and $\widetilde{\mathbf{\Lambda}}_2 = (-\frac{1}{4})$, so $\delta = \frac{9}{4}$. Since $\mathbf{E}\mathbf{e}_1 = (0, \frac{3}{4})$, the bound is $\frac{3}{4}/\frac{9}{4} = \frac{1}{3} \approx 0.333$, within six percent of the truth.

The corollary. The gap is $\gamma = 2$, and $\mathbf{E}$ is $\frac{3}{4}$ times a Hermitian matrix with eigenvalues ± 1, so $\|\mathbf{E}\|_2 = \frac{3}{4} < 1 = \gamma/2$. The bound is $\frac{3/4}{2 - 3/4} = \frac{3}{5} = 0.6$, and its cruder form is $2 \cdot \frac{3}{4}/2 = \frac{3}{4}$.

The theorem is nearly exact because it uses the perturbed eigenvalue $-\frac{1}{4}$ itself. The corollary knows only that this eigenvalue lies within $\frac{3}{4}$ of 0, and it must allow for the worst case $\frac{3}{4}$, which would leave a separation of only $\frac{5}{4}$.

20.9.7 Clusters: the subspace is stable when the vectors are not

Example (A tiny perturbation that turns the eigenvectors through 45°) looks like bad news for eigenvectors, and it is, for **individual** eigenvectors of close eigenvalues. But the theorem was stated for subspaces, and that is its real strength. The subspace spanned by the eigenvectors of a cluster of eigenvalues is governed by the gap between the cluster and the rest of the spectrum. The gaps inside the cluster do not enter.

Example. A stable plane with unstable lines inside it

Let $0 < \eta < 1$ and

$$\mathbf{A} = \begin{pmatrix} 1+\eta & 0 & 0 \\ 0 & 1-\eta & 0 \\ 0 & 0 & -2 \end{pmatrix}, \qquad \mathbf{E} = \eta \begin{pmatrix} -1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Compare what happens to the eigenvectors for $1 \pm \eta$ with what happens to the plane they span. Then bound the motion of that plane under **any** Hermitian perturbation $\mathbf{E}'$ with $\|\mathbf{E}'\|_2 = \varepsilon < 3 - \eta$.

Solution. The upper-left 2×2 block is Example (A tiny perturbation that turns the eigenvectors through 45°). So the eigenvalues of $\widetilde{\mathbf{A}} = \mathbf{A} + \mathbf{E}$ are again $1 + \eta, 1 - \eta, -2$, and the eigenvectors for $1 \pm \eta$ turn through $\pi/4$, to $\frac{1}{\sqrt{2}}(1, \pm 1, 0)$. But they turn **inside** the plane $U = \operatorname{span}(\mathbf{e}_1, \mathbf{e}_2)$, which is invariant under both matrices. So the plane has not moved at all: both principal angles between U and the new plane $W = U$ are 0.

Theorem 20.9.3 with $k = 2$ sees this. Take $\mathbf{Q}_1 = \begin{pmatrix} \mathbf{e}_1 & \mathbf{e}_2 \end{pmatrix}$, $\mathbf{\Lambda}_1 = \operatorname{diag}(1 + \eta, 1 - \eta)$, and

$\widetilde{\mathbf{\Lambda}}_2 = (-2)$. The separation is $\delta = (1-\eta) - (-2) = 3 - \eta$, and $\|\mathbf{E}\mathbf{Q}_1\|_F = \|\mathbf{E}\|_F = 2\eta$, so

$$\|\sin\Theta\|_F \le \frac{2\eta}{3-\eta}.$$

This is small with η, and it has no 2η in its denominator. (The true value is 0; the theorem does not know that $\mathbf{E}$ happens to preserve U.)

For a general Hermitian $\mathbf{E}'$ with $\|\mathbf{E}'\|_2 = \varepsilon$, Corollary 17.3.6 puts the smallest eigenvalue of $\mathbf{A} + \mathbf{E}'$ at most $-2 + \varepsilon$, while the eigenvalues of $\mathbf{\Lambda}_1$ are at least $1 - \eta$. So case (i) of Theorem 20.9.4 holds with $a = 1 - \eta$ and $\delta = 3 - \eta - \varepsilon > 0$, and the plane moves by

$$\sin\theta_2 \le \frac{\varepsilon}{3-\eta-\varepsilon}.$$

The individual eigenvectors are covered by Corollary 20.9.5 only when $\varepsilon < \eta$, and then only with a bound of order $\varepsilon/(2\eta)$.

This is the practical lesson of the section. When a matrix has a cluster of close eigenvalues, do not ask for the individual eigenvectors of the cluster. They are ill-conditioned and may be meaningless at the precision of the data. Ask for an orthonormal basis of the invariant subspace of the whole cluster, which is as stable as the cluster is isolated. Numerical eigensolvers act on this advice.

20.9.8 Exercises

20.9.8.1 A. Check your understanding

Exercise. A1

(a) State the subspace perturbation equation Eq. 20.9.1, and say which matrix in it measures the principal angles.

(b) In Theorem 20.9.3, which two lists of eigenvalues must be separated by δ? Why not the eigenvalues of $\mathbf{A}$ inside and outside the subspace?

(c) True or false: "if $\|\mathbf{E}\|_2$ is tiny compared with $\|\mathbf{A}\|_2$, then every unit eigenvector of $\mathbf{A} + \mathbf{E}$ is close to a unit eigenvector of $\mathbf{A}$". Justify.

(d) Why is $\|\sin\Theta\|_2 \le \|\sin\Theta\|_F$, and when does Theorem 20.9.4 improve on what the Frobenius theorem gives for $\|\sin\Theta\|_2$?

(e) Which earlier result turns the bound on the matrix $\mathbf{X} = \widetilde{\mathbf{Q}}_2^*\mathbf{Q}_1$ into a bound on angles?

Solution.

(a) $\widetilde{\mathbf{\Lambda}}_2\mathbf{X} - \mathbf{X}\mathbf{\Lambda}_1 = \widetilde{\mathbf{Q}}_2^*\mathbf{E}\mathbf{Q}_1$, with $\mathbf{X} = \widetilde{\mathbf{Q}}_2^*\mathbf{Q}_1$; by Lemma 20.9.2 the singular values of $\mathbf{X}$ are the sines of the principal angles between $\mathrm{col}(\mathbf{Q}_1)$ and $\mathrm{col}(\widetilde{\mathbf{Q}}_1)$, together with zeros.

(b) The diagonal of $\widetilde{\mathbf{\Lambda}}_2$, the eigenvalues of the **perturbed** matrix belonging to the complement, against the diagonal of $\mathbf{\Lambda}_1$, the eigenvalues of the **unperturbed** matrix belonging to the subspace. These are the two coefficient matrices of the Sylvester equation, and the proof divides by their separation, which Theorem 20.7.2 computes as the smallest of those differences. A gap inside the spectrum of $\mathbf{A}$ alone is useful only after Corollary 17.3.6 converts it, as in Corollary 20.9.5.

(c) False. In Example (A tiny perturbation that turns the eigenvectors through 45°), $\|\mathbf{E}\|_2 = \sqrt{2}\,\eta$ can be as small as we like compared with $\|\mathbf{A}\|_2 = 1 + \eta$, yet each unit eigenvector of $\mathbf{A} + \mathbf{E}$ makes the angle $\pi/4$ with every unit eigenvector of $\mathbf{A}$, which are the multiples of $\mathbf{e}_1$ and $\mathbf{e}_2$. What controls eigenvectors is the perturbation relative to the gap, not relative to the matrix.

(d) $\|\sin\Theta\|_2 = \sin\theta_k$ is one of the numbers whose squares sum to $\|\sin\Theta\|_F^2$. The Frobenius theorem gives $\sin\theta_k \le \|\mathbf{E}\mathbf{Q}_1\|_F/\delta$, and $\|\mathbf{E}\mathbf{Q}_1\|_F$ can be as large as $\sqrt{k}\,\|\mathbf{E}\mathbf{Q}_1\|_2$. When the two spectra lie on opposite sides of a gap of width δ, Theorem 20.9.4 replaces it by $\|\mathbf{E}\mathbf{Q}_1\|_2$, which is never worse and can be $\sqrt{k}$ times smaller.

(e) Theorem 20.8.4, applied with $W = \operatorname{col}(\widetilde{\mathbf{Q}}_1)$ and the orthonormal basis $\widetilde{\mathbf{Q}}_2$ of $W^{\perp}$.

20.9.8.2 B. Practice

Exercise. B1: How far does the top eigenvector move?

Let

$$\mathbf{A} = \begin{pmatrix} 5 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & -1 \end{pmatrix}, \qquad \mathbf{E} = \frac{1}{2}\begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix},$$

and let $\widetilde{\mathbf{x}}$ be a unit eigenvector of $\widetilde{\mathbf{A}} = \mathbf{A} + \mathbf{E}$ for its largest eigenvalue.

(a) Compute $\|\mathbf{E}\|_2$ and the gap γ of $\lambda_1(\mathbf{A}) = 5$.

(b) Use Corollary 20.9.5 to bound $\sin\theta(\mathbf{e}_1, \widetilde{\mathbf{x}})$.

(c) Use Theorem 20.9.3, with the separation estimated by Corollary 17.3.6, to obtain a better bound. Hence show that the angle is less than 13°.

(d) Check that 2 is an eigenvalue of $\widetilde{\mathbf{A}}$ and find the other two. Recompute the bound of (c) with the exact separation.

Solution.

(a) The matrix $\mathbf{M} = 2\mathbf{E}$ is symmetric with characteristic polynomial $t^3 - 2t = t(t^2 - 2)$, so its eigenvalues are $0, \pm\sqrt{2}$ and $\|\mathbf{E}\|_2 = \frac{1}{2}\sqrt{2} = \frac{\sqrt{2}}{2} \approx 0.707$ by Lemma 17.3.5. The eigenvalues of $\mathbf{A}$ are $5, 2, -1$, so $\gamma = 5 - 2 = 3$.

(b) $\|\mathbf{E}\|_2 \approx 0.707 < 1.5 = \gamma/2$, so the corollary applies:

$$\sin\theta \le \frac{\sqrt{2}/2}{3 - \sqrt{2}/2} \approx 0.308.$$

(c) By Corollary 17.3.6, $\lambda_2(\widetilde{\mathbf{A}}) \le 2 + \frac{\sqrt{2}}{2}$ and $\lambda_3(\widetilde{\mathbf{A}}) \le -1 + \frac{\sqrt{2}}{2}$, so both lie at distance at least $\delta = 3 - \frac{\sqrt{2}}{2} \approx 2.293$ from 5. Apply Theorem 20.9.3 with $k = 1$, $\mathbf{Q}_1 = \mathbf{e}_1$, $\mathbf{\Lambda}_1 = (5)$, and $\widetilde{\mathbf{\Lambda}}_2 = \operatorname{diag}(\lambda_2(\widetilde{\mathbf{A}}), \lambda_3(\widetilde{\mathbf{A}}))$. Since $\mathbf{E}\mathbf{e}_1 = (0, \frac{1}{2}, 0)$,

$$\sin\theta \le \frac{\|\mathbf{E}\mathbf{e}_1\|}{\delta} = \frac{1/2}{3 - \sqrt{2}/2} \approx 0.218.$$

The gain over (b) comes from $\|\mathbf{E}\mathbf{e}_1\| = \frac{1}{2}$ replacing $\|\mathbf{E}\|_2 \approx 0.707$. Since $\sin 13° \approx 0.225 > 0.218$ and sin increases on $[0, \pi/2]$, the angle is less than 13°.

(d) $\widetilde{\mathbf{A}} - 2\mathbf{I} = \begin{pmatrix} 3 & 1/2 & 0 \\ 1/2 & 0 & 1/2 \\ 0 & 1/2 & -3 \end{pmatrix}$ sends $(1, -6, -1)$ to $(3-3,\ \frac{1}{2}-\frac{1}{2},\ -3+3) = \mathbf{0}$, so 2 is an eigenvalue. The other two have sum $\operatorname{tr}\widetilde{\mathbf{A}} - 2 = 4$ and product $\det\widetilde{\mathbf{A}}/2$, where expanding along the first row gives $\det\widetilde{\mathbf{A}} = 5(-2-\frac{1}{4}) - \frac{1}{2}(-\frac{1}{2}) = -11$. So they are the roots of $t^2 - 4t - \frac{11}{2}$, namely $2 \pm \sqrt{19/2} \approx 5.082, -1.082$. The exact separation of $\{2, 2-\sqrt{19/2}\}$ from 5 is 3, and the bound becomes $\frac{1/2}{3} = \frac{1}{6} \approx 0.167$. (A direct computation of $\tilde{\mathbf{x}}$ gives $\sin\theta \approx 0.163$.)

Exercise. B2: Which theorem applies?

In each case $\mathbf{A}$ and $\widetilde{\mathbf{A}}$ are Hermitian 4×4 matrices, $k = 2$, $\mathbf{\Lambda}_1$ lists the eigenvalues of $\mathbf{A}$ belonging to $\mathbf{Q}_1$, $\widetilde{\mathbf{\Lambda}}_2$ lists the eigenvalues of $\widetilde{\mathbf{A}}$ belonging to $\widetilde{\mathbf{Q}}_2$, and $\|\mathbf{E}\mathbf{Q}_1\|_2 = 0.2$, $\|\mathbf{E}\mathbf{Q}_1\|_F = 0.25$. Determine which of Theorem 20.9.3 and Theorem 20.9.4 apply, and give the best bound on $\sin\theta_k$ that they yield. Justify your answer.

(a) $\mathbf{\Lambda}_1 = \operatorname{diag}(3, 2)$, $\widetilde{\mathbf{\Lambda}}_2 = \operatorname{diag}(0.4, -1)$.

(b) $\mathbf{\Lambda}_1 = \operatorname{diag}(3, 0)$, $\widetilde{\mathbf{\Lambda}}_2 = \operatorname{diag}(1.9, 5)$.

(c) $\mathbf{\Lambda}_1 = \operatorname{diag}(2, 1)$, $\widetilde{\mathbf{\Lambda}}_2 = \operatorname{diag}(2, -1)$.

(d) $\mathbf{\Lambda}_1 = \operatorname{diag}(-1, -2)$, $\widetilde{\mathbf{\Lambda}}_2 = \operatorname{diag}(0.5, 4)$.

Solution.

(a) Every eigenvalue in $\mathbf{\Lambda}_1$ is ≥ 2 and both in $\widetilde{\mathbf{\Lambda}}_2$ are $\leq 0.4 = 2 - 1.6$. So case (i) of the one-sided theorem holds with $a = 2$, $\delta = 1.6$, and $\sin\theta_2 \leq 0.2/1.6 = 0.125$. The Frobenius theorem also applies, with the smallest of the four differences, $\delta = 2 - 0.4 = 1.6$, but gives only $\sin\theta_2 \leq \|\sin\Theta\|_F \leq 0.25/1.6 \approx 0.156$.

(b) The value 1.9 lies **between** 0 and 3, so no point has one list entirely above it and the other entirely below, and the one-sided theorem does not apply. The Frobenius theorem applies with $\delta = \min(1.1, 1.9, 2, 5) = 1.1$, the four differences being $|1.9-3|$, $|1.9-0|$, $|5-3|$, and $|5-0|$. It gives $\sin\theta_2 \leq \|\sin\Theta\|_F \leq 0.25/1.1 \approx 0.227$.

(c) $|2-2| = 0$, so no $\delta > 0$ exists and **neither** theorem applies. The check after Lemma 20.9.2 shows why nothing can be said in general: an eigenvalue of $\widetilde{\mathbf{A}}$ outside the subspace coincides with one of $\mathbf{A}$ inside it, and then the equation need not determine $\mathbf{X}$.

(d) Every eigenvalue in $\mathbf{\Lambda}_1$ is ≤ -1 and both in $\widetilde{\mathbf{\Lambda}}_2$ are $\geq 0.5 = -1 + 1.5$, so case (ii) holds with $a = -1$ and $\delta = 1.5$: $\sin\theta_2 \leq 0.2/1.5 \approx 0.133$. The Frobenius theorem, with $\delta = 0.5 - (-1) = 1.5$ again, gives $0.25/1.5 \approx 0.167$.

Exercise. B3: The same bounds for the projections

In the setting of Theorem 20.9.3, let $U = \operatorname{col}(\mathbf{Q}_1)$ and $W = \operatorname{col}(\widetilde{\mathbf{Q}}_1)$, with orthogonal projections P_U and P_W. Prove that

$$\|P_U - P_W\|_F \leq \frac{\sqrt{2}\,\|\mathbf{E}\mathbf{Q}_1\|_F}{\delta},$$

and that under the hypotheses of Theorem 20.9.4, $\|P_U - P_W\|_2 \le \|\mathbf{E}\mathbf{Q}_1\|_2/\delta$.

Solution. Both subspaces have dimension k, so Theorem 20.8.5 applies: $\|P_U - P_W\|_F = \sqrt{2}\,\|\sin\Theta\|_F$ and $\|P_U - P_W\|_2 = \sin\theta_k$. By Theorem 20.9.3, $\sqrt{2}\,\|\sin\Theta\|_F \le \sqrt{2}\,\|\mathbf{E}\mathbf{Q}_1\|_F/\delta$, which is the first claim. By Theorem 20.9.4, $\sin\theta_k \le \|\mathbf{E}\mathbf{Q}_1\|_2/\delta$, which is the second.

20.9.8.3 C. Going deeper

Exercise. C1: A bound that needs only the perturbed matrix

Let $\widetilde{\mathbf{A}} \in M_n(\mathbb{C})$ be Hermitian with $\widetilde{\mathbf{Q}} = \begin{pmatrix}\widetilde{\mathbf{Q}}_1 & \widetilde{\mathbf{Q}}_2\end{pmatrix}$ and $\widetilde{\mathbf{\Lambda}}_2$ as in the text. Let $\mathbf{Z} \in M_{n\times k}(\mathbb{C})$ be **any** matrix with orthonormal columns, not necessarily eigenvectors of anything, and put

$$\mathbf{M} = \mathbf{Z}^*\widetilde{\mathbf{A}}\mathbf{Z} \in M_k(\mathbb{C}), \qquad \mathbf{R} = \widetilde{\mathbf{A}}\mathbf{Z} - \mathbf{Z}\mathbf{M}.$$

(a) Prove that if every eigenvalue of the Hermitian matrix $\mathbf{M}$ differs from every diagonal entry of $\widetilde{\mathbf{\Lambda}}_2$ by at least $\delta > 0$, then the principal angles between $\mathrm{col}(\mathbf{Z})$ and $\mathrm{col}(\widetilde{\mathbf{Q}}_1)$ satisfy $\|\sin\Theta\|_F \le \|\mathbf{R}\|_F/\delta$.

(b) Deduce: let $\lambda_i(\widetilde{\mathbf{A}})$ be a **simple** eigenvalue of $\widetilde{\mathbf{A}}$, let $\mathbf{z}$ be a unit vector, $\mu = \mathbf{z}^*\widetilde{\mathbf{A}}\mathbf{z}$, and $\mathbf{r} = \widetilde{\mathbf{A}}\mathbf{z} - \mu\mathbf{z}$. If every eigenvalue of $\widetilde{\mathbf{A}}$ other than $\lambda_i(\widetilde{\mathbf{A}})$ lies at distance at least δ from μ, then a unit eigenvector $\tilde{\mathbf{x}}$ for $\lambda_i(\widetilde{\mathbf{A}})$ satisfies $\sin\theta(\mathbf{z}, \tilde{\mathbf{x}}) \le \|\mathbf{r}\|/\delta$.

Hint: diagonalize $\mathbf{M}$ *and compute* $\widetilde{\mathbf{Q}}_2^*\mathbf{R}$ *in two ways.*

Solution.

(a) $\mathbf{M}^* = \mathbf{Z}^*\widetilde{\mathbf{A}}^*\mathbf{Z} = \mathbf{M}$, so by Corollary 12.4.3 and Theorem 12.2.1, $\mathbf{M} = \mathbf{V}\mathbf{D}\mathbf{V}^*$ with $\mathbf{V} \in M_k(\mathbb{C})$ unitary and $\mathbf{D}$ real diagonal, holding the eigenvalues of $\mathbf{M}$. Put $\mathbf{Z}' = \mathbf{Z}\mathbf{V}$, which has orthonormal columns, since $\mathbf{Z}'^*\mathbf{Z}' = \mathbf{V}^*\mathbf{V} = \mathbf{I}_k$, and the same column space as $\mathbf{Z}$, since $\mathbf{V}$ is invertible. Then $\mathbf{R}\mathbf{V} = \widetilde{\mathbf{A}}\mathbf{Z}' - \mathbf{Z}\mathbf{V}\mathbf{D} = \widetilde{\mathbf{A}}\mathbf{Z}' - \mathbf{Z}'\mathbf{D}$. As in the proof of Lemma 20.9.2, $\widetilde{\mathbf{Q}}_2^*\widetilde{\mathbf{A}} = \widetilde{\mathbf{\Lambda}}_2\widetilde{\mathbf{Q}}_2^*$, so with $\mathbf{X}' = \widetilde{\mathbf{Q}}_2^*\mathbf{Z}'$,

$$\widetilde{\mathbf{Q}}_2^*\mathbf{R}\mathbf{V} = \widetilde{\mathbf{\Lambda}}_2\mathbf{X}' - \mathbf{X}'\mathbf{D}.$$

Both coefficient matrices are real diagonal, hence normal, and their diagonal entries differ by at least δ, so, as in the proof of Theorem 20.9.3, Theorem 20.7.2 and Proposition 20.7.1 (d) give $\|\mathbf{X}'\|_F \le \|\widetilde{\mathbf{Q}}_2^*\mathbf{R}\mathbf{V}\|_F/\delta$. By Lemma 20.9.1 (c), $\|\widetilde{\mathbf{Q}}_2^*\mathbf{R}\mathbf{V}\|_F \le \|\mathbf{R}\mathbf{V}\|_F = \|\mathbf{R}\|_F$. Finally, by Theorem 20.8.4 applied as in Lemma 20.9.2, with $\mathrm{col}(\mathbf{Z}') = \mathrm{col}(\mathbf{Z})$ in place of $\mathrm{col}(\mathbf{Q}_1)$, $\|\mathbf{X}'\|_F = \|\sin\Theta\|_F$. This proves (a).

(b) Take $k = 1$ and $\mathbf{Z} = \mathbf{z}$. Then $\mathbf{M} = (\mu)$, $\mathbf{R} = \mathbf{r}$, and the only eigenvalue of $\mathbf{M}$ is μ. Since $\lambda_i(\widetilde{\mathbf{A}})$ is simple, its eigenspace is the line $\mathrm{span}(\tilde{\mathbf{x}})$, so we may choose $\widetilde{\mathbf{Q}}$ with first column $\tilde{\mathbf{x}}$ and $\widetilde{\mathbf{\Lambda}}_2$ holding the other eigenvalues of $\widetilde{\mathbf{A}}$, as in the proof of Corollary 20.9.5. Part (a) gives $\sin\theta(\mathbf{z}, \tilde{\mathbf{x}}) \le \|\mathbf{r}\|/\delta$. This is the eigenvector companion of a residual bound: everything on the right is computable from $\mathbf{z}$ and $\widetilde{\mathbf{A}}$ and a gap estimate, with no unperturbed matrix in sight.

Exercise. C2: The constant in the eigenvector bound is sharp

Let $g > 0$ and $e > 0$, and let $\mathbf{A} = \operatorname{diag}(g, 0)$, $\mathbf{E} = e\left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$.

(a) Show that a unit eigenvector $\tilde{\mathbf{x}}$ of $\mathbf{A} + \mathbf{E}$ for its larger eigenvalue makes an angle $\theta \in [0, \pi/4)$ with $\mathbf{e}_1$ satisfying $\tan 2\theta = 2e/g$.

(b) Deduce that $\sin\theta \cdot g/e \to 1$ as $e \to 0^+$.

(c) Hence show that for no constant $c < 1$ is it true that $\sin\theta(\mathbf{x}, \tilde{\mathbf{x}}) \le c\,\|\mathbf{E}\|_2/\gamma$ for all sufficiently small Hermitian $\mathbf{E}$, in the setting of Corollary 20.9.5.

Solution.

(a) The larger eigenvalue of $\widetilde{\mathbf{A}} = \left(\begin{smallmatrix} g & e \\ e & 0 \end{smallmatrix}\right)$ is $\lambda_+ = h + s$, where $h = g/2$ and $s = \sqrt{h^2 + e^2}$. The first row of $\widetilde{\mathbf{A}} - \lambda_+ \mathbf{I}$ is $(g - \lambda_+, e) = (h - s, e)$, so $(e, s - h)$ spans the eigenline, and $s - h > 0$. The angle with $\mathbf{e}_1$ therefore satisfies $\tan\theta = t := (s - h)/e > 0$. Using $s^2 = h^2 + e^2$,

$$1 - t^2 = \frac{e^2 - (s-h)^2}{e^2} = \frac{2h(s-h)}{e^2} > 0,$$

so $t < 1$, that is $\theta < \pi/4$, and

$$\tan 2\theta = \frac{2t}{1-t^2} = \frac{2(s-h)}{e} \cdot \frac{e^2}{2h(s-h)} = \frac{e}{h} = \frac{2e}{g}.$$

(b) From (a), $\cos 2\theta = 1/\sqrt{1 + \tan^2 2\theta} = g/\sqrt{g^2 + 4e^2}$, since $2\theta \in [0, \pi/2)$. Hence

$$\sin^2\theta = \frac{1 - \cos 2\theta}{2} = \frac{\sqrt{g^2+4e^2} - g}{2\sqrt{g^2+4e^2}} = \frac{2e^2}{\sqrt{g^2+4e^2}\,(\sqrt{g^2+4e^2} + g)},$$

multiplying above and below by $\sqrt{g^2 + 4e^2} + g$. So

$$\left(\frac{g \sin\theta}{e}\right)^2 = \frac{2g^2}{\sqrt{g^2+4e^2}\,(\sqrt{g^2+4e^2} + g)},$$

which tends to $2g^2/(g \cdot 2g) = 1$ as $e \to 0^+$, by continuity of the square root.

(c) Here $\mathbf{x} = \mathbf{e}_1$, $\gamma = g$, and $\|\mathbf{E}\|_2 = e$, since $\mathbf{E}/e$ is Hermitian with eigenvalues ± 1. If $\sin\theta \le c\,e/g$ held for all small $e > 0$ with some $c < 1$, then $\sin\theta \cdot g/e \le c$ for those e, contradicting (b). So to first order in $\|\mathbf{E}\|_2$, the bound $\|\mathbf{E}\|_2/(\gamma - \|\mathbf{E}\|_2)$ of Corollary 20.9.5 cannot be improved by any constant factor.

20.10 Singular Values and Polar Factors

Chapter 17 §09 proved that singular values are perfectly conditioned: $|\sigma_i(\mathbf{A}+\mathbf{E}) - \sigma_i(\mathbf{A})| \le \|\mathbf{E}\|_2$, for any matrices at all. This section finishes that story and then starts a new one. First the Frobenius-norm version, Mirsky's theorem, which follows from Section 5's Hoffman–Wielandt

inequality once the Hermitian dilation of Chapter 17 turns singular values into eigenvalues. Then the polar decomposition $\mathbf{A} = \mathbf{UP}$: how far do the unitary factor and the positive factor move when $\mathbf{A}$ does? Both answers come from one tool used twice, the bound of Section 7 for a Sylvester equation with positive coefficients.

Throughout, the field is $\mathbb{C}$, $\|\cdot\|_2$ is the spectral norm and $\|\cdot\|_F$ the Frobenius norm, and Lemma 20.9.1 of Section 9 is used freely to move unitary factors and conjugate transposes in and out of either norm. For $\mathbf{A} \in M_{m\times n}(\mathbb{C})$ we put $p = \min(m, n)$, and $\sigma_1(\mathbf{A}) \geq \cdots \geq \sigma_p(\mathbf{A}) \geq 0$ are its singular values (Definition 13.9.1).

20.10.1 Singular values in the Frobenius norm

Here is where Chapter 17 left the question. Its Corollary 17.9.5 says that for **any** $\mathbf{A}, \mathbf{E} \in M_{m\times n}(\mathbb{C})$,

$$|\sigma_i(\mathbf{A} + \mathbf{E}) - \sigma_i(\mathbf{A})| \leq \|\mathbf{E}\|_2 \qquad (1 \leq i \leq p),$$

with no symmetry, squareness, or smallness assumed. In the language of Section 5 this is a bound on the largest entry of the vector of differences, and the matching is simply by index. What it does not give is a good bound on **all** the differences together. Squaring and adding p copies of it gives

$$\sum_{i=1}^{p} \big(\sigma_i(\mathbf{A} + \mathbf{E}) - \sigma_i(\mathbf{A})\big)^2 \leq p\,\|\mathbf{E}\|_2^2,$$

and the factor p grows with the size of the matrices. For Hermitian eigenvalues, Section 5 removed that factor: Corollary 20.5.4 bounds the sum of squares by $\|\mathbf{E}\|_F^2$, with no dependence on n. Chapter 17 recorded, in Proposition 17.9.3, a device for carrying any theorem about Hermitian eigenvalues over to singular values, and promised that this chapter would lean on it. Here it is.

Theorem 20.10.1. Mirsky's Theorem

Let $\mathbf{A}, \mathbf{B} \in M_{m\times n}(\mathbb{C})$ and $p = \min(m, n)$. Then

$$\sum_{i=1}^{p} \big(\sigma_i(\mathbf{A}) - \sigma_i(\mathbf{B})\big)^2 \;\leq\; \|\mathbf{A} - \mathbf{B}\|_F^2.$$

Idea. Replace each matrix by its Hermitian dilation $\mathcal{H}(\cdot)$, whose eigenvalues are $\pm\sigma_i$ and zeros. Listed decreasingly, the eigenvalues of $\mathcal{H}(\mathbf{A})$ are $\sigma_1(\mathbf{A}), \ldots, \sigma_p(\mathbf{A})$, then the zeros, then $-\sigma_p(\mathbf{A}), \ldots, -\sigma_1(\mathbf{A})$, and the positions depend only on m and n, so $\mathcal{H}(\mathbf{B})$ has the same pattern. Apply the Hermitian Hoffman–Wielandt inequality to the two dilations. On the left each difference $\sigma_i(\mathbf{A}) - \sigma_i(\mathbf{B})$ appears twice, once from the top and once from the bottom, while the zeros cancel against zeros. On the right each entry of $\mathbf{A} - \mathbf{B}$ appears twice, once in each off-diagonal block. The two factors of 2 cancel.

Proof. For $\mathbf{M} \in M_{m\times n}(\mathbb{C})$, $\mathcal{H}(\mathbf{M}) = \left(\begin{smallmatrix} \mathbf{O} & \mathbf{M} \\ \mathbf{M}^* & \mathbf{O} \end{smallmatrix}\right)$ is Hermitian (Proposition 17.9.3). Since conjugate transposition is additive,

$$\mathcal{H}(\mathbf{A}) - \mathcal{H}(\mathbf{B}) = \begin{pmatrix} \mathbf{O} & \mathbf{A} - \mathbf{B} \\ (\mathbf{A} - \mathbf{B})^* & \mathbf{O} \end{pmatrix} = \mathcal{H}(\mathbf{A} - \mathbf{B}),$$

and since the entries of $\mathcal{H}(\mathbf{M})$ are those of $\mathbf{M}$, those of $\mathbf{M}^*$, and zeros,

$$\|\mathcal{H}(\mathbf{M})\|_F^2 = \|\mathbf{M}\|_F^2 + \|\mathbf{M}^*\|_F^2 = 2\|\mathbf{M}\|_F^2. \tag{20.10.1}$$

By Proposition 17.9.3, $\mathcal{H}(\mathbf{A}) \in M_{m+n}(\mathbb{C})$ has eigenvalues $\pm\sigma_1(\mathbf{A}), \dots, \pm\sigma_p(\mathbf{A})$ and $m+n-2p$ zeros, with multiplicity. Since every $\sigma_i(\mathbf{A}) \geq 0$, the list

$$\begin{aligned} \sigma_1(\mathbf{A}) \geq \cdots \geq \sigma_p(\mathbf{A}) \;&\geq\; \underbrace{0 \geq \cdots \geq 0}_{m+n-2p} \\ &\geq\; -\sigma_p(\mathbf{A}) \geq \cdots \geq -\sigma_1(\mathbf{A}) \end{aligned}$$

is decreasing, and a multiset of real numbers has only one decreasing listing. So, with $N = m + n$,

$$\lambda_i(\mathcal{H}(\mathbf{A})) = \begin{cases} \sigma_i(\mathbf{A}) & 1 \leq i \leq p, \\ 0 & p < i \leq N - p, \\ -\sigma_{N+1-i}(\mathbf{A}) & N - p < i \leq N. \end{cases}$$

The same formula holds for $\mathbf{B}$, with the **same** three ranges of i, because m, n, and p are the same. Hence, summing over the three ranges, and writing $j = N + 1 - i$ in the third,

$$\begin{aligned} \sum_{i=1}^{N} \big(\lambda_i(\mathcal{H}(\mathbf{A})) - \lambda_i(\mathcal{H}(\mathbf{B}))\big)^2 &= \sum_{i=1}^{p} \big(\sigma_i(\mathbf{A}) - \sigma_i(\mathbf{B})\big)^2 + 0 \\ &\quad + \sum_{j=1}^{p} \big(-\sigma_j(\mathbf{A}) + \sigma_j(\mathbf{B})\big)^2 \\ &= 2\sum_{i=1}^{p} \big(\sigma_i(\mathbf{A}) - \sigma_i(\mathbf{B})\big)^2. \end{aligned}$$

By Corollary 20.5.4, applied to the Hermitian matrices $\mathcal{H}(\mathbf{A})$ and $\mathcal{H}(\mathbf{B})$, the left-hand side is at most $\|\mathcal{H}(\mathbf{A}) - \mathcal{H}(\mathbf{B})\|_F^2 = \|\mathcal{H}(\mathbf{A} - \mathbf{B})\|_F^2 = 2\|\mathbf{A} - \mathbf{B}\|_F^2$, by Eq. 20.10.1. Dividing by 2 proves the theorem.

■

This is the payment Chapter 17 §09 promised when it said that Chapter 20 leans on the dilation for perturbation bounds. Together with Corollary 17.9.5, it bounds the vector of singular-value differences in the two norms one reaches for first: its largest entry by $\|\mathbf{A} - \mathbf{B}\|_2$, its Euclidean length by $\|\mathbf{A} - \mathbf{B}\|_F$. Chapter 21 puts both inside one statement covering every unitarily invariant norm.

Note what was **not** assumed. Theorem 20.5.2 needs normal matrices, and Section 5 showed by a Jordan-block pair that it fails without normality. Mirsky's theorem holds for every pair of matrices, square or not, because the dilation manufactures the Hermitian hypothesis instead of assuming it.

Example. Two matrices that differ in one entry

Let $\mathbf{A} = \left(\begin{smallmatrix} 2 & 1 \\ 1 & 2 \end{smallmatrix}\right)$ and $\mathbf{B} = \left(\begin{smallmatrix} 2 & 1 \\ 0 & 2 \end{smallmatrix}\right)$. Compute the singular values of both, and compare the two sides of Theorem 20.10.1 with what Corollary 17.9.5 alone would give.

Solution. $\mathbf{A}$ is symmetric with $\mathbf{A}(1,1) = 3(1,1)$ and $\mathbf{A}(1,-1) = (1,-1)$, so $\mathbf{A}^\top\mathbf{A} = \mathbf{A}^2$ has eigenvalues 9 and 1, and $\sigma(\mathbf{A}) = (3,1)$. For $\mathbf{B}$,

$$\mathbf{B}^\top\mathbf{B} = \begin{pmatrix} 4 & 2 \\ 2 & 5 \end{pmatrix},$$

with trace 9 and determinant 16, so its eigenvalues are $\frac{9\pm\sqrt{17}}{2}$. These are perfect squares: $\left(\frac{\sqrt{17}\pm 1}{2}\right)^2 = \frac{18\pm 2\sqrt{17}}{4} = \frac{9\pm\sqrt{17}}{2}$. Hence

$$\sigma_1(\mathbf{B}) = \tfrac{\sqrt{17}+1}{2} \approx 2.562, \qquad \sigma_2(\mathbf{B}) = \tfrac{\sqrt{17}-1}{2} \approx 1.562.$$

The difference $\mathbf{A} - \mathbf{B} = \left(\begin{smallmatrix} 0 & 0 \\ 1 & 0 \end{smallmatrix}\right)$ has $\|\mathbf{A}-\mathbf{B}\|_F = \|\mathbf{A}-\mathbf{B}\|_2 = 1$. The left side of Mirsky's inequality is

$$\left(\tfrac{5-\sqrt{17}}{2}\right)^2 + \left(\tfrac{3-\sqrt{17}}{2}\right)^2 = \tfrac{(42-10\sqrt{17})+(26-6\sqrt{17})}{4} = 17 - 4\sqrt{17} \approx 0.508,$$

comfortably below 1. The index-by-index bound gives $|\sigma_i(\mathbf{A}) - \sigma_i(\mathbf{B})| \le 1$ for each i; the actual values are about 0.438 and 0.562. On its own it bounds the sum of squares only by 2, and Mirsky halves that.

Quick check. Let $\mathbf{A} = \operatorname{diag}(3,1)$ and $\mathbf{B} = \operatorname{diag}(1,3)$. Compute both sides of Theorem 20.10.1. Why is the left side not $(3-1)^2 + (1-3)^2$?

Solution. Both matrices have singular values $(3,1)$ in decreasing order, so the left side is 0, while $\|\mathbf{A}-\mathbf{B}\|_F^2 = \|\operatorname{diag}(2,-2)\|_F^2 = 8$. The theorem compares the **sorted** lists, the i-th largest with the i-th largest. Pairing the diagonal entries by position would give 8, which is also allowed, but it is not what the theorem measures.

Warning. **Stable singular values do not mean stable singular vectors.** Take $0 < \eta < 1$, $\mathbf{A} = \operatorname{diag}(1+\eta, 1-\eta)$ and $\mathbf{B} = \left(\begin{smallmatrix} 1 & \eta \\ \eta & 1 \end{smallmatrix}\right)$, the pair of Example (A tiny perturbation that turns the eigenvectors through 45°). Both are positive definite with eigenvalues $1 \pm \eta$, so both have singular values $1 \pm \eta$, and the left side of Mirsky's inequality is 0. But the right singular vector for σ_1 is $\mathbf{e}_1$ for $\mathbf{A}$ and $\frac{1}{\sqrt{2}}(1,1)$ for $\mathbf{B}$, a 45° turn, while $\|\mathbf{A}-\mathbf{B}\|_F = 2\eta$ is as small as we like. As with eigenvectors, singular vectors are controlled by gaps. Applying Section 9 to the dilations $\mathcal{H}(\mathbf{A})$ and $\mathcal{H}(\mathbf{B})$, whose eigenvectors are built from the singular vectors, is one way to bound them.

20.10.2 The square root is Lipschitz away from zero

The second half of the section is about the polar decomposition, and its positive factor is a square root, $|\mathbf{A}| = (\mathbf{A}^*\mathbf{A})^{1/2}$. So we start with square roots.

For positive numbers there is an exact formula for how far a square root moves:

$$\sqrt{p}-\sqrt{q}=\frac{p-q}{\sqrt{p}+\sqrt{q}},$$

the usual trick of multiplying by the conjugate. It shows that the square root is Lipschitz on $[c,\infty)$ with constant $1/(2\sqrt{c})$, and that no constant works near 0. For matrices the trick fails as written: with $\mathbf{S}=\mathbf{P}^{1/2}$ and $\mathbf{T}=\mathbf{Q}^{1/2}$, the product $(\mathbf{S}-\mathbf{T})(\mathbf{S}+\mathbf{T})=\mathbf{S}^2+\mathbf{S}\mathbf{T}-\mathbf{T}\mathbf{S}-\mathbf{T}^2$ is $\mathbf{P}-\mathbf{Q}$ only when $\mathbf{S}$ and $\mathbf{T}$ commute. But a different grouping of the same terms works with no commuting at all:

$$\mathbf{S}(\mathbf{S}-\mathbf{T})+(\mathbf{S}-\mathbf{T})\mathbf{T}=\mathbf{S}^2-\mathbf{S}\mathbf{T}+\mathbf{S}\mathbf{T}-\mathbf{T}^2=\mathbf{P}-\mathbf{Q}. \tag{20.10.2}$$

This says that $\mathbf{X}=\mathbf{S}-\mathbf{T}$ solves the Sylvester equation $\mathbf{S}\mathbf{X}+\mathbf{X}\mathbf{T}=\mathbf{P}-\mathbf{Q}$, whose coefficients are positive definite. That is exactly the situation of Corollary 20.7.4.

Theorem 20.10.2. Perturbation of the Square Root

Let $\mathbf{P},\mathbf{Q}\in M_n(\mathbb{C})$ be positive definite. Then

$$\|\mathbf{P}^{1/2}-\mathbf{Q}^{1/2}\| \;\le\; \frac{\|\mathbf{P}-\mathbf{Q}\|}{\lambda_n(\mathbf{P})^{1/2}+\lambda_n(\mathbf{Q})^{1/2}},$$

where $\|\cdot\|$ is either $\|\cdot\|_2$ or $\|\cdot\|_F$.

Proof. Put $\mathbf{S}=\mathbf{P}^{1/2}$ and $\mathbf{T}=\mathbf{Q}^{1/2}$, the positive square roots of Theorem 13.2.1. By that theorem the eigenvalues of $\mathbf{S}$ are the nonnegative square roots of those of $\mathbf{P}$, and $t\mapsto\sqrt{t}$ is increasing, so the smallest eigenvalue of $\mathbf{S}$ is $\alpha:=\lambda_n(\mathbf{P})^{1/2}$, which is positive because $\mathbf{P}\succ\mathbf{O}$. Every eigenvalue of the Hermitian matrix $\mathbf{S}-\alpha\mathbf{I}$ is therefore nonnegative, so $\mathbf{S}\succeq\alpha\mathbf{I}$ by Theorem 13.1.4 (b). In the same way $\mathbf{T}\succeq\beta\mathbf{I}$ with $\beta:=\lambda_n(\mathbf{Q})^{1/2}>0$.

By Eq. 20.10.2, $\mathbf{X}=\mathbf{S}-\mathbf{T}$ satisfies $\mathbf{S}\mathbf{X}+\mathbf{X}\mathbf{T}=\mathbf{P}-\mathbf{Q}$. Since $\alpha+\beta>0$, Corollary 20.7.4 applies to this equation, whose solution is therefore unique and equal to $\mathbf{X}$, and gives

$$\|\mathbf{S}-\mathbf{T}\|\le\frac{\|\mathbf{P}-\mathbf{Q}\|}{\alpha+\beta}$$

in both the spectral and the Frobenius norm. This proves the theorem. ∎

For $n=1$ this is the scalar identity, as an inequality that happens to be an equality. It is sharp for every n: with $\mathbf{P}=\operatorname{diag}(4,100)$ and $\mathbf{Q}=\operatorname{diag}(1,100)$, $\mathbf{P}^{1/2}-\mathbf{Q}^{1/2}=\operatorname{diag}(1,0)$ has both norms equal to 1, and the bound is $3/(2+1)=1$; padding both matrices with a further block $100\,\mathbf{I}_{n-2}$ changes neither side of the inequality, so the equality holds for every $n\ge2$. The difference sits exactly in the direction where both matrices are smallest, which is the worst place for it.

Remark. The proof used only $\alpha+\beta>0$, so the conclusion holds as soon as one of $\mathbf{P}$ and $\mathbf{Q}$ is positive definite and the other positive semidefinite.

Warning. **The square root is not Lipschitz on all positive semidefinite matrices.** Take $\mathbf{P} = \varepsilon \mathbf{I}_n$ with $0 < \varepsilon < 1$ and $\mathbf{Q} = \mathbf{O}$. Then $\|\mathbf{P} - \mathbf{Q}\|_2 = \varepsilon$, while $\mathbf{P}^{1/2} - \mathbf{Q}^{1/2} = \sqrt{\varepsilon}\,\mathbf{I}_n$ has spectral norm $\sqrt{\varepsilon}$, which is larger by the factor $1/\sqrt{\varepsilon}$. That factor is unbounded as $\varepsilon \to 0^+$, so no inequality $\|\mathbf{P}^{1/2} - \mathbf{Q}^{1/2}\|_2 \le C\|\mathbf{P} - \mathbf{Q}\|_2$ can hold for all $\mathbf{P}, \mathbf{Q} \succeq \mathbf{O}$. The denominator of the theorem is what records this. By the remark the theorem still applies, with denominator $\sqrt{\varepsilon} + 0$, and the bound $\varepsilon/\sqrt{\varepsilon} = \sqrt{\varepsilon}$ is attained. Exercise (C1: The square root is continuous everywhere) shows that the square root is nevertheless continuous everywhere, with a square-root modulus.

20.10.3 The positive polar factor

Recall Theorem 13.10.1: every $\mathbf{A} \in M_n(\mathbb{C})$ is a unitary matrix times the positive semidefinite $|\mathbf{A}| = (\mathbf{A}^*\mathbf{A})^{1/2}$, and when $\mathbf{A}$ is invertible both factors are unique. In this section we write the decomposition as

$$\mathbf{A} = \mathbf{U}\mathbf{P}, \qquad \mathbf{U} \text{ unitary}, \qquad \mathbf{P} = |\mathbf{A}|,$$

keeping the letter $\mathbf{W}$ for another use below. For invertible $\mathbf{A}$ the matrix $\mathbf{A}^*\mathbf{A}$ is positive definite, since $\mathbf{x}^*\mathbf{A}^*\mathbf{A}\mathbf{x} = \|\mathbf{A}\mathbf{x}\|^2 > 0$ for $\mathbf{x} \neq \mathbf{0}$. Its eigenvalues are the $\sigma_i(\mathbf{A})^2$ (Definition 13.9.1), so

$$\lambda_n(\mathbf{A}^*\mathbf{A}) = \sigma_n(\mathbf{A})^2 > 0, \qquad \mathbf{P} = |\mathbf{A}| \succeq \sigma_n(\mathbf{A})\,\mathbf{I},$$

the second because the eigenvalues of $|\mathbf{A}|$ are the square roots $\sigma_i(\mathbf{A})$ (Theorem 13.2.1) and Theorem 13.1.4 (b) applies to $|\mathbf{A}| - \sigma_n(\mathbf{A})\mathbf{I}$ as before.

Theorem 20.10.3. Perturbation of the Positive Polar Factor

Let $\mathbf{A}, \mathbf{B} \in M_n(\mathbb{C})$ be invertible. Then

$$\| \,|\mathbf{A}| - |\mathbf{B}|\, \| \;\le\; \frac{(\|\mathbf{A}\|_2 + \|\mathbf{B}\|_2)\,\|\mathbf{A} - \mathbf{B}\|}{\sigma_n(\mathbf{A}) + \sigma_n(\mathbf{B})},$$

where $\|\cdot\|$ is either $\|\cdot\|_2$ or $\|\cdot\|_F$.

Proof. Both $\mathbf{A}^*\mathbf{A}$ and $\mathbf{B}^*\mathbf{B}$ are positive definite, with $\lambda_n(\mathbf{A}^*\mathbf{A})^{1/2} = \sigma_n(\mathbf{A})$ and $\lambda_n(\mathbf{B}^*\mathbf{B})^{1/2} = \sigma_n(\mathbf{B})$, as just shown. So Theorem 20.10.2 gives

$$\| \,|\mathbf{A}| - |\mathbf{B}|\, \| \le \frac{\|\mathbf{A}^*\mathbf{A} - \mathbf{B}^*\mathbf{B}\|}{\sigma_n(\mathbf{A}) + \sigma_n(\mathbf{B})}.$$

Now $\mathbf{A}^*\mathbf{A} - \mathbf{B}^*\mathbf{B} = \mathbf{A}^*(\mathbf{A} - \mathbf{B}) + (\mathbf{A} - \mathbf{B})^*\mathbf{B}$, as expanding the right side shows. By Lemma 20.9.1 (b) and (a),

$$\|\mathbf{A}^*(\mathbf{A} - \mathbf{B})\| \le \|\mathbf{A}^*\|_2 \|\mathbf{A} - \mathbf{B}\| = \|\mathbf{A}\|_2 \|\mathbf{A} - \mathbf{B}\|,$$

and likewise $\|(\mathbf{A} - \mathbf{B})^*\mathbf{B}\| \le \|(\mathbf{A} - \mathbf{B})^*\|\,\|\mathbf{B}\|_2 = \|\mathbf{A} - \mathbf{B}\|\,\|\mathbf{B}\|_2$. The triangle inequality for the norm (Theorem 16.3.2 (b) for $\|\cdot\|_2$, Corollary 11.1.6 for the Frobenius norm, which is induced by an inner product) gives $\|\mathbf{A}^*\mathbf{A} - \mathbf{B}^*\mathbf{B}\| \le (\|\mathbf{A}\|_2 + \|\mathbf{B}\|_2)\|\mathbf{A} - \mathbf{B}\|$, and substituting proves the theorem.

■

When $\mathbf{B}$ is close to $\mathbf{A}$, the constant is about $2\|\mathbf{A}\|_2/(2\sigma_n(\mathbf{A})) = \kappa_2(\mathbf{A})$, the condition number of Chapter 16. So by this estimate $|\mathbf{A}|$ is as sensitive as $\mathbf{A}$ is ill-conditioned. It is a crude estimate, as the example below shows. Sharper ones are known, but this is the one the Sylvester bound gives with no extra work.

20.10.4 The unitary polar factor

For a nonzero complex number, the unitary polar factor is the phase $z/|z|$. If two numbers z and w have the same modulus r, their phases differ by exactly $|z-w|/r$. So the phase moves at the rate of the perturbation divided by the modulus, and it is undefined at $z=0$. The theorem below is the matrix form of this, with σ_n as the modulus.

Theorem 20.10.4. Perturbation of the Unitary Polar Factor

Let $\mathbf{A}, \mathbf{B} \in M_n(\mathbb{C})$ be invertible, with polar decompositions (Theorem 13.10.1) $\mathbf{A} = \mathbf{U}\mathbf{P}$ and $\mathbf{B} = \mathbf{V}\mathbf{Q}$, where $\mathbf{U}, \mathbf{V}$ are unitary, $\mathbf{P} = |\mathbf{A}|$, and $\mathbf{Q} = |\mathbf{B}|$. Then

$$\|\mathbf{U} - \mathbf{V}\| \;\le\; \frac{2\,\|\mathbf{A} - \mathbf{B}\|}{\sigma_n(\mathbf{A}) + \sigma_n(\mathbf{B})},$$

where $\|\cdot\|$ is either $\|\cdot\|_2$ or $\|\cdot\|_F$.

Idea. Measure $\mathbf{U}$ against $\mathbf{V}$ by the single unitary matrix $\mathbf{W} = \mathbf{V}^*\mathbf{U}$, which equals $\mathbf{I}$ exactly when $\mathbf{U} = \mathbf{V}$; then $\|\mathbf{U} - \mathbf{V}\| = \|\mathbf{W} - \mathbf{I}\|$. Now write $\mathbf{A} - \mathbf{B}$ in the two polar frames. Seen through $\mathbf{V}^*$ on the left it is $\mathbf{W}\mathbf{P} - \mathbf{Q}$. Its conjugate transpose, seen through $\mathbf{U}$ on the right, is $\mathbf{P} - \mathbf{Q}\mathbf{W}$. The difference of these two expressions is $(\mathbf{W} - \mathbf{I})\mathbf{P} + \mathbf{Q}(\mathbf{W} - \mathbf{I})$, a Sylvester equation for $\mathbf{W} - \mathbf{I}$ with positive definite coefficients $\mathbf{Q}$ and $\mathbf{P}$, and its right-hand side has norm at most $2\|\mathbf{A} - \mathbf{B}\|$.

Proof. Put $\mathbf{W} = \mathbf{V}^*\mathbf{U}$, which is unitary. Since $\mathbf{U}^*\mathbf{U} = \mathbf{V}^*\mathbf{V} = \mathbf{I}$, $\mathbf{P}^* = \mathbf{P}$, and $\mathbf{Q}^* = \mathbf{Q}$,

$$\begin{aligned} \mathbf{V}^*(\mathbf{A} - \mathbf{B}) &= \mathbf{V}^*\mathbf{U}\mathbf{P} - \mathbf{V}^*\mathbf{V}\mathbf{Q} = \mathbf{W}\mathbf{P} - \mathbf{Q}, \\ (\mathbf{A} - \mathbf{B})^*\mathbf{U} &= \mathbf{P}\mathbf{U}^*\mathbf{U} - \mathbf{Q}\mathbf{V}^*\mathbf{U} = \mathbf{P} - \mathbf{Q}\mathbf{W}. \end{aligned}$$

Subtracting the second identity from the first,

$$\begin{aligned} \mathbf{V}^*(\mathbf{A} - \mathbf{B}) - (\mathbf{A} - \mathbf{B})^*\mathbf{U} &= \mathbf{W}\mathbf{P} - \mathbf{P} + \mathbf{Q}\mathbf{W} - \mathbf{Q} \\ &= \mathbf{Q}(\mathbf{W} - \mathbf{I}) + (\mathbf{W} - \mathbf{I})\mathbf{P}. \end{aligned} \tag{20.10.3}$$

So $\mathbf{X} = \mathbf{W} - \mathbf{I}$ solves $\mathbf{Q}\mathbf{X} + \mathbf{X}\mathbf{P} = \mathbf{Y}$, with $\mathbf{Y}$ the left side of Eq. 20.10.3. As shown before Theorem 20.10.3, $\mathbf{Q} \succeq \sigma_n(\mathbf{B})\mathbf{I}$ and $\mathbf{P} \succeq \sigma_n(\mathbf{A})\mathbf{I}$, and $\sigma_n(\mathbf{A}) + \sigma_n(\mathbf{B}) > 0$. So Corollary 20.7.4, applied with $\mathbf{Q}$ and $\mathbf{P}$ as its two coefficients, gives

$$\|\mathbf{W} - \mathbf{I}\| \le \frac{\|\mathbf{Y}\|}{\sigma_n(\mathbf{A}) + \sigma_n(\mathbf{B})}$$

in both norms. By Lemma 20.9.1 (c) and (a), $\|\mathbf{V}^*(\mathbf{A} - \mathbf{B})\| = \|\mathbf{A} - \mathbf{B}\|$ and $\|(\mathbf{A} - \mathbf{B})^*\mathbf{U}\| = \|(\mathbf{A} - \mathbf{B})^*\| = \|\mathbf{A} - \mathbf{B}\|$, so $\|\mathbf{Y}\| \le 2\|\mathbf{A} - \mathbf{B}\|$ by the triangle inequality in either norm

(Theorem 16.3.2 (b), Corollary 11.1.6). Finally $\mathbf{U}-\mathbf{V}=\mathbf{V}(\mathbf{V}^*\mathbf{U}-\mathbf{I})=\mathbf{V}(\mathbf{W}-\mathbf{I})$, so $\|\mathbf{U}-\mathbf{V}\|=\|\mathbf{W}-\mathbf{I}\|$ by Lemma 20.9.1 (c). Combining the three displays proves the theorem.

■

Notice the difference from Theorem 20.10.3. There the constant was a condition number, but here only σ_n appears. The unitary factor feels how close $\mathbf{A}$ is to the singular matrices, which is $\sigma_n(\mathbf{A})$ by Corollary 17.9.7, and nothing else.

Quick check. For $n=1$, what does the theorem say about two nonzero complex numbers z and w? Test it on $z=1$, $w=-1$.

Solution. A nonzero $z\in\mathbb{C}=M_1(\mathbb{C})$ has polar decomposition $z=\frac{z}{|z|}\cdot|z|$, with $\sigma_1(z)=|z|$, and both norms are the modulus. So the theorem says

$$\left|\frac{z}{|z|}-\frac{w}{|w|}\right|\le\frac{2|z-w|}{|z|+|w|}.$$

For $z=1$, $w=-1$ both sides equal 2. So the constant 2 in the theorem cannot be lowered, even for $n=1$.

Example. The polar factors of two nearby matrices

Let $\mathbf{A}=\left(\begin{smallmatrix}3&0\\4&5\end{smallmatrix}\right)$, whose polar decomposition Example (A polar decomposition) found to be $\mathbf{U}=\frac{1}{\sqrt5}\left(\begin{smallmatrix}2&-1\\1&2\end{smallmatrix}\right)$, $|\mathbf{A}|=\sqrt5\left(\begin{smallmatrix}2&1\\1&2\end{smallmatrix}\right)$, with $\sigma(\mathbf{A})=(3\sqrt5,\sqrt5)$. Let $\mathbf{B}=\left(\begin{smallmatrix}3&0\\6&5\end{smallmatrix}\right)$. Find the polar decomposition of $\mathbf{B}$, compute $\|\mathbf{U}-\mathbf{V}\|$, and compare it with Theorem 20.10.4.

Solution. Finding $\mathbf{V}$. For a real 2×2 matrix of positive determinant, look for $\mathbf{V}$ among the rotations $\mathbf{R}_\varphi=\left(\begin{smallmatrix}c&-s\\s&c\end{smallmatrix}\right)$, with $c=\cos\varphi$, $s=\sin\varphi$. For $\mathbf{B}=(b_{ij})$, the off-diagonal entries of $\mathbf{R}_\varphi^{\top}\mathbf{B}$ are $cb_{12}+sb_{22}$ and $-sb_{11}+cb_{21}$, and they agree exactly when $(b_{11}+b_{22})s=(b_{21}-b_{12})c$. For our $\mathbf{B}$ this reads $8s=6c$, so $(c,s)=(\frac45,\frac35)$, and

$$\mathbf{V}=\tfrac15\begin{pmatrix}4&-3\\3&4\end{pmatrix},$$

$$\mathbf{V}^{\top}\mathbf{B}=\tfrac15\begin{pmatrix}4&3\\-3&4\end{pmatrix}\begin{pmatrix}3&0\\6&5\end{pmatrix}=\begin{pmatrix}6&3\\3&4\end{pmatrix}.$$

This symmetric matrix has trace 10 and determinant 15, so both its eigenvalues, $5\pm\sqrt{10}$, are positive, and it is positive definite. Thus $\mathbf{B}=\mathbf{V}(\mathbf{V}^{\top}\mathbf{B})$ is a unitary matrix times a positive definite one. Squaring gives $(\mathbf{V}^{\top}\mathbf{B})^2=\mathbf{B}^{\top}\mathbf{V}\mathbf{V}^{\top}\mathbf{B}=\mathbf{B}^{\top}\mathbf{B}$, so $\mathbf{V}^{\top}\mathbf{B}=|\mathbf{B}|$ by uniqueness of the positive square root (Theorem 13.2.1). Since $\mathbf{B}$ is invertible, Theorem 13.10.1 now gives $\mathbf{V}$ as its unitary polar factor. Its singular values are the eigenvalues of $|\mathbf{B}|$, $\sigma(\mathbf{B})=(5+\sqrt{10},5-\sqrt{10})\approx(8.162,1.838)$. The same recipe applied to $\mathbf{A}$ gives $8s=4c$, which recovers $\mathbf{U}$.

The distance. $\mathbf{U}$ and $\mathbf{V}$ are the rotations through angles $\varphi_{\mathbf{A}}$ and $\varphi_{\mathbf{B}}$ with

$$\cos(\varphi_{\mathbf{B}} - \varphi_{\mathbf{A}}) = \tfrac{2}{\sqrt{5}} \cdot \tfrac{4}{5} + \tfrac{1}{\sqrt{5}} \cdot \tfrac{3}{5} = \tfrac{11}{5\sqrt{5}}.$$

For two rotations, $(\mathbf{U}-\mathbf{V})^{\top}(\mathbf{U}-\mathbf{V}) = 2\mathbf{I} - (\mathbf{U}^{\top}\mathbf{V} + \mathbf{V}^{\top}\mathbf{U}) = (2 - 2\cos(\varphi_{\mathbf{B}} - \varphi_{\mathbf{A}}))\mathbf{I}$, so both singular values of $\mathbf{U}-\mathbf{V}$ equal $\sqrt{2 - 22/(5\sqrt{5})}$ and

$$\|\mathbf{U}-\mathbf{V}\|_2 \approx 0.180, \qquad \|\mathbf{U}-\mathbf{V}\|_F \approx 0.254.$$

The bound. $\mathbf{A}-\mathbf{B} = \left(\begin{smallmatrix} 0 & 0 \\ -2 & 0 \end{smallmatrix}\right)$, so $\|\mathbf{A}-\mathbf{B}\|_2 = \|\mathbf{A}-\mathbf{B}\|_F = 2$, and

$$\frac{2 \cdot 2}{\sigma_2(\mathbf{A}) + \sigma_2(\mathbf{B})} = \frac{4}{\sqrt{5} + 5 - \sqrt{10}} \approx 0.982$$

in both norms. The theorem holds, with a factor of about 5 to spare in the spectral norm and 4 in the Frobenius norm. For the positive factors, Theorem 20.10.3 gives

$$\| \, |\mathbf{A}| - |\mathbf{B}| \, \| \le \frac{(3\sqrt{5} + 5 + \sqrt{10}) \cdot 2}{\sqrt{5} + 5 - \sqrt{10}} \approx 7.30,$$

while $|\mathbf{A}| - |\mathbf{B}| = \left(\begin{smallmatrix} 2\sqrt{5}-6 & \sqrt{5}-3 \\ \sqrt{5}-3 & 2\sqrt{5}-4 \end{smallmatrix}\right)$ has spectral norm about 1.79 and Frobenius norm about 1.93. That bound is correct but crude, as promised.

Warning. **At a singular matrix the unitary factor jumps, and no bound of this kind survives.** For $0 < \varepsilon \le 1$ let

$$\mathbf{A}_\varepsilon = \operatorname{diag}(1, \varepsilon) = \mathbf{I} \operatorname{diag}(1, \varepsilon),$$
$$\mathbf{B}_\varepsilon = \operatorname{diag}(1, -\varepsilon) = \operatorname{diag}(1, -1) \operatorname{diag}(1, \varepsilon).$$

These are polar decompositions: the unitary factors are $\mathbf{U} = \mathbf{I}$ and $\mathbf{V} = \operatorname{diag}(1, -1)$, and both positive factors are $\operatorname{diag}(1, \varepsilon)$. As $\varepsilon \to 0^+$, both matrices tend to the singular $\operatorname{diag}(1, 0)$, and $\|\mathbf{A}_\varepsilon - \mathbf{B}_\varepsilon\| = 2\varepsilon \to 0$ in both norms. Yet $\mathbf{U} - \mathbf{V} = \operatorname{diag}(0, 2)$ has spectral and Frobenius norm 2 for every ε. So the unitary factor has no limit at $\operatorname{diag}(1, 0)$, and it cannot be chosen continuously there. This matches the nonuniqueness in Theorem 13.10.1: both $\mathbf{I}$ and $\operatorname{diag}(1, -1)$ are unitary polar factors of $\operatorname{diag}(1, 0)$. The theorem, with $\sigma_2(\mathbf{A}_\varepsilon) = \sigma_2(\mathbf{B}_\varepsilon) = \varepsilon$, gives the bound $(2 \cdot 2\varepsilon)/(2\varepsilon) = 2$, which is attained. Padding both matrices with an identity block, $\mathbf{A}_\varepsilon \oplus \mathbf{I}_{n-2}$ and $\mathbf{B}_\varepsilon \oplus \mathbf{I}_{n-2}$, gives the same equality for every $n \ge 2$: the constant 2 cannot be improved, and the denominator cannot be dropped.

Corollary 13.10.2 identified the unitary polar factor as the unitary matrix nearest to $\mathbf{A}$ in the Frobenius norm. Theorem 20.10.4 adds that for invertible $\mathbf{A}$ this nearest unitary matrix moves continuously with $\mathbf{A}$, at a rate no worse than $2/(\sigma_n(\mathbf{A}) + \sigma_n(\mathbf{B}))$. A matrix that has drifted from unitarity can therefore be put back reliably, as long as it stays well away from the singular matrices.

20.10.5 Exercises

20.10.5.1 A. Check your understanding

Exercise. A1

(a) State Mirsky's theorem. In its proof by the Hermitian dilation of Proposition 17.9.3, where does each of the two factors of 2 come from?

(b) True or false: $|\sigma_i(\mathbf{A}) - \sigma_i(\mathbf{B})| \le \|\mathbf{A} - \mathbf{B}\|_F$ for every i. Justify.

(c) True or false: the unitary polar factor is a continuous function of $\mathbf{A}$ on the invertible matrices, and it extends continuously to all of $M_n(\mathbb{C})$. Justify.

(d) Which identity turns the perturbation of a square root into a Sylvester equation, and which result then bounds it?

(e) Why does Theorem 20.10.4 hold in both the spectral and the Frobenius norm?

Solution.

(a) Theorem 20.10.1: $\sum_{i=1}^{p} (\sigma_i(\mathbf{A}) - \sigma_i(\mathbf{B}))^2 \le \|\mathbf{A} - \mathbf{B}\|_F^2$. On the eigenvalue side, each difference $\sigma_i(\mathbf{A}) - \sigma_i(\mathbf{B})$ occurs twice in the sorted spectra of the dilations, once as σ_i near the top and once as $-\sigma_i$ near the bottom. On the norm side, each entry of $\mathbf{A} - \mathbf{B}$ occurs twice in $\mathcal{H}(\mathbf{A} - \mathbf{B})$, once in $\mathbf{A} - \mathbf{B}$ and once, conjugated, in $(\mathbf{A} - \mathbf{B})^*$ (Eq. 20.10.1).

(b) True. Every term of the sum in Mirsky's theorem is nonnegative, so each one is at most the sum, which is at most $\|\mathbf{A} - \mathbf{B}\|_F^2$. (It also follows from Corollary 17.9.5 and $\|\cdot\|_2 \le \|\cdot\|_F$, Proposition 16.3.4.)

(c) False as a whole. The first half is true: by Theorem 20.10.4, $\|\mathbf{U} - \mathbf{V}\|_2 \le 2\|\mathbf{A} - \mathbf{B}\|_2/(\sigma_n(\mathbf{A}) + \sigma_n(\mathbf{B}))$, which tends to 0 as $\mathbf{B} \to \mathbf{A}$ for fixed invertible $\mathbf{A}$, because $\sigma_n(\mathbf{B}) \ge 0$. The second half is false: in the last warning of the section, $\mathbf{A}_\varepsilon$ and $\mathbf{B}_\varepsilon$ both tend to $\operatorname{diag}(1, 0)$ while their unitary factors stay at distance 2, so no value at $\operatorname{diag}(1, 0)$ makes the factor continuous there.

(d) $\mathbf{S}(\mathbf{S} - \mathbf{T}) + (\mathbf{S} - \mathbf{T})\mathbf{T} = \mathbf{S}^2 - \mathbf{T}^2$ (Eq. 20.10.2), which exhibits $\mathbf{S} - \mathbf{T}$ as the solution of $\mathbf{SX} + \mathbf{XT} = \mathbf{P} - \mathbf{Q}$. Section 7's Corollary 20.7.4 bounds that solution.

(e) Each step holds in both norms: Corollary 20.7.4 is stated for both, and so are the facts of Lemma 20.9.1 used to remove the unitary factors and conjugate transposes.

20.10.5.2 B. Practice

Exercise. B1: Mirsky by hand

Let $\mathbf{A} = \left(\begin{smallmatrix} 3 & 0 \\ 4 & 5 \end{smallmatrix}\right)$, with $\sigma(\mathbf{A}) = (3\sqrt{5}, \sqrt{5})$ from Example (A polar decomposition), and $\mathbf{B} = \left(\begin{smallmatrix} 3 & 0 \\ 4 & 3 \end{smallmatrix}\right)$.

(a) Show that $\sigma(\mathbf{B}) = (\sqrt{13} + 2, \sqrt{13} - 2)$.

(b) Verify Theorem 20.10.1 for this pair.

(c) What bound on $\sum_i (\sigma_i(\mathbf{A}) - \sigma_i(\mathbf{B}))^2$ does Corollary 17.9.5 give? Hence say what Mirsky's theorem gains here.

Solution.

(a) $\mathbf{B}^\top\mathbf{B} = \left(\begin{smallmatrix} 25 & 12 \\ 12 & 9 \end{smallmatrix}\right)$ has trace 34 and determinant $225 - 144 = 81$, so its eigenvalues are $17 \pm \sqrt{289 - 81} = 17 \pm 4\sqrt{13}$. Since $(\sqrt{13} \pm 2)^2 = 17 \pm 4\sqrt{13}$ and $\sqrt{13} - 2 > 0$, the

singular values are $\sqrt{13} \pm 2 \approx 5.606, 1.606$.

(b) $\mathbf{A} - \mathbf{B} = \operatorname{diag}(0, 2)$, so $\|\mathbf{A} - \mathbf{B}\|_F^2 = 4$. The left side is

$$\begin{aligned}(3\sqrt{5} - \sqrt{13} - 2)^2 + (\sqrt{5} - \sqrt{13} + 2)^2 \\ = 84 - 8\sqrt{65} - 8\sqrt{5} \approx 1.613,\end{aligned}$$

expanding each square (the $\sqrt{13}$ terms cancel). Indeed $1.613 \le 4$.

(c) $\|\mathbf{A} - \mathbf{B}\|_2 = 2$, so each $|\sigma_i(\mathbf{A}) - \sigma_i(\mathbf{B})| \le 2$, and the sum of the two squares is at most 8. Mirsky's theorem gives 4, half as much. The true value is about 1.613: the differences are about 1.103 and 0.631.

Exercise. B2: A square root that does not commute

Let $\mathbf{S} = \left(\begin{smallmatrix} 3 & 2 \\ 2 & 3 \end{smallmatrix}\right)$, $\mathbf{T} = \left(\begin{smallmatrix} 3 & 2 \\ 2 & 4 \end{smallmatrix}\right)$, $\mathbf{P} = \mathbf{S}^2$, and $\mathbf{Q} = \mathbf{T}^2$.

(a) Compute $\mathbf{P}$ and $\mathbf{Q}$, and explain why $\mathbf{S} = \mathbf{P}^{1/2}$ and $\mathbf{T} = \mathbf{Q}^{1/2}$.

(b) Find $\lambda_2(\mathbf{P})$ and $\lambda_2(\mathbf{Q})^{1/2}$.

(c) Compute $\|\mathbf{P} - \mathbf{Q}\|_2$ and $\|\mathbf{P} - \mathbf{Q}\|_F$, and hence the two bounds of Theorem 20.10.2. Compare them with $\|\mathbf{S} - \mathbf{T}\|$.

Solution.

(a) Squaring, $\mathbf{P} = \left(\begin{smallmatrix} 13 & 12 \\ 12 & 13 \end{smallmatrix}\right)$ and $\mathbf{Q} = \left(\begin{smallmatrix} 13 & 14 \\ 14 & 20 \end{smallmatrix}\right)$. $\mathbf{S}$ is symmetric with eigenvalues 5 and 1 (eigenvectors $(1, \pm 1)$), and $\mathbf{T}$ is symmetric with trace 7 and determinant 8, so both eigenvalues of each are positive, and $\mathbf{S}, \mathbf{T} \succ \mathbf{O}$. By the uniqueness in Theorem 13.2.1, $\mathbf{S} = \mathbf{P}^{1/2}$ and $\mathbf{T} = \mathbf{Q}^{1/2}$.

(b) The eigenvalues of $\mathbf{P}$ are 25 and 1, so $\lambda_2(\mathbf{P}) = 1$. The eigenvalues of $\mathbf{T}$ are $\frac{7 \pm \sqrt{17}}{2}$, and by Theorem 13.2.1 those of $\mathbf{Q}$ are their squares, so $\lambda_2(\mathbf{Q})^{1/2} = \lambda_2(\mathbf{T}) = \frac{7 - \sqrt{17}}{2} \approx 1.438$.

(c) $\mathbf{P} - \mathbf{Q} = \left(\begin{smallmatrix} 0 & -2 \\ -2 & -7 \end{smallmatrix}\right)$ is symmetric with characteristic polynomial $t^2 + 7t - 4$, so its eigenvalues are $\frac{-7 \pm \sqrt{65}}{2}$, and by Lemma 17.3.5 $\|\mathbf{P} - \mathbf{Q}\|_2 = \frac{7 + \sqrt{65}}{2} \approx 7.531$. Also $\|\mathbf{P} - \mathbf{Q}\|_F = \sqrt{4 + 4 + 49} = \sqrt{57} \approx 7.550$. The denominator is $1 + 1.438 = 2.438$, so the bounds are about 3.089 (spectral) and 3.096 (Frobenius). Since $\mathbf{S} - \mathbf{T} = \operatorname{diag}(0, -1)$, both of its norms equal 1. Hence the theorem holds here with a factor of about 3 to spare. It is not an equality because $\mathbf{S} - \mathbf{T}$ does not sit along the eigenvector of $\mathbf{P}$ and $\mathbf{Q}$ where both are smallest.

Exercise. B3: A nearly sharp instance

Let $\mathbf{A} = 2\mathbf{I}_2$ and $\mathbf{B} = \left(\begin{smallmatrix} 2 & -1 \\ 1 & 2 \end{smallmatrix}\right)$.

(a) Find the polar decompositions of $\mathbf{A}$ and $\mathbf{B}$.

(b) Compute $\|\mathbf{U} - \mathbf{V}\|_2$ and $\|\mathbf{U} - \mathbf{V}\|_F$, and compare them with the bounds of Theorem 20.10.4.

(c) Compute $\| \, |\mathbf{A}| - |\mathbf{B}| \, \|_2$ and the bound of Theorem 20.10.3.

Solution.

(a) $\mathbf{A} = \mathbf{I} \cdot 2\mathbf{I}$, with $\mathbf{U} = \mathbf{I}$ and $|\mathbf{A}| = 2\mathbf{I} \succ \mathbf{O}$. The columns of $\mathbf{B}$ are orthogonal, each of length $\sqrt{5}$, so $\mathbf{V} = \frac{1}{\sqrt{5}}\mathbf{B}$ is orthogonal and $\mathbf{B} = \mathbf{V} \cdot \sqrt{5}\,\mathbf{I}$, with $\sqrt{5}\,\mathbf{I} \succ \mathbf{O}$. Both matrices are invertible, so these are their polar decompositions (Theorem 13.10.1), and $\sigma_2(\mathbf{A}) = 2$, $\sigma_2(\mathbf{B}) = \sqrt{5}$.

(b) $\mathbf{V}$ is the rotation with $\cos\varphi = \frac{2}{\sqrt{5}}$, so, as in Example (The polar factors of two nearby matrices), $(\mathbf{U}-\mathbf{V})^{\top}(\mathbf{U}-\mathbf{V}) = (2 - 2\cos\varphi)\mathbf{I} = (2 - \frac{4}{\sqrt{5}})\mathbf{I}$. Hence $\|\mathbf{U}-\mathbf{V}\|_2 = \sqrt{2 - 4/\sqrt{5}} \approx 0.4595$ and $\|\mathbf{U}-\mathbf{V}\|_F = \sqrt{2}\,\|\mathbf{U}-\mathbf{V}\|_2 \approx 0.6498$. Next, $\mathbf{A}-\mathbf{B} = \left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)$ is orthogonal, so $\|\mathbf{A}-\mathbf{B}\|_2 = 1$ and $\|\mathbf{A}-\mathbf{B}\|_F = \sqrt{2}$. The bounds are

$$\frac{2 \cdot 1}{2+\sqrt{5}} \approx 0.4721, \qquad \frac{2\sqrt{2}}{2+\sqrt{5}} \approx 0.6677.$$

Both hold, and each is within 3% of the true value.

(c) $|\mathbf{A}| - |\mathbf{B}| = (2-\sqrt{5})\mathbf{I}$, of spectral norm $\sqrt{5}-2 \approx 0.236$. The bound is $(\|\mathbf{A}\|_2 + \|\mathbf{B}\|_2)\|\mathbf{A}-\mathbf{B}\|_2/(2+\sqrt{5}) = (2+\sqrt{5}) \cdot 1/(2+\sqrt{5}) = 1$, a factor of about 4 larger.

20.10.5.3 C. Going deeper

Exercise. C1: The square root is continuous everywhere

Let $\mathbf{S}, \mathbf{T} \in M_n(\mathbb{C})$ be positive semidefinite.

(a) Let λ be an eigenvalue of the Hermitian matrix $\mathbf{S}-\mathbf{T}$, with unit eigenvector $\mathbf{x}$. Prove that $\mathbf{S}^2 - \mathbf{T}^2 = (\mathbf{S}-\mathbf{T})^2 + (\mathbf{S}-\mathbf{T})\mathbf{T} + \mathbf{T}(\mathbf{S}-\mathbf{T})$, and that $\mathbf{x}^*(\mathbf{S}^2-\mathbf{T}^2)\mathbf{x} = \lambda^2 + 2\lambda\,\mathbf{x}^*\mathbf{T}\mathbf{x}$.

(b) Deduce that if $\lambda \geq 0$ then $\lambda^2 \leq \|\mathbf{S}^2 - \mathbf{T}^2\|_2$.

(c) Prove that $\|\mathbf{P}^{1/2} - \mathbf{Q}^{1/2}\|_2 \leq \|\mathbf{P}-\mathbf{Q}\|_2^{1/2}$ for all positive semidefinite $\mathbf{P}, \mathbf{Q} \in M_n(\mathbb{C})$, and show that the exponent $\frac{1}{2}$ cannot be replaced by anything larger.

Hint: for (c), apply (b) to the pair $(\mathbf{S}, \mathbf{T})$ or to the pair $(\mathbf{T}, \mathbf{S})$, according to the sign of the eigenvalue of $\mathbf{S}-\mathbf{T}$ of largest modulus.

Solution.

(a) Expanding,

$$\begin{aligned}(\mathbf{S}-\mathbf{T})^2 + (\mathbf{S}-\mathbf{T})\mathbf{T} + \mathbf{T}(\mathbf{S}-\mathbf{T}) \\ = \mathbf{S}^2 - \mathbf{S}\mathbf{T} - \mathbf{T}\mathbf{S} + \mathbf{T}^2 + \mathbf{S}\mathbf{T} - \mathbf{T}^2 + \mathbf{T}\mathbf{S} - \mathbf{T}^2 = \mathbf{S}^2 - \mathbf{T}^2.\end{aligned}$$

Since $(\mathbf{S}-\mathbf{T})\mathbf{x} = \lambda\mathbf{x}$ with λ real (Theorem 12.2.1) and $\mathbf{S}-\mathbf{T}$ Hermitian, also $\mathbf{x}^*(\mathbf{S}-\mathbf{T}) = \lambda\mathbf{x}^*$. Hence $\mathbf{x}^*(\mathbf{S}-\mathbf{T})^2\mathbf{x} = \lambda^2\mathbf{x}^*\mathbf{x} = \lambda^2$, $\mathbf{x}^*(\mathbf{S}-\mathbf{T})\mathbf{T}\mathbf{x} = \lambda\,\mathbf{x}^*\mathbf{T}\mathbf{x}$, and $\mathbf{x}^*\mathbf{T}(\mathbf{S}-\mathbf{T})\mathbf{x} = \lambda\,\mathbf{x}^*\mathbf{T}\mathbf{x}$. Adding gives the claim.

(b) $\mathbf{x}^*\mathbf{T}\mathbf{x} \geq 0$ because $\mathbf{T} \succeq \mathbf{O}$, so if $\lambda \geq 0$, (a) gives $\lambda^2 \leq \mathbf{x}^*(\mathbf{S}^2-\mathbf{T}^2)\mathbf{x}$. By Theorem 11.1.5 and Theorem 16.3.2 (a), $\mathbf{x}^*\mathbf{M}\mathbf{x} \leq \|\mathbf{M}\mathbf{x}\|\|\mathbf{x}\| \leq \|\mathbf{M}\|_2$ for any $\mathbf{M}$ and unit $\mathbf{x}$. So $\lambda^2 \leq \|\mathbf{S}^2 - \mathbf{T}^2\|_2$.

(c) Put $\mathbf{S} = \mathbf{P}^{1/2}$ and $\mathbf{T} = \mathbf{Q}^{1/2}$, so $\mathbf{S}^2 - \mathbf{T}^2 = \mathbf{P} - \mathbf{Q}$. By Lemma 17.3.5, $\|\mathbf{S}-\mathbf{T}\|_2 = |\lambda|$ for some eigenvalue λ of $\mathbf{S}-\mathbf{T}$. If $\lambda \geq 0$, (b) gives $\lambda^2 \leq \|\mathbf{P}-\mathbf{Q}\|_2$. If $\lambda < 0$,

then $-\lambda > 0$ is an eigenvalue of $\mathbf{T} - \mathbf{S}$, and (b) applied to the pair $(\mathbf{T}, \mathbf{S})$ gives $\lambda^2 \le \|\mathbf{T}^2 - \mathbf{S}^2\|_2 = \|\mathbf{P} - \mathbf{Q}\|_2$. In both cases $\|\mathbf{P}^{1/2} - \mathbf{Q}^{1/2}\|_2 = |\lambda| \le \|\mathbf{P} - \mathbf{Q}\|_2^{1/2}$.

For the exponent, take $\mathbf{P} = \varepsilon\mathbf{I}$ and $\mathbf{Q} = \mathbf{O}$ with $0 < \varepsilon < 1$, as in the warning after Theorem 20.10.2. Then $\|\mathbf{P}^{1/2} - \mathbf{Q}^{1/2}\|_2 = \varepsilon^{1/2}$ and $\|\mathbf{P} - \mathbf{Q}\|_2 = \varepsilon$. If $\varepsilon^{1/2} \le C\varepsilon^r$ held for all small ε with some $r > \frac{1}{2}$, then $C \ge \varepsilon^{1/2-r} \to \infty$ as $\varepsilon \to 0^+$, which is impossible. So the square root, which is not Lipschitz near singular matrices, is still continuous on all positive semidefinite matrices, with modulus $t \mapsto t^{1/2}$.

Exercise. C2: Both Eckart–Young lower bounds from perturbation theory

Let $\mathbf{A} \in M_{m\times n}(\mathbb{C})$, let $0 \le k < p = \min(m, n)$, and let $\mathbf{B} \in M_{m\times n}(\mathbb{C})$ have rank at most k.

(a) Explain why $\sigma_i(\mathbf{B}) = 0$ for $i > k$.

(b) Prove, using Theorem 20.10.1, that $\|\mathbf{A} - \mathbf{B}\|_F^2 \ge \sum_{i=k+1}^{p} \sigma_i(\mathbf{A})^2$.

(c) Prove, using Corollary 17.9.5, that $\|\mathbf{A} - \mathbf{B}\|_2 \ge \sigma_{k+1}(\mathbf{A})$.

(d) Compare with Theorem 13.11.3 and Theorem 17.9.6.

Solution.

(a) By Theorem 13.9.3 the number of nonzero singular values of $\mathbf{B}$ is $\operatorname{rank}\mathbf{B} \le k$, and they come first in the decreasing list, so $\sigma_i(\mathbf{B}) = 0$ for $i > k$.

(b) By Theorem 20.10.1, and dropping the nonnegative terms with $i \le k$,

$$\begin{aligned}\|\mathbf{A} - \mathbf{B}\|_F^2 &\ge \sum_{i=1}^{p} \big(\sigma_i(\mathbf{A}) - \sigma_i(\mathbf{B})\big)^2 \\ &\ge \sum_{i=k+1}^{p} \big(\sigma_i(\mathbf{A}) - 0\big)^2,\end{aligned}$$

using (a) in the last step.

(c) By Corollary 17.9.5 with $i = k + 1 \le p$ and $\mathbf{E} = \mathbf{B} - \mathbf{A}$,

$$\sigma_{k+1}(\mathbf{A}) = |\sigma_{k+1}(\mathbf{A}) - \sigma_{k+1}(\mathbf{B})| \le \|\mathbf{B} - \mathbf{A}\|_2 = \|\mathbf{A} - \mathbf{B}\|_2,$$

using (a) and $\sigma_{k+1}(\mathbf{A}) \ge 0$.

(d) These are exactly the lower-bound halves of the Frobenius and the spectral Eckart–Young theorems. Both theorems also show that the bounds are attained, by the truncation $\mathbf{A}_k$, which has singular values $\sigma_1(\mathbf{A}), \dots, \sigma_k(\mathbf{A}), 0, \dots, 0$, so that it makes every term with $i \le k$ vanish. So "the best rank-k approximation" is a perturbation theorem read backwards: no matrix of rank k can be closer to $\mathbf{A}$ than the singular values of $\mathbf{A}$ allow, because singular values cannot move further than the perturbation.

Exercise. C3: The nearest unitary to an almost-unitary matrix

Let $\mathbf{E} \in M_n(\mathbb{C})$ with $\|\mathbf{E}\|_2 < 1$, let $\mathbf{A} = \mathbf{I} + \mathbf{E}$, and let $\mathbf{U}$ be the unitary polar factor of $\mathbf{A}$.

(a) Prove that $\sigma_n(\mathbf{A}) \ge 1 - \|\mathbf{E}\|_2 > 0$, so that $\mathbf{A}$ is invertible.

(b) Prove that $\|\mathbf{U}-\mathbf{I}\| \le 2\|\mathbf{E}\|/(2-\|\mathbf{E}\|_2)$ in both the spectral and the Frobenius norm.

(c) Deduce that if $\|\mathbf{E}\|_2 \le \frac{1}{2}$ then $\|\mathbf{U}-\mathbf{I}\|_F \le \frac{4}{3}\|\mathbf{E}\|_F$. Compare with the bound $\|\mathbf{U}-\mathbf{I}\|_F \le 2\|\mathbf{E}\|_F$ that Corollary 13.10.2 and the triangle inequality give with no computation.

Solution.

(a) The identity has all singular values equal to 1. By Corollary 17.9.5 with $i = n$, $|\sigma_n(\mathbf{I}+\mathbf{E}) - 1| \le \|\mathbf{E}\|_2$, so $\sigma_n(\mathbf{A}) \ge 1 - \|\mathbf{E}\|_2 > 0$. An $n \times n$ matrix with n nonzero singular values has rank n (Theorem 13.9.3), so $\mathbf{A}$ is invertible.

(b) Apply Theorem 20.10.4 to $\mathbf{A}$ and $\mathbf{B} = \mathbf{I}$, whose polar decomposition is $\mathbf{I} = \mathbf{I} \cdot \mathbf{I}$, with $\mathbf{V} = \mathbf{I}$ and $\sigma_n(\mathbf{I}) = 1$. Since $\mathbf{A} - \mathbf{I} = \mathbf{E}$,

$$\|\mathbf{U}-\mathbf{I}\| \le \frac{2\|\mathbf{E}\|}{\sigma_n(\mathbf{A})+1} \le \frac{2\|\mathbf{E}\|}{2-\|\mathbf{E}\|_2},$$

using (a) to bound the denominator from below. This holds in both norms because the theorem does.

(c) If $\|\mathbf{E}\|_2 \le \frac{1}{2}$ then $2-\|\mathbf{E}\|_2 \ge \frac{3}{2}$, and (b) in the Frobenius norm gives $\|\mathbf{U}-\mathbf{I}\|_F \le \frac{4}{3}\|\mathbf{E}\|_F$.

For comparison, Corollary 13.10.2 says that $\mathbf{U}$ is at least as close to $\mathbf{A}$ as any unitary matrix is, and $\mathbf{I}$ is one, so $\|\mathbf{A}-\mathbf{U}\|_F \le \|\mathbf{A}-\mathbf{I}\|_F = \|\mathbf{E}\|_F$. The triangle inequality (Corollary 11.1.6) then gives

$$\|\mathbf{U}-\mathbf{I}\|_F \le \|\mathbf{U}-\mathbf{A}\|_F + \|\mathbf{A}-\mathbf{I}\|_F \le 2\|\mathbf{E}\|_F.$$

So the soft argument already locates $\mathbf{U}$ within $2\|\mathbf{E}\|_F$ of $\mathbf{I}$. Part (b) improves the constant to $\frac{4}{3}$ when $\|\mathbf{E}\|_2 \le \frac{1}{2}$, and to $2/(2-\|\mathbf{E}\|_2)$, which tends to 1, as $\mathbf{E} \to \mathbf{O}$. It also gives the spectral-norm statement, about which Corollary 13.10.2 says nothing.

20.11 Condition Numbers of Eigenvalues

The Bauer–Fike theorem of §04 bounds the movement of every eigenvalue of a diagonalizable matrix at once, by $\kappa_2(\mathbf{X})\|\mathbf{E}\|_2$ (Theorem 20.4.2). That is one number for the whole spectrum, and it depends on a choice of eigenvector matrix $\mathbf{X}$. §04 warned that a well-behaved eigenvalue can sit inside a badly conditioned $\mathbf{X}$, so the bound can be far too pessimistic for some eigenvalues. This section asks about **one** eigenvalue at a time. If λ is a simple eigenvalue of $\mathbf{A}$, how fast does it move when $\mathbf{A}$ is pushed in a direction $\mathbf{E}$? The answer is an exact derivative, $\mathbf{y}^*\mathbf{E}\mathbf{x}/(\mathbf{y}^*\mathbf{x})$, built from a right and a left eigenvector. Its worst size over all $\mathbf{E}$ with $\|\mathbf{E}\|_2 = 1$ is the condition number of that eigenvalue.

The field is $\mathbb{C}$. Throughout, $\mathbf{A}, \mathbf{E} \in M_n(\mathbb{C})$, and $\|\cdot\|$ on vectors is the Euclidean norm.

20.11.1 Left and right eigenvectors

Chapter 9 defined a left eigenvector as a nonzero $\mathbf{y}$ with $\mathbf{y}^\top\mathbf{A} = \lambda\mathbf{y}^\top$ (Definition 9.10.1). Over $\mathbb{C}$ it is more convenient to put a conjugate on it. In this section, a **left eigenvector** of $\mathbf{A}$ for λ is a nonzero $\mathbf{y} \in \mathbb{C}^n$ with

$$\mathbf{y}^*\mathbf{A} = \lambda\,\mathbf{y}^*,$$

which is the same as $\mathbf{A}^*\mathbf{y} = \overline{\lambda}\,\mathbf{y}$. Since $\mathbf{y}^* = \overline{\mathbf{y}}^\top$, the equation $\mathbf{y}^*\mathbf{A} = \lambda\mathbf{y}^*$ says exactly that $\overline{\mathbf{y}}$ is a left eigenvector in the sense of Definition 9.10.1. So the two conventions differ only by

conjugating the vector. The conjugate is worth it for one reason: the number $\mathbf{y}^*\mathbf{x}$ is then the standard inner product $\langle \mathbf{x}, \mathbf{y} \rangle$ of Chapter 11, and Cauchy–Schwarz applies to it.

For a **simple** eigenvalue, that is, one of algebraic multiplicity 1 (Definition 9.3.1), the right and left eigenvectors are as rigid as they can be, and they are never orthogonal. That last fact is what the whole section divides by.

Lemma 20.11.1. Left and Right Eigenvectors of a Simple Eigenvalue

Let $\mathbf{A} \in M_n(\mathbb{C})$ and let λ be a simple eigenvalue of $\mathbf{A}$, with a right eigenvector $\mathbf{x}$ and a left eigenvector $\mathbf{y}$.

(a) The right eigenvectors of $\mathbf{A}$ for λ are exactly the nonzero multiples of $\mathbf{x}$, and the left eigenvectors are exactly the nonzero multiples of $\mathbf{y}$.

(b) $\mathbf{y}^*\mathbf{x} \neq 0$.

Idea. Split λ off from the rest of the spectrum. In a basis that begins with $\mathbf{x}$, the matrix becomes block triangular with the 1×1 block (λ) in the corner and a block $\mathbf{B}$ with $\lambda \notin \operatorname{spec}(\mathbf{B})$. Chapter 12 §10 clears the off-diagonal corner whenever the two diagonal blocks have disjoint spectra, so $\mathbf{A}$ is similar to $(\lambda) \oplus \mathbf{B}$. In that picture everything is visible: $\mathbf{e}_1$ is the right eigenvector, $\mathbf{e}_1^*$ the left one, and $\mathbf{e}_1^*\mathbf{e}_1 = 1$.

Proof. If $n = 1$, both claims hold because every nonzero vector of $\mathbb{C}^1$ is a multiple of every other, and $\mathbf{y}^*\mathbf{x} = \overline{y_1}x_1$ is a product of two nonzero numbers. Let $n \geq 2$.

By Theorem 1.6.5, extend the independent list $(\mathbf{x})$ to a basis of $\mathbb{C}^n$, and let $\mathbf{P}_0$ be the invertible matrix with these columns, so that $\mathbf{P}_0\mathbf{e}_1 = \mathbf{x}$. Then $\mathbf{P}_0^{-1}\mathbf{A}\mathbf{P}_0\mathbf{e}_1 = \mathbf{P}_0^{-1}\lambda\mathbf{x} = \lambda\mathbf{e}_1$, so

$$\mathbf{P}_0^{-1}\mathbf{A}\mathbf{P}_0 = \begin{pmatrix} \lambda & \mathbf{c} \\ \mathbf{0} & \mathbf{B} \end{pmatrix}$$

for some $\mathbf{c} \in M_{1\times(n-1)}(\mathbb{C})$ and $\mathbf{B} \in M_{n-1}(\mathbb{C})$. By Theorem 7.10.3 and Theorem 7.8.4, $p_{\mathbf{A}}(x) = (x - \lambda)\, p_{\mathbf{B}}(x)$. If $p_{\mathbf{B}}(\lambda) = 0$, then $x - \lambda$ would divide $p_{\mathbf{B}}$ by the Factor Theorem (Lemma 3.7.1), so $(x - \lambda)^2$ would divide $p_{\mathbf{A}}$, against $a_{\mathbf{A}}(\lambda) = 1$. Hence $\lambda \notin \operatorname{spec}(\mathbf{B})$ (Theorem 9.2.1). The spectra $\{\lambda\}$ and $\operatorname{spec}(\mathbf{B})$ are disjoint, so by Corollary 12.10.3 (c) and (a) there is $\mathbf{Q} = \left(\begin{smallmatrix} 1 & \mathbf{z} \\ \mathbf{0} & \mathbf{I}_{n-1} \end{smallmatrix}\right)$ with $\mathbf{Q}^{-1}\mathbf{P}_0^{-1}\mathbf{A}\mathbf{P}_0\mathbf{Q} = (\lambda) \oplus \mathbf{B}$. Put $\mathbf{P} = \mathbf{P}_0\mathbf{Q}$. Then

$$\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = (\lambda) \oplus \mathbf{B}, \qquad \mathbf{P}\mathbf{e}_1 = \mathbf{P}_0\mathbf{Q}\mathbf{e}_1 = \mathbf{P}_0\mathbf{e}_1 = \mathbf{x}.$$

(a) Let $\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$ and write $\mathbf{P}^{-1}\mathbf{v} = (u_1, \mathbf{u}')$ with $\mathbf{u}' \in \mathbb{C}^{n-1}$. Then $((\lambda) \oplus \mathbf{B})\mathbf{P}^{-1}\mathbf{v} = \lambda\mathbf{P}^{-1}\mathbf{v}$, whose last $n - 1$ rows say $(\mathbf{B} - \lambda\mathbf{I})\mathbf{u}' = \mathbf{0}$. Since $\lambda \notin \operatorname{spec}(\mathbf{B})$, the matrix $\mathbf{B} - \lambda\mathbf{I}$ is invertible (Theorem 9.2.1), so $\mathbf{u}' = \mathbf{0}$ and $\mathbf{v} = u_1\mathbf{P}\mathbf{e}_1 = u_1\mathbf{x}$. For the left side, let $\mathbf{w}^*\mathbf{A} = \lambda\mathbf{w}^*$ and write $\mathbf{w}^*\mathbf{P} = (w_1, \mathbf{w}'^*)$. Then $(\mathbf{w}^*\mathbf{P})((\lambda) \oplus \mathbf{B}) = \lambda\,\mathbf{w}^*\mathbf{P}$, whose last $n - 1$ entries say $\mathbf{w}'^*(\mathbf{B} - \lambda\mathbf{I}) = \mathbf{0}$. Multiplying on the right by $(\mathbf{B} - \lambda\mathbf{I})^{-1}$ gives $\mathbf{w}' = \mathbf{0}$, so $\mathbf{w}^* = w_1\mathbf{e}_1^*\mathbf{P}^{-1}$. Every left eigenvector is therefore a multiple of the nonzero row $\mathbf{e}_1^*\mathbf{P}^{-1}$, and in particular every left eigenvector is a multiple of any other. This proves (a).

(b) By the computation just made, $\mathbf{y}^* = w_1\mathbf{e}_1^*\mathbf{P}^{-1}$ with $w_1 \neq 0$, because $\mathbf{y} \neq \mathbf{0}$. Hence

$$\mathbf{y}^*\mathbf{x} = w_1\, \mathbf{e}_1^*\mathbf{P}^{-1}\mathbf{P}\mathbf{e}_1 = w_1\, \mathbf{e}_1^*\mathbf{e}_1 = w_1 \neq 0.$$

This proves the lemma.

■

The proof produced a little more than was asked: the left eigenvector $\mathbf{e}_1^*\mathbf{P}^{-1}$ is normalized against $\mathbf{x}$, in the sense that $(\mathbf{e}_1^*\mathbf{P}^{-1})\mathbf{x} = 1$. This is the one-eigenvalue version of the biorthogonal normalization $\mathbf{y}_i^\top \mathbf{v}_i = 1$ of Proposition 9.10.7, and it exists for a simple eigenvalue of any matrix, diagonalizable or not.

Warning. **Simple is essential: at a Jordan block, the two eigenvectors are orthogonal.** For $\mathbf{J}_2(0) = \left(\begin{smallmatrix} 0 & 1 \\ 0 & 0 \end{smallmatrix}\right)$ the only eigenvalue is 0, twice. The right eigenvectors are the multiples of $\mathbf{x} = \mathbf{e}_1$, since $\mathbf{J}_2(0)\mathbf{e}_1 = \mathbf{0}$. The left eigenvectors are the multiples of $\mathbf{y} = \mathbf{e}_2$, since $\mathbf{e}_2^*\mathbf{J}_2(0)$ is the second row, which is zero. So $\mathbf{y}^*\mathbf{x} = 0$. Every formula in this section that divides by $\mathbf{y}^*\mathbf{x}$ is meaningless here, and §03's example shows why it must be: the eigenvalues of $\mathbf{J}_2(0) + t\mathbf{E}_{21}$ are $\pm\sqrt{t}$, which have no derivative at $t = 0$.

Quick check. Let $\mathbf{A} = \left(\begin{smallmatrix} 1 & 1 \\ 0 & 2 \end{smallmatrix}\right)$. Find a right and a left eigenvector for the eigenvalue 1, and check that $\mathbf{y}^*\mathbf{x} \neq 0$.

Solution. $\mathbf{A}\mathbf{e}_1 = \mathbf{e}_1$, so $\mathbf{x} = \mathbf{e}_1$. A left eigenvector solves $\mathbf{y}^*(\mathbf{A} - \mathbf{I}) = \mathbf{0}$, and $\mathbf{A} - \mathbf{I} = \left(\begin{smallmatrix} 0 & 1 \\ 0 & 1 \end{smallmatrix}\right)$; for $\mathbf{y}^* = (a, b)$ this reads $(0, a + b) = (0, 0)$, so $\mathbf{y} = (1, -1)$ will do. Then $\mathbf{y}^*\mathbf{x} = 1 \neq 0$, as Lemma 20.11.1 (b) requires. Here $\mathbf{y}$ is not a multiple of $\mathbf{x}$: the matrix is not normal.

20.11.2 The derivative of a simple eigenvalue

Now move $\mathbf{A}$ along a straight line $\mathbf{A} + t\mathbf{E}$, with t real. Chapter 16 proved that the eigenvalues move continuously (Corollary 16.7.4), but only as an unnumbered list, and Proposition 16.7.2 showed that no single eigenvalue can be followed continuously in general. A simple eigenvalue is the exception: it stays isolated from the rest for small t, so "the eigenvalue near λ" makes sense. The theorem says that this eigenvalue is differentiable at $t = 0$, and it names the derivative.

Theorem 20.11.2. Derivative of a Simple Eigenvalue

Let $\mathbf{A} \in M_n(\mathbb{C})$, let λ be a simple eigenvalue of $\mathbf{A}$ with a right eigenvector $\mathbf{x}$ and a left eigenvector $\mathbf{y}$, and let $\mathbf{E} \in M_n(\mathbb{C})$. Then there are $r > 0$ and $t_0 > 0$ with the following properties.

(a) For every **real** t with $|t| < t_0$, the matrix $\mathbf{A} + t\mathbf{E}$ has **exactly one** eigenvalue, counted with algebraic multiplicity, in the open disc $\{z \in \mathbb{C} : |z - \lambda| < r\}$. Call it $\lambda(t)$. Thus $\lambda(0) = \lambda$, and $\lambda(t)$ is a simple eigenvalue of $\mathbf{A} + t\mathbf{E}$.

(b) $\lambda(t) \to \lambda$ as $t \to 0$.

(c) The function $t \mapsto \lambda(t)$ is differentiable at $t = 0$, and

$$\lambda'(0) = \lim_{t \to 0} \frac{\lambda(t) - \lambda}{t} = \frac{\mathbf{y}^*\mathbf{E}\mathbf{x}}{\mathbf{y}^*\mathbf{x}}.$$

In other words, $\lambda(t) = \lambda + t\,\mathbf{y}^*\mathbf{E}\mathbf{x}/(\mathbf{y}^*\mathbf{x}) + \rho(t)$ with $\rho(t)/t \to 0$. The denominator is nonzero by Lemma 20.11.1 (b), and the quotient does not change when $\mathbf{x}$ or $\mathbf{y}$ is rescaled, which by part (a) of the lemma is the only freedom there is.

Idea. The target identity falls out of one line if we have eigenvectors $\mathbf{x}_t$ of $\mathbf{A} + t\mathbf{E}$ that converge to an eigenvector of $\mathbf{A}$: multiply $(\mathbf{A} + t\mathbf{E})\mathbf{x}_t = \lambda(t)\mathbf{x}_t$ on the left by $\mathbf{y}^*$ and use $\mathbf{y}^*\mathbf{A} = \lambda\mathbf{y}^*$, to get

$$(\lambda(t) - \lambda)\,\mathbf{y}^*\mathbf{x}_t = t\,\mathbf{y}^*\mathbf{E}\mathbf{x}_t.$$

Divide by $t\,\mathbf{y}^*\mathbf{x}_t$ and let $t \to 0$. So the plan has three steps.

① Isolate $\lambda(t)$ from the rest of the spectrum with Corollary 16.7.4.

② Produce the eigenvectors $\mathbf{x}_t$ **by a formula**. A column of the adjugate $\operatorname{adj}(\lambda(t)\mathbf{I} - \mathbf{A} - t\mathbf{E})$ is an eigenvector, and its entries are polynomials in $\lambda(t)$ and t, so it converges because $\lambda(t)$ does.

③ Check that the limit is a nonzero multiple of $\mathbf{x}$, and pass to the limit in the quotient.

Proof. Let $\lambda_1 = \lambda, \lambda_2, \dots, \lambda_n$ be the eigenvalues of $\mathbf{A}$, listed with algebraic multiplicity. Since λ is simple, $\lambda_j \neq \lambda$ for $j \geq 2$.

Step 1: the isolated eigenvalue. Put $r = \frac{1}{2}\min_{j\geq 2}|\lambda_j - \lambda|$ if $n \geq 2$, and $r = 1$ if $n = 1$; in both cases $r > 0$. By Corollary 16.7.4 with the norm $\|\cdot\|_2$ and with $\varepsilon = r$, there is $\delta > 0$ such that every $\mathbf{B}$ with $\|\mathbf{B} - \mathbf{A}\|_2 < \delta$ has eigenvalues $\mu_1, \dots, \mu_n$ that can be numbered so that $|\mu_i - \lambda_i| < r$ for every i. Put $t_0 = \delta/(\|\mathbf{E}\|_2 + 1)$. If $|t| < t_0$, then $\|(\mathbf{A} + t\mathbf{E}) - \mathbf{A}\|_2 = |t|\|\mathbf{E}\|_2 < \delta$, so the eigenvalues of $\mathbf{A} + t\mathbf{E}$ can be so numbered. Then $|\mu_1 - \lambda| < r$, while for $j \geq 2$

$$|\mu_j - \lambda| \geq |\lambda_j - \lambda| - |\mu_j - \lambda_j| > 2r - r = r$$

by Theorem 0.6.4. So exactly one entry of the list lies in the open disc of radius r about λ. This proves (a).

Step 2: continuity. Let $0 < \varepsilon < r$. By Corollary 16.7.4 again, now with this ε, there is $\delta' > 0$ such that for $|t|\|\mathbf{E}\|_2 < \delta'$ the eigenvalues of $\mathbf{A} + t\mathbf{E}$ can be numbered as $\mu'_1, \dots, \mu'_n$ with $|\mu'_i - \lambda_i| < \varepsilon$. If moreover $|t| < t_0$, then μ'_1 is an eigenvalue of $\mathbf{A} + t\mathbf{E}$ in the disc of radius $\varepsilon < r$ about λ, so $\mu'_1 = \lambda(t)$ by (a). Hence $|\lambda(t) - \lambda| < \varepsilon$ whenever $|t| < \min\bigl(t_0, \delta'/(\|\mathbf{E}\|_2 + 1)\bigr)$. This proves (b).

Step 3: an eigenvector given by a formula. Put $\mathbf{N}(t) = \lambda(t)\mathbf{I} - \mathbf{A} - t\mathbf{E}$ for $|t| < t_0$, so $\mathbf{N}(0) = \lambda\mathbf{I} - \mathbf{A}$. We first find a column of $\operatorname{adj}\mathbf{N}(0)$ that is not zero. If $n = 1$, then $\operatorname{adj}\mathbf{N}(0) = (1)$ by Definition 7.6.2. If $n \geq 2$, then $1 \leq g_{\mathbf{A}}(\lambda) \leq a_{\mathbf{A}}(\lambda) = 1$ by Theorem 9.3.1, so $\operatorname{rank}\mathbf{N}(0) = n - 1 \geq 1$ by Theorem 3.5.7. By Theorem 7.7.1 (a), $\mathbf{N}(0)$ has a nonzero $(n-1)\times(n-1)$ minor M_{ij}. Then the cofactor $C_{ij} = (-1)^{i+j}M_{ij}$ is nonzero, and it is the (j, i) entry of $\operatorname{adj}\mathbf{N}(0)$ (Definition 7.6.2). In either case some column k of $\operatorname{adj}\mathbf{N}(0)$ is nonzero. Define

$$\mathbf{x}_t = \operatorname{adj}(\mathbf{N}(t))\,\mathbf{e}_k \qquad (|t| < t_0).$$

Since $\lambda(t)$ is an eigenvalue of $\mathbf{A} + t\mathbf{E}$, we have $\det\mathbf{N}(t) = p_{\mathbf{A}+t\mathbf{E}}(\lambda(t)) = 0$ (Theorem 9.2.1). So Theorem 7.6.2 gives $\mathbf{N}(t)\operatorname{adj}\mathbf{N}(t) = \mathbf{O}$, and in particular

$$(\mathbf{A} + t\mathbf{E})\,\mathbf{x}_t = \lambda(t)\,\mathbf{x}_t \qquad (|t| < t_0).$$

At $t = 0$ this says $\mathbf{A}\mathbf{x}_0 = \lambda \mathbf{x}_0$, and $\mathbf{x}_0 \neq \mathbf{0}$ by the choice of k. By Lemma 20.11.1 (a), $\mathbf{x}_0 = c\,\mathbf{x}$ for some $c \neq 0$.

Next, $\mathbf{x}_t \to \mathbf{x}_0$ as $t \to 0$. Each entry of $\operatorname{adj}\mathbf{B}$ is $\pm$ the determinant of a submatrix of $\mathbf{B}$, or the constant 1 when $n = 1$, hence a polynomial in the entries of $\mathbf{B}$ by Theorem 7.4.3. So $\mathbf{B} \mapsto \operatorname{adj}\mathbf{B}$ is continuous by Lemma 16.7.1. And $\|\mathbf{N}(t) - \mathbf{N}(0)\|_2 \leq |\lambda(t) - \lambda| + |t|\|\mathbf{E}\|_2$, which tends to 0 by (b). Hence $\operatorname{adj}\mathbf{N}(t) \to \operatorname{adj}\mathbf{N}(0)$, and $\|\mathbf{x}_t - \mathbf{x}_0\| \leq \|\operatorname{adj}\mathbf{N}(t) - \operatorname{adj}\mathbf{N}(0)\|_2$ by Theorem 16.3.2 (a), since $\|\mathbf{e}_k\| = 1$. So $\mathbf{x}_t \to \mathbf{x}_0$.

Step 4: the derivative. Multiply the eigen-equation of Step 3 on the left by $\mathbf{y}^*$, and use $\mathbf{y}^*\mathbf{A} = \lambda \mathbf{y}^*$:

$$\lambda\,\mathbf{y}^*\mathbf{x}_t + t\,\mathbf{y}^*\mathbf{E}\mathbf{x}_t = \lambda(t)\,\mathbf{y}^*\mathbf{x}_t, \qquad \text{so} \qquad (\lambda(t) - \lambda)\,\mathbf{y}^*\mathbf{x}_t = t\,\mathbf{y}^*\mathbf{E}\mathbf{x}_t.$$

By Cauchy–Schwarz (Theorem 11.1.5), $|\mathbf{y}^*\mathbf{x}_t - \mathbf{y}^*\mathbf{x}_0| \leq \|\mathbf{y}\|\|\mathbf{x}_t - \mathbf{x}_0\| \to 0$, and likewise $\mathbf{y}^*\mathbf{E}\mathbf{x}_t \to \mathbf{y}^*\mathbf{E}\mathbf{x}_0$, because $\|\mathbf{E}\mathbf{x}_t - \mathbf{E}\mathbf{x}_0\| \leq \|\mathbf{E}\|_2\|\mathbf{x}_t - \mathbf{x}_0\|$. The limit $\mathbf{y}^*\mathbf{x}_0 = c\,\mathbf{y}^*\mathbf{x}$ is nonzero by Lemma 20.11.1 (b). So there is $t_1 \in (0, t_0]$ with $|\mathbf{y}^*\mathbf{x}_t - \mathbf{y}^*\mathbf{x}_0| < |\mathbf{y}^*\mathbf{x}_0|$, and hence $\mathbf{y}^*\mathbf{x}_t \neq 0$, for $|t| < t_1$. For $0 < |t| < t_1$ we may divide:

$$\frac{\lambda(t) - \lambda}{t} = \frac{\mathbf{y}^*\mathbf{E}\mathbf{x}_t}{\mathbf{y}^*\mathbf{x}_t} \longrightarrow \frac{\mathbf{y}^*\mathbf{E}\mathbf{x}_0}{\mathbf{y}^*\mathbf{x}_0} \qquad (t \to 0),$$
$$\frac{\mathbf{y}^*\mathbf{E}\mathbf{x}_0}{\mathbf{y}^*\mathbf{x}_0} = \frac{c\,\mathbf{y}^*\mathbf{E}\mathbf{x}}{c\,\mathbf{y}^*\mathbf{x}} = \frac{\mathbf{y}^*\mathbf{E}\mathbf{x}}{\mathbf{y}^*\mathbf{x}},$$

by the algebra of limits, the denominator's limit being nonzero. This proves (c).

■

No implicit function theorem was needed. The usual textbook route to this formula differentiates $\det(z\mathbf{I} - \mathbf{A} - t\mathbf{E}) = 0$ implicitly, which presupposes that a differentiable solution $z = \lambda(t)$ exists; that is the implicit function theorem, a piece of analysis this book has not imported. The proof above needed only the continuity of Corollary 16.7.4 and the fact that the adjugate writes down an eigenvector whose entries are polynomials. The analysis consists of limits of polynomials, which is the algebra of limits.

Remark. Since $\lambda(s)$ is a simple eigenvalue of $\mathbf{A} + s\mathbf{E}$ for every $|s| < t_0$, the theorem applies again at $\mathbf{A} + s\mathbf{E}$ in place of $\mathbf{A}$. The eigenvalue it isolates there is $\lambda(s + h)$ for small h, by the uniqueness in (a). So λ is differentiable on the whole interval $(-t_0, t_0)$, with $\lambda'(s) = \mathbf{y}_s^*\mathbf{E}\mathbf{x}_s/(\mathbf{y}_s^*\mathbf{x}_s)$ for right and left eigenvectors $\mathbf{x}_s, \mathbf{y}_s$ of $\mathbf{A} + s\mathbf{E}$ for $\lambda(s)$. We need only $s = 0$.

The formula has a transparent special case. If $\mathbf{A}$ is Hermitian, then $\mathbf{A}^*\mathbf{x} = \mathbf{A}\mathbf{x} = \lambda\mathbf{x}$ with λ real, so $\mathbf{x}$ is also a left eigenvector and we may take $\mathbf{y} = \mathbf{x}$:

$$\lambda'(0) = \frac{\mathbf{x}^*\mathbf{E}\mathbf{x}}{\mathbf{x}^*\mathbf{x}}.$$

When $\mathbf{E}$ is Hermitian too, this is the Rayleigh quotient $R_{\mathbf{E}}(\mathbf{x})$ of Chapter 17 (Definition 17.1.1), which lies in $[\lambda_n(\mathbf{E}), \lambda_1(\mathbf{E})]$ by Proposition 17.1.1. So $|\lambda'(0)| \leq \|\mathbf{E}\|_2$ by Lemma 17.3.5, which is the first-order shadow of Weyl's bound $|\lambda_i(\mathbf{A} + \mathbf{E}) - \lambda_i(\mathbf{A})| \leq \|\mathbf{E}\|_2$ (Corollary 17.3.6).

Quick check. Let $\mathbf{A} = \operatorname{diag}(d_1, \dots, d_n)$ with distinct d_i, and let $\mathbf{E} = (e_{ij})$ be arbitrary. What is the derivative at $t = 0$ of the eigenvalue of $\mathbf{A} + t\mathbf{E}$ near d_i?

Solution. The eigenvalue d_i is simple, with $\mathbf{x} = \mathbf{e}_i$ and $\mathbf{y} = \mathbf{e}_i$, since $\mathbf{e}_i^*\mathbf{A} = d_i\mathbf{e}_i^*$. So $\lambda'(0) = \mathbf{e}_i^*\mathbf{E}\mathbf{e}_i/(\mathbf{e}_i^*\mathbf{e}_i) = e_{ii}$. To first order, the eigenvalues of a perturbed diagonal matrix with distinct diagonal entries move by the diagonal of the perturbation, and the off-diagonal entries of $\mathbf{E}$ enter only at second order.

20.11.3 The condition number of a simple eigenvalue

The derivative $\mathbf{y}^*\mathbf{E}\mathbf{x}/(\mathbf{y}^*\mathbf{x})$ depends on the direction $\mathbf{E}$. To measure how sensitive λ itself is, bound it over all directions of a given size. By Cauchy–Schwarz and the definition of $\|\mathbf{E}\|_2$,

$$|\lambda'(0)| = \frac{|\mathbf{y}^*\mathbf{E}\mathbf{x}|}{|\mathbf{y}^*\mathbf{x}|} \le \frac{\|\mathbf{y}\|\,\|\mathbf{E}\mathbf{x}\|}{|\mathbf{y}^*\mathbf{x}|} \le \frac{\|\mathbf{x}\|\,\|\mathbf{y}\|}{|\mathbf{y}^*\mathbf{x}|}\,\|\mathbf{E}\|_2.$$

The factor in front of $\|\mathbf{E}\|_2$ depends on $\mathbf{A}$ and λ only. It recurs in every estimate below, so we give it a name.

The condition number of an eigenvalue is the largest factor by which a perturbation of the matrix can move that eigenvalue, to first order.

Definition 20.11.1. Condition Number of a Simple Eigenvalue

Let $\mathbf{A} \in M_n(\mathbb{C})$ and let λ be a **simple** eigenvalue of $\mathbf{A}$, with a right eigenvector $\mathbf{x}$ and a left eigenvector $\mathbf{y}$, so $\mathbf{y}^*\mathbf{A} = \lambda\mathbf{y}^*$. The **condition number** of λ is

$$\kappa_{\mathbf{A}}(\lambda) := \frac{\|\mathbf{x}\|\,\|\mathbf{y}\|}{|\mathbf{y}^*\mathbf{x}|}.$$

In words: take any right eigenvector and any left eigenvector for λ, multiply their lengths, and divide by the modulus of their inner product. The definition is made only for a **simple** eigenvalue, which is what makes the denominator nonzero (Lemma 20.11.1 (b)). It is also what makes the number well defined: by part (a) of the lemma, any other choice is $\alpha\mathbf{x}$ and $\beta\mathbf{y}$ with $\alpha, \beta \neq 0$, and then

$$\frac{\|\alpha\mathbf{x}\|\,\|\beta\mathbf{y}\|}{|(\beta\mathbf{y})^*(\alpha\mathbf{x})|} = \frac{|\alpha|\,|\beta|\,\|\mathbf{x}\|\,\|\mathbf{y}\|}{|\overline{\beta}\alpha|\,|\mathbf{y}^*\mathbf{x}|} = \kappa_{\mathbf{A}}(\lambda).$$

The subscript $\mathbf{A}$ and the argument λ are part of the notation. They keep $\kappa_{\mathbf{A}}(\lambda)$ apart from the condition number $\kappa(\mathbf{A}) = \|\mathbf{A}\|\|\mathbf{A}^{-1}\|$ of a matrix (Definition 16.8.1).

Examples.

- **$\mathbf{A}$ 1×1 matrix.** For $\mathbf{A} = (a)$ the eigenvalue a is simple, with $x = y = 1$, so $\kappa_{\mathbf{A}}(a) = 1$. This is the degenerate case, and it is the right answer: the eigenvalue of $(a + te)$ is $a + te$, which moves exactly as fast as the matrix.
- **A diagonal matrix with distinct entries.** For d_i, the Quick check above found $\mathbf{x} = \mathbf{y} = \mathbf{e}_i$, so $\kappa_{\mathbf{A}}(d_i) = 1$. Every eigenvalue is perfectly conditioned.

- **A triangular 2×2.** For $\mathbf{A} = \left(\begin{smallmatrix} 1 & 1 \\ 0 & 2 \end{smallmatrix}\right)$ and the eigenvalue 1, the Quick check after the lemma found $\mathbf{x} = \mathbf{e}_1$ and $\mathbf{y} = (1, -1)$, with $\mathbf{y}^*\mathbf{x} = 1$. So $\kappa_{\mathbf{A}}(1) = 1 \cdot \sqrt{2}/1 = \sqrt{2}$. The single off-diagonal entry has made the eigenvalue somewhat more sensitive.

Nonexample by minimal change. Change the entry 2 of the last example to $1 + s$, and let $s \to 0$. For real $s \neq 0$ the matrix $\mathbf{B}_s = \left(\begin{smallmatrix} 1 & 1 \\ 0 & 1+s \end{smallmatrix}\right)$ has the simple eigenvalue 1, with $\mathbf{x} = \mathbf{e}_1$ and $\mathbf{y} = (s, -1)$, so $\kappa_{\mathbf{B}_s}(1) = \sqrt{1 + s^2}/|s|$. At $s = 0$ the matrix is $\mathbf{J}_2(1)$, and the eigenvalue 1 is no longer simple. The right and left eigenvectors still exist, $\mathbf{e}_1$ and $\mathbf{e}_2$, but $\mathbf{y}^*\mathbf{x} = 0$, so the clause "simple" fails and the quotient is undefined. The formula for $s \neq 0$ shows that it blows up on the way.

Why this definition. Three choices in it can be checked.

- **Both lengths go in the numerator.** This makes $\kappa_{\mathbf{A}}(\lambda)$ independent of the scaling of $\mathbf{x}$ and $\mathbf{y}$, as just computed. A normalization such as $\mathbf{y}^*\mathbf{x} = 1$ alone would leave one length free.
- **The norm is the Euclidean one,** because then the bound $|\lambda'(0)| \leq \kappa_{\mathbf{A}}(\lambda)\|\mathbf{E}\|_2$ above is attained by some $\mathbf{E}$, as part (b) of the proposition below shows. So $\kappa_{\mathbf{A}}(\lambda)$ is the exact worst case, not merely an upper bound. The same number is also the worst case over $\|\mathbf{E}\|_F = 1$: the bound survives because $\|\mathbf{E}\|_2 \leq \|\mathbf{E}\|_F$ (Proposition 16.3.4), and the extremal $\mathbf{E}$ has rank one, so its two norms agree by the same proposition.
- **The name** matches Definition 16.8.1. There, $\kappa(\mathbf{A})$ is the worst factor by which $\mathbf{A}$ magnifies a relative error in $\mathbf{b}$ into a relative error in the solution of $\mathbf{A}\mathbf{x} = \mathbf{b}$. Here $\kappa_{\mathbf{A}}(\lambda)$ is the worst factor by which a change in the matrix is magnified into a change in one eigenvalue, to first order.

Warning. **A condition number is a first-order statement, not a bound.** It governs $\lambda(t)$ only for small t, through $\lambda(t) - \lambda = t\lambda'(0) + \rho(t)$, and the remainder $\rho(t)$ can be as large as the first-order term while $t\|\mathbf{E}\|_2$ still looks small. Take $\mathbf{A} = \left(\begin{smallmatrix} 1 & M \\ 0 & 2 \end{smallmatrix}\right)$ with $M > 0$, for which $\kappa_{\mathbf{A}}(1) = \sqrt{1 + M^2}$, as the worked example below computes, and $\mathbf{E} = \mathbf{E}_{21}$. At $t = -1/(4M)$ the matrix $\mathbf{A} + t\mathbf{E}$ has trace 3 and determinant $2 + \frac{1}{4}$, so both its eigenvalues equal $\frac{3}{2}$. The eigenvalue 1 has moved by $\frac{1}{2}$, while the first-order prediction is $\kappa_{\mathbf{A}}(1)\,|t|\,\|\mathbf{E}\|_2 = \sqrt{1 + M^2}/(4M)$, which is close to $\frac{1}{4}$ for large M. At that point the eigenvalue is no longer simple, and the square root in the exact formula $(3 - \sqrt{1 + 4Mt})/2$ has no derivative there.

The first payoff is a list of properties, which together say what kind of number $\kappa_{\mathbf{A}}(\lambda)$ is.

Proposition 20.11.3. Properties of the Eigenvalue Condition Number

Let $\mathbf{A} \in M_n(\mathbb{C})$ and let λ be a simple eigenvalue of $\mathbf{A}$ with right eigenvector $\mathbf{x}$ and left eigenvector $\mathbf{y}$.

(a) $\kappa_{\mathbf{A}}(\lambda)$ does not depend on the choice of $\mathbf{x}$ and $\mathbf{y}$, and $\kappa_{\mathbf{A}}(\lambda) \geq 1$, with equality if and only if $\mathbf{y}$ is a multiple of $\mathbf{x}$.

(b) For every $\mathbf{E} \in M_n(\mathbb{C})$, the eigenvalue $\lambda(t)$ of Theorem 20.11.2 satisfies $|\lambda'(0)| \leq \kappa_{\mathbf{A}}(\lambda)\|\mathbf{E}\|_2$. Equality holds for $\mathbf{E} = \mathbf{y}\mathbf{x}^*/(\|\mathbf{x}\|\|\mathbf{y}\|)$, which has $\|\mathbf{E}\|_2 = 1$. Hence $\kappa_{\mathbf{A}}(\lambda) = \max\{|\lambda'(0)| : \|\mathbf{E}\|_2 = 1\}$.

(c) If $\mathbf{A}$ is normal, then $\kappa_{\mathbf{A}}(\mu) = 1$ for every simple eigenvalue μ of $\mathbf{A}$.

(d) If $\mathbf{A} = \mathbf{X}\mathbf{\Lambda}\mathbf{X}^{-1}$ with $\mathbf{X}$ invertible and $\mathbf{\Lambda} = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$, and λ_i is simple, then $\kappa_{\mathbf{A}}(\lambda_i) \leq \kappa_2(\mathbf{X})$.

(e) $\kappa_{\mathbf{A}}(\lambda) = 1/\cos\theta$, where $\theta \in [0, \pi/2)$ is the principal angle between the lines $\operatorname{span}(\mathbf{x})$ and $\operatorname{span}(\mathbf{y})$ (Definition 20.8.1).

Proof.

(a) Independence of the choice was checked after Definition 20.11.1. By Theorem 11.1.5, $|\mathbf{y}^*\mathbf{x}| = |\langle \mathbf{x}, \mathbf{y}\rangle| \le \|\mathbf{x}\|\|\mathbf{y}\|$, so $\kappa_{\mathbf{A}}(\lambda) \ge 1$, with equality if and only if $(\mathbf{x}, \mathbf{y})$ is linearly dependent, that is, since both are nonzero, if and only if $\mathbf{y}$ is a multiple of $\mathbf{x}$.

(b) The inequality is the display before Definition 20.11.1, with $\lambda'(0)$ given by Theorem 20.11.2 (c). For the equality, put $\mathbf{E} = \mathbf{y}\mathbf{x}^*/(\|\mathbf{x}\|\|\mathbf{y}\|)$. For every $\mathbf{v}$, $\mathbf{y}\mathbf{x}^*\mathbf{v} = (\mathbf{x}^*\mathbf{v})\,\mathbf{y}$, so $\|\mathbf{y}\mathbf{x}^*\mathbf{v}\| = |\langle \mathbf{v}, \mathbf{x}\rangle|\,\|\mathbf{y}\| \le \|\mathbf{x}\|\|\mathbf{y}\|\|\mathbf{v}\|$ by Cauchy–Schwarz, with equality at $\mathbf{v} = \mathbf{x}$. Hence $\|\mathbf{y}\mathbf{x}^*\|_2 = \|\mathbf{x}\|\|\mathbf{y}\|$ by Definition 16.3.2, and $\|\mathbf{E}\|_2 = 1$. Moreover

$$\mathbf{y}^*\mathbf{E}\mathbf{x} = \frac{(\mathbf{y}^*\mathbf{y})(\mathbf{x}^*\mathbf{x})}{\|\mathbf{x}\|\|\mathbf{y}\|} = \|\mathbf{x}\|\,\|\mathbf{y}\|, \qquad \text{so} \qquad |\lambda'(0)| = \frac{\|\mathbf{x}\|\,\|\mathbf{y}\|}{|\mathbf{y}^*\mathbf{x}|} = \kappa_{\mathbf{A}}(\lambda).$$

The maximum formula restates the inequality and its equality case.

(c) Let $\mathbf{A}\mathbf{x} = \mu\mathbf{x}$ with $\mathbf{x} \neq \mathbf{0}$. By Theorem 12.3.3, $\mathbf{A}^*\mathbf{x} = \overline{\mu}\,\mathbf{x}$, and taking conjugate transposes gives $\mathbf{x}^*\mathbf{A} = \mu\,\mathbf{x}^*$. So $\mathbf{x}$ is also a left eigenvector for μ, and (a) gives $\kappa_{\mathbf{A}}(\mu) = 1$.

(d) Write $\mathbf{x}_i = \mathbf{X}\mathbf{e}_i$ and $\mathbf{y}_i^* = \mathbf{e}_i^*\mathbf{X}^{-1}$. From $\mathbf{A}\mathbf{X} = \mathbf{X}\mathbf{\Lambda}$ and $\mathbf{X}^{-1}\mathbf{A} = \mathbf{\Lambda}\mathbf{X}^{-1}$, column i and row i give $\mathbf{A}\mathbf{x}_i = \lambda_i\mathbf{x}_i$ and $\mathbf{y}_i^*\mathbf{A} = \lambda_i\mathbf{y}_i^*$, and both vectors are nonzero because $\mathbf{X}$ is invertible. Also $\mathbf{y}_i^*\mathbf{x}_i = \mathbf{e}_i^*\mathbf{X}^{-1}\mathbf{X}\mathbf{e}_i = 1$. Now $\|\mathbf{x}_i\| \le \|\mathbf{X}\|_2\|\mathbf{e}_i\| = \|\mathbf{X}\|_2$ by Theorem 16.3.2 (a). For $\mathbf{y}_i$, Cauchy–Schwarz gives

$$\|\mathbf{y}_i\|^2 = \mathbf{y}_i^*\mathbf{y}_i = \mathbf{e}_i^*\mathbf{X}^{-1}\mathbf{y}_i \le \|\mathbf{X}^{-1}\mathbf{y}_i\| \le \|\mathbf{X}^{-1}\|_2\,\|\mathbf{y}_i\|,$$

so $\|\mathbf{y}_i\| \le \|\mathbf{X}^{-1}\|_2$, after dividing by $\|\mathbf{y}_i\| > 0$. Therefore $\kappa_{\mathbf{A}}(\lambda_i) = \|\mathbf{x}_i\|\|\mathbf{y}_i\|/1 \le \|\mathbf{X}\|_2\|\mathbf{X}^{-1}\|_2 = \kappa_2(\mathbf{X})$, by (a) and Definition 16.8.1.

(e) The vectors $\mathbf{x}/\|\mathbf{x}\|$ and $\mathbf{y}/\|\mathbf{y}\|$ are orthonormal bases of the two lines, so by Definition 20.8.1 the single principal angle satisfies $\cos\theta = \sigma_1(c)$, where $c = \mathbf{y}^*\mathbf{x}/(\|\mathbf{x}\|\|\mathbf{y}\|)$ is a 1×1 matrix. Its only singular value is $|c|$, the square root of the eigenvalue $\overline{c}c$ of c^*c (Definition 13.9.1). Hence $\cos\theta = |\mathbf{y}^*\mathbf{x}|/(\|\mathbf{x}\|\|\mathbf{y}\|) = 1/\kappa_{\mathbf{A}}(\lambda)$, which is positive by Lemma 20.11.1 (b), so $\theta < \pi/2$. This proves the proposition.

■

Part (e) is the picture to keep. An eigenvalue is well conditioned when its right and left eigenvectors point in nearly the same direction, and ill conditioned when they are nearly perpendicular. For a normal matrix they coincide, by (c). At a Jordan block they are exactly perpendicular, by the warning after the lemma.

Part (d) places the new number beside §04. Bauer–Fike bounds the movement of every eigenvalue by $\kappa_2(\mathbf{X})\|\mathbf{E}\|_2$, for every size of $\mathbf{E}$. Part (d) says that, to first order, each simple eigenvalue individually does at least as well, with its own constant $\kappa_{\mathbf{A}}(\lambda_i)$, and that constant involves no choice of $\mathbf{X}$. This is the refinement §04's warning promised. The first-order statement is sharper. The Bauer–Fike bound is global and needs no smallness.

20.11.4 A worked example

Upper triangular 2×2 matrices show everything, because their eigenvectors can be written down at once.

Example. The condition numbers of a triangular 2×2 matrix

Let $a, b, d \in \mathbb{C}$ with $a \neq d$, and $\mathbf{A} = \left(\begin{smallmatrix} a & b \\ 0 & d \end{smallmatrix}\right)$.

(a) Compute $\kappa_{\mathbf{A}}(a)$ and $\kappa_{\mathbf{A}}(d)$.

(b) For $a = 1$, $b = M > 0$, $d = 2$, compare with the smallest value of $\kappa_2(\mathbf{X})$ over all diagonalizing $\mathbf{X}$.

(c) For the same matrix and $\mathbf{E} = \mathbf{E}_{21}$, compute $\lambda'(0)$ for the eigenvalue 1 from Theorem 20.11.2, and compare with the exact eigenvalues of $\mathbf{A} + t\mathbf{E}$ for $M = 100$ and $t = 10^{-4}$.

Solution.

(a) Both eigenvalues are simple, since $a \neq d$. For a: $\mathbf{A}\mathbf{e}_1 = a\mathbf{e}_1$, so $\mathbf{x} = \mathbf{e}_1$. A left eigenvector solves $\mathbf{y}^*(\mathbf{A} - a\mathbf{I}) = \mathbf{0}$, where $\mathbf{A} - a\mathbf{I} = \left(\begin{smallmatrix} 0 & b \\ 0 & d-a \end{smallmatrix}\right)$. For a row $\mathbf{y}^* = (u, v)$ the product is $(0,\ ub + v(d-a))$, so $\mathbf{y}^* = (d-a, -b)$ works. Then $\mathbf{y}^*\mathbf{x} = d - a$, and

$$\kappa_{\mathbf{A}}(a) = \frac{1 \cdot \sqrt{|d-a|^2 + |b|^2}}{|d-a|} = \sqrt{1 + \frac{|b|^2}{|a-d|^2}}.$$

For d: $\mathbf{y} = \mathbf{e}_2$ is a left eigenvector, because $\mathbf{e}_2^*\mathbf{A} = (0, d) = d\,\mathbf{e}_2^*$. A right eigenvector is $\mathbf{x} = (b, d-a)$, since $(\mathbf{A} - d\mathbf{I})\mathbf{x} = \big((a-d)b + b(d-a),\ 0\big) = \mathbf{0}$. Then $\mathbf{y}^*\mathbf{x} = d - a$, and $\kappa_{\mathbf{A}}(d)$ is given by the same formula. The two condition numbers are equal. They are large exactly when the off-diagonal entry is large compared with the gap between the eigenvalues.

(b) Here $\kappa_{\mathbf{A}}(1) = \kappa_{\mathbf{A}}(2) = \sqrt{1 + M^2}$. §04 computed κ_2 for two diagonalizing matrices of this $\mathbf{A}$ (Example (A large corner, separated eigenvalues)), and the better one, with unit columns, gave $M + \sqrt{1 + M^2}$. We show that no choice does better. The eigenvectors are the multiples of $\mathbf{e}_1$ and of $(M, 1)$, and the columns of a diagonalizing $\mathbf{X}$ are eigenvectors (Theorem 9.4.1). So, up to the order of its columns, which does not change $\kappa_2(\mathbf{X})$, every diagonalizing $\mathbf{X}$ has the form $\mathbf{X} = \left(\begin{smallmatrix} \alpha & \beta M \\ 0 & \beta \end{smallmatrix}\right)$ with $\alpha, \beta \neq 0$, by Lemma 20.11.1 (a). Multiplying $\mathbf{X}$ by a scalar does not change $\kappa_2(\mathbf{X})$, so we may take $\alpha = 1$. Let $\sigma_1 \geq \sigma_2 > 0$ be the singular values of $\mathbf{X}$. They are the square roots of the eigenvalues of $\mathbf{X}^*\mathbf{X}$ (Definition 13.9.1), so $\sigma_1^2\sigma_2^2 = \det(\mathbf{X}^*\mathbf{X}) = |\det \mathbf{X}|^2$, giving $\sigma_1\sigma_2 = |\beta|$, and $\sigma_1^2 + \sigma_2^2 = \operatorname{tr}(\mathbf{X}^*\mathbf{X}) = 1 + |\beta|^2(1 + M^2)$. With $k = \kappa_2(\mathbf{X}) = \sigma_1/\sigma_2$,

$$k + \frac{1}{k} = \frac{\sigma_1^2 + \sigma_2^2}{\sigma_1\sigma_2} = \frac{1}{|\beta|} + |\beta|(1 + M^2) \geq 2\sqrt{1 + M^2},$$

by the inequality $u + v \geq 2\sqrt{uv}$ for $u, v > 0$, with equality exactly when $|\beta| = 1/\sqrt{1 + M^2}$, that is, when both columns of $\mathbf{X}$ are unit vectors. The function $k \mapsto k + 1/k$ is increasing for $k \geq 1$, so the smallest possible $\kappa_2(\mathbf{X})$ is the root $k \geq 1$ of $k + 1/k = 2\sqrt{1 + M^2}$, namely

$$\min_{\mathbf{X}} \kappa_2(\mathbf{X}) = M + \sqrt{1 + M^2}.$$

Indeed $(M+\sqrt{1+M^2})^{-1} = \sqrt{1+M^2}-M$, and the two add up to $2\sqrt{1+M^2}$. So the per-eigenvalue constant $\sqrt{1+M^2}$ is smaller than the best Bauer–Fike constant, by a factor that tends to 2 as M grows, in agreement with Proposition 20.11.3 (d).

(c) With $\mathbf{x} = \mathbf{e}_1$ and $\mathbf{y}^* = (1, -M)$ from (a), $\mathbf{y}^*\mathbf{E}_{21}\mathbf{x} = \mathbf{y}^*\mathbf{e}_2 = -M$ and $\mathbf{y}^*\mathbf{x} = 1$, so

$$\lambda'(0) = -M, \qquad |\lambda'(0)| = M \le \sqrt{1+M^2}\,\|\mathbf{E}_{21}\|_2 = \kappa_{\mathbf{A}}(1)\,\|\mathbf{E}\|_2.$$

Exactly: $\mathbf{A}+t\mathbf{E} = \left(\begin{smallmatrix} 1 & M \\ t & 2 \end{smallmatrix}\right)$ has trace 3 and determinant $2-Mt$, so its eigenvalues are

$$\lambda_{\pm}(t) = \frac{3 \pm \sqrt{1+4Mt}}{2},$$

and the one near 1 is $\lambda_-(t)$. For $M = 100$ and $t = 10^{-4}$ we get $4Mt = 0.04$, and $\lambda_-(t) = (3-\sqrt{1.04})/2 \approx 0.990098$. The eigenvalue has moved by about 0.0099, and the first-order prediction $|t\lambda'(0)| = Mt$ is 0.01. Differentiating $\lambda_-(t)$ directly gives $\lambda_-'(0) = -\frac{1}{4}\cdot 4M = -M$, as the theorem says.

§04 set this matrix beside Chapter 12's Example (A perturbation that moves the eigenvalues a thousand times as far), which perturbed the same corner of $\left(\begin{smallmatrix} 1 & M \\ 0 & 1 \end{smallmatrix}\right)$. There the diagonal entries are equal, and the eigenvalue 1 moved by $\sqrt{Mt}$; with $M = 100$ and $t = 10^{-4}$ that is 0.1, ten times the movement found in (c). The derivative explains the difference. With a gap of 1 between the eigenvalues, the movement is linear in t to first order, with slope M, at most $\kappa_{\mathbf{A}}(1) = \sqrt{1+M^2}$. With no gap there is no derivative at all. Chapter 12's closing remark, that a matrix with well-separated eigenvalues and a large defect behaves less dramatically, is now a formula.

20.11.5 When eigenvalues coalesce

The two computations just compared are the ends of a single family. Put

$$\mathbf{A}_s = \begin{pmatrix} 1 & M \\ 0 & 1+s \end{pmatrix} \qquad (M > 0,\ s > 0).$$

By Example (The condition numbers of a triangular 2×2 matrix) (a), with $b = M$ and $a-d = -s$,

$$\kappa_{\mathbf{A}_s}(1) = \kappa_{\mathbf{A}_s}(1+s) = \sqrt{1+\frac{M^2}{s^2}} \longrightarrow \infty \qquad (s \to 0^+).$$

As the two eigenvalues move together, the left eigenvector $(s, -M)$ of the eigenvalue 1 turns towards $\mathbf{e}_2$, perpendicular to the right eigenvector $\mathbf{e}_1$. At $s = 0$ the matrix is a Jordan block in disguise, as Chapter 12 observed, the product $\mathbf{y}^*\mathbf{x}$ is 0, and the derivative no longer exists. Instead the eigenvalue moves by $\sqrt{Mt}$ (Example (A perturbation that moves the eigenvalues a thousand times as far)), the case $n = 2$ of the n-th root behavior of §03 (Example (The Jordan block meets Elsner's bound)).

So the first-order theory and the Hölder behavior of §03 are not two separate phenomena. In this family a large condition number is the sign of a nearby coalescence. Near such a coalescence, the linear regime $|\lambda(t)-\lambda| \approx \kappa|t|\|\mathbf{E}\|_2$ holds only for smaller and smaller t, and beyond it the square root takes over. The warning above showed the same thing from the other side: for $\left(\begin{smallmatrix} 1 & M \\ 0 & 2 \end{smallmatrix}\right)$, a perturbation of size $1/(4M)$, about $1/(4\kappa)$, is enough to make the two eigenvalues collide.

Remark. This completes the thread that Chapter 10 §03 started when it warned that the Jordan form is not stable and said that "Chapter 20 returns to this point". §03 measured the instability, a movement of $\varepsilon^{1/n}$ for a block of size n, and showed that the exponent is the worst possible (Example (The Jordan block meets Elsner's bound), Theorem 20.3.3). §04 bounded it for every matrix, with an m-th root for a largest block of size m (Theorem 20.4.5). This section adds the first-order view. A **defective** eigenvalue, one with a Jordan block of size $m \geq 2$, has no finite condition number in any sense: along $\mathbf{J}_m(\lambda)+t\mathbf{E}_{m1}$ every eigenvalue satisfies $|\lambda(t) - \lambda|/|t| = |t|^{1/m-1}$, which is unbounded as $t \to 0$. Any defective eigenvalue inherits this by similarity: if $\mathbf{A} = \mathbf{SJS}^{-1}$ with $\mathbf{J}$ in Jordan form (Theorem 10.3.1) and $\mathbf{E} = \mathbf{SE}'\mathbf{S}^{-1}$, where $\mathbf{E}'$ carries a single 1 in the lower-left corner of a block $\mathbf{J}_m(\lambda)$ and zeros elsewhere, then $\mathbf{A} + t\mathbf{E}$ is similar to $\mathbf{J} + t\mathbf{E}'$, so among its eigenvalues are the m roots of $(z - \lambda)^m = t$, at distance $|t|^{1/m}$ from λ. And, as the family $\mathbf{A}_s$ shows, simple eigenvalues that are about to merge into such a block have condition numbers that blow up. A multiple eigenvalue that is **not** defective is different. For $\mathbf{A} = \mathbf{I}_2$, every eigenvalue of $\mathbf{I}_2 + t\mathbf{E}$ lies within $|t|\|\mathbf{E}\|_2$ of 1 by Corollary 20.4.3 (a), so nothing is infinite. What fails there is only the numbering: the two eigenvalues of $\mathbf{I}_2 + t\mathbf{E}$ cannot in general be told apart as "the one that came from 1", so Definition 20.11.1 does not apply.

20.11.6 Summary and transfer

The chapter set out to answer four questions about a perturbed matrix $\mathbf{A} + \mathbf{E}$: where its eigenvalues are, how far they move, which one goes where, and what happens to eigenvectors, invariant subspaces, singular values, and polar factors.

- **Where.** Gershgorin's discs (§01) and the other inclusion regions (§02) locate the eigenvalues without computing any of them.
- **How far.** The continuity of roots (§03) is now a theorem of the book, proved by compactness. It comes with Elsner's Hölder bound, whose exponent $1/n$ cannot be improved, and with a matching version of that bound. Bauer–Fike (§04) replaces the exponent by 1 for diagonalizable matrices, at the price of the factor $\kappa_2(\mathbf{X})$.
- **Which goes where.** Hoffman–Wielandt (§05) matches the eigenvalues of two normal matrices in the Frobenius norm.
- **Hermitian matrices** (§06) got sharper tools: residual bounds, a quadratic bound for the Rayleigh quotient, and relative bounds under congruence.
- **Eigenvectors and subspaces.** The Sylvester equation (§07) and principal angles (§08) together give Davis–Kahan (§09): an invariant subspace moves by at most the perturbation divided by the gap.
- **Singular values and polar factors** (§10) follow by the dilation and by the Sylvester bounds.
- **To first order.** This section gave each simple eigenvalue its own derivative and condition number.

Nine moves did the work.

- **Dominant coordinate plus the triangle inequality.** Look at the row of the eigen-equation where the eigenvector is largest (§01), or at the two largest coordinates (§02). *Transfer:* to locate an eigenvalue without computing it, read the eigen-equation at the coordinate where the eigenvector peaks. There the eigenvalue cannot hide behind the other terms.
- **A determinant bounded two ways.** The product of the distances from μ to the

eigenvalues of $\mathbf{A}$ is $|\det(\mathbf{A}-\mu\mathbf{I})|$, and Hadamard's inequality bounds that determinant by column lengths (§03, Elsner). *Transfer:* when a product of unknown quantities is a determinant, bound the determinant from above and the product from below, and read off the smallest factor.

- **Compactness plus uniqueness.** Roots are bounded, so a subsequence of root vectors converges. Its limit is a root vector of the limit polynomial, and unique factorization says which one (§03). *Transfer:* to prove that a solution depends continuously on the data, bound the solutions, extract a convergent subsequence, and let uniqueness of the limiting solution do the rest.
- **An integer-valued continuous function is constant.** Deform $\mathbf{A}$ to its diagonal, and count the eigenvalues in a region whose boundary no eigenvalue crosses (§03, the counting theorem). *Transfer:* to count objects that move continuously, move to a case where the count is obvious, along a path that never lets one of them escape.
- **Factor out the resolvent.** Write $\mathbf{A}+\mathbf{E}-\mu\mathbf{I} = (\mathbf{A}-\mu\mathbf{I})\big(\mathbf{I}+(\mathbf{A}-\mu\mathbf{I})^{-1}\mathbf{E}\big)$, so that the Neumann series makes the second factor invertible whenever $\|(\mathbf{A}-\mu\mathbf{I})^{-1}\mathbf{E}\| < 1$ (§04). *Transfer:* to show a perturbed matrix is invertible, factor out the unperturbed one and ask whether what remains is close to $\mathbf{I}$.
- **Birkhoff turns a doubly stochastic weight into a permutation.** The weights $|w_{ij}|^2$ of a unitary matrix form a doubly stochastic matrix, and a linear function on those attains its minimum at a permutation matrix (§05). *Transfer:* to optimize over matchings, relax to doubly stochastic matrices. A linear objective has its optimum at a vertex, and the vertices are the matchings.
- **The perturbation of a subspace is a Sylvester equation.** The block that couples the old invariant subspace to the new complement satisfies $\widetilde{\mathbf{\Lambda}}_2\mathbf{X}-\mathbf{X}\mathbf{\Lambda}_1 =$ (a piece of $\mathbf{E}$), so the separation of the two spectra is the conditioning (§07, §09). *Transfer:* to measure how far an invariant subspace moves, write down the coupling block and recognize the equation it solves.
- **The singular-pair trick.** Test a matrix equation against the top singular pair of the unknown $\mathbf{X}$. This turns $\|\mathbf{X}\|_2$ into a scalar that the hypotheses bound directly (§07, used twice in §10). *Transfer:* to bound the spectral norm of the solution of a linear matrix equation, evaluate the equation at the vectors where that norm is attained.
- **The adjugate as a polynomial eigenvector.** A column of $\operatorname{adj}(\mu\mathbf{I}-\mathbf{B})$ is an eigenvector of $\mathbf{B}$ whenever μ is an eigenvalue, and its entries are polynomials in μ and the entries of $\mathbf{B}$. So it moves continuously when they do (this section). *Transfer:* when an object must be chosen continuously and there is no canonical choice, look for a formula that is polynomial in the data. The implicit function theorem then becomes unnecessary.

The chapter also paid promises made in earlier chapters.

The promise	Made in	Paid by
"This chapter quotes this theorem, and Chapter 20 proves it" (continuous dependence of roots on coefficients)	Chapter 16 §07	§03 (Continuity of roots)
"Chapter 20 returns to this point" (the Jordan form is not stable)	Chapter 10 §03	§03, §04, and this section
"Chapter 20 develops both and proves it" (Bauer–Fike)	Chapter 12 §11	§04 (The Bauer–Fike theorem)

The promise	Made in	Paid by
"Chapter 20 proves the theorem of Bauer and Fike"	Chapter 16 §07	§04
"the theorem of Bauer and Fike, in Chapter 20"	Chapter 17 §03	§04
"the bounds that do exist, Weyl's and Bauer–Fike's, are Chapters 17 and 20"	Chapter 16 §08	§04
"Bounds for the eigenvalues of non-Hermitian matrices under perturbation, beginning with Bauer–Fike, are Chapter 20's"	Chapter 17 §11	§§03–05
"its quantitative form appears in Chapter 20" (the Sylvester equation)	Chapter 8 §04	§07 (The Sylvester equation)
"belongs with perturbation theory in Chapter 20" (conditioning of the Sylvester equation)	Chapter 12 §10	§07
"Chapter 20 leans on it for perturbation bounds" (the Hermitian dilation)	Chapter 17 §09	§10 (Singular values and polar factors)

Some things the chapter did not do. The matching form of Elsner's bound in §03 carries a factor $2n-1$, which is not the best possible, and the better constants were not pursued. The spectral-norm form of Davis–Kahan for a cluster lying between two parts of the rest of the spectrum was stated without proof (§09), and nothing depends on it. The bounds of §05, and Mirsky's theorem in §10, hold in the Frobenius norm; §10's square-root and polar bounds hold in the spectral norm as well. Extending the eigenvalue and singular value bounds to every unitarily invariant norm belongs to Chapter 21; the square-root and polar bounds are not extended there, and stay in the two norms proved here. Finally, the residual bounds of §04 and §06 say that a computed eigenpair is an exact eigenpair of a nearby matrix. Turning that into a theory of backward error, and into algorithms that compute eigenvalues, is the subject of Chapter 24.

20.11.7 Exercises

20.11.7.1 A. Check your understanding

Exercise. A1

(a) Define the condition number $\kappa_{\mathbf{A}}(\lambda)$, saying for which λ it is defined.

(b) State the formula for the derivative of a simple eigenvalue of $\mathbf{A}+t\mathbf{E}$ at $t=0$.

(c) True or false: for every eigenvalue of every $\mathbf{A}\in M_n(\mathbb{C})$, a right eigenvector $\mathbf{x}$ and a left eigenvector $\mathbf{y}$ satisfy $\mathbf{y}^*\mathbf{x}\neq 0$. Justify your answer.

(d) True or false: if $\mathbf{A}=\mathbf{X}\mathbf{\Lambda}\mathbf{X}^{-1}$ with $\kappa_2(\mathbf{X})\geq 10^6$, then some eigenvalue of $\mathbf{A}$ has $\kappa_{\mathbf{A}}(\lambda)\geq 10^6$. Justify your answer.

(e) True or false: $\kappa_{2\mathbf{A}+3\mathbf{I}}(2\lambda+3)=\kappa_{\mathbf{A}}(\lambda)$ for every simple eigenvalue λ of $\mathbf{A}$. Justify your

answer.

(f) The proof of Theorem 20.11.2 needed eigenvectors of $\mathbf{A}+t\mathbf{E}$ that converge as $t\to 0$. Where did they come from?

Solution.

(a) For a **simple** eigenvalue λ of $\mathbf{A}\in M_n(\mathbb{C})$ with right eigenvector $\mathbf{x}$ and left eigenvector $\mathbf{y}$ ($\mathbf{y}^*\mathbf{A}=\lambda\mathbf{y}^*$), $\kappa_{\mathbf{A}}(\lambda)=\|\mathbf{x}\|\|\mathbf{y}\|/|\mathbf{y}^*\mathbf{x}|$ (Definition 20.11.1).

(b) $\lambda'(0)=\mathbf{y}^*\mathbf{E}\mathbf{x}/(\mathbf{y}^*\mathbf{x})$, where $\lambda(t)$ is the unique eigenvalue of $\mathbf{A}+t\mathbf{E}$ near λ (Theorem 20.11.2).

(c) False. For $\mathbf{J}_2(0)$ the eigenvectors are the multiples of $\mathbf{x}=\mathbf{e}_1$ and of $\mathbf{y}=\mathbf{e}_2$, and $\mathbf{y}^*\mathbf{x}=0$, as in the warning after Lemma 20.11.1. The eigenvalue 0 is not simple, and simplicity is the hypothesis of that lemma.

(d) False. Let $\mathbf{A}=\operatorname{diag}(1,2)$ and $\mathbf{X}=\operatorname{diag}(1,10^6)$. Then $\mathbf{X}\mathbf{A}\mathbf{X}^{-1}=\mathbf{A}$ because diagonal matrices commute, so $\mathbf{A}=\mathbf{X}\mathbf{A}\mathbf{X}^{-1}$, and $\kappa_2(\mathbf{X})=10^6$. But $\mathbf{A}$ is normal, so both eigenvalues have condition number 1 by Proposition 20.11.3 (c). A large $\kappa_2(\mathbf{X})$ may just reflect a poor choice of $\mathbf{X}$.

(e) True. The eigenvalue $2\lambda+3$ of $2\mathbf{A}+3\mathbf{I}$ is simple, since $p_{2\mathbf{A}+3\mathbf{I}}(x)=2^n p_{\mathbf{A}}((x-3)/2)$. From $\mathbf{A}\mathbf{x}=\lambda\mathbf{x}$ and $\mathbf{y}^*\mathbf{A}=\lambda\mathbf{y}^*$ we get $(2\mathbf{A}+3\mathbf{I})\mathbf{x}=(2\lambda+3)\mathbf{x}$ and $\mathbf{y}^*(2\mathbf{A}+3\mathbf{I})=(2\lambda+3)\mathbf{y}^*$. So the same $\mathbf{x}$ and $\mathbf{y}$ serve, and the formula gives the same number.

(f) From the adjugate: a nonzero column of $\operatorname{adj}(\lambda\mathbf{I}-\mathbf{A})$ was fixed, and the same column of $\operatorname{adj}(\lambda(t)\mathbf{I}-\mathbf{A}-t\mathbf{E})$ is an eigenvector of $\mathbf{A}+t\mathbf{E}$ for $\lambda(t)$ by Theorem 7.6.2. Its entries are polynomials in $\lambda(t)$, t, and the entries of $\mathbf{A}$ and $\mathbf{E}$, so it converges because $\lambda(t)$ does (Corollary 16.7.4). No implicit function theorem was used.

20.11.7.2 B. Practice

Exercise. B1: Three condition numbers

Let

$$\mathbf{A}=\begin{pmatrix}1&3&0\\0&5&4\\0&0&2\end{pmatrix}.$$

(a) For each eigenvalue of $\mathbf{A}$, find a right and a left eigenvector with integer entries, and compute $\kappa_{\mathbf{A}}(1)$, $\kappa_{\mathbf{A}}(2)$, and $\kappa_{\mathbf{A}}(5)$.

(b) Hence decide which eigenvalue is the most sensitive to first order, and find $\mathbf{E}$ with $\|\mathbf{E}\|_2=1$ for which that eigenvalue of $\mathbf{A}+t\mathbf{E}$ has $|\lambda'(0)|$ as large as possible.

Solution.

(a) The matrix is upper triangular with diagonal $1,5,2$, so these are the eigenvalues, each simple.

Eigenvalue 1. $\mathbf{x}=\mathbf{e}_1$. For $\mathbf{y}^*=(u,v,w)$, the equation $\mathbf{y}^*(\mathbf{A}-\mathbf{I})=\mathbf{0}$ reads $(0,\ 3u+4v,\ 4v+w)=\mathbf{0}$. Take $u=4$, $v=-3$, $w=12$, so $\mathbf{y}=(4,-3,12)$, with $\|\mathbf{y}\|=\sqrt{16+9+144}=13$ and $\mathbf{y}^*\mathbf{x}=4$. Hence $\kappa_{\mathbf{A}}(1)=13/4$.

Eigenvalue 2. $\mathbf{y}=\mathbf{e}_3$, since $\mathbf{e}_3^*\mathbf{A}=(0,0,2)$. For $\mathbf{x}=(p,q,r)$, the equation $(\mathbf{A}-2\mathbf{I})\mathbf{x}=\mathbf{0}$ reads $-p+3q=0$, $3q+4r=0$. Take $r=3$, $q=-4$, $p=-12$, so $\mathbf{x}=(-12,-4,3)$,

with $\|\mathbf{x}\| = \sqrt{144+16+9} = 13$ and $\mathbf{y}^*\mathbf{x} = 3$. Hence $\kappa_{\mathbf{A}}(2) = 13/3$.
Eigenvalue 5. For $\mathbf{x} = (p, q, 0)$, $(\mathbf{A} - 5\mathbf{I})\mathbf{x} = (-4p + 3q, 0, 0)$, so $\mathbf{x} = (3, 4, 0)$. For $\mathbf{y}^* = (0, v, w)$, $\mathbf{y}^*(\mathbf{A} - 5\mathbf{I}) = (0, 0, 4v - 3w)$, so $\mathbf{y} = (0, 3, 4)$. Both have length 5, and $\mathbf{y}^*\mathbf{x} = 12$. Hence $\kappa_{\mathbf{A}}(5) = 25/12$.

(b) Since $13/3 > 13/4 > 25/12$, the eigenvalue 2 is the most sensitive, although the eigenvalue 1 is exactly as close to its nearest neighbor and less sensitive: the gaps alone do not decide. By Proposition 20.11.3 (b), the worst direction is

$$\mathbf{E} = \frac{\mathbf{y}\mathbf{x}^*}{\|\mathbf{x}\|\|\mathbf{y}\|} = \frac{1}{13}\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ -12 & -4 & 3 \end{pmatrix},$$

with $\|\mathbf{E}\|_2 = 1$ and $\lambda'(0) = \|\mathbf{x}\|\|\mathbf{y}\|/(\mathbf{y}^*\mathbf{x}) = 13/3$. Hence, to first order, a perturbation of size 10^{-3} in this direction moves the eigenvalue 2 by about 4.3×10^{-3}.

Exercise. B2: A derivative checked against the exact answer

Let $\mathbf{A} = \left(\begin{smallmatrix} 0 & 1 \\ -2 & 3 \end{smallmatrix}\right)$ and $\mathbf{E} = \mathbf{E}_{12}$.

(a) Find the eigenvalues of $\mathbf{A}$ with right and left eigenvectors, and compute both condition numbers.

(b) Use Theorem 20.11.2 to find $\lambda'(0)$ for each eigenvalue of $\mathbf{A} + t\mathbf{E}$.

(c) Find the eigenvalues of $\mathbf{A} + t\mathbf{E}$ exactly. Hence compare the first-order estimate with the exact value at $t = 10^{-3}$.

Solution.

(a) $p_{\mathbf{A}}(x) = x^2 - 3x + 2 = (x-1)(x-2)$. For 1: $\mathbf{A} - \mathbf{I} = \left(\begin{smallmatrix} -1 & 1 \\ -2 & 2 \end{smallmatrix}\right)$, so $\mathbf{x}_1 = (1, 1)$, and $\mathbf{y}^*(\mathbf{A} - \mathbf{I}) = \mathbf{0}$ for $\mathbf{y}^* = (u, v)$ means $-u - 2v = 0$, so $\mathbf{y}_1 = (2, -1)$. Then $\mathbf{y}_1^*\mathbf{x}_1 = 1$ and $\kappa_{\mathbf{A}}(1) = \sqrt{2}\sqrt{5} = \sqrt{10}$. For 2: $\mathbf{A} - 2\mathbf{I} = \left(\begin{smallmatrix} -2 & 1 \\ -2 & 1 \end{smallmatrix}\right)$, so $\mathbf{x}_2 = (1, 2)$ and $\mathbf{y}_2 = (1, -1)$, with $\mathbf{y}_2^*\mathbf{x}_2 = -1$ and $\kappa_{\mathbf{A}}(2) = \sqrt{5}\sqrt{2} = \sqrt{10}$.

(b) $\mathbf{y}^*\mathbf{E}_{12}\mathbf{x} = \overline{y_1}\, x_2$. For 1: $2 \cdot 1/1 = 2$. For 2: $1 \cdot 2/(-1) = -2$.

(c) $\mathbf{A} + t\mathbf{E} = \left(\begin{smallmatrix} 0 & 1+t \\ -2 & 3 \end{smallmatrix}\right)$ has trace 3 and determinant $2 + 2t$, so its eigenvalues are $(3 \mp \sqrt{1 - 8t})/2$, for $t < 1/8$. Differentiating at $t = 0$ gives ± 2, as in (b). At $t = 10^{-3}$, the first-order estimates are 1.002 and 1.998, and the exact values are $(3 \mp \sqrt{0.992})/2 \approx 1.002004$ and 1.997996. Hence the error of the first-order estimate is about 4×10^{-6}, which is of the order of t^2, as expected of a first-order approximation.

Exercise. B3: Invariance of the condition number

Let λ be a simple eigenvalue of $\mathbf{A} \in M_n(\mathbb{C})$. Prove that:

(a) $\overline{\lambda}$ is a simple eigenvalue of $\mathbf{A}^*$, and $\kappa_{\mathbf{A}^*}(\overline{\lambda}) = \kappa_{\mathbf{A}}(\lambda)$;

(b) $\kappa_{\mathbf{U}^*\mathbf{A}\mathbf{U}}(\lambda) = \kappa_{\mathbf{A}}(\lambda)$ for every unitary $\mathbf{U}$.

(c) Explain why (b) fails for a general invertible $\mathbf{S}$ in place of $\mathbf{U}$, using $\mathbf{A} = \operatorname{diag}(1, 2)$.

Solution.

(a) The characteristic polynomial of $\mathbf{A}^*$ has as roots the conjugates of those of $\mathbf{A}$, with the same multiplicities: $p_{\mathbf{A}^*}(x) = \det(x\mathbf{I} - \mathbf{A}^*) = \overline{\det(\overline{x}\mathbf{I} - \mathbf{A})}$, because the determinant of the conjugate transpose is the conjugate of the determinant. So $\overline{\lambda}$ is a simple eigenvalue of $\mathbf{A}^*$. If $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$ and $\mathbf{y}^*\mathbf{A} = \lambda\mathbf{y}^*$, then taking conjugate transposes gives $\mathbf{x}^*\mathbf{A}^* = \overline{\lambda}\mathbf{x}^*$ and $\mathbf{A}^*\mathbf{y} = \overline{\lambda}\mathbf{y}$. So for $\mathbf{A}^*$ and $\overline{\lambda}$, the vector $\mathbf{y}$ is a right eigenvector and $\mathbf{x}$ a left one. Hence $\kappa_{\mathbf{A}^*}(\overline{\lambda}) = \|\mathbf{y}\|\|\mathbf{x}\|/|\mathbf{x}^*\mathbf{y}| = \kappa_{\mathbf{A}}(\lambda)$, since $|\mathbf{x}^*\mathbf{y}| = |\overline{\mathbf{y}^*\mathbf{x}}| = |\mathbf{y}^*\mathbf{x}|$.

(b) $\mathbf{U}^*\mathbf{A}\mathbf{U}$ has the same characteristic polynomial as $\mathbf{A}$ (Theorem 7.10.3), so λ is simple for it. From $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$ and $\mathbf{y}^*\mathbf{A} = \lambda\mathbf{y}^*$ we get $(\mathbf{U}^*\mathbf{A}\mathbf{U})(\mathbf{U}^*\mathbf{x}) = \lambda\,\mathbf{U}^*\mathbf{x}$ and $(\mathbf{U}^*\mathbf{y})^*(\mathbf{U}^*\mathbf{A}\mathbf{U}) = \mathbf{y}^*\mathbf{A}\mathbf{U} = \lambda\,(\mathbf{U}^*\mathbf{y})^*$. Since $\mathbf{U}^*$ is unitary, $\|\mathbf{U}^*\mathbf{x}\| = \|\mathbf{x}\|$ and $\|\mathbf{U}^*\mathbf{y}\| = \|\mathbf{y}\|$ (Theorem 11.7.3), and $(\mathbf{U}^*\mathbf{y})^*(\mathbf{U}^*\mathbf{x}) = \mathbf{y}^*\mathbf{U}\mathbf{U}^*\mathbf{x} = \mathbf{y}^*\mathbf{x}$. So all three ingredients of Definition 20.11.1 are unchanged.

(c) For $\mathbf{S} = \left(\begin{smallmatrix}1 & 1\\ 0 & 1\end{smallmatrix}\right)$, $\mathbf{S}^{-1}\mathbf{A}\mathbf{S} = \left(\begin{smallmatrix}1 & -1\\ 0 & 1\end{smallmatrix}\right)\left(\begin{smallmatrix}1 & 1\\ 0 & 2\end{smallmatrix}\right) = \left(\begin{smallmatrix}1 & -1\\ 0 & 2\end{smallmatrix}\right)$. By Example (The condition numbers of a triangular 2×2 matrix) (a) with $b = -1$ and $|a-d| = 1$, its eigenvalue 1 has condition number $\sqrt{2}$, whereas $\kappa_{\mathbf{A}}(1) = 1$ because $\mathbf{A}$ is normal. A nonunitary similarity changes lengths and angles, so it changes the angle between $\mathbf{x}$ and $\mathbf{y}$ in Proposition 20.11.3 (e).

20.11.7.3 C. Going deeper

Exercise. C1: Perfect conditioning characterizes normality

(a) Let $\mathbf{A} \in M_n(\mathbb{C})$ have n distinct eigenvalues. Prove that $\mathbf{A}$ is normal if and only if $\kappa_{\mathbf{A}}(\lambda) = 1$ for every eigenvalue λ.

(b) Show that the hypothesis of distinct eigenvalues cannot be replaced by "$\kappa_{\mathbf{A}}(\lambda) = 1$ for every **simple** eigenvalue λ": give a nonnormal $\mathbf{A} \in M_3(\mathbb{C})$ with a simple eigenvalue, all of whose simple eigenvalues have condition number 1.

Hint: in (a), what does $\kappa = 1$ say about $\mathbf{y}$, and what does the computation in Theorem 9.10.6 say about $\mathbf{y}_j^*\mathbf{x}_i$?

Solution.

(a) ($\Rightarrow$) This is Proposition 20.11.3 (c).
($\Leftarrow$) Let $\lambda_1, \dots, \lambda_n$ be the distinct eigenvalues, with right eigenvectors $\mathbf{x}_1, \dots, \mathbf{x}_n$. Since $\kappa_{\mathbf{A}}(\lambda_j) = 1$, Proposition 20.11.3 (a) says that a left eigenvector for λ_j is a multiple of $\mathbf{x}_j$, so $\mathbf{x}_j$ itself is one: $\mathbf{x}_j^*\mathbf{A} = \lambda_j\mathbf{x}_j^*$. Let $i \neq j$. Computing $\mathbf{x}_j^*\mathbf{A}\mathbf{x}_i$ in two ways,

$$\lambda_j\,\mathbf{x}_j^*\mathbf{x}_i = (\mathbf{x}_j^*\mathbf{A})\mathbf{x}_i = \mathbf{x}_j^*(\mathbf{A}\mathbf{x}_i) = \lambda_i\,\mathbf{x}_j^*\mathbf{x}_i,$$

so $(\lambda_j - \lambda_i)\mathbf{x}_j^*\mathbf{x}_i = 0$, and $\mathbf{x}_j^*\mathbf{x}_i = 0$ because $\lambda_i \neq \lambda_j$. So the vectors $\mathbf{x}_i/\|\mathbf{x}_i\|$ form an orthonormal list of n vectors in $\mathbb{C}^n$, hence an orthonormal basis of eigenvectors. With $\mathbf{U}$ the unitary matrix having these columns, $\mathbf{A} = \mathbf{U}\,\mathrm{diag}(\lambda_1, \dots, \lambda_n)\mathbf{U}^*$, so $\mathbf{A}$ is normal by Corollary 12.4.3.

(b) Let $\mathbf{A} = \mathbf{J}_2(0) \oplus (1) = \left(\begin{smallmatrix}0 & 1 & 0\\ 0 & 0 & 0\\ 0 & 0 & 1\end{smallmatrix}\right)$. Its eigenvalues are $0, 0, 1$, so 1 is the only simple eigenvalue. Its right and left eigenvectors are both $\mathbf{e}_3$, since $\mathbf{A}\mathbf{e}_3 = \mathbf{e}_3$ and $\mathbf{e}_3^*\mathbf{A} = \mathbf{e}_3^*$, so $\kappa_{\mathbf{A}}(1) = 1$. But $\mathbf{A}$ is not normal: $\mathbf{A}^*\mathbf{A} = \mathrm{diag}(0, 1, 1)$ while $\mathbf{A}\mathbf{A}^* = \mathrm{diag}(1, 0, 1)$.

Exercise. C2: The derivatives add up to the trace

Let $\mathbf{A} \in M_n(\mathbb{C})$ have n distinct eigenvalues $\lambda_1, \dots, \lambda_n$, and let $\mathbf{E} \in M_n(\mathbb{C})$. Write $\lambda_i(t)$ for the eigenvalue of $\mathbf{A} + t\mathbf{E}$ near λ_i, from Theorem 20.11.2.

(a) Prove that $\lambda_1'(0) + \cdots + \lambda_n'(0) = \operatorname{tr} \mathbf{E}$.

(b) Deduce this again from $\operatorname{tr}(\mathbf{A} + t\mathbf{E}) = \lambda_1(t) + \cdots + \lambda_n(t)$, valid for all small t, and explain why the two arguments agree.

Hint: in (a), use a diagonalizing $\mathbf{X}$ *and the rows of* $\mathbf{X}^{-1}$ *as left eigenvectors.*

Solution.

(a) $\mathbf{A}$ is diagonalizable by Corollary 9.4.2, say $\mathbf{A} = \mathbf{X}\mathbf{\Lambda}\mathbf{X}^{-1}$. As in the proof of Proposition 20.11.3 (d), $\mathbf{x}_i = \mathbf{X}\mathbf{e}_i$ and $\mathbf{y}_i^* = \mathbf{e}_i^*\mathbf{X}^{-1}$ are right and left eigenvectors for λ_i, with $\mathbf{y}_i^*\mathbf{x}_i = 1$. So by Theorem 20.11.2,

$$\sum_{i=1}^{n} \lambda_i'(0) = \sum_{i=1}^{n} \mathbf{e}_i^*\mathbf{X}^{-1}\mathbf{E}\mathbf{X}\mathbf{e}_i = \operatorname{tr}(\mathbf{X}^{-1}\mathbf{E}\mathbf{X}) = \operatorname{tr} \mathbf{E},$$

where the middle equality sums the diagonal entries of $\mathbf{X}^{-1}\mathbf{E}\mathbf{X}$ and the last uses $\operatorname{tr}(\mathbf{B}\mathbf{C}) = \operatorname{tr}(\mathbf{C}\mathbf{B})$ (Theorem 0.9.6 (3)).

(b) Let g be the smallest distance between two of the λ_i. By Theorem 20.11.2 (b), $|\lambda_i(t) - \lambda_i| < g/2$ for every i once $|t|$ is small enough, so $\lambda_1(t), \dots, \lambda_n(t)$ are n distinct eigenvalues of $\mathbf{A} + t\mathbf{E}$. They are therefore all of its eigenvalues, each simple, and Theorem 9.3.3 gives $\operatorname{tr}(\mathbf{A} + t\mathbf{E}) = \sum_i \lambda_i(t)$. The left side is $\operatorname{tr} \mathbf{A} + t \operatorname{tr} \mathbf{E}$ and $\operatorname{tr} \mathbf{A} = \sum_i \lambda_i$, so $\sum_i (\lambda_i(t) - \lambda_i)/t = \operatorname{tr} \mathbf{E}$ for small $t \neq 0$. Letting $t \to 0$ and using (c) of the theorem for each i gives the claim. The two arguments agree because both express the same fact: the sum of the eigenvalues is a linear function of the matrix, so its derivative in the direction $\mathbf{E}$ is $\operatorname{tr} \mathbf{E}$, whatever the individual derivatives are.

Exercise. C3: The adjugate at a simple eigenvalue

Let $n \geq 2$, let λ be a simple eigenvalue of $\mathbf{A} \in M_n(\mathbb{C})$ with right eigenvector $\mathbf{x}$ and left eigenvector $\mathbf{y}$, and put $\mathbf{N} = \lambda\mathbf{I} - \mathbf{A}$.

(a) Prove that $\operatorname{adj} \mathbf{N} = c\,\mathbf{x}\mathbf{y}^*$ for some scalar $c \neq 0$.

(b) Prove that $\operatorname{tr}(\operatorname{adj} \mathbf{N}) = p_{\mathbf{A}}'(\lambda)$, the derivative of the characteristic polynomial at λ.

(c) Deduce a second proof that $\mathbf{y}^*\mathbf{x} \neq 0$, and the formula

$$\operatorname{adj}(\lambda\mathbf{I} - \mathbf{A}) = p_{\mathbf{A}}'(\lambda)\,\frac{\mathbf{x}\mathbf{y}^*}{\mathbf{y}^*\mathbf{x}}.$$

Hint: for (a), use both products in Theorem 7.6.2; for (b), apply Theorem 16.6.6 to $t \mapsto (\lambda + t)\mathbf{I} - \mathbf{A}$.

Solution.

(a) Since λ is an eigenvalue, $\det \mathbf{N} = 0$, so Theorem 7.6.2 gives $\mathbf{N} \operatorname{adj} \mathbf{N} = \mathbf{O}$ and $\operatorname{adj} \mathbf{N}\, \mathbf{N} =$

$\mathbf{O}$. The first says every column of $\operatorname{adj}\mathbf{N}$ lies in $\ker(\lambda\mathbf{I}-\mathbf{A})$, so by Lemma 20.11.1 (a) every column is a multiple of $\mathbf{x}$: $\operatorname{adj}\mathbf{N} = \mathbf{x}\mathbf{w}^{*}$ for some $\mathbf{w}\in\mathbb{C}^{n}$, whose conjugated entries are those multiples. Step 3 of the proof of Theorem 20.11.2 showed that $\operatorname{adj}\mathbf{N}\neq\mathbf{O}$, so $\mathbf{w}\neq\mathbf{0}$. The second identity gives $\mathbf{x}(\mathbf{w}^{*}\mathbf{N}) = \mathbf{O}$, and since $\mathbf{x}\neq\mathbf{0}$ this forces $\mathbf{w}^{*}\mathbf{N} = \mathbf{0}^{*}$, that is, $\mathbf{w}^{*}\mathbf{A} = \lambda\mathbf{w}^{*}$. So $\mathbf{w}$ is a left eigenvector, hence $\mathbf{w} = \bar{c}\,\mathbf{y}$ for some $c\neq 0$ by Lemma 20.11.1 (a), and $\operatorname{adj}\mathbf{N} = c\,\mathbf{x}\mathbf{y}^{*}$.

(b) Let $\mathbf{B}(t) = (\lambda+t)\mathbf{I}-\mathbf{A}$ for real t, so $\mathbf{B}'(0) = \mathbf{I}$ and $\mathbf{B}(0) = \mathbf{N}$. By Theorem 16.6.6, $t\mapsto\det\mathbf{B}(t)$ is differentiable at 0 with derivative $\operatorname{tr}(\operatorname{adj}\mathbf{N}\,\mathbf{I}) = \operatorname{tr}(\operatorname{adj}\mathbf{N})$. On the other hand $\det\mathbf{B}(t) = p_{\mathbf{A}}(\lambda+t)$. Expanding each power $(\lambda+t)^{k}$ by the binomial theorem gives $p_{\mathbf{A}}(\lambda+t) = p_{\mathbf{A}}(\lambda) + p_{\mathbf{A}}'(\lambda)\,t + t^{2}h(t)$ for a polynomial h, because the coefficient of t in $\sum_{k}c_{k}(\lambda+t)^{k}$ is $\sum_{k}kc_{k}\lambda^{k-1} = p_{\mathbf{A}}'(\lambda)$. So the difference quotient at 0 is $p_{\mathbf{A}}'(\lambda) + t\,h(t)\to p_{\mathbf{A}}'(\lambda)$. The two derivatives agree: $\operatorname{tr}(\operatorname{adj}\mathbf{N}) = p_{\mathbf{A}}'(\lambda)$.

(c) By (a) and (b), $p_{\mathbf{A}}'(\lambda) = \operatorname{tr}(c\,\mathbf{x}\mathbf{y}^{*}) = c\,\mathbf{y}^{*}\mathbf{x}$, since the trace of $\mathbf{x}\mathbf{y}^{*}$ is $\sum_{i}x_{i}\overline{y_{i}} = \mathbf{y}^{*}\mathbf{x}$. Because λ is a simple root of $p_{\mathbf{A}}$, Theorem 6.4.5 (a) gives $p_{\mathbf{A}}'(\lambda)\neq 0$. Hence $\mathbf{y}^{*}\mathbf{x}\neq 0$, which reproves Lemma 20.11.1 (b) from (a) of that lemma. Moreover $c = p_{\mathbf{A}}'(\lambda)/(\mathbf{y}^{*}\mathbf{x})$, which is the displayed formula. For example, for $\mathbf{A} = \left(\begin{smallmatrix}1 & M\\ 0 & 2\end{smallmatrix}\right)$ and $\lambda = 1$, we have $\operatorname{adj}(\mathbf{I}-\mathbf{A}) = \left(\begin{smallmatrix}-1 & M\\ 0 & 0\end{smallmatrix}\right)$ and $p_{\mathbf{A}}'(1) = -1$, while $\mathbf{x}\mathbf{y}^{*}/(\mathbf{y}^{*}\mathbf{x}) = \left(\begin{smallmatrix}1 & -M\\ 0 & 0\end{smallmatrix}\right)$, in agreement.

Chapter 21

Matrix Inequalities

Five chapters have left inequalities unpaid, and they turn out to be one inequality seen from five sides. Chapter 13 §10 and §11 said that the truncated singular value decomposition and the unitary polar factor are nearest in "every unitarily invariant norm", a phrase it could not define. Chapter 16 §08 admitted that it had never said which norms those are. Chapter 17 §06 found a family of functionals that fail to be norms on Hermitian matrices and guessed the repair; §07 proved a majorization and left its consequences open; §11 stated Loewner's theorem once, promised this chapter's proof four times, and proved none of it. Chapter 18 §10 and §11 deferred the convexity of $\mathbf{A} \mapsto \operatorname{tr} f(\mathbf{A})$. Chapter 19 §07 and §09 deferred the theorem of Hardy, Littlewood, and Pólya. Chapter 20 §05, §10, and §11 wanted Weyl's, Mirsky's, and Hoffman–Wielandt's bounds inside one statement.

The organizing fact behind all of them is **majorization**, the order $\mathbf{x} \prec \mathbf{y}$ that Chapter 17 §08 defined. This chapter closes the circle in four movements.

- **Majorization is a doubly stochastic map.** $\mathbf{x} \prec \mathbf{y}$ exactly when $\mathbf{x} = \mathbf{D}\mathbf{y}$ for a doubly stochastic $\mathbf{D}$, and with Birkhoff's theorem, exactly when $\mathbf{x}$ lies in the convex hull of the permutations of $\mathbf{y}$. Convexity then acts one row at a time, and every convex function respects the order.
- **A majorization becomes a norm inequality.** The norms that cannot tell coordinates apart are the symmetric gauge functions; composing one with the singular values gives a unitarily invariant norm, and every unitarily invariant norm arises this way exactly once. Ky Fan dominance then says that "$|||\mathbf{A}||| \le |||\mathbf{B}|||$ for every unitarily invariant norm" and "$\sigma(\mathbf{A}) \prec_w \sigma(\mathbf{B})$" are the same statement.
- **The deferred theorems fall out.** Eckart–Young, the nearest unitary matrix, Lidskii, and Mirsky, each in every unitarily invariant norm at once.
- **The order-theoretic side.** Which functions preserve the Loewner order $\mathbf{A} \succeq \mathbf{B}$? Very few: Loewner's theorem says they are exactly the functions with a particular integral representation, and the powers t^p qualify only for $0 \le p \le 1$. The chapter ends with the exponential, where Golden and Thompson's trace inequality replaces the identity $e^{a+b} = e^a e^b$ that matrices cannot have.

Matrices are complex unless a statement says otherwise, and f applied to a Hermitian matrix always means the spectral functional calculus of Chapter 12 §08 (Definition 12.8.2). Four notational conventions hold throughout.

- **$\boldsymbol{\lambda}(\mathbf{A})$ is the decreasing eigenvalue list** of a Hermitian $\mathbf{A}$, $\sigma(\mathbf{A})$ the decreasing singular value list, and $\mathbf{x}^{\downarrow}$ the decreasing rearrangement of a real vector, as in Chapter 17.
- **Majorization is Chapter 17 §08's** (Definition 17.8.1), with $\prec_w$ the top-sums version.

- **A unitarily invariant norm is written $|||\mathbf{A}|||$**, with triple bars, to keep it apart from any particular norm. **Symmetric gauge functions are Φ**; Chapter 18 §04's gauge p_K of a convex set is a different object with the same word, and §03 says so once.
- **The order on matrices is always the Loewner order $\mathbf{A} \succeq \mathbf{B}$** of Chapter 13 §05 (Definition 13.5.1). No matrix is ever compared entrywise: the order of Chapters 18 and 19 is used here only on **vectors**, as $\mathbf{x} \leq \mathbf{y}$ in §03, which says so at its head.

21.0.1 The analysis this chapter imports

The list is Chapter 16's, facts **(A1)**–**(A6)** of that chapter's introduction — completeness, monotone convergence, compactness of closed bounded sets, the extreme value theorem, the intermediate value theorem, and the mean value theorem — together with the **algebra of limits**. Each is named again where it is used, except the intermediate value theorem **(A5)**, which this chapter never needs. The weight falls on three: **(A1)** and **(A3)** build §10's integral, completeness for the convergence of the Riemann sums and compactness for the uniform continuity of the integrand, and **(A6)** is how §09 turns a nonnegative derivative into monotonicity along a segment. Section 9 also uses **(A3)** and **(A4)** for the same two jobs, uniform continuity and boundedness on a compact interval, and **(A2)** bounds the improper integral of §10 that represents t^p.

One fact is quoted beyond that list, and it is this chapter's own. Loewner's theorem cannot be proved from (A1)–(A6): its analytic half is a theorem about functions of a complex variable. The book's practice, as with Taylor's theorem in Chapter 14 §06 and the theorem of Abel and Ruffini in Chapter 11, is to quote one named result at its point of use and prove everything else. So:

(A7) (Pick and Nevanlinna.) *Let $I \subseteq \mathbb{R}$ be an open interval and $f\colon I \to \mathbb{R}$. Suppose the Loewner matrix $\left[f[t_i, t_j]\right]_{i,j=1}^{n}$ of divided differences is positive semidefinite for every n and all $t_1, \ldots, t_n \in I$. Then f is the restriction to I of a function that is analytic on $\mathbb{C} \setminus (\mathbb{R} \setminus I)$ and maps the upper half plane into its closure; and every such function has the integral representation of Chapter 17 §11's Theorem 17.11.9 (ii). In particular such an f is automatically infinitely differentiable on I.*

(A7) is stated once, in §10, and used three times there. Its clause (c) supplies the smoothness that §09's criterion asks for, its clause (a) the implication (i) $\Rightarrow$ (iii) of Loewner's theorem, and its clause (b) the implication (iii) $\Rightarrow$ (ii). Together they give the one thing the chapter cannot prove: that a function which is operator monotone of every order has the integral representation. No other section uses it, and everything else in the chapter is proved.

What this means in practice is worth spelling out, because the reader should know which results are conditional.

- **Resting on (A7):** only the implication just named, and therefore only the "hard" direction of Theorem 17.11.9 — that (i) implies (ii). Section 10 marks each step where it enters.
- **Not resting on (A7):** everything in §§01–08, §11, and §12; the whole of §09, including the criterion that, for a continuously differentiable f, monotonicity of order n is exactly positivity of every $n \times n$ Loewner matrix, which is the elementary heart of Loewner's theorem and is proved in full; the **easy** direction of Loewner's theorem, that the integral representation forces operator monotonicity, which §10 proves outright; the operator monotonicity of t^p for $0 \leq p \leq 1$ and of $\log t$, both obtained from that easy direction together with an explicit representation; and the three negative results — that t^p fails for every $p > 1$, that t^2 fails already at order 2, and that e^t is not operator monotone — which come from §09's criterion and from explicit two-by-two witnesses.

Section 10 also integrates a continuous matrix-valued function over a compact interval. That is **not** imported: the section builds the integral entrywise as a limit of Riemann sums, using only (A1)–(A4), and proves the one property it needs, that an integral of positive semidefinite matrices is positive semidefinite. Beyond the list, the only analytic fact quoted is the fundamental theorem of calculus, which §10 uses twice and flags both times: in a remark, to check the logarithm's integral representation, and in an exercise, to evaluate the constant $K_{1/2}$. Neither use is a step in a proof. In particular §09's derivative of a matrix function is obtained without the implicit function theorem, exactly as Chapter 20 §11's derivative of a simple eigenvalue was.

21.0.2 What you need

- **Chapter 0 §09** — matrix multiplication, and the trace with its cyclic invariance: Definition 0.9.9, Theorem 0.9.6.
- **Chapter 3** — the invertibility criteria and permutation matrices: Theorem 3.4.4, Definition 3.4.2, Lemma 3.4.7.
- **Chapter 6 §06** — Lagrange interpolation: Theorem 6.6.1.
- **Chapter 7 §05** — multiplicativity of the determinant and the determinant test for invertibility, in the exercises of §§06 and 11 only; no proof in the chapter needs a determinant: Theorem 7.5.4, Theorem 7.5.3.
- **Chapter 8 §01** — block multiplication: Theorem 8.1.1.
- **Chapter 9** — the trace and determinant read off the eigenvalues, and the spectra of **AB** and **BA**: Theorem 9.3.3, Theorem 9.10.4.
- **Chapter 10 §09** — the matrix exponential: Definition 10.9.2, Theorem 10.9.4.
- **Chapter 11 §01 and §07** — the Frobenius inner product, Pythagoras and Cauchy–Schwarz, and the unitary invariance of the Frobenius norm: Theorem 11.1.3, Theorem 11.1.5, Lemma 11.7.5.
- **Chapter 12** — the spectral theorem, the spectral resolution, and the functional calculus: Corollary 12.1.4, Theorem 12.2.1, Corollary 12.4.3, Corollary 12.5.2, Definition 12.8.1, Theorem 12.8.1, Theorem 12.8.2, Definition 12.8.2, Theorem 12.8.3, Theorem 12.8.5.
- **Chapter 13** — positive semidefiniteness, square roots, the Loewner order, the Schur product theorem, and the singular value and polar decompositions: Definition 13.1.1, Proposition 13.1.2, Theorem 13.1.4, Theorem 13.1.5, Theorem 13.2.1, Definition 13.5.1, Proposition 13.5.1, Lemma 13.5.2, Theorem 13.5.3, Example (Squaring does not preserve the Loewner order), Proposition 13.5.4, Theorem 13.6.1, Definition 13.8.1, Theorem 13.8.2, Definition 13.9.1, Theorem 13.9.1, Theorem 13.9.2, Theorem 13.10.1, Corollary 13.10.2, Definition 13.11.1, Theorem 13.11.3.
- **Chapter 15 §04** — the pairing Chapter 15 calls a **contraction**, a word §08 gives a different meaning: Definition 15.4.1 (named so that the clash is visible, not used).
- **Chapter 16** — norms, the operator norm, entrywise convergence, bounds for the exponential, and the derivative of a matrix-valued function: Definition 16.1.1, Example (The 1-, 2-, and ∞-norms on F^n), Corollary 16.2.4, Definition 16.3.2, Theorem 16.3.2, Theorem 16.3.3, Theorem 16.6.1, Definition 16.6.1.
- **Chapter 17** — the variational descriptions, Weyl's and Lidskii's inequalities, majorization, the Hermitian dilation, and the operator monotone functions: Corollary 17.2.4, Lemma 17.3.5, Corollary 17.3.6, Theorem 17.6.2, Corollary 17.6.5, Exercise (C2: Partial sums that behave like a norm), Theorem 17.7.2, Definition 17.8.1, Example (A chain, the two extremes, and size one), Proposition 17.8.1, Lemma 17.8.3, Lemma 17.8.6, Lemma 17.9.1, Theorem 17.9.2, Proposition 17.9.3, Theorem 17.9.4, Corollary 17.9.5, Theorem 17.9.6, Definition 17.11.1, Proposition 17.11.1, Proposition 17.11.2, Proposition 17.11.3, Proposition 17.11.4, Corollary 17.11.5, Proposition 17.11.8, Theorem 17.11.9.
- **Chapter 18** — convex sets and hulls, convex functions, Jensen's inequality, gauges and

duality, and the classical inequalities: Example (The vectors majorized by a fixed vector), Proposition 18.1.1, Lemma 18.1.2, Definition 18.1.4, Theorem 18.1.3, Definition 18.4.1, Definition 18.4.4, Proposition 18.4.6, Proposition 18.4.7, Theorem 18.4.8, Definition 18.8.3, Definition 18.8.4, Definition 18.10.1, Example (First convex functions), Theorem 18.10.2, Proposition 18.10.3, Lemma 18.10.5, Theorem 18.10.6, Lemma 18.11.1, Corollary 18.11.5, Theorem 18.11.6.

- **Chapter 19 §07** — Birkhoff's theorem and the half of Hardy–Littlewood–Pólya already proved: Theorem 19.7.2, Proposition 19.7.6.
- **Chapter 20** — the derivative of a simple eigenvalue, the norm toolkit behind Davis–Kahan, and the perturbation bounds this chapter upgrades: Corollary 20.5.4, Lemma 20.9.1, Theorem 20.10.1, Theorem 20.11.2.

21.0.3 Roadmap

- **Doubly stochastic maps.** T-transforms average two coordinates, and finitely many of them turn $\mathbf{y}$ into any $\mathbf{x} \prec \mathbf{y}$. That is the theorem of Hardy, Littlewood, and Pólya, and with Birkhoff's theorem it says that the majorization order is containment of permutation polytopes.
- **Convex functions of a majorization.** Karamata's inequality, its converse, Schur-convex and Schur-concave functions, and the convexity of $\mathbf{A} \mapsto \operatorname{tr} f(\mathbf{A})$ for every convex f.
- **Symmetric gauge functions.** Norms on $\mathbb{R}^n$ blind to the order and the signs of the coordinates. They are monotone under weak majorization, and they have a duality theory of their own.
- **Unitarily invariant norms.** The norms unchanged by unitary factors on either side: the Frobenius and spectral norms, the Ky Fan k-norms, and the Schatten p-norms. Von Neumann's correspondence matches them exactly with the symmetric gauges, through $|||\mathbf{A}||| = \Phi(\sigma(\mathbf{A}))$.
- **The trace inequality.** $|\operatorname{tr}(\mathbf{A}^*\mathbf{B})| \le \sum_i \sigma_i(\mathbf{A})\sigma_i(\mathbf{B})$, and the duality it gives between a unitarily invariant norm and the norm of the dual gauge.
- **Ky Fan dominance.** Weak majorization of singular values is equivalent to domination in every unitarily invariant norm. Eckart–Young and the nearest unitary matrix follow, in every such norm.
- **Lidskii in majorization form.** Chapter 17's majorization, run through the previous three sections, yields one bound for every convex function and one for every unitarily invariant norm, with Hoffman–Wielandt and Mirsky as special cases.
- **Operator convex functions.** The Hansen–Pedersen–Jensen inequality $f(\mathbf{V}^*\mathbf{A}\mathbf{V}) \preceq \mathbf{V}^* f(\mathbf{A})\mathbf{V}$ for a contraction $\mathbf{V}$, proved by dilating $\mathbf{V}$ to a unitary matrix, and the contrast with the far weaker trace convexity of §02.
- **Loewner matrices.** The derivative of f at a diagonal matrix is a Hadamard product with the matrix of divided differences. With the Schur product theorem this makes monotonicity of order n equivalent to positivity of every $n \times n$ Loewner matrix.
- **Loewner's theorem.** The three conditions of Chapter 17 §11 are equivalent. One implication quotes (A7); the rest is proved, including the integral machinery. The powers t^p are operator monotone exactly for $0 \le p \le 1$, $\log t$ is, and e^t is not.
- **Means and matrix Cauchy–Schwarz.** Cauchy–Schwarz and the arithmetic–geometric mean inequality for matrices, each holding in every unitarily invariant norm at once; then the matrix geometric mean $\mathbf{A}\#\mathbf{B}$, its surprising symmetry, and the chain from the harmonic through the geometric to the arithmetic mean in the Loewner order.
- **Golden–Thompson.** Lie's product formula, a trace inequality for powers of a product of two positive matrices, and $\operatorname{tr} e^{\mathbf{A}+\mathbf{B}} \le \operatorname{tr}(e^{\mathbf{A}} e^{\mathbf{B}})$. Three matrices break it. Then Klein's inequality and the relative entropy, Lieb's concavity theorem stated and not proved, and

a summary that closes the chapter and Part V.

21.0.4 Named moves

- **A majorization is a doubly stochastic map.** Turn the family of partial-sum inequalities into a single linear map, and let convexity act one row at a time.
- **A symmetric gauge turns a weak majorization into a norm inequality.** To prove one inequality for a whole family of norms, find the combinatorial statement about the underlying numbers that every member of the family respects.
- **A Hermitian dilation turns singular values into eigenvalues.** Every theorem about the eigenvalues of a Hermitian matrix has a singular-value twin, obtained by applying it to $\mathcal{H}(\mathbf{A})$.
- **The Schur product theorem turns a Loewner matrix into monotonicity.** Differentiate a matrix function, recognize the derivative as an entrywise product with a positive semidefinite matrix, and integrate along a segment.
- **A limit of Riemann sums carries the Loewner order through an integral.** An order defined by a closed convex cone survives every limit, so it survives integration.
- **Lie's product formula turns a sum in the exponent into a product.** Split one operation into k small steps, make the error of a single step $O(1/k^2)$, and let telescoping absorb the factor k.

21.1 Doubly Stochastic Maps

Chapter 17 §08 defined majorization and Chapter 19 §07 proved one half of the fact that makes it useful: a doubly stochastic matrix can only even a vector out, $\mathbf{Sy} \prec \mathbf{y}$. The other half was promised twice. Chapter 19 §07 ended by saying that "this converse of the proposition is part of the theorem of Hardy, Littlewood, and Pólya, which Chapter 21 takes up", and Chapter 19 §09 listed among its unfinished business "the theorem of Hardy, Littlewood, and Pólya, that $\mathbf{x} \prec \mathbf{y}$ exactly when $\mathbf{x} = \mathbf{Dy}$ for a doubly stochastic $\mathbf{D}$". This section pays both. The proof builds $\mathbf{D}$ out of one very simple move, repeated: average two coordinates and leave the rest alone.

Throughout, vectors are real, $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, and $\prec$ and $\prec_w$ are Chapter 17 §08's Definition 17.8.1. Recall $\mathbf{x}^{\downarrow}$, the decreasing rearrangement, and Ω_n, the set of doubly stochastic matrices of Definition 18.8.4: nonnegative entries, every row and every column adding up to 1. Recall also Proposition 19.7.6, the half already proved: if $\mathbf{S} \in \Omega_n$ and $\mathbf{y} \in \mathbb{R}^n$, then $\mathbf{Sy} \prec \mathbf{y}$.

21.1.1 What is still missing

Chapter 19 got the easy direction from Birkhoff's theorem in three lines. A doubly stochastic $\mathbf{S}$ is an average of permutation matrices, a permutation only rearranges $\mathbf{y}$, and majorization cannot see a rearrangement; so $\mathbf{Sy}$ is an average of vectors all majorized by $\mathbf{y}$, and the set of such vectors is convex.

Read backwards, that argument gives nothing. It starts from a matrix and produces a majorization. We now want to start from the majorization $\mathbf{x} \prec \mathbf{y}$, which is a list of $n-1$ inequalities about sorted partial sums together with one equality of totals, and manufacture a matrix. Nothing in Definition 17.8.1 mentions a matrix, so one has to be built.

One such matrix is in fact already available. Chapter 19 §07 observed that Theorem 17.8.5 realizes any $\mathbf{x} \prec \mathbf{y}$ as the diagonal of a Hermitian $\mathbf{A}$ with $\boldsymbol{\lambda}(\mathbf{A}) = \mathbf{y}^{\downarrow}$, while Lemma 19.7.5 writes that diagonal as $\mathbf{Sy}^{\downarrow}$ for a doubly stochastic $\mathbf{S}$; composing with the permutation matrix

that sorts $\mathbf{y}$ gives $\mathbf{x} = \mathbf{D}\mathbf{y}$ with $\mathbf{D} \in \Omega_n$. What that route does not give is any grip on $\mathbf{D}$, which arrives as the squared moduli of the entries of a unitary matrix supplied by an existence theorem. The construction below builds $\mathbf{D}$ by hand instead, as a product of at most $n-1$ factors, each of which averages a single pair of coordinates and fixes the rest.

Here is the smallest case, done by hand, to see what a construction might look like. Let $n = 2$ and $\mathbf{x} \prec \mathbf{y}$ with both decreasing, so $x_1 + x_2 = y_1 + y_2$ and $x_1 \le y_1$. Also $x_2 \ge y_2$, since subtracting $x_1 \le y_1$ from the equal totals reverses the inequality. So $y_2 \le x_1 \le y_1$, and x_1 is a weighted average of y_1 and y_2: if $y_1 = y_2$ then $\mathbf{x} = \mathbf{y}$, and otherwise

$$x_1 = (1-t)y_1 + ty_2, \qquad t = \frac{y_1 - x_1}{y_1 - y_2} \in [0,1].$$

The equal totals then force $x_2 = ty_1 + (1-t)y_2$. In matrix form,

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1-t & t \\ t & 1-t \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix},$$

and the matrix is doubly stochastic. So for $n = 2$ and sorted vectors the converse holds, with a $\mathbf{D}$ of a very special shape.

This is the same computation as Lemma 17.8.3, where the weights were c^2 and s^2 for a rotation and the averaged numbers were two eigenvalues. There the averaging was performed by conjugating a diagonal matrix; here it is performed on the vector directly. Both say: *any number between two others is an average of them*, and that single fact is the engine of this section.

21.1.2 Averaging two coordinates

The 2×2 matrix above can be planted inside $\mathbf{I}_n$, at any pair of positions. That gives the move we will repeat.

A T-transform mixes two coordinates and leaves every other coordinate alone.

Definition 21.1.1. T-transform

Let $n \ge 2$, let $1 \le j < k \le n$, and let $\mathbf{Q} \in M_n(\mathbb{R})$ be the permutation matrix (Definition 3.4.2) of the transposition interchanging j and k. For a real number t with $0 \le t \le 1$, the matrix

$$\mathbf{T} = (1-t)\mathbf{I} + t\mathbf{Q}$$

is called a **T-transform**, and we say that it **acts on the pair** $\{j, k\}$.

In words: $\mathbf{T}$ is the convex combination, with weights $1-t$ and t, of doing nothing and swapping two coordinates. Applied to a vector,

$$\begin{aligned} (\mathbf{T}\mathbf{y})_j &= (1-t)y_j + ty_k, \\ (\mathbf{T}\mathbf{y})_k &= ty_j + (1-t)y_k, \\ (\mathbf{T}\mathbf{y})_i &= y_i \qquad (i \ne j, k). \end{aligned} \tag{21.1.1}$$

The two moved entries are averages of the old y_j and y_k, with the **same** pair of weights used in both slots, exchanged. In particular $(\mathbf{T}\mathbf{y})_j + (\mathbf{T}\mathbf{y})_k = y_j + y_k$: a T-transform conserves the total, and indeed conserves the total of the pair it touches.

The first thing to check is that these matrices belong in the discussion at all.

Proposition 21.1.1. T-transforms Are Doubly Stochastic

Every T-transform lies in Ω_n.

Proof. Let $\mathbf{T} = (1-t)\mathbf{I} + t\mathbf{Q}$ with $0 \le t \le 1$, and let $\{j, k\}$ be the pair $\mathbf{Q}$ interchanges. In the two positions (j, j) and (k, k) the entry of $\mathbf{T}$ is $1-t$, in the two positions (j, k) and (k, j) it is t, in the remaining $n-2$ diagonal positions both $\mathbf{I}$ and $\mathbf{Q}$ carry a 1, so the entry is 1, and everywhere else it is 0. Each entry is therefore 1, $1-t$, t, or 0, hence nonnegative, since $t \ge 0$ and $1-t \ge 0$. A permutation matrix has exactly one 1 in each row and each column and zeros elsewhere (the remark after Definition 3.4.2), so $\mathbf{Q}\mathbf{1} = \mathbf{1}$ and $\mathbf{1}^\top\mathbf{Q} = \mathbf{1}^\top$; the same is true of $\mathbf{I}$, which is the permutation matrix of the identity permutation. Hence

$$\mathbf{T}\mathbf{1} = (1-t)\mathbf{1} + t\mathbf{1} = \mathbf{1}, \qquad \mathbf{1}^\top\mathbf{T} = (1-t)\mathbf{1}^\top + t\mathbf{1}^\top = \mathbf{1}^\top,$$

which says that every row and every column of $\mathbf{T}$ adds up to 1. By Definition 18.8.4, $\mathbf{T} \in \Omega_n$.

■

Example. Three T-transforms, and two impostors

Take $n = 3$ and $\mathbf{y} = (6, 3, 0)$.

(a) Compute $\mathbf{T}\mathbf{y}$ for the T-transform on $\{1, 3\}$ with $t = \frac{1}{3}$, and for the ones with $t = 0$ and $t = 1$.

(b) Is the matrix $\frac{1}{2}(\mathbf{I} + \mathbf{P})$, with $\mathbf{P}$ the permutation matrix of the 3-cycle $1 \mapsto 2 \mapsto 3 \mapsto 1$, a T-transform?

(c) Is $-\mathbf{I} + 2\mathbf{Q}$, with $\mathbf{Q}$ the transposition matrix of $\{1, 2\}$, a T-transform?

Solution.

(a) With $j = 1$, $k = 3$, and $t = \frac{1}{3}$, Eq. 21.1.1 gives

$$\mathbf{T}\mathbf{y} = (\tfrac{2}{3}\cdot 6 + \tfrac{1}{3}\cdot 0,\ 3,\ \tfrac{1}{3}\cdot 6 + \tfrac{2}{3}\cdot 0) = (4, 3, 2).$$

At $t = 0$, $\mathbf{T} = \mathbf{I}$ and $\mathbf{T}\mathbf{y} = \mathbf{y}$; at $t = 1$, $\mathbf{T} = \mathbf{Q}$ and $\mathbf{T}\mathbf{y} = (0, 3, 6)$. So the endpoints of the family are "do nothing" and "swap".

(b) No. It is doubly stochastic, by the computation in Proposition 21.1.1 with $\mathbf{Q}$ replaced by $\mathbf{P}$, but Definition 21.1.1 asks for the permutation matrix of a **transposition**. Concretely, $\frac{1}{2}(\mathbf{I} + \mathbf{P})$ changes all three coordinates of $\mathbf{y}$, while a T-transform fixes at least $n-2$ of them.

(c) No: $t = 2$ is outside $[0, 1]$, and the matrix has the entry -1 in position $(1, 1)$, so it is not even doubly stochastic. Applied to $\mathbf{y}$ it gives $(0, 9, 0)$, whose first two entries are further apart than 6 and 3: with $t \notin [0, 1]$ the "average" is not between the two numbers, and the move spreads rather than evens.

Quick check. Let $\mathbf{T}$ be the T-transform on $\{1, 2\}$ with $t = \frac{1}{4}$, and let $\mathbf{y} = (8, 4, 1)$. Compute $\mathbf{T}\mathbf{y}$, and verify $\mathbf{T}\mathbf{y} \prec \mathbf{y}$ directly from Definition 17.8.1.

Solution. $(\mathbf{Ty})_1 = \frac{3}{4}\cdot 8 + \frac{1}{4}\cdot 4 = 7$ and $(\mathbf{Ty})_2 = \frac{1}{4}\cdot 8 + \frac{3}{4}\cdot 4 = 5$, so $\mathbf{Ty} = (7, 5, 1)$. Both vectors are decreasing with total 13, and the running totals are $7, 12$ against $8, 12$. Since $7 \le 8$ and $12 \le 12$, (M1) holds, and (M2) is the equal totals. So $\mathbf{Ty} \prec \mathbf{y}$, as Proposition 19.7.6 already promised, since $\mathbf{T} \in \Omega_n$.

Warning. **A product of T-transforms is usually not a T-transform.** T-transforms are doubly stochastic, and so are their products (this is checked inside the proof of the main theorem below), but the product of two of them can move every coordinate. Averaging coordinates 1 and 2, then coordinates 2 and 3, changes all three. The theorem below produces a product, not a single T-transform.

21.1.3 One averaging step

Now the engine. Given $\mathbf{x} \prec \mathbf{y}$, we want to push $\mathbf{y}$ toward $\mathbf{x}$ by one T-transform, and we want a measure of progress that cannot stall. The measure is the number of coordinates in which the two vectors already agree.

Lemma 21.1.2. One Averaging Step

Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ be **decreasing**, with $\mathbf{x} \prec \mathbf{y}$ and $\mathbf{x} \neq \mathbf{y}$. Then there is a T-transform $\mathbf{T}$ such that $\mathbf{y}' = \mathbf{Ty}$ satisfies

(a) $\mathbf{y}'$ is decreasing;
(b) $\mathbf{x} \prec \mathbf{y}'$;
(c) $y'_i = x_i$ for strictly more indices i than $y_i = x_i$ does.

Idea. Both vectors are sorted, so compare them entry by entry. Somewhere $\mathbf{y}$ is too big and somewhere it is too small. Let j be the **last** place where $\mathbf{y}$ is too big and k the **first** place after j where it is too small; between them the two vectors already agree. Averaging coordinates j and k moves y_j down and y_k up by the same amount, and we move by exactly as much as the tighter of the two deficits allows, so that one of the two positions lands on $\mathbf{x}$ exactly. Two things then need checking, and they are not equally hard. The partial sums are the easy one: the only ones that drop are those between j and k, and they have slack at least $y_j - x_j$ to give away, while we give away no more than that. The delicate one is that $\mathbf{y}'$ must come out **decreasing**, since otherwise the lemma cannot be applied to it again; that is what taking j last buys, because it makes the entries between j and k agree with $\mathbf{x}$'s.

Proof. Write $X_m = x_1 + \cdots + x_m$ and $Y_m = y_1 + \cdots + y_m$, with $X_0 = Y_0 = 0$. Since both vectors are decreasing, Definition 17.8.1 says $X_m \le Y_m$ for $m = 1, \dots, n-1$ and $X_n = Y_n$.

Step 1: the two indices. Some index has $y_i > x_i$: otherwise $y_i \le x_i$ for every i, and since $\mathbf{x} \neq \mathbf{y}$ at least one inequality is strict, so $Y_n < X_n$, contradicting $X_n = Y_n$. Let j be the **largest** index with $y_j > x_j$.

Next, some index $i > j$ has $y_i < x_i$. Suppose not. By the choice of j, $y_i \le x_i$ for every $i > j$;

so $y_i = x_i$ for every $i > j$, and subtracting these from $X_n = Y_n$ gives $X_j = Y_j$. Then

$$X_{j-1} = X_j - x_j > Y_j - y_j = Y_{j-1},$$

because $x_j < y_j$. For $j \geq 2$ this contradicts $X_{j-1} \leq Y_{j-1}$, and for $j = 1$ it reads $0 > 0$, which is false. So we may let k be the **smallest** index with $k > j$ and $y_k < x_k$.

By the maximality of j we have $y_i \geq x_i$ for $j < i < k$, and by the minimality of k we have $y_i \leq x_i$ there. Hence

$$y_i = x_i \qquad (j < i < k). \tag{$*$}$$

Step 2: the transform. Put

$$\delta = \min\{\, y_j - x_j,\ x_k - y_k \,\} > 0.$$

Since $\mathbf{x}$ is decreasing and $j < k$, $x_j \geq x_k$, so

$$y_j > x_j \geq x_k > y_k,$$

and in particular $y_j - y_k > 0$. Also $\delta \leq y_j - x_j \leq y_j - y_k$, using $x_j \geq x_k > y_k$. So

$$t = \frac{\delta}{y_j - y_k}$$

satisfies $0 < t \leq 1$, and Definition 21.1.1 gives a T-transform $\mathbf{T}$ acting on $\{j, k\}$ with this t. By Eq. 21.1.1,

$$y'_j = y_j - t(y_j - y_k) = y_j - \delta, \qquad y'_k = y_k + \delta,$$

and $y'_i = y_i$ for every other i.

Step 3: (c), the count. By the choice of δ, either $\delta = y_j - x_j$, and then $y'_j = x_j$; or $\delta = x_k - y_k$, and then $y'_k = x_k$. So $\mathbf{y}'$ agrees with $\mathbf{x}$ in at least one of the positions j, k. Neither position agreed before, since $y_j > x_j$ and $y_k < x_k$. Every other position is untouched. Hence the number of agreements strictly increases.

Step 4: (a), the sorted order. Only positions j and k change, and $\delta > 0$ lowers y_j and raises y_k. From $\delta \leq y_j - x_j$ and $\delta \leq x_k - y_k$,

$$y'_j \geq x_j, \qquad y'_k \leq x_k.$$

If $j \geq 2$, then $y'_{j-1} = y_{j-1} \geq y_j > y'_j$. If $j + 1 < k$, then $y'_{j+1} = y_{j+1} = x_{j+1}$ by $(*)$, and $x_{j+1} \leq x_j \leq y'_j$. If $j + 1 = k$, then $y'_{j+1} = y'_k \leq x_k \leq x_j \leq y'_j$. Symmetrically, if $k - 1 > j$ then $y'_{k-1} = x_{k-1} \geq x_k \geq y'_k$, and if $k \leq n - 1$ then $y'_{k+1} = y_{k+1} \leq y_k < y'_k$. Every remaining consecutive pair consists of two untouched entries of the decreasing $\mathbf{y}$. So $\mathbf{y}'$ is decreasing.

Step 5: (b), the majorization. Write $Y'_m = y'_1 + \cdots + y'_m$. Since $\mathbf{y}'$ differs from $\mathbf{y}$ only at j and k, by $-\delta$ and $+\delta$,

$$Y'_m = \begin{cases} Y_m - \delta & (j \leq m \leq k - 1), \\ Y_m & (\text{otherwise}). \end{cases}$$

In particular $Y'_n = Y_n = X_n$, which is (M2). For $m < j$ and for $m \geq k$ we have $Y'_m = Y_m \geq X_m$, which is (M1) there. For $j \leq m \leq k - 1$, the identity $(*)$ gives $y_i = x_i$ for $j < i \leq m$, hence

$$Y_m - X_m = Y_j - X_j.$$

Now $Y_j - X_j = (Y_{j-1} - X_{j-1}) + (y_j - x_j) \geq y_j - x_j \geq \delta$, using $Y_{j-1} \geq X_{j-1}$ (which is (M1), or $0 \geq 0$ when $j = 1$) and the definition of δ. Therefore

$$Y'_m - X_m = (Y_m - X_m) - \delta = (Y_j - X_j) - \delta \geq 0,$$

which is (M1) at m. So $\mathbf{x} \prec \mathbf{y}'$. This proves the lemma.

■

The choice of the two indices is the only delicate part, and it is worth saying why the obvious alternative fails, and where. Suppose j were taken as the *smallest* index with $y_j > x_j$, keeping k as the smallest index after j with $y_k < x_k$. Step 5 would survive: the minimality of k still gives $y_i \geq x_i$ for $j < i < k$, so the slack $Y_m - X_m$ is nondecreasing on $j \leq m \leq k-1$ instead of constant, and it is therefore at least its value $Y_j - X_j \geq y_j - x_j \geq \delta$ at $m = j$. What breaks is Step 4. The entries strictly between j and k need no longer agree with $\mathbf{x}$'s, and lowering y_j by δ can push it below y_{j+1}, so that $\mathbf{y}'$ comes out unsorted and Lemma 21.1.2 cannot be applied to it a second time. Take $\mathbf{x} = (5, 5, 0)$ and $\mathbf{y} = (6, 6, -2)$, both decreasing, with $\mathbf{x} \prec \mathbf{y}$ since $5 \leq 6$, $10 \leq 12$, and both totals are 10. Choosing j smallest gives $j = 1$ and $k = 3$, so $\delta = \min\{6-5,\ 0-(-2)\} = 1$ and $\mathbf{y}' = (5, 6, -1)$, which is not decreasing; the rule of the lemma gives $j = 2$ instead, and $\mathbf{y}' = (6, 5, -1)$, which is. Taking j last and k first after it makes the interval between them a stretch where the two vectors already coincide, and that stretch is what protects the order.

21.1.4 The theorem

Theorem 21.1.3. Hardy–Littlewood–Pólya Theorem

Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$. Then

$$\mathbf{x} \prec \mathbf{y} \quad \Longleftrightarrow \quad \mathbf{x} = \mathbf{D}\mathbf{y} \ \text{ for some } \mathbf{D} \in \Omega_n.$$

Moreover, if $\mathbf{x}$ and $\mathbf{y}$ are both decreasing, $\mathbf{D}$ may be taken to be a product of at most $n-1$ T-transforms.

Idea. ($\Leftarrow$) is Chapter 19's Proposition 19.7.6 and needs nothing new.

For ($\Rightarrow$), sort both vectors first, since neither side of the equivalence sees the order. Then count agreements. Lemma 21.1.2 raises that count by at least one, keeps both hypotheses alive, and therefore may be applied again; the count is at most n, so the process stops, at $\mathbf{y}$ turned into $\mathbf{x}$. The bound $n-1$ comes from noticing that the count can never be exactly $n-1$: two vectors with equal totals that agree in all but one coordinate agree in that one too.

Proof. ($\Leftarrow$) If $\mathbf{x} = \mathbf{D}\mathbf{y}$ with $\mathbf{D} \in \Omega_n$, then $\mathbf{x} \prec \mathbf{y}$ by Proposition 19.7.6.

($\Rightarrow$) Suppose $\mathbf{x} \prec \mathbf{y}$.

Step 1: the decreasing case. Assume first that $\mathbf{x}$ and $\mathbf{y}$ are both decreasing. For $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$ write $\operatorname{dis}(\mathbf{u}, \mathbf{v})$ for the number of indices i with $u_i \neq v_i$. We claim:

Claim. If $\mathbf{x}$ and $\mathbf{y}$ are decreasing with $\mathbf{x} \prec \mathbf{y}$ and $d = \operatorname{dis}(\mathbf{x}, \mathbf{y})$, then $\mathbf{x} = \mathbf{T}_r \cdots \mathbf{T}_1 \mathbf{y}$ for some T-transforms $\mathbf{T}_1, \dots, \mathbf{T}_r$ with $r = 0$ if $d = 0$, and $r \le d - 1$ otherwise.

Proof. First, $d \neq 1$: if $\mathbf{x}$ and $\mathbf{y}$ agreed in all indices but one, say i_0, then subtracting those $n - 1$ equalities from the equal totals (M2) would give $x_{i_0} = y_{i_0}$ too, so $d = 0$. We induct on d.

If $d = 0$ then $\mathbf{x} = \mathbf{y}$, and the empty product, $r = 0$, works. Let $d \ge 2$, and assume the claim for every smaller value. By Lemma 21.1.2 there is a T-transform $\mathbf{T}_1$ with $\mathbf{y}' = \mathbf{T}_1 \mathbf{y}$ decreasing, $\mathbf{x} \prec \mathbf{y}'$, and $d' = \operatorname{dis}(\mathbf{x}, \mathbf{y}') \le d - 1$. If $d' = 0$, then $\mathbf{x} = \mathbf{T}_1 \mathbf{y}$ and $r = 1 \le d - 1$. Otherwise $d' \ge 2$, since $d' \neq 1$ by the first paragraph applied to $\mathbf{x}$ and $\mathbf{y}'$; the inductive hypothesis gives $\mathbf{x} = \mathbf{T}_r \cdots \mathbf{T}_2 \mathbf{y}'$ with $r - 1 \le d' - 1$, so $r \le d' \le d - 1$. This proves the claim.

■

Since $d \le n$, the claim gives $r \le d - 1 \le n - 1$ in every case. This proves the "moreover" clause.

Step 2: products stay doubly stochastic. If $\mathbf{A}, \mathbf{B} \in \Omega_n$, then $\mathbf{AB}$ has nonnegative entries, and $\mathbf{AB1} = \mathbf{A1} = \mathbf{1}$, $\mathbf{1}^\top \mathbf{AB} = \mathbf{1}^\top \mathbf{B} = \mathbf{1}^\top$; so $\mathbf{AB} \in \Omega_n$ by Definition 18.8.4. By induction, any finite product of doubly stochastic matrices is doubly stochastic. With Proposition 21.1.1, the matrix $\mathbf{D}_0 = \mathbf{T}_r \cdots \mathbf{T}_1$ of Step 1 lies in Ω_n.

Step 3: the general case. Let $\mathbf{x} \prec \mathbf{y}$ be arbitrary. Majorization sees only rearrangements (the remark after Definition 17.8.1), so $\mathbf{x}^\downarrow \prec \mathbf{y}^\downarrow$, and Step 1 gives $\mathbf{D}_0 \in \Omega_n$ with $\mathbf{x}^\downarrow = \mathbf{D}_0 \mathbf{y}^\downarrow$. Choose permutations $\sigma, \tau \in S_n$ with $\mathbf{y}^\downarrow = \mathbf{P}_\sigma \mathbf{y}$ and $\mathbf{x} = \mathbf{P}_\tau \mathbf{x}^\downarrow$; these exist because $\mathbf{y}^\downarrow$ is a rearrangement of $\mathbf{y}$ and $\mathbf{x}$ of $\mathbf{x}^\downarrow$. Each $\mathbf{P}_\sigma$ is doubly stochastic, having exactly one 1 in each row and each column and zeros elsewhere (the remark after Definition 3.4.2), so by Step 2

$$\mathbf{D} = \mathbf{P}_\tau \mathbf{D}_0 \mathbf{P}_\sigma \in \Omega_n, \qquad \mathbf{D}\mathbf{y} = \mathbf{P}_\tau \mathbf{D}_0 \mathbf{y}^\downarrow = \mathbf{P}_\tau \mathbf{x}^\downarrow = \mathbf{x}.$$

This proves the theorem.

■

Two remarks on what the theorem does and does not say. It converts a family of $n-1$ inequalities and one equality into a single algebraic identity $\mathbf{x} = \mathbf{D}\mathbf{y}$, and an identity is much easier to feed into another argument, as the next section does, one row of $\mathbf{D}$ at a time. And the $n - 1$ in the "moreover" clause is sharp; Exercise (C3: The count $n - 1$ cannot be improved) exhibits a pair needing that many.

Remark. The word **decreasing** in the "moreover" clause is not decoration. For $\mathbf{y} = (6, 4, 2, 0)$ and the unsorted $\mathbf{x} = (\frac{3}{2}, 1, \frac{9}{2}, 5)$ we have $\mathbf{x} \prec \mathbf{y}$, since $\mathbf{x}^\downarrow = (5, \frac{9}{2}, \frac{3}{2}, 1)$ has running totals $5, \frac{19}{2}, 11$ against $6, 10, 12$ and both totals are 12; yet no product of three T-transforms equals $\mathbf{x}$, since for each of the finitely many choices of the three pairs the best the three parameters can do still misses it. Four transforms suffice: the transforms on $\{1,2\}$, $\{3,4\}$, $\{1,3\}$, $\{2,4\}$ with $t = \frac{1}{2}, \frac{1}{2}, \frac{7}{8}, 1$ carry $\mathbf{y}$ through $(5, 5, 2, 0)$, $(5, 5, 1, 1)$, and $(\frac{3}{2}, 5, \frac{9}{2}, 1)$ to $\mathbf{x}$. Step 3 of the proof handles an unsorted pair with permutations instead, and a permutation matrix is a T-transform only when it is a transposition.

Example. Four coordinates, two transforms

Let $\mathbf{y} = (7, 5, 3, 1)$ and $\mathbf{x} = (5, 5, 4, 2)$. Verify $\mathbf{x} \prec \mathbf{y}$, run the construction of Theorem 21.1.3, and write out the resulting $\mathbf{D}$.

Solution. Both vectors are decreasing with total 16, and the running totals are $5, 10, 14$ for $\mathbf{x}$ against $7, 12, 15$ for $\mathbf{y}$. So (M1) and (M2) hold and $\mathbf{x} \prec \mathbf{y}$.

First step. Here $\mathbf{y} - \mathbf{x} = (2, 0, -1, -1)$, so the only index with $y_i > x_i$ is $i = 1$, giving $j = 1$; the smallest index above it with $y_i < x_i$ is $k = 3$. Then

$$\delta = \min\{7 - 5,\ 4 - 3\} = 1, \qquad t = \frac{1}{7 - 3} = \tfrac{1}{4},$$

and $\mathbf{T}_1$ is the T-transform on $\{1, 3\}$ with $t = \frac{1}{4}$:

$$\mathbf{y}' = \mathbf{T}_1\mathbf{y} = (\tfrac{3}{4} \cdot 7 + \tfrac{1}{4} \cdot 3,\ 5,\ \tfrac{1}{4} \cdot 7 + \tfrac{3}{4} \cdot 3,\ 1) = (6, 5, 4, 1).$$

Now $\mathbf{y}'$ agrees with $\mathbf{x}$ in positions 2 and 3, where before it agreed only in position 2.

Second step. $\mathbf{y}' - \mathbf{x} = (1, 0, 0, -1)$, so $j = 1$ and $k = 4$, with $\delta = \min\{1, 1\} = 1$ and $t = 1/(6 - 1) = \frac{1}{5}$. The T-transform $\mathbf{T}_2$ on $\{1, 4\}$ gives

$$\mathbf{T}_2\mathbf{y}' = (\tfrac{4}{5} \cdot 6 + \tfrac{1}{5} \cdot 1,\ 5,\ 4,\ \tfrac{1}{5} \cdot 6 + \tfrac{4}{5} \cdot 1) = (5, 5, 4, 2) = \mathbf{x}.$$

Two transforms, and $2 \le n - 1 = 3$.

The matrix. Multiplying $\mathbf{T}_2\mathbf{T}_1$ row by row,

$$\mathbf{D} = \mathbf{T}_2\mathbf{T}_1 = \begin{pmatrix} \frac{3}{5} & 0 & \frac{1}{5} & \frac{1}{5} \\ 0 & 1 & 0 & 0 \\ \frac{1}{4} & 0 & \frac{3}{4} & 0 \\ \frac{3}{20} & 0 & \frac{1}{20} & \frac{4}{5} \end{pmatrix}.$$

Every row adds up to 1, and so does every column: the first is $\frac{3}{5} + \frac{1}{4} + \frac{3}{20} = \frac{12+5+3}{20} = 1$, the third is $\frac{1}{5} + \frac{3}{4} + \frac{1}{20} = \frac{4+15+1}{20} = 1$, and the second and fourth are visibly 1. Finally

$$\mathbf{D}\mathbf{y} = (\tfrac{21+3+1}{5},\ 5,\ \tfrac{7+9}{4},\ \tfrac{21+3+16}{20}) = (5, 5, 4, 2) = \mathbf{x}.$$

Note that $\mathbf{D}$ is not a T-transform: it moves three of the four coordinates.

Warning. **The matrix $\mathbf{D}$ is far from unique, and "doubly stochastic" cannot be relaxed to "row-stochastic".** For $\mathbf{y} = (2, 1, 0)$ and $\mathbf{x} = (1, 1, 1)$, both

$$\mathbf{D}_1 = \tfrac{1}{3}\mathbf{J}, \qquad \mathbf{D}_2 = \begin{pmatrix} 0 & 1 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} \end{pmatrix}$$

are doubly stochastic with $\mathbf{D}_i\mathbf{y} = \mathbf{x}$. And if only the **rows** are required to add up to 1, the conclusion fails: with $\mathbf{y} = (1, 0)$ and the row-stochastic $\mathbf{S} = \left(\begin{smallmatrix} 1 & 0 \\ 1 & 0 \end{smallmatrix}\right)$, $\mathbf{S}\mathbf{y} = (1, 1)$, whose total is $2 \neq 1$, so $\mathbf{S}\mathbf{y} \not\prec \mathbf{y}$. The column condition is what enforces the equal totals.

21.1.5 Majorization as a polytope

Birkhoff's theorem turns the doubly stochastic matrix of Theorem 21.1.3 into an average of permutations, and the statement becomes geometric.

Corollary 21.1.4. Majorization Is Membership in a Permutation Polytope

Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$. Then

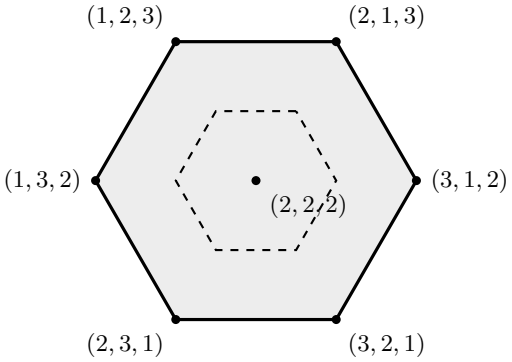
$$\mathbf{x} \prec \mathbf{y} \quad \Longleftrightarrow \quad \mathbf{x} \in \operatorname{conv}\{\mathbf{P}_\sigma \mathbf{y} : \sigma \in S_n\}.$$

Proof. ($\Rightarrow$) By Theorem 21.1.3, $\mathbf{x} = \mathbf{D}\mathbf{y}$ with $\mathbf{D} \in \Omega_n$. By Theorem 19.7.2, $\mathbf{D} \in \operatorname{conv}\{\mathbf{P}_\sigma : \sigma \in S_n\}$, so by Theorem 18.1.3, $\mathbf{D} = \sum_{i=1}^{m} t_i \mathbf{P}_{\sigma_i}$ with $t_i \geq 0$ and $\sum_i t_i = 1$. Hence $\mathbf{x} = \sum_i t_i \mathbf{P}_{\sigma_i}\mathbf{y}$, a convex combination of the vectors $\mathbf{P}_\sigma \mathbf{y}$, which lies in their convex hull by Theorem 18.1.3.

($\Leftarrow$) By Theorem 18.1.3, $\mathbf{x} = \sum_i t_i \mathbf{P}_{\sigma_i}\mathbf{y}$ with $(t_1, \ldots, t_m)$ a list of nonnegative weights adding up to 1. Put $\mathbf{D} = \sum_i t_i \mathbf{P}_{\sigma_i}$. Its entries are nonnegative, and $\mathbf{D}\mathbf{1} = \sum_i t_i \mathbf{1} = \mathbf{1}$, $\mathbf{1}^\top \mathbf{D} = \mathbf{1}^\top$, so $\mathbf{D} \in \Omega_n$ and $\mathbf{x} = \mathbf{D}\mathbf{y}$. By Proposition 19.7.6, $\mathbf{x} \prec \mathbf{y}$.

■

The set $\operatorname{conv}\{\mathbf{P}_\sigma \mathbf{y}\}$ is the **permutation polytope** of $\mathbf{y}$: the convex hull of the $n!$ rearrangements of $\mathbf{y}$. It lies in the affine subspace where the coordinate sum equals $\sum_i y_i$, so its dimension is at most $n - 1$. For $n = 3$ and distinct entries it is a hexagon.



The solid hexagon is the permutation polytope of $\mathbf{y} = (3, 2, 1)$, drawn inside the plane $x_1 + x_2 + x_3 = 6$; the dashed one, at half the size, is the polytope of $(\frac{5}{2}, 2, \frac{3}{2})$. By Corollary 21.1.4, $\mathbf{x} \prec \mathbf{y}$ says that $\mathbf{x}$ lies in the solid hexagon, and $\mathbf{x} \prec (\frac{5}{2}, 2, \frac{3}{2}) \prec \mathbf{y}$ says that the dashed hexagon sits inside the solid one. *The majorization order is containment of permutation polytopes.* The center $(2, 2, 2)$, the flat vector, lies in every one of them, which is Example (A chain, the two extremes, and size one) (b) again.

Quick check. Use Corollary 21.1.4 to give a second proof that $\{\mathbf{x} : \mathbf{x} \prec \mathbf{y}\}$ is convex, a fact Chapter 18 proved by hand in Example (The vectors majorized by a fixed vector).

Solution. By the corollary, $\{\mathbf{x} : \mathbf{x} \prec \mathbf{y}\} = \operatorname{conv}\{\mathbf{P}_\sigma \mathbf{y} : \sigma \in S_n\}$. A convex hull is convex (Theorem 18.1.3, or Definition 18.1.4 directly). So the set is convex.

21.1.6 Exercises

21.1.6.1 A. Check your understanding

Exercise. A1

(a) Define a T-transform, and say which coordinates of a vector it can change.

(b) State the Hardy–Littlewood–Pólya theorem in full.

(c) True or false: if $\mathbf{x} = \mathbf{S}\mathbf{y}$ with $\mathbf{S}$ row-stochastic (nonnegative entries, every row adding up to 1), then $\mathbf{x} \prec \mathbf{y}$. Justify your answer.

(d) True or false: if $\mathbf{x} \prec \mathbf{y}$, the doubly stochastic $\mathbf{D}$ with $\mathbf{x} = \mathbf{D}\mathbf{y}$ is unique. Justify your answer.

(e) In Lemma 21.1.2, why is j chosen as the **largest** index with $y_j > x_j$ rather than the smallest?

(f) Why can the number of coordinates in which two decreasing vectors $\mathbf{x} \prec \mathbf{y}$ agree never be exactly $n-1$?

Solution.

(a) For $1 \le j < k \le n$ and $0 \le t \le 1$, it is $\mathbf{T} = (1-t)\mathbf{I} + t\mathbf{Q}$ with $\mathbf{Q}$ the permutation matrix of the transposition of j and k (Definition 21.1.1). It changes only the coordinates j and k, replacing each by an average of the two old values.

(b) For $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, $\mathbf{x} \prec \mathbf{y}$ if and only if $\mathbf{x} = \mathbf{D}\mathbf{y}$ for some doubly stochastic $\mathbf{D}$; and when $\mathbf{x}$ and $\mathbf{y}$ are both decreasing, $\mathbf{D}$ may be taken to be a product of at most $n-1$ T-transforms (Theorem 21.1.3).

(c) False. With $\mathbf{y} = (1, 0)$ and $\mathbf{S} = \left(\begin{smallmatrix} 1 & 0 \\ 1 & 0 \end{smallmatrix}\right)$, $\mathbf{S}\mathbf{y} = (1, 1)$ has total 2 while $\mathbf{y}$ has total 1, so (M2) fails. Row sums alone control nothing but the individual entries; the column sums are what preserve the total.

(d) False. For $\mathbf{y} = (2, 1, 0)$ and $\mathbf{x} = (1, 1, 1)$, both $\frac{1}{3}\mathbf{J}$ and the matrix $\mathbf{D}_2$ of the warning after Example (Four coordinates, two transforms) send $\mathbf{y}$ to $\mathbf{x}$.

(e) Because it makes the entries strictly between j and k agree with $\mathbf{x}$'s, which is what Step 4 needs in order to keep $\mathbf{y}'$ decreasing. The majorization would survive the other choice: with j smallest, the slack $Y_m - X_m$ is nondecreasing on $j \le m \le k-1$ rather than constant, hence still at least δ throughout, and Step 5 goes through. The sorted order is what fails, and with it the possibility of applying the lemma again. For $\mathbf{x} = (5, 5, 0)$ and $\mathbf{y} = (6, 6, -2)$, taking j smallest gives $j = 1$, $k = 3$, $\delta = 1$, and $\mathbf{y}' = (5, 6, -1)$, which is not decreasing, while the rule of the lemma gives $j = 2$ and $\mathbf{y}' = (6, 5, -1)$, which is.

(f) Suppose they agree in all indices but i_0. Subtracting the $n-1$ equalities $x_i = y_i$ from the equal totals of (M2) leaves $x_{i_0} = y_{i_0}$, so in fact they agree everywhere.

21.1.6.2 B. Practice

Exercise. B1: Run the construction

Let $\mathbf{y} = (9, 5, 4)$ and $\mathbf{x} = (6, 6, 6)$. Verify $\mathbf{x} \prec \mathbf{y}$, produce T-transforms $\mathbf{T}_1, \mathbf{T}_2$ with $\mathbf{x} = \mathbf{T}_2\mathbf{T}_1\mathbf{y}$ by the recipe of Lemma 21.1.2, and write out $\mathbf{D} = \mathbf{T}_2\mathbf{T}_1$. Hence check that $\mathbf{D}$ is

doubly stochastic.

Solution. Both vectors are decreasing with total 18, and the running totals are $6, 12$ against $9, 14$. So (M1) and (M2) hold and $\mathbf{x} \prec \mathbf{y}$.

First step. $\mathbf{y} - \mathbf{x} = (3, -1, -2)$, so $j = 1$ and $k = 2$. Then $\delta = \min\{3, 1\} = 1$ and $t = 1/(9-5) = \frac{1}{4}$, and $\mathbf{T}_1$ acts on $\{1, 2\}$:

$$\mathbf{y}' = (\tfrac{3}{4} \cdot 9 + \tfrac{1}{4} \cdot 5,\ \tfrac{1}{4} \cdot 9 + \tfrac{3}{4} \cdot 5,\ 4) = (8, 6, 4).$$

Second step. $\mathbf{y}' - \mathbf{x} = (2, 0, -2)$, so $j = 1$, $k = 3$, $\delta = \min\{2, 2\} = 2$, and $t = 2/(8-4) = \frac{1}{2}$. Then

$$\mathbf{T}_2 \mathbf{y}' = (\tfrac{1}{2} \cdot 8 + \tfrac{1}{2} \cdot 4,\ 6,\ \tfrac{1}{2} \cdot 8 + \tfrac{1}{2} \cdot 4) = (6, 6, 6) = \mathbf{x}.$$

The matrix. $\mathbf{T}_1$ has rows $(\frac{3}{4}, \frac{1}{4}, 0)$, $(\frac{1}{4}, \frac{3}{4}, 0)$, $(0, 0, 1)$, and $\mathbf{T}_2$ has rows $(\frac{1}{2}, 0, \frac{1}{2})$, $(0, 1, 0)$, $(\frac{1}{2}, 0, \frac{1}{2})$. Multiplying,

$$\mathbf{D} = \mathbf{T}_2 \mathbf{T}_1 = \begin{pmatrix} \frac{3}{8} & \frac{1}{8} & \frac{1}{2} \\ \frac{1}{4} & \frac{3}{4} & 0 \\ \frac{3}{8} & \frac{1}{8} & \frac{1}{2} \end{pmatrix}.$$

Each row adds up to 1, and the columns add up to $\frac{3}{8} + \frac{1}{4} + \frac{3}{8} = 1$, $\frac{1}{8} + \frac{3}{4} + \frac{1}{8} = 1$, and $\frac{1}{2} + 0 + \frac{1}{2} = 1$. So $\mathbf{D} \in \Omega_3$, and $\mathbf{D}\mathbf{y} = (\frac{27}{8} + \frac{5}{8} + 2,\ \frac{9}{4} + \frac{15}{4},\ \frac{27}{8} + \frac{5}{8} + 2) = (6, 6, 6)$.

Exercise. B2: Determine which are majorizations

For each pair, determine whether $\mathbf{x} \prec \mathbf{y}$. Where it holds, exhibit a doubly stochastic $\mathbf{D}$ with $\mathbf{x} = \mathbf{D}\mathbf{y}$; where it fails, name the clause of Definition 17.8.1 and the value of k at which it fails. Justify your answer.

(a) $\mathbf{x} = (2, 2, 2)$, $\mathbf{y} = (4, 1, 1)$.

(b) $\mathbf{x} = (3, 1, 0)$, $\mathbf{y} = (2, 2, 0)$.

(c) $\mathbf{x} = (1, 3, 0)$, $\mathbf{y} = (2, 0, 2)$.

(d) $\mathbf{x} = (1, 1)$, $\mathbf{y} = (2, 1)$.

Solution.

(a) Both decreasing, totals $6 = 6$; running totals $2, 4$ against $4, 5$. So $\mathbf{x} \prec \mathbf{y}$. Here $\mathbf{x}$ is the flat vector, and $\mathbf{D} = \frac{1}{3}\mathbf{J}$ works: every entry of $\frac{1}{3}\mathbf{J}\mathbf{y}$ is $\frac{1}{3}(4 + 1 + 1) = 2$.

(b) Totals $4 = 4$, but at $k = 1$ the sorted first entries are $3 > 2$, so (M1) fails at $k = 1$ and $\mathbf{x} \not\prec \mathbf{y}$.

(c) Sorted, $\mathbf{x}^{\downarrow} = (3, 1, 0)$ and $\mathbf{y}^{\downarrow} = (2, 2, 0)$; this is (b) again, so $\mathbf{x} \not\prec \mathbf{y}$, with (M1) failing at $k = 1$.

(d) Totals $2 \neq 3$, so (M2) fails and $\mathbf{x} \not\prec \mathbf{y}$. (Both running totals do satisfy (M1), so $\mathbf{x} \prec_w \mathbf{y}$.)

Exercise. B3: Into the polytope

Let $\mathbf{y} = (4, 2, 0)$ and $\mathbf{x} = (2, 2, 2)$. Using Corollary 21.1.4, write $\mathbf{x}$ explicitly as a convex combination of rearrangements of $\mathbf{y}$, in two different ways.

Solution. The six rearrangements of $\mathbf{y}$ are the vectors with entries $4, 2, 0$ in some order. Averaging the three cyclic shifts,

$$\tfrac{1}{3}[(4,2,0) + (0,4,2) + (2,0,4)] = \tfrac{1}{3}(6,6,6) = (2,2,2) = \mathbf{x}.$$

Alternatively, average a vector with its reversal:

$$\tfrac{1}{2}[(4,2,0) + (0,2,4)] = \tfrac{1}{2}(4,4,4) = (2,2,2) = \mathbf{x}.$$

Both are convex combinations, with weights $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ and $(\frac{1}{2}, \frac{1}{2})$. This is the nonuniqueness of the warning after Example (Four coordinates, two transforms), seen on the polytope: $\mathbf{x}$ is the center of the hexagon, lying strictly inside it rather than on an edge, and such a point has many representations. (The hexagon itself has empty interior in $\mathbb{R}^3$, since it lies in the plane $x_1 + x_2 + x_3 = 6$; "inside" here means inside the hexagon, within that plane.)

21.1.6.3 C. Going deeper

Exercise. C1: Row-stochastic is not enough, twice over

Let $\mathbf{S} \in M_n(\mathbb{R})$ be **row-stochastic**: every entry is nonnegative and every row adds up to 1.

(a) Give such an $\mathbf{S}$ and a $\mathbf{y} \in \mathbb{R}^2$ with $\mathbf{Sy} \not\prec \mathbf{y}$.

(b) Prove that if $\mathbf{Sy} \prec \mathbf{y}$ for **every** $\mathbf{y} \in \mathbb{R}^n$, then $\mathbf{S}$ is doubly stochastic.

Hint for (b): test the hypothesis on the standard basis vectors.

Solution.

(a) Take $\mathbf{S} = \left(\begin{smallmatrix} 1 & 0 \\ 1 & 0 \end{smallmatrix}\right)$ and $\mathbf{y} = (1, 0)$. Then $\mathbf{Sy} = (1, 1)$, whose total is 2, while $\mathbf{y}$ has total 1. So (M2) fails.

(b) Apply the hypothesis with $\mathbf{y} = \mathbf{e}_j$. The vector $\mathbf{Se}_j$ is the j-th column of $\mathbf{S}$, by Definition 0.9.9. From $\mathbf{Se}_j \prec \mathbf{e}_j$, clause (M2) gives

$$\sum_{i=1}^{n} s_{ij} = \sum_{i=1}^{n} (\mathbf{e}_j)_i = 1.$$

So every column of $\mathbf{S}$ adds up to 1. The entries are nonnegative and the rows add up to 1 by hypothesis, so $\mathbf{S}$ is doubly stochastic by Definition 18.8.4.

Exercise. C2: Transitivity, a second time

(a) Prove that if $\mathbf{A}, \mathbf{B} \in \Omega_n$ then $\mathbf{AB} \in \Omega_n$.

(b) Hence deduce that $\mathbf{x} \prec \mathbf{y}$ and $\mathbf{y} \prec \mathbf{z}$ imply $\mathbf{x} \prec \mathbf{z}$, and compare this route with the

proof of Proposition 17.8.1 (b).

Solution.

(a) The entries of $\mathbf{AB}$ are $\sum_k a_{ik}b_{kj}$, sums of products of nonnegative numbers, hence nonnegative. Since $\mathbf{A1} = \mathbf{1}$ and $\mathbf{B1} = \mathbf{1}$, we get $(\mathbf{AB})\mathbf{1} = \mathbf{A}(\mathbf{B1}) = \mathbf{A1} = \mathbf{1}$, so every row of $\mathbf{AB}$ adds up to 1. Since $\mathbf{1}^\top\mathbf{A} = \mathbf{1}^\top$ and $\mathbf{1}^\top\mathbf{B} = \mathbf{1}^\top$, we get $\mathbf{1}^\top(\mathbf{AB}) = (\mathbf{1}^\top\mathbf{A})\mathbf{B} = \mathbf{1}^\top\mathbf{B} = \mathbf{1}^\top$, so every column does too. By Definition 18.8.4, $\mathbf{AB} \in \Omega_n$.

(b) By Theorem 21.1.3 there are $\mathbf{D}, \mathbf{E} \in \Omega_n$ with $\mathbf{x} = \mathbf{Dy}$ and $\mathbf{y} = \mathbf{Ez}$. Then $\mathbf{x} = \mathbf{DEz}$, and $\mathbf{DE} \in \Omega_n$ by (a). By Theorem 21.1.3 again, $\mathbf{x} \prec \mathbf{z}$.

Proposition 17.8.1 (b) proved the same thing in one line, by chaining the inequalities $X_k \le Y_k \le Z_k$ between running totals. That proof is shorter and uses nothing. The route here is longer, but it explains transitivity rather than verifying it: a majorization is an averaging, and averaging twice is averaging.

Exercise. C3: The count $n-1$ cannot be improved

Let $n \ge 2$, $\mathbf{y} = (n, 0, \dots, 0)$, and $\mathbf{x} = (1, 1, \dots, 1)$ in $\mathbb{R}^n$. For $\mathbf{z} \in \mathbb{R}^n$, write $\operatorname{supp}(\mathbf{z}) = \{i : z_i \neq 0\}$.

(a) Show that $\mathbf{x} \prec \mathbf{y}$.

(b) Let $\mathbf{T}$ be a T-transform. Prove that $|\operatorname{supp}(\mathbf{Tz})| \le |\operatorname{supp}(\mathbf{z})| + 1$ for every $\mathbf{z} \in \mathbb{R}^n$.

(c) Deduce that no product of fewer than $n-1$ T-transforms sends $\mathbf{y}$ to $\mathbf{x}$.

Solution.

(a) Both vectors are decreasing with total n. Every running total of $\mathbf{y}$ is n, and the k-th running total of $\mathbf{x}$ is $k \le n$. So (M1) and (M2) hold.

(b) Let $\mathbf{T}$ act on $\{j, k\}$ with parameter t. By Eq. 21.1.1, $(\mathbf{Tz})_i = z_i$ for $i \notin \{j, k\}$, so

$$\operatorname{supp}(\mathbf{Tz}) \subseteq \operatorname{supp}(\mathbf{z}) \cup \{j, k\}.$$

If both $j, k \in \operatorname{supp}(\mathbf{z})$, the right side is $\operatorname{supp}(\mathbf{z})$ and there is nothing to prove. If neither is, then $z_j = z_k = 0$, so $(\mathbf{Tz})_j = (\mathbf{Tz})_k = 0$ and again $\operatorname{supp}(\mathbf{Tz}) \subseteq \operatorname{supp}(\mathbf{z})$. In the remaining case exactly one of them, say j, lies in $\operatorname{supp}(\mathbf{z})$, and then $\operatorname{supp}(\mathbf{Tz}) \subseteq \operatorname{supp}(\mathbf{z}) \cup \{k\}$, a set with one more element. In all cases the bound holds.

(c) Suppose $\mathbf{x} = \mathbf{T}_r \cdots \mathbf{T}_1\mathbf{y}$. Put $\mathbf{z}^{(0)} = \mathbf{y}$ and $\mathbf{z}^{(i)} = \mathbf{T}_i\mathbf{z}^{(i-1)}$. Then $|\operatorname{supp}(\mathbf{z}^{(0)})| = 1$ and $|\operatorname{supp}(\mathbf{z}^{(r)})| = |\operatorname{supp}(\mathbf{x})| = n$. By (b), the size of the support grows by at most 1 at each step, so

$$n \le 1 + r, \qquad \text{that is,} \qquad r \ge n - 1.$$

Together with Theorem 21.1.3, which gives such a product with $r \le n-1$, exactly $n-1$ transforms are needed for this pair.

21.2 Convex Functions of a Majorization

Section 1 turned the n inequalities of a majorization into a single identity, $\mathbf{x} = \mathbf{D}\mathbf{y}$ with $\mathbf{D}$ doubly stochastic. This section spends it. Chapter 17 §07 left to this chapter "the general consequences of a majorization $\mathbf{x} \prec \mathbf{y}$, namely $\sum_i \phi(x_i) \le \sum_i \phi(y_i)$ for every convex $\phi\colon \mathbb{R} \to \mathbb{R}$, ..."; that is the first theorem below, and its converse says that the convex functions see nothing but the majorization. Two more promises are paid from it: Chapter 17 §08's claim that for nonnegative vectors "$\mathbf{x} \prec \mathbf{y}$ implies $x_1 \cdots x_n \ge y_1 \cdots y_n$", and Chapters 18 §10 and §11's claim that "$\mathbf{A} \mapsto \operatorname{tr} f(\mathbf{A}) = \sum_i f(\lambda_i(\mathbf{A}))$ is convex on the Hermitian matrices with spectrum in that interval" for every convex f.

Conventions. Vectors are real and $\prec$, $\prec_w$ are Definition 17.8.1. A **convex** function is one in the sense of Definition 18.10.1, and an interval of $\mathbb{R}$ is a convex subset of $\mathbb{R}$, since it contains every number lying between two of its points. In the last subsection $F = \mathbb{R}$ or $\mathbb{C}$, matrices in $M_n(F)$ written $\mathbf{A}, \mathbf{B}, \mathbf{C}$ are **Hermitian**, and $f(\mathbf{A})$ is the spectral functional calculus of Chapter 12 §08, Definition 12.8.2. There $\mathbf{P}_j$, with an integer subscript, is an orthogonal projection of a spectral resolution; $\mathbf{P}_\sigma$ always means a permutation matrix. No analysis is imported here: the only facts used from outside linear algebra reach us through Chapter 18's already proved results.

21.2.1 Karamata's inequality

A majorization is a statement about sums of the largest entries. A convex function is one whose graph lies below its chords. The link between them is that an averaging matrix turns one into the other, and §01 says that a majorization *is* an averaging matrix.

Theorem 21.2.1. Karamata's Inequality

Let $I \subseteq \mathbb{R}$ be an interval, let $\phi\colon I \to \mathbb{R}$ be **convex**, and let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ satisfy $\mathbf{x} \prec \mathbf{y}$, with every entry of $\mathbf{y}$ lying in I. Then every entry of $\mathbf{x}$ lies in I, and

$$\sum_{i=1}^{n} \phi(x_i) \;\le\; \sum_{i=1}^{n} \phi(y_i).$$

Idea. Write $\mathbf{x} = \mathbf{D}\mathbf{y}$ with $\mathbf{D}$ doubly stochastic. Each x_i is then a weighted average of the y_j, with the weights in row i of $\mathbf{D}$; Jensen's inequality bounds $\phi(x_i)$ by the same average of the $\phi(y_j)$. Add these n inequalities. The coefficient of $\phi(y_j)$ on the right is the sum of column j of $\mathbf{D}$, which is 1. The rows and the columns do two different jobs, and this is where "doubly" earns its keep.

Proof. By Theorem 21.1.3 there is a doubly stochastic $\mathbf{D} = (d_{ij})$ with $\mathbf{x} = \mathbf{D}\mathbf{y}$, that is,

$$x_i = \sum_{j=1}^{n} d_{ij} y_j \qquad (i = 1, \dots, n).$$

Fix i. The numbers $d_{i1}, \dots, d_{in}$ are nonnegative and add up to 1, since row i of $\mathbf{D}$ is a row of a doubly stochastic matrix (Definition 18.8.4). So x_i is a convex combination of $y_1, \dots, y_n$, all of which lie in the convex set I; by Lemma 18.1.2, $x_i \in I$. Now Theorem 18.10.2, applied

to the convex ϕ on I with the points $y_1, \dots, y_n$ and the weights $d_{i1}, \dots, d_{in}$, gives

$$\phi(x_i) = \phi\Big(\sum_j d_{ij} y_j\Big) \;\le\; \sum_{j=1}^{n} d_{ij}\, \phi(y_j).$$

Summing over i and exchanging the order of the two finite sums,

$$\sum_{i=1}^{n} \phi(x_i) \;\le\; \sum_{j=1}^{n} \Big(\sum_{i=1}^{n} d_{ij}\Big) \phi(y_j) = \sum_{j=1}^{n} \phi(y_j),$$

because every column of $\mathbf{D}$ adds up to 1. This proves the theorem.

■

The proof is three lines and uses Theorem 21.1.3 once. That is what §01 was for: without it we would be comparing sorted partial sums with the values of an arbitrary convex function, and there is no route from one to the other.

Example. One majorization, four inequalities

Let $\mathbf{y} = (5, 3, 1, 1)$ and $\mathbf{x} = (3, 3, 2, 2)$. Check that $\mathbf{x} \prec \mathbf{y}$, and write down what Theorem 21.2.1 gives for $\phi(t) = t^2$, for $\phi(t) = |t - 2|$, and for $\phi(t) = \max\{t - 2, 0\}$.

Solution. Both vectors are decreasing with total 10, and the running totals are $3, 6, 8$ for $\mathbf{x}$ against $5, 8, 9$ for $\mathbf{y}$. So (M1) and (M2) hold, and $\mathbf{x} \prec \mathbf{y}$.

All three functions are convex on $I = \mathbb{R}$: t^2 by Lemma 18.10.5, since its derivative $2t$ is increasing, and the other two as pointwise maxima of affine functions, by Proposition 18.10.3 ($|t - 2| = \max\{t - 2,\ 2 - t\}$). So:

$$\begin{aligned} \textstyle\sum_i x_i^2 = 9 + 9 + 4 + 4 = 26 \;&\le\; 25 + 9 + 1 + 1 = 36 = \textstyle\sum_i y_i^2, \\ \textstyle\sum_i |x_i - 2| = 1 + 1 + 0 + 0 = 2 \;&\le\; 3 + 1 + 1 + 1 = 6, \\ \textstyle\sum_i \max\{x_i - 2, 0\} = 1 + 1 = 2 \;&\le\; 3 + 1 = 4. \end{aligned}$$

Each is an instance of the one theorem. The last one measures how much of each vector sits above the level 2, and it is the test function the converse below is built from.

Quick check. Does Theorem 21.2.1 apply to $\phi(t) = t^3$ on $I = \mathbb{R}$? Test it on $\mathbf{x} = (0, 0, 0) \prec \mathbf{y} = (1, 1, -2)$.

Solution. No. On $\mathbb{R}$ the function t^3 is not convex: at $a = -2$, $b = 0$, and $t = \frac{1}{2}$ the midpoint value is $\phi(-1) = -1$, while the chord gives $\frac{1}{2}\phi(-2) + \frac{1}{2}\phi(0) = -4$, and $-1 > -4$. The conclusion of the theorem does fail here. Both vectors have total 0, and the running totals $0, 0$ of $\mathbf{x}$ are at most $1, 2$, so $\mathbf{x} \prec \mathbf{y}$. But $\sum_i x_i^3 = 0$ while $\sum_i y_i^3 = 1 + 1 - 8 = -6$, and $0 > -6$. On $I = (0, \infty)$, where t^3 **is** convex, since its derivative $3t^2$ is increasing there (Lemma 18.10.5), the theorem does apply, and no such failure is possible for vectors with strictly positive entries.

21.2.2 The converse, and the weak version

Karamata's inequality is not merely a consequence of majorization; it is equivalent to it. Two families of test functions do the work: the affine ones pin down the totals, and the "hockey sticks" $t \mapsto \max\{t-c, 0\}$ pin down the partial sums.

Proposition 21.2.2. Convex Functions Detect Majorization

Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$. If

$$\sum_{i=1}^{n} \phi(x_i) \;\le\; \sum_{i=1}^{n} \phi(y_i)$$

for **every** convex $\phi\colon \mathbb{R} \to \mathbb{R}$, then $\mathbf{x} \prec \mathbf{y}$.

Idea. Use $\phi(t) = t$ and $\phi(t) = -t$ to get the equal totals. For the k-th partial sum, use the function that is 0 below a level c and rises with slope 1 above it. Summed over a vector it always exceeds "top k sum minus kc", and it *equals* that when c is the k-th largest entry. Setting $c = y_k^{\downarrow}$ makes the right-hand side exact and the left-hand side an over-estimate, which is the direction we need.

Proof. Both $\phi(t) = t$ and $\phi(t) = -t$ are affine, hence convex by Example (First convex functions) (a). They give $\sum_i x_i \le \sum_i y_i$ and $-\sum_i x_i \le -\sum_i y_i$, so the totals are equal. This is (M2).

Fix k with $1 \le k \le n-1$, put $c = y_k^{\downarrow}$, and let

$$\phi_c(t) = \max\{t - c,\; 0\},$$

the pointwise maximum of the two affine functions $t \mapsto t - c$ and $t \mapsto 0$, hence convex by Proposition 18.10.3. For any $\mathbf{z} \in \mathbb{R}^n$ the sum $\sum_i \phi_c(z_i)$ does not depend on the order of the entries, so it equals $\sum_i \phi_c(z_i^{\downarrow})$. Every term is ≥ 0, and each of the first k terms satisfies $\phi_c(z_i^{\downarrow}) \ge z_i^{\downarrow} - c$. Hence

$$\sum_{i=1}^{n} \phi_c(z_i) \;\ge\; \sum_{i=1}^{k} z_i^{\downarrow} - kc. \tag{$*$}$$

For $\mathbf{z} = \mathbf{y}$ and this particular c, $(*)$ is an equality: for $i \le k$ we have $y_i^{\downarrow} \ge y_k^{\downarrow} = c$, so $\phi_c(y_i^{\downarrow}) = y_i^{\downarrow} - c$; and for $i > k$ we have $y_i^{\downarrow} \le c$, so $\phi_c(y_i^{\downarrow}) = 0$. Therefore

$$\sum_{i=1}^{k} x_i^{\downarrow} - kc \overset{(*)}{\le} \sum_{i=1}^{n} \phi_c(x_i) \;\le\; \sum_{i=1}^{n} \phi_c(y_i) = \sum_{i=1}^{k} y_i^{\downarrow} - kc,$$

the middle step being the hypothesis applied to ϕ_c. Canceling kc gives (M1) at k. As k was arbitrary, $\mathbf{x} \prec \mathbf{y}$.

■

Together with Theorem 21.2.1 this is an equivalence: $\mathbf{x} \prec \mathbf{y}$ **if and only if** $\sum_i \phi(x_i) \le \sum_i \phi(y_i)$ for every convex $\phi\colon \mathbb{R} \to \mathbb{R}$. So majorization is exactly the order that all convex functions agree on. The proof also shows how few test functions are needed: the two affine ones and the $n-1$

hockey sticks at the levels $y_1^{\downarrow}, \dots, y_{n-1}^{\downarrow}$.

Weak majorization has a matching statement, with the class of test functions cut down.

Proposition 21.2.3. Karamata for Weak Majorization

Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ with $\mathbf{x} \prec_w \mathbf{y}$, and let $\phi\colon \mathbb{R} \to \mathbb{R}$ be convex and **increasing**. Then

$$\sum_{i=1}^{n} \phi(x_i) \;\le\; \sum_{i=1}^{n} \phi(y_i).$$

Idea. The only thing wrong with $\mathbf{y}$ is that its total may be too big. Take the excess away from the **smallest** entry of $\mathbf{y}$. That keeps the vector sorted, leaves all the partial sums before the last one untouched, and produces a genuine majorization, to which Theorem 21.2.1 applies. Lowering an entry can only lower ϕ, which is where "increasing" is used.

Proof. Let $s = \sum_i y_i - \sum_i x_i$. Taking $k = n$ in (M1) for $\mathbf{x} \prec_w \mathbf{y}$ gives $s \ge 0$. Define $\mathbf{w} \in \mathbb{R}^n$ by

$$w_i = y_i^{\downarrow} \quad (i \le n-1), \qquad w_n = y_n^{\downarrow} - s.$$

Only the last entry has changed, and it has been lowered, so $\mathbf{w}$ is still decreasing and $\mathbf{w}^{\downarrow} = \mathbf{w}$. Its total is $\sum_i y_i - s = \sum_i x_i$, which is (M2) for $\mathbf{x} \prec \mathbf{w}$. For $k \le n-1$, $\sum_{i \le k} w_i = \sum_{i \le k} y_i^{\downarrow} \ge \sum_{i \le k} x_i^{\downarrow}$ by (M1) for $\mathbf{x} \prec_w \mathbf{y}$. So $\mathbf{x} \prec \mathbf{w}$.

By Theorem 21.2.1 with $I = \mathbb{R}$, $\sum_i \phi(x_i) \le \sum_i \phi(w_i)$. Since $w_i \le y_i^{\downarrow}$ for every i and ϕ is increasing, $\phi(w_i) \le \phi(y_i^{\downarrow})$, so

$$\sum_i \phi(w_i) \;\le\; \sum_i \phi(y_i^{\downarrow}) = \sum_i \phi(y_i).$$

Chaining the two inequalities proves the proposition.

■

Warning. **"Increasing" cannot be dropped from Proposition 21.2.3.** Take $\mathbf{x} = (0, 0)$ and $\mathbf{y} = (1, 0)$. The running totals $0, 0$ of $\mathbf{x}$ are at most $1, 1$, so $\mathbf{x} \prec_w \mathbf{y}$. The function $\phi(t) = -t$ is convex but decreasing, and $\sum_i \phi(x_i) = 0$ while $\sum_i \phi(y_i) = -1$. A weak majorization allows $\mathbf{y}$ to carry more total than $\mathbf{x}$, and a decreasing ϕ punishes exactly that.

21.2.3 Schur-convex functions

Theorem 21.2.1 is a statement about one particular shape of function, $\mathbf{x} \mapsto \sum_i \phi(x_i)$. The property it establishes deserves a name of its own, because many functions have it that are not of that shape.

A function is Schur-convex when spreading a vector out can only increase it.

The domain needs a moment's thought. If G is to be compared at $\mathbf{x}$ and $\mathbf{y}$ whenever $\mathbf{x} \prec \mathbf{y}$,

then the domain must contain $\mathbf{x}$ as soon as it contains $\mathbf{y}$. Call a set $A \subseteq \mathbb{R}^n$ **symmetric** if $\mathbf{P}_\sigma \mathbf{x} \in A$ for every $\mathbf{x} \in A$ and every $\sigma \in S_n$. A symmetric **convex** set has the property we need: if $\mathbf{y} \in A$ and $\mathbf{x} \prec \mathbf{y}$, then $\mathbf{x} \in \operatorname{conv}\{\mathbf{P}_\sigma \mathbf{y}\}$ by Corollary 21.1.4, every $\mathbf{P}_\sigma \mathbf{y}$ lies in A by symmetry, and a convex set contains the convex combinations of its points (Lemma 18.1.2), so $\mathbf{x} \in A$.

Definition 21.2.1. Schur-Convex Function

Let $A \subseteq \mathbb{R}^n$ be **symmetric and convex**, and let $G\colon A \to \mathbb{R}$. Then G is **Schur-convex** if

$$\mathbf{x} \prec \mathbf{y} \quad \Longrightarrow \quad G(\mathbf{x}) \le G(\mathbf{y})$$

for **all** $\mathbf{x}, \mathbf{y} \in A$. It is **strictly Schur-convex** if the inequality is **strict** whenever $\mathbf{x}^{\downarrow} \neq \mathbf{y}^{\downarrow}$.

Definition 21.2.2. Schur-Concave Function

With A as in Definition 21.2.1, a function $G\colon A \to \mathbb{R}$ is **Schur-concave** if $-G$ is Schur-convex, that is, if

$$\mathbf{x} \prec \mathbf{y} \quad \Longrightarrow \quad G(\mathbf{x}) \ge G(\mathbf{y})$$

for all $\mathbf{x}, \mathbf{y} \in A$; and **strictly Schur-concave** if $-G$ is strictly Schur-convex.

In words: a Schur-convex function is one that respects the majorization order, and a Schur-concave one reverses it. The two standing examples of A below are $\mathbb{R}^n$ itself and the open orthant $\{\mathbf{x} : x_i > 0 \text{ for every } i\}$, both symmetric and convex.

A Schur-convex function is automatically **symmetric**, meaning $G(\mathbf{P}_\sigma \mathbf{x}) = G(\mathbf{x})$ for every permutation. Indeed $\mathbf{x}$ and $\mathbf{P}_\sigma \mathbf{x}$ are rearrangements of each other, so each majorizes the other (Proposition 17.8.1 (a) and the remark after Definition 17.8.1), and two applications of the definition give $G(\mathbf{x}) \le G(\mathbf{P}_\sigma \mathbf{x}) \le G(\mathbf{x})$.

Example. Four functions on $\mathbb{R}^n$

Decide which of the following are Schur-convex, and which are Schur-concave, on $A = \mathbb{R}^n$.

(a) $G(\mathbf{x}) = \sum_i x_i^2$.
(b) $G(\mathbf{x}) = \max_i x_i$.
(c) $G(\mathbf{x}) = \sum_i x_i$.
(d) $G(\mathbf{x}) = \sum_i x_i^3$, for $n = 3$.

Solution.

(a) Schur-convex. $\phi(t) = t^2$ is convex on $\mathbb{R}$, as in Example (One majorization, four inequalities), so Theorem 21.2.1 gives $\sum_i x_i^2 \le \sum_i y_i^2$ whenever $\mathbf{x} \prec \mathbf{y}$.

(b) Schur-convex. $\max_i x_i = x_1^{\downarrow}$, and (M1) at $k = 1$ says exactly $x_1^{\downarrow} \le y_1^{\downarrow}$. No convexity is needed here; the definition of majorization already contains the statement.

(c) Both Schur-convex and Schur-concave, since (M2) makes it constant on each majorization class. This is the degenerate case, and it shows that the two notions are not exclusive.

(d) Neither, and one pair of comparisons settles both. The flat vector is majorized by every vector with the same total (Example (A chain, the two extremes, and size one) (b)),

so

$$(0,0,0) \prec (2,-1,-1), \qquad (0,0,0) \prec (1,1,-2),$$

both totals being 0. Now $G(0,0,0) = 0$, while $G(2,-1,-1) = 8 - 1 - 1 = 6$ and $G(1,1,-2) = 1 + 1 - 8 = -6$. Schur-convexity applied to the second majorization would demand $0 \le -6$, and Schur-concavity applied to the first would demand $0 \ge 6$. Both fail.

Part (d) is the nonexample by minimal change. The function $\sum_i \phi(x_i)$ with $\phi(t) = t^3$ differs from the Schur-convex $\sum_i \phi(x_i)$ with $\phi(t) = t^2$ only in the exponent, and t^3 fails the one clause that matters, convexity on the interval where the entries live.

The next theorem explains why so many natural functions turn out Schur-convex: convexity plus symmetry is enough, and there is no need to be of the form $\sum_i \phi(x_i)$.

Theorem 21.2.4. A Symmetric Convex Function Is Schur-Convex

Let $A \subseteq \mathbb{R}^n$ be symmetric and convex, and let $G\colon A \to \mathbb{R}$ be **convex** and **symmetric**, that is, $G(\mathbf{P}_\sigma \mathbf{x}) = G(\mathbf{x})$ for every $\mathbf{x} \in A$ and every $\sigma \in S_n$. Then G is Schur-convex.

Idea. $\mathbf{x} \prec \mathbf{y}$ puts $\mathbf{x}$ inside the permutation polytope of $\mathbf{y}$, whose corners all carry the **same** value of G, by symmetry. A convex function on a convex combination is at most the corresponding combination of its values, and an average of copies of one number is that number.

Proof. Let $\mathbf{x}, \mathbf{y} \in A$ with $\mathbf{x} \prec \mathbf{y}$. By Corollary 21.1.4 and Theorem 18.1.3,

$$\mathbf{x} = \sum_{i=1}^{m} t_i \, \mathbf{P}_{\sigma_i} \mathbf{y}$$

for some $\sigma_1, \dots, \sigma_m \in S_n$ and weights $t_i \ge 0$ with $\sum_i t_i = 1$. Each $\mathbf{P}_{\sigma_i}\mathbf{y}$ lies in A, since A is symmetric. By Theorem 18.10.2 applied to the convex G on A,

$$G(\mathbf{x}) \;\le\; \sum_{i=1}^{m} t_i \, G(\mathbf{P}_{\sigma_i}\mathbf{y}) = \sum_{i=1}^{m} t_i \, G(\mathbf{y}) = G(\mathbf{y}),$$

the middle equality being the symmetry of G and the last one $\sum_i t_i = 1$. This shows that G is Schur-convex.

■

Now the promise from Chapter 17 §08. There the second proof of Hadamard's inequality used only the trace, and a remark recorded that a proof "by majorization in earnest" would need the fact that "for vectors with nonnegative entries, $\mathbf{x} \prec \mathbf{y}$ implies $x_1 \cdots x_n \ge y_1 \cdots y_n$ (the product is **Schur-concave**)". Here it is.

Corollary 21.2.5. The Product Is Schur-Concave

Let $\mathbf{y} \in \mathbb{R}^n$ have **nonnegative** entries and let $\mathbf{x} \prec \mathbf{y}$. Then every entry of $\mathbf{x}$ is nonnegative,

and

$$x_1 x_2 \cdots x_n \;\geq\; y_1 y_2 \cdots y_n.$$

Idea. Take $-\log$, which is convex, and apply Theorem 21.2.1: the product becomes a sum and the inequality flips. The logarithm needs positive entries, so the case of a zero entry is dealt with first — and it is the easy case, because a zero entry makes the right-hand side 0 while the left-hand side is a product of nonnegative numbers.

Proof. First, every entry of $\mathbf{x}$ is nonnegative. For $n = 1$, (M2) gives $\mathbf{x} = \mathbf{y}$. For $n \geq 2$, subtracting (M1) at $k = n - 1$ from (M2) gives

$$x_n^{\downarrow} \;\geq\; y_n^{\downarrow} \;\geq\; 0,$$

and $x_n^{\downarrow}$ is the smallest entry of $\mathbf{x}$.

Case 1: some $y_i = 0$. Then $y_1 \cdots y_n = 0$, while $x_1 \cdots x_n \geq 0$ as a product of nonnegative numbers. The inequality holds.

Case 2: every $y_i > 0$. Then $y_n^{\downarrow} > 0$, so $x_n^{\downarrow} > 0$ by the display above, and every entry of both vectors lies in $I = (0, \infty)$. By Lemma 18.11.1 (c) the function log is concave on I, so $\phi = -\log$ is convex there. By Theorem 21.2.1,

$$\sum_{i=1}^{n}(-\log x_i) \;\leq\; \sum_{i=1}^{n}(-\log y_i), \qquad \text{that is,} \qquad \sum_{i=1}^{n}\log y_i \;\leq\; \sum_{i=1}^{n}\log x_i.$$

By $\log(ab) = \log a + \log b$ (Lemma 18.11.1 (c)) and induction on n, $\sum_i \log x_i = \log(x_1 \cdots x_n)$ and likewise for $\mathbf{y}$. So $\log(y_1 \cdots y_n) \leq \log(x_1 \cdots x_n)$. Since log is strictly increasing on I (Lemma 18.11.1 (c)), two positive numbers with $\log u \leq \log v$ satisfy $u \leq v$; otherwise $u > v$ would give $\log u > \log v$. Hence $y_1 \cdots y_n \leq x_1 \cdots x_n$. This proves the corollary. ∎

So the flat vector, the bottom of the majorization order, has the largest product among nonnegative vectors with a fixed total — which is Lemma 17.8.6 again, now as a special case rather than a separate induction.

Warning. **Schur-convex does not mean convex.** By Corollary 21.2.5 the function $G(\mathbf{x}) = -x_1 x_2$ is Schur-convex on the symmetric convex set $A = \{\mathbf{x} \in \mathbb{R}^2 : x_1, x_2 \geq 0\}$. It is not convex: at $\mathbf{x} = (1, 1)$, $\mathbf{y} = (2, 2)$, and $t = \frac{1}{2}$,

$$G(\tfrac{3}{2}, \tfrac{3}{2}) = -\tfrac{9}{4} = -2.25, \qquad \tfrac{1}{2}G(1,1) + \tfrac{1}{2}G(2,2) = \tfrac{1}{2}(-1) + \tfrac{1}{2}(-4) = -2.5,$$

and $-2.25 > -2.5$. Theorem 21.2.4 is a one-way street.

Quick check. Is the function $G(\mathbf{x}) = x_1^{\downarrow} + x_2^{\downarrow}$ on $\mathbb{R}^3$ Schur-convex? Is it convex?

Solution. Schur-convex: the inequality $x_1^{\downarrow} + x_2^{\downarrow} \le y_1^{\downarrow} + y_2^{\downarrow}$ is (M1) at $k = 2$. Convex: yes, and this gives a second proof. The two largest entries have the largest sum among all pairs, so $G(\mathbf{x}) = \max_{i<j}(x_i + x_j)$, a maximum of three linear functions, convex by Proposition 18.10.3. It is also symmetric, so Theorem 21.2.4 applies. (This G is the Ky Fan sum s_2 of Chapter 17 §06, read on vectors instead of eigenvalues.)

21.2.4 Convex functions of a Hermitian matrix

The last payoff is about matrices. For a Hermitian $\mathbf{A}$ and a convex f on an interval containing $\operatorname{spec}(\mathbf{A})$, the number $\operatorname{tr} f(\mathbf{A})$ turns out to be $\sum_i f(\lambda_i(\mathbf{A}))$. Chapter 18 §10 stated, and deferred, that this is a convex function of $\mathbf{A}$. The proof needs two lemmas: a bookkeeping one about traces, and one saying that the quadratic form of $f(\mathbf{A})$ dominates f of the quadratic form of $\mathbf{A}$.

Lemma 21.2.6. Reading a Trace on an Orthonormal Basis

Let $F = \mathbb{R}$ or $\mathbb{C}$.

(a) Let $(\mathbf{u}_1, \dots, \mathbf{u}_n)$ be an orthonormal basis of F^n and $\mathbf{M} \in M_n(F)$. Then $\operatorname{tr} \mathbf{M} = \sum_{i=1}^n \langle \mathbf{M}\mathbf{u}_i, \mathbf{u}_i \rangle$.

(b) Let $I \subseteq \mathbb{R}$, let $\mathbf{A} \in M_n(F)$ be Hermitian with $\operatorname{spec}(\mathbf{A}) \subseteq I$, and let $g\colon I \to \mathbb{R}$. Then $g(\mathbf{A})\mathbf{u} = g(\lambda)\mathbf{u}$ for every eigenvector $\mathbf{u}$ of $\mathbf{A}$ with eigenvalue λ, and $\operatorname{tr} g(\mathbf{A}) = \sum_{i=1}^n g(\lambda_i(\mathbf{A}))$.

Proof.

(a) Let $\mathbf{U} \in M_n(F)$ be the matrix whose i-th column is $\mathbf{u}_i$. Then $(\mathbf{U}^*\mathbf{U})_{ij} = \mathbf{u}_i^* \mathbf{u}_j = \langle \mathbf{u}_j, \mathbf{u}_i \rangle = \delta_{ij}$, so $\mathbf{U}$ is unitary, and $(\mathbf{U}^*\mathbf{M}\mathbf{U})_{ii} = \mathbf{u}_i^* \mathbf{M} \mathbf{u}_i = \langle \mathbf{M}\mathbf{u}_i, \mathbf{u}_i \rangle$. By Theorem 0.9.6 (3), used with the factors $\mathbf{U}^*$ and $\mathbf{M}\mathbf{U}$, $\operatorname{tr}(\mathbf{U}^*\mathbf{M}\mathbf{U}) = \operatorname{tr}(\mathbf{M}\mathbf{U}\mathbf{U}^*) = \operatorname{tr} \mathbf{M}$. Summing the diagonal of $\mathbf{U}^*\mathbf{M}\mathbf{U}$ gives (a).

(b) Let $\mathbf{A} = \sum_j \mu_j \mathbf{P}_j$ be the spectral resolution of Definition 12.8.1, with $\mu_1, \dots, \mu_k$ the distinct eigenvalues; it exists by Theorem 12.8.1, since a Hermitian matrix has an orthonormal basis of eigenvectors (Corollary 12.4.3 over $\mathbb{C}$, Corollary 12.5.2 over $\mathbb{R}$). Let $\mathbf{A}\mathbf{u} = \lambda \mathbf{u}$ with $\mathbf{u} \neq \mathbf{0}$. Then $\lambda = \mu_{j_0}$ for exactly one j_0, and $\mathbf{u}$ lies in the eigenspace of μ_{j_0}, which is the image of $\mathbf{P}_{j_0}$; so $\mathbf{P}_{j_0}\mathbf{u} = \mathbf{u}$, and $\mathbf{P}_j \mathbf{u} = \mathbf{P}_j \mathbf{P}_{j_0} \mathbf{u} = \mathbf{0}$ for $j \neq j_0$ by Theorem 12.8.1 (b). Hence $g(\mathbf{A})\mathbf{u} = \sum_j g(\mu_j)\mathbf{P}_j \mathbf{u} = g(\lambda)\mathbf{u}$, by Definition 12.8.2.

Now take an orthonormal basis $(\mathbf{u}_1, \dots, \mathbf{u}_n)$ of F^n with $\mathbf{A}\mathbf{u}_i = \lambda_i(\mathbf{A})\mathbf{u}_i$, which exists by the spectral theorem just cited. By the previous paragraph, $\langle g(\mathbf{A})\mathbf{u}_i, \mathbf{u}_i \rangle = g(\lambda_i(\mathbf{A}))\langle \mathbf{u}_i, \mathbf{u}_i \rangle = g(\lambda_i(\mathbf{A}))$, and (a) adds these up to $\operatorname{tr} g(\mathbf{A})$.

■

Lemma 21.2.7. Peierls' Inequality

Let $I \subseteq \mathbb{R}$ be an interval, let $\mathbf{A} \in M_n(F)$ be Hermitian with $\operatorname{spec}(\mathbf{A}) \subseteq I$, let $f\colon I \to \mathbb{R}$ be convex, and let $\mathbf{u} \in F^n$ with $\|\mathbf{u}\| = 1$. Then $\langle \mathbf{A}\mathbf{u}, \mathbf{u} \rangle \in I$ and

$$f\big(\langle \mathbf{A}\mathbf{u}, \mathbf{u} \rangle\big) \;\le\; \langle f(\mathbf{A})\mathbf{u}, \mathbf{u} \rangle,$$

with **equality** when $\mathbf{u}$ is an eigenvector of $\mathbf{A}$.

Idea. Write $\mathbf{A} = \sum_j \mu_j \mathbf{P}_j$, its spectral resolution. Then $\langle \mathbf{A}\mathbf{u}, \mathbf{u}\rangle = \sum_j \mu_j \|\mathbf{P}_j\mathbf{u}\|^2$ is a weighted average of the eigenvalues, with weights $\|\mathbf{P}_j\mathbf{u}\|^2$ that add up to $\|\mathbf{u}\|^2 = 1$. The same weights appear in $\langle f(\mathbf{A})\mathbf{u}, \mathbf{u}\rangle = \sum_j f(\mu_j)\|\mathbf{P}_j\mathbf{u}\|^2$. So the two sides of the inequality are f of an average against the average of f, which is Jensen.

Proof. Let $\mathbf{A} = \mu_1\mathbf{P}_1 + \cdots + \mu_k\mathbf{P}_k$ be the spectral resolution of Definition 12.8.1, with $\mu_1, \dots, \mu_k$ the **distinct** eigenvalues of $\mathbf{A}$, which exists by Theorem 12.8.1 since a Hermitian matrix has an orthonormal basis of eigenvectors (Corollary 12.4.3 over $\mathbb{C}$, Corollary 12.5.2 over $\mathbb{R}$). Put

$$w_j = \langle \mathbf{P}_j\mathbf{u}, \mathbf{u}\rangle.$$

By Theorem 12.8.1 (b), $\mathbf{P}_j = \mathbf{P}_j^2$ and $\mathbf{P}_j^* = \mathbf{P}_j$, so

$$w_j = \langle \mathbf{P}_j\mathbf{P}_j\mathbf{u}, \mathbf{u}\rangle = \langle \mathbf{P}_j\mathbf{u}, \mathbf{P}_j^*\mathbf{u}\rangle = \langle \mathbf{P}_j\mathbf{u}, \mathbf{P}_j\mathbf{u}\rangle = \|\mathbf{P}_j\mathbf{u}\|^2 \;\geq\; 0,$$

and by Theorem 12.8.1 (c), $\sum_j w_j = \langle \mathbf{u}, \mathbf{u}\rangle = 1$. Hence, by Theorem 12.8.1 (d) and Definition 12.8.2,

$$\langle \mathbf{A}\mathbf{u}, \mathbf{u}\rangle = \sum_{j=1}^{k} \mu_j w_j, \qquad \langle f(\mathbf{A})\mathbf{u}, \mathbf{u}\rangle = \sum_{j=1}^{k} f(\mu_j) w_j.$$

The first is a convex combination of $\mu_1, \dots, \mu_k$, all of which lie in the convex set I, so $\langle \mathbf{A}\mathbf{u}, \mathbf{u}\rangle \in I$ by Lemma 18.1.2. Applying Theorem 18.10.2 to the convex f on I, with points μ_j and weights w_j,

$$f(\langle \mathbf{A}\mathbf{u}, \mathbf{u}\rangle) = f\Big(\sum_j w_j\mu_j\Big) \;\leq\; \sum_j w_j f(\mu_j) = \langle f(\mathbf{A})\mathbf{u}, \mathbf{u}\rangle.$$

For the equality clause, let $\mathbf{A}\mathbf{u} = \lambda\mathbf{u}$ with $\|\mathbf{u}\| = 1$. By Lemma 21.2.6 (b), $f(\mathbf{A})\mathbf{u} = f(\lambda)\mathbf{u}$, so the right side is $f(\lambda)\langle \mathbf{u}, \mathbf{u}\rangle = f(\lambda)$; and the left side is $f(\langle \lambda\mathbf{u}, \mathbf{u}\rangle) = f(\lambda)$ as well. This proves the lemma.

■

Quick check. Take $\mathbf{A} = \left(\begin{smallmatrix} 2 & 1 \\ 1 & 2 \end{smallmatrix}\right)$, $f(t) = t^2$, and $\mathbf{u} = \mathbf{e}_1$. Compute both sides of Lemma 21.2.7, and then repeat with $\mathbf{u} = \frac{1}{\sqrt{2}}(1, 1)$.

Solution. $\langle \mathbf{A}\mathbf{e}_1, \mathbf{e}_1\rangle = a_{11} = 2$, so the left side is $f(2) = 4$. Here $f(\mathbf{A}) = \mathbf{A}^2 = \left(\begin{smallmatrix} 5 & 4 \\ 4 & 5 \end{smallmatrix}\right)$, so the right side is $(\mathbf{A}^2)_{11} = 5$. Indeed $4 \leq 5$, and the inequality is strict because $\mathbf{e}_1$ is not an eigenvector of $\mathbf{A}$.

For $\mathbf{u} = \frac{1}{\sqrt{2}}(1, 1)$, an eigenvector with $\mathbf{A}\mathbf{u} = 3\mathbf{u}$, the left side is $f(3) = 9$ and the right side is $\langle \mathbf{A}^2\mathbf{u}, \mathbf{u}\rangle = 9\langle \mathbf{u}, \mathbf{u}\rangle = 9$. Equality, as the lemma says.

Theorem 21.2.8. The Trace of a Convex Function Is Convex

Let $I \subseteq \mathbb{R}$ be an interval and let $f\colon I \to \mathbb{R}$ be **convex**. Let $\mathbf{A}, \mathbf{B} \in M_n(F)$ be Hermitian with $\operatorname{spec}(\mathbf{A}) \subseteq I$ and $\operatorname{spec}(\mathbf{B}) \subseteq I$, and let $0 \le \theta \le 1$. Then $\mathbf{C} = \theta\mathbf{A} + (1-\theta)\mathbf{B}$ is Hermitian with $\operatorname{spec}(\mathbf{C}) \subseteq I$, and

$$\operatorname{tr} f(\mathbf{C}) \;\le\; \theta \operatorname{tr} f(\mathbf{A}) + (1-\theta) \operatorname{tr} f(\mathbf{B}).$$

In words: the Hermitian matrices with spectrum in I form a **convex** set, and $\mathbf{A} \mapsto \operatorname{tr} f(\mathbf{A})$ is a **convex** function on it.

Idea. Fix an orthonormal eigenbasis of the midpoint matrix $\mathbf{C} = \theta\mathbf{A} + (1-\theta)\mathbf{B}$. Along each of those n directions the value of f at the eigenvalue of $\mathbf{C}$ is exactly the left side of Peierls' inequality, because an eigenvector gives equality; and the eigenvalue itself splits as θ times a quadratic form of $\mathbf{A}$ plus $(1-\theta)$ times one of $\mathbf{B}$. Ordinary convexity of f handles that split, and Peierls' inequality, now as an inequality, converts the two quadratic forms into quadratic forms of $f(\mathbf{A})$ and $f(\mathbf{B})$. Summing over the basis turns quadratic forms into traces.

Proof. **Step 1: the spectrum of $\mathbf{C}$.** The matrix $\mathbf{C} = \theta\mathbf{A} + (1-\theta)\mathbf{B}$ is Hermitian, being a real combination of Hermitian matrices. Let μ be an eigenvalue of $\mathbf{C}$, with unit eigenvector $\mathbf{x}$. Then

$$\mu = \langle \mathbf{C}\mathbf{x}, \mathbf{x}\rangle = \theta\langle \mathbf{A}\mathbf{x}, \mathbf{x}\rangle + (1-\theta)\langle \mathbf{B}\mathbf{x}, \mathbf{x}\rangle.$$

By Lemma 13.5.2, $\langle \mathbf{A}\mathbf{x}, \mathbf{x}\rangle$ lies in $[\lambda_n(\mathbf{A}), \lambda_1(\mathbf{A})]$, an interval whose endpoints are eigenvalues of $\mathbf{A}$, hence contained in I; likewise $\langle \mathbf{B}\mathbf{x}, \mathbf{x}\rangle \in I$. So μ is a convex combination of two points of I and lies in I. Hence $\operatorname{spec}(\mathbf{C}) \subseteq I$, and $f(\mathbf{C})$ is defined.

Step 2: the inequality. Choose an orthonormal basis $(\mathbf{u}_1, \dots, \mathbf{u}_n)$ of F^n consisting of eigenvectors of $\mathbf{C}$, which exists by Corollary 12.4.3 over $\mathbb{C}$ and Corollary 12.5.2 over $\mathbb{R}$. Fix i. By Lemma 21.2.7 applied to $\mathbf{C}$ and the eigenvector $\mathbf{u}_i$, with equality,

$$\langle f(\mathbf{C})\mathbf{u}_i, \mathbf{u}_i\rangle = f\big(\langle \mathbf{C}\mathbf{u}_i, \mathbf{u}_i\rangle\big).$$

By Step 1, $\langle \mathbf{C}\mathbf{u}_i, \mathbf{u}_i\rangle = \theta\alpha_i + (1-\theta)\beta_i$ with $\alpha_i = \langle \mathbf{A}\mathbf{u}_i, \mathbf{u}_i\rangle \in I$ and $\beta_i = \langle \mathbf{B}\mathbf{u}_i, \mathbf{u}_i\rangle \in I$. So Definition 18.10.1 and then Lemma 21.2.7, now applied to $\mathbf{A}$ and to $\mathbf{B}$ with the unit vector $\mathbf{u}_i$, give

$$\begin{aligned}\langle f(\mathbf{C})\mathbf{u}_i, \mathbf{u}_i\rangle = f\big(\theta\alpha_i + (1-\theta)\beta_i\big) &\le \theta f(\alpha_i) + (1-\theta) f(\beta_i) \\ &\le \theta\langle f(\mathbf{A})\mathbf{u}_i, \mathbf{u}_i\rangle + (1-\theta)\langle f(\mathbf{B})\mathbf{u}_i, \mathbf{u}_i\rangle,\end{aligned}$$

where the last step uses $\theta \ge 0$ and $1-\theta \ge 0$. Summing over $i = 1, \dots, n$ and using Lemma 21.2.6 (a) three times, once for each of $f(\mathbf{C})$, $f(\mathbf{A})$, and $f(\mathbf{B})$,

$$\operatorname{tr} f(\mathbf{C}) \;\le\; \theta \operatorname{tr} f(\mathbf{A}) + (1-\theta) \operatorname{tr} f(\mathbf{B}).$$

Since $\mathbf{A}$, $\mathbf{B}$, and θ were arbitrary, this is Definition 18.10.1 for $\mathbf{A} \mapsto \operatorname{tr} f(\mathbf{A})$ on the set of Hermitian matrices with spectrum in I, which Step 1 shows is convex. This proves the theorem.

■

This pays Chapter 18 §10's "it is proved in Chapter 21, and nothing in this chapter uses it", and Chapter 18 §11's closing note that "the convexity of $\mathbf{A} \mapsto \operatorname{tr} f(\mathbf{A})$ for convex f, and the consequences of a majorization for every convex function, belong to Chapter 21".

Example. The logarithm of the determinant

Let $f(t) = -\log t$ on $I = (0, \infty)$, so that the matrices in play are the positive definite Hermitian ones. Identify $\operatorname{tr} f(\mathbf{A})$, and check the resulting inequality for

$$\mathbf{A} = \begin{pmatrix} 1 & 0 \\ 0 & 4 \end{pmatrix}, \qquad \mathbf{B} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}, \qquad \theta = \tfrac{1}{2}.$$

Solution. A Hermitian $\mathbf{A}$ has $\operatorname{spec}(\mathbf{A}) \subseteq (0, \infty)$ exactly when $\mathbf{A} \succ \mathbf{O}$, by Theorem 13.1.5 (b). For such an $\mathbf{A}$, Lemma 21.2.6 (b), then $\log(ab) = \log a + \log b$ (Lemma 18.11.1 (c)) with induction on n, and then Theorem 9.3.3 give

$$\operatorname{tr} f(\mathbf{A}) = -\sum_i \log \lambda_i(\mathbf{A}) = -\log \prod_i \lambda_i(\mathbf{A}) = -\log \det \mathbf{A}.$$

By Theorem 21.2.8, $\mathbf{A} \mapsto -\log\det \mathbf{A}$ **is convex on the positive definite matrices**; equivalently, det is log-concave there.

For the data given, $\det \mathbf{A} = 4$ and $\det \mathbf{B} = 4 - 1 = 3$, and

$$\mathbf{C} = \tfrac{1}{2}(\mathbf{A} + \mathbf{B}) = \begin{pmatrix} \frac{3}{2} & \frac{1}{2} \\ \frac{1}{2} & 3 \end{pmatrix}, \qquad \det \mathbf{C} = \tfrac{9}{2} - \tfrac{1}{4} = \tfrac{17}{4}.$$

Both $\mathbf{A}$ and $\mathbf{B}$ are positive definite ($\mathbf{A}$ is diagonal with positive entries; $\mathbf{B}$ has eigenvalues 3 and 1), so the theorem says $-\log \frac{17}{4} \le -\frac{1}{2}\log 4 - \frac{1}{2}\log 3$, that is, after multiplying by -1 and exponentiating,

$$\det \mathbf{C} = \tfrac{17}{4} = 4.25 \;\ge\; \sqrt{4 \cdot 3} = \sqrt{12} \approx 3.464,$$

which holds, since $4.25^2 = 18.0625 \ge 12$.

Remark. Theorem 21.2.8 holds for **every** convex f, which is a very large class. It is nonetheless much weaker than the operator convexity of Definition 17.11.1 (b), which asks for $f(\theta\mathbf{A} + (1-\theta)\mathbf{B}) \preceq \theta f(\mathbf{A}) + (1-\theta) f(\mathbf{B})$ in the Loewner order, a statement about matrices rather than about one number extracted from them. Taking traces turns the Loewner inequality into the one proved here, since the trace of a positive semidefinite matrix is the sum of its nonnegative eigenvalues (Theorem 13.1.4 (b), Theorem 9.3.3); so operator convexity implies trace convexity, but the converse fails badly, and a later section of this chapter exhibits convex functions that are not operator convex.

21.2.5 Exercises

21.2.5.1 A. Check your understanding

Exercise. A1

(a) State Karamata's inequality, with all its hypotheses.

(b) Which test functions prove Proposition 21.2.2, and which clause of Definition 17.8.1 does each one give?

(c) Define Schur-convex and Schur-concave.

(d) True or false: every Schur-convex function on $\mathbb{R}^n$ is convex. Justify your answer.

(e) True or false: every convex function on $\mathbb{R}^n$ is Schur-convex. Justify your answer.

(f) State what Theorem 21.2.8 says for $f(t) = t^2$, and identify the resulting function of $\mathbf{A}$.

Solution.

(a) If I is an interval, $\phi\colon I \to \mathbb{R}$ is convex, $\mathbf{x} \prec \mathbf{y}$ in $\mathbb{R}^n$, and every entry of $\mathbf{y}$ lies in I, then every entry of $\mathbf{x}$ lies in I and $\sum_i \phi(x_i) \le \sum_i \phi(y_i)$ (Theorem 21.2.1).

(b) The affine functions $\phi(t) = t$ and $\phi(t) = -t$ give the equal totals, (M2). The hockey stick $\phi_c(t) = \max\{t - c, 0\}$ with $c = y_k^{\downarrow}$ gives (M1) at k.

(c) For a symmetric convex $A \subseteq \mathbb{R}^n$, $G\colon A \to \mathbb{R}$ is Schur-convex if $\mathbf{x} \prec \mathbf{y}$ implies $G(\mathbf{x}) \le G(\mathbf{y})$ (Definition 21.2.1), and Schur-concave if $\mathbf{x} \prec \mathbf{y}$ implies $G(\mathbf{x}) \ge G(\mathbf{y})$ (Definition 21.2.2).

(d) False. $G(\mathbf{x}) = -x_1 x_2$ on the nonnegative quadrant is Schur-convex by Corollary 21.2.5, but the warning after that corollary shows that $G(\frac{3}{2}, \frac{3}{2}) = -2.25$ exceeds the average -2.5 of $G(1,1)$ and $G(2,2)$.

(e) False. A Schur-convex function must be symmetric, as shown after Definition 21.2.2, and $G(\mathbf{x}) = x_1$ is linear, hence convex, but not symmetric: $G(1,0) = 1 \neq 0 = G(0,1)$, while $(1,0)$ and $(0,1)$ majorize each other. Symmetry is exactly the missing hypothesis of Theorem 21.2.4.

(f) For $I = \mathbb{R}$ and $f(t) = t^2$, it says that $\mathbf{A} \mapsto \operatorname{tr}(\mathbf{A}^2)$ is convex on the Hermitian matrices. Since $\mathbf{A}$ is Hermitian, $\operatorname{tr}(\mathbf{A}^2) = \operatorname{tr}(\mathbf{A}^*\mathbf{A}) = \|\mathbf{A}\|_F^2$, the squared Frobenius norm.

21.2.5.2 B. Practice

Exercise. B1: Four consequences of one majorization

Let $\mathbf{y} = (6, 2, 1, 1)$ and $\mathbf{x} = (4, 3, 2, 1)$.

(a) Verify that $\mathbf{x} \prec \mathbf{y}$.

(b) Compute $\sum_i x_i^2$ and $\sum_i y_i^2$, and $\sum_i \max\{x_i - 3, 0\}$ and $\sum_i \max\{y_i - 3, 0\}$. Name the theorem each comparison illustrates.

(c) Compute $\prod_i x_i$ and $\prod_i y_i$. Hence check Corollary 21.2.5.

(d) Give a convex $\phi\colon \mathbb{R} \to \mathbb{R}$ for which the two sides are **equal**.

Solution.

(a) Both vectors are decreasing with total 10; the running totals are $4, 7, 9$ for $\mathbf{x}$ against $6, 8, 9$ for $\mathbf{y}$. So (M1) and (M2) hold.

(b) $\sum_i x_i^2 = 16 + 9 + 4 + 1 = 30$ and $\sum_i y_i^2 = 36 + 4 + 1 + 1 = 42$, so $30 \le 42$. Also

$\sum_i \max\{x_i - 3, 0\} = 1 + 0 + 0 + 0 = 1$ and $\sum_i \max\{y_i - 3, 0\} = 3$, so $1 \le 3$. Both illustrate Theorem 21.2.1, the first with $\phi(t) = t^2$ and the second with the hockey stick at level 3.

(c) $\prod_i x_i = 4 \cdot 3 \cdot 2 \cdot 1 = 24$ and $\prod_i y_i = 6 \cdot 2 \cdot 1 \cdot 1 = 12$. Indeed $24 \ge 12$, as Corollary 21.2.5 requires for nonnegative vectors.

(d) Any affine ϕ, for instance $\phi(t) = t$: both sides are the common total 10, by (M2). Affine functions are the ones on which majorization gives no information.

Exercise. B2: Which are Schur-convex?

Determine which of the following functions are Schur-convex on the stated symmetric convex set. Justify your answer, naming the theorem used or giving a counterexample.

(a) $G(\mathbf{x}) = \max_i x_i - \min_i x_i$ on $\mathbb{R}^n$.

(b) $G(\mathbf{x}) = \sum_i |x_i|$ on $\mathbb{R}^n$.

(c) $G(\mathbf{x}) = x_1 + 2x_2$ on $\mathbb{R}^2$.

(d) $G(\mathbf{x}) = \sum_i \sqrt{x_i}$ on $A = \{\mathbf{x} \in \mathbb{R}^n : x_i > 0 \text{ for every } i\}$.

Solution.

(a) Schur-convex. $G(\mathbf{x}) = x_1^{\downarrow} - x_n^{\downarrow}$, and $\mathbf{x} \prec \mathbf{y}$ gives $x_1^{\downarrow} \le y_1^{\downarrow}$ by (M1) at $k = 1$ and $x_n^{\downarrow} \ge y_n^{\downarrow}$ by subtracting (M1) at $k = n - 1$ from (M2). Subtracting the second from the first gives $G(\mathbf{x}) \le G(\mathbf{y})$.

(b) Schur-convex, by Theorem 21.2.1 with $\phi(t) = |t|$, convex as a maximum of two affine functions (Proposition 18.10.3).

(c) Not Schur-convex. It is not symmetric, and a Schur-convex function must be. Concretely $(1, 0)$ and $(0, 1)$ are rearrangements of each other, so $(1, 0) \prec (0, 1)$ and $(0, 1) \prec (1, 0)$. Schur-convexity would give $G(1, 0) \le G(0, 1)$ from the first and $G(0, 1) \le G(1, 0)$ from the second, hence $G(1, 0) = G(0, 1)$. But $G(1, 0) = 1$ and $G(0, 1) = 2$.

(d) Not Schur-convex; it is Schur-**concave**. First, $t \mapsto \sqrt{t}$ is concave on $(0, \infty)$: for $a, b > 0$ and $0 \le \theta \le 1$, Lemma 17.8.6 gives $2\sqrt{ab} \le a + b$, so

$$\begin{aligned}(\theta\sqrt{a} + (1-\theta)\sqrt{b})^2 &= \theta^2 a + (1-\theta)^2 b + 2\theta(1-\theta)\sqrt{ab} \\ &\le \theta^2 a + (1-\theta)^2 b + \theta(1-\theta)(a+b) \\ &= \theta a + (1-\theta) b,\end{aligned}$$

using $\theta^2 + \theta(1-\theta) = \theta$ and $(1-\theta)^2 + \theta(1-\theta) = 1 - \theta$; taking square roots of two nonnegative numbers preserves the inequality. So $\phi(t) = -\sqrt{t}$ is convex on $(0, \infty)$, and Theorem 21.2.1 turns $\mathbf{x} \prec \mathbf{y}$ into $\sum_i(-\sqrt{x_i}) \le \sum_i(-\sqrt{y_i})$, that is, $G(\mathbf{x}) \ge G(\mathbf{y})$. For a concrete instance, $(2, 2) \prec (3, 1)$ and $G(2, 2) = 2\sqrt{2} \approx 2.83$, while $G(3, 1) = \sqrt{3} + 1 \approx 2.73$, so G went **down** as the vector spread out.

Exercise. B3: Trace convexity on two-by-two matrices

(a) Let $f(t) = t^2$, $\mathbf{A} = \operatorname{diag}(2, 0)$, $\mathbf{B} = \left(\begin{smallmatrix} 1 & 1 \\ 1 & 1 \end{smallmatrix}\right)$, and $\theta = \frac{1}{2}$. Compute $\operatorname{tr} f(\mathbf{A})$, $\operatorname{tr} f(\mathbf{B})$, and $\operatorname{tr} f(\mathbf{C})$ for $\mathbf{C} = \frac{1}{2}(\mathbf{A} + \mathbf{B})$, and verify Theorem 21.2.8.

(b) Let $f(t) = -\log t$, $\mathbf{A} = \mathbf{I}_2$, $\mathbf{B} = \operatorname{diag}(4, \frac{1}{4})$, and $\theta = \frac{1}{2}$. Do the same, and state the inequality in terms of determinants.

Solution.

(a) Both matrices are real symmetric, hence Hermitian. $\mathbf{A}^2 = \operatorname{diag}(4, 0)$, so $\operatorname{tr} f(\mathbf{A}) = 4$; and $\mathbf{B}^2 = \left(\begin{smallmatrix} \frac{2}{2} & \frac{2}{2} \end{smallmatrix}\right)$, so $\operatorname{tr} f(\mathbf{B}) = 4$. Next

$$\mathbf{C} = \tfrac{1}{2}\begin{pmatrix} 3 & 1 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} \frac{3}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix}, \qquad \mathbf{C}^2 = \begin{pmatrix} \frac{5}{2} & 1 \\ 1 & \frac{1}{2} \end{pmatrix},$$

so $\operatorname{tr} f(\mathbf{C}) = \frac{5}{2} + \frac{1}{2} = 3$. The theorem asserts $3 \le \frac{1}{2} \cdot 4 + \frac{1}{2} \cdot 4 = 4$, which holds.

(b) Both are positive definite, so their spectra lie in $(0, \infty)$. By Example (The logarithm of the determinant), $\operatorname{tr} f(\mathbf{M}) = -\log\det \mathbf{M}$. Here $\det \mathbf{A} = 1$ and $\det \mathbf{B} = 4 \cdot \frac{1}{4} = 1$, so both traces are $-\log 1 = 0$. The midpoint is $\mathbf{C} = \operatorname{diag}(\frac{5}{2}, \frac{5}{8})$, with $\det \mathbf{C} = \frac{25}{16}$, so $\operatorname{tr} f(\mathbf{C}) = -\log \frac{25}{16}$. The theorem asserts

$$-\log \tfrac{25}{16} \;\le\; \tfrac{1}{2} \cdot 0 + \tfrac{1}{2} \cdot 0 = 0,$$

which holds because $\frac{25}{16} > 1$ and $\log$ is increasing with $\log 1 = 0$ (Lemma 18.11.1 (c)). In determinant form: $\det \mathbf{C} = \frac{25}{16} \ge 1 = \sqrt{\det \mathbf{A}\, \det \mathbf{B}}$.

21.2.5.3 C. Going deeper

Exercise. C1: Entropy

For $\mathbf{x}$ in $A = \{\mathbf{x} \in \mathbb{R}^n : x_i > 0 \text{ for every } i\}$, let $H(\mathbf{x}) = -\sum_i x_i \log x_i$.

(a) Prove that $g(t) = t \log t$ is convex on $(0, \infty)$. *Hint: apply Theorem 18.10.2 to the concave* log*, with the points $1/a$, $1/b$ and cleverly chosen weights.*

(b) Deduce that H is Schur-concave on A.

(c) Deduce that $H(\mathbf{x}) \le \log n$ for every $\mathbf{x} \in A$ with $\sum_i x_i = 1$, with equality when $\mathbf{x} = (\frac{1}{n}, \dots, \frac{1}{n})$.

Solution.

(a) Let $a, b > 0$ and $0 \le \theta \le 1$, and put $c = \theta a + (1-\theta)b > 0$. Set

$$w_1 = \frac{\theta a}{c}, \qquad w_2 = \frac{(1-\theta)b}{c},$$

which are nonnegative and satisfy $w_1 + w_2 = (\theta a + (1-\theta)b)/c = 1$. The function $-\log$ is convex on $(0, \infty)$ by Lemma 18.11.1 (c), so Theorem 18.10.2 with the points $1/a$ and $1/b$ gives

$$-\log\Big(w_1 \frac{1}{a} + w_2 \frac{1}{b}\Big) \;\le\; -w_1 \log \frac{1}{a} - w_2 \log \frac{1}{b}.$$

Now $w_1/a + w_2/b = \theta/c + (1-\theta)/c = 1/c$, and $\log(1/a) = -\log a$ since $\log a + \log(1/a) =$

$\log 1 = 0$ (Lemma 18.11.1 (c)). So the display reads

$$\log c \;\le\; w_1 \log a + w_2 \log b = \frac{\theta a \log a + (1-\theta) b \log b}{c}.$$

Multiplying by $c > 0$ gives $c \log c \le \theta a \log a + (1-\theta) b \log b$, which is Definition 18.10.1 for g.

(b) By (a) and Theorem 21.2.1 with $I = (0, \infty)$, $\mathbf{x} \prec \mathbf{y}$ in A implies $\sum_i x_i \log x_i \le \sum_i y_i \log y_i$, that is, $H(\mathbf{x}) \ge H(\mathbf{y})$. By Definition 21.2.2, H is Schur-concave. (The set A is symmetric and convex, so Definition 21.2.2 applies.)

(c) Let $\mathbf{x} \in A$ with $\sum_i x_i = 1$. Its mean is $\frac{1}{n}$, so the flat vector $\mathbf{f} = (\frac{1}{n}, \dots, \frac{1}{n})$ satisfies $\mathbf{f} \prec \mathbf{x}$ by Example (A chain, the two extremes, and size one) (b), and $\mathbf{f} \in A$. By (b), $H(\mathbf{f}) \ge H(\mathbf{x})$. Finally

$$H(\mathbf{f}) = -\sum_{i=1}^{n} \tfrac{1}{n} \log \tfrac{1}{n} = -\log \tfrac{1}{n} = \log n,$$

using $\log \frac{1}{n} = -\log n$. So $H(\mathbf{x}) \le \log n$, with equality at $\mathbf{x} = \mathbf{f}$.

Exercise. C2: Three norms under majorization

Let $\mathbf{x} \prec \mathbf{y}$ in $\mathbb{R}^n$. Prove that

$$\|\mathbf{x}\|_1 \le \|\mathbf{y}\|_1, \qquad \|\mathbf{x}\|_2 \le \|\mathbf{y}\|_2, \qquad \|\mathbf{x}\|_\infty \le \|\mathbf{y}\|_\infty.$$

Which of the three needs Theorem 21.2.1, and which follows from Definition 17.8.1 alone?

Solution. For $\|\cdot\|_1$, apply Theorem 21.2.1 with $\phi(t) = |t|$, convex on $\mathbb{R}$ by Proposition 18.10.3: $\sum_i |x_i| \le \sum_i |y_i|$.

For $\|\cdot\|_2$, apply Theorem 21.2.1 with $\phi(t) = t^2$, convex on $\mathbb{R}$ by Lemma 18.10.5, since its derivative $2t$ is increasing: $\sum_i x_i^2 \le \sum_i y_i^2$. Both sides are nonnegative and $t \mapsto \sqrt{t}$ is increasing on $[0, \infty)$, so $\|\mathbf{x}\|_2 \le \|\mathbf{y}\|_2$.

For $\|\cdot\|_\infty$, no convexity is needed. Since $\|\mathbf{z}\|_\infty = \max_i |z_i| = \max\{z_1^{\downarrow}, \; -z_n^{\downarrow}\}$, and (M1) at $k = 1$ gives $x_1^{\downarrow} \le y_1^{\downarrow}$ while subtracting (M1) at $k = n-1$ from (M2) gives $x_n^{\downarrow} \ge y_n^{\downarrow}$, hence $-x_n^{\downarrow} \le -y_n^{\downarrow}$. The maximum of the two smaller numbers is at most the maximum of the two larger ones.

So the ∞-norm follows from Definition 17.8.1 directly, and the other two from Theorem 21.2.1. A later section of this chapter proves the same for **every** norm that is invariant under permuting and changing the signs of the coordinates.

Exercise. C3: One convex function is not enough

Let $\mathbf{x} = (2, -1, -1)$ and $\mathbf{y} = (1, 1, -2)$.

(a) Show that $\sum_i x_i = \sum_i y_i$ and $\sum_i x_i^2 = \sum_i y_i^2$, yet neither $\mathbf{x} \prec \mathbf{y}$ nor $\mathbf{y} \prec \mathbf{x}$.

(b) Exhibit a convex ϕ with $\sum_i \phi(x_i) > \sum_i \phi(y_i)$, and another with the inequality reversed.

Explain how this is consistent with Proposition 21.2.2.

Solution.

(a) Both totals are 0, and $\sum_i x_i^2 = 4+1+1 = 6 = 1+1+4 = \sum_i y_i^2$. Both vectors are already decreasing, with running totals $2, 1$ for $\mathbf{x}$ and $1, 2$ for $\mathbf{y}$. (M1) at $k = 1$ fails for $\mathbf{x} \prec \mathbf{y}$, since $2 > 1$, and (M1) at $k = 2$ fails for $\mathbf{y} \prec \mathbf{x}$, since $2 > 1$.

(b) Take $\phi(t) = \max\{t-1, 0\}$, convex by Proposition 18.10.3. Then $\sum_i \phi(x_i) = 1+0+0 = 1$ and $\sum_i \phi(y_i) = 0+0+0 = 0$, so $1 > 0$. This is the test function of Proposition 21.2.2 at $c = y_1^{\downarrow} = 1$, and it certifies $\mathbf{x} \not\prec \mathbf{y}$.

Take $\psi(t) = \max\{t+1, 0\}$. Then $\sum_i \psi(x_i) = 3+0+0 = 3$ and $\sum_i \psi(y_i) = 2+2+0 = 4$, so $3 < 4$, certifying $\mathbf{y} \not\prec \mathbf{x}$ by the same proposition applied with the roles exchanged, at $c = x_2^{\downarrow} = -1$.

There is no contradiction: Proposition 21.2.2 asks the inequality to hold for **every** convex ϕ. Agreeing on the affine functions and on t^2 is far from agreeing on all of them, and here two hockey sticks pull in opposite directions.

21.3 Symmetric Gauge Functions

The last two sections turned a majorization into an inequality between sums of a convex function. This section turns it into an inequality between **norms**. The norms that work are the ones blind to exactly the information a majorization discards — the order of the entries and their signs — and Section 4 will show that these are precisely the norms a matrix inherits from its singular values.

Throughout, the field is $\mathbb{R}$ and the space is $\mathbb{R}^n$ with the dot product $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^\top \mathbf{y}$, $n \geq 1$. Majorization $\prec$ and weak majorization $\prec_w$ are Chapter 17 §08's (Definition 17.8.1), and $\mathbf{x}^{\downarrow}$ is the decreasing rearrangement. A **permutation matrix $\mathbf{P}_\sigma$** is Definition 3.4.2; $\mathbf{P}_\sigma \mathbf{x}$ rearranges the entries of $\mathbf{x}$. We write $|\mathbf{x}| = (|x_1|, \dots, |x_n|)$, and $\mathbf{x} \leq \mathbf{y}$ means $x_i \leq y_i$ for **every** i — the entrywise order of Chapter 18 §05, never the Loewner order. From §01 we use Corollary 21.1.4, which rests on Theorem 21.1.3; the exercises also use §02's Theorem 21.2.1.

21.3.1 What a norm must forget

Chapter 17 §06 measured a Hermitian matrix by its Ky Fan partial sum $s_k(\mathbf{A}) = \lambda_1(\mathbf{A}) + \cdots + \lambda_k(\mathbf{A})$, and Exercise (C2: Partial sums that behave like a norm) asked whether s_k is a norm. On the positive semidefinite matrices it behaves like one; on all Hermitian matrices it fails, because $s_1(\mathrm{diag}(0, -1)) = 0$ with $\mathrm{diag}(0, -1) \neq \mathbf{O}$. The exercise ended by naming the repair: apply s_k to the **singular** values instead of the eigenvalues. Singular values are nonnegative, so nothing can cancel, and the failure disappears.

That repair suggests the shape of the whole chapter's answer. A measurement of a matrix built from its singular values is a function of a list of n nonnegative numbers. Two things about that list are accidents of bookkeeping. Its **order** is one: the singular values of a matrix form a multiset, and Chapter 13 chose to write them decreasingly. Its **signs** are the other: a formula that is to extend from nonnegative lists to all of $\mathbb{R}^n$, and be a norm there, cannot notice which entries are negative, since $\mathbf{x}$ and the vector obtained by flipping one sign have the same absolute values and must be measured alike.

A symmetric gauge function is a norm that cannot tell the entries of a vector apart by where they sit or which way they point.

Definition 21.3.1. Symmetric Gauge Function

A **symmetric gauge function** on $\mathbb{R}^n$ is a norm $\Phi\colon \mathbb{R}^n \to \mathbb{R}$ (Definition 16.1.1) such that

(G1) $\Phi(\mathbf{P}_\sigma \mathbf{x}) = \Phi(\mathbf{x})$ for **every** permutation matrix $\mathbf{P}_\sigma$ and every $\mathbf{x} \in \mathbb{R}^n$ (**permutation invariance**);

(G2) $\Phi(\varepsilon_1 x_1, \dots, \varepsilon_n x_n) = \Phi(\mathbf{x})$ for **all** signs $\varepsilon_i \in \{+1, -1\}$ and every $\mathbf{x} \in \mathbb{R}^n$ (**sign invariance**).

In words: (G1) says that shuffling the entries does not change the measurement, and (G2) says that flipping any of their signs does not either. The word **every** does the work in both, and the two clauses are genuinely separate: below are a norm with (G1) and not (G2), and one with (G2) and not (G1). Together they say that $\Phi(\mathbf{x})$ depends only on the multiset $\{|x_1|, \dots, |x_n|\}$, that is,

$$\Phi(\mathbf{x}) = \Phi\big(|\mathbf{x}|^{\downarrow}\big) \qquad (\mathbf{x} \in \mathbb{R}^n), \tag{21.3.1}$$

since $|\mathbf{x}|$ is obtained from $\mathbf{x}$ by (G2) and $|\mathbf{x}|^{\downarrow}$ from $|\mathbf{x}|$ by (G1). Note also that (N2) of Definition 16.1.1 already gives $\Phi(-\mathbf{x}) = \Phi(\mathbf{x})$; (G2) is much stronger, because it flips the signs **one entry at a time**.

Warning. **The word "gauge" is used here in a different sense from Chapter 18 §04.** There, Definition 18.4.1 attached to a convex set K the function $p_K(\mathbf{x}) = \inf\{t > 0 : \mathbf{x} \in tK\}$, which measures a vector by inflating K. That object is not this one. The name "symmetric gauge function" is standard and traditional for Definition 21.3.1, and the two never appear in the same argument in this book; where confusion is possible, this chapter writes Φ and never p_K.

The first examples are the norms already at hand.

Example. The p-norms, the Ky Fan gauges, and size one

Check each against Definition 21.3.1.

(a) $\|\cdot\|_p$ for $1 \le p \le \infty$.

(b) For $1 \le k \le n$, the function $\Phi_k(\mathbf{x}) = \sum_{i \le k} |x|_i^{\downarrow}$, the sum of the k largest absolute values.

(c) On $\mathbb{R}^1$, the function $\Phi(x) = c|x|$ for a constant $c > 0$.

Solution.

(a) These are norms: Example (The 1-, 2-, and ∞-norms on F^n) for $p = 1, 2, \infty$, and Corollary 18.11.5 supplies the triangle inequality for the rest. The formulas $\big(\sum_i |x_i|^p\big)^{1/p}$ and $\max_i |x_i|$ are built from the absolute values alone and are unchanged by reordering the sum or the maximum, so (G1) and (G2) hold.

(b) Rewrite the definition as a maximum:

$$\Phi_k(\mathbf{x}) = \max\Big\{\sum_{i\in S}|x_i| \;:\; S\subseteq\{1,\dots,n\},\ |S|=k\Big\},$$

since the largest such sum is the one over the indices of the k largest absolute values. Each function $\mathbf{x}\mapsto\sum_{i\in S}|x_i|$ is nonnegative, absolutely homogeneous, and subadditive, so the maximum of finitely many of them is too, which gives (N2) and (N3). For (N1), every singleton is contained in some S, so $\Phi_k(\mathbf{x})\geq\max_i|x_i|$, and $\Phi_k(\mathbf{x})=0$ forces $\mathbf{x}=\mathbf{0}$. Both (G1) and (G2) are visible in the displayed formula. At the two ends, $\Phi_1=\|\cdot\|_\infty$ and $\Phi_n=\|\cdot\|_1$.

(c) A norm on $\mathbb{R}^1$ by inspection. There is only one permutation and $\Phi(-x)=\Phi(x)$, so (G1) and (G2) are automatic: on $\mathbb{R}^1$ **every** norm is a symmetric gauge function. The degenerate case is worth keeping in mind, because it shows that the two clauses carry no information until $n\geq 2$.

The functions in (b) will be used constantly, so they get a name of their own.

Definition 21.3.2. Ky Fan Gauge

For $1\leq k\leq n$, the **k-th Ky Fan gauge** on $\mathbb{R}^n$ is

$$\Phi_k(\mathbf{x}) = \sum_{i=1}^{k}|x|_i^{\downarrow},$$

the sum of the k largest absolute values of the entries of $\mathbf{x}$.

Nonexamples by minimal change. Each drops exactly one clause.

- *A norm with (G1) and not (G2).* On $\mathbb{R}^2$ let

$$\Phi(\mathbf{x}) = \max(|x_1+x_2|,\ |x_1|,\ |x_2|).$$

Each of the three expressions is nonnegative, absolutely homogeneous, and subadditive, so their maximum is too; and $\Phi(\mathbf{x})=0$ forces $x_1=x_2=0$, so Φ is a norm. Swapping x_1 and x_2 permutes the three expressions, so (G1) holds. But $\Phi(1,1)=2$ while $\Phi(1,-1)=1$: (G2) fails.
- *A norm with (G2) and not (G1).* On $\mathbb{R}^2$ let $\Phi(\mathbf{x})=|x_1|+2|x_2|$. This is a norm, and it is built from absolute values, so (G2) holds. But $\Phi(1,0)=1$ and $\Phi(0,1)=2$: (G1) fails.

Why this definition. The two clauses are exactly what a norm needs in order to be computable from an unordered list of nonnegative numbers, and nothing more is demanded — in particular, no relation to any other norm, and no smoothness. Dropping (G1) allows a norm to prefer one coordinate to another, which no measurement of singular values can do, since relabeling the singular values of a matrix is not an operation on the matrix at all. Dropping (G2) allows a norm to see cancellation between coordinates, which is fatal: Section 7 needs Φ evaluated at differences of eigenvalue lists and of singular value lists, whose entries have both signs (Theorem 21.7.2 and Theorem 21.7.4).

Quick check. Is $\Phi(\mathbf{x})=\max_i x_i$ (no absolute values) a symmetric gauge function on $\mathbb{R}^n$ for $n\geq 2$? Which clause fails first?

Solution. It is not, and it fails before (G1) or (G2) are reached: it is not a norm. Absolute homogeneity (N2) fails at $c = -1$, since $\max_i(-x_i) = -\min_i x_i$, which for $\mathbf{x} = (1, 0)$ is 0 while $\max_i x_i = 1$. Positive definiteness (N1) fails too: $\Phi(-1, -1) = -1 < 0$. Restricted to vectors with nonnegative entries the formula is perfectly reasonable, and it equals $\|\mathbf{x}\|_\infty$ there; the definition insists on a norm on all of $\mathbb{R}^n$ precisely so that this kind of one-sided formula is excluded.

21.3.2 A symmetric gauge is monotone

The first payoff is that a symmetric gauge respects the crudest comparison there is: if every entry of $\mathbf{x}$ is smaller in absolute value than the matching entry of $\mathbf{y}$, then $\Phi(\mathbf{x}) \le \Phi(\mathbf{y})$. This is not automatic for a norm. The function $N(\mathbf{x}) = |x_1 - x_2| + |x_1|$ is a norm on $\mathbb{R}^2$, and $N(1, 0) = 2$ is **larger** than $N(1, 1) = 1$ although $(1, 0)$ is dominated entrywise by $(1, 1)$. Sign invariance is what rules this out.

Proposition 21.3.1. Symmetric Gauges Are Monotone

Let Φ be a symmetric gauge function on $\mathbb{R}^n$, and let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ satisfy $|x_i| \le |y_i|$ for every i. Then $\Phi(\mathbf{x}) \le \Phi(\mathbf{y})$.

Idea. Change one coordinate at a time. Shrinking a single entry towards 0 is taking an average of two vectors: the vector itself and the one with that entry's sign flipped. Sign invariance says those two have the same Φ, and the triangle inequality says the average has no larger one.

Proof. **Step 1: one coordinate.** Suppose $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$ agree in every coordinate except the j-th, where $|u_j| \le |v_j|$. If $v_j = 0$ then $u_j = 0$ and $\mathbf{u} = \mathbf{v}$. Otherwise put

$$\theta = \frac{1}{2}\Big(1 + \frac{u_j}{v_j}\Big),$$

which lies in $[0, 1]$ because $|u_j / v_j| \le 1$. Let $\mathbf{v}'$ be $\mathbf{v}$ with its j-th coordinate negated. Then $\theta v_j + (1 - \theta)(-v_j) = u_j$, while in every other coordinate $\mathbf{v}$ and $\mathbf{v}'$ agree with $\mathbf{u}$, so

$$\mathbf{u} = \theta \mathbf{v} + (1 - \theta)\mathbf{v}'.$$

By (N3) and (N2) of Definition 16.1.1, $\Phi(\mathbf{u}) \le \theta\Phi(\mathbf{v}) + (1 - \theta)\Phi(\mathbf{v}')$, and $\Phi(\mathbf{v}') = \Phi(\mathbf{v})$ by (G2). Hence $\Phi(\mathbf{u}) \le \Phi(\mathbf{v})$.

Step 2: all coordinates. Put $\mathbf{z}^{(0)} = \mathbf{y}$ and, for $1 \le j \le n$,

$$\mathbf{z}^{(j)} = (x_1, \dots, x_j, y_{j+1}, \dots, y_n),$$

so that $\mathbf{z}^{(n)} = \mathbf{x}$. Consecutive vectors $\mathbf{z}^{(j-1)}$ and $\mathbf{z}^{(j)}$ agree except in coordinate j, where $|x_j| \le |y_j|$. By Step 1, $\Phi(\mathbf{z}^{(j)}) \le \Phi(\mathbf{z}^{(j-1)})$ for each j. Chaining the n inequalities gives $\Phi(\mathbf{x}) \le \Phi(\mathbf{y})$, as claimed.

■

Only (G2) was used. Permutation invariance is needed for the next step, and there it does all the work.

Quick check. Deduce from Proposition 21.3.1 that $\Phi(\mathbf{x}) \geq |x_j|\,\Phi(\mathbf{e}_j)$ for every j. What does the same proposition give as an upper bound, using $\mathbf{x} = \sum_j x_j \mathbf{e}_j$?

Solution. The vector $x_j\mathbf{e}_j$ has j-th entry x_j and all others 0, so its entries are dominated in absolute value by those of $\mathbf{x}$; Proposition 21.3.1 gives $|x_j|\Phi(\mathbf{e}_j) = \Phi(x_j\mathbf{e}_j) \leq \Phi(\mathbf{x})$. For the upper bound the triangle inequality suffices: $\Phi(\mathbf{x}) \leq \sum_j |x_j|\Phi(\mathbf{e}_j)$. By (G1) all the numbers $\Phi(\mathbf{e}_j)$ are equal, to $\Phi(\mathbf{e}_1)$ say, so

$$\Phi(\mathbf{e}_1)\|\mathbf{x}\|_\infty \;\leq\; \Phi(\mathbf{x}) \;\leq\; \Phi(\mathbf{e}_1)\|\mathbf{x}\|_1.$$

Every symmetric gauge is trapped between multiples of $\|\cdot\|_\infty$ and $\|\cdot\|_1$, with the same constant at both ends. Section 4 turns this into the statement that every **normalized** unitarily invariant norm lies between the spectral norm and the trace norm.

Both bounds are cited later, so they are recorded here.

Corollary 21.3.2. A Symmetric Gauge Is Trapped Between Two Norms

Let Φ be a symmetric gauge function on $\mathbb{R}^n$. Then $\Phi(\mathbf{e}_j) = \Phi(\mathbf{e}_1)$ for every j, and

$$\Phi(\mathbf{e}_1)\|\mathbf{x}\|_\infty \;\leq\; \Phi(\mathbf{x}) \;\leq\; \Phi(\mathbf{e}_1)\|\mathbf{x}\|_1 \qquad (\mathbf{x} \in \mathbb{R}^n).$$

Proof. Fix j. Some permutation matrix carries $\mathbf{e}_j$ to $\mathbf{e}_1$, so $\Phi(\mathbf{e}_j) = \Phi(\mathbf{e}_1)$ by (G1). The entries of $x_j\mathbf{e}_j$ are dominated in absolute value by those of $\mathbf{x}$, so Proposition 21.3.1 gives

$$|x_j|\,\Phi(\mathbf{e}_1) = \Phi(x_j\mathbf{e}_j) \leq \Phi(\mathbf{x}),$$

where the equality is (N2) of Definition 16.1.1; taking the largest $|x_j|$ gives the left-hand inequality. For the right-hand one, $\mathbf{x} = \sum_j x_j\mathbf{e}_j$, so (N3) and (N2) give $\Phi(\mathbf{x}) \leq \sum_j |x_j|\Phi(\mathbf{e}_j) = \Phi(\mathbf{e}_1)\|\mathbf{x}\|_1$. This proves the corollary. ∎

21.3.3 From weak majorization to a dominating vector

Now for the bridge. A majorization $\mathbf{x} \prec \mathbf{y}$ will be handled by Corollary 21.1.4, which writes $\mathbf{x}$ as an average of rearrangements of $\mathbf{y}$; a norm cannot grow under an average. Weak majorization loses the equality of totals, so $\mathbf{x}$ need not be such an average — $(1,0) \prec_w (2,0)$, but $(1,0)$ is no average of $(2,0)$ and $(0,2)$. What survives is that $\mathbf{x}$ is dominated **entrywise** by something that is such an average, and that is exactly the hypothesis Proposition 21.3.1 wants.

Lemma 21.3.3. Weak Majorization Is Domination plus Majorization

Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$. Then $\mathbf{x} \prec_w \mathbf{y}$ if and only if there is a vector $\mathbf{z} \in \mathbb{R}^n$ with

$$\mathbf{x} \le \mathbf{z} \quad \text{entrywise}, \qquad \mathbf{z} \prec \mathbf{y}.$$

If $\mathbf{x}$ has **nonnegative** entries, then $\mathbf{z}$ may be taken with nonnegative entries as well.

Idea. ($\Leftarrow$) is two easy monotonicity steps. For ($\Rightarrow$), sort both vectors and think of the deficit $\sum_i y_i - \sum_i x_i$ as water to be poured into $\mathbf{x}$. Pouring it into the **top** entries would break the partial-sum inequalities immediately, so pour it into the bottom: raise every entry of $\mathbf{x}$ that lies below a level c up to c, and choose c so that the totals match. Choosing c needs no analysis: on each interval between consecutive entries of $\mathbf{x}$ the total is an affine function of c, so the right c can be written down.

Proof. ($\Leftarrow$) Suppose $\mathbf{x} \le \mathbf{z}$ entrywise and $\mathbf{z} \prec \mathbf{y}$. Fix k and let I be a set of k indices carrying the k largest entries of $\mathbf{x}$, so that $\sum_{i \le k} x_i^{\downarrow} = \sum_{i \in I} x_i$. Then

$$\sum_{i \le k} x_i^{\downarrow} = \sum_{i \in I} x_i \;\le\; \sum_{i \in I} z_i \;\le\; \sum_{i \le k} z_i^{\downarrow},$$

the last step because the sum of **any** k entries of $\mathbf{z}$ is at most the sum of its k largest. Hence $\mathbf{x} \prec_w \mathbf{z}$, and $\mathbf{z} \prec \mathbf{y}$ gives $\sum_{i \le k} z_i^{\downarrow} \le \sum_{i \le k} y_i^{\downarrow}$ for $k < n$ and equality of totals at $k = n$; in both cases $\sum_{i \le k} z_i^{\downarrow} \le \sum_{i \le k} y_i^{\downarrow}$. Combining, $\mathbf{x} \prec_w \mathbf{y}$.

($\Rightarrow$) Suppose $\mathbf{x} \prec_w \mathbf{y}$. Majorization and weak majorization depend only on the decreasing rearrangements (Definition 17.8.1 and the remark after it), and the entrywise relation $\mathbf{x} \le \mathbf{z}$ is preserved by applying one permutation to both sides. So it is enough to construct $\mathbf{z}$ when $\mathbf{x} = \mathbf{x}^{\downarrow}$ and $\mathbf{y} = \mathbf{y}^{\downarrow}$: for general $\mathbf{x}$, choose $\mathbf{P}_\sigma$ with $\mathbf{P}_\sigma \mathbf{x} = \mathbf{x}^{\downarrow}$, build $\mathbf{z}'$ for the pair $(\mathbf{x}^{\downarrow}, \mathbf{y}^{\downarrow})$, and take $\mathbf{z} = \mathbf{P}_\sigma^{-1} \mathbf{z}'$.

So let $x_1 \ge \cdots \ge x_n$ and $y_1 \ge \cdots \ge y_n$ with $\sum_{i \le k} x_i \le \sum_{i \le k} y_i$ for every k. Write $S = \sum_{i=1}^{n} y_i$ and, for $0 \le k \le n-1$,

$$c_k = \frac{S - (x_1 + \cdots + x_k)}{n - k},$$

the level at which the last $n-k$ entries would have to sit for the total to be S. At $k = n-1$ we have $c_{n-1} = S - \sum_{i \le n-1} x_i \ge \sum_{i \le n} x_i - \sum_{i \le n-1} x_i = x_n$, using $\mathbf{x} \prec_w \mathbf{y}$ at $k = n$. So the set of indices $k \in \{0, \dots, n-1\}$ with $c_k \ge x_{k+1}$ is nonempty; let j be its **smallest** element and put $c = c_j$. Define

$$\mathbf{z} = (x_1, \dots, x_j, \underbrace{c, \dots, c}_{n-j}).$$

The vector $\mathbf{z}$ *dominates* $\mathbf{x}$. For $i \le j$ the entries agree. For $i > j$, $z_i = c \ge x_{j+1} \ge x_i$, since $\mathbf{x}$ is decreasing and $c \ge x_{j+1}$ by the choice of j. If in addition $\mathbf{x} \ge \mathbf{0}$, then $c \ge x_{j+1} \ge 0$ and every entry of $\mathbf{z}$ is nonnegative, which is the last claim of the lemma.

The vector $\mathbf{z}$ *is decreasing.* Only the step from x_j to c is in question, so assume $j \ge 1$. By

the minimality of j we have $c_{j-1} < x_j$. Multiplying out the two definitions,

$$(n-j+1)c_{j-1} = S - \sum_{i \le j-1} x_i = (n-j)c_j + x_j,$$

so $(n-j)c = (n-j+1)c_{j-1} - x_j < (n-j+1)x_j - x_j = (n-j)x_j$, and $n-j \ge 1$ gives $c < x_j$. Hence $\mathbf{z} = \mathbf{z}^{\downarrow}$.

The vector $\mathbf{z}$ *is majorized by* $\mathbf{y}$. Its total is $\sum_{i \le j} x_i + (n-j)c = S$ by the definition of c_j, which is (M2). For (M1) with $k \le j$, the first k entries of $\mathbf{z}$ are those of $\mathbf{x}$, so $\sum_{i \le k} z_i = \sum_{i \le k} x_i \le \sum_{i \le k} y_i$. For $k > j$, subtract from the totals: since $\sum_i z_i = \sum_i y_i = S$,

$$\sum_{i \le k} z_i - \sum_{i \le k} y_i = \sum_{i > k} y_i - \sum_{i > k} z_i = \sum_{i > k} y_i - (n-k)c,$$

so it suffices to show $\sum_{i>k} y_i \le (n-k)c$. First, at $k = j$ this holds: from $(n-j)c = S - \sum_{i \le j} x_i$ and $\sum_{i \le j} x_i \le \sum_{i \le j} y_i$,

$$(n-j)c \;\ge\; S - \sum_{i \le j} y_i = \sum_{i > j} y_i.$$

So c is at least the average of the $n-j$ smallest entries of $\mathbf{y}$. Now let $k > j$. Each of the entries $y_{j+1}, \dots, y_k$ is at least each of $y_{k+1}, \dots, y_n$, because $\mathbf{y}$ is decreasing; hence each of them is at least the average A of $y_{k+1}, \dots, y_n$, and therefore

$$\sum_{i > j} y_i = \sum_{i=j+1}^{k} y_i + (n-k)A \;\ge\; (k-j)A + (n-k)A = (n-j)A.$$

Combining, $(n-j)c \ge \sum_{i>j} y_i \ge (n-j)A$, so $c \ge A$ and $(n-k)c \ge (n-k)A = \sum_{i>k} y_i$, as required. This proves $\mathbf{z} \prec \mathbf{y}$ and the lemma.

■

The construction is short enough to run by hand.

Example. Filling from the bottom

Let $\mathbf{x} = (4, 1, 1)$ and $\mathbf{y} = (5, 3, 2)$. Verify $\mathbf{x} \prec_w \mathbf{y}$ and produce a $\mathbf{z}$ with $\mathbf{x} \le \mathbf{z}$ and $\mathbf{z} \prec \mathbf{y}$.

Solution. Both vectors are already decreasing. The running totals are $4, 5, 6$ for $\mathbf{x}$ and $5, 8, 10$ for $\mathbf{y}$, and $4 \le 5$, $5 \le 8$, $6 \le 10$, so $\mathbf{x} \prec_w \mathbf{y}$. The totals differ, so $\mathbf{x} \not\prec \mathbf{y}$.

Here $S = 10$ and $n = 3$. The levels are $c_0 = 10/3$ and $c_1 = (10-4)/2 = 3$. Since $c_0 = 10/3 < 4 = x_1$, the index $k = 0$ fails; since $c_1 = 3 \ge 1 = x_2$, the smallest admissible index is $j = 1$, and $c = 3$. Hence

$$\mathbf{z} = (4, 3, 3).$$

Check: $(4, 1, 1) \le (4, 3, 3)$ entrywise; $\mathbf{z}$ is decreasing; its running totals $4, 7, 10$ satisfy $4 \le 5$, $7 \le 8$, $10 = 10$, so $\mathbf{z} \prec \mathbf{y}$. The deficit of 4 between the totals was poured entirely into the two bottom entries, which is what keeps the top partial sums under control.

21.3.4 The inequality

Everything is now in place.

Theorem 21.3.4. Symmetric Gauges Respect Majorization

Let Φ be a symmetric gauge function on $\mathbb{R}^n$.

(a) If $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ and $\mathbf{x} \prec \mathbf{y}$, then $\Phi(\mathbf{x}) \le \Phi(\mathbf{y})$.

(b) If $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ have **nonnegative** entries and $\mathbf{x} \prec_w \mathbf{y}$, then $\Phi(\mathbf{x}) \le \Phi(\mathbf{y})$.

Idea. For (a), §01 turns $\mathbf{x} \prec \mathbf{y}$ into "$\mathbf{x}$ is an average of rearrangements of $\mathbf{y}$", and a norm never exceeds an average of the values it takes on the points averaged — those values all being $\Phi(\mathbf{y})$, by permutation invariance. For (b), Lemma 21.3.3 inserts a middle vector $\mathbf{z}$: from $\mathbf{x}$ to $\mathbf{z}$ use monotonicity, from $\mathbf{z}$ to $\mathbf{y}$ use (a).

Proof.

(a) By Corollary 21.1.4, $\mathbf{x} \in \operatorname{conv}\{\mathbf{P}_\sigma \mathbf{y} : \sigma \in S_n\}$, so there are weights $t_\sigma \ge 0$ with $\sum_\sigma t_\sigma = 1$ and $\mathbf{x} = \sum_\sigma t_\sigma \mathbf{P}_\sigma \mathbf{y}$. By the triangle inequality (N3) and absolute homogeneity (N2) of Definition 16.1.1, applied to this finite sum,

$$\Phi(\mathbf{x}) \;\le\; \sum_\sigma t_\sigma \, \Phi(\mathbf{P}_\sigma \mathbf{y}) = \sum_\sigma t_\sigma \, \Phi(\mathbf{y}) = \Phi(\mathbf{y}),$$

where the middle equality is (G1).

(b) By Lemma 21.3.3 there is a $\mathbf{z}$ **with nonnegative entries** satisfying $\mathbf{x} \le \mathbf{z}$ entrywise and $\mathbf{z} \prec \mathbf{y}$. Since $\mathbf{x}$ and $\mathbf{z}$ have nonnegative entries, $|x_i| = x_i \le z_i = |z_i|$ for every i, so Proposition 21.3.1 gives $\Phi(\mathbf{x}) \le \Phi(\mathbf{z})$. By part (a), $\Phi(\mathbf{z}) \le \Phi(\mathbf{y})$. This proves the theorem.

■

Part (a) needs no sign condition, and part (b) cannot do without one.

Warning. **Part (b) is false for vectors with negative entries.** Take $\mathbf{x} = (0, -10)$ and $\mathbf{y} = (0, 0)$ in $\mathbb{R}^2$. The decreasing rearrangements are $(0, -10)$ and $(0, 0)$, and the running totals $0, -10$ are at most $0, 0$, so $\mathbf{x} \prec_w \mathbf{y}$. But $\Phi(\mathbf{x}) > 0 = \Phi(\mathbf{y})$ for every norm Φ. The lemma is not at fault — $\mathbf{z} = \mathbf{0}$ does satisfy $\mathbf{x} \le \mathbf{z}$ and $\mathbf{z} \prec \mathbf{y}$. What fails is the step $\Phi(\mathbf{x}) \le \Phi(\mathbf{z})$, because Proposition 21.3.1 compares **absolute values**, and $|-10| > |0|$. Weak majorization lets a vector run far in the negative direction, and a norm measures that; only equality of totals, as in $\prec$, or nonnegativity of the entries rules it out. This is why Section 4 applies the theorem to singular values and never to eigenvalues.

The converse of part (b) says that the family of symmetric gauges is strong enough to *detect* weak majorization, and it costs nothing: the Ky Fan gauges Φ_k of Definition 21.3.2 are themselves symmetric gauges (Example (The p-norms, the Ky Fan gauges, and size one) (b)), so the inequalities $\Phi_k(\mathbf{x}) \le \Phi_k(\mathbf{y})$ for every k are among those hypothesized, and for nonnegative

vectors they are the definition of $\mathbf{x} \prec_w \mathbf{y}$. Exercise C1 below states the two directions together, and Section 6 turns them into a theorem about matrices.

21.3.5 The dual gauge

Chapter 18 §04 gave every norm on $\mathbb{R}^n$ a dual, $\|\mathbf{y}\|_* = \max\{\langle \mathbf{x}, \mathbf{y}\rangle : \|\mathbf{x}\| \le 1\}$ (Definition 18.4.4), and proved that taking the dual twice returns the norm (Theorem 18.4.8). Applied to a symmetric gauge, the construction stays inside the class.

Definition 21.3.3. Dual Gauge

Let Φ be a symmetric gauge function on $\mathbb{R}^n$. Its **dual gauge** is the dual norm of Definition 18.4.4,

$$\Phi^*(\mathbf{y}) = \max\{\langle \mathbf{x}, \mathbf{y}\rangle : \mathbf{x} \in \mathbb{R}^n,\ \Phi(\mathbf{x}) \le 1\}.$$

Theorem 21.3.5. The Dual of a Symmetric Gauge

Let Φ be a symmetric gauge function on $\mathbb{R}^n$. Then Φ^* is a symmetric gauge function, and $\Phi^{**} = \Phi$.

Idea. Being a norm and being its own double dual are Chapter 18's, quoted. Only (G1) and (G2) need an argument, and each is the same one: the transformation can be moved off $\mathbf{y}$ and onto $\mathbf{x}$, where it permutes the constraint set $\{\Phi \le 1\}$ onto itself.

Proof. That Φ^* is a norm is Proposition 18.4.6 (a), and $\Phi^{**} = \Phi$ is Theorem 18.4.8; both apply to any norm on $\mathbb{R}^n$.

(G1). Let $\mathbf{P} = \mathbf{P}_\sigma$ be a permutation matrix. By Lemma 3.4.7 (b), $\mathbf{P}^\top = \mathbf{P}^{-1}$, so $\langle \mathbf{x}, \mathbf{P}\mathbf{y}\rangle = \mathbf{x}^\top \mathbf{P}\mathbf{y} = (\mathbf{P}^\top \mathbf{x})^\top \mathbf{y} = \langle \mathbf{P}^\top \mathbf{x}, \mathbf{y}\rangle$. As $\mathbf{x}$ runs over $\{\Phi \le 1\}$, so does $\mathbf{x}' = \mathbf{P}^\top \mathbf{x}$: indeed $\Phi(\mathbf{P}^\top \mathbf{x}) = \Phi(\mathbf{x})$ by (G1) for Φ, and $\mathbf{x} \mapsto \mathbf{P}^\top \mathbf{x}$ is a bijection of $\mathbb{R}^n$. Hence the two maxima are over the same set of numbers, and $\Phi^*(\mathbf{P}\mathbf{y}) = \Phi^*(\mathbf{y})$.

(G2). Fix signs $\varepsilon_i = \pm 1$ and write $\mathbf{y}_\varepsilon = (\varepsilon_1 y_1, \dots, \varepsilon_n y_n)$. Then $\langle \mathbf{x}, \mathbf{y}_\varepsilon\rangle = \sum_i x_i \varepsilon_i y_i = \langle \mathbf{x}_\varepsilon, \mathbf{y}\rangle$, and $\mathbf{x} \mapsto \mathbf{x}_\varepsilon$ is a bijection of $\mathbb{R}^n$ with $\Phi(\mathbf{x}_\varepsilon) = \Phi(\mathbf{x})$ by (G2) for Φ. So again the two maxima range over the same numbers, and $\Phi^*(\mathbf{y}_\varepsilon) = \Phi^*(\mathbf{y})$. This proves the theorem. ∎

So duality is an involution of the set of symmetric gauge functions on $\mathbb{R}^n$. The examples are Chapter 18's: by Proposition 18.4.7 and Theorem 18.11.6,

$$(\|\cdot\|_p)^* = \|\cdot\|_q \qquad \left(\tfrac{1}{p} + \tfrac{1}{q} = 1,\ 1 \le p \le \infty\right),$$

so $\|\cdot\|_1$ and $\|\cdot\|_\infty$ are dual to each other and $\|\cdot\|_2$ is self-dual. In the notation of Definition 21.3.2 that says $\Phi_1^* = \Phi_n$ and $\Phi_n^* = \Phi_1$; exercise C2 below computes Φ_k^* for the k in between. Section 5 will find the matrix meaning of all of this: the dual of a unitarily invariant norm, taken with respect to the trace pairing, is the norm attached to the dual gauge.

21.3.6 Exercises

21.3.6.1 A. Check your understanding

Exercise. A1

(a) State Definition 21.3.1, naming both invariance clauses and the space the function is defined on.

(b) Explain why $\Phi(\mathbf{x}) = \Phi(|\mathbf{x}|^{\downarrow})$ for every symmetric gauge and every $\mathbf{x}$.

(c) True or false, with a reason: if $\mathbf{x} \prec_w \mathbf{y}$ in $\mathbb{R}^n$, then $\|\mathbf{x}\|_2 \le \|\mathbf{y}\|_2$.

(d) Name the hypothesis in Theorem 21.3.4 (b) that part (a) does without, and say where it is used in the proof.

Solution.

(a) A symmetric gauge function on $\mathbb{R}^n$ is a norm $\Phi\colon \mathbb{R}^n \to \mathbb{R}$ satisfying (G1) $\Phi(\mathbf{P}_\sigma \mathbf{x}) = \Phi(\mathbf{x})$ for every permutation matrix, and (G2) $\Phi(\varepsilon_1 x_1, \dots, \varepsilon_n x_n) = \Phi(\mathbf{x})$ for all choices of signs $\varepsilon_i = \pm 1$.

(b) This is Eq. 21.3.1. Replacing each x_i by $|x_i|$ is a choice of signs, so (G2) gives $\Phi(\mathbf{x}) = \Phi(|\mathbf{x}|)$; sorting $|\mathbf{x}|$ decreasingly is applying a permutation matrix, so (G1) gives $\Phi(|\mathbf{x}|) = \Phi(|\mathbf{x}|^{\downarrow})$.

(c) False in general. Take $\mathbf{x} = (0, -10)$ and $\mathbf{y} = (0, 0)$: then $\mathbf{x} \prec_w \mathbf{y}$ but $\|\mathbf{x}\|_2 = 10 > 0$. It is true when $\mathbf{x}$ and $\mathbf{y}$ have nonnegative entries, by Theorem 21.3.4 (b) with $\Phi = \|\cdot\|_2$.

(d) Nonnegativity of the entries. It is used to pass from $\mathbf{x} \le \mathbf{z}$ entrywise to $|x_i| \le |z_i|$, which is what Proposition 21.3.1 requires; without it, $x_i \le z_i$ says nothing about absolute values.

21.3.6.2 B. Practice

Exercise. B1: Which of these are symmetric gauges

Determine which of the following are symmetric gauge functions on $\mathbb{R}^3$. Justify your answer, naming the exact clause that fails when one does.

(a) $\Phi(\mathbf{x}) = |x_1| + |x_2| + |x_3|$.

(b) $\Phi(\mathbf{x}) = |x|_1^{\downarrow} + |x|_2^{\downarrow}$.

(c) $\Phi(\mathbf{x}) = \left(|x_1|^{1/2} + |x_2|^{1/2} + |x_3|^{1/2}\right)^2$.

(d) $\Phi(\mathbf{x}) = \max(|x_1|, |x_2|) + |x_3|$.

(e) $\Phi(\mathbf{x}) = |x_1 + x_2 + x_3| + |x_1 - x_2| + |x_2 - x_3|$.

Solution.

(a) Yes: this is $\|\cdot\|_1 = \Phi_3$, Example (The p-norms, the Ky Fan gauges, and size one) (a).

(b) Yes: this is the Ky Fan gauge Φ_2 of Definition 21.3.2, checked in Example (The p-norms, the Ky Fan gauges, and size one) (b).

(c) No. It is nonnegative, absolutely homogeneous, permutation invariant, and sign invariant, but (N3) fails: with $\mathbf{x} = (1, 0, 0)$ and $\mathbf{y} = (0, 1, 0)$ we get $\Phi(\mathbf{x}) = \Phi(\mathbf{y}) = 1$ while

$\Phi(\mathbf{x}+\mathbf{y}) = (1+1)^2 = 4 > 2$. It is not a norm, so it is not a symmetric gauge.

(d) No. It is a norm — a sum of two seminorms, positive definite because each coordinate appears — and it is sign invariant, so (G2) holds. But (G1) fails: $\Phi(1,1,0) = 1$ while $\Phi(1,0,1) = 2$.

(e) No. It is a norm: it is nonnegative and absolutely homogeneous, subadditive as a sum of three seminorms, and $\Phi(\mathbf{x}) = 0$ forces $x_1 = x_2 = x_3$ and $|3x_1| = 0$, hence $\mathbf{x} = \mathbf{0}$. Both invariance clauses fail. For (G1), $\Phi(1,1,0) = 2{+}0{+}1 = 3$ while $\Phi(1,0,1) = 2{+}1{+}1 = 4$. For (G2), $\Phi(1,1,1) = 3+0+0 = 3$ while $\Phi(1,-1,1) = 1+2+2 = 5$.

Exercise. B2: Filling from the bottom

Let $\mathbf{x} = (6,2,2,0)$ and $\mathbf{y} = (7,4,3,2)$ in $\mathbb{R}^4$.

(a) Verify that $\mathbf{x} \prec_w \mathbf{y}$ and that $\mathbf{x} \not\prec \mathbf{y}$.

(b) Carry out the construction in the proof of Lemma 21.3.3 to produce $\mathbf{z}$ with $\mathbf{x} \le \mathbf{z}$ and $\mathbf{z} \prec \mathbf{y}$.

(c) Hence bound $\|\mathbf{x}\|_2$ by $\|\mathbf{y}\|_2$ without computing either, and then check the bound numerically.

Solution.

(a) Both are decreasing. Running totals: $6, 8, 10, 10$ for $\mathbf{x}$ and $7, 11, 14, 16$ for $\mathbf{y}$. Each is at most the corresponding one, so $\mathbf{x} \prec_w \mathbf{y}$; the totals 10 and 16 differ, so $\mathbf{x} \not\prec \mathbf{y}$.

(b) Here $S = 16$ and $n = 4$, so $c_0 = 16/4 = 4$, $c_1 = (16-6)/3 = 10/3$, $c_2 = (16-8)/2 = 4$, $c_3 = 16-10 = 6$. Test $c_k \ge x_{k+1}$: $c_0 = 4 < 6 = x_1$, so $k = 0$ fails; $c_1 = 10/3 \ge 2 = x_2$, so $j = 1$ and $c = 10/3$. Hence

$$\mathbf{z} = (6, \tfrac{10}{3}, \tfrac{10}{3}, \tfrac{10}{3}).$$

Check: $\mathbf{x} \le \mathbf{z}$ entrywise; $\mathbf{z}$ is decreasing; running totals $6, \frac{28}{3}, \frac{38}{3}, 16$, and $6 \le 7$, $\frac{28}{3} \approx 9.33 \le 11$, $\frac{38}{3} \approx 12.67 \le 14$, $16 = 16$. So $\mathbf{z} \prec \mathbf{y}$.

(c) Both vectors have nonnegative entries, so Theorem 21.3.4 (b) with $\Phi = \|\cdot\|_2$ gives $\|\mathbf{x}\|_2 \le \|\mathbf{y}\|_2$. Numerically $\|\mathbf{x}\|_2 = \sqrt{36+4+4} = \sqrt{44}$ and $\|\mathbf{y}\|_2 = \sqrt{49+16+9+4} = \sqrt{78}$, and $44 \le 78$. The intermediate vector satisfies $\|\mathbf{z}\|_2^2 = 36 + 3 \cdot \frac{100}{9} = \frac{208}{3} \approx 69.3$, between the two, as the proof predicts.

Exercise. B3: Monotonicity used twice

Let Φ be a symmetric gauge function on $\mathbb{R}^3$ with $\Phi(\mathbf{e}_1) = 1$.

(a) Prove that $1 \le \Phi(1,1,0) \le 2$.

(b) Prove that $\Phi(2,1,1) \le \Phi(3,1,0)$.

(c) Give a symmetric gauge for which the inequality in (b) is strict, and one for which it is an equality.

Solution.

(a) The Quick check following Proposition 21.3.1 gives $\Phi(\mathbf{e}_1)\|\mathbf{x}\|_\infty \le \Phi(\mathbf{x}) \le \Phi(\mathbf{e}_1)\|\mathbf{x}\|_1$. With $\mathbf{x} = (1,1,0)$ and $\Phi(\mathbf{e}_1) = 1$ this reads $1 \le \Phi(1,1,0) \le 2$.

(b) Both vectors are decreasing with total 4, and the running totals $2, 3$ of $(2,1,1)$ are at most the running totals $3, 4$ of $(3,1,0)$, so $(2,1,1) \prec (3,1,0)$. Now Theorem 21.3.4 (a) applies.

(c) Strict for $\Phi = \|\cdot\|_2$: $\sqrt{6} < \sqrt{10}$. It is also strict for $\Phi = \Phi_2$, where the values are 3 and 4. Equality holds for $\Phi = \|\cdot\|_1 = \Phi_3$, where both values are 4: any two nonnegative vectors with the same total have the same 1-norm, so a majorization can never be detected by Φ_3 alone. This is the general shape of the equality case — the top gauge sees only the total, which a majorization preserves.

21.3.6.3 C. Going deeper

Exercise. C1: Symmetric gauges detect weak majorization

Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ have nonnegative entries.

(a) Prove that $\mathbf{x} \prec_w \mathbf{y}$ if and only if $\Phi(\mathbf{x}) \le \Phi(\mathbf{y})$ for **every** symmetric gauge function Φ on $\mathbb{R}^n$.

(b) Prove that $\mathbf{x} \prec \mathbf{y}$ if and only if $\mathbf{x} \prec_w \mathbf{y}$ and $\sum_i x_i = \sum_i y_i$.

(c) Deduce that $\mathbf{x} \prec \mathbf{y}$ if and only if $\Phi(\mathbf{x}) \le \Phi(\mathbf{y})$ for every symmetric gauge and $\sum_i x_i = \sum_i y_i$.

Hint for (a): only finitely many gauges are needed.

Solution.

(a) ($\Rightarrow$) is Theorem 21.3.4 (b). ($\Leftarrow$) Apply the hypothesis to the Ky Fan gauges Φ_k of Definition 21.3.2, $k = 1, \dots, n$. Since $\mathbf{x}$ and $\mathbf{y}$ have nonnegative entries, $|\mathbf{x}|^\downarrow = \mathbf{x}^\downarrow$ and likewise for $\mathbf{y}$, so $\Phi_k(\mathbf{x}) = \sum_{i \le k} x_i^\downarrow$ and $\Phi_k(\mathbf{y}) = \sum_{i \le k} y_i^\downarrow$. The inequalities $\Phi_k(\mathbf{x}) \le \Phi_k(\mathbf{y})$ for $k = 1, \dots, n$ are exactly (M1) extended to $k = n$, which is Definition 17.8.1's $\mathbf{x} \prec_w \mathbf{y}$.

(b) This is a restatement of Definition 17.8.1: $\mathbf{x} \prec \mathbf{y}$ means (M1) for $k \le n-1$ together with equality of totals, and $\mathbf{x} \prec_w \mathbf{y}$ means (M1) for $k \le n$. Given equality of totals, the case $k = n$ of (M1) is automatic, so the two lists of conditions agree.

(c) Combine (a) and (b): $\mathbf{x} \prec \mathbf{y}$ holds exactly when $\mathbf{x} \prec_w \mathbf{y}$ and the totals are equal, and by (a) the first of these is equivalent to $\Phi(\mathbf{x}) \le \Phi(\mathbf{y})$ for every symmetric gauge.

Exercise. C2: The dual of a Ky Fan gauge

Fix $1 \le k \le n$ and let Φ_k be the Ky Fan gauge of Definition 21.3.2.

(a) Prove that $\Phi_k^*(\mathbf{y}) \ge \max(\|\mathbf{y}\|_\infty, \ \frac{1}{k}\|\mathbf{y}\|_1)$ by exhibiting two competitors.

(b) Let $\Phi_k(\mathbf{x}) \le 1$ and let $\mathbf{y}$ have nonnegative, decreasing entries. Put $t = |x|_k^\downarrow$ and split each $|x_i|$ as $(|x_i| - t)_+ + \min(|x_i|, t)$. Prove that $\sum_i (|x_i| - t)_+ \le 1 - kt$ and that

$0 \le t \le 1/k$.

(c) Deduce $\Phi_k^*(\mathbf{y}) = \max(\|\mathbf{y}\|_\infty, \frac{1}{k}\|\mathbf{y}\|_1)$, and check the cases $k = 1$ and $k = n$ against Proposition 18.4.7.

Solution.

(a) The vector $\mathbf{e}_1$ has $\Phi_k(\mathbf{e}_1) = 1$; replacing it by $\pm\mathbf{e}_j$ with the sign of y_j, for a j where $|y_j|$ is largest, gives $\langle \mathbf{x}, \mathbf{y}\rangle = \|\mathbf{y}\|_\infty$. The vector $\mathbf{x}$ with $x_i = \pm 1/k$, the sign of y_i, has $\Phi_k(\mathbf{x}) = k \cdot \frac{1}{k} = 1$ and $\langle \mathbf{x}, \mathbf{y}\rangle = \frac{1}{k}\|\mathbf{y}\|_1$. So $\Phi_k^*(\mathbf{y})$ is at least both numbers.

(b) Since $|x|^\downarrow$ is decreasing, $(|x|_i^\downarrow - t)_+ = 0$ for $i > k$, so

$$\sum_i (|x_i| - t)_+ = \sum_{i \le k} (|x|_i^\downarrow - t) = \Phi_k(\mathbf{x}) - kt \le 1 - kt.$$

Also $kt \le \sum_{i \le k} |x|_i^\downarrow = \Phi_k(\mathbf{x}) \le 1$, so $0 \le t \le 1/k$.

(c) By Theorem 21.3.5, Φ_k^* is a symmetric gauge, so its value at $\mathbf{y}$ depends only on $|\mathbf{y}|^\downarrow$; and replacing $\mathbf{x}$ by $|\mathbf{x}|$ leaves $\Phi_k(\mathbf{x})$ unchanged and does not decrease $\langle \mathbf{x}, \mathbf{y}\rangle$ when $\mathbf{y} \ge \mathbf{0}$. So we may assume $\mathbf{y}$ has nonnegative decreasing entries and $\mathbf{x} \ge \mathbf{0}$. With t as in (b), and using $y_i \le y_1 = \|\mathbf{y}\|_\infty$ in the first sum and $\min(x_i, t) \le t$ in the second,

$$\langle \mathbf{x}, \mathbf{y}\rangle \le \|\mathbf{y}\|_\infty (1 - kt) + t\|\mathbf{y}\|_1 = \|\mathbf{y}\|_\infty + t(\|\mathbf{y}\|_1 - k\|\mathbf{y}\|_\infty).$$

If $\|\mathbf{y}\|_1 \le k\|\mathbf{y}\|_\infty$, the bracket is at most 0 and $t \ge 0$, so the right side is at most $\|\mathbf{y}\|_\infty$. If $\|\mathbf{y}\|_1 > k\|\mathbf{y}\|_\infty$, the right side increases with t, and $t \le 1/k$ gives the bound $\|\mathbf{y}\|_\infty + \frac{1}{k}\|\mathbf{y}\|_1 - \|\mathbf{y}\|_\infty = \frac{1}{k}\|\mathbf{y}\|_1$. In both cases $\langle \mathbf{x}, \mathbf{y}\rangle \le \max(\|\mathbf{y}\|_\infty, \frac{1}{k}\|\mathbf{y}\|_1)$, which with (a) gives equality.

For $k = 1$, $\Phi_1 = \|\cdot\|_\infty$ and the formula gives $\max(\|\mathbf{y}\|_\infty, \|\mathbf{y}\|_1) = \|\mathbf{y}\|_1$, matching Proposition 18.4.7 (b). For $k = n$, $\Phi_n = \|\cdot\|_1$ and the formula gives $\max(\|\mathbf{y}\|_\infty, \frac{1}{n}\|\mathbf{y}\|_1) = \|\mathbf{y}\|_\infty$, matching part (a) of the same proposition.

Exercise. C3: Where each hypothesis is used

(a) Determine whether there is a norm on $\mathbb{R}^2$ that is permutation invariant and satisfies $\Phi(\mathbf{x}) \le \Phi(\mathbf{y})$ whenever $|x_i| \le |y_i|$ for all i, but is **not** a symmetric gauge function. Justify your answer.

(b) Theorem 21.3.4 (b) needs nonnegative entries, while Lemma 21.3.3 does not. Locate the exact step of the proof of the theorem that uses the hypothesis, and explain why $\mathbf{x} = (0, -10) \prec_w (0, 0) = \mathbf{y}$ breaks it even though the lemma supplies a perfectly good $\mathbf{z}$.

(c) Let $\mathbf{x} \prec_w \mathbf{y}$ with $\mathbf{x} = \mathbf{x}^\downarrow$, and let $\mathbf{z}$ be the vector built in the proof of Lemma 21.3.3. Prove that $\mathbf{z} \prec \mathbf{z}'$ for **every** $\mathbf{z}'$ with $\mathbf{x} \le \mathbf{z}'$ and $\sum_i z_i' = \sum_i y_i$, and deduce that $\sum_i \phi(z_i) \le \sum_i \phi(z_i')$ for every convex ϕ. *Hint: compare top partial sums; Theorem 21.2.1 does the last step.*

Solution.

(a) There is none: monotonicity in the stated sense already forces (G2). If $|x_i| \le |y_i|$

implies $\Phi(\mathbf{x}) \le \Phi(\mathbf{y})$, then for any signs the vectors $\mathbf{x}$ and $(\varepsilon_1 x_1, \dots, \varepsilon_n x_n)$ dominate each other entrywise in absolute value, so their Φ-values are equal, which is (G2). With permutation invariance this is Definition 21.3.1. So the conclusion of Proposition 21.3.1 is not merely implied by (G2): for any norm the two are equivalent.

(b) The step is $\Phi(\mathbf{x}) \le \Phi(\mathbf{z})$, which invokes Proposition 21.3.1 and therefore needs $|x_i| \le |z_i|$, not $x_i \le z_i$. For $\mathbf{x} = (0, -10)$ and $\mathbf{y} = (0, 0)$ the lemma does supply a $\mathbf{z}$, namely $\mathbf{z} = \mathbf{0}$: indeed $\mathbf{z} \prec \mathbf{y}$ and $\mathbf{x} \le \mathbf{z}$. But $|x_2| = 10 > 0 = |z_2|$, so Proposition 21.3.1 does not apply, and it must not: $\Phi(\mathbf{x}) > 0 = \Phi(\mathbf{z})$ for every norm. When the entries are nonnegative, $x_i \le z_i$ and $|x_i| \le |z_i|$ say the same thing, and the gap closes.

(c) Write $S = \sum_i y_i$ and recall $\mathbf{z} = (x_1, \dots, x_j, c, \dots, c)$, decreasing, with $\sum_i z_i = S$. Let $\mathbf{z}'$ satisfy $\mathbf{x} \le \mathbf{z}'$ and $\sum_i z'_i = S$, and write $M_1 \ge \cdots \ge M_n$ for its entries in decreasing order. Since the totals agree, (M2) holds, and it remains to check (M1) for $1 \le k \le n-1$.

Case $k \le j$. Then $\sum_{i \le k} z_i = \sum_{i \le k} x_i \le \sum_{i \le k} z'_i \le \sum_{i \le k} M_i$, the first inequality from $\mathbf{x} \le \mathbf{z}'$ and the second because no k entries of $\mathbf{z}'$ add up to more than its k largest.

Case $k > j$. Here $\sum_{i \le k} z_i = S - (n-k)c$, so the claim is $\sum_{i > k} M_i \le (n-k)c$. Suppose not. Then the average of $M_{k+1}, \dots, M_n$ exceeds c, and M_{k+1}, being the largest of them, satisfies $M_{k+1} > c$; hence $M_i > c$ for every $i \le k+1$. Also $\sum_{i \le j} M_i \ge \sum_{i \le j} x_i$, by the argument of the previous case. Therefore

$$\begin{aligned} S = \sum_i M_i &= \sum_{i \le j} M_i + \sum_{i=j+1}^{k} M_i + \sum_{i > k} M_i \\ &> \sum_{i \le j} x_i + (k-j)c + (n-k)c = S, \end{aligned}$$

the last equality being $\sum_{i \le j} x_i + (n-j)c = S$, which defined c. This contradiction proves the case.

So $\mathbf{z} \prec \mathbf{z}'$, and Theorem 21.2.1 gives $\sum_i \phi(z_i) \le \sum_i \phi(z'_i)$ for every convex ϕ. In words: among all admissible ways of raising $\mathbf{x}$ to the required total, the flat filling is the most even one.

21.4 Unitarily Invariant Norms

Chapter 16 ended its survey of norms with a list of what it had left undone, and one entry read: "It has not said which norms are invariant under unitary multiplication, which is Chapter 21." This section says it. The answer is that such a norm is nothing but a symmetric gauge function applied to the singular values, so §03's inequality transfers wholesale from vectors to matrices — and the Ky Fan functionals that failed to be norms on Hermitian matrices become norms as soon as they are read on singular values instead.

Throughout, $F = \mathbb{R}$ **or** $F = \mathbb{C}$, $\mathbf{A}, \mathbf{B} \in M_{m \times n}(F)$, $p = \min(m, n)$, and the decreasing list of singular values (Definition 13.9.1) is $\sigma(\mathbf{A}) = (\sigma_1(\mathbf{A}), \dots, \sigma_p(\mathbf{A}))$. Symmetric gauge functions are §03's (Definition 21.3.1) and live on $\mathbb{R}^p$. For $\mathbf{x} \in \mathbb{R}^p$ we write

$$\operatorname{diag}_{m,n}(\mathbf{x}) \in M_{m \times n}(F)$$

for the matrix with $x_1, \dots, x_p$ down the main diagonal and every other entry 0; for $m = n$ this is the usual $\operatorname{diag}(x_1, \dots, x_n)$. Finally, one consequence of Theorem 13.9.1 and Theorem 13.9.2 is used constantly: if $\mathbf{A} = \mathbf{W}\mathbf{\Sigma}\mathbf{X}^*$ is a singular value decomposition and $\mathbf{U} \in M_m(F)$, $\mathbf{V} \in M_n(F)$

are unitary, then

$$\mathbf{UAV} = (\mathbf{UW})\,\mathbf{\Sigma}\,(\mathbf{V}^*\mathbf{X})^*$$

is again of that shape, with $\mathbf{UW}$ and $\mathbf{V}^*\mathbf{X}$ unitary, so

$$\sigma(\mathbf{UAV}) = \sigma(\mathbf{A}). \tag{21.4.1}$$

21.4.1 Norms that do not see the unitary factors

Two matrix norms in the book already ignore unitary factors. The Frobenius norm does, by Lemma 11.7.5; the spectral norm does, by Lemma 20.9.1 (c). Both facts were proved separately and both were used the same way: to replace $\mathbf{A}$ by $\mathbf{\Sigma}$ and reduce a question about a matrix to a question about a list of numbers. That is a property worth isolating, and it is the property Chapter 16 left unnamed.

A norm is unitarily invariant when rotating the domain and the codomain does not change the measurement.

Definition 21.4.1. Unitarily Invariant Norm

A **unitarily invariant norm** on $M_{m\times n}(F)$ is a norm $|||\cdot|||$ on $M_{m\times n}(F)$, regarded as a vector space over F, such that

$$|||\mathbf{UAV}||| = |||\mathbf{A}|||$$

for **every** $\mathbf{A} \in M_{m\times n}(F)$ and **all** unitary $\mathbf{U} \in M_m(F)$, $\mathbf{V} \in M_n(F)$.

The triple bars are standard and are used from here on for a norm known to be unitarily invariant; a single bar with a subscript, as in $\|\mathbf{A}\|_2$ or $\|\mathbf{A}\|_F$, keeps naming a specific one. In words: the condition is not invariance under similarity, and not invariance under transposition; it is invariance under multiplying on the **left** by any unitary and on the **right** by any unitary, independently. The two unitaries are unrelated, which is what makes the condition strong. Asking only for invariance under $\mathbf{A} \mapsto \mathbf{UAU}^*$, with the same unitary on both sides, is a much weaker demand: $|\operatorname{tr}\mathbf{A}|$ has it, since $\operatorname{tr}(\mathbf{UAU}^*) = \operatorname{tr}\mathbf{A}$ by Theorem 0.9.6 (3), yet it is not determined by $\sigma(\mathbf{A})$, because $\mathbf{I}_2$ and $\operatorname{diag}(1,-1)$ both have singular values $1, 1$ while their traces have absolute values 2 and 0.

Example. The two familiar norms

Check each against Definition 21.4.1.

(a) The Frobenius norm $\|\mathbf{A}\|_F$.

(b) The spectral norm $\|\mathbf{A}\|_2$.

Solution.

(a) It is a norm, being the norm of the Frobenius inner product of Chapter 11 §01, and $\|\mathbf{UAV}\|_F = \|\mathbf{A}\|_F$ is Lemma 11.7.5.

(b) It is a norm by Theorem 16.3.2 (b), and Lemma 20.9.1 (c) gives $\|\mathbf{UAV}\|_2 = \|\mathbf{A}\|_2$ for unitary $\mathbf{U}, \mathbf{V}$. By Theorem 16.3.3 (c), $\|\mathbf{A}\|_2 = \sigma_1(\mathbf{A})$, so this norm too is a function of $\sigma(\mathbf{A})$ alone — and Eq. 21.4.1 gives a second proof of its invariance.

Both are built from $\sigma(\mathbf{A})$. The largest singular value and the square root of the sum of their squares are two entries in a whole family, and the family is the point of the section.

Definition 21.4.2. Ky Fan k-Norms

For $1 \le k \le p$, the **Ky Fan k-norm** of $\mathbf{A} \in M_{m\times n}(F)$ is

$$|||\mathbf{A}|||_{(k)} = \sum_{i=1}^{k} \sigma_i(\mathbf{A}),$$

the sum of the k largest singular values. For $k = 1$ it is the spectral norm $\|\mathbf{A}\|_2$; for $k = p$ it is called the **trace norm**.

Chapter 17 §06 measured a Hermitian matrix by $s_k(\mathbf{A}) = \lambda_1(\mathbf{A}) + \cdots + \lambda_k(\mathbf{A})$ and found in Exercise (C2: Partial sums that behave like a norm) that this is no norm: $s_1(\mathrm{diag}(0,-1)) = 0$ although $\mathrm{diag}(0,-1) \neq \mathbf{O}$. The exercise said that "the repair is to apply N_k [its name there for s_k] to the singular values instead of the eigenvalues, which is what Chapter 21 does". Definition 21.4.2 is that repair, and the repair works. Positive definiteness now holds because singular values are nonnegative, so they can only sum to 0 by all vanishing, which by Theorem 13.9.1 forces $\mathbf{A} = \mathbf{O}$; absolute homogeneity holds because $(c\mathbf{A})^*(c\mathbf{A}) = |c|^2\mathbf{A}^*\mathbf{A}$ gives $\sigma(c\mathbf{A}) = |c|\sigma(\mathbf{A})$ through Definition 13.9.1; and unitary invariance is Eq. 21.4.1. The triangle inequality is the one axiom still outstanding, and Lemma 21.4.2 below supplies it.

Nonexamples by minimal change. Both are honest norms on $M_2(\mathbb{R})$ that fail only the invariance clause, and one witness refutes both. Let

$$\mathbf{U} = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix},$$

a rotation through half a right angle, which is orthogonal since its columns are orthonormal.

- *The entrywise maximum* $N(\mathbf{A}) = \max_{i,j}|a_{ij}|$ is a norm (it is the ∞-norm of the entry list). But $N(\mathbf{I}_2) = 1$ while $N(\mathbf{U}\mathbf{I}_2\mathbf{I}_2) = N(\mathbf{U}) = 1/\sqrt{2}$.
- *The operator* 1*-norm* $\|\mathbf{A}\|_1$, the largest column sum of absolute values (Theorem 16.3.3 (a)), is a norm. But $\|\mathbf{I}_2\|_1 = 1$ while $\|\mathbf{U}\|_1 = 1/\sqrt{2} + 1/\sqrt{2} = \sqrt{2}$.

Both failures are visible in the same picture: $\mathbf{I}_2$ and $\mathbf{U}$ have the same singular values, namely $1, 1$, because both are orthogonal, and a unitarily invariant norm is therefore forced to give them the same value. A norm that reads individual entries or individual columns is reading a basis, and a unitary change of basis is exactly what Definition 21.4.1 forbids it to notice.

Why this definition. The clause is exactly what makes a norm a function of the singular values and nothing else, as the next lemma shows; that is its whole purpose. Weakening it to one-sided invariance $|||\mathbf{U}\mathbf{A}||| = |||\mathbf{A}|||$ is not enough. The largest Euclidean length among the columns of $\mathbf{A}$ is a norm with that property, since a unitary matrix preserves the length of each column; but it gives $\mathrm{diag}(2,1)$ the value 2, while $\mathrm{diag}(2,1)\mathbf{U}^\top = \frac{1}{\sqrt{2}}\left(\begin{smallmatrix} 2 & 2 \\ -1 & 1 \end{smallmatrix}\right)$ has both columns of length $\sqrt{5/2}$, and the two matrices have the same singular values. Strengthening it to include all invertible $\mathbf{U}, \mathbf{V}$ is far too much, since every nonzero matrix of a given rank would then have the same norm.

Quick check. Is $\mathbf{A} \mapsto \sigma_2(\mathbf{A})$ a unitarily invariant norm on $M_2(\mathbb{R})$? Is $\mathbf{A} \mapsto \sigma_1(\mathbf{A}) - \sigma_2(\mathbf{A})$?

Solution. Neither, and both fail (N1) of Definition 16.1.1 rather than the invariance clause: they are unitarily invariant by Eq. 21.4.1, but they are not norms. The matrix $\mathbf{A} = \mathbf{e}_1\mathbf{e}_1^\top$ has $\sigma(\mathbf{A}) = (1, 0)$, so $\sigma_2(\mathbf{A}) = 0$ with $\mathbf{A} \neq \mathbf{O}$. The identity has $\sigma(\mathbf{I}_2) = (1, 1)$, so $\sigma_1(\mathbf{I}_2) - \sigma_2(\mathbf{I}_2) = 0$ with $\mathbf{I}_2 \neq \mathbf{O}$. This is the same failure Chapter 17 met with s_k, and it says that not every symmetric-looking formula in the singular values is a norm; the correspondence below says exactly which are.

21.4.2 A unitarily invariant norm is a function of the singular values

Lemma 21.4.1. The Gauge of a Unitarily Invariant Norm

Let $\|\cdot\|$ be a unitarily invariant norm on $M_{m\times n}(F)$ and define

$$\Phi(\mathbf{x}) = \left\|\operatorname{diag}_{m,n}(\mathbf{x})\right\| \qquad (\mathbf{x} \in \mathbb{R}^p).$$

Then Φ is a symmetric gauge function on $\mathbb{R}^p$, and

$$\|\mathbf{A}\| = \Phi(\sigma(\mathbf{A})) \qquad \text{for every } \mathbf{A} \in M_{m\times n}(F).$$

In particular $\|\mathbf{A}\|$ depends on $\mathbf{A}$ only through $\sigma(\mathbf{A})$.

Idea. The map $\mathbf{x} \mapsto \operatorname{diag}_{m,n}(\mathbf{x})$ is linear and injective, so the norm axioms for Φ are inherited entry for entry. The two invariance clauses are the two families of unitary matrices that act diagonally: a permutation matrix on each side rearranges the diagonal, a diagonal matrix of signs on one side flips its signs. The last formula is the singular value decomposition, read once.

Proof. **Φ is a norm.** The map $L(\mathbf{x}) = \operatorname{diag}_{m,n}(\mathbf{x})$ is linear from $\mathbb{R}^p$ to $M_{m\times n}(F)$ and injective, since $L(\mathbf{x}) = \mathbf{O}$ reads off as $\mathbf{x} = \mathbf{0}$. Hence $\Phi = \|\cdot\| \circ L$ satisfies (N1) — $\Phi(\mathbf{x}) = 0$ gives $L(\mathbf{x}) = \mathbf{O}$ and $\mathbf{x} = \mathbf{0}$ — and (N2) for real scalars, and (N3), all directly from the corresponding axioms for $\|\cdot\|$.

(G1). The letter σ is taken here, so write π for a permutation of $\{1, \dots, p\}$. Extend it to a permutation $\widehat{\pi}$ of $\{1, \dots, m\}$ and to one $\widetilde{\pi}$ of $\{1, \dots, n\}$ by fixing every index greater than p, and let $\mathbf{P} \in M_m(F)$ and $\mathbf{Q} \in M_n(F)$ be their permutation matrices (Definition 3.4.2), which are unitary by Lemma 3.4.7 (b). Writing $\mathbf{D} = \operatorname{diag}_{m,n}(\mathbf{x})$, multiplying by $\mathbf{P}$ on the left moves row k to row $\widehat{\pi}(k)$ and multiplying by $\mathbf{Q}^\top$ on the right moves column l to column $\widetilde{\pi}(l)$; no entry is created or destroyed. The only nonzero entries of $\mathbf{D}$ are $d_{kk} = x_k$ for $k \le p$, and they land in positions $(\widehat{\pi}(k), \widetilde{\pi}(k)) = (\pi(k), \pi(k))$, again on the main diagonal. Hence

$$\mathbf{P}\,\operatorname{diag}_{m,n}(\mathbf{x})\,\mathbf{Q}^\top = \operatorname{diag}_{m,n}(\mathbf{P}_\pi \mathbf{x}).$$

Taking $\|\cdot\|$ of both sides and using Definition 21.4.1 gives $\Phi(\mathbf{P}_\pi \mathbf{x}) = \Phi(\mathbf{x})$.

(G2). Let $\varepsilon_1, \dots, \varepsilon_p$ be signs. The matrix $\operatorname{diag}(\varepsilon_1, \dots, \varepsilon_p, 1, \dots, 1) \in M_m(F)$ is real diagonal with entries of modulus 1, hence unitary, and multiplying on the left by it multiplies row k

by ε_k. So it carries $\operatorname{diag}_{m,n}(\mathbf{x})$ to $\operatorname{diag}_{m,n}(\varepsilon_1 x_1, \dots, \varepsilon_p x_p)$, and invariance gives (G2).

The formula. Let $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$ be a singular value decomposition (Theorem 13.9.1). Then $\mathbf{\Sigma} = \operatorname{diag}_{m,n}(\sigma(\mathbf{A}))$, and $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$ with $\mathbf{U}$ and $\mathbf{V}^*$ unitary, so $|||\mathbf{A}||| = |||\mathbf{\Sigma}||| = \Phi(\sigma(\mathbf{A}))$. This proves the lemma.

■

So the map $|||\cdot||| \mapsto \Phi$ loses nothing. The question is whether it is surjective: does every symmetric gauge come from a unitarily invariant norm? Evaluating Φ at $\sigma(\mathbf{A})$ certainly produces a unitarily invariant *function*, by Eq. 21.4.1, and (N1) and (N2) are easy. Everything hangs on the triangle inequality, and the triangle inequality hangs on one majorization.

21.4.3 Singular values are subadditive

Lemma 21.4.2. Weak Subadditivity of the Singular Values

Let $\mathbf{A}, \mathbf{B} \in M_{m\times n}(F)$. Then

$$\sigma(\mathbf{A}+\mathbf{B}) \prec_w \sigma(\mathbf{A}) + \sigma(\mathbf{B}),$$

that is, $\displaystyle\sum_{i=1}^{k} \sigma_i(\mathbf{A}+\mathbf{B}) \le \sum_{i=1}^{k} (\sigma_i(\mathbf{A}) + \sigma_i(\mathbf{B}))$ for every $1 \le k \le p$.

Idea. The statement is about singular values, and the tool — Corollary 17.6.5, which says that the top k eigenvalues of a Hermitian sum are subadditive — is about eigenvalues. The Hermitian dilation $\mathcal{H}$ of Proposition 17.9.3 converts one into the other, and it does so **linearly**: $\mathcal{H}(\mathbf{A}+\mathbf{B}) = \mathcal{H}(\mathbf{A}) + \mathcal{H}(\mathbf{B})$, because the blocks add. Since the top p eigenvalues of $\mathcal{H}(\mathbf{A})$ are exactly $\sigma_1(\mathbf{A}), \dots, \sigma_p(\mathbf{A})$, the corollary transfers verbatim.

Proof. Write $\mathcal{H}(\mathbf{M}) = \left(\begin{smallmatrix} \mathbf{O} & \mathbf{M} \\ \mathbf{M}^* & \mathbf{O} \end{smallmatrix}\right) \in M_{m+n}(F)$ for the Hermitian dilation of Proposition 17.9.3. Adding block by block, $\mathcal{H}(\mathbf{A}+\mathbf{B}) = \mathcal{H}(\mathbf{A}) + \mathcal{H}(\mathbf{B})$.

By Proposition 17.9.3 the eigenvalues of $\mathcal{H}(\mathbf{M})$, with multiplicity, are $\sigma_1(\mathbf{M}), \dots, \sigma_p(\mathbf{M})$, then $m+n-2p$ zeros, then $-\sigma_p(\mathbf{M}), \dots, -\sigma_1(\mathbf{M})$. Every singular value is ≥ 0 and every listed negative number is ≤ 0, so this list, read in the order written, is already decreasing. Hence

$$\lambda_i(\mathcal{H}(\mathbf{M})) = \sigma_i(\mathbf{M}) \qquad (1 \le i \le p),$$

for every $\mathbf{M} \in M_{m\times n}(F)$.

Now fix $1 \le k \le p$. The matrices $\mathcal{H}(\mathbf{A})$ and $\mathcal{H}(\mathbf{B})$ are Hermitian of the same size, so Corollary 17.6.5 applies to them:

$$\sum_{i=1}^{k} \lambda_i(\mathcal{H}(\mathbf{A}) + \mathcal{H}(\mathbf{B})) \;\le\; \sum_{i=1}^{k} \lambda_i(\mathcal{H}(\mathbf{A})) + \sum_{i=1}^{k} \lambda_i(\mathcal{H}(\mathbf{B})).$$

Replacing $\mathcal{H}(\mathbf{A}) + \mathcal{H}(\mathbf{B})$ by $\mathcal{H}(\mathbf{A}+\mathbf{B})$ on the left and using the displayed identity three

times turns this into

$$\sum_{i=1}^{k}\sigma_i(\mathbf{A}+\mathbf{B}) \;\le\; \sum_{i=1}^{k}\sigma_i(\mathbf{A})+\sum_{i=1}^{k}\sigma_i(\mathbf{B}).$$

Finally, $\sigma(\mathbf{A})+\sigma(\mathbf{B})$ is a sum of two decreasing vectors, hence decreasing, so it is its own decreasing rearrangement, and the displayed inequalities for $k=1,\dots,p$ are precisely (M1) for $\prec_w$ in Definition 17.8.1. This proves the lemma.

■

Two remarks on the route. The dilation was chosen because it reduces the statement to a corollary already proved, in four lines and with no new variational argument; the alternative is to establish a Ky Fan maximum principle for singular values directly, maximizing $\sum_{i\le k}|\langle \mathbf{A}\mathbf{w}_i,\mathbf{q}_i\rangle|$ over pairs of orthonormal lists, which is a genuine theorem in its own right and is not needed here. And the conclusion really is only **weak** majorization: for $\mathbf{A}=\mathbf{I}_2$ and $\mathbf{B}=\operatorname{diag}(-1,1)$ we have $\sigma(\mathbf{A})=\sigma(\mathbf{B})=(1,1)$ and $\mathbf{A}+\mathbf{B}=\operatorname{diag}(0,2)$, so $\sigma(\mathbf{A}+\mathbf{B})=(2,0)$ has total 2 while $\sigma(\mathbf{A})+\sigma(\mathbf{B})=(2,2)$ has total 4. Section 6 records this as a negative result: no strengthening to $\prec$ is available.

The debts of the last two examples are now paid. By Lemma 21.4.2 with k fixed,

$$\|\mathbf{A}+\mathbf{B}\|_{(k)} \;\le\; \|\mathbf{A}\|_{(k)}+\|\mathbf{B}\|_{(k)},$$

so the Ky Fan k-norms of Definition 21.4.2 really are norms, and so is the trace norm, the case $k=p$. Chapter 17 §06's Ky Fan functionals have been repaired exactly as that section promised.

21.4.4 The correspondence

Theorem 21.4.3. Unitarily Invariant Norms Are Symmetric Gauges

Let $m,n\ge 1$ and $p=\min(m,n)$. The assignments

$$|||\cdot||| \;\longmapsto\; \Phi_{|||\cdot|||}(\mathbf{x})=\left|\left|\left|\operatorname{diag}_{m,n}(\mathbf{x})\right|\right|\right|, \qquad \Phi \;\longmapsto\; \|\mathbf{A}\|_{\Phi}=\Phi(\sigma(\mathbf{A}))$$

are mutually inverse bijections between the unitarily invariant norms on $M_{m\times n}(F)$ and the symmetric gauge functions on $\mathbb{R}^p$.

Idea. Lemma 21.4.1 is one direction and already contains the fact that the round trip $|||\cdot||| \mapsto \Phi \mapsto |||\cdot|||$ is the identity. For the other direction the only issue is the triangle inequality for $\mathbf{A}\mapsto\Phi(\sigma(\mathbf{A}))$, and it splits into two steps that have just been prepared: $\sigma(\mathbf{A}+\mathbf{B})\prec_w \sigma(\mathbf{A})+\sigma(\mathbf{B})$, then §03's theorem, then the triangle inequality for Φ itself.

Proof. **Step 1: $\|\cdot\|_{\Phi}$ is a unitarily invariant norm.** Let Φ be a symmetric gauge function on $\mathbb{R}^p$ and put $\|\mathbf{A}\|_{\Phi}=\Phi(\sigma(\mathbf{A}))$.

(N1). $\Phi(\sigma(\mathbf{A}))\ge 0$, and it vanishes only if $\sigma(\mathbf{A})=\mathbf{0}$, which by Theorem 13.9.1 forces $\mathbf{A}=\mathbf{U}\mathbf{O}\mathbf{V}^{*}=\mathbf{O}$.

(N2). For $c\in F$, $(c\mathbf{A})^{*}(c\mathbf{A})=|c|^2\mathbf{A}^{*}\mathbf{A}$, so each eigenvalue of $(c\mathbf{A})^{*}(c\mathbf{A})$ is $|c|^2$ times the matching eigenvalue of $\mathbf{A}^{*}\mathbf{A}$ and Definition 13.9.1 gives $\sigma(c\mathbf{A})=|c|\sigma(\mathbf{A})$. Hence $\|c\mathbf{A}\|_{\Phi}=$

$\Phi(|c|\sigma(\mathbf{A})) = |c|\,\|\mathbf{A}\|_\Phi$.

(N3). The vectors $\sigma(\mathbf{A}+\mathbf{B})$ and $\sigma(\mathbf{A})+\sigma(\mathbf{B})$ have nonnegative entries, and Lemma 21.4.2 gives $\sigma(\mathbf{A}+\mathbf{B}) \prec_w \sigma(\mathbf{A})+\sigma(\mathbf{B})$. So Theorem 21.3.4 (b) applies, and then the triangle inequality for the norm Φ:

$$\begin{aligned}\|\mathbf{A}+\mathbf{B}\|_\Phi = \Phi(\sigma(\mathbf{A}+\mathbf{B})) &\le \Phi(\sigma(\mathbf{A})+\sigma(\mathbf{B})) \\ &\le \Phi(\sigma(\mathbf{A})) + \Phi(\sigma(\mathbf{B})) = \|\mathbf{A}\|_\Phi + \|\mathbf{B}\|_\Phi\,.\end{aligned}$$

Invariance. By Eq. 21.4.1, $\sigma(\mathbf{UAV}) = \sigma(\mathbf{A})$ for unitary $\mathbf{U},\mathbf{V}$, so $\|\mathbf{UAV}\|_\Phi = \|\mathbf{A}\|_\Phi$.

Step 2: the round trips. Start from a unitarily invariant norm $\|\cdot\|$. By Lemma 21.4.1 its $\Phi = \Phi_{\|\cdot\|}$ is a symmetric gauge and $\|\mathbf{A}\| = \Phi(\sigma(\mathbf{A})) = \|\mathbf{A}\|_\Phi$ for every $\mathbf{A}$, so the second assignment undoes the first.

Start instead from a symmetric gauge Φ on $\mathbb{R}^p$, and let Ψ be the gauge of the norm $\|\cdot\|_\Phi$. For $\mathbf{x}\in\mathbb{R}^p$, the matrix $\operatorname{diag}_{m,n}(\mathbf{x})$ has $\operatorname{diag}_{m,n}(\mathbf{x})^*\operatorname{diag}_{m,n}(\mathbf{x}) = \operatorname{diag}(x_1^2,\dots,x_p^2,0,\dots,0)$, whose eigenvalue list in decreasing order is $(|\mathbf{x}|^{\downarrow})^2$ padded with zeros; so Definition 13.9.1 gives $\sigma(\operatorname{diag}_{m,n}(\mathbf{x})) = |\mathbf{x}|^{\downarrow}$. Hence

$$\Psi(\mathbf{x}) = \left\|\operatorname{diag}_{m,n}(\mathbf{x})\right\|_\Phi = \Phi(|\mathbf{x}|^{\downarrow}) = \Phi(\mathbf{x}),$$

the last equality by Eq. 21.3.1. So the first assignment undoes the second, and the two maps are mutually inverse bijections. This proves the theorem.

■

The theorem is a machine for producing matrix norms: choose any symmetric gauge and it hands back a unitarily invariant norm, with the triangle inequality free of charge. Feeding it the gauges of Example (The p-norms, the Ky Fan gauges, and size one) (a) gives the most-used family.

Example. The Schatten norms

The letter p is reserved in this section for $\min(m,n)$, so write the exponent as r. For $1\le r\le\infty$, the **Schatten r-norm** is

$$\|\mathbf{A}\|_{S_r} = \|\sigma(\mathbf{A})\|_r = \Big(\sum_{i=1}^{p}\sigma_i(\mathbf{A})^r\Big)^{1/r},$$

read as $\|\mathbf{A}\|_{S_\infty} = \sigma_1(\mathbf{A})$ at $r=\infty$. Identify S_1, S_2, and S_∞, and explain why each $\|\cdot\|_{S_r}$ is a unitarily invariant norm.

Solution. By Example (The p-norms, the Ky Fan gauges, and size one) (a), $\|\cdot\|_r$ is a symmetric gauge function on $\mathbb{R}^p$ for every $1\le r\le\infty$. So Theorem 21.4.3 makes $\mathbf{A}\mapsto\|\sigma(\mathbf{A})\|_r$ a unitarily invariant norm, with no further checking. At the three familiar exponents: $\|\cdot\|_{S_1}$ is the trace norm, since $\|\sigma(\mathbf{A})\|_1 = \sum_i\sigma_i(\mathbf{A})$; $\|\cdot\|_{S_2}$ is the Frobenius norm, since $\|\mathbf{A}\|_F^2 = \|\mathbf{\Sigma}\|_F^2 = \sum_i\sigma_i(\mathbf{A})^2$ by Lemma 11.7.5 applied to $\mathbf{A}=\mathbf{U\Sigma V}^*$; and $\|\cdot\|_{S_\infty}$ is the spectral norm, by Theorem 16.3.3 (c). The three norms Chapter 16 measured matrices with are three points of one family. A concrete reading: the symmetric $\mathbf{A} = \left(\begin{smallmatrix}2&1\\1&2\end{smallmatrix}\right)$ has eigenvalues 3 and 1, both positive, so $\mathbf{A}^*\mathbf{A} = \mathbf{A}^2$ has eigenvalues 9 and 1 and $\sigma(\mathbf{A}) = (3,1)$ by Defini-

tion 13.9.1. Hence $|||\mathbf{A}|||_{S_1} = 4$, $|||\mathbf{A}|||_{S_2} = \sqrt{10}$ — which matches $\|\mathbf{A}\|_F = \sqrt{4+1+1+4}$ — and $|||\mathbf{A}|||_{S_\infty} = 3$.

21.4.5 What every unitarily invariant norm satisfies

Chapter 20 §09 collected the facts it needed about moving factors in and out of the spectral and Frobenius norms (Lemma 20.9.1). Part (b) of that lemma — $\|\mathbf{XMY}\| \le \|\mathbf{X}\|_2\|\mathbf{M}\|\|\mathbf{Y}\|_2$ — was proved separately for the two norms. It holds for all of them at once.

Proposition 21.4.4. Basic Properties

Let $|||\cdot|||$ be a unitarily invariant norm on $M_{m\times n}(F)$, with gauge Φ as in Lemma 21.4.1, and let $\mathbf{A} \in M_{m\times n}(F)$.

(a) $\sigma(\mathbf{A}^*) = \sigma(\mathbf{A})$; consequently, if $|||\cdot|||'$ is the unitarily invariant norm on $M_{n\times m}(F)$ attached to the same gauge Φ by Theorem 21.4.3, then $|||\mathbf{A}^*|||' = |||\mathbf{A}|||$. For $m = n$ this reads $|||\mathbf{A}^*||| = |||\mathbf{A}|||$.

(b) $|||\mathbf{XA}||| \le \|\mathbf{X}\|_2\, |||\mathbf{A}|||$ for every $\mathbf{X} \in M_m(F)$, and $|||\mathbf{AY}||| \le |||\mathbf{A}|||\, \|\mathbf{Y}\|_2$ for every $\mathbf{Y} \in M_n(F)$.

(c) If $|||\cdot|||$ is **normalized**, meaning $|||\mathbf{e}_1\mathbf{e}_1^*||| = 1$ for the $m \times n$ matrix $\mathbf{e}_1\mathbf{e}_1^*$ with a single 1 in position $(1,1)$, then

$$\|\mathbf{A}\|_2 \;\le\; |||\mathbf{A}||| \;\le\; |||\mathbf{A}|||_{(p)}\,.$$

(d) A normalized unitarily invariant norm on $M_n(F)$ is submultiplicative: $|||\mathbf{AB}||| \le |||\mathbf{A}|||\, |||\mathbf{B}|||$.

Idea. Everything reduces to one inequality between singular values, $\sigma_i(\mathbf{XA}) \le \|\mathbf{X}\|_2\sigma_i(\mathbf{A})$, and then to §03's monotonicity Proposition 21.3.1. That inequality is one line of the min–max theorem: the subspace on which $\mathbf{A}$ stretches by at most $\sigma_i(\mathbf{A})$ also works for $\mathbf{XA}$, with the factor $\|\mathbf{X}\|_2$ paid once. Part (c) is Corollary 21.3.2, read through the correspondence, and (d) is (b) and (c) combined.

Proof.

(a) By Theorem 13.9.1 write $\mathbf{A} = \mathbf{U\Sigma V}^*$; then $\mathbf{A}^* = \mathbf{V\Sigma}^*\mathbf{U}^*$, and $\mathbf{\Sigma}^* = \mathrm{diag}_{n,m}(\sigma(\mathbf{A}))$ has the same diagonal. By Theorem 13.9.2 applied to $\mathbf{A}^*$, $\sigma(\mathbf{A}^*) = \sigma(\mathbf{A})$. The consequence is Lemma 21.4.1 used twice: $|||\mathbf{A}^*|||' = \Phi(\sigma(\mathbf{A}^*)) = \Phi(\sigma(\mathbf{A})) = |||\mathbf{A}|||$.

(b) Everything rests on one inequality between singular values.
Claim. $\sigma_k(\mathbf{XA}) \le \|\mathbf{X}\|_2\, \sigma_k(\mathbf{A})$ for every $1 \le k \le n$, with the convention $\sigma_k = 0$ for $k > p$.
Proof. Fix a singular value decomposition of $\mathbf{A}$ with right singular vectors $\mathbf{v}_1, \dots, \mathbf{v}_n$, and let $T_k = \mathrm{span}(\mathbf{v}_k, \dots, \mathbf{v}_n)$, a subspace of F^n of dimension $n - k + 1$. By Lemma 17.9.1 (a), $\|\mathbf{Ax}\| \le \sigma_k(\mathbf{A})\|\mathbf{x}\|$ for every $\mathbf{x} \in T_k$. By Theorem 16.3.2 (a), $\|\mathbf{X}(\mathbf{Ax})\| \le \|\mathbf{X}\|_2\|\mathbf{Ax}\|$, so

$$\|(\mathbf{XA})\mathbf{x}\| \le \|\mathbf{X}\|_2\, \sigma_k(\mathbf{A})\, \|\mathbf{x}\| \qquad (\mathbf{x} \in T_k).$$

The matrices $\mathbf{XA}$ and $\mathbf{A}$ both have n columns, so Theorem 17.9.2 applies to $\mathbf{XA}$ with subspaces of F^n: $\sigma_k(\mathbf{XA})$ is the **minimum** over subspaces of dimension $n-k+1$ of the largest stretch on them, and T_k is one competitor. Hence $\sigma_k(\mathbf{XA}) \le \|\mathbf{X}\|_2 \sigma_k(\mathbf{A})$. ■

Now $\mathbf{XA}$ and $\mathbf{A}$ lie in the same space $M_{m\times n}(F)$, and the claim says $|\sigma_i(\mathbf{XA})| \le |\|\mathbf{X}\|_2 \sigma_i(\mathbf{A})|$ for $1 \le i \le p$. By Proposition 21.3.1 and then (N2) for Φ,

$$\|\mathbf{XA}\| = \Phi(\sigma(\mathbf{XA})) \le \Phi(\|\mathbf{X}\|_2 \sigma(\mathbf{A})) = \|\mathbf{X}\|_2 \, \Phi(\sigma(\mathbf{A})) = \|\mathbf{X}\|_2 \, \|\mathbf{A}\| .$$

For the second inequality, apply the first to $\mathbf{A}^*$ and $\mathbf{Y}^*$ inside $M_{n\times m}(F)$, with the norm $\|\cdot\|'$ of part (a): $\|(\mathbf{AY})^*\|' = \|\mathbf{Y}^*\mathbf{A}^*\|' \le \|\mathbf{Y}^*\|_2 \, \|\mathbf{A}^*\|'$. By (a) the left side is $\|\mathbf{AY}\|$ and $\|\mathbf{A}^*\|' = \|\mathbf{A}\|$, while $\|\mathbf{Y}^*\|_2 = \|\mathbf{Y}\|_2$ by Lemma 20.9.1 (a).

(c) The matrix $\mathbf{e}_1\mathbf{e}_1^*$ is $\operatorname{diag}_{m,n}(\mathbf{e}_1)$, so the normalization says $\Phi(\mathbf{e}_1) = 1$. By Corollary 21.3.2, $\Phi(\mathbf{e}_1)\|\mathbf{x}\|_\infty \le \Phi(\mathbf{x}) \le \Phi(\mathbf{e}_1)\|\mathbf{x}\|_1$ for every $\mathbf{x} \in \mathbb{R}^p$. At $\mathbf{x} = \sigma(\mathbf{A})$ this reads

$$\sigma_1(\mathbf{A}) \;\le\; \Phi(\sigma(\mathbf{A})) \;\le\; \sum_{i=1}^{p} \sigma_i(\mathbf{A}),$$

which is the claim, since $\sigma_1(\mathbf{A}) = \|\mathbf{A}\|_2$ by Theorem 16.3.3 (c) and the right-hand side is $\|\mathbf{A}\|_{(p)}$ by Definition 21.4.2.

(d) By (b) and then (c), $\|\mathbf{AB}\| \le \|\mathbf{A}\|_2 \, \|\mathbf{B}\| \le \|\mathbf{A}\| \, \|\mathbf{B}\|$. This proves the proposition.

■

Applying (b) twice, once on each side, gives $\|\mathbf{XAY}\| \le \|\mathbf{X}\|_2 \, \|\mathbf{A}\| \, \|\mathbf{Y}\|_2$. That is Lemma 20.9.1 (b) with the two named norms replaced by any unitarily invariant one, so the debt announced at the head of this subsection is paid. Taking $\mathbf{X} = \mathbf{U}$ and $\mathbf{Y} = \mathbf{V}$ unitary recovers Definition 21.4.1 itself: a unitary matrix has orthonormal columns, so $\|\mathbf{U}\|_2 = \|\mathbf{V}\|_2 = 1$ by Lemma 20.9.1 (c), and the bound reads $\|\mathbf{UAV}\| \le \|\mathbf{A}\|$; applying it again to $\mathbf{U}^*(\mathbf{UAV})\mathbf{V}^* = \mathbf{A}$ gives $\|\mathbf{A}\| \le \|\mathbf{UAV}\|$, and the two together give equality.

Warning. **A unitarily invariant norm need not be submultiplicative, and the normalization in (d) is what supplies the lower bound that rules it out.** Half the spectral norm, $\|\mathbf{A}\| = \frac{1}{2}\|\mathbf{A}\|_2$, is a unitarily invariant norm — it is $\Phi(\sigma(\mathbf{A}))$ for the symmetric gauge $\Phi = \frac{1}{2}\|\cdot\|_\infty$. At $\mathbf{A} = \mathbf{B} = \mathbf{I}_n$ it gives $\|\mathbf{AB}\| = \frac{1}{2}$ while $\|\mathbf{A}\| \, \|\mathbf{B}\| = \frac{1}{4}$, so submultiplicativity fails. Doubling instead of halving is harmless: $2\|\cdot\|_2$ is submultiplicative. What Proposition 21.4.4 (d) really needs is $\|\mathbf{A}\|_2 \le \|\mathbf{A}\|$, and Corollary 21.3.2 delivers that as soon as $\|\mathbf{e}_1\mathbf{e}_1^*\| \ge 1$, normalized or not; here $\|\mathbf{e}_1\mathbf{e}_1^*\| = \frac{1}{2}$, and the bound runs the wrong way. Submultiplicativity is a property of the **scaling** of a norm, not of its shape.

Theorem 21.4.3 is the section's payment of Chapter 16 §08's debt: the norms invariant under unitary multiplication are exactly the symmetric gauges of the singular values, and there are as many of them as there are norms on $\mathbb{R}^p$ that ignore order and sign. What the correspondence buys is that any inequality between two singular value lists in the weak majorization order becomes an inequality between the matrices in **every** one of these norms at once. Section 6 makes that the definition of dominance and uses it to settle low-rank approximation and the nearest unitary matrix; Section 7 does the same for Lidskii's and Mirsky's theorems.

21.4.6 Exercises

21.4.6.1 A. Check your understanding

Exercise. A1

(a) State Definition 21.4.1, saying exactly which matrices $\mathbf{U}$ and $\mathbf{V}$ range over.

(b) Explain in one sentence why a unitarily invariant norm can be computed from $\sigma(\mathbf{A})$ alone.

(c) True or false, with a reason: $\mathbf{A} \mapsto \max_{i,j} |a_{ij}|$ is a unitarily invariant norm on $M_2(\mathbb{R})$.

(d) Which axiom of Definition 16.1.1 is the only hard one for $\mathbf{A} \mapsto \Phi(\sigma(\mathbf{A}))$, and which two earlier results prove it?

Solution.

(a) A norm $\|\cdot\|$ on $M_{m\times n}(F)$ with $\|\mathbf{UAV}\| = \|\mathbf{A}\|$ for every $\mathbf{A}$ and all unitary $\mathbf{U} \in M_m(F)$ and $\mathbf{V} \in M_n(F)$, the two chosen independently.

(b) Because $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$ with $\mathbf{\Sigma} = \operatorname{diag}_{m,n}(\sigma(\mathbf{A}))$, so invariance gives $\|\mathbf{A}\| = \|\mathbf{\Sigma}\|$, a quantity built from $\sigma(\mathbf{A})$ alone (Lemma 21.4.1).

(c) False. It is a norm but not unitarily invariant: $\mathbf{I}_2$ and the rotation $\mathbf{U} = \frac{1}{\sqrt{2}}\left(\begin{smallmatrix} 1 & -1 \\ 1 & 1 \end{smallmatrix}\right)$ have entrywise maxima 1 and $1/\sqrt{2}$, while $\mathbf{U} = \mathbf{U}\mathbf{I}_2\mathbf{I}_2$.

(d) The triangle inequality (N3). It follows from Lemma 21.4.2, which gives $\sigma(\mathbf{A}+\mathbf{B}) \prec_w \sigma(\mathbf{A}) + \sigma(\mathbf{B})$, together with Theorem 21.3.4 (b), which turns a weak majorization of nonnegative vectors into an inequality of gauges.

21.4.6.2 B. Practice

Exercise. B1: Which of these are unitarily invariant norms

Determine which of the following are unitarily invariant norms on $M_2(\mathbb{R})$. Justify your answer, naming the clause that fails when one does.

(a) $\|\mathbf{A}\| = \sigma_1(\mathbf{A}) + \sigma_2(\mathbf{A})$.

(b) $\|\mathbf{A}\| = |a_{11}| + |a_{12}| + |a_{21}| + |a_{22}|$.

(c) $\|\mathbf{A}\| = \left(\sigma_1(\mathbf{A})^3 + \sigma_2(\mathbf{A})^3\right)^{1/3}$.

(d) $\|\mathbf{A}\| = |\operatorname{tr} \mathbf{A}|$.

(e) $\|\mathbf{A}\| = \max\left(\sigma_1(\mathbf{A}),\ 2\sigma_2(\mathbf{A})\right)$.

Solution.

(a) Yes: this is the trace norm, the Ky Fan 2-norm of Definition 21.4.2, equivalently $\|\cdot\|_{S_1}$ of Example (The Schatten norms). Alternatively, it is $\Phi(\sigma(\mathbf{A}))$ for $\Phi = \|\cdot\|_1$, so Theorem 21.4.3 applies.

(b) No. It is a norm — the 1-norm of the entry list — but it is not unitarily invariant: it gives $\mathbf{I}_2$ the value 2 and the rotation $\mathbf{U}$ above the value $4/\sqrt{2} = 2\sqrt{2}$, although $\sigma(\mathbf{I}_2) = \sigma(\mathbf{U}) = (1, 1)$.

(c) Yes: it is $\Phi(\sigma(\mathbf{A}))$ for $\Phi = \|\cdot\|_3$, a symmetric gauge by Example (The p-norms, the Ky Fan gauges, and size one) (a), so Theorem 21.4.3 applies. This is the Schatten 3-norm.

(d) No. It is unitarily *similarity* invariant but not a norm: $\operatorname{tr}\left(\begin{smallmatrix} 1 & 0 \\ 0 & -1 \end{smallmatrix}\right) = 0$ with the matrix nonzero, so (N1) fails. It is not unitarily invariant either, since $\operatorname{tr}(\mathbf{UAV})$ generally differs from $\operatorname{tr}\mathbf{A}$.

(e) No. By Theorem 21.4.3 the question is whether $\Phi(\mathbf{x}) = \max(|x|_1^{\downarrow},\ 2|x|_2^{\downarrow})$ is a symmetric gauge on $\mathbb{R}^2$, and it is not a norm: subadditivity fails. Take $\mathbf{x} = (1, \frac{1}{2})$ and $\mathbf{y} = (\frac{1}{2}, 1)$. Then $\Phi(\mathbf{x}) = \max(1, 1) = 1$ and $\Phi(\mathbf{y}) = 1$, while $\mathbf{x} + \mathbf{y} = (\frac{3}{2}, \frac{3}{2})$ gives $\Phi(\mathbf{x} + \mathbf{y}) = \max(\frac{3}{2}, 3) = 3 > 2$. The failure is easy to see in the matrices as well: $\operatorname{diag}(1, \frac{1}{2})$ and $\operatorname{diag}(\frac{1}{2}, 1)$ each get the value 1, and their sum $\frac{3}{2}\mathbf{I}_2$ gets 3. The trap is that $\mathbf{x} \mapsto 2|x|_2^{\downarrow}$ is not a norm — it vanishes at $\mathbf{e}_1$ — so the maximum of it with Φ_1 inherits nothing. Contrast exercise C1 below, where both members of the maximum are norms.

Exercise. B2: Measuring one matrix in every way

Let

$$\mathbf{A} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}, \qquad \mathbf{B} = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}.$$

(a) Compute $\sigma(\mathbf{A})$ and $\sigma(\mathbf{B})$, and then $\|\mathbf{A}\|_2$, $\|\mathbf{A}\|_F$, $|||\mathbf{A}|||_{(1)}$, and $|||\mathbf{A}|||_{(2)}$.

(b) Compute $\sigma(\mathbf{A}+\mathbf{B})$ and verify $\sigma(\mathbf{A}+\mathbf{B}) \prec_w \sigma(\mathbf{A}) + \sigma(\mathbf{B})$ directly, saying for which k the inequality is strict.

(c) Verify Proposition 21.4.4 (c) for $\mathbf{A}$ and the Frobenius norm, and check that the Frobenius norm is normalized.

Solution.

(a) $\mathbf{A}$ is real symmetric with eigenvectors $(1, 1)$ and $(1, -1)$ for the eigenvalues 3 and 1. Hence $\mathbf{A}^*\mathbf{A} = \mathbf{A}^2$ has eigenvalues 9 and 1, and Definition 13.9.1 gives $\sigma(\mathbf{A}) = (3, 1)$, as in Example (The Schatten norms). Likewise $\mathbf{B}$ has eigenvectors $(1, -1)$ and $(1, 1)$ for 3 and 1, so $\sigma(\mathbf{B}) = (3, 1)$. Hence $\|\mathbf{A}\|_2 = 3$, $\|\mathbf{A}\|_F = \sqrt{9+1} = \sqrt{10}$ (which agrees with $\sqrt{4+1+1+4}$), $|||\mathbf{A}|||_{(1)} = 3$ and $|||\mathbf{A}|||_{(2)} = 4$.

(b) $\mathbf{A}+\mathbf{B} = \operatorname{diag}(4, 4)$, so $\sigma(\mathbf{A}+\mathbf{B}) = (4, 4)$, while $\sigma(\mathbf{A})+\sigma(\mathbf{B}) = (6, 2)$. The partial sums are $4, 8$ against $6, 8$: strict at $k = 1$, equality at $k = 2$. So $\sigma(\mathbf{A}+\mathbf{B}) \prec_w \sigma(\mathbf{A}) + \sigma(\mathbf{B})$, and here the totals happen to agree, so the relation is even a majorization.

(c) The Frobenius norm has $\|\mathbf{e}_1\mathbf{e}_1^{\top}\|_F = 1$, so it is normalized. The chain reads $3 = \|\mathbf{A}\|_2 \le \sqrt{10} \le 4 = |||\mathbf{A}|||_{(2)}$, and $9 \le 10 \le 16$ confirms it.

Exercise. B3: Moving a factor across

Let $\mathbf{A} \in M_{m\times n}(F)$, let $|||\cdot|||$ be a unitarily invariant norm on $M_{m\times n}(F)$, and let P be an orthogonal projection of F^m onto a subspace, with matrix $\mathbf{P}$.

(a) Prove that $\|\mathbf{P}\|_2 \le 1$.

(b) Deduce that $|||\mathbf{PA}||| \le |||\mathbf{A}|||$.

(c) Give an example with $m = n = 2$ in which the inequality in (b) is strict for the Frobenius norm.

Solution.

(a) For $\mathbf{x} \in F^m$, write $\mathbf{x} = \mathbf{P}\mathbf{x} + (\mathbf{x} - \mathbf{P}\mathbf{x})$, an orthogonal decomposition, so $\|\mathbf{x}\|^2 = \|\mathbf{P}\mathbf{x}\|^2 + \|\mathbf{x} - \mathbf{P}\mathbf{x}\|^2 \geq \|\mathbf{P}\mathbf{x}\|^2$ by Theorem 11.1.3. Hence $\|\mathbf{P}\mathbf{x}\| \leq \|\mathbf{x}\|$ for every $\mathbf{x}$, and $\|\mathbf{P}\|_2 \leq 1$ by Definition 16.3.2.

(b) By Proposition 21.4.4 (b) with $\mathbf{X} = \mathbf{P}$, $|||\mathbf{P}\mathbf{A}||| \leq \|\mathbf{P}\|_2 \, |||\mathbf{A}||| \leq |||\mathbf{A}|||$.

(c) Take $\mathbf{P} = \mathbf{e}_1\mathbf{e}_1^\top$ and $\mathbf{A} = \mathbf{I}_2$. Then $\mathbf{P}\mathbf{A} = \mathbf{e}_1\mathbf{e}_1^\top$ has $\|\mathbf{P}\mathbf{A}\|_F = 1$, while $\|\mathbf{I}_2\|_F = \sqrt{2}$.

21.4.6.3 C. Going deeper

Exercise. C1: A gauge built from two Ky Fan gauges

On $\mathbb{R}^p$ let $\Phi(\mathbf{x}) = \max\big(\Phi_1(\mathbf{x}),\ \frac{2}{k}\Phi_k(\mathbf{x})\big)$ for a fixed $2 \leq k \leq p$, with Φ_1 and Φ_k the Ky Fan gauges of Definition 21.3.2.

(a) Prove that Φ is a symmetric gauge function.

(b) Write down the unitarily invariant norm it induces on $M_{m\times n}(F)$, and decide whether it is normalized.

(c) For $p = k = 2$, show that $\Phi(\mathbf{x}) = \max(|x|_1^{\downarrow},\ |x_1| + |x_2|) = |x_1| + |x_2|$, and conclude that the norm of (b) is the trace norm in that case.

Solution.

(a) A maximum of finitely many norms is a norm: it is nonnegative, absolutely homogeneous, and subadditive because each member is, and it vanishes only where all members do, hence only at $\mathbf{0}$. Here both members are symmetric gauges by Example (The p-norms, the Ky Fan gauges, and size one) (b), and a maximum of permutation invariant and sign invariant functions has the same invariances. So Φ satisfies Definition 21.3.1.

(b) By Theorem 21.4.3 the induced norm is

$$|||\mathbf{A}||| = \max\Big(\sigma_1(\mathbf{A}),\ \tfrac{2}{k}\sum_{i\leq k}\sigma_i(\mathbf{A})\Big).$$

It is normalized: at $\mathbf{A} = \mathbf{e}_1\mathbf{e}_1^*$ the singular values are $1, 0, \ldots, 0$, so the two candidates are 1 and $2/k \leq 1$, and the maximum is 1.

(c) With $p = k = 2$, $\frac{2}{2}\Phi_2(\mathbf{x}) = |x_1| + |x_2|$, which is at least $\Phi_1(\mathbf{x}) = |x|_1^{\downarrow}$. So $\Phi = \|\cdot\|_1$, and the induced norm is $\sigma_1(\mathbf{A}) + \sigma_2(\mathbf{A})$, the trace norm.

Exercise. C2: How much scaling submultiplicativity can stand

Let $c > 0$ and consider $N_c(\mathbf{A}) = c\,\|\mathbf{A}\|_2$ on $M_n(F)$, $n \geq 1$.

(a) Prove that N_c is a unitarily invariant norm for every $c > 0$, and compute its gauge.

(b) Determine exactly which c make N_c submultiplicative. *Hint: test at* $\mathbf{A} = \mathbf{B} = \mathbf{I}_n$.

(c) Do the same for $M_c(\mathbf{A}) = c\,|||\mathbf{A}|||_{(n)}$, the scaled trace norm. Show that $\mathbf{A} = \mathbf{B} = \mathbf{I}_n$ is now the **wrong** test matrix, and find one that gives the sharp answer.

Solution.

(a) A positive multiple of a norm is a norm, and multiplying by c does not disturb $N_c(\mathbf{UAV}) = N_c(\mathbf{A})$. Its gauge is $\mathbf{x} \mapsto N_c(\mathrm{diag}(\mathbf{x})) = c\|\mathbf{x}\|_\infty$, since the largest singular value of $\mathrm{diag}(\mathbf{x})$ is $\|\mathbf{x}\|_\infty$.

(b) Exactly $c \geq 1$. If $c \geq 1$, then $N_c(\mathbf{AB}) = c\|\mathbf{AB}\|_2 \leq c\|\mathbf{A}\|_2\|\mathbf{B}\|_2 \leq c^2\|\mathbf{A}\|_2\|\mathbf{B}\|_2 = N_c(\mathbf{A})N_c(\mathbf{B})$, using submultiplicativity of the spectral norm (Theorem 16.3.2 (d)) and $c \leq c^2$. If $c < 1$, take $\mathbf{A} = \mathbf{B} = \mathbf{I}_n$: then $N_c(\mathbf{AB}) = c$ while $N_c(\mathbf{A})N_c(\mathbf{B}) = c^2 < c$.

(c) The answer is again exactly $c \geq 1$. Sufficiency: by Proposition 21.4.4 (b) and (c), $|||\mathbf{AB}|||_{(n)} \leq \|\mathbf{A}\|_2\,|||\mathbf{B}|||_{(n)} \leq |||\mathbf{A}|||_{(n)}\,|||\mathbf{B}|||_{(n)}$, so for $c \geq 1$,

$$M_c(\mathbf{AB}) = c\,|||\mathbf{AB}|||_{(n)} \leq c\,|||\mathbf{A}|||_{(n)}\,|||\mathbf{B}|||_{(n)} \leq c^2\,|||\mathbf{A}|||_{(n)}\,|||\mathbf{B}|||_{(n)} = M_c(\mathbf{A})M_c(\mathbf{B}).$$

Necessity: at $\mathbf{A} = \mathbf{B} = \mathbf{I}_n$ one gets $M_c(\mathbf{I}_n) = cn$ and the requirement $cn \leq c^2n^2$, that is $c \geq 1/n$ — a weaker condition, so for $n \geq 2$ the identity is the wrong test matrix. Take instead $\mathbf{A} = \mathbf{B} = \mathbf{e}_1\mathbf{e}_1^\top$, which is idempotent with singular values $1, 0, \ldots, 0$. Then $M_c(\mathbf{AB}) = c$ while $M_c(\mathbf{A})M_c(\mathbf{B}) = c^2$, and $c \leq c^2$ forces $c \geq 1$. The moral: submultiplicativity is tested by matrices of the **smallest** rank, where the trace norm and the spectral norm agree, not by the identity, where they differ by a factor of n.

Exercise. C3: The correspondence in the smallest case

(a) Describe all unitarily invariant norms on $M_{1\times n}(F)$, the row vectors. *Hint: what is p?*

(b) Deduce that on $M_{1\times n}(F)$ every unitarily invariant norm is a positive multiple of the Euclidean norm of the row.

(c) Explain why the answer to (a) does not contradict the existence of many different unitarily invariant norms on $M_2(F)$.

Solution.

(a) Here $p = \min(1, n) = 1$, so Theorem 21.4.3 matches unitarily invariant norms on $M_{1\times n}(F)$ with symmetric gauge functions on $\mathbb{R}^1$. By Example (The p-norms, the Ky Fan gauges, and size one) (c), every norm on $\mathbb{R}^1$ is such a gauge, and each is $\Phi(t) = c|t|$ with $c > 0$. So the unitarily invariant norms are exactly $\mathbf{A} \mapsto c\,\sigma_1(\mathbf{A})$, $c > 0$.

(b) A row vector $\mathbf{a}^\top$ has $\mathbf{A}^*\mathbf{A} = \overline{\mathbf{a}}\mathbf{a}^\top$ of rank at most 1, with the single nonzero eigenvalue $\mathbf{a}^*\mathbf{a} = \|\mathbf{a}\|_2^2$; so $\sigma_1(\mathbf{A}) = \|\mathbf{a}\|_2$. Hence every unitarily invariant norm on $M_{1\times n}(F)$ is $c\|\mathbf{a}\|_2$.

(c) Because the correspondence is with norms on $\mathbb{R}^p$, and $p = 1$ leaves only one shape. For $M_2(F)$, $p = 2$, and there are many norms on $\mathbb{R}^2$ that are permutation and sign invariant — the whole family $\|\cdot\|_q$ for $1 \leq q \leq \infty$, for instance, giving the Schatten norms of Example (The Schatten norms). The rigidity in (a) is a statement about $\mathbb{R}^1$, not about matrices.

21.5 The Trace Inequality

Sections 3 and 4 built a dictionary: a unitarily invariant norm is a symmetric gauge applied to the list of singular values, and a symmetric gauge respects weak majorization. To use that dictionary we need a supply of inequalities between singular values, and the cleanest source of them is a single scalar inequality about how two matrices pair against each other. This section proves it. The pairing is the Frobenius one, $\operatorname{tr}(\mathbf{A}^*\mathbf{B})$, and the theorem says that it is largest when the two matrices are *aligned*: when their singular value decompositions can be written with the same unitary factors.

21.5.1 How large can a pairing be?

Chapter 11 §01 made $M_{m\times n}(F)$ an inner product space with $\langle \mathbf{A}, \mathbf{B}\rangle = \operatorname{tr}(\mathbf{B}^*\mathbf{A})$, whose norm is the Frobenius norm $\|\mathbf{A}\|_F$. So $\operatorname{tr}(\mathbf{A}^*\mathbf{B}) = \langle \mathbf{B}, \mathbf{A}\rangle$ is an inner product of two matrices, and Cauchy–Schwarz (Theorem 11.1.5) already bounds it:

$$|\operatorname{tr}(\mathbf{A}^*\mathbf{B})| \;\le\; \|\mathbf{A}\|_F\,\|\mathbf{B}\|_F = \Big(\sum_i \sigma_i(\mathbf{A})^2\Big)^{1/2}\Big(\sum_i \sigma_i(\mathbf{B})^2\Big)^{1/2},$$

using $\|\mathbf{A}\|_F^2 = \sum_i \sigma_i(\mathbf{A})^2$. That bound depends only on the two lists of singular values, which is the right kind of answer; but it is not the best one.

Here is the smallest example. Let $\mathbf{A} = \operatorname{diag}(2,1)$ and $\mathbf{B} = \mathbf{I}_2$. Then $\operatorname{tr}(\mathbf{A}^*\mathbf{B}) = 3$, while Cauchy–Schwarz allows $\sqrt{5}\cdot\sqrt{2} \approx 3.162$. The number 3 is $\sigma_1(\mathbf{A})\sigma_1(\mathbf{B}) + \sigma_2(\mathbf{A})\sigma_2(\mathbf{B})$, the two lists of singular values multiplied **term by term in decreasing order**. That is the true bound, and the pairing $\sum_i \sigma_i\tau_i$ is never larger than the Cauchy–Schwarz product, by Cauchy–Schwarz applied to the two lists.

Two matrices pair as well as their singular values allow, and no better; and the best case is when the two matrices stretch the same directions by the same ranking.

The proof takes the two singular value decompositions apart and finds, sitting between them, a square array of nonnegative weights whose rows and columns each add up to at most 1. Everything then reduces to a fact about such arrays, which is worth isolating first.

21.5.2 Weights that do not add up to more than one

Definition 21.5.1. Doubly Substochastic Matrix

A matrix $\mathbf{D} = (d_{kl}) \in M_p(\mathbb{R})$ is **doubly substochastic** if

(S1) $d_{kl} \ge 0$ for all k, l;
(S2) $\sum_{l=1}^{p} d_{kl} \le 1$ for **every** row index k; and
(S3) $\sum_{k=1}^{p} d_{kl} \le 1$ for **every** column index l.

In words: the entries are nonnegative, and no row and no column adds up to more than one. Replacing both "≤ 1" by "$= 1$" gives the doubly stochastic matrices of Definition 18.8.4, so every doubly stochastic matrix is doubly substochastic, and in particular so is every permutation matrix. The zero matrix is doubly substochastic, which the stochastic condition excludes; that degenerate case matters, because it is what lets a substochastic matrix throw weight away instead

of merely moving it. A nonexample by minimal change: $\left(\begin{smallmatrix} 1 & 1 \\ 0 & 0 \end{smallmatrix}\right)$ satisfies (S1) and (S3) but its first row adds to 2, so (S2) fails.

The lemma we need says that such an array of weights cannot help two decreasing nonnegative lists pair better than they do term by term.

Lemma 21.5.1. Substochastic Pairing

Let $x_1 \geq \cdots \geq x_p \geq 0$ and $y_1 \geq \cdots \geq y_p \geq 0$ be real numbers, and let $\mathbf{D} \in M_p(\mathbb{R})$ be doubly substochastic. Then

$$\sum_{k=1}^{p} \sum_{l=1}^{p} d_{kl}\, x_k y_l \;\leq\; \sum_{k=1}^{p} x_k y_k.$$

Idea. Decreasing nonnegative lists are nonnegative combinations of the "staircase" lists $(1, \ldots, 1, 0, \ldots, 0)$: write $x_k = \sum_{i \geq k} (x_i - x_{i+1})$. Summation by parts in both variables at once turns each side into a nonnegative combination of the same coefficients, and all that is left to compare is what multiplies them: on the left a partial sum of $\mathbf{D}$ over a top-left rectangle, on the right the number $\min(i, j)$. The two substochastic conditions say exactly that an $i \times j$ rectangle of $\mathbf{D}$ holds at most $\min(i, j)$.

Proof. Put $x_{p+1} = y_{p+1} = 0$ and, for $1 \leq i \leq p$, write $\Delta x_i = x_i - x_{i+1}$ and $\Delta y_i = y_i - y_{i+1}$. These are ≥ 0, since the lists are decreasing and $x_p, y_p \geq 0$, and telescoping gives

$$x_k = \sum_{i=k}^{p} \Delta x_i, \qquad y_l = \sum_{j=l}^{p} \Delta y_j \qquad (1 \leq k, l \leq p).$$

Substituting both and exchanging the order of the four finite sums,

$$\sum_{k,l} d_{kl} x_k y_l = \sum_{i=1}^{p} \sum_{j=1}^{p} \Delta x_i \, \Delta y_j \, S_{ij}, \qquad S_{ij} \coloneqq \sum_{k \leq i} \sum_{l \leq j} d_{kl},$$

because the pair (k, l) contributes to (i, j) exactly when $k \leq i$ and $l \leq j$. The same substitution with $\mathbf{D}$ replaced by the identity matrix gives

$$\sum_{k=1}^{p} x_k y_k = \sum_{i=1}^{p} \sum_{j=1}^{p} \Delta x_i \, \Delta y_j \, \min(i, j),$$

since the number of $k \leq \min(i, j)$ is $\min(i, j)$.

It remains to compare S_{ij} with $\min(i, j)$. Summing (S2) over the i rows $k \leq i$ and discarding the nonnegative entries with $l > j$ gives $S_{ij} \leq i$; summing (S3) over the j columns $l \leq j$ gives $S_{ij} \leq j$. Hence $S_{ij} \leq \min(i, j)$ for all i, j. Since every Δx_i and every Δy_j is ≥ 0, replacing S_{ij} by $\min(i, j)$ can only increase the double sum. This proves the lemma.

■

Taking $\mathbf{D}$ to be the permutation matrix of $\pi \in S_p$ gives $\sum_k x_k y_{\pi(k)} \leq \sum_k x_k y_k$: two decreasing lists pair best when they are paired in order. That is the **rearrangement inequality**, and it is the special case that carries the meaning of the lemma. The general case allows the weights

to be spread out and partly thrown away, and says that neither helps.

Quick check. Where does the proof use $x_p \geq 0$ and $y_p \geq 0$, rather than just the decrease? Find a doubly substochastic $\mathbf{D}$ and a decreasing pair with a negative entry for which the conclusion fails.

Solution. It is used to make $\Delta x_p = x_p$ and $\Delta y_p = y_p$ nonnegative, which is what allows S_{ij} to be replaced by the larger $\min(i, j)$ in the last step. Without it the conclusion fails: take $p = 2$, $\mathbf{x} = (1, -1)$, $\mathbf{y} = (1, 1)$, and $\mathbf{D} = \operatorname{diag}(1, 0)$, which is doubly substochastic. The left side is 1 and the right side is $1 - 1 = 0$, so the asserted inequality $1 \leq 0$ is false. The proof breaks at the step named: here $\Delta x_2 = x_2 = -1 < 0$ and $\Delta y_2 = 1$, so replacing $S_{22} = 1$ by $\min(2, 2) = 2$ lowers the double sum instead of raising it. Throwing weight away helps when a pairing is negative, and that is exactly what nonnegativity rules out.

21.5.3 The trace inequality

Theorem 21.5.2. Von Neumann's Trace Inequality

Let $\mathbf{A}, \mathbf{B} \in M_{m\times n}(F)$ with $F = \mathbb{R}$ or $\mathbb{C}$, and let $p = \min(m, n)$. Then

$$\left|\operatorname{tr}(\mathbf{A}^*\mathbf{B})\right| \;\leq\; \sum_{i=1}^{p} \sigma_i(\mathbf{A})\,\sigma_i(\mathbf{B}).$$

The bound is attained whenever $\mathbf{A}$ and $\mathbf{B}$ have **aligned** singular value decompositions, that is, whenever there are unitary $\mathbf{U} \in M_m(F)$ and $\mathbf{V} \in M_n(F)$ with $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}_\mathbf{A}\mathbf{V}^*$ and $\mathbf{B} = \mathbf{U}\mathbf{\Sigma}_\mathbf{B}\mathbf{V}^*$, where $\mathbf{\Sigma}_\mathbf{A}, \mathbf{\Sigma}_\mathbf{B} \in M_{m\times n}(F)$ carry $\sigma_i(\mathbf{A})$ and $\sigma_i(\mathbf{B})$ on the main diagonal and zeros elsewhere.

Idea. Write both singular value decompositions and push the four unitary factors into two: one, $\mathbf{M}$, sits between the two diagonal matrices, the other, $\mathbf{N}$, closes the trace up. Multiplying out, the trace becomes $\sum_{k,l} \sigma_k(\mathbf{A})\sigma_l(\mathbf{B}) m_{kl} n_{lk}$: every pairing of a singular value of $\mathbf{A}$ with one of $\mathbf{B}$ occurs, weighted by one entry of each unitary. The weights $|m_{kl} n_{lk}|$ are exactly where unitarity is spent — each row of $\mathbf{M}$ and each column of $\mathbf{N}$ is a unit vector, so Cauchy–Schwarz makes the weight array doubly substochastic — and Lemma 21.5.1 finishes.

Proof. Fix singular value decompositions $\mathbf{A} = \mathbf{U}_1\mathbf{\Sigma}_\mathbf{A}\mathbf{V}_1^*$ and $\mathbf{B} = \mathbf{U}_2\mathbf{\Sigma}_\mathbf{B}\mathbf{V}_2^*$ (Theorem 13.9.1), with $\mathbf{U}_1, \mathbf{U}_2 \in M_m(F)$ and $\mathbf{V}_1, \mathbf{V}_2 \in M_n(F)$ unitary and $\mathbf{\Sigma}_\mathbf{A}, \mathbf{\Sigma}_\mathbf{B}$ as in the statement. Put

$$\mathbf{M} := \mathbf{U}_1^*\mathbf{U}_2 \in M_m(F), \qquad \mathbf{N} := \mathbf{V}_2^*\mathbf{V}_1 \in M_n(F),$$

both unitary, being products of unitary matrices. Then $\mathbf{A}^*\mathbf{B} = \mathbf{V}_1\mathbf{\Sigma}_\mathbf{A}^*\mathbf{M}\mathbf{\Sigma}_\mathbf{B}\mathbf{V}_2^*$, so by Theorem 0.9.6 (3), moving the factor $\mathbf{V}_1$ to the back,

$$\operatorname{tr}(\mathbf{A}^*\mathbf{B}) = \operatorname{tr}\big(\mathbf{\Sigma}_\mathbf{A}^*\,\mathbf{M}\,\mathbf{\Sigma}_\mathbf{B}\,\mathbf{N}\big).$$

Now compute that trace entry by entry. The matrix $\mathbf{\Sigma}_\mathbf{A}^* \in M_{n\times m}(F)$ has (k, a) entry

$\sigma_k(\mathbf{A})\delta_{ka}$, which vanishes unless $k = a \le p$, and $\mathbf{\Sigma_B}$ has (b,l) entry $\sigma_b(\mathbf{B})\delta_{bl}$, which vanishes unless $b = l \le p$. Hence

$$\operatorname{tr}(\mathbf{A}^*\mathbf{B}) = \sum_{k=1}^{p}\sum_{l=1}^{p} \sigma_k(\mathbf{A})\,\sigma_l(\mathbf{B})\,m_{kl}\,n_{lk}, \tag{21.5.1}$$

all indices being in range because $p \le m$ and $p \le n$. By the triangle inequality (Theorem 0.6.4),

$$\left|\operatorname{tr}(\mathbf{A}^*\mathbf{B})\right| \;\le\; \sum_{k=1}^{p}\sum_{l=1}^{p} \sigma_k(\mathbf{A})\,\sigma_l(\mathbf{B})\,d_{kl}, \qquad d_{kl} \coloneqq |m_{kl}|\,|n_{lk}|.$$

Claim. $\mathbf{D} = (d_{kl})_{k,l\le p}$ is doubly substochastic.

Proof. Each $d_{kl} \ge 0$, which is (S1). Fix $k \le p$. By Cauchy–Schwarz (Theorem 11.1.5) in $\mathbb{R}^p$, and then by extending both sums to their full ranges $l \le m$ and $l \le n$, which only adds nonnegative terms,

$$\sum_{l=1}^{p}|m_{kl}||n_{lk}| \le \Big(\sum_{l=1}^{m}|m_{kl}|^2\Big)^{1/2}\Big(\sum_{l=1}^{n}|n_{lk}|^2\Big)^{1/2} = 1,$$

because row k of the unitary $\mathbf{M}$ and column k of the unitary $\mathbf{N}$ are unit vectors. This is (S2), and (S3) is the same computation with the roles of the two indices exchanged: column l of $\mathbf{M}$ and row l of $\mathbf{N}$ are unit vectors as well.

■

The lists $\sigma_1(\mathbf{A}) \ge \cdots \ge \sigma_p(\mathbf{A}) \ge 0$ and $\sigma_1(\mathbf{B}) \ge \cdots \ge \sigma_p(\mathbf{B}) \ge 0$ are decreasing and nonnegative (Definition 13.9.1), so Lemma 21.5.1 applies and gives

$$\sum_{k,l}\sigma_k(\mathbf{A})\sigma_l(\mathbf{B})d_{kl} \;\le\; \sum_{k=1}^{p}\sigma_k(\mathbf{A})\sigma_k(\mathbf{B}),$$

which is the stated bound.

For the last sentence, suppose $\mathbf{A} = \mathbf{U\Sigma_A V}^*$ and $\mathbf{B} = \mathbf{U\Sigma_B V}^*$ with the same $\mathbf{U}$ and $\mathbf{V}$. Then $\mathbf{A}^*\mathbf{B} = \mathbf{V\Sigma_A^*U}^*\mathbf{U\Sigma_B V}^* = \mathbf{V\Sigma_A^*\Sigma_B V}^*$, and $\mathbf{\Sigma_A^*\Sigma_B}$ is the $n\times n$ diagonal matrix with entries $\sigma_i(\mathbf{A})\sigma_i(\mathbf{B})$ for $i \le p$ and zeros after. Taking traces and using Theorem 0.9.6 (3) again, $\operatorname{tr}(\mathbf{A}^*\mathbf{B}) = \sum_{i\le p}\sigma_i(\mathbf{A})\sigma_i(\mathbf{B})$. This proves the theorem.

■

Three things about the proof are worth keeping. The weight matrix $\mathbf{D}$ is where all the information about the two sets of singular directions has gone; the only thing about it that survives is a pair of inequalities. Unitarity is spent exactly once, on the unit rows and columns inside the claim, and nothing else about $\mathbf{M}$ and $\mathbf{N}$ is used. And the equality case shows the bound is not merely an upper bound to be improved later: it is attained for every prescribed pair of singular value lists.

Warning. The singular values are paired **in matching order**, and no other pairing gives an upper bound. For $\mathbf{A} = \operatorname{diag}(2,1)$ and $\mathbf{B} = \operatorname{diag}(20,15)$ the reversed pairing is $2\cdot 15 + 1\cdot 20 = 50$, while $\operatorname{tr}(\mathbf{A}^*\mathbf{B}) = 40 + 15 = 55$. What is true for every permutation π is the **lower** bound

$\sum_i \sigma_i(\mathbf{A})\sigma_{\pi(i)}(\mathbf{B}) \le \sum_i \sigma_i(\mathbf{A})\sigma_i(\mathbf{B})$, which is the rearrangement inequality above.

Example. A strict case and an equality case

Let

$$\mathbf{A} = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}, \qquad \mathbf{B} = \begin{pmatrix} 12 & -12 \\ 16 & 9 \end{pmatrix}, \qquad \mathbf{B}' = \begin{pmatrix} 20 & 0 \\ 0 & 15 \end{pmatrix}.$$

Check that $\mathbf{B}$ and $\mathbf{B}'$ have the same singular values, and compare the two sides of Theorem 21.5.2 for the pairs $(\mathbf{A}, \mathbf{B})$ and $(\mathbf{A}, \mathbf{B}')$.

Solution. For $\mathbf{B}$,

$$\mathbf{B}^\top \mathbf{B} = \begin{pmatrix} 144 + 256 & -144 + 144 \\ -144 + 144 & 144 + 81 \end{pmatrix} = \begin{pmatrix} 400 & 0 \\ 0 & 225 \end{pmatrix},$$

so $\sigma(\mathbf{B}) = (20, 15)$ by Definition 13.9.1, and $\sigma(\mathbf{B}') = (20, 15)$ as well since $\mathbf{B}'$ is diagonal with positive decreasing entries. Also $\sigma(\mathbf{A}) = (2, 1)$. So both pairs have the same right-hand side,

$$\sigma_1(\mathbf{A})\sigma_1(\mathbf{B}) + \sigma_2(\mathbf{A})\sigma_2(\mathbf{B}) = 2 \cdot 20 + 1 \cdot 15 = 55.$$

On the left, $\operatorname{tr}(\mathbf{A}^\top \mathbf{B}') = 2 \cdot 20 + 1 \cdot 15 = 55$: equality. And $\operatorname{tr}(\mathbf{A}^\top \mathbf{B}) = 2 \cdot 12 + 1 \cdot 9 = 33 < 55$: strict.

The reason is visible in the decompositions. Both $\mathbf{A}$ and $\mathbf{B}'$ are diagonal with decreasing nonnegative entries, so $\mathbf{A} = \mathbf{I}\mathbf{A}\mathbf{I}^*$ and $\mathbf{B}' = \mathbf{I}\mathbf{B}'\mathbf{I}^*$ are aligned. For $\mathbf{B}$, on the other hand, $\mathbf{B} = \mathbf{U}_2 \operatorname{diag}(20, 15)$ with

$$\mathbf{U}_2 = \tfrac{1}{5} \begin{pmatrix} 3 & -4 \\ 4 & 3 \end{pmatrix},$$

a rotation, so in the notation of the proof $\mathbf{M} = \mathbf{U}_1^* \mathbf{U}_2 = \mathbf{U}_2$ and $\mathbf{N} = \mathbf{I}$. The weight matrix is $\mathbf{D} = \operatorname{diag}(\tfrac{3}{5}, \tfrac{3}{5})$, doubly substochastic and far from a permutation, and Eq. 21.5.1 reads $2 \cdot 20 \cdot \tfrac{3}{5} + 1 \cdot 15 \cdot \tfrac{3}{5} = 24 + 9 = 33$. The rotation has thrown away two fifths of each weight, and that is exactly the loss.

21.5.4 The dual of a unitarily invariant norm

The trace inequality is a statement about a pairing, so it is really a statement about duality. Recall the dual norm of Definition 18.4.4: it measures a vector by the largest value of the pairing against vectors the given norm calls short. On matrices, the pairing to use is the Frobenius one, and the result is exactly as clean as one could hope: dualizing a unitarily invariant norm dualizes its symmetric gauge and nothing else.

Corollary 21.5.3. Duality of Unitarily Invariant Norms

Let $\|\cdot\|$ be a unitarily invariant norm on $M_{m\times n}(F)$ (Definition 21.4.1) with symmetric gauge Φ on $\mathbb{R}^p$, $p = \min(m, n)$, so that $\|\mathbf{A}\| = \Phi(\sigma(\mathbf{A}))$ (Lemma 21.4.1). Define

$$\|\mathbf{B}\|^* := \max\{|\operatorname{tr}(\mathbf{B}^*\mathbf{A})| \ : \ \mathbf{A} \in M_{m\times n}(F),\ \|\mathbf{A}\| \le 1\}.$$

Then $\|\mathbf{B}\|^* = \Phi^*(\sigma(\mathbf{B}))$, where Φ^* is the dual gauge of Definition 21.3.3. In particular $\|\cdot\|^*$ is again a unitarily invariant norm, the trace norm $\|\cdot\|_{(p)}$ of Definition 21.4.2 and the spectral norm $\|\cdot\|_2$ are dual to each other, and the Frobenius norm is self-dual.

Idea. Both inequalities are one line each once the trace inequality is available. For $\leq$, bound the pairing by $\sum_i \sigma_i(\mathbf{A})\sigma_i(\mathbf{B})$ and then by $\Phi(\sigma(\mathbf{A}))\Phi^*(\sigma(\mathbf{B}))$, which is what a dual norm is for. For $\geq$, build the competitor $\mathbf{A}$ out of $\mathbf{B}$: take the vector where $\Phi^*(\sigma(\mathbf{B}))$ is attained and hang it on $\mathbf{B}$'s own singular vectors, which makes the pair aligned and turns the trace inequality into an equality.

Proof. Write $\sigma(\mathbf{B}) \in \mathbb{R}^p$ for the singular value list of $\mathbf{B}$, and fix a singular value decomposition $\mathbf{B} = \mathbf{U}\mathbf{\Sigma}_\mathbf{B}\mathbf{V}^*$ (Theorem 13.9.1).

($\leq$) Let $\|\mathbf{A}\| \leq 1$. By Theorem 21.5.2 and then by the defining property of a dual norm (Proposition 18.4.6 (b), applied to the norm Φ of Definition 21.3.3), which gives $\langle \mathbf{x}, \mathbf{y}\rangle \leq \Phi(\mathbf{x})\Phi^*(\mathbf{y})$ for all $\mathbf{x}, \mathbf{y}$,

$$\begin{aligned} |\operatorname{tr}(\mathbf{B}^*\mathbf{A})| &\leq \sum_{i=1}^{p} \sigma_i(\mathbf{A})\sigma_i(\mathbf{B}) \leq \Phi(\sigma(\mathbf{A}))\, \Phi^*(\sigma(\mathbf{B})) \\ &= \|\mathbf{A}\|\ \Phi^*(\sigma(\mathbf{B})) \leq \Phi^*(\sigma(\mathbf{B})). \end{aligned}$$

($\geq$) By Definition 21.3.3 there is $\mathbf{x} \in \mathbb{R}^p$ with $\Phi(\mathbf{x}) \leq 1$ and $\langle \mathbf{x}, \sigma(\mathbf{B})\rangle = \Phi^*(\sigma(\mathbf{B}))$. We may assume $x_1 \geq \cdots \geq x_p \geq 0$. Indeed, replacing each x_i by $|x_i|$ leaves $\Phi(\mathbf{x})$ unchanged, by the sign invariance in Definition 21.3.1, and does not decrease $\langle \mathbf{x}, \sigma(\mathbf{B})\rangle$ because every $\sigma_i(\mathbf{B}) \geq 0$; sorting the result decreasingly leaves Φ unchanged, by the permutation invariance in Definition 21.3.1, and does not decrease $\langle \mathbf{x}, \sigma(\mathbf{B})\rangle$ by the rearrangement inequality (Lemma 21.5.1 with a permutation matrix). Since $\Phi^*(\sigma(\mathbf{B}))$ is by definition the largest value of $\langle \cdot, \sigma(\mathbf{B})\rangle$ on $\{\Phi \leq 1\}$, the value is still exactly $\Phi^*(\sigma(\mathbf{B}))$. Now put

$$\mathbf{A} \coloneqq \mathbf{U}\,\mathbf{X}\,\mathbf{V}^*, \qquad \mathbf{X} \in M_{m\times n}(F) \text{ with } (\mathbf{X})_{ii} = x_i \ (i \leq p)$$

and all other entries 0. Since $x_1 \geq \cdots \geq x_p \geq 0$, this is a singular value decomposition of $\mathbf{A}$, so $\sigma(\mathbf{A}) = \mathbf{x}$ and $\|\mathbf{A}\| = \Phi(\mathbf{x}) \leq 1$. It is aligned with the chosen decomposition of $\mathbf{B}$, so the equality case of Theorem 21.5.2 gives

$$|\operatorname{tr}(\mathbf{B}^*\mathbf{A})| = \sum_{i=1}^{p} x_i\sigma_i(\mathbf{B}) = \Phi^*(\sigma(\mathbf{B})).$$

Hence the maximum is at least $\Phi^*(\sigma(\mathbf{B}))$, and the two bounds together prove the formula. (The maximum is attained, by the competitor just built.)

The remaining claims follow. By Theorem 21.3.5, Φ^* is a symmetric gauge, so $\|\cdot\|^*$ is the unitarily invariant norm it corresponds to under Theorem 21.4.3. The trace norm is the one with $\Phi = \|\cdot\|_1$ and the spectral norm the one with $\Phi = \|\cdot\|_\infty$, by Lemma 21.4.1 and Theorem 16.3.3 (c); these gauges are dual to each other by Proposition 18.4.7 (a), (b). The Frobenius norm has $\Phi = \|\cdot\|_2$, which is self-dual by Proposition 18.4.7 (c). This proves the corollary.

■

So the two extreme **normalized** unitarily invariant norms of Section 4 — the smallest, $\|\cdot\|_2$, and the largest, the trace norm, extreme among the normalized ones by Proposition 21.4.4 (c) — are exchanged by duality, and the Frobenius norm sits at the fixed point in the middle. Since $\Phi^{**} = \Phi$ (Theorem 21.3.5), dualizing twice returns the original norm.

Quick check. What does Corollary 21.5.3 say for $\mathbf{B} = \mathbf{e}_1\mathbf{e}_1^*$, a single matrix unit, and $|||\cdot|||$ the spectral norm?

Solution. Here $\sigma(\mathbf{B}) = (1, 0, \dots, 0)$, and the gauge of the spectral norm is $\|\cdot\|_\infty$, whose dual is $\|\cdot\|_1$. So $|||\mathbf{B}|||^* = \|(1, 0, \dots, 0)\|_1 = 1$: the largest value of $|\mathrm{tr}(\mathbf{B}^*\mathbf{A})| = |a_{11}|$ over matrices with $\|\mathbf{A}\|_2 \le 1$ is 1, attained at $\mathbf{A} = \mathbf{e}_1\mathbf{e}_1^*$. An entry of a matrix never exceeds its spectral norm, and for a matrix unit that is sharp.

21.5.5 Exercises

21.5.5.1 A. Check your understanding

Exercise. A1

(a) Define a doubly substochastic matrix, and say which doubly substochastic matrices are doubly stochastic.

(b) State Theorem 21.5.2, with its hypotheses, and say what "aligned" means.

(c) Determine whether the following is true for all $\mathbf{A}, \mathbf{B} \in M_n(\mathbb{C})$, and justify your answer: $|\mathrm{tr}(\mathbf{AB})| \le \sum_i \sigma_i(\mathbf{A})\sigma_i(\mathbf{B})$.

(d) Where in the proof of Theorem 21.5.2 is the unitarity of $\mathbf{M}$ and $\mathbf{N}$ used, and what exactly is deduced from it?

(e) Determine whether the following is true, and justify your answer: if $\mathrm{tr}(\mathbf{A}^*\mathbf{B}) = \sum_i \sigma_i(\mathbf{A})\sigma_i(\mathbf{B})$ then $\mathbf{A} = \mathbf{B}$.

Solution.

(a) A square real matrix with nonnegative entries whose every row sum and every column sum is at most 1 (Definition 21.5.1). It is doubly stochastic exactly when all those sums equal 1.

(b) For $\mathbf{A}, \mathbf{B} \in M_{m\times n}(F)$ and $p = \min(m, n)$, $|\mathrm{tr}(\mathbf{A}^*\mathbf{B})| \le \sum_{i\le p} \sigma_i(\mathbf{A})\sigma_i(\mathbf{B})$. The pair is aligned when there are unitary $\mathbf{U}, \mathbf{V}$ with $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}_\mathbf{A}\mathbf{V}^*$ and $\mathbf{B} = \mathbf{U}\mathbf{\Sigma}_\mathbf{B}\mathbf{V}^*$, the **same** $\mathbf{U}$ and $\mathbf{V}$ for both; then equality holds.

(c) True. Apply Theorem 21.5.2 to the pair $\mathbf{A}^*$ and $\mathbf{B}$, using $\mathrm{tr}(\mathbf{AB}) = \mathrm{tr}((\mathbf{A}^*)^*\mathbf{B})$ and $\sigma(\mathbf{A}^*) = \sigma(\mathbf{A})$, which is Proposition 21.4.4 (a).

(d) Only in the claim, and only to say that each row of $\mathbf{M}$ and each column of $\mathbf{N}$ is a unit vector. Cauchy–Schwarz then makes the weight array doubly substochastic. Nothing else about $\mathbf{M}$ and $\mathbf{N}$ is used.

(e) False. Take $\mathbf{A} = \mathbf{I}_2$ and $\mathbf{B} = 2\mathbf{I}_2$: both sides are 4, and $\mathbf{A} \ne \mathbf{B}$. The theorem says that alignment forces equality, and $\mathbf{A}$ and $\mathbf{B}$ are aligned, with $\mathbf{U} = \mathbf{V} = \mathbf{I}$; it says nothing that would force the two matrices to be equal.

21.5.5.2 B. Practice

Exercise. B1: Both sides of the inequality

Let

$$\mathbf{A} = \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix}, \qquad \mathbf{B} = \begin{pmatrix} 0 & 4 \\ 2 & 0 \end{pmatrix}.$$

Compute $\operatorname{tr}(\mathbf{A}^\top \mathbf{B})$, the singular values of both matrices, and both sides of Theorem 21.5.2. Then find a matrix $\mathbf{B}''$ with the same singular values as $\mathbf{B}$ for which equality holds.

Solution. $\mathbf{A}^\top \mathbf{B} = \left(\begin{smallmatrix} 0 & 12 \\ 2 & 0 \end{smallmatrix}\right)$, so $\operatorname{tr}(\mathbf{A}^\top \mathbf{B}) = 0$. Since $\mathbf{A}$ is diagonal with positive decreasing entries, $\sigma(\mathbf{A}) = (3, 1)$. For $\mathbf{B}$, $\mathbf{B}^\top \mathbf{B} = \operatorname{diag}(4, 16)$, so $\sigma(\mathbf{B}) = (4, 2)$. The right-hand side is $3 \cdot 4 + 1 \cdot 2 = 14$, and $0 \le 14$ with a lot to spare: $\mathbf{B}$ sends the direction $\mathbf{A}$ stretches most to the one it stretches least.

For equality, align: take $\mathbf{B}'' = \operatorname{diag}(4, 2)$. Then $\operatorname{tr}(\mathbf{A}^\top \mathbf{B}'') = 12 + 2 = 14$, and $\sigma(\mathbf{B}'') = (4, 2)$.

Exercise. B2: The trace against the trace norm

Let $\mathbf{A} \in M_n(\mathbb{C})$.

(a) Prove that $|\operatorname{tr} \mathbf{A}| \le \sum_{i=1}^{n} \sigma_i(\mathbf{A})$.

(b) Determine all $\mathbf{A}$ for which equality holds with $\mathbf{A}$ positive semidefinite.

Solution.

(a) Apply Theorem 21.5.2 with the pair $\mathbf{I}_n$ and $\mathbf{A}$, in that order: $\operatorname{tr}(\mathbf{I}^* \mathbf{A}) = \operatorname{tr} \mathbf{A}$, and $\sigma_i(\mathbf{I}) = 1$ for every i, so $|\operatorname{tr} \mathbf{A}| \le \sum_i 1 \cdot \sigma_i(\mathbf{A})$.

(b) For $\mathbf{A} \succeq \mathbf{O}$ the eigenvalues are nonnegative (Theorem 13.1.4) and, by the spectral theorem, $\mathbf{A} = \mathbf{U} \operatorname{diag}(\lambda_1, \dots, \lambda_n) \mathbf{U}^*$ with $\lambda_i \ge 0$ decreasing; this is a singular value decomposition, so $\sigma_i(\mathbf{A}) = \lambda_i(\mathbf{A})$. Hence $\operatorname{tr} \mathbf{A} = \sum_i \lambda_i = \sum_i \sigma_i(\mathbf{A})$ by Corollary 12.1.4: equality holds for **every** positive semidefinite $\mathbf{A}$. (Consistently with the theorem, $\mathbf{I}$ and such an $\mathbf{A}$ are aligned, with $\mathbf{V} = \mathbf{U}$.)

Exercise. B3: A weight matrix

Let $\mathbf{A} = \operatorname{diag}(5, 1)$ and let $\mathbf{B} = \mathbf{U} \operatorname{diag}(2, 1) \mathbf{U}^*$ with $\mathbf{U} = \frac{1}{\sqrt{2}} \left(\begin{smallmatrix} 1 & -1 \\ 1 & 1 \end{smallmatrix}\right)$. Write down the matrices $\mathbf{M}$ and $\mathbf{N}$ of the proof of Theorem 21.5.2 for this pair, the weight matrix $\mathbf{D}$, and verify Eq. 21.5.1 and the inequality numerically.

Solution. Take $\mathbf{U}_1 = \mathbf{V}_1 = \mathbf{I}$ for $\mathbf{A}$, and $\mathbf{U}_2 = \mathbf{U}$, $\mathbf{V}_2 = \mathbf{U}$ for $\mathbf{B}$ (the decomposition given is a singular value decomposition, since $\operatorname{diag}(2, 1)$ is diagonal with positive decreasing entries). Then $\mathbf{M} = \mathbf{U}_1^* \mathbf{U}_2 = \mathbf{U}$ and $\mathbf{N} = \mathbf{V}_2^* \mathbf{V}_1 = \mathbf{U}^*$, so

$$m_{kl} n_{lk} = u_{kl}\, \overline{u_{kl}} = |u_{kl}|^2 = \tfrac{1}{2}$$

for all k, l, and $\mathbf{D} = \frac{1}{2}\left(\begin{smallmatrix} 1 & 1 \\ 1 & 1 \end{smallmatrix}\right)$, which is doubly stochastic, hence doubly substochastic. Now Eq. 21.5.1 gives

$$\operatorname{tr}(\mathbf{A}^*\mathbf{B}) = \tfrac{1}{2}(5 \cdot 2 + 5 \cdot 1 + 1 \cdot 2 + 1 \cdot 1) = \tfrac{18}{2} = 9,$$

which checks against $\mathbf{B} = \frac{1}{2}\left(\begin{smallmatrix} 3 & 1 \\ 1 & 3 \end{smallmatrix}\right)$ and $\operatorname{tr}(\mathbf{AB}) = \frac{1}{2}(5{\cdot}3{+}1{\cdot}3) = 9$. The bound is $5{\cdot}2{+}1{\cdot}1 = 11$, and $9 \leq 11$.

21.5.5.3 C. Going deeper

Exercise. C1: The triangle inequality for the trace norm

Write $\|\mathbf{A}\|_{(p)} = \sum_i \sigma_i(\mathbf{A})$ for the trace norm of Definition 21.4.2, $\mathbf{A} \in M_{m\times n}(\mathbb{C})$.

(a) Prove that $\|\mathbf{A}\|_{(p)} = \max\{|\operatorname{tr}(\mathbf{A}^*\mathbf{Q})| : \mathbf{Q} \in M_{m\times n}(\mathbb{C}),\ \|\mathbf{Q}\|_2 \leq 1\}$.

(b) Hence deduce that $\|\cdot\|_{(p)}$ satisfies the triangle inequality, without using Lemma 21.4.2.

Hint: for (a), Corollary 21.5.3 with the spectral norm, read the other way round.

Solution.

(a) The spectral norm is unitarily invariant, with gauge $\Phi = \|\cdot\|_\infty$ (Lemma 21.4.1, Theorem 16.3.3 (c)). By Corollary 21.5.3 applied to that norm, the maximum in the statement equals $\Phi^*(\sigma(\mathbf{A})) = \|\sigma(\mathbf{A})\|_1 = \sum_i \sigma_i(\mathbf{A})$, using Proposition 18.4.7 (b).

(b) Conjugate transposition is additive, so $\operatorname{tr}((\mathbf{A}+\mathbf{B})^*\mathbf{Q}) = \operatorname{tr}(\mathbf{A}^*\mathbf{Q}) + \operatorname{tr}(\mathbf{B}^*\mathbf{Q})$ for every $\mathbf{Q}$. Given $\mathbf{A}, \mathbf{B}$, choose by (a) a $\mathbf{Q}$ with $\|\mathbf{Q}\|_2 \leq 1$ attaining the maximum for $\mathbf{A}+\mathbf{B}$; then, by the triangle inequality for the modulus (Theorem 0.6.4) and (a) again,

$$\|\mathbf{A}+\mathbf{B}\|_{(p)} = |\operatorname{tr}((\mathbf{A}+\mathbf{B})^*\mathbf{Q})| \leq |\operatorname{tr}(\mathbf{A}^*\mathbf{Q})| + |\operatorname{tr}(\mathbf{B}^*\mathbf{Q})| \leq \|\mathbf{A}\|_{(p)} + \|\mathbf{B}\|_{(p)}\,.$$

No step used Lemma 21.4.2: the route ran through the trace inequality and duality alone.

Exercise. C2: The nearest unitary, again

Let $\mathbf{A} \in M_n(\mathbb{C})$ with polar decomposition $\mathbf{A} = \mathbf{W}|\mathbf{A}|$ (Theorem 13.10.1).

(a) Prove that $\max\{\operatorname{Re}\operatorname{tr}(\mathbf{Q}^*\mathbf{A}) : \mathbf{Q} \in M_n(\mathbb{C}) \text{ unitary}\} = \sum_i \sigma_i(\mathbf{A})$, attained at $\mathbf{Q} = \mathbf{W}$.

(b) Deduce Corollary 13.10.2, that $\mathbf{W}$ minimizes $\|\mathbf{A} - \mathbf{Q}\|_F$ over unitary $\mathbf{Q}$.

Solution.

(a) For unitary $\mathbf{Q}$, $\sigma_i(\mathbf{Q}) = 1$ for every i, so Theorem 21.5.2 gives $\operatorname{Re}\operatorname{tr}(\mathbf{Q}^*\mathbf{A}) \leq |\operatorname{tr}(\mathbf{Q}^*\mathbf{A})| \leq \sum_i \sigma_i(\mathbf{A})$. For attainment, fix a singular value decomposition $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$ and recall from the proof of Theorem 13.10.1 that $\mathbf{W} = \mathbf{U}\mathbf{V}^*$. Then $\mathbf{W}^*\mathbf{A} = \mathbf{V}\mathbf{U}^*\mathbf{U}\mathbf{\Sigma}\mathbf{V}^* = \mathbf{V}\mathbf{\Sigma}\mathbf{V}^*$, whose trace is $\operatorname{tr}\mathbf{\Sigma} = \sum_i \sigma_i(\mathbf{A})$ by Theorem 0.9.6 (3); it is real.

(b) For unitary $\mathbf{Q}$, expanding in the Frobenius inner product,

$$\|\mathbf{A} - \mathbf{Q}\|_F^2 = \|\mathbf{A}\|_F^2 + n - 2\operatorname{Re}\operatorname{tr}(\mathbf{Q}^*\mathbf{A}),$$

since $\|\mathbf{Q}\|_F^2 = \operatorname{tr}(\mathbf{Q}^*\mathbf{Q}) = n$. Only the last term depends on $\mathbf{Q}$, and it enters with a

minus sign, so minimizing the left side is maximizing $\operatorname{Re}\operatorname{tr}(\mathbf{Q}^*\mathbf{A})$. By (a) the maximum is attained at $\mathbf{Q} = \mathbf{W}$, which is Corollary 13.10.2.

21.6 Ky Fan Dominance

Three earlier chapters have deferred the same kind of statement to this one: that a matrix is closest to another in **every** unitarily invariant norm at once. Such a statement looks like infinitely many theorems, one for each norm. This section shows that it is one theorem about singular values, and that it can be tested on a short finite list of norms. Then it collects the debts.

21.6.1 One weak majorization, every norm

Section 3 proved that a symmetric gauge respects weak majorization, and Section 4 matched the gauges with the unitarily invariant norms. Put together, they say that a weak majorization of singular values forces an inequality of norms. What is new here is the **converse**: nothing else forces it, and the Ky Fan k-norms already see everything.

To compare two matrices in every unitarily invariant norm, compare the running totals of their singular values.

Theorem 21.6.1. Ky Fan Dominance Theorem

Let $\mathbf{A}, \mathbf{B} \in M_{m\times n}(F)$ with $F = \mathbb{R}$ or $\mathbb{C}$, and let $p = \min(m, n)$. The following are equivalent.

(a) $\sigma(\mathbf{A}) \prec_w \sigma(\mathbf{B})$, that is, $\sum_{i\le k} \sigma_i(\mathbf{A}) \le \sum_{i\le k} \sigma_i(\mathbf{B})$ for every $k = 1, \dots, p$.

(b) $|||\mathbf{A}||| \le |||\mathbf{B}|||$ for **every** unitarily invariant norm $|||\cdot|||$ on $M_{m\times n}(F)$.

Idea.

(a) $\Rightarrow$ (b) is the dictionary of Sections 3 and 4 read in one direction: a unitarily invariant norm is a symmetric gauge evaluated at the singular values, and a symmetric gauge is monotone under weak majorization of nonnegative vectors. For (b) $\Rightarrow$ (a) we need, for each k, a unitarily invariant norm that computes the k-th running total. The Ky Fan k-norm is exactly that, so the p inequalities in (a) are p of the infinitely many inequalities in (b).

Proof.

(a) $\Rightarrow$ (b). Let $|||\cdot|||$ be a unitarily invariant norm and let Φ be its symmetric gauge, so that $|||\mathbf{M}||| = \Phi(\sigma(\mathbf{M}))$ for every $\mathbf{M}$ (Theorem 21.4.3). The vectors $\sigma(\mathbf{A})$ and $\sigma(\mathbf{B})$ have nonnegative entries (Definition 13.9.1), so Theorem 21.3.4 (b) applies to the hypothesis $\sigma(\mathbf{A}) \prec_w \sigma(\mathbf{B})$ and gives $\Phi(\sigma(\mathbf{A})) \le \Phi(\sigma(\mathbf{B}))$, which is $|||\mathbf{A}||| \le |||\mathbf{B}|||$.

(b) $\Rightarrow$ (a). Fix $k \in \{1, \dots, p\}$. Section 3's k-th Ky Fan gauge $\Phi_k(\mathbf{x}) = \sum_{i\le k} |x|_i^{\downarrow}$ (Defini-

tion 21.3.2) is a symmetric gauge function, so by Theorem 21.4.3

$$\|\mathbf{M}\|_{(k)} := \Phi_k(\sigma(\mathbf{M})) = \sum_{i=1}^{k} \sigma_i(\mathbf{M})$$

is a unitarily invariant norm on $M_{m\times n}(F)$ — the Ky Fan k-norm of Definition 21.4.2 — the second equality holding because $\sigma(\mathbf{M})$ is already nonnegative and decreasing, so that its k largest absolute values are its first k entries. Applying (b) to this one norm gives

$$\sum_{i=1}^{k} \sigma_i(\mathbf{A}) \;\le\; \sum_{i=1}^{k} \sigma_i(\mathbf{B}).$$

As k was arbitrary and both lists are decreasing, this is (a), by Definition 17.8.1. This proves the theorem.

■

The theorem is a translation device, and it will be used in one direction almost every time: to prove an inequality for every unitarily invariant norm, produce the weak majorization. Note also what it says about testing: the condition "for every unitarily invariant norm" is equivalent to "for the p Ky Fan norms", so a statement quantified over an infinite family of norms is decided by finitely many numbers.

Warning. One norm is never enough. For $\mathbf{A} = \operatorname{diag}(1,1)$ and $\mathbf{B} = \operatorname{diag}(1,0)$ the spectral norms agree, $\|\mathbf{A}\|_2 = \|\mathbf{B}\|_2 = 1$, yet $\sigma(\mathbf{A}) = (1,1)$ is not weakly majorized by $\sigma(\mathbf{B}) = (1,0)$, and indeed the trace norms are 2 and 1. Dominance needs **all** the running totals, and it compares matrices of the **same size**, since otherwise the two sides are norms on different spaces.

Quick check. Let $\mathbf{A}, \mathbf{B} \in M_{m\times n}(F)$ satisfy $\sigma_i(\mathbf{A}) \le \sigma_i(\mathbf{B})$ for every i. Does $\|\mathbf{A}\| \le \|\mathbf{B}\|$ hold for every unitarily invariant norm?

Solution. Yes. Entrywise domination of two decreasing nonnegative lists gives $\sum_{i\le k} \sigma_i(\mathbf{A}) \le \sum_{i\le k} \sigma_i(\mathbf{B})$ for every k, which is $\sigma(\mathbf{A}) \prec_w \sigma(\mathbf{B})$; now apply Theorem 21.6.1. The converse fails: $\sigma(\mathbf{A}) = (1,1)$ and $\sigma(\mathbf{B}) = (2,0)$ satisfy $\sigma(\mathbf{A}) \prec_w \sigma(\mathbf{B})$ with $\sigma_2(\mathbf{A}) > \sigma_2(\mathbf{B})$.

21.6.2 Low-rank approximation in every unitarily invariant norm

Chapter 13 §11 proved that the truncated singular value decomposition $\mathbf{A}_k$ is the best rank-k approximation in the Frobenius norm, and Chapter 17 §09 proved the same for the spectral norm. Both promised more. Chapter 13 §11 said that the two statements are "instances of one theorem: $\mathbf{A}_k$ is a best rank-k approximation in **every** unitarily invariant norm", and predicted the shape of the proof: "it needs inequalities between the singular values of $\mathbf{A}$ and of $\mathbf{A} - \mathbf{B}$, not a projection argument". Chapter 16 §08 recorded the same debt — "The version covering

every unitarily invariant norm at once is Chapter 21's" — and Chapter 17 §09 said that the common minimizer "is optimal in every unitarily invariant norm at once, which is Chapter 21's theorem". Here it is, by the predicted route.

Corollary 21.6.2. Eckart–Young Theorem in Every Unitarily Invariant Norm

Let $\mathbf{A} \in M_{m\times n}(F)$ with singular values $\sigma_1 \geq \cdots \geq \sigma_p$, $p = \min(m, n)$, let $0 \leq k \leq p$, and let $\mathbf{A}_k$ be a truncation of Definition 13.11.1. Then for **every** $\mathbf{B} \in M_{m\times n}(F)$ with $\operatorname{rank}\mathbf{B} \leq k$ and **every** unitarily invariant norm on $M_{m\times n}(F)$,

$$|||\mathbf{A} - \mathbf{A}_k||| \;\leq\; |||\mathbf{A} - \mathbf{B}|||,$$

and the minimum value is $|||\mathbf{A} - \mathbf{A}_k||| = \Phi(\sigma_{k+1}, \ldots, \sigma_p, 0, \ldots, 0)$, where Φ is the symmetric gauge of the norm.

Idea. The singular values of the error $\mathbf{A} - \mathbf{A}_k$ are the discarded $\sigma_{k+1}, \ldots, \sigma_p$, padded with zeros. For an arbitrary competitor $\mathbf{B}$ of rank at most k, Weyl's inequality for singular values, applied with the splitting $\mathbf{A} = (\mathbf{A} - \mathbf{B}) + \mathbf{B}$, says that $\mathbf{B}$ can shift the singular value list of $\mathbf{A}$ by at most k places, because $\sigma_{k+1}(\mathbf{B}) = 0$. That is precisely $\sigma_i(\mathbf{A} - \mathbf{B}) \geq \sigma_{k+i}(\mathbf{A})$: the truncation wins index by index, which is more than dominance needs.

Proof. **The error of the truncation.** Fix the singular value decomposition $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$ (Theorem 13.9.1) used to form $\mathbf{A}_k = \sum_{i\leq k} \sigma_i \mathbf{u}_i \mathbf{v}_i^*$ (Definition 13.11.1). Then

$$\mathbf{A} - \mathbf{A}_k = \sum_{i>k} \sigma_i \mathbf{u}_i \mathbf{v}_i^* = \mathbf{U}\mathbf{\Sigma}^{(k)}\mathbf{V}^*,$$

where $\mathbf{\Sigma}^{(k)}$ is $\mathbf{\Sigma}$ with its first k diagonal entries replaced by 0. Since $\mathbf{U}$ and $\mathbf{V}$ are unitary, $(\mathbf{A} - \mathbf{A}_k)^*(\mathbf{A} - \mathbf{A}_k) = \mathbf{V}(\mathbf{\Sigma}^{(k)})^*\mathbf{\Sigma}^{(k)}\mathbf{V}^*$ has eigenvalues $\sigma_{k+1}^2, \ldots, \sigma_p^2$ together with zeros, so by Definition 13.9.1

$$\sigma_i(\mathbf{A} - \mathbf{A}_k) = \sigma_{k+i}(\mathbf{A}) \qquad (1 \leq i \leq p), \tag{21.6.1}$$

with the convention $\sigma_j(\mathbf{A}) = 0$ for **every** $j > p$. Chapter 17 §09 states it for $p < j \leq n$; we extend it here, because the indices $k + i$ above run as far as $2p$, which may exceed n.

The competitor. Let $\operatorname{rank}\mathbf{B} \leq k$. Then $\sigma_{k+1}(\mathbf{B}) = 0$, since a matrix has exactly $\operatorname{rank}\mathbf{B}$ nonzero singular values (Theorem 13.9.1). Fix $i \leq p$. If $i + k > n$, then $\sigma_{k+i}(\mathbf{A}) = 0 \leq \sigma_i(\mathbf{A} - \mathbf{B})$ by the convention. Otherwise $i + (k + 1) - 1 = i + k \leq n$, so Theorem 17.9.4 applies to the sum $\mathbf{A} = (\mathbf{A} - \mathbf{B}) + \mathbf{B}$ with indices i and $k + 1$ and gives

$$\sigma_{k+i}(\mathbf{A}) \;\leq\; \sigma_i(\mathbf{A} - \mathbf{B}) + \sigma_{k+1}(\mathbf{B}) = \sigma_i(\mathbf{A} - \mathbf{B}).$$

Either way, with Eq. 21.6.1,

$$\sigma_i(\mathbf{A} - \mathbf{A}_k) \;\leq\; \sigma_i(\mathbf{A} - \mathbf{B}) \qquad (1 \leq i \leq p).$$

Summing the first k' of these for each $k' \leq p$ gives $\sigma(\mathbf{A}-\mathbf{A}_k) \prec_w \sigma(\mathbf{A}-\mathbf{B})$, so Theorem 21.6.1 gives $|||\mathbf{A} - \mathbf{A}_k||| \leq |||\mathbf{A} - \mathbf{B}|||$ for every unitarily invariant norm.

The value. By Theorem 21.4.3 and Eq. 21.6.1,

$$|||\mathbf{A} - \mathbf{A}_k||| = \Phi(\sigma(\mathbf{A} - \mathbf{A}_k)) = \Phi(\sigma_{k+1}, \dots, \sigma_p, 0, \dots, 0).$$

This proves the corollary.

■

The proof gives more than was asked: the error of the truncation is smaller than that of any competitor **singular value by singular value**, not merely in every norm. Specializing Φ recovers the two known cases — $\Phi = \|\cdot\|_\infty$ gives σ_{k+1}, which is Theorem 17.9.6, and $\Phi = \|\cdot\|_2$ gives $(\sum_{i>k} \sigma_i^2)^{1/2}$, which is Theorem 13.11.3 — and now every other norm comes with them.

Example. One matrix, four norms

Let

$$\mathbf{A} = \begin{pmatrix} 6 & -4 & 0 \\ 8 & 3 & 0 \\ 0 & 0 & 1 \end{pmatrix} \in M_3(\mathbb{R}).$$

Find $\mathbf{A}_1$ and $\mathbf{A}_2$, and tabulate the approximation errors in the spectral, Frobenius, trace, and Ky Fan 2-norms. Compare $\mathbf{A}_1$ with the rank-one competitor $\mathbf{C} = \operatorname{diag}(0, 0, 1)$.

Solution. Write $\mathbf{Q} = \frac{1}{5}\begin{pmatrix} 3 & -4 & 0 \\ 4 & 3 & 0 \\ 0 & 0 & 5 \end{pmatrix}$, which is orthogonal, since its columns are orthonormal. Then $\mathbf{A} = \mathbf{Q}\operatorname{diag}(10, 5, 1)$, so this is a singular value decomposition with $\mathbf{U} = \mathbf{Q}$, $\mathbf{V} = \mathbf{I}$, and $\sigma(\mathbf{A}) = (10, 5, 1)$. Hence

$$\mathbf{A}_1 = \begin{pmatrix} 6 & 0 & 0 \\ 8 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \qquad \mathbf{A}_2 = \begin{pmatrix} 6 & -4 & 0 \\ 8 & 3 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

with errors $\mathbf{A} - \mathbf{A}_1 = \mathbf{Q}\operatorname{diag}(0, 5, 1)$ and $\mathbf{A} - \mathbf{A}_2 = \mathbf{Q}\operatorname{diag}(0, 0, 1)$, whose singular value lists are $(5, 1, 0)$ and $(1, 0, 0)$, as Eq. 21.6.1 predicts.

error measured in	$\mathbf{A} - \mathbf{A}_1$	$\mathbf{A} - \mathbf{A}_2$	$\mathbf{A} - \mathbf{C}$
spectral $\|\cdot\|_2$	5	1	10
Frobenius $\|\cdot\|_F$	$\sqrt{26}$	1	$\sqrt{125}$
trace norm	6	1	15
Ky Fan 2-norm (Definition 21.4.2)	6	1	15

The competitor $\mathbf{C} = \operatorname{diag}(0, 0, 1)$ has rank 1, and $\mathbf{A} - \mathbf{C} = \mathbf{A}_2 = \mathbf{Q}\operatorname{diag}(10, 5, 0)$ has singular values $(10, 5, 0)$. It loses to $\mathbf{A}_1$ in all four norms at once, as it must: $(5, 1, 0)$ is dominated entry by entry by $(10, 5, 0)$. Note also that the Ky Fan 2-norm and the trace norm agree on these three errors, because each has at most two nonzero singular values.

21.6.3 The nearest unitary matrix

Chapter 13 §10 proved that the unitary polar factor $\mathbf{W}$ of $\mathbf{A} = \mathbf{W}|\mathbf{A}|$ is the nearest unitary matrix in the Frobenius norm (Corollary 13.10.2) and warned: "It is true, though not proved here, that $\mathbf{W}$ also minimizes the distance in every unitarily invariant norm; Chapter 21 takes up that family of norms." That is this corollary.

The proof passes through a symmetric list of the form $(z_1, \dots, z_q, 0, \dots, 0, -z_q, \dots, -z_1)$, and has to read off its top-k sums. Section 7 meets the same list again, so we settle the counting once.

Lemma 21.6.3. Top Sums of a Symmetric List

Let $z_1, \dots, z_q$ be real numbers, let $r \geq 0$ be an integer, and let $\mathbf{c} \in \mathbb{R}^{2q+r}$ be the list

$$\mathbf{c} = (z_1, \dots, z_q, \underbrace{0, \dots, 0}_{r}, -z_q, \dots, -z_1).$$

Then for every $k \leq q$, the largest sum of k entries of $\mathbf{c}$, which is the running total $\sum_{j \leq k} c_j^{\downarrow}$, equals the sum of the k largest of the numbers $|z_1|, \dots, |z_q|$.

Proof. Fix $k \leq q$. Write S for the sum of the k largest of the $|z_i|$. Since the largest sum of k entries of $\mathbf{c}$ is obtained by taking the k largest entries, it is $\sum_{j \leq k} c_j^{\downarrow}$, and the two descriptions agree.

($\geq$) For each index i, the number $|z_i|$ is an entry of $\mathbf{c}$: it is z_i in position i if $z_i \geq 0$, and $-z_i$ in position $2q + r + 1 - i$ otherwise. Distinct indices i give distinct positions, so the k largest of the $|z_i|$ occur at k distinct positions of $\mathbf{c}$, and their sum S is one of the sums being maximized.

($\leq$) Let k positions be given. The entries in the middle range contribute 0, and the remaining positions belong to at most k of the indices i. For one such index, the chosen positions contribute z_i, or $-z_i$, or (if both are chosen) 0, hence at most $|z_i|$ in every case. So the sum of the chosen entries is at most $\sum_{i \in J} |z_i|$ for a set J of at most k indices, and since every $|z_i| \geq 0$ that is at most S. This proves the lemma. ∎

Corollary 21.6.4. The Polar Factor Is Nearest in Every Unitarily Invariant Norm

Let $\mathbf{A} \in M_n(\mathbb{C})$ with polar decomposition $\mathbf{A} = \mathbf{W}|\mathbf{A}|$ (Theorem 13.10.1). Then for every unitary $\mathbf{Q} \in M_n(\mathbb{C})$ and every unitarily invariant norm on $M_n(\mathbb{C})$,

$$|||\mathbf{A} - \mathbf{W}||| \;\leq\; |||\mathbf{A} - \mathbf{Q}|||.$$

Idea. Peel off $\mathbf{W}$. Unitary invariance turns the comparison of $\mathbf{A} - \mathbf{Q}$ with $\mathbf{A} - \mathbf{W}$ into a comparison of $|\mathbf{A}| - \mathbf{U}$ with $|\mathbf{A}| - \mathbf{I}$, where $\mathbf{U} = \mathbf{W}^*\mathbf{Q}$ runs over all unitary matrices: the question is why the identity is the unitary matrix closest to a positive semidefinite one. By dominance it is enough to weakly majorize singular values, and singular values of a difference of two matrices are eigenvalues of a difference of two Hermitian ones, through the dilation

of Chapter 17 §09. There, Lidskii's theorem is waiting: it already says that the vector of eigenvalue shifts is majorized by the eigenvalues of the difference, and choosing the right k entries of that majorization is the whole computation.

Proof. **Step 1: reduce to $|\mathbf{A}|$ and $\mathbf{I}$.** Let $\mathbf{Q}$ be unitary and put $\mathbf{U} \coloneqq \mathbf{W}^*\mathbf{Q}$, which is unitary. Then

$$\mathbf{A}-\mathbf{Q} = \mathbf{W}|\mathbf{A}| - \mathbf{Q} = \mathbf{W}(|\mathbf{A}| - \mathbf{W}^*\mathbf{Q}) = \mathbf{W}(|\mathbf{A}| - \mathbf{U}),$$

so $\|\mathbf{A}-\mathbf{Q}\| = \||\mathbf{A}| - \mathbf{U}\|$ by unitary invariance (Definition 21.4.1). Taking $\mathbf{Q} = \mathbf{W}$ gives $\mathbf{U} = \mathbf{I}$ and $\|\mathbf{A}-\mathbf{W}\| = \||\mathbf{A}| - \mathbf{I}\|$. Writing $\mathbf{S} = |\mathbf{A}| \succeq \mathbf{O}$, it suffices to prove

$$\|\mathbf{S}-\mathbf{I}\| \le \|\mathbf{S}-\mathbf{U}\| \qquad \text{for every unitary } \mathbf{U}.$$

By Theorem 21.6.1 this is the weak majorization $\sigma(\mathbf{S}-\mathbf{I}) \prec_w \sigma(\mathbf{S}-\mathbf{U})$.

Step 2: the singular values of $\mathbf{S}-\mathbf{I}$. Let $t_1 \ge \cdots \ge t_n \ge 0$ be the eigenvalues of $\mathbf{S}$, which are nonnegative because $\mathbf{S} \succeq \mathbf{O}$ (Theorem 13.1.4). The spectral theorem writes $\mathbf{S} = \mathbf{Y}\,\mathrm{diag}(t_1,\dots,t_n)\mathbf{Y}^*$ with $\mathbf{Y}$ unitary, and this is a singular value decomposition, so $\sigma_i(\mathbf{S}) = t_i$. The matrix $\mathbf{S}-\mathbf{I}$ is Hermitian with eigenvalues $t_i - 1$, and $(\mathbf{S}-\mathbf{I})^*(\mathbf{S}-\mathbf{I}) = (\mathbf{S}-\mathbf{I})^2$ has eigenvalues $(t_i-1)^2$, so by Definition 13.9.1

$$\sigma(\mathbf{S}-\mathbf{I}) = \big(|t_1-1|,\dots,|t_n-1|\big)^{\downarrow}.$$

Write $s_j = \sigma_j(\mathbf{S}-\mathbf{U})$ for the numbers we must bound from below.

Step 3: dilate. Let $\mathcal{H}(\cdot)$ be the Hermitian dilation of Proposition 17.9.3. Conjugate transposition is additive, so $\mathcal{H}(\mathbf{S}) - \mathcal{H}(\mathbf{U}) = \mathcal{H}(\mathbf{S}-\mathbf{U})$, and all three are Hermitian matrices in $M_{2n}(\mathbb{C})$. By Proposition 17.9.3, the eigenvalues of $\mathcal{H}(\mathbf{M})$ are $\pm\sigma_i(\mathbf{M})$, and since the σ_i are nonnegative and decreasing, the decreasing list is

$$\boldsymbol{\lambda}(\mathcal{H}(\mathbf{M})) = (\sigma_1(\mathbf{M}),\dots,\sigma_n(\mathbf{M}),-\sigma_n(\mathbf{M}),\dots,-\sigma_1(\mathbf{M})).$$

Applied to $\mathbf{S}$, to $\mathbf{U}$ (all of whose singular values are 1) and to $\mathbf{S}-\mathbf{U}$, this gives the three lists

$$\begin{aligned}\boldsymbol{\lambda}(\mathcal{H}(\mathbf{S})) &= (t_1,\dots,t_n,-t_n,\dots,-t_1),\\ \boldsymbol{\lambda}(\mathcal{H}(\mathbf{U})) &= (1,\dots,1,-1,\dots,-1),\\ \boldsymbol{\lambda}(\mathcal{H}(\mathbf{S}-\mathbf{U})) &= (s_1,\dots,s_n,-s_n,\dots,-s_1).\end{aligned}$$

Step 4: apply Lidskii's theorem. Chapter 17 §07 proved, in the remark following Theorem 17.7.2, that $\boldsymbol{\lambda}(\mathbf{X}) - \boldsymbol{\lambda}(\mathbf{Y}) \prec \boldsymbol{\lambda}(\mathbf{X}-\mathbf{Y})$ for Hermitian $\mathbf{X},\mathbf{Y}$. Applied to $\mathbf{X} = \mathcal{H}(\mathbf{S})$ and $\mathbf{Y} = \mathcal{H}(\mathbf{U})$,

$$\boldsymbol{\lambda}(\mathcal{H}(\mathbf{S})) - \boldsymbol{\lambda}(\mathcal{H}(\mathbf{U})) \;\prec\; \boldsymbol{\lambda}(\mathcal{H}(\mathbf{S}-\mathbf{U})).$$

By Step 3 the vector on the left is

$$\mathbf{d} = (t_1-1,\dots,t_n-1,\; -(t_n-1),\dots,-(t_1-1)) \in \mathbb{R}^{2n},$$

which is the symmetric list of Lemma 21.6.3 with $q = n$, $r = 0$, and $z_i = t_i - 1$. Fix $k \le n$. By that lemma the k-th running total of $\mathbf{d}^{\downarrow}$ is the sum of the k largest of the numbers $|t_i - 1|$,

which by Step 2 is $\sum_{j\le k}\sigma_j(\mathbf{S}-\mathbf{I})$. The right-hand list is decreasing with nonnegative first half, so its k-th running total is $\sum_{j\le k}s_j$. So (M1) of Definition 17.8.1, at the index k, reads

$$\sum_{j=1}^{k}\sigma_j(\mathbf{S}-\mathbf{I}) \;\le\; \sum_{j=1}^{k}\sigma_j(\mathbf{S}-\mathbf{U}).$$

As this holds for every $k \le n$, it is the weak majorization required in Step 1. This proves the corollary.

■

The mechanism deserves a name, because Section 7 runs it again in general: **a Hermitian dilation turns a statement about singular values into one about eigenvalues**, where Lidskii's theorem applies. What was special here is that both matrices were **square**, so that their dilations had no zeros in the middle of the spectrum and no bookkeeping was needed; that one of the two was unitary is what made its dilated eigenvalue list the constant ± 1.

Remark. The inequality of Corollary 21.6.4 cannot be improved to $\sigma_j(\mathbf{A}-\mathbf{W}) \le \sigma_j(\mathbf{A}-\mathbf{Q})$ index by index. Take $\mathbf{A}=\mathbf{S}=\operatorname{diag}(2,0)$, which is positive semidefinite, so that $|\mathbf{A}|=\mathbf{A}$ and $\mathbf{W}=\mathbf{I}$ is a polar factor, and take $\mathbf{Q}=\left(\begin{smallmatrix}0&1\\1&0\end{smallmatrix}\right)$. Then the matrix $\mathbf{S}-\mathbf{Q}=\left(\begin{smallmatrix}2&-1\\-1&0\end{smallmatrix}\right)$ is symmetric with eigenvalues $1\pm\sqrt{2}$, so $\sigma(\mathbf{S}-\mathbf{Q})=(1+\sqrt{2},\sqrt{2}-1)\approx(2.414,0.414)$, while $\sigma(\mathbf{S}-\mathbf{I})=(1,1)$. The second singular value is smaller for $\mathbf{Q}$, and only the running totals, $1\le 2.414$ and $2\le 2.828$, go the right way. So the weak majorization, and not something cruder, is what the corollary needs.

21.6.4 Two smaller consequences

A compression never increases a unitarily invariant norm.

Corollary 21.6.5. Compressions Are Contractions

Let $\mathbf{A}\in M_n(F)$ and let $\mathbf{P}\in M_n(F)$ be an orthogonal projection. Then $|||\mathbf{PAP}||| \le |||\mathbf{A}|||$ for every unitarily invariant norm.

Proof. If $\mathbf{P}=\mathbf{O}$ the left side is $|||\mathbf{O}|||=0\le|||\mathbf{A}|||$. Otherwise $\mathbf{P}$ is Hermitian with eigenvalues in $\{0,1\}$, at least one of them equal to 1, so $\|\mathbf{P}\|_2=1$ by Lemma 17.3.5. Applying the two bounds of Proposition 21.4.4 (b) in turn,

$$|||\mathbf{PAP}||| \le \|\mathbf{P}\|_2\,|||\mathbf{AP}||| \le \|\mathbf{P}\|_2\,|||\mathbf{A}|||\,\|\mathbf{P}\|_2 = |||\mathbf{A}|||\,,$$

as claimed.

■

And a negative result, about the subadditivity of singular values. Section 4 proved $\sigma(\mathbf{A}+\mathbf{B}) \prec_w \sigma(\mathbf{A})+\sigma(\mathbf{B})$ (Lemma 21.4.2), where the right-hand side is the entrywise sum of the two decreasing lists. The weak majorization cannot be upgraded to majorization.

Example. Weak majorization is the most that holds

Let $\mathbf{A} = \operatorname{diag}(1,0)$ and $\mathbf{B} = \operatorname{diag}(-1,0)$. Then $\sigma(\mathbf{A}) = \sigma(\mathbf{B}) = (1,0)$, so $\sigma(\mathbf{A}) + \sigma(\mathbf{B}) = (2,0)$, while $\mathbf{A} + \mathbf{B} = \mathbf{O}$ has $\sigma(\mathbf{A} + \mathbf{B}) = (0,0)$. The totals are 0 and 2, so $\sigma(\mathbf{A} + \mathbf{B}) \prec \sigma(\mathbf{A}) + \sigma(\mathbf{B})$ is false, although the weak majorization holds with room to spare.

The failure is not an accident of the example. Take $\mathbf{B} = -\mathbf{A}$ for any $\mathbf{A} \neq \mathbf{O}$: then $\sigma(\mathbf{B}) = \sigma(\mathbf{A})$, so the right-hand list has total $2\sum_i \sigma_i(\mathbf{A}) > 0$, while $\sigma(\mathbf{A} + \mathbf{B}) = \mathbf{0}$ has total 0. Cancellation destroys singular value mass, and clause (M2) of Definition 17.8.1 — equal totals — therefore fails for every nonzero $\mathbf{A}$. Weak majorization is the most Lemma 21.4.2 can assert.

21.6.5 Exercises

21.6.5.1 A. Check your understanding

Exercise. A1

(a) State Theorem 21.6.1, and say which family of norms proves the direction that is not a direct consequence of Sections 3 and 4.

(b) Define the Ky Fan k-norm and say why it is a unitarily invariant norm.

(c) Determine whether the following is true, and justify your answer: if $\|\mathbf{A}\|_F \leq \|\mathbf{B}\|_F$ and $\|\mathbf{A}\|_2 \leq \|\mathbf{B}\|_2$, then $|||\mathbf{A}||| \leq |||\mathbf{B}|||$ for every unitarily invariant norm.

(d) In the proof of Corollary 21.6.2, where is the hypothesis $\operatorname{rank}\mathbf{B} \leq k$ used?

(e) Determine whether the following is true, and justify your answer: for $\mathbf{A} \in M_n(\mathbb{C})$ invertible, the nearest unitary matrix in the trace norm is the same as the nearest one in the Frobenius norm.

Solution.

(a) For $\mathbf{A}, \mathbf{B}$ of the same size, $\sigma(\mathbf{A}) \prec_w \sigma(\mathbf{B})$ if and only if $|||\mathbf{A}||| \leq |||\mathbf{B}|||$ for every unitarily invariant norm. The direction (b) $\Rightarrow$ (a) is the one needing new input, and it is proved by applying the hypothesis to the Ky Fan k-norms.

(b) $|||\mathbf{M}|||_{(k)} = \sum_{i \leq k} \sigma_i(\mathbf{M})$ (Definition 21.4.2). It is the unitarily invariant norm attached by Theorem 21.4.3 to the k-th Ky Fan gauge Definition 21.3.2, evaluated at $\sigma(\mathbf{M})$.

(c) False. Two norms do not decide the family. Take $\mathbf{A} = \operatorname{diag}(1,1,0)$ and $\mathbf{B} = \operatorname{diag}(\sqrt{2},0,0)$: then $\|\mathbf{A}\|_F = \sqrt{2} = \|\mathbf{B}\|_F$ and $\|\mathbf{A}\|_2 = 1 \leq \sqrt{2} = \|\mathbf{B}\|_2$, but the trace norms are 2 and $\sqrt{2}$, so $|||\mathbf{A}||| \leq |||\mathbf{B}|||$ fails for the trace norm.

(d) Only to give $\sigma_{k+1}(\mathbf{B}) = 0$, which is what makes the Weyl inequality with second index $k+1$ say something about $\mathbf{A} - \mathbf{B}$ alone.

(e) False: in the trace norm the minimizer need not be unique, so there is no *the* nearest unitary matrix. The polar factor $\mathbf{W}$ is nearest in both norms, by Corollary 21.6.4, and for invertible $\mathbf{A}$ it is the **only** Frobenius minimizer (Corollary 13.10.2); in the trace norm it can share the minimum. Take $\mathbf{A} = \operatorname{diag}(2, \frac{1}{2})$, which is positive definite, so $|\mathbf{A}| = \mathbf{A}$ and $\mathbf{W} = \mathbf{I}$. Then $\mathbf{A} - \mathbf{I} = \operatorname{diag}(1, -\frac{1}{2})$ has trace norm $\frac{3}{2}$. The rotation $\mathbf{Q} = \frac{1}{5}\left(\begin{smallmatrix} 4 & -3 \\ 3 & 4 \end{smallmatrix}\right)$ is unitary, and

$$\mathbf{A} - \mathbf{Q} = \tfrac{1}{10}\begin{pmatrix} 12 & 6 \\ -6 & -3 \end{pmatrix}$$

has determinant 0 and is nonzero, so it has rank 1: its second singular value is 0, and the first is $\|\mathbf{A}-\mathbf{Q}\|_F = \frac{1}{10}\sqrt{144+36+36+9} = \frac{3}{2}$. Its trace norm is therefore $\frac{3}{2}$ as well, so $\mathbf{I}$ and $\mathbf{Q}$ are both nearest in the trace norm. In the Frobenius norm they are not tied: $\|\mathbf{A}-\mathbf{I}\|_F = \frac{\sqrt{5}}{2} < \frac{3}{2} = \|\mathbf{A}-\mathbf{Q}\|_F$.

21.6.5.2 B. Practice

Exercise. B1: Dominance by hand

Let $\mathbf{A}, \mathbf{B} \in M_3(\mathbb{R})$ have singular values $\sigma(\mathbf{A}) = (3,2,2)$ and $\sigma(\mathbf{B}) = (4,3,1)$.

(a) Determine whether $|||\mathbf{A}||| \le |||\mathbf{B}|||$ for every unitarily invariant norm.

(b) Answer the same question for $\sigma(\mathbf{A}') = (3,3,2)$ in place of $\sigma(\mathbf{A})$.

Solution.

(a) The running totals are $3,5,7$ for $\mathbf{A}$ and $4,7,8$ for $\mathbf{B}$, and $3 \le 4$, $5 \le 7$, $7 \le 8$. So $\sigma(\mathbf{A}) \prec_w \sigma(\mathbf{B})$ and Theorem 21.6.1 gives the inequality for every unitarily invariant norm.

(b) The running totals of $\sigma(\mathbf{A}')$ are $3,6,8$, all at most $4,7,8$, so again yes. This one is worth noticing, because the index-by-index comparison **fails**: $\sigma_3(\mathbf{A}') = 2 > 1 = \sigma_3(\mathbf{B})$. Dominance is a statement about running totals, and it is strictly weaker than entrywise domination.

Exercise. B2: Best approximations of a 2 by 2 matrix

Let $\mathbf{A} = \left(\begin{smallmatrix} 3 & 4 \\ 4 & -3 \end{smallmatrix}\right)$. Compute $\sigma(\mathbf{A})$, find the best rank-one approximation, and compute the error in the spectral, Frobenius, and trace norms. What is unusual about this matrix?

Solution. $\mathbf{A}^\top\mathbf{A} = \left(\begin{smallmatrix} 25 & 0 \\ 0 & 25 \end{smallmatrix}\right)$, so $\sigma(\mathbf{A}) = (5,5)$. By Corollary 21.6.2 the best rank-one approximation is $\mathbf{A}_1 = 5\mathbf{u}_1\mathbf{v}_1^*$ for a singular value decomposition of $\mathbf{A}$. One such decomposition is $\mathbf{A} = \mathbf{U}\,\mathrm{diag}(5,5)\mathbf{V}^\top$ with $\mathbf{V} = \mathbf{I}$ and $\mathbf{U} = \mathbf{A}/5$, which is orthogonal because $\mathbf{A}^\top\mathbf{A} = 25\mathbf{I}$; it gives $\mathbf{u}_1 = \frac{1}{5}(3,4)$, $\mathbf{v}_1 = \mathbf{e}_1$ and

$$\mathbf{A}_1 = \begin{pmatrix} 3 & 0 \\ 4 & 0 \end{pmatrix}, \qquad \mathbf{A} - \mathbf{A}_1 = \begin{pmatrix} 0 & 4 \\ 0 & -3 \end{pmatrix}.$$

The error has singular values $(\sigma_2, 0) = (5,0)$, as Corollary 21.6.2 predicts: spectral error 5, Frobenius error 5, trace-norm error 5. What is unusual is that $\sigma_1 = \sigma_2$, so the truncation is not unique: $\mathbf{A}/5$ is symmetric and orthogonal, and **every** unit vector is a right singular vector. Every rank-one matrix of the form $5\mathbf{u}\mathbf{v}^*$ arising from a singular value decomposition is a best approximation, and they are genuinely different matrices.

Exercise. B3: A compression

Let $\mathbf{A} = \left(\begin{smallmatrix} 2 & 1 \\ 1 & 3 \end{smallmatrix}\right)$ and let $\mathbf{P} = \mathbf{e}_1\mathbf{e}_1^\top$. Compute $|||\mathbf{PAP}|||$ and $|||\mathbf{A}|||$ for the trace norm, and check

Corollary 21.6.5.

Solution. $\mathbf{PAP} = \left(\begin{smallmatrix} 2 & 0 \\ 0 & 0 \end{smallmatrix}\right)$, whose singular values are $(2, 0)$, so its trace norm is 2. The matrix $\mathbf{A}$ is symmetric with $\operatorname{tr}\mathbf{A} = 5$ and $\det\mathbf{A} = 5$, so its eigenvalues are $(5 \pm \sqrt{5})/2$, both positive; hence $\mathbf{A} \succ \mathbf{O}$, its singular values are its eigenvalues, and its trace norm is $\operatorname{tr}\mathbf{A} = 5$. Indeed $2 \le 5$.

21.6.5.3 C. Going deeper

Exercise. C1: Dominance for sums

Let $\mathbf{A}, \mathbf{B} \in M_{m\times n}(\mathbb{C})$.

(a) Prove that $|||\mathbf{A}+\mathbf{B}||| \le |||\mathbf{A}||| + |||\mathbf{B}|||$ follows from Lemma 21.4.2 together with Theorem 21.3.4 (b), without using that $|||\cdot|||$ is a norm.

(b) Deduce that $\sum_{i\le k}\sigma_i(\mathbf{A}+\mathbf{B}) \le \sum_{i\le k}\sigma_i(\mathbf{A}) + \sum_{i\le k}\sigma_i(\mathbf{B})$ for every k.

Solution.

(a) By Lemma 21.4.2, $\sigma(\mathbf{A}+\mathbf{B}) \prec_w \sigma(\mathbf{A}) + \sigma(\mathbf{B})$, and all three vectors have nonnegative entries. Let Φ be the gauge of the norm. By Theorem 21.3.4 (b), $\Phi(\sigma(\mathbf{A}+\mathbf{B})) \le \Phi(\sigma(\mathbf{A}) + \sigma(\mathbf{B}))$, and by the triangle inequality for the norm Φ on $\mathbb{R}^p$ the right side is at most $\Phi(\sigma(\mathbf{A})) + \Phi(\sigma(\mathbf{B}))$. By Theorem 21.4.3 these are $|||\mathbf{A}+\mathbf{B}|||$, and $|||\mathbf{A}||| + |||\mathbf{B}|||$. (The triangle inequality used is that of Φ, a norm on $\mathbb{R}^p$, not that of $|||\cdot|||$.)

(b) Apply (a) to the Ky Fan k-norm, whose value at $\mathbf{M}$ is $\sum_{i\le k}\sigma_i(\mathbf{M})$.

Exercise. C2: Distance to the singular matrices, in every norm

Let $\mathbf{A} \in M_n(\mathbb{C})$ be invertible with singular values $\sigma_1 \ge \cdots \ge \sigma_n > 0$, and let $|||\cdot|||$ be a unitarily invariant norm with gauge Φ.

(a) Prove that $\min\{|||\mathbf{A}-\mathbf{B}||| : \mathbf{B} \in M_n(\mathbb{C}) \text{ singular}\} = \Phi(\sigma_n, 0, \dots, 0)$.

(b) Deduce that this minimum equals $\sigma_n\,\Phi(\mathbf{e}_1)$, and identify it for the spectral, Frobenius, and trace norms.

Hint: a matrix is singular exactly when its rank is at most $n-1$.

Solution.

(a) By Theorem 7.5.3 (d), $\mathbf{B}$ is singular if and only if $\operatorname{rank}\mathbf{B} \le n-1$. So the minimum is over the set appearing in Corollary 21.6.2 with $k = n-1$, and that corollary gives the value $\Phi(\sigma_n, 0, \dots, 0)$, attained at the truncation $\mathbf{A}_{n-1}$, which is singular because its rank is $n-1$.

(b) The vector $(\sigma_n, 0, \dots, 0)$ is $\sigma_n\mathbf{e}_1$, and Φ is a norm, so $\Phi(\sigma_n\mathbf{e}_1) = \sigma_n\Phi(\mathbf{e}_1)$ as $\sigma_n > 0$. For the spectral norm $\Phi = \|\cdot\|_\infty$ and $\Phi(\mathbf{e}_1) = 1$; for the Frobenius norm $\Phi = \|\cdot\|_2$ and $\Phi(\mathbf{e}_1) = 1$; for the trace norm $\Phi = \|\cdot\|_1$ and $\Phi(\mathbf{e}_1) = 1$. All three give σ_n. This is not an accident: the normalization $\Phi(\mathbf{e}_1) = 1$ of Section 4 is exactly the condition that makes

every unitarily invariant norm agree on rank-one matrices with one nonzero singular value.

Exercise. C3: Unitary matrices are far apart

Let $\mathbf{U}, \mathbf{Q} \in M_n(\mathbb{C})$ be unitary with $\mathbf{U} \neq \mathbf{Q}$.

(a) Prove that $\sigma(\mathbf{U} - \mathbf{Q}) = (|\mu_1 - 1|, \dots, |\mu_n - 1|)^{\downarrow}$, where $\mu_1, \dots, \mu_n$ are the eigenvalues of $\mathbf{Q}^*\mathbf{U}$.

(b) Deduce that $\|\mathbf{U} - \mathbf{Q}\| = \|\operatorname{diag}(\mu_1 - 1, \dots, \mu_n - 1)\|$ for every unitarily invariant norm, and that the trace-norm distance between two distinct unitary matrices can be as large as $2n$.

Solution.

(a) $\mathbf{U} - \mathbf{Q} = \mathbf{Q}(\mathbf{Q}^*\mathbf{U} - \mathbf{I})$, and multiplying on the left by a unitary matrix does not change singular values, since $(\mathbf{QM})^*(\mathbf{QM}) = \mathbf{M}^*\mathbf{M}$ (Definition 13.9.1). The matrix $\mathbf{Z} := \mathbf{Q}^*\mathbf{U}$ is unitary, hence normal, so by the spectral theorem $\mathbf{Z} = \mathbf{Y} \operatorname{diag}(\mu_1, \dots, \mu_n)\mathbf{Y}^*$ with $\mathbf{Y}$ unitary, and $\mathbf{Z} - \mathbf{I} = \mathbf{Y} \operatorname{diag}(\mu_i - 1)\mathbf{Y}^*$. Then $(\mathbf{Z} - \mathbf{I})^*(\mathbf{Z} - \mathbf{I}) = \mathbf{Y} \operatorname{diag}(|\mu_i - 1|^2)\mathbf{Y}^*$, so the singular values are the numbers $|\mu_i - 1|$ sorted decreasingly (Definition 13.9.1).

(b) Both matrices in the claimed equality have the same singular value list, so Theorem 21.4.3 gives equal norms. Each μ_i has modulus 1, so $|\mu_i - 1| \leq 2$, with equality when $\mu_i = -1$. Taking $\mathbf{Q} = \mathbf{I}$ and $\mathbf{U} = -\mathbf{I}$ gives every $\mu_i = -1$ and trace-norm distance $\sum_i 2 = 2n$.

21.7 Lidskii's Theorem in Majorization Form

Chapter 17 §07 proved Lidskii's inequality and then translated the whole family of inequalities into one majorization. It could go no further, and said so. This section spends what it banked: the majorization is already in hand, and Sections 2, 3, and 6 turn it into a statement about every convex function and every unitarily invariant norm. The last part of the section carries all of it across to singular values.

21.7.1 What Chapter 17 proved, and what it left

Here is the statement we start from. For Hermitian $\mathbf{A}, \mathbf{B} \in M_n(F)$, Theorem 17.7.2 says that for every index set $1 \leq i_1 < \cdots < i_k \leq n$,

$$\sum_{j=1}^{k} \lambda_{i_j}(\mathbf{A}) - \sum_{j=1}^{k} \lambda_{i_j}(\mathbf{B}) \;\leq\; \sum_{j=1}^{k} \lambda_j(\mathbf{A} - \mathbf{B}),$$

and the remark that closes that section observes that, since the largest k entries of a vector are the best choice of k of them, this family of inequalities is exactly clause (M1) of Definition 17.8.1 for the pair of vectors below, while clause (M2) is the identity $\operatorname{tr}\mathbf{A} - \operatorname{tr}\mathbf{B} = \operatorname{tr}(\mathbf{A} - \mathbf{B})$:

$$\boldsymbol{\lambda}(\mathbf{A}) - \boldsymbol{\lambda}(\mathbf{B}) \;\prec\; \boldsymbol{\lambda}(\mathbf{A} - \mathbf{B}). \tag{21.7.1}$$

This is proved, in Chapter 17, and we do not prove it again. What Chapter 17 could not do it named exactly: "Not proved here, and left to Chapter 21: the general consequences of a majorization $\mathbf{x} \prec \mathbf{y}$, namely $\sum_i \phi(x_i) \le \sum_i \phi(y_i)$ for every convex $\phi\colon \mathbb{R} \to \mathbb{R}$, together with the matching statement for the unitarily invariant norms of that chapter."

Both consequences now exist. Section 2 proved the first, as Theorem 21.2.1. Section 3 proved that a symmetric gauge is monotone under majorization, as Theorem 21.3.4, and Section 4 identified the gauges with the unitarily invariant norms. All that is left is to apply them, which takes two short proofs, and then to repeat the exercise for singular values, which takes one idea.

Example. A pair to keep in mind

Let

$$\mathbf{A} = \begin{pmatrix} 2 & 2 \\ 2 & 2 \end{pmatrix}, \qquad \mathbf{B} = \begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix}.$$

Compute the three eigenvalue lists and check Eq. 21.7.1.

Solution. $\mathbf{A}$ is symmetric with $\operatorname{tr}\mathbf{A} = 4$ and $\det\mathbf{A} = 0$, so $\boldsymbol{\lambda}(\mathbf{A}) = (4, 0)$. $\mathbf{B}$ is symmetric with $\operatorname{tr}\mathbf{B} = 0$ and $\det\mathbf{B} = -1$, so $\boldsymbol{\lambda}(\mathbf{B}) = (1, -1)$. Their difference

$$\mathbf{A} - \mathbf{B} = \begin{pmatrix} 2 & 3 \\ 3 & 2 \end{pmatrix}$$

has trace 4 and determinant $4 - 9 = -5$, so $\boldsymbol{\lambda}(\mathbf{A} - \mathbf{B}) = (5, -1)$.

The vector of shifts is $\boldsymbol{\lambda}(\mathbf{A}) - \boldsymbol{\lambda}(\mathbf{B}) = (4 - 1,\ 0 - (-1)) = (3, 1)$, already decreasing. Now compare with $(5, -1)$: the top sums are $3 \le 5$, and the totals are $4 = 4$. So $(3, 1) \prec (5, -1)$, as Eq. 21.7.1 promises, and the majorization is strict in its one nontrivial inequality.

21.7.2 Every convex function at once

Theorem 21.7.1. Lidskii's Theorem, Convex Form

Let $\mathbf{A}, \mathbf{B} \in M_n(F)$ be Hermitian, $F = \mathbb{R}$ or $\mathbb{C}$. Then for every convex $\phi\colon \mathbb{R} \to \mathbb{R}$,

$$\sum_{i=1}^{n} \phi(\lambda_i(\mathbf{A}) - \lambda_i(\mathbf{B})) \ \le \ \sum_{i=1}^{n} \phi(\lambda_i(\mathbf{A} - \mathbf{B})).$$

Proof. By Eq. 21.7.1, $\boldsymbol{\lambda}(\mathbf{A}) - \boldsymbol{\lambda}(\mathbf{B}) \prec \boldsymbol{\lambda}(\mathbf{A} - \mathbf{B})$. Apply Theorem 21.2.1 to this majorization, with $I = \mathbb{R}$ and the given ϕ. This proves the theorem. ∎

Three readings are worth having. With $\phi(t) = |t|$, which is convex, the theorem says that the total movement of the eigenvalues is at most the total size of the eigenvalues of the perturbation. With $\phi(t) = t^2$ it is the Hoffman–Wielandt inequality, as the next corollary spells out. With $\phi(t) = \max(t, 0)$ it says that the total **upward** movement of the eigenvalues is at most the total

of the positive eigenvalues of the perturbation, and with $\phi(t) = e^{ct}$ it gives bounds no earlier chapter could have stated.

Warning. The left side pairs $\lambda_i(\mathbf{A})$ with $\lambda_i(\mathbf{B})$ at the **same index**, both lists being in decreasing order, and the theorem claims nothing about any other pairing. For $\mathbf{A} = \operatorname{diag}(4, 0)$ and $\mathbf{B} = \left(\begin{smallmatrix} 2 & 2 \\ 2 & 2 \end{smallmatrix}\right)$ the spectra are equal, $\boldsymbol{\lambda}(\mathbf{A}) = \boldsymbol{\lambda}(\mathbf{B}) = (4, 0)$, so the sorted pairing gives 0 with $\phi(t) = t^2$. The reversed pairing gives $(4-0)^2 + (0-4)^2 = 32$, while $\mathbf{A} - \mathbf{B} = \left(\begin{smallmatrix} 2 & -2 \\ -2 & -2 \end{smallmatrix}\right)$ has $\sum_i \lambda_i(\mathbf{A}-\mathbf{B})^2 = \|\mathbf{A}-\mathbf{B}\|_F^2 = 16$. So the same inequality with a different matching is simply false.

21.7.3 Every unitarily invariant norm at once

To state the norm version we need to put a vector of shifts inside a matrix, and Section 4 has already done it. By Lemma 21.4.1 the gauge of a unitarily invariant norm is recovered from the norm by

$$\left\|\operatorname{diag}_{m,n}(\mathbf{x})\right\| = \Phi(\mathbf{x}) \qquad (\mathbf{x} \in \mathbb{R}^p), \tag{21.7.2}$$

where $\operatorname{diag}_{m,n}(\mathbf{x})$ carries $\mathbf{x}$ down the main diagonal and zeros elsewhere, and is written $\operatorname{diag}(\mathbf{x})$ when $m = n$. One more reading will be used throughout: if $\mathbf{H} \in M_n(F)$ is Hermitian then $\mathbf{H}^*\mathbf{H} = \mathbf{H}^2$ has eigenvalues $\lambda_i(\mathbf{H})^2$, so $\sigma(\mathbf{H}) = |\boldsymbol{\lambda}(\mathbf{H})|^{\downarrow}$ by Definition 13.9.1, and hence $\|\mathbf{H}\| = \Phi(\sigma(\mathbf{H})) = \Phi(\boldsymbol{\lambda}(\mathbf{H}))$, the last step by the permutation and sign invariance of Definition 21.3.1.

Theorem 21.7.2. Lidskii's Theorem, Norm Form

Let $\mathbf{A}, \mathbf{B} \in M_n(F)$ be Hermitian. Then for **every** unitarily invariant norm on $M_n(F)$,

$$\left\|\operatorname{diag}(\boldsymbol{\lambda}(\mathbf{A}) - \boldsymbol{\lambda}(\mathbf{B}))\right\| \;\le\; \|\mathbf{A} - \mathbf{B}\|.$$

Idea. Both sides are the same symmetric gauge evaluated at two real vectors: the vector of eigenvalue shifts on the left, the eigenvalue vector of $\mathbf{A} - \mathbf{B}$ on the right. Chapter 17 says the first is majorized by the second, and Section 3 says a symmetric gauge respects that.

Proof. Let Φ be the symmetric gauge of the norm, so that $\|\mathbf{M}\| = \Phi(\sigma(\mathbf{M}))$ (Theorem 21.4.3). Put $\mathbf{x} = \boldsymbol{\lambda}(\mathbf{A}) - \boldsymbol{\lambda}(\mathbf{B})$ and $\mathbf{y} = \boldsymbol{\lambda}(\mathbf{A} - \mathbf{B})$, both in $\mathbb{R}^n$. By Eq. 21.7.2 the left-hand side is $\Phi(\mathbf{x})$; and since $\mathbf{A} - \mathbf{B}$ is Hermitian, the discussion above gives $\|\mathbf{A} - \mathbf{B}\| = \Phi(\mathbf{y})$. By Eq. 21.7.1, $\mathbf{x} \prec \mathbf{y}$, so Theorem 21.3.4 (a) — the part for arbitrary real vectors, which is what majorization rather than weak majorization buys — gives $\Phi(\mathbf{x}) \le \Phi(\mathbf{y})$. This proves the theorem. ∎

That is the second half of Chapter 17 §07's promise, and it is also what Chapter 20 §11 was waiting for when it listed, among the things that chapter did not do, that its bounds "hold in the Frobenius norm" and that "extending the eigenvalue and singular value bounds to every

unitarily invariant norm belongs to Chapter 21". Two specializations make the point. With the spectral norm, whose gauge is $\|\cdot\|_\infty$, the theorem reads $\max_i|\lambda_i(\mathbf{A})-\lambda_i(\mathbf{B})| \le \|\mathbf{A}-\mathbf{B}\|_2$, which is Weyl's perturbation bound Corollary 17.3.6. With the Frobenius norm, whose gauge is $\|\cdot\|_2$, it is Hoffman–Wielandt. One statement, two classical corollaries, and every norm in between.

Corollary 21.7.3. Hoffman–Wielandt for Hermitian Matrices, Again

Let $\mathbf{A},\mathbf{B}\in M_n(F)$ be Hermitian. Then

$$\sum_{i=1}^{n}\big(\lambda_i(\mathbf{A})-\lambda_i(\mathbf{B})\big)^2 \;\le\; \|\mathbf{A}-\mathbf{B}\|_F^2.$$

Proof. The function $\phi(t)=t^2$ is convex, by Theorem 18.10.6 (a) with $\phi''=2>0$. By Theorem 21.7.1,

$$\sum_i\big(\lambda_i(\mathbf{A})-\lambda_i(\mathbf{B})\big)^2 \le \sum_i \lambda_i(\mathbf{A}-\mathbf{B})^2.$$

The matrix $\mathbf{E}=\mathbf{A}-\mathbf{B}$ is Hermitian, so the spectral theorem writes $\mathbf{E}=\mathbf{U}\,\mathrm{diag}(\boldsymbol{\lambda}(\mathbf{E}))\mathbf{U}^*$ with $\mathbf{U}$ unitary, and Lemma 11.7.5 gives $\|\mathbf{E}\|_F^2=\|\mathrm{diag}(\boldsymbol{\lambda}(\mathbf{E}))\|_F^2=\sum_i\lambda_i(\mathbf{E})^2$. This proves the corollary. ∎

This is Corollary 20.5.4 of Chapter 20 §05, reached by a different road. That section said so in advance: "Chapter 21 shows that a majorization $\mathbf{x}\prec\mathbf{y}$ gives $\sum_i\phi(x_i)\le\sum_i\phi(y_i)$ for every convex ϕ; with $\phi(t)=t^2$ this is Corollary 20.5.4 again, and Chapter 21 turns the same majorization into one bound for every unitarily invariant norm. Nothing here depends on that route; the proof above goes through Birkhoff's theorem . . . " So the two proofs are independent: Chapter 20 argued from Birkhoff's theorem applied to the matrix of squared moduli of the entries of a unitary change of basis, and never used Lidskii's inequality; this one uses Lidskii and never mentions doubly stochastic matrices. Neither is a special case of the other, and Chapter 20 reaches further in one direction: its Corollary 20.5.4 is the Hermitian case of Theorem 20.5.2, which holds for all **normal** $\mathbf{A},\mathbf{B}$. What that theorem gives there is weaker in kind — some permutation matching the two eigenvalue lists, not the sorted one, since complex eigenvalues have no decreasing order — and the argument here, resting on Eq. 21.7.1, stays with Hermitian matrices throughout.

Quick check. Check Theorem 21.7.2 and Corollary 21.7.3 for the pair of Example (A pair to keep in mind), in the spectral, Frobenius, and trace norms.

Solution. There $\boldsymbol{\lambda}(\mathbf{A})-\boldsymbol{\lambda}(\mathbf{B})=(3,1)$ and $\boldsymbol{\lambda}(\mathbf{A}-\mathbf{B})=(5,-1)$, so $\mathrm{diag}(3,1)$ has singular values $(3,1)$ and $\mathbf{A}-\mathbf{B}$ has singular values $(5,1)$. The three norms of the left and right sides are: spectral, $3\le5$; Frobenius, $\sqrt{10}\le\sqrt{26}$; trace, $4\le6$. The Frobenius case squared is $9+1=10\le26$, which is Corollary 21.7.3.

21.7.4 The same theorem for singular values

Chapter 20 §10 proved Mirsky's theorem in the Frobenius norm and closed with: "Chapter 21 puts both inside one statement covering every unitarily invariant norm", the two being the index-by-index bound Corollary 17.9.5 and the Frobenius bound Theorem 20.10.1. The route is the one that section used and Section 6 used again: **a Hermitian dilation turns singular values into eigenvalues**, and then Lidskii's theorem applies. The only delicate point is the bookkeeping of the zeros, because a rectangular matrix dilates to a matrix with extra zero eigenvalues in the middle of its spectrum.

Theorem 21.7.4. Mirsky's Theorem in Every Unitarily Invariant Norm

Let $\mathbf{A}, \mathbf{B} \in M_{m\times n}(\mathbb{C})$ and $p = \min(m, n)$. Then for **every** unitarily invariant norm on $M_{m\times n}(\mathbb{C})$,

$$\left\| \operatorname{diag}_{m,n}(\sigma(\mathbf{A}) - \sigma(\mathbf{B})) \right\| \;\le\; |||\mathbf{A} - \mathbf{B}|||.$$

Idea. Dilate all three matrices. The eigenvalue list of $\mathcal{H}(\mathbf{M})$ is $\sigma(\mathbf{M})$, then $m+n-2p$ zeros, then $-\sigma(\mathbf{M})$ reversed, and the **positions** of the three blocks depend only on m and n — so $\mathbf{A}$ and $\mathbf{B}$ have their singular values in the same slots, and subtracting the two lists subtracts them index by index. Lidskii's majorization then compares a symmetric list $\pm(\sigma_i(\mathbf{A})-\sigma_i(\mathbf{B}))$ with a symmetric list $\pm\sigma_j(\mathbf{A}-\mathbf{B})$ in $\mathbb{R}^{m+n}$. Reading its top-k sums for $k \le p$ turns it into a weak majorization of the two lists we care about, in $\mathbb{R}^p$, and Section 3 finishes.

Proof. Write $N = m + n$, let $\delta_i = \sigma_i(\mathbf{A}) - \sigma_i(\mathbf{B})$ for $i \le p$ and $\boldsymbol{\delta} = (\delta_1, \dots, \delta_p)$, and let Φ be the symmetric gauge of the norm.

Step 1: what has to be proved. By Eq. 21.7.2 and the sign invariance of Definition 21.3.1, $\left\| \operatorname{diag}_{m,n}(\boldsymbol{\delta}) \right\| = \Phi(\boldsymbol{\delta}) = \Phi(|\boldsymbol{\delta}|)$; and $|||\mathbf{A} - \mathbf{B}||| = \Phi(\sigma(\mathbf{A} - \mathbf{B}))$ by Theorem 21.4.3. Both $|\boldsymbol{\delta}|$ and $\sigma(\mathbf{A} - \mathbf{B})$ have nonnegative entries, so by Theorem 21.3.4 (b) it is enough to prove the weak majorization

$$|\boldsymbol{\delta}| \;\prec_w\; \sigma(\mathbf{A} - \mathbf{B}) \qquad \text{in } \mathbb{R}^p. \tag{21.7.3}$$

Step 2: the dilations and their spectra. For $\mathbf{M} \in M_{m\times n}(\mathbb{C})$ let $\mathcal{H}(\mathbf{M}) = \left(\begin{smallmatrix} \mathbf{O} & \mathbf{M} \\ \mathbf{M}^* & \mathbf{O} \end{smallmatrix}\right) \in M_N(\mathbb{C})$ be the Hermitian dilation of Proposition 17.9.3. Conjugate transposition is additive, so

$$\mathcal{H}(\mathbf{A}) - \mathcal{H}(\mathbf{B}) = \mathcal{H}(\mathbf{A} - \mathbf{B}),$$

and all three matrices are Hermitian. By Proposition 17.9.3 the eigenvalues of $\mathcal{H}(\mathbf{M})$ are $\pm\sigma_1(\mathbf{M}), \dots, \pm\sigma_p(\mathbf{M})$ together with $N - 2p$ zeros; since the $\sigma_i(\mathbf{M})$ are nonnegative and decreasing, the decreasing list is

$$\lambda_i(\mathcal{H}(\mathbf{M})) = \begin{cases} \sigma_i(\mathbf{M}) & 1 \le i \le p, \\ 0 & p < i \le N - p, \\ -\sigma_{N+1-i}(\mathbf{M}) & N - p < i \le N. \end{cases} \tag{21.7.4}$$

The three ranges depend only on m and n, so the formula applies to $\mathbf{A}$, to $\mathbf{B}$, and to $\mathbf{A} - \mathbf{B}$

with the same ranges. This is the bookkeeping the rectangular case needs: the zeros of $\mathcal{H}(\mathbf{A})$ sit in exactly the positions of the zeros of $\mathcal{H}(\mathbf{B})$, and so cancel when the two lists are subtracted.

Step 3: Lidskii. By Eq. 21.7.1 applied to the Hermitian pair $\mathcal{H}(\mathbf{A})$, $\mathcal{H}(\mathbf{B})$,

$$\boldsymbol{\lambda}(\mathcal{H}(\mathbf{A})) - \boldsymbol{\lambda}(\mathcal{H}(\mathbf{B})) \;\prec\; \boldsymbol{\lambda}(\mathcal{H}(\mathbf{A}-\mathbf{B})).$$

By Eq. 21.7.4 the vector on the left has entries δ_i in positions $i \le p$, zeros in the middle range, and $-\delta_{N+1-i}$ in the last p positions; the vector on the right has $\sigma_j(\mathbf{A}-\mathbf{B})$, zeros, and the negatives of the $\sigma_j(\mathbf{A}-\mathbf{B})$ in the same three ranges.

Step 4: read the running totals. Fix $k \le p$. The left-hand vector is the symmetric list of Lemma 21.6.3 with $q = p$, $r = N - 2p$, and $z_i = \delta_i$, so by that lemma its k-th running total — the largest sum of k of its entries — is $\sum_{j\le k} |\boldsymbol{\delta}|_j^{\downarrow}$. The right-hand vector is already decreasing, so its k-th running total is $\sum_{j\le k}\sigma_j(\mathbf{A}-\mathbf{B})$. Since $k \le p < N$, clause (M1) of Definition 17.8.1 applies at this k and gives

$$\sum_{j=1}^{k} |\boldsymbol{\delta}|_j^{\downarrow} \;\le\; \sum_{j=1}^{k} \sigma_j(\mathbf{A}-\mathbf{B}).$$

As $k \le p$ was arbitrary, this is Eq. 21.7.3, and Step 1 finishes the proof.

■

Remark. **Where the factor of two went.** Chapter 20 §10 proved the Frobenius case by taking a **norm** of the dilation, and paid for it: $\|\mathcal{H}(\mathbf{M})\|_F^2 = 2\|\mathbf{M}\|_F^2$, so a factor 2 appeared on each side and canceled. That cancellation is special to a norm which adds squares, and for a general unitarily invariant norm there is no such identity — indeed $|||\mathcal{H}(\mathbf{M})|||$ is a norm on a different matrix space, computed by a gauge on $\mathbb{R}^N$ rather than on $\mathbb{R}^p$. The proof above never takes a norm of a dilation. It uses the dilation only to obtain a **majorization of eigenvalue lists**, and then reads off the p running totals it needs. No factor of 2 arises, and none has to cancel.

Specializing the gauge recovers both statements that Chapter 20 §10 wanted inside one: $\Phi = \|\cdot\|_\infty$ gives $\max_i|\sigma_i(\mathbf{A}) - \sigma_i(\mathbf{B})| \le \|\mathbf{A}-\mathbf{B}\|_2$, which is Corollary 17.9.5, and $\Phi = \|\cdot\|_2$ gives Theorem 20.10.1. Chapter 20 §11's ledger entry — "extending the eigenvalue and singular value bounds to every unitarily invariant norm belongs to Chapter 21" — is now paid for the eigenvalue bounds by Theorem 21.7.2 and for the singular value bounds by Theorem 21.7.4. Chapter 20 §10's square-root and polar bounds (Theorem 20.10.2, Theorem 20.10.3, Theorem 20.10.4) are **not** extended here: they still hold only in the two norms Chapter 20 proved them in, the spectral and the Frobenius.

The pair of Example (A pair to keep in mind) illustrates the singular value form as well. There $\mathbf{A} = \left(\begin{smallmatrix} 2 & 2 \\ 2 & 2 \end{smallmatrix}\right)$ is positive semidefinite with eigenvalues $4, 0$, so $\sigma(\mathbf{A}) = (4, 0)$; $\mathbf{B}$ has eigenvalues ± 1, so $\sigma(\mathbf{B}) = (1, 1)$; and $\mathbf{A}-\mathbf{B}$ has eigenvalues $5, -1$, so $\sigma(\mathbf{A}-\mathbf{B}) = (5, 1)$. The vector of differences is $\sigma(\mathbf{A}) - \sigma(\mathbf{B}) = (3, -1)$, whose absolute values $(3, 1)$ have running totals $3 \le 5$ and $4 \le 6$: the weak majorization of Eq. 21.7.3, and with it the norm inequality in every unitarily invariant norm.

Warning. Both Theorem 21.7.2 and Theorem 21.7.4 compare **sorted** lists, and the matching they use is the sorted one. No claim is made for any other pairing, and Theorem 21.7.4 in particular is not a statement about eigenvalues: for non-Hermitian matrices the eigenvalues can be wildly unstable while the singular values are not, which is the whole content of Chapter 20 §10's remark that Mirsky's theorem needs no normality hypothesis at all.

21.7.5 Exercises

21.7.5.1 A. Check your understanding

Exercise. A1

(a) State the majorization Eq. 21.7.1, and say which chapter proves it.

(b) State Theorem 21.7.2, and say what $\operatorname{diag}(\boldsymbol{\lambda}(\mathbf{A}) - \boldsymbol{\lambda}(\mathbf{B}))$ means.

(c) Determine whether the following is true for Hermitian $\mathbf{A}, \mathbf{B}$, and justify your answer: $\sum_i |\lambda_i(\mathbf{A}) - \lambda_i(\mathbf{B})| \le \sum_i |\lambda_i(\mathbf{A} - \mathbf{B})|$.

(d) Determine whether the following is true, and justify your answer: Theorem 21.7.4 holds with the singular values of $\mathbf{A}$ and $\mathbf{B}$ listed in **increasing** order on the left.

(e) Where in the proof of Theorem 21.7.4 is it used that $\mathbf{A}$ and $\mathbf{B}$ have the same size?

Solution.

(a) For Hermitian $\mathbf{A}, \mathbf{B} \in M_n(F)$, $\boldsymbol{\lambda}(\mathbf{A}) - \boldsymbol{\lambda}(\mathbf{B}) \prec \boldsymbol{\lambda}(\mathbf{A} - \mathbf{B})$. It is proved in Chapter 17 §07, from Theorem 17.7.2 over all index sets together with the equality of traces.

(b) For every unitarily invariant norm, $|||\operatorname{diag}(\boldsymbol{\lambda}(\mathbf{A}) - \boldsymbol{\lambda}(\mathbf{B}))||| \le |||\mathbf{A} - \mathbf{B}|||$. The matrix $\operatorname{diag}(\mathbf{x})$ is the diagonal matrix in $M_n(F)$ whose i-th diagonal entry is x_i; here $x_i = \lambda_i(\mathbf{A}) - \lambda_i(\mathbf{B})$, the shift of the i-th eigenvalue, the lists being in decreasing order.

(c) True. Take $\phi(t) = |t|$ in Theorem 21.7.1; it is convex, being a norm on $\mathbb{R}$. Alternatively it is Theorem 21.7.2 for the trace norm, using Eq. 21.7.2.

(d) True, and for a cheap reason: reversing both lists changes the vector $\boldsymbol{\delta}$ only by reversing it, and the gauge of a unitarily invariant norm is permutation-invariant (Definition 21.3.1), so the left-hand side is unchanged. What would **not** be true is to sort one list increasingly and the other decreasingly; see the warning after Theorem 21.7.1.

(e) Twice. The two matrices must lie in the same space for $\mathbf{A} - \mathbf{B}$ to be defined and for one norm to apply to all three matrices; and in Step 2, m, n, and p being the same for $\mathbf{A}$ and $\mathbf{B}$ is what puts the zeros of the two dilated spectra in the same positions, so that the difference of the two lists has the shape used in Step 4.

21.7.5.2 B. Practice

Exercise. B1: Running the three consequences

Let $\mathbf{A} = \operatorname{diag}(5, 1, 0)$ and let $\mathbf{B} = \mathbf{A} - \mathbf{E}$ with

$$\mathbf{E} = \begin{pmatrix} 0 & 0 & 2 \\ 0 & 0 & 0 \\ 2 & 0 & 0 \end{pmatrix}.$$

(a) Compute $\boldsymbol{\lambda}(\mathbf{B})$ and $\boldsymbol{\lambda}(\mathbf{E})$.

(b) Verify Eq. 21.7.1 for this pair.

(c) Verify Theorem 21.7.2 in the spectral, Frobenius, and trace norms, and Corollary 21.7.3.

Solution.

(a) $\mathbf{E}$ acts as $\left(\begin{smallmatrix} 0 & 2 \\ 2 & 0 \end{smallmatrix}\right)$ on $\operatorname{span}(\mathbf{e}_1, \mathbf{e}_3)$ and as 0 on $\mathbf{e}_2$, so $\boldsymbol{\lambda}(\mathbf{E}) = (2, 0, -2)$. Next,

$$\mathbf{B} = \mathbf{A} - \mathbf{E} = \begin{pmatrix} 5 & 0 & -2 \\ 0 & 1 & 0 \\ -2 & 0 & 0 \end{pmatrix}.$$

On $\operatorname{span}(\mathbf{e}_1, \mathbf{e}_3)$ it is $\left(\begin{smallmatrix} 5 & -2 \\ -2 & 0 \end{smallmatrix}\right)$, with trace 5 and determinant -4, so its eigenvalues there are $(5 \pm \sqrt{41})/2$; together with the eigenvalue 1 from $\mathbf{e}_2$,

$$\boldsymbol{\lambda}(\mathbf{B}) = \left(\tfrac{5+\sqrt{41}}{2},\ 1,\ \tfrac{5-\sqrt{41}}{2}\right) \approx (5.702,\ 1,\ -0.702).$$

(b) The shifts are

$$\begin{aligned} \boldsymbol{\lambda}(\mathbf{A}) - \boldsymbol{\lambda}(\mathbf{B}) &= (5 - \tfrac{5+\sqrt{41}}{2},\ 0,\ \tfrac{\sqrt{41}-5}{2}) \\ &= (\tfrac{5-\sqrt{41}}{2},\ 0,\ \tfrac{\sqrt{41}-5}{2}) \approx (-0.702,\ 0,\ 0.702). \end{aligned}$$

Sorted decreasingly this is $(0.702, 0, -0.702)$, with running totals $0.702,\ 0.702,\ 0$. The list $\boldsymbol{\lambda}(\mathbf{E}) = (2, 0, -2)$ has running totals $2, 2, 0$. Each of the first two is larger, and the totals agree at 0. So the majorization holds.

(c) Write $c = \frac{\sqrt{41}-5}{2} \approx 0.702$. The left-hand matrix is $\operatorname{diag}(-c, 0, c)$, whose singular values are $(c, c, 0)$; the right-hand matrix is $\mathbf{E}$, whose singular values are $(2, 2, 0)$. Spectral: $c \le 2$. Frobenius: $c\sqrt{2} \approx 0.993 \le 2\sqrt{2} \approx 2.828$. Trace: $2c \approx 1.403 \le 4$. Squaring the Frobenius comparison gives

$$2c^2 = \tfrac{(\sqrt{41}-5)^2}{2} = \tfrac{66-10\sqrt{41}}{2} = 33 - 5\sqrt{41} \approx 0.984$$

on the left and $\|\mathbf{E}\|_F^2 = 8$ on the right, which is Corollary 21.7.3.

Exercise. B2: Singular values of a one-entry change

Let $\mathbf{A} = \left(\begin{smallmatrix} 3 & 1 \\ 1 & 3 \end{smallmatrix}\right)$ and $\mathbf{B} = \left(\begin{smallmatrix} 3 & 1 \\ 0 & 3 \end{smallmatrix}\right)$. Compute both sides of Theorem 21.7.4 in the trace norm, and compare with what Corollary 17.9.5 gives on its own.

Solution. $\mathbf{A}$ is symmetric with eigenvalues 4 and 2, both positive, so $\sigma(\mathbf{A}) = (4, 2)$. For $\mathbf{B}$, $\mathbf{B}^\top\mathbf{B} = \left(\begin{smallmatrix} 9 & 3 \\ 3 & 10 \end{smallmatrix}\right)$, with trace 19 and determinant 81, so its eigenvalues are $\frac{19 \pm \sqrt{37}}{2}$; these are

the squares of $\frac{\sqrt{37}\pm1}{2}$, since $\left(\frac{\sqrt{37}\pm1}{2}\right)^2 = \frac{38\pm2\sqrt{37}}{4}$. Hence

$$\sigma(\mathbf{B}) = \left(\tfrac{\sqrt{37}+1}{2},\ \tfrac{\sqrt{37}-1}{2}\right) \approx (3.541,\ 2.541).$$

The differences are $\sigma(\mathbf{A})-\sigma(\mathbf{B}) = (\frac{7-\sqrt{37}}{2},\ \frac{5-\sqrt{37}}{2}) \approx (0.459,\ -0.541)$, so the left side in the trace norm is

$$\tfrac{7-\sqrt{37}}{2} + \tfrac{\sqrt{37}-5}{2} = 1.$$

On the right, $\mathbf{A}-\mathbf{B} = \left(\begin{smallmatrix}0&0\\1&0\end{smallmatrix}\right)$ has singular values $(1,0)$, so $|||\mathbf{A}-\mathbf{B}||| = 1$ in the trace norm. The inequality holds with **equality** here.

Corollary 17.9.5 gives only $|\sigma_i(\mathbf{A}) - \sigma_i(\mathbf{B})| \le \|\mathbf{A}-\mathbf{B}\|_2 = 1$ for each i separately, hence at best 2 for the sum. Mirsky's theorem halves that, and here it is exact.

Exercise. B3: A bound with no norm in sight

Let $\mathbf{A},\mathbf{B} \in M_n(\mathbb{C})$ be Hermitian with $\|\mathbf{A}-\mathbf{B}\|_2 \le \varepsilon$. Prove that $\sum_i |\lambda_i(\mathbf{A}) - \lambda_i(\mathbf{B})| \le n\varepsilon$, and give a pair for which this is an equality.

Solution. By Theorem 21.7.1 with $\phi(t) = |t|$,

$$\sum_i |\lambda_i(\mathbf{A}) - \lambda_i(\mathbf{B})| \le \sum_i |\lambda_i(\mathbf{A}-\mathbf{B})| \le n \max_i |\lambda_i(\mathbf{A}-\mathbf{B})| = n\|\mathbf{A}-\mathbf{B}\|_2 \le n\varepsilon,$$

the last equality by Lemma 17.3.5. For equality take $\mathbf{B} = \mathbf{A} - \varepsilon\mathbf{I}$: then $\lambda_i(\mathbf{B}) = \lambda_i(\mathbf{A}) - \varepsilon$ for every i, the left side is $n\varepsilon$, and $\|\mathbf{A}-\mathbf{B}\|_2 = \varepsilon$.

21.7.5.3 C. Going deeper

Exercise. C1: The Schatten norms

Recall the Schatten r-norm $|||\mathbf{M}|||_{S_r} = \left(\sum_i \sigma_i(\mathbf{M})^r\right)^{1/r}$ of Example (The Schatten norms), the unitarily invariant norm with gauge $\|\cdot\|_r$, and let $1 \le r < \infty$.

(a) Prove that for Hermitian $\mathbf{A},\mathbf{B} \in M_n(\mathbb{C})$, $\sum_i |\lambda_i(\mathbf{A}) - \lambda_i(\mathbf{B})|^r \le \sum_i |\lambda_i(\mathbf{A}-\mathbf{B})|^r$, in two ways: from Theorem 21.7.1 and from Theorem 21.7.2.

(b) Deduce the same statement for singular values, with σ in place of λ and $\mathbf{A},\mathbf{B}$ arbitrary of the same size.

Solution.

(a) *From Theorem 21.7.1.* First, $\phi(t) = |t|^r$ is convex on $\mathbb{R}$. For $r = 1$ this is $|t| = \max\{t,-t\}$, convex by Proposition 18.10.3. Let $r > 1$, let $a,b \in \mathbb{R}$, and let $0 < \theta < 1$ (the cases $\theta = 0,1$ are equalities); write $\theta' = 1-\theta$ and let $q = r/(r-1)$ be the exponent conjugate to r. Apply Theorem 18.11.4 in $\mathbb{R}^2$ to

$$\mathbf{x} = (\theta^{1/r}|a|,\ (\theta')^{1/r}|b|), \qquad \mathbf{y} = (\theta^{1/q},\ (\theta')^{1/q}).$$

Since $\frac{1}{r} + \frac{1}{q} = 1$, the rules for powers of Chapter 18 §11 give $\langle \mathbf{x},\mathbf{y}\rangle = \theta|a| + \theta'|b|$ and

$\|\mathbf{y}\|_q = (\theta+\theta')^{1/q} = 1$, so

$$\theta|a| + \theta'|b| \;\le\; \|\mathbf{x}\|_r = \left(\theta|a|^r + \theta'|b|^r\right)^{1/r}.$$

The function $u \mapsto u^r$ is increasing on $[0,\infty)$ (Chapter 18 §11), and $|\theta a+\theta' b| \le \theta|a|+\theta'|b|$ by Theorem 0.6.4, so raising the displayed inequality to the power r gives

$$|\theta a + \theta' b|^r \;\le\; \left(\theta|a| + \theta'|b|\right)^r \;\le\; \theta|a|^r + \theta'|b|^r,$$

which is Definition 18.10.1 for ϕ on $\mathbb{R}$. Now apply Theorem 21.7.1.

From Theorem 21.7.2. Take the Schatten r-norm, whose gauge is $\|\cdot\|_r$. The left-hand side of Theorem 21.7.2 is $\left(\sum_i |\lambda_i(\mathbf{A}) - \lambda_i(\mathbf{B})|^r\right)^{1/r}$ by Eq. 21.7.2, and the right-hand side is $\left(\sum_i |\lambda_i(\mathbf{A}-\mathbf{B})|^r\right)^{1/r}$, since $\mathbf{A}-\mathbf{B}$ is Hermitian. Raise to the power r.

(b) Apply Theorem 21.7.4 with the Schatten r-norm and raise to the power r, reading both sides through Theorem 21.4.3 as above.

Exercise. C2: What the dilation costs

Let $\mathbf{A}, \mathbf{B} \in M_{m\times n}(\mathbb{C})$ and let $\mathcal{H}(\cdot)$ be the Hermitian dilation of Proposition 17.9.3.

(a) Prove that $\|\mathcal{H}(\mathbf{A})\|$ and $\|\mathbf{A}\|$ cannot be compared in general, by computing both for the Frobenius norm and for the spectral norm with $\mathbf{A} = \mathbf{I}_2$.

(b) Explain, in one or two sentences, why the proof of Theorem 21.7.4 is nevertheless free of any such factor.

(c) Deduce Corollary 21.6.2 from Theorem 21.7.4, and say why Section 6 nevertheless proved it another way.

Hint: for (c), a matrix of rank at most k has $\sigma_i = 0$ for $i > k$.

Solution.

(a) For $\mathbf{A} = \mathbf{I}_2$, $\mathcal{H}(\mathbf{A}) \in M_4(\mathbb{C})$ has singular values $(1,1,1,1)$, so $\|\mathcal{H}(\mathbf{A})\|_F = 2$ while $\|\mathbf{A}\|_F = \sqrt{2}$: the Frobenius norm grows by $\sqrt{2}$. The spectral norms, however, are both 1. So the ratio $\|\mathcal{H}(\mathbf{A})\| \,/\, \|\mathbf{A}\|$ depends on the norm, and there is no single factor to divide out. (The two norms also live on different spaces, M_4 and M_2.)

(b) Because the dilation is used only to produce a majorization between **eigenvalue lists**, from which the required inequality between running totals of the p numbers $|\delta_i|$ and the p numbers $\sigma_j(\mathbf{A}-\mathbf{B})$ is read off directly. No norm is ever applied to a dilated matrix.

(c) Let $\operatorname{rank}\mathbf{B} \le k$, so $\sigma_i(\mathbf{B}) = 0$ for $i > k$ and hence $|\delta_i| = \sigma_i(\mathbf{A})$ for $i > k$, in the notation of the proof of Theorem 21.7.4. Let $\mathbf{z} \in \mathbb{R}^p$ have $z_i = 0$ for $i \le k$ and $z_i = \sigma_i(\mathbf{A})$ for $i > k$. Then $|z_i| \le |\delta_i|$ for every i, so Proposition 21.3.1 gives $\Phi(\mathbf{z}) \le \Phi(\boldsymbol{\delta})$; and $\Phi(\mathbf{z}) = \Phi(\sigma_{k+1}, \dots, \sigma_p, 0, \dots, 0)$ by the permutation invariance of Definition 21.3.1. Combining with Theorem 21.7.4,

$$\Phi(\sigma_{k+1}, \dots, \sigma_p, 0, \dots, 0) \le \Phi(\boldsymbol{\delta}) = \left\|\operatorname{diag}_{m,n}(\boldsymbol{\delta})\right\| \le \|\mathbf{A}-\mathbf{B}\|,$$

which is the bound of Corollary 21.6.2, attained at $\mathbf{A}_k$. Section 6 proved it earlier in the chapter, so it could not use this section; and its route gives more, namely

$\sigma_i(\mathbf{A}-\mathbf{A}_k) \le \sigma_i(\mathbf{A}-\mathbf{B})$ index by index, which is the route Chapter 13 §11 predicted.

21.8 Operator Convex Functions

Chapter 17 §11 defined two notions side by side and then spent itself on one of them. Operator monotonicity got four families of examples, a warning, and a classification to look forward to; operator convexity got one proposition — that $t \mapsto t^2$ is operator convex (Proposition 17.11.8) — and then the chapter moved on. This section repays the debt. It collects the examples that are within reach, and it proves the one structural theorem the notion has: an inequality of Jensen's type that survives being squeezed by a contraction. That theorem is what makes operator convexity useful rather than merely curious.

Throughout, $F = \mathbb{R}$ **or** $F = \mathbb{C}$, every matrix whose spectrum is named is Hermitian, and I denotes an **interval** of real numbers. The letter I is plain italic and is not the identity matrix $\mathbf{I}$. For $f\colon I \to \mathbb{R}$ and Hermitian $\mathbf{A}$ with $\operatorname{spec}(\mathbf{A}) \subseteq I$, the matrix $f(\mathbf{A})$ is the one built by the functional calculus of Chapter 12 §08 (Definition 12.8.2), and it is Hermitian because f is real-valued (Theorem 12.8.3 (c)).

21.8.1 What operator convexity asks

Recall the definition. A function $f\colon I \to \mathbb{R}$ is **operator convex on** I if, for **every** size $n \ge 1$, all Hermitian $\mathbf{A}, \mathbf{B} \in M_n(F)$ with spectra in I, and every $\theta \in [0,1]$,

$$f(\theta\mathbf{A} + (1-\theta)\mathbf{B}) \;\preceq\; \theta f(\mathbf{A}) + (1-\theta) f(\mathbf{B})$$

in the Loewner order (Definition 17.11.1 (b), Definition 13.5.1). It is **operator concave** when $-f$ is operator convex. Chapter 17 §11 checked that the left-hand side makes sense: the spectrum of a convex combination of two Hermitian matrices with spectra in I again lies in I.

Taking $n = 1$ turns every matrix into a number and the Loewner order into $\ge$, so an operator convex function is convex in the sense of Definition 18.10.1. As with monotonicity, the content lies in the sizes $n \ge 2$, and the clause "for every n" is what makes the class small.

Two facts from Chapter 17 §11 are the starting stock. Affine functions are operator convex and operator concave at once, with equality in the definition (Proposition 17.11.2), and $t \mapsto t^2$ is operator convex on $\mathbb{R}$, because

$$\theta\mathbf{A}^2 + (1-\theta)\mathbf{B}^2 - \big(\theta\mathbf{A} + (1-\theta)\mathbf{B}\big)^2 = \theta(1-\theta)(\mathbf{A}-\mathbf{B})^2$$

and the square of a Hermitian matrix is positive semidefinite (Proposition 17.11.8).

Before adding to the stock we record three bookkeeping identities. Each says that a familiar operation on matrices passes through the functional calculus untouched, and each is used below without further comment.

Lemma 21.8.1. Three Ways to Move the Functional Calculus

Let $I \subseteq \mathbb{R}$ be an interval and $f\colon I \to \mathbb{R}$.

(a) **(Shift.)** Let $s \in \mathbb{R}$, let $g(t) = f(t+s)$ on $I - s$, and let $\mathbf{A}$ be Hermitian with $\operatorname{spec}(\mathbf{A}) \subseteq I - s$. Then $g(\mathbf{A}) = f(\mathbf{A} + s\mathbf{I})$.

(b) **(Unitary conjugation.)** Let $\mathbf{A}$ be Hermitian with $\operatorname{spec}(\mathbf{A}) \subseteq I$ and let $\mathbf{U}$ be unitary

of the same size. Then $\operatorname{spec}(\mathbf{U}^*\mathbf{A}\mathbf{U}) = \operatorname{spec}(\mathbf{A})$ and

$$f(\mathbf{U}^*\mathbf{A}\mathbf{U}) = \mathbf{U}^* f(\mathbf{A})\mathbf{U}.$$

(c) **(Direct sums.)** Let $\mathbf{A}, \mathbf{B}$ be Hermitian with spectra in I. Then $f(\mathbf{A} \oplus \mathbf{B}) = f(\mathbf{A}) \oplus f(\mathbf{B})$.

Proof. Write the spectral resolution $\mathbf{A} = \sum_j \mu_j \mathbf{P}_j$ with distinct μ_j (Definition 12.8.1, Theorem 12.8.1).

(a) The numbers $\mu_j + s$ are distinct, and $\mathbf{A} + s\mathbf{I} = \sum_j (\mu_j + s)\mathbf{P}_j$ by Theorem 12.8.1 (c) and (d). The $\mathbf{P}_j$ are nonzero self-adjoint idempotents, pairwise annihilating, and summing to $\mathbf{I}$ (Theorem 12.8.1 (b), (c)), so this is the spectral resolution of $\mathbf{A} + s\mathbf{I}$ by Theorem 12.8.2. Hence $f(\mathbf{A} + s\mathbf{I}) = \sum_j f(\mu_j + s)\mathbf{P}_j = \sum_j g(\mu_j)\mathbf{P}_j = g(\mathbf{A})$, using Definition 12.8.2 twice.

(b) Put $\mathbf{Q}_j = \mathbf{U}^*\mathbf{P}_j\mathbf{U}$. Each $\mathbf{Q}_j$ is self-adjoint, $\mathbf{Q}_j^2 = \mathbf{U}^*\mathbf{P}_j\mathbf{U}\mathbf{U}^*\mathbf{P}_j\mathbf{U} = \mathbf{Q}_j$, and $\mathbf{Q}_i\mathbf{Q}_j = \mathbf{O}$ for $i \neq j$, all because $\mathbf{U}\mathbf{U}^* = \mathbf{I}$; moreover $\sum_j \mathbf{Q}_j = \mathbf{U}^*\mathbf{I}\mathbf{U} = \mathbf{I}$ and $\mathbf{Q}_j \neq \mathbf{O}$ since $\mathbf{U}$ is invertible. Also $\mathbf{U}^*\mathbf{A}\mathbf{U} = \sum_j \mu_j \mathbf{Q}_j$. So this is the spectral resolution of $\mathbf{U}^*\mathbf{A}\mathbf{U}$ by Theorem 12.8.2, its distinct eigenvalues are the μ_j, and

$$f(\mathbf{U}^*\mathbf{A}\mathbf{U}) = \sum_j f(\mu_j)\mathbf{Q}_j = \mathbf{U}^*\Big(\sum_j f(\mu_j)\mathbf{P}_j\Big)\mathbf{U} = \mathbf{U}^* f(\mathbf{A})\mathbf{U}.$$

(c) Let $\nu_1, \dots, \nu_r$ be the distinct elements of $\operatorname{spec}(\mathbf{A}) \cup \operatorname{spec}(\mathbf{B})$. For each ν, let $\mathbf{P}_\nu^{\mathbf{A}}$ be the orthogonal projection onto $E_\nu(\mathbf{A})$, understood as $\mathbf{O}$ when $\nu \notin \operatorname{spec}(\mathbf{A})$, and likewise $\mathbf{P}_\nu^{\mathbf{B}}$. A vector $(\mathbf{x}, \mathbf{y})$ satisfies $(\mathbf{A} \oplus \mathbf{B})(\mathbf{x}, \mathbf{y}) = \nu(\mathbf{x}, \mathbf{y})$ exactly when $\mathbf{A}\mathbf{x} = \nu\mathbf{x}$ and $\mathbf{B}\mathbf{y} = \nu\mathbf{y}$, so $E_\nu(\mathbf{A} \oplus \mathbf{B}) = E_\nu(\mathbf{A}) \oplus E_\nu(\mathbf{B})$, whose orthogonal projection is $\mathbf{P}_\nu^{\mathbf{A}} \oplus \mathbf{P}_\nu^{\mathbf{B}}$. Hence $\operatorname{spec}(\mathbf{A} \oplus \mathbf{B}) = \{\nu_1, \dots, \nu_r\} \subseteq I$ and, by Definition 12.8.2,

$$f(\mathbf{A} \oplus \mathbf{B}) = \sum_\nu f(\nu)(\mathbf{P}_\nu^{\mathbf{A}} \oplus \mathbf{P}_\nu^{\mathbf{B}}) = f(\mathbf{A}) \oplus f(\mathbf{B}).$$

■

21.8.2 The examples we can prove

Operator convex functions form a cone, exactly as the operator monotone ones do, and for the same reason.

Proposition 21.8.2. Combining Operator Convex Functions

Let f and g be operator convex on I.

(a) For all $c \geq 0$ and $\alpha, \beta \in \mathbb{R}$, the functions $f + g$ and $\alpha + \beta t + cf(t)$ are operator convex on I.

(b) For every $s \in \mathbb{R}$, the translate $t \mapsto f(t + s)$ is operator convex on $I - s$.

Proof.

(a) Fix Hermitian $\mathbf{A}, \mathbf{B}$ of the same size with spectra in I, fix $\theta \in [0, 1]$ and put $\mathbf{C} =$

$\theta\mathbf{A} + (1-\theta)\mathbf{B}$. By Theorem 12.8.3 (b), the defect

$$D_h \coloneqq \theta h(\mathbf{A}) + (1-\theta)h(\mathbf{B}) - h(\mathbf{C})$$

satisfies $D_{f+g} = D_f + D_g$ and $D_{cf} = c\,D_f$. Both D_f and D_g are $\succeq \mathbf{O}$ by hypothesis, so for every $\mathbf{x}$ the quadratic forms add and scale to something ≥ 0, giving $D_{f+g} \succeq \mathbf{O}$ and $D_{cf} \succeq \mathbf{O}$. Adding an affine function changes nothing, since $D_h = \mathbf{O}$ for affine h by Proposition 17.11.2.

(b) Let $g(t) = f(t+s)$ and let $\mathbf{A}, \mathbf{B}$ be Hermitian with spectra in $I - s$. Then $\mathbf{A} + s\mathbf{I}$ and $\mathbf{B} + s\mathbf{I}$ have spectra in I, and $\theta(\mathbf{A} + s\mathbf{I}) + (1-\theta)(\mathbf{B} + s\mathbf{I}) = \mathbf{C} + s\mathbf{I}$. By Lemma 21.8.1 (a) applied three times,

$$g(\mathbf{C}) = f(\mathbf{C} + s\mathbf{I}) \preceq \theta f(\mathbf{A} + s\mathbf{I}) + (1-\theta) f(\mathbf{B} + s\mathbf{I}) = \theta g(\mathbf{A}) + (1-\theta) g(\mathbf{B}).$$

■

The new example is the reciprocal. Chapter 17 §11 left it as an exercise; part (d) of the theorem below rests on it, so it is proved here. The proof is the one the exercise's hint pointed at, and the whole content is that a positive definite matrix and its inverse sit in a single positive semidefinite block matrix.

Theorem 21.8.3. Four Operator Convex Functions

(a) Every affine function $t \mapsto \alpha + \beta t$ is operator convex and operator concave on $\mathbb{R}$.

(b) $t \mapsto t^2$ is operator convex on $\mathbb{R}$, hence on every interval.

(c) $t \mapsto 1/t$ is operator convex on $(0, \infty)$; equivalently, $t \mapsto -1/t$ is operator concave there.

(d) For every $c > 0$, the function $t \mapsto 1/(t+c)$ is operator convex on $(0, \infty)$, and $t \mapsto t/(t+c)$ is operator **concave** there.

Idea. Only (c) is new. For numbers, $1/t$ is convex because $1/t$ is the smallest u with $tu \geq 1$, and "$tu \geq 1$" is a condition that survives averaging. The matrix version of "$tu \geq 1$" is that the block matrix $\left(\begin{smallmatrix} \mathbf{A} & \mathbf{I} \\ \mathbf{I} & \mathbf{U} \end{smallmatrix}\right)$ is positive semidefinite, and Chapter 13 §06 reads that off a Schur complement. Averaging two such blocks averages the corners, and reading the Schur complement backwards turns the average back into an inequality between matrices.

Proof.

(a) is Proposition 17.11.2 and (b) is Proposition 17.11.8. Restricting to a smaller interval only shrinks the supply of matrices, so both persist on every interval.

(b) **Claim.** For every $\mathbf{G} \succ \mathbf{O}$ in $M_n(F)$, the Hermitian block matrix

$$\mathbf{M}_{\mathbf{G}} = \begin{pmatrix} \mathbf{G} & \mathbf{I}_n \\ \mathbf{I}_n & \mathbf{G}^{-1} \end{pmatrix}$$

is positive semidefinite.

Proof. The inverse of a Hermitian invertible matrix is Hermitian, since $(\mathbf{G}^{-1})^* = (\mathbf{G}^*)^{-1} = \mathbf{G}^{-1}$, so $\mathbf{M}_{\mathbf{G}}$ is Hermitian. Its corner $\mathbf{G}$ is $\succ \mathbf{O}$, and its Schur complement is $\mathbf{M}_{\mathbf{G}}/\mathbf{G} =$

$\mathbf{G}^{-1} - \mathbf{I}_n\mathbf{G}^{-1}\mathbf{I}_n = \mathbf{O} \succeq \mathbf{O}$. By Theorem 13.6.1 (b), $\mathbf{M}_{\mathbf{G}} \succeq \mathbf{O}$.

■

Now let $\mathbf{A}, \mathbf{B} \in M_n(F)$ be Hermitian with spectra in $(0, \infty)$, so $\mathbf{A} \succ \mathbf{O}$ and $\mathbf{B} \succ \mathbf{O}$ by Theorem 13.1.5 (b), and let $\theta \in [0, 1]$. Put $\mathbf{C} = \theta\mathbf{A} + (1-\theta)\mathbf{B}$. For $\mathbf{x} \neq \mathbf{0}$,

$$\langle \mathbf{C}\mathbf{x}, \mathbf{x}\rangle = \theta\langle \mathbf{A}\mathbf{x}, \mathbf{x}\rangle + (1-\theta)\langle \mathbf{B}\mathbf{x}, \mathbf{x}\rangle > 0,$$

since both terms are ≥ 0 and at least one coefficient is positive; hence $\mathbf{C} \succ \mathbf{O}$. By the Claim, $\mathbf{M}_{\mathbf{A}} \succeq \mathbf{O}$ and $\mathbf{M}_{\mathbf{B}} \succeq \mathbf{O}$, so their convex combination

$$\theta\mathbf{M}_{\mathbf{A}} + (1-\theta)\mathbf{M}_{\mathbf{B}} = \begin{pmatrix} \mathbf{C} & \mathbf{I}_n \\ \mathbf{I}_n & \theta\mathbf{A}^{-1} + (1-\theta)\mathbf{B}^{-1} \end{pmatrix}$$

has nonnegative quadratic form at every vector and is therefore $\succeq \mathbf{O}$. Its corner $\mathbf{C}$ is $\succ \mathbf{O}$, so Theorem 13.6.1 (b) applies in the other direction and gives

$$\theta\mathbf{A}^{-1} + (1-\theta)\mathbf{B}^{-1} - \mathbf{C}^{-1} \succeq \mathbf{O}.$$

Since $t \mapsto 1/t$ sends a positive definite matrix to its inverse (Definition 12.8.2, as recalled in Chapter 17 §11), this is exactly the inequality required of $f(t) = 1/t$. The size n was arbitrary.

(d) By (c) and Proposition 21.8.2 (b), $t \mapsto 1/(t+c)$ is operator convex on $(0, \infty) - c = (-c, \infty)$, hence on the smaller interval $(0, \infty)$. Finally $t/(t+c) = 1 - c \cdot \frac{1}{t+c}$, which is an affine function minus a nonnegative multiple of an operator convex function, so it is operator concave by Proposition 21.8.2 (a) applied to $-f$.

■

Remark. Compare the list with Chapter 17 §11's list of operator monotone functions. The two overlap only in the affine ones: t^2 is operator convex and not operator monotone, $-1/t$ is operator monotone and not even convex, and $t/(t+c)$ is both operator monotone (Corollary 17.11.5) and operator concave. Two further examples, $-\log t$ and $t \log t$, are operator convex on $(0, \infty)$; both proofs need an integral representation that §10 of this chapter supplies, so they are postponed to there.

21.8.3 Convex is not operator convex

The gap between convexity and operator convexity is as wide as the gap between increasing and operator monotone, and a single cube exposes it.

Example. The Cube Is Convex and Not Operator Convex

Let

$$\mathbf{A} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \qquad \mathbf{B} = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}.$$

Show that $\mathbf{A} \succeq \mathbf{O}$ and $\mathbf{B} \succ \mathbf{O}$, and that with $\theta = \frac{1}{2}$ the inequality required of $f(t) = t^3$ on $[0, \infty)$ fails.

Solution. $\mathbf{A}$ is diagonal with nonnegative entries, so $\mathbf{A} \succeq \mathbf{O}$ with spectrum $\{0, 1\}$. $\mathbf{B}$ is real symmetric with leading principal minors 2 and $2 - 1 = 1$, both positive, so $\mathbf{B} \succ \mathbf{O}$ by Sylvester's criterion (Theorem 13.1.5 (d)). Both spectra lie in $[0, \infty)$, and so does that of $\frac{1}{2}(\mathbf{A} + \mathbf{B})$, a convex combination of two positive semidefinite matrices.

Now compute. $\mathbf{A}^3 = \mathbf{A}$. From $\mathbf{B}^2 = \left(\begin{smallmatrix} 5 & 3 \\ 3 & 2 \end{smallmatrix}\right)$ we get $\mathbf{B}^3 = \left(\begin{smallmatrix} 13 & 8 \\ 8 & 5 \end{smallmatrix}\right)$. Also $\mathbf{A} + \mathbf{B} = \left(\begin{smallmatrix} 3 & 1 \\ 1 & 1 \end{smallmatrix}\right)$, whose square is $\left(\begin{smallmatrix} 10 & 4 \\ 4 & 2 \end{smallmatrix}\right)$ and whose cube is $\left(\begin{smallmatrix} 34 & 14 \\ 14 & 6 \end{smallmatrix}\right)$. Hence, clearing the denominator by a factor 8,

$$8\Big(\tfrac{1}{2}(\mathbf{A}^3 + \mathbf{B}^3) - \big(\tfrac{1}{2}(\mathbf{A} + \mathbf{B})\big)^3\Big) = 4(\mathbf{A}^3 + \mathbf{B}^3) - (\mathbf{A} + \mathbf{B})^3 = \begin{pmatrix} 22 & 18 \\ 18 & 14 \end{pmatrix}.$$

Its determinant is $22 \cdot 14 - 18^2 = 308 - 324 = -16 < 0$, so it is not positive semidefinite (Theorem 13.1.4 (e)). A witness needs no eigenvalues: with $\mathbf{x} = (2, -3)$,

$$\mathbf{x}^\top \begin{pmatrix} 22 & 18 \\ 18 & 14 \end{pmatrix} \mathbf{x} = 4 \cdot 22 - 12 \cdot 18 + 9 \cdot 14 = 88 - 216 + 126 = -2 < 0.$$

By Theorem 12.8.3 (a), $t \mapsto t^3$ sends a Hermitian matrix to its cube, so t^3 is not operator convex on $[0, \infty)$. It is nevertheless convex there in the ordinary sense, since its second derivative $6t$ is ≥ 0 (Theorem 18.10.6).

Warning. **Convexity of f is not enough, already at size 2.** Every t^p with $p \geq 1$ is convex on $[0, \infty)$, and the exponents 3 and 4 already fail to be operator convex. The trace does not notice the difference: §02's Theorem 21.2.8 makes $\mathbf{A} \mapsto \operatorname{tr} f(\mathbf{A})$ convex for **every** convex f, and in Example (The Cube Is Convex and Not Operator Convex) the trace of the displayed matrix is $22 + 14 = 36 > 0$, exactly as that theorem predicts, while the matrix itself is indefinite. A convex function controls the sum of the eigenvalues; an operator convex function controls the matrix.

Quick check. Is $t \mapsto t^4$ operator convex on $[0, \infty)$, because $t^4 = (t^2)^2$ is a composite of two operator convex functions?

Solution. No. Proposition 21.8.2 covers sums and nonnegative multiples, never composites, and there is no reason it should: operator convexity of g constrains g on convex combinations of matrices, while $(g \circ f)(\mathbf{A}) = g(f(\mathbf{A}))$ asks about g at the matrix $f(\mathbf{A})$, which is a different question. The same pair as in Example (The Cube Is Convex and Not Operator Convex) settles it. From $\mathbf{A}^4 = \mathbf{A}$, $\mathbf{B}^4 = \left(\begin{smallmatrix} 34 & 21 \\ 21 & 13 \end{smallmatrix}\right)$, and $(\mathbf{A} + \mathbf{B})^4 = \left(\begin{smallmatrix} 116 & 48 \\ 48 & 20 \end{smallmatrix}\right)$,

$$4\Big(\tfrac{1}{2}(\mathbf{A}^4 + \mathbf{B}^4) - \big(\tfrac{1}{2}(\mathbf{A} + \mathbf{B})\big)^4\Big) = \begin{pmatrix} 41 & 30 \\ 30 & 21 \end{pmatrix},$$

whose determinant is $861 - 900 = -39 < 0$. A witness is $\mathbf{x} = (3, -4)$, giving $9 \cdot 41 - 24 \cdot 30 + 16 \cdot 21 = -15 < 0$.

21.8.4 Contractions and their dilations

The theorem this section is heading for compares $f(\mathbf{V}^*\mathbf{A}\mathbf{V})$ with $\mathbf{V}^* f(\mathbf{A})\mathbf{V}$. The matrices $\mathbf{V}$ for which this can possibly work are the ones that do not stretch: otherwise $\mathbf{V}^*\mathbf{A}\mathbf{V}$ can leave the interval where f lives.

A contraction is a matrix that never lengthens a vector.

Definition 21.8.1. Contraction

Let $\mathbf{V} \in M_{n\times m}(F)$. Then $\mathbf{V}$ is a **contraction** if $\mathbf{V}^*\mathbf{V} \preceq \mathbf{I}_m$, that is, if $\|\mathbf{V}\mathbf{x}\| \le \|\mathbf{x}\|$ for every $\mathbf{x} \in F^m$. It is an **isometry** if $\mathbf{V}^*\mathbf{V} = \mathbf{I}_m$.

The word is overloaded: Chapter 15 §04 uses **contraction** for the pairing $V^* \otimes V \to F$ (Definition 15.4.1), which is a different object entirely, and the two never appear together. The two phrasings above agree because $\langle(\mathbf{I}_m - \mathbf{V}^*\mathbf{V})\mathbf{x}, \mathbf{x}\rangle = \|\mathbf{x}\|^2 - \|\mathbf{V}\mathbf{x}\|^2$. Examples: every unitary matrix; every orthogonal projection $\mathbf{P}$, since $\mathbf{P}^*\mathbf{P} = \mathbf{P}^2 = \mathbf{P} \preceq \mathbf{I}$ (Definition 13.5.1); the matrix $\frac{1}{2}\mathbf{I}$; and any matrix whose columns are an orthonormal list, which is an isometry. A nonexample by minimal change: $2\mathbf{I}$ fails, since $\mathbf{I} - (2\mathbf{I})^*(2\mathbf{I}) = -3\mathbf{I}$ is not positive semidefinite.

The key structural fact is that a contraction is always the corner of a unitary matrix of twice the size. This is the **dilation**, and it is how a statement about contractions gets converted into a statement about unitary conjugation, which the functional calculus understands.

Lemma 21.8.4. Unitary Dilation of a Contraction

Let $\mathbf{V} \in M_{n\times m}(F)$ be a contraction, and set

$$\mathbf{D}_1 = (\mathbf{I}_m - \mathbf{V}^*\mathbf{V})^{1/2}, \qquad \mathbf{D}_2 = (\mathbf{I}_n - \mathbf{V}\mathbf{V}^*)^{1/2}.$$

Then $\mathbf{I}_m - \mathbf{V}^*\mathbf{V} \succeq \mathbf{O}$ and $\mathbf{I}_n - \mathbf{V}\mathbf{V}^* \succeq \mathbf{O}$, so both square roots exist, and the $(n+m)\times(n+m)$ matrix

$$\mathbf{U} = \begin{pmatrix} \mathbf{V} & \mathbf{D}_2 \\ \mathbf{D}_1 & -\mathbf{V}^* \end{pmatrix}$$

is unitary.

Idea. Only one of the four block identities needs an idea: $\mathbf{V}^*\mathbf{D}_2 = \mathbf{D}_1\mathbf{V}^*$, which says that $\mathbf{V}^*$ intertwines the two square roots. It holds for $\mathbf{V}^*(\mathbf{I} - \mathbf{V}\mathbf{V}^*) = (\mathbf{I} - \mathbf{V}^*\mathbf{V})\mathbf{V}^*$ by one line of algebra, hence for every polynomial in the two matrices, and the positive square root **is** a polynomial (Theorem 13.2.1 (a)) — provided one polynomial can be chosen to serve both matrices at once, which Lagrange interpolation allows.

Proof. Write $\mathbf{M} = \mathbf{I}_m - \mathbf{V}^*\mathbf{V}$ and $\mathbf{N} = \mathbf{I}_n - \mathbf{V}\mathbf{V}^*$. By hypothesis $\mathbf{M} \succeq \mathbf{O}$, which says $\|\mathbf{V}\mathbf{x}\| \le \|\mathbf{x}\|$ for every $\mathbf{x} \in F^m$. For $\mathbf{N}$, fix $\mathbf{y} \in F^n$ and put $\mathbf{z} = \mathbf{V}^*\mathbf{y}$. Then, by Cauchy–Schwarz (Theorem 11.1.5) and the contraction property applied to $\mathbf{z}$,

$$\|\mathbf{z}\|^2 = \langle \mathbf{V}\mathbf{z}, \mathbf{y}\rangle \;\le\; \|\mathbf{V}\mathbf{z}\|\,\|\mathbf{y}\| \;\le\; \|\mathbf{z}\|\,\|\mathbf{y}\|,$$

the first equality because $\langle \mathbf{V}\mathbf{z}, \mathbf{y}\rangle = \mathbf{y}^*\mathbf{V}\mathbf{V}^*\mathbf{y} = \|\mathbf{V}^*\mathbf{y}\|^2$. If $\mathbf{z} \neq \mathbf{0}$ we may divide by $\|\mathbf{z}\| > 0$

to get $\|\mathbf{V}^*\mathbf{y}\| \le \|\mathbf{y}\|$, and if $\mathbf{z} = \mathbf{0}$ that inequality is immediate. Hence $\langle \mathbf{N}\mathbf{y}, \mathbf{y}\rangle = \|\mathbf{y}\|^2 - \|\mathbf{V}^*\mathbf{y}\|^2 \ge 0$ for every $\mathbf{y}$, so $\mathbf{N} \succeq \mathbf{O}$. Both square roots now exist and are Hermitian, by Theorem 13.2.1.

Claim. $\mathbf{V}^*\mathbf{N}^{1/2} = \mathbf{M}^{1/2}\mathbf{V}^*$.

Proof. First, $\mathbf{V}^*\mathbf{N} = \mathbf{V}^* - \mathbf{V}^*\mathbf{V}\mathbf{V}^* = \mathbf{M}\mathbf{V}^*$. Hence $\mathbf{V}^*\mathbf{N}^k = \mathbf{M}^k\mathbf{V}^*$ for every integer $k \ge 0$, by induction: the case $k = 0$ is trivial, and $\mathbf{V}^*\mathbf{N}^{k+1} = (\mathbf{V}^*\mathbf{N})\mathbf{N}^k = \mathbf{M}(\mathbf{V}^*\mathbf{N}^k) = \mathbf{M}^{k+1}\mathbf{V}^*$. Therefore $\mathbf{V}^* q(\mathbf{N}) = q(\mathbf{M})\mathbf{V}^*$ for every polynomial q with real coefficients.

The set $\operatorname{spec}(\mathbf{M}) \cup \operatorname{spec}(\mathbf{N})$ is finite and contained in $[0, \infty)$, so by Lagrange interpolation (Theorem 6.6.1) there is a real polynomial q with $q(\nu) = \sqrt{\nu}$ for every ν in it. By Theorem 12.8.3 (a) and Definition 12.8.2, $q(\mathbf{M}) = \sum_j q(\mu_j)\mathbf{P}_j = \sum_j \sqrt{\mu_j}\mathbf{P}_j = \mathbf{M}^{1/2}$, and likewise $q(\mathbf{N}) = \mathbf{N}^{1/2}$. Substituting q into the displayed identity proves the claim.

■

Taking adjoints in the claim, and using that $\mathbf{M}^{1/2}$ and $\mathbf{N}^{1/2}$ are Hermitian, gives $\mathbf{N}^{1/2}\mathbf{V} = \mathbf{V}\mathbf{M}^{1/2}$. Now compute, in blocks (Theorem 8.1.1),

$$\begin{aligned}\mathbf{U}^*\mathbf{U} &= \begin{pmatrix} \mathbf{V}^* & \mathbf{D}_1 \\ \mathbf{D}_2 & -\mathbf{V} \end{pmatrix}\begin{pmatrix} \mathbf{V} & \mathbf{D}_2 \\ \mathbf{D}_1 & -\mathbf{V}^* \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{V}^*\mathbf{V} + \mathbf{D}_1^2 & \mathbf{V}^*\mathbf{D}_2 - \mathbf{D}_1\mathbf{V}^* \\ \mathbf{D}_2\mathbf{V} - \mathbf{V}\mathbf{D}_1 & \mathbf{D}_2^2 + \mathbf{V}\mathbf{V}^* \end{pmatrix}.\end{aligned}$$

The top-left block is $\mathbf{V}^*\mathbf{V} + \mathbf{M} = \mathbf{I}_m$ and the bottom-right is $\mathbf{N} + \mathbf{V}\mathbf{V}^* = \mathbf{I}_n$, both by the definition of $\mathbf{M}$ and $\mathbf{N}$. The top-right block is $\mathbf{O}$ by the Claim, and the bottom-left is $\mathbf{O}$ by its adjoint form. So $\mathbf{U}^*\mathbf{U} = \mathbf{I}_{m+n}$, and since $\mathbf{U}$ is square this makes $\mathbf{U}$ unitary (Theorem 3.4.4).

■

Quick check. Take $m = n = 1$ and $\mathbf{V} = (c)$ with c real, $|c| \le 1$. What is the dilation?

Solution. $\mathbf{D}_1 = \mathbf{D}_2 = (\sqrt{1-c^2})$, so $\mathbf{U} = \left(\begin{smallmatrix} c & \sqrt{1-c^2} \\ \sqrt{1-c^2} & -c \end{smallmatrix}\right)$. Writing $c = \cos\varphi$ with $\varphi \in [0, \pi]$, so that $\sin\varphi \ge 0$ and $\sqrt{1-c^2} = \sin\varphi$, this is the reflection matrix $\left(\begin{smallmatrix} \cos\varphi & \sin\varphi \\ \sin\varphi & -\cos\varphi \end{smallmatrix}\right)$, which is orthogonal with determinant -1. So the lemma is the statement that every number in $[-1, 1]$ is the corner of a 2×2 reflection — the sentence "$\cos^2\varphi + \sin^2\varphi = 1$", read as a matrix identity.

21.8.5 Jensen's inequality for operators

Here is the theorem. In words: **squeezing first and then applying f is at most applying f first and then squeezing**, provided f is operator convex and does not reward the squeezing at the origin.

Theorem 21.8.5. Hansen–Pedersen–Jensen Inequality

Let $I \subseteq \mathbb{R}$ be an interval with $0 \in I$, let $f\colon I \to \mathbb{R}$ be operator convex on I, let $\mathbf{A} \in M_n(F)$ be Hermitian with $\operatorname{spec}(\mathbf{A}) \subseteq I$, and let $\mathbf{V} \in M_{n\times m}(F)$ be a contraction. Then $\operatorname{spec}(\mathbf{V}^*\mathbf{A}\mathbf{V}) \subseteq I$, and

$$f(\mathbf{V}^*\mathbf{A}\mathbf{V}) \preceq \mathbf{V}^* f(\mathbf{A})\mathbf{V} + f(0)(\mathbf{I}_m - \mathbf{V}^*\mathbf{V}).$$

In particular:

(a) if $f(0) \le 0$, then $f(\mathbf{V}^*\mathbf{A}\mathbf{V}) \preceq \mathbf{V}^* f(\mathbf{A})\mathbf{V}$;

(b) if $\mathbf{V}$ is an isometry, then $f(\mathbf{V}^*\mathbf{A}\mathbf{V}) \preceq \mathbf{V}^* f(\mathbf{A})\mathbf{V}$, with no condition on $f(0)$.

Idea. Dilate. Once $\mathbf{V}$ sits in the corner of a unitary $\mathbf{U}$, the matrix $\mathbf{V}^*\mathbf{A}\mathbf{V}$ sits in the corner of $\mathbf{U}^*(\mathbf{A}\oplus\mathbf{O})\mathbf{U}$, which is a unitary conjugate of $\mathbf{A}\oplus\mathbf{O}$ and therefore has the same spectrum. Conjugating a second time by the sign matrix $\mathbf{W} = \mathbf{I}_m \oplus (-\mathbf{I}_n)$ flips the off-diagonal blocks, so **averaging the two conjugates deletes them**: the average is block diagonal, with $\mathbf{V}^*\mathbf{A}\mathbf{V}$ in the corner. Operator convexity applied to that average, and then reading off the top-left block, is the whole proof. The term $f(0)(\mathbf{I} - \mathbf{V}^*\mathbf{V})$ is the residue of the zero block we padded with.

Proof. **The spectrum.** Let $\mathbf{x} \in F^m$ be a unit eigenvector of the Hermitian matrix $\mathbf{V}^*\mathbf{A}\mathbf{V}$, with eigenvalue μ. If $\mathbf{V}\mathbf{x} = \mathbf{0}$, then $\mu = \langle \mathbf{V}^*\mathbf{A}\mathbf{V}\mathbf{x}, \mathbf{x}\rangle = 0 \in I$. Otherwise put $\mathbf{y} = \mathbf{V}\mathbf{x}/\|\mathbf{V}\mathbf{x}\|$ and $\tau = \|\mathbf{V}\mathbf{x}\|^2$, so $0 < \tau \le 1$ because $\mathbf{V}$ is a contraction, and

$$\mu = \langle \mathbf{A}\mathbf{V}\mathbf{x}, \mathbf{V}\mathbf{x}\rangle = \tau\langle \mathbf{A}\mathbf{y}, \mathbf{y}\rangle.$$

By Lemma 13.5.2, $\langle \mathbf{A}\mathbf{y}, \mathbf{y}\rangle$ lies between $\lambda_n(\mathbf{A})$ and $\lambda_1(\mathbf{A})$, hence in I, since I is an interval containing the spectrum. Then $\mu = \tau\langle \mathbf{A}\mathbf{y}, \mathbf{y}\rangle + (1-\tau)\cdot 0$ is a convex combination of two points of I, so $\mu \in I$. The same argument applies to $\mathbf{D}_2 = (\mathbf{I}_n - \mathbf{V}\mathbf{V}^*)^{1/2}$: it is Hermitian with $\mathbf{D}_2^*\mathbf{D}_2 = \mathbf{I}_n - \mathbf{V}\mathbf{V}^* \preceq \mathbf{I}_n$, because $\mathbf{V}\mathbf{V}^* \succeq \mathbf{O}$, so it is a contraction. Hence $\operatorname{spec}(\mathbf{D}_2\mathbf{A}\mathbf{D}_2) \subseteq I$.

The two conjugates. Let $\mathbf{U}$ be the dilation of Lemma 21.8.4 and put

$$\mathbf{M} = \mathbf{A} \oplus \mathbf{O}_m, \qquad \mathbf{W} = \mathbf{I}_m \oplus (-\mathbf{I}_n),$$

so that $\mathbf{W}$ is unitary and Hermitian, and $\operatorname{spec}(\mathbf{M}) = \operatorname{spec}(\mathbf{A}) \cup \{0\} \subseteq I$. Multiplying out in blocks,

$$\mathbf{U}^*\mathbf{M}\mathbf{U} = \begin{pmatrix} \mathbf{V}^*\mathbf{A}\mathbf{V} & \mathbf{V}^*\mathbf{A}\mathbf{D}_2 \\ \mathbf{D}_2\mathbf{A}\mathbf{V} & \mathbf{D}_2\mathbf{A}\mathbf{D}_2 \end{pmatrix},$$

and conjugating by $\mathbf{W}$ negates the two off-diagonal blocks. Hence, with $\mathbf{X} = \mathbf{U}^*\mathbf{M}\mathbf{U}$ and $\mathbf{Y} = (\mathbf{U}\mathbf{W})^*\mathbf{M}(\mathbf{U}\mathbf{W}) = \mathbf{W}^*\mathbf{X}\mathbf{W}$,

$$\tfrac{1}{2}(\mathbf{X} + \mathbf{Y}) = (\mathbf{V}^*\mathbf{A}\mathbf{V}) \oplus (\mathbf{D}_2\mathbf{A}\mathbf{D}_2). \tag{$*$}$$

Both $\mathbf{X}$ and $\mathbf{Y}$ are unitary conjugates of $\mathbf{M}$, so both are Hermitian with spectrum $\operatorname{spec}(\mathbf{M}) \subseteq I$ by Lemma 21.8.1 (b).

Convexity. Apply Definition 17.11.1 (b) with $\theta = \tfrac{1}{2}$ to $\mathbf{X}$ and $\mathbf{Y}$, then Lemma 21.8.1 (b) twice:

$$f\big(\tfrac{1}{2}(\mathbf{X} + \mathbf{Y})\big) \preceq \tfrac{1}{2}(\mathbf{U}^* f(\mathbf{M})\mathbf{U} + \mathbf{W}^*\mathbf{U}^* f(\mathbf{M})\mathbf{U}\mathbf{W}).$$

By Lemma 21.8.1 (c), $f(\mathbf{M}) = f(\mathbf{A}) \oplus f(0)\mathbf{I}_m$. The same block computation as before, now with $f(\mathbf{A})$ in place of $\mathbf{A}$ and $f(0)\mathbf{I}_m$ in place of $\mathbf{O}_m$, gives

$$\mathbf{U}^* f(\mathbf{M})\mathbf{U} = \begin{pmatrix} \mathbf{V}^* f(\mathbf{A})\mathbf{V} + f(0)\mathbf{D}_1^2 & * \\ * & * \end{pmatrix},$$

and averaging with its $\mathbf{W}$-conjugate again deletes the off-diagonal blocks. On the left, Lemma 21.8.1 (c) and $(*)$ give $f(\frac{1}{2}(\mathbf{X}+\mathbf{Y})) = f(\mathbf{V}^*\mathbf{A}\mathbf{V}) \oplus f(\mathbf{D}_2\mathbf{A}\mathbf{D}_2)$. So the inequality reads, between two block diagonal matrices,

$$f(\mathbf{V}^*\mathbf{A}\mathbf{V}) \oplus f(\mathbf{D}_2\mathbf{A}\mathbf{D}_2) \preceq (\mathbf{V}^* f(\mathbf{A})\mathbf{V} + f(0)\mathbf{D}_1^2) \oplus \mathbf{Z}$$

for some Hermitian $\mathbf{Z}$. The difference of the two sides is positive semidefinite, so its top-left $m \times m$ block is too, by Theorem 13.6.1 (a). Since $\mathbf{D}_1^2 = \mathbf{I}_m - \mathbf{V}^*\mathbf{V}$, that block inequality is the displayed statement of the theorem.

(a) If $f(0) \le 0$, then $f(0)(\mathbf{I}_m - \mathbf{V}^*\mathbf{V}) \preceq \mathbf{O}$, because $\mathbf{I}_m - \mathbf{V}^*\mathbf{V} \succeq \mathbf{O}$ and a nonpositive multiple of a positive semidefinite matrix has nonpositive quadratic form. Adding it can only decrease, so $\mathbf{V}^* f(\mathbf{A})\mathbf{V} + f(0)(\mathbf{I}_m - \mathbf{V}^*\mathbf{V}) \preceq \mathbf{V}^* f(\mathbf{A})\mathbf{V}$, and transitivity of $\preceq$ (Proposition 13.5.1) finishes.

(b) If $\mathbf{V}^*\mathbf{V} = \mathbf{I}_m$ the extra term is $\mathbf{O}$.

■

The special case worth remembering separately takes $\mathbf{V}$ to be a projection. It says that **an operator convex function commutes with compression, up to the right inequality**.

Corollary 21.8.6. Compression

Let I be an interval with $0 \in I$, let $f\colon I \to \mathbb{R}$ be operator convex with $f(0) \le 0$, let $\mathbf{A} \in M_n(F)$ be Hermitian with $\operatorname{spec}(\mathbf{A}) \subseteq I$, and let $\mathbf{P} \in M_n(F)$ be the matrix of an orthogonal projection. Then $\operatorname{spec}(\mathbf{PAP}) \subseteq I$ and

$$f(\mathbf{PAP}) \preceq \mathbf{P} f(\mathbf{A})\mathbf{P}.$$

Proof. $\mathbf{P}$ is Hermitian with $\mathbf{P}^*\mathbf{P} = \mathbf{P}^2 = \mathbf{P} \preceq \mathbf{I}$, so $\mathbf{P}$ is a contraction (Definition 21.8.1). Apply Theorem 21.8.5 (a) with $\mathbf{V} = \mathbf{P}$, noting $\mathbf{P}^*\mathbf{A}\mathbf{P} = \mathbf{P}\mathbf{A}\mathbf{P}$.

■

Warning. **The condition $f(0) \le 0$ is not decoration.** Take $f(t) = t^2 + 1$, operator convex on $\mathbb{R}$ by Proposition 21.8.2 (a), with $f(0) = 1 > 0$. Take $\mathbf{A} = \mathbf{O}_n$ and $\mathbf{V} = \mathbf{O}_{n\times m}$, a contraction. Then $f(\mathbf{V}^*\mathbf{A}\mathbf{V}) = f(\mathbf{O}_m) = \mathbf{I}_m$ while $\mathbf{V}^* f(\mathbf{A})\mathbf{V} = \mathbf{O}$, so the conclusion of Theorem 21.8.5 (a) fails as badly as it can. The full inequality of the theorem survives: its right-hand side is $\mathbf{O} + 1 \cdot \mathbf{I}_m = \mathbf{I}_m$, with equality.

Quick check. Why can the hypothesis "$\mathbf{V}$ is a contraction" not be dropped in Corollary 21.8.6 by rescaling — that is, why does the corollary for $\mathbf{P}$ not give a version for $c\mathbf{P}$ with $c > 1$?

Solution. Because both sides scale differently. For $f(t) = t^2$ and $\mathbf{V} = c\mathbf{I}$ the claimed inequality would read $c^4\mathbf{A}^2 \preceq c^2\mathbf{A}^2$, which fails for every $c > 1$ and every $\mathbf{A}$ with $\mathbf{A}^2 \neq \mathbf{O}$. The contraction hypothesis is what keeps the left-hand side from being inflated, and it enters the proof twice: it puts $\operatorname{spec}(\mathbf{V}^*\mathbf{A}\mathbf{V})$ back inside I, which is what makes $f(\mathbf{V}^*\mathbf{A}\mathbf{V})$ defined at all, and it gives the square roots of Lemma 21.8.4 something to be square roots of.

21.8.6 Where this sits

Two convexity statements about matrices are now on the table, and they are of very different strengths.

- §02's Theorem 21.2.8: for **every** convex f, the function $\mathbf{A} \mapsto \operatorname{tr} f(\mathbf{A})$ is convex on the Hermitian matrices with spectrum in I. This is a statement about one real number attached to the matrix.
- Operator convexity: for a much smaller class of f, the matrices themselves are ordered. Example (The Cube Is Convex and Not Operator Convex) shows t^3 in the first class and not the second.

The Hansen–Pedersen–Jensen inequality is the reason the smaller class is worth isolating. A trace inequality cannot be compressed: knowing $\operatorname{tr} f(\mathbf{A})$ says nothing about $\operatorname{tr} f(\mathbf{P}\mathbf{A}\mathbf{P})$, since the compression changes the eigenvalues in ways the trace does not track. What Corollary 21.8.6 compares is not the traces but the matrices, $f(\mathbf{P}\mathbf{A}\mathbf{P})$ with $\mathbf{P}f(\mathbf{A})\mathbf{P}$, and it is the step that later turns an operator convex function into an inequality between matrices of different sizes. Section 10 of this chapter adds $-\log t$ and $t \log t$ to the list of operator convex functions, and §11 uses the operator monotonicity of $t^{1/2}$ to build the geometric mean of two positive definite matrices.

21.8.7 Exercises

21.8.7.1 A. Check your understanding

Exercise. A1

(a) State what it means for $f\colon I \to \mathbb{R}$ to be operator convex, and say what the clause "for every n" ranges over.

(b) Explain why an operator convex function is convex, and name a convex function that is not operator convex.

(c) Decide whether the following is correct, and justify your answer: if f is operator convex on I and $c < 0$, then cf is operator convex on I.

(d) State the Hansen–Pedersen–Jensen inequality, and say which hypothesis fails for $f(t) = t^2 + 1$.

(e) Is $t \mapsto -1/t$ operator convex on $(0, \infty)$? Is it operator concave there?

Solution.

(a) For every $n \geq 1$, all Hermitian $\mathbf{A}, \mathbf{B} \in M_n(F)$ with spectra in I and every $\theta \in [0,1]$, $f(\theta\mathbf{A} + (1-\theta)\mathbf{B}) \preceq \theta f(\mathbf{A}) + (1-\theta) f(\mathbf{B})$ (Definition 17.11.1 (b)). The clause ranges over all matrix sizes at once.

(b) Taking $n = 1$ turns the matrices into numbers and $\preceq$ into $\leq$, which is Definition 18.10.1. The function $t \mapsto t^3$ is convex on $[0, \infty)$ and not operator convex there (Example (The Cube Is Convex and Not Operator Convex)).

(c) Incorrect: Proposition 21.8.2 (a) requires $c \geq 0$. For $c = -1$ and $f(t) = t^2$, the claim would make $-t^2$ operator convex, that is t^2 operator concave; but Proposition 17.11.8 gives $\theta\mathbf{A}^2 + (1-\theta)\mathbf{B}^2 - (\theta\mathbf{A} + (1-\theta)\mathbf{B})^2 = \theta(1-\theta)(\mathbf{A}-\mathbf{B})^2$, which is nonzero for $\mathbf{A} \neq \mathbf{B}$ and $0 < \theta < 1$, so the reverse inequality fails.

(d) Theorem 21.8.5: for f operator convex on an interval containing 0, $\mathbf{A}$ Hermitian with spectrum in that interval and $\mathbf{V}$ a contraction, $f(\mathbf{V}^*\mathbf{A}\mathbf{V}) \preceq \mathbf{V}^* f(\mathbf{A})\mathbf{V} + f(0)(\mathbf{I} - \mathbf{V}^*\mathbf{V})$. The last term vanishes when $\mathbf{V}$ is an isometry, and when $f(0) \leq 0$ it is $\preceq \mathbf{O}$ and may be discarded, which leaves $f(\mathbf{V}^*\mathbf{A}\mathbf{V}) \preceq \mathbf{V}^* f(\mathbf{A})\mathbf{V}$ in both cases. For $f(t) = t^2 + 1$ the failing hypothesis is $f(0) \leq 0$.

(e) It is not operator convex: at $n = 1$ the scalar function $-1/t$ is not convex, since at 1, 3 and their midpoint 2 the value $-\frac{1}{2}$ exceeds the chord's $-\frac{2}{3}$. It **is** operator concave, by Theorem 21.8.3 (c).

21.8.7.2 B. Practice

Exercise. B1: Which are operator convex?

Determine which of the following functions are operator convex on the interval given. Justify your answer.

(i) $f(t) = 7 - 3t$ on $\mathbb{R}$.

(ii) $f(t) = 2t^2 + 5t - 1$ on $\mathbb{R}$.

(iii) $f(t) = \dfrac{4}{t+3}$ on $(0, \infty)$.

(iv) $f(t) = \dfrac{t}{t+3}$ on $(0, \infty)$.

(v) $f(t) = t^3$ on $[0, \infty)$.

Solution.

(i) Operator convex (and concave), by Theorem 21.8.3 (a).

(ii) Operator convex: it is 2 times t^2 plus the affine function $5t - 1$, so Proposition 21.8.2 (a) applies with $c = 2 \geq 0$.

(iii) Operator convex, by Theorem 21.8.3 (d) with $c = 3$ and Proposition 21.8.2 (a) with the factor $4 \geq 0$.

(iv) Not operator convex; it is operator **concave** by Theorem 21.8.3 (d). Operator convexity would force ordinary convexity at $n = 1$, and $t/(t+3) = 1 - 3/(t+3)$ has second derivative $-6/(t+3)^3 < 0$ on $(0, \infty)$, so it is strictly concave there (Theorem 18.10.6).

(v) Not operator convex, by Example (The Cube Is Convex and Not Operator Convex).

Exercise. B2: A dilation by hand

Let $\mathbf{V} = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & 0 \end{pmatrix}$.

(a) Verify that $\mathbf{V}$ is a contraction and compute $\mathbf{D}_1 = (\mathbf{I}_2 - \mathbf{V}^*\mathbf{V})^{1/2}$ and $\mathbf{D}_2 = (\mathbf{I}_2 - \mathbf{V}\mathbf{V}^*)^{1/2}$.

(b) Write down the 4×4 matrix $\mathbf{U}$ of Lemma 21.8.4 and verify directly that its first column is a unit vector orthogonal to its third column.

Solution.

(a) $\mathbf{V}^*\mathbf{V} = \mathbf{V}\mathbf{V}^* = \operatorname{diag}(\frac{1}{4}, 0)$, so $\mathbf{I}_2 - \mathbf{V}^*\mathbf{V} = \operatorname{diag}(\frac{3}{4}, 1) \succeq \mathbf{O}$, and $\mathbf{V}$ is a contraction. Both $\mathbf{I}_2 - \mathbf{V}^*\mathbf{V}$ and $\mathbf{I}_2 - \mathbf{V}\mathbf{V}^*$ equal $\operatorname{diag}(\frac{3}{4}, 1)$, a diagonal matrix with nonnegative entries, so

$$\mathbf{D}_1 = \mathbf{D}_2 = \operatorname{diag}\left(\tfrac{\sqrt{3}}{2},\, 1\right),$$

the unique positive semidefinite square root (Theorem 13.2.1).

(b) With the blocks in the order of the lemma,

$$\mathbf{U} = \begin{pmatrix} \frac{1}{2} & 0 & \frac{\sqrt{3}}{2} & 0 \\ 0 & 0 & 0 & 1 \\ \frac{\sqrt{3}}{2} & 0 & -\frac{1}{2} & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}.$$

Its first column is $(\frac{1}{2}, 0, \frac{\sqrt{3}}{2}, 0)$, of squared length $\frac{1}{4} + \frac{3}{4} = 1$. Its third column is $(\frac{\sqrt{3}}{2}, 0, -\frac{1}{2}, 0)$, and the inner product of the two is $\frac{1}{2} \cdot \frac{\sqrt{3}}{2} + \frac{\sqrt{3}}{2} \cdot (-\frac{1}{2}) = 0$.

Exercise. B3: Compression of a square

Let

$$\mathbf{A} = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}, \qquad \mathbf{P} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}.$$

Compute $(\mathbf{PAP})^2$ and $\mathbf{PA}^2\mathbf{P}$, and verify the inequality of Corollary 21.8.6 for $f(t) = t^2$. Hence state what the inequality says about the entry in position $(1, 1)$.

Solution. $\mathbf{P}$ is the orthogonal projection onto $\operatorname{span}(\mathbf{e}_1)$, and $\mathbf{PAP} = \operatorname{diag}(2, 0)$, so $(\mathbf{PAP})^2 = \operatorname{diag}(4, 0)$. From $\mathbf{A}^2 = \left(\begin{smallmatrix} 5 & 3 \\ 3 & 2 \end{smallmatrix}\right)$ we get $\mathbf{PA}^2\mathbf{P} = \operatorname{diag}(5, 0)$. Hence

$$\mathbf{PA}^2\mathbf{P} - (\mathbf{PAP})^2 = \operatorname{diag}(1, 0) \;\succeq\; \mathbf{O},$$

as Corollary 21.8.6 predicts for $f(t) = t^2$, which is operator convex with $f(0) = 0$. In position $(1, 1)$ the inequality says $(\mathbf{A}^2)_{11} \geq (a_{11})^2$, that is $5 \geq 4$: the $(1, 1)$ entry of $\mathbf{A}^2$ is $\sum_j |a_{1j}|^2$, which exceeds $|a_{11}|^2$ by the contribution of the off-diagonal entries.

21.8.7.3 C. Going deeper

Exercise. C1: The square, without the dilation

Let $\mathbf{A} \in M_n(F)$ be Hermitian and let $\mathbf{V} \in M_{n\times m}(F)$ be a contraction.

(a) Prove that $(\mathbf{V}^*\mathbf{A}\mathbf{V})^2 \preceq \mathbf{V}^*\mathbf{A}^2\mathbf{V}$ directly, without using Theorem 21.8.5.

(b) Which hypothesis of Theorem 21.8.5 does your proof use, and which does it not?

Hint for (a): insert $\mathbf{I} - \mathbf{V}\mathbf{V}^*$.

Solution.

(a) Compute the difference and factor it:

$$\mathbf{V}^*\mathbf{A}^2\mathbf{V} - (\mathbf{V}^*\mathbf{A}\mathbf{V})^2 = \mathbf{V}^*\mathbf{A}\mathbf{A}\mathbf{V} - \mathbf{V}^*\mathbf{A}\mathbf{V}\mathbf{V}^*\mathbf{A}\mathbf{V} = \mathbf{V}^*\mathbf{A}(\mathbf{I}_n - \mathbf{V}\mathbf{V}^*)\mathbf{A}\mathbf{V}.$$

As in the proof of Lemma 21.8.4, $\mathbf{I}_n - \mathbf{V}\mathbf{V}^* \succeq \mathbf{O}$ because $\mathbf{V}$ is a contraction. Writing $\mathbf{S} = \mathbf{A}\mathbf{V}$, the right-hand side is $\mathbf{S}^*(\mathbf{I}_n - \mathbf{V}\mathbf{V}^*)\mathbf{S}$, which is $\succeq \mathbf{O}$ by Proposition 13.1.2 (a). Hence $(\mathbf{V}^*\mathbf{A}\mathbf{V})^2 \preceq \mathbf{V}^*\mathbf{A}^2\mathbf{V}$.

(b) It uses the contraction hypothesis, in the single step $\mathbf{I}_n - \mathbf{V}\mathbf{V}^* \succeq \mathbf{O}$. It does not use $f(0) \le 0$ — but it does not need to, since $f(t) = t^2$ has $f(0) = 0$.

Exercise. C2: Shrinking the matrix

Let I be an interval with $0 \in I$, let f be operator convex on I with $f(0) \le 0$, and let $\mathbf{A} \in M_n(F)$ be Hermitian with $\operatorname{spec}(\mathbf{A}) \subseteq I$.

(a) Prove that $f(\theta\mathbf{A}) \preceq \theta f(\mathbf{A})$ for every $\theta \in [0,1]$, assuming $\operatorname{spec}(\theta\mathbf{A}) \subseteq I$.

(b) Give an operator convex f with $f(0) > 0$ for which (a) fails, and say at which θ.

(c) Let $\mathbf{V} \in M_{n\times 1}(F)$ be a unit column vector and $g(t) = t^2 + 1$ on $\mathbb{R}$. Write out what Theorem 21.8.5 (b) says here, and verify it for $\mathbf{A} = \left(\begin{smallmatrix} 2 & 1 \\ 1 & 1 \end{smallmatrix}\right)$ and $\mathbf{V} = \mathbf{e}_1$.

Hint for (a): a scalar multiple of the identity is a contraction when the scalar is small.

Solution.

(a) Put $\mathbf{V} = \sqrt{\theta}\,\mathbf{I}_n$, which is Hermitian with $\mathbf{V}^*\mathbf{V} = \theta\mathbf{I}_n \preceq \mathbf{I}_n$, so $\mathbf{V}$ is a contraction (Definition 21.8.1). Then $\mathbf{V}^*\mathbf{A}\mathbf{V} = \theta\mathbf{A}$ and $\mathbf{V}^*f(\mathbf{A})\mathbf{V} = \theta f(\mathbf{A})$. By Theorem 21.8.5 (a),

$$f(\theta\mathbf{A}) = f(\mathbf{V}^*\mathbf{A}\mathbf{V}) \;\preceq\; \mathbf{V}^*f(\mathbf{A})\mathbf{V} = \theta f(\mathbf{A}).$$

(b) Take $f(t) = t^2 + 1$, operator convex on $\mathbb{R}$ by Proposition 21.8.2 (a), with $f(0) = 1 > 0$. At $\theta = 0$ the claim reads $f(\mathbf{O}) \preceq \mathbf{O}$, that is $\mathbf{I}_n \preceq \mathbf{O}$, which is false. In fact it fails at every $\theta < 1$: the claim is $\theta^2\mathbf{A}^2 + \mathbf{I} \preceq \theta\mathbf{A}^2 + \theta\mathbf{I}$, which rearranges to $(1-\theta)\mathbf{I} \preceq (\theta - \theta^2)\mathbf{A}^2$, and for $\mathbf{A} = \mathbf{O}$ the right-hand side is $\mathbf{O}$ while the left-hand side is $\succ \mathbf{O}$.

(c) A unit column $\mathbf{V}$ satisfies $\mathbf{V}^*\mathbf{V} = (1) = \mathbf{I}_1$, so it is an isometry and (b) of the theorem applies with no condition on $g(0) = 1$. It says $g(\mathbf{V}^*\mathbf{A}\mathbf{V}) \preceq \mathbf{V}^*g(\mathbf{A})\mathbf{V}$, which for 1×1 matrices is the number inequality $(\mathbf{V}^*\mathbf{A}\mathbf{V})^2 + 1 \le \mathbf{V}^*(\mathbf{A}^2 + \mathbf{I})\mathbf{V}$. With $\mathbf{V} = \mathbf{e}_1$: $\mathbf{e}_1^*\mathbf{A}\mathbf{e}_1 = 2$, so the left side is 5; and $\mathbf{A}^2 = \left(\begin{smallmatrix} 5 & 3 \\ 3 & 2 \end{smallmatrix}\right)$, so the right side is $5 + 1 = 6$. Indeed $5 \le 6$.

Exercise. C3: What the trace does not see

Let $\mathbf{A} = \operatorname{diag}(1, 0)$ and $\mathbf{B} = \left(\begin{smallmatrix} 2 & 1 \\ 1 & 1 \end{smallmatrix}\right)$ be the matrices of Example (The Cube Is Convex and Not Operator Convex), and let $f(t) = t^3$ on $[0, \infty)$.

(a) Compute $\operatorname{tr}(\frac{1}{2}(\mathbf{A}^3 + \mathbf{B}^3)) - \operatorname{tr}((\frac{1}{2}(\mathbf{A} + \mathbf{B}))^3)$, and say which theorem makes its sign predictable.

(b) The same defect matrix is indefinite. Explain, in one sentence, why (a) and that fact are consistent.

(c) Give a convex function g on $[0, \infty)$ for which the defect matrix at this pair $\mathbf{A}, \mathbf{B}$ is positive semidefinite, and name the reason.

Solution.

(a) From Example (The Cube Is Convex and Not Operator Convex) the defect is $\frac{1}{8}\left(\begin{smallmatrix} 22 & 18 \\ 18 & 14 \end{smallmatrix}\right)$, whose trace is $\frac{1}{8}(22 + 14) = \frac{9}{2} > 0$. Its sign is predictable from Theorem 21.2.8: $t \mapsto t^3$ is convex on $[0, \infty)$, so $\mathbf{A} \mapsto \operatorname{tr}(\mathbf{A}^3)$ is a convex function of the Hermitian matrix, and the defect of a convex function at a midpoint is ≥ 0.

(b) A Hermitian 2×2 matrix can have positive trace and a negative eigenvalue; the trace only sees the **sum** of the eigenvalues, and here they are approximately 4.55 and -0.055.

(c) Take $g(t) = t^2$. By Proposition 17.11.8 the defect equals $\frac{1}{4}(\mathbf{A} - \mathbf{B})^2$, a nonnegative multiple of the square of a Hermitian matrix, hence $\succeq \mathbf{O}$. Explicitly $\mathbf{A} - \mathbf{B} = \left(\begin{smallmatrix} -1 & -1 \\ -1 & -1 \end{smallmatrix}\right)$, so the defect is $\frac{1}{4}\left(\begin{smallmatrix} 2 & 2 \\ 2 & 2 \end{smallmatrix}\right) = \left(\begin{smallmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{smallmatrix}\right) \succeq \mathbf{O}$.

21.9 Loewner Matrices

Chapter 17 §11 left operator monotonicity as an infinite list of conditions: one for every matrix size, each quantified over all pairs of Hermitian matrices of that size. Nothing in that list can be checked. This section replaces the condition at size n by a condition on **numbers**: a single symmetric $n \times n$ matrix built out of f alone, at n chosen points, must be positive semidefinite. Two ingredients do the work — a formula for the derivative of $s \mapsto f(\mathbf{A} + s\mathbf{H})$, and the Schur product theorem of Chapter 13 §08. They are the elementary half of Loewner's theorem, and they are proved here in full.

Throughout, $F = \mathbb{R}$ **or** $F = \mathbb{C}$, every matrix whose spectrum is named is Hermitian, and I denotes an **interval** of real numbers with more than one point. The letter I is plain italic and is not the identity matrix $\mathbf{I}$. For $f\colon I \to \mathbb{R}$ and Hermitian $\mathbf{A}$ with $\operatorname{spec}(\mathbf{A}) \subseteq I$, the matrix $f(\mathbf{A})$ is the one built by the functional calculus of Chapter 12 §08 (Definition 12.8.2). We write $f \in C^1(I)$ to mean that f is differentiable on I, with one-sided derivatives at any endpoint belonging to I, and that f' is continuous on I.

21.9.1 Slopes between two points

Operator monotonicity compares $f(\mathbf{A})$ with $f(\mathbf{B})$, and the difference of two values of f is what a slope measures. Everything below is assembled from slopes.

The divided difference of f at two points is the slope of the chord joining them, and at a repeated

point it is the slope of the tangent.

Definition 21.9.1. Divided Difference

Let $f\colon I \to \mathbb{R}$ be differentiable. For $s, t \in I$, the **divided difference** of f is

$$f[s,t] := \begin{cases} \dfrac{f(s)-f(t)}{s-t}, & s \neq t, \\ f'(t), & s = t. \end{cases}$$

In words: for two distinct points it is the slope of the chord through $(s, f(s))$ and $(t, f(t))$; the second clause is the value that makes the first one continuous as s approaches t, which is exactly what the derivative is. It is symmetric, $f[s,t] = f[t,s]$, since swapping s and t negates both numerator and denominator, and it is additive in f: $(f+g)[s,t] = f[s,t] + g[s,t]$ and $(cf)[s,t] = c\,f[s,t]$.

Examples.

- $f(t) = \alpha + \beta t$: then $f[s,t] = \beta$ for all s, t, including $s = t$. The divided difference of an affine function is constant — it has one slope.
- $f(t) = t^2$: then $f[s,t] = (s^2 - t^2)/(s-t) = s + t$ for $s \neq t$, and $f[t,t] = 2t = t + t$. So $f[s,t] = s + t$ throughout.
- $f(t) = t^k$ for $k \geq 1$: dividing $s^k - t^k$ by $s - t$ gives $f[s,t] = \sum_{p=0}^{k-1} s^p t^{k-1-p}$, which at $s = t$ reads $kt^{k-1} = f'(t)$, so again one formula covers both clauses.
- $f(t) = -1/t$ on $(0, \infty)$: then $-1/s + 1/t = (s-t)/(st)$, so $f[s,t] = 1/(st)$, and at $s = t$ this is $1/t^2 = f'(t)$.
- $f(t) = \sqrt{t}$ on $(0, \infty)$: rationalizing, $(\sqrt{s} - \sqrt{t})/(s-t) = 1/(\sqrt{s} + \sqrt{t})$, and at $s = t$ this is $1/(2\sqrt{t}) = f'(t)$.

Nonexample by minimal change. Drop differentiability and keep continuity: for $f(t) = |t|$ on $\mathbb{R}$ the first clause still makes sense, but $f[0,0]$ has no value, since the chord slopes $f[s,0] = \operatorname{sgn}(s)$ do not settle down as $s \to 0$. Differentiability fills in the diagonal; continuous differentiability makes the filled-in function continuous.

Lemma 21.9.1. Divided Differences Are Continuous

Let $f \in C^1(I)$. If (s_k) and (t_k) are sequences in I with $s_k \to s_0 \in I$ and $t_k \to t_0 \in I$, then $f[s_k, t_k] \to f[s_0, t_0]$. In particular, for each fixed $\alpha \in I$ the function $t \mapsto f[\alpha, t]$ is continuous on I.

Proof. *Case 1:* $s_0 \neq t_0$. For large k we have $s_k \neq t_k$, since $s_k - t_k \to s_0 - t_0 \neq 0$. So $f[s_k, t_k] = (f(s_k) - f(t_k))/(s_k - t_k)$ for large k, and the algebra of limits applies, the denominator having nonzero limit: the quotient tends to $(f(s_0) - f(t_0))/(s_0 - t_0) = f[s_0, t_0]$, using that f is continuous because it is differentiable.

Case 2: $s_0 = t_0$. Let $\varepsilon > 0$. Since f' is continuous at t_0, there is $\delta > 0$ with $|f'(c) - f'(t_0)| < \varepsilon$ for every $c \in I$ with $|c - t_0| < \delta$. Take k large enough that $|s_k - t_0| < \delta$ and $|t_k - t_0| < \delta$. If $s_k = t_k$ then $f[s_k, t_k] = f'(s_k)$, which is within ε of $f'(t_0)$. If $s_k \neq t_k$, then by the mean value theorem (A6), applied to f on the interval with endpoints s_k and t_k — an interval contained

in I — there is c_k strictly between them with

$$f[s_k, t_k] = \frac{f(s_k) - f(t_k)}{s_k - t_k} = f'(c_k).$$

Then $|c_k - t_0| < \delta$, so again $|f[s_k, t_k] - f'(t_0)| < \varepsilon$. Since $f[t_0, t_0] = f'(t_0)$, this proves the claim.

■

21.9.2 The Loewner matrix

Collect the slopes between all pairs drawn from a list of points, and a matrix appears.

The Loewner matrix of f at n points is the table of all the chord slopes between them.

Definition 21.9.2. Loewner Matrix

Let $f\colon I \to \mathbb{R}$ be differentiable and let $t_1, \dots, t_n \in I$. The **Loewner matrix** of f at these points is

$$L_f(t_1, \dots, t_n) \coloneqq \left(f[t_i, t_j]\right)_{i,j=1}^{n} \in M_n(\mathbb{R}).$$

Every entry is real and the matrix is **symmetric**, because the divided difference is; its diagonal entries are $f'(t_1), \dots, f'(t_n)$. A real symmetric matrix is Hermitian, so it makes sense to ask whether it is positive semidefinite — the only question this section asks about it. The points need not be distinct; if $t_i = t_j$ then rows i and j coincide.

Examples. All four use the divided differences computed above.

- $f(t) = \alpha + \beta t$: every entry is β, so $L_f = \beta \mathbf{J}$, where $\mathbf{J}$ is the all-ones matrix. Since $\mathbf{J} = \mathbf{1}\mathbf{1}^{\top}$ is positive semidefinite, $L_f \succeq \mathbf{O}$ exactly when $\beta \geq 0$ — which is precisely when the affine function is operator monotone, by Proposition 17.11.2 one way and Proposition 17.11.1 the other.
- $f(t) = -1/t$ on $(0, \infty)$: the (i, j) entry is $1/(t_i t_j)$, so $L_f = \mathbf{d}\mathbf{d}^{\top}$ with $\mathbf{d} = (1/t_1, \dots, 1/t_n)$. A matrix of the form $\mathbf{d}\mathbf{d}^{\top}$ is positive semidefinite, so **every** Loewner matrix of $-1/t$ is $\succeq \mathbf{O}$, at every n. Compare Proposition 17.11.3.
- $f(t) = -1/(t + c)$ with $c \geq 0$: the same computation with $t_i + c$ in place of t_i gives the rank-one matrix with entries $1/((t_i + c)(t_j + c))$, again $\succeq \mathbf{O}$. Compare Corollary 17.11.5.
- $f(t) = \sqrt{t}$ on $(0, \infty)$: the (i, j) entry is $1/(\sqrt{t_i} + \sqrt{t_j})$. A matrix of the form $(1/(x_i + x_j))$ with $x_i > 0$ is called a **Cauchy matrix**.

Nonexample by minimal change. In the first example take $\alpha = 0$ and $\beta = 1$, so that $f(t) = t$ and $L_f = \mathbf{J} \succeq \mathbf{O}$. Raise the exponent from 1 to 2, and positivity is lost at once.

Example. The Loewner Matrix of the Square

Let $f(t) = t^2$ on $[0, \infty)$ and let $s \neq t$ be points of $[0, \infty)$. Compute $L_f(s, t)$ and decide whether it is positive semidefinite.

Solution. Since $f[s,t] = s+t$ and $f'(t) = 2t$,

$$L_f(s,t) = \begin{pmatrix} 2s & s+t \\ s+t & 2t \end{pmatrix}, \qquad \det L_f(s,t) = 4st - (s+t)^2 = -(s-t)^2.$$

The determinant is a principal minor, and it is < 0 because $s \neq t$. By Theorem 13.1.4 (e), $L_f(s,t)$ is not positive semidefinite — it is indefinite, its determinant being negative. So the square already fails at $n = 2$, at **every** pair of distinct points, which is the numerical shadow of Example (Squaring does not preserve the Loewner order).

Example. A Cauchy Matrix

Let $f(t) = \sqrt{t}$ on $(0,\infty)$ and take the points $1, 4, 9$. Compute $L_f(1,4,9)$ and decide whether it is positive semidefinite.

Solution. The square roots are $1, 2, 3$, and the (i,j) entry is $1/(\sqrt{t_i} + \sqrt{t_j})$, so

$$L_f(1,4,9) = \begin{pmatrix} \frac{1}{2} & \frac{1}{3} & \frac{1}{4} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} \\ \frac{1}{4} & \frac{1}{5} & \frac{1}{6} \end{pmatrix}.$$

The first two leading principal minors are

$$\tfrac{1}{2} \qquad \text{and} \qquad \tfrac{1}{2}\cdot\tfrac{1}{4} - \left(\tfrac{1}{3}\right)^2 = \tfrac{1}{8} - \tfrac{1}{9} = \tfrac{1}{72}.$$

For the third, expand along the first row. The three minors are

$$\tfrac{1}{24} - \tfrac{1}{25} = \tfrac{1}{600}, \qquad \tfrac{1}{18} - \tfrac{1}{20} = \tfrac{1}{180}, \qquad \tfrac{1}{15} - \tfrac{1}{16} = \tfrac{1}{240},$$

so, with the alternating signs,

$$\det L_f(1,4,9) = \tfrac{1}{2}\cdot\tfrac{1}{600} - \tfrac{1}{3}\cdot\tfrac{1}{180} + \tfrac{1}{4}\cdot\tfrac{1}{240} = \tfrac{1}{1200} - \tfrac{1}{540} + \tfrac{1}{960}.$$

Over the common denominator 43200 the three terms are 36, 80, and 45, and $36 - 80 + 45 = 1$, so $\det L_f(1,4,9) = 1/43200$. All three leading principal minors are positive, so $L_f(1,4,9) \succ \mathbf{O}$ by Sylvester's criterion (Theorem 13.1.5 (d)), and in particular $\succeq \mathbf{O}$.

The contrast between the two examples is the point of the section: $\sqrt{t}$ is operator monotone and t^2 is not, and the Loewner matrices see the difference with no matrix inequality tested. The theorem below says this is no coincidence.

Warning. **One size at a time.** A Loewner matrix is built from **one** choice of n points. Positivity for one choice says nothing about another, and positivity at all choices of n points is not by itself a statement about $n+1$ points: $f(t) = t^2$ has $L_f(t) = (2t) \succeq \mathbf{O}$ at every point of $[0,\infty)$, and already at two points it fails (Example (The Loewner Matrix of the Square)). The criterion below is an equivalence *size by size*: it matches the $n \times n$ Loewner matrices

with order-preservation on $n \times n$ matrices, and operator monotonicity is the conjunction of all of these.

Since the criterion works one size at a time, the property it characterizes needs a name.

Definition 21.9.3. Monotone of Order n

Let $n \geq 1$ and let $f\colon I \to \mathbb{R}$. Then f is **monotone of order n on I** if, for all Hermitian $\mathbf{A}, \mathbf{B} \in M_n(F)$ whose spectra lie in I,

$$\mathbf{A} \succeq \mathbf{B} \quad \Longrightarrow \quad f(\mathbf{A}) \succeq f(\mathbf{B}).$$

Comparing with Definition 17.11.1 (a): **f is operator monotone on I exactly when it is monotone of order n for every $n \geq 1$.** Order 1 is plain monotonicity (Proposition 17.11.1), and order 2 is already a strictly stronger demand: $t \mapsto t^2$ is increasing on $[0, \infty)$ and fails at $n = 2$ (Example (Squaring does not preserve the Loewner order)). Whether each further order is in turn strictly stronger than the one before is not decided here.

21.9.3 Differentiating a matrix function

The link between the Loewner matrix and the order is a derivative, and the derivative comes out of one exact algebraic identity. Write both matrices in their spectral resolutions, insert the identity on both sides, and the difference of the two functional calculi factors through the difference of the matrices, with divided differences as the coefficients.

Lemma 21.9.2. The Two-Resolution Identity

Let $f\colon I \to \mathbb{R}$ be differentiable and let $\mathbf{A}, \mathbf{B} \in M_n(F)$ be Hermitian with spectra in I. Write the spectral resolutions

$$\mathbf{A} = \sum_i \alpha_i \mathbf{P}_i, \qquad \mathbf{B} = \sum_k \beta_k \mathbf{Q}_k$$

with α_i the distinct eigenvalues of $\mathbf{A}$ and β_k those of $\mathbf{B}$ (Theorem 12.8.1). Then

$$f(\mathbf{B}) - f(\mathbf{A}) = \sum_{i,k} f[\alpha_i, \beta_k]\, \mathbf{P}_i (\mathbf{B} - \mathbf{A}) \mathbf{Q}_k.$$

Idea. The resolutions of the identity $\sum_i \mathbf{P}_i = \mathbf{I}$ and $\sum_k \mathbf{Q}_k = \mathbf{I}$ can be inserted on the two sides of any matrix, which splits it into $i \times k$ pieces. On the piece $\mathbf{P}_i \cdot \mathbf{Q}_k$, the matrix $\mathbf{A}$ acts as the number α_i from the left and $\mathbf{B}$ acts as β_k from the right. So both $f(\mathbf{B}) - f(\mathbf{A})$ and $\mathbf{B} - \mathbf{A}$ become numbers on each piece, namely $f(\beta_k) - f(\alpha_i)$ and $\beta_k - \alpha_i$, and the ratio of those two numbers is the divided difference.

Proof. By Theorem 12.8.1 (c), $\sum_i \mathbf{P}_i = \mathbf{I}$ and $\sum_k \mathbf{Q}_k = \mathbf{I}$. By Theorem 12.8.1 (e), $\mathbf{P}_i \mathbf{A} = \alpha_i \mathbf{P}_i$ and $\mathbf{B}\mathbf{Q}_k = \beta_k \mathbf{Q}_k$; and $\mathbf{P}_i f(\mathbf{A}) = f(\alpha_i)\mathbf{P}_i$, $f(\mathbf{B})\mathbf{Q}_k = f(\beta_k)\mathbf{Q}_k$ follow from

Definition 12.8.2 together with Theorem 12.8.1 (b). Hence

$$\begin{aligned} f(\mathbf{B}) - f(\mathbf{A}) &= \sum_{i,k} \mathbf{P}_i (f(\mathbf{B}) - f(\mathbf{A})) \mathbf{Q}_k \\ &= \sum_{i,k} (f(\beta_k) - f(\alpha_i)) \mathbf{P}_i \mathbf{Q}_k, \end{aligned}$$

and in the same way

$$\mathbf{P}_i (\mathbf{B} - \mathbf{A}) \mathbf{Q}_k = (\beta_k - \alpha_i)\, \mathbf{P}_i \mathbf{Q}_k.$$

Fix i and k. If $\alpha_i \neq \beta_k$, then by Definition 21.9.1

$$(f(\beta_k) - f(\alpha_i)) \mathbf{P}_i \mathbf{Q}_k = f[\alpha_i, \beta_k](\beta_k - \alpha_i) \mathbf{P}_i \mathbf{Q}_k = f[\alpha_i, \beta_k]\, \mathbf{P}_i (\mathbf{B} - \mathbf{A}) \mathbf{Q}_k.$$

If $\alpha_i = \beta_k$, then both $f(\beta_k) - f(\alpha_i)$ and $\beta_k - \alpha_i$ are 0, so both sides of that equation are the zero matrix and it holds as well. Summing over i and k gives the identity.

■

The identity is exact, with no limit in it. To turn it into a derivative we let $\mathbf{B}$ approach $\mathbf{A}$, which needs the functional calculus to be continuous. It is, and the proof shows what a mixed pair of eigenbases is for.

Lemma 21.9.3. Continuity of the Functional Calculus

Let $J = [a, b]$ be a closed bounded interval, let $g\colon J \to \mathbb{R}$ be continuous, and let $n \geq 1$. For every $\varepsilon > 0$ there is $\eta > 0$ such that all Hermitian $\mathbf{X}, \mathbf{Y} \in M_n(F)$ with spectra in J and $\|\mathbf{X} - \mathbf{Y}\|_F < \eta$ satisfy

$$\|g(\mathbf{X}) - g(\mathbf{Y})\|_F < \varepsilon.$$

Idea. Diagonalize $\mathbf{X}$ and $\mathbf{Y}$ separately, and read both $g(\mathbf{X}) - g(\mathbf{Y})$ and $\mathbf{X} - \mathbf{Y}$ in the **mixed** basis, one eigenbasis on the left and the other on the right. In that basis the two matrices have entries $(g(a_i) - g(b_k))w_{ik}$ and $(a_i - b_k)w_{ik}$ with the same w_{ik}, so they differ only in how the scalar factor is built. Where a_i and b_k are close, continuity of g makes the first factor small; where they are far apart, the second factor is large and can pay for the first. The Frobenius norm is the right yardstick because it is computed from the entries and is unchanged by a unitary factor on either side.

Proof. By the extreme value theorem (A4), g is bounded on the compact interval J; let $M = \max_{t \in J} |g(t)|$. If $M = 0$, then g is identically 0 and there is nothing to prove, so assume $M > 0$.

Step 1: g is uniformly continuous on J. Suppose not. Then there is $\varepsilon_0 > 0$ such that for every $k \geq 1$ there are $a_k, b_k \in J$ with $|a_k - b_k| \leq 1/k$ and $|g(a_k) - g(b_k)| \geq \varepsilon_0$. By compactness (A3), a subsequence a_{k_j} converges to some $a \in J$; then $b_{k_j} \to a$ too, since $|b_{k_j} - a_{k_j}| \to 0$. Continuity of g at a gives $g(a_{k_j}) \to g(a)$ and $g(b_{k_j}) \to g(a)$, so $|g(a_{k_j}) - g(b_{k_j})| \to 0$, contradicting the lower bound ε_0.

Step 2: the mixed basis. Let $\mathbf{X}$ and $\mathbf{Y}$ be Hermitian with spectra in J. By Corollary 12.4.3 over $\mathbb{C}$, or Corollary 12.5.2 over $\mathbb{R}$, there are unitary $\mathbf{U}, \mathbf{V} \in M_n(F)$ whose columns $\mathbf{u}_1, \ldots, \mathbf{u}_n$

and $\mathbf{v}_1, \dots, \mathbf{v}_n$ are orthonormal eigenbases, with $\mathbf{X}\mathbf{u}_i = a_i\mathbf{u}_i$ and $\mathbf{Y}\mathbf{v}_k = b_k\mathbf{v}_k$, all $a_i, b_k \in J$. Put $w_{ik} = \mathbf{u}_i^*\mathbf{v}_k$, the entries of the unitary matrix $\mathbf{W} = \mathbf{U}^*\mathbf{V}$; then $\sum_{i,k}|w_{ik}|^2 = \|\mathbf{W}\|_F^2 = n$.

Write $\mathbf{X} = \sum_j \mu_j \mathbf{P}_j$ for the spectral resolution. Exactly one μ_j equals a_i, and $\mathbf{u}_i$ lies in its eigenspace, so $\mathbf{P}_j\mathbf{u}_i = \mathbf{u}_i$ for that j and $\mathbf{P}_j\mathbf{u}_i = \mathbf{0}$ for the others, the eigenspaces being orthogonal (Theorem 12.8.1 (a)). Hence $g(\mathbf{X})\mathbf{u}_i = g(a_i)\mathbf{u}_i$ by Definition 12.8.2, and likewise $g(\mathbf{Y})\mathbf{v}_k = g(b_k)\mathbf{v}_k$; since $g(a_i)$ is real and $g(\mathbf{X})$ is Hermitian, also $\mathbf{u}_i^* g(\mathbf{X}) = g(a_i)\mathbf{u}_i^*$. Hence the (i,k) entry of $\mathbf{U}^*(g(\mathbf{X}) - g(\mathbf{Y}))\mathbf{V}$ is

$$\mathbf{u}_i^*(g(\mathbf{X}) - g(\mathbf{Y}))\mathbf{v}_k = (g(a_i) - g(b_k))w_{ik},$$

and the same computation with g replaced by the identity function gives $\mathbf{u}_i^*(\mathbf{X} - \mathbf{Y})\mathbf{v}_k = (a_i - b_k)w_{ik}$. By Lemma 11.7.5,

$$\|g(\mathbf{X}) - g(\mathbf{Y})\|_F^2 = \sum_{i,k}|g(a_i) - g(b_k)|^2|w_{ik}|^2,$$
$$\|\mathbf{X} - \mathbf{Y}\|_F^2 = \sum_{i,k}|a_i - b_k|^2|w_{ik}|^2.$$

Step 3: the split. Let $\varepsilon > 0$. By Step 1 there is $\delta > 0$ with $|g(a) - g(b)| \le \varepsilon/(2\sqrt{n})$ whenever $a, b \in J$ and $|a - b| < \delta$. Split the index pairs (i,k) into those with $|a_i - b_k| < \delta$ and those with $|a_i - b_k| \ge \delta$. The first group contributes at most

$$\frac{\varepsilon^2}{4n}\sum_{i,k}|w_{ik}|^2 = \frac{\varepsilon^2}{4n}\cdot n = \frac{\varepsilon^2}{4}.$$

In the second group $|g(a_i) - g(b_k)| \le 2M \le (2M/\delta)|a_i - b_k|$, so that group contributes at most $(2M/\delta)^2\|\mathbf{X} - \mathbf{Y}\|_F^2$. Therefore

$$\|g(\mathbf{X}) - g(\mathbf{Y})\|_F^2 \;\le\; \frac{\varepsilon^2}{4} + \Big(\frac{2M}{\delta}\Big)^2\|\mathbf{X} - \mathbf{Y}\|_F^2.$$

Take $\eta = \varepsilon\delta/(4M) > 0$. If $\|\mathbf{X} - \mathbf{Y}\|_F < \eta$, the second term is less than $\varepsilon^2/4$, so $\|g(\mathbf{X}) - g(\mathbf{Y})\|_F^2 < \varepsilon^2/2 < \varepsilon^2$. This proves the lemma.

■

Now the theorem. With the two lemmas in hand it is bookkeeping.

Theorem 21.9.4. The Derivative of a Matrix Function

Let $f \in C^1(I)$, let $\mathbf{A}, \mathbf{H} \in M_n(F)$ be Hermitian, and suppose there are $s_0 > 0$ and a closed bounded interval $J \subseteq I$ with

$$\operatorname{spec}(\mathbf{A} + s\mathbf{H}) \subseteq J \qquad \text{for every real } s \text{ with } |s| < s_0.$$

Then the matrix-valued function $\mathbf{G}(s) := f(\mathbf{A} + s\mathbf{H})$, defined for $|s| < s_0$, is differentiable at $s = 0$ in the sense of Definition 16.6.1, and:

(a) If $\mathbf{A} = \operatorname{diag}(t_1, \dots, t_n)$, then

$$\mathbf{G}'(0) = L_f(t_1, \dots, t_n) \circ \mathbf{H},$$

the Hadamard product of Definition 13.8.1.

(b) In general, if $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{U}^*$ with $\mathbf{U}$ unitary and $\mathbf{D} = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$, then

$$\mathbf{G}'(0) = \mathbf{U}\Big(L_f(\lambda_1, \dots, \lambda_n) \circ (\mathbf{U}^*\mathbf{H}\mathbf{U})\Big)\mathbf{U}^*.$$

Remark. The hypothesis on J is automatic when I is **open**. Write $I = (c, d)$ with $-\infty \le c < d \le \infty$. Then $c < \lambda_n(\mathbf{A}) \le \lambda_1(\mathbf{A}) < d$, so there is $\rho > 0$ with $[\lambda_n(\mathbf{A}) - \rho, \lambda_1(\mathbf{A}) + \rho] \subseteq I$; taking $s_0 = \rho/(\|\mathbf{H}\|_2 + 1)$ and J that closed interval works, because $|\lambda_i(\mathbf{A} + s\mathbf{H}) - \lambda_i(\mathbf{A})| \le |s|\|\mathbf{H}\|_2 < \rho$ by Corollary 17.3.6.

Idea. Apply Lemma 21.9.2 with $\mathbf{B} = \mathbf{A} + s\mathbf{H}$. The right-hand side is a sum over pairs (eigenvalue of $\mathbf{A}$, eigenvalue of $\mathbf{B}$), and the sum over the second index is nothing but a functional calculus: freezing the first argument of the divided difference turns $f[\alpha_i, \cdot]$ into an ordinary continuous function g_i, and $\sum_k g_i(\beta_k)\mathbf{Q}_k$ is $g_i(\mathbf{B})$. So the identity collapses to

$$f(\mathbf{A} + s\mathbf{H}) - f(\mathbf{A}) = s \sum_i \mathbf{P}_i \, \mathbf{H} \, g_i(\mathbf{A} + s\mathbf{H}),$$

with the factor s already pulled out. Now let $s \to 0$: by Lemma 21.9.3, $g_i(\mathbf{A} + s\mathbf{H}) \to g_i(\mathbf{A})$, and the limit of the sum is the Hadamard product in disguise.

Proof. Write $\mathbf{A} = \sum_i \alpha_i \mathbf{P}_i$ for the spectral resolution, with $\alpha_1, \dots, \alpha_r$ the distinct eigenvalues of $\mathbf{A}$, and put $\mathbf{B}_s = \mathbf{A} + s\mathbf{H}$, so that $\mathbf{G}(s) = f(\mathbf{B}_s)$, for $|s| < s_0$. For each i define

$$g_i \colon J \to \mathbb{R}, \qquad g_i(x) = f[\alpha_i, x],$$

which is continuous by Lemma 21.9.1, since $f \in C^1(I)$ and $J \subseteq I$.

Step 1: an exact formula. Let $\mathbf{B}_s = \sum_k \beta_k(s)\mathbf{Q}_k(s)$ be the spectral resolution of $\mathbf{B}_s$. By Lemma 21.9.2 with $\mathbf{B} = \mathbf{B}_s$, and $\mathbf{B}_s - \mathbf{A} = s\mathbf{H}$,

$$\begin{aligned} f(\mathbf{B}_s) - f(\mathbf{A}) &= \sum_{i,k} f[\alpha_i, \beta_k(s)] \; \mathbf{P}_i(s\mathbf{H})\mathbf{Q}_k(s) \\ &= s \sum_i \mathbf{P}_i \mathbf{H} \Big(\sum_k g_i(\beta_k(s))\mathbf{Q}_k(s) \Big) = s \sum_i \mathbf{P}_i \, \mathbf{H} \, g_i(\mathbf{B}_s), \end{aligned}$$

the last equality being Definition 12.8.2 applied to g_i and $\mathbf{B}_s$, whose spectrum lies in J, the domain of g_i.

Step 2: the limit. Since $\|\mathbf{B}_s - \mathbf{A}\|_F = |s| \, \|\mathbf{H}\|_F \to 0$ as $s \to 0$, and since $\mathbf{A}$ and every $\mathbf{B}_s$ are Hermitian with spectra in J, Lemma 21.9.3 gives $\|g_i(\mathbf{B}_s) - g_i(\mathbf{A})\|_F \to 0$ for each i. By Definition 12.8.2,

$$g_i(\mathbf{A}) = \sum_j g_i(\alpha_j)\mathbf{P}_j = \sum_j f[\alpha_i, \alpha_j] \, \mathbf{P}_j.$$

Dividing the formula of Step 1 by $s \neq 0$ and letting $s \to 0$, the algebra of limits (applied

entry by entry, and using $|m_{pq}| \le \|\mathbf{M}\|_F$ for any matrix) gives

$$\frac{\mathbf{G}(s) - \mathbf{G}(0)}{s} \longrightarrow \sum_{i,j} f[\alpha_i, \alpha_j]\, \mathbf{P}_i \mathbf{H} \mathbf{P}_j .$$

Entrywise, this says exactly that each entry of $\mathbf{G}$ is differentiable at $s = 0$, with derivative the corresponding entry of the right-hand side. So $\mathbf{G}'(0)$ exists in the sense of Definition 16.6.1 and equals $\sum_{i,j} f[\alpha_i, \alpha_j] \mathbf{P}_i \mathbf{H} \mathbf{P}_j$.

Step 3: reading it as a Hadamard product. Let $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{U}^*$ with $\mathbf{U}$ unitary, $\mathbf{D} = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$, and let $\mathbf{u}_1, \dots, \mathbf{u}_n$ be the columns of $\mathbf{U}$, so $\mathbf{A}\mathbf{u}_p = \lambda_p \mathbf{u}_p$. Each $\mathbf{P}_i$ is the orthogonal projection onto the eigenspace of α_i, which is spanned by those $\mathbf{u}_p$ with $\lambda_p = \alpha_i$; hence $\mathbf{P}_i = \sum_{p : \lambda_p = \alpha_i} \mathbf{u}_p \mathbf{u}_p^*$, and these sums partition $\{1, \dots, n\}$. Therefore

$$\sum_{i,j} f[\alpha_i, \alpha_j] \mathbf{P}_i \mathbf{H} \mathbf{P}_j = \sum_{p,q} f[\lambda_p, \lambda_q]\, \mathbf{u}_p \mathbf{u}_p^* \mathbf{H} \mathbf{u}_q \mathbf{u}_q^* .$$

Now $\mathbf{u}_p^* \mathbf{H} \mathbf{u}_q = (\mathbf{U}^* \mathbf{H} \mathbf{U})_{pq}$, and for any $\mathbf{M} \in M_n(F)$ we have $\mathbf{U}\mathbf{M}\mathbf{U}^* = \sum_{p,q} m_{pq} \mathbf{u}_p \mathbf{u}_q^*$. Taking $\mathbf{M} = L_f(\lambda_1, \dots, \lambda_n) \circ (\mathbf{U}^* \mathbf{H} \mathbf{U})$, whose (p,q) entry is $f[\lambda_p, \lambda_q](\mathbf{U}^* \mathbf{H} \mathbf{U})_{pq}$, the displayed sum is $\mathbf{U}\mathbf{M}\mathbf{U}^*$. This proves (b), and (a) is the case $\mathbf{U} = \mathbf{I}$.

■

Remark. Two sanity checks. For $f(t) = t$ the Loewner matrix is $\mathbf{J}$, the identity for $\circ$, and the formula returns $\mathbf{H}$ — as it must, since $f(\mathbf{A} + s\mathbf{H}) = \mathbf{A} + s\mathbf{H}$. And the diagonal entries of (a) are $f'(t_i) h_{ii}$: to first order the i-th eigenvalue of $\operatorname{diag}(t_1, \dots, t_n) + s\mathbf{H}$ moves by h_{ii} when the t_i are distinct (Theorem 20.11.2, as computed in Chapter 20 §11), and the chain rule then moves f of it by $f'(t_i) h_{ii}$. What the theorem adds is the **off-diagonal** part, where the eigenvectors turn, and there the chord slope $f[t_i, t_j]$ appears in place of the tangent slope.

Quick check. Take $n = 2$, $\mathbf{A} = \operatorname{diag}(t_1, t_2)$ with $t_1 \neq t_2$, and $\mathbf{H} = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$. What does Theorem 21.9.4 (a) predict, and what does it say about the eigenvalues of $\mathbf{A} + s\mathbf{H}$ to first order?

Solution. The Hadamard product with $\mathbf{H}$ keeps only the off-diagonal entries, so the derivative is $\left(\begin{smallmatrix} 0 & f[t_1,t_2] \\ f[t_1,t_2] & 0 \end{smallmatrix}\right)$. Its diagonal is zero, which matches the eigenvalue statement: $h_{11} = h_{22} = 0$, so to first order neither eigenvalue of $\mathbf{A} + s\mathbf{H}$ moves. All of the first-order change in $f(\mathbf{A} + s\mathbf{H})$ is in the eigenvectors, and the coefficient that measures it is the chord slope between t_1 and t_2, not a derivative.

21.9.4 The criterion

Everything is now in place. One direction walks along the segment from $\mathbf{B}$ to $\mathbf{A}$, using the Schur product theorem to keep the derivative positive semidefinite; the other reads the derivative off in a single well-chosen direction.

Theorem 21.9.5. Loewner's Matrix Criterion

Let I be an interval with more than one point, let $f \in C^1(I)$ and let $n \geq 1$. The following are equivalent.

(i) f is monotone of order n on I.
(ii) $L_f(t_1, \ldots, t_n) \succeq \mathbf{O}$ for all $t_1, \ldots, t_n \in I$.

Idea. For (ii) $\Rightarrow$ (i): walk from $\mathbf{B}$ to $\mathbf{A}$ along the segment $\mathbf{B}_s = \mathbf{B} + s(\mathbf{A} - \mathbf{B})$, and watch the number $\varphi(s) = \langle f(\mathbf{B}_s)\mathbf{x}, \mathbf{x}\rangle$ for a fixed $\mathbf{x}$. Its derivative is, in the eigenbasis of $\mathbf{B}_s$, the quadratic form of a Hadamard product of two positive semidefinite matrices: the Loewner matrix, positive by hypothesis, and the direction $\mathbf{A} - \mathbf{B}$, positive because $\mathbf{A} \succeq \mathbf{B}$. The Schur product theorem says that such a product is positive semidefinite, so $\varphi' \geq 0$, and the mean value theorem does the rest.

For (i) $\Rightarrow$ (ii): the derivative formula turns order-preservation into $L_f \circ \mathbf{H} \succeq \mathbf{O}$ for every $\mathbf{H} \succeq \mathbf{O}$. Choose the one $\mathbf{H}$ that leaves L_f alone, namely the all-ones matrix $\mathbf{J} = \mathbf{1}\mathbf{1}^\top$, which is positive semidefinite and is the identity for $\circ$.

Proof. **(ii) $\Rightarrow$ (i).** Let $\mathbf{A}, \mathbf{B} \in M_n(F)$ be Hermitian with spectra in I and $\mathbf{A} \succeq \mathbf{B}$. Put $\mathbf{H} = \mathbf{A} - \mathbf{B} \succeq \mathbf{O}$ and $\mathbf{B}_s = \mathbf{B} + s\mathbf{H}$ for $s \in [0, 1]$. For such s, both $\mathbf{A} - \mathbf{B}_s = (1 - s)\mathbf{H}$ and $\mathbf{B}_s - \mathbf{B} = s\mathbf{H}$ are $\succeq \mathbf{O}$, so $\mathbf{A} \succeq \mathbf{B}_s \succeq \mathbf{B}$ and Corollary 17.2.4 gives

$$\lambda_i(\mathbf{B}) \;\leq\; \lambda_i(\mathbf{B}_s) \;\leq\; \lambda_i(\mathbf{A}) \qquad (i = 1, \ldots, n).$$

Hence $\operatorname{spec}(\mathbf{B}_s) \subseteq J := [\lambda_n(\mathbf{B}), \lambda_1(\mathbf{A})]$ for every $s \in [0, 1]$, and $J \subseteq I$ because its two endpoints are eigenvalues of $\mathbf{B}$ and of $\mathbf{A}$, hence lie in I, and I is an interval.

Fix $\mathbf{x} \in F^n$ and put $\varphi(s) = \langle f(\mathbf{B}_s)\mathbf{x}, \mathbf{x}\rangle$ for $s \in [0, 1]$, a real number because $f(\mathbf{B}_s)$ is Hermitian (Theorem 12.8.3 (c)).

φ is continuous on $[0, 1]$. We have $\|\mathbf{B}_s - \mathbf{B}_{s'}\|_F = |s - s'|\|\mathbf{H}\|_F$, so Lemma 21.9.3 applied to f on J makes $s \mapsto f(\mathbf{B}_s)$ continuous in the Frobenius norm, and $|\langle \mathbf{M}\mathbf{x}, \mathbf{x}\rangle| \leq \|\mathbf{M}\|_F \|\mathbf{x}\|^2$ transfers this to φ.

$\varphi'(s) \geq 0$ for $0 < s < 1$. Fix such an s and put $s_0 = \min(s, 1 - s) > 0$. For $|\sigma| < s_0$ we have $\mathbf{B}_s + \sigma\mathbf{H} = \mathbf{B}_{s+\sigma}$ with $s + \sigma \in (0, 1)$, so its spectrum lies in J. Thus Theorem 21.9.4 applies to the pair $\mathbf{B}_s, \mathbf{H}$. Write $\mathbf{B}_s = \mathbf{U}\mathbf{D}\mathbf{U}^*$ with $\mathbf{U}$ unitary and $\mathbf{D} = \operatorname{diag}(\lambda_1(\mathbf{B}_s), \ldots, \lambda_n(\mathbf{B}_s))$, and set

$$\mathbf{C} = L_f(\lambda_1(\mathbf{B}_s), \ldots, \lambda_n(\mathbf{B}_s)), \qquad \mathbf{M} = \mathbf{U}^*\mathbf{H}\mathbf{U}.$$

Then $\mathbf{C} \succeq \mathbf{O}$ by hypothesis (ii), since the $\lambda_i(\mathbf{B}_s)$ lie in $J \subseteq I$, and $\mathbf{M} \succeq \mathbf{O}$ by Proposition 13.1.2 (a). By the Schur product theorem (Theorem 13.8.2 (a)), $\mathbf{C} \circ \mathbf{M} \succeq \mathbf{O}$. A quadratic form $\langle \mathbf{X}\mathbf{x}, \mathbf{x}\rangle$ is a fixed linear combination of the entries of $\mathbf{X}$, so an entrywise derivative of a matrix-valued function differentiates it; hence φ is differentiable at s with

$$\varphi'(s) = \langle \mathbf{U}(\mathbf{C} \circ \mathbf{M})\mathbf{U}^*\mathbf{x}, \mathbf{x}\rangle = \langle (\mathbf{C} \circ \mathbf{M})\mathbf{y}, \mathbf{y}\rangle \;\geq\; 0,$$

where $\mathbf{y} = \mathbf{U}^*\mathbf{x}$.

Conclusion. Let $0 < a < b < 1$. By the mean value theorem (A6) applied to φ on $(0,1)$, there is c between a and b with $\varphi(b) - \varphi(a) = \varphi'(c)(b-a) \geq 0$. Letting $a \to 0^+$ and $b \to 1^-$ and using the continuity of φ on $[0,1]$, we get $\varphi(1) - \varphi(0) \geq 0$, that is

$$\langle f(\mathbf{A})\mathbf{x}, \mathbf{x}\rangle \;\geq\; \langle f(\mathbf{B})\mathbf{x}, \mathbf{x}\rangle.$$

As $\mathbf{x}$ was arbitrary, $f(\mathbf{A}) \succeq f(\mathbf{B})$ by Definition 13.5.1.

(i) $\Rightarrow$ (ii). Suppose first that $t_1, \dots, t_n$ all lie in the interior of I, an open interval (c,d). Put

$$\mathbf{A} = \operatorname{diag}(t_1, \dots, t_n), \qquad \mathbf{H} = \mathbf{J} = \mathbf{1}\mathbf{1}^{\top},$$

so $\mathbf{H} \succeq \mathbf{O}$. Let $\rho > 0$ be small enough that $[\min_i t_i - \rho, \max_i t_i + \rho] \subseteq (c,d)$; this is possible because $c < \min_i t_i \leq \max_i t_i < d$. Put $s_0 = \rho/(\|\mathbf{H}\|_2 + 1)$ and $J = [\min_i t_i - \rho, \max_i t_i + \rho]$. By Corollary 17.3.6, $\operatorname{spec}(\mathbf{A} + s\mathbf{H}) \subseteq J \subseteq I$ for $|s| < s_0$, so Theorem 21.9.4 applies.

For $0 < s < s_0$ we have $(\mathbf{A} + s\mathbf{H}) - \mathbf{A} = s\mathbf{H} \succeq \mathbf{O}$, and both matrices have spectra in I, so (i) gives $f(\mathbf{A} + s\mathbf{H}) \succeq f(\mathbf{A})$, that is

$$\frac{f(\mathbf{A} + s\mathbf{H}) - f(\mathbf{A})}{s} \;\succeq\; \mathbf{O}.$$

For each fixed $\mathbf{x}$ the number $\langle \cdot\,\mathbf{x}, \mathbf{x}\rangle$ of the left-hand side is ≥ 0, and it converges as $s \to 0^+$ to the quadratic form of the derivative, by Theorem 21.9.4 (a); a nonstrict inequality survives a limit, so the derivative is $\succeq \mathbf{O}$. But the derivative is $L_f(t_1, \dots, t_n) \circ \mathbf{J} = L_f(t_1, \dots, t_n)$, because every entry of $\mathbf{J}$ is 1, so $\mathbf{J}$ is the identity for the Hadamard product (Definition 13.8.1). Hence $L_f(t_1, \dots, t_n) \succeq \mathbf{O}$.

Now let $t_1, \dots, t_n$ be arbitrary points of I. Since I has more than one point, it has an interior point m; put $t_i^{(k)} = (1 - 1/k)t_i + (1/k)m$ for $k \geq 1$. Each $t_i^{(k)}$ lies in I, being a convex combination of two of its points, and it is an **interior** point: if $(m - \varepsilon, m + \varepsilon) \subseteq I$, then every point within ε/k of $t_i^{(k)}$ is of the form $(1 - 1/k)t_i + (1/k)(m + u)$ with $|u| < \varepsilon$, hence lies in I. Also $t_i^{(k)} \to t_i$. By the case just proved, $L_f(t_1^{(k)}, \dots, t_n^{(k)}) \succeq \mathbf{O}$ for every k, and by Lemma 21.9.1 each entry converges to the corresponding entry of $L_f(t_1, \dots, t_n)$. For each $\mathbf{x}$, the quadratic forms are ≥ 0 and converge to the quadratic form of $L_f(t_1, \dots, t_n)$, which is therefore ≥ 0. This proves the theorem.

■

Remark. The hypothesis $f \in C^1(I)$ enters in three places: it gives the diagonal entries $f'(t_i)$ of the Loewner matrix a meaning, it makes Lemma 21.9.1 available, and through that lemma it supplies the continuous functions g_i that the derivative formula needs. Whether it can be dropped is not settled here: a monotone-of-order-n function is not obviously differentiable. The next section returns to that point for operator monotone functions, where the answer turns out to be yes, at the price of an imported theorem.

Taking $n = 1$ recovers Proposition 17.11.1: $L_f(t) = (f'(t))$, so monotone of order 1 means $f' \geq 0$, which for a C^1 function is the same as increasing, by the mean value theorem (A6). The first genuinely new case is $n = 2$, and there the criterion is two inequalities about numbers.

Corollary 21.9.6. The Criterion at Order Two

Let I be an interval with more than one point and let $f \in C^1(I)$. Then f is monotone of order 2 on I if and only if

$$f'(t) \geq 0 \ \text{ for all } t \in I, \qquad \text{and} \qquad f[s,t]^2 \leq f'(s)f'(t) \ \text{ for all } s,t \in I.$$

Proof. By Theorem 21.9.5 with $n = 2$, monotonicity of order 2 is equivalent to $L_f(s,t) \succeq \mathbf{O}$ for all $s,t \in I$. The matrix

$$L_f(s,t) = \begin{pmatrix} f'(s) & f[s,t] \\ f[s,t] & f'(t) \end{pmatrix}$$

is real symmetric, and its principal minors are the two diagonal entries $f'(s), f'(t)$ and the determinant $f'(s)f'(t) - f[s,t]^2$. By Theorem 13.1.4 (e), $L_f(s,t) \succeq \mathbf{O}$ if and only if all three are ≥ 0. Ranging over all $s,t \in I$, these are exactly the two displayed conditions; the case $s = t$ of the second is the tautology $f'(t)^2 \leq f'(t)^2$.

■

In words: f **increases, and the slope of every chord is at most the geometric mean of the slopes of the tangents at its ends.** Three of Chapter 17 §11's examples can now be re-derived, and its nonexample refuted, without touching a matrix.

- $f(t) = -1/t$ on $(0,\infty)$: $f'(t) = 1/t^2 \geq 0$, and $f[s,t] = 1/(st)$, so $f[s,t]^2 = 1/(s^2t^2) = f'(s)f'(t)$. The condition holds with **equality** everywhere, which is the numerical form of the fact that L_f has rank one.
- $f(t) = \sqrt{t}$ on $(0,\infty)$: $f'(t) = 1/(2\sqrt{t}) \geq 0$, and $f[s,t] = 1/(\sqrt{s}+\sqrt{t})$, so

$$f'(s)f'(t) - f[s,t]^2 = \frac{(\sqrt{s}+\sqrt{t})^2 - 4\sqrt{st}}{4\sqrt{st}\,(\sqrt{s}+\sqrt{t})^2} = \frac{(\sqrt{s}-\sqrt{t})^2}{4\sqrt{st}\,(\sqrt{s}+\sqrt{t})^2} \ \geq \ 0.$$

- $f(t) = \alpha + \beta t$: $f' = \beta$ and $f[s,t] = \beta$, so the conditions are $\beta \geq 0$ and $\beta^2 \leq \beta^2$.
- $f(t) = t^2$ on $[0,\infty)$: $f' = 2t \geq 0$, but $f[s,t]^2 - f'(s)f'(t) = (s+t)^2 - 4st = (s-t)^2 > 0$ for $s \neq t$. So t^2 is **not** monotone of order 2, and therefore not operator monotone. This is Example (Squaring does not preserve the Loewner order) again, now with the failure located at every pair of distinct points rather than at one pair of matrices.

Warning. **Order two is not "increasing and concave".** The second condition of Corollary 21.9.6 looks like a concavity statement and is not one, in either direction. The function $f(t) = t - t^2/2$ on $(0,1)$ is increasing ($f' = 1 - t > 0$) and concave, yet

$$f'(s)f'(t) - f[s,t]^2 = (1-s)(1-t) - \Big(1 - \frac{s+t}{2}\Big)^2 = st - \Big(\frac{s+t}{2}\Big)^2,$$

which equals $-\big((s-t)/2\big)^2$, negative for $s \neq t$; so f is not monotone of order 2. In the other direction, $g(t) = -\log(1-t)$ on $(0,1)$ is increasing and **convex** and *is* monotone of order 2. Writing $a = 1-s$ and $b = 1-t$, so that $s - t = b - a$, we have $g'(s)g'(t) = 1/(ab)$ and $g[s,t] = (\log b - \log a)/(b-a)$, so the second condition reads

$$(\log b - \log a)^2 \ \leq \ \frac{(b-a)^2}{ab},$$

an inequality about the logarithm alone; Exercise (C2: The logarithm at order two) proves it.

21.9.5 What this gives, and what it does not

The criterion is an equivalence, so it reads either way. Right to left it manufactures operator monotone functions: exhibit positive semidefinite Loewner matrices at every order and the function preserves the order at every size. Left to right it refutes them cheaply, since one bad pair of points kills order 2 and hence operator monotonicity.

What it does not give is the classification. The conditions "$L_f(t_1, \dots, t_n) \succeq \mathbf{O}$ for all n and all points" are still infinitely many, and turning them into the integral representation of Theorem 17.11.9 is the analytic half of Loewner's theorem — the business of the next section, and the one place where the chapter spends an imported theorem. What is proved here needs no analysis beyond three items of the chapter's list: the mean value theorem (A6), which turns a nonnegative derivative into monotonicity along the segment; and, inside Lemma 21.9.3, the extreme value theorem (A4), which bounds g, and the compactness of a closed bounded interval (A3), which makes g uniformly continuous.

21.9.6 Exercises

21.9.6.1 A. Check your understanding

Exercise. A1

(a) Define $f[s, t]$ for $s = t$, and say what hypothesis on f the definition needs.

(b) Write down $L_f(t_1, t_2, t_3)$ in terms of divided differences, and say why it is Hermitian.

(c) State Theorem 21.9.5, and say which of its two directions uses the Schur product theorem.

(d) Decide whether the following is correct, and justify your answer: if $L_f(t_1, t_2) \succeq \mathbf{O}$ for all $t_1, t_2 \in I$, then f is operator monotone on I.

(e) What does Theorem 21.9.4 give when $\mathbf{H}$ is diagonal?

Solution.

(a) $f[t, t] = f'(t)$ (Definition 21.9.1); the definition needs f to be differentiable on I, and the continuity statement Lemma 21.9.1 needs f' continuous as well.

(b) It is the 3×3 matrix whose (i, j) entry is $f[t_i, t_j]$, with diagonal $f'(t_1), f'(t_2), f'(t_3)$. It is real and symmetric, because f is real-valued and $f[s, t] = f[t, s]$; a real symmetric matrix is Hermitian.

(c) For $f \in C^1(I)$ and $n \geq 1$: f is monotone of order n on I if and only if $L_f(t_1, \dots, t_n) \succeq \mathbf{O}$ for all $t_1, \dots, t_n \in I$. The Schur product theorem (Theorem 13.8.2) is used in the direction (ii) $\Rightarrow$ (i), to show that the derivative along the segment is positive semidefinite.

(d) Incorrect. Positivity of all 2×2 Loewner matrices gives monotonicity of order 2 only (Corollary 21.9.6); operator monotonicity demands every order. The warning after Example (A Cauchy Matrix) makes the point, and Definition 17.11.1 (a) quantifies over all n.

(e) If $\mathbf{H} = \operatorname{diag}(h_1, \dots, h_n)$ and $\mathbf{A} = \operatorname{diag}(t_1, \dots, t_n)$, then $L_f \circ \mathbf{H}$ is the diagonal matrix $\operatorname{diag}(f'(t_1)h_1, \dots, f'(t_n)h_n)$: the two matrices commute, the problem is scalar in each coordinate, and the answer is the ordinary chain rule n times.

21.9.6.2 B. Practice

Exercise. B1: Loewner matrices by hand

Compute $L_f(1, 2)$ for each of the following, and determine whether it is positive semidefinite. Justify your answer.

(i) $f(t) = 4 + 3t$ on $\mathbb{R}$.
(ii) $f(t) = t^2$ on $(0, \infty)$.
(iii) $f(t) = -1/t$ on $(0, \infty)$.
(iv) $f(t) = \sqrt{t}$ on $(0, \infty)$, at the points 1 and 4 instead.

Solution.

(i) Every entry is 3, so $L_f(1, 2) = 3\mathbf{J} = \left(\begin{smallmatrix} 3 & 3 \\ 3 & 3 \end{smallmatrix}\right)$. Its principal minors are $3, 3$ and 0, all ≥ 0, so it is $\succeq \mathbf{O}$ (Theorem 13.1.4 (e)).

(ii) $f'(t) = 2t$ and $f[1, 2] = 3$, so $L_f(1, 2) = \left(\begin{smallmatrix} 2 & 3 \\ 3 & 4 \end{smallmatrix}\right)$, with determinant $8 - 9 = -1 < 0$. Not positive semidefinite.

(iii) $f'(t) = 1/t^2$ and $f[1, 2] = 1/2$, so $L_f(1, 2) = \left(\begin{smallmatrix} 1 & 1/2 \\ 1/2 & 1/4 \end{smallmatrix}\right)$, with determinant $\frac{1}{4} - \frac{1}{4} = 0$ and diagonal entries > 0. Positive semidefinite, of rank one; indeed it is $\mathbf{d}\mathbf{d}^\top$ with $\mathbf{d} = (1, \frac{1}{2})$.

(iv) $f'(t) = 1/(2\sqrt{t})$ and $f[1, 4] = 1/(1 + 2) = \frac{1}{3}$, so $L_f(1, 4) = \left(\begin{smallmatrix} 1/2 & 1/3 \\ 1/3 & 1/4 \end{smallmatrix}\right)$, with determinant $\frac{1}{8} - \frac{1}{9} = \frac{1}{72} > 0$ and positive diagonal. Positive definite, and in particular $\succeq \mathbf{O}$; it is the top-left corner of the matrix of Example (A Cauchy Matrix).

Exercise. B2: A derivative, two ways

Let $\mathbf{A} = \operatorname{diag}(1, 4)$, let $\mathbf{H} = \left(\begin{smallmatrix} 1 & 2 \\ 2 & 3 \end{smallmatrix}\right)$, and let $f(t) = t^2$ on $(0, \infty)$.

(a) Compute $L_f(1, 4) \circ \mathbf{H}$.
(b) Expand $(\mathbf{A} + s\mathbf{H})^2$ and differentiate at $s = 0$ directly. Check that the two answers agree.

Solution.

(a) $L_f(1, 4) = \left(\begin{smallmatrix} 2 & 5 \\ 5 & 8 \end{smallmatrix}\right)$, so

$$L_f(1, 4) \circ \mathbf{H} = \begin{pmatrix} 2 \cdot 1 & 5 \cdot 2 \\ 5 \cdot 2 & 8 \cdot 3 \end{pmatrix} = \begin{pmatrix} 2 & 10 \\ 10 & 24 \end{pmatrix}.$$

(b) $(\mathbf{A}+s\mathbf{H})^2 = \mathbf{A}^2 + s(\mathbf{AH}+\mathbf{HA}) + s^2\mathbf{H}^2$, so the derivative at $s=0$ is $\mathbf{AH}+\mathbf{HA}$. Now

$$\mathbf{AH} = \begin{pmatrix} 1 & 2 \\ 8 & 12 \end{pmatrix}, \qquad \mathbf{HA} = \begin{pmatrix} 1 & 8 \\ 2 & 12 \end{pmatrix}, \qquad \mathbf{AH}+\mathbf{HA} = \begin{pmatrix} 2 & 10 \\ 10 & 24 \end{pmatrix},$$

which agrees with (a). In general $(\mathbf{AH}+\mathbf{HA})_{ij} = (t_i + t_j)h_{ij}$ for diagonal $\mathbf{A}$, and $t_i + t_j$ is exactly the divided difference of t^2.

Exercise. B3: Refuting order two

For each function, use Corollary 21.9.6 to decide whether it is monotone of order 2 on the interval given. Justify your answer.

(i) $f(t) = t/(t+1)$ on $(0,\infty)$.
(ii) $f(t) = t^2 - 6t$ on $(0,1)$.
(iii) $f(t) = 2t - 1/t$ on $(0,\infty)$.
(iv) $f(t) = t + t^2$ on $(0,\infty)$.

Solution.

(i) Monotone of order 2. Here $f(t) = 1 - 1/(t+1)$, so $f'(t) = 1/(t+1)^2 \geq 0$ and, by the computation for $-1/(t+c)$ with $c=1$, $f[s,t] = 1/((s+1)(t+1))$. Hence $f[s,t]^2 = f'(s)f'(t)$, with equality. (In fact the Loewner matrices of every order are rank one and $\succeq \mathbf{O}$, which recovers Corollary 17.11.5.)

(ii) Not monotone of order 2 — indeed it fails already at order 1, since $f'(t) = 2t - 6 < 0$ on $(0,1)$, so the first condition fails.

(iii) Monotone of order 2. Here $f'(t) = 2 + 1/t^2 > 0$ and $f[s,t] = 2 + 1/(st)$, by additivity of divided differences and the two computations after Definition 21.9.1. So

$$f'(s)f'(t) - f[s,t]^2 = \frac{2}{s^2} + \frac{2}{t^2} - \frac{4}{st} = 2\Big(\frac{1}{s} - \frac{1}{t}\Big)^2 \geq 0,$$

the terms 4 and $1/(s^2t^2)$ canceling between the two products.

(iv) Not monotone of order 2. Here $f'(t) = 1 + 2t > 0$ and $f[s,t] = 1 + s + t$, so

$$\begin{aligned} f'(s)f'(t) - f[s,t]^2 &= (1+2s)(1+2t) - (1+s+t)^2 \\ &= 4st - (s+t)^2 = -(s-t)^2, \end{aligned}$$

which is < 0 for $s \neq t$. Adding the operator monotone function t to the nonmonotone t^2 does not repair it.

21.9.6.3 C. Going deeper

Exercise. C1: Powers, directly

Let $k \geq 1$, let $\mathbf{A} = \operatorname{diag}(t_1, \dots, t_n)$, and let $\mathbf{H} \in M_n(F)$ be Hermitian.

(a) Expand $(\mathbf{A}+s\mathbf{H})^k$ and show that the coefficient of s is $\sum_{p+q=k-1} \mathbf{A}^p\mathbf{H}\mathbf{A}^q$, the sum being over integers $p, q \geq 0$.

(b) Compute the (i,j) entry of that matrix and identify it with $f[t_i,t_j]h_{ij}$ for $f(t)=t^k$.

(c) Hence prove Theorem 21.9.4 (a) for every polynomial f, without using Lemma 21.9.3.

Solution.

(a) Expanding the product $(\mathbf{A}+s\mathbf{H})(\mathbf{A}+s\mathbf{H})\cdots(\mathbf{A}+s\mathbf{H})$ of k factors and keeping the order of every product, the terms are indexed by the choice of $\mathbf{A}$ or $s\mathbf{H}$ in each factor. The terms with no $\mathbf{H}$ give $\mathbf{A}^k$; the terms with exactly one $\mathbf{H}$, in position $p+1$, give $s\,\mathbf{A}^p\mathbf{H}\mathbf{A}^{k-1-p}$ for $p=0,\dots,k-1$; the remaining terms carry a factor s^m with $m\ge 2$. So

$$(\mathbf{A}+s\mathbf{H})^k=\mathbf{A}^k+s\sum_{p+q=k-1}\mathbf{A}^p\mathbf{H}\mathbf{A}^q+s^2\mathbf{R}(s)$$

with $\mathbf{R}$ a matrix whose entries are polynomials in s.

(b) Since $\mathbf{A}^p=\operatorname{diag}(t_1^p,\dots,t_n^p)$, the matrix $\mathbf{A}^p\mathbf{H}\mathbf{A}^q$ has (i,j) entry $t_i^p h_{ij} t_j^q$. Summing,

$$\Big(\sum_{p+q=k-1}\mathbf{A}^p\mathbf{H}\mathbf{A}^q\Big)_{ij}=h_{ij}\sum_{p=0}^{k-1}t_i^p t_j^{k-1-p}=h_{ij}\,f[t_i,t_j],$$

the last equality by the third example after Definition 21.9.1, which covers $t_i=t_j$ as well.

(c) Each entry of $(\mathbf{A}+s\mathbf{H})^k$ is a polynomial in s, so it is differentiable at $s=0$, and by (a) and (b) its derivative there is $f[t_i,t_j]h_{ij}$, that is, the (i,j) entry of $L_f(t_1,\dots,t_n)\circ\mathbf{H}$. By Theorem 12.8.3 (a) the functional calculus of t^k is the k-th power, so this is the assertion for $f(t)=t^k$. For a general polynomial $f=\sum_k c_k t^k$, both $\mathbf{A}\mapsto f(\mathbf{A})$ and $f\mapsto L_f$ are linear in f (Theorem 12.8.3 (b) and the additivity of divided differences), so the identity for the powers gives it for f.

Exercise. C2: The logarithm at order two

(a) Prove that $r-1/r-2\log r\ge 0$ for every $r\ge 1$.

(b) Deduce that $\log$ is monotone of order 2 on $(0,\infty)$.

(c) Deduce that $f(t)=-\log(1-t)$ is monotone of order 2 on $(0,1)$, and verify that f is convex there.

Hint for (a): differentiate, and recognize a square.

Solution.

(a) Let $h(r)=r-1/r-2\log r$ on $[1,\infty)$. Then $h(1)=0$ and, for $r>0$,

$$h'(r)=1+\frac{1}{r^2}-\frac{2}{r}=\Big(1-\frac{1}{r}\Big)^2\ \ge\ 0.$$

For $r>1$, the mean value theorem (A6) gives $h(r)-h(1)=h'(c)(r-1)$ for some $c\in(1,r)$, and both factors are ≥ 0. Hence $h(r)\ge 0$.

(b) By Corollary 21.9.6 with $f=\log$ we must check $f'(t)=1/t\ge 0$, which holds on $(0,\infty)$,

and

$$\Big(\frac{\log s - \log t}{s - t}\Big)^2 \;\le\; \frac{1}{st} \qquad (s \neq t),$$

the case $s = t$ being the tautology. Both sides are unchanged by swapping s and t, so assume $s > t > 0$ and write $s = tr^2$ with $r > 1$. Then $\log s - \log t = 2\log r$, while

$$\frac{s - t}{\sqrt{st}} = \frac{t(r^2 - 1)}{tr} = r - \frac{1}{r}.$$

Both $2\log r$ and $r - 1/r$ are ≥ 0, and (a) says $2\log r \le r - 1/r$. Squaring the inequality between nonnegative numbers gives $(\log s - \log t)^2 \le (s-t)^2/(st)$, which is the required inequality after dividing by $(s - t)^2 > 0$.

(c) Put $a = 1 - s$ and $b = 1 - t$ for $s, t \in (0, 1)$, so $a, b \in (0, 1)$ and $s - t = b - a$. Then $f'(t) = 1/(1 - t)$, so $f'(s)f'(t) = 1/(ab)$, and

$$f[s, t] = \frac{-\log(1 - s) + \log(1 - t)}{s - t} = \frac{\log b - \log a}{b - a},$$

which is the divided difference of log at a and b. By (b) its square is at most $1/(ab) = f'(s)f'(t)$, and $f' > 0$, so Corollary 21.9.6 applies. Finally $f''(t) = 1/(1 - t)^2 > 0$, so f is convex on $(0, 1)$ by Theorem 18.10.6. So an increasing **convex** function can be monotone of order 2.

Exercise. C3: Orders are nested

Let $1 \le m \le n$ and let $f\colon I \to \mathbb{R}$; in (b) and (c) assume in addition that $f \in C^1(I)$, so that the Loewner matrices are defined and Theorem 21.9.5 applies.

(a) Prove that if f is monotone of order n on I, then it is monotone of order m on I.

(b) Give the corresponding statement about Loewner matrices, and prove it directly from Definition 21.9.2, without using Theorem 21.9.5.

(c) Explain why (a) and (b) together mean that a single failing pair of points rules out operator monotonicity at every order $n \ge 2$.

Hint for (a): pad with a scalar block.

Solution.

(a) Let $\mathbf{A}, \mathbf{B} \in M_m(F)$ be Hermitian with spectra in I, $\mathbf{A} \succeq \mathbf{B}$. Fix $c \in I$ and put

$$\widetilde{\mathbf{A}} = \mathbf{A} \oplus c\mathbf{I}_{n-m}, \qquad \widetilde{\mathbf{B}} = \mathbf{B} \oplus c\mathbf{I}_{n-m},$$

both Hermitian in $M_n(F)$ with spectra in I. Their difference is $(\mathbf{A} - \mathbf{B}) \oplus \mathbf{O}$, whose quadratic form at $(\mathbf{x}, \mathbf{y})$ is $\langle (\mathbf{A} - \mathbf{B})\mathbf{x}, \mathbf{x}\rangle \ge 0$, so $\widetilde{\mathbf{A}} \succeq \widetilde{\mathbf{B}}$. By hypothesis $f(\widetilde{\mathbf{A}}) \succeq f(\widetilde{\mathbf{B}})$. By Lemma 21.8.1 (c), $f(\widetilde{\mathbf{A}}) = f(\mathbf{A}) \oplus f(c)\mathbf{I}_{n-m}$ and likewise for $\mathbf{B}$, so the difference is $(f(\mathbf{A}) - f(\mathbf{B})) \oplus \mathbf{O}$. Evaluating its quadratic form at $(\mathbf{x}, \mathbf{0})$ gives $\langle (f(\mathbf{A}) - f(\mathbf{B}))\mathbf{x}, \mathbf{x}\rangle \ge 0$ for every $\mathbf{x} \in F^m$, that is $f(\mathbf{A}) \succeq f(\mathbf{B})$.

(b) If $L_f(t_1, \dots, t_n) \succeq \mathbf{O}$ for all $t_1, \dots, t_n \in I$, then $L_f(t_1, \dots, t_m) \succeq \mathbf{O}$ for all $t_1, \dots, t_m \in I$. Indeed, given $t_1, \dots, t_m$, pad the list with any $t_{m+1}, \dots, t_n \in I$; then $L_f(t_1, \dots, t_m)$ is the leading principal $m \times m$ submatrix of $L_f(t_1, \dots, t_n)$, because the (i, j) entry of each is $f[t_i, t_j]$ for $i, j \le m$. A principal submatrix of a positive semidefinite matrix is

positive semidefinite, since its quadratic form is the quadratic form of the big matrix at vectors supported on those coordinates (compare the proof of Theorem 13.1.4).

(c) By (b), a pair $s,t \in I$ with $L_f(s,t) \not\succeq \mathbf{O}$ forces $L_f(t_1,\dots,t_n) \not\succeq \mathbf{O}$ for the padded list at every $n \geq 2$, so by Theorem 21.9.5 f is monotone of no order $n \geq 2$. Equivalently, by (a), failure at order 2 propagates upward — which is why Corollary 21.9.6 is the cheapest refutation available.

21.10 Loewner's Theorem

Chapter 17 §11 stated Loewner's theorem, said four times that this chapter would prove it, and named the one thing it was missing. Section 9 supplied the elementary half: a continuously differentiable function is monotone of order n exactly when every Loewner matrix of size n built from it is positive semidefinite (Theorem 21.9.5). What is left is to turn that infinite family of conditions into a single formula, and to run the formula back. Turning it into a formula is the one thing this book does not prove; running it back costs about a page of integration, which we build first.

Throughout, $F = \mathbb{R}$ **or** $F = \mathbb{C}$, every matrix whose eigenvalues are indexed is Hermitian, and $f(\mathbf{A})$ is the functional calculus of Definition 12.8.2. The letter I denotes an **interval** of real numbers, not the identity matrix $\mathbf{I}$, as in Chapter 17 §11. Three things from the analysis list of Chapter 16's introduction are used: completeness (A1), monotone convergence (A2) and the compactness of closed bounded sets (A3). Everything imported beyond those is named where it appears.

21.10.1 Integrating a matrix-valued function

Chapter 17 §11 read the integral formula of Theorem 17.11.9 (ii) and said what stopped it there:

> What this book lacks is the analysis that carries the Loewner order through an integral, which is not on the list (A1)–(A6) in Chapter 16's introduction.

That analysis is one definition and one lemma. It is not a theory of integration; it is the least that makes the sentence "an integral of positive semidefinite matrices is positive semidefinite" mean something.

A **tagged partition** of a compact interval $[a,b]$ is a list of division points $a = s_0 < s_1 < \cdots < s_N = b$ together with **tags** $\tau_i \in [s_{i-1}, s_i]$; we call the whole package P. Its **mesh** is $\max_i (s_i - s_{i-1})$. For $\mathbf{M}\colon [a,b] \to M_{p\times q}(\mathbb{C})$ the associated **Riemann sum** is

$$S(\mathbf{M}, P) := \sum_{i=1}^{N} \mathbf{M}(\tau_i)\,(s_i - s_{i-1}).$$

Call $\mathbf{M}$ **continuous** when each of its pq entries is a continuous complex-valued function of s; by Corollary 16.2.4 this is the same as continuity for any norm on $M_{p\times q}(\mathbb{C})$.

Proposition 21.10.1. Riemann Sums of a Continuous Matrix Function Converge

Let $\mathbf{M}\colon [a,b] \to M_{p\times q}(\mathbb{C})$ be continuous. Then there is a matrix $J \in M_{p\times q}(\mathbb{C})$ such that $S(\mathbf{M}, P_k) \to J$ for **every** sequence (P_k) of tagged partitions of $[a,b]$ whose meshes tend to 0.

Idea. Two Riemann sums of small mesh are close because each is close to the sum over their common refinement, and on a refinement the tag moves by less than the mesh. So the sums form a Cauchy sequence and converge. The one analytic input is that a continuous function on a compact interval does not merely move a little near each point but moves a little **uniformly**, which is compactness.

Proof.

Claim. For every $\varepsilon > 0$ there is $\delta > 0$ such that $\|\mathbf{M}(s) - \mathbf{M}(s')\|_F \le \varepsilon$ whenever $s, s' \in [a,b]$ and $|s - s'| < \delta$.

Proof. Suppose not. Then for some $\varepsilon > 0$ and every $k \ge 1$ the value $\delta = 1/k$ fails, so there are $s_k, s_k' \in [a,b]$ with $|s_k - s_k'| < 1/k$ and $\|\mathbf{M}(s_k) - \mathbf{M}(s_k')\|_F > \varepsilon$. The interval $[a,b]$ is a closed bounded subset of $\mathbb{R}$, so by the compactness of closed bounded sets, **fact (A3) of Chapter 16's introduction**, a subsequence s_{k_j} converges to some $s \in [a,b]$. Then $s_{k_j}' \to s$ as well, since $|s_{k_j}' - s| \le |s_{k_j}' - s_{k_j}| + |s_{k_j} - s| \to 0$. Continuity of $\mathbf{M}$ gives $\mathbf{M}(s_{k_j}) \to \mathbf{M}(s)$ and $\mathbf{M}(s_{k_j}') \to \mathbf{M}(s)$, so $\|\mathbf{M}(s_{k_j}) - \mathbf{M}(s_{k_j}')\|_F \to 0$, contradicting the lower bound ε.

■

Fix $\varepsilon > 0$ and let δ be as in the claim. Let P be a tagged partition of mesh $< \delta$, and let Q be a tagged partition whose division points include those of P. Each interval q of Q lies inside exactly one interval of P; write $\tau(q)$ for the tag of that interval of P and $\rho(q)$ for the tag of q. Splitting each term of $S(\mathbf{M}, P)$ along the intervals of Q inside it,

$$S(\mathbf{M}, P) - S(\mathbf{M}, Q) = \sum_q \big(\mathbf{M}(\tau(q)) - \mathbf{M}(\rho(q))\big)|q|,$$

where $|q|$ is the length of q. Both $\tau(q)$ and $\rho(q)$ lie in one interval of P, of length $< \delta$, so each bracket has Frobenius norm at most ε. The triangle inequality gives

$$\|S(\mathbf{M}, P) - S(\mathbf{M}, Q)\|_F \;\le\; \varepsilon \sum_q |q| \;=\; \varepsilon(b-a).$$

If P and P' both have mesh $< \delta$, take for Q the partition whose division points are those of P together with those of P', with any tags; applying the bound twice,

$$\|S(\mathbf{M}, P) - S(\mathbf{M}, P')\|_F \;\le\; 2\varepsilon(b-a). \tag{$*$}$$

Now let (P_k) be tagged partitions with meshes tending to 0. Given $\eta > 0$, apply $(*)$ with $\varepsilon = \eta/(2(b-a)+1)$: for all k, l large enough that both meshes are $< \delta$, $\|S(\mathbf{M}, P_k) - S(\mathbf{M}, P_l)\|_F < \eta$. So each entry of $S(\mathbf{M}, P_k)$ is a Cauchy sequence of complex numbers, hence convergent by **completeness, fact (A1)**; call the entrywise limit J. If (P_k') is a second such sequence, interleaving the two produces a third sequence of meshes tending to 0, whose sums converge; both subsequences must have that common limit, so the limits agree. This proves the proposition.

■

Definition 21.10.1. Integral of a Continuous Matrix-Valued Function

Let $\mathbf{M}\colon [a,b] \to M_{p\times q}(\mathbb{C})$ be continuous. Its **integral**

$$\int_a^b \mathbf{M}(s)\,\mathrm{d}s \;\in\; M_{p\times q}(\mathbb{C})$$

is the matrix J of Proposition 21.10.1. For a continuous $\mathbf{M}$ on an open or half-open interval with endpoints $a < b$, where $a = -\infty$ or $b = +\infty$ is allowed, the **improper integral** $\int_a^b \mathbf{M}(s)\,\mathrm{d}s$ is the matrix J', if one exists, such that

$$\int_{\alpha_k}^{\beta_k} \mathbf{M}(s)\,\mathrm{d}s \;\longrightarrow\; J' \quad \text{entrywise}$$

for every pair of sequences with $\alpha_k \to a$, $\beta_k \to b$, and $a < \alpha_k < \beta_k < b$. The integral is then said to **converge**.

Five properties are used below, and each is read off the Riemann sums together with the algebra of limits.

- **Entrywise.** The (i,j) entry of $\int_a^b \mathbf{M}$ is $\int_a^b m_{ij}$, because that is true of every Riemann sum. In particular Definition 21.10.1 for $p = q = 1$ is the ordinary Riemann integral of a continuous complex function.
- **Linear.** $\int(\mathbf{M}+\mathbf{N}) = \int \mathbf{M} + \int \mathbf{N}$ and $\int c\mathbf{M} = c\int \mathbf{M}$ for a scalar c.
- **Additive in the interval.** For $a < c < b$, $\int_a^b \mathbf{M} = \int_a^c \mathbf{M} + \int_c^b \mathbf{M}$: joining a tagged partition of $[a,c]$ to one of $[c,b]$ gives a tagged partition of $[a,b]$ whose Riemann sum is the sum of the two, and the meshes tend to 0 together.
- **Constant factors pass out.** For constant $\mathbf{X} \in M_{r\times p}(\mathbb{C})$, $\mathbf{Y} \in M_{q\times r'}(\mathbb{C})$, $\int_a^b \mathbf{X}\mathbf{M}(s)\mathbf{Y}\,\mathrm{d}s = \mathbf{X}\big(\int_a^b \mathbf{M}\big)\mathbf{Y}$, since $S(\mathbf{X}\mathbf{M}\mathbf{Y}, P) = \mathbf{X}\,S(\mathbf{M},P)\,\mathbf{Y}$.
- **Pairing passes in.** Taking $\mathbf{X} = \mathbf{x}^*$ and $\mathbf{Y} = \mathbf{x}$ in the previous item, $\mathbf{x}^*\big(\int_a^b \mathbf{M}\big)\mathbf{x} = \int_a^b \mathbf{x}^*\mathbf{M}(s)\mathbf{x}\,\mathrm{d}s$.

All five hold for a convergent improper integral, by taking the limit in α and β.

Now the lemma the whole section is for.

Lemma 21.10.2. The Integral of a Positive Semidefinite Integrand

Let $\mathbf{M}\colon [a,b] \to M_n(\mathbb{C})$ be continuous with $\mathbf{M}(s) \succeq \mathbf{O}$ for every $s \in [a,b]$. Then

$$\int_a^b \mathbf{M}(s)\,\mathrm{d}s \;\succeq\; \mathbf{O}.$$

The same conclusion holds for a **convergent improper** integral of a continuous integrand that is $\succeq \mathbf{O}$ at every point of the interval.

Proof. Let $\mathbf{x} \in \mathbb{C}^n$. For a tagged partition P,

$$\mathbf{x}^* S(\mathbf{M},P)\mathbf{x} = \sum_{i=1}^{N} (\mathbf{x}^*\mathbf{M}(\tau_i)\mathbf{x})(s_i - s_{i-1}),$$

a sum of products of two nonnegative real numbers, since $\mathbf{M}(\tau_i) \succeq \mathbf{O}$ (Definition 13.1.1) and

the interval lengths are positive. Hence $\mathbf{x}^* S(\mathbf{M}, P)\mathbf{x} \geq 0$ for every P. Taking any sequence of tagged partitions with meshes tending to 0 and using Proposition 21.10.1, $\mathbf{x}^* S(\mathbf{M}, P_k)\mathbf{x} \to \mathbf{x}^* \left(\int_a^b \mathbf{M}\right)\mathbf{x}$; a nonstrict inequality survives a limit, so that number is ≥ 0.

It remains to check that $\int_a^b \mathbf{M}$ is Hermitian. Each $\mathbf{M}(s)$ is Hermitian, so each Riemann sum is, and the entrywise limit of Hermitian matrices is Hermitian: $\overline{\left(\int m_{ji}\right)} = \int \overline{m_{ji}} = \int m_{ij}$, since conjugation is continuous and commutes with sums. So $\int_a^b \mathbf{M} \succeq \mathbf{O}$.

For the improper case, each proper integral $\int_\alpha^\beta \mathbf{M}$ is $\succeq \mathbf{O}$ by what has just been proved, and $\mathbf{x}^* \left(\int_a^b \mathbf{M}\right)\mathbf{x}$ is the limit of the numbers $\mathbf{x}^* \left(\int_\alpha^\beta \mathbf{M}\right)\mathbf{x} \geq 0$, and the limit is Hermitian for the same reason.

■

That is the whole of the promised analysis. Nothing in it is deep: the positive semidefinite matrices are closed under sums and under multiplication by nonnegative numbers, and they are closed under limits, so a limit of sums of them stays inside.

Quick check. Let $\mathbf{M}(t) = \begin{pmatrix} 1 & t \\ t & t^2 \end{pmatrix}$ for $t \in [0, 1]$. Is $\mathbf{M}(t) \succeq \mathbf{O}$ for each t? Compute $\int_0^1 \mathbf{M}(t)\,\mathrm{d}t$ and check that the answer is $\succeq \mathbf{O}$. Does it have rank 1?

Solution. For each t, $\mathbf{M}(t) = \mathbf{v}(t)\mathbf{v}(t)^\top$ with $\mathbf{v}(t) = (1, t)$, so $\mathbf{x}^\top \mathbf{M}(t)\mathbf{x} = (x_1 + t x_2)^2 \geq 0$ and $\mathbf{M}(t) \succeq \mathbf{O}$. Integrating entrywise, and using $\int_0^1 t\,\mathrm{d}t = \frac{1}{2}$ and $\int_0^1 t^2\,\mathrm{d}t = \frac{1}{3}$ (for a continuous integrand these can be read off the Riemann sums $N^{-2}\sum_{i \leq N} i$ and $N^{-3}\sum_{i \leq N} i^2$),

$$\int_0^1 \mathbf{M}(t)\,\mathrm{d}t = \begin{pmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{3} \end{pmatrix},$$

whose determinant is $\frac{1}{3} - \frac{1}{4} = \frac{1}{12} > 0$ and whose $(1,1)$ entry is positive, so it is positive definite by Sylvester's criterion (Theorem 13.1.5 (d)). It has rank 2, not 1: every integrand was a rank-one matrix, and the integral is not. Lemma 21.10.2 preserves positivity and nothing else.

21.10.2 The one fact this chapter quotes

Two statements in this chapter are not proved from Chapters 1–20: the one below, and Lieb's concavity theorem, which §12 states and nothing in the book uses. Beside them this section quotes two standard facts of one-variable calculus, the fundamental theorem of calculus and the substitution rule, each named at the point where it is used. The book's practice for such a fact is fixed — Chapter 11 with the theorem of Abel and Ruffini, Chapter 14 §06 with Taylor's theorem — and it is followed here: the statement is given in full, in a block of its own, named at every use, and never buried inside a proof that presents itself as self-contained.

Theorem 21.10.3. (A7) The Theorem of Pick and Nevanlinna

Let $I \subseteq \mathbb{R}$ be an open interval and let $f\colon I \to \mathbb{R}$.

(a) Suppose that f is differentiable and that for every $n \geq 1$ and all $t_1, \dots, t_n \in I$ the Loewner matrix $L_f(t_1, \dots, t_n)$ of Definition 21.9.2 is positive semidefinite. Then f is the restriction to I of a function analytic on $\mathbb{C} \setminus (\mathbb{R} \setminus I)$ that maps every point of the upper half-plane $\{z : \operatorname{Im} z > 0\}$ to a point with nonnegative imaginary part.

(b) Conversely, every function analytic on $\mathbb{C} \setminus (-\infty, 0]$ that maps the upper half-plane into $\{\operatorname{Im} z \geq 0\}$ has, on $I = (0, \infty)$, the integral representation of Theorem 17.11.9 (ii): there are $\alpha \in \mathbb{R}$, $\beta \geq 0$, and a positive measure ν on $[0, \infty)$ with $\int (1+s^2)^{-1} \mathrm{d}\nu(s) < \infty$ such that

$$f(t) = \alpha + \beta t + \int_{[0,\infty)} \Big(\frac{s}{1+s^2} - \frac{1}{t+s} \Big) \mathrm{d}\nu(s)$$

for every $t > 0$.

(c) A function that is operator monotone on an open interval is continuously differentiable there.

What it gives. Exactly one thing: it converts an infinite family of matrix conditions — one positive semidefinite Loewner matrix for each size and each choice of points — into a single formula with two constants and one measure. Nothing in the statement is about matrices at all after clause (a) is applied.

What it costs. Clause (a) is complex analysis: one shows that the positivity of the Loewner matrices forces f to extend analytically across I, and that the extension preserves the upper half-plane. Clause (b) is the Herglotz–Riesz representation of such a function, which produces the measure ν by a compactness argument in a space of measures. Clause (c) is a regularity theorem. None of these is within reach of Chapters 1–20, and Chapter 17 §11 said so when it stated the theorem: "Its proof needs complex analysis and some measure theory, which lie outside Chapter 16's list (A1)–(A6)."

A second cost, worth naming separately. The integral in (b) is taken against a measure, and this book constructs no such integral. Definition 21.10.1 covers the case $\mathrm{d}\nu(s) = w(s)\,\mathrm{d}s$ with w continuous, and that is the only case the corollaries below need: each of them starts from a representation written down and verified by hand, with a continuous density. For a general ν, which is what (b) allows, the argument uses exactly two properties of the integral: it is **linear** in the integrand, and it sends a **nonnegative** integrand to a nonnegative number. Both are part of the meaning of "integral against a positive measure". For the integral the book does construct they are proved — linearity in the list of properties after Definition 21.10.1, positivity in Lemma 21.10.2; for a general measure they are quoted, along with (A7) itself.

Which half rests on it. Only the hard half. In Theorem 21.10.4 below, the implication (ii) $\Rightarrow$ (i) is proved here from Chapter 17 §11 and Lemma 21.10.2, with no appeal to (A7) — subject only to the second cost just named, since for a general ν the linearity and the positivity of the integral are quoted rather than constructed. The implications (i) $\Rightarrow$ (iii) $\Rightarrow$ (ii) — the classification itself — use (A7) and cannot be had without it. Nothing else in this chapter uses (A7); in particular Corollary 21.10.8, Corollary 21.10.9, and Corollary 21.10.10 below, which are what the earlier chapters actually asked for, are proved without it.

21.10.3 Loewner's theorem

Theorem 21.10.4. Loewner's Theorem

Let $f\colon (0, \infty) \to \mathbb{R}$. The following are equivalent.

(i) f is operator monotone on $(0,\infty)$.

(ii) There are $\alpha \in \mathbb{R}$, $\beta \geq 0$, and a positive measure ν on $[0,\infty)$ with $\int (1+s^2)^{-1} d\nu(s) < \infty$ such that for every $t > 0$

$$f(t) = \alpha + \beta t + \int_{[0,\infty)} \Big(\frac{s}{1+s^2} - \frac{1}{t+s} \Big) d\nu(s).$$

(iii) f is the restriction to $(0,\infty)$ of a function analytic on $\mathbb{C} \setminus (-\infty, 0]$ that sends every point of the upper half-plane to a point with nonnegative imaginary part.

These are the three conditions of Theorem 17.11.9.

Idea. Three steps, in a cycle.

① (ii) ⇒ (i). At each fixed $s \geq 0$ the integrand is $-1/(t+s)$ plus a constant, and Corollary 17.11.5 already knows those are operator monotone. Subtracting $f(\mathbf{B})$ from $f(\mathbf{A})$ kills the constants and leaves an integral of positive semidefinite matrices, which Lemma 21.10.2 declares positive semidefinite.

② (i) ⇒ (iii). Operator monotonicity is monotonicity of every order, and §09 converts that into positivity of every Loewner matrix. (A7) converts that into analyticity.

③ (iii) ⇒ (ii). (A7) again.

Proof. **Step 1: (ii) ⇒ (i).** Let $\mathbf{A}, \mathbf{B} \in M_n(F)$ be Hermitian with spectra in $(0,\infty)$ and $\mathbf{A} \succeq \mathbf{B}$. For $s \geq 0$ put $g_s(t) = -1/(t+s)$ on $(0,\infty)$. By Corollary 17.11.5, g_s is operator monotone there, so

$$\mathbf{R}(s) \coloneqq g_s(\mathbf{A}) - g_s(\mathbf{B}) = (\mathbf{B} + s\mathbf{I})^{-1} - (\mathbf{A} + s\mathbf{I})^{-1}$$

satisfies $\mathbf{R}(s) \succeq \mathbf{O}$ for every $s \geq 0$; the identification of $g_s(\mathbf{A})$ with $-(\mathbf{A}+s\mathbf{I})^{-1}$ is Definition 12.8.2 applied to the spectral resolution of $\mathbf{A}$.

Write $\mathbf{A} = \sum_j \mu_j \mathbf{P}_j$ for that resolution (Theorem 12.8.1). By Definition 12.8.2 and the hypothesis (ii),

$$\begin{aligned} f(\mathbf{A}) &= \sum_j f(\mu_j)\mathbf{P}_j \\ &= \alpha\mathbf{I} + \beta\mathbf{A} \\ &\quad + \int_{[0,\infty)} \Big(\frac{s}{1+s^2}\mathbf{I} - (\mathbf{A}+s\mathbf{I})^{-1} \Big) d\nu(s), \end{aligned}$$

where the second line substitutes the formula (ii) for each $f(\mu_j)$ and moves the finite sum $\sum_j (\cdot)\mathbf{P}_j$ inside the integral, which is legitimate because the integral is **linear** in its integrand and the $\mathbf{P}_j$ do not depend on s; the identities $\sum_j \mathbf{P}_j = \mathbf{I}$ and $\sum_j \mu_j \mathbf{P}_j = \mathbf{A}$ are Theorem 12.8.1 (c) and (d). The same computation applies to $\mathbf{B}$. Subtracting, the terms $\alpha\mathbf{I}$ and $\frac{s}{1+s^2}\mathbf{I}$ cancel, and

$$f(\mathbf{A}) - f(\mathbf{B}) = \beta(\mathbf{A} - \mathbf{B}) + \int_{[0,\infty)} \mathbf{R}(s)\, d\nu(s).$$

Let $\mathbf{x} \in F^n$. Pairing passes inside the integral, so

$$\mathbf{x}^*(f(\mathbf{A}) - f(\mathbf{B}))\mathbf{x} = \beta\,\mathbf{x}^*(\mathbf{A} - \mathbf{B})\mathbf{x} + \int_{[0,\infty)} \mathbf{x}^*\mathbf{R}(s)\mathbf{x}\,\mathrm{d}\nu(s) \;\geq\; 0,$$

because $\beta \geq 0$, $\mathbf{x}^*(\mathbf{A}-\mathbf{B})\mathbf{x} \geq 0$, and the integrand $\mathbf{x}^*\mathbf{R}(s)\mathbf{x}$ is nonnegative for every s. This last step is Lemma 21.10.2 when $\mathrm{d}\nu(s) = w(s)\mathrm{d}s$ with w continuous, and in general uses only the linearity and positivity of the integral against a positive measure, as recorded with Theorem 21.10.3. Hence $f(\mathbf{A}) \succeq f(\mathbf{B})$. The size n was arbitrary, so f is operator monotone (Definition 17.11.1 (a)).

Step 2: (i) ⇒ (iii). Let f be operator monotone on $(0,\infty)$. By Theorem 21.10.3 (c), f is continuously differentiable there. Fix $n \geq 1$ and $t_1, \ldots, t_n \in (0,\infty)$. Being operator monotone, f is in particular monotone of order n, so Theorem 21.9.5 gives $L_f(t_1,\ldots,t_n) \succeq \mathbf{O}$. As n and the points were arbitrary, the hypothesis of Theorem 21.10.3 (a) holds with $I = (0,\infty)$. Since $\mathbb{R} \setminus (0,\infty) = (-\infty, 0]$, that clause produces an analytic function on $\mathbb{C} \setminus (-\infty, 0]$ restricting to f and mapping the upper half-plane into its closure. That is (iii).

Step 3: (iii) ⇒ (ii). This is Theorem 21.10.3 (b).

The cycle (i) ⇒ (iii) ⇒ (ii) ⇒ (i) closes, so the three conditions are equivalent.

■

Step 1 is the step Chapter 17 §11 identified as missing, and it turned out to cost one lemma. Steps 2 and 3 are where the content is, and they are quoted. It is worth being precise about the division: **the book proves that every function of the form (ii) is operator monotone, and quotes that there are no others.**

Warning. **Analytic and increasing is not enough.** Condition (iii) is not "extends analytically"; it is "extends analytically **and keeps the upper half-plane**". The function e^t extends to an entire function and is increasing on $\mathbb{R}$, yet $\operatorname{Im} e^z = e^x \sin y$ for $z = x + iy$, which is negative at $z = 3\pi i/2$, where $e^z = -i$. So e^t fails (iii), and Example (The exponential is not operator monotone) below confirms the failure directly with two 2×2 matrices.

21.10.4 The powers

Chapter 13 §05 announced the answer and deferred the proof:

> The dividing line is at $s = 1$: the functions preserving the order are called **operator monotone**, and $t \mapsto t^s$ is operator monotone on $[0,\infty)$ exactly for $0 \leq s \leq 1$. Chapter 21 proves this, along with the companion fact that $t \mapsto \log t$ is operator monotone while $t \mapsto e^t$ is not.

Chapter 17 §11 repeated it, adding that "the operator monotonicity of t^p for every $p \in (0,1)$ needs a new idea, an integral representation of the kind in Theorem 17.11.9". Here is the representation, and it is elementary. Real powers $t^r = e^{r\log t}$ are Chapter 18 §11's (Lemma 18.11.1).

Lemma 21.10.5. The Integral That Represents a Power

Let $0 < p < 1$. Then the improper integral

$$K_p := \int_0^\infty \frac{u^{p-1}}{1+u}\,\mathrm{d}u$$

converges, with $0 < K_p < \infty$, and for every $\lambda > 0$

$$\int_0^\infty s^{p-1}\,\frac{\lambda}{\lambda+s}\,\mathrm{d}s \;=\; K_p\,\lambda^p.$$

Idea. Both halves are one idea. For convergence, cut $(0,\infty)$ at the powers of 2: on each piece the integrand is bounded by its value at one endpoint, and the resulting bounds form a geometric series, convergent exactly because $0 < p < 1$. For the identity, substitute $s = \lambda u$; since the substitution is linear it can be made inside the Riemann sums, where it is visible as a relabeling of the partition together with a factor λ^p.

Proof. **Convergence.** The integrand is positive and continuous on $(0,\infty)$, so interval additivity makes $\delta \mapsto \int_\delta^1$ increase as δ decreases and $R \mapsto \int_1^R$ increase with R; the same property splits each range at the powers of 2 below. It is enough to bound both families above: bounded above, they have a least upper bound L by **completeness, fact (A1) of Chapter 16's introduction**, and monotonicity in the parameter makes L the limit along *every* sequence of parameters tending to the endpoint, which is the form Definition 21.10.1 asks for; along a decreasing sequence $\delta_k \to 0^+$ that limit is **monotone convergence, fact (A2)**. Since $p - 1 < 0$, the function u^{p-1} is decreasing, so on $[2^{-k-1}, 2^{-k}]$ it is at most $(2^{-k-1})^{p-1} = 2^{(k+1)(1-p)}$; as $1/(1+u) \le 1$ and the interval has length 2^{-k-1}, the integral over it is at most $2^{-(k+1)p}$. Summing the geometric series,

$$\int_\delta^1 \frac{u^{p-1}}{1+u}\,\mathrm{d}u \;\le\; \sum_{k\ge 0} 2^{-(k+1)p} = \frac{1}{2^p - 1},$$

for every $\delta \in (0,1)$. For the other end use $u^{p-1}/(1+u) \le u^{p-2}$, which on $[2^k, 2^{k+1}]$ is at most $2^{k(p-2)}$; the interval has length 2^k, so the integral over it is at most $2^{k(p-1)}$, and

$$\int_1^R \frac{u^{p-1}}{1+u}\,\mathrm{d}u \;\le\; \sum_{k\ge 0} 2^{k(p-1)} = \frac{1}{1 - 2^{p-1}},$$

which is finite because $p - 1 < 0$. So $K_p < \infty$, and $K_p > 0$ because the integrand is positive and $\int_1^2$ is already positive.

The identity. Fix $\lambda > 0$ and $0 < \delta < R$. Let P be a tagged partition $\delta/\lambda = u_0 < \cdots < u_N = R/\lambda$ with tags v_i. Multiplying every point by λ gives a tagged partition λP of $[\delta, R]$, with tags λv_i and mesh λ times that of P. Writing $g(s) = s^{p-1}\lambda/(\lambda+s)$ and $h(u) = u^{p-1}/(1+u)$,

$$\begin{aligned} g(\lambda v_i)\cdot\lambda(u_i - u_{i-1}) &= \frac{\lambda^{p-1} v_i^{p-1}\lambda}{\lambda(1+v_i)}\,\lambda(u_i - u_{i-1}) \\ &= \lambda^p\, h(v_i)(u_i - u_{i-1}), \end{aligned}$$

using $(\lambda v)^{p-1} = \lambda^{p-1} v^{p-1}$ (Lemma 18.11.1 and the rules for real powers recorded after it). Summing over i, $S(g, \lambda P) = \lambda^p S(h, P)$. Letting the mesh of P tend to 0 — which sends that

of λP to 0 as well — and using Proposition 21.10.1 on both sides,

$$\int_{\delta}^{R} g(s)\,\mathrm{d}s = \lambda^{p} \int_{\delta/\lambda}^{R/\lambda} h(u)\,\mathrm{d}u.$$

Now let $\delta \to 0^{+}$ and $R \to \infty$. The right side tends to $\lambda^{p} K_{p}$ by the first part, so the left side converges to it. This proves the lemma.

■

Remark. The number K_{p} equals $\pi / \sin(p\pi)$. That evaluation is a standard exercise in contour integration, and the constant $\sin(p\pi)/\pi$ is the one Chapter 17 §11 displayed. Nothing below needs its value: only $0 < K_{p} < \infty$ is used, and that was proved. Exercise B2 computes $K_{1/2} = \pi$ by hand.

Two more small tools, both elementary, and then the corollary.

Lemma 21.10.6. Composites of Operator Monotone Functions

Let I, J be intervals, let $g\colon I \to \mathbb{R}$ be operator monotone on I with $g(I) \subseteq J$, and let $h\colon J \to \mathbb{R}$ be operator monotone on J. Then $h \circ g$ is operator monotone on I.

Proof. Let $\mathbf{A}, \mathbf{B} \in M_{n}(F)$ be Hermitian with spectra in I and $\mathbf{A} \succeq \mathbf{B}$. Then $g(\mathbf{A}) \succeq g(\mathbf{B})$, and both are Hermitian (Theorem 12.8.3 (c)) with spectra $g(\operatorname{spec}(\mathbf{A}))$ and $g(\operatorname{spec}(\mathbf{B}))$ by Theorem 12.8.3 (d), hence contained in $g(I) \subseteq J$. So $h(g(\mathbf{A})) \succeq h(g(\mathbf{B}))$.

It remains to see that $h(g(\mathbf{A})) = (h \circ g)(\mathbf{A})$. Let $\mathbf{A} = \sum_{j} \mu_{j} \mathbf{P}_{j}$ be the spectral resolution. Group the indices by the value of g: let $\nu_{1}, \dots, \nu_{r}$ be the distinct numbers among $g(\mu_{1}), \dots, g(\mu_{k})$ and put $\mathbf{Q}_{l} = \sum_{j : g(\mu_{j}) = \nu_{l}} \mathbf{P}_{j}$. Each $\mathbf{Q}_{l}$ is a nonzero orthogonal projection, $\mathbf{Q}_{l}\mathbf{Q}_{l'} = \mathbf{O}$ for $l \neq l'$, and $\sum_{l} \mathbf{Q}_{l} = \mathbf{I}$, by Theorem 12.8.1 (b) and (c). Since $g(\mathbf{A}) = \sum_{l} \nu_{l} \mathbf{Q}_{l}$ with the ν_{l} distinct, Theorem 12.8.2 says this is the spectral resolution of $g(\mathbf{A})$. Hence, by Definition 12.8.2 applied twice,

$$h(g(\mathbf{A})) = \sum_{l} h(\nu_{l}) \mathbf{Q}_{l} = \sum_{j} h(g(\mu_{j})) \mathbf{P}_{j} = (h \circ g)(\mathbf{A}).$$

The same holds for $\mathbf{B}$, so $(h \circ g)(\mathbf{A}) \succeq (h \circ g)(\mathbf{B})$.

■

Lemma 21.10.7. Pointwise Limits Stay Operator Monotone

Let $f_{k}\colon I \to \mathbb{R}$ be operator monotone on I for every k, and suppose $f_{k}(t) \to f(t)$ for every $t \in I$. Then f is operator monotone on I.

Proof. Let $\mathbf{A} \succeq \mathbf{B}$ be Hermitian of the same size with spectra in I, with spectral resolutions $\mathbf{A} = \sum_{j} \mu_{j} \mathbf{P}_{j}$ and $\mathbf{B} = \sum_{l} \nu_{l} \mathbf{Q}_{l}$. By Definition 12.8.2, $f_{k}(\mathbf{A}) = \sum_{j} f_{k}(\mu_{j}) \mathbf{P}_{j}$, a fixed finite

combination of fixed matrices with scalar coefficients converging to $f(\mu_j)$; so $f_k(\mathbf{A}) \to f(\mathbf{A})$ entrywise, and likewise $f_k(\mathbf{B}) \to f(\mathbf{B})$. For each $\mathbf{x}$ and each k, $\mathbf{x}^*(f_k(\mathbf{A}) - f_k(\mathbf{B}))\mathbf{x} \geq 0$, and the left side tends to $\mathbf{x}^*(f(\mathbf{A}) - f(\mathbf{B}))\mathbf{x}$. A nonstrict inequality survives a limit, so $f(\mathbf{A}) \succeq f(\mathbf{B})$.

■

Corollary 21.10.8. The Operator Monotone Powers

Let $p \in \mathbb{R}$. The function $t \mapsto t^p$ is operator monotone on $(0, \infty)$ **if and only if** $0 \leq p \leq 1$. For those p the statement extends to $[0, \infty)$: if $\mathbf{A} \succeq \mathbf{B} \succeq \mathbf{O}$ are Hermitian of the same size, then $\mathbf{A}^p \succeq \mathbf{B}^p$.

Idea. For $0 < p < 1$, Lemma 21.10.5 rewrites t^p as a positive superposition of the functions $t/(t+s)$ that Chapter 17 §11 already proved operator monotone, and the superposition is an honest Riemann integral, so Lemma 21.10.2 finishes it. For $p > 1$ the trick is that operator monotone functions compose: if t^p were operator monotone for one $p > 1$, enough compositions with each other and with a small power would manufacture t^2, which is not.

Proof. **The cases $p = 0$ and $p = 1$.** These are the constant 1 and the identity, operator monotone by Proposition 17.11.2.

The case $p < 0$. Then $t \mapsto t^p$ is strictly decreasing on $(0, \infty)$, since $t \mapsto t^{-p}$ is strictly increasing there (Chapter 18 §11). An operator monotone function is increasing (Proposition 17.11.1), so it is not operator monotone.

The case $0 < p < 1$, on $(0, \infty)$. Let $\mathbf{A} \succeq \mathbf{B}$ be Hermitian in $M_n(F)$ with spectra in $(0, \infty)$. Write $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{U}^*$ with $\mathbf{U}$ unitary and $\mathbf{D} = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ the eigenvalues, all positive (Corollary 12.4.3, Corollary 12.5.2, Theorem 13.1.5 (b)). For $s > 0$,

$$s^{p-1}\mathbf{A}(\mathbf{A} + s\mathbf{I})^{-1} = \mathbf{U}\,\operatorname{diag}\Big(\frac{s^{p-1}\lambda_i}{\lambda_i + s}\Big)\mathbf{U}^*,$$

where we are using that $f(\mathbf{A}) = \mathbf{U}\operatorname{diag}(f(\lambda_1), \ldots, f(\lambda_n))\mathbf{U}^*$ for every function f on $\operatorname{spec}(\mathbf{A})$: writing $\mathbf{u}_1, \ldots, \mathbf{u}_n$ for the columns of $\mathbf{U}$, the spectral projection $\mathbf{P}_j$ of Definition 12.8.1 is $\sum_{i:\lambda_i = \mu_j} \mathbf{u}_i\mathbf{u}_i^*$, so the two expressions for $f(\mathbf{A})$ have the same terms. Every entry of the integrand is therefore a fixed linear combination of the n scalar functions of Lemma 21.10.5. By that lemma each of their improper integrals converges, so the matrix integral converges, and by the linearity of the integral together with the rule that constant factors pass out,

$$\int_0^\infty s^{p-1}\mathbf{A}(\mathbf{A} + s\mathbf{I})^{-1}\mathrm{d}s = K_p\,\mathbf{U}\operatorname{diag}(\lambda_i^p)\mathbf{U}^* = K_p\,\mathbf{A}^p,$$

the last equality by Definition 12.8.2. The same holds for $\mathbf{B}$. Subtracting,

$$\mathbf{A}^p - \mathbf{B}^p = \frac{1}{K_p}\int_0^\infty s^{p-1}(\mathbf{C}(s) - \mathbf{Y}(s))\mathrm{d}s,$$

with $\mathbf{C}(s) = \mathbf{A}(\mathbf{A} + s\mathbf{I})^{-1}$ and $\mathbf{Y}(s) = \mathbf{B}(\mathbf{B} + s\mathbf{I})^{-1}$. Fix $s > 0$. The function $t \mapsto t/(t+s)$ is operator monotone on $(0, \infty)$ by Corollary 17.11.5, and it sends $\mathbf{A}$ to $\mathbf{C}(s)$ and $\mathbf{B}$ to $\mathbf{Y}(s)$;

hence $\mathbf{C}(s) - \mathbf{Y}(s) \succeq \mathbf{O}$, and multiplying by the positive number s^{p-1} keeps it so. The integrand is continuous in s on $(0,\infty)$ and the integral converges, so Lemma 21.10.2 in its improper form gives $\mathbf{A}^p - \mathbf{B}^p \succeq \mathbf{O}$. As $K_p > 0$ and n was arbitrary, $t \mapsto t^p$ is operator monotone on $(0,\infty)$.

The case $p > 1$. Suppose, for a contradiction, that $t \mapsto t^p$ is operator monotone on $(0,\infty)$ for some $p > 1$. It maps $(0,\infty)$ into itself, so by Lemma 21.10.6 it may be composed with itself: $t \mapsto t^{p^k}$ is operator monotone for every $k \geq 1$. Choose k with $p^k \geq 2$ and put $q = 2/p^k \in (0,1]$. By the cases already proved, $t \mapsto t^q$ is operator monotone on $(0,\infty)$ and maps it into itself, so Lemma 21.10.6 applied to $g(t) = t^q$ and $h(u) = u^{p^k}$ makes

$$t \longmapsto (t^q)^{p^k} = t^{\,qp^k} = t^2$$

operator monotone on $(0,\infty)$.

Now let $\mathbf{A} \succeq \mathbf{B} \succeq \mathbf{O}$ be Hermitian and let $\varepsilon > 0$. Then $\mathbf{A} + \varepsilon\mathbf{I} \succeq \mathbf{B} + \varepsilon\mathbf{I} \succ \mathbf{O}$, both with spectra in $(0,\infty)$, so $(\mathbf{A} + \varepsilon\mathbf{I})^2 \succeq (\mathbf{B} + \varepsilon\mathbf{I})^2$. Expanding both squares, that says

$$\mathbf{A}^2 - \mathbf{B}^2 + 2\varepsilon(\mathbf{A} - \mathbf{B}) \;\succeq\; \mathbf{O}.$$

Fix $\mathbf{x}$ and let $\varepsilon \to 0^+$: the number $\mathbf{x}^*(\mathbf{A}^2 - \mathbf{B}^2)\mathbf{x} + 2\varepsilon\,\mathbf{x}^*(\mathbf{A} - \mathbf{B})\mathbf{x}$ is ≥ 0 for every $\varepsilon > 0$, so its limit $\mathbf{x}^*(\mathbf{A}^2 - \mathbf{B}^2)\mathbf{x}$ is ≥ 0. Hence $\mathbf{A}^2 \succeq \mathbf{B}^2$ whenever $\mathbf{A} \succeq \mathbf{B} \succeq \mathbf{O}$, contradicting Example (Squaring does not preserve the Loewner order). So no $p > 1$ works.

Extension to $[0,\infty)$ for $0 \leq p \leq 1$. Let $\mathbf{A} \succeq \mathbf{B} \succeq \mathbf{O}$ and $\varepsilon > 0$. As above $\mathbf{A} + \varepsilon\mathbf{I} \succeq \mathbf{B} + \varepsilon\mathbf{I} \succ \mathbf{O}$, so $(\mathbf{A} + \varepsilon\mathbf{I})^p \succeq (\mathbf{B} + \varepsilon\mathbf{I})^p$. If $\mathbf{A} = \sum_j \mu_j \mathbf{P}_j$ is the spectral resolution, then $\mathbf{A} + \varepsilon\mathbf{I} = \sum_j (\mu_j + \varepsilon)\mathbf{P}_j$ is the spectral resolution of $\mathbf{A} + \varepsilon\mathbf{I}$ (Theorem 12.8.2, as the shifted eigenvalues are still distinct), so

$$(\mathbf{A} + \varepsilon\mathbf{I})^p = \sum_j (\mu_j + \varepsilon)^p \mathbf{P}_j \;\longrightarrow\; \sum_j \mu_j^p \mathbf{P}_j = \mathbf{A}^p$$

as $\varepsilon \to 0^+$: for $\mu_j > 0$ this is the continuity of $u \mapsto e^{p \log u}$ on $(0,\infty)$ (Lemma 18.11.1), and for $\mu_j = 0$ it says $\varepsilon^p \to 0$, which holds because $u \mapsto u^p$ is a strictly increasing bijection of $[0,\infty)$ with itself, so $\varepsilon^p < \eta$ as soon as $\varepsilon < \eta^{1/p}$. The same for $\mathbf{B}$. Pairing with $\mathbf{x}$ and letting $\varepsilon \to 0^+$ gives $\mathbf{x}^*(\mathbf{A}^p - \mathbf{B}^p)\mathbf{x} \geq 0$. This proves the corollary.

■

This pays Chapter 13 §05 and Chapter 17 §11 in full for the powers, and it is worth recording what it cost: Corollary 17.11.5 from Chapter 17 §11, Lemma 21.10.2 from the first part of this section, and facts (A1)–(A3). **Neither the theorem of Pick and Nevanlinna nor Theorem 21.10.4 itself was used** — Chapter 17 §11 expected the powers to arrive with Loewner's theorem, and they arrived without it. What (A7) adds is the knowledge that the representation exists for *every* operator monotone function; for the particular function t^p the representation was written down and verified by hand.

Quick check. In Theorem 21.10.4 (ii), why does the term $s/(1+s^2)$ inside the integral play no part in the proof that f is operator monotone?

Solution. Because it does not depend on t. Applying the functional calculus turns it into $(s/(1+s^2))\mathbf{I}$, the same matrix for $\mathbf{A}$ as for $\mathbf{B}$, so it cancels in $f(\mathbf{A})-f(\mathbf{B})$. Its only job is to make the integral converge: without it the integrand $-1/(t+s)$ alone need not be integrable at $s=\infty$ against ν, while the difference $s/(1+s^2)-1/(t+s)$ decays like s^{-2}.

21.10.5 The logarithm and the exponential

Corollary 21.10.9. The Logarithm Is Operator Monotone

The function $t \mapsto \log t$ is operator monotone on $(0,\infty)$.

Idea. The logarithm is a limit of powers: $(t^p-1)/p \to \log t$ as $p \to 0^+$, because that quotient is a difference quotient for the exponential. Each $(t^p-1)/p$ is an increasing affine function of an operator monotone function, and Lemma 21.10.7 passes to the limit.

Proof. For $0<p<1$ put $f_p(t)=(t^p-1)/p$ on $(0,\infty)$. By Corollary 21.10.8, $t \mapsto t^p$ is operator monotone there, and $f_p=-\frac{1}{p}+\frac{1}{p}\,t^p$ with $1/p>0$, so f_p is operator monotone by Proposition 17.11.4 (a).

Fix $t>0$. If $t=1$ then $f_p(1)=0=\log 1$ for every p. If $t\neq 1$, write $c=\log t\neq 0$ and $x=pc$, so that $t^p=e^{pc}=e^x$ by the definition of real powers (Chapter 18 §11) and

$$f_p(t)=\frac{e^x-1}{p}=c\,\frac{e^x-1}{x}.$$

As $p\to 0^+$ we have $x\to 0$, and $(e^x-1)/x\to 1$, this being the difference quotient of exp at 0, whose value is $e^0=1$ (Theorem 10.9.4 (b) in size 1, as recorded in the proof of Lemma 18.11.1). Hence $f_p(t)\to c=\log t$.

Taking $p=1/k$ and letting $k\to\infty$, Lemma 21.10.7 applies and log is operator monotone on $(0,\infty)$.

■

Remark. The logarithm's place in Theorem 21.10.4 (ii) is as simple as it can be: $\alpha=0$, $\beta=0$, and ν the ordinary length measure, so that

$$\log t=\int_0^\infty\Big(\frac{s}{1+s^2}-\frac{1}{t+s}\Big)\mathrm{d}s$$

for every $t>0$. To verify this, take $0<\delta<R$ and use the fundamental theorem of calculus — an imported fact of one-variable calculus, quoted here and named again in the one proof that uses this representation, Corollary 21.10.10 — with the antiderivative $\frac{1}{2}\log(1+s^2)-\log(t+s)$, whose derivative is the integrand. The value at R is $\log(\sqrt{1+R^2}/(t+R))\to 0$, and the value at δ is $\log(\sqrt{1+\delta^2}/(t+\delta))\to\log(1/t)=-\log t$. The difference is $\log t$.

That representation has one more use, and it pays a debt of §08. The list of operator convex functions there stopped short of two, $-\log t$ and $t\log t$, because each needs an integral representation. The display above is that representation, and it settles both at once.

Corollary 21.10.10. Two Operator Convex Functions from the Logarithm

The functions $t \mapsto -\log t$ and $t \mapsto t\log t$ are operator convex on $(0,\infty)$.

Idea. Read the representation of $\log t$ for a fixed s instead of for a fixed t. The integrand of $-\log t$ is then $1/(t+s)$ plus a constant, and Theorem 21.8.3 (d) already knows that function is operator convex; multiplying the representation through by t turns the integrand into $-t/(t+s)$ plus an affine function, and the same clause knows $t/(t+s)$ is operator **concave**. Either way the convexity defect is an integral of positive semidefinite matrices, which Lemma 21.10.2 declares positive semidefinite.

Proof. Write $c(s) = s/(1+s^2)$. The representation in the remark above says that for every $t>0$

$$-\log t = \int_0^\infty \Big(\frac{1}{t+s} - c(s)\Big)\mathrm{d}s, \qquad t\log t = \int_0^\infty \Big(t\,c(s) - \frac{t}{t+s}\Big)\mathrm{d}s,$$

the second display being the first multiplied by $-t$, a constant as far as s is concerned.

Let $\mathbf{A},\mathbf{B}\in M_n(F)$ be Hermitian with spectra in $(0,\infty)$, let $\theta\in[0,1]$ and put $\mathbf{C}=\theta\mathbf{A}+(1-\theta)\mathbf{B}$, whose spectrum again lies in $(0,\infty)$, as Chapter 17 §11 checked.

The logarithm. Write $\mathbf{A}=\sum_j \mu_j \mathbf{P}_j$ for the spectral resolution (Theorem 12.8.1). Substituting the first display for each $-\log\mu_j$ and moving the finite sum $\sum_j(\cdot)\mathbf{P}_j$ inside the integral — legitimate because the integral is linear in its integrand and the $\mathbf{P}_j$ do not depend on s — gives

$$-\log\mathbf{A} = \int_0^\infty \big((\mathbf{A}+s\mathbf{I})^{-1} - c(s)\mathbf{I}\big)\mathrm{d}s,$$

where $\sum_j \mathbf{P}_j=\mathbf{I}$ is Theorem 12.8.1 (c) and the identification of $\sum_j(\mu_j+s)^{-1}\mathbf{P}_j$ with $(\mathbf{A}+s\mathbf{I})^{-1}$ is Definition 12.8.2, exactly as in Step 1 of Theorem 21.10.4. Every entry of the integrand is a fixed linear combination of the n convergent scalar integrals above, so the matrix integral converges. The same holds for $\mathbf{B}$ and for $\mathbf{C}$. Forming $\theta(\cdot)+(1-\theta)(\cdot)-(\cdot)$, the terms $c(s)\mathbf{I}$ cancel at each s, and the three convergent integrals combine into one:

$$\begin{aligned}&\theta(-\log\mathbf{A})+(1-\theta)(-\log\mathbf{B})-(-\log\mathbf{C})\\ &\quad=\int_0^\infty\big(\theta(\mathbf{A}+s\mathbf{I})^{-1}+(1-\theta)(\mathbf{B}+s\mathbf{I})^{-1}-(\mathbf{C}+s\mathbf{I})^{-1}\big)\mathrm{d}s.\end{aligned}$$

Fix $s>0$ and put $g_s(t)=1/(t+s)$, so that $g_s(\mathbf{A})=(\mathbf{A}+s\mathbf{I})^{-1}$ by Definition 12.8.2. By Theorem 21.8.3 (d), g_s is operator convex on $(0,\infty)$, so the integrand is $\theta g_s(\mathbf{A})+(1-\theta)g_s(\mathbf{B})-g_s(\mathbf{C})\succeq\mathbf{O}$. It is continuous in s on $(0,\infty)$ and the integral converges, so Lemma 21.10.2 in its improper form makes the left side $\succeq\mathbf{O}$. Since n, the two matrices and θ were arbitrary, $-\log t$ is operator convex on $(0,\infty)$.

The function $t\log t$. Its functional calculus at $\mathbf{A}$ is $\mathbf{A}\log\mathbf{A}$, the product of $\mathbf{A}$ and $\log\mathbf{A}$,

by Theorem 12.8.3 (b). The same substitution, now with the second display, gives

$$\mathbf{A}\log\mathbf{A} = \int_0^\infty \bigl(c(s)\mathbf{A} - \mathbf{A}(\mathbf{A}+s\mathbf{I})^{-1}\bigr)\,\mathrm{d}s,$$

its entries again convergent, and likewise for $\mathbf{B}$ and $\mathbf{C}$. Forming the same combination, the terms $c(s)\mathbf{A}$ contribute $c(s)(\theta\mathbf{A}+(1-\theta)\mathbf{B}-\mathbf{C}) = \mathbf{O}$ at each s, so

$$\begin{aligned}&\theta\,\mathbf{A}\log\mathbf{A} + (1-\theta)\,\mathbf{B}\log\mathbf{B} - \mathbf{C}\log\mathbf{C}\\ &\quad= \int_0^\infty \bigl(\theta h_s(\mathbf{A}) + (1-\theta)h_s(\mathbf{B}) - h_s(\mathbf{C})\bigr)\,\mathrm{d}s,\end{aligned}$$

where $h_s(t) = -t/(t+s)$ and $h_s(\mathbf{A}) = -\mathbf{A}(\mathbf{A}+s\mathbf{I})^{-1}$. By Theorem 21.8.3 (d) the function $t/(t+s)$ is operator concave on $(0,\infty)$, so h_s is operator convex and each integrand is $\succeq \mathbf{O}$. By Lemma 21.10.2 again the left side is $\succeq \mathbf{O}$, and $t\log t$ is operator convex on $(0,\infty)$.

■

Both proofs quote the fundamental theorem of calculus once, through the representation of $\log t$ that the remark verified with it; nothing else outside Chapters 1–20 enters, and in particular Theorem 21.10.3 does not. With this, the two functions §08 left over are on its list.

The companion negative fact is the one Chapter 13 §05 named beside it.

Example. The exponential is not operator monotone

Show that $t \mapsto e^t$ is not operator monotone on $\mathbb{R}$, and exhibit two 2×2 Hermitian matrices witnessing the failure.

Solution. Order two already fails. The exponential is infinitely differentiable, so Theorem 21.9.5 applies. Take $t_1 = 1$ and $t_2 = -1$. The divided differences (Definition 21.9.1) are $f[1,1] = e$, $f[-1,-1] = e^{-1}$ and

$$f[1,-1] = \frac{e - e^{-1}}{1-(-1)} = \sinh 1,$$

so the Loewner matrix and its determinant are

$$\begin{aligned}L_{\exp}(1,-1) &= \begin{pmatrix} e & \sinh 1 \\ \sinh 1 & e^{-1}\end{pmatrix},\\ \det L_{\exp}(1,-1) &= 1 - (\sinh 1)^2.\end{aligned}$$

Subtracting the series for e^{-1} from that for e leaves only the odd terms, so $\sinh 1 = 1 + \tfrac{1}{3!} + \tfrac{1}{5!} + \cdots > 1$. The determinant is therefore negative, and $L_{\exp}(1,-1)$ is not positive semidefinite. By Theorem 21.9.5, exp is not monotone of order 2, hence not operator monotone.

An explicit pair. Take

$$\mathbf{A} = \begin{pmatrix} 2 & 1 \\ 1 & 0\end{pmatrix}, \qquad \mathbf{B} = \begin{pmatrix} 1 & 0 \\ 0 & -1\end{pmatrix},$$

so that $\mathbf{A}-\mathbf{B} = \left(\begin{smallmatrix} 1 & 1 \\ 1 & 1\end{smallmatrix}\right) \succeq \mathbf{O}$ and $\mathbf{A} \succeq \mathbf{B}$. Write $\mathbf{A} = \mathbf{I} + \mathbf{N}$ with $\mathbf{N} = \left(\begin{smallmatrix} 1 & 1 \\ 1 & -1\end{smallmatrix}\right)$. Then $\mathbf{N}^2 = 2\mathbf{I}$,

so $\mathbf{N}^{2m} = 2^m\mathbf{I}$ and $\mathbf{N}^{2m+1} = 2^m\mathbf{N}$, and summing the series of Definition 10.9.2 in two halves,

$$e^{\mathbf{N}} = (\cosh\sqrt{2})\,\mathbf{I} + \frac{\sinh\sqrt{2}}{\sqrt{2}}\,\mathbf{N}.$$

Since $\mathbf{I}$ and $\mathbf{N}$ commute, $e^{\mathbf{A}} = e \cdot e^{\mathbf{N}}$ by Theorem 10.9.4 (d), while $e^{\mathbf{B}} = \operatorname{diag}(e, e^{-1})$. Writing $c = \cosh\sqrt{2}$ and $d = (\sinh\sqrt{2})/\sqrt{2}$, a direct expansion of the 2×2 determinant, using $c^2 - 2d^2 = \cosh^2\sqrt{2} - \sinh^2\sqrt{2} = 1$, gives

$$\det(e^{\mathbf{A}} - e^{\mathbf{B}}) = e^2(1 - c + d) + (1 - c - d).$$

With $c = 2.17818\ldots$ and $d = 1.36830\ldots$ this is $-1.1417\ldots$, a negative number, so $e^{\mathbf{A}} - e^{\mathbf{B}}$ has one positive and one negative eigenvalue and $e^{\mathbf{A}} \not\succeq e^{\mathbf{B}}$.

Remark. The two halves of Chapter 13 §05's "companion fact" are now both in hand: log is operator monotone on $(0, \infty)$ (Corollary 21.10.9) and e^t is not operator monotone (Example (The exponential is not operator monotone)). Neither proof used Theorem 21.10.3. The logarithm and the exponential are inverse to each other and increasing, which shows that operator monotonicity is not inherited by inverse functions.

21.10.6 Exercises

21.10.6.1 A. Check your understanding

Exercise. A1

(a) State what Theorem 21.10.3 gives, and say which implication of Theorem 21.10.4 rests on it.

(b) Explain in one sentence why an integral of positive semidefinite matrices is positive semidefinite, naming the two properties of the positive semidefinite matrices that the argument uses.

(c) Decide whether the following is correct, and justify your answer: "a function that extends to an entire function of z and is increasing on $(0, \infty)$ is operator monotone there".

(d) For which real p is $t \mapsto t^p$ operator monotone on $(0, \infty)$? Name the result and say which of its cases needed (A7).

(e) Where in the proof of Theorem 21.10.4 is the hypothesis $\beta \geq 0$ used, and what would go wrong without it?

Solution.

(a) It gives two things: that positivity of all the Loewner matrices of f forces f to extend analytically across the interval preserving the upper half-plane, and that such a function has the integral representation (ii); a third clause says an operator monotone function is automatically C^1. The implications (i) $\Rightarrow$ (iii) and (iii) $\Rightarrow$ (ii) rest on it. The implication (ii) $\Rightarrow$ (i) does not.

(b) Every Riemann sum is a sum of nonnegative multiples of the values of the integrand,

and the integral is a limit of such sums: the positive semidefinite matrices are closed under nonnegative combinations and closed under limits (Lemma 21.10.2).

(c) Incorrect. e^t is entire and increasing on $(0,\infty)$, and is not operator monotone there: Example (The exponential is not operator monotone) gives a witness with spectra in $\mathbb{R}$, and Exercise (B3: Which are operator monotone?) (vi) shifts it into $(0,\infty)$. Condition (iii) of Theorem 21.10.4 asks in addition that the upper half-plane be mapped into its closure, which e^z fails at $z = 3\pi i/2$.

(d) Exactly for $0 \le p \le 1$ (Corollary 21.10.8). No case needed (A7): the positive cases came from Lemma 21.10.5 and Lemma 21.10.2, and the failure for $p > 1$ from Lemma 21.10.6 together with Example (Squaring does not preserve the Loewner order).

(e) In the final display of Step 1, where $\beta\,\mathbf{x}^*(\mathbf{A}-\mathbf{B})\mathbf{x}$ is asserted to be ≥ 0. With $\beta < 0$ that term is ≤ 0 and can dominate: at $n = 1$ the function $f(t) = \beta t$ with $\beta < 0$ is decreasing, hence not operator monotone (Proposition 17.11.1).

21.10.6.2 B. Practice

Exercise. B1: One matrix integral

Let

$$\mathbf{M}(t) = \begin{pmatrix} 2+t & 1 \\ 1 & 2-t \end{pmatrix}, \qquad t \in [-1,1].$$

Show that $\mathbf{M}(t) \succeq \mathbf{O}$ for every $t \in [-1,1]$, compute $\int_{-1}^{1} \mathbf{M}(t)\,\mathrm{d}t$, and confirm the conclusion of Lemma 21.10.2 for it.

Solution. $\mathbf{M}(t)$ is real symmetric. Its leading entry is $2+t \ge 1 > 0$ and its determinant is $(2+t)(2-t) - 1 = 3 - t^2 \ge 2 > 0$ for $|t| \le 1$, so $\mathbf{M}(t) \succ \mathbf{O}$ by Sylvester's criterion (Theorem 13.1.5 (d)); in particular $\mathbf{M}(t) \succeq \mathbf{O}$.

Integrating entrywise, $\int_{-1}^{1}(2 \pm t)\,\mathrm{d}t = 4$ since $\int_{-1}^{1} t\,\mathrm{d}t = 0$, and $\int_{-1}^{1} 1\,\mathrm{d}t = 2$. So

$$\int_{-1}^{1} \mathbf{M}(t)\,\mathrm{d}t = \begin{pmatrix} 4 & 2 \\ 2 & 4 \end{pmatrix},$$

whose eigenvalues are 6 and 2, both positive; so it is $\succeq \mathbf{O}$ by Theorem 13.1.4 (b), as Lemma 21.10.2 predicts.

Exercise. B2: The constant for the square root

(a) Show that $K_{1/2} = \int_0^\infty u^{-1/2}(1+u)^{-1}\mathrm{d}u = \pi$. *Hint: substitute* $u = w^2$.

(b) Let $\mathbf{A} = \left(\begin{smallmatrix} 5 & 4 \\ 4 & 5 \end{smallmatrix}\right)$. Using Lemma 21.10.5 with $p = \frac{1}{2}$, evaluate

$$\frac{1}{\pi}\int_0^\infty s^{-1/2}\mathbf{A}(\mathbf{A}+s\mathbf{I})^{-1}\,\mathrm{d}s$$

and check the answer against the positive square root of $\mathbf{A}$.

Solution.

(a) With $u = w^2$, so $\mathrm{d}u = 2w\,\mathrm{d}w$ and $u^{-1/2} = 1/w$ for $w > 0$, the substitution rule of one-variable calculus gives

$$\int_0^\infty \frac{u^{-1/2}}{1+u}\mathrm{d}u = \int_0^\infty \frac{1}{w}\cdot\frac{2w}{1+w^2}\mathrm{d}w = 2\int_0^\infty \frac{\mathrm{d}w}{1+w^2}.$$

The last integral is $\lim_{R\to\infty}\arctan R - \arctan 0 = \pi/2$, by the fundamental theorem of calculus with the antiderivative arctan. Hence $K_{1/2} = \pi$. (Both calculus facts are quoted, not proved here.)

(b) $\mathbf{A}$ has eigenvalues 9 and 1, with orthonormal eigenvectors $\frac{1}{\sqrt{2}}(1,1)$ and $\frac{1}{\sqrt{2}}(1,-1)$, so its spectral resolution is $\mathbf{A} = 9\mathbf{P}_1 + \mathbf{P}_2$ with

$$\mathbf{P}_1 = \tfrac{1}{2}\begin{pmatrix}1 & 1\\ 1 & 1\end{pmatrix}, \qquad \mathbf{P}_2 = \tfrac{1}{2}\begin{pmatrix}1 & -1\\ -1 & 1\end{pmatrix}.$$

By the computation in the proof of Corollary 21.10.8, the integral equals $K_{1/2}\mathbf{A}^{1/2}$, so dividing by π gives

$$\mathbf{A}^{1/2} = 9^{1/2}\mathbf{P}_1 + 1^{1/2}\mathbf{P}_2 = 3\mathbf{P}_1 + \mathbf{P}_2 = \begin{pmatrix}2 & 1\\ 1 & 2\end{pmatrix}.$$

Squaring confirms it: $\left(\begin{smallmatrix}2 & 1\\ 1 & 2\end{smallmatrix}\right)^2 = \left(\begin{smallmatrix}5 & 4\\ 4 & 5\end{smallmatrix}\right) = \mathbf{A}$, and the matrix is positive definite, so it is *the* positive square root (Theorem 13.2.1).

Exercise. B3: Which are operator monotone?

Determine which of the following are operator monotone on $(0,\infty)$. Justify your answer.

(i) $f(t) = t^{2/3}$.
(ii) $f(t) = t^{4/3}$.
(iii) $f(t) = 3 + 2\log t$.
(iv) $f(t) = -t^{-1/3}$.
(v) $f(t) = \log(1+t)$.
(vi) $f(t) = e^t$.

Solution.

(i) Operator monotone: $0 \le \frac{2}{3} \le 1$, so Corollary 21.10.8 applies.

(ii) Not operator monotone: $\frac{4}{3} > 1$ (Corollary 21.10.8).

(iii) Operator monotone: log is (Corollary 21.10.9), and $3+2\log t$ is $\alpha + c\log t$ with $c = 2 \ge 0$, so Proposition 17.11.4 (a) applies.

(iv) Operator monotone. The function $g(t) = t^{1/3}$ is operator monotone on $(0,\infty)$ with $g((0,\infty)) = (0,\infty)$, and $h(u) = -1/u$ is operator monotone on $(0,\infty)$ (Proposition 17.11.3); by Lemma 21.10.6 the composite $h(g(t)) = -t^{-1/3}$ is operator monotone.

(v) Operator monotone. The affine map $g(t) = t + 1$ is operator monotone on $(0,\infty)$ (Proposition 17.11.2) with image $(1,\infty) \subseteq (0,\infty)$, and log is operator monotone on

$(0,\infty)$; apply Lemma 21.10.6.

(vi) Not operator monotone, by Example (The exponential is not operator monotone). (The witness there has spectra in $\mathbb{R}$, not $(0,\infty)$; restrict it by replacing $\mathbf{A}$ and $\mathbf{B}$ with $\mathbf{A}+2\mathbf{I}$ and $\mathbf{B}+2\mathbf{I}$, which are $\succ \mathbf{O}$ and satisfy $e^{\mathbf{A}+2\mathbf{I}} = e^2 e^{\mathbf{A}}$, so the same determinant is negative after multiplication by $e^4 > 0$.)

21.10.6.3 C. Going deeper

Exercise. C1: When the integral loses information

(a) Give a continuous $\mathbf{M}\colon [0,1] \to M_2(\mathbb{R})$ with every $\mathbf{M}(t)$ symmetric, $\int_0^1 \mathbf{M} \succeq \mathbf{O}$, and $\mathbf{M}(t) \not\succeq \mathbf{O}$ for some t. Hence the converse of Lemma 21.10.2 fails.

(b) Prove that if $\mathbf{M}\colon [a,b] \to M_n(\mathbb{C})$ is continuous with $\mathbf{M}(t) \succeq \mathbf{O}$ for every t and $\int_a^b \mathbf{M} = \mathbf{O}$, then $\mathbf{M}(t) = \mathbf{O}$ for every t.

Hint for (b): a continuous nonnegative scalar function with zero integral is zero.

Solution.

(a) Take $\mathbf{M}(t) = \operatorname{diag}(1, 2t-1)$. It is continuous and symmetric, $\mathbf{M}(0) = \operatorname{diag}(1,-1)$ is indefinite, and $\int_0^1 \mathbf{M} = \operatorname{diag}(1,0) \succeq \mathbf{O}$.

(b) Fix $\mathbf{x} \in \mathbb{C}^n$ and put $g(t) = \mathbf{x}^* \mathbf{M}(t) \mathbf{x}$, a continuous real-valued function with $g \geq 0$. Because pairing passes inside the integral, $\int_a^b g = \mathbf{x}^* \big(\int_a^b \mathbf{M}\big)\mathbf{x} = 0$. Suppose $g(t_0) > 0$ for some $t_0 \in [a,b]$. By continuity there is a subinterval $[c,d] \subseteq [a,b]$ with $d > c$ on which $g \geq g(t_0)/2$. By interval additivity, $\int_a^b g = \int_a^c g + \int_c^d g + \int_d^b g$, with the outer terms absent if $c = a$ or $d = b$; each outer term is ≥ 0, because $g \geq 0$ and so every Riemann sum over that piece is ≥ 0. On $[c,d]$ every Riemann sum of g is at least $\frac{1}{2} g(t_0)(d-c)$, since each tag value is at least $\frac{1}{2} g(t_0)$ and the interval lengths sum to $d-c$; a nonstrict inequality survives a limit, so $\int_c^d g$ is at least that too. Hence

$$\int_a^b g \;\geq\; \int_c^d g \;\geq\; \frac{g(t_0)}{2}(d-c) \;>\; 0,$$

a contradiction. So $g \equiv 0$, that is $\mathbf{x}^* \mathbf{M}(t) \mathbf{x} = 0$ for every t and every $\mathbf{x}$. A positive semidefinite matrix with $\mathbf{x}^* \mathbf{M} \mathbf{x} = 0$ for every $\mathbf{x}$ is $\mathbf{O}$: by Theorem 13.1.4 (c) write $\mathbf{M}(t) = \mathbf{C}^* \mathbf{C}$, so $\|\mathbf{C}\mathbf{x}\|^2 = 0$ for every $\mathbf{x}$, hence $\mathbf{C} = \mathbf{O}$.

Exercise. C2: The cube, by a Loewner matrix

(a) Write down the Loewner matrix $L_f(t_1, t_2)$ for $f(t) = t^3$ and general $t_1 \neq t_2$, and evaluate its determinant at $t_1 = 1$, $t_2 = 2$. What does Theorem 21.9.5 conclude?

(b) Do the same for $f(t) = t^{1/2}$ at $t_1 = 1$, $t_2 = 4$, and say which result of this section the sign of the determinant illustrates.

Solution.

(a) For $f(t) = t^3$ we have $f'(t) = 3t^2$ and, factoring the difference of cubes,

$$f[t_1, t_2] = \frac{t_1^3 - t_2^3}{t_1 - t_2} = t_1^2 + t_1 t_2 + t_2^2,$$

so $L_f(t_1, t_2)$ has diagonal $(3t_1^2, 3t_2^2)$ and off-diagonal $t_1^2 + t_1 t_2 + t_2^2$ (Definition 21.9.2). At $t_1 = 1$, $t_2 = 2$ the entries are $3, 12$ and $1 + 2 + 4 = 7$, so

$$\det L_f(1, 2) = 3 \cdot 12 - 7^2 = 36 - 49 = -13 < 0.$$

So $L_f(1, 2) \not\succeq \mathbf{O}$, and by Theorem 21.9.5 $t \mapsto t^3$ is not monotone of order 2, hence not operator monotone on $(0, \infty)$. This agrees with Corollary 21.10.8, since $3 > 1$.

(b) For $f(t) = t^{1/2}$, $f'(t) = \frac{1}{2} t^{-1/2}$ and

$$f[t_1, t_2] = \frac{\sqrt{t_1} - \sqrt{t_2}}{t_1 - t_2} = \frac{1}{\sqrt{t_1} + \sqrt{t_2}}.$$

At $t_1 = 1$, $t_2 = 4$ the matrix is $\left(\begin{smallmatrix} 1/2 & 1/3 \\ 1/3 & 1/4 \end{smallmatrix}\right)$, with determinant $\frac{1}{8} - \frac{1}{9} = \frac{1}{72} > 0$ and positive diagonal, so it is positive semidefinite. That is consistent with Corollary 21.10.8 at $p = \frac{1}{2}$: an operator monotone function has all its Loewner matrices positive semidefinite, for every size and all points, and this is one instance of it.

Exercise. C3: Half of the representation, by hand

Let $0 < p < 1$, let $\mathbf{A} \succeq \mathbf{B} \succ \mathbf{O}$ be Hermitian in $M_n(F)$, and let $0 < \delta < R$. Put

$$\mathbf{Z}(\delta, R) = \int_\delta^R s^{p-1} (\mathbf{A}(\mathbf{A} + s\mathbf{I})^{-1} - \mathbf{B}(\mathbf{B} + s\mathbf{I})^{-1}) \mathrm{d}s.$$

(a) Prove that $\mathbf{Z}(\delta, R) \succeq \mathbf{O}$ for every such δ, R, **without** any statement about improper integrals.

(b) Deduce $\mathbf{A}^p \succeq \mathbf{B}^p$ from (a) and the convergence statement of Lemma 21.10.5.

(c) Explain why the same argument does not show $\mathbf{A}^2 \succeq \mathbf{B}^2$, by pointing at the step that fails.

Solution.

(a) For each $s \in [\delta, R]$, the function $t \mapsto t/(t+s)$ is operator monotone on $(0, \infty)$ by Corollary 17.11.5, and it sends $\mathbf{A}$ to $\mathbf{A}(\mathbf{A} + s\mathbf{I})^{-1}$ and $\mathbf{B}$ to $\mathbf{B}(\mathbf{B} + s\mathbf{I})^{-1}$ (Definition 12.8.2). So the bracket is $\succeq \mathbf{O}$, and $s^{p-1} > 0$ keeps it so. The integrand is continuous on the compact interval $[\delta, R]$, because each entry is a rational function of s with nonvanishing denominator times s^{p-1}. Hence Lemma 21.10.2 gives $\mathbf{Z}(\delta, R) \succeq \mathbf{O}$.

(b) By Lemma 21.10.5 and the diagonalization in the proof of Corollary 21.10.8, $\mathbf{Z}(\delta, R) \to K_p(\mathbf{A}^p - \mathbf{B}^p)$ entrywise as $\delta \to 0^+$ and $R \to \infty$. For each $\mathbf{x}$, the numbers $\mathbf{x}^* \mathbf{Z}(\delta, R) \mathbf{x}$ are ≥ 0 by (a), and a nonstrict inequality survives a limit, so $K_p\, \mathbf{x}^*(\mathbf{A}^p - \mathbf{B}^p)\mathbf{x} \geq 0$; as $K_p > 0$, $\mathbf{A}^p \succeq \mathbf{B}^p$.

(c) The failure is at the very start. For $p = 2$ the integral $\int_0^\infty s^{p-1} \lambda/(\lambda + s)\, \mathrm{d}s$ diverges at $s = \infty$: the integrand behaves like $\lambda s^{p-2} = \lambda$, which is not integrable there. There is no

representation of t^2 as a positive superposition of the functions $t/(t+s)$ to begin with — and by Corollary 21.10.8 there cannot be one, since t^2 is not operator monotone.

21.11 Means and Matrix Cauchy–Schwarz

Two inequalities about numbers refuse to die: $|\bar{a}b|^2 \le |a|^2|b|^2$, and $2|\bar{a}b| \le |a|^2 + |b|^2$. For matrices the products do not commute and the absolute values are matrices themselves, so it is not obvious what either statement should say. Sections 4 and 6 have given the right language: measure a matrix by a unitarily invariant norm, and compare two matrices by weak majorization of their singular values. In that language both inequalities survive, for **every** unitarily invariant norm at once. The second half of the section then builds a genuine geometric mean of two positive definite matrices, which is where two order facts from the earlier chapters pay off again: the operator monotonicity of $t \mapsto t^{1/2}$ from Chapter 13 §05 and that of $t \mapsto -1/t$ from Chapter 17 §11.

Throughout, $F = \mathbb{R}$ or $F = \mathbb{C}$ and all matrices are $n \times n$ over F, with $|||\cdot|||$ an arbitrary unitarily invariant norm on $M_n(F)$ (Definition 21.4.1). As always $\sigma_1(\mathbf{A}) \ge \cdots \ge \sigma_n(\mathbf{A}) \ge 0$ are the singular values in decreasing order and $\lambda_1(\mathbf{A}) \ge \cdots \ge \lambda_n(\mathbf{A})$ the eigenvalues of a Hermitian $\mathbf{A}$; for $\mathbf{A} \succeq \mathbf{O}$ the two lists agree. The order $\succeq$ is the Loewner order (Definition 13.5.1).

21.11.1 An off-diagonal block is small

Everything in the first half comes from one observation about a positive semidefinite matrix cut into four blocks: the corner blocks control the off-diagonal one, and they do so eigenvalue by eigenvalue.

Lemma 21.11.1. The Off-Diagonal Block of a Positive Semidefinite Matrix

Let $k, l \ge 1$, let $\mathbf{P} \in M_k(F)$, $\mathbf{Q} \in M_l(F)$, and $\mathbf{X} \in M_{k\times l}(F)$, and suppose that

$$\mathbf{M} = \begin{pmatrix} \mathbf{P} & \mathbf{X} \\ \mathbf{X}^* & \mathbf{Q} \end{pmatrix} \succeq \mathbf{O}.$$

Then $2\sigma_j(\mathbf{X}) \le \lambda_j(\mathbf{M})$ for every $j = 1, \dots, \min(k, l)$.

Idea. Flipping the sign of the second block of coordinates is a unitary congruence, so it turns $\mathbf{M}$ into another positive semidefinite matrix — the same one with $\mathbf{X}$ replaced by $-\mathbf{X}$. Half the difference of the two is the Hermitian dilation of $\mathbf{X}$, whose eigenvalues are exactly $\pm\sigma_i(\mathbf{X})$ and zeros. Since what was subtracted is itself positive semidefinite, the dilation is dominated by $\frac{1}{2}\mathbf{M}$ in the Loewner order, and Chapter 17 §02 turns that into a comparison of eigenvalues index by index.

Proof. Put $\mathbf{S} = \mathbf{I}_k \oplus (-\mathbf{I}_l)$, which is Hermitian and unitary. By Theorem 8.1.1,

$$\mathbf{S}^*\mathbf{M}\mathbf{S} = \begin{pmatrix} \mathbf{P} & -\mathbf{X} \\ -\mathbf{X}^* & \mathbf{Q} \end{pmatrix},$$

and $\mathbf{S}^*\mathbf{M}\mathbf{S} \succeq \mathbf{O}$ because a congruence by an invertible matrix preserves positive semidefi-

niteness (Proposition 13.1.2 (a)). Subtracting,

$$\tfrac{1}{2}(\mathbf{M} - \mathbf{S}^{*}\mathbf{M}\mathbf{S}) = \begin{pmatrix} \mathbf{O} & \mathbf{X} \\ \mathbf{X}^{*} & \mathbf{O} \end{pmatrix} = \mathcal{H}(\mathbf{X}),$$

the Hermitian dilation of $\mathbf{X}$ (Proposition 17.9.3). Since $\tfrac{1}{2}\mathbf{S}^{*}\mathbf{M}\mathbf{S} \succeq \mathbf{O}$, this says

$$\mathcal{H}(\mathbf{X}) = \tfrac{1}{2}\mathbf{M} - \tfrac{1}{2}\mathbf{S}^{*}\mathbf{M}\mathbf{S} \preceq \tfrac{1}{2}\mathbf{M}.$$

By Corollary 17.2.4 the Loewner order compares every eigenvalue, so $\lambda_j(\mathcal{H}(\mathbf{X})) \le \lambda_j(\tfrac{1}{2}\mathbf{M}) = \tfrac{1}{2}\lambda_j(\mathbf{M})$ for every j.

Finally, by Proposition 17.9.3, $\mathcal{H}(\mathbf{X})$ has eigenvalues $\pm\sigma_1(\mathbf{X}), \dots, \pm\sigma_p(\mathbf{X})$ and $k + l - 2p$ zeros, with multiplicity, where $p = \min(k, l)$. Sorted decreasingly, the first p of them are $\sigma_1(\mathbf{X}) \ge \cdots \ge \sigma_p(\mathbf{X})$, since every remaining entry is $\le 0 \le \sigma_p(\mathbf{X})$. Hence $\lambda_j(\mathcal{H}(\mathbf{X})) = \sigma_j(\mathbf{X})$ for $j \le p$, and the displayed inequality reads $2\sigma_j(\mathbf{X}) \le \lambda_j(\mathbf{M})$.

■

21.11.2 A product against a sum of squares

The lemma becomes an inequality about products as soon as one remembers where positive semidefinite block matrices come from: $\mathbf{D}^{*}\mathbf{D}$, for $\mathbf{D}$ the block row built from the two matrices in question.

Theorem 21.11.2. Arithmetic–Geometric Mean Inequality for Matrices

Let $\mathbf{A}, \mathbf{B} \in M_n(F)$. Then

$$2\sigma_j(\mathbf{A}^{*}\mathbf{B}) \le \lambda_j(\mathbf{A}\mathbf{A}^{*} + \mathbf{B}\mathbf{B}^{*}) \qquad (j = 1, \dots, n),$$

and consequently, for every unitarily invariant norm on $M_n(F)$,

$$2\,|||\mathbf{A}^{*}\mathbf{B}||| \le |||\mathbf{A}\mathbf{A}^{*} + \mathbf{B}\mathbf{B}^{*}|||.$$

Idea. Put the two matrices side by side in one $n \times 2n$ block row $\mathbf{D} = (\mathbf{A} \;\; \mathbf{B})$. Then $\mathbf{D}^{*}\mathbf{D}$ is the $2n \times 2n$ positive semidefinite matrix whose off-diagonal block is exactly $\mathbf{A}^{*}\mathbf{B}$, so Lemma 21.11.1 applies; and $\mathbf{D}\mathbf{D}^{*}$, which has the same nonzero eigenvalues, is the sum $\mathbf{A}\mathbf{A}^{*} + \mathbf{B}\mathbf{B}^{*}$. The norm statement is then the singular value statement read through §06's dominance theorem.

Proof. Let $\mathbf{D} = (\mathbf{A} \;\; \mathbf{B}) \in M_{n \times 2n}(F)$, the block row with $\mathbf{A}$ then $\mathbf{B}$. By Theorem 8.1.1,

$$\mathbf{D}^{*}\mathbf{D} = \begin{pmatrix} \mathbf{A}^{*}\mathbf{A} & \mathbf{A}^{*}\mathbf{B} \\ \mathbf{B}^{*}\mathbf{A} & \mathbf{B}^{*}\mathbf{B} \end{pmatrix}, \qquad \mathbf{D}\mathbf{D}^{*} = \mathbf{A}\mathbf{A}^{*} + \mathbf{B}\mathbf{B}^{*}.$$

The matrix $\mathbf{D}^{*}\mathbf{D}$ is Hermitian and satisfies $\mathbf{x}^{*}\mathbf{D}^{*}\mathbf{D}\mathbf{x} = \|\mathbf{D}\mathbf{x}\|^2 \ge 0$ for every $\mathbf{x} \in F^{2n}$, so $\mathbf{D}^{*}\mathbf{D} \succeq \mathbf{O}$ (Definition 13.1.1). Its off-diagonal block is $\mathbf{A}^{*}\mathbf{B}$, so Lemma 21.11.1 with $k = l = n$ gives

$$2\sigma_j(\mathbf{A}^{*}\mathbf{B}) \le \lambda_j(\mathbf{D}^{*}\mathbf{D}) \qquad (j = 1, \dots, n).$$

It remains to identify $\lambda_j(\mathbf{D}^*\mathbf{D})$ with $\lambda_j(\mathbf{D}\mathbf{D}^*)$ for $j \le n$. Applying Theorem 9.10.4 to the pair $\mathbf{D}^* \in M_{2n\times n}$ and $\mathbf{D} \in M_{n\times 2n}$ gives $x^n p_{\mathbf{D}^*\mathbf{D}}(x) = x^{2n} p_{\mathbf{D}\mathbf{D}^*}(x)$, that is $p_{\mathbf{D}^*\mathbf{D}} = x^n p_{\mathbf{D}\mathbf{D}^*}$: the eigenvalue list of $\mathbf{D}^*\mathbf{D}$ is that of $\mathbf{D}\mathbf{D}^*$ with n extra zeros. Both matrices are positive semidefinite, so all their eigenvalues are ≥ 0 (Theorem 13.1.4 (b)) and the first n entries of the two decreasing lists agree. This proves the first display.

For the second, both $2\mathbf{A}^*\mathbf{B}$ and $\mathbf{A}\mathbf{A}^* + \mathbf{B}\mathbf{B}^*$ lie in $M_n(F)$, and the singular values of the positive semidefinite $\mathbf{A}\mathbf{A}^* + \mathbf{B}\mathbf{B}^*$ are its eigenvalues. So the first display says $\sigma_j(2\mathbf{A}^*\mathbf{B}) \le \sigma_j(\mathbf{A}\mathbf{A}^* + \mathbf{B}\mathbf{B}^*)$ for every j. Two decreasing lists of nonnegative numbers with the first dominated entry by entry satisfy $\sigma(2\mathbf{A}^*\mathbf{B}) \prec_w \sigma(\mathbf{A}\mathbf{A}^* + \mathbf{B}\mathbf{B}^*)$, since each top-$k$ partial sum is dominated term by term (Definition 17.8.1). Now Theorem 21.6.1 turns that weak majorization into $|||2\mathbf{A}^*\mathbf{B}||| \le |||\mathbf{A}\mathbf{A}^* + \mathbf{B}\mathbf{B}^*|||$, and $|||2\mathbf{A}^*\mathbf{B}||| = 2\,|||\mathbf{A}^*\mathbf{B}|||$ by homogeneity of a norm.

■

Warning. **The two sums are not interchangeable.** It is tempting to write $\mathbf{A}^*\mathbf{A} + \mathbf{B}^*\mathbf{B}$ on the right, since that is what pairs with $\mathbf{A}^*\mathbf{B}$ in the block matrix $\mathbf{D}^*\mathbf{D}$. The resulting statement is **false**. Take

$$\mathbf{A} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \qquad \mathbf{B} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}.$$

Then $\mathbf{A}^*\mathbf{A} = \operatorname{diag}(1,0)$ and $\mathbf{B}^*\mathbf{B} = \operatorname{diag}(0,1)$, so $\mathbf{A}^*\mathbf{A} + \mathbf{B}^*\mathbf{B} = \mathbf{I}$ and $\|\mathbf{I}\|_2 = 1$; while $\mathbf{A}^*\mathbf{B} = \mathbf{B}$ has $\|\mathbf{B}\|_2 = 1$, so $2\,|||\mathbf{A}^*\mathbf{B}||| = 2 > 1$ in the spectral norm. The correct right-hand side is $\mathbf{A}\mathbf{A}^* + \mathbf{B}\mathbf{B}^* = \operatorname{diag}(2,0)$, of spectral norm 2 — and here the inequality is an equality.

21.11.3 Cauchy–Schwarz

Theorem 21.11.3. Cauchy–Schwarz for Unitarily Invariant Norms

Let $\mathbf{A}, \mathbf{B} \in M_n(F)$. Then for every unitarily invariant norm on $M_n(F)$,

$$|||\mathbf{A}^*\mathbf{B}|||^2 \;\le\; |||\mathbf{A}^*\mathbf{A}|||\;\,|||\mathbf{B}^*\mathbf{B}|||.$$

Idea. The scalar inequality $2|\bar{a}b| \le |a|^2 + |b|^2$ becomes $|\bar{a}b| \le |a||b|$ by the classical trick of replacing a by ta and b by b/t and choosing t to balance the two terms. The product $\mathbf{A}^*\mathbf{B}$ is unchanged by that substitution, so the same trick upgrades Theorem 21.11.2. The one thing to check is that the two sides of the balance are the right quantities, and that is the identity $|||\mathbf{A}\mathbf{A}^*||| = |||\mathbf{A}^*\mathbf{A}|||$.

Proof. Put $a = |||\mathbf{A}^*\mathbf{A}|||$ and $b = |||\mathbf{B}^*\mathbf{B}|||$. If $a = 0$, then $\mathbf{A}^*\mathbf{A} = \mathbf{O}$ by the positivity clause of Definition 16.1.1, so $\|\mathbf{A}\mathbf{x}\|^2 = \mathbf{x}^*\mathbf{A}^*\mathbf{A}\mathbf{x} = 0$ for every $\mathbf{x}$ and $\mathbf{A} = \mathbf{O}$; then both sides of the asserted inequality are 0. Likewise if $b = 0$. So assume $a > 0$ and $b > 0$.

First, $|||\mathbf{A}\mathbf{A}^*||| = |||\mathbf{A}^*\mathbf{A}|||$. Indeed $\mathbf{A}\mathbf{A}^*$ and $\mathbf{A}^*\mathbf{A}$ are positive semidefinite with the same eigenvalues, namely $\sigma_i(\mathbf{A})^2$ (Definition 13.9.1 and Theorem 9.10.4 (b), the two matrices being

of equal size n), hence the same singular values; and a unitarily invariant norm depends only on the singular values (Lemma 21.4.1).

Let $t > 0$ and apply Theorem 21.11.2 to the pair $t\mathbf{A}$ and $t^{-1}\mathbf{B}$. Their product is $(t\mathbf{A})^*(t^{-1}\mathbf{B}) = \mathbf{A}^*\mathbf{B}$, unchanged, while

$$(t\mathbf{A})(t\mathbf{A})^* + (t^{-1}\mathbf{B})(t^{-1}\mathbf{B})^* = t^2\mathbf{A}\mathbf{A}^* + t^{-2}\mathbf{B}\mathbf{B}^*.$$

So, using the triangle inequality and homogeneity of $|||\cdot|||$ and then the identity just proved,

$$2\,|||\mathbf{A}^*\mathbf{B}||| \;\le\; t^2\,|||\mathbf{A}\mathbf{A}^*||| + t^{-2}\,|||\mathbf{B}\mathbf{B}^*||| = t^2 a + t^{-2} b.$$

By the arithmetic–geometric mean inequality for two positive numbers (Lemma 17.8.6), $t^2 a + t^{-2} b \ge 2\sqrt{ab}$, with equality when $t^2 a = t^{-2} b$; choosing $t = (b/a)^{1/4}$, which is legitimate since $a, b > 0$, makes the two terms equal and gives

$$2\,|||\mathbf{A}^*\mathbf{B}||| \;\le\; 2\sqrt{ab}.$$

Dividing by 2 and squaring proves the theorem.

■

Remark. Three familiar inequalities are special cases. With $|||\cdot||| = \|\cdot\|_F$ it is $\|\mathbf{A}^*\mathbf{B}\|_F^2 \le \|\mathbf{A}^*\mathbf{A}\|_F \|\mathbf{B}^*\mathbf{B}\|_F$; with the spectral norm it is $\|\mathbf{A}^*\mathbf{B}\|_2^2 \le \|\mathbf{A}\|_2^2 \|\mathbf{B}\|_2^2$, which also follows from submultiplicativity; and with the trace norm $|||\cdot|||_{(n)} = \sum_i \sigma_i(\cdot)$ it compares $\sum_i \sigma_i(\mathbf{A}^*\mathbf{B})$ with $\left(\sum_i \sigma_i(\mathbf{A})^2\right)^{1/2}\left(\sum_i \sigma_i(\mathbf{B})^2\right)^{1/2}$. At $n = 1$ all of them are $|\bar{a}b|^2 \le |a|^2|b|^2$.

21.11.4 The geometric mean of two positive definite matrices

The inequalities above were about *norms* of products. The rest of the section is about an operation: given $\mathbf{A}, \mathbf{B} \succ \mathbf{O}$, produce a positive definite matrix that deserves the name $\sqrt{\mathbf{A}\mathbf{B}}$.

The arithmetic mean gives no trouble: $\tfrac{1}{2}(\mathbf{A} + \mathbf{B})$ is positive definite, symmetric in the two arguments, and monotone in each. The geometric mean is the problem. The naive candidate is $(\mathbf{A}\mathbf{B})^{1/2}$, and it does not exist: $\mathbf{A}\mathbf{B}$ need not be Hermitian, so Theorem 13.2.1 has nothing to say about it. The next candidate, $\mathbf{A}^{1/2}\mathbf{B}^{1/2}$, is a matrix, but usually not a Hermitian one. What works is to measure $\mathbf{B}$ in the coordinates that make $\mathbf{A}$ the identity, take the square root there, and translate back.

The geometric mean of two positive definite matrices is the square root of the second one, computed after the first has been turned into the identity, and then translated back.

Definition 21.11.1. Geometric Mean of Positive Definite Matrices

Let $\mathbf{A}, \mathbf{B} \in M_n(F)$ with $\mathbf{A} \succ \mathbf{O}$ and $\mathbf{B} \succ \mathbf{O}$. Their **geometric mean** is

$$\mathbf{A}\#\mathbf{B} \;:=\; \mathbf{A}^{1/2}\left(\mathbf{A}^{-1/2}\mathbf{B}\mathbf{A}^{-1/2}\right)^{1/2}\mathbf{A}^{1/2},$$

where $\mathbf{A}^{1/2}$ is the positive square root of Theorem 13.2.1, $\mathbf{A}^{-1/2} := (\mathbf{A}^{1/2})^{-1}$, and the outer exponent $\frac{1}{2}$ is again the positive square root.

In words: conjugate $\mathbf{B}$ by $\mathbf{A}^{-1/2}$ — this is the **whitening** move of Chapter 13 §05, which replaces the pair $(\mathbf{A}, \mathbf{B})$ by $(\mathbf{I}, \mathbf{C})$ with $\mathbf{C} = \mathbf{A}^{-1/2}\mathbf{B}\mathbf{A}^{-1/2}$ — take the positive square root of $\mathbf{C}$, and undo the conjugation.

Everything in the definition makes sense. $\mathbf{A} \succ \mathbf{O}$ gives $\mathbf{A}^{1/2} \succ \mathbf{O}$ (its eigenvalues are the positive square roots of those of $\mathbf{A}$, by Theorem 13.2.1), so $\mathbf{A}^{1/2}$ is invertible and $\mathbf{A}^{-1/2}$ is Hermitian. Hence $\mathbf{C} = \mathbf{A}^{-1/2}\mathbf{B}\mathbf{A}^{-1/2} = (\mathbf{A}^{-1/2})^{*}\mathbf{B}\mathbf{A}^{-1/2}$ is a congruence of $\mathbf{B} \succ \mathbf{O}$ by an invertible matrix and is therefore $\succ \mathbf{O}$ (Proposition 13.1.2 (b)); so $\mathbf{C}^{1/2}$ exists, is unique, and is $\succ \mathbf{O}$. Finally $\mathbf{A}\#\mathbf{B} = (\mathbf{A}^{1/2})^{*}\mathbf{C}^{1/2}\mathbf{A}^{1/2}$ is another such congruence, so $\mathbf{A}\#\mathbf{B} \succ \mathbf{O}$. In particular $\mathbf{A}\#\mathbf{B}$ is Hermitian, which $\mathbf{A}^{1/2}\mathbf{B}^{1/2}$ is not.

Examples.

- **Numbers.** For $n = 1$ with $\mathbf{A} = (a)$, $\mathbf{B} = (b)$, and $a, b > 0$, the formula reads $\sqrt{a}\,(b/a)^{1/2}\sqrt{a} = \sqrt{ab}$. This is where the name comes from.
- **One argument the identity.** $\mathbf{A}\#\mathbf{I} = \mathbf{A}^{1/2}(\mathbf{A}^{-1})^{1/2}\mathbf{A}^{1/2} = \mathbf{A}^{1/2}\mathbf{A}^{-1/2}\mathbf{A}^{1/2} = \mathbf{A}^{1/2}$, using $(\mathbf{A}^{-1})^{1/2} = \mathbf{A}^{-1/2}$, which holds because $(\mathbf{A}^{-1/2})^2 = \mathbf{A}^{-1}$ and $\mathbf{A}^{-1/2} \succ \mathbf{O}$, so uniqueness in Theorem 13.2.1 applies. Likewise $\mathbf{A}\#\mathbf{A} = \mathbf{A}^{1/2}\mathbf{I}\mathbf{A}^{1/2} = \mathbf{A}$.
- **Commuting arguments.** If $\mathbf{A}\mathbf{B} = \mathbf{B}\mathbf{A}$ then $\mathbf{A}\#\mathbf{B} = (\mathbf{A}\mathbf{B})^{1/2}$. Indeed $\mathbf{A}^{1/2}$ commutes with $\mathbf{B}$ (Theorem 13.2.1 (b)), hence so does $\mathbf{A}^{-1/2}$, so $\mathbf{C} = \mathbf{B}\mathbf{A}^{-1}$ commutes with $\mathbf{A}$, and $\mathbf{C}^{1/2}$, being a polynomial in $\mathbf{C}$ (Theorem 13.2.1 (a)), commutes with $\mathbf{A}$ too. Therefore

$$\begin{aligned}(\mathbf{A}\#\mathbf{B})^2 &= \mathbf{A}^{1/2}\mathbf{C}^{1/2}\mathbf{A}\,\mathbf{C}^{1/2}\mathbf{A}^{1/2} = \mathbf{A}^{1/2}\mathbf{A}\,\mathbf{C}\,\mathbf{A}^{1/2} \\ &= \mathbf{A}^{1/2}\mathbf{A}\mathbf{B}\mathbf{A}^{-1}\mathbf{A}^{1/2} = \mathbf{A}^{1/2}\mathbf{B}\mathbf{A}^{1/2} = \mathbf{A}\mathbf{B},\end{aligned}$$

  using $\mathbf{C} = \mathbf{B}\mathbf{A}^{-1}$ and $\mathbf{A}\mathbf{B} = \mathbf{B}\mathbf{A}$ once more; and $\mathbf{A}\#\mathbf{B} \succ \mathbf{O}$, so it is the positive square root of $\mathbf{A}\mathbf{B}$. For instance $\operatorname{diag}(4,1)\#\operatorname{diag}(1,9) = \operatorname{diag}(2,3)$.
- **A noncommuting pair, exactly.** Let

$$\mathbf{A} = \begin{pmatrix} 5 & 4 \\ 4 & 5 \end{pmatrix}, \qquad \mathbf{B} = \begin{pmatrix} 8 & 10 \\ 10 & 17 \end{pmatrix}.$$

  Then $\mathbf{A}^{1/2} = \left(\begin{smallmatrix} 2 & 1 \\ 1 & 2 \end{smallmatrix}\right)$, and $\mathbf{A}^{-1/2} = \frac{1}{3}\left(\begin{smallmatrix} 2 & -1 \\ -1 & 2 \end{smallmatrix}\right)$, so $\mathbf{C} = \mathbf{A}^{-1/2}\mathbf{B}\mathbf{A}^{-1/2} = \operatorname{diag}(1,4)$ and $\mathbf{C}^{1/2} = \operatorname{diag}(1,2)$. Hence

$$\mathbf{A}\#\mathbf{B} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} = \begin{pmatrix} 6 & 6 \\ 6 & 9 \end{pmatrix}.$$

  Here $\mathbf{A}\mathbf{B} \neq \mathbf{B}\mathbf{A}$, as the $(1,2)$ entries 118 and 82 show. Note $\det(\mathbf{A}\#\mathbf{B}) = 18 = \sqrt{9 \cdot 36} = \sqrt{\det\mathbf{A}\,\det\mathbf{B}}$; Exercise C1 explains why.

A nonexample by minimal change. Drop the middle square root and take $\mathbf{A}^{1/2}\mathbf{B}^{1/2}$. With $\mathbf{A} = \operatorname{diag}(4,1)$ and $\mathbf{B} = \left(\begin{smallmatrix} 5 & 4 \\ 4 & 5 \end{smallmatrix}\right)$, the two square roots are $\operatorname{diag}(2,1)$ and $\left(\begin{smallmatrix} 2 & 1 \\ 1 & 2 \end{smallmatrix}\right)$, and

$$\mathbf{A}^{1/2}\mathbf{B}^{1/2} = \begin{pmatrix} 4 & 2 \\ 1 & 2 \end{pmatrix},$$

which is not Hermitian, so it is not $\succ \mathbf{O}$ and is not a mean of anything. Its square is $\left(\begin{smallmatrix} 18 & 12 \\ 6 & 6 \end{smallmatrix}\right)$, not $\mathbf{A}\mathbf{B} = \left(\begin{smallmatrix} 20 & 16 \\ 4 & 5 \end{smallmatrix}\right)$.

Warning. **Half of A times half of B is not a mean.** $\mathbf{A}^{1/2}\mathbf{B}^{1/2}$ is Hermitian only when $\mathbf{A}$ and $\mathbf{B}$ commute: $(\mathbf{A}^{1/2}\mathbf{B}^{1/2})^* = \mathbf{B}^{1/2}\mathbf{A}^{1/2}$, and if this equals $\mathbf{A}^{1/2}\mathbf{B}^{1/2}$, then squaring gives $\mathbf{AB} = \mathbf{BA}$. The sandwich in Definition 21.11.1 exists precisely to repair this, and the price is that the formula no longer *looks* symmetric in $\mathbf{A}$ and $\mathbf{B}$. It is; that is the content of Theorem 21.11.4 (b), and it is the one property here that needs an argument.

Quick check. Compute $\mathbf{A}\#\mathbf{B}$ for $\mathbf{A} = \operatorname{diag}(4,1)$ and $\mathbf{B} = \operatorname{diag}(1,9)$ straight from Definition 21.11.1, and check the answer against $(\mathbf{AB})^{1/2}$.

Solution. $\mathbf{A}^{1/2} = \operatorname{diag}(2,1)$ and $\mathbf{A}^{-1/2} = \operatorname{diag}(\tfrac{1}{2},1)$, so $\mathbf{C} = \mathbf{A}^{-1/2}\mathbf{B}\mathbf{A}^{-1/2} = \operatorname{diag}(\tfrac{1}{4},9)$ and $\mathbf{C}^{1/2} = \operatorname{diag}(\tfrac{1}{2},3)$. Hence

$$\mathbf{A}\#\mathbf{B} = \operatorname{diag}(2,1)\operatorname{diag}(\tfrac{1}{2},3)\operatorname{diag}(2,1) = \operatorname{diag}(2,3).$$

The two matrices commute, and $\mathbf{AB} = \operatorname{diag}(4,9)$, whose positive square root is $\operatorname{diag}(2,3)$. The two agree.

21.11.5 What the geometric mean does

The key is a description of $\mathbf{A}\#\mathbf{B}$ that never mentions $\mathbf{A}^{1/2}$. Chapter 13 §06's block test supplies it.

Theorem 21.11.4. Properties of the Geometric Mean

Let $\mathbf{A},\mathbf{B},\mathbf{A}',\mathbf{B}' \in M_n(F)$ be positive definite and let $\mathbf{S} \in M_n(F)$ be invertible.

(a) **Extremal characterization.** $\mathbf{A}\#\mathbf{B}$ is the largest Hermitian $\mathbf{X} \in M_n(F)$ for which

$$\begin{pmatrix} \mathbf{A} & \mathbf{X} \\ \mathbf{X} & \mathbf{B} \end{pmatrix} \succeq \mathbf{O}:$$

this holds for $\mathbf{X} = \mathbf{A}\#\mathbf{B}$, and every Hermitian $\mathbf{X}$ for which it holds satisfies $\mathbf{X} \preceq \mathbf{A}\#\mathbf{B}$.

(b) **Symmetry.** $\mathbf{A}\#\mathbf{B} = \mathbf{B}\#\mathbf{A}$.

(c) **Congruence.** $(\mathbf{S}^*\mathbf{A}\mathbf{S})\#(\mathbf{S}^*\mathbf{B}\mathbf{S}) = \mathbf{S}^*(\mathbf{A}\#\mathbf{B})\mathbf{S}$.

(d) **Monotonicity.** If $\mathbf{A} \succeq \mathbf{A}'$ and $\mathbf{B} \succeq \mathbf{B}'$, then $\mathbf{A}\#\mathbf{B} \succeq \mathbf{A}'\#\mathbf{B}'$.

(e) **Arithmetic–geometric mean, and inversion.** $\mathbf{A}\#\mathbf{B} \preceq \tfrac{1}{2}(\mathbf{A}+\mathbf{B})$, and $(\mathbf{A}\#\mathbf{B})^{-1} = \mathbf{A}^{-1}\#\mathbf{B}^{-1}$.

Idea. Everything rests on (a), and (a) rests on whitening. With $\mathbf{A} \succ \mathbf{O}$ the block test Theorem 13.6.1 turns the block condition into $\mathbf{B} - \mathbf{X}\mathbf{A}^{-1}\mathbf{X} \succeq \mathbf{O}$; substituting $\mathbf{X} = \mathbf{A}^{1/2}\mathbf{Y}\mathbf{A}^{1/2}$ and conjugating by $\mathbf{A}^{-1/2}$ turns *that* into $\mathbf{Y}^2 \preceq \mathbf{C}$, with $\mathbf{C} = \mathbf{A}^{-1/2}\mathbf{B}\mathbf{A}^{-1/2}$. The largest Hermitian $\mathbf{Y}$ with $\mathbf{Y}^2 \preceq \mathbf{C}$ is $\mathbf{C}^{1/2}$, because $\mathbf{Y} \preceq |\mathbf{Y}| = (\mathbf{Y}^2)^{1/2}$ and the square root is operator monotone. Once (a) is available, (b), (c), and (d) are all statements about the *set* of admissible $\mathbf{X}$, and they read off the symmetry, the congruence-invariance, and the

monotonicity of that set without any further computation.

Proof. Write $\mathbf{C} = \mathbf{A}^{-1/2}\mathbf{B}\mathbf{A}^{-1/2} \succ \mathbf{O}$, so that $\mathbf{A}\#\mathbf{B} = \mathbf{A}^{1/2}\mathbf{C}^{1/2}\mathbf{A}^{1/2}$, and for Hermitian $\mathbf{X}$ put

$$\mathbf{M}(\mathbf{X}) = \begin{pmatrix} \mathbf{A} & \mathbf{X} \\ \mathbf{X} & \mathbf{B} \end{pmatrix},$$

which is Hermitian because $\mathbf{X}^* = \mathbf{X}$. Call $\mathbf{X}$ **admissible** for $(\mathbf{A}, \mathbf{B})$ when $\mathbf{M}(\mathbf{X}) \succeq \mathbf{O}$.

(a). Since $\mathbf{A} \succ \mathbf{O}$, Theorem 13.6.1 (b) says that $\mathbf{X}$ is admissible exactly when the Schur complement $\mathbf{B} - \mathbf{X}\mathbf{A}^{-1}\mathbf{X}$ is $\succeq \mathbf{O}$. Every Hermitian $\mathbf{X}$ is uniquely $\mathbf{X} = \mathbf{A}^{1/2}\mathbf{Y}\mathbf{A}^{1/2}$ with $\mathbf{Y} = \mathbf{A}^{-1/2}\mathbf{X}\mathbf{A}^{-1/2}$ Hermitian, and then

$$\mathbf{X}\mathbf{A}^{-1}\mathbf{X} = \mathbf{A}^{1/2}\mathbf{Y}\mathbf{A}^{1/2}\mathbf{A}^{-1}\mathbf{A}^{1/2}\mathbf{Y}\mathbf{A}^{1/2} = \mathbf{A}^{1/2}\mathbf{Y}^2\mathbf{A}^{1/2}.$$

Conjugating by the invertible Hermitian $\mathbf{A}^{-1/2}$ and using Proposition 13.1.2 (a) in both directions,

$$\mathbf{B} - \mathbf{A}^{1/2}\mathbf{Y}^2\mathbf{A}^{1/2} \succeq \mathbf{O} \quad\Longleftrightarrow\quad \mathbf{C} - \mathbf{Y}^2 \succeq \mathbf{O}.$$

Claim. If $\mathbf{Y}$ is Hermitian and $\mathbf{Y}^2 \preceq \mathbf{C}$, then $\mathbf{Y} \preceq \mathbf{C}^{1/2}$.

Proof. Let $\mathbf{Y} = \sum_j \mu_j \mathbf{P}_j$ be the spectral resolution (Theorem 12.8.1), so that $\mathbf{Y}^2 = \sum_j \mu_j^2 \mathbf{P}_j$ and the matrix $\sum_j |\mu_j| \mathbf{P}_j$ is $\succeq \mathbf{O}$ with square $\mathbf{Y}^2$; by uniqueness in Theorem 13.2.1 it is $(\mathbf{Y}^2)^{1/2}$. Then

$$(\mathbf{Y}^2)^{1/2} - \mathbf{Y} = \sum_j (|\mu_j| - \mu_j)\mathbf{P}_j \;\succeq\; \mathbf{O},$$

because $\mathbf{x}^*\mathbf{P}_j\mathbf{x} = \|\mathbf{P}_j\mathbf{x}\|^2 \geq 0$ and every coefficient is ≥ 0. So $\mathbf{Y} \preceq (\mathbf{Y}^2)^{1/2}$. Also $\mathbf{Y}^2 \preceq \mathbf{C}$ with $\mathbf{Y}^2 \succeq \mathbf{O}$, so Proposition 13.5.4 gives $(\mathbf{Y}^2)^{1/2} \preceq \mathbf{C}^{1/2}$. Chaining the two (Proposition 13.5.1) gives $\mathbf{Y} \preceq \mathbf{C}^{1/2}$.

■

The choice $\mathbf{Y} = \mathbf{C}^{1/2}$ itself satisfies $\mathbf{C} - \mathbf{Y}^2 = \mathbf{O} \succeq \mathbf{O}$. So among the Hermitian $\mathbf{Y}$ with $\mathbf{C} - \mathbf{Y}^2 \succeq \mathbf{O}$ the largest is $\mathbf{C}^{1/2}$. Translating back by the order-preserving bijection $\mathbf{Y} \mapsto \mathbf{A}^{1/2}\mathbf{Y}\mathbf{A}^{1/2}$ (again Proposition 13.1.2 (a), in both directions), the largest admissible $\mathbf{X}$ is $\mathbf{A}^{1/2}\mathbf{C}^{1/2}\mathbf{A}^{1/2} = \mathbf{A}\#\mathbf{B}$.

(b). Let $\mathbf{W} = \left(\begin{smallmatrix} \mathbf{O} & \mathbf{I} \\ \mathbf{I} & \mathbf{O} \end{smallmatrix}\right)$, which is Hermitian and unitary, hence invertible. By Theorem 8.1.1,

$$\mathbf{W}^* \begin{pmatrix} \mathbf{A} & \mathbf{X} \\ \mathbf{X} & \mathbf{B} \end{pmatrix} \mathbf{W} = \begin{pmatrix} \mathbf{B} & \mathbf{X} \\ \mathbf{X} & \mathbf{A} \end{pmatrix},$$

so by Proposition 13.1.2 (a), applied to $\mathbf{W}$ and to $\mathbf{W}^{-1}$, a Hermitian $\mathbf{X}$ is admissible for $(\mathbf{A}, \mathbf{B})$ if and only if it is admissible for $(\mathbf{B}, \mathbf{A})$. The two admissible sets coincide, so they have the same largest element; by (a) that element is $\mathbf{A}\#\mathbf{B}$ on one side and $\mathbf{B}\#\mathbf{A}$ on the other, and a partial order has at most one largest element (Proposition 13.5.1).

(c). Put $\mathbf{T} = \mathbf{S} \oplus \mathbf{S}$. By Theorem 8.1.1,

$$\mathbf{T}^* \begin{pmatrix} \mathbf{A} & \mathbf{X} \\ \mathbf{X} & \mathbf{B} \end{pmatrix} \mathbf{T} = \begin{pmatrix} \mathbf{S}^*\mathbf{A}\mathbf{S} & \mathbf{S}^*\mathbf{X}\mathbf{S} \\ \mathbf{S}^*\mathbf{X}\mathbf{S} & \mathbf{S}^*\mathbf{B}\mathbf{S} \end{pmatrix},$$

and $\mathbf{T}$ is invertible, so $\mathbf{X}$ is admissible for $(\mathbf{A},\mathbf{B})$ exactly when $\mathbf{S}^*\mathbf{X}\mathbf{S}$ is admissible for $(\mathbf{S}^*\mathbf{A}\mathbf{S},\mathbf{S}^*\mathbf{B}\mathbf{S})$ (Proposition 13.1.2 (a), both ways). The map $\mathbf{X}\mapsto\mathbf{S}^*\mathbf{X}\mathbf{S}$ is a bijection of the Hermitian matrices that preserves $\preceq$ in both directions, so it carries the largest element of one admissible set to the largest element of the other. Both $\mathbf{S}^*\mathbf{A}\mathbf{S}$ and $\mathbf{S}^*\mathbf{B}\mathbf{S}$ are $\succ\mathbf{O}$ by Proposition 13.1.2 (b), so (a) applies to them, and the claim follows.

(d). Let $\mathbf{X}$ be admissible for $(\mathbf{A}',\mathbf{B}')$. Then

$$\begin{pmatrix}\mathbf{A}&\mathbf{X}\\ \mathbf{X}&\mathbf{B}\end{pmatrix}=\begin{pmatrix}\mathbf{A}'&\mathbf{X}\\ \mathbf{X}&\mathbf{B}'\end{pmatrix}+\begin{pmatrix}\mathbf{A}-\mathbf{A}'&\mathbf{O}\\ \mathbf{O}&\mathbf{B}-\mathbf{B}'\end{pmatrix},$$

a sum of two positive semidefinite matrices — the second because its quadratic form is $\mathbf{y}_1^*(\mathbf{A}-\mathbf{A}')\mathbf{y}_1+\mathbf{y}_2^*(\mathbf{B}-\mathbf{B}')\mathbf{y}_2\ge 0$. So $\mathbf{X}$ is admissible for $(\mathbf{A},\mathbf{B})$. Taking $\mathbf{X}=\mathbf{A}'\#\mathbf{B}'$, which is admissible for $(\mathbf{A}',\mathbf{B}')$ by (a), gives $\mathbf{A}'\#\mathbf{B}'\preceq\mathbf{A}\#\mathbf{B}$ by the maximality in (a).

(e). Since $\mathbf{C}^{1/2}$ is Hermitian, $(\mathbf{I}-\mathbf{C}^{1/2})^2\succeq\mathbf{O}$ by Theorem 13.1.4 (c), and $(\mathbf{C}^{1/2})^2=\mathbf{C}$ gives

$$\tfrac{1}{2}(\mathbf{I}+\mathbf{C})-\mathbf{C}^{1/2}=\tfrac{1}{2}(\mathbf{I}-\mathbf{C}^{1/2})^2\ \succeq\ \mathbf{O}.$$

Conjugating by $\mathbf{A}^{1/2}$ and using $\mathbf{A}^{1/2}\mathbf{C}\mathbf{A}^{1/2}=\mathbf{B}$,

$$\tfrac{1}{2}(\mathbf{A}+\mathbf{B})-\mathbf{A}\#\mathbf{B}=\tfrac{1}{2}\mathbf{A}^{1/2}(\mathbf{I}-\mathbf{C}^{1/2})^2\mathbf{A}^{1/2}\ \succeq\ \mathbf{O},$$

by Proposition 13.1.2 (a). For the inversion formula, $\mathbf{A}\#\mathbf{B}=\mathbf{A}^{1/2}\mathbf{C}^{1/2}\mathbf{A}^{1/2}$ is a product of invertible matrices, with inverse $\mathbf{A}^{-1/2}\mathbf{C}^{-1/2}\mathbf{A}^{-1/2}$, where $\mathbf{C}^{-1/2}=(\mathbf{C}^{1/2})^{-1}$. On the other side, using $(\mathbf{A}^{-1})^{1/2}=\mathbf{A}^{-1/2}$ and $(\mathbf{A}^{-1})^{-1/2}=\mathbf{A}^{1/2}$,

$$\mathbf{A}^{-1}\#\mathbf{B}^{-1}=\mathbf{A}^{-1/2}(\mathbf{A}^{1/2}\mathbf{B}^{-1}\mathbf{A}^{1/2})^{1/2}\mathbf{A}^{-1/2},$$

and $\mathbf{A}^{1/2}\mathbf{B}^{-1}\mathbf{A}^{1/2}=(\mathbf{A}^{-1/2}\mathbf{B}\mathbf{A}^{-1/2})^{-1}=\mathbf{C}^{-1}$, whose positive square root is $\mathbf{C}^{-1/2}$, since $(\mathbf{C}^{-1/2})^2=\mathbf{C}^{-1}$ and $\mathbf{C}^{-1/2}\succ\mathbf{O}$. So $\mathbf{A}^{-1}\#\mathbf{B}^{-1}=\mathbf{A}^{-1/2}\mathbf{C}^{-1/2}\mathbf{A}^{-1/2}=(\mathbf{A}\#\mathbf{B})^{-1}$.

■

Part (a) is the statement to remember, because it is the one with no square roots in it: $\mathbf{A}\#\mathbf{B}$ is the best Hermitian matrix that can sit in the off-diagonal corner above $\mathbf{A}$ and $\mathbf{B}$. Read beside Lemma 21.11.1, which says that such a corner cannot be too large in singular value, part (a) says exactly how large it can be in the Loewner order.

Quick check. Take $\mathbf{B}=\mathbf{A}$. Part (a) then says that $\mathbf{A}$ is the largest Hermitian $\mathbf{X}$ with $\left(\begin{smallmatrix}\mathbf{A}&\mathbf{X}\\ \mathbf{X}&\mathbf{A}\end{smallmatrix}\right)\succeq\mathbf{O}$. Check by hand that $\mathbf{X}=\mathbf{A}$ is admissible and that $\mathbf{X}=2\mathbf{A}$ is not.

Solution. For $\mathbf{X}=\mathbf{A}$ the quadratic form at $(\mathbf{y}_1,\mathbf{y}_2)$ is

$$\mathbf{y}_1^*\mathbf{A}\mathbf{y}_1+\mathbf{y}_1^*\mathbf{A}\mathbf{y}_2+\mathbf{y}_2^*\mathbf{A}\mathbf{y}_1+\mathbf{y}_2^*\mathbf{A}\mathbf{y}_2=(\mathbf{y}_1+\mathbf{y}_2)^*\mathbf{A}(\mathbf{y}_1+\mathbf{y}_2)\ \ge\ 0,$$

so $\mathbf{X}=\mathbf{A}$ is admissible — as it must be, since $\mathbf{A}\#\mathbf{A}=\mathbf{A}$ by the second example after Definition 21.11.1. For $\mathbf{X}=2\mathbf{A}$ take $\mathbf{y}_2=-\mathbf{y}_1$: the four terms give $\mathbf{y}_1^*\mathbf{A}\mathbf{y}_1-4\mathbf{y}_1^*\mathbf{A}\mathbf{y}_1+\mathbf{y}_1^*\mathbf{A}\mathbf{y}_1=-2\mathbf{y}_1^*\mathbf{A}\mathbf{y}_1$, which is <0 for $\mathbf{y}_1\neq\mathbf{0}$ because $\mathbf{A}\succ\mathbf{O}$. So $2\mathbf{A}$ is not admissible.

21.11.6 The chain of three means

Beside the arithmetic mean $\frac{1}{2}(\mathbf{A}+\mathbf{B})$ and the geometric mean $\mathbf{A}\#\mathbf{B}$ sits a third. The **harmonic mean** of two positive definite matrices is the inverse of the arithmetic mean of their inverses,

$$\Big(\frac{\mathbf{A}^{-1}+\mathbf{B}^{-1}}{2}\Big)^{-1},$$

which at $n=1$ is the familiar $2ab/(a+b)$. For two positive numbers the three are ordered harmonic $\le$ geometric $\le$ arithmetic, and the same order holds for matrices.

Corollary 21.11.5. Harmonic, Geometric, and Arithmetic

Let $\mathbf{A},\mathbf{B}\in M_n(F)$ be positive definite. Then

$$\Big(\frac{\mathbf{A}^{-1}+\mathbf{B}^{-1}}{2}\Big)^{-1} \preceq \mathbf{A}\#\mathbf{B} \preceq \frac{\mathbf{A}+\mathbf{B}}{2}.$$

Proof. The right inequality is Theorem 21.11.4 (e). For the left, apply that same part to the positive definite matrices $\mathbf{A}^{-1}$ and $\mathbf{B}^{-1}$:

$$\mathbf{A}^{-1}\#\mathbf{B}^{-1} \preceq \tfrac{1}{2}(\mathbf{A}^{-1}+\mathbf{B}^{-1}).$$

Both sides are positive definite — the left by the remark after Definition 21.11.1, the right as a positive combination of positive definite matrices. Since $t\mapsto -1/t$ is operator monotone on $(0,\infty)$ (Proposition 17.11.3), applying it reverses the inequality between the inverses:

$$\Big(\tfrac{1}{2}(\mathbf{A}^{-1}+\mathbf{B}^{-1})\Big)^{-1} \preceq (\mathbf{A}^{-1}\#\mathbf{B}^{-1})^{-1}.$$

By Theorem 21.11.4 (e) the right side is $\mathbf{A}\#\mathbf{B}$.

■

The three means agree when $\mathbf{A}=\mathbf{B}$, and for the pair of the worked example they are

$$\tfrac{1}{5}\begin{pmatrix}28&26\\26&37\end{pmatrix} \preceq \begin{pmatrix}6&6\\6&9\end{pmatrix} \preceq \begin{pmatrix}\frac{13}{2}&7\\7&11\end{pmatrix},$$

the two differences being $\frac{2}{5}\left(\begin{smallmatrix}1&2\\2&4\end{smallmatrix}\right)$ and $\frac{1}{2}\left(\begin{smallmatrix}1&2\\2&4\end{smallmatrix}\right)$, each a nonnegative multiple of a rank-one positive semidefinite matrix.

Remark. The geometric mean is the first genuinely noncommutative *average* in the book, and it is built out of two order facts proved long before this chapter: the operator monotonicity of $t\mapsto t^{1/2}$, which is Chapter 13 §05's Proposition 13.5.4, used in the claim inside Theorem 21.11.4, and the operator monotonicity of $t\mapsto -1/t$, which is Chapter 17 §11's Proposition 17.11.3, used in Corollary 21.11.5. Without the first, the extremal characterization would fail; without the second, the chain would have no left-hand end.

21.11.7 Exercises

21.11.7.1 A. Check your understanding

Exercise. A1

(a) Define $\mathbf{A}\#\mathbf{B}$ for positive definite $\mathbf{A}, \mathbf{B}$, and say why each square root in the formula exists.

(b) State the extremal characterization of $\mathbf{A}\#\mathbf{B}$, and explain in one sentence how it makes the symmetry $\mathbf{A}\#\mathbf{B} = \mathbf{B}\#\mathbf{A}$ obvious.

(c) Decide whether the following is correct, and justify your answer: "$\mathbf{A}^{1/2}\mathbf{B}^{1/2}$ is positive definite whenever $\mathbf{A}$ and $\mathbf{B}$ are".

(d) State the matrix Cauchy–Schwarz inequality and the arithmetic–geometric mean inequality for singular values, Theorem 21.11.2 — not the one for the geometric mean $\mathbf{A}\#\mathbf{B}$ — and say which one implies the other and how.

(e) Which of $\mathbf{A}^*\mathbf{A} + \mathbf{B}^*\mathbf{B}$ and $\mathbf{A}\mathbf{A}^* + \mathbf{B}\mathbf{B}^*$ bounds $2\,|||\mathbf{A}^*\mathbf{B}|||$, and what goes wrong with the other?

Solution.

(a) $\mathbf{A}\#\mathbf{B} = \mathbf{A}^{1/2}(\mathbf{A}^{-1/2}\mathbf{B}\mathbf{A}^{-1/2})^{1/2}\mathbf{A}^{1/2}$ (Definition 21.11.1). The outer square root exists because $\mathbf{A} \succ \mathbf{O}$ (Theorem 13.2.1), and $\mathbf{A}^{1/2} \succ \mathbf{O}$ is invertible; the inner one because $\mathbf{A}^{-1/2}\mathbf{B}\mathbf{A}^{-1/2}$ is a congruence of $\mathbf{B} \succ \mathbf{O}$ by an invertible matrix, hence $\succ \mathbf{O}$ (Proposition 13.1.2 (b)).

(b) It is the largest Hermitian $\mathbf{X}$ with $\left(\begin{smallmatrix}\mathbf{A} & \mathbf{X}\\ \mathbf{X} & \mathbf{B}\end{smallmatrix}\right) \succeq \mathbf{O}$ (Theorem 21.11.4 (a)). Swapping the two diagonal blocks is a congruence by the unitary $\left(\begin{smallmatrix}\mathbf{O} & \mathbf{I}\\ \mathbf{I} & \mathbf{O}\end{smallmatrix}\right)$, which changes neither the condition nor $\mathbf{X}$, so $(\mathbf{A}, \mathbf{B})$ and $(\mathbf{B}, \mathbf{A})$ have the same admissible set and therefore the same largest element.

(c) Incorrect. It need not even be Hermitian: with $\mathbf{A} = \operatorname{diag}(4, 1)$ and $\mathbf{B} = \left(\begin{smallmatrix}5 & 4\\ 4 & 5\end{smallmatrix}\right)$, $\mathbf{A}^{1/2}\mathbf{B}^{1/2} = \left(\begin{smallmatrix}4 & 2\\ 1 & 2\end{smallmatrix}\right)$. It is Hermitian exactly when $\mathbf{A}\mathbf{B} = \mathbf{B}\mathbf{A}$.

(d) Cauchy–Schwarz: $|||\mathbf{A}^*\mathbf{B}|||^2 \le |||\mathbf{A}^*\mathbf{A}|||\,|||\mathbf{B}^*\mathbf{B}|||$ (Theorem 21.11.3). Arithmetic–geometric mean for singular values: $2\,|||\mathbf{A}^*\mathbf{B}||| \le |||\mathbf{A}\mathbf{A}^* + \mathbf{B}\mathbf{B}^*|||$ (Theorem 21.11.2). The second implies the first: replace $\mathbf{A}$ by $t\mathbf{A}$ and $\mathbf{B}$ by $t^{-1}\mathbf{B}$, which leaves $\mathbf{A}^*\mathbf{B}$ alone, use the triangle inequality, and pick t to balance the two terms.

(e) $\mathbf{A}\mathbf{A}^* + \mathbf{B}\mathbf{B}^*$ does (Theorem 21.11.2). With $\mathbf{A}^*\mathbf{A} + \mathbf{B}^*\mathbf{B}$ the statement is false: the warning after that theorem gives $\mathbf{A}, \mathbf{B}$ with $\mathbf{A}^*\mathbf{A} + \mathbf{B}^*\mathbf{B} = \mathbf{I}$ of spectral norm 1 and $2\|\mathbf{A}^*\mathbf{B}\|_2 = 2$.

21.11.7.2 B. Practice

Exercise. B1: A geometric mean, exactly

Let

$$\mathbf{A} = \begin{pmatrix} 5 & 4 \\ 4 & 5 \end{pmatrix}, \qquad \mathbf{B} = \begin{pmatrix} 13 & 20 \\ 20 & 37 \end{pmatrix}.$$

(a) Verify that $\mathbf{A}^{1/2} = \left(\begin{smallmatrix}2 & 1\\ 1 & 2\end{smallmatrix}\right)$ and compute $\mathbf{C} = \mathbf{A}^{-1/2}\mathbf{B}\mathbf{A}^{-1/2}$ and $\mathbf{A}\#\mathbf{B}$.

(b) Check the arithmetic–geometric mean inequality $\mathbf{A}\#\mathbf{B} \preceq \frac{1}{2}(\mathbf{A} + \mathbf{B})$ directly, by com-

puting the difference and its eigenvalues.

(c) Verify the extremal characterization for your answer by computing the Schur complement $\mathbf{B} - \mathbf{X}\mathbf{A}^{-1}\mathbf{X}$ with $\mathbf{X} = \mathbf{A}\#\mathbf{B}$.

Solution.

(a) $\left(\begin{smallmatrix} 2 & 1 \\ 1 & 2 \end{smallmatrix}\right)^2 = \left(\begin{smallmatrix} 5 & 4 \\ 4 & 5 \end{smallmatrix}\right) = \mathbf{A}$, and $\left(\begin{smallmatrix} 2 & 1 \\ 1 & 2 \end{smallmatrix}\right)$ has eigenvalues 3 and 1, both positive, so it is $\succ \mathbf{O}$ and is the positive square root (Theorem 13.2.1). Its determinant is 3, so $\mathbf{A}^{-1/2} = \frac{1}{3}\left(\begin{smallmatrix} 2 & -1 \\ -1 & 2 \end{smallmatrix}\right)$. Then

$$\mathbf{A}^{-1/2}\mathbf{B} = \tfrac{1}{3}\begin{pmatrix} 6 & 3 \\ 27 & 54 \end{pmatrix} = \begin{pmatrix} 2 & 1 \\ 9 & 18 \end{pmatrix},$$

and multiplying on the right by $\mathbf{A}^{-1/2}$ gives $\mathbf{C} = \frac{1}{3}\left(\begin{smallmatrix} 3 & 0 \\ 0 & 27 \end{smallmatrix}\right) = \operatorname{diag}(1, 9)$. Hence $\mathbf{C}^{1/2} = \operatorname{diag}(1, 3)$ and

$$\mathbf{A}\#\mathbf{B} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix}\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} = \begin{pmatrix} 7 & 8 \\ 8 & 13 \end{pmatrix}.$$

(b) $\frac{1}{2}(\mathbf{A} + \mathbf{B}) = \left(\begin{smallmatrix} 9 & 12 \\ 12 & 21 \end{smallmatrix}\right)$, so the difference is

$$\tfrac{1}{2}(\mathbf{A} + \mathbf{B}) - \mathbf{A}\#\mathbf{B} = \begin{pmatrix} 2 & 4 \\ 4 & 8 \end{pmatrix} = 2\begin{pmatrix} 1 \\ 2 \end{pmatrix}\begin{pmatrix} 1 & 2 \end{pmatrix},$$

of trace 10 and determinant $16 - 16 = 0$, so its eigenvalues are 10 and 0 and it is $\succeq \mathbf{O}$ (Theorem 13.1.4 (b)). The inequality holds, with the difference singular.

(c) $\det \mathbf{A} = 9$, so $\mathbf{A}^{-1} = \frac{1}{9}\left(\begin{smallmatrix} 5 & -4 \\ -4 & 5 \end{smallmatrix}\right)$. With $\mathbf{X} = \left(\begin{smallmatrix} 7 & 8 \\ 8 & 13 \end{smallmatrix}\right)$,

$$\mathbf{X}\mathbf{A}^{-1} = \tfrac{1}{9}\begin{pmatrix} 3 & 12 \\ -12 & 33 \end{pmatrix} = \tfrac{1}{3}\begin{pmatrix} 1 & 4 \\ -4 & 11 \end{pmatrix},$$

and $\mathbf{X}\mathbf{A}^{-1}\mathbf{X} = \frac{1}{3}\left(\begin{smallmatrix} 39 & 60 \\ 60 & 111 \end{smallmatrix}\right) = \left(\begin{smallmatrix} 13 & 20 \\ 20 & 37 \end{smallmatrix}\right) = \mathbf{B}$. So the Schur complement is $\mathbf{O}$, which is $\succeq \mathbf{O}$: $\mathbf{X}$ is admissible, and the block matrix is singular, as it must be at the largest admissible $\mathbf{X}$.

Exercise. B2: The two inequalities on one pair

Let $\mathbf{A} = \left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$ and $\mathbf{B} = \left(\begin{smallmatrix} 1 & 0 \\ 1 & 1 \end{smallmatrix}\right)$ in $M_2(\mathbb{R})$.

(a) Compute $\mathbf{A}^*\mathbf{B}$ and $\mathbf{A}\mathbf{A}^* + \mathbf{B}\mathbf{B}^*$, and verify $2\sigma_j(\mathbf{A}^*\mathbf{B}) \le \lambda_j(\mathbf{A}\mathbf{A}^* + \mathbf{B}\mathbf{B}^*)$ for $j = 1, 2$ using traces and determinants.

(b) Verify Theorem 21.11.3 for this pair in the Frobenius norm.

Solution.

(a) $\mathbf{A}^* = \left(\begin{smallmatrix} 1 & 0 \\ 1 & 1 \end{smallmatrix}\right)$, so $\mathbf{A}^*\mathbf{B} = \left(\begin{smallmatrix} 1 & 0 \\ 2 & 1 \end{smallmatrix}\right)$, and $\mathbf{A}\mathbf{A}^* = \left(\begin{smallmatrix} 2 & 1 \\ 1 & 1 \end{smallmatrix}\right)$, $\mathbf{B}\mathbf{B}^* = \left(\begin{smallmatrix} 1 & 1 \\ 1 & 2 \end{smallmatrix}\right)$, so $\mathbf{A}\mathbf{A}^* + \mathbf{B}\mathbf{B}^* = \left(\begin{smallmatrix} 3 & 2 \\ 2 & 3 \end{smallmatrix}\right)$, with eigenvalues 5 and 1.
The singular values of $\mathbf{M} = \mathbf{A}^*\mathbf{B}$ come from $\mathbf{M}^*\mathbf{M} = \left(\begin{smallmatrix} 5 & 2 \\ 2 & 1 \end{smallmatrix}\right)$, of trace 6 and determinant 1, so $\sigma_1^2, \sigma_2^2 = 3 \pm 2\sqrt{2} = (\sqrt{2} \pm 1)^2$ and $\sigma_1 = \sqrt{2} + 1$, $\sigma_2 = \sqrt{2} - 1$. Then $2\sigma_1 = 2\sqrt{2} + 2 \approx 4.83 \le 5$ and $2\sigma_2 = 2\sqrt{2} - 2 \approx 0.83 \le 1$. Both hold; the first is $2\sqrt{2} \le 3$, that is $8 \le 9$, and the second is $2\sqrt{2} \le 3$ again.

(b) $\mathbf{A}^*\mathbf{A} = \left(\begin{smallmatrix} 1 & 1 \\ 1 & 2 \end{smallmatrix}\right)$ and $\mathbf{B}^*\mathbf{B} = \left(\begin{smallmatrix} 2 & 1 \\ 1 & 1 \end{smallmatrix}\right)$, each of Frobenius norm $\sqrt{1+1+1+4} = \sqrt{7}$. And $\|\mathbf{A}^*\mathbf{B}\|_F^2 = 1+4+1 = 6$. So the inequality reads $6 \le \sqrt{7}\cdot\sqrt{7} = 7$, which holds.

Exercise. B3: Means of commuting matrices

Let $\mathbf{A} = \operatorname{diag}(a_1,\dots,a_n)$ and $\mathbf{B} = \operatorname{diag}(b_1,\dots,b_n)$ with all $a_i, b_i > 0$. Compute the three means of Corollary 21.11.5 and deduce the classical chain of inequalities between the harmonic, geometric, and arithmetic means of two positive numbers.

Solution. The matrices commute, so $\mathbf{A}\#\mathbf{B} = (\mathbf{AB})^{1/2} = \operatorname{diag}(\sqrt{a_i b_i})$ by the third example after Definition 21.11.1. Also $\frac{1}{2}(\mathbf{A}+\mathbf{B}) = \operatorname{diag}\left(\frac{a_i+b_i}{2}\right)$ and

$$\left(\frac{\mathbf{A}^{-1}+\mathbf{B}^{-1}}{2}\right)^{-1} = \operatorname{diag}\left(\frac{2a_ib_i}{a_i+b_i}\right).$$

A diagonal matrix is $\succeq \mathbf{O}$ exactly when its entries are ≥ 0, so Corollary 21.11.5 says precisely

$$\frac{2a_ib_i}{a_i+b_i} \;\le\; \sqrt{a_ib_i} \;\le\; \frac{a_i+b_i}{2}$$

for every i, the classical chain for the two positive numbers a_i, b_i. The right inequality is Lemma 17.8.6 with two terms, and the left is the right one applied to $1/a_i$ and $1/b_i$ and inverted.

21.11.7.3 C. Going deeper

Exercise. C1: The determinant of a geometric mean

Let $\mathbf{A}, \mathbf{B} \in M_n(F)$ be positive definite.

(a) Prove that $\det(\mathbf{A}\#\mathbf{B}) = \sqrt{\det\mathbf{A}\,\det\mathbf{B}}$.

(b) Deduce from Corollary 21.11.5 that $\sqrt{\det\mathbf{A}\,\det\mathbf{B}} \le \det\left(\frac{1}{2}(\mathbf{A}+\mathbf{B})\right)$.

Hint for (b): the Loewner order compares determinants of positive definite matrices.

Solution.

(a) Write $\mathbf{C} = \mathbf{A}^{-1/2}\mathbf{B}\mathbf{A}^{-1/2}$. Taking determinants in Definition 21.11.1 and using multiplicativity (Theorem 7.5.4),

$$\det(\mathbf{A}\#\mathbf{B}) = \det(\mathbf{A}^{1/2})^2\det(\mathbf{C}^{1/2}) = \det\mathbf{A}\cdot\det(\mathbf{C}^{1/2}).$$

Now $(\mathbf{C}^{1/2})^2 = \mathbf{C}$ gives $\det(\mathbf{C}^{1/2})^2 = \det\mathbf{C}$, and $\det(\mathbf{C}^{1/2}) > 0$ because $\mathbf{C}^{1/2} \succ \mathbf{O}$ has positive eigenvalues, so $\det(\mathbf{C}^{1/2}) = \sqrt{\det\mathbf{C}}$. Finally $\det\mathbf{C} = \det(\mathbf{A}^{-1/2})^2\det\mathbf{B} = \det\mathbf{B}/\det\mathbf{A}$. Hence

$$\det(\mathbf{A}\#\mathbf{B}) = \det\mathbf{A}\cdot\sqrt{\det\mathbf{B}/\det\mathbf{A}} = \sqrt{\det\mathbf{A}\,\det\mathbf{B}}.$$

(b) By Corollary 21.11.5, $\mathbf{A}\#\mathbf{B} \preceq \frac{1}{2}(\mathbf{A}+\mathbf{B})$, and both are positive definite. For positive

definite $\mathbf{X} \preceq \mathbf{Y}$ we have $\det \mathbf{X} \leq \det \mathbf{Y}$: by Corollary 17.2.4, $\lambda_i(\mathbf{Y}) \geq \lambda_i(\mathbf{X}) > 0$ for every i, and the determinant of a Hermitian matrix is the product of its eigenvalues with multiplicity (Theorem 9.3.3, whose hypothesis holds because the characteristic polynomial of a Hermitian matrix splits over $\mathbb{R}$ by Theorem 12.2.1), a product of positive numbers increasing in each factor. Applying this and part (a) gives the inequality. At $n = 1$ it is again the classical arithmetic–geometric mean inequality.

Exercise. C2: The geometric mean solves an equation

Let $\mathbf{A}, \mathbf{B} \in M_n(F)$ be positive definite.

(a) Prove that $\mathbf{X} = \mathbf{A}\#\mathbf{B}$ is the **only** positive definite solution of $\mathbf{X}\mathbf{A}^{-1}\mathbf{X} = \mathbf{B}$.

(b) Use (a) to give a second proof that $\mathbf{A}\#\mathbf{B} = \mathbf{B}\#\mathbf{A}$.

Solution.

(a) Write $\mathbf{C} = \mathbf{A}^{-1/2}\mathbf{B}\mathbf{A}^{-1/2}$. A positive definite $\mathbf{X}$ may be written uniquely as $\mathbf{X} = \mathbf{A}^{1/2}\mathbf{Y}\mathbf{A}^{1/2}$ with $\mathbf{Y} = \mathbf{A}^{-1/2}\mathbf{X}\mathbf{A}^{-1/2}$, and $\mathbf{Y} \succ \mathbf{O}$ by Proposition 13.1.2 (b). As computed in the proof of Theorem 21.11.4 (a), $\mathbf{X}\mathbf{A}^{-1}\mathbf{X} = \mathbf{A}^{1/2}\mathbf{Y}^2\mathbf{A}^{1/2}$. So $\mathbf{X}\mathbf{A}^{-1}\mathbf{X} = \mathbf{B}$ is equivalent to $\mathbf{A}^{1/2}\mathbf{Y}^2\mathbf{A}^{1/2} = \mathbf{A}^{1/2}\mathbf{C}\mathbf{A}^{1/2}$, that is to $\mathbf{Y}^2 = \mathbf{C}$, since $\mathbf{A}^{1/2}$ is invertible. By the uniqueness of the positive square root (Theorem 13.2.1), the only positive definite solution is $\mathbf{Y} = \mathbf{C}^{1/2}$, that is $\mathbf{X} = \mathbf{A}\#\mathbf{B}$.

(b) Let $\mathbf{X} = \mathbf{A}\#\mathbf{B}$, so $\mathbf{X}\mathbf{A}^{-1}\mathbf{X} = \mathbf{B}$ by (a) and $\mathbf{X} \succ \mathbf{O}$ is invertible. Multiplying that identity on the left and on the right by $\mathbf{X}^{-1}$ gives $\mathbf{A}^{-1} = \mathbf{X}^{-1}\mathbf{B}\mathbf{X}^{-1}$, and inverting both sides gives $\mathbf{A} = \mathbf{X}\mathbf{B}^{-1}\mathbf{X}$. So $\mathbf{X}$ is a positive definite solution of $\mathbf{X}\mathbf{B}^{-1}\mathbf{X} = \mathbf{A}$, which by (a), with the roles of $\mathbf{A}$ and $\mathbf{B}$ exchanged, has $\mathbf{B}\#\mathbf{A}$ as its only positive definite solution. Hence $\mathbf{A}\#\mathbf{B} = \mathbf{B}\#\mathbf{A}$.

Exercise. C3: Cauchy–Schwarz for the Ky Fan norms, directly

Let $\mathbf{A}, \mathbf{B} \in M_n(F)$ and $1 \leq k \leq n$.

(a) Let $\mathbf{A}^*\mathbf{B} = \mathbf{W}\mathbf{\Sigma}\mathbf{Z}^*$ be a singular value decomposition (Theorem 13.9.1), and let $\mathbf{U}_k$ and $\mathbf{V}_k$ be the matrices of the first k columns of $\mathbf{W}$ and of $\mathbf{Z}$. Show that $\operatorname{tr}(\mathbf{U}_k^*\mathbf{A}^*\mathbf{B}\mathbf{V}_k) = \sum_{j \leq k} \sigma_j(\mathbf{A}^*\mathbf{B})$.

(b) Deduce, using the Cauchy–Schwarz inequality for the Frobenius inner product (Theorem 11.1.5) and the Ky Fan maximum principle (Theorem 17.6.2), that

$$\sum_{j \leq k} \sigma_j(\mathbf{A}^*\mathbf{B}) \leq \Big(\sum_{j \leq k} \lambda_j(\mathbf{A}^*\mathbf{A})\Big)^{1/2} \Big(\sum_{j \leq k} \lambda_j(\mathbf{B}^*\mathbf{B})\Big)^{1/2}.$$

Solution.

(a) The columns of $\mathbf{W}$ and $\mathbf{Z}$ are orthonormal, so $\mathbf{U}_k^*\mathbf{W} = (\mathbf{I}_k \;\; \mathbf{O})$ and $\mathbf{Z}^*\mathbf{V}_k = \begin{pmatrix} \mathbf{I}_k \\ \mathbf{O} \end{pmatrix}$. Hence

$$\mathbf{U}_k^*\mathbf{A}^*\mathbf{B}\mathbf{V}_k = \mathbf{U}_k^*\mathbf{W}\mathbf{\Sigma}\mathbf{Z}^*\mathbf{V}_k = \operatorname{diag}(\sigma_1, \ldots, \sigma_k),$$

the top-left $k \times k$ corner of $\mathbf{\Sigma}$, whose trace is $\sum_{j \le k} \sigma_j(\mathbf{A}^*\mathbf{B})$.

(b) Writing the trace as the Frobenius inner product $\langle \mathbf{X}, \mathbf{Y} \rangle = \operatorname{tr}(\mathbf{Y}^*\mathbf{X})$ of Chapter 11 §01,

$$\operatorname{tr}(\mathbf{U}_k^*\mathbf{A}^*\mathbf{B}\mathbf{V}_k) = \operatorname{tr}((\mathbf{A}\mathbf{U}_k)^*(\mathbf{B}\mathbf{V}_k)) = \langle \mathbf{B}\mathbf{V}_k, \mathbf{A}\mathbf{U}_k \rangle,$$

so by Theorem 11.1.5 it is at most $\|\mathbf{A}\mathbf{U}_k\|_F \|\mathbf{B}\mathbf{V}_k\|_F$. Now $\mathbf{U}_k^*\mathbf{U}_k = \mathbf{I}_k$, so $\mathbf{U}_k$ is an isometry and

$$\|\mathbf{A}\mathbf{U}_k\|_F^2 = \operatorname{tr}(\mathbf{U}_k^*\mathbf{A}^*\mathbf{A}\mathbf{U}_k) \;\le\; \sum_{j \le k} \lambda_j(\mathbf{A}^*\mathbf{A})$$

by Theorem 17.6.2 applied to the Hermitian matrix $\mathbf{A}^*\mathbf{A}$. The same bound holds for $\|\mathbf{B}\mathbf{V}_k\|_F^2$ with $\mathbf{B}^*\mathbf{B}$. Combining with (a) gives the display. (Taking $k = n$ recovers Theorem 21.11.3 for the trace norm.)

21.12 Golden–Thompson

For numbers, $e^{a+b} = e^a e^b$. For matrices this fails the moment $\mathbf{A}$ and $\mathbf{B}$ stop commuting, and Chapter 10 §09 was careful to say so: Theorem 10.9.4 (d) buys the identity only at the price of $\mathbf{AB} = \mathbf{BA}$. The question this section answers is what survives without that price. The answer is that the **trace** still knows the right inequality, in one direction, and that the direction is the useful one.

The section also closes the chapter, and Part V with it. After Golden–Thompson come two of its relatives — Klein's inequality and the nonnegativity of the relative entropy — a counterexample showing that three matrices are one too many, and one theorem that is quoted rather than proved.

21.12.1 Sums in the exponent, products in the answer

A product $e^{\mathbf{A}}e^{\mathbf{B}}$ is not $e^{\mathbf{A}+\mathbf{B}}$, but it is not far off either: to first order in $\mathbf{A}$ and $\mathbf{B}$ the two agree, since both are $\mathbf{I} + \mathbf{A} + \mathbf{B} + \dots$, and the discrepancy is quadratic. So if the exponents are made small by a factor $1/k$ and the resulting error is repaired k times over, the discrepancy should vanish. It does, and this is the one tool that turns a sum in the exponent into a product.

Lemma 21.12.1. Lie's Product Formula

Let $\mathbf{A}, \mathbf{B} \in M_n(\mathbb{C})$. Then

$$e^{\mathbf{A}+\mathbf{B}} = \lim_{k \to \infty} \left(e^{\mathbf{A}/k} e^{\mathbf{B}/k}\right)^k,$$

the limit taken entrywise, and also in any norm on $M_n(\mathbb{C})$.

Idea. Write $\mathbf{S}_k = e^{(\mathbf{A}+\mathbf{B})/k}$ and $\mathbf{T}_k = e^{\mathbf{A}/k}e^{\mathbf{B}/k}$. Then $\mathbf{S}_k^k$ is exactly $e^{\mathbf{A}+\mathbf{B}}$, so the whole task is to compare $\mathbf{S}_k^k$ with $\mathbf{T}_k^k$. Two estimates do it. Both $\mathbf{S}_k$ and $\mathbf{T}_k$ have norm at most $e^{(\|\mathbf{A}\|+\|\mathbf{B}\|)/k}$, so no power of either can be large; and $\mathbf{S}_k - \mathbf{T}_k$ is of size $1/k^2$, because the two agree through the linear term. A telescoping identity turns a difference of k-th powers into k copies of $\mathbf{S}_k - \mathbf{T}_k$, and $k \cdot 1/k^2$ tends to 0.

Proof. Fix the spectral norm $\|\cdot\|_2$, which is submultiplicative by Theorem 16.3.2 (d). Put $a = \|\mathbf{A}\|_2$, $b = \|\mathbf{B}\|_2$, and $c = a + b$, and let k be an integer with $k \geq \max(1, c)$. Write

$$\mathbf{S}_k = e^{(\mathbf{A}+\mathbf{B})/k}, \qquad \mathbf{T}_k = e^{\mathbf{A}/k} e^{\mathbf{B}/k}.$$

Step 1. $\mathbf{S}_k^k = e^{\mathbf{A}+\mathbf{B}}$. The matrix $(\mathbf{A}+\mathbf{B})/k$ commutes with itself, so Theorem 10.9.4 (d) gives $e^{j(\mathbf{A}+\mathbf{B})/k} e^{(\mathbf{A}+\mathbf{B})/k} = e^{(j+1)(\mathbf{A}+\mathbf{B})/k}$ for every $j \geq 0$. Induction on j yields $\mathbf{S}_k^j = e^{j(\mathbf{A}+\mathbf{B})/k}$, and $j = k$ is the claim.

Step 2. Norm bounds. By Theorem 16.6.1 (a), $\|\mathbf{S}_k\|_2 \leq e^{\|\mathbf{A}+\mathbf{B}\|_2/k} \leq e^{c/k}$, using $\|\mathbf{A}+\mathbf{B}\|_2 \leq c$; and $\|\mathbf{T}_k\|_2 \leq e^{a/k} e^{b/k} = e^{c/k}$ by submultiplicativity. Since $k \geq c$, both are at most e.

Step 3. $\mathbf{S}_k - \mathbf{T}_k$ is of size $1/k^2$. For any $\mathbf{X} \in M_n(\mathbb{C})$ put $\mathbf{R}(\mathbf{X}) = e^{\mathbf{X}} - \mathbf{I} - \mathbf{X}$, so that Theorem 16.6.1 (c) gives

$$\|\mathbf{R}(\mathbf{X})\|_2 \leq \tfrac{1}{2} \|\mathbf{X}\|_2^2 \, e^{\|\mathbf{X}\|_2}.$$

Multiplying out $\mathbf{T}_k = (\mathbf{I} + \frac{\mathbf{A}}{k} + \mathbf{R}(\frac{\mathbf{A}}{k}))(\mathbf{I} + \frac{\mathbf{B}}{k} + \mathbf{R}(\frac{\mathbf{B}}{k}))$, and using $e^{\mathbf{B}/k} = \mathbf{I} + \frac{\mathbf{B}}{k} + \mathbf{R}(\frac{\mathbf{B}}{k})$ to collect the three terms that begin with $\mathbf{R}(\frac{\mathbf{A}}{k})$,

$$\mathbf{T}_k = \mathbf{I} + \frac{\mathbf{A}+\mathbf{B}}{k} + \mathbf{R}\Big(\frac{\mathbf{B}}{k}\Big) + \frac{\mathbf{A}\mathbf{B}}{k^2} + \frac{\mathbf{A}}{k}\mathbf{R}\Big(\frac{\mathbf{B}}{k}\Big) + \mathbf{R}\Big(\frac{\mathbf{A}}{k}\Big) e^{\mathbf{B}/k}.$$

Subtracting this from $\mathbf{S}_k = \mathbf{I} + \frac{\mathbf{A}+\mathbf{B}}{k} + \mathbf{R}(\frac{\mathbf{A}+\mathbf{B}}{k})$, the constant and linear terms cancel, and the triangle inequality with $a/k, b/k, c/k \leq 1$ gives

$$\begin{aligned} \|\mathbf{S}_k - \mathbf{T}_k\|_2 &\leq \frac{c^2 e}{2k^2} + \frac{b^2 e}{2k^2} + \frac{ab}{k^2} + \frac{ab^2 e}{2k^3} + \frac{a^2 e^2}{2k^2} \\ &\leq \frac{M}{k^2}, \qquad M = \frac{(c^2+b^2)e}{2} + ab + \frac{ab^2 e}{2} + \frac{a^2 e^2}{2}, \end{aligned}$$

where the fourth term used $1/k^3 \leq 1/k^2$. The constant M depends only on $\mathbf{A}$ and $\mathbf{B}$.

Step 4. Telescoping. For any $\mathbf{M}, \mathbf{N} \in M_n(\mathbb{C})$ and any $k \geq 1$,

$$\mathbf{M}^k - \mathbf{N}^k = \sum_{j=0}^{k-1} \mathbf{M}^j (\mathbf{M} - \mathbf{N}) \mathbf{N}^{k-1-j},$$

since the j-th summand is $\mathbf{M}^{j+1}\mathbf{N}^{k-1-j} - \mathbf{M}^j \mathbf{N}^{k-j}$ and the sum telescopes. Applying this to $\mathbf{M} = \mathbf{S}_k$, $\mathbf{N} = \mathbf{T}_k$ and using Steps 1 and 2,

$$\|e^{\mathbf{A}+\mathbf{B}} - \mathbf{T}_k^k\|_2 \leq \sum_{j=0}^{k-1} e^{jc/k} \, \|\mathbf{S}_k - \mathbf{T}_k\|_2 \, e^{(k-1-j)c/k} \leq k \, e^{c} \, \frac{M}{k^2} = \frac{e^c M}{k}.$$

The right side tends to 0, so $\mathbf{T}_k^k \to e^{\mathbf{A}+\mathbf{B}}$ in the spectral norm, hence entrywise and in every norm by Corollary 16.2.4. This proves the formula.

■

Nothing in the proof used that the matrices are Hermitian: the formula holds for all complex matrices. What it costs is a limit, and every consequence below is obtained by pushing an inequality through that limit.

21.12.2 The trace of a product of two positive matrices

Golden–Thompson compares $\operatorname{tr} e^{\mathbf{A}+\mathbf{B}}$ with $\operatorname{tr}(e^{\mathbf{A}}e^{\mathbf{B}})$. Before comparing them we should know that the second number is real, which is not obvious: the product of two Hermitian matrices is almost never Hermitian. One identity settles this, and it will be used again.

Lemma 21.12.2. Powers of a Product of Two Positive Matrices

Let $\mathbf{X}, \mathbf{Y} \in M_n(\mathbb{C})$ be positive semidefinite and let $k \geq 1$. Then

$$\operatorname{tr}\big((\mathbf{XY})^k\big) = \operatorname{tr}\big((\mathbf{X}^{1/2}\mathbf{Y}\mathbf{X}^{1/2})^k\big),$$

and this number is real and ≥ 0.

Proof. Let $\mathbf{X}^{1/2}$ be the positive square root of Theorem 13.2.1, and put $\mathbf{S} = \mathbf{X}^{1/2}\mathbf{Y}\mathbf{X}^{1/2}$. We claim

$$(\mathbf{XY})^k = \mathbf{X}^{1/2}\,\mathbf{S}^{k-1}\,\mathbf{X}^{1/2}\mathbf{Y} \qquad (k \geq 1).$$

For $k = 1$ the right side is $\mathbf{X}^{1/2}\mathbf{X}^{1/2}\mathbf{Y} = \mathbf{XY}$. Assuming it for k, and using $\mathbf{X}^{1/2}\mathbf{Y}\mathbf{X}\mathbf{Y} = \mathbf{S}\,\mathbf{X}^{1/2}\mathbf{Y}$,

$$(\mathbf{XY})^{k+1} = \mathbf{X}^{1/2}\mathbf{S}^{k-1}(\mathbf{X}^{1/2}\mathbf{Y}\mathbf{X}\mathbf{Y}) = \mathbf{X}^{1/2}\mathbf{S}^{k}\mathbf{X}^{1/2}\mathbf{Y},$$

which is the claim for $k+1$. Taking traces and moving the leading $\mathbf{X}^{1/2}$ to the back, which Theorem 0.9.6 (3) permits,

$$\operatorname{tr}\big((\mathbf{XY})^k\big) = \operatorname{tr}(\mathbf{S}^{k-1}\mathbf{X}^{1/2}\mathbf{Y}\mathbf{X}^{1/2}) = \operatorname{tr}(\mathbf{S}^{k}).$$

Now $\mathbf{S} = (\mathbf{X}^{1/2})^{*}\mathbf{Y}\mathbf{X}^{1/2} \succeq \mathbf{O}$ by Proposition 13.1.2 (a), since $\mathbf{X}^{1/2}$ is Hermitian. So $\mathbf{S}^k$ is Hermitian with eigenvalues the k-th powers of those of $\mathbf{S}$, all ≥ 0 by Theorem 13.1.4 (b). Its trace is the sum of its eigenvalues (Theorem 9.3.3), hence real and ≥ 0.

■

Quick check. Why is $\operatorname{tr}(e^{\mathbf{A}}e^{\mathbf{B}})$ a real number for Hermitian $\mathbf{A}$ and $\mathbf{B}$, even though $e^{\mathbf{A}}e^{\mathbf{B}}$ is not Hermitian?

Solution. For Hermitian $\mathbf{A}$ with spectral resolution $\mathbf{A} = \sum_i \lambda_i \mathbf{P}_i$, Theorem 12.8.5 (a) gives $e^{\mathbf{A}} = \sum_i e^{\lambda_i}\mathbf{P}_i$, which is Hermitian with positive eigenvalues, hence positive definite by Theorem 13.1.4 (b). The same holds for $e^{\mathbf{B}}$. So Lemma 21.12.2 with $k = 1$ applies: $\operatorname{tr}(e^{\mathbf{A}}e^{\mathbf{B}})$ is real and ≥ 0. The product itself is not Hermitian unless $\mathbf{A}$ and $\mathbf{B}$ commute, but its trace does not notice.

The engine of Golden–Thompson is the next inequality. It says that in a trace, the powers may be moved off the product and onto the two factors — provided the exponent is a power of two. That restriction is not an accident of the proof technique here; it is what makes a repeated Cauchy–Schwarz argument close.

Lemma 21.12.3. The Trace Power Inequality

Let $r \geq 0$ and $m = 2^r$.

(a) $\operatorname{tr}\big((\mathbf{A}^*)^m \mathbf{A}^m\big) \leq \operatorname{tr}\big((\mathbf{A}\mathbf{A}^*)^m\big)$ for every $\mathbf{A} \in M_n(\mathbb{C})$.

(b) $\operatorname{tr}\big((\mathbf{X}\mathbf{Y})^m\big) \leq \operatorname{tr}\big(\mathbf{X}^m \mathbf{Y}^m\big)$ for all positive semidefinite $\mathbf{X}, \mathbf{Y} \in M_n(\mathbb{C})$.

Idea. Both halves are proved together by induction on r, because each step of one needs the other at the previous stage. The only inequality used is Cauchy–Schwarz for the Frobenius inner product, in the form $|\operatorname{tr}(\mathbf{M}^2)| \leq \operatorname{tr}(\mathbf{M}^*\mathbf{M})$: a trace of a square is bounded by a trace of a product of a matrix with its adjoint. Everything else is cyclic rearrangement of words in $\mathbf{X}$ and $\mathbf{Y}$, or in $\mathbf{A}$ and $\mathbf{A}^*$. The role of (a) is to absorb the mismatch that Cauchy–Schwarz creates: it turns $(\mathbf{A}^*)^m \mathbf{A}^m$ — the wrong grouping — into $(\mathbf{A}\mathbf{A}^*)^m$, which is a word one can feed back into (b).

Proof. Throughout, $\langle \mathbf{M}, \mathbf{N} \rangle = \operatorname{tr}(\mathbf{N}^*\mathbf{M})$ is the Frobenius inner product of Chapter 11 §01, with induced norm $\|\mathbf{M}\|_F$.

Claim 1. $|\operatorname{tr}(\mathbf{M}^2)| \leq \operatorname{tr}(\mathbf{M}^*\mathbf{M})$ *for every* $\mathbf{M} \in M_n(\mathbb{C})$. Since $(\mathbf{M}^*)^* = \mathbf{M}$, we have $\operatorname{tr}(\mathbf{M}^2) = \operatorname{tr}\big((\mathbf{M}^*)^*\mathbf{M}\big) = \langle \mathbf{M}, \mathbf{M}^* \rangle$. By Theorem 11.1.5, $|\langle \mathbf{M}, \mathbf{M}^* \rangle| \leq \|\mathbf{M}\|_F \|\mathbf{M}^*\|_F$. Finally $\|\mathbf{M}^*\|_F^2 = \operatorname{tr}(\mathbf{M}\mathbf{M}^*) = \operatorname{tr}(\mathbf{M}^*\mathbf{M}) = \|\mathbf{M}\|_F^2$ by Theorem 0.9.6 (3), so the bound is $\|\mathbf{M}\|_F^2 = \operatorname{tr}(\mathbf{M}^*\mathbf{M})$.

Claim 2. *For positive semidefinite* $\mathbf{X}, \mathbf{Y}$ *and* $q \geq 1$, $\operatorname{tr}\big((\mathbf{X}\mathbf{Y}^2\mathbf{X})^q\big) = \operatorname{tr}\big((\mathbf{X}^2\mathbf{Y}^2)^q\big)$*; and for any* $\mathbf{A}$ *and* $q \geq 1$, $\operatorname{tr}\big((\mathbf{A}^2(\mathbf{A}^*)^2)^q\big) = \operatorname{tr}\big((\mathbf{P}\mathbf{Q})^q\big)$*, where* $\mathbf{P} = \mathbf{A}\mathbf{A}^*$ *and* $\mathbf{Q} = \mathbf{A}^*\mathbf{A}$. Both are cyclic rearrangements, justified by Theorem 0.9.6 (3). Induction on q gives $(\mathbf{X}\mathbf{Y}^2\mathbf{X})^q = \mathbf{X}(\mathbf{Y}^2\mathbf{X}^2)^{q-1}\mathbf{Y}^2\mathbf{X}$; moving the leading $\mathbf{X}$ to the back turns its trace into $\operatorname{tr}\big((\mathbf{Y}^2\mathbf{X}^2)^q\big)$, which is $\operatorname{tr}\big((\mathbf{X}^2\mathbf{Y}^2)^q\big)$ by one more cyclic shift. For the second identity, $\mathbf{A}^2(\mathbf{A}^*)^2 = \mathbf{A}\mathbf{P}\mathbf{A}^*$, and induction on q gives $(\mathbf{A}\mathbf{P}\mathbf{A}^*)^q = \mathbf{A}\,\mathbf{P}(\mathbf{Q}\mathbf{P})^{q-1}\,\mathbf{A}^*$; moving the leading $\mathbf{A}$ to the back turns its trace into $\operatorname{tr}\big(\mathbf{P}(\mathbf{Q}\mathbf{P})^{q-1}\mathbf{Q}\big) = \operatorname{tr}\big((\mathbf{P}\mathbf{Q})^q\big)$.

We now induct on r, proving (a) and (b) simultaneously.

Base $r = 0$**.** Part (a) reads $\operatorname{tr}(\mathbf{A}^*\mathbf{A}) \leq \operatorname{tr}(\mathbf{A}\mathbf{A}^*)$, an equality by Theorem 0.9.6 (3). Part (b) reads $\operatorname{tr}(\mathbf{X}\mathbf{Y}) \leq \operatorname{tr}(\mathbf{X}\mathbf{Y})$.

Step. Assume (a) and (b) for r, and write $q = 2^r$.

Part (b) for $2q$. Put $\mathbf{A} = \mathbf{X}\mathbf{Y}$, so $\mathbf{A}^* = \mathbf{Y}\mathbf{X}$ because $\mathbf{X}$ and $\mathbf{Y}$ are Hermitian, and put $\mathbf{M} = \mathbf{A}^q = (\mathbf{X}\mathbf{Y})^q$. Then $\mathbf{M}^2 = (\mathbf{X}\mathbf{Y})^{2q}$ and $\mathbf{M}^*\mathbf{M} = (\mathbf{A}^*)^q\mathbf{A}^q$. The number $\operatorname{tr}(\mathbf{M}^2) = \operatorname{tr}\big((\mathbf{X}\mathbf{Y})^{2q}\big)$ is real and ≥ 0 by Lemma 21.12.2, so it equals its own absolute value, and Claim 1 gives

$$\operatorname{tr}\big((\mathbf{X}\mathbf{Y})^{2q}\big) \leq \big|\operatorname{tr}(\mathbf{M}^2)\big| \leq \operatorname{tr}\big((\mathbf{A}^*)^q\mathbf{A}^q\big).$$

By the inductive hypothesis (a),

$$\operatorname{tr}\big((\mathbf{A}^*)^q\mathbf{A}^q\big) \leq \operatorname{tr}\big((\mathbf{A}\mathbf{A}^*)^q\big) = \operatorname{tr}\big((\mathbf{X}\mathbf{Y}^2\mathbf{X})^q\big) = \operatorname{tr}\big((\mathbf{X}^2\mathbf{Y}^2)^q\big),$$

the last step by Claim 2. Now $\mathbf{X}^2 = \mathbf{X}^*\mathbf{X} \succeq \mathbf{O}$ and $\mathbf{Y}^2 \succeq \mathbf{O}$ by Theorem 13.1.4 (c), so the

inductive hypothesis (b) applies to the pair $\mathbf{X}^2, \mathbf{Y}^2$ and gives

$$\operatorname{tr}\big((\mathbf{X}^2\mathbf{Y}^2)^q\big) \le \operatorname{tr}(\mathbf{X}^{2q}\mathbf{Y}^{2q}).$$

Chaining the three displays proves (b) for $2q$.

Part (a) for $2q$. Put $\mathbf{B} = \mathbf{A}^2$, so $\mathbf{B}^* = (\mathbf{A}^*)^2$, $\mathbf{A}^{2q} = \mathbf{B}^q$, and $(\mathbf{A}^*)^{2q} = (\mathbf{B}^*)^q$. The inductive hypothesis (a), applied to $\mathbf{B}$, and then Claim 2 give

$$\operatorname{tr}\big((\mathbf{A}^*)^{2q}\mathbf{A}^{2q}\big) = \operatorname{tr}\big((\mathbf{B}^*)^q\mathbf{B}^q\big) \le \operatorname{tr}\big((\mathbf{B}\mathbf{B}^*)^q\big) = \operatorname{tr}\big((\mathbf{P}\mathbf{Q})^q\big),$$

with $\mathbf{P} = \mathbf{A}\mathbf{A}^*$ and $\mathbf{Q} = \mathbf{A}^*\mathbf{A}$, both positive semidefinite by Theorem 13.1.4 (c). The inductive hypothesis (b) applied to $\mathbf{P}, \mathbf{Q}$ gives $\operatorname{tr}\big((\mathbf{P}\mathbf{Q})^q\big) \le \operatorname{tr}(\mathbf{P}^q\mathbf{Q}^q)$. Since $\mathbf{P}^q$ and $\mathbf{Q}^q$ are Hermitian, $\operatorname{tr}(\mathbf{P}^q\mathbf{Q}^q) = \langle \mathbf{P}^q, \mathbf{Q}^q\rangle$, and this number is real and ≥ 0 by Lemma 21.12.2, both matrices being positive semidefinite. So Theorem 11.1.5 gives

$$\operatorname{tr}(\mathbf{P}^q\mathbf{Q}^q) = \langle \mathbf{P}^q, \mathbf{Q}^q\rangle \le \|\mathbf{P}^q\|_F \|\mathbf{Q}^q\|_F = \operatorname{tr}(\mathbf{P}^{2q})^{1/2}\operatorname{tr}(\mathbf{Q}^{2q})^{1/2}.$$

Write $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$ for a singular value decomposition (Theorem 13.9.1), with $\mathbf{\Sigma} = \operatorname{diag}(\sigma_1, \dots, \sigma_n)$. Then $\mathbf{P} = \mathbf{U}\mathbf{\Sigma}^2\mathbf{U}^*$ and $\mathbf{Q} = \mathbf{V}\mathbf{\Sigma}^2\mathbf{V}^*$, so $\mathbf{P}$ and $\mathbf{Q}$ are both unitarily similar to $\mathbf{\Sigma}^2$ and therefore have the same eigenvalues with multiplicity. Hence $\operatorname{tr}(\mathbf{P}^{2q}) = \operatorname{tr}(\mathbf{Q}^{2q})$, the displayed bound is $\operatorname{tr}(\mathbf{P}^{2q})$, and

$$\operatorname{tr}\big((\mathbf{A}^*)^{2q}\mathbf{A}^{2q}\big) \le \operatorname{tr}(\mathbf{P}^{2q}) = \operatorname{tr}\big((\mathbf{A}\mathbf{A}^*)^{2q}\big).$$

This is (a) for $2q$, and the induction is complete.

■

The restriction to $m = 2^r$ is a restriction of this proof, not of the truth: (b) holds for every integer $m \ge 1$. It even holds for every real $m \ge 1$, once $(\mathbf{X}\mathbf{Y})^m$ is read as $(\mathbf{X}^{1/2}\mathbf{Y}\mathbf{X}^{1/2})^m$, which Lemma 21.12.2 says is the same thing for integer exponents and which is the only reading available for the others, since $\mathbf{X}\mathbf{Y}$ is neither Hermitian nor positive semidefinite. The general case is a theorem of Lieb and Thirring and is not proved here; nothing below needs it, because a limit along the powers of two is all Lie's formula asks for.

21.12.3 The inequality of Golden and Thompson

Now the two halves fit together. Lie's formula writes $e^{\mathbf{A}+\mathbf{B}}$ as a limit of m-th powers of $e^{\mathbf{A}/m}e^{\mathbf{B}/m}$, and the trace power inequality moves the exponent m off the product and onto the two factors, where it reassembles $e^{\mathbf{A}}$ and $e^{\mathbf{B}}$.

Theorem 21.12.4. The Golden–Thompson Inequality

Let $\mathbf{A}, \mathbf{B} \in M_n(\mathbb{C})$ be Hermitian. Then

$$\operatorname{tr} e^{\mathbf{A}+\mathbf{B}} \;\le\; \operatorname{tr}(e^{\mathbf{A}}e^{\mathbf{B}}).$$

Both sides are real and positive, and equality holds if $\mathbf{A}\mathbf{B} = \mathbf{B}\mathbf{A}$.

Idea. Run the trace power inequality at the m-th root of each exponential and let m run through the powers of two. At each m the right-hand side is exactly $\operatorname{tr}(e^{\mathbf{A}}e^{\mathbf{B}})$, independent

of m, while the left-hand side converges to $\operatorname{tr} e^{\mathbf{A}+\mathbf{B}}$ by Lie's formula. A nonstrict inequality survives a limit.

Proof. Fix $r \geq 0$ and put $m = 2^r$, $\mathbf{X} = e^{\mathbf{A}/m}$, and $\mathbf{Y} = e^{\mathbf{B}/m}$. Since $\mathbf{A}/m$ is Hermitian with spectral resolution $\sum_i \mu_i \mathbf{P}_i$, Theorem 12.8.5 (a) gives $\mathbf{X} = \sum_i e^{\mu_i}\mathbf{P}_i$, which is Hermitian with all eigenvalues $e^{\mu_i} > 0$; so $\mathbf{X} \succ \mathbf{O}$ by Theorem 13.1.5 (b), and likewise $\mathbf{Y} \succ \mathbf{O}$. As in Step 1 of the proof of Lemma 21.12.1, $\mathbf{X}^m = e^{\mathbf{A}}$ and $\mathbf{Y}^m = e^{\mathbf{B}}$.

By Lemma 21.12.3 (b),

$$\operatorname{tr}\Big(\big(e^{\mathbf{A}/m}e^{\mathbf{B}/m}\big)^m\Big) \leq \operatorname{tr}(\mathbf{X}^m\mathbf{Y}^m) = \operatorname{tr}(e^{\mathbf{A}}e^{\mathbf{B}}).$$

The right side does not depend on r. By Lemma 21.12.1 the matrices $\big(e^{\mathbf{A}/m}e^{\mathbf{B}/m}\big)^m$ converge entrywise to $e^{\mathbf{A}+\mathbf{B}}$ as $m = 2^r \to \infty$, so their traces, being sums of n entries, converge to $\operatorname{tr} e^{\mathbf{A}+\mathbf{B}}$. A nonstrict inequality survives a limit, so $\operatorname{tr} e^{\mathbf{A}+\mathbf{B}} \leq \operatorname{tr}(e^{\mathbf{A}}e^{\mathbf{B}})$.

Both sides are real and ≥ 0: the left because $\mathbf{A}+\mathbf{B}$ is Hermitian, so $e^{\mathbf{A}+\mathbf{B}} \succ \mathbf{O}$ by the argument just given, and the right by Lemma 21.12.2 with $k = 1$. They are in fact positive, since a positive definite matrix has positive trace. If $\mathbf{AB} = \mathbf{BA}$, then $e^{\mathbf{A}}e^{\mathbf{B}} = e^{\mathbf{A}+\mathbf{B}}$ by Theorem 10.9.4 (d), and the two sides are equal. ∎

A two-by-two family shows that the inequality is genuinely an inequality, and lets the reader see what the two sides look like.

Example. The inequality on a two-by-two family

For real a, b let

$$\mathbf{A} = \begin{pmatrix} a & 0 \\ 0 & -a \end{pmatrix}, \qquad \mathbf{B} = \begin{pmatrix} 0 & b \\ b & 0 \end{pmatrix}.$$

Compute both sides of Golden–Thompson, and decide when the inequality is strict.

Solution. Both matrices are real symmetric. Write $\mathbf{M} = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$, so that $\mathbf{B} = b\mathbf{M}$ and $\mathbf{M}^2 = \mathbf{I}$. Then $\mathbf{B}^{2m} = b^{2m}\mathbf{I}$ and $\mathbf{B}^{2m+1} = b^{2m+1}\mathbf{M}$, so splitting the exponential series into even and odd terms gives

$$e^{\mathbf{B}} = \cosh(b)\,\mathbf{I} + \sinh(b)\,\mathbf{M} = \begin{pmatrix} \cosh b & \sinh b \\ \sinh b & \cosh b \end{pmatrix},$$

for every real b. Since $e^{\mathbf{A}} = \operatorname{diag}(e^{a}, e^{-a})$, the diagonal entries of $e^{\mathbf{A}}e^{\mathbf{B}}$ are $e^{a}\cosh b$ and $e^{-a}\cosh b$, whence

$$\operatorname{tr}(e^{\mathbf{A}}e^{\mathbf{B}}) = (e^{a} + e^{-a})\cosh b = 2\cosh a\,\cosh b.$$

For the left side, $\mathbf{A}+\mathbf{B} = \left(\begin{smallmatrix} a & b \\ b & -a \end{smallmatrix}\right)$ has trace 0 and determinant $-(a^2+b^2)$, so its eigenvalues are $\pm d$ with $d = \sqrt{a^2+b^2}$. By Theorem 12.8.5 (a) the eigenvalues of $e^{\mathbf{A}+\mathbf{B}}$ are e^{d} and e^{-d}, so $\operatorname{tr} e^{\mathbf{A}+\mathbf{B}} = 2\cosh d$ by Theorem 9.3.3. The inequality therefore reads

$$\cosh\sqrt{a^2+b^2} \;\leq\; \cosh a\,\cosh b,$$

a statement about numbers, and it can be checked term by term. Expanding both sides as power series with nonnegative terms, the coefficient of $a^{2i}b^{2j}$ is $\binom{i+j}{i}/(2i+2j)!$ on the left and $1/((2i)!(2j)!)$ on the right, so what is needed is $\binom{i+j}{i} \le \binom{2i+2j}{2i}$. That holds because doubling each of i chosen items out of $i+j$ produces $2i$ chosen items out of $2i+2j$, injectively; and it is strict as soon as $i, j \ge 1$, since then some $2i$-subset splits a doubled pair and so is not in the image.

Equality therefore forces every term with $i, j \ge 1$ to vanish, that is $a = 0$ or $b = 0$ — exactly the cases where $\mathbf{A}$ and $\mathbf{B}$ commute. For $a = b = 1$ the two sides are $2\cosh\sqrt{2} \approx 4.3564$ and $2\cosh^2 1 \approx 4.7622$.

Warning. **Golden–Thompson is a statement about traces and about nothing else.** It is tempting to guess the Loewner-order version $e^{\mathbf{A}+\mathbf{B}} \preceq e^{\mathbf{A}/2}e^{\mathbf{B}}e^{\mathbf{A}/2}$, whose right side is Hermitian and positive semidefinite and has the same trace as $e^{\mathbf{A}}e^{\mathbf{B}}$. It is false in general. Indeed both sides have the same determinant, namely $e^{\operatorname{tr}\mathbf{A}+\operatorname{tr}\mathbf{B}}$ by Theorem 10.9.4 (e) — and if $\mathbf{M} \succeq \mathbf{N} \succ \mathbf{O}$ with $\det\mathbf{M} = \det\mathbf{N}$, then $\mathbf{Z} = \mathbf{N}^{-1/2}\mathbf{M}\mathbf{N}^{-1/2} \succeq \mathbf{I}$ by Theorem 13.5.3 (a), so all eigenvalues of $\mathbf{Z}$ are ≥ 1 while their product is $\det\mathbf{Z} = \det\mathbf{M}/\det\mathbf{N} = 1$; hence they are all 1, the Hermitian matrix $\mathbf{Z}$ equals $\mathbf{I}$, and $\mathbf{M} = \mathbf{N}^{1/2}\mathbf{Z}\mathbf{N}^{1/2} = \mathbf{N}$. So the Loewner inequality would force $e^{\mathbf{A}+\mathbf{B}} = e^{\mathbf{A}/2}e^{\mathbf{B}}e^{\mathbf{A}/2}$, and taking traces would force equality in Golden–Thompson, which Example (The inequality on a two-by-two family) refutes for $a = b = 1$.

21.12.4 Three matrices are one too many

Golden–Thompson looks as though it should iterate. It does not. The three-matrix statement

$$\operatorname{tr} e^{\mathbf{A}+\mathbf{B}+\mathbf{C}} \;\le\; \operatorname{tr}(e^{\mathbf{A}}e^{\mathbf{B}}e^{\mathbf{C}})$$

is **false**, and the failure is not marginal: the right side can be negative while the left side, a trace of a positive definite matrix, is positive. A family of two-by-two real symmetric matrices shows it, exactly.

Example. The three-matrix version fails

Let $t > 0$ and let $\mathbf{P}_1, \mathbf{P}_2, \mathbf{P}_3$ be the orthogonal projections of $\mathbb{R}^2$ onto the three lines through the origin at angles 0, $2\pi/3$, and $4\pi/3$:

$$\mathbf{P}_1 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad \mathbf{P}_2 = \frac{1}{4}\begin{pmatrix} 1 & -\sqrt{3} \\ -\sqrt{3} & 3 \end{pmatrix}, \quad \mathbf{P}_3 = \frac{1}{4}\begin{pmatrix} 1 & \sqrt{3} \\ \sqrt{3} & 3 \end{pmatrix}.$$

Put $\mathbf{A} = t\mathbf{P}_1$, $\mathbf{B} = t\mathbf{P}_2$, $\mathbf{C} = t\mathbf{P}_3$, all symmetric. Compute both sides of the three-matrix inequality and show that it fails.

Solution. Each $\mathbf{P}_i$ is $\mathbf{u}_i\mathbf{u}_i^{\top}$ for a unit vector $\mathbf{u}_i$, with $\mathbf{u}_1 = (1, 0)$, $\mathbf{u}_2 = (-\frac{1}{2}, \frac{\sqrt{3}}{2})$, $\mathbf{u}_3 = (-\frac{1}{2}, -\frac{\sqrt{3}}{2})$; each is symmetric with $\mathbf{P}_i^2 = \mathbf{P}_i$. Adding the three matrices gives $\mathbf{P}_1+\mathbf{P}_2+\mathbf{P}_3 = \frac{3}{2}\mathbf{I}$, so $\mathbf{A}+\mathbf{B}+\mathbf{C} = \frac{3t}{2}\mathbf{I}$ and

$$\operatorname{tr} e^{\mathbf{A}+\mathbf{B}+\mathbf{C}} = 2e^{3t/2}.$$

For the right side, $\mathbf{P}_i^2 = \mathbf{P}_i$ makes the exponential series collapse: $e^{t\mathbf{P}} = \mathbf{I} + \alpha\mathbf{P}$ with $\alpha = e^t - 1$, because $\sum_{j\ge1} t^j\mathbf{P}^j/j! = (\sum_{j\ge1} t^j/j!)\mathbf{P}$. Expanding the product of three such factors and using $\operatorname{tr}\mathbf{I} = 2$, $\operatorname{tr}\mathbf{P}_i = 1$, $\operatorname{tr}(\mathbf{P}_i\mathbf{P}_j) = (\mathbf{u}_i\cdot\mathbf{u}_j)^2$, and $\operatorname{tr}(\mathbf{P}_1\mathbf{P}_2\mathbf{P}_3) = (\mathbf{u}_1\cdot\mathbf{u}_2)(\mathbf{u}_2\cdot\mathbf{u}_3)(\mathbf{u}_3\cdot\mathbf{u}_1)$,

$$\operatorname{tr}(e^{\mathbf{A}}e^{\mathbf{B}}e^{\mathbf{C}}) = 2 + 3\alpha + \tfrac{3}{4}\alpha^2 - \tfrac{1}{8}\alpha^3,$$

since each pairwise dot product is $-\frac{1}{2}$, so each $\operatorname{tr}(\mathbf{P}_i\mathbf{P}_j) = \frac{1}{4}$ and the triple product is $-\frac{1}{8}$. In terms of t this is

$$\operatorname{tr}(e^{\mathbf{A}}e^{\mathbf{B}}e^{\mathbf{C}}) = \tfrac{1}{8}(-e^{3t} + 9e^{2t} + 9e^t - 1).$$

The cubic term is negative, and for large t it wins. Precisely, if $e^t \ge 10$ then $e^{2t} \ge 10e^t$ and $e^{3t} \ge 10e^{2t} = 9e^{2t} + e^{2t} \ge 9e^{2t} + 10e^t$, so the bracket is negative and $\operatorname{tr}(e^{\mathbf{A}}e^{\mathbf{B}}e^{\mathbf{C}}) < 0$, while $\operatorname{tr} e^{\mathbf{A}+\mathbf{B}+\mathbf{C}} = 2e^{3t/2} > 0$. Taking $t = 3$, where $e^3 > 20$, already does this: the two sides are $2e^{4.5} \approx 180.0$ and $\frac{1}{8}(-e^9 + 9e^6 + 9e^3 - 1) \approx -536.6$.

The failure is not confined to large t. At $t = 1$ the two sides are $2e^{3/2} \approx 8.9634$ and $\frac{1}{8}(-e^3+9e^2+9e-1) \approx 8.7351$, and at $t = 2$ they are $2e^3 \approx 40.171$ and $\frac{1}{8}(-e^6+9e^4+9e^2-1) \approx 19.182$.

The numbers in the example are exact, and the two closed forms may be compared by hand; only their decimal expansions are approximations.

Remark. There is a second, more basic obstruction. For three Hermitian matrices the number $\operatorname{tr}(e^{\mathbf{A}}e^{\mathbf{B}}e^{\mathbf{C}})$ need not be **real**, so no inequality can even be written down. With

$$\mathbf{A} = a\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \mathbf{B} = b\begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad \mathbf{C} = c\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix},$$

each of the three squares is a multiple of $\mathbf{I}$, so each exponential splits into a hyperbolic cosine times $\mathbf{I}$ plus a hyperbolic sine times the matrix, and multiplying out gives

$$\operatorname{tr}(e^{\mathbf{A}}e^{\mathbf{B}}e^{\mathbf{C}}) = 2\cosh a\cosh b\cosh c + 2i\sinh a\sinh b\sinh c,$$

because the product of the three matrices above is $i\mathbf{I}$ while every shorter product among them has trace 0. The imaginary part is nonzero whenever a, b, c are all nonzero. With two matrices this cannot happen, by Lemma 21.12.2.

Remark. Two of the three matrices commuting is enough to restore the inequality, and exercise Exercise (C1: Two out of three is enough) asks for the three cases. So the example above is as economical as it can be: no two of $\mathbf{P}_1, \mathbf{P}_2, \mathbf{P}_3$ commute.

21.12.5 Klein's inequality and relative entropy

Golden–Thompson is one of a family of trace inequalities whose proofs all end in the same place: a scalar convexity inequality summed against nonnegative weights that add to one. The next theorem is the cleanest member of the family. It needs one fact about convex functions of a real variable, which Chapter 18 recorded only in an exercise, so we prove it here.

Lemma 21.12.5. A Convex Function Lies Above Its Tangent

Let $I \subseteq \mathbb{R}$ be an open interval and let $f\colon I \to \mathbb{R}$ be convex and differentiable. Then

$$f(s) \;\geq\; f(t) + f'(t)(s-t) \qquad \text{for all } s,t \in I.$$

Proof. Fix $s,t \in I$. If $s = t$ both sides are $f(t)$. Assume $s \neq t$ and let $0 < \theta \leq 1$. The point $t + \theta(s-t) = (1-\theta)t + \theta s$ lies in I, and Definition 18.10.1 gives

$$f\big(t + \theta(s-t)\big) \leq (1-\theta) f(t) + \theta f(s),$$

that is, $\big(f(t+\theta(s-t)) - f(t)\big)/\theta \leq f(s) - f(t)$. Writing $h = \theta(s-t)$, which is nonzero and tends to 0 as $\theta \to 0^+$, the left side equals

$$(s-t)\cdot\frac{f(t+h)-f(t)}{h} \;\longrightarrow\; (s-t)f'(t),$$

by differentiability of f at t. A nonstrict inequality survives a limit, so $(s-t)f'(t) \leq f(s) - f(t)$, which is the claim.

■

Theorem 21.12.6. Klein's Inequality

Let $I \subseteq \mathbb{R}$ be an open interval and let $f\colon I \to \mathbb{R}$ be convex and differentiable. Let $\mathbf{A}, \mathbf{B} \in M_n(\mathbb{C})$ be Hermitian with $\operatorname{spec}(\mathbf{A}) \subseteq I$ and $\operatorname{spec}(\mathbf{B}) \subseteq I$. Then

$$\operatorname{tr}\big(f(\mathbf{A}) - f(\mathbf{B}) - (\mathbf{A}-\mathbf{B})f'(\mathbf{B})\big) \;\geq\; 0,$$

where $f(\mathbf{A})$, $f(\mathbf{B})$, and $f'(\mathbf{B})$ are the functional calculus of Definition 12.8.2.

Idea. Compute the trace in an orthonormal eigenbasis of $\mathbf{B}$. Each diagonal entry then involves $\mathbf{B}$ only through one eigenvalue μ_j, and involves $\mathbf{A}$ only through the numbers $c_{ij} = |\langle \mathbf{v}_j, \mathbf{u}_i\rangle|^2$, the overlaps between the two eigenbases. Those numbers are nonnegative and each column of the array sums to 1, which is exactly what is needed to write the whole trace as a weighted average of the scalar quantities $f(\lambda_i) - f(\mu_j) - f'(\mu_j)(\lambda_i - \mu_j)$. Every one of those is ≥ 0 by Lemma 21.12.5.

Proof. By Corollary 12.4.3 there are orthonormal bases $(\mathbf{u}_1,\dots,\mathbf{u}_n)$ and $(\mathbf{v}_1,\dots,\mathbf{v}_n)$ of $\mathbb{C}^n$ with $\mathbf{A}\mathbf{u}_i = \lambda_i \mathbf{u}_i$ and $\mathbf{B}\mathbf{v}_j = \mu_j \mathbf{v}_j$; all λ_i and μ_j lie in I. Put $c_{ij} = |\langle \mathbf{v}_j, \mathbf{u}_i\rangle|^2 \geq 0$. Expanding $\mathbf{v}_j$ in the orthonormal basis $(\mathbf{u}_i)$ gives $\sum_i c_{ij} = \|\mathbf{v}_j\|^2 = 1$ for each j.

For any $\mathbf{M} \in M_n(\mathbb{C})$ we have $\operatorname{tr}\mathbf{M} = \sum_j \langle \mathbf{M}\mathbf{v}_j, \mathbf{v}_j\rangle$: with $\mathbf{V}$ the unitary matrix whose columns are the $\mathbf{v}_j$, the sum is $\operatorname{tr}(\mathbf{V}^*\mathbf{M}\mathbf{V}) = \operatorname{tr}(\mathbf{M}\mathbf{V}\mathbf{V}^*) = \operatorname{tr}\mathbf{M}$ by Theorem 0.9.6 (3). We evaluate the three terms at $\mathbf{v}_j$.

By Definition 12.8.2, $f(\mathbf{A})\mathbf{u}_i = f(\lambda_i)\mathbf{u}_i$, so expanding $\mathbf{v}_j = \sum_i \langle \mathbf{v}_j, \mathbf{u}_i\rangle \mathbf{u}_i$ and using orthonor-

mality,

$$\langle f(\mathbf{A})\mathbf{v}_j, \mathbf{v}_j\rangle = \sum_i f(\lambda_i)\, c_{ij}, \qquad \langle \mathbf{A}\mathbf{v}_j, \mathbf{v}_j\rangle = \sum_i \lambda_i\, c_{ij}.$$

Also $f(\mathbf{B})\mathbf{v}_j = f(\mu_j)\mathbf{v}_j$ and $f'(\mathbf{B})\mathbf{v}_j = f'(\mu_j)\mathbf{v}_j$, so $\langle f(\mathbf{B})\mathbf{v}_j, \mathbf{v}_j\rangle = f(\mu_j)$ and

$$\langle (\mathbf{A}-\mathbf{B})f'(\mathbf{B})\mathbf{v}_j, \mathbf{v}_j\rangle = f'(\mu_j)\Big(\sum_i \lambda_i c_{ij} - \mu_j\Big).$$

Summing over j, and using $\sum_i c_{ij} = 1$ to write $f(\mu_j) = \sum_i c_{ij} f(\mu_j)$ and $\mu_j = \sum_i c_{ij}\mu_j$,

$$\operatorname{tr}\big(f(\mathbf{A}) - f(\mathbf{B}) - (\mathbf{A}-\mathbf{B})f'(\mathbf{B})\big) = \sum_{i,j} c_{ij}\,\big[\, f(\lambda_i) - f(\mu_j) - f'(\mu_j)(\lambda_i - \mu_j)\,\big].$$

Every bracket is ≥ 0 by Lemma 21.12.5, and every $c_{ij} \geq 0$. This proves the inequality.

■

The corollary that made Klein's inequality famous concerns density matrices, the positive semidefinite matrices of trace 1 that Chapter 18 §08 met as a convex body (Definition 18.8.3). It needs the derivative of the logarithm, which the book has not yet computed.

Lemma 21.12.7. The Derivative of the Logarithm

The function $\log\colon (0,\infty) \to \mathbb{R}$ of Lemma 18.11.1 is differentiable with $\log'(y) = 1/y$. Consequently $g(t) = t\log t$ is differentiable on $(0,\infty)$ with $g'(t) = \log t + 1$ and $g''(t) = 1/t$, and g is strictly convex there.

Proof. Continuity. By Lemma 18.11.1 (a) and (b), $\exp\colon \mathbb{R} \to (0,\infty)$ is a strictly increasing bijection, so log is strictly increasing. Fix $y > 0$, put $s = \log y$, and let $\varepsilon > 0$. Then $e^{s-\varepsilon} < y < e^{s+\varepsilon}$, and for y' in that open interval, applying the increasing log gives $s - \varepsilon < \log y' < s + \varepsilon$. So log is continuous at y.

Differentiability. Let $h \neq 0$ be small enough that $y + h > 0$, and put $u = \log(y+h) - s$, so that $y + h = e^{s+u} = ye^u$ and $h = y(e^u - 1)$. Since log is injective, $u \neq 0$, and $u \to 0$ as $h \to 0$ by continuity. Hence

$$\frac{\log(y+h) - \log y}{h} = \frac{u}{y(e^u - 1)} \longrightarrow \frac{1}{y},$$

because $(e^u - 1)/u \to 1$: the exponential is its own derivative (Theorem 10.9.4 (b) with $n = 1$), and its value at 0 is 1.

The function $t\log t$. For $t > 0$ and small $h \neq 0$,

$$\frac{(t+h)\log(t+h) - t\log t}{h} = t \cdot \frac{\log(t+h) - \log t}{h} + \log(t+h),$$

which tends to $t \cdot (1/t) + \log t = 1 + \log t$, using continuity of log. Differentiating once more, $g''(t) = 1/t > 0$, so g is strictly convex by Lemma 18.10.5.

■

Corollary 21.12.8. Nonnegativity of the Relative Entropy

Let $\mathbf{A}, \mathbf{B} \in M_n(\mathbb{C})$ be positive definite with $\operatorname{tr}\mathbf{A} = \operatorname{tr}\mathbf{B} = 1$. Then

$$\operatorname{tr}\big(\mathbf{A}(\log\mathbf{A} - \log\mathbf{B})\big) \;\geq\; 0.$$

Proof. Apply Theorem 21.12.6 on $I = (0,\infty)$ to $f(t) = t\log t$, which is convex and differentiable with $f'(t) = \log t + 1$ by Lemma 21.12.7. Both spectra lie in I by Theorem 13.1.5 (b). By Theorem 12.8.3 (b), $f'(\mathbf{B}) = \log\mathbf{B} + \mathbf{I}$. So Klein's inequality reads

$$\operatorname{tr}(\mathbf{A}\log\mathbf{A}) - \operatorname{tr}(\mathbf{B}\log\mathbf{B}) - \operatorname{tr}\big((\mathbf{A}-\mathbf{B})(\log\mathbf{B} + \mathbf{I})\big) \;\geq\; 0.$$

Expanding the last trace by linearity, it equals $\operatorname{tr}(\mathbf{A}\log\mathbf{B}) - \operatorname{tr}(\mathbf{B}\log\mathbf{B}) + \operatorname{tr}\mathbf{A} - \operatorname{tr}\mathbf{B}$. Substituting and canceling $\operatorname{tr}(\mathbf{B}\log\mathbf{B})$,

$$\operatorname{tr}(\mathbf{A}\log\mathbf{A}) - \operatorname{tr}(\mathbf{A}\log\mathbf{B}) - \operatorname{tr}\mathbf{A} + \operatorname{tr}\mathbf{B} \;\geq\; 0,$$

and $\operatorname{tr}\mathbf{A} = \operatorname{tr}\mathbf{B} = 1$ removes the last two terms.

■

The quantity $\operatorname{tr}(\mathbf{A}(\log\mathbf{A} - \log\mathbf{B}))$ is the **relative entropy** of $\mathbf{A}$ with respect to $\mathbf{B}$, and the corollary says it is a genuine measure of discrepancy: never negative, and zero when $\mathbf{A} = \mathbf{B}$. For diagonal $\mathbf{A}$ and $\mathbf{B}$ it is the familiar sum $\sum_i a_i(\log a_i - \log b_i)$ over two probability vectors, so the corollary contains the scalar statement as its commuting case.

Remark. Klein's inequality is the trace-level shadow of §02's Lemma 21.2.7 and Theorem 21.2.8: all three say that a convex function applied to a Hermitian matrix behaves, after a trace, like a convex function applied to a weighted average of eigenvalues, the weights being the overlaps between two eigenbases. The array (c_{ij}) in the proof above is doubly stochastic (Definition 18.8.4), and Chapter 20 §05 used the same array for Hoffman–Wielandt. It is the same object doing the same job.

21.12.6 What is stated here and not proved

One theorem belongs in this section by subject and is beyond its means.

Lieb's concavity theorem. *Let $\mathbf{X} \in M_n(\mathbb{C})$ be fixed and let $p, q \geq 0$ with $p + q \leq 1$. Then the map*

$$(\mathbf{A}, \mathbf{B}) \;\longmapsto\; \operatorname{tr}(\mathbf{X}^*\mathbf{A}^p\mathbf{X}\mathbf{B}^q)$$

is jointly concave on pairs of positive definite matrices. This is a theorem of Lieb, and it is **not proved in this book**. What it needs is machinery this chapter does not build: either the theory of tensor products of positive maps, or a differentiation argument through the integral representations of §10 that would have to control two variables at once. It is recorded here because it is the natural strengthening of everything above — Golden–Thompson and the concavity of $(\mathbf{A}, \mathbf{B}) \mapsto \operatorname{tr} e^{\log\mathbf{A} + \log\mathbf{B}}$ both follow from it — and because a reader who meets it elsewhere should know where it sits.

Nothing in this book depends on it. No statement in Chapters 0 to 21 is proved using

Lieb's theorem, and no later chapter is promised a consequence of it.

21.12.7 Summary and transfer

Five earlier chapters left inequalities unproved, each of them an inequality that was supposed to hold for **every** unitarily invariant norm, or for **every** convex function. This chapter proved them, and the reason they could all be proved at once is that they are all the same statement about vectors: a majorization.

- **Majorization is a doubly stochastic map.** §01 proved the theorem of Hardy, Littlewood, and Pólya, Theorem 21.1.3, that $\mathbf{x} \prec \mathbf{y}$ exactly when $\mathbf{x} = \mathbf{D}\mathbf{y}$ for a doubly stochastic $\mathbf{D}$, and with Birkhoff's theorem from Chapter 19, exactly when $\mathbf{x}$ lies in the convex hull of the permutations of $\mathbf{y}$ (Corollary 21.1.4).
- **Hence every convex function.** §02's Theorem 21.2.1 turned that into $\sum_i \phi(x_i) \le \sum_i \phi(y_i)$, and Theorem 21.2.8 turned it into the convexity of $\mathbf{A} \mapsto \operatorname{tr} f(\mathbf{A})$.
- **Hence every unitarily invariant norm.** §03 identified the symmetric gauge functions and showed that they are monotone under weak majorization (Theorem 21.3.4); §04's Theorem 21.4.3 matched them one for one with the unitarily invariant norms of Definition 21.4.1, through $\|\mathbf{A}\| = \Phi(\sigma(\mathbf{A}))$; §05 proved von Neumann's trace inequality; and §06's Theorem 21.6.1 closed the circle, making $\sigma(\mathbf{A}) \prec_w \sigma(\mathbf{B})$ *equivalent* to $\|\mathbf{A}\| \le \|\mathbf{B}\|$ for every such norm.
- **The applications.** §06 deduced Eckart–Young (Corollary 21.6.2) and the nearest-unitary theorem (Corollary 21.6.4) in every unitarily invariant norm; §07 did the same for Lidskii (Theorem 21.7.2) and Mirsky (Theorem 21.7.4).
- **The order-theoretic half.** §08 studied operator convexity and proved the Hansen–Pedersen–Jensen inequality (Theorem 21.8.5); §09 proved that operator monotonicity of order n is exactly positivity of every $n \times n$ Loewner matrix (Theorem 21.9.5); §10 assembled Loewner's theorem (Theorem 21.10.4), and with it Corollary 21.10.8 and Corollary 21.10.9; §11 built the matrix geometric mean (Definition 21.11.1).
- **The exponential.** This section proved Lie's product formula, the trace power inequality, Golden–Thompson, and Klein's inequality, and showed that three matrices break the pattern.

Six moves did the work of the chapter.

- **A majorization is a doubly stochastic map.** To prove something for all $\mathbf{x} \prec \mathbf{y}$, write $\mathbf{x} = \mathbf{D}\mathbf{y}$ and let convexity act one row at a time (§01, §02). *Transfer:* when a hypothesis is a family of inequalities between partial sums, look for the linear map that realizes it; the inequalities are usually the vertices of a polytope in disguise.
- **A symmetric gauge turns a weak majorization into a norm inequality.** A norm that cannot tell coordinates apart, and cannot see their signs, is monotone along $\prec_w$ (§03, §04). *Transfer:* to prove one inequality for a whole family of norms, find the single combinatorial statement about the underlying numbers that all of them respect.
- **A Hermitian dilation turns singular values into eigenvalues.** Chapter 17 §09's $\mathcal{H}(\mathbf{A})$ has eigenvalues $\pm\sigma_i(\mathbf{A})$, so every theorem about eigenvalues of Hermitian matrices has a singular-value twin (§04, §07). *Transfer:* when a result is known for Hermitian matrices and wanted for all matrices, look for the Hermitian matrix whose spectrum encodes the data.
- **The Schur product theorem turns a Loewner matrix into monotonicity.** The derivative of f at a diagonal matrix is a Hadamard product with the Loewner matrix, and Chapter 13 §08 says that a Hadamard product of positive semidefinite matrices is positive semidefinite (§09). *Transfer:* to show that a matrix-valued function is increasing,

differentiate it and recognize the derivative as an entrywise product.

- **A limit of Riemann sums carries the Loewner order through an integral.** The positive semidefinite cone is closed, so an integral of positive semidefinite matrices is positive semidefinite (§10). *Transfer:* an order defined by a closed convex cone survives any limit, so it survives integration; the only work is building the integral.
- **Lie's product formula turns a sum in the exponent into a product.** Lemma 21.12.1 replaces $e^{\mathbf{A}+\mathbf{B}}$, which nothing can be said about directly, by a limit of products, each of which can be attacked with Cauchy–Schwarz (this section). *Transfer:* when two operations fail to commute, split one of them into k small steps, estimate the error of a single step as $O(1/k^2)$, and let telescoping absorb the factor k.

The chapter paid the following promises, in the words in which they were made.

The promise	Made in	Paid by
Hardy–Littlewood–Pólya, "$\mathbf{x} \prec \mathbf{y}$ exactly when $\mathbf{x} = \mathbf{D}\mathbf{y}$ for a doubly stochastic $\mathbf{D}$"	Chapter 19 §09	§01
"the general consequences of a majorization $\mathbf{x} \prec \mathbf{y}$, namely $\sum_i \phi(x_i) \le \sum_i \phi(y_i)$ for every convex ϕ"	Chapter 17 §07	§02 and §07
"for the whole class of functions that reverse majorization" (the product is Schur-concave)	Chapter 17 §08	§02
"$\mathbf{A} \mapsto \operatorname{tr} f(\mathbf{A})$ … is proved in Chapter 21"	Chapter 18 §10	§02
"It has not said which norms are invariant under unitary multiplication, which is Chapter 21"	Chapter 16 §08	§04
"The repair is to apply N_k to the singular values instead of the eigenvalues, which is what Chapter 21 does"	Chapter 17 §06	§04
Eckart–Young "in **every** unitarily invariant norm", by a route that "needs inequalities between the singular values of $\mathbf{A}$ and of $\mathbf{A} - \mathbf{B}$, not a projection argument"	Chapter 13 §11, Chapter 16 §08, Chapter 17 §09	§06
"$\mathbf{W}$ also minimizes the distance in every unitarily invariant norm"	Chapter 13 §10	§06
"with $\phi(t) = t^2$ this is Corollary 20.5.4 again"	Chapter 20 §05	§07

The promise	Made in	Paid by
"one statement covering every unitarily invariant norm" (the index-by-index singular-value bound and Mirsky)	Chapter 20 §10 and §11	§07
"$t \mapsto t^s$ is operator monotone on $[0,\infty)$ exactly for $0 \le s \le 1$", with "the companion fact that $t \mapsto \log t$ is operator monotone while $t \mapsto e^t$ is not"	Chapter 13 §05	§10, modulo (A7)

Some things the chapter did not do. Loewner's theorem is proved here only up to one quoted fact, **(A7)**, the Nevanlinna–Pick characterization of Pick functions; §10 says precisely which implication rests on it, and §09's elementary half — that for a continuously differentiable f, monotonicity of every order is exactly positivity of every Loewner matrix — is proved in full. The Lieb–Thirring inequality, which would remove the restriction to $m = 2^r$ in Lemma 21.12.3, is not proved, and nothing here needs it. Lieb's concavity theorem is stated above and not proved, and nothing in the book depends on it. The equality cases of von Neumann's trace inequality and of Golden–Thompson were treated only in the directions that were needed.

Part V ends here. It began in Chapter 16 with a complaint: nine chapters had said "close" and "small" with nothing to measure by. A norm supplied the yardstick, and the five chapters since have been spent finding out what a yardstick is good for. Chapter 17 made the eigenvalues of a Hermitian matrix into the values of an optimization problem, so that they could be compared. Chapter 18 supplied the language of convexity in which such comparisons live. Chapter 19 studied the one class of non-Hermitian matrices — the nonnegative ones — where a comparable theory exists, and produced Birkhoff's theorem, which this chapter used once. Chapter 20 asked how much the answers move when the data move. And Chapter 21 has collected the inequalities that all five chapters kept needing, and shown that they are one inequality seen from different sides. What none of these chapters has done is compute anything: every proof here is an existence or a comparison, and not one of them is an algorithm. That is the subject of Chapter 24.

21.12.8 Exercises

21.12.8.1 A. Check your understanding

Exercise. A1

(a) State the Golden–Thompson inequality, with its hypotheses.

(b) State Lie's product formula, and say for which matrices it holds.

(c) True or false: for Hermitian $\mathbf{A}, \mathbf{B}$ the matrix $e^{\mathbf{A}}e^{\mathbf{B}}$ is positive semidefinite. Justify your answer.

(d) True or false: $\operatorname{tr} e^{\mathbf{A}+\mathbf{B}} \le \operatorname{tr}(e^{\mathbf{B}}e^{\mathbf{A}})$ for Hermitian $\mathbf{A}, \mathbf{B}$. Justify your answer.

(e) True or false: $\operatorname{tr} e^{\mathbf{A}+\mathbf{B}+\mathbf{C}} \le \operatorname{tr}(e^{\mathbf{A}}e^{\mathbf{B}}e^{\mathbf{C}})$ for Hermitian $\mathbf{A}, \mathbf{B}, \mathbf{C}$. Justify your answer.

(f) Where in the proof of Theorem 21.12.4 is the hypothesis that $\mathbf{A}$ and $\mathbf{B}$ are Hermitian used?

Solution.

(a) For Hermitian $\mathbf{A}, \mathbf{B} \in M_n(\mathbb{C})$, $\operatorname{tr} e^{\mathbf{A}+\mathbf{B}} \le \operatorname{tr}(e^{\mathbf{A}} e^{\mathbf{B}})$ (Theorem 21.12.4). No commutativity is assumed; if $\mathbf{AB} = \mathbf{BA}$ the two sides are equal.

(b) $e^{\mathbf{A}+\mathbf{B}} = \lim_k (e^{\mathbf{A}/k} e^{\mathbf{B}/k})^k$ for **all** $\mathbf{A}, \mathbf{B} \in M_n(\mathbb{C})$, Hermitian or not (Lemma 21.12.1).

(c) False, since positive semidefinite includes Hermitian (Definition 13.1.1 (P1)) and $(e^{\mathbf{A}} e^{\mathbf{B}})^* = e^{\mathbf{B}} e^{\mathbf{A}}$, which equals $e^{\mathbf{A}} e^{\mathbf{B}}$ only when the two exponentials commute. In Example (The inequality on a two-by-two family) with $a = b = 1$,

$$e^{\mathbf{A}} e^{\mathbf{B}} = \begin{pmatrix} e \cosh 1 & e \sinh 1 \\ e^{-1} \sinh 1 & e^{-1} \cosh 1 \end{pmatrix},$$

which is not symmetric because $e \sinh 1 \neq e^{-1} \sinh 1$.

(d) True. $\operatorname{tr}(e^{\mathbf{B}} e^{\mathbf{A}}) = \operatorname{tr}(e^{\mathbf{A}} e^{\mathbf{B}})$ by Theorem 0.9.6 (3), and $\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}$, so this is Theorem 21.12.4 with the roles exchanged.

(e) False; see Example (The three-matrix version fails), where the right side is negative and the left side positive. The right side need not even be real (the remark after that example).

(f) In two places. First, to know that $e^{\mathbf{A}/m}$ and $e^{\mathbf{B}/m}$ are positive semidefinite, which is the hypothesis of Lemma 21.12.3 (b). Second, in Lemma 21.12.2, which is what makes $\operatorname{tr}(e^{\mathbf{A}} e^{\mathbf{B}})$ real so that the inequality is a statement about real numbers.

21.12.8.2 B. Practice

Exercise. B1: Both sides, exactly

Let

$$\mathbf{A} = \begin{pmatrix} 0 & 0 \\ 0 & -2 \end{pmatrix}, \qquad \mathbf{B} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

(a) Compute $e^{\mathbf{A}}$, $e^{\mathbf{B}}$, and $\operatorname{tr}(e^{\mathbf{A}} e^{\mathbf{B}})$ exactly.

(b) Compute $\operatorname{tr} e^{\mathbf{A}+\mathbf{B}}$ exactly, and verify Theorem 21.12.4 numerically to three decimal places.

Solution.

(a) $\mathbf{A}$ is diagonal, so $e^{\mathbf{A}} = \operatorname{diag}(1, e^{-2})$. Since $\mathbf{B}^2 = \mathbf{I}$, splitting the series into even and odd terms gives $e^{\mathbf{B}} = \cosh(1)\mathbf{I} + \sinh(1)\mathbf{B}$, that is

$$e^{\mathbf{B}} = \begin{pmatrix} \cosh 1 & \sinh 1 \\ \sinh 1 & \cosh 1 \end{pmatrix}.$$

Hence $e^{\mathbf{A}} e^{\mathbf{B}}$ has diagonal entries $\cosh 1$ and $e^{-2} \cosh 1$, so

$$\operatorname{tr}(e^{\mathbf{A}} e^{\mathbf{B}}) = (1 + e^{-2}) \cosh 1 = \frac{(1 + e^{-2})(e + e^{-1})}{2}.$$

Numerically $(1 + 0.135335)(1.543081) = 1.751914$.

(b) $\mathbf{A}+\mathbf{B} = \left(\begin{smallmatrix} 0 & 1 \\ 1 & -2 \end{smallmatrix}\right)$ has trace -2 and determinant -1, so its characteristic polynomial is $x^2 + 2x - 1$ and its eigenvalues are $-1 \pm \sqrt{2}$. A Hermitian matrix has $\operatorname{tr} e^{\mathbf{M}} = \sum_i e^{\lambda_i}$ by Theorem 12.8.5 (a) and Theorem 9.3.3, so

$$\operatorname{tr} e^{\mathbf{A}+\mathbf{B}} = e^{-1+\sqrt{2}} + e^{-1-\sqrt{2}} = 2e^{-1}\cosh\sqrt{2}.$$

Numerically $2(0.367879)(2.178183) = 1.602618$. Indeed $1.603 \le 1.752$, and the inequality is strict because $\mathbf{A}$ and $\mathbf{B}$ do not commute: $\mathbf{AB} = \left(\begin{smallmatrix} 0 & 0 \\ -2 & 0 \end{smallmatrix}\right)$ while $\mathbf{BA} = \left(\begin{smallmatrix} 0 & -2 \\ 0 & 0 \end{smallmatrix}\right)$.

Exercise. B2: The trace power inequality on integers

Let $\mathbf{X} = \operatorname{diag}(2, 1)$ and $\mathbf{Y} = \left(\begin{smallmatrix} 1 & 1 \\ 1 & 1 \end{smallmatrix}\right)$.

(a) Check that $\mathbf{X}$ and $\mathbf{Y}$ are positive semidefinite, and compute $\operatorname{tr}\big((\mathbf{XY})^m\big)$ and $\operatorname{tr}(\mathbf{X}^m\mathbf{Y}^m)$ for $m = 1, 2, 4$.

(b) Verify Lemma 21.12.3 (b) in these three cases, and say for which of them it is strict.

Solution.

(a) $\mathbf{X}$ is diagonal, so its eigenvalues are 2 and 1; $\mathbf{Y}$ is symmetric with trace 2 and determinant 0, so its eigenvalues are 2 and 0. Both lists are nonnegative, so $\mathbf{X} \succeq \mathbf{O}$ and $\mathbf{Y} \succeq \mathbf{O}$ by Theorem 13.1.4 (b).
Now $\mathbf{XY} = \left(\begin{smallmatrix} 2 & 2 \\ 1 & 1 \end{smallmatrix}\right)$, which has trace 3 and determinant 0, so its eigenvalues are 3 and 0 and $\operatorname{tr}\big((\mathbf{XY})^m\big) = 3^m$: the values are $3, 9, 81$. On the other side $\mathbf{Y}^m = 2^{m-1}\mathbf{Y}$, since $\mathbf{Y}^2 = 2\mathbf{Y}$, and $\mathbf{X}^m = \operatorname{diag}(2^m, 1)$, so

$$\operatorname{tr}(\mathbf{X}^m\mathbf{Y}^m) = 2^{m-1}\operatorname{tr}\big(\operatorname{diag}(2^m, 1)\mathbf{Y}\big) = 2^{m-1}(2^m + 1),$$

giving $3, 10, 136$ for $m = 1, 2, 4$.

(b) The inequality $3^m \le 2^{m-1}(2^m + 1)$ holds in all three cases. It is an equality at $m = 1$, as the base case of the induction says it must be, and strict at $m = 2$ ($9 < 10$) and $m = 4$ ($81 < 136$).

Exercise. B3: Klein with the exponential

Let $\mathbf{A}, \mathbf{B} \in M_n(\mathbb{C})$ be Hermitian. Use Theorem 21.12.6 with $f(t) = e^t$ to prove

$$\operatorname{tr} e^{\mathbf{A}} \;\ge\; \operatorname{tr} e^{\mathbf{B}} + \operatorname{tr}\big((\mathbf{A}-\mathbf{B})e^{\mathbf{B}}\big),$$

and check the inequality for $\mathbf{A} = \operatorname{diag}(1, 0)$, $\mathbf{B} = \mathbf{O}$.

Solution. The function $f(t) = e^t$ is differentiable on $I = \mathbb{R}$ with $f' = f$, and it is convex by Lemma 18.11.1 (a). Both spectra are real, hence contained in I. So Theorem 21.12.6 gives

$$\operatorname{tr}\big(e^{\mathbf{A}} - e^{\mathbf{B}} - (\mathbf{A}-\mathbf{B})e^{\mathbf{B}}\big) \;\ge\; 0,$$

and the trace is linear (Theorem 0.9.6 (1)), so this rearranges to the stated inequality.

For $\mathbf{A} = \operatorname{diag}(1,0)$ and $\mathbf{B} = \mathbf{O}$: $\operatorname{tr} e^{\mathbf{A}} = e + 1 \approx 3.71828$, $\operatorname{tr} e^{\mathbf{B}} = \operatorname{tr} \mathbf{I} = 2$, and $\operatorname{tr}(\mathbf{A} e^{\mathbf{O}}) = \operatorname{tr} \mathbf{A} = 1$. So the inequality reads $3.71828 \geq 3$, and it is strict.

21.12.8.3 C. Going deeper

Exercise. C1: Two out of three is enough

Let $\mathbf{A}, \mathbf{B}, \mathbf{C} \in M_n(\mathbb{C})$ be Hermitian. Prove that

$$\operatorname{tr} e^{\mathbf{A}+\mathbf{B}+\mathbf{C}} \;\leq\; \operatorname{tr}\!\left(e^{\mathbf{A}} e^{\mathbf{B}} e^{\mathbf{C}}\right)$$

holds whenever **some two** of the three matrices commute, treating the three cases separately. Explain why this does not contradict Example (The three-matrix version fails).

Hint: Theorem 10.9.4 (d), and the cyclic invariance of the trace.

Solution. Case 1: $\mathbf{BC} = \mathbf{CB}$. Then $e^{\mathbf{B}} e^{\mathbf{C}} = e^{\mathbf{B}+\mathbf{C}}$ by Theorem 10.9.4 (d), so $\operatorname{tr}(e^{\mathbf{A}} e^{\mathbf{B}} e^{\mathbf{C}}) = \operatorname{tr}(e^{\mathbf{A}} e^{\mathbf{B}+\mathbf{C}})$. Since $\mathbf{B}+\mathbf{C}$ is Hermitian, Theorem 21.12.4 applied to the pair $\mathbf{A}, \mathbf{B}+\mathbf{C}$ gives $\operatorname{tr} e^{\mathbf{A}+\mathbf{B}+\mathbf{C}} \leq \operatorname{tr}(e^{\mathbf{A}} e^{\mathbf{B}+\mathbf{C}})$, as required.

Case 2: $\mathbf{AB} = \mathbf{BA}$. Then $e^{\mathbf{A}} e^{\mathbf{B}} = e^{\mathbf{A}+\mathbf{B}}$, so $\operatorname{tr}(e^{\mathbf{A}} e^{\mathbf{B}} e^{\mathbf{C}}) = \operatorname{tr}(e^{\mathbf{A}+\mathbf{B}} e^{\mathbf{C}})$, and Theorem 21.12.4 applied to $\mathbf{A}+\mathbf{B}$ and $\mathbf{C}$ finishes.

Case 3: $\mathbf{AC} = \mathbf{CA}$. By Theorem 0.9.6 (3), $\operatorname{tr}(e^{\mathbf{A}} e^{\mathbf{B}} e^{\mathbf{C}}) = \operatorname{tr}(e^{\mathbf{C}} e^{\mathbf{A}} e^{\mathbf{B}}) = \operatorname{tr}(e^{\mathbf{A}+\mathbf{C}} e^{\mathbf{B}})$, the last step because $\mathbf{A}$ and $\mathbf{C}$ commute. Now Theorem 21.12.4 applied to $\mathbf{A}+\mathbf{C}$ and $\mathbf{B}$ gives $\operatorname{tr} e^{\mathbf{A}+\mathbf{B}+\mathbf{C}} \leq \operatorname{tr}(e^{\mathbf{A}+\mathbf{C}} e^{\mathbf{B}})$.

There is no contradiction: in Example (The three-matrix version fails) no two of $\mathbf{P}_1, \mathbf{P}_2, \mathbf{P}_3$ commute. Indeed, suppose two orthogonal projections $\mathbf{P}, \mathbf{Q}$ onto distinct lines commute. Then $\mathbf{PQ}$ is Hermitian, since $(\mathbf{PQ})^* = \mathbf{QP} = \mathbf{PQ}$, and idempotent, since $\mathbf{PQPQ} = \mathbf{P}^2\mathbf{Q}^2 = \mathbf{PQ}$; so it is an orthogonal projection, and its image lies in the image of $\mathbf{P}$ and in that of $\mathbf{Q}$, which meet only in $\mathbf{0}$. Hence $\mathbf{PQ} = \mathbf{O}$, which for $\mathbf{P} = \mathbf{u}\mathbf{u}^{\top}$ and $\mathbf{Q} = \mathbf{w}\mathbf{w}^{\top}$ says $\mathbf{u} \cdot \mathbf{w} = 0$. But the three unit vectors above pair to $-\tfrac{1}{2}$, never 0. For instance $\mathbf{P}_1\mathbf{P}_2 = \tfrac{1}{4}\left(\begin{smallmatrix} 1 & -\sqrt{3} \\ 0 & 0 \end{smallmatrix}\right) \neq \mathbf{O}$.

Exercise. C2: The symmetric Lie formula

Let $\mathbf{A}, \mathbf{B} \in M_n(\mathbb{C})$.

(a) Prove that $e^{\mathbf{A}+\mathbf{B}} = \lim_{k\to\infty} \left(e^{\mathbf{A}/2k} e^{\mathbf{B}/k} e^{\mathbf{A}/2k}\right)^k$.

(b) Deduce that for Hermitian $\mathbf{A}, \mathbf{B}$ one also has $\operatorname{tr} e^{\mathbf{A}+\mathbf{B}} \leq \operatorname{tr}\!\left(e^{\mathbf{A}/2} e^{\mathbf{B}} e^{\mathbf{A}/2}\right)$, and explain why this is not a new inequality.

Hint: for (a), rerun Steps 2–4 of the proof of Lemma 21.12.1 with the new $\mathbf{T}_k$*; only the constant* M *changes.*

Solution.

(a) Keep $\mathbf{S}_k = e^{(\mathbf{A}+\mathbf{B})/k}$ and set $\mathbf{T}_k = e^{\mathbf{A}/2k} e^{\mathbf{B}/k} e^{\mathbf{A}/2k}$. Step 1 of the proof of Lemma 21.12.1 is unchanged: $\mathbf{S}_k^k = e^{\mathbf{A}+\mathbf{B}}$. For Step 2, submultiplicativity and Theorem 16.6.1 (a) give

$\|\mathbf{T}_k\|_2 \le e^{a/2k}e^{b/k}e^{a/2k} = e^{c/k}$, the same bound as before, with $a = \|\mathbf{A}\|_2$, $b = \|\mathbf{B}\|_2$, $c = a + b$. For Step 3, write $e^{\mathbf{X}} = \mathbf{I} + \mathbf{X} + \mathbf{R}(\mathbf{X})$ as there and multiply out the three factors. Choosing the $\mathbf{I}$ from all three gives $\mathbf{I}$; choosing exactly one nonidentity term gives $\frac{\mathbf{A}}{2k} + \frac{\mathbf{B}}{k} + \frac{\mathbf{A}}{2k} = \frac{\mathbf{A}+\mathbf{B}}{k}$ together with the three remainders $\mathbf{R}(\frac{\mathbf{A}}{2k})$, $\mathbf{R}(\frac{\mathbf{B}}{k})$, $\mathbf{R}(\frac{\mathbf{A}}{2k})$. Every other term is a product of at least two nonidentity factors. Once $k \ge \max(1, c)$, each $\mathbf{R}$ has norm at most $c^2 e/(2k^2)$ by Theorem 16.6.1 (c), so every nonidentity factor has norm at most K/k with $K = c + c^2 e/2$; a product of j of them therefore has norm at most K^j/k^j, which for $2 \le j \le 3$ and $k \ge 1$ is at most $\max(K^2, K^3)/k^2$. There are 27 terms in all. Subtracting from $\mathbf{S}_k = \mathbf{I} + \frac{\mathbf{A}+\mathbf{B}}{k} + \mathbf{R}(\frac{\mathbf{A}+\mathbf{B}}{k})$, the $\mathbf{I}$ and the linear terms cancel and what is left is a sum of at most 27 terms, each of norm at most a constant over k^2, so $\|\mathbf{S}_k - \mathbf{T}_k\|_2 \le M'/k^2$ for a constant M' depending only on $\mathbf{A}$ and $\mathbf{B}$. Step 4 is verbatim, and yields $\|e^{\mathbf{A}+\mathbf{B}} - \mathbf{T}_k^k\|_2 \le e^c M'/k \to 0$.

(b) By Theorem 0.9.6 (3), $\operatorname{tr}(e^{\mathbf{A}/2}e^{\mathbf{B}}e^{\mathbf{A}/2}) = \operatorname{tr}(e^{\mathbf{A}/2}e^{\mathbf{A}/2}e^{\mathbf{B}}) = \operatorname{tr}(e^{\mathbf{A}}e^{\mathbf{B}})$, the last step by Theorem 10.9.4 (d) since $\mathbf{A}/2$ commutes with itself. So the two right-hand sides are the same number and the inequality is Theorem 21.12.4 restated. What is new is only the form: the middle expression is a positive semidefinite matrix, being $(e^{\mathbf{A}/2})^* e^{\mathbf{B}} e^{\mathbf{A}/2}$ with $e^{\mathbf{A}/2}$ Hermitian (Proposition 13.1.2 (a)), whereas $e^{\mathbf{A}}e^{\mathbf{B}}$ is not Hermitian. The warning after Example (The inequality on a two-by-two family) shows that the Loewner-order strengthening of this form is nevertheless false.

Exercise. C3: When is the relative entropy zero?

Let $\mathbf{A}, \mathbf{B} \in M_n(\mathbb{C})$ be positive definite with $\operatorname{tr}\mathbf{A} = \operatorname{tr}\mathbf{B} = 1$.

(a) Show that equality holds in Lemma 21.12.5 for $f(t) = t\log t$ only when $s = t$.

(b) Deduce that $\operatorname{tr}(\mathbf{A}(\log\mathbf{A} - \log\mathbf{B})) = 0$ forces $\mathbf{A} = \mathbf{B}$.

Hint: for (b), look at which coefficients c_{ij} in the proof of Theorem 21.12.6 can be nonzero.

Solution.

(a) By Lemma 21.12.7, $f(t) = t\log t$ is **strictly** convex on $(0, \infty)$. Let $s \ne t$ in $(0, \infty)$ and let $m = \frac{1}{2}(s + t)$ be their midpoint, which also lies in $(0, \infty)$. Strict convexity gives

$$f(m) < \tfrac{1}{2}f(t) + \tfrac{1}{2}f(s),$$

while Lemma 21.12.5 applied at the point t with m in place of s gives

$$f(m) \;\ge\; f(t) + f'(t)(m - t) = f(t) + \tfrac{1}{2}f'(t)(s - t).$$

Combining and multiplying by 2, $2f(t) + f'(t)(s - t) < f(t) + f(s)$, that is $f(s) > f(t) + f'(t)(s - t)$. So the inequality is strict whenever $s \ne t$.

(b) In the proof of Theorem 21.12.6 with $f(t) = t\log t$, the total is $\sum_{i,j} c_{ij}\big[f(\lambda_i) - f(\mu_j) - f'(\mu_j)(\lambda_i - \mu_j)\big]$, with all $c_{ij} \ge 0$ and every bracket ≥ 0. The computation in Corollary 21.12.8 shows that this total equals $\operatorname{tr}(\mathbf{A}(\log\mathbf{A} - \log\mathbf{B}))$ when $\operatorname{tr}\mathbf{A} = \operatorname{tr}\mathbf{B}$. If it is 0, then $c_{ij} = 0$ whenever the bracket is nonzero, that is, by (a), whenever $\lambda_i \ne \mu_j$.

So $\langle \mathbf{v}_j, \mathbf{u}_i\rangle = 0$ whenever $\lambda_i \ne \mu_j$. Fix j and expand $\mathbf{v}_j = \sum_i \langle \mathbf{v}_j, \mathbf{u}_i\rangle \mathbf{u}_i$: only indices i with $\lambda_i = \mu_j$ contribute, so $\mathbf{v}_j$ lies in the eigenspace of $\mathbf{A}$ for the eigenvalue μ_j, giving $\mathbf{A}\mathbf{v}_j = \mu_j\mathbf{v}_j = \mathbf{B}\mathbf{v}_j$. This holds for every j, and the $\mathbf{v}_j$ form a basis, so $\mathbf{A} = \mathbf{B}$.

Chapter 22

Affine and Projective Geometry

Twenty chapters have studied the vector space, and a vector space has an origin. Geometry does not. Nothing a geometer says — these three points are collinear, these two lines are parallel, this point is the midpoint of that segment — mentions a zero vector, and the sets that geometry actually hands us are usually not subspaces at all. Chapter 3's solution set of $\mathbf{Ax} = \mathbf{b}$ with $\mathbf{b} \neq \mathbf{0}$ is the first example in the book: it misses $\mathbf{0}$, the sum of two of its points is not one of its points, and yet it is a perfectly good flat piece of space. Something is being described there that the word "subspace" does not reach.

There are two classical repairs, and this chapter is both of them. **Affine geometry** deletes the origin and keeps what survives: points, differences of points, flats, parallelism, ratios along a line, and the maps that send lines to lines. **Projective geometry** goes the other way and *adds* points — one for every direction — so that two lines in a plane always meet, the dimension formula for a join loses its second case, every quadric in a plane is one of a very short list, and interchanging "point" with "hyperplane" turns every theorem into another theorem.

Everything here is proved with linear algebra, and mostly with the linear algebra of Chapters 1 to 5. An affine space is a vector space acting freely and transitively on a set; a projective space is the set of lines through the origin of a vector space; a projectivity is an invertible linear map read on those lines; a conic is a quadratic form; a line in space is a wedge of two vectors. The geometry is the payoff, not the method.

22.0.1 Conventions

- **Points are not bold.** P, Q, R are points of an affine space $\mathcal{A}$, and $\overrightarrow{PQ} \in V$ is the vector from P to Q; vectors follow the book's rule by type, bold in coordinates, $\mathbf{x} \in F^n$, and plain in an abstract space, $v \in V$. We write $\mathbb{A}^n$ for affine n-space, the set F^n with its origin deliberately forgotten.
- **A flat is an affine subspace**, and $\vec{A}$ is its direction space, a subspace of V. The join $A \vee B$ is the smallest flat containing both, the affine counterpart of $U + W$.
- $\mathbb{P}(V)$ **is the projective space of** V, of dimension $\dim V - 1$, with $\mathbb{P}^n(F) = \mathbb{P}(F^{n+1})$; $[\mathbf{x}] = [x_0 : x_1 : \cdots : x_n]$ is the point determined by $\mathbf{x} \neq \mathbf{0}$. The empty projective space has dimension -1, and that convention is used, not merely tolerated. The projective linear group is $\mathrm{PGL}(V)$.
- **The cross-ratio** $(A, B; C, D)$ **is a point of** $\mathbb{P}^1(F)$, not a scalar. Over a field where the affine chart applies it is the familiar number, and Section 7 gives the formula; keeping it projective is what lets it take the value $[1 : 0]$ without apology.
- V^* **is Chapter 5's dual space** and $\mathbb{P}(V^*)$ the dual projective space, whose points are

the hyperplanes of $\mathbb{P}(V)$. Section 8 writes S^0 for the subspace of $\mathbb{P}(V^*)$ dual to S, built from the annihilator.

- $Z(q)$ **is the zero set of a quadratic form**, and "conic" means $Z(q)$ in a projective plane. Section 5's affine quadric is instead the zero set of a degree-two polynomial $\mathbf{x}^\top\mathbf{A}\mathbf{x}+2\mathbf{b}^\top\mathbf{x}+c$, and carries the $(n+1)\times(n+1)$ matrix $\widetilde{\mathbf{A}}$ alongside the small matrix $\mathbf{A}$.

22.0.1.1 Which field, and where the characteristic bites

Sections 1, 2, 6, 7, 8 and 9 work over an **arbitrary field** F, with no hypothesis on the characteristic at all. That includes the chapter's finite-field computations: the four-point affine plane over $\mathbb{F}_2$ in Section 1, and the count of $\mathbb{P}^n(\mathbb{F}_q)$ with the Fano plane in Section 6.

Characteristic 2 is barred exactly where quadratic forms appear, which is Section 5 throughout and Section 10 up to the Plücker material. The reason is Chapter 14's: over $\mathbb{F}_2$ a quadratic form is not determined by a symmetric matrix, and Section 5 opens by showing that there is then nothing left to classify — every $\mathbf{x}^\top\mathbf{A}\mathbf{x}$ with $\mathbf{A}$ symmetric collapses to the square of a linear form. Section 5 names the hypothesis at each of the three places it does work, and Section 10 states it at the head and lifts it explicitly for the Plücker subsection, which needs no assumption on F. Two smaller restrictions are local and flagged where they occur: Section 2's centroid needs $3\neq 0$ as well as $2\neq 0$, and over $\mathbb{F}_3$ the three medians really are parallel; Section 7's negative result about $n+1$ points needs F to have at least three elements, and Section 7 proves that over $\mathbb{F}_2$ the negative result is false.

Three sections are metric, and they are over $\mathbb{R}$ with the standard inner product: Section 4 entirely, the second half of Section 5, and Section 3, which inherits Chapter 18's setting of $\mathbb{R}^n$ with the dot product. Everywhere else a distance would be meaningless: a projectivity does not preserve one.

22.0.2 What the chapter quotes, and what it does not prove

The chapter imports no analysis. It develops none of its own — no limits, no continuity, no compactness argument — and it adds nothing to the list of analytic facts **(A1)–(A6)** that Chapter 16 introduced. It uses that list exactly once, at one remove: Section 3's proof that a bounded polyhedron is the hull of its vertices cites Chapter 16's Corollary 16.2.5 to supply the compactness hypothesis of Chapter 18's Theorem 18.8.5. That is the whole of it. Section 3 says so at its head, and uses the words "closed", "bounded", and "compact" only inside the results it cites.

Topology is likewise absent. $\mathbb{P}^2(\mathbb{R})$ is a famous topological object, and a reader who has met it will expect the words "compact" and "nonorientable"; Section 6 says plainly that this chapter does not use them, because every argument here is a statement about subspaces of a vector space, valid over any field, including finite ones where topology has nothing to say.

Six things are quoted and not proved. Each is flagged again where it appears, and **nothing in the chapter rests on any of them.**

- **Section 9** — the coordinatization theory of axiomatic projective planes: that Desargues's theorem holds in a plane exactly when the plane comes from a division ring, that Pappus's theorem holds exactly when that ring is commutative, that Pappus therefore implies Desargues, and that non-Desarguesian planes exist. This book has built neither division rings nor synthetic planes, and "projective plane" has meant $\mathbb{P}(V)$ throughout. Section 9 states these with credit in one remark and proves Desargues and Pappus itself, over an arbitrary field, by linear algebra.

- **Section 8** — the duality principle is claimed as a **working method, not a meta-theorem**. What Section 8 proves is a list of five equivalences about the correspondence $S \mapsto S^0$; to turn "every incidence statement has a dual" into a theorem one has to define a formal language and induct on it, which this book has not set up. So every dual statement the chapter uses is proved individually, and Section 8 says this in as many words.
- **Section 4** — the classification of motions is proved in full for $\mathbb{R}^2$ and $\mathbb{R}^3$, both the direct and the opposite cases. The general shape in $\mathbb{R}^n$ — a translation along a flat together with mutually orthogonal rotation planes — is described in a warning and is **not** proved; the warning says so, and gives $\mathbf{R}_{\pi/2} \oplus \mathbf{R}_{\pi/3}$ in $\mathbb{R}^4$ as the witness that the three-dimensional list does not extend.
- **Section 3** — that a dual variable is the *exact* rate at which the optimal value gains when a constraint is loosened. Section 3 proves the inequality, that y_i **bounds** the gain, and then exhibits a two-row program where the exact reading fails. It also declines to develop the finite description of an unbounded polyhedron as a polytope plus a cone, pointing at Chapter 18's exercise instead.
- **Section 10** — *which* quadric the Klein quadric is. The main text records that it is smooth, with inertia $(3, 3, 0)$ over $\mathbb{R}$; the Gram matrix that shows this is computed in Exercise (C2: The Klein quadric), and nothing in the section rests on the answer.
- **Section 10** — which proportionality factors turn the parametrization of the circle into *primitive* Pythagorean triples. That is arithmetic about parities and common factors, not linear algebra, and Section 10 says so where it stops.

Two exercises from earlier chapters are cited, both **for credit only**: Chapter 15 §10's Exercise (C1: The Plücker relation), which reaches the Plücker relation by the $\omega \wedge \omega$ route, and Chapter 18 §09's Exercise (C1: Unbounded polyhedra). Section 10 proves the Plücker relation in both directions from results in the text, and its proof needs no hypothesis on the characteristic, which the exercise's route does. No theorem here rests on an exercise.

One debt is paid. Chapter 15 §08 hinted at Plücker coordinates without naming the chapter that would build them; Section 10 does.

22.0.3 What you need

The chapter leans hard on Chapters 0 to 5 and on Chapters 12, 14, and 18, and lightly on everything else. **Chapters 9, 10, 13, 19, 20, and 21 are not used at all.** Eigenvalues appear only through Chapter 12's spectral theorem and Chapter 14's inertia; no canonical form of an operator, no positive semidefinite matrix, no singular value, and no matrix inequality is needed anywhere.

- **Chapter 0** — sets, maps, and the algebra the constructions rest on: functions and their inverses for Section 4, equivalence relations and quotient sets for Section 6's construction of $\mathbb{P}(V)$, the characteristic of a field for Section 5, and the group axioms, which the affine, Euclidean, and projective groups each verify: Theorem 0.4.7, Theorem 0.4.2, Definition 0.5.2, Definition 0.5.5, Theorem 0.5.2, Definition 0.7.2, Theorem 0.9.3, Theorem 0.9.2, Definition 0.10.1, Definition 0.10.2, Definition 0.10.3, Example (Groups), Theorem 0.10.1, Theorem 0.10.3.
- **Chapter 1** — the chapter's constant companion, and above all the dimension formula for a sum of subspaces, which is what both the affine join formula and the projective dimension formula reduce to: Theorem 1.2.1, Theorem 1.2.2, Theorem 1.3.1, Definition 1.4.1, Lemma 1.5.3, Theorem 1.5.1, Theorem 1.6.5, Theorem 1.6.8, Theorem 1.6.4, Theorem 1.6.3, Theorem 1.7.1, Theorem 1.7.4, Theorem 1.7.3, Theorem 1.8.2.
- **Chapter 3** — the object that opens Section 1, the solution set of a system with $\mathbf{b} \neq \mathbf{0}$; the

rank facts Sections 2, 3 and 10 count with; and the observation that a homogeneous system with more unknowns than equations has a nontrivial solution, which is the whole existence half of "five points determine a conic": Theorem 3.1.1, Theorem 3.1.3, Corollary 3.2.5, Theorem 3.4.4, Theorem 3.5.7, Theorem 3.5.8, Theorem 3.5.5, Lemma 3.7.2.

- **Chapter 2** — linear maps, which are the linear parts of affine maps and the representatives of projectivities: Example (Translation), Theorem 2.1.6, Theorem 2.1.5, Theorem 2.2.1, Theorem 2.2.2, Theorem 2.2.3, Theorem 2.2.5, Theorem 2.2.4, Theorem 2.3.1, Corollary 2.3.2, Theorem 2.4.2, Theorem 2.5.1, Theorem 2.5.7, Theorem 2.5.3, Theorem 2.5.4, Corollary 2.5.5.
- **Chapter 4** — the rank of a map read off its matrix, and the cosets that flats turn out to be: Theorem 4.1.5, Definition 4.4.2, Lemma 4.4.3, Theorem 4.4.5, Theorem 4.5.1.
- **Chapter 5** — the whole of Section 8 and half of Sections 2 and 6. The annihilator is what makes projective duality a bijection, and the hyperplane-as-kernel-of-a-functional is what makes a flat a solution set: Definition 5.1.2, Theorem 5.1.3, Theorem 5.1.1, Proposition 5.1.2, Corollary 5.1.4, Definition 5.2.1, Theorem 5.2.3, Theorem 5.2.2, Corollary 5.3.4, Corollary 5.3.5, Theorem 5.3.2, Definition 5.4.1, Theorem 5.4.1, Definition 5.5.2, Theorem 5.5.3, Theorem 5.6.2, Lemma 5.6.1, Corollary 5.6.6, Proposition 5.6.4, Example (Equations for a plane in $\mathbb{R}^4$).
- **Chapter 6 §01** — one fact: a nonzero polynomial of degree at most d over a field has at most d roots, which is how an exercise of Section 5 knows that a line meeting a quadric in three points lies inside it. Section 10 uses Chapter 3 §07's version of the same bound for the same purpose, and that one is in the text: Corollary 6.1.4.
- **Chapter 7** — determinants, for the direct-or-opposite dichotomy of Section 4 and for the coordinate computations; the Vandermonde determinant is how Section 7 certifies a list of points in general position: Corollary 7.3.3, Theorem 7.4.5, Theorem 7.5.4, Theorem 7.5.2, Theorem 7.8.1, Corollary 7.8.2, Theorem 7.10.3.
- **Chapter 8 §01** — block multiplication, which is what the $(n+1) \times (n+1)$ matrix of an affine map is for: Theorem 8.1.1.
- **Chapter 11** — the Euclidean sections' toolkit: the inner product and Cauchy–Schwarz, orthogonal decomposition, the characterizations of an isometry, and the polarization identity that turns distance-preservation into linearity: Theorem 11.1.5, Theorem 11.1.2, Theorem 11.2.1, Corollary 11.2.3, Theorem 11.3.2, Theorem 11.6.5, Lemma 11.7.2, Theorem 11.7.3, Proposition 11.7.4, Theorem 11.7.6, Definition 11.8.1, Proposition 11.8.2.
- **Chapter 12** — the real spectral theorem, which rotates a quadric onto its principal axes, and the canonical form of an orthogonal matrix, which is the entire engine of Section 4's two classification theorems: Corollary 12.5.2, Theorem 12.5.5, Example (Finding the axes of a conic), Theorem 12.6.1, Corollary 12.6.2, Theorem 12.6.4, Theorem 12.6.5, Theorem 12.6.7.
- **Chapter 14** — the classification of quadratic forms, which Sections 5 and 10 turn into the classification of quadrics. Congruence is what an affine change of coordinates does to the big matrix, and Sylvester's law is what separates an ellipse from an empty set: Definition 14.1.2, Proposition 14.1.2, Definition 14.2.1, Proposition 14.2.2, Proposition 14.2.4, Theorem 14.2.3, Definition 14.3.2, Theorem 14.3.5, Theorem 14.3.3, Proposition 14.3.2, Theorem 14.4.2, Proposition 14.4.3, Corollary 14.4.4, Theorem 14.5.1, Corollary 14.5.2, Definition 14.5.1, Theorem 14.5.3, Definition 14.7.1, Corollary 14.9.3, Example (The Pfaffian in Sizes 2 and 4).
- **Chapter 15 §§08–10** — exterior powers, for Plücker coordinates alone: the basis of $\Lambda^2 F^4$, that a wedge is nonzero exactly when the factors are independent, and that its coordinates are the minors: Lemma 15.8.1, Theorem 15.8.5, Theorem 15.8.7, Corollary 15.8.6, Theorem 15.9.5, Exercise (C1: The Plücker relation) (credited, not used).
- **Chapter 16** — two facts, both in the metric sections: the reverse triangle inequality for Section 4's nonexample, and the compactness of closed bounded sets, which is the chapter's

only contact with analysis: Lemma 16.1.1, Corollary 16.2.5.

- **Chapter 17** — the Rayleigh quotient, which reads Section 5's semi-axes off the eigenvalues, and the inertia triple: Proposition 17.1.1, Definition 17.10.1.
- **Chapter 18** — all of Section 3, and the affine vocabulary of Sections 1, 2 and 5. Convex hulls and affine hulls, separation, extreme points and faces, polytopes and Minkowski–Weyl, and linear programming duality: Definition 18.1.5, Proposition 18.1.4, Theorem 18.1.3, Proposition 18.1.1, Definition 18.1.2, Definition 18.2.1, Lemma 18.2.1, Theorem 18.3.6, Definition 18.6.2, Theorem 18.6.1, Theorem 18.6.3, Corollary 18.6.2, Theorem 18.6.5, Definition 18.8.1, Definition 18.8.2, Proposition 18.8.4, Lemma 18.8.3, Theorem 18.8.5, Corollary 18.8.6, Definition 18.9.2, Proposition 18.9.1, Theorem 18.9.5, Theorem 18.9.2, Corollary 18.9.3, Example (The cube and the cross-polytope), Exercise (C1: Unbounded polyhedra) (pointed at, not used).

22.0.4 Roadmap

- **Affine Spaces and Affine Maps.** A set on which a vector space acts freely and transitively: differences of points are vectors, sums of points are meaningless, and Chasles's relation is the one law. Choosing an origin identifies the space with V, and the identification is canonical only after the choice. Affine combinations are exactly the combinations whose coefficients sum to 1, and no other coefficient sum survives a change of origin. An affine map is a linear map plus a translation, for a unique linear part, and the whole affine group fits in one $(n+1) \times (n+1)$ block matrix — the device the rest of the chapter computes with.
- **Flats, Independence, and Coordinates.** Flats are the cosets, parallelism is an equivalence relation on flats of equal dimension, and two flats either meet or lie in two distinct parallel hyperplanes. The dimension of a join has two cases: the subspace formula when they meet, and a corrected formula with a $+1$ when they miss, in which the intersection of **directions** takes over. No convention for $\dim \emptyset$ merges the two. Then affine frames, barycentric coordinates, and an affine map pinned down by its values on an affine frame of $n+1$ points; on any set lying inside a proper flat — in particular on any n points — two distinct affine maps agree. The medians of a triangle meet at the centroid, in one line of barycentric coordinates — and over $\mathbb{F}_3$ they do not meet at all.
- **Convex Polyhedra, Geometrically.** A reading of Chapter 18, not a reproof: it says so in its first paragraph. A face is what a supporting hyperplane cuts out; faces are polyhedra and there are finitely many; vertex, extreme point, and basic feasible solution are three names for one thing. The face lattices of a cube and of a simplex are computed in full. A linear program is a family of parallel hyperplanes sweeping across a solid, the optimal set is a whole face, and the dual variables are the nonnegative weights expressing the objective as a combination of the active outer normals. Negative cases: an unbounded slab with no vertex at all, whose optimum is nonetheless attained on a face that contains none, and a quadrant that has a vertex and no optimum. No algorithm appears here or anywhere in this book; Chapter 24 is where algorithms live.
- **Euclidean Motions.** Distance alone forces everything: a map of $\mathbb{R}^n$ preserving distances is affine with orthogonal linear part, and **surjectivity is not assumed** — it is proved. Polarization is the move, and the Euclidean norm is doing real work, as a distance-preserving nonaffine map for the maximum norm shows. Motions form a group with the translations normal in it and the orthogonal group as the quotient's content, recorded without quotient groups, which this book has not built. Then both classifications in full: in the plane, translation, rotation, reflection, glide reflection; in space, translation, rotation, screw, reflection, glide, rotatory reflection, told apart by the fixed set. From $\mathbb{R}^4$ on the list breaks, and the section says so.

- **Affine and Euclidean Quadrics.** A quadric is a quadratic form that has forgotten where the origin is. An affine substitution acts on the $(n+1)\times(n+1)$ matrix by congruence, a center exists exactly when the linear vector lies in the column space of the small matrix, and the parabola is precisely the centerless case. Over $\mathbb{R}$ every quadric is affinely equivalent to exactly one of three families, separated by the pair of inertias; for $n=2$ that is nine normal forms, two of which have empty zero sets and are still inequivalent, and for $n=3$ it is seventeen. Allowing only motions instead keeps the eigenvalues, and they are the reciprocal squared semi-axes — the Rayleigh quotient read on an ellipsoid.
- **Projective Space.** Stop working with the points of a vector space and start working with its lines through the origin. The affine chart splits $\mathbb{P}^n(F)$ into a copy of $\mathbb{A}^n$ and a hyperplane at infinity of directions, and the two parallel lines of the opening now meet, at a point whose coordinates are computed. The dimension formula $\dim(S\vee T)+\dim(S\cap T)=\dim S+\dim T$ has **no exceptional case**, which is the whole reason for the construction; two distinct lines of a projective plane meet in exactly one point. Finite fields give $|\mathbb{P}^n(\mathbb{F}_q)|=(q^{n+1}-1)/(q-1)$, and the Fano plane is drawn and checked.
- **Projective Transformations.** An invertible linear map descends to $\mathbb{P}(V)$, and two of them give the same projectivity exactly when they differ by a nonzero scalar — so $\mathrm{PGL}(V)$ is $\mathrm{GL}(V)$ modulo scalars, said without quotient groups. The fundamental theorem: $n+2$ points in general position can be carried to any other such list by exactly one projectivity. $n+1$ points are not enough, and the explicit failure is a diagonal matrix; over $\mathbb{F}_2$, and only there, $n+1$ points do suffice, which is proved. Then the cross-ratio, invariant under projectivities and a complete invariant of four points on a line, computed exactly; perspectivities; and the sentence that ties the chapter together — affine maps are exactly the projectivities fixing the hyperplane at infinity as a set.
- **Duality.** A point of $\mathbb{P}(V^*)$ is a hyperplane of $\mathbb{P}(V)$. The annihilator gives an inclusion-reversing bijection $S\mapsto S^0$ between the projective subspaces of the two spaces, with $\dim S^0=n-1-\dim S$, exchanging joins with meets and the two extreme subspaces with each other. Dualizing twice returns the original. The duality principle is then stated as five equivalences and used as a recipe, each application being its own proof; "two distinct points lie on a unique line" dualizes to "two distinct lines meet in a unique point", and the projective dimension formula turns out to be self-dual. The trap is spelled out: the dual of "point" is "hyperplane", and in $\mathbb{P}^3$ two lines need not meet.
- **Desargues and Pappus.** Two triangles are perspective from a point if and only if they are perspective from a line, proved inside $\mathbb{P}^2$ by choosing representatives so that the hypothesis reads $a'=a+o$, with every scaling step verified — that is where such proofs go wrong. The converse costs no second theorem: the same lemma, applied to a pair of auxiliary triangles built out of the same picture. Pappus follows in coordinates, with the frame chosen by the previous section's fundamental theorem to make the computation short, and the degenerate configurations excluded by hypothesis. Both hold over every field, and the remark that closes the section records, with credit and without proof, that Desargues means "division ring" and Pappus means "commutative".
- **Conics and Lines in Space.** Over an algebraically closed field a projective quadric is classified by rank alone; over $\mathbb{R}$ by rank and inertia up to sign, giving five conics in $\mathbb{P}^2(\mathbb{R})$: empty, a point, a line, two lines, a smooth conic. Five points in general position lie on exactly one conic, and the independence of the five conditions is proved, not assumed. A smooth conic with one point is in bijection with $\mathbb{P}^1$ by projection from that point, which turns one Pythagorean triple into all of them. Pole, polar, and the dual conic use Section 8's duality. Finally Plücker coordinates, over any field: a line of $\mathbb{P}^3$ is a point of $\mathbb{P}^5$, the six coordinates satisfy $p_{01}p_{23}-p_{02}p_{13}+p_{03}p_{12}=0$, and **every** point of that Klein quadric is a line — both directions proved. A summary closes the chapter.

22.0.5 Named moves

- **Choose an origin, then say that you chose it.** An affine space becomes a vector space the moment a point is picked, and every affine statement is then a linear statement about the difference vectors. The discipline is to record which conclusions survive a different choice: those are the affine ones, and the coefficient sum 1 is exactly the test.
- **Add a coordinate to linearize.** An affine map of F^n is not linear, but $\mathbf{x} \mapsto (1, \mathbf{x})$ — Section 1 writes the extra coordinate last instead — makes it the restriction of a linear map of F^{n+1}, with matrix $\left(\begin{smallmatrix} 1 & \mathbf{0}^\top \\ \mathbf{b} & \mathbf{M} \end{smallmatrix}\right)$. The same extra coordinate is what turns a quadric polynomial into a quadratic form and an affine space into a chart of a projective space. Three of this chapter's theorems are that one observation in three costumes.
- **Pass from points to lines through the origin.** Scaling is the enemy of a well-defined point and the friend of a well-defined condition: $q(\lambda x) = \lambda^2 q(x)$ vanishes when $q(x)$ does. Anything homogeneous descends to $\mathbb{P}(V)$, which is why every object in the second half of the chapter is the solution set of homogeneous polynomial equations.
- **Choose the representatives so the hypothesis becomes an equation.** A projective point is a whole line of vectors, and the content of most projective proofs is the scaling chosen at the start: $a' = a + o$ in Desargues, a normalized frame in the fundamental theorem, the second intersection point in the conic parametrization. State out loud which scalings remain free afterwards.
- **Split the translation and ask what $\mathbf{I}-\mathbf{Q}$ can absorb.** To classify a map $\mathbf{x} \mapsto \mathbf{Q}\mathbf{x}+\mathbf{b}$, decompose $\mathbf{b}$ along the kernel of $\mathbf{I} - \mathbf{Q}$ and its complement. The part in the image is absorbed by moving the fixed set; the part in the kernel survives and is the glide. That single question produces both of Section 4's classification theorems.
- **Read an old theorem geometrically instead of proving a new one.** Section 3 adds no convexity theorem to Chapter 18. It renames the objects — face, vertex, edge, facet — and the theorems become statements about a solid with flat sides. Knowing when a chapter is a translation rather than a construction is worth as much as the construction.
- **Count the conditions, then prove they are independent.** Five points impose five linear conditions on a 6-dimensional space of forms, so a conic exists by counting alone. Uniqueness is the real theorem, and it is where general position earns its keep: a second independent solution would let us push a conic through a sixth point on one of the lines, a conic meeting a line three times contains it, so it splits into two lines, and the second line then carries three of the original five. The count is cheap; the independence is the content.
- **Dualize by the dimension formula, never by the words.** The dual of a d-dimensional subspace has dimension $n-1-d$. Compute that number before translating any statement: in a plane it sends points to lines, but in $\mathbb{P}^3$ it sends points to planes and lines to lines, and translating "point" to "line" there produces a false theorem.

22.1 Affine Spaces and Affine Maps

Every space in this book so far has had a distinguished point: the zero vector. Geometry does not. A plane drawn on paper has no privileged spot, and none of the statements a geometer cares about — three points are collinear, two lines are parallel, this point is the midpoint of that segment — mentions one. This section builds the object that has all the structure of a vector space except the origin, and the maps between such objects. Everything in the rest of the chapter is written in this language.

22.1.1 A set with no zero in it

Chapter 3 already produced such an object without naming it. Let $\mathbf{A} \in M_{m\times n}(F)$ and $\mathbf{b} \in F^m$, and suppose $\mathbf{Ax} = \mathbf{b}$ is consistent. By Theorem 3.1.3 the solution set is $\mathbf{p}+N$, where $N = \mathrm{nul}(\mathbf{A})$ and $\mathbf{p}$ is any one solution. When $\mathbf{b} \neq \mathbf{0}$ this set is **not** a subspace: it misses $\mathbf{0}$, and the sum of two solutions is not a solution. Yet it is not shapeless either. The difference of two solutions is a vector of N, and adding a vector of N to a solution gives a solution. So the solution set carries an action of N, and N records exactly how any two solutions differ.

That is the whole structure, and it deserves a name. The only operation is: *subtract two points and get a vector; add a vector to a point and get a point.* Adding two points is meaningless, and so is multiplying a point by a scalar.

An affine space is a set of points in which differences of points are vectors, and every vector moves every point to exactly one other.

Definition 22.1.1. Affine Space

Let V be a vector space over a field F. An **affine space with direction space** V is a **nonempty** set $\mathcal{A}$, whose elements are called **points**, together with a map

$$\mathcal{A} \times \mathcal{A} \to V, \qquad (P, Q) \mapsto \overrightarrow{PQ},$$

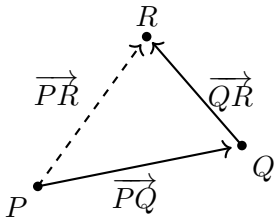
such that:

(AS1) (**Chasles's relation**) $\overrightarrow{PQ} + \overrightarrow{QR} = \overrightarrow{PR}$ **for all** $P, Q, R \in \mathcal{A}$;

(AS2) **for every** $P \in \mathcal{A}$, the map $\mathcal{A} \to V$, $Q \mapsto \overrightarrow{PQ}$, is a **bijection**.

The **dimension** of $\mathcal{A}$ is $\dim \mathcal{A} := \dim V$. For $P \in \mathcal{A}$ and $v \in V$ we write $P + v$ for the unique point Q with $\overrightarrow{PQ} = v$, which exists and is unique by (AS2).

Read the two clauses. (AS1) says that traveling from P to Q and then from Q to R is traveling from P to R; it is the one law that makes the arrows behave like displacements. (AS2) says that from any fixed point P, every vector of V is the displacement to exactly one point: **existence** makes $\mathcal{A}$ big enough (the action is *transitive*), **uniqueness** makes it no bigger (the action is *free*). The set $\mathcal{A}$ is not a vector space and has no zero; V is a vector space and has one.



Two consequences are immediate and will be used constantly.

Lemma 22.1.1. First consequences of Chasles

Let $\mathcal{A}$ be an affine space with direction space V. For all $P, Q \in \mathcal{A}$ and $u, w \in V$:

(a) $\overrightarrow{PP} = 0$, and $\overrightarrow{QP} = -\overrightarrow{PQ}$;

(b) $P + 0 = P$ and $(P + u) + w = P + (u + w)$;

(c) $\overrightarrow{PQ} = 0$ if and only if $P = Q$.

Proof.

(a) Taking $P = Q = R$ in (AS1) gives $\overrightarrow{PP} + \overrightarrow{PP} = \overrightarrow{PP}$, so $\overrightarrow{PP} = 0$ by cancellation in the group $(V, +)$ (Theorem 0.10.1). Taking $R = P$ in (AS1) gives $\overrightarrow{PQ} + \overrightarrow{QP} = \overrightarrow{PP} = 0$.

(b) $P + 0 = P$ restates $\overrightarrow{PP} = 0$. For the second identity, put $Q = P + u$ and $R = Q + w$, so that $\overrightarrow{PQ} = u$ and $\overrightarrow{QR} = w$. By (AS1), $\overrightarrow{PR} = u + w$, which says $R = P + (u + w)$.

(c) ($\Leftarrow$) is (a). ($\Rightarrow$) If $\overrightarrow{PQ} = 0 = \overrightarrow{PP}$, then $Q = P$ by the injectivity in (AS2).

■

The clauses in part (b) are the axioms of an action of the additive group of V on $\mathcal{A}$, and the definition is often given that way instead. The two presentations carry the same information.

Proposition 22.1.2. The two forms of the definition agree

Let V be a vector space over F and $\mathcal{A}$ a nonempty set. Giving a map $\mathcal{A} \times \mathcal{A} \to V$ satisfying (AS1) and (AS2) is the same as giving a map $\mathcal{A} \times V \to \mathcal{A}$, written $(P, v) \mapsto P + v$, such that

(AC1) $P + 0 = P$ for every $P \in \mathcal{A}$;

(AC2) $(P + u) + w = P + (u + w)$ for all $P \in \mathcal{A}$ and $u, w \in V$;

(AC3) for all $P, Q \in \mathcal{A}$ there is **exactly one** $v \in V$ with $P + v = Q$.

The two constructions are inverse to each other.

Idea. There is nothing to find: in one direction the addition was already defined after Definition 22.1.1, in the other the vector $\overrightarrow{PQ}$ is the one whose uniqueness (AC3) asserts. The work is checking the axioms, and each check is one line.

Proof. Suppose first that $(P, Q) \mapsto \overrightarrow{PQ}$ satisfies (AS1) and (AS2), and define $P + v$ as after Definition 22.1.1. Then (AC1) and (AC2) are Lemma 22.1.1 (b), and (AC3) is (AS2), since $P + v = Q$ says $\overrightarrow{PQ} = v$.

Conversely, suppose $(P, v) \mapsto P + v$ satisfies (AC1)–(AC3), and let $\overrightarrow{PQ}$ be the unique v with $P + v = Q$. For (AS1), put $u = \overrightarrow{PQ}$ and $w = \overrightarrow{QR}$; then by (AC2),

$$P + (u + w) = (P + u) + w = Q + w = R,$$

so $\overrightarrow{PR} = u + w$ by the uniqueness in (AC3). For (AS2), fix P. The map $Q \mapsto \overrightarrow{PQ}$ is surjective because v is the image of $P + v$, and injective because $\overrightarrow{PQ} = \overrightarrow{PQ'} = v$ gives $Q = P + v = Q'$.

The two passages are inverse: starting from $\overrightarrow{\cdot\,\cdot}$, the constructed addition has $P + v = Q$ exactly when $\overrightarrow{PQ} = v$, so the difference map it returns is the original one; and starting from an addition, the same equivalence returns it unchanged. This proves the proposition.

■

Remark. The converse half used only (AC2) and (AC3), and it could not have used more: (AC1) follows from the other two. Indeed (AC3) with $Q = P$ gives exactly one v with $P + v = P$, and that v has $P + (v + v) = (P + v) + v = P + v = P$ by (AC2), so $v + v = v$ by the same uniqueness and $v = 0$. We list (AC1) anyway, because that is how the axioms of a group action are usually written.

Now the examples. The first is the one every other example is modeled on.

Example. Four affine spaces

Check each of the following against Definition 22.1.1, and say what its direction space and dimension are.

(a) A vector space V over F, with $\overrightarrow{xy} := y - x$.

(b) A coset $p + U$ of a subspace U of V (Definition 4.4.2), with the same difference map.

(c) The solution set of a **consistent** system $\mathbf{Ax} = \mathbf{b}$, $\mathbf{A} \in M_{m\times n}(F)$.

(d) The set $\mathbb{F}_2^2$ of four points, with $\overrightarrow{\mathbf{xy}} = \mathbf{y} - \mathbf{x}$.

Solution.

(a) Chasles reads $(y - x) + (z - y) = z - x$, which is true in any vector space. For fixed x, the map $y \mapsto y - x$ is a bijection $V \to V$ with inverse $v \mapsto x + v$. Direction space V; dimension $\dim V$. Here the point x and the vector x are the same object, and only the language distinguishes them.

(b) The set $p + U$ is nonempty (it contains p) and the difference of two of its elements is $(p + u) - (p + w) = u - w \in U$, so the map does land in U. Chasles is again the identity in (a). For fixed $x \in p + U$, the map $y \mapsto y - x$ is a bijection $p + U \to U$: it is injective, and $u \in U$ is the image of $x + u$, which lies in $p + U$ because U is closed under addition. Direction space U; dimension $\dim U$. Over $\mathbb{R}$ these are exactly Chapter 18's affine subspaces of V (Definition 18.1.5 (b)), with the same dimension.

(c) By Theorem 3.1.3 the solution set is $\mathbf{p} + \operatorname{nul}(\mathbf{A})$ for any one solution $\mathbf{p}$, so this is a special case of (b). Direction space $\operatorname{nul}(\mathbf{A})$; dimension $\operatorname{nullity}(\mathbf{A}) = n - \operatorname{rank}\mathbf{A}$ by Theorem 3.5.7. Consistency is what makes the set nonempty, and Theorem 3.5.8 says when it holds.

(d) A special case of (a) with $V = \mathbb{F}_2^2$: four points, direction space $\mathbb{F}_2^2$, dimension 2. It is worth seeing because it is so small. Every pair of distinct points $P \neq Q$ has $\overrightarrow{PQ} = \overrightarrow{QP}$, since $-1 = 1$ in $\mathbb{F}_2$; the next section will call a coset of a one-dimensional subspace a **line**, and here each such coset has exactly two points. So every one of the $\binom{4}{2} = 6$ pairs of distinct points is itself a line, and the six lines fall into three pairs with no point in common.

A nonexample, by minimal change. Delete one point from the plane: let $\mathcal{A} = \mathbb{R}^2 \setminus \{\mathbf{0}\}$, with $\overrightarrow{\mathbf{xy}} = \mathbf{y} - \mathbf{x}$. Chasles still holds, because it is an identity about subtraction. But (AS2) fails: for $P = (1, 0)$, no point Q of $\mathcal{A}$ has $\overrightarrow{PQ} = (-1, 0)$, since the only candidate is $\mathbf{0}$, which we removed. The map $Q \mapsto \overrightarrow{PQ}$ is injective but not surjective. It is exactly the transitivity of the action that fails: the set has a hole.

Warning. **A point is not a vector.** In an affine space there is no zero point; and over a field in which $2 \neq 0$ and $3 \neq 0$ there is no sum $P+Q$ of two points and no multiple $3P$. The only legal expressions are $\overrightarrow{PQ} \in V$, $P+v \in \mathcal{A}$, and the affine combinations defined below. If you find yourself writing $P+Q$, you have silently chosen an origin; the next subsection says what that costs.

22.1.2 Choosing an origin

An affine space looks like its direction space as soon as one point is nominated as the zero. This is the chapter's first named move, and it is used in almost every proof that follows: *to compute, pick an origin; to state a result, do not.*

Proposition 22.1.3. Choice of Origin

Let $\mathcal{A}$ be an affine space with direction space V, and let $O \in \mathcal{A}$. Then

$$\theta_O \colon \mathcal{A} \to V, \qquad \theta_O(P) = \overrightarrow{OP},$$

is a bijection with $\theta_O(O) = 0$, and it carries the affine structure of $\mathcal{A}$ to that of V in Example (Four affine spaces) (a):

$$\overrightarrow{PQ} = \theta_O(Q) - \theta_O(P) \qquad (P, Q \in \mathcal{A}).$$

If $O' \in \mathcal{A}$ is another point, then $\theta_{O'}(P) = \theta_O(P) - \theta_O(O')$ for every P: the two identifications differ by a translation of V.

Proof. θ_O is a bijection by (AS2), and $\theta_O(O) = \overrightarrow{OO} = 0$ by Lemma 22.1.1 (a). By Chasles and the same lemma,

$$\theta_O(Q) - \theta_O(P) = \overrightarrow{OQ} - \overrightarrow{OP} = \overrightarrow{OQ} + \overrightarrow{PO} = \overrightarrow{PQ}.$$

For the last statement, apply this with P and O' in place of Q and P: $\theta_{O'}(P) = \overrightarrow{O'P} = \theta_O(P) - \theta_O(O')$.

■

So an n-dimensional affine space over F is a copy of F^n with the labels rubbed off. Choosing O writes the labels back on, and any two choices differ by a translation. We write $\mathbb{A}^n(F)$ for F^n regarded as an affine space over itself, and $\mathbb{A}^n$ when $F = \mathbb{R}$; its points are written $P = (p_1, \dots, p_n)$, in plain type, even though the same list of numbers written $\mathbf{p}$ would be a vector.

Warning. **The identification is never canonical.** Proposition 22.1.3 does not say that $\mathcal{A}$ "is" V; it says that each point of $\mathcal{A}$ gives one isomorphism, and there are as many as there are points. A statement proved after choosing O is a statement about $\mathcal{A}$ only if it does not depend on which O was chosen. Checking that independence is a real obligation, and the next subsection is nothing but that check, done once for all.

22.1.3 Affine combinations

Sums of points are forbidden, and yet the centroid $\frac{1}{3}(P_1 + P_2 + P_3)$ of a triangle is a genuine point of the plane, independent of any coordinate system. Something with weights adding up to 1 survives the loss of the origin. Here is the computation that says which expressions survive.

Fix $O \in \mathcal{A}$, points $P_1, \dots, P_k \in \mathcal{A}$ and scalars $t_1, \dots, t_k \in F$ with $s := t_1 + \cdots + t_k$. Let $O' \in \mathcal{A}$ and $w = \overrightarrow{OO'}$. Since $\overrightarrow{O'P_i} = \overrightarrow{O'O} + \overrightarrow{OP_i} = \overrightarrow{OP_i} - w$,

$$\sum_{i=1}^{k} t_i \overrightarrow{O'P_i} = \sum_{i=1}^{k} t_i \overrightarrow{OP_i} - sw, \tag{22.1.1}$$

and therefore, adding this vector to the point $O' = O + w$ and using Lemma 22.1.1 (b), we get the companion identity for points.

$$O' + \sum_{i=1}^{k} t_i \overrightarrow{O'P_i} = \Big(O + \sum_{i=1}^{k} t_i \overrightarrow{OP_i}\Big) + (1 - s)w. \tag{22.1.2}$$

The vector on the left of Eq. 22.1.1 does not depend on the origin exactly when $s = 0$; the point on the left of Eq. 22.1.2 does not depend on it exactly when $s = 1$. Both cases are worth a name.

Weights adding to one make a point; weights adding to zero make a vector; nothing else means anything.

Definition 22.1.2. Affine Combination

Let $\mathcal{A}$ be an affine space with direction space V, let $P_1, \dots, P_k \in \mathcal{A}$ with $k \geq 1$, and let $t_1, \dots, t_k \in F$.

(a) If $t_1 + \cdots + t_k = 1$, the **affine combination** $\sum_i t_i P_i$ is the **point** $O + \sum_i t_i \overrightarrow{OP_i}$, for any $O \in \mathcal{A}$.

(b) If $t_1 + \cdots + t_k = 0$, the expression $\sum_i t_i P_i$ denotes the **vector** $\sum_i t_i \overrightarrow{OP_i} \in V$, for any $O \in \mathcal{A}$.

For no other value of $t_1 + \cdots + t_k$ is $\sum_i t_i P_i$ defined.

In words: the same symbol means a point when the weights add up to 1 and a vector when they add up to 0, and the type of the answer is read off from the weights. The simplest instance of (b) is $k = 2$ with weights $(-1, 1)$: the vector $Q - P$ is $\overrightarrow{PQ}$. The simplest instance of (a) is $k = 1$ with $t_1 = 1$, which is P_1.

Proposition 22.1.4. Affine combinations are exactly the origin-free ones

With the notation of Definition 22.1.2, put $s = t_1 + \cdots + t_k$.

(a) If $s = 1$, the point $O + \sum_i t_i \overrightarrow{OP_i}$ is the same for every $O \in \mathcal{A}$.

(b) If $s = 0$, the vector $\sum_i t_i \overrightarrow{OP_i}$ is the same for every $O \in \mathcal{A}$.

(c) Suppose $\dim \mathcal{A} \geq 1$. If $s \neq 1$, there are points $P_1, \dots, P_k$ and origins O, O' for which

$O + \sum_i t_i \overrightarrow{OP_i}$ and $O' + \sum_i t_i \overrightarrow{O'P_i}$ differ; and if $s \neq 0$, the same is true of the vectors $\sum_i t_i \overrightarrow{OP_i}$.

Proof.

(a) and (b) are Eq. 22.1.2 and Eq. 22.1.1 with $1 - s = 0$ and $s = 0$ respectively, since O' was an arbitrary point of $\mathcal{A}$.

(b) Since $\dim \mathcal{A} \geq 1$, the direction space V contains some $w \neq 0$. Take any $O \in \mathcal{A}$, put $O' = O + w$, and take $P_1 = \cdots = P_k = O$, so that $\overrightarrow{OP_i} = 0$ for every i. Then Eq. 22.1.2 reads

$$O' + \sum_i t_i \overrightarrow{O'P_i} = O + (1 - s)w,$$

which differs from $O + \sum_i t_i \overrightarrow{OP_i} = O$ when $(1 - s)w \neq 0$, that is, when $s \neq 1$, by Lemma 22.1.1 (c). Likewise Eq. 22.1.1 reads $\sum_i t_i \overrightarrow{O'P_i} = -sw$, which differs from $\sum_i t_i \overrightarrow{OP_i} = 0$ when $s \neq 0$.

■

Taking $\mathcal{A} = V$ over $\mathbb{R}$ and $O = 0$, part (a) becomes the ordinary vector expression $\sum_i t_i x_i$ with $\sum_i t_i = 1$, which is Chapter 18's affine combination (Definition 18.1.5 (a)). So Definition 22.1.2 extends that notion from a real vector space to an affine space over any field, and adds the companion vector-valued case. Chapter 18's affine hull $\operatorname{aff} S$ and its description as a coset (Proposition 18.1.4) are the real case of the flats of the next section.

Quick check. Let $f : \mathbb{R} \to \mathbb{R}$ be $f(x) = 2x + 3$, which is not linear. Compute $f(3 \cdot 1 - 2 \cdot 2)$ and $3f(1) - 2f(2)$; then compute $f(1 + 2)$ and $f(1) + f(2)$. What do the two pairs show?

Solution. First pair: $3 \cdot 1 - 2 \cdot 2 = -1$ and $f(-1) = 1$, while $3f(1) - 2f(2) = 3 \cdot 5 - 2 \cdot 7 = 1$. They agree, and the weights 3 and -2 add up to 1. Second pair: $f(3) = 9$, while $f(1) + f(2) = 5 + 7 = 12$. They differ, and the weights 1 and 1 add up to 2. So f respects combinations whose weights add up to 1 and destroys the others: it is affine but not linear. That is the content of the next subsection.

22.1.4 Affine maps

Linear maps are the functions that respect the structure of a vector space. Affine maps should be the functions that respect the structure of an affine space, and by Definition 22.1.2 that structure is: which affine combinations equal which points. So there is only one reasonable definition.

An affine map is one that can be applied inside an affine combination.

Definition 22.1.3. Affine Map

Let $\mathcal{A}$ and $\mathcal{B}$ be affine spaces over the same field F, with direction spaces V and W. A map

$f\colon \mathcal{A} \to \mathcal{B}$ is **affine** if

$$f\Big(\sum_{i=1}^{k} t_i P_i\Big) = \sum_{i=1}^{k} t_i f(P_i)$$

for every $k \geq 1$, all $P_1, \dots, P_k \in \mathcal{A}$ and all $t_1, \dots, t_k \in F$ with $t_1 + \cdots + t_k = 1$.

Both sides are points, by Definition 22.1.2 (a) applied in $\mathcal{A}$ and in $\mathcal{B}$. Nothing is asked about weights summing to anything else, and by the Quick check above nothing could be.

The definition quantifies over all k, which is awkward to verify. The next theorem replaces it by a single linear map, and is the result the rest of the chapter uses.

Theorem 22.1.5. An Affine Map Is Linear Plus a Translation

Let $\mathcal{A}, \mathcal{B}$ be affine spaces over F with direction spaces V, W, and let $f\colon \mathcal{A} \to \mathcal{B}$. The following are equivalent.

(a) f is affine.

(b) There is a linear map $T\colon V \to W$ with $\overrightarrow{f(P)f(Q)} = T(\overrightarrow{PQ})$ for all $P, Q \in \mathcal{A}$.

(c) For **some** $O \in \mathcal{A}$ there is a linear map $T\colon V \to W$ with $f(P) = f(O) + T(\overrightarrow{OP})$ for all $P \in \mathcal{A}$.

The map T in (b) and (c) is unique, is the same in both, and does not depend on O. It is called the **linear part** of f and written $\vec{f}$.

Idea. The cycle is (a) $\Rightarrow$ (c) $\Rightarrow$ (b) $\Rightarrow$ (a). For (a) $\Rightarrow$ (c) we must manufacture a linear map out of a condition about weights, so we need the two vector-space axioms written as affine combinations. They are

$$O + (u + w) = (O + u) + (O + w) - O, \qquad O + \lambda u = \lambda(O + u) + (1 - \lambda)O,$$

whose weights add up to $1 - 1 + 1 = 1$ and $\lambda + (1 - \lambda) = 1$. Applying f to each is additivity and homogeneity of T.

Proof.

(a) $\Rightarrow$ (c). Fix $O \in \mathcal{A}$ and define $T\colon V \to W$ by $T(v) = \overrightarrow{f(O)\,f(O+v)}$; this is a map because $O + v$ is a well-defined point. Then $f(O + v) = f(O) + T(v)$ by the definition of $+$ in $\mathcal{B}$, so once T is shown to be linear, taking $v = \overrightarrow{OP}$ gives (c).

Additivity. Let $u, w \in V$. The three points $O+u$, $O+w$, O with weights $1, 1, -1$ have weight sum 1, and their affine combination, computed with origin O, is $O+(u+w-0) = O+(u+w)$. Applying (a) and computing the right-hand side with origin $f(O)$,

$$\begin{aligned} f(O + u + w) &= f(O + u) + f(O + w) - f(O) \\ &= f(O) + \big(T(u) + T(w)\big), \end{aligned}$$

so $T(u + w) = T(u) + T(w)$.

Homogeneity. Let $\lambda \in F$. The two points $O + u$ and O with weights λ and $1 - \lambda$ have weight

sum 1, and their affine combination with origin O is $O+\lambda u$. Applying (a) and computing with origin $f(O)$,

$$f(O+\lambda u)=\lambda f(O+u)+(1-\lambda)f(O)=f(O)+\lambda T(u),$$

so $T(\lambda u)=\lambda T(u)$. Hence T is linear.

(c) $\Rightarrow$ (b). Let O and T be as in (c) and let $P,Q\in\mathcal{A}$. Then $f(P)=f(O)+T(\overrightarrow{OP})$ and $f(Q)=f(O)+T(\overrightarrow{OQ})$, so by Proposition 22.1.3 applied in $\mathcal{B}$ with origin $f(O)$, and then linearity of T,

$$\overrightarrow{f(P)f(Q)}=T(\overrightarrow{OQ})-T(\overrightarrow{OP})=T(\overrightarrow{OQ}-\overrightarrow{OP})=T(\overrightarrow{PQ}),$$

the last equality by Proposition 22.1.3 in $\mathcal{A}$.

(d) $\Rightarrow$ (a). Let T be as in (b), let $\sum_i t_i=1$, and put $X=\sum_i t_iP_i$. Choose any $O\in\mathcal{A}$, so that $\overrightarrow{OX}=\sum_i t_i\overrightarrow{OP_i}$ by Definition 22.1.2. Applying (b) to the pairs (O,X) and (O,P_i), and then linearity,

$$\overrightarrow{f(O)f(X)}=T(\overrightarrow{OX})=\sum_i t_i\,T(\overrightarrow{OP_i})=\sum_i t_i\overrightarrow{f(O)f(P_i)}.$$

The right-hand side is, by Definition 22.1.2 computed in $\mathcal{B}$ with origin $f(O)$, the displacement from $f(O)$ to $\sum_i t_i f(P_i)$. Hence $f(X)=\sum_i t_i f(P_i)$, which is (a).

Uniqueness. Suppose T and T' both satisfy (b). Given $v\in V$, fix O and put $Q=O+v$; then $T(v)=\overrightarrow{f(O)f(Q)}=T'(v)$. So T is determined by f, independently of any origin; and a T satisfying (c) for one O satisfies (b) by the implication above, hence is this same map. This proves the theorem.

■

So an affine map is *one linear map plus one displacement*, no more. In $\mathbb{A}^n(F)\to\mathbb{A}^m(F)$ it is

$$f(\mathbf{x})=\mathbf{A}\mathbf{x}+\mathbf{b},\qquad \mathbf{A}\in M_{m\times n}(F),\ \mathbf{b}\in F^m,$$

with $\mathbf{A}$ the matrix of $\vec{f}$ in the standard bases and $\mathbf{b}=f(\mathbf{0})$.

Warning. **Affine is weaker than linear, and the difference is exactly the origin.** An affine map $f\colon V\to W$ between *vector spaces* is linear if and only if $f(0)=0$: if $f(x)=T(x)+b$ then $f(0)=b$, and f is linear exactly when $b=0$. The translation $\tau(x)=x+b$ with $b\neq 0$ is affine with linear part id_V and is **not** linear — Chapter 2 recorded it as Example (Translation), the standard nonexample of a linear map. Conversely $x\mapsto x^2$ on $\mathbb{R}$ fixes 0 and is not affine: it sends the affine combination $2\cdot 1+(-1)\cdot 0=1$ to 1, while $2f(1)-f(0)=2$.

An affine map is determined by very little, but not by too little. Here is the negative half; the positive half is the subject of the next section.

Example. Two points are not enough in the plane

Let $f,g\colon\mathbb{A}^2\to\mathbb{A}^2$ be $g(x,y)=(x,y)$ and $f(x,y)=(x+y,\ y)$. Both are affine, being linear. They agree at the two distinct points $(0,0)$ and $(1,0)$ — indeed at every point of the line

$y = 0$ — yet $f(0,1) = (1,1) \neq (0,1) = g(0,1)$. So two points, even distinct ones, do not determine an affine map of the plane. The count to remember is $n+1$: the next section proves that 3 suitably chosen points do determine an affine map of the plane, and that 2 never will.

22.1.5 The affine group, and the matrix that computes it

Affine maps compose, and the linear part is functorial. That single fact organizes every computation in the chapter.

Theorem 22.1.6. Composition, Inverses, and the Affine Group

Let $\mathcal{A}, \mathcal{B}, \mathcal{C}$ be affine spaces over F, and write V and W for the direction spaces of $\mathcal{A}$ and $\mathcal{B}$.

(a) If $f\colon \mathcal{A} \to \mathcal{B}$ and $g\colon \mathcal{B} \to \mathcal{C}$ are affine, then $g \circ f$ is affine and $\overrightarrow{g \circ f} = \vec{g} \circ \vec{f}$.

(b) An affine $f\colon \mathcal{A} \to \mathcal{B}$ is bijective if and only if $\vec{f}$ is bijective, and then f^{-1} is affine with $\overrightarrow{f^{-1}} = (\vec{f})^{-1}$.

(c) The bijective affine maps $\mathcal{A} \to \mathcal{A}$ form a group $\mathrm{Aff}(\mathcal{A})$ under composition, and $f \mapsto \vec{f}$ is a surjective group homomorphism $\mathrm{Aff}(\mathcal{A}) \to \mathrm{GL}(V)$ whose kernel is the set of **translations** $\tau_v\colon P \mapsto P + v$, $v \in V$.

Idea. Everything is read off from the identity $\overrightarrow{f(P)f(Q)} = \vec{f}(\overrightarrow{PQ})$ of Theorem 22.1.5 (b), which turns statements about points into statements about vectors. For (c), the kernel consists of the affine maps whose linear part is the identity, and such a map moves every point by the same vector.

Proof.

(a) For $P, Q \in \mathcal{A}$, applying Theorem 22.1.5 (b) twice,

$$\overrightarrow{g(f(P))\,g(f(Q))} = \vec{g}(\overrightarrow{f(P)f(Q)}) = \vec{g}(\vec{f}(\overrightarrow{PQ})).$$

Since $\vec{g} \circ \vec{f}$ is linear (Theorem 2.4.2), condition (b) of that theorem holds for $g \circ f$ with this map, so $g \circ f$ is affine and, by the uniqueness clause, $\overrightarrow{g \circ f} = \vec{g} \circ \vec{f}$.

(b) ($\Leftarrow$) Suppose $\vec{f}$ is bijective. If $f(P) = f(Q)$ then $\vec{f}(\overrightarrow{PQ}) = \overrightarrow{f(P)f(Q)} = 0$, so $\overrightarrow{PQ} = 0$ and $P = Q$ by Lemma 22.1.1 (c); thus f is injective. For surjectivity, fix $O \in \mathcal{A}$ and let $R \in \mathcal{B}$. As $\vec{f}$ is surjective, $\overrightarrow{f(O)R} = \vec{f}(v)$ for some $v \in V$, and then $f(O+v) = f(O) + \vec{f}(v) = R$ by part (c) of Theorem 22.1.5.
($\Rightarrow$) Suppose f is bijective and fix $O \in \mathcal{A}$. If $\vec{f}(v) = 0$, then $f(O+v) = f(O)+0 = f(O)$, so $O + v = O$ and $v = 0$; thus $\vec{f}$ is injective. Given $w \in W$, the point $f(O) + w$ is $f(P)$ for some P, and then $w = \overrightarrow{f(O)f(P)} = \vec{f}(\overrightarrow{OP})$; thus $\vec{f}$ is surjective.
For the inverse, let $R, S \in \mathcal{B}$ and put $P = f^{-1}(R)$, $Q = f^{-1}(S)$. Then $\vec{f}(\overrightarrow{PQ}) = \overrightarrow{RS}$, so $\overrightarrow{f^{-1}(R)f^{-1}(S)} = (\vec{f})^{-1}(\overrightarrow{RS})$. Since $(\vec{f})^{-1}$ is linear (Theorem 2.5.1), f^{-1} is affine with the stated linear part.

(c) By (a) and (b), $\mathrm{Aff}(\mathcal{A})$ is closed under composition and inverses and contains $\mathrm{id}_{\mathcal{A}}$,

whose linear part is id_V; composition of functions is associative, so it is a group (Definition 0.10.1). By (a) the map $f \mapsto \vec{f}$ preserves composition, and by (b) it lands in $\mathrm{GL}(V)$; so it is a homomorphism (Definition 0.10.3). It is surjective: given $T \in \mathrm{GL}(V)$, fix O and set $f(P) = O + T(\overrightarrow{OP})$, which is affine with $\vec{f} = T$ by Theorem 22.1.5 (c) and bijective by (b).

Its kernel is $\{f : \vec{f} = \mathrm{id}_V\}$. Each translation τ_v lies there, since $\overrightarrow{\tau_v(P)\tau_v(Q)} = \overrightarrow{PQ}$ by Chasles. Conversely suppose $\vec{f} = \mathrm{id}_V$, fix O and put $v = \overrightarrow{Of(O)}$. For any P, Chasles gives

$$\overrightarrow{Pf(P)} = \overrightarrow{PO} + \overrightarrow{Of(O)} + \overrightarrow{f(O)f(P)} = \overrightarrow{PO} + v + \overrightarrow{OP} = v,$$

so $f(P) = P + v = \tau_v(P)$. This proves the theorem.

■

In coordinates the homomorphism of (c) is visible in the shape of a matrix. Identify $\mathbf{x} \in F^n$ with the column $\binom{\mathbf{x}}{1} \in F^{n+1}$ — the points of $\mathbb{A}^n(F)$ become the vectors of F^{n+1} whose last coordinate is 1.

Proposition 22.1.7. The Block Matrix of an Affine Map

Let $f\colon \mathbb{A}^n(F) \to \mathbb{A}^m(F)$ be affine, $f(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{b}$ with $\mathbf{A} \in M_{m\times n}(F)$ the matrix of $\vec{f}$ and $\mathbf{b} = f(\mathbf{0})$. Put

$$\widehat{f} = \begin{pmatrix} \mathbf{A} & \mathbf{b} \\ \mathbf{0}^{\top} & 1 \end{pmatrix} \in M_{(m+1)\times(n+1)}(F).$$

Then $\widehat{f}\binom{\mathbf{x}}{1} = \binom{f(\mathbf{x})}{1}$ for every $\mathbf{x} \in F^n$; and for affine $f\colon \mathbb{A}^n(F) \to \mathbb{A}^m(F)$ and $g\colon \mathbb{A}^m(F) \to \mathbb{A}^p(F)$ one has $\widehat{g \circ f} = \widehat{g}\,\widehat{f}$. For $m = n$, f is bijective if and only if $\mathbf{A}$ is invertible, and then

$$\widehat{f^{-1}} = (\widehat{f})^{-1} = \begin{pmatrix} \mathbf{A}^{-1} & -\mathbf{A}^{-1}\mathbf{b} \\ \mathbf{0}^{\top} & 1 \end{pmatrix}.$$

Proof. Block multiplication (Theorem 8.1.1) gives

$$\begin{pmatrix} \mathbf{A} & \mathbf{b} \\ \mathbf{0}^{\top} & 1 \end{pmatrix} \begin{pmatrix} \mathbf{x} \\ 1 \end{pmatrix} = \begin{pmatrix} \mathbf{A}\mathbf{x} + \mathbf{b} \\ 1 \end{pmatrix},$$

which is the first claim. Writing $g(\mathbf{y}) = \mathbf{C}\mathbf{y} + \mathbf{d}$, the same rule gives

$$\begin{pmatrix} \mathbf{C} & \mathbf{d} \\ \mathbf{0}^{\top} & 1 \end{pmatrix} \begin{pmatrix} \mathbf{A} & \mathbf{b} \\ \mathbf{0}^{\top} & 1 \end{pmatrix} = \begin{pmatrix} \mathbf{C}\mathbf{A} & \mathbf{C}\mathbf{b} + \mathbf{d} \\ \mathbf{0}^{\top} & 1 \end{pmatrix},$$

and $(g \circ f)(\mathbf{x}) = \mathbf{C}(\mathbf{A}\mathbf{x} + \mathbf{b}) + \mathbf{d} = (\mathbf{C}\mathbf{A})\mathbf{x} + (\mathbf{C}\mathbf{b} + \mathbf{d})$, so the two agree. For $m = n$: f is bijective exactly when $\vec{f}$ is, by Theorem 22.1.6 (b), and $\vec{f}$ is bijective exactly when its matrix $\mathbf{A}$ is invertible, by Theorem 4.1.5 (b). In that case $\mathbf{x} \mapsto \mathbf{A}^{-1}\mathbf{x} - \mathbf{A}^{-1}\mathbf{b}$ is affine and is a two-sided inverse of f, and its block matrix is the one displayed, which is therefore $(\widehat{f})^{-1}$ by the composition rule.

■

So composing affine maps is multiplying $(n+1)\times(n+1)$ matrices, and $\mathrm{Aff}(\mathbb{A}^n(F))$ is a group of such matrices. The last row $(\mathbf{0}^\top \mid 1)$ is what records "these are points, not vectors"; the projective sections later in this chapter read that row again, as the equation of the hyperplane at infinity, after the extra coordinate is moved from last place to first.

Example. Composing a shear and a quarter turn

In $\mathbb{A}^2$ let $f(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{b}$ and $g(\mathbf{y}) = \mathbf{C}\mathbf{y} + \mathbf{d}$, where

$$\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \quad \mathbf{C} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \quad \mathbf{d} = \begin{pmatrix} 2 \\ 0 \end{pmatrix}.$$

Compute $g \circ f$ in block form, and find f^{-1}.

Solution. The block matrices multiply as

$$\begin{pmatrix} 0 & -1 & 2 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 & 1 \\ 0 & 1 & -1 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & -1 & 3 \\ 1 & 2 & 1 \\ 0 & 0 & 1 \end{pmatrix},$$

so $(g \circ f)(x, y) = (-y + 3,\ x + 2y + 1)$. **Check directly:** $f(x, y) = (x + 2y + 1,\ y - 1)$, and applying g gives $(-(y-1) + 2,\ x + 2y + 1) = (3 - y,\ x + 2y + 1)$, the same.

For the inverse, $\mathbf{A}^{-1} = \begin{pmatrix} 1 & -2 \\ 0 & 1 \end{pmatrix}$ and $-\mathbf{A}^{-1}\mathbf{b} = -\begin{pmatrix} 1 & -2 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 1 \\ -1 \end{pmatrix} = \begin{pmatrix} -3 \\ 1 \end{pmatrix}$, so $f^{-1}(x, y) = (x - 2y - 3,\ y + 1)$. **Check:** $f^{-1}(f(x, y)) = ((x + 2y + 1) - 2(y - 1) - 3,\ (y - 1) + 1) = (x, y)$.

Two things have been set up. Affine maps are linear maps that have forgotten where the origin is, and they are computed by one matrix of size one larger. The next section uses both to describe the subsets of an affine space that are themselves affine spaces.

22.1.6 Exercises

22.1.6.1 A. Check your understanding

Exercise. A1

(a) State Chasles's relation, and deduce $\overrightarrow{QP} = -\overrightarrow{PQ}$ from it.

(b) In an affine space over a field in which $2 \neq 0$ and $3 \neq 0$, which of $P + Q$, $3P$, $\overrightarrow{PQ} + \overrightarrow{RS}$, $P + \overrightarrow{QR}$, $\frac{1}{2}P + \frac{1}{2}Q$ are defined? For each defined one, say whether it is a point or a vector.

(c) True or false, with a reason: a map $f\colon \mathbb{R}^2 \to \mathbb{R}^2$ with $f(\mathbf{0}) = \mathbf{0}$ is affine if and only if it is linear.

(d) An affine map $f\colon \mathbb{A}^3 \to \mathbb{A}^3$ has $\vec{f}$ of rank 2. Is f injective? Justify your answer.

Solution.

(a) $\overrightarrow{PQ} + \overrightarrow{QR} = \overrightarrow{PR}$ for all points P, Q, R. Taking $R = P$ gives $\overrightarrow{PQ} + \overrightarrow{QP} = \overrightarrow{PP}$, and

$\overrightarrow{PP} = 0$ by Lemma 22.1.1 (a); hence $\overrightarrow{QP} = -\overrightarrow{PQ}$.

(b) $P + Q$ and $3P$ are **not** defined: the weights add up to 2 and to 3, and under the hypotheses on F neither sum is 0 or 1, so Definition 22.1.2 gives these expressions no meaning; Proposition 22.1.4 (c) says why nothing could, since for $\dim \mathcal{A} \geq 1$ no such expression is origin-free. $\overrightarrow{PQ} + \overrightarrow{RS}$ is a vector (a sum in V). $P + \overrightarrow{QR}$ is a point. $\frac{1}{2}P + \frac{1}{2}Q$ is a point, the midpoint, and $\frac{1}{2}$ exists because $2 \neq 0$. (The hypotheses on F are needed: over $\mathbb{F}_2$ the weights of $P + Q$ add up to $2 = 0$, and Definition 22.1.2 (b) makes $P + Q$ the vector $\overrightarrow{OP} + \overrightarrow{OQ} = \overrightarrow{PQ}$.)

(c) True. An affine map is $f(\mathbf{x}) = T(\mathbf{x}) + \mathbf{b}$ with T linear and $\mathbf{b} = f(\mathbf{0})$ (Theorem 22.1.5), so $f(\mathbf{0}) = \mathbf{0}$ forces $\mathbf{b} = \mathbf{0}$ and $f = T$. Conversely every linear map is affine, with $\vec{f} = f$, and sends $\mathbf{0}$ to $\mathbf{0}$.

(d) No. By Theorem 22.1.6 (b), f is bijective exactly when $\vec{f}$ is, and for an operator on a finite-dimensional space injective, surjective, and bijective agree (Theorem 2.5.7). Here $\operatorname{rank} \vec{f} = 2 < 3$, so $\vec{f}$ is not injective; taking $\mathbf{v} \neq \mathbf{0}$ in $\ker \vec{f}$ and any P, the points P and $P + \mathbf{v}$ have the same image, since $\overrightarrow{f(P)f(P+\mathbf{v})} = \vec{f}(\mathbf{v}) = \mathbf{0}$.

22.1.6.2 B. Practice

Exercise. B1: Determine which are affine

Determine which of the following maps are affine. Justify your answer; for each affine one give its linear part, and for each other one name a failing affine combination.

(a) $f\colon \mathbb{A}^2 \to \mathbb{A}^2$, $f(x, y) = (2x - y + 1,\ 5)$.

(b) $f\colon \mathbb{A}^2 \to \mathbb{A}^1$, $f(x, y) = xy$.

(c) $f\colon \mathbb{A}^1 \to \mathbb{A}^2$, $f(t) = (1 - t,\ 2 + 3t)$.

(d) $f\colon \mathbb{A}^2 \to \mathbb{A}^2$, $f(x, y) = (x + 1,\ |y|)$.

Solution.

(a) Affine: $f(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{b}$ with $\mathbf{A} = \begin{pmatrix} 2 & -1 \\ 0 & 0 \end{pmatrix}$ and $\mathbf{b} = (1, 5)$. Its linear part is $\mathbf{x} \mapsto \mathbf{A}\mathbf{x}$, of rank 1.

(b) Not affine. Take the affine combination $2(1, 1) - 1 \cdot (0, 0) = (2, 2)$, with weights $2 + (-1) = 1$. Then $f(2, 2) = 4$, while $2f(1, 1) - f(0, 0) = 2$.

(c) Affine: $f(t) = \mathbf{A}t + \mathbf{b}$ with $\mathbf{A} = \begin{pmatrix} -1 \\ 3 \end{pmatrix}$ and $\mathbf{b} = (1, 2)$. Linear part $t \mapsto (-t, 3t)$, of rank 1.

(d) Not affine. Take the affine combination $2(0, 1) - 1 \cdot (0, 3) = (0, -1)$, with weights $2 + (-1) = 1$. Then $f(0, -1) = (1, 1)$, while $2f(0, 1) - f(0, 3) = 2(1, 1) - (1, 3) = (1, -1)$. The two differ, so f is not affine. (The first coordinate is affine in x; it is $|y|$ that fails.)

Exercise. B2: Block matrices

In $\mathbb{A}^2$, let r be the affine map with $\widehat{r} = \begin{pmatrix} 0 & -1 & 1 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ and let τ be the translation by $(2, 3)$.

(a) Write r and τ in the form $\mathbf{x} \mapsto \mathbf{A}\mathbf{x} + \mathbf{b}$.

(b) Compute $\widehat{\tau \circ r}$ and $\widehat{r \circ \tau}$, and say whether the two maps are equal.

(c) Find all fixed points of r.

Solution.

(a) $r(x, y) = (-y+1,\ x)$, that is, $\mathbf{A}_r = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$, $\mathbf{b}_r = (1, 0)$. And $\tau(x, y) = (x+2,\ y+3)$, with $\mathbf{A}_\tau = \mathbf{I}_2$, $\mathbf{b}_\tau = (2, 3)$.

(b) By Proposition 22.1.7, $\widehat{\tau \circ r} = \widehat{\tau}\,\widehat{r}$ and $\widehat{r \circ \tau} = \widehat{r}\,\widehat{\tau}$:

$$\widehat{\tau}\,\widehat{r} = \begin{pmatrix} 0 & -1 & 3 \\ 1 & 0 & 3 \\ 0 & 0 & 1 \end{pmatrix}, \qquad \widehat{r}\,\widehat{\tau} = \begin{pmatrix} 0 & -1 & -2 \\ 1 & 0 & 2 \\ 0 & 0 & 1 \end{pmatrix}.$$

They differ, so $\tau \circ r \neq r \circ \tau$. (The linear parts agree; the translation vectors do not, because $\vec{r}$ moves $\mathbf{b}_\tau$.)

(c) $r(x, y) = (x, y)$ reads $-y + 1 = x$ and $x = y$, so $x = y = \frac{1}{2}$. The unique fixed point is $(\frac{1}{2}, \frac{1}{2})$. Equivalently, $(\mathbf{I} - \mathbf{A}_r)\mathbf{x} = \mathbf{b}_r$ with $\mathbf{I} - \mathbf{A}_r = \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}$ invertible.

Exercise. B3: The four-point plane

Let $\mathcal{A} = \mathbb{F}_2^2$ as in Example (Four affine spaces) (d), with points $O = (0, 0)$, $P = (1, 0)$, $Q = (0, 1)$, $R = (1, 1)$.

(a) Compute $\overrightarrow{PQ}$ and $\overrightarrow{QP}$, and explain why they are equal.

(b) How many affine combinations of the three points O, P, Q are there, and which points do they give? (A combination is counted by its list of weights.)

(c) How many bijective affine maps $\mathcal{A} \to \mathcal{A}$ are there?

Solution.

(a) $\overrightarrow{PQ} = Q - P = (0, 1) - (1, 0) = (1, 1)$ and $\overrightarrow{QP} = P - Q = (1, 1)$. They are equal because $\overrightarrow{QP} = -\overrightarrow{PQ}$ (Lemma 22.1.1 (a)) and $-\mathbf{v} = \mathbf{v}$ for every $\mathbf{v}$ over $\mathbb{F}_2$.

(b) A weight list $(t_0, t_1, t_2) \in \mathbb{F}_2^3$ with $t_0 + t_1 + t_2 = 1$ has an odd number of ones, so it is one of $(1, 0, 0), (0, 1, 0), (0, 0, 1), (1, 1, 1)$: four lists. The first three give O, P, Q, and $(1, 1, 1)$ gives $O + P + Q = (1, 1) = R$. So the four affine combinations give the four points, each exactly once.

(c) By Theorem 22.1.6 (c) an affine bijection is a translation composed with a linear one, and the number of pairs $(T, \mathbf{v}) \in \mathrm{GL}(\mathbb{F}_2^2) \times \mathbb{F}_2^2$ is $|\mathrm{GL}_2(\mathbb{F}_2)| \cdot 4$. A matrix in $\mathrm{GL}_2(\mathbb{F}_2)$ has a nonzero first column (3 choices) and a second column outside the span of the first ($4 - 2 = 2$ choices), so $|\mathrm{GL}_2(\mathbb{F}_2)| = 6$ and there are 24 affine bijections. Since $|\mathcal{A}| = 4$,

these are **all** $4! = 24$ permutations of the four points.

22.1.6.3 C. Going deeper

Exercise. C1: Fixed points

Let $f\colon \mathbb{A}^n \to \mathbb{A}^n$ be affine, $f(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{b}$.

(a) Prove that f has a fixed point if and only if $\mathbf{b} \in \operatorname{col}(\mathbf{I} - \mathbf{A})$, and that the set of fixed points, when nonempty, is a coset of $\operatorname{nul}(\mathbf{I} - \mathbf{A})$.

(b) Deduce that if 1 is not an eigenvalue of $\mathbf{A}$, then f has exactly one fixed point.

(c) Give an affine map of $\mathbb{A}^2$ with no fixed point and linear part $\neq \mathrm{id}$.

Solution.

(a) $f(\mathbf{x}) = \mathbf{x}$ says $\mathbf{A}\mathbf{x} + \mathbf{b} = \mathbf{x}$, that is, $(\mathbf{I} - \mathbf{A})\mathbf{x} = \mathbf{b}$. This system is consistent exactly when $\mathbf{b} \in \operatorname{col}(\mathbf{I} - \mathbf{A})$ (Theorem 3.1.1), and then its solution set is $\mathbf{p} + \operatorname{nul}(\mathbf{I} - \mathbf{A})$ for any one solution $\mathbf{p}$, by Theorem 3.1.3.

(b) If 1 is not an eigenvalue of $\mathbf{A}$, then $\operatorname{nul}(\mathbf{I} - \mathbf{A}) = \{\mathbf{0}\}$, so $\mathbf{I} - \mathbf{A}$ is invertible (Theorem 3.4.4). Hence $(\mathbf{I} - \mathbf{A})\mathbf{x} = \mathbf{b}$ has the unique solution $\mathbf{x} = (\mathbf{I} - \mathbf{A})^{-1}\mathbf{b}$.

(c) Take $\mathbf{A} = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$ and $\mathbf{b} = (1, 0)$, so $f(x, y) = (x + 1,\ 2y)$. Then $\mathbf{I} - \mathbf{A} = \begin{pmatrix} 0 & 0 \\ 0 & -1 \end{pmatrix}$ and $\operatorname{col}(\mathbf{I} - \mathbf{A}) = \operatorname{span}(\mathbf{e}_2)$, which does not contain $\mathbf{b} = \mathbf{e}_1$; so there is no fixed point, by (a). And $\mathbf{A} \neq \mathbf{I}$.

Exercise. C2: Affine maps on two points, over two fields

Let $f\colon \mathcal{A} \to \mathcal{B}$ be a map of affine spaces over F that preserves affine combinations of **two** points, that is, $f((1-t)P + tQ) = (1-t)f(P) + tf(Q)$ for all $P, Q \in \mathcal{A}$ and $t \in F$.

(a) Prove that if $\operatorname{char} F \neq 2$, then f is affine. *Hint: run the proof of Theorem 22.1.5, replacing the three-point combination by two two-point ones.*

(b) Show that over $\mathbb{F}_2$ the hypothesis is empty, and give a nonaffine map $\mathbb{F}_2^2 \to \mathbb{F}_2$ satisfying it.

Solution.

(a) Fix $O \in \mathcal{A}$ and define $T(v) = \overrightarrow{f(O)f(O+v)}$ as in Theorem 22.1.5. Homogeneity is the hypothesis with $P = O + v$, $Q = O$: $O + \lambda v = \lambda(O + v) + (1 - \lambda)O$, so $f(O + \lambda v) = \lambda f(O + v) + (1 - \lambda)f(O) = f(O) + \lambda T(v)$, giving $T(\lambda v) = \lambda T(v)$.
For additivity, let $u, w \in V$ and let $M = \frac{1}{2}(O + u) + \frac{1}{2}(O + w)$, which is legal since $2 \neq 0$ in F. With origin O, $M = O + \frac{1}{2}(u + w)$. By the hypothesis, $f(M) = \frac{1}{2}f(O + u) + \frac{1}{2}f(O + w)$, which with origin $f(O)$ reads $f(M) = f(O) + \frac{1}{2}(T(u) + T(w))$. So $T\big(\frac{1}{2}(u + w)\big) = \frac{1}{2}(T(u) + T(w))$. Applying homogeneity with $\lambda = 2$ to the left-hand side gives $T(u + w) = T(u) + T(w)$. Hence T is linear, and $f(P) = f(O) + T(\overrightarrow{OP})$ by construction, so f is affine by Theorem 22.1.5 (c).

(b) Over $\mathbb{F}_2$ the only scalars are $t = 0$ and $t = 1$, for which $(1 - t)P + tQ$ is P or Q; the

required identity then reads $f(P) = f(P)$ or $f(Q) = f(Q)$. So every map satisfies the hypothesis. Take $f\colon \mathbb{F}_2^2 \to \mathbb{F}_2$ with $f(0,0) = f(1,0) = f(0,1) = 0$ and $f(1,1) = 1$. It is not affine: by Exercise (B3: The four-point plane) (b), $(1,1)$ is the affine combination $O + P + Q$ of the other three points with weights $(1,1,1)$, so an affine f would have $f(1,1) = f(O) + f(P) + f(Q) = 0$.

Exercise. C3: Conjugating a translation

Let $\mathcal{A}$ be an affine space with direction space V, let $f \in \mathrm{Aff}(\mathcal{A})$ and let $v \in V$.

(a) Prove that $f \circ \tau_v \circ f^{-1} = \tau_{\vec{f}(v)}$.

(b) Hence prove that the translations form a subgroup of $\mathrm{Aff}(\mathcal{A})$ that is carried to itself by every conjugation, and that $v \mapsto \tau_v$ is an isomorphism from $(V, +)$ onto it.

(c) Deduce that if $\dim \mathcal{A} \geq 2$, then $\mathrm{Aff}(\mathcal{A})$ is not abelian.

Solution.

(a) Both sides are affine maps $\mathcal{A} \to \mathcal{A}$. By Theorem 22.1.6 (a) the linear part of the left-hand side is $\vec{f} \circ \mathrm{id}_V \circ (\vec{f})^{-1} = \mathrm{id}_V$, so by the kernel description in Theorem 22.1.6 (c) it is a translation τ_u for some u. To identify u, fix $O \in \mathcal{A}$ and put $P = f^{-1}(O)$. Then

$$u = \overrightarrow{O\,(f \circ \tau_v \circ f^{-1})(O)} = \overrightarrow{f(P)\,f(P+v)} = \vec{f}(v),$$

using Theorem 22.1.5 (b) and $\overrightarrow{P(P+v)} = v$.

(b) The translations are the kernel of $f \mapsto \vec{f}$ (Theorem 22.1.6 (c)), hence a subgroup by Theorem 0.10.3, and (a) says every conjugate of a translation is a translation. The map $v \mapsto \tau_v$ satisfies $\tau_u \circ \tau_w = \tau_{u+w}$ by Lemma 22.1.1 (b), so it is a homomorphism $(V, +) \to \mathrm{Aff}(\mathcal{A})$; it is surjective onto the translations by definition, and injective because $\tau_v = \mathrm{id}$ forces $P + v = P$, hence $v = 0$.

(c) Since $\dim V \geq 2$, pick independent $v, w \in V$ and $T \in \mathrm{GL}(V)$ with $T(v) = w$ (extend $\{v, w\}$ to a basis of V by Theorem 1.8.2, and map that basis to itself with v and w swapped, using Theorem 2.1.6; this T is invertible because $T \circ T$ fixes every basis vector and so is id_V, by the uniqueness in that theorem). By Theorem 22.1.6 (c) there is $f \in \mathrm{Aff}(\mathcal{A})$ with $\vec{f} = T$. If $\mathrm{Aff}(\mathcal{A})$ were abelian then $f \circ \tau_v \circ f^{-1} = \tau_v$, while (a) gives $\tau_{T(v)} = \tau_w$; by the injectivity in (b) that forces $v = w$, contradicting independence.

22.2 Flats, Independence, and Coordinates

An affine space has no origin, so it has no subspaces in the sense of Chapter 1: a subspace must contain 0, and there is no 0. What it has instead are its **flats** — the lines, planes, and hyperplanes that sit inside it. This section defines them, settles when two of them meet and how big the smallest flat containing both is, puts coordinates on an affine space using $n + 1$ points rather than n vectors, and shows that in $\mathbb{A}^n(F)$ the flats are exactly the solution sets of linear systems. The join formula has two cases, and the second one is where affine geometry parts company with linear algebra.

Throughout, $\mathcal{A}$ is an affine space over a field F with direction space V (Definition 22.1.1), and

from the dimension formula onward $\dim \mathcal{A} = \dim V = n$ is finite, except where a statement says otherwise.

22.2.1 Flats

A subset of $\mathcal{A}$ deserves to be called flat when it is an affine space in its own right, using the same difference map. By Definition 22.1.2 there is only one thing to require.

A flat is a nonempty set of points closed under taking affine combinations.

Definition 22.2.1. Flat

A subset $A \subseteq \mathcal{A}$ is a **flat**, or an **affine subspace** of $\mathcal{A}$, if it is **nonempty** and contains every affine combination of its own points: **whenever** $k \geq 1$, $P_1, \dots, P_k \in A$ and $t_1, \dots, t_k \in F$ satisfy $t_1 + \cdots + t_k = 1$, the point $\sum_{i=1}^{k} t_i P_i$ lies in A.

The requirement that A be nonempty is a convention, and it is the useful one: a flat is going to have a direction space and a dimension, and the empty set has neither. Intersections of flats may of course be empty, and we will say so each time.

Proposition 22.2.1. Flats Are the Cosets

Let $A \subseteq \mathcal{A}$ be nonempty. Then A is a flat if and only if there are $P \in \mathcal{A}$ and a subspace U of V with $A = P + U \coloneqq \{P + u : u \in U\}$. In that case:

(a) $U = \{\overrightarrow{XY} : X, Y \in A\}$ is determined by A alone; it is written $\vec{A}$ and called the **direction space** of A.

(b) $A = Q + \vec{A}$ for **every** $Q \in A$, and A, with the difference map of $\mathcal{A}$ restricted to it, is an affine space with direction space $\vec{A}$.

The **dimension** of A is $\dim A \coloneqq \dim \vec{A}$. A flat of dimension 0 is a single **point**, one of dimension 1 is a **line**, one of dimension 2 is a **plane**, and, when $\dim \mathcal{A} = n$ is finite, one of dimension $n - 1$ is a **hyperplane** of $\mathcal{A}$.

Idea. One direction is a computation with origin P. For the other, fix $P \in A$ and let U be the set of displacements from P to points of A; the three subspace axioms come from three affine combinations, with weight lists (1), $(\lambda, 1-\lambda)$ and $(1, 1, -1)$.

Proof. ($\Leftarrow$) Let $A = P + U$ with U a subspace, and let $P + u_1, \dots, P + u_k \in A$ with $\sum_i t_i = 1$. Computing the affine combination with origin P (Definition 22.1.2),

$$\sum_i t_i (P + u_i) = P + \sum_i t_i u_i \in P + U,$$

since U is closed under linear combinations. And $A \ni P$ is nonempty.

($\Rightarrow$) Let A be a flat, fix $P \in A$ and put $U = \{\overrightarrow{PX} : X \in A\}$. Then $0 = \overrightarrow{PP} \in U$. If $u = \overrightarrow{PX} \in U$ and $\lambda \in F$, the point $\lambda X + (1-\lambda)P$ lies in A and equals $P + \lambda u$ when computed with origin P, so $\lambda u \in U$. If also $w = \overrightarrow{PY} \in U$, the point $X + Y - P$ lies in A (weights $1 + 1 - 1 = 1$) and equals $P + (u + w)$, so $u + w \in U$. By Theorem 1.2.1, U is a

subspace, and $A = P + U$ by construction.

(a) Let $X, Y \in A$. Then $\overrightarrow{XY} = \overrightarrow{XP} + \overrightarrow{PY} \in U$ by Chasles, since U is closed under negation and addition; so $\{\overrightarrow{XY} : X, Y \in A\} \subseteq U$. Conversely each $u \in U$ is $\overrightarrow{PX}$ for some $X \in A$, and $P \in A$. Hence the two sets are equal, and the right-hand side mentions only A.

(b) Let $Q \in A$. For $X \in A$ we have $\overrightarrow{QX} \in \vec{A}$ by (a), so $A \subseteq Q + \vec{A}$; and for $u \in \vec{A}$, writing $u = \overrightarrow{XY}$ with $X, Y \in A$, the point $Q + u$ is the affine combination $Q + Y - X$ of points of A, hence lies in A. So $A = Q + \vec{A}$. Finally the restricted difference map takes values in $\vec{A}$ by (a), Chasles is inherited, and for fixed $P \in A$ the map $X \mapsto \overrightarrow{PX}$ is a bijection $A \to \vec{A}$, being the restriction of the bijection of (AS2) to $A = P + \vec{A}$. So (AS1) and (AS2) hold and A is an affine space with direction space $\vec{A}$.

■

Over $\mathbb{R}$, with $\mathcal{A} = V$, a flat is exactly Chapter 18's affine subspace of V (Definition 18.1.5 (b)), and the two dimensions agree: Chapter 18 also reads the subspace off the coset as the set of differences. Chapter 5's **affine hyperplane** $\mathbf{p} + H$ (the paragraph before Proposition 5.6.4) is the case $\dim A = n - 1$ in $\mathbb{A}^n(F)$. A one-point set $\{P\} = P + \{0\}$ is a flat, and $\mathcal{A} = P + V$ is a flat: the two degenerate cases.

A nonexample, by minimal change. A line is a flat; the union of two lines is not. Let $X = \{(x, y) \in \mathbb{A}^2 : xy = 0\}$, the union of the two coordinate axes. It is nonempty, and it contains every affine combination of points taken from one axis, each axis being a flat. But Definition 22.2.1 asks for all of them: the points $P = (1, 0)$ and $Q = (0, 1)$ lie in X, the weights of $\frac{1}{2}P + \frac{1}{2}Q$ add up to 1, and $\frac{1}{2}P + \frac{1}{2}Q = (\frac{1}{2}, \frac{1}{2})$ has $xy = \frac{1}{4} \neq 0$. So the closure clause fails for one pair of points, and X is not a coset of a subspace, although it is a union of two of them.

Proposition 22.2.2. Intersections of Flats

Let $(A_i)_{i \in I}$ be a nonempty family of flats of $\mathcal{A}$ and put $A = \bigcap_{i \in I} A_i$. If $A \neq \emptyset$, then A is a flat and $\vec{A} = \bigcap_{i \in I} \vec{A}_i$.

Proof. Suppose $A \neq \emptyset$ and fix $P \in A$. If $P_1, \dots, P_k \in A$ and $\sum_j t_j = 1$, then the point $\sum_j t_j P_j$ lies in each A_i, because each A_i is a flat containing all the P_j; so it lies in A. Hence A is a flat. By Proposition 22.2.1 (b), $A_i = P + \vec{A}_i$ for each i and $A = P + \vec{A}$, so

$$\vec{A} = \{\overrightarrow{PX} : X \in A\} = \bigcap_{i \in I} \{\overrightarrow{PX} : X \in A_i\} = \bigcap_{i \in I} \vec{A}_i,$$

where the middle equality holds because $X \mapsto \overrightarrow{PX}$ is injective (AS2).

■

Since $\mathcal{A}$ itself is a flat containing every subset, the following is well defined.

Definition 22.2.2. Join of Flats

Let A, B be flats of $\mathcal{A}$. Their **join** $A \vee B$ is the intersection of all flats of $\mathcal{A}$ containing $A \cup B$.

The intersection contains A, so it is nonempty, and it is a flat by Proposition 22.2.2: $A \vee B$ is the **smallest** flat containing both. A single point P is a flat, and we write $P \vee Q$, $A \vee P$ for joins in which a one-point flat $\{P\}$ appears. This is the affine analogue of $U + W$, and Chapter 18's affine hull is the same construction over $\mathbb{R}$ (Proposition 18.1.4).

Parallelism is what an affine space has and a vector space does not notice.

Definition 22.2.3. Parallel Flats

Two flats A, B of $\mathcal{A}$ are **parallel**, written $A \parallel B$, if $\vec{A} = \vec{B}$.

Proposition 22.2.3. Parallelism

Parallelism is an equivalence relation on the set of flats of $\mathcal{A}$. Two parallel flats are either equal or disjoint, and each equivalence class consists of the flats $P + U$, $P \in \mathcal{A}$, for one fixed subspace U.

Proof. Equality of direction spaces is an equivalence relation, so $\parallel$ is one. Let $A \parallel B$ with common direction U, and suppose $X \in A \cap B$. Then $A = X + U = B$ by Proposition 22.2.1 (b). So A and B are equal or disjoint. Finally, the flats with direction space U are exactly the sets $P + U$ with $P \in \mathcal{A}$, again by Proposition 22.2.1.

■

Parallel flats have the same dimension. Some authors also call a line parallel to a plane containing its direction; we do not, and we say "$\vec{A} \subseteq \vec{B}$" when we mean that.

22.2.2 When two flats meet, and how big their join is

Here is the central computation of the section. Two subspaces of a vector space always meet, in 0 at least. Two flats need not, and the dimension formula splits accordingly.

Theorem 22.2.4. Intersection and Join of Two Flats

Let A, B be flats of $\mathcal{A}$, and choose $P \in A$ and $Q \in B$. Parts (a) and (b) need no hypothesis on $\dim \mathcal{A}$.

(a) $A \vee B = P + (\vec{A} + \vec{B} + \operatorname{span}(\overrightarrow{PQ}))$.

(b) $A \cap B \neq \emptyset$ if and only if $\overrightarrow{PQ} \in \vec{A} + \vec{B}$; and then $A \cap B$ is a flat whose direction space is $\vec{A} \cap \vec{B}$.

(c) Let $\dim \mathcal{A} = n$ be finite. If $A \cap B \neq \emptyset$, then

$$\dim(A \vee B) = \dim A + \dim B - \dim(A \cap B).$$

If $A \cap B = \emptyset$, then

$$\dim(A \vee B) = \dim A + \dim B - \dim(\vec{A} \cap \vec{B}) + 1.$$

(d) Let $\dim \mathcal{A} = n$ be finite. If $A \cap B = \emptyset$, there is a hyperplane H of V with $A \subseteq P + H$ and $B \subseteq Q + H$, and the two flats $P + H$, $Q + H$ are parallel, distinct, and disjoint.

Idea. Everything is read off from one subspace, $W = \vec{A} + \vec{B} + \operatorname{span}(\overrightarrow{PQ})$. The three summands are forced: the join must contain the directions of A and of B, and it must contain a way of getting from A to B. Whether the third summand is already inside the first two is exactly whether A and B meet — and that one extra dimension, present precisely when they miss each other, is the whole difference between (c)'s two cases.

Proof. Put $W = \vec{A} + \vec{B} + \operatorname{span}(\overrightarrow{PQ})$, a subspace of V, and $C = P + W$, a flat by Proposition 22.2.1.

(a) First, $A \cup B \subseteq C$. Indeed $A = P + \vec{A} \subseteq P + W$. And each point of B is $Q + b$ with $b \in \vec{B}$, which equals $P + (\overrightarrow{PQ} + b)$ by Lemma 22.1.1 (b), and $\overrightarrow{PQ} + b \in W$. So C is one of the flats being intersected in Definition 22.2.2, whence $A \vee B \subseteq C$.
Conversely, $A \vee B$ is a flat containing P, so $A \vee B = P + U$ by Proposition 22.2.1 (b), where U is the direction space of $A \vee B$. Since $A \subseteq A \vee B$ and $P \in A$, part (a) of that proposition gives $\vec{A} \subseteq U$. Since $Q \in A \vee B$, also $\overrightarrow{PQ} \in U$. And for $b \in \vec{B}$ the point $Q + b$ lies in $B \subseteq A \vee B$, so $\overrightarrow{PQ} + b = \overrightarrow{P\,(Q + b)} \in U$, and therefore $b \in U$. Hence $W \subseteq U$, that is, $C \subseteq A \vee B$. The two inclusions give (a).

(b) ($\Leftarrow$) Suppose $\overrightarrow{PQ} = a + b$ with $a \in \vec{A}$, $b \in \vec{B}$. Put $X = P + a$, a point of A. Then

$$\overrightarrow{QX} = \overrightarrow{QP} + \overrightarrow{PX} = -(a + b) + a = -b \in \vec{B},$$

so $X \in Q + \vec{B} = B$. Hence $X \in A \cap B$.
($\Rightarrow$) If $X \in A \cap B$, then $\overrightarrow{PX} \in \vec{A}$ and $\overrightarrow{XQ} \in \vec{B}$ by Proposition 22.2.1 (a), so $\overrightarrow{PQ} = \overrightarrow{PX} + \overrightarrow{XQ} \in \vec{A} + \vec{B}$. The last clause is Proposition 22.2.2 applied to the two flats A and B.

(c) *Case 1:* $A \cap B \neq \emptyset$. By (b), $\overrightarrow{PQ} \in \vec{A} + \vec{B}$, so $W = \vec{A} + \vec{B}$. By (a) and Theorem 1.7.3,

$$\begin{aligned} \dim(A \vee B) &= \dim(\vec{A} + \vec{B}) \\ &= \dim \vec{A} + \dim \vec{B} - \dim(\vec{A} \cap \vec{B}), \end{aligned}$$

and $\dim(\vec{A} \cap \vec{B}) = \dim(A \cap B)$ by (b), which is the first formula.
Case 2: $A \cap B = \emptyset$. By (b), $\overrightarrow{PQ} \notin \vec{A} + \vec{B}$; in particular $\overrightarrow{PQ} \neq 0$. Then

$$\operatorname{span}(\overrightarrow{PQ}) \cap (\vec{A} + \vec{B}) = \{0\},$$

since a nonzero $\lambda\overrightarrow{PQ}$ lying in $\vec{A} + \vec{B}$ would give $\overrightarrow{PQ} = \lambda^{-1}(\lambda\overrightarrow{PQ}) \in \vec{A} + \vec{B}$, λ being nonzero and hence invertible. So Theorem 1.7.3 applied to the two subspaces $\vec{A} + \vec{B}$ and $\operatorname{span}(\overrightarrow{PQ})$ gives $\dim W = \dim(\vec{A} + \vec{B}) + 1$, and one more application, as in Case 1, gives the second formula.

(d) Put $U = \vec{A} + \vec{B}$ and $r = \dim U$. By (b), $\overrightarrow{PQ} \notin U$, so appending $\overrightarrow{PQ}$ to a basis $(u_1, \dots, u_r)$ of U leaves an independent list (Lemma 1.5.3). Extend it to a basis

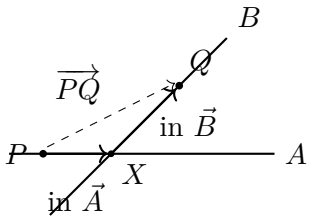
$$(u_1, \dots, u_r,\ \overrightarrow{PQ},\ w_1, \dots, w_{n-r-1})$$

of V by Theorem 1.6.5, and let $H = \operatorname{span}(u_1, \dots, u_r, w_1, \dots, w_{n-r-1})$. Then $\dim H = n-1$, so $\dim(V/H) = 1$ by Theorem 4.4.5 and H is a hyperplane (Definition 5.5.2). Now $\vec{A} \subseteq U \subseteq H$ gives $A = P + \vec{A} \subseteq P + H$, and likewise $B \subseteq Q + H$. The flats $P + H$ and $Q + H$ are parallel by Definition 22.2.3. They are distinct: if they were equal, then $Q \in P + H$ would give $\overrightarrow{PQ} \in H$, contradicting the independence of the displayed basis. Being parallel and distinct, they are disjoint by Proposition 22.2.3. This proves the theorem.

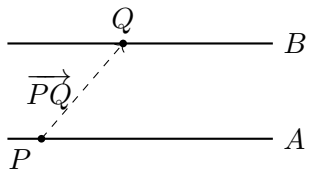
■

The two formulas in (c) are best read side by side. When the flats meet, the count is the subspace count of Theorem 1.7.3, with $A \cap B$ in the role of $\vec{A} \cap \vec{B}$. When they miss, the correction term is taken among the **directions**, where the flats always do meet, and a $+1$ is paid for the displacement that carries one flat to the other.

Part (b) is what the two pictures below show. On the left the displacement $\overrightarrow{PQ}$ splits as a step inside $\vec{A}$ followed by a step inside $\vec{B}$, and the flats meet; on the right it does not split, so it contributes a direction of its own to the join.



$A \cap B \neq \emptyset$



$A \cap B = \emptyset$

Example. Two lines in space, twice

In $\mathbb{A}^3$ let

$$A = (1,0,0) + \operatorname{span}((1,1,0)),$$
$$B = (0,0,1) + \operatorname{span}((1,-1,0)),$$
$$B' = (0,0,1) + \operatorname{span}((1,1,0)).$$

Decide in each case whether the lines meet, and compute $\dim(A \vee B)$ and $\dim(A \vee B')$.

Solution. Here $P = (1,0,0)$, $Q = (0,0,1)$ and $\overrightarrow{PQ} = (-1,0,1)$.

***A* and *B*.** The directions $(1,1,0)$ and $(1,-1,0)$ are independent, and $\vec{A} + \vec{B}$ is the plane of vectors with third coordinate 0. Since $\overrightarrow{PQ} = (-1,0,1)$ has third coordinate 1, it is not in $\vec{A} + \vec{B}$, so $A \cap B = \emptyset$ by Theorem 22.2.4 (b): the lines are **skew**, that is, disjoint and not parallel. (Directly: $(1+s, s, 0) = (t, -t, 1)$ forces $0 = 1$.) Also $\vec{A} \cap \vec{B} = \{\mathbf{0}\}$, the two directions being independent. The second formula of (c) gives

$$\dim(A \vee B) = 1 + 1 - 0 + 1 = 3,$$

so $A \vee B = \mathbb{A}^3$: two skew lines span the whole space.

***A* and *B′*.** Now $\vec{A} = \vec{B'}$, so the lines are parallel, and they are distinct because $\overrightarrow{PQ} = (-1,0,1) \notin \vec{A}$; hence they are disjoint (Proposition 22.2.3). Here $\vec{A} \cap \vec{B'} = \vec{A}$ has dimension

1, so

$$\dim(A \vee B') = 1 + 1 - 1 + 1 = 2,$$

and $A \vee B'$ is the plane $(1,0,0) + \operatorname{span}((1,1,0),(-1,0,1))$, by part (a).

Warning. **There is no convention for $\dim \emptyset$ that makes one formula do both cases.** It is tempting to set $\dim \emptyset = -1$ and keep writing $\dim(A \vee B) = \dim A + \dim B - \dim(A \cap B)$. For the skew lines of Example (Two lines in space, twice) that gives $1 + 1 - (-1) = 3$, which is right; for the parallel lines A and B' it gives 3 as well, and the correct answer is 2. The two disjoint cases differ in $\dim(\vec{A} \cap \vec{B})$, which the empty intersection cannot record. Use the second formula of Theorem 22.2.4 (c), and always check first whether the flats meet.

Quick check. In $\mathbb{A}^4$, let A and B be planes (flats of dimension 2) with $\vec{A} \cap \vec{B} = \{\mathbf{0}\}$. Must they meet?

Solution. Yes. By Theorem 1.7.3, $\dim(\vec{A} + \vec{B}) = 2 + 2 - 0 = 4 = \dim V$, so $\vec{A} + \vec{B} = V$ by Theorem 1.6.8. Hence $\overrightarrow{PQ} \in \vec{A} + \vec{B}$ for any $P \in A$, $Q \in B$, and Theorem 22.2.4 (b) gives $A \cap B \neq \emptyset$. (The dimension formula agrees: the disjoint case would predict $2+2-0+1 = 5 > 4$, which is impossible.) Their intersection is then a single point, since its direction space is $\vec{A} \cap \vec{B} = \{\mathbf{0}\}$.

22.2.3 Affine frames and barycentric coordinates

To put coordinates on an n-dimensional vector space we choose n vectors. To put coordinates on an n-dimensional affine space we must also say where the origin is, so we choose $n+1$ points — and if we refuse to single out one of them, the coordinates come out symmetric in all $n+1$. That symmetry is what makes the centroid computation below one line.

Chapter 18 §02 defined a list $(x_0, \dots, x_m)$ in a real vector space to be **affinely independent** when the only scalars with $\sum_i \lambda_i = 0$ and $\sum_i \lambda_i x_i = 0$ are all zero (Definition 18.2.1). By Definition 22.1.2 (b) the expression $\sum_i \lambda_i P_i$ is a well-defined vector for points of an affine space whenever $\sum_i \lambda_i = 0$, so the same words apply here verbatim. The next lemma is Lemma 18.2.1 (a) in that setting; its proof uses nothing about $\mathbb{R}$, so we give it again over an arbitrary field.

Lemma 22.2.5. Affine Independence by Differences

Let $P_0, P_1, \dots, P_m \in \mathcal{A}$. The following are equivalent.

(a) The only $\lambda_0, \dots, \lambda_m \in F$ with $\sum_{i=0}^m \lambda_i = 0$ and $\sum_{i=0}^m \lambda_i P_i = 0$ are $\lambda_0 = \cdots = \lambda_m = 0$.

(b) The list $(\overrightarrow{P_0P_1}, \dots, \overrightarrow{P_0P_m})$ is linearly independent in V.

A list satisfying these is called **affinely independent**. Condition (a) does not mention P_0 specially, so (b) holds with any P_j in place of P_0 as soon as it holds for one.

Proof. By Definition 22.1.2 (b) computed with origin P_0, and since $\overrightarrow{P_0P_0} = 0$,

$$\sum_{i=0}^{m} \lambda_i P_i = \sum_{i=0}^{m} \lambda_i \overrightarrow{P_0P_i} = \sum_{i=1}^{m} \lambda_i \overrightarrow{P_0P_i}.$$

Given $\lambda_1, \dots, \lambda_m$, the condition $\sum_{i=0}^{m} \lambda_i = 0$ holds for exactly one λ_0, namely $\lambda_0 = -\sum_{i \geq 1} \lambda_i$; and $(\lambda_0, \dots, \lambda_m)$ is the zero list exactly when $(\lambda_1, \dots, \lambda_m)$ is. So the relations forbidden by (a) correspond, one for one, to the nontrivial linear relations among $\overrightarrow{P_0P_1}, \dots, \overrightarrow{P_0P_m}$, and (a) holds if and only if there are none, which is (b).

■

Over $\mathbb{R}$ with $\mathcal{A} = V$ this is Lemma 18.2.1 (a), and Chapter 18's warning applies unchanged: affine independence is **not** linear independence. The list $(\mathbf{0}, \mathbf{e}_1, \mathbf{e}_2)$ in $\mathbb{R}^2$ is linearly dependent and affinely independent, which is why the count below is $n+1$ and not n.

Definition 22.2.4. Affine Frame

Let $\dim \mathcal{A} = n$. An **affine frame** of $\mathcal{A}$ is a list $(P_0, P_1, \dots, P_n)$ of $n+1$ affinely independent points of $\mathcal{A}$.

By Lemma 22.2.5 the list $(\overrightarrow{P_0P_1}, \dots, \overrightarrow{P_0P_n})$ is then independent with $n = \dim V$ entries, hence a basis of V (Theorem 1.6.4); conversely, if it is a basis, the frame condition holds. So an affine frame is "a point together with a basis", written symmetrically. In $\mathbb{A}^n(F)$ the **standard frame** is $(\mathbf{0}, \mathbf{e}_1, \dots, \mathbf{e}_n)$.

Theorem 22.2.6. Barycentric Coordinates

Let $\dim \mathcal{A} = n$ and let $(P_0, \dots, P_n)$ be an affine frame of $\mathcal{A}$. For every $X \in \mathcal{A}$ there are **unique** scalars $t_0, \dots, t_n \in F$ with

$$t_0 + t_1 + \cdots + t_n = 1 \qquad \text{and} \qquad X = \sum_{i=0}^{n} t_i P_i.$$

They are called the **barycentric coordinates** of X with respect to the frame.

Idea. Existence is the basis expansion of $\overrightarrow{P_0X}$, with t_0 invented to make the weights add up to 1. Uniqueness is affine independence: the difference of two weight lists is an affine dependence.

Proof. Existence. The list $(\overrightarrow{P_0P_1}, \dots, \overrightarrow{P_0P_n})$ is a basis of V, so there are unique $t_1, \dots, t_n \in F$ with $\overrightarrow{P_0X} = \sum_{i=1}^{n} t_i \overrightarrow{P_0P_i}$ (Theorem 1.5.1). Put $t_0 = 1 - (t_1 + \cdots + t_n)$, so that $\sum_{i=0}^{n} t_i = 1$. Computing the affine combination with origin P_0,

$$\sum_{i=0}^{n} t_i P_i = P_0 + \sum_{i=0}^{n} t_i \overrightarrow{P_0P_i} = P_0 + \overrightarrow{P_0X} = X.$$

Uniqueness. Suppose $\sum_i t_i P_i = X = \sum_i t'_i P_i$ with both weight lists summing to 1, and put $\lambda_i = t_i - t'_i$. Then $\sum_i \lambda_i = 0$, so $\sum_i \lambda_i P_i$ is the vector $\sum_i \lambda_i \overrightarrow{P_0 P_i}$ of Definition 22.1.2 (b), and computing both combinations with origin P_0 gives

$$\sum_i \lambda_i \overrightarrow{P_0 P_i} = \overrightarrow{P_0 X} - \overrightarrow{P_0 X} = 0.$$

By Lemma 22.2.5 (a), every $\lambda_i = 0$. This proves the theorem.

■

Barycentric coordinates come as a **list**, not as separate numbers: they are pinned down only by the requirement that they add up to 1, so no one of them means anything on its own. Their virtue is that the frame's points are treated alike: P_j has coordinates $(0, \dots, 1, \dots, 0)$ with the 1 in slot j, for every j.

Example. Reading a triangle

In $\mathbb{A}^2$ take the frame $P_0 = (0,0)$, $P_1 = (4,0)$, $P_2 = (0,3)$.

(a) Check that it is a frame, and find the barycentric coordinates (Theorem 22.2.6) of $X = (1,1)$.

(b) Find the point with barycentric coordinates $(\frac{1}{2}, \frac{1}{4}, \frac{1}{4})$.

(c) Describe the line $P_1 \vee P_2$ in barycentric coordinates.

Solution.

(a) $\overrightarrow{P_0 P_1} = (4,0)$ and $\overrightarrow{P_0 P_2} = (0,3)$ are independent, so by Lemma 22.2.5 the list is affinely independent, and three points in a plane make a frame. For $X = (1,1)$: from $(1,1) = t_1(4,0) + t_2(0,3)$ we get $t_1 = \frac{1}{4}$ and $t_2 = \frac{1}{3}$, so $t_0 = 1 - \frac{1}{4} - \frac{1}{3} = \frac{5}{12}$. The coordinates are $(\frac{5}{12}, \frac{1}{4}, \frac{1}{3})$. **Check:** $\frac{5}{12}(0,0) + \frac{1}{4}(4,0) + \frac{1}{3}(0,3) = (1,1)$, and $\frac{5}{12} + \frac{3}{12} + \frac{4}{12} = 1$.

(b) $\frac{1}{2}(0,0) + \frac{1}{4}(4,0) + \frac{1}{4}(0,3) = (1, \frac{3}{4})$.

(c) The join of the two distinct points P_1, P_2 is $P_1 + \operatorname{span}(\overrightarrow{P_1 P_2})$ by Theorem 22.2.4 (a), a line. Its points are the affine combinations $(1-t)P_1 + tP_2$, whose barycentric coordinates are $(0, 1-t, t)$; conversely a point with $t_0 = 0$ is such a combination. So $P_1 \vee P_2 = \{t_0 = 0\}$. The point X of (a) has $t_0 = \frac{5}{12} \neq 0$ and so is not on it.

The next result is the advertisement for the whole apparatus: a classical theorem of plane geometry, proved in one line of arithmetic with weights.

Proposition 22.2.7. The Medians Meet at the Centroid

Let F be a field in which $2 \neq 0$ and $3 \neq 0$, let $\mathcal{A}$ be an affine space over F, and let (P_1, P_2, P_3) be an affinely independent list in $\mathcal{A}$. Put

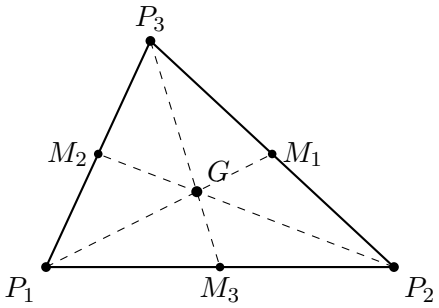
$$M_1 = \tfrac{1}{2}P_2 + \tfrac{1}{2}P_3, \qquad M_2 = \tfrac{1}{2}P_1 + \tfrac{1}{2}P_3, \qquad M_3 = \tfrac{1}{2}P_1 + \tfrac{1}{2}P_2,$$

the **midpoints** of the sides, and let $m_i = P_i \vee M_i$ be the **medians**. Then m_1, m_2, m_3 are

lines meeting in exactly one point, the **centroid**

$$G = \tfrac{1}{3}P_1 + \tfrac{1}{3}P_2 + \tfrac{1}{3}P_3,$$

and $\overrightarrow{P_iG} = 2\,\overrightarrow{GM_i}$ for each i.



Idea. Work in the plane $\pi = (P_1 \vee P_2) \vee P_3$, where (P_1, P_2, P_3) is an affine frame, and give every point its barycentric coordinates. A median is then the set of coordinate triples of one shape, and the three shapes have exactly one triple in common.

Proof. The three points are affinely independent, so $(\overrightarrow{P_1P_2}, \overrightarrow{P_1P_3})$ is independent by Lemma 22.2.5 and, by Theorem 22.2.4 (a) applied twice, $\pi := (P_1 \vee P_2) \vee P_3 = P_1 + \operatorname{span}(\overrightarrow{P_1P_2}, \overrightarrow{P_1P_3})$ is a plane in which (P_1, P_2, P_3) is an affine frame. All the points named lie in π, being affine combinations of P_1, P_2, P_3, so we may use the barycentric coordinates of Theorem 22.2.6 throughout; the weights $\frac{1}{2}$ and $\frac{1}{3}$ exist because 2 and 3 are nonzero, hence invertible, in F.

Each m_i is a line, and G lies on it. Take $i = 1$. In coordinates $P_1 = (1,0,0)$ and $M_1 = (0, \frac{1}{2}, \frac{1}{2})$, so $P_1 \neq M_1$ by the uniqueness in Theorem 22.2.6, and $m_1 = P_1 \vee M_1 = P_1 + \operatorname{span}(\overrightarrow{P_1M_1})$ by Theorem 22.2.4 (a), a flat of dimension 1 because $\overrightarrow{P_1M_1} \neq 0$. Its points are the affine combinations $(1-s)P_1 + sM_1$, $s \in F$, with coordinates

$$(1 - s,\ \tfrac{s}{2},\ \tfrac{s}{2}).$$

At $s = \frac{2}{3}$ this is $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$, the coordinate triple of G. So $G \in m_1$, and by the same computation with the indices permuted, $G \in m_2$ and $G \in m_3$.

No other common point. A point of $m_1 \cap m_2$ has coordinates $(1-s, \frac{s}{2}, \frac{s}{2})$ and $(\frac{s'}{2}, 1-s', \frac{s'}{2})$ for some $s, s' \in F$. Uniqueness in Theorem 22.2.6 makes the two triples equal entry by entry. The third entries give $s = s'$, and then the first two give $1 - s = \frac{s}{2}$, that is, $\frac{3}{2}s = 1$ and $s = \frac{2}{3}$. So $m_1 \cap m_2 = \{G\}$, and a fortiori $m_1 \cap m_2 \cap m_3 = \{G\}$.

The ratio. With origin P_1 we have $\overrightarrow{P_1M_1} = \frac{1}{2}\overrightarrow{P_1P_2} + \frac{1}{2}\overrightarrow{P_1P_3}$ and $\overrightarrow{P_1G} = \frac{1}{3}\overrightarrow{P_1P_2} + \frac{1}{3}\overrightarrow{P_1P_3}$, so $\overrightarrow{P_1G} = \frac{2}{3}\overrightarrow{P_1M_1}$ and $\overrightarrow{GM_1} = \overrightarrow{P_1M_1} - \overrightarrow{P_1G} = \frac{1}{3}\overrightarrow{P_1M_1}$. Hence $\overrightarrow{P_1G} = 2\overrightarrow{GM_1}$, and likewise for $i = 2, 3$. This proves the proposition.

■

Remark. The hypotheses on the field are not decoration. Over $\mathbb{F}_2$ there are no midpoints, because $\frac{1}{2}$ does not exist; over $\mathbb{F}_3$ there are midpoints but no centroid, because $\frac{1}{3}$ does not exist — and indeed in $\mathbb{A}^2(\mathbb{F}_3)$ the three medians of an affinely independent triple are parallel and pairwise disjoint, as Exercise C3 below computes.

22.2.4 An affine map is fixed by a frame

Barycentric coordinates make the next theorem immediate, and it is the affine counterpart of "a linear map is determined by its values on a basis".

Theorem 22.2.8. An Affine Map Is Determined by a Frame

Let $\mathcal{A}$ have dimension n with affine frame $(P_0, \dots, P_n)$, let $\mathcal{B}$ be an affine space over the same field with direction space W, and let $Q_0, \dots, Q_n \in \mathcal{B}$ be **any** points. Then there is exactly one affine map $f\colon \mathcal{A} \to \mathcal{B}$ with $f(P_i) = Q_i$ for $i = 0, \dots, n$. Moreover f carries the barycentric coordinates of Theorem 22.2.6 across:

$$f\Big(\sum_{i=0}^{n} t_i P_i\Big) = \sum_{i=0}^{n} t_i Q_i \qquad \Big(\sum_i t_i = 1\Big).$$

Proof. *Existence.* The list $(\overrightarrow{P_0P_1}, \dots, \overrightarrow{P_0P_n})$ is a basis of V, so by Theorem 2.1.6 there is a linear $T\colon V \to W$ with $T(\overrightarrow{P_0P_i}) = \overrightarrow{Q_0Q_i}$ for $i = 1, \dots, n$. Define $f(X) = Q_0 + T(\overrightarrow{P_0X})$; it is affine with $\vec{f} = T$ by Theorem 22.1.5 (c). Then $f(P_0) = Q_0 + T(0) = Q_0$, and $f(P_i) = Q_0 + \overrightarrow{Q_0Q_i} = Q_i$.

Uniqueness. Let g be affine with $g(P_i) = Q_i$ for all i. By Theorem 22.1.5 (b),

$$\vec{g}(\overrightarrow{P_0P_i}) = \overrightarrow{g(P_0)g(P_i)} = \overrightarrow{Q_0Q_i} = T(\overrightarrow{P_0P_i}) \qquad (i = 1, \dots, n),$$

so $\vec{g}$ and T agree on a basis and are equal (Theorem 2.1.5). Hence $g(X) = g(P_0) + \vec{g}(\overrightarrow{P_0X}) = Q_0 + T(\overrightarrow{P_0X}) = f(X)$ for every X.

The displayed identity is Definition 22.1.3 applied to $\sum_i t_i P_i$, using $f(P_i) = Q_i$.

■

So $n + 1$ points in the right position determine an affine map of an n-dimensional space, and prescribe it freely. Fewer points never do, and not because our proof is weak.

Proposition 22.2.9. Points inside a proper flat do not determine an affine map

Let $\dim \mathcal{A} = n \geq 1$ and let $S \subseteq \mathcal{A}$ be contained in a flat A with $\dim A < n$ — for instance, any set of at most n points. Then there are two **distinct** affine maps $\mathcal{A} \to \mathcal{A}$ that agree at every point of S.

Proof. Let $P \in A$ and let $(u_1, \dots, u_d)$ be a basis of $\vec{A}$, where $d = \dim A < n$. Extend it to a basis $(u_1, \dots, u_d, w_1, \dots, w_{n-d})$ of V (Theorem 1.6.5); since $d < n$, the vector w_1 exists. Let

$T\colon V \to V$ be the linear map fixing each u_j and each w_k with $k \geq 2$, and sending w_1 to 0 (Theorem 2.1.6), and put $g(X) = P + T(\overrightarrow{PX})$, an affine map by Theorem 22.1.5 (c).

For $X \in A$ we have $\overrightarrow{PX} \in \vec{A} = \operatorname{span}(u_1, \dots, u_d)$, on which T is the identity, so $g(X) = P + \overrightarrow{PX} = X$. Hence g agrees with $\mathrm{id}_{\mathcal{A}}$ on A, and so on S. But $g \neq \mathrm{id}_{\mathcal{A}}$, since $g(P + w_1) = P + T(w_1) = P \neq P + w_1$ by Lemma 22.1.1 (c).

Finally, a nonempty set of at most n points does lie in such a flat. Put

$$D = P_0 + \operatorname{span}(\overrightarrow{P_0P_1}, \dots, \overrightarrow{P_0P_m}).$$

This is the smallest flat containing $P_0, \dots, P_m$: it is a flat by Proposition 22.2.1 and contains every P_i; and any flat containing all the P_i contains P_0, hence equals $P_0 + U$ with each $\overrightarrow{P_0P_i} \in U$, so it contains D by Theorem 1.3.1. Thus $\dim D \leq m$, and $m + 1 \leq n$ gives $\dim D \leq n - 1 < n$. (An empty S lies in the flat $\{P\}$ for any P.)

■

Example (Two points are not enough in the plane) is an instance of this proposition, with $n = 2$ and $S = \{(0,0), (1,0)\}$; it exhibits its own pair of maps rather than the pair the proof constructs. The moral is the count: $n + 1$, never n. It is the same count that will govern projective transformations later in this chapter, where the answer turns out to be $n + 2$.

22.2.5 A flat is a solution set

Chapter 3 produced flats as solution sets; Chapter 5 proved that every subspace of F^n is a null space. Putting the two together identifies the flats of $\mathbb{A}^n(F)$ completely.

Theorem 22.2.10. Flats Are Exactly the Solution Sets

Let $A \subseteq \mathbb{A}^n(F)$ be nonempty and $0 \leq k \leq n$. The following are equivalent.

(a) A is a flat of dimension k.

(b) $A = \{\mathbf{x} \in F^n : \mathbf{M}\mathbf{x} = \mathbf{c}\}$ for some $\mathbf{M} \in M_{(n-k)\times n}(F)$ with linearly independent rows and some $\mathbf{c} \in F^{n-k}$.

In that case no system with fewer than $n - k$ equations has solution set A, and for $k < n$ the flat A is an intersection of exactly $n - k$ affine hyperplanes.

Idea. A flat is a coset $\mathbf{p} + U$ by Proposition 22.2.1, Chapter 5 turns U into a null space, and Chapter 3's "particular plus homogeneous" turns the coset into the solution set of the inhomogeneous system with the same matrix and right-hand side $\mathbf{M}\mathbf{p}$.

Proof.

(a) $\Rightarrow$ (b). By Proposition 22.2.1, $A = \mathbf{p} + U$ with $U = \vec{A}$ and $\dim U = k$. By Theorem 5.6.2 (a) there is $\mathbf{M} \in M_{(n-k)\times n}(F)$ with $\operatorname{nul}(\mathbf{M}) = U$ whose rows $\mathbf{a}_1^{\top}, \dots, \mathbf{a}_{n-k}^{\top}$ are the ones corresponding to a basis $(\varphi_{\mathbf{a}_1}, \dots, \varphi_{\mathbf{a}_{n-k}})$ of U^0 under the isomorphism $\mathbf{a} \mapsto \varphi_{\mathbf{a}}$ of Theorem 5.1.1; an injective linear map preserves independence (Theorem 2.2.5), so

those rows are linearly independent. Put $\mathbf{c} = \mathbf{M}\mathbf{p}$. Then $\mathbf{p}$ solves $\mathbf{M}\mathbf{x} = \mathbf{c}$, so by Theorem 3.1.3 the solution set is $\mathbf{p} + \operatorname{nul}(\mathbf{M}) = \mathbf{p} + U = A$.

(b) $\Rightarrow$ (a). The set A is nonempty, so it has an element $\mathbf{p}$, and Theorem 3.1.3 gives $A = \mathbf{p} + \operatorname{nul}(\mathbf{M})$, a coset, hence a flat by Proposition 22.2.1 with $\vec{A} = \operatorname{nul}(\mathbf{M})$. The rows of $\mathbf{M}$ are independent, so $\operatorname{rank}\mathbf{M} = \dim \operatorname{row}(\mathbf{M}) = n - k$ by Theorem 3.5.5, and Theorem 3.5.7 gives $\dim A = \operatorname{nullity}(\mathbf{M}) = n - (n - k) = k$.

Minimality. Suppose $\mathbf{B} \in M_{m\times n}(F)$ and $\mathbf{d} \in F^m$ satisfy $\{\mathbf{x} : \mathbf{B}\mathbf{x} = \mathbf{d}\} = A$. Fix $\mathbf{p} \in A$. For any $\mathbf{x}$, $\mathbf{B}\mathbf{x} = \mathbf{d}$ holds if and only if $\mathbf{B}(\mathbf{x} - \mathbf{p}) = \mathbf{0}$, so $\mathbf{x} \in A$ if and only if $\mathbf{x} - \mathbf{p} \in \operatorname{nul}(\mathbf{B})$; that is, $A = \mathbf{p} + \operatorname{nul}(\mathbf{B})$. Comparing with $A = \mathbf{p} + \vec{A}$ and using Proposition 22.2.1 (a), $\operatorname{nul}(\mathbf{B}) = \vec{A} = U$. By Theorem 5.6.2 (b), $m \geq n - k$.

Hyperplanes. Let $k < n$ and write $\mathbf{a}_1^\top, \dots, \mathbf{a}_{n-k}^\top$ for the rows of a matrix $\mathbf{M}$ as in (b). Each $\varphi_{\mathbf{a}_i}$ vanishes on $U = \operatorname{nul}(\mathbf{M})$, hence lies in U^0 (Definition 5.2.1), and the $\varphi_{\mathbf{a}_i}$ are independent because the rows are and $\mathbf{a} \mapsto \varphi_{\mathbf{a}}$ is an isomorphism (Theorem 5.1.1); as $\dim U^0 = n - k$ by Theorem 5.2.2, they are a basis of U^0 (Theorem 1.6.4). So $\mathbf{M}$ is a matrix as in Theorem 5.6.2 (a), and Proposition 5.6.4 applied to U and $\mathbf{p}$ writes $A = \mathbf{p} + U$ as the intersection of the $n - k$ affine hyperplanes $\{\mathbf{x} : \mathbf{a}_i^\top \mathbf{x} = \mathbf{a}_i^\top \mathbf{p}\}$.

■

For $k = n$ the system is empty and $A = F^n$; for $k = 0$ it has n independent equations and A is a single point. In between, the count is the familiar one: **dimension counts parameters, codimension counts independent equations**, and now it holds for flats as it did for subspaces.

Example. Equations for a plane in four-space

In Example (Equations for a plane in $\mathbb{R}^4$) the subspace $U = \operatorname{span}((1,1,0,1),(0,1,1,1))$ of $\mathbb{R}^4$ was found to be $\{\mathbf{x} : x_1 - x_2 + x_3 = 0,\ x_4 - x_2 = 0\}$. Find equations for the flat $A = \mathbf{p} + U$ with $\mathbf{p} = (1,0,0,0)$, and give two of its points.

Solution. Take $\mathbf{M} = \begin{pmatrix} 1 & -1 & 1 & 0 \\ 0 & -1 & 0 & 1 \end{pmatrix}$, whose null space is U, and $\mathbf{c} = \mathbf{M}\mathbf{p} = (1,0)$. By Theorem 22.2.10,

$$A = \{\mathbf{x} \in \mathbb{R}^4 : x_1 - x_2 + x_3 = 1, \quad x_4 - x_2 = 0\},$$

a flat of dimension $4 - 2 = 2$. Two points: $\mathbf{p} = (1,0,0,0)$, with $1 - 0 + 0 = 1$ and $0 - 0 = 0$; and $\mathbf{p} + (1,1,0,1) = (2,1,0,1)$, with $2 - 1 + 0 = 1$ and $1 - 1 = 0$. By the minimality clause, no single equation cuts out A, since $n - k = 2$.

That closes the affine part of the chapter's foundations. Flats are cosets, they meet or they do not, their joins obey a two-case formula, they carry symmetric coordinates, and in $\mathbb{A}^n(F)$ they are the solution sets of linear systems. The one irritant is the second case of the join formula, and the reason two coplanar lines can fail to meet: in the affine plane, parallel lines have nowhere to intersect. The projective sections later in this chapter remove that exception by adding the missing points.

22.2.6 Exercises

22.2.6.1 A. Check your understanding

Exercise. A1

(a) Define a flat and its direction space, and say why the direction space does not depend on the point used to compute it.

(b) State both cases of the dimension formula for $\dim(A \vee B)$, and say which one applies to two distinct parallel planes in $\mathbb{A}^3$.

(c) True or false, with a reason: the intersection of two flats is a flat.

(d) What are the barycentric coordinates of P_2 with respect to the frame (P_0, P_1, P_2, P_3) of a three-dimensional affine space?

Solution.

(a) A flat is a nonempty subset closed under affine combinations (Definition 22.2.1); equivalently a coset $P + U$ with U a subspace of V (Proposition 22.2.1). Its direction space is $\vec{A} = \{\overrightarrow{XY} : X, Y \in A\}$, a description that mentions only A, so no choice of point enters.

(b) If $A \cap B \neq \emptyset$, $\dim(A \vee B) = \dim A + \dim B - \dim(A \cap B)$; if $A \cap B = \emptyset$, $\dim(A \vee B) = \dim A + \dim B - \dim(\vec{A} \cap \vec{B}) + 1$. Two distinct parallel planes in $\mathbb{A}^3$ are disjoint (Proposition 22.2.3), so the second applies: $2 + 2 - 2 + 1 = 3$, and their join is $\mathbb{A}^3$.

(c) False as stated, because the intersection may be empty, and the empty set is not a flat (Definition 22.2.1). It is true that a **nonempty** intersection of flats is a flat (Proposition 22.2.2). Two parallel lines in $\mathbb{A}^2$ give the counterexample.

(d) $(0, 0, 1, 0)$: the combination $0 \cdot P_0 + 0 \cdot P_1 + 1 \cdot P_2 + 0 \cdot P_3$ has weights summing to 1 and equals P_2, and the coordinates are unique by Theorem 22.2.6.

22.2.6.2 B. Practice

Exercise. B1: Two flats in four-space

In $\mathbb{A}^4$ let $A = (1, 0, 0, 0) + \operatorname{span}\big((1, 1, 0, 0), (0, 0, 1, 1)\big)$, and let

$$B = (0, 1, 0, 0) + \operatorname{span}\big((1, 0, 1, 0)\big), \qquad B' = (2, 1, 0, 0) + \operatorname{span}\big((1, 0, 1, 0)\big).$$

For each of B and B', decide whether it meets A, and compute $\dim(A \vee \cdot)$. Justify your answers with Theorem 22.2.4.

Solution. Here $P = (1, 0, 0, 0)$, $\vec{A} = \operatorname{span}\big((1, 1, 0, 0), (0, 0, 1, 1)\big)$, $\vec{B} = \vec{B'} = \operatorname{span}\big((1, 0, 1, 0)\big)$. The three spanning vectors are independent (the first two have disjoint supports, and $a(1, 1, 0, 0) + b(0, 0, 1, 1) = (1, 0, 1, 0)$ would need $a = 1$ and $a = 0$), so $\dim(\vec{A} + \vec{B}) = 3$ and $\vec{A} \cap \vec{B} = \{\mathbf{0}\}$.

B. Here $Q = (0, 1, 0, 0)$ and $\overrightarrow{PQ} = (-1, 1, 0, 0)$. If $\overrightarrow{PQ} = a(1, 1, 0, 0) + b(0, 0, 1, 1) + c(1, 0, 1, 0)$, the second coordinate gives $a = 1$ and the first gives $a + c = -1$, so $c = -2$; the third then gives $b + c = 0$, so $b = 2$, and the fourth gives $b = 0$, a contradiction. So $\overrightarrow{PQ} \notin \vec{A} + \vec{B}$ and

$A \cap B = \emptyset$ by part (b) of the theorem. Part (c) gives $\dim(A \vee B) = 2 + 1 - 0 + 1 = 4$, so $A \vee B = \mathbb{A}^4$.

B'. Here $Q' = (2,1,0,0)$ and $\overrightarrow{PQ'} = (1,1,0,0) \in \vec{A} \subseteq \vec{A} + \vec{B'}$, so $A \cap B' \neq \emptyset$. Following the proof of part (b), the meeting point is $P + (1,1,0,0) = (2,1,0,0) = Q'$, and indeed $Q' \in A \cap B'$. By part (b) the direction space of $A \cap B'$ is $\vec{A} \cap \vec{B'} = \{\mathbf{0}\}$, so the intersection is the single point $\{Q'\}$, of dimension 0, and $\dim(A \vee B') = 2 + 1 - 0 = 3$.

Exercise. B2: Barycentric arithmetic

In $\mathbb{A}^2$ take the frame $P_0 = (1,1)$, $P_1 = (3,1)$, $P_2 = (1,4)$.

(a) Verify that it is a frame.

(b) Find the barycentric coordinates (Theorem 22.2.6) of $X = (2,2)$.

(c) Find the centroid of (P_0, P_1, P_2) in ordinary coordinates, and check its barycentric coordinates.

Solution.

(a) $\overrightarrow{P_0P_1} = (2,0)$ and $\overrightarrow{P_0P_2} = (0,3)$ are independent, so the list is affinely independent by Lemma 22.2.5, and it has $3 = 2 + 1$ entries.

(b) $\overrightarrow{P_0X} = (1,1) = t_1(2,0) + t_2(0,3)$ gives $t_1 = \frac{1}{2}$, $t_2 = \frac{1}{3}$, so $t_0 = 1 - \frac{1}{2} - \frac{1}{3} = \frac{1}{6}$. The coordinates are $(\frac{1}{6}, \frac{1}{2}, \frac{1}{3})$. **Check:** $\frac{1}{6}(1,1) + \frac{1}{2}(3,1) + \frac{1}{3}(1,4) = (\frac{1}{6} + \frac{3}{2} + \frac{1}{3},\ \frac{1}{6} + \frac{1}{2} + \frac{4}{3}) = (2,2)$.

(c) $G = \frac{1}{3}((1,1) + (3,1) + (1,4)) = (\frac{5}{3}, 2)$. Its barycentric coordinates are $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ by definition, and by uniqueness these are the only ones. **Check against (b)'s method:** $\overrightarrow{P_0G} = (\frac{2}{3}, 1) = \frac{1}{3}(2,0) + \frac{1}{3}(0,3)$, so $t_1 = t_2 = \frac{1}{3}$ and $t_0 = \frac{1}{3}$.

Exercise. B3: From parameters to equations and back

(a) Find a system of equations whose solution set is the line $A = (1,2,3) + \operatorname{span}((1,1,1))$ in $\mathbb{A}^3$, using as few equations as possible.

(b) The set $B = \{\mathbf{x} \in \mathbb{R}^3 : x_1 + x_2 + x_3 = 6\}$ is a flat. Find its dimension, a point of it, and its direction space.

(c) Decide whether $A \subseteq B$, and compute $\dim(A \vee B)$.

Solution.

(a) Here $k = 1$ and $n = 3$, so Theorem 22.2.10 calls for $3 - 1 = 2$ equations, and no fewer will do. A functional $\mathbf{a}^\top \mathbf{x}$ vanishes on $\operatorname{span}((1,1,1))$ exactly when $a_1 + a_2 + a_3 = 0$; two independent such rows are $(1,-1,0)$ and $(0,1,-1)$. With $\mathbf{p} = (1,2,3)$ the right-hand sides are $1 - 2 = -1$ and $2 - 3 = -1$, so

$$A = \{\mathbf{x} : x_1 - x_2 = -1,\ x_2 - x_3 = -1\}.$$

Check: $\mathbf{p}$ satisfies both, and so does $\mathbf{p} + (1,1,1) = (2,3,4)$.

(b) One equation with a nonzero row, so $\dim B = 3 - 1 = 2$ by Theorem 22.2.10, a plane. The point $(6, 0, 0)$ lies in it, and $\vec{B} = \operatorname{nul}\begin{pmatrix} 1 & 1 & 1 \end{pmatrix} = \{\mathbf{x} : x_1 + x_2 + x_3 = 0\}$.

(c) $\mathbf{p} = (1, 2, 3)$ has $1 + 2 + 3 = 6$, so $\mathbf{p} \in B$; but $(1, 1, 1) \notin \vec{B}$, since $1 + 1 + 1 = 3 \neq 0$. So $A \not\subseteq B$, while $A \cap B \neq \emptyset$ (it contains $\mathbf{p}$). Since $\vec{A} \cap \vec{B} = \{\mathbf{0}\}$, the intersection is the single point $\{\mathbf{p}\}$ and $\dim(A \vee B) = 1 + 2 - 0 = 3$: the line and the plane together span $\mathbb{A}^3$.

22.2.6.3 C. Going deeper

Exercise. C1: Affine maps and flats

Let $f \colon \mathcal{A} \to \mathcal{B}$ be affine and let $A \subseteq \mathcal{A}$ be a flat.

(a) Prove that $f(A)$ is a flat of $\mathcal{B}$ whose direction space is $\vec{f}(\vec{A})$, and deduce $\dim f(A) \leq \dim A$.

(b) Prove that if $B \subseteq \mathcal{B}$ is a flat with $f^{-1}(B) \neq \emptyset$, then $f^{-1}(B)$ is a flat with direction space $(\vec{f})^{-1}(\vec{B})$.

(c) Hence prove that an **injective** affine map sends flats to flats of the same dimension, and lines to lines.

Solution.

(a) $f(A)$ is nonempty. Let $P \in A$, so $A = P + \vec{A}$ (Proposition 22.2.1). For $u \in \vec{A}$, Theorem 22.1.5 (c) gives $f(P + u) = f(P) + \vec{f}(u)$. Hence $f(A) = f(P) + \vec{f}(\vec{A})$, and $\vec{f}(\vec{A})$ is a subspace of W, being the image of the linear map $\vec{f}$ restricted to the subspace $\vec{A}$ (Theorem 2.2.2 applied to that restriction). So $f(A)$ is a flat by Proposition 22.2.1, with the stated direction space, and $\dim f(A) = \dim \vec{f}(\vec{A}) \leq \dim \vec{A} = \dim A$, because the images of a basis of $\vec{A}$ span $\vec{f}(\vec{A})$ (Theorem 2.2.4) and a spanning list is at least as long as the dimension (Theorem 1.6.3 (b)).

(b) Let $P \in f^{-1}(B)$, so $f(P) \in B$. For $u \in V$, $f(P+u) = f(P)+\vec{f}(u)$ lies in $B = f(P)+\vec{B}$ exactly when $\vec{f}(u) \in \vec{B}$. So $f^{-1}(B) = P + (\vec{f})^{-1}(\vec{B})$, and $(\vec{f})^{-1}(\vec{B})$ is a subspace: it contains 0 and is closed under linear combinations by linearity of $\vec{f}$ and the subspace test (Theorem 1.2.1). By Proposition 22.2.1, $f^{-1}(B)$ is a flat with that direction space.

(c) If f is injective, then $\vec{f}$ is injective: $\vec{f}(u) = 0$ gives $f(P + u) = f(P)$, hence $P + u = P$ and $u = 0$. An injective linear map carries a basis of $\vec{A}$ to a linearly independent list (Theorem 2.2.5) that spans $\vec{f}(\vec{A})$, hence to a basis of it; so $\dim f(A) = \dim \vec{f}(\vec{A}) = \dim \vec{A} = \dim A$ by (a). A line has dimension 1, so its image is a flat of dimension 1, a line.

Exercise. C2: The midpoint quadrilateral

Let F be a field with $2 \neq 0$, let $\mathcal{A}$ be an affine space over F, and let $P_1, P_2, P_3, P_4 \in \mathcal{A}$ be any four points. Put $M_{12} = \frac{1}{2}P_1 + \frac{1}{2}P_2$, and define M_{23}, M_{34}, M_{41} likewise.

(a) Prove that $\overrightarrow{M_{12}M_{23}} = \overrightarrow{M_{41}M_{34}} = \frac{1}{2}\overrightarrow{P_1P_3}$.

(b) Deduce that if $M_{12}, M_{23}, M_{34}, M_{41}$ are distinct, then $M_{12} \vee M_{23}$ and $M_{41} \vee M_{34}$ are parallel lines, and likewise for the other pair of opposite sides: the four midpoints form a parallelogram.

(c) What does the statement say when P_1, P_2, P_3, P_4 are **not** coplanar, for instance in $\mathbb{A}^3$?

Solution.

(a) Fix any origin O and write $p_i = \overrightarrow{OP_i}$. By Definition 22.1.2, $\overrightarrow{OM_{12}} = \frac{1}{2}(p_1 + p_2)$ and similarly for the others. Hence, by Proposition 22.1.3,

$$\overrightarrow{M_{12}M_{23}} = \tfrac{1}{2}(p_2 + p_3) - \tfrac{1}{2}(p_1 + p_2) = \tfrac{1}{2}(p_3 - p_1) = \tfrac{1}{2}\overrightarrow{P_1P_3},$$

and

$$\overrightarrow{M_{41}M_{34}} = \tfrac{1}{2}(p_3 + p_4) - \tfrac{1}{2}(p_4 + p_1) = \tfrac{1}{2}(p_3 - p_1).$$

The answer does not depend on O, as it must not.

(b) If $M_{12} \neq M_{23}$, the join $M_{12} \vee M_{23}$ is the line $M_{12} + \operatorname{span}(\overrightarrow{M_{12}M_{23}})$ by Theorem 22.2.4 (a). By (a) the two lines $M_{12} \vee M_{23}$ and $M_{41} \vee M_{34}$ have the same direction space $\operatorname{span}(\frac{1}{2}\overrightarrow{P_1P_3})$, so they are parallel (Definition 22.2.3). Running (a) with the indices shifted by one gives $\overrightarrow{M_{23}M_{34}} = \overrightarrow{M_{12}M_{41}} = \frac{1}{2}\overrightarrow{P_2P_4}$, so the other two sides are parallel too.

(c) Nothing in (a) used coplanarity, so the conclusion holds for a skew quadrilateral in $\mathbb{A}^3$ as well: the four midpoints of a skew quadrilateral are coplanar and form a parallelogram. Four noncoplanar points are affinely independent: otherwise $(\overrightarrow{P_1P_2}, \overrightarrow{P_1P_3}, \overrightarrow{P_1P_4})$ would be dependent (Lemma 22.2.5) and span a subspace of dimension at most 2, putting all four points in a flat of dimension at most 2. The four midpoints are then distinct, since the difference of two of them is $\sum_i \lambda_i P_i$ with $\sum_i \lambda_i = 0$ and the λ_i not all zero — for M_{12} and M_{23} the weights are $\frac{1}{2}$ at P_1 and $-\frac{1}{2}$ at P_3 — hence is a nonzero vector; so (b) applies. The two parallel lines $M_{12} \vee M_{23}$ and $M_{41} \vee M_{34}$ are also distinct: were they equal, their common direction space would contain $\overrightarrow{M_{12}M_{41}}$, which is $\frac{1}{2}\overrightarrow{P_2P_4}$ by (b), so $\frac{1}{2}\overrightarrow{P_2P_4} = \mu \cdot \frac{1}{2}\overrightarrow{P_1P_3}$ for some $\mu \in F$ (Proposition 22.2.1 (a)) — a relation $\sum_i \lambda_i P_i = 0$ with weights summing to 0 and $-\frac{1}{2} \neq 0$ at P_2, again against affine independence. Their join is therefore a plane, of dimension $1 + 1 - 1 + 1 = 2$ by Theorem 22.2.4 (c), and it contains all four midpoints.

Exercise. C3: Medians over the field with three elements

Work in $\mathbb{A}^2(\mathbb{F}_3)$ with the affinely independent list $P_1 = (0,0)$, $P_2 = (1,0)$, $P_3 = (0,1)$. (Nothing is lost by this choice: by Theorem 22.2.8 any other affinely independent triple is carried to this one by an affine bijection, which preserves midpoints, being an affine combination, and carries lines to lines by Exercise (C1: Affine maps and flats).)

(a) Compute the three midpoints M_i and the three medians $m_i = P_i \vee M_i$.

(b) Prove that the three medians are pairwise parallel and pairwise disjoint, so that Proposition 22.2.7 fails over $\mathbb{F}_3$.

(c) Explain which step of the proof of Proposition 22.2.7 breaks, and why the hypothesis $3 \neq 0$ cannot be dropped.

Solution.

(a) In $\mathbb{F}_3$ we have $2^{-1} = 2$, since $2 \cdot 2 = 4 = 1$. So

$$\begin{aligned} M_1 &= 2P_2 + 2P_3 = 2(1,0) + 2(0,1) = (2,2), \\ M_2 &= 2P_1 + 2P_3 = (0,2), \\ M_3 &= 2P_1 + 2P_2 = (2,0). \end{aligned}$$

Then $\overrightarrow{P_1M_1} = (2,2)$, $\overrightarrow{P_2M_2} = (0,2) - (1,0) = (2,2)$ and $\overrightarrow{P_3M_3} = (2,0) - (0,1) = (2,2)$. So

$$\begin{aligned} m_1 &= (0,0) + \operatorname{span}((2,2)), \\ m_2 &= (1,0) + \operatorname{span}((2,2)), \\ m_3 &= (0,1) + \operatorname{span}((2,2)). \end{aligned}$$

(b) All three have the same direction space $\operatorname{span}((2,2)) = \operatorname{span}((1,1))$, so they are parallel (Definition 22.2.3). They are distinct: $\overrightarrow{P_1P_2} = (1,0) \notin \operatorname{span}((1,1))$, and likewise for the other two pairs, so no two of them share a point by Proposition 22.2.1 (b). Distinct parallel flats are disjoint (Proposition 22.2.3). Hence the three medians have no common point, and the conclusion of Proposition 22.2.7 is false here.

(c) The proof solved $\frac{3}{2}s = 1$ for s, which needs 3 to be invertible. Over $\mathbb{F}_3$ the point $\frac{1}{3}(P_1 + P_2 + P_3)$ does not exist, because $3 = 0$ has no inverse; the weights $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ cannot be written down at all. Concretely, the equation $1 - s = \frac{s}{2}$ becomes $1 - s = 2s$, that is, $1 = 3s = 0$, which has no solution: the parametrizations of m_1 and m_2 never agree. So the hypothesis $3 \neq 0$ is not an artifact of the proof; it is needed for the statement.

22.3 Convex Polyhedra, Geometrically

Chapter 18 built the theory of convex sets and proved its main theorems: that a compact convex set is the hull of its extreme points, that a bounded polyhedron and a polytope are the same thing, and that a feasible, bounded linear program and its dual have the same optimal value. **This section proves none of that again.** It reads those theorems as statements about the geometry of a solid with flat sides: what its faces are, how they fit together, and what a linear program looks like when the objective is a family of parallel hyperplanes sliding across it. Everything here is a corollary of Chapter 18 or a short argument on top of it, and every citation points back there.

Throughout, the field is $\mathbb{R}$ and the space is $\mathbb{R}^n$ with the dot product, as in Chapter 18. A polyhedron is a set $P = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{A}\mathbf{x} \leq \mathbf{b}\}$ (Definition 18.9.2), with $\mathbf{A} \in M_{m\times n}(\mathbb{R})$ whose rows we write $\mathbf{a}_1^\top, \dots, \mathbf{a}_m^\top$, so that $\mathbf{A}\mathbf{x} \leq \mathbf{b}$ is the list of inequalities $\mathbf{a}_i^\top \mathbf{x} \leq b_i$. A **vertex** of P is an extreme point of it (Definition 18.8.1), and $\operatorname{ext} P$ is the set of them. We use Chapters 16 and 18's topological words — closed, bounded, compact — only inside the results we cite; nothing here develops them further.

22.3.1 Faces, cut by a supporting hyperplane

A cube has six square sides, twelve edges, and eight corners, and each of them is the part of the cube touched by a plane that has the whole cube on one side. That is the picture to formalize. The tool is Chapter 18's supporting hyperplane (Theorem 18.3.6): a hyperplane through a point of a convex set with the set in one of its closed half-spaces.

A face of a polyhedron is the part of it touched by a hyperplane that does not cut into it.

Definition 22.3.1. Face of a Polyhedron

Let $P \subseteq \mathbb{R}^n$ be a polyhedron. A **face** of P is a set of the form

$$F = \{\mathbf{x} \in P : \mathbf{c}^\top \mathbf{x} = \beta\}$$

for some $\mathbf{c} \in \mathbb{R}^n$ and $\beta \in \mathbb{R}$ such that $\mathbf{c}^\top \mathbf{x} \leq \beta$ **for every** $\mathbf{x} \in P$. A face other than P itself is called **proper**.

Clause by clause. The pair $(\mathbf{c}, \beta)$ must be a **valid inequality** for P: the whole polyhedron lies in the closed half-space $\{\mathbf{c}^\top \mathbf{x} \leq \beta\}$. The face is then what the boundary hyperplane $\{\mathbf{c}^\top \mathbf{x} = \beta\}$ actually touches. Note that $\mathbf{c}$ is allowed to be $\mathbf{0}$.

Two faces come free. Taking $\mathbf{c} = \mathbf{0}$ and $\beta = 0$ makes the inequality $0 \leq 0$, valid everywhere, and $F = P$: **the whole polyhedron is a face of itself**. Taking $\mathbf{c} = \mathbf{0}$ and $\beta = 1$ makes the inequality $0 \leq 1$, again valid, and $F = \emptyset$: **the empty set is a face**. These two are the degenerate cases, and they are kept because the theorems below are cleaner with them than without.

Chapter 18 §08 already has a notion of face for an arbitrary convex set, Definition 18.8.2: a convex subset $F \subseteq C$ that swallows every segment of C passing through it at an inner point. The two agree here in one direction at once.

Remark. Every face in the sense of Definition 22.3.1 is a face in the sense of Definition 18.8.2. That is exactly Proposition 18.8.4 (a), applied with $\mathbf{w} = \mathbf{c}$ and $\alpha = \beta$, since on $\mathbb{R}^n$ the inner product $\langle \mathbf{c}, \mathbf{x} \rangle$ is $\mathbf{c}^\top \mathbf{x}$. So everything Chapter 18 proves about faces applies to ours; in particular Lemma 18.8.3 gives $\operatorname{ext} F = F \cap \operatorname{ext} P$. The converse — that every face of a **polyhedron** in Chapter 18's sense is cut out by a supporting hyperplane — is also true, and Exercise (C2: The two notions of face agree for a polyhedron) proves it. It uses flatness: the same exercise ends with a round convex set having a one-point face in the sense of Definition 18.8.2 that no supporting hyperplane isolates, so the converse fails for general convex sets.

Examples, in $\mathbb{R}^2$, with P the triangle $\{x_1 \geq 0,\ x_2 \geq 0,\ x_1 + x_2 \leq 3\}$ of Chapter 18 §09.

1. $\mathbf{c} = (1, 1)$, $\beta = 3$. The inequality is valid by the third row, and the face is the slanted edge from $(3, 0)$ to $(0, 3)$.
2. $\mathbf{c} = (1, 0)$, $\beta = 3$. Valid, since $x_1 \leq x_1 + x_2 \leq 3$ on P. The face is the single point $(3, 0)$: the hyperplane $x_1 = 3$ touches the triangle at one corner.
3. $\mathbf{c} = (-1, 0)$, $\beta = 0$. Valid by the first row, and the face is the vertical edge $\{0\} \times [0, 3]$.
4. $\mathbf{c} = (1, 1)$, $\beta = 4$. Valid, but never attained: the face is $\emptyset$.

A nonexample by minimal change. In example 1 replace $\beta = 3$ by $\beta = 2$. The set $\{\mathbf{x} \in P : x_1 + x_2 = 2\}$ is still a segment inside P, still convex, and still a slice by a hyperplane. What fails is the one clause the definition insists on: $x_1 + x_2 \leq 2$ is **not** valid for P, since $(3, 0) \in P$ violates it. The hyperplane cuts through the triangle instead of resting against it, and the slice is not a face. It also fails Definition 18.8.2: the segment from $(0, 0)$ to $(3, 0)$ passes through $(2, 0)$ at an inner point but is not contained in the slice.

The reason faces are useful is that there are only finitely many of them, and each is again a polyhedron described by the **same** rows, some of them turned into equations.

Proposition 22.3.1. Faces Are Polyhedra, and There Are Finitely Many

Let $P = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{A}\mathbf{x} \le \mathbf{b}\}$ with m rows. For $J \subseteq \{1, \dots, m\}$ put

$$P_J = \{\mathbf{x} \in P : \mathbf{a}_i^\top \mathbf{x} = b_i \text{ for every } i \in J\}.$$

Then:

(a) every P_J is a face of P, and it is a polyhedron;

(b) every **nonempty** face of P equals P_J for some J;

(c) consequently P has at most 2^m nonempty faces, and every face of P is a polyhedron.

Idea. For (a), add up the rows indexed by J: the sum $\sum_{i\in J} \mathbf{a}_i^\top \mathbf{x}$ is at most $\sum_{i\in J} b_i$ on P, and a sum of numbers each at most its bound hits the total bound only if every one of them does. So one valid inequality encodes all of J at once.

For (b), let F be a nonempty face and let J collect the rows that are tight at **every** point of F. Certainly $F \subseteq P_J$. For the other inclusion we need a point $\mathbf{x}^*$ of F at which every row outside J is **slack**; averaging finitely many points of F, one per slack row, produces it. Then, given $\mathbf{y} \in P_J$ with $\mathbf{c}^\top \mathbf{y} < \beta$, step from $\mathbf{x}^*$ **away** from $\mathbf{y}$. The rows in J are equations satisfied by both points, so they survive the step exactly; the rows outside J have room at $\mathbf{x}^*$, so they survive a short step. But the step raises $\mathbf{c}^\top \mathbf{x}$ above β, which no point of P may do.

Proof.

(a) Put $\mathbf{c} = \sum_{i\in J} \mathbf{a}_i$ and $\beta = \sum_{i\in J} b_i$, with the empty sums $\mathbf{0}$ and 0. For $\mathbf{x} \in P$,

$$\mathbf{c}^\top \mathbf{x} = \sum_{i\in J} \mathbf{a}_i^\top \mathbf{x} \le \sum_{i\in J} b_i = \beta,$$

so the inequality is valid. If $\mathbf{x} \in P$ and $\mathbf{c}^\top \mathbf{x} = \beta$, then $\sum_{i\in J}(b_i - \mathbf{a}_i^\top \mathbf{x}) = 0$ is a sum of nonnegative numbers, so each is zero and $\mathbf{x} \in P_J$; conversely every $\mathbf{x} \in P_J$ has $\mathbf{c}^\top \mathbf{x} = \beta$. Hence $P_J = \{\mathbf{x} \in P : \mathbf{c}^\top \mathbf{x} = \beta\}$ is a face. It is a polyhedron because each equation $\mathbf{a}_i^\top \mathbf{x} = b_i$ is the pair of inequalities $\mathbf{a}_i^\top \mathbf{x} \le b_i$ and $-\mathbf{a}_i^\top \mathbf{x} \le -b_i$, so P_J is cut out by $m + |J|$ inequalities.

(b) Let $F = \{\mathbf{x} \in P : \mathbf{c}^\top \mathbf{x} = \beta\}$ be nonempty, with $\mathbf{c}^\top \mathbf{x} \le \beta$ on P. Put

$$J = \{i : \mathbf{a}_i^\top \mathbf{x} = b_i \text{ for every } \mathbf{x} \in F\}.$$

Then $F \subseteq P_J$ by the definition of J, and it remains to prove $P_J \subseteq F$.
First note that F is convex: it is the intersection of the convex set P (Proposition 18.9.1) with the solution set of a linear equation, which is convex as the preimage of a point under $\mathbf{x} \mapsto \mathbf{c}^\top \mathbf{x}$ (Proposition 18.1.1 (b)); an intersection of convex sets is convex by Proposition 18.1.1 (a).
Step 1: a point of F with every row outside J slack. For each $i \notin J$ choose, by the definition of J, a point $\mathbf{x}_i \in F$ with $\mathbf{a}_i^\top \mathbf{x}_i < b_i$. If $J = \{1, \dots, m\}$ let $\mathbf{x}^*$ be any point of F. Otherwise let $k \ge 1$ be the number of indices outside J and let $\mathbf{x}^* = \frac{1}{k}\sum_{i\notin J} \mathbf{x}_i$, which lies in F by convexity. For each $i \notin J$, every term of $\mathbf{a}_i^\top \mathbf{x}^* = \frac{1}{k}\sum_{j\notin J} \mathbf{a}_i^\top \mathbf{x}_j$ is at most b_i and the term $j = i$ is strictly less, so $\mathbf{a}_i^\top \mathbf{x}^* < b_i$.

Step 2: the step away from $\mathbf{y}$. Let $\mathbf{y} \in P_J$ and suppose, for a contradiction, that $\mathbf{c}^\top \mathbf{y} < \beta$. For $t > 0$ put $\mathbf{z}_t = \mathbf{x}^* + t(\mathbf{x}^* - \mathbf{y})$. For $i \in J$ we have $\mathbf{a}_i^\top \mathbf{x}^* = b_i$, because $\mathbf{x}^* \in F \subseteq P_J$, and $\mathbf{a}_i^\top \mathbf{y} = b_i$, because $\mathbf{y} \in P_J$; hence

$$\mathbf{a}_i^\top \mathbf{z}_t = b_i + t(b_i - b_i) = b_i \le b_i.$$

For $i \notin J$ put $\delta = \min_{i \notin J}(b_i - \mathbf{a}_i^\top \mathbf{x}^*) > 0$ and $M = \max_{i \notin J} |\mathbf{a}_i^\top \mathbf{x}^* - \mathbf{a}_i^\top \mathbf{y}|$; both are minima and maxima of finitely many numbers. Choose $t > 0$ with $tM < \delta$, which is possible (any t will do if $M = 0$). Then for $i \notin J$,

$$\mathbf{a}_i^\top \mathbf{z}_t = \mathbf{a}_i^\top \mathbf{x}^* + t(\mathbf{a}_i^\top \mathbf{x}^* - \mathbf{a}_i^\top \mathbf{y}) \le \mathbf{a}_i^\top \mathbf{x}^* + tM < \mathbf{a}_i^\top \mathbf{x}^* + \delta \le b_i.$$

So $\mathbf{z}_t \in P$. But

$$\mathbf{c}^\top \mathbf{z}_t = \mathbf{c}^\top \mathbf{x}^* + t(\mathbf{c}^\top \mathbf{x}^* - \mathbf{c}^\top \mathbf{y}) = \beta + t(\beta - \mathbf{c}^\top \mathbf{y}) > \beta,$$

since $t > 0$ and $\beta - \mathbf{c}^\top \mathbf{y} > 0$. This contradicts the validity of $\mathbf{c}^\top \mathbf{x} \le \beta$ on P. Hence $\mathbf{c}^\top \mathbf{y} = \beta$ and $\mathbf{y} \in F$, so $P_J \subseteq F$ and $F = P_J$.

(c) There are 2^m subsets J, so by (b) at most that many nonempty faces. Each of them is a polyhedron by (a), and the empty face is the polyhedron $\{\mathbf{x} : 0 \le -1\}$. This proves the proposition.

■

The map $J \mapsto P_J$ is surjective onto the nonempty faces but is very far from injective: adding to J a row that is already tight everywhere on P_J changes nothing, and adding two incompatible rows gives $\emptyset$.

Warning. **A row of the system need not be a side of the solid.** In $P = \{\mathbf{x} \in \mathbb{R}^2 : x_1 \le 1,\ x_2 \le 1,\ x_1 + x_2 \le 3\}$ the third row is implied by the first two, since $x_1 + x_2 \le 2$ on P. It defines no face except the empty one: $P_{\{3\}} = \{\mathbf{x} \in P : x_1 + x_2 = 3\} = \emptyset$. So a polyhedron with m rows can have far fewer than m sides, and counting rows is not counting geometry. Conversely, Definition 22.3.1 never mentions the rows at all, which is why Proposition 22.3.1 has to prove that the faces can be read off them.

22.3.2 Vertices, edges, facets

A nonempty face F is a nonempty convex set, so it has a dimension: $\dim F = \dim \operatorname{aff} F$, as in Chapter 18 §01, and $\operatorname{aff} F$ is the flat $\mathbf{p} + U$ of Proposition 18.1.4 — a flat in the sense of Definition 22.2.1. The graded vocabulary follows.

Definition 22.3.2. Edge, Facet

Let $P \subseteq \mathbb{R}^n$ be a nonempty polyhedron. A face of P of dimension 1 is an **edge**; a face of dimension $\dim P - 1$ is a **facet**. A face of dimension 0 is a single point, and the next proposition shows that these are exactly the vertices.

Proposition 22.3.2. Three Descriptions of a Corner

Let $P = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{A}\mathbf{x} \leq \mathbf{b}\}$ and $\mathbf{v} \in P$. The following are equivalent.

(a) $\mathbf{v}$ is a vertex of P, that is, an extreme point.

(b) $\{\mathbf{v}\}$ is a face of P.

(c) Among the rows active at $\mathbf{v}$ there are n linearly independent ones, that is, $\mathbf{v}$ is a basic feasible point.

Idea.

(a) and (c) are Theorem 18.9.2, already proved. A one-point face is an extreme point by Proposition 18.8.4 (b). For the remaining implication, take J to be the active set: if the active rows have rank n, the system $\mathbf{A}_J\mathbf{x} = \mathbf{b}_J$ has $\mathbf{v}$ as its only solution, so $P_J = \{\mathbf{v}\}$, and Proposition 22.3.1 makes that a face.

Proof.

(a) $\Leftrightarrow$ (c) is Theorem 18.9.2.

(b) $\Rightarrow$ (a). Write $\{\mathbf{v}\} = \{\mathbf{x} \in P : \mathbf{c}^\top\mathbf{x} = \beta\}$ with $\mathbf{c}^\top\mathbf{x} \leq \beta$ on P. By Proposition 18.8.4 (b), applied with $\mathbf{w} = \mathbf{c}$ and $\alpha = \beta$, $\mathbf{v} \in \operatorname{ext} P$.

(c) $\Rightarrow$ (b). Let $I = I(\mathbf{v})$ be the active set and $\mathbf{A}_I$ the matrix of active rows, of rank n by hypothesis. Its null space is $\{\mathbf{0}\}$ by Theorem 3.5.7, so $\mathbf{A}_I\mathbf{x} = \mathbf{b}_I$ has at most one solution; since $\mathbf{v}$ is one, it is the only one. Hence

$$P_I = \{\mathbf{x} \in P : \mathbf{A}_I\mathbf{x} = \mathbf{b}_I\} = \{\mathbf{v}\},$$

and P_I is a face by Proposition 22.3.1 (a). This proves the equivalence.

■

So a corner of a polyhedron can be recognized three ways: by convexity (it is not an average), by geometry (a hyperplane touches P there and nowhere else), or by algebra (enough constraints are tight). The three are used interchangeably from here on.

Bounded polyhedra are the ones with enough corners to rebuild them. This is the bounded half of the Minkowski–Weyl theorem, and it is the sentence this whole section is arranged around.

Theorem 22.3.3. A Bounded Polyhedron Is the Hull of Its Vertices

Let $P \subseteq \mathbb{R}^n$ be a bounded polyhedron. Then $\operatorname{ext} P$ is finite and

$$P = \operatorname{conv}(\operatorname{ext} P).$$

Conversely, every polytope is a bounded polyhedron.

Proof. If $P = \emptyset$, both sides are $\emptyset$ and $\operatorname{ext} P = \emptyset$. Otherwise P is closed and convex by Proposition 18.9.1, and bounded by hypothesis, so it is compact by Corollary 16.2.5. By Theorem 18.8.5, $P = \operatorname{conv}(\operatorname{ext} P)$, and $\operatorname{ext} P$ is finite by Corollary 18.9.3. The converse is Theorem 18.9.5 (a).

■

The same theorem applied to a face is what makes an edge a segment between two corners — the statement that lets one speak of walking along the edges of a solid.

Proposition 22.3.4. An Edge Joins Two Vertices

Let $P \subseteq \mathbb{R}^n$ be a bounded polyhedron and F an edge of P. Then $F = [\mathbf{v}, \mathbf{w}]$ for two distinct vertices $\mathbf{v}, \mathbf{w}$ of P.

Idea. An edge is itself a bounded polyhedron, so it is the hull of its own vertices; those are vertices of P because a face passes extremeness upwards. Being 1-dimensional, it sits on a line, and the hull of finitely many points of a line is the segment between the outermost two.

Proof. By Proposition 22.3.1, F is a polyhedron, and it is bounded because $F \subseteq P$. By Theorem 22.3.3, $F = \operatorname{conv}(\operatorname{ext} F)$ with $\operatorname{ext} F$ finite, and $F \neq \emptyset$ since it has a dimension. By the remark after Definition 22.3.1, F is a face of P in the sense of Definition 18.8.2, so Lemma 18.8.3 gives $\operatorname{ext} F = F \cap \operatorname{ext} P$: every vertex of F is a vertex of P.

Fix $\mathbf{p} \in F$. By Proposition 18.1.4, $\operatorname{aff} F = \mathbf{p} + U$ with $U = \operatorname{span}\{\mathbf{x} - \mathbf{p} : \mathbf{x} \in F\}$, and $\dim U = \dim F = 1$, so $U = \operatorname{span}(\mathbf{u})$ for some $\mathbf{u} \neq \mathbf{0}$. Every point of F lies in $\operatorname{aff} F$, hence is $\mathbf{p} + t\mathbf{u}$ for exactly one $t \in \mathbb{R}$, the map $t \mapsto \mathbf{p} + t\mathbf{u}$ being injective because $\mathbf{u} \neq \mathbf{0}$. Let $t_1 \leq \cdots \leq t_k$ be the parameters of the points of $\operatorname{ext} F$, a nonempty finite list. A convex combination $\sum_j s_j(\mathbf{p} + t_j\mathbf{u})$ equals $\mathbf{p} + (\sum_j s_j t_j)\mathbf{u}$, and $\sum_j s_j t_j$ lies in $[t_1, t_k]$; conversely each $t \in [t_1, t_k]$ is $(1-\lambda)t_1 + \lambda t_k$ for some $\lambda \in [0,1]$. Hence, writing $[\mathbf{v}, \mathbf{w}]$ for the segment of Chapter 18 §01,

$$F = \{\mathbf{p} + t\mathbf{u} : t_1 \leq t \leq t_k\} = [\mathbf{v}, \mathbf{w}], \qquad \mathbf{v} = \mathbf{p} + t_1\mathbf{u}, \quad \mathbf{w} = \mathbf{p} + t_k\mathbf{u}.$$

If $t_1 = t_k$ then F is a single point and $\dim F = 0$, contrary to hypothesis; so $\mathbf{v} \neq \mathbf{w}$, and both are vertices of P. This proves the proposition.

■

22.3.3 Two face lattices, computed

Faces are ordered by inclusion, and the resulting picture — which face lies in which — is the **face lattice** of the polyhedron. It is the combinatorial skeleton of the solid. Here it is in full for the two standard examples.

Example. The faces of a cube

Let $B_\infty = [-1,1]^n$, described by the $2n$ rows $\mathbf{e}_i^\top \mathbf{x} \leq 1$ and $-\mathbf{e}_i^\top \mathbf{x} \leq 1$. Find all faces of B_∞ and their dimensions.

Solution. By Proposition 22.3.1 (b) every nonempty face is some P_J. A set J of rows says, for each coordinate i, either nothing, or $x_i = 1$, or $x_i = -1$, or both; the last makes $P_J = \emptyset$.

So the nonempty faces are indexed by **sign vectors** $\mathbf{s} \in \{-1, 0, 1\}^n$:

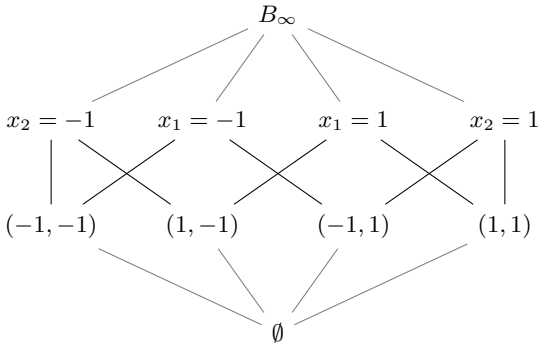
$$F_{\mathbf{s}} = \{\mathbf{x} \in B_\infty : x_i = s_i \text{ for every } i \text{ with } s_i \neq 0\}.$$

Each $F_{\mathbf{s}}$ is nonempty: the point with $x_i = s_i$ where $s_i \neq 0$ and $x_i = 0$ elsewhere lies in it. Distinct sign vectors give distinct faces, because $s_i = \pm 1$ exactly when the coordinate x_i is constantly ± 1 on $F_{\mathbf{s}}$. So B_∞ has exactly 3^n nonempty faces, one for each sign vector, together with $\emptyset$.

Let $Z = \{i : s_i = 0\}$ and $z = |Z|$. Then $\operatorname{aff} F_{\mathbf{s}}$ is the flat $\{\mathbf{x} : x_i = s_i \text{ for } i \notin Z\}$, a coset of $\operatorname{span}\{\mathbf{e}_i : i \in Z\}$: it contains $F_{\mathbf{s}}$, and conversely $F_{\mathbf{s}}$ contains the point $\mathbf{p}$ with $p_i = s_i$ for $i \notin Z$ and $p_i = 0$ for $i \in Z$, together with $\mathbf{p} + \mathbf{e}_i$ for every $i \in Z$, whose differences from $\mathbf{p}$ span $\operatorname{span}\{\mathbf{e}_i : i \in Z\}$. By Proposition 18.1.4, $\dim F_{\mathbf{s}} = z$.

So the faces of dimension k are the sign vectors with exactly k zeros, and there are $\binom{n}{k}2^{n-k}$ of them. For $n = 3$: 8 vertices ($k = 0$), 12 edges ($k = 1$), 6 facets ($k = 2$), and the cube itself, and $8 + 12 + 6 + 1 = 27 = 3^3$. Every vertex is a sign vector ± 1 in each coordinate, matching what Example (The cube and the cross-polytope) found by the rank test.

Here is the resulting lattice for the square $B_\infty \subseteq \mathbb{R}^2$, with $\emptyset$ at the bottom and the square at the top, and a line drawn when the lower face is contained in the upper one.



The face lattice of the square: one empty face, four vertices, four edges, and the square itself. Nine non-empty faces, 3^2

Example. The faces of a simplex

Let $\Delta_n = \{\mathbf{t} \in \mathbb{R}^n : t_i \geq 0,\ \sum_i t_i = 1\}$ be the standard simplex (Definition 18.1.2), described by the n rows $-t_i \leq 0$ together with $\mathbf{1}^\top \mathbf{t} \leq 1$ and $-\mathbf{1}^\top \mathbf{t} \leq -1$. Find all faces of Δ_n.

Solution. The last two rows are tight at every point of Δ_n, so including them in J changes nothing. A face is therefore P_J for J determined by a set $S \subseteq \{1, \ldots, n\}$ of coordinates forced to zero:

$$F_S = \{\mathbf{t} \in \Delta_n : t_i = 0 \text{ for every } i \in S\}.$$

Write $T = \{1, \ldots, n\} \setminus S$. If $T = \emptyset$ the conditions $\mathbf{t} = \mathbf{0}$ and $\sum_i t_i = 1$ clash, so $F_S = \emptyset$. If $T \neq \emptyset$, then F_S consists of the vectors $\mathbf{t} \geq \mathbf{0}$ supported in T with $\sum_{i \in T} t_i = 1$, which by Theorem 18.1.3 is exactly $\operatorname{conv}\{\mathbf{e}_i : i \in T\}$, a copy of $\Delta_{|T|}$. Its dimension is $|T| - 1$, by the computation of $\dim \Delta_k$ after Proposition 18.1.4. Distinct nonempty T give distinct faces, since $\mathbf{e}_i \in F_S$ exactly for $i \in T$.

So the nonempty faces of Δ_n are indexed by the **nonempty subsets** $T \subseteq \{1, \dots, n\}$, and there are $2^n - 1$ of them: n vertices, $\binom{n}{2}$ edges, and so on up to Δ_n itself. For $n = 4$, the tetrahedron $\Delta_4 \subseteq \mathbb{R}^4$ of dimension 3 has $4 + 6 + 4 + 1 = 15 = 2^4 - 1$ nonempty faces.

The contrast is the point. The cube's lattice is the set of sign vectors; that of the simplex is the set of nonempty subsets. A cube in dimension n has 2^n vertices and $2n$ facets; a simplex has n vertices and n facets. The cube's corners grow exponentially with n, those of the simplex only linearly.

Quick check. How many faces does a triangle in $\mathbb{R}^2$ have, counting the empty face and the triangle itself?

Solution. Eight: $\emptyset$, three vertices, three edges, and the triangle. Count directly from Proposition 22.3.1 (b), writing the triangle as the polyhedron cut out by its three sides: $P_\emptyset$ is the triangle, each single row gives one side, each pair of rows gives the corner where those two sides meet, and all three rows together give $\emptyset$. That is $1 + 3 + 3$ nonempty faces, plus $\emptyset$. The count matches Example (The faces of a simplex) with $n = 3$, whose $2^3 - 1 = 7$ nonempty faces are indexed by the nonempty subsets of the three corners.

22.3.4 A linear program, seen

Fix $\mathbf{c} \in \mathbb{R}^n$ with $\mathbf{c} \neq \mathbf{0}$. The level sets $\{\mathbf{x} : \mathbf{c}^\top \mathbf{x} = \gamma\}$ of the objective are parallel affine hyperplanes, one for each γ, and they sweep across $\mathbb{R}^n$ as γ grows, moving in the direction $\mathbf{c}$. Maximizing $\mathbf{c}^\top \mathbf{x}$ over a polyhedron P means pushing the sweeping hyperplane as far as it will go while still touching P. Two things can happen: it escapes to infinity, and the program is unbounded; or there is a last hyperplane that touches, and what it touches is a face.

Proposition 22.3.5. The Optimal Set Is a Face

Let $P \subseteq \mathbb{R}^n$ be a polyhedron, $\mathbf{c} \in \mathbb{R}^n$, and suppose $\mathbf{c}^\top \mathbf{x}$ attains a maximum $\gamma^\star$ on P. Then the set $F^\star$ of maximizers is a nonempty face of P. If P is bounded, $F^\star$ contains a vertex of P.

Proof. By the definition of a maximum, $\mathbf{c}^\top \mathbf{x} \leq \gamma^\star$ for every $\mathbf{x} \in P$, and $F^\star = \{\mathbf{x} \in P : \mathbf{c}^\top \mathbf{x} = \gamma^\star\}$ is nonempty. That is Definition 22.3.1 verbatim, so $F^\star$ is a face.

Suppose P is bounded. Then $F^\star \subseteq P$ is a bounded polyhedron by Proposition 22.3.1, and nonempty, so $\operatorname{ext} F^\star \neq \emptyset$ by Theorem 22.3.3 (a nonempty set is not the hull of the empty set). By the remark after Definition 22.3.1 and Lemma 18.8.3, $\operatorname{ext} F^\star = F^\star \cap \operatorname{ext} P$, so any point of $\operatorname{ext} F^\star$ is a vertex of P lying in $F^\star$.

■

The last sentence is Corollary 18.8.6 read geometrically, and it says a little more than that corollary: not only is *some* maximizer a vertex, but the *whole* set of maximizers is a face — a single vertex when the optimum is unique, and otherwise an edge or a larger face, up to all of P

when $\mathbf{c} = \mathbf{0}$. A linear program with a nonunique optimum has a positive-dimensional optimal face, and every point of it is optimal.

An algorithm now suggests itself: start at a vertex, and as long as some edge leaving it increases the objective, walk along that edge to the vertex at its other end (Proposition 22.3.4 guarantees there is one, when P is bounded). That is the idea of the **simplex method**, sketched in Chapter 18 §09 after Corollary 18.9.3. **No algorithm is developed here or anywhere in this book:** Chapter 24 is where this book's algorithms live, and it treats the numerical linear algebra of elimination, least squares, and eigenvalues, not linear programming. What this section supplies is the geometry an algorithm would exploit.

Example. A linear program and its picture

Maximize $3x_1 + 2x_2$ over

$$P = \{\mathbf{x} \in \mathbb{R}^2 : x_1 \geq 0,\ x_2 \geq 0,\ x_1 + x_2 \leq 5,\ 2x_1 + x_2 \leq 8\}.$$

Find the vertices, the optimum, the optimal face, and the dual optimal solution.

Solution. Number the rows (1) $-x_1 \leq 0$, (2) $-x_2 \leq 0$, (3) $x_1 + x_2 \leq 5$, (4) $2x_1 + x_2 \leq 8$. No two rows are multiples of each other, so each of the $\binom{4}{2} = 6$ pairs gives one basic point, which we test against the other two rows:

pair	basic point	verdict
(1), (2)	$(0,0)$	feasible
(1), (3)	$(0,5)$	(4): $5 \leq 8$: feasible
(1), (4)	$(0,8)$	(3): $8 \leq 5$ fails
(2), (3)	$(5,0)$	(4): $10 \leq 8$ fails
(2), (4)	$(4,0)$	(3): $4 \leq 5$: feasible
(3), (4)	$(3,2)$	feasible

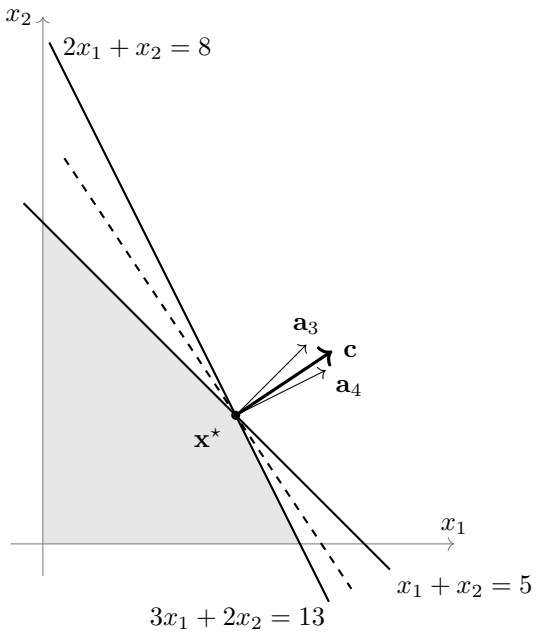
For the last pair, subtracting (3) from (4) gives $x_1 = 3$, hence $x_2 = 2$. By Proposition 22.3.2 the vertices are $(0,0)$, $(0,5)$, $(4,0)$ and $(3,2)$. P is bounded, since $0 \leq x_1, x_2 \leq 5$ on it, so by Theorem 22.3.3 it is the quadrilateral $\operatorname{conv}\{(0,0),(4,0),(3,2),(0,5)\}$.

The objective takes the value 0 at $(0,0)$, 10 at $(0,5)$, 12 at $(4,0)$, and 13 at $(3,2)$, so by Proposition 22.3.5 the maximum is 13, at $\mathbf{x}^\star = (3,2)$, and the optimal face is $\{\mathbf{x} \in P : 3x_1 + 2x_2 = 13\}$. Since only one vertex achieves 13 and the face is the hull of its vertices, the optimal face is the single point $\{(3,2)\}$: the optimum is unique.

The rows active at $\mathbf{x}^\star$ are (3) and (4), with outer normals $\mathbf{a}_3 = (1,1)$ and $\mathbf{a}_4 = (2,1)$. And

$$\mathbf{c} = (3,2) = 1 \cdot (1,1) + 1 \cdot (2,1) = 1 \cdot \mathbf{a}_3 + 1 \cdot \mathbf{a}_4,$$

so $\mathbf{y} = (1,1)$ is a candidate for the dual. In the notation of Definition 18.6.2, with $\mathbf{A}$ the rows $(1,1)$ and $(2,1)$ and $\mathbf{b} = (5,8)$, the dual asks to minimize $5y_1 + 8y_2$ subject to $y_1 + 2y_2 \geq 3$, $y_1 + y_2 \geq 2$, $\mathbf{y} \geq \mathbf{0}$. At $\mathbf{y} = (1,1)$ both constraints are equalities and the value is $5 + 8 = 13$, which matches. By Corollary 18.6.2 (a), $\mathbf{x}^\star$ and $\mathbf{y}$ are both optimal, and no further checking is needed.



The feasible quadrilateral, the level line of the objective through the optimum, and the two active outer normals whose non-negative combination is $\mathbf{c} = \mathbf{a}_3 + \mathbf{a}_4$

22.3.5 What the dual variables are, geometrically

The example ended with an identity worth isolating: the objective vector was a **nonnegative** combination of the outer normals of the constraints active at the optimum, and the coefficients were the dual optimal solution. That is the geometric content of duality, and Chapter 18 has already done all the work.

Recall Chapter 18 §06's pair of programs, with $\mathbf{A} \in M_{m\times n}(\mathbb{R})$, $\mathbf{b} \in \mathbb{R}^m$, $\mathbf{c} \in \mathbb{R}^n$:

$$\text{(P)}\quad \text{maximize } \mathbf{c}^\top \mathbf{x} \ \text{ subject to } \mathbf{A}\mathbf{x} \le \mathbf{b},\ \mathbf{x} \ge \mathbf{0};$$
$$\text{(D)}\quad \text{minimize } \mathbf{b}^\top \mathbf{y} \ \text{ subject to } \mathbf{A}^\top \mathbf{y} \ge \mathbf{c},\ \mathbf{y} \ge \mathbf{0}.$$

The feasible region of (P) is the polyhedron with the m rows $\mathbf{a}_i^\top \mathbf{x} \le b_i$ and the n rows $(-\mathbf{e}_j)^\top \mathbf{x} \le 0$. The **outer normal** of a row is the vector on its left-hand side: $\mathbf{a}_i$ for the i-th constraint and $-\mathbf{e}_j$ for the sign constraint on x_j.

Proposition 22.3.6. Dual Variables Are the Weights of the Active Normals

Let $\mathbf{x}^\star$ be an optimal solution of (P). Then there are $\mathbf{y} \in \mathbb{R}^m$ and $\mathbf{s} \in \mathbb{R}^n$ with $\mathbf{y} \ge \mathbf{0}$ and $\mathbf{s} \ge \mathbf{0}$ such that

$$\mathbf{c} = \sum_{i=1}^{m} y_i \mathbf{a}_i + \sum_{j=1}^{n} s_j (-\mathbf{e}_j),$$

where $y_i = 0$ whenever $\mathbf{a}_i^\top \mathbf{x}^\star < b_i$, and $s_j = 0$ whenever $x_j^\star > 0$. That is, $\mathbf{c}$ is a nonnegative combination of the outer normals of the constraints **active at** $\mathbf{x}^\star$, and the weights are an optimal solution of (D).

Idea. Strong duality hands over an optimal $\mathbf{y}$ for (D); the slack $\mathbf{s} = \mathbf{A}^\top \mathbf{y} - \mathbf{c}$ is nonnegative because $\mathbf{y}$ is dual feasible, and rearranging is the displayed identity. Complementary slackness

is exactly the statement that the weights vanish on the inactive constraints.

Proof. Since $\mathbf{x}^\star$ is feasible, (P) is feasible, and $\mathbf{c}^\top\mathbf{x} \le \mathbf{c}^\top\mathbf{x}^\star$ for every feasible $\mathbf{x}$, so (P) is not unbounded. By Theorem 18.6.3 (b), (D) is therefore feasible, and by Theorem 18.6.3 (a) both programs have optimal solutions with equal optimal values; let $\mathbf{y}$ be an optimal solution of (D). Then $\mathbf{y} \ge \mathbf{0}$ and $\mathbf{A}^\top\mathbf{y} \ge \mathbf{c}$, so $\mathbf{s} \coloneqq \mathbf{A}^\top\mathbf{y} - \mathbf{c} \ge \mathbf{0}$. Reading the j-th coordinate of $\mathbf{A}^\top\mathbf{y}$ as $\sum_i y_i a_{ij}$, we get $\mathbf{A}^\top\mathbf{y} = \sum_i y_i \mathbf{a}_i$, and hence

$$\mathbf{c} = \mathbf{A}^\top\mathbf{y} - \mathbf{s} = \sum_{i=1}^{m} y_i \mathbf{a}_i + \sum_{j=1}^{n} s_j(-\mathbf{e}_j).$$

Both $\mathbf{x}^\star$ and $\mathbf{y}$ are optimal, so Theorem 18.6.5 applies: its condition (i) says $y_i > 0$ forces $(\mathbf{A}\mathbf{x}^\star)_i = b_i$, which is the contrapositive of "$\mathbf{a}_i^\top\mathbf{x}^\star < b_i$ forces $y_i = 0$"; its condition (ii) says $x_j^\star > 0$ forces $(\mathbf{A}^\top\mathbf{y})_j = c_j$, that is $s_j = 0$. This proves the proposition. ∎

So a dual optimal solution is a **certificate written in normals**. The hyperplane $\{\mathbf{x} : \mathbf{c}^\top\mathbf{x} = \mathbf{b}^\top\mathbf{y}\}$ supports the feasible region at $\mathbf{x}^\star$, by Theorem 18.6.1 and equality of the optimal values, and Proposition 22.3.6 says how its normal $\mathbf{c}$ is assembled from the normals of the facets meeting at $\mathbf{x}^\star$. Increasing b_i loosens the i-th constraint, and y_i **bounds** the gain: the dual feasible region $\{\mathbf{y} \ge \mathbf{0} : \mathbf{A}^\top\mathbf{y} \ge \mathbf{c}\}$ does not mention $\mathbf{b}$ at all, so $\mathbf{y}$ stays dual feasible when b_i is raised to $b_i + \varepsilon$, and Theorem 18.6.1 then caps the objective of every point feasible for the loosened program at $\mathbf{b}^\top\mathbf{y} + \varepsilon y_i$, the old optimal value plus εy_i. That y_i is also the **exact** rate of gain, its *shadow price* in the economists' phrase, says more, and this book does not prove it; nothing here uses it, and it can fail. Take $n = 1$, $m = 2$, $\mathbf{A} = \left(\begin{smallmatrix}1\\1\end{smallmatrix}\right)$, $\mathbf{b} = (1, 1)$, and $\mathbf{c} = 1$, so that (P) maximizes x subject to $x \le 1$ written twice and $x \ge 0$. Its optimum is 1, and $\mathbf{y} = (1, 0)$ is dual feasible with $\mathbf{b}^\top\mathbf{y} = 1$, hence dual optimal by Corollary 18.6.2 (a); but raising b_1 leaves the second copy binding and the optimum at 1, a gain of 0 against $y_1 = 1$. The same pair shows which half of complementary slackness is available: Theorem 18.6.5 (i) says $y_i > 0$ forces row i tight, so a constraint **slack** at the optimum has $y_i = 0$, and that is what Proposition 22.3.6 records. The converse is the trap — here $y_2 = 0$ while row 2 is tight.

22.3.6 Polyhedra without corners

Boundedness is doing real work in Theorem 22.3.3, and the failures are worth seeing.

Example. A slab, and a quadrant

(a) Find the faces of the slab $S = \{\mathbf{x} \in \mathbb{R}^3 : 0 \le x_3 \le 1\}$, and its vertices.

(b) Show that $\max\{x_1 + x_2 : \mathbf{x} \in Q\}$ does not exist for the quadrant $Q = \{\mathbf{x} \in \mathbb{R}^2 : x_1 \ge 0,\ x_2 \ge 0\}$, although Q has a vertex.

Solution.

(a) S has the two rows $-x_3 \le 0$ and $x_3 \le 1$. By Proposition 22.3.1 the nonempty faces are among $P_\emptyset = S$, $P_{\{1\}} = \{x_3 = 0\}$, $P_{\{2\}} = \{x_3 = 1\}$ and $P_{\{1,2\}} = \emptyset$. So there are three nonempty faces: S and two parallel planes, of dimensions 3, 2, 2. There is no

face of dimension 0 or 1. In particular S has **no vertices**: at any $\mathbf{x} \in S$ the active rows are among $\pm\mathbf{e}_3$, of rank at most $1 < 3$, so Theorem 18.9.2 denies a vertex. Hence $\operatorname{conv}(\operatorname{ext} S) = \operatorname{conv} \emptyset = \emptyset \neq S$, and Theorem 18.8.5 indeed needs compactness. Note also that $\max\{-x_3 : \mathbf{x} \in S\} = 0$ **is** attained, on the whole face $\{x_3 = 0\}$, which contains no vertex: the last sentence of Proposition 22.3.5 needs boundedness too.

(b) The point $(0, 0)$ is a vertex of Q: the two active rows $-\mathbf{e}_1, -\mathbf{e}_2$ are independent, so Theorem 18.9.2 applies. It is the only one, since at any other point of Q at most one row is active. But $(t, t) \in Q$ for every $t \geq 0$ and $x_1 + x_2 = 2t$ there, so the objective is unbounded above and no maximum exists. A polyhedron can have corners and still have no optimum.

Warning. **"Vertex" and "corner of the picture" can both be empty words for an unbounded polyhedron.** The slab above is a perfectly ordinary convex solid, closed, with a genuine boundary, and it has no vertices, no edges, and no bounded face. Nothing in this section rebuilds it from finitely many points, and nothing should: by Theorem 18.9.5 (a) a polyhedron is a polytope exactly when it is bounded. The general finite description of an unbounded polyhedron needs directions as well as points — every nonempty polyhedron is a polytope plus a finitely generated cone — which Chapter 18 derives from Theorem 18.9.5 (b) in Exercise (C1: Unbounded polyhedra), and which is not pursued here.

22.3.7 Exercises

22.3.7.1 A. Check your understanding

Exercise. A1

(a) State Definition 22.3.1, and say which pair $(\mathbf{c}, \beta)$ exhibits P itself as a face and which exhibits $\emptyset$.

(b) True or false: a polyhedron given by m inequalities in $\mathbb{R}^n$ has exactly 2^m nonempty faces. Justify your answer.

(c) Give the three equivalent descriptions of a vertex from Proposition 22.3.2.

(d) True or false: if a linear program over a bounded nonempty polyhedron has two distinct optimal solutions, it has infinitely many. Justify your answer.

(e) How many faces of dimension 2 does the cube $[-1, 1]^4$ have?

Solution.

(a) A face of a polyhedron P is a set $\{\mathbf{x} \in P : \mathbf{c}^\top \mathbf{x} = \beta\}$ for some $\mathbf{c}, \beta$ with $\mathbf{c}^\top \mathbf{x} \leq \beta$ valid on all of P. The pair $(\mathbf{0}, 0)$ gives $F = P$; the pair $(\mathbf{0}, 1)$ gives $F = \emptyset$.

(b) False; 2^m is only an upper bound (Proposition 22.3.1 (c)). The map $J \mapsto P_J$ need not be injective, and some P_J may be empty. The warning after that proposition has $m = 3$ rows in $\mathbb{R}^2$ and only four nonempty faces: the polyhedron, two edges, and one vertex.

(c) $\mathbf{v} \in P$ is a vertex when: it is an extreme point of P; or $\{\mathbf{v}\}$ is a face of P; or among the rows active at $\mathbf{v}$ there are n linearly independent ones.

(d) True. The optimal set is a face (Proposition 22.3.5), hence convex, so it contains the whole segment joining the two optimal points, and a segment with distinct ends has infinitely many points.

(e) By Example (The faces of a cube) with $n = 4$, $k = 2$: $\binom{4}{2}2^{4-2} = 6 \cdot 4 = 24$.

22.3.7.2 B. Practice

Exercise. B1: Vertices and faces of a region

Let

$$P = \{\mathbf{x} \in \mathbb{R}^2 : x_1 \geq 0,\ x_2 \geq 0,\ x_1 + 2x_2 \leq 8,\ 2x_1 + x_2 \leq 10\}.$$

Find all vertices of P, list all of its faces, and maximize $x_1 + 3x_2$ over P.

Solution. Number the rows (1) $-x_1 \leq 0$, (2) $-x_2 \leq 0$, (3) $x_1 + 2x_2 \leq 8$, (4) $2x_1 + x_2 \leq 10$. No two are multiples, so each of the six pairs gives a basic point:

pair	point	verdict
(1), (2)	$(0, 0)$	feasible
(1), (3)	$(0, 4)$	(4): $4 \leq 10$: feasible
(1), (4)	$(0, 10)$	(3): $20 \leq 8$ fails
(2), (3)	$(8, 0)$	(4): $16 \leq 10$ fails
(2), (4)	$(5, 0)$	(3): $5 \leq 8$: feasible
(3), (4)	$(4, 2)$	feasible

For the last pair, twice (3) minus (4) gives $3x_2 = 6$, so $x_2 = 2$ and $x_1 = 8 - 4 = 4$. By Proposition 22.3.2 the vertices are $(0, 0)$, $(0, 4)$, $(5, 0)$ and $(4, 2)$, so P is a quadrilateral, not a pentagon; two of the six basic points are infeasible. P is bounded ($0 \leq x_1 \leq 5$, $0 \leq x_2 \leq 4$ on it), so $P = \operatorname{conv}\{(0, 0), (5, 0), (4, 2), (0, 4)\}$ by Theorem 22.3.3.

By Proposition 22.3.1 (b) every nonempty face is some P_J. The four singletons give the four edges, computed directly: $P_{\{1\}} = \{x_1 = 0,\ 0 \leq x_2 \leq 4\}$; $P_{\{2\}} = \{x_2 = 0,\ 0 \leq x_1 \leq 5\}$; $P_{\{3\}} = \{x_1 + 2x_2 = 8\} \cap P$, where $x_1 = 8 - 2x_2 \geq 0$ gives $x_2 \leq 4$ and $2x_1 + x_2 = 16 - 3x_2 \leq 10$ gives $x_2 \geq 2$, a segment from $(4, 2)$ to $(0, 4)$; and $P_{\{4\}} = \{2x_1 + x_2 = 10\} \cap P$, which in the same way is the segment from $(4, 2)$ to $(5, 0)$. The six pairs give $P_{\{1,2\}} = \{(0, 0)\}$, $P_{\{1,3\}} = \{(0, 4)\}$, $P_{\{2,4\}} = \{(5, 0)\}$, $P_{\{3,4\}} = \{(4, 2)\}$, and $P_{\{1,4\}} = P_{\{2,3\}} = \emptyset$, the two infeasible basic points; every J with three or more rows gives $\emptyset$, since no point of P has three rows active. So P has ten faces in all: $\emptyset$, four vertices, four edges, and P itself.

The objective $x_1 + 3x_2$ takes the values 0, 5, 10, 12 at $(0, 0)$, $(5, 0)$, $(4, 2)$, $(0, 4)$. By Proposition 22.3.5 the maximum is 12, attained at $(0, 4)$ alone.

Exercise. B2: Reading the dual off the normals

For the program of Exercise (B1: Vertices and faces of a region) with objective $x_1 + 3x_2$, use Proposition 22.3.6 to write $\mathbf{c} = (1, 3)$ as a nonnegative combination of the outer normals of the constraints active at the optimum, and check the resulting dual solution against Corollary 18.6.2.

Solution. The optimum is $\mathbf{x}^\star = (0,4)$. The active rows there are (1), with outer normal $-\mathbf{e}_1 = (-1,0)$, and (3), with outer normal $\mathbf{a}_3 = (1,2)$. In the notation of Definition 18.6.2 the rows of $\mathbf{A}$ are (3) and (4) in that order, so the weight on $\mathbf{a}_3$ is y_1. Writing $\mathbf{c} = y_1\mathbf{a}_3 + s_1(-\mathbf{e}_1)$ with $\mathbf{a}_3 = (1,2)$:

$$(1,3) = y_1(1,2) + s_1(-1,0)$$

gives $2y_1 = 3$, so $y_1 = \frac{3}{2}$, and $1 = \frac{3}{2} - s_1$, so $s_1 = \frac{1}{2}$. Both are nonnegative, as Proposition 22.3.6 requires. Row (4) is inactive at $\mathbf{x}^\star$, so its weight is $y_2 = 0$; and $x_2^\star = 4 > 0$, so the slack on the second sign constraint is $s_2 = 0$.

So $\mathbf{A}$ has rows $(1,2)$ and $(2,1)$, with $\mathbf{b} = (8,10)$ and $\mathbf{c} = (1,3)$. The dual solution read off is $\mathbf{y} = (\frac{3}{2}, 0)$, which satisfies $\mathbf{y} \geq \mathbf{0}$ and

$$\mathbf{A}^\top\mathbf{y} = \tfrac{3}{2}(1,2) = (\tfrac{3}{2},3) \geq (1,3) = \mathbf{c},$$

so it is dual feasible, with value $\mathbf{b}^\top\mathbf{y} = 8 \cdot \frac{3}{2} = 12$. Since the primal value at $(0,4)$ is also 12, Corollary 18.6.2 (a) confirms that both are optimal.

Exercise. B3: Faces of the cross-polytope

Let $B_1 = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{s}^\top\mathbf{x} \leq 1 \text{ for every sign vector } \mathbf{s} \in \{-1,1\}^n\}$, the cross-polytope of Example (The cube and the cross-polytope). Determine the face P_J for J a single sign vector $\mathbf{s}$, and its dimension. How many facets does B_1 have?

Solution. For a single sign vector $\mathbf{s}$, $P_{\{\mathbf{s}\}} = \{\mathbf{x} \in B_1 : \mathbf{s}^\top\mathbf{x} = 1\}$. Now $\mathbf{s}^\top\mathbf{x} = \sum_i s_i x_i \leq \sum_i |x_i| = \|\mathbf{x}\|_1$, term by term, so equality holds exactly when $s_i x_i = |x_i|$ for every i; and $\|\mathbf{x}\|_1 = \max_{\mathbf{s}} \mathbf{s}^\top\mathbf{x} \leq 1$ on B_1, as Example (The cube and the cross-polytope) records. So $\mathbf{s}^\top\mathbf{x} = 1$ forces $\|\mathbf{x}\|_1 = 1$ and $s_i x_i = |x_i| \geq 0$ for every i; conversely those conditions give $\mathbf{s}^\top\mathbf{x} = \|\mathbf{x}\|_1 = 1$. Writing $\mathbf{x} = \sum_i x_i\mathbf{e}_i$ with $x_i = s_i|x_i|$, we get

$$P_{\{\mathbf{s}\}} = \Big\{\sum_i t_i(s_i\mathbf{e}_i) : t_i \geq 0,\ \sum_i t_i = 1\Big\} = \operatorname{conv}\{s_1\mathbf{e}_1, \ldots, s_n\mathbf{e}_n\},$$

using Theorem 18.1.3. This is a copy of Δ_n, of dimension $n-1$ by the computation after Proposition 18.1.4. Since $\dim B_1 = n$ — it contains $\mathbf{0}$ and every $\pm\mathbf{e}_i$, and the differences $\mathbf{e}_i - \mathbf{0}$ span $\mathbb{R}^n$, so Proposition 18.1.4 gives $\operatorname{aff} B_1 = \mathbb{R}^n$ — each $P_{\{\mathbf{s}\}}$ is a facet. Distinct $\mathbf{s}$ give distinct facets, since $\mathbf{s}$ is recovered from $P_{\{\mathbf{s}\}}$ by the signs of the vertices $s_i\mathbf{e}_i$.

There are no others. Every nonempty face is P_J for a set J of sign vectors, by Proposition 22.3.1 (b), and if $\mathbf{s} \neq \mathbf{s}'$ both lie in J then $P_J \subseteq P_{\{\mathbf{s}\}} \cap P_{\{\mathbf{s}'\}}$. A point of that intersection has $x_i = s_i|x_i| = s_i'|x_i|$, so $x_i = 0$ whenever $s_i \neq s_i'$; hence the intersection is $\operatorname{conv}\{s_i\mathbf{e}_i : s_i = s_i'\}$, of dimension at most $n-2$, since at least one index is missing. So B_1 has exactly 2^n facets: in $\mathbb{R}^n$ it has $2n$ vertices and 2^n facets, exactly the reverse of the cube.

22.3.7.3 C. Going deeper

Exercise. C1: A face of a face

Let P be a polyhedron and F a face of P.

(a) Prove that every face of F is a face of P.

(b) Deduce that the faces of P contained in a fixed face F are exactly the faces of F.

Hint: for (a), use Proposition 22.3.1 to write both F and the face of F in the form P_J.

Solution.

(a) We may assume the face G of F is nonempty, since $\emptyset$ is a face of P. Fix a system $\mathbf{A}\mathbf{x} \le \mathbf{b}$ with $P = \{\mathbf{x} : \mathbf{A}\mathbf{x} \le \mathbf{b}\}$ and rows $\mathbf{a}_i^\top \mathbf{x} \le b_i$, $i = 1, \dots, m$. By Proposition 22.3.1 (b), $F = P_J$ for some J. By part (a) of the same proposition, F is cut out by the m original rows together with the rows $-\mathbf{a}_i^\top \mathbf{x} \le -b_i$ for $i \in J$. By (b) applied to F with **this** system, $G = F_K$ for some set K of those rows; since the rows $-\mathbf{a}_i^\top \mathbf{x} \le -b_i$ with $i \in J$ are tight everywhere on F, including them in K changes nothing, so we may take $K \subseteq \{1, \dots, m\}$. Then

$$G = \{\mathbf{x} \in F : \mathbf{a}_i^\top \mathbf{x} = b_i \text{ for } i \in K\} = \{\mathbf{x} \in P : \mathbf{a}_i^\top \mathbf{x} = b_i \text{ for } i \in J \cup K\} = P_{J \cup K},$$

which is a face of P by Proposition 22.3.1 (a).

(b) By (a), every face of F is a face of P, and it is contained in F. Conversely let G be a face of P with $G \subseteq F$, say $G = \{\mathbf{x} \in P : \mathbf{c}^\top \mathbf{x} = \beta\}$ with $\mathbf{c}^\top \mathbf{x} \le \beta$ on P. Then $\mathbf{c}^\top \mathbf{x} \le \beta$ holds in particular on $F \subseteq P$, and

$$\{\mathbf{x} \in F : \mathbf{c}^\top \mathbf{x} = \beta\} = \{\mathbf{x} \in P : \mathbf{c}^\top \mathbf{x} = \beta\} \cap F = G \cap F = G,$$

so G is a face of F.

Exercise. C2: The two notions of face agree for a polyhedron

Let $P = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{A}\mathbf{x} \le \mathbf{b}\}$ and let $F \subseteq P$ be a nonempty face in the sense of Definition 18.8.2: F is convex, and whenever $t\mathbf{y} + (1-t)\mathbf{z} \in F$ with $\mathbf{y}, \mathbf{z} \in P$ and $0 < t < 1$, both $\mathbf{y}$ and $\mathbf{z}$ lie in F.

(a) Let $J = \{i : \mathbf{a}_i^\top \mathbf{x} = b_i \text{ for every } \mathbf{x} \in F\}$. Prove that $F = P_J$.

(b) Deduce that F is a face in the sense of Definition 22.3.1, so the two notions agree for polyhedra.

Hint: for (a), Step 1 of the proof of Proposition 22.3.1 produces a point of F at which every row outside J is slack; use it as the midpoint of a short segment.

Solution.

(a) $F \subseteq P_J$ by the definition of J. For the reverse inclusion, note first that F is convex, so Step 1 of the proof of Proposition 22.3.1 applies verbatim and yields $\mathbf{x}^* \in F$ with $\mathbf{a}_i^\top \mathbf{x}^* < b_i$ for every $i \notin J$.
Let $\mathbf{y} \in P_J$. If $\mathbf{y} = \mathbf{x}^*$ we are done, so assume $\mathbf{y} \ne \mathbf{x}^*$ and put $\mathbf{z}_t = \mathbf{x}^* + t(\mathbf{x}^* - \mathbf{y})$ for $t > 0$. Exactly as in Step 2 of that proof, $\mathbf{a}_i^\top \mathbf{z}_t = b_i$ for $i \in J$ and, for small enough $t > 0$, $\mathbf{a}_i^\top \mathbf{z}_t < b_i$ for $i \notin J$; so $\mathbf{z}_t \in P$. Fix such a t. Then

$$\mathbf{x}^* = \frac{t}{1+t}\,\mathbf{y} + \frac{1}{1+t}\,\mathbf{z}_t,$$

a combination with positive coefficients adding to 1: the right side is $\frac{t\mathbf{y}+\mathbf{x}^*+t\mathbf{x}^*-t\mathbf{y}}{1+t} = \mathbf{x}^*$. Both $\mathbf{y}$ and $\mathbf{z}_t$ lie in P and $\mathbf{x}^* \in F$, so the face property of Definition 18.8.2 forces $\mathbf{y} \in F$. Hence $P_J \subseteq F$ and $F = P_J$.

(b) By Proposition 22.3.1 (a), P_J is a face in the sense of Definition 22.3.1. Together with the remark after Definition 22.3.1, which gives the other direction for every face, the two notions coincide on polyhedra.

They do not coincide for general convex sets, and one round set shows it. Let $D = \{\mathbf{x} \in \mathbb{R}^2 : \|\mathbf{x}\| \le 1\}$, let $\mathbf{q} = (2, 0)$, let $C = \operatorname{conv}(D \cup \{\mathbf{q}\})$, and put $\mathbf{p} = (\frac{1}{2}, \frac{\sqrt{3}}{2})$, so that $\|\mathbf{p}\| = 1$ and $\langle \mathbf{p}, \mathbf{q} \rangle = 1$. Since D is convex, Theorem 18.1.3 writes every point of C as $(1-s)\mathbf{d} + s\mathbf{q}$ with $\mathbf{d} \in D$ and $s \in [0, 1]$, whence

$$\langle \mathbf{p}, (1-s)\mathbf{d} + s\mathbf{q} \rangle = (1-s)\langle \mathbf{p}, \mathbf{d} \rangle + s \le (1-s)\|\mathbf{d}\| + s \le 1,$$

the first inequality by Theorem 11.1.5. Equality with $s < 1$ forces $\langle \mathbf{p}, \mathbf{d} \rangle = \|\mathbf{d}\| = 1$, hence $(\mathbf{p}, \mathbf{d})$ dependent and $\mathbf{d} = \mathbf{p}$; so $\{\mathbf{x} \in C : \langle \mathbf{p}, \mathbf{x} \rangle = 1\} = [\mathbf{p}, \mathbf{q}]$. Now $\{\mathbf{p}\}$ **is** a face of C in the sense of Definition 18.8.2: if $\mathbf{p} = t\mathbf{y} + (1-t)\mathbf{z}$ with $\mathbf{y}, \mathbf{z} \in C$ and $0 < t < 1$, then $1 = t\langle \mathbf{p}, \mathbf{y} \rangle + (1-t)\langle \mathbf{p}, \mathbf{z} \rangle$, and both inner products are at most 1, so both equal 1 and $\mathbf{y} = \mathbf{p} + a(\mathbf{q} - \mathbf{p})$, $\mathbf{z} = \mathbf{p} + b(\mathbf{q} - \mathbf{p})$ with $a, b \in [0, 1]$; then $ta + (1-t)b = 0$ forces $a = b = 0$. But no supporting hyperplane isolates $\mathbf{p}$. If $\langle \mathbf{c}, \mathbf{x} \rangle \le \beta$ on C with $\langle \mathbf{c}, \mathbf{p} \rangle = \beta$, then $\mathbf{c} \neq \mathbf{0}$, since $\mathbf{c} = \mathbf{0}$ gives the face C or $\emptyset$; testing the inequality at the point $\mathbf{c}/\|\mathbf{c}\|$ of D gives $\|\mathbf{c}\| \le \beta = \langle \mathbf{c}, \mathbf{p} \rangle \le \|\mathbf{c}\|$, so equality holds in Theorem 11.1.5 and $\mathbf{c} = \|\mathbf{c}\|\mathbf{p}$, $\beta = \|\mathbf{c}\|$. Then $\langle \mathbf{c}, \mathbf{q} \rangle = \|\mathbf{c}\|\langle \mathbf{p}, \mathbf{q} \rangle = \beta$, so $\mathbf{q}$ lies in the face too, and the face is not $\{\mathbf{p}\}$.

Exercise. C3: The optimal face has no better vertex outside it

Let $P \subseteq \mathbb{R}^n$ be a nonempty bounded polyhedron, $\mathbf{c} \in \mathbb{R}^n$, and let $F^\star$ be the optimal face of Proposition 22.3.5.

(a) Prove that $F^\star = \operatorname{conv}(F^\star \cap \operatorname{ext} P)$.

(b) Deduce that the optimal value of the program equals $\max\{\mathbf{c}^\top \mathbf{v} : \mathbf{v} \in \operatorname{ext} P\}$, a maximum over a finite set, and that this finite maximum is attained exactly at the vertices lying in $F^\star$.

Solution.

(a) $F^\star$ is a nonempty bounded polyhedron by Proposition 22.3.1, so Theorem 22.3.3 gives $F^\star = \operatorname{conv}(\operatorname{ext} F^\star)$. By the remark after Definition 22.3.1, $F^\star$ is a face of P in the sense of Definition 18.8.2, so Lemma 18.8.3 gives $\operatorname{ext} F^\star = F^\star \cap \operatorname{ext} P$. Substituting proves (a).

(b) By Corollary 18.9.3, $\operatorname{ext} P$ is finite, and it is nonempty by Theorem 22.3.3 since $P \neq \emptyset$. Write $\gamma^\star$ for the optimal value. Every vertex lies in P, so $\mathbf{c}^\top \mathbf{v} \le \gamma^\star$ for $\mathbf{v} \in \operatorname{ext} P$; and by (a) $F^\star$ contains a vertex, at which $\mathbf{c}^\top \mathbf{v} = \gamma^\star$. So the finite maximum equals $\gamma^\star$. A vertex $\mathbf{v}$ attains it exactly when $\mathbf{c}^\top \mathbf{v} = \gamma^\star$, which is exactly the condition $\mathbf{v} \in F^\star$ defining the optimal face.

22.4 Euclidean Motions

Affine geometry knows about lines, parallels, and ratios along a line, but not about length. Put the standard inner product back on $\mathbb{R}^n$ and a new question appears: which maps of space leave every distance unchanged? One might expect a large and unruly collection, since the condition says nothing about linearity, nothing about continuity, and nothing about where the origin goes. The first theorem of this section says otherwise. A map that preserves distances is forced to be affine, with an orthogonal linear part — and it is forced without assuming that it is surjective, which most treatments quietly do. The rest of the section turns that structure theorem into a complete list of the motions of the plane and of space.

Throughout, the field is $\mathbb{R}$, the space is $\mathbb{R}^n$ with the standard inner product, and the **distance** between $\mathbf{x}$ and $\mathbf{y}$ is $\|\mathbf{x}-\mathbf{y}\|$. Matrices are bold, points of $\mathbb{R}^n$ are bold vectors as usual, and we write $\mathbf{R}_\theta$ and the reflections of Theorem 11.7.6 for the elements of $\mathrm{O}(2)$.

22.4.1 Distance alone

Chapter 11 studied the **linear** maps preserving distance: the isometries, which in $\mathbb{R}^n$ are the orthogonal matrices (Theorem 11.7.3). But a translation $\mathbf{x} \mapsto \mathbf{x}+\mathbf{b}$ with $\mathbf{b} \neq \mathbf{0}$ also preserves every distance and is not linear, since it moves the origin. So the linear theory misses exactly what affine geometry was invented to supply, and the right notion drops linearity from the hypotheses altogether.

A Euclidean motion is a map of space that changes no distance.

Definition 22.4.1. Euclidean Motion

A **Euclidean motion** of $\mathbb{R}^n$, or simply a **motion**, is a map $f\colon \mathbb{R}^n \to \mathbb{R}^n$ such that

$$\|f(\mathbf{x})-f(\mathbf{y})\| = \|\mathbf{x}-\mathbf{y}\| \qquad \text{for all } \mathbf{x},\mathbf{y} \in \mathbb{R}^n.$$

No linearity, continuity, or surjectivity is assumed.

In words: the only requirement is that the distance between any two points equals the distance between their images. The quantifier is **for all** pairs, not for pairs at some special distance, and that turns out to be exactly the right amount of information.

Examples.

1. **Translations.** $t_{\mathbf{b}}(\mathbf{x}) = \mathbf{x}+\mathbf{b}$ satisfies $t_{\mathbf{b}}(\mathbf{x}) - t_{\mathbf{b}}(\mathbf{y}) = \mathbf{x}-\mathbf{y}$, so it is a motion for every $\mathbf{b}$. It is linear only when $\mathbf{b}=\mathbf{0}$.
2. **Orthogonal maps.** For $\mathbf{Q} \in \mathrm{O}(n)$, the map $\mathbf{x} \mapsto \mathbf{Q}\mathbf{x}$ is a motion: $\|\mathbf{Q}\mathbf{x}-\mathbf{Q}\mathbf{y}\| = \|\mathbf{Q}(\mathbf{x}-\mathbf{y})\| = \|\mathbf{x}-\mathbf{y}\|$, the last equality by Theorem 11.7.3 (a) applied to $T_{\mathbf{Q}}$, which is an isometry by part (f) of that theorem.
3. **Compositions.** If f and g are motions then so is $g \circ f$, since $\|g(f(\mathbf{x}))-g(f(\mathbf{y}))\| = \|f(\mathbf{x})-f(\mathbf{y})\| = \|\mathbf{x}-\mathbf{y}\|$. In particular $\mathbf{x} \mapsto \mathbf{Q}\mathbf{x}+\mathbf{b}$ is a motion for every $\mathbf{Q} \in \mathrm{O}(n)$ and $\mathbf{b} \in \mathbb{R}^n$.
4. **The degenerate case** $n=0$ is the one-point space and its only motion is the identity; for $n=1$ the motions are $x \mapsto \pm x + b$, as the theorem below will confirm.

A nonexample by minimal change. Take $f(\mathbf{x}) = 2\mathbf{x}$. It is linear, continuous, and bijective, and it multiplies every distance by the same factor 2. The one clause of Definition 22.4.1 it fails is the only one there is: $\|f(\mathbf{x})-f(\mathbf{y})\| = 2\|\mathbf{x}-\mathbf{y}\| \neq \|\mathbf{x}-\mathbf{y}\|$ as soon as $\mathbf{x} \neq \mathbf{y}$. A map that

rescales all distances equally is a **similarity**, not a motion, and this book does not pursue them.

Why this definition, and why it is enough. Nothing in it mentions the vector space structure of $\mathbb{R}^n$; it is a statement about $\mathbb{R}^n$ as a set with a distance. The content of the next theorem is that the vector space structure is recoverable anyway. That is worth pausing on: distances alone remember which points are midpoints, hence which maps are affine, hence all of the linear algebra.

Theorem 22.4.1. Every Motion Is Affine

Let $f\colon \mathbb{R}^n \to \mathbb{R}^n$ be a Euclidean motion. Then there are a unique $\mathbf{Q} \in \mathrm{O}(n)$ and a unique $\mathbf{b} \in \mathbb{R}^n$ with

$$f(\mathbf{x}) = \mathbf{Q}\mathbf{x} + \mathbf{b} \qquad \text{for every } \mathbf{x} \in \mathbb{R}^n.$$

In particular f is an affine map (Definition 22.1.3) whose linear part $\vec{f}$ (Theorem 22.1.5) is the isometry $T_{\mathbf{Q}}$.

Idea. Three steps. ① Move the origin out of the way: $g(\mathbf{x}) = f(\mathbf{x}) - f(\mathbf{0})$ is a motion fixing $\mathbf{0}$, so it preserves norms as well as distances. ② A map preserving both norms and distances preserves inner products, because the inner product is built from those two by polarization: $2\langle \mathbf{x}, \mathbf{y}\rangle = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 - \|\mathbf{x}-\mathbf{y}\|^2$. ③ A map preserving inner products is linear, and the way to see it is the reflex for showing a vector is zero: pair the vector $g(\mathbf{x}+\mathbf{y}) - g(\mathbf{x}) - g(\mathbf{y})$ with itself. Expanding the square turns it into a combination of inner products, each of which g has just been shown to preserve, and the combination is the square of $(\mathbf{x}+\mathbf{y}) - \mathbf{x} - \mathbf{y} = \mathbf{0}$.

Proof. **Step 1: reduce to a motion fixing the origin.** Put $\mathbf{b} = f(\mathbf{0})$ and $g(\mathbf{x}) = f(\mathbf{x}) - \mathbf{b}$. For all $\mathbf{x}, \mathbf{y}$,

$$\|g(\mathbf{x}) - g(\mathbf{y})\| = \|f(\mathbf{x}) - f(\mathbf{y})\| = \|\mathbf{x} - \mathbf{y}\|,$$

so g is a motion, and $g(\mathbf{0}) = \mathbf{0}$. Taking $\mathbf{y} = \mathbf{0}$ gives

$$\|g(\mathbf{x})\| = \|g(\mathbf{x}) - g(\mathbf{0})\| = \|\mathbf{x} - \mathbf{0}\| = \|\mathbf{x}\|$$

for every $\mathbf{x}$: g preserves norms.

Step 2: g preserves inner products. By Theorem 11.1.2 (c), for real vectors $\mathbf{u}, \mathbf{v}$,

$$\|\mathbf{u} - \mathbf{v}\|^2 = \|\mathbf{u}\|^2 - 2\langle \mathbf{u}, \mathbf{v}\rangle + \|\mathbf{v}\|^2,$$
$$\text{so} \qquad 2\langle \mathbf{u}, \mathbf{v}\rangle = \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2 - \|\mathbf{u} - \mathbf{v}\|^2.$$

This is the polarization identity of Lemma 11.7.2 in the form that uses the distance: the inner product is determined by norms and distances alone. Applying it to $g(\mathbf{x}), g(\mathbf{y})$ and then to $\mathbf{x}, \mathbf{y}$, and using Step 1 three times,

$$\begin{aligned} 2\langle g(\mathbf{x}), g(\mathbf{y})\rangle &= \|g(\mathbf{x})\|^2 + \|g(\mathbf{y})\|^2 - \|g(\mathbf{x}) - g(\mathbf{y})\|^2 \\ &= \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 - \|\mathbf{x} - \mathbf{y}\|^2 = 2\langle \mathbf{x}, \mathbf{y}\rangle. \end{aligned}$$

So $\langle g(\mathbf{x}), g(\mathbf{y})\rangle = \langle \mathbf{x}, \mathbf{y}\rangle$ for all $\mathbf{x}, \mathbf{y}$.

Step 3: g is linear. Fix $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ and expand the squared norm of $\mathbf{d} = g(\mathbf{x}+\mathbf{y}) - g(\mathbf{x}) - g(\mathbf{y})$.

For any three vectors $\mathbf{r}, \mathbf{s}, \mathbf{z}$, repeated use of Theorem 11.1.2 (c) gives

$$\begin{aligned}\|\mathbf{r}-\mathbf{s}-\mathbf{z}\|^2 = \|\mathbf{r}\|^2 + \|\mathbf{s}\|^2 + \|\mathbf{z}\|^2 \\ - 2\langle \mathbf{r}, \mathbf{s}\rangle - 2\langle \mathbf{r}, \mathbf{z}\rangle + 2\langle \mathbf{s}, \mathbf{z}\rangle.\end{aligned}$$

Apply this with $\mathbf{r} = g(\mathbf{x}+\mathbf{y})$, $\mathbf{s} = g(\mathbf{x})$, $\mathbf{z} = g(\mathbf{y})$. Every term on the right is a norm or an inner product of values of g, and Steps 1 and 2 replace each by the corresponding quantity for $\mathbf{x}+\mathbf{y}$, $\mathbf{x}$, $\mathbf{y}$. The result is the same expression with $\mathbf{r} = \mathbf{x}+\mathbf{y}$, $\mathbf{s} = \mathbf{x}$, $\mathbf{z} = \mathbf{y}$, that is, $\|(\mathbf{x}+\mathbf{y})-\mathbf{x}-\mathbf{y}\|^2 = 0$. Hence $\|\mathbf{d}\|^2 = 0$ and $\mathbf{d} = \mathbf{0}$ by Theorem 11.1.2 (a), so

$$g(\mathbf{x}+\mathbf{y}) = g(\mathbf{x}) + g(\mathbf{y}).$$

For scalars, let $\lambda \in \mathbb{R}$ and expand in the same way:

$$\begin{aligned}\|g(\lambda\mathbf{x}) - \lambda g(\mathbf{x})\|^2 &= \|g(\lambda\mathbf{x})\|^2 - 2\lambda\langle g(\lambda\mathbf{x}), g(\mathbf{x})\rangle + \lambda^2\|g(\mathbf{x})\|^2 \\ &= \|\lambda\mathbf{x}\|^2 - 2\lambda\langle \lambda\mathbf{x}, \mathbf{x}\rangle + \lambda^2\|\mathbf{x}\|^2 \\ &= \lambda^2\|\mathbf{x}\|^2 - 2\lambda^2\|\mathbf{x}\|^2 + \lambda^2\|\mathbf{x}\|^2 = 0,\end{aligned}$$

so $g(\lambda\mathbf{x}) = \lambda g(\mathbf{x})$. Therefore g is linear.

Step 4: the matrix. Let $\mathbf{Q}$ be the matrix of g in the standard basis, so $g(\mathbf{x}) = \mathbf{Q}\mathbf{x}$. By Step 2, g preserves inner products, so $\mathbf{Q} \in \mathrm{O}(n)$ by Theorem 11.7.3 (b) and (f), the standard basis being orthonormal. Then $f(\mathbf{x}) = g(\mathbf{x}) + \mathbf{b} = \mathbf{Q}\mathbf{x} + \mathbf{b}$.

Uniqueness. If $\mathbf{Q}\mathbf{x} + \mathbf{b} = \mathbf{Q}'\mathbf{x} + \mathbf{b}'$ for every $\mathbf{x}$, then $\mathbf{x} = \mathbf{0}$ gives $\mathbf{b} = \mathbf{b}'$, and then $(\mathbf{Q} - \mathbf{Q}')\mathbf{x} = \mathbf{0}$ for every $\mathbf{x}$ gives $\mathbf{Q} = \mathbf{Q}'$. This proves the theorem.

■

Remark. **Surjectivity was never used.** The definition of a motion asks only that distances be preserved, and the proof never produces a preimage of anything. Most treatments define a motion to be a distance-preserving **bijection**, and then the bijectivity is an unused hypothesis — as the next corollary shows, it comes for free.

Corollary 22.4.2. Motions Are Bijections

Every Euclidean motion f of $\mathbb{R}^n$ is a bijection, and f^{-1} is again a motion. The linear part of f^{-1} is $\mathbf{Q}^{\top}$, where $\mathbf{Q}$ is the linear part of f.

Proof. Write $f(\mathbf{x}) = \mathbf{Q}\mathbf{x} + \mathbf{b}$ by Theorem 22.4.1. Since $\mathbf{Q} \in \mathrm{O}(n)$, $\mathbf{Q}$ is invertible with $\mathbf{Q}^{-1} = \mathbf{Q}^{\top}$ (Theorem 11.7.3 (g1)), so the map $h(\mathbf{y}) = \mathbf{Q}^{\top}(\mathbf{y} - \mathbf{b}) = \mathbf{Q}^{\top}\mathbf{y} - \mathbf{Q}^{\top}\mathbf{b}$ satisfies $h(f(\mathbf{x})) = \mathbf{x}$ and $f(h(\mathbf{y})) = \mathbf{y}$. Hence f is a bijection with inverse h, which is a motion by examples 2 and 3 after Definition 22.4.1, with linear part $\mathbf{Q}^{\top}$.

■

Warning. **The Euclidean norm is doing real work.** Step 2 is the only place where the special shape of $\|\cdot\|$ is used, and for other norms the phenomenon fails. Here is a distance-

preserving map that is not affine, for a norm of Chapter 16 and between spaces of different dimensions. Give $\mathbb{R}^2$ the norm $\|(y_1, y_2)\|_\infty = \max(|y_1|, |y_2|)$ of Chapter 16 and consider $\varphi\colon \mathbb{R} \to \mathbb{R}^2$, $\varphi(t) = (t, |t|)$. For $s, t \in \mathbb{R}$, Lemma 16.1.1 applied to the absolute value as a norm on $\mathbb{R}$ gives $\big||s| - |t|\big| \le |s - t|$, so

$$\|\varphi(s) - \varphi(t)\|_\infty = \max\big(|s - t|,\ \big||s| - |t|\big|\big) = |s - t|.$$

So φ preserves distances exactly, and it is visibly not affine: $\varphi(-1) = (-1, 1)$, $\varphi(0) = (0, 0)$, and $\varphi(1) = (1, 1)$, yet $\varphi(0)$ is not the midpoint of $\varphi(-1)$ and $\varphi(1)$, which is $(0, 1)$. Replacing $|t|$ by any function $\mathbb{R} \to \mathbb{R}$ that never changes by more than its input does gives another one. What makes Theorem 22.4.1 true is polarization, an identity special to the norms that come from an inner product, and not any property shared by all norms.

The determinant of the linear part now splits the motions into two halves, exactly as it splits $\mathrm{O}(n)$.

Definition 22.4.2. Direct and Opposite Motions

Let $f(\mathbf{x}) = \mathbf{Q}\mathbf{x}+\mathbf{b}$ be a motion of $\mathbb{R}^n$. It is **direct**, or orientation-preserving, if $\det \mathbf{Q} = 1$, and **opposite**, or orientation-reversing, if $\det \mathbf{Q} = -1$. These are the only possibilities: $\mathbf{Q}^{\top}\mathbf{Q} = \mathbf{I}_n$, and $\det \mathbf{Q}^{\top} = \det \mathbf{Q}$ (Theorem 7.4.5), so $(\det \mathbf{Q})^2 = \det(\mathbf{Q}^{\top}\mathbf{Q}) = 1$ by Theorem 7.5.4.

Quick check. For which n is the map $f(\mathbf{x}) = -\mathbf{x} + \mathbf{b}$ a direct motion of $\mathbb{R}^n$? Describe it for $n = 2$ and $n = 3$.

Solution. It is a motion with linear part $-\mathbf{I}_n$, and $\det(-\mathbf{I}_n) = (-1)^n$, so it is direct exactly when n is even. For $n = 2$ it is direct: $-\mathbf{I}_2 = \mathbf{R}_\pi$, and f is the half-turn about the point $\frac{1}{2}\mathbf{b}$, which is its unique fixed point, since $-\mathbf{x} + \mathbf{b} = \mathbf{x}$ forces $\mathbf{x} = \frac{1}{2}\mathbf{b}$. For $n = 3$ it is opposite: it is the point reflection $\mathbf{x} \mapsto 2(\frac{1}{2}\mathbf{b}) - \mathbf{x}$ in the point $\frac{1}{2}\mathbf{b}$, which sends each point to the point diametrically opposite it through that center.

22.4.2 The group of motions

Write $\mathrm{E}(n)$ for the set of Euclidean motions of $\mathbb{R}^n$ and $\mathrm{Trans}(n) = \{t_{\mathbf{b}} : \mathbf{b} \in \mathbb{R}^n\}$ for the translations. One proposition records how these fit together, and every part of it is a line of computation.

Proposition 22.4.3. The Euclidean Group

(a) $\mathrm{E}(n)$ is a group under composition.

(b) The map $\lambda\colon \mathrm{E}(n) \to \mathrm{O}(n)$ sending a motion to its linear part is a surjective group homomorphism.

(c) $\ker \lambda = \mathrm{Trans}(n)$, and $\mathbf{b} \mapsto t_{\mathbf{b}}$ is a group isomorphism $(\mathbb{R}^n, +) \to \mathrm{Trans}(n)$.

(d) For every motion f with linear part $\mathbf{Q}$ and every $\mathbf{b} \in \mathbb{R}^n$,

$$f \circ t_{\mathbf{b}} \circ f^{-1} = t_{\mathbf{Q}\mathbf{b}}.$$

In particular a conjugate of a translation is a translation.

Idea. Write everything in the form $\mathbf{x} \mapsto \mathbf{Q}\mathbf{x} + \mathbf{b}$. Composing two such maps multiplies the matrices and moves the translation vectors through, so the matrix is exactly the part that composes without interference — which is what a homomorphism is. Part (d) is one substitution, and it explains where the translations sit: they are the motions that every other motion pushes around among themselves.

Proof. By Theorem 22.4.1 every motion is $\mathbf{x} \mapsto \mathbf{Q}\mathbf{x}+\mathbf{b}$ with $\mathbf{Q} \in \mathrm{O}(n)$, and conversely every such map is a motion (examples 2 and 3 after Definition 22.4.1). Write $f_{\mathbf{Q},\mathbf{b}}$ for it. A direct computation gives the composition rule:

$$f_{\mathbf{R},\mathbf{c}}(f_{\mathbf{Q},\mathbf{b}}(\mathbf{x})) = \mathbf{R}(\mathbf{Q}\mathbf{x}+\mathbf{b})+\mathbf{c} = (\mathbf{R}\mathbf{Q})\mathbf{x}+(\mathbf{R}\mathbf{b}+\mathbf{c}),$$

so

$$f_{\mathbf{R},\mathbf{c}} \circ f_{\mathbf{Q},\mathbf{b}} = f_{\mathbf{R}\mathbf{Q},\ \mathbf{R}\mathbf{b}+\mathbf{c}}. \tag{$*$}$$

(a) By Corollary 22.4.2 every motion is a bijection of $\mathbb{R}^n$, so $\mathrm{E}(n)$ is a subset of the group of all bijections $\mathbb{R}^n \to \mathbb{R}^n$ under composition, which is a group by the argument of Example (Groups) for S_n: composition of functions is associative by Theorem 0.4.2, the identity map is a two-sided identity, and a bijection has a two-sided inverse (Theorem 0.4.7), none of which used finiteness. It contains $\mathrm{id} = f_{\mathbf{I},\mathbf{0}}$, it is closed under composition by $(*)$ and because $\mathbf{R}\mathbf{Q} \in \mathrm{O}(n)$ (Proposition 11.7.4), and it is closed under inverses by Corollary 22.4.2. By Definition 0.10.2 it is a subgroup, hence a group.

(b) With $\lambda(f_{\mathbf{Q},\mathbf{b}}) = \mathbf{Q}$, well defined by the uniqueness in Theorem 22.4.1, $(*)$ reads $\lambda(g \circ f) = \lambda(g)\lambda(f)$, so λ is a homomorphism (Definition 0.10.3). It is surjective because $\lambda(f_{\mathbf{Q},\mathbf{0}}) = \mathbf{Q}$ for every $\mathbf{Q} \in \mathrm{O}(n)$.

(c) $\lambda(f_{\mathbf{Q},\mathbf{b}}) = \mathbf{I}$ if and only if $\mathbf{Q} = \mathbf{I}$, that is, $f = t_{\mathbf{b}}$; so $\ker\lambda = \mathrm{Trans}(n)$. By $(*)$ with $\mathbf{R} = \mathbf{Q} = \mathbf{I}$, $t_{\mathbf{c}} \circ t_{\mathbf{b}} = t_{\mathbf{b}+\mathbf{c}}$, so $\mathbf{b} \mapsto t_{\mathbf{b}}$ is a homomorphism from $(\mathbb{R}^n, +)$; it is injective because $t_{\mathbf{b}} = \mathrm{id}$ forces $\mathbf{b} = t_{\mathbf{b}}(\mathbf{0}) = \mathbf{0}$, and its image is $\mathrm{Trans}(n)$ by definition.

(d) By Corollary 22.4.2, $f^{-1}(\mathbf{x}) = \mathbf{Q}^{\top}(\mathbf{x}-\mathbf{c})$ where $f = f_{\mathbf{Q},\mathbf{c}}$. Then

$$f(t_{\mathbf{b}}(f^{-1}(\mathbf{x}))) = \mathbf{Q}(\mathbf{Q}^{\top}(\mathbf{x}-\mathbf{c})+\mathbf{b})+\mathbf{c} = (\mathbf{x}-\mathbf{c})+\mathbf{Q}\mathbf{b}+\mathbf{c} = \mathbf{x}+\mathbf{Q}\mathbf{b},$$

which is $t_{\mathbf{Q}\mathbf{b}}(\mathbf{x})$. This proves the proposition.

■

Remark. In the vocabulary of group theory, (c) and (d) say that $\mathrm{Trans}(n)$ is a **normal** subgroup of $\mathrm{E}(n)$ and that the quotient group $\mathrm{E}(n)/\mathrm{Trans}(n)$ is isomorphic to $\mathrm{O}(n)$. This book has not built quotient groups — Chapter 4 §04's quotients are quotients of **vector spaces**, and $\mathrm{E}(n)$ is not one — so the proposition records instead the exact content such a quotient would package: a surjective homomorphism onto $\mathrm{O}(n)$ whose kernel is exactly the translations. Concretely, the motions with a given linear part $\mathbf{Q}$ are exactly the maps $t_{\mathbf{b}} \circ f_{\mathbf{Q},\mathbf{0}}$, one for each $\mathbf{b}$, which is the statement that the fibers of λ are the translates of its kernel.

Warning. $\mathrm{E}(n)$ **is not the direct product of** $\mathrm{O}(n)$ **and** $\mathrm{Trans}(n)$**.** Part (d) shows why: conjugating $t_{\mathbf{b}}$ by a rotation gives $t_{\mathbf{Qb}}$, not $t_{\mathbf{b}}$, so the two subgroups do **not** commute with one another. Concretely in $\mathbb{R}^2$, with $\mathbf{Q} = \mathbf{R}_{\pi/2}$ and $\mathbf{b} = \mathbf{e}_1$, the composite $f_{\mathbf{Q},\mathbf{0}} \circ t_{\mathbf{e}_1}$ sends $\mathbf{0}$ to $\mathbf{Qe}_1 = \mathbf{e}_2$, while $t_{\mathbf{e}_1} \circ f_{\mathbf{Q},\mathbf{0}}$ sends $\mathbf{0}$ to $\mathbf{e}_1$. A rotation followed by a translation is not the same motion as the translation followed by the rotation.

22.4.3 Motions of the plane

Now the classification. Four kinds of map appear, and each is built from a linear one by moving the origin to a chosen point.

Definition 22.4.3. Translation, Rotation, Reflection, Glide Reflection

Let $\mathbf{p} \in \mathbb{R}^2$.

(a) The **translation** by $\mathbf{b}$ is $t_{\mathbf{b}}(\mathbf{x}) = \mathbf{x} + \mathbf{b}$.

(b) The **rotation about** $\mathbf{p}$ **by** θ is $\mathbf{x} \mapsto \mathbf{R}_\theta(\mathbf{x} - \mathbf{p}) + \mathbf{p}$.

(c) Let $L = \mathbf{p} + \mathrm{span}(\mathbf{u})$ be a line, $\mathbf{u} \neq \mathbf{0}$, and let $\mathbf{Q}_L \in \mathrm{O}(2)$ be the linear reflection fixing $\mathrm{span}(\mathbf{u})$ and negating $\mathrm{span}(\mathbf{u})^\perp$. The **reflection in** L is $s_L(\mathbf{x}) = \mathbf{Q}_L(\mathbf{x} - \mathbf{p}) + \mathbf{p}$.

(d) Let $\mathbf{v} \neq \mathbf{0}$ be parallel to L, that is, $\mathbf{v} \in \mathrm{span}(\mathbf{u})$. The **glide reflection** with axis L and vector $\mathbf{v}$ is $t_{\mathbf{v}} \circ s_L$.

The reflection s_L does not depend on which point $\mathbf{p} \in L$ is used: replacing $\mathbf{p}$ by $\mathbf{p} + c\mathbf{u}$ changes $\mathbf{Q}_L(\mathbf{x}-\mathbf{p})+\mathbf{p}$ by $-c\mathbf{Q}_L\mathbf{u}+c\mathbf{u} = \mathbf{0}$, since $\mathbf{Q}_L\mathbf{u} = \mathbf{u}$. All four maps are motions, being composites of translations and orthogonal maps. Their linear parts are $\mathbf{I}$, $\mathbf{R}_\theta$, $\mathbf{Q}_L$, and $\mathbf{Q}_L$, so the first two are direct and the last two are opposite.

Theorem 22.4.4. Classification of the Motions of the Plane

Let $f(\mathbf{x}) = \mathbf{Qx} + \mathbf{b}$ be a motion of $\mathbb{R}^2$. Then **exactly one** of the following holds.

(a) $f = \mathrm{id}$.

(b) $f = t_{\mathbf{b}}$ is a translation with $\mathbf{b} \neq \mathbf{0}$; the vector $\mathbf{b}$ is unique.

(c) f is a rotation about a point $\mathbf{p}$ by an angle $\theta \in (0, 2\pi)$; the point $\mathbf{p}$ and the angle θ are unique.

(d) f is a reflection s_L in a line L; the line L is unique.

(e) f is a glide reflection with axis L and vector $\mathbf{v} \neq \mathbf{0}$; the pair $(L, \mathbf{v})$ is unique.

The five cases are told apart by the determinant of $\mathbf{Q}$ and the fixed-point set $\mathrm{Fix}(f) = \{\mathbf{x} : f(\mathbf{x}) = \mathbf{x}\}$:

	(a)	(b)	(c)	(d)	(e)
$\det \mathbf{Q}$	1	1	1	-1	-1
$\mathrm{Fix}(f)$	$\mathbb{R}^2$	$\emptyset$	a point	a line	$\emptyset$

Idea. Split on $\det \mathbf{Q}$, which Theorem 11.7.6 turns into "$\mathbf{Q}$ is a rotation $\mathbf{R}_\theta$" or "$\mathbf{Q}$ is a linear reflection".

For $\det \mathbf{Q} = 1$ with $\mathbf{Q} \neq \mathbf{I}$, the matrix $\mathbf{I} - \mathbf{Q}$ is invertible, and solving $(\mathbf{I} - \mathbf{Q})\mathbf{p} = \mathbf{b}$ is exactly finding a fixed point; once there is one, moving the origin to it makes f linear.

For $\det \mathbf{Q} = -1$ there may be no fixed point, so instead split $\mathbf{b}$ into its components along the mirror and across it. The **across** part is absorbed by sliding the mirror halfway; the **along** part is what is left over, and it is the glide.

Proof. *Case* $\det \mathbf{Q} = 1$. By Theorem 11.7.6 (a), $\mathbf{Q} = \mathbf{R}_\theta$ for a unique $\theta \in [0, 2\pi)$.

If $\theta = 0$ then $\mathbf{Q} = \mathbf{I}$ and $f = t_{\mathbf{b}}$, which is case (a) when $\mathbf{b} = \mathbf{0}$ and case (b) otherwise; the vector is $\mathbf{b} = f(\mathbf{0})$, hence unique. Its fixed-point set is $\mathbb{R}^2$ or $\emptyset$.

If $\theta \neq 0$, then

$$\det(\mathbf{I} - \mathbf{R}_\theta) = (1 - \cos\theta)^2 + \sin^2\theta = 2 - 2\cos\theta.$$

Were this 0 we would have $\cos\theta = 1$, hence $\sin^2\theta = 1 - \cos^2\theta = 0$ and $\mathbf{R}_\theta = \mathbf{I} = \mathbf{R}_0$, so $\theta = 0$ by the uniqueness in Theorem 11.7.6 (a), a contradiction. So $\mathbf{I} - \mathbf{Q}$ is invertible and there is a unique $\mathbf{p}$ with $(\mathbf{I} - \mathbf{Q})\mathbf{p} = \mathbf{b}$, that is, with $f(\mathbf{p}) = \mathbf{Q}\mathbf{p} + \mathbf{b} = \mathbf{p}$. For every $\mathbf{x}$,

$$f(\mathbf{x}) - \mathbf{p} = \mathbf{Q}\mathbf{x} + \mathbf{b} - \mathbf{p} = \mathbf{Q}\mathbf{x} - \mathbf{Q}\mathbf{p} = \mathbf{Q}(\mathbf{x} - \mathbf{p}),$$

so f is the rotation about $\mathbf{p}$ by θ: case (c). Its fixed points satisfy $(\mathbf{I} - \mathbf{Q})(\mathbf{x} - \mathbf{p}) = \mathbf{0}$, hence $\mathbf{x} = \mathbf{p}$, so $\mathrm{Fix}(f) = \{\mathbf{p}\}$ and $\mathbf{p}$ is determined by f; θ is determined because $\mathbf{Q}$ is.

Case $\det \mathbf{Q} = -1$. By Theorem 11.7.6 (b), $\mathbf{Q}$ fixes a line $\mathrm{span}(\mathbf{u})$ pointwise and negates $\mathrm{span}(\mathbf{u})^\perp = \mathrm{span}(\mathbf{w})$, where $(\mathbf{u}, \mathbf{w})$ is an orthonormal basis of $\mathbb{R}^2$. By Theorem 11.3.2 write

$$\mathbf{b} = \mathbf{b}_\parallel + \mathbf{b}_\perp, \qquad \mathbf{b}_\parallel = \langle \mathbf{b}, \mathbf{u} \rangle \mathbf{u}, \quad \mathbf{b}_\perp = \langle \mathbf{b}, \mathbf{w} \rangle \mathbf{w}.$$

Put $\mathbf{c} = \frac{1}{2}\mathbf{b}_\perp$ and $L = \mathbf{c} + \mathrm{span}(\mathbf{u})$. The reflection in L is $s_L(\mathbf{x}) = \mathbf{Q}(\mathbf{x} - \mathbf{c}) + \mathbf{c} = \mathbf{Q}\mathbf{x} + (\mathbf{I} - \mathbf{Q})\mathbf{c}$, and $\mathbf{Q}\mathbf{c} = -\mathbf{c}$ because $\mathbf{c} \in \mathrm{span}(\mathbf{w})$, so $(\mathbf{I} - \mathbf{Q})\mathbf{c} = 2\mathbf{c} = \mathbf{b}_\perp$ and

$$s_L(\mathbf{x}) = \mathbf{Q}\mathbf{x} + \mathbf{b}_\perp, \qquad f(\mathbf{x}) = \mathbf{Q}\mathbf{x} + \mathbf{b}_\perp + \mathbf{b}_\parallel = (t_{\mathbf{b}_\parallel} \circ s_L)(\mathbf{x}).$$

If $\mathbf{b}_\parallel = \mathbf{0}$, then $f = s_L$: case (d). Its fixed set is L, since $\mathbf{Q}(\mathbf{x} - \mathbf{c}) = \mathbf{x} - \mathbf{c}$ holds exactly when $\mathbf{x} - \mathbf{c} \in \mathrm{span}(\mathbf{u})$; so L is determined by f.

If $\mathbf{b}_\parallel \neq \mathbf{0}$, then f is the glide reflection with axis L and vector $\mathbf{b}_\parallel$, which is parallel to L: case (e). For its fixed set, compute the square of f:

$$f(f(\mathbf{x})) = \mathbf{Q}(\mathbf{Q}\mathbf{x} + \mathbf{b}) + \mathbf{b} = \mathbf{x} + (\mathbf{Q}\mathbf{b} + \mathbf{b}) = \mathbf{x} + 2\mathbf{b}_\parallel,$$

using $\mathbf{Q}^2 = \mathbf{I}$ and $\mathbf{Q}\mathbf{b} = \mathbf{b}_\parallel - \mathbf{b}_\perp$. So $f \circ f = t_{2\mathbf{b}_\parallel} \neq \mathrm{id}$, and a fixed point of f would be a fixed point of $f \circ f$; hence $\mathrm{Fix}(f) = \emptyset$.

Uniqueness in case (e). Suppose also $f = t_{\mathbf{v}} \circ s_{L'}$ with $\mathbf{v} \neq \mathbf{0}$ parallel to L'. Picking $\mathbf{p}' \in L'$ gives $f(\mathbf{x}) = \mathbf{Q}_{L'}\mathbf{x} + (\mathbf{I} - \mathbf{Q}_{L'})\mathbf{p}' + \mathbf{v}$, so comparing linear parts (unique by Theorem 22.4.1) gives $\mathbf{Q}_{L'} = \mathbf{Q}$, and hence L' has direction $\mathrm{span}(\mathbf{u})$ and $\mathbf{v} \in \mathrm{span}(\mathbf{u})$. Now $\mathbf{I} - \mathbf{Q}$ sends $\mathbf{u}$ to $\mathbf{0}$ and $\mathbf{w}$ to $2\mathbf{w}$, so $(\mathbf{I} - \mathbf{Q})\mathbf{p}' \in \mathrm{span}(\mathbf{w})$, and

$$\mathbf{b} = (\mathbf{I} - \mathbf{Q})\mathbf{p}' + \mathbf{v}, \qquad (\mathbf{I} - \mathbf{Q})\mathbf{p}' \in \mathrm{span}(\mathbf{w}), \quad \mathbf{v} \in \mathrm{span}(\mathbf{u}).$$

Comparing with $\mathbf{b} = \mathbf{b}_{\parallel} + \mathbf{b}_{\perp}$ and using the uniqueness of the decomposition in Theorem 11.3.2, $\mathbf{v} = \mathbf{b}_{\parallel}$. Finally $s_{L'} = t_{-\mathbf{v}} \circ f = s_L$, and a reflection determines its axis as its fixed set, so $L' = L$.

Exactly one case holds. The table follows from what was computed in each case, and it separates the five: within $\det \mathbf{Q} = 1$ the fixed sets $\mathbb{R}^2$, $\emptyset$, a single point are distinct, and within $\det \mathbf{Q} = -1$ a line and $\emptyset$ are distinct. This proves the theorem.

■

So there are only four shapes of plane motion, and the surprise is the fourth. A student asked to list them usually produces translations, rotations, and reflections, and stops; the glide reflection is the one that has no fixed point and no visible center, and it is forced on us by the two-line computation $f \circ f = t_{2\mathbf{b}_{\parallel}}$.

Example. Naming a plane motion

Classify the motion $f(\mathbf{x}) = \mathbf{Q}\mathbf{x} + \mathbf{b}$ of $\mathbb{R}^2$ with

$$\mathbf{Q} = \tfrac{1}{5}\begin{pmatrix} -3 & 4 \\ 4 & 3 \end{pmatrix}, \qquad \mathbf{b} = (5, 0).$$

Solution. The linear part. The columns $\frac{1}{5}(-3, 4)$ and $\frac{1}{5}(4, 3)$ have norm $\frac{1}{5}\sqrt{9+16} = 1$ and inner product $\frac{1}{25}(-12 + 12) = 0$, so $\mathbf{Q} \in \mathrm{O}(2)$ by Theorem 11.7.3 (f), and $\det \mathbf{Q} = \frac{1}{25}(-9 - 16) = -1$. So f is opposite, and we are in case (d) or (e) of Theorem 22.4.4.

The mirror direction. Solve $\mathbf{Q}\mathbf{x} = \mathbf{x}$: the first row gives $-3x_1 + 4x_2 = 5x_1$, that is $x_2 = 2x_1$, and the second row gives $4x_1 + 3x_2 = 5x_2$, the same condition. So $\mathbf{Q}$ fixes $\operatorname{span}(\mathbf{u}')$ with $\mathbf{u}' = (1, 2)$, and negates $\operatorname{span}(\mathbf{w}')$ with $\mathbf{w}' = (2, -1)$: indeed $\mathbf{Q}(2, -1) = \frac{1}{5}(-6 - 4,\ 8 - 3) = (-2, 1)$.

Splitting $\mathbf{b}$. Since $\|\mathbf{u}'\|^2 = 5$,

$$\mathbf{b}_{\parallel} = \frac{\langle \mathbf{b}, \mathbf{u}' \rangle}{5}\mathbf{u}' = \frac{5}{5}(1, 2) = (1, 2), \qquad \mathbf{b}_{\perp} = \mathbf{b} - \mathbf{b}_{\parallel} = (4, -2) = 2\,\mathbf{w}'.$$

As $\mathbf{b}_{\parallel} \neq \mathbf{0}$, f is a **glide reflection**. Its axis is $L = \mathbf{c} + \operatorname{span}((1, 2))$ with $\mathbf{c} = \frac{1}{2}\mathbf{b}_{\perp} = (2, -1)$, and its vector is $(1, 2)$.

Checks. The point $(2, -1)$ should go to $(2, -1) + (1, 2) = (3, 1)$, and indeed $f(2, -1) = \frac{1}{5}(-6 - 4,\ 8 - 3) + (5, 0) = (-2, 1) + (5, 0) = (3, 1)$, which lies on L. And $f \circ f$ should be the translation by $2\mathbf{b}_{\parallel} = (2, 4)$: $f(f(\mathbf{x})) = \mathbf{x} + \mathbf{Q}\mathbf{b} + \mathbf{b}$, with $\mathbf{Q}\mathbf{b} = \frac{1}{5}(-15, 20) = (-3, 4)$, so $\mathbf{Q}\mathbf{b} + \mathbf{b} = (2, 4)$.

The next example is the one that explains where reflections sit in the group: they generate everything, and two of them make a direct motion.

Example. Two mirrors

(a) Reflect in the line $L_1 = \operatorname{span}((1, 0))$ and then in $L_2 = \operatorname{span}((1, 1))$. Identify the

composite.

(b) Reflect in $M_1 = \{x_2 = 0\}$ and then in $M_2 = \{x_2 = 1\}$. Identify the composite.

(c) Prove the general statements: two reflections in lines through a common point $\mathbf{p}$ meeting at angle α compose to the rotation about $\mathbf{p}$ by 2α; two reflections in parallel lines compose to a translation by twice the perpendicular vector between them.

Solution.

(a) The reflection in the first coordinate axis is $\mathbf{Q}_1 = \operatorname{diag}(1, -1)$, and the reflection in the line $x_2 = x_1$ swaps the coordinates, $\mathbf{Q}_2 = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$ (each fixes its line pointwise and negates the orthogonal complement, which is the description in Theorem 11.7.6 (b)). Then

$$\mathbf{Q}_2\mathbf{Q}_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} = \mathbf{R}_{\pi/2}.$$

The two lines meet at angle $\pi/4$, and the composite is the rotation by $2 \cdot \pi/4 = \pi/2$ about their common point $\mathbf{0}$.

(b) $s_{M_1}(x_1, x_2) = (x_1, -x_2)$, and $s_{M_2}(x_1, x_2) = (x_1, 2 - x_2)$, since reflecting in the line $x_2 = 1$ fixes the first coordinate and sends x_2 to $2 - x_2$. Composing, $s_{M_2}(s_{M_1}(\mathbf{x})) = (x_1, 2 + x_2)$, the translation by $(0, 2)$ — twice the perpendicular vector $(0, 1)$ from M_1 to M_2.

(c) Let $L_i = \mathbf{p} + \operatorname{span}(\mathbf{u}_i)$ with $\mathbf{u}_i$ unit vectors, meeting at $\mathbf{p}$. Moving the origin to $\mathbf{p}$ (that is, conjugating by $t_{-\mathbf{p}}$) we may assume $\mathbf{p} = \mathbf{0}$, because $s_{L_i} = t_{\mathbf{p}} \circ \mathbf{Q}_{L_i} \circ t_{-\mathbf{p}}$ and the two translations cancel in the middle of the composite. By Theorem 11.7.6 (b) the linear reflection in the line through $\mathbf{0}$ at angle φ is $\mathbf{M}_{2\varphi}$, the matrix with rows $(\cos 2\varphi, \sin 2\varphi)$ and $(\sin 2\varphi, -\cos 2\varphi)$. Writing $\mathbf{D} = \operatorname{diag}(1, -1)$, one checks $\mathbf{M}_\psi = \mathbf{R}_\psi\mathbf{D}$ and $\mathbf{D}\mathbf{R}_\psi\mathbf{D} = \mathbf{R}_{-\psi}$, the second because $\mathbf{D}$ changes the sign of $\sin\psi$ in both off-diagonal entries. Hence, for lines at angles φ_1 and φ_2 with $\alpha = \varphi_2 - \varphi_1$,

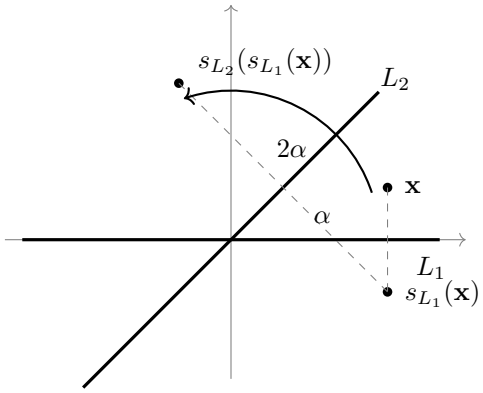
$$\mathbf{M}_{2\varphi_2}\mathbf{M}_{2\varphi_1} = \mathbf{R}_{2\varphi_2}\mathbf{D}\mathbf{R}_{2\varphi_1}\mathbf{D} = \mathbf{R}_{2\varphi_2}\mathbf{R}_{-2\varphi_1} = \mathbf{R}_{2\alpha},$$

a rotation about $\mathbf{0}$ by 2α; conjugating back gives the rotation about $\mathbf{p}$.

For parallel lines $L_i = \mathbf{c}_i + \operatorname{span}(\mathbf{u})$ the linear part is the same reflection $\mathbf{Q}$ for both, and $s_{L_i}(\mathbf{x}) = \mathbf{Q}\mathbf{x} + (\mathbf{I} - \mathbf{Q})\mathbf{c}_i$. Therefore

$$\begin{aligned} s_{L_2}(s_{L_1}(\mathbf{x})) &= \mathbf{Q}(\mathbf{Q}\mathbf{x} + (\mathbf{I} - \mathbf{Q})\mathbf{c}_1) + (\mathbf{I} - \mathbf{Q})\mathbf{c}_2 \\ &= \mathbf{x} + (\mathbf{Q} - \mathbf{I})\mathbf{c}_1 + (\mathbf{I} - \mathbf{Q})\mathbf{c}_2 = \mathbf{x} + (\mathbf{I} - \mathbf{Q})(\mathbf{c}_2 - \mathbf{c}_1), \end{aligned}$$

using $\mathbf{Q}^2 = \mathbf{I}$. Since $\mathbf{I} - \mathbf{Q}$ kills $\operatorname{span}(\mathbf{u})$ and doubles $\operatorname{span}(\mathbf{u})^\perp$, the vector $(\mathbf{I} - \mathbf{Q})(\mathbf{c}_2 - \mathbf{c}_1)$ is twice the component of $\mathbf{c}_2 - \mathbf{c}_1$ perpendicular to the lines. In (b), $\mathbf{c}_2 - \mathbf{c}_1 = (0, 1)$ and the translation was by $(0, 2)$.



Two mirrors meeting at angle α compose to the rotation through 2α about their common point

22.4.4 Motions of space

In $\mathbb{R}^3$ the same method works, with Corollary 12.6.2 supplying the normal form of the linear part. Two new names are needed.

Definition 22.4.4. Screw Motion, Rotatory Reflection

Let $L = \mathbf{p} + \operatorname{span}(\mathbf{u})$ be a line in $\mathbb{R}^3$ with $\|\mathbf{u}\| = 1$, and let $\Pi = \mathbf{p} + W$ be a plane.

(a) A **rotation about the axis** L is a motion $\mathbf{x} \mapsto \mathbf{Q}(\mathbf{x} - \mathbf{p}) + \mathbf{p}$ with $\mathbf{Q} \in \mathrm{SO}(3)$, $\mathbf{Q} \neq \mathbf{I}$ and $\mathbf{Q}\mathbf{u} = \mathbf{u}$. It fixes L pointwise, and by Corollary 12.6.2 it turns the plane $\operatorname{span}(\mathbf{u})^{\perp}$ by a single angle $\theta \in (0, \pi]$, the **angle** of the rotation. The requirement $\mathbf{Q} \neq \mathbf{I}$ keeps the identity out: it is a translation, and a translation should not also count as a rotation about every line at once.

(b) A **screw motion** is $t_{\mathbf{v}} \circ r$, where r is a rotation about L — so its angle is again nonzero — and $\mathbf{v} \neq \mathbf{0}$ is parallel to L.

(c) The **reflection in the plane** Π is $\mathbf{x} \mapsto \mathbf{Q}_{\Pi}(\mathbf{x} - \mathbf{p}) + \mathbf{p}$, with $\mathbf{Q}_{\Pi}$ the linear map fixing W pointwise and negating $W^{\perp}$; a **glide reflection** is $t_{\mathbf{v}} \circ s_{\Pi}$ with $\mathbf{v} \neq \mathbf{0}$ lying in W.

(d) A **rotatory reflection about the point** $\mathbf{p}$ is the composite of a rotation about an axis $L = \mathbf{p} + \operatorname{span}(\mathbf{u})$ by an angle $\theta \in (0, \pi]$ with the reflection in the plane $\mathbf{p} + \operatorname{span}(\mathbf{u})^{\perp}$ perpendicular to that axis. The two commute, and the composite is $\mathbf{x} \mapsto \mathbf{Q}(\mathbf{x} - \mathbf{p}) + \mathbf{p}$ where $\mathbf{Q}$ is $[-1] \oplus \mathbf{R}_{\theta}$ in an orthonormal basis beginning with $\mathbf{u}$; in particular $\det \mathbf{Q} = -1$. The case $\theta = \pi$ gives $\mathbf{Q} = -\mathbf{I}_3$ and the **point reflection** $\mathbf{x} \mapsto 2\mathbf{p} - \mathbf{x}$.

Theorem 22.4.5. Classification of the Motions of Space

Let $f(\mathbf{x}) = \mathbf{Q}\mathbf{x} + \mathbf{b}$ be a motion of $\mathbb{R}^3$.

(a) If f is direct, then f is a translation (possibly the identity), a rotation about an axis, or a screw motion, according as $\mathbf{Q} = \mathbf{I}$, or $\mathbf{Q} \neq \mathbf{I}$ and $\mathbf{b}$ is orthogonal to the line fixed by $\mathbf{Q}$, or $\mathbf{Q} \neq \mathbf{I}$ and it is not.

(b) If f is opposite, then f is a reflection in a plane, a glide reflection, or a rotatory reflection about a point.

In both parts the three cases are mutually exclusive, and are told apart by $\operatorname{Fix}(f)$ together with whether $\mathbf{Q} = \mathbf{I}$:

$\det \mathbf{Q} = 1$	translation	rotation	screw
$\operatorname{Fix}(f)$	$\mathbb{R}^3$ or $\emptyset$	a line	$\emptyset$
$\mathbf{Q}$	$\mathbf{I}$	$\neq \mathbf{I}$	$\neq \mathbf{I}$

$\det \mathbf{Q} = -1$	reflection	glide	rotatory
$\operatorname{Fix}(f)$	a plane	$\emptyset$	a point

Idea. Corollary 12.6.2 gives an orthonormal basis $(\mathbf{u}, \mathbf{w}_1, \mathbf{w}_2)$ in which $\mathbf{Q}$ is $[1] \oplus \mathbf{R}_\theta$ or $[-1] \oplus \mathbf{R}_\theta$ with $\theta \in [0, \pi]$. In every case, split $\mathbf{b}$ into its component along $\mathbf{u}$ and its component in $W = \operatorname{span}(\mathbf{u})^\perp$, and ask whether $\mathbf{I} - \mathbf{Q}$ can absorb each piece.

For a direct motion with $\theta \neq 0$, $\mathbf{I} - \mathbf{Q}$ kills $\mathbf{u}$ but is invertible on W; so the W-part of $\mathbf{b}$ is absorbed by moving the axis, and the $\mathbf{u}$-part survives as the glide along the axis. That is Chasles's screw.

For an opposite motion the roles swap. When $\theta = 0$, $\mathbf{Q}$ is a plane reflection: $\mathbf{I} - \mathbf{Q}$ is invertible on $\operatorname{span}(\mathbf{u})$ and kills W, so the $\mathbf{u}$-part of $\mathbf{b}$ is absorbed by sliding the mirror plane and the W-part survives as the glide. When $\theta \neq 0$, $\mathbf{I} - \mathbf{Q}$ is invertible outright, so **all** of $\mathbf{b}$ is absorbed and f has a fixed point.

Proof. By Corollary 12.6.2 there are an orthonormal basis $(\mathbf{u}, \mathbf{w}_1, \mathbf{w}_2)$ of $\mathbb{R}^3$ and $\theta \in [0, \pi]$ such that, in that basis, $\mathbf{Q}$ is $[1] \oplus \mathbf{R}_\theta$ when $\det \mathbf{Q} = 1$ and $[-1] \oplus \mathbf{R}_\theta$ when $\det \mathbf{Q} = -1$. Put $W = \operatorname{span}(\mathbf{u})^\perp = \operatorname{span}(\mathbf{w}_1, \mathbf{w}_2)$, a $\mathbf{Q}$-invariant subspace on which $\mathbf{Q}$ acts as $\mathbf{R}_\theta$ in the basis $(\mathbf{w}_1, \mathbf{w}_2)$, and write, by Theorem 11.3.2,

$$\mathbf{b} = \mathbf{b}_\parallel + \mathbf{b}_\perp, \qquad \mathbf{b}_\parallel = \langle \mathbf{b}, \mathbf{u} \rangle \mathbf{u}, \quad \mathbf{b}_\perp \in W.$$

As in the proof of Theorem 22.4.4, $\det(\mathbf{I}_2 - \mathbf{R}_\theta) = 2 - 2\cos\theta$, which is nonzero exactly when $\mathbf{R}_\theta \neq \mathbf{I}_2$, that is, when $\theta \neq 0$.

(a) Direct motions. Here $\mathbf{Q}\mathbf{u} = \mathbf{u}$.

If $\theta = 0$, then $\mathbf{Q} = \mathbf{I}$ and $f = t_{\mathbf{b}}$, the identity when $\mathbf{b} = \mathbf{0}$. Its fixed set is $\mathbb{R}^3$ or $\emptyset$.

If $\theta \neq 0$, then $\mathbf{Q} \neq \mathbf{I}$ and the restriction of $\mathbf{I} - \mathbf{Q}$ to W has matrix $\mathbf{I}_2 - \mathbf{R}_\theta$, which is invertible. Since $\mathbf{b}_\perp \in W$, there is a unique $\mathbf{c} \in W$ with $(\mathbf{I} - \mathbf{Q})\mathbf{c} = \mathbf{b}_\perp$. Put $L = \mathbf{c} + \operatorname{span}(\mathbf{u})$ and $r(\mathbf{x}) = \mathbf{Q}(\mathbf{x} - \mathbf{c}) + \mathbf{c} = \mathbf{Q}\mathbf{x} + (\mathbf{I} - \mathbf{Q})\mathbf{c} = \mathbf{Q}\mathbf{x} + \mathbf{b}_\perp$. Then r fixes L pointwise, because $r(\mathbf{c} + s\mathbf{u}) = \mathbf{Q}(s\mathbf{u}) + \mathbf{c} = s\mathbf{u} + \mathbf{c}$, so r is the rotation about the axis L. And

$$f(\mathbf{x}) = \mathbf{Q}\mathbf{x} + \mathbf{b}_\perp + \mathbf{b}_\parallel = (t_{\mathbf{b}_\parallel} \circ r)(\mathbf{x}),$$

with $\mathbf{b}_\parallel$ parallel to L. So $f = r$ is a rotation about L when $\mathbf{b}_\parallel = \mathbf{0}$, and a screw motion otherwise.

For the fixed sets, note that $(\mathbf{I} - \mathbf{Q})\mathbf{x} \in W$ for every $\mathbf{x}$, because $(\mathbf{I} - \mathbf{Q})\mathbf{u} = \mathbf{0}$ and $\mathbf{I} - \mathbf{Q}$ maps W into W. Hence the $\mathbf{u}$-component of $f(\mathbf{x}) - \mathbf{x} = \mathbf{b} - (\mathbf{I} - \mathbf{Q})\mathbf{x}$ is $\mathbf{b}_\parallel$. If $\mathbf{b}_\parallel \neq \mathbf{0}$ there is no fixed point; if $\mathbf{b}_\parallel = \mathbf{0}$ the fixed points are the $\mathbf{x}$ with $(\mathbf{I} - \mathbf{Q})\mathbf{x} = \mathbf{b}_\perp$, whose W-component is the unique $\mathbf{c}$ above and whose $\mathbf{u}$-component is free, that is, the line L.

(b) Opposite motions. Here $\mathbf{Qu} = -\mathbf{u}$.

If $\theta = 0$, then $\mathbf{Q}$ fixes W pointwise and negates $\operatorname{span}(\mathbf{u})$: it is the linear reflection in the plane W. Put $\mathbf{c} = \frac{1}{2}\mathbf{b}_{\parallel} \in \operatorname{span}(\mathbf{u})$ and $\Pi = \mathbf{c} + W$. Since $\mathbf{Qc} = -\mathbf{c}$ we get $(\mathbf{I} - \mathbf{Q})\mathbf{c} = 2\mathbf{c} = \mathbf{b}_{\parallel}$, so the reflection in Π is

$$s_{\Pi}(\mathbf{x}) = \mathbf{Q}(\mathbf{x} - \mathbf{c}) + \mathbf{c} = \mathbf{Qx} + \mathbf{b}_{\parallel}, \qquad f(\mathbf{x}) = (t_{\mathbf{b}_{\perp}} \circ s_{\Pi})(\mathbf{x}),$$

with $\mathbf{b}_{\perp} \in W$ parallel to Π. So f is the reflection in Π when $\mathbf{b}_{\perp} = \mathbf{0}$, and a glide reflection otherwise. For the fixed sets: $(\mathbf{I} - \mathbf{Q})\mathbf{x} \in \operatorname{span}(\mathbf{u})$ for every $\mathbf{x}$, since $\mathbf{I} - \mathbf{Q}$ kills W, so the W-component of $f(\mathbf{x}) - \mathbf{x}$ is $\mathbf{b}_{\perp}$. If $\mathbf{b}_{\perp} \neq \mathbf{0}$ there is no fixed point; if $\mathbf{b}_{\perp} = \mathbf{0}$ the fixed points are the $\mathbf{x}$ with $(\mathbf{I} - \mathbf{Q})\mathbf{x} = \mathbf{b}_{\parallel}$, that is, with $\mathbf{u}$-component $\mathbf{c}$, which is the plane Π.

If $\theta \neq 0$, then in the basis $(\mathbf{u}, \mathbf{w}_1, \mathbf{w}_2)$ the matrix $\mathbf{I}_3 - \mathbf{Q}$ is $[2] \oplus (\mathbf{I}_2 - \mathbf{R}_{\theta})$, whose determinant is $2(2 - 2\cos\theta) \neq 0$. So $\mathbf{I} - \mathbf{Q}$ is invertible and there is a unique $\mathbf{p}$ with $(\mathbf{I} - \mathbf{Q})\mathbf{p} = \mathbf{b}$, that is, $f(\mathbf{p}) = \mathbf{p}$. Then $f(\mathbf{x}) - \mathbf{p} = \mathbf{Q}(\mathbf{x} - \mathbf{p})$ for every $\mathbf{x}$, as in the plane case, and f is a rotatory reflection about $\mathbf{p}$: in the basis $(\mathbf{u}, \mathbf{w}_1, \mathbf{w}_2)$ its linear part is $[-1] \oplus \mathbf{R}_{\theta}$, the rotation by θ about $\operatorname{span}(\mathbf{u})$ followed by the reflection in W, and these two commute. The fixed set is $\{\mathbf{p}\}$, since $\mathbf{I} - \mathbf{Q}$ is injective. When $\theta = \pi$, $\mathbf{R}_{\pi} = -\mathbf{I}_2$ and $\mathbf{Q} = -\mathbf{I}_3$, giving the point reflection $\mathbf{x} \mapsto 2\mathbf{p} - \mathbf{x}$.

Finally, exactly one case holds in each part. Among the direct motions a translation has $\mathbf{Q} = \mathbf{I}$ while a rotation and a screw do not, and those two differ by having a line of fixed points or none; among the opposite motions the three fixed sets, a plane, the empty set, and a single point, are pairwise different. This proves the theorem.

■

Remark. Part (a) is **Chasles's theorem**: every direct motion of space is a screw, provided the two degenerate cases barred by Definition 22.4.4 are allowed back in — angle $\theta = 0$, which is a pure translation, and $\mathbf{v} = \mathbf{0}$, which is a pure rotation. The corresponding statement in the plane, Theorem 22.4.4, has no screw because a rotation of $\mathbb{R}^2$ leaves no axis to glide along.

Warning. **From $\mathbb{R}^4$ on, the list breaks down.** The proof rested entirely on Corollary 12.6.2, which in turn rested on the arithmetic $p + q + 2m = 3$ of Theorem 12.6.1. In $\mathbb{R}^4$ the matrix $\mathbf{R}_{\pi/2} \oplus \mathbf{R}_{\pi/3}$ has determinant 1 and fixes no nonzero vector, so a direct motion of $\mathbb{R}^4$ need not have an axis and need not be a screw. The right general statement replaces "one axis and one angle" by the full canonical form: a **direct** motion of $\mathbb{R}^n$ decomposes the space into a flat, along which it translates, and a list of mutually orthogonal planes, each turned by its own angle. The word **direct** is needed there: that description hides the $-\mathbf{I}_q$ block of Theorem 12.6.1, which splits into rotations by π only when q is even, and an opposite motion keeps one leftover direction and negates it. The glide reflection of Theorem 22.4.5 (b) is the smallest witness, with $p = 2$, $q = 1$, $m = 0$: it has no rotation plane, yet it is not a pure translation. This book proves the classification only for $n \leq 3$.

Example. Naming a space motion

Classify the motion $f(\mathbf{x}) = \mathbf{Q}\mathbf{x} + \mathbf{b}$ of $\mathbb{R}^3$ with

$$\mathbf{Q} = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}, \qquad \mathbf{b} = (4, 1, -2).$$

Solution. The linear part. The columns of $\mathbf{Q}$ are $\mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_1$, an orthonormal list, so $\mathbf{Q} \in \mathrm{O}(3)$ by Theorem 11.7.3 (f). Expanding along the first row, $\det \mathbf{Q} = 0 - 0 + 1 \cdot (1 \cdot 1 - 0 \cdot 0) = 1$, so f is direct. By Theorem 12.6.4 (c), $\mathbf{Q} = \mathbf{R}_{\mathbf{n},\theta}$ for some unit $\mathbf{n}$ and $\theta \in [0, \pi]$. Its trace is 0, so by Theorem 12.6.5 (a) the angle is given by $\cos\theta = \frac{1}{2}(0 - 1) = -\frac{1}{2}$, that is $\theta = 2\pi/3$, and in particular $\mathbf{Q} \neq \mathbf{I}$. The axis direction is fixed by $\mathbf{Q}$: reading $\mathbf{Q}\mathbf{x} = (x_3, x_1, x_2)$, the equation $\mathbf{Q}\mathbf{x} = \mathbf{x}$ says $x_1 = x_2 = x_3$, so the axis direction is $\mathbf{u}' = (1, 1, 1)$.

Splitting $\mathbf{b}$. With $\|\mathbf{u}'\|^2 = 3$ and $\langle \mathbf{b}, \mathbf{u}' \rangle = 4 + 1 - 2 = 3$,

$$\mathbf{b}_{\parallel} = \tfrac{3}{3}(1, 1, 1) = (1, 1, 1), \qquad \mathbf{b}_{\perp} = \mathbf{b} - \mathbf{b}_{\parallel} = (3, 0, -3).$$

Since $\mathbf{b}_{\parallel} \neq \mathbf{0}$, Theorem 22.4.5 (a) says f is a **screw motion** with glide vector $(1, 1, 1)$.

The axis. We need $\mathbf{c}$ perpendicular to $\mathbf{u}'$ with $(\mathbf{I} - \mathbf{Q})\mathbf{c} = \mathbf{b}_{\perp}$. Writing $(\mathbf{I} - \mathbf{Q})\mathbf{c} = (c_1 - c_3,\ c_2 - c_1,\ c_3 - c_2)$, the equations are $c_1 - c_3 = 3$, $c_2 - c_1 = 0$, $c_3 - c_2 = -3$, together with $c_1 + c_2 + c_3 = 0$. From the second, $c_2 = c_1$; from the first, $c_3 = c_1 - 3$; the sum gives $3c_1 - 3 = 0$, so $\mathbf{c} = (1, 1, -2)$. The axis is $L = (1, 1, -2) + \mathrm{span}((1, 1, 1))$.

Checks. The rotation part $r(\mathbf{x}) = \mathbf{Q}\mathbf{x} + (3, 0, -3)$ should fix $\mathbf{c}$: $\mathbf{Q}(1, 1, -2) = (-2, 1, 1)$, and $(-2, 1, 1) + (3, 0, -3) = (1, 1, -2)$. And f applied three times should be the translation by $3\mathbf{b}_{\parallel} = (3, 3, 3)$, since $\mathbf{Q}^3 = \mathbf{I}$:

$$f^3(\mathbf{x}) = \mathbf{Q}^3\mathbf{x} + \mathbf{Q}^2\mathbf{b} + \mathbf{Q}\mathbf{b} + \mathbf{b} = \mathbf{x} + (\mathbf{I} + \mathbf{Q} + \mathbf{Q}^2)\mathbf{b}.$$

Here $\mathbf{I} + \mathbf{Q} + \mathbf{Q}^2$ is the all-ones matrix $\mathbf{J}$, so $(\mathbf{I} + \mathbf{Q} + \mathbf{Q}^2)\mathbf{b} = (4 + 1 - 2)\mathbf{1} = (3, 3, 3)$, as predicted.

22.4.5 Exercises

22.4.5.1 A. Check your understanding

Exercise. A1

(a) State Definition 22.4.1, and say which of linearity, continuity, and surjectivity it assumes.

(b) True or false: every motion of $\mathbb{R}^n$ with a fixed point is linear. Justify your answer.

(c) Name the four kinds of motion of $\mathbb{R}^2$ and give the fixed-point set of each.

(d) True or false: the composite of two opposite motions of $\mathbb{R}^3$ is opposite. Justify your answer.

(e) A motion of $\mathbb{R}^3$ has exactly one fixed point. What kind is it?

Solution.

(a) A motion is a map $f\colon \mathbb{R}^n \to \mathbb{R}^n$ with $\|f(\mathbf{x}) - f(\mathbf{y})\| = \|\mathbf{x} - \mathbf{y}\|$ for all $\mathbf{x}, \mathbf{y}$. It assumes none of the three: Theorem 22.4.1 shows that a motion is automatically affine, and Corollary 22.4.2 that it is automatically a bijection.

(b) False as stated, but only just: by Theorem 22.4.1, $f(\mathbf{x}) = \mathbf{Q}\mathbf{x} + \mathbf{b}$, and $f(\mathbf{p}) = \mathbf{p}$ gives $\mathbf{b} = \mathbf{p} - \mathbf{Q}\mathbf{p}$, so $f(\mathbf{x}) = \mathbf{Q}(\mathbf{x} - \mathbf{p}) + \mathbf{p}$, which is linear precisely when $\mathbf{p}$ can be taken to be $\mathbf{0}$. For example the half-turn of $\mathbb{R}^2$ about $(1, 0)$, $\mathbf{x} \mapsto -\mathbf{x} + (2, 0)$, has a fixed point but sends $\mathbf{0}$ to $(2, 0) \neq \mathbf{0}$, so it is not linear. What is true is that a motion fixing $\mathbf{0}$ is linear, which is Steps 1 to 3 of Theorem 22.4.1.

(c) Translations (fixed set $\mathbb{R}^2$ if the identity, otherwise $\emptyset$); rotations about a point (one point); reflections in a line (that line); glide reflections ($\emptyset$). This is Theorem 22.4.4.

(d) False. By Proposition 22.4.3 (b) the linear part of a composite is the product of the linear parts, and det is multiplicative (Theorem 7.5.4), so the composite has det $= (-1)(-1) = 1$ and is **direct**. What composes is the parity, not the word "opposite": Example (Two mirrors) is the two-dimensional instance, where two reflections make a rotation.

(e) A rotatory reflection about that point, by Theorem 22.4.5: it is the only kind whose fixed set is a single point. In particular it is opposite.

22.4.5.2 B. Practice

Exercise. B1: Classify three plane motions

Classify each motion $f(\mathbf{x}) = \mathbf{Q}\mathbf{x} + \mathbf{b}$ of $\mathbb{R}^2$, giving the center and angle, the axis, or the axis and glide vector, as appropriate.

(a) $\mathbf{Q} = \left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$, $\mathbf{b} = (2, 4)$.

(b) $\mathbf{Q} = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$, $\mathbf{b} = (3, 3)$.

(c) $\mathbf{Q} = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$, $\mathbf{b} = (1, -1)$.

Solution.

(a) $\det \mathbf{Q} = 1$ and $\mathbf{Q} = \mathbf{R}_{\pi/2}$, so f is a rotation by $\pi/2$. Its center solves $(\mathbf{I} - \mathbf{Q})\mathbf{p} = \mathbf{b}$, that is $p_1 + p_2 = 2$ and $-p_1 + p_2 = 4$; adding gives $2p_2 = 6$, so $\mathbf{p} = (-1, 3)$. Check: $f(-1, 3) = (-3, -1) + (2, 4) = (-1, 3)$. So f is the rotation about $(-1, 3)$ by $\pi/2$.

(b) $\det \mathbf{Q} = -1$, and $\mathbf{Q}$ swaps the coordinates: it is the reflection in $\operatorname{span}((1, 1))$, negating $\operatorname{span}((1, -1))$. With $\mathbf{u}' = (1, 1)$, $\langle \mathbf{b}, \mathbf{u}' \rangle = 6$ and $\|\mathbf{u}'\|^2 = 2$, so $\mathbf{b}_{\|} = 3(1, 1) = (3, 3) = \mathbf{b}$ and $\mathbf{b}_{\perp} = \mathbf{0}$. Since $\mathbf{b}_{\|} \neq \mathbf{0}$, f is a glide reflection with axis $\mathbf{c} + \operatorname{span}((1, 1))$, where $\mathbf{c} = \frac{1}{2}\mathbf{b}_{\perp} = \mathbf{0}$: the axis is the line $x_2 = x_1$ and the glide vector is $(3, 3)$. Check: $f(0, 0) = (3, 3)$, which is on the axis.

(c) Same $\mathbf{Q}$. Now $\langle \mathbf{b}, \mathbf{u}' \rangle = 0$, so $\mathbf{b}_{\|} = \mathbf{0}$ and $\mathbf{b}_{\perp} = (1, -1)$. Hence f is a **reflection**, in the line $L = \mathbf{c} + \operatorname{span}((1, 1))$ with $\mathbf{c} = \frac{1}{2}(1, -1)$. Check: $f(\frac{1}{2}, -\frac{1}{2}) = (-\frac{1}{2}, \frac{1}{2}) + (1, -1) = (\frac{1}{2}, -\frac{1}{2})$, a fixed point, and $f(\frac{3}{2}, \frac{1}{2}) = (\frac{1}{2}, \frac{3}{2}) + (1, -1) = (\frac{3}{2}, \frac{1}{2})$, another one; the two lie on L.

Exercise. B2: A space motion

Classify the motion $f(\mathbf{x}) = \mathbf{Q}\mathbf{x} + \mathbf{b}$ of $\mathbb{R}^3$ with

$$\mathbf{Q} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix}, \qquad \mathbf{b} = (0, 2, 0),$$

giving the axis and the glide vector if there is one.

Solution. The columns are $\mathbf{e}_1, \mathbf{e}_3, -\mathbf{e}_2$, orthonormal, so $\mathbf{Q} \in \mathrm{O}(3)$; expanding along the first row, $\det \mathbf{Q} = 1 \cdot (0 + 1) = 1$, so f is direct. The fixed vectors of $\mathbf{Q}$ satisfy x_1 free, $-x_3 = x_2$, $x_2 = x_3$, hence $x_2 = x_3 = 0$: the axis direction is $\mathbf{u} = \mathbf{e}_1$, and $\mathbf{Q}$ is the quarter turn about it.

Now $\mathbf{b}_{\parallel} = \langle \mathbf{b}, \mathbf{e}_1 \rangle \mathbf{e}_1 = \mathbf{0}$ and $\mathbf{b}_{\perp} = (0, 2, 0)$. Since $\mathbf{b}_{\parallel} = \mathbf{0}$, Theorem 22.4.5 (a) says f is a **pure rotation** about an axis, with no glide. The axis is $\mathbf{c} + \mathrm{span}(\mathbf{e}_1)$ where $\mathbf{c} \in \mathrm{span}(\mathbf{e}_2, \mathbf{e}_3)$ solves $(\mathbf{I} - \mathbf{Q})\mathbf{c} = \mathbf{b}_{\perp}$:

$$(\mathbf{I} - \mathbf{Q})(0, c_2, c_3) = (0,\ c_2 + c_3,\ -c_2 + c_3) = (0, 2, 0)$$

gives $c_2 + c_3 = 2$ and $c_3 = c_2$, so $\mathbf{c} = (0, 1, 1)$. The axis is $\{(t, 1, 1) : t \in \mathbb{R}\}$. Check: $f(0, 1, 1) = (0, -1, 1) + (0, 2, 0) = (0, 1, 1)$, and $f(5, 1, 1) = (5, -1, 1) + (0, 2, 0) = (5, 1, 1)$.

Exercise. B3: A motion is determined by few points

Let f and g be motions of $\mathbb{R}^n$ that agree on $\mathbf{0}, \mathbf{e}_1, \dots, \mathbf{e}_n$. Prove that $f = g$.

Hint: consider $h = g^{-1} \circ f$ and use Theorem 22.4.1.

Solution. By Corollary 22.4.2, g^{-1} is a motion, so $h = g^{-1} \circ f$ is a motion (example 3 after Definition 22.4.1). It fixes $\mathbf{0}$ and each $\mathbf{e}_i$, since f and g agree there. By Theorem 22.4.1, $h(\mathbf{x}) = \mathbf{Q}\mathbf{x} + \mathbf{b}$ with $\mathbf{Q} \in \mathrm{O}(n)$; from $h(\mathbf{0}) = \mathbf{0}$ we get $\mathbf{b} = \mathbf{0}$, and then $\mathbf{Q}\mathbf{e}_i = h(\mathbf{e}_i) = \mathbf{e}_i$ for each i, so $\mathbf{Q}$ agrees with $\mathbf{I}$ on a basis and $\mathbf{Q} = \mathbf{I}$. Hence $h = \mathrm{id}$, and composing with g gives $f = g$.

(The $n + 1$ points $\mathbf{0}, \mathbf{e}_1, \dots, \mathbf{e}_n$ are affinely independent, and the same argument works for any affinely independent list of $n + 1$ points, which is the affine statement of §02.)

22.4.5.3 C. Going deeper

Exercise. C1: Reflections generate the motions

(a) Prove that every motion of $\mathbb{R}^n$ is a composite of at most $n + 1$ reflections in affine hyperplanes, where the reflection in the hyperplane $\mathbf{p} + \mathrm{span}(\mathbf{w})^{\perp}$ is $\mathbf{x} \mapsto \mathbf{H}_{\mathbf{w}}(\mathbf{x} - \mathbf{p}) + \mathbf{p}$ with $\mathbf{H}_{\mathbf{w}}$ the Householder reflection of Definition 11.8.1.

(b) Deduce that a translation of $\mathbb{R}^n$ by $\mathbf{b} \neq \mathbf{0}$ is a composite of exactly two such reflections, and identify them.

Hint: for (a), first make the motion fix a point, then quote Theorem 12.6.7.

Solution.

(a) Write $f(\mathbf{x}) = \mathbf{Q}\mathbf{x} + \mathbf{b}$ by Theorem 22.4.1. If $\mathbf{b} = \mathbf{0}$, then $\mathbf{Q}$ is a product of at most n Householder reflections by Theorem 12.6.7, and each of those is the reflection in the hyperplane $\mathbf{0} + \operatorname{span}(\mathbf{w})^\perp$ in the above sense, so f is a composite of at most $n \le n+1$ reflections.
If $\mathbf{b} \neq \mathbf{0}$, let σ be the reflection in the hyperplane $\frac{1}{2}\mathbf{b} + \operatorname{span}(\mathbf{b})^\perp$, that is, $\sigma(\mathbf{x}) = \mathbf{H}_\mathbf{b}(\mathbf{x} - \frac{1}{2}\mathbf{b}) + \frac{1}{2}\mathbf{b}$. Since $\mathbf{H}_\mathbf{b}\mathbf{b} = -\mathbf{b}$ (Proposition 11.8.2 (d)), this simplifies to $\sigma(\mathbf{x}) = \mathbf{H}_\mathbf{b}\mathbf{x} + \mathbf{b}$, and then, using $\mathbf{H}_\mathbf{b}^2 = \mathbf{I}$ (Proposition 11.8.2 (b)),

$$\sigma(\sigma(\mathbf{x})) = \mathbf{H}_\mathbf{b}(\mathbf{H}_\mathbf{b}\mathbf{x} + \mathbf{b}) + \mathbf{b} = \mathbf{x} - \mathbf{b} + \mathbf{b} = \mathbf{x},$$

so $\sigma^{-1} = \sigma$. Therefore

$$(\sigma \circ f)(\mathbf{x}) = \mathbf{H}_\mathbf{b}(\mathbf{Q}\mathbf{x} + \mathbf{b}) + \mathbf{b} = \mathbf{H}_\mathbf{b}\mathbf{Q}\mathbf{x} - \mathbf{b} + \mathbf{b} = \mathbf{H}_\mathbf{b}\mathbf{Q}\mathbf{x},$$

a linear orthogonal map, which by Theorem 12.6.7 is a product of at most n Householder reflections. Hence $f = \sigma \circ (\sigma \circ f)$ is a composite of at most $n+1$ reflections.

(b) Let $f = t_\mathbf{b}$ with $\mathbf{b} \neq \mathbf{0}$. It is not a single reflection, since a reflection has a fixed point and $t_\mathbf{b}$ has none. Let σ_0 be the reflection in $\operatorname{span}(\mathbf{b})^\perp$ and σ_1 the reflection in $\frac{1}{2}\mathbf{b} + \operatorname{span}(\mathbf{b})^\perp$. By the computation in (a), $\sigma_0(\mathbf{x}) = \mathbf{H}_\mathbf{b}\mathbf{x}$ and $\sigma_1(\mathbf{x}) = \mathbf{H}_\mathbf{b}\mathbf{x} + \mathbf{b}$, so

$$(\sigma_1 \circ \sigma_0)(\mathbf{x}) = \mathbf{H}_\mathbf{b}(\mathbf{H}_\mathbf{b}\mathbf{x}) + \mathbf{b} = \mathbf{x} + \mathbf{b} = t_\mathbf{b}(\mathbf{x}).$$

So exactly two reflections suffice, in two parallel hyperplanes perpendicular to $\mathbf{b}$ and at distance $\frac{1}{2}\|\mathbf{b}\|$ apart — the n-dimensional form of Example (Two mirrors) (b).

Exercise. C2: Matching two congruent frames

Let $\mathbf{p}_0, \dots, \mathbf{p}_n$ and $\mathbf{q}_0, \dots, \mathbf{q}_n$ be points of $\mathbb{R}^n$ with $\|\mathbf{p}_i - \mathbf{p}_j\| = \|\mathbf{q}_i - \mathbf{q}_j\|$ for all i, j, and suppose $\mathbf{p}_1 - \mathbf{p}_0, \dots, \mathbf{p}_n - \mathbf{p}_0$ is a basis of $\mathbb{R}^n$.

(a) Prove that there is a motion f of $\mathbb{R}^n$ with $f(\mathbf{p}_i) = \mathbf{q}_i$ for every i.

(b) Prove that f is unique, and give an example with fewer points where it is not.

Hint: for (a), translate both lists so that $\mathbf{p}_0 = \mathbf{q}_0 = \mathbf{0}$, then compare the inner products of the two lists.

Solution.

(a) Replacing $\mathbf{p}_i$ by $\mathbf{p}_i - \mathbf{p}_0$ and $\mathbf{q}_i$ by $\mathbf{q}_i - \mathbf{q}_0$ changes no distance, and undoing the two translations at the end changes a motion into a motion; so assume $\mathbf{p}_0 = \mathbf{q}_0 = \mathbf{0}$. Put $\mathbf{v}_i = \mathbf{p}_i$ and $\mathbf{w}_i = \mathbf{q}_i$ for $i \ge 1$. Taking $j = 0$ in the hypothesis gives $\|\mathbf{v}_i\| = \|\mathbf{w}_i\|$, and then for $i, j \ge 1$, by Theorem 11.1.2 (c),

$$\begin{aligned} 2\langle \mathbf{v}_i, \mathbf{v}_j \rangle &= \|\mathbf{v}_i\|^2 + \|\mathbf{v}_j\|^2 - \|\mathbf{v}_i - \mathbf{v}_j\|^2 \\ &= \|\mathbf{w}_i\|^2 + \|\mathbf{w}_j\|^2 - \|\mathbf{w}_i - \mathbf{w}_j\|^2 = 2\langle \mathbf{w}_i, \mathbf{w}_j \rangle. \end{aligned}$$

So the two lists have the same inner products. Since $(\mathbf{v}_1, \dots, \mathbf{v}_n)$ is a basis, Theorem 2.1.5 gives a unique linear T with $T\mathbf{v}_i = \mathbf{w}_i$, and for $\mathbf{x} = \sum_i x_i \mathbf{v}_i$,

$$\|T\mathbf{x}\|^2 = \sum_{i,j} x_i x_j \langle \mathbf{w}_i, \mathbf{w}_j \rangle = \sum_{i,j} x_i x_j \langle \mathbf{v}_i, \mathbf{v}_j \rangle = \|\mathbf{x}\|^2,$$

where both ends use Theorem 11.1.2 (c) repeatedly to expand a squared norm of a sum. So T is an isometry, and $f(\mathbf{x}) = T\mathbf{x}$ is a motion with $f(\mathbf{p}_i) = \mathbf{q}_i$ for $i \geq 1$ and $f(\mathbf{0}) = \mathbf{0}$.

(b) If f and g are motions with $f(\mathbf{p}_i) = g(\mathbf{p}_i) = \mathbf{q}_i$ for all i, then $h = g^{-1} \circ f$ is a motion (Corollary 22.4.2) fixing every $\mathbf{p}_i$. Write $h(\mathbf{x}) = \mathbf{Q}\mathbf{x} + \mathbf{b}$ by Theorem 22.4.1. Subtracting $h(\mathbf{p}_0) = \mathbf{p}_0$ from $h(\mathbf{p}_i) = \mathbf{p}_i$ gives $\mathbf{Q}(\mathbf{p}_i - \mathbf{p}_0) = \mathbf{p}_i - \mathbf{p}_0$ for $i \geq 1$, so $\mathbf{Q}$ fixes a basis and $\mathbf{Q} = \mathbf{I}$; then $\mathbf{b} = \mathbf{p}_0 - \mathbf{p}_0 = \mathbf{0}$ and $h = \mathrm{id}$, so $f = g$.

With fewer points the spanning hypothesis fails and uniqueness goes with it. In $\mathbb{R}^2$, take the single pair $\mathbf{p}_0 = \mathbf{q}_0 = (0, 0)$ and $\mathbf{p}_1 = \mathbf{q}_1 = (1, 0)$. Both the identity and the reflection in the first coordinate axis are motions carrying $\mathbf{p}_i$ to $\mathbf{q}_i$, and they differ at $(0, 1)$. Here $\mathbf{p}_1 - \mathbf{p}_0 = \mathbf{e}_1$ spans only a line.

Exercise. C3: Motions of a finite set

Let $S \subseteq \mathbb{R}^n$ be a finite nonempty set and let $G = \{f \in \mathrm{E}(n) : f(S) = S\}$ be its symmetry group.

(a) Prove that G is a subgroup of $\mathrm{E}(n)$, and that every $f \in G$ fixes the **centroid** $\mathbf{c} = \frac{1}{|S|} \sum_{\mathbf{p} \in S} \mathbf{p}$.

(b) Deduce that G is isomorphic to a subgroup of $\mathrm{O}(n)$, and that G contains no translation other than the identity.

Solution.

(a) The identity lies in G. If $f, g \in G$ then $(g \circ f)(S) = g(S) = S$, and $f(S) = S$ with f a bijection (Corollary 22.4.2) gives $f^{-1}(S) = S$; so G is a subgroup by Definition 0.10.2, using Proposition 22.4.3 (a).
Let $f \in G$, with $f(\mathbf{x}) = \mathbf{Q}\mathbf{x} + \mathbf{b}$. Since f restricts to a bijection of the finite set S,

$$\sum_{\mathbf{p} \in S} f(\mathbf{p}) = \sum_{\mathbf{q} \in S} \mathbf{q} = |S|\, \mathbf{c},$$

while on the other hand $\sum_{\mathbf{p} \in S} f(\mathbf{p}) = \mathbf{Q}(\sum_{\mathbf{p} \in S} \mathbf{p}) + |S|\mathbf{b} = |S|(\mathbf{Q}\mathbf{c} + \mathbf{b}) = |S| f(\mathbf{c})$. Dividing by $|S| \neq 0$ gives $f(\mathbf{c}) = \mathbf{c}$.

(b) Conjugating by the translation $t_{-\mathbf{c}}$ gives an injective homomorphism $G \to \mathrm{E}(n)$, $f \mapsto t_{-\mathbf{c}} \circ f \circ t_{\mathbf{c}}$, whose image consists of motions fixing $\mathbf{0}$, hence of linear ones by Theorem 22.4.1; those are the orthogonal maps. So G is isomorphic to a subgroup of $\mathrm{O}(n)$, namely $\{\lambda(f) : f \in G\}$ in the notation of Proposition 22.4.3. If $t_{\mathbf{b}} \in G$ then $t_{\mathbf{b}}(\mathbf{c}) = \mathbf{c}$ by (a), so $\mathbf{b} = \mathbf{0}$. Geometrically: a finite figure cannot be carried onto itself by a nontrivial translation, because a translation would move its centroid.

22.5 Affine and Euclidean Quadrics

Chapter 14 classified quadratic forms up to congruence, and the answer over $\mathbb{R}$ was the inertia. A curve like $5x^2 + 4xy + 5y^2 - 14x - 14y - 7 = 0$ is not a quadratic form, though: it has linear terms and a constant, and its zero set does not pass through the origin. This section asks what the inertia classification becomes once the origin is given up, which is exactly the question affine geometry was built for. Two answers come out, one for affine changes of coordinates and one for the rigid motions of Definition 22.4.1, and the difference between them is the difference between "it is an ellipse" and "its semi-axes are $\sqrt{7}$ and $\sqrt{3}$".

22.5.1 Where characteristic two bites

Everything in this section needs a field in which $2 \neq 0$, and it is worth seeing why before any definition is written down.

A degree-two polynomial in $x_1, \dots, x_n$ has a homogeneous quadratic part, and we want to write that part as $\mathbf{x}^\top \mathbf{A}\mathbf{x}$ with $\mathbf{A}$ **symmetric**, so that Chapter 14 applies to it. For a symmetric $\mathbf{A}$ the coefficient of $x_i x_j$ with $i \neq j$ is $a_{ij} + a_{ji} = 2a_{ij}$. Recovering a_{ij} from the polynomial means halving, which is the division by $1 + 1$ that Theorem 14.3.5 performs, and on which the uniqueness of the symmetric matrix attached to a quadratic form (Definition 14.3.2) depends.

Over $\mathbb{F}_2$ the halving is not merely inconvenient; the target is not there. Every symmetric $\mathbf{A} \in M_n(\mathbb{F}_2)$ gives

$$\mathbf{x}^\top \mathbf{A}\mathbf{x} = \sum_i a_{ii} x_i^2 + 2 \sum_{i<j} a_{ij} x_i x_j = \sum_i a_{ii} x_i^2,$$

and since $a^2 = a$ for $a \in \mathbb{F}_2$ and squaring is additive there, this is $\left(\sum_i a_{ii} x_i\right)^2$, the square of a **linear** form. So over $\mathbb{F}_2$ no expression $\mathbf{x}^\top \mathbf{A}\mathbf{x}$ with $\mathbf{A}$ symmetric ever has an xy term, the polynomial xy is not of that shape at all, and the ones that are have already collapsed to squares of linear forms. There is nothing left to classify.

So: throughout this section F is a field with $\operatorname{char} F \neq 2$ (Definition 0.7.2), and from the affine classification onward $F = \mathbb{R}$. The hypothesis will be invoked by name at each of the three places where it does work: writing the linear part as $2\mathbf{b}^\top \mathbf{x}$, reading off a symmetric $\mathbf{A}$, and solving for a center.

22.5.2 The equation of a quadric

Fix an origin O and a basis of the direction space, which by Proposition 22.1.3 identifies the affine space $\mathcal{A}$ of dimension n with F^n. Every polynomial of degree 2 in the resulting coordinates can then be written in one standard shape.

A quadric is what you get when a quadratic form is allowed to forget where the origin is.

Definition 22.5.1. Quadric

Let $\operatorname{char} F \neq 2$ and let $\mathcal{A}$ be an affine space of dimension n over F, identified with F^n by a choice of origin and basis (Proposition 22.1.3). A **quadric polynomial** on $\mathcal{A}$ is a function

$$Q(\mathbf{x}) = \mathbf{x}^\top \mathbf{A}\mathbf{x} + 2\mathbf{b}^\top \mathbf{x} + c$$

with $\mathbf{A} \in M_n(F)$ **symmetric and nonzero**, $\mathbf{b} \in F^n$ and $c \in F$. Its **quadratic part** is the

quadratic form $q(\mathbf{x}) = \mathbf{x}^\top \mathbf{A}\mathbf{x}$. The **quadric** defined by Q is its zero set

$$\{\, \mathbf{x} \in F^n : Q(\mathbf{x}) = 0 \,\}.$$

For $n = 2$ a quadric is called a **conic**.

Read the clauses. "**Symmetric**" is a normalization, not a restriction: $\mathbf{x}^\top \mathbf{A}\mathbf{x} = \mathbf{x}^\top \mathbf{A}^\top \mathbf{x}$ because a 1×1 matrix equals its transpose, so $\mathbf{A}$ may always be replaced by $\frac{1}{2}(\mathbf{A} + \mathbf{A}^\top)$, and once symmetric it is determined by q (Theorem 14.3.5). "**Nonzero**" excludes $\mathbf{A} = \mathbf{O}$, where Q would be affine and its zero set a flat (Definition 22.2.1) already handled in Section 2; it is the clause that makes the object genuinely quadratic. The factor 2 in $2\mathbf{b}^\top \mathbf{x}$ is a convention we may adopt because 2 is invertible: a linear part $\beta_1 x_1 + \cdots + \beta_n x_n$ is $2\mathbf{b}^\top \mathbf{x}$ with $b_i = \beta_i/2$. It pays for itself in every formula below.

The data are recoverable from the function. Definition 22.5.1 calls Q a function, so before speaking of "its" quadratic part or "its" constant we must check that $\mathbf{A}$, $\mathbf{b}$ and c are determined by Q. They are, over any field of characteristic $\neq 2$. First $Q(\mathbf{0}) = c$. Next

$$Q(\mathbf{x}) - Q(-\mathbf{x}) = 4\mathbf{b}^\top \mathbf{x} \qquad \text{for every } \mathbf{x} \in F^n,$$

and $4 = 2 \cdot 2 \neq 0$, so taking $\mathbf{x} = \mathbf{e}_i$ recovers b_i for each i. Finally $\mathbf{x}^\top \mathbf{A}\mathbf{x} = Q(\mathbf{x}) - 2\mathbf{b}^\top \mathbf{x} - c$ is then determined as a function of $\mathbf{x}$, and a **symmetric** $\mathbf{A}$ is determined by that quadratic form (Theorem 14.3.5). So the triple $(\mathbf{A}, \mathbf{b}, c)$ is unique.

Examples. Over $\mathbb{R}$ with $n = 2$:

- $x^2 + y^2 - 1$, with $\mathbf{A} = \mathbf{I}_2$, $\mathbf{b} = \mathbf{0}$, $c = -1$: the unit circle.
- $x^2 - y$, with $\mathbf{A} = \operatorname{diag}(1, 0)$, $\mathbf{b} = (0, -\frac{1}{2})$, $c = 0$: a parabola. Here the halving convention is visible, and $\mathbf{A}$ is singular.
- $x^2 + y^2$: the zero set is the single point $\mathbf{0}$, although the polynomial has degree 2.
- x^2: the zero set is the line $x = 0$, traced "twice" in the sense that $\mathbf{A} = \operatorname{diag}(1, 0)$ has rank 1 and the polynomial is a square. This is the degenerate case, and it matters because the classification must find room for it.
- $x^2 + y^2 + 1$: the zero set is **empty**. Over $\mathbb{R}$ a quadric may have no points at all, and the classification below has two different normal forms whose zero sets are both empty.

Nonexample by minimal change. Replace $x^2 + y^2 - 1$ by $x^4 + y^2 - 1$: the quadratic part is no longer a quadratic form, and there is no $\mathbf{A}$. Replace it instead by $2x + 3y - 1$: now $\mathbf{A} = \mathbf{O}$, the clause "nonzero" fails, and the zero set is a line, not a quadric.

Warning. **A quadric is classified through its polynomial, not through its point set.** Over $\mathbb{R}$ the polynomials $x^2 + y^2 + 1$ and $x^2 + 1$ both have empty zero set, and no coordinate change can be expected to tell empty from empty; yet the two polynomials will land in different classes below, and rightly so, because their behavior over $\mathbb{C}$ and their intersections with lines differ. Everything in the affine classification below is a statement about the polynomial Q up to an invertible affine substitution and a nonzero scalar factor. When the zero set determines Q up to a scalar it will be said explicitly.

22.5.3 The matrix of a quadric

The whole point of writing the linear part as $2\mathbf{b}^{\top}\mathbf{x}$ is that Q then becomes a single quadratic form in one extra variable. Write

$$\tilde{\mathbf{x}} = \begin{pmatrix} \mathbf{x} \\ 1 \end{pmatrix} \in F^{n+1}, \qquad \widetilde{\mathbf{A}} = \begin{pmatrix} \mathbf{A} & \mathbf{b} \\ \mathbf{b}^{\top} & c \end{pmatrix} \in M_{n+1}(F),$$

so that $\widetilde{\mathbf{A}}$ is symmetric and

$$\tilde{\mathbf{x}}^{\top}\widetilde{\mathbf{A}}\,\tilde{\mathbf{x}} = \mathbf{x}^{\top}\mathbf{A}\mathbf{x} + \mathbf{x}^{\top}\mathbf{b} + \mathbf{b}^{\top}\mathbf{x} + c = Q(\mathbf{x}),$$

the two middle terms being equal because each is a 1×1 matrix and one is the transpose of the other. We call $\widetilde{\mathbf{A}}$ the **big matrix** of Q and $\mathbf{A}$ its **small matrix**. The extra coordinate is the same trick that turned an affine map into an $(n+1)\times(n+1)$ block matrix in Proposition 22.1.7, and the two devices fit together exactly.

Proposition 22.5.1. Affine Substitution Is Congruence of the Big Matrix

Let Q have big matrix $\widetilde{\mathbf{A}}$, let $\mathbf{M} \in M_n(F)$ be invertible, let $\mathbf{t} \in F^n$, and let $g(\mathbf{y}) = \mathbf{M}\mathbf{y} + \mathbf{t}$ be the corresponding invertible affine map (Definition 22.1.3), with block matrix (Proposition 22.1.7)

$$\widehat{g} = \begin{pmatrix} \mathbf{M} & \mathbf{t} \\ \mathbf{0}^{\top} & 1 \end{pmatrix}.$$

Then $Q \circ g$ is again a quadric polynomial, its big matrix is $\widehat{g}^{\top}\widetilde{\mathbf{A}}\widehat{g}$, and its small matrix, linear vector, and constant are

$$\mathbf{A}' = \mathbf{M}^{\top}\mathbf{A}\mathbf{M}, \qquad \mathbf{b}' = \mathbf{M}^{\top}(\mathbf{A}\mathbf{t} + \mathbf{b}), \qquad c' = Q(\mathbf{t}).$$

Proof. By Proposition 22.1.7, $\widehat{g}\,\tilde{\mathbf{y}} = (\mathbf{M}\mathbf{y} + \mathbf{t},\, 1) = \widetilde{g(\mathbf{y})}$. Hence

$$Q(g(\mathbf{y})) = \widetilde{g(\mathbf{y})}^{\top}\,\widetilde{\mathbf{A}}\,\widetilde{g(\mathbf{y})} = \tilde{\mathbf{y}}^{\top}(\widehat{g}^{\top}\widetilde{\mathbf{A}}\widehat{g})\tilde{\mathbf{y}},$$

and $\widehat{g}^{\top}\widetilde{\mathbf{A}}\widehat{g}$ is symmetric, being a congruence of a symmetric matrix (Proposition 14.2.4). Multiplying the blocks out,

$$\begin{aligned} \begin{pmatrix} \mathbf{M}^{\top} & \mathbf{0} \\ \mathbf{t}^{\top} & 1 \end{pmatrix}\begin{pmatrix} \mathbf{A} & \mathbf{b} \\ \mathbf{b}^{\top} & c \end{pmatrix}\begin{pmatrix} \mathbf{M} & \mathbf{t} \\ \mathbf{0}^{\top} & 1 \end{pmatrix} &= \begin{pmatrix} \mathbf{M}^{\top}\mathbf{A} & \mathbf{M}^{\top}\mathbf{b} \\ \mathbf{t}^{\top}\mathbf{A} + \mathbf{b}^{\top} & \mathbf{t}^{\top}\mathbf{b} + c \end{pmatrix}\begin{pmatrix} \mathbf{M} & \mathbf{t} \\ \mathbf{0}^{\top} & 1 \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{M}^{\top}\mathbf{A}\mathbf{M} & \mathbf{M}^{\top}(\mathbf{A}\mathbf{t} + \mathbf{b}) \\ (\mathbf{A}\mathbf{t} + \mathbf{b})^{\top}\mathbf{M} & \mathbf{t}^{\top}\mathbf{A}\mathbf{t} + 2\mathbf{b}^{\top}\mathbf{t} + c \end{pmatrix}, \end{aligned}$$

where the bottom-left block was transposed using $\mathbf{A}^{\top} = \mathbf{A}$. Reading off the blocks gives $\mathbf{A}'$, $\mathbf{b}'$ and $c' = \mathbf{t}^{\top}\mathbf{A}\mathbf{t} + 2\mathbf{b}^{\top}\mathbf{t} + c = Q(\mathbf{t})$. Finally $\mathbf{A}' = \mathbf{M}^{\top}\mathbf{A}\mathbf{M}$ is nonzero, since $\mathbf{M}$ is invertible and congruence preserves rank (Theorem 14.2.3), so $Q \circ g$ is a quadric polynomial. This proves the proposition. ∎

Three consequences are worth naming at once. The new constant is Q evaluated at the new origin — obvious in hindsight, since $\mathbf{y} = \mathbf{0}$ means $\mathbf{x} = \mathbf{t}$. The new small matrix is a congruence

of the old one, so **the rank of $\mathbf{A}$ is an affine invariant of Q** by Theorem 14.2.3, and over $\mathbb{R}$ **so is its inertia** by Corollary 14.5.2 (b) — the latter only up to the simultaneous exchange of n_+ with n_- that a negative scalar factor forces, as Step 4 of the classification theorem below records. And $\widetilde{\mathbf{A}}$ itself only undergoes congruences by matrices whose last row is $(\mathbf{0}^\top, 1)$, which is a strictly smaller set than all congruences: that restriction is exactly why the affine classification is finer than Chapter 14's.

Replacing Q by λQ with $\lambda \neq 0$ does not change the zero set and replaces $\widetilde{\mathbf{A}}$ by $\lambda\widetilde{\mathbf{A}}$, so a scalar factor costs nothing geometrically and is worth allowing alongside the substitution.

Definition 22.5.2. Affine Equivalence of Quadric Polynomials

Two quadric polynomials Q_1, Q_2 on F^n are **affinely equivalent** when

$$Q_2 = \lambda\,(Q_1 \circ g)$$

for some invertible affine map g (Definition 22.1.3) and some $\lambda \in F$ with $\lambda \neq 0$.

This is an equivalence relation, because the invertible affine maps form a group and the nonzero scalars do too.

22.5.4 Centers

A circle has a center and a parabola does not, and the difference is visible in the equation before any picture is drawn.

A center is a point about which the equation is symmetric.

Definition 22.5.3. Center of a Quadric

Let Q be a quadric polynomial on F^n. A point $\mathbf{p} \in F^n$ is a **center** of Q if

$$Q(\mathbf{p}+\mathbf{v}) = Q(\mathbf{p}-\mathbf{v}) \qquad \text{for every } \mathbf{v} \in F^n.$$

Q is **central** if it has at least one center, and **centerless** otherwise.

In words: reflecting in $\mathbf{p}$, the map $\mathbf{p}+\mathbf{v} \mapsto \mathbf{p}-\mathbf{v}$, carries the quadric to itself, and more than that, leaves the whole polynomial unchanged. The next proposition turns this into a linear system.

Proposition 22.5.2. When a Center Exists

Let $\operatorname{char} F \neq 2$ and let $Q(\mathbf{x}) = \mathbf{x}^\top\mathbf{A}\mathbf{x} + 2\mathbf{b}^\top\mathbf{x} + c$ be a quadric polynomial on F^n. Then:

(a) $\mathbf{p}$ is a center of Q if and only if $\mathbf{A}\mathbf{p} + \mathbf{b} = \mathbf{0}$;

(b) Q is central if and only if $\mathbf{b} \in \operatorname{col}(\mathbf{A})$; in particular Q is central whenever $\mathbf{A}$ is invertible, and the center is then the unique point $-\mathbf{A}^{-1}\mathbf{b}$;

(c) if $\mathbf{p}$ is a center then $Q(\mathbf{p}+\mathbf{v}) = Q(\mathbf{p}) + \mathbf{v}^\top\mathbf{A}\mathbf{v}$ for every $\mathbf{v}$, and the value $Q(\mathbf{p})$ is the same at every center.

Idea. Expand $Q(\mathbf{p} \pm \mathbf{v})$ and subtract. Everything even in $\mathbf{v}$ cancels and what is left is $4\mathbf{v}^\top(\mathbf{A}\mathbf{p} + \mathbf{b})$, a linear functional of $\mathbf{v}$. A linear functional vanishes identically exactly when its vector vanishes, so the condition on $\mathbf{p}$ is one linear equation per coordinate. The 4 is where $\operatorname{char} F \neq 2$ is spent.

Proof.

(a) Expanding and using $\mathbf{A}^\top = \mathbf{A}$, so that $\mathbf{p}^\top\mathbf{A}\mathbf{v} = \mathbf{v}^\top\mathbf{A}\mathbf{p}$,

$$Q(\mathbf{p} \pm \mathbf{v}) = \mathbf{p}^\top\mathbf{A}\mathbf{p} \pm 2\mathbf{v}^\top\mathbf{A}\mathbf{p} + \mathbf{v}^\top\mathbf{A}\mathbf{v} + 2\mathbf{b}^\top\mathbf{p} \pm 2\mathbf{b}^\top\mathbf{v} + c.$$

Subtracting the two,

$$Q(\mathbf{p} + \mathbf{v}) - Q(\mathbf{p} - \mathbf{v}) = 4\mathbf{v}^\top\mathbf{A}\mathbf{p} + 4\mathbf{b}^\top\mathbf{v} = 4\,\mathbf{v}^\top(\mathbf{A}\mathbf{p} + \mathbf{b}). \tag{22.5.1}$$

Since $\operatorname{char} F \neq 2$, the element $4 = 2 \cdot 2$ is nonzero. So $\mathbf{p}$ is a center exactly when $\mathbf{v}^\top(\mathbf{A}\mathbf{p} + \mathbf{b}) = 0$ for every $\mathbf{v}$. Taking $\mathbf{v} = \mathbf{e}_i$ shows that this forces every entry of $\mathbf{A}\mathbf{p} + \mathbf{b}$ to vanish; conversely $\mathbf{A}\mathbf{p} + \mathbf{b} = \mathbf{0}$ makes Eq. 22.5.1 zero for every $\mathbf{v}$. Hence (a).

(b) By (a), a center is a solution of $\mathbf{A}\mathbf{p} = -\mathbf{b}$, and such a solution exists exactly when $-\mathbf{b}$, equivalently $\mathbf{b}$, lies in $\operatorname{col}(\mathbf{A})$ (Theorem 0.9.3). If $\mathbf{A}$ is invertible then $\operatorname{col}(\mathbf{A}) = F^n$ and $\mathbf{p} = -\mathbf{A}^{-1}\mathbf{b}$ is the only solution.

(c) With $\mathbf{A}\mathbf{p} + \mathbf{b} = \mathbf{0}$, the expansion above with the sign $+$ reads

$$Q(\mathbf{p} + \mathbf{v}) = (\mathbf{p}^\top\mathbf{A}\mathbf{p} + 2\mathbf{b}^\top\mathbf{p} + c) + 2\mathbf{v}^\top(\mathbf{A}\mathbf{p} + \mathbf{b}) + \mathbf{v}^\top\mathbf{A}\mathbf{v} = Q(\mathbf{p}) + \mathbf{v}^\top\mathbf{A}\mathbf{v}.$$

If $\mathbf{p}'$ is a second center then $\mathbf{A}(\mathbf{p}' - \mathbf{p}) = (-\mathbf{b}) - (-\mathbf{b}) = \mathbf{0}$, so putting $\mathbf{v} = \mathbf{p}' - \mathbf{p}$ in the last identity gives $Q(\mathbf{p}') = Q(\mathbf{p}) + \mathbf{v}^\top\mathbf{A}\mathbf{v} = Q(\mathbf{p}) + \mathbf{v}^\top\mathbf{0} = Q(\mathbf{p})$. This proves the proposition.

■

Part (c) is the useful half. Translating the origin to a center — that is, applying Proposition 22.5.1 with $\mathbf{M} = \mathbf{I}$ and $\mathbf{t} = \mathbf{p}$ — gives $\mathbf{b}' = \mathbf{A}\mathbf{p} + \mathbf{b} = \mathbf{0}$, so the equation becomes $\mathbf{v}^\top\mathbf{A}\mathbf{v} + Q(\mathbf{p}) = 0$ with no linear term at all, and Chapter 14 applies verbatim. That single move is the engine of the classification.

Example. The parabola is centerless

Show that $Q(x, y) = x^2 - y$ over $\mathbb{R}$ has no center, and find all centers of $Q_1(x, y) = x^2 - 1$.

Solution. For Q, $\mathbf{A} = \operatorname{diag}(1, 0)$ and $\mathbf{b} = (0, -\frac{1}{2})$. Then $\operatorname{col}(\mathbf{A}) = \operatorname{span}(\mathbf{e}_1)$ and $\mathbf{b} \notin \operatorname{col}(\mathbf{A})$ because its second entry is nonzero, so by Proposition 22.5.2 (b) there is no center. Concretely, $Q(\mathbf{p}+\mathbf{v}) - Q(\mathbf{p}-\mathbf{v}) = 4\mathbf{v}^\top(\mathbf{A}\mathbf{p}+\mathbf{b}) = 4(v_1p_1 - \frac{1}{2}v_2)$, which is $-2 \neq 0$ at $\mathbf{v} = \mathbf{e}_2$ whatever $\mathbf{p}$ is.

For Q_1, $\mathbf{A} = \operatorname{diag}(1, 0)$ and $\mathbf{b} = \mathbf{0} \in \operatorname{col}(\mathbf{A})$, so the centers are the solutions of $\mathbf{A}\mathbf{p} = \mathbf{0}$, namely the whole y-axis. That is right: Q_1 defines the pair of lines $x = \pm 1$, and reflecting in any point of the y-axis swaps them. And $Q_1(\mathbf{p}) = -1$ at every one of those centers, as

Proposition 22.5.2 (c) promises.

Quick check. A quadric polynomial on F^n with $\operatorname{rank}\mathbf{A} = n$ has exactly one center, and one with $\operatorname{rank}\mathbf{A} = n-1$ has either none or a whole line of them. Which is it for x^2+y^2-2x on $\mathbb{R}^2$, and which for x^2+2y?

Solution. For x^2+y^2-2x we have $\mathbf{A} = \mathbf{I}_2$ of rank 2 and $\mathbf{b} = (-1,0)$, so the unique center is $-\mathbf{A}^{-1}\mathbf{b} = (1,0)$; indeed the curve is the circle $(x-1)^2+y^2 = 1$. For x^2+2y we have $\mathbf{A} = \operatorname{diag}(1,0)$ of rank 1 and $\mathbf{b} = (0,1) \notin \operatorname{col}(\mathbf{A}) = \operatorname{span}(\mathbf{e}_1)$, so there is no center at all: it is a parabola. The line of centers occurs instead for x^2-1, where $\mathbf{b} = \mathbf{0}$ does lie in $\operatorname{col}(\mathbf{A})$.

22.5.5 The affine classification over the reals

Now the theorem. Over $\mathbb{R}$ we may diagonalize the small matrix by Sylvester (Corollary 14.5.2), translate to kill as much of the linear part as a center allows, and then scale. What survives is a short list.

Theorem 22.5.3. Affine Classification of Real Quadrics

Let $Q(\mathbf{x}) = \mathbf{x}^\top\mathbf{A}\mathbf{x} + 2\mathbf{b}^\top\mathbf{x} + c$ be a quadric polynomial on $\mathbb{R}^n$ and let $r = \operatorname{rank}\mathbf{A}$, which is ≥ 1 because $\mathbf{A} \neq \mathbf{O}$ (Definition 22.5.1). Then Q is affinely equivalent to **exactly one** polynomial from the following list, for one value of p in the range shown, where $\sigma_{r,p}(\mathbf{y}) = y_1^2+\cdots+y_p^2-y_{p+1}^2-\cdots-y_r^2$:

(I) $\sigma_{r,p}(\mathbf{y}) - 1$, with $0 \leq p \leq r$;
(II) $\sigma_{r,p}(\mathbf{y})$, with $r \leq 2p \leq 2r$;
(III) $\sigma_{r,p}(\mathbf{y}) - y_{r+1}$, with $r < n$ and $r \leq 2p \leq 2r$.

Case (III) occurs exactly when Q is centerless. The parameter p that occurs is $n_+(\mathbf{A})$ or $n_-(\mathbf{A})$, the two being interchanged by a negative scalar factor (Definition 14.5.1).

Idea. Three moves, in this order, and each is a result already proved.

① **Diagonalize the quadratic part.** Sylvester's law (Corollary 14.5.2) produces coordinates in which the quadratic part is $\sigma_{r,p}$; by Proposition 22.5.1 this changes $\mathbf{b}$ and c but nothing else.

② **Complete the square in the first r variables.** A translation kills the linear coefficients $b_1,\dots,b_r$, because those are exactly the coordinates in which $\mathbf{A}$ is invertible. This is Proposition 22.5.2 read one coordinate at a time.

③ **Deal with what is left,** a polynomial $\sigma_{r,p} + 2\sum_{i>r} b_i y_i + c$ in which the y_i with $i > r$ appear only linearly. If they do not appear at all, scaling clears the constant to 0 or -1, giving (II) or (I). If one of them does appear, it can be used as a new coordinate that swallows the constant too, giving (III).

That the list has no repetitions is a separate, purely computational matter: the pair (inertia

of $\mathbf{A}$, inertia of $\widetilde{\mathbf{A}}$) is an affine invariant by Proposition 22.5.1, and reading it off the three families separates them.

Proof. **Step 1: diagonalize.** Write $(p, r-p, n-r)$ for the inertia of $\mathbf{A}$ (Definition 14.5.1); this p is the one carried through Steps 1 to 3. By Corollary 14.5.2 (a) there is an invertible $\mathbf{M}_1$ with $\mathbf{M}_1^{\top}\mathbf{A}\mathbf{M}_1 = \mathbf{D}$, where $\mathbf{D} = \operatorname{diag}(\epsilon_1, \dots, \epsilon_n)$ has $\epsilon_i = 1$ for $i \le p$, $\epsilon_i = -1$ for $p < i \le r$, and $\epsilon_i = 0$ for $i > r$. Substituting $\mathbf{x} = \mathbf{M}_1\mathbf{y}$ and applying Proposition 22.5.1 with $\mathbf{t} = \mathbf{0}$, the polynomial becomes

$$Q_1(\mathbf{y}) = \sigma_{r,p}(\mathbf{y}) + 2\mathbf{b}_1^{\top}\mathbf{y} + c, \qquad \mathbf{b}_1 = \mathbf{M}_1^{\top}\mathbf{b}.$$

Step 2: complete the square. Let $\mathbf{t}$ have entries $t_i = -\epsilon_i b_{1,i}$ for $i \le r$ and $t_i = 0$ for $i > r$, and substitute $\mathbf{y} = \mathbf{z} + \mathbf{t}$. By Proposition 22.5.1 with $\mathbf{M} = \mathbf{I}$, the new linear vector is $\mathbf{b}_2 = \mathbf{D}\mathbf{t} + \mathbf{b}_1$, whose i-th entry for $i \le r$ is $\epsilon_i t_i + b_{1,i} = -\epsilon_i^2 b_{1,i} + b_{1,i} = 0$, since $\epsilon_i^2 = 1$ there; for $i > r$ it is $b_{1,i}$, unchanged. The new constant is $c_2 = Q_1(\mathbf{t})$. So

$$Q_2(\mathbf{z}) = \sigma_{r,p}(\mathbf{z}) + 2\sum_{i>r} b_{1,i} z_i + c_2.$$

Step 3, Case A: $b_{1,i} = 0$ for every $i > r$. Then $Q_2 = \sigma_{r,p} + c_2$.

If $c_2 = 0$, this is $\sigma_{r,p}$. If $r \le 2p$ we are in case (II). Otherwise multiply by $\lambda = -1$ and reverse the order of the first r coordinates, which turns $\sigma_{r,p}$ into $\sigma_{r,\,r-p}$ with $r \le 2(r-p)$; again case (II).

If $c_2 \neq 0$, substitute $z_i = |c_2|^{1/2} w_i$ for $i \le r$ and $z_i = w_i$ otherwise, an invertible linear change, and then multiply by $\lambda = 1/|c_2|$. The result is $\sigma_{r,p}(\mathbf{w}) + \operatorname{sgn}(c_2)$. If $\operatorname{sgn}(c_2) = -1$ this is case (I) with parameter p. If $\operatorname{sgn}(c_2) = +1$, multiply by -1 and reverse the order of the first r coordinates to get $\sigma_{r,\,r-p} - 1$, case (I) with parameter $r-p$.

Step 3, Case B: $b_{1,i} \neq 0$ for some $i > r$. Here $r < n$. Put $\mathbf{c} = (b_{1,r+1}, \dots, b_{1,n}) \neq \mathbf{0}$ in $\mathbb{R}^{n-r}$. Extend $2\mathbf{c}$ to a basis of $\mathbb{R}^{n-r}$ (Theorem 1.6.5) and let $\mathbf{N} \in M_{n-r}(\mathbb{R})$ be the invertible matrix whose **first row** is $2\mathbf{c}^{\top}$ and whose remaining rows complete it to an invertible matrix; such rows exist because a basis of $\mathbb{R}^{n-r}$ written as rows forms an invertible matrix. Define new coordinates by

$$u_i = z_i \ (i \le r), \qquad (u_{r+1}, \dots, u_n)^{\top} = \mathbf{N}(z_{r+1}, \dots, z_n)^{\top} + c_2\,\mathbf{e}_1,$$

an invertible affine substitution, under which $u_{r+1} = 2\mathbf{c}^{\top}(z_{r+1}, \dots, z_n) + c_2$. Hence $Q_2 = \sigma_{r,p}(\mathbf{u}) + u_{r+1}$. Replacing u_{r+1} by $-y_{r+1}$ and keeping the other coordinates gives $\sigma_{r,p}(\mathbf{y}) - y_{r+1}$. If $r > 2p$, multiply by -1, reverse the order of the first r coordinates and replace y_{r+1} by $-y_{r+1}$; the result is $\sigma_{r,\,r-p}(\mathbf{y}) - y_{r+1}$ with $r \le 2(r-p)$. Either way we reach case (III).

Step 4: the list has no repetitions. By Proposition 22.5.1 an affine equivalence replaces $\mathbf{A}$ by $\lambda\mathbf{M}^{\top}\mathbf{A}\mathbf{M}$ and $\widetilde{\mathbf{A}}$ by $\lambda\widehat{g}^{\top}\widetilde{\mathbf{A}}\widehat{g}$, with $\mathbf{M}$ and $\widehat{g}$ invertible. A congruence preserves inertia (Corollary 14.5.2 (b)), and multiplying a real symmetric matrix by λ leaves the inertia alone

if $\lambda > 0$ and exchanges n_+ with n_- if $\lambda < 0$. Hence the **invariant pair**

$$(\operatorname{In}\mathbf{A},\ \operatorname{In}\widetilde{\mathbf{A}}),$$
taken up to exchanging n_+ with n_-
in both entries simultaneously,

in which In is the inertia triple of Definition 17.10.1, is an invariant of the affine equivalence class. Write $\mathbf{D}_{r,p} = \operatorname{diag}(\epsilon_1, \dots, \epsilon_n)$ for the small matrix shared by all three families. In family (I) the linear vector is $\mathbf{0}$ and the constant is -1, so $\widetilde{\mathbf{A}} = \mathbf{D}_{r,p} \oplus (-1)$; in family (II) the constant is 0, so $\widetilde{\mathbf{A}} = \mathbf{D}_{r,p} \oplus 0$. In family (III) the linear vector is $-\frac{1}{2}\mathbf{e}_{r+1}$ and the constant is 0, so $\widetilde{\mathbf{A}}$ is diagonal except in the four positions indexed by $r+1$ and $n+1$, where it carries the block $\begin{pmatrix} 0 & -\frac{1}{2} \\ -\frac{1}{2} & 0 \end{pmatrix}$, whose eigenvalues are $\frac{1}{2}$ and $-\frac{1}{2}$; the remaining diagonal entries are $\epsilon_1, \dots, \epsilon_r$ and $n - r - 1$ zeros. So

$$\begin{aligned}
\text{(I)}: &\quad \operatorname{In}\mathbf{A} = (p,\ r-p,\ n-r), &\quad \operatorname{In}\widetilde{\mathbf{A}} = (p,\ r-p+1,\ n-r);\\
\text{(II)}: &\quad \operatorname{In}\mathbf{A} = (p,\ r-p,\ n-r), &\quad \operatorname{In}\widetilde{\mathbf{A}} = (p,\ r-p,\ n-r+1);\\
\text{(III)}: &\quad \operatorname{In}\mathbf{A} = (p,\ r-p,\ n-r), &\quad \operatorname{In}\widetilde{\mathbf{A}} = (p+1,\ r-p+1,\ n-r-1).
\end{aligned}$$

Therefore $\operatorname{rank}\mathbf{A} = r$ in all three, while $\operatorname{rank}\widetilde{\mathbf{A}}$ is $r+1$, r and $r+2$ respectively. Ranks are unaffected by the exchange of n_+ and n_-, so the pair $(\operatorname{rank}\mathbf{A}, \operatorname{rank}\widetilde{\mathbf{A}})$ already determines both the family and r.

It remains to separate two members of one family with the same r. If the invariants agree without the exchange, then $\operatorname{In}\mathbf{A}$ gives $p = p'$ at once. Suppose instead they agree after the exchange, so that $p' = r - p$.

Family (I). The exchanged $\operatorname{In}\widetilde{\mathbf{A}}$ of the p-member is $(r-p+1,\ p,\ n-r)$, while the p'-member has $(p',\ r-p'+1,\ n-r) = (r-p,\ p+1,\ n-r)$. Comparing first entries gives $r-p+1 = r-p$, which is false. So family (I) has no repetitions.

Families (II) and (III). Here both members satisfy $r \le 2p$ and $r \le 2p' = 2(r-p)$, that is $2p \le r$. Hence $2p = r = 2p'$ and $p = p'$.

So distinct entries of the list are affinely inequivalent, and Step 3 showed every Q meets one. The normal form reached has small matrix $\mathbf{D}_{r,p}$, of inertia $(p,\ r-p,\ n-r)$; since $\operatorname{In}\mathbf{A}$ is preserved by affine equivalence up to the exchange of n_+ with n_-, that p is $n_+(\mathbf{A})$ or $n_-(\mathbf{A})$, as claimed. Finally, Q is centerless exactly in case (III): by Proposition 22.5.2 (b) centrality is the condition $\mathbf{b} \in \operatorname{col}(\mathbf{A})$, which is preserved by affine equivalence because centers are (if $\mathbf{p}$ is a center of Q and g is invertible affine, then $g^{-1}(\mathbf{p})$ is a center of $\lambda\, Q \circ g$, since $\mathbf{v} \mapsto \mathbf{M}\mathbf{v}$ is a bijection of $\mathbb{R}^n$); and $\mathbf{0}$ is a center in cases (I) and (II), while in case (III) the small matrix is $\mathbf{D}_{r,p}$ and the linear vector is $-\frac{1}{2}\mathbf{e}_{r+1} \notin \operatorname{col}(\mathbf{D}_{r,p}) = \operatorname{span}(\mathbf{e}_1, \dots, \mathbf{e}_r)$. This proves the theorem.

■

22.5.5.1 The nine real conics

For $n = 2$ the parameters run over $r \in \{1, 2\}$ and $0 \le p \le r$, and the list has nine entries. Here they are, with the zero set of each. The last column records the classical names, and so does the paragraph on $n = 3$ below; those names are not defined anywhere in this book, and each set-theoretic reading is a one-line check from the normal form — for instance $x^2 + y^2 = -1$ has

no real solution because a sum of squares is nonnegative, and $x^2 - y^2 = 0$ is the pair of lines $y = \pm x$, by factoring.

Normal form	(r, p)	Family	In $\mathbf{A}$	In $\widetilde{\mathbf{A}}$	Zero set
$x^2 + y^2 = 1$	$(2, 2)$	I	$(2, 0, 0)$	$(2, 1, 0)$	ellipse
$x^2 - y^2 = 1$	$(2, 1)$	I	$(1, 1, 0)$	$(1, 2, 0)$	hyperbola
$x^2 + y^2 = -1$	$(2, 0)$	I	$(2, 0, 0)$	$(3, 0, 0)$	empty
$x^2 + y^2 = 0$	$(2, 2)$	II	$(2, 0, 0)$	$(2, 0, 1)$	one point
$x^2 - y^2 = 0$	$(2, 1)$	II	$(1, 1, 0)$	$(1, 1, 1)$	two crossing lines
$x^2 = 1$	$(1, 1)$	I	$(1, 0, 1)$	$(1, 1, 1)$	two parallel lines
$x^2 = -1$	$(1, 0)$	I	$(1, 0, 1)$	$(2, 0, 1)$	empty
$x^2 = 0$	$(1, 1)$	II	$(1, 0, 1)$	$(1, 0, 2)$	one line
$x^2 = y$	$(1, 1)$	III	$(1, 0, 1)$	$(2, 1, 0)$	parabola

Two rows of the table are written with the parameter p of the negated polynomial: $x^2 + y^2 = -1$ means $-x^2 - y^2 - 1 = 0$, which is $\sigma_{2,0} - 1$ after multiplying by -1, and likewise for $x^2 = -1$. Reading the two inertia columns up to a simultaneous exchange of the first two entries confirms that no two rows agree: rows 1, 3, 4 have $\operatorname{In}\mathbf{A} = (2, 0, 0)$ and differ in $\operatorname{In}\widetilde{\mathbf{A}}$; rows 2 and 5 differ in $\operatorname{In}\widetilde{\mathbf{A}}$; and rows 6, 7, 8, 9 have $\operatorname{In}\mathbf{A} = (1, 0, 1)$ with four different $\operatorname{In}\widetilde{\mathbf{A}}$. In each of those two groups the exchange changes $\operatorname{In}\mathbf{A}$, so it cannot rescue a match. The one group where $\operatorname{In}\mathbf{A}$ survives the exchange is rows 2 and 5, with $\operatorname{In}\mathbf{A} = (1, 1, 0)$; there $\operatorname{In}\widetilde{\mathbf{A}}$ is $(1, 2, 0)$ against $(1, 1, 1)$, and the exchanged forms $(2, 1, 0)$ and $(1, 1, 1)$ still disagree.

Rows 3 and 7 are the honest surprise: both zero sets are empty, and both are needed, because the polynomials are not affinely equivalent. Over $\mathbb{C}$ they are quite different objects. Row 7 is $x^2 + 1 = (x - i)(x + i)$, a pair of distinct lines, while row 3 has an invertible small matrix. Exercise (C2: Over another field) (a) below classifies affine quadrics over $\mathbb{C}$, where rank alone is the invariant and the two rows stay apart: row 3 has $\operatorname{rank}\mathbf{A} = 2$ and $\operatorname{rank}\widetilde{\mathbf{A}} = 3$, row 7 has 1 and 2. Read projectively, Section 10's rank criterion makes row 3 a smooth conic.

For $n = 3$ the same count gives $4 + 7 + 6 = 17$ normal forms. The new shapes are the ellipsoid $x^2 + y^2 + z^2 = 1$, the hyperboloids of one and two sheets $x^2 + y^2 - z^2 = \pm 1$, the cone $x^2 + y^2 - z^2 = 0$, the elliptic and hyperbolic paraboloids $x^2 \pm y^2 = z$, and the three cylinders $x^2 + y^2 = 1$, $x^2 - y^2 = 1$, $x^2 = y$, which are the conics of the table dragged along the missing coordinate. The rest are a pair of intersecting planes, a pair of parallel planes, a double plane, a line, a point, and three empty sets.

Warning. **The rank of $\widetilde{\mathbf{A}}$ does not classify; the inertia does.** The ellipse $x^2 + y^2 - 1$, the hyperbola $x^2 - y^2 - 1$, and the empty conic $x^2 + y^2 + 1$ all have $\operatorname{rank}\mathbf{A} = 2$ and $\operatorname{rank}\widetilde{\mathbf{A}} = 3$. Their small matrices already differ in inertia in the hyperbola's case, and the ellipse and the empty conic are told apart only by $\operatorname{In}\widetilde{\mathbf{A}} = (2, 1, 0)$ against $(3, 0, 0)$. A count of nonzero eigenvalues cannot distinguish a curve from nothing at all.

22.5.6 Euclidean quadrics and principal axes

Affine equivalence forgets shape: every ellipse is affinely a circle. To keep shape, restrict the allowed substitutions to the Euclidean motions of $\mathbb{R}^n$, which by Theorem 22.4.1 are exactly the maps $\mathbf{x} = \mathbf{U}\mathbf{y} + \mathbf{t}$ with $\mathbf{U} \in \mathrm{O}(n)$; that theorem gives one direction, and the other is that such a map preserves distances, since $\mathbf{U}$ preserves norms (Theorem 11.7.3). Call two quadric polynomials **Euclidean equivalent** when one is $Q \circ g$ for such a g. Note that no scalar factor is allowed now, and that is deliberate: a scalar factor would rescale the lengths we are trying to measure.

The gain is immediate. Under $\mathbf{x} = \mathbf{U}\mathbf{y} + \mathbf{t}$ the small matrix becomes $\mathbf{U}^{\top}\mathbf{A}\mathbf{U}$, which is both a congruence and a **similarity**, since $\mathbf{U}^{\top} = \mathbf{U}^{-1}$. So the eigenvalues of $\mathbf{A}$, not merely their signs, are Euclidean invariants. This is the observation Chapter 12 made in Theorem 12.5.5, and the classification below is that theorem with the origin allowed to move.

Theorem 22.5.4. Euclidean Classification of Real Quadrics

Let $Q(\mathbf{x}) = \mathbf{x}^{\top}\mathbf{A}\mathbf{x} + 2\mathbf{b}^{\top}\mathbf{x} + c$ be a quadric polynomial on $\mathbb{R}^n$, let $r = \operatorname{rank}\mathbf{A}$, and let $\lambda_1, \dots, \lambda_r$ be the nonzero eigenvalues of $\mathbf{A}$, listed with multiplicity. Then there is a Euclidean motion $\mathbf{x} = \mathbf{U}\mathbf{y} + \mathbf{t}$, with $\mathbf{U} \in \mathrm{O}(n)$, after which Q becomes exactly one of

(a) $\lambda_1 y_1^2 + \cdots + \lambda_r y_r^2 + c_0$, when Q is central, where $c_0 = Q(\mathbf{p})$ at any center $\mathbf{p}$;
(b) $\lambda_1 y_1^2 + \cdots + \lambda_r y_r^2 + 2\mu\, y_{r+1}$, when Q is centerless, where $\mu = \|\mathbf{b} - P_{\mathrm{col}(\mathbf{A})}\mathbf{b}\| > 0$.

The columns of $\mathbf{U}$ are an orthonormal basis of eigenvectors of $\mathbf{A}$, the first r of them for $\lambda_1, \dots, \lambda_r$; these are the **principal axes** of Q. The multiset $\{\lambda_1, \dots, \lambda_r\}$ and the number c_0, respectively μ, are unchanged by every Euclidean motion, so the normal form is unique up to the order of its terms.

Idea. The spectral theorem replaces Step 1 of Theorem 22.5.3, and it is the only change: instead of scaling the eigenvalues away to ± 1 we keep them, because an orthogonal matrix is not allowed to rescale. Steps 2 and 3 then go through with one adjustment. In Step 3 we may no longer stretch a coordinate to absorb $\mathbf{c}$; instead we **rotate** the last $n - r$ coordinates so that $\mathbf{c}$ points along one of them, which is legitimate because those coordinates carry the eigenvalue 0 and a rotation among them leaves the quadratic part untouched.

Proof. **Step 1.** By Corollary 12.5.2 there is $\mathbf{U}_1 \in \mathrm{O}(n)$ with $\mathbf{U}_1^{\top}\mathbf{A}\mathbf{U}_1 = \mathbf{D} = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$, the eigenvalues of $\mathbf{A}$ with multiplicity, and we order the basis so that $\lambda_1, \dots, \lambda_r$ are the nonzero ones; the count is right because $\mathbf{D}$ is similar to $\mathbf{A}$, so $\operatorname{rank}\mathbf{D} = \operatorname{rank}\mathbf{A} = r$, and a diagonal matrix has rank equal to its number of nonzero entries. Substituting $\mathbf{x} = \mathbf{U}_1\mathbf{y}$ and using Proposition 22.5.1 with $\mathbf{t} = \mathbf{0}$ gives

$$Q_1(\mathbf{y}) = \sum_{i=1}^{r} \lambda_i y_i^2 + 2\mathbf{b}_1^{\top}\mathbf{y} + c, \qquad \mathbf{b}_1 = \mathbf{U}_1^{\top}\mathbf{b}.$$

Step 2. Translate by $\mathbf{t}_1$ with entries $-b_{1,i}/\lambda_i$ for $i \le r$ and 0 otherwise, which is legitimate because those λ_i are nonzero. As in Theorem 22.5.3, Step 2, the new linear vector has i-th entry $\lambda_i t_{1,i} + b_{1,i} = 0$ for $i \le r$ and $b_{1,i}$ for $i > r$. Write $\mathbf{c} = (b_{1,r+1}, \dots, b_{1,n})$ and c' for the

new constant, so

$$Q_2(\mathbf{z}) = \sum_{i=1}^{r} \lambda_i z_i^2 + 2\mathbf{c}^\top (z_{r+1}, \dots, z_n) + c'.$$

Claim: $\mathbf{c} = \mathbf{0}$ *if and only if* Q *is central, and in general* $\|\mathbf{c}\| = \|\mathbf{b} - P_{\mathrm{col}(\mathbf{A})}\mathbf{b}\|$. The last $n - r$ columns $\mathbf{q}_{r+1}, \dots, \mathbf{q}_n$ of $\mathbf{U}_1$ are an orthonormal basis of the eigenspace of $\mathbf{A}$ for 0, that is of $\mathrm{nul}(\mathbf{A})$, and the entries of $\mathbf{c}$ are $\mathbf{q}_i^\top \mathbf{b} = \langle \mathbf{b}, \mathbf{q}_i \rangle$. By Theorem 11.6.5 (b) applied to the self-adjoint operator $\mathbf{x} \mapsto \mathbf{A}\mathbf{x}$, we have $\mathrm{col}(\mathbf{A}) = \mathrm{nul}(\mathbf{A})^\perp$, so by Theorem 11.3.2 the component of $\mathbf{b}$ in $\mathrm{nul}(\mathbf{A})$ is $\mathbf{b} - P_{\mathrm{col}(\mathbf{A})}\mathbf{b}$, and by Theorem 11.2.1 its norm is $\|\mathbf{c}\|$. Hence $\mathbf{c} = \mathbf{0}$ exactly when $\mathbf{b} \in \mathrm{col}(\mathbf{A})$, which is centrality by Proposition 22.5.2 (b). This proves the claim.

Step 3, central case. Here $\mathbf{c} = \mathbf{0}$, so $Q_2 = \sum_{i \le r} \lambda_i z_i^2 + c'$, which is form (a). The new origin $\mathbf{p}$ satisfies $\mathbf{A}\mathbf{p} + \mathbf{b} = \mathbf{0}$ by Proposition 22.5.2 (a), because the linear part has vanished, so $\mathbf{p}$ is a center and $c' = Q(\mathbf{p}) = c_0$ by Proposition 22.5.1; Proposition 22.5.2 (c) says the value does not depend on which center.

Step 3, centerless case. Here $\mathbf{c} \neq \mathbf{0}$. By Corollary 11.2.3 the unit vector $\mathbf{c}/\|\mathbf{c}\|$ extends to an orthonormal basis of $\mathbb{R}^{n-r}$; let $\mathbf{R} \in \mathrm{O}(n-r)$ have that basis as its columns, so $\mathbf{R}\mathbf{e}_1 = \mathbf{c}/\|\mathbf{c}\|$ and therefore $\mathbf{R}^\top \mathbf{c} = \|\mathbf{c}\|\,\mathbf{e}_1$. Put $\mathbf{U}_2 = \mathbf{I}_r \oplus \mathbf{R}$, which lies in $\mathrm{O}(n)$. Substituting $\mathbf{z} = \mathbf{U}_2 \mathbf{w}$ leaves the quadratic part alone, since

$$\begin{aligned} \mathbf{U}_2^\top \big(\mathrm{diag}(\lambda_1, \dots, \lambda_r) \oplus \mathbf{O}\big)\mathbf{U}_2 &= \mathrm{diag}(\lambda_1, \dots, \lambda_r) \oplus \mathbf{R}^\top \mathbf{O}\mathbf{R} \\ &= \mathrm{diag}(\lambda_1, \dots, \lambda_r) \oplus \mathbf{O}, \end{aligned}$$

and turns the linear part into $2\mathbf{c}^\top \mathbf{R}(w_{r+1}, \dots, w_n)^\top = 2\|\mathbf{c}\| w_{r+1}$. Finally translate w_{r+1} by $-c'/(2\|\mathbf{c}\|)$, which changes no quadratic term because the coefficient of w_{r+1}^2 is $\lambda_{r+1} = 0$. The result is form (b) with $\mu = \|\mathbf{c}\| > 0$.

Step 4: invariance. Let $g(\mathbf{y}) = \mathbf{U}\mathbf{y} + \mathbf{t}$ with $\mathbf{U} \in \mathrm{O}(n)$. By Proposition 22.5.1, $Q \circ g$ has small matrix $\mathbf{U}^\top \mathbf{A}\mathbf{U}$, which is similar to $\mathbf{A}$ and so has the same eigenvalues with multiplicity (Theorem 7.10.3). Its linear vector is $\mathbf{b}' = \mathbf{U}^\top(\mathbf{A}\mathbf{t} + \mathbf{b})$. For the central case, $g^{-1}(\mathbf{p})$ is a center of $Q \circ g$ whenever $\mathbf{p}$ is a center of Q, as noted at the end of Theorem 22.5.3, and the value there is $(Q \circ g)(g^{-1}(\mathbf{p})) = Q(\mathbf{p})$; so c_0 is unchanged. For the centerless case, $\mathrm{nul}(\mathbf{U}^\top \mathbf{A}\mathbf{U}) = \mathbf{U}^\top \mathrm{nul}(\mathbf{A})$, and $\mathbf{A}\mathbf{t} \in \mathrm{col}(\mathbf{A}) = \mathrm{nul}(\mathbf{A})^\perp$, so the component of $\mathbf{b}'$ in $\mathrm{nul}(\mathbf{U}^\top \mathbf{A}\mathbf{U})$ is $\mathbf{U}^\top$ applied to the component of $\mathbf{b}$ in $\mathrm{nul}(\mathbf{A})$; an orthogonal matrix preserves norms (Theorem 11.7.3), so μ is unchanged. This proves the theorem.

■

When $\mathbf{A}$ is invertible and $c_0 \neq 0$, dividing form (a) by $-c_0$ puts it as

$$\frac{y_1^2}{a_1^2} + \cdots + \frac{y_n^2}{a_n^2} = 1 \quad \text{with} \quad \frac{1}{a_i^2} = \frac{\lambda_i}{-c_0},$$

whenever all the ratios $\lambda_i/(-c_0)$ are positive, which is the ellipsoid case. The **semi-axes** $a_i = \sqrt{-c_0/\lambda_i}$ lie along the eigenvectors, and a large eigenvalue gives a short axis, exactly as in Example (Finding the axes of a conic).

Remark. That last sentence is the Rayleigh quotient in disguise. For a centered ellipsoid $\mathbf{x}^\top \mathbf{A}\mathbf{x} = 1$ with $\mathbf{A}$ positive definite, any point $\mathbf{x}$ on it has $R_{\mathbf{A}}(\mathbf{x}) = \mathbf{x}^\top \mathbf{A}\mathbf{x}/\|\mathbf{x}\|^2 = 1/\|\mathbf{x}\|^2$. By Proposition 17.1.1 the quotient attains its maximum $\lambda_{\max}$ at a top eigenvector, and since

$R_{\mathbf{A}}$ is unchanged by scaling while every nonzero vector scales to a point of the ellipsoid, the maximum over the ellipsoid is that same $\lambda_{\max}$, attained at the scaled top eigenvector; so $\|\mathbf{x}\|$ attains its **minimum** $1/\sqrt{\lambda_{\max}}$ there: the shortest semi-axis points along the eigenvector of the largest eigenvalue. The longest does the same with the smallest eigenvalue. Chapter 24 uses the same picture for a vibrating chain whose masses are all equal, where the eigenvalues are squared frequencies and the principal axes are the shapes in which the chain vibrates.

Example. One conic, two classifications

Over $\mathbb{R}$, consider

$$Q(x,y) = 5x^2 + 4xy + 5y^2 - 14x - 14y - 7.$$

Classify its zero set affinely, and then find its center, principal axes, and semi-axes.

Solution. The small and big matrices are

$$\mathbf{A} = \begin{pmatrix} 5 & 2 \\ 2 & 5 \end{pmatrix}, \qquad \widetilde{\mathbf{A}} = \begin{pmatrix} 5 & 2 & -7 \\ 2 & 5 & -7 \\ -7 & -7 & -7 \end{pmatrix},$$

the off-diagonal 2 being half of 4 and the -7 in the last column being half of -14.

Center. $\det \mathbf{A} = 21 \neq 0$, so by Proposition 22.5.2 (b) there is exactly one center, the solution of $\mathbf{Ap} = -\mathbf{b} = (7,7)$, namely $\mathbf{p} = (1,1)$. Then $c_0 = Q(1,1) = 5+4+5-14-14-7 = -21$.

Affine route. By Proposition 22.5.2 (c), translating to $\mathbf{p}$ gives $5X^2 + 4XY + 5Y^2 - 21$ with $X = x-1$, $Y = y-1$. Completing the square,

$$5X^2 + 4XY + 5Y^2 = 5\Big(X + \tfrac{2}{5}Y\Big)^2 + \tfrac{21}{5}Y^2,$$

since $5 - \frac{4}{5} = \frac{21}{5}$. So with $s = X + \frac{2}{5}Y$ and $t = Y$ the equation is $5s^2 + \frac{21}{5}t^2 = 21$, and the further scaling $w_1 = \sqrt{5/21}\,s$, $w_2 = t/\sqrt{5}$ gives $w_1^2 + w_2^2 = 1$. The inertia of $\mathbf{A}$ is $(2,0,0)$ and that of $\widetilde{\mathbf{A}}$ is $(2,1,0)$, since $\det \widetilde{\mathbf{A}} = (\det \mathbf{A})c_0 = -441 < 0$ forces one negative eigenvalue and $\mathbf{A}$ supplies two positive ones. Row 1 of the table: an **ellipse**.

Euclidean route. The characteristic polynomial of $\mathbf{A}$ is $x^2 - 10x + 21 = (x-3)(x-7)$, so the eigenvalues are 3 and 7 with unit eigenvectors

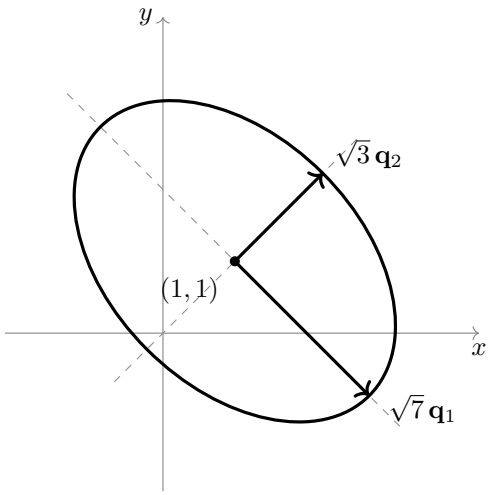
$$\mathbf{q}_1 = \tfrac{1}{\sqrt{2}}(1,-1), \qquad \mathbf{q}_2 = \tfrac{1}{\sqrt{2}}(1,1),$$

as $\mathbf{A}(1,-1) = (3,-3)$ and $\mathbf{A}(1,1) = (7,7)$. By Theorem 22.5.4 the motion $\mathbf{x} = \mathbf{Uy} + \mathbf{p}$ with $\mathbf{U} = (\mathbf{q}_1\ \mathbf{q}_2)$ turns Q into $3y_1^2 + 7y_2^2 - 21$, that is

$$\frac{y_1^2}{7} + \frac{y_2^2}{3} = 1.$$

So the ellipse is centered at $(1,1)$, with semi-axis $\sqrt{7}$ along $\mathbf{q}_1$ and semi-axis $\sqrt{3}$ along $\mathbf{q}_2$. Check the longer one: the point $(1,1) + \sqrt{7}\,\mathbf{q}_1 = (1+\sqrt{7/2},\ 1-\sqrt{7/2})$ has $X = -Y = \sqrt{7/2}$, and $5X^2 + 4XY + 5Y^2 = (5-4+5)\cdot\frac{7}{2} = 21$, as required.

What each route saw. Both say "ellipse". Only the Euclidean one says which ellipse: the affine route ended at a circle of radius 1, having first sheared the plane by $s = X + \frac{2}{5}Y$ and then stretched it by the unequal factors $\sqrt{5/21}$ and $1/\sqrt{5}$ to get there.



$5x^2 + 4xy + 5y^2 - 14x - 14y - 7 = 0$: center $(1, 1)$, axes along the eigenvectors of $\left(\begin{smallmatrix} 5 & 2 \\ 2 & 5 \end{smallmatrix}\right)$

The two classifications answer two different questions, and it is worth being blunt about which is which. The affine one asks what the quadric is *made of* — a curve, a pair of lines, a point, nothing — and its invariants are two inertias. The Euclidean one asks what the quadric *looks like*, and its invariants are the eigenvalues themselves. Section 10 asks a third question, the projective one, and the answer there is shorter than both.

22.5.7 Exercises

22.5.7.1 A. Check your understanding

Exercise. A1

(a) State Definition 22.5.1, including every clause on $\mathbf{A}$, and say what goes wrong if the clause "nonzero" is dropped.

(b) Write down the big matrix $\widetilde{\mathbf{A}}$ of $3x^2 - 2xy + y^2 + 4x - 2$ over $\mathbb{R}$.

(c) Name the three places in this section where $\operatorname{char} F \neq 2$ is used.

(d) True or false: two quadric polynomials with the same zero set are affinely equivalent. Justify your answer.

(e) A quadric on $\mathbb{R}^5$ has $\operatorname{rank} \mathbf{A} = 3$ and $\operatorname{rank} \widetilde{\mathbf{A}} = 5$. Which of the three families of Theorem 22.5.3 does it belong to, and is it central?

Solution.

(a) $Q(\mathbf{x}) = \mathbf{x}^{\top}\mathbf{A}\mathbf{x} + 2\mathbf{b}^{\top}\mathbf{x} + c$ with $\mathbf{A} \in M_n(F)$ symmetric and nonzero, $\mathbf{b} \in F^n$, $c \in F$, over a field of characteristic $\neq 2$. If $\mathbf{A} = \mathbf{O}$ were allowed, Q would be an affine function and its zero set a flat or all of F^n; nothing quadratic would remain, and Theorem 22.5.3, whose first step diagonalizes a small matrix of rank $r \geq 1$, would have nothing to work on.

(b) The coefficient of xy is -2, so $a_{12} = -1$; the coefficient of x is 4, so $b_1 = 2$; there is no y term, so $b_2 = 0$. Hence

$$\widetilde{\mathbf{A}} = \begin{pmatrix} 3 & -1 & 2 \\ -1 & 1 & 0 \\ 2 & 0 & -2 \end{pmatrix}.$$

(c) Writing a linear part $\sum \beta_i x_i$ as $2\mathbf{b}^\top \mathbf{x}$; recovering a **symmetric** $\mathbf{A}$ from the quadratic part, where the off-diagonal entries are halves (Theorem 14.3.5); and dividing by 4 in Eq. 22.5.1 to get the center equation $\mathbf{Ap} + \mathbf{b} = \mathbf{0}$.

(d) False. Over $\mathbb{R}$, $x^2 + y^2 + 1$ and $x^2 + 1$ both have empty zero set, but $\operatorname{rank} \mathbf{A}$ is 2 for the first and 1 for the second, and rank is an affine invariant by Proposition 22.5.1.

(e) $\operatorname{rank} \widetilde{\mathbf{A}} = \operatorname{rank} \mathbf{A} + 2$, so it is family (III), and family (III) is exactly the centerless case by Theorem 22.5.3.

22.5.7.2 B. Practice

Exercise. B1: Centers

For each of the following quadric polynomials over $\mathbb{R}$, determine the set of centers. Justify your answer.

(a) $2x^2 + 3y^2 - 4x + 6y + 1$.

(b) $x^2 - 4xy + 4y^2 + 2x - 4y + 5$.

(c) $x^2 - 4xy + 4y^2 + 2x + y$.

Solution. In each case solve $\mathbf{Ap} = -\mathbf{b}$ (Proposition 22.5.2 (a)).

(a) $\mathbf{A} = \operatorname{diag}(2, 3)$, $\mathbf{b} = (-2, 3)$. Since $\mathbf{A}$ is invertible the center is unique: $\mathbf{p} = -\mathbf{A}^{-1}\mathbf{b} = (1, -1)$.

(b) $\mathbf{A} = \left(\begin{smallmatrix} 1 & -2 \\ -2 & 4 \end{smallmatrix}\right)$, which has rank 1 with $\operatorname{col}(\mathbf{A}) = \operatorname{span}((1, -2))$, and $\mathbf{b} = (1, -2) \in \operatorname{col}(\mathbf{A})$. So centers exist, and they solve $p_1 - 2p_2 = -1$: the line $x - 2y + 1 = 0$. (Consistently, the polynomial is $(x - 2y + 1)^2 + 4$, whose zero set is empty but whose equation is symmetric about that line.)

(c) Same $\mathbf{A}$, but $\mathbf{b} = (1, \frac{1}{2})$. If $\mathbf{b}$ were in $\operatorname{col}(\mathbf{A}) = \operatorname{span}((1, -2))$ its second entry would be -2 times its first, and $\frac{1}{2} \neq -2$. So there is no center; this is a parabola.

Exercise. B2: Affine type

Classify the conic $x^2 + 6xy + y^2 - 4x - 4y + 3 = 0$ over $\mathbb{R}$: find $\operatorname{In} \mathbf{A}$ and $\operatorname{In} \widetilde{\mathbf{A}}$, name the row of the table, and exhibit the invertible affine substitution that reaches the normal form.

Solution. Here $\mathbf{A} = \left(\begin{smallmatrix} 1 & 3 \\ 3 & 1 \end{smallmatrix}\right)$ and $\mathbf{b} = (-2, -2)$, $c = 3$. The eigenvalues of $\mathbf{A}$ are 4 and -2, from $x^2 - 2x - 8 = (x - 4)(x + 2)$, so $\operatorname{In} \mathbf{A} = (1, 1, 0)$ by Theorem 14.5.3, and $\det \mathbf{A} = -8 \neq 0$.

The unique center solves $\mathbf{Ap} = (2, 2)$, giving $\mathbf{p} = (\frac{1}{2}, \frac{1}{2})$, and $c_0 = Q(\mathbf{p}) = \frac{1}{4} + \frac{3}{2} + \frac{1}{4} - 2 - 2 + 3 = 1$. Translating, the equation becomes $X^2 + 6XY + Y^2 + 1 = 0$ with $X = x - \frac{1}{2}$, $Y = y - \frac{1}{2}$. Completing the square, $X^2 + 6XY + Y^2 = (X + 3Y)^2 - 8Y^2$, so with $s = X + 3Y$, $t = Y$

the equation is $s^2 - 8t^2 = -1$. Multiplying by -1 and setting $w_1 = 2\sqrt{2}\,t$, $w_2 = s$ gives $w_1^2 - w_2^2 = 1$.

So $\widetilde{\mathbf{A}}$ is congruent to $\operatorname{diag}(1, -8, 1)$ up to the translation, giving $\operatorname{In}\widetilde{\mathbf{A}} = (2, 1, 0)$; up to the simultaneous exchange this is the invariant $((1, 1, 0), (1, 2, 0))$ of row 2. The conic is a **hyperbola**.

Exercise. B3: Principal axes

Find the center, principal axes, and semi-axes of $2x^2 - 4xy + 5y^2 - 4x + 2y - 7 = 0$ over $\mathbb{R}$.

Solution. $\mathbf{A} = \left(\begin{smallmatrix} 2 & -2 \\ -2 & 5 \end{smallmatrix}\right)$, $\mathbf{b} = (-2, 1)$, $c = -7$. Since $\det \mathbf{A} = 6 \neq 0$ the center solves $\mathbf{A}\mathbf{p} = (2, -1)$: from $2p_1 - 2p_2 = 2$ and $-2p_1 + 5p_2 = -1$ we get $p_2 = \frac{1}{3}$, $p_1 = \frac{4}{3}$. Then

$$c_0 = Q(\mathbf{p}) = \mathbf{b}^\top \mathbf{p} + c = (-2)\tfrac{4}{3} + 1 \cdot \tfrac{1}{3} - 7 = -\tfrac{28}{3},$$

using $Q(\mathbf{p}) = \mathbf{p}^\top(\mathbf{A}\mathbf{p} + \mathbf{b}) + \mathbf{b}^\top \mathbf{p} + c = \mathbf{b}^\top \mathbf{p} + c$, valid because $\mathbf{A}\mathbf{p} + \mathbf{b} = \mathbf{0}$.

The characteristic polynomial is $x^2 - 7x + 6 = (x-1)(x-6)$, so $\lambda = 1, 6$ with unit eigenvectors $\mathbf{q}_1 = \frac{1}{\sqrt{5}}(2, 1)$ and $\mathbf{q}_2 = \frac{1}{\sqrt{5}}(1, -2)$; indeed $\mathbf{A}(2, 1) = (2, 1)$ and $\mathbf{A}(1, -2) = (6, -12)$. By Theorem 22.5.4 the normal form is $y_1^2 + 6y_2^2 - \frac{28}{3} = 0$, that is

$$\frac{y_1^2}{28/3} + \frac{y_2^2}{14/9} = 1.$$

So the curve is an ellipse centered at $(\frac{4}{3}, \frac{1}{3})$, with semi-axis $2\sqrt{21}/3 = \sqrt{28/3}$ along $\mathbf{q}_1$ and semi-axis $\sqrt{14}/3$ along $\mathbf{q}_2$.

22.5.7.3 C. Going deeper

Exercise. C1: Which quadrics are determined by their points

Work over $\mathbb{R}$ with $n = 2$.

(a) Give two quadric polynomials with the same zero set that are **not** proportional.

(b) Prove that if Q is a quadric polynomial whose zero set contains three points of some line L, then Q vanishes on all of L.

(c) Deduce that the zero set of a quadric polynomial is never exactly a set of three non-collinear points.

Hint for (b): restrict Q to a parametrization of L.

Solution.

(a) Take $Q_1 = x^2 + y^2$ and $Q_2 = x^2 + 2y^2$. Both vanish exactly at $\mathbf{0}$, since a sum of squares of reals is zero only when each is, but $Q_2 \neq \lambda Q_1$ for any λ: comparing the coefficients of x^2 forces $\lambda = 1$, and then the coefficients of y^2 disagree.

(b) Write $L = \{\mathbf{p} + t\mathbf{v} : t \in \mathbb{R}\}$ with $\mathbf{v} \neq \mathbf{0}$, a flat of dimension 1 (Definition 22.2.1). Then

$$f(t) = Q(\mathbf{p} + t\mathbf{v}) = (\mathbf{v}^\top \mathbf{A}\mathbf{v})t^2 + 2(\mathbf{v}^\top(\mathbf{A}\mathbf{p} + \mathbf{b}))t + Q(\mathbf{p})$$

by the expansion in the proof of Proposition 22.5.2, so f is a polynomial of degree at most 2 in t. If Q vanishes at three distinct points of L, then f has three distinct roots. A nonzero polynomial of degree at most 2 over a field has at most 2 roots (Corollary 6.1.4), so f is the zero polynomial and Q vanishes on all of L.

(c) Suppose, for a contradiction, that the zero set of a quadric polynomial Q is exactly $\{P_1, P_2, P_3\}$ with the three points not collinear, that is affinely independent (Definition 18.2.1). Then (P_1, P_2, P_3) is an affine frame of $\mathbb{R}^2$ (Definition 22.2.4), so by Theorem 22.2.8 there is an affine map f carrying it to the frame $((0,0),(1,0),(0,1))$. That f is invertible: by Theorem 22.1.5 (b) its linear part sends $\overrightarrow{P_1P_2}$ to $\mathbf{e}_1$ and $\overrightarrow{P_1P_3}$ to $\mathbf{e}_2$, so it carries a basis to a basis and is invertible, whence f is bijective by Proposition 22.1.7. Substituting by f^{-1}, we may therefore assume the three points **are** $(0,0)$, $(1,0)$, $(0,1)$, since an invertible affine substitution carries the zero set bijectively to the zero set (Proposition 22.5.1).

Write $Q = ax^2 + 2hxy + by^2 + 2dx + 2ey + c$. Then $Q(0,0) = c = 0$, $Q(1,0) = a + 2d = 0$ and $Q(0,1) = b + 2e = 0$, so

$$Q = a\,x(x-1) + b\,y(y-1) + 2hxy.$$

Restricting to the line $y = 0$ gives $a\,x(x-1)$. If $a = 0$ this vanishes identically and the whole line lies in the zero set, which is impossible for a three-point set; so $a \neq 0$, and by the same argument on $x = 0$, $b \neq 0$.

Now fix $m \in \mathbb{R}$ with $m \neq 0$ and restrict to the line $y = mx$:

$$Q(t, mt) = t\big[(a + 2hm + bm^2)\,t - (a + bm)\big].$$

The only point of $\{P_1, P_2, P_3\}$ on that line is $(0,0)$, because $(1,0)$ needs $m = 0$ and $(0,1)$ is not of the form (t, mt). So $t = 0$ must be the only root. If $a + 2hm + bm^2 \neq 0$, the bracket vanishes at $t = (a + bm)/(a + 2hm + bm^2)$, and that must be 0, forcing $a + bm = 0$. Hence for every $m \neq 0$,

$$a + 2hm + bm^2 = 0 \qquad \text{or} \qquad a + bm = 0.$$

The first equation is a quadratic in m with leading coefficient $b \neq 0$, so it has at most 2 roots, and the second has exactly one since $b \neq 0$ (Corollary 6.1.4). That is at most three permitted values of m, while $\mathbb{R} \setminus \{0\}$ is infinite. Contradiction. Hence no quadric polynomial over $\mathbb{R}$ has exactly three noncollinear points as its zero set.

Exercise. C2: Over another field

(a) Over $\mathbb{C}$, prove that every quadric polynomial in n variables is equivalent, under an invertible affine substitution and a nonzero scalar, to one of $\sigma_r(\mathbf{y}) - 1$, $\sigma_r(\mathbf{y})$, or $\sigma_r(\mathbf{y}) - y_{r+1}$ with $r < n$, where $\sigma_r(\mathbf{y}) = y_1^2 + \cdots + y_r^2$. How many normal forms are there for $n = 2$?

(b) Over $\mathbb{F}_3$, how many points does the conic $x^2 + y^2 - 1 = 0$ have in $\mathbb{F}_3^2$? And $x^2 + y^2 = 0$?

Solution.

(a) Repeat the proof of Theorem 22.5.3, replacing Corollary 14.5.2 by Corollary 14.4.4, which says that over $\mathbb{C}$ a symmetric matrix of rank r is congruent to $\mathbf{I}_r \oplus \mathbf{O}_{n-r}$: rank alone is the congruence invariant. Step 1 gives $\sigma_r + 2\mathbf{b}_1^\top \mathbf{y} + c$; Step 2 is unchanged; and in Step 3 Case A with $c_2 \neq 0$ we substitute $z_i = \gamma w_i$ for $i \leq r$ with $\gamma^2 = -c_2$, possible because every complex number has a square root, and obtain $\sigma_r - 1$ after dividing by $-c_2$. No sign remains to normalize, and no p survives, so the invariants reduce to $(\operatorname{rank}\mathbf{A}, \operatorname{rank}\widetilde{\mathbf{A}}) = (r, r+1), (r, r), (r, r+2)$, which are pairwise distinct. For $n = 2$: $r = 1$ gives three forms and $r = 2$ gives two (family (III) needs $r < n$), so **five** in all, against nine over $\mathbb{R}$. The four forms that disappear are accounted for by inertia: over $\mathbb{C}$ the ellipse, the hyperbola, and the empty conic $x^2 + y^2 + 1$ all become $\sigma_2 - 1$; the point and the crossing lines both become σ_2; and $x^2 = 1$ and $x^2 = -1$ both become $\sigma_1 - 1$, the last two because $i^2 = -1$.

(b) The squares in $\mathbb{F}_3$ are $0^2 = 0$, $1^2 = 1$, $2^2 = 1$, so $x^2 \in \{0, 1\}$ and $x^2 = 1$ has two solutions while $x^2 = 0$ has one. For $x^2 + y^2 = 1$: either $x^2 = 1$ and $y^2 = 0$, giving $2 \cdot 1 = 2$ points, or $x^2 = 0$ and $y^2 = 1$, giving $1 \cdot 2 = 2$ points; and $1 + 1 = 2 \neq 1$ rules out the remaining combination. Total **4** points, namely $(\pm 1, 0)$ and $(0, \pm 1)$. For $x^2 + y^2 = 0$: the options are $0 + 0$ only, since $1 + 1 = 2 \neq 0$ and $1 + 0 = 1 \neq 0$. Total **1** point, the origin. So over $\mathbb{F}_3$ this "circle" behaves like a real circle in the first case and like the real point-conic in the second, because -1 is not a square in $\mathbb{F}_3$.

Exercise. C3: The center as a projection

Let $Q(\mathbf{x}) = \mathbf{x}^\top \mathbf{A}\mathbf{x} + 2\mathbf{b}^\top \mathbf{x} + c$ be a central quadric polynomial on $\mathbb{R}^n$ with $r = \operatorname{rank}\mathbf{A}$.

(a) Prove that the set of centers is a flat of dimension $n - r$ (Definition 22.2.1).

(b) Prove that the set of centers is contained in the zero set of Q if and only if $Q(\mathbf{p}) = 0$ for one, hence every, center $\mathbf{p}$.

(c) Give an example on $\mathbb{R}^3$ with a line of centers, none of which lies on the quadric.

Solution.

(a) By Proposition 22.5.2 (a) the centers are the solutions of $\mathbf{A}\mathbf{p} = -\mathbf{b}$, a consistent linear system. By Theorem 3.1.3 its solution set is $\mathbf{p}_0 + \operatorname{nul}(\mathbf{A})$ for any particular solution $\mathbf{p}_0$, a coset of $\operatorname{nul}(\mathbf{A})$, hence a flat with direction space $\operatorname{nul}(\mathbf{A})$. Its dimension is $\dim \operatorname{nul}(\mathbf{A}) = n - \operatorname{rank}\mathbf{A} = n - r$ by the Rank-Nullity Theorem (Theorem 2.3.1).

(b) By Proposition 22.5.2 (c) the value $Q(\mathbf{p})$ is the same at every center, so "for one" and "for every" agree; and the set of centers lies in the zero set exactly when that common value is 0.

(c) Take $Q(x, y, z) = x^2 + y^2 - 1$ on $\mathbb{R}^3$. Here $\mathbf{A} = \operatorname{diag}(1, 1, 0)$ has rank 2, $\mathbf{b} = \mathbf{0} \in \operatorname{col}(\mathbf{A})$, and the centers are the solutions of $\mathbf{A}\mathbf{p} = \mathbf{0}$, namely the z-axis, a line as (a) predicts with $n - r = 1$. The common value is $Q(0, 0, z) = -1 \neq 0$, so no center lies on the quadric, which is the cylinder of radius 1 about the z-axis.

22.6 Projective Space

The affine geometry of the first two sections has one flaw, and it is visible in the statement of Theorem 22.2.4: the formula for the dimension of a join has two cases, because two flats may fail to meet. This section builds a geometry in which that never happens. The construction is entirely linear — a projective space is the set of lines through the origin of a vector space — and the payoff is a dimension formula with no exceptional case at all. Everything here works over an arbitrary field F, and neither analysis nor topology enters.

22.6.1 Two lines that do not meet

In the affine plane $\mathbb{A}^2(\mathbb{R})$, take

$$L_1 = \{(y_1, y_2) : y_2 - y_1 = 0\}, \qquad L_2 = \{(y_1, y_2) : y_2 - y_1 = 1\}.$$

A point in both would satisfy $0 = 1$, so $L_1 \cap L_2 = \emptyset$. Both are flats of dimension 1 with the same direction space $\operatorname{span}((1,1))$, and their join is the whole plane. So Theorem 22.2.4 (c) is in its second case: $\dim(L_1 \vee L_2) = 1 + 1 - 1 + 1 = 2$, with the extra $+1$ paid for by the failure to meet, and there is nothing to substitute for $\dim(L_1 \cap L_2)$, because the empty set has no dimension.

The two lines do have something in common: a direction. They are going the same way, and the natural thing to say is that they meet "at infinity". The whole of this section is the work of turning that phrase into mathematics without losing linearity. The repair, in one sentence: *stop working with points of a vector space and start working with its lines through the origin.*

22.6.2 The projective space of a vector space

Let V be a vector space over F. On $V \setminus \{0\}$ define

$$x \sim y \quad \text{if and only if} \quad y = \lambda x \text{ for some } \lambda \in F,\ \lambda \neq 0.$$

This is an equivalence relation (Definition 0.5.2): it is reflexive with $\lambda = 1$; symmetric, because $y = \lambda x$ with $\lambda \neq 0$ gives $x = \lambda^{-1} y$; and transitive, because $y = \lambda x$ and $z = \mu y$ give $z = (\mu\lambda)x$ with $\mu\lambda \neq 0$, a field having no zero divisors. Note that 0 had to be removed: it is equivalent only to itself, and keeping it would put a single extra point into every picture with no geometric job.

A point of projective space is a line through the origin, with the origin itself thrown away.

Definition 22.6.1. Projective Space

Let V be a finite-dimensional vector space over a field F with $\dim V = n+1$. The **projective space** of V is the quotient set (Definition 0.5.5)

$$\mathbb{P}(V) = (V \setminus \{0\})/\!\sim,$$

whose elements are called **points**. The point determined by $x \neq 0$ is written $[x]$. The **dimension** of $\mathbb{P}(V)$ is

$$\dim \mathbb{P}(V) = \dim V - 1 = n,$$

so that $\mathbb{P}(\{0\}) = \emptyset$ has dimension -1. We write $\mathbb{P}^n(F) = \mathbb{P}(F^{n+1})$ for **projective n-space over F**. A projective space of dimension 1 is a **projective line**, and one of dimension 2 a

projective plane.

Two clauses need comment. The shift by one in the dimension is not a decoration: it is what makes a projective line behave like a line and a projective plane like a plane, as Theorem 22.6.2 is about to confirm. And the convention $\dim \emptyset = -1$ looks like a dodge, but it is forced by the same shift — the empty set is $\mathbb{P}(\{0\})$ and $\dim\{0\} = 0$ — and it will do real work in Theorem 22.6.6.

Proposition 22.6.1. Points Are Lines Through the Origin

Let V be a finite-dimensional vector space over F. The map $[x] \mapsto \operatorname{span}(x)$ is a bijection from $\mathbb{P}(V)$ onto the set of 1-dimensional subspaces of V.

Proof. *Well defined.* If $[x] = [y]$ then $y = \lambda x$ with $\lambda \neq 0$, so $\operatorname{span}(y) = \operatorname{span}(x)$ because each of x, y is a scalar multiple of the other. Also $x \neq 0$, so $\operatorname{span}(x)$ has dimension 1.

Injective. If $\operatorname{span}(x) = \operatorname{span}(y)$ with $x, y \neq 0$, then $y = \lambda x$ for some $\lambda \in F$, and $\lambda \neq 0$ since $y \neq 0$. Hence $[x] = [y]$.

Surjective. A 1-dimensional subspace has a basis (x) with $x \neq 0$, and then it is $\operatorname{span}(x)$, the image of $[x]$. This proves the proposition.

■

So a point of $\mathbb{P}^2(F)$ is a line through the origin of F^3, and a "line" of the projective plane will turn out to be a plane through the origin of F^3. Every statement below can be read twice: once about $\mathbb{P}(V)$ and once about subspaces of V.

Definition 22.6.2. Homogeneous Coordinates

Let $\mathbf{x} = (x_0, x_1, \dots, x_n) \in F^{n+1}$ be nonzero. The **homogeneous coordinates** of the point $[\mathbf{x}] \in \mathbb{P}^n(F)$ are written

$$[\mathbf{x}] = [x_0 : x_1 : \cdots : x_n],$$

the colons recording that only the ratios are determined: for every $\lambda \neq 0$,

$$[x_0 : x_1 : \cdots : x_n] = [\lambda x_0 : \lambda x_1 : \cdots : \lambda x_n],$$

and not all of $x_0, \dots, x_n$ are zero.

Coordinates run from 0. In F^{n+1} we index the entries and the standard basis from 0, writing $\mathbf{x} = (x_0, x_1, \dots, x_n)$ and $\mathbf{e}_0, \dots, \mathbf{e}_n$, and every projective section of this chapter does the same. Subscripts start at 1 everywhere else in the book; the exception pays for itself here, because it keeps the n of $\mathbb{P}^n(F)$ equal to the dimension while the coordinates number $n + 1$.

Examples.

- $\mathbb{P}^0(F) = \mathbb{P}(F^1)$ has exactly one point, $[1]$, since every nonzero scalar is a nonzero multiple of 1. Its dimension is 0, as a single point should have.
- $\mathbb{P}^1(F)$ consists of the points $[1 : a]$ for $a \in F$, all distinct, together with the single point $[0 : 1]$. Indeed if $x_0 \neq 0$ then $[x_0 : x_1] = [1 : x_1/x_0]$, and if $x_0 = 0$ then $x_1 \neq 0$ and $[0 : x_1] = [0 : 1]$.
- $\mathbb{P}^1(\mathbb{F}_2) = \{[1 : 0], [1 : 1], [0 : 1]\}$ has three points. The three lines through the origin of $\mathbb{F}_2^2$

are $\mathrm{span}((1,0))$, $\mathrm{span}((1,1))$, $\mathrm{span}((0,1))$, matching Proposition 22.6.1.
- The degenerate case $\mathbb{P}(\{0\}) = \emptyset$. It is worth keeping, because intersections produce it.

Nonexample by minimal change. Over $\mathbb{R}$, replace the relation by $x \approx y$ when $y = \lambda x$ with $\lambda > 0$. This is still an equivalence relation, but the quotient is not $\mathbb{P}(V)$: every projective point $[x]$ splits into the two classes of x and $-x$. The clause that fails is the one allowing **every** nonzero λ, and what is lost is precisely the identification of a line with itself traversed backwards.

Warning. **A single homogeneous coordinate means nothing.** The expression "$x_1 = 3$" is not a statement about a point of $\mathbb{P}^n(F)$, since $[1:3:0] = [2:6:0]$ and the second coordinate changed. Only two kinds of statement survive scaling: a ratio such as x_1/x_0 where $x_0 \neq 0$, and the vanishing or nonvanishing of a **homogeneous** expression, such as $x_1 = 0$ or $x_0x_2 - x_1^2 = 0$. Whenever a formula below names a coordinate, check that it survives replacing $\mathbf{x}$ by $\lambda\mathbf{x}$.

22.6.3 The affine chart, and where the missing points live

Here is the theorem that makes the construction worth doing: projective space is an affine space plus a smaller projective space.

Theorem 22.6.2. The Standard Affine Chart

Let $n \geq 1$ and let

$$U_0 = \{[x_0 : \cdots : x_n] \in \mathbb{P}^n(F) : x_0 \neq 0\},$$
$$H_\infty = \{[x_0 : \cdots : x_n] \in \mathbb{P}^n(F) : x_0 = 0\}.$$

Then:

(a) both conditions are independent of the representative, and $\mathbb{P}^n(F) = U_0 \sqcup H_\infty$, a disjoint union;

(b) the map $\iota\colon F^n \to U_0$, $\iota(a_1, \ldots, a_n) = [1 : a_1 : \cdots : a_n]$, is a bijection, with inverse $[x_0 : \cdots : x_n] \mapsto (x_1/x_0, \ldots, x_n/x_0)$;

(c) $H_\infty = \mathbb{P}(W)$ with $W = \{\mathbf{x} \in F^{n+1} : x_0 = 0\}$, so H_∞ has dimension $n-1$, and $[0 : x_1 : \cdots : x_n] \mapsto [x_1 : \cdots : x_n]$ is a bijection $H_\infty \to \mathbb{P}^{n-1}(F)$.

U_0 is the **standard affine chart** and H_∞ the **hyperplane at infinity**.

Idea. Scaling multiplies x_0 by a nonzero number, so it cannot create or destroy a zero there: that single observation makes both halves well defined. On the half where $x_0 \neq 0$ we may **normalize** x_0 to 1, and once $x_0 = 1$ the remaining coordinates are no longer ambiguous, which is why the affine points come back one for one. On the other half x_0 carries no information at all, so it may be deleted, and what is left is one dimension smaller.

Proof.

(a) If $\mathbf{y} = \lambda\mathbf{x}$ with $\lambda \neq 0$ then $y_0 = \lambda x_0$, and a product of two elements of a field is zero only if one of them is, so $y_0 = 0$ exactly when $x_0 = 0$. Hence both defining conditions

depend only on the point. Every point satisfies exactly one of them, so the union is disjoint and exhausts $\mathbb{P}^n(F)$.

(b) *Well defined:* $(1, a_1, \dots, a_n) \neq \mathbf{0}$, and its zeroth entry is $1 \neq 0$, so $\iota(\mathbf{a}) \in U_0$.
Injective: if $[1 : \mathbf{a}] = [1 : \mathbf{a}']$ then $(1, \mathbf{a}') = \lambda(1, \mathbf{a})$ for some $\lambda \neq 0$; comparing zeroth entries gives $\lambda = 1$, hence $\mathbf{a}' = \mathbf{a}$.
Surjective: let $[\mathbf{x}] \in U_0$, so $x_0 \neq 0$ and x_0^{-1} exists. Then $x_0^{-1}\mathbf{x} = (1, x_1/x_0, \dots, x_n/x_0)$, so $[\mathbf{x}] = \iota(x_1/x_0, \dots, x_n/x_0)$. The displayed formula for the inverse is exactly this computation, and it is independent of the representative because replacing $\mathbf{x}$ by $\lambda\mathbf{x}$ multiplies numerator and denominator of each ratio by λ.

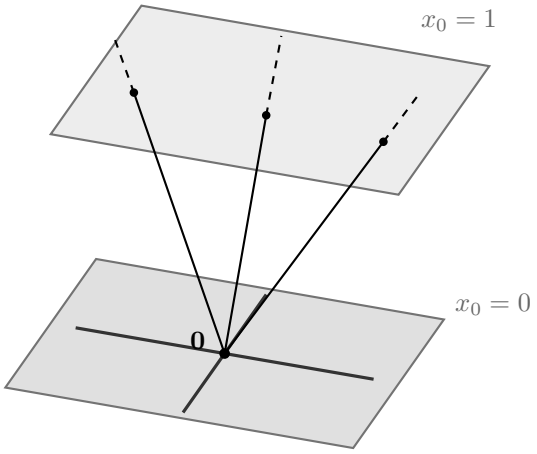
(c) W is a subspace of F^{n+1}, being the kernel of the linear functional $\mathbf{x} \mapsto x_0$, and $\dim W = n$ by the Rank-Nullity Theorem (Theorem 2.3.1), the functional being nonzero and hence surjective onto F (Proposition 5.1.2). A nonzero $\mathbf{x}$ has $x_0 = 0$ exactly when $\mathbf{x} \in W \setminus \{\mathbf{0}\}$, so $H_\infty = \mathbb{P}(W)$ and $\dim H_\infty = n - 1$ by Definition 22.6.1. The stated map is well defined, because deleting the zeroth coordinate turns $\lambda\mathbf{x}$ into λ times the deletion and cannot produce $\mathbf{0}$ from a nonzero $\mathbf{x}$ with $x_0 = 0$; it is injective and surjective because inserting a zeroth coordinate 0 inverts it. This proves the theorem.

■

For $n = 2$ the statement reads: **the projective plane is the affine plane together with a projective line.** Writing it out, every point of $\mathbb{P}^2(F)$ is exactly one of

$$[1 : a : b] \ \ (a, b \in F), \qquad [0 : 1 : m] \ \ (m \in F), \qquad [0 : 0 : 1],$$

the first family being the affine plane through ι, and the last two being the line at infinity $H_\infty \cong \mathbb{P}^1(F)$, which has $|F| + 1$ points when F is finite. The label m is chosen deliberately: the next proposition shows $[0 : 1 : m]$ is the point where all affine lines of slope m meet.



each line through $\mathbf{0}$ is one point of $\mathbb{P}^2(F)$:
those meeting the plane $x_0 = 1$ give the affine chart,
those lying inside the plane $x_0 = 0$ give the line at infinity

Proposition 22.6.3. Completing an Affine Line

Let $a_1, a_2 \in F$ be not both zero, let $b \in F$, and let

$$L = \{(y_1, y_2) \in F^2 : a_1 y_1 + a_2 y_2 = b\},$$

a flat of dimension 1 in $\mathbb{A}^2$ by Theorem 22.2.10, the single row (a_1, a_2) being nonzero, with

direction space $\operatorname{span}((-a_2, a_1))$. Let $\varphi(x_0, x_1, x_2) = a_1x_1 + a_2x_2 - bx_0$ and $\bar{L} = \mathbb{P}(\ker \varphi) \subseteq \mathbb{P}^2(F)$. Then:

(a) $\bar{L}$ is a projective line;

(b) $\iota(L) = \bar{L} \cap U_0$, so $\bar{L}$ contains a faithful copy of L;

(c) $\bar{L} \setminus U_0$ is the single point $[0 : -a_2 : a_1]$;

(d) two such lines L, L' have the same direction space if and only if $\bar{L}$ and $\bar{L}'$ meet H_∞ in the same point.

Proof.

(a) φ is linear and nonzero, since $\varphi(0, 1, 0) = a_1$ and $\varphi(0, 0, 1) = a_2$ are not both zero. So φ is surjective onto F (Proposition 5.1.2) and $\dim \ker \varphi = 3 - 1 = 2$ by Theorem 2.3.1. Hence $\bar{L} = \mathbb{P}(\ker \varphi)$ has dimension 1.

(b) A point of U_0 is $\iota(y_1, y_2) = [1 : y_1 : y_2]$ by Theorem 22.6.2 (b), and $\varphi(1, y_1, y_2) = a_1y_1 + a_2y_2 - b$. This vanishes exactly when $(y_1, y_2) \in L$. Since φ is homogeneous of degree 1, the vanishing does not depend on the representative.

(c) A point of $\bar{L} \setminus U_0$ has $x_0 = 0$ and $a_1x_1 + a_2x_2 = 0$ with $(x_1, x_2) \neq (0, 0)$. The solution set of that single nontrivial equation in F^2 is $\operatorname{span}((-a_2, a_1))$, of dimension 1, so there is exactly one such projective point, namely $[0 : -a_2 : a_1]$.

(d) By (c) the point in question is $[0 : -a_2 : a_1]$, which by Proposition 22.6.1 determines and is determined by $\operatorname{span}((-a_2, a_1))$, the direction space of L. This proves the proposition.

■

Now return to the hook. The lines $L_1 : y_2 - y_1 = 0$ and $L_2 : y_2 - y_1 = 1$ have $(a_1, a_2) = (-1, 1)$ and $b = 0$ respectively $b = 1$, so

$$\bar{L}_1 : x_2 - x_1 = 0, \qquad \bar{L}_2 : x_2 - x_1 - x_0 = 0.$$

Subtracting the two equations gives $x_0 = 0$, and then $x_2 = x_1$ with $(x_1, x_2) \neq (0, 0)$. So

$$\bar{L}_1 \cap \bar{L}_2 = \{[0 : 1 : 1]\},$$

exactly one point, and it lies on the line at infinity. By Proposition 22.6.3 (c) that point is $[0 : -a_2 : a_1] = [0 : -1 : -1] = [0 : 1 : 1]$ for both lines, which is their common direction $\operatorname{span}((1, 1))$ read as a point. Two parallel lines meet at their shared direction, and nowhere else.

Quick check. In $\mathbb{P}^2(\mathbb{R})$, which point of the line at infinity do all the **vertical** affine lines $y_1 = c$ pass through, and which point do all lines of slope m pass through?

Solution. A vertical line $y_1 = c$ has $(a_1, a_2) = (1, 0)$, so by Proposition 22.6.3 (c) its point at infinity is $[0 : -a_2 : a_1] = [0 : 0 : 1]$. A line $y_2 = my_1 + k$, that is $my_1 - y_2 = -k$, has $(a_1, a_2) = (m, -1)$ and point at infinity $[0 : 1 : m]$. So the points of H_∞ are $[0 : 1 : m]$ for each slope $m \in \mathbb{R}$, plus $[0 : 0 : 1]$ for the vertical direction: one point per direction, and the vertical direction is no longer a special case.

22.6.4 Projective subspaces, and a dimension formula with one case

Definition 22.6.3. Projective Subspace

Let V be a finite-dimensional vector space over F. A **projective subspace** of $\mathbb{P}(V)$ is a subset of the form

$$\mathbb{P}(U) = \{[x] : x \in U \setminus \{0\}\}$$

for a subspace U of V. Its **dimension** is $\dim U - 1$. A projective subspace of dimension 0 is a **point**, of dimension 1 a **line**, of dimension 2 a **plane**, and of dimension $\dim \mathbb{P}(V) - 1$ a **hyperplane**.

The dimension is only well defined if U can be recovered from $\mathbb{P}(U)$. It can, and the reason is worth isolating.

Lemma 22.6.4. A Projective Subspace Remembers Its Subspace

Let U and W be subspaces of V. Then $\mathbb{P}(U) \subseteq \mathbb{P}(W)$ if and only if $U \subseteq W$; in particular $\mathbb{P}(U) = \mathbb{P}(W)$ forces $U = W$.

Proof. ($\Leftarrow$) Immediate from the definition.

($\Rightarrow$) Suppose $\mathbb{P}(U) \subseteq \mathbb{P}(W)$ and let $x \in U$. If $x = 0$ then $x \in W$. Otherwise $[x] \in \mathbb{P}(U) \subseteq \mathbb{P}(W)$, so $[x] = [w]$ for some $w \in W \setminus \{0\}$, that is $x = \lambda w$ with $\lambda \neq 0$; hence $x \in W$, W being a subspace. So $U \subseteq W$.

For the last claim, apply the equivalence in both directions and use that $\subseteq$ is antisymmetric. This proves the lemma.

■

Proposition 22.6.5. Meet and Join

Let $S = \mathbb{P}(U)$ and $T = \mathbb{P}(W)$ be projective subspaces of $\mathbb{P}(V)$. Then:

(a) $S \cap T = \mathbb{P}(U \cap W)$, so an intersection of projective subspaces is a projective subspace;

(b) among all projective subspaces containing $S \cup T$ there is a smallest one, namely $\mathbb{P}(U + W)$.

We write $S \vee T = \mathbb{P}(U + W)$ and call it the **join** of S and T.

Proof.

(a) Let $x \neq 0$. By Lemma 22.6.4 applied to $\operatorname{span}(x)$, we have $[x] \in \mathbb{P}(U)$ exactly when $x \in U$. Hence $[x] \in S \cap T$ exactly when $x \in U$ and $x \in W$, that is $x \in (U \cap W) \setminus \{0\}$, which says $[x] \in \mathbb{P}(U \cap W)$.

(b) $U \subseteq U+W$ and $W \subseteq U+W$, so $\mathbb{P}(U+W)$ contains $S \cup T$ by Lemma 22.6.4. Conversely,

if $\mathbb{P}(Y) \supseteq S \cup T$ for a subspace Y, then the same lemma gives $U \subseteq Y$ and $W \subseteq Y$, hence $U + W \subseteq Y$ because $U + W$ is the smallest subspace containing both (Theorem 1.7.1), and so $\mathbb{P}(U + W) \subseteq \mathbb{P}(Y)$. This proves the proposition.

■

Theorem 22.6.6. Projective Dimension Formula

Let S and T be projective subspaces of a finite-dimensional projective space. Then

$$\dim(S \vee T) + \dim(S \cap T) = \dim S + \dim T. \tag{22.6.1}$$

There is **no exceptional case**: the identity holds for all S and T, including when $S \cap T = \emptyset$, where $\dim(S \cap T) = -1$.

Idea. There is nothing to invent. Write S and T as $\mathbb{P}(U)$ and $\mathbb{P}(W)$, apply Chapter 1's dimension formula to U and W, and subtract 1 four times. The only thing that needs watching is the bookkeeping of those four -1's, and the reason the formula has no second case is that two of them cancel against the other two even when $U \cap W = \{0\}$ — which is precisely the situation the affine version could not handle.

Proof. By Definition 22.6.3 write $S = \mathbb{P}(U)$ and $T = \mathbb{P}(W)$; the subspaces U and W are uniquely determined by Lemma 22.6.4. By Proposition 22.6.5, $S \vee T = \mathbb{P}(U + W)$ and $S \cap T = \mathbb{P}(U \cap W)$. Hence, using Definition 22.6.3 four times and Theorem 1.7.3 once,

$$\begin{aligned} \dim(S \vee T) + \dim(S \cap T) &= \big(\dim(U + W) - 1\big) + \big(\dim(U \cap W) - 1\big) \\ &= \big(\dim U + \dim W\big) - 2 \\ &= \big(\dim U - 1\big) + \big(\dim W - 1\big) \\ &= \dim S + \dim T. \end{aligned}$$

The middle step is Theorem 1.7.3, which holds for **all** subspaces U, W with no hypothesis beyond finite dimension. In particular when $U \cap W = \{0\}$ the second term on the first line is $0 - 1 = -1$, and the computation is unaffected. This proves the theorem.

■

Compare the two formulas side by side. For flats A, B of an affine space, Theorem 22.2.4 (c) gives

$$\dim(A \vee B) = \begin{cases} \dim A + \dim B - \dim(A \cap B), & A \cap B \neq \emptyset, \\ \dim A + \dim B - \dim(\vec{A} \cap \vec{B}) + 1, & A \cap B = \emptyset, \end{cases}$$

where $\vec{A}, \vec{B}$ are the direction spaces. Two cases, and in the second the intersection has dropped out of the formula entirely, replaced by an intersection of direction spaces and an apologetic $+1$. Eq. 22.6.1 has one case and needs no directions. That is the whole reason for building projective space, and everything the rest of this chapter does with it rests on this one line.

Corollary 22.6.7. Two Lines in a Projective Plane Meet

Let P be a projective plane, that is $P = \mathbb{P}(V)$ with $\dim V = 3$. Then any two distinct lines of P meet in exactly one point.

Proof. Let $S = \mathbb{P}(U)$ and $T = \mathbb{P}(W)$ be distinct lines, so $\dim U = \dim W = 2$ and $U \neq W$ by Lemma 22.6.4. Then $U + W$ contains U properly: otherwise $W \subseteq U$, and two subspaces of the same dimension with one inside the other are equal (Theorem 1.6.8 applied inside U), contradicting $U \neq W$. So $\dim(U + W) \geq 3$, and $\dim(U + W) \leq \dim V = 3$, whence $\dim(U + W) = 3$ and $\dim(S \vee T) = 2$.

By Theorem 22.6.6,

$$\dim(S \cap T) = \dim S + \dim T - \dim(S \vee T) = 1 + 1 - 2 = 0,$$

and a projective subspace of dimension 0 is $\mathbb{P}(Y)$ with $\dim Y = 1$, which by Proposition 22.6.1 is a single point. This proves the corollary.

■

Proposition 22.6.8. Two Points Determine a Line

Let $\mathbb{P}(V)$ be a projective space of dimension at least 1. Any two distinct points of $\mathbb{P}(V)$ lie on exactly one line.

Proof. Let $[x] \neq [y]$. Then x and y are linearly independent: a dependence relation with $x, y \neq 0$ would give $y = \lambda x$ with $\lambda \neq 0$, hence $[x] = [y]$. So $U = \operatorname{span}(x, y)$ has dimension 2 and $\mathbb{P}(U)$ is a line containing both points.

If $\mathbb{P}(Y)$ is any line containing both, then $x, y \in Y$ by Lemma 22.6.4, so $U \subseteq Y$; and $\dim Y = 2 = \dim U$, so $U = Y$ by Theorem 1.6.8. This proves the proposition.

■

The two statements just proved are each other's mirror images: "two distinct points lie on a unique line" and "two distinct lines meet in a unique point". Section 8 will show that this mirroring is not a coincidence.

Remark. The set $\mathbb{P}^2(\mathbb{R})$ is also a famous topological object, the real projective plane, and readers who have met it will be expecting words like "compact" and "nonorientable". This chapter does not use them. Nothing here is topological; every argument has been a statement about subspaces of a vector space, valid over any field, including finite ones where topology has nothing to say.

22.6.5 Counting points

Over a finite field the whole space can be counted, and the count is a useful check on the constructions above.

Proposition 22.6.9. The Number of Points of a Finite Projective Space

Let F be a finite field with q elements and let $n \geq 0$. Then

$$|\mathbb{P}^n(F)| = 1 + q + q^2 + \cdots + q^n = \frac{q^{n+1} - 1}{q - 1}.$$

More generally a projective subspace of dimension d has $(q^{d+1} - 1)/(q - 1)$ points.

Idea. Count the nonzero vectors and divide by the size of one equivalence class. Every class has the same size, $q - 1$, because scaling a fixed nonzero vector by distinct scalars gives distinct vectors. The alternative count, chart by chart, gives the geometric sum directly, and the two agreeing is a check on Theorem 22.6.2.

Proof. A vector of F^{n+1} is a list of $n+1$ entries, each chosen from F independently, so there are q^{n+1} of them and $q^{n+1} - 1$ nonzero ones.

Fix $\mathbf{x} \neq \mathbf{0}$. Its equivalence class is $\{\lambda\mathbf{x} : \lambda \in F, \lambda \neq 0\}$, and the map $\lambda \mapsto \lambda\mathbf{x}$ is injective there: if $\lambda\mathbf{x} = \mu\mathbf{x}$ then $(\lambda - \mu)\mathbf{x} = \mathbf{0}$, and $\mathbf{x} \neq \mathbf{0}$ forces $\lambda = \mu$. So every class has exactly $q - 1$ elements.

The classes partition $F^{n+1} \setminus \{\mathbf{0}\}$ (Theorem 0.5.2), so the number of classes is $(q^{n+1} - 1)/(q - 1)$. That this equals $1 + q + \cdots + q^n$ follows from

$$\begin{aligned}(q - 1)(1 + q + \cdots + q^n) &= (q + q^2 + \cdots + q^{n+1}) - (1 + q + \cdots + q^n) \\ &= q^{n+1} - 1,\end{aligned}$$

the middle terms canceling in pairs. For a projective subspace $\mathbb{P}(U)$ of dimension d, the space U has dimension $d+1$, hence q^{d+1} elements, since a choice of basis puts it in bijection with F^{d+1} (Corollary 2.5.5); the same argument applies. This proves the proposition. ∎

The geometric sum is Theorem 22.6.2 in disguise: the chart contributes q^n points and the hyperplane at infinity contributes $|\mathbb{P}^{n-1}(F)|$, so $|\mathbb{P}^n| = q^n + |\mathbb{P}^{n-1}|$, and unwinding that recursion from $|\mathbb{P}^0| = 1$ gives $1 + q + \cdots + q^n$.

Example. The Fano plane

Describe $\mathbb{P}^2(\mathbb{F}_2)$ completely: its points, its lines, how many points lie on each line, and how many lines pass through each point. Verify Corollary 22.6.7 and Proposition 22.6.8 by direct count.

Solution. By Proposition 22.6.9 there are $1+2+4=7$ points. Since $\mathbb{F}_2^\times = \{1\}$, each point has a **unique** nonzero representative, so the points are simply the seven nonzero vectors of $\mathbb{F}_2^3$. Number the point $[x_0 : x_1 : x_2]$ by $4x_0 + 2x_1 + x_2$:

$$1 = [0{:}0{:}1], \quad 2 = [0{:}1{:}0], \quad 3 = [0{:}1{:}1], \quad 4 = [1{:}0{:}0],$$
$$5 = [1{:}0{:}1], \quad 6 = [1{:}1{:}0], \quad 7 = [1{:}1{:}1].$$

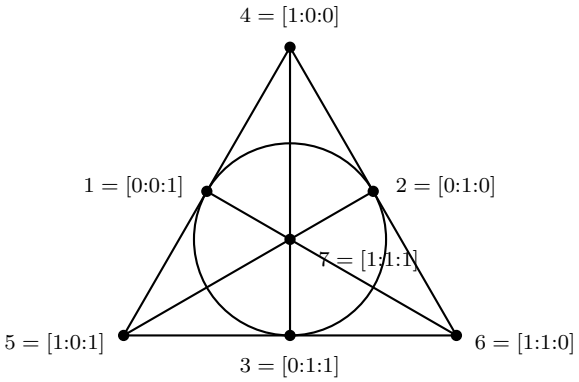
A line is $\mathbb{P}(U)$ with $\dim U = 2$, and by Proposition 22.6.9 it has $(2^2-1)/(2-1) = 3$ points. Since $U = \{\mathbf{0}, \mathbf{u}, \mathbf{w}, \mathbf{u}+\mathbf{w}\}$ for any two distinct nonzero $\mathbf{u}, \mathbf{w} \in U$, the lines are exactly the triples $\{\mathbf{u}, \mathbf{w}, \mathbf{u}+\mathbf{w}\}$ of nonzero vectors. There are seven:

$$\{1,2,3\} : x_0 = 0, \qquad \{1,4,5\} : x_1 = 0,$$
$$\{2,4,6\} : x_2 = 0, \qquad \{1,6,7\} : x_0 + x_1 = 0,$$
$$\{2,5,7\} : x_0 + x_2 = 0, \qquad \{3,4,7\} : x_1 + x_2 = 0,$$
$$\{3,5,6\} : x_0 + x_1 + x_2 = 0.$$

Each is the kernel of one of the seven nonzero linear functionals on $\mathbb{F}_2^3$, and different functionals have different kernels here because a functional is determined up to a nonzero scalar by its kernel — if $\ker\varphi = \ker\psi = U$ with $\dim U = 2$, then φ and ψ both vanish on U and so lie in the annihilator U^0 (Definition 5.2.1), which has dimension $3-2=1$ by Theorem 5.2.2 — and the only nonzero scalar here is 1. So there are exactly seven lines.

Incidences. Each of the seven labels appears in exactly three of the seven displayed triples, so every point lies on exactly 3 lines. The total is consistent: counting incident pairs (point, line through it) by lines gives $7 \cdot 3 = 21$, and counting them by points gives $7 \cdot 3 = 21$ as well.

The two theorems. There are $\binom{7}{2} = 21$ pairs of points, and each of the 7 lines contains $\binom{3}{2} = 3$ pairs, for 21 pairs in total; since no pair can be covered twice without two lines sharing two points — which Proposition 22.6.8 forbids — each pair is covered exactly once. Dually there are $\binom{7}{2} = 21$ pairs of lines, and each of the 7 points lies on 3 lines and so accounts for $\binom{3}{2} = 3$ of those pairs, again 21 in total; no pair of lines is accounted for twice, since two points on both lines would again contradict Proposition 22.6.8. So every pair of lines meets in exactly one point, as Corollary 22.6.7 says.



the Fano plane $\mathbb{P}^2(\mathbb{F}_2)$: 7 points, 7 lines
(three sides, three medians, one circle), 3 points on each line

In the picture the "midpoint" of two marked points really is their sum: $4+5 = [0{:}0{:}1] = 1$, and likewise all round. The circle is the line $x_0 = 0$, which under Theorem 22.6.2 is the line at infinity, so the picture also shows $\mathbb{P}^2(\mathbb{F}_2)$ as the four affine points $4, 5, 6, 7$ of $\mathbb{F}_2^2$ together with

the three directions $1, 2, 3$.

Warning. **A drawing of a finite projective plane is a mnemonic, not a picture.** The seven points of the Fano plane are not points of the Euclidean plane, and the "circle" is a line, not a curve. The only content of the drawing is the incidence data: which triples of the seven labels form a line. Any picture with the same seven triples is the same object, and no metric statement about the drawing means anything.

22.6.6 Exercises

22.6.6.1 A. Check your understanding

Exercise. A1

(a) State Definition 22.6.1, and say why 0 is removed before taking the quotient.

(b) What is $\dim \mathbb{P}(V)$ when $\dim V = 5$? What is the dimension of a hyperplane of that projective space, and of what subspace of V is it the image?

(c) Decide whether $[2:4:6] = [1:2:3]$ in $\mathbb{P}^2(\mathbb{Q})$, and whether $[0:1] = [0:2]$ in $\mathbb{P}^1(\mathbb{F}_3)$. Justify your answers.

(d) True or false: the condition $x_1 + x_2 = 1$ defines a subset of $\mathbb{P}^2(\mathbb{R})$. Justify your answer.

(e) How many points does $\mathbb{P}^3(\mathbb{F}_5)$ have?

Solution.

(a) $\mathbb{P}(V) = (V \setminus \{0\})/\sim$ with $x \sim y$ when $y = \lambda x$ for some nonzero λ, and $\dim \mathbb{P}(V) = \dim V - 1$. The vector 0 is removed because its class is $\{0\}$ alone, so it would contribute one point whose span $\operatorname{span}(0) = \{0\}$ is the zero subspace and not a line, and Proposition 22.6.1 would fail.

(b) $\dim \mathbb{P}(V) = 4$. A hyperplane has dimension 3 and is $\mathbb{P}(U)$ for a subspace U of V with $\dim U = 4$.

(c) Yes: $(2,4,6) = 2\,(1,2,3)$ and $2 \neq 0$ in $\mathbb{Q}$. Yes again: $(0,2) = 2\,(0,1)$ and $2 \neq 0$ in $\mathbb{F}_3$.

(d) False. The expression is not homogeneous: $(1,1,0)$ satisfies it but $(2,2,0)$, which represents the same point, does not. Only the vanishing of a homogeneous expression is a condition on points.

(e) $(5^4 - 1)/(5 - 1) = 624/4 = 156$, by Proposition 22.6.9.

22.6.6.2 B. Practice

Exercise. B1: Meeting at infinity

In $\mathbb{P}^2(\mathbb{R})$, complete each pair of affine lines and find their intersection point, saying in each case whether it lies in the chart U_0 or at infinity.

(a) $y_2 = 2y_1 + 1$ and $y_2 = 2y_1 - 3$.

(b) $y_2 = 2y_1 + 1$ and $y_1 + y_2 = 4$.

(c) $y_1 = 5$ and $y_1 = -2$.

Solution.

(a) Writing them as $2y_1 - y_2 = -1$ and $2y_1 - y_2 = 3$, Proposition 22.6.3 gives $\bar{L}_1 : 2x_1 - x_2 + x_0 = 0$ and $\bar{L}_2 : 2x_1 - x_2 - 3x_0 = 0$. Subtracting, $4x_0 = 0$, so $x_0 = 0$ and $x_2 = 2x_1$: the point $[0 : 1 : 2]$, at infinity. This agrees with part (c) of that proposition, since $(a_1, a_2) = (2, -1)$ gives $[0 : 1 : 2]$.

(b) Here $\bar{L}_1 : 2x_1 - x_2 + x_0 = 0$ and $\bar{L}_2 : x_1 + x_2 - 4x_0 = 0$. Adding, $3x_1 - 3x_0 = 0$, so $x_1 = x_0$; then $x_2 = 4x_0 - x_1 = 3x_0$. Taking $x_0 = 1$ gives $[1 : 1 : 3]$, which lies in U_0 and corresponds to the affine point $(1, 3)$. Check: $3 = 2 \cdot 1 + 1$ and $1 + 3 = 4$.

(c) $\bar{L}_1 : x_1 - 5x_0 = 0$ and $\bar{L}_2 : x_1 + 2x_0 = 0$. Subtracting, $7x_0 = 0$, so $x_0 = 0$ and then $x_1 = 0$, forcing $x_2 \neq 0$: the point $[0 : 0 : 1]$, at infinity, the vertical direction.

Exercise. B2: Joins and meets

In $\mathbb{P}^4(F)$, let S be a plane and T a plane.

(a) What are the possible values of $\dim(S \cap T)$? Justify each.

(b) Show that two planes in $\mathbb{P}^4(F)$ always meet.

(c) Give an example in $\mathbb{P}^5(F)$ of two planes that do not meet.

Solution. Throughout, $S = \mathbb{P}(U)$ and $T = \mathbb{P}(W)$ with $\dim U = \dim W = 3$ inside a space of dimension 5 (for $\mathbb{P}^4$).

(a) By Theorem 22.6.6, $\dim(S \cap T) = 4 - \dim(S \vee T)$. Now $S \vee T$ contains S, so $\dim(S \vee T) \geq 2$, and it is contained in $\mathbb{P}^4$, so $\dim(S \vee T) \leq 4$. Hence $\dim(S \cap T) \in \{0, 1, 2\}$, and all three occur: 2 when $S = T$; 1 when $U \cap W$ has dimension 2, for instance $U = \operatorname{span}(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$, $W = \operatorname{span}(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_4)$; and 0 for $U = \operatorname{span}(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$, $W = \operatorname{span}(\mathbf{e}_3, \mathbf{e}_4, \mathbf{e}_5)$.

(b) The computation in (a) shows $\dim(S \cap T) \geq 4 - 4 = 0 > -1$, so $S \cap T \neq \emptyset$.

(c) In $\mathbb{P}^5(F) = \mathbb{P}(F^6)$ take $U = \operatorname{span}(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$ and $W = \operatorname{span}(\mathbf{e}_4, \mathbf{e}_5, \mathbf{e}_6)$. Then $U \cap W = \{\mathbf{0}\}$, so $S \cap T = \mathbb{P}(\{\mathbf{0}\}) = \emptyset$, and indeed Theorem 22.6.6 gives $\dim(S \vee T) + (-1) = 2 + 2$, that is $\dim(S \vee T) = 5$, the whole space.

Exercise. B3: A projective plane over three elements

Let $F = \mathbb{F}_3$.

(a) How many points does $\mathbb{P}^2(\mathbb{F}_3)$ have, and how many lines?

(b) How many points lie on each line, and how many lines pass through each point?

(c) List the four points of the line $x_0 = 0$.

Solution.

(a) By Proposition 22.6.9, $|\mathbb{P}^2(\mathbb{F}_3)| = 1 + 3 + 9 = 13$. A line is $\mathbb{P}(U)$ with $\dim U = 2$, equivalently the kernel of a nonzero functional on $\mathbb{F}_3^3$, and two nonzero functionals have the same kernel exactly when they are proportional (Corollary 5.6.6 applied to the 1×3 matrices, or directly: a kernel of dimension 2 determines the functional up to scale). So the lines correspond to the points of $\mathbb{P}((\mathbb{F}_3^3)^*)$, of which there are again 13.

(b) Each line has $(3^2 - 1)/(3 - 1) = 4$ points. For the lines through a fixed point $[\mathbf{x}]$: they are the $\mathbb{P}(U)$ with $\operatorname{span}(\mathbf{x}) \subseteq U$ and $\dim U = 2$, and every such U is $\operatorname{span}(\mathbf{x}, \mathbf{y})$ for one of the $3^3 - 3 = 24$ vectors $\mathbf{y} \notin \operatorname{span}(\mathbf{x})$, while two choices $\mathbf{y}, \mathbf{y}'$ give the same U exactly when $\mathbf{y}' \in U \setminus \operatorname{span}(\mathbf{x})$, a set of $3^2 - 3 = 6$ vectors. So each point lies on $24/6 = 4$ lines. (Consistently, the incident pairs number $13 \cdot 4 = 52$ counted by lines and $13 \cdot 4 = 52$ counted by points.)

(c) The points with $x_0 = 0$ are $[0 : x_1 : x_2]$ with $(x_1, x_2) \neq (0, 0)$, that is the points of $\mathbb{P}^1(\mathbb{F}_3)$: $[0 : 1 : 0]$, $[0 : 1 : 1]$, $[0 : 1 : 2]$ and $[0 : 0 : 1]$. Four points, as (b) requires.

22.6.6.3 C. Going deeper

Exercise. C1: Other charts

Work in $\mathbb{P}^n(F)$ and fix $j \in \{0, 1, \ldots, n\}$. Let $U_j = \{[\mathbf{x}] : x_j \neq 0\}$.

(a) Prove that U_j is well defined and in bijection with F^n.

(b) Prove that $U_0 \cup U_1 \cup \cdots \cup U_n = \mathbb{P}^n(F)$.

(c) For $n = 1$ and $F = \mathbb{R}$, describe the point of U_1 that is missing from U_0, and the point of U_0 missing from U_1. Conclude that the "point at infinity" is not a property of a point, but of a chart.

Solution.

(a) Both parts repeat Theorem 22.6.2 with the index 0 replaced by j. Scaling multiplies x_j by a nonzero scalar, so the condition $x_j \neq 0$ is independent of the representative. The bijection is

$$(a_1, \ldots, a_n) \mapsto [a_1 : \cdots : a_j : 1 : a_{j+1} : \cdots : a_n],$$

with the 1 in position j; it is injective because comparing the j-th entries of two proportional representatives forces the scalar to be 1, and surjective because $[\mathbf{x}] = [x_j^{-1}\mathbf{x}]$ when $x_j \neq 0$.

(b) A point is $[\mathbf{x}]$ with $\mathbf{x} \neq \mathbf{0}$, so some coordinate x_j is nonzero, and then $[\mathbf{x}] \in U_j$.

(c) $\mathbb{P}^1(\mathbb{R}) = \{[1 : a] : a \in \mathbb{R}\} \cup \{[0 : 1]\}$. The chart U_0 omits exactly $[0 : 1]$, and the chart $U_1 = \{[b : 1] : b \in \mathbb{R}\} \cup \{[1 : 0]\}$ omits exactly $[1 : 0]$. Both omitted points are perfectly ordinary points of $\mathbb{P}^1(\mathbb{R})$; each is at infinity for one chart and in the affine part of the other. So "at infinity" describes a relation between a point and a chosen chart, not the point itself.

Exercise. C2: Counting subspaces

Let F be a finite field with q elements.

(a) Prove that the number of lines in $\mathbb{P}^2(F)$ equals the number of points, namely q^2+q+1. *Hint: a line is the kernel of a nonzero functional.*

(b) Deduce that every point of $\mathbb{P}^2(F)$ lies on exactly $q+1$ lines.

(c) How many points does a hyperplane of $\mathbb{P}^n(F)$ have, and how many points of $\mathbb{P}^n(F)$ lie outside it?

Solution.

(a) A line of $\mathbb{P}^2(F)$ is $\mathbb{P}(U)$ with $\dim U = 2$, and by Lemma 5.6.1 such a U is the kernel of a nonzero functional φ on F^3, since $\dim U^0 = 3-2 = 1$ (Theorem 5.2.2) and U is the common kernel of its annihilator. Two nonzero functionals have the same kernel exactly when they are proportional: if $\ker\varphi = \ker\psi = U$, then φ and ψ both lie in the 1-dimensional space U^0, so one is a scalar multiple of the other, and the scalar is nonzero. So lines correspond to points of $\mathbb{P}((F^3)^*)$, and $(F^3)^*$ has dimension 3 (Corollary 5.1.4); by Proposition 22.6.9 there are q^2+q+1 of them.

(b) Count incident pairs (point, line through it) in two ways. Each line has $q+1$ points by Proposition 22.6.9, so the count is $(q^2+q+1)(q+1)$. By symmetry among the q^2+q+1 points — or simply dividing — each point lies on $(q^2+q+1)(q+1)/(q^2+q+1) = q+1$ lines, provided every point lies on the same number. That last point needs an argument. The lines through a fixed $[\mathbf{x}]$ are the $\mathbb{P}(U)$ with $\operatorname{span}(\mathbf{x}) \subseteq U$ and $\dim U = 2$, and every such U is $\operatorname{span}(\mathbf{x},\mathbf{y})$ for some $\mathbf{y} \notin \operatorname{span}(\mathbf{x})$, of which there are q^3-q. Two choices $\mathbf{y},\mathbf{y}'$ give the same U exactly when $\mathbf{y}' \in U \setminus \operatorname{span}(\mathbf{x})$, a set of q^2-q vectors. So the number of such U is $(q^3-q)/(q^2-q) = q+1$, independently of $[\mathbf{x}]$.

(c) A hyperplane is $\mathbb{P}(U)$ with $\dim U = n$, so it has $(q^n-1)/(q-1) = 1+q+\cdots+q^{n-1}$ points. Subtracting from $1+q+\cdots+q^n$ leaves q^n points outside, in agreement with Theorem 22.6.2, whose chart U_0 is exactly the complement of a hyperplane and is in bijection with F^n.

Exercise. C3: Where the affine formula goes wrong

Let A and B be flats of $\mathbb{A}^3(F)$.

(a) Give two lines in $\mathbb{A}^3(F)$ that do not meet and are not parallel, and compute $\dim(A \vee B)$ from Theorem 22.2.4.

(b) Complete both lines to projective lines in $\mathbb{P}^3(F)$ by the recipe of Theorem 22.6.2, that is by adding the points with $x_0 = 0$ to the projective subspace each spans, and verify Eq. 22.6.1 for the completed pair.

(c) Explain in one sentence why the projective formula could not have an exceptional case, given Proposition 22.6.5.

Solution.

(a) Take $A = \{(t,0,0) : t \in F\}$ and $B = \{(0,s,1) : s \in F\}$. A common point would need third coordinate 0 and 1, so $A \cap B = \emptyset$. Their direction spaces are $\operatorname{span}(\mathbf{e}_1)$ and

span($\mathbf{e}_2$), which are different, so they are not parallel and $\vec{A} \cap \vec{B} = \{\mathbf{0}\}$. The second case of Theorem 22.2.4 (c) gives $\dim(A \vee B) = 1 + 1 - 0 + 1 = 3$, the whole of $\mathbb{A}^3(F)$.

(b) Under ι of Theorem 22.6.2, A becomes $\{[1 : t : 0 : 0]\}$, whose points together with $[0 : 1 : 0 : 0]$ form $\mathbb{P}(U)$ with $U = \operatorname{span}((1,0,0,0),(0,1,0,0))$; and B becomes $\{[1 : 0 : s : 1]\}$, giving $\mathbb{P}(W)$ with $W = \operatorname{span}((1,0,0,1),(0,0,1,0))$. Now $U \cap W = \{\mathbf{0}\}$: a vector of U has third and fourth entries 0, while a vector $a(1,0,0,1) + b(0,0,1,0)$ of W has third entry b and fourth entry a, so both a and b vanish. Hence $S \cap T = \emptyset$ of dimension -1, and $U + W$ has dimension $2 + 2 - 0 = 4$, so $\dim(S \vee T) = 3$. Then $3 + (-1) = 1 + 1$, as Eq. 22.6.1 requires. The two projective lines still do not meet, which is correct: Corollary 22.6.7 is about a projective **plane**, and these lines do not lie in one.

(c) Because by Proposition 22.6.5 both the meet and the join of projective subspaces are again projective subspaces, coming from $U \cap W$ and $U + W$, and the vector-space formula Theorem 1.7.3 holds for those with no hypothesis whatever, the empty intersection being simply the case $U \cap W = \{0\}$.

22.7 Projective Transformations

Section 06 built $\mathbb{P}(V)$ out of the lines through the origin of V, and a construction is only as good as the maps that go with it. Here the maps are already in hand: an invertible linear operator sends lines through the origin to lines through the origin, so it permutes the points of $\mathbb{P}(V)$. This section shows that two operators give the same permutation exactly when they differ by a nonzero scalar, that such a map is pinned down by its values on $n+2$ points in general position — and, over any field with more than two elements, **not** on $n+1$ — and that the failure at $n+1$ is precisely what makes the cross-ratio of four points on a line a genuine invariant.

Throughout, F is an arbitrary field and V, W are finite-dimensional vector spaces over F. Nothing in this section needs a characteristic assumption, an ordering or an inner product. In F^{n+1} we index coordinates from 0, writing $\mathbf{x} = (x_0, x_1, \dots, x_n)$ and $\mathbf{e}_0, \dots, \mathbf{e}_n$ for the standard basis, so that homogeneous coordinates read $[x_0 : x_1 : \cdots : x_n]$ as in Definition 22.6.2. We write $F^\times = F \setminus \{0\}$ for the multiplicative group of F (Example (Groups)), and $\mathrm{GL}(V)$ for the set of invertible linear operators $V \to V$; composition makes $\mathrm{GL}(V)$ a group (Definition 0.10.1), with identity id_V, exactly as it makes $\mathrm{GL}_n(F)$ one.

22.7.1 Maps induced by linear isomorphisms

We need maps $\mathbb{P}(V) \to \mathbb{P}(W)$, and the only structure a projective space has is its supply of points, that is, of one-dimensional subspaces (Definition 22.6.1). So we want maps that carry one-dimensional subspaces to one-dimensional subspaces. A linear map T does this as soon as it kills nothing: $T(\operatorname{span}(x)) = \operatorname{span}(Tx)$, and the right-hand side is a line precisely when $Tx \neq 0$.

A projective transformation is an invertible linear map, read on lines instead of on vectors.

Definition 22.7.1. Projective Transformation

Let $T\colon V \to W$ be an **isomorphism** of vector spaces over F. The **projective transformation**, or **projectivity**, induced by T is the map

$$[T]\colon \mathbb{P}(V) \to \mathbb{P}(W), \qquad [T]([x]) := [Tx].$$

A map $\mathbb{P}(V) \to \mathbb{P}(W)$ is a projectivity if it equals $[T]$ for **some** isomorphism T.

Two things must be checked before the formula means anything, and both use invertibility. First, $[Tx]$ has to be a point at all: a point of $\mathbb{P}(W)$ is $[w]$ for a **nonzero** w, and if $x \neq 0$ then $Tx \neq 0$, because T is injective (Theorem 2.2.3). Second, the recipe must not depend on the representative: if $[x] = [y]$, then $y = \lambda x$ for some $\lambda \in F^\times$ (Definition 22.6.2), so $Ty = \lambda Tx$ and $[Ty] = [Tx]$.

Proposition 22.7.1. Projectivities Compose and Invert

Let $T\colon V \to W$ and $S\colon W \to X$ be isomorphisms of finite-dimensional vector spaces over F. Then $[S] \circ [T] = [ST]$, the map $[T]$ is a bijection with inverse $[T^{-1}]$, and $[\mathrm{id}_V] = \mathrm{id}_{\mathbb{P}(V)}$.

Proof. For $x \in V$ with $x \neq 0$, $([S] \circ [T])([x]) = [S(Tx)] = [(ST)x] = [ST]([x])$, and ST is an isomorphism as a composition of isomorphisms (Theorem 2.5.3). Taking $S = T^{-1}$ gives $[T^{-1}] \circ [T] = [\mathrm{id}_V]$, and $[\mathrm{id}_V]([x]) = [x]$, so $[\mathrm{id}_V] = \mathrm{id}_{\mathbb{P}(V)}$; taking the composition in the other order gives $[T] \circ [T^{-1}] = \mathrm{id}_{\mathbb{P}(W)}$. Hence $[T]$ is a bijection with inverse $[T^{-1}]$. ∎

Examples. Take $V = W = F^{n+1}$, so that a projectivity of $\mathbb{P}^n(F)$ is $[\mathbf{x}] \mapsto [\mathbf{Ax}]$ for an invertible $\mathbf{A} \in \mathrm{GL}_{n+1}(F)$.

- $\mathbf{A} = \mathbf{I}_{n+1}$ gives the identity, and so does $\mathbf{A} = \lambda \mathbf{I}_{n+1}$ for any $\lambda \neq 0$, since $[\lambda \mathbf{x}] = [\mathbf{x}]$. This degenerate case is the whole content of the next subsection.
- On $\mathbb{P}^1(F)$, the matrix $\mathbf{A} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ sends $[x_0 : x_1]$ to $[x_1 : x_0]$. In the chart $x_1 \neq 0$, where $[t : 1]$ is the scalar t, this is $t \mapsto 1/t$ for $t \neq 0$, and it exchanges $[1 : 0]$ with $[0 : 1]$. A projectivity repairs the one defect of $t \mapsto 1/t$: on the projective line there is no missing value.
- On $\mathbb{P}^2(F)$, a permutation matrix permutes the three coordinate points $[1 : 0 : 0]$, $[0 : 1 : 0]$, $[0 : 0 : 1]$.

Nonexample by minimal change. Drop invertibility. On $\mathbb{P}^1(F)$ put $\mathbf{A} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$. For $[1 : 0]$ the recipe gives $[1 : 0]$, which is fine; but $\mathbf{A}(0, 1) = (0, 0)$, and $[\mathbf{0}]$ is not a point of $\mathbb{P}^1(F)$. The formula fails on exactly the points of $\mathbb{P}(\ker \mathbf{A})$, and injectivity is the clause that rules them out. The recipe does still make sense on the remaining points $[1 : t]$, but it is no longer injective there: $\mathbf{A}(1, t) = (1, 0)$ for every t, so all of them go to the single point $[1 : 0]$.

Projectivities respect everything $\mathbb{P}(V)$ is made of.

Proposition 22.7.2. Projectivities Preserve Subspaces, Joins, and Meets

Let $T\colon V \to W$ be an isomorphism and let U be a subspace of V. Then

$$[T](\mathbb{P}(U)) = \mathbb{P}(T(U)),$$

a projective subspace of $\mathbb{P}(W)$ of the same dimension as $\mathbb{P}(U)$, and the restriction of $[T]$ to $\mathbb{P}(U)$ is the projectivity induced by the isomorphism $T|_U\colon U \to T(U)$. Consequently $[T]$

preserves inclusion, intersections, and joins of projective subspaces.

Proof. By Definition 22.6.3, $\mathbb{P}(U) = \{[u] : u \in U,\ u \neq 0\}$ and $\dim \mathbb{P}(U) = \dim U - 1$. A point of $[T](\mathbb{P}(U))$ is $[Tu]$ with $0 \neq u \in U$, and these are exactly the points $[w]$ with $0 \neq w \in T(U)$, since T is a bijection from U onto $T(U)$. Hence $[T](\mathbb{P}(U)) = \mathbb{P}(T(U))$. The restriction $T|_U$ is linear and injective with image $T(U)$, hence an isomorphism onto $T(U)$, so $\dim T(U) = \dim U$ (Theorem 2.5.4) and the two projective subspaces have equal dimension; and $[T|_U]$ agrees with $[T]$ on $\mathbb{P}(U)$ by the same formula.

For the last sentence, T preserves inclusion of subspaces and satisfies $T(U_1 \cap U_2) = T(U_1) \cap T(U_2)$ and $T(U_1 + U_2) = T(U_1) + T(U_2)$, the first because T is injective and the second because T is linear and surjective onto its image. Passing to projective subspaces, inclusion transfers in both directions by Lemma 22.6.4, intersections by Proposition 22.6.5 (a), and joins by Proposition 22.6.5 (b), in each case applied on both sides of T.

■

22.7.2 The projective linear group

Different matrices can induce the same projectivity: $\mathbf{I}_2$ and $2\mathbf{I}_2$ both act as the identity on $\mathbb{P}^1(\mathbb{Q})$. The next theorem says that this is the only way it happens, and it is the reason the group of projectivities is smaller than GL.

Theorem 22.7.3. Projectivities and Scalars

Let $\dim V = \dim W \geq 1$ and let $S, T\colon V \to W$ be isomorphisms.

(a) $[S] = [T]$ if and only if $S = \lambda T$ for some $\lambda \in F^\times$.

(b) The set $\mathrm{PGL}(V) := \{[T] : T \in \mathrm{GL}(V)\}$ of projectivities of $\mathbb{P}(V)$ is a group under composition, the **projective linear group** of V.

(c) The map $\Pi\colon \mathrm{GL}(V) \to \mathrm{PGL}(V)$, $T \mapsto [T]$, is a surjective group homomorphism, and $\ker \Pi = \{\lambda \,\mathrm{id}_V : \lambda \in F^\times\}$.

For $V = F^{n+1}$ we write $\mathrm{PGL}_{n+1}(F)$ for $\mathrm{PGL}(F^{n+1})$; part (a) then says that $\mathbf{A}, \mathbf{B} \in \mathrm{GL}_{n+1}(F)$ induce the same projectivity of $\mathbb{P}^n(F)$ exactly when $\mathbf{B} = \lambda\mathbf{A}$ for some $\lambda \neq 0$.

Idea. One direction is the computation $[\lambda Tx] = [Tx]$. For the other, put $R = T^{-1}S$, so that the hypothesis says $[Rx] = [x]$ for every $x \neq 0$: **every nonzero vector is an eigenvector of** R. That is far stronger than it looks. Test it on x, y and $x + y$: if x and y are independent, comparing coefficients forces their eigenvalues to agree, and if they are dependent, the eigenvalue is forced directly. So R is a scalar. Parts (b) and (c) are then bookkeeping with Proposition 22.7.1.

Proof.

(a) ($\Leftarrow$) If $S = \lambda T$ with $\lambda \neq 0$, then for every $x \neq 0$ we have $Sx = \lambda(Tx)$, so $[Sx] = [Tx]$ by Definition 22.6.2. Hence $[S] = [T]$.

($\Rightarrow$) Suppose $[S] = [T]$. Put $R = T^{-1}S \in \mathrm{GL}(V)$. For every $x \neq 0$, $[Sx] = [Tx]$ gives $Sx = \mu\, Tx$ for some $\mu \in F^\times$, hence $Rx = T^{-1}Sx = \mu x$. So for each $x \neq 0$ there is a scalar $\lambda_x \in F^\times$ with $Rx = \lambda_x x$.
We show λ_x does not depend on x. Let $x, y \neq 0$.
Case 1: (x, y) is dependent. Take a dependence $sx + ty = 0$ with s, t not both zero. If $t = 0$ then $s \neq 0$ and $x = 0$, which is false; so $t \neq 0$ and $y = cx$ with $c = -s/t$, and $c \neq 0$ because $y \neq 0$. Then

$$\lambda_y y = Ry = cRx = c\lambda_x x = \lambda_x y,$$

and $y \neq 0$ gives $\lambda_y = \lambda_x$.
Case 2: (x, y) is independent. Then $x + y \neq 0$, and

$$\lambda_{x+y} x + \lambda_{x+y} y = R(x + y) = Rx + Ry = \lambda_x x + \lambda_y y.$$

Independence lets us compare coefficients, giving $\lambda_x = \lambda_{x+y} = \lambda_y$.
So there is one scalar $\lambda \in F^\times$ with $Rx = \lambda x$ for all $x \neq 0$, and also for $x = 0$. Hence $R = \lambda\, \mathrm{id}_V$ and $S = TR = \lambda T$.

(b) By Proposition 22.7.1, $\mathrm{PGL}(V)$ is closed under composition and under inverses and contains $\mathrm{id}_{\mathbb{P}(V)} = [\mathrm{id}_V]$; composition of maps is associative. So (G1)–(G3) of Definition 0.10.1 hold.

(c) Surjectivity is the definition of $\mathrm{PGL}(V)$, and $\Pi(ST) = [ST] = [S][T] = \Pi(S)\Pi(T)$ by Proposition 22.7.1, so Π is a homomorphism (Definition 0.10.3). Its kernel is $\{T : [T] = [\mathrm{id}_V]\}$, which by (a) is $\{\lambda\, \mathrm{id}_V : \lambda \in F^\times\}$. This proves the theorem.

■

Remark. Part (c) is the sentence "$\mathrm{PGL}(V)$ is $\mathrm{GL}(V)$ modulo scalars", made precise without quotient groups, which this book has not built. By (a), the fiber of Π over $[T]$ is $\{\lambda T : \lambda \in F^\times\}$, and the fibers of any map partition its domain. The picture is the one Chapter 2 drew for the cosets $v + U$ of Definition 4.4.2, with multiplication in place of addition and Lemma 4.4.3 replaced by part (a) above.

Warning. **A projectivity has no determinant, no trace, and no eigenvalues.** Only a matrix has those, and a projectivity of $\mathbb{P}^n(F)$ is a matrix only up to a scalar. On $\mathbb{P}^2(\mathbb{Q})$, the matrices $\mathbf{I}_3$ and $2\mathbf{I}_3$ induce the same projectivity but have determinants 1 and 8, and traces 3 and 6. In general $\det(\lambda\mathbf{A}) = \lambda^{n+1} \det \mathbf{A}$, so the most one can hope to define is $\det \mathbf{A}$ modulo $(n+1)$-th powers. What **is** well defined is a fixed point: $[\mathbf{Ax}] = [\mathbf{x}]$ says $\mathbf{x}$ is an eigenvector of $\mathbf{A}$, and being an eigenvector does not depend on scaling $\mathbf{A}$.

22.7.3 Frames, and how many points pin a projectivity down

A linear map is determined by its values on a basis. A projectivity of $\mathbb{P}^n(F)$ is induced by an $(n+1) \times (n+1)$ matrix, which has $n+1$ columns, so the first guess is that $n+1$ points should determine it. Over any field with more than two elements the guess is wrong — the

two-element field is a genuine exception, and we return to it below — and the reason is visible in Theorem 22.7.3: knowing $[\mathbf{A}\mathbf{e}_i]$ fixes each column of $\mathbf{A}$ only up to its own scalar, which leaves $n+1$ unknown scalars, while changing $\mathbf{A}$ to $\lambda\mathbf{A}$ costs only one. One extra point supplies the missing information.

First, the hypothesis under which the extra point helps.

Definition 22.7.2. General Position

Let $\dim V = n+1$. A list $P_1, \dots, P_m$ of points of $\mathbb{P}(V)$ is **in general position** if for **every** subset $I \subseteq \{1, \dots, m\}$ with $|I| \le n+1$, and any choice of representatives v_i with $P_i = [v_i]$, the list $(v_i)_{i \in I}$ is linearly independent.

The clause "any choice of representatives" costs nothing: replacing v_i by $c_i v_i$ with $c_i \neq 0$ does not change whether a list is independent, so the condition may be tested on one choice. For $m = n+2$, which is the case we need, general position says exactly that **every $n+1$ of the points have independent representatives**; no subset of size $n+2$ can be independent in an $(n+1)$-dimensional space, which is why the bound $|I| \le n+1$ is there.

Examples. In $\mathbb{P}^2(F)$, the four points $[1:0:0]$, $[0:1:0]$, $[0:0:1]$, $[1:1:1]$ are in general position: the first three representatives form a basis, and any three of the four are independent, since each of the other three triples has determinant ± 1. In $\mathbb{P}^1(F)$, general position for three points means simply that they are pairwise distinct. The same holds in every dimension: the $n+2$ points $[\mathbf{e}_0], \dots, [\mathbf{e}_n]$, $[1 : \cdots : 1]$ of $\mathbb{P}^n(F)$ are in general position. Indeed $n+1$ of the representatives $\mathbf{e}_0, \dots, \mathbf{e}_n, (1, \dots, 1)$ are either the $\mathbf{e}_i$ themselves, with determinant 1, or all of them but one, say $\mathbf{e}_j$, together with $(1, \dots, 1)$; in that matrix row j is zero except for the entry 1 in the last column, and expanding along row j leaves the determinant of a permutation matrix, so the value is ± 1. In $\mathbb{P}^n(F)$, a single point is always in general position, since one nonzero vector is independent.

Nonexample by minimal change. Replace $[1:1:1]$ by $[1:1:0]$. The list $[1:0:0]$, $[0:1:0]$, $[0:0:1]$, $[1:1:0]$ fails: the subset $\{1,2,4\}$ has representatives $(1,0,0)$, $(0,1,0)$, $(1,1,0)$, and the third is the sum of the first two. Geometrically, three of the four points are collinear. No projectivity carries the first list to this one: a projectivity preserves joins and their dimensions (Proposition 22.7.2), and here $[1:0:0] \vee [0:1:0] \vee [1:1:0]$ is a line, while $[1:0:0] \vee [0:1:0] \vee [1:1:1]$ is the whole plane.

The key step is that general position lets us choose representatives once and for all.

Lemma 22.7.4. Normalizing a Frame

Let $\dim V = n+1 \ge 1$ and let $P_1, \dots, P_{n+2}$ be points of $\mathbb{P}(V)$ in general position. Then there are representatives v_i with $P_i = [v_i]$ such that

$$(v_1, \dots, v_{n+1}) \text{ is a basis of } V \qquad \text{and} \qquad v_{n+2} = v_1 + \cdots + v_{n+1}.$$

Moreover such a list is unique up to one common scalar: if $(v'_1, \dots, v'_{n+2})$ is another, then $v'_i = a v_i$ for all i, with a single $a \in F^\times$.

Idea. Start from arbitrary representatives u_i. The first $n+1$ form a basis, so u_{n+2} has coordinates $c_1, \dots, c_{n+1}$ in it. General position is exactly the statement that **no coordinate is zero**: a vanishing c_j would put u_{n+2} in the span of the other n basis vectors, making $n+1$

of our points dependent. Nonzero coordinates can be absorbed into the basis vectors, and then the last vector is the sum of the others. For uniqueness, write the second list as $a_i v_i$ and read off from the relation $\sum v_i = v_{n+2}$ that all a_i coincide.

Proof. Choose any representatives u_i with $P_i = [u_i]$. By Definition 22.7.2 the list $(u_1, \dots, u_{n+1})$ is independent, and it has $n+1 = \dim V$ entries, so it is a basis of V (Theorem 1.6.4). Write

$$u_{n+2} = c_1 u_1 + \cdots + c_{n+1} u_{n+1}, \qquad c_i \in F.$$

Every c_j is nonzero. Suppose $c_j = 0$ for some $j \le n+1$. Then

$$1 \cdot u_{n+2} - \sum_{i \neq j,\ i \le n+1} c_i u_i = 0$$

is a linear relation whose coefficient on u_{n+2} is $1 \neq 0$, so the list consisting of u_{n+2} together with the n vectors u_i ($i \neq j$, $i \le n+1$) is dependent (Definition 1.4.1). It has $n+1$ entries and its index set has $n+1$ elements, contradicting Definition 22.7.2.

Now set $v_i := c_i u_i$ for $i \le n+1$ and $v_{n+2} := u_{n+2}$. Each $c_i \neq 0$, so $[v_i] = [u_i] = P_i$, and $(v_1, \dots, v_{n+1})$ is still a basis, being the old basis with each vector scaled by a nonzero scalar. By construction $v_{n+2} = \sum_{i=1}^{n+1} c_i u_i = \sum_{i=1}^{n+1} v_i$.

Uniqueness. Let $(v_1', \dots, v_{n+2}')$ be another such list. Since $[v_i'] = P_i = [v_i]$, there are $a_i \in F^\times$ with $v_i' = a_i v_i$. Applying the defining relation to both lists,

$$a_{n+2} \sum_{i=1}^{n+1} v_i = a_{n+2} v_{n+2} = v_{n+2}' = \sum_{i=1}^{n+1} v_i' = \sum_{i=1}^{n+1} a_i v_i.$$

The list $(v_1, \dots, v_{n+1})$ is independent, so comparing coefficients gives $a_i = a_{n+2}$ for every $i \le n+1$. Hence all a_i equal one scalar a.

■

The theorem is now three lines of linear algebra.

Theorem 22.7.5. Fundamental Theorem of Projective Geometry

Let $\dim V = \dim W = n+1 \ge 1$. Let $P_1, \dots, P_{n+2}$ be points of $\mathbb{P}(V)$ in general position and $Q_1, \dots, Q_{n+2}$ points of $\mathbb{P}(W)$ in general position. Then there is **exactly one** projectivity $f \colon \mathbb{P}(V) \to \mathbb{P}(W)$ with

$$f(P_i) = Q_i \qquad \text{for } i = 1, \dots, n+2.$$

Idea. Normalize both lists by Lemma 22.7.4. The first $n+1$ normalized vectors on each side are bases, so there is a unique linear map sending one basis to the other, and the normalization was arranged so that this map automatically takes the $(n+2)$-nd vector to the $(n+2)$-nd. For uniqueness, a second projectivity differs from the first by an operator that scales each of $v_1, \dots, v_{n+1}$ and also their sum; the same coefficient comparison as in the lemma forces all the scalars to agree.

Proof. By Lemma 22.7.4 choose representatives $v_1, \dots, v_{n+2}$ of the P_i with $(v_1, \dots, v_{n+1})$ a basis of V and $v_{n+2} = v_1 + \cdots + v_{n+1}$, and likewise $w_1, \dots, w_{n+2}$ for the Q_i.

Existence. By Theorem 2.1.5 there is a unique linear $T\colon V \to W$ with $Tv_i = w_i$ for $i \le n+1$. Its image is $\operatorname{span}(w_1, \dots, w_{n+1}) = W$ by Theorem 2.2.4, so T is surjective, and since $\dim V = \dim W$ is finite, T is an isomorphism by Corollary 2.3.2 (e). By linearity,

$$Tv_{n+2} = T\Big(\sum_{i=1}^{n+1} v_i\Big) = \sum_{i=1}^{n+1} w_i = w_{n+2}.$$

Hence $[T](P_i) = [Tv_i] = [w_i] = Q_i$ for all $i \le n+2$.

Uniqueness. Let $[S]$ be a projectivity with $[S](P_i) = Q_i$ for all i, where $S\colon V \to W$ is an isomorphism. Then $[Sv_i] = [w_i]$, so there are $\mu_i \in F^\times$ with $Sv_i = \mu_i w_i$. Put $R = T^{-1}S \in \mathrm{GL}(V)$; then $Rv_i = \mu_i v_i$ for every $i \le n+2$. Applying R to $v_{n+2} = \sum_{i \le n+1} v_i$,

$$\mu_{n+2} \sum_{i=1}^{n+1} v_i = Rv_{n+2} = \sum_{i=1}^{n+1} Rv_i = \sum_{i=1}^{n+1} \mu_i v_i.$$

By independence of $(v_1, \dots, v_{n+1})$, $\mu_i = \mu_{n+2} =: \mu$ for all $i \le n+1$. So R agrees with $\mu\,\mathrm{id}_V$ on a basis, hence $R = \mu\,\mathrm{id}_V$ (Theorem 2.1.5), and $S = TR = \mu T$. By Theorem 22.7.3 (a), $[S] = [T]$. This proves the theorem.

■

Existence and uniqueness are both used constantly: existence says we may always **choose coordinates** so that $n+2$ prescribed points in general position become $[\mathbf{e}_0], \dots, [\mathbf{e}_n]$ and $[1 : \cdots : 1]$; uniqueness says that having done so, nothing is left to choose.

Example. Moving the standard frame

Work in $\mathbb{P}^2(\mathbb{Q})$. Find the projectivity carrying

$$[1:0:0], \quad [0:1:0], \quad [0:0:1], \quad [1:1:1]$$

to

$$Q_0 = [1:1:0], \quad Q_1 = [0:1:1], \quad Q_2 = [1:0:1], \quad Q_3 = [4:3:5],$$

after checking that the second list is in general position.

Solution. General position. Form the 3×3 determinant of each triple of representatives, taken as columns. For the triples (Q_0, Q_1, Q_2), (Q_0, Q_1, Q_3), (Q_0, Q_2, Q_3), (Q_1, Q_2, Q_3) these determinants are 2, 6, -4, 2. All are nonzero in $\mathbb{Q}$, so every triple is independent (Theorem 7.5.2 together with the fact that the columns of an invertible matrix are independent), and the four points are in general position.

Normalizing. Following Lemma 22.7.4, we need scalars $\lambda_0, \lambda_1, \lambda_2$ with

$$\lambda_0(1,1,0) + \lambda_1(0,1,1) + \lambda_2(1,0,1) = (4,3,5),$$

that is, $\lambda_0 + \lambda_2 = 4$, $\lambda_0 + \lambda_1 = 3$, $\lambda_1 + \lambda_2 = 5$. Adding all three gives $\lambda_0 + \lambda_1 + \lambda_2 = 6$, so

$\lambda_0 = 6-5 = 1$, $\lambda_1 = 6-4 = 2$, $\lambda_2 = 6-3 = 3$. All are nonzero, as the lemma promised.

The matrix. The standard frame is already normalized, with $\mathbf{e}_0 + \mathbf{e}_1 + \mathbf{e}_2 = (1,1,1)$. So the map sends $\mathbf{e}_i$ to $\lambda_i \mathbf{q}_i$, that is,

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 3 \\ 1 & 2 & 0 \\ 0 & 2 & 3 \end{pmatrix}, \qquad \det \mathbf{A} = 12 \neq 0.$$

Check. The columns are $(1,1,0)$, $(0,2,2)$, $(3,0,3)$, which represent Q_0, Q_1, Q_2; and $\mathbf{A}(1,1,1) = (4,3,5)$, which represents Q_3. By Theorem 22.7.5 this $[\mathbf{A}]$ is the only projectivity doing this, so every other matrix that works is a scalar multiple of $\mathbf{A}$.

Now the negative result promised above. It is not that the proof of Theorem 22.7.5 is wasteful; as long as F has more than two elements, $n+1$ points genuinely do not suffice.

Proposition 22.7.6. What Fixes the Coordinate Points

Let $n \geq 1$. A projectivity of $\mathbb{P}^n(F)$ fixes each of the $n+1$ points $[\mathbf{e}_0], \ldots, [\mathbf{e}_n]$ if and only if it is $[\mathbf{D}]$ for an invertible **diagonal** matrix $\mathbf{D}$. The group of such projectivities is trivial if and only if $F = \mathbb{F}_2$.

Proof. ($\Leftarrow$) If $\mathbf{D} = \operatorname{diag}(d_0, \ldots, d_n)$ with all $d_i \neq 0$, then $\mathbf{D}\mathbf{e}_i = d_i \mathbf{e}_i$, so $[\mathbf{D}][\mathbf{e}_i] = [\mathbf{e}_i]$.

($\Rightarrow$) Let $[\mathbf{A}]$ fix each $[\mathbf{e}_i]$. Then $\mathbf{A}\mathbf{e}_i = d_i \mathbf{e}_i$ for some $d_i \in F^\times$. But $\mathbf{A}\mathbf{e}_i$ is the i-th column of $\mathbf{A}$ (Theorem 0.9.2), so every column of $\mathbf{A}$ has a single nonzero entry, on the diagonal: $\mathbf{A} = \operatorname{diag}(d_0, \ldots, d_n)$.

For the last sentence, by Theorem 22.7.3 (a) we have $[\mathbf{D}] = [\mathbf{I}_{n+1}]$ if and only if $\mathbf{D} = \lambda \mathbf{I}_{n+1}$, that is, if and only if $d_0 = \cdots = d_n$. So the group is trivial exactly when every tuple $(d_0, \ldots, d_n) \in (F^\times)^{n+1}$ is constant, which, since $n \geq 1$, happens exactly when $F^\times$ has one element, that is, when $F = \mathbb{F}_2$.

■

Remark. $n+1$ **points do not determine a projectivity.** Let F have an element $c \neq 0, 1$ — every field except $\mathbb{F}_2$. In $\mathbb{P}^1(F)$ the two points $[1:0]$ and $[0:1]$ are in general position, and both $[\mathbf{I}_2]$ and $[\operatorname{diag}(1,c)]$ fix them, yet they are different maps: the second sends $[1:1]$ to $[1:c] \neq [1:1]$. The same works in $\mathbb{P}^n(F)$ with $\mathbf{D} = \operatorname{diag}(1, \ldots, 1, c)$ and the point $[1:\cdots:1]$. So the count of $n+2$ in Theorem 22.7.5 is exact.

Over $\mathbb{F}_2$ this counterexample is unavailable, and Proposition 22.7.6 already shows why: the group it describes is trivial there. So the negative result needs the hypothesis that F has at least three elements, and over the one excluded field the situation is genuinely different.

Proposition 22.7.7. Over the Two-Element Field, $n+1$ Points Suffice

Let $n \geq 1$ and $F = \mathbb{F}_2$. Given two lists $P_0, \ldots, P_n$ and $Q_0, \ldots, Q_n$ of $n+1$ points of $\mathbb{P}^n(\mathbb{F}_2)$,

each in general position, there is exactly one projectivity of $\mathbb{P}^n(\mathbb{F}_2)$ carrying P_i to Q_i for every i.

Proof. Over $\mathbb{F}_2$ the only nonzero scalar is 1, so each point has exactly one representative: write $P_i = [\mathbf{v}_i]$ and $Q_i = [\mathbf{w}_i]$. By Definition 22.7.2 the lists $(\mathbf{v}_0, \dots, \mathbf{v}_n)$ and $(\mathbf{w}_0, \dots, \mathbf{w}_n)$ are independent, and each has $n+1 = \dim \mathbb{F}_2^{n+1}$ entries, hence each is a basis (Theorem 1.6.4). By Theorem 2.1.5 there is a unique linear T with $T\mathbf{v}_i = \mathbf{w}_i$ for all i; its image is $\operatorname{span}(\mathbf{w}_0, \dots, \mathbf{w}_n) = \mathbb{F}_2^{n+1}$ (Theorem 2.2.4), so it is surjective and hence an isomorphism (Corollary 2.3.2 (e)). Thus $[T]$ carries P_i to Q_i.

If $[S]$ does too, then $[S\mathbf{v}_i] = [\mathbf{w}_i]$, so $S\mathbf{v}_i = \mu_i \mathbf{w}_i$ with $\mu_i \in \mathbb{F}_2 \setminus \{0\} = \{1\}$, that is, $S\mathbf{v}_i = \mathbf{w}_i = T\mathbf{v}_i$ for every i. Two linear maps agreeing on a basis are equal (Theorem 2.1.5), so $S = T$ and $[S] = [T]$.

■

So the theorem to remember is Theorem 22.7.5, with the count $n+2$ that works over every field; $\mathbb{F}_2$ is the one field where a smaller count also works, for the accidental reason that it has no scalars to spare.

Quick check. Exhibit two different projectivities of $\mathbb{P}^2(\mathbb{Q})$ that agree on $[1:0:0]$, $[0:1:0]$ and $[0:0:1]$, and name a point where they disagree.

Solution. Take $[\mathbf{I}_3]$ and $[\mathbf{D}]$ with $\mathbf{D} = \operatorname{diag}(1,1,2)$. Both fix the three coordinate points. They differ at $[1:1:1]$, which $[\mathbf{D}]$ sends to $[1:1:2]$; and $[1:1:2] \neq [1:1:1]$, since $(1,1,2)$ is not a scalar multiple of $(1,1,1)$. By Theorem 22.7.3 (a) the two projectivities are different, because $\mathbf{D}$ is not a scalar multiple of $\mathbf{I}_3$.

22.7.4 The cross-ratio

Three distinct points on a projective line can be moved to any other three, by Theorem 22.7.5 with $n = 1$. A fourth point cannot: the theorem has used up all its freedom. What is left over is one invariant, and it is a point of $\mathbb{P}^1(F)$.

Definition 22.7.3. Cross-Ratio

Let L be a projective line, that is, a projective space of dimension 1 over F, let $A, B, C \in L$ be **pairwise distinct**, and let $D \in L$ be arbitrary. Let $f\colon L \to \mathbb{P}^1(F)$ be the unique projectivity with

$$f(A) = [1:0], \qquad f(B) = [0:1], \qquad f(C) = [1:1].$$

The **cross-ratio** of the ordered quadruple is the point

$$(A, B; C, D) := f(D) \in \mathbb{P}^1(F).$$

The projectivity f exists and is unique because three pairwise distinct points of a line are in general position (Definition 22.7.2 with $n = 1$), on both sides. The cross-ratio is a **point of**

$\mathbb{P}^1(F)$, not a scalar; in the chart $[\lambda : \mu] \mapsto \lambda/\mu$ it becomes a scalar whenever $\mu \neq 0$, and the one excluded value, $[1 : 0]$, occurs exactly when $D = A$, since f is injective. Likewise $(A, B; C, D) = [0 : 1]$ exactly when $D = B$, and $[1 : 1]$ exactly when $D = C$.

A formula makes this computable.

Proposition 22.7.8. Computing the Cross-Ratio

In the situation of Definition 22.7.3, write $L = \mathbb{P}(U)$ with $\dim U = 2$, choose representatives $A = [a]$, $B = [b]$, and expand

$$c = c_1 a + c_2 b, \qquad d = d_1 a + d_2 b$$

for representatives c of C and d of D. Then $c_1 \neq 0$, $c_2 \neq 0$, and

$$(A, B; C, D) = [\, d_1 c_2 : d_2 c_1 \,].$$

If moreover $L = \mathbb{P}^1(F)$ and the four points lie in the chart $\mu \neq 0$, with scalars a, b, c, d given by $[t : 1] \mapsto t$, then

$$(A, B; C, D) = \Big[\, (c - a)(d - b) \;:\; (c - b)(d - a) \,\Big].$$

Idea. Since $A \neq B$, the pair (a, b) is a basis of U, so the expansions exist. The projectivity of Definition 22.7.3 is the one induced by the linear map sending $c_1 a \mapsto \mathbf{e}_0$ and $c_2 b \mapsto \mathbf{e}_1$, because that map sends c to $\mathbf{e}_0 + \mathbf{e}_1$ — this is Lemma 22.7.4 for $n = 1$, done by hand. Then evaluate it at d. For the chart formula, take $\mathbf{a} = (a, 1)$, $\mathbf{b} = (b, 1)$ and solve two 2×2 systems.

Proof. The points $A \neq B$ give independent representatives (Definition 22.7.2), and $\dim U = 2$, so (a, b) is a basis of U (Theorem 1.6.4) and the expansions exist and are unique. If $c_1 = 0$ then $c = c_2 b$, so $C = B$; if $c_2 = 0$ then $C = A$. Both are excluded, so $c_1 c_2 \neq 0$.

Let $T \colon U \to F^2$ be the linear map with $T(c_1 a) = \mathbf{e}_0$ and $T(c_2 b) = \mathbf{e}_1$ (Theorem 2.1.5; $(c_1 a, c_2 b)$ is a basis since $c_1 c_2 \neq 0$). Its image is $\operatorname{span}(\mathbf{e}_0, \mathbf{e}_1) = F^2$ (Theorem 2.2.4), so it is surjective, hence an isomorphism by Corollary 2.3.2 (e). Then

$$[T](A) = [Ta] = [c_1^{-1}\mathbf{e}_0] = [1 : 0], \qquad [T](B) = [0 : 1],$$

and $[T](C) = [T(c_1 a + c_2 b)] = [\mathbf{e}_0 + \mathbf{e}_1] = [1 : 1]$. By the uniqueness in Theorem 22.7.5, $[T]$ is the map f of Definition 22.7.3. Finally

$$f(D) = [T(d_1 a + d_2 b)] = [\, d_1 c_1^{-1}\mathbf{e}_0 + d_2 c_2^{-1}\mathbf{e}_1 \,] = [\, d_1 c_1^{-1} : d_2 c_2^{-1} \,],$$

and multiplying both entries by $c_1 c_2 \neq 0$ gives $[\, d_1 c_2 : d_2 c_1 \,]$.

For the chart formula take $\mathbf{a} = (a, 1)$, $\mathbf{b} = (b, 1)$, $\mathbf{c} = (c, 1)$, $\mathbf{d} = (d, 1)$; these are representatives of the four points, and $a \neq b$ because $A \neq B$. Solving $c_1 a + c_2 b = c$, $c_1 + c_2 = 1$ gives

$$c_1 = \frac{c - b}{a - b}, \qquad c_2 = \frac{a - c}{a - b},$$

and the same computation with d in place of c gives d_1, d_2. Hence

$$d_1c_2 = \frac{(d-b)(a-c)}{(a-b)^2}, \quad d_2c_1 = \frac{(a-d)(c-b)}{(a-b)^2},$$

and clearing the common factor $(a-b)^{-2}$ and the common sign gives $(A,B;C,D) = [\,(c-a)(d-b) : (c-b)(d-a)\,]$, as claimed.

■

Quick check. Proposition 22.7.8 insists that A, B and C be pairwise distinct, but puts no condition at all on D. Where in the proof is each of these used, and why is $[\,d_1c_2 : d_2c_1\,]$ a point of $\mathbb{P}^1(F)$ for every D?

Solution. $A \neq B$ makes (a,b) independent, hence a basis, which is what lets c and d be expanded at all. $C \neq A$ and $C \neq B$ are what force $c_2 \neq 0$ and $c_1 \neq 0$, which is what makes (c_1a, c_2b) a basis and lets us divide by c_1 and c_2. For D nothing is needed: $d \neq 0$, so $(d_1,d_2) \neq (0,0)$, and since $c_1c_2 \neq 0$ the pair (d_1c_2, d_2c_1) is again not $(0,0)$. So it names a point of $\mathbb{P}^1(F)$, whichever point D is.

Theorem 22.7.9. Invariance and Completeness of the Cross-Ratio

Let L and L' be projective lines over F.

(a) **(Invariance)** If $g\colon L \to L'$ is a projectivity and $A,B,C \in L$ are pairwise distinct, then for every $D \in L$, $(g(A),g(B);g(C),g(D)) = (A,B;C,D)$.

(b) **(Completeness)** Let $A,B,C \in L$ be pairwise distinct and $A',B',C' \in L'$ pairwise distinct, and let $D \in L$, $D' \in L'$. Then there is a projectivity $L \to L'$ carrying A,B,C,D to A',B',C',D' if and only if $(A,B;C,D) = (A',B';C',D')$.

Idea. Both halves come from the uniqueness in Theorem 22.7.5, with no computation. For (a), $f \circ g^{-1}$ normalizes the image triple, so it **is** the normalizing map of that triple; evaluate it at $g(D)$. For (b), let h be the unique projectivity matching the triples; by (a) it preserves the cross-ratio, and the cross-ratio determines the fourth point because the normalizing map is injective.

Proof.

(a) Let $f\colon L \to \mathbb{P}^1(F)$ be the projectivity of Definition 22.7.3 for (A,B,C). Since g is a bijection (Proposition 22.7.1), $g(A), g(B), g(C)$ are pairwise distinct, so the corresponding map $f'\colon L' \to \mathbb{P}^1(F)$ exists. The composition $f \circ g^{-1}\colon L' \to \mathbb{P}^1(F)$ is a projectivity (Proposition 22.7.1) and sends $g(A) \mapsto [1:0]$, $g(B) \mapsto [0:1]$, $g(C) \mapsto [1:1]$. By the uniqueness in Theorem 22.7.5, $f' = f \circ g^{-1}$. Hence

$$(g(A),g(B);g(C),g(D)) = f'(g(D)) = f(D) = (A,B;C,D).$$

(b) ($\Rightarrow$) is (a). ($\Leftarrow$) Let $h\colon L \to L'$ be the unique projectivity with $h(A) = A'$, $h(B) = B'$, $h(C) = C'$ (Theorem 22.7.5, $n = 1$). Let f' be the normalizing map for (A', B', C'). By (a),

$$f'(h(D)) = (A', B'; C', h(D)) = (A, B; C, D) = (A', B'; C', D') = f'(D').$$

Since f' is injective, $h(D) = D'$. This proves the theorem.

■

So four points on a line carry exactly one piece of data that survives every projectivity, and two quadruples are projectively equivalent precisely when that datum agrees.

Example. A cross-ratio, before and after

On $\mathbb{P}^1(\mathbb{Q})$, use the chart $[t : 1] \mapsto t$ and take the points with parameters $a = 0$, $b = 1$, $c = 3$, $d = 7$. Compute $(A, B; C, D)$. Then apply the projectivity $[\mathbf{A}]$ with $\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 1 & 3 \end{pmatrix}$ and compute the cross-ratio of the images.

Solution. By Proposition 22.7.8,

$$(A, B; C, D) = \big[\,(3-0)(7-1) : (3-1)(7-0)\,\big] = [18 : 14] = [9 : 7],$$

which is the scalar $9/7$ in the chart.

The matrix $\mathbf{A}$ has $\det \mathbf{A} = 3 - 2 = 1 \neq 0$, so $[\mathbf{A}]$ is a projectivity. It sends $[t : 1]$ to $[t + 2 : t + 3]$, which lies in the chart when $t \neq -3$ and is then the scalar $(t+2)/(t+3)$. The four parameters become

$$0 \mapsto \tfrac{2}{3}, \qquad 1 \mapsto \tfrac{3}{4}, \qquad 3 \mapsto \tfrac{5}{6}, \qquad 7 \mapsto \tfrac{9}{10}.$$

The images are not equally spaced in any sense, and the gaps have changed from $1, 2, 4$ to $\tfrac{1}{12}, \tfrac{1}{12}, \tfrac{1}{15}$. The cross-ratio, however, is

$$\Big[\big(\tfrac{5}{6} - \tfrac{2}{3}\big)\big(\tfrac{9}{10} - \tfrac{3}{4}\big) : \big(\tfrac{5}{6} - \tfrac{3}{4}\big)\big(\tfrac{9}{10} - \tfrac{2}{3}\big)\Big] = \Big[\tfrac{1}{40} : \tfrac{7}{360}\Big] = [9 : 7],$$

after multiplying both entries by 360. This is Theorem 22.7.9 (a) in one instance.

As a second check, run Proposition 22.7.8 in its first form with $\mathbf{a} = (0, 1)$ and $\mathbf{b} = (1, 1)$: solving gives $(c_1, c_2) = (-2, 3)$ and $(d_1, d_2) = (-6, 7)$, so $[\,d_1c_2 : d_2c_1\,] = [-18 : -14] = [9 : 7]$.

Warning. **The cross-ratio depends on the order of the four points.** It is a function of an **ordered** quadruple. Swapping the first two points inverts it: with the notation of Proposition 22.7.8, $(B, A; C, D) = [\,d_2c_1 : d_1c_2\,]$, the two entries exchanged. With $a, b, c, d = 0, 1, 3, 7$ as above, $(A, B; C, D) = [9 : 7]$ while $(B, A; C, D) = [7 : 9]$, and $9/7 \neq 7/9$. Writing "the cross-ratio of four points" without fixing an order names six different values in general.

22.7.5 Perspectivities

The classical source of projectivities is projection from a point, and it is worth seeing that the synthetic construction is the linear-algebraic one.

Proposition 22.7.10. Projection From a Point Is a Projectivity

Let $\dim V = 3$, let $L = \mathbb{P}(U)$ and $L' = \mathbb{P}(U')$ be lines in the projective plane $\mathbb{P}(V)$, and let $O = [o]$ be a point lying on neither. For $P \in L$, let $\pi(P)$ be the unique point in which the line $O \vee P$ meets L'. Then $\pi\colon L \to L'$ is well defined, and it is the projectivity induced by the restriction to U of the projection of V onto U' along $\operatorname{span}(o)$. In particular π preserves cross-ratio.

Idea. Everything is forced by dimensions in a 3-dimensional V. Since $o \notin U'$ and $\dim U' = 2$, we have $V = U' \oplus \operatorname{span}(o)$, so every vector splits uniquely as "a piece in U' plus a multiple of o". Throwing away the multiple of o is a linear map, and on U it is injective because $o \notin U$. That map is the projection, and the line $O \vee P$ is exactly where the discarded multiple of o lives.

Proof. Since $o \neq 0$ and $o \notin U'$, the sum $U' + \operatorname{span}(o)$ properly contains U', so its dimension exceeds 2 and it equals V. Since $\dim U' + \dim \operatorname{span}(o) = 2 + 1 = 3 = \dim(U' + \operatorname{span}(o))$, the sum is direct by Theorem 1.7.4, so $V = U' \oplus \operatorname{span}(o)$. Let $\rho\colon V \to V$ be the projection onto U' along $\operatorname{span}(o)$ (Theorem 4.5.1 (b)), so that

$$x = \rho(x) + t(x)\, o, \qquad \rho(x) \in U',\ t(x) \in F,$$

is the unique such splitting, and ρ is linear.

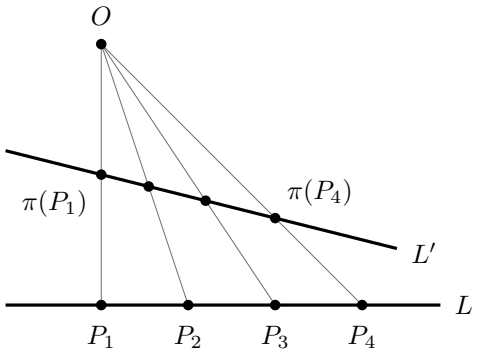
The restriction $T := \rho|_U\colon U \to U'$ is an isomorphism. If $u \in U$ and $\rho(u) = 0$, then $u = t(u)o$; since $o \notin U$, this forces $t(u) = 0$ and $u = 0$. So T is injective, and $\dim U = \dim U' = 2$ makes it an isomorphism (Corollary 2.3.2 (e)).

$[T] = \pi$. Let $P = [u]$ with $0 \neq u \in U$. Since $o \notin U$, the vector o is not a multiple of u, so $\operatorname{span}(u, o)$ has dimension 2. The projective subspace $\mathbb{P}(\operatorname{span}(u, o))$ contains O and P, and any projective subspace $\mathbb{P}(W)$ containing both has $u, o \in W$, hence $\operatorname{span}(u, o) \subseteq W$; so $O \vee P = \mathbb{P}(\operatorname{span}(u, o))$, a line. Now $Tu = u - t(u)o \in \operatorname{span}(u, o) \cap U'$, and $Tu \neq 0$, so $[Tu]$ lies on both $O \vee P$ and L'. These two lines are distinct, because O lies on the first and not on the second, so

$$\dim\big(\operatorname{span}(u, o) \cap U'\big) = 2 + 2 - \dim\big(\operatorname{span}(u, o) + U'\big) = 4 - 3 = 1$$

by Theorem 1.7.3, the sum being all of V since it properly contains U'. Hence the intersection point is unique and equals $[Tu]$. So π is well defined and $\pi = [T]$, a projectivity. Cross-ratio invariance is Theorem 22.7.9 (a).

■



projection from O carries L to L':
the four points are equally spaced on L and not on L',
but both quadruples have cross-ratio $[4:3]$

The two lines need not be distinct. If $L = L'$, then for $P \in L$ the line $O \vee P$ is not L, because O lies on the first and not on the second, so the two meet in exactly one point (Corollary 22.6.7) and that point is P: the construction gives $\pi = \mathrm{id}_L$. The classical statement takes the lines distinct, but the proof above never uses it.

Informally, this is what a photograph does. As a physical model rather than a theorem: a camera records, for each point of the scene, the point where the line from that point to the lens meets the film, and along any one line of the scene that is exactly the map π. So a photograph of a straight railway track distorts the spacing of the sleepers but not their cross-ratio. Four equally spaced sleepers, with parameters $0, 1, 2, 3$ along the track, have cross-ratio

$$\big[\,(2-0)(3-1) : (2-1)(3-0)\,\big] = [4:3],$$

and the four images on the film, bunched ever closer together toward the horizon, have cross-ratio $[4:3]$ as well. Equal spacing is not a projective notion; $[4:3]$ is.

22.7.6 Affine maps inside the projective group

The chapter began with affine geometry, and we can now say exactly how it sits inside projective geometry. By Theorem 22.6.2, $\mathbb{P}^n(F)$ is the disjoint union of the chart $\{x_0 \neq 0\}$, identified with affine n-space by

$$[x_0 : x_1 : \cdots : x_n] \longmapsto \Big(\frac{x_1}{x_0}, \ldots, \frac{x_n}{x_0}\Big),$$

and the hyperplane at infinity $H_\infty = \{[\mathbf{x}] : x_0 = 0\} = \mathbb{P}(\mathrm{span}(\mathbf{e}_1, \ldots, \mathbf{e}_n))$.

Theorem 22.7.11. Affine Maps Are the Projectivities Fixing Infinity

Let $n \geq 1$ and let g be a projectivity of $\mathbb{P}^n(F)$. Then $g(H_\infty) = H_\infty$ if and only if $g = [\mathbf{A}]$ for a matrix of the block form

$$\mathbf{A} = \begin{pmatrix} 1 & \mathbf{0}^\top \\ \mathbf{b} & \mathbf{M} \end{pmatrix}, \qquad \mathbf{b} \in F^n, \quad \mathbf{M} \in \mathrm{GL}_n(F).$$

In that case g maps the chart $\{x_0 \neq 0\}$ onto itself, and in the affine coordinates above it acts as

$$\mathbf{x} \longmapsto \mathbf{M}\mathbf{x} + \mathbf{b},$$

an invertible affine map with linear part $\mathbf{x} \mapsto \mathbf{M}\mathbf{x}$ (Definition 22.1.3, Theorem 22.1.5).

Conversely every invertible affine map of F^n arises from exactly one such $\mathbf{A}$.

Idea. "g fixes H_∞ as a set" says that the matrix maps $\operatorname{span}(\mathbf{e}_1, \dots, \mathbf{e}_n)$ into itself, which is a statement about the top entry of n columns. Invertibility then forces the remaining corner entry to be nonzero, and since a projectivity only knows its matrix up to a scalar, we may divide by that entry. What is left is the block matrix that Section 01 already introduced for affine maps — the same matrix, with the extra coordinate moved from last place to first.

Proof. Write $H = \operatorname{span}(\mathbf{e}_1, \dots, \mathbf{e}_n) = \{\mathbf{x} \in F^{n+1} : x_0 = 0\}$, so $H_\infty = \mathbb{P}(H)$, and let $g = [\mathbf{A}]$ with $\mathbf{A} \in \mathrm{GL}_{n+1}(F)$.

($\Rightarrow$) Suppose $g(H_\infty) = H_\infty$. By Proposition 22.7.2, $\mathbb{P}(\mathbf{A}H) = \mathbb{P}(H)$, and hence $\mathbf{A}H = H$: a nonzero $\mathbf{h} \in \mathbf{A}H$ has $[\mathbf{h}] \in \mathbb{P}(H)$, so $\mathbf{h} \in H$, and conversely. Thus $\mathbf{A}\mathbf{e}_j \in H$ for $j = 1, \dots, n$, which says the columns $1, \dots, n$ of $\mathbf{A}$ have zero entry in row 0. Writing column 0 as $(a, \mathbf{b})$ with $a \in F$ and $\mathbf{b} \in F^n$, and the lower right $n \times n$ block as $\mathbf{M}$,

$$\mathbf{A} = \begin{pmatrix} a & \mathbf{0}^\top \\ \mathbf{b} & \mathbf{M} \end{pmatrix}.$$

Here $\mathbf{M}$ is invertible: $\mathbf{A}$ restricts to an injective map $H \to H$, which is therefore surjective (Corollary 2.3.2 (e)), and that restriction is $\mathbf{M}$ in the basis $(\mathbf{e}_1, \dots, \mathbf{e}_n)$ of H. And $a \neq 0$: if $a = 0$ then column 0 of $\mathbf{A}$ lies in $H = \mathbf{A}H$, so $\mathbf{A}\mathbf{e}_0 = \mathbf{A}\mathbf{h}$ for some $\mathbf{h} \in H$, and injectivity would give $\mathbf{e}_0 \in H$, which is false. Replacing $\mathbf{A}$ by $a^{-1}\mathbf{A}$, which by Theorem 22.7.3 (a) induces the same projectivity, puts $\mathbf{A}$ in the stated form with 1 in the corner.

($\Leftarrow$) For $\mathbf{A}$ of the stated form and $\mathbf{x} \in H$, the first entry of $\mathbf{A}\mathbf{x}$ is $1 \cdot 0 + \mathbf{0}^\top \mathbf{x}' = 0$ where $\mathbf{x}'$ collects the last n entries, so $\mathbf{A}H \subseteq H$; and $\mathbf{A}$ is invertible, since $\mathbf{A}\mathbf{x} = \mathbf{0}$ forces $x_0 = 0$ and then $\mathbf{M}\mathbf{x}' = \mathbf{0}$, so $\mathbf{x}' = \mathbf{0}$. By Corollary 2.3.2 (e) again $\mathbf{A}H = H$, and Proposition 22.7.2 gives $g(H_\infty) = H_\infty$.

The action on the chart. Since g is a bijection of $\mathbb{P}^n(F)$ mapping H_∞ onto H_∞, it maps the complement $\{x_0 \neq 0\}$ onto itself. A point of the chart is $[1 : \mathbf{x}]$ with $\mathbf{x} \in F^n$, and

$$\mathbf{A}\begin{pmatrix} 1 \\ \mathbf{x} \end{pmatrix} = \begin{pmatrix} 1 \\ \mathbf{M}\mathbf{x} + \mathbf{b} \end{pmatrix},$$

whose affine coordinates are $\mathbf{M}\mathbf{x} + \mathbf{b}$. By Theorem 22.1.5 this is the affine map with linear part $\mathbf{x} \mapsto \mathbf{M}\mathbf{x}$ and value $\mathbf{b}$ at the origin, and it is invertible because $\mathbf{M}$ is.

Converse. Given an invertible affine map $\mathbf{x} \mapsto \mathbf{M}\mathbf{x} + \mathbf{b}$ of F^n, the matrix displayed above induces a projectivity fixing H_∞ and restricting to it. If two such matrices $\mathbf{A}$, $\mathbf{A}'$ induce the same projectivity, then $\mathbf{A}' = \lambda\mathbf{A}$ by Theorem 22.7.3 (a), and comparing the corner entries gives $\lambda = 1$, so $\mathbf{A}' = \mathbf{A}$. This proves the theorem.

■

So the affine group of F^n is the subgroup of $\mathrm{PGL}_{n+1}(F)$ stabilizing one hyperplane. Projective geometry is affine geometry with the hyperplane at infinity forgotten, and every affine notion that a projectivity can destroy — parallelism, ratios of lengths along a line, midpoints — is a notion that mentions H_∞. The cross-ratio never does, which is why it survives.

22.7.7 Exercises

22.7.7.1 A. Check your understanding

Exercise. A1

(a) State precisely when two isomorphisms $S, T\colon V \to W$ induce the same projectivity.

(b) How many points, and satisfying what condition, determine a projectivity of $\mathbb{P}^n(F)$ uniquely?

(c) True or false, with a reason: a projectivity of $\mathbb{P}^2(\mathbb{Q})$ fixing four points in general position is the identity.

(d) True or false, with a reason: det is a well-defined function on $\mathrm{PGL}_3(\mathbb{Q})$.

(e) What is $(A, B; C, B)$, and why?

Solution.

(a) Exactly when $S = \lambda T$ for some $\lambda \in F^\times$ (Theorem 22.7.3 (a)).

(b) $n+2$ points in general position, that is, $n+2$ points of which every $n+1$ have linearly independent representatives (Definition 22.7.2, Theorem 22.7.5).

(c) True. Here $n = 2$, so four points in general position are $n+2$ points in general position. The identity is one projectivity fixing them, and by the uniqueness in Theorem 22.7.5 it is the only one.

(d) False. The matrices $\mathbf{I}_3$ and $2\mathbf{I}_3$ induce the same projectivity but have determinants 1 and 8.

(e) $(A, B; C, B) = [0 : 1]$. By Definition 22.7.3 the cross-ratio is $f(D)$ with $D = B$, and $f(B) = [0 : 1]$ by the definition of f.

22.7.7.2 B. Practice

Exercise. B1: A projectivity of the line

Find a matrix inducing the projectivity of $\mathbb{P}^1(\mathbb{Q})$ that sends $[1 : 0]$, $[0 : 1]$, $[1 : 1]$ to $[1 : 1]$, $[1 : 2]$, $[2 : 3]$ respectively, and verify the three conditions.

Solution. The source triple is already normalized in the sense of Lemma 22.7.4: $(\mathbf{e}_0, \mathbf{e}_1)$ is a basis and $(1, 1) = \mathbf{e}_0 + \mathbf{e}_1$. Normalize the target: we need $\lambda_1(1, 1) + \lambda_2(1, 2) = (2, 3)$, that is, $\lambda_1 + \lambda_2 = 2$ and $\lambda_1 + 2\lambda_2 = 3$. Subtracting, $\lambda_2 = 1$, then $\lambda_1 = 1$. Both are nonzero, as Lemma 22.7.4 guarantees.

So the map sends $\mathbf{e}_0 \mapsto (1, 1)$ and $\mathbf{e}_1 \mapsto (1, 2)$, that is,

$$\mathbf{A} = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}, \qquad \det \mathbf{A} = 1 \neq 0.$$

Check: $\mathbf{A}\mathbf{e}_0 = (1, 1)$, so $[1 : 0] \mapsto [1 : 1]$; $\mathbf{A}\mathbf{e}_1 = (1, 2)$, so $[0 : 1] \mapsto [1 : 2]$; $\mathbf{A}(1, 1) = (2, 3)$, so $[1 : 1] \mapsto [2 : 3]$. By Theorem 22.7.5 no other projectivity does this, so every other correct answer is $\lambda\mathbf{A}$.

Exercise. B2: General position

Determine which of the following lists of points of $\mathbb{P}^2(\mathbb{Q})$ are in general position. Justify your answer.

(a) $[1:0:0]$, $[1:1:0]$, $[1:1:1]$, $[0:1:1]$.
(b) $[1:0:0]$, $[0:1:0]$, $[1:2:0]$, $[0:0:1]$.
(c) $[1:1:1]$, $[1:2:4]$, $[1:3:9]$, $[1:4:16]$.

Solution. Here $n = 2$, so general position for four points means every triple of representatives is independent, tested by a 3×3 determinant.

(a) The four determinants, in the order of the triples $(1,2,3)$, $(1,2,4)$, $(1,3,4)$, $(2,3,4)$, are

$$\det \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix} = 1, \quad \det \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix} = 1,$$

$$\det \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{pmatrix} = 0, \quad \det \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \end{pmatrix} = -1.$$

The third is zero, so the list is **not** in general position: $[1:0:0]$, $[1:1:1]$, $[0:1:1]$ are collinear, since $(1,1,1) = (1,0,0) + (0,1,1)$.

(b) Not in general position: $(1,2,0) = (1,0,0) + 2(0,1,0)$, so the first three points are collinear.

(c) In general position. The four representatives are $(1, t, t^2)$ for $t = 1, 2, 3, 4$, so each triple gives a Vandermonde matrix, whose determinant is $\prod_{i<j}(t_j - t_i)$ (Theorem 7.8.1, and Corollary 7.8.2 for the conclusion). Explicitly, $t = 1, 2, 3$ gives $1 \cdot 2 \cdot 1 = 2$; $t = 1, 2, 4$ gives $1 \cdot 3 \cdot 2 = 6$; $t = 1, 3, 4$ gives $2 \cdot 3 \cdot 1 = 6$; $t = 2, 3, 4$ gives $1 \cdot 2 \cdot 1 = 2$. All are nonzero, so every triple is independent.

Exercise. B3: Cross-ratio

In $\mathbb{P}^1(\mathbb{Q})$ with the chart $[t:1] \mapsto t$, let A, B, C, D have parameters $1, 2, 4, 8$.

(a) Compute $(A, B; C, D)$.

(b) Compute the images of the four points under the projectivity induced by $\begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$, and recompute the cross-ratio.

Solution.

(a) By Proposition 22.7.8,

$$(A, B; C, D) = \big[\,(4-1)(8-2) : (4-2)(8-1)\,\big] = [18:14] = [9:7].$$

(b) The matrix has determinant 1, and sends $[t:1]$ to $[t:t+1]$, which in the chart is

$t/(t+1)$. The four parameters become $\frac{1}{2}, \frac{2}{3}, \frac{4}{5}, \frac{8}{9}$. Then

$$\frac{4}{5}-\frac{1}{2}=\frac{3}{10}, \quad \frac{8}{9}-\frac{2}{3}=\frac{2}{9},$$
$$\frac{4}{5}-\frac{2}{3}=\frac{2}{15}, \quad \frac{8}{9}-\frac{1}{2}=\frac{7}{18},$$

so the cross-ratio is $\left[\frac{3}{10}\cdot\frac{2}{9} : \frac{2}{15}\cdot\frac{7}{18}\right] = \left[\frac{1}{15} : \frac{7}{135}\right] = [9:7]$, after multiplying both entries by 135. This agrees with (a), as Theorem 22.7.9 (a) requires.

22.7.7.3 C. Going deeper

Exercise. C1: Sharp triple transitivity, and a count

(a) Prove that for any two triples of pairwise distinct points of $\mathbb{P}^1(F)$ there is exactly one projectivity carrying the first to the second.

(b) Hence compute $|\mathrm{PGL}_2(\mathbb{F}_q)|$ for a finite field $\mathbb{F}_q$, and check your answer against $q=2$ by listing the projectivities of $\mathbb{P}^1(\mathbb{F}_2)$.

Solution.

(a) On a line, $n=1$, so $n+2=3$ and general position for three points means pairwise distinct (Definition 22.7.2). The statement is Theorem 22.7.5.

(b) By (a), $\mathrm{PGL}_2(\mathbb{F}_q)$ acts on the set of ordered triples of pairwise distinct points of $\mathbb{P}^1(\mathbb{F}_q)$ in such a way that exactly one group element carries any given triple to any other. Fixing one triple, the map $g \mapsto (g(A), g(B), g(C))$ is therefore a bijection from $\mathrm{PGL}_2(\mathbb{F}_q)$ onto the set of such triples. Now $\mathbb{P}^1(\mathbb{F}_q)$ has $q+1$ points: there are q^2-1 nonzero vectors in $\mathbb{F}_q^2$, and each point has exactly $q-1$ representatives, so the count is $(q^2-1)/(q-1) = q+1$. Hence

$$|\mathrm{PGL}_2(\mathbb{F}_q)| = (q+1)\,q\,(q-1).$$

For $q=2$ this is $3\cdot 2\cdot 1 = 6$. And indeed $\mathbb{P}^1(\mathbb{F}_2)$ has the three points $[1:0]$, $[0:1]$, $[1:1]$; by (a) every permutation of them is realized by exactly one projectivity, giving $3! = 6$. (Consistently, $\mathrm{GL}_2(\mathbb{F}_2)$ has 6 elements and the scalars are only $\mathbf{I}_2$, so Π of Theorem 22.7.3 is injective here.)

Exercise. C2: Rearranging the four points

Let A, B, C be pairwise distinct points of a projective line and D a fourth point.

(a) Prove that $(B, A; C, D)$ is obtained from $(A, B; C, D) = [\lambda : \mu]$ by exchanging the two entries.

(b) Suppose all four points lie in a chart, with parameters a, b, c, d, and that $(A, B; C, D)$ is the scalar χ. Prove that $(A, C; B, D) = 1-\chi$.

Solution.

(a) Use Proposition 22.7.8 with the representatives a, b and the expansions $c = c_1 a + c_2 b$, $d = d_1 a + d_2 b$. Computing $(B, A; C, D)$ means expanding in the basis (b, a) instead,

which exchanges the roles of the two coordinates: the coefficients of c are now (c_2, c_1) and those of d are (d_2, d_1). The formula gives $[\,d_2c_1 : d_1c_2\,]$, the entries of $[\,d_1c_2 : d_2c_1\,]$ exchanged.

(b) By Proposition 22.7.8 in the chart,

$$\chi = \frac{(c-a)(d-b)}{(c-b)(d-a)}, \qquad (A, C; B, D) = \frac{(b-a)(d-c)}{(b-c)(d-a)}.$$

Put both over the common denominator $(c-b)(d-a)$, noting $(b-c) = -(c-b)$:

$$\chi + (A, C; B, D) = \frac{(c-a)(d-b) - (b-a)(d-c)}{(c-b)(d-a)}.$$

Expand the numerator:

$$\begin{aligned}(c-a)(d-b) - (b-a)(d-c) &= cd - cb - ad + ab - bd + bc + ad - ac \\ &= cd + ab - bd - ac = (c-b)(d-a).\end{aligned}$$

Hence $\chi + (A, C; B, D) = 1$, as claimed.

Exercise. C3: Cross-ratio characterizes projectivities of a line

Let L and L' be projective lines over F and let $g\colon L \to L'$ be an **injection** such that

$$\big(g(A), g(B); g(C), g(D)\big) = (A, B; C, D)$$

for all $A, B, C, D \in L$ with A, B, C pairwise distinct. Prove that g is a projectivity.

Hint: pin g down on three points first.

Solution. Write $L = \mathbb{P}(U)$ with $\dim U = 2$ and pick a basis (u_1, u_2) of U; the three points $[u_1]$, $[u_2]$, $[u_1 + u_2]$ are pairwise distinct, since no two of $u_1, u_2, u_1 + u_2$ are proportional. So we may choose pairwise distinct $A, B, C \in L$. Since g is injective, $g(A), g(B), g(C)$ are pairwise distinct. By Theorem 22.7.5 with $n = 1$ there is a unique projectivity $h\colon L \to L'$ with $h(A) = g(A)$, $h(B) = g(B)$, $h(C) = g(C)$.

Let $D \in L$. By the hypothesis on g and by Theorem 22.7.9 (a) applied to h,

$$\big(g(A), g(B); g(C), g(D)\big) = (A, B; C, D) = \big(h(A), h(B); h(C), h(D)\big).$$

The two triples on the outside are the same, namely $(g(A), g(B), g(C))$. Writing f' for the normalizing projectivity of that triple (Definition 22.7.3), the displayed equality reads $f'(g(D)) = f'(h(D))$, and f' is injective, so $g(D) = h(D)$. As D was arbitrary, $g = h$, a projectivity.

Note where each hypothesis went: injectivity of g was needed only to make $g(A), g(B), g(C)$ distinct, so that the cross-ratio hypothesis is applicable at all. Surjectivity was never assumed, and comes out at the end: $g = h$ and h is a bijection (Proposition 22.7.1).

22.8 Duality

Two facts about the projective plane sit side by side. Two distinct points lie on exactly one line; two distinct lines meet in exactly one point. Read either sentence aloud with "point" and "line" exchanged and you get the other. This is not a coincidence and it is not a philosophy: Chapter 5 already built the machine that produces it. A hyperplane of $\mathbb{P}(V)$ is the set of points killed by a linear functional, the functional is determined by the hyperplane up to a scalar, and so the hyperplanes of $\mathbb{P}(V)$ are themselves the points of a projective space, namely $\mathbb{P}(V^*)$. This section turns that observation into an inclusion-reversing dictionary between the projective subspaces of $\mathbb{P}(V)$ and those of $\mathbb{P}(V^*)$, gets the dimension bookkeeping right at both ends, and says exactly how much of the "duality principle" we are entitled to claim.

Throughout, F is an arbitrary field and V is a finite-dimensional vector space over F with $\dim V = n + 1 \geq 1$, so that $\mathbb{P}(V)$ has dimension n (Definition 22.6.1). No hypothesis on the characteristic is needed anywhere in this section, and no analysis is used. We keep the two extreme projective subspaces in play: $\mathbb{P}(\{0\}) = \emptyset$, of dimension -1, and $\mathbb{P}(V)$ itself, of dimension n. The correspondence below exchanges them, and it is precisely at those two ends that a duality argument is easiest to get wrong.

22.8.1 Projective subspaces remember their subspaces

Everything rests on being able to move between a projective subspace $\mathbb{P}(U)$ and the subspace U it came from. That is the content of a short lemma, proved in Section 6 and recalled here because almost every proof below uses it.

Lemma 22.8.1. Passing Between U and $\mathbb{P}(U)$

Let U, U_1, U_2 be subspaces of V.

(a) $\mathbb{P}(U_1) \subseteq \mathbb{P}(U_2)$ if and only if $U_1 \subseteq U_2$. In particular $U \mapsto \mathbb{P}(U)$ is injective, so a projective subspace determines its subspace.

(b) $\mathbb{P}(U_1 \cap U_2) = \mathbb{P}(U_1) \cap \mathbb{P}(U_2)$.

(c) $\mathbb{P}(U_1) \vee \mathbb{P}(U_2) = \mathbb{P}(U_1 + U_2)$, where $\vee$ is the join, the smallest projective subspace containing both (Proposition 22.6.5).

Proof. Part (a) is Lemma 22.6.4, and parts (b) and (c) are Proposition 22.6.5 (a) and (b), the latter being where $\vee$ was defined. ∎

One consequence will be used repeatedly: if $S \subseteq T$ are projective subspaces of the same finite dimension, then $S = T$. Indeed $U_S \subseteq U_T$ by (a) and $\dim U_S = \dim U_T$, so $U_S = U_T$ by Theorem 1.6.8.

22.8.2 The points of the dual space are the hyperplanes

A **hyperplane** of $\mathbb{P}(V)$ is a projective subspace of dimension $n-1$, that is, $\mathbb{P}(H)$ for a hyperplane H of V in the sense of Definition 5.5.2. In $\mathbb{P}^2$ these are the lines; in $\mathbb{P}^3$, the planes.

Now the hook. A hyperplane of V is the kernel of a nonzero functional (Theorem 5.5.3), and the functional is determined by its kernel up to a scalar, which is the injectivity half of the next proposition. "A nonzero vector of V^*, up to a nonzero scalar" is exactly the description of a point of $\mathbb{P}(V^*)$. So the hyperplanes of $\mathbb{P}(V)$ are not merely a set: they are the points of a projective space of their own, of the same dimension.

The dual projective space is the space whose points are the hyperplanes of the original.

Definition 22.8.1. Dual Projective Space

Let V be a finite-dimensional vector space over F with $\dim V = n+1 \geq 1$. The **dual projective space** of $\mathbb{P}(V)$ is the projective space

$$\mathbb{P}(V^*)$$

of the dual space V^* (Definition 5.1.2). A point of $\mathbb{P}(V^*)$ is a class $[\varphi]$ with $\varphi \in V^*$ **nonzero**, and $[\varphi] = [\psi]$ exactly when $\psi = \lambda\varphi$ for some $\lambda \in F \setminus \{0\}$.

Its dimension is $\dim V^* - 1 = n$, by Corollary 5.1.4: *the dual projective space has the same dimension as the original.* That is worth pausing on, because the correspondence we are about to build reverses dimensions inside a space of unchanged size.

Nonexample by minimal change. Take $\varphi = 0$. It is a perfectly good element of V^*, but $[0]$ is not a point of $\mathbb{P}(V^*)$ — the definition excludes the zero vector, as Definition 22.6.1 does. And indeed $\ker 0 = V$, which is not a hyperplane: it has codimension 0, not 1 (Definition 5.5.2). The clause "nonzero" is exactly what keeps the following bijection honest.

Proposition 22.8.2. Points of the Dual Are Hyperplanes

Let $\dim V = n+1 \geq 1$. The map

$$[\varphi] \longmapsto \mathbb{P}(\ker \varphi)$$

is a well-defined bijection from $\mathbb{P}(V^*)$ onto the set of hyperplanes of $\mathbb{P}(V)$.

Idea. Three checks. Well defined: proportional functionals have the same kernel. Surjective: a hyperplane $\mathbb{P}(H)$ has $\dim H^0 = 1$ by the annihilator count, so it is cut out by some nonzero functional. Injective: two functionals with the same kernel H both lie in the one-dimensional space H^0, so they are proportional. The counting does all the work.

Proof. Well defined. If $\varphi \neq 0$ then $\ker \varphi$ is a hyperplane of V by Theorem 5.5.3, so $\mathbb{P}(\ker \varphi)$ is a projective subspace of dimension $\dim \ker \varphi - 1 = n-1$, a hyperplane of $\mathbb{P}(V)$. If $[\psi] = [\varphi]$, then $\psi = \lambda\varphi$ with $\lambda \neq 0$, and $\psi(v) = 0$ if and only if $\varphi(v) = 0$, so $\ker \psi = \ker \varphi$.

Surjective. Let $\mathbb{P}(H)$ be a hyperplane of $\mathbb{P}(V)$, so $\dim H = n$. By Theorem 5.2.2, $\dim H^0 = (n+1) - n = 1$, so there is a nonzero $\varphi \in H^0$. Then $H \subseteq \ker \varphi$ by Definition 5.2.1, and both have dimension n — the second by Theorem 5.5.3 and Theorem 4.4.5 — so $H = \ker \varphi$ by Theorem 1.6.8.

Injective. Suppose $\ker \varphi = \ker \psi = H$ with $\varphi, \psi \neq 0$. Then $\varphi, \psi \in H^0$, which has dimension 1 as above. Two nonzero vectors of a one-dimensional space are proportional, so $\psi = \lambda\varphi$ with $\lambda \neq 0$, that is, $[\psi] = [\varphi]$.

■

In coordinates this is completely concrete. Take $V = F^{n+1}$ with coordinates $x_0, \dots, x_n$. By Theorem 5.1.1 every functional is $\varphi_{\mathbf{a}}(\mathbf{x}) = a_0x_0 + \cdots + a_nx_n$ for a unique $\mathbf{a} \in F^{n+1}$, and this identification is a linear isomorphism $F^{n+1} \to (F^{n+1})^*$. So a point of the dual plane is a class $[a_0 : \cdots : a_n]$, and the hyperplane it names is

$$\{\, [\mathbf{x}] \in \mathbb{P}^n(F) \;:\; a_0x_0 + \cdots + a_nx_n = 0 \,\}.$$

The equation is homogeneous, so it does not matter which representative $\mathbf{x}$ we substitute; and scaling $\mathbf{a}$ does not change the solution set. Both scalings are exactly the ones the projective spaces divide out.

Example. A line and its point

In $\mathbb{P}^2(\mathbb{Q})$, find the point of the dual plane naming the line through $[1:1:0]$ and $[0:1:1]$. Then find the point where the lines $x_0 - x_1 + x_2 = 0$ and $x_0 + x_2 = 0$ meet.

Solution. The line. Its subspace is $U = \operatorname{span}\big((1,1,0),(0,1,1)\big)$, of dimension 2. By Theorem 5.2.2, $\dim U^0 = 3 - 2 = 1$, so one nonzero functional does it. We need $a_0 + a_1 = 0$ and $a_1 + a_2 = 0$, whose solutions are the multiples of $\mathbf{a} = (1,-1,1)$. Check: $1 - 1 + 0 = 0$ and $0 - 1 + 1 = 0$. So the line is $x_0 - x_1 + x_2 = 0$, named by the dual point $[1:-1:1]$.

The meet. Dually, we need $\mathbf{x}$ with $x_0 - x_1 + x_2 = 0$ and $x_0 + x_2 = 0$. Subtracting, $x_1 = 0$; then $x_0 = -x_2$. The solution space is spanned by $(-1,0,1)$, so the two lines meet in the single point $[-1:0:1]$. Check: $-1 - 0 + 1 = 0$ and $-1 + 1 = 0$.

The two computations are the same computation, run in V and in V^*. That is the whole of duality in one example.

Warning. **There is no canonical duality between $\mathbb{P}(V)$ and itself.** The rule "the point $[a_0 : a_1 : a_2]$ names the line $a_0x_0 + a_1x_1 + a_2x_2 = 0$" looks like a self-duality of $\mathbb{P}^2(F)$, but it depends on the coordinates. In $\mathbb{P}^2(\mathbb{Q})$ the rule assigns to $[1:0:1]$ the line $x_0 + x_2 = 0$. Change coordinates by $y_0 = x_0$, $y_1 = x_1$, $y_2 = 2x_2$. The same point is $[1:0:2]$ in the new coordinates, and the same rule now assigns to it the line $y_0 + 2y_2 = 0$, that is, $x_0 + 4x_2 = 0$. These are different lines: the point $[1:0:-1]$ lies on the first, since $1 - 1 = 0$, and not on the second, since $1 - 4 = -3 \neq 0$. What is canonical is the correspondence between $\mathbb{P}(V)$ and $\mathbb{P}(V^*)$, not one between $\mathbb{P}(V)$ and itself. Identifying the two requires a choice — a basis, or a nondegenerate form, which is how the pole and polar of a conic will be built in Section 10 of this chapter.

22.8.3 The duality correspondence

We now extend $[\varphi] \mapsto \mathbb{P}(\ker \varphi)$ from hyperplanes to all projective subspaces. The right tool is the annihilator, and we keep Chapter 5's notation: for a subspace $U \subseteq V$, $U^0 = \{\varphi \in V^* : \varphi(u) = 0 \text{ for all } u \in U\}$ (Definition 5.2.1).

For a projective subspace $S = \mathbb{P}(U)$ of $\mathbb{P}(V)$, define

$$S^0 := \mathbb{P}(U^0) \subseteq \mathbb{P}(V^*).$$

This is well defined because S determines U, by Lemma 22.8.1 (a).

Before the theorem, the dimension bookkeeping, because it is the step where such a correspondence is usually stated wrongly. Projective dimension is one less than linear dimension. So from $\dim U + \dim U^0 = n + 1$ we get

$$(\dim S + 1) + (\dim S^0 + 1) = n + 1, \qquad \text{that is} \qquad \dim S^0 = n - 1 - \dim S.$$

Test it at both ends. If $S = \emptyset$, so $U = \{0\}$ and $\dim S = -1$, the formula gives $\dim S^0 = n$, and indeed $\{0\}^0 = V^*$, so $S^0 = \mathbb{P}(V^*)$. If $S = \mathbb{P}(V)$, with $\dim S = n$, the formula gives -1, and indeed $V^0 = \{0\}$, so $S^0 = \emptyset$. A point ($\dim S = 0$) goes to a hyperplane ($\dim S^0 = n - 1$), and a hyperplane goes to a point, agreeing with Proposition 22.8.2. The formula is symmetric in the sense that $d \mapsto n - 1 - d$ is its own inverse on $\{-1, 0, \ldots, n\}$.

Theorem 22.8.3. The Duality Correspondence

Let $\dim V = n + 1 \geq 1$, and let S, T be projective subspaces of $\mathbb{P}(V)$, the empty set and $\mathbb{P}(V)$ included. Then:

(a) $\dim S^0 = n - 1 - \dim S$;
(b) $S \subseteq T$ if and only if $T^0 \subseteq S^0$;
(c) $(S \vee T)^0 = S^0 \cap T^0$ and $(S \cap T)^0 = S^0 \vee T^0$;
(d) $S \mapsto S^0$ is a bijection from the set of projective subspaces of $\mathbb{P}(V)$ onto the set of projective subspaces of $\mathbb{P}(V^*)$.

Idea. Every part is a Chapter 5 statement on annihilators, translated through Lemma 22.8.1, with the dimension shifted by one at each end. The only part that needs thought is the reverse implication in (b) and the surjectivity in (d): both go up a level to V^{**} and come back with Corollary 5.3.4, which is available because V is finite-dimensional.

Proof. Write $S = \mathbb{P}(U)$ and $T = \mathbb{P}(U')$, which is legitimate and unambiguous by Lemma 22.8.1 (a).

(a) By Theorem 5.2.2, $\dim U^0 = (n + 1) - \dim U$. Hence

$$\dim S^0 = \dim U^0 - 1 = n - \dim U = n - (\dim S + 1) = n - 1 - \dim S.$$

(b) ($\Rightarrow$) If $S \subseteq T$, then $U \subseteq U'$ by Lemma 22.8.1 (a), so $(U')^0 \subseteq U^0$ by Theorem 5.2.3 (a), so $T^0 \subseteq S^0$ by the same lemma. ($\Leftarrow$) Suppose $T^0 \subseteq S^0$, so $(U')^0 \subseteq U^0$. Applying Theorem 5.2.3 (a) inside V^* gives $U^{00} \subseteq (U')^{00}$. By Corollary 5.3.4, $U^{00} = \mathrm{ev}_V(U)$ and $(U')^{00} = \mathrm{ev}_V(U')$, so $\mathrm{ev}_V(U) \subseteq \mathrm{ev}_V(U')$. Since ev_V is injective (Theorem 5.3.2), $U \subseteq U'$, and $S \subseteq T$.

(c) By Lemma 22.8.1 (c), $S \vee T = \mathbb{P}(U + U')$, and $(U+U')^0 = U^0 \cap (U')^0$ by Theorem 5.2.3 (c). Applying $\mathbb{P}(\cdot)$ and Lemma 22.8.1 (b),

$$(S \vee T)^0 = \mathbb{P}(U^0 \cap (U')^0) = \mathbb{P}(U^0) \cap \mathbb{P}((U')^0) = S^0 \cap T^0.$$

For the second identity, $S \cap T = \mathbb{P}(U \cap U')$ by Lemma 22.8.1 (b), and $(U \cap U')^0 = U^0 + (U')^0$ by Theorem 5.2.3 (d), which needs V finite-dimensional. Applying $\mathbb{P}(\cdot)$ and part (c) of the same lemma gives $(S \cap T)^0 = S^0 \vee T^0$.

(d) *Injective.* If $S^0 = T^0$ then $U^0 = (U')^0$ by Lemma 22.8.1 (a), so $U = U'$ by Corollary 5.3.5, so $S = T$.

Surjective. Let $R = \mathbb{P}(W)$ be a projective subspace of $\mathbb{P}(V^*)$, where W is a subspace of V^*. Put

$$U := \mathrm{ev}_V^{-1}(W^0) = \{v \in V : \varphi(v) = 0 \text{ for every } \varphi \in W\},$$

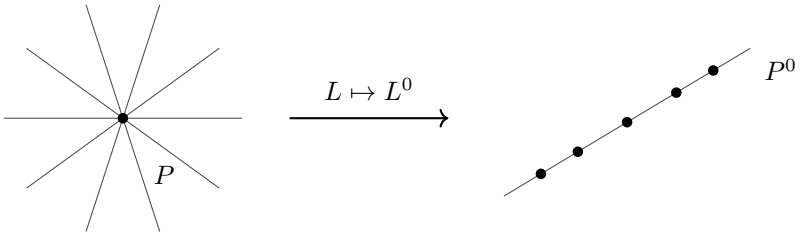
a subspace of V: it is the intersection of the kernels $\ker \varphi$ over $\varphi \in W$, each a subspace by Theorem 2.2.1, and an intersection of subspaces is a subspace by Theorem 1.2.2. Since ev_V is surjective (Theorem 5.3.2), $\mathrm{ev}_V(U) = W^0$. By Corollary 5.3.4, $(U^0)^0 = U^{00} = \mathrm{ev}_V(U) = W^0$. Now U^0 and W are subspaces of the finite-dimensional space V^* with the same annihilator, so $U^0 = W$ by Corollary 5.3.5 applied in V^*. Hence $\mathbb{P}(U)^0 = \mathbb{P}(W) = R$. This proves the theorem.

■

Quick check. Let S be a line in $\mathbb{P}^3(F)$. What is $\dim S^0$, and what kind of object is S^0? What if S is a point?

Solution. Here $n = 3$ and $\dim S = 1$, so $\dim S^0 = 3 - 1 - 1 = 1$ by Theorem 22.8.3 (a): S^0 is again a line, this time in the dual space $\mathbb{P}((F^4)^*)$. Lines in $\mathbb{P}^3$ are self-dual in dimension, which is one reason they are the interesting objects there. If instead S is a point, $\dim S = 0$ and $\dim S^0 = 3 - 1 - 0 = 2$: a plane in the dual space. Concretely, it is the set of all planes of $\mathbb{P}^3$ through that point.

Remark. The last sentence of the solution is the general reading of the correspondence. Since $S \subseteq P$ if and only if $P^0 \subseteq S^0$, the hyperplanes containing S are exactly the points of S^0: *a projective subspace is named, on the dual side, by the family of hyperplanes that contain it.* For a point, that family is a hyperplane's worth of hyperplanes.



in $\mathbb{P}(V)$ in $\mathbb{P}(V^*)$

each line L through P is one point L^0 of the dual plane,
and those points fill the line P^0

22.8.4 Dualizing twice

Applying the correspondence twice lands in $\mathbb{P}(V^{**})$, not back in $\mathbb{P}(V)$. Chapter 5 already told us what to do: the evaluation map is a canonical isomorphism $V \to V^{**}$, and it induces a projectivity.

Theorem 22.8.4. Dualizing Twice Returns the Original

Let $\dim V = n+1 \geq 1$. The evaluation map $\mathrm{ev}_V \colon V \to V^{**}$ is an isomorphism (Theorem 5.3.2), so it induces a projectivity $[\mathrm{ev}_V] \colon \mathbb{P}(V) \to \mathbb{P}(V^{**})$ (Definition 22.7.1). For every projective subspace S of $\mathbb{P}(V)$,

$$S^{00} = [\mathrm{ev}_V](S),$$

so that under the identification $\mathbb{P}(V) = \mathbb{P}(V^{**})$ given by $[\mathrm{ev}_V]$ we have $S^{00} = S$. Moreover the inverse of the bijection of Theorem 22.8.3 (d) is $R \mapsto [\mathrm{ev}_V]^{-1}(R^0)$.

Proof. Write $S = \mathbb{P}(U)$. By the definition of the correspondence applied twice, $S^{00} = \mathbb{P}((U^0)^0) = \mathbb{P}(\mathrm{ev}_V(U))$, using Corollary 5.3.4. By Proposition 22.7.2, $\mathbb{P}(\mathrm{ev}_V(U)) = [\mathrm{ev}_V](\mathbb{P}(U)) = [\mathrm{ev}_V](S)$.

For the last statement, let R be a projective subspace of $\mathbb{P}(V^*)$ and put $S \coloneqq [\mathrm{ev}_V]^{-1}(R^0)$, a projective subspace of $\mathbb{P}(V)$ by Proposition 22.7.2 applied to $[\mathrm{ev}_V]^{-1}$. Then $[\mathrm{ev}_V](S) = R^0$, so by the first part $(S^0)^0 = S^{00} = R^0$. Both S^0 and R are projective subspaces of $\mathbb{P}(V^*)$, and Theorem 22.8.3 (d) applied to the space V^* says that the correspondence there is injective, so $S^0 = R$.

■

Duality is also compatible with the maps of Section 07, which is what makes it more than a coordinate trick.

Proposition 22.8.5. Duality and Projectivities

Let $T \colon V \to W$ be an isomorphism of $(n+1)$-dimensional spaces, with dual map $T' \colon W^* \to V^*$ (Definition 5.4.1), which is an isomorphism by Theorem 5.4.1 (d). Then for every projective subspace S of $\mathbb{P}(V)$,

$$\big([T](S)\big)^0 = [(T')^{-1}](S^0).$$

Proof. Write $S = \mathbb{P}(U)$, so $[T](S) = \mathbb{P}(T(U))$ by Proposition 22.7.2. For $\psi \in W^*$,

$$\psi \in T(U)^0 \iff \psi(Tu) = 0 \text{ for all } u \in U \iff (T'\psi)(u) = 0 \text{ for all } u \in U,$$

the second step by Definition 5.4.1, and the right-hand condition says $T'\psi \in U^0$. Hence $T(U)^0 = (T')^{-1}(U^0)$. Applying $\mathbb{P}(\cdot)$ and Proposition 22.7.2 to the isomorphism $(T')^{-1}$ gives the claim.

■

22.8.5 The duality principle

We can now state precisely what "every theorem has a dual" amounts to in this book.

Theorem 22.8.6. The Duality Principle

Let $\dim V = n + 1 \geq 1$, and let $S \mapsto S^0$ be the correspondence of Theorem 22.8.3, from the projective subspaces of the n-dimensional space $\mathbb{P}(V)$ onto those of the n-dimensional space $\mathbb{P}(V^*)$. Let R, S, T be projective subspaces of $\mathbb{P}(V)$ and let d be an integer. Then:

(a) $S \subseteq T$ if and only if $T^0 \subseteq S^0$; and $S = T$ if and only if $S^0 = T^0$;

(b) $\dim S = d$ if and only if $\dim S^0 = n - 1 - d$;

(c) $R = S \vee T$ if and only if $R^0 = S^0 \cap T^0$;

(d) $R = S \cap T$ if and only if $R^0 = S^0 \vee T^0$;

(e) as S runs over all projective subspaces of $\mathbb{P}(V)$, the subspace S^0 runs exactly once over all projective subspaces of $\mathbb{P}(V^*)$.

Proof.

(a) The first claim is Theorem 22.8.3 (b), and the second follows by applying it in both directions, or from injectivity in Theorem 22.8.3 (d).

(b) By Theorem 22.8.3 (a), $\dim S^0 = n - 1 - \dim S$. So $\dim S = d$ gives $\dim S^0 = n - 1 - d$; conversely $\dim S^0 = n - 1 - d$ gives $n - 1 - \dim S = n - 1 - d$, hence $\dim S = d$.

(c) ($\Rightarrow$) is Theorem 22.8.3 (c). ($\Leftarrow$) If $R^0 = S^0 \cap T^0 = (S \vee T)^0$, then $R = S \vee T$ by the equality clause of (a).

(d) The same argument with the second identity of Theorem 22.8.3 (c).

(e) This is Theorem 22.8.3 (d).

■

How the principle is used, and what we have not proved. Every clause of Theorem 22.8.6 is an "if and only if", and by (e) the correspondence is a bijection between the two supplies of subspaces. So a proved statement about $\mathbb{P}(V)$ whose hypotheses and conclusion are built **only** from the relations $\subseteq$ and $=$, the operations $\vee$ and $\cap$, and the dimensions of the subspaces involved, can be translated clause by clause: reverse every inclusion, exchange $\vee$ with $\cap$, and replace every dimension d by $n - 1 - d$. The translated statement then holds in $\mathbb{P}(V^*)$; and since every $(n+1)$-dimensional space W satisfies $W \cong (W^*)^*$ (Theorem 5.3.2), a statement true

in $\mathbb{P}(V^*)$ for **every** V is true in $\mathbb{P}(W)$ for every W, because projectivities preserve all of this structure (Proposition 22.7.2).

This is a recipe, carried out below on the statements this chapter makes, and each carrying out is a proof. It is **not** a theorem here. To make "for every statement built only from ..." itself a theorem, one has to define a formal language of incidence statements and induct on the way a statement is built, and this book has set up no such language. So we claim the principle only as a working method, and we prove each dual statement we use.

Corollary 22.8.7. Two Points, Two Lines

Let $\dim V = 3$, so that $\mathbb{P}(V)$ is a projective plane.

(a) Two distinct points of $\mathbb{P}(V)$ lie on exactly one line.

(b) Two distinct lines of $\mathbb{P}(V)$ meet in exactly one point.

Idea. Prove (a) by hand: two distinct points have independent representatives, so their span is a plane in V and a line in $\mathbb{P}(V)$. Then get (b) from (a) by the dictionary, in $\mathbb{P}(V^*)$, which is a projective plane of its own. Section 06 already proved both statements directly from the dimension formula (Proposition 22.6.8, Corollary 22.6.7). The point of doing (b) again this way is that duality delivers it with no geometry at all: every step is a clause of Theorem 22.8.6 with $n = 2$.

Proof.

(a) Let $P = [v] \neq Q = [w]$. The list (v, w) is independent: otherwise w would be a scalar multiple of v, giving $P = Q$. So $\operatorname{span}(v, w)$ has dimension 2, and by Lemma 22.8.1 (c), $P \vee Q = \mathbb{P}(\operatorname{span}(v, w))$, a projective subspace of dimension 1, that is, a line containing both points. If L is any line containing P and Q, then $P \vee Q \subseteq L$ because the join is the smallest such subspace, and both have dimension 1, so $L = P \vee Q$ by the consequence noted after Lemma 22.8.1.

(b) Here $n = 2$. Let $L_1 \neq L_2$ be lines of $\mathbb{P}(V)$. By Theorem 22.8.6 (b), $\dim L_i^0 = 2-1-1 = 0$, so L_1^0 and L_2^0 are points of $\mathbb{P}(V^*)$, and they are distinct by the equality clause of Theorem 22.8.6 (a). The space $\mathbb{P}(V^*)$ is a projective plane, since $\dim V^* = 3$ by Corollary 5.1.4, so part (a), applied in $\mathbb{P}(V^*)$, says that $L_1^0 \vee L_2^0$ is a line, of dimension 1.

By Theorem 22.8.6 (d), $(L_1 \cap L_2)^0 = L_1^0 \vee L_2^0$. Hence $\dim(L_1 \cap L_2)^0 = 1$, and Theorem 22.8.6 (b) read backwards gives $\dim(L_1 \cap L_2) = 2 - 1 - 1 = 0$. A projective subspace of dimension 0 is a single point. This proves that L_1 and L_2 meet in exactly one point.

■

The same bookkeeping shows that the dimension formula of Theorem 22.6.6 is **self-dual**: replacing each of $\dim(S \vee T)$, $\dim(S \cap T)$, $\dim S$, $\dim T$ by $n - 1$ minus itself and exchanging $\vee$ with $\cap$ turns

$$\dim(S \vee T) + \dim(S \cap T) = \dim S + \dim T$$

into

$$\begin{aligned}&(n-1-\dim(S\cap T))+(n-1-\dim(S\vee T))\\&=(n-1-\dim S)+(n-1-\dim T),\end{aligned}$$

which is the same equation after canceling $2(n-1)$. A self-dual theorem is one whose dual gives nothing new, and that is a useful thing to be able to detect.

Warning. **The dual of "point" is not "line"; it is "hyperplane", and which dimension that is depends on n.** In the projective plane, $n=2$, and $2-1-0=1$: points dualize to lines, which is why Corollary 22.8.7 looks like a swap of the two words. In $\mathbb{P}^3$, $n=3$, so points dualize to **planes** $(3-1-0=2)$ and lines dualize to lines $(3-1-1=1)$. The correct dual of "two distinct points lie on a unique line" in $\mathbb{P}^3$ is "two distinct planes meet in a unique line", and that is true. The careless dual, "two distinct lines meet in a point", is **false** in $\mathbb{P}^3$: the lines $\mathbb{P}(\operatorname{span}(\mathbf{e}_0,\mathbf{e}_1))$ and $\mathbb{P}(\operatorname{span}(\mathbf{e}_2,\mathbf{e}_3))$ have $\operatorname{span}(\mathbf{e}_0,\mathbf{e}_1)\cap\operatorname{span}(\mathbf{e}_2,\mathbf{e}_3)=\{\mathbf{0}\}$, so they do not meet at all. Always compute $n-1-d$; never translate by the words alone.

Pole and polar — the correspondence a conic sets up between the points and the lines of a plane — is the one piece of duality that needs a quadratic form rather than only the annihilator. It is treated in Section 10 of this chapter, where conics are, and nothing in this section depends on it.

22.8.6 Exercises

22.8.6.1 A. Check your understanding

Exercise. A1

(a) Define the dual projective space of $\mathbb{P}(V)$, and say what its dimension is.

(b) Let S be a plane in $\mathbb{P}^4(F)$. Compute $\dim S^0$.

(c) True or false, with a reason: the duality correspondence of Theorem 22.8.3 is a bijection from the points of $\mathbb{P}(V)$ to the points of $\mathbb{P}(V)$.

(d) In $\mathbb{P}^2$, what is the dual of the statement "three distinct points are collinear"?

(e) What are $\emptyset^0$ and $\mathbb{P}(V)^0$?

Solution.

(a) It is $\mathbb{P}(V^*)$, the projective space of the dual space (Definition 22.8.1). If $\dim\mathbb{P}(V)=n$, then $\dim\mathbb{P}(V^*)=n$ as well, by Corollary 5.1.4.

(b) A plane has dimension 2 and $n=4$, so $\dim S^0=4-1-2=1$: a line in the dual space.

(c) False, twice over. It is a bijection between **all** projective subspaces of $\mathbb{P}(V)$ and all projective subspaces of $\mathbb{P}(V^*)$, and it sends points of $\mathbb{P}(V)$ to hyperplanes of $\mathbb{P}(V^*)$, not to points; and its target is $\mathbb{P}(V^*)$, not $\mathbb{P}(V)$ (see the warning after Example (A line and its point)).

(d) With $n=2$, points dualize to lines and a line dualizes to a point. "Three distinct points lie on one line" becomes "three distinct lines pass through one point", that is, the three lines are concurrent.

(e) $\emptyset^0 = \mathbb{P}(\{0\}^0) = \mathbb{P}(V^*)$, of dimension $n = n - 1 - (-1)$; and $\mathbb{P}(V)^0 = \mathbb{P}(V^0) = \mathbb{P}(\{0\}) = \emptyset$, of dimension $-1 = n - 1 - n$.

22.8.6.2 B. Practice

Exercise. B1: Lines and dual points

Work in $\mathbb{P}^2(\mathbb{Q})$.

(a) Find the dual point of the line through $[1 : 2 : 1]$ and $[0 : 1 : 3]$.

(b) Find the point where the lines $x_0 + x_1 - x_2 = 0$ and $2x_0 - x_1 = 0$ meet.

Solution.

(a) We need $\mathbf{a} = (a_0, a_1, a_2)$ with $a_0 + 2a_1 + a_2 = 0$ and $a_1 + 3a_2 = 0$. From the second, $a_1 = -3a_2$; substituting, $a_0 = -2a_1 - a_2 = 6a_2 - a_2 = 5a_2$. Taking $a_2 = 1$ gives $\mathbf{a} = (5, -3, 1)$, so the line is $5x_0 - 3x_1 + x_2 = 0$ and the dual point is $[5 : -3 : 1]$. Check: $5 - 6 + 1 = 0$ and $0 - 3 + 3 = 0$. By Theorem 5.2.2 the solution space is one-dimensional, so this is the only answer up to scaling.

(b) Solve $x_0 + x_1 - x_2 = 0$ and $2x_0 - x_1 = 0$. The second gives $x_1 = 2x_0$, and then $x_2 = x_0 + x_1 = 3x_0$. Taking $x_0 = 1$, the meet is $[1 : 2 : 3]$. Check: $1 + 2 - 3 = 0$ and $2 - 2 = 0$. The solution space is one-dimensional, consistent with Corollary 22.8.7 (b).

Exercise. B2: Dimension bookkeeping

Let $n = 5$, so $\mathbb{P}(V) = \mathbb{P}^5(F)$. For each of the following S, give $\dim S^0$, and say which of the two is larger.

(a) S a point.

(b) S a line.

(c) S of dimension 2.

(d) S a hyperplane.

(e) $S = \emptyset$.

Then determine whether there is any S with $\dim S^0 = \dim S$. Justify your answer.

Solution. By Theorem 22.8.3 (a), $\dim S^0 = 5 - 1 - \dim S = 4 - \dim S$.

(a) $\dim S = 0$, $\dim S^0 = 4$; the dual is larger. (b) 1 and 3; the dual is larger. (c) 2 and 2; equal. (d) 4 and 0; the original is larger. (e) -1 and 5; the dual is larger.

For the last part, $\dim S^0 = \dim S$ means $4 - \dim S = \dim S$, that is, $\dim S = 2$. So the self-dual dimension in $\mathbb{P}^5$ is 2, and part (c) is an instance. In general $\dim S^0 = \dim S$ forces $2 \dim S = n - 1$, which has a solution exactly when n is odd.

Exercise. B3: Dualize a statement

For each statement about $\mathbb{P}^3(F)$, write the dual statement, using Theorem 22.8.6 with $n = 3$, and say whether both are true.

(a) A point and a line not through it lie in exactly one plane.

(b) Three planes with no common line meet in at most one point.

Solution. With $n = 3$, the dimension rule is $d \mapsto 2 - d$: points (0) and planes (2) exchange, lines (1) stay lines, $\emptyset$ (-1) and the whole space (3) exchange. Joins become meets.

(a) Dual: *a plane and a line not contained in it meet in exactly one point.* Both are true. For the original: if $P \not\subseteq L$ then $P \cap L = \emptyset$, so by Theorem 22.6.6 $\dim(P \vee L) = 0 + 1 - (-1) = 2$, a plane, and it is unique because any plane containing both contains $P \vee L$ and has the same dimension. Dualizing each clause by Theorem 22.8.6 gives the dual statement, which is therefore also true; directly, if the plane Π does not contain L, then $\Pi \vee L$ has dimension 3, so $\dim(\Pi \cap L) = 2 + 1 - 3 = 0$.

(b) Dualize clause by clause, as the warning after Corollary 22.8.7 insists. The hypothesis "the three planes contain no common line" reads: there is no L with $\dim L = 1$ and $L \subseteq \Pi_1 \cap \Pi_2 \cap \Pi_3$. Writing $P_i = \Pi_i^0$, which are points by Theorem 22.8.6 (b), and using (a) and (d) of that theorem, the hypothesis becomes: there is no L' with $\dim L' = 1$ and $P_1 \vee P_2 \vee P_3 \subseteq L'$, that is, the three points are **not** collinear. The conclusion "$\Pi_1 \cap \Pi_2 \cap \Pi_3$ has dimension at most 0" becomes "$P_1 \vee P_2 \vee P_3$ has dimension at least 2". So the dual statement is: *three points not on a common line span a projective subspace of dimension at least* 2, that is, they span a plane. Both are true: three noncollinear points $[v_1], [v_2], [v_3]$ have independent representatives, since a dependence would put all three in a subspace of dimension at most 2 and hence on a common line, so their join has dimension 2 by Lemma 22.8.1 (c). The original then follows by dualizing back.

22.8.6.3 C. Going deeper

Exercise. C1: Counting in a finite projective plane

Let q be a prime power and $F = \mathbb{F}_q$.

(a) Show that a projective space of dimension d over $\mathbb{F}_q$ has $(q^{d+1} - 1)/(q - 1)$ points.

(b) Deduce that $\mathbb{P}^2(\mathbb{F}_q)$ has as many lines as points, and that every point lies on exactly $q + 1$ lines.

Solution.

(a) Such a space is $\mathbb{P}(U)$ with $\dim U = d + 1$. The set $U \setminus \{0\}$ has $q^{d+1} - 1$ elements, and two nonzero vectors give the same point exactly when they differ by a scalar in $\mathbb{F}_q \setminus \{0\}$, a set of size $q - 1$. So each point has exactly $q - 1$ representatives, and the number of points is $(q^{d+1} - 1)/(q - 1)$.

(b) By Proposition 22.8.2 the lines of $\mathbb{P}^2(\mathbb{F}_q)$ are in bijection with the points of $\mathbb{P}((\mathbb{F}_q^3)^*)$, which by Corollary 5.1.4 is again a projective plane over $\mathbb{F}_q$. So by (a) with $d = 2$ both counts are $N := q^2 + q + 1$.

For the last claim, fix a point P. The lines of $\mathbb{P}^2(\mathbb{F}_q)$ containing P are, by the remark after Theorem 22.8.3, exactly the points of P^0, and $\dim P^0 = 2-1-0 = 1$ by Theorem 22.8.3 (a). So by (a) with $d=1$ there are $q+1$ of them. As a check, count the incident pairs (P, L) with $P \in L$ in two ways: each line is a projective space of dimension 1 and so carries $q+1$ points, giving $N(q+1)$ pairs, and the other count gives $N(q+1)$ as well.

Exercise. C2: A form makes a self-duality

Let β be a nondegenerate symmetric bilinear form on V, with $\dim V = n+1$.

(a) Show that $R_\beta \colon V \to V^*$, $v \mapsto \beta(\cdot, v)$, is an isomorphism, and deduce that it induces a projectivity $[R_\beta] \colon \mathbb{P}(V) \to \mathbb{P}(V^*)$.

(b) Hence define, for a projective subspace S of $\mathbb{P}(V)$, a projective subspace $S^\perp$ of $\mathbb{P}(V)$ with $\dim S^\perp = n - 1 - \dim S$, and show that the assignment $S \mapsto S^\perp$ is an inclusion-reversing bijection of the set of projective subspaces of $\mathbb{P}(V)$ with itself.

(c) Give an example over $\mathbb{R}$ with $n = 2$ in which $S \subseteq S^\perp$ for some point S, and explain why this cannot happen if β is positive definite.

Solution.

(a) By Proposition 14.1.2, β is nondegenerate exactly when R_β is an isomorphism $V \to V^*$. A projectivity is induced by any isomorphism (Definition 22.7.1), so $[R_\beta]$ is one.

(b) Define $S^\perp := [R_\beta]^{-1}(S^0)$. This is a projective subspace of $\mathbb{P}(V)$ by Proposition 22.7.2, of the same dimension as S^0, namely $n-1-\dim S$ by Theorem 22.8.3 (a). Since $[R_\beta]^{-1}$ is a bijection preserving inclusion (Proposition 22.7.2) and $S \mapsto S^0$ is an inclusion-reversing bijection onto the projective subspaces of $\mathbb{P}(V^*)$ (Theorem 22.8.3 (b), (d)), the composite is an inclusion-reversing bijection of the projective subspaces of $\mathbb{P}(V)$ with themselves.

(c) Take $V = \mathbb{R}^3$ and $\beta(\mathbf{x}, \mathbf{y}) = x_0y_0 + x_1y_1 - x_2y_2$, which is nondegenerate since its matrix $\operatorname{diag}(1, 1, -1)$ is invertible. Let $\mathbf{v} = (1, 0, 1)$, so $\beta(\mathbf{v}, \mathbf{v}) = 1 + 0 - 1 = 0$, and let $S = \{[\mathbf{v}]\}$. Then $R_\beta(\mathbf{v}) = \beta(\cdot, \mathbf{v})$ kills $\mathbf{v}$, so $[\mathbf{v}] \in S^\perp$, that is, $S \subseteq S^\perp$. If β were positive definite, then $\beta(\mathbf{v}, \mathbf{v}) > 0$ for every $\mathbf{v} \neq \mathbf{0}$, so no nonzero vector is annihilated by its own functional, and $S \subseteq S^\perp$ is impossible for a point S. The nonzero vectors with $\beta(\mathbf{v}, \mathbf{v}) = 0$ are the isotropic ones of Definition 14.7.1, taken for the quadratic form $q(\mathbf{v}) = \beta(\mathbf{v}, \mathbf{v})$; they are exactly the points lying on their own duals.

Exercise. C3: What the correspondence does to a chain

Let $\dim V = n+1$ and let

$$\emptyset = S_{-1} \subsetneq S_0 \subsetneq S_1 \subsetneq \cdots \subsetneq S_n = \mathbb{P}(V)$$

be projective subspaces with $\dim S_d = d$ for each d.

(a) Prove that $S_n^0 \subsetneq S_{n-1}^0 \subsetneq \cdots \subsetneq S_{-1}^0$ is again such a chain, and identify $\dim S_d^0$.

(b) Deduce that $S \mapsto S^0$ carries the chains of this shape in $\mathbb{P}(V)$ bijectively to those in $\mathbb{P}(V^*)$, and that it reverses their order.

Solution.

(a) By Theorem 22.8.6 (a), $S_d \subseteq S_{d+1}$ gives $S^0_{d+1} \subseteq S^0_d$, and the inclusion is strict: if $S^0_{d+1} = S^0_d$ then $S_{d+1} = S_d$ by the equality clause, contradicting strictness. By Theorem 22.8.6 (b), $\dim S^0_d = n - 1 - d$. As d runs from -1 to n, the value $n - 1 - d$ runs from n down to -1, taking each value once. So writing $T_e \coloneqq S^0_{n-1-e}$ gives $\dim T_e = e$ and $T_{-1} \subsetneq T_0 \subsetneq \cdots \subsetneq T_n$, a chain of the same shape in $\mathbb{P}(V^*)$.

(b) The assignment is injective because $S \mapsto S^0$ is (Theorem 22.8.6 (e)), applied term by term. It is surjective: given a chain $T_{-1} \subsetneq T_0 \subsetneq \cdots \subsetneq T_n$ with $\dim T_e = e$ in $\mathbb{P}(V^*)$, set $S_d \coloneqq [\mathrm{ev}_V]^{-1}(T^0_{n-1-d})$, which by Theorem 22.8.4 is the unique projective subspace of $\mathbb{P}(V)$ with $S^0_d = T_{n-1-d}$. By Theorem 22.8.6 (b), $\dim S_d = n - 1 - (n - 1 - d) = d$; and if $d < d'$ then $n - 1 - d > n - 1 - d'$, so $T_{n-1-d} \supsetneq T_{n-1-d'}$, whence $S_d \subsetneq S_{d'}$ by Theorem 22.8.6 (a). So (S_d) is a chain of the required shape whose image is the given one, and the order has been reversed.

22.9 Desargues and Pappus

Two theorems from the drawing board, both of them older than coordinates. Both speak only of points, lines, and incidence: no lengths, no angles, nothing a projective transformation could disturb. Both are proved here by one algebraic move, the one that makes projective arguments short — choosing, out of the whole line of vectors that represents a point, the single vector that makes the hypothesis into an equation.

22.9.1 Perspective from a point, perspective from a line

Draw a triangle on a sheet of glass and let a lamp cast its shadow on the floor. The two triangles that result have their corresponding vertices joined by three lines through the lamp. That is the first of the two relations below. The second is what the same picture looks like when you attend to the sides instead of the vertices: corresponding sides, extended, meet the ground in three points, and those three points lie on one line.

Both relations need the words "the line through two points", and there is exactly one: distinct points $A = [a]$ and $B = [b]$ of $\mathbb{P}(V)$ have independent representatives and lie on the unique line $\mathbb{P}(\mathrm{span}(a, b))$, by Proposition 22.6.8. We write AB for it, and we use that description of it, as $\mathbb{P}(\mathrm{span}(a, b))$, in every proof below.

Definition 22.9.1. Triangles, and the Two Perspectives

Let $\dim V = 3$, so that $\mathbb{P}(V)$ is a projective plane (Definition 22.6.3). A **triangle** in $\mathbb{P}(V)$ is an ordered triple (A, B, C) of **noncollinear** points, called its **vertices**; its **sides** are the three lines AB, BC, CA.

Let (A, B, C) and (A', B', C') be triangles with $A \neq A'$, $B \neq B'$ and $C \neq C'$. They are:

(a) **perspective from a point** if some point O lies on **all three** of the lines AA', BB', CC'; such an O is a **center**;

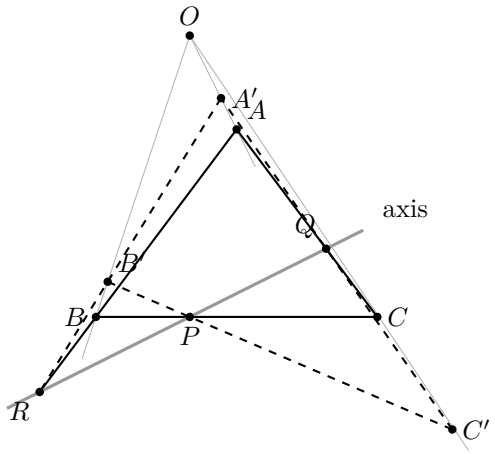
(b) **perspective from a line** if the three pairs of corresponding sides $\{AB, A'B'\}$,

$\{BC, B'C'\}$, $\{CA, C'A'\}$ consist of **distinct** lines and the three intersection points

$$AB \cap A'B', \qquad BC \cap B'C', \qquad CA \cap C'A'$$

are collinear; a line through all three is an **axis**.

In (b) each intersection is a single point: two **distinct** lines in a projective plane meet in exactly one point (Corollary 22.6.7). That is why the clause "distinct in each pair" is part of the definition and not an afterthought; without it the phrase "the point $AB \cap A'B'$" names nothing.



The picture shows both relations at once: the three gray rays make the triangles perspective from O, and the three marked points $R = AB \cap A'B'$, $P = BC \cap B'C'$, $Q = CA \cap C'A'$ lie on the thick line. Desargues's theorem says that this coincidence is no coincidence.

Warning. **Two corresponding sides may be the same line, and then part (b) is not even a statement.** Take a line ℓ, four distinct points A, B, A', B' on it and a point O on it, and any $C \notin \ell$. Fix representatives o of O and c of C and put $C' = [o + c]$, a point of the line OC other than O and C; it does not lie on ℓ, since $OC \neq \ell$ and the two meet only at O (Corollary 22.6.7). The triples (A, B, C) and (A', B', C') are triangles, the six vertices are distinct, and the three lines AA', BB', CC' all pass through O. But $AB = \ell = A'B'$, so "the point $AB \cap A'B'$" does not exist. Exercise (C3: Why the side hypothesis is there) writes this out in coordinates. The hypothesis that rules it out is the one in the theorem below: no vertex of either triangle lies on a side of the other.

22.9.2 Choosing the representatives

A point of $\mathbb{P}(V)$ is a whole line of vectors, and every proof has to pick one vector per point. The pick is free, and freedom used well is the whole art. Here the hypothesis is three collinearities, and each of them becomes a clean equation as soon as the two representatives involved are scaled against each other.

Lemma 22.9.1. Scaling Three Collinear Points

Let $W \subseteq V$ be a 2-dimensional subspace and let O, A, A' be three **distinct** points of the line $\mathbb{P}(W)$. Let o be any nonzero vector representing O. Then there are vectors a representing A

and a' representing A' with

$$a' = a + o.$$

Idea. Two distinct points of a line give a basis of W, so the third is a combination of them. Both coefficients are nonzero, precisely because the third point is neither of the first two; and scaling the two representatives by the right amounts turns that combination into a sum with coefficients 1 and 1.

Proof. Pick any nonzero a_0 representing A. Since $O \neq A$, the pair (o, a_0) is independent, and it lies in W with $\dim W = 2$, so it is a basis of W. As $A' \in \mathbb{P}(W)$, a representative a'_0 of A' can be written

$$a'_0 = \mu\, a_0 + \nu\, o, \qquad \mu, \nu \in F.$$

Here $\nu \neq 0$: otherwise $a'_0 = \mu a_0$ with $\mu \neq 0$, giving $A' = A$. And $\mu \neq 0$: otherwise $a'_0 = \nu o$ with $\nu \neq 0$, giving $A' = O$. Both are excluded because the three points are distinct.

Put $a := (\mu/\nu)\, a_0$ and $a' := (1/\nu)\, a'_0$, which is legitimate since μ/ν and $1/\nu$ are nonzero scalars, so a still represents A and a' still represents A'. Then

$$a' = \tfrac{1}{\nu}(\mu a_0 + \nu o) = \tfrac{\mu}{\nu} a_0 + o = a + o.$$

This proves the lemma.

■

It is Lemma 22.7.4 for $n = 1$, with the one common scalar spent on fixing o first. The lemma is used three times in a row, always with the **same** o. That is the point: o is chosen once and never rescaled again, while each pair (a, a'), (b, b'), (c, c') is scaled to match it.

Quick check. In the lemma, which conclusions survive if we drop the hypothesis $A' \neq O$? And if we drop $A' \neq A$?

Solution. Neither survives. If $A' = O$ we may take $a'_0 = o$, so $\mu = 0$; and the conclusion fails, because every representative a' of $A' = O$ is a multiple of o, so $a = a' - o$ would lie in $\operatorname{span}(o)$, forcing $a = 0$ or $A = [a] = O$, and both are impossible. If $A' = A$ we may take $a'_0 = a_0$, so $\nu = 0$, and $a' = a + o$ would say that two representatives of the same point A differ by o; subtracting, o would be a multiple of a, making $O = A$, again excluded. Each of the two scalars μ, ν is guarded by exactly one of the two hypotheses.

22.9.3 Desargues's theorem

The forward half of Desargues's theorem needs less than the full configuration, and stating it with its minimal hypotheses is what makes it reusable: the converse will apply it to a **different** pair of triangles built from the same picture.

Lemma 22.9.2. Perspective from a Point Gives an Axis

Let $\dim V = 3$. Let (A, B, C) and (A', B', C') be triangles in $\mathbb{P}(V)$ with $A \neq A'$, $B \neq B'$, $C \neq C'$, and suppose:

(i) some point O, **distinct from all six vertices**, lies on each of AA', BB', CC';
(ii) $AB \neq A'B'$, $BC \neq B'C'$ and $CA \neq C'A'$.

Then the three points $AB \cap A'B'$, $BC \cap B'C'$, $CA \cap C'A'$ are collinear.

Idea. Fix one representative o of O. By Lemma 22.9.1 the three collinearities can be written as $a' = a + o$, $b' = b + o$, $c' = c + o$ for suitable representatives. Subtract two of them and o cancels: $a - b = a' - b'$. One vector, lying in $\operatorname{span}(a, b)$ and in $\operatorname{span}(a', b')$ at once, so its point lies on both AB and $A'B'$. The three such vectors add to 0, so they span at most a plane; the triangles being genuine keeps them from collapsing to a line, so they span exactly a plane, and a plane of V is a line of $\mathbb{P}(V)$.

Proof. Fix a nonzero o representing O. The points O, A, A' are distinct, by (i) and by $A \neq A'$, and they are collinear, since $O \in AA'$ and $AA' = \mathbb{P}(W)$ with $\dim W = 2$. So Lemma 22.9.1 gives representatives a of A and a' of A' with $a' = a + o$. The same lemma, with the **same** o, gives $b' = b + o$ and $c' = c + o$.

Subtracting, $a - b = a' - b'$. Now $a - b \neq 0$, since $a = b$ would give $A = B$, contradicting that (A, B, C) is a triangle. Since $a - b \in \operatorname{span}(a, b)$ and $AB = \mathbb{P}(\operatorname{span}(a, b))$, the point $[a - b]$ lies on AB; since $a - b = a' - b' \in \operatorname{span}(a', b')$, it lies on $A'B'$ as well. By (ii) these two lines are distinct, so by Corollary 22.6.7 they meet in exactly one point, and therefore

$$AB \cap A'B' = [a - b].$$

The same argument, with the letters advanced cyclically, gives $BC \cap B'C' = [b - c]$ and $CA \cap C'A' = [c - a]$.

Let $U = \operatorname{span}(a - b,\ b - c,\ c - a)$. The three vectors sum to 0, so $U = \operatorname{span}(a - b,\ b - c)$ and $\dim U \leq 2$. Moreover $\dim U = 2$: if $a - b = t(b - c)$ for some $t \in F$, then $a = (1 + t)b - tc \in \operatorname{span}(b, c)$, so A would lie on BC, contradicting that (A, B, C) is a triangle; and $b - c \neq 0$ for the same reason as before. So $\mathbb{P}(U)$ is a line, and it contains all three points. This proves the lemma.

■

Now the configuration of Definition 22.9.1, with the one extra hypothesis that kept the warning above at bay.

Theorem 22.9.3. Desargues's Theorem

Let $\dim V = 3$ and let (A, B, C) and (A', B', C') be triangles in $\mathbb{P}(V)$ such that

(i) the six vertices are distinct, and
(ii) no vertex of either triangle lies on a side of the other.

Then the two triangles are perspective from a point if and only if they are perspective from a line.

Idea. ① The hypotheses (i) and (ii) are exactly what is needed to feed Lemma 22.9.2: (ii) forces corresponding sides apart, and — less obviously — it also forces a center to miss all six vertices, because a center sitting on a vertex would drag a vertex of one triangle onto a side of the other. ② For the converse there is no second theorem to prove: the same lemma is applied to the auxiliary triangles (A, A', Q) and (B, B', P), which are perspective from R because the axis passes through R. Their corresponding sides meet in $AA' \cap BB'$, in C', and in C; collinearity of those three puts $AA' \cap BB'$ on the line CC', which is concurrency.

Proof. Throughout, (ii) is used in the form: none of A, B, C lies on $A'B'$, $B'C'$ or $C'A'$, and none of A', B', C' lies on AB, BC or CA.

First, the three pairs of corresponding sides are distinct. If $AB = A'B'$ then $A' \in AB$, contrary to (ii); the other two pairs are the same argument with the letters advanced.

($\Rightarrow$) Suppose O lies on AA', BB', CC'. We check that O is none of the six vertices. If $O = A$, then $A \in BB'$; since $A \neq B$, the line through A and B is AB, and it contains B', so $B' \in AB$, contrary to (ii). If $O = B$, then $B \in AA'$ gives $A' \in AB$ in the same way, and if $O = C$ then $C \in AA'$ gives $A' \in CA$; both contradict (ii). If $O = A'$, then $A' \in BB'$; since $A' \neq B'$, the line through them is $A'B'$, and it contains B, so $B \in A'B'$, contrary to (ii). If $O = B'$ then $B' \in AA'$ gives $A \in A'B'$, and if $O = C'$ then $C' \in AA'$ gives $A \in C'A'$; both contradict (ii). So Lemma 22.9.2 applies, and its conclusion is that the two triangles are perspective from a line.

($\Leftarrow$) Suppose the three points

$$R = AB \cap A'B', \qquad P = BC \cap B'C', \qquad Q = CA \cap C'A'$$

are collinear. We apply Lemma 22.9.2 to the triangles (A, A', Q) and (B, B', P), with center R, and verify its hypotheses one at a time.

The points Q and P are not vertices. If $Q = A$ then $A \in C'A'$, and if $Q = A'$ then $A' \in CA$; both contradict (ii). Likewise $P \neq B$ and $P \neq B'$.

Both triples are triangles. Suppose A, A', Q were collinear. Since $Q \in CA$ and $Q \neq A$, the line QA is CA; so $AA' = QA = CA$, whence $A' \in CA$, contrary to (ii). The triple (B, B', P) is handled the same way.

Corresponding vertices are distinct. $A \neq B$ and $A' \neq B'$ by (i). If $Q = P$, that point lies on CA and on CB; these two lines are distinct (else A, B, C would be collinear) and both contain C, so by Corollary 22.6.7 they meet only at C, giving $Q = C$ and hence $C \in C'A'$, contrary to (ii).

The center R is none of the six vertices A, A', Q, B, B', P. If $R = A$ then $A \in A'B'$; if $R = B$ then $B \in A'B'$; if $R = A'$ then $A' \in AB$; if $R = B'$ then $B' \in AB$. All four contradict (ii). If $R = Q$, then this point lies on AB and on CA, two distinct lines through A, so $R = A$, already excluded; if $R = P$, the same argument with AB and BC gives $R = B$, also excluded.

R lies on the three joins. It lies on AB and on $A'B'$ by definition. Since $Q \neq P$, the line QP

is defined, and it contains R because P, Q, R are collinear.

Corresponding sides of the two auxiliary triangles are distinct. If $AA' = BB'$, this one line contains A and B, hence equals AB, and then $A' \in AB$, contrary to (ii). Next, $Q \in C'A'$ and $Q \neq A'$ give $A'Q = C'A'$, and $P \in B'C'$ with $P \neq B'$ give $B'P = B'C'$; these are distinct, since $C'A' = B'C'$ would make A', B', C' collinear. Finally $QA = CA$ and $PB = CB$, distinct for the same reason.

By Lemma 22.9.2 the three points

$$X = AA' \cap BB',$$
$$A'Q \cap B'P = C'A' \cap B'C' = C',$$
$$QA \cap PB = CA \cap CB = C$$

are collinear, the last two evaluations being Corollary 22.6.7 applied to two distinct lines with an evident common point. Since $C \neq C'$ by (i), the line through C and C' is CC', and collinearity puts X on it. So X lies on AA', on BB', and on CC': the triangles are perspective from the point X. This proves the theorem.

■

Two routes to Desargues's theorem are standard. One embeds the plane in $\mathbb{P}^3$, moves one triangle off the plane, and uses that two distinct planes of $\mathbb{P}^3$ meet in a line; Exercise (C1: Desargues in space) carries that route out. The proof above is the other, purely two-dimensional route, and the step that decides whether it works is the scaling of Lemma 22.9.1 — one representative o fixed first, every other representative matched to it afterwards.

Remark. The two halves of the theorem are dual statements: interchange "point" with "line", "lies on" with "passes through", and "perspective from a point" becomes "perspective from a line". Read through the correspondence of Theorem 22.8.3, the converse should therefore be the dual of the direct half rather than an independent fact; **we do not carry that dualization out here**, and nothing depends on it. The proof above does not use it; it derives the converse from the direct half by applying it to a second pair of triangles, so that nothing rests on how far the duality principle of Theorem 22.8.6 can be pushed as a general statement.

Example. Desargues with exact coordinates

In $\mathbb{P}^2(\mathbb{Q})$ take $O = [1:1:1]$ and the triangle $A = [1:0:0]$, $B = [0:1:0]$, $C = [0:0:1]$, and on the three lines through O take

$$A' = [3:1:1], \qquad B' = [1:4:1], \qquad C' = [1:1:2].$$

Verify the hypotheses, scale the representatives as in the proof, and find the axis.

Solution. Each primed point does lie on the corresponding line: $(3,1,1) = 2(1,0,0)+(1,1,1)$, $(1,4,1) = 3(0,1,0) + (1,1,1)$, and $(1,1,2) = (0,0,1) + (1,1,1)$. Both triples are triangles,

since $(\mathbf{e}_0, \mathbf{e}_1, \mathbf{e}_2)$ is a basis and

$$\det \begin{pmatrix} 3 & 1 & 1 \\ 1 & 4 & 1 \\ 1 & 1 & 2 \end{pmatrix} = 17 \neq 0.$$

The six vertices are distinct. No vertex of either lies on a side of the other: the sides of ABC are the coordinate lines $\{x_2 = 0\}$, $\{x_0 = 0\}$, $\{x_1 = 0\}$, and each primed point has all three coordinates nonzero; in the other direction the primed sides are

$$\begin{aligned} A'B' &= \{3x_0 + 2x_1 - 11x_2 = 0\}, \\ B'C' &= \{7x_0 - x_1 - 3x_2 = 0\}, \\ C'A' &= \{-x_0 + 5x_1 - 2x_2 = 0\}, \end{aligned}$$

and no coefficient in any of them is zero, so none of $\mathbf{e}_0, \mathbf{e}_1, \mathbf{e}_2$ satisfies any of the three equations.

Now scale. With $\mathbf{o} = (1, 1, 1)$ the relation $\mathbf{a}'_0 = 2\mathbf{a}_0 + \mathbf{o}$ has $\mu = 2$, $\nu = 1$, so the lemma takes $\mathbf{a} = 2(1, 0, 0) = (2, 0, 0)$ and $\mathbf{a}' = (3, 1, 1)$; similarly $\mathbf{b} = (0, 3, 0)$, $\mathbf{b}' = (1, 4, 1)$ and $\mathbf{c} = (0, 0, 1)$, $\mathbf{c}' = (1, 1, 2)$. Each check is one subtraction:

$$\mathbf{a}' - \mathbf{a} = \mathbf{b}' - \mathbf{b} = \mathbf{c}' - \mathbf{c} = (1, 1, 1) = \mathbf{o}.$$

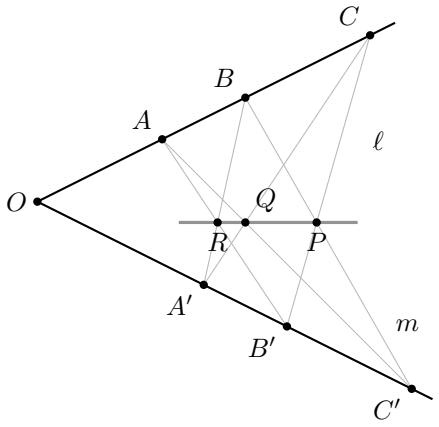
The three intersection points are therefore

$$\begin{aligned} R &= [\mathbf{a} - \mathbf{b}] = [2 : -3 : 0], \\ P &= [\mathbf{b} - \mathbf{c}] = [0 : 3 : -1], \\ Q &= [\mathbf{c} - \mathbf{a}] = [-2 : 0 : 1], \end{aligned}$$

and indeed $(2, -3, 0) + (0, 3, -1) + (-2, 0, 1) = \mathbf{0}$. A linear form vanishing on the first two is $3x_0 + 2x_1 + 6x_2$, and it vanishes on the third as well, so the axis is the line $\{3x_0 + 2x_1 + 6x_2 = 0\}$.

22.9.4 Pappus's theorem

Desargues's theorem uses two triangles. Pappus's uses two lines, with three points on each, and concludes that three points built by cross-joining them lie on a line. Nothing is assumed about the two triples beyond being distinct and avoiding the one point the two lines share.



The six thin lines are the cross-joins; they meet in pairs at R, Q and P, and the theorem says

those three points always lie on one line, the thick one.

Theorem 22.9.4. Pappus's Theorem

Let $\dim V = 3$, let ℓ and m be **distinct** lines of $\mathbb{P}(V)$ and let O be their common point. Let A, B, C be three **distinct** points of ℓ, none equal to O, and let A', B', C' be three **distinct** points of m, none equal to O. Then the three points

$$\begin{aligned} R &= AB' \cap A'B, \\ Q &= AC' \cap A'C, \\ P &= BC' \cap B'C \end{aligned}$$

exist and are collinear.

Idea. Coordinates, in the one basis the data offers: the common point O and one direction along each line. Every point of the configuration then costs a single parameter, each cross-join is a 2×2 system, and the three answers come out as explicit vectors. What finishes the proof is not a determinant but an identity visible on the page: the vector for Q is the sum of the vectors for P and R.

Proof. Write $\ell = \mathbb{P}(U)$ and $m = \mathbb{P}(W)$ with $\dim U = \dim W = 2$. Since $\ell \neq m$ we have $U \neq W$ (Lemma 22.6.4), and two distinct subspaces of the same dimension satisfy $U \not\subseteq W$, so $U + W$ strictly contains W and has dimension $3 = \dim V$. By the dimension formula Theorem 1.7.3, $\dim(U \cap W) = 2 + 2 - 3 = 1$. Thus $U \cap W = \operatorname{span}(o)$ for a nonzero o, and $O = [o]$ is the unique common point of ℓ and m.

Choose $u \in U \setminus \operatorname{span}(o)$ and $w \in W \setminus \operatorname{span}(o)$. Then (o, u) is a basis of U and (o, w) is a basis of W, and (u, w, o) is a basis of V: it has three members and spans $U + W = V$. Every point of ℓ other than O is represented by a vector $su + to$ with $s \neq 0$, hence, after dividing by s, by exactly one vector of the form $u + to$; and likewise on m. So there are scalars with

$$\begin{aligned} A &= [u + ao], & B &= [u + bo], & C &= [u + co], \\ A' &= [w + a'o], & B' &= [w + b'o], & C' &= [w + c'o], \end{aligned}$$

where a, b, c are distinct and a', b', c' are distinct, because the representatives above are uniquely determined by the points.

The three intersections exist. The lines AB' and $A'B$ are defined, since $A \neq B'$ and $A' \neq B$: a point of ℓ other than O is not a point of m, as $\ell \cap m = \{O\}$. They are distinct: if $AB' = A'B$, this line would contain A and B, two distinct points of ℓ, hence be ℓ; then $A' \in \ell \cap m = \{O\}$, contradicting $A' \neq O$. So $R = AB' \cap A'B$ is a single point (Corollary 22.6.7), and the same argument gives Q and P.

Computing R. A vector in $\operatorname{span}(u + ao,\ w + b'o)$ has the form

$$\lambda u + \mu w + (\lambda a + \mu b')o,$$

and a vector in $\operatorname{span}(w + a'o,\ u + bo)$ has the form $\lambda' u + \mu' w + (\lambda' b + \mu' a')o$. A nonzero vector lying in both must have $\lambda = \lambda'$ and $\mu = \mu'$, the triple (u, w, o) being a basis, and then

$$\lambda a + \mu b' = \lambda b + \mu a' \quad \Longleftrightarrow \quad \lambda(a - b) = \mu(a' - b').$$

Taking $\lambda = a' - b'$ and $\mu = a - b$ solves it. The resulting vector is the same read in either span: in the first the coefficient of o is $(a' - b')a + (a - b)b' = aa' - bb'$, and in the second it is $(a' - b')b + (a - b)a' = aa' - bb'$ as well. So

$$r = (a' - b')\,u + (a - b)\,w + (aa' - bb')\,o$$

lies on both lines, and it is nonzero since $a \neq b$. The two lines being distinct, Corollary 22.6.7 makes $[r]$ their unique common point: $R = [r]$. Advancing the letters, $Q = [q]$ and $P = [p]$ with

$$q = (a' - c')\,u + (a - c)\,w + (aa' - cc')\,o,$$
$$p = (b' - c')\,u + (b - c)\,w + (bb' - cc')\,o.$$

The finish. Comparing the three displayed vectors coordinate by coordinate in the basis (u, w, o),

$$p + r - q = 0.$$

So $\operatorname{span}(p, q, r) = \operatorname{span}(p, r)$ has dimension at most 2, and it has dimension at least 1 since $r \neq 0$. Any subspace of dimension 1 or 2 is contained in some 2-dimensional subspace Z of V, and then $\mathbb{P}(Z)$ is a line containing P, Q and R. This proves the theorem.

■

The hypotheses did exactly three jobs, and it is worth saying which. That $\ell \neq m$ gave the basis. That none of the six points is O let every one of them be written as $u + to$ or $w + to$ with the leading coefficient 1; a point equal to O has no such representative, and the parametrization would break down. That a, b, c are distinct and a', b', c' are distinct made the six cross-joins exist and the three vectors nonzero. Nothing else was needed — in particular the six points may interleave along their two lines in any order, and no further nondegeneracy hypothesis is required.

Remark. Over $\mathbb{F}_2$ the theorem is vacuous: a projective line over $\mathbb{F}_2$ has $2 + 1 = 3$ points (Proposition 22.6.9), so after removing O only two remain and three distinct points A, B, C cannot be chosen. Over $\mathbb{F}_3$ the hypotheses can just be met, and Exercise (C2: Pappus over the two smallest fields) checks the conclusion there by hand.

Example. Pappus with exact coordinates

In $\mathbb{P}^2(\mathbb{Q})$ let $\ell = \{x_1 = 0\}$ and $m = \{x_0 = 0\}$, and take

$$A = [1:0:0], \quad B = [1:0:1], \quad C = [1:0:2],$$
$$A' = [0:1:0], \quad B' = [0:1:1], \quad C' = [0:1:3].$$

Find R, Q, P and the line through them.

Solution. Here $\mathbf{o} = \mathbf{e}_2$, $\mathbf{u} = \mathbf{e}_0$, $\mathbf{w} = \mathbf{e}_1$, and the parameters are $a = 0$, $b = 1$, $c = 2$, $a' = 0$,

$b' = 1$, $c' = 3$. The formulas of the proof give

$$\mathbf{r} = (0-1)\mathbf{e}_0 + (0-1)\mathbf{e}_1 + (0-1)\mathbf{e}_2 = -(1,1,1),$$
$$\mathbf{q} = (0-3)\mathbf{e}_0 + (0-2)\mathbf{e}_1 + (0-6)\mathbf{e}_2 = -(3,2,6),$$
$$\mathbf{p} = (1-3)\mathbf{e}_0 + (1-2)\mathbf{e}_1 + (1-6)\mathbf{e}_2 = -(2,1,5),$$

so $R = [1:1:1]$, $Q = [3:2:6]$, $P = [2:1:5]$, and $\mathbf{p} + \mathbf{r} - \mathbf{q} = \mathbf{0}$ as it must be. A linear form vanishing at R and P is $4x_0 - 3x_1 - x_2$, and $4 \cdot 3 - 3 \cdot 2 - 6 = 0$, so all three lie on the line $\{4x_0 - 3x_1 - x_2 = 0\}$.

As a check on the formulas, compute R directly: AB' is spanned by $(1,0,0)$ and $(0,1,1)$, so it is $\{x_1 = x_2\}$; $A'B$ is spanned by $(0,1,0)$ and $(1,0,1)$, so it is $\{x_0 = x_2\}$. Their common point is $[1:1:1]$.

22.9.5 What the two theorems measure

Both theorems have been proved here inside $\mathbb{P}(V)$ for a vector space V over a field F, and the proofs used the field only through its arithmetic. It is worth recording what happens if either assumption is loosened, because the answers are among the most quoted facts in the subject — and because this book has built neither the tools to prove them nor the language to state them carefully.

Remark. An axiomatic **projective plane** is a set of points and lines satisfying incidence axioms, with no vector space given in advance. Hilbert's coordinatization theory attaches an algebraic system to such a plane and reads geometric axioms off it. Two results of that theory: Desargues's theorem holds in a plane exactly when the plane comes from a **division ring** — a field except that multiplication need not commute — and Pappus's theorem holds exactly when that division ring is **commutative**, hence a field. Consequently Pappus implies Desargues, a purely geometric implication first proved by Hessenberg; and there are non-Desarguesian planes, the Moulton plane being the standard example. None of this is proved here, and nothing in this book depends on it: we have not constructed division rings, and "projective plane" has meant $\mathbb{P}(V)$ throughout. The honest summary is that our proof of Pappus's theorem multiplied scalars in whatever order was convenient — and that commuting them is exactly what the general theory says cannot be dispensed with.

22.9.6 Exercises

22.9.6.1 A. Check your understanding

Exercise. A1

(a) Define "perspective from a point" and "perspective from a line" for two triangles in a projective plane.

(b) In Lemma 22.9.1, which hypothesis makes $\mu \neq 0$ and which makes $\nu \neq 0$?

(c) State Theorem 22.9.3, with both of its hypotheses on the configuration.

(d) True or false, with a reason: in Theorem 22.9.4 the six points are required to be pairwise

distinct.

(e) Name the one place in the proof of Theorem 22.9.4 where the hypothesis $\ell \neq m$ is used.

Solution.

(a) Definition 22.9.1: perspective from a point means some point lies on all three lines AA', BB', CC'; perspective from a line means the three pairs of corresponding sides are pairs of distinct lines and their three intersection points are collinear.

(b) The hypothesis $A' \neq O$ makes $\mu \neq 0$, and $A' \neq A$ makes $\nu \neq 0$.

(c) Two triangles whose six vertices are distinct and with no vertex of either lying on a side of the other are perspective from a point if and only if they are perspective from a line.

(d) True, though it is stated in two pieces: A, B, C are required distinct and A', B', C' are required distinct, and no point of $\ell \setminus \{O\}$ can equal a point of $m \setminus \{O\}$, since $\ell \cap m = \{O\}$. So all six are distinct.

(e) In the first paragraph, to get $\dim(U \cap W) = 1$ and hence the basis (u, w, o). If $\ell = m$ then $U = W$ and there is no such basis.

22.9.6.2 B. Practice

Exercise. B1: A Desargues configuration

In $\mathbb{P}^2(\mathbb{Q})$ let $O = [1:1:1]$, $A = [1:0:0]$, $B = [0:1:0]$, $C = [0:0:1]$, and

$$A' = [0:1:1], \qquad B' = [1:0:1], \qquad C' = [1:1:0].$$

Check that each primed point lies on the corresponding line through O, scale the representatives as in Lemma 22.9.2, and compute the axis.

Solution. $(0,1,1) = -(1,0,0) + (1,1,1)$, and similarly $(1,0,1) = -(0,1,0) + (1,1,1)$ and $(1,1,0) = -(0,0,1) + (1,1,1)$. In each case $\mu = -1$ and $\nu = 1$, so Lemma 22.9.1 takes $\mathbf{a} = -\mathbf{e}_0$, $\mathbf{a}' = (0,1,1)$, and likewise $\mathbf{b} = -\mathbf{e}_1$, $\mathbf{b}' = (1,0,1)$, $\mathbf{c} = -\mathbf{e}_2$, $\mathbf{c}' = (1,1,0)$. Each difference $\mathbf{a}' - \mathbf{a}$ is $(1,1,1)$, as required.

Hence

$$AB \cap A'B' = [\mathbf{a} - \mathbf{b}] = [-1:1:0],$$
$$BC \cap B'C' = [\mathbf{b} - \mathbf{c}] = [0:-1:1],$$
$$CA \cap C'A' = [\mathbf{c} - \mathbf{a}] = [1:0:-1],$$

and the three vectors sum to $\mathbf{0}$. A linear form vanishing at the first two is $x_0 + x_1 + x_2$, which also vanishes at the third, so the axis is $\{x_0 + x_1 + x_2 = 0\}$. (Both triples are triangles, since $\det(\mathbf{a}' \mid \mathbf{b}' \mid \mathbf{c}') = 2 \neq 0$.)

Exercise. B2: A Pappus configuration

In $\mathbb{P}^2(\mathbb{Q})$ let $\ell = \{x_1 = 0\}$ and $m = \{x_0 = 0\}$, and take

$$A = [1:0:1], \quad B = [1:0:2], \quad C = [1:0:4],$$
$$A' = [0:1:1], \quad B' = [0:1:3], \quad C' = [0:1:4].$$

Compute R, Q, P and the line through them, and verify $\mathbf{p} + \mathbf{r} - \mathbf{q} = \mathbf{0}$.

Solution. With $\mathbf{u} = \mathbf{e}_0$, $\mathbf{w} = \mathbf{e}_1$, $\mathbf{o} = \mathbf{e}_2$ the parameters are $a = 1$, $b = 2$, $c = 4$ and $a' = 1$, $b' = 3$, $c' = 4$. By the formulas in the proof of Theorem 22.9.4,

$$\mathbf{r} = (1-3)\mathbf{e}_0 + (1-2)\mathbf{e}_1 + (1-6)\mathbf{e}_2 = (-2,-1,-5),$$
$$\mathbf{q} = (1-4)\mathbf{e}_0 + (1-4)\mathbf{e}_1 + (1-16)\mathbf{e}_2 = (-3,-3,-15),$$
$$\mathbf{p} = (3-4)\mathbf{e}_0 + (2-4)\mathbf{e}_1 + (6-16)\mathbf{e}_2 = (-1,-2,-10).$$

Then $\mathbf{p} + \mathbf{r} = (-3,-3,-15) = \mathbf{q}$, so $\mathbf{p} + \mathbf{r} - \mathbf{q} = \mathbf{0}$. The points are $R = [2:1:5]$, $Q = [1:1:5]$, $P = [1:2:10]$. A linear form vanishing at R and Q is $5x_1 - x_2$, and $5 \cdot 2 - 10 = 0$, so the three lie on $\{5x_1 = x_2\}$.

Exercise. B3: Concurrency from collinearity

In $\mathbb{P}^2(\mathbb{Q})$ let $A = [1:0:0]$, $B = [0:1:0]$, $C = [0:0:1]$ and

$$A' = [1:2:2], \qquad B' = [2:1:2], \qquad C' = [2:2:1].$$

Verify the hypotheses of Theorem 22.9.3, then decide whether AA', BB', CC' are concurrent by computing the three points $AB \cap A'B'$, $BC \cap B'C'$, $CA \cap C'A'$ and testing them for collinearity. Do **not** compute the three lines AA', BB', CC'.

Solution. Both triples are triangles: $(\mathbf{e}_0, \mathbf{e}_1, \mathbf{e}_2)$ is a basis, and the determinant of the matrix with rows $(1,2,2)$, $(2,1,2)$, $(2,2,1)$ equals $5 \neq 0$. The six vertices are distinct. No vertex of ABC lies on a side of $A'B'C'$, and conversely: the sides of ABC are the three coordinate lines and every primed point has all coordinates nonzero, while $A'B' = \{2x_0 + 2x_1 - 3x_2 = 0\}$, $B'C' = \{-3x_0 + 2x_1 + 2x_2 = 0\}$, $C'A' = \{2x_0 - 3x_1 + 2x_2 = 0\}$, none of which is satisfied by $\mathbf{e}_0$, $\mathbf{e}_1$ or $\mathbf{e}_2$.

Now $AB = \{x_2 = 0\}$, and a point of $A'B'$ with $x_2 = 0$ satisfies $2x_0 + 2x_1 = 0$, so $AB \cap A'B' = [1:-1:0]$. Cyclically, $BC \cap B'C' = [0:1:-1]$ and $CA \cap C'A' = [-1:0:1]$. The three vectors sum to $\mathbf{0}$, so they are dependent and the points are collinear, on $\{x_0 + x_1 + x_2 = 0\}$. By the ($\Leftarrow$) half of Theorem 22.9.3 the three lines AA', BB', CC' are therefore concurrent.

22.9.6.3 C. Going deeper

Exercise. C1: Desargues in space

Let $\dim V = 4$, so $\mathbb{P}(V) = \mathbb{P}^3$. Let (A, B, C) and (A', B', C') be triangles spanning **distinct** planes Π and Π' of $\mathbb{P}^3$, with the six vertices distinct, and suppose a point $O \notin \Pi \cup \Pi'$ lies on each of AA', BB', CC'.

(a) Prove that A, B, A', B' lie in one plane of $\mathbb{P}^3$, and deduce that the lines AB and $A'B'$ meet.

(b) Prove that $\Pi \cap \Pi'$ is a line.

(c) Deduce that the three points $AB \cap A'B'$, $BC \cap B'C'$, $CA \cap C'A'$ lie on that line, and so are collinear.

Hint: for (a), all four points lie on lines through O; for (b), use the dimension formula for subspaces.

Solution.

(a) Since $O \notin \Pi \cup \Pi'$ and all six vertices lie in $\Pi \cup \Pi'$, the point O is none of them; and $A \neq A'$, $B \neq B'$ because the six are distinct. Let o represent O. The points O, A, A' are distinct and collinear, so Lemma 22.9.1 gives representatives with $a' = a + o$, and the same o gives $b' = b + o$. All four of a, b, a', b' then lie in $Z = \operatorname{span}(o, a, b)$, of dimension at most 3, so the four points lie in the projective subspace $\mathbb{P}(Z)$, of dimension at most 2 — a plane or less. Subtracting, $a - b = a' - b'$, a vector that is nonzero because $A \neq B$ and that lies in $\operatorname{span}(a, b) \cap \operatorname{span}(a', b')$; so $[a - b]$ lies on both AB and $A'B'$, and the two lines meet.

(b) $\Pi = \mathbb{P}(S)$ and $\Pi' = \mathbb{P}(S')$ with $\dim S = \dim S' = 3$ and $S \neq S'$, inside a 4-dimensional V. Then $S + S' = V$, so by Theorem 1.7.3, $\dim(S \cap S') = 3 + 3 - 4 = 2$, and $\mathbb{P}(S \cap S')$ is a line.

(c) First, $AB \neq A'B'$: otherwise A and A' would both lie in Π', hence so would the line AA' and the point O on it, contradicting $O \notin \Pi \cup \Pi'$. Two distinct lines have at most one common point, since two distinct points determine their line (Proposition 22.6.8), and by (a) these two lines do meet; so $AB \cap A'B'$ is a single point. It lies on $AB \subseteq \Pi$ and on $A'B' \subseteq \Pi'$, hence in $\Pi \cap \Pi'$, which is the line of (b). The same holds for the other two points. Three points on one line are collinear.

Exercise. C2: Pappus over the two smallest fields

(a) Prove that the hypotheses of Theorem 22.9.4 cannot be satisfied when $F = \mathbb{F}_2$.

(b) Over $F = \mathbb{F}_3$, take ℓ, m, $\mathbf{o}$, $\mathbf{u}$, $\mathbf{w}$ as in the proof of Theorem 22.9.4 and the parameters $a = a' = 0$, $b = b' = 1$, $c = c' = 2$. Compute $\mathbf{r}, \mathbf{q}, \mathbf{p}$ and exhibit a line containing all three points.

(c) Explain why the choice of parameters in (b) is the only one available over $\mathbb{F}_3$, up to renaming the points.

Solution.

(a) A line of $\mathbb{P}^2(\mathbb{F}_q)$ is $\mathbb{P}(U)$ with $\dim U = 2$, and $U \setminus \{\mathbf{0}\}$ has $q^2 - 1$ vectors falling into $(q^2 - 1)/(q - 1) = q + 1$ points (Proposition 22.6.9). For $q = 2$ a line has 3 points, so $\ell \setminus \{O\}$ has only two, and three distinct points A, B, C of ℓ, none equal to O, do not exist.

(b) With $a = a' = 0$, $b = b' = 1$, $c = c' = 2$ the formulas give

$$\begin{aligned}\mathbf{r} &= (0-1)\mathbf{u} + (0-1)\mathbf{w} + (0-1)\mathbf{o} = (2,2,2),\\ \mathbf{q} &= (0-2)\mathbf{u} + (0-2)\mathbf{w} + (0-4)\mathbf{o} = (1,1,2),\\ \mathbf{p} &= (1-2)\mathbf{u} + (1-2)\mathbf{w} + (1-4)\mathbf{o} = (2,2,0),\end{aligned}$$

all read in $\mathbb{F}_3$ and in the basis $(\mathbf{u}, \mathbf{w}, \mathbf{o})$. So $R = [1:1:1]$, $Q = [1:1:2]$, $P = [1:1:0]$, and every one of them satisfies $x_0 - x_1 = 0$. That linear form defines a line, and the three points lie on it. As a check, $\mathbf{p} + \mathbf{r} - \mathbf{q} = (2+2-1,\ 2+2-1,\ 0+2-2) = (0,0,0)$ in $\mathbb{F}_3$.

(c) A line of $\mathbb{P}^2(\mathbb{F}_3)$ has $3 + 1 = 4$ points, so $\ell \setminus \{O\}$ has exactly three, namely the ones with parameters $0, 1, 2$; the same on m. The three points A, B, C must therefore be those three in some order, and likewise A', B', C'. Renaming amounts to permuting the parameters, which permutes P, Q, R and changes nothing about their collinearity.

Exercise. C3: Why the side hypothesis is there

In $\mathbb{P}^2(\mathbb{Q})$ let

$$\begin{aligned} A &= [1:0:0], \quad B = [0:1:0], \quad C = [0:0:1],\\ A' &= [1:1:0], \quad B' = [1:2:0], \quad C' = [1:3:1],\end{aligned}$$

and let $O = [1:3:0]$.

(a) Prove that (A, B, C) and (A', B', C') are triangles with six distinct vertices, and that O lies on each of AA', BB', CC'.

(b) Prove that the two triangles are **not** perspective from a line, in the strict sense of Definition 22.9.1.

(c) Say which clause of hypothesis (ii) of Theorem 22.9.3 fails here.

Solution.

(a) $(\mathbf{e}_0, \mathbf{e}_1, \mathbf{e}_2)$ is a basis, so (A, B, C) is a triangle, and

$$\det \begin{pmatrix} 1 & 1 & 0 \\ 1 & 2 & 0 \\ 1 & 3 & 1 \end{pmatrix} = 1 \neq 0,$$

so (A', B', C') is one too. The six coordinate vectors are pairwise nonproportional, so the vertices are distinct. Now A, A', O all have last coordinate 0, hence lie on the line $\{x_2 = 0\}$, and so are collinear; the same for B, B', O. Finally $(1,3,1) = (0,0,1) + (1,3,0)$, so C' lies on the line through C and O.

(b) The line AB is $\{x_2 = 0\}$, since $\mathbf{e}_0$ and $\mathbf{e}_1$ span its underlying plane. The line $A'B'$ is spanned by $(1,1,0)$ and $(1,2,0)$, an independent pair inside the same plane, so $A'B' = \{x_2 = 0\} = AB$ as well. The pair $\{AB, A'B'\}$ therefore does not consist of distinct lines, and clause (b) of Definition 22.9.1 fails at its first requirement: there is no point "$AB \cap A'B'$" to speak of, the intersection being a whole line.

(c) The clause that no vertex of one triangle lies on a side of the other. Here $A' = [1:1:0]$ lies on $AB = \{x_2 = 0\}$, and so do B' and O.

22.10 Conics and Lines in Space

Chapter 14 classified symmetric bilinear forms and never drew a picture. This section is the picture. A quadratic form on V has a zero set in $\mathbb{P}(V)$, that zero set is unchanged when the form is scaled, and Sylvester's law becomes a list of shapes. Then a second theme, which looks unrelated and is not: a line in three-dimensional projective space is a plane in F^4, a plane in F^4 is a wedge of two vectors, and the wedges among all of $\Lambda^2 F^4$ are picked out by one quadratic equation. The lines in space are themselves a quadric.

Throughout this section $\operatorname{char} F \neq 2$, except in the Plücker material at the end, which needs no hypothesis on F and says so where it begins. The reason is Chapter 14's: in characteristic 2 a quadratic form is not determined by a symmetric matrix, and every classification theorem we lean on fails.

22.10.1 Quadrics in projective space

A quadratic form q on V is not a function on $\mathbb{P}(V)$: its value at λx is λ^2 times its value at x, so a point has no well-defined q-value. But whether the value is **zero** does not move, and that is all a zero set needs.

A projective quadric is where a quadratic form vanishes; scaling the vector cannot change whether it does.

Definition 22.10.1. Projective Quadric, Conic, Smooth Quadric

Let $\operatorname{char} F \neq 2$, let V be a finite-dimensional vector space over F and let q be a **nonzero** quadratic form on V (Definition 14.3.2), with associated symmetric bilinear form β (Theorem 14.3.5). The **projective quadric defined by** q is

$$Z(q) \coloneqq \{\, [x] \in \mathbb{P}(V) : q(x) = 0 \,\}.$$

The **rank** of the quadric is $\operatorname{rank} \beta$, and the quadric is **smooth** (or **nondegenerate**) when β is nondegenerate, that is, when the rank is $\dim V$. A quadric in a projective **plane** is called a **conic**, the same word Section 5 used for the affine case (Definition 22.5.1).

The set is well defined: if $x' = \lambda x$ with $\lambda \neq 0$, then $q(x') = \beta(\lambda x, \lambda x) = \lambda^2 q(x)$, which is zero exactly when $q(x)$ is. The word **nonzero** in the definition rules out $q = 0$, whose zero set is all of $\mathbb{P}(V)$ and carries no information. And the rank is well defined because β is: over a field of characteristic $\neq 2$ a quadratic form has exactly one symmetric bilinear form behind it (Theorem 14.3.5).

Fix a basis of V. Then $q(\mathbf{x}) = \mathbf{x}^{\top}\mathbf{A}\mathbf{x}$ for a unique symmetric matrix $\mathbf{A}$, the matrix of β (Definition 14.1.2), and $\operatorname{rank} \mathbf{A}$ is the rank of the quadric. Three examples in $\mathbb{P}^2$, written in homogeneous coordinates $[x_0 : x_1 : x_2]$.

- $q = x_0^2 + x_1^2 - x_2^2$, with $\mathbf{A} = \operatorname{diag}(1, 1, -1)$, rank 3, smooth. Over $\mathbb{R}$ its points are the projective version of the unit circle; $[1 : 0 : 1]$ is one of them.
- $q = x_0 x_1$, with $\mathbf{A} = \begin{pmatrix} 0 & 1/2 & 0 \\ 1/2 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$, rank 2. Its zero set is the union of the two lines $\{x_0 = 0\}$ and $\{x_1 = 0\}$.
- $q = x_0^2$, rank 1. Its zero set is the single line $\{x_0 = 0\}$, because a field has no nonzero element whose square is 0. The form remembers something the set forgets, which is the first warning of the section.

Warning. **A quadric is named by its form, not by its set.** Over $\mathbb{R}$ the distinct forms $x_0^2+x_1^2+x_2^2$ and $2x_0^2+5x_1^2+x_2^2$ have the same zero set in $\mathbb{P}^2(\mathbb{R})$, namely the empty set, which records nothing at all about either of them. So "the quadric $Z(q)$" is always shorthand for "the quadric defined by q", and the theorem below classifies **forms** up to projective equivalence, not sets.

A projective transformation $[x] \mapsto [Tx]$ (Definition 22.7.1) carries $Z(q)$ to $Z(q \circ T^{-1})$, and $Z(q) = Z(cq)$ for every $c \neq 0$. So the right relation on forms is: q and q' are **projectively equivalent** when $q' = c\,(q \circ T)$ for some $c \in F^\times$ and some $T \in \mathrm{GL}(V)$. In matrices: $\mathbf{A}' = c\,\mathbf{P}^\top \mathbf{A}\mathbf{P}$ with $\mathbf{P}$ invertible, which is congruence (Definition 14.2.1) followed by a scaling.

Theorem 22.10.1. Classification of Projective Quadrics

Let $\operatorname{char} F \neq 2$ and $\dim V = n+1$, and let q, q' be nonzero quadratic forms on V with matrices $\mathbf{A}, \mathbf{A}'$ in some basis.

(a) If F is algebraically closed, then q and q' are projectively equivalent if and only if $\operatorname{rank}\mathbf{A} = \operatorname{rank}\mathbf{A}'$. There are exactly $n+1$ classes, one for each rank $r = 1, \dots, n+1$, with normal form $x_0^2 + \cdots + x_{r-1}^2$.

(b) If $F = \mathbb{R}$, then q and q' are projectively equivalent if and only if their inertias (n_+, n_-, n_0) and (n'_+, n'_-, n'_0) agree **up to interchanging** n_+ **with** n_-. A normal form is $x_0^2 + \cdots + x_{n_+-1}^2 - x_{n_+}^2 - \cdots - x_{r-1}^2$ with $n_+ \geq n_-$, $r = n_+ + n_-$.

Idea. Both parts reduce the extra freedom — the scalar c — to something already classified. Over an algebraically closed field c has a square root, so scaling is itself a congruence and adds nothing: the answer is Chapter 14's. Over $\mathbb{R}$ a positive c is again a congruence, and the only new move is $c = -1$, which turns q into $-q$ and swaps the two counts.

Proof.

(a) ($\Leftarrow$) It is enough to show that a symmetric $\mathbf{A}$ of rank $r \geq 1$ is congruent to $\mathbf{I}_r \oplus \mathbf{O}$, since a congruence is the case $c = 1$ and congruence is an equivalence relation (Proposition 14.2.2). By Theorem 14.4.2, $\mathbf{A} \simeq \mathbf{D} = \operatorname{diag}(d_1, \dots, d_{n+1})$, and congruence preserves rank (Theorem 14.2.3), so exactly r of the d_i are nonzero; permuting the basis, a congruence by a permutation matrix, we may assume these are $d_1, \dots, d_r$. Since F is algebraically closed, choose s_i with $s_i^2 = d_i$ for $i \leq r$ and put $t_i = 1/s_i$ for $i \leq r$, $t_i = 1$ otherwise; by Proposition 14.4.3 the congruence by $\operatorname{diag}(t_1, \dots, t_{n+1})$ replaces d_i by $t_i^2 d_i$, giving $\mathbf{I}_r \oplus \mathbf{O}$. (For $F = \mathbb{C}$ this is Corollary 14.4.4.)
($\Rightarrow$) Suppose $\mathbf{A}' = c\,\mathbf{P}^\top\mathbf{A}\mathbf{P}$ with $c \neq 0$. Choose $s \in F$ with $s^2 = c$; then $\mathbf{A}' = (s\mathbf{P})^\top \mathbf{A}(s\mathbf{P})$ with $s\mathbf{P}$ invertible, so $\mathbf{A}' \simeq \mathbf{A}$ and Theorem 14.2.3 gives $\operatorname{rank}\mathbf{A}' = \operatorname{rank}\mathbf{A}$. The ranks of a nonzero form run over $1, \dots, n+1$, and the normal form is the $\mathbf{I}_r \oplus \mathbf{O}$ produced above, whose quadratic form is $x_0^2 + \cdots + x_{r-1}^2$.

(b) ($\Leftarrow$) If the inertias agree, Corollary 14.5.2 (b) gives $\mathbf{A}' \simeq \mathbf{A}$, which is the case $c = 1$. If they agree after interchanging n_+ and n_-, then $\mathbf{A}'$ has the same inertia as $-\mathbf{A}$, so $\mathbf{A}' \simeq -\mathbf{A} = (-1)\mathbf{A}$, which is the case $c = -1$ composed with a congruence.

($\Rightarrow$) Let $\mathbf{A}' = c\,\mathbf{P}^\top\mathbf{A}\mathbf{P}$. If $c > 0$, put $s = \sqrt{c}$ and argue as in (a): $\mathbf{A}' \simeq \mathbf{A}$, so the inertias

agree by Corollary 14.5.2 (b). If $c < 0$, put $s = \sqrt{-c}$; then $\mathbf{A}' = (s\mathbf{P})^{\top}(-\mathbf{A})(s\mathbf{P}) \simeq -\mathbf{A}$. Diagonalizing $\mathbf{A}$ with n_+ positive and n_- negative entries diagonalizes $-\mathbf{A}$ with the two counts interchanged, so the inertia of $-\mathbf{A}$, hence of $\mathbf{A}'$, is (n_-, n_+, n_0) by Theorem 14.5.1. This proves the theorem.

■

For $n = 2$ the list is short enough to write out, and worth memorizing.

Corollary 22.10.2. The Conics of a Projective Plane

Let $\operatorname{char} F \neq 2$ and let q be a nonzero quadratic form on a 3-dimensional space.

(a) Over an algebraically closed F there are three classes, with zero sets: a line (rank 1), two distinct lines (rank 2), a smooth conic (rank 3).

(b) Over $\mathbb{R}$ there are five classes, listed by (n_+, n_-) with $n_+ \geq n_-$, with zero sets:

$(1,0)$: a line, $(2,0)$: a single point,
$(1,1)$: two distinct lines, $(3,0)$: empty,
$(2,1)$: a smooth conic, nonempty.

Proof. The classes are those of Theorem 22.10.1; only the zero sets need identifying, and each is read off a normal form.

$x_0^2 = 0$ holds exactly when $x_0 = 0$, a field having no nonzero square root of 0; the set is the line $\{x_0 = 0\}$. Next, $x_0^2 + x_1^2$ factors over an algebraically closed field as $(x_0 + ix_1)(x_0 - ix_1)$ with $i^2 = -1$, and the two linear forms are not proportional since $i \neq -i$ (as $\operatorname{char} F \neq 2$), so the set is two distinct lines; over $\mathbb{R}$ instead $x_0^2 + x_1^2 = 0$ forces $x_0 = x_1 = 0$, leaving the single point $[0:0:1]$. The form $x_0^2 - x_1^2 = (x_0 - x_1)(x_0 + x_1)$ gives two distinct lines over every field of characteristic $\neq 2$. Over $\mathbb{R}$, $x_0^2 + x_1^2 + x_2^2 = 0$ forces $\mathbf{x} = \mathbf{0}$, which is not a point of $\mathbb{P}^2$, so the set is empty. Finally $x_0^2 + x_1^2 - x_2^2$ has rank 3, so its quadric is smooth, and it contains $[1:0:1]$. This proves the corollary.

■

Remark. Section 5 classified **affine** quadrics and produced a longer real list, because an affine classification must also record whether a center exists and whether the constant can be cleared. There is no conflict: the affine group is smaller than the projective one — an invertible affine substitution is exactly a projectivity fixing the hyperplane at infinity (Theorem 22.7.11) — so the affine classification is the finer of the two, and the list above is what Section 5 pointed forward to for the complex case.

Warning. **Rank alone does not classify real conics; inertia does.** The forms $x_0^2+x_1^2+x_2^2$ and $x_0^2 + x_1^2 - x_2^2$ both have rank 3, and one zero set is empty while the other has infinitely many points. One rank lower, $x_0^2 + x_1^2$ and $x_0^2 - x_1^2$ both have rank 2, and the zero sets are a point and a pair of lines. Over $\mathbb{C}$ each pair collapses: rank is a complete invariant there, so the two forms of each pair become projectively equivalent.

Example. Classifying a conic by hand

Classify the conic $x_0^2 + 4x_0x_1 + 5x_1^2 + 2x_1x_2 + 2x_2^2 = 0$ in $\mathbb{P}^2(\mathbb{R})$, and say what happens over $\mathbb{C}$.

Solution. Complete the square in x_0 first, since x_0^2 appears:

$$x_0^2 + 4x_0x_1 = (x_0 + 2x_1)^2 - 4x_1^2,$$

so the form equals $(x_0 + 2x_1)^2 + x_1^2 + 2x_1x_2 + 2x_2^2$. Completing the square in x_1 in what is left, $x_1^2 + 2x_1x_2 = (x_1 + x_2)^2 - x_2^2$, and therefore

$$q = (x_0 + 2x_1)^2 + (x_1 + x_2)^2 + x_2^2.$$

The three linear forms $x_0 + 2x_1$, $x_1 + x_2$, x_2 are independent — the matrix with these rows is upper triangular with determinant 1 — so this is a diagonalization with inertia $(3, 0, 0)$. By Corollary 22.10.2 (b) the conic is **empty** over $\mathbb{R}$.

Over $\mathbb{C}$ the same three squares have rank 3, so by (a) the conic is smooth, and it is not empty: take $x_2 = 1$ and $x_1 = -1$, so that the second square vanishes and the equation becomes $(x_0 - 2)^2 = -1$, solved by $x_0 = 2 + i$. The point $[2 + i : -1 : 1]$ lies on it. The real conic was empty not because the form was degenerate but because it was definite, which is a statement no algebraically closed field can make.

Quick check. A conic in $\mathbb{P}^2(\mathbb{R})$ consists of exactly one point. What is the rank of its form, and what is its inertia?

Solution. By the list in Corollary 22.10.2 (b) the only class with a one-point zero set is $(n_+, n_-) = (2, 0)$, of rank 2; up to sign the form is $x_0^2 + x_1^2$ in suitable coordinates. Rank 1 gives a line and rank 3 gives either nothing or a curve, so no other class can produce one point.

22.10.2 Five points determine a conic

Six coefficients, five conditions, one answer. The existence half is a count; the uniqueness half is where the work is, and it is exactly where "in general position" earns its keep.

"In general position" is Definition 22.7.2 read with $m = 5$ and $n = 2$: every subset of at most three of the five points has independent representatives. In a projective plane that says exactly that the five points are pairwise distinct and that no three of them are collinear.

The lemma below is what forbids a conic from swallowing a line, and it is used twice more later.

Lemma 22.10.3. A Conic Containing a Line Is Degenerate

Let $\operatorname{char} F \neq 2$, let $\dim V = 3$ and let $q \neq 0$ be a quadratic form on V vanishing on every vector of a 2-dimensional subspace W. Then $q = \varphi\chi$ for two nonzero linear functionals φ, χ

on V with $\ker\varphi = W$, and the rank of q is at most 2. In particular a **smooth** conic contains no line.

Proof. Let β be the symmetric form of q and choose a basis (w_0, w_1, v_2) of V with $W = \operatorname{span}(w_0, w_1)$. Evaluating q at w_0, at w_1, and at $w_0 + w_1$, all of which lie in W, gives

$$\beta(w_0, w_0) = 0, \quad \beta(w_1, w_1) = 0, \quad 2\beta(w_0, w_1) = 0,$$

the third because $q(w_0 + w_1) = \beta(w_0, w_0) + 2\beta(w_0, w_1) + \beta(w_1, w_1)$. Since 2 is invertible, $\beta(w_0, w_1) = 0$ too. Put $c_0 = \beta(w_0, v_2)$, $c_1 = \beta(w_1, v_2)$ and $c = \beta(v_2, v_2)$. In the coordinates $\mathbf{x} = (x_0, x_1, x_2)$ of this basis,

$$q(\mathbf{x}) = 2c_0x_0x_2 + 2c_1x_1x_2 + cx_2^2 = x_2(2c_0x_0 + 2c_1x_1 + cx_2),$$

a product $\varphi\chi$ with $\varphi(\mathbf{x}) = x_2$, whose kernel is W, and χ the second factor, which is nonzero because $q \neq 0$. The Gram matrix of β in this basis (Definition 14.1.2) has rows $(0, 0, c_0)$, $(0, 0, c_1)$ and (c_0, c_1, c); the first two lie in a 1-dimensional subspace, so the three of them span at most a 2-dimensional one and the rank is at most 2.

If q is smooth its rank is 3, so no such W exists, and a line $\mathbb{P}(W)$ contained in $Z(q)$ would supply one. This proves the lemma.

■

Theorem 22.10.4. Five Points Determine a Conic

Let $\operatorname{char} F \neq 2$ and let $P_1, \dots, P_5$ be five points of $\mathbb{P}^2$ in general position (Definition 22.7.2), that is, pairwise distinct with no three collinear. Then there is a nonzero quadratic form q on F^3 vanishing at all five, and it is unique up to a nonzero scalar. Consequently exactly one conic passes through the five points.

Idea. The symmetric 3×3 matrices form a 6-dimensional space and each point imposes one linear equation on it, so five points leave at least a line of solutions: that is existence, and it is only counting. For uniqueness suppose the solution space had dimension 2. Then a whole pencil of conics passes through the five points, and a pencil is big enough to be forced through a **sixth** point, chosen on the line P_1P_2. A conic through three points of a line contains the line, so by Lemma 22.10.3 it splits into two lines; one of them is P_1P_2, and the other must then carry P_3, P_4, P_5 — three collinear points, which general position forbids.

Proof. Let $S \subseteq M_3(F)$ be the subspace of symmetric matrices. A symmetric 3×3 matrix is determined by its three diagonal entries and its three entries above the diagonal, each of which may be chosen freely, so $\dim S = 6$. By Theorem 14.3.5 the map $\mathbf{A} \mapsto q_{\mathbf{A}}$, $q_{\mathbf{A}}(\mathbf{x}) = \mathbf{x}^{\top}\mathbf{A}\mathbf{x}$, is a bijection from S onto the quadratic forms on F^3, and it is linear.

Existence. Write $P_i = [\mathbf{x}_i]$. For each i the map $\mathbf{A} \mapsto \mathbf{x}_i^{\top}\mathbf{A}\mathbf{x}_i$ is a linear functional on S, because the expression is linear in the entries of $\mathbf{A}$. The five functionals give a homogeneous linear system of 5 equations in the 6 coordinates of $\mathbf{A}$, which by Corollary 3.2.5 (a) has a nontrivial solution $\mathbf{A} \neq \mathbf{O}$. The corresponding $q_{\mathbf{A}}$ is a nonzero quadratic form vanishing

at every $\mathbf{x}_i$, hence at every P_i. (Vanishing does not depend on the representative, by the homogeneity noted after Definition 22.10.1.)

Uniqueness. Let $N \subseteq S$ be the solution space of that system, and suppose $\dim N \geq 2$. Pick independent $\mathbf{A}_1, \mathbf{A}_2 \in N$, and write q_1, q_2 for the corresponding forms, which are independent as forms since $\mathbf{A} \mapsto q_{\mathbf{A}}$ is a bijection that is linear.

The line $L = P_1P_2$ contains at least three points: it is $\mathbb{P}(U)$ with $\dim U = 2$, so it contains $[\mathbf{u}_1]$, $[\mathbf{u}_2]$ and $[\mathbf{u}_1 + \mathbf{u}_2]$ for a basis $(\mathbf{u}_1, \mathbf{u}_2)$ of U, and these are three distinct points. Choose $R = [\mathbf{y}]$ on L with $R \neq P_1$ and $R \neq P_2$. The condition $\lambda\, \mathbf{y}^{\top}\mathbf{A}_1\mathbf{y} + \mu\, \mathbf{y}^{\top}\mathbf{A}_2\mathbf{y} = 0$ is one homogeneous linear equation in (λ, μ), so by Corollary 3.2.5 (a) it has a solution with $(\lambda, \mu) \neq (0, 0)$. Put $\mathbf{A} = \lambda\mathbf{A}_1 + \mu\mathbf{A}_2$, which is nonzero because $\mathbf{A}_1, \mathbf{A}_2$ are independent, and let $q = q_{\mathbf{A}}$, a nonzero quadratic form.

Now q vanishes at the three distinct points P_1, P_2, R of L. Restrict q to U: in the basis $(\mathbf{u}_1, \mathbf{u}_2)$ it reads $g(s,t) = \alpha s^2 + \gamma st + \delta t^2$. A nonzero g vanishes at at most two points of $\mathbb{P}(U)$. Indeed, if $\alpha \neq 0$ then $g(s,0) = \alpha s^2 \neq 0$ whenever $s \neq 0$, so every zero has $t \neq 0$ and is therefore $[m : 1]$ for a scalar m with $\alpha m^2 + \gamma m + \delta = 0$ — a polynomial in m of degree at most 2, nonzero because its leading coefficient is α, so with at most two roots (Lemma 3.7.2); if $\alpha = 0$ then $g = t(\gamma s + \delta t)$, whose zeros are $[1 : 0]$ together with the at most one point where $\gamma s + \delta t = 0$, and $(\gamma, \delta) \neq (0, 0)$ since $g \neq 0$. Three zeros therefore force $g = 0$, that is, q vanishes on all of U.

By Lemma 22.10.3, $q = \varphi\chi$ with $\ker\varphi = U$. For $i = 3, 4, 5$ we have $q(\mathbf{x}_i) = 0$, and $\varphi(\mathbf{x}_i) \neq 0$, since $\varphi(\mathbf{x}_i) = 0$ would put P_i on $L = P_1P_2$, making P_1, P_2, P_i collinear. Hence $\chi(\mathbf{x}_i) = 0$ for $i = 3, 4, 5$, so P_3, P_4, P_5 all lie on the line $\{\chi = 0\}$ — three collinear points among the five, contradicting general position.

So $\dim N = 1$: the form is unique up to a nonzero scalar, and $Z(q)$ is the unique conic through the five points. This proves the theorem.

■

Example. A conic through five points

Find the conic of $\mathbb{P}^2(\mathbb{Q})$ through $[1 : 0 : 0]$, $[0 : 1 : 0]$, $[0 : 0 : 1]$, $[1 : 1 : 1]$ and $[1 : 2 : 4]$, and decide whether it is smooth.

Solution. Write $q = ax_0^2 + bx_1^2 + cx_2^2 + dx_0x_1 + ex_0x_2 + fx_1x_2$. The first three points give $a = b = c = 0$ at once. Then $[1 : 1 : 1]$ gives $d + e + f = 0$ and $[1 : 2 : 4]$ gives $2d + 4e + 8f = 0$, that is $d + 2e + 4f = 0$. Subtracting, $e + 3f = 0$, so $e = -3f$ and $d = -e - f = 2f$. Taking $f = 1$,

$$q = 2x_0x_1 - 3x_0x_2 + x_1x_2,$$

and the solution space is indeed one-dimensional, as Theorem 22.10.4 promises. (No three of the five points are collinear: taking the ten triples in the order 123, 124, 125, 134, 135, 145, 234, 235, 245, 345, the determinants of the corresponding coordinate matrices are

$$1,\ 1,\ 4,\ -1,\ -2,\ 2,\ 1,\ 1,\ -3,\ 1,$$

none of them zero.)

Its matrix is

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & -3/2 \\ 1 & 0 & 1/2 \\ -3/2 & 1/2 & 0 \end{pmatrix}, \qquad \det \mathbf{A} = -\tfrac{3}{2} \neq 0,$$

so the conic is smooth. Over $\mathbb{R}$ its inertia is $(2, 1, 0)$: the three eigenvalues are real (Corollary 12.5.2) and the inertia counts their signs (Theorem 14.5.3); their product is $\det \mathbf{A} < 0$, so an odd number of them is negative, and $\operatorname{tr} \mathbf{A} = 0$ rules out all three, leaving exactly one.

22.10.3 Projecting a conic onto a line

Here is the payoff a reader can take home. A smooth conic looks like a curve, and the next theorem says that as a set it **is** a projective line, provided it has one point to start from. Over $\mathbb{Q}$ that converts one solution of a Diophantine equation into all of them; over $\mathbb{F}_q$ it says a smooth conic with a point has exactly $q + 1$ of them.

Theorem 22.10.5. Projection from a Point of a Conic

Let $\operatorname{char} F \neq 2$, let $C = Z(q)$ be a **smooth** conic in $\mathbb{P}(V)$, $\dim V = 3$, and suppose C contains a point $P_0 = [p]$. Let ℓ be any line of $\mathbb{P}(V)$ with $P_0 \notin \ell$. Then the map

$$\psi \colon \ell \to C, \qquad \psi([x]) = \left[\, -q(x)\, p + 2\beta(p, x)\, x \,\right]$$

is a well-defined bijection. In particular C is in bijection with $\mathbb{P}^1$.

Idea. Each point X of ℓ is joined to P_0, and the line P_0X meets C in P_0 and in exactly one further point — "further" understood so that a tangent line counts P_0 twice. Expanding $q(\lambda p + \mu x)$ and using $q(p) = 0$ makes μ a factor, so the quadratic equation is linear once the known root is removed, and the displayed formula is that root written out. Every line through P_0 other than ℓ itself meets ℓ once, which is where the inverse comes from.

Proof. Write $\ell = \mathbb{P}(U)$ with $\dim U = 2$; since $P_0 \notin \ell$, we have $p \notin U$, so $x \notin \operatorname{span}(p)$ for every $0 \neq x \in U$.

The formula is well defined. Replacing x by cx with $c \neq 0$ multiplies $-q(x)p + 2\beta(p, x)x$ by c^2, so the point is unchanged. The vector is nonzero: if both $q(x) = 0$ and $\beta(p, x) = 0$, then for all λ, μ,

$$q(\lambda p + \mu x) = \lambda^2 q(p) + 2\lambda\mu\, \beta(p, x) + \mu^2 q(x) = 0,$$

so q vanishes on the 2-dimensional subspace $\operatorname{span}(p, x)$, contradicting Lemma 22.10.3 since C is smooth. If instead $\beta(p, x) \neq 0$ the vector has a nonzero component along x modulo $\operatorname{span}(p)$; and if $\beta(p, x) = 0$ with $q(x) \neq 0$, the vector is $-q(x)p \neq 0$.

The value lies on C. Put $z = -q(x)p + 2\beta(p, x)x$. Using $q(p) = 0$ and bilinearity,

$$\begin{aligned} q(z) &= q(x)^2 q(p) - 2 \cdot 2q(x)\beta(p, x)\, \beta(p, x) + 4\beta(p, x)^2 q(x) \\ &= 0 - 4q(x)\beta(p, x)^2 + 4\beta(p, x)^2 q(x) = 0. \end{aligned}$$

A fact used twice. If $[y] \in C$ and $[y] \neq P_0$, then $\beta(p, y) \neq 0$; otherwise q would vanish

on $\operatorname{span}(p, y)$, a 2-dimensional subspace since p, y represent distinct points, contradicting Lemma 22.10.3 again.

Surjectivity. Let $[y] \in C$. Suppose first $[y] \neq P_0$. The line $P_0[y]$ is distinct from ℓ, since $P_0 \notin \ell$, so by Corollary 22.6.7 it meets ℓ in a single point $[x]$, and $x = \lambda p + \mu y$ with $\mu \neq 0$ (else $[x] = P_0 \in \ell$). Then $q(x) = 2\lambda\mu\,\beta(p, y)$ and $\beta(p, x) = \mu\,\beta(p, y)$, so

$$\begin{aligned} -q(x)p + 2\beta(p, x)x &= -2\lambda\mu\beta(p, y)\,p + 2\mu\beta(p, y)(\lambda p + \mu y) \\ &= 2\mu^2\beta(p, y)\,y, \end{aligned}$$

which is a nonzero multiple of y by the fact just proved. So $\psi([x]) = [y]$. Now suppose $[y] = P_0$. The set $T = \{[z] : \beta(p, z) = 0\}$ is a line, β being nondegenerate so that $\beta(p, \cdot)$ is a nonzero functional; it contains P_0, because $\beta(p, p) = q(p) = 0$, hence $T \neq \ell$, and T meets ℓ in one point $[x]$. There $\beta(p, x) = 0$ and $q(x) \neq 0$ by the nonvanishing argument above, so $\psi([x]) = [-q(x)p] = P_0$.

Injectivity. Suppose $\psi([x']) = \psi([x'']) = [y]$, with $[x'], [x''] \in \ell$. For any x the vector $-q(x)p + 2\beta(p, x)x$ lies in $\operatorname{span}(p, x)$, so $[y]$ lies on the line joining P_0 to $[x]$. If $[y] \neq P_0$, then both $[x']$ and $[x'']$ lie on the line joining P_0 to $[y]$, and both lie on ℓ; these two lines are distinct because $P_0 \notin \ell$, so by Corollary 22.6.7 they meet in one point and $[x'] = [x'']$. If $[y] = P_0$, then $2\beta(p, x')x' \in \operatorname{span}(p)$; since $x' \notin \operatorname{span}(p)$ and 2 is invertible, $\beta(p, x') = 0$, and likewise for x''. So $[x']$ and $[x'']$ both lie on $T \cap \ell$, again a single point.

Finally ℓ is a projective line, hence in bijection with $\mathbb{P}^1$ by any choice of basis of U. This proves the theorem.

■

Example. Pythagorean triples from one point

Parametrize the rational points of the conic $x_0^2 + x_1^2 = x_2^2$ in $\mathbb{P}^2(\mathbb{Q})$, starting from $P_0 = [-1 : 0 : 1]$.

Solution. Here $q(\mathbf{x}) = x_0^2 + x_1^2 - x_2^2$, with $\beta(\mathbf{x}, \mathbf{y}) = x_0y_0 + x_1y_1 - x_2y_2$, and $\mathbf{A} = \operatorname{diag}(1, 1, -1)$ is invertible, so the conic is smooth. Also $q(-1, 0, 1) = 1 - 1 = 0$, so $P_0 \in C$. Take $\ell = \{x_0 = 0\}$, which misses P_0, and write its points as $[0 : u : v]$ with $(u, v) \neq (0, 0)$.

With $\mathbf{x} = (0, u, v)$ we get $q(\mathbf{x}) = u^2 - v^2$ and $\beta(\mathbf{p}, \mathbf{x}) = -v$, so

$$\begin{aligned} \psi([0 : u : v]) &= \left[-(u^2 - v^2)(-1, 0, 1) + 2(-v)(0, u, v)\right] \\ &= \left[\,u^2 - v^2 \;:\; -2uv \;:\; -u^2 - v^2\,\right] \\ &= \left[\,v^2 - u^2 \;:\; 2uv \;:\; u^2 + v^2\,\right]. \end{aligned}$$

Taking u, v to be integers, not both zero, the three coordinates are integers, and

$$(v^2 - u^2)^2 + (2uv)^2 = u^4 + 2u^2v^2 + v^4 = (u^2 + v^2)^2,$$

a Pythagorean triple; the identity is exactly the statement that ψ lands in C.

By Theorem 22.10.5 every point of C arises from exactly one point $[0 : u : v]$ of ℓ. So if

(a, b, c) are integers with $a^2 + b^2 = c^2$ and $c \neq 0$, then $[a : b : c]$ is a rational point of C, hence

$$(a, b, c) = \lambda\,(v^2 - u^2,\ 2uv,\ u^2 + v^2)$$

for some rational $\lambda \neq 0$ and integers u, v: every Pythagorean triple is proportional to one of this shape. (If $c = 0$ then $a^2 + b^2 = 0$ forces $a = b = 0$.) Deciding **which** proportionality factors give primitive triples is arithmetic about parities and common factors, not linear algebra, and we do not do it here.

Two checks. At $[u : v] = [1 : 0]$, that is at the point $[0 : 1 : 0]$ of ℓ, the formula returns $[-1 : 0 : 1] = P_0$. That is as it should be: the line joining P_0 to $[0 : 1 : 0]$ consists of the points $[-\lambda : \mu : \lambda]$, and $q(-\lambda, \mu, \lambda) = \lambda^2 + \mu^2 - \lambda^2 = \mu^2$, which vanishes only for $\mu = 0$. So that line meets C at P_0 alone, and P_0 is the only value ψ can take there. At $[u : v] = [1 : 2]$ the formula gives $[3 : 4 : 5]$.

22.10.4 Pole, polar, and the dual conic

A smooth conic carries an invertible symmetric matrix, and an invertible symmetric matrix is a perfect pairing. Every point therefore names a line, and the dictionary this sets up is where tangency comes from.

Definition 22.10.2. Polar Line, Tangent Line

Let $C = Z(q)$ be a smooth conic in $\mathbb{P}(V)$, $\dim V = 3$, with symmetric bilinear form β, and let $P = [x]$. The **polar** of P with respect to C is

$$P^\beta := \{\, [y] \in \mathbb{P}(V) : \beta(x, y) = 0 \,\},$$

and P is the **pole** of that line. For $P \in C$, the polar P^β is called the **tangent line** to C at P.

Because β is nondegenerate, $\beta(x, \cdot)$ is a nonzero functional (Proposition 14.1.2 (c)), so its kernel is 2-dimensional and P^β really is a line; and replacing x by cx scales the functional, leaving the line alone. The next proposition says the word "tangent" is deserved.

Proposition 22.10.6. Reciprocity and Tangency

With the notation of Definition 22.10.2, let $P = [x]$ and $Q = [y]$.

(a) $Q \in P^\beta$ if and only if $P \in Q^\beta$.

(b) $P \in C$ if and only if $P \in P^\beta$.

(c) Let $P \in C$ and let L be a line through P. Then $L \cap C = \{P\}$ if and only if $L = P^\beta$; every other line through P meets C in exactly two points.

Proof.

(a) Both conditions say $\beta(x, y) = 0$, since β is symmetric.

(b) $P \in P^\beta$ says $\beta(x, x) = 0$, which is $q(x) = 0$.

(c) Write $L = \mathbb{P}(\operatorname{span}(x, y))$ with $y \notin \operatorname{span}(x)$. A point of L is $[\lambda x + \mu y]$, and since $q(x) = 0$,

$$q(\lambda x + \mu y) = \mu\big(2\lambda\,\beta(x, y) + \mu\, q(y)\big).$$

Case 1: $L = P^{\beta}$, that is, $\beta(x, y) = 0$. Then $q(\lambda x + \mu y) = \mu^2 q(y)$, and $q(y) \neq 0$: otherwise q would vanish on $\operatorname{span}(x, y)$, contradicting Lemma 22.10.3. So the expression vanishes only when $\mu = 0$, giving the single point P.

Case 2: $L \neq P^{\beta}$, so $\beta(x, y) \neq 0$. Then the factorization above vanishes when $\mu = 0$, giving P, and when $2\lambda\beta(x, y) + \mu q(y) = 0$, which has the solution $(\lambda, \mu) = (-q(y),\, 2\beta(x, y))$ with $\mu \neq 0$, giving a point of C different from P. Every point of $L \cap C$ is one of these two, since the two cases $\mu = 0$ and $2\lambda\beta(x, y) + \mu q(y) = 0$ exhaust the factorization and each determines $[\lambda : \mu]$. This proves the proposition.

■

Corollary 22.10.7. The Dual Conic

Let C be a smooth conic in $\mathbb{P}(V)$ with matrix $\mathbf{A}$ in some basis, and identify each line of $\mathbb{P}(V)$ with the point $[\varphi]$ of $\mathbb{P}(V^*)$ for which the line is $\mathbb{P}(\ker \varphi)$, as Proposition 22.8.2 allows. Then the set of tangent lines of C is the smooth conic of $\mathbb{P}(V^*)$ with matrix $\mathbf{A}^{-1}$.

Proof. The tangent at $[\mathbf{x}] \in C$ is the zero set of the functional $\mathbf{y} \mapsto \beta(\mathbf{x}, \mathbf{y}) = \mathbf{x}^{\top}\mathbf{A}\mathbf{y}$, whose coordinate vector in the dual basis (Theorem 5.1.3) is $\mathbf{A}\mathbf{x}$. As $[\mathbf{x}]$ runs over C, the corresponding point of $\mathbb{P}(V^*)$ is $[\mathbf{A}\mathbf{x}]$ with $\mathbf{x}^{\top}\mathbf{A}\mathbf{x} = 0$. Since $\mathbf{A}$ is invertible, putting $\mathbf{z} = \mathbf{A}\mathbf{x}$ gives $\mathbf{x} = \mathbf{A}^{-1}\mathbf{z}$ and

$$0 = \mathbf{x}^{\top}\mathbf{A}\mathbf{x} = \mathbf{z}^{\top}(\mathbf{A}^{-1})^{\top}\mathbf{A}\mathbf{A}^{-1}\mathbf{z} = \mathbf{z}^{\top}\mathbf{A}^{-1}\mathbf{z},$$

using $(\mathbf{A}^{-1})^{\top} = (\mathbf{A}^{\top})^{-1} = \mathbf{A}^{-1}$. Conversely every nonzero $\mathbf{z}$ with $\mathbf{z}^{\top}\mathbf{A}^{-1}\mathbf{z} = 0$ arises this way, from $\mathbf{x} = \mathbf{A}^{-1}\mathbf{z}$, which satisfies $\mathbf{x}^{\top}\mathbf{A}\mathbf{x} = \mathbf{z}^{\top}\mathbf{A}^{-1}\mathbf{z} = 0$. The matrix $\mathbf{A}^{-1}$ is symmetric and invertible, so the set is a smooth conic. This proves the corollary.

■

22.10.5 Lines in space: Plücker coordinates

The definition, the proposition, the theorem, and the warning that follow work over **any** field; no characteristic hypothesis is needed, and none is used. Only the paragraph closing the subsection returns to quadrics, and it re-imposes $\operatorname{char} F \neq 2$ where it does.

Chapter 15 built the exterior powers and hinted that the $k \times k$ minors of an $n \times k$ matrix turn a k-dimensional subspace into a single point. Here it is. A line of $\mathbb{P}^3 = \mathbb{P}(F^4)$ is $\mathbb{P}(W)$ with $\dim W = 2$; a basis $(\mathbf{u}, \mathbf{w})$ of W has a wedge $\mathbf{u} \wedge \mathbf{w} \in \Lambda^2 F^4$, which is nonzero by Theorem 15.8.7; and $\dim \Lambda^2 F^4 = \binom{4}{2} = 6$ by Theorem 15.8.5. So a line in space becomes a point of $\mathbb{P}^5$. The only questions are whether the point depends on the basis, whether different lines give different points, and which points of $\mathbb{P}^5$ occur.

Index convention. In this part we number the standard basis of F^4 as $\mathbf{e}_0, \mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$, matching the homogeneous coordinates $[x_0 : x_1 : x_2 : x_3]$ on $\mathbb{P}^3$, and we write $e_{ij} = \mathbf{e}_i \wedge \mathbf{e}_j$ for $i < j$ and

$e_{0123} = \mathbf{e}_0 \wedge \mathbf{e}_1 \wedge \mathbf{e}_2 \wedge \mathbf{e}_3$.

Definition 22.10.3. Plücker Coordinates of a Line

Let $L = \mathbb{P}(W)$ be a line of $\mathbb{P}^3$ and let $(\mathbf{u}, \mathbf{w})$ be a basis of W. The **Plücker coordinates** of L relative to that basis are the six scalars

$$p_{ij} \coloneqq u_i w_j - u_j w_i, \qquad 0 \le i < j \le 3,$$

that is, the 2×2 minors of the 4×2 matrix with columns $\mathbf{u}$ and $\mathbf{w}$. By Theorem 15.9.5 they are the coordinates of the wedge:

$$\mathbf{u} \wedge \mathbf{w} = \sum_{i<j} p_{ij}\, e_{ij}.$$

The point $[\mathbf{u} \wedge \mathbf{w}] \in \mathbb{P}(\Lambda^2 F^4) = \mathbb{P}^5$ is the **Plücker point** of L, written $[L]$.

Proposition 22.10.8. The Plücker Point Determines the Line

Let $L = \mathbb{P}(W)$ be a line of $\mathbb{P}^3$ and let $(\mathbf{u}, \mathbf{w})$ be a basis of W.

(a) $\mathbf{u} \wedge \mathbf{w} \neq 0$, and $[\mathbf{u} \wedge \mathbf{w}]$ does not depend on the choice of basis $(\mathbf{u}, \mathbf{w})$ of W, so $[L]$ is well defined.

(b) $W = \{\, \mathbf{v} \in F^4 : \mathbf{v} \wedge \mathbf{u} \wedge \mathbf{w} = 0 \,\}$. Consequently $L \mapsto [L]$ is injective on lines.

Proof.

(a) A basis is independent, so Theorem 15.8.7 gives $\mathbf{u} \wedge \mathbf{w} \neq 0$. Any other basis is $\mathbf{u}' = a\mathbf{u} + c\mathbf{w}$, $\mathbf{w}' = b\mathbf{u} + d\mathbf{w}$ for a change-of-basis matrix with columns (a, c) and (b, d), which is invertible, so $ad - bc \neq 0$. Expanding,

$$\mathbf{u}' \wedge \mathbf{w}' = ad\, \mathbf{u} \wedge \mathbf{w} + cb\, \mathbf{w} \wedge \mathbf{u} = (ad - bc)\, \mathbf{u} \wedge \mathbf{w},$$

the terms $\mathbf{u} \wedge \mathbf{u}$ and $\mathbf{w} \wedge \mathbf{w}$ vanishing. A nonzero scalar multiple represents the same point of $\mathbb{P}^5$.

(b) By Theorem 15.8.7, $\mathbf{v} \wedge \mathbf{u} \wedge \mathbf{w} = 0$ exactly when $(\mathbf{v}, \mathbf{u}, \mathbf{w})$ is dependent, and since $(\mathbf{u}, \mathbf{w})$ is independent this happens exactly when $\mathbf{v} \in \operatorname{span}(\mathbf{u}, \mathbf{w}) = W$. So W, hence L, is recovered from any representative of $[L]$; and if $[L] = [L']$ their representatives differ by a scalar, which does not change the set in the display, so $W = W'$ and $L = L'$. This proves the proposition.

■

Which points of $\mathbb{P}^5$ are Plücker points? Exactly one equation decides it.

Theorem 22.10.9. The Plücker Relation and the Klein Quadric

Let F be any field and let $0 \neq \omega = \sum_{i<j} p_{ij} e_{ij} \in \Lambda^2 F^4$. Then $\omega = \mathbf{u} \wedge \mathbf{w}$ for some $\mathbf{u}, \mathbf{w} \in F^4$ if and only if

$$p_{01}p_{23} - p_{02}p_{13} + p_{03}p_{12} = 0. \tag{22.10.1}$$

Consequently $L \mapsto [L]$ is a **bijection** from the set of lines of $\mathbb{P}^3$ onto the set

$$\mathcal{K} \coloneqq \{\, [\omega] \in \mathbb{P}^5 : p_{01}p_{23} - p_{02}p_{13} + p_{03}p_{12} = 0 \,\},$$

the **Klein quadric**.

Idea. ($\Rightarrow$) is an identity: substitute $p_{ij} = u_i w_j - u_j w_i$ and watch twelve terms cancel in six pairs. ($\Leftarrow$) needs a way to produce the two vectors, and the right one is to look for them in a kernel. The map $\mathbf{v} \mapsto \mathbf{v} \wedge \omega$ has a matrix, and pairing it against a second vector turns it into a bilinear form $B(\mathbf{v}, \mathbf{v}')\, e_{0123} = \mathbf{v} \wedge \mathbf{v}' \wedge \omega$ — which is alternating, because $\mathbf{v} \wedge \mathbf{v} = 0$. Two facts about alternating matrices then finish it: the determinant of a 4×4 one is the square of its Pfaffian, which here is the left-hand side of Eq. 22.10.1, and the rank of an alternating matrix is even. A vanishing Pfaffian therefore drops the rank from 4 to 2, leaving a kernel of dimension 2, and the two vectors are there.

Proof. ($\Rightarrow$) Suppose $p_{ij} = u_i w_j - u_j w_i$. Expanding the three products,

$$\begin{aligned}
p_{01}p_{23} &= u_0u_2w_1w_3 - u_0u_3w_1w_2 - u_1u_2w_0w_3 + u_1u_3w_0w_2,\\
p_{02}p_{13} &= u_0u_1w_2w_3 - u_0u_3w_1w_2 - u_1u_2w_0w_3 + u_2u_3w_0w_1,\\
p_{03}p_{12} &= u_0u_1w_2w_3 - u_0u_2w_1w_3 - u_1u_3w_0w_2 + u_2u_3w_0w_1.
\end{aligned}$$

Each of the six monomials appears in exactly two of the three lines, and forming $p_{01}p_{23} - p_{02}p_{13} + p_{03}p_{12}$ cancels both copies every time. Monomial by monomial: $u_0u_2w_1w_3$ carries $+1$ in the first line and -1 in the third, contributing $(+1) - 0 + (-1) = 0$; $u_0u_3w_1w_2$ carries -1 in the first and -1 in the second, contributing $(-1) - (-1) + 0 = 0$; $u_1u_2w_0w_3$ likewise $(-1) - (-1) + 0 = 0$; $u_1u_3w_0w_2$ gives $(+1) - 0 + (-1) = 0$; $u_0u_1w_2w_3$ gives $0 - 1 + 1 = 0$; and $u_2u_3w_0w_1$ gives $0 - 1 + 1 = 0$. The total is 0. This direction used nothing about F and did not need $\mathbf{u} \wedge \mathbf{w} \neq 0$.

($\Leftarrow$) Suppose $\omega \neq 0$ satisfies Eq. 22.10.1. Since $\dim \Lambda^4 F^4 = 1$ with basis e_{0123} (Corollary 15.8.6), there is a unique scalar $B(\mathbf{v}, \mathbf{v}')$ with

$$\mathbf{v} \wedge \mathbf{v}' \wedge \omega = B(\mathbf{v}, \mathbf{v}')\, e_{0123},$$

where, expanding ω in the basis, $\mathbf{v} \wedge \mathbf{v}' \wedge \omega = \sum_{i<j} p_{ij}\, \mathbf{v} \wedge \mathbf{v}' \wedge \mathbf{e}_i \wedge \mathbf{e}_j$ is a combination of wedges of **four** vectors. So B is bilinear, the wedge being linear in each slot, and alternating: $B(\mathbf{v}, \mathbf{v}) e_{0123} = \mathbf{v} \wedge \mathbf{v} \wedge \omega = 0$. Let $\mathbf{G}$ be its matrix in the basis $(\mathbf{e}_0, \mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$; by Proposition 14.3.2 (b), $\mathbf{G}$ is an alternating matrix. Each entry is read off $\mathbf{e}_a \wedge \mathbf{e}_b \wedge \mathbf{e}_i \wedge \mathbf{e}_j$, which is $\pm e_{0123}$ when $\{a, b, i, j\} = \{0, 1, 2, 3\}$ and 0 otherwise: for instance $B(\mathbf{e}_0, \mathbf{e}_1) = p_{23}$ because only the term $\mathbf{e}_0 \wedge \mathbf{e}_1 \wedge \mathbf{e}_2 \wedge \mathbf{e}_3 = e_{0123}$ survives, and $B(\mathbf{e}_0, \mathbf{e}_2) = -p_{13}$ because $\mathbf{e}_0 \wedge \mathbf{e}_2 \wedge \mathbf{e}_1 \wedge \mathbf{e}_3 = -e_{0123}$ by Lemma 15.8.1 (a). In full,

$$\mathbf{G} = \begin{pmatrix} 0 & p_{23} & -p_{13} & p_{12} \\ -p_{23} & 0 & p_{03} & -p_{02} \\ p_{13} & -p_{03} & 0 & p_{01} \\ -p_{12} & p_{02} & -p_{01} & 0 \end{pmatrix}.$$

By Example (The Pfaffian in Sizes 2 and 4), a 4×4 alternating matrix with entries a, b, c, d, e, f above the diagonal, read in the order $a_{12}, a_{13}, a_{14}, a_{23}, a_{24}, a_{34}$, has determinant $(af - be + cd)^2$. Here that reads

$$\det \mathbf{G} = \left(p_{23}p_{01} - (-p_{13})(-p_{02}) + p_{12}p_{03}\right)^2 = 0$$

by Eq. 22.10.1. So $\operatorname{rank} \mathbf{G} \le 3$, and $\operatorname{rank} \mathbf{G}$ is **even** by Corollary 14.9.3 (a), hence $\operatorname{rank} \mathbf{G} \le 2$.

Next, $\operatorname{rad}(B) = \{\mathbf{v} : \mathbf{v} \wedge \omega = 0\}$. Indeed if $\mathbf{v} \wedge \omega = 0$ then $B(\mathbf{v}, \mathbf{v}') = 0$ for every $\mathbf{v}'$. Conversely suppose $\eta = \mathbf{v} \wedge \omega \neq 0$ and write $\eta = \sum_S c_S e_S$ over the four 3-element subsets $S \subseteq \{0,1,2,3\}$, with some $c_S \neq 0$; let k be the index not in S. Then $\mathbf{e}_k \wedge \eta = \pm c_S\, e_{0123} \neq 0$, every other term having a repeated factor. Exchanging the first two arguments of a wedge of four vectors changes its sign (Lemma 15.8.1 (a)), so $\mathbf{v} \wedge \mathbf{e}_k \wedge \omega = -\,\mathbf{e}_k \wedge \mathbf{v} \wedge \omega = -\,\mathbf{e}_k \wedge \eta \neq 0$, and therefore $B(\mathbf{v}, \mathbf{e}_k) \neq 0$. (The two one-sided radicals agree because B is alternating, hence skew-symmetric by Theorem 14.3.3 (a).)

By Proposition 14.1.2 (b), $\dim \operatorname{rad}(B) = 4 - \operatorname{rank} \mathbf{G} \ge 2$. Choose independent $\mathbf{u}, \mathbf{w}$ with $\mathbf{u} \wedge \omega = \mathbf{w} \wedge \omega = 0$ and extend to a basis $(\mathbf{u}, \mathbf{w}, \mathbf{x}, \mathbf{y})$ of F^4 (Theorem 1.6.5). By Theorem 15.8.5 the six wedges of pairs from this basis form a basis of $\Lambda^2 F^4$, so

$$\omega = c_1\, \mathbf{u} \wedge \mathbf{w} + c_2\, \mathbf{u} \wedge \mathbf{x} + c_3\, \mathbf{u} \wedge \mathbf{y} + c_4\, \mathbf{w} \wedge \mathbf{x} + c_5\, \mathbf{w} \wedge \mathbf{y} + c_6\, \mathbf{x} \wedge \mathbf{y}.$$

Wedging with $\mathbf{u}$ kills the three terms containing $\mathbf{u}$ and leaves

$$0 = \mathbf{u} \wedge \omega = c_4\, \mathbf{u} \wedge \mathbf{w} \wedge \mathbf{x} + c_5\, \mathbf{u} \wedge \mathbf{w} \wedge \mathbf{y} + c_6\, \mathbf{u} \wedge \mathbf{x} \wedge \mathbf{y},$$

three distinct members of the basis of $\Lambda^3 F^4$ given by Theorem 15.8.5, hence independent; so $c_4 = c_5 = c_6 = 0$. Wedging what remains with $\mathbf{w}$,

$$0 = \mathbf{w} \wedge \omega = -c_2\, \mathbf{u} \wedge \mathbf{w} \wedge \mathbf{x} - c_3\, \mathbf{u} \wedge \mathbf{w} \wedge \mathbf{y},$$

so $c_2 = c_3 = 0$. Hence $\omega = c_1\, \mathbf{u} \wedge \mathbf{w}$ with $c_1 \neq 0$, and $\omega = (c_1 \mathbf{u}) \wedge \mathbf{w}$ is a wedge.

The bijection. If L is a line then $[L] = [\mathbf{u} \wedge \mathbf{w}]$ satisfies Eq. 22.10.1 by ($\Rightarrow$), and the relation is a homogeneous condition of degree 2, so it depends only on the point of $\mathbb{P}^5$. Injectivity is Proposition 22.10.8 (b). For surjectivity, let $[\omega] \in \mathcal{K}$. By ($\Leftarrow$), $\omega = \mathbf{u} \wedge \mathbf{w}$, and $\omega \neq 0$ forces $(\mathbf{u}, \mathbf{w})$ independent by Theorem 15.8.7, so $W = \operatorname{span}(\mathbf{u}, \mathbf{w})$ is 2-dimensional and $L = \mathbb{P}(W)$ is a line with $[L] = [\omega]$. This proves the theorem.

■

Remark. Chapter 15 reaches the same equation by a different route, in Exercise (C1: The Plücker relation): there $\omega \wedge \omega = 2(p_{01}p_{23} - p_{02}p_{13} + p_{03}p_{12})e_{0123}$, so in characteristic $\neq 2$ a decomposable ω, for which $\omega \wedge \omega = 0$, forces the bracket to vanish. That route is credited here, not used. The proof above stands on results proved in the text, and it needs no assumption on the characteristic, which the $\omega \wedge \omega$ route does; Exercise (C2: The Klein quadric) develops the $\omega \wedge \omega$ identity for its own sake.

Warning. **Not every point of $\mathbb{P}^5$ is a line.** The element $\omega = e_{01} + e_{23}$ has $p_{01} = p_{23} = 1$ and all other coordinates 0, so the left-hand side of Eq. 22.10.1 equals 1, and $1 \neq 0$ in every field. It is therefore not a wedge, and $[\omega] \notin \mathcal{K}$. Six numbers are two too many to describe a

line freely; the relation removes one dimension and the projective scaling another.

When $\operatorname{char} F \neq 2$, the left-hand side of Eq. 22.10.1 is a quadratic form on the 6-dimensional space $\Lambda^2 F^4$, so $\mathcal{K}$ is a projective quadric in $\mathbb{P}^5$ in the exact sense of Definition 22.10.1. That is the sentence the section was built to reach: the set of all lines in three-dimensional space is itself a quadric, and everything proved about quadrics in the first half applies to it. Which quadric it is, Exercise (C2: The Klein quadric) works out: it computes the Gram matrix and finds the quadric smooth, with inertia $(3, 3, 0)$ over $\mathbb{R}$. That computation is left to the exercise, and nothing here rests on it.

Example. A line, its Plücker point, and back

Find the Plücker coordinates of the line L of $\mathbb{P}^3(\mathbb{Q})$ through $[1:0:1:0]$ and $[0:1:0:2]$, check the relation, and recover L from the coordinates.

Solution. With $\mathbf{u} = (1, 0, 1, 0)$ and $\mathbf{w} = (0, 1, 0, 2)$, the six minors are

$$\begin{aligned} p_{01} &= 1, & p_{02} &= 0, & p_{03} &= 2, \\ p_{12} &= -1, & p_{13} &= 0, & p_{23} &= 2, \end{aligned}$$

so $\mathbf{u} \wedge \mathbf{w} = e_{01} + 2e_{03} - e_{12} + 2e_{23}$. The relation holds:

$$p_{01}p_{23} - p_{02}p_{13} + p_{03}p_{12} = 1 \cdot 2 - 0 \cdot 0 + 2 \cdot (-1) = 0.$$

To go back, build the matrix $\mathbf{G}$ of the proof:

$$\mathbf{G} = \begin{pmatrix} 0 & 2 & 0 & -1 \\ -2 & 0 & 2 & 0 \\ 0 & -2 & 0 & 1 \\ 1 & 0 & -1 & 0 \end{pmatrix},$$

whose rank is 2: row 3 is -1 times row 1 and row 4 is $-\frac{1}{2}$ times row 2, while rows 1 and 2 are independent. Its null space is cut out by $2v_1 - v_3 = 0$ and $v_0 - v_2 = 0$, that is, by $v_2 = v_0$ and $v_3 = 2v_1$, and is spanned by $(1, 0, 1, 0)$ and $(0, 1, 0, 2)$ — the two points we started from, as Proposition 22.10.8 (b) promises.

22.10.6 Summary and transfer

The chapter began by removing the origin and ended by adding points at infinity, and both moves paid for themselves in theorems that have no exceptional cases. Four habits are worth carrying away.

A geometry is a group acting on a set, and the invariants name the subject. Affine maps preserve parallelism and ratios along a line; Euclidean motions preserve distance; projectivities preserve only incidence and the cross-ratio. Each time, the theorems of the geometry are exactly the statements its group cannot disturb, and the group was a matrix group in disguise: an $(n+1) \times (n+1)$ block matrix for the affine group, an orthogonal matrix plus a translation for motions, and $\mathrm{GL}(V)$ modulo scalars for projectivities.

Adding points removes cases. The affine formula for the dimension of a join needed two cases; its projective counterpart needs none. Two lines in a projective plane always meet. A quadratic form in three variables has one classification list, not a longer one split by whether a center exists. The extra points cost nothing, since $\mathbb{P}(V)$ is built from the same vector space.

Choose the representative. A point of $\mathbb{P}(V)$ is a line of vectors, and nearly every projective proof in this chapter began by picking one vector per point so that the hypothesis became an equation: $a' = a + o$ in Desargues, $u + to$ in Pappus, the normalized second intersection point in the conic parametrization. The choice is where the content is, and it is worth stating out loud each time which scalings are still free afterwards.

Homogeneous means algebraic. A condition that survives scaling is a condition on $\mathbb{P}(V)$: $q(x) = 0$, the Plücker relation, membership of a subspace. That is why each of the objects in this chapter turned out to be the solution set of homogeneous polynomial equations — and the last of them, the Klein quadric, made the lines of $\mathbb{P}^3$ into points of a quadric in $\mathbb{P}^5$, so that a question about lines becomes a question about a symmetric matrix. Chapter 14's classification of forms and Chapter 15's exterior powers were the two tools that made it possible, and neither was built with geometry in mind.

What the chapter did **not** do is also worth recording. It developed no analysis of its own: no limits and no continuity arguments. The only topological words used inside a proof, in Section 3, are quoted from Chapters 16 and 18's theorems, and nothing is said about the topology of $\mathbb{P}^2(\mathbb{R})$, which is a subject of its own. Quadrics required $\operatorname{char} F \neq 2$ everywhere, as Chapter 14 warned they would; Desargues, Pappus, and the Plücker correspondence did not. And the axiomatic side of projective geometry — planes not of the form $\mathbb{P}(V)$, coordinatization by division rings — was pointed at once, in Section 9, and proved nowhere.

22.10.7 Exercises

22.10.7.1 A. Check your understanding

Exercise. A1

(a) State Definition 22.10.1 and say why the zero set is well defined on $\mathbb{P}(V)$.

(b) Which invariant classifies quadrics over an algebraically closed field, and which over $\mathbb{R}$?

(c) List the five projective equivalence classes of conics in $\mathbb{P}^2(\mathbb{R})$ with their zero sets.

(d) State the Plücker relation, and say what the six coordinates are the minors of.

(e) True or false, with a reason: a smooth conic over $\mathbb{R}$ is never empty.

Solution.

(a) The zero set $Z(q) = \{[x] : q(x) = 0\}$ of a nonzero quadratic form q, over a field of characteristic $\neq 2$. It is well defined because $q(\lambda x) = \lambda^2 q(x)$, so for $\lambda \neq 0$ one value is zero exactly when the other is.

(b) Rank alone over an algebraically closed field (Theorem 22.10.1 (a)); rank together with inertia up to interchanging n_+ and n_- over $\mathbb{R}$ (part (b) of the same theorem).

(c) By Corollary 22.10.2 (b): $(1, 0)$ a line; $(2, 0)$ a point; $(1, 1)$ two distinct lines; $(3, 0)$ the empty set; $(2, 1)$ a smooth nonempty conic.

(d) $p_{01}p_{23} - p_{02}p_{13} + p_{03}p_{12} = 0$, where the p_{ij} are the six 2×2 minors of the 4×2 matrix

whose columns are two points spanning the line.

(e) False. $x_0^2 + x_1^2 + x_2^2$ has rank 3, so its conic is smooth, and the zero set in $\mathbb{P}^2(\mathbb{R})$ is empty.

22.10.7.2 B. Practice

Exercise. B1: Classify these conics

Classify each of the following conics of $\mathbb{P}^2$ over $\mathbb{R}$ and over $\mathbb{C}$, giving the rank, the real inertia, and the zero set. Justify your answer.

(a) $2x_0x_1 + 2x_0x_2 + 2x_1x_2 = 0$.

(b) $x_0^2 + 2x_0x_1 + x_1^2 = 0$.

(c) $x_0^2 + 2x_0x_1 + x_1^2 + x_2^2 = 0$.

Solution.

(a) Put $y_0 = x_0 + x_1$ and $y_1 = x_0 - x_1$, so $2x_0x_1 = \frac{1}{2}(y_0^2 - y_1^2)$ and $2x_0x_2 + 2x_1x_2 = 2y_0x_2$. Then

$$q = \tfrac{1}{2}(y_0^2 + 4y_0x_2) - \tfrac{1}{2}y_1^2 = \tfrac{1}{2}(y_0 + 2x_2)^2 - 2x_2^2 - \tfrac{1}{2}y_1^2.$$

The three forms $y_0 + 2x_2 = x_0 + x_1 + 2x_2$, $y_1 = x_0 - x_1$ and x_2 are independent, the matrix with those rows having determinant -2. So the rank is 3 and the real inertia is $(1, 2, 0)$, which is $(2, 1, 0)$ after a sign change; by Corollary 22.10.2 the zero set is a smooth nonempty conic over $\mathbb{R}$ and a smooth conic over $\mathbb{C}$. For instance $[1 : 0 : 0]$ lies on it, every term of q having a factor x_1 or x_2.

(b) $q = (x_0 + x_1)^2$, rank 1, inertia $(1, 0, 2)$. The zero set is the line $\{x_0 + x_1 = 0\}$ over either field.

(c) $q = (x_0 + x_1)^2 + x_2^2$, rank 2, inertia $(2, 0, 1)$. Over $\mathbb{R}$ the zero set is the single point $[1 : -1 : 0]$. Over $\mathbb{C}$ the form factors as $(x_0 + x_1 + ix_2)(x_0 + x_1 - ix_2)$, so the zero set is two distinct lines.

Exercise. B2: A conic through five points

Find the conic of $\mathbb{P}^2(\mathbb{Q})$ through $[1 : 0 : 0]$, $[0 : 1 : 0]$, $[0 : 0 : 1]$, $[1 : 1 : 1]$ and $[1 : -1 : 2]$, verify that no three of the five points are collinear, and decide whether the conic is smooth.

Solution. Write $q = ax_0^2 + bx_1^2 + cx_2^2 + dx_0x_1 + ex_0x_2 + fx_1x_2$. The three standard basis points give $a = b = c = 0$. Then $[1 : 1 : 1]$ gives $d + e + f = 0$ and $[1 : -1 : 2]$ gives $-d + 2e - 2f = 0$. Adding, $3e - f = 0$, so $f = 3e$ and $d = -e - f = -4e$. Taking $e = 1$,

$$q = -4x_0x_1 + x_0x_2 + 3x_1x_2.$$

No three of the five are collinear: the determinants of the 3×3 matrices formed by any three of the five coordinate vectors are, in the order 123, 124, 125, 134, 135, 145, 234, 235, 245, 345,

$$1,\ 1,\ 2,\ -1,\ 1,\ 3,\ 1,\ 1,\ -1,\ -2,$$

none of them zero.

The matrix is

$$\mathbf{A} = \begin{pmatrix} 0 & -2 & 1/2 \\ -2 & 0 & 3/2 \\ 1/2 & 3/2 & 0 \end{pmatrix}, \qquad \det \mathbf{A} = -3 \neq 0,$$

so the conic is smooth, and by the argument of Example (A conic through five points) (product of eigenvalues negative, trace zero) its real inertia is $(2, 1, 0)$.

Exercise. B3: Plücker coordinates

(a) Compute the Plücker coordinates of the line of $\mathbb{P}^3(\mathbb{Q})$ through $[1 : 1 : 0 : 0]$ and $[0 : 0 : 1 : 1]$, and check Eq. 22.10.1.

(b) Determine whether $[\omega]$ with $(p_{01}, p_{02}, p_{03}, p_{12}, p_{13}, p_{23}) = (1, 0, 0, 0, 0, 1)$ is the Plücker point of a line. Justify your answer.

(c) The six numbers $(1, 1, 0, 0, -1, -1)$, in the same order, satisfy Eq. 22.10.1. Find the line they come from.

Solution.

(a) With $\mathbf{u} = (1, 1, 0, 0)$ and $\mathbf{w} = (0, 0, 1, 1)$,

$$\begin{aligned} p_{01} &= 0, \quad p_{02} = 1, \quad p_{03} = 1, \\ p_{12} &= 1, \quad p_{13} = 1, \quad p_{23} = 0, \end{aligned}$$

and $0 \cdot 0 - 1 \cdot 1 + 1 \cdot 1 = 0$.

(b) No. The left-hand side of Eq. 22.10.1 is $1 \cdot 1 - 0 + 0 = 1 \neq 0$, so by Theorem 22.10.9 the element $e_{01} + e_{23}$ is not a wedge and its point is not on the Klein quadric.

(c) First, $1 \cdot (-1) - 1 \cdot (-1) + 0 \cdot 0 = 0$, as stated. Build $\mathbf{G}$ as in the proof of Theorem 22.10.9:

$$\mathbf{G} = \begin{pmatrix} 0 & -1 & 1 & 0 \\ 1 & 0 & 0 & -1 \\ -1 & 0 & 0 & 1 \\ 0 & 1 & -1 & 0 \end{pmatrix}.$$

Its null space is given by $-v_1 + v_2 = 0$ and $v_0 - v_3 = 0$, so it is spanned by $(1, 0, 0, 1)$ and $(0, 1, 1, 0)$. The line is the one through $[1 : 0 : 0 : 1]$ and $[0 : 1 : 1 : 0]$; as a check, its minors are $p_{01} = 1$, $p_{02} = 1$, $p_{03} = 0$, $p_{12} = 0$, $p_{13} = -1$, $p_{23} = -1$, as given.

22.10.7.3 C. Going deeper

Exercise. C1: All rational points of a circle

Work in $\mathbb{P}^2(\mathbb{Q})$ with $C = Z(x_0^2 + x_1^2 - x_2^2)$.

(a) Prove that C is smooth and that $P_0 = [0 : 1 : 1]$ lies on it.

(b) Using Theorem 22.10.5 with $\ell = \{x_1 = 0\}$, compute the resulting parametrization of

C explicitly.

(c) Deduce that every solution of $a^2+b^2=c^2$ in integers with $c\neq 0$ is a rational multiple of $(2uv,\ u^2-v^2,\ u^2+v^2)$ for some integers u,v, and exhibit the triple obtained at $[u:v]=[2:1]$.

Solution.

(a) The matrix is $\operatorname{diag}(1,1,-1)$, invertible, so C is smooth; and $0+1-1=0$, so $P_0\in C$. Also $P_0\notin\ell$, since its x_1-coordinate is $1\neq 0$.

(b) Here $\mathbf{p}=(0,1,1)$ and $\beta(\mathbf{x},\mathbf{y})=x_0y_0+x_1y_1-x_2y_2$. A point of ℓ is $[u:0:v]$, with $q(\mathbf{x})=u^2-v^2$ and $\beta(\mathbf{p},\mathbf{x})=-v$. So

$$\begin{aligned}\psi([u:0:v]) &= [-(u^2-v^2)(0,1,1)-2v(u,0,v)]\\ &= [-2uv\ :\ v^2-u^2\ :\ -u^2-v^2]\\ &= [2uv\ :\ u^2-v^2\ :\ u^2+v^2].\end{aligned}$$

Indeed $(2uv)^2+(u^2-v^2)^2=(u^2+v^2)^2$.

(c) Let $a^2+b^2=c^2$ with a,b,c integers and $c\neq 0$. Then $[a:b:c]$ is a point of C with rational coordinates, so by Theorem 22.10.5 it equals $\psi([u:0:v])$ for exactly one point of ℓ, which has rational, hence after clearing denominators integer, coordinates u,v. Therefore (a,b,c) is a rational multiple of $(2uv,\ u^2-v^2,\ u^2+v^2)$. At $[u:v]=[2:1]$ this is $(4,3,5)$.

Exercise. C2: The Klein quadric

Let $\operatorname{char}F\neq 2$, write $Q(\omega)=p_{01}p_{23}-p_{02}p_{13}+p_{03}p_{12}$ for $\omega\in\Lambda^2F^4$, and let β_Q be the symmetric bilinear form with $\beta_Q(\omega,\omega)=Q(\omega)$, which exists and is unique by Theorem 14.3.5.

(a) Prove that $\omega\wedge\omega=2Q(\omega)\,e_{0123}$ for every $\omega\in\Lambda^2F^4$, and deduce $\omega\wedge\omega'=2\beta_Q(\omega,\omega')\,e_{0123}$.

(b) Prove that two lines L,L' of $\mathbb{P}^3$ meet if and only if $\beta_Q([L],[L'])=0$, the value being computed from any representatives.

(c) Compute the matrix of β_Q in the basis ordered $(e_{01},e_{23},e_{02},e_{13},e_{03},e_{12})$. Deduce that Q has rank 6, so that the Klein quadric is smooth, and that over $\mathbb{R}$ its inertia is $(3,3,0)$.

Hint: for (a), expand and note that only complementary pairs of index sets survive; the same computation appears in Exercise (C1: The Plücker relation) (a).

Solution.

(a) Expanding $\omega\wedge\omega=\sum_{S,T}p_Sp_T\,e_S\wedge e_T$ over 2-element index sets, a term vanishes unless S and T are disjoint, so only the pairs $\{01\},\{23\}$, $\{02\},\{13\}$, $\{03\},\{12\}$ and their reverses survive. A reversal does not change $e_S\wedge e_T$ for $|S|=|T|=2$: passing each of the two factors of e_T across the two factors of e_S costs four transpositions, and a permutation that is a product of four transpositions has sign $+1$ (Corollary 7.3.3), so by Lemma 15.8.1 (b) the value is unchanged. Sorting, $e_{01}\wedge e_{23}=e_{0123}$, $e_{02}\wedge e_{13}=-e_{0123}$ (one transposition), and $e_{03}\wedge e_{12}=e_{0123}$ (two). So $\omega\wedge\omega=2(p_{01}p_{23}-p_{02}p_{13}+p_{03}p_{12})e_{0123}$.
For the second identity expand $(\omega+\omega')\wedge(\omega+\omega')$ in two ways. Wedging is symmetric

on 2-vectors, by bilinearity and the remark just made, so the cross terms combine: $2Q(\omega+\omega')e_{0123} = 2Q(\omega)e_{0123} + 2\,\omega\wedge\omega' + 2Q(\omega')e_{0123}$. Since $\beta_Q(\omega,\omega') = \frac{1}{2}(Q(\omega+\omega') - Q(\omega) - Q(\omega'))$ by Theorem 14.3.5, this rearranges to $\omega\wedge\omega' = 2\beta_Q(\omega,\omega')e_{0123}$.

(b) Write $L = \mathbb{P}(W)$, $L' = \mathbb{P}(W')$ with bases $(\mathbf{u},\mathbf{w})$ and $(\mathbf{u}',\mathbf{w}')$. By (a), $\beta_Q = 0$ on the pair exactly when $\mathbf{u}\wedge\mathbf{w}\wedge\mathbf{u}'\wedge\mathbf{w}' = 0$, which by Theorem 15.8.7 happens exactly when $(\mathbf{u},\mathbf{w},\mathbf{u}',\mathbf{w}')$ is dependent, that is, when $\dim(W+W') \le 3$. By Theorem 1.7.3 this says $\dim(W\cap W') \ge 2+2-3 = 1$, which says $L\cap L' \neq \emptyset$. Rescaling a representative rescales β_Q by a nonzero factor, so vanishing does not depend on the choice.

(c) In that order the matrix is block diagonal with the three 2×2 blocks

$$\begin{pmatrix} 0 & 1/2 \\ 1/2 & 0 \end{pmatrix}, \qquad \begin{pmatrix} 0 & -1/2 \\ -1/2 & 0 \end{pmatrix}, \qquad \begin{pmatrix} 0 & 1/2 \\ 1/2 & 0 \end{pmatrix},$$

since Q is a sum of three products of complementary coordinates. Each block has determinant $-1/4 \neq 0$, so the whole matrix is invertible and $\operatorname{rank} Q = 6$: the quadric is smooth. Over $\mathbb{R}$ each block has eigenvalues $\pm 1/2$, one positive and one negative, so the inertia is $(3,3,0)$.

Chapter 23

Algebras and Representations

Twenty-one chapters have studied operators one at a time, or in commuting families. Chapter 9 asked for a basis in which a single operator is simple and found eigenvectors; Chapter 10 finished the question with the Jordan and rational forms; Chapters 11 to 13 answered it again with an inner product to respect. Then the question widened once, in Chapter 9 §09 and Chapter 12 §07: one basis good for a whole commuting family at a time. This chapter widens it once more, and to the last setting in which the question still has a clean answer — a set of operators closed under addition, under scalars, and under composition, with no assumption that its members commute. Such a set is an **algebra**, and the chapter's main theorem says that over an algebraically closed field there is essentially one building block, a full matrix algebra, and that everything semisimple is a product of those.

Nothing forced the earlier stopping point except habit. The polynomials in a fixed operator form an algebra; so do the circulants of Chapter 12 §09, the matrices commuting with a given matrix, the upper triangular matrices, the block diagonal matrices, and the linear symmetries of a figure. Every one of those the book has met and handled separately. What they have in common is the only thing this chapter needs, and naming it is most of the work.

The second half is the same statement read for groups. A group cannot be added, so linear algebra cannot reach it directly; but the formal combinations of group elements do form an algebra, the **group algebra** $F[G]$, and a group acting by invertible operators is exactly a representation of that algebra. Everything proved about algebras then applies word for word. Maschke's theorem says $F[G]$ is semisimple as soon as the characteristic of F does not divide $|G|$; the structure theorem splits it into matrix blocks; and the traces of those blocks — the **characters** — turn out to be an orthonormal basis of the functions that are constant on conjugacy classes. One square array of numbers then records every complex representation of the group, up to equivalence, and the number of rows is the number of conjugacy classes.

One debt is paid in full. For a finite abelian group every block has size one, the character table is a single unitary matrix, and for the cyclic group that matrix is Chapter 12 §09's Fourier matrix, entry for entry. The discrete Fourier transform, the convolution theorem, and "every circulant is a polynomial in the shift" were representation theory all along, and §09 proves each of the three identifications.

This is the last of the book's chapters of general theory. The chapter that follows is about arithmetic, algorithms, and applications, and nothing here depends on it.

23.0.1 Conventions

- **Algebras are associative, unital, and finite-dimensional over** F, unless a statement says otherwise. The one standing exception is $F[x]$, kept as a source of counterexamples and always named as infinite-dimensional where it is used that way. Nothing requires commutativity, and nothing requires a nonzero element to have an inverse: an algebra is not a group, and §01 closes with a warning to that effect.
- **The definition is already in the book.** Chapter 14 §12's Definition 14.12.1 is used as it stands; §01 builds on it rather than restating it, adding Definition 23.1.1, Definition 23.1.2, Definition 23.1.3, and Definition 23.1.4.
- $\mathcal{L}(V)$ **and** $\mathrm{End}(V)$ **name the same object.** The book's default is $\mathcal{L}(V)$; §§05, 06, and 07 write $\mathrm{End}(V)$ where the algebra structure is the point. Likewise $\mathrm{End}_A(V) \subseteq \mathcal{L}(V)$, with the subscript never dropped, is the set of maps respecting an action of A.
- $F\langle S\rangle$ **is the subalgebra generated by** S and $\langle p\rangle$, without the F, is Chapter 6's ideal generated by p. They are different and usually different sizes; §01 says so where the notation is introduced.
- S' **is the commutant**, the set of elements commuting with every member of S; $S'' = (S')'$ is the double commutant, and $Z(A) = A'$ is the center. The definition asks nothing of S — not that it be a subspace, not that its own members commute.
- **"Simple" is two words.** An **algebra** is simple when its only two-sided ideals are $\{0\}$ and itself; a **space** on which an algebra acts is simple, equivalently irreducible, when its only invariant subspaces are $\{0\}$ and itself. Neither implies the other, and §03 opens with the warning and the witness. **"Irreducible" is three words**, since Chapter 6 calls a polynomial irreducible (Definition 6.3.1) and Chapter 19 §02 calls a nonnegative matrix irreducible; §03's proof of Theorem 23.3.7 uses the first and the second within two lines of each other, and the argument is what tells them apart.
- **"Semisimple" is three words, and §03 reconciles them.** An algebra **acts semisimply on a space** when every invariant subspace has an invariant complement; an **algebra is semisimple** when it acts semisimply on itself, that is, when every left ideal has a complementary left ideal; and Chapter 10 §08 called an **operator** semisimple when its minimal polynomial is a product of distinct irreducibles. Theorem 23.3.7 proves the third is the first, applied to $F[T]$.
- $\mathrm{rad}\, A$ **is the radical of an algebra** (Definition 23.3.5), the set of elements killed by every irreducible representation. Chapter 14 §01's $\mathrm{rad}(\beta)$ is the radical of a bilinear form; the argument tells them apart and they never meet.
- $\mathbf{E}_{ij}$ **is the matrix unit** with a 1 in position (i, j), and δ_g is the basis vector of $F[G]$ attached to the group element g, so that $\delta_g\delta_h = \delta_{gh}$. The Kronecker delta always carries two numerical subscripts, δ_{jk}, and no group element is ever written as a pair of numbers.
- **Groups are finite from §07 on**, written multiplicatively. $\chi_\rho(g) = \mathrm{tr}\,\rho(g)$ is the character of a representation; only for an abelian group is a character a homomorphism, and §09 says so before adopting that reading and writing $\widehat{G}$ for the group of them.

23.0.1.1 Which field, and where the hypothesis bites

Most theorems in this chapter are false over the wrong field, and it is worth saying once, exactly, which ones need what. §10 sets the same information out as a table, with a counterexample in every row.

Sections 1, 2, and 3 work over an arbitrary field F**, with no hypothesis at all.** That includes the chapter's first real theorem, Theorem 23.2.2, whose proof uses nothing about F beyond the existence of inverses of nonzero scalars; the classification of the left ideals of $M_n(F)$; the first isomorphism theorem for algebras; and Theorem 23.3.3, which is the section's engine and holds over $\mathbb{F}_2$ as readily as over $\mathbb{C}$.

Algebraic closure is asked for at exactly five places. The second half of Schur's lemma, Theorem 23.4.1 (b), and its three corollaries Corollary 23.4.3, Corollary 23.4.4, and Corollary 23.4.5; Burnside's theorem Theorem 23.5.2; and Wedderburn's structure theorem Theorem 23.6.1. In every one of them the hypothesis buys the same single thing — an operator on a nonzero space has an eigenvalue — and that is the only thing it buys. The **first** half of Schur's lemma needs nothing of the field and is used over arbitrary fields. **The double commutant theorem needs nothing of the field either**: Theorem 23.5.5 holds over every F, and what it asks instead is that A act semisimply and contain id_V. Both of those are shown by §05 to be necessary.

One matrix is the witness for every failure of algebraic closure. Rotation of $\mathbb{R}^2$ through $2\pi/n$ with $n \geq 3$ has no real eigenvalue. The algebra it generates is commutative, it acts irreducibly on $\mathbb{R}^2$, its commutant is a copy of $\mathbb{C}$ rather than $\mathbb{R}$, and it is a two-dimensional subalgebra of $M_2(\mathbb{R})$ and so is not all of $\mathrm{End}(\mathbb{R}^2)$. That one example breaks Schur's second half, breaks the corollary that irreducible plus commutative means one-dimensional, and breaks Burnside. §04 computes the commutant in full.

A hypothesis on the characteristic enters only with groups, and only in one theorem. Theorem 23.7.3 needs $\mathrm{char}\, F \nmid |G|$, because its proof divides by $|G|$; the witness is $\mathbb{Z}/p\mathbb{Z}$ acting on $\mathbb{F}_p^2$ by unipotent matrices, where the invariant line has no invariant complement. Finiteness of G is used twice over — to average and to count — and §07 shows $\mathbb{Z}$ acting on $\mathbb{Q}^2$ failing in characteristic zero for the other reason. Everything in §07 up to that point is over an arbitrary field.

Sections 8 and 9 fix $F = \mathbb{C}$ throughout, where both hypotheses hold at once and neither can fail: $\mathbb{C}$ is algebraically closed, so Schur's lemma gives scalars, and $\mathrm{char}\, \mathbb{C} = 0$ divides no group order, so the averaging is legal. That is why the complex theory is so much cleaner than the general one. The modular case, where the characteristic does divide $|G|$, is a different subject and the chapter does not enter it.

One further hypothesis is local. Proposition 23.4.6, which says an algebra closed under adjoints acts semisimply, lives on a real or complex inner product space; it is the reason Chapter 12 §07 needed no averaging, and it is used nowhere else.

23.0.2 What the chapter quotes, and what it does not prove

The chapter imports no analysis. No limit, no continuity, no compactness argument appears in it, and it adds nothing to the list of analytic facts **(A1)–(A6)** that Chapter 16 introduced — Chapter 16's numbering, not §01's algebra axioms (A1)–(A3) — it does not use that list at all. Roots of unity are Chapter 0 §06's complex numbers and nothing more.

Six things are quoted with credit and proved nowhere in the book. Each is flagged again where it appears, and **nothing in the chapter rests on any of them.** No theorem here rests on an exercise either; where an exercise is pointed at, the text says that an exercise is not a proof.

- **Section 6 — the Wedderburn–Artin theorem over a division ring**: that a finite-dimensional semisimple algebra over *any* field is a product $M_{n_1}(D_1) \times \cdots \times M_{n_k}(D_k)$ with each D_i a division algebra. The chapter proves only the algebraically closed case, where every D_i is F. The reason given is the honest one: the book has never built division rings, and the quaternions are one example, not a theory.
- **Section 6 — Wedderburn's little theorem**, that a finite division ring is a field. Same reason. Its effect on the theorem above is recorded and then not used.

- **Section 8 — that the character table does not determine the group.** Two nonisomorphic groups of order 8, the symmetries of a square and the unit quaternions, have identical tables. §08 names them in a warning and constructs neither; the book has neither the group theory to do it nor any need of the fact. What the character table *does* determine — every complex representation of the group up to equivalence — is proved, in Corollary 23.8.8.
- **Section 9 — the structure theorem for finite abelian groups**, that every one is a direct product of cyclic groups. What §09 actually proves is $|\widehat{G}| = |G|$ for every finite abelian G (Proposition 23.9.3), and $\widehat{G} \cong G$ for cyclic groups and for finite direct products of them (Theorem 23.9.4). The general isomorphism is stated as true and not proved, and §09 says which of its later statements would need it: none. The remark points at Chapter 10 §07's Smith normal form as the linear-algebra shadow of the missing argument, and at an exercise there, and adds that the exercise assumes an input it does not prove.
- **Section 3 — the converse of Proposition 23.3.6**, that a finite-dimensional algebra with zero radical is semisimple. Not proved and never used: whenever semisimplicity is needed later it is established directly, by Maschke's theorem or by exhibiting a decomposition.
- **Section 1 — when the commutant of a single matrix is the algebra of polynomials in it.** The exact condition is that the minimal and characteristic polynomials agree. §01 computes two commutants where they do and one where they do not, states the criterion in a remark with credit, and proves nothing about it.

Two further names are mentioned and no statement of theirs is used. §05 records that its proof of the double commutant theorem is the finite-dimensional shadow of Jacobson's density theorem, and that there is an analytic theorem of the same shape, von Neumann's, for operators on a Hilbert space. Neither is proved and neither is needed.

Four neighboring subjects are left outside on purpose, and §10 lists them with the reason in each case: induced representations and Frobenius reciprocity, which need cosets of a subgroup that Chapter 0 does not have; modular representation theory, which is what happens when Maschke's theorem fails; character tables as a computational craft, of which the chapter computes three small ones; and representations of infinite groups, where every proof here loses both of its uses of finiteness.

23.0.3 What you need

The chapter leans on Chapters 0 to 6 and on Chapters 9, 10, and 12, and takes one or two facts each from Chapters 7, 8, 11, 14, and 15. **Chapters 13 and 16 to 22 are not cited anywhere in it.** No norm, no singular value, no positive semidefinite matrix, no matrix inequality, no convexity, and no geometry appears at any point. Eigenvalues enter only through Chapter 9, and almost always through one fact: over an algebraically closed field, an operator on a nonzero space has one.

Chapter 0 §10 is all the group theory the book has — groups, subgroups, homomorphisms, the symmetric group and cycle notation, and nothing else. There is no Lagrange's theorem, no coset of a subgroup, and no quotient group. Where this chapter needs more than that, it builds it in place and says so: conjugacy classes, and the proof that they partition the group, in §08; the order of an element and the cyclic subgroup it generates, and the direct product of two groups, in §09; and, in an exercise of §01, the class sums that span the center of $F[G]$.

- **Chapter 0** — matrix multiplication and the three views of a product, which carry §01's and §02's computations with matrix units; the characteristic of a field and the arithmetic

of integer multiples of 1_F, which is exactly Maschke's hypothesis; the complex numbers and their roots of unity, which are every value a character takes; and equivalence relations and partitions, for the conjugacy classes: Definition 0.9.9, Theorem 0.9.4, Theorem 0.9.2, Definition 0.7.2, Lemma 0.7.4, Theorem 0.8.5, Example (Roots of Unity), Theorem 0.6.4, Definition 0.5.2, Theorem 0.5.2, Theorem 0.2.3, Theorem 0.2.10, Definition 0.10.1, Definition 0.10.2, Definition 0.10.3, Example (Groups), Example (Computing with cycles), Theorem 0.10.1, Theorem 0.10.3, Theorem 0.10.2, Theorem 0.10.4.

- **Chapter 1** — the constant companion: subspaces, spans, bases, and dimension, and above all the direct sum criteria, since every decomposition theorem in the chapter is a statement that some sum is direct. Theorem 1.7.6 supplies the ordinary complement that Maschke's averaging repairs, and Theorem 1.6.8 is how a minimal invariant subspace is recognized as irreducible: Example (The Standard Vector Spaces), Theorem 1.2.1, Theorem 1.2.2, Definition 1.3.2, Theorem 1.3.1, Example (Standard Bases), Theorem 1.6.5, Theorem 1.6.8, Theorem 1.6.4, Theorem 1.7.6, Theorem 1.7.4, Theorem 1.7.5, Theorem 1.7.3.
- **Chapter 3** — three facts. Permutation matrices are how the symmetric group becomes operators, and the theorem that a one-sided inverse of a square matrix is two-sided is what turns the row orthogonality of characters into column orthogonality in one line: Definition 3.4.2, Lemma 3.4.7, Theorem 3.4.5.
- **Chapter 2** — the workhorse, and the chapter's largest debt after Chapter 1. Kernels and images give every invariance argument, and Rank–Nullity is used in four sections: Theorem 2.1.6, Theorem 2.2.1, Theorem 2.2.2, Theorem 2.2.3, Theorem 2.3.1, Theorem 2.4.2, Theorem 2.5.1, Theorem 2.5.3, Theorem 2.5.4.
- **Chapter 4** — products and quotients of spaces give the quotient algebra, the first isomorphism theorem, and the n-fold copy in the proof of the double commutant theorem; the matrix of a map turns $\mathcal{L}(V)$ into $M_n(F)$ as algebras; and projections and the trace are the whole of §08: Theorem 4.1.4, Theorem 4.1.2, Definition 4.4.1, Definition 4.4.3, Theorem 4.4.1, Theorem 4.4.5, Theorem 4.4.7, Definition 4.5.1, Theorem 4.5.1, Definition 4.5.2, Theorem 4.5.3, Theorem 4.5.4.
- **Chapter 5** — the dual space, in two places. Burnside's proof runs on rank-one operators written as $v\varphi$, so it needs a dual basis and $\dim V^* = \dim V$; and the annihilator and the dual map appear in the exercises of §§02 and 07, which classify the right ideals of $M_n(F)$ and build the contragredient representation: Theorem 5.1.1, Definition 5.1.3, Corollary 5.1.4, Definition 5.2.1, Theorem 5.2.3, Theorem 5.2.2, Corollary 5.3.5, Definition 5.4.1, Theorem 5.4.1, Theorem 5.6.7.
- **Chapter 6** — polynomials. Division with remainder and the minimal polynomial give $\dim F[T] = \deg m_T$; the ideals of $F[x]$ are the commutative model §02 generalizes; unique factorization and the primary decomposition are what §03 needs to identify Chapter 10's semisimple operator with this chapter's; the Fundamental Theorem of Algebra is where "algebraically closed" is grounded; and Lagrange interpolation writes the projections onto the eigenspaces of a diagonalizable operator as polynomials in it, which is how §05 computes a commutant: Theorem 6.1.1, Theorem 6.1.3, Definition 6.2.1, Example (First Ideals), Proposition 6.2.1, Definition 6.3.1, Theorem 6.3.5, Theorem 6.5.1, Theorem 6.6.5, Theorem 6.6.1, Theorem 6.7.3, Theorem 6.7.1.
- **Chapter 7 §03** — one fact, used for one representation: the sign of a permutation is multiplicative, which is what makes the sign representation of S_n a representation: Definition 7.3.2, Theorem 7.3.1. It is named where §07 lists its examples, and checked in one of that section's exercises; no proof here uses it.
- **Chapter 8 §01** — one fact: block diagonal matrices multiply blockwise, which is how §06's worked example recognizes a product of matrix algebras inside $M_n(F)$: Theorem 8.1.4. It is used there and nowhere else, and not inside a proof.
- **Chapter 9** — invariant subspaces, eigenvalues, and the commuting-operator theorems

this chapter re-derives and explains. Theorem 9.2.1 and Theorem 9.2.4 are the single fact algebraic closure is spent on; Theorem 9.8.1 and the cyclic machinery settle §03's comparison with Chapter 10; and §§04 and 10 look back at Theorem 9.9.3 and Theorem 9.9.5 from the new height: Definition 9.1.1, Proposition 9.1.2, Example (A Rotation Has No Invariant Lines), Theorem 9.2.1, Theorem 9.2.4, Theorem 9.2.2, Theorem 9.4.1, Theorem 9.5.1, Theorem 9.8.1, Theorem 9.8.2, Corollary 9.8.3, Lemma 9.9.2, Theorem 9.9.1, Theorem 9.9.3, Theorem 9.9.5.

- **Chapter 10** — three separate debts. The cyclic subspace machinery of §05 is what §03 needs to prove that a semisimple operator is one whose generated algebra acts semisimply; the Weyr commutant is re-read in §05 and again in §10 as the case where the double commutant theorem has nothing to say; and the Smith normal form is what §09 points at for the gap it cannot close: Definition 10.5.2, Proposition 10.5.1, Theorem 10.5.2, Theorem 10.5.3, Proposition 10.4.6, Theorem 10.7.3, Exercise (C1: The same algorithm over the integers) (pointed at, not used), Definition 10.8.1, Proposition 10.8.2.
- **Chapter 11** — the inner product, in three places and never as a standing assumption: §04's proposition that an adjoint-closed algebra acts semisimply, §08's inner product on class functions, and §09's proof that the character matrix is unitary: Definition 11.1.1, Theorem 11.1.4, Definition 11.3.1, Theorem 11.3.2, Definition 11.6.1, Theorem 11.6.1, Theorem 11.7.3.
- **Chapter 12** — the chapter this one explains. Section 9's circulants, the Fourier matrix, the convolution theorem, and the diagonalization of every circulant at once are §09's subject; §07's commuting normal families and §10's disjoint-spectra commutant are re-read in §§04 and 10: Corollary 12.4.5, Theorem 12.7.1, Definition 12.9.1, Corollary 12.9.5, Theorem 12.9.1, Theorem 12.9.4, Theorem 12.9.6, Theorem 12.9.3, Theorem 12.9.2, Example (The Fourier matrix of size four), Example (The running example, diagonalized), Corollary 12.10.4.
- **Chapter 14 §12** — the definition of an algebra, which this chapter uses as it stands, and the quaternions, which are the chapter's only noncommutative division algebra and are named as such twice: Definition 14.12.1, Definition 14.12.3. Both are carried over as definitions, in §01's own definition of an algebra homomorphism and in its examples; no proof here cites Chapter 14.
- **Chapter 15** — the free vector space on a set, which is what the group algebra is built from; the fact that a multilinear map may be prescribed freely on bases, which is how the group algebra's multiplication is shown to exist and be unique; and the two-sided ideal and quotient algebra of §10, which §02 takes over without reproving: Theorem 15.1.1, Definition 15.2.3, Lemma 15.2.2, Definition 15.10.3, Lemma 15.10.5.

23.0.4 Roadmap

- **Associative Algebras.** The vocabulary, and the two examples the rest of the chapter runs on. Subalgebra, with the identity demanded by name and a set exhibited that satisfies every other clause without being one; homomorphism, kernel, and image, with transposition as the nonexample that reverses products; the subalgebra generated by a set, which for one operator is $F[T]$ of dimension $\deg m_T$. Then the group algebra $F[G]$, built as the free vector space on G with the multiplication forced by bilinearity, and commutative exactly when G is abelian. Cayley's theorem for algebras: left multiplication embeds any finite-dimensional algebra in the operators on itself, so matrices are the general case and not a special one. Finally commutants, three of them computed by comparing entries, and the proof over every field that the center of $M_n(F)$ is the scalars.
- **Ideals and Quotients.** In a noncommutative algebra there are three kinds of ideal, and which kind is meant decides every later statement. One product rule does the work —

sandwiching a matrix between two matrix units extracts one entry and puts it anywhere — and out of it comes the chapter's first real theorem: over **every** field, $M_n(F)$ has no two-sided ideals but $\{0\}$ and itself. The left ideals are another matter, and they are classified completely, by subspaces of row vectors, with the column ideals minimal among them and the transposed statement for right ideals recorded so the two are not confused. Then quotient algebras and the first isomorphism theorem, with two instances: the group algebra of a cyclic group is $F[x]$ with one relation, and the upper triangular 2×2 matrices modulo their strictly upper part are $F \times F$.

- **Simple and Semisimple Algebras.** A second sense in which something cannot be broken apart — this time a space rather than an algebra — and the section opens by warning that both are spelled "simple". Representation of an algebra, invariant subspace, irreducible space. Then the theorem, over any field: every invariant subspace has an invariant complement if and only if the space is a sum of irreducibles, if and only if it is a direct sum of finitely many. An algebra semisimple in its own regular representation acts semisimply on every space. The radical collects what no irreducible representation can see; a nilpotent ideal always lies inside it, a semisimple algebra has none, and the upper triangular matrices have a one-dimensional one, computed. Finally the reconciliation with Chapter 10: an operator is semisimple exactly when $F[T]$ acts semisimply — proved with no hypothesis on the field, because the argument never leaves F.
- **Schur's Lemma.** Maps that respect the action, and how few of them there are. Part (a), that a map between irreducibles is zero or an isomorphism, needs nothing of the field and says the commutant of an irreducible action is a division algebra. Part (b), that over an algebraically closed field the division algebra is F, needs an eigenvalue and nothing else. The failure over $\mathbb{R}$ is then exhibited in full, the commutant of the rotation representation computed by multiplying out and identified with $\mathbb{C}$, every nonzero element invertible exactly as part (a) predicts. Two corollaries: an operator commuting with an irreducible family is a scalar, and an irreducible space for a commutative algebra is a line. Then the look back — Chapter 9 §09's common eigenvector was this, its clever trick Schur's lemma unnamed — and a proposition that an adjoint-closed algebra acts semisimply, which recovers Chapter 12 §07 for commuting self-adjoint families out of two general facts. A warning closes it: "the commutant is the scalars" follows from irreducibility and is never a test for it.
- **Burnside and the Double Commutant.** Two theorems about how much of an algebra its commutant records. Burnside: over an algebraically closed field, a subalgebra of $\mathrm{End}(V)$ containing the identity and acting irreducibly is **all** of $\mathrm{End}(V)$ — irreducibility, a statement about subspaces, forces the largest dimension there is. The proof takes a member of least rank, shows that rank is one, and spreads one rank-one operator over the whole space in two directions; a version for a multiplicative set rather than an algebra is recorded alongside it. The double commutant theorem: if A acts semisimply and contains id_V then $A'' = A$, over **any** field. Its engine is matching a single vector, and the upgrade to a whole basis is done by copying the space n times. Both hypotheses are shown necessary and neither is shown necessary in the converse direction. The section ends by re-reading Chapter 10's Weyr commutant and Chapter 12's disjoint-spectra commutant as computations of A', and by doing the two extremes, the diagonal matrices and the scalars.
- **The Structure Theorem.** Products of algebras, and the observation that such a product has dimension $\sum n_i^2$. The regular representation, whose invariant subspaces are the left ideals. Then Wedderburn in the case the book can carry: over an algebraically closed field a nonzero semisimple algebra is a product $M_{n_1}(F) \times \cdots \times M_{n_k}(F)$, and k and the multiset of the n_i are determined, being the number of isomorphism classes of irreducible spaces and their dimensions. The proof splits the regular representation, keeps one representative of each class, computes the commutant of the image with Schur's lemma, and lets the

double commutant theorem close the gap; Burnside is **not** used. The corollary reads off $\dim A = \sum n_i^2$ and $\dim Z(A) = k$, which between them pin down the block sizes of a small algebra by arithmetic alone. The general theorem over a division ring and Wedderburn's little theorem get their own subsection, stated with credit and proved nowhere. Two decompositions are computed: $\mathbb{C}[C_4]$ with its four idempotents, and $M_2(\mathbb{C})$, already in Wedderburn form.

- **Group Representations.** A homomorphism $G \to \mathrm{GL}(V)$, with the trivial, sign, permutation, rotation, and regular representations, and a nonexample that fails only by not being well defined on a quotient. Subrepresentations, equivalence, irreducibility, direct sums. Then the bridge: representations of G over F correspond exactly to algebra homomorphisms $F[G] \to \mathrm{End}(V)$, carrying invariant subspaces to invariant subspaces and intertwining maps to $F[G]$-maps, so §§01 to 06 apply verbatim. Maschke's theorem: average a projection over the group and the average is a projection onto the same subspace that commutes with the action. The hypothesis $\mathrm{char}\, F \nmid |G|$ is spent on one division, and without it the unnormalized sum annihilates the subspace instead of fixing it; $\mathbb{Z}/p\mathbb{Z}$ on $\mathbb{F}_p^2$ is worked out, and a warning drops finiteness instead and breaks the theorem the other way. Complete reducibility follows by induction, so $F[G]$ is semisimple. The section closes with the two families §08 will compute characters for: all the irreducible complex representations of $\mathbb{Z}/n\mathbb{Z}$, and three for S_3 with every matrix written out.
- **Characters and Orthogonality.** Conjugacy classes are built here, since the book has not met them, and shown to partition the group. A character is one number per group element: constant on classes, additive over direct sums, equal to the dimension at the identity and largest there in absolute value, conjugated by inversion — and **not** a homomorphism. The class functions form an inner product space whose dimension is the number of classes. Two averaging lemmas — the average of a representation is the projection onto the fixed space, and the trace of $f \mapsto BfA$ is $(\mathrm{tr}\, A)(\mathrm{tr}\, B)$ — combine into a count of intertwining maps, and Schur's lemma turns that count into a 0 or a 1: the irreducible characters are orthonormal. The criterion $\langle \chi, \chi \rangle = 1$ then decides irreducibility without looking at a vector, and decomposes the permutation representation of S_3 in three lines. The deepest result is that the characters **span** the class functions, by a third averaging argument run on the regular representation; so the number of irreducible representations is the number of conjugacy classes. Out of it come the sum of the squares of the degrees — the identity §06 proved by another route entirely — and column orthogonality, from a one-sided inverse. The tables of S_3 and $\mathbb{Z}/4\mathbb{Z}$ are computed and every relation checked.
- **Finite Abelian Groups and the Fourier Transform.** The order of an element, the cyclic group, and the direct product of groups are built first, Chapter 0 having none of them. Over $\mathbb{C}$ an abelian group has only one-dimensional representations, so a character is a homomorphism $G \to \mathbb{C}^\times$ whose values are roots of unity, and the characters form a group $\widehat{G}$ under pointwise multiplication. It has the same order as G, always; it is isomorphic to G for cyclic groups and for products of them, and the section stops there and says exactly what it is not proving. Lining G up against $\widehat{G}$ gives a square character matrix, which is unitary, and inversion and Plancherel are what a unitary matrix always gives. Then the three identifications: for $\mathbb{Z}/n\mathbb{Z}$ the character matrix **is** the Fourier matrix; the group algebra of any finite abelian group splits into the functions on $\widehat{G}$, which for the cyclic group is the convolution theorem; and the circulants are the group algebra of $\mathbb{Z}/n\mathbb{Z}$ acting by multiplication, with the basis reversal that Chapter 12's indexing forces spelled out in a warning. The case $n = 4$ is checked against Chapter 12's numbers.
- **What the Chapter Buys.** A closing section that introduces nothing. Four moves are named and located; four earlier theorems are read again from here — simultaneous diagonalization, commuting normal families, circulants and the discrete Fourier transform, and the two commutant computations of Chapters 10 and 12. A table gives, for each main

theorem, the hypothesis on the field it needs and the counterexample that shows the hypothesis is not decoration, with three comments on what the table reveals. Then the list of what was quoted, the list of what was left outside and why, and a last look back at what a clean answer cost.

23.0.5 Named moves

- **Average over the group to build an invariant object.** Start with something that depends on a choice — a projection, an arbitrary linear map, a vector — and replace it by the average of its translates. The average is invariant because translating permutes the summands. This is the whole of Maschke's theorem, the whole of the orthogonality relations, and the whole of the proof that the characters span. It costs one division by $|G|$, and that division is the chapter's only hypothesis on the characteristic.
- **Use Schur's lemma to force a scalar, then take a trace to find which one.** Build an operator by some construction, check that it commutes with everything in sight, conclude it is λ id, and compute λ from the trace. Two lines that replace an unknown operator by a number, over and over, and the exact price is that F be algebraically closed.
- **Sandwich with matrix units.** The identity $\mathbf{E}_{ij}\mathbf{X}\mathbf{E}_{kl} = x_{jk}\mathbf{E}_{il}$ extracts one entry of an arbitrary matrix and puts it wherever you like. An ideal that contains anything nonzero therefore contains every matrix unit, which is the simplicity of $M_n(F)$ in three lines and over any field.
- **Let the algebra act on itself.** An abstract algebra has no vectors until you give it some, and the cheapest supply is the algebra. Left multiplication makes it an algebra of operators (which is why matrices are general), and the invariant subspaces of that action are the left ideals — so a question about ideals and a question about representations are one question.
- **Copy the space to upgrade one vector to a basis.** A density argument that matches an operator on a single vector becomes a proof that two operators are equal, once it is run on $V \oplus \cdots \oplus V$ with the basis of V packed into one vector of the copy. The only work is checking that the copied algebra inherits the hypotheses, and semisimplicity is the hypothesis that survives copying.
- **Count with $\sum n_i^2$.** The dimension identity of the structure theorem, read for a group algebra as $|G| = \sum_i (\dim V_i)^2$, pins down the degrees of all the irreducible representations of a small group from the number of conjugacy classes alone, before a single matrix is written. It does for this chapter what Rank–Nullity does for Chapter 2.
- **Read a family of operators as the algebra it generates.** A commuting family generates a commutative algebra; a group of invertible operators spans one; a single operator generates $F[T]$. Invariant subspaces, commutants, and irreducibility are unchanged by the passage, so a theorem about algebras is a theorem about every family the book has studied. That is what lets the last section re-prove four earlier results without a new idea in any of them.

23.1 Associative Algebras

Since Chapter 9 the book has looked at one operator at a time, or at a family whose members commute. But operators rarely arrive alone. The polynomials in a fixed operator, the circulants of Chapter 12 §09, the matrices commuting with a given matrix, the linear symmetries of a figure: each of these is a whole set of operators, closed under addition, under scalars, and under composition. This section makes such a system a single object of study. It fixes the vocabulary — subalgebra, homomorphism, generated algebra, commutant — and builds the two examples

the rest of the chapter runs on, the matrix algebra and the group algebra.

23.1.1 What an algebra is, and what we assume about it

The definition is already in the book. Chapter 14 §12 introduced an **algebra over** F (Definition 14.12.1) as a vector space A over F with a multiplication $A \times A \to A$ that is bilinear (A1), associative (A2), and has an identity 1_A (A3). It also fixed the notion of a homomorphism of F-algebras. We build on that definition; the homomorphism clause is restated below only so that later sections have a label to cite.

Three standing conventions, stated once and used everywhere in this chapter.

- **Associative and unital.** Every algebra satisfies (A1), (A2), and (A3). Lie algebras and the cross product, which fail (A2) and (A3), are outside the chapter.
- **Finite-dimensional.** Unless a statement says otherwise, $\dim_F A < \infty$. The one standing exception is $F[x]$, which we keep as a source of counterexamples and always name as infinite-dimensional when we use it that way.
- **Not commutative, not invertible.** Nothing requires $xy = yx$, and nothing requires a nonzero element to have an inverse. An algebra has an addition and a multiplication with all the usual laws except commutativity and division; it is not a group.

The examples are the ones Chapter 14 listed, and it is worth seeing why each satisfies the axioms. In $M_n(F)$, matrix multiplication is bilinear and associative with identity $\mathbf{I}_n$ (Theorem 0.9.4), so $M_n(F)$ is an F-algebra of dimension n^2. In $\mathcal{L}(V)$ for a vector space V over F, composition is bilinear, associative, and has the identity id_V (Theorem 2.4.2), so $\mathcal{L}(V)$ is an F-algebra; when $\dim V = n$ it has dimension n^2 (Theorem 4.1.4). The polynomial ring $F[x]$ is an algebra of infinite dimension. The quaternions $\mathbb{H}$ form a real algebra of dimension 4 (Definition 14.12.3). The field F itself is an algebra of dimension 1, the degenerate case, and it matters: it is the smallest algebra that is not the zero space, and every nonzero algebra contains a copy of it.

Remark. Many books write $\mathrm{End}(V)$, for the **endomorphisms** of V, where this book writes $\mathcal{L}(V)$, and $\mathrm{Hom}(V, W)$ where this book writes $\mathcal{L}(V, W)$. The names mean the same objects. This book writes $\mathcal{L}$ by default, and later sections of this chapter write $\mathrm{End}(V)$ when the algebra structure is the point.

23.1.2 Subalgebras

After an object come its subobjects. We already know several sets of matrices that are closed under the three operations: the diagonal matrices, the upper triangular matrices, the circulants (Corollary 12.9.5), the polynomials in a fixed matrix. In each case we checked closure by hand. The right move is to name the condition once.

A subalgebra is a subspace that is closed under multiplication and contains the identity.

Definition 23.1.1. Subalgebra

Let A be an algebra over F. A subset $B \subseteq A$ is a **subalgebra** of A if

(SA1) B is a **subspace** of A;

(SA2) $xy \in B$ **for all** $x, y \in B$;
(SA3) $1_A \in B$.

Clause by clause. (SA1) is the linear condition: B is closed under addition and under scalars, so it is a vector space in its own right. (SA2) says the multiplication of A restricts to a map $B \times B \to B$. (SA3) asks for **the** identity of A, not merely for some element of B that acts as an identity on B; the difference is real, and the warning below shows a set that satisfies (SA1) and (SA2), has its own identity, and is not a subalgebra.

Well-definedness. A subalgebra is an algebra. Bilinearity, associativity, and the identity law are equations among elements of B, and they hold there because they hold in A. So nothing needs to be re-checked once (SA1)–(SA3) are known — which is exactly the labor that naming the condition saves.

Throughout, $\mathbf{E}_{ij}$ denotes the **matrix unit** with a 1 in position (i, j) and zeros elsewhere; the n^2 of them are a basis of $M_n(F)$ (Example (Standard Bases)).

Example. Six subalgebras

Decide which of the following are subalgebras, and of which algebra.

(a) The scalars $F1_A = \{c1_A : c \in F\}$ inside any algebra A.
(b) The diagonal matrices and the upper triangular matrices inside $M_n(F)$. We write $\mathcal{T}_n$ for the second of these, since it returns below and in the next section.
(c) The circulants $\mathcal{C}_n$ inside $M_n(\mathbb{C})$ (Chapter 12 §09).
(d) The quaternions $\mathbb{H}$ inside $M_2(\mathbb{C})$, as an algebra over $\mathbb{R}$.
(e) The **strictly** upper triangular matrices inside $M_n(F)$.
(f) The matrices of trace 0 inside $M_2(F)$.

Solution.

(a) A subalgebra. It is the image of the linear map $c \mapsto c1_A$, hence a subspace; $(c1_A)(d1_A) = (cd)1_A$ by bilinearity; and $1_A = 1 \cdot 1_A$. Its dimension is 1 unless $A = \{0\}$: if $1_A = 0$, then $x = 1_A x = 0$ for every x. This is the degenerate example, and every subalgebra contains it.
(b) Both are subalgebras. Sums and scalar multiples of diagonal (respectively upper triangular) matrices are again such, the product of two upper triangular matrices is upper triangular because $(\mathbf{AB})_{ij} = \sum_k a_{ik} b_{kj}$ vanishes when $i > j$ — every term has $k < i$ or $k > j$ — and $\mathbf{I}_n$ is diagonal. The diagonal matrices have dimension n, with basis the $\mathbf{E}_{ii}$, and $\mathcal{T}_n$ has dimension $n(n+1)/2$, with basis the $\mathbf{E}_{ij}$ for $i \le j$.
(c) A subalgebra, of dimension n: this is exactly Corollary 12.9.5 (a), (b), together with the fact that $\mathbf{I}$ is the circulant whose first row is $(1, 0, \dots, 0)$.
(d) A subalgebra of $M_2(\mathbb{C})$ **over** $\mathbb{R}$, of real dimension 4, as Definition 14.12.3 records: the real span of $\mathbf{I}_2, i, j, k$ is closed under multiplication by the quaternion relations, and it contains $\mathbf{I}_2$. It is **not** a subalgebra over $\mathbb{C}$, since it is not closed under multiplication by i: $i\mathbf{I}_2$ has both diagonal entries equal to i, while every element of $\mathbb{H}$ has diagonal entries $a + bi$ and $a - bi$, forcing $a = 0$ and $b = 1$ and $b = -1$ at once.
(e) **Not** a subalgebra. It is a subspace and it is closed under products, by the computation in (b) with strict inequalities. But $\mathbf{I}_n$ is not strictly upper triangular for $n \ge 1$, so

(SA3) fails. This is the nonexample by minimal change: drop the diagonal from (b) and exactly one clause breaks.

(f) **Not** a subalgebra. $\mathbf{E}_{12}\mathbf{E}_{21} = \mathbf{E}_{11}$ has trace $1 \neq 0$, so (SA2) fails. Clause (SA3) fails too, except in characteristic 2, where $\operatorname{tr}\mathbf{I}_2 = 2 = 0$.

Warning. (SA3) is not a formality. In $M_2(F)$, the set $B = \{a\mathbf{E}_{11} : a \in F\}$ is a subspace, is closed under products, and even has an identity of its own, namely $\mathbf{E}_{11}$, since $\mathbf{E}_{11}(a\mathbf{E}_{11}) = a\mathbf{E}_{11}$. It is still not a subalgebra of $M_2(F)$, because $\mathbf{I}_2 \notin B$. Subspaces that are closed under multiplication but miss the identity are the shape of the *ideals* of the next section — though this particular one is not an ideal either, as Section 2 checks. They are a different kind of object, and no theorem about subalgebras applies to them.

Quick check. In $M_2(\mathbb{F}_2)$, is the set of matrices of trace 0 closed under multiplication? Does it contain the identity?

Solution. It contains $\mathbf{I}_2$, because $\operatorname{tr}\mathbf{I}_2 = 1 + 1 = 0$ in $\mathbb{F}_2$, so (SA3) holds here. But it is still not a subalgebra: $\mathbf{E}_{12}$ and $\mathbf{E}_{21}$ have trace 0 and $\mathbf{E}_{12}\mathbf{E}_{21} = \mathbf{E}_{11}$ has trace $1 \neq 0$, so (SA2) fails. One clause can hold over one field and fail over another; the other clause fails over every field.

23.1.3 Homomorphisms, isomorphisms, and the algebra generated by a set

Definition 14.12.1 already says what a homomorphism of F-algebras is: an F-linear map $\varphi\colon A \to B$ with $\varphi(xy) = \varphi(x)\varphi(y)$ and $\varphi(1_A) = 1_B$. We name it, so that later sections can cite it, and add the two subsets it produces.

Definition 23.1.2. Algebra Homomorphism, Kernel, and Image

Let A and B be algebras over F. An **algebra homomorphism** $\varphi\colon A \to B$ is an F-linear map with

$$\varphi(xy) = \varphi(x)\varphi(y) \quad \text{for all } x, y \in A, \qquad \varphi(1_A) = 1_B,$$

as in Definition 14.12.1. Its **kernel** is $\ker\varphi = \{x \in A : \varphi(x) = 0\}$ and its **image** is $\operatorname{im}\varphi = \{\varphi(x) : x \in A\}$, the kernel and image of φ as a linear map.

The unital clause $\varphi(1_A) = 1_B$ does not follow from the other two. The zero map $M_2(F) \to M_2(F)$ is linear and multiplicative and sends $\mathbf{I}_2$ to $\mathbf{O} \neq \mathbf{I}_2$, so it is **not** an algebra homomorphism. We will never call it one.

Definition 23.1.3. Isomorphic Algebras

An **algebra isomorphism** is a bijective algebra homomorphism. Algebras A and B are **isomorphic**, written $A \cong B$, if there is an algebra isomorphism $A \to B$.

The inverse of an algebra isomorphism φ is again one: it is linear (Theorem 2.5.1), it satisfies $\varphi^{-1}(uv) = \varphi^{-1}(u)\varphi^{-1}(v)$ — apply the injective map φ to both sides — and $\varphi^{-1}(1_B) = 1_A$. So $\cong$ is symmetric, and it is reflexive and transitive for the same reasons as in Theorem 2.5.3.

Two homomorphisms to keep in mind. First, taking matrices: for a finite-dimensional V with ordered basis $\mathcal{B}$, the map $T \mapsto [T]_{\mathcal{B}}^{\mathcal{B}}$ is an algebra isomorphism $\mathcal{L}(V) \to M_n(F)$, being a linear bijection (Theorem 4.1.4) that turns composition into matrix multiplication (Theorem 4.1.2) and sends id_V to $\mathbf{I}_n$. Second, a nonexample: transposition $\mathbf{A} \mapsto \mathbf{A}^\top$ on $M_n(F)$ is a linear bijection fixing $\mathbf{I}_n$, but $(\mathbf{AB})^\top = \mathbf{B}^\top\mathbf{A}^\top$, so it reverses products. For $n \geq 2$ it is not a homomorphism, since $(\mathbf{E}_{12}\mathbf{E}_{21})^\top = \mathbf{E}_{11}$ while $\mathbf{E}_{12}^\top\mathbf{E}_{21}^\top = \mathbf{E}_{21}\mathbf{E}_{12} = \mathbf{E}_{22}$.

Now the construction we need constantly: the smallest subalgebra containing a given set of operators.

Definition 23.1.4. The Subalgebra Generated by a Set

Let A be an algebra over F and $S \subseteq A$. The **subalgebra generated by** S, written $F\langle S\rangle$, is the intersection of all subalgebras of A containing S.

The letter F in $F\langle S\rangle$ records that scalars are allowed; the next section writes $\langle S\rangle$, without the F, for the **ideal** generated by S, which is a different and usually larger set. The intersection of any collection of subalgebras is a subalgebra — each clause of Definition 23.1.1 is preserved by intersections — and A itself is one of them, so $F\langle S\rangle$ is a subalgebra containing S, and it is contained in every other one. That is what "generated" means. It also has a concrete description.

Proposition 23.1.1. What the Generated Subalgebra Contains

Let A be an algebra over F and $S \subseteq A$. Then $F\langle S\rangle$ is the span of 1_A together with all finite products $s_1 s_2 \cdots s_k$ with $k \geq 1$ and $s_1, \ldots, s_k \in S$.

Proof. Let P be that span. Every subalgebra containing S contains each product $s_1 \cdots s_k$ by (SA2) and induction on k, contains 1_A by (SA3), and is a subspace by (SA1); hence it contains P. In particular $P \subseteq F\langle S\rangle$.

Conversely P is itself a subalgebra: it is a span, hence a subspace; it contains 1_A; and a product of two spanning elements is either 1_A, or a product $s_1 \cdots s_k$, or a concatenation $(s_1 \cdots s_k)(t_1 \cdots t_l)$, which is again such a product by associativity. By bilinearity, a product of two elements of P is a combination of such products, so it lies in P. Since $S \subseteq P$, minimality gives $F\langle S\rangle \subseteq P$. This proves the proposition. ∎

For a single operator the products are just the powers, and the minimal polynomial says how many of them are independent.

Proposition 23.1.2. The Algebra Generated by One Operator

Let $V \neq \{0\}$ be finite-dimensional over F and let $T \in \mathcal{L}(V)$ have minimal polynomial m_T, of degree d. Then

$$F[T] := \{p(T) : p \in F[x]\}$$

is a **commutative** subalgebra of $\mathcal{L}(V)$, it equals $F\langle\{T\}\rangle$, and $(\mathrm{id}_V, T, T^2, \dots, T^{d-1})$ is a basis of it. In particular $\dim F[T] = \deg m_T$.

Idea. Division with remainder cuts every power of T down to degree less than d, which gives spanning; and a shorter relation among the powers would be an annihilating polynomial of degree less than d, which is what minimality forbids.

Proof. By Theorem 6.7.1 the map $p \mapsto p(T)$ is linear and multiplicative, so $F[T]$ is a subspace closed under products, and it contains $\mathrm{id}_V = 1(T)$; it is therefore a subalgebra. Any two of its elements commute, since $p(T)q(T) = (pq)(T) = (qp)(T) = q(T)p(T)$. By Proposition 23.1.1, $F\langle\{T\}\rangle$ is the span of id_V and the powers T^k with $k \geq 1$, which is $F[T]$.

Since V is finite-dimensional, m_T exists (Theorem 6.7.3 (b), (c)); and $d \geq 1$, because the only monic polynomial of degree 0 is 1, whose value at T is $\mathrm{id}_V \neq 0$ since $V \neq \{0\}$.

Spanning. Let $p \in F[x]$. Dividing by m_T (Theorem 6.1.1) gives $p = qm_T + r$ with $\deg r < d$, and $p(T) = q(T)m_T(T) + r(T) = r(T)$, a combination of $\mathrm{id}_V, T, \dots, T^{d-1}$.

Independence. Suppose $c_0 \mathrm{id}_V + c_1 T + \cdots + c_{d-1}T^{d-1} = 0$ with the c_i not all zero. Then $r = c_0 + c_1 x + \cdots + c_{d-1}x^{d-1}$ is a nonzero polynomial with $r(T) = 0$ and $\deg r < d$, contradicting the fact that m_T has least degree among the nonzero annihilating polynomials (Theorem 6.7.3 (b)).

So the d listed operators are a basis, and $\dim F[T] = d$. This proves the proposition.

■

Example. Two polynomial algebras

Compute $\dim F[\mathbf{A}]$ for $\mathbf{A} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$ and for $\mathbf{B} = \mathrm{diag}(1, 2, 3)$ over $\mathbb{Q}$.

Solution. For $\mathbf{A}$: $\mathbf{A}^2 = \mathbf{O}$ and $\mathbf{A} \neq \mathbf{O}$, so x^2 annihilates $\mathbf{A}$ and no polynomial of degree 1 does (if $a\mathbf{I} + b\mathbf{A} = \mathbf{O}$, reading the $(1,1)$ and $(1,2)$ entries gives $a = b = 0$). Hence $m_{\mathbf{A}} = x^2$ and $\dim \mathbb{Q}[\mathbf{A}] = 2$, with basis $(\mathbf{I}, \mathbf{A})$. Note that $\dim \mathbb{Q}[\mathbf{A}] = 2$ is much smaller than $\dim M_3(\mathbb{Q}) = 9$.

For $\mathbf{B}$: $(x-1)(x-2)(x-3)$ annihilates $\mathbf{B}$, since the product of the three diagonal matrices $\mathbf{B} - k\mathbf{I}$ is $\mathbf{O}$. No polynomial r of degree at most 2 does, unless $r = 0$: $r(\mathbf{B}) = \mathrm{diag}(r(1), r(2), r(3)) = \mathbf{O}$ would give r three distinct roots. So $m_{\mathbf{B}}$ has degree 3 and $\dim \mathbb{Q}[\mathbf{B}] = 3$, with basis $(\mathbf{I}, \mathbf{B}, \mathbf{B}^2)$.

23.1.4 The group algebra

Here is the object that turns a group of symmetries into linear algebra. Suppose a finite group G acts on a space by invertible operators $\rho(g)$. Then every F-combination $\sum_g a_g \rho(g)$ is again an operator, and the products of such combinations are governed entirely by the multiplication table of G. It is worth building, once and for all, the algebra in which those formal combinations live.

The group algebra is the set of formal F-combinations of group elements, multiplied by the group law.

Definition 23.1.5. Group Algebra

Let G be a **finite** group with identity e, and let F be a field. The **group algebra** $F[G]$ is the free vector space $F^{(G)}$ on the set G (Definition 15.2.3), with basis $\{\delta_g : g \in G\}$, equipped with the unique bilinear multiplication satisfying

$$\delta_g \, \delta_h = \delta_{gh} \qquad \text{for all } g, h \in G.$$

We write g for δ_g and 1 for δ_e, so that a general element is a formal sum $\sum_{g \in G} a_g \, g$ with all $a_g \in F$, two such sums are equal exactly when all their coefficients agree, and

$$\Big(\sum_g a_g \, g\Big)\Big(\sum_h b_h \, h\Big) = \sum_{k \in G} \Big(\sum_{g \in G} a_g \, b_{g^{-1}k}\Big) k.$$

Well-definedness. Choose an ordering $g_1, \dots, g_N$ of G, so that $(\delta_{g_1}, \dots, \delta_{g_N})$ is an ordered basis of $F^{(G)}$ (Lemma 15.2.2 (a)). By Theorem 15.1.1 with $k = 2$, there is **exactly one** bilinear map $F[G] \times F[G] \to F[G]$ taking the prescribed values $\delta_{g_i g_j}$ on basis pairs. The displayed formula for the product is that bilinear map written out: expanding $\sum_{g,h} a_g b_h \, gh$ and collecting the terms with $gh = k$, which means $h = g^{-1}k$, gives the stated coefficient of k.

The axioms hold. Bilinearity is (A1) by construction. For associativity (A2), both $(xy)z$ and $x(yz)$ are trilinear in (x, y, z), so by Theorem 15.1.1 it is enough to compare them on basis triples, where $(\delta_g \delta_h)\delta_k = \delta_{(gh)k} = \delta_{g(hk)} = \delta_g(\delta_h \delta_k)$ by the associativity axiom (G1) of Definition 0.10.1. For (A3), $\delta_e \delta_g = \delta_{eg} = \delta_g = \delta_g \delta_e$ on basis elements, hence everywhere by bilinearity. So $F[G]$ is an F-algebra, of dimension $|G|$.

In words: an element of $F[G]$ is a bookkeeping device, a list of $|G|$ scalars indexed by the group, and multiplying two of them convolves the lists along the group law. The group sits inside $F[G]$: the map $g \mapsto \delta_g$ is injective (distinct basis vectors) and multiplicative, and each δ_g is invertible with inverse $\delta_{g^{-1}}$.

Example. Three group algebras

(a) $G = \{e\}$.
(b) $G = \mathbb{Z}/n\mathbb{Z}$, written multiplicatively as $\{1, g, \dots, g^{n-1}\}$ with $g^n = 1$.
(c) $G = S_3$.

In each case give the dimension and decide whether $F[G]$ is commutative.

Solution.

(a) $F[\{e\}] = F\delta_e \cong F$, of dimension 1, and commutative. This is the degenerate case.

(b) $\dim F[G] = n$, and $F[G]$ is commutative because $g^i g^j = g^{i+j} = g^j g^i$ on basis elements and multiplication is bilinear. Moreover the map $\varepsilon\colon F[x] \to F[G]$ sending $p = \sum_i c_i x^i$ to $\sum_i c_i g^i$ is an algebra homomorphism: it is linear by construction, it sends 1 to $g^0 = 1$, and for $p = \sum_i c_i x^i$ and $q = \sum_j d_j x^j$,

$$\varepsilon(pq) = \sum_k \Big(\sum_{i+j=k} c_i d_j \Big) g^k = \Big(\sum_i c_i g^i \Big) \Big(\sum_j d_j g^j \Big) = \varepsilon(p)\varepsilon(q),$$

where the middle equality is bilinearity together with $g^i g^j = g^{i+j}$. It is surjective because the powers $1, g, \dots, g^{n-1}$ are a basis. Its kernel is exactly $\langle x^n - 1 \rangle$: dividing p by $x^n - 1$ (Theorem 6.1.1) gives $p = q(x^n - 1) + r$ with $\deg r < n$, and since $\varepsilon(x^n - 1) = g^n - 1 = 0$ we get $\varepsilon(p) = \varepsilon(r)$, which vanishes only when all coefficients of r vanish, the g^i being a basis. The next section turns this into an isomorphism $F[\mathbb{Z}/n\mathbb{Z}] \cong F[x]/\langle x^n - 1 \rangle$.

(c) $\dim F[S_3] = 6$, and $F[S_3]$ is **not** commutative: with $\sigma = (1\ 2)$ and $\tau = (1\ 2\ 3)$ we have $\sigma\tau \neq \tau\sigma$ in S_3, so $\delta_\sigma \delta_\tau \neq \delta_\tau \delta_\sigma$ are distinct basis vectors. In general $F[G]$ is commutative if and only if G is abelian: the displayed argument gives one direction, and if $F[G]$ is commutative then $\delta_{gh} = \delta_g \delta_h = \delta_h \delta_g = \delta_{hg}$, and distinct group elements give distinct basis vectors, so $gh = hg$.

Warning. The elements of $F[G]$ are formal sums, not group elements, and most of them are **not** invertible. In $\mathbb{Q}[\mathbb{Z}/2\mathbb{Z}]$ with $\mathbb{Z}/2\mathbb{Z} = \{1, g\}$, the element $1 + g$ satisfies $(1+g)^2 = 1 + 2g + g^2 = 2(1+g)$, so $\frac{1}{2}(1+g)$ is a nonzero element equal to its own square and different from 1. Multiplying it by $1 - g$ gives 0, so neither factor has an inverse. Passing from G to $F[G]$ buys linearity at the price of invertibility.

23.1.5 Every finite-dimensional algebra is an algebra of matrices

Matrices look like a special case of an algebra. They are not: every finite-dimensional algebra is one, and the embedding is the most economical thing imaginable — let the algebra act on itself.

Theorem 23.1.3. Left Multiplication Embeds an Algebra in Its Own Operators

Let A be an algebra over F. For $a \in A$ define $L_a\colon A \to A$ by $L_a(x) = ax$. Then

$$L\colon A \to \mathcal{L}(A), \qquad a \mapsto L_a,$$

is an **injective** algebra homomorphism. If $\dim_F A = n < \infty$, then A is isomorphic to a subalgebra of $M_n(F)$.

Idea. Each axiom of an algebra says one thing about L: bilinearity in the second slot makes L_a linear, bilinearity in the first makes $a \mapsto L_a$ linear, associativity makes L multiplicative,

and the identity makes L unital and — this is the point — injective, because an operator that kills everything in particular kills 1_A, and $L_a(1_A) = a$.

Proof. Each L_a is linear, because multiplication is linear in its second argument (A1). The map L is linear, because multiplication is linear in its first argument: $L_{a+ca'}(x) = (a+ca')x = ax + c(a'x)$ for all x.

L is multiplicative: for $a, b, x \in A$, associativity (A2) gives

$$L_{ab}(x) = (ab)x = a(bx) = L_a(L_b(x)) = (L_aL_b)(x),$$

so $L_{ab} = L_aL_b$. It is unital: $L_{1_A}(x) = 1_Ax = x$, so $L_{1_A} = \mathrm{id}_A$.

L is injective: if $L_a = 0$ then, evaluating at 1_A, $a = a1_A = L_a(1_A) = 0$. So $\ker L = \{0\}$ and L is injective (Theorem 2.2.3).

Suppose now $\dim_F A = n < \infty$. Then $\operatorname{im} L$ is a subalgebra of $\mathcal{L}(A)$: it is a subspace, it is closed under products since $L_aL_b = L_{ab}$, and it contains $\mathrm{id}_A = L_{1_A}$. The map $L\colon A \to \operatorname{im} L$ is a bijective algebra homomorphism, hence an algebra isomorphism. Fixing an ordered basis of A and taking matrices is an algebra isomorphism $\mathcal{L}(A) \to M_n(F)$, as recorded after Definition 23.1.3, and it carries $\operatorname{im} L$ to a subalgebra of $M_n(F)$ isomorphic to A. This proves the theorem.

■

The theorem is the reason the rest of the chapter may think in matrices without loss. It is also a genuine computation: to see an abstract algebra concretely, write down the matrices of left multiplication in a basis.

Example. Left multiplication for the upper triangular algebra

Let $A = \mathcal{T}_2$, the upper triangular matrices in $M_2(F)$, with ordered basis $\mathcal{B} = (\mathbf{E}_{11}, \mathbf{E}_{12}, \mathbf{E}_{22})$. Compute the matrix of $L_{\mathbf{U}}$ for $\mathbf{U} = \begin{pmatrix} a & b \\ 0 & c \end{pmatrix}$, and read off what the embedding of Theorem 23.1.3 does.

Solution. Multiply each basis element on the left by $\mathbf{U}$:

$$\mathbf{U}\mathbf{E}_{11} = a\mathbf{E}_{11}, \qquad \mathbf{U}\mathbf{E}_{12} = a\mathbf{E}_{12}, \qquad \mathbf{U}\mathbf{E}_{22} = b\mathbf{E}_{12} + c\mathbf{E}_{22}.$$

Putting the coordinate vectors in as columns,

$$[L_{\mathbf{U}}]_{\mathcal{B}}^{\mathcal{B}} = \begin{pmatrix} a & 0 & 0 \\ 0 & a & b \\ 0 & 0 & c \end{pmatrix}.$$

So $\mathcal{T}_2$, an algebra of 2×2 matrices of dimension 3, is isomorphic to the algebra of 3×3 matrices of this shape. The embedding is faithful but wasteful: $\mathbf{U} \mapsto \mathbf{U}$ already embeds $\mathcal{T}_2$ in $M_2(F)$. Left multiplication is the embedding that always exists, not the smallest one.

23.1.6 Commutants

The last construction of the section is the one that will carry Schur's lemma later. Given a set of operators, ask which operators commute with all of them. Chapter 9 §09 asked this about a pair, Chapter 10 §04 about a Weyr matrix, Chapter 12 §07 about a normal family. It is the same question each time.

The commutant of a set is everything that commutes with all of it.

Definition 23.1.6. Commutant and Center

Let A be an algebra over F and let $S \subseteq A$ be **any subset**. The **commutant** of S in A is

$$S' := \{a \in A : as = sa \text{ for every } s \in S\}.$$

The **center** of A is $Z(A) := A'$, the set of elements commuting with **every** element of A. We write S'' for $(S')'$.

In words: to belong to S', an element must commute with each member of S — one failure disqualifies it. The members of S need not commute with each other, and S need not be a subspace or a subalgebra; the definition asks nothing of S.

Proposition 23.1.4. Properties of the Commutant

Let A be an algebra over F and $S, T \subseteq A$.

(a) S' is a subalgebra of A; in particular $Z(A)$ is a commutative subalgebra.
(b) If $S \subseteq T$, then $T' \subseteq S'$.
(c) $S \subseteq S''$.
(d) $S' = S'''$.

Proof.

(a) Fix $s \in S$. The set $\{a : as = sa\}$ is the kernel of the linear map $a \mapsto as - sa$, hence a subspace (Theorem 2.2.1); S' is the intersection of these subspaces over $s \in S$, hence a subspace. If $a, b \in S'$ then $(ab)s = a(bs) = a(sb) = (as)b = (sa)b = s(ab)$, using associativity at each step, so $ab \in S'$. And $1_A s = s = s1_A$, so $1_A \in S'$. Thus S' is a subalgebra. For the center, every element of $Z(A)$ commutes with every element of A, so in particular any two of them commute.

(b) If a commutes with every element of T and $S \subseteq T$, then a commutes with every element of S.

(c) Let $s \in S$. Every $a \in S'$ satisfies $as = sa$, which is the condition for s to lie in $(S')'$.

(d) Applying (c) to S' gives $S' \subseteq S'''$. Applying (b) to the inclusion $S \subseteq S''$ of (c) gives $(S'')' \subseteq S'$, that is, $S''' \subseteq S'$.

■

Three commutants, computed. Each is a short exercise in comparing entries, and the pattern they show is the one Schur's lemma will explain.

Example. Commutants of three matrices

Compute the commutant S' inside the stated algebra for

(a) $S = \{\mathbf{D}\}$ inside $M_3(F)$, where $\mathbf{D} = \operatorname{diag}(d_1, d_2, d_3)$ has d_1, d_2, d_3 **distinct**;
(b) $S = \{\mathbf{N}\}$ inside $M_3(F)$, where $\mathbf{N} = \mathbf{J}_3(0)$ is the nilpotent Jordan block;
(c) $S = M_3(\mathbb{Q})$ inside $M_3(\mathbb{Q})$.

Solution.

(a) *Distinct diagonal.* Comparing the (i,j) entries, $(\mathbf{DX})_{ij} = d_i x_{ij}$ and $(\mathbf{XD})_{ij} = x_{ij} d_j$, so $\mathbf{DX} = \mathbf{XD}$ says $(d_i - d_j)x_{ij} = 0$ for all i, j. Since the d_i are distinct, $x_{ij} = 0$ whenever $i \neq j$. So $\{\mathbf{D}\}'$ is the algebra of diagonal matrices in $M_3(F)$, of dimension 3. Here $m_{\mathbf{D}}$ has degree 3, since $(x - d_1)(x - d_2)(x - d_3)$ annihilates $\mathbf{D}$ while no nonzero polynomial of degree at most 2 does, such a polynomial having the three distinct roots d_1, d_2, d_3; so by Proposition 23.1.2 $\dim F[\mathbf{D}] = 3$ as well. Every polynomial in $\mathbf{D}$ commutes with $\mathbf{D}$, so $F[\mathbf{D}] \subseteq \{\mathbf{D}\}'$, and equal dimensions give $\{\mathbf{D}\}' = F[\mathbf{D}]$.

(b) *Jordan block.* Write $\mathbf{N} = \mathbf{E}_{12} + \mathbf{E}_{23}$. Then $(\mathbf{NX})_{ij} = x_{i+1,j}$ (read as 0 when $i + 1 > 3$) and $(\mathbf{XN})_{ij} = x_{i,j-1}$ (read as 0 when $j - 1 < 1$). Setting these equal for all i, j: the entries with $j = 1$ give $x_{21} = x_{31} = 0$, the entries with $i = 3$ give $x_{31} = x_{32} = 0$, and the four remaining equations give $x_{22} = x_{11}$, $x_{23} = x_{12}$, $x_{32} = x_{21}$, and $x_{33} = x_{22}$. So

$$\{\mathbf{N}\}' = \left\{ \begin{pmatrix} a & b & c \\ 0 & a & b \\ 0 & 0 & a \end{pmatrix} : a, b, c \in F \right\} = F[\mathbf{N}],$$

of dimension 3, since $\mathbf{I}, \mathbf{N}, \mathbf{N}^2$ are exactly the three matrices of this shape with one of a, b, c equal to 1 and the others 0. Again $\deg m_{\mathbf{N}} = 3$, so this is $F[\mathbf{N}]$ by Proposition 23.1.2.

(c) *The whole algebra.* Let $\mathbf{X} \in M_3(\mathbb{Q})'$. Commuting with $\mathbf{D} = \operatorname{diag}(1, 2, 3)$ forces $\mathbf{X}$ diagonal by the first computation, say $\mathbf{X} = \operatorname{diag}(x_1, x_2, x_3)$; and then $\mathbf{XE}_{12} = x_1 \mathbf{E}_{12}$ while $\mathbf{E}_{12}\mathbf{X} = x_2 \mathbf{E}_{12}$, so $x_1 = x_2$, and similarly $x_2 = x_3$. Hence $\mathbf{X}$ is a scalar matrix. Conversely scalar matrices commute with everything. So $Z(M_3(\mathbb{Q})) = \mathbb{Q}\mathbf{I}_3$, of dimension 1. The argument used that $1, 2, 3$ are distinct scalars; over a field of characteristic 2 or 3 they are not, which is why the general statement below is proved with matrix units instead.

The last computation deserves a clean statement over every field, because later sections use it.

Proposition 23.1.5. The Center of the Matrix Algebra Is the Scalars

Let F be **any** field and $n \geq 1$. Then $Z(M_n(F)) = F\mathbf{I}_n$. Equivalently, for a finite-dimensional $V \neq \{0\}$, the operators commuting with every operator on V are exactly the scalar multiples of id_V.

Proof. ($\supseteq$) A scalar matrix $c\mathbf{I}_n$ satisfies $(c\mathbf{I}_n)\mathbf{X} = c\mathbf{X} = \mathbf{X}(c\mathbf{I}_n)$ by Theorem 0.9.4 (4).

($\subseteq$) For $n = 1$ the claim reads $F = F\mathbf{I}_1$, which is true, so assume $n \geq 2$. Let $\mathbf{X} = (x_{pq}) \in$

$Z(M_n(F))$. Writing $(\mathbf{E}_{ij})_{pq} = \delta_{pi}\delta_{qj}$ and expanding both products with Theorem 0.9.2 (1),

$$(\mathbf{E}_{ij}\mathbf{X})_{pq} = \delta_{pi}\, x_{jq}, \qquad (\mathbf{X}\mathbf{E}_{ij})_{pq} = x_{pi}\, \delta_{qj}.$$

Now fix j and choose any $i \neq j$, which is possible because $n \geq 2$. Reading $\mathbf{E}_{ij}\mathbf{X} = \mathbf{X}\mathbf{E}_{ij}$ at $(p, q) = (i, q)$ with $q \neq j$ gives $x_{jq} = x_{ii}\delta_{qj} = 0$, so every off-diagonal entry in row j of $\mathbf{X}$ vanishes; and reading it at $(p, q) = (i, j)$ gives $x_{jj} = x_{ii}$. Letting j run over all indices makes $\mathbf{X}$ diagonal, and applying the second conclusion to every pair $i \neq j$ makes all its diagonal entries equal. So $\mathbf{X} = x_{11}\mathbf{I}_n$.

The statement for operators follows by choosing a basis and applying the algebra isomorphism $\mathcal{L}(V) \to M_n(F)$ recorded after Definition 23.1.3, which carries id_V to $\mathbf{I}_n$, and under which an element of the center corresponds to an element of the center. This proves the proposition.

■

Quick check. In $M_2(F)$, what is $\{\mathbf{I}_2\}'$, and what is $\{\mathbf{I}_2\}''$? Does $S'' = S$ hold for $S = \{\mathbf{I}_2\}$?

Solution. Everything commutes with $\mathbf{I}_2$, so $\{\mathbf{I}_2\}' = M_2(F)$. Then $\{\mathbf{I}_2\}'' = Z(M_2(F)) = F\mathbf{I}_2$ by Proposition 23.1.5. So $S \subsetneq S''$ here: S'' is the one-dimensional space of scalar matrices, while S is a single matrix. The inclusion of Proposition 23.1.4 (c) is genuinely an inclusion, and S'' should be read as "the smallest thing S cannot be distinguished from by commutation".

Remark. Both commutants in Example (Commutants of three matrices) (a), (b) came out equal to the algebra generated by the matrix. That is not automatic. For $\mathbf{A} = \mathbf{I}_n$ with $n \geq 2$, the algebra $F[\mathbf{A}]$ is the one-dimensional space of scalar matrices, while $\{\mathbf{A}\}' = M_n(F)$ has dimension n^2. The exact condition is that the minimal and characteristic polynomials of $\mathbf{A}$ agree, which is Chapter 10 §05's condition for a cyclic vector to exist; **that equivalence is quoted here and not proved**, and nothing in this chapter depends on it.

Warning. An algebra is not a group, and the theory of this chapter never assumes an inverse. In $F[x]$ the only elements with a multiplicative inverse are the nonzero constants, since $\deg(pq) = \deg p + \deg q$; in $M_n(F)$ the invertible elements are the nonsingular matrices, a proper subset; in $F[G]$ the warning above produced two nonzero elements with product 0. Every argument below that wants to divide must first produce an inverse, and most of them cannot.

23.1.7 Exercises

23.1.7.1 A. Check your understanding

Exercise. A1

(a) State the three clauses in the definition of a subalgebra, and say which one the strictly upper triangular matrices fail.

(b) Give the definition of the commutant S' of a subset S of an algebra A, and say why S itself need not be a subalgebra.

(c) True or false, with a reason: a linear bijection $\varphi\colon A \to B$ with $\varphi(xy) = \varphi(x)\varphi(y)$ for all x, y is an algebra isomorphism.

(d) What is $\dim F[G]$, and what is $\dim F[T]$ for an operator T on a finite-dimensional space?

Solution.

(a) (SA1) a subspace, (SA2) closed under products, (SA3) contains 1_A. The strictly upper triangular matrices fail (SA3), since $\mathbf{I}_n$ has nonzero diagonal.

(b) $S' = \{a \in A : as = sa \text{ for every } s \in S\}$ (Definition 23.1.6). The definition quantifies over the elements of S one at a time and asks nothing of S as a set, so S may be a single element, a finite list, or any subset at all.

(c) True — but note that $\varphi(1_A) = 1_B$ is a separate requirement in general, which here follows from bijectivity: for every $u \in B$, writing $u = \varphi(x)$, we get $\varphi(1_A)u = \varphi(1_A x) = u$ and likewise on the other side, so $\varphi(1_A)$ is an identity of B and equals 1_B by uniqueness of identities. The statement is true for bijective φ; it is false without bijectivity, as the zero map shows.

(d) $\dim F[G] = |G|$ (Definition 23.1.5) and $\dim F[T] = \deg m_T$ (Proposition 23.1.2).

23.1.7.2 B. Practice

Exercise. B1: Determine which are subalgebras

Determine which of the following are subalgebras of the stated algebra. Justify your answer, naming the exact clause that fails when one does.

(a) $\{\mathbf{A} \in M_2(F) : a_{21} = 0\}$ inside $M_2(F)$.

(b) $\{\mathbf{A} \in M_2(F) : a_{12} = a_{21}\}$ inside $M_2(F)$.

(c) $\{p \in F[x] : p(0) = p(1)\}$ inside $F[x]$.

(d) $\{c\,\mathrm{id}_V + N : c \in F,\ N^2 = 0\}$ inside $\mathcal{L}(V)$, for $\dim V = 3$.

Solution.

(a) A subalgebra: this is $\mathcal{T}_2$, handled in Example (Six subalgebras) (b).

(b) **Not** a subalgebra. The set is the symmetric matrices, a subspace containing $\mathbf{I}_2$, so (SA1) and (SA3) hold. But (SA2) fails: $\left(\begin{smallmatrix}0&1\\1&0\end{smallmatrix}\right)\left(\begin{smallmatrix}1&0\\0&0\end{smallmatrix}\right) = \left(\begin{smallmatrix}0&0\\1&0\end{smallmatrix}\right)$, which is not symmetric.

(c) A subalgebra. It is the kernel of the linear functional $p \mapsto p(0) - p(1)$, hence a subspace; it contains 1; and if $p(0) = p(1)$ and $q(0) = q(1)$ then $(pq)(0) = p(0)q(0) = p(1)q(1) = (pq)(1)$ by Theorem 0.8.5.

(d) **Not** a subalgebra, because it is not even a subspace: (SA1) fails. Take $N_1 = \mathbf{E}_{12}$ and $N_2 = \mathbf{E}_{23}$, both squaring to 0. Their sum $N = \mathbf{E}_{12} + \mathbf{E}_{23}$ has $N^2 = \mathbf{E}_{13} \neq 0$ and $N^3 = 0$, so N is not of the listed form: if $N = c\,\mathrm{id}_V + M$ with $M^2 = 0$, then $M = N - c\,\mathrm{id}_V$ and $0 = M^2 = N^2 - 2cN + c^2\,\mathrm{id}_V$, whose diagonal entries in the

standard basis are all c^2; so $c = 0$, and then $M = N$ with $N^2 \neq 0$.

Exercise. B2: A commutant

Let $\mathbf{D} = \operatorname{diag}(1, 1, 2) \in M_3(\mathbb{Q})$. Compute $\{\mathbf{D}\}'$ and its dimension, and compare it with $\dim \mathbb{Q}[\mathbf{D}]$. Hence show that $\{\mathbf{D}\}' \neq \mathbb{Q}[\mathbf{D}]$.

Solution. As in Example (Commutants of three matrices), $\mathbf{DX} = \mathbf{XD}$ says $(d_i - d_j)x_{ij} = 0$ with $(d_1, d_2, d_3) = (1, 1, 2)$. The coefficient vanishes exactly for $(i, j) = (1, 1)$, $(1, 2)$, $(2, 1)$, $(2, 2)$, and $(3, 3)$, so

$$\{\mathbf{D}\}' = \left\{ \begin{pmatrix} p & q & 0 \\ r & s & 0 \\ 0 & 0 & t \end{pmatrix} \right\},$$

of dimension 5.

For $\mathbb{Q}[\mathbf{D}]$: the polynomial $(x - 1)(x - 2)$ annihilates $\mathbf{D}$, and no polynomial of degree 1 does, since $a\mathbf{I} + b\mathbf{D} = \mathbf{O}$ would give $a + b = 0$ and $a + 2b = 0$, hence $a = b = 0$. So $m_{\mathbf{D}} = (x - 1)(x - 2)$ and $\dim \mathbb{Q}[\mathbf{D}] = 2$ by Proposition 23.1.2. Since $5 \neq 2$, the two algebras differ; concretely $\mathbf{E}_{12} \in \{\mathbf{D}\}'$ is not a polynomial in $\mathbf{D}$, because every such polynomial is diagonal.

Exercise. B3: Cayley for the quaternions

Let $\mathbb{H}$ be the quaternions with ordered basis $\mathcal{B} = (1, i, j, k)$ over $\mathbb{R}$, where 1 denotes $\mathbf{I}_2$. Compute the matrices of L_i and L_j in $\mathcal{B}$, and verify that they satisfy $L_i^2 = -\mathbf{I}_4$ and $L_i L_j = L_k$.

Solution. By Definition 14.12.3, $i \cdot 1 = i$, $ii = -1$, $ij = k$, and $ik = -j$. Writing the coordinate vectors of these four images as columns,

$$[L_i]_{\mathcal{B}}^{\mathcal{B}} = \begin{pmatrix} 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

Likewise $j \cdot 1 = j$, $ji = -k$, $jj = -1$, $jk = i$, so

$$[L_j]_{\mathcal{B}}^{\mathcal{B}} = \begin{pmatrix} 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{pmatrix}.$$

Squaring the first gives $-\mathbf{I}_4$, and multiplying the two in this order gives the matrix with columns $(0, 0, 0, 1), (0, 0, 1, 0), (0, -1, 0, 0), (-1, 0, 0, 0)$, which is the matrix of L_k computed the same way from $k \cdot 1 = k$, $ki = j$, $kj = -i$, $kk = -1$. Both checks also follow from Theorem 23.1.3 without computing: L is multiplicative, so $L_i L_i = L_{ii} = L_{-1} = -\mathbf{I}_4$ and $L_i L_j = L_{ij} = L_k$. This exhibits $\mathbb{H}$ as a subalgebra of $M_4(\mathbb{R})$.

23.1.7.3 C. Going deeper

Exercise. C1: The center of a group algebra

Let G be a finite group and F a field. For $g \in G$, the **conjugacy class** of g is $K_g = \{hgh^{-1} : h \in G\}$, and the **class sum** is $z_g = \sum_{k \in K_g} k \in F[G]$.

(a) Prove that $x = \sum_g a_g\, g$ lies in $Z(F[G])$ if and only if $a_{hgh^{-1}} = a_g$ for all $g, h \in G$.

(b) Deduce that the distinct class sums form a basis of $Z(F[G])$, so that $\dim Z(F[G])$ is the number of conjugacy classes of G.

(c) Compute $\dim Z(F[S_3])$.

Hint for (a): by bilinearity it is enough to test commutation against the basis elements $h \in G$.

Solution.

(a) By bilinearity, x is central if and only if $xh = hx$ for every $h \in G$, equivalently $h^{-1}xh = x$ for every h (multiply by h^{-1}, which exists in G). Now

$$h^{-1}xh = \sum_g a_g\, h^{-1}gh = \sum_u a_{huh^{-1}}\, u,$$

where the second equality substitutes $u = h^{-1}gh$, that is, $g = huh^{-1}$; this is a bijection of G, so it merely relabels the sum. Comparing coefficients with $x = \sum_u a_u u$, which is legitimate because the u are a basis, gives $h^{-1}xh = x$ for all h exactly when $a_{huh^{-1}} = a_u$ for all u, h.

(b) Being conjugate is an equivalence relation on G: $g = ege^{-1}$; if $g' = hgh^{-1}$ then $g = h^{-1}g'(h^{-1})^{-1}$; and conjugating twice is conjugating by the product. So G is the disjoint union of its distinct classes $K_1, \dots, K_r$. By (a), a central element has a constant coefficient on each class, so $Z(F[G])$ is exactly the span of the class sums $z_1, \dots, z_r$; and these are independent, because they are sums over **disjoint** nonempty sets of basis vectors. Hence $\dim Z(F[G]) = r$.

(c) Conjugate elements satisfy the same equations $g^k = e$: expanding and canceling the inner pairs $h^{-1}h$ gives $(hgh^{-1})^k = hg^kh^{-1}$, which equals e exactly when $g^k = e$. Now S_3 splits into three sets by this test: $\{e\}$; the three transpositions, which are the elements with $g^2 = e \neq g$; and the two 3-cycles, which are the elements with $g^3 = e \neq g$. No conjugacy class can meet two of these sets. Each of them is a single class: $(1\ 3)(1\ 2)(1\ 3)^{-1} = (2\ 3)$ and $(2\ 3)(1\ 2)(2\ 3)^{-1} = (1\ 3)$ make the three transpositions conjugate, and $(1\ 2)(1\ 2\ 3)(1\ 2)^{-1} = (1\ 3\ 2)$ makes the two 3-cycles conjugate. So $r = 3$ and $\dim Z(F[S_3]) = 3$.

Exercise. C2: Generated by two matrix units

Work in $M_2(F)$.

(a) Compute $F\langle\{\mathbf{E}_{12}\}\rangle$ and its dimension.

(b) Compute $F\langle\{\mathbf{E}_{12}, \mathbf{E}_{21}\}\rangle$ and its dimension.

(c) Deduce that adding one generator can change the commutant from an algebra of dimension 2 to the scalars, and compute both commutants.

Solution.

(a) $\mathbf{E}_{12}^2 = \mathbf{O}$, so by Proposition 23.1.1 the generated algebra is $\operatorname{span}(\mathbf{I}_2, \mathbf{E}_{12})$, of dimension 2. It is $F[\mathbf{E}_{12}]$, which matches Proposition 23.1.2 since $m_{\mathbf{E}_{12}} = x^2$.

(b) The words include $\mathbf{E}_{12}\mathbf{E}_{21} = \mathbf{E}_{11}$ and $\mathbf{E}_{21}\mathbf{E}_{12} = \mathbf{E}_{22}$, so the generated algebra contains all four matrix units and is $M_2(F)$, of dimension 4.

(c) $\{\mathbf{E}_{12}\}'$: writing $\mathbf{X} = (x_{ij})$, $\mathbf{E}_{12}\mathbf{X}$ has first row (x_{21}, x_{22}) and second row zero, while $\mathbf{X}\mathbf{E}_{12}$ has second column (x_{11}, x_{21}) and first column zero. Equality forces $x_{21} = 0$ and $x_{22} = x_{11}$, so $\{\mathbf{E}_{12}\}' = \operatorname{span}(\mathbf{I}_2, \mathbf{E}_{12})$, of dimension 2. By Proposition 23.1.4 (b), $\{\mathbf{E}_{12}, \mathbf{E}_{21}\}' \subseteq \{\mathbf{E}_{12}\}'$, and it equals $Z(M_2(F)) = F\mathbf{I}_2$ by part (b) and Proposition 23.1.5. One extra generator cuts the commutant from dimension 2 to dimension 1.

Exercise. C3: Is the group algebra ever a field?

Let G be a finite group with $|G| \geq 2$ and let F be any field.

(a) Pick $g \neq e$ in G of order $m \geq 2$, meaning $g^m = e$ and $g^i \neq e$ for $1 \leq i < m$. (Such an m exists: the powers $g, g^2, \ldots$ cannot all be distinct in a finite group, and canceling gives $g^i = e$ for some $i \geq 1$.) Show that $x = 1 - g$ and $y = 1 + g + \cdots + g^{m-1}$ satisfy $xy = 0$ with $x \neq 0$ and $y \neq 0$.

(b) Deduce that $F[G]$ is never a field, and never isomorphic to a subalgebra of a field.

(c) Is $F[G]$ ever commutative for $|G| \geq 2$? Give an example or a proof that it is not.

Solution.

(a) The powers $1, g, \ldots, g^{m-1}$ are distinct elements of G: if $g^i = g^j$ with $0 \leq i < j \leq m-1$ then $g^{j-i} = e$ with $1 \leq j - i < m$, contradicting minimality of m (cancellation is Theorem 0.10.1 (4)). So $x \neq 0$ and $y \neq 0$, both being nonzero combinations of distinct basis vectors. Expanding,

$$xy = \sum_{i=0}^{m-1} g^i - \sum_{i=0}^{m-1} g^{i+1} = 1 - g^m = 1 - 1 = 0.$$

(b) In a field, and more generally in any subalgebra of a field, a product of two nonzero elements is nonzero: if $xy = 0$ and $x \neq 0$ then $y = x^{-1}(xy) = 0$. Part (a) produces a violation, so $F[G]$ is not a field and embeds in none. (An algebra with such a pair x, y is said to have *zero divisors*.)

(c) Yes, whenever G is abelian: then $g_i g_j = g_j g_i$ on basis elements, and bilinearity extends the identity to all of $F[G]$. For example $F[\mathbb{Z}/2\mathbb{Z}]$ is commutative of dimension 2. Part (b) says it is still not a field, which is consistent: $\mathbb{Q}[\mathbb{Z}/2\mathbb{Z}] \cong \mathbb{Q} \times \mathbb{Q}$, a commutative algebra with zero divisors.

23.2 Ideals and Quotients

Subalgebras are the parts of an algebra one can multiply inside. Ideals are the parts one can divide by. Chapter 15 §10 introduced two-sided ideals for one purpose, to cut the Clifford algebra

out of the tensor algebra; here they become the tool for taking an algebra apart. Two things are new. First, in a noncommutative algebra there are **three** kinds of ideal, and the difference decides every statement below. Second, the main theorem of the section says that the matrix algebra has no two-sided ideals at all — which is why matrix algebras are the atoms out of which the rest of the chapter builds everything.

23.2.1 Three kinds of ideal

Section 1 left a set unexplained. In $M_2(F)$ the matrices $a\mathbf{E}_{11}$ form a subspace closed under products, but not a subalgebra: the identity is missing. What kind of object is it? Look instead at the slightly bigger set

$$C_1 = \{\mathbf{X} \in M_2(F) : \text{column 2 of } \mathbf{X} \text{ is } \mathbf{0}\} = \operatorname{span}(\mathbf{E}_{11}, \mathbf{E}_{21}).$$

Multiplying on the **left** by anything keeps a matrix in C_1, because $\mathbf{AX}$ has k-th column $\mathbf{Ax}_k$ (Theorem 0.9.2), so a zero column stays zero. Multiplying on the **right** does not: $\mathbf{E}_{11}\mathbf{E}_{12} = \mathbf{E}_{12} \notin C_1$. Absorption on one side is a genuine phenomenon, and it needs its own name.

An ideal is a subspace that absorbs multiplication: a left ideal from the left, a right ideal from the right, a two-sided ideal from both.

Definition 23.2.1. Left and Right Ideals

Let A be an algebra over F and let $I \subseteq A$ be a **subspace**. Then I is

(a) a **left ideal** of A if $ax \in I$ **for every** $a \in A$ and **every** $x \in I$;

(b) a **right ideal** of A if $xa \in I$ **for every** $a \in A$ and **every** $x \in I$.

A **two-sided ideal** is a subspace that is both, which is Definition 15.10.3. When A is commutative the three notions coincide, and we say **ideal**.

Read the quantifiers. The multiplier a runs over the **whole** algebra, not over I; being closed among one's own elements, which is clause (SA2) of a subalgebra, is a different and much weaker condition. And the definition is not symmetric: in C_1 above, "every a" works on the left and fails on the right.

Example. Ideals in five algebras

(a) In any algebra A: $\{0\}$ and A.

(b) In $F[x]$: the sets $\langle p \rangle$ of Definition 6.2.1.

(c) In $M_n(F)$: the column ideal C_j of matrices whose columns other than the j-th are zero.

(d) In $M_n(F)$: the set $R_U = \{\mathbf{X} : \operatorname{col}(\mathbf{X}) \subseteq U\}$ for a subspace $U \subseteq F^n$.

(e) In any algebra A: the kernel of an algebra homomorphism $\varphi\colon A \to B$.

Say in each case which of the three kinds of ideal is meant.

Solution.

(a) Two-sided, and these are the degenerate cases: $\{0\}$ absorbs everything because $a0 = 0$, and A absorbs everything because there is nowhere else to go.

(b) Two-sided, because $F[x]$ is commutative. Chapter 6's (I1)–(I3) are exactly Definition 23.2.1 read in a commutative algebra, with (I1) and (I2) replaced by the stronger requirement that I be a subspace — which is no change, since $cf = (c1)f \in I$ for a scalar c by (I3).

(c) A **left** ideal, by the computation above, of dimension n. Not a right ideal for $n \geq 2$: $\mathbf{E}_{jj}\mathbf{E}_{jk} = \mathbf{E}_{jk} \notin C_j$ for $k \neq j$.

(d) A **right** ideal: $\mathrm{col}(\mathbf{XA}) \subseteq \mathrm{col}(\mathbf{X}) \subseteq U$, because each column of $\mathbf{XA}$ is a combination of the columns of $\mathbf{X}$ (Theorem 0.9.2). Not a left ideal in general: for $U = \mathrm{span}(\mathbf{e}_1) \subseteq F^2$, $\mathbf{E}_{11} \in R_U$ but $\mathbf{E}_{21}\mathbf{E}_{11} = \mathbf{E}_{21}$ has column space $\mathrm{span}(\mathbf{e}_2) \not\subseteq U$.

(e) Two-sided: this is Lemma 15.10.5 (c).

The nonexample is the one Chapter 15 already gave, and it is worth repeating because it is the mistake to avoid. The diagonal matrices in $M_2(F)$ form a subspace closed under products — a subalgebra, in fact — and are **not** an ideal of any kind: $\mathbf{E}_{11}$ is diagonal, while $\mathbf{E}_{21}\mathbf{E}_{11} = \mathbf{E}_{21}$ and $\mathbf{E}_{11}\mathbf{E}_{12} = \mathbf{E}_{12}$ are not. Closure under one's own products is not absorption.

Why insist on absorbing products with the whole algebra? Because that is exactly the condition under which the quotient vector space A/I inherits a multiplication, as Lemma 15.10.5 (a) shows: the computation $x'y' = xy + uy + xw + uw$ needs uy and xw to lie in I when u, w do, with x, y arbitrary. The condition is necessary as well: taking $u = 0$ leaves xw, which must lie in I for every $x \in A$ and $w \in I$, and taking $w = 0$ leaves uy, which must lie in I for every $u \in I$ and $y \in A$. A one-sided ideal gives only half of that, and half is not enough.

Warning. A two-sided ideal is almost never a subalgebra, and a subalgebra is almost never an ideal. If a left ideal I contains an invertible element u, then for every $a \in A$ we have $a = (au^{-1})u \in I$, so $I = A$. In particular a **proper** ideal of any kind contains no invertible element, hence does not contain 1_A, hence fails (SA3). The only ideal that is a subalgebra is A itself.

Quick check. In $M_2(F)$, is $\mathrm{span}(\mathbf{E}_{11})$ a left ideal? A right ideal?

Solution. Neither. It is a subspace, but $\mathbf{E}_{21}\mathbf{E}_{11} = \mathbf{E}_{21} \notin \mathrm{span}(\mathbf{E}_{11})$, so it is not a left ideal, and $\mathbf{E}_{11}\mathbf{E}_{12} = \mathbf{E}_{12} \notin \mathrm{span}(\mathbf{E}_{11})$, so it is not a right ideal. It is closed under its own products, which is why it looked like something in the warning of Section 1; it is neither a subalgebra nor an ideal. The smallest left ideal containing it is C_1, and the smallest two-sided ideal containing it is all of $M_2(F)$, by the theorem below.

23.2.2 Matrix units multiply in one line

Everything about ideals of $M_n(F)$ follows from a single product rule. Recall that the **matrix unit** $\mathbf{E}_{ij}$ has a 1 in position (i, j) and zeros elsewhere, and that the n^2 matrix units are a basis of $M_n(F)$ (Example (Standard Bases)).

Lemma 23.2.1. The Matrix Unit Product Rule

Let F be a field, $n \geq 1$, and $1 \leq i, j, k, l \leq n$. Then, in $M_n(F)$,

$$\mathbf{E}_{ij}\mathbf{E}_{kl} = \delta_{jk}\,\mathbf{E}_{il}, \qquad \text{and} \qquad \mathbf{E}_{ij}\,\mathbf{X}\,\mathbf{E}_{kl} = x_{jk}\,\mathbf{E}_{il} \quad \text{for every } \mathbf{X} = (x_{pq}).$$

Moreover $\mathbf{E}_{11} + \mathbf{E}_{22} + \cdots + \mathbf{E}_{nn} = \mathbf{I}_n$.

Proof. Write $(\mathbf{E}_{ij})_{pq} = \delta_{pi}\delta_{qj}$, which is the definition of the matrix unit. By Theorem 0.9.2 (1),

$$(\mathbf{E}_{ij}\mathbf{E}_{kl})_{pq} = \sum_{r} \delta_{pi}\delta_{rj}\,\delta_{rk}\delta_{ql} = \delta_{pi}\delta_{ql}\sum_{r}\delta_{rj}\delta_{rk} = \delta_{jk}\,\delta_{pi}\delta_{ql},$$

since the only surviving term of the last sum is $r = j$, which contributes 1 when $k = j$ and 0 otherwise. The right-hand side is the (p, q) entry of $\delta_{jk}\mathbf{E}_{il}$.

For the second identity, expand $\mathbf{X} = \sum_{r,s} x_{rs}\mathbf{E}_{rs}$ in the basis of matrix units and use bilinearity together with the first identity twice:

$$\mathbf{E}_{ij}\mathbf{X}\mathbf{E}_{kl} = \sum_{r,s} x_{rs}\,\mathbf{E}_{ij}\mathbf{E}_{rs}\mathbf{E}_{kl} = \sum_{r,s} x_{rs}\,\delta_{jr}\delta_{sk}\,\mathbf{E}_{il} = x_{jk}\,\mathbf{E}_{il}.$$

The last statement is the observation that $\sum_i \mathbf{E}_{ii}$ has (p, q) entry $\sum_i \delta_{pi}\delta_{qi} = \delta_{pq}$. This proves the lemma.

■

The second identity is the one to remember: *sandwiching a matrix between two matrix units extracts one entry and moves it to a chosen place.* Choosing i and l freely, a single nonzero entry of $\mathbf{X}$ can be moved anywhere.

23.2.3 The matrix algebra is simple

An algebra whose only two-sided ideals are $\{0\}$ and itself is called **simple**; Section 3 makes this a definition, adding the clause that the algebra be nonzero. The theorem says that matrix algebras are simple over every field.

Theorem 23.2.2. The Matrix Algebra Has No Proper Two-Sided Ideals

Let F be **any** field and $n \geq 1$. The only two-sided ideals of $M_n(F)$ are $\{\mathbf{O}\}$ and $M_n(F)$. Consequently, for a finite-dimensional vector space $V \neq \{0\}$ over F, the only two-sided ideals of $\mathcal{L}(V)$ are $\{0\}$ and $\mathcal{L}(V)$.

Idea. An ideal that contains anything nonzero contains a matrix with some entry $x_{jk} \neq 0$. Sandwich it: Lemma 23.2.1 turns that one entry into an arbitrary matrix unit, up to the nonzero scalar x_{jk}, which we may divide out. Once the ideal has all n^2 matrix units, it has their span, and it also has their sum, which is $\mathbf{I}_n$.

Proof. Both $\{\mathbf{O}\}$ and $M_n(F)$ are two-sided ideals. Let $\mathcal{I}$ be a two-sided ideal with $\mathcal{I} \neq \{\mathbf{O}\}$, and choose $\mathbf{X} \in \mathcal{I}$ with $\mathbf{X} \neq \mathbf{O}$. Then some entry x_{jk} is nonzero; fix such a pair (j, k).

Let i and l be arbitrary. Since $\mathcal{I}$ absorbs multiplication on both sides, $\mathbf{E}_{ij}\mathbf{X}\mathbf{E}_{kl} \in \mathcal{I}$, and by Lemma 23.2.1 this element is $x_{jk}\mathbf{E}_{il}$. As $x_{jk} \neq 0$ it is invertible in F, and $\mathcal{I}$ is a subspace, so

$$\mathbf{E}_{il} = x_{jk}^{-1}(x_{jk}\mathbf{E}_{il}) \in \mathcal{I}.$$

Hence $\mathcal{I}$ contains every matrix unit. The matrix units span $M_n(F)$ (Example (Standard Bases)) and $\mathcal{I}$ is a subspace, so $\mathcal{I} = M_n(F)$.

For $\mathcal{L}(V)$, fix an ordered basis of V and use the algebra isomorphism $\mathcal{L}(V) \to M_n(F)$ recorded after Definition 23.1.3. An algebra isomorphism carries two-sided ideals to two-sided ideals, since it is a linear bijection and $\varphi(ax) = \varphi(a)\varphi(x)$ in both directions. This proves the theorem.

■

Nothing in the proof used anything about F beyond the existence of x_{jk}^{-1}, which is the field axiom. The theorem holds over $\mathbb{Q}$, over $\mathbb{C}$, and over $\mathbb{F}_2$ alike, and it is the reason matrix algebras will turn out to be the building blocks in this chapter's structure theorem: they cannot be broken.

One consequence is worth recording; §03 proves it again for every simple algebra.

Corollary 23.2.3. Homomorphisms out of a Matrix Algebra Are Injective

Let F be a field, $n \geq 1$, and let $\varphi\colon M_n(F) \to B$ be an algebra homomorphism into any F-algebra $B \neq \{0\}$. Then φ is injective.

Proof. By Lemma 15.10.5 (c), $\ker\varphi$ is a two-sided ideal of $M_n(F)$, so by Theorem 23.2.2 it is $\{\mathbf{O}\}$ or $M_n(F)$. If $\ker\varphi = M_n(F)$, then $1_B = \varphi(\mathbf{I}_n) = 0$, and then $b = 1_B b = 0$ for every $b \in B$, contradicting $B \neq \{0\}$. So $\ker\varphi = \{\mathbf{O}\}$ and φ is injective (Theorem 2.2.3).

■

Warning. "Simple" says nothing about one-sided ideals. $M_n(F)$ has no two-sided ideals other than the two obvious ones, and for $n \geq 2$ it has a whole lattice of left ideals, as the next result shows. Whenever a statement in this chapter says "ideal", check which kind is meant before using it.

23.2.4 The left ideals of the matrix algebra

The left ideals are not trivial, and they are worth knowing exactly: they are indexed by subspaces of row vectors, and the smallest of them are the column ideals.

Proposition 23.2.4. The Left Ideals of $M_n(F)$

Let F be a field and $n \geq 1$. For a subspace $W \subseteq M_{1\times n}(F)$ of row vectors, put

$$L_W := \{\mathbf{X} \in M_n(F) : \operatorname{row}(\mathbf{X}) \subseteq W\}.$$

(a) L_W is a left ideal of $M_n(F)$, and $\dim L_W = n \dim W$.

(b) Every left ideal L of $M_n(F)$ equals L_W for exactly one W, namely the span of the union of the row spaces of the elements of L:

$$W = \operatorname{span}\Big(\bigcup_{\mathbf{X}\in L} \operatorname{row}(\mathbf{X})\Big).$$

(c) $W \mapsto L_W$ is an inclusion-preserving bijection from subspaces of $M_{1\times n}(F)$ onto left ideals of $M_n(F)$. The minimal nonzero left ideals are the L_W with $\dim W = 1$.

(d) Taking $W = \operatorname{span}(\mathbf{e}_j^\top)$ gives the column ideal C_j of Example (Ideals in five algebras) (c), of dimension n, and $M_n(F) = C_1 \oplus \cdots \oplus C_n$ as a direct sum of left ideals.

Idea. Two facts drive everything. First, left multiplication can only shrink the row space: row i of $\mathbf{AX}$ is a combination of the rows of $\mathbf{X}$. Second, left multiplication can produce **every** matrix whose rows lie in the row space of $\mathbf{X}$, because the rows of $\mathbf{AX}$ are arbitrary combinations of the rows of $\mathbf{X}$, one combination per row and chosen independently. So a left ideal is determined by the row vectors it can reach, and it contains every matrix built from them.

Proof. Throughout we use two consequences of Theorem 0.9.2 (3): row i of $\mathbf{AX}$ is (row i of $\mathbf{A}$) $\mathbf{X}$; and for a row vector $\mathbf{a}^\top \in M_{1\times n}(F)$, the product $\mathbf{a}^\top\mathbf{X}$ is the combination of the rows of $\mathbf{X}$ with weights $a_1, \ldots, a_n$, so that

$$\operatorname{row}(\mathbf{X}) = \{\, \mathbf{a}^\top\mathbf{X} \;:\; \mathbf{a}^\top \in M_{1\times n}(F) \,\}. \tag{$*$}$$

(a) L_W is a subspace: every row of $\mathbf{X}+c\mathbf{Y}$ is the corresponding row of $\mathbf{X}$ plus c times that of $\mathbf{Y}$, hence lies in W when both do. It is a left ideal: each row of $\mathbf{AX}$ lies in $\operatorname{row}(\mathbf{X}) \subseteq W$ by the first consequence. For the dimension, the map $M_n(F) \to (M_{1\times n}(F))^n$ sending $\mathbf{X}$ to its list of rows is a linear bijection, and it carries L_W onto $W \times \cdots \times W$ — the rows may be chosen from W independently of one another. So $\dim L_W = n \dim W$ by Theorem 4.4.1.

(b) Let L be a left ideal and put $W := \operatorname{span}(\bigcup_{\mathbf{X}\in L} \operatorname{row}(\mathbf{X}))$, a subspace of $M_{1\times n}(F)$ by Theorem 1.3.1; then $L \subseteq L_W$ by construction.
For the reverse inclusion, choose $\mathbf{X}_1, \ldots, \mathbf{X}_m \in L$ with $W = \operatorname{row}(\mathbf{X}_1) + \cdots + \operatorname{row}(\mathbf{X}_m)$. Such a choice exists: W is a subspace of the n-dimensional space $M_{1\times n}(F)$, so it has a finite basis; each basis vector is by Definition 1.3.2 a finite combination of row vectors drawn from the spaces $\operatorname{row}(\mathbf{X})$ with $\mathbf{X} \in L$, and each such space is a subspace, so that basis vector lies in the sum of the finitely many $\operatorname{row}(\mathbf{X})$ involved. Collecting the $\mathbf{X}$ used across the whole basis gives $\mathbf{X}_1, \ldots, \mathbf{X}_m \in L$ with $W \subseteq \operatorname{row}(\mathbf{X}_1) + \cdots + \operatorname{row}(\mathbf{X}_m)$, and the reverse containment holds because each $\operatorname{row}(\mathbf{X}_t) \subseteq W$. Let $\mathbf{Y} \in L_W$ and let $\mathbf{y}_1^\top, \ldots, \mathbf{y}_n^\top$ be its rows, all in W. Each $\mathbf{y}_p^\top$ lies in $W = \operatorname{row}(\mathbf{X}_1) + \cdots + \operatorname{row}(\mathbf{X}_m)$, hence is a sum of elements of the $\operatorname{row}(\mathbf{X}_t)$, so by $(*)$ there are row vectors $\mathbf{a}_{pt}^\top$ with

$$\mathbf{y}_p^\top = \sum_{t=1}^{m} \mathbf{a}_{pt}^\top \mathbf{X}_t \qquad (1 \leq p \leq n).$$

Let $\mathbf{A}_t \in M_n(F)$ be the matrix whose p-th row is $\mathbf{a}_{pt}^\top$. Then row p of $\sum_t \mathbf{A}_t\mathbf{X}_t$ is $\sum_t \mathbf{a}_{pt}^\top\mathbf{X}_t = \mathbf{y}_p^\top$, so $\mathbf{Y} = \sum_t \mathbf{A}_t\mathbf{X}_t$. Each $\mathbf{A}_t\mathbf{X}_t$ lies in L because L is a left ideal, and L is a subspace, so $\mathbf{Y} \in L$. Hence $L = L_W$.
Uniqueness: L_W determines W, because $W = \operatorname{span}\big(\bigcup_{\mathbf{X}\in L_W} \operatorname{row}(\mathbf{X})\big)$. Indeed that span is contained in W by definition, and it contains every $\mathbf{a}^\top \in W$, since the matrix $\mathbf{e}_1\mathbf{a}^\top$, whose first row is $\mathbf{a}^\top$ and whose other rows are zero, lies in L_W.

(c) If $W \subseteq W'$ then $L_W \subseteq L_{W'}$; conversely $L_W \subseteq L_{W'}$ forces $W \subseteq W'$, by the description of W just given. With (b), the map is a bijection, and it preserves and reflects inclusion, so it matches minimal nonzero left ideals with minimal nonzero subspaces, which are the lines.

(d) A matrix has all its rows in $\operatorname{span}(\mathbf{e}_j^\top)$ exactly when every entry outside column j is zero, which is the description of C_j; its dimension is $n \cdot 1 = n$ by (a). Every $\mathbf{X}$ is the sum of its columns placed in the corresponding C_j, and a matrix lying in C_j and in the sum of the others is zero, so the sum is direct. This proves the proposition.

■

Remark. Everything transposes. Since $(\mathbf{XA})^\top = \mathbf{A}^\top\mathbf{X}^\top$, the linear bijection $\mathbf{X} \mapsto \mathbf{X}^\top$ carries left ideals onto right ideals and back. Applying it to the proposition, and writing $U = \{\mathbf{a} \in F^n : \mathbf{a}^\top \in W\}$, the right ideals of $M_n(F)$ are exactly the sets $R_U = \{\mathbf{X} : \operatorname{col}(\mathbf{X}) \subseteq U\}$ of Example (Ideals in five algebras) (d), for subspaces $U \subseteq F^n$, with $\dim R_U = n \dim U$. It is worth keeping the two statements apart: column spaces classify **right** ideals, row spaces classify **left** ideals.

23.2.5 Quotients, and the first isomorphism theorem

Chapter 15 §10 did the work. Lemma 15.10.5 says that the quotient **space** $A/\mathcal{I}$ of Chapter 2 (Definition 4.4.3) carries exactly one multiplication making $\pi\colon A \to A/\mathcal{I}$ multiplicative, namely $(x+\mathcal{I})(y+\mathcal{I}) = xy+\mathcal{I}$, and that with it $A/\mathcal{I}$ is an algebra with identity $1_A + \mathcal{I}$. We give the object its name.

Definition 23.2.2. Quotient Algebra

Let A be an algebra over F and $\mathcal{I}$ a **two-sided** ideal of A. The **quotient algebra** $A/\mathcal{I}$ is the quotient vector space $A/\mathcal{I}$ with the multiplication $(x+\mathcal{I})(y+\mathcal{I}) = xy+\mathcal{I}$ of Lemma 15.10.5, and $\pi\colon A \to A/\mathcal{I}$, $\pi(x) = x+\mathcal{I}$, is the **quotient homomorphism**.

Two facts come for free. The ideal must be two-sided: a *merely* one-sided ideal gives a quotient vector space, but the product of cosets is not well defined on it. And when A is finite-dimensional, $\dim(A/\mathcal{I}) = \dim A - \dim \mathcal{I}$ by Theorem 4.4.5.

Theorem 23.2.5. First Isomorphism Theorem for Algebras

Let $\varphi\colon A \to B$ be a homomorphism of F-algebras. Then $\ker\varphi$ is a two-sided ideal of A, $\operatorname{im}\varphi$

is a subalgebra of B, and the rule

$$\bar{\varphi}\colon A/\ker\varphi \to B, \qquad \bar{\varphi}(x+\ker\varphi) = \varphi(x),$$

defines an injective algebra homomorphism with image $\operatorname{im}\varphi$. Hence

$$A/\ker\varphi \cong \operatorname{im}\varphi$$

as F-algebras.

Idea. Chapter 4 already produced the linear isomorphism; all that is left is to see that it respects products, and that is one line, because the product on the quotient was **defined** so that π is multiplicative.

Proof. $\ker\varphi$ is a two-sided ideal by Lemma 15.10.5 (c). $\operatorname{im}\varphi$ is a subalgebra of B: it is a subspace (Theorem 2.2.2), it is closed under products since $\varphi(x)\varphi(y) = \varphi(xy)$, and it contains $1_B = \varphi(1_A)$.

Write $K = \ker\varphi$. By Theorem 4.4.7, $\bar{\varphi}$ is a well-defined **linear** map, it is injective, its image is $\operatorname{im}\varphi$, and $\bar{\varphi}\circ\pi = \varphi$.

It is multiplicative. Every element of A/K is $\pi(x)$ for some $x \in A$, and by Definition 23.2.2 and the multiplicativity of φ,

$$\bar{\varphi}(\pi(x)\pi(y)) = \bar{\varphi}(\pi(xy)) = \varphi(xy) = \varphi(x)\varphi(y) = \bar{\varphi}(\pi(x))\,\bar{\varphi}(\pi(y)).$$

It is unital: $\bar{\varphi}(1_{A/K}) = \bar{\varphi}(\pi(1_A)) = \varphi(1_A) = 1_B$. So $\bar{\varphi}$ is an injective algebra homomorphism onto $\operatorname{im}\varphi$, hence an algebra isomorphism $A/K \to \operatorname{im}\varphi$. This proves the theorem.

■

The theorem turns every surjective homomorphism into a quotient, and that is how quotient algebras are recognized in practice: build a homomorphism, compute its kernel, and read off the answer. Here are two instances.

Example. The group algebra of a cyclic group

Let F be a field, $n \geq 1$, and $G = \mathbb{Z}/n\mathbb{Z}$ written multiplicatively as $\{1, g, \dots, g^{n-1}\}$ with $g^n = 1$. Show that

$$F[G] \cong F[x]/\langle x^n - 1\rangle.$$

Solution. Example (Three group algebras) (b) produced an algebra homomorphism $\varepsilon\colon F[x] \to F[G]$ with $\varepsilon(p) = \sum_i c_i g^i$ for $p = \sum_i c_i x^i$, showed that it is surjective, and identified $\ker\varepsilon = \langle x^n - 1\rangle$. Note that $\langle x^n - 1\rangle$ is a two-sided ideal of $F[x]$, since $F[x]$ is commutative. By Theorem 23.2.5,

$$F[x]/\langle x^n - 1\rangle \cong \operatorname{im}\varepsilon = F[G].$$

Both sides have dimension n: the left by Theorem 6.6.5, the right by Definition 23.1.5. This

pays off the promise made in Section 1, and it says something the group alone does not: the group algebra of a cyclic group is a ring of polynomials with one relation.

The second instance needs a two-element algebra built from F. For F-algebras A and B, the **product algebra** $A \times B$ is the product vector space (Definition 4.4.1) with componentwise multiplication $(a, b)(a', b') = (aa', bb')$. Bilinearity, associativity, and the identity $(1_A, 1_B)$ are inherited componentwise, so $A \times B$ is an algebra, of dimension $\dim A + \dim B$ (Theorem 4.4.1). We need only $F \times F$, of dimension 2.

Example. Upper triangular matrices modulo their strictly upper part

Let $\mathcal{T}_2$ be the upper triangular matrices in $M_2(F)$ and let $\mathcal{N} = \operatorname{span}(\mathbf{E}_{12})$ be the strictly upper triangular ones.

(a) Show that $\mathcal{N}$ is a two-sided ideal of $\mathcal{T}_2$.

(b) Identify $\mathcal{T}_2/\mathcal{N}$.

Solution. Write a general element of $\mathcal{T}_2$ as $\mathbf{U} = \left(\begin{smallmatrix} a & b \\ 0 & c \end{smallmatrix}\right)$ and note

$$\begin{pmatrix} a & b \\ 0 & c \end{pmatrix}\begin{pmatrix} p & q \\ 0 & r \end{pmatrix} = \begin{pmatrix} ap & aq + br \\ 0 & cr \end{pmatrix}. \tag{$\dagger$}$$

(a) $\mathcal{N}$ is the set of elements of $\mathcal{T}_2$ with $a = c = 0$, a subspace. Putting $a = c = 0$ in ($\dagger$) gives a product with both diagonal entries 0, and putting $p = r = 0$ does the same; so $\mathcal{N}$ absorbs multiplication on both sides.

(b) Define $\varphi\colon \mathcal{T}_2 \to F \times F$ by $\varphi(\mathbf{U}) = (a, c)$. It is linear, it sends $\mathbf{I}_2$ to $(1, 1)$, and by ($\dagger$) it is multiplicative, since the diagonal of a product of upper triangular matrices is the product of the diagonals. It is surjective, since $\varphi(\operatorname{diag}(a, c)) = (a, c)$, and its kernel is exactly $\mathcal{N}$. By Theorem 23.2.5,

$$\mathcal{T}_2/\mathcal{N} \cong F \times F.$$

A dimension check: $\dim \mathcal{T}_2 = 3$, $\dim \mathcal{N} = 1$, $\dim(F \times F) = 2 = 3 - 1$, as Theorem 4.4.5 requires.

So $\mathcal{T}_2$ is **not** simple: it has a proper nonzero two-sided ideal, and dividing by it leaves the most transparent algebra there is, a product of copies of F. Theorem 23.2.2 says $M_2(F)$ admits no such surgery. That contrast — an algebra with an ideal to divide out, against one with none — is the division the rest of the chapter runs on.

23.2.6 Exercises

23.2.6.1 A. Check your understanding

Exercise. A1

(a) Define left ideal, right ideal, and two-sided ideal, and say which of the three is needed to form a quotient algebra.

(b) State the product rule for matrix units, and say which step of the proof of Theorem 23.2.2 uses that F is a field.

(c) True or false, with a reason: a two-sided ideal of an algebra A is a subalgebra of A.

(d) Give a left ideal of $M_3(F)$ of dimension 6, and say which subspace of row vectors it corresponds to.

Solution.

(a) See Definition 23.2.1: a subspace absorbing multiplication from the left, from the right, or from both. Forming a quotient algebra needs a **two-sided** ideal, because the product of cosets is otherwise not well defined (Lemma 15.10.5 (a)).

(b) $\mathbf{E}_{ij}\mathbf{E}_{kl} = \delta_{jk}\mathbf{E}_{il}$, and $\mathbf{E}_{ij}\mathbf{X}\mathbf{E}_{kl} = x_{jk}\mathbf{E}_{il}$ (Lemma 23.2.1). The field is used when dividing by the nonzero scalar x_{jk} to extract $\mathbf{E}_{il}$ from $x_{jk}\mathbf{E}_{il}$.

(c) False, unless $\mathcal{I} = A$. A proper ideal contains no invertible element, in particular not 1_A, so (SA3) fails; see the warning after Example (Ideals in five algebras).

(d) L_W with $W = \operatorname{span}(\mathbf{e}_1^\top, \mathbf{e}_2^\top)$, the matrices whose third column is zero; $\dim L_W = 3 \cdot 2 = 6$ by Proposition 23.2.4 (a).

23.2.6.2 B. Practice

Exercise. B1: Determine which are ideals

Determine, for each of the following subspaces, whether it is a left ideal, a right ideal, a two-sided ideal, or none. Justify your answer.

(a) $\{\mathbf{X} \in M_2(F) : x_{11} + x_{22} = 0\}$.

(b) $\{\mathbf{X} \in M_2(F) : \text{row 2 of } \mathbf{X} \text{ is } \mathbf{0}\}$.

(c) $\{\mathbf{X} \in M_3(F) : \mathbf{X}\mathbf{e}_1 = \mathbf{0}\}$.

(d) $\{p \in F[x] : p(0) = 0\}$ inside $F[x]$.

Solution.

(a) None. It is a subspace, but $\mathbf{E}_{12}$ lies in it while $\mathbf{E}_{21}\mathbf{E}_{12} = \mathbf{E}_{22}$ does not, so it is not a left ideal; and $\mathbf{E}_{12}\mathbf{E}_{21} = \mathbf{E}_{11}$ is not in it, so it is not a right ideal. (One may also argue from Theorem 23.2.2: a two-sided ideal of dimension 3 is impossible.)

(b) A **right** ideal, not a left ideal. The set is a subspace, and right multiplication keeps the condition: by Theorem 0.9.2 (3), row 2 of $\mathbf{X}\mathbf{A}$ is (row 2 of $\mathbf{X}$) $\mathbf{A} = \mathbf{0}$. Left multiplication does not: $\mathbf{E}_{11}$ has second row $\mathbf{0}$, while $\mathbf{E}_{21}\mathbf{E}_{11} = \mathbf{E}_{21}$ does not.

(c) A **left** ideal, not a right ideal. It is L_W for $W = \operatorname{span}(\mathbf{e}_2^\top, \mathbf{e}_3^\top)$: a matrix kills $\mathbf{e}_1$ exactly when its first column is zero, that is, when every row lies in W. Left multiplication keeps the condition since $(\mathbf{A}\mathbf{X})\mathbf{e}_1 = \mathbf{A}(\mathbf{X}\mathbf{e}_1) = \mathbf{0}$. Right multiplication does not: $\mathbf{E}_{12}\mathbf{e}_1 = \mathbf{0}$, while $(\mathbf{E}_{12}\mathbf{E}_{21})\mathbf{e}_1 = \mathbf{E}_{11}\mathbf{e}_1 = \mathbf{e}_1 \neq \mathbf{0}$.

(d) A two-sided ideal, since $F[x]$ is commutative: it is $\langle x \rangle$ by Theorem 6.1.3, and $(hp)(0) = h(0)p(0) = 0$.

Exercise. B2: A quotient computed

Let $A = F[x]/\langle x^2 \rangle$.

(a) Write down a basis of A and its multiplication table.

(b) Show that $\mathcal{I} = \operatorname{span}(x + \langle x^2 \rangle)$ is a two-sided ideal of A and identify $A/\mathcal{I}$.

(c) Show that A is not isomorphic to $F \times F$.

Solution.

(a) By Theorem 6.6.5, $\dim A = 2$, with basis $(1 + \langle x^2 \rangle,\ x + \langle x^2 \rangle)$. Write 1 and t for these. The table is $1 \cdot 1 = 1$, $1 \cdot t = t \cdot 1 = t$, and $t^2 = x^2 + \langle x^2 \rangle = 0$.

(b) $\mathcal{I} = \operatorname{span}(t)$ is a subspace, and $(a + bt)t = at \in \mathcal{I}$, $t(a + bt) = at \in \mathcal{I}$, so it absorbs on both sides. The map $\varphi\colon A \to F$, $a + bt \mapsto a$, is linear, sends 1 to 1, and is multiplicative because $(a+bt)(a'+b't) = aa' + (ab' + a'b)t$. Its kernel is $\mathcal{I}$, so $A/\mathcal{I} \cong F$ by Theorem 23.2.5.

(c) In $F \times F$ the equation $y^2 = 0$ forces $y = 0$, since a product of scalars vanishes only when a factor does. In A the nonzero element t satisfies $t^2 = 0$. An algebra isomorphism preserves squares and sends 0 to 0, so no isomorphism exists.

Exercise. B3: An ideal generated

Inside $M_3(F)$, let $\mathcal{I} = \langle \mathbf{E}_{23} \rangle$ be the two-sided ideal generated by $\mathbf{E}_{23}$ in the sense of Definition 15.10.3. Compute $\mathcal{I}$, and do the same for the **left** ideal generated by $\mathbf{E}_{23}$, that is, $\operatorname{span}\{\mathbf{A}\mathbf{E}_{23} : \mathbf{A} \in M_3(F)\}$.

Solution. By Definition 15.10.3, $\mathcal{I}$ is spanned by the elements $\mathbf{A}\mathbf{E}_{23}\mathbf{B}$. Taking $\mathbf{A} = \mathbf{E}_{i2}$ and $\mathbf{B} = \mathbf{E}_{3l}$ gives $\mathbf{E}_{i2}\mathbf{E}_{23}\mathbf{E}_{3l} = \mathbf{E}_{il}$ by Lemma 23.2.1, for all i and l. So $\mathcal{I}$ contains every matrix unit and $\mathcal{I} = M_3(F)$, of dimension 9. This is Theorem 23.2.2 in a single instance.

The left ideal is different. By $(*)$ in the proof of Proposition 23.2.4, $\operatorname{row}(\mathbf{E}_{23}) = \operatorname{span}(\mathbf{e}_3^{\top})$, so $\{\mathbf{A}\mathbf{E}_{23}\}$ is exactly the set of matrices all of whose rows are multiples of $\mathbf{e}_3^{\top}$, that is, the column ideal C_3, of dimension 3. Generating on one side and on two sides give a three-dimensional and a nine-dimensional answer.

23.2.6.3 C. Going deeper

Exercise. C1: All the ideals of the upper triangular algebra

Let $\mathcal{T}_2$ be the upper triangular matrices in $M_2(F)$, with basis $(\mathbf{E}_{11}, \mathbf{E}_{12}, \mathbf{E}_{22})$.

(a) Prove that $\mathcal{J}_1 = \operatorname{span}(\mathbf{E}_{11}, \mathbf{E}_{12})$ and $\mathcal{J}_2 = \operatorname{span}(\mathbf{E}_{12}, \mathbf{E}_{22})$ are two-sided ideals.

(b) Prove that $\{\mathbf{O}\}$, $\operatorname{span}(\mathbf{E}_{12})$, $\mathcal{J}_1$, $\mathcal{J}_2$, and $\mathcal{T}_2$ are the **only** two-sided ideals of $\mathcal{T}_2$.

(c) Identify $\mathcal{T}_2/\mathcal{J}_1$ and $\mathcal{T}_2/\mathcal{J}_2$.

Hint for (b): if an ideal contains an element with $a \neq 0$, multiply it by $\mathbf{E}_{11}$ on both sides.

Solution. Throughout write $\mathbf{U} = \left(\begin{smallmatrix} a & b \\ 0 & c \end{smallmatrix}\right)$ and use (†) of Example (Upper triangular matrices modulo their strictly upper part).

(a) $\mathcal{J}_1$ is the set of $\mathbf{U}$ with $c = 0$. Putting $c = 0$ in (†) gives a product with lower right entry $cr = 0$; putting $r = 0$ does the same. So $\mathcal{J}_1$ absorbs on both sides. Symmetrically, $\mathcal{J}_2$ is the set with $a = 0$, and (†) gives upper left entry ap, which vanishes if $a = 0$ or $p = 0$.

(b) Let $\mathcal{I}$ be a two-sided ideal and take $\mathbf{U} \in \mathcal{I}$. By (†), $\mathbf{E}_{11}\mathbf{U}\mathbf{E}_{11} = a\mathbf{E}_{11}$ and $\mathbf{E}_{22}\mathbf{U}\mathbf{E}_{22} = c\mathbf{E}_{22}$, both in $\mathcal{I}$. Also $\mathbf{E}_{11}\mathbf{U}\mathbf{E}_{22} = b\mathbf{E}_{12} \in \mathcal{I}$.
Case 1: some element of $\mathcal{I}$ has $a \neq 0$. Then $\mathbf{E}_{11} \in \mathcal{I}$, hence $\mathbf{E}_{12} = \mathbf{E}_{11}\mathbf{E}_{12} \in \mathcal{I}$, so $\mathcal{J}_1 \subseteq \mathcal{I}$. If in addition some element has $c \neq 0$, then $\mathbf{E}_{22} \in \mathcal{I}$ and $\mathcal{I} = \mathcal{T}_2$; otherwise every element has $c = 0$ and $\mathcal{I} = \mathcal{J}_1$.
Case 2: every element of $\mathcal{I}$ has $a = 0$. If some element has $c \neq 0$, then $\mathbf{E}_{22} \in \mathcal{I}$, hence $\mathbf{E}_{12} = \mathbf{E}_{12}\mathbf{E}_{22} \in \mathcal{I}$, so $\mathcal{J}_2 \subseteq \mathcal{I} \subseteq \mathcal{J}_2$ and $\mathcal{I} = \mathcal{J}_2$. Otherwise every element has $a = c = 0$, so $\mathcal{I} \subseteq \operatorname{span}(\mathbf{E}_{12})$, and $\mathcal{I}$ is $\{\mathbf{O}\}$ or $\operatorname{span}(\mathbf{E}_{12})$.
These five exhaust the possibilities.

(c) $\mathbf{U} \mapsto c$ is a surjective algebra homomorphism $\mathcal{T}_2 \to F$ by (†), with kernel $\mathcal{J}_1$; so $\mathcal{T}_2/\mathcal{J}_1 \cong F$ by Theorem 23.2.5. Symmetrically $\mathbf{U} \mapsto a$ gives $\mathcal{T}_2/\mathcal{J}_2 \cong F$.

Exercise. C2: Simplicity over a small field, and a failure

(a) Verify Theorem 23.2.2 by hand for $M_2(\mathbb{F}_2)$, starting from the ideal generated by $\mathbf{E}_{12} + \mathbf{E}_{21}$.

(b) Let $A = F \times F$. Show that A has exactly four two-sided ideals, and conclude that a commutative algebra of dimension 2 can fail to be simple.

(c) Deduce that $M_2(F)$ and $F^4 = F \times F \times F \times F$ are not isomorphic as F-algebras, although they have the same dimension.

Solution.

(a) Let $\mathbf{X} = \mathbf{E}_{12} + \mathbf{E}_{21}$ and let $\mathcal{I}$ be the two-sided ideal it generates. Its $(1,2)$ entry is $1 \neq 0$, so by Lemma 23.2.1, $\mathbf{E}_{i1}\mathbf{X}\mathbf{E}_{2l} = x_{12}\mathbf{E}_{il} = \mathbf{E}_{il}$ for all $i, l \in \{1,2\}$. Hence $\mathcal{I}$ contains $\mathbf{E}_{11}, \mathbf{E}_{12}, \mathbf{E}_{21}, \mathbf{E}_{22}$ and so all of $M_2(\mathbb{F}_2)$. Over $\mathbb{F}_2$ no division was needed, because the nonzero entry was already 1.

(b) Let $\mathcal{I} \subseteq F \times F$ be an ideal and put $e_1 = (1,0)$, $e_2 = (0,1)$. For $(u,v) \in \mathcal{I}$ we get $(u,0) = e_1(u,v) \in \mathcal{I}$ and $(0,v) \in \mathcal{I}$. So $\mathcal{I}$ is determined by whether it contains a pair with $u \neq 0$ and whether it contains a pair with $v \neq 0$, giving $\{0\}$, $F \times \{0\}$, $\{0\} \times F$, and $F \times F$. Each of these four is indeed an ideal, by componentwise multiplication. The two middle ones are proper and nonzero, so $F \times F$ is not simple.

(c) $M_2(F)$ has exactly two two-sided ideals (Theorem 23.2.2), while F^4 has more: the same argument as in (b) shows that $F \times \{0\} \times \{0\} \times \{0\}$ is a proper nonzero ideal. An algebra isomorphism carries two-sided ideals bijectively to two-sided ideals, so no isomorphism can exist. Both algebras have dimension 4, so dimension alone classifies nothing.

Exercise. C3: Left ideals and null spaces

Let F be a field and $n \geq 1$.

(a) For a subspace $U \subseteq F^n$, prove that $N_U = \{\mathbf{X} \in M_n(F) : \mathbf{X}\mathbf{u} = \mathbf{0} \text{ for every } \mathbf{u} \in U\}$ is a left ideal of $M_n(F)$.

(b) Prove that $\dim N_U = n(n - \dim U)$.

(c) Prove that $U \mapsto N_U$ is an **inclusion-reversing** bijection from subspaces of F^n onto left ideals of $M_n(F)$, and describe N_U in the language of Proposition 23.2.4.

Hint for (b): a matrix kills every vector of U exactly when each of its rows annihilates U.

Solution.

(a) N_U is the intersection over $\mathbf{u} \in U$ of the null spaces of the linear maps $\mathbf{X} \mapsto \mathbf{X}\mathbf{u}$, hence a subspace. If $\mathbf{X} \in N_U$ and $\mathbf{A} \in M_n(F)$, then $(\mathbf{A}\mathbf{X})\mathbf{u} = \mathbf{A}(\mathbf{X}\mathbf{u}) = \mathbf{0}$, so $\mathbf{A}\mathbf{X} \in N_U$.

(b) Let $\mathbf{a}^\top$ be a row of $\mathbf{X}$. The entries of $\mathbf{X}\mathbf{u}$ are the numbers $\mathbf{a}^\top\mathbf{u}$ as $\mathbf{a}^\top$ runs over the rows, so $\mathbf{X}\mathbf{u} = \mathbf{0}$ for every $\mathbf{u} \in U$ exactly when every row of $\mathbf{X}$ annihilates U, that is, when $\operatorname{row}(\mathbf{X}) \subseteq W_U$, where W_U is the space of row vectors annihilating U. Identifying a row vector $\mathbf{a}^\top$ with the functional $\mathbf{u} \mapsto \mathbf{a}^\top\mathbf{u}$, which is a linear bijection $M_{1\times n}(F) \to (F^n)^*$ by Theorem 5.1.1, the space W_U corresponds to U^0, so $\dim W_U = n - \dim U$ by Theorem 5.2.2. Hence $N_U = L_{W_U}$, and Proposition 23.2.4 (a) gives $\dim N_U = n(n - \dim U)$.

(c) It is enough to show that $U \mapsto W_U$, which under the identification of (b) is $U \mapsto U^0$, is an inclusion-reversing bijection from subspaces of F^n onto subspaces of $M_{1\times n}(F)$; composing with the inclusion-preserving bijection $W \mapsto L_W$ of Proposition 23.2.4 (c) then gives the claim, since $N_U = L_{W_U}$ by (b).

It reverses inclusions by Theorem 5.2.3 (a), and it is injective by Corollary 5.3.5. For surjectivity, let W be a subspace of $(F^n)^*$ of dimension k, with basis $\varphi_1, \dots, \varphi_k$, and put $U = \ker\varphi_1 \cap \cdots \cap \ker\varphi_k$. By Theorem 5.6.7, U has codimension k, so $\dim U = n - k$ and $\dim U^0 = k$ by Theorem 5.2.2. Every φ_i vanishes on U, so $W \subseteq U^0$; equal dimensions give $W = U^0$.

23.3 Simple and Semisimple Algebras

Section 2 found that $M_n(F)$ has no two-sided ideals except the two unavoidable ones: as an algebra it cannot be collapsed. This section is about a second kind of object that is indecomposable — not the algebra, but a space the algebra acts on — and about what happens when a space does break apart. The main theorem says that "every invariant subspace can be split off" and "the space is a sum of unbreakable pieces" are one condition wearing two faces, and that condition is what the rest of the chapter runs on. At the end we check the new vocabulary against the old: an operator is semisimple in Chapter 10's sense exactly when the algebra it generates acts semisimply.

Throughout, F is **any** field. Nothing in this section asks it to be algebraically closed or to have a particular characteristic; the hypotheses start biting in the next section. Algebras are associative, unital, and finite-dimensional over F, as agreed in Section 1.

23.3.1 Two meanings of the word simple

Section 2 proved that $M_n(F)$ has exactly two two-sided ideals (Theorem 23.2.2), while $F[x]$ — the infinite-dimensional standing exception of Section 1 — has infinitely many. That difference deserves a name.

A simple algebra is one that cannot be collapsed: forming the quotient by an ideal either changes nothing or destroys everything.

Definition 23.3.1. Simple Algebra

An F-algebra A is **simple** if $A \neq \{0\}$ and the only two-sided ideals (Definition 15.10.3) of A are $\{0\}$ and A itself.

In words: two conditions, and the first is not decoration. The zero algebra has only one ideal, so without the clause $A \neq \{0\}$ it would count as simple and every theorem below would need an exception for it.

Examples.

- **The field itself.** $A = F$ is simple: a nonzero ideal contains some $c \neq 0$, hence contains $c^{-1}c = 1$, hence everything.
- **Matrix algebras.** $M_n(F)$ is simple for every $n \geq 1$ (Theorem 23.2.2). This is the example the whole chapter is aimed at.
- **A degenerate case.** $n = 1$ recovers $M_1(F) = F$, so the two examples above are one.

Nonexample by minimal change. Replace $M_n(F)$ by the algebra A of **upper triangular** 2×2 matrices over F. It is a subalgebra of $M_2(F)$ of dimension 3, with basis $\mathbf{E}_{11}, \mathbf{E}_{12}, \mathbf{E}_{22}$, and the set

$$\mathcal{N} = \left\{ \begin{pmatrix} 0 & b \\ 0 & 0 \end{pmatrix} : b \in F \right\}$$

is a two-sided ideal of A: it is a subspace, and

$$\begin{pmatrix} a & c \\ 0 & d \end{pmatrix} \begin{pmatrix} 0 & b \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & ab \\ 0 & 0 \end{pmatrix},$$
$$\begin{pmatrix} 0 & b \\ 0 & 0 \end{pmatrix} \begin{pmatrix} a & c \\ 0 & d \end{pmatrix} = \begin{pmatrix} 0 & bd \\ 0 & 0 \end{pmatrix}.$$

Since $\{\mathbf{O}\} \neq \mathcal{N} \neq A$, the algebra A is not simple. The clause that fails is the second one; $A \neq \{0\}$ is fine.

- **Another nonexample, the infinite-dimensional one.** $F[x]$ has the ideal $\langle x \rangle$ (Proposition 6.2.1), the polynomials with zero constant term (Example (First Ideals) (b)), which is neither $\{0\}$ nor $F[x]$.

Why this definition. Section 2 identified ideals with kernels: the kernel of an algebra homomorphism out of A is a two-sided ideal, and conversely every two-sided ideal is the kernel of the map onto the quotient algebra (Definition 23.2.2, Theorem 23.2.5). Simplicity therefore says that A admits no interesting image under a homomorphism, which is the following one-line consequence.

Proposition 23.3.1. Maps Out of a Simple Algebra

Let A be a simple F-algebra, let $B \neq \{0\}$ be an F-algebra, and let $\varphi\colon A \to B$ be an algebra homomorphism (Definition 23.1.2). Then φ is injective.

Proof. Since $B \neq \{0\}$ we have $1_B \neq 0$: otherwise $y = 1_B y = 0$ for every $y \in B$. The kernel of φ is a two-sided ideal of A, and it does not contain 1_A, because $\varphi(1_A) = 1_B \neq 0$. So $\ker\varphi \neq A$, and simplicity forces $\ker\varphi = \{0\}$. A linear map with zero kernel is injective (Theorem 2.2.3). ∎

Section 2 had already seen the case $A = M_n(F)$, which is simple by Theorem 23.2.2: that is Corollary 23.2.3. So a simple algebra has only one image under a homomorphism worth the name: itself. Section 6 will turn this around and recognize the **semisimple** algebras, over an algebraically closed field, as products of matrix algebras; its Exercise C1 picks out the simple ones among them as the single matrix algebras.

23.3.2 Algebras acting on spaces

An algebra is already a set of operators, by Section 1's embedding theorem (Theorem 23.1.3). But the same abstract algebra acts on many different spaces — $M_n(F)$ acts on F^n, on itself, on $M_n(F)$ by conjugation — and the differences between those actions are the subject. We need a word for one of them.

Definition 23.3.2. Representation of an Algebra, and Invariant Subspaces

Let A be an F-algebra. A **representation** of A on a finite-dimensional vector space V over F is an algebra homomorphism

$$\rho\colon A \to \mathcal{L}(V),$$

so that ρ is linear, $\rho(ab) = \rho(a)\rho(b)$, and $\rho(1_A) = \mathrm{id}_V$. We then call V an **A-space** and abbreviate $\rho(a)(v)$ to av. A subspace $U \subseteq V$ is **A-invariant** if $au \in U$ for every $a \in A$ and every $u \in U$.

In words: the multiplication of A is carried to composition of operators, and — the clause that is easy to overlook — **the identity of A must act as the identity of V**. Two standing examples:

- **The tautological representation.** If $A \subseteq \mathcal{L}(V)$ is a subalgebra containing id_V (Definition 23.1.1), the inclusion is a representation. This is the case we usually mean when we say "an algebra of operators on V".
- **The regular representation.** A acts on itself by left multiplication: $\rho(a)(b) = ab$. Associativity gives $\rho(ab) = \rho(a)\rho(b)$, bilinearity gives linearity, and $\rho(1_A) = \mathrm{id}_A$. Here the A-invariant subspaces are precisely the **left ideals** of Section 2.

Nonexample by minimal change. Take $A = F$, $V = F^2$, and $\rho(c) = 0$ for every c. This is linear and multiplicative, but $\rho(1) = 0 \neq \mathrm{id}_V$, so it is not a representation. The clause that fails is unitality, and it is not a technicality: it is what makes $v = 1v$ lie in Av, a step used twice below.

Now the indecomposable actions.

An irreducible A-space is one with nothing inside it for A to see.

Definition 23.3.3. Irreducible, or Simple, A-space

An A-space V is **irreducible** — equivalently, **simple** — if $V \neq \{0\}$ and the only A-invariant subspaces of V are $\{0\}$ and V.

Both words are standard and we use them interchangeably. Again two clauses, and again the first is needed: $\{0\}$ has only one subspace and must be excluded by hand, exactly as the zero algebra was.

Examples.

- **Lines.** Any A-space of dimension 1 is irreducible, whatever A is, since its only subspaces are $\{0\}$ and itself. These are the smallest possible pieces, and Section 4 will show that over an algebraically closed field a commutative algebra has no others.
- **The column space of a matrix algebra.** F^n is an irreducible $M_n(F)$-space. Indeed, let $U \neq \{\mathbf{0}\}$ be invariant and pick $\mathbf{u} \in U$ with $\mathbf{u} \neq \mathbf{0}$. Given any $\mathbf{w} \in F^n$, extend $\mathbf{u}$ to a basis of F^n (Theorem 1.6.5) and let $\mathbf{X}$ be the matrix of the linear map sending $\mathbf{u}$ to $\mathbf{w}$ and the other basis vectors to $\mathbf{0}$ (Theorem 2.1.6). Then $\mathbf{w} = \mathbf{X}\mathbf{u} \in U$. So $U = F^n$.

Nonexample by minimal change. Shrink $M_2(F)$ to the upper triangular algebra A above, still acting on F^2. The line $F\mathbf{e}_1$ is A-invariant, because an upper triangular $\mathbf{X}$ has $\mathbf{X}\mathbf{e}_1 = x_{11}\mathbf{e}_1$. So F^2 is not an irreducible A-space; the clause that fails is the second.

Warning. **Two different words are both spelled "simple".** An algebra is simple when it has no proper nonzero **two-sided ideal**; an A-space is simple when it has no proper nonzero **invariant subspace**. The first does not imply the second. For $n \geq 2$ the algebra $M_n(F)$ is simple, yet $M_n(F)$ as a space over itself is **not**: the matrices whose columns other than the first are zero form a left ideal that is neither $\{0\}$ nor all of $M_n(F)$, that is, a proper nonzero invariant subspace of the regular representation.

Quick check. Let $A \subseteq M_2(F)$ be the algebra of **diagonal** matrices and let $V = F^2$. Which subspaces of V are A-invariant? Is V irreducible?

Solution. The invariant subspaces are $\{\mathbf{0}\}$, $F\mathbf{e}_1$, $F\mathbf{e}_2$, and F^2. Each of these four is invariant because $\mathrm{diag}(d_1, d_2)$ fixes each coordinate line. No other line $F\mathbf{v}$ with $\mathbf{v} = (x, y)$ is: $\mathbf{E}_{11}\mathbf{v} = (x, 0)$ must be a multiple $c(x, y)$ of $\mathbf{v}$, and if both $x \neq 0$ and $y \neq 0$ then $cy = 0$ forces $c = 0$ and hence $x = 0$, a contradiction. So V is not irreducible: $F\mathbf{e}_1$ is a proper nonzero invariant subspace.

23.3.3 Splitting a space into pieces

Here is the phenomenon the section is built on. Take $V = F^2$ and the single operator $\mathbf{N} = \mathbf{J}_2(0)$, that is, $\mathbf{N}\mathbf{e}_1 = \mathbf{0}$ and $\mathbf{N}\mathbf{e}_2 = \mathbf{e}_1$. The line $F\mathbf{e}_1$ is invariant. Is there an invariant complement? A

complement of a line in F^2 is a line $F\mathbf{v}$ with $\mathbf{v} = (x, y)$ and $y \neq 0$, and $\mathbf{Nv} = (y, 0)$ is a nonzero element of $F\mathbf{e}_1$, which meets $F\mathbf{v}$ only in $\mathbf{0}$. So $\mathbf{Nv} \notin F\mathbf{v}$: there is **no** invariant complement, and F^2 does not break into invariant lines.

Compare $\mathbf{D} = \mathrm{diag}(1, 2)$, where $F^2 = F\mathbf{e}_1 \oplus F\mathbf{e}_2$ splits perfectly. The difference between these two pictures is the definition we want.

An algebra acts semisimply when every invariant subspace can be split off.

Definition 23.3.4. Semisimple Action, Semisimple Algebra

Let A be an F-algebra and V an A-space. We say A **acts semisimply on** V if **every** A-invariant subspace $U \subseteq V$ has an A-invariant complement: a subspace $W \subseteq V$ with

$$V = U \oplus W \qquad \text{and} \qquad W \text{ is } A\text{-invariant.}$$

The algebra A is **semisimple** if it acts semisimply on **itself** in the regular representation — that is, if every left ideal of A has a complementary left ideal.

In words: the first notion is a property of an action, the second a property of the algebra alone, obtained by feeding the algebra its own most faithful space. The examples come after the theorem that makes the definition usable.

Warning. **Acting semisimply depends on the space, not only on the algebra.** The upper triangular algebra A does not act semisimply on F^2, as we check below. Yet it does act semisimply on the one-dimensional space F through $\mathbf{X} \mapsto x_{11}$, for the silly reason that a one-dimensional space has no invariant subspace to split off. Always say **which space**. The unqualified phrase "A is semisimple" is reserved for the regular representation, and a proposition below shows that this is the right choice: a semisimple algebra acts semisimply on every space.

The first thing to know is that the property passes to invariant subspaces. Without this, no induction on dimension gets started.

Lemma 23.3.2. Splitting Is Inherited by Invariant Subspaces

Let A act semisimply on a finite-dimensional A-space V, and let $W \subseteq V$ be A-invariant. Then W, with the action $a \mapsto \rho(a)|_W$, is an A-space, and A acts semisimply on W.

Idea. An invariant subspace U of W is also one of V, so it has an invariant complement U' in V. That complement is too big; cut it down to $U' \cap W$ and check that nothing was lost.

Proof. Since W is A-invariant, $\rho(a)|_W \in \mathcal{L}(W)$ for every $a \in A$, and $a \mapsto \rho(a)|_W$ is linear, multiplicative, and sends 1_A to id_W; so it is a representation.

Let $U \subseteq W$ be A-invariant. Then U is an A-invariant subspace of V, so there is an A-invariant U' with $V = U \oplus U'$. Put $W' = U' \cap W$, which is A-invariant because U' and W are.

$U + W' = W$: the inclusion $\subseteq$ holds since $U \subseteq W$ and $W' \subseteq W$. Conversely, let $w \in W$ and write $w = u + u'$ with $u \in U$, $u' \in U'$. Then $u' = w - u \in W$, because both w and u lie in W; hence $u' \in U' \cap W = W'$ and $w \in U + W'$.

The sum is direct: $U \cap W' \subseteq U \cap U' = \{0\}$.

Therefore $W = U \oplus W'$ with W' an A-invariant complement, which is what semisimplicity of the action on W asks for.

■

Now the theorem. It is the section's real content, and every later section uses it.

Theorem 23.3.3. Complete Reducibility

Let A be an F-algebra and let V be a finite-dimensional A-space. The following are equivalent:

(a) A acts semisimply on V;
(b) V is a sum of irreducible A-invariant subspaces;
(c) V is a direct sum of finitely many irreducible A-invariant subspaces.

Here a sum over the empty family is $\{0\}$, so all three hold when $V = \{0\}$. No hypothesis is placed on F.

Idea. For (a) $\Rightarrow$ (c), induct on $\dim V$. The smallest nonzero invariant subspace is automatically irreducible: it has no room for a proper one. Split it off using (a), and apply the inductive hypothesis to the complement — which is legitimate precisely because of Lemma 23.3.2. For (b) $\Rightarrow$ (a), given an invariant U, build a complement greedily out of the given irreducible pieces: add pieces to U as long as the sum stays direct, and then show that a piece left outside would have to meet the sum in $\{0\}$, so it could have been added too.

Proof.

(a) $\Rightarrow$ (c). Induction on $\dim V$. If $\dim V = 0$ then V is the empty direct sum and there is nothing to prove. Let $\dim V \geq 1$ and suppose the implication holds for every A-space of smaller dimension.

Among the nonzero A-invariant subspaces of V — there is at least one, namely V — choose W of smallest dimension. Then W is irreducible: it is nonzero, and if $U \subseteq W$ is A-invariant with $U \neq \{0\}$, then U is a nonzero A-invariant subspace of V, so $\dim U \geq \dim W$ by the choice of W, and $U = W$ by Theorem 1.6.8.

By (a) there is an A-invariant W' with $V = W \oplus W'$. By Lemma 23.3.2, W' is an A-space on which A acts semisimply, and $\dim W' = \dim V - \dim W < \dim V$ because $\dim W \geq 1$. The inductive hypothesis writes $W' = W_1 \oplus \cdots \oplus W_r$ with each W_i an irreducible A-invariant subspace of W'. Invariance and irreducibility refer only to the action of A, not to the ambient space, so each W_i is an irreducible A-invariant subspace of V as well. Hence

$$V = W \oplus W_1 \oplus \cdots \oplus W_r,$$

a direct sum of finitely many irreducibles.

(c) $\Rightarrow$ (b). A direct sum is in particular a sum.

(d) $\Rightarrow$ (a). Write $V = \sum_{i\in I} W_i$ with each W_i an irreducible A-invariant subspace, and let $U \subseteq V$ be A-invariant. Call a finite subset $J \subseteq I$ **good** if the sum $U + \sum_{j\in J} W_j$ is direct. The empty set is good. Every good J satisfies

$$|J| \le \dim U + \sum_{j\in J} \dim W_j = \dim\Big(U \oplus \bigoplus_{j\in J} W_j\Big) \le \dim V,$$

where the first inequality uses $\dim W_j \ge 1$ and the middle equality is Theorem 1.7.5. So the sizes of good sets are bounded, and we may choose a good J with $|J|$ as large as possible. Put

$$W = \sum_{j\in J} W_j,$$

an A-invariant subspace, since a sum of A-invariant subspaces is A-invariant. Write $S = U \oplus W$.

Claim: $W_i \subseteq S$ for every $i \in I$. Fix i. The subspace $W_i \cap S$ is A-invariant and contained in the irreducible W_i, so it is $\{0\}$ or W_i. Suppose it were $\{0\}$. Then $i \notin J$, since $j \in J$ gives $W_j \subseteq S$ and $W_j \neq \{0\}$. We show $J \cup \{i\}$ is good, contradicting maximality. Let

$$u + \sum_{j\in J} w_j + w_i = 0, \qquad u \in U,\ w_j \in W_j,\ w_i \in W_i.$$

Then $w_i = -(u + \sum_j w_j) \in W_i \cap S = \{0\}$, so $w_i = 0$; and then $u + \sum_j w_j = 0$ forces $u = 0$ and every $w_j = 0$, because J is good (Theorem 1.7.5 (b)). So the sum over $J \cup \{i\}$ is direct, which is the contradiction. Hence $W_i \cap S = W_i$, that is, $W_i \subseteq S$.

The claim gives $V = \sum_i W_i \subseteq S$, so $V = S = U \oplus W$ with W invariant. This proves (a).

■

Two features of the proof are worth keeping. The complement produced in (b) $\Rightarrow$ (a) is a sum of some of the given pieces, so nothing new has to be constructed. And the pieces in (c) are far from unique: F^2 under the scalar matrices is a direct sum of **any** two distinct lines.

Example. Upper Triangular Matrices Are Not Semisimple

Let $A \subseteq M_2(F)$ be the algebra of upper triangular matrices, of dimension 3. Show that A does not act semisimply on F^2, and list the irreducible A-subspaces of F^2.

Solution. The A-invariant subspaces of F^2 are $\{\mathbf{0}\}$, $F\mathbf{e}_1$, and F^2, and no others. Certainly all three are invariant, the middle one because $\mathbf{X}\mathbf{e}_1 = x_{11}\mathbf{e}_1$ for upper triangular $\mathbf{X}$. Conversely, let $F\mathbf{v}$ be an invariant line with $\mathbf{v} = (x, y)$. Since $\mathbf{E}_{12} \in A$ and $\mathbf{E}_{12}\mathbf{v} = (y, 0)$, invariance gives $(y, 0) = c(x, y)$ for some $c \in F$. If $y \neq 0$ then $cy = 0$ gives $c = 0$, hence $y = 0$ after all; so $y = 0$ and $F\mathbf{v} = F\mathbf{e}_1$.

Now $F\mathbf{e}_1$ is an invariant subspace with no invariant complement, since a complement of a line in F^2 is a line, and $F\mathbf{e}_1$ is the only invariant one. So A does not act semisimply on

F^2. The only irreducible A-subspace is $F\mathbf{e}_1$; in particular F^2 is not a sum of irreducibles, in agreement with Theorem 23.3.3.

For contrast, here is an action that does split, and it is the one Section 6 will build on.

Example. Matrix Algebras Are Semisimple

Show that $A = M_n(F)$ is a semisimple algebra for every $n \geq 1$.

Solution. Section 2 did the work. For $j = 1, \dots, n$ let $C_j \subseteq A$ be the column ideal: the matrices all of whose columns except the j-th are zero. By Proposition 23.2.4 (d), each C_j is a left ideal and $A = C_1 \oplus \cdots \oplus C_n$. Moreover $C_j = L_W$ for the one-dimensional space of row vectors $W = \operatorname{span}(\mathbf{e}_j^{\top})$, so C_j is a **minimal** nonzero left ideal by part (c) of the same proposition.

The A-invariant subspaces of the regular representation are the left ideals of A, so "minimal nonzero left ideal" says exactly "irreducible A-subspace". Hence the regular representation of A is a direct sum of irreducibles, and Theorem 23.3.3 makes A semisimple.

The definition of a semisimple **algebra** only looked at one space. It controls all of them.

Proposition 23.3.4. A Semisimple Algebra Has No Other Kind of Representation

Let A be a semisimple F-algebra and let V be any finite-dimensional A-space. Then A acts semisimply on V.

Idea. Every vector v lies in Av, because 1_A acts as the identity; so V is the sum of the subspaces Av. Chop A itself into irreducible left ideals L and show each Lv is either zero or irreducible. Then V is a sum of irreducibles and Theorem 23.3.3 finishes.

Proof. If $V = \{0\}$ there is nothing to prove, so let $\dim V \geq 1$. Since A is semisimple, Theorem 23.3.3 applied to the regular representation gives $A = L_1 \oplus \cdots \oplus L_k$ with each L_m an irreducible left ideal.

Claim: for a left ideal L that is irreducible and a vector $v \in V$, the subspace $Lv = \{av : a \in L\}$ is either $\{0\}$ or an irreducible A-invariant subspace of V. It is a subspace, being the image of the linear map $a \mapsto av$ restricted to L, and it is A-invariant because $b(av) = (ba)v$ with $ba \in L$. Suppose $Lv \neq \{0\}$ and let $U \subseteq Lv$ be A-invariant with $U \neq \{0\}$. Put

$$L_0 = \{a \in L : av \in U\}.$$

This is a subspace of L, and it is a left ideal: for $b \in A$ and $a \in L_0$ we have $ba \in L$ and $(ba)v = b(av) \in U$ since U is invariant. As L is irreducible, $L_0 = \{0\}$ or $L_0 = L$. If $L_0 = \{0\}$, then every $u \in U$ has the form $u = av$ with $a \in L$, whence $a \in L_0 = \{0\}$ and $u = 0$, contradicting $U \neq \{0\}$. So $L_0 = L$, which says $Lv \subseteq U$ and therefore $U = Lv$. This proves the claim.

Now let $v \in V$. Since $1_A v = v$, we have $v \in Av = L_1 v + \cdots + L_k v$. Hence

$$V = \sum_{v \in V} Av = \sum_{v \in V} \sum_{m=1}^{k} L_m v,$$

and discarding the summands that are $\{0\}$ exhibits V as a sum of irreducible A-invariant subspaces. By Theorem 23.3.3, A acts semisimply on V.

■

23.3.4 How far from semisimple: the radical

The upper triangular algebra failed to act semisimply on F^2, and the ideal $\mathcal{N}$ was visibly to blame: it is nonzero, yet it squares to zero, so no irreducible space can feel it. That suggests measuring the failure by collecting everything invisible.

We write $\operatorname{rad} A$ for what follows. Chapter 14 §01 writes $\operatorname{rad}(\beta)$ for the radical of a bilinear form; here the argument is an algebra and the two never meet.

Definition 23.3.5. Radical of an Algebra

Let A be an F-algebra. Its **radical** is

$$\operatorname{rad} A := \bigcap_{V} \{\, a \in A : aV = \{0\} \,\},$$

where V runs over all irreducible A-spaces and $aV = \{av : v \in V\}$.

In words: the elements that every irreducible representation sends to 0. The quantifier ranges over a manageable supply: if V is irreducible and $v \in V$ is nonzero, then Av is a nonzero invariant subspace, so $V = Av$ is the image of the linear map $a \mapsto av$, and $\dim V \le \dim A$.

The radical is a two-sided ideal. It is a subspace, being an intersection of kernels of linear maps; and if $a \in \operatorname{rad} A$, $x, y \in A$, and V is irreducible, then $(xay)v = x(a(yv)) = x0 = 0$ for every $v \in V$. Taking $y = 1_A$ and $x = 1_A$ in turn shows xa and ay lie in $\operatorname{rad} A$.

The first fact we want is the one the triangular example suggested.

Lemma 23.3.5. A Nilpotent Ideal Is Invisible

Let A be an F-algebra and let $\mathcal{I} \subseteq A$ be a two-sided ideal which is **nilpotent**: for some $m \ge 1$, every product $a_1 a_2 \cdots a_m$ of m elements of $\mathcal{I}$ is 0. Then $\mathcal{I} \subseteq \operatorname{rad} A$.

Proof. Let V be an irreducible A-space and write $\mathcal{I}V = \operatorname{span}\{av : a \in \mathcal{I},\ v \in V\}$, and more generally $\mathcal{I}^k V = \operatorname{span}\{a_1 \cdots a_k v : a_i \in \mathcal{I},\ v \in V\}$.

The subspace $\mathcal{I}V$ is A-invariant, since $b(av) = (ba)v$ with $ba \in \mathcal{I}$. By irreducibility, $\mathcal{I}V = \{0\}$ or $\mathcal{I}V = V$.

Suppose $\mathcal{I}V = V$. Applying $\mathcal{I}$ to both sides, $\mathcal{I}^2 V = \mathcal{I}(\mathcal{I}V) = \mathcal{I}V = V$, and repeating gives $\mathcal{I}^k V = V$ for every $k \ge 1$. But $\mathcal{I}^m V = \{0\}$ by nilpotency, so $V = \{0\}$, contradicting irreducibility.

Hence $\mathcal{I}V = \{0\}$, that is, every $a \in \mathcal{I}$ acts as 0 on V. Since V was an arbitrary irreducible A-space, $\mathcal{I} \subseteq \operatorname{rad} A$.

■

And the radical really does obstruct semisimplicity.

Proposition 23.3.6. A Semisimple Algebra Has Zero Radical

Let A be a semisimple F-algebra. Then $\operatorname{rad} A = \{0\}$.

Proof. By Theorem 23.3.3 applied to the regular representation, $A = L_1 \oplus \cdots \oplus L_k$ with each L_m an irreducible A-space. Let $a \in \operatorname{rad} A$. Then a acts as 0 on each L_m, so $aL_m = \{0\}$ for every m, and therefore $aA = \{0\}$. In particular $a = a1_A = 0$.

■

Remark. The converse is also true for finite-dimensional algebras: $\operatorname{rad} A = \{0\}$ forces A to be semisimple. We do not prove it, and nothing in this book uses it; whenever semisimplicity is needed later it is established directly, by Maschke's theorem in Section 7 or by exhibiting a decomposition.

We can now finish the running example.

Quick check. Let A be the algebra of upper triangular 2×2 matrices and $\mathcal{N}$ its ideal of strictly upper triangular matrices. Using the two results above, identify $\operatorname{rad} A$ and decide whether A is a semisimple algebra. *Hint: the maps* $\mathbf{X} \mapsto x_{11}$ *and* $\mathbf{X} \mapsto x_{22}$ *go from* A *to* F.

Solution. $\operatorname{rad} A = \mathcal{N}$, and A is not semisimple.

($\supseteq$) We saw that $\mathcal{N}$ is a two-sided ideal, and $\mathcal{N}^2 = \{\mathbf{O}\}$ because $\mathbf{E}_{12}\mathbf{E}_{12} = \mathbf{O}$. By Lemma 23.3.5, $\mathcal{N} \subseteq \operatorname{rad} A$.

($\subseteq$) For upper triangular $\mathbf{X}, \mathbf{Y}$ the product has $(\mathbf{XY})_{11} = x_{11}y_{11}$ and $(\mathbf{XY})_{22} = x_{22}y_{22}$. So $\rho_1(\mathbf{X}) = x_{11}$ and $\rho_2(\mathbf{X}) = x_{22}$ are algebra homomorphisms $A \to F$, each sending $\mathbf{I}_2$ to 1; they make F an A-space in two ways, irreducible because $\dim F = 1$. An element of $\operatorname{rad} A$ must be killed by both, so it has $x_{11} = x_{22} = 0$ and lies in $\mathcal{N}$.

Since $\operatorname{rad} A = \mathcal{N} \neq \{\mathbf{O}\}$, Proposition 23.3.6 shows A is not a semisimple algebra.

23.3.5 Back to a single operator

Chapter 10 §08 already used the word semisimple, for an operator: T is semisimple when m_T is a product of **distinct** monic irreducible polynomials (Definition 10.8.1). That definition was about a polynomial. This section's is about invariant subspaces. They agree, and proving so is the best check that the new definition is the right one.

Fix a finite-dimensional $V \neq \{0\}$ over F and $T \in \mathcal{L}(V)$, and let

$$F[T] \coloneqq \{\, p(T) : p \in F[x] \,\} \subseteq \mathcal{L}(V),$$

which is a commutative subalgebra of $\mathcal{L}(V)$ containing id_V (Proposition 23.1.2). A subspace U is $F[T]$-invariant if and only if it is T-invariant: one direction is the case $p = x$, and the other holds because $TU \subseteq U$ gives $T^k U \subseteq U$ for every k, hence $p(T)U \subseteq U$ for every p.

Theorem 23.3.7. The Two Meanings of Semisimple Agree

Let $V \neq \{0\}$ be a finite-dimensional vector space over F and let $T \in \mathcal{L}(V)$. The following are equivalent:

(a) T is semisimple in the sense of Definition 10.8.1, that is, m_T is a product of distinct monic irreducible polynomials;

(b) $F[T]$ acts semisimply on V: every T-invariant subspace has a T-invariant complement;

(c) V is a sum of irreducible $F[T]$-invariant subspaces.

No hypothesis is placed on F.

Idea.

(b) and (c) are Theorem 23.3.3 for the algebra $F[T]$, so the work is the polynomial statement (a). Forwards: the primary decomposition splits V into pieces on which the minimal polynomial is a single irreducible p, and inside such a piece **every** nonzero vector generates a cyclic subspace of dimension $\deg p$ — which leaves no room for a proper invariant subspace. Backwards: an irreducible piece is cyclic, and a factorization of its minimal polynomial would manufacture a smaller invariant subspace inside it; so each piece contributes one irreducible polynomial, and their product, which is squarefree, kills T.

Proof. By the remark preceding the theorem, "T-invariant" and "$F[T]$-invariant" are the same condition, so (b) $\Leftrightarrow$ (c) is Theorem 23.3.3 applied to $A = F[T]$.

(a) $\Rightarrow$ (c). Write $m_T = p_1 \cdots p_k$ with $p_1, \dots, p_k$ distinct monic irreducibles, so that every exponent in the factorization is 1. By Theorem 9.8.1, $V = V_1 \oplus \cdots \oplus V_k$ with $V_i = \ker p_i(T)$ a nonzero T-invariant subspace on which $m_{T|_{V_i}} = p_i$.

Fix i and let $v \in V_i$ with $v \neq 0$. Then $p_i(T)v = 0$, so the T-annihilator $m_{T,v}$ divides p_i (Definition 10.5.2); it is monic of degree at least 1 because $v \neq 0$, and p_i is irreducible, so $m_{T,v} = p_i$. By Theorem 10.5.2 (a), the cyclic subspace $Z(v;T)$ has dimension $\deg p_i$, and it is T-invariant (Proposition 10.5.1).

$Z(v;T)$ is irreducible. Let $U \subseteq Z(v;T)$ be T-invariant with $U \neq \{0\}$ and pick $u \in U$ with $u \neq 0$. Since $u \in V_i$, the same argument gives $m_{T,u} = p_i$ and $\dim Z(u;T) = \deg p_i$. As $Z(u;T) \subseteq U \subseteq Z(v;T)$ and the two ends have equal dimension, Theorem 1.6.8 gives $U = Z(v;T)$.

Every nonzero $v \in V_i$ lies in $Z(v;T) \subseteq V_i$, so V_i is a sum of irreducible T-invariant subspaces, and hence so is $V = V_1 + \cdots + V_k$.

(c) $\Rightarrow$ (a). Write $V = \sum_{i \in I} W_i$ with each W_i irreducible.

Each $m_{T|_{W_i}}$ is irreducible. Pick $w \in W_i$ with $w \neq 0$. Then $Z(w;T)$ is a nonzero T-invariant subspace of W_i, so $Z(w;T) = W_i$, and Theorem 10.5.2 (a), (c) give $q := m_{T|_{W_i}} = m_{T,w}$ with $\deg q = \dim W_i \geq 1$. Suppose q were not irreducible. Since $\deg q \geq 1$, Definition 6.3.1 then gives $q = fg$ with both f and g nonconstant, and we may take them monic after dividing each by its leading coefficient. So $1 \leq \deg f < \deg q$ and $1 \leq \deg g < \deg q$. Put $u = g(T)w$. Then $u \neq 0$: otherwise $g(T)w = 0$ would give $m_{T,w} \mid g$ (Definition 10.5.2), impossible since $g \neq 0$ has degree less than $\deg m_{T,w}$. Also $f(T)u = (fg)(T)w = q(T)w = 0$, so $m_{T,u} \mid f$ and, by Theorem 10.5.2 (a), $\dim Z(u;T) \leq \deg f < \dim W_i$. But $Z(u;T)$ is a nonzero T-invariant subspace of W_i, hence equal to W_i by irreducibility — a contradiction. So no such factorization exists and q is irreducible.

Finitely many pieces suffice. Start with one W_{i_1}; as long as the sum of the chosen pieces is not V, some W_i is not contained in it, and adding that one strictly increases the dimension. The process stops, giving $V = W_{i_1} + \cdots + W_{i_r}$.

Conclusion. Let $q_1, \ldots, q_s$ be the distinct polynomials among $m_{T|_{W_{i_1}}}, \ldots, m_{T|_{W_{i_r}}}$ and put $q = q_1 \cdots q_s$, a product of distinct monic irreducibles. For each j, the polynomial $m_{T|_{W_{i_j}}}$ divides q, so $q(T)$ vanishes on W_{i_j}; as these subspaces span V, $q(T) = 0$ and therefore $m_T \mid q$ (Theorem 9.5.1). If some irreducible p had $p^2 \mid m_T$, then $p^2 \mid q$, which contradicts the uniqueness of the factorization of q into irreducibles (Theorem 6.3.5), every exponent there being 1. So no irreducible square divides m_T, which is (a).

■

Two things follow at once. First, the operator case is nothing but the case $A = F[T]$ of Theorem 23.3.3, so Chapter 10's vocabulary was this chapter's all along. Second, over an algebraically closed field every irreducible polynomial is linear — a monic polynomial of degree at least 2 has a root λ, hence the nonconstant factor $x - \lambda$ (Theorem 6.1.3) — so "semisimple" there means "a product of distinct **linear** factors", which is diagonalizability (Theorem 9.8.2): over such a field, $F[T]$ acts semisimply on V exactly when T is diagonalizable. The shear $\mathbf{J}_2(0)$ fails on both counts, as it must. Chapter 10 §08 made a finer statement in the same direction, Proposition 10.8.2: in characteristic zero, semisimple means diagonalizable after the field has been enlarged enough to split m_T. That proposition needs its characteristic hypothesis; Theorem 23.3.7 does not, because it never leaves F.

23.3.6 Exercises

23.3.6.1 A. Check your understanding

Exercise. A1

(a) State what it means for an algebra to be **simple** and what it means for an A-space to be **irreducible**, and give one object that is the first but whose regular representation is not the second.

(b) True or false: if A acts semisimply on V, then every A-invariant subspace of V is irreducible. Justify your answer.

(c) Where exactly does the proof of Proposition 23.3.4 use the requirement $\rho(1_A) = \mathrm{id}_V$ built into Definition 23.3.2? And why is the step $a = a1_A$ in the proof of Proposition 23.3.6 **not** an instance of that requirement?

(d) Name $\operatorname{rad} A$ for the algebra A of upper triangular 2×2 matrices, and say which two

results identify it.

Solution.

(a) A is simple if $A \neq \{0\}$ and its only two-sided ideals are $\{0\}$ and A (Definition 23.3.1); an A-space V is irreducible if $V \neq \{0\}$ and its only A-invariant subspaces are $\{0\}$ and V (Definition 23.3.3). For $n \geq 2$ the algebra $M_n(F)$ is simple, but its regular representation is not irreducible, since the first column ideal C_1 of Example (Matrix Algebras Are Semisimple) is a proper nonzero left ideal.

(b) False. Take $A = F$ acting on $V = F^2$ by scalars. Every subspace is invariant and every subspace has a complement, so A acts semisimply; but V itself is an invariant subspace of dimension 2, and it is not irreducible.

(c) In the final paragraph, at the words "Since $1_A v = v$": that is the requirement $\rho(1_A) = \mathrm{id}_V$ and nothing else, and without it v need not lie in Av, so V need not be the sum of the subspaces Av. The step $a = a1_A$ in Proposition 23.3.6 is a different statement: it is axiom (A3) of Definition 14.12.1, which makes 1_A a two-sided identity of A itself, and it holds with no representation in sight.

(d) $\operatorname{rad} A = \mathcal{N}$, the strictly upper triangular matrices: the inclusion $\supseteq$ is Lemma 23.3.5 applied to the nilpotent ideal $\mathcal{N}$, and $\subseteq$ comes from the two one-dimensional representations $\mathbf{X} \mapsto x_{11}$ and $\mathbf{X} \mapsto x_{22}$.

23.3.6.2 B. Practice

Exercise. B1: Which actions split?

Determine which of the following subalgebras of $M_2(F)$, acting on $V = F^2$, act semisimply. Justify your answer in each case, and when the action is semisimple exhibit V as a direct sum of irreducibles.

(a) the scalar matrices $\{c\mathbf{I}_2 : c \in F\}$;

(b) the diagonal matrices;

(c) $F[\mathbf{N}] = \{a\mathbf{I}_2 + b\mathbf{N} : a, b \in F\}$, where $\mathbf{N} = \mathbf{J}_2(0)$;

(d) $M_2(F)$ itself.

Solution.

(a) Semisimple. Every subspace is invariant, and every subspace of a finite-dimensional space has a complement (Theorem 1.7.6). $V = F\mathbf{e}_1 \oplus F\mathbf{e}_2$, and each line is irreducible because it has dimension 1.

(b) Semisimple. By the Quick check above, the invariant subspaces are $\{\mathbf{0}\}$, $F\mathbf{e}_1$, $F\mathbf{e}_2$, F^2; the two lines are complements of each other and the two extreme cases are complements of each other. Again $V = F\mathbf{e}_1 \oplus F\mathbf{e}_2$ with both summands irreducible.

(c) Not semisimple. The invariant subspaces are the $\mathbf{N}$-invariant ones, and the only invariant line is $F\mathbf{e}_1$: if $\mathbf{v} = (x, y)$ with $y \neq 0$ then $\mathbf{N}\mathbf{v} = (y, 0)$ is a nonzero element of $F\mathbf{e}_1$, which is not a multiple of $\mathbf{v}$. So $F\mathbf{e}_1$ has no invariant complement. (Alternatively: $m_{\mathbf{N}} = x^2$ is not squarefree, so Theorem 23.3.7 applies.)

(d) Semisimple, trivially: V is itself irreducible, as shown after Definition 23.3.3, so V is a one-term direct sum of irreducibles.

Exercise. B2: A two-dimensional algebra

Let $\mathbf{N} = \mathbf{J}_2(0) \in M_2(F)$ and $A = \{a\mathbf{I}_2 + b\mathbf{N} : a, b \in F\}$.

(a) Prove that A is a commutative subalgebra of $M_2(F)$ of dimension 2.

(b) Prove that $F\mathbf{N}$ is a two-sided ideal of A and compute $\operatorname{rad} A$.

(c) Is A a semisimple algebra? Justify your answer.

Solution.

(a) $A = \operatorname{span}(\mathbf{I}_2, \mathbf{N})$ is a subspace, and $\mathbf{I}_2, \mathbf{N}$ are independent since $\mathbf{N}$ is not a scalar. It contains $\mathbf{I}_2$, and

$$(a\mathbf{I}_2 + b\mathbf{N})(a'\mathbf{I}_2 + b'\mathbf{N}) = aa'\mathbf{I}_2 + (ab' + a'b)\mathbf{N},$$

using $\mathbf{N}^2 = \mathbf{O}$; the right-hand side lies in A and is symmetric in the two factors, so A is a commutative subalgebra of dimension 2.

(b) The displayed product shows $(a\mathbf{I}_2 + b\mathbf{N})(c\mathbf{N}) = ac\mathbf{N} \in F\mathbf{N}$, and A is commutative, so $F\mathbf{N}$ is a two-sided ideal; $(F\mathbf{N})^2 = \{\mathbf{O}\}$ since $\mathbf{N}^2 = \mathbf{O}$. By Lemma 23.3.5, $F\mathbf{N} \subseteq \operatorname{rad} A$. Conversely $\varphi(a\mathbf{I}_2 + b\mathbf{N}) = a$ is an algebra homomorphism $A \to F$ by the same display, and $\varphi(\mathbf{I}_2) = 1$, so F is an irreducible A-space; an element of $\operatorname{rad} A$ must have $a = 0$. Hence $\operatorname{rad} A = F\mathbf{N}$.

(c) No: $\operatorname{rad} A \neq \{\mathbf{O}\}$, so Proposition 23.3.6 rules it out.

Exercise. B3: Semisimple operators

For each of the following, decide whether $F[T]$ acts semisimply on V, using Theorem 23.3.7.

(a) $F = \mathbb{R}$, $V = \mathbb{R}^2$, T the rotation matrix $\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$;

(b) $F = \mathbb{Q}$, $V = \mathbb{Q}^3$, $T = \operatorname{diag}(1, 1, 2)$;

(c) $F = \mathbb{Q}$, $V = \mathbb{Q}^3$, $T = \mathbf{J}_3(0)$;

(d) $F = \mathbb{R}$, $V = \mathbb{R}^4$, $T = \mathbf{C}((x^2 + 1)^2)$.

Solution.

(a) Yes. $m_T = x^2 + 1$, irreducible over $\mathbb{R}$ and occurring once, so T is semisimple. Note that T is **not** diagonalizable over $\mathbb{R}$; semisimplicity is the weaker condition.

(b) Yes. $m_T = (x - 1)(x - 2)$, two distinct irreducibles.

(c) No. $m_T = x^3$ repeats the irreducible x.

(d) No. $m_T = (x^2 + 1)^2$ by Theorem 10.5.3, and $x^2 + 1$ is repeated.

23.3.6.3 C. Going deeper

Exercise. C1: The diagonal algebra

Let $A \subseteq M_n(F)$ be the algebra of all diagonal matrices, acting on $V = F^n$.

(a) Prove that every A-invariant subspace of V is the span of a subset of $\{\mathbf{e}_1, \dots, \mathbf{e}_n\}$. *Hint: consider* $\mathbf{E}_{ii}\mathbf{v}$.

(b) Deduce that A acts semisimply on V and identify the irreducible A-subspaces.

Solution.

(a) Let U be A-invariant and put $S = \{i : u_i \neq 0 \text{ for some } \mathbf{u} \in U\}$. For $\mathbf{u} \in U$ and each i, the matrix $\mathbf{E}_{ii}$ is diagonal, so $\mathbf{E}_{ii}\mathbf{u} = u_i\mathbf{e}_i \in U$; if $i \in S$, choosing $\mathbf{u}$ with $u_i \neq 0$ gives $\mathbf{e}_i \in U$. Hence $\operatorname{span}\{\mathbf{e}_i : i \in S\} \subseteq U$. Conversely every $\mathbf{u} \in U$ is $\sum_{i \in S} u_i\mathbf{e}_i$, since $u_i = 0$ for $i \notin S$ by the definition of S. So $U = \operatorname{span}\{\mathbf{e}_i : i \in S\}$.

(b) By (a) the invariant subspaces are indexed by subsets $S \subseteq \{1, \dots, n\}$, and the complement of $\operatorname{span}\{\mathbf{e}_i : i \in S\}$ inside V given by the complementary subset is again invariant. So A acts semisimply, and $V = F\mathbf{e}_1 \oplus \cdots \oplus F\mathbf{e}_n$. The irreducible A-subspaces are exactly the coordinate lines $F\mathbf{e}_i$: they are irreducible because they have dimension 1, and by (a) no other subspace is both invariant and free of a proper nonzero invariant subspace.

Exercise. C2: The radical is never everything

Let $A \neq \{0\}$ be a finite-dimensional F-algebra.

(a) Prove that A has at least one irreducible A-space. *Hint: among the nonzero left ideals, take one of smallest dimension.*

(b) Prove that every irreducible A-space has dimension at most $\dim A$.

(c) Deduce that $1_A \notin \operatorname{rad} A$, so $\operatorname{rad} A \neq A$.

Solution.

(a) A itself is a nonzero left ideal, so among the nonzero left ideals there is one, L, of smallest dimension. A nonzero left ideal contained in L is a nonzero left ideal of A, hence has dimension at least $\dim L$, hence equals L by Theorem 1.6.8. Since the A-invariant subspaces of the regular representation are the left ideals, L is an irreducible A-space.

(b) Let V be irreducible and $v \in V$ nonzero. The set Av is an A-invariant subspace containing $v = 1_A v$, so $Av = V$. It is the image of the linear map $A \to V$, $a \mapsto av$, so $\dim V \leq \dim A$ by Theorem 2.3.1.

(c) Let L be as in (a). Since $L \neq \{0\}$, pick $a \in L$ with $a \neq 0$; then $1_A a = a \neq 0$, so 1_A does not act as 0 on L and therefore $1_A \notin \operatorname{rad} A$. Hence $\operatorname{rad} A \neq A$. (Note the contrast with Proposition 23.3.6: the radical is always a proper ideal, but it is $\{0\}$ only in the semisimple case.)

23.4 Schur's Lemma

Section 3 cut a space into irreducible pieces. The next question is the one that always follows a decomposition: what are the maps between the pieces? The answer is startling in its poverty — there are almost none, and the ones that exist are invertible — and that poverty is exactly what makes the rest of the chapter work. A one-line lemma will replace an unknown operator by a scalar, over and over.

The lemma has two parts, and they do **not** have the same hypotheses. The first holds over every field. The second holds only when the field is algebraically closed, and the section ends by exhibiting the failure over $\mathbb{R}$ in full. Along the way we get the explanation of why commuting operators share eigenvectors, a fact Chapter 9 proved without explaining.

Recall that a field F is **algebraically closed** if every nonconstant polynomial in $F[x]$ has a root in F. The Fundamental Theorem of Algebra (Theorem 6.5.1) says $\mathbb{C}$ is one; $\mathbb{R}$ and $\mathbb{Q}$ are not, as x^2+1 and x^2-2 show. Unless a statement says otherwise, F is an arbitrary field, A is an F-algebra, and A-spaces are finite-dimensional, as in Definition 23.3.2.

23.4.1 Maps that respect the action

We have objects: A-spaces. We have subobjects: invariant subspaces. What is missing is the maps, and the requirement almost writes itself. Chapter 9 §09 got everything it had out of a single hypothesis, $ST = TS$; read S as a map and T as the action, and that hypothesis is the one below.

An A-map is a linear map for which acting and then mapping is the same as mapping and then acting.

Definition 23.4.1. A-Map, and Isomorphism of A-spaces

Let A be an F-algebra and let V and W be A-spaces. A linear map $\varphi\colon V \to W$ is an **A-map** if

$$\varphi(av) = a\,\varphi(v) \qquad \text{for every } a \in A \text{ and every } v \in V.$$

We write $\mathrm{End}_A(V)$ for the set of A-maps from V to itself. A **bijective** A-map is an **isomorphism of A-spaces**, and V and W are **isomorphic** if one exists.

In words: the two actions, on the source and on the target, are tied together by φ. Note that the letter a on the left acts on V and the one on the right acts on W; the condition is a statement about two different operators. The subscript in $\mathrm{End}_A(V)$ is part of the symbol and is never dropped: the book writes $\mathcal{L}(V)$ for **all** linear maps $V \to V$, and $\mathrm{End}_A(V) \subseteq \mathcal{L}(V)$ is the subset of those that respect the action.

Three routine facts, each one line. $\mathrm{End}_A(V)$ **is a subalgebra of $\mathcal{L}(V)$ containing** id_V**:** the defining condition is linear in φ, the identity satisfies it, and if φ, ψ satisfy it then $(\varphi\psi)(av) = \varphi(a\psi(v)) = a(\varphi\psi)(v)$. **The inverse of a bijective A-map is an A-map:** given $w \in W$ write $w = \varphi(v)$; then $\varphi^{-1}(aw) = \varphi^{-1}(\varphi(av)) = av = a\varphi^{-1}(w)$. **A composite of A-maps is an A-map**, by the same computation as the first.

Examples.

- **Scalars.** $\lambda\,\mathrm{id}_V \in \mathrm{End}_A(V)$ for every $\lambda \in F$, since the action is linear. These are the A-maps we get for free, and Schur's lemma will say that over an algebraically closed field

they are all of them.

- **The commutant, again.** If $A \subseteq \mathcal{L}(V)$ is a subalgebra containing id_V, acting tautologically, then φ is an A-map exactly when $\varphi T = T\varphi$ for every $T \in A$. So $\mathrm{End}_A(V) = A'$, the commutant of Section 1 (Definition 23.1.6, Proposition 23.1.4). The two notions are the same notion.
- **A projection onto a summand.** If $V = U \oplus W$ with both summands A-invariant, the projection of V onto U along W is an A-map: write $v = u + w$, and $av = au + aw$ is the corresponding splitting of av, so the projection of av is au.

Nonexample by minimal change. Take $A = F[\mathbf{N}] \subseteq \mathcal{L}(F^2)$ with $\mathbf{N} = \mathbf{J}_2(0)$, and $\varphi = \mathbf{E}_{21}$. Then $\varphi(\mathbf{N}\mathbf{e}_2) = \varphi(\mathbf{e}_1) = \mathbf{e}_2$ while $\mathbf{N}\varphi(\mathbf{e}_2) = \mathbf{N}\mathbf{0} = \mathbf{0}$. Linear, but not an A-map: the clause that fails is the displayed one, at $a = \mathbf{N}$ and $\mathbf{v} = \mathbf{e}_2$.

23.4.2 The lemma

Everything now follows from one observation: an A-map has an A-invariant kernel and an A-invariant image, and an irreducible A-space (Definition 23.3.3) has almost none of either.

Theorem 23.4.1. Schur's Lemma

Let A be an F-algebra.

(a) Let V and W be **irreducible** A-spaces and let $\varphi\colon V \to W$ be an A-map. Then either $\varphi = 0$ or φ is an isomorphism of A-spaces. **No hypothesis on F is needed.**

(b) Suppose in addition that F is **algebraically closed**, and let V be an irreducible A-space. Then every A-map $\varphi\colon V \to V$ is a scalar: $\varphi = \lambda\,\mathrm{id}_V$ for some $\lambda \in F$. Equivalently, $\mathrm{End}_A(V) = \{\lambda\,\mathrm{id}_V : \lambda \in F\}$.

Idea. For (a): the kernel and the image are invariant, so each is one of the two extremes, and a single word — "$\varphi \neq 0$" — rules out the wrong extreme on both sides. For (b): we want to compare φ with a scalar, so we need a scalar to compare it with, and the only source of one is an eigenvalue. That is the whole role of algebraic closure. Once λ is in hand, $\varphi - \lambda\,\mathrm{id}_V$ is an A-map that is **not** injective, and part (a) says a nonzero one would have to be.

Proof.

(a) The kernel is A-invariant: if $\varphi(v) = 0$ then $\varphi(av) = a\varphi(v) = 0$. The image is A-invariant: $a\varphi(v) = \varphi(av) \in \mathrm{im}\,\varphi$.
Suppose $\varphi \neq 0$. Then $\ker\varphi \neq V$, and since V is irreducible its only other invariant subspace is $\{0\}$, so $\ker\varphi = \{0\}$ and φ is injective (Theorem 2.2.3). Also $\mathrm{im}\,\varphi \neq \{0\}$, and since W is irreducible, $\mathrm{im}\,\varphi = W$, so φ is surjective. Hence φ is bijective, and its inverse is an A-map by the remark above; so φ is an isomorphism of A-spaces.

(b) Since V is irreducible, $V \neq \{0\}$, so $n = \dim V \geq 1$ and the characteristic polynomial p_φ is monic of degree n, hence nonconstant. As F is algebraically closed, p_φ has a root $\lambda \in F$, and λ is then an eigenvalue of φ (Theorem 9.2.1).

Put $\psi = \varphi - \lambda\,\mathrm{id}_V$. It is an A-map, because $\mathrm{End}_A(V)$ is a subspace containing id_V. Its kernel is $E_\lambda(\varphi) \neq \{0\}$, so ψ is not injective and therefore not an isomorphism. By (a), applied with $W = V$, the only remaining possibility is $\psi = 0$, that is, $\varphi = \lambda\,\mathrm{id}_V$. Conversely every scalar

multiple of id_V is an A-map, so $\mathrm{End}_A(V)$ consists exactly of these. This proves the lemma.

■

Part (a) says something worth naming even when the field is small: for an irreducible V, every nonzero element of $\mathrm{End}_A(V)$ is invertible. An algebra with that property is called a **division algebra**, and part (a) is the statement that the commutant of an irreducible action is one. Part (b) says that over an algebraically closed field the only division algebra that can occur is F itself. That is a genuine restriction on the field, not a formality.

Warning. **Only the second part needs algebraic closure.** Part (a) was proved over an arbitrary field and will be used over arbitrary fields. Part (b) used exactly one thing that (a) did not: the existence of an eigenvalue of φ. Quoting "Schur's lemma says the commutant is the scalars" without the hypothesis is the standard error, and the next example is the standard witness.

23.4.3 What goes wrong over the reals

Example. A Real Irreducible Representation with Commutant $\mathbb{C}$

Let $n \geq 3$, put $\theta = 2\pi/n$, $c = \cos\theta$, $s = \sin\theta$, and let

$$\mathbf{R} = \begin{pmatrix} c & -s \\ s & c \end{pmatrix} \in M_2(\mathbb{R}).$$

Since $\mathbf{R}^n = \mathbf{I}_2$, the powers of $\mathbf{R}$ repeat with period dividing n; Section 7 will read $k \mapsto \mathbf{R}^k$ as a representation of a cyclic group, and that is what we will then call the **rotation representation**. Let $A = \mathbb{R}[\mathbf{R}] \subseteq M_2(\mathbb{R})$ be the algebra of polynomials in $\mathbf{R}$, so that the A-invariant subspaces of $\mathbb{R}^2$ are exactly the $\mathbf{R}$-invariant ones. Show that $\mathbb{R}^2$ is an irreducible A-space and compute $\mathrm{End}_A(\mathbb{R}^2)$.

Solution. **Irreducible.** A proper nonzero invariant subspace of $\mathbb{R}^2$ would be a line $\mathbb{R}\mathbf{v}$ with $\mathbf{R}\mathbf{v} \in \mathbb{R}\mathbf{v}$, that is, a real eigenvector of $\mathbf{R}$. Now

$$p_{\mathbf{R}}(x) = x^2 - 2cx + 1,$$

whose discriminant is $4c^2 - 4 = -4s^2$. For $n \geq 3$ we have $0 < \theta \leq 2\pi/3 < \pi$, so $s = \sin\theta > 0$ and the discriminant is negative: $p_{\mathbf{R}}$ has no real root, hence $\mathbf{R}$ has no real eigenvalue (Theorem 9.2.1), hence no invariant line. So the only invariant subspaces are $\{\mathbf{0}\}$ and $\mathbb{R}^2$, and $\mathbb{R}^2 \neq \{\mathbf{0}\}$.

The commutant. Since A acts tautologically, $\mathrm{End}_A(\mathbb{R}^2)$ is the commutant A' (Definition 23.1.6): the set of $\mathbf{X} \in M_2(\mathbb{R})$ commuting with every element of A; since A consists of

polynomials in $\mathbf{R}$, this is the set of $\mathbf{X}$ with $\mathbf{XR} = \mathbf{RX}$. Write $\mathbf{X} = \left(\begin{smallmatrix} p & q \\ r & t \end{smallmatrix}\right)$. Multiplying out,

$$\mathbf{XR} = \begin{pmatrix} pc + qs & -ps + qc \\ rc + ts & -rs + tc \end{pmatrix},$$
$$\mathbf{RX} = \begin{pmatrix} cp - sr & cq - st \\ sp + cr & sq + ct \end{pmatrix},$$

so that

$$\mathbf{XR} - \mathbf{RX} = s \begin{pmatrix} q + r & t - p \\ t - p & -(q + r) \end{pmatrix}.$$

Since $s \neq 0$, the equation $\mathbf{XR} = \mathbf{RX}$ says exactly $r = -q$ and $t = p$. Writing $\mathbf{Q} = \left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$ for the quarter turn, we get

$$\mathrm{End}_A(\mathbb{R}^2) = \{\, p\mathbf{I}_2 + r\mathbf{Q} : p, r \in \mathbb{R} \,\},$$

a two-dimensional real algebra. It is a copy of $\mathbb{C}$: the map $p + ri \mapsto p\mathbf{I}_2 + r\mathbf{Q}$ is linear and bijective (it carries the basis $1, i$ to the independent $\mathbf{I}_2, \mathbf{Q}$), it sends 1 to $\mathbf{I}_2$, and since $\mathbf{Q}^2 = -\mathbf{I}_2$,

$$(p\mathbf{I}_2 + r\mathbf{Q})(p'\mathbf{I}_2 + r'\mathbf{Q}) = (pp' - rr')\mathbf{I}_2 + (pr' + rp')\mathbf{Q},$$

which matches $(p + ri)(p' + r'i) = (pp' - rr') + (pr' + rp')i$. So it is an isomorphism of $\mathbb{R}$-algebras.

What this shows. Part (b) of Theorem 23.4.1 fails over $\mathbb{R}$: here $\mathrm{End}_A(\mathbb{R}^2)$ has dimension 2, not 1. Part (a) does not fail, and we can see it directly: $\det(p\mathbf{I}_2 + r\mathbf{Q}) = p^2 + r^2$, which is nonzero unless $p = r = 0$, so every nonzero element of the commutant is invertible, exactly as (a) predicts. The commutant is a division algebra; it is just not $\mathbb{R}$.

Remark. The same computation with $n = 4$ gives $\mathbf{R} = \mathbf{Q}$, and the commutant $\mathbb{R}[\mathbf{Q}]$ is then the algebra generated by $\mathbf{R}$ itself. Over $\mathbb{C}$ the picture collapses: $\mathbf{R}$ has eigenvalues $e^{\pm i\theta}$ there, so the two eigenlines are invariant and $\mathbb{C}^2$ is **not** an irreducible $\mathbb{C}[\mathbf{R}]$-space. Irreducibility, like semisimplicity, is a statement about a space **and** a field.

23.4.4 Families of operators

The corollaries below are about a family of operators — a set of matrices, the image of a group — rather than about an algebra given in advance. The translation costs one lemma, and Section 1's generated subalgebra is what it produces.

Lemma 23.4.2. A Family and the Algebra It Generates

Let V be a finite-dimensional vector space over F and let $\mathcal{F} \subseteq \mathcal{L}(V)$ be any set of operators. Write $A(\mathcal{F}) := F\langle\mathcal{F}\rangle$ for the subalgebra of $\mathcal{L}(V)$ generated by $\mathcal{F}$ (Definition 23.1.4). Then:

(a) $A(\mathcal{F})$ contains $\mathcal{F}$ and id_V, and is the span of id_V together with all products $S_1 S_2 \cdots S_m$ with $m \geq 1$ and $S_1, \ldots, S_m \in \mathcal{F}$;

(b) a subspace $U \subseteq V$ is invariant under every member of $\mathcal{F}$ if and only if it is $A(\mathcal{F})$-invariant;

(c) an operator $S \in \mathcal{L}(V)$ commutes with every member of $\mathcal{F}$ if and only if it commutes with every element of $A(\mathcal{F})$;

(d) if the members of $\mathcal{F}$ commute with one another, then $A(\mathcal{F})$ is commutative.

Proof.

(a) A generated subalgebra contains its generators, and every subalgebra contains the identity (Definition 23.1.1), which in $\mathcal{L}(V)$ is id_V. The span description is Proposition 23.1.1 applied to the set $\mathcal{F}$ inside the algebra $\mathcal{L}(V)$.

(b) ($\Leftarrow$) holds because $\mathcal{F} \subseteq A(\mathcal{F})$. ($\Rightarrow$) If $SU \subseteq U$ for every $S \in \mathcal{F}$, then applying the members one at a time gives $S_1 \cdots S_m U \subseteq U$ for every word, and $\mathrm{id}_V\, U \subseteq U$; a subspace is closed under linear combinations, so $A(\mathcal{F})U \subseteq U$.

(c) ($\Leftarrow$) again holds because $\mathcal{F} \subseteq A(\mathcal{F})$. ($\Rightarrow$) If S commutes with each S_i then it commutes with $S_1 \cdots S_m$, by moving S across one factor at a time; it commutes with id_V; and the set of operators commuting with S is a subspace, so it contains the span.

(d) Let $S \in \mathcal{F}$. By hypothesis S commutes with every member of $\mathcal{F}$, so by (c) it commutes with every element of $A(\mathcal{F})$. Now fix $a \in A(\mathcal{F})$. What we have just shown says that every member of $\mathcal{F}$ commutes with a; applying (c) to a in place of S, the element a commutes with every element of $A(\mathcal{F})$. As a was arbitrary, $A(\mathcal{F})$ is commutative.

■

Say that a family $\mathcal{F} \subseteq \mathcal{L}(V)$ **acts irreducibly** if $V \neq \{0\}$ and the only subspaces invariant under every member of $\mathcal{F}$ are $\{0\}$ and V. By (a) and (b), this says exactly that V is an irreducible $A(\mathcal{F})$-space.

Corollary 23.4.3. An Operator Commuting with an Irreducible Family Is a Scalar

Let F be **algebraically closed**, let V be a finite-dimensional vector space over F, and let $\mathcal{F} \subseteq \mathcal{L}(V)$ act irreducibly. If $S \in \mathcal{L}(V)$ commutes with every member of $\mathcal{F}$, then $S = \lambda\, \mathrm{id}_V$ for some $\lambda \in F$.

Proof. Put $A = A(\mathcal{F})$, a subalgebra of $\mathcal{L}(V)$ containing id_V by Lemma 23.4.2 (a), so V is an A-space under the tautological representation, and by (b) it is irreducible. By (c), S commutes with every element of A, that is, $S \in \mathrm{End}_A(V)$. Since F is algebraically closed, Theorem 23.4.1 (b) gives $S = \lambda\, \mathrm{id}_V$.

■

Corollary 23.4.4. Over an Algebraically Closed Field, Commutative Means One-Dimensional

Let F be **algebraically closed**, let A be a **commutative** F-algebra, and let V be an irreducible A-space. Then $\dim V = 1$.

Idea. Commutativity makes every operator of the action an A-map, so Schur turns all of them into scalars at once. But an algebra of scalars leaves every subspace invariant, and

only a line can survive that.

Proof. Let $a \in A$. For every $b \in A$ we have $\rho(a)\rho(b) = \rho(ab) = \rho(ba) = \rho(b)\rho(a)$, so $\rho(a)$ is an A-map from V to itself. By Theorem 23.4.1 (b) there is $\lambda(a) \in F$ with $\rho(a) = \lambda(a)\,\mathrm{id}_V$.

Hence $av = \lambda(a)v$ for every $a \in A$ and $v \in V$, so **every** subspace of V is A-invariant. Since V is irreducible, $V \neq \{0\}$; pick $v \neq 0$. Then Fv is a nonzero A-invariant subspace, so $Fv = V$ and $\dim V = 1$.

■

Both hypotheses are doing work. Commutativity is needed: $\mathbb{C}^2$ is an irreducible $M_2(\mathbb{C})$-space of dimension 2. Algebraic closure is needed: in Example (A Real Irreducible Representation with Commutant $\mathbb{C}$) the algebra $A = \mathbb{R}[\mathbf{R}]$ is commutative, being generated by one operator, and $\mathbb{R}^2$ is an irreducible A-space of dimension 2.

Quick check. The algebra $A = \mathbb{R}[\mathbf{R}]$ of Example (A Real Irreducible Representation with Commutant $\mathbb{C}$) is commutative and $\mathbb{R}^2$ is an irreducible A-space of dimension 2. Which hypothesis of Corollary 23.4.4 fails, and where exactly does the proof break?

Solution. Algebraic closure fails: $\mathbb{R}$ is not algebraically closed. The proof breaks at its first move, the appeal to Theorem 23.4.1 (b): the operator $\rho(\mathbf{R}) = \mathbf{R}$ is indeed an A-map, but it is not a scalar, because its characteristic polynomial $x^2 - 2cx + 1$ has no root in $\mathbb{R}$ and so $\mathbf{R}$ has no eigenvalue to be a scalar multiple of. Everything before that point — commutativity, irreducibility, the fact that each $\rho(a)$ is an A-map — is still true.

Section 9 will use Corollary 23.4.4 in exactly the form stated: over $\mathbb{C}$, every irreducible representation of a commutative algebra, and hence of an abelian group, is one-dimensional, so its "matrix" is a single complex number.

23.4.5 Looking back at commuting operators

Chapter 9 §09 proved that a commuting family of operators, each with a splitting characteristic polynomial, has a common eigenvector (Theorem 9.9.3), and built simultaneous diagonalization on it. At the time the proof looked like a clever trick: take a minimal invariant subspace and show every operator is a scalar on it. We can now say what the trick was.

Corollary 23.4.5. Commuting Operators Have a Common Eigenvector

Let F be **algebraically closed**, let $V \neq \{0\}$ be finite-dimensional over F, and let $\mathcal{F} \subseteq \mathcal{L}(V)$ be a family whose members commute with one another. Then there is $v \neq 0$ that is an eigenvector of every $T \in \mathcal{F}$.

Proof. Among the nonzero subspaces of V invariant under every member of $\mathcal{F}$ — there is one, namely V — choose W of smallest dimension. Put $A = A(\mathcal{F})$, which is a commutative

subalgebra of $\mathcal{L}(V)$ containing id_V by Lemma 23.4.2 (a), (d), and note that W is A-invariant by part (b) of the same lemma. So $a \mapsto a|_W$ is a representation of A on W.

W is an irreducible A-space: it is nonzero, and a nonzero A-invariant subspace $U \subseteq W$ is, by Lemma 23.4.2 (b), invariant under every member of $\mathcal{F}$, hence has $\dim U \geq \dim W$ by the choice of W, hence equals W (Theorem 1.6.8).

By Corollary 23.4.4, $\dim W = 1$. Write $W = Fv$ with $v \neq 0$. For every $T \in \mathcal{F}$ we have $Tv \in W = Fv$, so v is an eigenvector of T. This proves the corollary.

■

That is Chapter 9's argument, with its steps given their names: "minimal nonzero invariant subspace" is "irreducible", and "every operator is a scalar on it" is Schur's lemma. The version proved there is the stronger one — it asks only that each p_T split over F, not that F be algebraically closed — and getting that extra generality is exactly what its Lemma 9.9.2 was for. What this section adds is not strength but explanation: the scalars appeared because a commutative algebra acting irreducibly over an algebraically closed field has nowhere to act except on a line.

The other half of Chapter 9 §09, simultaneous diagonalization, needs one more ingredient: that the whole space splits into such lines, not just that one of them exists. Section 3's language names that ingredient — the algebra must act semisimply — and in the presence of an inner product it comes for free.

Proposition 23.4.6. Closed Under Adjoints Implies Semisimple

Optional: nothing later depends on this.

Let V be a finite-dimensional inner product space over $F = \mathbb{R}$ or $F = \mathbb{C}$, and let $A \subseteq \mathcal{L}(V)$ be a subalgebra containing id_V and closed under adjoints: $T \in A$ implies $T^* \in A$ (Definition 11.6.1). Then:

(a) for every A-invariant subspace U, the orthogonal complement $U^\perp$ is A-invariant, and $V = U \oplus U^\perp$; in particular A acts semisimply on V;

(b) V is a direct sum of **pairwise orthogonal** irreducible A-invariant subspaces.

Proof.

(a) Let U be A-invariant, let $T \in A$ and $w \in U^\perp$. For every $u \in U$,

$$\langle Tw, u\rangle = \langle w, T^*u\rangle = 0,$$

because $T^* \in A$ gives $T^*u \in U$, while $w \in U^\perp$. So $Tw \in U^\perp$, and $U^\perp$ is A-invariant. The splitting $V = U \oplus U^\perp$ is Theorem 11.3.2 (a). Hence every A-invariant subspace has an A-invariant complement (Definition 23.3.4).

(b) Induction on $\dim V$. If $V = \{0\}$ it is the empty direct sum. Otherwise choose, among the nonzero A-invariant subspaces, one of smallest dimension, W; as in the proof of Theorem 23.3.3 it is irreducible. By (a), $V = W \oplus W^\perp$ with $W^\perp$ invariant and $\dim W^\perp < \dim V$.

The restricted algebra $A|_{W^\perp} = \{T|_{W^\perp} : T \in A\} \subseteq \mathcal{L}(W^\perp)$ satisfies the same hypotheses on

the inner product space $W^\perp$: it is a subalgebra containing $\mathrm{id}_{W^\perp}$, and for $w, w' \in W^\perp$,

$$\langle T|_{W^\perp} w, w' \rangle = \langle Tw, w' \rangle = \langle w, T^* w' \rangle = \langle w, T^*|_{W^\perp} w' \rangle,$$

so $(T|_{W^\perp})^* = T^*|_{W^\perp} \in A|_{W^\perp}$ by uniqueness of the adjoint (Theorem 11.6.1). By the inductive hypothesis $W^\perp = W_1 \oplus \cdots \oplus W_r$ with the W_i pairwise orthogonal and irreducible, and each $W_i \subseteq W^\perp$ is orthogonal to W. Hence $V = W \oplus W_1 \oplus \cdots \oplus W_r$ as required.

■

Now take a commuting family $\mathcal{F}$ of **self-adjoint** operators on a finite-dimensional complex inner product space $V \neq \{0\}$, and put $A = A(\mathcal{F})$. It is commutative by Lemma 23.4.2 (d). It is closed under adjoints, because $(S_1 \cdots S_m)^* = S_m^* \cdots S_1^* = S_m \cdots S_1 = S_1 \cdots S_m$, the last step by commutativity, and because $\varphi \mapsto \varphi^*$ is conjugate-linear, so the span of a set of self-adjoint operators is closed under it. By Proposition 23.4.6, V is an orthogonal direct sum of irreducible A-subspaces; by Corollary 23.4.4 those subspaces are lines; and a line invariant under every member of $\mathcal{F}$ is spanned by a common eigenvector. Normalizing gives an orthonormal basis of common eigenvectors — Chapter 12 §07's conclusion (Theorem 12.7.1) for a commuting self-adjoint family, obtained here by putting two general facts together rather than by an induction tailored to the situation. Chapter 12 proves more, for commuting **normal** families, and pays for it with a longer argument.

Warning. **The converse of Corollary 23.4.3 is false.** A family whose commutant is only the scalars need not act irreducibly, and the witness is short. Take $F = \mathbb{C}$ and let $\mathcal{F}$ be the algebra of upper triangular matrices in $M_2(\mathbb{C})$, acting on $\mathbb{C}^2$. A matrix commuting with $\mathbf{E}_{11}$ must be diagonal, and one commuting with $\mathbf{E}_{12}$ as well must have equal diagonal entries, so the commutant is $\{p\mathbf{I}_2 : p \in \mathbb{C}\}$, the scalars; exercise C2 does the two computations. Yet $\mathbb{C}\mathbf{e}_1$ is invariant under every upper triangular matrix, so $\mathcal{F}$ does not act irreducibly. Thus "the commutant is the scalars" is a consequence of irreducibility, never a test for it. The genuine converse statement, that an irreducible subalgebra is **all** of $\mathcal{L}(V)$, is Burnside's theorem in Section 5, and it needs more than the commutant.

23.4.6 Exercises

23.4.6.1 A. Check your understanding

Exercise. A1

(a) State both parts of Schur's lemma and say precisely which one needs F to be algebraically closed, and what that hypothesis is used for.

(b) True or false: if V and W are irreducible A-spaces with $\dim V \neq \dim W$, then the only A-map $V \to W$ is 0. Justify your answer.

(c) Explain why, for an irreducible A-space V, no element of $\mathrm{End}_A(V)$ can be nonzero and noninvertible.

(d) Give $\mathrm{End}_A(\mathbb{R}^2)$ for the rotation representation of Example (A Real Irreducible Representation with Commutant $\mathbb{C}$), and say which part of Schur's lemma it illustrates and which it refutes.

Solution.

(a) Part (a): an A-map between irreducible A-spaces is 0 or an isomorphism; no hypothesis on F. Part (b): over an algebraically closed F, every A-map from an irreducible V to itself is a scalar. Algebraic closure is used once, to produce a root of the characteristic polynomial of φ, that is, an eigenvalue.

(b) True. A nonzero A-map would be an isomorphism by Theorem 23.4.1 (a), and an isomorphism of vector spaces preserves dimension (Theorem 2.5.4), contradicting $\dim V \neq \dim W$.

(c) By Theorem 23.4.1 (a) with $W = V$: a nonzero A-map $V \to V$ is an isomorphism, hence invertible.

(d) $\mathrm{End}_A(\mathbb{R}^2) = \{p\mathbf{I}_2 + r\mathbf{Q}\} \cong \mathbb{C}$. It illustrates part (a) — every nonzero element has determinant $p^2 + r^2 \neq 0$, so the commutant is a division algebra — and refutes part (b) read without its hypothesis, since the commutant is two-dimensional over $\mathbb{R}$.

23.4.6.2 B. Practice

Exercise. B1: An irreducibility test

Let $\mathbf{D} = \mathrm{diag}(i, -i) \in M_2(\mathbb{C})$, which satisfies $\mathbf{D}^4 = \mathbf{I}_2$, and let $A = \mathbb{C}[\mathbf{D}]$ be the algebra of polynomials in $\mathbf{D}$.

(a) Compute $\mathrm{End}_A(\mathbb{C}^2)$.

(b) Deduce from Theorem 23.4.1 (b) that $\mathbb{C}^2$ is not an irreducible A-space, and confirm this by exhibiting a proper nonzero invariant subspace.

Solution.

(a) An operator is an A-map exactly when it commutes with $\mathbf{D}$, since the elements of A are the polynomials in $\mathbf{D}$. Writing $\mathbf{X} = \left(\begin{smallmatrix} p & q \\ r & t \end{smallmatrix}\right)$,

$$\mathbf{XD} - \mathbf{DX} = \begin{pmatrix} 0 & -2iq \\ 2ir & 0 \end{pmatrix},$$

which vanishes exactly when $q = r = 0$. So $\mathrm{End}_A(\mathbb{C}^2)$ is the algebra of diagonal matrices, of dimension 2.

(b) If $\mathbb{C}^2$ were irreducible then, $\mathbb{C}$ being algebraically closed, Theorem 23.4.1 (b) would force $\mathrm{End}_A(\mathbb{C}^2)$ to have dimension 1. It has dimension 2, so $\mathbb{C}^2$ is not irreducible. Indeed $\mathbb{C}\mathbf{e}_1$ is invariant, since $\mathbf{D}\mathbf{e}_1 = i\mathbf{e}_1$.

Exercise. B2: The quarter turn, over two fields

Let $\mathbf{Q} = \left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$, the rotation by a quarter turn, and let $A_F = F[\mathbf{Q}]$ for $F = \mathbb{R}$ and for $F = \mathbb{C}$.

(a) Show that $\mathbb{R}^2$ is an irreducible $A_{\mathbb{R}}$-space and that $\mathrm{End}_{A_{\mathbb{R}}}(\mathbb{R}^2) \cong \mathbb{C}$.

(b) Show that $\mathbb{C}^2$ is **not** an irreducible $A_{\mathbb{C}}$-space, and find two invariant lines.

(c) Explain why (a) and (b) together are consistent with Schur's lemma.

Solution.

(a) This is Example (A Real Irreducible Representation with Commutant $\mathbb{C}$) with $n = 4$, so $\theta = \pi/2$, $c = 0$, $s = 1$: the characteristic polynomial x^2+1 has no real root, so there is no invariant line and $\mathbb{R}^2$ is irreducible; and the commutant is $\{p\mathbf{I}_2 + r\mathbf{Q}\} \cong \mathbb{C}$.

(b) Over $\mathbb{C}$ the polynomial $x^2 + 1$ has the roots $\pm i$, so $\mathbf{Q}$ has eigenvalues i and $-i$, with eigenvectors $(1, -i)$ and $(1, i)$ respectively: from $\mathbf{Q}(1, -i) = (i, 1) = i(1, -i)$ and $\mathbf{Q}(1, i) = (-i, 1) = -i(1, i)$. Each of $\mathbb{C}(1, i)$ and $\mathbb{C}(1, -i)$ is a proper nonzero invariant line.

(c) Part (b) of Schur's lemma applies only over an algebraically closed field. Over $\mathbb{C}$ it does apply, and it is not contradicted: the space is reducible there, so the lemma says nothing about it. Over $\mathbb{R}$ the space is irreducible but the field is not algebraically closed, so only part (a) applies — and part (a) holds, since every nonzero element of $\{p\mathbf{I}_2 + r\mathbf{Q}\}$ is invertible.

Exercise. B3: A common eigenvector

Let

$$\mathbf{A} = \begin{pmatrix} 2 & 1 \\ 0 & 2 \end{pmatrix}, \qquad \mathbf{B} = \begin{pmatrix} 3 & 5 \\ 0 & 3 \end{pmatrix} \in M_2(\mathbb{C}).$$

(a) Verify that $\mathbf{AB} = \mathbf{BA}$.

(b) Find a common eigenvector, and check that it spans the unique minimal nonzero invariant subspace, as in the proof of Corollary 23.4.5.

Solution.

(a) The diagonal entries of both products are $2 \cdot 3 = 6$, the lower left entry of both is 0, and the upper right entries are $(\mathbf{AB})_{12} = 2 \cdot 5 + 1 \cdot 3 = 13$ and $(\mathbf{BA})_{12} = 3 \cdot 1 + 5 \cdot 2 = 13$. So

$$\mathbf{AB} = \mathbf{BA} = \begin{pmatrix} 6 & 13 \\ 0 & 6 \end{pmatrix}.$$

(b) $\mathbf{Ae}_1 = 2\mathbf{e}_1$ and $\mathbf{Be}_1 = 3\mathbf{e}_1$, so $\mathbf{e}_1$ is a common eigenvector and $\mathbb{C}\mathbf{e}_1$ is invariant. It is the only invariant line: if $\mathbf{v} = (x, y)$ with $y \neq 0$ spanned an invariant line, then $\mathbf{Av} = (2x + y, 2y)$ would be a multiple $c\mathbf{v}$, forcing $c = 2$ from the second coordinate and then $y = 0$ from the first. So $\mathbb{C}\mathbf{e}_1$ is the minimal nonzero invariant subspace the proof of Corollary 23.4.5 produces, and it is a line, as Corollary 23.4.4 requires.

23.4.6.3 C. Going deeper

Exercise. C1: A commutant that is a field

Let $\mathbf{A} = \left(\begin{smallmatrix} 0 & 2 \\ 1 & 0 \end{smallmatrix}\right) \in M_2(\mathbb{Q})$, the companion matrix of $x^2 - 2$, and let $A = \mathbb{Q}[\mathbf{A}]$.

(a) Prove that $\mathbb{Q}^2$ is an irreducible A-space.

(b) Compute $\mathrm{End}_A(\mathbb{Q}^2)$ and prove that it is a field in which 2 has a square root.

(c) Explain why this does not contradict Theorem 23.4.1.

Solution.

(a) $p_{\mathbf{A}} = x^2 - 2$ by Theorem 10.5.3, and $x^2 - 2$ has no rational root, since $\sqrt{2}$ is irrational (Theorem 0.2.3). So $\mathbf{A}$ has no eigenvalue in $\mathbb{Q}$ (Theorem 9.2.1), hence no invariant line, and the only invariant subspaces of $\mathbb{Q}^2$ are $\{\mathbf{0}\}$ and $\mathbb{Q}^2$.

(b) An A-map is an $\mathbf{X} = \left(\begin{smallmatrix} p & q \\ r & t \end{smallmatrix}\right)$ commuting with $\mathbf{A}$. Computing,

$$\mathbf{XA} - \mathbf{AX} = \begin{pmatrix} q - 2r & 2p - 2t \\ t - p & 2r - q \end{pmatrix},$$

which vanishes exactly when $t = p$ and $q = 2r$. So

$$\mathrm{End}_A(\mathbb{Q}^2) = \{\, p\mathbf{I}_2 + r\mathbf{A} : p, r \in \mathbb{Q} \,\},$$

using $\mathbf{A} = \left(\begin{smallmatrix} 0 & 2 \\ 1 & 0 \end{smallmatrix}\right)$. Since $\mathbf{A}^2 = 2\mathbf{I}_2$, this is a commutative algebra in which $\mathbf{A}$ is a square root of $2\mathbf{I}_2$; and every nonzero element is invertible, because $\det(p\mathbf{I}_2 + r\mathbf{A}) = p^2 - 2r^2$, which vanishes for rational p, r only when $p = r = 0$: if $r = 0$ then $p^2 = 0$, and if $r \neq 0$ then $(p/r)^2 = 2$ with $p/r \in \mathbb{Q}$, contradicting Theorem 0.2.3. A commutative algebra in which every nonzero element is invertible is a field.

(c) Part (a) of Theorem 23.4.1 predicts exactly this: the commutant of an irreducible action is a division algebra. Part (b) does not apply, because $\mathbb{Q}$ is not algebraically closed.

Exercise. C2: Scalar commutant without irreducibility

Let F be **algebraically closed** — say $F = \mathbb{C}$, so that Corollary 23.4.3 really applies — and let $A \subseteq M_2(F)$ be the algebra of upper triangular matrices.

(a) Compute the set of $\mathbf{X} \in M_2(F)$ commuting with every element of A. (The computation uses nothing about F.)

(b) Deduce that the converse of Corollary 23.4.3 fails: a family with commutant the scalars need not act irreducibly.

Solution.

(a) It suffices to commute with the basis $\mathbf{E}_{11}, \mathbf{E}_{12}, \mathbf{E}_{22}$. Write $\mathbf{X} = \left(\begin{smallmatrix} p & q \\ r & t \end{smallmatrix}\right)$. From

$$\mathbf{XE}_{11} - \mathbf{E}_{11}\mathbf{X} = \begin{pmatrix} 0 & -q \\ r & 0 \end{pmatrix}$$

we get $q = r = 0$, and then

$$\mathbf{XE}_{12} - \mathbf{E}_{12}\mathbf{X} = \begin{pmatrix} 0 & p - t \\ 0 & 0 \end{pmatrix}$$

gives $p = t$. Conversely a scalar matrix commutes with everything. So the commutant is $\{p\mathbf{I}_2 : p \in F\}$, the scalars.

(b) The family $\mathcal{F} = A$ has commutant the scalars by (a), yet it does not act irreducibly on F^2: the line $F\mathbf{e}_1$ is invariant, as computed in Example (Upper Triangular Matrices Are Not Semisimple). So the implication in Corollary 23.4.3 cannot be reversed. (The hypothesis is not vacuous either: over an algebraically closed field the corollary does apply whenever the family is irreducible.)

23.5 Burnside and the Double Commutant

Twice already the book has computed a commutant: Chapter 10 §04 found the matrices commuting with a basic Weyr matrix, and Chapter 12 §10 found the matrices commuting with a block diagonal matrix whose blocks have separated spectra. Both times the question was *given one operator, who commutes with it?* This section turns the question round. Given a whole algebra of operators, how much of the algebra can be read off from its commutant, and when is the algebra as large as it could possibly be? Two theorems answer this. Burnside's theorem says that over an algebraically closed field an algebra with no invariant subspace to hide in must be everything; the double commutant theorem says that a semisimple algebra is exactly the set of operators commuting with everything that commutes with it.

Throughout, V is a finite-dimensional vector space over a field F, and $\mathrm{End}(V)$ is the space $\mathcal{L}(V)$ of operators regarded as an algebra under composition. A **subalgebra** $A \subseteq \mathrm{End}(V)$ contains id_V by (SA3) of Definition 23.1.1; this is not decoration, and the warning after Theorem 23.5.5 shows what goes wrong without it.

23.5.1 Commutants, and when an operator is a scalar

For a subalgebra $A \subseteq \mathrm{End}(V)$, Definition 23.1.6 sets

$$A' = \{t \in \mathrm{End}(V) : ta = at \text{ for all } a \in A\},$$

and Proposition 23.1.4 says A' is again a subalgebra containing id_V.

Applying the prime twice gives the **double commutant** $A'' := (A')'$. One inclusion is free and worth recording at once: by Proposition 23.1.4 (c),

$$A \subseteq A''.$$

The content of this section is that the reverse inclusion holds under a hypothesis we can name exactly.

Both theorems below end by proving that some operator is a scalar, and §01 has already supplied the fact that lets us conclude this. By Proposition 23.1.5 the center of $M_n(F)$ consists of the scalar matrices or, in the operator form recorded there, for a finite-dimensional $V \neq \{0\}$ the operators commuting with every operator on V are exactly the scalar multiples of id_V:

$$\mathrm{End}(V)' = \{\lambda\,\mathrm{id}_V : \lambda \in F\}.$$

Nothing new is proved here; it is the operator form that the proofs below call on.

23.5.2 Algebras that act irreducibly

Recall from §03 that V is **irreducible** as an A-space (§03 calls such a space simple, Definition 23.3.3) when $V \neq \{0\}$ and the only A-invariant subspaces of V are $\{0\}$ and V. The following one-line consequence is used in every proof of this section, so we isolate it.

Lemma 23.5.1. An Irreducible Algebra Moves One Vector Everywhere

Let $A \subseteq \mathrm{End}(V)$ be a subalgebra with $\mathrm{id}_V \in A$, acting irreducibly on V. Then $Av := \{av : a \in A\} = V$ for every $v \neq 0$.

Proof. The set Av is the image of the linear map $A \to V$, $a \mapsto av$, hence a subspace. It is A-invariant: $b(av) = (ba)v \in Av$ for $b \in A$, since $ba \in A$. It contains $\mathrm{id}_V v = v \neq 0$, so it is not $\{0\}$. Irreducibility leaves only $Av = V$. ∎

Warning. **The assumption $\mathrm{id}_V \in A$ is doing real work here.** Take $V = F$, a line, and $A = \{0\}$, a subspace closed under composition but not a subalgebra, since it does not contain id_V. The only subspaces of V at all are $\{0\}$ and V, so A satisfies the letter of irreducibility, yet $Av = \{0\} \neq V$ and $A \neq \mathrm{End}(V) = F$. Every statement below about an algebra acting irreducibly assumes the identity belongs to the algebra.

23.5.3 Burnside's theorem

Here is the question. An algebra $A \subseteq \mathrm{End}(V)$ that acts irreducibly has nowhere to hide: no proper nonzero subspace survives it. Does that force A to be all of $\mathrm{End}(V)$? Over $\mathbb{C}$ the answer is yes. The proof is worth the trouble: the corollary below, which says that an irreducibly acting **set** of operators already spans $\mathrm{End}(V)$, is what the representation theory later in this chapter uses to turn statements about a group into statements about all matrices.

Theorem 23.5.2. Burnside's Theorem

Let F be an **algebraically closed** field, that is, one in which every nonconstant polynomial in $F[x]$ has a root. Let V be a vector space over F with $1 \leq \dim V < \infty$, and let $A \subseteq \mathrm{End}(V)$ be a subalgebra with $\mathrm{id}_V \in A$. If V is irreducible as an A-space, then

$$A = \mathrm{End}(V).$$

Idea. **Step roadmap.** ① Among the nonzero members of A, let r be the smallest rank, attained at a. ② Show $r = 1$. If $r \geq 2$, the image $\mathrm{im}\, a$ has two independent vectors au_1, au_2, and irreducibility supplies $b \in A$ carrying au_1 to u_2. The operator ab maps $\mathrm{im}\, a$ into itself; an eigenvalue λ of that restriction — this is the one place algebraic closure is spent — makes $aba - \lambda a$ a member of A of strictly smaller rank, hence zero, and then $au_2 = \lambda\, au_1$ contradicts independence. ③ One rank-one operator $v_0\varphi_0$ in A gives all of them: multiplying on the left

moves v_0 anywhere, and multiplying on the right moves φ_0 anywhere. ④ Rank-one operators span $\operatorname{End}(V)$.

Proof. Write $v\varphi$ for the operator $x \mapsto \varphi(x)v$, where $v \in V$ and $\varphi \in V^*$; it has rank 1 exactly when $v \neq 0$ and $\varphi \neq 0$, and every operator of rank at most 1 has this form, since an operator with image $\operatorname{span}(v)$, $v \neq 0$, sends each x to a unique multiple $\varphi(x)v$, and φ so defined is linear.

Step 1: a member of least rank. Since $\mathrm{id}_V \in A$ and $\dim V \geq 1$, the set $\{a \in A : a \neq 0\}$ is nonempty. Let

$$r = \min\{\operatorname{rank} a : a \in A,\ a \neq 0\} \geq 1,$$

and fix $a \in A$ with $a \neq 0$ and $\operatorname{rank} a = r$.

Step 2: $r = 1$. Suppose $r \geq 2$. Then $\dim \operatorname{im} a \geq 2$, so there are $u_1, u_2 \in V$ with au_1 and au_2 independent; in particular $au_1 \neq 0$. By Lemma 23.5.1 there is $b \in A$ with $b(au_1) = u_2$.

Put $W = \operatorname{im} a$, a subspace of dimension r. Then $ab(W) \subseteq \operatorname{im} a = W$, so $s := (ab)|_W$ is an operator on W. Its characteristic polynomial p_s is monic of degree $r \geq 1$, hence nonconstant, hence has a root $\lambda \in F$ because F is algebraically closed; by Theorem 9.2.1, λ is an eigenvalue of s, so $\ker(s - \lambda\,\mathrm{id}_W) \neq \{0\}$.

Consider $c := aba - \lambda a$, which lies in A because A is closed under products and linear combinations. For every $x \in V$,

$$c\,x = (ab - \lambda\,\mathrm{id}_V)(ax) = (s - \lambda\,\mathrm{id}_W)(ax),$$

the second equality because $ax \in W$. Hence $\operatorname{im} c = (s - \lambda\,\mathrm{id}_W)(W)$, and Theorem 2.3.1 applied to $s - \lambda\,\mathrm{id}_W$ on W gives

$$\operatorname{rank} c = r - \dim\ker(s - \lambda\,\mathrm{id}_W) \leq r - 1.$$

By the minimality of r, an element of A of rank less than r must be 0, so $c = 0$, that is, $aba = \lambda a$. Applying both sides to u_1 and using $bau_1 = u_2$,

$$au_2 = (aba)u_1 = \lambda\, au_1,$$

which contradicts the independence of au_1 and au_2. Therefore $r = 1$.

Step 3: A contains every operator of rank at most one. By Step 2 there is $a_0 \in A$ with $\operatorname{rank} a_0 = 1$; write $a_0 = v_0\varphi_0$ with $v_0 \neq 0$ and $\varphi_0 \neq 0$.

Left multiplication. For $b \in A$ and $x \in V$ we get $b\,a_0\,x = \varphi_0(x)\,bv_0$, that is, $ba_0 = (bv_0)\varphi_0$. By Lemma 23.5.1, $Av_0 = V$, so

$$v\varphi_0 \in A \qquad \text{for every } v \in V.$$

Right multiplication. For $c \in A$ we get $a_0c = v_0(\varphi_0 \circ c)$. Put

$$S = \{\, \varphi_0 \circ c : c \in A \,\} \subseteq V^*,$$

the image of the linear map $A \to V^*$, $c \mapsto \varphi_0 \circ c$, and hence a subspace of V^*.

Claim. $S = V^*$.

Proof. Let $U = \{x \in V : \varphi(x) = 0 \text{ for every } \varphi \in S\}$, a subspace of V. It is A-invariant: if $x \in U$ and $d \in A$, then for every $c \in A$ we have $cd \in A$ and so $\varphi_0(c(dx)) = \varphi_0((cd)x) = 0$, whence $dx \in U$. It is not all of V: taking $c = \mathrm{id}_V$ shows $\varphi_0 \in S$, and $\varphi_0 \neq 0$ gives some x with $\varphi_0(x) \neq 0$. Irreducibility therefore forces $U = \{0\}$.

Let $d = \dim S$ and pick a basis $\varphi_1, \dots, \varphi_d$ of S. The linear map $\Theta\colon V \to F^d$, $\Theta(x) = (\varphi_1(x), \dots, \varphi_d(x))$, has kernel exactly $U = \{0\}$, since a vector killed by a basis of S is killed by all of S. So Θ is injective and Theorem 2.3.1 gives $\dim V \leq d$. On the other hand S is a subspace of V^*, and $\dim V^* = \dim V$ by Corollary 5.1.4, so $d \leq \dim V$. Hence $d = \dim V^*$ and $S = V^*$ by Theorem 1.6.8.

■

Now let $v \in V$ and $\varphi \in V^*$ be arbitrary. Choose $c \in A$ with $\varphi = \varphi_0 \circ c$, which the Claim allows, and $b \in A$ with $bv_0 = v$, which Lemma 23.5.1 allows when $v \neq 0$ (and for $v = 0$ take $b = 0$). Then for every x,

$$(b\, a_0\, c)x = b(\varphi_0(cx)\, v_0) = \varphi(x)\, v,$$

so $v\varphi = ba_0c \in A$.

Step 4: rank-one operators span. Let $t \in \mathrm{End}(V)$, let $(v_1, \dots, v_n)$ be a basis of V and $(\varphi_1, \dots, \varphi_n)$ its dual basis (Definition 5.1.3). The operator $\sum_{i=1}^{n} (tv_i)\varphi_i$ sends v_j to tv_j for each j, because $\varphi_i(v_j) = \delta_{ij}$; two operators agreeing on a basis are equal, so $t = \sum_i (tv_i)\varphi_i$. Each summand lies in A by Step 3, and A is a subspace, so $t \in A$. This proves $A = \mathrm{End}(V)$.

■

The theorem is stronger than it may look. It says that irreducibility, a statement about *subspaces*, forces a statement about *dimension*: over an algebraically closed field, an irreducibly acting subalgebra of $\mathrm{End}(V)$ has dimension exactly $(\dim V)^2$, the largest possible. There is no room between "acts irreducibly" and "is everything".

The version about sets rather than algebras is worth recording alongside it, since a group of invertible operators is a set closed under composition and not a subspace.

Corollary 23.5.3. An Irreducible Multiplicative Set Spans the Whole Algebra

Let F be algebraically closed, V a vector space over F with $1 \leq \dim V < \infty$, and let $\mathcal{S} \subseteq \mathrm{End}(V)$ be a set with $\mathrm{id}_V \in \mathcal{S}$ that is closed under composition. If the only subspaces $U \subseteq V$ with $s(U) \subseteq U$ for all $s \in \mathcal{S}$ are $\{0\}$ and V, then $\mathrm{span}(\mathcal{S}) = \mathrm{End}(V)$.

Proof. Put $A = \mathrm{span}(\mathcal{S})$. It is a subspace containing id_V, and it is closed under multiplication: a product of two linear combinations of members of $\mathcal{S}$ expands, by bilinearity of composition, into a linear combination of products of members of $\mathcal{S}$, which lie in $\mathcal{S}$. So A is a subalgebra. A subspace U is A-invariant if and only if it is invariant under every member of $\mathcal{S}$, since invariance is preserved by linear combinations. So V is irreducible as an A-space, and Theorem 23.5.2 gives $A = \mathrm{End}(V)$.

■

Warning. **Burnside is false over a field that is not algebraically closed.** Let $F = \mathbb{R}$, $V = \mathbb{R}^2$ and

$$A = \left\{ \begin{pmatrix} x & -y \\ y & x \end{pmatrix} : x, y \in \mathbb{R} \right\},$$

the span of $\mathbf{I}_2$ and $\mathbf{Q} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$. Since $\mathbf{Q}^2 = -\mathbf{I}_2$, this is a subalgebra, and it acts irreducibly: an invariant line would be spanned by a real eigenvector of $\mathbf{Q}$, and $p_{\mathbf{Q}} = x^2 + 1$ has no real root. Yet $\dim A = 2$, not 4, so $A \neq M_2(\mathbb{R})$. This is the same witness §04 used against the second half of Schur's lemma (Theorem 23.4.1), and the reason is the same: an operator on a real space need have no eigenvalue, and Step 2 of the proof needs one.

Quick check. Let $A \subseteq M_2(\mathbb{C})$ be the subalgebra generated by $\mathbf{X} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ and $\mathbf{Z} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$. Without computing products, what must $\dim A$ be, and why?

Solution. It must be 4, so $A = M_2(\mathbb{C})$. A one-dimensional invariant subspace would be a common eigenvector of $\mathbf{X}$ and $\mathbf{Z}$. The eigenvectors of $\mathbf{Z}$ are the multiples of $\mathbf{e}_1$ and of $\mathbf{e}_2$, and $\mathbf{X}\mathbf{e}_1 = \mathbf{e}_2$ is not a multiple of $\mathbf{e}_1$, nor is $\mathbf{X}\mathbf{e}_2 = \mathbf{e}_1$ a multiple of $\mathbf{e}_2$. So A acts irreducibly on $\mathbb{C}^2$, and Theorem 23.5.2 gives $A = M_2(\mathbb{C})$. Directly: $\mathbf{I}_2, \mathbf{X}, \mathbf{Z}$ and $\mathbf{XZ} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ are independent, which confirms it.

23.5.4 The double commutant theorem

Burnside handles the irreducible case. The general case is not "A is everything" — a diagonal algebra is not everything — but a statement about how faithfully A is recorded by its commutant. The right hypothesis is semisimplicity of the action, and it cannot be weakened to nothing, as the warning below shows.

The engine is a single step: an operator in A'' can be matched by a member of A on **one** given vector.

Lemma 23.5.4. Matching on One Vector

Let W be a finite-dimensional vector space over F and let $B \subseteq \mathrm{End}(W)$ be a subalgebra with $\mathrm{id}_W \in B$ which acts **semisimply** on W: every B-invariant subspace of W has a B-invariant complement (Definition 23.3.4). Then for every $t \in B''$ and every $w \in W$ there is $b \in B$ with $bw = tw$.

Idea. The subspace Bw is invariant, so it has an invariant complement, and the projection p onto Bw along that complement commutes with B. Being in B'', the operator t commutes with p, hence cannot move w out of the image of p, which is Bw.

Proof. Put $U = Bw = \{bw : b \in B\}$, a subspace of W (the image of $b \mapsto bw$) which is B-invariant, since $b_1(bw) = (b_1 b)w$ for $b_1 \in B$. By hypothesis there is a B-invariant subspace U_1 with $W = U \oplus U_1$. Let $p \in \mathrm{End}(W)$ be the projection onto U along U_1 (Theorem 4.5.1 (b)).

$p \in B'$. Indeed, let $b \in B$ and write $x = u + u_1$ with $u \in U$, $u_1 \in U_1$. Then $bu \in U$ and $bu_1 \in U_1$ by invariance, so $bx = bu + bu_1$ is the decomposition of bx, giving $p(bx) = bu = b(px)$.

Since $t \in B''$ and $p \in B'$, we have $tp = pt$. Also $w = \mathrm{id}_W\, w \in Bw = U$, so $pw = w$. Therefore

$$tw = t(pw) = p(tw) \in \mathrm{im}\, p = U = Bw,$$

so $tw = bw$ for some $b \in B$.

■

Matching on one vector is not yet matching everywhere. The trick that upgrades it is to run the lemma not on V but on n copies of V at once, where one cleverly chosen vector carries a whole basis.

Theorem 23.5.5. Double Commutant Theorem

Let F be **any** field, let V be a finite-dimensional vector space over F, and let $A \subseteq \mathrm{End}(V)$ be a subalgebra with $\mathrm{id}_V \in A$. Assume that A acts **semisimply** on V: every A-invariant subspace of V has an A-invariant complement (Definition 23.3.4). Then

$$A'' = A.$$

Idea. **Step roadmap.** ① Lemma 23.5.4 matches $t \in A''$ with some $a \in A$ on one vector. ② To match on a whole basis at once, let A act diagonally on $V^n = V \oplus \cdots \oplus V$ ($n = \dim V$) and apply the lemma there to the single vector whose n entries are the basis vectors. ③ For that we must check two things about the copied algebra $\tilde{A}$: that it still acts semisimply, which Theorem 23.3.3 supplies, and that $\tilde{t}$ lies in $\tilde{A}''$, which is a computation with the $n \times n$ array of components of an operator on V^n.

Proof. The inclusion $A \subseteq A''$ was noted above, so only $A'' \subseteq A$ needs proof. Let $t \in A''$, let $n = \dim V$; if $n = 0$ there is nothing to prove, so let $n \geq 1$.

Step 1: the copied algebra. Let $V^n = V \oplus \cdots \oplus V$ be the product of n copies of V (Definition 4.4.1), and for $s \in \mathrm{End}(V)$ let $\tilde{s} \in \mathrm{End}(V^n)$ be given by $\tilde{s}(x_1, \ldots, x_n) = (sx_1, \ldots, sx_n)$. The map $s \mapsto \tilde{s}$ is linear and satisfies $\widetilde{s_1 s_2} = \tilde{s}_1 \tilde{s}_2$ and $\widetilde{\mathrm{id}_V} = \mathrm{id}_{V^n}$, so $\tilde{A} = \{\tilde{a} : a \in A\}$ is a subalgebra of $\mathrm{End}(V^n)$ containing the identity.

Step 2: $\tilde{A}$ acts semisimply on V^n. Since A acts semisimply on V, Theorem 23.3.3 writes $V = W_1 + \cdots + W_q$ with each W_j an irreducible A-invariant subspace. For $1 \leq k \leq n$ let $\iota_k \colon V \to V^n$ place a vector in the k-th slot and 0 elsewhere; it is injective and satisfies $\iota_k(sx) = \tilde{s}\, \iota_k(x)$. Hence $\iota_k(W_j)$ is $\tilde{A}$-invariant, and it is irreducible, because ι_k carries the A-invariant subspaces of W_j bijectively onto the $\tilde{A}$-invariant subspaces of $\iota_k(W_j)$. Every

element of V^n is the sum of its n slots, so

$$V^n = \sum_{k=1}^{n} \sum_{j=1}^{q} \iota_k(W_j),$$

a sum of irreducible $\tilde{A}$-invariant subspaces. By the converse half of Theorem 23.3.3, $\tilde{A}$ acts semisimply on V^n.

Step 3: $\tilde{t} \in \tilde{A}''$. Let $\pi_k \colon V^n \to V$ be the k-th coordinate map. Every $s \in \mathrm{End}(V^n)$ is determined by its components $s_{kl} = \pi_k \circ s \circ \iota_l \in \mathrm{End}(V)$, through $\pi_k(sx) = \sum_l s_{kl} x_l$. Comparing components in $s\tilde{a} = \tilde{a}s$ gives

$$(s\tilde{a})_{kl} = s_{kl}\, a, \qquad (\tilde{a}s)_{kl} = a\, s_{kl},$$

so $s \in \tilde{A}'$ if and only if every s_{kl} lies in A'. Now let $s \in \tilde{A}'$. Then $(\tilde{t}s)_{kl} = t\, s_{kl}$ and $(s\tilde{t})_{kl} = s_{kl}\, t$, and these agree because $s_{kl} \in A'$ and $t \in A''$. Hence $\tilde{t}$ commutes with every $s \in \tilde{A}'$, that is, $\tilde{t} \in \tilde{A}''$.

Step 4: conclude. Let $(v_1, \dots, v_n)$ be a basis of V and put $\tilde{v} = (v_1, \dots, v_n) \in V^n$. By Steps 1–3 the hypotheses of Lemma 23.5.4 hold for $B = \tilde{A}$ on $W = V^n$, with $\tilde{t} \in \tilde{A}''$. So there is $a \in A$ with $\tilde{a}\,\tilde{v} = \tilde{t}\,\tilde{v}$, that is, $av_i = tv_i$ for every i. Two operators agreeing on a basis are equal, so $t = a \in A$. This proves $A'' \subseteq A$, and with the reverse inclusion, $A'' = A$.

■

Notice which hypotheses were used where. Algebraic closure was **not** used: the double commutant theorem holds over every field, unlike Burnside. Finite dimension was used for the basis in Step 4 and inside Theorem 23.3.3. Semisimplicity was used exactly once, to produce the invariant complement in Lemma 23.5.4.

Warning. **Neither hypothesis can simply be dropped.**

Without the identity. Take $A = \{0\} \subseteq \mathrm{End}(F^2)$, a subspace closed under composition but with no identity element of $\mathrm{End}(F^2)$ in it. Every subspace of F^2 is A-invariant and every subspace has a complement, so the condition on invariant complements holds; but $A' = M_2(F)$ and $A'' = \{\lambda \mathbf{I}_2\}$ by Proposition 23.1.5, so $A'' \neq A$.

Without semisimplicity. Let $A \subseteq M_2(F)$ be the upper triangular matrices. The line $\mathrm{span}(\mathbf{e}_1)$ is A-invariant, and it is the **only** invariant line: if $\mathbf{v} = (v_1, v_2)$ has $v_2 \neq 0$, then $\mathbf{E}_{12}\mathbf{v} = v_2\mathbf{e}_1$ is a nonzero multiple of $\mathbf{e}_1$ and does not lie in $\mathrm{span}(\mathbf{v})$. A complement of $\mathrm{span}(\mathbf{e}_1)$ would be an invariant line other than $\mathrm{span}(\mathbf{e}_1)$, so there is none and A does not act semisimply. Here $A' = \{\lambda \mathbf{I}_2\}$: a matrix commuting with $\mathbf{E}_{11}$ and $\mathbf{E}_{22}$ is diagonal, and commuting with $\mathbf{E}_{12}$ then forces the two diagonal entries to agree. Hence $A'' = \{\lambda \mathbf{I}_2\}' = M_2(F)$, which is strictly larger than A.

Remark. Semisimplicity is sufficient but not necessary. Let $\mathbf{N} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$ and $A = \{x\mathbf{I}_2 + y\mathbf{N}\}$. Solving $\mathbf{XN} = \mathbf{NX}$ gives exactly the matrices $x\mathbf{I}_2 + y\mathbf{N}$, so $A' = A$ and therefore $A'' = A$ — even though A does not act semisimply, the invariant line $\mathrm{span}(\mathbf{e}_1) = \ker \mathbf{N}$ again being the only one.

Remark. The name has two sources. In finite dimensions the result above is due to Wedderburn, and the argument through n copies of V is the finite-dimensional shadow of Jacobson's density theorem, which we do not prove. For algebras of operators on a Hilbert space there is an analytic statement of the same shape, von Neumann's double commutant theorem, which is outside this book.

23.5.5 The commutant of one operator, revisited

The two earlier computations of the book now read as computations of A' for the smallest interesting algebra: the one generated by a single operator.

Corollary 23.5.6. The Commutant of a Single Operator

Let V be a finite-dimensional vector space over F, let $T \in \mathcal{L}(V)$, and let $A = F[T] \subseteq \mathrm{End}(V)$ be the algebra of all polynomials in T. Then

$$A' = \{S \in \mathcal{L}(V) : ST = TS\}.$$

Consequently the following two earlier results are computations of A'.

(a) If $F = \mathbb{C}$ and the matrix of T in some basis is $\mathbf{A}_1 \oplus \cdots \oplus \mathbf{A}_r$ with the $\mathbf{A}_p$ of pairwise disjoint spectra, then in that same basis every member of A' has a block **diagonal** matrix for the partition into those block sizes (Corollary 12.10.4).

(b) If the matrix of T in some basis is a basic Weyr matrix of Weyr structure $(w_1, \ldots, w_s)$, then in that same basis every member of A' has a block **upper triangular** matrix for the partition $w_1 + \cdots + w_s$ (Proposition 10.4.6).

If moreover T is diagonalizable with distinct eigenvalues $\lambda_1, \ldots, \lambda_k$ and $d_i = \dim E_{\lambda_i}(T)$, then A acts semisimply, $\dim A = k$, $\dim A' = \sum_{i=1}^{k} d_i^2$, and $A'' = A$.

Idea. Nothing in (a) and (b) is new: whatever commutes with T commutes with every polynomial in it, so the two earlier computations were computations of A' all along. The diagonalizable case is where the second commutant enters. The Lagrange polynomials evaluated at T are the projections onto the eigenspaces, and they turn out to be a basis of A; a commuting operator is exactly one that preserves each eigenspace and is otherwise unconstrained, which counts A'; and the same projections assemble an invariant complement for any invariant subspace, which is the hypothesis Theorem 23.5.5 needs.

Proof. An operator commutes with every polynomial in T as soon as it commutes with T, and $T \in A$; this gives the displayed description of A'. Parts (a) and (b) are then the cited results verbatim, read for the matrix of T in the given basis.

For the last sentence, let T be diagonalizable, $E_i = E_{\lambda_i}(T)$, so $V = E_1 \oplus \cdots \oplus E_k$ by Theorem 9.4.1 (c). Let $\ell_i \in F[x]$ be the Lagrange polynomial with $\ell_i(\lambda_j) = \delta_{ij}$ (Theorem 6.6.1) and $P_i = \ell_i(T)$. Evaluating on each E_j, where T acts as λ_j, shows P_i is the projection onto E_i along the other summands. Since $T^m = \sum_i \lambda_i^m P_i$ for every $m \geq 0$, the algebra A is spanned by $P_1, \ldots, P_k$, which are independent because $P_i P_j = \delta_{ij} P_i$ and each $P_i \neq 0$: applying $\sum_i c_i P_i = 0$ to a nonzero vector of E_j gives $c_j = 0$. So $\dim A = k$.

By Theorem 9.9.1, every $S \in A'$ maps each E_i into itself; conversely such an S commutes with T, because both ST and TS act as $\lambda_i S$ on E_i. So A' is the set of operators preserving all the E_i, and restricting gives a linear bijection $A' \to \mathrm{End}(E_1) \times \cdots \times \mathrm{End}(E_k)$; hence $\dim A' = \sum_i d_i^2$. Each P_i preserves every E_j, so $P_i \in A'$.

Finally, A acts semisimply. Let U be an A-invariant subspace. Then U is T-invariant, and $T|_U$ is diagonalizable (Corollary 9.8.3), so U has a basis of eigenvectors of T (Theorem 9.4.1, (a) $\Leftrightarrow$ (b)) and therefore $U = (U \cap E_1) \oplus \cdots \oplus (U \cap E_k)$. Choose a complement C_i of $U \cap E_i$ inside E_i for each i (Theorem 1.7.6) and put $C = C_1 \oplus \cdots \oplus C_k$. Then $V = U \oplus C$, and C is A-invariant because $P_j(C) = C_j \subseteq C$ and the P_j span A. So Theorem 23.5.5 applies and gives $A'' = A$.

■

So the diagonalizable case is an instance of Theorem 23.5.5, reached here by hand; the Weyr case in (b) is **not**, since $F[\mathbf{W}]$ does not act semisimply once $s \geq 2$. That is the honest shape of the relationship: the two earlier sections computed first commutants in cases where the second commutant theorem has nothing to say, and this section explains what extra hypothesis would make the computation reversible.

Example. The Two Extremes

Inside $M_n(F)$ with $n \geq 2$, compute A' and A'' for (i) A the diagonal matrices, and (ii) A the scalar matrices.

Solution.

(i) Let $A = \{\mathrm{diag}(c_1, \ldots, c_n)\}$. Then $\mathbf{X}$ commutes with $\mathbf{E}_{ii}$ for every i exactly when $\mathbf{X}$ is diagonal, since $(\mathbf{X}\mathbf{E}_{ii})_{kl} = x_{ki}\delta_{il}$ and $(\mathbf{E}_{ii}\mathbf{X})_{kl} = \delta_{ik}x_{il}$, and comparing at (k, i) with $k \neq i$ gives $x_{ki} = 0$. So $A' = A$, and applying the prime again, $A'' = A' = A$. This agrees with Theorem 23.5.5: the diagonal algebra acts semisimply, F^n being the direct sum of the n coordinate lines, each of which is invariant and one-dimensional, hence irreducible.

(ii) Let $A = \{\lambda \mathbf{I}_n\}$. Every matrix commutes with a scalar matrix, so $A' = M_n(F)$, and $A'' = M_n(F)' = \{\lambda \mathbf{I}_n\} = A$ by Proposition 23.1.5. Again $A'' = A$, as it must be: the scalars act semisimply because every subspace is invariant and every subspace has a complement.

The two examples sit at opposite ends. In (i) the commutant is as small as the algebra; in (ii) the commutant is everything. In both, applying the prime a second time returns to the start, which is what the theorem promises whenever the action is semisimple.

23.5.6 Exercises

23.5.6.1 A. Check your understanding

Exercise. A1

(a) State Burnside's theorem, with every hypothesis.

(b) Give the field hypothesis of Theorem 23.5.5, and say why it differs from the one in (a).

(c) True or false, with a reason: if $A \subseteq \mathrm{End}(V)$ is a subalgebra containing id_V, then $A''' = A'$.

(d) Explain in one sentence why $\{0\}$ and V being the only A-invariant subspaces does **not** by itself imply $Av = V$ for $v \neq 0$.

Solution.

(a) F algebraically closed, V finite-dimensional over F with $\dim V \geq 1$, $A \subseteq \mathrm{End}(V)$ a subalgebra containing id_V, and V irreducible as an A-space; the conclusion is $A = \mathrm{End}(V)$.

(b) Theorem 23.5.5 has no field hypothesis at all: it holds over every field. Algebraic closure entered Burnside only to produce an eigenvalue of $(ab)|_{\mathrm{im}\, a}$, and the double commutant proof never needs an eigenvalue — it needs an invariant complement, which semisimplicity hands over directly.

(c) True, and it holds for any subset: it is Proposition 23.1.4 (d). The two ingredients are (c) of that proposition, $\mathcal{S} \subseteq \mathcal{S}''$, and (b), that $\mathcal{S} \subseteq \mathcal{T}$ implies $\mathcal{T}' \subseteq \mathcal{S}'$. Applying the prime to $A \subseteq A''$ gives $A''' \subseteq A'$, and applying the inclusion $B \subseteq B''$ to $B = A'$ gives $A' \subseteq A'''$.

(d) Because Av could be $\{0\}$ if A contained no operator sending v anywhere nonzero; the identity in A is what rules this out, as in the warning after Lemma 23.5.1.

23.5.6.2 B. Practice

Exercise. B1: Determine which are irreducible

For each subalgebra of $M_2(\mathbb{C})$, determine whether it acts irreducibly on $\mathbb{C}^2$, and hence whether it equals $M_2(\mathbb{C})$. Justify your answer.

(a) The algebra generated by $\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$.

(b) The algebra generated by $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ and $\begin{pmatrix} 1 & 0 \\ 0 & i \end{pmatrix}$.

(c) The algebra of all matrices $\begin{pmatrix} a & b \\ 0 & a \end{pmatrix}$.

Solution.

(a) Not irreducible. With $\mathbf{N} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$ we have $\mathbf{N}^2 = \mathbf{O}$, so the algebra is $\{x\mathbf{I}_2 + y\mathbf{N}\}$, and $\mathrm{span}(\mathbf{e}_1) = \mathrm{im}\, \mathbf{N}$ is invariant, proper, and nonzero. It is not $M_2(\mathbb{C})$; its dimension is 2.

(b) Irreducible, so it equals $M_2(\mathbb{C})$ by Theorem 23.5.2. The eigenvectors of $\mathrm{diag}(1, i)$ are the multiples of $\mathbf{e}_1$ and of $\mathbf{e}_2$ (the diagonal entries are distinct), and the swap matrix maps each of those lines to the other, so no line is invariant under both.

(c) Not irreducible: this is the algebra of (a), written out. The invariant line is again

$\operatorname{span}(\mathbf{e}_1)$.

Exercise. B2: A commutant and a double commutant

Let $\mathbf{A} = \operatorname{diag}(1,1,2) \in M_3(\mathbb{R})$ and let $A = \mathbb{R}[\mathbf{A}] \subseteq M_3(\mathbb{R})$.

(a) Compute $\dim A$.

(b) Describe A' explicitly and compute $\dim A'$.

(c) Compute A'', and check the answer against Theorem 23.5.5.

Solution.

(a) The eigenvalues are 1 and 2, each with a nonzero eigenspace, so $m_{\mathbf{A}} = (x-1)(x-2)$ and, by the count in Corollary 23.5.6 with $k = 2$, $\dim A = 2$: a basis is $\mathbf{I}_3, \mathbf{A}$, equivalently $\mathbf{P}_1 = \operatorname{diag}(1,1,0)$, $\mathbf{P}_2 = \operatorname{diag}(0,0,1)$.

(b) $E_1 = \operatorname{span}(\mathbf{e}_1, \mathbf{e}_2)$ and $E_2 = \operatorname{span}(\mathbf{e}_3)$. Every matrix in A' preserves both (Theorem 9.9.1), and conversely a matrix preserving both commutes with $\mathbf{A}$, since $\mathbf{A}$ acts as a scalar on each; so a matrix lies in A' exactly when it has the shape

$$\begin{pmatrix} p & q & 0 \\ u & v & 0 \\ 0 & 0 & w \end{pmatrix}.$$

Its dimension is $2^2 + 1^2 = 5$.

(c) A matrix commuting with all of these must commute in particular with $\operatorname{diag}(1,1,0)$ and $\operatorname{diag}(0,0,1)$, hence be of the same block shape, and its 2×2 block must commute with all of $M_2(\mathbb{R})$, hence be scalar (Proposition 23.1.5). So $A'' = \{\alpha \mathbf{P}_1 + \beta \mathbf{P}_2\} = A$, of dimension 2. Since $\mathbf{A}$ is diagonalizable, A acts semisimply, and Theorem 23.5.5 predicts exactly $A'' = A$.

Exercise. B3: One theorem, once

Let F be algebraically closed, let V be finite-dimensional over F with $\dim V \geq 2$, and let $A \subseteq \operatorname{End}(V)$ be a **commutative** subalgebra containing id_V. Prove that A does not act irreducibly on V.

Solution. Suppose it did. By Theorem 23.5.2, $A = \operatorname{End}(V)$. But $\operatorname{End}(V)$ is not commutative when $\dim V \geq 2$: choosing a basis and the operators with matrices $\mathbf{E}_{12}$ and $\mathbf{E}_{21}$ gives $\mathbf{E}_{12}\mathbf{E}_{21} = \mathbf{E}_{11} \neq \mathbf{E}_{22} = \mathbf{E}_{21}\mathbf{E}_{12}$. This contradicts commutativity of A, so A is not irreducible.

23.5.6.3 C. Going deeper

Exercise. C1: The rank-one step needs its hypothesis

In Step 2 of the proof of Theorem 23.5.2, the operator ab was restricted to $W = \operatorname{im} a$ and an eigenvalue was extracted.

(a) Prove that $ab(W) \subseteq W$, and give an example of $a, b \in M_2(F)$ with $ba(\operatorname{im} a) \not\subseteq \operatorname{im} a$, so that the order of the product matters.

(b) Take $F = \mathbb{R}$, $A = \{x\mathbf{I}_2 + y\mathbf{Q}\}$ with $\mathbf{Q} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$. Compute the minimal rank r of a nonzero member of A and confirm that the conclusion $r = 1$ fails.

(c) Which sentence of Step 2 fails for this A? Say exactly where the hypothesis on F was used.

Solution.

(a) $ab(W) = a(b(W)) \subseteq \operatorname{im} a = W$, since a is applied last. For the other order take $a = \mathbf{E}_{11}$ and $b = \mathbf{E}_{21}$ in $M_2(F)$. Then $\operatorname{im} a = \operatorname{span}(\mathbf{e}_1)$ and $ba\,\mathbf{e}_1 = \mathbf{E}_{21}\mathbf{e}_1 = \mathbf{e}_2 \notin \operatorname{span}(\mathbf{e}_1)$.

(b) A nonzero $x\mathbf{I}_2 + y\mathbf{Q}$ has determinant $x^2 + y^2 > 0$ for real x, y not both zero, so it is invertible and has rank 2. Hence $r = 2$, and the conclusion $r = 1$ fails; consistently, A contains no rank-one matrix and is not all of $M_2(\mathbb{R})$.

(c) The sentence "its characteristic polynomial has a root $\lambda \in F$ because F is algebraically closed". With $a = \mathbf{I}_2$ and $b = \mathbf{Q}$ we get $W = \mathbb{R}^2$ and $s = \mathbf{Q}$, whose characteristic polynomial $x^2 + 1$ has no real root. Everything before that sentence is valid over any field.

Exercise. C2: Commutants under a change of basis

Let V be finite-dimensional over F, let $A \subseteq \operatorname{End}(V)$ be a subalgebra containing id_V, and let $g \in \operatorname{End}(V)$ be invertible. Put $gAg^{-1} = \{gag^{-1} : a \in A\}$.

(a) Prove that gAg^{-1} is a subalgebra containing id_V and that $(gAg^{-1})' = gA'g^{-1}$.

(b) Deduce that if $A'' = A$ then $(gAg^{-1})'' = gAg^{-1}$.

(c) Hence show that the conclusion of Theorem 23.5.5 is a property of the algebra up to similarity, and use this to give, for each $n \geq 2$, a subalgebra of $M_n(F)$ with $A'' \neq A$.

Solution.

(a) The map $a \mapsto gag^{-1}$ is linear, sends id_V to id_V, and satisfies $(ga_1g^{-1})(ga_2g^{-1}) = g(a_1a_2)g^{-1}$, so the image is a subalgebra containing the identity. For the commutant: t commutes with every gag^{-1} if and only if $g^{-1}tg$ commutes with every a, that is, if and only if $g^{-1}tg \in A'$, that is, $t \in gA'g^{-1}$.

(b) Apply (a) twice: $(gAg^{-1})'' = (gA'g^{-1})' = gA''g^{-1} = gAg^{-1}$.

(c) By (b), whether $A'' = A$ holds is unchanged by replacing A with a similar copy, so it is a property of the similarity class. For a failure in $M_n(F)$ take $A_n =$ all upper triangular matrices. The same computation as in the warning after Theorem 23.5.5, done with the matrix units $\mathbf{E}_{ii}$ and $\mathbf{E}_{i,i+1}$, gives $A_n' = \{\lambda\mathbf{I}_n\}$, hence $A_n'' = M_n(F) \neq A_n$; and by (b) every similar copy gA_ng^{-1} fails too.

Exercise. C3: Is this look-alike a theorem?

Let V be finite-dimensional over an algebraically closed F and let $A \subseteq \operatorname{End}(V)$ be a subal-

gebra with $\mathrm{id}_V \in A$.

(a) Prove: if V is irreducible as an A-space then $A' = \{\lambda\ \mathrm{id}_V\}$.

(b) Is the converse true? That is, does $A' = \{\lambda\ \mathrm{id}_V\}$ force V to be irreducible? Justify your answer.

Hint for (b): consider the upper triangular matrices in $M_2(F)$.

Solution.

(a) By Theorem 23.5.2, $A = \mathrm{End}(V)$, so $A' = \mathrm{End}(V)' = \{\lambda\ \mathrm{id}_V\}$ by Proposition 23.1.5. (This recovers the second half of Schur's lemma, Theorem 23.4.1, in this special case.)

(b) No. Let A be the upper triangular matrices in $M_2(F)$. The warning after Theorem 23.5.5 computes $A' = \{\lambda \mathbf{I}_2\}$, yet $\mathrm{span}(\mathbf{e}_1)$ is a proper nonzero invariant subspace, so F^2 is not irreducible as an A-space — and indeed $A \neq M_2(F)$. A trivial commutant is therefore strictly weaker than irreducibility; it is semisimplicity, not a small commutant, that makes the commutant remember the algebra.

23.6 The Structure Theorem

§05 settled the extreme case: over an algebraically closed field, an algebra of operators with no invariant subspace to hide in is the whole of $\mathrm{End}(V)$. Most algebras are not like that. This section shows that, over an algebraically closed field and under one hypothesis — semisimplicity — that extreme case is nevertheless the only building block there is. Every such algebra is a product of full matrix algebras, and the sizes of the blocks are determined by the algebra. That is the structure theorem, and it is the last piece of general theory the chapter needs: from here on the work is to apply it to groups.

Throughout, F is a field and A is an associative algebra with identity, finite-dimensional over F, and with $A \neq \{0\}$.

23.6.1 Products of matrix algebras

Before proving that something is a product of matrix algebras, we should say what the product is and look at a few.

A product of algebras is a list of algebras run side by side, with nothing passing between the slots.

Definition 23.6.1. Product of algebras

Let $A_1, \dots, A_k$ be F-algebras. Their **product** $A_1 \times \cdots \times A_k$ is the product vector space (Definition 4.4.1) with multiplication defined **slot by slot**,

$$(x_1, \dots, x_k)(y_1, \dots, y_k) = (x_1 y_1, \dots, x_k y_k),$$

and identity $(1_{A_1}, \dots, 1_{A_k})$.

The axioms of Definition 14.12.1 hold slot by slot because they hold in each A_i, and $\dim(A_1 \times \cdots \times A_k) = \sum_i \dim A_i$ by Theorem 4.4.1. Two features of a product will be used constantly. First, the element e_i with 1_{A_i} in slot i and 0 elsewhere satisfies $e_i^2 = e_i$, $e_i e_j = 0$ for $i \neq j$,

and $e_1 + \cdots + e_k = 1$. Second, the product is almost never commutative, but its center always contains the k-dimensional span of $e_1, \ldots, e_k$.

Example. Three products, seen from two sides

Identify each of the following with a product of matrix algebras over F, and give its dimension: (i) the diagonal matrices in $M_n(F)$; (ii) $M_n(F)$ itself; (iii) the block diagonal matrices in $M_5(F)$ with blocks of sizes $2, 2, 1$, where the **two** 2×2 blocks are allowed to differ.

Solution.

(i) The map $\mathrm{diag}(c_1, \ldots, c_n) \mapsto (c_1, \ldots, c_n)$ is a bijection, is linear, sends $\mathbf{I}_n$ to $(1, \ldots, 1)$, and is multiplicative because diagonal matrices multiply entry by entry. So the diagonal matrices form $F \times \cdots \times F = M_1(F)^n$, of dimension $n = \sum_{i=1}^{n} 1^2$.

(ii) $M_n(F)$ is the product with one slot, $k = 1$ and $n_1 = n$, of dimension n^2.

(iii) Writing a block diagonal matrix as the list of its blocks gives $M_2(F) \times M_2(F) \times M_1(F)$, since block diagonal matrices multiply blockwise (Theorem 8.1.4, applied twice). Its dimension is $4 + 4 + 1 = 9 = 2^2 + 2^2 + 1^2$.

In all three the dimension is $\sum_i n_i^2$. The theorem below says that this is not a coincidence of the examples.

23.6.2 The regular representation

An abstract algebra has no vectors to act on until we give it some, and the cheapest supply is the algebra itself.

Remark. The vocabulary is §03's and §04's. An **A-space** is a vector space V with an algebra homomorphism $A \to \mathrm{End}(V)$, written av (Definition 23.3.2); **invariant** subspace and **irreducible** A-space are as in Definition 23.3.3; and a **homomorphism of A-spaces** $f\colon V \to W$ is what §04 calls an A-map, a linear map with $f(av) = af(v)$ (Definition 23.4.1). A bijective one is an **isomorphism**, and we then write $V \cong W$.

The **regular representation** of A, introduced in §03 (Definition 23.3.2), is A acting on itself by left multiplication: $L\colon A \to \mathrm{End}(A)$, $L_a(x) = ax$. It is injective because $L_a(1_A) = a$; this is exactly Theorem 23.1.3. Its invariant subspaces are the subspaces $I \subseteq A$ with $AI \subseteq I$, which is to say the **left ideals** of §02. So the irreducible invariant subspaces of the regular representation are the **minimal left ideals**: the nonzero left ideals containing no smaller nonzero left ideal.

We call A **semisimple** when the regular representation is semisimple in the sense of §03 (Definition 23.3.4): every left ideal of A has a complementary left ideal.

23.6.3 The structure theorem

Theorem 23.6.1. Wedderburn's Structure Theorem, Algebraically Closed Case

Let F be an **algebraically closed** field and let $A \neq \{0\}$ be a **semisimple** finite-dimensional

F-algebra: every left ideal of A has a complementary left ideal. Then there are an integer $k \geq 1$, integers $n_1, \ldots, n_k \geq 1$, and an isomorphism of F-algebras

$$A \;\cong\; M_{n_1}(F) \times \cdots \times M_{n_k}(F).$$

Moreover k and the multiset $\{n_1, \ldots, n_k\}$ depend only on A: they are the number of isomorphism classes of irreducible A-spaces and their dimensions.

Idea. **Step roadmap.** ① Break the regular representation into minimal left ideals. ② Keep one from each isomorphism class, say $V_1, \ldots, V_k$, and map A into $\mathrm{End}(V_1) \times \cdots \times \mathrm{End}(V_k)$ by letting it act on each. ③ The map is injective because an element acting as zero on every V_i acts as zero on all of A, hence kills 1. ④ It is surjective: on $W = V_1 \oplus \cdots \oplus V_k$, Schur's lemma computes the commutant of the image to be the scalars-on-each-slot, and Theorem 23.5.5 turns that into the statement that the image is everything preserving the slots. ⑤ Uniqueness: read the n_i off as the dimensions of the irreducible A-spaces, which no choice in the proof can change.

Proof. **Step 1: split the regular representation.** Regard A as an A-space by left multiplication. Its invariant subspaces are the left ideals, and by hypothesis every one of them has an invariant complement. By Theorem 23.3.3, (a) $\Rightarrow$ (c), A is a **direct** sum of finitely many irreducible invariant subspaces:

$$A = L_1 \oplus \cdots \oplus L_m, \qquad L_j \text{ a minimal left ideal.}$$

Since $A \neq \{0\}$ we have $m \geq 1$.

Step 2: one representative per class. Being isomorphic is an equivalence relation on the list $L_1, \ldots, L_m$; pick one member from each class and call the chosen spaces $V_1, \ldots, V_k$. They are pairwise nonisomorphic, and every L_j is isomorphic to exactly one of them. Put $n_i = \dim V_i \geq 1$ and let $\rho_i \colon A \to \mathrm{End}(V_i)$ be the homomorphism giving the action. Then $k \geq 1$.

Step 3: the map, and its injectivity. Let

$$\Phi \colon A \to \mathrm{End}(V_1) \times \cdots \times \mathrm{End}(V_k), \qquad \Phi(a) = (\rho_1(a), \ldots, \rho_k(a)).$$

Each ρ_i is an algebra homomorphism and the product has slot-by-slot operations (Definition 23.6.1), so Φ is an algebra homomorphism, and $\Phi(1_A)$ is the identity.

Suppose $\Phi(a) = 0$. Fix j, let i be the index with $L_j \cong V_i$, and let $f \colon L_j \to V_i$ be an isomorphism of A-spaces. For $x \in L_j$ we get $f(ax) = \rho_i(a) f(x) = 0$, and f is injective, so $ax = 0$. Thus $aL_j = \{0\}$ for every j, and since $A = L_1 + \cdots + L_m$ we get $aA = \{0\}$. In particular $a = a1_A = 0$. So Φ is injective.

Step 4: surjectivity. Let $W = V_1 \oplus \cdots \oplus V_k$ be the product space, with A acting slot by slot, and let $B \subseteq \mathrm{End}(W)$ be the image of A under this action; B is a subalgebra containing id_W. Let $D \subseteq \mathrm{End}(W)$ be the set of operators mapping each V_i into itself. Restricting to the slots identifies D with $\mathrm{End}(V_1) \times \cdots \times \mathrm{End}(V_k)$ as an algebra, and under that identification $\Phi(A)$ corresponds to B. So it suffices to prove $B = D$.

B acts semisimply on W. Each V_i is B-invariant, and a subspace of V_i is B-invariant exactly

when it is A-invariant, so each V_i is an irreducible B-invariant subspace; and $W = V_1 + \cdots + V_k$. By Theorem 23.3.3, B acts semisimply.

The commutant of B. Let $\iota_j \colon V_j \to W$ and $\pi_i \colon W \to V_i$ be the slot inclusions and projections, and for $s \in \mathrm{End}(W)$ put $s_{ij} = \pi_i \circ s \circ \iota_j \colon V_j \to V_i$. Then s commutes with the action of every $a \in A$ if and only if

$$s_{ij}\,\rho_j(a) = \rho_i(a)\,s_{ij} \qquad \text{for all } i, j, a,$$

that is, if and only if every s_{ij} is a homomorphism of A-spaces. For $i \neq j$ the spaces V_i and V_j are irreducible and nonisomorphic, so Theorem 23.4.1 gives $s_{ij} = 0$. For $i = j$ the field is algebraically closed and V_i is irreducible, so the second part of Theorem 23.4.1 gives $s_{ii} = \lambda_i \,\mathrm{id}_{V_i}$ for some $\lambda_i \in F$. Conversely every operator of that shape commutes with B. Hence

$$B' = \{\, \lambda_1 \,\mathrm{id}_{V_1} \oplus \cdots \oplus \lambda_k \,\mathrm{id}_{V_k} : \lambda_i \in F \,\}.$$

The double commutant of B. Let $p_i \in \mathrm{End}(W)$ be the projection onto V_i along the other slots; it is the member of B' with $\lambda_i = 1$ and the other $\lambda_j = 0$. If $t \in B''$, then t commutes with p_i, so

$$t(V_i) = t(\operatorname{im} p_i) = \operatorname{im}(tp_i) = \operatorname{im}(p_i t) \subseteq \operatorname{im} p_i = V_i,$$

which says $t \in D$. Conversely every $t \in D$ commutes with each $\lambda_1 \,\mathrm{id} \oplus \cdots \oplus \lambda_k \,\mathrm{id}$, so $D \subseteq B''$. Hence $B'' = D$.

Conclusion. B is a subalgebra of $\mathrm{End}(W)$ containing id_W and acting semisimply, so Theorem 23.5.5 gives $B = B'' = D$. Therefore Φ is surjective, and with Step 3 it is an isomorphism of algebras onto $\mathrm{End}(V_1) \times \cdots \times \mathrm{End}(V_k)$.

Step 5: matrices. Choosing a basis $\mathcal{B}_i$ of V_i makes $T \mapsto [T]^{\mathcal{B}_i}_{\mathcal{B}_i}$ a linear bijection $\mathrm{End}(V_i) \to M_{n_i}(F)$ (Theorem 4.1.4) which is multiplicative (Theorem 4.1.2) and sends id_{V_i} to $\mathbf{I}_{n_i}$, hence an isomorphism of algebras. Doing this in each slot gives

$$A \cong M_{n_1}(F) \times \cdots \times M_{n_k}(F).$$

Step 6: uniqueness. We first identify all irreducible A-spaces.

Claim. If $A = \sum_j L_j$ with each L_j an irreducible left ideal, then every irreducible A-space is isomorphic to some L_j.

Proof. Let U be an irreducible A-space and pick $u \in U$, $u \neq 0$. The map $f \colon A \to U$, $f(a) = au$, is a homomorphism of A-spaces from the regular representation, since $f(ba) = (ba)u = b f(a)$. Its image is an invariant subspace containing $f(1_A) = u \neq 0$, hence equals U by irreducibility. From $A = \sum_j L_j$ we get $U = \sum_j f(L_j)$, so $f(L_{j_0}) \neq \{0\}$ for some j_0. Then $f|_{L_{j_0}} \colon L_{j_0} \to U$ is a nonzero homomorphism between irreducible A-spaces, hence an isomorphism by Theorem 23.4.1.

■

Now let $B_0 = M_{p_1}(F) \times \cdots \times M_{p_l}(F)$ be any product of matrix algebras, and let $C_j = F^{p_j}$ carry the action of B_0 through the j-th slot. Each C_j is irreducible: given $\mathbf{v} \neq \mathbf{0}$ in F^{p_j} and any $\mathbf{w}$, extend $\mathbf{v}$ to a basis (Theorem 1.6.5) and let $\mathbf{X}$ be the matrix sending $\mathbf{v} \mapsto \mathbf{w}$ and the other basis vectors to $\mathbf{0}$; then $M_{p_j}(F)\mathbf{v} = F^{p_j}$, so no proper nonzero invariant subspace exists. They are pairwise nonisomorphic: the idempotent $e_j \in B_0$ acts as the identity on C_j and as 0 on C_i for $i \neq j$, so an isomorphism $f \colon C_j \to C_i$ would give $f(\mathbf{v}) = f(e_j\mathbf{v}) = e_j f(\mathbf{v}) = \mathbf{0}$ for every $\mathbf{v}$, forcing $f = 0$. Finally, for each j and each column index c let $C_{j,c} \subseteq B_0$ consist

of the elements whose slots other than j are $\mathbf{O}$ and whose j-th slot has all columns but the c-th equal to $\mathbf{0}$. Each $C_{j,c}$ is a left ideal, the map $x \mapsto x_j \mathbf{e}_c$ is an isomorphism of B_0-spaces $C_{j,c} \to C_j$, and B_0 is the direct sum of the $C_{j,c}$. So by the Claim the irreducible B_0-spaces are, up to isomorphism, exactly $C_1, \dots, C_l$: there are l of them and their dimensions are $p_1, \dots, p_l$.

An isomorphism of algebras $A \to B_0$ turns B_0-spaces into A-spaces, preserving irreducibility and isomorphism. So if $A \cong M_{n_1}(F) \times \cdots \times M_{n_k}(F)$, then A has exactly k isomorphism classes of irreducible spaces and their dimensions are $n_1, \dots, n_k$. Both quantities are determined by the isomorphism class of A alone, so they do not depend on the choices made in Steps 1 and 2. This proves the theorem.

■

Remark. The case $k = 1$ says that a semisimple algebra with only one irreducible space, up to isomorphism, is a full matrix algebra. Burnside's theorem (Theorem 23.5.2) is the concrete counterpart of that sentence: there the algebra arrives already written as operators on an irreducible space, its hypothesis is irreducibility rather than semisimplicity, and its conclusion is the sharper equality $A = \mathrm{End}(V)$ in place of an isomorphism. The proof above does not use Theorem 23.5.2; what it borrows from §05 is the double commutant theorem.

Two hypotheses carried the proof, and it is worth saying again where. **Semisimplicity** was used twice: to split the regular representation in Step 1, and to apply the double commutant theorem in Step 4. **Algebraic closure** was used once, in the second half of Schur's lemma, to know that a homomorphism of an irreducible space to itself is a scalar. Drop algebraic closure and the diagonal slots become division algebras rather than F; that is the general theorem, quoted below and not proved.

Corollary 23.6.2. Counting Dimensions in the Wedderburn Form

In the situation of Theorem 23.6.1,

$$\dim_F A = n_1^2 + \cdots + n_k^2,$$

and the center $Z(A) = \{z \in A : za = az \text{ for all } a \in A\}$ has $\dim_F Z(A) = k$.

Proof. The dimension of a product is the sum of the dimensions (Theorem 4.4.1) and $\dim M_{n_i}(F) = n_i^2$, which gives the first formula. An isomorphism of algebras carries the center onto the center, so we may compute in $M_{n_1}(F) \times \cdots \times M_{n_k}(F)$. Multiplication is slot by slot, so $(z_1, \dots, z_k)$ is central exactly when each z_i is central in $M_{n_i}(F)$, and by Proposition 23.1.5, read through the identification of $M_{n_i}(F)$ with $\mathrm{End}(F^{n_i})$, that means $z_i = \lambda_i \mathbf{I}_{n_i}$. So $Z(A)$ is isomorphic to F^k as a vector space, of dimension k.

■

The two numbers in the corollary are the whole invariant content of a semisimple algebra over an algebraically closed field: k is read from the center, and then the n_i are k positive integers with $\sum n_i^2 = \dim A$. When A is the group algebra $\mathbb{C}[G]$ of a finite group, whose dimension is

$|G|$ by Definition 23.1.5, this becomes the sum-of-squares formula $|G| = \sum_i n_i^2$ for the dimensions of the irreducible representations, and k turns out to be the number of conjugacy classes. Both statements are proved later in this chapter, once Maschke's theorem supplies the missing hypothesis that $\mathbb{C}[G]$ is semisimple.

Quick check. A semisimple algebra A over $\mathbb{C}$ has $\dim A = 10$ and $\dim Z(A) = 4$. What are the possible block sizes?

Solution. By Corollary 23.6.2 we need $k = 4$ positive integers with $n_1^2 + \cdots + n_4^2 = 10$. No n_i can be 3 or more: that block alone contributes at least 9, and the other three contribute at least 3, for a total of at least 12. So every n_i is 1 or 2; if t of them equal 2, the sum is $4t + (4 - t) = 10$, so $t = 2$. Up to order the block sizes are $1, 1, 2, 2$ and $A \cong \mathbb{C} \times \mathbb{C} \times M_2(\mathbb{C}) \times M_2(\mathbb{C})$.

23.6.4 What is quoted and not proved

The theorem above is the case this book can carry. The general statement is older and stronger, and we record it honestly.

Remark. **The Wedderburn–Artin theorem.** Over an arbitrary field F, a finite-dimensional semisimple F-algebra is isomorphic to a product

$$M_{n_1}(D_1) \times \cdots \times M_{n_k}(D_k),$$

where each D_i is a **division algebra** over F in the sense of §04 — an F-algebra, here finite-dimensional and not assumed commutative, in which every nonzero element has a multiplicative inverse. The result is due to Wedderburn, and Artin extended it beyond finite dimension. **We do not prove it.** The reason is not that the argument is long — it follows the same route, with Schur's lemma supplying $\mathrm{End}_A(V_i) = D_i$ in place of F — but that this book has never built the theory of division rings. The only noncommutative division algebra it has met is the quaternions $\mathbb{H}$ over $\mathbb{R}$ (Definition 14.12.3), and one example is not a theory. Nothing in this chapter depends on the general statement.

Remark. **Wedderburn's little theorem.** Every **finite** division ring is commutative, hence a field. This is a separate theorem of Wedderburn's, and we do not prove it either; the standard proofs use counting arguments in group theory or cyclotomic polynomials, neither of which this book has developed. Its effect on the theorem above is that over a finite field the division algebras D_i are all fields, so the blocks are matrix algebras over field extensions of F.

Warning. **Semisimplicity is not automatic, so the theorem does not apply to every algebra.** The upper triangular matrices in $M_2(F)$ form a three-dimensional algebra which is **not** a product of matrix algebras: a product $M_{n_1}(F) \times \cdots \times M_{n_k}(F)$ of total dimension 3 would need $\sum n_i^2 = 3$, hence $k = 3$ and all $n_i = 1$, hence a commutative algebra; but

$\mathbf{E}_{11}\mathbf{E}_{12} = \mathbf{E}_{12}$ while $\mathbf{E}_{12}\mathbf{E}_{11} = \mathbf{O}$. What fails is semisimplicity: as §05 noted, $\operatorname{span}(\mathbf{e}_1)$ is an invariant subspace with no invariant complement. §03 identifies the obstruction as the radical of the algebra (Definition 23.3.5).

23.6.5 Two worked decompositions

Example. The group algebra of a cyclic group of order four

Let $C_4 = \{1, g, g^2, g^3\}$ be a cyclic group of order 4, so $g^4 = 1$; it exists, for instance as the subgroup $\{1, i, -1, -i\}$ of $(\mathbb{C} \setminus \{0\}, \cdot)$ with $g = i$ (Theorem 0.10.2). Find the Wedderburn form of the group algebra $\mathbb{C}[C_4]$ (Definition 23.1.5) explicitly, and exhibit the idempotents $e_1, \dots, e_4$ of the product.

Solution. Let $\omega_1 = 1$, $\omega_2 = i$, $\omega_3 = -1$, $\omega_4 = -i$ be the four complex numbers with $\omega^4 = 1$. Define

$$\Phi\colon \mathbb{C}[C_4] \to \mathbb{C}^4, \qquad \Phi\Big(\sum_{t=0}^{3} a_t g^t\Big) = \Big(\sum_t a_t \omega_r^t\Big)_{r=1}^{4}.$$

It is linear by construction. It is multiplicative: it is enough to check on the basis g^s, g^t, where $g^s g^t = g^{s+t}$ with the exponent read modulo 4, and $\omega_r^{s+t} = \omega_r^s \omega_r^t$ with $\omega_r^4 = 1$ making the reduction harmless. It sends $1 = g^0$ to $(1,1,1,1)$. Its matrix in the basis $(1, g, g^2, g^3)$ of the source and the standard basis of $\mathbb{C}^4$ is

$$\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & i & -1 & -i \\ 1 & -1 & 1 & -1 \\ 1 & -i & -1 & i \end{pmatrix},$$

whose determinant is $-16i \neq 0$, so Φ is bijective. Hence

$$\mathbb{C}[C_4] \cong \mathbb{C} \times \mathbb{C} \times \mathbb{C} \times \mathbb{C} = M_1(\mathbb{C})^4,$$

with $k = 4$ and $n_1 = n_2 = n_3 = n_4 = 1$; the dimension count $\sum n_i^2 = 4 = |C_4|$ checks out, and $Z(\mathbb{C}[C_4])$ is everything, of dimension 4, as it must be for a commutative algebra.

The idempotents pull back to

$$e_r = \tfrac{1}{4} \sum_{t=0}^{3} \overline{\omega_r^t}\, g^t, \qquad r = 1, \dots, 4,$$

explicitly

$$e_1 = \tfrac{1}{4}(1 + g + g^2 + g^3), \quad e_2 = \tfrac{1}{4}(1 - ig - g^2 + ig^3),$$
$$e_3 = \tfrac{1}{4}(1 - g + g^2 - g^3), \quad e_4 = \tfrac{1}{4}(1 + ig - g^2 - ig^3).$$

A direct expansion confirms $e_r^2 = e_r$, $e_r e_s = 0$ for $r \neq s$, and $e_1 + e_2 + e_3 + e_4 = 1$. For instance

$$\begin{aligned} 16\, e_2^2 &= (1 - ig - g^2 + ig^3)^2 \\ &= 4 - 4ig - 4g^2 + 4ig^3 = 16\, e_2, \end{aligned}$$

where the middle equality collects the sixteen products $(\pm 1, \pm i) g^{s+t}$ with $g^4 = 1$.

The matrix in that example is, up to the factor $\frac{1}{2}$ that normalizes its columns, the Fourier matrix of size four (Example (The Fourier matrix of size four), Theorem 12.9.3). That is not a coincidence, and a later section of this chapter explains it: the characters of a finite abelian group assemble into a Fourier matrix, and the isomorphism above is the discrete Fourier transform.

Example. An algebra already in Wedderburn form

What is the Wedderburn form of $M_2(\mathbb{C})$, and what are its irreducible spaces?

Solution. It is $M_2(\mathbb{C})$ itself: $k = 1$ and $n_1 = 2$, with $\dim = 4 = 2^2$ and $\dim Z = 1$, the scalars (Proposition 23.1.5). To see that the theorem applies, note that $M_2(\mathbb{C}) = C_1 \oplus C_2$ where C_c is the set of matrices whose only nonzero column is the c-th; each C_c is a left ideal isomorphic to $\mathbb{C}^2$ as an $M_2(\mathbb{C})$-space, and $\mathbb{C}^2$ is irreducible, so Theorem 23.3.3 makes the algebra semisimple. There is exactly one isomorphism class of irreducible space, namely $\mathbb{C}^2$, of dimension 2 — which is Step 6's description of k and n_1. Note that $M_2(\mathbb{C})$ is **not** a product of two smaller algebras: it is simple (Theorem 23.2.2), while a product with $k \geq 2$ has the proper nonzero two-sided ideal $e_1 A$.

23.6.6 Exercises

23.6.6.1 A. Check your understanding

Exercise. A1

(a) State Theorem 23.6.1 with all of its hypotheses, and say which hypothesis fails for the upper triangular matrices in $M_2(\mathbb{C})$.

(b) Where in the proof is algebraic closure used, and what replaces the conclusion when it is dropped?

(c) A semisimple algebra over $\mathbb{C}$ has dimension 5. List all possibilities for its Wedderburn form.

(d) True or false, with a reason: a commutative semisimple algebra over $\mathbb{C}$ of dimension n is isomorphic to $\mathbb{C}^n$.

Solution.

(a) F algebraically closed, $A \neq \{0\}$ finite-dimensional over F, and A semisimple, that is, every left ideal has a complementary left ideal. For the upper triangular matrices semisimplicity fails: $\operatorname{span}(\mathbf{e}_1)$ is an invariant subspace of the natural action with no invariant complement, and correspondingly the left ideal of matrices with second row zero has no complementary left ideal.

(b) In Step 4, in the second half of Theorem 23.4.1, to know that s_{ii} is a scalar. Without algebraic closure the slots become $M_{n_i}(D_i)$ for division algebras D_i, which is the Wedderburn–Artin theorem quoted above and not proved here.

(c) We need positive integers with $\sum n_i^2 = 5$: either $1 + 4$, giving $\mathbb{C} \times M_2(\mathbb{C})$, or $1 + 1 + 1 + 1 + 1$, giving $\mathbb{C}^5$.

(d) True. By Theorem 23.6.1, $A \cong \prod_i M_{n_i}(\mathbb{C})$, and $M_{n_i}(\mathbb{C})$ is commutative only for $n_i = 1$ (otherwise $\mathbf{E}_{11}\mathbf{E}_{12} \neq \mathbf{E}_{12}\mathbf{E}_{11}$). So every $n_i = 1$ and $k = \sum n_i^2 = n$.

23.6.6.2 B. Practice

Exercise. B1: Determine which are semisimple

Determine which of the following algebras over $\mathbb{C}$ are semisimple, and give the Wedderburn form of those that are. Justify your answer.

(a) The diagonal matrices in $M_3(\mathbb{C})$.
(b) $\mathbb{C}[x]/(x^2)$.
(c) $\mathbb{C}[x]/(x^2 - 1)$.

Solution.

(a) Semisimple, and equal to $\mathbb{C} \times \mathbb{C} \times \mathbb{C}$ by Example (Three products, seen from two sides) (i). As a check on semisimplicity: the algebra is the direct sum of the three left ideals $\mathbb{C}\mathbf{E}_{ii}$, each of dimension 1 and hence irreducible, so Theorem 23.3.3 applies.

(b) Not semisimple. Both algebras here are quotients of the infinite-dimensional algebra $\mathbb{C}[x]$ by the ideal generated by one polynomial (Definition 6.2.1, Definition 23.2.2), and each has dimension 2, a basis being the classes of 1 and x. Write ε for the class of x in $\mathbb{C}[x]/(x^2)$, so the algebra is $\{a + b\varepsilon\}$ with $\varepsilon^2 = 0$. The line $\mathbb{C}\varepsilon$ is a left ideal, and it is the **only** one-dimensional one: a left ideal spanned by $a + b\varepsilon$ with $a \neq 0$ contains $\varepsilon(a+b\varepsilon) = a\varepsilon$, hence ε, hence is at least two-dimensional. So $\mathbb{C}\varepsilon$ has no complementary left ideal, which would have to be one-dimensional. Consistently, a product of matrix algebras of dimension 2 would be $\mathbb{C} \times \mathbb{C}$, which has no nonzero element squaring to 0.

(c) Semisimple, with Wedderburn form $\mathbb{C} \times \mathbb{C}$. The map $p + (x^2 - 1) \mapsto (p(1), p(-1))$ is linear, multiplicative, and unital, and on the representatives $a+bx$ it reads $(a+b, a-b)$, which vanishes only for $a = b = 0$; an injective linear map between spaces of dimension 2 is bijective (Theorem 2.3.1). A product of copies of $\mathbb{C}$ is semisimple by the argument in (a).

Exercise. B2: A dimension count

Let A be a semisimple algebra over $\mathbb{C}$ with $\dim A = 8$ and $\dim Z(A) = 5$. Determine the Wedderburn form of A, and the dimensions of the irreducible A-spaces.

Solution. By Corollary 23.6.2, $k = \dim Z(A) = 5$ and $n_1^2 + \cdots + n_5^2 = 8$ with every $n_i \geq 1$. Writing $8 = 5 + 3$, we need $\sum(n_i^2 - 1) = 3$, and $n^2 - 1$ takes the values $0, 3, 8, \ldots$ for $n = 1, 2, 3, \ldots$. So exactly one n_i equals 2 and the rest equal 1. Hence

$$A \cong \mathbb{C} \times \mathbb{C} \times \mathbb{C} \times \mathbb{C} \times M_2(\mathbb{C}),$$

and the irreducible A-spaces have dimensions $1, 1, 1, 1, 2$.

Exercise. B3: The group algebra of a group of order three

Let $C_3 = \{1, g, g^2\}$ with $g^3 = 1$. Imitating Example (The group algebra of a cyclic group of order four), produce an explicit algebra isomorphism $\mathbb{C}[C_3] \to \mathbb{C}^3$ and write down the three idempotents.

Solution. Let $\omega = e^{2\pi i/3}$, so $1, \omega, \omega^2$ are the cube roots of 1 and $1 + \omega + \omega^2 = 0$. Define

$$\Phi\Big(\sum_{t=0}^{2} a_t g^t\Big) = \Big(\sum_t a_t,\ \sum_t a_t \omega^t,\ \sum_t a_t \omega^{2t}\Big).$$

As in the example, Φ is linear, multiplicative (because $\omega^{s+t} = \omega^s \omega^t$ and $\omega^3 = 1$), and unital. Its matrix has rows $(1,1,1)$, $(1,\omega,\omega^2)$, $(1,\omega^2,\omega)$; it is a Vandermonde matrix in the distinct nodes $1, \omega, \omega^2$, hence invertible. So $\mathbb{C}[C_3] \cong \mathbb{C}^3$, and $\sum n_i^2 = 3 = |C_3|$. The idempotents are

$$\begin{aligned} e_1 &= \tfrac{1}{3}(1 + g + g^2), \\ e_2 &= \tfrac{1}{3}(1 + \omega^2 g + \omega g^2), \\ e_3 &= \tfrac{1}{3}(1 + \omega g + \omega^2 g^2), \end{aligned}$$

the r-th being $\frac{1}{3}\sum_t \overline{\omega^{(r-1)t}} g^t$.

23.6.6.3 C. Going deeper

Exercise. C1: Simple plus semisimple

Let F be algebraically closed and let $A \neq \{0\}$ be a finite-dimensional semisimple F-algebra.

(a) Prove that A is simple (has no two-sided ideal but $\{0\}$ and A) if and only if $k = 1$, that is, $A \cong M_n(F)$ for some n.

(b) Deduce that a finite-dimensional simple semisimple algebra over $\mathbb{C}$ has square dimension.

Hint for (a): in a product, the set of elements whose slots after the first are zero is a two-sided ideal.

Solution.

(a) ($\Leftarrow$) If $k = 1$ then $A \cong M_n(F)$, which is simple by Theorem 23.2.2, and simplicity is preserved by an algebra isomorphism.
($\Rightarrow$) Suppose $k \geq 2$ and write $A \cong M_{n_1}(F) \times \cdots \times M_{n_k}(F)$. The set I of elements whose slots $2, \ldots, k$ are $\mathbf{O}$ is closed under addition and under multiplication by any element on either side, because multiplication is slot by slot; so I is a two-sided ideal. It is nonzero, since $e_1 \in I$, and it is not all of A, since $e_2 \notin I$. So A is not simple. Contrapositively, simple forces $k = 1$.

(b) By (a), $A \cong M_n(F)$, so $\dim A = n^2$ by Corollary 23.6.2 with $k = 1$.

Exercise. C2: How often each irreducible appears

Keep the notation of Theorem 23.6.1 and of its proof, so that $A = L_1 \oplus \cdots \oplus L_m$ with each L_j isomorphic to exactly one V_i. Let m_i be the number of indices j with $L_j \cong V_i$. For A-spaces V, W write $\mathrm{Hom}_A(V, W)$ for the vector space of homomorphisms of A-spaces $V \to W$.

(a) Prove that $f \mapsto f(1_A)$ is a linear bijection $\mathrm{Hom}_A(A, V) \to V$ for every A-space V, where A carries the regular representation.

(b) Prove that $\dim \mathrm{Hom}_A(L_j, V_i)$ is 1 if $L_j \cong V_i$ and 0 otherwise.

(c) Deduce that $m_i = n_i$: each irreducible appears in the regular representation exactly as many times as its dimension. Check the result against Corollary 23.6.2.

Hint for (b): Theorem 23.4.1, both parts.

Solution.

(a) The map is linear. It is injective: if $f(1_A) = 0$ then $f(a) = f(a1_A) = af(1_A) = 0$ for every a. It is surjective: given $v \in V$, the map $a \mapsto av$ is linear and satisfies $(ba)v = b(av)$, so it is a homomorphism of A-spaces sending 1_A to v.

(b) If $L_j \not\cong V_i$, the first part of Theorem 23.4.1 says every homomorphism between these irreducible spaces is 0, so the dimension is 0. If $L_j \cong V_i$, fix an isomorphism h; then $f \mapsto h^{-1}f$ is a linear bijection $\mathrm{Hom}_A(L_j, V_i) \to \mathrm{Hom}_A(L_j, L_j)$, and by the second part of Theorem 23.4.1, valid because F is algebraically closed, every self-homomorphism of the irreducible space L_j is a scalar. So the dimension is 1.

(c) A homomorphism out of a direct sum is the list of its restrictions to the summands, so $\mathrm{Hom}_A(A, V_i)$ is linearly isomorphic to $\bigoplus_j \mathrm{Hom}_A(L_j, V_i)$, of dimension m_i by (b). By (a) that dimension is $\dim V_i = n_i$. Hence $m_i = n_i$. As a check, $\dim A = \sum_j \dim L_j = \sum_i m_i n_i = \sum_i n_i^2$, which is Corollary 23.6.2.

Exercise. C3: Change the field

Let $A = \{x\mathbf{I}_2 + y\mathbf{Q} : x, y \in \mathbb{R}\} \subseteq M_2(\mathbb{R})$, where $\mathbf{Q} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$.

(a) Prove that A is a semisimple $\mathbb{R}$-algebra of dimension 2.

(b) Show that A is **not** isomorphic to a product $M_{n_1}(\mathbb{R}) \times \cdots \times M_{n_k}(\mathbb{R})$, and identify which hypothesis of Theorem 23.6.1 fails.

(c) Prove that A is itself a division algebra over $\mathbb{R}$ and identify it.

Solution.

(a) $\mathbf{Q}^2 = -\mathbf{I}_2$, so A is closed under multiplication and is a subalgebra containing $\mathbf{I}_2$, of dimension 2. Every nonzero element $x\mathbf{I}_2 + y\mathbf{Q}$ has determinant $x^2 + y^2 > 0$, hence is invertible in $M_2(\mathbb{R})$, and its inverse $(x^2 + y^2)^{-1}(x\mathbf{I}_2 - y\mathbf{Q})$ lies in A. So every nonzero left ideal contains an invertible element and hence equals A; the only left ideals are $\{\mathbf{O}\}$ and A, and each has a complementary left ideal (the other one). So A is semisimple.

(b) A product with $\sum n_i^2 = 2$ must be $\mathbb{R} \times \mathbb{R}$, which contains the nonzero element $(1, 0)$

with $(1,0)(0,1) = (0,0)$; but in A the product of two nonzero elements is nonzero, since both are invertible. So no such isomorphism exists. The hypothesis that fails is algebraic closure: $\mathbb{R}$ is not algebraically closed.

(c) By (a) every nonzero element of A is invertible, and its inverse again lies in A, so A is a division algebra over $\mathbb{R}$; it is a copy of $\mathbb{C}$, the map $x\mathbf{I}_2 + y\mathbf{Q} \mapsto x + yi$ being a bijective algebra homomorphism. This is the D_1 that would appear in the quoted Wedderburn–Artin form of A.

23.7 Group Representations

Sections 01–06 studied an algebra of operators in the abstract. A group is not an algebra — its elements cannot be added — but a group can be made to *act* by operators, and then the operators it produces span an algebra. This section sets up that translation: a representation of a group G over F is the same thing as a representation of the group algebra $F[G]$, so everything proved about algebras applies at once. The section's main theorem, Maschke's, says when the resulting algebra is semisimple, and its hypothesis on the field is not decoration: we exhibit a group and a field where the conclusion fails outright.

Throughout, G is a **finite** group, written multiplicatively, with its identity element written 1, and F is a field. All vector spaces are finite-dimensional over F. Where a statement needs F to be $\mathbb{C}$, or needs a condition on $\operatorname{char} F$, the statement says so.

23.7.1 Letting a group act by operators

The book's groups so far have been groups *of* matrices or *of* permutations: $\mathrm{GL}_n(F)$ (Example (Groups)), the symmetric group S_n, the classical groups of Chapter 14. Such a group arrives with its linear algebra attached. An abstract group does not. If all we know about G is its multiplication table, the only way to bring linear algebra to bear is to manufacture operators out of it.

The manufacture has one requirement. Group multiplication must become composition of operators, since composition is the only product on $\mathcal{L}(V)$, and every group element must become an **invertible** operator, since every group element can be undone. That is exactly the definition of a homomorphism into the group $\mathrm{GL}(V)$ of invertible operators on V, a group under composition by the same argument that makes $\mathrm{GL}_n(F)$ one (Example (Groups)).

A representation is a way of realizing the elements of a group as invertible operators, so that multiplying in the group means composing the operators.

Definition 23.7.1. Representation of a group

Let G be a group, F a field, and V a finite-dimensional vector space over F. A **representation** of G on V over F is a group homomorphism

$$\rho\colon G \to \mathrm{GL}(V),$$

that is, a map assigning to each $g \in G$ an invertible operator $\rho(g)$ on V such that $\rho(gh) = \rho(g)\rho(h)$ **for all** $g, h \in G$. The space V is the **representation space**, and $\dim V$ is the **degree** of ρ. The representation is **faithful** if ρ is injective. When $V = F^n$ we speak of a **matrix representation** and write $\rho(g) \in \mathrm{GL}_n(F)$.

In words: to give a representation is to attach to each $g \in G$ an invertible operator $\rho(g)$ on one fixed space V, in such a way that the operator attached to a product is the composition of the operators attached to the factors. Two conditions that look like extra demands are already implied by Theorem 0.10.3: $\rho(1) = \mathrm{id}_V$, and $\rho(g^{-1}) = \rho(g)^{-1}$ for every g. So we never have to check them.

Note what is **not** required. Nothing says ρ is injective, and nothing says the operators $\rho(g)$ are distinct. The constant map $g \mapsto \mathrm{id}_V$ is a perfectly good representation; it simply carries no information about G.

Examples. Fix a finite group G and a field F.

- **The trivial representation.** $V = F$ and $\rho(g) = \mathrm{id}_F$ for every g. It is a representation because $\mathrm{id}_F \cdot \mathrm{id}_F = \mathrm{id}_F$, matching $\rho(gh) = \rho(g)\rho(h)$. Its degree is 1. This is the degenerate case, and it is not a curiosity: it is one of the basic building blocks, and in Section 08 it is the piece that detects averages.
- **The sign representation of S_n.** $V = F$ and $\rho(\sigma) = \mathrm{sgn}(\sigma)\ \mathrm{id}_F$, where sgn is the sign of a permutation (Definition 7.3.2). It is a homomorphism because $\mathrm{sgn}(\sigma\tau) = \mathrm{sgn}(\sigma)\,\mathrm{sgn}(\tau)$ (Theorem 7.3.1). Over a field of characteristic 2 we have $-1 = 1$, so the sign representation coincides with the trivial one; over $\mathbb{C}$ it does not. It is faithful on S_2, whose two elements have signs 1 and -1, but not on S_n for $n \geq 3$, where $\mathrm{sgn}((1\ 2\ 3)) = 1 = \mathrm{sgn}(\mathrm{id})$ while $(1\ 2\ 3) \neq \mathrm{id}$.
- **The permutation representation of S_n.** $V = F^n$ and $\rho(\sigma)\mathbf{e}_i = \mathbf{e}_{\sigma(i)}$, extended linearly (Theorem 2.1.6). Then $\rho(\sigma)\rho(\tau)\mathbf{e}_i = \rho(\sigma)\mathbf{e}_{\tau(i)} = \mathbf{e}_{\sigma(\tau(i))} = \rho(\sigma\tau)\mathbf{e}_i$, where the last step uses that $\sigma\tau$ means "first τ, then σ". So ρ is a homomorphism, and each $\rho(\sigma)$ is invertible with inverse $\rho(\sigma^{-1})$. The matrix of $\rho(\sigma)$ in the standard basis is the permutation matrix $\mathbf{P}_\sigma$ of Definition 3.4.2. The same recipe works for any subgroup of S_n.
- **The rotation representation of $\mathbb{Z}/n\mathbb{Z}$ over $\mathbb{R}$.** $V = \mathbb{R}^2$ and $\rho([k])$ is rotation by $2\pi k/n$, with matrix $\begin{pmatrix} \cos(2\pi k/n) & -\sin(2\pi k/n) \\ \sin(2\pi k/n) & \cos(2\pi k/n) \end{pmatrix}$. Composing rotations adds angles, so $\rho([k])\rho([l]) = \rho([k+l])$; and rotating by $2\pi n/n = 2\pi$ is the identity, so the recipe does not depend on the representative k. For $n \geq 3$ this representation has no invariant line. A line invariant under $\rho([1])$ would give that matrix a real eigenvalue (Proposition 9.1.2), but with $c = \cos(2\pi/n)$ and $s = \sin(2\pi/n)$ its characteristic polynomial is $x^2 - 2cx + 1$, of discriminant $4c^2 - 4 = -4s^2$, which is negative because $0 < 2\pi/n < \pi$ forces $s \neq 0$ when $n \geq 3$. (The quarter-turn case is Example (A Rotation Has No Invariant Lines).) Section 04 turned this into the standard warning about Schur's lemma over $\mathbb{R}$.
- **The regular representation.** $V = F[G]$, the group algebra of Definition 23.1.5: the free vector space on the set G (Definition 15.2.3), with basis $\{\delta_g : g \in G\}$ and multiplication determined by $\delta_g\delta_h = \delta_{gh}$. Its dimension is $|G|$. Put $\rho_{\mathrm{reg}}(g)\delta_h = \delta_{gh}$, extended linearly. Then $\rho_{\mathrm{reg}}(g)\rho_{\mathrm{reg}}(g')\delta_h = \delta_{gg'h} = \rho_{\mathrm{reg}}(gg')\delta_h$, so ρ_{reg} is a homomorphism, and $\rho_{\mathrm{reg}}(g)$ is invertible with inverse $\rho_{\mathrm{reg}}(g^{-1})$. It is faithful: if $\rho_{\mathrm{reg}}(g) = \mathrm{id}$ then $\delta_g = \rho_{\mathrm{reg}}(g)\delta_1 = \delta_1$, so $g = 1$ because the δ_h are distinct basis vectors. Its degree is $|G|$.

Nonexample by minimal change. Keep $V = F^2$ and $G = \mathbb{Z}/4\mathbb{Z}$, and try $\rho([k]) = \begin{pmatrix} 1 & k \\ 0 & 1 \end{pmatrix}$ over $\mathbb{Q}$. Each matrix is invertible, and $\rho([k])\rho([l]) = \begin{pmatrix} 1 & k+l \\ 0 & 1 \end{pmatrix}$, so multiplicativity holds on representatives. But the recipe is not **well defined** on $\mathbb{Z}/4\mathbb{Z}$: the class $[0]$ is also $[4]$, and $\begin{pmatrix} 1 & 4 \\ 0 & 1 \end{pmatrix} \neq \mathbf{I}_2$. The clause that fails is that ρ must be a function on G. Over $\mathbb{F}_2$ the same recipe

on $\mathbb{Z}/2\mathbb{Z}$ *is* well defined, because $\begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix} = \mathbf{I}_2$ there; we return to that example when Maschke's theorem breaks.

Warning. **A representation need not remember the group.** The trivial representation of any group is a representation, and it forgets everything. Conversely, a noninjective ρ can still be useful: the sign representation of S_n has a huge kernel yet separates even from odd permutations. When a statement needs ρ to see all of G, it must say **faithful**; none of the theorems below does.

Quick check. Is $\rho\colon \mathbb{Z}/3\mathbb{Z} \to \mathrm{GL}_1(\mathbb{R}) = \mathbb{R}\setminus\{0\}$, $\rho([k]) = (-1)^k$, a representation?

Solution. No: it is not well defined. We have $[0] = [3]$ in $\mathbb{Z}/3\mathbb{Z}$, but $(-1)^0 = 1$ while $(-1)^3 = -1$. Equivalently, a representation of degree 1 is a homomorphism $\mathbb{Z}/3\mathbb{Z} \to \mathbb{R}\setminus\{0\}$, and $\rho([1])^3 = \rho([0]) = 1$ forces $\rho([1]) = 1$, since 1 is the only real cube root of 1. So the only degree-1 real representation of $\mathbb{Z}/3\mathbb{Z}$ is the trivial one.

23.7.2 Subrepresentations, equivalence, and irreducibility

Having made operators, we ask the question this book always asks of a family of operators: can the space be broken into smaller pieces that the family preserves? Chapter 9 called such a piece an invariant subspace (Definition 9.1.1). Here the family is $\{\rho(g) : g \in G\}$.

Definition 23.7.2. Subrepresentation

Let ρ be a representation of G on V. A subspace $U \subseteq V$ is **G-invariant** if $\rho(g)u \in U$ **for every** $g \in G$ and every $u \in U$. In that case the restrictions $\rho(g)|_U$ define a representation $\rho|_U$ of G on U, the **subrepresentation** of ρ on U.

Two checks make this legitimate. First, $\rho(g)|_U$ really is an operator **on** U, by invariance. Second, it is invertible on U: applying invariance to g^{-1} gives $\rho(g)^{-1}U = \rho(g^{-1})U \subseteq U$, so $\rho(g)|_U$ has a two-sided inverse on U, namely $\rho(g^{-1})|_U$. Multiplicativity is inherited. The subspaces $\{0\}$ and V are always G-invariant; we call them the **trivial** invariant subspaces.

Definition 23.7.3. Equivalence of representations

Let ρ and ρ' be representations of G on V and V', over the same field. A linear map $S\colon V \to V'$ **intertwines** ρ and ρ' if

$$S\,\rho(g) = \rho'(g)\,S \qquad \text{for every } g \in G.$$

The representations are **equivalent**, written $\rho \cong \rho'$, if some intertwining S is an isomorphism; equivalently, if $\rho'(g) = S\rho(g)S^{-1}$ for all g.

Equivalence is what "the same representation written in another basis" means. Indeed, if $V = V' = F^n$ and S is an invertible matrix, then $\rho'(g) = S\rho(g)S^{-1}$ is simultaneous similarity of the whole family $\{\rho(g)\}$ by one matrix. It is an equivalence relation: id_V intertwines ρ with itself; if S intertwines ρ with ρ' and is invertible then S^{-1} intertwines ρ' with ρ, since $S\rho(g) = \rho'(g)S$ gives $\rho(g)S^{-1} = S^{-1}\rho'(g)$; and a composition of intertwining isomorphisms intertwines.

Definition 23.7.4. Irreducible representation

A representation ρ of G on V is **irreducible** if $V \neq \{0\}$ and the only G-invariant subspaces of V are $\{0\}$ and V. A representation with $V \neq \{0\}$ that is not irreducible is **reducible**.

Read the two clauses separately. The zero space is excluded by decree, exactly as 1 is excluded from the primes: keeping it would break every statement of the form "every representation is built from irreducible ones". The second clause is the real content.

Examples and a nonexample.

- Every representation of degree 1 is irreducible: a one-dimensional space has no subspaces but $\{0\}$ and itself.
- The permutation representation of S_3 on F^3 is **reducible** for every F. The line $L = \mathrm{span}(\mathbf{e}_1+\mathbf{e}_2+\mathbf{e}_3)$ is S_3-invariant, since $\rho(\sigma)(\mathbf{e}_1+\mathbf{e}_2+\mathbf{e}_3) = \mathbf{e}_{\sigma(1)}+\mathbf{e}_{\sigma(2)}+\mathbf{e}_{\sigma(3)} = \mathbf{e}_1+\mathbf{e}_2+\mathbf{e}_3$, the same three basis vectors in another order. And $L \neq \{\mathbf{0}\}$, $L \neq F^3$.
- So is the **sum-zero** subspace $W = \{\mathbf{x} \in F^3 : x_1+x_2+x_3 = 0\}$, of dimension 2: permuting the coordinates of $\mathbf{x}$ does not change their sum. We will see below that over $\mathbb{C}$ the subrepresentation on W is irreducible; this is the **standard representation** of S_3.
- **Nonexample by minimal change.** Replace S_3 by the subgroup $H = \{\mathrm{id}, (1\ 2)\} \subseteq S_3$ and keep $V = F^3$. Then $\mathrm{span}(\mathbf{e}_3)$ is H-invariant, though it is not S_3-invariant. Invariance depends on the whole group, not on one element of it.

Finally, the operation that puts representations together. If ρ_1, ρ_2 are representations of G on V_1, V_2, their **direct sum** is the representation $\rho_1 \oplus \rho_2$ on $V_1 \oplus V_2$ given by $(\rho_1 \oplus \rho_2)(g)(v_1, v_2) = (\rho_1(g)v_1, \rho_2(g)v_2)$. In a basis adapted to the decomposition its matrix is block diagonal, $\rho_1(g) \oplus \rho_2(g)$. Conversely, if $V = U \oplus W$ with both summands G-invariant, then the map $U \oplus W \to V$, $(u, w) \mapsto u + w$, is an isomorphism intertwining $\rho|_U \oplus \rho|_W$ with ρ, so $\rho \cong \rho|_U \oplus \rho|_W$. So "decomposing a representation" and "finding an invariant complement" are the same task.

23.7.3 Representations of G and representations of $F[G]$

Here is the bridge that puts Sections 01–06 to work. The group algebra $F[G]$ was built in Definition 23.1.5 precisely so that the next theorem is true.

Theorem 23.7.1. Representations of G are representations of $F[G]$

Let G be a finite group, F a field, and V a finite-dimensional vector space over F. The assignment

$$\rho \longmapsto \widetilde{\rho}, \qquad \widetilde{\rho}\Big(\sum_{g\in G} a_g \delta_g\Big) := \sum_{g\in G} a_g\, \rho(g),$$

is a bijection from the set of representations $\rho\colon G \to \mathrm{GL}(V)$ onto the set of algebra homomorphisms $F[G] \to \mathrm{End}(V)$ (Definition 23.1.2) — that is, onto the ways of making V an

$F[G]$-space in the sense of Definition 23.3.2. Moreover a subspace $U \subseteq V$ is G-invariant for ρ if and only if $\widetilde{\rho}(a)U \subseteq U$ for every $a \in F[G]$.

Idea. An element of $F[G]$ is a formal linear combination of group elements, so a linear map out of $F[G]$ is determined by, and may be prescribed freely on, the basis $\{\delta_g\}$. The only thing to check is that "multiplicative on the basis" upgrades to "multiplicative everywhere", and that is bilinearity of both products. The inverse map recovers ρ by restricting to the basis; the one point needing care is why $\widetilde{\rho}(\delta_g)$ is automatically **invertible**, which is where $\delta_g\delta_{g^{-1}} = \delta_1$ earns its keep.

Proof. Let $\rho\colon G \to \mathrm{GL}(V)$ be a representation. The formula defines a linear map $\widetilde{\rho}\colon F[G] \to \mathrm{End}(V)$, because $\{\delta_g\}$ is a basis of $F[G]$ and a linear map may be prescribed on a basis (Theorem 2.1.6). It sends δ_1 to $\rho(1) = \mathrm{id}_V$ (Theorem 0.10.3). For multiplicativity, take $a = \sum_g a_g\delta_g$ and $b = \sum_h b_h\delta_h$. Multiplication in $F[G]$ and composition in $\mathrm{End}(V)$ are both bilinear, so

$$\begin{aligned}\widetilde{\rho}(ab) &= \widetilde{\rho}\Big(\sum_{g,h} a_g b_h\,\delta_{gh}\Big) = \sum_{g,h} a_g b_h\,\rho(gh)\\ &= \sum_{g,h} a_g b_h\,\rho(g)\rho(h) = \Big(\sum_g a_g\rho(g)\Big)\Big(\sum_h b_h\rho(h)\Big) = \widetilde{\rho}(a)\widetilde{\rho}(b),\end{aligned}$$

where the third equality is the homomorphism property of ρ. Hence $\widetilde{\rho}$ is an algebra homomorphism.

Conversely, let $\varphi\colon F[G] \to \mathrm{End}(V)$ be an algebra homomorphism, and set $\rho(g) := \varphi(\delta_g)$. Since δ_1 is the identity of $F[G]$ (Definition 23.1.5) and id_V is the identity of $\mathrm{End}(V)$, the unital clause of Definition 23.1.2 gives $\varphi(\delta_1) = \mathrm{id}_V$. Then $\rho(g)\rho(h) = \varphi(\delta_g\delta_h) = \varphi(\delta_{gh}) = \rho(gh)$. In particular $\rho(g)\rho(g^{-1}) = \varphi(\delta_1) = \mathrm{id}_V$ and likewise in the other order, so each $\rho(g)$ is invertible and ρ maps into $\mathrm{GL}(V)$. Thus ρ is a representation, and $\widetilde{\rho} = \varphi$ because the two agree on the basis $\{\delta_g\}$ and both are linear. Since also $\widetilde{\rho}(\delta_g) = \rho(g)$, the two constructions are mutually inverse, and the assignment is a bijection.

For the last sentence, suppose U is G-invariant. For $a = \sum_g a_g\delta_g$ and $u \in U$, each $\rho(g)u$ lies in U, and U is a subspace, so $\widetilde{\rho}(a)u = \sum_g a_g\rho(g)u \in U$. Conversely, if $\widetilde{\rho}(a)U \subseteq U$ for all a, take $a = \delta_g$. This proves the theorem. ∎

So the two languages are interchangeable, and we use whichever is shorter. A G-invariant subspace is an $F[G]$-invariant subspace; an irreducible representation of G on V is exactly a simple $F[G]$-space in the sense of Definition 23.3.3; and a map intertwining two representations of G is exactly an $F[G]$-map in the sense of Definition 23.4.1, since the condition $S\rho(g) = \rho'(g)S$ for all g extends by linearity to all of $F[G]$. In particular Schur's lemma (Theorem 23.4.1) applies verbatim to irreducible representations of groups, and we use it that way in Section 08.

Remark. Working with $F[G]$ rather than with G buys one thing that matters: the ability to **add** group elements. Every averaging argument below lives in $F[G]$, not in G. Chapter 10 §07 made the same move for a single operator, replacing T by the algebra $F[T]$ and the

space by an $F[x]$-module.

23.7.4 Maschke's theorem

We now know that a reducible representation has a proper nonzero invariant subspace U. To break V into pieces we need more: an invariant **complement**, a G-invariant W with $V = U \oplus W$. An ordinary complement always exists (Theorem 1.7.6), but it has no reason to be invariant.

The idea for repairing it is the one move this chapter uses over and over. Start with an object that is not invariant, and average its G-translates. The average is invariant because translating it permutes the terms of the sum. Averaging divides by $|G|$, and that is the whole hypothesis.

Lemma 23.7.2. When the group order is invertible

Let F be a field and m a positive integer. Then $m \cdot 1_F \neq 0$ in F if and only if $\operatorname{char} F$ does not divide m.

Proof. If $\operatorname{char} F = 0$, then $m \cdot 1_F \neq 0$ for every positive integer m (Definition 0.7.2), and 0 divides no positive integer, so both sides of the equivalence hold. Assume from now on that $p := \operatorname{char} F > 0$, so p is the **smallest** positive integer with $p \cdot 1_F = 0$.

($\Rightarrow$) We prove the contrapositive: if $p \mid m$, then $m \cdot 1_F = 0$. Write $m = qp$ with $q \geq 1$. By Lemma 0.7.4, $m \cdot 1_F = (qp) \cdot 1_F = (q \cdot 1_F)(p \cdot 1_F) = (q \cdot 1_F) \cdot 0 = 0$.

($\Leftarrow$) Suppose, for a contradiction, that some positive integer m has $m \cdot 1_F = 0$ and $p \nmid m$. Choose such an m as small as possible. Now $m \neq p$, since $p \mid p$, and $m > p$, since p is the smallest positive integer killing 1_F. So $m - p$ is a positive integer, and by Lemma 0.7.4

$$(m - p) \cdot 1_F + p \cdot 1_F = m \cdot 1_F = 0,$$

whence $(m - p) \cdot 1_F = 0$. Also $p \nmid m - p$, for otherwise p would divide $(m - p) + p = m$. This contradicts the minimality of m. This proves the lemma. ∎

Theorem 23.7.3. Maschke's Theorem

Let G be a **finite** group and F a field whose characteristic does **not** divide $|G|$. Let ρ be a representation of G on a finite-dimensional space V over F, and let $U \subseteq V$ be a G-invariant subspace. Then there is a G-invariant subspace $W \subseteq V$ with

$$V = U \oplus W.$$

Idea. Pick **any** projection P of V onto U; its kernel is a complement, but the wrong one, because P need not commute with the $\rho(g)$. Repair it by averaging the conjugates $\rho(g) P \rho(g)^{-1}$ over $g \in G$. ① Each conjugate is again a projection onto U, because U is invariant. ② Their average P_0 is therefore again a projection onto U — here we divide by $|G|$. ③ Conjugating P_0 by $\rho(h)$ only permutes the terms of the sum, so P_0 commutes with every $\rho(h)$, and then its kernel is invariant.

Proof. Write $n = |G|$. By hypothesis and Lemma 23.7.2, $n \cdot 1_F \neq 0$, so the scalar $\frac{1}{n} := (n \cdot 1_F)^{-1}$ exists in F. By Theorem 1.7.6 there is a subspace W_0 with $V = U \oplus W_0$; let $P \in \mathcal{L}(V)$ be the projection onto U along W_0 (Theorem 4.5.1 (b)), so $\operatorname{im} P = U$ and $Pu = u$ for all $u \in U$. Define

$$P_0 := \frac{1}{n} \sum_{g \in G} \rho(g)\, P\, \rho(g)^{-1} \ \in \mathcal{L}(V).$$

Step 1: $\operatorname{im} P_0 \subseteq U$. Let $v \in V$ and $g \in G$. Then $P\rho(g)^{-1}v \in \operatorname{im} P = U$, and U is G-invariant, so $\rho(g)P\rho(g)^{-1}v \in U$. A subspace is closed under sums and scalar multiples, so $P_0 v \in U$.

Step 2: $P_0 u = u$ for $u \in U$. Let $u \in U$ and $g \in G$. Since U is G-invariant and $\rho(g)^{-1} = \rho(g^{-1})$, we have $\rho(g)^{-1}u \in U$, so $P\rho(g)^{-1}u = \rho(g)^{-1}u$ and hence $\rho(g)P\rho(g)^{-1}u = u$. Summing over the n elements of G gives $\sum_g \rho(g)P\rho(g)^{-1}u = (n \cdot 1_F)u$, and multiplying by $\frac{1}{n}$ gives $P_0 u = u$.

By Steps 1 and 2, $P_0 v \in U$ for every v, and P_0 fixes U pointwise; therefore $P_0^2 v = P_0(P_0 v) = P_0 v$, so P_0 is a projection (Definition 4.5.1) with $\operatorname{im} P_0 = U$.

Step 3: P_0 commutes with every $\rho(h)$. Fix $h \in G$. Then

$$\rho(h)P_0\rho(h)^{-1} = \frac{1}{n} \sum_{g \in G} \rho(hg)\, P\, \rho(hg)^{-1},$$

using $\rho(h)\rho(g) = \rho(hg)$ and $\rho(g)^{-1}\rho(h)^{-1} = (\rho(h)\rho(g))^{-1} = \rho(hg)^{-1}$. The map $g \mapsto hg$ is a bijection $G \to G$, with inverse $g \mapsto h^{-1}g$ by Theorem 0.10.1, so as g runs over G so does hg, and the sum is unchanged: $\rho(h)P_0\rho(h)^{-1} = P_0$. Multiplying on the right by $\rho(h)$ gives $\rho(h)P_0 = P_0\rho(h)$.

Step 4: conclusion. Put $W := \ker P_0$. By Theorem 4.5.1 (a), $V = \operatorname{im} P_0 \oplus \ker P_0 = U \oplus W$. And W is G-invariant: if $P_0 w = 0$ and $h \in G$, then by Step 3, $P_0\rho(h)w = \rho(h)P_0 w = 0$, so $\rho(h)w \in W$. This proves the theorem.

■

Two hypotheses did real work, and it is worth naming where. **Finiteness** of G is what makes the sum $\sum_g$ a finite sum, hence an operator at all. The condition on the **characteristic** is what makes $\frac{1}{n}$ exist, and Step 2 is where it is spent: without the division, the computation in Step 2 gives $\sum_g \rho(g)P\rho(g)^{-1}u = (n \cdot 1_F)u$, which is 0 for every $u \in U$ when $\operatorname{char} F$ divides n. The unnormalized sum then **annihilates** U instead of fixing it, and there is nothing to rescue. Both failures are real, not artifacts of the proof.

Example. Maschke fails when the characteristic divides the order

Let p be a prime, $F = \mathbb{F}_p$, $G = \mathbb{Z}/p\mathbb{Z}$, and $V = \mathbb{F}_p^2$. Define

$$\rho([k]) := \begin{pmatrix} 1 & k \\ 0 & 1 \end{pmatrix} \in \mathrm{GL}_2(\mathbb{F}_p).$$

Show that ρ is a well-defined representation, that $U = \operatorname{span}(\mathbf{e}_1)$ is G-invariant, and that U has **no** G-invariant complement.

Solution. Well defined. The entry k is read in $\mathbb{F}_p$, so the matrix depends only on k modulo p; and $\left(\begin{smallmatrix}1 & k\\ 0 & 1\end{smallmatrix}\right)\left(\begin{smallmatrix}1 & l\\ 0 & 1\end{smallmatrix}\right) = \left(\begin{smallmatrix}1 & k+l\\ 0 & 1\end{smallmatrix}\right)$, so $\rho([k])\rho([l]) = \rho([k]+[l])$. Each matrix is invertible, with inverse $\rho([-k])$. So ρ is a representation of the additive group $\mathbb{Z}/p\mathbb{Z}$. Here $|G| = p = \operatorname{char}\mathbb{F}_p$, so the characteristic **does** divide the order and the hypothesis of Theorem 23.7.3 fails.

The invariant line. $\rho([k])\mathbf{e}_1 = \mathbf{e}_1$ for every k, so $U = \operatorname{span}(\mathbf{e}_1)$ is G-invariant, and $U \neq \{\mathbf{0}\}, V$.

No invariant complement. If $V = U \oplus W$, then $\dim W = 2 - 1 = 1$ by Theorem 1.7.3, so $W = \operatorname{span}(\mathbf{v})$ for some $\mathbf{v} = (a, b) \neq \mathbf{0}$; and $b \neq 0$, since $b = 0$ would force $a \neq 0$ and $W = \operatorname{span}(\mathbf{e}_1) = U$, which is not a complement of itself. Suppose W is G-invariant. Then $\rho([1])\mathbf{v} = (a + b,\, b)$ must be a scalar multiple $\lambda\mathbf{v} = (\lambda a, \lambda b)$. Comparing second coordinates and dividing by $b \neq 0$ gives $\lambda = 1$; comparing first coordinates then gives $a + b = a$, so $b = 0$, a contradiction. Hence U has no G-invariant complement. In fact U is the **only** nontrivial invariant subspace, so V is neither irreducible nor a direct sum of irreducibles.

Warning. **Do not drop finiteness either.** Let $G = \mathbb{Z}$ (infinite) act on $\mathbb{Q}^2$ by $\rho(k) = \left(\begin{smallmatrix}1 & k\\ 0 & 1\end{smallmatrix}\right)$ — now well defined, since $\mathbb{Z}$ has no relations to respect. The computation just made, with $b \neq 0$ in $\mathbb{Q}$, again shows $\operatorname{span}(\mathbf{e}_1)$ is invariant with no invariant complement, and here the characteristic is 0. Averaging over $\mathbb{Z}$ is not an option: the sum has infinitely many terms.

Once one invariant complement can always be found, the full decomposition follows by induction.

Corollary 23.7.4. Complete Reducibility of Group Representations

Let G be a finite group and F a field with $\operatorname{char} F \nmid |G|$ — for instance $F = \mathbb{C}$. Then every representation ρ of G on a nonzero finite-dimensional space V over F decomposes as

$$V = V_1 \oplus V_2 \oplus \cdots \oplus V_r$$

with each V_i a G-invariant subspace on which the subrepresentation is irreducible.

Proof. We argue by strong induction on $\dim V \geq 1$ (Theorem 0.2.10). If ρ is irreducible, take $r = 1$ and $V_1 = V$. Otherwise V has a G-invariant subspace U with $U \neq \{0\}$ and $U \neq V$. By Theorem 23.7.3 there is a G-invariant W with $V = U \oplus W$, and $W \neq \{0\}$ since $U \neq V$. Both $\dim U$ and $\dim W$ are at least 1, and since $U \cap W = \{0\}$ (Theorem 1.7.4) the dimension formula Theorem 1.7.3 gives $\dim U + \dim W = \dim V$; hence both are smaller than $\dim V$. The induction hypothesis decomposes U and W into irreducible invariant subspaces, and concatenating the two lists decomposes V, because a subspace of U invariant under $\rho|_U$ is invariant under ρ. This proves the corollary. ∎

In the language of Section 03, the corollary says that $F[G]$ acts semisimply on every representation (Definition 23.3.4, Theorem 23.3.3); applied to the regular representation, it says $F[G]$ is a semisimple algebra. Section 06's structure theorem then applies to $\mathbb{C}[G]$, and Section 08 cashes that in.

Quick check. Where exactly does the proof of Corollary 23.7.4 use that $V \neq \{0\}$?

Solution. In the base of the induction and in the meaning of the conclusion. If $V = \{0\}$ it is not irreducible (Definition 23.7.4 excludes the zero space) and it has no nonzero invariant subspace to split off, so neither branch of the argument applies. The usual convention is that $\{0\}$ is the empty direct sum, $r = 0$; the corollary is stated with $V \neq \{0\}$ so that $r \geq 1$.

23.7.5 Two families worked out

We close with the two examples Section 08 will compute characters for. Both are over $\mathbb{C}$, where Corollary 23.7.4 applies to every finite group.

Example. The irreducible complex representations of a cyclic group

Let $G = \mathbb{Z}/n\mathbb{Z}$ and $F = \mathbb{C}$. Determine all irreducible representations of G up to equivalence.

Solution. Every irreducible is one-dimensional. Let ρ be a representation of G on $V \neq \{0\}$ over $\mathbb{C}$, and set $A = \rho([1])$. By Theorem 9.2.4, A has an eigenvector $v \neq 0$, say $Av = \lambda v$. The line $L = \operatorname{span}(v)$ is invariant under A, hence under every power $A^k = \rho([k])$, so L is G-invariant. If ρ is irreducible then $L = V$, so $\dim V = 1$. (For abelian G the algebra $\mathbb{C}[G]$ is commutative, since $\delta_g\delta_h = \delta_{gh} = \delta_{hg} = \delta_h\delta_g$ and multiplication is bilinear; so this is also Corollary 23.4.4 read through Theorem 23.7.1.)

Listing them. A one-dimensional representation is a homomorphism $\chi\colon \mathbb{Z}/n\mathbb{Z} \to \mathbb{C} \setminus \{0\}$, determined by $z = \chi([1])$, which must satisfy $z^n = \chi([n]) = \chi([0]) = 1$. So z is an n-th root of unity, $z = \omega^k$ with $\omega = e^{2\pi i/n}$ and $0 \leq k \leq n-1$ (Example (Roots of Unity)). Conversely each such z does define a homomorphism $\chi_k([j]) = \omega^{jk}$: the recipe respects representatives because $\omega^n = 1$, and $\omega^{(j+l)k} = \omega^{jk}\omega^{lk}$.

They are pairwise inequivalent. Two one-dimensional representations are equivalent exactly when they are equal, since on a one-dimensional space every operator is λ id for a scalar λ, and $S(\lambda\ \mathrm{id})S^{-1} = \lambda\ \mathrm{id}$ for every isomorphism S. And $\chi_k \neq \chi_l$ for $0 \leq k < l \leq n-1$, because $\chi_k([1]) = \omega^k \neq \omega^l = \chi_l([1])$, the n-th roots of unity being distinct.

So $\mathbb{Z}/n\mathbb{Z}$ has exactly n irreducible complex representations, all of degree 1, namely $\chi_0, \ldots, \chi_{n-1}$. Their degrees satisfy $\sum_k 1^2 = n = |G|$, the count Corollary 23.6.2 gives for the semisimple algebra $\mathbb{C}[G]$.

Example. The irreducible complex representations of the symmetric group on three letters

Write down three pairwise inequivalent irreducible complex representations of S_3, with explicit matrices for the two-dimensional one.

Solution. Trivial and sign. $\rho_{\mathrm{triv}}(\sigma) = 1$ and $\rho_{\mathrm{sgn}}(\sigma) = \operatorname{sgn}(\sigma)$, both on $\mathbb{C}$, both irreducible

because their degree is 1. They are inequivalent: $\rho_{\text{triv}}((1\ 2)) = 1 \neq -1 = \rho_{\text{sgn}}((1\ 2))$.

The standard representation. Take the permutation representation on $\mathbb{C}^3$ and restrict it to the invariant plane $W = \{\mathbf{x} : x_1+x_2+x_3 = 0\}$. A basis of W is $\mathcal{B} = (\mathbf{f}_1, \mathbf{f}_2)$ with $\mathbf{f}_1 = \mathbf{e}_1 - \mathbf{e}_2$ and $\mathbf{f}_2 = \mathbf{e}_2 - \mathbf{e}_3$: they lie in W, they are independent, and $\dim W = 2$ because W is the kernel of the surjective functional $\mathbf{x} \mapsto x_1 + x_2 + x_3$ on $\mathbb{C}^3$ (Theorem 2.3.1). Since S_3 is generated by $(1\ 2)$ and $(1\ 2\ 3)$ — every element of S_3 is a product of transpositions (Theorem 0.10.4), and $(1\ 3) = (1\ 2\ 3)(1\ 2)$, $(2\ 3) = (1\ 2)(1\ 2\ 3)$ — it is enough to record two matrices. With $\rho(\sigma)\mathbf{e}_i = \mathbf{e}_{\sigma(i)}$:

$$\begin{aligned} \rho((1\ 2))\mathbf{f}_1 &= \mathbf{e}_2 - \mathbf{e}_1 = -\mathbf{f}_1, \\ \rho((1\ 2))\mathbf{f}_2 &= \mathbf{e}_1 - \mathbf{e}_3 = \mathbf{f}_1 + \mathbf{f}_2, \\ \rho((1\ 2\ 3))\mathbf{f}_1 &= \mathbf{e}_2 - \mathbf{e}_3 = \mathbf{f}_2, \\ \rho((1\ 2\ 3))\mathbf{f}_2 &= \mathbf{e}_3 - \mathbf{e}_1 = -\mathbf{f}_1 - \mathbf{f}_2. \end{aligned}$$

Writing $R_\sigma := \rho(\sigma)|_W$, the two matrices in the basis $\mathcal{B}$ are

$$[R_{(1\ 2)}]^{\mathcal{B}}_{\mathcal{B}} = \begin{pmatrix} -1 & 1 \\ 0 & 1 \end{pmatrix}, \qquad [R_{(1\ 2\ 3)}]^{\mathcal{B}}_{\mathcal{B}} = \begin{pmatrix} 0 & -1 \\ 1 & -1 \end{pmatrix}.$$

The remaining four matrices follow by multiplying: $\mathbf{I}_2$ for id, $\left(\begin{smallmatrix} 0 & -1 \\ -1 & 0 \end{smallmatrix}\right)$ for $(1\ 3)$, $\left(\begin{smallmatrix} 1 & 0 \\ 1 & -1 \end{smallmatrix}\right)$ for $(2\ 3)$, and $\left(\begin{smallmatrix} -1 & 1 \\ -1 & 0 \end{smallmatrix}\right)$ for $(1\ 3\ 2)$.

Irreducibility. A proper nonzero invariant subspace of the plane W would be a line $\text{span}(\mathbf{v})$, and $\mathbf{v}$ would be an eigenvector of **both** displayed matrices. The first has eigenvalues -1 and 1, with eigenvectors spanned by $(1, 0)$ and $(1, 2)$ respectively. The second has characteristic polynomial $x^2 + x + 1$, whose roots are the two primitive cube roots of unity; neither $(1, 0)$ nor $(1, 2)$ is an eigenvector of it, since $\left(\begin{smallmatrix} 0 & -1 \\ 1 & -1 \end{smallmatrix}\right)(1, 0) = (0, 1)$ and $\left(\begin{smallmatrix} 0 & -1 \\ 1 & -1 \end{smallmatrix}\right)(1, 2) = (-2, -1)$ are not multiples of $(1, 0)$, $(1, 2)$. So there is no common eigenvector, and the standard representation is irreducible.

Inequivalent. It has degree 2, the others degree 1, and equivalent representations act on spaces of the same dimension.

The degrees satisfy $1^2 + 1^2 + 2^2 = 6 = |S_3|$, which is the count Corollary 23.6.2 gives for the semisimple algebra $\mathbb{C}[S_3]$. That count runs over **all** isomorphism classes of irreducible $\mathbb{C}[S_3]$-spaces (Theorem 23.6.1), and a further class would add at least 1 to a total already equal to 6; so these three are all of the irreducible complex representations of S_3. Section 08 reaches the same conclusion by counting conjugacy classes, which does not require the list to be found first.

23.7.6 Exercises

23.7.6.1 A. Check your understanding

Exercise. A1

(a) State the definition of a representation of a group G on a vector space V, and say which two properties of ρ come for free.

(b) State Maschke's theorem with all of its hypotheses.

(c) Determine whether the following is true: "every representation of degree 1 is irre-

ducible." Justify your answer.

(d) Give an example of a reducible representation of S_3 over $\mathbb{C}$, and name an invariant subspace.

(e) Explain why $\rho\colon \mathbb{Z}/2\mathbb{Z} \to \mathrm{GL}_1(\mathbb{Q})$, $\rho([k]) = 2^k$, is **not** a representation.

Solution.

(a) A representation is a group homomorphism $\rho\colon G \to \mathrm{GL}(V)$ with V finite-dimensional over F, that is, $\rho(gh) = \rho(g)\rho(h)$ for all $g, h \in G$ with each $\rho(g)$ invertible (Definition 23.7.1). Free: $\rho(1) = \mathrm{id}_V$ and $\rho(g^{-1}) = \rho(g)^{-1}$, by Theorem 0.10.3.

(b) If G is finite, F is a field with $\mathrm{char}\, F \nmid |G|$, V is finite-dimensional over F, and $U \subseteq V$ is G-invariant, then U has a G-invariant complement W, so $V = U \oplus W$ (Theorem 23.7.3).

(c) True. If $\dim V = 1$ then $V \neq \{0\}$ and the only subspaces of V are $\{0\}$ and V, so in particular the only invariant ones are the trivial ones.

(d) The permutation representation of S_3 on $\mathbb{C}^3$; the line $\mathrm{span}(\mathbf{e}_1 + \mathbf{e}_2 + \mathbf{e}_3)$ is invariant, as is the sum-zero plane.

(e) It is not well defined: $[0] = [2]$ in $\mathbb{Z}/2\mathbb{Z}$, but $2^0 = 1 \neq 4 = 2^2$. (Equivalently, a degree-1 representation of $\mathbb{Z}/2\mathbb{Z}$ needs $\rho([1])^2 = 1$, and the only rational solutions of $z^2 = 1$ are $z = \pm 1$.)

23.7.6.2 B. Practice

Exercise. B1: Which of these are representations?

Determine which of the following are representations. Justify your answer in each case.

(a) $G = \mathbb{Z}/6\mathbb{Z}$, $F = \mathbb{C}$, $\rho([k]) = \begin{pmatrix} \omega^k & 0 \\ 0 & \omega^{-k} \end{pmatrix}$ with $\omega = e^{\pi i/3}$.

(b) $G = S_3$, $F = \mathbb{R}$, $\rho(\sigma) = \mathrm{sgn}(\sigma)\,\mathbf{P}_\sigma$, where $\mathbf{P}_\sigma$ is the permutation matrix with $\mathbf{P}_\sigma \mathbf{e}_i = \mathbf{e}_{\sigma(i)}$ (Definition 3.4.2).

(c) $G = S_3$, $F = \mathbb{R}$, $\rho(\sigma) = \mathbf{P}_\sigma^{\top}$ with $\mathbf{P}_\sigma$ as in (b).

(d) $G = \mathbb{Z}/4\mathbb{Z}$, $F = \mathbb{R}$, $\rho([k]) = \begin{pmatrix} 1 & 0 \\ 0 & k+1 \end{pmatrix}$.

Solution.

(a) Yes. Here $\omega^6 = e^{2\pi i} = 1$, so the matrix depends only on k modulo 6 and the map is well defined; it is diagonal, hence invertible since $\omega^{\pm k} \neq 0$; and $\rho([k])\rho([l]) = \mathrm{diag}(\omega^{k+l}, \omega^{-k-l}) = \rho([k] + [l])$.

(b) Yes. Both factors are multiplicative: $\mathrm{sgn}(\sigma\tau) = \mathrm{sgn}(\sigma)\,\mathrm{sgn}(\tau)$ (Theorem 7.3.1) and $\mathbf{P}_\sigma \mathbf{P}_\tau = \mathbf{P}_{\sigma\tau}$ (Lemma 3.4.7 (a)). Scalars commute with matrices, so $\rho(\sigma)\rho(\tau) = \mathrm{sgn}(\sigma\tau)\mathbf{P}_{\sigma\tau} = \rho(\sigma\tau)$, and $\rho(\sigma)$ is invertible because $\mathbf{P}_\sigma$ is and $\mathrm{sgn}(\sigma) \neq 0$.

(c) No. Taking transposes reverses products: $\rho(\sigma)\rho(\tau) = \mathbf{P}_\sigma^{\top}\mathbf{P}_\tau^{\top} = (\mathbf{P}_\tau \mathbf{P}_\sigma)^{\top} = \mathbf{P}_{\tau\sigma}^{\top} = \rho(\tau\sigma)$. Since S_3 is not abelian this differs from $\rho(\sigma\tau)$: with $\sigma = (1\ 2)$, $\tau = (1\ 3)$ we get $\tau\sigma = (1\ 2\ 3) \neq (1\ 3\ 2) = \sigma\tau$ (Example (Computing with cycles)), and distinct permutations have distinct permutation matrices. (The repair is $\sigma \mapsto \mathbf{P}_{\sigma^{-1}}^{\top}$, or equivalently

$\mathbf{P}_\sigma$ itself.)

(d) No, on two counts. It is not well defined: $[0] = [4]$ but $\operatorname{diag}(1,1) \neq \operatorname{diag}(1,5)$. And it is not multiplicative even on representatives: $\rho([1])\rho([1]) = \operatorname{diag}(1,4)$ while $\rho([2]) = \operatorname{diag}(1,3)$.

Exercise. B2: Averaging an inner product

Let G be a finite group and ρ a representation of G on a complex space V carrying an inner product $\langle \cdot, \cdot \rangle$ (Definition 11.1.1). Define

$$\langle u, v \rangle_G := \frac{1}{|G|} \sum_{g \in G} \langle \rho(g)u, \rho(g)v \rangle.$$

Prove that $\langle \cdot, \cdot \rangle_G$ is an inner product on V and that every $\rho(g)$ is an isometry for it. Hence give a second proof of Theorem 23.7.3 over $\mathbb{C}$.

Solution. (IP1). For fixed v, the map $u \mapsto \langle \rho(g)u, \rho(g)v \rangle$ is linear in u, being a composition of the linear map $\rho(g)$ with a slot that is linear by (IP1) for $\langle \cdot, \cdot \rangle$. A sum of linear maps times a scalar is linear.

(IP2). $\langle v, u \rangle_G = \frac{1}{|G|} \sum_g \langle \rho(g)v, \rho(g)u \rangle = \frac{1}{|G|} \sum_g \overline{\langle \rho(g)u, \rho(g)v \rangle} = \overline{\langle u, v \rangle_G}$, since conjugation is additive.

(IP3). For $v \neq 0$, each $\rho(g)v \neq 0$ because $\rho(g)$ is invertible, so each term $\langle \rho(g)v, \rho(g)v \rangle$ is > 0, and $|G| > 0$ is a positive real. Hence $\langle v, v \rangle_G > 0$.

Isometry. For $h \in G$, $\langle \rho(h)u, \rho(h)v \rangle_G = \frac{1}{|G|} \sum_g \langle \rho(gh)u, \rho(gh)v \rangle = \langle u, v \rangle_G$, because $g \mapsto gh$ is a bijection of G (Theorem 0.10.1).

Second proof of Maschke over $\mathbb{C}$. Let U be G-invariant and let $W = U^\perp$ be its orthogonal complement for $\langle \cdot, \cdot \rangle_G$. Then $V = U \oplus W$ by Theorem 11.3.2. If $w \in W$, $u \in U$, and $h \in G$, then, writing $u = \rho(h)(\rho(h)^{-1}u)$ and using that $\rho(h)$ preserves $\langle \cdot, \cdot \rangle_G$,

$$\langle \rho(h)w, u \rangle_G = \langle \rho(h)w, \rho(h)\rho(h)^{-1}u \rangle_G = \langle w, \rho(h)^{-1}u \rangle_G = 0,$$

the last step because $\rho(h)^{-1}u = \rho(h^{-1})u \in U$ and $w \in U^\perp$. So $\rho(h)w \in W$, and W is G-invariant. Note that this route needs $F = \mathbb{R}$ or $\mathbb{C}$, where inner products live, whereas Theorem 23.7.3 needs only a condition on the characteristic.

Exercise. B3: Decomposing a permutation representation

Number the coordinates of $\mathbb{C}^3$ as $0, 1, 2$, as in Chapter 12 §09, and let $G = \mathbb{Z}/3\mathbb{Z}$ act by $\rho([k]) = \mathbf{S}^k$, where $\mathbf{S}$ is the cyclic shift, $\mathbf{S}\mathbf{x} = (x_1, x_2, x_0)$. Check that ρ is a representation, decompose $\mathbb{C}^3$ into one-dimensional invariant subspaces, and identify each with one of the χ_k of Example (The irreducible complex representations of a cyclic group).

Solution. A representation. Shifting three times restores every entry, so $\mathbf{S}^3 = \mathbf{I}$, and therefore $\mathbf{S}^k$ depends only on k modulo 3: the recipe is well defined on $\mathbb{Z}/3\mathbb{Z}$. It is multiplicative

because $\mathbf{S}^k\mathbf{S}^l = \mathbf{S}^{k+l}$, and each $\mathbf{S}^k$ is invertible with inverse $\mathbf{S}^{3-k}$.

Decomposition. Let $\omega = e^{2\pi i/3}$. By Theorem 12.9.2 the vectors $\mathbf{f}_j = \frac{1}{\sqrt{3}}(1, \omega^j, \omega^{2j})$, $j = 0, 1, 2$, satisfy $\mathbf{S}\mathbf{f}_j = \omega^j \mathbf{f}_j$. The three eigenvalues $1, \omega, \omega^2$ are distinct, so the $\mathbf{f}_j$ are independent (Theorem 9.2.2), and three independent vectors in $\mathbb{C}^3$ form a basis (Theorem 1.6.4). Hence

$$\mathbb{C}^3 = \operatorname{span}(\mathbf{f}_0) \oplus \operatorname{span}(\mathbf{f}_1) \oplus \operatorname{span}(\mathbf{f}_2),$$

and each line is invariant under $\mathbf{S}$, hence under every $\rho([k]) = \mathbf{S}^k$.

Identification. On $\operatorname{span}(\mathbf{f}_j)$ the element $[k]$ acts by the scalar ω^{jk}, which is exactly $\chi_j([k])$ in the notation of Example (The irreducible complex representations of a cyclic group). So the three summands realize χ_0, χ_1, χ_2: every irreducible representation of $\mathbb{Z}/3\mathbb{Z}$ occurs exactly once.

23.7.6.3 C. Going deeper

Exercise. C1: One-dimensional representations and the commutator

Let G be a finite group and F a field.

(a) Prove that every representation of degree 1 satisfies $\rho(ghg^{-1}h^{-1}) = 1$ for all $g, h \in G$.

(b) Deduce that the only degree-1 complex representation of $G = S_3$ other than the trivial one is the sign representation. *Hint: write the 3-cycle* $(1\ 2\ 3)$ *in the form* $ghg^{-1}h^{-1}$ *with* g *and* h *transpositions.*

Solution.

(a) A degree-1 representation takes values in $F \setminus \{0\}$, which is abelian under multiplication. So $\rho(ghg^{-1}h^{-1}) = \rho(g)\rho(h)\rho(g)^{-1}\rho(h)^{-1} = \rho(g)\rho(g)^{-1}\rho(h)\rho(h)^{-1} = 1$, using Theorem 0.10.3 for the inverses.

(b) Let ρ have degree 1 over $\mathbb{C}$. Transpositions are self-inverse, so $(1\ 2)(1\ 3)(1\ 2)^{-1}(1\ 3)^{-1} = (1\ 2)(1\ 3)(1\ 2)(1\ 3)$. Applying this product to each point, rightmost factor first: $1 \mapsto 3 \mapsto 3 \mapsto 1 \mapsto 2$, $2 \mapsto 2 \mapsto 1 \mapsto 3 \mapsto 3$, and $3 \mapsto 1 \mapsto 2 \mapsto 2 \mapsto 1$. So the product is $(1\ 2\ 3)$. By (a), $\rho((1\ 2\ 3)) = 1$, and then $\rho((1\ 3\ 2)) = \rho((1\ 2\ 3))^2 = 1$. Let $c = \rho((1\ 2))$. Since $(1\ 2)^2 = \mathrm{id}$, $c^2 = 1$, so $c = \pm 1$. Finally $(1\ 3) = (1\ 2\ 3)(1\ 2)$ and $(2\ 3) = (1\ 2)(1\ 2\ 3)$, so $\rho((1\ 3)) = \rho((2\ 3)) = c$ as well. Thus ρ is determined by c: for $c = 1$ it is trivial, and for $c = -1$ it agrees with sgn on every element. Both values do occur, so there are exactly two.

Exercise. C2: Maschke over a field of the right characteristic

Let $G = \mathbb{Z}/2\mathbb{Z}$ and $V = F^2$, with $\rho([1]) = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$.

(a) For $F = \mathbb{Q}$, find two G-invariant lines whose direct sum is V.

(b) For $F = \mathbb{F}_2$, prove that V has exactly one nontrivial G-invariant subspace, and conclude that Theorem 23.7.3 fails here. Which hypothesis is violated?

(c) For $F = \mathbb{F}_3$, decompose V into irreducible subrepresentations.

Solution.

(a) $\rho([1])$ swaps the coordinates, so $\mathbf{u} = (1,1)$ and $\mathbf{w} = (1,-1)$ satisfy $\rho([1])\mathbf{u} = \mathbf{u}$ and $\rho([1])\mathbf{w} = -\mathbf{w}$. Both lines are invariant, since $\rho([0]) = \mathbf{I}_2$ as well. The two vectors are independent, because the matrix with columns $\mathbf{u}, \mathbf{w}$ has determinant $-2 \neq 0$ in $\mathbb{Q}$; hence $\operatorname{span}(\mathbf{u}) \cap \operatorname{span}(\mathbf{w}) = \{\mathbf{0}\}$ and the sum is direct (Theorem 1.7.4). Its dimension is $2 = \dim V$, so $V = \operatorname{span}(\mathbf{u}) \oplus \operatorname{span}(\mathbf{w})$.

(b) Over $\mathbb{F}_2$ we have $-1 = 1$, so $\mathbf{w} = \mathbf{u} = (1,1)$ and the two lines coincide. A nontrivial invariant subspace is a line $\operatorname{span}((a,b))$ with $(a,b) \neq (0,0)$, and invariance requires $(b,a) = \lambda(a,b)$. Over $\mathbb{F}_2$ the only nonzero scalar is $\lambda = 1$, forcing $a = b$, so $(a,b) = (1,1)$ and the only nontrivial invariant subspace is $\operatorname{span}((1,1))$. A complement of a line in the plane V is again a line, and the only invariant line is $\operatorname{span}((1,1))$ itself, which is not a complement of itself. So this invariant subspace has no invariant complement, and V is not a direct sum of irreducibles. The violated hypothesis is on the characteristic: $\operatorname{char} \mathbb{F}_2 = 2 = |G|$.

(c) Over $\mathbb{F}_3$, $-1 = 2 \neq 1$, so $\mathbf{u} = (1,1)$ and $\mathbf{w} = (1,2)$ are again independent, with $\rho([1])\mathbf{u} = \mathbf{u}$ and $\rho([1])\mathbf{w} = 2\mathbf{w} = -\mathbf{w}$. Hence $V = \operatorname{span}(\mathbf{u}) \oplus \operatorname{span}(\mathbf{w})$, each summand of dimension 1 and so irreducible. Here $\operatorname{char} \mathbb{F}_3 = 3 \nmid 2 = |G|$, and Theorem 23.7.3 applies.

Exercise. C3: The contragredient representation

Let G be a finite group and ρ a representation of G on V over F. For $g \in G$ define $\tau(g) := (\rho(g)^{-1})' \in \mathcal{L}(V^*)$, where T' denotes the dual map (Definition 5.4.1).

(a) Prove that τ is a representation of G on V^*, the **contragredient** of ρ.

(b) Prove that if $U \subseteq V$ is G-invariant for ρ, then its annihilator $U^0 \subseteq V^*$ (Definition 5.2.1) is G-invariant for τ.

(c) Hence deduce that if τ is irreducible, so is ρ.

Solution.

(a) Each $\rho(g)^{-1}$ is invertible, so $\tau(g)$ is invertible by Theorem 5.4.1 (d). For multiplicativity, put $S = \rho(h)^{-1}$ and $T = \rho(g)^{-1}$. By Theorem 5.4.1 (b),

$$\begin{aligned}\tau(g)\tau(h) = T'S' = (ST)' &= (\rho(h)^{-1}\rho(g)^{-1})' \\ &= ((\rho(g)\rho(h))^{-1})' = (\rho(gh)^{-1})' = \tau(gh),\end{aligned}$$

where the fourth equality is Theorem 0.10.1 (3) applied in $\mathrm{GL}(V)$. So τ is a homomorphism into $\mathrm{GL}(V^*)$.

(b) Let $\varphi \in U^0$, $g \in G$, and $u \in U$. By Definition 5.4.1, $(\tau(g)\varphi)(u) = \varphi(\rho(g)^{-1}u)$. Since $\rho(g)^{-1} = \rho(g^{-1})$ and U is G-invariant, $\rho(g)^{-1}u \in U$, so the value is 0. Hence $\tau(g)\varphi$ annihilates U, that is, $\tau(g)\varphi \in U^0$.

(c) Suppose τ is irreducible. Then $V^* \neq \{0\}$, so $V \neq \{0\}$ by Corollary 5.1.4. Let U be a G-invariant subspace of V with $U \neq \{0\}$ and $U \neq V$. By (b), U^0 is G-invariant, and by Theorem 5.2.2, $\dim U^0 = \dim V - \dim U$. From $0 < \dim U < \dim V$ we get $0 < \dim U^0 < \dim V = \dim V^*$, so U^0 is a nontrivial invariant subspace of V^*, contradicting irreducibility. Hence ρ has no such U, and ρ is irreducible.

23.8 Characters and Orthogonality

Complete reducibility (Corollary 23.7.4) promises that every complex representation of a finite group breaks into irreducible pieces, but the promise is not constructive: it says nothing about which pieces, or how many of each. This section produces a small table of numbers that answers both questions. To each representation we attach one function $G \to \mathbb{C}$, its character, and we prove that the characters of the irreducible representations are an **orthonormal basis** of a space of functions we can compute the dimension of by counting. Everything a complex representation is, up to equivalence, is then read off from one row of numbers.

Throughout, G is a **finite** group and the field is $\mathbb{C}$. The hypothesis matters twice over: $\operatorname{char}\mathbb{C} = 0$ never divides $|G|$, so Theorem 23.7.3 applies; and $\mathbb{C}$ is algebraically closed, so Schur's lemma (Theorem 23.4.1) gives that the only operators commuting with an irreducible representation are the scalars. The theory over a field whose characteristic divides $|G|$ — modular representation theory — is a different subject, and this book does not enter it.

23.8.1 Conjugacy classes

Before characters we need the sets they are constant on. The book has met them only in an exercise of Section 1 (Exercise (C1: The center of a group algebra)), so we build them here.

Definition 23.8.1. Conjugate elements, conjugacy class

Let G be a group. Elements $g, h \in G$ are **conjugate** if $h = xgx^{-1}$ for **some** $x \in G$. The **conjugacy class** of g is

$$C(g) \coloneqq \{xgx^{-1} : x \in G\}.$$

Conjugacy is an equivalence relation (Definition 0.5.2): $g = 1g1^{-1}$ gives reflexivity; if $h = xgx^{-1}$, then $g = x^{-1}h(x^{-1})^{-1}$, which gives symmetry using Theorem 0.10.1; and if $h = xgx^{-1}$ and $k = yhy^{-1}$, then $k = (yx)g(yx)^{-1}$, which gives transitivity. Hence the conjugacy classes are the equivalence classes, and by Theorem 0.5.2 they **partition** G: every element lies in exactly one class. We write k for the number of conjugacy classes of G, a finite number since G is finite.

In an abelian group $xgx^{-1} = g$ for every x, so every class is a single element and $k = |G|$. At the other extreme, $C(1) = \{1\}$ always.

Example. The conjugacy classes of the symmetric group on three letters

Find the conjugacy classes of S_3.

Solution. First a rule that does the work. For $\sigma \in S_n$ and a cycle $(i_1\ i_2\ \cdots\ i_r)$,

$$\sigma\,(i_1\ i_2\ \cdots\ i_r)\,\sigma^{-1} = (\sigma(i_1)\ \sigma(i_2)\ \cdots\ \sigma(i_r)).$$

Both sides are permutations, so it is enough to check they agree at every point. At $\sigma(i_j)$: the left side applies σ^{-1} to get i_j, then the cycle to get i_{j+1} (indices read cyclically), then σ to get $\sigma(i_{j+1})$; the right side does the same in one step. At a point $\sigma(m)$ with $m \notin \{i_1, \ldots, i_r\}$: the left side gives $\sigma(m)$ back, and so does the right side, since $\sigma(m)$ is not among the $\sigma(i_j)$ by injectivity. Every point of $\{1, \ldots, n\}$ is $\sigma(m)$ for exactly one m, so the two agree.

So conjugation relabels the entries of a cycle. In S_3, conjugating $(1\ 2)$ by the six elements produces $(\sigma(1)\ \sigma(2))$, and as σ runs over S_3 the pair $\{\sigma(1), \sigma(2)\}$ runs over all three two-element subsets; so $C((1\ 2)) = \{(1\ 2), (1\ 3), (2\ 3)\}$. Conjugating $(1\ 2\ 3)$ gives $(\sigma(1)\ \sigma(2)\ \sigma(3))$, which is $(1\ 2\ 3)$ or $(1\ 3\ 2)$ according to the cyclic order; both occur, for $\sigma = \mathrm{id}$ and $\sigma = (1\ 2)$, so $C((1\ 2\ 3)) = \{(1\ 2\ 3), (1\ 3\ 2)\}$. The three classes are

$$\{\mathrm{id}\}, \qquad \{(1\ 2), (1\ 3), (2\ 3)\}, \qquad \{(1\ 2\ 3), (1\ 3\ 2)\},$$

of sizes $1, 3, 2$, which add to $6 = |S_3|$ as a partition must. So $k = 3$ for S_3 — the same number as the three irreducible complex representations found in Example (The irreducible complex representations of the symmetric group on three letters). That is no coincidence, and the theorem explaining it is the main result of this section.

23.8.2 The character of a representation

A representation ρ on V is a whole family of matrices, and the matrices change when the basis does. To compare representations we want something attached to ρ that does not move. Chapter 2 already supplies the candidate: the trace of an operator does not depend on the basis (Definition 4.5.2), and similar operators have equal traces.

The character of a representation records one number per group element: the trace of the operator that element acts by.

Definition 23.8.2. Character

Let ρ be a representation of a finite group G on a finite-dimensional complex vector space V. The **character** of ρ is the function

$$\chi_\rho \colon G \to \mathbb{C}, \qquad \chi_\rho(g) \coloneqq \operatorname{tr} \rho(g).$$

Its **degree** is $\chi_\rho(1) = \dim V$. A character is **irreducible** if ρ is irreducible (Definition 23.7.4).

In words: a character is a list of $|G|$ complex numbers, one per group element, obtained by computing traces. It throws away a great deal — the whole matrix becomes one number — and the content of this section is that it throws away exactly the information we did not want.

Examples. Let G be finite.

- The trivial representation has $\chi(g) = \operatorname{tr}(\mathrm{id}_{\mathbb{C}}) = 1$ for every g. We write χ_{triv} for the constant function 1.
- The sign representation of S_n has $\chi(\sigma) = \operatorname{sgn}(\sigma)$.
- The permutation representation of S_n on $\mathbb{C}^n$ has $\chi(\sigma) = \#\{i : \sigma(i) = i\}$, the number of **fixed points**. Indeed the matrix of $\rho(\sigma)$ in the standard basis has (i,i)-entry equal to 1 exactly when $\sigma(i) = i$, and 0 otherwise.
- The regular representation on $\mathbb{C}[G]$ has $\chi_{\mathrm{reg}}(1) = |G|$ and $\chi_{\mathrm{reg}}(g) = 0$ for $g \neq 1$. The matrix of $\rho_{\mathrm{reg}}(g)$ in the basis $\{\delta_h\}$ has (h,h)-entry equal to 1 exactly when $gh = h$, that is, when $g = 1$ by cancellation (Theorem 0.10.1).

One fact about finite groups is needed first, and Chapter 0 did not record it. It costs four lines, and Section 9 will use it constantly.

Lemma 23.8.1. Order of an Element

Let G be a finite group and $g \in G$. Then there is a least integer $m \geq 1$ with $g^m = 1$, called the **order** of g. Moreover $g^k = g^l$ if and only if m divides $k - l$, and

$$\langle g \rangle \coloneqq \{g^k : k \in \mathbb{Z}\} = \{1, g, g^2, \dots, g^{m-1}\}$$

is a subgroup of G with exactly m elements.

Proof. The $|G| + 1$ elements $g^0, g^1, \dots, g^{|G|}$ cannot all be distinct, so $g^i = g^j$ for some $0 \leq i < j$. Canceling g^i (Theorem 0.10.1) gives $g^{j-i} = 1$ with $j - i \geq 1$. Hence the set of positive integers p with $g^p = 1$ is nonempty, and it has a least element m.

Suppose $g^k = g^l$, so $g^{k-l} = 1$. Divide: $k - l = qm + r$ with $0 \leq r < m$. Then

$$g^r = (g^m)^{-q} g^{k-l} = 1,$$

and minimality of m forces $r = 0$, that is $m \mid k - l$. Conversely $m \mid k - l$ gives $g^{k-l} = (g^m)^{(k-l)/m} = 1$, so $g^k = g^l$.

Consequently g^k depends only on k modulo m, so $\langle g \rangle = \{1, g, \dots, g^{m-1}\}$, and those m elements are distinct because $m \nmid k - l$ whenever $0 \leq l < k \leq m - 1$. Finally $1 \in \langle g \rangle$, $g^k g^l = g^{k+l} \in \langle g \rangle$, and $(g^k)^{-1} = g^{-k} \in \langle g \rangle$, so $\langle g \rangle$ is a subgroup (Definition 0.10.2).

■

The next proposition collects the properties we use constantly. Chapter 4's trace results do all the work.

Proposition 23.8.2. Basic properties of characters

Let ρ be a representation of a finite group G on a complex space V, with character $\chi = \chi_\rho$. Then:

(a) $\chi(1) = \dim V$;
(b) χ is a **class function**: $\chi(xgx^{-1}) = \chi(g)$ for all $g, x \in G$;
(c) equivalent representations have equal characters;
(d) if $V = U \oplus W$ with both summands G-invariant, then $\chi = \chi_{\rho|_U} + \chi_{\rho|_W}$;
(e) $\chi(g^{-1}) = \overline{\chi(g)}$ for every $g \in G$, and $|\chi(g)| \leq \chi(1)$.

Idea. Parts (a)–(d) are the four standard facts about the trace read in this setting: the trace of the identity, invariance under conjugation, invariance under similarity, and additivity over a block decomposition. Part (e) is different: it needs to know that $\rho(g)$ is diagonalizable with eigenvalues on the unit circle, which comes from g having finite order.

Proof.

(a) $\rho(1) = \mathrm{id}_V$ (Theorem 0.10.3), and $\operatorname{tr} \mathrm{id}_V = (\dim V) \cdot 1 = \dim V$ in $\mathbb{C}$ (Theorem 4.5.3).

(b) $\chi(xgx^{-1}) = \operatorname{tr}(\rho(x)\rho(g)\rho(x)^{-1}) = \operatorname{tr}(\rho(g)\rho(x)^{-1}\rho(x)) = \operatorname{tr}\rho(g) = \chi(g)$, where the second equality is $\operatorname{tr}(ST) = \operatorname{tr}(TS)$ (Theorem 4.5.3) applied with $S = \rho(x)$ and $T = \rho(g)\rho(x)^{-1}$.

(c) If $\rho'(g) = S\rho(g)S^{-1}$ for an isomorphism S (Definition 23.7.3), the same computation as in (b) gives $\operatorname{tr}\rho'(g) = \operatorname{tr}\rho(g)$.

(d) Choose a basis of V formed by a basis of U followed by a basis of W. Since both are invariant, the matrix of $\rho(g)$ in this basis is block diagonal, with the matrices of $\rho(g)|_U$ and $\rho(g)|_W$ on the diagonal. The diagonal entries of the big matrix are those of the two blocks, so the traces add (Definition 4.5.2).

(e) We may assume $V \neq \{0\}$, since otherwise χ is the zero function and both claims are clear.

Finite order. By Lemma 23.8.1 there is an integer $d \geq 1$ with $g^d = 1$, namely the order of g.

Diagonalizability. Then $\rho(g)^d = \rho(g^d) = \mathrm{id}_V$, so the polynomial $x^d - 1$ annihilates $\rho(g)$. Over $\mathbb{C}$ it equals $\prod_{m=0}^{d-1}(x - \zeta^m)$ with $\zeta = e^{2\pi i/d}$, a product of **distinct** linear factors (Example (Roots of Unity)). By Theorem 9.8.2, (c) $\Rightarrow$ (a), $\rho(g)$ is diagonalizable, and every eigenvalue λ satisfies $\lambda^d = 1$, hence $|\lambda| = 1$.

Conclusion. Let $\lambda_1, \dots, \lambda_n$ be the diagonal entries in an eigenbasis, $n = \dim V$. Then $\chi(g) = \sum_j \lambda_j$. In the same basis $\rho(g)^{-1} = \rho(g^{-1})$ is diagonal with entries λ_j^{-1}, and $|\lambda_j| = 1$ gives $\lambda_j^{-1} = \overline{\lambda_j}$. Hence $\chi(g^{-1}) = \sum_j \overline{\lambda_j} = \overline{\chi(g)}$. Finally $|\chi(g)| \leq \sum_j |\lambda_j| = n = \chi(1)$ by the triangle inequality for complex numbers (Theorem 0.6.4), applied repeatedly. This proves the proposition.

■

Warning. **A character is not a homomorphism.** For the standard representation of S_3 (Example (The irreducible complex representations of the symmetric group on three letters)), $\chi((1\ 2)) = 0$ and $\chi((1\ 3)) = 0$, while $(1\ 2)(1\ 3) = (1\ 3\ 2)$ has $\chi((1\ 3\ 2)) = -1 \neq 0 \cdot 0$. Only characters of **degree** 1 are homomorphisms, because then $\chi(g)$ is the whole operator. Nor need the values of a character be real: for $\mathbb{Z}/4\mathbb{Z}$ they include $\pm i$.

Quick check. A representation of G has character χ with $\chi(1) = 3$ and $\chi(g) = 4$ for some g. What is wrong?

Solution. Nothing can have such a character: by Proposition 23.8.2 (e), $|\chi(g)| \leq \chi(1) = 3$ for every g, and $4 > 3$. The bound says a character is largest, in absolute value, at the identity.

23.8.3 Class functions and their inner product

Part (b) of Proposition 23.8.2 says every character lives in one particular space of functions, and that space is small enough to be a useful home.

Definition 23.8.3. Class function

Let G be a finite group. A function $f\colon G \to \mathbb{C}$ is a **class function** if $f(xgx^{-1}) = f(g)$ for all $g, x \in G$; equivalently, if f is constant on each conjugacy class. The set of class functions is written $\mathrm{Cl}(G)$.

The functions $G \to \mathbb{C}$ form a complex vector space under the pointwise operations (Example (The Standard Vector Spaces) (d)). The zero function is a class function, and sums and scalar multiples of class functions are class functions, so $\mathrm{Cl}(G)$ is a subspace of it (Theorem 1.2.1). Its dimension is exactly what we want to count.

Nonexample by minimal change. On S_3, the function f that is 1 at $(1\ 2)$ and 0 elsewhere is **not** a class function: $(1\ 3) = \sigma(1\ 2)\sigma^{-1}$ for $\sigma = (2\ 3)$, so f would have to take the same value at $(1\ 3)$, and it takes 0 there. Changing f to be 1 on **all three** transpositions repairs it; that is the indicator function of a whole class, and those are the building blocks below.

Proposition 23.8.3. The class functions have dimension the number of classes

Let G be a finite group with conjugacy classes $C_1, \dots, C_k$. For each j let $1_{C_j}\colon G \to \mathbb{C}$ be the function that is 1 on C_j and 0 elsewhere. Then $(1_{C_1}, \dots, 1_{C_k})$ is a basis of $\mathrm{Cl}(G)$, and $\dim \mathrm{Cl}(G) = k$.

Proof. Each 1_{C_j} is constant on every class — it is 1 on C_j and 0 on the others — so it lies in $\mathrm{Cl}(G)$.

Spanning. Let $f \in \mathrm{Cl}(G)$ and pick $g_j \in C_j$ for each j. Every $g \in G$ lies in exactly one class C_j (the classes partition G), and then $f(g) = f(g_j)$ while $\sum_l f(g_l) 1_{C_l}(g) = f(g_j)$. So $f = \sum_j f(g_j) 1_{C_j}$.

Independence. Suppose $\sum_j c_j 1_{C_j} = 0$. Evaluating at $g_j \in C_j$ gives $c_j = 0$, since every other 1_{C_l} vanishes there. Hence the list is a basis and $\dim \mathrm{Cl}(G) = k$.

■

Now the pairing. It is the standard inner product of $\mathbb{C}^{|G|}$, with the coordinates indexed by G and normalized by $|G|$ so that the trivial character has length 1.

Definition 23.8.4. The inner product of class functions

For functions $f, f'\colon G \to \mathbb{C}$ on a finite group G, set

$$\langle f, f' \rangle := \frac{1}{|G|} \sum_{g \in G} f(g)\, \overline{f'(g)}.$$

This satisfies (IP1)–(IP3) of Definition 11.1.1 on the space of all functions $G \to \mathbb{C}$, hence on the subspace $\mathrm{Cl}(G)$: linearity in the first slot and conjugate symmetry are immediate from the formula, and $\langle f, f \rangle = \frac{1}{|G|} \sum_g |f(g)|^2 > 0$ whenever $f \neq 0$, since some term is positive and none is negative. So $\mathrm{Cl}(G)$ is a complex inner product space of dimension k, and all of Chapter 11 is available in it.

When f and f' are class functions the sum can be shortened to one term per class: with $c_j = |C_j|$ and $g_j \in C_j$,

$$\langle f, f' \rangle = \frac{1}{|G|} \sum_{j=1}^{k} c_j \, f(g_j) \, \overline{f'(g_j)}. \tag{23.8.1}$$

That is the formula used for hand computation, and it is the only place the class sizes enter.

23.8.4 Two averaging lemmas

Both proofs of orthogonality run through the same device as Maschke's theorem: average over G, and the average is G-invariant. Here we also need its trace.

Lemma 23.8.4. The average of a representation is the projection onto the fixed space

Let ρ be a representation of a finite group G on a complex space V, and set

$$V^G := \{ v \in V : \rho(g)v = v \text{ for all } g \in G \}, \qquad P := \frac{1}{|G|} \sum_{g \in G} \rho(g).$$

Then V^G is a subspace, P is a projection with $\operatorname{im} P = V^G$, and

$$\frac{1}{|G|} \sum_{g \in G} \chi_\rho(g) = \dim V^G.$$

Proof. $V^G = \bigcap_{g \in G} \ker(\rho(g) - \mathrm{id}_V)$ is an intersection of subspaces, hence a subspace (Theorem 1.2.2).

Fix $h \in G$. Then $\rho(h)P = \frac{1}{|G|} \sum_g \rho(hg) = P$, because $g \mapsto hg$ is a bijection of G (Theorem 0.10.1) and so only reorders the terms. Hence $\rho(h)(Pv) = Pv$ for every v, giving $\operatorname{im} P \subseteq V^G$. Conversely, if $v \in V^G$ then $Pv = \frac{1}{|G|} \sum_g v = v$. So P is the identity on V^G; combined with $\operatorname{im} P \subseteq V^G$ this gives $P^2 = P$ and $\operatorname{im} P = V^G$.

Thus P is a projection (Definition 4.5.1) of rank $\dim V^G$, and $\operatorname{tr} P = \dim V^G$ by Theorem 4.5.4, the field being $\mathbb{C}$. On the other hand the trace is linear (Theorem 4.5.3), so $\operatorname{tr} P = \frac{1}{|G|} \sum_g \operatorname{tr} \rho(g) = \frac{1}{|G|} \sum_g \chi_\rho(g)$. This proves the lemma. ∎

Lemma 23.8.5. The trace of a two-sided multiplication

Let V and W be finite-dimensional vector spaces over a field F, let $A \in \mathcal{L}(V)$ and $B \in \mathcal{L}(W)$, and let

$$\Phi \colon \mathcal{L}(V, W) \to \mathcal{L}(V, W), \qquad \Phi(f) = B \circ f \circ A.$$

Then Φ is linear and $\operatorname{tr} \Phi = (\operatorname{tr} A)(\operatorname{tr} B)$.

Idea. Write everything in coordinates. In the basis of $\mathcal{L}(V,W)$ given by the matrix units $\mathbf{E}_{ij}$, the map $\mathbf{X}\mapsto \mathbf{BXA}$ sends $\mathbf{E}_{ij}$ to a matrix whose (i,j)-entry is $b_{ii}a_{jj}$; adding those diagonal coefficients factorizes as a product of two sums.

Proof. Linearity of Φ is immediate from bilinearity of composition. Fix bases of V and W and identify $\mathcal{L}(V,W)$ with $M_{m\times n}(F)$, where $n=\dim V$ and $m=\dim W$ (Theorem 4.1.4); let $\mathbf{A}=(a_{jl})\in M_n(F)$ and $\mathbf{B}=(b_{ki})\in M_m(F)$ be the matrices of A and B. Under the identification Φ becomes $\mathbf{X}\mapsto\mathbf{BXA}$.

Let $\mathbf{E}_{ij}\in M_{m\times n}(F)$ be the matrix unit with a 1 in position (i,j) and zeros elsewhere — the rectangular version of the matrix units of Lemma 23.2.1; these mn matrices form a basis. By Definition 0.9.9,

$$(\mathbf{B}\,\mathbf{E}_{ij}\,\mathbf{A})_{kl}=\sum_{p,q} b_{kp}(\mathbf{E}_{ij})_{pq}a_{ql}=b_{ki}\,a_{jl}.$$

In the basis $(\mathbf{E}_{ij})$, the coefficient of $\mathbf{E}_{ij}$ in $\Phi(\mathbf{E}_{ij})$ is the (i,j)-entry of $\mathbf{B}\mathbf{E}_{ij}\mathbf{A}$, namely $b_{ii}a_{jj}$. These coefficients are the diagonal entries of the matrix of Φ, so by Definition 4.5.2,

$$\operatorname{tr}\Phi=\sum_{i=1}^{m}\sum_{j=1}^{n}b_{ii}a_{jj}=\Big(\sum_i b_{ii}\Big)\Big(\sum_j a_{jj}\Big)=(\operatorname{tr}\mathbf{B})(\operatorname{tr}\mathbf{A}),$$

and $\operatorname{tr}\mathbf{A}=\operatorname{tr}A$, $\operatorname{tr}\mathbf{B}=\operatorname{tr}B$ by Definition 4.5.2. This proves the lemma.

■

Now put the two together. Given representations ρ on V and σ on W, make G act on the space $\mathcal{L}(V,W)$ of linear maps by

$$\pi(g)f\coloneqq\sigma(g)\circ f\circ\rho(g)^{-1}.$$

Proposition 23.8.6. The representation on the space of linear maps

Let ρ,σ be representations of a finite group G on complex spaces V,W. Then π above is a representation of G on $\mathcal{L}(V,W)$, its character is $\chi_\pi(g)=\chi_\sigma(g)\overline{\chi_\rho(g)}$, and its fixed space is the space

$$\mathcal{L}_G(V,W)\coloneqq\{f\in\mathcal{L}(V,W):f\rho(g)=\sigma(g)f\ \text{for all } g\in G\}$$

of maps intertwining ρ and σ (Definition 23.7.3); for $W=V$ and $\sigma=\rho$ this is the commutant $\operatorname{End}_{\mathbb{C}[G]}(V)$ of Definition 23.4.1. Consequently

$$\frac{1}{|G|}\sum_{g\in G}\chi_\sigma(g)\overline{\chi_\rho(g)}=\dim\mathcal{L}_G(V,W).\tag{23.8.2}$$

Proof. A representation. Each $\pi(g)$ is linear in f, and $\pi(g)\pi(h)f=\sigma(g)\sigma(h)f\rho(h)^{-1}\rho(g)^{-1}=\sigma(gh)f\rho(gh)^{-1}=\pi(gh)f$, using the homomorphism property of ρ and σ and Theorem 0.10.1 for the inverse of a product. Taking $h=g^{-1}$ shows $\pi(g)$ is invertible with inverse $\pi(g^{-1})$.

Its character. Apply Lemma 23.8.5 with $A = \rho(g)^{-1}$ and $B = \sigma(g)$:

$$\begin{aligned}\chi_\pi(g) &= \operatorname{tr}\pi(g) = \operatorname{tr}(\rho(g)^{-1})\ \operatorname{tr}(\sigma(g)) \\ &= \chi_\rho(g^{-1})\chi_\sigma(g) = \overline{\chi_\rho(g)}\,\chi_\sigma(g),\end{aligned}$$

the last step by Proposition 23.8.2 (e).

Its fixed space. $\pi(g)f = f$ means $\sigma(g)f\rho(g)^{-1} = f$, and composing with $\rho(g)$ on the right turns this into $\sigma(g)f = f\rho(g)$. So the fixed space is exactly $\mathcal{L}_G(V, W)$.

The count. Apply Lemma 23.8.4 to π and substitute the character just computed. This proves the proposition.

■

23.8.5 Orthogonality of the irreducible characters

Everything is now in place. Schur's lemma converts the right-hand side of Eq. 23.8.2 into a 0 or a 1.

Theorem 23.8.7. Orthogonality of Irreducible Characters

Let G be a finite group and let ρ on V and σ on W be **irreducible** complex representations of G. Then

$$\langle \chi_\sigma, \chi_\rho \rangle = \begin{cases} 1, & \text{if } \rho \cong \sigma, \\ 0, & \text{if } \rho \ncong \sigma. \end{cases}$$

In particular the characters of pairwise inequivalent irreducible complex representations form an orthonormal list in $\mathrm{Cl}(G)$.

Idea. By Eq. 23.8.2, $\langle \chi_\sigma, \chi_\rho \rangle$ counts the dimension of the space of intertwining maps $V \to W$. Schur's lemma says that space is $\{0\}$ when the two are inequivalent, because a nonzero intertwiner would be an isomorphism. When they are equivalent, fix one isomorphism S_0; every other intertwiner differs from it by an intertwiner of V with itself, which over $\mathbb{C}$ is a scalar. So the space is the line $\mathbb{C}S_0$.

Proof. By Eq. 23.8.2, $\langle \chi_\sigma, \chi_\rho \rangle = \dim \mathcal{L}_G(V, W)$, which is a nonnegative integer. It remains to compute that dimension.

Case 1: $\rho \ncong \sigma$. By Theorem 23.7.1, V and W are irreducible $\mathbb{C}[G]$-spaces; and since the condition $f\rho(g) = \sigma(g)f$ for all $g \in G$ extends by linearity to all of $\mathbb{C}[G]$, the members of $\mathcal{L}_G(V, W)$ are exactly the $\mathbb{C}[G]$-maps $V \to W$ (Definition 23.4.1). Let $f \in \mathcal{L}_G(V, W)$ be nonzero. By Theorem 23.4.1 (a) it is an isomorphism, making ρ and σ equivalent (Definition 23.7.3) — a contradiction. Hence $\mathcal{L}_G(V, W) = \{0\}$ and the inner product is 0.

Case 2: $\rho \cong \sigma$. Fix an isomorphism $S_0 \in \mathcal{L}_G(V, W)$, so $S_0\rho(g) = \sigma(g)S_0$ for all g. Applying S_0^{-1} on both sides gives $\rho(g)S_0^{-1} = S_0^{-1}\sigma(g)$, so S_0^{-1} intertwines σ with ρ. Let $f \in \mathcal{L}_G(V, W)$ and set $u := S_0^{-1}f \in \mathcal{L}(V)$. Then for every g,

$$u\rho(g) = S_0^{-1}f\rho(g) = S_0^{-1}\sigma(g)f = \rho(g)S_0^{-1}f = \rho(g)u,$$

so u commutes with every $\rho(g)$. Since ρ is irreducible and the field is algebraically closed, Corollary 23.4.3 gives $u = \lambda \,\mathrm{id}_V$ for some $\lambda \in \mathbb{C}$, and hence $f = S_0 u = \lambda S_0$. Therefore $\mathcal{L}_G(V, W) = \mathbb{C}S_0$, which has dimension 1 because $S_0 \neq 0$.

The last sentence of the statement follows: distinct members of a list of pairwise inequivalent irreducibles pair to 0, and each pairs with itself to 1. This proves the theorem.

■

Two consequences come for free, and both are used in every computation below.

Corollary 23.8.8. Irreducibility criterion

Let ρ be a complex representation of a finite group G on $V \neq \{0\}$, with character χ. Write $V = V_1 \oplus \cdots \oplus V_r$ with each V_i irreducible (Corollary 23.7.4), and let $\chi_1, \dots, \chi_s$ be the distinct characters occurring among the $\chi_{\rho|_{V_i}}$, with multiplicities $m_1, \dots, m_s \geq 1$. Then

$$\chi = \sum_{i=1}^{s} m_i \chi_i, \qquad m_i = \langle \chi, \chi_i \rangle, \qquad \langle \chi, \chi \rangle = \sum_{i=1}^{s} m_i^2.$$

In particular $\langle \chi, \chi \rangle = 1$ if and only if ρ is irreducible, and two complex representations of G with the same character are equivalent.

Proof. Repeated use of Proposition 23.8.2 (d) gives $\chi = \sum_{i=1}^{r} \chi_{\rho|_{V_i}}$, and grouping equal terms gives $\chi = \sum_i m_i \chi_i$. Distinct characters come from inequivalent representations, since equivalent ones have equal characters (Proposition 23.8.2 (c)); so by Theorem 23.8.7 the χ_i are orthonormal, and $\langle \chi, \chi_i \rangle = m_i$ and, expanding the pairing in the first slot and then in the second, $\langle \chi, \chi \rangle = \sum_i m_i^2$.

If ρ is irreducible then $s = 1$ and $m_1 = 1$, so $\langle \chi, \chi \rangle = 1$. Conversely, a sum of squares of positive integers equals 1 only when there is a single term equal to 1; then $r = 1$ and $V = V_1$ is irreducible.

For the last claim, suppose ρ and ρ' have the same character. Every irreducible character ψ of G then has $\langle \chi_\rho, \psi \rangle = \langle \chi_{\rho'}, \psi \rangle$, so each irreducible occurs with the same multiplicity in both decompositions. So the two decompositions into irreducibles use the same irreducibles with the same multiplicities, and the summands can be matched in pairs with equal characters. Two irreducible summands with equal characters are equivalent: were they inequivalent, Theorem 23.8.7 would make their characters orthogonal, hence unequal, since an irreducible character pairs with itself to 1 and not to 0. The direct sum of those equivalences is an isomorphism $V \to V'$ intertwining ρ and ρ'. This proves the corollary.

■

Example. Checking the standard representation by its character

Use Corollary 23.8.8 to confirm that the two-dimensional representation of S_3 built in Example (The irreducible complex representations of the symmetric group on three letters) is irreducible, and decompose the permutation representation of S_3 on $\mathbb{C}^3$.

Solution. From Example (The irreducible complex representations of the symmetric group on three letters) the matrices give $\chi_{\mathrm{std}}(\mathrm{id}) = 2$, $\chi_{\mathrm{std}}((1\,2)) = -1+1 = 0$, and $\chi_{\mathrm{std}}((1\,2\,3)) = 0 + (-1) = -1$; by Proposition 23.8.2 (b) these three values determine χ_{std} on all of S_3, the classes being those of Example (The conjugacy classes of the symmetric group on three letters). Using Eq. 23.8.1 with class sizes $1, 3, 2$,

$$\langle \chi_{\mathrm{std}}, \chi_{\mathrm{std}} \rangle = \tfrac{1}{6}(1 \cdot 2^2 + 3 \cdot 0^2 + 2 \cdot (-1)^2) = \tfrac{4+0+2}{6} = 1,$$

so the representation is irreducible.

For the permutation representation, χ_{perm} counts fixed points: 3 at id, 1 at a transposition, 0 at a 3-cycle. Then

$$\begin{aligned} \langle \chi_{\mathrm{perm}}, \chi_{\mathrm{triv}} \rangle &= \tfrac{1}{6}(1 \cdot 3 + 3 \cdot 1 + 2 \cdot 0) = 1, \\ \langle \chi_{\mathrm{perm}}, \chi_{\mathrm{sgn}} \rangle &= \tfrac{1}{6}(1 \cdot 3 + 3 \cdot 1 \cdot (-1) + 2 \cdot 0) = 0, \\ \langle \chi_{\mathrm{perm}}, \chi_{\mathrm{std}} \rangle &= \tfrac{1}{6}(1 \cdot 3 \cdot 2 + 0 + 0) = 1. \end{aligned}$$

So $\chi_{\mathrm{perm}} = \chi_{\mathrm{triv}} + \chi_{\mathrm{std}}$, and by Corollary 23.8.8 the permutation representation is the direct sum of the trivial and the standard one. That matches the splitting $\mathbb{C}^3 = \mathrm{span}(\mathbf{e}_1+\mathbf{e}_2+\mathbf{e}_3) \oplus W$ found by hand in Example (The irreducible complex representations of the symmetric group on three letters), and it was found here without looking at a single vector.

23.8.6 The characters span the class functions

Orthonormality already bounds the number of irreducible representations: an orthonormal list in $\mathrm{Cl}(G)$ is independent (Theorem 11.1.4), so there are at most k of them up to equivalence. The next theorem shows there are exactly k. It is the deepest statement of the section, and its proof is the third and last averaging argument: this time we average the group elements themselves, weighted by a class function.

Theorem 23.8.9. The Irreducible Characters Are a Basis of the Class Functions

Let G be a finite group with k conjugacy classes. Then G has exactly k irreducible complex representations up to equivalence, and their characters $\chi_1, \dots, \chi_k$ form an orthonormal **basis** of $\mathrm{Cl}(G)$.

Idea. Suppose a class function f is orthogonal to every irreducible character; we must show $f = 0$. Build from f the operator $T_f = \sum_g \overline{f(g)}\rho(g)$ on any representation ρ. ① Because f is a class function, T_f commutes with every $\rho(h)$ — conjugating by $\rho(h)$ permutes the terms and leaves the coefficients alone. ② So on an irreducible ρ it is a scalar, and its trace, which is $|G|\langle \chi_\rho, f \rangle = 0$, forces that scalar to be 0. ③ Complete reducibility (Corollary 23.7.4) spreads this to every representation. ④ Apply it to the regular representation and evaluate at δ_1: the result is $\sum_g \overline{f(g)}\delta_g = 0$, and the δ_g are a basis.

Proof. Orthonormal lists are independent (Theorem 11.1.4) and $\dim \mathrm{Cl}(G) = k$ (Proposition 23.8.3), so by Theorem 23.8.7 any list of pairwise inequivalent irreducible representations has at most k members. Choose such a list $\rho_1, \dots, \rho_s$ with s as large as possible; by maximal-

ity every irreducible complex representation of G is equivalent to some ρ_i. Write $\chi_i = \chi_{\rho_i}$, and let $U = \operatorname{span}(\chi_1, \dots, \chi_s) \subseteq \mathrm{Cl}(G)$, using Proposition 23.8.2 (b).

Let $f \in U^{\perp}$, the orthogonal complement of U in $\mathrm{Cl}(G)$ (Definition 11.3.1), so f is a class function with $\langle f, \chi_i \rangle = 0$, hence also $\langle \chi_i, f \rangle = \overline{\langle f, \chi_i \rangle} = 0$, for every i. For **any** representation ρ of G on V, define

$$T_f^{\rho} := \sum_{g \in G} \overline{f(g)}\, \rho(g) \in \mathcal{L}(V).$$

Claim. **Claim 1.** T_f^{ρ} commutes with $\rho(h)$ for every $h \in G$.

Proof. Fix h. Then

$$\rho(h)\, T_f^{\rho}\, \rho(h)^{-1} = \sum_{g \in G} \overline{f(g)}\, \rho(hgh^{-1}).$$

The map $g \mapsto hgh^{-1}$ is a bijection $G \to G$, with inverse $g \mapsto h^{-1}gh$. Substituting $g' = hgh^{-1}$, so that $g = h^{-1}g'h$, the sum becomes $\sum_{g'} \overline{f(h^{-1}g'h)}\rho(g')$. Since f is a class function, $f(h^{-1}g'h) = f(g')$, so the sum is T_f^{ρ}. Multiplying on the right by $\rho(h)$ gives the claim.

■

Claim. **Claim 2.** If ρ is irreducible, then $T_f^{\rho} = 0$.

Proof. By Claim 1 and Corollary 23.4.3, over the algebraically closed field $\mathbb{C}$ we get $T_f^{\rho} = \lambda\, \mathrm{id}_V$ for some $\lambda \in \mathbb{C}$. Taking traces and using linearity of the trace (Theorem 4.5.3),

$$\lambda \dim V = \operatorname{tr} T_f^{\rho} = \sum_{g} \overline{f(g)}\, \chi_{\rho}(g) = |G|\, \langle \chi_{\rho}, f \rangle.$$

Now ρ is equivalent to some ρ_i, so $\chi_{\rho} = \chi_i$ by Proposition 23.8.2 (c), and $\langle \chi_i, f \rangle = 0$. Since $\dim V \geq 1$, $\lambda = 0$.

■

Claim. **Claim 3.** $T_f^{\rho} = 0$ for **every** representation ρ.

Proof. We may assume $V \neq \{0\}$. By Corollary 23.7.4, $V = V_1 \oplus \cdots \oplus V_r$ with each V_j invariant and $\rho|_{V_j}$ irreducible. Each $\rho(g)$ maps V_j into V_j, hence so does T_f^{ρ}, and its restriction to V_j is $\sum_g \overline{f(g)}\, \rho|_{V_j}(g) = T_f^{\rho|_{V_j}}$, which is 0 by Claim 2. So T_f^{ρ} vanishes on each V_j, hence on their sum V.

■

Now apply Claim 3 to the regular representation ρ_{reg} on $\mathbb{C}[G]$, where $\rho_{\mathrm{reg}}(g)\delta_h = \delta_{gh}$. Evaluating at the basis vector δ_1,

$$0 = T_f^{\rho_{\mathrm{reg}}} \delta_1 = \sum_{g \in G} \overline{f(g)}\, \delta_g.$$

The δ_g are a basis of $\mathbb{C}[G]$, so every coefficient vanishes: $\overline{f(g)} = 0$, that is $f(g) = 0$, for all g. Hence $U^{\perp} = \{0\}$.

By Theorem 11.3.2, $\mathrm{Cl}(G) = U \oplus U^{\perp} = U$. So $(\chi_1, \dots, \chi_s)$ is an orthonormal spanning list, hence a basis, and $s = \dim \mathrm{Cl}(G) = k$ by Proposition 23.8.3. Finally, inequivalent irreducibles have different characters, since orthonormality gives $\langle \chi_i, \chi_j \rangle = 0 \neq 1 = \langle \chi_i, \chi_i \rangle$ for $i \neq j$; so

the k characters correspond to the k equivalence classes of irreducible representations. This proves the theorem.

■

Corollary 23.8.10. The sum of the squares of the degrees

Let G be a finite group and let $\rho_1, \dots, \rho_k$ on $V_1, \dots, V_k$ be a complete list of pairwise inequivalent irreducible complex representations of G, one for each of the k conjugacy classes (Theorem 23.8.9). Then each ρ_i occurs in the regular representation with multiplicity exactly $\dim V_i$, and

$$|G| = \sum_{i=1}^{k} (\dim V_i)^2.$$

Proof. The regular character is $\chi_{\mathrm{reg}}(1) = |G|$ and $\chi_{\mathrm{reg}}(g) = 0$ for $g \neq 1$, as computed after Definition 23.8.2. Hence

$$\langle \chi_{\mathrm{reg}}, \chi_i \rangle = \frac{1}{|G|} \sum_{g} \chi_{\mathrm{reg}}(g)\overline{\chi_i(g)} = \frac{1}{|G|}\,|G|\,\overline{\chi_i(1)} = \dim V_i,$$

the value $\chi_i(1) = \dim V_i$ being a real number (Proposition 23.8.2 (a)). By Corollary 23.8.8 this inner product is the multiplicity of ρ_i in ρ_{reg}, so $\chi_{\mathrm{reg}} = \sum_i (\dim V_i)\chi_i$. Evaluating at $g = 1$ gives $|G| = \sum_i (\dim V_i)\chi_i(1) = \sum_i (\dim V_i)^2$.

■

This is Corollary 23.6.2, the count $\dim A = \sum n_i^2$ of the structure theorem, applied to $A = \mathbb{C}[G]$: the two proofs are different and the answer is the same.

Corollary 23.8.11. Orthogonality of the columns

Let G be a finite group with conjugacy classes $C_1, \dots, C_k$, representatives $g_j \in C_j$ and sizes $c_j = |C_j|$, and irreducible characters $\chi_1, \dots, \chi_k$. Then for all j, l,

$$\sum_{i=1}^{k} \chi_i(g_j)\,\overline{\chi_i(g_l)} = \begin{cases} |G|/c_j, & j = l, \\ 0, & j \neq l. \end{cases}$$

Idea. Theorem 23.8.9 makes the table of character values a **square** array. Rescale it so that orthonormality of the rows says the matrix is unitary; then its columns are orthonormal too, because a one-sided inverse of a square matrix is two-sided.

Proof. Let $\mathbf{U} \in M_k(\mathbb{C})$ have entries $u_{ij} = \sqrt{c_j/|G|}\;\chi_i(g_j)$; the square root is a positive real.

By Eq. 23.8.1 and Theorem 23.8.7,

$$(\mathbf{U}\mathbf{U}^{*})_{ii'} = \sum_{j=1}^{k} u_{ij}\overline{u_{i'j}} = \frac{1}{|G|}\sum_{j=1}^{k} c_j\,\chi_i(g_j)\overline{\chi_{i'}(g_j)} = \langle \chi_i, \chi_{i'} \rangle,$$

which is 1 for $i = i'$ and 0 otherwise. Hence $\mathbf{U}\mathbf{U}^{*} = \mathbf{I}_k$. By Theorem 3.4.5, $\mathbf{U}^{*}\mathbf{U} = \mathbf{I}_k$ as well, so for all j, l

$$\sum_{i=1}^{k} \overline{u_{ij}}\,u_{il} = \frac{\sqrt{c_j c_l}}{|G|}\sum_{i=1}^{k} \overline{\chi_i(g_j)}\,\chi_i(g_l)$$

equals 1 when $j = l$ and 0 otherwise. For $j = l$ this gives $\sum_i |\chi_i(g_j)|^2 = |G|/c_j$, and for $j \neq l$ it gives $\sum_i \overline{\chi_i(g_j)}\chi_i(g_l) = 0$. Conjugating the second identity, which is a sum of complex numbers, gives the stated form. This proves the corollary.

■

23.8.7 Two character tables

The **character table** of G is the $k \times k$ array whose (i, j)-entry is $\chi_i(g_j)$: one row per irreducible character, one column per conjugacy class. By Theorem 23.8.9 it is square, by Theorem 23.8.7 its rows are orthonormal, and by Corollary 23.8.11 its columns are orthogonal.

Example. The character table of the symmetric group on three letters

Write down the character table of S_3 and verify every orthogonality relation.

Solution. By Example (The conjugacy classes of the symmetric group on three letters) there are $k = 3$ classes, with representatives id, (1 2), (1 2 3) and sizes $1, 3, 2$. So there are three irreducible complex representations (Theorem 23.8.9), and Example (The irreducible complex representations of the symmetric group on three letters) produced three: the trivial, the sign, and the standard one. Their characters, from the definitions and the traces computed in Example (Checking the standard representation by its character), are:

	id	(1 2)	(1 2 3)
class size	1	3	2
χ_{triv}	1	1	1
χ_{sgn}	1	-1	1
χ_{std}	2	0	-1

Degrees. $1^2 + 1^2 + 2^2 = 6 = |S_3|$, as Corollary 23.8.10 requires.

Row orthogonality, from Eq. 23.8.1 with weights $1, 3, 2$ and divisor 6. All values are real, so conjugation does nothing:

$$\begin{aligned}
\langle \chi_{\text{triv}}, \chi_{\text{triv}} \rangle &= \tfrac{1}{6}(1+3+2) = 1, \\
\langle \chi_{\text{sgn}}, \chi_{\text{sgn}} \rangle &= \tfrac{1}{6}(1+3+2) = 1, \\
\langle \chi_{\text{std}}, \chi_{\text{std}} \rangle &= \tfrac{1}{6}(4+0+2) = 1, \\
\langle \chi_{\text{triv}}, \chi_{\text{sgn}} \rangle &= \tfrac{1}{6}(1-3+2) = 0, \\
\langle \chi_{\text{triv}}, \chi_{\text{std}} \rangle &= \tfrac{1}{6}(2+0-2) = 0, \\
\langle \chi_{\text{sgn}}, \chi_{\text{std}} \rangle &= \tfrac{1}{6}(2+0-2) = 0.
\end{aligned}$$

Column orthogonality, from Corollary 23.8.11, with no weights this time:

$$\begin{aligned}
\text{column id}: &\quad 1+1+4 = 6 = 6/1, \\
\text{column } (1\ 2): &\quad 1+1+0 = 2 = 6/3, \\
\text{column } (1\ 2\ 3): &\quad 1+1+1 = 3 = 6/2, \\
\text{id against } (1\ 2): &\quad 1\cdot 1 + 1\cdot(-1) + 2\cdot 0 = 0, \\
\text{id against } (1\ 2\ 3): &\quad 1+1-2 = 0, \\
(1\ 2) \text{ against } (1\ 2\ 3): &\quad 1-1+0 = 0.
\end{aligned}$$

Every relation holds.

Example. The character table of a cyclic group of order four

Write down the character table of $\mathbb{Z}/4\mathbb{Z}$ and verify the orthogonality relations.

Solution. The group is abelian, so every conjugacy class is a single element and $k = 4$. By Theorem 23.8.9 there are four irreducible complex representations, and Example (The irreducible complex representations of a cyclic group) produced four, all of degree 1: $\chi_m([j]) = i^{jm}$, using $\omega = e^{2\pi i/4} = i$. Their degrees satisfy $1^2 + 1^2 + 1^2 + 1^2 = 4 = |G|$.

	$[0]$	$[1]$	$[2]$	$[3]$
χ_0	1	1	1	1
χ_1	1	i	-1	$-i$
χ_2	1	-1	1	-1
χ_3	1	$-i$	-1	i

Row orthogonality. For characters of degree 1, $\overline{\chi_l([j])} = i^{-jl}$, so

$$\langle \chi_m, \chi_l \rangle = \frac{1}{4} \sum_{j=0}^{3} i^{j(m-l)}.$$

If $m = l$ every term is 1 and the sum is $4/4 = 1$. If $m \neq l$, put $d = m - l$, so $i^d \neq 1$ and $(i^d)^4 = 1$; then $(i^d - 1)\sum_{j=0}^{3} (i^d)^j = (i^d)^4 - 1 = 0$, and dividing by $i^d - 1 \neq 0$ gives $\sum_j i^{jd} = 0$. For instance $\langle \chi_1, \chi_0 \rangle = \frac{1}{4}(1 + i - 1 - i) = 0$ and $\langle \chi_3, \chi_1 \rangle = \frac{1}{4}(1 - 1 + 1 - 1) = 0$.

Column orthogonality. Every class has size 1, so Corollary 23.8.11 predicts 4 down the diagonal and 0 off it. The same geometric sum settles every one of them: column $[j]$ against column $[l]$ gives $\sum_{m=0}^{3} \chi_m([j])\overline{\chi_m([l])} = \sum_{m=0}^{3} i^{m(j-l)}$, which is 4 when $j = l$, and is 0 otherwise by the computation just made, read with $d = j - l$. Three of them written out. Column [1] against itself: $1 + |i|^2 + 1 + |-i|^2 = 4$. Column [0] against [2]: the entries of column [0] are $1, 1, 1, 1$ and those of column [2] are $1, -1, 1, -1$, so the sum is $1 \cdot 1 + 1 \cdot (-1) + 1 \cdot 1 + 1 \cdot (-1) = 0$. Column [1] against [3]: $1 + i \cdot \overline{(-i)} + (-1)\overline{(-1)} + (-i)\bar{i} = 1 + i \cdot i + 1 + (-i)(-i) = 1 - 1 + 1 - 1 = 0$.

Read as a matrix, the table is $(i^{jm})_{m,j}$, which up to the factor $1/2$ is the 4×4 Fourier matrix of Example (The Fourier matrix of size four). That is not a coincidence, and Section 09 explains it.

Warning. **The character table does not determine the group.** Two nonisomorphic groups can have identical character tables. The standard witness is a pair of groups of order 8: the eight symmetries of a square (four rotations and four reflections), and the group $\{\pm \mathbf{I}_2, \pm i, \pm j, \pm k\}$ sitting inside the quaternions $\mathbb{H}$ of Definition 14.12.3 under multiplication. Both have five conjugacy classes and degrees $1, 1, 1, 1, 2$, and their tables agree entry for entry, yet the groups are not isomorphic. **This is stated on credit and not proved here**; nothing below depends on it. What the character table *does* determine is every complex representation of the group, up to equivalence (Corollary 23.8.8).

23.8.8 Exercises

23.8.8.1 A. Check your understanding

Exercise. A1

(a) Define the character of a representation, and state why it does not depend on a choice of basis.

(b) State the orthogonality relation for irreducible complex characters.

(c) Determine whether the following is true: "if χ is a character of G then $\chi(gh) = \chi(g)\chi(h)$." Justify your answer.

(d) A group of order 12 has exactly four conjugacy classes. Determine the degrees of its irreducible complex representations.

(e) Explain why the number of irreducible complex representations of an abelian group G is $|G|$.

Solution.

(a) $\chi_\rho(g) = \operatorname{tr}\rho(g)$ (Definition 23.8.2). The trace of an operator is computed from its matrix in any basis and the answer is the same for all of them (Definition 4.5.2), so χ_ρ depends only on ρ.

(b) For irreducible complex representations ρ, σ of a finite group, $\langle \chi_\sigma, \chi_\rho \rangle$ is 1 if $\rho \cong \sigma$ and 0 otherwise (Theorem 23.8.7).

(c) False. See the warning after Proposition 23.8.2: for the standard representation of S_3, $\chi((1\ 2))\chi((1\ 3)) = 0$ but $\chi((1\ 3\ 2)) = -1$. It is true precisely for characters of degree 1.

(d) Four classes means four irreducible representations (Theorem 23.8.9), with degrees $d_1, \dots, d_4 \geq 1$ satisfying $d_1^2 + d_2^2 + d_3^2 + d_4^2 = 12$ (Corollary 23.8.10). Every group has the trivial representation, so we may take $d_1 = 1$ and need $d_2^2 + d_3^2 + d_4^2 = 11$ with each $d_i \geq 1$. The squares available are $1, 4, 9$, since $4^2 = 16 > 11$. A triple containing a 9 has its other two terms summing to 2, so it is $9+1+1 = 11$, which works; and a triple with no 9 is one of $4+4+4 = 12$, $4+4+1 = 9$, $4+1+1 = 6$, and $1+1+1 = 3$, none of them 11. So the degrees are $1, 1, 1, 3$.

(e) In an abelian group $xgx^{-1} = g$, so every conjugacy class is a singleton and $k = |G|$; now apply Theorem 23.8.9.

23.8.8.2 B. Practice

Exercise. B1: A character computed and decomposed

Let $\mathbf{P}_\tau$ be the permutation matrix of $\tau \in S_3$, so that $\mathbf{P}_\tau \mathbf{e}_i = \mathbf{e}_{\tau(i)}$ (Definition 3.4.2), and let π be the representation of S_3 on $\mathcal{L}(\mathbb{C}^3)$ given by $\pi(\tau)(f) = \mathbf{P}_\tau\, f\, \mathbf{P}_\tau^{-1}$.

(a) Compute the character of π on the three conjugacy classes.

(b) Decompose π into irreducibles.

Solution.

(a) This is the construction of Proposition 23.8.6 with both representations equal to the permutation representation ρ_{perm} on $\mathbb{C}^3$. So its character is $\chi_\pi(\tau) = \chi_{\text{perm}}(\tau)\overline{\chi_{\text{perm}}(\tau)} = |\chi_{\text{perm}}(\tau)|^2$. From Example (Checking the standard representation by its character), $\chi_{\text{perm}} = (3, 1, 0)$ on the classes of id, $(1\ 2)$, $(1\ 2\ 3)$, so $\chi_\pi = (9, 1, 0)$.

(b) With weights $1, 3, 2$ and divisor 6:

$$\begin{aligned} \langle \chi_\pi, \chi_{\text{triv}} \rangle &= \tfrac{1}{6}(9 + 3 + 0) = 2, \\ \langle \chi_\pi, \chi_{\text{sgn}} \rangle &= \tfrac{1}{6}(9 - 3 + 0) = 1, \\ \langle \chi_\pi, \chi_{\text{std}} \rangle &= \tfrac{1}{6}(9 \cdot 2 + 0 + 0) = 3. \end{aligned}$$

So $\chi_\pi = 2\chi_{\text{triv}} + \chi_{\text{sgn}} + 3\chi_{\text{std}}$, and the degrees check: $2 + 1 + 6 = 9 = \dim \mathcal{L}(\mathbb{C}^3)$. As a check on the arithmetic, $\langle \chi_\pi, \chi_\pi \rangle = \frac{1}{6}(81+3+0) = 14 = 2^2+1^2+3^2$, as Corollary 23.8.8 requires. In particular $\dim \mathcal{L}_G(\mathbb{C}^3, \mathbb{C}^3) = 2$: the G-equivariant operators on $\mathbb{C}^3$ form a plane, spanned by the identity and the all-ones matrix.

Exercise. B2: Which of these are characters?

Let $G = S_3$, with classes ordered id, (1 2), (1 2 3) of sizes $1, 3, 2$. Determine which of the following class functions are characters of complex representations of S_3. Justify your answer.

(a) $(4, 0, 1)$;
(b) $(3, 1, 0)$;
(c) $(2, 2, -1)$;
(d) $(1, 1, -1)$.

Solution. A class function is a character exactly when it is a **nonnegative integer** combination of $\chi_{\text{triv}} = (1, 1, 1)$, $\chi_{\text{sgn}} = (1, -1, 1)$, and $\chi_{\text{std}} = (2, 0, -1)$, by Corollary 23.7.4 and Proposition 23.8.2 (d); the coefficients are the inner products $\langle \cdot, \chi_i \rangle$.

(a) $\langle (4,0,1), \chi_{\text{triv}} \rangle = \frac{1}{6}(4 + 0 + 2) = 1$, $\langle \cdot, \chi_{\text{sgn}} \rangle = \frac{1}{6}(4 - 0 + 2) = 1$, $\langle \cdot, \chi_{\text{std}} \rangle = \frac{1}{6}(8 + 0 - 2) = 1$. All nonnegative integers, and $1 \cdot 1 + 1 \cdot 1 + 1 \cdot 2 = 4$ matches the degree. **Yes**: it is the character of the direct sum of the trivial, the sign, and the standard representation, equivalently of the direct sum of the permutation representation and the sign representation, since $(3, 1, 0) + (1, -1, 1) = (4, 0, 1)$.

(b) Inner products $1, 0, 1$ as computed in Example (Checking the standard representation by its character). **Yes**: the permutation representation.

(c) $\langle \cdot, \chi_{\text{triv}} \rangle = \frac{1}{6}(2 + 6 - 2) = 1$, $\langle \cdot, \chi_{\text{sgn}} \rangle = \frac{1}{6}(2 - 6 - 2) = -1$. A negative multiplicity is impossible, so **no**. (It is a difference of characters, called a virtual character, but not a character.)

(d) $\langle (1,1,-1), \chi_{\text{triv}} \rangle = \frac{1}{6}(1 + 3 - 2) = \frac{1}{3}$, which is not an integer. So **no**: it is not even a combination of the three irreducible characters with integer coefficients, let alone nonnegative ones.

Exercise. B3: Degree-one characters of a cyclic group

Let $G = \mathbb{Z}/6\mathbb{Z}$.

(a) Write down the character table.
(b) Verify $\langle \chi_1, \chi_4 \rangle = 0$ and $\langle \chi_2, \chi_2 \rangle = 1$ by direct computation.

Solution.

(a) The group is abelian of order 6, so there are six classes, six irreducible characters, all of degree 1, given by $\chi_m([j]) = \omega^{jm}$ with $\omega = e^{2\pi i/6}$ (Example (The irreducible complex representations of a cyclic group)). The table has (m, j)-entry ω^{jm}, with $\omega^6 = 1$.

(b) Here

$$\begin{aligned}\langle \chi_1, \chi_4 \rangle &= \tfrac{1}{6} \sum_{j=0}^{5} \omega^j \overline{\omega^{4j}} = \tfrac{1}{6} \sum_j \omega^{-3j} \\ &= \tfrac{1}{6} \sum_j (-1)^j = \tfrac{1}{6}(1 - 1 + 1 - 1 + 1 - 1) = 0,\end{aligned}$$

using $\omega^{-3} = \omega^3 = -1$. And $\langle \chi_2, \chi_2 \rangle = \frac{1}{6}\sum_j \omega^{2j}\overline{\omega^{2j}} = \frac{1}{6}\sum_j |\omega^{2j}|^2 = \frac{1}{6} \cdot 6 = 1$.

23.8.8.3 C. Going deeper

Exercise. C1: Counting fixed points

Let G be a finite group acting on $\{1, \dots, n\}$ through a homomorphism $\varphi\colon G \to S_n$, and let $\rho(g) = \mathbf{P}_{\varphi(g)}$ be the corresponding permutation representation on $\mathbb{C}^n$, where $\mathbf{P}_\sigma$ is the permutation matrix of Definition 3.4.2 (so $\mathbf{P}_\sigma \mathbf{P}_\tau = \mathbf{P}_{\sigma\tau}$ by Lemma 3.4.7 (a)).

(a) Prove that $\langle \chi_\rho, \chi_{\text{triv}} \rangle$ equals the average number of fixed points of the permutations $\varphi(g)$, and also equals $\dim(\mathbb{C}^n)^G$.

(b) Deduce that for every $n \geq 1$, the average number of fixed points of a permutation in S_n is exactly 1.

(c) For $G = S_3$ with $\varphi = \text{id}$, verify (a) numerically.

Solution.

(a) By the computation after Definition 23.8.2, $\chi_\rho(g)$ is the number of i with $\varphi(g)(i) = i$. Hence

$$\langle \chi_\rho, \chi_{\text{triv}} \rangle = \frac{1}{|G|} \sum_g \chi_\rho(g) \cdot \overline{1}$$

is exactly the average of those counts. By Lemma 23.8.4 applied to ρ, the same quantity is $\dim(\mathbb{C}^n)^G$.

(b) Take $G = S_n$ and $\varphi = \text{id}$. A vector $\mathbf{x} \in \mathbb{C}^n$ fixed by every $\rho(\tau)$ has $x_i = x_j$ for all $i \neq j$, by taking τ the transposition $(i\ j)$; conversely every vector with equal coordinates is fixed. So $(\mathbb{C}^n)^{S_n} = \text{span}(\mathbf{e}_1 + \cdots + \mathbf{e}_n)$, of dimension 1 (for $n = 1$ this is $\mathbb{C}$ itself). By (a) the average number of fixed points equals that dimension, namely 1.

(c) For S_3, $\chi_\rho = (3, 1, 0)$ on classes of sizes $1, 3, 2$, so the average number of fixed points is $\frac{1}{6}(1 \cdot 3 + 3 \cdot 1 + 2 \cdot 0) = 1$. Directly: id fixes three points, each of the three transpositions fixes one, and each of the two 3-cycles fixes none, giving $(3 + 3 + 0)/6 = 1$. And $\dim(\mathbb{C}^3)^{S_3} = 1$ by (b). The three agree.

Exercise. C2: Degree-one characters form a group

Let G be a finite group and let $X(G)$ be the set of characters of degree 1 of complex representations of G. (For abelian G this group is the subject of Section 09.)

(a) Prove that the pointwise product of two degree-one characters is a degree-one character, and that $X(G)$ is a group under this product.

(b) Prove that $|X(G)| \leq k$, the number of conjugacy classes, with equality if and only if G is abelian. *Hint: use Corollary 23.8.10 in one direction and Corollary 23.4.4 in the other.*

Solution.

(a) A degree-one character is a homomorphism $\chi\colon G \to \mathbb{C} \setminus \{0\}$ (the representation *is*

the scalar it multiplies by). If χ, ψ are two of them then $(\chi\psi)(gh) = \chi(gh)\psi(gh) = \chi(g)\psi(g)\chi(h)\psi(h) = (\chi\psi)(g)(\chi\psi)(h)$, using commutativity of $\mathbb{C}$; so $\chi\psi$ is again a homomorphism into $\mathbb{C}\setminus\{0\}$, hence a degree-one character. The product is associative because multiplication in $\mathbb{C}$ is; χ_{triv} is an identity; and $g \mapsto \chi(g)^{-1}$ is a homomorphism and satisfies $\chi \cdot \chi^{-1} = \chi_{\text{triv}}$. So $X(G)$ is a group (Definition 0.10.1).

(b) Degree-one characters are irreducible (a one-dimensional representation has no non-trivial invariant subspace) and pairwise distinct elements of $X(G)$ are inequivalent representations, since for degree one, equivalence means equality. So $X(G)$ is a subset of the k irreducible characters (Theorem 23.8.9) and $|X(G)| \le k$.

Equality says every irreducible character has degree 1. If G is abelian then $\mathbb{C}[G]$ is a commutative algebra, so every irreducible $\mathbb{C}[G]$-space is one-dimensional by Corollary 23.4.4; by Theorem 23.7.1 those are exactly the irreducible representations of G. Conversely, if every irreducible has degree 1 then $|G| = \sum_{i=1}^{k} 1 = k$ by Corollary 23.8.10, so every conjugacy class is a singleton; that is, $xgx^{-1} = g$ for all x, g, which says G is abelian.

Exercise. C3: The second orthogonality relation

Let G be a finite group with irreducible complex characters $\chi_1, \dots, \chi_k$.

(a) Prove that $\sum_{i=1}^{k} |\chi_i(g)|^2 = |G|/|C(g)|$ for every $g \in G$.

(b) Deduce that $\sum_i |\chi_i(g)|^2 = 1$ if and only if the class of g is all of G, and that this happens only for the trivial group. *Hint:* $C(1) = \{1\}$ *always.*

(c) For $G = S_3$ compute $\sum_i |\chi_i(g)|^2$ at each of the three classes and check (a).

Solution.

(a) This is the diagonal case $j = l$ of Corollary 23.8.11, with $g = g_j$ a representative of its own class and $c_j = |C(g)|$. The value is independent of the representative chosen, as it must be, since characters are class functions.

(b) By (a), the sum equals 1 exactly when $|C(g)| = |G|$, that is, when the class of g is all of G. But $C(1) = \{1\}$, and the classes partition G, so a class equal to G forces $G = \{1\}$. Conversely for the trivial group the only class is $\{1\} = G$ and the single character has $|\chi_1(1)|^2 = 1$.

(c) From Example (The character table of the symmetric group on three letters): at id, $1 + 1 + 4 = 6 = 6/1$; at $(1\ 2)$, $1 + 1 + 0 = 2 = 6/3$; at $(1\ 2\ 3)$, $1 + 1 + 1 = 3 = 6/2$. Each matches $|G|/|C(g)|$ with class sizes $1, 3, 2$.

23.9 Finite Abelian Groups and the Fourier Transform

Everything in this chapter has been building a machine, and for one family of groups the machine produces an object the book already owns. When G is abelian and the field is $\mathbb{C}$, Section 4 forces every irreducible representation down to a single dimension, Section 8 makes the resulting characters orthonormal, and what is left is one unitary matrix. For the cyclic group that matrix is the Fourier matrix of Chapter 12 Section 9, entry for entry. This section proves that, and two

companion statements: the convolution theorem is the group algebra splitting into a product of copies of $\mathbb{C}$, and a circulant is an element of that group algebra multiplying.

Throughout, G is a **finite abelian** group, with identity written e, and the field is $\mathbb{C}$ unless a statement says otherwise. When $G = \mathbb{Z}/n\mathbb{Z}$ we follow Chapter 12 Section 9 and run every index from 0 to $n-1$, reading subscripts modulo n. We write δ_g for the basis vector of the group algebra $\mathbb{C}[G]$ attached to $g \in G$ (Definition 23.1.5), so that $\delta_g \delta_h = \delta_{gh}$ and δ_e is the identity of the algebra; the Kronecker delta of Chapter 11 always carries two numerical subscripts, as in δ_{jk}, and no group element is ever written as a pair of numbers.

23.9.1 One dimension is all there is

Section 4 proved that over an algebraically closed field the irreducible representations of a commutative algebra are lines, and Section 7 turned representations of G into representations of $\mathbb{C}[G]$. Since G is abelian, $\mathbb{C}[G]$ is commutative, so the two results together leave no room: an irreducible complex representation of G is one-dimensional, and an operator on a line is a number. A character is then not merely a function on G; it is a homomorphism out of G.

Before that we recall the one fact about finite groups that Chapter 0 did not record: every g in a finite group has a least $m \geq 1$ with $g^m = e$, its **order**, and $\langle g \rangle = \{e, g, \dots, g^{m-1}\}$ is then a subgroup with exactly m elements. Section 8 proved it, as Lemma 23.8.1, writing the identity 1 where this section writes e; we use it constantly below.

A group of the form $\langle g \rangle$ gets a name, since it is the only shape of group we can handle in full.

Definition 23.9.1. Cyclic group

A group G is **cyclic** if $G = \langle g \rangle$ for some $g \in G$; such a g is called a **generator** of G. A cyclic group is abelian, since $g^k g^l = g^{k+l} = g^l g^k$.

One construction of groups from groups is needed before the examples, and Chapter 0 did not record it either.

Definition 23.9.2. Direct product of groups

Let G and H be groups. Their **direct product** $G \times H$ is the set of ordered pairs (g, h) with the componentwise operation $(g, h)(g', h') := (gg', hh')$. It is a group, with identity (e_G, e_H) and $(g, h)^{-1} = (g^{-1}, h^{-1})$; the three axioms are checked one component at a time. It is abelian when G and H are, and $|G \times H| = |G|\,|H|$.

By Lemma 23.8.1, a finite cyclic group $G = \langle g \rangle$ has $|G|$ equal to the order of g. The standard example is $\mathbb{Z}/n\mathbb{Z}$ under addition, generated by $[1]$, with $[1]$ of order n. The group $\mathbb{Z}$ is cyclic and infinite. The Klein four-group $\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$ is abelian but **not** cyclic: every element satisfies $x + x = 0$, so no element has order 4.

Now the theorem that starts the section.

Theorem 23.9.1. Characters of an Abelian Group Are Homomorphisms

Let G be a finite abelian group and let the field be $\mathbb{C}$. Then:

(a) every irreducible representation of G is one-dimensional;

(b) a one-dimensional representation of G is the same thing as a group homomorphism $\chi\colon G \to \mathbb{C}^\times$, and the character of that representation is χ itself;

(c) every one-dimensional representation is irreducible, and two of them are equivalent if and only if they are equal;

(d) for every such χ and every $g \in G$, the number $\chi(g)$ is an m-th root of unity, where m is the order of g (Lemma 23.8.1); in particular $|\chi(g)| = 1$ and $\chi(g)^{-1} = \overline{\chi(g)} = \chi(g^{-1})$.

Idea. Only (a) is deep, and Section 4 did it: in a commutative algebra every element commutes with the whole algebra, so Schur's lemma (Theorem 23.4.1) makes every element act as a scalar, and then every line is invariant, so an irreducible representation can only be a line. The rest is bookkeeping about 1×1 matrices, together with the observation that a finite group makes every value a root of unity.

Proof.

(a) A representation of G over $\mathbb{C}$ is a representation of the algebra $\mathbb{C}[G]$ (Theorem 23.7.1), and $\mathbb{C}[G]$ is commutative because G is abelian: the products $\delta_g\delta_h = \delta_{gh}$ and $\delta_h\delta_g = \delta_{hg}$ agree on basis vectors, and the general case follows by bilinearity. Since $\mathbb{C}$ is algebraically closed, Corollary 23.4.4 applies and every irreducible representation is one-dimensional.

(b) Let $\rho\colon G \to \mathrm{GL}(V)$ with $\dim V = 1$. For each g the operator $\rho(g)$ maps the line V to itself, so $\rho(g) = \chi(g)\,\mathrm{id}_V$ for a unique scalar $\chi(g)$, and $\chi(g) \neq 0$ because $\rho(g)$ is invertible. From $\rho(gh) = \rho(g)\rho(h)$ we get $\chi(gh) = \chi(g)\chi(h)$, so $\chi\colon G \to \mathbb{C}^\times$ is a homomorphism (Definition 0.10.3). In any basis of V the matrix of $\rho(g)$ is the 1×1 matrix with entry $\chi(g)$, so its trace, which is the character (Definition 23.8.2), is $\chi(g)$. Conversely a homomorphism $\chi\colon G \to \mathbb{C}^\times$ defines a representation on $V = \mathbb{C}$ by $\rho(g)z = \chi(g)z$.

(c) A one-dimensional V has no subspaces but $\{0\}$ and V, and $V \neq \{0\}$, so there is no proper nonzero subrepresentation and ρ is irreducible (Definition 23.7.4). Suppose ρ, ρ' are one-dimensional with scalars χ, χ' and $T\colon V \to V'$ is an isomorphism with $T\rho(g) = \rho'(g)T$ for all g. Applying both sides to a nonzero $v \in V$ gives $\chi(g)\,Tv = \chi'(g)\,Tv$, and $Tv \neq 0$, so $\chi = \chi'$. The converse is immediate with $T = \mathrm{id}$.

(d) By Lemma 23.8.1 there is $m \geq 1$ with $g^m = e$. Then

$$\chi(g)^m = \chi(g^m) = \chi(e) = 1$$

by Theorem 0.10.3, so $\chi(g)$ is an m-th root of unity (Example (Roots of Unity)). Taking absolute values, $|\chi(g)|^m = 1$ with $|\chi(g)|$ a positive real, so $|\chi(g)| = 1$ and therefore $\overline{\chi(g)} = \chi(g)^{-1}$. Finally $\chi(g)^{-1} = \chi(g^{-1})$, again by Theorem 0.10.3; this last identity is Proposition 23.8.2 (e) in the one-dimensional case.

■

Part (a) is where the field matters, and it matters twice over. It uses algebraic closure, and Section 4 exhibited what goes wrong without it: the rotation representation of a cyclic group on $\mathbb{R}^2$ is irreducible of dimension 2. Nothing in (a) needs a hypothesis on the characteristic; that hypothesis enters when we want representations to decompose, which is Maschke's theorem (Theorem 23.7.3), and again through Section 8, whose orthogonality relations are proved by

averaging over the group and need it too.

Example. The characters of a cyclic group

Let $G = \mathbb{Z}/n\mathbb{Z}$, written additively, and let $\omega := e^{2\pi i/n}$. Find all homomorphisms $G \to \mathbb{C}^\times$.

Solution. Section 7 did this, in Example (The irreducible complex representations of a cyclic group): such a χ is determined by $z := \chi([1])$, which satisfies $z^n = \chi([n]) = \chi([0]) = 1$ and so equals ω^j for exactly one $j \in \{0, \dots, n-1\}$ (Example (Roots of Unity)); and each j does give a homomorphism. So there are exactly n characters,

$$\chi_j([m]) = \omega^{jm} \qquad (j, m = 0, 1, \dots, n-1),$$

with χ_0 the constant function 1. What Section 7 obtained as a list of irreducible representations is, by Theorem 23.9.1 (b), the same list read as homomorphisms.

Example. The Klein four-group

Let $G = \mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$, with elements written (x, y) for $x, y \in \{0, 1\}$. Find all characters and write down the resulting table.

Solution. Every element satisfies $2(x, y) = (0, 0)$, so $\chi(x, y)^2 = \chi(2(x, y)) = \chi(0, 0) = 1$ by Theorem 23.9.1 (b), and each value $\chi(x, y)$ is ± 1. A homomorphism is determined by $\chi(1, 0)$ and $\chi(0, 1)$, and any choice of signs for those two works: setting $\chi_{a,b}(x, y) := (-1)^{ax+by}$ for $a, b \in \{0, 1\}$ gives a homomorphism, because $(-1)^{a(x+x')+b(y+y')}$ factors and the exponent only matters modulo 2. So there are exactly four characters, and the table of values, with rows the characters and columns the elements in the order $(0, 0), (1, 0), (0, 1), (1, 1)$, is

$$\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix},$$

the rows being $\chi_{0,0}, \chi_{1,0}, \chi_{0,1}, \chi_{1,1}$. Any two distinct rows agree in two places and disagree in two, so their entrywise product sums to 0: the rows are orthogonal, as Theorem 23.8.7 says they must be.

Warning. The word "character" is doing double duty, and only for abelian groups do the two jobs coincide. For a general finite group the character $\chi_\rho = \operatorname{tr} \rho$ of Definition 23.8.2 is **not** a homomorphism: $\chi_\rho(e) = \dim V$, which is 1 only when ρ is one-dimensional. The group S_3 has an irreducible complex representation of dimension 2 (Section 7), whose character takes the value 2 at the identity and is therefore not a homomorphism into $\mathbb{C}^\times$. Reserve "the characters of G" for the homomorphisms $G \to \mathbb{C}^\times$ only when G is abelian, which is what we do from here on.

Quick check. The group $\mathbb{Z}/3\mathbb{Z}$ has exactly one homomorphism into $\mathbb{R}^\times$. Which, and why does this not contradict Theorem 23.9.1?

Solution. If $\chi\colon \mathbb{Z}/3\mathbb{Z} \to \mathbb{R}^\times$ is a homomorphism, then $\chi([1])^3 = 1$ with $\chi([1])$ real, and the only real cube root of 1 is 1. So χ is trivial. There is no contradiction: Theorem 23.9.1 is a statement over $\mathbb{C}$, and its part (a) rests on algebraic closure. Over $\mathbb{R}$ the group $\mathbb{Z}/3\mathbb{Z}$ has an irreducible two-dimensional representation, by rotation through $2\pi/3$, and that representation has no one-dimensional pieces to contribute a character.

23.9.2 The dual group

The characters of G are now a set of functions on G, and that set has structure of its own: the pointwise product of two homomorphisms into an abelian group is again a homomorphism. So the characters form a group, and the rest of the section is about how that group compares with G.

The dual group of G is the group of ways to measure G by complex numbers of modulus one.

Definition 23.9.3. Dual group

Let G be a finite abelian group. The **dual group** of G is

$$\widehat{G} := \{\, \chi\colon G \to \mathbb{C}^\times : \chi \text{ is a group homomorphism} \,\},$$

with the **pointwise** operation $(\chi\psi)(g) := \chi(g)\psi(g)$. Its elements are the **characters** of G.

In words: an element of $\widehat{G}$ is one homomorphism out of G, and two of them are multiplied by multiplying their values at each point separately. Nothing here refers to a representation any more; by Theorem 23.9.1 this is exactly the set of irreducible characters of G, repackaged so that it carries a product.

Proposition 23.9.2. The Dual Group Is a Group

For a finite abelian group G, the set $\widehat{G}$ with the pointwise operation is a finite abelian group. Its identity is the **trivial character** χ_0, with $\chi_0(g) = 1$ for every g, and the inverse of χ is $\overline{\chi}$, the function $g \mapsto \overline{\chi(g)}$.

Proof. First, the operation lands in $\widehat{G}$: if $\chi, \psi \in \widehat{G}$ then

$$(\chi\psi)(gh) = \chi(g)\chi(h)\psi(g)\psi(h) = (\chi\psi)(g)\,(\chi\psi)(h),$$

where the first equality is that χ and ψ are homomorphisms and the second reorders the four factors, using that $\mathbb{C}^\times$ is commutative; so $\chi\psi \in \widehat{G}$. Associativity and commutativity hold pointwise because they hold in $\mathbb{C}^\times$. The constant function $\chi_0 = 1$ is a homomorphism and satisfies $\chi_0\chi = \chi$. For inverses, $\overline{\chi}$ is a homomorphism because conjugation is multiplicative, and $\chi(g)\overline{\chi(g)} = |\chi(g)|^2 = 1$ by Theorem 23.9.1 (d), so $\chi\overline{\chi} = \chi_0$. Finally $\widehat{G}$ is finite: a character is determined by its values on the $|G|$ elements of G, and each value $\chi(g)$ is an

m-th root of unity, where m is the order of g and $m \le |G|$ (Lemma 23.8.1); there are only finitely many such numbers. This proves the proposition.

■

How big is $\widehat{G}$? Section 8 already answered this, and the answer costs one observation about abelian groups.

Proposition 23.9.3. The Dual Group Has the Same Size

For every finite abelian group G, $|\widehat{G}| = |G|$.

Proof. Two elements $g, h \in G$ are conjugate when $h = xgx^{-1}$ for some $x \in G$ (Definition 23.8.1). In an abelian group $xgx^{-1} = gxx^{-1} = g$, so each conjugacy class is a single element and G has exactly $|G|$ conjugacy classes. By Theorem 23.8.9 the number of irreducible complex characters of G equals the number of conjugacy classes, hence it is $|G|$. By Theorem 23.9.1 (b) and (c), the irreducible characters are precisely the elements of $\widehat{G}$, each arising from one equivalence class of irreducible representations. So $|\widehat{G}| = |G|$.

■

That is a statement about cardinality. The stronger statement is that $\widehat{G}$ and G are isomorphic as groups, and here the book runs into a limit it should name rather than hide. We can prove it for cyclic groups, and we can prove that it survives taking products; and that is as far as Chapter 0's group theory reaches.

Theorem 23.9.4. The Dual of a Product of Cyclic Groups

Let G and H be finite abelian groups. Then:

(a) if G is cyclic, then $\widehat{G} \cong G$;

(b) $\widehat{G \times H} \cong \widehat{G} \times \widehat{H}$;

(c) consequently, if G is a direct product of finitely many finite cyclic groups, then $\widehat{G} \cong G$.

Idea. For (a), a homomorphism out of $\langle g \rangle$ is decided by one value, and that value may be any n-th root of unity; so the dual is listed by the same n exponents that list G, and matching exponent to exponent is the isomorphism. For (b), a character of a product splits into its two restrictions, and multiplying the restrictions back together recovers it. Part (c) is then induction on the number of factors.

Proof.

(a) Let $G = \langle g \rangle$ with $|G| = n$, so g has order n (Lemma 23.8.1), and put $\omega = e^{2\pi i/n}$. For $j \in \mathbb{Z}$ define $\chi_j(g^k) := \omega^{jk}$. This is well defined: if $g^k = g^l$ then $n \mid k - l$, so $\omega^{jk} = \omega^{jl}$. It is a homomorphism, since $\omega^{j(k+l)} = \omega^{jk}\omega^{jl}$, so $\chi_j \in \widehat{G}$.
Now define $\Theta \colon G \to \widehat{G}$ by $\Theta(g^j) := \chi_j$. It is well defined by the same divisibility:

$n \mid j - j'$ gives $\omega^{jk} = \omega^{j'k}$ for every k, so $\chi_j = \chi_{j'}$. It is a homomorphism, because $\chi_j \chi_{j'}$ sends g^k to $\omega^{jk}\omega^{j'k} = \omega^{(j+j')k}$, which is $\chi_{j+j'}(g^k)$. It is injective: if $\Theta(g^j) = \chi_0$ then $\omega^j = \chi_j(g) = 1$, so $n \mid j$ and $g^j = e$; and a homomorphism with trivial kernel is injective, since $\Theta(x) = \Theta(y)$ gives $\Theta(xy^{-1}) = \chi_0$ by Theorem 0.10.3. It is surjective: given $\chi \in \widehat{G}$, we have $\chi(g)^n = \chi(g^n) = \chi(e) = 1$, so $\chi(g) = \omega^j$ for some j (Example (Roots of Unity)), and then $\chi(g^k) = \omega^{jk} = \chi_j(g^k)$ for every k, so $\chi = \chi_j = \Theta(g^j)$. Hence Θ is an isomorphism.

(b) Define $\Xi\colon \widehat{G} \times \widehat{H} \to \widehat{G \times H}$ by letting $\Xi(\chi, \psi)$ be the function $(g, h) \mapsto \chi(g)\psi(h)$. That function is a homomorphism, since

$$\chi(gg')\psi(hh') = \big(\chi(g)\psi(h)\big)\big(\chi(g')\psi(h')\big),$$

using commutativity of $\mathbb{C}^\times$. The map Ξ is a homomorphism of groups because both sides are computed pointwise. It is injective: if $\chi(g)\psi(h) = 1$ for all g, h, then taking $h = e_H$ gives $\chi(g) = 1$ for all g, whence also $\psi = \chi_0$; so the kernel is trivial and Ξ is injective, for the reason given in (a). It is surjective: given $\theta \in \widehat{G \times H}$, set $\chi(g) \coloneqq \theta(g, e_H)$ and $\psi(h) \coloneqq \theta(e_G, h)$, both homomorphisms; then

$$\theta(g, h) = \theta\big((g, e_H)(e_G, h)\big) = \chi(g)\psi(h),$$

so $\theta = \Xi(\chi, \psi)$. Hence Ξ is an isomorphism.

(c) Induct on the number r of factors. For $r = 1$ this is (a). If $G = G_1 \times \cdots \times G_r$ with each G_i cyclic, write $G = G_1 \times K$ with $K = G_2 \times \cdots \times G_r$. By (b), $\widehat{G} \cong \widehat{G_1} \times \widehat{K}$, which by (a) and the inductive hypothesis is isomorphic to $G_1 \times K = G$.

■

Remark. The chapter can go no further here, and it should say so. It is true that **every** finite abelian group is a direct product of cyclic groups, so that $\widehat{G} \cong G$ for every finite abelian G. That structure theorem is **not proved in this book**, and nothing below depends on it: every later statement of this section is either about a cyclic group or uses only Proposition 23.9.3, which is proved. The linear-algebra shadow of the structure theorem is in Chapter 10 Section 7: Theorem 10.7.3 diagonalizes a matrix over $F[x]$ with a divisibility chain, and the identical argument over $\mathbb{Z}$ turns a matrix of relations between generators of a finite abelian group into a list of cyclic factors. Chapter 10 sets that out as Exercise (C1: The same algorithm over the integers), and an exercise is not a proof; the input it assumes, that a finite abelian group is presentable by an integer relation matrix at all, is also not proved there.

Warning. The isomorphism $\widehat{G} \cong G$ of Theorem 23.9.4 is **not canonical**: the map Θ in the proof depends on choosing a generator g of G and a primitive root of unity ω, and different choices give different isomorphisms. There is no way to remove the choice. What *is* canonical is the isomorphism from G to its **double** dual, $g \mapsto (\chi \mapsto \chi(g))$, which mentions no choices at all; Exercise (C2: The double dual needs no choices) asks you to prove it. This is the same phenomenon as Chapter 5's $V \cong V^*$ against $V \cong V^{**}$, and it is not a coincidence.

Quick check. Why does Proposition 23.9.3 fail for S_3 if "character" is read as "homomorphism into $\mathbb{C}^\times$"?

Solution. Because Proposition 23.9.3 used that every conjugacy class of an abelian group is a single element. In S_3 there are three classes, not six, so there are three irreducible characters; and of those only two are one-dimensional, so the set of homomorphisms $S_3 \to \mathbb{C}^\times$ has two elements, not six. The proof breaks at the first line, not the last.

23.9.3 The character matrix, the transform, and inversion

With $|\widehat{G}| = |G| = n$ we may line the two sets up and form a square matrix of values. Everything the section is heading towards is a statement about that one matrix.

Definition 23.9.4. Character matrix

Let G be a finite abelian group of order n. Fix an ordering $g_0, \dots, g_{n-1}$ of G and an ordering $\chi_0, \dots, \chi_{n-1}$ of $\widehat{G}$, which has n elements too (Proposition 23.9.3). The **character matrix** of G for those orderings is $\mathbf{U} \in M_n(\mathbb{C})$ with

$$(\mathbf{U})_{mj} := \frac{\chi_j(g_m)}{\sqrt{n}} \qquad (0 \le m, j \le n-1).$$

Its j-th column lists the values of χ_j, scaled by $n^{-1/2}$.

Section 8 rescaled the same table the other way up: the matrix in the proof of Corollary 23.8.11 has the characters along its **rows**, and is the transpose of this $\mathbf{U}$ once every class size c_j is 1, which for an abelian G it is.

The scaling is there for one reason, and the next theorem is that reason. Alongside it we name the transform.

Definition 23.9.5. Fourier transform on a finite abelian group

Let G be a finite abelian group of order n and let $f : G \to \mathbb{C}$. The **Fourier transform** of f is the function $\widehat{f} : \widehat{G} \to \mathbb{C}$ given by

$$\widehat{f}(\chi) := \sum_{g \in G} f(g)\,\chi(g).$$

Theorem 23.9.5. Fourier Inversion on a Finite Abelian Group

Let G be a finite abelian group of order n, with orderings fixed as in Definition 23.9.4 and character matrix $\mathbf{U}$. Then:

(a) $\mathbf{U}$ is unitary, and the n functions $n^{-1/2}\chi_j$ form an orthonormal basis of the space $\mathbb{C}^G$ of all functions $G \to \mathbb{C}$;

(b) **(Inversion)** every $f\colon G \to \mathbb{C}$ satisfies

$$f(g) = \frac{1}{n} \sum_{\chi \in \widehat{G}} \widehat{f}(\chi)\, \overline{\chi(g)} \qquad (g \in G);$$

(c) **(Plancherel)** $\displaystyle\sum_{g \in G} |f(g)|^2 = \frac{1}{n} \sum_{\chi \in \widehat{G}} |\widehat{f}(\chi)|^2.$

Idea. Identify a function on G with its column of values, so that $\mathbb{C}^G$ becomes $\mathbb{C}^n$ with the standard inner product. Orthonormality of the characters, which Section 8 proved by averaging and Schur's lemma (Theorem 23.4.1), then says exactly that the columns of $\mathbf{U}$ are an orthonormal basis. Inversion and Plancherel are what a unitary matrix always gives: expand in the new basis, and lengths do not change.

Proof.

(a) Identify $f \in \mathbb{C}^G$ with the column $\mathbf{x} \in \mathbb{C}^n$, $x_m = f(g_m)$. Under this identification the standard inner product on $\mathbb{C}^n$ is $\langle f, f' \rangle = \sum_g f(g)\overline{f'(g)}$. The j-th column of $\mathbf{U}$ is the function $n^{-1/2}\chi_j$, and for any j, k,

$$\langle \tfrac{\chi_j}{\sqrt{n}}, \tfrac{\chi_k}{\sqrt{n}} \rangle = \frac{1}{n} \sum_{g \in G} \chi_j(g)\overline{\chi_k(g)},$$

which is the inner product of class functions of Definition 23.8.4 applied to χ_j and χ_k. By Theorem 23.9.1 (b) and (c) these are irreducible characters, of inequivalent representations when $j \neq k$, so the value is δ_{jk} by Theorem 23.8.7. Hence the columns of $\mathbf{U}$ are orthonormal; there are n of them in $\mathbb{C}^n$, so they form an orthonormal basis (Theorem 11.1.4, Theorem 1.6.4), and $\mathbf{U}$ is unitary (Theorem 11.7.3).

(b) Write $\mathbf{y} \in \mathbb{C}^n$ for the column with $y_j = n^{-1/2}\widehat{f}(\chi_j)$. By Definition 23.9.5,

$$y_j = \frac{1}{\sqrt{n}} \sum_m f(g_m)\chi_j(g_m) = \sum_m (\mathbf{U})_{mj}\, x_m,$$

that is $\mathbf{y} = \mathbf{U}^{\top}\mathbf{x}$. Now $\mathbf{U}^{\top}$ is unitary as well, since $(\mathbf{U}^{\top})^{*}\mathbf{U}^{\top} = \overline{\mathbf{U}}\,\mathbf{U}^{\top} = \overline{\mathbf{U}\,\mathbf{U}^{*}} = \mathbf{I}$, so $\mathbf{x} = (\mathbf{U}^{\top})^{*}\mathbf{y} = \overline{\mathbf{U}}\,\mathbf{y}$. Reading off entry m,

$$f(g_m) = \sum_j \overline{(\mathbf{U})_{mj}}\, y_j = \frac{1}{n} \sum_j \widehat{f}(\chi_j)\, \overline{\chi_j(g_m)},$$

which is the stated formula.

(c) A unitary matrix preserves the norm (Theorem 11.7.3), so $\|\mathbf{y}\|^2 = \|\mathbf{x}\|^2$. Here $\|\mathbf{x}\|^2 = \sum_g |f(g)|^2$ and $\|\mathbf{y}\|^2 = \frac{1}{n}\sum_j |\widehat{f}(\chi_j)|^2$. This proves the theorem. ∎

Two remarks on conventions, both harmless. Some authors put a conjugate into the transform, defining $\widehat{f}(\chi) = \sum_g f(g)\overline{\chi(g)}$; since $\overline{\chi} = \chi^{-1} \in \widehat{G}$ (Proposition 23.9.2), this only relabels $\widehat{G}$ by inversion, and the inversion formula swaps its conjugate accordingly. We use the version without

the conjugate because it is the one Chapter 12 already wrote down. Second, the $n^{-1/2}$ sits in $\mathbf{U}$ and not in $\widehat{f}$, which is again Chapter 12's choice.

23.9.4 The cyclic group is Chapter 12

Now take $G = \mathbb{Z}/n\mathbb{Z}$, ordered by $g_m = [m]$, with the characters $\chi_j([m]) = \omega^{jm}$ of Example (The characters of a cyclic group) in that order, and $\omega = e^{2\pi i/n}$. Three theorems follow, and together they say that the discrete Fourier transform of Chapter 12 was representation theory all along.

Theorem 23.9.6. The Character Matrix Is the Fourier Matrix

For $G = \mathbb{Z}/n\mathbb{Z}$ with the orderings above, the character matrix of Definition 23.9.4 is exactly the Fourier matrix $\mathbf{F}$ of Theorem 12.9.3:

$$(\mathbf{U})_{mj} = \frac{\omega^{jm}}{\sqrt{n}} = (\mathbf{F})_{mj} \qquad (0 \le m, j \le n-1).$$

In particular the unitarity of $\mathbf{F}$ and the orthonormality of the characters of $\mathbb{Z}/n\mathbb{Z}$ are the same statement.

Proof. By Definition 23.9.4 and Example (The characters of a cyclic group), $(\mathbf{U})_{mj} = \chi_j(g_m)/\sqrt{n} = \omega^{jm}/\sqrt{n}$. By Theorem 12.9.3, $(\mathbf{F})_{mj} = \omega^{mj}/\sqrt{n}$. The two agree because $jm = mj$. The final sentence is Theorem 23.9.5 (a) read for this $\mathbf{U}$. ∎

That last sentence deserves to be dwelt on, because the book has now proved the same fact twice by genuinely different routes. Chapter 12 proved $\mathbf{F}^*\mathbf{F} = \mathbf{I}$ by summing a geometric series: for $z^n = 1$ and $z \neq 1$, $1 + z + \cdots + z^{n-1} = 0$. Section 8 proved orthonormality of irreducible characters by averaging an arbitrary linear map over the group and applying Schur's lemma. The first argument is a computation that happens to come out; the second says why it had to. Neither is redundant, and the second does not make the first obsolete, since the first is what you use to compute.

Theorem 23.9.7. The Group Algebra of an Abelian Group Splits

Let G be a finite abelian group of order n, and let $\mathbb{C}^{\widehat{G}}$ denote the algebra of all functions $\widehat{G} \to \mathbb{C}$ with pointwise operations, which is a product of n copies of $\mathbb{C}$. Then

$$\Phi \colon \mathbb{C}[G] \to \mathbb{C}^{\widehat{G}}, \qquad \Phi\Big(\sum_g c_g \delta_g\Big)(\chi) := \sum_g c_g\, \chi(g),$$

is an isomorphism of algebras. For $G = \mathbb{Z}/n\mathbb{Z}$ it is the discrete Fourier transform $\mathbf{c} \mapsto \widehat{\mathbf{c}} = \sqrt{n}\,\mathbf{F}\mathbf{c}$ of Chapter 12 Section 9, and its multiplicativity is Theorem 12.9.6 (b).

Idea. The map is multiplicative for one reason only: $\chi(gh) = \chi(g)\chi(h)$, so evaluating a character on a product of group algebra elements factors. Bijectivity is not a separate argument either; the matrix of Φ in the obvious bases is the character matrix up to transposing and scaling, and that matrix is unitary.

Proof. Write $a = \sum_g c_g \delta_g$ and $b = \sum_h d_h \delta_h$ in $\mathbb{C}[G]$ (Definition 23.1.5), so that $ab = \sum_{g,h} c_g d_h \, \delta_{gh}$. The map Φ is linear in a by construction. For multiplicativity, evaluate at $\chi \in \widehat{G}$:

$$\begin{aligned}\Phi(ab)(\chi) &= \sum_{g,h} c_g d_h \, \chi(gh) = \sum_{g,h} c_g d_h \, \chi(g)\chi(h) \\ &= \Big(\sum_g c_g \chi(g)\Big)\Big(\sum_h d_h \chi(h)\Big) = \Phi(a)(\chi)\,\Phi(b)(\chi),\end{aligned}$$

the second equality because χ is a homomorphism. The identity of $\mathbb{C}[G]$ is δ_e, and $\Phi(\delta_e)(\chi) = \chi(e) = 1$ for every χ, which is the identity of $\mathbb{C}^{\widehat{G}}$. So Φ is a homomorphism of algebras (Definition 23.1.2).

For bijectivity, order G and $\widehat{G}$ as in Definition 23.9.4. In the basis $(\delta_{g_0}, \dots, \delta_{g_{n-1}})$ of $\mathbb{C}[G]$ and the basis of point masses on $\widehat{G}$, the matrix of Φ has (j,m)-entry $\chi_j(g_m) = \sqrt{n}\,(\mathbf{U})_{mj}$; that is, the matrix is $\sqrt{n}\,\mathbf{U}^{\top}$. By Theorem 23.9.5 (a), $\mathbf{U}$ is unitary, hence invertible, hence so is $\sqrt{n}\,\mathbf{U}^{\top}$. A linear bijection that is a homomorphism of algebras is an isomorphism of algebras.

For $G = \mathbb{Z}/n\mathbb{Z}$, write $a = \sum_m c_m \delta_{[m]}$. Then $\Phi(a)(\chi_j) = \sum_m c_m \omega^{jm} = p_{\mathbf{c}}(\omega^j)$, the j-th entry of the vector $\widehat{\mathbf{c}} = \sqrt{n}\,\mathbf{F}\mathbf{c}$ of Chapter 12 Section 9, with $p_{\mathbf{c}}$ the symbol of Definition 12.9.1. Multiplicativity of Φ then reads $p_{\mathbf{c}*\mathbf{d}}(\omega^j) = p_{\mathbf{c}}(\omega^j) p_{\mathbf{d}}(\omega^j)$, which is Theorem 12.9.6 (b).

■

So the convolution theorem is a splitting theorem. Convolution is hard to look at because it is multiplication in $\mathbb{C}[G]$ written in the basis of point masses; in the basis of characters the same multiplication is diagonal, because there the algebra has fallen apart into n independent copies of $\mathbb{C}$. Seen from Section 6 this is no surprise: $\mathbb{C}[G]$ is semisimple by Theorem 23.7.3 applied to the regular representation, and a commutative semisimple algebra over $\mathbb{C}$ is a product of matrix algebras by Theorem 23.6.1, all of whose blocks must be 1×1 because $M_k(\mathbb{C})$ is commutative only for $k = 1$. Then $\dim \mathbb{C}[G] = |G|$ and Corollary 23.6.2 force the number of blocks to be $|G|$. The content of Theorem 23.9.7 over and above that is the explicit isomorphism.

The third identification puts Chapter 12's matrices back into the picture.

Theorem 23.9.8. Circulants Are the Group Algebra of the Cyclic Group

Let $n \geq 1$ and let $\mathcal{C}_n \subseteq M_n(\mathbb{C})$ be the algebra of circulants (Corollary 12.9.5). Let $\Psi : \mathbb{C}[\mathbb{Z}/n\mathbb{Z}] \to M_n(\mathbb{C})$ be the linear map with

$$\Psi(\delta_{[m]}) := \mathbf{S}^m \qquad (m = 0, 1, \dots, n-1),$$

where $\mathbf{S}$ is the cyclic shift. Then:

(a) Ψ is an isomorphism of algebras onto $\mathcal{C}_n$, and $\Psi\big(\sum_m c_m \delta_{[m]}\big)$ is the circulant $\mathbf{C}$ with first row $\mathbf{c}$;

(b) $\Psi(a)$ is the matrix of the map "multiply by a" on $\mathbb{C}[\mathbb{Z}/n\mathbb{Z}]$, in the ordered basis

$$\mathcal{B} = (\delta_{[0]},\ \delta_{[-1]},\ \delta_{[-2]},\ \dots,\ \delta_{[-(n-1)]});$$

(c) $\mathbf{F}^*\Psi(a)\mathbf{F}$ is the diagonal matrix whose j-th entry is $\Phi(a)(\chi_j)$; that is, Φ is Ψ followed by conjugating with $\mathbf{F}$ and reading off the diagonal.

Idea. Part (a) is almost a tautology once the right map is chosen: Ψ sends the basis element $\delta_{[m]}$ to $\mathbf{S}^m$, and the only relation in the group, $[m]+[k]=[m+k]$ read modulo n, matches the only relation among the powers, $\mathbf{S}^n=\mathbf{I}$. Part (b) is an index computation, and the reversal of the basis in it is forced by Chapter 12's convention $(\mathbf{C})_{jk}=c_{k-j}$. Part (c) is Theorem 12.9.4 restated.

Proof.

(a) By Theorem 12.9.1 (a), $\sum_m c_m\mathbf{S}^m = p_{\mathbf{c}}(\mathbf{S})$ is the circulant with first row $\mathbf{c}$, so Ψ has the stated values and its image is $\mathcal{C}_n$, which by Corollary 12.9.5 (a) has $\mathbf{I},\mathbf{S},\dots,\mathbf{S}^{n-1}$ as a basis. Hence Ψ carries a basis to a basis and is a linear bijection onto $\mathcal{C}_n$. For multiplicativity it suffices, by bilinearity of both products, to check it on basis vectors:

$$\Psi\big(\delta_{[m]}\delta_{[k]}\big)=\Psi\big(\delta_{[m+k]}\big)=\mathbf{S}^{(m+k)\bmod n}=\mathbf{S}^m\mathbf{S}^k,$$

the last equality because $\mathbf{S}^n=\mathbf{I}$. Finally $\Psi(\delta_{[0]})=\mathbf{I}$, so Ψ is an isomorphism of algebras onto $\mathcal{C}_n$.

(b) Write $\mathcal{B}=(u_0,\dots,u_{n-1})$ with $u_k=\delta_{[-k]}$; this is a reordering of the basis of point masses, hence a basis. For $a=\sum_m c_m\delta_{[m]}$,

$$a\,u_k=\sum_m c_m\,\delta_{[m-k]}=\sum_m c_m\,u_{k-m}=\sum_j c_{k-j}\,u_j,$$

substituting $j=k-m$ and reading all subscripts modulo n. So the (j,k)-entry of the matrix of multiplication by a in $\mathcal{B}$ is c_{k-j}, which is $(\mathbf{C})_{jk}$ by Definition 12.9.1. Hence that matrix is $\mathbf{C}=\Psi(a)$.

(c) By (a) and Theorem 12.9.4, $\mathbf{F}^*\Psi(a)\mathbf{F}=\operatorname{diag}(p_{\mathbf{c}}(1),\dots,p_{\mathbf{c}}(\omega^{n-1}))$, and $p_{\mathbf{c}}(\omega^j)=\Phi(a)(\chi_j)$ was computed at the end of the proof of Theorem 23.9.7.

■

Warning. Part (b) reverses the basis, and the reversal is not cosmetic. With Chapter 12's convention $(\mathbf{C})_{jk}=c_{k-j}$, applying a circulant to a vector is **not** convolution: $(\mathbf{C}\mathbf{x})_j=\sum_t c_t x_{j+t}$, with a plus sign, whereas $(\mathbf{c}*\mathbf{x})_j=\sum_t c_t x_{j-t}$. What is true is that $\mathbf{C}\mathbf{x}=\mathbf{c}^-*\mathbf{x}$, where $\mathbf{c}^-$ is the vector with $(\mathbf{c}^-)_t=c_{-t}$, and that the **product** of two circulants is the circulant of the convolution of their first rows (Theorem 12.9.6 (a)). The group algebra sees the product, which is why Theorem 23.9.8 (a) needs no reversal while (b) does. If you keep only one of these, keep the product statement.

Example. The case $n=4$, checked against Chapter 12

For $G=\mathbb{Z}/4\mathbb{Z}$, write down the character table, the character matrix, and the image under Φ of the element $a=\delta_{[0]}+2\delta_{[1]}+3\delta_{[2]}+4\delta_{[3]}$. Compare with Chapter 12 Section 9.

Solution. Here $\omega = e^{2\pi i/4} = i$ and $\chi_j([m]) = i^{jm}$. The table of values, rows $\chi_0, \chi_1, \chi_2, \chi_3$ and columns $[0], [1], [2], [3]$, is

$$\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & i & -1 & -i \\ 1 & -1 & 1 & -1 \\ 1 & -i & -1 & i \end{pmatrix}.$$

The array is symmetric, since $i^{jm} = i^{mj}$, so reading it with rows indexed by m and columns by j and dividing by $\sqrt{4} = 2$ gives the character matrix

$$\mathbf{U} = \frac{1}{2}\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & i & -1 & -i \\ 1 & -1 & 1 & -1 \\ 1 & -i & -1 & i \end{pmatrix}.$$

This is exactly the matrix $\mathbf{F}$ of Example (The Fourier matrix of size four).

By Theorem 23.9.8 (a), $\Psi(a)$ is the circulant with first row $(1, 2, 3, 4)$, the running example of Chapter 12 Section 9. By Theorem 23.9.7,

$$\begin{aligned} \Phi(a)(\chi_0) &= 1 + 2 + 3 + 4 = 10, \\ \Phi(a)(\chi_1) &= 1 + 2i - 3 - 4i = -2 - 2i, \\ \Phi(a)(\chi_2) &= 1 - 2 + 3 - 4 = -2, \\ \Phi(a)(\chi_3) &= 1 - 2i - 3 + 4i = -2 + 2i, \end{aligned}$$

using $i^2 = -1$ and $i^3 = -i$. These are the four eigenvalues found in Example (The running example, diagonalized), in the same order, and Theorem 23.9.8 (c) says that $\mathbf{F}^*\Psi(a)\mathbf{F} = \operatorname{diag}(10, -2 - 2i, -2, -2 + 2i)$, which is what that example computed. As a check on the first value, χ_0 is the trivial character, so $\Phi(a)(\chi_0)$ is the sum of the coefficients, which is the common row sum 10 of the circulant.

What has been bought? Chapter 12 said that a circulant is a polynomial in the shift, and proved it with an induction on the entries of $\mathbf{S}^m$. That statement is now structural: $\mathbb{C}[\mathbb{Z}/n\mathbb{Z}]$ is generated as an algebra by one element $\delta_{[1]}$, whose image is $\mathbf{S}$, so *every* element of the algebra is a polynomial in $\mathbf{S}$. What the structural reading does not hand over free is that those polynomials are exactly the circulants; that is still Chapter 12's entry computation (Theorem 12.9.1 (a)), quoted in the proof of Theorem 23.9.8 (a). The one matrix $\mathbf{F}$ that diagonalizes every circulant simultaneously is the character matrix, which exists before any circulant is named because the characters of G exist before any element of $\mathbb{C}[G]$ is named. And the reason the eigenvalues came out as n evaluations of one polynomial is that they are the n coordinates of a in the product decomposition $\mathbb{C}[G] \cong \mathbb{C}^{\widehat{G}}$.

23.9.5 Exercises

23.9.5.1 A. Check your understanding

Exercise. A1

(a) Define the dual group $\widehat{G}$ of a finite abelian group G, and say what its identity and inverses are.

(b) State why $|\widehat{G}| = |G|$, naming the two results used.

(c) Write down the character matrix of G and state what its unitarity says about the characters.

(d) Determine whether the following are true, with a reason. (i) Every irreducible complex representation of a finite abelian group is one-dimensional. (ii) The book has proved $\widehat{G} \cong G$ for every finite abelian G. (iii) The Fourier matrix of Chapter 12 Section 9 is the character matrix of $\mathbb{Z}/n\mathbb{Z}$. (iv) For $\mathbf{A} \in \mathcal{C}_n$ and $\mathbf{x} \in \mathbb{C}^n$, $\mathbf{Ax}$ is the convolution of the first row of $\mathbf{A}$ with $\mathbf{x}$.

(e) Explain in one sentence why the values of a character of a finite abelian group are roots of unity.

Solution.

(a) $\widehat{G}$ is the set of group homomorphisms $G \to \mathbb{C}^\times$ with the pointwise product (Definition 23.9.3). Its identity is the trivial character $\chi_0 \equiv 1$, and the inverse of χ is $\overline{\chi}$ (Proposition 23.9.2).

(b) Because G is abelian, each conjugacy class is a single element, so there are $|G|$ classes; Theorem 23.8.9 makes the number of irreducible characters equal to that, and Theorem 23.9.1 identifies the irreducible characters with the elements of $\widehat{G}$ (Proposition 23.9.3).

(c) $(\mathbf{U})_{mj} = \chi_j(g_m)/\sqrt{n}$ for fixed orderings of G and $\widehat{G}$ (Definition 23.9.4). Unitarity says the columns are orthonormal, that is, $\frac{1}{n}\sum_g \chi_j(g)\overline{\chi_k(g)} = \delta_{jk}$: the characters are an orthonormal basis of the functions on G (Theorem 23.9.5).

(d) (i) True over $\mathbb{C}$, by Theorem 23.9.1 (a); it is false over $\mathbb{R}$. (ii) False. It is proved here only for direct products of cyclic groups (Theorem 23.9.4), and the book does not prove that every finite abelian group is such a product. (iii) True, entry for entry (Theorem 23.9.6). (iv) False: $(\mathbf{Ax})_j = \sum_t c_t x_{j+t}$, while convolution has x_{j-t}. The product of two circulants, however, is the circulant of the convolution.

(e) Each g has finite order m (Lemma 23.8.1), and applying χ to $g^m = e$ gives $\chi(g)^m = 1$.

23.9.5.2 B. Practice

Exercise. B1: The character table of the group of order three

Let $G = \mathbb{Z}/3\mathbb{Z}$ and $\omega = e^{2\pi i/3}$.

(a) Write down the three characters and the character matrix.

(b) Verify one off-diagonal orthogonality relation by hand.

(c) Let $a = 2\delta_{[0]} - \delta_{[1]} - \delta_{[2]} \in \mathbb{C}[G]$. Compute $\Phi(a)$, write down the circulant $\Psi(a)$, and state its eigenvalues.

Solution.

(a) By Example (The characters of a cyclic group), $\chi_j([m]) = \omega^{jm}$ for $j = 0, 1, 2$, and

$$\mathbf{U} = \frac{1}{\sqrt{3}} \begin{pmatrix} 1 & 1 & 1 \\ 1 & \omega & \omega^2 \\ 1 & \omega^2 & \omega^4 \end{pmatrix} = \frac{1}{\sqrt{3}} \begin{pmatrix} 1 & 1 & 1 \\ 1 & \omega & \omega^2 \\ 1 & \omega^2 & \omega \end{pmatrix},$$

using $\omega^4 = \omega^3\omega = \omega$.

(b) Take $j = 1$, $k = 0$. Then

$$\tfrac{1}{3} \sum_{m=0}^{2} \chi_1([m])\overline{\chi_0([m])} = \tfrac{1}{3}(1 + \omega + \omega^2) = 0,$$

since $\omega \neq 1$ and $(\omega - 1)(1 + \omega + \omega^2) = \omega^3 - 1 = 0$.

(c) Here $\mathbf{c} = (2, -1, -1)$, so $\Phi(a)(\chi_j) = 2 - \omega^j - \omega^{2j}$. For $j = 0$ this is $2 - 1 - 1 = 0$. For $j = 1$ and $j = 2$ the two subtracted terms are ω and ω^2 in one order or the other, and $\omega + \omega^2 = -1$ by (b), so the value is $2 + 1 = 3$ both times. By Theorem 23.9.8 (a),

$$\Psi(a) = \begin{pmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{pmatrix},$$

with eigenvalues $0, 3, 3$. The eigenvalue 0 is visible directly: every row sums to 0, so the all-ones vector is in the kernel.

Exercise. B2: A product of two cyclic groups

Let $G = \mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/3\mathbb{Z}$ and $\omega = e^{2\pi i/3}$.

(a) Use Theorem 23.9.4 (b) to list the six characters of G explicitly.

(b) Show that G is cyclic by exhibiting a generator, and say which character of G corresponds to which character of $\mathbb{Z}/6\mathbb{Z}$ under your isomorphism.

Solution.

(a) By Theorem 23.9.4 (b) every character of G is $(x, y) \mapsto \chi(x)\psi(y)$ with $\chi \in \widehat{\mathbb{Z}/2\mathbb{Z}}$ and $\psi \in \widehat{\mathbb{Z}/3\mathbb{Z}}$. By Example (The characters of a cyclic group) these are $\chi_a(x) = (-1)^{ax}$ for $a \in \{0, 1\}$ and $\psi_b(y) = \omega^{by}$ for $b \in \{0, 1, 2\}$. So the six characters are

$$\theta_{a,b}(x, y) = (-1)^{ax}\omega^{by} \qquad (a \in \{0, 1\},\ b \in \{0, 1, 2\}),$$

and there are $2 \cdot 3 = 6 = |G|$ of them, as Proposition 23.9.3 requires.

(b) Let $g = ([1], [1])$. Its multiples are $([1], [1]), ([0], [2]), ([1], [0]), ([0], [1]), ([1], [2]), ([0], [0])$ for $k = 1, \ldots, 6$, all six distinct, so g has order 6 and $G = \langle g \rangle$ is cyclic. Send $[k] \in \mathbb{Z}/6\mathbb{Z}$ to kg; this is the isomorphism. Under it the character χ_j of $\mathbb{Z}/6\mathbb{Z}$ with $\chi_j([k]) = \zeta^{jk}$, $\zeta = e^{2\pi i/6}$, pulls back to $\theta_{a,b}$ with $\theta_{a,b}(g) = (-1)^a\omega^b = \zeta^{3a+2b}$, so $j \equiv 3a + 2b \pmod 6$. As (a, b) runs over the six pairs, $3a + 2b$ runs over $0, 2, 4, 3, 5, 1$ modulo 6, hitting each residue once.

Exercise. B3: A transform and its inversion

Let $G = \mathbb{Z}/4\mathbb{Z}$ and let f be the function with values $f([0]), f([1]), f([2]), f([3]) = 1, 1, -1, -1$.

(a) Compute $\widehat{f}(\chi_j)$ for $j = 0, 1, 2, 3$.

(b) Recover $f([1])$ from Theorem 23.9.5 (b).

(c) Check Plancherel's identity for f.

Solution.

(a) With $\chi_j([m]) = i^{jm}$ and Definition 23.9.5,

$$\begin{aligned}\widehat{f}(\chi_0) &= 1 + 1 - 1 - 1 = 0,\\ \widehat{f}(\chi_1) &= 1 + i - (-1) - (-i) = 2 + 2i,\\ \widehat{f}(\chi_2) &= 1 - 1 - 1 + 1 = 0,\\ \widehat{f}(\chi_3) &= 1 - i + 1 - i = 2 - 2i.\end{aligned}$$

For $j = 3$ the powers are $i^0, i^3, i^6, i^9 = 1, -i, -1, i$, so the sum is $1 - i + 1 - i$.

(b) Only $j = 1, 3$ contribute, and $\overline{\chi_1([1])} = \overline{i} = -i$, $\overline{\chi_3([1])} = \overline{-i} = i$. So

$$\tfrac{1}{4}\big((2 + 2i)(-i) + (2 - 2i)(i)\big) = \tfrac{1}{4}\big((2 - 2i) + (2 + 2i)\big) = 1,$$

using $(2 + 2i)(-i) = -2i - 2i^2 = 2 - 2i$ and $(2 - 2i)(i) = 2i - 2i^2 = 2 + 2i$. This is $f([1])$, as it should be.

(c) The left side is $1 + 1 + 1 + 1 = 4$. The right side is $\tfrac{1}{4}(0 + |2 + 2i|^2 + 0 + |2 - 2i|^2) = \tfrac{1}{4}(8 + 8) = 4$. They agree.

23.9.5.3 C. Going deeper

Exercise. C1: The idempotents that do the splitting

Let G be a finite abelian group of order n. For $\chi \in \widehat{G}$ put

$$\varepsilon_\chi := \frac{1}{n} \sum_{g \in G} \overline{\chi(g)}\, \delta_g \in \mathbb{C}[G].$$

(a) Prove that $\Phi(\varepsilon_\chi)(\psi)$ equals 1 when $\psi = \chi$ and 0 otherwise.

(b) Deduce that $\varepsilon_\chi \varepsilon_\psi = 0$ for $\chi \neq \psi$, that $\varepsilon_\chi^2 = \varepsilon_\chi$, and that $\sum_\chi \varepsilon_\chi = \delta_e$.

(c) For $G = \mathbb{Z}/4\mathbb{Z}$ write ε_{χ_1} out, and identify the matrix $\Psi(\varepsilon_{\chi_1})$ in terms of the Fourier vector $\mathbf{f}_1$ of Chapter 12 Section 9.

Hint for (a): the sum is an inner product of two characters.

Solution.

(a) By Definition 23.9.5 and the definition of Φ,

$$\Phi(\varepsilon_\chi)(\psi) = \frac{1}{n}\sum_g \overline{\chi(g)}\,\psi(g) = \frac{1}{n}\sum_g \psi(g)\overline{\chi(g)},$$

which is the inner product of the irreducible characters ψ and χ (Definition 23.8.4). By Theorem 23.8.7 it is 1 if $\psi = \chi$ and 0 otherwise.

(b) Φ is an isomorphism of algebras (Theorem 23.9.7), and in $\mathbb{C}^{\widehat{G}}$ the functions $\Phi(\varepsilon_\chi)$ are the point masses at the various χ. Point masses satisfy exactly the three relations claimed: the pointwise product of two distinct ones is 0, each is its own square, and their sum is the constant function 1, which is $\Phi(\delta_e)$. Applying Φ^{-1} transfers the relations to $\mathbb{C}[G]$.

(c) With $\chi_1([m]) = i^m$, $\overline{\chi_1([m])} = (-i)^m$, so

$$\varepsilon_{\chi_1} = \tfrac{1}{4}(\delta_{[0]} - i\,\delta_{[1]} - \delta_{[2]} + i\,\delta_{[3]}).$$

By Theorem 23.9.8 (a), $\Psi(\varepsilon_{\chi_1})$ is the circulant with first row $\frac{1}{4}(1,-i,-1,i)$, whose (j,k)-entry is $\frac{1}{4}(-i)^{k-j} = \frac{1}{4}i^{j-k}$. With $\mathbf{f}_1 = \frac{1}{2}(1,i,-1,-i)$, the matrix $\mathbf{f}_1\mathbf{f}_1^*$ has (j,k)-entry $\frac{1}{4}i^j\overline{i^k} = \frac{1}{4}i^{j-k}$. This ε_{χ_1} is the idempotent called e_2 in Example (The group algebra of a cyclic group of order four), whose list runs from 1 and is written in the generator $g = [1]$. So $\Psi(\varepsilon_{\chi_1}) = \mathbf{f}_1\mathbf{f}_1^*$, the orthogonal projection onto the line spanned by $\mathbf{f}_1$. The idempotents of the group algebra are the spectral projections of Chapter 12.

Exercise. C2: The double dual needs no choices

Let G be a finite abelian group. For $g \in G$ define $\mathrm{ev}_g\colon \widehat{G} \to \mathbb{C}^\times$ by $\mathrm{ev}_g(\chi) := \chi(g)$.

(a) Prove that $\mathrm{ev}_g \in \widehat{\widehat{G}}$ and that $g \mapsto \mathrm{ev}_g$ is a group homomorphism $G \to \widehat{\widehat{G}}$.

(b) Prove that it is injective. *Hint: the rows of the character matrix are indexed by the elements of G.*

(c) Deduce that it is an isomorphism, and explain why this is not a second proof of Theorem 23.9.4.

Solution.

(a) For $\chi, \psi \in \widehat{G}$, $\mathrm{ev}_g(\chi\psi) = (\chi\psi)(g) = \chi(g)\psi(g) = \mathrm{ev}_g(\chi)\,\mathrm{ev}_g(\psi)$, so ev_g is a homomorphism $\widehat{G} \to \mathbb{C}^\times$, that is an element of $\widehat{\widehat{G}}$. And $\mathrm{ev}_{gh}(\chi) = \chi(gh) = \chi(g)\chi(h) = (\mathrm{ev}_g\,\mathrm{ev}_h)(\chi)$, so $g \mapsto \mathrm{ev}_g$ is a homomorphism.

(b) Suppose ev_g is the trivial character of $\widehat{G}$, that is $\chi(g) = 1 = \chi(e)$ for every $\chi \in \widehat{G}$. Then the rows of the character matrix indexed by g and by e are equal, so $\mathbf{U}$ has two equal rows and is singular, contradicting Theorem 23.9.5 (a) unless $g = e$. Hence the kernel is trivial and the map is injective (Theorem 0.10.3 gives that the kernel is a subgroup, and a homomorphism with trivial kernel is injective by cancellation).

(c) By Proposition 23.9.3 applied twice, $|\widehat{\widehat{G}}| = |\widehat{G}| = |G|$, so an injective homomorphism between them is a bijection, hence an isomorphism. It is not a proof that $\widehat{G} \cong G$: it compares G with the dual of its dual, and gives no map between G and $\widehat{G}$ at all. The difference is exactly the one Chapter 5 drew between $V \cong V^*$, which needs a choice of

basis, and $V \cong V^{**}$, which does not.

Exercise. C3: The same theory over a finite field

Two properties of $\mathbb{C}$ were in use throughout this section: it contains n distinct n-th roots of unity, and n is invertible in it. They are not everything $\mathbb{C}$ contributed: algebraic closure was spent in Theorem 23.9.1 (a), and Section 8's orthogonality relations were proved for complex representations only. But for one group and one field the conclusions can be checked directly with nothing more than those two properties. Take $F = \mathbb{F}_5$, $G = \mathbb{Z}/4\mathbb{Z}$, and $\omega = 2 \in \mathbb{F}_5$.

(a) Verify that ω has order 4 in $\mathbb{F}_5^\times$, and that 4 is invertible in $\mathbb{F}_5$.

(b) Write down the 4×4 matrix $\mathbf{M}$ with $(\mathbf{M})_{mj} = \omega^{jm}$ over $\mathbb{F}_5$ and check that $4^{-1}\sum_m \omega^{jm}\omega^{-km} = \delta_{jk}$ for $j = 1$ and $k = 0, 1$.

(c) Let $\mathbf{C}$ be the circulant over $\mathbb{F}_5$ with first row $(1, 2, 3, 4)$. Compute $\mathbf{M}^{-1}\mathbf{C}\mathbf{M}$ without inverting $\mathbf{M}$, by evaluating the symbol at the powers of ω.

Hint for (c): the proof of Theorem 12.9.4 used only $\mathbf{S}^n = \mathbf{I}$ and the geometric-series identity, both of which hold over $\mathbb{F}_5$.

Solution.

(a) $2^1 = 2$, $2^2 = 4$, $2^3 = 3$, $2^4 = 1$ in $\mathbb{F}_5$, so $\omega = 2$ has order 4 and $1, 2, 4, 3$ are four distinct fourth roots of 1. Also $4 \cdot 4 = 16 = 1$, so 4 is invertible with $4^{-1} = 4$. Note $\operatorname{char}\mathbb{F}_5 = 5$ does not divide $|G| = 4$, which is Maschke's hypothesis.

(b) Reading rows $m = 0, 1, 2, 3$ and columns $j = 0, 1, 2, 3$,

$$\mathbf{M} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 4 & 3 \\ 1 & 4 & 1 & 4 \\ 1 & 3 & 4 & 2 \end{pmatrix}.$$

For $j = k = 1$ every term is $\omega^m\omega^{-m} = 1$, so the sum is $4^{-1} \cdot 4 = 1$. For $j = 1$, $k = 0$ it is $4^{-1}(1 + 2 + 4 + 3) = 4 \cdot 10 = 4 \cdot 0 = 0$ in $\mathbb{F}_5$.

(c) The symbol is $p_{\mathbf{c}}(x) = 1 + 2x + 3x^2 + 4x^3$. Evaluating over $\mathbb{F}_5$:

$$\begin{aligned} p_{\mathbf{c}}(1) &= 1 + 2 + 3 + 4 = 10 = 0, \\ p_{\mathbf{c}}(2) &= 1 + 4 + 12 + 32 = 49 = 4, \\ p_{\mathbf{c}}(4) &= 1 + 8 + 48 + 256 = 313 = 3, \\ p_{\mathbf{c}}(3) &= 1 + 6 + 27 + 108 = 142 = 2. \end{aligned}$$

The columns of $\mathbf{M}$ are the eigenvectors, so $\mathbf{M}^{-1}\mathbf{C}\mathbf{M} = \operatorname{diag}(0, 4, 3, 2)$. The matrix $\mathbf{C}$ is singular over $\mathbb{F}_5$, as it must be: its row sums are $10 = 0$. Every conclusion checked here holds over $\mathbb{F}_5$ for this G, except that "unitary" has no meaning and $\mathbf{M}$ is merely invertible; the orthogonality relations survive with $\bar{\cdot}$ replaced by inversion in $\mathbb{F}_5^\times$.

23.10 What the Chapter Buys

Nine sections is a lot of new vocabulary, and vocabulary is not what you should carry away. This closing section introduces nothing. It names the four moves the chapter actually used, reads four earlier theorems of the book again from the height we have reached, sets out in one place which hypothesis on the field each result needed and what breaks without it, and says plainly what was quoted rather than proved and what was left outside the door on purpose.

23.10.1 Four moves

Average over the group to build an invariant object. This is the whole of Theorem 23.7.3. Given a projection onto a subrepresentation that is not invariant, replace it by the average of its conjugates $\rho(g)P\rho(g)^{-1}$ over $g \in G$; the average is invariant because translating the summation index permutes the summands, and it is still a projection onto the same subspace. The hypothesis $\operatorname{char} F \nmid |G|$ is not decoration: the average is divided by $|G|$, and if that number is 0 in F there is nothing to divide by. The same move produced the intertwiner in Theorem 23.8.7, where an arbitrary linear map is averaged into an equivariant one.

Once you have seen it, you will recognize it elsewhere: whenever a construction depends on a choice and a finite group acts on the choices, averaging over the group removes the dependence.

Use Schur's lemma to force a scalar. Theorem 23.4.1 has two halves, and the useful one in practice is the second: over an algebraically closed field, an operator commuting with an irreducible family is a scalar (Corollary 23.4.3). The pattern of use is always the same. You have an operator built by some construction, you check it commutes with everything in sight, and you conclude it is λ id; then you take a trace to find λ. That single step is what turns the averaged intertwiner of the previous paragraph into the orthogonality relations, and it is what collapses the representations of an abelian group to lines (Corollary 23.4.4).

Count with $\sum n_i^2$. Wedderburn's theorem gives $\dim A = \sum_i n_i^2$ (Corollary 23.6.2), which for a group algebra reads $|G| = \sum_i (\dim V_i)^2$ (Corollary 23.8.10). This is a very strong constraint for a small group: once you know the number of conjugacy classes, the sum of squares often has only one solution, and you have the dimensions of all the irreducible representations before you have written down a single matrix. It plays the role Rank–Nullity plays in Chapter 2: an identity that pins down what you were otherwise going to have to compute.

Read a commuting family as a commutative algebra. Chapters 9 and 12 studied families $\mathcal{F} \subseteq \mathcal{L}(V)$ whose members commute. The algebra generated by $\mathcal{F}$ is then commutative, and the chapter's theorems apply to it directly. This is not a reformulation for its own sake; it is what makes the next four paragraphs possible.

23.10.2 Four old theorems, seen from here

Commuting operators have a common eigenvector (Theorem 9.9.3). Chapter 9 proved this by restricting to an eigenspace and inducting. From here: the algebra A generated by the family is commutative, and V contains an irreducible A-invariant subspace U, namely a nonzero invariant subspace of least dimension. By Corollary 23.4.4, over an algebraically closed field U is a line, and a line invariant under every member of the family is precisely a common eigenvector. Chapter 9's hypothesis — that each characteristic polynomial splits — is what algebraic closure was standing in for. Theorem 9.9.5 then follows by decomposing V into such lines.

One orthonormal eigenbasis for a commuting normal family (Theorem 12.7.1). The same statement with the inner product carried along. The commuting normal family generates

a commutative algebra, and that algebra is closed under adjoints because the adjoint of a normal operator is a polynomial in it (Corollary 12.4.5). Closure under adjoints is exactly what makes the invariant complement available without averaging: the orthogonal complement of an invariant subspace is invariant (Proposition 23.4.6). So the family acts semisimply for free, and over $\mathbb{C}$ — where Corollary 23.4.4 applies — the decomposition into lines is orthogonal.

Circulants and the discrete Fourier transform (Theorem 12.9.4, Corollary 12.9.5, Theorem 12.9.6). Section 9 did this one in full. Chapter 12 found a single unitary $\mathbf{F}$ diagonalizing every circulant at once and could not say where it came from; it came from the character table of $\mathbb{Z}/n\mathbb{Z}$ (Theorem 23.9.6). Chapter 12 found that the product of two circulants is the circulant of the convolution of their first rows; that is multiplication in the group algebra written in the wrong basis, and in the right one it is the coordinatewise product on $\mathbb{C}^{\widehat{G}}$ (Theorem 23.9.7). Chapter 12 found that every circulant is a polynomial in the shift; that is the statement that $\mathbb{C}[\mathbb{Z}/n\mathbb{Z}]$ is generated by one element (Theorem 23.9.8). Three computations, one structure.

The commutant of a Weyr matrix (Proposition 10.4.6), **and disjoint spectra splitting a commutant** (Corollary 12.10.4). These were computations with blocks. The chapter's name for what they compute is the commutant A' of the algebra A generated by the matrix, and Theorem 23.5.5 says what to expect of it when A is semisimple: the commutant of the commutant gives nothing new. Chapter 12's disjoint-spectra result is the case where the invariant subspaces are forced apart by the spectra alone, semisimple or not, so that the answer is block diagonal whatever the blocks are like inside; Chapter 10's Weyr computation is the opposite situation, one eigenvalue and a matrix that need not be semisimple, where the double commutant theorem does not apply and the answer is correspondingly untidy — block triangular rather than block diagonal.

Quick check. Chapter 12 proved that the Fourier matrix is unitary by summing a geometric series. Section 8 proved the same relations by averaging and Schur's lemma. What does the second proof need that the first does not?

Solution. The second needs the field to be $\mathbb{C}$ — algebraically closed, so that Schur's lemma gives scalars, and of characteristic not dividing $|G|$, so that the averaging is legal — and it needs the group to be finite. The first needs only the identity $(z-1)(1+z+\cdots+z^{n-1}) = z^n-1$, and it works over any field containing n distinct n-th roots of unity, as Exercise (C3: The same theory over a finite field) shows over $\mathbb{F}_5$. The second proof explains; the first computes. Neither replaces the other.

23.10.3 Which hypothesis did what

Almost every theorem in this chapter is false over the wrong field, and the table is worth having in one place. In it F is the field of the representation and G is a finite group.

Result	Needs	What fails without it
Theorem 23.2.2	nothing	—
Theorem 23.4.1, first half	nothing	—

Result	Needs	What fails without it
Theorem 23.4.1, second half	F algebraically closed	the rotation representation of a cyclic group on $\mathbb{R}^2$ is irreducible with commutant a copy of $\mathbb{C}$
Corollary 23.4.4	F algebraically closed	the same rotation representation, irreducible of dimension 2
Corollary 23.4.5	F algebraically closed	the same rotation matrix, a commuting family of one member with no real eigenvector
Theorem 23.5.2	F algebraically closed	the same example: a proper subalgebra acting irreducibly
Theorem 23.6.1, as proved here	F algebraically closed, A semisimple	without closure, $\mathbb{C}$ as an $\mathbb{R}$-algebra is semisimple and is not a product of matrix algebras over $\mathbb{R}$; without semisimplicity, the upper triangular matrices are not one either
Theorem 23.5.5	no field hypothesis; A acts semisimply and contains id_V	for the upper triangular 2×2 matrices, A' is the scalars and $A'' = M_2(F) \neq A$
Theorem 23.7.3	$\mathrm{char}\, F \nmid \lvert G \rvert$	$\mathbb{Z}/p\mathbb{Z}$ acting on $\mathbb{F}_p^2$ by unipotent matrices: the invariant line has no invariant complement
Theorem 23.8.7	$F = \mathbb{C}$	the modular case, which the chapter does not treat
Theorem 23.9.1	$F = \mathbb{C}$	over $\mathbb{R}$ the group $\mathbb{Z}/3\mathbb{Z}$ has only the trivial character

Three comments on the table. First, Theorem 23.5.5 is the one big theorem of the chapter that needs nothing of the field at all; what it needs is semisimplicity, and that is a hypothesis about the algebra and the space it acts on. Second, the two "algebraically closed" rows about Schur and Burnside are the same failure seen twice, and the witness is the same rotation matrix both times; that is a sign the hypothesis is doing genuine work rather than covering a gap in a proof. Third, the characteristic hypothesis and the algebraic-closure hypothesis are independent. Over $\mathbb{C}$ only the first can fail, and it cannot, since $\mathrm{char}\, \mathbb{C} = 0$; that is why the complex theory is so much cleaner than the general one, and why the chapter chose $\mathbb{C}$ whenever it could.

Warning. The most tempting misreading of this chapter is "every finite-dimensional algebra over $\mathbb{C}$ is a product of matrix algebras". It is not. Theorem 23.6.1 assumes the algebra is **semisimple**, and the upper triangular 2×2 matrices over $\mathbb{C}$ are a three-dimensional algebra that is not. Acting on $\mathbb{C}^2$, that algebra leaves the line $\mathrm{span}(\mathbf{e}_1)$ invariant with no invariant complement, since $\mathbf{E}_{12}$ sends any $(x, 1)$ to $(1, 0)$; and its radical is the line spanned by $\mathbf{E}_{12}$ (Definition 23.3.5). No product of matrix algebras has a nonzero radical. Before applying

Wedderburn, check semisimplicity — and for a group algebra that check is Theorem 23.7.3, which is a statement about the characteristic and nothing else.

23.10.4 Quoted, not proved

The chapter stated six things with credit and proved none of them. Nothing in the chapter rests on any of them, and each is flagged where it appears.

- **Wedderburn–Artin over a division ring.** A finite-dimensional semisimple algebra over any field is a product of matrix algebras over division rings. We proved only the algebraically closed case, where the division rings are the field itself. The book has not built division rings, and the general theorem needs them.
- **Wedderburn's little theorem**, that a finite division ring is a field. Same reason.
- **The character table does not determine the group.** There are two nonisomorphic groups of order 8 with identical character tables. The book has neither the group theory to construct them nor the need.
- **The structure theorem for finite abelian groups**, that every one is a direct product of cyclic groups. Section 9 proved $\widehat{G} \cong G$ only for such products, and said so. Chapter 10's Smith normal form is the linear-algebra shadow of the missing argument.
- **The converse of Proposition 23.3.6**, that a finite-dimensional algebra with zero radical is semisimple. Section 3 states it and does not use it; wherever semisimplicity is needed later it is established directly.
- **When the commutant of a single matrix is the algebra of polynomials in it**, namely when the minimal and characteristic polynomials agree. Section 1 states the criterion in a remark and computes examples on both sides of it, and proves nothing about it.

23.10.5 Left outside on purpose

The subject continues, and the chapter stopped where the book's tools stopped rather than where the subject does. Four directions, in rough order of how soon you would meet them.

Induced representations and Frobenius reciprocity. Given a subgroup $H \leq G$ and a representation of H, there is a way to build a representation of G, and the characters of the two are related by an adjunction. This is the standard next tool, and it is the standard way to produce the irreducible representations of a group rather than merely count them. It needs cosets of a subgroup, which Chapter 0 does not have.

Modular representation theory. What happens when $\operatorname{char} F$ divides $|G|$ and Theorem 23.7.3 fails. Representations no longer decompose into irreducibles, the group algebra is not semisimple, and a genuinely different theory takes over. Section 7's unipotent example is the whole of what we said about it.

Character tables as a computational subject. We computed a few small ones: the symmetric group on three letters, the cyclic group $\mathbb{Z}/n\mathbb{Z}$, and the Klein four-group. Computing them in general is a craft with its own techniques, and it belongs to a course on group theory.

Representations of infinite groups. Every theorem here used finiteness twice: to average, and to count. For compact groups an integral replaces the sum and most of the chapter survives; for other infinite groups almost none of it does.

23.10.6 One more look back

The book began with one linear map and asked for a basis making it simple. Chapter 9 found eigenvectors, Chapter 10 found the Jordan and rational forms, Chapters 11 to 13 found orthonormal bases when there was an inner product to respect. Then the question widened: Chapter 9 Section 9 and Chapter 12 Section 7 asked for a basis good for a whole commuting family at once. This chapter widened it once more, to a set of operators closed under composition, and found that over an algebraically closed field there is essentially one building block, a full matrix algebra, and that everything semisimple is a product of those.

That is a satisfying place to stop, but notice what it cost. The answer is clean only over an algebraically closed field, only for semisimple algebras, and — on the group side — only when the characteristic is coprime to the order. Every one of those three conditions was a real restriction with a two-by-two counterexample behind it. A structure theorem that applies to everything is usually a structure theorem that says nothing; this one says a great deal, and the price is printed on the label.

23.10.7 Exercises

23.10.7.1 A. Check your understanding

Exercise. A1

(a) Name the four moves of this chapter and, for each, one theorem in which it was used.

(b) State the field hypothesis needed by each of Theorem 23.4.1 (second half), Theorem 23.5.2, Theorem 23.6.1, and Theorem 23.7.3.

(c) Determine whether the following are true, with a reason. (i) Every finite-dimensional complex algebra is a product of matrix algebras. (ii) Every complex representation of a finite group decomposes into irreducibles. (iii) Every representation of a finite group over every field decomposes into irreducibles. (iv) The book proves that two nonisomorphic groups can share a character table.

(d) Why is $|G| = \sum_i (\dim V_i)^2$ useful before any matrix has been written down?

Solution.

(a) Averaging over the group, used in Theorem 23.7.3 and Theorem 23.8.7; forcing a scalar with Schur, used in Corollary 23.4.3 and Corollary 23.4.4; counting with $\sum n_i^2$, used in Corollary 23.6.2 and Corollary 23.8.10; reading a commuting family as a commutative algebra, used throughout Section 9 and in the rereading of Theorem 9.9.5 above.

(b) Schur's second half and Burnside both need F algebraically closed. Theorem 23.6.1, in the form proved here, needs F algebraically closed and the algebra semisimple. Theorem 23.7.3 needs $\operatorname{char} F \nmid |G|$ and nothing else.

(c) (i) False: the upper triangular 2×2 matrices are a counterexample, since they are not semisimple. (ii) True, by Corollary 23.7.4, since $\operatorname{char} \mathbb{C} = 0$ never divides $|G|$. (iii) False: $\mathbb{Z}/p\mathbb{Z}$ on $\mathbb{F}_p^2$ by unipotent matrices. (iv) False. The statement is quoted with credit and not proved.

(d) Because the number of summands is the number of conjugacy classes (Theorem 23.8.9), and for a small group the equation $|G| = \sum n_i^2$ with that many positive integer summands, one of which is 1 for the trivial representation, usually has a unique solution.

The dimensions come out of arithmetic rather than out of linear algebra.

23.10.7.2 B. Practice

Exercise. B1: Two group algebras of order six

(a) Write down the Wedderburn decomposition of $\mathbb{C}[\mathbb{Z}/6\mathbb{Z}]$, with a reason for each factor.

(b) The group S_3 has three conjugacy classes. Use Corollary 23.8.10 to find the dimensions of its irreducible complex representations, and write down the Wedderburn decomposition of $\mathbb{C}[S_3]$.

(c) Check both answers against $\dim \mathbb{C}[G] = |G|$.

Solution.

(a) $\mathbb{Z}/6\mathbb{Z}$ is abelian, so all irreducible representations are one-dimensional (Theorem 23.9.1) and there are six of them (Proposition 23.9.3). Hence

$$\mathbb{C}[\mathbb{Z}/6\mathbb{Z}] \cong \mathbb{C} \times \mathbb{C} \times \mathbb{C} \times \mathbb{C} \times \mathbb{C} \times \mathbb{C},$$

which is Theorem 23.9.7 for this group.

(b) Three conjugacy classes means three irreducible representations (Theorem 23.8.9), of dimensions $n_1, n_2, n_3 \geq 1$ with $n_1^2 + n_2^2 + n_3^2 = 6$ by Corollary 23.8.10. The only solution in positive integers, up to order, is $1 + 1 + 4$, so the dimensions are $1, 1, 2$. Therefore

$$\mathbb{C}[S_3] \cong \mathbb{C} \times \mathbb{C} \times M_2(\mathbb{C}).$$

(c) In (a), $6 \cdot 1 = 6$. In (b), $1 + 1 + 4 = 6$. Both match $\dim \mathbb{C}[G] = |G|$, which is the dimension count of Corollary 23.6.2.

Exercise. B2: What the sum of squares allows at order eight

Let G be a group of order 8.

(a) List all the possible multisets of dimensions of the irreducible complex representations of G. *Remember that the trivial representation is always present.*

(b) Say which of your lists corresponds to an abelian G, and why.

(c) Deduce how many conjugacy classes a nonabelian group of order 8 has.

Solution.

(a) By Corollary 23.8.10 the dimensions satisfy $\sum n_i^2 = 8$ with every $n_i \geq 1$. Let k be the number of n_i equal to 2 or more. Since $3^2 = 9 > 8$, every n_i is 1 or 2, so $8 = 4k + (\text{number of ones})$. For $k = 0$ we get eight 1s; for $k = 1$, four 1s and one 2; for $k = 2$ we would get no 1s at all, which is impossible because the trivial representation has dimension 1. So the two possibilities are $(1, 1, 1, 1, 1, 1, 1, 1)$ and $(1, 1, 1, 1, 2)$.

(b) The first. By Theorem 23.9.1 an abelian G has all irreducible representations one-dimensional, and conversely if all dimensions are 1 then by Theorem 23.6.1 the algebra $\mathbb{C}[G]$ is a product of copies of $\mathbb{C}$, hence commutative; since the elements δ_g commute

in $\mathbb{C}[G]$, the group is abelian.

(c) A nonabelian G of order 8 has dimension list $(1,1,1,1,2)$, so five irreducible representations, and by Theorem 23.8.9 five conjugacy classes.

23.10.7.3 C. Going deeper

Exercise. C1: The circulants as a semisimple algebra

Let $n \geq 2$ and let $\mathcal{C}_n \subseteq M_n(\mathbb{C})$ be the algebra of circulants of Chapter 12 Section 9.

(a) Explain why $\mathcal{C}_n$ acts semisimply on $\mathbb{C}^n$, and write down its Wedderburn form as an abstract algebra.

(b) Prove that an $\mathbf{X} \in M_n(\mathbb{C})$ commutes with every circulant if and only if $\mathbf{F}^*\mathbf{X}\mathbf{F}$ is diagonal, and deduce $\mathcal{C}_n' = \mathcal{C}_n$.

(c) Verify Theorem 23.5.5 for $A = \mathcal{C}_n$.

Hint for (b): first show that a matrix commuting with every diagonal matrix is diagonal, by testing against the matrix units $\mathbf{E}_{kk}$.

Solution.

(a) First, $\mathcal{C}_n$ acts semisimply on $\mathbb{C}^n$. By Theorem 12.9.1 (b) the adjoint of a circulant is a circulant, so if U is $\mathcal{C}_n$-invariant and $\mathbf{x} \in U^{\perp}$, then for every $\mathbf{C} \in \mathcal{C}_n$ and $\mathbf{u} \in U$ we have $\langle \mathbf{C}\mathbf{x}, \mathbf{u} \rangle = \langle \mathbf{x}, \mathbf{C}^*\mathbf{u} \rangle = 0$, because $\mathbf{C}^*\mathbf{u} \in U$; so $U^{\perp}$ is an invariant complement. Second, as an abstract algebra $\mathcal{C}_n \cong \mathbb{C}[\mathbb{Z}/n\mathbb{Z}]$ by Theorem 23.9.8, and by Theorem 23.9.7 its Wedderburn form is the product of n copies of $\mathbb{C}$; all the blocks of Theorem 23.6.1 are 1×1 because the algebra is commutative.

(b) First the hint. Suppose $\mathbf{Y} \in M_n(\mathbb{C})$ commutes with every diagonal matrix. Taking $\mathbf{D} = \mathbf{E}_{kk}$, the matrix $\mathbf{Y}\mathbf{E}_{kk}$ has k-th column equal to the k-th column of $\mathbf{Y}$ and zeros elsewhere, while $\mathbf{E}_{kk}\mathbf{Y}$ has k-th row equal to the k-th row of $\mathbf{Y}$ and zeros elsewhere. Equating them forces $y_{jk} = 0$ for $j \neq k$ and $y_{kj} = 0$ for $j \neq k$. Letting k run, $\mathbf{Y}$ is diagonal.
Now let $\mathbf{X}$ commute with every circulant and put $\mathbf{Y} = \mathbf{F}^*\mathbf{X}\mathbf{F}$. By Corollary 12.9.5 (c) the circulants are exactly the matrices $\mathbf{F}\mathbf{D}\mathbf{F}^*$ with $\mathbf{D}$ diagonal, and

$$\mathbf{X}(\mathbf{F}\mathbf{D}\mathbf{F}^*) = (\mathbf{F}\mathbf{D}\mathbf{F}^*)\mathbf{X} \iff \mathbf{Y}\mathbf{D} = \mathbf{D}\mathbf{Y},$$

after multiplying by $\mathbf{F}^*$ on the left and $\mathbf{F}$ on the right and using $\mathbf{F}^*\mathbf{F} = \mathbf{I}$ (Theorem 12.9.3). So $\mathbf{X}$ commutes with every circulant exactly when $\mathbf{Y}$ commutes with every diagonal matrix, that is exactly when $\mathbf{Y}$ is diagonal, that is exactly when $\mathbf{X} \in \mathcal{C}_n$. Hence $\mathcal{C}_n' = \mathcal{C}_n$.

(c) By (b), $\mathcal{C}_n'' = (\mathcal{C}_n')' = \mathcal{C}_n' = \mathcal{C}_n$. And $\mathcal{C}_n$ acts semisimply on $\mathbb{C}^n$ by (a) and contains $\mathbf{I}$, so Theorem 23.5.5 applies and predicts exactly this. The example also shows the theorem has content: for the algebra of scalar matrices, which is also semisimple and contains $\mathbf{I}$, the commutant is all of $M_n(\mathbb{C})$ and the double commutant is the scalars again — the same conclusion reached along a very different route.

Chapter 24

Computation and Applications

Chapter 21 closed Part V by naming a debt: "every proof here is an existence or a comparison, and not one of them is an algorithm. That is the subject of Chapter 24." It is the book's largest single debt — twenty-six promises across eighteen files — and it has been accumulating since Chapter 3, where partial pivoting was recommended for a reason the book could not yet give. Twenty chapters have built objects and proved things about them. Not one has said how to produce those objects from data of finite accuracy, in finitely many arithmetic operations, and not one has been able to say what an answer computed that way even means.

This chapter answers three questions in that order. **What does a computed answer mean?** It is the exact answer to a nearby question, and the whole of numerical linear algebra lies in making "nearby" precise: a model of arithmetic, backward error, and the rule that ties the two to Chapter 16's condition number, with Chapter 20's residual bounds supplying the error bars. **How are the objects of this book actually computed?** Elimination with pivoting, least squares by reflections, eigenvalues by iteration — because no finite eigenvalue algorithm can exist — Krylov subspaces, conjugate gradients, the fast Fourier transform, and fast multiplication. **What is linear algebra for?** Vibrations, principal components, and graphs, each an application in which a theorem already proved does the work.

The chapter is about arithmetic, algorithms, and cost. No product, library, or language is named anywhere in it: an algorithm here is a rule for producing numbers from numbers, and its cost is a count of arithmetic operations.

24.0.1 Conventions

Seven conventions hold throughout, and each is stated once where it is introduced.

- u **is the unit roundoff** and $\mathrm{fl}(\cdot)$ the rounding map of §01's model, with $\mathcal{R}$ the set of representable numbers and $\gamma_k = ku/(1-ku)$ the accumulated-rounding constant, defined in §01 and used in every section that mentions arithmetic.
- **A flop is one arithmetic operation**, and costs are leading-order counts, written $\frac{2}{3}n^3 + O(n^2)$. The constant matters and the exponent matters more; §§02 and 13 are where that sentence is made quantitative.
- **Perturbations of data are $\Delta\mathbf{A}$ and $\Delta\mathbf{b}$**, and a computed object carries a hat: $\widehat{\mathbf{x}}$, $\widehat{\mathrm{L}}$. Chapters 16, 17, and 20 wrote a perturbation $\mathbf{E}$; $\mathbf{E}$ is kept here for a perturbation whose size is given, and Δ for one produced by an algorithm.
- $\kappa(\mathbf{A})$ **always means Chapter 16 §08's condition number** (Definition 16.8.1). Chapter 20 §11's condition number of a simple eigenvalue does not appear in this chapter, so the two never have to be told apart.

- **Iteration indices are superscripts in parentheses**, $\mathbf{x}^{(k)}$, $\mathbf{A}^{(k)}$, leaving subscripts free for coordinates.
- **$|\mathbf{A}|$ is the entrywise absolute value** ($|a_{ij}|$), as in Chapter 19, and $\mathbf{A} \le \mathbf{B}$ the entrywise order. This is a local override of Chapter 13's $|\mathbf{A}| = (\mathbf{A}^*\mathbf{A})^{1/2}$, which does not appear in this chapter.
- **New symbols**, each defined at its first use: the growth factor g_n (§02), the Krylov space $\mathcal{K}_k(\mathbf{A}, \mathbf{b})$ and the Ritz values (§06), the energy norm $\|\cdot\|_{\mathbf{A}}$ and the Chebyshev polynomials T_k (§07, the same polynomials as Chapter 11 §10), the mass and stiffness matrices $\mathbf{M}$ and $\mathbf{K}$ (§09), the data matrix $\mathbf{X}$ with its centered form $\mathbf{X}_c$ and sample covariance $\mathbf{S}$ (§10), the graph Laplacian and its reduced form (§11, following Chapter 3 §07), the unnormalized Fourier matrix $\mathbf{W}_n = \sqrt{n}\,\mathbf{F}_n$ (§12), and the arithmetic cost $M(n)$ with the exponent ω (§13). Two letters are reused, each with a warning where it happens: Ł is the flowed matrix of §08 and the Laplacian of §11, and $\mathbf{K}$ is §09's stiffness matrix, not Chapter 19 §09's second-difference matrix $\mathbf{K}_n$.

24.0.2 What the chapter assumes

Every earlier chapter assumed only field axioms, and Part V added a short list of facts from analysis. This chapter adds assumptions of a different kind, and both are stated as definitions so that the reader can see exactly where mathematics stops.

Floating-point arithmetic is not a theorem of linear algebra. It is a model of a machine. A machine stores finitely many numbers and rounds every result to one of them, and the statement that it does so with a small relative error is a description of a device, not a fact to be proved. So §01 states it as a **definition**, Definition 24.1.1, in the register this book uses for definitions: a finite set of representable numbers, a rounding map fl, and the requirement

$$\mathrm{fl}(x \odot y) = (x \odot y)(1 + \delta), \qquad |\delta| \le u,$$

for $\odot \in \{+, -, \times, \div\}$ and for the square root, with overflow and underflow excluded by assumption. Everything after §01 is a theorem **about that model**. The model is an idealization: the arithmetic in common use satisfies it for those five operations, with u about 10^{-16}, and every rounding claim in this chapter is a consequence of it and of nothing else.

Section 9 does the same for a piece of physics. Newton's second law and Hooke's law are not theorems either, so Definition 24.9.1 names them as clauses (P1) and (P2), says that Hooke's law holds only for small displacements and that no real system is free of friction, and every statement about vibrations is then a theorem about that idealization. Nothing outside §09 uses it.

No new analysis is imported. The list remains Chapter 16's facts **(A1)–(A6)** — completeness, monotone convergence, compactness of closed bounded sets, the extreme value theorem, the intermediate value theorem, and the mean value theorem — together with the algebra of limits. Each is named again where it is used. **Chapter 21's quoted fact (A7), the Nevanlinna–Pick characterization behind Loewner's theorem, is not used anywhere in this chapter**, and neither is any consequence of it that Chapter 21 marked as resting on it.

24.0.3 What is quoted and not proved

Several classical results in this subject have proofs longer than the sections that need them. The book's practice is to prove everything or to name the result at its point of use and say plainly that it is not proved. Here is the complete list, section by section; each item is flagged again in the section that quotes it.

- **§02** — that Chapter 3's ordering of elimination, which updates the whole trailing block at each step, satisfies the same backward error bound as the ordering §02 displays. The section also declines to quantify the seam between that bound, which is about the factors the model computes, and the growth bound, which is about exact elimination. Separately, and not a quoted theorem at all: partial pivoting is **not** shown to be backward stable, and §02 says so rather than tidying it.
- **§03** — Wilkinson's accumulated backward error for a complete Householder factorization, of which §03 proves the one-reflection case; and the rates at which classical and modified Gram–Schmidt lose orthogonality, $u\kappa_2(\mathbf{A})^2$ and $u\kappa_2(\mathbf{A})$.
- **§04** — that Rayleigh quotient iteration converges at a quadratic rate, not a cubic one, for a general diagonalizable matrix. The cubic Hermitian rate is proved.
- **§05** — the theorem of Abel and Ruffini, quoted from Chapter 11 §08, which is the reason every eigenvalue method in this chapter iterates; and the general convergence theorem of the QR algorithm, of which §05 proves the first column and hence the whole 2×2 case.
- **§06** — the Kaniel–Paige–Saad estimate for the extreme Ritz values, of which §06 proves a clean special case; and the analysis of the loss of orthogonality in Lanczos, which §06 describes and names without proving.
- **§08** — the differentiability of $t \mapsto \mathbf{Q}(t)$, which the continuous form of the Toda statement rests on. The discrete form, that the QR iterates sit on the flow at integer times, is proved.
- **§10** — not a quoted theorem but a missing subject. Chapter 13 §11 promised an argument that a sample says something about a population, and that is statistics, which this book does not develop. Section 10 proves everything deterministic, supplies the finite-population reading, and says plainly where it stops.
- **§13** — the recursive form of Strassen's error bound; Winograd's lower bound of seven multiplications for the 2×2 product; the equality of the two definitions of ω, through all arithmetic operations and through multiplications alone; every bound on ω below $\log_2 7$; and Laderman's twenty-three-product identity for 3×3 matrices, which an exercise hands to the reader.

Nothing in this book is proved from any of them. Every theorem in the chapter either proves its own case or carries what it needs as an explicit hypothesis, and since all of these enter the book here, no earlier chapter can rest on them either. The one place a quoted result is used at all is exercise C1 (b) of §13, where Laderman's identity is given to the reader as a hypothesis and the deduction from it is the reader's own.

24.0.4 What you need

This chapter draws on almost the whole book. Chapters 5, 14, 15, 22, and 23 are the exceptions: duality, bilinear forms, tensor products, affine and projective geometry, and algebras and representations are not used here. Chapter 21 supplies the sentence the chapter answers, but none of its theorems.

- **Chapter 0 §06 and §09** — the complex numbers §07 and §12 work in, and the definition of the matrix product with its three views and the trace, which between them supply every operation count in the chapter: Theorem 0.6.2, Theorem 0.6.4, Example (Roots of Unity), Definition 0.9.9, Example (A product of square matrices), Theorem 0.9.2, Theorem 0.9.6.
- **Chapter 1 §06** — basis extension, which §06 uses to finish an orthonormal list, and "a subspace of the full dimension is the whole space", which §11 uses to identify a column space: Theorem 1.6.5, Theorem 1.6.8.
- **Chapter 3** — the invertibility criteria, rank, elimination and the LU factorization, and the incidence matrix of a network that §11 turns into the Laplacian: Theorem 3.4.4, Theorem 3.5.7, Theorem 3.5.5, Lemma 3.6.3, Lemma 3.6.1, Lemma 3.6.2, Theorem 3.6.4,

Theorem 3.6.5, Example (An invertible matrix with no LU factorization), Definition 3.7.1, Proposition 3.7.4, Corollary 3.7.5, Exercise (C3: Networks in several pieces).

- **Chapter 4 §02** — similarity, the relation every eigenvalue iteration in §§04–08 preserves: Definition 4.2.2.
- **Chapter 6 §01** — that a polynomial over an infinite field is determined by the function it defines, which is how §08 turns an equality of values into an equality of characteristic polynomials: Theorem 6.1.5.
- **Chapter 7** — determinants, almost all of them for §11: the alternating properties, the definition with the triangular and transpose rules, multiplicativity and the invertibility criterion, Laplace expansion and the adjugate identity, minors and rank, and the Gram determinant that §11 turns into the matrix-tree theorem. Section 6 also uses the tridiagonal recurrence and the Vandermonde determinant, and §§05, 07, and 08 the characteristic polynomial: Theorem 7.2.1, Definition 7.4.1, Theorem 7.4.5, Theorem 7.4.6, Theorem 7.5.4, Theorem 7.5.2, Theorem 7.6.1, Theorem 7.6.2, Theorem 7.7.1, Corollary 7.7.3, Exercise (C1: Counting spanning trees), Theorem 7.8.3, Theorem 7.8.1, Theorem 7.10.2, Theorem 7.10.3.
- **Chapter 8** — block multiplication, which is what makes both fast algorithms recurse, and Schur complements, which are what elimination is: Theorem 8.1.1, Definition 8.3.1, Theorem 8.3.1, Theorem 8.3.2.
- **Chapter 9** — diagonalization for §04, the eigenvalues on the diagonal of a triangular form for §05, and Cayley–Hamilton for §08: Theorem 9.4.1, Theorem 9.7.4, Theorem 9.6.1.
- **Chapter 10** — cyclic subspaces, which are Krylov subspaces under another name, and the matrix exponential, which §08 flows along and §09 solves with: Definition 10.5.1, Definition 10.5.4, Definition 10.5.2, Definition 10.5.3, Proposition 10.5.1, Theorem 10.5.2, Theorem 10.5.6, Definition 10.9.2, Theorem 10.9.4, Corollary 10.9.5, Theorem 10.9.6, Example (A Second-Order Equation with a Repeated Root).
- **Chapter 11** — the chapter's largest single source: inner products and Pythagoras, orthonormal bases and Gram–Schmidt, projections and best approximation, least squares, self-adjointness and isometries, reflections and rotations with the QR factorization and Hessenberg form, Parseval's identity, and the orthogonal polynomials behind the three-term recurrence: Theorem 11.1.5, Corollary 11.1.6, Theorem 11.1.3, Theorem 11.1.4, Definition 11.1.3, Theorem 11.2.2, Theorem 11.2.1, Definition 11.3.2, Definition 11.3.3, Theorem 11.3.2, Theorem 11.3.5, Theorem 11.4.2, Corollary 11.4.4, Proposition 11.6.7, Definition 11.7.2, Proposition 11.7.4, Theorem 11.7.3, Definition 11.8.1, Proposition 11.8.2, Theorem 11.8.3, Theorem 11.8.1, Theorem 11.8.4, Definition 11.8.2, Theorem 11.8.5, Theorem 11.9.2, Definition 11.10.1, Proposition 11.10.1, Theorem 11.10.3.
- **Chapter 12** — the spectral theorem in both forms, and the circulants and the discrete Fourier transform that §§11 and 12 rest on: Corollary 12.4.3, Corollary 12.5.2, Definition 12.9.1, Theorem 12.9.4, Corollary 12.9.5, Theorem 12.9.3, Theorem 12.9.6, Example (The Fourier matrix of size four), Exercise (B1: Eigenvalues of small circulants).
- **Chapter 13** — positive definiteness and Cholesky, simultaneous congruence, which is exactly §09's basis of normal modes, the Loewner order, the singular value decomposition with low-rank approximation for §10, and the pseudoinverse: Definition 13.1.1, Theorem 13.1.5, Theorem 13.1.4, Proposition 13.1.2, Theorem 13.2.3, Example (A Cholesky Factorization, and One That Fails), Definition 13.4.1, Theorem 13.4.2, Theorem 13.4.1, Example (A pencil that does not commute), Definition 13.5.1, Definition 13.9.1, Theorem 13.9.1, Definition 13.11.1, Theorem 13.11.3, Corollary 13.11.4, Definition 13.12.1.
- **Chapter 16** — norms and operator norms, the derivative of a matrix-valued function, and above all the condition number and the theorem that bounds a relative error by it: Lemma 16.1.1, Theorem 16.3.3, Theorem 16.3.2, Definition 16.6.1, Lemma 16.6.4, Theorem 16.6.5, Theorem 16.6.6, Definition 16.8.1, Proposition 16.8.2, Theorem 16.8.3,

Theorem 16.8.4.

- **Chapter 17** — the Rayleigh quotient and its critical points, the min–max description of the eigenvalues, Weyl's inequalities, interlacing, and the variational description of the singular values: Definition 17.1.1, Proposition 17.1.1, Proposition 17.1.2, Proposition 17.1.3, Theorem 17.2.3, Corollary 17.2.4, Corollary 17.3.6, Lemma 17.3.5, Theorem 17.4.2, Corollary 17.4.3, Lemma 17.4.5, Theorem 17.9.4, Corollary 17.9.5.
- **Chapter 18** — two small facts, one from each end of the chapter: the coset description of an affine hull, which §10 minimizes over, and the logarithm of a positive definite matrix, which §08 needs to know which matrices its theorem covers: Proposition 18.1.4, Lemma 18.11.1. Both are cited in the text, in §10's opening and §08's warning, and neither is a step in a proof.
- **Chapter 19** — the cyclic permutation that defeats the power method, the Google matrix and the PageRank rate that §04 explains, and the second-difference matrix that §09's uniform chain turns out to be: Example (The Cyclic Permutation), Theorem 19.8.2, Proposition 19.8.3, Proposition 19.9.6. These are examples the chapter re-reads rather than results it argues from: they appear in §04's and §09's examples, and in no proof.
- **Chapter 20** — the perturbation bounds that turn a residual into an error bar, which is what makes every iteration in this chapter stoppable: Theorem 20.4.4, Theorem 20.6.1, Theorem 20.6.2. They are quoted where they are needed — the warnings of §§01 and 06 — and applied in the examples and exercises of §§04, 05, and 06, but no proof here is built on one.

24.0.5 Roadmap

- **Floating point and backward error.** The model, the constant γ_k, the backward error of an inner product, and the three words the chapter runs on: backward stable, forward error, mixed stability. Then the rule, made a theorem: forward error at most condition number times backward error.
- **Elimination, pivoting, and cost.** One step of elimination is a Schur complement of a 1×1 pivot. The cost is $\frac{2}{3}n^3$, and half that for Cholesky. A triangular solve is backward stable, and so is the factorization, up to a growth factor that partial pivoting bounds by 2^{n-1} — a bound that is attained, which is why the section stops short of calling partial pivoting stable.
- **Least squares in practice.** The sensitivity of the least-squares solution involves the residual as well as $\kappa_2(\mathbf{A})$. Forming the normal equations loses twice the digits; the reflection route loses them once, because one reflection is backward stable, and classical Gram–Schmidt loses orthogonality at a rate this section states with credit.
- **The power method.** Repeated multiplication finds the dominant eigenvector at rate $|\lambda_2/\lambda_1|$, the Rayleigh quotient squares that for Hermitian matrices, a shift and an inverse pick out any eigenvalue, and updating the shift from the Rayleigh quotient makes the convergence cubic. It all fails when two eigenvalues share a modulus.
- **The QR algorithm.** Factor and reverse, repeatedly. The iterates are orthogonal iteration applied to the powers of the matrix, which is where the convergence comes from; Hessenberg form brings the cost of a step down to about $6n^2$ flops; and a shift is what turns slow convergence into fast.
- **Krylov subspaces.** The right home for an iteration that can afford only one matrix–vector product at a time. Gram–Schmidt inside that space is the Arnoldi process and gives a Hessenberg matrix, with a three-term recurrence — the Lanczos process — when the matrix is Hermitian; the eigenvalues of the small matrix interlace the large one's, and Chapter 20's bounds apply to them verbatim. In floating point the three-term economy costs orthogonality, and the section says so.

- **Steepest descent and conjugate gradients.** Solving $\mathbf{Ax} = \mathbf{b}$ by minimizing a quadratic. Steepest descent converges like $(\kappa - 1)/(\kappa + 1)$; conjugate gradients minimizes over the whole Krylov space and converges like $(\sqrt{\kappa} - 1)/(\sqrt{\kappa} + 1)$. The square root is the point, and Chebyshev polynomials, small on an interval and large just outside it, are where it comes from.
- **Lax pairs and the Toda flow.** The chapter's one piece of pure mathematics: a differential equation whose flow preserves the spectrum, and which the QR iterates sit on at integer times. That the flow's time-one map *is* a step of the algorithm is derived, but only after granting a differentiability this book has not set up. An algorithm and a differential equation can be the same object seen at two resolutions.
- **Vibrations.** Masses and springs give $\mathbf{M}\ddot{\mathbf{x}} = -\mathbf{K}\mathbf{x}$, the substitution $\mathbf{x} = e^{i\omega t}\mathbf{v}$ gives a generalized eigenproblem, and Chapter 13 gives the $\mathbf{M}$-orthonormal basis of normal modes. Stiffening raises every frequency, adding mass lowers them, and pinning a mass interlaces them.
- **Principal components.** The best-fitting k-dimensional affine subspace of a finite set of points is spanned by the leading singular vectors of the centered data matrix, and "variance explained" is a theorem rather than a convention. The covariance matrix is never formed, for §03's reason, and the inference from a sample to a population is handed back to statistics.
- **Graph Laplacians.** $Ł = \mathbf{N}\mathbf{N}^\top$ for the incidence matrix of Chapter 3 §07. The multiplicity of the eigenvalue 0 counts components, the second smallest eigenvalue measures connectivity and the signs of its eigenvector split the graph, and the determinant of a reduced Laplacian counts spanning trees, which completes Chapter 7 §07.
- **The fast Fourier transform.** The Fourier matrix factors into a permutation, two half-size copies of itself, and a diagonal factor. Solving the resulting recurrence gives $O(n \log n)$, and with it fast convolution, fast polynomial multiplication, and fast circulant solves.
- **Fast matrix multiplication.** Two 2×2 matrices multiply with seven products instead of eight, and the recursion gives $O(n^{\log_2 7})$. The exponent ω is at least 2, which is proved; everything known about it below $\log_2 7$ is stated with credit and not proved. Strassen's algorithm satisfies a norm error bound but no entrywise one, and a small entry really can come out with every digit wrong. The section closes the chapter.

24.0.6 Named moves

- **State the machine as a definition, then prove theorems about it.** When a subject rests on something that is not mathematics, isolate it in one definition, say what kind of statement it is, and derive the rest. Everything in this chapter that mentions rounding is a consequence of Definition 24.1.1 alone, and §09 treats its two physical laws the same way.
- **Measure an answer by a change in the data that makes it exact.** The forward error is what one wants and cannot see; the backward error is what one can bound. Prove the backward statement and let the condition number convert it.
- **Count the operations, and solve the recurrence exactly.** An asymptotic order hides the constant that decides whether an algorithm is worth using. Every cost in this chapter is a closed form or an explicit leading term.
- **A residual is an error bar.** Chapter 20's perturbation bounds turn a quantity the algorithm already has into a certificate, which is what allows an iteration to know when to stop.
- **Iterate in a subspace, not along a direction.** The Krylov space is generated by the one operation an iteration can afford, and it is large enough to contain the answer to a polynomial approximation problem whose solution is the algorithm.

- **A conserved spectrum comes from a commutator.** If $\dot{Ł} = [\mathbf{B}, Ł]$, then $Ł(t)$ is a similarity orbit and its eigenvalues do not move; differentiating a conjugation is the whole proof.
- **A fast algorithm is a factorization.** Factor the linear map (the Fourier matrix) or the bilinear map (the matrix product) into cheap pieces and recursive calls, then count the calls against the size reduction.

24.1 Floating Point and Backward Error

Chapter 16 ended with a number and a warning. The number was $\kappa(\mathbf{A})$, which bounds how much a relative error in the data can be magnified on the way to the answer. The warning was that κ says nothing about arithmetic: it "describes a problem", and an error bound of its shape "only becomes a statement about a computation once one knows how large a perturbation the computation itself introduces". That chapter closed the door with a sentence we now have to walk through: "Conditioning bounds what *any* method can promise; whether a given method keeps that promise is stability, and it is Chapter 24's subject."

This section builds the two things that were missing. First a **model** of machine arithmetic — an assumption, stated plainly as one, because floating-point arithmetic is not a theorem of linear algebra. Then the vocabulary that turns the model into verdicts about algorithms: forward error, backward error, and stability.

24.1.1 Two answers to one question

Chapter 16 §08 left a small experiment on the reader's desk. The system was

$$\begin{aligned} 10^{-15}x_1 + x_2 &= 1, \\ x_1 + x_2 &= 2, \end{aligned}$$

whose coefficient matrix has $\kappa_\infty < 5$: by every measure of Chapter 16 this problem is as well behaved as a problem can be, and its exact solution is $(1, 1)$ to fourteen places. Eliminating x_1 with the first equation as the pivot row, and rounding to twelve significant decimal digits at every step, returns $(0, 1)$. Exchanging the two equations first, and rounding in exactly the same way, returns $(1, 1)$.

Two algorithms, one problem, one arithmetic, and one of the two answers has no correct digit in it. Nothing in Chapter 16 can tell the two apart, because κ never mentions a method. Section 2 works that example in the model below and says exactly what the exchange achieved. This section builds the language it will be said in.

24.1.2 A model of machine arithmetic

A machine holds finitely many numbers. It cannot hold $1/3$, it cannot hold $\sqrt{2}$, and between any two numbers it does hold there are infinitely many it does not. So the result of an arithmetic operation on two stored numbers is usually not a stored number, and the machine must replace it by a neighbor.

We could try to describe which numbers a particular machine stores and exactly how it rounds. That would be a description of a machine, not mathematics, and it would be obsolete before the ink dried. What is wanted instead is the *one property* of rounding that every later argument uses, elevated into an assumption.

Every operation is performed exactly and then moved to the nearest stored number, so each operation multiplies its exact answer by something within u of 1.

Definition 24.1.1. The floating-point model

A **floating-point model** consists of a finite set $\mathcal{R} \subseteq \mathbb{R}$ of **representable numbers** with $0, 1 \in \mathcal{R}$ and $-\mathcal{R} = \mathcal{R}$, a map $\mathrm{fl}\colon \mathbb{R} \to \mathcal{R}$ sending each real number to **a nearest** element of $\mathcal{R}$, and a real number $u > 0$ called the **unit roundoff**, subject to the following.

(FP1) **(Rounding.)** For every $t \in \mathbb{R}$ in range, $\mathrm{fl}(t) = t(1+\delta)$ for some δ with $|\delta| \le u$.

(FP2) **(Arithmetic.)** For $x, y \in \mathcal{R}$ and $\odot \in \{+, -, \times, \div\}$ with $y \neq 0$ when $\odot$ is $\div$, the computed value of $x \odot y$ is $\mathrm{fl}(x \odot y)$; the computed value of $\sqrt{x}$ for $x \in \mathcal{R}$, $x \ge 0$, is $\mathrm{fl}(\sqrt{x}\,)$. Hence

$$\mathrm{fl}(x \odot y) = (x \odot y)(1+\delta), \qquad |\delta| \le u,$$

and likewise $\mathrm{fl}(\sqrt{x}) = \sqrt{x}\,(1+\delta)$ with $|\delta| \le u$.

(FP3) **(No overflow or underflow.)** Every quantity occurring in a computation we analyze is assumed to be **in range**, so that (FP1) applies to it.

Clause by clause. "A nearest element" is a mild requirement, and the only use we make of it is the observation that $|t| \le 1$ forces $|\mathrm{fl}(t)| \le 1$, because $1 \in \mathcal{R}$. (FP1) is the whole content of the word *roundoff*: the error made in storing a real number is **relative**, of size at most u, and that is essentially the only thing we will ever know about fl. (FP2) says an operation is as accurate as storing its answer would be — the machine does not do anything stupid in between. The displayed identity is what every proof in this chapter actually uses, and it is worth reading aloud: *the computed product is the exact product times* $1+\delta$. (FP3) is an honest exclusion. Numbers too large or too small for $\mathcal{R}$ break (FP1) completely, and analyzing them is a separate subject; we assume the problem away and say so each time it matters.

This is a definition, not a theorem. Nothing in Chapters 1 to 21 implies it, and nothing in this chapter proves it. It is an idealization of a machine, in the way that a frictionless plane is an idealization of a table. The arithmetic in common use satisfies it for the four operations and the square root, with u a little over 10^{-16}; a decimal machine carrying t significant digits satisfies it with $u = \frac{1}{2} \cdot 10^{1-t}$. Every statement in this chapter that mentions rounding is a theorem **about the model**, and carries exactly as much authority as the model does.

Example. The model we compute in

Throughout this chapter, a worked number comes from the **six-digit decimal model**: $\mathcal{R}$ is the set of numbers $\pm d_1.d_2d_3d_4d_5d_6 \times 10^e$ with digits d_i, $d_1 \neq 0$ and e in some fixed finite range, together with 0; and fl rounds to the nearest such number. (The range of e is what (FP3) lets us stop thinking about.) What is u, and what is $\mathrm{fl}(99999996)$?

Solution. Two consecutive representable numbers of the same exponent e differ by 10^{e-5}, and a number t with that exponent has $|t| \ge 10^e$. Rounding to the nearest moves t by at most half the gap, so

$$|\mathrm{fl}(t) - t| \le \tfrac{1}{2} \cdot 10^{e-5} \le \tfrac{1}{2} \cdot 10^{-5}|t|,$$

which is (FP1) with $u = \frac{1}{2} \cdot 10^{-5} = 5 \times 10^{-6}$.

For $\mathrm{fl}(99999996)$: the number is 9.9999996×10^7, and the two neighbors with six digits are

9.99999×10^7 and 1.00000×10^8. The second is nearer, so $\mathrm{fl}(99999996) = 10^8$. Note what happened: eight digits went in, six came out, and the two digits that were lost were the only two that distinguished 99999996 from 10^8.

Warning. **A model with an absolute error bound would be a different, and false, model.** Suppose we had written $\mathrm{fl}(x+y) = (x+y)+\delta$ with $|\delta| \le u$. Then $\mathrm{fl}(10^6+1)$ would have to be within $u = 5\times 10^{-6}$ of 10^6+1, whereas a six-digit machine returns 10^6, an absolute error of 1. The error a machine makes is proportional to the size of the number it is storing; a machine is accurate in *ratio*, never in *amount*. Confusing the two is the commonest mistake in the subject, and (FP1) exists to prevent it.

Quick check. In the six-digit model, is $\mathrm{fl}(x+y) = x+y$ possible when x and y are representable and both nonzero? Is $\mathrm{fl}(x-y) = x-y$ possible when $x \neq y$?

Solution. Yes to both, and easily. $1+2=3$ is representable, so the first addition is exact. For the second, $10^8 - 99999.9 = 99900000.1$, which is *not* representable, so that particular subtraction is inexact; but $1.00001 - 1.00000 = 10^{-5}$ is representable exactly, so that subtraction is exact. The model never claims an operation *is* inexact. It claims only a bound on how inexact it can be. That both extremes occur is what makes the subject interesting: the second subtraction is performed exactly, and it is nevertheless the shape of subtraction that destroys answers, as the last example of this section will show.

24.1.3 Accumulated rounding: the constant γ_k

A single operation costs a factor $1+\delta$. An algorithm performs many, and the factors pile up: an answer typically carries a product like $(1+\delta_1)(1+\delta_2)\cdots(1+\delta_k)$, sometimes with an inverted factor where a division occurred. Carrying such products around is unbearable, so we bound them once and for all. The constant that results is used in every remaining section of this chapter.

Lemma 24.1.1. Accumulated rounding

Let $k \ge 1$ and let $\delta_1, \dots, \delta_k$ satisfy $|\delta_i| \le u$, and let $\rho_1, \dots, \rho_k \in \{+1, -1\}$. If $ku < 1$, then

$$\prod_{i=1}^{k} (1+\delta_i)^{\rho_i} = 1+\theta \qquad \text{with} \qquad |\theta| \le \gamma_k := \frac{ku}{1-ku}.$$

Idea. Induction on k, with the two signs treated as two cases. The only thing to watch is that the recursion for γ_k survives division, where the denominator $1-u$ costs something; the definition $\gamma_k = ku/(1-ku)$ is exactly the bound that absorbs that cost.

Proof. Note first that $ku < 1$ forces $ju < 1$ for $j \le k$, so every γ_j with $j \le k$ is defined and positive, and $\gamma_1 \le \gamma_2 \le \cdots \le \gamma_k$.

Base case $k = 1$. If $\rho_1 = +1$, then $\theta = \delta_1$ and $|\theta| \le u \le u/(1-u) = \gamma_1$. If $\rho_1 = -1$, then

$$\theta = \frac{1}{1+\delta_1} - 1 = \frac{-\delta_1}{1+\delta_1}, \qquad |\theta| \le \frac{u}{1-u} = \gamma_1,$$

where $1 + \delta_1 \ge 1 - u > 0$ because $u < 1$.

Inductive step. Suppose the product of the first $k-1$ factors is $1+\theta'$ with $|\theta'| \le \gamma_{k-1}$, and write $\delta = \delta_k$.

If $\rho_k = +1$, then $1 + \theta = (1+\theta')(1+\delta)$, so $\theta = \theta' + \delta + \theta'\delta$ and

$$\begin{aligned} |\theta| &\le \gamma_{k-1} + u + \gamma_{k-1}u \\ &= \frac{(k-1)u + u(1-(k-1)u) + (k-1)u^2}{1-(k-1)u} = \frac{ku}{1-(k-1)u}, \end{aligned}$$

and $1 - (k-1)u \ge 1 - ku > 0$, so this is at most $ku/(1-ku) = \gamma_k$.

If $\rho_k = -1$, then $1+\theta = (1+\theta')/(1+\delta)$, so $\theta = (\theta' - \delta)/(1+\delta)$ and $|\theta| \le (\gamma_{k-1} + u)/(1-u)$. It remains to check that this is at most γ_k. Clearing the positive denominators $1-(k-1)u$, $1-u$, and $1-ku$, the inequality $(\gamma_{k-1}+u)/(1-u) \le \gamma_k$ is equivalent to

$$u(k-(k-1)u)(1-ku) \le ku(1-ku+(k-1)u^2),$$

and expanding both sides gives $k - k^2u - (k-1)u + k(k-1)u^2$ on the left against $k - k^2u + k(k-1)u^2$ on the right, after dividing by $u > 0$. The left side is the right side minus $(k-1)u$, which is at least 0, so the left side is at most the right side and the inequality holds. This proves the lemma.

■

Remark. Read γ_k as "about ku". Indeed if $ku \le 0.01$ then $\gamma_k \le ku/0.99 < 1.02\,ku$. The denominator is there only so that the induction above closes; it is never the point of a bound. We will also use, without further comment, that $\gamma_j \le \gamma_k$ for $j \le k$, and we set $\gamma_0 = 0$, which is what the formula gives and what the empty product deserves.

24.1.4 The first backward error analysis

Here is the simplest algorithm in the book: add up n products. What comes out is not $\mathbf{x}^\top\mathbf{y}$. But it is *exactly* $\mathbf{x}^\top\mathbf{y}$ for a slightly different $\mathbf{x}$ — and that turns out to be the useful thing to say.

Theorem 24.1.2. Backward Error of an Inner Product

Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ have representable entries, let $nu < 1$, and let the inner product be computed in the model of Definition 24.1.1 in the natural order,

$$s_1 = \mathrm{fl}(x_1 y_1), \qquad s_k = \mathrm{fl}\big(s_{k-1} + \mathrm{fl}(x_k y_k)\big) \quad (k = 2, \dots, n).$$

Then there is a vector $\Delta\mathbf{x} \in \mathbb{R}^n$ with

$$s_n = (\mathbf{x} + \Delta\mathbf{x})^{\top}\mathbf{y} \qquad \text{and} \qquad |\Delta x_i| \le \gamma_n |x_i| \quad \text{for every } i.$$

Idea. Unroll the recursion and see which rounding factors touch which term. The term $x_k y_k$ picks up one factor from its own multiplication and one from each addition it takes part in afterwards, so no term collects more than n factors. Lemma 24.1.1 converts "at most n factors" into "$1+\theta$ with $|\theta| \le \gamma_n$", and a factor attached to $x_k y_k$ can be blamed on x_k alone.

Proof. By (FP2) there are numbers δ_k and ε_k, all of modulus at most u, with

$$\begin{aligned} \mathrm{fl}(x_k y_k) &= x_k y_k (1+\delta_k) \quad (1 \le k \le n), \\ s_k &= (s_{k-1} + \mathrm{fl}(x_k y_k))(1+\varepsilon_k) \quad (2 \le k \le n). \end{aligned}$$

An induction on m gives

$$s_m = \sum_{k=1}^{m} x_k y_k (1+\delta_k) \prod_{j=\max(k,2)}^{m} (1+\varepsilon_j),$$

the empty product being 1: for $m = 1$ this is $s_1 = x_1 y_1 (1+\delta_1)$, and the step from $m-1$ to m multiplies every existing term by $1+\varepsilon_m$ and adds the new term $x_m y_m (1+\delta_m)(1+\varepsilon_m)$, which is what the formula says.

Take $m = n$ and count factors in the k-th term. For $k = 1$ there are $1 + (n-1) = n$ of them; for $k \ge 2$ there are $1 + (n-k+1) = n-k+2 \le n$. Every factor has the form $1+\delta$ with $|\delta| \le u$, and $nu < 1$, so Lemma 24.1.1 (with all $\rho_i = +1$) gives numbers θ_k with

$$s_n = \sum_{k=1}^{n} x_k y_k (1+\theta_k), \qquad |\theta_k| \le \gamma_n.$$

Put $\Delta x_k = x_k \theta_k$. Then $|\Delta x_k| \le \gamma_n |x_k|$ and $s_n = \sum_k (x_k + \Delta x_k) y_k = (\mathbf{x} + \Delta\mathbf{x})^{\top}\mathbf{y}$. This proves the theorem. $\blacksquare$

The conclusion is worth staring at. The computed number is not close to $\mathbf{x}^{\top}\mathbf{y}$ — if $\mathbf{x}^{\top}\mathbf{y}$ happens to be tiny compared with $\sum_k |x_k y_k|$, it can be wrong in every digit. What the theorem says is different and stronger in kind: *the computed number is an inner product, exactly, of data within a relative γ_n of the given data.* The algorithm has not solved a nearby problem approximately; it has solved a nearby problem exactly.

Remark. The same proof, with the roles of $\mathbf{x}$ and $\mathbf{y}$ exchanged, gives $s_n = \mathbf{x}^{\top}(\mathbf{y} + \Delta\mathbf{y})$ with $|\Delta\mathbf{y}| \le \gamma_n |\mathbf{y}|$ entrywise. What it does **not** give is a perturbation of both at once with half the constant, and it does not give a bound on $|s_n - \mathbf{x}^{\top}\mathbf{y}|$ relative to $|\mathbf{x}^{\top}\mathbf{y}|$. Exercise (C1: When backward stability says nothing) makes the last point precise. For complex $\mathbf{x}, \mathbf{y}$ the operations $+$, $\times$ are themselves built from several real operations, so the model must be applied to those; the resulting bound has the same shape with a modestly larger constant, and we do not prove it here or use it.

24.1.5 Forward error, backward error, stability

We can now name what Theorem 24.1.2 did, and the alternative it declined to do. Fix a **problem**: a map $f\colon D \to \mathbb{R}^m$ from data to answers, where D is a set of admissible data in some $\mathbb{R}^N$. The set D always consists of **real** data and is never cut down to the representable ones: for the problems of this chapter it is all of $\mathbb{R}^N$, or all invertible matrices, and it has to be, because the perturbed datum $\mathbf{a}+\Delta\mathbf{a}$ below must lie in it. An **algorithm** for f is a map $\widehat{f}$ — defined at the representable data, and sending each of them to whatever the computation actually returns in the model, rounding and all. Two numbers measure how the second differs from the first.

Definition 24.1.2. Forward and backward error

Let f be a problem, $\widehat{f}$ an algorithm for it, and $\mathbf{a} \in D$ a representable datum with $f(\mathbf{a}) \neq \mathbf{0}$. The **forward error** of $\widehat{f}$ at $\mathbf{a}$ is the relative error in the answer,

$$\frac{\|\widehat{f}(\mathbf{a}) - f(\mathbf{a})\|}{\|f(\mathbf{a})\|}.$$

The **backward error** of $\widehat{f}$ at $\mathbf{a}$ is the smallest relative perturbation of the data that accounts for the computed answer exactly,

$$\inf\Big\{ \frac{\|\Delta\mathbf{a}\|}{\|\mathbf{a}\|} \;:\; \mathbf{a}+\Delta\mathbf{a} \in D \text{ and } f(\mathbf{a}+\Delta\mathbf{a}) = \widehat{f}(\mathbf{a}) \Big\},$$

with the convention that it is ∞ when no such $\Delta\mathbf{a}$ exists.

The forward error is what a user wants to know and can almost never compute, since it needs $f(\mathbf{a})$. The backward error is what an analyst can bound, because bounding it only requires *exhibiting* one $\Delta\mathbf{a}$, which a proof like the one above hands over for free. An algorithm whose backward error is always as small as the arithmetic allows is the best one can hope for.

A backward stable algorithm returns the exact answer to a question you might as well have asked.

Definition 24.1.3. Backward stability

Let $f\colon D \to \mathbb{R}^m$ be a problem. An algorithm $\widehat{f}$ for f, running in a floating-point model with unit roundoff u, is **backward stable** if there is a constant c, depending only on the dimensions of the data and **not** on the data itself or on u, such that for every representable $\mathbf{a} \in D$ there exists $\Delta\mathbf{a} \in \mathbb{R}^N$ with

$$\mathbf{a}+\Delta\mathbf{a} \in D, \qquad \widehat{f}(\mathbf{a}) = f(\mathbf{a}+\Delta\mathbf{a}), \qquad \|\Delta\mathbf{a}\| \le c\,u\,\|\mathbf{a}\|.$$

The small words carry the weight. **For every a**: one lucky datum proves nothing. c **independent of the data**: a "constant" allowed to grow with $\|\mathbf{a}\|$ or with κ would make the definition vacuous. c **independent of** u: the bound must improve when the arithmetic does. The constant is allowed to depend on n, and in practice always does — γ_n is about nu, not u — which is why "backward stable" is a statement about a *family* of algorithms indexed by size, and why nobody worries much about a factor of n.

By Theorem 24.1.2, the natural inner product algorithm is backward stable as an algorithm in $\mathbf{x}$ for the problem $\mathbf{x} \mapsto \mathbf{x}^{\top}\mathbf{y}$ with $\mathbf{y}$ fixed, on the domain $D = \mathbb{R}^n$. Take $c = 1.02\,n$, which depends on the dimension alone. Indeed the entrywise bound $|\Delta\mathbf{x}| \le \gamma_n|\mathbf{x}|$ implies $\|\Delta\mathbf{x}\| \le \gamma_n\|\mathbf{x}\|$ in

each of the norms $\|\cdot\|_1$, $\|\cdot\|_2$, $\|\cdot\|_\infty$, and $\gamma_n \le 1.02\,nu$ in every model with $nu \le 0.01$, by the remark after Lemma 24.1.1. (Restricting to such models is how a bound stated with γ_n is turned into one with a constant independent of u, and we do it silently from here on.)

Some algorithms are not backward stable and are not bad either: they perturb the data a little *and* the answer a little, and neither perturbation alone accounts for the result. The weaker notion that covers them is worth naming, so that we never have to claim more than we have proved.

Definition 24.1.4. Mixed forward-backward stability

An algorithm $\widehat{f}$ for $f\colon D \to \mathbb{R}^m$ is **mixed forward-backward stable** if there is a constant c, depending only on the dimensions, such that for every representable $\mathbf{a} \in D$ there are $\Delta\mathbf{a}$ and $\Delta\mathbf{y}$ with $\mathbf{a} + \Delta\mathbf{a} \in D$ and

$$\widehat{f}(\mathbf{a}) = f(\mathbf{a} + \Delta\mathbf{a}) + \Delta\mathbf{y}, \qquad \|\Delta\mathbf{a}\| \le cu\|\mathbf{a}\|, \qquad \|\Delta\mathbf{y}\| \le cu\|f(\mathbf{a} + \Delta\mathbf{a})\|.$$

Backward stability is the special case $\Delta\mathbf{y} = \mathbf{0}$, so it is the stronger statement of the two. A mixed stable algorithm does **not** satisfy the hypothesis of the next theorem: its $\Delta\mathbf{y}$ contributes a further term to the forward error, and we do not estimate that term anywhere in this chapter. Where a method in this chapter is only mixed stable, we say so, and we claim nothing more about it than the definition gives.

Warning. **Backward stable does not mean accurate, and inaccurate does not mean unstable.** A backward stable algorithm commits to nothing about the forward error except what the *problem* allows, and for an ill-conditioned problem that is very little. Conversely, an algorithm can return an accurate answer on one datum by luck while being backward unstable. The two words describe different objects: conditioning belongs to f, stability belongs to $\widehat{f}$. Chapter 16 §08 promised exactly this: "Chapter 24 draws the distinction properly, defines backward stability, and proves which of the algorithms in this book have it." Definition 24.1.3 is the definition; the proofs occupy Sections 2 and 3, and the next theorem is the only bridge between the two sides.

24.1.6 Multiplying the two together

The rule of thumb everyone quotes is *forward error* $\lesssim$ *condition number* $\times$ *backward error*. For linear systems it is not a rule of thumb at all. It is Theorem 16.8.3 with a different reading of the same symbols.

Theorem 24.1.3. Forward Error from Backward Error

Let $\mathbf{A} \in M_n(\mathbb{C})$ be invertible, $\mathbf{b} \neq \mathbf{0}$, and $\mathbf{A}\mathbf{x} = \mathbf{b}$. Let $\hat{\mathbf{x}}$ be any vector which is the exact solution of a perturbed system,

$$(\mathbf{A} + \Delta\mathbf{A})\hat{\mathbf{x}} = \mathbf{b}, \qquad \frac{\|\Delta\mathbf{A}\|}{\|\mathbf{A}\|} \le \eta,$$

and suppose $r := \kappa(\mathbf{A})\,\eta < 1$. Then $\mathbf{A} + \Delta\mathbf{A}$ is invertible and

$$\frac{\|\widehat{\mathbf{x}} - \mathbf{x}\|}{\|\mathbf{x}\|} \;\le\; \frac{r}{1-r} \;=\; \frac{\kappa(\mathbf{A})\eta}{1 - \kappa(\mathbf{A})\eta}.$$

In particular, if $\eta = cu$ is admissible — as it is when the algorithm producing $\widehat{\mathbf{x}}$ is backward stable with constant c **for the problem $\mathbf{A} \mapsto \mathbf{A}^{-1}\mathbf{b}$ with $\mathbf{b}$ held fixed**, so that only $\mathbf{A}$ is perturbed — and if $\kappa(\mathbf{A})cu \le \frac{1}{2}$, then the forward relative error is at most $2\kappa(\mathbf{A})cu$.

Idea. There is nothing to do: backward stability says the computed answer solves a perturbed system exactly, and Chapter 16 already bounded the effect of perturbing $\mathbf{A}$. The content of the theorem is the *translation*, not the estimate.

Proof. Since $\|\Delta\mathbf{A}\|/\|\mathbf{A}\| \le \eta$, we have $\kappa(\mathbf{A})\|\Delta\mathbf{A}\|/\|\mathbf{A}\| \le r < 1$, so Theorem 16.8.3 (b), applied with $\mathbf{E} = \Delta\mathbf{A}$, says that $\mathbf{A} + \Delta\mathbf{A}$ is invertible and that

$$\frac{\|\widehat{\mathbf{x}} - \mathbf{x}\|}{\|\mathbf{x}\|} \le \frac{r'}{1-r'}, \qquad r' := \kappa(\mathbf{A})\frac{\|\Delta\mathbf{A}\|}{\|\mathbf{A}\|} \le r.$$

The function $s \mapsto s/(1-s)$ is increasing on $[0,1)$, since its derivative $(1-s)^{-2}$ is positive there, so $r'/(1-r') \le r/(1-r)$. For the last claim put $\eta = cu$; then $r = \kappa(\mathbf{A})cu \le \frac{1}{2}$ gives $1 - r \ge \frac{1}{2}$ and $r/(1-r) \le 2r$. This proves the theorem.

■

Remark. **When the right-hand side is perturbed too.** A solver for $\mathbf{A}\mathbf{x} = \mathbf{b}$ normally perturbs the pair $(\mathbf{A}, \mathbf{b})$, returning $\widehat{\mathbf{x}} \neq \mathbf{0}$ with

$$(\mathbf{A} + \Delta\mathbf{A})\widehat{\mathbf{x}} = \mathbf{b} + \Delta\mathbf{b}, \qquad \|\Delta\mathbf{A}\|_2 \le cu\|\mathbf{A}\|_2, \qquad \|\Delta\mathbf{b}\|_2 \le cu\|\mathbf{b}\|_2,$$

which is not literally the hypothesis above. One absorption repairs that. Put $\Delta\mathbf{A}' = \Delta\mathbf{A} - \Delta\mathbf{b}\,\widehat{\mathbf{x}}^*/\|\widehat{\mathbf{x}}\|_2^2$; then $(\mathbf{A} + \Delta\mathbf{A}')\widehat{\mathbf{x}} = \mathbf{b} + \Delta\mathbf{b} - \Delta\mathbf{b} = \mathbf{b}$, and the subtracted matrix is an outer product, of 2-norm $\|\Delta\mathbf{b}\|_2/\|\widehat{\mathbf{x}}\|_2$. So Theorem 24.1.3 applies in the 2-norm with

$$\eta = cu\Big(1 + \frac{\|\mathbf{b}\|_2}{\|\mathbf{A}\|_2\|\widehat{\mathbf{x}}\|_2}\Big) \;\le\; cu\Big(1 + \frac{\|\mathbf{x}\|_2}{\|\widehat{\mathbf{x}}\|_2}\Big),$$

the last step because $\|\mathbf{b}\|_2 = \|\mathbf{A}\mathbf{x}\|_2 \le \|\mathbf{A}\|_2\|\mathbf{x}\|_2$. That is a small multiple of cu as soon as the computed $\widehat{\mathbf{x}}$ is not much shorter than $\mathbf{x}$, and it is how every later appeal to the theorem for a linear solver is to be read.

Reading the bound in digits. Suppose the arithmetic has $u \approx 10^{-16}$ and the algorithm is backward stable with a modest c. The bound then reads: relative forward error at most about $\kappa(\mathbf{A}) \cdot 10^{-16}$. Taking logarithms, the computed answer may be wrong from about the $(16 - \log_{10}\kappa(\mathbf{A}))$-th significant digit onward. This is the origin of the sentence "a stable solve loses about $\log_{10}\kappa$ decimal digits". Two cautions, both real. It is an upper bound, so the loss may be smaller — Chapter 16 §08's warning that "the condition number bounds the error; it does not produce it" applies verbatim. And it is a statement about $\kappa(\mathbf{A})$ for the *problem as*

posed: an algorithm that quietly poses a different problem pays that problem's condition number instead.

That last point is the whole of Chapter 16 §08's remaining ledger. Its third unpaid item read: "**Hence the comparison of the two algorithms** — that a stable implementation of the QR route loses about $\log_{10}\kappa_2(\mathbf{A})$ decimal digits while forming and solving the normal equations loses about $2\log_{10}\kappa_2(\mathbf{A})$ — is still Chapter 24's to prove." The theorem above supplies the first half of the sentence for any backward stable method. Section 3 supplies the second half and the two algorithms, using Chapter 16's own Theorem 16.8.4 for the square.

Quick check. An algorithm for $\mathbf{Ax} = \mathbf{b}$ is backward stable, $\kappa(\mathbf{A}) = 10^8$, and $u = 10^{-16}$. A user reports that the computed $\widehat{\mathbf{x}}$ has only eight correct digits and concludes that the algorithm is broken. What is wrong with the conclusion?

Solution. Nothing at all is wrong with the algorithm: Theorem 24.1.3 — applied after absorbing any perturbation of $\mathbf{b}$ into $\Delta\mathbf{A}$, as in the remark following it — predicts a relative forward error of about $\kappa(\mathbf{A})u = 10^{-8}$, up to the modest constant c, which is exactly eight correct digits. The loss is charged to the problem, not to the method. Chapter 16 §08 put it as sharply as it can be put: with data of limited accuracy, an ill-conditioned answer *is not determined* to more digits than that, whatever anybody computes with.

24.1.7 Cancellation, and where the damage really happens

Backward stability is a property of a whole algorithm, and algorithms are usually wrecked in one place. That place is almost always a subtraction of two nearly equal numbers. The following example is the standard one, and it repays being done digit by digit.

Example. Two ways to solve a quadratic

In the six-digit decimal model of Example (The model we compute in), find the two roots of

$$x^2 - 10^4 x + 1 = 0$$

by the quadratic formula $x_\pm = \tfrac{1}{2}(10^4 \pm \sqrt{10^8 - 4})$, and then again with the smaller root computed as $x_- = 1/x_+$. Compare with the exact roots.

Solution. The exact roots. Their sum is 10^4 and their product is 1. Now $9999.9998^2 = 10^8 - 4 + 4\times 10^{-8}$, which is just above $10^8 - 4$, so $\sqrt{10^8 - 4} = 9999.99979999\ldots$ and

$$x_+ = 9999.99989999\ldots, \qquad x_- = \frac{1}{x_+} = 1.0000000100\ldots \times 10^{-4}.$$

The formula, step by step. Each line is one operation of the model.

- $\mathrm{fl}(10^4 \times 10^4) = 10^8$, exactly representable, no error.
- $\mathrm{fl}(10^8 - 4) = \mathrm{fl}(99999996) = 1.00000 \times 10^8$, by Example (The model we compute in).

Here the two informative digits are lost, and the relative error committed is about 4×10^{-8}, comfortably inside $u = 5 \times 10^{-6}$. The model is behaving perfectly.
- $\mathrm{fl}(\sqrt{10^8}) = 10^4$, exactly.
- $\mathrm{fl}(10^4 + 10^4) = 2 \times 10^4$ and $\mathrm{fl}(2 \times 10^4 \div 2) = 10^4$: the computed x_+ is 10^4, with relative error 10^{-8}. Excellent.
- $\mathrm{fl}(10^4 - 10^4) = 0$, exactly, and $\mathrm{fl}(0 \div 2) = 0$: the computed x_- is $\mathbf{0}$.

The repair. Compute $x_+ = 10^4$ as above, then $\mathrm{fl}(1 \div 10^4) = 10^{-4}$, whose relative error against $1.00000001 \times 10^{-4}$ is 10^{-8}.

What to conclude. The second route loses nothing; the first returns a root with no correct digit, indeed with no digits at all. And notice **where** the damage was done. The fatal subtraction $10^4 - 10^4$ was performed **exactly**: both operands were representable and so was their difference. The error was already present, manufactured four lines earlier when 99999996 was rounded to 10^8, and the subtraction merely deleted the digits the two operands had in common, which was all of them. Cancellation does not create error. It *exposes* error, by deleting the digits that were correct and promoting the ones that were not.

Warning. **A small residual is not a small error, and a small error is not a small residual.** For a computed solution $\widehat{\mathbf{x}}$ of $\mathbf{A}\mathbf{x} = \mathbf{b}$ with $\widehat{\mathbf{x}} \neq \mathbf{0}$, put $\mathbf{r} = \mathbf{b} - \mathbf{A}\widehat{\mathbf{x}}$ and

$$\Delta\mathbf{A} = \frac{\mathbf{r}\,\widehat{\mathbf{x}}^*}{\|\widehat{\mathbf{x}}\|_2^2}, \qquad \text{so that} \qquad (\mathbf{A} + \Delta\mathbf{A})\widehat{\mathbf{x}} = \mathbf{A}\widehat{\mathbf{x}} + \mathbf{r} = \mathbf{b},$$

and $\|\Delta\mathbf{A}\|_2 = \|\mathbf{r}\|_2 / \|\widehat{\mathbf{x}}\|_2$: Cauchy–Schwarz gives $\leq$, and $\Delta\mathbf{A}\widehat{\mathbf{x}} = \mathbf{r}$ attains it — the computation is the one in the proof of Theorem 20.4.4, with $\mathbf{b}$ in place of $\mu\widehat{\mathbf{x}}$. So a small relative residual *is* a small backward error, and nothing more: by Theorem 24.1.3 the forward error may still be $\kappa(\mathbf{A})$ times as large. In the other direction, an accurate $\widehat{\mathbf{x}}$ for a badly scaled $\mathbf{A}$ can have a residual far larger than one expects. Chapter 20 §04's Theorem 20.4.4 is the eigenvalue version of exactly this move, and that section named the technique: "An error analysis of this kind, which measures an answer by a change in the data that makes it exact, is a **backward error analysis**. Chapter 24 builds on it."

24.1.8 Exercises

24.1.8.1 A. Check your understanding

Exercise. A1

(a) State the three clauses of Definition 24.1.1, and say which one is an exclusion rather than an assumption about accuracy.

(b) Define the unit roundoff, and give its value for a decimal model carrying t significant digits.

(c) Define the forward error and the backward error of an algorithm at a datum, and say which of the two a user can usually measure.

(d) Decide whether each statement is correct, with a reason: (i) "a backward stable algorithm returns an answer with small forward error"; (ii) "if the computed $\widehat{\mathbf{x}}$ has a tiny residual, then $\widehat{\mathbf{x}}$ is close to $\mathbf{x}$".

(e) Why is Definition 24.1.1 called a model rather than a theorem?

Solution.

(a) (FP1) rounding is relative, with error at most u; (FP2) each operation returns the rounding of its exact result, so $\mathrm{fl}(x \odot y) = (x \odot y)(1+\delta)$ with $|\delta| \le u$; (FP3) no overflow or underflow. (FP3) is the exclusion: it does not say the arithmetic is good, it says we decline to analyze the case where it is not.

(b) u is the bound in (FP1): $|\mathrm{fl}(t) - t| \le u|t|$ for every in-range t. For a decimal model with t significant digits, $u = \frac{1}{2} \cdot 10^{1-t}$ (Example (The model we compute in) does the case $t = 6$).

(c) See Definition 24.1.2. The forward error is $\|\widehat{f}(\mathbf{a}) - f(\mathbf{a})\|/\|f(\mathbf{a})\|$, the backward error the smallest $\|\Delta\mathbf{a}\|/\|\mathbf{a}\|$ with $f(\mathbf{a} + \Delta\mathbf{a}) = \widehat{f}(\mathbf{a})$. A user can usually measure neither directly, but the backward error is the one an analyst can *bound*, because a bound needs only one exhibited $\Delta\mathbf{a}$; computing the forward error needs the exact answer, which is what the computation was for.

(d) (i) Incorrect. It returns an answer with small *backward* error; the forward error is then at most about κ times that, by Theorem 24.1.3 (after absorbing any perturbation of $\mathbf{b}$, as the remark after that theorem does), and for large κ that is no guarantee at all. (ii) Incorrect, for the same reason: a small residual is a small backward error (see the warning above), which bounds the forward error only after multiplication by $\kappa(\mathbf{A})$.

(e) Because it is an assumption about a physical machine, not a consequence of anything proved earlier in the book. Nothing in Chapters 1 to 21 mentions rounding. Everything this chapter proves about arithmetic is a theorem *about the model*, true of any machine that satisfies it and silent about one that does not.

24.1.8.2 B. Practice

Exercise. B1: The γ calculus

Assume $(j + k)u < 1$.

(a) Prove that $1 + \gamma_k = 1/(1 - ku)$.

(b) Hence prove that $\gamma_j + \gamma_k + \gamma_j\gamma_k \le \gamma_{j+k}$.

(c) Explain in one sentence why (b) is the statement one needs when two separately analyzed computations are performed one after the other.

Solution.

(a) $1 + \gamma_k = 1 + ku/(1 - ku) = ((1 - ku) + ku)/(1 - ku) = 1/(1 - ku)$.

(b) Using (a) twice,

$$1 + \gamma_j + \gamma_k + \gamma_j\gamma_k = (1 + \gamma_j)(1 + \gamma_k) = \frac{1}{(1 - ju)(1 - ku)}.$$

Now $(1 - ju)(1 - ku) = 1 - (j + k)u + jku^2 \ge 1 - (j + k)u > 0$, and inverting reverses the inequality between positive numbers, so the right-hand side is at most

$1/(1-(j+k)u) = 1+\gamma_{j+k}$. Subtracting 1 gives the claim.

(c) If one stage multiplies the exact answer by $1+\theta_1$ with $|\theta_1| \le \gamma_j$ and the next by $1+\theta_2$ with $|\theta_2| \le \gamma_k$, the combined factor is $1+\theta_1+\theta_2+\theta_1\theta_2$, and (b) says the total is still of the form $1+\theta$ with $|\theta| \le \gamma_{j+k}$ — the counts simply add.

Exercise. B2: Summation

Let $x_1, \dots, x_n$ be representable, let $nu < 1$, and let s_n be computed by $s_1 = x_1$, $s_k = \mathrm{fl}(s_{k-1} + x_k)$.

(a) Prove that $s_n = \sum_{k=1}^{n} x_k(1+\theta_k)$ with $|\theta_k| \le \gamma_{n-1}$.

(b) Deduce $|s_n - \sum_k x_k| \le \gamma_{n-1} \sum_k |x_k|$, and explain why this is **not** a bound on the relative error of s_n.

(c) Give three representable numbers in the six-digit model for which the computed sum has relative error larger than u.

Solution.

(a) By (FP2), $s_k = (s_{k-1} + x_k)(1+\varepsilon_k)$ for $2 \le k \le n$, with $|\varepsilon_k| \le u$. Unrolling as in the proof of Theorem 24.1.2,

$$s_n = \sum_{k=1}^{n} x_k \prod_{j=\max(k,2)}^{n} (1+\varepsilon_j).$$

The $k=1$ and $k=2$ terms carry $n-1$ factors and the rest carry fewer, so Lemma 24.1.1 gives $|\theta_k| \le \gamma_{n-1}$.

(b) Subtract $\sum_k x_k$ and use the triangle inequality: $|s_n - \sum_k x_k| = |\sum_k x_k\theta_k| \le \gamma_{n-1} \sum_k |x_k|$. It is not a relative error bound because the right-hand side involves $\sum_k |x_k|$, not $|\sum_k x_k|$, and the two differ by an arbitrarily large factor when the x_k have mixed signs and cancel.

(c) Take $x_1 = 1$, $x_2 = 10^{-7}$, $x_3 = -1$, summed in that order. Then $\mathrm{fl}(1 + 10^{-7}) = 1$, since $1 + 10^{-7}$ lies strictly between the representable neighbors 1 and 1.00001 and is nearer the first; and $\mathrm{fl}(1-1) = 0$. The computed sum is 0, the true sum is 10^{-7}, and the relative error is 1. This is (b) at work: $\sum |x_k| = 2$ while $|\sum x_k| = 10^{-7}$.

Exercise. B3: The model by hand

Work in the six-digit decimal model.

(a) Compute $\mathrm{fl}(1/3)$ and verify (FP1) for it.

(b) Compute the two sides of $\mathrm{fl}\big(\mathrm{fl}(a+b)+c\big)$ and $\mathrm{fl}\big(a+\mathrm{fl}(b+c)\big)$ for $a = 10^5$, $b = 1$, $c = 1$, and say what this shows about the model.

Solution.

(a) $1/3 = 3.333333\overline{3} \times 10^{-1}$, so $\mathrm{fl}(1/3) = 0.333333$ and $|\mathrm{fl}(1/3) - 1/3| = \frac{1}{3} \times 10^{-6}$. Dividing by $1/3$ gives a relative error of 10^{-6}, which is at most $u = 5 \times 10^{-6}$.

(b) Left: $\mathrm{fl}(10^5+1) = 100001$ exactly (six digits), and $\mathrm{fl}(100001+1) = 100002$. Right: $\mathrm{fl}(1+1) = 2$, and $\mathrm{fl}(10^5+2) = 100002$. Here the two agree. Change a to 10^6: left gives $\mathrm{fl}(10^6+1) = 10^6$, then $\mathrm{fl}(10^6+1) = 10^6$; right gives $\mathrm{fl}(1+1) = 2$, then $\mathrm{fl}(10^6+2) = 10^6$ as well, since 1000002 rounds to 1.00000×10^6. Try $a = 10^6$, $b = c = 3$: left is 10^6 then 10^6; right is 6 then $\mathrm{fl}(1000006) = 1.00001 \times 10^6$. The two differ. **Floating-point addition is commutative but not associative**, so the *order* of a summation is part of the algorithm, not an irrelevance — which is why Theorem 24.1.2 names the order it analyzes.

24.1.8.3 C. Going deeper

Exercise. C1: When backward stability says nothing

Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ and let s_n be the computed inner product of Theorem 24.1.2.

(a) Prove that $|s_n - \mathbf{x}^\top \mathbf{y}| \le \gamma_n |\mathbf{x}|^\top |\mathbf{y}|$, where $|\cdot|$ is entrywise.

(b) Deduce that if all $x_i y_i \ge 0$ then the *relative* error of s_n is at most γ_n.

(c) Give $\mathbf{x}, \mathbf{y} \in \mathbb{R}^3$ with representable entries, in the six-digit model, for which the relative error of s_n is at least 1. Explain why this does not contradict Theorem 24.1.2.

Solution.

(a) By the theorem's proof, $s_n = \sum_k x_k y_k (1+\theta_k)$ with $|\theta_k| \le \gamma_n$, so

$$|s_n - \mathbf{x}^\top \mathbf{y}| = \Big|\sum_k x_k y_k \theta_k\Big| \le \gamma_n \sum_k |x_k||y_k| = \gamma_n |\mathbf{x}|^\top |\mathbf{y}|.$$

(b) If every $x_k y_k \ge 0$ then $|\mathbf{x}|^\top |\mathbf{y}| = \sum_k x_k y_k = \mathbf{x}^\top \mathbf{y}$, so (a) reads $|s_n - \mathbf{x}^\top \mathbf{y}| \le \gamma_n \, \mathbf{x}^\top \mathbf{y}$, which is the relative bound (and $\mathbf{x}^\top \mathbf{y} > 0$ unless every term vanishes, in which case $s_n = 0$ too).

(c) Take $\mathbf{x} = (1, 10^{-7}, -1)$ and $\mathbf{y} = (1,1,1)$, all representable. Each product $x_k y_k$ is computed exactly, so $\delta_1 = \delta_2 = \delta_3 = 0$. Then $s_1 = 1$; next $s_2 = \mathrm{fl}(1 + 10^{-7}) = 1$, which is $(1+10^{-7})(1+\varepsilon_2)$ with

$$\varepsilon_2 = \frac{-10^{-7}}{1+10^{-7}}, \qquad |\varepsilon_2| < 10^{-7} < u;$$

finally $s_3 = \mathrm{fl}(1-1) = 0$, exactly, so $\varepsilon_3 = 0$ serves. The true value is $\mathbf{x}^\top \mathbf{y} = 10^{-7}$, so the forward relative error is 1.

There is no contradiction. Following the proof, $\theta_1 = \theta_2 = \varepsilon_2$ and $\theta_3 = 0$, all of modulus below $u \le \gamma_3$, and indeed

$$(1+\varepsilon_2) + 10^{-7}(1+\varepsilon_2) - 1 = (1+10^{-7})(1+\varepsilon_2) - 1 = 0 = s_3.$$

So the backward perturbation $\Delta\mathbf{x} = (\varepsilon_2, 10^{-7}\varepsilon_2, 0)$ is of relative size about 10^{-7} in every coordinate: the algorithm solved a neighboring problem exactly. What made the answer worthless is the *problem*, not the method: here $|\mathbf{x}|^\top |\mathbf{y}| = 2 + 10^{-7}$ while $\mathbf{x}^\top \mathbf{y} = 10^{-7}$, so part (a)'s bound is twenty million times the quantity being computed. This is the same phenomenon as Example (Two ways to solve a quadratic), seen through the backward error.

Exercise. C2: An unstable way to compute a variance

Given $x_1, \dots, x_N$ with mean $\overline{x} = \frac{1}{N}\sum_i x_i$, the quantity $S = \sum_i (x_i - \overline{x})^2$ can be computed in one pass as $S = \sum_i x_i^2 - \frac{1}{N}\left(\sum_i x_i\right)^2$, or in two passes by forming $\overline{x}$ first.

(a) Verify the identity $\sum_i (x_i - \overline{x})^2 = \sum_i x_i^2 - \frac{1}{N}(\sum_i x_i)^2$.

(b) In the six-digit model, run both routes on $x = (10000, 10001, 10002)$, showing every operation, and compare with the exact value $S = 2$.

(c) Say which route is to be preferred and why, in the vocabulary of this section.

Solution.

(a) Expand: $\sum_i (x_i - \overline{x})^2 = \sum_i x_i^2 - 2\overline{x}\sum_i x_i + N\overline{x}^2$. Since $\sum_i x_i = N\overline{x}$, the last two terms are $-2N\overline{x}^2 + N\overline{x}^2 = -N\overline{x}^2 = -\frac{1}{N}(\sum_i x_i)^2$.

(b) *One pass.* The squares are 10^8, 100020001, 100040004; rounding each to six digits gives 1.00000×10^8, 1.00020×10^8, and 1.00040×10^8, and accumulating gives $\mathrm{fl}(1.00000 \times 10^8 + 1.00020 \times 10^8) = 2.00020 \times 10^8$ and then 3.00060×10^8. The sum $\sum_i x_i = 30003$ is exact, its square is 900180009, which rounds to 9.00180×10^8, and dividing by 3 gives 3.00060×10^8. Subtracting: $\mathrm{fl}(3.00060 \times 10^8 - 3.00060 \times 10^8) = \mathbf{0}$.
Two passes. $\overline{x} = \mathrm{fl}(30003/3) = 10001$, exactly. The deviations $-1, 0, 1$ are computed exactly, their squares are $1, 0, 1$, and the sum is $\mathbf{2}$, the exact answer.

(c) The two-pass route. The one-pass route subtracts two nine-digit numbers that agree in their first six digits, so in a six-digit model it commits catastrophic cancellation and returns 0; worse, 0 is not the value of S for *any* small relative perturbation of the data. Indeed $S(\mathbf{x} + \Delta\mathbf{x}) = 0$ forces the three perturbed entries to be equal, say to v; and for $\mathbf{x} = (10000, 10001, 10002)$ the largest of $|x_i - v|$ is at least 1, whatever v is, because $x_3 - x_1 = 2$. So every $\Delta\mathbf{x}$ that accounts for the computed answer has

$$\frac{\|\Delta\mathbf{x}\|_\infty}{\|\mathbf{x}\|_\infty} \geq \frac{1}{10002} > 19\,u,$$

whereas Definition 24.1.3 allows cu. One datum bounds the constant rather than disposing of it — what this shows exactly is that no $c \leq 19$ will serve for the one-pass route — but the mechanism is general, and it is the one Example (Two ways to solve a quadratic) identified. The two-pass route performs the subtraction $x_i - \overline{x}$ *first*, while the operands are still exactly representable, and nothing is ever lost. Same identity, same arithmetic, different algorithm — which is precisely the distinction between a problem and a method.

24.2 Elimination, Pivoting, and Cost

Chapter 3 solved $\mathbf{Ax} = \mathbf{b}$ by elimination and packaged the work as $\mathbf{A} = \text{Ł}\mathbf{U}$ or $\mathbf{PA} = \text{Ł}\mathbf{U}$. It said nothing about how much arithmetic that costs, and almost nothing about what rounding does to it — only a remark, about a two-by-two matrix with a tiny corner entry, ending: "Chapter 24 makes this precise." Chapter 8 left a second note: "Chapter 24 reads Gaussian elimination itself as a sequence of Schur complements of 1×1 pivots." Chapter 13 left a third, that Cholesky "does about half the arithmetic of a general LU factorization" and needs no pivoting.

This section pays all three. Section 1 supplied the model and the word *backward stable*; here

that word is applied to the oldest algorithm in the book. Throughout, matrices are **real**, since the representable numbers of Definition 24.1.1 are real, and $|\mathbf{M}|$ denotes the entrywise absolute value ($|m_{ij}|$), with $|\mathbf{M}| \le |\mathbf{N}|$ read entrywise.

24.2.1 Elimination is a sequence of Schur complements

Chapter 8 built a machine for one block elimination: if $\mathbf{M} = \left(\begin{smallmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{C} & \mathbf{D}\end{smallmatrix}\right)$ with $\mathbf{A}$ invertible, then subtracting $\mathbf{C}\mathbf{A}^{-1}$ times the first block row from the second leaves $\left(\begin{smallmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{O} & \mathbf{M}/\mathbf{A}\end{smallmatrix}\right)$, where $\mathbf{M}/\mathbf{A} = \mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B}$ is the Schur complement (Definition 8.3.1, Theorem 8.3.1). Gaussian elimination is that machine run with a 1×1 block, $n-1$ times. Saying so exactly is the cleanest description of the algorithm there is, and it is what makes block versions of it possible.

Theorem 24.2.1. Elimination Produces Schur Complements

Let $\mathbf{A} \in M_n(F)$ and $1 \le k \le n-1$, and suppose the first k steps of elimination without row swaps (Chapter 3 §06) are carried out with **nonzero pivots**, producing $\mathbf{A}^{(k)}$. Partition

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_k & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}, \qquad \mathbf{A}^{(k)} = \begin{pmatrix} \mathbf{V} & \mathbf{B}' \\ \mathbf{O} & \mathbf{S} \end{pmatrix},$$

with $\mathbf{A}_k, \mathbf{V} \in M_k(F)$, the zero block being what the first k steps created. Then $\mathbf{A}_k$ is invertible and

$$\mathbf{S} = \mathbf{A}/\mathbf{A}_k = \mathbf{D} - \mathbf{C}\,\mathbf{A}_k^{-1}\mathbf{B}.$$

In particular, for $k = 1$ and $a_{11} \neq 0$, one step of elimination replaces the trailing $(n-1) \times (n-1)$ block by the Schur complement of the 1×1 pivot block.

Idea. The first k steps multiply $\mathbf{A}$ on the left by a unit lower triangular matrix whose columns $k+1, \dots, n$ are those of the identity — that is exactly what Lemma 3.6.3 (c) says, and it is the whole point. Such a matrix has block form $\left(\begin{smallmatrix}Ł_{11} & \mathbf{O}\\ Ł_{21} & \mathbf{I}\end{smallmatrix}\right)$. Multiply that by the partitioned $\mathbf{A}^{(k)}$, match the four blocks against $\mathbf{A}$, and the $Ł_{11}$ cancels out of the bottom-right block, leaving the Schur complement.

Proof. Write $\boldsymbol{\ell}_j$ for the j-th multiplier vector, whose first j entries are zero, and recall from Chapter 3 §06 that step j multiplies on the left by $\mathbf{G}_j(-\boldsymbol{\ell}_j)$, with inverse $\mathbf{G}_j(\boldsymbol{\ell}_j)$ by Lemma 3.6.3 (b). Hence $\mathbf{A} = Ł_k \mathbf{A}^{(k)}$ where, by Lemma 3.6.3 (c),

$$Ł_k = \mathbf{G}_1(\boldsymbol{\ell}_1) \cdots \mathbf{G}_k(\boldsymbol{\ell}_k) = \mathbf{I} + \sum_{j=1}^{k} \boldsymbol{\ell}_j \mathbf{e}_j^{\top}.$$

Only columns $1, \dots, k$ of $Ł_k$ differ from those of $\mathbf{I}$, so partitioning $Ł_k$ as $\mathbf{A}$ is partitioned gives

$$Ł_k = \begin{pmatrix} Ł_{11} & \mathbf{O} \\ Ł_{21} & \mathbf{I}_{n-k} \end{pmatrix},$$

with $Ł_{11} \in M_k(F)$ unit lower triangular, hence invertible by Lemma 3.6.1 (b).

By Theorem 8.1.1,

$$\begin{pmatrix} \mathbf{A}_k & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix} = \begin{pmatrix} Ł_{11} & \mathbf{O} \\ Ł_{21} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{V} & \mathbf{B}' \\ \mathbf{O} & \mathbf{S} \end{pmatrix} = \begin{pmatrix} Ł_{11}\mathbf{V} & Ł_{11}\mathbf{B}' \\ Ł_{21}\mathbf{V} & Ł_{21}\mathbf{B}' + \mathbf{S} \end{pmatrix}.$$

Comparing blocks: $\mathbf{A}_k = Ł_{11}\mathbf{V}$, $\mathbf{B} = Ł_{11}\mathbf{B}'$, $\mathbf{C} = Ł_{21}\mathbf{V}$, and $\mathbf{D} = Ł_{21}\mathbf{B}' + \mathbf{S}$.

Now $\mathbf{V}$ is upper triangular with the first k pivots on its diagonal, and those are nonzero by hypothesis, so $\mathbf{V}$ is invertible by Lemma 3.6.2. Therefore $\mathbf{A}_k = Ł_{11}\mathbf{V}$ is invertible, and $\mathbf{A}_k^{-1} = \mathbf{V}^{-1}Ł_{11}^{-1}$. Hence

$$\mathbf{C}\,\mathbf{A}_k^{-1}\mathbf{B} = (Ł_{21}\mathbf{V})(\mathbf{V}^{-1}Ł_{11}^{-1})(Ł_{11}\mathbf{B}') = Ł_{21}\mathbf{B}',$$

and so $\mathbf{S} = \mathbf{D} - Ł_{21}\mathbf{B}' = \mathbf{D} - \mathbf{C}\mathbf{A}_k^{-1}\mathbf{B} = \mathbf{A}/\mathbf{A}_k$. This proves the theorem.

■

Example. Two steps, two Schur complements

Let

$$\mathbf{A} = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 7 & 5 \\ -1 & 4 & 8 \end{pmatrix}.$$

Carry out two steps of elimination without row swaps, then compute $\mathbf{A}/\mathbf{A}_1$ and $\mathbf{A}/\mathbf{A}_2$ directly from Definition 8.3.1, and check Theorem 24.2.1 in both cases.

Solution. The elimination. The first multipliers are $\ell_{21} = 2/1 = 2$ and $\ell_{31} = -1/1 = -1$; subtracting ℓ_{i1} times row 1 from row i gives $\mathbf{A}^{(1)}$. The second multiplier is $\ell_{32} = 6/3 = 2$, and row 3 becomes $(0, 0, 11 - 2(-1)) = (0, 0, 13)$:

$$\mathbf{A}^{(1)} = \begin{pmatrix} 1 & 2 & 3 \\ 0 & 3 & -1 \\ 0 & 6 & 11 \end{pmatrix}, \qquad \mathbf{A}^{(2)} = \begin{pmatrix} 1 & 2 & 3 \\ 0 & 3 & -1 \\ 0 & 0 & 13 \end{pmatrix}.$$

The Schur complement of the 1×1 block. Here $\mathbf{A}_1 = (1)$, $\mathbf{B} = \begin{pmatrix} 2 & 3 \end{pmatrix}$, $\mathbf{C} = \begin{pmatrix} 2 & -1 \end{pmatrix}^{\top}$, and $\mathbf{D} = \left(\begin{smallmatrix} 7 & 5 \\ 4 & 8 \end{smallmatrix}\right)$, so

$$\mathbf{A}/\mathbf{A}_1 = \mathbf{D} - \mathbf{C}\mathbf{A}_1^{-1}\mathbf{B} = \begin{pmatrix} 7 & 5 \\ 4 & 8 \end{pmatrix} - \begin{pmatrix} 2 \\ -1 \end{pmatrix}\begin{pmatrix} 2 & 3 \end{pmatrix}$$
$$= \begin{pmatrix} 7 & 5 \\ 4 & 8 \end{pmatrix} - \begin{pmatrix} 4 & 6 \\ -2 & -3 \end{pmatrix} = \begin{pmatrix} 3 & -1 \\ 6 & 11 \end{pmatrix},$$

which is exactly the trailing block of $\mathbf{A}^{(1)}$.

The Schur complement of the 2×2 block. Now $\mathbf{A}_2 = \left(\begin{smallmatrix} 1 & 2 \\ 2 & 7 \end{smallmatrix}\right)$, with $\det \mathbf{A}_2 = 3$ and $\mathbf{A}_2^{-1} =$

$\frac{1}{3}\left(\begin{smallmatrix} 7 & -2 \\ -2 & 1 \end{smallmatrix}\right)$, while $\mathbf{C} = \begin{pmatrix} -1 & 4 \end{pmatrix}$ and $\mathbf{B} = \begin{pmatrix} 3 & 5 \end{pmatrix}^{\top}$. Then

$$\mathbf{C}\mathbf{A}_2^{-1} = \tfrac{1}{3}\begin{pmatrix} -15 & 6 \end{pmatrix} = \begin{pmatrix} -5 & 2 \end{pmatrix},$$
$$\mathbf{A}/\mathbf{A}_2 = 8 - \begin{pmatrix} -5 & 2 \end{pmatrix}\begin{pmatrix} 3 \\ 5 \end{pmatrix} = 8 - (-5) = 13,$$

the trailing 1×1 block of $\mathbf{A}^{(2)}$. Both readings agree with Theorem 24.2.1, and as an arithmetic check the pivots $1, 3, 13$ multiply to 39, which is $\det \mathbf{A}$ — the corollary below says why that is no coincidence.

Two consequences come out for free. The first identifies the pivots with ratios of minors, which is how Chapter 13 §01 wrote them in Sylvester's criterion.

Corollary 24.2.2. Pivots Are Ratios of Leading Minors

Let $\mathbf{A} \in M_n(F)$, let $1 \le k \le n-1$, and write $\mathbf{A}_j$ for the leading principal $j \times j$ submatrix of $\mathbf{A}$, with $\det \mathbf{A}_0 := 1$.

(a) If the first k pivots of elimination without row swaps are nonzero, then $\mathbf{A}_1, \dots, \mathbf{A}_k$ are invertible and the j-th pivot equals $\det \mathbf{A}_j / \det \mathbf{A}_{j-1}$ for $j \le k$.

(b) Conversely, if $\mathbf{A}_1, \dots, \mathbf{A}_{n-1}$ are invertible then elimination without row swaps succeeds on $\mathbf{A}$ with nonzero pivots, and $\mathbf{A} = Ł\mathbf{U}$ by Theorem 3.6.4.

Proof.

(a) Invertibility of $\mathbf{A}_1, \dots, \mathbf{A}_k$ is Theorem 24.2.1 applied with $1, \dots, k$ in place of k. For $j = 1$ the pivot is $a_{11} = \det \mathbf{A}_1 / \det \mathbf{A}_0$, as claimed. For $2 \le j \le k$, the j-th pivot is the (j,j) entry of $\mathbf{A}^{(j-1)}$, which by the theorem (with $j-1$ in place of k) is the $(1,1)$ entry of $\mathbf{A}/\mathbf{A}_{j-1}$. That entry is a_{jj} minus row j of $\mathbf{C}$ times $\mathbf{A}_{j-1}^{-1}$ times column j of $\mathbf{B}$, so it involves only rows and columns $1, \dots, j$ of $\mathbf{A}$: it equals the 1×1 matrix $\mathbf{A}_j/\mathbf{A}_{j-1}$. By Theorem 8.3.2 (a), $\det \mathbf{A}_j = \det \mathbf{A}_{j-1} \cdot \det(\mathbf{A}_j/\mathbf{A}_{j-1})$, which is the claim.

(b) We show by induction on j that the j-th pivot is nonzero, for $j = 1, \dots, n-1$. For $j = 1$ the pivot is $a_{11} = \det \mathbf{A}_1 \neq 0$. Let $2 \le j \le n-1$ and suppose the first $j-1$ pivots are nonzero. Then Theorem 24.2.1 applies with $j-1$ in place of k, and the calculation in (a) identifies the j-th pivot as $\det \mathbf{A}_j / \det \mathbf{A}_{j-1}$, which is nonzero because $\mathbf{A}_j$ and $\mathbf{A}_{j-1}$ are invertible (Theorem 7.5.2). So elimination never stops, and Theorem 3.6.4 gives $\mathbf{A} = Ł\mathbf{U}$.

■

Quick check. The matrix $\left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$ has no LU factorization (Example (An invertible matrix with no LU factorization)). Which hypothesis of Corollary 24.2.2 (b) fails, and what does Theorem 24.2.1 say about it?

Solution. $\mathbf{A}_1 = \mathbf{O}_1$ is not invertible, so (b) does not apply. Theorem 24.2.1 does not apply either: it assumes the first pivot is nonzero, and here it is 0, so the Schur complement $\mathbf{A}/\mathbf{A}_1$

is not even defined — Definition 8.3.1 requires the pivot block to be invertible. This is the whole content of Chapter 3's "elimination stops": there is no Schur complement to form.

24.2.2 What elimination costs

A **flop** is one arithmetic operation: one addition, subtraction, multiplication, or division, or one square root. Counting flops is crude — it ignores which operations are slow and how data moves — but it captures how the work grows with n. We quote leading terms, writing $\frac{2}{3}n^3+O(n^2)$ and meaning it. **The constant matters and the exponent matters more**: halving the constant is worth having, but a smaller exponent overtakes it at some size and stays ahead forever.

Proposition 24.2.3. The Cost of Elimination

Counting each of the five operations as one flop, and counting only arithmetic on matrix and vector entries:

(a) Elimination without row swaps on $\mathbf{A} \in M_n(\mathbb{R})$ costs

$$\frac{n(n-1)(4n+1)}{6} = \tfrac{2}{3}n^3 - \tfrac{1}{2}n^2 - \tfrac{1}{6}n$$

flops, that is $\frac{2}{3}n^3 + O(n^2)$.

(b) Forward substitution with a unit lower triangular Ł costs $n^2 - n$ flops and back substitution with an upper triangular $\mathbf{U}$ costs n^2, so solving $\mathbf{Ax} = \mathbf{b}$ from a factorization costs $2n^2 - n$ flops.

(c) The Cholesky recursion of Chapter 13 §02 on a positive definite $\mathbf{A} \in M_n(\mathbb{R})$ costs

$$\frac{n(n+1)(2n+1)}{6} = \tfrac{1}{3}n^3 + \tfrac{1}{2}n^2 + \tfrac{1}{6}n$$

flops, that is $\frac{1}{3}n^3 + O(n^2)$, of which n are square roots.

Proof.

(a) At step k the algorithm computes $n-k$ multipliers, one division each, and then replaces each of the $(n-k)^2$ entries of the trailing block by $a_{ij} - \ell_{ik}a_{kj}$, two flops each. The entries below the pivot are set to zero without arithmetic. So the cost is

$$\sum_{k=1}^{n-1}\Big[(n-k) + 2(n-k)^2\Big] = \sum_{j=1}^{n-1}(j + 2j^2) = \frac{(n-1)n}{2} + \frac{(n-1)n(2n-1)}{3},$$

using $\sum_{j\le m} j = m(m+1)/2$ and $\sum_{j\le m} j^2 = m(m+1)(2m+1)/6$ with $m = n-1$. Putting the two terms over 6 gives $n(n-1)(3+2(2n-1))/6 = n(n-1)(4n+1)/6$.

(b) Forward substitution computes $x_i = b_i - \sum_{j<i}\ell_{ij}x_j$ with no division, since $\ell_{ii} = 1$: that is $2(i-1)$ flops, and $\sum_{i\le n} 2(i-1) = n(n-1)$. Back substitution computes $x_i = (b_i - \sum_{j>i} u_{ij}x_j)/u_{ii}$, which is $2(n-i)+1$ flops, and $\sum_{i\le n}(2(n-i)+1) = n(n-1)+n = n^2$.

(c) Column j of the recursion computes $\ell_{jj} = \big(a_{jj} - \sum_{k<j}\ell_{jk}^2\big)^{1/2}$, costing $2(j-1)$ flops for the sum and one square root, and then $\ell_{ij} = (a_{ij} - \sum_{k<j}\ell_{ik}\ell_{jk})/\ell_{jj}$ for each of the

$n-j$ rows below, costing $2(j-1)+1$ flops each. The total is

$$\sum_{j=1}^{n}(n-j+1)(2j-1) = \sum_{i=1}^{n} i\,(2n-2i+1)$$
$$= (2n+1)\frac{n(n+1)}{2} - 2\cdot\frac{n(n+1)(2n+1)}{6},$$

substituting $i = n-j+1$; the two terms combine to $n(n+1)(2n+1)(\frac{1}{2}-\frac{1}{3}) = n(n+1)(2n+1)/6$.

■

Comparing (a) and (c): $\frac{1}{3}n^3$ against $\frac{2}{3}n^3$. This is Chapter 13 §02's "Cholesky does about half the arithmetic of a general LU factorization", now a count rather than a claim, for the reason that section gave — the recursion produces one triangular factor instead of two. Part (b) says the factorization is the expensive part: solving m systems with the same $\mathbf{A}$ costs $\frac{2}{3}n^3 + 2mn^2$, not m times $\frac{2}{3}n^3$.

24.2.3 One row at a time

Every algorithm below computes its entries the same way: accumulate a sum of products into a running total, then possibly divide. Analyzing that pattern once saves doing it three times.

Lemma 24.2.4. One Row of Substitution

Let $m \ge 0$, let $t_1,\dots,t_m$, $y_1,\dots,y_m$, c, and $d \ne 0$ be representable, and suppose

$$s_0 = c, \qquad s_p = \mathrm{fl}(s_{p-1} - \mathrm{fl}(t_p y_p)) \quad (p = 1,\dots,m), \qquad z = \mathrm{fl}(s_m/d)$$

are computed in the model of Definition 24.1.1, with $(m+1)u < 1$. Then there are $\theta_0,\dots,\theta_m$ with

$$c = d\,z\,(1+\theta_0) + \sum_{p=1}^{m} t_p y_p (1+\theta_p), \qquad |\theta_p| \le \gamma_{m+1} \ \text{ for all } p.$$

If the final division is omitted, so that $z = s_m$, the same identity holds with $d = 1$ and $|\theta_p| \le \gamma_m$.

Idea. Unroll the recursion, then divide the whole identity by the product of the accumulation factors. Dividing is the trick: it moves those factors onto c's neighbors as **inverse** factors, which is exactly what Lemma 24.1.1 was stated to allow.

Proof. By (FP2) there are $\delta_p, \varepsilon_p, \eta$, all of modulus at most u, with $\mathrm{fl}(t_p y_p) = t_p y_p(1+\delta_p)$, $s_p = (s_{p-1} - t_p y_p(1+\delta_p))(1+\varepsilon_p)$, and $z = (s_m/d)(1+\eta)$. Induction on m gives

$$s_m = c\prod_{p=1}^{m}(1+\varepsilon_p) - \sum_{p=1}^{m} t_p y_p (1+\delta_p)\prod_{q=p}^{m}(1+\varepsilon_q),$$

the case $m = 0$ being $s_0 = c$, and the step from $m-1$ to m multiplying everything by $1+\varepsilon_m$ and subtracting the new term.

Each $1+\varepsilon_p$ is nonzero, since $|\varepsilon_p| \le u < 1$, so we may divide the identity by $\prod_{p\le m}(1+\varepsilon_p)$ and rearrange:

$$c = s_m \prod_{p=1}^{m}(1+\varepsilon_p)^{-1} + \sum_{p=1}^{m} t_p y_p (1+\delta_p) \prod_{q=1}^{p-1}(1+\varepsilon_q)^{-1}.$$

From $z = (s_m/d)(1+\eta)$ we get $s_m = d\,z\,(1+\eta)^{-1}$. So the first term is $d\,z$ times a product of $m+1$ factors of the form $(1+\delta)^{\pm 1}$ with $|\delta| \le u$, which by Lemma 24.1.1 is $1+\theta_0$ with $|\theta_0| \le \gamma_{m+1}$. The p-th term of the sum carries $1+(p-1) = p \le m$ such factors, so it is $t_p y_p(1+\theta_p)$ with $|\theta_p| \le \gamma_p \le \gamma_{m+1}$.

If the division is omitted, the first term carries only the m inverse factors, so $|\theta_0| \le \gamma_m$, and the rest of the argument is unchanged. This proves the lemma.

■

24.2.4 Solving a triangular system

Here is the cleanest backward error analysis in the subject, and the model for every other one in this chapter. Note in advance what it does **not** assume: nothing about the size of the diagonal entries of $\mathbf{T}$, nothing about its conditioning, and no pivoting of any kind.

Theorem 24.2.5. Backward Error of a Triangular Solve

Let $\mathbf{T} \in M_n(\mathbb{R})$ be lower triangular with representable entries and nonzero diagonal, let $\mathbf{b} \in \mathbb{R}^n$ be representable, and let $nu < 1$. Let $\widehat{\mathbf{x}}$ be computed by forward substitution in the model: for each i, the products $t_{ij}\widehat{x}_j$ are subtracted from b_i one at a time in increasing order of j, and the result is divided by t_{ii} — that is, exactly the computation of Lemma 24.2.4 with $m = i-1$. Then there is a lower triangular $\Delta\mathbf{T}$ with

$$(\mathbf{T}+\Delta\mathbf{T})\widehat{\mathbf{x}} = \mathbf{b} \qquad \text{and} \qquad |\Delta t_{ij}| \le \gamma_i\,|t_{ij}| \le \gamma_n\,|t_{ij}| \quad (j \le i).$$

The same statement, with γ_{n+1-i} in place of γ_i, holds for back substitution with an upper triangular $\mathbf{T}$.

Idea. Row i *is* the pattern of Lemma 24.2.4, with $c = b_i$, the $m = i-1$ products $t_{ij}\widehat{x}_j$, and the final division by t_{ii}. The lemma hands back an identity in which every coefficient of $\widehat{x}_j$ is t_{ij} times something within γ_i of 1; that "something minus one", times t_{ij}, is Δt_{ij}.

Proof. Fix i. The quantities $\widehat{x}_1, \dots, \widehat{x}_{i-1}$ are already computed, hence representable, so Lemma 24.2.4 applies with $m = i-1$, $c = b_i$, $t_p = t_{ip}$, $y_p = \widehat{x}_p$, $d = t_{ii}$, and $z = \widehat{x}_i$. Since $m+1 = i \le n$ and $nu < 1$, it gives $\theta_0, \dots, \theta_{i-1}$ of modulus at most γ_i with

$$b_i = t_{ii}\widehat{x}_i(1+\theta_0) + \sum_{j<i} t_{ij}\widehat{x}_j(1+\theta_j).$$

Define $\Delta t_{ii} = t_{ii}\theta_0$, $\Delta t_{ij} = t_{ij}\theta_j$ for $j < i$, and $\Delta t_{ij} = 0$ for $j > i$. Then $\Delta\mathbf{T}$ is lower triangular, $|\Delta t_{ij}| \le \gamma_i|t_{ij}| \le \gamma_n|t_{ij}|$ since $\gamma_i \le \gamma_n$, and the display says precisely that row i of $(\mathbf{T}+\Delta\mathbf{T})\widehat{\mathbf{x}}$ equals b_i. As i was arbitrary, $(\mathbf{T}+\Delta\mathbf{T})\widehat{\mathbf{x}} = \mathbf{b}$.

For back substitution, row i has $n-i$ products instead of $i-1$, so the lemma applies with

$m = n - i$ and the bound is γ_{n+1-i}. This proves the theorem.

■

This is as strong as a backward error statement gets. The perturbation is bounded **entrywise**, and entrywise *relative to the entries of* $\mathbf{T}$ *itself*: a zero entry of $\mathbf{T}$ is not perturbed at all, and a tiny entry is perturbed only in its own last digits. So a triangular solve is backward stable in the sense of Definition 24.1.3 — with the constant $c = 1.02\,n$ in any model with $nu \le 0.01$, since $\gamma_n \le 1.02\,nu$ there — whatever $\kappa(\mathbf{T})$ may be.

Warning. **Backward stable is not accurate, and this theorem is where the point bites hardest.** Theorem 24.2.5 holds for every nonsingular triangular $\mathbf{T}$. It does not say $\widehat{\mathbf{x}}$ is near $\mathbf{x}$. By Theorem 24.1.3 the forward error can be about $\kappa(\mathbf{T})\gamma_n$, and triangular matrices can be spectacularly ill-conditioned. Take the $n \times n$ matrix $\mathbf{T}$ with 1 on the diagonal and -1 everywhere above it. Its row sums of moduli are $n, n-1, \dots, 1$, so $\|\mathbf{T}\|_\infty = n$. Its inverse has 1 on the diagonal and 2^{j-i-1} in position (i,j) for $j > i$: back substitution on $\mathbf{Tx} = \mathbf{e}_j$ gives $x_j = 1$ and $x_i = \sum_{p>i} x_p$ for $i < j$, and each partial sum of $1, 1, 2, 4, \dots$ is the next term. Hence $\|\mathbf{T}^{-1}\|_\infty = 1 + (1 + 2 + \cdots + 2^{n-2}) = 2^{n-1}$ and $\kappa_\infty(\mathbf{T}) = n2^{n-1}$, so for $n = 60$ a backward stable solve can still return nonsense. Stability is a promise about the *data*, and a promise about the data is worth only what the conditioning of the problem makes it worth.

24.2.5 The factorization itself

For the factorization we analyze the ordering that computes each entry of Ł and $\mathbf{U}$ by one accumulation. In exact arithmetic it produces the same factors as Chapter 3's row operations — for invertible $\mathbf{A}$ this is Theorem 3.6.5 — but it groups the arithmetic differently, and rounding notices the difference.

Algorithm. **Elimination, one entry at a time.** Input: $\mathbf{A} \in M_n(\mathbb{R})$. Set $\ell_{ii} = 1$. For $i = 1, 2, \dots, n$:

- for $j = i, \dots, n$, put $u_{ij} = a_{ij} - \sum_{k<i} \ell_{ik} u_{kj}$;
- if $u_{ii} = 0$, stop; otherwise for $r = i+1, \dots, n$, put $\ell_{ri} = (a_{ri} - \sum_{k<i} \ell_{rk} u_{ki})/u_{ii}$.

Theorem 24.2.6. Backward Error of a Factorization

Let $\mathbf{A} \in M_n(\mathbb{R})$ have representable entries, let $nu < 1$, and suppose the algorithm above runs to completion in the model of Definition 24.1.1, producing $\widehat{\text{Ł}}$ and $\widehat{\mathbf{U}}$. Then

$$\widehat{\text{Ł}}\widehat{\mathbf{U}} = \mathbf{A} + \Delta\mathbf{A} \qquad \text{with} \qquad |\Delta\mathbf{A}| \le \gamma_n \, |\widehat{\text{Ł}}| \, |\widehat{\mathbf{U}}|$$

entrywise.

Idea. Each entry is one application of Lemma 24.2.4: the entries of $\mathbf{U}$ with no division, the entries of Ł with one. In both cases the lemma returns a_{ij} as a sum of the products $\widehat{\ell}_{ip}\widehat{u}_{pj}$, each with its own factor $1 + \theta_p$. Collecting the terms is exactly forming the (i,j) entry of $\widehat{\text{Ł}}\widehat{\mathbf{U}}$, and the θ_p are what is left over.

Proof. *Case $i \le j$.* The computation of $\widehat{u}_{ij}$ is Lemma 24.2.4 with $m = i-1$, $c = a_{ij}$, $t_p = \widehat{\ell}_{ip}$, $y_p = \widehat{u}_{pj}$, no division, and $z = \widehat{u}_{ij}$. Writing θ_i for the θ_0 of the lemma, and using $\widehat{\ell}_{ii} = 1$,

$$a_{ij} = \sum_{p=1}^{i} \widehat{\ell}_{ip}\widehat{u}_{pj}(1+\theta_p), \qquad |\theta_p| \le \gamma_{i-1}.$$

(For $i = 1$ the sum has one term and $\theta_1 = 0$.) Since $\widehat{\ell}_{ip} = 0$ for $p > i$, the sum without the θ_p is $(\widehat{\mathrm{L}}\widehat{\mathbf{U}})_{ij}$. Hence

$$|a_{ij} - (\widehat{\mathrm{L}}\widehat{\mathbf{U}})_{ij}| = \Big|\sum_{p\le i} \widehat{\ell}_{ip}\widehat{u}_{pj}\theta_p\Big| \le \gamma_{i-1}\sum_{p\le i}|\widehat{\ell}_{ip}||\widehat{u}_{pj}| = \gamma_{i-1}\big(|\widehat{\mathrm{L}}||\widehat{\mathbf{U}}|\big)_{ij},$$

the last equality again because $\widehat{\ell}_{ip} = 0$ for $p > i$.

Case $i > j$. The computation of $\widehat{\ell}_{ij}$ is Lemma 24.2.4 with $m = j-1$, $c = a_{ij}$, $t_p = \widehat{\ell}_{ip}$, $y_p = \widehat{u}_{pj}$, $d = \widehat{u}_{jj}$, and $z = \widehat{\ell}_{ij}$. Writing θ_j for θ_0,

$$a_{ij} = \sum_{p=1}^{j} \widehat{\ell}_{ip}\widehat{u}_{pj}(1+\theta_p), \qquad |\theta_p| \le \gamma_j.$$

Since $\widehat{u}_{pj} = 0$ for $p > j$, the sum without the θ_p is again $(\widehat{\mathrm{L}}\widehat{\mathbf{U}})_{ij}$, and the same estimate gives $|a_{ij} - (\widehat{\mathrm{L}}\widehat{\mathbf{U}})_{ij}| \le \gamma_j(|\widehat{\mathrm{L}}||\widehat{\mathbf{U}}|)_{ij}$.

In both cases the constant is at most γ_n. This proves the theorem.

■

Remark. **What is proved and what is not.** The theorem is about the ordering displayed above. Chapter 3's ordering, which updates the whole trailing block at each step, satisfies the same bound $|\Delta\mathbf{A}| \le \gamma_n|\widehat{\mathrm{L}}||\widehat{\mathbf{U}}|$; the proof is a different induction of comparable length, and **we do not give it here**. Nothing later in this chapter uses it. Also, the theorem assumes the algorithm completes, which in the model means every computed $\widehat{u}_{ii}$ is nonzero; it says nothing about what happens when one of them is zero, or about the case where a pivot is nonzero only by a rounding accident. One more seam will appear below: the growth bound for partial pivoting, proved in the next subsection, concerns the **exact** elimination in Chapter 3's ordering, whereas the bound just proved concerns the factors this ordering computes in the model. In exact arithmetic, and for invertible $\mathbf{A}$, the two orderings produce the same Ł and $\mathbf{U}$, by Theorem 3.6.5; in the model they agree only to within rounding, and we do not quantify that here. Nothing below is allowed to cross that seam quietly. Eq. 24.2.1, proved next, measures growth in the **computed** factors, which is a quantity one can read off after the fact; the a priori bound 2^{n-1} of Theorem 24.2.8 is a statement about exact elimination. The three places that put the two together — the warning after Example (A matrix that really grows) and the solutions to A1 and B3 — say that they are doing so, and each of them uses the combination only to show that even its most favorable reading proves nothing.

The bound is useless until we know how big $|\widehat{\mathrm{L}}||\widehat{\mathbf{U}}|$ is, and *that* is not controlled by $\mathbf{A}$. It is controlled by what the algorithm does to $\mathbf{A}$, which is why elimination needs a pivoting rule at

all. The quantity to watch has a name.

Definition 24.2.1. Growth factor

Let elimination — with whatever pivoting rule is in force — be run on $\mathbf{A} \in M_n(\mathbb{R})$, producing an upper factor $\mathbf{U}$. The **growth factor** of that run is

$$g_n := \frac{\max_{i,j}|u_{ij}|}{\max_{i,j}|a_{ij}|},$$

the largest entry in modulus the algorithm produces, relative to the largest entry in modulus it was given. For a run in the model, $\mathbf{U}$ means the computed $\widehat{\mathbf{U}}$.

Remark. The classical definition takes the maximum over **every** intermediate matrix $\mathbf{A}^{(0)} = \mathbf{A}, \mathbf{A}^{(1)}, \dots, \mathbf{A}^{(n-1)} = \mathbf{U}$ of the row-operation ordering, not only over $\mathbf{U}$. That quantity is at least g_n, so any upper bound proved for it bounds g_n too, and it is what the growth theorem of the next subsection actually estimates. The ratio above is the form the backward error bound needs.

Suppose a pivoting rule keeps every multiplier at most 1 in modulus — the next subsection exhibits one, for the algorithm above as well as for Chapter 3's ordering. Then $|\widehat{\ell}_{ip}| \le 1$, so for all i, j

$$\left(|\widehat{\mathbf{L}}||\widehat{\mathbf{U}}|\right)_{ij} \le \sum_p |\widehat{u}_{pj}| \le n \max_{p,q}|\widehat{u}_{pq}| = n\, g_n \max_{p,q}|a_{pq}|,$$

and Theorem 24.2.6 becomes

$$\max_{i,j}|\Delta a_{ij}| \;\le\; n\,\gamma_n\, g_n\, \max_{i,j}|a_{ij}|. \tag{24.2.1}$$

Since γ_n is about nu, the relative backward error is about $n^2 g_n u$. The dimensional factor n^2 is pessimistic and nobody minds it. **Everything therefore depends on** g_n — and on nothing else.

24.2.6 Pivoting

Proposition 24.2.7. Partial Pivoting Bounds the Multipliers

Run elimination with **partial pivoting**: at step k, before eliminating, exchange row k with a row $r \ge k$ for which $|a_{rk}^{(k-1)}|$ is largest. Then every multiplier satisfies $|\ell_{ik}| \le 1$, in exact arithmetic; and in the model $|\widehat{\ell}_{ik}| \le 1$ as well.

Proof. After the exchange, the pivot $a_{kk}^{(k-1)}$ has the largest modulus in its column among rows $k, \dots, n$. If it is zero then the whole column is zero, every multiplier is set to 0, and there is nothing to prove. Otherwise $|a_{ik}^{(k-1)}| \le |a_{kk}^{(k-1)}|$ for $i > k$, so $|\ell_{ik}| = |a_{ik}^{(k-1)}/a_{kk}^{(k-1)}| \le 1$.

In the model, $\widehat{\ell}_{ik} = \mathrm{fl}(x)$ for a real x with $|x| \le 1$. Both 1 and -1 lie in $\mathcal{R}$, by Definition 24.1.1, and the one with the sign of x is at distance $1 - |x|$ from x; since $\mathrm{fl}(x)$ is a **nearest** element of $\mathcal{R}$, also $|\mathrm{fl}(x) - x| \le 1 - |x|$. If $|\mathrm{fl}(x)|$ were greater than 1 we would get $|\mathrm{fl}(x) - x| \ge$

$|\mathrm{fl}(x)| - |x| > 1 - |x|$, a contradiction. So $|\widehat{\ell}_{ik}| \le 1$.

■

Remark. **The same rule, one entry at a time.** The proposition is stated for Chapter 3's ordering, which has the column entries $a_{rk}^{(k-1)}$ in hand at step k. In the entry-at-a-time algorithm those entries appear as the candidate numerators $a_{ri} - \sum_{k<i} \widehat{\ell}_{rk}\widehat{u}_{ki}$ with $r \ge i$, and partial pivoting there means: form all of them, exchange rows i and r so that the largest of them in modulus becomes $\widehat{u}_{ii}$, and only **then** divide. The proof above applies word for word, with the candidate numerator in place of $a_{rk}^{(k-1)}$ — it uses nothing about those numbers except that the divisor is the largest of them in modulus — so every multiplier of that ordering satisfies $|\widehat{\ell}_{ri}| \le 1$ too. That is what licenses the hypothesis under which Eq. 24.2.1 was derived for the computed factors of Theorem 24.2.6.

Theorem 24.2.8. Growth Under Partial Pivoting

In exact arithmetic, elimination with partial pivoting on any $\mathbf{A} \in M_n(\mathbb{R})$ has growth factor $g_n \le 2^{n-1}$.

Idea. One step replaces an entry by itself minus a multiple, of modulus at most 1, of another entry. Two numbers of modulus at most M cannot make more than $2M$. Repeat $n-1$ times.

Proof. Write $M_k = \max_{i,j}|a_{ij}^{(k)}|$, and let $\mathbf{B} = (b_{ij})$ be $\mathbf{A}^{(k-1)}$ after the row exchange of step k; an exchange only permutes entries, so $\max_{i,j}|b_{ij}| = M_{k-1}$. Every entry of $\mathbf{A}^{(k)}$ is either an entry of $\mathbf{B}$ (the rows $1, \dots, k$ are left alone, and the entries below the pivot are set to 0) or of the form

$$a_{ij}^{(k)} = b_{ij} - \ell_{ik} b_{kj}, \qquad |a_{ij}^{(k)}| \le |b_{ij}| + |\ell_{ik}||b_{kj}| \le 2M_{k-1},$$

using $|\ell_{ik}| \le 1$ from Proposition 24.2.7. Hence $M_k \le 2M_{k-1}$ for every k, and by induction $M_k \le 2^k M_0$. The intermediate matrices are $\mathbf{A}^{(0)}, \dots, \mathbf{A}^{(n-1)}$, so

$$\max_{i,j,k}|a_{ij}^{(k)}| = \max_{0 \le k \le n-1} M_k \le 2^{n-1} M_0,$$

which is $g_n \le 2^{n-1}$. This proves the theorem.

■

Remark. One expects the same argument in the model to give $M_k \le 2(1+u)^2 M_{k-1}$, the subtraction and the multiplication each contributing a factor $1+\delta$, and hence $g_n \le \big(2(1+u)^2\big)^{n-1}$, which for any u worth using is 2^{n-1} in all but the last digits. **We have not written that induction out**, and nothing in this section cites it: Theorem 24.2.8, the exact-arithmetic statement, is the only growth bound we have, and the passages that carry it over to a computed factorization say that they are doing so.

Quick check. Run elimination on $\left(\begin{smallmatrix}1 & 2\\ 3 & 4\end{smallmatrix}\right)$ without row swaps, and again with partial pivoting, and compute g_2 each time. Does partial pivoting make the growth factor smaller?

Solution. Without a swap the multiplier is 3 and $u_{22} = 4 - 3 \cdot 2 = -2$, so $\mathbf{U} = \left(\begin{smallmatrix}1 & 2\\ 0 & -2\end{smallmatrix}\right)$ and $g_2 = 2/4 = \frac{1}{2}$. With partial pivoting the rows are exchanged, the multiplier is $\frac{1}{3}$, and $u_{22} = 2 - \frac{1}{3} \cdot 4 = \frac{2}{3}$, so $\mathbf{U} = \left(\begin{smallmatrix}3 & 4\\ 0 & 2/3\end{smallmatrix}\right)$ and $g_2 = 4/4 = 1$. Pivoting made the growth factor *larger*, and $g_n < 1$ is perfectly possible. Partial pivoting does not minimize growth; it bounds the multipliers, and Theorem 24.2.8 turns that into a worst-case bound — here $2^1 = 2$, which both runs respect.

The bound 2^{n-1} is dreadful, and it is attained.

Example. A matrix that really grows

Let $\mathbf{Y}_n \in M_n(\mathbb{R})$ have 1 on the diagonal, -1 strictly below it, 0 strictly above it except in the last column, and 1 throughout the last column. For $n = 4$,

$$\mathbf{Y}_4 = \begin{pmatrix} 1 & 0 & 0 & 1 \\ -1 & 1 & 0 & 1 \\ -1 & -1 & 1 & 1 \\ -1 & -1 & -1 & 1 \end{pmatrix}.$$

Run elimination with partial pivoting on $\mathbf{Y}_4$ and on $\mathbf{Y}_5$, and compute the growth factor of each.

Solution. The runs below are in exact arithmetic, so the factors are Ł and $\mathbf{U}$, with no hats. Every entry of $\mathbf{Y}_n$ has modulus 1, so g_n is simply the largest modulus of an entry of the resulting $\mathbf{U}$.

Step 1 on $\mathbf{Y}_4$. Column 1 is $(1, -1, -1, -1)$; all four entries have modulus 1, so the largest is already in place and no exchange is made. The multipliers are -1, so each lower row has row 1 **added** to it. Rows $2, 3, 4$ become $(0, 1, 0, 2)$, $(0, -1, 1, 2)$, $(0, -1, -1, 2)$: the last column has doubled.

Step 2. Column 2 below the diagonal is $(-1, -1)$ against the pivot 1; again no exchange, multipliers -1. Rows $3, 4$ become $(0, 0, 1, 4)$ and $(0, 0, -1, 4)$.

Step 3. Multiplier -1 again; row 4 becomes $(0, 0, 0, 8)$. So

$$\mathbf{U} = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 1 & 4 \\ 0 & 0 & 0 & 8 \end{pmatrix}, \qquad g_4 = \frac{8}{1} = 8 = 2^3.$$

For $n = 5$ the same three observations repeat once more: no exchange is ever made, every multiplier is -1, and the last column runs $1, 2, 4, 8, 16$ down the diagonal of $\mathbf{U}$, giving $g_5 = 16 = 2^4$.

Every entry of $\mathbf{Y}_n$ is ± 1, and partial pivoting does exactly what it is told; for these two sizes the bound of Theorem 24.2.8 is attained, not approached. The same three observations run by induction for every n, and Exercise (C2: The growth bound is attained for every n) carries that induction out.

Warning. **Partial pivoting is not proved stable by anything in this chapter.** Eq. 24.2.1 bounds the backward error by $n\gamma_n g_n \max_{i,j}|a_{ij}|$, where g_n is the growth in the **computed** factors, and the only bound we have on growth is Theorem 24.2.8's 2^{n-1}, for the **exact** elimination; the remark after Theorem 24.2.6 records that gap and we have not closed it. Grant it anyway, and the resulting backward error bound is about $n^2 2^{n-1} u$, which for $n = 60$ exceeds 1 and says nothing at all — so even the optimistic reading proves nothing. What is true in practice is that g_n is almost always a small multiple of 1 — but "almost always" is a statement about the matrices people meet, not a theorem, and Example (A matrix that really grows), with Exercise (C2: The growth bound is attained for every n) for general n, shows there is no theorem to be had. Partial pivoting is used because it is cheap and because its failures are rare and constructed, not because it is backward stable in the sense of Definition 24.1.3. That is an honest summary of the state of the subject, and the reader should not accept a tidier one.

24.2.7 The example Chapter 3 promised

Chapter 3 §06 ended its discussion of pivoting with a two-by-two matrix and a promise: after describing how $\left(\begin{smallmatrix}\varepsilon & 1\\ 1 & 1\end{smallmatrix}\right)$ makes elimination produce a computed $\text{Ł}\mathbf{U}$ whose $(2,2)$ entry is 0 instead of 1, it wrote "Swapping the rows first gives the harmless multiplier ε. Chapter 24 makes this precise." Here it is, made precise.

Example. The tiny pivot, in the model

Work in the six-digit decimal model of Example (The model we compute in), where $u = 5 \times 10^{-6}$, and take

$$\mathbf{A} = \begin{pmatrix} \varepsilon & 1 \\ 1 & 1 \end{pmatrix}, \qquad \mathbf{b} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \qquad \varepsilon = 10^{-8}.$$

Factor and solve without an exchange, then with one. In each case compute $\widehat{\text{Ł}}\widehat{\mathbf{U}}$ exactly and report the backward error.

Solution. The exact answer. Subtracting the first equation's ε-multiple, $(1-\varepsilon)x_1 = 1$, so

$$\mathbf{x} = \left(\frac{1}{1-\varepsilon}, \frac{1-2\varepsilon}{1-\varepsilon}\right) = (1.00000001\ldots,\ 0.99999999\ldots).$$

Also $\kappa_\infty(\mathbf{A}) = 4/(1-\varepsilon) < 5$, exactly as in Chapter 16 §08: the problem is well conditioned.

Without the exchange. The multiplier is $\text{fl}(1/\varepsilon) = 10^8$, and

$$\widehat{u}_{22} = \text{fl}(1 - \text{fl}(10^8 \cdot 1)) = \text{fl}(1 - 10^8) = \text{fl}(-99999999) = -10^8,$$

the last step by rounding to six digits. Forward substitution gives $\widehat{y}_1 = 1$ and $\widehat{y}_2 = \text{fl}(2 -$

$10^8) = -10^8$; back substitution gives $\widehat{x}_2 = \mathrm{fl}(-10^8/-10^8) = 1$ and then

$$\widehat{x}_1 = \mathrm{fl}((1-1)/\varepsilon) = 0.$$

The computed answer is $(0,1)$ — Chapter 16 §08's $(0,1)$, reproduced here in a model rather than by hand-waving about digits. Multiplying the computed factors *exactly*,

$$\begin{aligned}\widehat{\mathrm{L}}\widehat{\mathrm{U}} &= \begin{pmatrix} 1 & 0 \\ 10^8 & 1 \end{pmatrix}\begin{pmatrix} 10^{-8} & 1 \\ 0 & -10^8 \end{pmatrix} \\ &= \begin{pmatrix} 10^{-8} & 1 \\ 1 & 0 \end{pmatrix} = \mathbf{A} + \Delta\mathbf{A}, \qquad \Delta\mathbf{A} = \begin{pmatrix} 0 & 0 \\ 0 & -1 \end{pmatrix}.\end{aligned}$$

So $\|\Delta\mathbf{A}\|_\infty/\|\mathbf{A}\|_\infty = \frac{1}{2}$. The backward error is not $O(u)$; it is $O(1)$. This factorization is **not** backward stable, and it is not even close.

With the exchange. Partial pivoting compares $|\varepsilon|$ with 1 and swaps, so we factor $\mathbf{PA} = \left(\begin{smallmatrix} 1 & 1 \\ \varepsilon & 1 \end{smallmatrix}\right)$ with $\mathbf{Pb} = (2,1)$. The multiplier is $\mathrm{fl}(\varepsilon/1) = \varepsilon$, and $\widehat{u}_{22} = \mathrm{fl}(1 - \mathrm{fl}(\varepsilon \cdot 1)) = \mathrm{fl}(1 - 10^{-8}) = 1$, since $1 - 10^{-8}$ lies between the representable neighbors 0.999999 and 1 and is far nearer the second. Forward substitution gives $\widehat{y}_1 = 2$ and $\widehat{y}_2 = \mathrm{fl}(1 - \mathrm{fl}(2\varepsilon)) = \mathrm{fl}(1 - 2 \times 10^{-8}) = 1$; back substitution gives $\widehat{x}_2 = 1$ and $\widehat{x}_1 = \mathrm{fl}((2-1)/1) = 1$. The computed answer is $(1,1)$, whose relative error against the exact $\mathbf{x}$ is about 10^{-8}, far *better* than u. Multiplying the computed factors exactly,

$$\begin{aligned}\widehat{\mathrm{L}}\widehat{\mathrm{U}} &= \begin{pmatrix} 1 & 0 \\ 10^{-8} & 1 \end{pmatrix}\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} 1 & 1 \\ 10^{-8} & 1 + 10^{-8} \end{pmatrix} = \mathbf{PA} + \Delta\mathbf{A}, \qquad \Delta\mathbf{A} = \begin{pmatrix} 0 & 0 \\ 0 & 10^{-8} \end{pmatrix},\end{aligned}$$

so $\|\Delta\mathbf{A}\|_\infty/\|\mathbf{A}\|_\infty = 5 \times 10^{-9}$, comfortably below u. This factorization is backward stable.

What the exchange achieved. Exactly what Eq. 24.2.1 predicts. Without it the growth factor is $g_2 = 10^8$: an entry of modulus 10^8 is created out of a matrix whose largest entry is 1, and creating it destroys the information in $a_{22} = 1$, which lies beneath its last digit. With the exchange, Proposition 24.2.7 keeps the multiplier at $\varepsilon \le 1$, $g_2 = 1$, and nothing large is ever made. Note what was **not** the problem: $\kappa_\infty(\mathbf{A}) < 5$ throughout. Chapter 16 §08 said so: "Conditioning bounds what *any* method can promise; whether a given method keeps that promise is stability."

24.2.8 Cholesky needs no pivoting

For a positive definite matrix the whole difficulty disappears, and the reason is a one-line inequality about the entries of the factor. Chapter 13 §02 asserted it — "it needs no pivoting: the recursion above never divides by something that positive definiteness allows to be zero" — and left the arithmetic half here.

Theorem 24.2.9. Cholesky Does Not Grow

Let $\mathbf{A} \in M_n(\mathbb{R})$ be symmetric positive definite with representable entries, and let L be its Cholesky factor, $\mathbf{A} = \mathrm{L}\mathrm{L}^\top$ with $\ell_{ii} > 0$ (Theorem 13.2.3).

(a) $|\ell_{ij}| \le \sqrt{a_{ii}}$ for all $j \le i$, and

$$\left(|Ł||Ł^{\top}|\right)_{ik} \le \sqrt{a_{ii}a_{kk}} \le \max_{p,q}|a_{pq}| \qquad \text{for all } i,k.$$

(b) Suppose $(n+1)u < 1$ and the recursion of Chapter 13 §02 runs to completion in the model, producing $\widehat{Ł}$. Then

$$\widehat{Ł}\widehat{Ł}^{\top} = \mathbf{A} + \Delta\mathbf{A}, \qquad |\Delta\mathbf{A}| \le \gamma_{n+1}\,|\widehat{Ł}|\,|\widehat{Ł}^{\top}|.$$

(c) If moreover $\gamma_{n+1} < 1$, then

$$\max_{i,j}|\Delta a_{ij}| \le \frac{\gamma_{n+1}}{1-\gamma_{n+1}} \max_{i,j}|a_{ij}|.$$

Idea. For (a): the rows of Ł have known lengths, because $(ŁŁ^{\top})_{ii} = a_{ii}$ *is* the squared length of row i. An entry is no longer than its row, and Cauchy–Schwarz compares two rows. For (b): each off-diagonal entry is Lemma 24.2.4 again, and the diagonal entry is the same lemma followed by a square root, which contributes two more factors. For (c): part (b) bounds the row lengths of $\widehat{Ł}$ in terms of $\mathbf{A}$, and then (a)'s argument runs again on $\widehat{Ł}$.

Proof.

(a) Row i of Ł has $\sum_p \ell_{ip}^2 = (ŁŁ^{\top})_{ii} = a_{ii}$, so every single term satisfies $\ell_{ij}^2 \le a_{ii}$. By the Cauchy–Schwarz inequality (Theorem 11.1.5) applied to the rows i and k of $|Ł|$,

$$\left(|Ł||Ł^{\top}|\right)_{ik} = \sum_p |\ell_{ip}||\ell_{kp}| \le \Big(\sum_p \ell_{ip}^2\Big)^{1/2}\Big(\sum_p \ell_{kp}^2\Big)^{1/2} = \sqrt{a_{ii}a_{kk}}.$$

Finally $\sqrt{a_{ii}a_{kk}} \le \max_p a_{pp}$, and $\max_p a_{pp} = \max_{p,q}|a_{pq}|$: the diagonal entries are positive (take $\mathbf{x} = \mathbf{e}_p$ in Definition 13.1.1), and $|a_{pq}| = |(ŁŁ^{\top})_{pq}| \le \left(|Ł||Ł^{\top}|\right)_{pq} \le \sqrt{a_{pp}a_{qq}} \le \max_p a_{pp}$, the middle step by the display just proved.

(b) Fix $i > j$. The computation $\widehat{\ell}_{ij} = \mathrm{fl}\big((a_{ij} - \sum_{k<j}\widehat{\ell}_{ik}\widehat{\ell}_{jk})/\widehat{\ell}_{jj}\big)$ is Lemma 24.2.4 with $m = j-1$ and $d = \widehat{\ell}_{jj}$, so, writing θ_j for θ_0,

$$a_{ij} = \sum_{p=1}^{j} \widehat{\ell}_{ip}\widehat{\ell}_{jp}(1+\theta_p), \qquad |\theta_p| \le \gamma_j \le \gamma_{n+1}.$$

Since $\widehat{\ell}_{jp} = 0$ for $p > j$, the sum without the θ_p is $(\widehat{Ł}\widehat{Ł}^{\top})_{ij}$, so

$$|a_{ij} - (\widehat{Ł}\widehat{Ł}^{\top})_{ij}| \le \gamma_{n+1}\sum_{p\le j}|\widehat{\ell}_{ip}||\widehat{\ell}_{jp}| = \gamma_{n+1}\left(|\widehat{Ł}||\widehat{Ł}^{\top}|\right)_{ij}.$$

For $i = j$, let $\widehat{s}$ be the value accumulated before the square root, so $\widehat{\ell}_{jj} = \mathrm{fl}(\sqrt{\widehat{s}}) = \sqrt{\widehat{s}}(1+\eta)$ with $|\eta| \le u$, and $\widehat{s} = \widehat{\ell}_{jj}^2(1+\eta)^{-2}$. By the **proof** of Lemma 24.2.4 with $m = j-1$ and no division — we need not just the size of θ_0' but its shape, and the proof produces $1+\theta_0'$ as the product of the $j-1$ factors $(1+\varepsilon_p)^{-1}$ —

$$a_{jj} = \widehat{s}\,(1+\theta_0') + \sum_{p<j}\widehat{\ell}_{jp}^2(1+\theta_p), \qquad |\theta_0'|, |\theta_p| \le \gamma_{j-1}.$$

Substituting $\widehat{s}$, the first term is $\widehat{\ell}_{jj}^{\,2}$ times a product of $(j-1)+2 \le n+1$ factors of the form $(1+\delta)^{\pm 1}$, which by Lemma 24.1.1 is $1+\theta_j$ with $|\theta_j| \le \gamma_{n+1}$. So again $a_{jj} = \sum_{p \le j} \widehat{\ell}_{jp}\widehat{\ell}_{jp}(1+\theta_p)$ with all $|\theta_p| \le \gamma_{n+1}$, and the displayed estimate follows as before. Since $\widehat{\mathrm{Ł}}$ is lower triangular, the entries with $i < j$ are the transposes of these, and $\Delta\mathbf{A}$ is symmetric.

(c) The diagonal case of (b) reads $|a_{ii} - (\widehat{\mathrm{Ł}}\widehat{\mathrm{Ł}}^{\top})_{ii}| \le \gamma_{n+1}(\widehat{\mathrm{Ł}}\widehat{\mathrm{Ł}}^{\top})_{ii}$, because $(|\widehat{\mathrm{Ł}}||\widehat{\mathrm{Ł}}^{\top}|)_{ii} = \sum_p \widehat{\ell}_{ip}^{\,2} = (\widehat{\mathrm{Ł}}\widehat{\mathrm{Ł}}^{\top})_{ii}$. Hence $(\widehat{\mathrm{Ł}}\widehat{\mathrm{Ł}}^{\top})_{ii}(1-\gamma_{n+1}) \le a_{ii}$, that is

$$\sum_p \widehat{\ell}_{ip}^{\,2} \le \frac{a_{ii}}{1-\gamma_{n+1}}.$$

Cauchy–Schwarz as in (a), now for $\widehat{\mathrm{Ł}}$, gives $(|\widehat{\mathrm{Ł}}||\widehat{\mathrm{Ł}}^{\top}|)_{ik} \le \sqrt{a_{ii}a_{kk}}/(1-\gamma_{n+1}) \le \max_{p,q}|a_{pq}|/(1-\gamma_{n+1})$. Substituting into (b) gives (c). This proves the theorem.

■

Compare (c) with Eq. 24.2.1. For a general matrix the backward error carried the factor g_n, which partial pivoting bounds only by 2^{n-1}, sharply. Here the same quantity is bounded by the largest entry of $\mathbf{A}$ itself, with **no growth factor and no pivoting**: the Cholesky recursion is unconditionally backward stable. The reason sits entirely in part (a) — the rows of the factor cannot be longer than the diagonal of $\mathbf{A}$ allows — and that is a fact about positive definiteness, not about arithmetic. Chapter 13 §02 called positive definiteness "the one structural hypothesis under which elimination is unconditionally safe"; this is the proof.

Warning. **Part (a) bounds $|\ell_{ij}|$ by $\sqrt{a_{ii}}$, the diagonal entry of its own row, not of its column.** The column version is false. For

$$\mathbf{A} = \begin{pmatrix} 0.01 & 0.09 \\ 0.09 & 100 \end{pmatrix},$$

which is positive definite since $0.01 > 0$ and $\det \mathbf{A} = 1 - 0.0081 > 0$ (Theorem 13.1.5 (d)), the factor is $\ell_{11} = 0.1$, $\ell_{21} = 0.9$, and $|\ell_{21}| = 0.9$ is nine times $\sqrt{a_{11}} = 0.1$. It is at most $\sqrt{a_{22}} = 10$, as (a) requires. An entry of Ł can be far larger than the pivot it was divided by; what it cannot exceed is the size of its own row.

24.2.9 Exercises

24.2.9.1 A. Check your understanding

Exercise. A1

(a) Define a flop and the growth factor g_n.

(b) State the partial pivoting rule, and say what Proposition 24.2.7 gets from it.

(c) State the leading term of the flop count for an LU factorization, for two triangular solves and for Cholesky.

(d) Decide whether each statement is correct, with a reason: (i) "a triangular solve needs pivoting when the diagonal entries are small"; (ii) "elimination with partial pivoting is backward stable"; (iii) "a backward stable factorization gives an accurate solution".

(e) In one sentence, what does Theorem 24.2.1 say about the trailing block after k steps?

Solution.

(a) A flop is one arithmetic operation — an addition, subtraction, multiplication, division, or square root. The growth factor is $g_n = \max_{i,j}|u_{ij}| / \max_{i,j}|a_{ij}|$, the largest entry in modulus the elimination produces relative to the largest in modulus it was given (Definition 24.2.1); the classical version maximizes over every intermediate matrix as well, and is at least as large.

(b) At step k, exchange row k with the row $r \geq k$ maximizing $|a_{rk}^{(k-1)}|$. The consequence is $|\ell_{ik}| \leq 1$ for every multiplier, in exact arithmetic and in the model.

(c) $\frac{2}{3}n^3$, $2n^2$, and $\frac{1}{3}n^3$ respectively (Proposition 24.2.3); the third is half the first.

(d) (i) Incorrect. Theorem 24.2.5 holds for every nonsingular triangular matrix, with no hypothesis on the diagonal and no pivoting. (ii) Incorrect as stated. The only growth bound available is Theorem 24.2.8's $g_n \leq 2^{n-1}$, proved for the exact elimination, and Example (A matrix that really grows) shows it is attained; even granting its transfer to the computed factors — the gap recorded in the remark after Theorem 24.2.6 — the resulting backward error bound $n^2 2^{n-1} u$ is vacuous for moderate n. (iii) Incorrect. It gives an accurate solution only when the problem is well conditioned; Theorem 24.1.3 supplies the missing factor $\kappa(\mathbf{A})$.

(e) It is the Schur complement $\mathbf{A}/\mathbf{A}_k$ of the leading $k \times k$ block — so elimination is the repeated formation of Schur complements of 1×1 pivots.

24.2.9.2 B. Practice

Exercise. B1: A Schur complement by two routes

Let

$$\mathbf{A} = \begin{pmatrix} 2 & 1 & -1 \\ 4 & 5 & -1 \\ -2 & 8 & 2 \end{pmatrix}.$$

(a) Carry out one step of elimination without row swaps and read off the trailing 2×2 block.

(b) Compute $\mathbf{A}/\mathbf{A}_1 = \mathbf{D} - \mathbf{C}a_{11}^{-1}\mathbf{B}$ directly from Definition 8.3.1 and check that the two agree.

(c) Carry out the second step and verify Corollary 24.2.2 (a) by computing $\det \mathbf{A}_1$, $\det \mathbf{A}_2$, and $\det \mathbf{A}_3$.

Solution.

(a) The multipliers are $\ell_{21} = 4/2 = 2$ and $\ell_{31} = -2/2 = -1$. Subtracting gives rows $(0, 3, 1)$ and $(0, 9, 1)$, so the trailing block is $\left(\begin{smallmatrix} 3 & 1 \\ 9 & 1 \end{smallmatrix}\right)$.

(b) Here $\mathbf{A}_1 = (2)$, $\mathbf{B} = (1, -1)$, $\mathbf{C} = (4, -2)^\top$, $\mathbf{D} = \left(\begin{smallmatrix} 5 & -1 \\ 8 & 2 \end{smallmatrix}\right)$, so

$$\mathbf{C}a_{11}^{-1}\mathbf{B} = \tfrac{1}{2}\begin{pmatrix} 4 \\ -2 \end{pmatrix}\begin{pmatrix} 1 & -1 \end{pmatrix} = \begin{pmatrix} 2 & -2 \\ -1 & 1 \end{pmatrix}, \qquad \mathbf{A}/\mathbf{A}_1 = \begin{pmatrix} 3 & 1 \\ 9 & 1 \end{pmatrix},$$

which agrees with (a), as Theorem 24.2.1 says it must.

(c) The second multiplier is $9/3 = 3$, and the second step turns $(0, 9, 1)$ into $(0, 0, 1-3) = (0, 0, -2)$. So the pivots are $2, 3, -2$. Now $\det \mathbf{A}_1 = 2$, $\det \mathbf{A}_2 = 10 - 4 = 6$, and $\det \mathbf{A} = 2(10+8) - 1(8-2) + (-1)(32+10) = 36 - 6 - 42 = -12$. The ratios are $2/1 = 2$, $6/2 = 3$, and $-12/6 = -2$: the pivots, as claimed.

Exercise. B2: Growth, with and without a swap

Let $\mathbf{A} = \begin{pmatrix} 10^{-4} & 1 \\ 1 & 1 \end{pmatrix}$, and work in the six-digit decimal model.

(a) Factor $\mathbf{A}$ without an exchange, compute g_2, and compute $\widehat{Ł}\widehat{\mathbf{U}}$ exactly.

(b) Do the same after exchanging the rows.

(c) Compare the two backward errors with the bound of Eq. 24.2.1.

Solution.

(a) The multiplier is $\mathrm{fl}(1/10^{-4}) = 10^4$, and $\widehat{u}_{22} = \mathrm{fl}(1 - 10^4) = -9999.00$, which is exactly representable. So

$$\widehat{Ł}\widehat{\mathbf{U}} = \begin{pmatrix} 1 & 0 \\ 10^4 & 1 \end{pmatrix} \begin{pmatrix} 10^{-4} & 1 \\ 0 & -9999 \end{pmatrix} = \begin{pmatrix} 10^{-4} & 1 \\ 1 & 1 \end{pmatrix} = \mathbf{A},$$

so here $\Delta\mathbf{A} = \mathbf{O}$: with six digits, $1 - 10^4$ is *not* rounded, and nothing is lost. The growth factor is $g_2 = 9999/1 = 9999$.

(b) After the exchange the multiplier is 10^{-4} and $\widehat{u}_{22} = \mathrm{fl}(1 - 10^{-4}) = 0.999900$, again exact, so $\widehat{Ł}\widehat{\mathbf{U}} = \mathbf{PA}$ exactly and $g_2 = 1$.

(c) Both backward errors are zero, so both satisfy the bound. For $n = 2$ the entry-at-a-time algorithm and Chapter 3's row operations perform the very same three operations, so Eq. 24.2.1 applies to these runs as it stands, with g_2 read off the computed factors. The point is the *bound*, not the outcome: in (a) it is $2\gamma_2 g_2 \approx 4 \times 9999\, u \approx 0.2$, which permits a catastrophe, while in (b) it is about $4u$, which does not. Large growth does not force a large backward error; it removes the guarantee that there is none, and Example (The tiny pivot, in the model) shows what happens when the permission is taken up.

Exercise. B3: The Cholesky entry bound

Let $\mathbf{A} = \begin{pmatrix} 4 & -2 & 6 \\ -2 & 5 & -1 \\ 6 & -1 & 26 \end{pmatrix}$, whose Cholesky factor was computed in Example (A Cholesky Factorization, and One That Fails) to be Ł with rows $(2, 0, 0)$, $(-1, 2, 0)$, $(3, 1, 4)$.

(a) Check Theorem 24.2.9 (a) entry by entry.

(b) Compute $|Ł||Ł^{\top}|$ and check that every entry is at most $\max_{p,q} |a_{pq}|$.

(c) What would the bound of Eq. 24.2.1 require if $\mathbf{A}$ were factored by general elimination with partial pivoting instead?

Solution.

(a) The bound is $|\ell_{ij}| \le \sqrt{a_{ii}}$. Row 1: $2 \le \sqrt{4} = 2$, with equality. Row 2: $1, 2 \le \sqrt{5} \approx 2.236$. Row 3: $3, 1, 4 \le \sqrt{26} \approx 5.099$. All hold, and the row sums of squares are $4, 5, 26$, which are a_{11}, a_{22}, a_{33} exactly.

(b) $|Ł|$ has rows $(2,0,0)$, $(1,2,0)$, $(3,1,4)$, so

$$|Ł||Ł^\top| = \begin{pmatrix} 4 & 2 & 6 \\ 2 & 5 & 5 \\ 6 & 5 & 26 \end{pmatrix}.$$

The largest entry of $\mathbf{A}$ in modulus is 26, and every entry above is at most 26 — indeed the (i,k) entry is at most $\sqrt{a_{ii}a_{kk}}$: for instance $5 \le \sqrt{5 \cdot 26} \approx 11.4$.

(c) It would require knowing g_3 for the computed factors, and the only bound available is Theorem 24.2.8's exact-arithmetic $2^2 = 4$; granting the transfer between the two (the remark after Theorem 24.2.6), Eq. 24.2.1 permits $\max|\Delta a_{ij}| \le 3\gamma_3 \cdot 4 \cdot 26 = 12\gamma_3 \cdot 26$. Theorem 24.2.9 (c) permits $(\gamma_4/(1-\gamma_4)) \cdot 26$ instead: the factor $3g_3 = 12$ disappears entirely, and the price is the step from γ_3 to $\gamma_4/(1-\gamma_4)$, which is about $\frac{4}{3}\gamma_3$. No pivoting is needed to get it.

24.2.9.3 C. Going deeper

Exercise. C1: Column diagonal dominance needs no pivoting

Call $\mathbf{A} \in M_n(\mathbb{R})$ **strictly column diagonally dominant** if $|a_{jj}| > \sum_{i \ne j}|a_{ij}|$ for every j.

(a) Show that partial pivoting makes no exchange at the first step.

(b) Show that $\mathbf{A}/\mathbf{A}_1$ is again strictly column diagonally dominant, and that each of its column sums $\sum_{i \ge 2}|a'_{ij}|$ is at most the corresponding column sum $\sum_{i \ge 1}|a_{ij}|$ of $\mathbf{A}$.

(c) Deduce that elimination without any exchanges succeeds and that $g_n < 2$.

Hint: for (c), bound every entry of every $\mathbf{A}^{(k)}$ by a column sum of $\mathbf{A}$.

Solution.

(a) Column 1 satisfies $|a_{11}| > \sum_{i \ge 2}|a_{i1}| \ge |a_{i1}|$ for each $i \ge 2$, so the entry of largest modulus in column 1 is already a_{11}, and it is nonzero.

(b) Write $a'_{ij} = a_{ij} - (a_{i1}/a_{11})a_{1j}$ for $i, j \ge 2$. For the column sums, the triangle inequality gives

$$\sum_{i \ge 2}|a'_{ij}| \le \sum_{i \ge 2}|a_{ij}| + \frac{|a_{1j}|}{|a_{11}|}\sum_{i \ge 2}|a_{i1}| \le \sum_{i \ge 2}|a_{ij}| + |a_{1j}| = \sum_{i \ge 1}|a_{ij}|,$$

using $\sum_{i \ge 2}|a_{i1}| < |a_{11}|$. For dominance, fix $j \ge 2$. Then

$$\sum_{\substack{i \ge 2 \\ i \ne j}}|a'_{ij}| \le \sum_{\substack{i \ge 2 \\ i \ne j}}|a_{ij}| + \frac{|a_{1j}|}{|a_{11}|}\sum_{\substack{i \ge 2 \\ i \ne j}}|a_{i1}| < (|a_{jj}| - |a_{1j}|) + \frac{|a_{1j}|}{|a_{11}|}(|a_{11}| - |a_{j1}|),$$

where the two strict bounds are column-j and column-1 dominance. The right side equals $|a_{jj}| - |a_{1j}||a_{j1}|/|a_{11}|$, which is at most $|a_{jj} - a_{j1}a_{1j}/a_{11}| = |a'_{jj}|$ by the reverse

triangle inequality. So $\sum_{i\geq 2, i\neq j}|a'_{ij}| < |a'_{jj}|$.

(c) By (a) and (b) and induction, every Schur complement $\mathbf{A}^{(k)}$ restricted to its active block is strictly column diagonally dominant, so no exchange is ever made and no pivot is zero. By (b) and induction, the column sums of the active blocks never increase, so every entry of every $\mathbf{A}^{(k)}$ is bounded by $\max_j \sum_i |a_{ij}|$ — and the rows that have been frozen into $\mathbf{U}$ were bounded the same way when they were created. Finally, dominance gives $\sum_i |a_{ij}| < 2|a_{jj}| \leq 2\max_{p,q}|a_{pq}|$, so $g_n < 2$. Hence such matrices need no pivoting at all.

Exercise. C2: The growth bound is attained for every n

Let $\mathbf{Y}_n$ be the matrix of Example (A matrix that really grows): $y_{ii} = 1$, $y_{ij} = -1$ for $i > j$, $y_{in} = 1$ for all i, and $y_{ij} = 0$ for $j > i$ with $j < n$.

(a) Prove that partial pivoting makes no exchange at any step, and that every multiplier equals -1.

(b) Prove by induction on k that after k steps the active trailing block is again of the form $\mathbf{Y}_{n-k}$, except that its last column is 2^k times the all-ones vector.

(c) Deduce $g_n = 2^{n-1}$, and explain why this does not contradict the practical success of partial pivoting.

Solution.

(a) and (b) together, by induction. At step $k = 1$, column 1 is $(1, -1, \ldots, -1)$, every entry of modulus 1, so the pivot $y_{11} = 1$ is already of largest modulus and no exchange is made; the multipliers are $-1/1 = -1$. Subtracting -1 times row 1 from row i adds row 1 to row i. Row 1 is $(1, 0, \ldots, 0, 1)$, so for $i \geq 2$ the entries in columns $2, \ldots, n-1$ are unchanged, the entry in column 1 becomes 0, and the entry in column n becomes $1+1 = 2$. The trailing $(n-1)\times(n-1)$ block therefore has 1 on its diagonal, -1 below, 0 above except in the last column, and 2 throughout its last column: it is $\mathbf{Y}_{n-1}$ with the last column doubled. Since doubling one column changes neither which entry of a column has largest modulus nor the ratios within the other columns, the same applies at every subsequent step, and after k steps the active block is $\mathbf{Y}_{n-k}$ with last column 2^k times the all-ones vector.

(b) Every entry of $\mathbf{Y}_n$ has modulus 1, so $M_0 = 1$. By (b) the largest entry ever created is the (n, n) entry after step $n-1$, namely 2^{n-1}. So $g_n = 2^{n-1}$, meeting Theorem 24.2.8 exactly.

This does not contradict anything practical: $\mathbf{Y}_n$ is a constructed matrix, and the theorem only says the *worst case* is 2^{n-1}. What the example does destroy is the hope of a theorem: no bound better than 2^{n-1} can be proved for partial pivoting, so any claim that partial pivoting is backward stable is a claim about the matrices one meets, not about all matrices. That is why Theorem 24.2.9 matters — it is a genuine theorem, for a class of matrices identified in advance.

24.3 Least Squares in Practice

Chapter 11 solved the least-squares problem twice. Section 4 of that chapter produced the normal equations $\mathbf{A}^*\mathbf{A}\mathbf{x} = \mathbf{A}^*\mathbf{b}$ (Theorem 11.4.2), and Section 8 rewrote them, once a $\mathbf{QR}$ factorization was available, as the triangular system $\mathbf{R}\mathbf{x} = \mathbf{Q}^*\mathbf{b}$. It then recommended the second route and postponed the reason: "forming $\mathbf{A}^*\mathbf{A}$ squares the sensitivity of the problem to rounding, while (†) never forms that product at all. Chapter 24 makes 'sensitivity' precise and proves the comparison; here we record only the practice."

Chapter 16 supplied half of the reason. Its Theorem 16.8.4 proved $\kappa_2(\mathbf{A}^*\mathbf{A}) = \kappa_2(\mathbf{A})^2$ and then listed what was still missing: the conditioning of the fitting problem itself, which involves the residual; the arithmetic, which needed a model; and the comparison of the two algorithms in digits. Sections 1 and 2 of this chapter built the model and the vocabulary. This section pays the three debts, in that order, and ends with the reason the reflections of Chapter 11 are the right way to produce $\mathbf{Q}$ and $\mathbf{R}$.

Throughout, $F = \mathbb{R}$ or $F = \mathbb{C}$, $\mathbf{A} \in M_{m\times n}(F)$ with $m \geq n$ and $\operatorname{rank}\mathbf{A} = n$, and $\mathbf{b} \in F^m$ is nonzero. By Corollary 11.4.4 there is then exactly one least-squares solution $\mathbf{x}$, and $\mathbf{A}^*\mathbf{A}$ is invertible. All norms are $\|\cdot\|_2$ unless a subscript says otherwise.

24.3.1 How sensitive is the problem?

Before any algorithm is judged, the problem has to be measured. Write

$$\mathbf{r} = \mathbf{b} - \mathbf{A}\mathbf{x}, \qquad \mathbf{A}\mathbf{x} = P_{\operatorname{col}(\mathbf{A})}\mathbf{b},$$

the residual and the projection of Theorem 11.4.2 (b). Since $\mathbf{r} \perp \operatorname{col}(\mathbf{A})$ and $\mathbf{A}\mathbf{x} \in \operatorname{col}(\mathbf{A})$, Pythagoras (Theorem 11.1.3) gives $\|\mathbf{b}\|^2 = \|\mathbf{A}\mathbf{x}\|^2 + \|\mathbf{r}\|^2$. So there is a unique $\theta \in [0, \pi/2]$ with

$$\cos\theta = \frac{\|\mathbf{A}\mathbf{x}\|}{\|\mathbf{b}\|}, \qquad \sin\theta = \frac{\|\mathbf{r}\|}{\|\mathbf{b}\|}, \tag{24.3.1}$$

the angle between $\mathbf{b}$ and the column space of $\mathbf{A}$. It is 0 exactly when the system is consistent and $\pi/2$ exactly when $\mathbf{A}^*\mathbf{b} = \mathbf{0}$, in which case $\mathbf{x} = \mathbf{0}$ and no relative error can be spoken of. **We assume $\theta < \pi/2$ throughout**, so $\mathbf{x} \neq \mathbf{0}$.

One number is now missing from Chapter 16's account, and Eq. 24.3.1 supplies it. Here is the sensitivity of $\mathbf{x}$ to a perturbation of the data.

Theorem 24.3.1. Conditioning of the Least-Squares Problem

Let $\mathbf{A} \in M_{m\times n}(F)$ have rank $n \leq m$, let $\mathbf{b} \neq \mathbf{0}$, let $\mathbf{x}$ be the least-squares solution of $\mathbf{A}\mathbf{z} = \mathbf{b}$, and let $\theta < \pi/2$ be as in Eq. 24.3.1. Let $\mathbf{E} \in M_{m\times n}(F)$ and $\mathbf{f} \in F^m$, and for real t near 0 let $\mathbf{x}(t)$ be the least-squares solution of $(\mathbf{A} + t\mathbf{E})\mathbf{z} = \mathbf{b} + t\mathbf{f}$. Then $\mathbf{x}(t)$ is defined and differentiable near $t = 0$, with $\mathbf{x}(0) = \mathbf{x}$, and

$$\frac{\|\mathbf{x}'(0)\|}{\|\mathbf{x}\|} \;\leq\; \kappa_2(\mathbf{A})\Big(\varepsilon_{\mathbf{A}} + \frac{\varepsilon_{\mathbf{b}}}{\cos\theta}\Big) \;+\; \kappa_2(\mathbf{A})^2\,\varepsilon_{\mathbf{A}}\tan\theta,$$

where $\varepsilon_{\mathbf{A}} = \|\mathbf{E}\|/\|\mathbf{A}\|$ and $\varepsilon_{\mathbf{b}} = \|\mathbf{f}\|/\|\mathbf{b}\|$.

Idea. The least-squares solution is $\mathbf{x}(t) = \mathbf{N}(t)^{-1}\mathbf{A}(t)^*\mathbf{b}(t)$ with $\mathbf{N}(t) = \mathbf{A}(t)^*\mathbf{A}(t)$, which is a differentiable function of t because $\mathbf{N}(0)$ is invertible. Rather than differentiate that formula, differentiate the equation $\mathbf{N}(t)\mathbf{x}(t) = \mathbf{A}(t)^*\mathbf{b}(t)$ it comes from: the product rule gives three terms, two of them combine into $\mathbf{E}^*(\mathbf{b}-\mathbf{A}\mathbf{x}) = \mathbf{E}^*\mathbf{r}$, and the residual appears by itself. Then read every norm off a singular value decomposition: $\|(\mathbf{A}^*\mathbf{A})^{-1}\| = \sigma_n^{-2}$ is where the square comes from, and the only inequality used to finish is $\sigma_1\|\mathbf{x}\| \geq \|\mathbf{A}\mathbf{x}\| = \|\mathbf{b}\|\cos\theta$.

Proof. Write $\mathbf{A}(t) = \mathbf{A} + t\mathbf{E}$, $\mathbf{b}(t) = \mathbf{b} + t\mathbf{f}$, and $\mathbf{N}(t) = \mathbf{A}(t)^*\mathbf{A}(t)$. Every entry of $\mathbf{N}(t)$ is a polynomial in t, so $\mathbf{N}$ is differentiable in the sense of Definition 16.6.1, and $\mathbf{N}(0) = \mathbf{A}^*\mathbf{A}$ is invertible by Corollary 11.4.4 (c). By Theorem 16.6.5, $\mathbf{N}(t)$ is invertible on an open interval around 0 and $t \mapsto \mathbf{N}(t)^{-1}$ is differentiable there. On that interval $\operatorname{rank}\mathbf{A}(t) = n$, again by Corollary 11.4.4, so Theorem 11.4.2 has the unique solution

$$\mathbf{x}(t) = \mathbf{N}(t)^{-1}\mathbf{A}(t)^*\mathbf{b}(t),$$

differentiable by Lemma 16.6.4. (Conjugation is applied entrywise and commutes with differentiation in the real variable t, so $(\mathbf{A}(t)^*)' = \mathbf{E}^*$.)

Differentiate the identity $\mathbf{N}(t)\mathbf{x}(t) = \mathbf{A}(t)^*\mathbf{b}(t)$ at $t = 0$ with Lemma 16.6.4. Since $\mathbf{N}'(0) = \mathbf{E}^*\mathbf{A} + \mathbf{A}^*\mathbf{E}$,

$$(\mathbf{E}^*\mathbf{A} + \mathbf{A}^*\mathbf{E})\mathbf{x} + \mathbf{A}^*\mathbf{A}\,\mathbf{x}'(0) = \mathbf{E}^*\mathbf{b} + \mathbf{A}^*\mathbf{f}.$$

Move the first bracket across and collect the two terms carrying $\mathbf{E}^*$:

$$\mathbf{A}^*\mathbf{A}\,\mathbf{x}'(0) = \mathbf{E}^*(\mathbf{b}-\mathbf{A}\mathbf{x}) - \mathbf{A}^*(\mathbf{E}\mathbf{x}) + \mathbf{A}^*\mathbf{f} = \mathbf{E}^*\mathbf{r} - \mathbf{A}^*(\mathbf{E}\mathbf{x}) + \mathbf{A}^*\mathbf{f}.$$

Hence

$$\mathbf{x}'(0) = (\mathbf{A}^*\mathbf{A})^{-1}\mathbf{E}^*\mathbf{r} \;-\; (\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^*(\mathbf{E}\mathbf{x}) \;+\; (\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^*\mathbf{f}. \tag{$*$}$$

Now the two norms. Write $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$ as in Theorem 13.9.1, with $\sigma_1 \geq \cdots \geq \sigma_n > 0$. As in the proof of Theorem 16.8.4, $\mathbf{A}^*\mathbf{A} = \mathbf{V}\operatorname{diag}(\sigma_1^2,\dots,\sigma_n^2)\mathbf{V}^*$, so

$$(\mathbf{A}^*\mathbf{A})^{-1} = \mathbf{V}\operatorname{diag}(\sigma_1^{-2},\dots,\sigma_n^{-2})\mathbf{V}^*, \qquad (\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^* = \mathbf{V}\mathbf{\Sigma}^+\mathbf{U}^*,$$

where $\mathbf{\Sigma}^+$ is the $n \times m$ matrix with $\sigma_1^{-1},\dots,\sigma_n^{-1}$ on its diagonal — the second identity because $\operatorname{diag}(\sigma_i^{-2})\mathbf{\Sigma}^* = \mathbf{\Sigma}^+$. Unitary factors do not change singular values, rectangular ones included: for unitary $\mathbf{U}_1, \mathbf{V}_1$ of the right sizes, $(\mathbf{U}_1\mathbf{M}\mathbf{V}_1)^*(\mathbf{U}_1\mathbf{M}\mathbf{V}_1) = \mathbf{V}_1^*(\mathbf{M}^*\mathbf{M})\mathbf{V}_1$ is unitarily similar to $\mathbf{M}^*\mathbf{M}$, so the two have the same eigenvalues and Definition 13.9.1 reads the same singular values off them. Hence by Theorem 16.3.3 (c),

$$\|(\mathbf{A}^*\mathbf{A})^{-1}\| = \sigma_n^{-2}, \qquad \|(\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^*\| = \sigma_n^{-1}.$$

Applying these to $(*)$ with Theorem 16.3.2 (a),

$$\|\mathbf{x}'(0)\| \;\leq\; \frac{\|\mathbf{E}\|\,\|\mathbf{r}\|}{\sigma_n^2} \;+\; \frac{\|\mathbf{E}\|\,\|\mathbf{x}\|}{\sigma_n} \;+\; \frac{\|\mathbf{f}\|}{\sigma_n}.$$

Divide by $\|\mathbf{x}\|$ and use $\|\mathbf{A}\| = \sigma_1$ (Theorem 16.3.3 (c)) and $\kappa_2(\mathbf{A}) = \sigma_1/\sigma_n$ (Definition 16.8.1). The middle term becomes $\|\mathbf{E}\|/\sigma_n = \kappa_2(\mathbf{A})\varepsilon_{\mathbf{A}}$ exactly. For the other two, note that $\sigma_1\|\mathbf{x}\| \geq \|\mathbf{A}\mathbf{x}\| = \|\mathbf{b}\|\cos\theta$, so

$$\frac{\|\mathbf{r}\|}{\sigma_1\|\mathbf{x}\|} \leq \frac{\|\mathbf{b}\|\sin\theta}{\|\mathbf{b}\|\cos\theta} = \tan\theta, \qquad \frac{\|\mathbf{b}\|}{\sigma_1\|\mathbf{x}\|} \leq \frac{1}{\cos\theta}.$$

Therefore

$$\frac{\|\mathbf{E}\|\|\mathbf{r}\|}{\sigma_n^2\|\mathbf{x}\|} = \frac{\sigma_1^2}{\sigma_n^2}\cdot\frac{\|\mathbf{E}\|}{\sigma_1}\cdot\frac{\|\mathbf{r}\|}{\sigma_1\|\mathbf{x}\|} \leq \kappa_2(\mathbf{A})^2\varepsilon_{\mathbf{A}}\tan\theta,$$
$$\frac{\|\mathbf{f}\|}{\sigma_n\|\mathbf{x}\|} = \frac{\sigma_1}{\sigma_n}\cdot\frac{\|\mathbf{f}\|}{\|\mathbf{b}\|}\cdot\frac{\|\mathbf{b}\|}{\sigma_1\|\mathbf{x}\|} \leq \frac{\kappa_2(\mathbf{A})\varepsilon_{\mathbf{b}}}{\cos\theta}.$$

Adding the three estimates gives the stated bound. This proves the theorem.

■

What is linearized, and what is not. The theorem itself is exact: it bounds a derivative, and no term has been discarded. What is an approximation is the use one makes of it. Reading it as

$$\frac{\|\Delta\mathbf{x}\|}{\|\mathbf{x}\|} \lesssim \left[\kappa_2(\mathbf{A})\left(1+\tfrac{1}{\cos\theta}\right)+\kappa_2(\mathbf{A})^2\tan\theta\right]\varepsilon \tag{24.3.2}$$

for data perturbed by $\|\Delta\mathbf{A}\| \leq \varepsilon\|\mathbf{A}\|$ and $\|\Delta\mathbf{b}\| \leq \varepsilon\|\mathbf{b}\|$ drops a remainder of size $O(\varepsilon^2)$, exactly as Theorem 16.8.3 (b) drops the factor $1/(1-r)$. For the $\varepsilon \approx 10^{-16}$ of the floating-point model that remainder is invisible, and every statement below that carries a $\lesssim$ is a statement of this kind.

Two readings of the bound matter.

- **When the residual is zero**, $\theta = 0$ and the bound is $2\kappa_2(\mathbf{A})\varepsilon$: the least-squares problem is exactly as sensitive as a square linear system with the same condition number. A square invertible $\mathbf{A}$ is always this case, and Theorem 16.8.3 is recovered.
- **When the residual is large**, the term $\kappa_2(\mathbf{A})^2\tan\theta$ dominates the bound, and what it measures belongs to the *problem*: no algorithm can return an answer more accurate than data of that quality determine. This is the sentence Chapter 16 §08 promised, where the least-squares problem was said to have "a sensitivity that depends on $\kappa_2(\mathbf{A})$, on the residual $\|\mathbf{A}\mathbf{x}^+ - \mathbf{b}\|$ and on the angle between $\mathbf{b}$ and $\mathrm{col}(\mathbf{A})$". Exercise (C1: The squared term is attained) exhibits data on which the $\kappa_2(\mathbf{A})^2\varepsilon_{\mathbf{A}}\tan\theta$ term is attained exactly, so it is not an artifact of the estimates.

Quick check. Take $m = n$ with $\mathbf{A}$ invertible. What does Eq. 24.3.1 give for θ, and what does Eq. 24.3.2 reduce to? Compare with Theorem 16.8.3.

Solution. For an invertible square $\mathbf{A}$ the system $\mathbf{A}\mathbf{z} = \mathbf{b}$ is consistent, so its least-squares solution has residual $\mathbf{r} = \mathbf{0}$ and Eq. 24.3.1 gives $\sin\theta = 0$, that is $\theta = 0$. Then $\cos\theta = 1$ and $\tan\theta = 0$, and Eq. 24.3.2 reads $\|\Delta\mathbf{x}\|/\|\mathbf{x}\| \lesssim 2\kappa(\mathbf{A})\varepsilon$. That is Theorem 16.8.3 (a) and (b) added together, with the factor $1/(1-r)$ replaced by the $O(\varepsilon^2)$ that the linearization discards. The square case carries no κ^2 at all.

24.3.2 Two routes, and the digits they lose

The problem is measured; now the two algorithms. Both are run in the arithmetic of Definition 24.1.1, with u the unit roundoff and $\gamma_k = ku/(1-ku)$ the constant of Lemma 24.1.1.

The **normal-equation route** forms $\mathbf{C} = \mathrm{fl}(\mathbf{A}^*\mathbf{A})$ and $\mathbf{d} = \mathrm{fl}(\mathbf{A}^*\mathbf{b})$ entry by entry as inner products, then solves $\mathbf{Cz} = \mathbf{d}$. The matrix it hands to the solver need **not** be positive definite. The exact $\mathbf{A}^*\mathbf{A}$ is: $\mathbf{x}^*\mathbf{A}^*\mathbf{A}\mathbf{x} = \|\mathbf{A}\mathbf{x}\|^2 > 0$ for $\mathbf{x} \neq \mathbf{0}$, the columns of $\mathbf{A}$ being independent. The computed $\mathbf{C}$ is Hermitian, only one triangle of it being formed, but its eigenvalues are those of $\mathbf{A}^*\mathbf{A}$ displaced by the rounding, and the smallest of them, $\sigma_n(\mathbf{A})^2$, can be displaced past 0 — Example (A matrix where one route survives and the other does not) below exhibits a $\mathbf{C}$ that comes out singular, on which the Cholesky factorization of Theorem 13.2.3 fails outright. **When $\mathbf{C}$ is positive definite** — which it is while the rounding stays below $\sigma_n(\mathbf{A})^2$ — Cholesky, and the analysis Section 2 gives it, apply.

The cost is mn^2 flops to form the product — there are $\frac{1}{2}n(n+1)$ entries to compute, each an inner product of length m costing $2m$ flops — plus $\frac{1}{3}n^3$ for the factorization (Proposition 24.2.3 (c)) and $2n^2$ for the two triangular solves. The **QR route** never forms that product: it computes a factorization $\mathbf{A} = \mathbf{QR}$ by reflections (Theorem 11.8.4), applies $\mathbf{Q}^*$ to $\mathbf{b}$, and solves the triangular system $\mathbf{Rx} = \mathbf{Q}^*\mathbf{b}$ by back substitution, which Theorem 24.2.5 showed to be backward stable. Its cost is $2mn^2 - \frac{2}{3}n^3$ flops by a sum of the same kind as Proposition 24.2.3, which we do not write out: about twice the first when $m \gg n$.

The rule of thumb of Theorem 24.1.3 says: forward relative error $\leq$ condition number $\times$ backward relative error. The two routes solve *different problems*, and that is the whole story. The first solves a square linear system whose matrix is $\mathbf{A}^*\mathbf{A}$, of condition number $\kappa_2(\mathbf{A})^2$ by Theorem 16.8.4. The second solves the least-squares problem itself, of sensitivity given by Theorem 24.3.1. Here is the comparison. Part (a) is Theorem 24.1.3 applied to the system $\mathbf{A}^*\mathbf{A}\mathbf{x} = \mathbf{A}^*\mathbf{b}$, except that here the right-hand side is perturbed as well, so we give the short direct argument instead of appealing to it.

Theorem 24.3.2. The Normal Equations Lose Twice the Digits

Let $\mathbf{A} \in M_{m\times n}(F)$ have rank $n \leq m$, let $\mathbf{b} \neq \mathbf{0}$, let $\mathbf{x}$ be the least-squares solution, and let $\theta < \pi/2$ be as in Eq. 24.3.1. Work in the model of Definition 24.1.1, and suppose $mu \leq \frac{1}{2}$.

(a) **(Normal equations.)** Suppose $\mathbf{C} = \mathrm{fl}(\mathbf{A}^*\mathbf{A})$ and $\mathbf{d} = \mathrm{fl}(\mathbf{A}^*\mathbf{b})$ are formed by inner products, and that $\mathbf{Cz} = \mathbf{d}$ is then solved by a method that is backward stable in the sense of Definition 24.1.3, so that the computed $\widehat{\mathbf{x}}$ satisfies $(\mathbf{C}+\Delta\mathbf{C})\widehat{\mathbf{x}} = \mathbf{d}+\Delta\mathbf{d}$ with $\|\Delta\mathbf{C}\| \leq c_s u\|\mathbf{C}\|$ and $\|\Delta\mathbf{d}\| \leq c_s u\|\mathbf{d}\|$. Then there is a constant c, depending only on m, n, and c_s and of size $O(mn)$, such that if $c\,u\,\kappa_2(\mathbf{A})^2 < 1$ then

$$\frac{\|\widehat{\mathbf{x}}-\mathbf{x}\|}{\|\mathbf{x}\|} \;\leq\; \frac{c\,u\,\kappa_2(\mathbf{A})^2\left(1+\frac{1}{\cos\theta}\right)}{1-c\,u\,\kappa_2(\mathbf{A})^2}.$$

The bound carries $\kappa_2(\mathbf{A})^2$ **even when the residual is zero**.

(b) **(A backward stable least-squares solver.)** Suppose instead that $\widehat{\mathbf{x}}$ is the exact least-squares solution of $(\mathbf{A}+\Delta\mathbf{A})\mathbf{z} = \mathbf{b}+\Delta\mathbf{b}$ for some $\Delta\mathbf{A}, \Delta\mathbf{b}$ with $\|\Delta\mathbf{A}\| \leq c'u\|\mathbf{A}\|$ and $\|\Delta\mathbf{b}\| \leq c'u\|\mathbf{b}\|$. Then, to first order in u as in Eq. 24.3.2,

$$\frac{\|\widehat{\mathbf{x}}-\mathbf{x}\|}{\|\mathbf{x}\|} \;\lesssim\; c'u\Big[\kappa_2(\mathbf{A})\left(1+\tfrac{1}{\cos\theta}\right) + \kappa_2(\mathbf{A})^2\tan\theta\Big].$$

Idea. For (a), the two sources of error — forming the product, and solving the system — are both perturbations of the *exact* normal equations, and both are bounded by a constant times $u\|\mathbf{A}\|^2$ or $u\|\mathbf{A}\|\|\mathbf{b}\|$. Subtract the exact equations from the perturbed ones and multiply by $(\mathbf{A}^*\mathbf{A})^{-1}$, whose norm is σ_n^{-2}: the two factors of σ_1/σ_n appear there and nowhere else. The one nuisance is that the estimate produces $\|\widehat{\mathbf{x}}\|$ rather than $\|\mathbf{x}\|$; subadditivity of the norm (Corollary 11.1.6) trades one for the other at the cost of the denominator, exactly as in Theorem 16.8.3 (b). For (b) there is nothing to do: the hypothesis says the computed answer is the exact answer to perturbed data, which is what Theorem 24.3.1 measures.

Proof. (a). We argue over $\mathbb{R}$, the field in which Theorem 24.1.2 is stated; over $\mathbb{C}$ each complex multiplication and addition is a bounded number of real ones, and the same argument gives the same bound with a larger constant c. Write $\mathbf{a}_1,\dots,\mathbf{a}_n$ for the columns of $\mathbf{A}$, so that $(\mathbf{A}^*\mathbf{A})_{ij}=\mathbf{a}_i^*\mathbf{a}_j$ and $(\mathbf{A}^*\mathbf{b})_i=\mathbf{a}_i^*\mathbf{b}$, each an inner product of length m. By Theorem 24.1.2, $\mathrm{fl}(\mathbf{a}_i^*\mathbf{a}_j)=(\mathbf{a}_i+\Delta\mathbf{a})^*\mathbf{a}_j$ with $|\Delta\mathbf{a}|\le\gamma_m|\mathbf{a}_i|$ entrywise, so

$$|c_{ij}-\mathbf{a}_i^*\mathbf{a}_j|\le\gamma_m|\mathbf{a}_i|^{\top}|\mathbf{a}_j|\le\gamma_m\|\mathbf{a}_i\|\|\mathbf{a}_j\|,$$

the last step by Cauchy–Schwarz (Theorem 11.1.5). Summing squares and using $\|\mathbf{M}\|\le\|\mathbf{M}\|_F$ — true because $\|\mathbf{M}\|_F^2=\operatorname{tr}(\mathbf{M}^*\mathbf{M})$ is the sum of the eigenvalues of $\mathbf{M}^*\mathbf{M}$, hence of the $\sigma_i(\mathbf{M})^2$, while $\|\mathbf{M}\|=\sigma_1(\mathbf{M})$ by Theorem 16.3.3 (c) —

$$\|\mathbf{C}-\mathbf{A}^*\mathbf{A}\|\le\|\mathbf{C}-\mathbf{A}^*\mathbf{A}\|_F\le\gamma_m\Big(\sum_i\|\mathbf{a}_i\|^2\Big)=\gamma_m\|\mathbf{A}\|_F^2\le n\gamma_m\|\mathbf{A}\|^2.$$

The same argument on $\mathbf{d}$ gives $\|\mathbf{d}-\mathbf{A}^*\mathbf{b}\|\le\gamma_m\|\mathbf{b}\|\|\mathbf{A}\|_F\le\sqrt{n}\,\gamma_m\|\mathbf{A}\|\|\mathbf{b}\|$.

Now add the solver's backward error. Put $\boldsymbol{\Delta}=(\mathbf{C}+\Delta\mathbf{C})-\mathbf{A}^*\mathbf{A}$ and $\mathbf{h}=(\mathbf{d}+\Delta\mathbf{d})-\mathbf{A}^*\mathbf{b}$. Since $\|\mathbf{C}\|\le(1+n\gamma_m)\|\mathbf{A}\|^2$ and $\|\mathbf{d}\|\le(1+\sqrt{n}\gamma_m)\|\mathbf{A}\|\|\mathbf{b}\|$,

$$\|\boldsymbol{\Delta}\|\le c\,u\|\mathbf{A}\|^2,\qquad\|\mathbf{h}\|\le c\,u\|\mathbf{A}\|\|\mathbf{b}\|,$$

with $c=2nm+c_s(1+2nm)$ serving, say, since the hypothesis $mu\le\frac12$ gives $\gamma_m\le2mu$, and $\sqrt{n}\le n$.

The exact solution satisfies $\mathbf{A}^*\mathbf{A}\mathbf{x}=\mathbf{A}^*\mathbf{b}$ and the computed one satisfies $(\mathbf{A}^*\mathbf{A}+\boldsymbol{\Delta})\widehat{\mathbf{x}}=\mathbf{A}^*\mathbf{b}+\mathbf{h}$. Subtracting,

$$\mathbf{A}^*\mathbf{A}(\widehat{\mathbf{x}}-\mathbf{x})=\mathbf{h}-\boldsymbol{\Delta}\widehat{\mathbf{x}},\qquad\|\widehat{\mathbf{x}}-\mathbf{x}\|\le\sigma_n^{-2}\big(c\,u\|\mathbf{A}\|\|\mathbf{b}\|+c\,u\|\mathbf{A}\|^2\|\widehat{\mathbf{x}}\|\big),$$

using $\|(\mathbf{A}^*\mathbf{A})^{-1}\|=\sigma_n^{-2}$ from the proof of Theorem 24.3.1. Divide by $\|\mathbf{x}\|$, write $\rho=\|\widehat{\mathbf{x}}-\mathbf{x}\|/\|\mathbf{x}\|$, and use $\sigma_1\|\mathbf{x}\|\ge\|\mathbf{b}\|\cos\theta$ together with $\|\widehat{\mathbf{x}}\|\le\|\mathbf{x}\|(1+\rho)$:

$$\rho\;\le\;c\,u\,\kappa_2(\mathbf{A})^2\Big(\frac{1}{\cos\theta}+1+\rho\Big).$$

Since $c\,u\,\kappa_2(\mathbf{A})^2<1$ we may collect the ρ terms and divide, which gives the stated bound. Nothing in the argument used $\mathbf{r}$, so it holds for $\theta=0$ as it stands.

(b). By hypothesis $\widehat{\mathbf{x}}=\mathbf{x}(1)$ for the family of Theorem 24.3.1 with $\mathbf{E}=\Delta\mathbf{A}$ and $\mathbf{f}=\Delta\mathbf{b}$, and $\varepsilon_{\mathbf{A}},\varepsilon_{\mathbf{b}}\le c'u$. The stated bound is Eq. 24.3.2 for that family. This proves the theorem.

■

In digits. A relative error of η means about $\log_{10}(1/\eta)$ correct decimal digits, and an arithmetic with unit roundoff $u \approx 10^{-16}$ offers about 16 of them. Part (a) says the normal-equation route returns an answer with relative error about $u\kappa_2(\mathbf{A})^2$: it **loses about** $2\log_{10}\kappa_2(\mathbf{A})$ **digits**, whatever the residual. Part (b) says a backward stable least-squares solver returns an answer with relative error about $u\kappa_2(\mathbf{A})$ when the residual is small, and so **loses about** $\log_{10}\kappa_2(\mathbf{A})$ **digits**.

The comparison Chapter 16 §08 left open has two halves, and only one of them is settled here. The half that is proved is the *implication*: **a** backward stable least-squares solver loses about $\log_{10}\kappa_2(\mathbf{A})$ digits, while forming $\mathbf{A}^*\mathbf{A}$ loses about twice that — and part (a), which carries no stability hypothesis at all, is unconditional.

The half that is **not** proved is that the reflection route of Chapter 11 §08 supplies such a solver. That is the accumulated statement of the remark after Theorem 24.3.3 below, which this book records with credit and does not prove; what Theorem 24.3.3 itself proves is the backward stability of **one** reflection. A further qualification is the content of the next warning.

Warning. **The two-versus-one comparison is a small-residual statement, and the reason is not symmetric.** For a large residual the bound in (b) also carries a $\kappa_2(\mathbf{A})^2$, through the term $\kappa_2(\mathbf{A})^2\tan\theta$. The difference is *where that square comes from.* In (b) it is the sensitivity of the problem, certified by Theorem 24.3.1, and no method whatever can avoid it. In (a) it is manufactured by the algorithm: it is there when $\theta = 0$, where the problem's own sensitivity is only $\kappa_2(\mathbf{A})$. So the honest form of the comparison is: **a backward stable method loses what the problem costs; forming $\mathbf{A}^*\mathbf{A}$ loses $\kappa_2(\mathbf{A})$ times more than that whenever $\tan\theta \lesssim 1/\kappa_2(\mathbf{A})$, and never less.**

Example. A matrix where one route survives and the other does not

Let $\varepsilon > 0$ and

$$\mathbf{A} = \begin{pmatrix} 1 & 1 \\ \varepsilon & 0 \\ 0 & \varepsilon \end{pmatrix}, \qquad \mathbf{b} = \begin{pmatrix} 1 \\ 2\varepsilon \\ -\varepsilon \end{pmatrix}.$$

Find $\kappa_2(\mathbf{A})$, the exact least-squares solution, and the residual. Then run both routes in the model with $u = 2^{-53}$, for $\varepsilon = 10^{-6}$ and for $\varepsilon = 10^{-9}$, and compare.

Solution. Exact data. $\mathbf{A}^\top\mathbf{A} = \left(\begin{smallmatrix} 1+\varepsilon^2 & 1 \\ 1 & 1+\varepsilon^2 \end{smallmatrix}\right)$, which sends $(1,1)$ to $(2+\varepsilon^2)(1,1)$ and $(1,-1)$ to $\varepsilon^2(1,-1)$. So its eigenvalues are $2+\varepsilon^2$ and ε^2, the singular values of $\mathbf{A}$ are $\sqrt{2+\varepsilon^2}$ and ε, and

$$\kappa_2(\mathbf{A}) = \frac{\sqrt{2+\varepsilon^2}}{\varepsilon}.$$

Also $\mathbf{A}(2,-1) = (2-1,\ 2\varepsilon,\ -\varepsilon) = \mathbf{b}$, so the system is **consistent**: the least-squares solution is $\mathbf{x} = (2,-1)$ exactly, the residual is $\mathbf{r} = \mathbf{0}$, and $\theta = 0$. By Theorem 24.3.1 the problem's sensitivity is exactly $2\kappa_2(\mathbf{A})$ — there is no κ_2^2 in the problem at all.

$\varepsilon = 10^{-6}$. Here $\kappa_2(\mathbf{A}) = 1.4142136 \cdot 10^6$, so $u\kappa_2(\mathbf{A}) = 1.6 \cdot 10^{-10}$ and $u\kappa_2(\mathbf{A})^2 = 2.2 \cdot 10^{-4}$. Forming the product rounds $1+\varepsilon^2 = 1+10^{-12}$ to 1.0000000000010000889; the last four digits are the rounding, and they are a relative perturbation of about 10^{-4} in the *difference*

ε^2 that carries the small eigenvalue. Cholesky and the two triangular solves then return

$$\widehat{\mathbf{x}} = (1.99983348134991,\ -0.99983348134991),$$

a relative error of $\|\widehat{\mathbf{x}} - \mathbf{x}\|/\|\mathbf{x}\| = 1.05 \cdot 10^{-4}$ — about four correct digits out of sixteen, and within a factor of 2 of the bound $u\kappa_2(\mathbf{A})^2 = 2.2 \cdot 10^{-4}$. (Entry by entry the relative errors are $8.3 \cdot 10^{-5}$ and $1.7 \cdot 10^{-4}$; it is the 2-norm figure that the theorem bounds.) The reflection route returns $(2, -1)$, every digit correct.

$\varepsilon = 10^{-9}$. Now $\kappa_2(\mathbf{A}) = 1.4142136 \cdot 10^9$ and $u\kappa_2(\mathbf{A})^2 = 222$, so part (a) promises nothing at all — and nothing is what it delivers. Since $\varepsilon^2 = 10^{-18}$ is below u, the computed product is

$$\mathbf{C} = \mathrm{fl}(\mathbf{A}^\top \mathbf{A}) = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix},$$

a **singular** matrix: its Cholesky factorization (Theorem 13.2.3) fails at the second pivot, and the route cannot even be run. The reflection route again returns $(2, -1)$.

The bound is an upper bound. For $\varepsilon = 10^{-9}$ the reflection route is entitled to lose $\log_{10} \kappa_2(\mathbf{A}) \approx 9$ digits and in fact loses none: the answer is exactly representable and the arithmetic happens to reach it. Part (b) is a worst case over perturbations, as Theorem 16.8.3 was. Part (a) is an upper bound too, and it is not attained exactly either; what this example shows is that it is not a pessimistic one — the normal-equation route does lose its $2 \log_{10} \kappa_2(\mathbf{A})$ digits here, to within a factor of 2 in the error.

Warning. **Forming $\mathbf{A}^*\mathbf{A}$ is not forbidden, it is limited.** If $\kappa_2(\mathbf{A}) = 10^3$ then $u\kappa_2(\mathbf{A})^2 \approx 10^{-10}$, and the normal equations return ten correct digits for half the arithmetic. The rule is quantitative, not moral: the route is safe exactly while $u\kappa_2(\mathbf{A})^2$ is small enough for the purpose, and it collapses — as Example (A matrix where one route survives and the other does not) shows — once $\kappa_2(\mathbf{A})$ approaches $u^{-1/2} \approx 10^8$. What is never acceptable is forming the product without knowing which side of that threshold one is on.

24.3.3 Why the reflection route is the stable one

Part (b) above assumed a backward stable least-squares solver. That hypothesis has to be earned, and earning it is what Chapter 11 §08's reflections are for. The claim to be supported is: applying a Householder reflection by the formula

$$\mathbf{H}_{\mathbf{w}}\mathbf{v} = \mathbf{v} - \frac{2\langle \mathbf{v}, \mathbf{w} \rangle}{\|\mathbf{w}\|^2}\,\mathbf{w} \tag{24.3.3}$$

of Chapter 11 §08 — **applying it, never forming the matrix** — commits a backward error of size u, with no dependence on the conditioning of anything. Here is that statement, proved.

Theorem 24.3.3. Backward Stability of One Reflection

Work in the model of Definition 24.1.1 over $\mathbb{R}$, and suppose $(10m + 16)u \le \frac{1}{4}$. Let $\mathbf{w} \in \mathbb{R}^m$ be nonzero and let $\mathbf{H} = \mathbf{H}_{\mathbf{w}}$ be the reflection of Definition 11.8.1, and suppose $\mathbf{w}$ and the input are stored exactly.

(a) For $\mathbf{v} \in \mathbb{R}^m$, let $\widehat{\mathbf{y}}$ be the vector computed from $\mathbf{v}$ and $\mathbf{w}$ by Eq. 24.3.3. Then

$$\widehat{\mathbf{y}} = \mathbf{H}(\mathbf{v} + \Delta\mathbf{v}), \qquad \|\Delta\mathbf{v}\|_2 \le (10m + 16)\, u\, \|\mathbf{v}\|_2.$$

(b) For $\mathbf{A} \in M_{m\times n}(\mathbb{R})$, let $\widehat{\mathbf{B}}$ be the matrix whose columns are computed from those of $\mathbf{A}$ in this way. Then

$$\widehat{\mathbf{B}} = \mathbf{H}(\mathbf{A} + \Delta\mathbf{A}), \qquad \|\Delta\mathbf{A}\|_F \le (10m + 16)\, u\, \|\mathbf{A}\|_F.$$

Idea. Three quantities are computed before anything is subtracted: $\|\mathbf{w}\|^2$, the inner product $\langle \mathbf{v}, \mathbf{w}\rangle$, and their quotient β. Each is an inner product or a division, so Theorem 24.1.2 and Lemma 24.1.1 bound them at once, and the crude estimate $|\beta| \le 2\|\mathbf{v}\|/\|\mathbf{w}\|$ — which is Cauchy–Schwarz and nothing more — keeps every error proportional to $\|\mathbf{v}\|$, not to $\|\mathbf{v}\|/\|\mathbf{w}\|$. Then the forward error $\|\widehat{\mathbf{y}} - \mathbf{H}\mathbf{v}\|$ is converted into a backward error for free, because $\mathbf{H}$ is an isometry and its own inverse: $\Delta\mathbf{v} = \mathbf{H}(\widehat{\mathbf{y}} - \mathbf{H}\mathbf{v})$ has exactly the same length. Part (b) is part (a) column by column, since the Frobenius norm adds the columns in squares.

Proof. *(a).* Write $\nu = \|\mathbf{w}\|_2$, $\beta = 2\langle \mathbf{v}, \mathbf{w}\rangle/\nu^2$, so that $\mathbf{H}\mathbf{v} = \mathbf{v} - \beta\mathbf{w}$, and abbreviate $\gamma_k = ku/(1-ku)$. The hypothesis $(10m+16)u \le \frac{1}{4}$ gives in particular $(m+2)u \le \frac{1}{4}$, so $\gamma_k \le \frac{4}{3}ku \le \frac{1}{3}$ for every $k \le m+2$.

Step 1: the two inner products. The computed $s = \mathrm{fl}(\mathbf{w}^*\mathbf{w})$ is a sum of m nonnegative terms w_i^2, each carrying at most m rounding factors, so by Lemma 24.1.1 $s = \sum_i w_i^2(1+\theta_i)$ with $|\theta_i| \le \gamma_m$; as every $w_i^2 \ge 0$ this gives $s = \nu^2(1+\theta)$ with $|\theta| \le \gamma_m$. By Theorem 24.1.2 the computed $p = \mathrm{fl}(\mathbf{v}^*\mathbf{w})$ satisfies $p = \langle \mathbf{v}, \mathbf{w}\rangle + \eta$ with

$$|\eta| \le \gamma_m |\mathbf{v}|^\top |\mathbf{w}| \le \gamma_m \|\mathbf{v}\|\nu,$$

by Cauchy–Schwarz (Theorem 11.1.5).

Step 2: the coefficient. The computed $\widehat{\beta}$ is $(2p/s)$ times two rounding factors, so $\widehat{\beta} = (2p/s)(1+\theta_2)$ with $|\theta_2| \le \gamma_2$, by Lemma 24.1.1. Writing $2p/s = (\beta + \zeta)/(1+\theta)$ with $\zeta = 2\eta/\nu^2$ and $|\zeta| \le 2\gamma_m\|\mathbf{v}\|/\nu$,

$$\left|\frac{1+\theta_2}{1+\theta} - 1\right| \le \frac{\gamma_m + \gamma_2}{1-\gamma_m} \le \tfrac{3}{2}(\gamma_m + \gamma_2) \le 2(m+2)u,$$

using $1 - \gamma_m \ge \frac{2}{3}$ and $\gamma_k \le \frac{4}{3}ku$; in particular the factor itself has modulus at most 2. Since $|\beta| \le 2\|\mathbf{v}\|/\nu$ by Cauchy–Schwarz,

$$|\widehat{\beta} - \beta| \le |\beta| \cdot 2(m+2)u + 2|\zeta| \le \frac{\|\mathbf{v}\|}{\nu}\big(4(m+2)u + 6mu\big) \le K\, u\, \frac{\|\mathbf{v}\|}{\nu},$$

with $K = 10m + 8$. In particular $|\widehat{\beta}| \le 3\|\mathbf{v}\|/\nu$, since $Ku \le (10m+16)u \le \frac{1}{4}$.

Step 3: the subtraction. Entry i is computed as $\widehat{y}_i = (v_i - \widehat{\beta}w_i(1+\delta_i))(1+\delta_i')$ with $|\delta_i|, |\delta_i'| \le u$. Subtracting $(\mathbf{H}\mathbf{v})_i = v_i - \beta w_i$ and regrouping,

$$\widehat{y}_i - (\mathbf{H}\mathbf{v})_i = (v_i - \widehat{\beta}w_i)\delta_i' + (\beta - \widehat{\beta})w_i - \widehat{\beta}w_i\delta_i(1+\delta_i').$$

Take 2-norms over i, one term at a time. The first is at most $u(\|\mathbf{v}\| + |\widehat{\beta}|\nu) \le 4u\|\mathbf{v}\|$; the second is $|\beta - \widehat{\beta}|\nu \le Ku\|\mathbf{v}\|$; the third is at most $|\widehat{\beta}|\nu\, u(1+u) \le 4u\|\mathbf{v}\|$. Hence

$$\|\widehat{\mathbf{y}} - \mathbf{H}\mathbf{v}\|_2 \le (K+8)u\|\mathbf{v}\|_2 = (10m+16)u\|\mathbf{v}\|_2.$$

Step 4: forward to backward. Put $\Delta\mathbf{v} = \mathbf{H}(\widehat{\mathbf{y}} - \mathbf{H}\mathbf{v})$. Then $\mathbf{H}(\mathbf{v}+\Delta\mathbf{v}) = \mathbf{H}\mathbf{v} + \mathbf{H}^2(\widehat{\mathbf{y}} - \mathbf{H}\mathbf{v}) = \widehat{\mathbf{y}}$ by Proposition 11.8.2 (b), and $\|\Delta\mathbf{v}\|_2 = \|\widehat{\mathbf{y}} - \mathbf{H}\mathbf{v}\|_2$ because $\mathbf{H}$ is unitary (Proposition 11.8.2 (c)). This is (a).

(b). Apply (a) to each column $\mathbf{a}_j$, obtaining $\Delta\mathbf{a}_j$ with $\|\Delta\mathbf{a}_j\|_2 \le (10m+16)u\|\mathbf{a}_j\|_2$, and let $\Delta\mathbf{A}$ have these columns. Then $\widehat{\mathbf{B}} = \mathbf{H}(\mathbf{A} + \Delta\mathbf{A})$ column by column, and squaring and summing over j gives $\|\Delta\mathbf{A}\|_F \le (10m+16)u\|\mathbf{A}\|_F$. This proves the theorem.

■

The constants are crude on purpose; what matters is their shape. **The bound is a multiple of u and of m, and of nothing else.** No condition number appears, and none can: a reflection is an isometry, so it neither stretches nor shrinks, and there is no small quantity anywhere for a rounding error to be measured against.

Remark. **What the accumulation says, and that we do not prove it.** Running $\min(n, m-1)$ reflections as in Theorem 11.8.4, one obtains computed $\widehat{\mathbf{R}}$ and stored reflection vectors. The classical theorem — due to Wilkinson, and proved by inducting the argument above over the reflections while tracking how the perturbations compose — states that there is an **exactly** unitary $\widetilde{\mathbf{Q}} \in M_m(\mathbb{R})$, not in general the product of the computed reflections, with

$$\widetilde{\mathbf{Q}}\,\widehat{\mathbf{R}} = \mathbf{A} + \Delta\mathbf{A}, \qquad \|\Delta\mathbf{A}\|_F \le c(m,n)\, u\, \|\mathbf{A}\|_F,$$

and that the resulting least-squares solution is the exact solution of a problem with data perturbed by $O(u)$ — the hypothesis of Theorem 24.3.2 (b). **This accumulated statement is not proved here**, and nothing in this book depends on it: every theorem above either proves its own case or carries backward stability as a stated hypothesis. The induction is not deep, but it is long, and the honest thing is to say so.

24.3.4 What goes wrong with Gram–Schmidt

Chapter 11 §02 warned that "the classical process is numerically fragile" and promised that this chapter would say why the reflection route is the stable one. Half of the answer is Theorem 24.3.3. The other half is what the Gram–Schmidt recursion of Theorem 11.2.2 does instead.

The recursion subtracts from $\mathbf{a}_k$ its components along the vectors already produced. When the columns are nearly dependent, what is left is a small difference of large quantities, and the leading digits cancel, exactly as in the subtraction of nearly equal numbers examined in Section 1. The computed $\mathbf{q}_k$ is then a unit vector in a noticeably wrong direction, and, worse, the error is not random: it is inherited by every later step. The measured consequence is a loss of *orthogonality*: the computed $\widehat{\mathbf{Q}}$ satisfies $\widehat{\mathbf{Q}}^{\top}\widehat{\mathbf{Q}} \ne \mathbf{I}$ by an amount that grows with $\kappa_2(\mathbf{A})$. The classical analyses, which we state with credit and **do not prove here**, say how fast:

- for the **classical** process, $\|\widehat{\mathbf{Q}}^{\top}\widehat{\mathbf{Q}} - \mathbf{I}\|$ is of order $u\,\kappa_2(\mathbf{A})^2$ (Kielbasinski; Giraud, Langou, and Rozloznik);
- for the **modified** process, in which each new vector is projected against the $\mathbf{q}_i$ one at a

time as it is formed rather than all at once from the original $\mathbf{a}_k$, it is of order $u\,\kappa_2(\mathbf{A})$ (Bjorck);
- for the reflection route it is of order u, independently of $\kappa_2(\mathbf{A})$, which is the content of Theorem 24.3.3.

Example. Three columns and three answers

Let $\varepsilon = 10^{-6}$ and

$$\mathbf{A} = \begin{pmatrix} 1 & 1 & 1 \\ \varepsilon & 0 & 0 \\ 0 & \varepsilon & 0 \end{pmatrix}.$$

Show that $\kappa_2(\mathbf{A}) = 3/\varepsilon + O(\varepsilon)$, then run the classical and the modified recursion in the model with $u = 2^{-53}$ and compare the inner products of the computed vectors.

Solution. The condition number. $\mathbf{A}^{\top}\mathbf{A} = \begin{pmatrix} 1+\varepsilon^2 & 1 & 1 \\ 1 & 1+\varepsilon^2 & 1 \\ 1 & 1 & 1 \end{pmatrix}$. Its characteristic polynomial factors as

$$\det(t\mathbf{I} - \mathbf{A}^{\top}\mathbf{A}) = (t - \varepsilon^2)(t^2 - (3+\varepsilon^2)t + \varepsilon^2),$$

as one checks by expanding both sides. The quadratic has roots $t_{\pm}$ with $t_+ t_- = \varepsilon^2$ and $t_+ + t_- = 3 + \varepsilon^2$, so $t_+ = 3 + O(\varepsilon^2)$ and $t_- = \varepsilon^2/3 + O(\varepsilon^4)$. Hence $\sigma_1 = \sqrt{3} + O(\varepsilon^2)$ and $\sigma_3 = \varepsilon/\sqrt{3} + O(\varepsilon^3)$, and

$$\kappa_2(\mathbf{A}) = \frac{\sqrt{3}}{\varepsilon/\sqrt{3}} + O(\varepsilon) = \frac{3}{\varepsilon} + O(\varepsilon) = 3.000000 \cdot 10^{6}.$$

So $u\kappa_2(\mathbf{A}) = 3.3 \cdot 10^{-10}$ and $u\kappa_2(\mathbf{A})^2 = 1.0 \cdot 10^{-3}$.

The computed vectors. Both recursions produce the same $\mathbf{q}_1$ and $\mathbf{q}_2$, with $\langle \mathbf{q}_1, \mathbf{q}_2 \rangle = 6.3 \cdot 10^{-11}$: that is the unavoidable $u\kappa_2(\mathbf{A})$, and it is already five to six digits worse than u, being larger by a factor of $5.7 \cdot 10^5$. The two differ at the third step, which is where the cancellation bites:

$$\begin{aligned} \text{classical:} \quad & \langle \mathbf{q}_2, \mathbf{q}_3 \rangle = -8.89 \cdot 10^{-5}, \\ \text{modified:} \quad & \langle \mathbf{q}_2, \mathbf{q}_3 \rangle = 0. \end{aligned}$$

The classical figure is within a factor of 12 of $u\kappa_2(\mathbf{A})^2 = 1.0 \cdot 10^{-3}$; the modified figure is at the level of $u\kappa_2(\mathbf{A})$ or below. Two vectors at inner product $9 \cdot 10^{-5}$ are at about 89.995 degrees to each other, and a basis that is orthonormal only to five digits is worth five digits in everything computed from it.

24.3.5 The sign in the reflection, quantified

Chapter 11 §08 chose $\mathbf{w} = \mathbf{x} + \alpha\|\mathbf{x}\|\mathbf{e}_1$ rather than $\mathbf{w}' = \mathbf{x} - \alpha\|\mathbf{x}\|\mathbf{e}_1$, where α is the phase of x_1, and said: "Chapter 24 quantifies the loss; the summary here is that both signs are correct mathematics and only one is correct arithmetic." Here is the quantity.

Take $\mathbf{x} \in \mathbb{R}^m$ nonzero with $x_1 > 0$, so $\alpha = 1$, and write $\widetilde{\mathbf{x}} = (x_2, \dots, x_m)$. Expanding as in Theorem 11.8.3,

$$\|\mathbf{w}\|^2 = 2\|\mathbf{x}\|(\|\mathbf{x}\| + x_1), \qquad \|\mathbf{w}'\|^2 = 2\|\mathbf{x}\|(\|\mathbf{x}\| - x_1),$$

and $(\|\mathbf{x}\| + x_1)(\|\mathbf{x}\| - x_1) = \|\widetilde{\mathbf{x}}\|^2$. Both vectors are built by the same two operations: compute $\|\mathbf{x}\|$, then add it to or subtract it from x_1. The computed norm carries an error of absolute size about $u\|\mathbf{x}\|$, and every other entry is copied from $\mathbf{x}$ exactly. So in both cases the computed vector is $\mathbf{w} + \Delta\mathbf{w}$ with $\|\Delta\mathbf{w}\| \approx u\|\mathbf{x}\|$ — *the same absolute error*. What differs is what that error is measured against, because $\mathbf{H}_{\mathbf{w}}$ depends on $\mathbf{w}$ only through the line it spans. Writing $\mathbf{p} = \mathbf{w}/\|\mathbf{w}\|$ and $\mathbf{q} = (\mathbf{w} + \Delta\mathbf{w})/\|\mathbf{w} + \Delta\mathbf{w}\|$, Definition 11.8.1 gives $\mathbf{H}_{\mathbf{w}} - \mathbf{H}_{\mathbf{w}+\Delta\mathbf{w}} = -2(\mathbf{p}\mathbf{p}^* - \mathbf{q}\mathbf{q}^*)$, and

$$\|\mathbf{p}\mathbf{p}^* - \mathbf{q}\mathbf{q}^*\| \le \|\mathbf{p}(\mathbf{p}-\mathbf{q})^*\| + \|(\mathbf{p}-\mathbf{q})\mathbf{q}^*\| = 2\|\mathbf{p}-\mathbf{q}\| \le \frac{8\|\Delta\mathbf{w}\|}{\|\mathbf{w}\|}.$$

For the last step, abbreviate $\mathbf{s} = \Delta\mathbf{w}$ and split

$$\mathbf{p} - \mathbf{q} = \mathbf{w}\Big(\frac{1}{\|\mathbf{w}\|} - \frac{1}{\|\mathbf{w}+\mathbf{s}\|}\Big) - \frac{\mathbf{s}}{\|\mathbf{w}+\mathbf{s}\|}.$$

The second piece has norm $\|\mathbf{s}\|/\|\mathbf{w}+\mathbf{s}\|$, and the first has norm $\big|\|\mathbf{w}+\mathbf{s}\| - \|\mathbf{w}\|\big|/\|\mathbf{w}+\mathbf{s}\|$, which is at most the same quantity by the reverse triangle inequality (Lemma 16.1.1); and $\|\mathbf{w}+\mathbf{s}\| \ge \|\mathbf{w}\|/2$ as soon as $\|\mathbf{s}\| \le \|\mathbf{w}\|/2$, which is certainly the case here, $\|\mathbf{s}\|$ being of size $u\|\mathbf{x}\|$. So the computed reflection differs from the intended one by a small multiple of

$$u\,\frac{\|\mathbf{x}\|}{\|\mathbf{w}\|}, \qquad \text{where} \qquad \frac{\|\mathbf{x}\|}{\|\mathbf{w}\|} = \sqrt{\frac{\|\mathbf{x}\|}{2(\|\mathbf{x}\| + x_1)}} \;\le\; \frac{1}{\sqrt{2}}.$$

The safe sign therefore never amplifies: its reflection is accurate to a small multiple of u, no matter what $\mathbf{x}$ is. The other sign amplifies by the factor

$$\frac{\|\mathbf{w}\|}{\|\mathbf{w}'\|} = \sqrt{\frac{\|\mathbf{x}\| + x_1}{\|\mathbf{x}\| - x_1}} = \frac{\|\mathbf{x}\| + x_1}{\|\widetilde{\mathbf{x}}\|},$$

which is **unbounded**: it blows up precisely as $\mathbf{x}$ approaches a positive multiple of $\mathbf{e}_1$, where $\|\widetilde{\mathbf{x}}\| \to 0$ and $\mathbf{w}' \to \mathbf{0}$.

Example. The same vector, the two signs

Let $\varepsilon = 10^{-8}$ and $\mathbf{x} = (1, \varepsilon, \varepsilon) \in \mathbb{R}^3$. Predict the amplification factor for each sign, then apply each computed reflection to $\mathbf{x}$ in the model with $u = 2^{-53}$ and measure how flat the answer is.

Solution. Prediction. $\|\widetilde{\mathbf{x}}\| = \sqrt{2}\,\varepsilon = 1.41 \cdot 10^{-8}$ and $\|\mathbf{x}\| = \sqrt{1 + 2\varepsilon^2}$, so $\|\mathbf{x}\|/\|\mathbf{w}\| = 0.5$ for the safe sign, while

$$\frac{\|\mathbf{x}\|}{\|\mathbf{w}'\|} = \sqrt{\frac{\|\mathbf{x}\|}{2(\|\mathbf{x}\| - 1)}} = 7.07 \cdot 10^{7}, \qquad u\,\frac{\|\mathbf{x}\|}{\|\mathbf{w}'\|} = 7.9 \cdot 10^{-9}.$$

The arithmetic. The computed $\|\mathbf{x}\|$ is exactly 1: the true value exceeds 1 by $\varepsilon^2 = 10^{-16}$, which is below the unit roundoff, so the addition returns 1. Hence

$$\widehat{\mathbf{w}} = (2, \varepsilon, \varepsilon), \qquad \widehat{\mathbf{w}}' = (0, \varepsilon, \varepsilon).$$

The first is right to full relative accuracy in every entry. The second has lost its first entry

altogether: it should be -10^{-16} and it is 0. Applying Eq. 24.3.3,

$$\mathbf{H}_{\widehat{\mathbf{w}}}\,\mathbf{x} = (-1, 0, 0), \qquad \mathbf{H}_{\widehat{\mathbf{w}}'}\,\mathbf{x} = (1,\ -10^{-8},\ -10^{-8}).$$

The safe sign flattens $\mathbf{x}$ exactly: the two entries that were supposed to be annihilated are true zeros. The other sign leaves behind entries of size 10^{-8} beside a vector of norm 1 — a failure of $1.4 \cdot 10^{-8}$ where 10^{-16} was available, and within a factor of 2 of the estimate $u\|\mathbf{x}\|/\|\mathbf{w}'\| = 7.9 \cdot 10^{-9}$. Eight digits, thrown away by a sign.

For $\varepsilon = 10^{-6}$ and $\varepsilon = 10^{-4}$ the same computation leaves residues of $1.3 \cdot 10^{-10}$ and $1.5 \cdot 10^{-13}$, against estimates $7.9 \cdot 10^{-11}$ and $7.9 \cdot 10^{-13}$. The estimates fall by exactly 100 each time, as $1/\|\widetilde{\mathbf{x}}\|$ does; the residues fall by 112 and then by 825, because a residue depends on the actual rounding error committed in $\|\mathbf{x}\|$, which varies with the binary expansion of $\|\mathbf{x}\|$ and not only with its size. Over four orders of magnitude the estimate stays within a factor of 6 of the measurement, which is all an order-of-magnitude estimate promises.

Nonexample by minimal change. Change one character of Example (The same vector, the two signs): take $\mathbf{x} = (-1, \varepsilon, \varepsilon)$ with $\varepsilon = 10^{-8}$. The arithmetic is unaltered — the computed $\|\mathbf{x}\|$ is again exactly 1 — and the rule of Theorem 11.8.3 still works: the phase of x_1 is now $\alpha = -1$, so $\mathbf{w} = \mathbf{x} + \alpha\|\mathbf{x}\|\mathbf{e}_1 = (-2, \varepsilon, \varepsilon)$, again with $\|\mathbf{x}\|/\|\mathbf{w}\| = 0.5$ and again flattening $\mathbf{x}$ to $(1, 0, 0)$ exactly. What fails is the rule with the phase dropped, $\mathbf{w} = \mathbf{x} + \|\mathbf{x}\|\mathbf{e}_1$: it returns $(0, \varepsilon, \varepsilon)$, and that reflection leaves $\mathbf{x}$ with the same residue $1.4 \cdot 10^{-8}$ as before. So the clause carrying the weight is not the plus sign but the **phase** α: what the rule requires is that $\alpha\|\mathbf{x}\|$ be added to x_1 so that the first entry grows in modulus, never so that it cancels. Reading the plus sign literally is correct mathematics for every $\mathbf{x}$ and correct arithmetic only for $x_1 > 0$.

Warning. **The bad sign is not merely inaccurate; on one input it does not exist.** If $\mathbf{x} = c\,\mathbf{e}_1$ with $c > 0$, then $\mathbf{w}' = \mathbf{0}$ and Definition 11.8.1 does not apply at all. That is not an accident of the arithmetic but the limit of the cancellation: the quantity whose leading digits are being destroyed is $\|\mathbf{x}\| - x_1$, and on that input it is zero. An algorithm that is undefined at a point is inaccurate near it, and the sign rule of Theorem 11.8.3 is exactly the rule that keeps the point out of reach.

24.3.6 Exercises

24.3.6.1 A. Check your understanding

Exercise. A1

(a) Define θ for a least-squares problem $\mathbf{Az} \approx \mathbf{b}$ with $\operatorname{rank}\mathbf{A} = n$, and say what $\theta = 0$ means about the system.

(b) State Theorem 24.3.1, and say which of its two terms survives when the residual is zero.

(c) Explain in one sentence why $\kappa_2(\mathbf{A})^2$ appears in the normal-equation route even for a consistent system, and why it need not appear in the reflection route.

(d) True or false: "a matrix with $\kappa_2(\mathbf{A}) = 10^3$ should never be handled by the normal equations". Justify your answer.

(e) What quantity does Theorem 24.3.3 bound, and which quantity does it **not** depend

on?

Solution.

(a) By Eq. 24.3.1, $\theta \in [0, \pi/2]$ is determined by $\sin\theta = \|\mathbf{r}\|/\|\mathbf{b}\|$ with $\mathbf{r} = \mathbf{b} - \mathbf{A}\mathbf{x}$ the residual, equivalently $\cos\theta = \|\mathbf{A}\mathbf{x}\|/\|\mathbf{b}\|$; it is the angle between $\mathbf{b}$ and $\mathrm{col}(\mathbf{A})$. Then $\theta = 0$ says $\mathbf{r} = \mathbf{0}$, that is, the system $\mathbf{A}\mathbf{z} = \mathbf{b}$ is consistent.

(b) See Theorem 24.3.1: the relative sensitivity is at most $\kappa_2(\mathbf{A})(\varepsilon_{\mathbf{A}} + \varepsilon_{\mathbf{b}}/\cos\theta) + \kappa_2(\mathbf{A})^2\varepsilon_{\mathbf{A}}\tan\theta$. With $\mathbf{r} = \mathbf{0}$ we get $\tan\theta = 0$, and only the $\kappa_2(\mathbf{A})$ term survives, with $\cos\theta = 1$.

(c) The normal-equation route solves a *different* linear system, the one with matrix $\mathbf{A}^*\mathbf{A}$, whose condition number is $\kappa_2(\mathbf{A})^2$ by Theorem 16.8.4, and the rounding committed in forming that matrix is magnified by its condition number, not by $\mathbf{A}$'s; the reflection route never forms the product, and $\kappa_2(\mathbf{R}) = \kappa_2(\mathbf{A})$.

(d) False. By Theorem 24.3.2 (a) the loss is about $2\log_{10}\kappa_2(\mathbf{A}) = 6$ digits out of sixteen, which is usually ample; the route costs about half the arithmetic. The warning after Example (A matrix where one route survives and the other does not) states the rule quantitatively: what matters is whether $u\kappa_2(\mathbf{A})^2$ is small enough for the purpose.

(e) It bounds the **backward** error of applying one reflection by the formula Eq. 24.3.3: the computed output is the exact output for an input perturbed by at most $(10m + 16)u$ relatively. It does not depend on any condition number, only on m and u.

24.3.6.2 B. Practice

Exercise. B1: Reading the bound

Let $\mathbf{A} \in M_{50\times 3}(\mathbb{R})$ have $\kappa_2(\mathbf{A}) = 10^5$, and suppose the data are accurate to $\varepsilon = 10^{-10}$ relatively. Using Eq. 24.3.2 with $u \approx 10^{-16}$ where needed:

(a) Bound the relative error of the exact least-squares solution caused by the data error, when the residual is zero.

(b) Do the same when $\tan\theta = 1$.

(c) In case (b), how many digits does the normal-equation route lose on top of the data error, and how many does a backward stable method lose?

Solution.

(a) With $\theta = 0$, Eq. 24.3.2 gives $2\kappa_2(\mathbf{A})\varepsilon = 2 \cdot 10^5 \cdot 10^{-10} = 2 \cdot 10^{-5}$: about five correct digits, and no algorithm can restore them.

(b) With $\tan\theta = 1$ we have $\cos\theta = 1/\sqrt{2}$, so the bracket is $\kappa_2(1 + \sqrt{2}) + \kappa_2^2 = 2.4 \cdot 10^5 + 10^{10} \approx 10^{10}$, and the bound is $10^{10} \cdot 10^{-10} = 1$. The data alone determine nothing: the $\kappa_2^2\tan\theta$ term has swallowed the problem.

(c) The arithmetic contributes $u\kappa_2(\mathbf{A})^2 = 10^{-16} \cdot 10^{10} = 10^{-6}$ for the normal equations and $u\,(\kappa_2 + \kappa_2^2\tan\theta) \approx 10^{-6}$ for a backward stable method as well, by Theorem 24.3.2. In this case the two lose the same six digits — because the κ_2^2 is the problem's, not the algorithm's. Both are far below the 10^0 already lost to the data, so the choice of method is irrelevant here; it is for $\tan\theta$ small that it decides everything.

Exercise. B2: Where the product goes singular

Let $\mathbf{A} = \left(\begin{smallmatrix} 1 & 1 \\ \varepsilon & 0 \\ 0 & \varepsilon \end{smallmatrix}\right)$ as in Example (A matrix where one route survives and the other does not), in the model with $u = 2^{-53}$.

(a) For which ε does $\mathrm{fl}(1+\varepsilon^2) = 1$ hold, and what is $\mathrm{fl}(\mathbf{A}^\top\mathbf{A})$ then?

(b) Express the threshold in (a) in terms of $\kappa_2(\mathbf{A})$ and u.

(c) Explain why the reflection route is untroubled at that threshold.

Solution.

(a) By Definition 24.1.1, $\mathrm{fl}(1+\delta) = 1$ exactly when δ is at most the largest quantity that rounds away beside 1, which is $u = 2^{-53} \approx 1.11 \cdot 10^{-16}$. So $\mathrm{fl}(1+\varepsilon^2) = 1$ for $\varepsilon^2 \le u$, that is $\varepsilon \lesssim 1.05 \cdot 10^{-8}$. Then every entry of $\mathrm{fl}(\mathbf{A}^\top\mathbf{A})$ is 1 and the computed matrix is $\left(\begin{smallmatrix} 1 & 1 \\ 1 & 1 \end{smallmatrix}\right)$, which is singular: the information distinguishing the two columns has been rounded away.

(b) From Example (A matrix where one route survives and the other does not), $\kappa_2(\mathbf{A}) = \sqrt{2+\varepsilon^2}/\varepsilon \approx \sqrt{2}/\varepsilon$, so $\varepsilon^2 \le u$ says $\kappa_2(\mathbf{A})^2 \ge 2/u$, that is $u\kappa_2(\mathbf{A})^2 \ge 2$. The normal equations collapse exactly when the bound of Theorem 24.3.2 (a) exceeds 1, which is the same as saying that the route promises no correct digit.

(c) The reflection route never computes $1+\varepsilon^2$. It works with $\mathbf{A}$ itself, whose entries record ε and not ε^2, and by Theorem 24.3.3 each reflection perturbs the data relatively by $O(u)$. A relative perturbation of 10^{-16} in $\varepsilon = 10^{-9}$ leaves ε intact; a relative perturbation of 10^{-16} in 1 destroys $\varepsilon^2 = 10^{-18}$. Squaring is what moves the small quantity below the rounding level.

Exercise. B3: One reflection, by hand

Let $\mathbf{x} = (3, 0, 4) \in \mathbb{R}^3$.

(a) Compute $\mathbf{w}$ and $\mathbf{w}'$ as in Theorem 11.8.3, and the two amplification factors $\|\mathbf{x}\|/\|\mathbf{w}\|$ and $\|\mathbf{x}\|/\|\mathbf{w}'\|$.

(b) Verify the identity $\|\mathbf{w}\|^2\|\mathbf{w}'\|^2 = 4\|\mathbf{x}\|^2\|\widetilde{\mathbf{x}}\|^2$ on this example.

(c) Is the sign choice important for this $\mathbf{x}$? Justify your answer.

Solution.

(a) $\|\mathbf{x}\| = 5$ and $x_1 = 3 > 0$, so $\alpha = 1$, $\mathbf{w} = (8, 0, 4)$, and $\mathbf{w}' = (-2, 0, 4)$. Then $\|\mathbf{w}\|^2 = 80$ and $\|\mathbf{w}'\|^2 = 20$, so

$$\frac{\|\mathbf{x}\|}{\|\mathbf{w}\|} = \frac{5}{\sqrt{80}} = \frac{\sqrt{5}}{4} \approx 0.559, \qquad \frac{\|\mathbf{x}\|}{\|\mathbf{w}'\|} = \frac{5}{\sqrt{20}} = \frac{\sqrt{5}}{2} \approx 1.118.$$

Both are of order 1. (The general formulas agree: $\|\mathbf{w}\|^2 = 2 \cdot 5 \cdot (5+3) = 80$ and $\|\mathbf{w}'\|^2 = 2 \cdot 5 \cdot (5-3) = 20$.)

(b) $\widetilde{\mathbf{x}} = (0, 4)$ has $\|\widetilde{\mathbf{x}}\|^2 = 16$, and $4\|\mathbf{x}\|^2\|\widetilde{\mathbf{x}}\|^2 = 4 \cdot 25 \cdot 16 = 1600 = 80 \cdot 20$. The identity is $\|\mathbf{w}\|^2\|\mathbf{w}'\|^2 = 4\|\mathbf{x}\|^2(\|\mathbf{x}\| + x_1)(\|\mathbf{x}\| - x_1) = 4\|\mathbf{x}\|^2\|\widetilde{\mathbf{x}}\|^2$.

(c) No. The amplification factor of the bad sign is $(\|\mathbf{x}\| + x_1)/\|\widetilde{\mathbf{x}}\| = 8/4 = 2$, so the two choices differ by a factor of 2 in accuracy, which is nothing. The sign matters only when $\mathbf{x}$ is close to a multiple of $\mathbf{e}_1$, that is when $\|\widetilde{\mathbf{x}}\| \ll \|\mathbf{x}\|$; here $\|\widetilde{\mathbf{x}}\| = 4$ is comparable with $\|\mathbf{x}\| = 5$.

24.3.6.3 C. Going deeper

Exercise. C1: The squared term is attained

Let $\varepsilon, \delta > 0$ and take

$$\mathbf{A} = \begin{pmatrix} 1 & 0 \\ 0 & \varepsilon \\ 0 & 0 \end{pmatrix}, \qquad \mathbf{b} = \begin{pmatrix} 1 \\ 0 \\ \delta \end{pmatrix}, \qquad \mathbf{E} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 1 \end{pmatrix}, \qquad \mathbf{f} = \mathbf{0}.$$

(a) Compute $\kappa_2(\mathbf{A})$, the least-squares solution $\mathbf{x}$, the residual $\mathbf{r}$, and $\tan\theta$.

(b) Using $(*)$ in the proof of Theorem 24.3.1, compute $\mathbf{x}'(0)$ exactly.

(c) Hence show that the squared term of Theorem 24.3.1 is attained, in the sense that $\|\mathbf{x}'(0)\|/\|\mathbf{x}\|$ equals $\kappa_2(\mathbf{A})^2 \varepsilon_{\mathbf{A}} \tan\theta$ exactly, and explain what this says about any algorithm for this problem.

Solution.

(a) $\mathbf{A}^\top\mathbf{A} = \operatorname{diag}(1, \varepsilon^2)$, so the singular values are 1 and ε and $\kappa_2(\mathbf{A}) = 1/\varepsilon$ by the rectangular clause of Definition 16.8.1. The normal equations $\operatorname{diag}(1, \varepsilon^2)\mathbf{x} = \mathbf{A}^\top\mathbf{b} = (1, 0)$ give $\mathbf{x} = (1, 0)$. Then $\mathbf{A}\mathbf{x} = (1, 0, 0)$, so $\mathbf{r} = \mathbf{b} - \mathbf{A}\mathbf{x} = (0, 0, \delta)$ and, by Eq. 24.3.1,

$$\tan\theta = \frac{\|\mathbf{r}\|}{\|\mathbf{A}\mathbf{x}\|} = \delta.$$

Also $\|\mathbf{E}\| = 1 = \|\mathbf{A}\|$, so $\varepsilon_{\mathbf{A}} = 1$, and $\varepsilon_{\mathbf{b}} = 0$.

(b) The three terms of $(*)$. First, $\mathbf{E}^\top\mathbf{r} = \mathbf{e}_2(\mathbf{e}_3^\top\mathbf{r}) = \delta\,\mathbf{e}_2$, and $(\mathbf{A}^\top\mathbf{A})^{-1} = \operatorname{diag}(1, \varepsilon^{-2})$, so the first term is $(\delta/\varepsilon^2)\mathbf{e}_2$. Second, $\mathbf{E}\mathbf{x} = \mathbf{e}_3(\mathbf{e}_2^\top\mathbf{x}) = \mathbf{0}$ because $x_2 = 0$, so the second term vanishes. Third, $\mathbf{f} = \mathbf{0}$. Hence

$$\mathbf{x}'(0) = \Big(0, \frac{\delta}{\varepsilon^2}\Big).$$

(c) $\|\mathbf{x}\| = 1$ and $\|\mathbf{x}'(0)\| = \delta/\varepsilon^2$, while

$$\kappa_2(\mathbf{A})^2\,\varepsilon_{\mathbf{A}} \tan\theta = \frac{1}{\varepsilon^2} \cdot 1 \cdot \delta = \frac{\delta}{\varepsilon^2}.$$

The two agree exactly: the squared term is **attained**. The bound as a whole is not, and the reason is that the other two terms are nonnegative — here the first of them contributes a further $\kappa_2(\mathbf{A})\varepsilon_{\mathbf{A}} = 1/\varepsilon$, so the bound exceeds $\|\mathbf{x}'(0)\|/\|\mathbf{x}\|$ by exactly that much. It is the squared term that matters, and what it says is that the sensitivity is a property of the data $(\mathbf{A}, \mathbf{b})$ alone: a perturbation of $\mathbf{A}$ of relative size $\varepsilon_{\mathbf{A}}$ really does move the exact answer by $\kappa_2(\mathbf{A})^2 \tan\theta\,\varepsilon_{\mathbf{A}}$ relatively, so **no** algorithm, however stable, can return an answer better than that from data known only to within $\varepsilon_{\mathbf{A}}$. Compare Theorem 24.3.2 (b): a backward stable method achieves this and no more.

Exercise. C2: Why the residual has to appear

Let $\mathbf{A} \in M_{m\times n}(F)$ have rank n, and let $\mathbf{x}^+ = \mathbf{A}^+\mathbf{b}$ be the least-squares solution, where $\mathbf{A}^+$ is the pseudoinverse of Definition 13.12.1.

(a) Prove that $\mathbf{A}^+ = (\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^*$ when $\operatorname{rank}\mathbf{A} = n$, and deduce $\|\mathbf{A}^+\| = 1/\sigma_n(\mathbf{A})$.

(b) Deduce from $(*)$ in the proof of Theorem 24.3.1 that if $\mathbf{E} = \mathbf{O}$ — only $\mathbf{b}$ is perturbed — then no $\kappa_2(\mathbf{A})^2$ term can appear, whatever the residual.

(c) Explain, in one or two sentences, why perturbing $\mathbf{A}$ is different: which term of $(*)$ carries the residual, and why it cannot be present when $\mathbf{E} = \mathbf{O}$.

Solution.

(a) Write $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$ (Theorem 13.9.1). Since $\operatorname{rank}\mathbf{A} = n$, $\sigma_1, \dots, \sigma_n > 0$, so $\mathbf{\Sigma}^+$ of Definition 13.12.1 has σ_i^{-1} on its diagonal and $\mathbf{A}^+ = \mathbf{V}\mathbf{\Sigma}^+\mathbf{U}^*$. On the other hand $\mathbf{A}^*\mathbf{A} = \mathbf{V}(\mathbf{\Sigma}^*\mathbf{\Sigma})\mathbf{V}^* = \mathbf{V}\operatorname{diag}(\sigma_i^2)\mathbf{V}^*$, so

$$(\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^* = \mathbf{V}\operatorname{diag}(\sigma_i^{-2})\mathbf{V}^*\,\mathbf{V}\mathbf{\Sigma}^*\mathbf{U}^* = \mathbf{V}(\operatorname{diag}(\sigma_i^{-2})\mathbf{\Sigma}^*)\mathbf{U}^* = \mathbf{V}\mathbf{\Sigma}^+\mathbf{U}^* = \mathbf{A}^+,$$

because $\operatorname{diag}(\sigma_i^{-2})\mathbf{\Sigma}^*$ is the $n\times m$ matrix with entries $\sigma_i^{-2}\sigma_i = \sigma_i^{-1}$ on the diagonal and zeros elsewhere. Unitary factors leave singular values alone — $(\mathbf{V}\mathbf{\Sigma}^+\mathbf{U}^*)^*(\mathbf{V}\mathbf{\Sigma}^+\mathbf{U}^*) = \mathbf{U}(\mathbf{\Sigma}^{+*}\mathbf{\Sigma}^+)\mathbf{U}^*$ is unitarily similar to $\mathbf{\Sigma}^{+*}\mathbf{\Sigma}^+$, so the two have the same eigenvalues and Definition 13.9.1 reads the same singular values off them — so $\|\mathbf{A}^+\| = \|\mathbf{\Sigma}^+\| = \max_i \sigma_i^{-1} = 1/\sigma_n$ by Theorem 16.3.3 (c).

(b) With $\mathbf{E} = \mathbf{O}$, $(*)$ reduces to $\mathbf{x}'(0) = \mathbf{A}^+\mathbf{f}$, so by (a)

$$\frac{\|\mathbf{x}'(0)\|}{\|\mathbf{x}\|} \le \frac{\|\mathbf{f}\|}{\sigma_n\|\mathbf{x}\|} = \kappa_2(\mathbf{A})\,\varepsilon_{\mathbf{b}}\,\frac{\|\mathbf{b}\|}{\sigma_1\|\mathbf{x}\|} \le \frac{\kappa_2(\mathbf{A})\,\varepsilon_{\mathbf{b}}}{\cos\theta},$$

a single factor of $\kappa_2(\mathbf{A})$ however large $\|\mathbf{r}\|$ is. (The residual enters only through $1/\cos\theta$, which is bounded as long as θ is bounded away from $\pi/2$.)

(c) The residual sits in the term $(\mathbf{A}^*\mathbf{A})^{-1}\mathbf{E}^*\mathbf{r}$, and that is the only term with $(\mathbf{A}^*\mathbf{A})^{-1}$ *not* paired with an $\mathbf{A}^*$ — which is exactly why it carries σ_n^{-2} rather than σ_n^{-1}. It is proportional to $\mathbf{E}$, so it vanishes when $\mathbf{E} = \mathbf{O}$. Perturbing $\mathbf{b}$ tilts the target inside a fixed column space; perturbing $\mathbf{A}$ tilts the column space itself, and the residual is the lever arm by which that tilt moves the answer.

Exercise. C3: A reflection applied to a matrix

Let $\mathbf{H} = \mathbf{H}_{\mathbf{w}}$ be a Householder reflection in $\mathbb{R}^m$ and let $\mathbf{A} \in M_{m\times n}(\mathbb{R})$.

(a) Using Theorem 24.3.3 (b) and $\|\mathbf{M}\| \le \|\mathbf{M}\|_F \le \sqrt{n}\,\|\mathbf{M}\|$ for $\mathbf{M} \in M_{m\times n}$, show that the computed $\widehat{\mathbf{B}}$ satisfies $\|\widehat{\mathbf{B}} - \mathbf{H}\mathbf{A}\| \le (10m+16)\sqrt{n}\,u\,\|\mathbf{A}\|$.

(b) Deduce that $\sigma_i(\widehat{\mathbf{B}})$ differs from $\sigma_i(\mathbf{A})$ by at most $(10m+16)\sqrt{n}\,u\,\|\mathbf{A}\|$ for every i. *Hint: Chapter 20 §10 or Chapter 17 §09 bounds the perturbation of singular values by the norm of the perturbation.*

(c) Explain why the corresponding statement is false for the classical Gram–Schmidt process, citing Example (Three columns and three answers).

Solution.

(a) By Theorem 24.3.3 (b), $\widehat{\mathbf{B}} - \mathbf{HA} = \mathbf{H}\,\Delta\mathbf{A}$ with $\|\Delta\mathbf{A}\|_F \le (10m+16)u\|\mathbf{A}\|_F$. Since $\mathbf{H}$ is unitary (Proposition 11.8.2 (c)) it leaves the 2-norm unchanged, so

$$\|\widehat{\mathbf{B}} - \mathbf{HA}\| = \|\Delta\mathbf{A}\| \le \|\Delta\mathbf{A}\|_F \le (10m+16)u\|\mathbf{A}\|_F \le (10m+16)\sqrt{n}\,u\,\|\mathbf{A}\|.$$

(b) $\mathbf{HA}$ has the same singular values as $\mathbf{A}$: $(\mathbf{HA})^*(\mathbf{HA}) = \mathbf{A}^*\mathbf{H}^*\mathbf{HA} = \mathbf{A}^*\mathbf{A}$, so Definition 13.9.1 reads the same list off the same matrix. Weyl's inequality for singular values (Theorem 17.9.4 with $j = 1$, or Corollary 17.9.5 directly) gives $|\sigma_i(\widehat{\mathbf{B}}) - \sigma_i(\mathbf{HA})| \le \|\widehat{\mathbf{B}} - \mathbf{HA}\|$ for every i, and the right-hand side is bounded in (a).

(c) The Gram–Schmidt recursion of Theorem 11.2.2 does not apply a unitary matrix to $\mathbf{A}$; it *builds* a matrix $\widehat{\mathbf{Q}}$ out of $\mathbf{A}$ and hopes it is unitary. Example (Three columns and three answers) measures the damage: for $\kappa_2(\mathbf{A}) = 3 \cdot 10^6$ the computed vectors have inner products of size $9 \cdot 10^{-5}$, so $\widehat{\mathbf{Q}}$ is unitary only to five digits, and $\|\widehat{\mathbf{Q}}^\top\widehat{\mathbf{Q}} - \mathbf{I}\|$ exceeds u by eleven orders of magnitude on that one example. (That it does so at the rate $u\kappa_2(\mathbf{A})^2$ in general is the classical estimate quoted above, which this book does not prove.) Any conclusion drawn from "unitary factors leave singular values alone" then inherits that error instead of none. The reflection route wins because the unitary matrix is never in question: it is exactly unitary by construction, and only its *input* is perturbed.

24.4 The Power Method

Chapter 19 computed a Perron vector three times and each time stopped at the same place. Section 4 bracketed $\rho(\mathbf{A})$ by the ratios of $\mathbf{Ax}$ to $\mathbf{x}$ and observed: "the iteration $\mathbf{x} \mapsto \mathbf{Ax}$ is the power method, and Chapter 24 studies how fast it converges." Section 8 said the same of PageRank — "Chapter 24 studies it for general matrices, where the rate is governed by the ratio of the two largest eigenvalue moduli" — and Section 9's ledger listed "the power method, the algorithm that computes a Perron vector at the size of the web" among the debts. Chapter 17 §01 and Chapter 20 §06 each promised, in the same words, that "Chapter 24 builds an iteration on" the Rayleigh quotient.

This section pays all five. It is about one idea used three times: **multiply by a matrix repeatedly and the largest eigenvalue takes over.** Applied to $\mathbf{A}$ itself, the idea finds the eigenvalue of largest modulus. Applied to $(\mathbf{A} - \mu\mathbf{I})^{-1}$, it finds the eigenvalue nearest μ — any eigenvalue we choose to aim at. Applied to $(\mathbf{A} - \mu\mathbf{I})^{-1}$ with μ refreshed at every step from the Rayleigh quotient, it converges cubically, and that is the iteration Chapters 17 and 20 promised.

Throughout, $F = \mathbb{R}$ or $F = \mathbb{C}$, matrices are in $M_n(F)$ read inside $M_n(\mathbb{C})$, and $\|\cdot\| = \|\cdot\|_2$. Iteration indices are superscripts in parentheses, so that subscripts stay free for coordinates.

24.4.1 Multiplying by the matrix

Take $\mathbf{A}$ diagonalizable with eigenbasis $\mathbf{v}_1, \dots, \mathbf{v}_n$ and eigenvalues $\lambda_1, \dots, \lambda_n$, and expand a starting vector:

$$\mathbf{x} = c_1\mathbf{v}_1 + \cdots + c_n\mathbf{v}_n \quad \Longrightarrow \quad \mathbf{A}^k\mathbf{x} = c_1\lambda_1^k\mathbf{v}_1 + \cdots + c_n\lambda_n^k\mathbf{v}_n.$$

Divide by λ_1^k, and suppose $|\lambda_1|$ is strictly the largest. Every coefficient except the first is multiplied by $(\lambda_i/\lambda_1)^k$, a number of modulus less than 1 raised to the k-th power. So the sum collapses onto its first term. Nothing here needs to be discovered; it needs to be named, and its rate measured.

The power method is: multiply, rescale, repeat, and read off the dominant eigenvector.

Definition 24.4.1. Power Iteration

Let $\mathbf{A} \in M_n(F)$ and let $\mathbf{x}^{(0)} \in F^n$ with $\|\mathbf{x}^{(0)}\| = 1$. The **power iteration** generated by $\mathbf{A}$ and $\mathbf{x}^{(0)}$ is the sequence defined for $k \geq 0$ by

$$\mathbf{x}^{(k+1)} = \frac{\mathbf{A}\mathbf{x}^{(k)}}{\|\mathbf{A}\mathbf{x}^{(k)}\|},$$

which is defined **as long as** $\mathbf{A}\mathbf{x}^{(k)} \neq \mathbf{0}$, together with the **eigenvalue estimates**

$$\mu_k = (\mathbf{x}^{(k)})^* \mathbf{A}\, \mathbf{x}^{(k)}.$$

Clause by clause. The vectors are **unit** vectors: the normalization has no mathematical content, since $\mathbf{x}^{(k)}$ spans the same line as $\mathbf{A}^k\mathbf{x}^{(0)}$, and its only purpose is to stop the entries from growing or shrinking like λ_1^k until the arithmetic can no longer hold them. The iteration is **partial**: if $\mathbf{A}\mathbf{x}^{(k)} = \mathbf{0}$ the sequence stops, and it stops exactly when $\mathbf{x}^{(k)} \in \mathrm{nul}(\mathbf{A})$ — which, as we shall see, a good starting vector avoids. The estimate μ_k is the natural number to attach to a unit vector $\mathbf{x}$ that is nearly an eigenvector: if $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$ exactly then $\mathbf{x}^*\mathbf{A}\mathbf{x} = \lambda$. **For Hermitian $\mathbf{A}$ it is the Rayleigh quotient $R_{\mathbf{A}}(\mathbf{x}^{(k)})$ of Definition 17.1.1**, whose denominator is 1 here; for a general $\mathbf{A}$ the book reserves that name for the Hermitian case, so we write the quadratic form out.

Examples.

- **The identity.** $\mathbf{A} = \mathbf{I}$: every unit vector is fixed, $\mathbf{x}^{(k)} = \mathbf{x}^{(0)}$ for all k, and $\mu_k = 1$. The degenerate case is worth a glance, because it shows that the method converges to *an* eigenvector, never to a distinguished one.
- **A diagonal matrix.** $\mathbf{A} = \mathrm{diag}(3, 1)$ and $\mathbf{x}^{(0)} = \frac{1}{\sqrt{2}}(1, 1)$: then $\mathbf{A}^k\mathbf{x}^{(0)} \propto (3^k, 1)$, and the normalized vector runs to $\mathbf{e}_1$ with second coordinate 3^{-k}. The rate 3^{-k} is $|\lambda_2/\lambda_1|^k$.
- **A stochastic matrix.** For the Google matrix $\mathbf{G}$ of Chapter 19 §08, $\lambda_1 = 1$ and every other eigenvalue has modulus at most α (Theorem 19.8.2), so the iteration $\mathbf{x}_{k+1} = \mathbf{G}\mathbf{x}_k$ of Proposition 19.8.3 is a power iteration with rate at most α. Chapter 19 proved the rate directly from stochasticity, with no hypothesis on $\mathbf{G}$ beyond that; the theorem below proves it from the eigenvalues, for any diagonalizable matrix.

Nonexample by minimal change. Replace $\mathrm{diag}(3, 1)$ by $\mathrm{diag}(3, -3)$. The eigenvalues now have **equal modulus**, the collapse never happens, and from $\mathbf{x}^{(0)} = \frac{1}{\sqrt{2}}(1, 1)$ the iterates alternate between $\frac{1}{\sqrt{2}}(1, 1)$ and $\frac{1}{\sqrt{2}}(1, -1)$ forever. One clause of the hypothesis has been dropped — strictness in $|\lambda_1| > |\lambda_2|$ — and the conclusion is gone with it.

Here is the theorem, with the rate.

Theorem 24.4.1. The Power Method

Let $\mathbf{A} \in M_n(F)$ be diagonalizable, with a basis $\mathbf{v}_1, \ldots, \mathbf{v}_n$ of eigenvectors, $\mathbf{A}\mathbf{v}_i = \lambda_i\mathbf{v}_i$,

ordered by **modulus** so that

$$|\lambda_1| > |\lambda_2| \geq \cdots \geq |\lambda_n|.$$

Let $\mathbf{x}^{(0)} = c_1\mathbf{v}_1 + \cdots + c_n\mathbf{v}_n$ be a unit vector with $c_1 \neq 0$, and put $\rho = |\lambda_2/\lambda_1| < 1$, $\mathbf{q}_1 = \mathbf{v}_1/\|\mathbf{v}_1\|$ and

$$C = \frac{1}{\|\mathbf{v}_1\|}\sum_{i\geq 2}\Big|\frac{c_i}{c_1}\Big|\,\|\mathbf{v}_i\|.$$

Then the power iteration of Definition 24.4.1 is defined for every k, and for every k with $C\rho^k \leq \frac{1}{2}$ there is a scalar ω_k of modulus 1 with

(a) $\|\omega_k\mathbf{x}^{(k)} - \mathbf{q}_1\| \leq 4C\rho^k$;
(b) $|\mu_k - \lambda_1| \leq 8\|\mathbf{A}\|\,C\rho^k$.

In particular $\mu_k \to \lambda_1$, and the iterates approach the eigenline $\operatorname{span}(\mathbf{v}_1)$ in the sense of (a), both at the rate ρ^k.

Idea. The whole content is the display at the head of this subsection: divide $\mathbf{A}^k\mathbf{x}^{(0)}$ by $c_1\lambda_1^k$ and what is left is $\mathbf{v}_1$ plus a tail of size ρ^k. Two pieces of bookkeeping then stand between that and the statement. First, normalizing: a bound on $\mathbf{y} - \mathbf{v}_1$ is not yet a bound on $\mathbf{y}/\|\mathbf{y}\| - \mathbf{v}_1/\|\mathbf{v}_1\|$, and the conversion costs a factor 2 and the hypothesis $C\rho^k \leq \frac{1}{2}$, which only says that we are past the point where the tail could swamp $\mathbf{v}_1$. Second, phases: over $\mathbb{C}$ the scalar $c_1\lambda_1^k$ has a modulus and an argument, and normalizing kills the modulus but not the argument. That argument is the ω_k. It cancels out of μ_k, because a quadratic form does not see it.

Proof. **The iterates are defined, and what they are.** Since $c_1\lambda_1^k \neq 0$, the vector $\mathbf{A}^k\mathbf{x}^{(0)} = \sum_i c_i\lambda_i^k\mathbf{v}_i$ has a nonzero coefficient on $\mathbf{v}_1$, so $\mathbf{A}^k\mathbf{x}^{(0)} \neq \mathbf{0}$ for every k, the eigenvectors being a basis. An induction on k then gives $\mathbf{x}^{(k)} = \mathbf{A}^k\mathbf{x}^{(0)}/\|\mathbf{A}^k\mathbf{x}^{(0)}\|$: true at $k = 0$, and if it holds at k then $\mathbf{A}\mathbf{x}^{(k)}$ is a positive multiple of $\mathbf{A}^{k+1}\mathbf{x}^{(0)} \neq \mathbf{0}$, so normalizing it gives the same unit vector as normalizing $\mathbf{A}^{k+1}\mathbf{x}^{(0)}$.

The tail. Put

$$\mathbf{y}_k = \frac{\mathbf{A}^k\mathbf{x}^{(0)}}{c_1\lambda_1^k} = \mathbf{v}_1 + \mathbf{e}_k, \qquad \mathbf{e}_k = \sum_{i\geq 2}\frac{c_i}{c_1}\Big(\frac{\lambda_i}{\lambda_1}\Big)^k\mathbf{v}_i.$$

Since $|\lambda_i/\lambda_1| \leq \rho$ for $i \geq 2$, the triangle inequality (Corollary 11.1.6) gives

$$\|\mathbf{e}_k\| \leq \rho^k\sum_{i\geq 2}\Big|\frac{c_i}{c_1}\Big|\|\mathbf{v}_i\| = C\rho^k\|\mathbf{v}_1\|.$$

Normalizing. Let $\omega_k = \overline{\sigma_k}$, where $\sigma_k = c_1\lambda_1^k/|c_1\lambda_1^k|$ has modulus 1. Then $\mathbf{A}^k\mathbf{x}^{(0)} = |c_1\lambda_1^k|\,\sigma_k\mathbf{y}_k$, so $\mathbf{x}^{(k)} = \sigma_k\mathbf{y}_k/\|\mathbf{y}_k\|$ and $\omega_k\mathbf{x}^{(k)} = \mathbf{y}_k/\|\mathbf{y}_k\|$. For any nonzero $\mathbf{a}, \mathbf{b}$,

$$\frac{\mathbf{a}}{\|\mathbf{a}\|} - \frac{\mathbf{b}}{\|\mathbf{b}\|} = \frac{\mathbf{a}-\mathbf{b}}{\|\mathbf{a}\|} + \mathbf{b}\Big(\frac{1}{\|\mathbf{a}\|} - \frac{1}{\|\mathbf{b}\|}\Big),$$

$$\Big|\frac{1}{\|\mathbf{a}\|} - \frac{1}{\|\mathbf{b}\|}\Big|\|\mathbf{b}\| = \frac{\big|\|\mathbf{b}\| - \|\mathbf{a}\|\big|}{\|\mathbf{a}\|} \leq \frac{\|\mathbf{a}-\mathbf{b}\|}{\|\mathbf{a}\|},$$

the last step by the reverse triangle inequality (Lemma 16.1.1). Hence $\|\mathbf{a}/\|\mathbf{a}\| - \mathbf{b}/\|\mathbf{b}\|\| \le 2\|\mathbf{a}-\mathbf{b}\|/\|\mathbf{a}\|$. Apply this with $\mathbf{a} = \mathbf{y}_k$ and $\mathbf{b} = \mathbf{v}_1$. If $C\rho^k \le \tfrac12$ then $\|\mathbf{e}_k\| \le \tfrac12\|\mathbf{v}_1\|$, so $\|\mathbf{y}_k\| \ge \|\mathbf{v}_1\| - \|\mathbf{e}_k\| \ge \tfrac12\|\mathbf{v}_1\|$, and

$$\|\omega_k \mathbf{x}^{(k)} - \mathbf{q}_1\| \le \frac{2\|\mathbf{e}_k\|}{\|\mathbf{y}_k\|} \le \frac{4\|\mathbf{e}_k\|}{\|\mathbf{v}_1\|} \le 4C\rho^k,$$

which is (a).

The estimates. Write $\mathbf{z} = \omega_k \mathbf{x}^{(k)}$, a unit vector, and note $\mu_k = (\mathbf{x}^{(k)})^* \mathbf{A}\mathbf{x}^{(k)} = \mathbf{z}^*\mathbf{A}\mathbf{z}$ because $\overline{\omega_k}\omega_k = 1$. Since $\mathbf{A}\mathbf{q}_1 = \lambda_1\mathbf{q}_1$ and $\|\mathbf{q}_1\| = 1$, we have $\mathbf{q}_1^*\mathbf{A}\mathbf{q}_1 = \lambda_1$, so

$$\mu_k - \lambda_1 = \mathbf{z}^*\mathbf{A}(\mathbf{z}-\mathbf{q}_1) + (\mathbf{z}-\mathbf{q}_1)^*\mathbf{A}\mathbf{q}_1.$$

By Cauchy–Schwarz (Theorem 11.1.5) and Theorem 16.3.2 (a), each term is at most $\|\mathbf{A}\|\|\mathbf{z}-\mathbf{q}_1\|$, so $|\mu_k - \lambda_1| \le 2\|\mathbf{A}\|\|\mathbf{z}-\mathbf{q}_1\| \le 8\|\mathbf{A}\|C\rho^k$ by (a). This proves the theorem.

■

Aftermath. One matrix–vector product per step: $2n^2$ flops for a dense $\mathbf{A}$, and far fewer when $\mathbf{A}$ is sparse — for the hyperlink matrix of Chapter 19 §08, one operation per link. Nothing is factorized and nothing is stored but a vector, which is why the method survives at sizes where the matrix itself could never be written down. What it buys with that cheapness is a *linear* rate: to gain one decimal digit takes $1/\log_{10}(1/\rho)$ steps, and for $\rho = 0.99$ that is 230 of them.

Example. Every iterate exactly

Let $\mathbf{A} = \left(\begin{smallmatrix} 2 & 1 \\ 1 & 2 \end{smallmatrix}\right)$ and $\mathbf{x}^{(0)} = \mathbf{e}_1$. Compute $\mathbf{A}^k\mathbf{e}_1$, the estimates μ_k, and the errors, in closed form.

Solution. $\mathbf{A}$ has eigenvectors $(1,1)$ and $(1,-1)$ with eigenvalues 3 and 1, so $\rho = \tfrac13$. Writing $\mathbf{e}_1 = \tfrac12(1,1) + \tfrac12(1,-1)$,

$$\mathbf{A}^k\mathbf{e}_1 = \tfrac12 3^k(1,1) + \tfrac12(1,-1) = \Big(\frac{3^k+1}{2},\ \frac{3^k-1}{2}\Big).$$

So the iterates are proportional to $(1,0), (2,1), (5,4), (14,13), (41,40), (122,121), \ldots$, visibly running to the direction $(1,1)$. Put $t = 3^k$ and $\mathbf{x} = (a,b)$ with $a = (t+1)/2$, $b = (t-1)/2$. Then $a^2 + b^2 = (t^2+1)/2$ and $ab = (t^2-1)/4$, so

$$\mu_k = \frac{2a^2 + 2ab + 2b^2}{a^2+b^2} = \frac{(t^2+1) + \tfrac12(t^2-1)}{\tfrac12(t^2+1)} = \frac{3t^2+1}{t^2+1} = 3 - \frac{2}{9^k+1}.$$

The values are $2,\ \tfrac{14}{5},\ \tfrac{122}{41},\ \tfrac{1094}{365},\ \tfrac{9842}{3281}$, with errors $1,\ \tfrac15,\ \tfrac{1}{41},\ \tfrac{1}{365},\ \tfrac{1}{3281}$ — falling by factors $5,\ 8.2,\ 8.9,\ 9.0$, so by about a ninth in the limit, not a third.

The eigenvector error tells the other half. The component of $\mathbf{x} = (a,b)$ orthogonal to $(1,1)$ is $\tfrac12(a-b)(1,-1)$, of length $|a-b|/\sqrt{2} = 1/\sqrt{2}$, and $\|\mathbf{x}\| = \sqrt{(t^2+1)/2}$, so the sine of the angle between $\mathbf{x}^{(k)}$ and $(1,1)$ is $(9^k+1)^{-1/2}$, which does shrink like $3^{-k} = \rho^k$. Exactly:

$$3 - \mu_k = \frac{2}{9^k+1} = 2\sin^2\theta_k.$$

The eigenvector converges at rate ρ^k and the eigenvalue estimate at rate ρ^{2k}, and the second is the square of the first. That is not an accident of this matrix.

The corollary explains it, and pays Chapter 17 §01 in its own words.

Corollary 24.4.2. Hermitian Matrices: the Rayleigh Quotient Converges Twice as Fast

Let $\mathbf{A} \in M_n(F)$ be Hermitian, with an orthonormal basis $\mathbf{q}_1, \dots, \mathbf{q}_n$ of eigenvectors, $\mathbf{A}\mathbf{q}_i = \lambda_i \mathbf{q}_i$, ordered by modulus so that $|\lambda_1| > |\lambda_2| \geq \cdots \geq |\lambda_n|$. Let $\mathbf{x}^{(0)} = \sum_i c_i \mathbf{q}_i$ be a unit vector with $c_1 \neq 0$, and put $\rho = |\lambda_2/\lambda_1|$. Then for every k,

$$|R_{\mathbf{A}}(\mathbf{x}^{(k)}) - \lambda_1| \;\leq\; D\,\rho^{2k}, \qquad D = \frac{1}{|c_1|^2}\sum_{i\geq 2} |c_i|^2\,|\lambda_i - \lambda_1|.$$

Idea. The Rayleigh quotient is unchanged by scaling, so evaluate it at the unnormalized $\mathbf{A}^k\mathbf{x}^{(0)}$, where the coefficients are just $c_i\lambda_i^k$. Both the numerator and the denominator are then sums of *nonnegative* multiples of λ_i^{2k} — the exponent is even, which is the whole point — so the denominator is bounded below by its first term alone, and the numerator, after λ_1 is subtracted, loses its first term entirely. What is left is λ_i^{2k} against λ_1^{2k}: one factor of ρ^2 per step.

Proof. Write $\mathbf{y} = \mathbf{A}^k\mathbf{x}^{(0)} = \sum_i c_i\lambda_i^k\mathbf{q}_i$. Since $c_1\lambda_1^k \neq 0$, $\mathbf{y} \neq \mathbf{0}$, and $\mathbf{x}^{(k)}$ is a scalar multiple of $\mathbf{y}$ as in the proof of Theorem 24.4.1, so $R_{\mathbf{A}}(\mathbf{x}^{(k)}) = R_{\mathbf{A}}(\mathbf{y})$ by the scale-invariance of the Rayleigh quotient (Definition 17.1.1). The basis is orthonormal and $\mathbf{A}\mathbf{y} = \sum_i c_i\lambda_i^{k+1}\mathbf{q}_i$, so by Theorem 11.2.1 (b) and (c),

$$R_{\mathbf{A}}(\mathbf{y}) = \frac{\langle \mathbf{A}\mathbf{y}, \mathbf{y}\rangle}{\langle \mathbf{y}, \mathbf{y}\rangle} = \frac{\sum_i |c_i|^2\lambda_i^{2k+1}}{\sum_i |c_i|^2\lambda_i^{2k}},$$

all the λ_i being real (Proposition 11.6.7). Subtract λ_1, which is the same as subtracting λ_1 times the denominator from the numerator; the term $i = 1$ cancels, leaving

$$R_{\mathbf{A}}(\mathbf{y}) - \lambda_1 = \frac{\sum_{i\geq 2} |c_i|^2\lambda_i^{2k}(\lambda_i - \lambda_1)}{\sum_i |c_i|^2\lambda_i^{2k}}.$$

Every λ_i^{2k} is nonnegative, so the denominator is at least its first term $|c_1|^2\lambda_1^{2k} = |c_1|^2|\lambda_1|^{2k}$, and the numerator is at most $|\lambda_2|^{2k}\sum_{i\geq 2}|c_i|^2|\lambda_i - \lambda_1|$ in modulus, since $|\lambda_i| \leq |\lambda_2|$ for $i \geq 2$. Dividing gives the stated bound with $\rho^{2k} = (|\lambda_2|/|\lambda_1|)^{2k}$. This proves the corollary. ∎

This is the "second-order accuracy" of Chapter 17 §01, seen along an iteration: the eigenvector error is ρ^k and the eigenvalue error is its square. Theorem 20.6.2 says the same thing structurally and without any iteration — a residual of size ε yields an eigenvalue error of at most ε^2 divided by the gap — and the corollary above is that estimate made explicit for the particular unit vectors the power method produces.

Quick check. In Example (Every iterate exactly) the estimates μ_k increase towards 3 and never overshoot. Which earlier result says that they cannot?

Solution. Proposition 17.1.1: for a Hermitian $\mathbf{A}$, the values of $R_{\mathbf{A}}$ fill exactly the interval $[\lambda_n(\mathbf{A}), \lambda_1(\mathbf{A})]$ between the smallest and largest eigenvalues. Here $\mathbf{A} = \left(\begin{smallmatrix} 2 & 1 \\ 1 & 2 \end{smallmatrix}\right)$ has eigenvalues 3 and 1, so every $\mu_k = R_{\mathbf{A}}(\mathbf{x}^{(k)})$ lies in $[1, 3]$ and no estimate can exceed 3. The power method on a Hermitian matrix with $\lambda_1 > 0$ dominant therefore produces a sequence of certified *lower* bounds for λ_1, which is what Chapter 19 §04's brackets were doing by hand.

Example. The method that never converges

Let $\mathbf{P} \in M_3(\mathbb{R})$ be the cyclic permutation matrix of Example (The Cyclic Permutation), with columns $\mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_1$, so that $\mathbf{P}(x_1, x_2, x_3) = (x_3, x_1, x_2)$. Run the power iteration from $\mathbf{x}^{(0)} = \mathbf{e}_1$ and from $\mathbf{x}^{(0)} = \frac{1}{\sqrt{14}}(1, 2, 3)$.

Solution. By Example (The Cyclic Permutation) the eigenvalues of $\mathbf{P}$ are the three cube roots of unity, all of modulus 1: the hypothesis $|\lambda_1| > |\lambda_2|$ fails as badly as it can. Every $\mathbf{P}^k$ is a permutation matrix, so no normalization is needed, and $\mathbf{P}^3 = \mathbf{I}$. From $\mathbf{e}_1$ the iterates are

$$\mathbf{e}_1 \to \mathbf{e}_2 \to \mathbf{e}_3 \to \mathbf{e}_1 \to \dots$$

and from $\frac{1}{\sqrt{14}}(1, 2, 3)$ they are $\frac{1}{\sqrt{14}}$ times $(1, 2, 3), (3, 1, 2), (2, 3, 1), (1, 2, 3), \dots$. In both cases the sequence takes three distinct values over and over, so neither it nor the line it spans has a limit. The estimates behave worse than that: they are *constant*, $\mu_k = 0$ for every k in the first case and $\mu_k = \frac{11}{14}$ in the second, so they converge perfectly well — to a number that is not an eigenvalue of $\mathbf{P}$ at all. Yet $\mathbf{P}$ does have the eigenvalue 1, with eigenvector $\mathbf{1}$, and the method never finds it.

The failure is exactly the one Chapter 19 §04 warned of for the swap $\left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$, whose brackets "never improve". It is not a defect of the arithmetic and no amount of care repairs it: the hypothesis of Theorem 24.4.1 is false, and its conclusion is false too.

Warning. **Two hypotheses, two different kinds of failure.** If $|\lambda_1| = |\lambda_2|$, as in Example (The method that never converges), the method fails permanently and no starting vector helps. If instead $c_1 = 0$ — the starting vector happens to miss the dominant eigendirection — the method converges, but to the wrong eigenvector. The second failure is fragile and the first is not: a starting vector with $c_1 = 0$ lies in a proper subspace, and in the arithmetic of Definition 24.1.1 the first rounding error generally puts a small nonzero c_1 back, after which Theorem 24.4.1 takes over, only C being large. This is the one situation in the chapter where rounding *helps*, and it is not a reason to rely on it: a starting vector chosen without regard to $\mathbf{A}$ is a better plan than a lucky rounding.

24.4.2 Shifting and inverting

The power method finds the eigenvalue of largest modulus, and that is usually not the one anybody wants. The repair is a change of matrix. If μ is not an eigenvalue of $\mathbf{A}$, then $\mathbf{A}-\mu\mathbf{I}$ is invertible, and the eigenvalues of its inverse are the numbers $(\lambda_i-\mu)^{-1}$ — **large exactly for the λ_i close to μ**. Multiplying repeatedly by $(\mathbf{A}-\mu\mathbf{I})^{-1}$ therefore singles out the eigenvalue nearest to μ, and μ is ours to choose.

Definition 24.4.2. Inverse Iteration

Let $\mathbf{A}\in M_n(F)$, let $\mu\in F$ be **not an eigenvalue of $\mathbf{A}$**, and let $\mathbf{x}^{(0)}$ be a unit vector. **Inverse iteration with shift** μ is the power iteration of Definition 24.4.1 generated by $(\mathbf{A}-\mu\mathbf{I})^{-1}$ and $\mathbf{x}^{(0)}$; that is,

$$(\mathbf{A}-\mu\mathbf{I})\,\mathbf{z}^{(k)}=\mathbf{x}^{(k)},\qquad \mathbf{x}^{(k+1)}=\frac{\mathbf{z}^{(k)}}{\|\mathbf{z}^{(k)}\|}.$$

The eigenvalue estimates are taken from $\mathbf{A}$ itself: $\mu_k=(\mathbf{x}^{(k)})^{*}\mathbf{A}\mathbf{x}^{(k)}$.

The shifted matrix is **never inverted**. One factorization of $\mathbf{A}-\mu\mathbf{I}$ by elimination with pivoting is computed once, at a cost of $\frac{2}{3}n^3$ flops (Section 2), and each step then costs two triangular solves, $2n^2$ flops — the same price per step as the power method, after a one-off payment.

Theorem 24.4.3. Inverse Iteration Converges to the Nearest Eigenvector

Let $\mathbf{A}\in M_n(F)$ be diagonalizable with eigenbasis $\mathbf{v}_1,\dots,\mathbf{v}_n$ and eigenvalues $\lambda_1,\dots,\lambda_n$, and let $\mu\in F$ be not an eigenvalue of $\mathbf{A}$. Suppose there is an index j with

$$|\lambda_j-\mu|<|\lambda_i-\mu|\qquad\text{for every } i\neq j,$$

so that λ_j is the unique eigenvalue nearest μ. Let $\mathbf{x}^{(0)}$ be a unit vector whose $\mathbf{v}_j$-coefficient is nonzero. Then inverse iteration with shift μ is defined for every k, the iterates approach the eigenline $\operatorname{span}(\mathbf{v}_j)$ in the sense of Theorem 24.4.1 (a), and $\mu_k\to\lambda_j$, both at the rate ρ_μ^{k}, where

$$\rho_\mu=\frac{|\lambda_j-\mu|}{\min_{i\neq j}|\lambda_i-\mu|}<1.$$

Idea. There is nothing to prove that Theorem 24.4.1 has not proved. The only work is to check that $(\mathbf{A}-\mu\mathbf{I})^{-1}$ satisfies its hypotheses — same eigenvectors, reciprocal shifted eigenvalues, a strict winner in modulus — and to read off what the ratio ρ of that theorem becomes after the reciprocals are taken: dividing by the largest reciprocal is multiplying by the smallest difference, so the ratio turns upside down.

Proof. Let $\mathbf{V}=(\mathbf{v}_1\mid\cdots\mid\mathbf{v}_n)$, invertible because its columns are a basis, and $\mathbf{\Lambda}=\operatorname{diag}(\lambda_1,\dots,\lambda_n)$, so that $\mathbf{A}=\mathbf{V}\mathbf{\Lambda}\mathbf{V}^{-1}$ (Theorem 9.4.1). Then

$$\mathbf{A}-\mu\mathbf{I}=\mathbf{V}(\mathbf{\Lambda}-\mu\mathbf{I})\mathbf{V}^{-1},$$

and $\mathbf{\Lambda}-\mu\mathbf{I}$ is invertible because no λ_i equals μ. Hence $\mathbf{B}:=(\mathbf{A}-\mu\mathbf{I})^{-1}=\mathbf{V}(\mathbf{\Lambda}-\mu\mathbf{I})^{-1}\mathbf{V}^{-1}$

is diagonalizable, with the **same** eigenvectors $\mathbf{v}_i$ and eigenvalues $\beta_i = (\lambda_i - \mu)^{-1}$.

Their moduli are $|\beta_i| = 1/|\lambda_i - \mu|$, so the hypothesis on j says exactly that $|\beta_j| > |\beta_i|$ for every $i \neq j$: the eigenvalue β_j is strictly dominant. Relabeling so that j comes first, the second largest modulus is $\max_{i\neq j}|\beta_i| = 1/\min_{i\neq j}|\lambda_i - \mu|$, so the ratio of Theorem 24.4.1 is

$$\frac{\max_{i\neq j}|\beta_i|}{|\beta_j|} = \frac{|\lambda_j - \mu|}{\min_{i\neq j}|\lambda_i - \mu|} = \rho_\mu.$$

Since the $\mathbf{v}_j$-coefficient of $\mathbf{x}^{(0)}$ is nonzero, Theorem 24.4.1 applies to $\mathbf{B}$ and $\mathbf{x}^{(0)}$ and gives unimodular ω_k with $\|\omega_k \mathbf{x}^{(k)} - \mathbf{q}_j\| \le 4C\rho_\mu^k$, where $\mathbf{q}_j = \mathbf{v}_j/\|\mathbf{v}_j\|$ and C is that theorem's constant for $\mathbf{B}$ and $\mathbf{x}^{(0)}$. That is the convergence of the line.

For the estimates, repeat the last paragraph of the proof of Theorem 24.4.1 with $\mathbf{A}$ in place of $\mathbf{B}$: with $\mathbf{z} = \omega_k \mathbf{x}^{(k)}$ we have $\mu_k = \mathbf{z}^*\mathbf{A}\mathbf{z}$ and $\mathbf{q}_j^*\mathbf{A}\mathbf{q}_j = \lambda_j$, so $|\mu_k - \lambda_j| \le 2\|\mathbf{A}\|\|\mathbf{z} - \mathbf{q}_j\| \le 8\|\mathbf{A}\|C\rho_\mu^k$. This proves the theorem.

■

Example. Aiming at the smaller eigenvalue

Let $\mathbf{A} = \left(\begin{smallmatrix} 2 & 1 \\ 1 & 2 \end{smallmatrix}\right)$ again, and take the shift $\mu = \frac{3}{4}$, which is nearer to the eigenvalue 1 than to 3. Run inverse iteration from $\mathbf{x}^{(0)} = \mathbf{e}_1$, and compare with Example (Every iterate exactly).

Solution. $\mathbf{A} - \frac{3}{4}\mathbf{I} = \left(\begin{smallmatrix} 5/4 & 1 \\ 1 & 5/4 \end{smallmatrix}\right)$ has determinant $\frac{25}{16} - 1 = \frac{9}{16}$, so

$$(\mathbf{A} - \tfrac{3}{4}\mathbf{I})^{-1} = \tfrac{16}{9}\begin{pmatrix} 5/4 & -1 \\ -1 & 5/4 \end{pmatrix} = \tfrac{1}{9}\begin{pmatrix} 20 & -16 \\ -16 & 20 \end{pmatrix}.$$

Ignoring positive scalars, each step is $(a, b) \mapsto (5a - 4b,\ -4a + 5b)$, so the iterates are

$$(1,0) \to (5,-4) \to (41,-40) \to (365,-364) \to (3281,-3280) \to \dots,$$

running to the direction $(1,-1)$ — the eigenvector of the eigenvalue 1. In closed form: the eigenvalues of $(\mathbf{A} - \frac{3}{4}\mathbf{I})^{-1}$ are $(3 - \frac{3}{4})^{-1} = \frac{4}{9}$ and $(1 - \frac{3}{4})^{-1} = 4$, so expanding $\mathbf{e}_1 = \frac{1}{2}(1,1) + \frac{1}{2}(1,-1)$ gives

$$(\mathbf{A} - \tfrac{3}{4}\mathbf{I})^{-k}\mathbf{e}_1 = \tfrac{1}{2}(\tfrac{4}{9})^k(1,1) + \tfrac{1}{2}4^k(1,-1) \ \propto\ (9^k + 1,\ -(9^k - 1)),$$

the last step on multiplying by $2(9/4)^k$. Halving, this is the list above. Writing $\mathbf{x} = (a, b)$ with $a + b = 1$, we have $2ab = 1 - (a^2 + b^2)$ and so

$$\mu_k = \frac{2a^2 + 2ab + 2b^2}{a^2 + b^2} = 1 + \frac{1}{a^2 + b^2} = 1 + \frac{2}{81^k + 1},$$

since $a^2 + b^2 = \frac{1}{2}(81^k + 1)$. The errors are $1, \frac{1}{41}, \frac{1}{3281}, \frac{1}{265721}, \frac{1}{21523361}, \dots$, each about $\frac{1}{81}$ of the last, while the sine of the angle to $(1,-1)$ is $(81^k + 1)^{-1/2}$, shrinking by $\frac{1}{9}$ per step. The

theorem predicts exactly that:

$$\rho_\mu = \frac{|1 - \frac{3}{4}|}{|3 - \frac{3}{4}|} = \frac{1/4}{9/4} = \frac{1}{9},$$

with the estimates again going at the squared rate $\rho_\mu^{2k} = 81^{-k}$, as in Example (Every iterate exactly) and for the same reason. Compare that example, where the same matrix and the same starting vector gave rate $\frac{1}{3}$ towards the *other* eigenvector: one well-chosen shift has both redirected the method and made it gain twice as many digits per step.

The objection, and the answer. If μ is close to λ_j, then $\mathbf{A} - \mu\mathbf{I}$ is close to singular and $\kappa_2(\mathbf{A} - \mu\mathbf{I})$ is enormous. For normal $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{U}^*$ (Corollary 12.4.3) we have $\mathbf{A} - \mu\mathbf{I} = \mathbf{U}(\mathbf{D} - \mu\mathbf{I})\mathbf{U}^*$, whose singular values are the numbers $|\lambda_i - \mu|$, so by Proposition 16.8.2 (d)

$$\kappa_2(\mathbf{A} - \mu\mathbf{I}) = \frac{\max_i |\lambda_i - \mu|}{|\lambda_j - \mu|},$$

which grows without bound as $\mu \to \lambda_j$. Section 3 spent its length on how much a large condition number costs. And the better the shift, the worse the conditioning: **the method is deliberately solving an ill-conditioned system, and it does so on purpose.**

It does no harm, and here is why. Suppose the linear solve is backward stable in the sense of Definition 24.1.3, so that the computed $\widehat{\mathbf{z}}$ is the exact solution of

$$(\mathbf{A} - \mu\mathbf{I} + \boldsymbol{\Delta})\widehat{\mathbf{z}} = \mathbf{x}, \qquad \|\boldsymbol{\Delta}\| \le c\,u\,\|\mathbf{A} - \mu\mathbf{I}\|,$$

for a unit $\mathbf{x}$. The computed $\widehat{\mathbf{z}}$ may be wrong in every digit — the forward error is bounded only by $\kappa_2(\mathbf{A} - \mu\mathbf{I})cu$, which may exceed 1. But rearrange the same equation as a residual: $(\mathbf{A} - \mu\mathbf{I})\widehat{\mathbf{z}} - \mathbf{x} = -\boldsymbol{\Delta}\widehat{\mathbf{z}}$, so the **normalized** vector $\widehat{\mathbf{q}} = \widehat{\mathbf{z}}/\|\widehat{\mathbf{z}}\|$ satisfies

$$\|\mathbf{A}\widehat{\mathbf{q}} - \mu\widehat{\mathbf{q}}\| \;\le\; \frac{1}{\|\widehat{\mathbf{z}}\|} + c\,u\,\|\mathbf{A} - \mu\mathbf{I}\|.$$

The first term is small when $\|\widehat{\mathbf{z}}\|$ is large, and that takes an argument and a hypothesis. Near-singularity alone will not supply it: for $\mathbf{A} = \operatorname{diag}(1, 2)$, $\mu = 1 + 10^{-10}$, and $\mathbf{x} = \mathbf{e}_2$, the matrix $\mathbf{A} - \mu\mathbf{I}$ is as near singular as one pleases and yet $\mathbf{z} = (\mathbf{A} - \mu\mathbf{I})^{-1}\mathbf{x}$ has norm 1 to ten digits. What is needed is that $\mathbf{x}$ *see* the near-singular direction.

Take $\mathbf{A}$ normal with orthonormal eigenbasis $\mathbf{q}_1, \dots, \mathbf{q}_n$ (Corollary 12.4.3), so that $\mathbf{A}^*\mathbf{q}_j = \overline{\lambda_j}\mathbf{q}_j$ and hence $\mathbf{q}_j^*(\mathbf{A} - \mu\mathbf{I}) = (\lambda_j - \mu)\mathbf{q}_j^*$. Multiplying the perturbed equation by $\mathbf{q}_j^*$ gives $\mathbf{q}_j^*\mathbf{x} = (\lambda_j - \mu)\mathbf{q}_j^*\widehat{\mathbf{z}} + \mathbf{q}_j^*\boldsymbol{\Delta}\widehat{\mathbf{z}}$, so by Cauchy–Schwarz (Theorem 11.1.5) and Theorem 16.3.2 (a),

$$\|\widehat{\mathbf{z}}\| \;\ge\; \frac{|\mathbf{q}_j^*\mathbf{x}|}{|\lambda_j - \mu| + \|\boldsymbol{\Delta}\|}.$$

So a shift close to λ_j does force $\|\widehat{\mathbf{z}}\|$ to be large, **provided the component $\mathbf{q}_j^*\mathbf{x}$ is not small** — which is the hypothesis Theorem 24.4.3 already makes on the starting vector. Granting it, the first term of the residual bound is small, and the second is at the rounding level. **The residual of the normalized answer is small even though the answer itself is inaccurate**, and a small residual is all that an eigenvector approximation has to have.

For Hermitian $\mathbf{A}$ one can say precisely how much it buys. Let $\widehat{\mathbf{q}}$ be a unit vector, $\widehat{\mu} = R_{\mathbf{A}}(\widehat{\mathbf{q}})$ its Rayleigh quotient, and $\varepsilon = \|\mathbf{A}\widehat{\mathbf{q}} - \widehat{\mu}\widehat{\mathbf{q}}\|$ its residual. Expand $\widehat{\mathbf{q}} = \sum_i d_i\mathbf{q}_i$ in an orthonormal

eigenbasis (Corollary 12.4.3); by Theorem 11.2.1 (c),

$$\varepsilon^2 = \sum_i |d_i|^2(\lambda_i - \widehat{\mu})^2.$$

If λ_j is the only eigenvalue within distance g of $\widehat{\mu}$, every term with $i \neq j$ has $(\lambda_i - \widehat{\mu})^2 \geq g^2$, so $\sum_{i\neq j}|d_i|^2 \leq \varepsilon^2/g^2$ and the sine of the angle between $\widehat{\mathbf{q}}$ and $\mathbf{q}_j$ is at most ε/g. Together with Theorem 20.6.1, which puts an eigenvalue of $\mathbf{A}$ within ε of $\widehat{\mu}$, and Theorem 20.6.2, which improves that to ε^2/g, the small residual certifies both halves of the answer.

Warning. **"Ill-conditioned" is a statement about the wrong question.** The system $(\mathbf{A} - \mu\mathbf{I})\mathbf{z} = \mathbf{x}$ is indeed almost unsolvable *if what is wanted is* $\mathbf{z}$. What is wanted is the direction of $\mathbf{z}$, and the direction is precisely the part the ill-conditioning leaves alone: the huge factor $1/(\lambda_j - \mu)$ multiplies the component along $\mathbf{v}_j$, which is the component being sought, and dividing by $\|\widehat{\mathbf{z}}\|$ removes it. The residual estimate above is the rigorous form of that sentence. The same reasoning does **not** license inverse iteration when the eigenvalue is not simple or two eigenvalues are nearly equal: then g is small, the bound ε/g says nothing, and the computed direction really is unreliable.

24.4.3 The shift that improves itself

Inverse iteration needs a shift, and a better shift converges faster. But the iteration is manufacturing better and better approximations to λ_j as it goes, in the estimates μ_k. Use them.

Definition 24.4.3. Rayleigh Quotient Iteration

Let $\mathbf{A} \in M_n(F)$ be Hermitian and let $\mathbf{x}^{(0)}$ be a unit vector. **Rayleigh quotient iteration** is the sequence defined for $k \geq 0$ by

$$\mu_k = R_{\mathbf{A}}(\mathbf{x}^{(k)}) = (\mathbf{x}^{(k)})^*\mathbf{A}\mathbf{x}^{(k)}, \qquad (\mathbf{A} - \mu_k\mathbf{I})\mathbf{z}^{(k)} = \mathbf{x}^{(k)}, \qquad \mathbf{x}^{(k+1)} = \frac{\mathbf{z}^{(k)}}{\|\mathbf{z}^{(k)}\|},$$

the step being **undefined, and the iteration stopped, when μ_k is an eigenvalue of $\mathbf{A}$** — in which case μ_k is the answer, exactly.

It is inverse iteration with the shift refreshed at every step, and it is what Chapter 17 §01 and Chapter 20 §06 both promised. The price is a new factorization of $\mathbf{A} - \mu_k\mathbf{I}$ at every step, $\frac{2}{3}n^3$ flops instead of $2n^2$. What is bought is the following rate.

Theorem 24.4.4. Rayleigh Quotient Iteration Converges Cubically

Let $\mathbf{A} \in M_n(F)$ be Hermitian with an orthonormal basis $\mathbf{q}_1, \dots, \mathbf{q}_n$ of eigenvectors, $\mathbf{A}\mathbf{q}_i = \lambda_i\mathbf{q}_i$, and fix an index j whose eigenvalue is **simple**. Put

$$g = \min_{i\neq j}|\lambda_i - \lambda_j| > 0, \qquad D = \max_{i\neq j}|\lambda_i - \lambda_j|,$$

and for a unit vector $\mathbf{x}$ write $s(\mathbf{x}) = \|\mathbf{x} - \mathbf{q}_j\mathbf{q}_j^*\mathbf{x}\|$, the length of the part of $\mathbf{x}$ orthogonal to $\mathbf{q}_j$.

Let $\mathbf{x} = \mathbf{x}^{(k)}$ be a unit iterate with $s = s(\mathbf{x})$ satisfying $s \leq \frac{1}{2}$ and $Ds^2 \leq \frac{g}{2}$. Then $|\mu_k - \lambda_j| \leq Ds^2$, and either $\mu_k = \lambda_j$ — the iteration has found the eigenvalue exactly — or the next

iterate is defined and

$$s\big(\mathbf{x}^{(k+1)}\big) \;\le\; \frac{4D}{g}\, s^3.$$

Consequently, if $s_0 = s(\mathbf{x}^{(0)})$ satisfies $s_0 \le \frac{1}{2}$ and $s_0^2 \le g/(8D)$, then $a_k \coloneqq \sqrt{4D/g}\; s(\mathbf{x}^{(k)})$ satisfies $a_k \le a_0^{3^k}$ with $a_0 \le 1/\sqrt{2}$, so $s(\mathbf{x}^{(k)}) \to 0$ and $\mu_k \to \lambda_j$.

Idea. Two facts, multiplied together. The Rayleigh quotient is accurate to *second* order in the eigenvector error, so a shift built from an iterate at distance s is wrong by only Ds^2 — that is Corollary 24.4.2's squaring again, in its noniterative form. And one step of inverse iteration with a shift wrong by δ multiplies the error by about δ/g. Put $\delta = Ds^2$ and the step multiplies s by Ds^2/g: a factor s^2 per step, which is one cube. Everything else is the arithmetic of keeping the shift far enough from the other eigenvalues for g to survive, and that is what the two smallness hypotheses are for.

Proof. Expand $\mathbf{x} = \sum_i c_i \mathbf{q}_i$ with $\sum_i |c_i|^2 = 1$ (Theorem 11.2.1 (c)). Then $\mathbf{q}_j \mathbf{q}_j^* \mathbf{x} = c_j \mathbf{q}_j$, so

$$s^2 = \sum_{i \ne j} |c_i|^2, \qquad |c_j|^2 = 1 - s^2.$$

Step 1: the shift is accurate to second order. By Theorem 11.2.1 (b), $\mu_k = \mathbf{x}^* \mathbf{A} \mathbf{x} = \sum_i |c_i|^2 \lambda_i$, and since the weights sum to 1,

$$\mu_k - \lambda_j = \sum_{i \ne j} |c_i|^2 (\lambda_i - \lambda_j), \qquad |\mu_k - \lambda_j| \le D \sum_{i \ne j} |c_i|^2 = Ds^2.$$

Step 2: the shift stays away from the other eigenvalues. By hypothesis $Ds^2 \le g/2$, so $|\mu_k - \lambda_j| \le g/2$ and, for $i \ne j$,

$$|\lambda_i - \mu_k| \ge |\lambda_i - \lambda_j| - |\lambda_j - \mu_k| \ge g - \tfrac{g}{2} = \tfrac{g}{2} > 0.$$

In particular the only eigenvalue μ_k can equal is λ_j. If it does, the iteration stops with the exact answer. Otherwise $\mathbf{A} - \mu_k \mathbf{I}$ is invertible and the step is defined.

Step 3: one step. Since $\mathbf{A}\mathbf{q}_i = \lambda_i \mathbf{q}_i$,

$$\mathbf{z}^{(k)} = (\mathbf{A} - \mu_k \mathbf{I})^{-1} \mathbf{x} = \sum_i \frac{c_i}{\lambda_i - \mu_k}\, \mathbf{q}_i.$$

The quantity $s(\cdot)$ is unchanged by scaling, so $s(\mathbf{x}^{(k+1)}) = s(\mathbf{z}^{(k)}/\|\mathbf{z}^{(k)}\|)$ is the ratio of the length of the part of $\mathbf{z}^{(k)}$ orthogonal to $\mathbf{q}_j$ to the length of $\mathbf{z}^{(k)}$:

$$s\big(\mathbf{x}^{(k+1)}\big)^2 = \frac{\displaystyle\sum_{i \ne j} \frac{|c_i|^2}{|\lambda_i - \mu_k|^2}}{\displaystyle\sum_i \frac{|c_i|^2}{|\lambda_i - \mu_k|^2}}.$$

Bound the numerator above using Step 2, and the denominator below by its $i = j$ term alone:

$$s\big(\mathbf{x}^{(k+1)}\big)^2 \;\le\; \frac{(2/g)^2\, s^2}{|c_j|^2/|\lambda_j - \mu_k|^2} \;=\; \frac{4s^2|\lambda_j - \mu_k|^2}{g^2\,(1-s^2)} \;\le\; \frac{4s^2 D^2 s^4}{g^2(1-s^2)},$$

using Step 1 at the end. Taking square roots and using $s \le \tfrac12$, so that $\sqrt{1-s^2} \ge \tfrac{\sqrt3}{2} \ge \tfrac12$,

$$s\big(\mathbf{x}^{(k+1)}\big) \;\le\; \frac{2D\,s^3}{g\sqrt{1-s^2}} \;\le\; \frac{4D}{g}\,s^3.$$

Step 4: the iteration. Put $C = 4D/g$ and $a_k = \sqrt{C}\,s(\mathbf{x}^{(k)})$. The hypotheses $s_0 \le \tfrac12$ and $s_0^2 \le g/(8D)$ give $a_0^2 = Cs_0^2 \le \tfrac12$, so $a_0 \le 1/\sqrt2 < 1$, and also $Ds_0^2 \le g/2$. Whenever the hypotheses hold at step k, Step 3 gives $a_{k+1} \le \sqrt{C}\,Cs_k^3 = a_k^3$, so $a_{k+1} \le a_k^3 \le a_k \le a_0$; hence $s_{k+1} \le s_k \le s_0 \le \tfrac12$ and $Ds_{k+1}^2 \le Ds_0^2 \le g/2$, and the hypotheses hold again. By induction $a_k \le a_0^{3^k}$ for every k. Since $a_0 < 1$, $a_k \to 0$ and so $s(\mathbf{x}^{(k)}) \to 0$; by Step 1, $|\mu_k - \lambda_j| \le Ds_k^2 \to 0$. This proves the theorem.

■

Cubic convergence is a different animal from the linear convergence of Theorem 24.4.1. Each step roughly **triples the number of correct digits**, so an iterate with two correct digits has six after one step and eighteen after two — which is more than the arithmetic holds. In practice the iteration is over in three or four steps from any reasonable start, and the cost is dominated by those three or four factorizations.

Example. Three steps, exactly

Run Rayleigh quotient iteration on $\mathbf{A} = \left(\begin{smallmatrix} 2 & 1 \\ 1 & 2 \end{smallmatrix}\right)$ from $\mathbf{x}^{(0)} \propto (2,1)$, in exact arithmetic.

Solution. Write each iterate as an integer vector, scaling being irrelevant. Take $\mathbf{q}_j = \frac{1}{\sqrt2}(1,1)$, so $\lambda_j = 3$, and here $g = D = 2$, giving $4D/g = 4$.

Step 0. $\mathbf{x} \propto (2,1)$. Then $\mu_0 = (2\cdot 2^2 + 2\cdot 2\cdot 1 + 2\cdot 1^2)/5 = \frac{14}{5}$, and $\mathbf{A} - \frac{14}{5}\mathbf{I} = \left(\begin{smallmatrix} -4/5 & 1 \\ 1 & -4/5 \end{smallmatrix}\right)$ has determinant $-\frac{9}{25}$. Solving gives $\mathbf{z} \propto (13,14)$.

Step 1. $\mathbf{x} \propto (13,14)$, $\mu_1 = \frac{1094}{365}$, and $3 - \mu_1 = \frac{1}{365}$. Solving again gives $\mathbf{z} \propto (9842, 9841)$.

Step 2. $\mu_2 = \frac{581130734}{193710245}$, with $3 - \mu_2 = \frac{1}{193710245}$.

Step 3. $\mathbf{x} \propto (3812798742493,\ 3812798742494)$ and $3 - \mu_3 = \frac{1}{29074868501520029845195085}$.

The eigenvalue errors are

$$2\cdot 10^{-1}, \quad 2.7\cdot 10^{-3}, \quad 5.2\cdot 10^{-9}, \quad 3.4\cdot 10^{-26},$$

and the eigenvector errors s_k are

$$3.16\cdot 10^{-1}, \quad 3.70\cdot 10^{-2}, \quad 5.08\cdot 10^{-5}, \quad 1.31\cdot 10^{-13}.$$

Check the theorem: $s_1/s_0^3 = 1.17$, $s_2/s_1^3 = 1.00$, $s_3/s_2^3 = 1.00$, all comfortably below the

predicted constant $4D/g = 4$. Four steps have produced twenty-five correct digits from a starting vector accurate to none.

Remark. **The non-Hermitian rate is quadratic, and we do not prove it.** For a general diagonalizable $\mathbf{A}$, with μ_k taken to be $(\mathbf{x}^{(k)})^* \mathbf{A}\mathbf{x}^{(k)}$, the same iteration converges locally at a *quadratic* rate rather than a cubic one. The reason is visible in Step 1 above: it used that μ_k is a weighted average of the eigenvalues with weights $|c_i|^2$, so that the first-order term in s is absent — and that identity is exactly the Hermitian spectral theorem at work. Without it the shift is accurate only to first order, $|\mu_k - \lambda_j| = O(s)$, and the step multiplies s by $O(s)$. **The general statement is classical and is not proved here**, and nothing in this book depends on it.

24.4.4 Deflation, and what comes next

One eigenpair is rarely enough. For Hermitian $\mathbf{A}$, once a unit eigenvector $\mathbf{q}_1$ and its eigenvalue λ_1 are known, the next one can be reached by removing the first: put

$$\mathbf{A}' = \mathbf{A} - \lambda_1 \mathbf{q}_1 \mathbf{q}_1^*.$$

Let $\mathbf{q}_1, \dots, \mathbf{q}_n$ be an orthonormal eigenbasis (Corollary 12.4.3). Then $\mathbf{A}'\mathbf{q}_1 = \lambda_1 \mathbf{q}_1 - \lambda_1 \mathbf{q}_1 (\mathbf{q}_1^* \mathbf{q}_1) = \mathbf{0}$, and for $i \neq 1$, $\mathbf{A}'\mathbf{q}_i = \lambda_i \mathbf{q}_i - \lambda_1 \mathbf{q}_1 (\mathbf{q}_1^* \mathbf{q}_i) = \lambda_i \mathbf{q}_i$ by orthonormality. So $\mathbf{A}'$ is Hermitian with the same eigenvectors and the eigenvalue list $0, \lambda_2, \dots, \lambda_n$: the power method applied to $\mathbf{A}'$ finds λ_2, provided λ_2 is now dominant in modulus.

Warning. **Deflation is exact only with an exact eigenvector.** If the computed $\widehat{\mathbf{q}}_1$ differs from $\mathbf{q}_1$ by an angle η, then the matrix actually formed differs from the intended $\mathbf{A}'$ by $\lambda_1(\mathbf{q}_1\mathbf{q}_1^* - \widehat{\mathbf{q}}_1\widehat{\mathbf{q}}_1^*)$, of norm of order $|\lambda_1|\eta$, and by the perturbation bounds of Chapters 17 and 20 that error is inherited by every eigenvalue and eigenvector computed afterwards. Deflating $n-1$ times accumulates $n-1$ such errors. For one or two extreme eigenvalues deflation is the right tool; for the whole spectrum it is not, and the next section builds the algorithm that is.

Everything in this section has been a single vector chasing a single eigenvector. The next section iterates on a whole basis at once, and the result — the QR algorithm — turns out to be the power method in disguise, its accumulated factors furnishing a $\mathbf{QR}$ factorization of $\mathbf{A}^k$, with the shifts of this section reappearing as the device that makes it fast.

24.4.5 Exercises

24.4.5.1 A. Check your understanding

Exercise. A1

(a) State the two hypotheses of Theorem 24.4.1 on the eigenvalues and on the starting vector, and say what the rate ρ is.

(b) For a Hermitian $\mathbf{A}$, at what rate does the eigenvector converge and at what rate does

the Rayleigh quotient? Which earlier result explains the relation between the two?

(c) Explain in one sentence why inverse iteration with shift μ finds the eigenvalue nearest μ.

(d) True or false: "the shifted matrix in inverse iteration is nearly singular, so the method is unreliable". Justify your answer.

(e) Give the per-step cost of the power method, of inverse iteration, and of Rayleigh quotient iteration, for a dense $n \times n$ matrix.

Solution.

(a) The dominant eigenvalue must be strictly dominant **in modulus**, $|\lambda_1| > |\lambda_2|$, and the starting vector must have a nonzero coefficient c_1 on the corresponding eigenvector. The rate is $\rho = |\lambda_2/\lambda_1|$, the ratio of the two largest moduli.

(b) The eigenvector converges at rate ρ^k (Theorem 24.4.1 (a)) and the Rayleigh quotient at rate ρ^{2k} (Corollary 24.4.2). The relation is the second-order accuracy of the Rayleigh quotient at a Hermitian matrix, quantified by Theorem 20.6.2: a residual of size ε gives an eigenvalue error of at most ε^2 divided by the gap.

(c) Because $(\mathbf{A} - \mu\mathbf{I})^{-1}$ has the same eigenvectors as $\mathbf{A}$ and the eigenvalues $(\lambda_i - \mu)^{-1}$, whose moduli are largest precisely for the λ_i nearest μ (Theorem 24.4.3).

(d) False. The forward error in $\mathbf{z}^{(k)}$ is indeed large, but only its **direction** is used, and normalizing removes the part of the error that the ill-conditioning creates; the residual $\|\mathbf{A}\widehat{\mathbf{q}} - \mu\widehat{\mathbf{q}}\|$ of the normalized answer is small, and that is what certifies an eigenvector. The method is unreliable for a different reason — a small gap g — as the warning after Theorem 24.4.3 says.

(e) Power method: one matrix–vector product, $2n^2$ flops. Inverse iteration: one factorization of $\mathbf{A} - \mu\mathbf{I}$ by elimination with pivoting, $\frac{2}{3}n^3$, computed once, then $2n^2$ per step. Rayleigh quotient iteration: a new factorization every step, $\frac{2}{3}n^3$ per step — paid for by the cubic rate, which needs only three or four steps.

24.4.5.2 B. Practice

Exercise. B1: A rate, read off the spectrum

Let $\mathbf{A} = \left(\begin{smallmatrix} 4 & 1 \\ 2 & 3 \end{smallmatrix}\right)$.

(a) Find the eigenvalues and eigenvectors of $\mathbf{A}$, and the rate ρ of the power method.

(b) Starting from $\mathbf{x}^{(0)} = \mathbf{e}_1$, compute $\mathbf{A}^k\mathbf{e}_1$ in closed form and verify that the direction converges at rate ρ^k.

(c) How many steps are needed to gain six decimal digits?

Solution.

(a) $\operatorname{tr}\mathbf{A} = 7$ and $\det\mathbf{A} = 10$, so the eigenvalues are 5 and 2. For $\lambda = 5$: $(\mathbf{A} - 5\mathbf{I})\mathbf{v} = \mathbf{0}$ reads $-v_1 + v_2 = 0$, giving $\mathbf{v}_1 = (1, 1)$. For $\lambda = 2$: $2v_1 + v_2 = 0$, giving $\mathbf{v}_2 = (1, -2)$. Hence $\rho = \frac{2}{5}$.

(b) $\mathbf{e}_1 = \frac{2}{3}(1,1) + \frac{1}{3}(1,-2)$, as one checks by adding. So

$$\mathbf{A}^k \mathbf{e}_1 = \tfrac{2}{3}5^k(1,1) + \tfrac{1}{3}2^k(1,-2).$$

Dividing by $\frac{2}{3}5^k$ gives $(1,1) + \frac{1}{2}(\frac{2}{5})^k(1,-2)$, so the distance from the direction $(1,1)$ is a constant times $(\frac{2}{5})^k = \rho^k$, as Theorem 24.4.1 says.

(c) Each step multiplies the error by $\rho = 0.4$, that is by $10^{-\log_{10}(2.5)} = 10^{-0.398}$. Six digits need $6/0.398 = 15.1$ steps, so sixteen (and indeed $\rho^{15} = 1.1 \cdot 10^{-6}$, $\rho^{16} = 4.3 \cdot 10^{-7}$).

Exercise. B2: Choosing a shift

Let $\mathbf{A} = \operatorname{diag}(10, 3, 1)$.

(a) What is the rate of the plain power method?

(b) Find the rate of inverse iteration with shift $\mu = 0$, and say which eigenvector it finds.

(c) Find the rate of inverse iteration with shift $\mu = \frac{5}{2}$, and say which eigenvector it finds.

(d) Which shift would you choose to compute the middle eigenvalue as fast as possible, and what rate does it give? Explain why no shift gives a rate of 0.

Solution.

(a) By Theorem 24.4.1, $\rho = |3/10| = 0.3$, and the iterates converge to $\mathbf{e}_1$.

(b) With $\mu = 0$ the eigenvalue nearest 0 is 1, at distance 1; the next nearest is 3, at distance 3. So $\rho_\mu = \frac{1}{3}$, and the method finds $\mathbf{e}_3$. (Inverse iteration with shift 0 is the power method on $\mathbf{A}^{-1}$.)

(c) Distances from $\frac{5}{2}$ are 7.5, 0.5, 1.5. The nearest eigenvalue is 3, the next nearest 1, so $\rho_\mu = 0.5/1.5 = \frac{1}{3}$, and the method finds $\mathbf{e}_2$, the **middle** eigenvector — which the plain power method can never reach.

(d) Take $\mu = 3 + \delta$ with $0 < \delta < 1$: the distances are $7 - \delta$, δ, $2 + \delta$, so $\rho_\mu = \delta/(2+\delta) \to 0$ as $\delta \to 0$. Any μ very close to 3 is excellent. No shift gives rate exactly 0, because $\mu = 3$ is an eigenvalue and Definition 24.4.2 then does not apply: $\mathbf{A} - 3\mathbf{I}$ is singular and there is nothing to solve with. The limiting case is not a fast method but no method — which is why Rayleigh quotient iteration, which walks the shift towards the eigenvalue, stops the moment it arrives.

Exercise. B3: Deflating a symmetric matrix

Let $\mathbf{A} = \left(\begin{smallmatrix} 2 & 1 \\ 1 & 2 \end{smallmatrix}\right)$.

(a) Write down $\mathbf{A}' = \mathbf{A} - 3\mathbf{q}_1\mathbf{q}_1^\top$ for the unit eigenvector $\mathbf{q}_1$ of the eigenvalue 3, and verify its eigenvalues directly.

(b) What does the power method of Theorem 24.4.1, applied to $\mathbf{A}'$ from $\mathbf{x}^{(0)} = \mathbf{e}_1$, produce, and at what rate?

(c) Explain why the answer to (b) makes deflation look better than it is.

Solution.

(a) $\mathbf{q}_1 = \frac{1}{\sqrt{2}}(1,1)$ and $\mathbf{q}_1\mathbf{q}_1^\top = \frac{1}{2}\left(\begin{smallmatrix}1&1\\1&1\end{smallmatrix}\right)$, so

$$\mathbf{A}' = \begin{pmatrix}2&1\\1&2\end{pmatrix} - \tfrac{3}{2}\begin{pmatrix}1&1\\1&1\end{pmatrix} = \tfrac{1}{2}\begin{pmatrix}1&-1\\-1&1\end{pmatrix}.$$

Then $\mathbf{A}'(1,1) = \mathbf{0}$ and $\mathbf{A}'(1,-1) = (1,-1)$: the eigenvalues are 0 and 1, as predicted, with the same eigenvectors.

(b) $\mathbf{A}'\mathbf{e}_1 = \frac{1}{2}(1,-1)$, which is already an eigenvector, so the power iteration of Definition 24.4.1 gives $\mathbf{x}^{(1)} = \frac{1}{\sqrt{2}}(1,-1)$ and every later iterate equals it: it converges in one step, with $\mu_1 = 1$ exactly.

(c) Because the deflation was exact. The eigenvector $\mathbf{q}_1 = \frac{1}{\sqrt{2}}(1,1)$ was known in closed form, so $\mathbf{A}'$ had the eigenvalue 0 exactly and the remaining eigenvector survived untouched. In a computation $\mathbf{q}_1$ is known only to within the accuracy of the first run, and by the warning after the deflation formula the error in $\mathbf{q}_1$ enters $\mathbf{A}'$ as a perturbation of size about $|\lambda_1|\eta$, which every later eigenpair inherits. One clean step here is not evidence about $n-1$ dirty ones.

24.4.5.3 C. Going deeper

Exercise. C1: What the squaring really needs

Let $\mathbf{A} \in M_n(F)$ be Hermitian with orthonormal eigenbasis $\mathbf{q}_1, \dots, \mathbf{q}_n$, and let $\mathbf{x}$ be a unit vector with $s = \|\mathbf{x} - \mathbf{q}_1\mathbf{q}_1^*\mathbf{x}\|$.

(a) Prove that $|R_{\mathbf{A}}(\mathbf{x}) - \lambda_1| \le s^2 \max_{i\neq 1}|\lambda_i - \lambda_1|$.

(b) Show by example that the corresponding statement for a non-Hermitian $\mathbf{A}$ is false: exhibit $\mathbf{A} \in M_2(\mathbb{C})$, an eigenvalue λ_1 with unit eigenvector $\mathbf{q}_1$, and unit vectors $\mathbf{x}$ with $s \to 0$ for which $|\mathbf{x}^*\mathbf{A}\mathbf{x} - \lambda_1|$ is **not** $O(s^2)$.

(c) Hence say which hypothesis of Theorem 24.4.4 the example violates, and which step of the proof breaks.

Solution.

(a) This is Step 1 of the proof of Theorem 24.4.4 with $j = 1$. Expanding $\mathbf{x} = \sum_i c_i\mathbf{q}_i$ with $\sum_i |c_i|^2 = 1$, we get $s^2 = \sum_{i\neq 1}|c_i|^2$ and, by Theorem 11.2.1 (b), $R_{\mathbf{A}}(\mathbf{x}) = \sum_i |c_i|^2\lambda_i$. Since the weights sum to 1,

$$R_{\mathbf{A}}(\mathbf{x}) - \lambda_1 = \sum_{i\neq 1}|c_i|^2(\lambda_i - \lambda_1),$$

whose modulus is at most $s^2 \max_{i\neq 1}|\lambda_i - \lambda_1|$.

(b) Take $\mathbf{A} = \left(\begin{smallmatrix}0&1\\0&0\end{smallmatrix}\right)$, the standard nondiagonalizable example, with $\lambda_1 = 0$ and $\mathbf{q}_1 = \mathbf{e}_1$. Let $\mathbf{x} = (\sqrt{1-s^2}, s)$ for small $s > 0$, a unit vector whose part orthogonal to $\mathbf{e}_1$ has length exactly s. Then

$$\mathbf{x}^*\mathbf{A}\mathbf{x} = \overline{x_1}x_2 = s\sqrt{1-s^2},$$

which is $s + O(s^3)$: **first** order in s, not second. So no bound of the form Cs^2 can hold as $s \to 0$.

(c) The example violates the hypothesis that $\mathbf{A}$ be Hermitian. The step that breaks is Step 1: for a Hermitian $\mathbf{A}$ the quantity $\mathbf{x}^*\mathbf{A}\mathbf{x}$ is the weighted average $\sum_i |c_i|^2\lambda_i$ of the eigenvalues, in which only the squared moduli of the coefficients appear and the first-order term is therefore absent. For a general $\mathbf{A}$ the cross terms $\overline{c_i}c_k\,\mathbf{q}_i^*\mathbf{A}\mathbf{q}_k$ with $i \neq k$ survive, and they are of first order in the small coefficients. With Step 1 weakened to $O(s)$, Step 3 multiplies s by $O(s)$ and the rate is quadratic, not cubic — which is the remark after Theorem 24.4.4.

Exercise. C2: Two eigenvalues of equal modulus, over the reals

Let $\mathbf{A} \in M_2(\mathbb{R})$ be the rotation matrix $\left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$.

(a) Find the eigenvalues of $\mathbf{A}$ and explain why Theorem 24.4.1 does not apply.

(b) Describe the power iteration from an arbitrary unit $\mathbf{x}^{(0)} \in \mathbb{R}^2$, and show that no line is a limit.

(c) Now take the shift $\mu = \frac{1}{2}$ and run inverse iteration over $\mathbb{C}$. Does the theorem apply? Compute ρ_μ. *Hint: the eigenvalues of* $\mathbf{A}$ *are* $\pm i$.

(d) Deduce a one-sentence moral about which of the two failures of Example (The method that never converges) a shift can repair.

Solution.

(a) $p_{\mathbf{A}}(x) = x^2 + 1$, so the eigenvalues are i and $-i$, both of modulus 1. The hypothesis $|\lambda_1| > |\lambda_2|$ fails.

(b) $\mathbf{A}$ is orthogonal, so $\|\mathbf{A}\mathbf{x}\| = \|\mathbf{x}\|$ and no normalization is needed: $\mathbf{x}^{(k)} = \mathbf{A}^k\mathbf{x}^{(0)}$. Since $\mathbf{A}^2 = -\mathbf{I}$, the iterates are $\mathbf{x}^{(0)}, \mathbf{A}\mathbf{x}^{(0)}, -\mathbf{x}^{(0)}, -\mathbf{A}\mathbf{x}^{(0)}, \mathbf{x}^{(0)}, \ldots$, of period 4, and the lines they span alternate between $\operatorname{span}(\mathbf{x}^{(0)})$ and $\operatorname{span}(\mathbf{A}\mathbf{x}^{(0)})$, which are distinct because $\mathbf{A}$ has no real eigenvector. A sequence taking two distinct values infinitely often has no limit.

(c) Over $\mathbb{C}$ the matrix is diagonalizable, with eigenvectors $(1, i)$ and $(1, -i)$. The shift $\mu = \frac{1}{2}$ is not an eigenvalue, and the distances are $|i - \frac{1}{2}| = |-i - \frac{1}{2}| = \sqrt{5}/2$ — **equal**. So there is no unique nearest eigenvalue and Theorem 24.4.3 still does not apply; $\rho_\mu = 1$. Any real shift fails in the same way, by symmetry in the conjugate pair; a shift such as $\mu = \frac{i}{2}$ does work, with distances $\frac{1}{2}$ and $\frac{3}{2}$ and $\rho_\mu = \frac{1}{3}$.

(d) A shift repairs a bad *rate* and a wrong *target*, by moving the origin from which moduli are measured; it cannot repair a tie, and over $\mathbb{R}$ a conjugate pair of eigenvalues is a tie that no real shift can break.

Exercise. C3: A small residual certifies an eigenpair

Let $\mathbf{A} \in M_n(\mathbb{C})$ be Hermitian with orthonormal eigenbasis $\mathbf{q}_1, \ldots, \mathbf{q}_n$, let $\widehat{\mathbf{q}}$ be a unit vector, let $\widehat{\mu} = R_{\mathbf{A}}(\widehat{\mathbf{q}})$ and $\varepsilon = \|\mathbf{A}\widehat{\mathbf{q}} - \widehat{\mu}\widehat{\mathbf{q}}\|$. Suppose λ_j is the only eigenvalue of $\mathbf{A}$ in the interval $(\widehat{\mu} - g, \widehat{\mu} + g)$.

(a) Prove that $\|\widehat{\mathbf{q}} - \mathbf{q}_j\mathbf{q}_j^*\widehat{\mathbf{q}}\| \leq \varepsilon/g$, assuming $\varepsilon < g$.

(b) Deduce a bound on $|\widehat{\mu} - \lambda_j|$ from Theorem 20.6.2, and compare it with the bound from Theorem 20.6.1.

(c) In the inverse iteration of Definition 24.4.2 with a backward stable solve, ε is about $1/\|\widehat{\mathbf{z}}\| + cu\|\mathbf{A} - \mu\mathbf{I}\|$. Explain what (a) then says about the computed direction, and in particular why the accuracy improves as the shift improves.

Solution.

(a) Expand $\widehat{\mathbf{q}} = \sum_i d_i \mathbf{q}_i$ with $\sum_i |d_i|^2 = 1$. Then $\mathbf{A}\widehat{\mathbf{q}} - \widehat{\mu}\widehat{\mathbf{q}} = \sum_i d_i(\lambda_i - \widehat{\mu})\mathbf{q}_i$, so by Theorem 11.2.1 (c),

$$\varepsilon^2 = \sum_i |d_i|^2(\lambda_i - \widehat{\mu})^2 \;\geq\; \sum_{i\neq j} |d_i|^2(\lambda_i - \widehat{\mu})^2 \;\geq\; g^2 \sum_{i\neq j} |d_i|^2,$$

because every λ_i with $i \neq j$ lies outside $(\widehat{\mu} - g, \widehat{\mu} + g)$. Since $\|\widehat{\mathbf{q}} - \mathbf{q}_j\mathbf{q}_j^*\widehat{\mathbf{q}}\|^2 = \sum_{i\neq j}|d_i|^2$, taking square roots gives the bound.

(b) Theorem 20.6.1 gives an eigenvalue within ε of $\widehat{\mu}$, and by hypothesis the only candidate in that range (once $\varepsilon < g$) is λ_j, so $|\widehat{\mu} - \lambda_j| \leq \varepsilon$. Theorem 20.6.2, applied with $\alpha = \widehat{\mu} - g$ and $\beta = \widehat{\mu} + g$, improves this to

$$|\widehat{\mu} - \lambda_j| \leq \frac{\varepsilon^2}{g},$$

which is smaller by a factor ε/g — the quadratic gain of the Rayleigh quotient once more.

(c) It says that the **direction** of the computed $\widehat{\mathbf{z}}$ is within ε/g of $\mathbf{q}_j$, whatever the relative error in $\widehat{\mathbf{z}}$ itself. As the shift μ approaches λ_j, the matrix $\mathbf{A} - \mu\mathbf{I}$ becomes more nearly singular, and — provided the right-hand side $\mathbf{x}^{(k)}$ has a component along $\mathbf{q}_j$ that is not small, by the bound $\|\widehat{\mathbf{z}}\| \geq |\mathbf{q}_j^*\mathbf{x}^{(k)}|/(|\lambda_j - \mu| + \|\mathbf{\Delta}\|)$ established in the discussion following Theorem 24.4.3 — $\|\widehat{\mathbf{z}}\|$ grows like $1/|\lambda_j - \mu|$. So the first part of ε shrinks in proportion, and the second part is already at the rounding level. **A worse-conditioned system gives a better answer**, because the quantity that grows is the one in the denominator. What the bound never forgives is a small g: two nearby eigenvalues leave ε/g large no matter how good the shift is.

24.5 The QR Algorithm

Chapter 11 §08 brought a square matrix to upper Hessenberg form with reflections and then stopped, with a warning attached: no finite procedure built from arithmetic and square roots can carry that reduction on to triangular form, because such a procedure would extract the roots of an arbitrary polynomial from its coefficients, and the theorem of Abel and Ruffini forbids it — a theorem this book quotes, there and here, and does not prove. Every eigenvalue method must therefore iterate, and nothing proved below rests on that quotation. The iteration this section develops is the one Chapter 11 named in a single sentence — factor $\mathbf{Z} = \mathbf{QR}$, then replace $\mathbf{Z}$ by $\mathbf{RQ}$ — and we now ask why it works, why it is always started from a Hessenberg matrix, and why it is always run with a shift.

24.5.1 Factor and multiply back

Section 4 iterated a matrix against a **vector**: the power method drives $\mathbf{x}^{(k)} = \mathbf{A}\mathbf{x}^{(k-1)}$ towards the dominant eigenvector, one eigenvalue at a time. The whole spectrum at once needs an

iteration on **matrices**, and it must be an iteration by similarity, or the spectrum would not survive it. Chapter 11's reflections already gave us a supply of similarities that cost little: $\mathbf{A} \mapsto \mathbf{Q}^*\mathbf{A}\mathbf{Q}$ with $\mathbf{Q}$ unitary. The question is which $\mathbf{Q}$ to take.

Here is the answer, and it looks like sleight of hand. Factor $\mathbf{A} = \mathbf{Q}\mathbf{R}$ and then multiply the two factors back **in the other order**. Since $\mathbf{Q}^*\mathbf{Q} = \mathbf{I}$, the product $\mathbf{R}\mathbf{Q}$ is $\mathbf{Q}^*(\mathbf{Q}\mathbf{R})\mathbf{Q} = \mathbf{Q}^*\mathbf{A}\mathbf{Q}$, so nothing has been lost; and the reversal shifts weight from the upper triangle into the lower one in a way that, repeated, empties the lower triangle altogether.

Factor the matrix, then put the factors back the wrong way round, and repeat.

Definition 24.5.1. The QR Iteration

Let $\mathbf{A} \in M_n(F)$, where $F = \mathbb{R}$ or $F = \mathbb{C}$. Set $\mathbf{A}^{(1)} = \mathbf{A}$, and for $k \geq 1$ define

$$\mathbf{A}^{(k)} = \mathbf{Q}^{(k)}\mathbf{R}^{(k)}, \qquad \mathbf{A}^{(k+1)} = \mathbf{R}^{(k)}\mathbf{Q}^{(k)},$$

where $\mathbf{A}^{(k)} = \mathbf{Q}^{(k)}\mathbf{R}^{(k)}$ is **a** factorization with $\mathbf{Q}^{(k)} \in M_n(F)$ unitary and $\mathbf{R}^{(k)} \in M_n(F)$ upper triangular. The sequence $\mathbf{A}^{(1)}, \mathbf{A}^{(2)}, \ldots$ is the **unshifted QR iteration** started at $\mathbf{A}$.

In words: at every step we factor the current matrix, then form the product of its factors in the reversed order, and that product is the next matrix. The superscript in parentheses counts steps, never coordinates.

Well-definedness. A factorization exists at every step, for every square matrix and with no hypothesis on its columns, by Theorem 11.8.4. It need not be unique: if $\mathbf{D}$ is diagonal with $|d_j| = 1$, then $(\mathbf{Q}\mathbf{D})(\mathbf{D}^*\mathbf{R})$ is another one. When $\mathbf{A}^{(k)}$ is **invertible** the ambiguity is exactly that, and demanding a positive real diagonal in $\mathbf{R}^{(k)}$ removes it, by Theorem 11.8.1; and if $\mathbf{A}$ is invertible then so is every $\mathbf{A}^{(k)}$, being similar to $\mathbf{A}$. We do not need the normalization for anything below, so the definition leaves the choice open.

Proposition 24.5.1. Every Step Is a Unitary Similarity

In the notation of Definition 24.5.1, $\mathbf{A}^{(k+1)} = (\mathbf{Q}^{(k)})^*\mathbf{A}^{(k)}\mathbf{Q}^{(k)}$ for every k. Consequently all the matrices $\mathbf{A}^{(k)}$ are unitarily similar to $\mathbf{A}$, and they share its characteristic polynomial, its eigenvalues with multiplicities, its trace, and its determinant.

Proof. Since $\mathbf{Q}^{(k)}$ is unitary, $(\mathbf{Q}^{(k)})^*\mathbf{Q}^{(k)} = \mathbf{I}_n$ (Definition 11.7.2), so

$$(\mathbf{Q}^{(k)})^*\mathbf{A}^{(k)}\mathbf{Q}^{(k)} = (\mathbf{Q}^{(k)})^*\mathbf{Q}^{(k)}\mathbf{R}^{(k)}\mathbf{Q}^{(k)} = \mathbf{R}^{(k)}\mathbf{Q}^{(k)} = \mathbf{A}^{(k+1)}.$$

A unitary matrix is invertible with $\mathbf{Q}^{-1} = \mathbf{Q}^*$, so this exhibits $\mathbf{A}^{(k+1)}$ as similar to $\mathbf{A}^{(k)}$ (Definition 4.2.2); composing these similarities for $1, \ldots, k-1$ makes $\mathbf{A}^{(k)}$ similar to $\mathbf{A}^{(1)} = \mathbf{A}$. Similar matrices have the same characteristic polynomial (Theorem 7.10.3), hence the same eigenvalues with multiplicities, and the same trace and determinant. This proves the proposition. ∎

Two examples fix what the iteration does and does not do.

Example. Two matrices the iteration cannot move

Determine the QR iterates of $\mathbf{A} = \operatorname{diag}(3, 1)$ and of the rotation $\mathbf{B} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$.

Solution. For $\mathbf{A}$: the factorization $\mathbf{A} = \mathbf{I}_2\mathbf{A}$ has $\mathbf{I}_2$ unitary and $\mathbf{A}$ upper triangular with positive diagonal, so by the uniqueness in Theorem 11.8.1 it is *the* factorization. Then $\mathbf{A}^{(2)} = \mathbf{A}\mathbf{I}_2 = \mathbf{A}$, and the iteration stands still. It has nothing to do: $\mathbf{A}$ is already triangular and its diagonal already lists the eigenvalues (Theorem 9.7.4).

For $\mathbf{B}$: here $\mathbf{B}$ is itself unitary, since its columns $(0, 1)$ and $(-1, 0)$ are orthonormal, so $\mathbf{B} = \mathbf{B}\,\mathbf{I}_2$ is a factorization with $\mathbf{I}_2$ upper triangular with positive diagonal — again the unique one. So $\mathbf{B}^{(2)} = \mathbf{I}_2\mathbf{B} = \mathbf{B}$, and the iteration stands still here too. This time it has everything to do: the eigenvalues of $\mathbf{B}$ are $\pm i$, and $\mathbf{B}$ is not triangular. The iteration is stuck.

The second matrix is the section's standing negative example, and the reason it is stuck is visible already: $|i| = |-i|$. We return to it after the convergence theorem, and repair it with a shift.

Warning. **Reversing the factors is not a similarity for just any factorization.** The step works because $\mathbf{Q}$ is unitary, so that $\mathbf{Q}^{-1} = \mathbf{Q}^*$ and $\mathbf{RQ} = \mathbf{Q}^*\mathbf{AQ}$. Do the same with an Ł$\mathbf{U}$ factorization and you get $\mathbf{U}$Ł $= \text{Ł}^{-1}\mathbf{A}$Ł, which is still a similarity — but with Ł^{-1} in place of Ł*, and Ł^{-1} can be enormous. That is the older Ł$\mathbf{R}$ iteration, and its instability under rounding is exactly the instability of elimination without pivoting met in Section 2.

24.5.2 The iteration is the power method in disguise

Nothing said so far suggests that the iterates go anywhere. The identity that explains them is the following, and it is the heart of the section: *the accumulated factors of k steps are a QR factorization of the k-th power of* $\mathbf{A}$. Powers of $\mathbf{A}$ are what the power method of Section 4 runs on, so the identity says the two methods are the same method, one applied to a vector and the other to a whole basis at once.

Write

$$\mathbf{U}_k = \mathbf{Q}^{(1)}\mathbf{Q}^{(2)}\cdots\mathbf{Q}^{(k)}, \qquad \mathbf{T}_k = \mathbf{R}^{(k)}\cdots\mathbf{R}^{(2)}\mathbf{R}^{(1)}.$$

The first is unitary, as a product of unitary matrices (Proposition 11.7.4), and the second is upper triangular, as a product of upper triangular matrices.

Theorem 24.5.2. QR Iterates Factor the Powers

Let $\mathbf{A} \in M_n(F)$ and let $\mathbf{A}^{(k)}$, $\mathbf{Q}^{(k)}$, $\mathbf{R}^{(k)}$ be as in Definition 24.5.1. Then for every $k \geq 1$,

$$\mathbf{A}^{(k+1)} = \mathbf{U}_k^*\,\mathbf{A}\,\mathbf{U}_k \qquad \text{and} \qquad \mathbf{A}^k = \mathbf{U}_k\mathbf{T}_k.$$

In particular $\mathbf{U}_k\mathbf{T}_k$ is a QR factorization of $\mathbf{A}^k$.

Idea. Induction on k, with the two statements proved together, because each feeds the other. The similarity statement is Proposition 24.5.1 telescoped. For the factorization, the move to find is the one that turns $\mathbf{A}\mathbf{U}_k$ into something with a $\mathbf{Q}$ at the front: the similarity statement, read backwards, says $\mathbf{A}\mathbf{U}_k = \mathbf{U}_k\mathbf{A}^{(k+1)}$, and $\mathbf{A}^{(k+1)}$ is by definition $\mathbf{Q}^{(k+1)}\mathbf{R}^{(k+1)}$. One line of bookkeeping then converts $\mathbf{A}\cdot\mathbf{A}^k$ into $\mathbf{U}_{k+1}\mathbf{T}_{k+1}$.

Proof. Both statements by induction on k.

For $k=1$: $\mathbf{U}_1 = \mathbf{Q}^{(1)}$ and $\mathbf{T}_1 = \mathbf{R}^{(1)}$, so $\mathbf{U}_1\mathbf{T}_1 = \mathbf{Q}^{(1)}\mathbf{R}^{(1)} = \mathbf{A}^{(1)} = \mathbf{A}$, and $\mathbf{U}_1^*\mathbf{A}\mathbf{U}_1 = \mathbf{A}^{(2)}$ by Proposition 24.5.1.

Let $k \geq 1$ and suppose both statements hold for k. By Proposition 24.5.1 applied at step $k+1$ and the inductive hypothesis,

$$\mathbf{A}^{(k+2)} = (\mathbf{Q}^{(k+1)})^*\mathbf{A}^{(k+1)}\mathbf{Q}^{(k+1)} = (\mathbf{Q}^{(k+1)})^*\mathbf{U}_k^*\mathbf{A}\,\mathbf{U}_k\mathbf{Q}^{(k+1)} = \mathbf{U}_{k+1}^*\mathbf{A}\,\mathbf{U}_{k+1},$$

where the last equality uses $\mathbf{U}_{k+1} = \mathbf{U}_k\mathbf{Q}^{(k+1)}$ and $(\mathbf{X}\mathbf{Y})^* = \mathbf{Y}^*\mathbf{X}^*$. That is the first statement for $k+1$.

For the second, multiply the inductive similarity $\mathbf{A}^{(k+1)} = \mathbf{U}_k^*\mathbf{A}\mathbf{U}_k$ on the left by $\mathbf{U}_k$, which is legitimate because $\mathbf{U}_k\mathbf{U}_k^* = \mathbf{I}_n$:

$$\mathbf{A}\,\mathbf{U}_k = \mathbf{U}_k\mathbf{A}^{(k+1)} = \mathbf{U}_k\mathbf{Q}^{(k+1)}\mathbf{R}^{(k+1)} = \mathbf{U}_{k+1}\mathbf{R}^{(k+1)}.$$

Hence, using the inductive factorization $\mathbf{A}^k = \mathbf{U}_k\mathbf{T}_k$,

$$\mathbf{A}^{k+1} = \mathbf{A}\,(\mathbf{U}_k\mathbf{T}_k) = (\mathbf{A}\,\mathbf{U}_k)\mathbf{T}_k = \mathbf{U}_{k+1}\mathbf{R}^{(k+1)}\mathbf{T}_k = \mathbf{U}_{k+1}\mathbf{T}_{k+1},$$

since $\mathbf{T}_{k+1} = \mathbf{R}^{(k+1)}\mathbf{T}_k$ by definition. As $\mathbf{U}_{k+1}$ is unitary and $\mathbf{T}_{k+1}$ upper triangular, this is a QR factorization of $\mathbf{A}^{k+1}$, and the induction is complete. ∎

The consequence to keep is about the **first column**. Comparing first columns in $\mathbf{A}^k = \mathbf{U}_k\mathbf{T}_k$ and writing $\tau_k = (\mathbf{T}_k)_{11}$, we get

$$\mathbf{A}^k\mathbf{e}_1 = \mathbf{U}_k\mathbf{T}_k\mathbf{e}_1 = \tau_k\,\mathbf{U}_k\mathbf{e}_1, \tag{24.5.1}$$

because the first column of an upper triangular matrix is $\tau_k\mathbf{e}_1$. So whenever $\tau_k \neq 0$, which the Quick check just below records for invertible $\mathbf{A}$, the first column of $\mathbf{U}_k$ is the normalized vector $\mathbf{A}^k\mathbf{e}_1/\|\mathbf{A}^k\mathbf{e}_1\|$, up to a scalar of modulus one — which is precisely the k-th iterate of the power method (Theorem 24.4.1) started at $\mathbf{e}_1$. The QR algorithm runs the power method in its first column, and, as the next theorem shows, that alone forces the first column of $\mathbf{A}^{(k+1)}$ to converge.

Quick check. Why is $\tau_k \neq 0$ when $\mathbf{A}$ is invertible?

Solution. Each $\mathbf{A}^{(j)}$ is similar to $\mathbf{A}$ (Proposition 24.5.1), hence invertible, and $\mathbf{R}^{(j)} = (\mathbf{Q}^{(j)})^*\mathbf{A}^{(j)}$ is then invertible as a product of invertible matrices. So $\mathbf{T}_k$ is invertible, being a product of invertible matrices; an upper triangular matrix is invertible exactly when its diagonal entries are all nonzero (Lemma 3.6.2), so $\tau_k = (\mathbf{T}_k)_{11} \neq 0$.

24.5.3 What converges, and how fast

Theorem 24.5.3. Convergence of the First Column

Let $\mathbf{A} \in M_n(\mathbb{C})$ be invertible and diagonalizable, say $\mathbf{A} = \mathbf{X}\mathbf{\Lambda}\mathbf{X}^{-1}$ with $\mathbf{\Lambda} = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$ and with $\mathbf{v}_1, \dots, \mathbf{v}_n$ the columns of $\mathbf{X}$. Assume a **single dominant eigenvalue**,

$$|\lambda_1| > |\lambda_2| \geq \cdots \geq |\lambda_n| > 0,$$

and write $\mathbf{e}_1 = c_1\mathbf{v}_1 + \cdots + c_n\mathbf{v}_n$, assuming $c_1 \neq 0$. Put $t = |\lambda_2/\lambda_1| < 1$. Then there is a constant C, depending on $\mathbf{A}$ and $\mathbf{X}$ but not on k, such that the unshifted QR iterates satisfy

$$\|\mathbf{A}^{(k+1)}\mathbf{e}_1 - \lambda_1\mathbf{e}_1\|_2 \leq C\,t^k \qquad (k \geq 1).$$

In particular $a_{11}^{(k)} \to \lambda_1$ and $a_{i1}^{(k)} \to 0$ for every $i \geq 2$, each at the rate t^k. For $n = 2$ this says that the iterates converge to upper triangular form, with $a_{11}^{(k)} \to \lambda_1$ and $a_{22}^{(k)} \to \lambda_2$.

Idea. By Eq. 24.5.1 the first column $\mathbf{u}_1$ of $\mathbf{U}_k$ is the power-method iterate, so it is close to the dominant eigenvector, and the vector $\mathbf{A}\mathbf{u}_1 - \lambda_1\mathbf{u}_1$ is small. Now read $\mathbf{A}^{(k+1)} = \mathbf{U}_k^*\mathbf{A}\mathbf{U}_k$ column by column: its first column is $\mathbf{U}_k^*\mathbf{A}\mathbf{u}_1$, and subtracting $\lambda_1\mathbf{e}_1 = \mathbf{U}_k^*(\lambda_1\mathbf{u}_1)$ leaves $\mathbf{U}_k^*(\mathbf{A}\mathbf{u}_1 - \lambda_1\mathbf{u}_1)$. A unitary matrix does not change lengths, so the whole error in the first column *equals* the length of that small vector. Nothing else is needed.

Proof. Write $\mathbf{u}_1 = \mathbf{U}_k\mathbf{e}_1$ for the first column of $\mathbf{U}_k$, suppressing k from the notation.

Step 1: the error in the first column is a residual. Since $\mathbf{U}_k$ is unitary and $\mathbf{A}^{(k+1)} = \mathbf{U}_k^*\mathbf{A}\mathbf{U}_k$ by Theorem 24.5.2,

$$\mathbf{A}^{(k+1)}\mathbf{e}_1 - \lambda_1\mathbf{e}_1 = \mathbf{U}_k^*\mathbf{A}\mathbf{u}_1 - \lambda_1\mathbf{U}_k^*\mathbf{u}_1 = \mathbf{U}_k^*(\mathbf{A}\mathbf{u}_1 - \lambda_1\mathbf{u}_1),$$

using $\mathbf{u}_1 = \mathbf{U}_k\mathbf{e}_1$ twice. A unitary matrix preserves the Euclidean norm (Theorem 11.7.3), so

$$\|\mathbf{A}^{(k+1)}\mathbf{e}_1 - \lambda_1\mathbf{e}_1\|_2 = \|\mathbf{A}\mathbf{u}_1 - \lambda_1\mathbf{u}_1\|_2. \tag{$*$}$$

Step 2: $\mathbf{u}_1$ is a power-method iterate. By Eq. 24.5.1, $\mathbf{A}^k\mathbf{e}_1 = \tau_k\mathbf{u}_1$ with $\tau_k \neq 0$, as the Quick check above records. Since $\mathbf{A}\mathbf{v}_i = \lambda_i\mathbf{v}_i$,

$$\mathbf{A}^k\mathbf{e}_1 = \sum_{i=1}^{n} c_i\lambda_i^k\mathbf{v}_i = c_1\lambda_1^k(\mathbf{v}_1 + \mathbf{z}_k), \qquad \mathbf{z}_k = \sum_{i\geq 2}\frac{c_i}{c_1}\Big(\frac{\lambda_i}{\lambda_1}\Big)^k\mathbf{v}_i,$$

which is legitimate because $c_1 \neq 0$ and $\lambda_1 \neq 0$. Every ratio obeys $|\lambda_i/\lambda_1| \leq t$, so

$$\|\mathbf{z}_k\|_2 \leq M t^k, \qquad M = \sum_{i \geq 2} \frac{|c_i|}{|c_1|} \|\mathbf{v}_i\|_2,$$

by the triangle inequality (Corollary 11.1.6). Put $\mathbf{y}_k = \mathbf{v}_1 + \mathbf{z}_k$. Then $\mathbf{u}_1 = \mathbf{A}^k \mathbf{e}_1 / \tau_k = \omega_k \, \mathbf{y}_k / \|\mathbf{y}_k\|_2$ for the scalar $\omega_k = c_1 \lambda_1^k \|\mathbf{y}_k\|_2 / \tau_k$, and $|\omega_k| = 1$ because $\|\mathbf{u}_1\|_2 = 1$.

Step 3: the residual is small. Since $\mathbf{A}\mathbf{v}_1 = \lambda_1 \mathbf{v}_1$, the vector $\mathbf{v}_1$ drops out of $(\mathbf{A} - \lambda_1 \mathbf{I})\mathbf{y}_k$, leaving

$$\mathbf{A}\mathbf{u}_1 - \lambda_1 \mathbf{u}_1 = \omega_k \, \frac{(\mathbf{A} - \lambda_1 \mathbf{I})\mathbf{z}_k}{\|\mathbf{y}_k\|_2},$$

$$\text{so} \qquad \|\mathbf{A}\mathbf{u}_1 - \lambda_1 \mathbf{u}_1\|_2 \leq \frac{(\|\mathbf{A}\|_2 + |\lambda_1|) M \, t^k}{\|\mathbf{y}_k\|_2}.$$

Choose k_0 with $M t^{k_0} \leq \frac{1}{2}\|\mathbf{v}_1\|_2$, possible because $t < 1$. For $k \geq k_0$ the triangle inequality gives $\|\mathbf{y}_k\|_2 \geq \|\mathbf{v}_1\|_2 - M t^k \geq \frac{1}{2}\|\mathbf{v}_1\|_2 > 0$, whence

$$\|\mathbf{A}\mathbf{u}_1 - \lambda_1 \mathbf{u}_1\|_2 \leq C_1 t^k, \qquad C_1 = \frac{2(\|\mathbf{A}\|_2 + |\lambda_1|) M}{\|\mathbf{v}_1\|_2}.$$

For the finitely many $k < k_0$ the crude bound $\|\mathbf{A}\mathbf{u}_1 - \lambda_1 \mathbf{u}_1\|_2 \leq \|\mathbf{A}\|_2 + |\lambda_1|$ holds, since $\|\mathbf{u}_1\|_2 = 1$. So the displayed estimate holds for all $k \geq 1$ with

$$C = \max\Big\{ C_1, \ (\|\mathbf{A}\|_2 + |\lambda_1|) t^{-k_0} \Big\},$$

and combining with $(*)$ proves the main estimate. The entrywise statements follow because each coordinate of a vector is at most its Euclidean norm.

For $n = 2$, the estimate forces $a_{21}^{(k)} \to 0$ and $a_{11}^{(k)} \to \lambda_1$; and $a_{22}^{(k)} = \operatorname{tr} \mathbf{A}^{(k)} - a_{11}^{(k)} = \lambda_1 + \lambda_2 - a_{11}^{(k)} \to \lambda_2$, the trace being preserved by Proposition 24.5.1. This proves the theorem.

■

What is proved here, and what is not. The theorem above settles the first column, for every n, and with it the whole 2×2 case. The general statement is this: if $\mathbf{A}$ is diagonalizable with $|\lambda_1| > |\lambda_2| > \cdots > |\lambda_n| > 0$, **and** if $\mathbf{X}^{-1}$ admits an Ł**U** factorization without row swaps (Theorem 3.6.4), then the whole of $\mathbf{A}^{(k)}$ below the diagonal tends to zero, with

$$a_{i+1,i}^{(k)} = O(|\lambda_{i+1}/\lambda_i|^k) \quad \text{and} \quad a_{ii}^{(k)} \to \lambda_i \quad (1 \leq i \leq n).$$

This is the classical convergence theorem of the algorithm, due to Rutishauser, Francis, and Kublanovskaya; **it is not proved here.** The proof runs through the factorization $\mathbf{A}^k = \mathbf{X}\mathbf{\Lambda}^k$Ł**U** of Theorem 24.5.2's identity, the observation that $\mathbf{\Lambda}^k$Ł$\mathbf{\Lambda}^{-k} \to \mathbf{I}$, and a continuity argument for the QR factorization that this book has not developed. Nothing in this section or later depends on it: the Hessenberg and shifting results below are proved outright, and the worked example at the end exhibits the rates rather than assuming them. The extra hypothesis on $\mathbf{X}^{-1}$ is not cosmetic, and the strict ordering of the moduli is not either — the next example shows what happens without it.

Example. The unshifted algorithm on a rotation

Let $\mathbf{B} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$. Explain why no unshifted QR iteration can bring $\mathbf{B}$ closer to triangular form, and say which hypothesis of Theorem 24.5.3 fails.

Solution. By Example (Two matrices the iteration cannot move), $\mathbf{B}^{(k)} = \mathbf{B}$ for every k: the iteration is stationary, and $\mathbf{B}$ is as far from triangular as a 2×2 matrix can be, with $b_{21} = 1$. The eigenvalues are $\pm i$, so $|\lambda_1| = |\lambda_2| = 1$ and the hypothesis $|\lambda_1| > |\lambda_2|$ fails. The failure is not an artifact of the proof. Over $\mathbb{R}$ the matrix is orthogonally similar to no real triangular matrix at all, since a real triangular matrix has real eigenvalues; and over $\mathbb{C}$ the power method has nothing to converge to, because $\mathbf{B}^k\mathbf{e}_1$ simply rotates around the plane forever (Theorem 24.4.1).

Warning. **Convergence is to triangular form, not to diagonal form.** Even when the iterates do converge, the strictly upper triangular part generally does not: only the entries **below** the diagonal die out. What is more, in the convergence theorem quoted above — the one this section does not prove — the order in which the eigenvalues appear on the diagonal is decided by the moduli, largest first, so that a run it covers reveals λ_1 in the top-left corner and λ_n in the bottom-right. With shifts, which is how the algorithm is actually run, even that ordering goes away.

24.5.4 Hessenberg first

As written, one step costs a QR factorization of a full $n \times n$ matrix and a matrix product: that is $O(n^3)$ flops, in the counting convention of Proposition 24.2.3, and the count really does grow like n^3, since the product alone touches n^2 entries with n terms each. The iteration needs many steps, and no amount of cleverness about the steps rescues a method that spends n^3 operations on each of them.

The remedy is Chapter 11's. Reduce $\mathbf{A}$ once to upper Hessenberg form by Theorem 11.8.5, at a cost of $O(n^3)$ paid a single time, and then iterate. This pays off only if the pattern survives the step, which is the content of the next proposition — and it is the sentence Chapter 11 §08 closed on: "Chapter 24 develops that iteration and explains why every implementation starts with the reduction proved here."

Proposition 24.5.4. A QR Step Preserves Hessenberg Form

Let $\mathbf{Z} \in M_n(\mathbb{R})$ be upper Hessenberg, $n \geq 2$. Then $\mathbf{Z}$ has a factorization $\mathbf{Z} = \mathbf{QR}$ with $\mathbf{Q}$ orthogonal and $\mathbf{R}$ upper triangular in which $\mathbf{Q}$ is a product of $n - 1$ Givens rotations and is **itself upper Hessenberg**; and for this factorization $\mathbf{RQ}$ is again upper Hessenberg. Producing $\mathbf{R}$ and then $\mathbf{RQ}$ costs about $6n^2$ flops.

Idea. A Hessenberg matrix has only $n-1$ entries in the way of triangularity, one per column, and Definition 11.8.2 removes one entry per rotation. So the natural factorization to use is the rotation one, not the reflection one — this is the case Chapter 11 §08 flagged, where

"the count reverses". Then two bookkeeping facts finish it: the accumulated product of the rotations is Hessenberg, and upper triangular times upper Hessenberg is upper Hessenberg.

Proof. For $1 \le i \le n-1$ let $\mathbf{G}_i = \mathbf{G}(i, i+1; c_i, s_i)$ be a Givens rotation in the coordinates $i, i+1$ (Definition 11.8.2). Working left to right, choose $\mathbf{G}_i$ to annihilate the entry in position $(i+1, i)$ of the matrix produced so far: if that matrix has entries a in position (i, i) and b in position $(i+1, i)$ with $(a, b) \neq (0, 0)$, take $c_i = a/r$ and $s_i = b/r$ with $r = \sqrt{a^2 + b^2}$, as in Chapter 11 §08; if $(a, b) = (0, 0)$, take $\mathbf{G}_i = \mathbf{I}_n$. This is the only entry below the diagonal in column i: the entries in rows $i+2, \dots, n$ of that column were zero in $\mathbf{Z}$, because $\mathbf{Z}$ is upper Hessenberg, and the earlier rotations $\mathbf{G}_1, \dots, \mathbf{G}_{i-1}$ touch only rows $1, \dots, i$. Left multiplication by $\mathbf{G}_i$ recombines rows i and $i+1$ only, so it cannot disturb the zeros already created: in a column $j \le i-1$ both of the rows being combined are already zero, since $i \ge j+1$. Hence

$$\mathbf{R} = \mathbf{G}_{n-1} \cdots \mathbf{G}_1 \mathbf{Z}$$

is upper triangular, and $\mathbf{Z} = \mathbf{Q}\mathbf{R}$ with $\mathbf{Q} = \mathbf{G}_1^{\top} \mathbf{G}_2^{\top} \cdots \mathbf{G}_{n-1}^{\top}$, orthogonal as a product of orthogonal matrices.

Claim. $\mathbf{P}_m = \mathbf{G}_1^{\top} \cdots \mathbf{G}_m^{\top}$ is upper Hessenberg for $1 \le m \le n-1$, and its columns $m+2, \dots, n$ are $\mathbf{e}_{m+2}, \dots, \mathbf{e}_n$.

Proof. Induction on m. Each $\mathbf{G}_i^{\top}$ fixes $\mathbf{e}_j$ for $j \notin \{i, i+1\}$, so $\mathbf{P}_m \mathbf{e}_j = \mathbf{e}_j$ whenever $j > m+1$, which is the second assertion. For $m = 1$, $\mathbf{P}_1 = \mathbf{G}_1^{\top}$ differs from $\mathbf{I}_n$ only in rows and columns $1, 2$, so its only nonzero entry below the diagonal is in position $(2, 1)$, and it is upper Hessenberg. Let $m \ge 2$ and suppose the claim holds for $m-1$. Right multiplication by $\mathbf{G}_m^{\top}$ changes only columns m and $m+1$, replacing them by the two combinations

$$c_m \, \mathbf{P}_{m-1} \mathbf{e}_m + s_m \, \mathbf{P}_{m-1} \mathbf{e}_{m+1} \quad \text{and} \quad -s_m \, \mathbf{P}_{m-1} \mathbf{e}_m + c_m \, \mathbf{P}_{m-1} \mathbf{e}_{m+1}.$$

By the inductive hypothesis $\mathbf{P}_{m-1} \mathbf{e}_m$ has zero entries in rows $m+2, \dots, n$, being column m of an upper Hessenberg matrix, and $\mathbf{P}_{m-1} \mathbf{e}_{m+1} = \mathbf{e}_{m+1}$ since $m+1 > (m-1)+1$. So both combinations vanish in rows $m+2, \dots, n$. Column m of $\mathbf{P}_m$ therefore vanishes below row $m+1$, as an upper Hessenberg matrix requires; and column $m+1$ vanishes below row $m+1$, which is more than required. All other columns are unchanged, so $\mathbf{P}_m$ is upper Hessenberg.

■

By the Claim, $\mathbf{Q} = \mathbf{P}_{n-1}$ is upper Hessenberg. Finally,

$$(\mathbf{R}\mathbf{Q})_{pq} = \sum_k r_{pk} q_{kq}$$

has a nonzero term only when $k \ge p$, because $\mathbf{R}$ is upper triangular, and $k \le q+1$, because $\mathbf{Q}$ is upper Hessenberg; together these force $p \le q+1$. So $\mathbf{R}\mathbf{Q}$ is upper Hessenberg.

Cost. Rotation $\mathbf{G}_i$ alters rows $i, i+1$ in columns $i, \dots, n$ only, since in the earlier columns both of those rows are already zero: that is $n - i + 1$ pairs of entries, each pair replaced by two combinations of two terms, so $6(n - i + 1)$ flops. Summing over $i = 1, \dots, n-1$ gives $3n^2 + O(n)$. Forming $\mathbf{R}\mathbf{Q} = \mathbf{R} \, \mathbf{G}_1^{\top} \cdots \mathbf{G}_{n-1}^{\top}$ applies $\mathbf{G}_i^{\top}$ on the right to columns $i, i+1$ of the matrix built so far, and the same induction as in the Claim shows that those two columns vanish below row $i+1$; so $6(i+1)$ flops suffice, and summing gives $3n^2 + O(n)$ again. The total is $6n^2 + O(n)$, which proves the proposition.

■

So the reduction is done once, for $O(n^3)$, and every subsequent step costs $O(n^2)$ — a saving of a factor of n on each of the many steps. That is the whole reason every implementation begins with Theorem 11.8.5.

Remark. The proposition is stated over $\mathbb{R}$ because Definition 11.8.2 is. Over $\mathbb{C}$ one uses the complex rotations described in Chapter 11 §08 and the argument is unchanged. The conclusion also does not depend on which QR factorization is used, as long as $\mathbf{Z}$ is invertible: two factorizations then differ by $\mathbf{Q} \mapsto \mathbf{QD}$ and $\mathbf{R} \mapsto \mathbf{D}^*\mathbf{R}$ with $\mathbf{D}$ unitary diagonal, and $(\mathbf{D}^*\mathbf{R})(\mathbf{QD}) = \mathbf{D}^*(\mathbf{RQ})\mathbf{D}$ has exactly the same pattern of zeros as $\mathbf{RQ}$. Invertibility is doing work there: for $\mathbf{Z} = \mathbf{e}_1\mathbf{e}_3^{\top} \in M_3(\mathbb{R})$, which is upper Hessenberg and singular, the reversal permutation $\mathbf{Q}$ (with $\mathbf{Q}\mathbf{e}_j = \mathbf{e}_{4-j}$) and $\mathbf{R} = \mathbf{e}_3\mathbf{e}_3^{\top}$ give an orthogonal $\mathbf{Q}$, an upper triangular $\mathbf{R}$, and $\mathbf{QR} = \mathbf{Z}$, while $\mathbf{RQ} = \mathbf{e}_3\mathbf{e}_1^{\top}$ has a nonzero $(3,1)$ entry and is not upper Hessenberg.

Quick check. A symmetric matrix is upper Hessenberg exactly when it is tridiagonal. What does Proposition 24.5.4 then say about a QR step applied to a symmetric tridiagonal matrix?

Solution. That the step returns a symmetric tridiagonal matrix. Symmetry survives because $\mathbf{RQ} = \mathbf{Q}^{\top}\mathbf{ZQ}$ is a congruence by an orthogonal matrix as well as a similarity, so $(\mathbf{Q}^{\top}\mathbf{ZQ})^{\top} = \mathbf{Q}^{\top}\mathbf{Z}^{\top}\mathbf{Q} = \mathbf{Q}^{\top}\mathbf{ZQ}$; and Hessenberg form survives by the proposition. A symmetric Hessenberg matrix has $z_{ij} = 0$ whenever $i > j+1$, hence also whenever $j > i+1$, which is tridiagonality. The cost then drops further, to $O(n)$ per step.

24.5.5 Shifts

The rate $|\lambda_{i+1}/\lambda_i|$ is a property of $\mathbf{A}$, and we cannot change $\mathbf{A}$. We can change the matrix we iterate on. For any $\mu \in F$, the matrix $\mathbf{A} - \mu\mathbf{I}$ has eigenvalues $\lambda_i - \mu$, so its ratios of moduli are $|\lambda_{i+1} - \mu|/|\lambda_i - \mu|$ — and a μ close to λ_n makes the last of these tiny. This is the same observation that turned the power method into inverse iteration in Section 4, where the shift μ was used to make $(\mathbf{A} - \mu\mathbf{I})^{-1}$ have a huge dominant eigenvalue (Theorem 24.4.3).

The **shifted QR step** is

$$\mathbf{A}^{(k)} - \mu_k\mathbf{I} = \mathbf{Q}^{(k)}\mathbf{R}^{(k)}, \qquad \mathbf{A}^{(k+1)} = \mathbf{R}^{(k)}\mathbf{Q}^{(k)} + \mu_k\mathbf{I},$$

and it is still a unitary similarity: $\mathbf{R}^{(k)}\mathbf{Q}^{(k)} = (\mathbf{Q}^{(k)})^*(\mathbf{A}^{(k)} - \mu_k\mathbf{I})\mathbf{Q}^{(k)}$, so adding $\mu_k\mathbf{I}$ back gives $\mathbf{A}^{(k+1)} = (\mathbf{Q}^{(k)})^*\mathbf{A}^{(k)}\mathbf{Q}^{(k)}$ exactly as in Proposition 24.5.1. The shift may be changed at every step, which is why the ratio can be driven down as the information improves. And since $\mathbf{A}^{(k)} - \mu_k\mathbf{I}$ is Hessenberg whenever $\mathbf{A}^{(k)}$ is, Proposition 24.5.4 applies to the shifted step: as it stands for a real μ_k, and in the complex form recorded in the Remark after it for a μ_k off the real line, which the second of the two shifts below produces.

Two shifts are standard. The **Rayleigh shift** takes $\mu_k = a_{nn}^{(k)}$, the trailing diagonal entry; when $\mathbf{A}$ is Hermitian this is the Rayleigh quotient $R_{\mathbf{A}^{(k)}}(\mathbf{e}_n)$ of Definition 17.1.1, hence a good eigenvalue estimate as soon as $\mathbf{e}_n$ is nearly an eigenvector. The **Wilkinson shift** takes μ_k to be

the eigenvalue of the trailing 2×2 block of $\mathbf{A}^{(k)}$ nearer to $a_{nn}^{(k)}$; it costs one quadratic formula and, unlike the Rayleigh shift, it can leave the real line, which is what lets it break the deadlock of a rotation.

The extreme case says exactly why shifting works.

Proposition 24.5.5. An Exact Shift Deflates in One Step

Let $\mathbf{Z} \in M_n(F)$ be upper Hessenberg, $n \geq 2$, with all its subdiagonal entries $z_{i+1,i}$ **nonzero** — such a matrix is called **unreduced**, the word Lemma 17.4.5 used for the tridiagonal case. Let $\mu \in F$ be an eigenvalue of $\mathbf{Z}$. Let $\mathbf{Z} - \mu\mathbf{I} = \mathbf{QR}$ with $\mathbf{Q}$ unitary and $\mathbf{R}$ upper triangular, and put $\mathbf{Z}' = \mathbf{RQ} + \mu\mathbf{I}$. Then the last row of $\mathbf{R}$ is zero, and the last row of $\mathbf{Z}'$ is $\mu\,\mathbf{e}_n^{\top}$. In particular $z'_{n,n-1} = 0$ and $z'_{nn} = \mu$.

Idea. An unreduced Hessenberg matrix cannot lose more than one dimension of rank: its first $n-1$ columns are independent whatever the matrix does, because the subdiagonal alone makes them so. Shifting by an eigenvalue makes the matrix singular, so the one missing dimension must be the last one, and in a QR factorization "the last dimension is missing" reads as a zero last row of $\mathbf{R}$.

Proof. Write $\mathbf{B} = \mathbf{Z} - \mu\mathbf{I}$, again upper Hessenberg with the same nonzero subdiagonal entries.

First, the columns $\mathbf{b}_1, \ldots, \mathbf{b}_{n-1}$ of $\mathbf{B}$ are linearly independent. Delete the first row: what remains is the $(n-1) \times (n-1)$ matrix with entries $b_{i+1,j}$ for $1 \leq i, j \leq n-1$, which is **lower** triangular, since $b_{i+1,j} = 0$ for $i+1 > j+1$, that is for $i > j$; and its diagonal entries $b_{j+1,j}$ are nonzero. A triangular matrix with nonzero diagonal is invertible (Lemma 3.6.2), so those $n-1$ truncated columns are independent, and the full columns are independent a fortiori.

Second, $\operatorname{rank}\mathbf{B} = n-1$: it is at least $n-1$ by the previous paragraph and at most $n-1$ because $\mathbf{B}$ is singular, μ being an eigenvalue of $\mathbf{Z}$.

Now $\mathbf{B} = \mathbf{QR}$ with $\mathbf{Q}$ invertible, so $\operatorname{rank}\mathbf{R} = \operatorname{rank}\mathbf{B} = n-1$ and the columns $\mathbf{Re}_1, \ldots, \mathbf{Re}_{n-1}$ are independent, being $\mathbf{Q}^*\mathbf{b}_1, \ldots, \mathbf{Q}^*\mathbf{b}_{n-1}$. Since $\mathbf{R}$ is upper triangular, its last row has the single possibly nonzero entry r_{nn}; if $r_{nn} \neq 0$, then $\mathbf{R}$ would be invertible (Lemma 3.6.2 applied after noting that independence of the first $n-1$ columns forces $r_{11}, \ldots, r_{n-1,n-1} \neq 0$), contradicting $\operatorname{rank}\mathbf{R} = n-1$. Hence $r_{nn} = 0$ and the last row of $\mathbf{R}$ is zero.

A matrix with zero last row keeps it under multiplication on the right: $(\mathbf{RQ})_{nj} = \sum_k r_{nk} q_{kj} = 0$. So the last row of $\mathbf{RQ}$ is zero and the last row of $\mathbf{Z}' = \mathbf{RQ} + \mu\mathbf{I}$ is $\mu\,\mathbf{e}_n^{\top}$, as claimed.

■

The proposition is the ideal case, and it explains the practice: the closer μ_k is to an eigenvalue, the smaller $z'_{n,n-1}$ is observed to become, and the algorithm watches that one entry. When it falls below a tolerance it is set to zero, the matrix splits into a smaller Hessenberg block and a known eigenvalue, and the iteration restarts on the block. This is **deflation**, and with it the algorithm is observed to finish after $O(n)$ steps in total rather than converging forever.

Two claims in that paragraph are descriptions of observed behavior, not theorems, and this book proves neither: that a nearer shift always gives a smaller subdiagonal entry,

and that the shifted algorithm terminates after $O(n)$ steps. Nothing in this section or later depends on them. What is proved is the exact case, Proposition 24.5.5, and the example below exhibits it.

Example. The shift repairs the rotation, and deflates a symmetric example

(a) Apply one shifted QR step with $\mu = i$ to $\mathbf{B} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$.

(b) Apply one shifted QR step with $\mu = 2$ to the symmetric tridiagonal matrix $\mathbf{A}$ with diagonal $(2, 2, 2)$ and off-diagonal entries 1.

Solution.

(a) Here $\mu = i$ is an eigenvalue, so Proposition 24.5.5 predicts an exact deflation. Indeed

$$\mathbf{B} - i\mathbf{I} = \begin{pmatrix} -i & -1 \\ 1 & -i \end{pmatrix},$$

whose second column is $-i$ times the first, so the matrix has rank 1. Its first column has norm $\sqrt{2}$, so take $\mathbf{q}_1 = \frac{1}{\sqrt{2}}(-i, 1)$ and $\mathbf{q}_2 = \frac{1}{\sqrt{2}}(1, -i)$, which is a unit vector orthogonal to $\mathbf{q}_1$ since $\overline{(-i)} \cdot 1 + \overline{1} \cdot (-i) = i - i = 0$. With $\mathbf{Q} = (\mathbf{q}_1 \mid \mathbf{q}_2)$ one computes $\mathbf{R} = \mathbf{Q}^*(\mathbf{B} - i\mathbf{I})$, which is

$$\mathbf{R} = \begin{pmatrix} \sqrt{2} & -i\sqrt{2} \\ 0 & 0 \end{pmatrix},$$

with zero last row as predicted. Then $\mathbf{RQ} = \operatorname{diag}(-2i, 0)$, and

$$\mathbf{B}' = \mathbf{RQ} + i\mathbf{I} = \operatorname{diag}(-i,\, i).$$

One step, and the eigenvalues are on the diagonal. The unshifted iteration would have run forever.

(b) The matrix is

$$\mathbf{A} = \begin{pmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{pmatrix}, \qquad \mathbf{A} - 2\mathbf{I} = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix},$$

and 2 is an eigenvalue of $\mathbf{A}$, since $(\mathbf{A} - 2\mathbf{I})(1, 0, -1) = \mathbf{0}$. Two Givens rotations triangularize $\mathbf{A} - 2\mathbf{I}$, as in the proof of Proposition 24.5.4. The first uses the entries 0 and 1 in positions $(1, 1)$ and $(2, 1)$, so $r = 1$, $c_1 = 0$, $s_1 = 1$, and $\mathbf{G}_1$ exchanges rows 1 and 2 with a sign, producing rows $(1, 0, 1)$, $(0, -1, 0)$, $(0, 1, 0)$. The second uses the entries -1 and 1 now in positions $(2, 2)$ and $(3, 2)$, so $r = \sqrt{2}$, $c_2 = -1/\sqrt{2}$, $s_2 = 1/\sqrt{2}$, and it turns those two rows into $(0, \sqrt{2}, 0)$ and $\mathbf{0}$. Hence

$$\mathbf{R} = \begin{pmatrix} 1 & 0 & 1 \\ 0 & \sqrt{2} & 0 \\ 0 & 0 & 0 \end{pmatrix}, \qquad \mathbf{Q} = \mathbf{G}_1^{\top}\mathbf{G}_2^{\top} = \begin{pmatrix} 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ 1 & 0 & 0 \\ 0 & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix},$$

with the last row of $\mathbf{R}$ zero, as Proposition 24.5.5 predicts. One checks $\mathbf{QR} = \mathbf{A} - 2\mathbf{I}$

column by column. Multiplying back and restoring the shift,

$$\mathbf{A}' = \mathbf{RQ} + 2\mathbf{I} = \begin{pmatrix} 2 & \sqrt{2} & 0 \\ \sqrt{2} & 2 & 0 \\ 0 & 0 & 2 \end{pmatrix}.$$

The last row and column have split off, carrying the eigenvalue 2, and the remaining 2×2 block has eigenvalues $2 \pm \sqrt{2}$ — the other two eigenvalues of $\mathbf{A}$. One shifted step, and the problem is finished.

24.5.6 Three unshifted steps, watched

Example. The subdiagonal shrinking at the predicted rate

Run the unshifted QR iteration on the symmetric tridiagonal matrix

$$\mathbf{A} = \begin{pmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{pmatrix},$$

whose eigenvalues are $2 + \sqrt{2}$, 2, and $2 - \sqrt{2}$. Carry out the first step exactly, and tabulate three steps.

Solution. Step one, by hand. Apply Theorem 11.2.2 to the columns $\mathbf{a}_1 = (2, 1, 0)$, $\mathbf{a}_2 = (1, 2, 1)$, $\mathbf{a}_3 = (0, 1, 2)$.

From $\mathbf{a}_1$: $r_{11} = \|\mathbf{a}_1\| = \sqrt{5}$ and $\mathbf{q}_1 = \frac{1}{\sqrt{5}}(2, 1, 0)$.

From $\mathbf{a}_2$: $r_{12} = \langle \mathbf{a}_2, \mathbf{q}_1 \rangle = 4/\sqrt{5}$, and

$$\mathbf{w}_2 = \mathbf{a}_2 - \tfrac{4}{5}(2, 1, 0) = \tfrac{1}{5}(-3, 6, 5), \qquad r_{22} = \|\mathbf{w}_2\| = \sqrt{14/5},$$

since $9 + 36 + 25 = 70$ and $\sqrt{70}/5 = \sqrt{14/5}$; so $\mathbf{q}_2 = \frac{1}{\sqrt{70}}(-3, 6, 5)$.

From $\mathbf{a}_3$: $r_{13} = 1/\sqrt{5}$ and $r_{23} = 16/\sqrt{70}$, and

$$\mathbf{w}_3 = \mathbf{a}_3 - \tfrac{1}{5}(2, 1, 0) - \tfrac{16}{70}(-3, 6, 5) = \tfrac{2}{7}(1, -2, 3), \qquad r_{33} = \tfrac{2\sqrt{14}}{7},$$

so $\mathbf{q}_3 = \frac{1}{\sqrt{14}}(1, -2, 3)$. Now $\mathbf{A}^{(2)} = \mathbf{RQ}$. Its $(1,1)$ entry is the first row of $\mathbf{R}$ against the first column of $\mathbf{Q}$, namely

$$\sqrt{5} \cdot \tfrac{2}{\sqrt{5}} + \tfrac{4}{\sqrt{5}} \cdot \tfrac{1}{\sqrt{5}} + \tfrac{1}{\sqrt{5}} \cdot 0 = 2 + \tfrac{4}{5} = \tfrac{14}{5},$$

and its $(2,1)$ entry is $\sqrt{14/5} \cdot \frac{1}{\sqrt{5}} + \frac{16}{\sqrt{70}} \cdot 0 = \sqrt{14}/5$. Continuing,

$$\mathbf{A}^{(2)} = \begin{pmatrix} \frac{14}{5} & \frac{\sqrt{14}}{5} & 0 \\ \frac{\sqrt{14}}{5} & \frac{82}{35} & \frac{2\sqrt{5}}{7} \\ 0 & \frac{2\sqrt{5}}{7} & \frac{6}{7} \end{pmatrix}.$$

Three checks: the matrix is symmetric and tridiagonal, as the Quick check after Proposition 24.5.4 predicted; its trace is $\frac{14}{5}+\frac{82}{35}+\frac{6}{7}=\frac{98+82+30}{35}=6=\operatorname{tr}\mathbf{A}$; and the entry $(3,1)$ is zero, not merely small.

Three steps, to eight places. The diagonal entries and the two subdiagonal entries run as follows.

k	$a_{11}^{(k)}$	$a_{22}^{(k)}$	$a_{33}^{(k)}$	$\lvert a_{21}^{(k)}\rvert$	$\lvert a_{32}^{(k)}\rvert$
1	2.00000000	2.00000000	2.00000000	1.000000	1.000000
2	2.80000000	2.34285714	0.85714286	0.748331	0.638877
3	3.14285714	2.24844720	0.60869565	0.559397	0.187848
4	3.30841121	2.10394692	0.58764187	0.372193	0.052177

The three diagonal entries are heading for $2+\sqrt{2}=3.41421356$, 2, and $2-\sqrt{2}=0.58578644$, largest modulus first — the order the convergence theorem quoted above predicts, exhibited here rather than proved. The two subdiagonal entries shrink by factors

$$0.748,\ 0.748,\ 0.665,\ldots \qquad \text{and} \qquad 0.639,\ 0.294,\ 0.278,\ldots$$

The predicted asymptotic factors are $|\lambda_2/\lambda_1| = 2/(2+\sqrt{2}) = 0.585786$ and $|\lambda_3/\lambda_2| = (2-\sqrt{2})/2 = 0.292893$. The second entry has already settled onto its rate; the first is still approaching it, and by step 10 its factor is 0.5859. Convergence is linear, and slow: each step buys about a quarter of a decimal digit in the $(2,1)$ entry. That is the cost of not shifting.

24.5.7 Exercises

24.5.7.1 A. Check your understanding

Exercise. A1

(a) Write down one step of the unshifted QR iteration, and one step of the shifted one, and prove in one line each that both are similarities.

(b) State Theorem 24.5.2, and say in one sentence what it has to do with the power method.

(c) True or false: the QR iterates of a matrix always converge to an upper triangular matrix. Justify your answer.

(d) Why is the iteration run on an upper Hessenberg matrix rather than on a full one? Give the two costs.

(e) In Proposition 24.5.5, which hypothesis on $\mathbf{Z}$ is doing the work, and where exactly is it used?

Solution.

(a) Unshifted: $\mathbf{A}^{(k)}=\mathbf{QR}$, $\mathbf{A}^{(k+1)}=\mathbf{RQ}$; then $\mathbf{RQ}=\mathbf{Q}^*\mathbf{QRQ}=\mathbf{Q}^*\mathbf{A}^{(k)}\mathbf{Q}$. Shifted: $\mathbf{A}^{(k)}-\mu\mathbf{I}=\mathbf{QR}$, $\mathbf{A}^{(k+1)}=\mathbf{RQ}+\mu\mathbf{I}$; then $\mathbf{RQ}+\mu\mathbf{I}=\mathbf{Q}^*(\mathbf{A}^{(k)}-\mu\mathbf{I})\mathbf{Q}+\mu\mathbf{Q}^*\mathbf{Q}=\mathbf{Q}^*\mathbf{A}^{(k)}\mathbf{Q}$.

(b) It says $\mathbf{A}^k=\mathbf{U}_k\mathbf{T}_k$ with $\mathbf{U}_k=\mathbf{Q}^{(1)}\cdots\mathbf{Q}^{(k)}$ unitary and $\mathbf{T}_k=\mathbf{R}^{(k)}\cdots\mathbf{R}^{(1)}$ upper triangular, and that $\mathbf{A}^{(k+1)}=\mathbf{U}_k^*\mathbf{A}\mathbf{U}_k$. Comparing first columns, the first column of $\mathbf{U}_k$ is $\mathbf{A}^k\mathbf{e}_1$ normalized, which is the k-th power-method iterate started at $\mathbf{e}_1$.

(c) False. The rotation $\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ is a fixed point of the unshifted iteration (Example (Two matrices the iteration cannot move)) and is not triangular; its two eigenvalues have the same modulus, which is exactly the hypothesis Theorem 24.5.3 needs.

(d) Because a QR step on a full matrix costs $O(n^3)$ flops, with the count genuinely growing like n^3, while a step on a Hessenberg matrix costs about $6n^2$ (Proposition 24.5.4), and the pattern is preserved, so the saving is repeated at every step. The reduction itself is $O(n^3)$, paid once (Theorem 11.8.5).

(e) That every subdiagonal entry of $\mathbf{Z}$ is nonzero. It is used exactly once, to show that the first $n-1$ columns of $\mathbf{Z}-\mu\mathbf{I}$ are independent: deleting the first row leaves a lower triangular matrix whose diagonal is that subdiagonal. Without it $\mathbf{Z}-\mu\mathbf{I}$ could drop rank in an earlier column, and the zero row of $\mathbf{R}$ need not be the last.

24.5.7.2 B. Practice

Exercise. B1: One step by hand

Let $\mathbf{A} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$. Compute the QR factorization of $\mathbf{A}$ in the sense of Theorem 11.8.1, then $\mathbf{A}^{(2)} = \mathbf{RQ}$. Verify that $\mathbf{A}^{(2)}$ is symmetric, has trace 4 and determinant 3, and that its off-diagonal entry is smaller than that of $\mathbf{A}$. Hence estimate how many steps are needed to make the off-diagonal entry smaller than 10^{-3}.

Solution. The columns are $\mathbf{a}_1 = (2,1)$ and $\mathbf{a}_2 = (1,2)$. Then $r_{11} = \sqrt{5}$, $\mathbf{q}_1 = \frac{1}{\sqrt{5}}(2,1)$, $r_{12} = \langle \mathbf{a}_2, \mathbf{q}_1 \rangle = 4/\sqrt{5}$, and

$$\mathbf{w}_2 = (1,2) - \tfrac{4}{5}(2,1) = \tfrac{3}{5}(-1,2), \qquad r_{22} = \tfrac{3}{\sqrt{5}}, \qquad \mathbf{q}_2 = \tfrac{1}{\sqrt{5}}(-1,2).$$

So

$$\mathbf{Q} = \tfrac{1}{\sqrt{5}} \begin{pmatrix} 2 & -1 \\ 1 & 2 \end{pmatrix}, \qquad \mathbf{R} = \tfrac{1}{\sqrt{5}} \begin{pmatrix} 5 & 4 \\ 0 & 3 \end{pmatrix},$$

and

$$\mathbf{A}^{(2)} = \mathbf{RQ} = \tfrac{1}{5} \begin{pmatrix} 5 & 4 \\ 0 & 3 \end{pmatrix} \begin{pmatrix} 2 & -1 \\ 1 & 2 \end{pmatrix} = \tfrac{1}{5} \begin{pmatrix} 14 & 3 \\ 3 & 6 \end{pmatrix}.$$

It is symmetric; its trace is $20/5 = 4 = \operatorname{tr} \mathbf{A}$; its determinant is $(84-9)/25 = 3 = \det \mathbf{A}$. The off-diagonal entry fell from 1 to $3/5$.

The eigenvalues are 3 and 1, so Theorem 24.5.3 gives the asymptotic factor $t = 1/3$ per step. To take $3/5$ below 10^{-3} needs $(1/3)^m \le 10^{-3}/(3/5)$, that is $m \ge \log(600)/\log 3 \approx 5.8$: about six further steps, so seven in all. (The first steps are a little slower than the asymptotic rate, so eight is a safer count.)

Exercise. B2: Fixed points

Determine which of the following matrices are fixed points of the unshifted QR iteration, that is, satisfy $\mathbf{A}^{(2)} = \mathbf{A}^{(1)}$. Justify your answer; for those that are not, name what changes.

(a) any upper triangular $\mathbf{A}$ with positive real diagonal entries;
(b) any unitary $\mathbf{A}$;
(c) $\mathbf{A} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$;
(d) any normal $\mathbf{A}$.

Solution.

(a) Yes. Then $\mathbf{A} = \mathbf{I}\mathbf{A}$ is a factorization with $\mathbf{I}$ unitary and $\mathbf{A}$ upper triangular with positive diagonal, hence the unique one (Theorem 11.8.1), and $\mathbf{A}^{(2)} = \mathbf{A}\mathbf{I} = \mathbf{A}$.

(b) Yes. Then $\mathbf{A} = \mathbf{A}\mathbf{I}$ is a factorization with $\mathbf{A}$ unitary and $\mathbf{I}$ upper triangular with positive diagonal, hence the unique one, and $\mathbf{A}^{(2)} = \mathbf{I}\mathbf{A} = \mathbf{A}$.

(c) Yes, by (b): the matrix is a permutation matrix, so its columns are orthonormal. This is worth noticing because its eigenvalues are 1 and -1, of equal modulus — another instance of the obstruction of Example (The unshifted algorithm on a rotation), and this one is symmetric, so unlike the rotation it *is* orthogonally similar to a diagonal matrix. The unshifted iteration simply will not find it.

(d) No. A normal matrix need not be fixed: $\mathbf{A} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$ is symmetric, hence normal, and Exercise (B1: One step by hand) computed $\mathbf{A}^{(2)} = \frac{1}{5}\begin{pmatrix} 14 & 3 \\ 3 & 6 \end{pmatrix} \neq \mathbf{A}$. What is preserved is normality, not the matrix: $\mathbf{A}^{(2)} = \mathbf{Q}^{*}\mathbf{A}\mathbf{Q}$ is unitarily similar to $\mathbf{A}$, and normality is invariant under unitary similarity.

Exercise. B3: A shift that deflates

Let $\mathbf{Z} = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$. Take μ to be either eigenvalue of $\mathbf{Z}$ and carry out one shifted QR step. Verify the conclusion of Proposition 24.5.5. *Hint: you need only the last row of* $\mathbf{R}$.

Solution. The characteristic polynomial is $x^2 - 5x - 2$, with roots $\mu_{\pm} = (5 \pm \sqrt{33})/2$. Every 2×2 matrix is upper Hessenberg, and the single subdiagonal entry 3 is nonzero, so Proposition 24.5.5 applies with $n = 2$. Take $\mu = \mu_{+}$. Then

$$\mathbf{Z} - \mu\mathbf{I} = \begin{pmatrix} 1-\mu & 2 \\ 3 & 4-\mu \end{pmatrix}$$

is singular, of rank 1 since it is nonzero. Its first column has norm $r = \sqrt{(1-\mu)^2 + 9} \neq 0$, so a QR factorization has $\mathbf{q}_1 = \frac{1}{r}(1-\mu, 3)$ and $r_{11} = r \neq 0$. Because the rank is 1, the second column of $\mathbf{R}$ is a multiple of $\mathbf{e}_1$, so $r_{22} = 0$ and the last row of $\mathbf{R}$ vanishes. Then the last row of $\mathbf{R}\mathbf{Q}$ vanishes too, and $\mathbf{Z}' = \mathbf{R}\mathbf{Q} + \mu\mathbf{I}$ has last row $(0, \mu)$. Its $(1,1)$ entry is then $\operatorname{tr}\mathbf{Z} - \mu = 5 - \mu = \mu_{-}$, since the trace is preserved. One step produced both eigenvalues exactly.

24.5.7.3 C. Going deeper

Exercise. C1: Hessenberg form is not enough by itself

(a) Prove that if $\mathbf{Z}$ is upper Hessenberg with $z_{m+1,m} = 0$ for some $1 \le m \le n-1$, then $\operatorname{spec}(\mathbf{Z}) = \operatorname{spec}(\mathbf{Z}_{11}) \cup \operatorname{spec}(\mathbf{Z}_{22})$, where $\mathbf{Z}_{11}$ is the leading $m \times m$ block and $\mathbf{Z}_{22}$ the trailing $(n-m) \times (n-m)$ block.

(b) Deduce that a QR implementation may, at every step, replace the problem by two smaller ones as soon as one subdiagonal entry vanishes.

(c) Give a 3×3 upper Hessenberg matrix with no zero subdiagonal entry for which the unshifted iteration never reduces any subdiagonal entry at all.

Solution.

(a) With $z_{m+1,m} = 0$ and $\mathbf{Z}$ upper Hessenberg, every entry z_{ij} with $i > m \ge j$ vanishes: for $i > m+1$ this is $i > j+1$, and for $i = m+1$ it is either $j < m$, again $i > j+1$, or $j = m$, the assumed entry. So $\mathbf{Z}$ is block upper triangular with diagonal blocks $\mathbf{Z}_{11}$ and $\mathbf{Z}_{22}$. By Theorem 8.1.1 and the determinant of a block triangular matrix (Theorem 7.8.4), $p_{\mathbf{Z}}(x) = p_{\mathbf{Z}_{11}}(x)\, p_{\mathbf{Z}_{22}}(x)$, so the eigenvalues of $\mathbf{Z}$ are those of the two blocks together.

(b) The two blocks are again upper Hessenberg, and by (a) computing their spectra computes that of $\mathbf{Z}$. Each step on a block of size m costs about $6m^2$ rather than $6n^2$ (Proposition 24.5.4), so splitting is a strict gain; and part (a) guarantees no eigenvalue is lost. This is deflation.

(c) Take the cyclic permutation matrix

$$\mathbf{Z} = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix},$$

which is upper Hessenberg with subdiagonal entries $1, 1$, and is orthogonal. By Exercise (B2: Fixed points) (b) it is a fixed point of the unshifted iteration, so no subdiagonal entry ever moves. Its eigenvalues are the three cube roots of 1, all of modulus 1 — the obstruction of Theorem 24.5.3 in its purest form.

Exercise. C2: Hermitian iterates and the Rayleigh quotient

Let $\mathbf{A} \in M_n(\mathbb{C})$ be Hermitian and let $\mathbf{A}^{(k)}$ be its unshifted QR iterates.

(a) Prove that every $\mathbf{A}^{(k)}$ is Hermitian with the same eigenvalues as $\mathbf{A}$.

(b) Prove that $a_{nn}^{(k)} = R_{\mathbf{A}}(\mathbf{u}_n^{(k)})$, where $\mathbf{u}_n^{(k)}$ is the last column of $\mathbf{U}_{k-1}$ and $R_{\mathbf{A}}$ is the Rayleigh quotient of Definition 17.1.1.

(c) Hence explain, citing Theorem 20.6.1, why the Rayleigh shift $\mu_k = a_{nn}^{(k)}$ is a certified eigenvalue estimate, and say what the certificate is.

Solution.

(a) By Theorem 24.5.2, $\mathbf{A}^{(k)} = \mathbf{U}_{k-1}^{*}\mathbf{A}\mathbf{U}_{k-1}$, with $\mathbf{U}_0 = \mathbf{I}$ at $k = 1$, so $(\mathbf{A}^{(k)})^{*} =$

$\mathbf{U}_{k-1}^{*}\mathbf{A}^{*}\mathbf{U}_{k-1} = \mathbf{A}^{(k)}$, and unitary similarity preserves eigenvalues.

(b) Writing $\mathbf{u}_n = \mathbf{U}_{k-1}\mathbf{e}_n$, which is a unit vector since $\mathbf{U}_{k-1}$ is unitary,

$$a_{nn}^{(k)} = \mathbf{e}_n^{*}\mathbf{A}^{(k)}\mathbf{e}_n = \mathbf{e}_n^{*}\mathbf{U}_{k-1}^{*}\mathbf{A}\mathbf{U}_{k-1}\mathbf{e}_n = \mathbf{u}_n^{*}\mathbf{A}\mathbf{u}_n = R_{\mathbf{A}}(\mathbf{u}_n),$$

the last equality because $R_{\mathbf{A}}(\mathbf{x}) = \mathbf{x}^{*}\mathbf{A}\mathbf{x}/\mathbf{x}^{*}\mathbf{x}$ and $\mathbf{u}_n^{*}\mathbf{u}_n = 1$.

(c) Put $\mu = a_{nn}^{(k)}$ and $\mathbf{r} = \mathbf{A}\mathbf{u}_n - \mu\mathbf{u}_n$. Then μ is real, $\mathbf{A}$ being Hermitian, and Theorem 20.6.1 says that $\mathbf{A}$ has an eigenvalue in $[\mu - \|\mathbf{r}\|_2, \mu + \|\mathbf{r}\|_2]$. The certificate is $\|\mathbf{r}\|_2$, and it costs nothing extra: by the computation in (b), $\mathbf{U}_{k-1}^{*}\mathbf{r} = \mathbf{A}^{(k)}\mathbf{e}_n - \mu\mathbf{e}_n$ is the last column of $\mathbf{A}^{(k)}$ with its diagonal entry removed, so $\|\mathbf{r}\|_2$ is the Euclidean norm of the entries of that column other than the diagonal one. If $\mathbf{A}$ was first reduced to Hessenberg form, so that every $\mathbf{A}^{(k)}$ is Hermitian and upper Hessenberg, hence tridiagonal, that norm is the single entry $|a_{n-1,n}^{(k)}|$. The algorithm is already watching that number.

Exercise. C3: Why the powers are the right object

Let $\mathbf{A} \in M_n(\mathbb{C})$ be invertible, with unshifted QR iterates as in Definition 24.5.1.

(a) Prove that for $1 \le j \le n$, the subspace spanned by the first j columns of $\mathbf{U}_k$ equals $\mathbf{A}^k(\operatorname{span}(\mathbf{e}_1, \dots, \mathbf{e}_j))$.

(b) Deduce that if the first j columns of $\mathbf{U}_k$ converge to a basis of an $\mathbf{A}$-invariant subspace W, then the entries of $\mathbf{A}^{(k+1)}$ in rows $j+1, \dots, n$ of columns $1, \dots, j$ tend to 0.

Hint for (b): write the entries as inner products of $\mathbf{A}$ *applied to one column against another.*

Solution.

(a) By Theorem 24.5.2, $\mathbf{A}^k = \mathbf{U}_k\mathbf{T}_k$ with $\mathbf{T}_k$ upper triangular and invertible (Quick check above), so $\mathbf{A}^k\mathbf{e}_i = \mathbf{U}_k\mathbf{T}_k\mathbf{e}_i$ lies in $\operatorname{span}(\mathbf{U}_k\mathbf{e}_1, \dots, \mathbf{U}_k\mathbf{e}_i)$ for each i, since $\mathbf{T}_k\mathbf{e}_i$ has zeros below row i. Hence $\mathbf{A}^k(\operatorname{span}(\mathbf{e}_1, \dots, \mathbf{e}_j)) \subseteq \operatorname{span}(\mathbf{U}_k\mathbf{e}_1, \dots, \mathbf{U}_k\mathbf{e}_j)$. Both sides have dimension j: the left because $\mathbf{A}^k$ is invertible, the right because the columns of a unitary matrix are independent. One inclusion plus equal dimension gives equality.

(b) Write $\mathbf{u}_i = \mathbf{U}_k\mathbf{e}_i$. For $p > j \ge q$,

$$a_{pq}^{(k+1)} = \mathbf{e}_p^{*}\mathbf{U}_k^{*}\mathbf{A}\mathbf{U}_k\mathbf{e}_q = \mathbf{u}_p^{*}\mathbf{A}\mathbf{u}_q = \langle \mathbf{A}\mathbf{u}_q, \mathbf{u}_p \rangle.$$

Suppose $\mathbf{u}_1, \dots, \mathbf{u}_j$ converge to an orthonormal basis $\mathbf{w}_1, \dots, \mathbf{w}_j$ of an $\mathbf{A}$-invariant subspace W, let P be the orthogonal projection onto W (Definition 11.3.2) and put $\delta_k = \max_{q \le j} \|\mathbf{u}_q - \mathbf{w}_q\|_2$, so $\delta_k \to 0$. Since $\mathbf{A}\mathbf{w}_q \in W$, we have $(\mathbf{I} - P)\mathbf{A}\mathbf{w}_q = \mathbf{0}$ and hence $\|(\mathbf{I} - P)\mathbf{A}\mathbf{u}_q\|_2 = \|(\mathbf{I} - P)\mathbf{A}(\mathbf{u}_q - \mathbf{w}_q)\|_2 \le \|\mathbf{A}\|_2\delta_k$. Since $\langle \mathbf{u}_p, \mathbf{u}_q \rangle = 0$ for $q \le j < p$, we have $|\langle \mathbf{u}_p, \mathbf{w}_q \rangle| = |\langle \mathbf{u}_p, \mathbf{w}_q - \mathbf{u}_q \rangle| \le \delta_k$ by Cauchy–Schwarz (Theorem 11.1.5), so $\|P\mathbf{u}_p\|_2 \le j\,\delta_k$. Splitting $\mathbf{u}_p$ and using that P is self-adjoint,

$$\begin{aligned} |\langle \mathbf{A}\mathbf{u}_q, \mathbf{u}_p \rangle| &= |\langle (\mathbf{I} - P)\mathbf{A}\mathbf{u}_q, \mathbf{u}_p \rangle + \langle \mathbf{A}\mathbf{u}_q, P\mathbf{u}_p \rangle| \\ &\le \|(\mathbf{I} - P)\mathbf{A}\mathbf{u}_q\|_2 + \|\mathbf{A}\|_2\|P\mathbf{u}_p\|_2 \;\le\; (1 + j)\|\mathbf{A}\|_2\,\delta_k, \end{aligned}$$

by Cauchy–Schwarz again and $\|\mathbf{u}_p\|_2 = \|\mathbf{u}_q\|_2 = 1$. The right-hand side tends to 0, so $a_{pq}^{(k+1)} \to 0$. This is Theorem 24.5.3 for $j = 1$, read as a statement about invariant

subspaces rather than about eigenvectors, and it is the form in which the general convergence theorem quoted above is proved.

24.6 Krylov Subspaces

Sections 4 and 5 both consist of multiplying by $\mathbf{A}$ over and over: the power method forms $\mathbf{b}, \mathbf{Ab}, \mathbf{A}^2\mathbf{b}, \dots$ and keeps only the newest vector, and the QR algorithm, by Theorem 24.5.2, does the same thing to $\mathbf{e}_1$ inside its first column. In both cases everything computed before the last step is thrown away, and this section keeps it instead. The span of the vectors produced by k matrix-vector products is a subspace of dimension up to k, and the best approximations it contains are far better than its last vector alone. That subspace, the process that builds an orthonormal basis of it, and the eigenvalue estimates it yields are the subject here.

Throughout, $F = \mathbb{R}$ or $F = \mathbb{C}$, and F^n carries the standard inner product, linear in the first slot, with $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{y}^*\mathbf{x}$ and $\|\cdot\| = \|\cdot\|_2$.

24.6.1 The space that k products span

A large matrix is often not available as a table of n^2 numbers at all. What is available is a rule for computing $\mathbf{Ax}$ from $\mathbf{x}$ — a sparse pattern of neighbors on a grid, a graph, a differential operator discretized. For such a matrix, forming $\mathbf{Ax}$ is cheap and forming $\mathbf{A}^{-1}$, or even looking at an entry of $\mathbf{A}^2$, is not. So the honest question is: after k multiplications, what do we possess?

We possess $\mathbf{b}, \mathbf{Ab}, \dots, \mathbf{A}^{k-1}\mathbf{b}$, and hence every linear combination of them. The power method uses one of these vectors. That is wasteful, and the waste has a name only because the span deserves one.

The Krylov space is everything that k matrix-vector products can reach.

Definition 24.6.1. Krylov Subspace

Let $\mathbf{A} \in M_n(F)$, let $\mathbf{b} \in F^n$, and let $k \geq 1$ be an integer. The **k-th Krylov subspace** of the pair $(\mathbf{A}, \mathbf{b})$ is

$$\mathcal{K}_k(\mathbf{A}, \mathbf{b}) \coloneqq \operatorname{span}\left(\mathbf{b},\ \mathbf{Ab},\ \mathbf{A}^2\mathbf{b},\ \dots,\ \mathbf{A}^{k-1}\mathbf{b}\right) \subseteq F^n.$$

In words: we list the first k vectors of the sequence obtained by repeatedly applying $\mathbf{A}$ to $\mathbf{b}$ — **starting at $\mathbf{b}$ itself**, so that the last one is $\mathbf{A}^{k-1}\mathbf{b}$ and not $\mathbf{A}^k\mathbf{b}$ — and take their span. It is a subspace for the usual reason, being a span. The index k counts **vectors**, not powers; the off-by-one is a convention, and it is the one that makes $\dim \mathcal{K}_k \leq k$ with equality in the typical case.

Examples. Take $F = \mathbb{R}$ throughout.

- **The identity.** For $\mathbf{A} = \mathbf{I}_n$ and any $\mathbf{b} \neq \mathbf{0}$, every $\mathbf{A}^j\mathbf{b}$ equals $\mathbf{b}$, so $\mathcal{K}_k(\mathbf{I}_n, \mathbf{b}) = \operatorname{span}(\mathbf{b})$ for every k. One product buys nothing, and neither do a hundred.
- **A companion matrix.** Let $\mathbf{Ae}_j = \mathbf{e}_{j+1}$ for $j \leq n-1$, as for the companion matrix of Definition 10.5.3, and let $\mathbf{b} = \mathbf{e}_1$. Then $\mathbf{A}^j\mathbf{e}_1 = \mathbf{e}_{j+1}$ for $j \leq n-1$, so $\mathcal{K}_k(\mathbf{A}, \mathbf{e}_1) = \operatorname{span}(\mathbf{e}_1, \dots, \mathbf{e}_k)$ for $k \leq n$, of dimension exactly k. Here every product buys a new dimension.
- **A diagonal matrix.** Let $\mathbf{A} = \operatorname{diag}(4, 3, 2, 1)$ and $\mathbf{b} = (1, 1, 1, 1)$. Then $\mathbf{A}^j\mathbf{b} =$

$(4^j, 3^j, 2^j, 1)$, and these four vectors for $j = 0, 1, 2, 3$ form a Vandermonde matrix with distinct nodes, hence are independent (Theorem 7.8.1). So $\dim \mathcal{K}_k = k$ for $k \le 4$, and $\mathcal{K}_4 = \mathbb{R}^4$.

- **Degenerate cases.** $\mathcal{K}_1(\mathbf{A}, \mathbf{b}) = \operatorname{span}(\mathbf{b})$; and $\mathcal{K}_k(\mathbf{A}, \mathbf{0}) = \{\mathbf{0}\}$ for every k, which is why $\mathbf{b} \neq \mathbf{0}$ is assumed from here on.

Nonexample by minimal change. Drop the middle vector of the list. Let $\mathbf{A} \in M_3(\mathbb{R})$ satisfy $\mathbf{A}\mathbf{e}_1 = \mathbf{e}_2$, $\mathbf{A}\mathbf{e}_2 = \mathbf{e}_3$, $\mathbf{A}\mathbf{e}_3 = \mathbf{0}$, and let $\mathbf{b} = \mathbf{e}_1$. The set $\operatorname{span}(\mathbf{b}, \mathbf{A}^2\mathbf{b}) = \operatorname{span}(\mathbf{e}_1, \mathbf{e}_3)$ is a perfectly good subspace, and it is not $\mathcal{K}_k(\mathbf{A}, \mathbf{b})$ for any k, since those are $\operatorname{span}(\mathbf{e}_1)$, $\operatorname{span}(\mathbf{e}_1, \mathbf{e}_2)$, and $\mathbb{R}^3$. The clause that fails is that the powers be **consecutive**, and what it costs is everything: $\mathbf{A}\operatorname{span}(\mathbf{e}_1, \mathbf{e}_3) = \operatorname{span}(\mathbf{e}_2)$, which meets $\operatorname{span}(\mathbf{e}_1, \mathbf{e}_3)$ only at $\mathbf{0}$. The Krylov spaces, by contrast, are nested and almost invariant, which is the next proposition.

Why this definition. Chapter 10 already met the union of all of them. The cyclic subspace $Z(\mathbf{b}; \mathbf{A}) = \operatorname{span}(\mathbf{b}, \mathbf{A}\mathbf{b}, \mathbf{A}^2\mathbf{b}, \dots)$ of Definition 10.5.1 is the smallest $\mathbf{A}$-invariant subspace containing $\mathbf{b}$ (Proposition 10.5.1), and Theorem 10.5.2 says that the list stops growing exactly at the degree of the $\mathbf{A}$-annihilator $m_{\mathbf{A},\mathbf{b}}$. The Krylov space is that construction **truncated at** k, and the truncation is the whole point: $Z(\mathbf{b}; \mathbf{A})$ is a theoretical object, while $\mathcal{K}_k(\mathbf{A}, \mathbf{b})$ is what k units of work actually deliver.

Proposition 24.6.1. Krylov Spaces: Nesting, Dimension, Invariance

Let $\mathbf{A} \in M_n(F)$ and $\mathbf{b} \in F^n$ with $\mathbf{b} \neq \mathbf{0}$, and let $d = \deg m_{\mathbf{A},\mathbf{b}} \ge 1$ be the degree of the $\mathbf{A}$-annihilator of $\mathbf{b}$ (Definition 10.5.2). Then, for every $k \ge 1$:

(a) $\mathcal{K}_k(\mathbf{A}, \mathbf{b}) = \{\, p(\mathbf{A})\mathbf{b} : p \in F[x],\ \deg p \le k-1 \,\}$, and $\mathcal{K}_k \subseteq \mathcal{K}_{k+1}$ and $\mathbf{A}\,\mathcal{K}_k \subseteq \mathcal{K}_{k+1}$;

(b) $\dim \mathcal{K}_k(\mathbf{A}, \mathbf{b}) = \min(k, d)$;

(c) $\mathcal{K}_{k+1} = \mathcal{K}_k$ if and only if $k \ge d$, and in that case $\mathcal{K}_k = Z(\mathbf{b}; \mathbf{A})$ is $\mathbf{A}$-invariant.

Idea. Everything is Theorem 10.5.2 read one index at a time. That theorem says $(\mathbf{b}, \mathbf{A}\mathbf{b}, \dots, \mathbf{A}^{d-1}\mathbf{b})$ is a basis of $Z(\mathbf{b}; \mathbf{A})$; so the first d vectors of our list are independent, and from $\mathbf{A}^d\mathbf{b}$ on they add nothing.

Proof.

(a) A linear combination $\sum_{j<k} a_j \mathbf{A}^j\mathbf{b}$ is $p(\mathbf{A})\mathbf{b}$ for $p = \sum_{j<k} a_j x^j$, and conversely, which is the first statement. The inclusion $\mathcal{K}_k \subseteq \mathcal{K}_{k+1}$ holds because the spanning list of $\mathcal{K}_k$ is part of that of $\mathcal{K}_{k+1}$. And $\mathbf{A}\mathbf{A}^j\mathbf{b} = \mathbf{A}^{j+1}\mathbf{b} \in \mathcal{K}_{k+1}$ for $j \le k-1$, so $\mathbf{A}$ maps a spanning list of $\mathcal{K}_k$ into $\mathcal{K}_{k+1}$, hence maps $\mathcal{K}_k$ into $\mathcal{K}_{k+1}$.

(b) By Theorem 10.5.2 (a), the list $(\mathbf{b}, \mathbf{A}\mathbf{b}, \dots, \mathbf{A}^{d-1}\mathbf{b})$ is a basis of $Z(\mathbf{b}; \mathbf{A})$. If $k \le d$, the first k entries of that list are a sublist of a basis, hence independent, so $\dim \mathcal{K}_k = k$. If $k > d$, then $\mathcal{K}_k \subseteq Z(\mathbf{b}; \mathbf{A})$, because every $\mathbf{A}^j\mathbf{b}$ lies there (Proposition 10.5.1 (a)); and $\mathcal{K}_k \supseteq \mathcal{K}_d = Z(\mathbf{b}; \mathbf{A})$, since $\mathcal{K}_d$ already contains a basis of it. So $\mathcal{K}_k = Z(\mathbf{b}; \mathbf{A})$ and $\dim \mathcal{K}_k = d$. In both cases $\dim \mathcal{K}_k = \min(k, d)$.

(c) ($\Leftarrow$) If $k \ge d$, then (b) gives $\dim \mathcal{K}_{k+1} = d = \dim \mathcal{K}_k$, and with $\mathcal{K}_k \subseteq \mathcal{K}_{k+1}$ from (a) this forces equality; and the common space is $Z(\mathbf{b}; \mathbf{A})$, $\mathbf{A}$-invariant by Proposition 10.5.1 (b). ($\Rightarrow$) If $k < d$, then (b) gives $\dim \mathcal{K}_{k+1} = k+1 > k = \dim \mathcal{K}_k$, so the two differ.

This proves the proposition.

■

Warning. **The list $\mathbf{b}, \mathbf{Ab}, \dots, \mathbf{A}^{k-1}\mathbf{b}$ is a basis of $\mathcal{K}_k$ and is useless as one.** By Section 4's reading of the power method (Theorem 24.4.1), $\mathbf{A}^j\mathbf{b}$ turns towards the dominant eigenvector as j grows, so the later vectors of the list are nearly parallel to each other. Even for $\mathbf{A} = \operatorname{diag}(4,3,2,1)$ and $\mathbf{b} = (1,1,1,1)$, with its modest ratio $\lambda_1/\lambda_2 = 4/3$, the angle between consecutive vectors of the list falls from 24 degrees to 8.9 degrees across those four vectors, and the matrix with those four columns has 2-norm condition number about 1200 — for a 4×4 problem with one-digit data. Independent they may be; in the arithmetic of Definition 24.1.1 the coordinates of a vector in that basis are computed with errors magnified by that condition number, which grows quickly with k. **A Krylov method never uses this basis.** It builds an orthonormal one instead, and that is the next subsection.

Quick check. Let $\mathbf{A} \in M_n(F)$ be invertible with $\mathbf{A}^2 = \mathbf{I}_n$, and let $\mathbf{b} \neq \mathbf{0}$ be a vector that is not an eigenvector of $\mathbf{A}$. What is $\dim \mathcal{K}_k(\mathbf{A}, \mathbf{b})$ for each $k \geq 1$?

Solution. It is 1 for $k = 1$ and 2 for every $k \geq 2$. The polynomial $x^2 - 1$ kills $\mathbf{b}$, so $m_{\mathbf{A},\mathbf{b}}$ divides it and has degree at most 2 (Definition 10.5.2). Degree 1 would mean $m_{\mathbf{A},\mathbf{b}} = x - \lambda$, that is $\mathbf{Ab} = \lambda\mathbf{b}$, making $\mathbf{b}$ an eigenvector, which is excluded; and $\deg m_{\mathbf{A},\mathbf{b}} \geq 1$ because $\mathbf{b} \neq \mathbf{0}$. So $d = 2$, and Proposition 24.6.1 (b) gives $\min(k, 2)$.

24.6.2 Arnoldi: Gram–Schmidt with a purpose

We want an orthonormal basis $\mathbf{q}_1, \dots, \mathbf{q}_k$ of $\mathcal{K}_k$, with the nesting preserved: $\operatorname{span}(\mathbf{q}_1, \dots, \mathbf{q}_j) = \mathcal{K}_j$ for every $j \leq k$. Running Theorem 11.2.2 on the list $\mathbf{b}, \mathbf{Ab}, \dots, \mathbf{A}^{k-1}\mathbf{b}$ would do it, and would inherit exactly the defect the warning describes, since it needs those vectors first.

The repair is one line long, and it is the idea of the whole subsection. Do not orthogonalize $\mathbf{A}^j\mathbf{b}$. Orthogonalize $\mathbf{Aq}_j$. The vector $\mathbf{Aq}_j$ lies in $\mathcal{K}_{j+1}$ by Proposition 24.6.1 (a), and subtracting its components along $\mathbf{q}_1, \dots, \mathbf{q}_j$ leaves a vector in $\mathcal{K}_{j+1}$ orthogonal to $\mathcal{K}_j$ — which is what the next basis vector must be. The powers of $\mathbf{A}$ are never formed.

Algorithm. **The Arnoldi process.** Given $\mathbf{A} \in M_n(F)$ and $\mathbf{b} \neq \mathbf{0}$, set $\mathbf{q}_1 = \mathbf{b}/\|\mathbf{b}\|$. For $j = 1, 2, \dots$:

$$h_{ij} = \langle \mathbf{Aq}_j, \mathbf{q}_i \rangle \quad (1 \leq i \leq j), \qquad \mathbf{w}_j = \mathbf{Aq}_j - \sum_{i=1}^{j} h_{ij}\mathbf{q}_i,$$

$$h_{j+1,j} = \|\mathbf{w}_j\|, \qquad \mathbf{q}_{j+1} = \mathbf{w}_j / h_{j+1,j} \quad \text{if } h_{j+1,j} \neq 0.$$

If $h_{j+1,j} = 0$ the process stops; this is called a **breakdown**, and the theorem below says it is good news.

Write $\mathbf{Q}_j = (\mathbf{q}_1 \mid \cdots \mid \mathbf{q}_j) \in M_{n\times j}(F)$, let $\mathbf{H}_k \in M_k(F)$ have entries h_{ij} for $i \le j+1$ and zeros elsewhere, and let $\widetilde{\mathbf{H}}_k \in M_{(k+1)\times k}(F)$ be $\mathbf{H}_k$ with the extra row $h_{k+1,k}\mathbf{e}_k^\top$ appended at the bottom.

Theorem 24.6.2. The Arnoldi Process

Let $\mathbf{A} \in M_n(F)$, let $\mathbf{b} \in F^n$ be nonzero, let $d = \deg m_{\mathbf{A},\mathbf{b}}$, and run the Arnoldi process. Then the **Arnoldi relation**

$$\mathbf{A}\mathbf{Q}_k = \mathbf{Q}_k\mathbf{H}_k + h_{k+1,k}\,\mathbf{q}_{k+1}\mathbf{e}_k^\top \qquad (1 \le k \le d) \tag{24.6.1}$$

holds, the last term being $\mathbf{O}$ when $k = d$. Moreover:

(a) For every $j \le d$, the list $(\mathbf{q}_1, \dots, \mathbf{q}_j)$ is an orthonormal basis of $\mathcal{K}_j(\mathbf{A}, \mathbf{b})$; moreover $h_{j+1,j} > 0$ for $j < d$, and $h_{d+1,d} = 0$, so the process breaks down at step d and at no earlier step.

(b) For $1 \le k \le d$ we have $\mathbf{Q}_k^*\mathbf{Q}_k = \mathbf{I}_k$, and $\mathbf{H}_k = \mathbf{Q}_k^*\mathbf{A}\mathbf{Q}_k$ is **upper Hessenberg**. For $k < d$, Eq. 24.6.1 can also be written $\mathbf{A}\mathbf{Q}_k = \mathbf{Q}_{k+1}\widetilde{\mathbf{H}}_k$.

(c) A breakdown means an invariant subspace: if $h_{k+1,k} = 0$, then $\mathbf{A}\mathcal{K}_k \subseteq \mathcal{K}_k$, and every eigenvalue of $\mathbf{H}_k$ is an eigenvalue of $\mathbf{A}$.

Idea. Only one thing has to be checked, and it is the nesting: if $\mathbf{q}_1, \dots, \mathbf{q}_j$ is an orthonormal basis of $\mathcal{K}_j$, then $\mathbf{A}\mathbf{q}_j$ lies in $\mathcal{K}_{j+1}$ but generally not in $\mathcal{K}_j$, and what the step subtracts is exactly its $\mathcal{K}_j$-component. The matrix identity is then just the definition of $\mathbf{w}_j$ rearranged, column by column, and the Hessenberg shape is the statement that $\mathbf{A}\mathbf{q}_j$ needs only $\mathbf{q}_1, \dots, \mathbf{q}_{j+1}$ — never $\mathbf{q}_{j+2}$ — because $\mathbf{A}\mathcal{K}_j \subseteq \mathcal{K}_{j+1}$.

Proof.

(a) Induction on j. For $j = 1$, $\mathbf{q}_1 = \mathbf{b}/\|\mathbf{b}\|$ is a unit vector spanning $\mathcal{K}_1 = \operatorname{span}(\mathbf{b})$. Let $1 \le j$ and suppose $(\mathbf{q}_1, \dots, \mathbf{q}_j)$ is an orthonormal basis of $\mathcal{K}_j$. The subtraction in the process makes $\mathbf{w}_j$ orthogonal to each $\mathbf{q}_i$ with $i \le j$: for such i,

$$\langle \mathbf{w}_j, \mathbf{q}_i \rangle = \langle \mathbf{A}\mathbf{q}_j, \mathbf{q}_i \rangle - \sum_{l=1}^{j} h_{lj}\langle \mathbf{q}_l, \mathbf{q}_i \rangle = h_{ij} - h_{ij} = 0,$$

using orthonormality. Also $\mathbf{A}\mathbf{q}_j \in \mathbf{A}\mathcal{K}_j \subseteq \mathcal{K}_{j+1}$ by Proposition 24.6.1 (a), and each $\mathbf{q}_i \in \mathcal{K}_j \subseteq \mathcal{K}_{j+1}$, so $\mathbf{w}_j \in \mathcal{K}_{j+1}$.

Claim. $\mathcal{K}_{j+1} = \mathcal{K}_j + \operatorname{span}(\mathbf{A}\mathbf{q}_j)$.

Proof. ($\supseteq$) Both summands lie in $\mathcal{K}_{j+1}$, by Proposition 24.6.1 (a). ($\subseteq$) Since $\mathcal{K}_{j+1} = \mathcal{K}_j + \operatorname{span}(\mathbf{A}^j\mathbf{b})$, it is enough to place $\mathbf{A}^j\mathbf{b}$ on the right. Write $\mathbf{A}^{j-1}\mathbf{b} = \sum_{i\le j} c_i\mathbf{q}_i$, possible because $\mathbf{A}^{j-1}\mathbf{b} \in \mathcal{K}_j$ and the $\mathbf{q}_i$ are a basis of it. Applying $\mathbf{A}$,

$$\mathbf{A}^j\mathbf{b} = \sum_{i\le j} c_i\,\mathbf{A}\mathbf{q}_i,$$

and for $i \le j-1$ we have $\mathbf{A}\mathbf{q}_i \in \mathbf{A}\mathcal{K}_i \subseteq \mathcal{K}_{i+1} \subseteq \mathcal{K}_j$. So only the term $c_j\mathbf{A}\mathbf{q}_j$ leaves $\mathcal{K}_j$.

■

Now $\mathbf{w}_j = \mathbf{0}$ if and only if $\mathbf{A}\mathbf{q}_j \in \mathcal{K}_j$, and by the Claim that happens if and only if $\mathcal{K}_{j+1} = \mathcal{K}_j$. Suppose first that $j < d$. Then $\dim \mathcal{K}_{j+1} = j+1 > j = \dim \mathcal{K}_j$ by Proposition 24.6.1 (b), so $\mathcal{K}_{j+1} \neq \mathcal{K}_j$ and $\mathbf{w}_j \neq \mathbf{0}$. Hence $h_{j+1,j} > 0$, and $\mathbf{q}_{j+1} = \mathbf{w}_j / h_{j+1,j}$ is a unit vector in $\mathcal{K}_{j+1}$ orthogonal to $\mathbf{q}_1, \dots, \mathbf{q}_j$. An orthogonal list of nonzero vectors is independent, so $(\mathbf{q}_1, \dots, \mathbf{q}_{j+1})$ is an orthonormal list of $j+1$ vectors in the $(j+1)$-dimensional space $\mathcal{K}_{j+1}$, hence a basis of it.

If instead $j = d$, then $\mathcal{K}_{d+1} = \mathcal{K}_d$ by Proposition 24.6.1 (c), so $\mathbf{A}\mathbf{q}_d \in \mathbf{A}\mathcal{K}_d \subseteq \mathcal{K}_{d+1} = \mathcal{K}_d$ and $\mathbf{w}_d = \mathbf{0}$. This gives (a).

(b), and Eq. 24.6.1. The columns of $\mathbf{Q}_k$ are orthonormal by (a), which is $\mathbf{Q}_k^* \mathbf{Q}_k = \mathbf{I}_k$. Rearranging the definition of $\mathbf{w}_j$,

$$\mathbf{A}\mathbf{q}_j = \sum_{i=1}^{j} h_{ij}\mathbf{q}_i + h_{j+1,j}\mathbf{q}_{j+1} = \sum_{i=1}^{j+1} h_{ij}\mathbf{q}_i,$$

for each $j \le k$, where for $j = d$ the last term is $\mathbf{0}$. For $j < k$ the right-hand side is $\mathbf{Q}_k \mathbf{H}_k \mathbf{e}_j$, since column j of $\mathbf{H}_k$ has entries $h_{1j}, \dots, h_{j+1,j}$ and zeros below and $j+1 \le k$; for $j = k$ it is $\mathbf{Q}_k \mathbf{H}_k \mathbf{e}_k + h_{k+1,k}\mathbf{q}_{k+1}$. Collecting the columns gives Eq. 24.6.1, and the form $\mathbf{A}\mathbf{Q}_k = \mathbf{Q}_{k+1}\widetilde{\mathbf{H}}_k$ is the same identity with the extra column of $\mathbf{Q}_{k+1}$ and the extra row of $\widetilde{\mathbf{H}}_k$ absorbing the last term. The entries h_{ij} with $i > j+1$ are zero by definition, which is exactly upper Hessenberg form. Finally, multiplying Eq. 24.6.1 on the left by $\mathbf{Q}_k^*$ and using $\mathbf{Q}_k^* \mathbf{Q}_k = \mathbf{I}_k$ and $\mathbf{Q}_k^* \mathbf{q}_{k+1} = \mathbf{0}$,

$$\mathbf{Q}_k^* \mathbf{A} \mathbf{Q}_k = \mathbf{H}_k + h_{k+1,k}\, \mathbf{Q}_k^* \mathbf{q}_{k+1} \mathbf{e}_k^{\top} = \mathbf{H}_k.$$

(c) If $h_{k+1,k} = 0$, then Eq. 24.6.1 reads $\mathbf{A}\mathbf{Q}_k = \mathbf{Q}_k \mathbf{H}_k$, so $\mathbf{A}\mathbf{q}_j \in \mathcal{K}_k$ for every $j \le k$ and hence $\mathbf{A}\mathcal{K}_k \subseteq \mathcal{K}_k$. If $\mathbf{H}_k \mathbf{y} = \theta \mathbf{y}$ with $\mathbf{y} \neq \mathbf{0}$, put $\mathbf{x} = \mathbf{Q}_k \mathbf{y}$, which is nonzero because the columns of $\mathbf{Q}_k$ are independent; then $\mathbf{A}\mathbf{x} = \mathbf{A}\mathbf{Q}_k\mathbf{y} = \mathbf{Q}_k\mathbf{H}_k\mathbf{y} = \theta \mathbf{Q}_k \mathbf{y} = \theta \mathbf{x}$. This proves the theorem.

■

The cost of k steps is k matrix-vector products plus the orthogonalizations, which use about $4nj$ flops at step j — j inner products and j subtractions, about $2n$ flops each — so $O(nk^2)$ in total; and all $k+1$ vectors must be stored. Both grow with k, which is why Arnoldi is restarted in practice. For Hermitian $\mathbf{A}$ neither grows, and that is the next theorem.

24.6.3 Lanczos: the recurrence collapses to three terms

Theorem 24.6.3. The Lanczos Recurrence

Let $\mathbf{A} \in M_n(\mathbb{C})$ be **Hermitian**, let $\mathbf{b} \neq \mathbf{0}$, and run the Arnoldi process. Then $\mathbf{H}_k$ is Hermitian for every $k \le d$, hence **tridiagonal**, with real diagonal entries; writing $\alpha_j = h_{jj} \in \mathbb{R}$ and $\beta_j = h_{j+1,j}$, which is > 0 for $j < d$ and 0 for $j = d$, we have $h_{j,j+1} = \beta_j$ and

$$\mathbf{A}\mathbf{q}_j = \beta_{j-1}\mathbf{q}_{j-1} + \alpha_j \mathbf{q}_j + \beta_j \mathbf{q}_{j+1} \qquad (1 \le j \le d), \tag{24.6.2}$$

with $\beta_0\mathbf{q}_0$ read as $\mathbf{0}$ and $\beta_d\mathbf{q}_{d+1}$ as $\mathbf{0}$. Consequently $\mathbf{H}_k = \mathbf{T}_k$ is the real symmetric tridiagonal matrix with diagonal $(\alpha_1, \dots, \alpha_k)$ and off-diagonal $(\beta_1, \dots, \beta_{k-1})$, and each step of the process needs only $\mathbf{q}_{j-1}$ and $\mathbf{q}_j$.

Idea. Two facts collide. The matrix $\mathbf{H}_k = \mathbf{Q}_k^*\mathbf{A}\mathbf{Q}_k$ is Hermitian because $\mathbf{A}$ is, and it is upper Hessenberg by Theorem 24.6.2. A Hermitian matrix with zeros below the subdiagonal has zeros above the superdiagonal too, by conjugating. Everything else is bookkeeping. The reason is worth naming separately: $h_{ij} = \langle \mathbf{A}\mathbf{q}_j, \mathbf{q}_i \rangle$ lets $\mathbf{A}$ move to the other slot for free, and $\mathbf{A}\mathbf{q}_i$ lives one index further along than $\mathbf{q}_i$ does.

Proof. By Theorem 24.6.2 (b), $\mathbf{H}_k = \mathbf{Q}_k^*\mathbf{A}\mathbf{Q}_k$, so $\mathbf{H}_k^* = \mathbf{Q}_k^*\mathbf{A}^*\mathbf{Q}_k = \mathbf{Q}_k^*\mathbf{A}\mathbf{Q}_k = \mathbf{H}_k$ and $\mathbf{H}_k$ is Hermitian; in particular its diagonal entries are real.

Let $i \le j-2$. Since $\mathbf{A}^* = \mathbf{A}$ is self-adjoint,

$$h_{ij} = \langle \mathbf{A}\mathbf{q}_j, \mathbf{q}_i \rangle = \langle \mathbf{q}_j, \mathbf{A}\mathbf{q}_i \rangle.$$

Now $\mathbf{q}_i \in \mathcal{K}_i$, so $\mathbf{A}\mathbf{q}_i \in \mathcal{K}_{i+1} \subseteq \mathcal{K}_{j-1}$ by Proposition 24.6.1 (a) and $i+1 \le j-1$, while $\mathbf{q}_j$ is orthogonal to $\mathcal{K}_{j-1}$ by Theorem 24.6.2 (a). Hence $h_{ij} = 0$ whenever $i \le j-2$. Together with $h_{ij} = 0$ for $i > j+1$, this leaves only $i \in \{j-1, j, j+1\}$, so $\mathbf{H}_k$ is tridiagonal. Being Hermitian, $h_{j,j+1} = \overline{h_{j+1,j}} = \beta_j$, which is real and positive. The displayed recurrence Eq. 24.6.2 is then the expansion $\mathbf{A}\mathbf{q}_j = \sum_i h_{ij}\mathbf{q}_i$ of Theorem 24.6.2 (b) with only three surviving terms; and $\mathbf{H}_k$ has real entries, is symmetric and is tridiagonal, as claimed. Since the step needs only $h_{j-1,j} = \beta_{j-1}$, $h_{jj} = \alpha_j$, and $\mathbf{q}_{j-1}, \mathbf{q}_j$, this proves the theorem. ∎

This is Chapter 11's three-term recurrence, and for the same reason. Chapter 11 §10 built the monic orthogonal polynomials for an inner product on $\mathbb{R}[x]$ and found (Theorem 11.10.3) that $p_{k+1} = (x - a_k)p_k - b_k p_{k-1}$, with only three terms; the reason given there was Proposition 11.10.1, "multiplication by x moves from one slot to the other for free". The two statements are the same statement. Given Hermitian $\mathbf{A}$ and $\mathbf{b} \neq \mathbf{0}$, define on $\mathbb{R}[x]$

$$\langle p, q \rangle_{\mathbf{b}} := \langle p(\mathbf{A})\mathbf{b}, q(\mathbf{A})\mathbf{b} \rangle.$$

This is symmetric and bilinear, and on $\mathbb{R}[x]_{\le d-1}$ it is positive definite: $\langle p, p \rangle_{\mathbf{b}} = \|p(\mathbf{A})\mathbf{b}\|^2$ vanishes only when $p(\mathbf{A})\mathbf{b} = \mathbf{0}$, which by Definition 10.5.2 means $m_{\mathbf{A},\mathbf{b}} \mid p$, impossible for a nonzero p of degree below d. Multiplication by x is symmetric for it, since $\mathbf{A}$ is Hermitian:

$$\langle xp, q \rangle_{\mathbf{b}} = \langle \mathbf{A}\,p(\mathbf{A})\mathbf{b}, q(\mathbf{A})\mathbf{b} \rangle = \langle p(\mathbf{A})\mathbf{b}, \mathbf{A}\,q(\mathbf{A})\mathbf{b} \rangle = \langle p, xq \rangle_{\mathbf{b}}.$$

Now Proposition 24.6.1 (a) says $\mathbf{q}_{j+1}$ is $r(\mathbf{A})\mathbf{b}$ for some r of degree at most j, and orthogonality to $\mathcal{K}_j$ says $\langle r, s \rangle_{\mathbf{b}} = 0$ for every s of degree below j. So for $j \le d-1$ the polynomial r is a scalar multiple of the j-th monic orthogonal polynomial p_j for $\langle \cdot, \cdot \rangle_{\mathbf{b}}$, and Eq. 24.6.2, at those indices, **is** the recurrence of Theorem 11.10.3 rewritten in terms of the vectors $p_j(\mathbf{A})\mathbf{b}$ — read, as that theorem's node inner product is read, only in the range of degrees where the form is positive definite. The last step $j = d$ lies outside that range: there the monic polynomial of degree d orthogonal to everything of lower degree is $m_{\mathbf{A},\mathbf{b}}$ itself, and $\langle m_{\mathbf{A},\mathbf{b}}, m_{\mathbf{A},\mathbf{b}} \rangle_{\mathbf{b}} = 0$. A third appearance of the same three terms is Chapter 7's: by Theorem 7.8.3, the characteristic polynomials $D_k(x) = \det(x\mathbf{I}_k - \mathbf{T}_k)$ of the leading blocks satisfy

$$D_k = (x - \alpha_k)D_{k-1} - \beta_{k-1}^2\, D_{k-2},$$

which is Theorem 11.10.3 with $a_{k-1} = \alpha_k$ and $b_{k-1} = \beta_{k-1}^2 > 0$. The Lanczos coefficients, the orthogonal polynomials, and the determinants of the tridiagonal blocks are one object seen three ways.

Example. A breakdown at step two

Let $\mathbf{A} = \begin{pmatrix} 3 & 1 & 1 \\ 1 & 3 & 1 \\ 1 & 1 & 3 \end{pmatrix}$ and $\mathbf{b} = \mathbf{e}_1$. Run the Lanczos recurrence of Theorem 24.6.3, and identify what it finds.

Solution. $\mathbf{q}_1 = \mathbf{e}_1$, and $\mathbf{A}\mathbf{q}_1 = (3,1,1)$. So $\alpha_1 = \langle \mathbf{A}\mathbf{q}_1, \mathbf{q}_1 \rangle = 3$ and

$$\mathbf{w}_1 = (3,1,1) - 3(1,0,0) = (0,1,1), \qquad \beta_1 = \sqrt{2}, \qquad \mathbf{q}_2 = \tfrac{1}{\sqrt{2}}(0,1,1).$$

Next, $\mathbf{A}\mathbf{q}_2 = \frac{1}{\sqrt{2}}(2,4,4)$, so $\alpha_2 = \langle \mathbf{A}\mathbf{q}_2, \mathbf{q}_2 \rangle = \frac{1}{2}(0+4+4) = 4$ and

$$\mathbf{w}_2 = \tfrac{1}{\sqrt{2}}(2,4,4) - 4 \cdot \tfrac{1}{\sqrt{2}}(0,1,1) - \sqrt{2}\,(1,0,0) = \tfrac{1}{\sqrt{2}}(2,0,0) - \sqrt{2}(1,0,0) = \mathbf{0}.$$

The process breaks down at step 2, so $d = 2$ and $\mathcal{K}_2$ is $\mathbf{A}$-invariant, by Theorem 24.6.2 (a), (c). Then

$$\mathbf{T}_2 = \begin{pmatrix} 3 & \sqrt{2} \\ \sqrt{2} & 4 \end{pmatrix}, \qquad p_{\mathbf{T}_2}(x) = (x-3)(x-4) - 2 = x^2 - 7x + 10,$$

with roots 5 and 2. And indeed $\operatorname{spec}(\mathbf{A}) = \{5, 2\}$, with 5 simple and 2 of multiplicity 2: the two steps have produced both eigenvalues exactly. This is Theorem 24.6.2 (c) in action, and it is also Proposition 24.6.1 (b) with $d = \deg m_{\mathbf{A},\mathbf{e}_1} = 2$, since $m_{\mathbf{A}} = (x-5)(x-2)$ and neither factor alone kills $\mathbf{e}_1$.

24.6.4 Ritz values

An invariant subspace would give eigenvalues exactly. A Krylov space is an **almost** invariant subspace — by Eq. 24.6.1 it fails to be invariant only through the single vector $h_{k+1,k}\mathbf{q}_{k+1}$ — so it should give almost-eigenvalues, and it should be possible to say how almost.

Definition 24.6.2. Ritz Values and Ritz Vectors

Let $\mathbf{A} \in M_n(F)$, let $\mathbf{Q}_k \in M_{n\times k}(F)$ have orthonormal columns, and put $\mathbf{H}_k = \mathbf{Q}_k^* \mathbf{A} \mathbf{Q}_k$. The eigenvalues $\theta_1, \dots, \theta_k$ of $\mathbf{H}_k$ are the **Ritz values** of $\mathbf{A}$ from the subspace $\operatorname{span}(\mathbf{q}_1, \dots, \mathbf{q}_k)$, and for an eigenvector $\mathbf{y}$ of $\mathbf{H}_k$ with $\mathbf{H}_k\mathbf{y} = \theta\mathbf{y}$ and $\|\mathbf{y}\| = 1$, the vector $\mathbf{x} = \mathbf{Q}_k\mathbf{y}$ is the corresponding **Ritz vector**; the pair $(\theta, \mathbf{x})$ is a **Ritz pair**.

In words: project the problem onto the subspace, solve the small problem exactly, and lift the answer back. The Ritz values depend on the subspace and not on which orthonormal basis of it we chose: replacing $\mathbf{Q}_k$ by $\mathbf{Q}_k\mathbf{V}$ with $\mathbf{V}$ unitary replaces $\mathbf{H}_k$ by $\mathbf{V}^*\mathbf{H}_k\mathbf{V}$, which has the same eigenvalues. Note that $\|\mathbf{x}\| = \|\mathbf{Q}_k\mathbf{y}\| = \|\mathbf{y}\| = 1$, since $\mathbf{Q}_k$ has orthonormal columns.

Theorem 24.6.4. Ritz Values Interlace, and Their Residual Is One Number

Let $\mathbf{A} \in M_n(\mathbb{C})$ be Hermitian with eigenvalues $\lambda_1(\mathbf{A}) \geq \cdots \geq \lambda_n(\mathbf{A})$, let $1 \leq k \leq d$, and let $\theta_1 \geq \cdots \geq \theta_k$ be the Ritz values from $\mathcal{K}_k(\mathbf{A}, \mathbf{b})$, that is, the eigenvalues of $\mathbf{T}_k$.

(a) $\lambda_i(\mathbf{A}) \;\geq\; \theta_i \;\geq\; \lambda_{i+n-k}(\mathbf{A})$ for $1 \leq i \leq k$.

(b) If $\mathbf{T}_k\mathbf{y} = \theta\mathbf{y}$ with $\|\mathbf{y}\| = 1$ and $\mathbf{x} = \mathbf{Q}_k\mathbf{y}$, then

$$\mathbf{A}\mathbf{x} - \theta\mathbf{x} = \beta_k\, y_k\, \mathbf{q}_{k+1}, \qquad \text{so} \qquad \|\mathbf{A}\mathbf{x} - \theta\mathbf{x}\| = \beta_k\, |y_k|,$$

where y_k is the last coordinate of $\mathbf{y}$ and $\beta_k = h_{k+1,k}$, with $\beta_d\mathbf{q}_{d+1}$ read as $\mathbf{0}$.

Idea. For (a), the obstacle is that $\mathbf{T}_k = \mathbf{Q}_k^*\mathbf{A}\mathbf{Q}_k$ is a compression, not a principal submatrix, and Chapter 17's interlacing is about principal submatrices. The repair is to complete $\mathbf{Q}_k$ to a unitary matrix: then $\mathbf{T}_k$ *is* the leading principal submatrix of a matrix unitarily similar to $\mathbf{A}$, and Corollary 17.4.3 applies unchanged. For (b), just apply Eq. 24.6.1 to $\mathbf{y}$ and watch the two $\mathbf{Q}_k\mathbf{T}_k\mathbf{y}$ terms cancel.

Proof.

(a) Extend $\mathbf{q}_1, \ldots, \mathbf{q}_k$ to an orthonormal basis $\mathbf{q}_1, \ldots, \mathbf{q}_n$ of $\mathbb{C}^n$, which is possible by Theorem 11.2.2 applied to any extension of the list to a basis (Theorem 1.6.5), and let $\mathbf{U} = (\mathbf{q}_1 \mid \cdots \mid \mathbf{q}_n)$, which is unitary. Then $\mathbf{B} = \mathbf{U}^*\mathbf{A}\mathbf{U}$ is Hermitian and unitarily similar to $\mathbf{A}$, so $\lambda_i(\mathbf{B}) = \lambda_i(\mathbf{A})$ for every i. Its entries are $b_{ij} = \langle \mathbf{A}\mathbf{q}_j, \mathbf{q}_i \rangle$, and for $i, j \leq k$ these are exactly the entries of $\mathbf{Q}_k^*\mathbf{A}\mathbf{Q}_k = \mathbf{T}_k$. So $\mathbf{T}_k = \mathbf{B}_{I,I}$ with $I = \{1, \ldots, k\}$, a principal submatrix of $\mathbf{B}$ obtained by deleting $m = n - k$ rows and the same columns. Now Corollary 17.4.3 gives

$$\lambda_i(\mathbf{B}) \;\geq\; \lambda_i(\mathbf{T}_k) \;\geq\; \lambda_{i+m}(\mathbf{B}) \qquad (1 \leq i \leq k),$$

which is the assertion, since $\lambda_i(\mathbf{T}_k) = \theta_i$ and $\lambda_i(\mathbf{B}) = \lambda_i(\mathbf{A})$.

(b) By Eq. 24.6.1, $\mathbf{A}\mathbf{Q}_k = \mathbf{Q}_k\mathbf{T}_k + \beta_k\mathbf{q}_{k+1}\mathbf{e}_k^{\top}$. Applying both sides to $\mathbf{y}$ and using $\mathbf{T}_k\mathbf{y} = \theta\mathbf{y}$ and $\mathbf{e}_k^{\top}\mathbf{y} = y_k$,

$$\mathbf{A}\mathbf{x} = \mathbf{A}\mathbf{Q}_k\mathbf{y} = \mathbf{Q}_k\mathbf{T}_k\mathbf{y} + \beta_k\, y_k\, \mathbf{q}_{k+1} = \theta\, \mathbf{Q}_k\mathbf{y} + \beta_k y_k \mathbf{q}_{k+1} = \theta\mathbf{x} + \beta_k y_k \mathbf{q}_{k+1}.$$

Subtracting $\theta\mathbf{x}$ gives the displayed identity, and taking norms gives the second form, since $\|\mathbf{q}_{k+1}\| = 1$ and $\beta_k \geq 0$. (For $k = d$ the right-hand side is $\mathbf{0}$, and the Ritz pair is an exact eigenpair, as Theorem 24.6.2 (c) already said.) This proves the theorem.

■

Part (b) is where the chapter's perturbation theory pays off, and it pays off for free. The residual $\|\mathbf{A}\mathbf{x} - \theta\mathbf{x}\|$ is the one quantity Chapter 20 §06 asks for, and here it is available **without forming $\mathbf{A}\mathbf{x}$**: it is the product of the number β_k the process has just computed and one coordinate of the small eigenvector $\mathbf{y}$. So Theorem 20.6.1 applies verbatim, and says that $\mathbf{A}$ has an eigenvalue within $\beta_k|y_k|$ of θ. Better still, $\theta = \langle \mathbf{A}\mathbf{x}, \mathbf{x} \rangle$ is the Rayleigh quotient of $\mathbf{x}$ — because $\langle \mathbf{A}\mathbf{x}, \mathbf{x} \rangle = \mathbf{y}^*\mathbf{Q}_k^*\mathbf{A}\mathbf{Q}_k\mathbf{y} = \mathbf{y}^*\mathbf{T}_k\mathbf{y} = \theta$ — so Theorem 20.6.2 applies too, and turns the same number into a bound of size $(\beta_k|y_k|)^2$ divided by the distance to the rest of the spectrum.

Nothing in this section had to be proved twice.

Warning. **A Ritz value that happens to equal an eigenvalue is not a converged one, and for a non-Hermitian $\mathbf{A}$ there is no interlacing to fall back on.** Take

$$\mathbf{A} = \begin{pmatrix} 2 & 1 & 1 \\ 0 & 2 & 1 \\ 0 & 0 & 1 \end{pmatrix}, \qquad \mathbf{b} = \mathbf{e}_3 .$$

Two Arnoldi steps give $\mathbf{q}_1 = \mathbf{e}_3$, $\mathbf{q}_2 = \tfrac{1}{\sqrt{2}}(1,1,0)$, and $\mathbf{H}_2 = \begin{pmatrix} 1 & 0 \\ \sqrt{2} & \tfrac{5}{2} \end{pmatrix}$, hence the Ritz values $\tfrac{5}{2}$ and 1; Exercise (B2: Two Arnoldi steps) works the details. The second is an exact eigenvalue of $\mathbf{A}$ — and its Ritz vector $\tfrac{1}{\sqrt{17}}(-2,-2,3)$ is not an eigenvector of $\mathbf{A}$, the residual being $\sqrt{34}/17 \approx 0.343$, nowhere near zero. What certifies convergence is the residual, never the agreement of a number with a number. Part (a) of Theorem 24.6.4 needs $\mathbf{A}$ Hermitian; without that hypothesis there is no interlacing and, by Theorem 20.4.4, a residual bound carries the factor $\kappa_2(\mathbf{X})$.

Example. Ritz values closing in on the extremes

Let $\mathbf{A} = \operatorname{diag}(4,3,2,1)$ and $\mathbf{b} = (1,1,1,1)$. Compute the Lanczos coefficients, the Ritz values for $k = 1,2,3$, and the residual of the largest Ritz pair at $k = 3$.

Solution. Here $\mathbf{q}_1 = \tfrac{1}{2}(1,1,1,1)$ and $\mathbf{A}\mathbf{q}_1 = \tfrac{1}{2}(4,3,2,1)$, so $\alpha_1 = \tfrac{1}{4}(4+3+2+1) = \tfrac{5}{2}$. Then

$$\mathbf{w}_1 = \tfrac{1}{2}(4,3,2,1) - \tfrac{5}{2}\cdot\tfrac{1}{2}(1,1,1,1) = \tfrac{1}{4}(3,1,-1,-3), \qquad \beta_1 = \frac{\sqrt{20}}{4} = \frac{\sqrt{5}}{2},$$

and $\mathbf{q}_2 = \tfrac{1}{\sqrt{20}}(3,1,-1,-3)$. Continuing in the same way gives $\alpha_2 = \alpha_3 = \alpha_4 = \tfrac{5}{2}$ as well, together with $\beta_2 = \tfrac{2}{\sqrt{5}}$ and $\beta_3 = \tfrac{3}{2\sqrt{5}}$.

The repetition of $\tfrac{5}{2}$ is a symmetry, not an accident. Let $\mathbf{S}$ be the reversal permutation matrix, $\mathbf{S}\mathbf{e}_j = \mathbf{e}_{5-j}$, and put $\mathbf{B} = \mathbf{A} - \tfrac{5}{2}\mathbf{I} = \operatorname{diag}(\tfrac{3}{2}, \tfrac{1}{2}, -\tfrac{1}{2}, -\tfrac{3}{2})$. Then $\mathbf{S}$ is orthogonal with $\mathbf{S}^2 = \mathbf{I}$, and $\mathbf{S}\mathbf{B}\mathbf{S} = -\mathbf{B}$, $\mathbf{S}\mathbf{b} = \mathbf{b}$. We show by induction, simultaneously, that $\mathbf{S}\mathbf{q}_j = \varepsilon_j \mathbf{q}_j$ with $\varepsilon_j = (-1)^{j-1}$ and that $\alpha_j = \tfrac{5}{2}$. The first holds at $j = 1$. Given it at j,

$$\alpha_j - \tfrac{5}{2} = \langle \mathbf{B}\mathbf{q}_j, \mathbf{q}_j \rangle = \langle \mathbf{S}\mathbf{B}\mathbf{q}_j, \mathbf{S}\mathbf{q}_j \rangle = \langle \mathbf{S}\mathbf{B}\mathbf{S}(\mathbf{S}\mathbf{q}_j), \mathbf{S}\mathbf{q}_j \rangle = -\varepsilon_j^2 \langle \mathbf{B}\mathbf{q}_j, \mathbf{q}_j \rangle,$$

and $\varepsilon_j^2 = 1$, so this number equals its own negative and $\alpha_j = \tfrac{5}{2}$. The Lanczos step then reads $\mathbf{w}_j = \mathbf{B}\mathbf{q}_j - \beta_{j-1}\mathbf{q}_{j-1}$, and applying $\mathbf{S}$ with $\varepsilon_{j-1} = -\varepsilon_j$ gives $\mathbf{S}\mathbf{w}_j = -\varepsilon_j \mathbf{w}_j$, so $\mathbf{S}\mathbf{q}_{j+1} = \varepsilon_{j+1}\mathbf{q}_{j+1}$. The Ritz values are the eigenvalues of the leading blocks of $\mathbf{T}_4$:

k	Ritz values	check against $4,3,2,1$
1	$\tfrac{5}{2}$	$4 \ge 2.5 \ge 1$
2	$\tfrac{5 \pm \sqrt{5}}{2} \approx 3.6180,\ 1.3820$	$4 \ge 3.618 \ge 2$; $3 \ge 1.382 \ge 1$
3	$\tfrac{5}{2} \pm \tfrac{\sqrt{205}}{10} \approx 3.9318,\ 1.0682,$ and $\tfrac{5}{2}$	$4 \ge 3.932 \ge 3$; $3 \ge 2.5 \ge 2$; $2 \ge 1.068 \ge 1$
4	$4,3,2,1$	exact

Every row obeys Theorem 24.6.4 (a), with $m = n - k$ shifting the index on the right. The extreme Ritz values close in on 4 and 1 fast, and the interior ones lag: at $k = 3$ the outer two are within 0.07 of 4 and 1, while the single interior value $\frac{5}{2}$ is half a unit from each of 3 and 2.

The residual at $k = 3$. The eigenvector of $\mathbf{T}_3$ for $\theta_1 = \frac{5}{2} + \frac{\sqrt{205}}{10}$ is, normalized,

$$\mathbf{y} = \left(\tfrac{5}{\sqrt{82}},\ \tfrac{1}{\sqrt{2}},\ \tfrac{4}{\sqrt{82}}\right),$$

so Theorem 24.6.4 (b) gives $\|\mathbf{A}\mathbf{x} - \theta_1\mathbf{x}\| = \beta_3|y_3| = \frac{3}{2\sqrt{5}} \cdot \frac{4}{\sqrt{82}} = \frac{3\sqrt{410}}{205} \approx 0.2963$. By Theorem 20.6.1, $\mathbf{A}$ has an eigenvalue within 0.2963 of 3.9318; the truth is 4, at distance 0.0682, so the certificate costs a factor of about 4. Suppose we also know that $\mathbf{A}$ has no eigenvalue in the open interval $(3, 4)$. Then Theorem 20.6.2 (b) with $\alpha = 3$ improves the upper end to

$$\theta_1 + \frac{(0.2963)^2}{3.9318 - 3} = 3.9318 + 0.0942 = 4.0260,$$

and the lower end is θ_1 itself by interlacing. The bracket $[3.9318,\ 4.0260]$ has width 0.094 where the residual bound gave 0.593. That is the quadratic gain of Chapter 20 §06, obtained here from one number the process had already produced.

24.6.5 How fast the extreme Ritz values converge

The example suggests the extremes converge quickly. Here is why, and how much of it we can prove.

Theorem 24.6.5. A Rate for the Largest Ritz Value

Let $\mathbf{A} \in M_n(\mathbb{C})$ be Hermitian with eigenvalues $\lambda_1 > \lambda_2 \geq \cdots \geq \lambda_n$ and an orthonormal eigenbasis $\mathbf{v}_1, \ldots, \mathbf{v}_n$. Let $\mathbf{b}$ be a unit vector with $\mathbf{b} = \sum_i c_i \mathbf{v}_i$ and $c_1 \neq 0$, and put

$$\tau^2 = \frac{\sum_{i \geq 2} |c_i|^2}{|c_1|^2}, \qquad \rho = \frac{\lambda_2 - \lambda_n}{2\lambda_1 - \lambda_2 - \lambda_n} \in [0, 1).$$

Then the largest Ritz value $\theta_1^{(k)}$ from $\mathcal{K}_k(\mathbf{A}, \mathbf{b})$ satisfies, for $1 \leq k \leq d$,

$$0 \ \leq\ \lambda_1 - \theta_1^{(k)} \ \leq\ (\lambda_1 - \lambda_n)\, \tau^2\, \rho^{2(k-1)},$$

and $\theta_1^{(k)}$ is nondecreasing in k.

Idea. Three moves. ① The largest Ritz value is the largest Rayleigh quotient **over the Krylov space**, by the Rayleigh principle transported through $\mathbf{Q}_k$. ② Every vector of the Krylov space is $p(\mathbf{A})\mathbf{b}$ for a polynomial of degree below k, so we may choose p and get a bound for free: whichever p makes p large at λ_1 and small at the other eigenvalues gives a good one. ③ Take the simplest such p, the $(k-1)$-st power of x minus the midpoint of $[\lambda_n, \lambda_2]$. That is not the best choice, and the best one is the subject of the next section.

Proof. **Step 1.** $\theta_1^{(k)} = \max\{R_{\mathbf{A}}(\mathbf{x}) : \mathbf{x} \in \mathcal{K}_k(\mathbf{A},\mathbf{b}),\ \mathbf{x} \neq \mathbf{0}\}$. Indeed $\theta_1^{(k)} = \lambda_1(\mathbf{T}_k) = \max_{\mathbf{y}\neq\mathbf{0}} R_{\mathbf{T}_k}(\mathbf{y})$ by Proposition 17.1.1, and the map $\mathbf{y} \mapsto \mathbf{Q}_k\mathbf{y}$ is a bijection from $\mathbb{C}^k \setminus \{\mathbf{0}\}$ onto $\mathcal{K}_k \setminus \{\mathbf{0}\}$, because the columns of $\mathbf{Q}_k$ are a basis of $\mathcal{K}_k$ (Theorem 24.6.2 (a)). Under it, with $\mathbf{x} = \mathbf{Q}_k\mathbf{y}$,

$$\langle \mathbf{A}\mathbf{x}, \mathbf{x}\rangle = \mathbf{y}^*\mathbf{Q}_k^*\mathbf{A}\mathbf{Q}_k\mathbf{y} = \mathbf{y}^*\mathbf{T}_k\mathbf{y}, \qquad \langle \mathbf{x}, \mathbf{x}\rangle = \mathbf{y}^*\mathbf{Q}_k^*\mathbf{Q}_k\mathbf{y} = \mathbf{y}^*\mathbf{y},$$

so $R_{\mathbf{A}}(\mathbf{x}) = R_{\mathbf{T}_k}(\mathbf{y})$ and the two maxima agree. Monotonicity in k follows at once from $\mathcal{K}_k \subseteq \mathcal{K}_{k+1}$ (Proposition 24.6.1 (a)), and $\theta_1^{(k)} \le \lambda_1$ from Proposition 17.1.1 applied on all of $\mathbb{C}^n$ — which is also Theorem 24.6.4 (a) at $i = 1$.

Step 2. Let $p \in \mathbb{R}[x]$ have $\deg p \le k-1$ and $p(\lambda_1) \neq 0$, and put $\mathbf{x} = p(\mathbf{A})\mathbf{b}$. Then $\mathbf{x} \in \mathcal{K}_k$ by Proposition 24.6.1 (a), and $\mathbf{x} = \sum_i c_i p(\lambda_i)\mathbf{v}_i$ since $\mathbf{A}\mathbf{v}_i = \lambda_i\mathbf{v}_i$. The list $(\mathbf{v}_i)$ being orthonormal, Theorem 11.2.1 gives

$$R_{\mathbf{A}}(\mathbf{x}) = \frac{\sum_i |c_i|^2 p(\lambda_i)^2 \lambda_i}{\sum_i |c_i|^2 p(\lambda_i)^2},$$

and $\mathbf{x} \neq \mathbf{0}$ because its first coefficient $c_1 p(\lambda_1)$ is nonzero. Subtracting from λ_1,

$$\lambda_1 - R_{\mathbf{A}}(\mathbf{x}) = \frac{\sum_{i\ge 2} |c_i|^2 p(\lambda_i)^2 (\lambda_1 - \lambda_i)}{\sum_i |c_i|^2 p(\lambda_i)^2},$$

the term $i = 1$ having vanished from the numerator. Every factor $\lambda_1 - \lambda_i$ is at most $\lambda_1 - \lambda_n$; and the denominator is at least its single term $|c_1|^2 p(\lambda_1)^2$, all terms being nonnegative. Hence

$$\begin{aligned} \lambda_1 - R_{\mathbf{A}}(\mathbf{x}) &\le (\lambda_1 - \lambda_n)\, \frac{\sum_{i\ge 2} |c_i|^2 p(\lambda_i)^2}{|c_1|^2 p(\lambda_1)^2} \\ &\le (\lambda_1 - \lambda_n)\, \tau^2 \max_{i\ge 2} \Big(\frac{p(\lambda_i)}{p(\lambda_1)}\Big)^2. \end{aligned} \tag{$*$}$$

By Step 1, $\lambda_1 - \theta_1^{(k)} \le \lambda_1 - R_{\mathbf{A}}(\mathbf{x})$.

Step 3. Take $\mu = \tfrac{1}{2}(\lambda_2 + \lambda_n)$, the midpoint of the interval holding the unwanted eigenvalues, and $p(x) = (x-\mu)^{k-1}$, of degree $k-1$. For $i \ge 2$ we have $\lambda_n \le \lambda_i \le \lambda_2$, so $|\lambda_i - \mu| \le \tfrac{1}{2}(\lambda_2 - \lambda_n)$. And

$$\lambda_1 - \mu = \tfrac{1}{2}(\lambda_2 - \lambda_n) + (\lambda_1 - \lambda_2) > \tfrac{1}{2}(\lambda_2 - \lambda_n) \ge 0,$$

using $\lambda_1 > \lambda_2$; in particular $p(\lambda_1) = (\lambda_1 - \mu)^{k-1} \neq 0$. Therefore

$$\max_{i\ge 2} \Big|\frac{p(\lambda_i)}{p(\lambda_1)}\Big| \le \Big(\frac{(\lambda_2 - \lambda_n)/2}{\lambda_1 - \mu}\Big)^{k-1} = \rho^{\,k-1},$$

since $2(\lambda_1 - \mu) = 2\lambda_1 - \lambda_2 - \lambda_n$. Substituting into $(*)$ gives the bound, and $\rho < 1$ because $\lambda_1 > \lambda_2$. This proves the theorem.

■

For $\mathbf{A} = \operatorname{diag}(4,3,2,1)$ and $\mathbf{b} = (1,1,1,1)/2$ the constants are $\tau^2 = 3$, $\lambda_1 - \lambda_n = 3$, and $\rho = \tfrac{1}{2}$, so the bound reads $9 \cdot 4^{-(k-1)}$: the values $9, 2.25, 0.5625$ for $k = 1,2,3$, against the true gaps 1.5, 0.382, 0.068 of Example (Ritz values closing in on the extremes). The bound is correct and loose, and the looseness is in Step 3.

It is worth naming what ρ is. Assume for the rest of this subsection that $\lambda_2 > \lambda_n$, which the theorem does not require; if instead $\lambda_2 = \lambda_n$, then $\rho = 0$ and the bound of Theorem 24.6.5 is already as strong as this route can make it. Put

$$\gamma = \frac{\lambda_1 - \lambda_2}{\lambda_2 - \lambda_n},$$

the size of the gap above λ_2 measured against the spread of the rest. Then $2\lambda_1 - \lambda_2 - \lambda_n = 2(\lambda_1 - \lambda_2) + (\lambda_2 - \lambda_n) = (\lambda_2 - \lambda_n)(1 + 2\gamma)$, so

$$\rho = \frac{1}{1 + 2\gamma}.$$

(For $\mathbf{A} = \operatorname{diag}(4, 3, 2, 1)$, $\gamma = \frac{1}{2}$ and $\rho = \frac{1}{2}$.) So Step 3 buys a factor $(1 + 2\gamma)^{-(k-1)}$ per unit of k, and a small gap makes it nearly worthless.

What is not proved here. The polynomial of Step 3 is not the best one. Minimizing $\max_{i \geq 2} |p(\lambda_i)/p(\lambda_1)|$ over the polynomials of degree at most $k - 1$ with $p(\lambda_1) = 1$ is a question about how small a polynomial can be kept on an interval while staying large at a point outside it, and its answer is the Chebyshev polynomial T_{k-1}, shifted and scaled to that interval. The resulting estimate, due to Kaniel, Paige, and Saad, is

$$0 \leq \lambda_1 - \theta_1^{(k)} \leq \frac{(\lambda_1 - \lambda_n)\,\tau^2}{\left(T_{k-1}(1 + 2\gamma)\right)^2}.$$

The next section defines the Chebyshev polynomials and proves the two facts this bound rests on: that T_k stays within 1 on $[-1, 1]$, and that it grows geometrically just outside it; the substitution into $(*)$ is then immediate, but **we do not carry it out here, and nothing in this section depends on it.** The reason it is worth stating is the shape of the improvement. For small γ, $T_{k-1}(1 + 2\gamma)$ grows roughly like $(1 + 2\sqrt{\gamma})^{k-1}$, where Step 3 managed only $(1 + 2\gamma)^{k-1}$. A small gap costs a square root, not a full factor. That is the same square root that the next section extracts for conjugate gradients, and it is the reason Krylov methods are used at all.

24.6.6 Losing orthogonality

Everything above is exact arithmetic. In the floating-point model of Definition 24.1.1, one feature of the Lanczos recurrence fails, and it fails in a way that is worth describing honestly rather than proving badly.

The three-term recurrence never re-examines $\mathbf{q}_1, \ldots, \mathbf{q}_{j-2}$. In exact arithmetic it does not need to: Theorem 24.6.3 proves that $\mathbf{w}_j$ is automatically orthogonal to them. In floating-point arithmetic the computed vectors satisfy the recurrence to within a small multiple of $u\|\mathbf{A}\|$, and nothing removes the small components along the earlier vectors that rounding introduces. Those components do not stay small.

The mechanism is specific, and it is not random drift. The component of the computed $\mathbf{q}_{k+1}$ along a Ritz vector $\mathbf{x}$ grows in inverse proportion to that Ritz pair's residual $\beta_k |y_k|$. So orthogonality is lost first, and most, in exactly the direction of a Ritz pair that has **converged** — the better the approximation, the faster the basis is corrupted in its direction. The visible symptom is that the computed $\mathbf{T}_k$ acquires a second copy of a Ritz value that converged several steps earlier, and later a third: the spurious duplicates, which correspond to no extra multiplicity in $\mathbf{A}$. An algorithm that counts eigenvalues by counting Ritz values will count wrong.

This paragraph is a description of observed behavior, not a theorem, and this book does not prove it. The analysis that makes it precise — that the loss of orthogonality is

confined to the converged directions, and that the computed $\mathbf{T}_k$ is the exact tridiagonal matrix of a larger problem — is due to Paige, and it is outside the scope of this chapter. Nothing above depends on it. The standard remedies are to reorthogonalize each new vector against all the earlier ones, which restores the guarantee and abandons the three-term economy, or to reorthogonalize only against the converged Ritz vectors, which is cheaper and rests on precisely the analysis we are not giving. Both are engineering, not theorems, and we label them as such.

24.6.7 Exercises

24.6.7.1 A. Check your understanding

Exercise. A1

(a) Define $\mathcal{K}_k(\mathbf{A}, \mathbf{b})$ and state its dimension in terms of $m_{\mathbf{A},\mathbf{b}}$.

(b) State the Arnoldi identity of Theorem 24.6.2 (b), naming the shape of every matrix in it.

(c) True or false: for Hermitian $\mathbf{A}$, the Arnoldi process needs all of $\mathbf{q}_1, \dots, \mathbf{q}_j$ to produce $\mathbf{q}_{j+1}$. Justify your answer.

(d) What does a breakdown $h_{k+1,k} = 0$ mean, and why is it good news?

(e) Give the formula for the residual of a Ritz pair, and say why it costs nothing to evaluate.

Solution.

(a) $\mathcal{K}_k(\mathbf{A}, \mathbf{b}) = \operatorname{span}(\mathbf{b}, \mathbf{A}\mathbf{b}, \dots, \mathbf{A}^{k-1}\mathbf{b})$, and $\dim \mathcal{K}_k = \min(k, d)$ with $d = \deg m_{\mathbf{A},\mathbf{b}}$ (Proposition 24.6.1).

(b) $\mathbf{A}\mathbf{Q}_k = \mathbf{Q}_{k+1}\widetilde{\mathbf{H}}_k$, where $\mathbf{Q}_k \in M_{n\times k}$ and $\mathbf{Q}_{k+1} \in M_{n\times(k+1)}$ have orthonormal columns and $\widetilde{\mathbf{H}}_k \in M_{(k+1)\times k}$ is upper Hessenberg; equivalently $\mathbf{A}\mathbf{Q}_k = \mathbf{Q}_k\mathbf{H}_k + h_{k+1,k}\mathbf{q}_{k+1}\mathbf{e}_k^\top$ with $\mathbf{H}_k = \mathbf{Q}_k^*\mathbf{A}\mathbf{Q}_k \in M_k$ upper Hessenberg.

(c) False. By Theorem 24.6.3 the recurrence has three terms, so only $\mathbf{q}_{j-1}$ and $\mathbf{q}_j$ are needed. The reason is that $\mathbf{H}_k$ is both Hermitian and upper Hessenberg, hence tridiagonal.

(d) It means $\mathcal{K}_k$ is $\mathbf{A}$-invariant (Theorem 24.6.2 (c)), equivalently $k = d$. It is good news because every eigenvalue of $\mathbf{H}_k$ is then an exact eigenvalue of $\mathbf{A}$: the method has finished, not failed.

(e) $\|\mathbf{A}\mathbf{x} - \theta\mathbf{x}\| = h_{k+1,k}|y_k|$ (Theorem 24.6.4 (b)). Both factors are already known — $h_{k+1,k}$ from the last step of the process, y_k from the small eigenvector — so no multiplication by $\mathbf{A}$ is needed.

24.6.7.2 B. Practice

Exercise. B1: Dimensions

For each pair $(\mathbf{A}, \mathbf{b})$, determine $\dim \mathcal{K}_k(\mathbf{A}, \mathbf{b})$ for every $k \geq 1$. Justify your answer.

(a) $\mathbf{A} = \operatorname{diag}(2, 2, 5)$, $\mathbf{b} = (1, 1, 1)$.

(b) $\mathbf{A} = \operatorname{diag}(2, 2, 5)$, $\mathbf{b} = (1, -1, 0)$.

(c) $\mathbf{A} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$, $\mathbf{b} = \mathbf{e}_2$.

(d) $\mathbf{A}$ the 4×4 matrix with $\mathbf{A}\mathbf{e}_j = \mathbf{e}_{j+1}$ for $j \le 3$ and $\mathbf{A}\mathbf{e}_4 = \mathbf{0}$, $\mathbf{b} = \mathbf{e}_1$.

Solution. By Proposition 24.6.1 (b) the answer is always $\min(k, d)$ with $d = \deg m_{\mathbf{A},\mathbf{b}}$.

(a) $\mathbf{b}$ has nonzero components in both eigenspaces, so no single factor of $(x-2)(x-5)$ kills it, while the product does: $d = 2$, and the dimensions are $1, 2, 2, 2, \ldots$.

(b) $\mathbf{A}\mathbf{b} = (2, -2, 0) = 2\mathbf{b}$, so $\mathbf{b}$ is an eigenvector and $m_{\mathbf{A},\mathbf{b}} = x - 2$, $d = 1$. The dimensions are $1, 1, 1, \ldots$: every further product is wasted.

(c) $\mathbf{A}\mathbf{e}_2 = \mathbf{e}_1 \neq \mathbf{0}$ and $\mathbf{A}^2\mathbf{e}_2 = \mathbf{0}$, so $m_{\mathbf{A},\mathbf{e}_2} = x^2$ and $d = 2$: dimensions $1, 2, 2, \ldots$.

(d) $\mathbf{A}^j\mathbf{e}_1 = \mathbf{e}_{j+1}$ for $j \le 3$ and $\mathbf{A}^4\mathbf{e}_1 = \mathbf{0}$, so $m_{\mathbf{A},\mathbf{e}_1} = x^4$ and $d = 4$: dimensions $1, 2, 3, 4, 4, \ldots$. Here the Krylov spaces grow as long as they possibly can.

Exercise. B2: Two Arnoldi steps

Let

$$\mathbf{A} = \begin{pmatrix} 2 & 1 & 1 \\ 0 & 2 & 1 \\ 0 & 0 & 1 \end{pmatrix}, \qquad \mathbf{b} = \mathbf{e}_3.$$

Carry out two Arnoldi steps. Write down $\mathbf{Q}_2$, $\widetilde{\mathbf{H}}_2$, and $\mathbf{H}_2$, verify Eq. 24.6.1, and compute the two Ritz values with their residuals. Hence say which, if either, is a converged eigenvalue estimate.

Solution. $\mathbf{q}_1 = \mathbf{e}_3$ and $\mathbf{A}\mathbf{q}_1 = (1, 1, 1)$, so $h_{11} = \langle \mathbf{A}\mathbf{q}_1, \mathbf{q}_1 \rangle = 1$ and

$$\mathbf{w}_1 = (1, 1, 1) - (0, 0, 1) = (1, 1, 0), \qquad h_{21} = \sqrt{2}, \qquad \mathbf{q}_2 = \tfrac{1}{\sqrt{2}}(1, 1, 0).$$

Next $\mathbf{A}\mathbf{q}_2 = \frac{1}{\sqrt{2}}(3, 2, 0)$, so $h_{12} = \langle \mathbf{A}\mathbf{q}_2, \mathbf{q}_1 \rangle = 0$ and $h_{22} = \frac{1}{2}(3 + 2) = \frac{5}{2}$, and

$$\mathbf{w}_2 = \tfrac{1}{\sqrt{2}}(3, 2, 0) - \tfrac{5}{2} \cdot \tfrac{1}{\sqrt{2}}(1, 1, 0) = \tfrac{1}{2\sqrt{2}}(1, -1, 0), \qquad h_{32} = \tfrac{1}{2},$$

with $\mathbf{q}_3 = \frac{1}{\sqrt{2}}(1, -1, 0)$. Hence

$$\mathbf{Q}_2 = \begin{pmatrix} 0 & \frac{1}{\sqrt{2}} \\ 0 & \frac{1}{\sqrt{2}} \\ 1 & 0 \end{pmatrix}, \qquad \widetilde{\mathbf{H}}_2 = \begin{pmatrix} 1 & 0 \\ \sqrt{2} & \frac{5}{2} \\ 0 & \frac{1}{2} \end{pmatrix}, \qquad \mathbf{H}_2 = \begin{pmatrix} 1 & 0 \\ \sqrt{2} & \frac{5}{2} \end{pmatrix}.$$

Check Eq. 24.6.1 column by column: $\mathbf{A}\mathbf{q}_1 = (1, 1, 1) = 1 \cdot \mathbf{q}_1 + \sqrt{2}\,\mathbf{q}_2$, and $\mathbf{A}\mathbf{q}_2 = \frac{1}{\sqrt{2}}(3, 2, 0) = 0 \cdot \mathbf{q}_1 + \frac{5}{2}\mathbf{q}_2 + \frac{1}{2}\mathbf{q}_3$, both correct.

$\mathbf{H}_2$ is lower triangular as well as upper Hessenberg, so its eigenvalues are $\theta = 1$ and $\theta = \frac{5}{2}$. For $\theta = \frac{5}{2}$: solving $\mathbf{H}_2\mathbf{y} = \frac{5}{2}\mathbf{y}$ gives $y_1 = 0$, so $\mathbf{y} = (0, 1)$, $y_2 = 1$, and the residual is $h_{32}|y_2| = \frac{1}{2}$. For $\theta = 1$: $\sqrt{2}\,y_1 + \frac{3}{2}y_2 = 0$, so $\mathbf{y} = \frac{1}{\sqrt{17}}(3, -2\sqrt{2})$ and the residual is $\frac{1}{2} \cdot \frac{2\sqrt{2}}{\sqrt{17}} = \frac{\sqrt{34}}{17} \approx 0.343$.

Neither is converged. The eigenvalues of $\mathbf{A}$ are $2, 2, 1$, so the Ritz value 1 is exactly an eigenvalue — yet its residual is 0.343, not 0, and its Ritz vector $\mathbf{Q}_2\mathbf{y} = \frac{1}{\sqrt{17}}(-2, -2, 3)$ is not an eigenvector: the eigenvectors for 1 are the multiples of $(0, 1, -1)$, which is not in $\mathcal{K}_2 = \operatorname{span}(\mathbf{e}_3, (1, 1, 0))$. The agreement is a coincidence of two numbers, and the residual is what says so.

Exercise. B3: Lanczos on a tridiagonal matrix

Let $\mathbf{A} \in M_n(\mathbb{R})$ be symmetric tridiagonal with every subdiagonal entry nonzero, and let $\mathbf{b} = \mathbf{e}_1$. Prove that the Lanczos process produces $\mathbf{q}_j = \pm\mathbf{e}_j$ for every j, and hence that $\mathbf{T}_n = \mathbf{A}$ up to signs. What does this say about the cost of Lanczos on a matrix that is already tridiagonal?

Solution. We show by induction on j that

$$\mathbf{A}^j\mathbf{e}_1 = \sum_{i \le j+1} c_i^{(j)}\mathbf{e}_i \qquad \text{with} \qquad c_{j+1}^{(j)} = a_{21}a_{32}\cdots a_{j+1,j} \neq 0,$$

for $0 \le j \le n-1$. At $j = 0$ the sum is $\mathbf{e}_1$ and the coefficient is the empty product 1. Suppose it holds at $j-1$. Since $\mathbf{A}$ is tridiagonal, $\mathbf{A}\mathbf{e}_i \in \operatorname{span}(\mathbf{e}_{i-1}, \mathbf{e}_i, \mathbf{e}_{i+1})$, so applying $\mathbf{A}$ to $\mathbf{A}^{j-1}\mathbf{e}_1 = \sum_{i \le j} c_i^{(j-1)}\mathbf{e}_i$ lands in $\operatorname{span}(\mathbf{e}_1, \dots, \mathbf{e}_{j+1})$; and the only term contributing to $\mathbf{e}_{j+1}$ is $c_j^{(j-1)}\mathbf{A}\mathbf{e}_j$, whose $\mathbf{e}_{j+1}$-coefficient is $c_j^{(j-1)}a_{j+1,j} \neq 0$, both factors being nonzero. This proves the induction.

Consequently the change-of-basis matrix from $(\mathbf{e}_1, \dots, \mathbf{e}_j)$ to $(\mathbf{e}_1, \mathbf{A}\mathbf{e}_1, \dots, \mathbf{A}^{j-1}\mathbf{e}_1)$ is triangular with nonzero diagonal, so $\mathcal{K}_j(\mathbf{A}, \mathbf{e}_1) = \operatorname{span}(\mathbf{e}_1, \dots, \mathbf{e}_j)$ for $1 \le j \le n$.

By Theorem 24.6.2 (a), $\mathbf{q}_j$ is a unit vector in $\mathcal{K}_j$ orthogonal to $\mathcal{K}_{j-1} = \operatorname{span}(\mathbf{e}_1, \dots, \mathbf{e}_{j-1})$, hence a unit multiple of $\mathbf{e}_j$; the sign is fixed by $\beta_{j-1} > 0$. Therefore $(\mathbf{T}_n)_{ij} = \langle \mathbf{A}\mathbf{q}_j, \mathbf{q}_i \rangle = \pm a_{ij}$, and since $\mathbf{q}_j = \varepsilon_j\mathbf{e}_j$ with $\varepsilon_j = \pm 1$, $\mathbf{T}_n = \mathbf{D}\mathbf{A}\mathbf{D}$ with $\mathbf{D} = \operatorname{diag}(\varepsilon_1, \dots, \varepsilon_n)$ — a similarity, so the two have the same eigenvalues.

The cost lesson is that Lanczos does no work at all here beyond rediscovering the matrix: on a tridiagonal matrix with $\mathbf{b} = \mathbf{e}_1$ it reproduces $\mathbf{A}$ and finds nothing early. The method earns its keep when n is large and only the first few extreme eigenvalues are wanted, so that $k \ll n$.

24.6.7.3 C. Going deeper

Exercise. C1: Ritz values from the other end

Let $\mathbf{A} \in M_n(\mathbb{C})$ be Hermitian and let $\theta_1^{(k)} \ge \cdots \ge \theta_k^{(k)}$ be the Ritz values from $\mathcal{K}_k(\mathbf{A}, \mathbf{b})$, $1 \le k \le d$.

(a) Prove that $\theta_k^{(k)}$ is nonincreasing in k and that $\theta_k^{(k)} \ge \lambda_n(\mathbf{A})$.

(b) Deduce a bound for $\theta_k^{(k)} - \lambda_n(\mathbf{A})$ from Theorem 24.6.5, with no new argument.

(c) Show that at $k = d$ the Ritz values are d of the eigenvalues of $\mathbf{A}$, and say which.

Solution.

(a) The argument of Step 1 in the proof of Theorem 24.6.5 applies verbatim with minima in place of maxima, using the second half of Proposition 17.1.1: $\theta_k^{(k)} = \min\{R_{\mathbf{A}}(\mathbf{x}) : \mathbf{x} \in \mathcal{K}_k,\ \mathbf{x} \neq \mathbf{0}\}$. A minimum over a larger set is no larger, and $\mathcal{K}_k \subseteq \mathcal{K}_{k+1}$, so the sequence is nonincreasing; and the minimum over all of $\mathbb{C}^n$ is $\lambda_n(\mathbf{A})$, so it is bounded below by that. (Alternatively, both statements are Theorem 24.6.4 (a) at $i = k$.)

(b) Apply Theorem 24.6.5 to $-\mathbf{A}$. Its eigenvalues are $-\lambda_n \geq -\lambda_{n-1} \geq \cdots \geq -\lambda_1$, with the same orthonormal eigenbasis in reversed order. Its Krylov spaces from $\mathbf{b}$ are the same subspaces, because $(-\mathbf{A})^j\mathbf{b} = (-1)^j\mathbf{A}^j\mathbf{b}$ and scaling the vectors of a list does not change its span. And $R_{-\mathbf{A}} = -R_{\mathbf{A}}$, so by Step 1 of the proof the largest Ritz value of $-\mathbf{A}$ from $\mathcal{K}_k$ is $-\theta_k^{(k)}$. The theorem gives

$$\theta_k^{(k)} - \lambda_n \leq (\lambda_1 - \lambda_n)\,\tilde{\tau}^2\,\tilde{\rho}^{\,2(k-1)}, \qquad \tilde{\rho} = \frac{\lambda_1 - \lambda_{n-1}}{\lambda_1 + \lambda_{n-1} - 2\lambda_n},$$

with $\tilde{\tau}^2 = \sum_{i \leq n-1} |c_i|^2/|c_n|^2$, provided $\lambda_{n-1} > \lambda_n$ and $c_n \neq 0$.

(c) At $k = d$ the process breaks down, so $\mathcal{K}_d$ is $\mathbf{A}$-invariant and every Ritz value is an eigenvalue of $\mathbf{A}$ by Theorem 24.6.2 (c). Which ones: writing $\mathbf{b} = \sum_i c_i\mathbf{v}_i$ in an orthonormal eigenbasis and grouping the eigenvectors by eigenvalue, $p(\mathbf{A})\mathbf{b} = \sum_i c_i p(\lambda_i)\mathbf{v}_i$, so $\mathcal{K}_d$ is contained in the span of those $\mathbf{v}_i$ with $c_i \neq 0$. The Ritz values are therefore the distinct eigenvalues λ of $\mathbf{A}$ for which $\mathbf{b}$ has a nonzero component in $E_\lambda(\mathbf{A})$, each occurring once; and d is their number, as Example (A breakdown at step two) illustrates with $d = 2$ for a matrix with eigenvalues $5, 2, 2$.

Exercise. C2: Shift invariance, and what it costs

Let $\mathbf{A} \in M_n(F)$, $\mathbf{b} \neq \mathbf{0}$, and let $\sigma, c \in F$ with $c \neq 0$.

(a) Prove that $\mathcal{K}_k(c\mathbf{A} + \sigma\mathbf{I}, \mathbf{b}) = \mathcal{K}_k(\mathbf{A}, \mathbf{b})$ for every k.

(b) Deduce that the Arnoldi vectors $\mathbf{q}_j$ are the same for $\mathbf{A}$ and for $c\mathbf{A} + \sigma\mathbf{I}$, up to a scalar of modulus one, and that the Ritz values transform by $\theta \mapsto c\theta + \sigma$.

(c) Contrast this with Section 5: explain in two sentences why shifting accelerates the QR algorithm but does **not** accelerate a Krylov method, and what one must do instead to get the effect.

Solution.

(a) Put $\mathbf{B} = c\mathbf{A} + \sigma\mathbf{I}$. By Proposition 24.6.1 (a), $\mathcal{K}_k(\mathbf{B}, \mathbf{b}) = \{p(\mathbf{B})\mathbf{b} : \deg p \leq k-1\}$ and $p(\mathbf{B}) = p(c\mathbf{A} + \sigma\mathbf{I}) = \tilde{p}(\mathbf{A})$ where $\tilde{p}(x) = p(cx + \sigma)$ has the same degree, $c \neq 0$. The map $p \mapsto \tilde{p}$ is a bijection of $F[x]_{\leq k-1}$ onto itself, with inverse $q \mapsto q((x-\sigma)/c)$. So the two sets of vectors coincide.

(b) The Arnoldi vectors are determined by the flag $\mathcal{K}_1 \subseteq \mathcal{K}_2 \subseteq \ldots$ alone, up to a scalar of modulus one: $\mathbf{q}_j$ is a unit vector in $\mathcal{K}_j$ orthogonal to $\mathcal{K}_{j-1}$ (Theorem 24.6.2 (a)), and the unit vectors of $\mathcal{K}_j \cap \mathcal{K}_{j-1}^{\perp}$ are exactly the modulus-one multiples of one of them. By (a) the flag is the same for $\mathbf{B}$ as for $\mathbf{A}$, so the run on $\mathbf{B}$ produces $\mathbf{Q}_k^{\mathbf{B}} = \mathbf{Q}_k\mathbf{D}$ with $\mathbf{D} = \operatorname{diag}(d_1, \ldots, d_k)$ and $|d_j| = 1$. The scalars need not all be 1: each run fixes its own by $h_{j,j-1} > 0$, and for $c = -1$, $\sigma = 0$, $\mathbf{A} = \operatorname{diag}(4, 3, 2, 1)$, and $\mathbf{b} = (1, 1, 1, 1)$ the

two runs give $\mathbf{q}_j^{\mathbf{B}} = (-1)^{j-1}\mathbf{q}_j$. Consequently

$$(\mathbf{Q}_k^{\mathbf{B}})^*\mathbf{B}\,\mathbf{Q}_k^{\mathbf{B}} = \mathbf{D}^*(c\,\mathbf{Q}_k^*\mathbf{A}\mathbf{Q}_k + \sigma\mathbf{Q}_k^*\mathbf{Q}_k)\mathbf{D} = \mathbf{D}^*(c\mathbf{H}_k + \sigma\mathbf{I}_k)\mathbf{D},$$

whose eigenvalues, $\mathbf{D}$ being unitary, are those of $c\mathbf{H}_k + \sigma\mathbf{I}_k$, namely $c\theta_i + \sigma$.

(c) The QR algorithm converges at a rate governed by ratios $|\lambda_{i+1} - \mu|/|\lambda_i - \mu|$, and those ratios genuinely change with μ; a Krylov space does not change at all under a shift, by (a), so nothing it produces can change either. To get the analogous effect one must change the space, not the matrix: apply the method to $(\mathbf{A}-\mu\mathbf{I})^{-1}$, whose Krylov spaces are different, which is the Krylov version of the inverse iteration of Theorem 24.4.3 and costs a factorization of $\mathbf{A} - \mu\mathbf{I}$ rather than a matrix-vector product.

Exercise. C3: When the Krylov space is everything

Let $\mathbf{A} \in M_n(F)$.

(a) Prove that there exists $\mathbf{b}$ with $\mathcal{K}_n(\mathbf{A}, \mathbf{b}) = F^n$ if and only if $m_{\mathbf{A}} = p_{\mathbf{A}}$.

(b) Prove that in that case the Arnoldi process run to $k = n$ produces a unitary $\mathbf{Q}_n$ with $\mathbf{Q}_n^*\mathbf{A}\mathbf{Q}_n$ upper Hessenberg, all of whose subdiagonal entries are nonzero.

(c) Hence give a second proof of Theorem 11.8.5 over $\mathbb{C}$ for such $\mathbf{A}$, and explain why it does not replace Chapter 11's.

Hint for (a): Chapter 10 §05 calls such a $\mathbf{b}$ *a cyclic vector.*

Solution.

(a) $\mathcal{K}_n(\mathbf{A}, \mathbf{b}) = F^n$ means $\dim \mathcal{K}_n = n$, which by Proposition 24.6.1 (b) means $\deg m_{\mathbf{A},\mathbf{b}} = n$, that is, $Z(\mathbf{b}; \mathbf{A}) = F^n$ and $\mathbf{b}$ is a cyclic vector for $\mathbf{A}$ (Definition 10.5.4). By Theorem 10.5.6, such a vector exists exactly when $m_{\mathbf{A}} = p_{\mathbf{A}}$.

(b) With $d = n$, Theorem 24.6.2 (a) says that $h_{j+1,j} \neq 0$ for every $j < n$ and that $(\mathbf{q}_1, \dots, \mathbf{q}_n)$ is an orthonormal basis of F^n; so $\mathbf{Q}_n$ is unitary, and $\mathbf{Q}_n^*\mathbf{A}\mathbf{Q}_n = \mathbf{H}_n$ is upper Hessenberg by Theorem 24.6.2 (b). Its subdiagonal entries are the $h_{j+1,j}$, all nonzero. (A Hessenberg matrix with this property is called **unreduced**, and Proposition 24.5.5 is about exactly these.)

(c) Part (b) exhibits a unitary $\mathbf{Q}_n$ with $\mathbf{Q}_n^*\mathbf{A}\mathbf{Q}_n$ upper Hessenberg, which is the conclusion of Theorem 11.8.5. It does not replace that theorem for two reasons. First, it needs $m_{\mathbf{A}} = p_{\mathbf{A}}$ by (a), while Theorem 11.8.5 holds for every square matrix; when no cyclic vector exists, the process breaks down at some $d < n$ and delivers an invariant subspace instead of a full basis. Second, Theorem 11.8.5 is proved by reflections, each applied to the whole matrix, and its $O(n^3)$ cost is what Section 5 assumed; the Arnoldi route needs n matrix-vector products and $O(n^3)$ orthogonalization work as well, and in floating-point arithmetic the orthogonality it produces is exactly what the last subsection says can be lost. The two constructions agree in exact arithmetic when both apply, and the reflection route is the one that is run.

24.7 Steepest Descent and Conjugate Gradients

Section 2 solved $\mathbf{Ax} = \mathbf{b}$ by elimination: about $\frac{2}{3}n^3$ operations, an exact answer after finitely many of them, and a factorization that fills in most of the zeros the matrix had. For a large sparse matrix that is the wrong bargain. This section builds two iterations that touch $\mathbf{A}$ only through the product $\mathbf{x} \mapsto \mathbf{Ax}$, for the case where $\mathbf{A}$ is positive definite, and counts what each of them buys. The first costs about κ steps per decimal digit of accuracy, the second about $\sqrt{\kappa}$. Where that square root comes from is the point of the section.

Throughout, $\mathbf{A} \in M_n(\mathbb{R})$ is symmetric and positive definite, $\mathbf{b} \in \mathbb{R}^n$, and $\mathbf{x}^\star = \mathbf{A}^{-1}\mathbf{b}$ is the exact solution. Iteration indices are superscripts in parentheses. We work over $\mathbb{R}$ because the function this section minimizes is built from $\mathbf{b}^\top\mathbf{x}$, which over $\mathbb{C}$ is not a real number; everything below survives over $\mathbb{C}$ with $\mathbf{b}^*\mathbf{x}$ replaced by its real part, and nothing else changes.

24.7.1 A linear system is a minimization problem

Positive definiteness is a statement about the number $\mathbf{x}^\top\mathbf{Ax}$ (Definition 13.1.1), so it invites us to look for a *quantity* that the solution makes small, rather than an equation the solution satisfies. There is one, and it is the simplest possible candidate: a quadratic minus a linear term.

Solving $\mathbf{Ax} = \mathbf{b}$ is the same as sitting at the bottom of a bowl.

Definition 24.7.1. The Energy Inner Product and the Energy Norm

Let $\mathbf{A} \in M_n(\mathbb{R})$ be symmetric positive definite. The **energy inner product** is

$$\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{A}} \coloneqq \mathbf{y}^\top \mathbf{Ax} \qquad (\mathbf{x}, \mathbf{y} \in \mathbb{R}^n),$$

and the **energy norm** is $\|\mathbf{x}\|_{\mathbf{A}} \coloneqq \langle \mathbf{x}, \mathbf{x} \rangle_{\mathbf{A}}^{1/2} = (\mathbf{x}^\top \mathbf{Ax})^{1/2}$. Two vectors are **A-orthogonal**, or **conjugate**, when $\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{A}} = 0$.

This is the inner product of Chapter 13 §04, written there as $\mathbf{y}^*\mathbf{Ax}$. It really is one: it is bilinear because matrix multiplication is; it is symmetric because the 1×1 matrix $\mathbf{y}^\top\mathbf{Ax}$ equals its own transpose $\mathbf{x}^\top\mathbf{Ay}$; and $\langle \mathbf{x}, \mathbf{x} \rangle_{\mathbf{A}} > 0$ for $\mathbf{x} \neq \mathbf{0}$ is clause (P2) of Definition 13.1.1 in its strict form. So $\|\cdot\|_{\mathbf{A}}$ is the induced norm of Definition 11.1.3, and every fact about norms and orthogonality from Chapter 11 applies verbatim. That is the reason for introducing it: it lets us reuse the Pythagorean theorem in a geometry bent by $\mathbf{A}$.

The word *energy* comes from the springs of Section 9, where $\frac{1}{2}\mathbf{x}^\top\mathbf{Kx}$ is the potential energy stored in a displacement $\mathbf{x}$.

Proposition 24.7.1. The Solution Minimizes the Energy Functional

Let $\mathbf{A} \in M_n(\mathbb{R})$ be symmetric positive definite, $\mathbf{b} \in \mathbb{R}^n$, and define

$$\phi(\mathbf{x}) \coloneqq \tfrac{1}{2}\,\mathbf{x}^\top\mathbf{Ax} - \mathbf{b}^\top\mathbf{x} \qquad (\mathbf{x} \in \mathbb{R}^n).$$

Then, with $\mathbf{x}^\star = \mathbf{A}^{-1}\mathbf{b}$,

$$\phi(\mathbf{x}) - \phi(\mathbf{x}^\star) = \tfrac{1}{2}\|\mathbf{x} - \mathbf{x}^\star\|_{\mathbf{A}}^2 \qquad \text{for every } \mathbf{x} \in \mathbb{R}^n.$$

In particular $\mathbf{x}^\star$ is the **unique** minimizer of ϕ.

Proof. $\mathbf{A}$ is invertible, since $\mathbf{A} \succ \mathbf{O}$ has only positive eigenvalues (Theorem 13.1.5, (a) $\Rightarrow$ (b)), so $\mathbf{x}^\star$ is defined and $\mathbf{A}\mathbf{x}^\star = \mathbf{b}$. Expand, using $\mathbf{A}^\top = \mathbf{A}$:

$$\begin{aligned}\tfrac{1}{2}\|\mathbf{x} - \mathbf{x}^\star\|_{\mathbf{A}}^2 &= \tfrac{1}{2}\mathbf{x}^\top\mathbf{A}\mathbf{x} - \mathbf{x}^\top\mathbf{A}\mathbf{x}^\star + \tfrac{1}{2}(\mathbf{x}^\star)^\top\mathbf{A}\mathbf{x}^\star \\ &= \tfrac{1}{2}\mathbf{x}^\top\mathbf{A}\mathbf{x} - \mathbf{b}^\top\mathbf{x} + \tfrac{1}{2}\,\mathbf{b}^\top\mathbf{x}^\star,\end{aligned}$$

where the second line uses $\mathbf{A}\mathbf{x}^\star = \mathbf{b}$ twice and $\mathbf{x}^\top\mathbf{b} = \mathbf{b}^\top\mathbf{x}$. The first two terms are $\phi(\mathbf{x})$. Putting $\mathbf{x} = \mathbf{x}^\star$ gives $0 = \phi(\mathbf{x}^\star) + \tfrac{1}{2}\mathbf{b}^\top\mathbf{x}^\star$, so the leftover constant is $-\phi(\mathbf{x}^\star)$, which is the identity. Since $\|\cdot\|_{\mathbf{A}}$ is a norm, the right-hand side is ≥ 0 with equality only at $\mathbf{x} = \mathbf{x}^\star$. This proves the proposition. ∎

So every iteration below may be read twice: as an attempt to solve a linear system, and as an attempt to walk downhill on a bowl whose shape is $\mathbf{A}$. The identity also tells us which error to measure. The natural yardstick is not $\|\mathbf{x} - \mathbf{x}^\star\|_2$ but $\|\mathbf{x} - \mathbf{x}^\star\|_{\mathbf{A}}$, because that is the one the algorithms actually make small.

Two quantities recur. The **error** is $\boldsymbol{\varepsilon} = \mathbf{x} - \mathbf{x}^\star$ and the **residual** is $\mathbf{r} = \mathbf{b} - \mathbf{A}\mathbf{x}$, and they are related by

$$\mathbf{r} = \mathbf{b} - \mathbf{A}\mathbf{x} = \mathbf{A}\mathbf{x}^\star - \mathbf{A}\mathbf{x} = -\mathbf{A}\boldsymbol{\varepsilon}. \tag{24.7.1}$$

The residual is computable and the error is not; Eq. 24.7.1 is the exchange rate between them.

24.7.2 Steepest descent

Fix $\mathbf{x}$ and a direction $\mathbf{d} \neq \mathbf{0}$, and ask what happens along the line $\mathbf{x} + t\mathbf{d}$. Expanding ϕ and using $\mathbf{A}^\top = \mathbf{A}$,

$$\phi(\mathbf{x} + t\mathbf{d}) = \phi(\mathbf{x}) - t\,\mathbf{d}^\top\mathbf{r} + \tfrac{1}{2}t^2\,\mathbf{d}^\top\mathbf{A}\mathbf{d}, \tag{24.7.2}$$

where $\mathbf{r} = \mathbf{b} - \mathbf{A}\mathbf{x}$. Since $\mathbf{d}^\top\mathbf{A}\mathbf{d} > 0$, the right-hand side is a parabola in t opening upwards, so it has one minimum, at

$$t = \frac{\mathbf{d}^\top\mathbf{r}}{\mathbf{d}^\top\mathbf{A}\mathbf{d}}, \qquad \text{with decrease} \quad \frac{(\mathbf{d}^\top\mathbf{r})^2}{2\,\mathbf{d}^\top\mathbf{A}\mathbf{d}}. \tag{24.7.3}$$

Choosing t this way is an **exact line search**; it costs one product $\mathbf{A}\mathbf{d}$.

Which $\mathbf{d}$ should we take? Read Eq. 24.7.2 for small t: the decrease is $t\,\mathbf{d}^\top\mathbf{r}$ to first order, and among unit vectors $\mathbf{d}$ the quantity $\mathbf{d}^\top\mathbf{r}$ is largest exactly at $\mathbf{d} = \mathbf{r}/\|\mathbf{r}\|_2$, by Cauchy–Schwarz (Theorem 11.1.5) with its equality case. So the residual is the direction in which ϕ falls fastest *at the starting point*, measured in the ordinary Euclidean norm.

Warning. **"Steepest" is a statement about the first instant, not about the step.** The direction that minimizes ϕ over the whole line is not $\mathbf{r}$ but $\mathbf{A}^{-1}\mathbf{r}$: with $\mathbf{d} = \mathbf{A}^{-1}\mathbf{r}$, Eq. 24.7.3 gives $t = 1$ and $\mathbf{x} + \mathbf{d} = \mathbf{x} + \mathbf{A}^{-1}(\mathbf{b} - \mathbf{A}\mathbf{x}) = \mathbf{x}^\star$, the exact answer in one step. That direction is unavailable, because producing it means solving the very system we are trying to solve. Every method in this section is an attempt to approximate $\mathbf{A}^{-1}\mathbf{r}$ cheaply.

Definition 24.7.2. Steepest Descent

Let $\mathbf{A} \succ \mathbf{O}$ and $\mathbf{b} \in \mathbb{R}^n$, and let $\mathbf{x}^{(0)} \in \mathbb{R}^n$. **Steepest descent** is the iteration: for $k = 0, 1, 2, \ldots$, set $\mathbf{r}^{(k)} = \mathbf{b} - \mathbf{A}\mathbf{x}^{(k)}$; if $\mathbf{r}^{(k)} = \mathbf{0}$, stop, since $\mathbf{x}^{(k)} = \mathbf{x}^\star$; otherwise put

$$t_k = \frac{(\mathbf{r}^{(k)})^\top \mathbf{r}^{(k)}}{(\mathbf{r}^{(k)})^\top \mathbf{A}\mathbf{r}^{(k)}}, \qquad \mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + t_k\, \mathbf{r}^{(k)}.$$

The denominator is positive because $\mathbf{r}^{(k)} \neq \mathbf{0}$ and $\mathbf{A} \succ \mathbf{O}$, so t_k is defined, and $t_k > 0$. One step costs one matrix–vector product.

To see how fast this is we need one inequality, and it is a statement about weighted averages, not about matrices at all.

Lemma 24.7.2. Kantorovich's Inequality

Let $\mathbf{A} \in M_n(\mathbb{R})$ be symmetric positive definite with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n > 0$. Then for every $\mathbf{y} \in \mathbb{R}^n$,

$$(\mathbf{y}^\top \mathbf{A}\mathbf{y})(\mathbf{y}^\top \mathbf{A}^{-1}\mathbf{y}) \leq \frac{(\lambda_1 + \lambda_n)^2}{4\lambda_1\lambda_n}\,(\mathbf{y}^\top \mathbf{y})^2.$$

Idea. In an orthonormal eigenbasis both factors become weighted averages of the eigenvalues, one of λ and one of $1/\lambda$, with the same weights. The obstacle is that $1/\lambda$ is not linear, so the second average is not determined by the first. It is, however, *bounded* by it: on the interval $[\lambda_n, \lambda_1]$ the curve $1/\lambda$ lies below the straight line joining its endpoints. Replacing the curve by that line makes the second average a linear function of the first, and the product becomes a scalar quadratic that we maximize by hand.

Proof. Write $a = \lambda_1$ and $c = \lambda_n$. For every λ with $c \leq \lambda \leq a$,

$$\frac{a + c - \lambda}{ac} - \frac{1}{\lambda} = \frac{\lambda(a + c - \lambda) - ac}{ac\lambda} = \frac{-(\lambda - a)(\lambda - c)}{ac\lambda} \;\geq\; 0, \tag{$*$}$$

because $\lambda - a \leq 0$, $\lambda - c \geq 0$, and $ac\lambda > 0$.

We may assume $\mathbf{y} \neq \mathbf{0}$, the case $\mathbf{y} = \mathbf{0}$ being an equality $0 = 0$, and by homogeneity (both sides scale by s^4 under $\mathbf{y} \mapsto s\mathbf{y}$) we may assume $\mathbf{y}^\top \mathbf{y} = 1$. By the spectral theorem (Corollary 12.5.2) there is an orthonormal basis $\mathbf{q}_1, \ldots, \mathbf{q}_n$ of $\mathbb{R}^n$ with $\mathbf{A}\mathbf{q}_i = \lambda_i \mathbf{q}_i$. Let $u_i = \mathbf{q}_i^\top \mathbf{y}$ be the coordinates of $\mathbf{y}$ in that basis, and put $w_i = u_i^2$. Then $\sum_i w_i = \|\mathbf{y}\|_2^2 = 1$

by Parseval's identity (Theorem 11.9.2), every $w_i \geq 0$, and

$$\mathbf{y}^{\top}\mathbf{A}\mathbf{y} = \sum_i \lambda_i w_i =: \mu, \qquad \mathbf{y}^{\top}\mathbf{A}^{-1}\mathbf{y} = \sum_i \frac{w_i}{\lambda_i},$$

the second because $\mathbf{A}^{-1}\mathbf{q}_i = \lambda_i^{-1}\mathbf{q}_i$. Since the w_i are weights summing to 1 and every $\lambda_i \in [c, a]$, the number μ lies in $[c, a]$. Averaging $(*)$ with the weights w_i,

$$\mathbf{y}^{\top}\mathbf{A}^{-1}\mathbf{y} \leq \sum_i w_i \, \frac{a + c - \lambda_i}{ac} = \frac{a + c - \mu}{ac}.$$

Hence $(\mathbf{y}^{\top}\mathbf{A}\mathbf{y})(\mathbf{y}^{\top}\mathbf{A}^{-1}\mathbf{y}) \leq \mu(a + c - \mu)/(ac)$. Finally, for any real μ,

$$\mu(a + c - \mu) = \frac{(a+c)^2}{4} - \Big(\frac{a+c}{2} - \mu\Big)^2 \leq \frac{(a+c)^2}{4},$$

and dividing by ac gives the bound. This proves the lemma.

■

Remark. Equality is attainable: take $\mathbf{y} = (\mathbf{q}_1 + \mathbf{q}_n)/\sqrt{2}$, so that $w_1 = w_n = \tfrac{1}{2}$ and every other weight is 0. Then $(*)$ is an equality at each λ_i carrying weight, since those are a and c, and $\mu = (a+c)/2$ makes the last step an equality too. Only the weights $w_i = u_i^2$ enter, so $(\mathbf{q}_1 - \mathbf{q}_n)/\sqrt{2}$ serves just as well, and so does any nonzero multiple of either. Example (The bound is attained) does exactly this, up to sign and scale.

We also record the condition number in the form we shall use it. For $\mathbf{A} \succ \mathbf{O}$, Lemma 17.3.5 gives $\|\mathbf{A}\|_2 = \lambda_1$, since all eigenvalues are positive; the eigenvalues of $\mathbf{A}^{-1}$ are the λ_i^{-1}, so the same lemma gives $\|\mathbf{A}^{-1}\|_2 = \lambda_n^{-1}$. Hence, by Definition 16.8.1,

$$\kappa := \kappa_2(\mathbf{A}) = \frac{\lambda_1}{\lambda_n} \geq 1. \tag{24.7.4}$$

Theorem 24.7.3. Convergence of Steepest Descent

Let $\mathbf{A} \in M_n(\mathbb{R})$ be symmetric positive definite with $\kappa = \kappa_2(\mathbf{A})$, let $\mathbf{b} \in \mathbb{R}^n$, and let $\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, \ldots$ be the steepest descent iterates. Then for every k for which $\mathbf{x}^{(k)}$ is defined,

$$\|\mathbf{x}^{(k)} - \mathbf{x}^{\star}\|_{\mathbf{A}} \leq \Big(\frac{\kappa - 1}{\kappa + 1}\Big)^{k} \|\mathbf{x}^{(0)} - \mathbf{x}^{\star}\|_{\mathbf{A}}.$$

Idea. One step at a time. Write the new energy-norm error in terms of the old one; the exact line search makes the difference a single explicit fraction, whose numerator and denominator are exactly the three quantities Kantorovich's inequality compares. What comes out is $1 - 4\lambda_1\lambda_n/(\lambda_1 + \lambda_n)^2$, and that number is the square of $(\kappa - 1)/(\kappa + 1)$.

Proof. It is enough to prove the estimate for one step and then iterate. Fix k and write $\mathbf{x} = \mathbf{x}^{(k)}$, $\mathbf{r} = \mathbf{r}^{(k)}$, $t = t_k$, and $\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}^{(k)} := \mathbf{x}^{(k)} - \mathbf{x}^\star$, and assume $\mathbf{r} \neq \mathbf{0}$. By Eq. 24.7.1, $\mathbf{r} = -\mathbf{A}\boldsymbol{\varepsilon}$, so

$$\|\boldsymbol{\varepsilon}\|_{\mathbf{A}}^2 = \boldsymbol{\varepsilon}^{\top}\mathbf{A}\boldsymbol{\varepsilon} = (\mathbf{A}^{-1}\mathbf{r})^{\top}\mathbf{A}(\mathbf{A}^{-1}\mathbf{r}) = \mathbf{r}^{\top}\mathbf{A}^{-1}\mathbf{r},$$

using $\mathbf{A}^{\top} = \mathbf{A}$ and $(\mathbf{A}^{-1})^{\top} = \mathbf{A}^{-1}$. The new error is $\boldsymbol{\varepsilon} + t\mathbf{r}$, so

$$\begin{aligned}\|\boldsymbol{\varepsilon} + t\mathbf{r}\|_{\mathbf{A}}^2 &= \|\boldsymbol{\varepsilon}\|_{\mathbf{A}}^2 + 2t\,\mathbf{r}^{\top}\mathbf{A}\boldsymbol{\varepsilon} + t^2\,\mathbf{r}^{\top}\mathbf{A}\mathbf{r} \\ &= \|\boldsymbol{\varepsilon}\|_{\mathbf{A}}^2 - 2t\,\mathbf{r}^{\top}\mathbf{r} + t^2\,\mathbf{r}^{\top}\mathbf{A}\mathbf{r},\end{aligned}$$

the second line because $\mathbf{A}\boldsymbol{\varepsilon} = -\mathbf{r}$. Substituting $t = (\mathbf{r}^{\top}\mathbf{r})/(\mathbf{r}^{\top}\mathbf{A}\mathbf{r})$ collapses the last two terms:

$$\|\boldsymbol{\varepsilon} + t\mathbf{r}\|_{\mathbf{A}}^2 = \|\boldsymbol{\varepsilon}\|_{\mathbf{A}}^2 - \frac{(\mathbf{r}^{\top}\mathbf{r})^2}{\mathbf{r}^{\top}\mathbf{A}\mathbf{r}} = \|\boldsymbol{\varepsilon}\|_{\mathbf{A}}^2\Big(1 - \frac{(\mathbf{r}^{\top}\mathbf{r})^2}{(\mathbf{r}^{\top}\mathbf{A}\mathbf{r})(\mathbf{r}^{\top}\mathbf{A}^{-1}\mathbf{r})}\Big),$$

where the last step divides and multiplies by $\|\boldsymbol{\varepsilon}\|_{\mathbf{A}}^2 = \mathbf{r}^{\top}\mathbf{A}^{-1}\mathbf{r}$, which is nonzero because $\mathbf{r} \neq \mathbf{0}$. By Lemma 24.7.2 applied to $\mathbf{y} = \mathbf{r}$,

$$\frac{(\mathbf{r}^{\top}\mathbf{r})^2}{(\mathbf{r}^{\top}\mathbf{A}\mathbf{r})(\mathbf{r}^{\top}\mathbf{A}^{-1}\mathbf{r})} \;\geq\; \frac{4\lambda_1\lambda_n}{(\lambda_1 + \lambda_n)^2},$$

so the bracket is at most

$$1 - \frac{4\lambda_1\lambda_n}{(\lambda_1+\lambda_n)^2} = \frac{(\lambda_1 - \lambda_n)^2}{(\lambda_1+\lambda_n)^2} = \Big(\frac{\kappa - 1}{\kappa + 1}\Big)^2,$$

the last equality on dividing numerator and denominator by λ_n^2 and using Eq. 24.7.4. Since $\boldsymbol{\varepsilon} + t\mathbf{r} = \mathbf{x}^{(k+1)} - \mathbf{x}^\star = \boldsymbol{\varepsilon}^{(k+1)}$, taking square roots gives $\|\boldsymbol{\varepsilon}^{(k+1)}\|_{\mathbf{A}} \leq \frac{\kappa-1}{\kappa+1}\|\boldsymbol{\varepsilon}^{(k)}\|_{\mathbf{A}}$, and an induction on k proves the theorem. ∎

Example. The bound is attained

Let $\mathbf{A} = \operatorname{diag}(9, 1)$, $\mathbf{b} = (9, 1)$, and $\mathbf{x}^{(0)} = (2, -8)$. Run three steps of steepest descent and compare with Theorem 24.7.3.

Solution. Here $\mathbf{x}^\star = (1, 1)$ and $\kappa = 9$, so the bound's factor is $\frac{4}{5}$. The first residual is

$$\mathbf{r}^{(0)} = \mathbf{b} - \mathbf{A}\mathbf{x}^{(0)} = (9, 1) - (18, -8) = (-9, 9),$$

so $\mathbf{r}^{\top}\mathbf{r} = 162$ and $\mathbf{r}^{\top}\mathbf{A}\mathbf{r} = 9 \cdot 81 + 81 = 810$, giving $t_0 = \frac{1}{5}$. Then

$$\mathbf{x}^{(1)} = (2, -8) + \tfrac{1}{5}(-9, 9) = \big(\tfrac{1}{5}, -\tfrac{31}{5}\big), \qquad \mathbf{r}^{(1)} = \big(\tfrac{36}{5}, \tfrac{36}{5}\big).$$

The two components of $\mathbf{r}^{(1)}$ are again equal in absolute value, so $t_1 = \frac{1}{5}$ as well, and $\mathbf{x}^{(2)} = \big(\frac{41}{25}, -\frac{119}{25}\big)$, $\mathbf{x}^{(3)} = \big(\frac{61}{125}, -\frac{451}{125}\big)$. The squared energy-norm errors are

$$90, \qquad \tfrac{288}{5}, \qquad \tfrac{4608}{125}, \qquad \tfrac{73728}{3125},$$

and each is exactly $\frac{16}{25}$ times the one before. So the ratio is $\frac{4}{5}$ at every step: the bound of Theorem 24.7.3 holds with equality here, and no better constant is possible.

The picture behind the example is a staircase. On an elongated bowl the steepest direction points across the valley rather than along it, so the iterates cross and re-cross. The residuals $\mathbf{r}^{(0)}$ and $\mathbf{r}^{(2)}$ in Example (The bound is attained) are parallel: the method keeps searching in directions it has already used.

Quick check. Show that consecutive steepest descent residuals are always orthogonal, $(\mathbf{r}^{(k+1)})^{\top}\mathbf{r}^{(k)} = 0$. What does that say about the shape of the path?

Solution. $\mathbf{r}^{(k+1)} = \mathbf{b} - \mathbf{A}\mathbf{x}^{(k+1)} = \mathbf{b} - \mathbf{A}\mathbf{x}^{(k)} - t_k\mathbf{A}\mathbf{r}^{(k)} = \mathbf{r}^{(k)} - t_k\mathbf{A}\mathbf{r}^{(k)}$, so

$$(\mathbf{r}^{(k+1)})^{\top}\mathbf{r}^{(k)} = \mathbf{r}^{\top}\mathbf{r} - t_k\,\mathbf{r}^{\top}\mathbf{A}\mathbf{r} = 0$$

by the choice of t_k, writing $\mathbf{r} = \mathbf{r}^{(k)}$. So each step turns a right angle: the path is a staircase. In two dimensions that forces every second direction to be parallel, which is why Example (The bound is attained) never escapes its worst case.

24.7.3 Conjugate directions

The defect is now visible: each line search spoils the minimality achieved along the previous direction. We want directions $\mathbf{p}^{(0)}, \mathbf{p}^{(1)}, \dots$ for which a search along $\mathbf{p}^{(k)}$ leaves the earlier work intact, so that after k steps the iterate is optimal over the whole k-dimensional affine piece explored so far.

Read the requirement off the goal. Being optimal over $\mathbf{x}^{(0)} + \operatorname{span}(\mathbf{p}^{(0)}, \dots, \mathbf{p}^{(k-1)})$ means, by Proposition 24.7.1, that the error is $\mathbf{A}$-orthogonal to that span; and one step changes the error by $t_k\mathbf{p}^{(k)}$. So the condition that preserves the earlier orthogonality is that $\mathbf{p}^{(k)}$ itself be $\mathbf{A}$-orthogonal to $\mathbf{p}^{(0)}, \dots, \mathbf{p}^{(k-1)}$. That is forced, not guessed. Which conjugate directions? The cheapest answer: take the residual and remove only what $\mathbf{A}$-orthogonality forbids.

Definition 24.7.3. The Conjugate Gradient Iteration

Let $\mathbf{A} \in M_n(\mathbb{R})$ be symmetric positive definite, $\mathbf{b} \in \mathbb{R}^n$, and $\mathbf{x}^{(0)} \in \mathbb{R}^n$. Set

$$\mathbf{r}^{(0)} = \mathbf{b} - \mathbf{A}\mathbf{x}^{(0)}, \qquad \mathbf{p}^{(0)} = \mathbf{r}^{(0)}.$$

For $k = 0, 1, 2, \dots$: if $\mathbf{r}^{(k)} = \mathbf{0}$, stop. Otherwise put

$$\alpha_k = \frac{(\mathbf{r}^{(k)})^{\top}\mathbf{r}^{(k)}}{(\mathbf{p}^{(k)})^{\top}\mathbf{A}\mathbf{p}^{(k)}}, \qquad \mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k\mathbf{p}^{(k)},$$

$$\mathbf{r}^{(k+1)} = \mathbf{r}^{(k)} - \alpha_k\mathbf{A}\mathbf{p}^{(k)}, \qquad \beta_k = \frac{(\mathbf{r}^{(k+1)})^{\top}\mathbf{r}^{(k+1)}}{(\mathbf{r}^{(k)})^{\top}\mathbf{r}^{(k)}},$$

and $\mathbf{p}^{(k+1)} = \mathbf{r}^{(k+1)} + \beta_k\mathbf{p}^{(k)}$. This is the **conjugate gradient** iteration, or **CG**.

A nonexample, by dropping one clause. Keep everything about the recursion and only weaken $\mathbf{A} \succ \mathbf{O}$ to "symmetric and invertible". Take

$$\mathbf{A} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \qquad \mathbf{b} = (1, 1), \qquad \mathbf{x}^{(0)} = \mathbf{0}.$$

Then $\mathbf{p}^{(0)} = \mathbf{r}^{(0)} = (1, 1)$ and $(\mathbf{p}^{(0)})^\top \mathbf{A}\mathbf{p}^{(0)} = 1 - 1 = 0$: the very first α_0 divides by zero, and the iteration does not start. What failed is exactly clause (P2) of Definition 13.1.1 in its strict form, which is what guarantees $\mathbf{p}^\top \mathbf{A}\mathbf{p} > 0$ for $\mathbf{p} \neq \mathbf{0}$. Symmetry alone is not enough, and nor is invertibility; both Proposition 24.7.1 (the bowl has a bottom) and Theorem 24.7.4 (the search directions are usable) need definiteness.

Three remarks on the shape of the recursion. The update of $\mathbf{r}$ is not a fresh evaluation of $\mathbf{b} - \mathbf{A}\mathbf{x}$ but a consequence of the update of $\mathbf{x}$; the two agree, since $\mathbf{b} - \mathbf{A}\mathbf{x}^{(k+1)} = \mathbf{r}^{(k)} - \alpha_k \mathbf{A}\mathbf{p}^{(k)}$. Only one product $\mathbf{A}\mathbf{p}^{(k)}$ is formed per step, serving both α_k and $\mathbf{r}^{(k+1)}$. And the new direction corrects the residual by a multiple of the **single** previous direction, not of all of them — the surprise, and the reason CG is short.

Theorem 24.7.4. Conjugate Gradients: Orthogonality and Optimality

Let $\mathbf{A} \succ \mathbf{O}$, $\mathbf{b} \in \mathbb{R}^n$, $\mathbf{x}^{(0)} \in \mathbb{R}^n$, and write $\mathcal{K}_k = \mathcal{K}_k(\mathbf{A}, \mathbf{r}^{(0)})$ for the Krylov subspaces of Definition 24.6.1, with $\mathcal{K}_0 = \{\mathbf{0}\}$. Suppose the iteration of Definition 24.7.3 has produced nonzero $\mathbf{r}^{(0)}, \dots, \mathbf{r}^{(k)}$. Then:

(a) $\mathbf{p}^{(0)}, \dots, \mathbf{p}^{(k)}$ are nonzero and pairwise $\mathbf{A}$-orthogonal, and every α_j and β_j with $j \leq k$ is defined, with $\alpha_j > 0$;

(b) $\mathbf{r}^{(0)}, \dots, \mathbf{r}^{(k)}$ are pairwise orthogonal, and $(\mathbf{r}^{(k+1)})^\top \mathbf{p}^{(j)} = 0$ for every $j \leq k$;

(c) $\operatorname{span}(\mathbf{r}^{(0)}, \dots, \mathbf{r}^{(k)}) = \operatorname{span}(\mathbf{p}^{(0)}, \dots, \mathbf{p}^{(k)}) = \mathcal{K}_{k+1}$, a space of dimension $k + 1$;

(d) $\mathbf{x}^{(k+1)}$ is the **unique** minimizer of $\|\mathbf{x} - \mathbf{x}^\star\|_{\mathbf{A}}$ over the affine set $\mathbf{x}^{(0)} + \mathcal{K}_{k+1}$.

Idea. Everything is one induction, and the two coefficients are exactly what the induction needs. Step k has to deliver two orthogonality relations: the new residual against the old directions, and the new direction against the old directions in the $\mathbf{A}$-inner product. The first is forced by α_k, the second by β_k, and each forces it only against the *latest* predecessor; the older ones come free, because $\mathbf{A}\mathbf{p}^{(j)}$ is a combination of $\mathbf{r}^{(j)}$ and $\mathbf{r}^{(j+1)}$, and the new residual is already orthogonal to both. Optimality (d) is then the Pythagorean theorem, applied in the energy inner product.

Proof. Parts (a), (b), and (c) are proved together by induction on k. Throughout, $\mathbf{A}^\top = \mathbf{A}$ is used silently to move $\mathbf{A}$ across a dot product, and $\mathbf{A}\mathcal{K}_j \subseteq \mathcal{K}_{j+1}$ directly from Definition 24.6.1.

Base $k = 0$**.** Here $\mathbf{p}^{(0)} = \mathbf{r}^{(0)} \neq \mathbf{0}$, so $(\mathbf{p}^{(0)})^\top \mathbf{A}\mathbf{p}^{(0)} > 0$ and $\alpha_0 = \|\mathbf{r}^{(0)}\|_2^2 / (\mathbf{p}^{(0)})^\top \mathbf{A}\mathbf{p}^{(0)}$ is defined and positive; that is (a). Part (c) is $\operatorname{span}(\mathbf{r}^{(0)}) = \operatorname{span}(\mathbf{p}^{(0)}) = \mathcal{K}_1$, which is Definition 24.6.1, and the dimension is 1 because $\mathbf{r}^{(0)} \neq \mathbf{0}$. For (b), a single vector is vacuously an orthogonal list, and

$$(\mathbf{r}^{(1)})^\top \mathbf{p}^{(0)} = (\mathbf{r}^{(0)})^\top \mathbf{p}^{(0)} - \alpha_0 (\mathbf{p}^{(0)})^\top \mathbf{A}\mathbf{p}^{(0)} = \|\mathbf{r}^{(0)}\|_2^2 - \|\mathbf{r}^{(0)}\|_2^2 = 0,$$

using $\mathbf{r}^{(1)} = \mathbf{r}^{(0)} - \alpha_0 \mathbf{A}\mathbf{p}^{(0)}$, $\mathbf{p}^{(0)} = \mathbf{r}^{(0)}$, and the definition of α_0.

Inductive step. Assume (a), (b), (c) for k, and assume $\mathbf{r}^{(k+1)} \neq \mathbf{0}$, so that β_k and $\mathbf{p}^{(k+1)}$ are formed. We prove (a), (b), (c) for $k+1$.

Step 1: the residuals stay orthogonal. By (b), $\mathbf{r}^{(k+1)}$ is orthogonal to $\mathbf{p}^{(0)}, \dots, \mathbf{p}^{(k)}$, and by (c) that list spans the same space as $\mathbf{r}^{(0)}, \dots, \mathbf{r}^{(k)}$. Hence $\mathbf{r}^{(k+1)} \perp \mathbf{r}^{(j)}$ for every $j \leq k$, and with (b) this makes $\mathbf{r}^{(0)}, \dots, \mathbf{r}^{(k+1)}$ pairwise orthogonal.

Step 2: (c) for $k+1$. From $\mathbf{r}^{(k+1)} = \mathbf{r}^{(k)} - \alpha_k \mathbf{A}\mathbf{p}^{(k)}$ with $\mathbf{r}^{(k)}, \mathbf{p}^{(k)} \in \mathcal{K}_{k+1}$, we get $\mathbf{r}^{(k+1)} \in \mathcal{K}_{k+2}$, and then $\mathbf{p}^{(k+1)} = \mathbf{r}^{(k+1)} + \beta_k \mathbf{p}^{(k)} \in \mathcal{K}_{k+2}$ as well. Moreover

$$\operatorname{span}(\mathbf{p}^{(0)}, \dots, \mathbf{p}^{(k+1)}) = \operatorname{span}(\mathbf{p}^{(0)}, \dots, \mathbf{p}^{(k)}, \mathbf{r}^{(k+1)}) = \operatorname{span}(\mathbf{r}^{(0)}, \dots, \mathbf{r}^{(k+1)}),$$

the first equality because $\mathbf{p}^{(k+1)}$ differs from $\mathbf{r}^{(k+1)}$ by a multiple of $\mathbf{p}^{(k)}$, the second by (c). By Step 1 the vectors $\mathbf{r}^{(0)}, \dots, \mathbf{r}^{(k+1)}$ are nonzero and pairwise orthogonal, hence independent (Theorem 11.1.4), so this common span has dimension $k+2$. It sits inside $\mathcal{K}_{k+2}$, which is spanned by $k+2$ vectors and so has dimension at most $k+2$; equal dimensions force equal spaces.

Step 3: $\mathbf{p}^{(k+1)} \neq \mathbf{0}$, and α_{k+1} is defined. By (b), $(\mathbf{r}^{(k+1)})^\top \mathbf{p}^{(k)} = 0$, so

$$(\mathbf{p}^{(k+1)})^\top \mathbf{r}^{(k+1)} = \|\mathbf{r}^{(k+1)}\|_2^2 + \beta_k (\mathbf{p}^{(k)})^\top \mathbf{r}^{(k+1)} = \|\mathbf{r}^{(k+1)}\|_2^2 > 0. \tag{$\dagger$}$$

Hence $\mathbf{p}^{(k+1)} \neq \mathbf{0}$, so $(\mathbf{p}^{(k+1)})^\top \mathbf{A}\mathbf{p}^{(k+1)} > 0$ and α_{k+1} is defined and positive.

Step 4: (a) for $k+1$. We must show $(\mathbf{p}^{(k+1)})^\top \mathbf{A}\mathbf{p}^{(j)} = 0$ for $j \leq k$. For $j = k$, use $\mathbf{A}\mathbf{p}^{(k)} = (\mathbf{r}^{(k)} - \mathbf{r}^{(k+1)})/\alpha_k$:

$$(\mathbf{p}^{(k+1)})^\top \mathbf{A}\mathbf{p}^{(k)} = (\mathbf{r}^{(k+1)})^\top \mathbf{A}\mathbf{p}^{(k)} + \beta_k (\mathbf{p}^{(k)})^\top \mathbf{A}\mathbf{p}^{(k)}.$$

By Step 1 the first term is $(0 - \|\mathbf{r}^{(k+1)}\|_2^2)/\alpha_k$, and $(\mathbf{p}^{(k)})^\top \mathbf{A}\mathbf{p}^{(k)} = \|\mathbf{r}^{(k)}\|_2^2/\alpha_k$ by the definition of α_k. The two cancel exactly when $\beta_k = \|\mathbf{r}^{(k+1)}\|_2^2/\|\mathbf{r}^{(k)}\|_2^2$, which is how β_k was defined. For $j < k$, the term $\beta_k (\mathbf{p}^{(k)})^\top \mathbf{A}\mathbf{p}^{(j)}$ vanishes by (a), and

$$\mathbf{A}\mathbf{p}^{(j)} = \frac{\mathbf{r}^{(j)} - \mathbf{r}^{(j+1)}}{\alpha_j} \in \operatorname{span}(\mathbf{r}^{(0)}, \dots, \mathbf{r}^{(k)}),$$

so $(\mathbf{r}^{(k+1)})^\top \mathbf{A}\mathbf{p}^{(j)} = 0$ by Step 1.

Step 5: (b) for $k+1$. The first half is Step 1. For the second, $\mathbf{r}^{(k+2)} = \mathbf{r}^{(k+1)} - \alpha_{k+1}\mathbf{A}\mathbf{p}^{(k+1)}$, so for $j = k+1$,

$$(\mathbf{r}^{(k+2)})^\top \mathbf{p}^{(k+1)} = (\mathbf{r}^{(k+1)})^\top \mathbf{p}^{(k+1)} - \alpha_{k+1} (\mathbf{p}^{(k+1)})^\top \mathbf{A}\mathbf{p}^{(k+1)} = 0$$

by ($\dagger$) and the definition of α_{k+1}; and for $j \leq k$,

$$(\mathbf{r}^{(k+2)})^\top \mathbf{p}^{(j)} = (\mathbf{r}^{(k+1)})^\top \mathbf{p}^{(j)} - \alpha_{k+1} (\mathbf{p}^{(k+1)})^\top \mathbf{A}\mathbf{p}^{(j)} = 0 - 0 = 0$$

by (b) and Step 4. This completes the induction.

Proof of (d). By construction $\mathbf{x}^{(k+1)} = \mathbf{x}^{(0)} + \sum_{j \leq k} \alpha_j \mathbf{p}^{(j)}$, which lies in $\mathbf{x}^{(0)} + \mathcal{K}_{k+1}$ by (c). Let $\boldsymbol{\varepsilon} = \mathbf{x}^{(k+1)} - \mathbf{x}^\star$. By Eq. 24.7.1, $\mathbf{A}\boldsymbol{\varepsilon} = -\mathbf{r}^{(k+1)}$, so for every $\mathbf{w} \in \mathcal{K}_{k+1} = \operatorname{span}(\mathbf{p}^{(0)}, \dots, \mathbf{p}^{(k)})$,

$$\langle \boldsymbol{\varepsilon}, \mathbf{w} \rangle_{\mathbf{A}} = \mathbf{w}^\top \mathbf{A}\boldsymbol{\varepsilon} = -\mathbf{w}^\top \mathbf{r}^{(k+1)} = 0$$

by (b). Now take any $\mathbf{x} = \mathbf{x}^{(0)} + \mathbf{z}$ with $\mathbf{z} \in \mathcal{K}_{k+1}$, and write $\mathbf{x} - \mathbf{x}^\star = \boldsymbol{\varepsilon} + \mathbf{w}$ with $\mathbf{w} = \mathbf{z} - \sum_{j \le k} \alpha_j \mathbf{p}^{(j)} \in \mathcal{K}_{k+1}$. By the Pythagorean theorem (Theorem 11.1.3) in the energy inner product,

$$\|\mathbf{x} - \mathbf{x}^\star\|_{\mathbf{A}}^2 = \|\boldsymbol{\varepsilon}\|_{\mathbf{A}}^2 + \|\mathbf{w}\|_{\mathbf{A}}^2 \;\ge\; \|\boldsymbol{\varepsilon}\|_{\mathbf{A}}^2,$$

with equality only for $\mathbf{w} = \mathbf{0}$, that is only for $\mathbf{x} = \mathbf{x}^{(k+1)}$. This proves the theorem.

■

Corollary 24.7.5. Finite Termination

In exact arithmetic the conjugate gradient iteration reaches $\mathbf{r}^{(m)} = \mathbf{0}$, and hence $\mathbf{x}^{(m)} = \mathbf{x}^\star$, for some $m \le n$. If $\mathbf{A}$ has only d distinct eigenvalues, then $m \le d$.

Proof. Suppose $\mathbf{r}^{(0)}, \dots, \mathbf{r}^{(k)}$ are all nonzero. By Theorem 24.7.4 (c), $\dim \mathcal{K}_{k+1} = k+1$. Since $\mathcal{K}_{k+1} \subseteq \mathbb{R}^n$, this forces $k+1 \le n$; so the residuals cannot all be nonzero for $k = n$, and $\mathbf{r}^{(m)} = \mathbf{0}$ for some $m \le n$. Then $\mathbf{A}\mathbf{x}^{(m)} = \mathbf{b}$, so $\mathbf{x}^{(m)} = \mathbf{x}^\star$.

For the refinement, let $\mu_1, \dots, \mu_d$ be the distinct eigenvalues and put $q(x) = (x - \mu_1) \cdots (x - \mu_d)$. By Corollary 12.5.2, $\mathbf{A} = \mathbf{Q}\mathbf{D}\mathbf{Q}^\top$ with $\mathbf{Q}$ orthogonal and $\mathbf{D}$ diagonal carrying the eigenvalues, so $q(\mathbf{A}) = \mathbf{Q}\, q(\mathbf{D})\, \mathbf{Q}^\top = \mathbf{O}$, every diagonal entry of $q(\mathbf{D})$ being q at an eigenvalue. Hence $\mathbf{A}^d \mathbf{r}^{(0)}$ is a combination of $\mathbf{r}^{(0)}, \mathbf{A}\mathbf{r}^{(0)}, \dots, \mathbf{A}^{d-1}\mathbf{r}^{(0)}$, so $\mathcal{K}_{d+1} = \mathcal{K}_d$ and $\dim \mathcal{K}_k \le d$ for every k. If $\mathbf{r}^{(0)}, \dots, \mathbf{r}^{(d)}$ were all nonzero we would get $\dim \mathcal{K}_{d+1} = d+1 > d$, a contradiction. So $m \le d$.

■

Example. Two distinct eigenvalues, two steps

Let $\mathbf{A} = \mathbf{I}_3 + \mathbf{J}$, where $\mathbf{J}$ is the all-ones matrix, let $\mathbf{b} = \mathbf{e}_1$ and $\mathbf{x}^{(0)} = \mathbf{0}$. Run CG.

Solution. $\mathbf{A}$ has rows $(2,1,1)$, $(1,2,1)$, $(1,1,2)$. Since $\mathbf{J}\mathbf{1} = 3\mathbf{1}$ and $\mathbf{J}$ kills $\mathbf{1}^\perp$, the eigenvalues of $\mathbf{A}$ are 4 (once) and 1 (twice): $d = 2$. Also $\mathbf{A}^{-1} = \mathbf{I} - \tfrac{1}{4}\mathbf{J}$, so $\mathbf{x}^\star = \mathbf{e}_1 - \tfrac{1}{4}\mathbf{1} = (\tfrac{3}{4}, -\tfrac{1}{4}, -\tfrac{1}{4})$.

Step 0. $\mathbf{r}^{(0)} = \mathbf{p}^{(0)} = (1,0,0)$, $(\mathbf{r}^{(0)})^\top \mathbf{r}^{(0)} = 1$, $\mathbf{A}\mathbf{p}^{(0)} = (2,1,1)$, $(\mathbf{p}^{(0)})^\top \mathbf{A}\mathbf{p}^{(0)} = 2$, so $\alpha_0 = \tfrac{1}{2}$ and

$$\mathbf{x}^{(1)} = (\tfrac{1}{2}, 0, 0), \qquad \mathbf{r}^{(1)} = (1,0,0) - \tfrac{1}{2}(2,1,1) = (0, -\tfrac{1}{2}, -\tfrac{1}{2}).$$

Then $\beta_0 = \tfrac{1}{2}/1 = \tfrac{1}{2}$ and $\mathbf{p}^{(1)} = (0, -\tfrac{1}{2}, -\tfrac{1}{2}) + \tfrac{1}{2}(1,0,0) = (\tfrac{1}{2}, -\tfrac{1}{2}, -\tfrac{1}{2})$.

Step 1. The rows of $\mathbf{A}$ applied to $\mathbf{p}^{(1)}$ give $1 - \tfrac{1}{2} - \tfrac{1}{2} = 0$, then $\tfrac{1}{2} - 1 - \tfrac{1}{2} = -1$, then $\tfrac{1}{2} - \tfrac{1}{2} - 1 = -1$; so $\mathbf{A}\mathbf{p}^{(1)} = (0, -1, -1)$ and $(\mathbf{p}^{(1)})^\top \mathbf{A}\mathbf{p}^{(1)} = 1$. With $(\mathbf{r}^{(1)})^\top \mathbf{r}^{(1)} = \tfrac{1}{2}$ this gives $\alpha_1 = \tfrac{1}{2}$ and

$$\mathbf{x}^{(2)} = (\tfrac{1}{2}, 0, 0) + \tfrac{1}{2}(\tfrac{1}{2}, -\tfrac{1}{2}, -\tfrac{1}{2}) = (\tfrac{3}{4}, -\tfrac{1}{4}, -\tfrac{1}{4}) = \mathbf{x}^\star,$$

with $\mathbf{r}^{(2)} = (0, -\tfrac{1}{2}, -\tfrac{1}{2}) - \tfrac{1}{2}(0, -1, -1) = \mathbf{0}$. Two steps, as Corollary 24.7.5 predicts, for a 3×3 system. Checks: $(\mathbf{r}^{(1)})^\top \mathbf{r}^{(0)} = 0$ and $(\mathbf{p}^{(1)})^\top \mathbf{A}\mathbf{p}^{(0)} = (\tfrac{1}{2}, -\tfrac{1}{2}, -\tfrac{1}{2}) \cdot (2,1,1) = 1 - \tfrac{1}{2} - \tfrac{1}{2} = 0$.

Warning. **Finite termination is a theorem about exact arithmetic, and it is not how CG is used.** In the floating-point model of Section 1 the computed residuals lose their orthogonality, as the Lanczos vectors of Section 6 do. The two algorithms are closely related: started from the same $\mathbf{A}$ and the same vector they build bases of the same Krylov spaces, and the tridiagonal matrix of Theorem 24.6.3 can be assembled from the CG coefficients — which are *not* the α_j and β_j of Theorem 24.6.3, those being the entries of that tridiagonal matrix, but different numbers. **That correspondence, due to Lanczos and to Hestenes and Stiefel, and the rounding behavior it carries across, due to Paige and Greenbaum, are described here and not proved; nothing in this section depends on either.** A run of CG on a matrix of order 10^6 is stopped after a few hundred steps, and the theorem that matters is Theorem 24.7.7 below.

24.7.4 Chebyshev polynomials

Theorem 24.7.4 (d) has an equivalent form that turns the whole question of speed into a question about polynomials, and that question has a classical answer. We need one family of polynomials and two facts about it.

The Chebyshev polynomials stay within 1 *on* $[-1,1]$ *and then grow geometrically outside it.*

Definition 24.7.4. Chebyshev Polynomials

The **Chebyshev polynomials** $T_0, T_1, T_2, \dots \in \mathbb{R}[x]$ are defined by

$$T_0 = 1, \qquad T_1 = x, \qquad T_{k+1} = 2x\,T_k - T_{k-1} \quad (k \geq 1).$$

The name comes from Chapter 11. Its table at the end of §10 lists the **monic** orthogonal polynomial sequence of Definition 11.10.1 for the weight $(1-t^2)^{-1/2}$ on $(-1,1)$, and the T_k above are its members rescaled — which is why the recurrence here is not the one of Theorem 11.10.3, that theorem being stated for the monic normalization, where no factor 2 appears. The book has not proved that identification and nothing below uses it. Here the recurrence *is* the definition, and we prove what we need from it directly. The first few are

$$T_2 = 2x^2 - 1, \quad T_3 = 4x^3 - 3x, \quad T_4 = 8x^4 - 8x^2 + 1.$$

We shall need the degrees, and the recurrence supplies them by induction on k. For $k = 0$ and $k = 1$, $\deg T_k = k$, with leading coefficient 1 in both cases. Let $k \geq 1$ and suppose $\deg T_{k-1} = k-1$ and $\deg T_k = k$ with leading coefficient c_k. Then $2xT_k$ has degree $k+1$ with leading coefficient $2c_k$, while $\deg T_{k-1} = k-1 < k+1$, so no cancellation is possible in $T_{k+1} = 2xT_k - T_{k-1}$ and $\deg T_{k+1} = k+1$ with leading coefficient $2c_k$. Hence $\deg T_k = k$ for every k, and since $c_1 = 1$ the doubling gives $c_k = 2^{k-1}$ for $k \geq 1$.

The one trick in the subject is a change of variable. Every complex number x can be written as $\frac{1}{2}(w + w^{-1})$ for some $w \neq 0$: the equation $w^2 - 2xw + 1 = 0$ has two roots in $\mathbb{C}$, neither of them 0 since their product is 1. In that variable the recurrence becomes trivial.

Lemma 24.7.6. Chebyshev Polynomials: Size Inside and Growth Outside

For every $k \geq 0$ and every $w \in \mathbb{C} \setminus \{0\}$,

$$T_k\Big(\frac{w+w^{-1}}{2}\Big) = \frac{w^k + w^{-k}}{2}. \tag{24.7.5}$$

Consequently:

(a) $|T_k(x)| \le 1$ for every real $x \in [-1,1]$, and $T_k(1) = 1$;

(b) for every real $z > 1$,

$$T_k(z) = \tfrac{1}{2}\Big((z+\sqrt{z^2-1})^k + (z+\sqrt{z^2-1})^{-k}\Big) \;\ge\; \tfrac{1}{2}(z+\sqrt{z^2-1})^k,$$

and, separately, $T_k(z) \ge 1$.

Idea. Under $x = \frac{1}{2}(w+w^{-1})$ the recurrence of Definition 24.7.4 says nothing more than $w^{k+1} + w^{-(k+1)} = (w+w^{-1})(w^k + w^{-k}) - (w^{k-1} + w^{-(k-1)})$, which is true because the cross terms cancel. The two facts are then two choices of w: on $[-1,1]$ take $|w| = 1$, so that w^k and w^{-k} are again on the unit circle; outside, take w real and greater than 1, so that w^k grows geometrically.

Proof. The identity Eq. 24.7.5. Induct on k. For $k = 0$ both sides are 1, and for $k = 1$ both are $\frac{1}{2}(w+w^{-1})$. Assume the identity for $k-1$ and k, with $k \ge 1$, and write $x = \frac{1}{2}(w+w^{-1})$. Then

$$\begin{aligned} T_{k+1}(x) &= 2x\,T_k(x) - T_{k-1}(x) \\ &= (w+w^{-1})\frac{w^k + w^{-k}}{2} - \frac{w^{k-1} + w^{-(k-1)}}{2} \\ &= \frac{w^{k+1} + w^{-(k-1)} + w^{k-1} + w^{-(k+1)} - w^{k-1} - w^{-(k-1)}}{2}, \end{aligned}$$

which is $\frac{1}{2}(w^{k+1} + w^{-(k+1)})$, as required.

(a). Let $x \in [-1,1]$ and put $w = x + i\sqrt{1-x^2}$. Then $|w|^2 = x^2 + (1-x^2) = 1$, so $w \ne 0$ and $w^{-1} = \overline{w}$ (Theorem 0.6.2), whence $\frac{1}{2}(w+w^{-1}) = \frac{1}{2}(w+\overline{w}) = x$. By Eq. 24.7.5 and $|u+v| \le |u| + |v|$ (Theorem 0.6.4),

$$|T_k(x)| = \tfrac{1}{2}|w^k + w^{-k}| \le \tfrac{1}{2}(|w|^k + |w|^{-k}) = 1.$$

For $x = 1$ take $w = 1$, giving $T_k(1) = \frac{1}{2}(1+1) = 1$.

(b). Let $z > 1$ and put $w = z + \sqrt{z^2-1}$, a real number > 1. Then

$$w(z - \sqrt{z^2-1}) = z^2 - (z^2-1) = 1,$$

so $w^{-1} = z - \sqrt{z^2-1}$ and $\frac{1}{2}(w+w^{-1}) = z$. Now Eq. 24.7.5 gives the displayed formula. Both w^k and w^{-k} are positive, so dropping the second gives the displayed inequality. For the separate bound, $\frac{1}{2}(s+s^{-1}) \ge 1$ for every $s > 0$, because $s + s^{-1} - 2 = (s^{1/2} - s^{-1/2})^2 \ge 0$; applying this with $s = w^k > 0$ gives $T_k(z) = \frac{1}{2}(w^k + w^{-k}) \ge 1$. This proves the lemma.

■

Warning. **Part (a) is false the moment x leaves $[-1,1]$, and spectacularly so.** $T_{10}(1.1)$ is already about 42.2, and $T_{50}(1.1)$ exceeds 2×10^9. That explosion is not a defect; it is the whole resource. The polynomial in the next theorem is small where the eigenvalues live and large at 0, and the ratio of the two is what CG converts into speed.

Remark. That the Chebyshev polynomials are also **extremal** — among monic polynomials of degree k, the scaled $T_k/2^{k-1}$ has the smallest maximum modulus on $[-1,1]$ — is classical, is not proved here, and is not needed. What Theorem 24.7.7 uses is only that T_k, rescaled, is an admissible competitor in the minimum of Eq. 24.7.6: small on the interval where the eigenvalues lie, and large at 0.

24.7.5 The square root

Theorem 24.7.7. Convergence of Conjugate Gradients

Let $\mathbf{A}\in M_n(\mathbb{R})$ be symmetric positive definite with $\kappa=\kappa_2(\mathbf{A})$, let $\mathbf{b}\in\mathbb{R}^n$, and let $\mathbf{x}^{(k)}$ be the conjugate gradient iterates from $\mathbf{x}^{(0)}$, with $\mathbf{x}^{(k)}:=\mathbf{x}^\star$ once the iteration has terminated. Then for every $k\ge 0$,

$$\|\mathbf{x}^{(k)}-\mathbf{x}^\star\|_{\mathbf{A}} \;\le\; 2\Big(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\Big)^{k}\|\mathbf{x}^{(0)}-\mathbf{x}^\star\|_{\mathbf{A}}.$$

Idea. Three moves. ① Theorem 24.7.4 says the CG error is the smallest among all errors reachable from the Krylov space; writing out what "reachable" means turns that into: the smallest value of $\|q(\mathbf{A})\boldsymbol{\varepsilon}^{(0)}\|_{\mathbf{A}}$ over polynomials q of degree $\le k$ with $q(0)=1$. ② In an eigenbasis that quantity is at most $\max_i|q(\lambda_i)|$ times the initial error, so we need one polynomial that is 1 at 0 and small on $[\lambda_n,\lambda_1]$. ③ Chebyshev's polynomial, rescaled so that the interval becomes $[-1,1]$, is that polynomial, and Lemma 24.7.6 measures it.

Proof. If $\kappa=1$ then $\lambda_1=\lambda_n$, so $\mathbf{A}=\lambda_1\mathbf{I}$ by the spectral theorem, and $\mathbf{p}^{(0)}=\mathbf{r}^{(0)}$ gives $\alpha_0=1/\lambda_1$ and $\mathbf{x}^{(1)}=\mathbf{x}^\star$; the bound then reads $0\le 0$ for $k\ge1$ and $\|\boldsymbol{\varepsilon}^{(0)}\|_{\mathbf{A}}\le 2\|\boldsymbol{\varepsilon}^{(0)}\|_{\mathbf{A}}$ for $k=0$, both true. So assume $\kappa>1$. We may also assume the iteration has not terminated before step k, since once it has, the left-hand side is 0.

Step 1: the error is a polynomial in A applied to the initial error. Write $\boldsymbol{\varepsilon}^{(0)}=\mathbf{x}^{(0)}-\mathbf{x}^\star$, so $\mathbf{r}^{(0)}=-\mathbf{A}\boldsymbol{\varepsilon}^{(0)}$ by Eq. 24.7.1. A vector of $\mathbf{x}^{(0)}+\mathcal{K}_k$ has the form $\mathbf{x}=\mathbf{x}^{(0)}+\sum_{j=0}^{k-1}c_j\mathbf{A}^j\mathbf{r}^{(0)}$, and then

$$\mathbf{x}-\mathbf{x}^\star=\boldsymbol{\varepsilon}^{(0)}-\sum_{j=0}^{k-1}c_j\mathbf{A}^{j+1}\boldsymbol{\varepsilon}^{(0)}=q(\mathbf{A})\,\boldsymbol{\varepsilon}^{(0)},\qquad q(x)=1-\sum_{j=0}^{k-1}c_jx^{j+1}.$$

As the c_j range over $\mathbb{R}^k$, q ranges over **all** polynomials of degree at most k with $q(0)=1$. So Theorem 24.7.4 (d) says

$$\|\boldsymbol{\varepsilon}^{(k)}\|_{\mathbf{A}} = \min\{\|q(\mathbf{A})\boldsymbol{\varepsilon}^{(0)}\|_{\mathbf{A}} : \deg q \le k,\ q(0) = 1\}. \tag{24.7.6}$$

Step 2: one eigenbasis estimate. By Corollary 12.5.2 take an orthonormal basis $\mathbf{q}_1, \dots, \mathbf{q}_n$ with $\mathbf{A}\mathbf{q}_i = \lambda_i \mathbf{q}_i$, and write $\boldsymbol{\varepsilon}^{(0)} = \sum_i u_i \mathbf{q}_i$. Then $q(\mathbf{A})\mathbf{q}_i = q(\lambda_i)\mathbf{q}_i$, so

$$\|q(\mathbf{A})\boldsymbol{\varepsilon}^{(0)}\|_{\mathbf{A}}^2 = \sum_i \lambda_i\, q(\lambda_i)^2 u_i^2 \le \Big(\max_i |q(\lambda_i)|\Big)^2 \sum_i \lambda_i u_i^2,$$

and $\sum_i \lambda_i u_i^2 = \|\boldsymbol{\varepsilon}^{(0)}\|_{\mathbf{A}}^2$. Every λ_i lies in $[\lambda_n, \lambda_1]$, so with Eq. 24.7.6, for any admissible q,

$$\|\boldsymbol{\varepsilon}^{(k)}\|_{\mathbf{A}} \le \Big(\max_{\lambda_n \le \lambda \le \lambda_1} |q(\lambda)|\Big)\|\boldsymbol{\varepsilon}^{(0)}\|_{\mathbf{A}}. \tag{24.7.7}$$

Step 3: the polynomial. Put

$$z = \frac{\lambda_1 + \lambda_n}{\lambda_1 - \lambda_n} = \frac{\kappa + 1}{\kappa - 1} > 1, \qquad q(\lambda) = \frac{1}{T_k(z)}\, T_k\Big(\frac{\lambda_1 + \lambda_n - 2\lambda}{\lambda_1 - \lambda_n}\Big).$$

The inner expression is an affine function of λ carrying λ_n to 1 and λ_1 to -1, so it maps $[\lambda_n, \lambda_1]$ into $[-1, 1]$; and it sends $\lambda = 0$ to z. By Lemma 24.7.6 (b), $T_k(z) \ge 1 > 0$, so q is a well-defined polynomial of degree at most k, and $q(0) = T_k(z)/T_k(z) = 1$. By Lemma 24.7.6 (a),

$$\max_{\lambda_n \le \lambda \le \lambda_1} |q(\lambda)| \le \frac{1}{T_k(z)}.$$

Step 4: measuring $T_k(z)$. Since $\kappa > 1$,

$$z^2 - 1 = \frac{(\kappa+1)^2 - (\kappa-1)^2}{(\kappa-1)^2} = \frac{4\kappa}{(\kappa-1)^2}, \qquad \sqrt{z^2-1} = \frac{2\sqrt{\kappa}}{\kappa - 1},$$

so, using $\kappa - 1 = (\sqrt{\kappa} - 1)(\sqrt{\kappa} + 1)$,

$$z + \sqrt{z^2-1} = \frac{\kappa + 1 + 2\sqrt{\kappa}}{\kappa - 1} = \frac{(\sqrt{\kappa}+1)^2}{(\sqrt{\kappa}-1)(\sqrt{\kappa}+1)} = \frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}.$$

By Lemma 24.7.6 (b), $T_k(z) \ge \tfrac{1}{2}\big((\sqrt{\kappa}+1)/(\sqrt{\kappa}-1)\big)^k$, so $1/T_k(z) \le 2\big((\sqrt{\kappa}-1)/(\sqrt{\kappa}+1)\big)^k$. Feeding this into Eq. 24.7.7 proves the theorem.

■

The square root is the whole gain. Take $\kappa = 100$. Steepest descent contracts by at most $99/101 \approx 0.9802$ per step, so reducing the energy-norm error by 10^{-6} needs about 691 steps; conjugate gradients contracts by $9/11 \approx 0.8182$, and $2(9/11)^k \le 10^{-6}$ already at $k = 73$. Both cost one product with $\mathbf{A}$ per step, so the ratio of the counts is the ratio of the work.

Quick check. Why does the factor 2 in Theorem 24.7.7 cost so little, and why is it there at all?

Solution. It costs little because it is a constant against a geometric factor: absorbing it changes the step count by $\log 2/\log((\sqrt{\kappa}+1)/(\sqrt{\kappa}-1))$, about 3.5 steps out of 73 when $\kappa = 100$. It is there because the lower bound for $T_k(z)$ in Lemma 24.7.6 (b) throws away the term w^{-k}; the surviving $\frac{1}{2}$ is exactly this 2.

Remark. The bound uses only the two ends of the spectrum, and it is often pessimistic. Eq. 24.7.7 holds for *every* admissible polynomial, so if the eigenvalues fall into a few tight clusters one may take q small near each cluster and do far better; Corollary 24.7.5 is the extreme case. Conjugate gradients sees the whole spectrum, not just its ends.

24.7.6 Two footnotes: Chebyshev iteration and preconditioning

The polynomial in Step 3 depends only on λ_1 and λ_n, so one can *fix* it in advance and update the iterates by the recurrence of Definition 24.7.4. This is the **Chebyshev iteration**. Three things are said about it: that it attains the bound of Theorem 24.7.7 rather than beating it; that it needs estimates of λ_1 and λ_n which CG does not; and that in exchange it forms no inner products, which matters when the work is split across many processors. **None of the three is proved here.** The method and its analysis are due to Golub and Varga, and nothing in this section depends on them.

Since the cost is governed by $\sqrt{\kappa}$, it pays to replace $\mathbf{A}\mathbf{x} = \mathbf{b}$ by an equivalent system of smaller condition number. Choosing a symmetric positive definite $\mathbf{M}$ close to $\mathbf{A}$, and writing $\mathbf{M} = \mathbf{C}\mathbf{C}^{\top}$, one applies CG to

$$(\mathbf{C}^{-1}\mathbf{A}\mathbf{C}^{-\top})\mathbf{y} = \mathbf{C}^{-1}\mathbf{b}, \qquad \mathbf{x} = \mathbf{C}^{-\top}\mathbf{y},$$

whose matrix is symmetric positive definite (Proposition 13.1.2) and whose eigenvalues are the generalized eigenvalues of the pencil $\mathbf{A} - \lambda\mathbf{M}$ of Definition 13.4.1, as Exercise C3 below asks you to check. This is **preconditioning**, and it is a subject rather than a theorem: which $\mathbf{M}$ to choose depends on where $\mathbf{A}$ came from, and no theorem of linear algebra answers that. We state the mechanism and stop.

24.7.7 Exercises

24.7.7.1 A. Check your understanding

Exercise. A1

(a) Define the energy inner product of a symmetric positive definite $\mathbf{A}$, and say which clause of Definition 13.1.1 makes it an inner product.

(b) State Proposition 24.7.1, and explain in one sentence why it makes $\|\cdot\|_{\mathbf{A}}$ the natural yardstick for the error.

(c) State the optimality property of the k-th conjugate gradient iterate, naming the set over which it is optimal.

(d) Determine whether the following statement is correct, and justify your answer: in exact arithmetic, conjugate gradients applied to a 1000×1000 positive definite matrix with 5 distinct eigenvalues terminates after at most 5 steps.

(e) Write down the two Chebyshev facts of Lemma 24.7.6, and say which of the two supplies the $\sqrt{\kappa}$ in Theorem 24.7.7.

Solution.

(a) $\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{A}} = \mathbf{y}^{\top} \mathbf{A} \mathbf{x}$. Symmetry and bilinearity hold for any symmetric $\mathbf{A}$; what makes it an inner product is $\langle \mathbf{x}, \mathbf{x} \rangle_{\mathbf{A}} = \mathbf{x}^{\top} \mathbf{A} \mathbf{x} > 0$ for $\mathbf{x} \neq \mathbf{0}$, which is clause (P2) of Definition 13.1.1 in its strict form.

(b) $\phi(\mathbf{x}) - \phi(\mathbf{x}^{\star}) = \tfrac{1}{2} \|\mathbf{x} - \mathbf{x}^{\star}\|_{\mathbf{A}}^{2}$, where $\phi(\mathbf{x}) = \tfrac{1}{2} \mathbf{x}^{\top} \mathbf{A} \mathbf{x} - \mathbf{b}^{\top} \mathbf{x}$ and $\mathbf{x}^{\star} = \mathbf{A}^{-1} \mathbf{b}$. So the energy norm of the error *is* the excess of ϕ over its minimum, up to the factor $\tfrac{1}{2}$: it is the quantity every descent method is built to decrease.

(c) By Theorem 24.7.4 (d), $\mathbf{x}^{(k)}$ is the unique minimizer of $\|\mathbf{x} - \mathbf{x}^{\star}\|_{\mathbf{A}}$ over the affine set $\mathbf{x}^{(0)} + \mathcal{K}_k(\mathbf{A}, \mathbf{r}^{(0)})$.

(d) Correct. By Corollary 24.7.5 the number of steps is at most the number d of distinct eigenvalues, here 5, and the order 1000 plays no part. The proof is that $q(\mathbf{A}) = \mathbf{O}$ for the degree-5 polynomial with the distinct eigenvalues as roots, so the Krylov spaces stop growing at dimension 5.

(e) $|T_k| \leq 1$ on $[-1, 1]$, and $T_k(z) \geq \tfrac{1}{2}(z + \sqrt{z^2 - 1})^k$ for $z > 1$. The second supplies the square root: at $z = (\kappa + 1)/(\kappa - 1)$ the number $z + \sqrt{z^2 - 1}$ equals $(\sqrt{\kappa} + 1)/(\sqrt{\kappa} - 1)$, and the square root enters through $\sqrt{z^2 - 1} = 2\sqrt{\kappa}/(\kappa - 1)$.

24.7.7.2 B. Practice

Exercise. B1: Steepest descent by hand

Let $\mathbf{A} = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}$, $\mathbf{b} = (2, 1)$, and $\mathbf{x}^{(0)} = (0, 0)$.

(a) Compute $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$ by steepest descent.

(b) Compute $\|\mathbf{x}^{(k)} - \mathbf{x}^{\star}\|_{\mathbf{A}}^{2}$ for $k = 0, 1, 2$ and compare the ratios with the bound of Theorem 24.7.3.

Solution. Here $\mathbf{x}^{\star} = (1, 1)$ and $\kappa = 2$, so the bound's factor is $(\kappa - 1)/(\kappa + 1) = \tfrac{1}{3}$.

(a) $\mathbf{r}^{(0)} = (2, 1)$, $\mathbf{r}^{\top} \mathbf{r} = 5$, $\mathbf{r}^{\top} \mathbf{A} \mathbf{r} = 2 \cdot 4 + 1 = 9$, so $t_0 = \tfrac{5}{9}$ and $\mathbf{x}^{(1)} = (\tfrac{10}{9}, \tfrac{5}{9})$. Then $\mathbf{r}^{(1)} = (2, 1) - \mathbf{A}\mathbf{x}^{(1)} = (2 - \tfrac{20}{9},\ 1 - \tfrac{5}{9}) = (-\tfrac{2}{9}, \tfrac{4}{9})$. Now $\mathbf{r}^{\top} \mathbf{r} = \tfrac{20}{81}$ and $\mathbf{r}^{\top} \mathbf{A} \mathbf{r} = 2 \cdot \tfrac{4}{81} + \tfrac{16}{81} = \tfrac{24}{81}$, so $t_1 = \tfrac{20}{24} = \tfrac{5}{6}$ and

$$\mathbf{x}^{(2)} = \left(\tfrac{10}{9} - \tfrac{5}{6} \cdot \tfrac{2}{9},\ \tfrac{5}{9} + \tfrac{5}{6} \cdot \tfrac{4}{9} \right) = \left(\tfrac{25}{27}, \tfrac{25}{27} \right).$$

(b) The errors are $(-1, -1)$, $(\tfrac{1}{9}, -\tfrac{4}{9})$, $(-\tfrac{2}{27}, -\tfrac{2}{27})$, with squared energy norms

$$3, \qquad 2 \cdot \tfrac{1}{81} + \tfrac{16}{81} = \tfrac{18}{81} = \tfrac{2}{9}, \qquad 3 \cdot \tfrac{4}{729} = \tfrac{4}{243}.$$

The ratios are $\tfrac{2}{27}$ and $\tfrac{2}{27}$, both well below the bound's $(\tfrac{1}{3})^2 = \tfrac{1}{9}$. The bound is not attained here because the initial residual does not split its weight equally between the two eigenvector directions, which is what equality in Lemma 24.7.2 demands.

Exercise. B2: Conjugate gradients by hand

Let $\mathbf{A} = \begin{pmatrix} 4 & 1 \\ 1 & 3 \end{pmatrix}$, $\mathbf{b} = (1, 2)$, and $\mathbf{x}^{(0)} = (0, 0)$. Run conjugate gradients to termination, and verify at each stage that the residuals are orthogonal and the directions $\mathbf{A}$-orthogonal.

Solution. $\det \mathbf{A} = 11$, so $\mathbf{x}^\star = \frac{1}{11}(3 \cdot 1 - 1 \cdot 2,\ -1 \cdot 1 + 4 \cdot 2) = (\frac{1}{11}, \frac{7}{11})$.

Step 0. $\mathbf{r}^{(0)} = \mathbf{p}^{(0)} = (1, 2)$; $(\mathbf{r}^{(0)})^\top \mathbf{r}^{(0)} = 5$; $\mathbf{A}\mathbf{p}^{(0)} = (6, 7)$; $(\mathbf{p}^{(0)})^\top \mathbf{A}\mathbf{p}^{(0)} = 6 + 14 = 20$. So $\alpha_0 = \frac{1}{4}$, $\mathbf{x}^{(1)} = (\frac{1}{4}, \frac{1}{2})$ and

$$\mathbf{r}^{(1)} = (1, 2) - \tfrac{1}{4}(6, 7) = (-\tfrac{1}{2}, \tfrac{1}{4}).$$

Check: $(\mathbf{r}^{(1)})^\top \mathbf{r}^{(0)} = -\frac{1}{2} + \frac{1}{2} = 0$. Then $(\mathbf{r}^{(1)})^\top \mathbf{r}^{(1)} = \frac{1}{4} + \frac{1}{16} = \frac{5}{16}$, $\beta_0 = \frac{5/16}{5} = \frac{1}{16}$ and

$$\mathbf{p}^{(1)} = (-\tfrac{1}{2}, \tfrac{1}{4}) + \tfrac{1}{16}(1, 2) = (-\tfrac{7}{16}, \tfrac{3}{8}).$$

Check: $(\mathbf{p}^{(1)})^\top \mathbf{A}\mathbf{p}^{(0)} = (-\frac{7}{16})6 + \frac{3}{8} \cdot 7 = -\frac{42}{16} + \frac{42}{16} = 0$.

Step 1. $\mathbf{A}\mathbf{p}^{(1)} = (4(-\frac{7}{16}) + \frac{3}{8},\ -\frac{7}{16} + 3 \cdot \frac{3}{8}) = (-\frac{11}{8}, \frac{11}{16})$, so

$$(\mathbf{p}^{(1)})^\top \mathbf{A}\mathbf{p}^{(1)} = \tfrac{77}{128} + \tfrac{33}{128} = \tfrac{55}{64}, \qquad \alpha_1 = \frac{5/16}{55/64} = \frac{4}{11}.$$

Then

$$\mathbf{x}^{(2)} = (\tfrac{1}{4}, \tfrac{1}{2}) + \tfrac{4}{11}(-\tfrac{7}{16}, \tfrac{3}{8}) = (\tfrac{11}{44} - \tfrac{7}{44},\ \tfrac{11}{22} + \tfrac{3}{22}) = (\tfrac{1}{11}, \tfrac{7}{11}) = \mathbf{x}^\star,$$

and

$$\mathbf{r}^{(2)} = (-\tfrac{1}{2}, \tfrac{1}{4}) - \tfrac{4}{11}(-\tfrac{11}{8}, \tfrac{11}{16}) = (-\tfrac{1}{2} + \tfrac{1}{2},\ \tfrac{1}{4} - \tfrac{1}{4}) = \mathbf{0}.$$

Two steps for a 2×2 system, which is what Corollary 24.7.5 allows.

Exercise. B3: Chebyshev by recurrence

(a) Compute T_5 from Definition 24.7.4.

(b) Evaluate $T_5(1)$, $T_5(-1)$, and $T_5(0)$, and check them against Lemma 24.7.6 (a).

(c) Evaluate $T_5(2)$ from the polynomial, and check it against the closed formula of Lemma 24.7.6 (b).

Solution.

(a) $T_2 = 2x^2 - 1$, $T_3 = 2x(2x^2 - 1) - x = 4x^3 - 3x$, $T_4 = 2x(4x^3 - 3x) - (2x^2 - 1) = 8x^4 - 8x^2 + 1$, and

$$T_5 = 2x(8x^4 - 8x^2 + 1) - (4x^3 - 3x) = 16x^5 - 20x^3 + 5x.$$

(b) $T_5(1) = 16 - 20 + 5 = 1$, $T_5(-1) = -1$, $T_5(0) = 0$. All three have absolute value at most 1, as (a) of the lemma requires, and the first is the equality $T_k(1) = 1$.

(c) $T_5(2) = 16 \cdot 32 - 20 \cdot 8 + 10 = 512 - 160 + 10 = 362$. By the closed formula with $z = 2$,

$w = 2+\sqrt{3}$ and $w^{-1} = 2-\sqrt{3}$, so

$$T_5(2) = \tfrac{1}{2}\big((2+\sqrt{3})^5 + (2-\sqrt{3})^5\big).$$

Expanding, $(2\pm\sqrt{3})^5 = 362 \pm 209\sqrt{3}$, so the half-sum is 362. The two agree, and the growth is visible: T_5 is at most 1 on $[-1,1]$ and already 362 at 2.

24.7.7.3 C. Going deeper

Exercise. C1: Sharpness of Kantorovich

Let $a \geq c > 0$ and let $\mathbf{A} = \operatorname{diag}(a, c)$.

(a) Prove that for $\mathbf{y} = (1,1)$ the inequality of Lemma 24.7.2 is an equality.

(b) Hence deduce that the constant $(\kappa-1)/(\kappa+1)$ in Theorem 24.7.3 cannot be replaced by any smaller one.

Solution.

(a) $\mathbf{y}^\top\mathbf{A}\mathbf{y} = a+c$, $\mathbf{y}^\top\mathbf{A}^{-1}\mathbf{y} = a^{-1}+c^{-1} = (a+c)/(ac)$, and $\mathbf{y}^\top\mathbf{y} = 2$. So the left side is $(a+c)^2/(ac)$ and the right side is $\big((a+c)^2/(4ac)\big)\cdot 4 = (a+c)^2/(ac)$. They agree.

(b) Take $\mathbf{b} = (a,c)$, so that $\mathbf{x}^\star = (1,1)$, and $\mathbf{x}^{(0)} = (1-\frac{1}{a},\ 1-\frac{1}{c})$, so that $\boldsymbol{\varepsilon}^{(0)} = (-\frac{1}{a}, -\frac{1}{c})$. Then $\mathbf{r}^{(0)} = -\mathbf{A}\boldsymbol{\varepsilon}^{(0)} = (1,1)$, which is the vector of part (a). By the proof of Theorem 24.7.3 the one-step ratio is

$$1 - \frac{(\mathbf{r}^\top\mathbf{r})^2}{(\mathbf{r}^\top\mathbf{A}\mathbf{r})(\mathbf{r}^\top\mathbf{A}^{-1}\mathbf{r})} = 1 - \frac{4ac}{(a+c)^2} = \Big(\frac{\kappa-1}{\kappa+1}\Big)^2,$$

so the first step attains the bound exactly. Moreover $\mathbf{r}^{(1)} = \mathbf{r}^{(0)} - t_0\mathbf{A}\mathbf{r}^{(0)} = (1-t_0 a,\ 1-t_0 c)$ with $t_0 = 2/(a+c)$, giving $\mathbf{r}^{(1)} = \big((c-a)/(a+c)\big)(1,-1)$; its two components again have equal absolute value, so part (a) applies to it up to sign and scale. That is enough: replacing $\mathbf{y}$ by $s\mathbf{y}$ multiplies both sides of Lemma 24.7.2 by s^4, and changing the sign of a coordinate changes neither side, since $\mathbf{A}$ and $\mathbf{A}^{-1}$ are diagonal. So every step attains the bound. Hence no constant smaller than $(\kappa-1)/(\kappa+1)$ can work. (Example (The bound is attained) is the case $a = 9$, $c = 1$, rescaled.)

Exercise. C2: Finite termination, from polynomials

Give a second proof of the second statement of Corollary 24.7.5 — that d distinct eigenvalues force termination in at most d steps — using Eq. 24.7.6 instead of a dimension count. *Hint: consider* $q(x) = \prod_{i=1}^{d}(1 - x/\mu_i)$.

Solution. Let $\mu_1, \dots, \mu_d$ be the distinct eigenvalues of $\mathbf{A}$; all are positive, so $q(x) = \prod_{i=1}^{d}(1 - x/\mu_i)$ is defined, has degree d, and satisfies $q(0) = 1$. It is therefore admissible in Eq. 24.7.6 with $k = d$. By Step 2 of the proof of Theorem 24.7.7,

$$\|\boldsymbol{\varepsilon}^{(d)}\|_{\mathbf{A}}^2 \leq \sum_i \lambda_i\, q(\lambda_i)^2 u_i^2 = 0,$$

since every eigenvalue λ_i of $\mathbf{A}$ equals some μ_j and so is a root of q. Hence $\boldsymbol{\varepsilon}^{(d)} = \mathbf{0}$ and $\mathbf{x}^{(d)} = \mathbf{x}^\star$. (Eq. 24.7.6 was stated for iterates before termination; if the iteration stops earlier the conclusion is immediate.)

Exercise. C3: What preconditioning does to the spectrum

Let $\mathbf{A}, \mathbf{M} \in M_n(\mathbb{R})$ be symmetric positive definite and write $\mathbf{M} = \mathbf{C}\mathbf{C}^\top$ with $\mathbf{C}$ invertible, which is possible by Theorem 13.2.3.

(a) Prove that $\mathbf{C}^{-1}\mathbf{A}\mathbf{C}^{-\top}$ is symmetric positive definite.

(b) Prove that $\mathbf{C}^{-1}\mathbf{A}\mathbf{C}^{-\top}$ and $\mathbf{M}^{-1}\mathbf{A}$ are similar, and deduce that their eigenvalues coincide.

(c) Hence show that those eigenvalues are exactly the generalized eigenvalues of the pencil $\mathbf{A} - \lambda\mathbf{M}$ in the sense of Definition 13.4.1, and say what Theorem 24.7.7 then predicts for preconditioned conjugate gradients.

Solution.

(a) It is symmetric because $(\mathbf{C}^{-1}\mathbf{A}\mathbf{C}^{-\top})^\top = \mathbf{C}^{-1}\mathbf{A}^\top\mathbf{C}^{-\top} = \mathbf{C}^{-1}\mathbf{A}\mathbf{C}^{-\top}$. It is positive definite by Proposition 13.1.2 (b) applied with $\mathbf{S} = \mathbf{C}^{-\top}$, which is invertible.

(b) Put $\mathbf{P} = \mathbf{C}^{-\top}$, which is invertible. Then

$$\mathbf{P}(\mathbf{C}^{-1}\mathbf{A}\mathbf{C}^{-\top})\mathbf{P}^{-1} = \mathbf{C}^{-\top}\mathbf{C}^{-1}\mathbf{A}\mathbf{C}^{-\top}\mathbf{C}^{\top} = \mathbf{C}^{-\top}\mathbf{C}^{-1}\mathbf{A} = (\mathbf{C}\mathbf{C}^\top)^{-1}\mathbf{A} = \mathbf{M}^{-1}\mathbf{A},$$

using $(\mathbf{C}\mathbf{C}^\top)^{-1} = \mathbf{C}^{-\top}\mathbf{C}^{-1}$. So the two matrices are similar, and similar matrices have the same characteristic polynomial (Theorem 7.10.3), hence the same eigenvalues with multiplicity.

(c) By Definition 13.4.1, λ is a generalized eigenvalue of the pencil $\mathbf{A} - \lambda\mathbf{M}$ exactly when $\mathbf{A}\mathbf{v} = \lambda\mathbf{M}\mathbf{v}$ for some $\mathbf{v} \neq \mathbf{0}$, that is when $\mathbf{M}^{-1}\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$; so the generalized eigenvalues are the eigenvalues of $\mathbf{M}^{-1}\mathbf{A}$, which by (b) are those of $\mathbf{C}^{-1}\mathbf{A}\mathbf{C}^{-\top}$. Running conjugate gradients on $(\mathbf{C}^{-1}\mathbf{A}\mathbf{C}^{-\top})\mathbf{y} = \mathbf{C}^{-1}\mathbf{b}$ therefore converges at the rate of Theorem 24.7.7 with κ replaced by the ratio of the largest to the smallest generalized eigenvalue of $\mathbf{A}$ relative to $\mathbf{M}$. Choosing $\mathbf{M}$ close to $\mathbf{A}$ makes that ratio close to 1; choosing $\mathbf{M} = \mathbf{A}$ makes it exactly 1, at the cost of having to solve with $\mathbf{A}$.

24.8 Lax Pairs and the Toda Flow

The QR algorithm of Section 5 is a sequence $\mathbf{A}^{(1)}, \mathbf{A}^{(2)}, \mathbf{A}^{(3)}, \dots$ of matrices, each similar to the last, drifting towards triangular form. Nothing in it looks like a differential equation. This section shows that it is one: there is a curve $Ł(t)$ of matrices, all with the same eigenvalues, that passes through the QR iterates at the integer times. The bridge is a class of matrix differential equations whose solutions cannot change their spectrum, and the proof that they cannot is a page of pure algebra. It is the one place in the chapter where nothing is computed and nothing is rounded.

Throughout, $I \subseteq \mathbb{R}$ is an open interval and matrix-valued functions of $t \in I$ are differentiated entrywise, as in Definition 16.6.1. **In this section the letter Ł names the matrix being**

flowed, following the universal notation for it; the graph Laplacian of Section 11 carries the same symbol and never appears here.

24.8.1 Which matrix differential equations preserve the spectrum?

Start from the answer we want and read the equation off it. Suppose $Ł(t)$ is a curve of matrices, all similar to $Ł(0)$ by a similarity that also moves with t:

$$Ł(t) = \mathbf{S}(t)\, Ł(0)\, \mathbf{S}(t)^{-1},$$

with $\mathbf{S}\colon I \to M_n(\mathbb{C})$ differentiable and everywhere invertible. Differentiate. By the product rule (Lemma 16.6.4) and the derivative of the inverse (Theorem 16.6.5),

$$\begin{aligned} Ł' &= \mathbf{S}'Ł(0)\mathbf{S}^{-1} - \mathbf{S}Ł(0)\mathbf{S}^{-1}\mathbf{S}'\mathbf{S}^{-1} \\ &= (\mathbf{S}'\mathbf{S}^{-1})Ł - Ł(\mathbf{S}'\mathbf{S}^{-1}), \end{aligned}$$

where the second line inserts $\mathbf{S}^{-1}\mathbf{S} = \mathbf{I}$ in the first term and recognizes $\mathbf{S}Ł(0)\mathbf{S}^{-1} = Ł$ in both. So with $\mathbf{B} = \mathbf{S}'\mathbf{S}^{-1}$ the curve satisfies $Ł' = \mathbf{B}Ł - Ł\mathbf{B}$. The commutator is not a clever choice: it is what a moving similarity is *forced* to satisfy.

A Lax pair is a matrix that moves, together with the infinitesimal similarity that moves it.

Definition 24.8.1. Lax Pair

Let $I \subseteq \mathbb{R}$ be an open interval. A **Lax pair** on I is a pair of functions $Ł, \mathbf{B}\colon I \to M_n(\mathbb{C})$, with $Ł$ differentiable, such that

$$Ł'(t) = [\mathbf{B}(t), Ł(t)] \coloneqq \mathbf{B}(t)Ł(t) - Ł(t)\mathbf{B}(t) \qquad \text{for every } t \in I.$$

The equation itself is called a **Lax equation**, and $Ł$ is said to **flow** under it.

Clause by clause. The two matrices are **square of the same size** n, or the bracket is not defined. Only $Ł$ is required to be **differentiable**; $\mathbf{B}$ may be any function of t whatsoever, and in the interesting cases it is built out of $Ł(t)$ itself, which makes the equation nonlinear. The bracket $[\mathbf{B}, Ł]$ is the commutator of the notation list, and the **order matters**: $[Ł, \mathbf{B}] = -[\mathbf{B}, Ł]$, so swapping the two letters runs the flow backwards.

Examples.

1. $\mathbf{B} = \mathbf{O}$**.** Then $Ł' = \mathbf{O}$, so $Ł$ is constant. The degenerate case, and it matters: it is the sanity check that the definition does not force motion.
2. **Constant** $\mathbf{B}$**.** Put $Ł(t) = e^{t\mathbf{B}}Ł_0e^{-t\mathbf{B}}$ for a fixed $Ł_0$. By Theorem 10.9.4 (b) and Lemma 16.6.4,

$$Ł'(t) = \mathbf{B}e^{t\mathbf{B}}Ł_0e^{-t\mathbf{B}} + e^{t\mathbf{B}}Ł_0(-\mathbf{B}e^{-t\mathbf{B}}) = \mathbf{B}Ł - Ł\mathbf{B},$$

the last step because $\mathbf{B}$ commutes with $e^{\pm t\mathbf{B}}$, again by Theorem 10.9.4 (b). So $(Ł, \mathbf{B})$ is a Lax pair on all of $\mathbb{R}$. Corollary 24.8.2 below says these are the only solutions when $\mathbf{B}$ is constant.

A third example is worth a block of its own, because it is the one to hold in mind.

Example. A curve of matrices with a frozen spectrum

Take $\mathbf{B} = \left(\begin{smallmatrix} 0 & 1 \\ 0 & 0 \end{smallmatrix}\right)$ and $Ł_0 = \left(\begin{smallmatrix} 0 & 0 \\ 1 & 0 \end{smallmatrix}\right)$. Since $\mathbf{B}^2 = \mathbf{O}$, Definition 10.9.2 gives $e^{t\mathbf{B}} = \mathbf{I} + t\mathbf{B}$, and the recipe of the previous example produces

$$Ł(t) = \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & -t \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} t & -t^2 \\ 1 & -t \end{pmatrix}.$$

Three of the four entries move — only the corner $\ell_{21} = 1$ stands still — and yet $\operatorname{tr} Ł(t) = 0$ and $\det Ł(t) = -t^2 + t^2 = 0$, so $p_{Ł(t)}(x) = x^2$ for every t. The curve stays inside one similarity class the whole time.

A nonexample, by dropping the second term. Replace the Lax equation by $Ł' = \mathbf{B}Ł$ with $\mathbf{B} = \mathbf{I}$. The unique solution with $Ł(0) = Ł_0$ is $Ł(t) = e^t Ł_0$ (Theorem 10.9.6, applied to each column). Its eigenvalues are $e^t\lambda_i$, which move. What was lost is precisely the second term: $\mathbf{B}Ł$ alone is not a commutator, and a commutator is what the differentiation of a similarity produced.

Warning. **Isospectral does not mean constant.** In Example (A curve of matrices with a frozen spectrum) the matrix $Ł(t)$ changes at every instant; what stays put is its list of eigenvalues. The eigen*vectors* move, and they must, since they are the columns of $\mathbf{S}(t)$ up to scale. A flow that fixed the eigenvectors as well would fix Ł. The whole use of a Lax pair — in Section 5's language, in this section's, and in the theory of integrable systems generally — is that something moves while something else does not.

24.8.2 The isospectral theorem

The derivation above suggests the proof: manufacture $\mathbf{S}(t)$ and conclude by similarity. That route runs into an obstacle the book cannot clear. The natural $\mathbf{S}$ is the solution of $\mathbf{S}' = \mathbf{B}\mathbf{S}$, $\mathbf{S}(t_0) = \mathbf{I}$, and for a **constant** $\mathbf{B}$ that is $e^{(t-t_0)\mathbf{B}}$ by Chapter 10 §09. For a $\mathbf{B}$ that depends on t — the only case of interest here — the existence of such an $\mathbf{S}$ is a theorem about linear differential equations with variable coefficients, and this book does not prove it. So we take a different route, which needs no flow at all.

Theorem 24.8.1. Lax Pairs Are Isospectral

Let $(Ł, \mathbf{B})$ be a Lax pair on an open interval I. Then the characteristic polynomial of $Ł(t)$ does not depend on t:

$$p_{Ł(t)} = p_{Ł(s)} \qquad \text{for all } s, t \in I.$$

In particular $Ł(t)$ has the same eigenvalues as $Ł(s)$, with the same algebraic multiplicities, and $\operatorname{tr} Ł(t)$ and $\det Ł(t)$ are constant.

Idea. Freeze $x \in \mathbb{C}$ and watch the single number $f(t) = \det(x\mathbf{I} - Ł(t))$. Jacobi's formula turns its derivative into a trace against the adjugate, and the adjugate is the one matrix that turns $\mathbf{M}$ into the scalar $\det \mathbf{M}$ from either side. Writing the bracket in terms of $\mathbf{M} = x\mathbf{I} - Ł$ rather than Ł — which is legitimate, because $x\mathbf{I}$ commutes with everything — makes the

two halves of the trace cancel, because a trace is unchanged by a cyclic shift. So $f' \equiv 0$, and a polynomial identity valid for every x is an identity of polynomials.

Proof. Fix $x \in \mathbb{C}$ and put $\mathbf{M}(t) = x\mathbf{I} - Ł(t)$, which is differentiable with $\mathbf{M}'(t) = -Ł'(t)$ (Definition 16.6.1). Let

$$f(t) = \det \mathbf{M}(t) = p_{Ł(t)}(x).$$

By Jacobi's formula (Theorem 16.6.6), f is differentiable on I with

$$f'(t) = \operatorname{tr}\big(\operatorname{adj} \mathbf{M}(t)\, \mathbf{M}'(t)\big).$$

Rewrite the bracket in terms of $\mathbf{M}$. Since $Ł = x\mathbf{I} - \mathbf{M}$ and $x\mathbf{I}$ commutes with $\mathbf{B}$,

$$[\mathbf{B}, Ł] = \mathbf{B}(x\mathbf{I} - \mathbf{M}) - (x\mathbf{I} - \mathbf{M})\mathbf{B} = \mathbf{MB} - \mathbf{BM},$$

so $\mathbf{M}' = -Ł' = -[\mathbf{B}, Ł] = \mathbf{BM} - \mathbf{MB}$. Writing $\mathbf{M} = \mathbf{M}(t)$ and using the linearity of the trace (Theorem 0.9.6 (1)),

$$f'(t) = \operatorname{tr}\big(\operatorname{adj} \mathbf{M}\, \mathbf{BM}\big) - \operatorname{tr}\big(\operatorname{adj} \mathbf{M}\, \mathbf{MB}\big).$$

A cyclic shift (Theorem 0.9.6 (3)) moves $\mathbf{M}$ to the front of the first trace, so

$$f'(t) = \operatorname{tr}\big(\mathbf{M} \operatorname{adj} \mathbf{M}\, \mathbf{B}\big) - \operatorname{tr}\big(\operatorname{adj} \mathbf{M}\, \mathbf{M}\, \mathbf{B}\big).$$

By Theorem 7.6.2, $\mathbf{M} \operatorname{adj} \mathbf{M} = \operatorname{adj} \mathbf{M}\, \mathbf{M} = (\det \mathbf{M})\mathbf{I}$, so the two traces are the same number and $f'(t) = 0$ for every $t \in I$.

A differentiable function on an interval with vanishing derivative is constant — the mean value theorem, fact (A6) of Chapter 16's introduction, applied to the real and imaginary parts of f. (Chapter 10 §09's fact (A4) says this for functions defined on all of $\mathbb{R}$, which is not the hypothesis here.) So $p_{Ł(t)}(x) = p_{Ł(s)}(x)$ for all $s, t \in I$.

Finally, $x \in \mathbb{C}$ was arbitrary, so the two polynomials $p_{Ł(t)}$ and $p_{Ł(s)}$ define the same function $\mathbb{C} \to \mathbb{C}$; since $\mathbb{C}$ is infinite, Theorem 6.1.5 (b) gives $p_{Ł(t)} = p_{Ł(s)}$ as polynomials. The eigenvalues are their roots and the multiplicities the multiplicities of those roots, and tr and det are two of the coefficients up to sign (Theorem 7.10.2). This proves the theorem. ∎

Remark. **What the proof did not need.** It never produced the similarity $\mathbf{S}(t)$, never solved a differential equation, and never used any property of $\mathbf{B}$ — not continuity, and not differentiability. The only analytic ingredients are two: (D1), the differentiation package Chapter 16 §06 quoted, which is what Jacobi's formula rests on, and the mean value theorem (A6) of Chapter 16's introduction, in the form "zero derivative on an interval implies constant". Everything else is the adjugate identity and the invariance of the trace under a cyclic shift.

When $\mathbf{B}$ *is* constant, the similarity is available and the theorem can be sharpened: the flow is completely determined.

Corollary 24.8.2. Constant $\mathbf{B}$: the flow is conjugation by the exponential

Let $\mathbf{B} \in M_n(\mathbb{C})$ be a fixed matrix and let $Ł\colon \mathbb{R} \to M_n(\mathbb{C})$ be differentiable with $Ł' = [\mathbf{B}, Ł]$. Then

$$Ł(t) = e^{t\mathbf{B}}\, Ł(0)\, e^{-t\mathbf{B}} \qquad (t \in \mathbb{R}),$$

so $Ł(t)$ is similar to $Ł(0)$ for every t.

Proof. Put $\mathbf{M}(t) = e^{-t\mathbf{B}} Ł(t) e^{t\mathbf{B}}$, a product of three differentiable matrix functions. By Lemma 16.6.4 and Theorem 10.9.4 (b), which also gives $\mathbf{B}e^{\pm t\mathbf{B}} = e^{\pm t\mathbf{B}}\mathbf{B}$,

$$\begin{aligned} \mathbf{M}'(t) &= -\mathbf{B}e^{-t\mathbf{B}} Ł e^{t\mathbf{B}} + e^{-t\mathbf{B}} Ł' e^{t\mathbf{B}} + e^{-t\mathbf{B}} Ł \mathbf{B} e^{t\mathbf{B}} \\ &= e^{-t\mathbf{B}}\big(-\mathbf{B}Ł + (\mathbf{B}Ł - Ł\mathbf{B}) + Ł\mathbf{B}\big)e^{t\mathbf{B}} = \mathbf{O}. \end{aligned}$$

By fact (A4) of Chapter 10 §09, every entry of $\mathbf{M}$ is constant, so $\mathbf{M}(t) = \mathbf{M}(0) = Ł(0)$. Multiplying on the left by $e^{t\mathbf{B}}$ and on the right by $e^{-t\mathbf{B}}$, and using $(e^{t\mathbf{B}})^{-1} = e^{-t\mathbf{B}}$ from Theorem 10.9.4 (c), gives the formula. This proves the corollary. ∎

Quick check. Prove directly from the Lax equation, without Theorem 24.8.1, that $\operatorname{tr} Ł(t)$ is constant.

Solution. The trace of a matrix is a linear function of its entries, so $(\operatorname{tr} Ł)'(t) = \operatorname{tr}(Ł'(t))$ by (D1) and Theorem 0.9.6 (1). Then

$$\operatorname{tr} Ł'(t) = \operatorname{tr}(\mathbf{B}Ł) - \operatorname{tr}(Ł\mathbf{B}) = 0$$

by Theorem 0.9.6 (3). So $\operatorname{tr} Ł$ has zero derivative on I and is constant by fact (A6) of Chapter 16's introduction. The same argument applied to $Ł^m$ gives $\operatorname{tr}(Ł^m)' = 0$ for every $m \geq 1$ (Exercise (C1: The power sums are constants of the motion)).

24.8.3 The Toda flow

Now for the flow that matters. Take $Ł$ real symmetric and tridiagonal, and take $\mathbf{B}$ to be the skew-symmetric matrix built from the off-diagonal entries of $Ł$. There are two sign conventions, differing by $\mathbf{B} \mapsto -\mathbf{B}$ and so running the flow in opposite directions; we fix the one that drives the largest eigenvalue towards the top-left corner, as Example (The 2×2 flow, twice) will confirm.

Definition 24.8.2. The Toda Flow

For $\mathbf{M} \in M_n(\mathbb{R})$ write $\mathbf{M}_+$ for its **strictly upper triangular part** and $\mathbf{M}_-$ for its **strictly lower triangular part**, so that $\mathbf{M}$ is the sum of $\mathbf{M}_-$, its diagonal part, and $\mathbf{M}_+$, and set

$$\mathbf{B}(\mathbf{M}) := \mathbf{M}_+ - \mathbf{M}_-.$$

The **Toda flow** is the Lax equation

$$Ł'(t) = [\mathbf{B}(Ł(t)),\ Ł(t)]$$

for a differentiable $Ł\colon I \to M_n(\mathbb{R})$ taking symmetric tridiagonal values.

Two clauses need checking. First, $\mathbf{B}(\mathbf{M})$ is **skew-symmetric** when $\mathbf{M}$ is symmetric: transposing turns the strictly upper part into the strictly lower part of $\mathbf{M}^\top = \mathbf{M}$, so $\mathbf{B}(\mathbf{M})^\top = \mathbf{M}_- - \mathbf{M}_+ = -\mathbf{B}(\mathbf{M})$. Second, the equation is **nonlinear**: $\mathbf{B}$ depends on Ł, and the right-hand side is quadratic in the entries of Ł. That is what makes it a genuine dynamical system rather than a rewriting of Corollary 24.8.2.

Writing the equation out shows what it does.

Proposition 24.8.3. The Toda Flow in Coordinates

Let $Ł \in M_n(\mathbb{R})$ be symmetric and tridiagonal, with diagonal entries $d_1, \dots, d_n$ and off-diagonal entries $e_1, \dots, e_{n-1}$, so $\ell_{i,i+1} = \ell_{i+1,i} = e_i$. Put $e_0 = e_n = 0$. Then $[\mathbf{B}(Ł), Ł]$ is again symmetric and tridiagonal, with

$$\big[\mathbf{B}(Ł), Ł\big]_{ii} = 2(e_i^2 - e_{i-1}^2), \qquad \big[\mathbf{B}(Ł), Ł\big]_{i,i+1} = e_i\,(d_{i+1} - d_i).$$

Consequently the Toda flow is the system

$$d_i' = 2(e_i^2 - e_{i-1}^2) \ \ (1 \le i \le n), \qquad e_i' = e_i\,(d_{i+1} - d_i) \ \ (1 \le i \le n-1). \tag{24.8.1}$$

Proof. Write $\mathbf{B} = \mathbf{B}(Ł)$, so that $b_{i,i+1} = e_i$, $b_{i+1,i} = -e_i$, and every other entry of $\mathbf{B}$ is 0. Then for all i, j,

$$(\mathbf{B}Ł)_{ij} = e_i\,\ell_{i+1,j} - e_{i-1}\,\ell_{i-1,j}, \qquad (Ł\mathbf{B})_{ij} = e_{j-1}\,\ell_{i,j-1} - e_j\,\ell_{i,j+1},$$

with the convention that a term with an index outside $1, \dots, n$ is 0; this is just the two nonzero entries in row i of $\mathbf{B}$, respectively in column j of $\mathbf{B}$. Write $\mathbf{C} = \mathbf{B}Ł - Ł\mathbf{B}$, so

$$c_{ij} = e_i\ell_{i+1,j} - e_{i-1}\ell_{i-1,j} - e_{j-1}\ell_{i,j-1} + e_j\ell_{i,j+1}. \tag{$*$}$$

$\mathbf{C}$ is symmetric: since $\mathbf{B}^\top = -\mathbf{B}$ and $Ł^\top = Ł$,

$$\mathbf{C}^\top = Ł^\top\mathbf{B}^\top - \mathbf{B}^\top Ł^\top = -Ł\mathbf{B} + \mathbf{B}Ł = \mathbf{C},$$

so it is enough to compute c_{ij} for $j \ge i$.

Case $j = i$. Every ℓ appearing in $(*)$ is then an off-diagonal neighbor: $\ell_{i+1,i} = e_i$, $\ell_{i-1,i} = e_{i-1}$, $\ell_{i,i-1} = e_{i-1}$, $\ell_{i,i+1} = e_i$. So

$$c_{ii} = e_i^2 - e_{i-1}^2 - e_{i-1}^2 + e_i^2 = 2(e_i^2 - e_{i-1}^2).$$

Case $j = i+1$. Now $\ell_{i+1,i+1} = d_{i+1}$ and $\ell_{i,i} = d_i$, while $\ell_{i-1,i+1} = 0$ and $\ell_{i,i+2} = 0$ because Ł is tridiagonal. So $c_{i,i+1} = e_i d_{i+1} - e_i d_i$.

Case $j = i+2$. Here $\ell_{i+1,i+2} = e_{i+1}$ and $\ell_{i,i+1} = e_i$, while $\ell_{i-1,i+2} = \ell_{i,i+3} = 0$. So

$$c_{i,i+2} = e_i e_{i+1} - 0 - e_{i+1} e_i + 0 = 0.$$

Case $j \geq i+3$. Each of the four entries $\ell_{i+1,j}$, $\ell_{i-1,j}$, $\ell_{i,j-1}$, $\ell_{i,j+1}$ has its two indices differing by at least 2, so all four are 0 and $c_{ij} = 0$.

Hence $\mathbf{C}$ is symmetric tridiagonal with the stated entries, and equating Ł′ with $\mathbf{C}$ entry by entry gives Eq. 24.8.1. This proves the proposition.

■

Read Eq. 24.8.1 for what it is. The off-diagonal entries feed the diagonal — mass e_i^2 flows from d_{i+1} up to d_i — and the diagonal drives the off-diagonal, each e_i growing or shrinking according to the sign of $d_{i+1} - d_i$. Summing the first family gives $\sum_i d_i' = 0$, the constancy of the trace, with the terms canceling in pairs. And an e_i that vanishes at some time has $e_i' = 0$ at that same time. It is tempting to read off that such an e_i stays 0 for ever — that the flow never re-couples two blocks which have already separated, the continuous shadow of deflation in Section 5 — but that is a statement about the *solutions* of the system, and getting it from the instantaneous one needs a uniqueness theorem for differential equations of the sort this section declined to assume above. Only the instantaneous statement is claimed here.

Example. The 2×2 flow, twice

Let $n = 2$, so $Ł = \left(\begin{smallmatrix} d_1 & e_1 \\ e_1 & d_2 \end{smallmatrix}\right)$ and Eq. 24.8.1 reads

$$d_1' = 2e_1^2, \qquad d_2' = -2e_1^2, \qquad e_1' = e_1(d_2 - d_1).$$

(a) Verify by hand that $\operatorname{tr} Ł$ and $\det Ł$ are constant along the flow.

(b) Solve the flow explicitly for $Ł(0) = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$.

Solution.

(a) $(\operatorname{tr} Ł)' = d_1' + d_2' = 2e_1^2 - 2e_1^2 = 0$. For the determinant,

$$(\det Ł)' = (d_1 d_2 - e_1^2)' = d_1' d_2 + d_1 d_2' - 2e_1 e_1',$$

by (D1) of Chapter 16 §06. Substituting,

$$2e_1^2 d_2 - 2d_1 e_1^2 - 2e_1^2(d_2 - d_1) = 0.$$

So both coefficients of $p_{Ł(t)}(x) = x^2 - (\operatorname{tr} Ł)x + \det Ł$ are constant, which is Theorem 24.8.1 for $n = 2$, got by hand.

(b) Put $u = e^{4t}$ and

$$d_1(t) = \frac{u-1}{u+1}, \qquad d_2(t) = -d_1(t), \qquad e_1(t) = \frac{2e^{2t}}{u+1}.$$

At $t = 0$, $u = 1$ gives $d_1 = 0$ and $e_1 = 1$, which is $Ł(0)$. Differentiating, with $u' = 4u$,

$$d_1' = \frac{4u(u+1) - (u-1)4u}{(u+1)^2} = \frac{8u}{(u+1)^2}, \qquad 2e_1^2 = \frac{8e^{4t}}{(u+1)^2} = \frac{8u}{(u+1)^2},$$

so the first equation holds; the second then holds because $d_2 = -d_1$. For the third,

$$e_1' = \frac{4e^{2t}(u+1) - 2e^{2t}\cdot 4u}{(u+1)^2} = \frac{4e^{2t}(1-u)}{(u+1)^2},$$

while $e_1(d_2 - d_1) = -2d_1e_1 = -2\cdot\frac{u-1}{u+1}\cdot\frac{2e^{2t}}{u+1}$, which is the same. Two checks: $d_1^2 + e_1^2 = ((u-1)^2 + 4u)/(u+1)^2 = 1$, so $\det Ł(t) = -1$ for every t, matching $\det Ł(0) = -1$; and as $t \to \infty$, $d_1 \to 1$ and $e_1 \to 0$, so $Ł(t) \to \operatorname{diag}(1, -1)$. The flow sorts the eigenvalues ± 1 onto the diagonal, largest first.

The last sentence of that example is the whole point of the section. Sorting the eigenvalues onto the diagonal in decreasing order, with the off-diagonal entries dying away, is exactly what the QR algorithm does. The next theorem says the resemblance is an identity.

24.8.4 The QR algorithm is the Toda flow at integer times

One piece of algebra first. The skew-symmetric matrices and the upper triangular matrices fill $M_n(\mathbb{R})$ between them, without overlap.

Lemma 24.8.4. Skew plus upper triangular, uniquely

Every $\mathbf{M} \in M_n(\mathbb{R})$ is $\mathbf{M} = \mathbf{K}+\mathbf{U}$ for exactly one skew-symmetric $\mathbf{K}$ and one upper triangular $\mathbf{U}$. If $\mathbf{M}$ is symmetric then $\mathbf{K} = \mathbf{M}_- - \mathbf{M}_+ = -\mathbf{B}(\mathbf{M})$.

Proof. *Uniqueness.* It suffices to show that a matrix $\mathbf{X}$ that is both skew-symmetric and upper triangular is $\mathbf{O}$. Its transpose $\mathbf{X}^\top = -\mathbf{X}$ is lower triangular and upper triangular at once, hence diagonal; so $\mathbf{X}$ is diagonal with $x_{ii} = -x_{ii}$, giving $\mathbf{X} = \mathbf{O}$. If $\mathbf{K}_1+\mathbf{U}_1 = \mathbf{K}_2+\mathbf{U}_2$ then $\mathbf{K}_1 - \mathbf{K}_2 = \mathbf{U}_2 - \mathbf{U}_1$ is such an $\mathbf{X}$.

Existence for symmetric $\mathbf{M}$, which is the only case we use. Put $\mathbf{K} = \mathbf{M}_- - \mathbf{M}_+$, skew-symmetric by the computation after Definition 24.8.2, and $\mathbf{U} = \mathbf{M} - \mathbf{K}$, whose strictly lower part is $\mathbf{M}_- - \mathbf{M}_- = \mathbf{O}$, so it is upper triangular. For a general $\mathbf{M}$, take $\mathbf{K} = \mathbf{M}_- - (\mathbf{M}_-)^\top$ and $\mathbf{U} = \mathbf{M} - \mathbf{K}$; the entries of $\mathbf{U}$ below the diagonal are $m_{ij} - m_{ij} = 0$ for $i > j$. This proves the lemma. ∎

Theorem 24.8.5. One QR Step Is One Unit of Toda Time

Let $Ł_0 \in M_n(\mathbb{R})$ be symmetric and let $\mathbf{A}^{(1)}, \mathbf{A}^{(2)}, \mathbf{A}^{(3)}, \ldots$ be the unshifted QR iterates of Definition 24.5.1 started at $\mathbf{A}^{(1)} = e^{Ł_0}$, with the factorizations

$$\mathbf{A}^{(k)} = \mathbf{Q}^{(k)}\mathbf{R}^{(k)}, \qquad \mathbf{A}^{(k+1)} = \mathbf{R}^{(k)}\mathbf{Q}^{(k)} \qquad (k \geq 1)$$

normalized so that each $\mathbf{R}^{(k)}$ has positive diagonal entries, and write $\mathbf{W}_k = \mathbf{Q}^{(1)}\mathbf{Q}^{(2)}\cdots\mathbf{Q}^{(k)}$, with $\mathbf{W}_0 = \mathbf{I}$. Then for every $k \geq 0$:

(a) the factorizations exist and are unique, and $\mathbf{A}^{(k+1)} = \mathbf{W}_k^{\top}\,\mathbf{A}^{(1)}\,\mathbf{W}_k$;

(b) $\mathbf{W}_k$ is the orthogonal factor in the QR factorization of $e^{k\text{Ł}_0}$;

(c) $\mathbf{A}^{(k+1)} = e^{\text{Ł}_k}$, where $\text{Ł}_k = \mathbf{W}_k^{\top}\,\text{Ł}_0\,\mathbf{W}_k$ is symmetric and has the same eigenvalues as Ł_0.

Idea. Each QR step is a similarity by the orthogonal factor, so after k steps the matrix has been conjugated by the accumulated product $\mathbf{W}_k$. Conjugating an exponential is exponentiating the conjugate, which turns the statement about $\mathbf{A}$ into a statement about Ł. The only thing left to identify is $\mathbf{W}_k$, and Section 5 has already done the work: Theorem 24.5.2 supplies the telescoping identity $(\mathbf{A}^{(1)})^k = \mathbf{W}_k\mathbf{R}^{(k)}\cdots\mathbf{R}^{(1)}$. A product of orthogonal matrices is orthogonal, a product of upper triangular matrices with positive diagonal is again one, so uniqueness of the QR factorization names $\mathbf{W}_k$ for us.

Proof. **Step 1: the factorizations exist and are unique.** By Theorem 10.9.4 (c), $\mathbf{A}^{(1)} = e^{\text{Ł}_0}$ is invertible, so its columns are independent and Theorem 11.8.1 applies, giving a unique $\mathbf{Q}^{(1)}$ with orthonormal columns — here square, hence orthogonal — and a unique upper triangular $\mathbf{R}^{(1)}$ with positive diagonal. Every $\mathbf{A}^{(k)}$ is similar to $\mathbf{A}^{(1)}$ by Proposition 24.5.1, hence invertible, so the same argument runs at every step. This positive-diagonal normalization is the choice Definition 24.5.1 deliberately left open, and it is the only ingredient here that Section 5 does not already supply.

Step 2: (a), and the telescoping identity. Every matrix here is real and each $\mathbf{Q}^{(k)}$ is orthogonal, so the conjugate transpose in Theorem 24.5.2 is a transpose. With $\mathbf{S}_k := \mathbf{R}^{(k)}\mathbf{R}^{(k-1)}\cdots\mathbf{R}^{(1)}$ and $\mathbf{S}_0 = \mathbf{I}$, that theorem — whose $\mathbf{U}_k$ and $\mathbf{T}_k$ are our $\mathbf{W}_k$ and $\mathbf{S}_k$ — gives, for every $k \geq 1$,

$$\mathbf{A}^{(k+1)} = \mathbf{W}_k^{\top}\,\mathbf{A}^{(1)}\,\mathbf{W}_k \qquad \text{and} \qquad (\mathbf{A}^{(1)})^k = \mathbf{W}_k\mathbf{S}_k,$$

and for $k = 0$ the two read $\mathbf{A}^{(1)} = \mathbf{A}^{(1)}$ and $\mathbf{I} = \mathbf{I}$, since $\mathbf{W}_0 = \mathbf{S}_0 = \mathbf{I}$. The first is (a).

Step 3: (b). Since Ł$_0$ commutes with itself, Theorem 10.9.4 (d) gives $(e^{\text{Ł}_0})^k = e^{k\text{Ł}_0}$, so $e^{k\text{Ł}_0} = \mathbf{W}_k\mathbf{S}_k$. Each $\mathbf{Q}^{(j)}$ is orthogonal, so $\mathbf{W}_k$ is orthogonal. Each $\mathbf{R}^{(j)}$ is upper triangular with positive diagonal, and a product of such matrices is upper triangular with diagonal entries the products of the diagonal entries, hence positive; so $\mathbf{S}_k$ is upper triangular with positive diagonal. By the uniqueness in Theorem 11.8.1, $\mathbf{W}_k\mathbf{S}_k$ **is** the QR factorization of $e^{k\text{Ł}_0}$.

Step 4: (c). $\mathbf{W}_k$ is orthogonal, so $\mathbf{W}_k^{\top} = \mathbf{W}_k^{-1}$, and Theorem 10.9.4 (a) gives

$$\mathbf{A}^{(k+1)} = \mathbf{W}_k^{-1}e^{\text{Ł}_0}\mathbf{W}_k = e^{\mathbf{W}_k^{-1}\text{Ł}_0\mathbf{W}_k} = e^{\text{Ł}_k}.$$

Finally $\text{Ł}_k^{\top} = \mathbf{W}_k^{\top}\text{Ł}_0^{\top}\mathbf{W}_k = \text{Ł}_k$, and Ł$_k$ is similar to Ł$_0$, so the two have the same characteristic polynomial (Theorem 7.10.3) and hence the same eigenvalues. This proves the theorem. ∎

So the QR algorithm, applied to the exponential of a symmetric matrix, is a sequence of or-

thogonal conjugations of $Ł_0$ whose k-th term is read off the QR factorization of $e^{kŁ_0}$. That last description makes sense for every real t, not only for integers, and it is the missing curve.

Remark. **The continuous statement, and the one thing in it we do not prove.** For $t \in \mathbb{R}$ let $e^{tŁ_0} = \mathbf{Q}(t)\mathbf{R}(t)$ be the QR factorization with $\mathbf{R}(t)$ of positive diagonal, and put $Ł(t) = \mathbf{Q}(t)^\top Ł_0 \mathbf{Q}(t)$. By Theorem 24.8.5 (b) and (c), $\mathbf{Q}(k) = \mathbf{W}_k$ and $Ł(k) = Ł_k$, so this curve passes through the QR iterates at the integer times. **Granting that $t \mapsto \mathbf{Q}(t)$ is differentiable** — it is, because the Gram–Schmidt formulas express its entries through arithmetic and square roots of positive quantities, but that is a fact of analysis this book has not set up — the curve satisfies the Lax equation $Ł' = [\mathbf{B}(Ł), Ł]$. Here is the derivation. Differentiating $e^{tŁ_0} = \mathbf{QR}$ with Theorem 10.9.4 (b) and Lemma 16.6.4 gives $Ł_0\mathbf{QR} = \mathbf{Q}'\mathbf{R} + \mathbf{QR}'$; multiplying by $\mathbf{R}^{-1}$ on the right and $\mathbf{Q}^\top$ on the left,

$$Ł(t) = \mathbf{Q}^\top Ł_0 \mathbf{Q} = \mathbf{Q}^\top\mathbf{Q}' + \mathbf{R}'\mathbf{R}^{-1}.$$

Now $\mathbf{Q}^\top\mathbf{Q} = \mathbf{I}$ differentiates to $(\mathbf{Q}')^\top\mathbf{Q} + \mathbf{Q}^\top\mathbf{Q}' = \mathbf{O}$, so $\mathbf{Q}^\top\mathbf{Q}'$ is skew-symmetric, while $\mathbf{R}'\mathbf{R}^{-1}$ is upper triangular. By Lemma 24.8.4 the splitting is unique, so $\mathbf{Q}^\top\mathbf{Q}' = -\mathbf{B}(Ł(t))$. Differentiating $Ł = \mathbf{Q}^\top Ł_0 \mathbf{Q}$, inserting $\mathbf{QQ}^\top = \mathbf{I}$ in each term, and using $(\mathbf{Q}')^\top\mathbf{Q} = -\mathbf{Q}^\top\mathbf{Q}'$,

$$Ł' = (\mathbf{Q}')^\top Ł_0\mathbf{Q} + \mathbf{Q}^\top Ł_0\mathbf{Q}' = -(\mathbf{Q}^\top\mathbf{Q}')Ł + Ł(\mathbf{Q}^\top\mathbf{Q}') = [\mathbf{B}(Ł), Ł].$$

That is a Lax equation, and $Ł_0$ was only assumed symmetric. It is the *Toda* flow of Definition 24.8.2 when $Ł(t)$ is in addition symmetric **tridiagonal**, which the definition demands, and it is so whenever $Ł_0$ is. Here is why. Since $\mathbf{R}(t)$ is upper triangular and invertible, the span of the first j columns of $\mathbf{Q}(t)$ is the span of the first j columns of $e^{tŁ_0}$, namely $e^{tŁ_0}\,\mathrm{span}(\mathbf{e}_1, \dots, \mathbf{e}_j)$. A tridiagonal $Ł_0$ carries $\mathrm{span}(\mathbf{e}_1, \dots, \mathbf{e}_j)$ into $\mathrm{span}(\mathbf{e}_1, \dots, \mathbf{e}_{j+1})$, and it commutes with $e^{tŁ_0}$ (Theorem 10.9.4 (b)), so it carries the span of the first j columns of $\mathbf{Q}(t)$ into the span of the first $j+1$. The i-th column of $\mathbf{Q}(t)$ is orthogonal to that span for $i > j+1$, so the (i,j) entry of $Ł(t) = \mathbf{Q}^\top Ł_0 \mathbf{Q}$ vanishes there: nothing below the subdiagonal, and symmetric, hence tridiagonal. For such an $Ł_0$ the curve is the Toda flow, Proposition 24.8.3 puts it in the coordinates Eq. 24.8.1, and, modulo that one differentiability, **the QR algorithm is the time-one map of the Toda flow**. The discrete half of the statement, Theorem 24.8.5, rests on nothing unproved; the continuous half is flagged here and nothing later depends on it.

Warning. **The interpolation is of Ł, not of A, and it needs the starting matrix to be an exponential.** Theorem 24.8.5 starts from $\mathbf{A}^{(1)} = e^{Ł_0}$, so it says nothing directly about the QR algorithm run on an arbitrary matrix. The matrices it does cover are exactly the symmetric positive definite ones. On the one hand $e^{Ł_0} = \mathbf{Q}e^{\mathbf{D}}\mathbf{Q}^\top$ is symmetric with positive eigenvalues e^{λ_i}, writing $Ł_0 = \mathbf{QDQ}^\top$ by Corollary 12.5.2 and using Theorem 10.9.4 (a) together with Corollary 10.9.5; on the other, a symmetric $\mathbf{A} \succ \mathbf{O}$ with eigenvalues $\mu_i > 0$ is $e^{Ł_0}$ for the symmetric $Ł_0 = \mathbf{Q}\,\mathrm{diag}(\log\mu_1, \dots, \log\mu_n)\mathbf{Q}^\top$, the logarithm being the one built in Chapter 18 §11 (Lemma 18.11.1). A symmetric matrix with a negative eigenvalue is therefore out of reach. The right reading is not "QR is Toda" but "QR on $e^{Ł_0}$ is Toda on $Ł_0$".

24.8.5 What this buys

Nothing in this section makes a computation faster. What it does is change what the QR algorithm *is*. As an algorithm it is a rule: factor, multiply back, repeat, and hope the subdiagonal shrinks. As a flow it is a curve in a bounded set — every entry of a symmetric matrix is at most its largest eigenvalue in absolute value, and the spectrum is frozen — and the questions one asks about it change accordingly. Why does the largest eigenvalue surface at the top left? Because $d_1' = 2e_1^2 \geq 0$ at all times. Why do the off-diagonal entries die away geometrically once the diagonal has almost settled? Because $e_i' = e_i(d_{i+1} - d_i)$, which near a diagonal limit is decay at the rate of the gap $d_i - d_{i+1}$ — the same eigenvalue gap that governs the QR algorithm's rate in Section 5. Neither reading is a proof, and neither is visible at all in the discrete language.

That is the general lesson, and it is worth naming as one of this chapter's moves: an algorithm and a differential equation can be the same object at two resolutions, and which resolution to use is a choice. Similar correspondences are known for other iterations, including the shifted QR algorithm; none of them is proved here. The subject they belong to is the theory of integrable systems, and it is the one place where this computational chapter touches pure mathematics.

24.8.6 Exercises

24.8.6.1 A. Check your understanding

Exercise. A1

(a) Define a Lax pair, and say which of the two functions is required to be differentiable.

(b) State Theorem 24.8.1, and name the two results from earlier chapters its proof uses.

(c) Determine whether the following statement is correct, and justify your answer: if $Ł' = [\mathbf{B}, Ł]$ then $Ł(t) = Ł(0)$ for all t.

(d) Explain in one sentence why the proof of Theorem 24.8.1 does not construct the conjugating matrix $\mathbf{S}(t)$.

(e) Write down the Toda equations Eq. 24.8.1 for $n = 3$, and check that the sum of the three diagonal equations is 0.

Solution.

(a) A pair $Ł, \mathbf{B}\colon I \to M_n(\mathbb{C})$ with Ł differentiable and $Ł'(t) = \mathbf{B}(t)Ł(t) - Ł(t)\mathbf{B}(t)$ for all $t \in I$. Only Ł must be differentiable.

(b) The characteristic polynomial of $Ł(t)$ is independent of t. The proof uses Jacobi's formula (Theorem 16.6.6) and the adjugate identity (Theorem 7.6.2), together with $\operatorname{tr}(\mathbf{XY}) = \operatorname{tr}(\mathbf{YX})$ (Theorem 0.9.6 (3)).

(c) Incorrect. Example (A curve of matrices with a frozen spectrum) has $Ł(t) = \left(\begin{smallmatrix} t & -t^2 \\ 1 & -t \end{smallmatrix}\right)$, which is nonconstant while its characteristic polynomial x^2 is constant. Isospectral is much weaker than constant.

(d) Because for a $\mathbf{B}$ depending on t that matrix is the solution of a linear differential equation with variable coefficients, whose existence this book does not prove; differentiating the determinant avoids needing it at all.

(e) With $e_0 = e_3 = 0$:

$$d_1' = 2e_1^2, \quad d_2' = 2(e_2^2 - e_1^2), \quad d_3' = -2e_2^2,$$

$$e_1' = e_1(d_2 - d_1), \quad e_2' = e_2(d_3 - d_2).$$

The three diagonal right-hand sides sum to $2e_1^2 + 2e_2^2 - 2e_1^2 - 2e_2^2 = 0$, which is the constancy of the trace.

24.8.6.2 B. Practice

Exercise. B1: A Lax pair verified

Let $\mathbf{B} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$ and $Ł(t) = \begin{pmatrix} t & -t^2 \\ 1 & -t \end{pmatrix}$.

(a) Verify directly that $Ł'(t) = [\mathbf{B}, Ł(t)]$.

(b) Compute $p_{Ł(t)}$ and confirm Theorem 24.8.1 in this case.

(c) Determine whether $Ł(t)$ and $Ł(0)$ are similar for every t, and exhibit the similarity.

Solution.

(a) $Ł'(t) = \left(\begin{smallmatrix} 1 & -2t \\ 0 & -1 \end{smallmatrix}\right)$. On the other side,

$$\mathbf{B}Ł = \begin{pmatrix} 1 & -t \\ 0 & 0 \end{pmatrix}, \qquad Ł\mathbf{B} = \begin{pmatrix} 0 & t \\ 0 & 1 \end{pmatrix},$$

so $[\mathbf{B}, Ł] = \left(\begin{smallmatrix} 1 & -2t \\ 0 & -1 \end{smallmatrix}\right) = Ł'(t)$.

(b) $\operatorname{tr} Ł(t) = t - t = 0$ and $\det Ł(t) = -t^2 - (-t^2) = 0$, so $p_{Ł(t)}(x) = x^2$ for every t. The spectrum is $\{0\}$ at all times.

(c) Yes. Since $\mathbf{B}$ is constant, Corollary 24.8.2 applies, and $e^{t\mathbf{B}} = \mathbf{I} + t\mathbf{B}$ because $\mathbf{B}^2 = \mathbf{O}$. So

$$Ł(t) = \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix} Ł(0) \begin{pmatrix} 1 & -t \\ 0 & 1 \end{pmatrix}, \qquad Ł(0) = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix},$$

which one checks by multiplying out. (Similarity is also forced by (b) alone here: both matrices are nonzero and nilpotent of size 2, hence both similar to $\mathbf{J}_2(0)$.)

Exercise. B2: The bracket by hand

Let

$$Ł = \begin{pmatrix} 1 & 2 & 0 \\ 2 & 3 & 1 \\ 0 & 1 & -1 \end{pmatrix}.$$

Compute $\mathbf{B}(Ł)$ and $[\mathbf{B}(Ł), Ł]$ directly, and check the answer against Proposition 24.8.3.

Solution. Here $d = (1, 3, -1)$ and $e = (2, 1)$. The strictly upper and lower parts give

$$\mathbf{B}(Ł) = \begin{pmatrix} 0 & 2 & 0 \\ -2 & 0 & 1 \\ 0 & -1 & 0 \end{pmatrix}.$$

Multiplying,

$$\mathbf{B}(Ł)Ł = \begin{pmatrix} 4 & 6 & 2 \\ -2 & -3 & -1 \\ -2 & -3 & -1 \end{pmatrix}, \qquad Ł\mathbf{B}(Ł) = \begin{pmatrix} -4 & 2 & 2 \\ -6 & 3 & 3 \\ -2 & 1 & 1 \end{pmatrix},$$

so

$$[\mathbf{B}(Ł), Ł] = \begin{pmatrix} 8 & 4 & 0 \\ 4 & -6 & -4 \\ 0 & -4 & -2 \end{pmatrix}.$$

Against Proposition 24.8.3: $2(e_1^2 - e_0^2) = 2\cdot 4 = 8$; $2(e_2^2 - e_1^2) = 2(1-4) = -6$; $2(e_3^2 - e_2^2) = -2$; $e_1(d_2 - d_1) = 2(3-1) = 4$; $e_2(d_3 - d_2) = 1(-1-3) = -4$. All five agree, the matrix is symmetric tridiagonal, and its trace is $8 - 6 - 2 = 0$, as it must be.

Exercise. B3: One QR step on an exponential

Let $Ł_0 = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$ and $\mathbf{A}^{(1)} = e^{Ł_0}$.

(a) Show that $e^{tŁ_0} = \left(\begin{smallmatrix} c(t) & s(t) \\ s(t) & c(t) \end{smallmatrix}\right)$ with $c(t) = \tfrac{1}{2}(e^t + e^{-t})$ and $s(t) = \tfrac{1}{2}(e^t - e^{-t})$.

(b) Find the QR factorization of $e^{tŁ_0}$ with positive diagonal, and hence $Ł(t) = \mathbf{Q}(t)^{\top} Ł_0 \mathbf{Q}(t)$.

(c) Compare your $Ł(1)$ with Example (The 2×2 flow, twice) (b).

Solution.

(a) $Ł_0^2 = \mathbf{I}$, so $Ł_0^{2m} = \mathbf{I}$ and $Ł_0^{2m+1} = Ł_0$. Splitting Definition 10.9.2 into even and odd terms,

$$e^{tŁ_0} = \Big(\sum_{m \ge 0} \frac{t^{2m}}{(2m)!}\Big)\mathbf{I} + \Big(\sum_{m \ge 0} \frac{t^{2m+1}}{(2m+1)!}\Big)Ł_0,$$

and the two scalar series are the even and odd halves of $e^t = \sum_j t^j/j!$, namely $c(t)$ and $s(t)$ by fact (A1) of Chapter 10 §09.

(b) Write $\rho = (c^2 + s^2)^{1/2} > 0$. Gram–Schmidt on the columns, as in Theorem 11.8.1: the first column (c, s) normalizes to $\mathbf{q}_1 = \rho^{-1}(c, s)$; the second column (s, c) has $\langle (s,c), \mathbf{q}_1 \rangle = 2cs/\rho$, and

$$(s,c) - \frac{2cs}{\rho^2}(c,s) = \frac{c^2 - s^2}{\rho^2}(-s,c) = \frac{1}{\rho^2}(-s,c),$$

because $c^2 - s^2 = 1$. That vector has length $1/\rho$, so $\mathbf{q}_2 = \rho^{-1}(-s, c)$. Hence $\mathbf{Q}(t) = \rho^{-1}\left(\begin{smallmatrix} c & -s \\ s & c \end{smallmatrix}\right)$, whose determinant is 1, and $\mathbf{R}(t) = \mathbf{Q}(t)^{\top} e^{tŁ_0}$ has positive diagonal. Then

$$Ł(t) = \mathbf{Q}^{\top} Ł_0 \mathbf{Q} = \frac{1}{\rho^2}\begin{pmatrix} 2cs & 1 \\ 1 & -2cs \end{pmatrix}.$$

Since $\rho^2 = c^2 + s^2 = \tfrac{1}{2}(e^{2t} + e^{-2t})$ and $2cs = \tfrac{1}{2}(e^{2t} - e^{-2t})$, multiplying numerator and denominator by $2e^{2t}$ gives

$$Ł(t) = \begin{pmatrix} \dfrac{e^{4t} - 1}{e^{4t} + 1} & \dfrac{2e^{2t}}{e^{4t} + 1} \\[2mm] \dfrac{2e^{2t}}{e^{4t} + 1} & -\dfrac{e^{4t} - 1}{e^{4t} + 1} \end{pmatrix}.$$

(c) This is exactly the solution written down in Example (The 2×2 flow, twice) (b). At $t = 1$ it is $d_1 = (e^4 - 1)/(e^4 + 1) \approx 0.9640$ and $e_1 = 2e^2/(e^4 + 1) \approx 0.2658$. One QR step on $e^{Ł_0}$ has moved Ł one unit of Toda time, and the off-diagonal entry has already fallen from 1 to about 0.27.

24.8.6.3 C. Going deeper

Exercise. C1: The power sums are constants of the motion

Let $(Ł, \mathbf{B})$ be a Lax pair on I.

(a) Prove that $t \mapsto \operatorname{tr}(Ł(t)^m)$ is constant for every integer $m \geq 1$. *Hint: the product rule, then Theorem 0.9.6 (3).*

(b) Deduce, for $n = 2$, that the characteristic polynomial of $Ł(t)$ is constant, without using Theorem 24.8.1.

Solution.

(a) By Lemma 16.6.4 and induction, $t \mapsto Ł(t)^m$ is differentiable with

$$(Ł^m)' = \sum_{j=0}^{m-1} Ł^j \, Ł' \, Ł^{m-1-j}.$$

The trace is linear (Theorem 0.9.6 (1)) and unchanged by a cyclic shift (Theorem 0.9.6 (3)), so each summand contributes $\operatorname{tr}(Ł^{m-1}Ł')$, and

$$(\operatorname{tr} Ł^m)' = m \operatorname{tr}(Ł^{m-1}Ł') = m \operatorname{tr}(Ł^{m-1}(\mathbf{B}Ł - Ł\mathbf{B})).$$

Now $\operatorname{tr}(Ł^{m-1}\mathbf{B}Ł) = \operatorname{tr}(ŁŁ^{m-1}\mathbf{B}) = \operatorname{tr}(Ł^m\mathbf{B})$ by a cyclic shift, and $\operatorname{tr}(Ł^{m-1}Ł\mathbf{B}) = \operatorname{tr}(Ł^m\mathbf{B})$ as well, so the bracket contributes 0. By fact (A6) of Chapter 16's introduction the function is constant on I.

(b) For $n = 2$, $p_{Ł(t)}(x) = x^2 - \tau x + \delta$ with $\tau = \operatorname{tr} Ł(t)$ and $\delta = \det Ł(t)$. By Cayley–Hamilton (Theorem 9.6.1), $Ł^2 - \tau Ł + \delta \mathbf{I} = \mathbf{O}$; taking traces gives $\operatorname{tr}(Ł^2) - \tau^2 + 2\delta = 0$, that is

$$\delta = \tfrac{1}{2}(\tau^2 - \operatorname{tr}(Ł^2)).$$

Part (a) with $m = 1$ makes τ constant and with $m = 2$ makes $\operatorname{tr}(Ł^2)$ constant, so δ is constant too. Hence $p_{Ł(t)}$ is constant.

Exercise. C2: How far B is from being determined

Let $(Ł, \mathbf{B})$ be a Lax pair on I and let $\mathbf{C}\colon I \to M_n(\mathbb{C})$ satisfy $[\mathbf{C}(t), Ł(t)] = \mathbf{O}$ for every t.

(a) Prove that $(Ł, \mathbf{B} + \mathbf{C})$ is also a Lax pair.

(b) Give an example with $\mathbf{C} \neq \mathbf{O}$, and conclude that a Lax equation does not determine $\mathbf{B}$.

(c) Prove, using Lemma 24.8.4, that there is nevertheless **exactly one** skew-symmetric $\mathbf{B}$ for which $\mathbf{B}+$ (upper triangular) equals a given symmetric Ł — that is, the $\mathbf{B}(Ł)$ of Definition 24.8.2 is the only natural candidate once "skew-symmetric" and "upper

triangular" are the two slots.

Solution.

(a) $[\mathbf{B}+\mathbf{C},Ł]=[\mathbf{B},Ł]+[\mathbf{C},Ł]=[\mathbf{B},Ł]+\mathbf{O}=Ł'$.

(b) Take $\mathbf{C}(t)=c\mathbf{I}$ for any scalar $c\neq 0$, or $\mathbf{C}(t)=q(Ł(t))$ for any polynomial q, since a polynomial in Ł commutes with Ł. So $\mathbf{B}$ can always be altered by a multiple of $\mathbf{I}$ without changing the flow, and the Lax equation determines $\mathbf{B}$ at most up to the commutant of Ł.

(c) Lemma 24.8.4 says $Ł=\mathbf{K}+\mathbf{U}$ for exactly one skew-symmetric $\mathbf{K}$ and one upper triangular $\mathbf{U}$, and computes $\mathbf{K}=Ł_{-}-Ł_{+}=-\mathbf{B}(Ł)$ for symmetric Ł. So the pair of slots (skew, upper triangular) picks out $\mathbf{B}(Ł)$ uniquely, up to the sign convention, and the remaining freedom found in (a) and (b) is freedom in a *different* decomposition, not in this one.

Exercise. C3: Symmetry constrains B

Let $(Ł,\mathbf{B})$ be a Lax pair of real matrices on I with $Ł(t)$ symmetric for every t.

(a) Prove that $[\mathbf{B}(t)+\mathbf{B}(t)^{\top},\ Ł(t)]=\mathbf{O}$ for every t.

(b) Deduce that $(Ł,\frac{1}{2}(\mathbf{B}-\mathbf{B}^{\top}))$ is also a Lax pair, so that $\mathbf{B}$ may always be taken skew-symmetric.

(c) Determine whether the Toda $\mathbf{B}(Ł)$ of Definition 24.8.2 satisfies the hypothesis of (a). Justify your answer.

Solution.

(a) Transposing $Ł'=\mathbf{B}Ł-Ł\mathbf{B}$ and using $Ł^{\top}=Ł$ — so that $(Ł')^{\top}=(Ł^{\top})'=Ł'$, differentiation being entrywise — gives

$$Ł'=(Ł')^{\top}=Ł^{\top}\mathbf{B}^{\top}-\mathbf{B}^{\top}Ł^{\top}=Ł\mathbf{B}^{\top}-\mathbf{B}^{\top}Ł=-[\mathbf{B}^{\top},Ł].$$

Adding this to $Ł'=[\mathbf{B},Ł]$ gives $2Ł'=[\mathbf{B},Ł]-[\mathbf{B}^{\top},Ł]$, while subtracting gives $\mathbf{O}=[\mathbf{B},Ł]+[\mathbf{B}^{\top},Ł]=[\mathbf{B}+\mathbf{B}^{\top},Ł]$.

(b) Put $\mathbf{C}=-\frac{1}{2}(\mathbf{B}+\mathbf{B}^{\top})$. By (a), $[\mathbf{B}+\mathbf{B}^{\top},Ł]=\mathbf{O}$, so $[\mathbf{C},Ł]=\mathbf{O}$, and Exercise (C2: How far **B** is from being determined) (a) says $(Ł,\mathbf{B}+\mathbf{C})$ is a Lax pair. But $\mathbf{B}+\mathbf{C}=\mathbf{B}-\frac{1}{2}(\mathbf{B}+\mathbf{B}^{\top})=\frac{1}{2}(\mathbf{B}-\mathbf{B}^{\top})$, which is skew-symmetric. So the skew part of $\mathbf{B}$ already drives the flow, and the symmetric part is inert.

(c) Yes, and trivially: $\mathbf{B}(Ł)$ is skew-symmetric for symmetric Ł, so $\mathbf{B}(Ł)+\mathbf{B}(Ł)^{\top}=\mathbf{O}$, which commutes with everything. Part (b) says that this is not a lucky feature of the Toda choice but the normal form of any Lax pair with symmetric Ł.

24.9 Vibrations

Chapter 13 §04 diagonalized two forms at once and then said, in one sentence, what the theorem was for: for a vibrating structure one matrix is the mass matrix, the other the stiffness matrix,

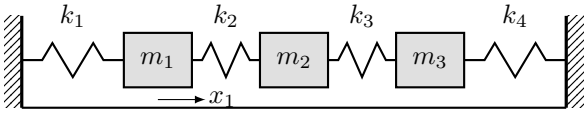
the generalized eigenvalues are the squared natural frequencies, and the eigenvectors are the normal modes. This section turns that sentence into a theorem and reads off its consequences. Everything in it is linear algebra already proved, applied to one physical model; the model is the only new ingredient, and it is stated as a model, not derived.

Throughout, the field is $\mathbb{R}$, and $n \geq 1$ is the number of moving masses.

A note on the letter K. In this section $\mathbf{K}$ is the stiffness matrix. Chapter 19 §09's second-difference matrix always carries its size as a subscript, $\mathbf{K}_n$, and it will appear below as exactly that; the two never collide.

24.9.1 A chain of masses and springs

Picture n carts on a line, joined to each other and to two fixed walls by springs. Each cart is at rest in some position, and we track only its **displacement** from that position, so that the rest state is the zero vector. The question is what the system does when it is pushed and released.



Two physical laws describe what happens, and both are assumptions, not theorems of this book. We give them a name and numbered clauses, so that the steps below can say which one they are using.

Definition 24.9.1. The mass–spring model

(P1) **(Newton's second law.)** A mass $m_i > 0$ whose displacement is $x_i(t)$ obeys $m_i x_i''(t) = f_i(t)$, where f_i is the total force on it.

(P2) **(Hooke's law.)** A spring of **stiffness** $k \geq 0$ whose two ends have moved by a and b pulls its ends together with force $k(b-a)$ acting on the end at a, and $-k(b-a)$ on the end at b. A wall is an end whose displacement is 0. A spring of stiffness 0 exerts no force whatever, so it is the same thing as no spring at all.

Hooke's law is an idealization: real springs obey it only for small displacements, and every real system also loses energy. We assume both laws exactly, and everything below is a theorem about that idealization; the energy loss reappears as the nonexample below.

Take the chain in the picture: three masses m_1, m_2, m_3 and four springs $k_1, \dots, k_4$, with k_1 joining the left wall to mass 1 and k_4 joining mass 3 to the right wall. The force on mass 1 comes from the springs on either side of it, so by (P1) and (P2) of Definition 24.9.1,

$$\begin{aligned} m_1 x_1'' &= -k_1 x_1 + k_2(x_2 - x_1), \\ m_2 x_2'' &= -k_2(x_2 - x_1) + k_3(x_3 - x_2), \\ m_3 x_3'' &= -k_3(x_3 - x_2) - k_4 x_3. \end{aligned}$$

Collecting the coefficients of x_1, x_2, x_3 on the right and moving the sign out front, this is $\mathbf{M}\mathbf{x}'' = -\mathbf{K}\mathbf{x}$ with

$$\mathbf{M} = \begin{pmatrix} m_1 & & \\ & m_2 & \\ & & m_3 \end{pmatrix}, \quad \mathbf{K} = \begin{pmatrix} k_1+k_2 & -k_2 & 0 \\ -k_2 & k_2+k_3 & -k_3 \\ 0 & -k_3 & k_3+k_4 \end{pmatrix}.$$

The matrix $\mathbf{M}$ is diagonal with positive entries, hence positive definite. The matrix $\mathbf{K}$ is symmetric, and it is positive semidefinite for a reason worth isolating, so we record the general shape first.

A mass–spring system is a pair of forms: one that measures inertia and one that measures the price of a deformation.

Definition 24.9.2. Mass–spring system

A **mass–spring system** with n degrees of freedom is a pair of real symmetric matrices $\mathbf{M}, \mathbf{K} \in M_n(\mathbb{R})$ with $\mathbf{M} \succ \mathbf{O}$ and $\mathbf{K} \succeq \mathbf{O}$, called the **mass matrix** and the **stiffness matrix**. A **motion** of the system is a twice-differentiable function $\mathbf{x}\colon \mathbb{R} \to \mathbb{R}^n$ with

$$\mathbf{M}\mathbf{x}''(t) = -\mathbf{K}\mathbf{x}(t) \qquad \text{for every } t \in \mathbb{R}. \tag{24.9.1}$$

The same equation is read for a function $\mathbf{x}\colon \mathbb{R} \to \mathbb{C}^n$ whenever a complex trial solution is convenient; a **motion** without qualification is real.

In words: $\mathbf{M} \succ \mathbf{O}$ says every direction of motion costs some inertia, so no coordinate is massless; $\mathbf{K} \succeq \mathbf{O}$ says no deformation releases energy, so the rest state $\mathbf{x} = \mathbf{0}$ is not a hilltop. The quantities $\tfrac{1}{2}(\mathbf{x}')^{\top}\mathbf{M}\mathbf{x}'$ and $\tfrac{1}{2}\mathbf{x}^{\top}\mathbf{K}\mathbf{x}$ are what the model calls the kinetic and potential energy of the state $(\mathbf{x}, \mathbf{x}')$ — a third physical identification, on the same footing as (P1) and (P2) of Definition 24.9.1 and used nowhere in a proof — and the two definiteness hypotheses say exactly that the first is positive and the second nonnegative away from rest.

For the chain, positive semidefiniteness of $\mathbf{K}$ is a factorization. Number the springs $1, \dots, n+1$ from left to right and let $\mathbf{B} \in M_{(n+1)\times n}(\mathbb{R})$ be the matrix whose row e records how much spring e is stretched: for the chain above,

$$\mathbf{B} = \begin{pmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & -1 & 1 \\ 0 & 0 & -1 \end{pmatrix}, \qquad \mathbf{K} = \mathbf{B}^{\top}\mathbf{D}_k\mathbf{B},$$

where $\mathbf{D}_k = \operatorname{diag}(k_1, \dots, k_{n+1})$. Multiplying out confirms the second identity entry by entry. Consequently

$$\mathbf{x}^{\top}\mathbf{K}\mathbf{x} = (\mathbf{B}\mathbf{x})^{\top}\mathbf{D}_k(\mathbf{B}\mathbf{x}) = \sum_{e} k_e\,(\mathbf{B}\mathbf{x})_e^2 \;\geq\; 0,$$

a sum of squares weighted by the stiffnesses: the potential energy stored in the springs. Together with $\mathbf{K}^{\top} = \mathbf{K}$ this is Definition 13.1.1, so $\mathbf{K} \succeq \mathbf{O}$ with no computation of eigenvalues. (The matrix $\mathbf{B}$ is, up to a transpose and the deletion of the two wall rows, the incidence matrix of Chapter 3 §07, Definition 3.7.1, of the chain together with a choice of orientation for each spring: the orientation is what fixes the sign pattern of the rows, and it is what Definition 3.7.1 records as the head and the tail of an edge. The last section of this chapter takes that observation seriously.)

Three instances, simplest first.

Example. Three systems

(a) $n = 1$: one mass m between two walls with springs k_1, k_2. Then $\mathbf{M} = (m)$ and $\mathbf{K} = (k_1 + k_2)$, both 1×1.

(b) $n = 3$ with all masses equal to m and all four stiffnesses equal to k. Then $\mathbf{M} = m\mathbf{I}_3$ and $\mathbf{K} = k\mathbf{K}_3$, where $\mathbf{K}_3$ is the second-difference matrix met in Proposition 19.9.6.

(c) The **free** chain: the same three masses joined to each other by k_2, k_3 but with **no walls**, so $k_1 = k_4 = 0$. Then $\mathbf{K}$ is still symmetric and still $\mathbf{B}^\top \mathbf{D}_k \mathbf{B}$, hence $\succeq \mathbf{O}$, but now $\mathbf{K}\mathbf{1} = \mathbf{0}$: sliding the whole chain sideways stretches nothing.

Case (c) is the degenerate one, and it is not a pathology to be excluded: it is why the definition asks only for $\mathbf{K} \succeq \mathbf{O}$. A system with a direction of zero stiffness has a motion that is not a vibration at all, and the theorem below will produce it.

Change one clause and the definition stops describing this situation. If a cart also feels friction proportional to its velocity, the equation becomes $\mathbf{M}\mathbf{x}'' = -\mathbf{K}\mathbf{x} - \mathbf{C}\mathbf{x}'$ with $\mathbf{C} \succeq \mathbf{O}$ the damping matrix; the right-hand side is no longer a function of $\mathbf{x}$ alone, Eq. 24.9.1 fails, and the analysis of this section does not apply to it. The models differ in exactly one term, and everything below uses the absence of that term.

Warning. **The natural frequencies are not the eigenvalues of $\mathbf{K}$ divided by the eigenvalues of $\mathbf{M}$.** Take $\mathbf{M} = \operatorname{diag}(1, 4)$ and $\mathbf{K} = \left(\begin{smallmatrix} 2 & -1 \\ -1 & 2 \end{smallmatrix}\right)$. The eigenvalues of $\mathbf{K}$ are 3 and 1, those of $\mathbf{M}$ are 4 and 1, so the four available quotients are $3, 3/4, 1$, and $1/4$. The squared natural frequencies, computed from $\det(\mathbf{K} - \lambda\mathbf{M}) = 4\lambda^2 - 10\lambda + 3$, are

$$\lambda = \tfrac{5 \pm \sqrt{13}}{4} \approx 2.151,\ 0.349,$$

and neither is on the list. Mass and stiffness interact through their eigenvectors, not only through their eigenvalues; only when $\mathbf{M}$ is a scalar multiple of $\mathbf{I}$ does the naive rule hold.

24.9.2 Normal modes

A motion of one mass on one spring is a single oscillation. The hope is that a motion of n masses is a superposition of n such oscillations, each with its own frequency and its own shape. That hope is exactly Theorem 13.4.1, and the only thing needed before it is the one-dimensional case.

Lemma 24.9.1. The scalar oscillator

Let $\omega \geq 0$ and let $y\colon \mathbb{R} \to \mathbb{C}$ be twice differentiable with $y'' = -\omega^2 y$. If $\omega > 0$, then

$$y(t) = y(0)\cos\omega t + \frac{y'(0)}{\omega}\sin\omega t$$

for every t; if $\omega = 0$, then $y(t) = y(0) + y'(0)t$. Conversely every function of these two shapes satisfies $y'' = -\omega^2 y$.

Idea. A second-order scalar equation is a first-order system in the variable (y, y'), which is the translation of Example (A Second-Order Equation with a Repeated Root). Once it is a system, Theorem 10.9.6 supplies both existence and, what we actually want, uniqueness: two solutions with the same value and the same derivative at 0 are the same function.

Proof. Put $\mathbf{N} = \left(\begin{smallmatrix} 0 & 1 \\ -\omega^2 & 0 \end{smallmatrix}\right)$ and $\mathbf{u} = (y, y')$. Then $\mathbf{u}' = (y', y'') = (y', -\omega^2 y) = \mathbf{N}\mathbf{u}$, and conversely a differentiable $\mathbf{u} = (u_1, u_2)$ with $\mathbf{u}' = \mathbf{N}\mathbf{u}$ has $u_2 = u_1'$ and $u_1'' = -\omega^2 u_1$. By Theorem 10.9.6 there is exactly one differentiable $\mathbf{u}\colon \mathbb{R} \to \mathbb{C}^2$ with $\mathbf{u}' = \mathbf{N}\mathbf{u}$ and a given value at 0.

For the converse half, $(e^{zt})' = ze^{zt}$ for $z \in \mathbb{C}$, which is the 1×1 case of Theorem 10.9.4 (b); applying it twice, $e^{\pm i\omega t}$ satisfies $y'' = -\omega^2 y$. Euler's formula — fact (A1) of Chapter 10 §09 — writes $\cos\omega t$ and $\sin\omega t$ as the combinations $\frac{1}{2}(e^{i\omega t} + e^{-i\omega t})$ and $\frac{1}{2i}(e^{i\omega t} - e^{-i\omega t})$, so both satisfy it too, as does every linear combination of them; and for $\omega = 0$ the functions 1 and t satisfy $y'' = 0$.

Now let y be any solution and let z be the function displayed in the statement. By the previous paragraph z is a solution. Differentiating, $z(0) = y(0)$ and $z'(0) = y'(0)$ in both cases, the case $\omega > 0$ using $(\sin\omega t)' = \omega\cos\omega t$ at $t = 0$. So (y, y') and (z, z') solve $\mathbf{u}' = \mathbf{N}\mathbf{u}$ with the same value at 0, and uniqueness gives $y = z$. This proves the lemma. ∎

Now the main theorem. Read the substitution in part (a) as a question: is there a motion in which every mass moves with the *same* time dependence, so that the shape of the deformation never changes and only its amplitude does?

Theorem 24.9.2. Normal Modes

Let $(\mathbf{M}, \mathbf{K})$ be a mass–spring system with n degrees of freedom (Definition 24.9.2).

(a) For $\omega \in \mathbb{R}$ and $\mathbf{v} \in \mathbb{R}^n$ with $\mathbf{v} \neq \mathbf{0}$, the complex-valued function $\mathbf{x}(t) = e^{i\omega t}\mathbf{v}$ satisfies Eq. 24.9.1 if and only if

$$\mathbf{K}\mathbf{v} = \omega^2 \mathbf{M}\mathbf{v}.$$

(b) There are real numbers $0 \le \omega_1 \le \cdots \le \omega_n$ and vectors $\mathbf{v}_1, \dots, \mathbf{v}_n \in \mathbb{R}^n$ with

$$\mathbf{K}\mathbf{v}_j = \omega_j^2 \mathbf{M}\mathbf{v}_j \quad \text{and} \quad \mathbf{v}_i^{\top}\mathbf{M}\mathbf{v}_j = \delta_{ij},$$

and $(\mathbf{v}_1, \dots, \mathbf{v}_n)$ is a basis of $\mathbb{R}^n$.

(c) Every motion of the system is

$$\mathbf{x}(t) = \sum_{j=1}^{n} c_j(t)\,\mathbf{v}_j,$$

where

$$c_j(t) = \begin{cases} a_j \cos\omega_j t + b_j \sin\omega_j t, & \omega_j > 0, \\ a_j + b_j t, & \omega_j = 0, \end{cases}$$

for uniquely determined real a_j, b_j; conversely every such function is a motion. The coefficients are $a_j = \mathbf{v}_j^{\top}\mathbf{M}\mathbf{x}(0)$, and $b_j = \mathbf{v}_j^{\top}\mathbf{M}\mathbf{x}'(0)/\omega_j$ when $\omega_j > 0$, $b_j = \mathbf{v}_j^{\top}\mathbf{M}\mathbf{x}'(0)$ when $\omega_j = 0$.

The numbers ω_j do not depend on the construction in (b): by Theorem 13.4.2 (b) the ω_j^2 are exactly the roots of $\det(\mathbf{K} - \lambda\mathbf{M})$, with multiplicity. They are the **natural frequencies**, and the pairs $(\omega_j, \mathbf{v}_j)$ are the **normal modes** of the system.

Idea. Part (a) is one differentiation. Part (b) is Chapter 13 §04 verbatim: $\mathbf{M} \succ \mathbf{O}$ is the hypothesis of Theorem 13.4.1, and the columns of the congruence it produces are $\mathbf{M}$-orthonormal generalized eigenvectors by Theorem 13.4.2; all that has to be added is that the generalized eigenvalues are ≥ 0, which is where $\mathbf{K} \succeq \mathbf{O}$ is spent. Part (c) then costs nothing: expanding $\mathbf{x}$ in the basis $(\mathbf{v}_j)$ turns Eq. 24.9.1 into n *separate* scalar equations, one per mode, and Lemma 24.9.1 solves each.

Proof.

(a) Differentiating twice, $\mathbf{x}''(t) = -\omega^2 e^{i\omega t}\mathbf{v}$, so

$$\mathbf{M}\mathbf{x}''(t) + \mathbf{K}\mathbf{x}(t) = e^{i\omega t}(\mathbf{K}\mathbf{v} - \omega^2\mathbf{M}\mathbf{v}).$$

Since $e^{i\omega t} \neq 0$ for every t, the left side vanishes for every t if and only if $\mathbf{K}\mathbf{v} = \omega^2\mathbf{M}\mathbf{v}$.

(b) The matrices $\mathbf{M}$ and $\mathbf{K}$ are real symmetric and $\mathbf{M} \succ \mathbf{O}$, so Theorem 13.4.1 gives a real invertible $\mathbf{S}$ with $\mathbf{S}^\top\mathbf{M}\mathbf{S} = \mathbf{I}_n$ and $\mathbf{S}^\top\mathbf{K}\mathbf{S} = \mathbf{D} = \operatorname{diag}(d_1, \dots, d_n)$, the d_j real. Let $\mathbf{s}_1, \dots, \mathbf{s}_n$ be the columns of $\mathbf{S}$. By Theorem 13.4.2 (a), $\mathbf{K}\mathbf{s}_j = d_j\mathbf{M}\mathbf{s}_j$ and $\mathbf{s}_i^\top\mathbf{M}\mathbf{s}_j = \delta_{ij}$. Each d_j is nonnegative: the (j,j) entry of $\mathbf{S}^\top\mathbf{K}\mathbf{S}$ is $d_j = \mathbf{s}_j^\top\mathbf{K}\mathbf{s}_j \geq 0$, because $\mathbf{K} \succeq \mathbf{O}$. Put $\omega_j = \sqrt{d_j} \geq 0$, and relabel the columns so that $\omega_1 \leq \cdots \leq \omega_n$; relabeling preserves both displayed properties. The list $(\mathbf{v}_1, \dots, \mathbf{v}_n)$ obtained is the list of columns of an invertible matrix, hence a basis of $\mathbb{R}^n$.

(c) *Every such function is a motion.* Let c_j be as displayed. By Lemma 24.9.1, $c_j'' = -\omega_j^2 c_j$. Hence, differentiating the sum termwise,

$$\mathbf{M}\mathbf{x}'' = \sum_j c_j''\,\mathbf{M}\mathbf{v}_j = -\sum_j \omega_j^2 c_j\,\mathbf{M}\mathbf{v}_j = -\sum_j c_j\,\mathbf{K}\mathbf{v}_j = -\mathbf{K}\mathbf{x},$$

where the third equality is (b).

Every motion is such a function. Let $\mathbf{x}$ be a motion. Since $(\mathbf{v}_j)$ is a basis, write $\mathbf{x}(t) = \sum_j u_j(t)\mathbf{v}_j$. Pairing with $\mathbf{M}\mathbf{v}_j$ and using (b) gives $u_j(t) = \mathbf{v}_j^\top\mathbf{M}\mathbf{x}(t)$, a fixed linear combination of the entries of $\mathbf{x}$, so each u_j is twice differentiable. Substituting into Eq. 24.9.1 and using (b),

$$\sum_j (u_j'' + \omega_j^2 u_j)\mathbf{M}\mathbf{v}_j = \mathbf{0}.$$

The vectors $\mathbf{M}\mathbf{v}_1, \dots, \mathbf{M}\mathbf{v}_n$ are independent, since $\mathbf{M}$ is invertible and the $\mathbf{v}_j$ are independent, so every coefficient vanishes: $u_j'' = -\omega_j^2 u_j$ for each j. By Lemma 24.9.1 each u_j has the displayed shape, with $a_j = u_j(0)$ and $b_j = u_j'(0)/\omega_j$ or $u_j'(0)$.

The coefficients. Pairing $\mathbf{x}(0) = \sum_i a_i\mathbf{v}_i$ with $\mathbf{M}\mathbf{v}_j$ and using $\mathbf{v}_j^\top\mathbf{M}\mathbf{v}_i = \delta_{ij}$ gives $\mathbf{v}_j^\top\mathbf{M}\mathbf{x}(0) = a_j$; the same computation on $\mathbf{x}'(0) = \sum_i u_i'(0)\mathbf{v}_i$ gives $\mathbf{v}_j^\top\mathbf{M}\mathbf{x}'(0) = u_j'(0)$. Uniqueness of a_j, b_j follows, since they are determined by $\mathbf{x}(0)$ and $\mathbf{x}'(0)$. This proves the theorem.

■

So the general motion has $2n$ real parameters, and the modes evolve independently. Mode j is a motion in which every mass oscillates at the single frequency ω_j, with relative displacements fixed by $\mathbf{v}_j$; the sign pattern of $\mathbf{v}_j$ says which masses move together and which move against each other.

Quick check. Why is the normalization $\mathbf{v}_i^\top \mathbf{M}\mathbf{v}_j = \delta_{ij}$, and not $\mathbf{v}_i^\top \mathbf{v}_j = \delta_{ij}$? Which step of the proof would fail?

Solution. The step reading off the coefficients. What is available is $\mathbf{S}^\top \mathbf{M}\mathbf{S} = \mathbf{I}$, which says the $\mathbf{v}_j$ are orthonormal for the inner product $\langle \mathbf{x}, \mathbf{y}\rangle_{\mathbf{M}} = \mathbf{y}^\top \mathbf{M}\mathbf{x}$ of Theorem 13.4.2 (a), not for the standard one. Generalized eigenvectors for *different* ω_j need not be orthogonal in the standard sense, and a genuine mass–spring system already shows it: for $\mathbf{M} = \operatorname{diag}(2,1)$ and $\mathbf{K} = \left(\begin{smallmatrix} 3 & -1 \\ -1 & 1 \end{smallmatrix}\right)$ the modes are the multiples of $(1,2)$, for $\omega^2 = \frac{1}{2}$, and of $(1,-1)$, for $\omega^2 = 2$; and $(1,2)^\top \mathbf{M}(1,-1) = 0$ while $(1,2)\cdot(1,-1) = -1 \neq 0$. So here the two normalizations cannot both be imposed: rescaling the $\mathbf{v}_j$ to standard unit length keeps them generalized eigenvectors but destroys $\mathbf{v}_i^\top \mathbf{M}\mathbf{v}_j = \delta_{ij}$, and the pairing step then has no δ_{ij} to use. (When $\mathbf{M} = c\mathbf{I}$ the two conditions differ by a constant only, and both can be met at once.)

Example. Three equal masses, four equal springs

Take $n = 3$ with every mass m and every stiffness k, so that $\mathbf{M} = m\mathbf{I}_3$ and $\mathbf{K} = k\mathbf{K}_3$ as in Example (Three systems) (b). Find the natural frequencies and the normal modes of Theorem 24.9.2.

Solution. *The pencil.* Here $\mathbf{K}\mathbf{v} = \omega^2 \mathbf{M}\mathbf{v}$ reads $k\mathbf{K}_3\mathbf{v} = m\omega^2\mathbf{v}$, an ordinary eigenvalue problem for the symmetric matrix

$$\mathbf{K}_3 = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix},$$

with $\omega^2 = (k/m)\mu$ for μ an eigenvalue of $\mathbf{K}_3$.

Eigenvalues. Expanding along the first row, $\det(\mathbf{K}_3 - \mu\mathbf{I}) = (2-\mu)^3 - 2(2-\mu) = (2-\mu)\big((2-\mu)^2 - 2\big)$, so

$$\mu = 2 - \sqrt{2}, \qquad \mu = 2, \qquad \mu = 2 + \sqrt{2}.$$

Eigenvectors. For $\mu = 2$ the first row of $\mathbf{K}_3 - 2\mathbf{I}$ gives $-v_2 = 0$ and the second gives $-v_1 - v_3 = 0$, so $\mathbf{v} \propto (1,0,-1)$. For $\mu = 2 \mp \sqrt{2}$, so that $2 - \mu = \pm\sqrt{2}$, the first row reads $(2-\mu)v_1 - v_2 = 0$ and the third reads $-v_2 + (2-\mu)v_3 = 0$, giving $v_2 = \pm\sqrt{2}\,v_1$ and $v_2 = \pm\sqrt{2}\,v_3$, so $\mathbf{v} \propto (1, \pm\sqrt{2}, 1)$. The middle row is then automatic: it reads $-1 + (2-\mu)(\pm\sqrt{2}) - 1 = -2 + 2 = 0$. Direct substitution confirms all three.

Normalizing. Since $\mathbf{M} = m\mathbf{I}_3$, the condition $\mathbf{v}^\top \mathbf{M}\mathbf{v} = 1$ is $m\|\mathbf{v}\|^2 = 1$. The squared lengths are $4, 2, 4$, so

$$\mathbf{v}_1 = \tfrac{1}{2\sqrt{m}} \begin{pmatrix} 1 \\ \sqrt{2} \\ 1 \end{pmatrix}, \quad \mathbf{v}_2 = \tfrac{1}{\sqrt{2m}} \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}, \quad \mathbf{v}_3 = \tfrac{1}{2\sqrt{m}} \begin{pmatrix} 1 \\ -\sqrt{2} \\ 1 \end{pmatrix},$$

and the natural frequencies are

$$\omega_1 = \sqrt{\tfrac{k}{m}}\sqrt{2-\sqrt{2}}, \quad \omega_2 = \sqrt{\tfrac{k}{m}}\sqrt{2}, \quad \omega_3 = \sqrt{\tfrac{k}{m}}\sqrt{2+\sqrt{2}},$$

approximately $0.765, 1.414$ and 1.848 times $\sqrt{k/m}$.

Reading the modes. In mode 1 all three masses move the same way, the middle one furthest: the slowest and softest motion. In mode 2 the middle mass stands still and the outer two move oppositely. In mode 3 neighbors always move oppositely, which stretches every spring at once and costs the most: the fastest motion. The number of sign changes along the chain increases with the frequency. That is a general phenomenon for such chains, an oscillation theorem for tridiagonal stiffness matrices; nothing in this section uses it, and this book does not prove it.

24.9.3 Frequencies as a minimum

Theorem 24.9.2 computes the frequencies exactly, at the price of solving an n-th degree equation. Often one wants instead a cheap estimate of the lowest frequency, and a way to say how it responds to a change in the design. Both come from a quotient.

Definition 24.9.3. Generalized Rayleigh quotient

Let $(\mathbf{M}, \mathbf{K})$ be a mass–spring system. Its **generalized Rayleigh quotient** is

$$\rho(\mathbf{v}) := \frac{\mathbf{v}^{\top}\mathbf{K}\mathbf{v}}{\mathbf{v}^{\top}\mathbf{M}\mathbf{v}},$$

defined for every $\mathbf{v} \in \mathbb{R}^n$ with $\mathbf{v} \neq \mathbf{0}$, and at no other vector.

The denominator is nonzero exactly because $\mathbf{M} \succ \mathbf{O}$, which is the role definiteness played in Definition 17.1.1; for $\mathbf{M} = \mathbf{I}$ the two definitions agree. Physically $\rho(\mathbf{v})$ is the potential energy of the deformation $\mathbf{v}$ divided by the kinetic energy of the velocity $\mathbf{v}$: the stiffness-to-inertia ratio of the shape $\mathbf{v}$. It is unchanged by scaling, since both numerator and denominator pick up c^2.

Corollary 24.9.3. Frequencies from the Rayleigh Quotient

Let $(\mathbf{M}, \mathbf{K})$ be a mass–spring system with natural frequencies $\omega_1 \le \cdots \le \omega_n$. Then for every j,

$$\begin{aligned}\omega_j^2 &= \min_{\dim W = j} \; \max_{\mathbf{0} \neq \mathbf{v} \in W} \; \rho(\mathbf{v}) \\ &= \max_{\dim W = n-j+1} \; \min_{\mathbf{0} \neq \mathbf{v} \in W} \; \rho(\mathbf{v}),\end{aligned}$$

W running over all subspaces of $\mathbb{R}^n$ of the stated dimension. In particular

$$\omega_1^2 = \min_{\mathbf{v} \neq \mathbf{0}} \rho(\mathbf{v}), \qquad \omega_n^2 = \max_{\mathbf{v} \neq \mathbf{0}} \rho(\mathbf{v}),$$

and every $\mathbf{v} \neq \mathbf{0}$ gives the upper bound $\omega_1^2 \le \rho(\mathbf{v})$, with equality exactly when $\mathbf{v}$ is a generalized eigenvector for ω_1.

Idea. The congruence $\mathbf{S}$ of Theorem 24.9.2 turns ρ into the ordinary Rayleigh quotient of the diagonal matrix $\mathbf{D}$: substituting $\mathbf{v} = \mathbf{S}\mathbf{y}$ replaces the numerator by $\mathbf{y}^{\top}\mathbf{D}\mathbf{y}$ and the denominator by $\mathbf{y}^{\top}\mathbf{y}$. Since $\mathbf{S}$ is invertible it matches subspaces of dimension j with subspaces of dimension j, so an optimization over subspaces is carried across unchanged, and Theorem 17.2.3 applies. The only care needed is the index: Chapter 17 orders eigenvalues **decreasingly** and frequencies are ordered **increasingly**, so $\omega_j^2 = \lambda_{n-j+1}(\mathbf{D})$.

Proof. Let $\mathbf{S}$ and $\mathbf{D} = \operatorname{diag}(d_1, \dots, d_n)$ be as in the proof of Theorem 24.9.2, so $\mathbf{S}^{\top}\mathbf{M}\mathbf{S} = \mathbf{I}_n$, $\mathbf{S}^{\top}\mathbf{K}\mathbf{S} = \mathbf{D}$, and the multiset $\{d_1, \dots, d_n\}$ is $\{\omega_1^2, \dots, \omega_n^2\}$. For $\mathbf{y} \neq \mathbf{0}$ put $\mathbf{v} = \mathbf{S}\mathbf{y}$, which is nonzero because $\mathbf{S}$ is invertible. Then

$$\rho(\mathbf{S}\mathbf{y}) = \frac{\mathbf{y}^{\top}\mathbf{S}^{\top}\mathbf{K}\mathbf{S}\mathbf{y}}{\mathbf{y}^{\top}\mathbf{S}^{\top}\mathbf{M}\mathbf{S}\mathbf{y}} = \frac{\mathbf{y}^{\top}\mathbf{D}\mathbf{y}}{\mathbf{y}^{\top}\mathbf{y}} = R_{\mathbf{D}}(\mathbf{y}),$$

the last step by Definition 17.1.1, $\mathbf{D}$ being real diagonal and hence Hermitian.

Fix j and put $k = n - j + 1$. Sorting $d_1, \dots, d_n$ decreasingly as in Chapter 17 makes $\lambda_k(\mathbf{D})$ the j-th smallest of them, that is $\lambda_k(\mathbf{D}) = \omega_j^2$. By Theorem 17.2.3,

$$\omega_j^2 = \lambda_k(\mathbf{D}) = \min_{\dim W' = n-k+1} \; \max_{\mathbf{0} \neq \mathbf{y} \in W'} R_{\mathbf{D}}(\mathbf{y}),$$

and $n - k + 1 = j$. The map $W' \mapsto \mathbf{S}W'$ is a bijection from the j-dimensional subspaces of $\mathbb{R}^n$ to themselves, since $\mathbf{S}$ is invertible, and it carries the nonzero vectors of W' onto those of $\mathbf{S}W'$; combined with the displayed identity $\rho(\mathbf{S}\mathbf{y}) = R_{\mathbf{D}}(\mathbf{y})$, the inner maximum over W' for $R_{\mathbf{D}}$ equals the inner maximum over $\mathbf{S}W'$ for ρ. Hence the two outer minima agree, which is the first line. The second line is the second equality of Theorem 17.2.3 transported in the same way, with $n - k + 1$ replaced by $k = n - j + 1$.

For the particular cases, take $j = 1$ and $j = n$ in the second and the first line respectively, where the only admissible W is $\mathbb{R}^n$ itself. Equality in $\omega_1^2 \leq \rho(\mathbf{v})$ holds exactly when $\mathbf{y} = \mathbf{S}^{-1}\mathbf{v}$ attains the minimum of $R_{\mathbf{D}}$, and that happens exactly at the eigenvectors of $\mathbf{D}$ for its smallest eigenvalue. In one direction, such an eigenvector attains the minimum by Proposition 17.1.1. Conversely, suppose $\mathbf{y} \neq \mathbf{0}$ attains it; replacing $\mathbf{y}$ by $\mathbf{y}/\|\mathbf{y}\|$ changes nothing, since $R_{\mathbf{D}}(c\mathbf{y}) = R_{\mathbf{D}}(\mathbf{y})$ for $c \neq 0$ directly from Definition 17.1.1, so we may take $\mathbf{y}$ to be a unit vector. Then Proposition 17.1.2 (b) applies in its minimum form and gives the condition of its part (a), whence $\mathbf{D}\mathbf{y} = R_{\mathbf{D}}(\mathbf{y})\mathbf{y}$ by that part, and $R_{\mathbf{D}}(\mathbf{y}) = \lambda_n(\mathbf{D})$ is the smallest eigenvalue by Proposition 17.1.1. Multiplying by $\mathbf{S}$ turns these into the generalized eigenvectors for ω_1, by Theorem 13.4.2 (a). This proves the corollary.

■

The practical content is the last sentence: **any** shape you can guess gives an upper bound for the lowest squared frequency, and the bound is good even when the guess is mediocre, because the quotient is flat at a minimum.

Example. Guessing the lowest mode of the chain

For the chain of Example (Three equal masses, four equal springs) with $k = m = 1$, estimate ω_1^2 by guessing a shape, without computing any eigenvalue.

Solution. The lowest mode should have all three masses moving together, so try $\mathbf{g} = (1, 1, 1)$. Then $\mathbf{K}_3\mathbf{g} = (1, 0, 1)$, so

$$\rho(\mathbf{g}) = \frac{\mathbf{g}^\top \mathbf{K}_3 \mathbf{g}}{\mathbf{g}^\top \mathbf{g}} = \frac{2}{3} \approx 0.6667.$$

The true value is $2 - \sqrt{2} \approx 0.5858$, so the guess overestimates by about 14% — and, by Corollary 24.9.3, it is certainly an overestimate.

Refine it: the middle mass should swing furthest, so try $\mathbf{g} = (2, 3, 2)$. Now $\mathbf{K}_3\mathbf{g} = (1, 2, 1)$, and

$$\rho(\mathbf{g}) = \frac{2 + 6 + 2}{4 + 9 + 4} = \frac{10}{17} \approx 0.5882.$$

The error has fallen to about 0.4%. The exact mode is $(1, \sqrt{2}, 1) \propto (2, 2\sqrt{2}, 2)$ and $2\sqrt{2} \approx 2.83$, so the guess $(2, 3, 2)$ is still off by about 6% in its middle entry: the *shape* is wrong by 6% and the *frequency* by 0.4%. Squaring the error in the shape is exactly what Proposition 17.1.2 predicts near a critical point of the quotient.

Quick check. The guess $\mathbf{g} = (1, 2, 1)$ also gives $\rho(\mathbf{g}) = 2/3$, the same as $(1, 1, 1)$, although $(1, 2, 1)$ is much closer to the true mode $(1, \sqrt{2}, 1)$. Is that a contradiction?

Solution. No. $\mathbf{K}_3(1, 2, 1) = (0, 2, 0)$, so $\rho = 4/6 = 2/3$. Both guesses lie on the same level set of ρ, and a level set is not a sphere around the minimizer: the quotient measures a ratio of energies, not a distance to $\mathbf{v}_1$. Being closer to the true mode in the Euclidean sense does not force a smaller value, only being closer *in the right directions* does. What Corollary 24.9.3 guarantees is one-sided, that both values are $\geq 2 - \sqrt{2}$, and both are.

24.9.4 Stiffer, heavier, pinned

The min–max description compares two systems, because the family of subspaces it runs over does not depend on the matrices. Three design changes can now be read off.

Proposition 24.9.4. Stiffening Raises, Loading Lowers

Let $(\mathbf{M}, \mathbf{K})$ and $(\mathbf{M}', \mathbf{K}')$ be mass–spring systems with n degrees of freedom, with natural frequencies ω_j and ω'_j.

(a) If $\mathbf{M}' = \mathbf{M}$ and $\mathbf{K}' \succeq \mathbf{K}$, then $\omega'_j \geq \omega_j$ for every j.
(b) If $\mathbf{K}' = \mathbf{K}$ and $\mathbf{M}' \succeq \mathbf{M}$, then $\omega'_j \leq \omega_j$ for every j.

Idea. Both are the same one-line observation: the hypothesis makes ρ' compare with ρ at **every** vector, and a min–max of a larger function is larger. In (b) the numerator must be nonnegative before enlarging the denominator helps, and that is exactly the hypothesis $\mathbf{K} \succeq \mathbf{O}$ from Definition 24.9.2.

Proof.

(a) Let $\mathbf{v} \neq \mathbf{0}$. By Definition 13.5.1, $\mathbf{v}^\top \mathbf{K}'\mathbf{v} \geq \mathbf{v}^\top \mathbf{K}\mathbf{v}$, and the two quotients have the same positive denominator $\mathbf{v}^\top \mathbf{M}\mathbf{v}$, so $\rho'(\mathbf{v}) \geq \rho(\mathbf{v})$. Fix j and a subspace W of dimension j. Then $\max_{\mathbf{0}\neq\mathbf{v}\in W} \rho'(\mathbf{v}) \geq \max_{\mathbf{0}\neq\mathbf{v}\in W} \rho(\mathbf{v})$, since the maximum of the larger function over the same set is larger; taking the minimum over all such W preserves the inequality. By Corollary 24.9.3 applied to each system, the two sides are $(\omega'_j)^2$ and ω_j^2. Both frequencies are ≥ 0, so $\omega'_j \geq \omega_j$.

(b) Let $\mathbf{v} \neq \mathbf{0}$. Now $\mathbf{v}^\top \mathbf{M}'\mathbf{v} \geq \mathbf{v}^\top \mathbf{M}\mathbf{v} > 0$ by Definition 13.5.1 and $\mathbf{M} \succ \mathbf{O}$, while the common numerator satisfies $\mathbf{v}^\top \mathbf{K}\mathbf{v} \geq 0$ because $\mathbf{K} \succeq \mathbf{O}$. Increasing a positive denominator does not increase a nonnegative fraction, so $\rho'(\mathbf{v}) \leq \rho(\mathbf{v})$, and the argument of (a) runs in reverse. This proves the proposition.

■

Part (b) needs $\mathbf{K} \succeq \mathbf{O}$ and not merely $\mathbf{K} = \mathbf{K}^\top$: for a 1×1 system with $\mathbf{K} = (-1)$, replacing $\mathbf{M} = (1)$ by $\mathbf{M}' = (4)$ moves the quotient from -1 up to $-1/4$. Such a $\mathbf{K}$ is not a stiffness matrix, and the definition excludes it for this reason among others.

The third change is to hold one mass still — to **pin** it. Constraining $x_p = 0$ leaves a system with $n - 1$ degrees of freedom whose matrices are obtained by deleting row and column p.

Proposition 24.9.5. Pinning Interlaces the Frequencies

Optional: nothing later depends on this.

Let $(\mathbf{M}, \mathbf{K})$ be a mass–spring system with $n \geq 2$ degrees of freedom and natural frequencies $\omega_1 \leq \cdots \leq \omega_n$. Fix p, let $I = \{1, \ldots, n\} \setminus \{p\}$, and let $\mathbf{M}' = \mathbf{M}_{I,I}$ and $\mathbf{K}' = \mathbf{K}_{I,I}$ be the principal submatrices on I. Then $(\mathbf{M}', \mathbf{K}')$ is a mass–spring system with $n - 1$ degrees of freedom, and its natural frequencies $\mu_1 \leq \cdots \leq \mu_{n-1}$ satisfy

$$\omega_j \;\leq\; \mu_j \;\leq\; \omega_{j+1} \qquad (1 \leq j \leq n-1).$$

Idea. Pinning restricts the competition to the hyperplane $H = \{\mathbf{v} : v_p = 0\}$. A subspace of H is a subspace of $\mathbb{R}^n$, so the min–max for the pinned system runs over *fewer* competitors than the one for the full system; that gives the left inequality at index j and, after the dimension count that makes Theorem 17.4.2 work, the right one at index $j + 1$.

Proof. Let $H = \{\mathbf{v} \in \mathbb{R}^n : v_p = 0\}$ and let $\varphi\colon H \to \mathbb{R}^{n-1}$ delete the p-th coordinate, a linear isomorphism. For $\mathbf{v} \in H$ with coordinates $\mathbf{y} = \varphi(\mathbf{v})$, the terms of $\mathbf{v}^\top \mathbf{M}\mathbf{v}$ involving index p all vanish, so

$$\mathbf{v}^\top \mathbf{M}\mathbf{v} = \mathbf{y}^\top \mathbf{M}'\mathbf{y}, \qquad \mathbf{v}^\top \mathbf{K}\mathbf{v} = \mathbf{y}^\top \mathbf{K}'\mathbf{y},$$

Both $\mathbf{M}'$ and $\mathbf{K}'$ are symmetric, being principal submatrices of symmetric matrices, and the displays give $\mathbf{y}^\top \mathbf{M}'\mathbf{y} > 0$ for $\mathbf{y} \neq \mathbf{0}$ and $\mathbf{y}^\top \mathbf{K}'\mathbf{y} \geq 0$ for all $\mathbf{y}$, since φ is surjective and $\varphi^{-1}(\mathbf{y}) \neq \mathbf{0}$ for $\mathbf{y} \neq \mathbf{0}$; so $\mathbf{M}' \succ \mathbf{O}$ and $\mathbf{K}' \succeq \mathbf{O}$ by Definition 13.1.1, and $(\mathbf{M}', \mathbf{K}')$ is a mass–spring system. Writing ρ and ρ' for the two generalized Rayleigh quotients, the two

displays give

$$\rho(\mathbf{v}) = \rho'(\varphi(\mathbf{v})) \qquad \text{for } \mathbf{0} \neq \mathbf{v} \in H.$$

Fix j with $1 \leq j \leq n-1$.

($\omega_j \leq \mu_j$) By Corollary 24.9.3 for the pinned system there is a subspace $W' \leq \mathbb{R}^{n-1}$ with $\dim W' = j$ and $\mu_j^2 = \max_{\mathbf{0}\neq\mathbf{y}\in W'} \rho'(\mathbf{y})$. Put $W = \varphi^{-1}(W') \leq H \leq \mathbb{R}^n$, of dimension j. Then $\max_{\mathbf{0}\neq\mathbf{v}\in W} \rho(\mathbf{v}) = \mu_j^2$, and W is one of the competitors in the min–max for ω_j^2, so $\omega_j^2 \leq \mu_j^2$.

($\mu_j \leq \omega_{j+1}$) By the max–min line of Corollary 24.9.3 for the pinned system there is a subspace $W' \leq \mathbb{R}^{n-1}$ with $\dim W' = (n-1) - j + 1 = n - j$ and $\mu_j^2 = \min_{\mathbf{0}\neq\mathbf{y}\in W'} \rho'(\mathbf{y})$. Put $W = \varphi^{-1}(W') \leq \mathbb{R}^n$, of dimension $n - j = n - (j+1) + 1$, which is the dimension the max–min line for ω_{j+1}^2 asks for. Hence

$$\omega_{j+1}^2 \;\geq\; \min_{\mathbf{0}\neq\mathbf{v}\in W} \rho(\mathbf{v}) = \mu_j^2.$$

All frequencies being ≥ 0, taking square roots preserves both inequalities. This proves the proposition.

■

Example. Pinning the middle mass

In Example (Three equal masses, four equal springs) with $k = m = 1$, hold mass 2 fixed. Deleting row and column 2 from $\mathbf{K}_3$ and from $\mathbf{I}_3$ gives $\mathbf{K}' = \operatorname{diag}(2, 2)$ and $\mathbf{M}' = \mathbf{I}_2$, so the pinned frequencies are $\mu_1 = \mu_2 = \sqrt{2}$. The unpinned ones are $\sqrt{2 - \sqrt{2}} \approx 0.765$, $\sqrt{2} \approx 1.414$, and $\sqrt{2 + \sqrt{2}} \approx 1.848$, and indeed

$$0.765 \leq 1.414 \leq 1.414 \leq 1.414 \leq 1.848,$$

which is Proposition 24.9.5, here with two of the four inequalities equalities. The answer is also visible: with the middle mass pinned, the two outer masses no longer interact, and each is a single mass between two unit springs, of frequency $\sqrt{2}$.

Three sentences summarize the section. A structure's frequencies are the generalized eigenvalues of a definite pencil, and its modes are an $\mathbf{M}$-orthonormal basis, so its motion decouples completely. Each frequency is a min–max of one explicit quotient, so a guessed shape bounds the lowest one from above. And because the min–max runs over a family of subspaces that does not mention the matrices, making a structure stiffer raises every frequency, making it heavier lowers every frequency, and pinning a point interlaces them.

24.9.5 Exercises

24.9.5.1 A. Check your understanding

Exercise. A1

(a) State the two hypotheses on $\mathbf{M}$ and $\mathbf{K}$ in Definition 24.9.2, and say which one guarantees that $\rho(\mathbf{v})$ is defined for every $\mathbf{v} \neq \mathbf{0}$.

(b) State Theorem 24.9.2 (b), including the normalization of the eigenvectors.

(c) A system has $\mathbf{K}$ invertible. What does Theorem 24.9.2 say about ω_1? What if $\mathbf{K}$ is singular?

(d) True or false: if every stiffness k_e of a chain is doubled and every mass is left alone, every natural frequency is multiplied by $\sqrt{2}$. Justify your answer.

(e) True or false: for $\mathbf{M} = \operatorname{diag}(1, 4)$ the squared natural frequencies are the eigenvalues of $\mathbf{K}$ divided by 1 and by 4 in some order. Justify your answer.

Solution.

(a) $\mathbf{M} \succ \mathbf{O}$ and $\mathbf{K} \succeq \mathbf{O}$, both symmetric. It is $\mathbf{M} \succ \mathbf{O}$ that makes the denominator $\mathbf{v}^{\top}\mathbf{M}\mathbf{v}$ nonzero, hence positive, for every $\mathbf{v} \neq \mathbf{0}$.

(b) There are $0 \leq \omega_1 \leq \cdots \leq \omega_n$ and a basis $\mathbf{v}_1, \ldots, \mathbf{v}_n$ of $\mathbb{R}^n$ with $\mathbf{K}\mathbf{v}_j = \omega_j^2 \mathbf{M}\mathbf{v}_j$ and $\mathbf{v}_i^{\top}\mathbf{M}\mathbf{v}_j = \delta_{ij}$: the eigenvectors are orthonormal for the $\mathbf{M}$-inner product, not for the standard one.

(c) If $\mathbf{K}$ is invertible then no ω_j is 0: $\omega_j = 0$ would give $\mathbf{K}\mathbf{v}_j = \mathbf{0}$ with $\mathbf{v}_j \neq \mathbf{0}$. So $\omega_1 > 0$ and every motion is a genuine oscillation. If $\mathbf{K}$ is singular then some $\omega_j = 0$, and by Theorem 24.9.2 (c) there are motions containing a term $b_j t\, \mathbf{v}_j$, which grow without bound: a drift, not a vibration.

(d) True. Replacing $\mathbf{K}$ by $2\mathbf{K}$ multiplies ρ by 2 at every vector, hence multiplies every min–max by 2 (Corollary 24.9.3), hence every ω_j^2 by 2 and every ω_j by $\sqrt{2}$.

(e) False; the warning above gives $\mathbf{K} = \left(\begin{smallmatrix} 2 & -1 \\ -1 & 2 \end{smallmatrix}\right)$ with $\mathbf{M} = \operatorname{diag}(1, 4)$, whose squared frequencies are $(5 \pm \sqrt{13})/4$, irrational, while the four candidate quotients $3, 3/4, 1, 1/4$ are rational.

24.9.5.2 B. Practice

Exercise. B1: Two masses, two springs

A cart of mass 2 is joined to a wall by a spring of stiffness 2, and a cart of mass 1 is joined to the first cart by a spring of stiffness 1; the second cart is free on the right.

(a) Write down $\mathbf{M}$ and $\mathbf{K}$.

(b) Find the natural frequencies and an $\mathbf{M}$-orthonormal basis of normal modes.

(c) Verify directly that the two modes are $\mathbf{M}$-orthogonal but **not** orthogonal for the standard inner product.

Solution.

(a) With x_1, x_2 the displacements, Newton and Hooke give $2x_1'' = -2x_1 + (x_2 - x_1)$ and $x_2'' = -(x_2 - x_1)$, so

$$\mathbf{M} = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}, \qquad \mathbf{K} = \begin{pmatrix} 3 & -1 \\ -1 & 1 \end{pmatrix}.$$

Both are symmetric, $\mathbf{M} \succ \mathbf{O}$, and $\mathbf{K} \succ \mathbf{O}$ since its leading minors are 3 and 2 (Theorem 13.1.5).

(b) Expanding the determinant of the pencil,

$$\begin{aligned}\det(\mathbf{K}-\lambda\mathbf{M}) &= (3-2\lambda)(1-\lambda)-1\\ &= 2\lambda^2-5\lambda+2=(2\lambda-1)(\lambda-2),\end{aligned}$$

so $\omega_1^2=\tfrac12$ and $\omega_2^2=2$, that is $\omega_1=1/\sqrt{2}$ and $\omega_2=\sqrt{2}$. For $\lambda=\tfrac12$, $\mathbf{K}-\tfrac12\mathbf{M}=\left(\begin{smallmatrix}2&-1\\-1&1/2\end{smallmatrix}\right)$ has kernel spanned by $(1,2)$; for $\lambda=2$, $\mathbf{K}-2\mathbf{M}=\left(\begin{smallmatrix}-1&-1\\-1&-1\end{smallmatrix}\right)$ has kernel spanned by $(1,-1)$. Since $(1,2)^{\top}\mathbf{M}(1,2)=2+4=6$ and $(1,-1)^{\top}\mathbf{M}(1,-1)=2+1=3$,

$$\mathbf{v}_1=\tfrac{1}{\sqrt{6}}(1,2),\qquad \mathbf{v}_2=\tfrac{1}{\sqrt{3}}(1,-1).$$

In the slow mode the two carts move together, the light one further; in the fast mode they move against each other.

(c) $(1,2)^{\top}\mathbf{M}(1,-1)=2\cdot1\cdot1+1\cdot2\cdot(-1)=0$, so they are $\mathbf{M}$-orthogonal, as Theorem 24.9.2 (b) requires. But $(1,2)\cdot(1,-1)=1-2=-1\neq0$: the standard inner product does not see the modes.

Exercise. B2: Bounding the slowest mode

For the system of Exercise B1, estimate ω_1^2 from above using the guesses $\mathbf{g}=(1,1)$ and $\mathbf{g}=(2,3)$, and compare with the exact value.

Solution. For $\mathbf{g}=(1,1)$: $\mathbf{g}^{\top}\mathbf{K}\mathbf{g}=3-1-1+1=2$ and $\mathbf{g}^{\top}\mathbf{M}\mathbf{g}=2+1=3$, so $\rho(\mathbf{g})=2/3\approx0.667$.

For $\mathbf{g}=(2,3)$: $\mathbf{K}\mathbf{g}=(6-3,-2+3)=(3,1)$, so $\mathbf{g}^{\top}\mathbf{K}\mathbf{g}=6+3=9$; and $\mathbf{g}^{\top}\mathbf{M}\mathbf{g}=8+9=17$, so $\rho(\mathbf{g})=9/17\approx0.529$.

The exact value is $\omega_1^2=1/2$. Both estimates exceed it, as Corollary 24.9.3 requires, and the second guess, being nearer the true mode $(1,2)\propto(2,4)$, is much nearer the true value.

Exercise. B3: Stiffening a chain

In the system of Exercise B1, the stiffness of the second spring is increased from 1 to 2, so that $\mathbf{K}$ becomes $\mathbf{K}'=\left(\begin{smallmatrix}4&-2\\-2&2\end{smallmatrix}\right)$.

(a) Verify that $\mathbf{K}'\succeq\mathbf{K}$.

(b) Compute the new squared frequencies and check Proposition 24.9.4 (a).

Solution.

(a) $\mathbf{K}'-\mathbf{K}=\left(\begin{smallmatrix}1&-1\\-1&1\end{smallmatrix}\right)$, whose trace is 2 and determinant 0, so its eigenvalues are 2 and 0 and it is positive semidefinite (Theorem 13.1.4). Hence $\mathbf{K}'\succeq\mathbf{K}$ by Definition 13.5.1. Directly: $\mathbf{v}^{\top}(\mathbf{K}'-\mathbf{K})\mathbf{v}=(v_1-v_2)^2\geq0$, which is the extra spring energy.

(b) $\det(\mathbf{K}'-\lambda\mathbf{M})=(4-2\lambda)(2-\lambda)-4=2\lambda^2-8\lambda+4$, so the squared frequencies are $2\pm\sqrt{2}$, about 0.586 and 3.414. The old ones were 0.5 and 2, and $0.586\geq0.5$, $3.414\geq2$,

as Proposition 24.9.4 (a) requires. Note the two increases are quite different in size: the proposition orders the frequencies, it does not move them by a common amount.

24.9.5.3 C. Going deeper

Exercise. C1: The free chain drifts

Let $(\mathbf{M}, \mathbf{K})$ be a mass–spring system.

(a) Prove that $\omega_1 = 0$ if and only if $\mathbf{K}$ is singular, and that in that case the generalized eigenvectors for $\omega = 0$ are exactly the nonzero vectors of $\ker \mathbf{K}$.

(b) Hence describe all motions of the free chain of Example (Three systems) (c) with three equal masses m and equal inner stiffnesses k, given that $\ker \mathbf{K}$ is spanned by $\mathbf{1}$.

(c) Deduce that the total momentum $\sum_i m_i x_i'(t)$ of the free chain is constant in time.

Solution.

(a) By Theorem 24.9.2 (b), $\omega_1 = 0$ if and only if some $\mathbf{v} \neq \mathbf{0}$ has $\mathbf{K}\mathbf{v} = 0 \cdot \mathbf{M}\mathbf{v} = \mathbf{0}$, that is if and only if $\ker \mathbf{K} \neq \{\mathbf{0}\}$, that is if and only if $\mathbf{K}$ is singular (Theorem 3.4.4). For such an ω the equation $\mathbf{K}\mathbf{v} = \omega^2 \mathbf{M}\mathbf{v}$ reads $\mathbf{K}\mathbf{v} = \mathbf{0}$, so its nonzero solutions are exactly the nonzero vectors of $\ker \mathbf{K}$.

(b) Here $\mathbf{M} = m\mathbf{I}_3$ and $\mathbf{K} = k\begin{pmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{pmatrix}$, whose kernel is $\operatorname{span}(\mathbf{1})$, one-dimensional. So $\omega_1 = 0$ with mode $\mathbf{v}_1 = \mathbf{1}/\sqrt{3m}$, and the other two frequencies are positive. By Theorem 24.9.2 (c) every motion is

$$\mathbf{x}(t) = (a_1 + b_1 t)\mathbf{v}_1 + \sum_{j=2}^{3} (a_j \cos \omega_j t + b_j \sin \omega_j t)\mathbf{v}_j \,:$$

an oscillation superposed on a uniform drift of the whole chain. The drift is the motion Theorem 24.9.2 produces for a singular $\mathbf{K}$, and physically it is the chain floating away at constant speed.

(c) The total momentum is $\mathbf{1}^{\top}\mathbf{M}\mathbf{x}'(t)$. Differentiating and using Eq. 24.9.1, $\mathbf{1}^{\top}\mathbf{M}\mathbf{x}''(t) = -\mathbf{1}^{\top}\mathbf{K}\mathbf{x}(t) = -(\mathbf{K}\mathbf{1})^{\top}\mathbf{x}(t) = 0$, where $\mathbf{K}^{\top} = \mathbf{K}$ and $\mathbf{K}\mathbf{1} = \mathbf{0}$ were used. A differentiable function with vanishing derivative is constant, so $\mathbf{1}^{\top}\mathbf{M}\mathbf{x}'(t)$ does not depend on t.

Exercise. C2: Adding a mass cannot raise a frequency, and can lower it a lot

(a) Let $(\mathbf{M}, \mathbf{K})$ be a mass–spring system and let $t > 0$. Prove that the system $(\mathbf{M} + t\mathbf{e}_p\mathbf{e}_p^{\top}, \mathbf{K})$, obtained by adding mass t to the p-th degree of freedom, has $\omega_j'(t) \leq \omega_j$ for every j, and that $t \mapsto \omega_j'(t)$ is nonincreasing.

(b) For the single mass of Example (Three systems) (a) with $m = k_1 = k_2 = 1$, compute $\omega_1'(t)$ exactly and find $\lim_{t\to\infty} \omega_1'(t)$.

(c) Give an example of a mass–spring system and a p for which adding mass at p leaves the **largest** frequency unchanged. *Hint: look for a mode that does not move the p-th mass.*

Solution.

(a) The matrix $t\mathbf{e}_p\mathbf{e}_p^\top$ is symmetric with $\mathbf{v}^\top(t\mathbf{e}_p\mathbf{e}_p^\top)\mathbf{v} = tv_p^2 \geq 0$, so it is positive semidefinite and $\mathbf{M} + t\mathbf{e}_p\mathbf{e}_p^\top \succeq \mathbf{M}$. By Proposition 24.9.4 (b), $\omega_j'(t) \leq \omega_j$ for every j. For monotonicity in t, if $0 < s \leq t$ then $\mathbf{M} + t\mathbf{e}_p\mathbf{e}_p^\top \succeq \mathbf{M} + s\mathbf{e}_p\mathbf{e}_p^\top$, because the difference is $(t-s)\mathbf{e}_p\mathbf{e}_p^\top \succeq \mathbf{O}$, and applying (b) again to these two systems gives $\omega_j'(t) \leq \omega_j'(s)$.

(b) Here $n = 1$, $\mathbf{M} = (1+t)$, and $\mathbf{K} = (2)$, so the single squared frequency is $\rho(1) = 2/(1+t)$ and $\omega_1'(t) = \sqrt{2/(1+t)}$. It is decreasing in t and tends to 0: an infinitely heavy cart does not move.

(c) Let $n = 2$, $\mathbf{M} = \mathbf{I}_2$, and $\mathbf{K} = \operatorname{diag}(1, 4)$, so $\omega_1^2 = 1$, $\omega_2^2 = 4$ with modes $\mathbf{e}_1, \mathbf{e}_2$. Adding mass at $p = 1$ gives $\mathbf{M}' = \operatorname{diag}(1+t, 1)$, and $\det(\mathbf{K} - \lambda\mathbf{M}') = (1 - \lambda(1+t))(4-\lambda)$, whose roots are $1/(1+t)$ and 4. The largest squared frequency is still 4, because its mode $\mathbf{e}_2$ has no component at $p = 1$ and the added mass therefore never moves in that mode. This also shows the inequalities of Proposition 24.9.4 (b) can be equalities.

24.10 Principal Components

Chapter 13 §11 proved that the k-dimensional subspace **through the origin** closest to a cloud of points is spanned by the leading left singular vectors, and then said plainly what it had not proved: the centering, the covariance matrix, and the reading of $\sigma_i^2/\sum_l \sigma_l^2$ as a fraction of variance. This section supplies all three — and the variance fraction turns out to be a theorem, not a convention — and then marks exactly where linear algebra stops and statistics begins. The honest answer is: more is provable than one might expect, and less than the phrase "explained variance" usually suggests.

Throughout, the field is $\mathbb{R}$. The data are $N \geq 1$ points $\mathbf{x}_1, \ldots, \mathbf{x}_N \in \mathbb{R}^n$, written as the columns of a matrix, and $\|\cdot\|$ is the Euclidean norm.

24.10.1 Centering

An **affine subspace** of $\mathbb{R}^n$ is a coset $A = \mathbf{p} + U = \{\mathbf{p} + \mathbf{z} : \mathbf{z} \in U\}$ of a subspace U, and its dimension is $\dim U$; Chapter 18 §01 calls U the direction space and shows in Proposition 18.1.4 that every affine hull has this form. The subspaces of $\mathbb{R}^n$ are the affine subspaces containing $\mathbf{0}$.

Chapter 13 §11 fitted subspaces, which all pass through $\mathbf{0}$. That is the wrong family for data: a cloud of points far from the origin is badly described by any line through the origin, and the quantity minimized is then dominated by the distance from $\mathbf{0}$ to the cloud rather than by its shape. Widening the competition to all affine subspaces fixes this, and it costs nothing: the extra freedom is used up in one place, and we can say in advance where.

Subtract the mean once, and the affine problem becomes the linear problem Chapter 13 solved.

Definition 24.10.1. Data matrix, mean, and centered data matrix

Let $\mathbf{x}_1, \ldots, \mathbf{x}_N \in \mathbb{R}^n$. The **data matrix** is the matrix $\mathbf{X} \in M_{n\times N}(\mathbb{R})$ whose j-th **column** is $\mathbf{x}_j$. The **mean** of the data is

$$\bar{\mathbf{x}} := \frac{1}{N}\sum_{j=1}^{N} \mathbf{x}_j \in \mathbb{R}^n,$$

and the **centered data matrix** is

$$\mathbf{X}_c \coloneqq \mathbf{X} - \bar{\mathbf{x}}\,\mathbf{1}^\top \in M_{n\times N}(\mathbb{R}),$$

whose j-th column is $\mathbf{c}_j \coloneqq \mathbf{x}_j - \bar{\mathbf{x}}$.

Two clauses deserve comment. **Columns** are the data points, so n is the number of measured quantities and N the number of records; the opposite convention is also in use, and every formula below would be transposed under it. And $\bar{\mathbf{x}}\,\mathbf{1}^\top$ is the $n \times N$ matrix all of whose columns equal $\bar{\mathbf{x}}$, since $\mathbf{1} \in \mathbb{R}^N$ is the all-ones vector.

The defining property of the centered matrix is that its own mean is zero:

$$\mathbf{X}_c\mathbf{1} = \mathbf{X}\mathbf{1} - \bar{\mathbf{x}}(\mathbf{1}^\top\mathbf{1}) = N\bar{\mathbf{x}} - N\bar{\mathbf{x}} = \mathbf{0},$$

using $\mathbf{1}^\top\mathbf{1} = N$. In words, each **row** of $\mathbf{X}_c$ sums to 0.

Example. Centering three data sets

(a) $N = 4$ points in $\mathbb{R}^2$: $(5,4), (4,1), (2,3), (1,0)$. Their sum is $(12,8)$, so $\bar{\mathbf{x}} = (3,2)$ and

$$\mathbf{X}_c = \begin{pmatrix} 2 & 1 & -1 & -2 \\ 2 & -1 & 1 & -2 \end{pmatrix}.$$

Both rows sum to 0.

(b) $N = 1$. Then $\bar{\mathbf{x}} = \mathbf{x}_1$ and $\mathbf{X}_c = \mathbf{O}$: a single point carries no shape at all. This degenerate case matters, because every ratio below has $0/0$ here and the statements must exclude it or say what they mean.

(c) Data already centered: if $\mathbf{X}\mathbf{1} = \mathbf{0}$ then $\bar{\mathbf{x}} = \mathbf{0}$ and $\mathbf{X}_c = \mathbf{X}$. Centering is idempotent, since $\mathbf{X}_c\mathbf{1} = \mathbf{0}$ always.

Change one clause and the object changes. Suppose one subtracts the mean of each *column* instead of forming the mean of the columns, replacing $\mathbf{x}_j$ by $\mathbf{x}_j - (\frac{1}{n}\mathbf{1}_n^\top\mathbf{x}_j)\mathbf{1}_n$ with $\mathbf{1}_n \in \mathbb{R}^n$. That removes the average of the measured quantities inside each record, which is a meaningful operation for some data and a meaningless one for most, and in any case is not the one below. The clause that matters is *which index is averaged over*.

24.10.2 The best-fitting affine subspace

Distances to an affine subspace reduce to distances to its direction space, and the reduction is Chapter 11's Best Approximation Theorem.

Lemma 24.10.1. Distance to a flat, and where its translate wants to be

Let $U \subseteq \mathbb{R}^n$ be a subspace, let $Q = \mathbf{I}_n - P_U$ be the orthogonal projection onto $U^\perp$, and let $\mathbf{p} \in \mathbb{R}^n$.

(a) For every $\mathbf{x} \in \mathbb{R}^n$, $d(\mathbf{x}, \mathbf{p}+U) = \|Q(\mathbf{x}-\mathbf{p})\|$, and the nearest point of $\mathbf{p}+U$ exists.

(b) For points $\mathbf{x}_1, \dots, \mathbf{x}_N$ with mean $\bar{\mathbf{x}}$,

$$\sum_{j=1}^{N} \|Q(\mathbf{x}_j - \mathbf{p})\|^2 = \sum_{j=1}^{N} \|Q(\mathbf{x}_j - \bar{\mathbf{x}})\|^2 + N\|Q(\bar{\mathbf{x}} - \mathbf{p})\|^2.$$

In particular the left side is smallest exactly when $\bar{\mathbf{x}} \in \mathbf{p} + U$.

Idea. Part (a) is the Best Approximation Theorem after sliding the whole picture by $-\mathbf{p}$. Part (b) is the one-dimensional identity "sum of squares about a point = sum of squares about the mean + N times the squared displacement", proved the way it always is: expand the square and notice that the cross term is a multiple of $\sum_j (\mathbf{x}_j - \bar{\mathbf{x}})$, which is $\mathbf{0}$ by the definition of the mean.

Proof.

(a) A point of $\mathbf{p}+U$ is $\mathbf{p}+\mathbf{z}$ with $\mathbf{z} \in U$, and $\|\mathbf{x}-(\mathbf{p}+\mathbf{z})\| = \|(\mathbf{x}-\mathbf{p})-\mathbf{z}\|$. By Theorem 11.3.5 the quantity $\|(\mathbf{x}-\mathbf{p})-\mathbf{z}\|$ over $\mathbf{z} \in U$ is smallest exactly at $\mathbf{z} = P_U(\mathbf{x}-\mathbf{p})$, where its value is $\|(\mathbf{x}-\mathbf{p}) - P_U(\mathbf{x}-\mathbf{p})\| = \|Q(\mathbf{x}-\mathbf{p})\|$. So the minimum is attained and equals $\|Q(\mathbf{x}-\mathbf{p})\|$, which is $d(\mathbf{x}, \mathbf{p}+U)$ by Definition 11.3.3 applied after the translation.

(b) Put $\mathbf{y}_j = Q(\mathbf{x}_j - \bar{\mathbf{x}})$ and $\mathbf{d} = Q(\bar{\mathbf{x}} - \mathbf{p})$, so that $Q(\mathbf{x}_j - \mathbf{p}) = \mathbf{y}_j + \mathbf{d}$ by linearity of Q. Since Q is linear and $\sum_j (\mathbf{x}_j - \bar{\mathbf{x}}) = N\bar{\mathbf{x}} - N\bar{\mathbf{x}} = \mathbf{0}$,

$$\sum_j \mathbf{y}_j = Q\Big(\sum_j (\mathbf{x}_j - \bar{\mathbf{x}})\Big) = \mathbf{0}.$$

Therefore, expanding each square,

$$\begin{aligned} \sum_j \|\mathbf{y}_j + \mathbf{d}\|^2 &= \sum_j \left(\|\mathbf{y}_j\|^2 + 2\langle \mathbf{y}_j, \mathbf{d}\rangle + \|\mathbf{d}\|^2\right) \\ &= \sum_j \|\mathbf{y}_j\|^2 + 2\Big\langle \sum_j \mathbf{y}_j, \mathbf{d}\Big\rangle + N\|\mathbf{d}\|^2 \\ &= \sum_j \|\mathbf{y}_j\|^2 + N\|\mathbf{d}\|^2, \end{aligned}$$

which is the identity. The second term is ≥ 0 and vanishes exactly when $\mathbf{d} = \mathbf{0}$, that is when $\bar{\mathbf{x}} - \mathbf{p} \in \ker Q = U$ (Theorem 11.3.2), that is when $\bar{\mathbf{x}} \in \mathbf{p} + U$. This proves the lemma.

■

Part (b) is the whole of "the best affine subspace passes through the mean". It holds for **every** direction space U, before any optimization over U has happened, which is why the affine problem costs nothing extra.

Theorem 24.10.2. The Best-Fitting Affine Subspace

Let $\mathbf{x}_1, \dots, \mathbf{x}_N \in \mathbb{R}^n$ have mean $\bar{\mathbf{x}}$ and centered data matrix $\mathbf{X}_c$ (Definition 24.10.1), let $\sigma_1 \geq \sigma_2 \geq \cdots \geq 0$ be the singular values of $\mathbf{X}_c$ (Definition 13.9.1), and let $\mathbf{u}_1, \mathbf{u}_2, \dots$ be the left singular vectors of a fixed singular value decomposition of $\mathbf{X}_c$ (Theorem 13.9.1). Let $1 \leq k \leq \min(n, N)$. Among all affine subspaces $A \subseteq \mathbb{R}^n$ of dimension at most k, the total squared distance

$$D(A) := \sum_{j=1}^{N} d(\mathbf{x}_j, A)^2$$

is smallest for

$$A_0 = \bar{\mathbf{x}} + \operatorname{span}(\mathbf{u}_1, \dots, \mathbf{u}_k),$$

and its smallest value is $\sum_{i>k} \sigma_i^2$. Moreover **every** minimizing A contains $\bar{\mathbf{x}}$.

Idea. Two optimizations in sequence, and Lemma 24.10.1 does the outer one first. ① For a *fixed* direction space U, the best translate is the one through $\bar{\mathbf{x}}$, whatever U is. ② With the translate fixed at $\bar{\mathbf{x}}$, the objective becomes the total squared distance from the **centered** points to the subspace U — which is exactly the quantity Chapter 13 §11 minimized. So nothing new has to be proved about subspaces; the work is entirely in noticing that step ① does not interact with step ②.

Proof. Write $A = \mathbf{p} + U$ with $\dim U \le k$, and let $Q = \mathbf{I}_n - P_U$. By Lemma 24.10.1 (a),

$$D(\mathbf{p} + U) = \sum_j \|Q(\mathbf{x}_j - \mathbf{p})\|^2,$$

and by part (b) of the same lemma this is at least $\sum_j \|Q(\mathbf{x}_j - \bar{\mathbf{x}})\|^2 = D(\bar{\mathbf{x}} + U)$, with equality exactly when $\bar{\mathbf{x}} \in \mathbf{p} + U$. So it suffices to minimize $D(\bar{\mathbf{x}} + U)$ over subspaces U with $\dim U \le k$.

Now $\|Q(\mathbf{x}_j - \bar{\mathbf{x}})\| = \|\mathbf{c}_j - P_U \mathbf{c}_j\| = d(\mathbf{c}_j, U)$ by Definition 11.3.3, where $\mathbf{c}_j = \mathbf{x}_j - \bar{\mathbf{x}}$ is the j-th column of $\mathbf{X}_c$. Hence

$$D(\bar{\mathbf{x}} + U) = \sum_{j=1}^{N} d(\mathbf{c}_j, U)^2,$$

which is the quantity $D(U)$ of Corollary 13.11.4 for the matrix $\mathbf{X}_c \in M_{n\times N}(\mathbb{R})$. That corollary, applicable because $1 \le k \le \min(n, N)$, says that the minimum over $\dim U \le k$ is $\sum_{i>k} \sigma_i^2$ and is attained at $U_0 = \operatorname{span}(\mathbf{u}_1, \dots, \mathbf{u}_k)$. Combining, $D(A) \ge \sum_{i>k} \sigma_i^2$ for every admissible A, with equality at $A_0 = \bar{\mathbf{x}} + U_0$.

Finally let $A = \mathbf{p} + U$ be any minimizer. Then $D(\mathbf{p} + U) = D(\bar{\mathbf{x}} + U)$, since the right side is admissible and no smaller; by the equality clause of Lemma 24.10.1 (b) this forces $\bar{\mathbf{x}} \in \mathbf{p} + U = A$. This proves the theorem.

■

Quick check. Is the minimizer unique? Take the four centered points $(1, 0)$, $(-1, 0)$, $(0, 1)$, $(0, -1)$ in $\mathbb{R}^2$ and $k = 1$.

Solution. No. Here $\bar{\mathbf{x}} = \mathbf{0}$ and $\mathbf{X}_c \mathbf{X}_c^{\top} = \operatorname{diag}(2, 2)$, so $\sigma_1 = \sigma_2 = \sqrt{2}$ and the theorem's minimum is $\sigma_2^2 = 2$. Every line through the origin attains it: for $\operatorname{span}(\mathbf{e}_1)$ the four squared distances are $0, 0, 1, 1$; for $\operatorname{span}((1, 1))$ each point is at squared distance $1 - \frac{1}{2} = \frac{1}{2}$. Total 2 in both cases. What the theorem asserts is that A_0 is *a* minimizer and that $\sum_{i>k} \sigma_i^2$ is *the* minimum value; uniqueness fails exactly when $\sigma_k = \sigma_{k+1}$, just as the truncation of Definition 13.11.1 is nonunique then.

24.10.3 The sample covariance matrix

Fix a direction $\mathbf{w} \in \mathbb{R}^n$ with $\mathbf{w} \neq \mathbf{0}$ and look at the data through it: each point $\mathbf{x}_j$ has the single coordinate $t_j = \mathbf{w}^\top \mathbf{x}_j$ along $\mathbf{w}$. The mean of these numbers is

$$\frac{1}{N}\sum_j \mathbf{w}^\top \mathbf{x}_j = \mathbf{w}^\top \bar{\mathbf{x}},$$

and their spread about that mean is the number

$$\begin{aligned}\frac{1}{N}\sum_{j=1}^{N}\left(\mathbf{w}^\top \mathbf{x}_j - \mathbf{w}^\top \bar{\mathbf{x}}\right)^2 &= \frac{1}{N}\sum_{j=1}^{N}\left(\mathbf{w}^\top \mathbf{c}_j\right)^2 \\ &= \frac{1}{N}\,\mathbf{w}^\top \mathbf{X}_c \mathbf{X}_c^\top\,\mathbf{w},\end{aligned}$$

the last step because $\sum_j (\mathbf{w}^\top \mathbf{c}_j)^2 = \|\mathbf{X}_c^\top \mathbf{w}\|^2 = \mathbf{w}^\top \mathbf{X}_c \mathbf{X}_c^\top \mathbf{w}$. The direction $\mathbf{w}$ enters only through a quadratic form, and the matrix of that form does not depend on $\mathbf{w}$. That is a recurring expression, so it gets a name.

The sample covariance matrix is the single matrix that answers "how spread out are the data in direction $\mathbf{w}$*?" for every* $\mathbf{w}$ *at once.*

Definition 24.10.2. Sample covariance matrix

Let $\mathbf{x}_1, \dots, \mathbf{x}_N \in \mathbb{R}^n$ with centered data matrix $\mathbf{X}_c$ (Definition 24.10.1). The **sample covariance matrix** of the data is

$$\mathbf{S} := \frac{1}{N}\,\mathbf{X}_c \mathbf{X}_c^\top \in M_n(\mathbb{R}),$$

with entries

$$s_{ab} = \frac{1}{N}\sum_{j=1}^{N}(x_{aj} - \bar{x}_a)(x_{bj} - \bar{x}_b) \qquad (1 \le a, b \le n).$$

For $\mathbf{w} \in \mathbb{R}^n$ with $\mathbf{w} \neq \mathbf{0}$, the **variance of the data in the direction** $\mathbf{w}$ is the number $\mathbf{w}^\top \mathbf{S}\mathbf{w}/\|\mathbf{w}\|^2$, which for a **unit** $\mathbf{w}$ is $\mathbf{w}^\top \mathbf{S}\mathbf{w}$; and the **total variance** of the data is $\operatorname{tr} \mathbf{S}$.

Clause by clause. The factor $1/N$ makes $\mathbf{S}$ an average, so it does not grow when data are duplicated. The entry s_{aa} is the ordinary variance of the a-th measured quantity, and s_{ab} for $a \neq b$ says whether quantities a and b tend to deviate from their means in the same direction (positive) or in opposite directions (negative). The centering is inside the definition, not optional: without it the same formula would measure spread about $\mathbf{0}$ rather than about the data.

Note also what $\mathbf{S}$ does **not** depend on: the order of the columns, since the defining sum runs over all of them. It does depend on the units of each coordinate, which is the subject of the warning below.

Example. Three covariance matrices

(a) The four points of Example (Centering three data sets) (a). Then

$$\mathbf{X}_c \mathbf{X}_c^\top = \begin{pmatrix} 10 & 6 \\ 6 & 10 \end{pmatrix}, \qquad \mathbf{S} = \begin{pmatrix} 5/2 & 3/2 \\ 3/2 & 5/2 \end{pmatrix}.$$

Both coordinates have variance $5/2$, and the positive off-diagonal entry says they rise

and fall together.

(b) Two points $\pm\mathbf{v}$ in $\mathbb{R}^n$ with $\mathbf{v} \neq \mathbf{0}$, so $N = 2$ and $\bar{\mathbf{x}} = \mathbf{0}$. Then $\mathbf{X}_c\mathbf{X}_c^\top = 2\mathbf{v}\mathbf{v}^\top$ and $\mathbf{S} = \mathbf{v}\mathbf{v}^\top$, of rank one: all the spread is in one direction, and $\mathbf{w}^\top\mathbf{S}\mathbf{w} = (\mathbf{w}^\top\mathbf{v})^2$ vanishes for every $\mathbf{w} \perp \mathbf{v}$.

(c) $N = 1$. Then $\mathbf{X}_c = \mathbf{O}$ and $\mathbf{S} = \mathbf{O}$: no direction has any spread.

Change the definition in one place and it stops being a covariance. Replacing $\mathbf{X}_c$ by $\mathbf{X}$ gives $\frac{1}{N}\mathbf{X}\mathbf{X}^\top$, whose (a, a) entry is the mean of x_{aj}^2 rather than the variance; for the data of (a) it is $\left(\begin{smallmatrix} 46/4 & 30/4 \\ 30/4 & 26/4 \end{smallmatrix}\right)$, which is not $\mathbf{S}$, and its top eigenvector is a multiple of $(1, 0.72\ldots)$ rather than of $(1, 1)$. The clause that fails is the subtraction of the mean, and with it every statement below.

Proposition 24.10.3. The Covariance Matrix and the Singular Value Decomposition

Let $\mathbf{X}_c \in M_{n\times N}(\mathbb{R})$ be a centered data matrix with singular values $\sigma_1 \geq \sigma_2 \geq \ldots$ (padded with zeros to length n), and let $\mathbf{X}_c = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ be a singular value decomposition with columns $\mathbf{u}_1, \ldots, \mathbf{u}_n$ of $\mathbf{U}$. Put $\mathbf{S} = \frac{1}{N}\mathbf{X}_c\mathbf{X}_c^\top$. Then:

(a) $\mathbf{S}$ is symmetric and $\mathbf{S} \succeq \mathbf{O}$;

(b) $\mathbf{S}\mathbf{u}_i = \dfrac{\sigma_i^2}{N}\,\mathbf{u}_i$ for every i, so the eigenvalues of $\mathbf{S}$ are $\sigma_1^2/N \geq \cdots \geq \sigma_n^2/N$ and the $\mathbf{u}_i$ are an orthonormal eigenbasis;

(c) for every $\mathbf{w} \neq \mathbf{0}$, the variance of the data in the direction $\mathbf{w}$ is the Rayleigh quotient $R_{\mathbf{S}}(\mathbf{w})$ of Definition 17.1.1;

(d) $\operatorname{tr}\mathbf{S} = \dfrac{1}{N}\sum_{j=1}^{N}\|\mathbf{x}_j - \bar{\mathbf{x}}\|^2 = \dfrac{1}{N}\sum_i \sigma_i^2$.

Idea. Every part is one line once $\mathbf{X}_c = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ is written down: (a) and (b) because $\mathbf{X}_c\mathbf{X}_c^\top = \mathbf{U}(\mathbf{\Sigma}\mathbf{\Sigma}^\top)\mathbf{U}^\top$ is already a diagonalization, (c) because the computation before Definition 24.10.2 *was* the Rayleigh quotient, and (d) because the trace does not notice the unitary factors and reads the same quantity off the rows and off the columns.

Proof.

(a) $\mathbf{S}^\top = \frac{1}{N}(\mathbf{X}_c\mathbf{X}_c^\top)^\top = \frac{1}{N}\mathbf{X}_c\mathbf{X}_c^\top = \mathbf{S}$, and for every $\mathbf{w}$, $\mathbf{w}^\top\mathbf{S}\mathbf{w} = \frac{1}{N}\|\mathbf{X}_c^\top\mathbf{w}\|^2 \geq 0$; with symmetry this is Definition 13.1.1.

(b) Since $\mathbf{V}$ is orthogonal, $\mathbf{V}^\top\mathbf{V} = \mathbf{I}_N$, so

$$\mathbf{X}_c\mathbf{X}_c^\top = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top\mathbf{V}\mathbf{\Sigma}^\top\mathbf{U}^\top = \mathbf{U}(\mathbf{\Sigma}\mathbf{\Sigma}^\top)\mathbf{U}^\top,$$

and $\mathbf{\Sigma}\mathbf{\Sigma}^\top = \operatorname{diag}(\sigma_1^2, \ldots, \sigma_n^2)$ by the shape of $\mathbf{\Sigma}$ in Theorem 13.9.1. Multiplying on the right by $\mathbf{u}_i$ and using $\mathbf{U}^\top\mathbf{u}_i = \mathbf{e}_i$ gives $\mathbf{X}_c\mathbf{X}_c^\top\mathbf{u}_i = \sigma_i^2\mathbf{u}_i$; divide by N. The $\mathbf{u}_i$ are orthonormal and there are n of them, so they are an orthonormal basis of $\mathbb{R}^n$, and the list σ_i^2/N is decreasing.

(c) By the computation preceding Definition 24.10.2, the variance in direction $\mathbf{w}$ is $\mathbf{w}^\top\mathbf{S}\mathbf{w}/\|\mathbf{w}\|^2$. Since $\mathbf{S}$ is symmetric, $R_{\mathbf{S}}(\mathbf{w}) = \langle\mathbf{S}\mathbf{w}, \mathbf{w}\rangle/\langle\mathbf{w}, \mathbf{w}\rangle = \mathbf{w}^\top\mathbf{S}\mathbf{w}/\|\mathbf{w}\|^2$ by Definition 17.1.1, which is the same number.

(d) The j-th diagonal entry of $\mathbf{X}_c^\top\mathbf{X}_c$ is $\|\mathbf{c}_j\|^2$, so $\operatorname{tr}(\mathbf{X}_c^\top\mathbf{X}_c) = \sum_j\|\mathbf{c}_j\|^2$; and $\operatorname{tr}(\mathbf{X}_c\mathbf{X}_c^\top) = \operatorname{tr}(\mathbf{X}_c^\top\mathbf{X}_c)$ by Theorem 0.9.6 (3). On the other hand the display in (b) and the same

trace identity give

$$\begin{aligned}\operatorname{tr}(\mathbf{X}_c\mathbf{X}_c^\top) &= \operatorname{tr}(\mathbf{U}(\mathbf{\Sigma}\mathbf{\Sigma}^\top)\mathbf{U}^\top)\\ &= \operatorname{tr}((\mathbf{\Sigma}\mathbf{\Sigma}^\top)\mathbf{U}^\top\mathbf{U}) = \sum_i \sigma_i^2,\end{aligned}$$

using $\mathbf{U}^\top\mathbf{U} = \mathbf{I}_n$. Dividing by N gives both equalities. This proves the proposition.

■

Part (c) connects this section to Chapter 17. The question "which direction has the most spread?" is the question "where is $R_{\mathbf{S}}$ largest?", and Proposition 17.1.1 answers it: the maximum is $\lambda_1(\mathbf{S}) = \sigma_1^2/N$, attained at $\mathbf{u}_1$. "Which direction has the most spread among those orthogonal to $\mathbf{u}_1, \dots, \mathbf{u}_{k-1}$?" is answered the same way by Proposition 17.1.3, with maximum σ_k^2/N at $\mathbf{u}_k$. So the leading left singular vectors are not only the best-fitting directions of Theorem 24.10.2; they are also, one at a time, the directions of greatest remaining spread. The two descriptions agree because they are the two halves of one Pythagorean identity, which is the next theorem.

Definition 24.10.3. Principal directions and components

In the situation of Proposition 24.10.3, the unit vectors $\mathbf{u}_1, \mathbf{u}_2, \dots$ are the **principal directions** of the data, and the numbers $\mathbf{u}_i^\top\mathbf{c}_1, \dots, \mathbf{u}_i^\top\mathbf{c}_N$ are the i-th **principal components**: the coordinates of the centered data along $\mathbf{u}_i$.

24.10.4 How much variance a subspace explains

Theorem 24.10.4. Variance Captured by the Principal Directions

In the situation of Proposition 24.10.3, let $1 \le k \le \min(n, N)$, let $U_0 = \operatorname{span}(\mathbf{u}_1, \dots, \mathbf{u}_k)$, and let P be the orthogonal projection onto a subspace U with $\dim U \le k$. Write

$$\operatorname{var}(U) := \frac{1}{N}\sum_{j=1}^{N} \|P(\mathbf{x}_j - \bar{\mathbf{x}})\|^2.$$

Then:

(a) $\operatorname{var}(U)$ is the total variance of the projected points $P\mathbf{x}_1, \dots, P\mathbf{x}_N$, whose mean is $P\bar{\mathbf{x}}$;

(b) $\operatorname{var}(U) + \dfrac{1}{N}D(\bar{\mathbf{x}} + U) = \operatorname{tr}\mathbf{S}$, where D is as in Theorem 24.10.2;

(c) $\operatorname{var}(U) \le \operatorname{var}(U_0) = \dfrac{1}{N}\sum_{i\le k}\sigma_i^2$;

(d) if $\mathbf{X}_c \neq \mathbf{O}$, the fraction of the total variance captured by U_0 is

$$\frac{\operatorname{var}(U_0)}{\operatorname{tr}\mathbf{S}} = \frac{\sigma_1^2 + \cdots + \sigma_k^2}{\sigma_1^2 + \sigma_2^2 + \dots}.$$

Idea. Everything is Pythagoras applied N times and then added up. For each centered point, $\|\mathbf{c}_j\|^2$ splits into the part P keeps and the part it throws away; averaging the first over j is $\operatorname{var}(U)$, averaging the second is the fitting error of Theorem 24.10.2, and averaging the total is $\operatorname{tr}\mathbf{S}$ by Proposition 24.10.3 (d). Once (b) is available, (c) and (d) are just the theorem

already proved, divided by N.

Proof.

(a) The mean of $P\mathbf{x}_1, \dots, P\mathbf{x}_N$ is $\frac{1}{N}\sum_j P\mathbf{x}_j = P\bar{\mathbf{x}}$ by linearity of P, so the j-th centered projected point is $P\mathbf{x}_j - P\bar{\mathbf{x}} = P\mathbf{c}_j$. By Proposition 24.10.3 (d) applied to these N points, their total variance is $\frac{1}{N}\sum_j \|P\mathbf{c}_j\|^2 = \operatorname{var}(U)$.

(b) Fix j. The vectors $P\mathbf{c}_j$ and $\mathbf{c}_j - P\mathbf{c}_j$ are orthogonal, the first lying in U and the second in $U^\perp$ (Theorem 11.3.2), so Theorem 11.1.3 gives

$$\|\mathbf{c}_j\|^2 = \|P\mathbf{c}_j\|^2 + \|\mathbf{c}_j - P\mathbf{c}_j\|^2.$$

Summing over j and dividing by N: the left side is $\operatorname{tr}\mathbf{S}$ by Proposition 24.10.3 (d), the first term on the right is $\operatorname{var}(U)$, and the second is $\frac{1}{N}\sum_j d(\mathbf{c}_j, U)^2 = \frac{1}{N}D(\bar{\mathbf{x}} + U)$, as computed in the proof of Theorem 24.10.2.

(c) By (b), maximizing $\operatorname{var}(U)$ over $\dim U \le k$ is the same as minimizing $D(\bar{\mathbf{x}} + U)$, and Theorem 24.10.2 says the minimum is $\sum_{i>k}\sigma_i^2$, attained at U_0. Hence $\operatorname{var}(U) \le \operatorname{var}(U_0)$, and by (b) again,

$$\begin{aligned}\operatorname{var}(U_0) &= \operatorname{tr}\mathbf{S} - \frac{1}{N}\sum_{i>k}\sigma_i^2 \\ &= \frac{1}{N}\sum_i \sigma_i^2 - \frac{1}{N}\sum_{i>k}\sigma_i^2 = \frac{1}{N}\sum_{i\le k}\sigma_i^2.\end{aligned}$$

(d) If $\mathbf{X}_c \neq \mathbf{O}$ then some $\sigma_i > 0$, so $\operatorname{tr}\mathbf{S} = \frac{1}{N}\sum_i \sigma_i^2 > 0$ and the quotient is defined; dividing the value in (c) by it cancels the $1/N$. This proves the theorem. ∎

So "the first k principal directions explain $(\sum_{i\le k}\sigma_i^2)/(\sum_i \sigma_i^2)$ of the variance" is a theorem about the N given points, not a convention: the numerator is the variance of the projected cloud, the denominator is the variance of the original cloud, and part (b) is what makes the two add up with nothing left over.

Example. Four points, done exactly

Take the four points $(5,4), (4,1), (2,3), (1,0)$ of Example (Centering three data sets) (a). Find the mean, the sample covariance matrix, the principal directions, the best-fitting line, the total squared distance to it, and the fraction of variance it explains.

Solution. Mean and centering. $\bar{\mathbf{x}} = (3,2)$, and $\mathbf{X}_c$ has columns $(2,2), (1,-1), (-1,1), (-2,-2)$, as computed above.

Covariance. $\mathbf{X}_c\mathbf{X}_c^\top = \left(\begin{smallmatrix}10 & 6\\ 6 & 10\end{smallmatrix}\right)$, so $\mathbf{S} = \left(\begin{smallmatrix}5/2 & 3/2\\ 3/2 & 5/2\end{smallmatrix}\right)$. Since $\mathbf{X}_c\mathbf{X}_c^\top$ has the form $a\mathbf{I} + b\mathbf{J}$ it sends $(1,1)$ to $16(1,1)$ and $(1,-1)$ to $4(1,-1)$, so its eigenvalues are 16 and 4 with

$$\mathbf{u}_1 = \tfrac{1}{\sqrt{2}}(1,1), \qquad \mathbf{u}_2 = \tfrac{1}{\sqrt{2}}(1,-1).$$

By Proposition 24.10.3 (b), $\sigma_1^2 = 16$ and $\sigma_2^2 = 4$, that is $\sigma_1 = 4$ and $\sigma_2 = 2$, and the eigenvalues of $\mathbf{S}$ are 4 and 1. Check: $\operatorname{tr}\mathbf{S} = 5$ and $4 + 1 = 5$.

Total variance, two ways. The squared distances of the centered points from $\mathbf{0}$ are $8, 2, 2, 8$, of average 5. And $\frac{1}{N}\sum_i \sigma_i^2 = \frac{1}{4}(16 + 4) = 5$, as Proposition 24.10.3 (d) requires.

Best-fitting line. By Theorem 24.10.2 with $k = 1$ it is

$$A_0 = (3, 2) + \operatorname{span}\big((1, 1)\big),$$

the line $y = x - 1$, and the total squared distance is $\sigma_2^2 = 4$. Directly: $(5, 4)$ and $(1, 0)$ lie on the line, contributing 0; for $(4, 1)$ the centered point is $(1, -1)$, whose projection onto $\operatorname{span}\big((1, 1)\big)$ is $\mathbf{0}$, so its squared distance is 2; likewise 2 for $(2, 3)$. Total 4, as predicted.

Variance explained. $\operatorname{var}(U_0) = \frac{1}{4}\sigma_1^2 = 4$, out of $\operatorname{tr}\mathbf{S} = 5$: the first principal direction explains $16/20 = 80\%$ of the variance, and the remaining 20% is the average squared distance $4/4 = 1$ to the line. The two add to 5, which is Theorem 24.10.4 (b).

Warning. **Principal directions are not "the important directions": they change when you change the units.** Take the four centered points $(1, 0), (-1, 0), (0, 2), (0, -2)$. Then $\mathbf{X}_c\mathbf{X}_c^\top = \operatorname{diag}(2, 8)$, so the first principal direction is $\mathbf{e}_2$. Now suppose the first coordinate was measured in meters and we re-express it in units three times smaller, which multiplies row 1 by 3. The data are the same data; the new centered matrix is $\mathbf{D}\mathbf{X}_c$ with $\mathbf{D} = \operatorname{diag}(3, 1)$, and

$$(\mathbf{D}\mathbf{X}_c)(\mathbf{D}\mathbf{X}_c)^\top = \mathbf{D}\,\operatorname{diag}(2, 8)\,\mathbf{D} = \operatorname{diag}(18, 8),$$

so the first principal direction is now $\mathbf{e}_1$. It has turned by a right angle because of a choice of units. Nothing is wrong with the theorem — $\mathbf{D}\mathbf{X}_c$ really is a different matrix — but "the direction of greatest variance" is a statement about a coordinate system, not about the data alone. Rescaling each coordinate to variance 1 before starting is one common repair; it is a modeling decision, and it is not forced by anything proved here.

24.10.5 Computing it, and what is not proved

Two closing matters, one computational and one about scope.

Never form $\mathbf{X}_c\mathbf{X}_c^\top$. Everything above is stated in terms of $\mathbf{S}$, and $\mathbf{S}$ is the natural object to state it with; but forming $\mathbf{S}$ and then finding its eigenvalues is the worst available way to compute the principal directions. The reason is Theorem 16.8.4: passing from $\mathbf{A}$ to $\mathbf{A}^*\mathbf{A}$ squares the condition number, $\kappa_2(\mathbf{A}^*\mathbf{A}) = \kappa_2(\mathbf{A})^2$, and Section 3 of this chapter turns that into a count of lost digits — Theorem 24.3.2 says the product route loses about $2\log_{10}\kappa_2$ decimal digits where a backward stable factorization of $\mathbf{A}$ itself loses about $\log_{10}\kappa_2$. The same arithmetic applies with $\mathbf{A} = \mathbf{X}_c^\top$ — but only when that matrix has full column rank, which is the hypothesis of Theorem 16.8.4. That asks for $\operatorname{rank}\mathbf{X}_c = n$, and $\mathbf{X}_c\mathbf{1} = \mathbf{0}$ already forces $\operatorname{rank}\mathbf{X}_c \le N - 1$, so it fails whenever $N \le n$: for three records of ten measured quantities there is nothing to apply. The count that follows needs no hypothesis on the rank, and it is the one to remember.

Here is the count in this setting. Each entry of $\mathbf{X}_c\mathbf{X}_c^\top$ is an inner product of two rows, each of length N, so by Theorem 24.1.2 the computed product is $\mathbf{X}_c\mathbf{X}_c^\top + \mathbf{\Delta}$ with $\|\mathbf{\Delta}\|_2$ of order $c(n, N)\,u\,\sigma_1^2$, where u is the unit roundoff of Section 1 and c is a modest dimensional factor.

It is a factor and not a constant because the entrywise bound has to be aggregated over the whole $n \times n$ product, by the same count as in Section 3, where the corresponding step reads $\|\mathbf{C} - \mathbf{A}^*\mathbf{A}\| \le \gamma_m \|\mathbf{A}\|_F^2 \le n\gamma_m \|\mathbf{A}\|^2$: the length of the inner products sits inside γ, and passing from the Frobenius norm to the spectral norm puts their number in front. Here that makes $c(n, N)$ of size about nN.

If the implementation forms only the upper triangle and reflects it — an assumption about the computation, not something proved here — then $\mathbf{\Delta}$ is symmetric, and Corollary 17.3.6 says a symmetric perturbation of that size moves every eigenvalue by at most $\|\mathbf{\Delta}\|_2$. So every σ_i^2 below $c(n, N)\, u\, \sigma_1^2$ — that is, up to the factor $\sqrt{c}$, every σ_i below $\sqrt{u}\,\sigma_1$ — is indistinguishable from 0 in the computed product, while a backward stable factorization of $\mathbf{X}_c$ itself keeps it. Those are exactly the singular values a low-rank fit has to decide about. The right computation therefore works on $\mathbf{X}_c$ directly and never builds the product.

What this chapter does **not** do is assemble that computation. The ingredients are here — the Householder reductions of Section 3, the shifted **QR** iteration of Section 5 — but no algorithm for the singular values of a given matrix is written down or analyzed anywhere in this book. So the paragraph above pays the parenthesis of the last clause of Chapter 13 §11's promise, that a singular value decomposition is "never" computed by forming $\mathbf{A}^*\mathbf{A}$ and finding its eigenvalues; the clause itself, how one is actually computed for a large matrix, is left unpaid.

The finite population, and the boundary. Chapter 13 §11 said that the statistics needs "a probabilistic model, a reason to prefer squared error, and an argument that the sample answer says something about the population". Here is exactly what has been delivered.

A probabilistic model, for the finite population of the data themselves. Regard the N columns as a population of N equally likely outcomes, each of weight $1/N$. Then "expectation" means the average $\frac{1}{N}\sum_j$, which needs no measure theory and no limit, and the mean of the population is $\bar{\mathbf{x}}$. For two real-valued quantities f, g on it, with averages $\bar{f} = \frac{1}{N}\sum_j f(\mathbf{x}_j)$ and $\bar{g} = \frac{1}{N}\sum_j g(\mathbf{x}_j)$, the **covariance** of the population is

$$\operatorname{Cov}(f, g) \coloneqq \frac{1}{N}\sum_{j=1}^{N} \big(f(\mathbf{x}_j) - \bar{f}\big)\big(g(\mathbf{x}_j) - \bar{g}\big),$$

and the variance of f is $\operatorname{Cov}(f, f)$. With that reading, the variance of $\mathbf{x} \mapsto \mathbf{w}^\top\mathbf{x}$ is $\mathbf{w}^\top\mathbf{S}\mathbf{w}$, which for a **unit** $\mathbf{w}$ is the variance in the direction $\mathbf{w}$ of Definition 24.10.2, hence the Rayleigh quotient of Proposition 24.10.3 (c); and the covariance of $\mathbf{x} \mapsto x_a$ with $\mathbf{x} \mapsto x_b$ is s_{ab}, which is the displayed formula of Definition 24.10.2 read backwards. So $\mathbf{S}$ **is** the covariance matrix of that population, and Theorem 24.10.4 is a statement about it. This much is a theorem, and it is what was proved above.

A reason to prefer squared error. Three, none of them a proof that squared error is the right criterion, which is a modeling question. (i) It does not depend on the coordinate frame: an orthogonal change of coordinates is an isometry, so it leaves every distance $d(\mathbf{x}_j, A)$, hence $D(A)$, unchanged. The largest entry of the error has no such property — the matrix $\operatorname{diag}(1, 0)$ has largest entry 1, while the same matrix read in the frame $\frac{1}{\sqrt{2}}(1, \pm 1)$ is $\frac{1}{2}\left(\begin{smallmatrix} 1 & 1 \\ 1 & 1 \end{smallmatrix}\right)$, with largest entry $\frac{1}{2}$ — and Chapter 13 §11's warning shows that it really does prefer a different answer. (ii) It splits: Theorem 24.10.4 (b) says captured plus lost equals total, exactly, which is what makes a *fraction* of variance meaningful at all. (iii) It has a closed-form minimizer, computable by a single factorization (Theorem 13.11.3); for the sum of **unsquared** distances $\sum_j d(\mathbf{x}_j, A)$, nothing proved in this book produces a minimizer at all.

An argument that the sample says something about a larger population. This is **not** proved here, and it cannot be with the tools of this book. Such an argument requires a probability

model in which the N points are drawn from a distribution on $\mathbb{R}^n$, an expectation defined by an integral rather than a finite sum, and theorems saying that $\mathbf{S}$ converges to the distribution's covariance and that its eigenvectors converge to the distribution's principal directions. Every one of those belongs to probability and statistics, which this book does not develop. Chapter 13 §11's sentence promised it; the honest correction is that linear algebra supplies the finite-population model and the reasons for squared error, and hands the inference to a subject this book does not enter.

Remark. **The divisor N versus $N-1$.** Statistics usually defines the sample covariance with $1/(N-1)$, which presupposes $N \geq 2$, because that makes it an unbiased estimator of a population covariance — a statement inside the probability model just excluded, and one we therefore cannot prove. With the finite-population reading the divisor is N, and that is the convention of Definition 24.10.2, which makes sense for every $N \geq 1$. For $N \geq 2$ the choice changes nothing here: it multiplies $\mathbf{S}$ by $N/(N-1)$, which scales every eigenvalue by the same factor, changes no eigenvector, and leaves every ratio such as Theorem 24.10.4 (d) exactly as it was.

24.10.6 Exercises

24.10.6.1 A. Check your understanding

Exercise. A1

(a) Define the centered data matrix and the sample covariance matrix, and say what $\mathbf{w}^\top \mathbf{S}\mathbf{w}$ means for a unit vector $\mathbf{w}$.

(b) State Theorem 24.10.2, including the value of the minimum and the claim about the mean.

(c) True or false: the principal directions of a data set are the eigenvectors of $\mathbf{X}\mathbf{X}^\top$. Justify your answer.

(d) True or false: if $N = 3$ points in $\mathbb{R}^{10}$ are given, then $\mathbf{S}$ has rank at most 2. Justify your answer.

(e) Why should a computation of the principal directions avoid forming $\mathbf{X}_c\mathbf{X}_c^\top$?

Solution.

(a) $\mathbf{X}_c = \mathbf{X} - \bar{\mathbf{x}}\mathbf{1}^\top$, the data matrix with the mean column subtracted from every column, and $\mathbf{S} = \frac{1}{N}\mathbf{X}_c\mathbf{X}_c^\top$ (Definition 24.10.1, Definition 24.10.2). For a unit $\mathbf{w}$, $\mathbf{w}^\top\mathbf{S}\mathbf{w}$ is the variance of the N numbers $\mathbf{w}^\top\mathbf{x}_j$, that is the spread of the data in the direction $\mathbf{w}$.

(b) Among all affine subspaces of $\mathbb{R}^n$ of dimension at most $k \leq \min(n, N)$, the total squared distance $\sum_j d(\mathbf{x}_j, A)^2$ is smallest at $A_0 = \bar{\mathbf{x}} + \operatorname{span}(\mathbf{u}_1, \dots, \mathbf{u}_k)$, the minimum value is $\sum_{i>k} \sigma_i^2$ with σ_i the singular values of $\mathbf{X}_c$, and every minimizer contains $\bar{\mathbf{x}}$.

(c) False in general; they are the eigenvectors of $\mathbf{X}_c\mathbf{X}_c^\top$. For the four points of Example (Four points, done exactly), $\mathbf{X}\mathbf{X}^\top = \left(\begin{smallmatrix} 46 & 30 \\ 30 & 26 \end{smallmatrix}\right)$, whose eigenvalues are $36 \pm 10\sqrt{10}$ and whose top eigenvector is not $(1,1)$, while the top principal direction is. The two **matrices** agree exactly when $\bar{\mathbf{x}} = \mathbf{0}$, since $\mathbf{X}\mathbf{X}^\top = \mathbf{X}_c\mathbf{X}_c^\top + N\bar{\mathbf{x}}\bar{\mathbf{x}}^\top$. Their **eigenvectors** can agree without that: for the two points $(0,0)$ and $(2,0)$ the mean is $(1,0) \neq \mathbf{0}$, yet $\mathbf{X}\mathbf{X}^\top = \operatorname{diag}(4,0)$ and $\mathbf{X}_c\mathbf{X}_c^\top = \operatorname{diag}(2,0)$ have the same eigenvectors.

(d) True. $\operatorname{rank}\mathbf{S} = \operatorname{rank}\mathbf{X}_c$ — since $\operatorname{nul}(\mathbf{X}_c^\top) = \operatorname{nul}(\mathbf{X}_c\mathbf{X}_c^\top)$, because $\mathbf{X}_c\mathbf{X}_c^\top\mathbf{w} = \mathbf{0}$ gives $\|\mathbf{X}_c^\top\mathbf{w}\|^2 = \mathbf{w}^\top\mathbf{X}_c\mathbf{X}_c^\top\mathbf{w} = 0$ — and $\mathbf{X}_c\mathbf{1} = \mathbf{0}$ with $\mathbf{X}_c$ having 3 columns forces $\operatorname{rank}\mathbf{X}_c \le 2$. Centering costs at most one from the rank and sometimes nothing:

$$\operatorname{rank}\mathbf{X} - 1 \;\le\; \operatorname{rank}\mathbf{X}_c \;\le\; \operatorname{rank}\mathbf{X},$$

the right inequality because $\mathbf{X}_c = \mathbf{X}(\mathbf{I}_N - \frac{1}{N}\mathbf{1}\mathbf{1}^\top)$ is a product (Theorem 3.5.9), the left because $\mathbf{X} = \mathbf{X}_c + \bar{\mathbf{x}}\mathbf{1}^\top$ adds to it a matrix of rank at most 1 (Exercise (C1: Rank of a sum)). The drop need not happen: for $\mathbf{x}_1 = (1,0)$, $\mathbf{x}_2 = (2,0)$, $\mathbf{x}_3 = (0,1)$ the mean is $(1, \frac{1}{3}) \neq \mathbf{0}$, yet $\operatorname{rank}\mathbf{X} = \operatorname{rank}\mathbf{X}_c = 2$.

(e) Because forming the product destroys the small singular values. Each entry of $\mathbf{X}_c\mathbf{X}_c^\top$ is an inner product, so the computed product is $\mathbf{X}_c\mathbf{X}_c^\top + \mathbf{\Delta}$ with $\|\mathbf{\Delta}\|_2$ of order $c(n,N)\,u\,\sigma_1^2$ (Theorem 24.1.2); if $\mathbf{\Delta}$ is symmetric, Corollary 17.3.6 moves every eigenvalue by at most that, so every σ_i below about $\sqrt{u}\,\sigma_1$ becomes indistinguishable from 0, while a backward stable factorization of $\mathbf{X}_c$ itself keeps it — and those are exactly the singular values a low-rank fit has to decide about. The conditioning statement $\kappa_2(\mathbf{X}_c\mathbf{X}_c^\top) = \kappa_2(\mathbf{X}_c)^2$ says the same thing in digits (Theorem 16.8.4, Theorem 24.3.2), but it needs $\mathbf{X}_c^\top$ to have full column rank, which fails in part (d) above; the Weyl count needs no hypothesis on the rank.

24.10.6.2 B. Practice

Exercise. B1: A four-point cloud

The points $(1,1), (3,5), (5,3), (7,7)$ are given in $\mathbb{R}^2$.

(a) Compute $\bar{\mathbf{x}}$, $\mathbf{X}_c$, $\mathbf{X}_c\mathbf{X}_c^\top$, and $\mathbf{S}$.

(b) Find the singular values of $\mathbf{X}_c$ and the principal directions.

(c) Find the best-fitting line, the total squared distance to it, and the fraction of variance it explains. Verify the total squared distance directly.

Solution.

(a) The sum of the four points is $(16,16)$, so $\bar{\mathbf{x}} = (4,4)$ and

$$\mathbf{X}_c = \begin{pmatrix} -3 & -1 & 1 & 3 \\ -3 & 1 & -1 & 3 \end{pmatrix}, \qquad \mathbf{X}_c\mathbf{X}_c^\top = \begin{pmatrix} 20 & 16 \\ 16 & 20 \end{pmatrix},$$

so $\mathbf{S} = \left(\begin{smallmatrix} 5 & 4 \\ 4 & 5 \end{smallmatrix}\right)$.

(b) $\mathbf{X}_c\mathbf{X}_c^\top$ sends $(1,1)$ to $36(1,1)$ and $(1,-1)$ to $4(1,-1)$, so $\sigma_1^2 = 36$, $\sigma_2^2 = 4$, that is $\sigma_1 = 6$, $\sigma_2 = 2$, with $\mathbf{u}_1 = \frac{1}{\sqrt{2}}(1,1)$ and $\mathbf{u}_2 = \frac{1}{\sqrt{2}}(1,-1)$. The eigenvalues of $\mathbf{S}$ are 9 and 1, which are $\sigma_i^2/4$ (Proposition 24.10.3 (b)).

(c) By Theorem 24.10.2 the best line is $(4,4) + \operatorname{span}\big((1,1)\big)$, that is $y = x$, with total squared distance $\sigma_2^2 = 4$. By Theorem 24.10.4 (d) it explains $36/40 = 90\%$ of the variance. Directly: $(1,1)$ and $(7,7)$ lie on $y = x$; the centered points $(-1,1)$ and $(1,-1)$ both project to $\mathbf{0}$ on $\operatorname{span}\big((1,1)\big)$, so each contributes 2. Total 4.

Exercise. B2: Variance direction by direction

For the data of Exercise B1, compute the variance in the directions $\mathbf{e}_1$, $\mathbf{e}_2$, $\frac{1}{\sqrt{2}}(1,1)$, and $\frac{1}{\sqrt{2}}(1,-1)$, and say which of them is largest and why that was predictable.

Solution. With $\mathbf{S} = \left(\begin{smallmatrix} 5 & 4 \\ 4 & 5 \end{smallmatrix}\right)$ and Definition 24.10.2, the variance in a unit direction $\mathbf{w}$ is $\mathbf{w}^\top \mathbf{S}\mathbf{w}$:

$$\mathbf{e}_1 \mapsto 5, \quad \mathbf{e}_2 \mapsto 5, \quad \tfrac{(1,1)}{\sqrt{2}} \mapsto 9, \quad \tfrac{(1,-1)}{\sqrt{2}} \mapsto 1.$$

The largest is 9, in the direction $\mathbf{u}_1$. This was predictable from Proposition 24.10.3 (c) and Proposition 17.1.1: the variance in a unit direction is the Rayleigh quotient of $\mathbf{S}$, whose maximum is $\lambda_1(\mathbf{S}) = 9$ and is attained at a unit eigenvector for λ_1. Note also $5+5 = 9+1 = \operatorname{tr}\mathbf{S}$: the total variance is the same however the orthonormal frame is chosen.

Exercise. B3: Changing the units

The centered data matrix of a two-coordinate data set is $\mathbf{X}_c = \left(\begin{smallmatrix} 3 & -3 & 0 & 0 \\ 0 & 0 & 4 & -4 \end{smallmatrix}\right)$. The second coordinate is then re-expressed in units c times smaller, which multiplies row 2 by $c > 0$.

(a) Find the principal directions before the change.

(b) Find the value of c at which the first principal direction switches from one coordinate axis to the other, and say what happens exactly at that value.

Solution.

(a) $\mathbf{X}_c\mathbf{X}_c^\top = \operatorname{diag}(18, 32)$, so the principal directions are $\mathbf{u}_1 = \mathbf{e}_2$ (variance $32/4 = 8$) and $\mathbf{u}_2 = \mathbf{e}_1$ (variance $18/4 = 4.5$).

(b) After the change the centered matrix is $\mathbf{D}\mathbf{X}_c$ with $\mathbf{D} = \operatorname{diag}(1, c)$, and $(\mathbf{D}\mathbf{X}_c)(\mathbf{D}\mathbf{X}_c)^\top = \operatorname{diag}(18, 32c^2)$. The first principal direction is $\mathbf{e}_2$ when $32c^2 > 18$, that is $c > 3/4$, and $\mathbf{e}_1$ when $c < 3/4$. At $c = 3/4$ the two eigenvalues are equal, $\sigma_1 = \sigma_2$, and **every** direction is a direction of greatest variance: the principal directions are not determined at all, and the best-fitting line of Theorem 24.10.2 is not unique, exactly as in the Quick check above.

24.10.6.3 C. Going deeper

Exercise. C1: The mean is forced, and the projection is not a shortcut

Let $\mathbf{x}_1, \dots, \mathbf{x}_N \in \mathbb{R}^n$ with mean $\bar{\mathbf{x}}$.

(a) Prove that the point $\mathbf{p} \in \mathbb{R}^n$ minimizing $\sum_j \|\mathbf{x}_j - \mathbf{p}\|^2$ is unique and equals $\bar{\mathbf{x}}$. (This is the case $k = 0$, which Theorem 24.10.2 excludes, so it has to be proved rather than quoted.)

(b) Hence deduce that for every subspace U, the affine subspace $\mathbf{p} + U$ minimizing D over translates of U is $\bar{\mathbf{x}} + U$.

(c) Give an example of data and a subspace U for which $\bar{\mathbf{x}} + U$ is **not** the best-fitting affine

subspace of its dimension, so that the choice of U still matters.

Solution.

(a) Apply Lemma 24.10.1 (b) with $U = \{\mathbf{0}\}$, so that $Q = \mathbf{I}_n$:

$$\sum_j \|\mathbf{x}_j - \mathbf{p}\|^2 = \sum_j \|\mathbf{x}_j - \bar{\mathbf{x}}\|^2 + N\|\bar{\mathbf{x}} - \mathbf{p}\|^2.$$

The first term does not depend on $\mathbf{p}$ and the second is ≥ 0, vanishing exactly when $\mathbf{p} = \bar{\mathbf{x}}$. So the minimum is attained at $\bar{\mathbf{x}}$ and nowhere else.

(b) By Lemma 24.10.1 (a), $D(\mathbf{p}+U) = \sum_j \|Q(\mathbf{x}_j - \mathbf{p})\|^2$, and by part (b) of that lemma this exceeds $D(\bar{\mathbf{x}}+U)$ unless $\bar{\mathbf{x}} \in \mathbf{p}+U$ — in which case $\mathbf{p}+U = \bar{\mathbf{x}}+U$, the two cosets having a common point and the same direction space. So the minimizing translate is $\bar{\mathbf{x}} + U$, and as a set it is unique.

(c) Take the four points of Example (Four points, done exactly) and $U = \operatorname{span}(\mathbf{e}_1)$. Then $\bar{\mathbf{x}}+U$ is the horizontal line $y = 2$, and the centered points $(2,2), (1,-1), (-1,1), (-2,-2)$ are at squared distances $4, 1, 1, 4$ from U, total 10. The best line is $\bar{\mathbf{x}} + \operatorname{span}((1,1))$, with total 4. Both pass through the mean; only one has the right direction. Choosing the translate is free, choosing the direction space is the whole problem.

Exercise. C2: Total variance is a trace, and what that buys

Let $\mathbf{x}_1, \dots, \mathbf{x}_N \in \mathbb{R}^n$ with sample covariance $\mathbf{S}$, and let $\mathbf{Q} \in M_n(\mathbb{R})$ be orthogonal.

(a) Prove that the data $\mathbf{Q}\mathbf{x}_1, \dots, \mathbf{Q}\mathbf{x}_N$ have sample covariance $\mathbf{Q}\mathbf{S}\mathbf{Q}^\top$, and deduce that the total variance $\operatorname{tr}\mathbf{S}$ is unchanged.

(b) Prove that for any orthonormal basis $(\mathbf{q}_1, \dots, \mathbf{q}_n)$ of $\mathbb{R}^n$, the sum of the variances of the data in the n directions $\mathbf{q}_i$ equals $\operatorname{tr}\mathbf{S}$.

(c) Deduce that no orthonormal frame can make every coordinate variance small: some $\mathbf{q}_i$ always has variance at least $\operatorname{tr}\mathbf{S}/n$. For the data of Example (Four points, done exactly), which frames attain that bound?

Solution.

(a) The mean of the transformed data is $\frac{1}{N}\sum_j \mathbf{Q}\mathbf{x}_j = \mathbf{Q}\bar{\mathbf{x}}$, so the transformed centered matrix is $\mathbf{Q}\mathbf{X} - \mathbf{Q}\bar{\mathbf{x}}\mathbf{1}^\top = \mathbf{Q}\mathbf{X}_c$. Hence its sample covariance is $\frac{1}{N}(\mathbf{Q}\mathbf{X}_c)(\mathbf{Q}\mathbf{X}_c)^\top = \mathbf{Q}\mathbf{S}\mathbf{Q}^\top$. By Theorem 0.9.6, $\operatorname{tr}(\mathbf{Q}\mathbf{S}\mathbf{Q}^\top) = \operatorname{tr}(\mathbf{S}\mathbf{Q}^\top\mathbf{Q}) = \operatorname{tr}\mathbf{S}$, since $\mathbf{Q}^\top\mathbf{Q} = \mathbf{I}_n$.

(b) By Definition 24.10.2 the variance in the unit direction $\mathbf{q}_i$ is $\mathbf{q}_i^\top\mathbf{S}\mathbf{q}_i$. Let $\mathbf{Q}$ be the orthogonal matrix with columns $\mathbf{q}_1, \dots, \mathbf{q}_n$. Then $\mathbf{q}_i^\top\mathbf{S}\mathbf{q}_i$ is the (i,i) entry of $\mathbf{Q}^\top\mathbf{S}\mathbf{Q}$, so the sum over i is $\operatorname{tr}(\mathbf{Q}^\top\mathbf{S}\mathbf{Q}) = \operatorname{tr}\mathbf{S}$ by part (a) applied to $\mathbf{Q}^\top$.

(c) A sum of n numbers equal to $\operatorname{tr}\mathbf{S}$ has a term at least the average $\operatorname{tr}\mathbf{S}/n$, since otherwise the sum would be $< \operatorname{tr}\mathbf{S}$. For Example (Four points, done exactly), $\operatorname{tr}\mathbf{S} = 5$ and the bound is $5/2$; the variances in a frame $(\mathbf{q}_1, \mathbf{q}_2)$ are $\mathbf{q}_1^\top\mathbf{S}\mathbf{q}_1$ and $\mathbf{q}_2^\top\mathbf{S}\mathbf{q}_2$, which sum to 5, so both equal $5/2$ exactly when $\mathbf{q}_1^\top\mathbf{S}\mathbf{q}_1 = 5/2$. Writing $\mathbf{q}_1 = a\mathbf{u}_1 + b\mathbf{u}_2$ with $a^2+b^2 = 1$ and using $\mathbf{S}\mathbf{u}_1 = 4\mathbf{u}_1$, $\mathbf{S}\mathbf{u}_2 = \mathbf{u}_2$, this reads $4a^2+b^2 = 5/2$, and subtracting $a^2+b^2 = 1$ gives $3a^2 = 3/2$, that is $a^2 = b^2 = 1/2$. So the bound is attained exactly by the frames whose vectors split their weight equally between $\mathbf{u}_1$ and $\mathbf{u}_2$ — for instance

the standard frame $(\mathbf{e}_1, \mathbf{e}_2)$, since $\mathbf{e}_1 = \frac{1}{\sqrt{2}}(\mathbf{u}_1 + \mathbf{u}_2)$; its two variances are indeed $5/2$ each. The extreme case of the spread is visible only in the principal frame; an arbitrary frame hides it.

24.11 Graph Laplacians

Chapter 3 built the incidence matrix of a network, used it to solve flow problems, and then set one object aside with a promise: the matrix $\mathbf{N}\mathbf{N}^\top$, the graph Laplacian, "returns in Chapter 24". Chapter 7 left a second debt at the same address. Its exercise Exercise (C1: Counting spanning trees) computed, for one four-vertex network, that $\det Ł_0$ equals the number of spanning trees, and named the general statement the matrix-tree theorem, leaving a proof "which Chapter 24 completes". This section pays both debts. It also does something the two earlier chapters could not: with the spectral theory of Chapters 12 and 17 in hand, we can read the eigenvalues of Ł and find the shape of the graph in them.

Throughout, a **graph** means what Chapter 3 §07 called a network with its orientation forgotten: a finite set of vertices $1, \dots, m$ and a finite set of edges $1, \dots, k$, each edge joining two **distinct** vertices. Two edges may join the same pair. An **orientation** chooses, for each edge e, a tail and a head, and turns the graph back into one of Chapter 3's networks with an incidence matrix $\mathbf{N} \in M_{m\times k}(\mathbb{R})$ (Definition 3.7.1).

24.11.1 The Laplacian of a graph

The matrix $\mathbf{N}\mathbf{N}^\top$ has already appeared three times without being studied: named and postponed in the closing sentence of Chapter 3 §07's flow-network subsection; implicitly in Exercise (C3: Networks in several pieces), where the kernel of $\mathbf{N}^\top$ — which, as we shall see, is the kernel of $\mathbf{N}\mathbf{N}^\top$ — turned out to be spanned by the indicator vectors of the components; and in Exercise (C1: Counting spanning trees), where its determinant counted something combinatorial. A product that keeps coming back deserves a name.

The Laplacian of a graph is the matrix that turns a labeling of the vertices into the total squared disagreement across the edges.

Definition 24.11.1. Graph Laplacian

Let G be a graph with vertices $1, \dots, m$ and $k \geq 1$ edges, and fix **any** orientation of G, with incidence matrix $\mathbf{N} \in M_{m\times k}(\mathbb{R})$. The **Laplacian** of G is

$$Ł := \mathbf{N}\mathbf{N}^\top \in M_m(\mathbb{R}).$$

In words: the (u, v)-entry of Ł is the dot product of row u and row v of $\mathbf{N}$, that is, the sum over all edges of the product of the two incidence entries. The definition names an orientation, and the first thing to check is that the orientation leaves no trace.

Write $\deg(v)$ for the number of edges meeting the vertex v, and a_{uv} for the number of edges joining u to v when $u \neq v$, with $a_{vv} = 0$. Let $\mathbf{D} = \operatorname{diag}(\deg(1), \dots, \deg(m))$ be the **degree matrix** and $\mathbf{A} = (a_{uv})$ the **adjacency matrix**. Neither mentions an orientation.

Proposition 24.11.1. First Properties of the Laplacian

Let G be a graph on m vertices with $k \geq 1$ edges and Laplacian Ł. Then:

(a) $Ł = \mathbf{D} - \mathbf{A}$; in particular Ł is symmetric and does **not** depend on the orientation chosen in Definition 24.11.1.

(b) For every $\mathbf{x} \in \mathbb{R}^m$,

$$\mathbf{x}^\top Ł \mathbf{x} = \sum_{e=1}^{k} (x_{u_e} - x_{v_e})^2,$$

where u_e and v_e are the two ends of the edge e. Consequently $Ł \succeq \mathbf{O}$.

(c) $Ł\mathbf{1} = \mathbf{0}$, where $\mathbf{1} = (1, \ldots, 1) \in \mathbb{R}^m$.

Idea. Everything is one line of $\mathbf{N}$. Part (a) reads the product $\mathbf{N}\mathbf{N}^\top$ entry by entry, using that a column of $\mathbf{N}$ has exactly one 1 and one -1. Part (b) is the observation that $\mathbf{x}^\top \mathbf{N}\mathbf{N}^\top \mathbf{x}$ is the squared length of $\mathbf{N}^\top \mathbf{x}$, and Chapter 3 already identified $\mathbf{N}^\top \mathbf{x}$ as the vector of differences across the edges. Part (c) is (b) applied to a constant vector, or the fact that each column of $\mathbf{N}$ sums to zero.

Proof.

(a) By Theorem 0.9.2, $(\mathbf{N}\mathbf{N}^\top)_{uv} = \sum_e \mathbf{N}_{ue}\mathbf{N}_{ve}$. If $u = v$, then $\mathbf{N}_{ue}^2$ is 1 when e meets u and 0 otherwise (Definition 3.7.1), so the sum is $\deg(u)$. If $u \neq v$, then $\mathbf{N}_{ue}\mathbf{N}_{ve} \neq 0$ only for an edge e with one end u and the other v, and such an edge contributes $1 \cdot (-1) = -1$ whichever way it is oriented; so the sum is $-a_{uv}$. Hence $Ł = \mathbf{D} - \mathbf{A}$. Both $\mathbf{D}$ and $\mathbf{A}$ are determined by G alone, and $\mathbf{A}$ is symmetric because $a_{uv} = a_{vu}$, so $Ł^\top = Ł$.

(b) As in the proof of Proposition 3.7.4, entry e of $\mathbf{N}^\top \mathbf{x}$ is $x_{\mathrm{tail}(e)} - x_{\mathrm{head}(e)}$, which is $\pm(x_{u_e} - x_{v_e})$. Therefore

$$\mathbf{x}^\top Ł \mathbf{x} = \mathbf{x}^\top \mathbf{N}\mathbf{N}^\top \mathbf{x} = (\mathbf{N}^\top \mathbf{x})^\top (\mathbf{N}^\top \mathbf{x}) = \|\mathbf{N}^\top \mathbf{x}\|^2,$$

and expanding the squared length coordinate by coordinate gives the stated sum, the signs disappearing because each term is squared. For the standard real inner product, $\langle Ł\mathbf{x}, \mathbf{x} \rangle = \mathbf{x}^\top Ł \mathbf{x}$, which is a sum of squares of real numbers and hence ≥ 0; and Ł is symmetric by (a). So $Ł \succeq \mathbf{O}$ by Definition 13.1.1.

(c) Each column of $\mathbf{N}$ has exactly one 1 and one -1, so $\mathbf{N}^\top \mathbf{1} = \mathbf{0}$, and $Ł\mathbf{1} = \mathbf{N}(\mathbf{N}^\top \mathbf{1}) = \mathbf{0}$.

■

Part (b) is the whole reason the Laplacian is useful. A vector $\mathbf{x} \in \mathbb{R}^m$ is a **labeling** of the vertices by numbers, and $\mathbf{x}^\top Ł \mathbf{x}$ measures how badly the labeling disagrees across the edges: it is zero exactly when the labeling is locally constant, and large when neighbors receive very different numbers. Every statement in this section is a consequence of that one reading.

Example. A Laplacian computed twice

Let G be the four-vertex graph of Exercise (C1: Counting spanning trees): vertices $1, 2, 3, 4$ and edges $e_1 = \{1, 2\}$, $e_2 = \{2, 3\}$, $e_3 = \{3, 4\}$, $e_4 = \{4, 1\}$, $e_5 = \{1, 3\}$. Compute Ł from $\mathbf{D} - \mathbf{A}$ and from $\mathbf{N}\mathbf{N}^\top$, and check them against each other.

Solution. The degrees are $\deg(1) = 3$, $\deg(2) = 2$, $\deg(3) = 3$, $\deg(4) = 2$, and every pair of vertices is joined by at most one edge, the missing pair being $\{2, 4\}$. Hence

$$\mathbf{D} - \mathbf{A} = \begin{pmatrix} 3 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 2 \end{pmatrix} - \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{pmatrix} = \begin{pmatrix} 3 & -1 & -1 & -1 \\ -1 & 2 & -1 & 0 \\ -1 & -1 & 3 & -1 \\ -1 & 0 & -1 & 2 \end{pmatrix}.$$

Orienting the edges as $1 \to 2$, $2 \to 3$, $3 \to 4$, $4 \to 1$, $1 \to 3$ gives the incidence matrix of Exercise (C1: Counting spanning trees),

$$\mathbf{N} = \begin{pmatrix} 1 & 0 & 0 & -1 & 1 \\ -1 & 1 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & -1 \\ 0 & 0 & -1 & 1 & 0 \end{pmatrix},$$

whose row 1 has squared length $1 + 1 + 1 = 3$, and whose rows 1 and 2 have dot product $1 \cdot (-1) = -1$. Continuing entry by entry reproduces the matrix above, so the two computations agree. Reversing the orientation of the edge $1 \to 2$ changes the first two rows of $\mathbf{N}$ in their first entry only, from 1 and -1 to -1 and 1; the diagonal entries are unchanged because they are sums of squares, and the $(1, 2)$-entry is still $(-1) \cdot 1 = -1$.

Warning. **Take the oriented incidence matrix, not the unoriented one.** If one replaces the -1 in Definition 3.7.1 by $+1$, obtaining the matrix $\mathbf{M}$ with a 1 at both ends of each edge, then $\mathbf{M}\mathbf{M}^\top = \mathbf{D} + \mathbf{A}$, not $\mathbf{D} - \mathbf{A}$. That matrix is also symmetric and positive semidefinite, but it is a different matrix with a different kernel: for the triangle, $\mathbf{D} - \mathbf{A}$ is singular while $\mathbf{D} + \mathbf{A}$ has determinant 4. The signs are what make Proposition 24.11.1 (b) a sum of **differences**, and the sign convention returns in Lemma 24.11.6, where it is load-bearing again.

Quick check. The graph of Example (A Laplacian computed twice) has the labeling $\mathbf{x} = (1, 1, 0, 0)$. Compute $\mathbf{x}^\top Ł\mathbf{x}$ from the edge formula, and say which edges contribute.

Solution. By Proposition 24.11.1 (b), each edge contributes the square of the difference of the labels at its ends. The edges $\{1, 2\}$ and $\{3, 4\}$ join equal labels and contribute 0; the edges $\{2, 3\}$, $\{4, 1\}$, and $\{1, 3\}$ each join a 1 to a 0 and contribute 1. So $\mathbf{x}^\top Ł\mathbf{x} = 3$, the number of edges crossing between $\{1, 2\}$ and $\{3, 4\}$. Directly: $Ł\mathbf{x} = (3 - 1, -1 + 2, -1 - 1, -1 + 0) = (2, 1, -2, -1)$ and $\mathbf{x}^\top Ł\mathbf{x} = 2 + 1 = 3$.

24.11.2 The kernel counts the components

Chapter 3's Exercise (C3: Networks in several pieces) computed $\operatorname{nul}(\mathbf{N}^\top)$ for a network in several pieces. That was an exercise, so we prove what we need here, in the form the spectral theory will use. Recall from Exercise (C3: Networks in several pieces) the notion we need: two vertices

are **linked** if some sequence of vertices joins them with consecutive vertices sharing an edge, and the equivalence classes of this relation are the **components** of G. That "linked" is an equivalence relation (Definition 0.5.2) takes one line: it is reflexive by the sequence of length zero, symmetric because a sequence reversed is again one, and transitive because two sequences may be concatenated. So the components partition the vertices (Theorem 0.5.2), which is what Step 3 below uses.

Theorem 24.11.2. The Kernel of the Laplacian

Let G be a graph on m vertices with $k \geq 1$ edges, with components $C_1, \ldots, C_c$, and let $\mathbf{1}_{C_r} \in \mathbb{R}^m$ be the vector with entry 1 at the vertices of C_r and 0 elsewhere. Then

$$\operatorname{nul}(Ł) = \operatorname{span}(\mathbf{1}_{C_1}, \ldots, \mathbf{1}_{C_c}),$$

this list is a basis, and the eigenvalue 0 of Ł has multiplicity exactly c. In particular Ł is singular, and $\operatorname{nul}(Ł) = \operatorname{span}(\mathbf{1})$ exactly when G is connected.

Idea. For a positive semidefinite matrix, lying in the kernel and making the quadratic form vanish are the same condition, and Proposition 24.11.1 (b) turns the second one into a statement with no linear algebra in it at all: the labeling is constant across every edge. Constant across every edge propagates along chains of edges, so it means constant on each component, which is exactly what the indicator vectors describe.

Proof. **Step 1: $\mathbf{x} \in \operatorname{nul}(Ł)$ if and only if $\mathbf{x}^\top Ł \mathbf{x} = 0$.** If $Ł\mathbf{x} = \mathbf{0}$, then $\mathbf{x}^\top Ł \mathbf{x} = 0$. Conversely, if $\mathbf{x}^\top Ł \mathbf{x} = 0$, then $\|\mathbf{N}^\top \mathbf{x}\|^2 = 0$ by the display in the proof of Proposition 24.11.1 (b), so $\mathbf{N}^\top \mathbf{x} = \mathbf{0}$ and $Ł\mathbf{x} = \mathbf{N}(\mathbf{N}^\top \mathbf{x}) = \mathbf{0}$.

Step 2: $\mathbf{x}^\top Ł \mathbf{x} = 0$ if and only if $\mathbf{x}$ is constant on each component. By Proposition 24.11.1 (b), $\mathbf{x}^\top Ł \mathbf{x}$ is a sum of squares of real numbers, so it vanishes if and only if every term does, that is, if and only if $x_u = x_v$ for every edge with ends u and v. Suppose this holds and let u, w lie in the same component. By the definition of linked there is a sequence $u = w_0, w_1, \ldots, w_t = w$ with consecutive vertices joined by an edge, so $x_u = x_{w_1} = \cdots = x_w$. Hence $\mathbf{x}$ is constant on each component. Conversely, if $\mathbf{x}$ is constant on each component, then the two ends of any edge lie in the same component and carry the same label, so every term vanishes.

Step 3: the indicators are a basis. By Steps 1 and 2, $\mathbf{x} \in \operatorname{nul}(Ł)$ if and only if $\mathbf{x} = \sum_{r=1}^{c} \lambda_r \mathbf{1}_{C_r}$, where λ_r is the common value of $\mathbf{x}$ on C_r; so the list spans $\operatorname{nul}(Ł)$. It is linearly independent: if $\sum_r \lambda_r \mathbf{1}_{C_r} = \mathbf{0}$, then reading the entry at any vertex of C_r gives $\lambda_r = 0$, since the components are disjoint and only the r-th vector is nonzero there. Hence $\dim \operatorname{nul}(Ł) = c$.

Step 4: multiplicity. Ł is symmetric by Proposition 24.11.1 (a), so by Corollary 12.5.2 it is orthogonally diagonalizable, and the number of times 0 occurs in its eigenvalue list is the dimension of its 0-eigenspace, which is $\operatorname{nul}(Ł)$. That dimension is $c \geq 1$, so Ł is singular. Finally, G is connected exactly when $c = 1$, and then C_1 is the whole vertex set and $\mathbf{1}_{C_1} = \mathbf{1}$.

■

This is the first of the two theorems that the Laplacian is for. Notice how little it costs: the

entire combinatorial content sits in Step 2, and the only property of Ł used is that its quadratic form is a sum of squares over the edges.

24.11.3 The Fiedler value

Theorem 24.11.2 says that 0 is always an eigenvalue and that its multiplicity detects whether the graph falls apart. The natural next question is quantitative: if the graph is connected, *how* connected is it? The eigenvalue just above 0 answers this, and the reason is Proposition 24.11.1 (b) again.

Because Ł $\succeq \mathbf{O}$, all its eigenvalues are ≥ 0 (Theorem 13.1.4 (b)), and it is convenient here to list them **increasingly**, against the book's usual convention. Write

$$\mu_1(Ł) \leq \mu_2(Ł) \leq \cdots \leq \mu_m(Ł)$$

for the eigenvalues of Ł with multiplicity in increasing order, so that $\mu_j(Ł) = \lambda_{m+1-j}(Ł)$ in the decreasing notation of Chapter 17. By Theorem 24.11.2, $\mu_1(Ł) = 0$.

Definition 24.11.2. Fiedler value

Let G be a graph on $m \geq 2$ vertices with at least one edge. The **Fiedler value**, or **algebraic connectivity**, of G is the **second smallest** eigenvalue $\mu_2(Ł)$ of its Laplacian, counted with multiplicity.

For the complete graph K_m, in which every pair of distinct vertices is joined by one edge, Proposition 24.11.1 (a) gives Ł $= m\mathbf{I} - \mathbf{J}$, where $\mathbf{J}$ is the all-ones matrix. Every row of $\mathbf{J}$ equals $\mathbf{1}^\top$, so $\operatorname{rank}\mathbf{J} = 1$ and 0 is an eigenvalue of $\mathbf{J}$ of multiplicity $m-1$ by Theorem 3.5.7; and $\mathbf{J1} = m\mathbf{1}$ supplies the last eigenvalue m. Hence Ł has eigenvalues 0 once and m with multiplicity $m-1$, and the Fiedler value of K_m is m. For a graph in two or more pieces it is 0, by Theorem 24.11.2. A single edge on two vertices has Laplacian $\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$ and Fiedler value 2. The degenerate case $m = 1$ is excluded because there is no second eigenvalue.

Proposition 24.11.3. Positive Fiedler Value Means Connected

Let G be a graph on $m \geq 2$ vertices with at least one edge. Then $\mu_2(Ł) > 0$ if and only if G is connected.

Proof. The eigenvalues of Ł are ≥ 0, so $\mu_2(Ł) = 0$ if and only if the eigenvalue 0 occurs at least twice in the list, that is, if and only if its multiplicity c is at least 2 (Theorem 24.11.2). That happens exactly when G has two or more components, that is, exactly when G is not connected. ∎

The number $\mu_2(Ł)$ is therefore a graded version of connectedness, and the grading has a meaning that can be read off directly.

Proposition 24.11.4. The Fiedler Value as a Minimum

Let G be a graph on $m \geq 2$ vertices with at least one edge. Then

$$\mu_2(Ł) = \min\left\{ \frac{\mathbf{x}^\top Ł \mathbf{x}}{\mathbf{x}^\top \mathbf{x}} \;:\; \mathbf{x} \in \mathbb{R}^m,\ \mathbf{x} \neq \mathbf{0},\ \mathbf{x}^\top \mathbf{1} = 0 \right\},$$

and the minimum is attained at an eigenvector for $\mu_2(Ł)$.

Idea. The quantity being minimized is the Rayleigh quotient $R_Ł$ of Chapter 17. Over all of $\mathbb{R}^m$ its minimum is $\mu_1(Ł) = 0$, attained at $\mathbf{1}$, so the constraint $\mathbf{x} \perp \mathbf{1}$ exactly removes the direction that achieves 0. What makes this work is that an orthonormal eigenbasis can be chosen with $\mathbf{1}/\sqrt{m}$ as its first member; after that the computation is the weighted-average argument of Chapter 17, run once.

Proof. Since Ł is symmetric, Corollary 12.5.2 supplies an orthonormal basis $(\mathbf{q}_1, \ldots, \mathbf{q}_m)$ of $\mathbb{R}^m$ with $Ł\mathbf{q}_j = \mu_j(Ł)\mathbf{q}_j$, the eigenvalues in the prescribed increasing order; moreover the basis may be chosen so that $\mathbf{q}_1 = \mathbf{1}/\sqrt{m}$. Indeed, let $E_0 = \mathrm{nul}(Ł)$, of dimension $c \geq 1$ by Theorem 24.11.2, and note $\mathbf{1} \in E_0$ and $\mathbf{1} \neq \mathbf{0}$. Applying Gram–Schmidt (Theorem 11.2.2) to a basis of E_0 beginning with $\mathbf{1}$ gives an orthonormal basis of E_0 whose first member is $\mathbf{1}/\sqrt{m}$; joining it to orthonormal bases of the other eigenspaces gives the required basis, because $\mu_1(Ł) = \cdots = \mu_c(Ł) = 0$ are the eigenvalues attached to E_0.

Now let $\mathbf{x} \neq \mathbf{0}$ with $\mathbf{x}^\top \mathbf{1} = 0$, and write $c_j = \langle \mathbf{x}, \mathbf{q}_j \rangle$, so that $\mathbf{x} = \sum_j c_j \mathbf{q}_j$ by Theorem 11.2.1 (a). The constraint says $c_1 = \langle \mathbf{x}, \mathbf{1} \rangle / \sqrt{m} = 0$. Hence, by Theorem 11.2.1 (b), (c),

$$\frac{\mathbf{x}^\top Ł \mathbf{x}}{\mathbf{x}^\top \mathbf{x}} = \frac{\sum_{j \geq 2} \mu_j(Ł)\, c_j^2}{\sum_{j \geq 2} c_j^2} \;\geq\; \mu_2(Ł),$$

a weighted average of $\mu_2(Ł), \ldots, \mu_m(Ł)$ with nonnegative weights, not all zero, being at least the smallest of them. Taking $\mathbf{x} = \mathbf{q}_2$, which satisfies $\mathbf{q}_2 \perp \mathbf{q}_1 = \mathbf{1}/\sqrt{m}$ and is nonzero, gives the value $\mu_2(Ł)$. So the set has minimum $\mu_2(Ł)$, attained at $\mathbf{q}_2$.

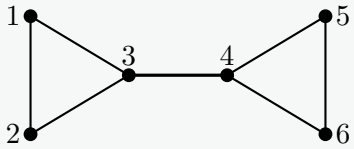
■

Remark. This is the $k = m - 1$ case of the max–min half of Theorem 17.2.3, read in the increasing indexing: there $\lambda_{m-1}(Ł) = \mu_2(Ł)$ is the largest value that any $(m-1)$-dimensional subspace can guarantee for $R_Ł$, and the proposition says that the particular subspace $\mathbf{1}^\perp$ already guarantees it.

A vector achieving the minimum is called a **Fiedler vector**. It must have entries of both signs, since it is orthogonal to $\mathbf{1}$, and Proposition 24.11.1 (b) says it is a labeling that keeps neighbors as close together as it can while being spread out overall. Splitting the vertices by the sign of a Fiedler vector is therefore a reasonable guess at a partition that cuts few edges.

Example. Two triangles and a bridge

Let G have vertices $1, \ldots, 6$ and edges $\{1,2\}$, $\{1,3\}$, $\{2,3\}$, $\{3,4\}$, $\{4,5\}$, $\{4,6\}$, $\{5,6\}$: two triangles joined by a single edge.



Find the Fiedler value and a Fiedler vector, and read off the partition.

Solution. The degrees are $2, 2, 3, 3, 2, 2$, so

$$Ł = \begin{pmatrix} 2 & -1 & -1 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 & 0 \\ -1 & -1 & 3 & -1 & 0 & 0 \\ 0 & 0 & -1 & 3 & -1 & -1 \\ 0 & 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & -1 & 2 \end{pmatrix}.$$

Rather than expand a 6×6 characteristic polynomial, guess the shape of the answer from the symmetry: the graph is carried to itself by the reflection $1 \leftrightarrow 6$, $2 \leftrightarrow 5$, $3 \leftrightarrow 4$, and by the swap $1 \leftrightarrow 2$. Look for an eigenvector reversed by the reflection and fixed by the swap, that is, of the form $\mathbf{x} = (1, 1, b, -b, -1, -1)$. Rows 1 and 3 of $Ł\mathbf{x} = \mu\mathbf{x}$ read

$$1 - b = \mu, \qquad -2 + 4b = \mu b,$$

the other four rows repeating these by symmetry. Substituting $\mu = 1 - b$ into the second gives $-2 + 4b = b - b^2$, that is $b^2 + 3b - 2 = 0$, whose positive root is $b = (\sqrt{17} - 3)/2$. Then

$$\mu = 1 - b = \frac{5 - \sqrt{17}}{2} \approx 0.4384.$$

The vector $\mathbf{x}$ satisfies $\mathbf{x}^\top \mathbf{1} = 2 + b - b - 2 = 0$, so by Proposition 24.11.4 its Rayleigh quotient, which is μ, is at least $\mu_2(Ł)$; and $\mu > 0$, so $\mu \neq \mu_1(Ł) = 0$. To see that μ really is the second smallest, note that $\mathbf{x}$ and $\mathbf{1}$ are eigenvectors for the distinct eigenvalues μ and 0, and check that the remaining four eigenvalues are larger: the vectors $(1, -1, 0, 0, 0, 0)$, $(0, 0, 0, 0, 1, -1)$, and $(1, 1, -2, -2, 1, 1)$ are eigenvectors for the eigenvalue 3, as one verifies row by row, and they are visibly independent, so together with $\mathbf{1}$ and $\mathbf{x}$ they account for five of the six. For the third of them, for instance, row 3 of Ł gives $-1 - 1 - 6 + 2 = -6 = 3 \cdot (-2)$ and row 1 gives $2 - 1 + 2 = 3 = 3 \cdot 1$. The trace of Ł is $2 + 2 + 3 + 3 + 2 + 2 = 14$, so the sixth eigenvalue is $14 - (0 + \mu + 3 + 3 + 3) = 5 - \mu = (5 + \sqrt{17})/2$, also larger than μ. Hence

$$\mu_2(Ł) = \frac{5 - \sqrt{17}}{2}, \qquad \mathbf{x} = \Big(1,\ 1,\ \tfrac{\sqrt{17}-3}{2},\ -\tfrac{\sqrt{17}-3}{2},\ -1,\ -1\Big)$$

is a Fiedler vector, and the signs of its entries split the vertices as $\{1, 2, 3\} \mid \{4, 5, 6\}$: exactly the two triangles, cut along the bridge.

Compare the Fiedler value $0.4384\ldots$ of this graph with the value 6 for K_6. Small means "nearly disconnected", and the bridge is what makes it small. One general fact in that direction is immediate from Chapter 17.

Proposition 24.11.5. Adding an Edge Raises No Eigenvalue

Optional: nothing later depends on this.

Let G be a graph on m vertices with at least one edge, and let G' be G with one extra edge joining $u \neq v$. Then $Ł' = Ł + \mathbf{w}\mathbf{w}^\top$ with $\mathbf{w} = \mathbf{e}_u - \mathbf{e}_v$, and

$$\mu_j(Ł') \;\geq\; \mu_j(Ł) \qquad (j = 1, \dots, m).$$

In particular the Fiedler value does not decrease when an edge is added.

Proof. The extra edge adds 1 to $\deg(u)$ and to $\deg(v)$ and 1 to a_{uv} and a_{vu}, so by Proposition 24.11.1 (a), $Ł' - Ł$ has entries 1 at (u,u) and (v,v), -1 at (u,v) and (v,u), and 0 elsewhere; that matrix is $\mathbf{w}\mathbf{w}^\top$. Now $\mathbf{w}\mathbf{w}^\top$ is symmetric, and $\mathbf{x}^\top \mathbf{w}\mathbf{w}^\top \mathbf{x} = (\mathbf{w}^\top \mathbf{x})^2 \geq 0$ for every $\mathbf{x}$, so $\mathbf{w}\mathbf{w}^\top \succeq \mathbf{O}$ and $Ł' \succeq Ł$ in the Loewner order (Definition 13.5.1). By Corollary 17.2.4, $\lambda_i(Ł') \geq \lambda_i(Ł)$ for every i, and rewriting the decreasing index $i = m + 1 - j$ as the increasing index j gives the claim.

■

24.11.4 Counting spanning trees

Now the second theorem, and the one Chapter 7 promised. Two more words are needed, and both are about the graph alone.

A **cycle** is a set of edges $\{f_1, \dots, f_t\}$ with $t \geq 2$ that can be listed together with distinct vertices $w_0, w_1, \dots, w_{t-1}$ so that, setting $w_t = w_0$, the edge f_i joins w_{i-1} to w_i for each i. Two edges joining the same pair of vertices form a cycle with $t = 2$. A **spanning tree** of G is a set T of edges such that the graph with all m vertices and only the edges of T is connected and contains no cycle. In Exercise (C1: Counting spanning trees) the phrase was "a set of 3 edges that connects all four vertices", and the proposition below reconciles the two descriptions.

The bridge from graphs to determinants is the following lemma. It is where such proofs usually go wrong, so read the third part carefully: it is a statement about $(m-1) \times (m-1)$ submatrices of $\mathbf{N}$ with one row deleted, not about arbitrary square submatrices, and the "exactly when" is an equivalence in both directions.

Lemma 24.11.6. Unimodularity of Incidence Minors

Let $\mathbf{N} \in M_{m \times k}(\mathbb{R})$ be the incidence matrix of an oriented graph G with m vertices and $k \geq 1$ edges. Write $\mathbf{n}_e \in \mathbb{R}^m$ for the column of $\mathbf{N}$ belonging to the edge e.

(a) Every square submatrix of $\mathbf{N}$ has determinant 0, 1, or -1.

(b) For a set S of edges, the list $(\mathbf{n}_e)_{e \in S}$ is linearly independent if and only if S contains no cycle.

(c) Fix a vertex r and put $I = [m] \setminus \{r\}$. For a set S of **exactly** $m - 1$ edges, S is a

spanning tree of G if and only if it contains no cycle, and

$$\det \mathbf{N}_{I,S} = \begin{cases} \pm 1 & \text{if } S \text{ is a spanning tree of } G, \\ 0 & \text{otherwise.} \end{cases}$$

Idea. ① For (a), induct on the size of the square submatrix and look at its columns. A column of $\mathbf{N}$ restricted to a subset of the rows keeps at most the one 1 and the one -1. If some column of the submatrix is zero, the determinant is 0; if some column has a single nonzero entry, expand along it and the size drops; and if **every** column has both, then the rows of the submatrix add up to zero, so they are dependent. ② For (b), a cycle gives an explicit dependence — walk around it and the telescoping sum collapses. Conversely a dependence, restricted to the edges with nonzero coefficient, forces every vertex to meet either none or at least two of them, and a finite graph with that property contains a cycle. ③ Part (c) combines the two, once one knows that deleting the row r costs no rank, which it does not, because the rows of $\mathbf{N}$ already sum to zero.

Proof. **(a)** We induct on t, the common number of rows and columns of the square submatrix $\mathbf{B} = \mathbf{N}_{I',S'}$. For $t = 1$, $\mathbf{B}$ is a single entry of $\mathbf{N}$, which is 0, 1, or -1 by Definition 3.7.1. Let $t \geq 2$ and assume the claim for $t - 1$. Each column of $\mathbf{N}$ has exactly one 1 and one -1, so each column of $\mathbf{B}$ has at most one 1 and at most one -1 and no other nonzero entries. Three cases.

Case 1: some column of $\mathbf{B}$ *is zero.* Then $\det \mathbf{B} = 0$ by Theorem 7.2.1.

Case 2: some column of $\mathbf{B}$ *has exactly one nonzero entry* $\varepsilon = \pm 1$*, in row* p *say.* Expanding along that column (Theorem 7.6.1) gives $\det \mathbf{B} = \pm\varepsilon \det \mathbf{B}'$, where $\mathbf{B}'$ is $\mathbf{B}$ with that row and column deleted. But $\mathbf{B}'$ is again a square submatrix of $\mathbf{N}$, of size $t-1$, so $\det \mathbf{B}' \in \{0, 1, -1\}$ by the induction hypothesis, and hence so is $\det \mathbf{B}$.

Case 3: every column of $\mathbf{B}$ *has both a* 1 *and a* -1. Then every column sums to 0, so the sum of all t rows of $\mathbf{B}$ is the zero row; the rows are linearly dependent, and $\det \mathbf{B} = 0$ by Theorem 7.2.1 and Theorem 7.4.5.

These three cases are exhaustive, which completes the induction.

(b) ($\Leftarrow$) We prove the contrapositive: suppose $\sum_{e \in S} a_e \mathbf{n}_e = \mathbf{0}$ with some $a_e \neq 0$, and let $S' = \{e \in S : a_e \neq 0\}$, a nonempty subset of S. Fix a vertex v. Entry v of the sum is $\sum_{e \in S'} a_e \mathbf{N}_{ve}$, and $\mathbf{N}_{ve} = \pm 1$ exactly for the edges of S' meeting v, and 0 otherwise. If exactly one edge $e_0 \in S'$ met v, that entry would be $\pm a_{e_0} \neq 0$, contradicting the sum being $\mathbf{0}$. So **every vertex meets either no edge of** S' **or at least two**.

Now build a walk. Pick any $f_1 \in S'$, with ends w_0 and w_1. Having reached w_i along f_i, the vertex w_i meets at least two edges of S', so we may choose $f_{i+1} \in S'$ with $f_{i+1} \neq f_i$ meeting w_i, and let w_{i+1} be its other end. Since there are only m vertices, some vertex repeats; let i be the smallest index with $w_i \in \{w_0, \ldots, w_{i-1}\}$, and let $j < i$ be the index with $w_j = w_i$. Then $w_j, w_{j+1}, \ldots, w_{i-1}$ are distinct, by the minimality of i, and $f_{j+1}, \ldots, f_i$ join them in a closed chain. Put $t = i - j$. First, $t \geq 2$: $t = 1$ would mean that f_{j+1} joins w_j to itself, and an edge joins two **distinct** vertices. Second, the t edges are distinct: for $t = 2$ that is the construction, $f_{j+2} \neq f_{j+1}$; and for $t \geq 3$ the pairs $\{w_{a-1}, w_a\}$ for $j + 1 \leq a \leq i$ are pairwise different pairs of the distinct vertices $w_j, \ldots, w_{i-1}$, so the edges joining them differ.

So $\{f_{j+1}, \dots, f_i\}$ is a cycle, and it lies in $S' \subseteq S$.

($\Rightarrow$) Again by contraposition. Let $\{f_1, \dots, f_t\} \subseteq S$ be a cycle with vertices $w_0, \dots, w_{t-1}$ and $w_t = w_0$ as in the definition. For each i set $\varepsilon_i = 1$ if f_i is oriented from w_{i-1} to w_i, and $\varepsilon_i = -1$ otherwise. By Definition 3.7.1 the column of an edge oriented from a to b is $\mathbf{e}_a - \mathbf{e}_b$, so in either case $\varepsilon_i \mathbf{n}_{f_i} = \mathbf{e}_{w_{i-1}} - \mathbf{e}_{w_i}$. Summing telescopes:

$$\sum_{i=1}^{t} \varepsilon_i \mathbf{n}_{f_i} = \mathbf{e}_{w_0} - \mathbf{e}_{w_t} = \mathbf{0},$$

a dependence with all coefficients ± 1, so the list $(\mathbf{n}_e)_{e \in S}$ is dependent.

(c) Write $\mathbf{N}_S = \mathbf{N}_{[m],S} \in M_{m \times (m-1)}(\mathbb{R})$ for the full columns and $\mathbf{N}_{I,S} \in M_{m-1}(\mathbb{R})$ for the same columns with row r deleted. Every column of $\mathbf{N}$ sums to 0, so row r of $\mathbf{N}_S$ is minus the sum of the other rows; deleting it therefore does not change the row space, and by Theorem 3.5.5

$$\operatorname{rank} \mathbf{N}_{I,S} = \operatorname{rank} \mathbf{N}_S.$$

Hence $\det \mathbf{N}_{I,S} \neq 0$ if and only if $\operatorname{rank} \mathbf{N}_S = m - 1$ (Theorem 3.4.4 and Theorem 7.5.2), that is, if and only if the $m - 1$ columns $(\mathbf{n}_e)_{e \in S}$ are linearly independent, which by (b) holds if and only if S contains no cycle. When it does not vanish, $\det \mathbf{N}_{I,S} = \pm 1$ by (a).

It remains to check the equivalence in the statement: for a set S of $m-1$ edges, "no cycle" and "spanning tree" agree. A spanning tree contains no cycle by definition. Conversely, suppose S has $m - 1$ edges and no cycle; we must show that the graph H with all m vertices and the edges of S is connected. By (b) the columns of $\mathbf{N}_S$ are independent, so $\operatorname{rank} \mathbf{N}_S = m - 1$ and $\dim \operatorname{col}(\mathbf{N}_S) = m - 1$. Each column of $\mathbf{N}_S$ lies in $W = \{\mathbf{z} \in \mathbb{R}^m : z_1 + \cdots + z_m = 0\}$, which has dimension $m - 1$ as computed in the proof of Corollary 3.7.5, so $\operatorname{col}(\mathbf{N}_S) = W$ by Theorem 1.6.8. Suppose H were disconnected, and let C be the vertex set of one of its components, so $\emptyset \neq C \neq [m]$. Every edge of S has both ends in C or both ends outside C, so $\mathbf{1}_C^{\top} \mathbf{n}_e = 0$ for every $e \in S$, and hence $\mathbf{1}_C^{\top} \mathbf{z} = 0$ for every $\mathbf{z} \in \operatorname{col}(\mathbf{N}_S) = W$. But choosing $u \in C$ and $v \notin C$ gives $\mathbf{z} = \mathbf{e}_u - \mathbf{e}_v \in W$ with $\mathbf{1}_C^{\top} \mathbf{z} = 1 \neq 0$, a contradiction. So H is connected, and S is a spanning tree.

■

Remark. Part (a) says that $\mathbf{N}$ is **totally unimodular**, the standard name for a matrix every square submatrix of which has determinant 0, 1, or -1. The orientation is essential: for the triangle with edges $\{1,2\}, \{2,3\}, \{3,1\}$, the unoriented matrix of the warning above is $\left(\begin{smallmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{smallmatrix}\right)$, whose determinant is 2. Case 3 of the proof is exactly where the -1 is spent.

One consequence of the last paragraph of the proof is worth recording separately, since it is what makes the sum in the theorem below a sum over spanning trees rather than over sets of every size.

Proposition 24.11.7. A Spanning Tree Has $m - 1$ Edges

Let G be a graph on $m \geq 2$ vertices. Every spanning tree of G has exactly $m - 1$ edges.

Conversely, if a set S of exactly $m-1$ edges is such that the graph on all m vertices with edge set S is connected, then S is a spanning tree.

Proof. Let T be a spanning tree and let H be the graph on all m vertices with edge set T. Then H is connected, so $T \neq \emptyset$ (for $m \geq 2$) and Proposition 3.7.4 applied to H gives $\operatorname{rank} \mathbf{N}_T = m-1$, where $\mathbf{N}_T$ is the incidence matrix of H with some orientation, that is, the columns of $\mathbf{N}$ indexed by T. Also T contains no cycle, so those columns are linearly independent by Lemma 24.11.6 (b), and therefore $|T| = \operatorname{rank} \mathbf{N}_T = m-1$.

Conversely, let S have exactly $m-1$ edges and let H be the graph on all m vertices with edge set S, assumed connected. By Proposition 3.7.4 applied to H, $\operatorname{rank} \mathbf{N}_S = m-1$, where $\mathbf{N}_S$ is the incidence matrix of H, that is, the columns of $\mathbf{N}$ indexed by S; it has $m-1$ columns, so Theorem 3.5.7 gives $\operatorname{nul}(\mathbf{N}_S) = \{\mathbf{0}\}$; since $\mathbf{N}_S\mathbf{a} = \sum_{e\in S} a_e \mathbf{n}_e$ by Theorem 0.9.3, the columns $(\mathbf{n}_e)_{e\in S}$ are linearly independent. By Lemma 24.11.6 (b), S contains no cycle, and H is connected, so S is a spanning tree. In particular the sets counted here are exactly the "sets of 3 edges that connect all four vertices" of Exercise (C1: Counting spanning trees).

■

Theorem 24.11.8. Matrix-Tree Theorem

Let G be a graph on $m \geq 2$ vertices with $k \geq 1$ edges and Laplacian Ł. Fix a vertex r, and let $Ł_0 = Ł_{I,I} \in M_{m-1}(\mathbb{R})$ be Ł with the row and the column of r deleted, where $I = [m]\setminus\{r\}$. Then

$$\det Ł_0 = \#\{\text{spanning trees of } G\}.$$

In particular the answer does not depend on which vertex r is deleted, and $\det Ł_0 = 0$ exactly when G is disconnected.

Idea. The determinant of $Ł_0$ is the determinant of a Gram matrix, because $Ł_0 = \mathbf{N}_0\mathbf{N}_0^{\top}$ for the reduced incidence matrix $\mathbf{N}_0$. Corollary 7.7.3 writes such a determinant as a sum of squares of $(m-1)\times(m-1)$ minors, one for each set of $m-1$ edges. Lemma 24.11.6 says that each of those squares is 1 for a spanning tree and 0 for anything else. So the sum counts spanning trees, one apiece. Every step has already been proved; the theorem is the three of them placed in a row.

Proof. Fix an orientation of G, with incidence matrix $\mathbf{N}$, and put $\mathbf{N}_0 = \mathbf{N}_{I,[k]} \in M_{(m-1)\times k}(\mathbb{R})$, the matrix $\mathbf{N}$ with the row of r deleted.

Step 1: $Ł_0 = \mathbf{N}_0\mathbf{N}_0^{\top}$. By Theorem 0.9.2, the (u,v)-entry of $\mathbf{N}\mathbf{N}^{\top}$ is row u of $\mathbf{N}$ dotted with row v of $\mathbf{N}$. Deleting the row and the column of r from $\mathbf{N}\mathbf{N}^{\top}$ therefore leaves exactly the dot products of the rows indexed by I, which is $\mathbf{N}_0\mathbf{N}_0^{\top}$.

Step 2: Cauchy–Binet. Applying Corollary 7.7.3 to $\mathbf{N}_0 \in M_{(m-1)\times k}(\mathbb{R})$,

$$\det Ł_0 = \sum_{\substack{S\subseteq[k]\\ |S|=m-1}} \left(\det(\mathbf{N}_0)_{[m-1],S}\right)^2,$$

one term for each set S of exactly $m-1$ edges. The submatrix $(\mathbf{N}_0)_{[m-1],S}$ keeps all rows of $\mathbf{N}_0$ and the columns S, so it is the submatrix of $\mathbf{N}$ on rows I and columns S, namely $\mathbf{N}_{I,S}$.

Step 3: identify the terms. By Lemma 24.11.6 (c), $\det \mathbf{N}_{I,S} = \pm 1$ when S is a spanning tree and 0 otherwise, so each term of the sum is 1 for a spanning tree and 0 otherwise. By Proposition 24.11.7 every spanning tree has exactly $m-1$ edges, so every spanning tree occurs as one of the sets S, and it occurs once. Hence the sum is the number of spanning trees.

The left-hand side of Step 2 was computed from a particular r, and the right-hand side does not mention r, so the value is the same for every choice. Finally, a disconnected graph has no spanning tree, since a spanning tree is connected on all m vertices, so the sum is empty and $\det Ł_0 = 0$. A connected graph has at least one: by Proposition 3.7.4, $\operatorname{rank} \mathbf{N} = m-1$, and deleting the row of r does not change the row space, since every column of $\mathbf{N}$ sums to 0, so $\operatorname{rank} \mathbf{N}_0 = m-1$ by Theorem 3.5.5. As $\mathbf{N}_0$ has $m-1$ rows, Theorem 7.7.1 (a) gives a nonzero $(m-1)\times(m-1)$ minor of $\mathbf{N}_0$, which uses all its rows and some set S of $m-1$ columns; that minor is $\det \mathbf{N}_{I,S} \neq 0$, so S is a spanning tree by Lemma 24.11.6 (c). Hence $\det Ł_0 = 0$ exactly when G is disconnected.

■

Example. The four-vertex graph, counted both ways

For the graph of Example (A Laplacian computed twice), compute $\det Ł_0$ with $r = 4$, and list the spanning trees.

Solution. Deleting the row and column of vertex 4 from the Laplacian computed in Example (A Laplacian computed twice) leaves the 3×3 matrix whose determinant Exercise (C1: Counting spanning trees) (a) expands along its first row: $\det Ł_0 = 8$.

Now count directly. A spanning tree has $4-1=3$ edges by Proposition 24.11.7, and there are $\binom{5}{3} = 10$ sets of three edges. By Lemma 24.11.6 (c) such a set fails to be a spanning tree exactly when it contains a cycle, and the only cycles with three or fewer edges are the triangles $\{e_1, e_2, e_5\}$ on vertices $1,2,3$ and $\{e_3, e_4, e_5\}$ on vertices $1,3,4$, in the edge numbering of Example (A Laplacian computed twice). So $10-2=8$ of the sets are spanning trees, matching $\det Ł_0 = 8$.

As a check on one term of the sum in the proof, the tree $\{e_1, e_2, e_3\}$ gives, on rows $I=\{1,2,3\}$,

$$\mathbf{N}_{I,\{1,2,3\}} = \begin{pmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & -1 & 1 \end{pmatrix},$$

lower triangular with determinant 1 (Theorem 7.4.6). The matching computation for a nontree, whose minor is 0, is Exercise (C1: Counting spanning trees) (d).

Warning. **Delete one row and the matching column, and no more.** The Laplacian itself is always singular by Theorem 24.11.2, so $\det Ł = 0$ says nothing. Deleting **two** vertices does not count anything either: for the four-cycle with edges $\{1,2\}, \{2,3\}, \{3,4\}, \{4,1\}$,

striking out the rows and columns of 3 and 4 leaves $\begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}$, of determinant 3, while the graph has 4 spanning trees. And the proof needs the deleted row and the deleted column to belong to the **same** vertex, since Step 1 produces a Gram matrix $\mathbf{N}_0\mathbf{N}_0^\top$ only then; that is a constraint on this route, and nothing above claims that a mismatched pair of deletions counts nothing. Note also that the theorem counts **sets of edges**, so parallel edges count as different trees: two vertices joined by three parallel edges have $Ł = \begin{pmatrix} 3 & -3 \\ -3 & 3 \end{pmatrix}$, $Ł_0 = (3)$, and three spanning trees, one per edge.

Two theorems, both proved from the single identity $\mathbf{x}^\top Ł\mathbf{x} = \sum_e (x_{u_e} - x_{v_e})^2$ and the single factorization $Ł = \mathbf{N}\mathbf{N}^\top$. The first reads the graph's components off the spectrum of Ł; the second reads the number of its spanning trees off one determinant, replacing a search over all $\binom{k}{m-1}$ edge sets by one $O(m^3)$ elimination (Proposition 24.2.3 (a)). Both are statements about a graph that no purely combinatorial argument delivers as cheaply.

24.11.5 Exercises

24.11.5.1 A. Check your understanding

Exercise. A1

(a) Define the Laplacian of a graph, and say why it does not depend on the orientation used to define it.

(b) Write down the formula for $\mathbf{x}^\top Ł\mathbf{x}$ as a sum over edges, and deduce that $Ł \succeq \mathbf{O}$.

(c) State the matrix-tree theorem, including all its hypotheses.

(d) True or false: a graph on m vertices is connected if and only if $\operatorname{rank} Ł = m - 1$. Justify your answer.

(e) True or false: if Ł has eigenvalue 0 with multiplicity 1, then $\det Ł_0 > 0$. Justify your answer.

Solution.

(a) With $\mathbf{N}$ the incidence matrix of any orientation, $Ł = \mathbf{N}\mathbf{N}^\top$ (Definition 24.11.1). By Proposition 24.11.1 (a), $Ł = \mathbf{D} - \mathbf{A}$, and the degree matrix and the adjacency matrix are determined by the graph alone.

(b) $\mathbf{x}^\top Ł\mathbf{x} = \sum_e (x_{u_e} - x_{v_e})^2$, where u_e, v_e are the ends of e (Proposition 24.11.1 (b)). This is a sum of squares of real numbers, hence ≥ 0, and Ł is symmetric, so $Ł \succeq \mathbf{O}$ by Definition 13.1.1.

(c) For a graph on $m \geq 2$ vertices with at least one edge, and any vertex r, the matrix $Ł_0$ obtained from Ł by deleting the row and the column of r satisfies $\det Ł_0 =$ the number of spanning trees (Theorem 24.11.8). Multiple edges are allowed; loops are not, since edges join distinct vertices.

(d) True. By Theorem 24.11.2, $\dim \operatorname{nul}(Ł) = c$, the number of components, so Theorem 3.5.7 gives $\operatorname{rank} Ł = m - c$; and $c = 1$ is connectedness.

(e) True. Multiplicity 1 means $c = 1$ by Theorem 24.11.2, so the graph is connected and has at least one spanning tree, and $\det Ł_0 \geq 1$ by Theorem 24.11.8.

24.11.5.2 B. Practice

Exercise. B1: The four-cycle

Let C_4 be the graph on vertices $1, 2, 3, 4$ with edges $\{1,2\}, \{2,3\}, \{3,4\}, \{4,1\}$.

(a) Write down Ł and explain why it is a circulant in the sense of Definition 12.9.1.

(b) Find all eigenvalues of Ł, the Fiedler value, and a Fiedler vector.

(c) Compute $\det Ł_0$ and list the spanning trees.

Solution.

(a) Every vertex has degree 2, and the vertices u, v are joined exactly when $v - u \equiv \pm 1 \pmod 4$. So, numbering rows and columns $0, 1, 2, 3$ as in Chapter 12 §09, the (j,k)-entry depends only on $k - j$ modulo 4, which is Definition 12.9.1; the first row is $\mathbf{c} = (2, -1, 0, -1)$ and

$$Ł = \begin{pmatrix} 2 & -1 & 0 & -1 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ -1 & 0 & -1 & 2 \end{pmatrix}.$$

(b) By Theorem 12.9.4 the eigenvalues are $p_{\mathbf{c}}(\omega^j)$ for $\omega = i$, with $p_{\mathbf{c}}(x) = 2 - x - x^3$. These were computed in Exercise (B1: Eigenvalues of small circulants) (c): $p_{\mathbf{c}}(1) = 0$, $p_{\mathbf{c}}(i) = 2$, $p_{\mathbf{c}}(-1) = 4$, $p_{\mathbf{c}}(-i) = 2$. In increasing order the eigenvalues are $0, 2, 2, 4$, so the Fiedler value is $\mu_2(Ł) = 2$. A real eigenvector for 2 is twice the real part of the Fourier vector $\mathbf{f}_1 = \frac{1}{2}(1, i, -1, -i)$, namely $(1, 0, -1, 0)$; it is orthogonal to $\mathbf{1}$, and $Ł(1, 0, -1, 0) = (2, 0, -2, 0)$, confirming the eigenvalue. So $(1, 0, -1, 0)$ is a Fiedler vector, splitting the cycle as $\{1\} \mid \{3\}$ with 2 and 4 undecided.

(c) Deleting the row and column of vertex 4,

$$\det Ł_0 = \det \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix} = 2(4 - 1) + 1(-2 - 0) = 4.$$

Directly: a spanning tree has 3 edges, and the only 3-edge subsets that are not spanning trees would have to contain a cycle; the only cycle is all four edges, so every one of the $\binom{4}{3} = 4$ three-edge subsets is a spanning tree. Both counts give 4.

Exercise. B2: Counting trees in a bowtie

Let G have vertices $1, \dots, 5$ and edges $\{1,2\}, \{1,3\}, \{2,3\}, \{3,4\}, \{3,5\}, \{4,5\}$: two triangles sharing the vertex 3.

(a) Write down Ł and compute $\det Ł_0$ with $r = 1$.

(b) Count the spanning trees directly, and check that the two answers agree.

(c) Verify that $\mathbf{x} = (1, 1, 0, -1, -1)$ is an eigenvector of Ł, and say what its eigenvalue tells you.

Solution.

(a) The degrees are $2, 2, 4, 2, 2$, so

$$Ł = \begin{pmatrix} 2 & -1 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ -1 & -1 & 4 & -1 & -1 \\ 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & -1 & 2 \end{pmatrix}, \qquad Ł_0 = \begin{pmatrix} 2 & -1 & 0 & 0 \\ -1 & 4 & -1 & -1 \\ 0 & -1 & 2 & -1 \\ 0 & -1 & -1 & 2 \end{pmatrix}.$$

Expanding $\det Ł_0$ along the first row: $2 \det \left(\begin{smallmatrix} 4 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{smallmatrix}\right) + 1 \cdot \det \left(\begin{smallmatrix} -1 & -1 & -1 \\ 0 & 2 & -1 \\ 0 & -1 & 2 \end{smallmatrix}\right)$. The first 3×3 determinant is $4(4-1) + 1(-2-1) - 1(1+2) = 12 - 3 - 3 = 6$, and the second is $-1(4-1) = -3$. So $\det Ł_0 = 12 - 3 = 9$.

(b) Removing the vertex 3 separates the graph into the two triangles, and a set of edges is a spanning tree exactly when it meets each triangle in a spanning tree of that triangle: the two triangles share only the vertex 3, so no cycle can use edges from both, and connectedness of the whole is connectedness of each half through 3. Each triangle has 3 spanning trees, namely its 3 pairs of edges. Hence there are $3 \times 3 = 9$ spanning trees, agreeing with (a).

(c) $Ł\mathbf{x}$ has entries $2 - 1 - 0 = 1$, $-1 + 2 - 0 = 1$, $-1 - 1 + 0 + 1 + 1 = 0$, $0 - 2 + 1 = -1$, $0 + 1 - 2 = -1$, so $Ł\mathbf{x} = \mathbf{x}$ and the eigenvalue is 1. Since $\mathbf{x}^\top \mathbf{1} = 0$, Proposition 24.11.4 gives $\mu_2(Ł) \le R_Ł(\mathbf{x}) = 1$. The Fiedler value is therefore at most 1, much smaller than the value 5 for the complete graph K_5, computed after Definition 24.11.2, and the sign pattern of $\mathbf{x}$ proposes the partition $\{1, 2\} \mid \{4, 5\}$ with the shared vertex 3 sitting on the fence: the graph is nearly two pieces.

Exercise. B3: Which minors are ± 1?

Let G be the graph on vertices $1, 2, 3$ with edges $e_1 = \{1, 2\}$, $e_2 = \{2, 3\}$, $e_3 = \{1, 3\}$, $e_4 = \{1, 2\}$ (so e_1 and e_4 are parallel). Orient every edge from the smaller vertex to the larger, let $r = 3$ and $I = \{1, 2\}$.

(a) Write down $\mathbf{N}$ and $\mathbf{N}_0 = \mathbf{N}_{I,[4]}$.

(b) Compute $\det \mathbf{N}_{I,S}$ for all six sets S of two edges, and say in each case whether S is a spanning tree.

(c) Hence find the number of spanning trees, and check it against $\det Ł_0$.

Solution.

(a) Column e has 1 at the tail and -1 at the head:

$$\mathbf{N} = \begin{pmatrix} 1 & 0 & 1 & 1 \\ -1 & 1 & 0 & -1 \\ 0 & -1 & -1 & 0 \end{pmatrix}, \qquad \mathbf{N}_0 = \begin{pmatrix} 1 & 0 & 1 & 1 \\ -1 & 1 & 0 & -1 \end{pmatrix}.$$

(b) Taking the two columns in increasing order each time,

$$\{e_1,e_2\}:\det\begin{pmatrix}1&0\\-1&1\end{pmatrix}=1,\qquad \{e_1,e_3\}:\det\begin{pmatrix}1&1\\-1&0\end{pmatrix}=1,$$
$$\{e_1,e_4\}:\det\begin{pmatrix}1&1\\-1&-1\end{pmatrix}=0,\quad \{e_2,e_3\}:\det\begin{pmatrix}0&1\\1&0\end{pmatrix}=-1,$$
$$\{e_2,e_4\}:\det\begin{pmatrix}0&1\\1&-1\end{pmatrix}=-1,\quad \{e_3,e_4\}:\det\begin{pmatrix}1&1\\0&-1\end{pmatrix}=-1.$$

Five of the six are ±1 and those five sets are spanning trees: each is a pair of edges covering all three vertices with no cycle. The exception is $\{e_1,e_4\}$, a cycle of length 2 formed by the parallel pair, which misses the vertex 3 entirely; its minor is 0, as Lemma 24.11.6 (c) predicts.

(c) So there are 5 spanning trees. The degrees are $\deg(1)=3$, $\deg(2)=3$, $\deg(3)=2$, and $a_{12}=2$, $a_{13}=a_{23}=1$, so

$$Ł=\begin{pmatrix}3&-2&-1\\-2&3&-1\\-1&-1&2\end{pmatrix},\qquad \det Ł_0=\det\begin{pmatrix}3&-2\\-2&3\end{pmatrix}=9-4=5.$$

The two answers agree.

24.11.5.3 C. Going deeper

Exercise. C1: Cayley's formula

Let K_m be the complete graph on $m\geq 2$ vertices, in which every pair of distinct vertices is joined by exactly one edge, and let $\mathbf{J}_t\in M_t(\mathbb{R})$ be the all-ones matrix.

(a) Show that the Laplacian of K_m is $m\mathbf{I}_m-\mathbf{J}_m$, and that $Ł_0=m\mathbf{I}_{m-1}-\mathbf{J}_{m-1}$ for every choice of deleted vertex.

(b) Prove that $\det(m\mathbf{I}_{m-1}-\mathbf{J}_{m-1})=m^{m-2}$.

(c) Deduce **Cayley's formula**: K_m has exactly m^{m-2} spanning trees. Check it for $m=2,3,4$ against a direct count.

Hint for (b): the eigenvalues of $\mathbf{J}_t$ are visible from its rank and its row sums.

Solution.

(a) Every vertex of K_m has degree $m-1$ and $a_{uv}=1$ for $u\neq v$, so by Proposition 24.11.1 (a),

$$Ł=(m-1)\mathbf{I}_m-(\mathbf{J}_m-\mathbf{I}_m)=m\mathbf{I}_m-\mathbf{J}_m.$$

Deleting the row and column of any one vertex from $m\mathbf{I}_m-\mathbf{J}_m$ leaves $m\mathbf{I}_{m-1}-\mathbf{J}_{m-1}$, since the pattern of entries is the same at every vertex.

(b) Put $t=m-1$. The matrix $\mathbf{J}_t$ is symmetric with every row equal to $\mathbf{1}^\top$, so $\operatorname{rank}\mathbf{J}_t=1$ and $\dim\operatorname{nul}(\mathbf{J}_t)=t-1$ by Theorem 3.5.7; hence 0 is an eigenvalue of multiplicity at least $t-1$. Also $\mathbf{J}_t\mathbf{1}=t\mathbf{1}$, so t is an eigenvalue. That accounts for t eigenvalues, so the eigenvalue list of $\mathbf{J}_t$ is $t,0,\dots,0$. Therefore $m\mathbf{I}_t-\mathbf{J}_t$ has eigenvalues $m-t=1$ once and m with multiplicity $t-1$, and since $\mathbf{J}_t$ is symmetric it is diagonalizable

(Corollary 12.5.2), so

$$\det(m\mathbf{I}_t - \mathbf{J}_t) = 1 \cdot m^{t-1} = m^{m-2}.$$

(c) By Theorem 24.11.8 and parts (a), (b), the number of spanning trees of K_m is $\det Ł_0 = m^{m-2}$. For $m = 2$: $2^0 = 1$, and indeed the single edge is the only spanning tree. For $m = 3$: $3^1 = 3$, and the triangle's spanning trees are its three pairs of edges. For $m = 4$: $4^2 = 16$; directly, K_4 has $\binom{6}{3} = 20$ sets of three edges, of which the 4 triangles are the only ones containing a cycle, leaving $20 - 4 = 16$.

Exercise. C2: Weighted graphs

Give each edge e of a graph G on $m \geq 2$ vertices a weight $w_e > 0$, let $\mathbf{W} = \operatorname{diag}(w_1, \dots, w_k)$, and define the **weighted Laplacian** $Ł^{\mathbf{w}} := \mathbf{N}\mathbf{W}\mathbf{N}^\top$, with $\mathbf{N}$ the incidence matrix of any orientation.

(a) Prove that $\mathbf{x}^\top Ł^{\mathbf{w}}\mathbf{x} = \sum_e w_e(x_{u_e} - x_{v_e})^2$, and deduce that $Ł^{\mathbf{w}} \succeq \mathbf{O}$ with the same kernel as Ł.

(b) Prove the weighted matrix-tree theorem: for any vertex r,

$$\det Ł_0^{\mathbf{w}} = \sum_T \prod_{e \in T} w_e,$$

the sum running over all spanning trees T of G.

(c) Check (b) on the graph of Example (A Laplacian computed twice) by comparing the coefficient of $w_1 w_2 w_3$ on both sides, and confirm that setting every $w_e = 1$ recovers Theorem 24.11.8.

Hint for (b): $\mathbf{N}\mathbf{W}\mathbf{N}^\top = (\mathbf{N}\mathbf{W}^{1/2})(\mathbf{N}\mathbf{W}^{1/2})^\top$.

Solution.

(a) Let $\mathbf{W}^{1/2} = \operatorname{diag}(\sqrt{w_1}, \dots, \sqrt{w_k})$, so $\mathbf{W} = \mathbf{W}^{1/2}\mathbf{W}^{1/2}$ and $\mathbf{W}^{1/2}$ is symmetric. Put $\mathbf{M} = \mathbf{N}\mathbf{W}^{1/2}$. Then $Ł^{\mathbf{w}} = \mathbf{M}\mathbf{M}^\top$ and, as in Proposition 24.11.1 (b),

$$\mathbf{x}^\top Ł^{\mathbf{w}}\mathbf{x} = \|\mathbf{M}^\top\mathbf{x}\|^2 = \sum_e w_e\big(x_{u_e} - x_{v_e}\big)^2,$$

since entry e of $\mathbf{M}^\top\mathbf{x} = \mathbf{W}^{1/2}\mathbf{N}^\top\mathbf{x}$ is $\sqrt{w_e}$ times entry e of $\mathbf{N}^\top\mathbf{x}$. The matrix is symmetric because $(\mathbf{M}\mathbf{M}^\top)^\top = \mathbf{M}\mathbf{M}^\top$, and the sum is ≥ 0, so $Ł^{\mathbf{w}} \succeq \mathbf{O}$. As in Steps 1 and 2 of the proof of Theorem 24.11.2, $\mathbf{x} \in \operatorname{nul}(Ł^{\mathbf{w}})$ if and only if this sum vanishes; every w_e is **strictly** positive, so the sum vanishes exactly when every difference $x_{u_e} - x_{v_e}$ does, which is the same condition as for Ł. Hence the kernels agree.

(b) With $\mathbf{M} = \mathbf{N}\mathbf{W}^{1/2}$ as above and $I = [m] \setminus \{r\}$, the argument of Step 1 of Theorem 24.11.8 gives $Ł_0^{\mathbf{w}} = \mathbf{M}_0\mathbf{M}_0^\top$ with $\mathbf{M}_0 = \mathbf{M}_{I,[k]}$. By Corollary 7.7.3,

$$\det Ł_0^{\mathbf{w}} = \sum_{|S| = m-1} \big(\det \mathbf{M}_{I,S}\big)^2.$$

Multiplying $\mathbf{N}$ on the right by the diagonal matrix $\mathbf{W}^{1/2}$ scales column e by $\sqrt{w_e}$, so

$\mathbf{M}_{I,S} = \mathbf{N}_{I,S}\,\mathbf{W}_{S,S}^{1/2}$ and, by Theorem 7.5.4 and Theorem 7.4.6,

$$(\det \mathbf{M}_{I,S})^2 = (\det \mathbf{N}_{I,S})^2 \prod_{e\in S} w_e.$$

By Lemma 24.11.6 (c) the first factor is 1 when S is a spanning tree and 0 otherwise, and by Proposition 24.11.7 every spanning tree has $m-1$ edges and so occurs among the sets S, once. This gives the formula.

(c) For that graph, with $r = 4$ and the edge numbering $e_1 = \{1,2\}$, $e_2 = \{2,3\}$, $e_3 = \{3,4\}$, $e_4 = \{4,1\}$, $e_5 = \{1,3\}$,

$$Ł_0^{\mathbf{w}} = \begin{pmatrix} w_1+w_4+w_5 & -w_1 & -w_5 \\ -w_1 & w_1+w_2 & -w_2 \\ -w_5 & -w_2 & w_2+w_3+w_5 \end{pmatrix}.$$

The weight w_3 occurs only in the $(3,3)$-entry, so in the Leibniz expansion (Definition 7.4.1) only the two permutations fixing 3 can produce a term containing w_3: the identity, contributing $(w_1+w_4+w_5)(w_1+w_2)(w_2+w_3+w_5)$, and the transposition of 1 and 2, contributing $-(-w_1)(-w_1)(w_2+w_3+w_5)$, every term of which carries a factor w_1^2. So the coefficient of $w_1w_2w_3$ comes from the diagonal product alone, and there it is 1. On the right, a spanning tree T contributes the single monomial $\prod_{e\in T} w_e$, and T is recovered from that monomial, so the coefficient of $w_1w_2w_3$ is 1 if $\{e_1,e_2,e_3\}$ is a spanning tree and 0 otherwise; it is one, as Example (The four-vertex graph, counted both ways) records. Setting every $w_e = 1$ gives $\mathbf{W} = \mathbf{I}$, hence $Ł^{\mathbf{w}} = Ł$, and the right-hand side becomes the number of spanning trees: Theorem 24.11.8.

24.12 The Fast Fourier Transform

Chapter 12 §09 built one unitary matrix, $\mathbf{F}$, that diagonalizes every circulant of its size, and then counted the cost of using it. Forming $\mathbf{Fb}$ as a matrix–vector product takes about n^2 multiplications, so the elegant three-step recipe "transform, divide, transform back" for a circulant system costs $O(n^2)$ — better than elimination, but not dramatically so. The section closed by naming the way out: the powers of ω repeat, splitting one transform of length n into two of length $n/2$ costs almost nothing, and iterating "brings the cost down to about $n \log n$". That algorithm is the **fast Fourier transform**, and Chapter 12 handed it here, "together with the cost accounting that makes 'quickly' precise".

This section pays that debt. There is no new theorem about $\mathbf{F}$ here; there is one matrix identity, proved by computing an entry, and one recurrence, solved exactly.

Throughout, as in Chapter 12 §09, **rows, columns, and vector entries are numbered** $0, 1, \ldots, n-1$, and $\omega_n := e^{2\pi i/n}$. We write $\mathbf{F}_n$ for the Fourier matrix of size n of Theorem 12.9.3, so $(\mathbf{F}_n)_{jk} = \omega_n^{jk}/\sqrt{n}$; Chapter 12 wrote it $\mathbf{F}$, the size being fixed there.

24.12.1 Counting the cost of a transform

Chapter 12 §09 defined the **discrete Fourier transform** of $\mathbf{c} \in \mathbb{C}^n$ to be $\widehat{\mathbf{c}} = \sqrt{n}\,\mathbf{F}_n\mathbf{c}$, the vector of eigenvalues $p_{\mathbf{c}}(\omega_n^j)$ of the circulant with first row $\mathbf{c}$. It is the unnormalized transform that the algorithm computes, so it is convenient to give its matrix a name of its own.

Definition 24.12.1. The transform matrix

For $n \geq 1$ let

$$\mathbf{W}_n \coloneqq \sqrt{n}\,\mathbf{F}_n \in M_n(\mathbb{C}), \qquad (\mathbf{W}_n)_{jk} = \omega_n^{jk} \quad (0 \leq j,k \leq n-1),$$

so that $\widehat{\mathbf{c}} = \mathbf{W}_n\mathbf{c}$ for every $\mathbf{c} \in \mathbb{C}^n$.

A **cost** in this section is a count of arithmetic operations, where one complex addition, one complex subtraction, one complex multiplication, and one complex division each count as **one operation**; a multiplication by the constant 1 is free, because it is not performed. What is being counted is the work of applying $\mathbf{W}_n$ to an arbitrary vector, not the work of writing $\mathbf{W}_n$ down.

Applying $\mathbf{W}_n$ as a plain matrix–vector product costs

$$n^2 \text{ multiplications} + n(n-1) \text{ additions} = 2n^2 - n$$

operations: each of the n entries of $\mathbf{W}_n\mathbf{c}$ is a sum of n products. This figure charges every one of those products, including the $2n-1$ whose factor is an entry $\omega_n^0 = 1$ of row 0 or column 0; it is the plain matrix–vector product as it is ordinarily performed, and the comparisons below use it as it stands. For $n = 1024$ that is 2 096 128 operations. The whole of this section is about replacing that number by about 15 000.

24.12.2 Splitting a transform in two

Where can a saving possibly come from? The entries of $\mathbf{W}_n$ are powers of ω_n, and there are only n distinct ones, so the n^2 products $\omega_n^{jk}c_k$ involve enormous repetition. The single algebraic fact that organizes the repetition is

$$\omega_n^2 = e^{4\pi i/n} = e^{2\pi i/m} = \omega_m, \qquad \text{where } n = 2m.$$

Squaring a primitive n-th root of unity gives a primitive m-th root of unity. So if we separate the even-numbered entries of $\mathbf{c}$ from the odd-numbered ones, the even half is transformed by $\mathbf{W}_m$, not by anything new. Two objects give the bookkeeping a name.

Definition 24.12.2. The even-odd permutation, the twiddle matrix, and the butterfly

Let $n = 2m$ with $m \geq 1$.

(a) The **even-odd permutation matrix** $\mathbf{P}_n \in M_n(\mathbb{C})$ is the permutation matrix determined by

$$(\mathbf{P}_n\mathbf{x})_t = x_{2t}, \qquad (\mathbf{P}_n\mathbf{x})_{m+t} = x_{2t+1} \qquad (0 \leq t \leq m-1);$$

it lists the even-numbered entries of $\mathbf{x}$ first, then the odd-numbered ones, each in their original order.

(b) The **twiddle matrix** is $\mathbf{D}_m \coloneqq \operatorname{diag}(1, \omega_n, \omega_n^2, \dots, \omega_n^{m-1}) \in M_m(\mathbb{C})$. Its entries are powers of ω_n, **not** of ω_m.

(c) The **butterfly matrix** is

$$\mathbf{B}_n \coloneqq \begin{pmatrix} \mathbf{I}_m & \mathbf{D}_m \\ \mathbf{I}_m & -\mathbf{D}_m \end{pmatrix} \in M_n(\mathbb{C}).$$

In words: $\mathbf{P}_n$ is pure data movement and does no arithmetic; $\mathbf{D}_m$ supplies the m twiddles that the odd half needs because it starts one place late; and $\mathbf{B}_n$ forms the m sums and the m differences that recombine the two halves. Applied to a vector $(\mathbf{u}, \mathbf{v})$ split into two halves of length m, the butterfly returns $(\mathbf{u} + \mathbf{D}_m\mathbf{v},\ \mathbf{u} - \mathbf{D}_m\mathbf{v})$.

Theorem 24.12.1. The Radix-Two Factorization

Let $n = 2m$ with $m \geq 1$. Then

$$\mathbf{W}_n = \mathbf{B}_n \begin{pmatrix} \mathbf{W}_m & \mathbf{O} \\ \mathbf{O} & \mathbf{W}_m \end{pmatrix} \mathbf{P}_n.$$

Equivalently, for the unitary Fourier matrices,

$$\mathbf{F}_n = \frac{1}{\sqrt{2}}\, \mathbf{B}_n \begin{pmatrix} \mathbf{F}_m & \mathbf{O} \\ \mathbf{O} & \mathbf{F}_m \end{pmatrix} \mathbf{P}_n.$$

Idea. Apply both sides to an arbitrary $\mathbf{x}$ and compute entry j of the left side. Split the defining sum $\sum_k \omega_n^{jk} x_k$ into its even $k = 2t$ and odd $k = 2t + 1$ parts; the relation $\omega_n^2 = \omega_m$ turns each part into a transform of length m, and the odd part carries one spare factor ω_n^j. That is the whole computation, and it already gives the top half of the answer. For the bottom half, write $j = m + i$; the relation $\omega_m^m = 1$ makes the two transforms of length m repeat what they gave at i, while $\omega_n^m = -1$ flips the sign of the spare factor. Sum and difference: the butterfly.

Proof. Let $\mathbf{x} \in \mathbb{C}^n$ and put

$$\mathbf{u} \coloneqq \mathbf{W}_m(x_0, x_2, \ldots, x_{n-2}), \qquad \mathbf{v} \coloneqq \mathbf{W}_m(x_1, x_3, \ldots, x_{n-1}),$$

so that $u_i = \sum_{t=0}^{m-1} \omega_m^{it} x_{2t}$ and $v_i = \sum_{t=0}^{m-1} \omega_m^{it} x_{2t+1}$ for $0 \leq i \leq m-1$, by Definition 24.12.1. By Definition 24.12.2 (a), $\mathbf{P}_n\mathbf{x}$ is exactly the pair of these two argument vectors stacked, so the right-hand side applied to $\mathbf{x}$ is $\mathbf{B}_n(\mathbf{u}, \mathbf{v}) = (\mathbf{u} + \mathbf{D}_m\mathbf{v},\ \mathbf{u} - \mathbf{D}_m\mathbf{v})$.

Now compute entry j of $\mathbf{W}_n\mathbf{x}$. Splitting the index k into $k = 2t$ and $k = 2t + 1$ with $0 \leq t \leq m - 1$, and using $\omega_n^2 = \omega_m$,

$$\begin{aligned} (\mathbf{W}_n\mathbf{x})_j &= \sum_{k=0}^{n-1} \omega_n^{jk} x_k = \sum_{t=0}^{m-1} \omega_n^{2jt} x_{2t} + \sum_{t=0}^{m-1} \omega_n^{j(2t+1)} x_{2t+1} \\ &= \sum_{t=0}^{m-1} \omega_m^{jt} x_{2t} + \omega_n^j \sum_{t=0}^{m-1} \omega_m^{jt} x_{2t+1}. \end{aligned}$$

Top half. For $0 \leq j \leq m - 1$ the two sums are u_j and v_j, so

$$(\mathbf{W}_n\mathbf{x})_j = u_j + \omega_n^j v_j = (\mathbf{u} + \mathbf{D}_m\mathbf{v})_j,$$

the last equality by Definition 24.12.2 (b), the j-th diagonal entry of $\mathbf{D}_m$ being ω_n^j.

Bottom half. For $j = m + i$ with $0 \le i \le m - 1$, we have $\omega_m^{jt} = \omega_m^{mt}\omega_m^{it} = \omega_m^{it}$, since $\omega_m^m = 1$; so both sums repeat their values at i, namely u_i and v_i. The spare factor is

$$\omega_n^{m+i} = \omega_n^m \, \omega_n^i = e^{\pi i} \, \omega_n^i = -\omega_n^i,$$

because $\omega_n^m = e^{2\pi i m/n} = e^{\pi i} = -1$. Hence

$$(\mathbf{W}_n \mathbf{x})_{m+i} = u_i - \omega_n^i v_i = (\mathbf{u} - \mathbf{D}_m \mathbf{v})_i.$$

The two halves together say that $\mathbf{W}_n \mathbf{x}$ equals the right-hand side applied to $\mathbf{x}$, for every $\mathbf{x} \in \mathbb{C}^n$; so the two matrices are equal. For the second display, substitute $\mathbf{W}_n = \sqrt{n}\,\mathbf{F}_n$ and $\mathbf{W}_m = \sqrt{m}\,\mathbf{F}_m$ and divide by $\sqrt{n}$, using $\sqrt{m}/\sqrt{n} = 1/\sqrt{2}$.

■

The identity is a factorization of one matrix into three, and each of the three is cheap for a different reason: $\mathbf{P}_n$ does no arithmetic, the middle factor is two copies of a **smaller instance of the same problem**, and $\mathbf{B}_n$ costs $O(n)$. That is the entire content of the fast Fourier transform. Everything below is bookkeeping.

Quick check. Write out Theorem 24.12.1 for $n = 2$ and check it directly.

Solution. Here $m = 1$, so $\mathbf{P}_2 = \mathbf{I}_2$ (it lists x_0 then x_1), $\mathbf{W}_1 = (1)$, and $\mathbf{D}_1 = (\omega_2^0) = (1)$. Hence

$$\mathbf{B}_2 = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix},$$

and the right-hand side is $\mathbf{B}_2\mathbf{I}_2\mathbf{I}_2 = \mathbf{B}_2$. And $(\mathbf{W}_2)_{jk} = \omega_2^{jk} = (-1)^{jk}$, which is the same matrix.

Example. The factorization for $n = 4$

Write out Theorem 24.12.1 for $n = 4$, and check it against the matrix $\mathbf{F}$ of Example (The Fourier matrix of size four).

Solution. Here $m = 2$ and $\omega_4 = i$, so $\mathbf{D}_2 = \operatorname{diag}(1, i)$ and

$$\mathbf{B}_4 = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & i \\ 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -i \end{pmatrix}, \quad \mathbf{P}_4 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad \mathbf{W}_2 = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}.$$

Multiplying, $\begin{pmatrix} \mathbf{W}_2 & \mathbf{O} \\ \mathbf{O} & \mathbf{W}_2 \end{pmatrix} \mathbf{P}_4$ has rows $(1,0,1,0)$, $(1,0,-1,0)$, $(0,1,0,1)$, $(0,1,0,-1)$. Ap-

plying $\mathbf{B}_4$ adds row 3 to row 1, adds i times row 4 to row 2, and so on, giving

$$\mathbf{B}_4 \begin{pmatrix} \mathbf{W}_2 & \mathbf{O} \\ \mathbf{O} & \mathbf{W}_2 \end{pmatrix} \mathbf{P}_4 = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & i & -1 & -i \\ 1 & -1 & 1 & -1 \\ 1 & -i & -1 & i \end{pmatrix} = \mathbf{W}_4,$$

whose (j,k)-entry is i^{jk}. Dividing by $\sqrt{4} = 2$ returns the matrix $\mathbf{F} = \frac{1}{2}\mathbf{W}_4$ written out in Example (The Fourier matrix of size four).

Warning. **The twiddles are powers of ω_n, not of ω_m.** If $\mathbf{D}_2$ in Example (The factorization for $n = 4$) is taken to be $\operatorname{diag}(1, \omega_2) = \operatorname{diag}(1, -1)$, the product becomes the **real** matrix with rows $(1,1,1,1)$, $(1,-1,-1,1)$, $(1,-1,1,-1)$, $(1,1,-1,-1)$, whereas the second row of $\mathbf{W}_4$ is $(1, i, -1, -i)$. The single factor ω_n^j in the proof is what remembers that the odd-numbered entries start one step later, and it is the only place where the transform of size n differs from two transforms of size m.

24.12.3 The cost, exactly

Read Theorem 24.12.1 from right to left as a procedure.

Algorithm. **Input.** $\mathbf{x} \in \mathbb{C}^n$, with $n = 2^p$ a power of two.

Output. $\mathbf{W}_n\mathbf{x}$.

1. If $n = 1$, return $\mathbf{x}$.
2. Split $\mathbf{x}$ into its even-numbered and odd-numbered entries (no arithmetic).
3. Compute $\mathbf{u}$ and $\mathbf{v}$, the transforms of the two halves, by applying this procedure to each of them.
4. Multiply: set $v'_0 := v_0$, and $v'_i := \omega_n^i v_i$ for $1 \le i \le m-1$, where $m = n/2$.
5. Return $(\mathbf{u} + \mathbf{v}',\ \mathbf{u} - \mathbf{v}')$.

The procedure returns $\mathbf{W}_n\mathbf{x}$ for every $n = 2^p$, by induction on p: for $p = 0$ step 1 returns $\mathbf{x} = \mathbf{W}_1\mathbf{x}$, and for $n = 2m$ the inductive hypothesis makes the $\mathbf{u}$ and $\mathbf{v}$ of step 3 the two transforms of length m named in the proof of Theorem 24.12.1, so steps 4 and 5 return $(\mathbf{u}+\mathbf{D}_m\mathbf{v},\ \mathbf{u}-\mathbf{D}_m\mathbf{v})$, which that proof identifies with $\mathbf{W}_n\mathbf{x}$. What is left is the cost.

Theorem 24.12.2. The Cost of the Fast Fourier Transform

Let $C(n)$ be the number of operations the procedure above performs on an input of length $n = 2^p$, with $p \ge 0$. Then $C(1) = 0$,

$$C(2m) = 2C(m) + 3m - 1 \qquad (m \ge 1),$$

and the solution of this recurrence is exactly

$$C(2^p) = \tfrac{3}{2}\, p\, 2^p - 2^p + 1, \qquad \text{that is} \qquad C(n) = \tfrac{3}{2}\, n \log_2 n - n + 1.$$

In particular $C(n) = O(n \log n)$.

Idea. Step 3 is where the recursion sits and contributes $2C(m)$; steps 4 and 5 are the only arithmetic outside it, and both are linear in n. Count them once, carefully: step 4 is $m-1$ multiplications, the factor $\omega_n^0 = 1$ at $i = 0$ not being applied at all, and step 5 is m additions and m subtractions. Then solve the recurrence by induction on p, which works because the recursion halves the length exactly.

Proof. For $n = 1$ the procedure returns its input and does nothing, so $C(1) = 0$.

Let $n = 2m$ with $m \geq 1$. Step 2 only moves data. Step 3 calls the procedure twice on length m, costing $2C(m)$. Step 4 copies v_0, which is no arithmetic, and forms $\omega_n^i v_i$ for $i = 1, \dots, m-1$, which is $m-1$ multiplications. Step 5 forms m sums and m differences, which is $2m$ operations. Adding,

$$C(2m) = 2C(m) + (m-1) + 2m = 2C(m) + 3m - 1.$$

Now induct on p. For $p = 0$, the formula gives $\frac{3}{2} \cdot 0 \cdot 1 - 1 + 1 = 0 = C(1)$. Let $p \geq 1$ and suppose $C(2^{p-1}) = \frac{3}{2}(p-1)2^{p-1} - 2^{p-1} + 1$. Taking $m = 2^{p-1}$ in the recurrence,

$$\begin{aligned} C(2^p) &= 2\Big(\tfrac{3}{2}(p-1)2^{p-1} - 2^{p-1} + 1\Big) + 3 \cdot 2^{p-1} - 1 \\ &= \tfrac{3}{2}(p-1)2^p - 2^p + 2 + \tfrac{3}{2} \cdot 2^p - 1 \\ &= \tfrac{3}{2} p\, 2^p - 2^p + 1, \end{aligned}$$

where the last line collects $\frac{3}{2}(p-1)2^p + \frac{3}{2}2^p = \frac{3}{2}p2^p$. This completes the induction. Substituting $n = 2^p$, so $p = \log_2 n$, gives the stated formula, and $\frac{3}{2} n \log_2 n - n + 1 \leq \frac{3}{2} n \log_2 n$ for $n \geq 1$, so $C(n) = O(n \log n)$.

■

The exact formula is worth tabulating against the $2n^2 - n$ of a plain matrix–vector product.

n	$C(n)$	$2n^2 - n$
2	2	6
8	29	120
32	209	2 016
1024	14 337	2 096 128

At $n = 1024$ the ratio is about 146, and it grows without bound, because $(2n^2 - n)/C(n)$ behaves like $\frac{4}{3} n / \log_2 n$. This is the sense in which the transform can be applied "quickly": the exponent drops from 2 to 1, up to a logarithm, and the constant $\frac{3}{2}$ in front is small.

Warning. **The recursion needs n even at every level, so $n = 2^p$ is the clean case.** For n odd and greater than 1 the factorization of Theorem 24.12.1 does not apply at all, and the procedure above cannot start. For $n = 2^a q$ with q odd, it descends a levels and then

must transform vectors of length q some other way; if q is large, nothing has been gained. The remedy is either to split by other factors of n, as in Exercise (C1: Splitting into three), or, for the convolution use below, to pad with zeros up to the next power of two — which at most doubles n and so changes the cost by a bounded factor.

Example. A transform of length eight

Compute $\mathbf{W}_8\mathbf{c}$ for $\mathbf{c} = (1,2,3,4,1,2,3,4)$ by the procedure, showing every level, and check the result against the direct definition at one entry.

Solution. Level 1. The even-numbered entries are $(1,3,1,3)$ and the odd-numbered ones are $(2,4,2,4)$.

Level 2, applied to $(1,3,1,3)$. Its even and odd halves are $(1,1)$ and $(3,3)$, whose transforms are $\mathbf{W}_2(1,1) = (2,0)$ and $\mathbf{W}_2(3,3) = (6,0)$. Here $n = 4$, so $\mathbf{D}_2 = \operatorname{diag}(1,i)$ and $\mathbf{D}_2(6,0) = (6,0)$. The butterfly gives $(2+6,\ 0+0,\ 2-6,\ 0-0) = (8,0,-4,0)$. So $\mathbf{u} = (8,0,-4,0)$.

Level 2, applied to $(2,4,2,4)$. Its even and odd halves are $(2,2)$ and $(4,4)$, whose transforms are $\mathbf{W}_2(2,2) = (4,0)$ and $\mathbf{W}_2(4,4) = (8,0)$. Again $n = 4$ and $\mathbf{D}_2(8,0) = (8,0)$, so the butterfly gives $(4+8,\ 0,\ 4-8,\ 0)$. So $\mathbf{v} = (12,0,-4,0)$.

Level 1, recombination. Now $n = 8$ and $\mathbf{D}_4 = \operatorname{diag}(1,\omega_8,i,\omega_8^3)$ with $\omega_8 = e^{\pi i/4}$, so

$$\mathbf{D}_4\mathbf{v} = (12,\ 0,\ -4i,\ 0),$$

the irrational entries of $\mathbf{D}_4$ never appearing because $v_1 = v_3 = 0$. The butterfly gives

$$\mathbf{W}_8\mathbf{c} = (20,\ 0,\ -4-4i,\ 0,\ -4,\ 0,\ -4+4i,\ 0).$$

Check at $j = 2$. By Definition 24.12.1, $(\mathbf{W}_8\mathbf{c})_2 = \sum_k \omega_8^{2k}c_k = \sum_k i^k c_k$ since $\omega_8^2 = i$. This is

$$1 + 2i - 3 - 4i + 1 + 2i - 3 - 4i = -4 - 4i,$$

matching. The zeros in the odd positions are not an accident. Since $\mathbf{c}$ repeats with period 4, that is $c_{k+4} = c_k$, splitting the defining sum at $k = 4$ gives

$$(\mathbf{W}_8\mathbf{c})_j = \sum_{k=0}^{3} c_k\big(\omega_8^{jk} + \omega_8^{j(k+4)}\big) = \big(1 + (-1)^j\big)\sum_{k=0}^{3}\omega_8^{jk}c_k,$$

using $\omega_8^{4j} = (e^{\pi i})^j = (-1)^j$. The factor $1 + (-1)^j$ is 0 at every odd j.

24.12.4 What the speed buys

Two corollaries, both immediate now, and both paying off Chapter 12 §09.

Chapter 12 §09 proved the Convolution Theorem: $\widehat{\mathbf{c} * \mathbf{d}} = \widehat{\mathbf{c}} \odot \widehat{\mathbf{d}}$ entrywise (Theorem 12.9.6 (b)). Before this section that identity converted an $O(n^2)$ convolution into three $O(n^2)$ transforms, which is no gain at all. With Theorem 24.12.2 it becomes a genuine algorithm.

Corollary 24.12.3. Convolution in $O(n \log n)$

Let $n = 2^p$. The cyclic convolution $\mathbf{c} * \mathbf{d}$ of two vectors in $\mathbb{C}^n$ can be computed in

$$3C(n) + 2n = \tfrac{9}{2}\, n \log_2 n - n + 3$$

operations, where C is as in Theorem 24.12.2. Consequently, the product of two polynomials in $\mathbb{C}[x]$ of degree at most d can be computed in $O(d \log d)$ operations.

Proof. **The inverse transform costs the same.** By Theorem 12.9.3, $\mathbf{F}_n^{-1} = \mathbf{F}_n^*$, and $\mathbf{F}_n$ is symmetric because $(\mathbf{F}_n)_{jk} = \omega_n^{jk}/\sqrt{n}$ is unchanged by swapping j and k; hence $\mathbf{F}_n^* = \overline{\mathbf{F}_n}$ and

$$\mathbf{W}_n^{-1} = \tfrac{1}{\sqrt{n}}\mathbf{F}_n^{-1} = \tfrac{1}{\sqrt{n}}\overline{\mathbf{F}_n} = \tfrac{1}{n}\overline{\mathbf{W}_n}.$$

Conjugating the identity of Theorem 24.12.1 entry by entry, and noting that $\mathbf{P}_n$ is real, gives

$$\overline{\mathbf{W}_n} = \overline{\mathbf{B}_n} \begin{pmatrix} \overline{\mathbf{W}_m} & \mathbf{O} \\ \mathbf{O} & \overline{\mathbf{W}_m} \end{pmatrix} \mathbf{P}_n,$$

which is the same factorization with every ω_n replaced by $\overline{\omega_n} = \omega_n^{-1}$. The procedure therefore runs verbatim with ω_n^{-1} in place of ω_n and computes $\overline{\mathbf{W}_n}\mathbf{y}$ in $C(n)$ operations.

The algorithm. By Theorem 12.9.6 (b) and the definition $\widehat{\mathbf{c}} = \mathbf{W}_n\mathbf{c}$,

$$\mathbf{W}_n(\mathbf{c} * \mathbf{d}) = (\mathbf{W}_n\mathbf{c}) \odot (\mathbf{W}_n\mathbf{d}), \qquad \text{so} \qquad \mathbf{c} * \mathbf{d} = \tfrac{1}{n}\,\overline{\mathbf{W}_n}\big((\mathbf{W}_n\mathbf{c}) \odot (\mathbf{W}_n\mathbf{d})\big).$$

Computing $\mathbf{W}_n\mathbf{c}$ and $\mathbf{W}_n\mathbf{d}$ costs $2C(n)$; the n entrywise products cost n; the inverse transform costs $C(n)$; the n divisions by n cost n. The total is $3C(n) + 2n$, and substituting the formula of Theorem 24.12.2 gives $\frac{9}{2}n \log_2 n - 3n + 3 + 2n$, as stated.

Polynomials. Let $p = \sum_{l=0}^{d} p_l x^l$ and $q = \sum_{l=0}^{d} q_l x^l$, and choose $n = 2^P$ with $n \geq 2d + 1$ and $n < 4d + 2$, which is possible since the powers of two are spaced by a factor of two. Let $\mathbf{c}, \mathbf{d} \in \mathbb{C}^n$ be the coefficient vectors padded with zeros. The coefficient of x^r in pq, for $0 \leq r \leq 2d$, is $\sum_l c_l d_{r-l}$ summed over $0 \leq l \leq r$, while $(\mathbf{c} * \mathbf{d})_r = \sum_{l=0}^{n-1} c_l d_{r-l}$ with the subscript read modulo n. The two agree: a term of the second sum with $l > r$ has d-subscript $r - l + n$, and it can be nonzero only if $l \leq d$ and $r - l + n \leq d$, which forces $n \leq d + l - r \leq 2d$, contradicting $n \geq 2d + 1$. So the coefficients of pq are the first $2d + 1$ entries of $\mathbf{c} * \mathbf{d}$, computable in $3C(n) + 2n$ operations with $n < 4d + 2$, which is $O(d \log d)$. ∎

Warning. **Convolution wraps around; polynomial multiplication does not.** Padding to length $n \geq 2d{+}1$ is not a convenience, it is the hypothesis. Take $p = q = 1{+}x$, so $d = 1$, and convolve the unpadded vectors $\mathbf{c} = \mathbf{d} = (1, 1)$ in $\mathbb{C}^2$: the answer is $(c_0d_0{+}c_1d_1,\ c_0d_1{+}c_1d_0) = (2, 2)$, the coefficient vector of $2 + 2x$, whereas $(1 + x)^2 = 1 + 2x + x^2$. The x^2 term has wrapped around onto the constant term, because in $\mathbb{C}^n$ the exponents are read modulo n and $x^2 \equiv 1$. Padding to $\mathbf{c} = \mathbf{d} = (1, 1, 0, 0)$ in $\mathbb{C}^4$ gives $(1, 2, 1, 0)$, which is right.

Corollary 24.12.4. Solving a Circulant System in $O(n \log n)$

Let $n = 2^p$, let $\mathbf{C} \in M_n(\mathbb{C})$ be an invertible circulant with first row $\mathbf{c}$, and let $\mathbf{b} \in \mathbb{C}^n$. Then the solution of $\mathbf{Cx} = \mathbf{b}$ can be computed in $3C(n) + 2n$ operations.

Proof. By Theorem 12.9.4, $\mathbf{C} = \mathbf{F}_n \mathbf{\Lambda} \mathbf{F}_n^*$ with $\mathbf{\Lambda} = \operatorname{diag}(p_{\mathbf{c}}(1), \dots, p_{\mathbf{c}}(\omega_n^{n-1}))$, and the diagonal of $\mathbf{\Lambda}$ is exactly $\widehat{\mathbf{c}} = \mathbf{W}_n \mathbf{c}$; the letter $\mathbf{D}$ is reserved here for the twiddle matrix of Definition 24.12.2. Since $\mathbf{C}$ is invertible, every $\widehat{c}_j \neq 0$ by Corollary 12.9.5 (d), so $\mathbf{\Lambda}^{-1}$ exists and

$$\mathbf{x} = \mathbf{C}^{-1}\mathbf{b} = \mathbf{F}_n \mathbf{\Lambda}^{-1} \mathbf{F}_n^* \mathbf{b} = \tfrac{1}{n}\, \mathbf{W}_n \Big((\overline{\mathbf{W}_n}\mathbf{b}) \oslash \widehat{\mathbf{c}} \Big),$$

where $\oslash$ is entrywise division and the last equality uses $\mathbf{F}_n = n^{-1/2}\mathbf{W}_n$ and $\mathbf{F}_n^* = n^{-1/2}\overline{\mathbf{W}_n}$, as in the proof of Corollary 24.12.3. The three transforms cost $C(n)$ each, by Theorem 24.12.2 and the first part of that proof; the n divisions and the n final scalings by $1/n$ cost $2n$. Sanity check at $\mathbf{C} = \mathbf{I}_n$: then $\mathbf{c} = \mathbf{e}_0$, $\widehat{\mathbf{c}} = (1, \dots, 1)$, and the formula returns $\frac{1}{n}\mathbf{W}_n \overline{\mathbf{W}_n}\mathbf{b} = \frac{1}{n}(n\mathbf{I}_n)\mathbf{b} = \mathbf{b}$, using $\mathbf{W}_n \overline{\mathbf{W}_n} = n\mathbf{F}_n\mathbf{F}_n^* = n\mathbf{I}_n$. ∎

Remark. None of this is a new theorem about the discrete Fourier transform. The matrix $\mathbf{F}_n$, the factorization $\mathbf{C} = \mathbf{F}_n \mathbf{\Lambda} \mathbf{F}_n^*$ of Theorem 12.9.4, and the Convolution Theorem (Theorem 12.9.6) are exactly as Chapter 12 left them; Theorem 24.12.1 is an identity between matrices all of which were already known, and Theorem 24.12.2 counts operations. The transform is a change of basis, and the fast Fourier transform is a way of applying that change of basis in $O(n \log n)$ operations instead of $O(n^2)$. What changes is not what can be computed but what can be computed at the sizes that occur.

24.12.5 Exercises

24.12.5.1 A. Check your understanding

Exercise. A1

(a) Write down the matrix $\mathbf{W}_n$, and say how it is related to the Fourier matrix $\mathbf{F}_n$ of Chapter 12.

(b) State Theorem 24.12.1, and say what each of the three factors on the right costs to apply.

(c) How many operations does $\mathbf{W}_n\mathbf{x}$ cost as a plain matrix–vector product, and how many by the fast algorithm when $n = 2^p$?

(d) True or false: the fast Fourier transform computes a different transform from the one in Chapter 12, which is why it is faster. Justify your answer.

(e) Why does the algorithm require n to be a power of two, and what is done when it is not?

Solution.

(a) $\mathbf{W}_n = \sqrt{n}\,\mathbf{F}_n$, with $(\mathbf{W}_n)_{jk} = \omega_n^{jk}$ for $0 \le j,k \le n-1$ and $\omega_n = e^{2\pi i/n}$ (Definition 24.12.1). It is the matrix of the discrete Fourier transform $\mathbf{c} \mapsto \widehat{\mathbf{c}}$ of Chapter 12 §09.

(b) $\mathbf{W}_n = \mathbf{B}_n \begin{pmatrix} \mathbf{W}_m & \mathbf{O} \\ \mathbf{O} & \mathbf{W}_m \end{pmatrix} \mathbf{P}_n$ for $n = 2m$. Applying $\mathbf{P}_n$ costs nothing, since it only moves entries; the middle factor costs two transforms of length m; and $\mathbf{B}_n$ costs $m-1$ multiplications and $2m$ additions and subtractions.

(c) Directly: n^2 multiplications and $n(n-1)$ additions, that is $2n^2 - n$. By the algorithm: $C(n) = \frac{3}{2} n \log_2 n - n + 1$ (Theorem 24.12.2).

(d) False. Theorem 24.12.1 is an identity between matrices, so the algorithm returns $\mathbf{W}_n\mathbf{x}$ exactly, in exact arithmetic. It is faster because it applies the same matrix by a cheaper route, not because it computes something else.

(e) The factorization needs $n = 2m$, and the recursion needs that at every level, so it needs $n = 2^p$. For other n one either pads with zeros to the next power of two, which is legitimate for convolution provided the padded length still exceeds the degree of the product (see the warning after Corollary 24.12.3), or splits by a different factor of n, as in Exercise (C1: Splitting into three).

24.12.5.2 B. Practice

Exercise. B1: A transform of length four

Let $\mathbf{c} = (1, 1, -1, -1) \in \mathbb{C}^4$.

(a) Compute $\mathbf{W}_4\mathbf{c}$ by the procedure of this section, showing the two halves and the twiddles.

(b) Check the answer by evaluating the symbol $p_{\mathbf{c}}$ at the fourth roots of unity.

(c) Count the operations your computation in (a) used, and compare with $C(4)$.

Solution.

(a) The even-numbered entries are $(1, -1)$ and the odd-numbered ones are $(1, -1)$. Both transform to $\mathbf{W}_2(1,-1) = (1-1,\ 1+1) = (0, 2)$, so $\mathbf{u} = \mathbf{v} = (0, 2)$. With $n = 4$, $\mathbf{D}_2 = \operatorname{diag}(1, i)$, so $\mathbf{D}_2\mathbf{v} = (0, 2i)$. The butterfly gives

$$\mathbf{W}_4\mathbf{c} = (0+0,\ 2+2i,\ 0-0,\ 2-2i) = (0,\ 2+2i,\ 0,\ 2-2i).$$

(b) The symbol is $p_{\mathbf{c}}(x) = 1 + x - x^2 - x^3$ (Definition 12.9.1), and $\widehat{c}_j = p_{\mathbf{c}}(i^j)$. Then $p_{\mathbf{c}}(1) = 1+1-1-1 = 0$; $p_{\mathbf{c}}(i) = 1+i+1+i = 2+2i$, using $-i^2 = 1$ and $-i^3 = i$; $p_{\mathbf{c}}(-1) = 1-1-1+1 = 0$; and $p_{\mathbf{c}}(-i) = 1-i+1-i = 2-2i$. The four values agree with (a).

(c) Each $\mathbf{W}_2$ costs 2 operations (one sum, one difference), so the two halves cost 4; the twiddles cost 1, since the factor 1 is free and only $i \cdot 2$ is a multiplication; the final butterfly costs 4. The total is 9, and $C(4) = \frac{3}{2}\cdot 4 \cdot 2 - 4 + 1 = 9$. A direct matrix–vector product would cost $2 \cdot 16 - 4 = 28$.

Exercise. B2: A transform of length eight

Let $\mathbf{c} = (2, 1, 0, 3, 2, 1, 0, 3) \in \mathbb{C}^8$. Compute $\mathbf{W}_8\mathbf{c}$ by the procedure, showing both levels, and verify entry $j = 2$ directly from Definition 24.12.1.

Solution. Level 1. Even-numbered entries $(2, 0, 2, 0)$; odd-numbered entries $(1, 3, 1, 3)$.

Level 2 on $(2, 0, 2, 0)$. Halves $(2, 2)$ and $(0, 0)$, with transforms $(4, 0)$ and $(0, 0)$. Here the inner n is 4, so $\mathbf{D}_2 = \mathrm{diag}(1, i)$ and $\mathbf{D}_2(0, 0) = (0, 0)$. The butterfly gives $\mathbf{u} = (4, 0, 4, 0)$.

Level 2 on $(1, 3, 1, 3)$. Halves $(1, 1)$ and $(3, 3)$, with transforms $(2, 0)$ and $(6, 0)$; $\mathbf{D}_2(6, 0) = (6, 0)$; the butterfly gives $\mathbf{v} = (8, 0, -4, 0)$.

Level 1, recombination. Now $n = 8$ and $\mathbf{D}_4 = \mathrm{diag}(1, \omega_8, i, \omega_8^3)$, so $\mathbf{D}_4\mathbf{v} = (8, 0, -4i, 0)$. Hence

$$\begin{aligned}\mathbf{W}_8\mathbf{c} &= (4 + 8,\ 0,\ 4 - 4i,\ 0,\ 4 - 8,\ 0,\ 4 + 4i,\ 0) \\ &= (12, 0, 4 - 4i, 0, -4, 0, 4 + 4i, 0).\end{aligned}$$

Check at $j = 2$. Since $\omega_8^2 = i$, Definition 24.12.1 gives $(\mathbf{W}_8\mathbf{c})_2 = \sum_k i^k c_k = 2 + i + 0 - 3i + 2 + i + 0 - 3i = 4 - 4i$, as computed.

Exercise. B3: A product of polynomials by convolution

Let $p = 1 + 2x + 3x^2$ and $q = 2 + x$ in $\mathbb{C}[x]$.

(a) What is the smallest power of two n for which the recipe in the proof of Corollary 24.12.3 is valid? Write down the padded coefficient vectors $\mathbf{c}$ and $\mathbf{d}$.

(b) Compute $\mathbf{c} * \mathbf{d}$ directly from the definition of cyclic convolution, and read off pq.

(c) Show that using $n = 3$ instead would give the wrong answer, and identify which coefficient wraps onto which.

Solution.

(a) The product pq has degree 3, so 4 coefficients are needed and any $n \geq 4$ works; the smallest power of two is $n = 4$. The padded vectors are $\mathbf{c} = (1, 2, 3, 0)$ and $\mathbf{d} = (2, 1, 0, 0)$.

(b) With subscripts modulo 4 and $d_0 = 2$, $d_1 = 1$, $d_2 = d_3 = 0$, the sum $(\mathbf{c} * \mathbf{d})_r = \sum_l c_l d_{r-l}$ keeps only $l = r$ and $l = r - 1$, so $(\mathbf{c} * \mathbf{d})_r = 2c_r + c_{r-1}$:

$$\mathbf{c} * \mathbf{d} = (2 + 0,\ 4 + 1,\ 6 + 2,\ 0 + 3) = (2, 5, 8, 3).$$

Hence $pq = 2 + 5x + 8x^2 + 3x^3$, which one checks by expanding: $(1 + 2x + 3x^2)(2 + x) = 2 + x + 4x + 2x^2 + 6x^2 + 3x^3$.

(c) Length 3 holds p and q but not the four coefficients of pq, and although 3 is not a power of two it shows the wrap plainly: with $\mathbf{c} = (1, 2, 3)$, $\mathbf{d} = (2, 1, 0)$, and subscripts modulo 3,

$$\mathbf{c} * \mathbf{d} = (2 + 3,\ 4 + 1,\ 6 + 2) = (5, 5, 8),$$

which is pq with the coefficient 3 of x^3 added onto the constant term, since $x^3 \equiv 1$ modulo $x^3 - 1$. The requirement $n \geq 2d + 1$ in Corollary 24.12.3 is exactly what rules this out.

24.12.5.3 C. Going deeper

Exercise. C1: Splitting into three

Let $n = 3m$ with $m \geq 1$, let $\omega_n = e^{2\pi i/n}$, and for $s = 0, 1, 2$ let

$$\mathbf{u}^{(s)} := \mathbf{W}_m\big(x_s,\ x_{3+s},\ \ldots,\ x_{3(m-1)+s}\big) \in \mathbb{C}^m.$$

(a) Prove that for $0 \leq q \leq 2$ and $0 \leq i \leq m-1$,

$$(\mathbf{W}_n\mathbf{x})_{qm+i} = \sum_{s=0}^{2} \omega_3^{qs}\, \omega_n^{is}\, u_i^{(s)},$$

where $\omega_3 = e^{2\pi i/3}$.

(b) Deduce a recursive procedure for $n = 3^p$, and show that its cost C_3 satisfies $C_3(3m) \leq 3C_3(m) + 12m$ and hence $C_3(3^p) \leq 4p\,3^p = 4n\log_3 n$.

Hint for (a): imitate the proof of Theorem 24.12.1, splitting the index k by its remainder modulo 3 and using $\omega_n^3 = \omega_m$ and $\omega_n^m = \omega_3$.

Solution.

(a) Write $k = 3t + s$ with $0 \leq s \leq 2$ and $0 \leq t \leq m-1$, which lists each $k \in \{0, \ldots, n-1\}$ exactly once. Then, using $\omega_n^3 = e^{6\pi i/n} = e^{2\pi i/m} = \omega_m$,

$$(\mathbf{W}_n\mathbf{x})_j = \sum_{s=0}^{2}\sum_{t=0}^{m-1} \omega_n^{j(3t+s)} x_{3t+s} = \sum_{s=0}^{2} \omega_n^{js} \sum_{t=0}^{m-1} \omega_m^{jt} x_{3t+s}.$$

Put $j = qm + i$. Since $\omega_m^m = 1$, we get $\omega_m^{jt} = \omega_m^{it}$, so the inner sum is $u_i^{(s)}$ by Definition 24.12.1. And $\omega_n^m = e^{2\pi i m/n} = e^{2\pi i/3} = \omega_3$, so

$$\omega_n^{js} = \omega_n^{(qm+i)s} = \big(\omega_n^m\big)^{qs}\omega_n^{is} = \omega_3^{qs}\,\omega_n^{is}.$$

Substituting gives the stated formula.

(b) The procedure is: split $\mathbf{x}$ by the remainder of the index modulo 3 (no arithmetic); transform the three parts recursively, getting $\mathbf{u}^{(0)}, \mathbf{u}^{(1)}, \mathbf{u}^{(2)}$; multiply $u_i^{(s)}$ by ω_n^{is} for $s = 1, 2$ and all i; and for each i apply $\mathbf{W}_3$ to the resulting triple, placing the three outputs at positions i, $m+i$, and $2m+i$. Correctness is (a).

Cost. The recursive calls cost $3C_3(m)$. The twiddles are $2m$ multiplications, the factors for $s = 0$ being 1. Each three-point transform $(a, b, c) \mapsto (a+b+c,\ a+\omega_3 b+\omega_3^2 c,\ a+\omega_3^2 b+\omega_3 c)$ costs at most 4 multiplications and 6 additions, that is 10 operations, and there are m of them. Hence $C_3(3m) \leq 3C_3(m) + 2m + 10m = 3C_3(m) + 12m$.

Now put $D_p = C_3(3^p)/3^p$. The inequality with $m = 3^{p-1}$ reads $3^p D_p \leq 3\cdot 3^{p-1} D_{p-1} + 12\cdot 3^{p-1}$, that is $D_p \leq D_{p-1} + 4$. Since $D_0 = C_3(1) = 0$, induction gives $D_p \leq 4p$, so $C_3(3^p) \leq 4p\,3^p$. With $n = 3^p$, that is $4n\log_3 n = O(n\log n)$.

Exercise. C2: The fourth power of the transform

Let $n \geq 1$ and let $\mathbf{R}_n \in M_n(\mathbb{C})$ be the permutation matrix with $(\mathbf{R}_n)_{jl} = 1$ when $j + l \equiv 0 \pmod{n}$ and 0 otherwise, so that $(\mathbf{R}_n \mathbf{x})_j = x_{n-j}$ with the subscript read modulo n.

(a) Prove that $\mathbf{W}_n^2 = n\,\mathbf{R}_n$.

(b) Deduce that $\mathbf{W}_n^4 = n^2 \mathbf{I}_n$ and $\mathbf{F}_n^4 = \mathbf{I}_n$, and that every eigenvalue of $\mathbf{F}_n$ lies in $\{1, i, -1, -i\}$.

(c) Deduce a second proof that $\mathbf{W}_n^{-1} = \frac{1}{n}\overline{\mathbf{W}_n}$.

Solution.

(a) By Definition 24.12.1 and the definition of the matrix product,

$$(\mathbf{W}_n^2)_{jl} = \sum_{k=0}^{n-1} \omega_n^{jk}\, \omega_n^{kl} = \sum_{k=0}^{n-1} \left(\omega_n^{j+l}\right)^k.$$

Put $z = \omega_n^{j+l}$, so $z^n = 1$. If n divides $j + l$, then $z = 1$ and the sum is n. Otherwise $z \neq 1$ by Example (Roots of Unity), and the geometric-series identity $(z-1)(1 + z + \cdots + z^{n-1}) = z^n - 1 = 0$ used in the proof of Theorem 12.9.3 gives $\sum_k z^k = 0$, since $z - 1 \neq 0$. So $(\mathbf{W}_n^2)_{jl} = n$ exactly when $j + l \equiv 0 \pmod{n}$ and 0 otherwise, which is $n\mathbf{R}_n$.

(b) The map $j \mapsto -j$ modulo n is its own inverse, so $\mathbf{R}_n^2 = \mathbf{I}_n$ and $\mathbf{W}_n^4 = (n\mathbf{R}_n)^2 = n^2\mathbf{I}_n$. Since $\mathbf{W}_n = \sqrt{n}\,\mathbf{F}_n$, we get $n^2\mathbf{F}_n^4 = n^2\mathbf{I}_n$, so $\mathbf{F}_n^4 = \mathbf{I}_n$. Hence $x^4 - 1$ annihilates $\mathbf{F}_n$; if $\mathbf{F}_n\mathbf{v} = \lambda\mathbf{v}$ with $\mathbf{v} \neq \mathbf{0}$, then $\mathbf{v} = \mathbf{F}_n^4\mathbf{v} = \lambda^4\mathbf{v}$, so $\lambda^4 = 1$ and $\lambda \in \{1, i, -1, -i\}$ by Example (Roots of Unity).

(c) From (a), $\mathbf{W}_n \cdot \frac{1}{n}\mathbf{W}_n\mathbf{R}_n = \mathbf{R}_n\mathbf{R}_n = \mathbf{I}_n$, so $\mathbf{W}_n^{-1} = \frac{1}{n}\mathbf{W}_n\mathbf{R}_n$. Its (j,l)-entry is $\frac{1}{n}$ times entry $(j, n-l)$ of $\mathbf{W}_n$, namely $\frac{1}{n}\omega_n^{j(n-l)} = \frac{1}{n}\omega_n^{-jl} = \frac{1}{n}\overline{\omega_n^{jl}}$, using $\omega_n^n = 1$ and $|\omega_n| = 1$. That is the (j,l)-entry of $\frac{1}{n}\overline{\mathbf{W}_n}$.

24.13 Fast Matrix Multiplication

Every algorithm in this chapter is built out of matrix products, and every cost we have counted has been counted against them. Section 12 broke one such count: the Fourier matrix is applied in about $n \log n$ operations, although the definition of a matrix–vector product asks for n^2. The definition of a matrix product asks for n^3 multiplications. This section asks whether n^3 is necessary, answers no, and then asks what the answer costs in accuracy.

24.13.1 What a fast algorithm would be

Count the operations that Definition 0.9.9 prescribes. The product $\mathbf{AB}$ of two matrices in $M_n(F)$ has n^2 entries, each of them a sum of n products, so the recipe performs n^3 multiplications and $n^2(n-1)$ additions:

$$n^3 + n^2(n-1) = 2n^3 - n^2$$

arithmetic operations in all. Chapter 3 §06 made the same kind of count for an LU factorization and got $\frac{2}{3}n^3$, so a product costs about three times a factorization. Nothing so far has suggested that either number could be smaller.

Before asking whether it can be, we need to say what "it" is. A claim that some procedure multiplies matrices in fewer than n^3 multiplications is empty until the permitted procedures are

fixed, and the fixing has to be done carefully: a procedure allowed to compare two numbers, or to branch on whether an entry is zero, is a different object from one that only computes.

A program is a list of arithmetic instructions, each using the inputs or earlier results, and its cost is the length of the list.

Definition 24.13.1. Arithmetic cost of a matrix product

Fix a field F and $n \geq 1$, and treat the $2n^2$ entries of $\mathbf{A}, \mathbf{B} \in M_n(F)$ as **indeterminates**. A **straight-line program** is a finite list $g_1, \dots, g_N$ in which each g_t is either

(i) one of the $2n^2$ indeterminates, or a constant of F (an **input step**), or

(ii) $g_r + g_s$, $g_r - g_s$, or $g_r g_s$ for some $r, s < t$ (an **arithmetic step**).

Each g_t is then a polynomial in the indeterminates. The program **computes the product** if for every i, j the polynomial $\sum_k a_{ik} b_{kj}$ occurs among the g_t. Its **cost** is the number of arithmetic steps, and

$$M(n) \coloneqq \min\{\, \text{cost of a program computing the } n \times n \text{ product} \,\}$$

is the **arithmetic cost** of the $n \times n$ product over F.

In words: input steps are free, because they only name the data; the three arithmetic operations each cost 1; and the program must produce all n^2 entries as **polynomial identities**, not merely for the particular numbers at hand. There are no comparisons, no branches, and no loops — a straight-line program does the same thing whatever the data are. Division is excluded, which costs nothing here: none of the algorithms below uses it.

Two examples fix the definition. For $n = 1$ the single instruction $g_3 = g_1 g_2$ computes $a_{11} b_{11}$, so $M(1) = 1$. For $n = 2$ the recipe of Definition 0.9.9 is the program that forms the eight products $a_{ik} b_{kj}$ and then four sums, at a cost of 12; so $M(2) \leq 12$. A degenerate case worth naming: a program that lists only input steps has cost 0 and computes nothing, since no entry of $\mathbf{AB}$ is an indeterminate.

The exponent of matrix multiplication is the smallest growth rate a family of such programs can have.

Definition 24.13.2. The exponent of matrix multiplication

The **exponent of matrix multiplication** over F is

$$\omega \coloneqq \inf\{\, \tau \in \mathbb{R} : M(n) = O(n^{\tau}) \text{ as } n \to \infty \,\}.$$

The infimum is over **all** real τ for which some constant c and some n_0 give $M(n) \leq c n^{\tau}$ for every $n \geq n_0$, so ω is a property of the whole family of problems, not of any one algorithm. The set is nonempty, since $\tau = 3$ belongs to it by the count above. It is bounded below, and the reason is the only thing about ω that this book proves.

Proposition 24.13.1. The exponent is at least two

$M(n) \geq n^2$ for every $n \geq 1$, and therefore $\omega \geq 2$.

Proof. Let $g_1, \dots, g_N$ be a program computing the $n \times n$ product, and fix i, j. The polynomial $\sum_k a_{ik} b_{kj}$ occurs as some g_t. It is not a constant and not an indeterminate, since it is a nonzero sum of degree-two monomials; so g_t is an arithmetic step. Distinct pairs (i, j) give distinct polynomials, because $\sum_k a_{ik} b_{kj}$ determines i and j from the indices appearing in it. Hence the n^2 pairs require n^2 distinct arithmetic steps, and the cost is at least n^2.

Now let τ be admissible, say $M(n) \le cn^\tau$ for all $n \ge n_0$. Then $n^2 \le cn^\tau$, so $n^{2-\tau} \le c$, for all $n \ge n_0$. If $\tau < 2$ the left side is unbounded, which is impossible; so $\tau \ge 2$. Every admissible τ is at least 2, and hence so is their infimum.

■

So $2 \le \omega \le 3$, and the question of the section is where in that interval ω lies.

Warning. **The exponent is an asymptotic statement and says nothing about any particular size.** $M(n) = O(n^\tau)$ hides a constant, and a program family with a small exponent may carry a constant so large that it loses to the $2n^3 - n^2$ recipe at every size anyone computes with. Section 12's $n \log n$ transform is the exception rather than the rule: there the constant is small. Below, every claim about a *size* is made with the exact operation count in hand, never from an exponent.

24.13.2 Seven products instead of eight

The classical program for $n = 2$ forms the eight products $a_{ik} b_{kj}$. Its cost is what makes the general recipe cost n^3, because a 2×2 block form turns one product of size n into eight of size $n/2$ — and $8 = 2^3$ is exactly the recursion that reproduces n^3. Reducing eight to seven would change the exponent.

Theorem 24.13.2. Strassen's identity

Let F be a field, let $m \ge 1$, and let $\mathbf{A}, \mathbf{B} \in M_{2m}(F)$ be partitioned into $m \times m$ blocks,

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix}, \qquad \mathbf{B} = \begin{pmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{pmatrix}.$$

Define seven products

$$\begin{aligned} \mathbf{P}_1 &= (\mathbf{A}_{11} + \mathbf{A}_{22})(\mathbf{B}_{11} + \mathbf{B}_{22}), & \mathbf{P}_2 &= (\mathbf{A}_{21} + \mathbf{A}_{22})\mathbf{B}_{11}, \\ \mathbf{P}_3 &= \mathbf{A}_{11}(\mathbf{B}_{12} - \mathbf{B}_{22}), & \mathbf{P}_4 &= \mathbf{A}_{22}(\mathbf{B}_{21} - \mathbf{B}_{11}), \\ \mathbf{P}_5 &= (\mathbf{A}_{11} + \mathbf{A}_{12})\mathbf{B}_{22}, & \mathbf{P}_6 &= (\mathbf{A}_{21} - \mathbf{A}_{11})(\mathbf{B}_{11} + \mathbf{B}_{12}), \\ \mathbf{P}_7 &= (\mathbf{A}_{12} - \mathbf{A}_{22})(\mathbf{B}_{21} + \mathbf{B}_{22}). \end{aligned}$$

Then, with the same block partition of $\mathbf{AB}$,

$$\mathbf{AB} = \begin{pmatrix} \mathbf{P}_1 + \mathbf{P}_4 - \mathbf{P}_5 + \mathbf{P}_7 & \mathbf{P}_3 + \mathbf{P}_5 \\ \mathbf{P}_2 + \mathbf{P}_4 & \mathbf{P}_1 + \mathbf{P}_3 - \mathbf{P}_2 + \mathbf{P}_6 \end{pmatrix}.$$

Only seven products of $m \times m$ matrices occur on the right, where Theorem 8.1.1 asks for eight.

Idea. There is no derivation to reproduce: the seven products were found by Strassen, and what a proof can do is check them. What deserves watching during the check is *which* rules it uses. Every step expands a product of sums by distributivity and then cancels equal terms. No two blocks are ever interchanged, and no block is ever inverted or divided by — only the ring operations of $M_m(F)$ appear. That is the whole point, because blocks do not commute.

Proof. Expand the seven products by distributivity, keeping every factor in its place:

$$\begin{aligned}
\mathbf{P}_1 &= \mathbf{A}_{11}\mathbf{B}_{11} + \mathbf{A}_{11}\mathbf{B}_{22} + \mathbf{A}_{22}\mathbf{B}_{11} + \mathbf{A}_{22}\mathbf{B}_{22},\\
\mathbf{P}_2 &= \mathbf{A}_{21}\mathbf{B}_{11} + \mathbf{A}_{22}\mathbf{B}_{11},\\
\mathbf{P}_3 &= \mathbf{A}_{11}\mathbf{B}_{12} - \mathbf{A}_{11}\mathbf{B}_{22},\\
\mathbf{P}_4 &= \mathbf{A}_{22}\mathbf{B}_{21} - \mathbf{A}_{22}\mathbf{B}_{11},\\
\mathbf{P}_5 &= \mathbf{A}_{11}\mathbf{B}_{22} + \mathbf{A}_{12}\mathbf{B}_{22},\\
\mathbf{P}_6 &= \mathbf{A}_{21}\mathbf{B}_{11} + \mathbf{A}_{21}\mathbf{B}_{12} - \mathbf{A}_{11}\mathbf{B}_{11} - \mathbf{A}_{11}\mathbf{B}_{12},\\
\mathbf{P}_7 &= \mathbf{A}_{12}\mathbf{B}_{21} + \mathbf{A}_{12}\mathbf{B}_{22} - \mathbf{A}_{22}\mathbf{B}_{21} - \mathbf{A}_{22}\mathbf{B}_{22}.
\end{aligned}$$

Take the four combinations in turn. For the $(1,1)$ block,

$$\begin{aligned}
\mathbf{P}_1 + \mathbf{P}_4 - \mathbf{P}_5 + \mathbf{P}_7 &= \mathbf{A}_{11}\mathbf{B}_{11} + \mathbf{A}_{11}\mathbf{B}_{22} + \mathbf{A}_{22}\mathbf{B}_{11} + \mathbf{A}_{22}\mathbf{B}_{22}\\
&\quad + \mathbf{A}_{22}\mathbf{B}_{21} - \mathbf{A}_{22}\mathbf{B}_{11}\\
&\quad - \mathbf{A}_{11}\mathbf{B}_{22} - \mathbf{A}_{12}\mathbf{B}_{22}\\
&\quad + \mathbf{A}_{12}\mathbf{B}_{21} + \mathbf{A}_{12}\mathbf{B}_{22} - \mathbf{A}_{22}\mathbf{B}_{21} - \mathbf{A}_{22}\mathbf{B}_{22}\\
&= \mathbf{A}_{11}\mathbf{B}_{11} + \mathbf{A}_{12}\mathbf{B}_{21},
\end{aligned}$$

since $\mathbf{A}_{11}\mathbf{B}_{22}$, $\mathbf{A}_{22}\mathbf{B}_{11}$, $\mathbf{A}_{22}\mathbf{B}_{22}$, $\mathbf{A}_{22}\mathbf{B}_{21}$, and $\mathbf{A}_{12}\mathbf{B}_{22}$ each occur once with each sign. For the $(1,2)$ block,

$$\mathbf{P}_3 + \mathbf{P}_5 = \mathbf{A}_{11}\mathbf{B}_{12} - \mathbf{A}_{11}\mathbf{B}_{22} + \mathbf{A}_{11}\mathbf{B}_{22} + \mathbf{A}_{12}\mathbf{B}_{22} = \mathbf{A}_{11}\mathbf{B}_{12} + \mathbf{A}_{12}\mathbf{B}_{22}.$$

For the $(2,1)$ block,

$$\mathbf{P}_2 + \mathbf{P}_4 = \mathbf{A}_{21}\mathbf{B}_{11} + \mathbf{A}_{22}\mathbf{B}_{11} + \mathbf{A}_{22}\mathbf{B}_{21} - \mathbf{A}_{22}\mathbf{B}_{11} = \mathbf{A}_{21}\mathbf{B}_{11} + \mathbf{A}_{22}\mathbf{B}_{21}.$$

For the $(2,2)$ block,

$$\begin{aligned}
\mathbf{P}_1 + \mathbf{P}_3 - \mathbf{P}_2 + \mathbf{P}_6 &= \mathbf{A}_{11}\mathbf{B}_{11} + \mathbf{A}_{11}\mathbf{B}_{22} + \mathbf{A}_{22}\mathbf{B}_{11} + \mathbf{A}_{22}\mathbf{B}_{22}\\
&\quad + \mathbf{A}_{11}\mathbf{B}_{12} - \mathbf{A}_{11}\mathbf{B}_{22}\\
&\quad - \mathbf{A}_{21}\mathbf{B}_{11} - \mathbf{A}_{22}\mathbf{B}_{11}\\
&\quad + \mathbf{A}_{21}\mathbf{B}_{11} + \mathbf{A}_{21}\mathbf{B}_{12} - \mathbf{A}_{11}\mathbf{B}_{11} - \mathbf{A}_{11}\mathbf{B}_{12}\\
&= \mathbf{A}_{21}\mathbf{B}_{12} + \mathbf{A}_{22}\mathbf{B}_{22}.
\end{aligned}$$

The four results are the four blocks of $\mathbf{AB}$ given by Theorem 8.1.1. This proves the identity.

■

Two features of that proof matter more than the result. First, it used only addition, subtraction, multiplication, and their laws in $M_m(F)$ — never the commutativity of a product, never an inverse, never a division. So the identity would hold with the blocks taken in any ring whatever.

Second, the case $m = 1$ is the ordinary 2×2 product with seven scalar multiplications instead of eight, and the case $m > 1$ is the same statement one level up. That self-similarity is what the next theorem exploits.

Example. Seven products on integer data

Multiply $\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$ and $\mathbf{B} = \begin{pmatrix} 5 & 6 \\ 7 & 8 \end{pmatrix}$ by Theorem 24.13.2, with blocks of size 1, so that each $\mathbf{P}_k$ is a number. Check against the classical product.

Solution. The seven products are

$$\begin{aligned} \mathbf{P}_1 &= (1+4)(5+8) = 65, & \mathbf{P}_2 &= (3+4)\cdot 5 = 35, \\ \mathbf{P}_3 &= 1\cdot(6-8) = -2, & \mathbf{P}_4 &= 4\cdot(7-5) = 8, \\ \mathbf{P}_5 &= (1+2)\cdot 8 = 24, & \mathbf{P}_6 &= (3-1)(5+6) = 22, \\ \mathbf{P}_7 &= (2-4)(7+8) = -30. \end{aligned}$$

Then

$$\begin{aligned} \mathbf{P}_1 + \mathbf{P}_4 - \mathbf{P}_5 + \mathbf{P}_7 &= 65 + 8 - 24 - 30 = 19, \\ \mathbf{P}_3 + \mathbf{P}_5 &= -2 + 24 = 22, \\ \mathbf{P}_2 + \mathbf{P}_4 &= 35 + 8 = 43, \\ \mathbf{P}_1 + \mathbf{P}_3 - \mathbf{P}_2 + \mathbf{P}_6 &= 65 - 2 - 35 + 22 = 50. \end{aligned}$$

The classical product is $\begin{pmatrix} 5+14 & 6+16 \\ 15+28 & 18+32 \end{pmatrix} = \begin{pmatrix} 19 & 22 \\ 43 & 50 \end{pmatrix}$, the same.

Warning. **The seven products are not seven of the eight classical ones.** Not one of $\mathbf{P}_1, \dots, \mathbf{P}_7$ equals a product $a_{ik}b_{kj}$; each is a product of two *combinations*. Dropping a classical product and repairing the rest is impossible, and the identity gives no license to do arithmetic in a different order in any other formula. What it gives is one algebraic fact about one bilinear map, and everything below is that fact applied to itself.

Quick check. Does Theorem 24.13.2 still hold over $\mathbb{F}_2$, where $1 + 1 = 0$?

Solution. Yes. The field F was arbitrary, and the proof used only distributivity, associativity, and the cancellation of a term against its negative. What changes over $\mathbb{F}_2$ is only cosmetic: every minus sign may be written as a plus sign, since $-x = x$ there.

24.13.3 The recursion, and its exact cost

Apply Theorem 24.13.2 with $m \times m$ blocks, and each of the seven products is a product of two $m \times m$ matrices, to be computed by the same rule. The recursion bottoms out at 1×1 blocks, where a product is one multiplication.

Theorem 24.13.3. The cost of the recursion

Let $n = 2^p$ with $p \geq 0$. Applying Theorem 24.13.2 recursively, with the 1×1 case done by a single multiplication, multiplies two matrices in $M_n(F)$ with exactly

$$T(n) = 7n^{\log_2 7} - 6n^2$$

arithmetic operations. In particular $T(n) = O(n^{\log_2 7})$ and $\omega \leq \log_2 7 = 2.8073\ldots$.

Idea. Two counts and one recurrence. Count the block additions in Theorem 24.13.2: ten to build the operands of the seven products, eight to assemble the four blocks of the answer, eighteen in all, each of them $(n/2)^2$ scalar additions. That gives $T(n) = 7T(n/2) + \frac{9}{2}n^2$. Then solve the recurrence exactly, by induction on p rather than by a growth estimate, so that the constants are visible.

Proof. **Step 1: the recurrence.** Let $n = 2m$ with $m \geq 1$, and partition $\mathbf{A}$ and $\mathbf{B}$ into four $m \times m$ blocks each. Forming the operands of $\mathbf{P}_1, \ldots, \mathbf{P}_7$ uses the block sums $\mathbf{A}_{11} + \mathbf{A}_{22}$, $\mathbf{B}_{11} + \mathbf{B}_{22}$, $\mathbf{A}_{21} + \mathbf{A}_{22}$, $\mathbf{B}_{12} - \mathbf{B}_{22}$, $\mathbf{B}_{21} - \mathbf{B}_{11}$, $\mathbf{A}_{11} + \mathbf{A}_{12}$, $\mathbf{A}_{21} - \mathbf{A}_{11}$, $\mathbf{B}_{11} + \mathbf{B}_{12}$, $\mathbf{A}_{12} - \mathbf{A}_{22}$, and $\mathbf{B}_{21} + \mathbf{B}_{22}$: ten of them. Assembling the answer uses three block additions for the $(1,1)$ block, one for $(1,2)$, one for $(2,1)$, and three for $(2,2)$: eight more. Each block addition is m^2 scalar additions, and the seven products cost $T(m)$ each. Hence

$$T(2m) = 7\,T(m) + 18m^2, \qquad T(1) = 1.$$

Step 2: the solution. We show $T(2^p) = 7 \cdot 7^p - 6 \cdot 4^p$ by induction on p. For $p = 0$ the right side is $7 - 6 = 1 = T(1)$. Suppose it holds for p. Taking $m = 2^p$, so that $m^2 = 4^p$,

$$\begin{aligned} T(2^{p+1}) &= 7(7 \cdot 7^p - 6 \cdot 4^p) + 18 \cdot 4^p \\ &= 7 \cdot 7^{p+1} - 42 \cdot 4^p + 18 \cdot 4^p \\ &= 7 \cdot 7^{p+1} - 24 \cdot 4^p = 7 \cdot 7^{p+1} - 6 \cdot 4^{p+1}, \end{aligned}$$

which is the statement for $p + 1$.

Step 3: reading it in n. With $n = 2^p$ we have $4^p = n^2$, and $7^p = 2^{p \log_2 7} = n^{\log_2 7}$. So $T(n) = 7n^{\log_2 7} - 6n^2$, which is at most $7n^{\log_2 7}$.

Finally, the recursion is a straight-line program in the sense of Definition 24.13.1: it uses only additions, subtractions, and multiplications, and by Theorem 24.13.2 it produces every entry of $\mathbf{AB}$ as a polynomial identity. Hence $M(n) \leq 7n^{\log_2 7}$ along the powers of two. For general n, pad $\mathbf{A}$ and $\mathbf{B}$ with zero rows and columns to the next power of two, which at most doubles n; the product of the padded matrices has $\mathbf{AB}$ in its leading block, so $M(n) \leq 7(2n)^{\log_2 7} < 49\, n^{\log_2 7}$ for every n. Therefore $\tau = \log_2 7$ is admissible in Definition 24.13.2, and $\omega \leq \log_2 7$.

■

The number $\log_2 7 = 2.8073\ldots$ is below 3, and that is the whole content: an operation that looked like an n^3 operation is not one. The mechanism is worth naming, because Section 12 used it too. There, the Fourier matrix $\mathbf{F}_n$ was *factored* into a permutation, two copies of $\mathbf{F}_{n/2}$,

and a diagonal factor, and the recurrence $C(n) = 2C(n/2) + O(n)$ followed. Here the bilinear map $(\mathbf{A}, \mathbf{B}) \mapsto \mathbf{AB}$ is factored into seven products and eighteen sums, and the recurrence $T(n) = 7T(n/2) + \frac{9}{2}n^2$ follows. In both cases the saving comes from a factorization, and the exponent is decided by the *number of recursive calls* against the *size reduction*: 2 calls at half size against a linear overhead gives $n \log n$; 7 calls at half size against a quadratic overhead gives $n^{\log_2 7}$.

Remark. The recursion is not a way to make $M(2)$ smaller than 12. Its own count at $n = 2$ is $T(2) = 49 - 24 = 25$: seven multiplications but eighteen additions, against eight and four. What it does is trade one multiplication for fourteen additions at *every* level, and the trade pays only because the multiplications are what recurse.

24.13.4 What is known about the exponent, and what is not

Theorem 24.13.3 gives $\omega \leq \log_2 7 < 2.8074$, and Proposition 24.13.1 gives $\omega \geq 2$. Both are proved above. The rest of this subsection is **stated and not proved here**, and nothing in this book depends on any of it.

Strassen's bound was the first below 3, and it has been improved many times since. The decisive technique is the *laser method* of Coppersmith and Winograd, which brought the bound below 2.376; successive refinements of that method by several authors have since pushed the published bound below 2.372. No lower bound better than the $\omega \geq 2$ proved above is known. **Whether $\omega = 2$ is an open problem.**

Two honest caveats belong with those numbers.

- The bounds below $\log_2 7$ are obtained by analyzing families of recursive constructions, not by exhibiting seven-product identities; the algorithms they yield carry constants so large that at every size that arises in practice they are slower than the $2n^3 - n^2$ recipe. This is a known feature of the constructions, and it is recorded here, not proved.
- ω may depend on the field F, and the definition above was made field by field for that reason. The bounds just quoted, and Strassen's, hold over every field.

Remark. The name **bilinear complexity** belongs to a related quantity: the least number of *multiplications* (as opposed to all arithmetic operations) needed to compute a bilinear map, which for the 2×2 product is exactly 7. The lower bound 7 is a theorem of Winograd and is not proved here; Theorem 24.13.2 is the upper half of it. The exponent ω is the same whether it is defined through Definition 24.13.1 or through multiplications alone, which is also not proved here and is not used.

24.13.5 Accuracy: a norm bound, but not an entrywise one

Everything so far was exact algebra. In the floating-point model of Definition 24.1.1 the two algorithms are not interchangeable, and the difference is sharper than "one is less accurate".

Start with the classical recipe, which inherits Section 1's analysis directly. Write $|\mathbf{A}|$ for the

entrywise absolute value ($|a_{ij}|$), as in Chapter 19, and $\widehat{\mathbf{C}}$ for the computed product. Everything below is real, because Definition 24.1.1 is.

Proposition 24.13.4. The classical product has a small entrywise error

Let $\mathbf{A}, \mathbf{B} \in M_n(\mathbb{R})$ have representable entries, let $nu < 1$, and let $\widehat{\mathbf{C}}$ be the product computed by Definition 0.9.9, each entry as one inner product taken in the natural order. Then

$$|\widehat{\mathbf{C}} - \mathbf{AB}| \;\le\; \gamma_n |\mathbf{A}||\mathbf{B}|$$

entrywise.

Proof. Fix i and j, and let $\mathbf{x}$ be row i of $\mathbf{A}$ and $\mathbf{y}$ column j of $\mathbf{B}$, both with representable entries. By Definition 0.9.9, $(\mathbf{AB})_{ij} = \mathbf{x}^{\top}\mathbf{y}$, and $\widehat{c}_{ij}$ is exactly the quantity s_n of Theorem 24.1.2. That theorem, applicable since $nu < 1$, gives a $\Delta\mathbf{x}$ with $\widehat{c}_{ij} = (\mathbf{x} + \Delta\mathbf{x})^{\top}\mathbf{y}$ and $|\Delta x_k| \le \gamma_n |x_k|$ for every k. Hence

$$|\widehat{c}_{ij} - (\mathbf{AB})_{ij}| = \Big|\sum_k \Delta x_k \, y_k\Big| \le \sum_k |\Delta x_k| \, |y_k| \le \gamma_n \sum_k |a_{ik}| \, |b_{kj}|,$$

where the first inequality is the triangle inequality and the second substitutes the bound on $\Delta\mathbf{x}$ together with $|x_k| = |a_{ik}|$. The right side is $\gamma_n(|\mathbf{A}||\mathbf{B}|)_{ij}$. Since i and j were arbitrary, this proves the bound.

■

The bound is **entrywise**, and that is its strength: the error in the (i, j) entry is measured against the size of the data that produced *that* entry. If $(|\mathbf{A}||\mathbf{B}|)_{ij}$ is tiny, the error there is tiny too. In particular, if $\mathbf{A}$ and $\mathbf{B}$ have nonnegative entries, then $(|\mathbf{A}||\mathbf{B}|)_{ij} = (\mathbf{AB})_{ij}$ and every entry is computed with small *relative* error.

Strassen's algorithm satisfies a weaker bound, in which the whole of $\mathbf{A}$ and the whole of $\mathbf{B}$ appear on the right. Here is the base case, proved.

Proposition 24.13.5. Strassen's two-by-two product satisfies a norm bound

Let $\mathbf{A}, \mathbf{B} \in M_2(\mathbb{R})$ have representable entries, put $\alpha = \max_{i,j}|a_{ij}|$ and $\beta = \max_{i,j}|b_{ij}|$, and let $\widehat{\mathbf{C}}$ be the result of evaluating Theorem 24.13.2 with $m = 1$, each $\mathbf{P}_k$ as written and each of the four combinations left to right. If $6u < 1$, then

$$\max_{i,j} |\widehat{c}_{ij} - (\mathbf{AB})_{ij}| \;\le\; 12\,\gamma_6\,\alpha\beta.$$

Idea. Each $\mathbf{P}_k$ is built by at most three operations, so the computed $\widehat{\mathbf{P}}_k$ carries at most three factors $1 + \delta$. Each output block is a sum of at most four of them, formed by at most three further additions, so each summand carries at most three more. Six factors in all, which is γ_6 by Lemma 24.1.1. The only other ingredient is a crude size bound on each $\mathbf{P}_k$.

Proof. **Sizes.** Each factor of $\mathbf{P}_k$ is either a single entry or a sum of two, so $|\mathbf{P}_1|, |\mathbf{P}_6|, |\mathbf{P}_7| \le (2\alpha)(2\beta) = 4\alpha\beta$, while $\mathbf{P}_2, \mathbf{P}_3, \mathbf{P}_4, \mathbf{P}_5$ each have a single entry as one factor and so satisfy $|\mathbf{P}_k| \le 2\alpha\beta$.

Rounding. Fix k. Forming $\mathbf{P}_k$ takes at most one addition on each side and then one multiplication: at most three operations, each contributing a factor $1+\delta$ with $|\delta| \le u$ by Definition 24.1.1. So $\widehat{\mathbf{P}}_k$ is $\mathbf{P}_k$ times a product of at most three such factors.

Assembly. Take $\widehat{c}_{11}$, computed as $((\widehat{\mathbf{P}}_1 + \widehat{\mathbf{P}}_4) - \widehat{\mathbf{P}}_5) + \widehat{\mathbf{P}}_7$. Each of the three additions contributes one factor $1+\delta$ to every summand already accumulated, so $\widehat{\mathbf{P}}_1$ and $\widehat{\mathbf{P}}_4$ pick up three such factors, $\widehat{\mathbf{P}}_5$ two, and $\widehat{\mathbf{P}}_7$ one. Combining with the previous paragraph, each $\mathbf{P}_m$ is multiplied by a product of at most six factors $1+\delta$, which by Lemma 24.1.1 is $1+\psi_m$ with $|\psi_m| \le \gamma_6$, legitimate because $6u < 1$. Hence

$$\widehat{c}_{11} = \sum_{m \in \{1,4,5,7\}} \pm\, \mathbf{P}_m (1+\psi_m),$$

with the signs of Theorem 24.13.2. Subtracting $c_{11} = \mathbf{P}_1 + \mathbf{P}_4 - \mathbf{P}_5 + \mathbf{P}_7$ and using the triangle inequality,

$$|\widehat{c}_{11} - c_{11}| \;\le\; \gamma_6 \sum_{m \in \{1,4,5,7\}} |\mathbf{P}_m| \;\le\; \gamma_6\,(4+2+2+4)\alpha\beta = 12\gamma_6\alpha\beta.$$

The same argument gives $|\widehat{c}_{22} - c_{22}| \le \gamma_6(4+2+2+4)\alpha\beta = 12\gamma_6\alpha\beta$ for the sum $\mathbf{P}_1 + \mathbf{P}_3 - \mathbf{P}_2 + \mathbf{P}_6$. The blocks $\widehat{c}_{12} = \widehat{\mathbf{P}}_3 + \widehat{\mathbf{P}}_5$ and $\widehat{c}_{21} = \widehat{\mathbf{P}}_2 + \widehat{\mathbf{P}}_4$ use one addition, so four factors and $\gamma_4 \le \gamma_6$ suffice, and $\sum_m |\mathbf{P}_m| \le 4\alpha\beta$ there. The largest of the four bounds is $12\gamma_6\alpha\beta$, which proves the proposition. ∎

A bound of this shape survives the recursion, with a constant that grows with n: for $n = 2^p$ there is a constant $c(n)$, polynomial in n, with

$$\max_{i,j} |\widehat{c}_{ij} - (\mathbf{AB})_{ij}| \;\le\; c(n)\, u \max_{i,j}|a_{ij}| \max_{i,j}|b_{ij}| + O(u^2).$$

This is not proved here. A bound of this form was first established by Brent, and the constant has been improved since; the induction is longer than this section, and nothing below uses it. What *is* proved below is that no improvement of it can be entrywise.

Theorem 24.13.6. Strassen's algorithm has no useful entrywise error bound

Work in a model as in Definition 24.1.1, and suppose the set $\mathcal{R}$ of representable numbers contains a number ε with

$$\varepsilon > 0, \qquad 1+\varepsilon \in \mathcal{R}, \qquad \varepsilon^2 \in \mathcal{R}, \qquad \mathrm{fl}\big((1+\varepsilon)^2\big) = 1 + 2\varepsilon.$$

Put $\mathbf{A} = \mathbf{B} = \mathrm{diag}(1, \varepsilon)$. Then the $\widehat{\mathbf{C}}$ of Proposition 24.13.5 has $\widehat{c}_{22} = 0$, while $(\mathbf{AB})_{22} = \varepsilon^2$. Consequently, any constant c for which

$$|\widehat{\mathbf{C}} - \mathbf{AB}| \;\le\; c\,u\,|\mathbf{A}||\mathbf{B}|$$

holds entrywise for this $\mathbf{A}$ and $\mathbf{B}$ must satisfy $c \ge 1/u$.

Idea. Choose the data so that the $(2,2)$ entry of the product is as small as the model can see, while the quantity $\mathbf{P}_1$ from which it is assembled has size 1. The term ε^2 is then the last bit of $(1+\varepsilon)^2$, and the fourth hypothesis says the rounding throws exactly that bit away. Nothing later can put it back, because the remaining steps are exact.

Proof. Write $a_{11} = b_{11} = 1$, $a_{22} = b_{22} = \varepsilon$, and $a_{12} = a_{21} = b_{12} = b_{21} = 0$; all are in $\mathcal{R}$, since $0, 1, \varepsilon \in \mathcal{R}$. Note first that fl fixes every element of $\mathcal{R}$, because a nearest element of $\mathcal{R}$ to a member of $\mathcal{R}$ is that member. By Definition 0.9.9, $(\mathbf{AB})_{22} = a_{21}b_{12} + a_{22}b_{22} = \varepsilon^2$.

Compute the four quantities that Theorem 24.13.2 combines into the $(2,2)$ entry. The operand sums are $\mathrm{fl}(a_{11} + a_{22}) = \mathrm{fl}(1+\varepsilon) = 1+\varepsilon$ and likewise for $\mathbf{B}$, both exact by hypothesis; so

$$\widehat{\mathbf{P}}_1 = \mathrm{fl}\big((1+\varepsilon)(1+\varepsilon)\big) = \mathrm{fl}\big((1+\varepsilon)^2\big) = 1+2\varepsilon,$$

by the fourth hypothesis. Next $\widehat{\mathbf{P}}_2 = \mathrm{fl}(\mathrm{fl}(0+\varepsilon)\cdot 1) = \varepsilon$, and $\widehat{\mathbf{P}}_3 = \mathrm{fl}(1\cdot \mathrm{fl}(0-\varepsilon)) = -\varepsilon$, using $-\mathcal{R} = \mathcal{R}$; and $\widehat{\mathbf{P}}_6 = \mathrm{fl}(\mathrm{fl}(0-1)\cdot \mathrm{fl}(1+0)) = \mathrm{fl}(-1) = -1$. Each of these values lies in $\mathcal{R}$, so each rounding is the identity.

Assemble $\mathbf{P}_1 + \mathbf{P}_3 - \mathbf{P}_2 + \mathbf{P}_6$ left to right. In order,

$$\mathrm{fl}\big((1+2\varepsilon)+(-\varepsilon)\big) = 1+\varepsilon, \qquad \mathrm{fl}\big((1+\varepsilon)-\varepsilon\big) = 1, \qquad \mathrm{fl}\big(1+(-1)\big) = 0,$$

every intermediate value being in $\mathcal{R}$ and so fixed by fl. Hence $\widehat{c}_{22} = 0$.

Finally $|\mathbf{A}||\mathbf{B}| = \mathrm{diag}(1, \varepsilon^2)$, so an entrywise bound at the $(2,2)$ position reads $\varepsilon^2 = |0-\varepsilon^2| \le cu\varepsilon^2$. Dividing by $\varepsilon^2 > 0$ gives $1 \le cu$, that is $c \ge 1/u$. This proves the theorem. ∎

The constant $1/u$ is about 10^{16} in the arithmetic in common use, so the conclusion is that no entrywise bound worth stating exists: Proposition 24.13.4 holds with $c = \gamma_2/u$, which is close to 2, and for Strassen's algorithm the same inequality at $n = 2$ needs a constant larger than the reciprocal of the unit roundoff. The hypotheses of the theorem are not exotic.

They are not vacuous either, and the fourth one fixes the scale of ε. Since $1+\varepsilon$ is representable, Definition 24.1.1 writes $\mathrm{fl}\big((1+\varepsilon)^2\big) = (1+\varepsilon)^2(1+\delta)$ with $|\delta| \le u$, and the fourth hypothesis says the left side is $1+2\varepsilon$, which forces

$$\delta = -\frac{\varepsilon^2}{(1+\varepsilon)^2}, \qquad \text{hence} \qquad \varepsilon^2 \le u\,(1+\varepsilon)^2.$$

So ε is at most about $\sqrt{u}$: the theorem asks for a quantity small enough that its square falls off the end of the significand, and for nothing else. The next example produces one.

Example. A small entry lost entirely

Work in binary arithmetic with 53-bit significands, rounding to nearest, so that $u = 2^{-53}$. Verify the hypotheses of Theorem 24.13.6 for $\varepsilon = 2^{-27}$, and say what each algorithm returns for the $(2,2)$ entry of the product of $\mathbf{A} = \mathbf{B} = \mathrm{diag}(1, 2^{-27})$.

Solution. The hypotheses. A number is representable when its significand fits in 53 bits. Now $\varepsilon = 2^{-27}$ is a power of two, so $\varepsilon \in \mathcal{R}$ and $\varepsilon^2 = 2^{-54} \in \mathcal{R}$. The number $1 + 2^{-27}$ needs bits 0 through 27, that is 28 of them, so it is representable. Finally

$$(1 + 2^{-27})^2 = 1 + 2^{-26} + 2^{-54},$$

which needs 55 bits and is not. In $[1, 2)$ the representable numbers are spaced 2^{-52} apart; our value lies 2^{-54} above $1 + 2^{-26}$ and $3 \cdot 2^{-54}$ below the next one up, so rounding to nearest gives $\mathrm{fl}((1 + 2^{-27})^2) = 1 + 2^{-26} = 1 + 2\varepsilon$. All four hypotheses hold.

The two answers. By Theorem 24.13.6, Strassen's algorithm returns $\widehat{c}_{22} = 0$, while the true value is 2^{-54}: the relative error is 1, not a loss of digits but a loss of all of them. The classical recipe computes $\mathrm{fl}(0 \cdot 0) = 0$ and then $\mathrm{fl}(0 + \mathrm{fl}(2^{-27} \cdot 2^{-27})) = \mathrm{fl}(2^{-54}) = 2^{-54}$, the exact answer.

The other three entries of the Strassen product also come out exactly right here, and the error 2^{-54} is invisible when measured against $\max|a_{ij}| \max|b_{ij}| = 1$, which is what Proposition 24.13.5 measures it against. That is the whole contrast in one example.

Warning. **"No entrywise bound" is not "unstable".** Proposition 24.13.5 is a genuine bound, and for a matrix whose entries are all of comparable size it says as much as Proposition 24.13.4 does. The two part company only when the entries of $\mathbf{AB}$ span many orders of magnitude, and then the classical algorithm delivers something Strassen's does not: small relative error in the small entries. Backward stability in the sense of Definition 24.1.3 is likewise not in question — the failure above is a failure of *entrywise* accuracy, exactly as Chapter 16 §08 warned that conditioning and stability answer different questions.

24.13.6 Whether to use it

The operation count and the error analysis pull in opposite directions, and the honest summary is a list of exact numbers rather than a verdict. Suppose the recursion is stopped at a base size n_0, below which the classical recipe is used, and let $n = n_0 2^p$. The same induction as in Theorem 24.13.3, started from $T(n_0) = 2n_0^3 - n_0^2$, gives

$$T(n_0 2^p) = 7^p(2n_0^3 - n_0^2) + 6n_0^2(7^p - 4^p).$$

Three readings of this formula. With $n_0 = 1$ the recursion runs to scalars, and it is *slower* than the classical recipe until n exceeds about 654; at $n = 512$ it costs 280 902 385 operations against 268 173 312, and at $n = 1024$ it costs 1 971 035 287 against 2 146 435 072. With a classical base case the picture changes completely: a *single* level of the recursion, $p = 1$, costs $14n_0^3 + 11n_0^2$ against the classical $16n_0^3 - 4n_0^2$ at size $2n_0$, and the first wins as soon as $2n_0^3 > 15n_0^2$, that is from $n_0 = 8$, so from $n = 16$. And at $n = 1024$ with $n_0 = 64$ the count is 1 301 696 512, about 0.61 of the classical one.

So on operation count alone the trade is favorable at sizes that are not large. Three things keep the classical algorithm the one in ordinary use at small and moderate sizes, and none of them is visible in an operation count.

- **Additions are not free in the way the count assumes.** The seven products must be stored, and the eighteen block sums read and write whole blocks that the classical recipe never touches. An operation count charges 1 for an addition and 1 for a multiplication

and charges nothing for moving a number; on a real machine the movement is often the expensive part, and the recursion moves much more data than it saves in arithmetic.

- **The error bound is weaker, and it is weaker in a way that matters exactly when it is not expected** — see Example (A small entry lost entirely).
- **The sizes have to cooperate.** The recursion wants even dimensions at every level; an odd size must be padded or handled specially, and the bookkeeping erodes the saving at precisely the moderate sizes where it is thinnest.

The conclusion is not that Theorem 24.13.2 is a curiosity. It is used, at large sizes, with a classical base case and with the accuracy caveat understood. What it is not is a replacement for a definition: **AB** still means what Definition 0.9.9 says it means, and every theorem in this book about products is untouched. A fast algorithm changes the cost of an operation, never the operation.

24.13.7 Summary and transfer

Chapter 21 closed Part V with a complaint and a promise: "every proof here is an existence or a comparison, and not one of them is an algorithm. That is the subject of Chapter 24." Thirteen sections later, the promise is paid, and what "compute" turned out to mean can be said in one sentence: a computed answer is the exact answer to a nearby question, and how near depends on the algorithm, while how much that nearness costs depends on the problem.

The chapter's one new assumption was the floating-point model of Definition 24.1.1, a description of a machine stated as a definition. Everything else was proved from it and from the linear algebra of Chapters 0 to 21.

- **What a computed answer means.** Section 1 stated the model, proved Lemma 24.1.1 and Theorem 24.1.2, and gave the words the rest of the chapter needs: Definition 24.1.3, Definition 24.1.2, and the rule Theorem 24.1.3, that the forward error is at most the condition number times the backward error.
- **Solving a system.** Section 2 read elimination as a sequence of Schur complements (Theorem 24.2.1), counted it (Proposition 24.2.3), and proved backward error analyses in the cases it could reach: Theorem 24.2.5 outright, and Theorem 24.2.6 for the ordering it displays, the same bound for Chapter 3's ordering being stated with credit. With it came the growth factor Definition 24.2.1, the bound Theorem 24.2.8 — attained, by Example (A matrix that really grows) — and the pivot-free safety of Cholesky (Theorem 24.2.9).
- **Least squares.** Section 3 proved that the sensitivity of the least-squares solution involves the residual as well as $\kappa_2(\mathbf{A})$ (Theorem 24.3.1), that the normal equations lose twice the digits (Theorem 24.3.2), and that a reflection does not (Theorem 24.3.3) — that last for one reflection applied to a matrix, the accumulation over the whole factorization being stated with credit.
- **Eigenvalues.** Section 4 built the power method (Theorem 24.4.1), inverse iteration (Theorem 24.4.3), and Rayleigh quotient iteration (Theorem 24.4.4); Section 5 identified the QR algorithm with orthogonal iteration (Theorem 24.5.2), proved its convergence (Theorem 24.5.3), and explained the Hessenberg reduction (Proposition 24.5.4) and the shifts.
- **Iterating in a subspace.** Section 6 built the Krylov space (Definition 24.6.1), the Arnoldi process (Theorem 24.6.2), and its Hermitian three-term form (Theorem 24.6.3), and showed that the Ritz values (Definition 24.6.2) interlace (Theorem 24.6.4) and carry Chapter 20's residual bounds. Section 7 minimized over the same space: Theorem 24.7.4, the Chebyshev bound Lemma 24.7.6, and the square root in Theorem 24.7.7.
- **One piece of pure mathematics.** Section 8 showed that an algorithm and a differential equation can be the same object: the Lax pair Definition 24.8.1, the isospectral theorem

Theorem 24.8.1, and the Toda flow's relation to the QR algorithm (Theorem 24.8.5).

- **Applications.** Section 9 turned a mass-and-spring system into a generalized eigenproblem (Theorem 24.9.2); Section 10 proved that the best-fitting affine subspace is spanned by the leading singular vectors of the centered data matrix (Theorem 24.10.2) and turned "variance explained" into a theorem (Theorem 24.10.4); Section 11 built the graph Laplacian (Definition 24.11.1), identified its kernel (Theorem 24.11.2), introduced the Fiedler value (Definition 24.11.2) with the minimum it solves (Proposition 24.11.4) and the partition the signs of a Fiedler vector give, and proved the matrix-tree theorem (Theorem 24.11.8).
- **Speed.** Section 12 factored the Fourier matrix (Theorem 24.12.1) and solved the resulting recurrence (Theorem 24.12.2); this section factored the two-by-two product (Theorem 24.13.2), solved the resulting recurrence (Theorem 24.13.3), proved the one lower bound on ω this book has (Proposition 24.13.1), and showed what the speed costs (Theorem 24.13.6).

Four moves carried the chapter.

- **Conditioning belongs to the problem, stability to the algorithm, and the two multiply.** Chapter 16 §08 stated the distinction and could not use it; Theorem 24.1.3 makes it a theorem, and every accuracy claim in the chapter is an instance. *Transfer:* when an answer is wrong, ask first whether any method could have done better. If the condition number says no, do not change the algorithm — change the problem.
- **Measure an answer by a change in the data that makes it exact.** A forward error is usually unknowable and a backward error is usually computable, so prove the backward statement and let Theorem 24.1.3 supply the rest (§§01, 02, 03). *Transfer:* when a quantity you want to bound is out of reach, look for a perturbation of the input that the computed output answers exactly; bounding that is often elementary.
- **Every eigenvalue algorithm iterates, because no finite one can exist.** Chapter 11 §08 gave the reason — a finite procedure using arithmetic and radicals would solve the general quintic — and Sections 4 to 7 are what is left once that door is shut: a sequence of subspaces, each cheap to extend, each carrying a certified error bar from Chapter 20. *Transfer:* when exact solution is impossible in principle, the design question becomes what to iterate *in*; the Krylov space is the answer whenever the only affordable operation is one matrix–vector product.
- **A fast algorithm is a factorization.** The Fourier matrix factors into sparse pieces and the two-by-two product factors into seven multiplications; in both cases the cost follows by solving a recurrence exactly, and in both cases the underlying object is unchanged (§§12, 13). *Transfer:* to make an operation fast, do not look for a shortcut in the arithmetic. Look for a factorization of the operation itself, and count the recursive calls against the size reduction.

The chapter paid the following promises, in the words in which they were made.

The promise	Made in	Paid by
"whether a given method keeps that promise is stability, and it is Chapter 24's subject"	Chapter 16 §08	§01

The promise	Made in	Paid by
"Chapter 24 draws the distinction properly, defines backward stability, and proves which of the algorithms in this book have it"	Chapter 16 §08	§01, §02, §03
"An error analysis of this kind, which measures an answer by a change in the data that makes it exact, is a **backward error analysis**. Chapter 24 builds on it."	Chapter 20 §04	§01
"Chapter 24 makes this precise" (the ε-pivot failure)	Chapter 3 §06	§02
"Chapter 24 reads Gaussian elimination itself as a sequence of Schur complements of 1×1 pivots"	Chapter 8 §03	§02
"Cholesky does about half the arithmetic of a general LU factorization ... it needs no pivoting"	Chapter 13 §02	§02
"Three things are *not* proved here, and all three belong to Chapter 24" (the second of them being that nothing in that chapter is about arithmetic)	Chapter 16 §08	§01, §02, §03
"The conditioning of the least-squares problem itself is not $\kappa_2(\mathbf{A})$ ... a term in $\kappa_2(\mathbf{A})^2$ appears even for the QR route"	Chapter 16 §08	§03
"Chapter 24 makes 'sensitivity' precise and proves the comparison"	Chapter 11 §08, quoted again in Chapter 16 §08	§03
"a stable implementation of the QR route loses about $\log_{10} \kappa_2(\mathbf{A})$ decimal digits while forming and solving the normal equations loses about $2 \log_{10} \kappa_2(\mathbf{A})$ — is still Chapter 24's to prove"	Chapter 16 §08	§03
"Chapter 24 quantifies the loss" (the Householder sign choice)	Chapter 11 §08	§03

The promise	Made in	Paid by
“Chapter 24 explains why it is the stable one” (reflections against classical Gram–Schmidt)	Chapter 11 §02	§03
“Chapter 24 builds an iteration on it” (the Rayleigh quotient)	Chapter 17 §01 and Chapter 20 §06	§04
“the iteration $\mathbf{x} \mapsto \mathbf{A}\mathbf{x}$ is the power method, and Chapter 24 studies how fast it converges”	Chapter 19 §04	§04
“Chapter 24 studies it for general matrices, where the rate is governed by the ratio of the two largest eigenvalue moduli”	Chapter 19 §08	§04
“the power method, the algorithm that computes a Perron vector at the size of the web, belongs to Chapter 24”	Chapter 19 §09	§04
“Chapter 24 develops that iteration and explains why every implementation starts with the reduction proved here”	Chapter 11 §08	§05
“algorithms that compute eigenvalues, is the subject of Chapter 24”	Chapter 20 §11	§04, §05, §06
“Those steps need a probabilistic model, a reason to prefer squared error, and an argument that the sample answer says something about the population. Chapter 24 supplies them”	Chapter 13 §11	§10, for the first two; the third is refused, below
“the graph Laplacian, returns in Chapter 24”	Chapter 3 §07	§11
the matrix-tree theorem, “which Chapter 24 completes”	Chapter 7 §07	§11
“Chapter 24 develops it, together with the cost accounting that makes ‘quickly’ precise” (the fast Fourier transform)	Chapter 12 §09	§12

Some things the chapter did not do, and the chapter’s index collects them in one list. Section 2

proved the backward error of a factorization for the ordering it displays and stated with credit that Chapter 3's ordering satisfies the same bound; it also showed that the growth bound for partial pivoting is far too large to prove stability, and said so rather than tidying it. Section 3 proved the backward error of a single reflection and stated with credit both the accumulation over a whole factorization and the rates at which classical and modified Gram–Schmidt lose orthogonality. Section 4 stated, and did not prove, that Rayleigh quotient iteration converges at a quadratic rate, not a cubic one, for a general diagonalizable matrix. Section 5 quoted the theorem of Abel and Ruffini, which is why every eigenvalue method here iterates, and proved the first column of the convergence theorem of the QR algorithm while stating the general form with credit. Section 6 stated the classical convergence estimate for the extreme Ritz values and proved the clean special case alone, and described the loss of orthogonality in Lanczos without analyzing it. Section 8 derived the continuous Toda statement only after granting that $t \mapsto \mathbf{Q}(t)$ is differentiable, which is analysis this book has not set up; its discrete half, Theorem 24.8.5, rests on nothing unproved. Section 10 proved every deterministic statement about principal components and said plainly that the inference from a sample to a larger population is statistics, which this book does not develop; Chapter 13 §11 promised that inference, and promised as well an account of how a singular value decomposition is actually computed for a large matrix, which no section here gives — §10 says which route to avoid and why, and points at the orthogonal reductions of §§03 and 05, and there it stops. Both over-promises are recorded rather than hidden. In this section, the recursive form of Strassen's error bound, Winograd's lower bound of seven multiplications, the equality of the two definitions of ω, everything known about ω below $\log_2 7$, and Laderman's twenty-three-product identity were stated with credit and not proved.

Nothing in this book is proved from any of them. Each is quoted exactly where a reader would otherwise expect a proof, and every theorem in the chapter either proves its own case or carries what it needs as an explicit hypothesis; since all of them enter the book here, no earlier chapter can rest on them either. The one place a quoted result is used at all is Exercise (C1: Would twenty-three products for three-by-three beat Strassen?) (b), where Laderman's identity is handed to the reader as a hypothesis and the deduction from it is the reader's own. Chapter 21's quoted fact **(A7)** was not used anywhere in this chapter, and no analysis beyond Chapter 16's **(A1)**–**(A6)** was imported.

The book's account of computation ends here. Twenty chapters built objects and proved things about them; this one asked what it costs to produce those objects from data of finite accuracy, and found that the answer splits cleanly in two. How much accuracy the data deserve is a question about the problem, answered by a condition number. How much of that the algorithm delivers is a question about the algorithm, answered by a backward error. Everything in this chapter — the pivoting, the reflections, the shifts, the Krylov spaces, the seven products — is an attempt to keep the second small when the first is out of our hands.

24.13.8 Exercises

24.13.8.1 A. Check your understanding

Exercise. A1

(a) State how many arithmetic operations Definition 0.9.9 uses for the product of two $n \times n$ matrices, separating multiplications from additions.

(b) Define the exponent ω, and state what is proved about it in this section.

(c) True or false: Theorem 24.13.2 shows that $M(2) \leq 7$. Justify your answer.

(d) Which properties of $m \times m$ matrices does the proof of Theorem 24.13.2 use? Say what

would go wrong if the identity had needed two blocks to commute.

(e) State the difference between the error bound of Proposition 24.13.4 and that of Proposition 24.13.5, and say for which matrices the difference matters.

Solution.

(a) n^3 multiplications and $n^2(n-1)$ additions, $2n^3 - n^2$ operations in all: each of the n^2 entries is a sum of n products, needing n multiplications and $n-1$ additions.

(b) $\omega = \inf\{\tau : M(n) = O(n^\tau)\}$, with $M(n)$ the arithmetic cost of Definition 24.13.1. Proved here: $\omega \geq 2$ (Proposition 24.13.1) and $\omega \leq \log_2 7$ (Theorem 24.13.3). Everything else quoted about ω is stated with credit and not proved.

(c) False, and for two reasons. The cost of Definition 24.13.1 counts additions as well as multiplications, and Theorem 24.13.2 uses 18 of them, so its cost at $n = 2$ is 25, not 7; indeed $M(2) \leq 12$ by the classical program, which is better. The number 7 counts multiplications only.

(d) Only the associativity and distributivity of matrix addition and multiplication and the existence of additive inverses, together with Theorem 8.1.1, which says that the blocks of $\mathbf{AB}$ are the eight products combined as for a 2×2 matrix. No block is ever inverted and no two are ever interchanged. If some cancellation had required, say, $\mathbf{A}_{11}\mathbf{B}_{22} = \mathbf{B}_{22}\mathbf{A}_{11}$, it would fail for general blocks, which do not commute, and the identity would be true for numbers and false for matrices — so the recursion of Theorem 24.13.3 would be unavailable, which is the only reason the identity is of any interest.

(e) Proposition 24.13.4 bounds the (i,j) error by $\gamma_n(|\mathbf{A}||\mathbf{B}|)_{ij}$, a quantity attached to that entry; Proposition 24.13.5 bounds every error by a multiple of $\max|a_{ij}| \max|b_{ij}|$, a quantity attached to the whole matrices. They agree in strength when all entries of $\mathbf{AB}$ are of comparable size, and differ when some entry of the product is far smaller than others, as in Example (A small entry lost entirely).

24.13.8.2 B. Practice

Exercise. B1: Seven products by hand

Let $\mathbf{A} = \begin{pmatrix} 2 & -1 \\ 0 & 3 \end{pmatrix}$ and $\mathbf{B} = \begin{pmatrix} 1 & 4 \\ -2 & 1 \end{pmatrix}$. Compute $\mathbf{P}_1, \dots, \mathbf{P}_7$ of Theorem 24.13.2 with blocks of size 1, assemble the four entries, and check the answer against Example (A product of square matrices). Hence state how many multiplications and how many additions you performed.

Solution. With $a_{11} = 2$, $a_{12} = -1$, $a_{21} = 0$, $a_{22} = 3$ and $b_{11} = 1$, $b_{12} = 4$, $b_{21} = -2$, $b_{22} = 1$,

$$\begin{aligned} \mathbf{P}_1 &= (2+3)(1+1) = 10, & \mathbf{P}_2 &= (0+3)\cdot 1 = 3, \\ \mathbf{P}_3 &= 2\cdot(4-1) = 6, & \mathbf{P}_4 &= 3\cdot(-2-1) = -9, \\ \mathbf{P}_5 &= (2-1)\cdot 1 = 1, & \mathbf{P}_6 &= (0-2)(1+4) = -10, \\ \mathbf{P}_7 &= (-1-3)(-2+1) = 4. \end{aligned}$$

Then

$$c_{11} = 10 - 9 - 1 + 4 = 4, \qquad c_{12} = 6 + 1 = 7,$$
$$c_{21} = 3 - 9 = -6, \qquad c_{22} = 10 + 6 - 3 - 10 = 3.$$

So $\mathbf{AB} = \begin{pmatrix} 4 & 7 \\ -6 & 3 \end{pmatrix}$, which is the product computed in Example (A product of square matrices).

The count: seven multiplications, and ten additions to form the operands together with eight to assemble the answer, so eighteen additions, twenty-five operations in all. The classical route uses eight multiplications and four additions, twelve operations. At this size the exchange is a loss.

Exercise. B2: One block, one level

Let

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 2 & 1 \\ 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 2 \\ 0 & 2 & 1 & 1 \end{pmatrix}, \qquad \mathbf{B} = \begin{pmatrix} 1 & 2 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 2 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \end{pmatrix},$$

partitioned into 2×2 blocks. Compute $\mathbf{P}_3$ and $\mathbf{P}_5$ of Theorem 24.13.2, hence the $(1,2)$ block of $\mathbf{AB}$, and verify it against the block formula $\mathbf{A}_{11}\mathbf{B}_{12} + \mathbf{A}_{12}\mathbf{B}_{22}$.

Solution. The blocks are

$$\mathbf{A}_{11} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \mathbf{A}_{12} = \begin{pmatrix} 2 & 1 \\ 1 & 0 \end{pmatrix}, \quad \mathbf{B}_{12} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \mathbf{B}_{22} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}.$$

Then $\mathbf{B}_{12} - \mathbf{B}_{22} = \begin{pmatrix} -1 & 0 \\ 1 & -1 \end{pmatrix}$ and $\mathbf{A}_{11} + \mathbf{A}_{12} = \begin{pmatrix} 3 & 1 \\ 1 & 1 \end{pmatrix}$, so

$$\mathbf{P}_3 = \mathbf{A}_{11}(\mathbf{B}_{12} - \mathbf{B}_{22}) = \begin{pmatrix} -1 & 0 \\ 1 & -1 \end{pmatrix}, \qquad \mathbf{P}_5 = \begin{pmatrix} 3 & 1 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 3 & 4 \\ 1 & 2 \end{pmatrix}.$$

Hence

$$\mathbf{P}_3 + \mathbf{P}_5 = \begin{pmatrix} 2 & 4 \\ 2 & 1 \end{pmatrix}.$$

Directly, $\mathbf{A}_{11}\mathbf{B}_{12} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ and $\mathbf{A}_{12}\mathbf{B}_{22} = \begin{pmatrix} 2 & 3 \\ 1 & 1 \end{pmatrix}$, whose sum is $\begin{pmatrix} 2 & 4 \\ 2 & 1 \end{pmatrix}$. The two agree, and by Theorem 8.1.1 this is rows $1, 2$ and columns $3, 4$ of $\mathbf{AB}$.

Exercise. B3: Counting the recursion

Let $n = 2^p$.

(a) How many multiplications does the recursion of Theorem 24.13.3 perform, as a function of p? How many does the classical recipe perform?

(b) Using $T(n) = 7n^{\log_2 7} - 6n^2$, compute $T(n)$ and $2n^3 - n^2$ for $n = 8$ and $n = 64$.

(c) Hence say, for these two sizes, which recipe uses fewer operations.

Solution.

(a) Each level replaces one product of size $2m$ by seven of size m, and a product of size 1 is one multiplication, so the recursion performs $7^p = n^{\log_2 7}$ multiplications. The classical recipe performs $n^3 = 8^p$.

(b) For $n = 8$, that is $p = 3$: $T(8) = 7 \cdot 343 - 6 \cdot 64 = 2401 - 384 = 2017$, while $2 \cdot 512 - 64 = 960$. For $n = 64$, that is $p = 6$: $T(64) = 7 \cdot 117\,649 - 6 \cdot 4096 = 823\,543 - 24\,576 = 798\,967$, while $2 \cdot 262\,144 - 4096 = 520\,192$.

(c) At both sizes the classical recipe wins, by a factor of about 2.1 at $n = 8$ and about 1.5 at $n = 64$. The ratio is shrinking, and it crosses 1 between $n = 512$ and $n = 1024$; recursing all the way to scalars is a losing strategy at any size a person would compute by hand.

24.13.8.3 C. Going deeper

Exercise. C1: Would twenty-three products for three-by-three beat Strassen?

Suppose a ring identity multiplies two 3×3 matrices with $q > 9$ multiplications of combinations of entries, together with some fixed number of additions, and is used recursively on 3×3 block matrices.

(a) Show that the recursion performs q^p multiplications on matrices of size $n = 3^p$, and deduce that it gives $\omega \le \log_3 q$.

(b) You are given that such an identity exists with $q = 23$; this is a result of Laderman, which is not proved here. Determine whether recursing on it beats Theorem 24.13.3. Justify your answer.

(c) Determine the largest q for which such an identity *would* beat Theorem 24.13.3.

Hint: compare logarithms to a common base.

Solution.

(a) One level replaces a product of size $3m$ by q products of size m, so by induction on p the recursion performs q^p multiplications at size $n = 3^p$, the case $p = 0$ being one multiplication. The additions at one level cost Cm^2 for a constant C depending only on the identity, so the operation count satisfies $T(3m) = qT(m) + Cm^2$ with $T(1) = 1$. Unrolling,

$$T(3^p) = q^p + C\sum_{j=0}^{p-1} q^j\, 9^{p-1-j} = q^p + C\,9^{p-1}\,\frac{(q/9)^p - 1}{(q/9) - 1} \;\le\; q^p\Big(1 + \frac{C}{q-9}\Big),$$

where the last step uses $q > 9$ and drops the -1. So $T(n) \le cn^{\log_3 q}$ with $c = 1 + C/(q-9)$, since $q^p = n^{\log_3 q}$. Padding to a power of three as in Theorem 24.13.3 extends this to all n, so $\omega \le \log_3 q$.

(b) No. $\log_3 23 = \ln 23 / \ln 3 = 3.13549\ldots/1.09861\ldots = 2.85404\ldots$, while $\log_2 7 =$

$2.80735\ldots$. Since $2.854 > 2.807$, the exponent obtained from 23 products at size 3 is worse than Strassen's.

(c) We need $\log_3 q < \log_2 7$, that is $q < 3^{\log_2 7}$. Now $3^{\log_2 7} = e^{(\log_2 7)\ln 3} = e^{3.08398\ldots} = 21.8498\ldots$, so the condition is $q \le 21$. Indeed $\log_3 21 = 3.04452\ldots/1.09861\ldots = 2.77124\ldots < \log_2 7$, while $\log_3 22 = 3.09104\ldots/1.09861\ldots = 2.81359\ldots > \log_2 7$. So an identity with 21 multiplications would beat Strassen and one with 22 would not.

Exercise. C2: Where the entrywise bound really fails

Let $\varepsilon > 0$ and $\mathbf{A} = \mathbf{B} = \begin{pmatrix} 1 & \varepsilon \\ \varepsilon & \varepsilon \end{pmatrix}$.

(a) Compute $\mathbf{AB}$ and $|\mathbf{A}||\mathbf{B}|$ exactly, and identify the entry of $\mathbf{AB}$ that is smallest as $\varepsilon \to 0$.

(b) Compute $\mathbf{P}_1, \ldots, \mathbf{P}_7$ exactly, and find which of them is largest as $\varepsilon \to 0$.

(c) Suppose $\varepsilon, 2\varepsilon, 1-\varepsilon, 1+\varepsilon, 1+2\varepsilon$ and $2\varepsilon^2$ are representable, and that the model rounds $(1+\varepsilon)^2$ to $1+2\varepsilon$ and $(\varepsilon-1)(1+\varepsilon)$ to -1, each time losing a trailing ε^2. Determine $\widehat{c}_{22}$, and the least constant c for which $|\widehat{\mathbf{C}} - \mathbf{AB}| \le cu|\mathbf{A}||\mathbf{B}|$ can hold at the $(2,2)$ position.

Hint: with $\varepsilon = 2^{-27}$ in 53-bit binary arithmetic, both roundings are what rounding to nearest does.

Solution.

(a) $\mathbf{AB} = \begin{pmatrix} 1+\varepsilon^2 & \varepsilon+\varepsilon^2 \\ \varepsilon+\varepsilon^2 & 2\varepsilon^2 \end{pmatrix}$, and since every entry of $\mathbf{A}$ and $\mathbf{B}$ is positive, $|\mathbf{A}||\mathbf{B}| = \mathbf{AB}$. The smallest entry as $\varepsilon \to 0$ is the $(2,2)$ entry $2\varepsilon^2$.

(b) With $a_{11} = b_{11} = 1$ and all other entries ε,

$$\mathbf{P}_1 = (1+\varepsilon)^2, \quad \mathbf{P}_2 = 2\varepsilon, \quad \mathbf{P}_3 = 0, \quad \mathbf{P}_4 = \varepsilon^2 - \varepsilon,$$
$$\mathbf{P}_5 = \varepsilon + \varepsilon^2, \quad \mathbf{P}_6 = \varepsilon^2 - 1, \quad \mathbf{P}_7 = 0.$$

As $\varepsilon \to 0$ the largest in modulus is $\mathbf{P}_1 \to 1$, with $\mathbf{P}_6 \to -1$ beside it; the other five all tend to 0.

(c) The $(2,2)$ entry is assembled as $\mathbf{P}_1 + \mathbf{P}_3 - \mathbf{P}_2 + \mathbf{P}_6$, and by (b) this is $(1 + 2\varepsilon + \varepsilon^2) + 0 - 2\varepsilon + (\varepsilon^2 - 1) = 2\varepsilon^2$, as it must be. Two of the four terms have modulus close to 1, and the two occurrences of ε^2 that survive the cancellation are exactly the two the model throws away: every operand that enters the $(2,2)$ entry — $a_{11}+a_{22} = b_{11}+b_{22} = b_{11}+b_{12} = 1+\varepsilon$, $a_{21}+a_{22} = 2\varepsilon$, $b_{12}-b_{22} = 0$, and $a_{21}-a_{11} = \varepsilon - 1$ — are all representable, so each is formed exactly, and then by hypothesis $\widehat{\mathbf{P}}_1 = 1 + 2\varepsilon$ and $\widehat{\mathbf{P}}_6 = -1$, while $\widehat{\mathbf{P}}_3 = 0$ and $\widehat{\mathbf{P}}_2 = 2\varepsilon$ are exact. Assembling left to right,

$$\mathrm{fl}\big((1+2\varepsilon)+0\big) = 1+2\varepsilon, \quad \mathrm{fl}\big((1+2\varepsilon)-2\varepsilon\big) = 1, \quad \mathrm{fl}\big(1+(-1)\big) = 0,$$

so $\widehat{c}_{22} = 0$ and the error is $2\varepsilon^2$. Since $(|\mathbf{A}||\mathbf{B}|)_{22} = 2\varepsilon^2$ by (a), an entrywise bound there reads $2\varepsilon^2 \le cu \cdot 2\varepsilon^2$, so the least admissible constant is $c = 1/u$, exactly as in Theorem 24.13.6. The mechanism is the same: a quantity of size ε^2 is carried inside numbers of size 1, where the model cannot see it.

Exercise. C3: Only the products recurse

Suppose someone proposes an identity that multiplies 2×2 matrices with eight multiplications but only two additions, and recurses on it.

(a) Write down the recurrence for its operation count with $T(1) = 1$, and solve it exactly for $n = 2^p$.

(b) Hence determine the exponent it gives, and compare with Theorem 24.13.3.

(c) Explain in one sentence what part of an identity the exponent depends on.

Solution.

(a) Two block additions of size $m \times m$ cost $2m^2$, so $T(2m) = 8T(m) + 2m^2$ with $T(1) = 1$. We claim $T(2^p) = \frac{3}{2}\cdot 8^p - \frac{1}{2}\cdot 4^p$. For $p = 0$ this is $\frac{3}{2} - \frac{1}{2} = 1$. Assuming it for p and taking $m = 2^p$, so $m^2 = 4^p$,

$$T(2^{p+1}) = 8\Big(\tfrac{3}{2}8^p - \tfrac{1}{2}4^p\Big) + 2\cdot 4^p = \tfrac{3}{2}8^{p+1} - 4\cdot 4^p + 2\cdot 4^p = \tfrac{3}{2}8^{p+1} - \tfrac{1}{2}4^{p+1},$$

since $-4 + 2 = -2$ and $-2\cdot 4^p = -\frac{1}{2}4^{p+1}$. (Check: $T(2) = 10$, $T(4) = 88$, $T(8) = 736$.)

(b) With $n = 2^p$ this is $T(n) = \frac{3}{2}n^3 - \frac{1}{2}n^2$, so the exponent is 3. Saving additions changed the constant from 2 to $\frac{3}{2}$ and nothing else; Theorem 24.13.3 changed the exponent from 3 to $\log_2 7$.

(c) Only the number of *multiplications* recurses, so only that number enters the exponent; the additions contribute to the constant.

Notation

This page is a reference, not a section: there is nothing to work through here, and it carries no exercises and no quick checks. Use it when you meet a symbol and want to know what it means and where the book fixed it. The symbols are grouped by the kind of thing they name, not by chapter, because a reader arrives with a symbol and not with a chapter number. Every entry links to the place where the book defines it; follow the link for the hypotheses, which a one-line gloss cannot carry.

Read the conventions first. Most of what looks like a clash between two symbols is resolved by one of them.

24.0.1 The standing conventions

The field. F is an arbitrary field throughout, and $\mathbb{F}_p$ is the field with p elements (Definition 0.7.1). Chapters 11 to 13 and much of Part V need $\mathbb{R}$ or $\mathbb{C}$, and say so where they do. Vector spaces are V, U, W, plain italic, and are finite-dimensional unless the statement says otherwise.

Bold is for matrices and coordinate vectors, never for maps, spaces, or abstract vectors. A matrix is a bold upright capital, $\mathbf{A}$; a coordinate vector, an element of F^n, is a bold lowercase letter, $\mathbf{v}$; an element of an abstract vector space is plain italic, v, so the letter shows whether a computation is in coordinates or in the structure. The test is exact: an object is bold

only when it is an element of a fixed, finite-dimensional F^n or $F^{m\times n}$ viewed in coordinates. So the zero matrix is bold, $\mathbf{O}$, as every matrix is; the zero vector is bold in F^n and plain, 0, in an abstract space; and a sequence in $F^{\mathbb{N}}$, a coset $v+U$, and the quaternion units i, j, k are plain. A coset's representative keeps its own type, so in F^n/U one writes $\mathbf{x}+U$; an operator or linear map is plain italic, T; a space is plain italic, V. Entries stay plain: a_{ij}, or $(\mathbf{A})_{ij}$. A Greek letter naming a matrix is bold too, as $\mathbf{\Lambda}$ or $\mathbf{\Sigma}$.

Maps and composition. Linear maps are $T, S\colon V \to W$, never f. Composition is written ST, and it means: first T, then S. For functions in general the book writes $g \circ f$.

Coordinates. $[v]_{\mathcal{B}}$ is the coordinate vector of v in the ordered basis $\mathcal{B}$ (Definition 1.5.2), and $[T]_{\mathcal{B}}^{\mathcal{C}}$ is the matrix of T with input basis $\mathcal{B}$ and output basis $\mathcal{C}$ (Definition 4.1.1). The subscript is the basis read into the first slot and the superscript the basis read into the second: for a map that is the input basis below and the output basis above, and for a bilinear form $[\beta]_{\mathcal{B}}^{\mathcal{B}}$ it is the basis used on the first argument below and the one used on the second above. Both slots are always written out, even when they hold the same basis. A basis is an **ordered** list, written in parentheses.

Transpose and adjoint. $\mathbf{A}^{\top}$ is the transpose (Definition 0.9.10) and $\mathbf{A}^{*}$ the conjugate transpose (Definition 11.1.2), which is also the symbol for the adjoint of an operator (Definition 11.6.1). The prime T' is the dual map (Definition 5.4.1), and never a derivative of a map.

Inner products are linear in the first slot and conjugate-linear in the second (Definition 11.1.1). Over $\mathbb{R}$ the distinction disappears.

Ordering of eigenvalues and singular values. For Hermitian $\mathbf{A}$, the eigenvalues are listed **decreasing**, $\lambda_1(\mathbf{A}) \geq \cdots \geq \lambda_n(\mathbf{A})$, with multiplicity. Singular values are listed decreasing as well, $\sigma_1(\mathbf{A}) \geq \sigma_2(\mathbf{A}) \geq \cdots \geq 0$ (Definition 13.9.1). Where a proof wants the increasing order it says so.

Indices start at 1, with three announced exceptions: Chapter 12 §09, where circulant and Fourier subscripts run from 0 to $n-1$ and are read modulo n; Chapter 22 §§06 to 10, where the coordinates of F^{n+1} run from 0 so that $\mathbb{P}^n(F)$ has dimension n (Definition 22.6.2); and Chapter 23 §09, which follows Chapter 12 §09 for group elements and characters.

Defining, not asserting. $\coloneqq$ means "is defined to be". A plain $=$ asserts an equality that has been or will be proved.

24.0.2 Numbers, sets, and fields

Symbol	Meaning	Defined at
$\mathbb{N}, \mathbb{Z}, \mathbb{Q}, \mathbb{R}, \mathbb{C}$	natural numbers (with 0), integers, rationals, reals, complex numbers	Definition 0.3.1
F, $\mathbb{F}_p$	a field; the field with p elements	Definition 0.7.1
$\operatorname{char} F$	characteristic of a field: the least $n>0$ with $n\cdot 1=0$, or 0	Definition 0.7.2
$\overline{z}$, $\lvert z\rvert$	complex conjugate and modulus	Definition 0.6.2
$\{x \in S : P(x)\}$	set-builder notation, with a colon	Definition 0.3.2
$\emptyset$, $S \subseteq T$	the empty set; inclusion	Definition 0.3.3, Definition 0.3.4

Symbol	Meaning	Defined at
$S \cup T$, $S \cap T$, $S \setminus T$, S^c	union, intersection, difference, complement	Definition 0.3.6, Definition 0.3.7, Definition 0.3.8
$S \times T$, $(x_1, \ldots, x_n)$	Cartesian product and n-tuples	Definition 0.3.10, Definition 0.3.11
$\mathcal{P}(S)$	power set, the set of all subsets	Definition 0.3.12
$f\colon X \to Y$, $x \mapsto f(x)$	a function and its rule	Definition 0.4.1
$f(A)$, $f^{-1}(B)$	image of a subset, preimage of a subset	Definition 0.4.2
$g \circ f$, id_X, f^{-1}	composition, identity function, inverse function	Definition 0.4.4, Definition 0.4.5, Definition 0.4.7
$x \sim y$, $[x]$, $X/\!\sim$	equivalence relation, equivalence class, quotient set	Definition 0.5.2, Definition 0.5.3, Definition 0.5.5
S_n, $\mathrm{sgn}(\sigma)$, $\mathrm{inv}(\sigma)$	symmetric group; sign of a permutation; number of inversions	Example (Groups), Definition 7.3.2, Definition 7.3.1
A_n	alternating group, the even permutations	Definition 7.3.3
$F[x]$, $F[x]_{\leq n}$	polynomials over F; those of degree at most n	Definition 0.8.2, Definition 0.8.4
$\deg p$, $p(c)$	degree; evaluation at a scalar	Definition 0.8.3, Definition 0.8.6
$p \mid q$, $\gcd(p,q)$, $\mathrm{lcm}(p,q)$	divisibility, greatest common divisor, least common multiple in $F[x]$	Definition 6.1.1, Definition 6.2.2, Definition 6.2.4
$\langle p \rangle$	the ideal of $F[x]$ generated by p: all multiples of p	Definition 6.2.1
$\mathrm{mult}_c(f)$	multiplicity of c as a root of f	Definition 6.4.1
p'	formal derivative of a polynomial	Definition 6.4.3
ℓ_i	Lagrange basis polynomial, $\ell_i(c_j) = \delta_{ij}$	Definition 6.6.1

24.0.3 Spaces

Symbol	Meaning	Defined at
V, U, W	vector spaces over F; U is usually a subspace	Definition 1.1.1, Definition 1.2.1
F^n, F^S	column vectors; **all** functions $S \to F$. An element of F^S, a sequence in $F^{\mathbb{N}}$ included, is plain	Definition 1.1.1
$F^{(S)}$	the free vector space on S: the **finitely supported** functions $S \to F$	Definition 15.2.3
$M_{m\times n}(F)$, $M_n(F)$	matrices over F	Definition 0.9.1
$\mathrm{span}(S)$	span; the span of the empty set is $\{0\}$	Definition 1.3.2
$\dim V$, $\dim_F V$	dimension, over F when the field matters	Definition 1.6.1

Symbol	Meaning	Defined at
$\mathcal{B} = (v_1, \ldots, v_n)$	an **ordered** basis; $\mathcal{E}$ is the standard basis	Definition 1.5.1, Example (Standard Bases)
$[v]_{\mathcal{B}}$	coordinate vector of v in $\mathcal{B}$	Definition 1.5.2
$U + W$, $U \oplus W$	sum of subspaces; internal direct sum	Definition 1.7.1, Definition 1.7.2
$V \times W$	product of vector spaces (the external direct sum)	Definition 4.4.1
$v + U$, V/U	coset, plain even in F^n/U (its representative keeps its own type: $\mathbf{x} + U$); quotient space	Definition 4.4.2, Definition 4.4.3
$\operatorname{codim} U$	codimension, $\dim V/U$	Definition 5.5.1
V^*, φ	dual space; a linear functional, always a Greek letter	Definition 5.1.2, Definition 5.1.1
$\varphi_1, \ldots, \varphi_n$	the dual basis of $\mathcal{B}$	Definition 5.1.3
U^0	annihilator of U inside V^*	Definition 5.2.1
V^{**}, ev_v	double dual; evaluation at v	Definition 5.3.1
$V_{\mathbb{C}}$	complexification of a real space	Definition 1.9.2
$\mathcal{F}$	a family of operators in $\mathcal{L}(V)$	Definition 9.9.1

24.0.4 Vectors, matrices, and their operations

Symbol	Meaning	Defined at
$\mathbf{v}, \mathbf{u}, \mathbf{w}, \mathbf{x}, \mathbf{y}$	coordinate vectors, in F^n: bold lowercase	Definition 1.1.2
v, u, w, x, y	vectors of an abstract space V: plain italic	Definition 1.1.1
$\mathbf{0}$, 0, $\mathbf{1}$	the zero vector of F^n, bold, and of an abstract space, plain; the all-ones vector	Definition 1.1.2, Definition 9.11.1
$\mathbf{e}_1, \ldots, \mathbf{e}_n$	the standard basis of F^n	Example (Standard Bases)
$\mathbf{E}_{ij}$	matrix unit: 1 in entry (i, j), zeros elsewhere	Example (Standard Bases)
$\mathbf{A}, \mathbf{B}, \mathbf{C}$, a_{ij}	matrices, bold upright capitals; their entries, plain	Definition 0.9.1
$\mathbf{O}_{m\times n}$, $\mathbf{O}_n$, $\mathbf{I}$, $\mathbf{I}_n$	zero matrix, a bold O like every matrix ($\mathbf{O}_n$ when square), also as a block; identity matrix	Definition 0.9.2, Definition 0.9.3
δ_{ij}	Kronecker delta: 1 when $i = j$, else 0	Definition 0.9.3
$\operatorname{diag}(d_1, \ldots, d_n)$	the diagonal matrix with these diagonal entries	Definition 0.9.4
$\mathbf{A}^\top$, $\mathbf{A}^*$	transpose; conjugate transpose	Definition 0.9.10, Definition 11.1.2
$\operatorname{tr} \mathbf{A}$	trace, the sum of the diagonal entries	Definition 0.9.12
$\det \mathbf{A}$	determinant	Definition 7.4.1
$\operatorname{adj} \mathbf{A}$	adjugate, the transposed matrix of cofactors	Definition 7.6.2
M_{ij}, C_{ij}	minor and cofactor $(-1)^{i+j} M_{ij}$	Definition 7.6.1
$\mathbf{A}^{-1}$	inverse of a nonsingular matrix	Definition 0.9.13
$p(\mathbf{A})$	a polynomial evaluated at a matrix	Definition 0.9.14

Symbol	Meaning	Defined at
$\mathbf{P}_{ij}$, $\mathbf{D}_i(c)$, $\mathbf{I}+c\mathbf{E}_{ij}$	the three elementary matrices: swap, scale, add	Definition 3.4.1
$\mathbf{P}_\sigma$	permutation matrix	Definition 3.4.2
$\mathbf{A}=\text{Ł}\mathbf{U}$, $\mathbf{PA}=\text{Ł}\mathbf{U}$	the LU and PLU factorizations	Definition 3.6.1, Definition 3.6.2
$\text{col}(\mathbf{A})$, $\text{row}(\mathbf{A})$, $\text{nul}(\mathbf{A})$	column space, row space, null space	Definition 3.5.1, Definition 3.5.2, Definition 3.5.3
$\text{rank}\,\mathbf{A}$, $\text{nullity}\,\mathbf{A}$	rank and nullity of a matrix	Definition 3.5.4, Definition 3.5.5
$\mathbf{A}\sim\mathbf{B}$	similar: $\mathbf{B}=\mathbf{P}^{-1}\mathbf{AP}$	Definition 4.2.2
$\mathbf{A}\simeq\mathbf{B}$	congruent: $\mathbf{B}=\mathbf{P}^{\top}\mathbf{AP}$	Definition 14.2.1
$\mathbf{A}\approx\mathbf{B}$	equivalent over $F[x]$: $\mathbf{B}=\mathbf{UAV}$ with $\mathbf{U},\mathbf{V}$ invertible over $F[x]$	Definition 10.7.2
equivalent matrices	$\mathbf{B}=\mathbf{QAP}$ with $\mathbf{P},\mathbf{Q}$ nonsingular	Definition 4.3.1
$\mathbf{A}\oplus\mathbf{B}$	block diagonal matrix with the given square diagonal blocks	Definition 8.1.2
$\mathbf{M}/\mathbf{A}$, $\mathbf{M}/\mathbf{D}$	Schur complement of an invertible block of a block matrix	Definition 8.3.1
$\mathbf{A}\otimes\mathbf{B}$	Kronecker product	Definition 8.4.1
$\text{vec}\,\mathbf{X}$	vectorization: the columns of $\mathbf{X}$ stacked into one column	Definition 8.4.2
$\mathbf{A}\circ\mathbf{B}$	Hadamard (entrywise) product	Definition 13.8.1
$[\mathbf{A},\mathbf{B}]=\mathbf{AB}-\mathbf{BA}$	commutator	Definition 8.5.1
$\mathbf{J}$	the all-ones matrix, every entry 1	Example (Orthogonally diagonalizing a 3 by 3 matrix)
$\mathbf{A}'(t)$, $\mathbf{x}'(t)$	entrywise derivative of a matrix- or vector-valued function of a real variable	Definition 16.6.1
$e^{\mathbf{A}}$, $f(\mathbf{A})$	matrix exponential; a function of a matrix through its Jordan form	Definition 10.9.2, Definition 10.9.1
$\lim_m \mathbf{A}^m$	entrywise convergence of matrices	Definition 10.10.1
$\mathbf{Ax}=\mathbf{b}$, $\mathbf{Ax}=\mathbf{0}$	a linear system and its homogeneous version	Definition 3.1.1, Definition 3.1.2
echelon form	the staircase shapes elimination produces	Definition 3.2.3, Definition 3.2.4

24.0.5 Maps and their spaces

Symbol	Meaning	Defined at
$T,S\colon V\to W$	linear maps; ST means first T, then S	Definition 2.1.1

Symbol	Meaning	Defined at
$T_{\mathbf{A}}$	the map $F^n \to F^m$, $\mathbf{x} \mapsto \mathbf{Ax}$	Example (Multiplication by a Matrix)
$\ker T$, $\operatorname{im} T$	kernel and image	Definition 2.2.1, Definition 2.2.2
$\operatorname{rank} T$, $\operatorname{nullity} T$	$\dim \operatorname{im} T$ and $\dim \ker T$	Definition 2.3.2, Definition 2.3.1
$\mathcal{L}(V,W)$, $\mathcal{L}(V)$	all linear maps $V \to W$; the operators on V	Definition 2.4.1
$\operatorname{End}(V)$	the same space $\mathcal{L}(V)$, named as an **algebra** under composition	Definition 23.1.1
$p(T)$, id_V	a polynomial in an operator; the identity operator	Definition 2.4.2, Definition 2.1.1
$[T]_{\mathcal{B}}^{\mathcal{C}}$	matrix of T from the input basis $\mathcal{B}$ to the output basis $\mathcal{C}$	Definition 4.1.1
$[\mathrm{id}]_{\mathcal{B}}^{\mathcal{C}}$	change-of-coordinates matrix from $\mathcal{B}$ to $\mathcal{C}$	Definition 4.2.1
$V \cong W$	isomorphic vector spaces	Definition 2.5.2, Definition 2.5.3
T'	dual map $W^* \to V^*$; the star is kept for adjoints	Definition 5.4.1
$\operatorname{tr} T$, $\det T$	trace and determinant of an operator	Definition 4.5.2, Definition 7.5.1
$T\|_U$, $\bar{T}$	restriction of T to an invariant subspace; the operator it induces on V/U	Definition 9.1.1, Definition 9.1.2
$\mathrm{GL}_n(F)$, $\mathrm{SL}_n(F)$	general and special linear groups	Example (Groups), Theorem 7.5.9
$\mathcal{M}(V_1,\dots,V_k;W)$	multilinear maps $V_1 \times \cdots \times V_k \to W$; the semicolon separates inputs from output	Definition 15.1.2

24.0.6 Inner products and norms

Symbol	Meaning	Defined at
$\langle u, v\rangle$	inner product: **linear in the first slot**, conjugate-linear in the second	Definition 11.1.1
$\langle \mathbf{A}, \mathbf{B}\rangle = \operatorname{tr}(\mathbf{B}^*\mathbf{A})$	the Frobenius inner product on matrices	Definition 11.1.1
$\|v\|$	norm; the induced norm $\sqrt{\langle v, v\rangle}$ unless said otherwise	Definition 11.1.3, Definition 16.1.1
$u \perp v$, θ	orthogonal vectors; the angle between two nonzero real vectors	Definition 11.1.4, Definition 11.1.5
$\mathbf{G}$	Gram matrix of a list, $(\mathbf{G})_{ij} = \langle v_j, v_i\rangle$	Definition 11.2.3
$U^{\perp}$	orthogonal complement	Definition 11.3.1
P_U, $d(v,U)$	orthogonal projection onto U; distance to U	Definition 11.3.2, Definition 11.3.3
T^*	adjoint of an operator	Definition 11.6.1
self-adjoint, normal	$T^* = T$; $T^*T = TT^*$	Definition 11.6.2, Definition 11.6.3

Symbol	Meaning	Defined at
$\mathrm{O}(n)$, $\mathrm{SO}(n)$, $\mathrm{U}(n)$	orthogonal, special orthogonal, and unitary groups	Definition 11.7.3
$\mathbf{H_w}$, $\mathbf{G}(i, j; c, s)$	Householder reflection; Givens rotation	Definition 11.8.1, Definition 11.8.2
P_n, T_n	Legendre and Chebyshev polynomials	Definition 11.10.1, Definition 24.7.4
$\|\mathbf{x}\|_p$	the p-norm; q is the conjugate exponent, $1/p + 1/q = 1$	Definition 16.1.1, Corollary 18.11.5
$\|f\|_\infty$	sup norm on $C[0, 1]$	Definition 16.1.1
$B_{\|\cdot\|}$, B_1, B_2, B_∞	closed unit ball of a norm; the balls of the three p-norms	Definition 16.1.2
$\|\mathbf{A}\|_F$	Frobenius norm, $\sqrt{\mathrm{tr}(\mathbf{A}^*\mathbf{A})}$; the subscript is obligatory	Lemma 11.7.5
$\|\mathbf{A}\|$, $\|\mathbf{A}\|_1$, $\|\mathbf{A}\|_\infty$	operator norm induced by a vector norm; largest column sum; largest row sum	Definition 16.3.2
$\kappa(\mathbf{A}) = \|\mathbf{A}\|\|\mathbf{A}^{-1}\|$	condition number	Definition 16.8.1
$\|\mathbf{y}\|_*$	dual norm, $\max_{\|\mathbf{x}\|\le 1}\langle \mathbf{x}, \mathbf{y}\rangle$	Definition 18.4.4
$\|\!\|\!\|\mathbf{A}\|\!\|\!\|$, $\|\!\|\!\|\mathbf{A}\|\!\|\!\|_{(k)}$	unitarily invariant norm; the Ky Fan k-norm $\sigma_1 + \cdots + \sigma_k$	Definition 21.4.1, Definition 21.4.2
Φ, Φ^*	symmetric gauge function and its dual, with $\|\!\|\!\|\mathbf{A}\|\!\|\!\| = \Phi(\sigma(\mathbf{A}))$	Definition 21.3.1, Definition 21.3.3
$\langle \mathbf{x}, \mathbf{y}\rangle_\mathbf{A}$, $\|\mathbf{x}\|_\mathbf{A}$	the inner product $\mathbf{y}^*\mathbf{A}\mathbf{x}$ of a positive definite $\mathbf{A}$, and its energy norm	Theorem 13.4.2, Definition 24.7.1
$\Theta(U, W)$, $\theta_1 \le \cdots \le \theta_k$	principal angles between two subspaces, in $[0, \pi/2]$	Definition 20.8.1

24.0.7 Bilinear forms, tensors, and exterior algebra

Symbol	Meaning	Defined at
$\beta\colon V \times V \to F$, $q(v) = \beta(v, v)$	bilinear form; the quadratic form it carries	Definition 14.1.1, Definition 14.3.2
$[\beta]_\mathcal{B}^\mathcal{B}$	Gram matrix of a form, $\beta(v_i, v_j)$, indices uncrossed; $\mathcal{B}$ is written twice because each slot of the form is given a basis	Definition 14.1.2
$\mathrm{rad}(\beta)$	radical of a form; β is nondegenerate exactly when it is $\{0\}$	Definition 14.1.3, Definition 14.1.4
discriminant	the square class of $\det[\beta]_\mathcal{B}^\mathcal{B}$ in $F^\times$, the same for every basis	Definition 14.2.2
(n_+, n_-, n_0)	inertia of a real symmetric form; the signature is the pair (n_+, n_-)	Definition 14.5.1
$\mathrm{In}(\mathbf{A})$	inertia triple of a Hermitian matrix, by sign of eigenvalue	Definition 17.10.1

Symbol	Meaning	Defined at
$\mathbf{H}_f(\mathbf{a})$	Hessian matrix of a twice-differentiable function at a point	Definition 14.6.2
isotropic, totally isotropic	$q(v) = 0$ with $v \neq 0$; a subspace on which β vanishes	Definition 14.7.1, Definition 14.7.2
hyperbolic plane, $\operatorname{ind}(\beta)$	the plane of a hyperbolic pair; the Witt index	Definition 14.7.3, Definition 14.7.4
$U \perp W$	orthogonal direct sum for a form: the summands pair to zero	Definition 14.7.3
$(V, \beta) \cong (V', \beta')$	**isometric** bilinear spaces (this use of $\cong$ is local to forms)	Definition 14.8.1
$\operatorname{Isom}(\beta)$	isometry group of a form, $\{\mathbf{P} : \mathbf{P}^\top \mathbf{A} \mathbf{P} = \mathbf{A}\}$	Definition 14.10.1
$\mathrm{O}(p,q)$, $\mathbf{I}_{p,q}$, $\mathrm{Sp}(2m, F)$, $\mathbf{\Omega}_{2m}$	the indefinite orthogonal and symplectic groups and their standard matrices	Definition 14.10.2
$V \otimes W$, $v \otimes w$	tensor product and a simple tensor	Definition 15.2.1, Definition 15.2.2
$S \otimes T$, $\mathcal{B} \otimes \mathcal{C}$	tensor product of maps; the product basis, first factor slow	Definition 15.3.1, Definition 15.3.2
Θ	the isomorphism $V^* \otimes W \to \mathcal{L}(V, W)$	Theorem 15.3.9
C, C_b^a	contraction $V^* \otimes V \to F$; contraction of one upper against one lower slot	Definition 15.4.1, Definition 15.4.3
$T_q^p(V)$, $T^{i_1 \ldots i_p}_{j_1 \ldots j_q}$	mixed tensors of type (p, q) and their components	Definition 15.4.2, Definition 15.5.1
$V^{\otimes k}$, $\mathrm{T}(V)$	k-th tensor power, with $V^{\otimes 0} = F$; the tensor algebra	Definition 15.6.1, Definition 15.6.2
$\operatorname{Sym}^k V$, $\operatorname{Sym}(V)$	k-th symmetric power; the symmetric algebra, written by juxtaposition	Definition 15.7.2, Definition 15.7.3
$(V^{\otimes k})^{S_k}$	the symmetric tensors, fixed by every permutation of the slots	Definition 15.7.5
$F[x_1, \ldots, x_n]$	the polynomial algebra in n variables	Definition 15.7.4
$\Lambda^k V$, $v_1 \wedge \cdots \wedge v_k$	k-th exterior power and a wedge	Definition 15.8.2
ΛV, e_S	exterior algebra; its basis element for an increasing index set	Definition 15.10.1
$\Lambda^k T$	exterior power of a linear map; on the top power it is $\det T$	Definition 15.9.1
ε_v	wedging by a fixed vector, $\omega \mapsto v \wedge \omega$	Lemma 15.9.1
$\mathcal{I}$, $\mathcal{I}_q$	a two-sided ideal of an algebra; the one whose quotient is the Clifford algebra	Definition 15.10.3, Definition 14.12.2
$\mathbb{H}$, i, j, k	the quaternions; its units, plain like the complex i	Definition 14.12.3

24.0.8 Spectra and canonical forms

Symbol	Meaning	Defined at
$p_T(x) = \det(x\mathbf{I} - T)$	characteristic polynomial, monic	Definition 7.10.1, Definition 7.10.2

Symbol	Meaning	Defined at
$m_T(x)$	minimal polynomial: the monic generator of the annihilating ideal	Definition 9.5.1, Definition 6.7.1
$\operatorname{spec}(T)$	spectrum, the set of eigenvalues	Definition 9.2.3
$E_\lambda(T) = \ker(T - \lambda I)$	eigenspace	Definition 9.2.1, Definition 9.2.2
$a_T(\lambda)$, $g_T(\lambda)$	algebraic and geometric multiplicity of λ	Definition 9.3.1, Definition 9.3.2
$G_\lambda(T)$	generalized eigenspace, $\ker(T - \lambda I)^n$	Definition 10.1.1, Definition 10.1.2
$\mathbf{J}_k(\lambda)$	$k \times k$ Jordan block: λ on the diagonal, 1 on the **super**diagonal	Definition 10.3.1, Definition 10.3.2
$\mathbf{\Lambda}(\lambda)$, $\mathbf{C}_k(\lambda)$	rotation-scaling block of a conjugate pair; the real Jordan block built from it	Definition 10.4.1
$\mathbf{C}(p)$	companion matrix of a monic p	Definition 10.5.3
$Z(v;T)$, $m_{T,v}$	cyclic subspace generated by v; the T-annihilator of v	Definition 10.5.1, Definition 10.5.2
$d_1 \mid d_2 \mid \cdots \mid d_r$	invariant factors, with $d_r = m_T$; their prime powers are the elementary divisors	Definition 10.6.1
$\rho(\mathbf{A})$	spectral radius, the largest modulus in the spectrum	Definition 16.4.1
$\mathbf{\Pi}$	the limit $\lim_m \mathbf{A}^m$ of a convergent matrix	Theorem 10.10.4
$\mathbf{A} \succeq \mathbf{O}$, $\mathbf{A} \succ \mathbf{O}$	positive semidefinite, positive definite; Hermitian is implied	Definition 13.1.1
$\mathbf{A} \succeq \mathbf{B}$	the Loewner order, $\mathbf{A} - \mathbf{B} \succeq \mathbf{O}$	Definition 13.5.1
$\mathbf{A}^{1/2}$	the unique positive semidefinite square root	Theorem 13.2.1
$\lvert\mathbf{A}\rvert = (\mathbf{A}^*\mathbf{A})^{1/2}$	absolute value of a matrix	Definition 13.2.1
$\sigma_1(\mathbf{A}) \geq \cdots \geq \sigma_{\min(m,n)}(\mathbf{A})$	singular values, decreasing	Definition 13.9.1
$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$	a singular value decomposition: $\mathbf{U}$ and $\mathbf{V}$ unitary, $\mathbf{\Sigma}$ the same size as $\mathbf{A}$ with the singular values on its diagonal and zeros elsewhere	Theorem 13.9.1
$\mathbf{A}^+$	Moore–Penrose pseudoinverse	Definition 13.12.1
$\lambda_1(\mathbf{A}) \geq \cdots \geq \lambda_n(\mathbf{A})$, $\boldsymbol{\lambda}(\mathbf{A})$	eigenvalues of a Hermitian matrix, decreasing; the vector of them	Definition 17.1.1, Definition 17.8.1
$R_{\mathbf{A}}(\mathbf{x})$, $R_T(v)$	Rayleigh quotient $\langle\mathbf{A}\mathbf{x},\mathbf{x}\rangle / \langle\mathbf{x},\mathbf{x}\rangle$	Definition 17.1.1
$s_k(\mathbf{A})$	Ky Fan partial sum $\lambda_1(\mathbf{A}) + \cdots + \lambda_k(\mathbf{A})$	Corollary 17.6.5
$\mathbf{d}(\mathbf{A})$	the diagonal vector $(a_{11}, \ldots, a_{nn})$ of a Hermitian matrix	Definition 17.8.1
$\mathcal{W} = (W_1, \ldots, W_k)$	a flag of subspaces, and a list adapted to it	Definition 17.7.1
$\mathcal{H}(\mathbf{A})$	Hermitian dilation, whose eigenvalues are $\pm\sigma_i(\mathbf{A})$ and zeros	Proposition 17.9.3
$r_i(\mathbf{A})$, $D_i(\mathbf{A})$	Gershgorin radius and disc	Definition 20.1.1
$O_{ij}(\mathbf{A})$	Brauer oval of Cassini	Theorem 20.2.1
$W(\mathbf{A})$	numerical range, $\{\mathbf{x}^*\mathbf{A}\mathbf{x} : \lVert\mathbf{x}\rVert_2 = 1\}$	Definition 20.2.1
$\mathbf{E}$, $\widetilde{\mathbf{A}}$	a perturbation, in $\mathbf{A} + \mathbf{E}$; a perturbed matrix when it is not of that shape	Definition 16.8.1
$\operatorname{sep}_F(\mathbf{A}, \mathbf{B})$	separation of two matrices in the Frobenius norm	Definition 20.7.1

Symbol	Meaning	Defined at
$\kappa_{\mathbf{A}}(\lambda)$	condition number of a simple eigenvalue	Definition 20.11.1
$\mathcal{S}_{\mathbf{A},\mathbf{B}}$	Sylvester operator $\mathbf{X} \mapsto \mathbf{AX} - \mathbf{XB}$	Definition 12.10.1
$\mathbf{B} - \lambda\mathbf{A}$	a matrix pencil; its generalized eigenvalues solve $\mathbf{Bx} = \lambda\mathbf{Ax}$	Definition 13.4.1

24.0.9 Orders, convexity, and nonnegative matrices

Symbol	Meaning	Defined at
$\mathbf{x} \geq \mathbf{0}$, $\mathbf{A} \geq \mathbf{O}$, $\mathbf{A} > \mathbf{O}$	the **entrywise** order; a positive matrix has every entry positive	Definition 19.1.1
$\mathbb{R}^m_{\geq 0}$	the nonnegative orthant	Definition 18.5.1
$\mathbf{x} \prec \mathbf{y}$, $\mathbf{x} \prec_w \mathbf{y}$, $\mathbf{x}^{\downarrow}$	majorization, weak majorization, decreasing rearrangement	Definition 17.8.1
$\operatorname{conv} S$, $\operatorname{aff} S$	convex hull; affine hull, a coset of a subspace	Definition 18.1.4, Definition 18.1.5
$[x, y]$	the segment from x to y	Definition 18.1.1
Δ_n	standard simplex: n nonnegative weights adding to 1	Definition 18.1.2
$\operatorname{int}(S)$, $\overline{S}$	interior and closure in a finite-dimensional real space	Definition 18.3.2, Definition 18.1.6
$P_C(x)$, $d(x, C)$	nearest point of a closed convex set, and the distance to it	Theorem 18.3.1
p_K, h_K, K°	gauge, support function, and polar set of a convex set	Definition 18.4.1, Definition 18.4.2, Definition 18.4.3
K^*	dual cone, inside V itself	Definition 18.5.3
$\operatorname{cone}(a_1, \ldots, a_m)$	finitely generated cone	Definition 18.5.2
$\operatorname{epi} f$	epigraph, $\{(x, s) : s \geq f(x)\}$	Definition 18.10.1
$\operatorname{ext} C$	extreme points of a convex set	Definition 18.8.1, Definition 18.8.2
$\mathcal{D}_n$, Ω_n	real density matrices; doubly stochastic matrices	Definition 18.8.3, Definition 18.8.4
$v(\mathbf{A})$	value of the matrix game with payoff matrix $\mathbf{A}$	Definition 18.7.2, Definition 18.7.1
$\mathbf{A}\#\mathbf{B}$	geometric mean of two positive definite matrices	Definition 21.11.1
$L_f(t_1, \ldots, t_n)$	Loewner matrix of divided differences	Definition 21.9.2, Definition 21.9.1
$G(\mathbf{A})$	directed graph of a square matrix: an edge $j \to i$ when $a_{ij} \neq 0$	Definition 19.2.1
irreducible, primitive	no proper invariant coordinate subspace; some power is positive	Definition 19.2.3, Definition 19.5.1
h, $\omega = e^{2\pi i/h}$	period of an irreducible nonnegative matrix, and the root of unity it names	Definition 19.5.2

Symbol	Meaning	Defined at
$\underline{r}_{\mathbf{A}}(\mathbf{x})$, $\overline{r}_{\mathbf{A}}(\mathbf{x})$	lower and upper Collatz–Wielandt functions; the bars are not conjugation	Definition 19.4.1
Perron root and vectors	$\rho(\mathbf{A})$, the right vector $\mathbf{v} > \mathbf{0}$ with $\mathbf{1}^{\top}\mathbf{v} = 1$, the left vector $\mathbf{w}$ with $\mathbf{w}^{\top}\mathbf{v} = 1$	Definition 19.1.2, Definition 19.3.1
$\mathbf{Q}$, $\mathbf{R}$, $\mathbf{N}$	transient block, absorbing block, and fundamental matrix of an absorbing chain	Definition 19.6.1
$\mathbf{G}$, $\mathbf{H}$, α	Google matrix, raw link matrix, damping factor	Definition 19.8.2, Definition 19.8.1
$\mathbf{K}_n$	the second-difference matrix, a nonsingular M-matrix	Definition 19.9.2, Definition 19.9.1
stochastic matrix, steady state	column sums 1 and nonnegative entries; a fixed probability vector	Definition 9.11.1, Definition 9.11.2

24.0.10 Algebras, groups, and representations

Symbol	Meaning	Defined at
A, 1_A	an associative algebra over F and its unit; plain italic, never bold	Definition 14.12.1
$\mathcal{T}_n$, $\mathcal{N}$	the upper triangular matrices in $M_n(F)$; its ideal of strictly upper triangular ones	Example (Six subalgebras)
$F\langle S\rangle$	the subalgebra generated by a set: the span of the words in S	Definition 23.1.4
$F[T]$, $F[\mathbf{A}]$	the algebra of polynomials in one operator or matrix, of dimension $\deg m_T$	Proposition 23.1.2
$F[G]$, δ_g	group algebra of a finite group and its basis of point masses	Definition 23.1.5
L_a	left multiplication by a; the regular representation is $a \mapsto L_a$	Theorem 23.1.3
S', S'', $Z(A)$	commutant, double commutant, center	Definition 23.1.6, Proposition 23.1.5
C_j, L_W, R_U	the column ideals of $M_n(F)$, and the left and right ideals of a subspace	Definition 23.2.1, Proposition 23.2.4
$A/\mathcal{I}$, π	quotient algebra by a two-sided ideal, and the quotient homomorphism	Definition 23.2.2
$\operatorname{rad} A$	radical of an algebra: killed by every irreducible representation	Definition 23.3.5
av, an A-space	an action of A on V, that is a homomorphism $A \to \operatorname{End}(V)$	Definition 23.3.2
simple, semisimple	an A-space with no proper nonzero invariant subspace; a direct sum of such	Definition 23.3.3, Definition 23.3.4
$\operatorname{End}_A(V)$, $\operatorname{Hom}_A(V,W)$	equivariant maps, $f(av) = af(v)$; the subscript is never dropped	Definition 23.4.1
$A(\mathcal{F})$	the subalgebra generated by a family of operators	Lemma 23.4.2
$v\varphi$	the rank-one operator $x \mapsto \varphi(x)v$	Theorem 23.5.2

Symbol	Meaning	Defined at
e_i	orthogonal idempotents of a product of algebras: italic, never bold	Definition 23.6.1
ρ, $\chi_\rho(g) = \operatorname{tr}\rho(g)$	a representation of a group and its character	Definition 23.7.1, Definition 23.8.2
irreducible, equivalent	no proper nonzero subrepresentation; conjugate by one isomorphism	Definition 23.7.4, Definition 23.7.3
$\mathrm{Cl}(G)$, $\langle f, f'\rangle$	class functions on G, and their averaged inner product	Definition 23.8.3, Definition 23.8.4
V^G	the fixed space of a representation	Lemma 23.8.4
$\langle g\rangle$, $G \times H$	the cyclic subgroup generated by g; a direct product of groups	Definition 23.9.1, Definition 23.9.2
$\widehat{G}$, $\widehat{f}$, ev_g	dual group of a finite abelian group, Fourier transform, evaluation at g	Definition 23.9.3, Definition 23.9.5
$\mathbf{U}$	character matrix, $(\mathbf{U})_{mj} = \chi_j(g_m)/\sqrt{n}$, indices from 0	Definition 23.9.4
$\mathbf{C}$, $\mathbf{F}$	circulant matrix; the Fourier matrix	Definition 12.9.1, Theorem 12.9.3
$\mathbf{c} * \mathbf{d}$	cyclic convolution, subscripts read modulo n	Definition 12.9.1

24.0.11 Geometry

Symbol	Meaning	Defined at
$\mathcal{A}, \mathcal{B}, \mathcal{C}$	affine spaces, plain calligraphic	Definition 22.1.1
P, Q, R, $\overrightarrow{PQ}$, $P + v$	points of an affine space, **not bold**; the vector between two points; translation of a point	Definition 22.1.1
$\mathbb{A}^n(F)$	affine n-space: F^n with its origin forgotten	Definition 22.1.1
θ_O	the choice-of-origin bijection $\mathcal{A} \to V$, $P \mapsto \overrightarrow{OP}$	Proposition 22.1.3
$\vec{f}$, $\widehat{f}$	linear part of an affine map; its block matrix, homogenizing coordinate **last**	Theorem 22.1.5, Proposition 22.1.7
$\mathrm{Aff}(\mathcal{A})$, τ_v	the affine group, and translation by v	Theorem 22.1.6
$\vec{A}$, $A \vee B$, $A \parallel B$	direction space of a flat; the join; parallelism	Definition 22.2.1, Definition 22.2.2, Definition 22.2.3
P_J	the face of a polyhedron cut out by making the rows $i \in J$ tight	Proposition 22.3.1, Definition 22.3.1
$\mathrm{E}(n)$, $t_{\mathbf{b}}$, $\mathrm{Fix}(f)$	the Euclidean group, translation by $\mathbf{b}$, fixed set of a motion	Definition 22.4.1
$\mathbf{R}_\theta$, $\mathbf{M}_\theta$	plane rotation and plane reflection matrices	Theorem 11.7.6
$\widetilde{\mathbf{A}}$, $\tilde{\mathbf{x}}$	big matrix and big vector of a quadric, with $Q(\mathbf{x}) = \tilde{\mathbf{x}}^\top \widetilde{\mathbf{A}} \tilde{\mathbf{x}}$	Definition 22.5.1
$\mathbb{P}(V)$, $\mathbb{P}^n(F)$	projective space of V, of dimension $\dim V - 1$; $\mathbb{P}(F^{n+1})$	Definition 22.6.1
$[\mathbf{x}] = [x_0 : \cdots : x_n]$	homogeneous coordinates; coordinates run from 0	Definition 22.6.2

Symbol	Meaning	Defined at
U_0, U_j, H_∞, ι	the standard affine chart, the other charts, the hyperplane at infinity, the chart bijection	Theorem 22.6.2
$S \vee T$, $S \cap T$	join and meet of projective subspaces	Proposition 22.6.5, Definition 22.6.3
$[T]$, $[\mathbf{A}]$	the projectivity induced by an isomorphism or an invertible matrix	Definition 22.7.1
$\mathrm{PGL}(V)$, $F^\times$	projective linear group; the multiplicative group of the field	Theorem 22.7.3
$(A, B; C, D)$	cross-ratio of four points: itself a point of $\mathbb{P}^1(F)$, not a scalar	Definition 22.7.3
S^0	projective annihilator, Chapter 5's U^0 read in $\mathbb{P}(V^*)$	Theorem 22.8.3, Definition 22.8.1
o	a chosen representative vector of a projective point	Lemma 22.9.1
$Z(q)$, P^β	zero set of a quadratic form; the polar of a point with respect to a conic	Definition 22.10.1, Definition 22.10.2
p_{ij}, $[L]$, $\mathcal{K}$	Plücker coordinates of a line, the point they form, and the Klein quadric	Definition 22.10.3
vol	volume: of a parallelepiped it is $\lvert\det\rvert$; extended to k-dimensional volume	Definition 7.9.1, Definition 13.3.1
$(\mathrm{In}\,\mathbf{A}, \mathrm{In}\,\widetilde{\mathbf{A}})$	the pair of inertia triples of a real quadric, an invariant of its affine equivalence class	Definition 22.5.2, Definition 17.10.1

24.0.12 Computation

Symbol	Meaning	Defined at
u, fl, $\mathcal{R}$, δ	unit roundoff, the rounding map, the representable numbers, one rounding	Definition 24.1.1
$\gamma_k = ku/(1-ku)$	the accumulated-rounding constant	Lemma 24.1.1
$\widehat{\mathbf{x}}$, $\Delta\mathbf{A}$	a computed object; a data perturbation produced by an algorithm	Definition 24.1.2, Definition 24.1.3
flop, g_n	one arithmetic operation; the growth factor of elimination	Definition 24.2.1, Definition 24.13.1
$\mathcal{K}_k(\mathbf{A}, \mathbf{b})$, $\theta_1, \ldots, \theta_k$	Krylov space $\mathrm{span}(\mathbf{b}, \mathbf{Ab}, \ldots, \mathbf{A}^{k-1}\mathbf{b})$; the Ritz values from it	Definition 24.6.1, Definition 24.6.2
Ł, Ł$_0$, μ_2(Ł)	graph Laplacian $\mathbf{D} - \mathbf{A}$, the reduced Laplacian, the Fiedler value	Definition 24.11.1, Definition 24.11.2
$\mathbf{X}$, $\mathbf{X}_c$, $\mathbf{S}$	data matrix, its centered form, the sample covariance	Definition 24.10.1, Definition 24.10.2
$\mathbf{M}$, $\mathbf{K}$, ω_j	mass matrix, stiffness matrix, natural frequencies	Definition 24.9.2
$\mathbf{W}_n$, $\mathbf{P}_n$, $\mathbf{D}_m$, $\mathbf{B}_n$	unnormalized transform matrix, even-odd permutation, twiddle matrix, butterfly	Definition 24.12.1, Definition 24.12.2
ω	the exponent of matrix multiplication	Definition 24.13.2

24.0.13 Where the book overrides itself

Twenty-four chapters do not have twenty-four alphabets. Several letters carry more than one meaning, and a few English words carry more than one definition. In every case the local meaning is declared where it is used, and in every case there is a tell — an argument, a subscript, a typeface — that decides which one is meant. The rows below list the clashes worth knowing before you meet them.

Symbol or word	The meanings	How to tell
U	a generic upper triangular matrix (Definition 0.9.5, Definition 3.6.1); a unitary matrix, in the Schur and singular value factorizations (Corollary 12.1.3, Definition 13.9.1); the character matrix of a finite abelian group (Definition 23.9.4)	the chapter: triangular in Chapters 3 and 23, unitary from Chapter 12 on, the character matrix only in Chapter 23 §09
V	the right factor of a singular value decomposition, $\mathbf{A} = \mathbf{U\Sigma V}^*$ (Theorem 13.9.1); the Vandermonde matrix $\mathbf{V}(x_1, \ldots, x_n)$ (Theorem 7.8.1); one of the two matrices invertible over $F[x]$ in $\mathbf{B} = \mathbf{UAV}$ (Definition 10.7.2). Plain V is a vector space	the Vandermonde matrix always carries its nodes; the typeface for the space, and Chapter 10 §07 is the one place **V** is invertible only over $F[x]$
irreducible	a polynomial with no nontrivial factorization (Definition 6.3.1); a nonnegative matrix with no invariant coordinate subspace (Definition 19.2.3); an A-space or a representation with no proper nonzero invariant subspace (Definition 23.3.3, Definition 23.7.4)	what the word is applied to. The three notions are unrelated
e_i versus $\mathbf{e}_i$	italic e_i is an idempotent of a product of algebras (Definition 23.6.1); bold $\mathbf{e}_i$ is the i-th standard basis vector (Example (Standard Bases))	the typeface, and that is the whole reason Chapter 23 chose italic
G	Gram matrix (Definition 11.2.3); Givens rotation $\mathbf{G}(i, j; c, s)$ (Definition 11.8.2); Google matrix (Definition 19.8.2); the alternating Gram matrix of the Plücker form (Definition 22.10.3). Plain $G(\mathbf{A})$ is the directed graph of a matrix (Definition 19.2.1)	the Givens rotation always carries its four arguments; the graph is plain italic with its matrix in brackets
H, **K**	Hessian $\mathbf{H}_f(\mathbf{a})$ (Definition 14.6.2); Householder reflection $\mathbf{H}_\mathbf{w}$ (Definition 11.8.1); the Hermitian and skew parts of a matrix (Lemma 20.2.3); the raw link matrix (Definition 19.8.1); the second-difference matrix $\mathbf{K}_n$ (Definition 19.9.2); the stiffness matrix (Definition 24.9.2). Calligraphic $\mathcal{K}_k(\mathbf{A}, \mathbf{b})$ is a Krylov space (Definition 24.6.1) and $\mathcal{K}$ the Klein quadric (Definition 22.10.3); blackboard $\mathbb{H}$ is the quaternions (Definition 14.12.3)	the Hessian and the Householder reflection always carry their argument, the second-difference matrix always its size $\mathbf{K}_n$, the Krylov space its subscript and arguments

Symbol or word	The meanings	How to tell
Ł	the lower triangular factor of an elimination (Definition 3.6.1); the graph Laplacian (Definition 24.11.1); the Lax matrix (Definition 24.8.1). Plain L_a is left multiplication in an algebra (Theorem 23.1.3) and L_f a Loewner matrix (Definition 21.9.2); calligraphic $\mathcal{L}(V, W)$ is a space of maps (Definition 2.4.1)	Chapter 24 §§08 and 11 each declare their use at the head of the section; the plain letters carry a subscript
R, r	the **R** of a QR factorization; the plane rotation $\mathbf{R}_\theta$ (Theorem 11.7.6); the rotation through $2\pi/n$ in Chapter 23; the absorbing block of a Markov chain (Definition 19.6.1). Plain $R_{\mathbf{A}}(\mathbf{x})$ is the Rayleigh quotient (Definition 17.1.1) and $r_i(\mathbf{A})$ a Gershgorin radius (Definition 20.1.1)	the Rayleigh quotient is plain italic and takes a vector; the Gershgorin radius is plain and takes an index
Q	the orthogonal factor of a QR factorization; the orthogonal linear part of a Euclidean motion (Definition 22.4.1); the quarter turn in Chapter 23; the transient block of an absorbing chain (Definition 19.6.1). Blackboard $\mathbb{Q}$ is the rationals (Definition 0.3.1)	the chapter, and the blackboard face for the field
S, $\mathcal{S}$	a congruence or a similarity matrix; a Schur complement in Chapter 24 §02; the sample covariance (Definition 24.10.2); the cyclic shift in Chapter 12 §09. Calligraphic $\mathcal{S}_{\mathbf{A},\mathbf{B}}$ is the Sylvester operator (Definition 12.10.1)	each section of Chapter 24 says which it means; the Sylvester operator carries both its matrices
N, $\mathcal{N}$	the fundamental matrix $(\mathbf{I} - \mathbf{Q})^{-1}$ of an absorbing chain (Definition 19.6.1); an incidence matrix, with Ł $= \mathbf{N}\mathbf{N}^\top$ (Definition 3.7.1); the strictly upper triangular matrices $\mathcal{N}$ (Example (Six subalgebras)). Blackboard $\mathbb{N}$ is the natural numbers	the calligraphic face for the algebra, the blackboard face for the number set
J	the all-ones matrix (Example (Orthogonally diagonalizing a 3 by 3 matrix)); the Jordan block $\mathbf{J}_k(\lambda)$ (Definition 10.3.1)	the Jordan block always carries its size and its eigenvalue
C	the companion matrix $\mathbf{C}(p)$ (Definition 10.5.3); the real Jordan block $\mathbf{C}_k(\lambda)$ (Definition 10.4.1); a circulant (Definition 12.9.1). Plain C and C_b^a are contractions (Definition 15.4.1) and C_{ij} a cofactor (Definition 7.6.1); blackboard $\mathbb{C}$ is the complex numbers	the companion matrix always carries its polynomial, the Jordan blocks their size and eigenvalue; a bare **C** in Chapter 23 is a circulant
W, W	the Wielandt matrix $\mathbf{W}_n$ (Theorem 19.5.6); the unnormalized transform matrix $\mathbf{W}_n$ of Chapter 24 §12; the accumulated orthogonal factor $\mathbf{W}_k$ of Chapter 24 §08. Plain W is a subspace, $W(\mathbf{A})$ the numerical range (Definition 20.2.1), and W_n the zero-sum matrices of Chapter 19 §07 (Lemma 19.7.1)	the numerical range always carries its matrix as an argument; the subspace never carries one

Symbol or word	The meanings	How to tell
Θ, θ	the isomorphism $\mathbb{C} \to \mathbb{R}^2$ (Example (The Standard Isomorphisms)); the basis-dependent $V \to V^*$ (Theorem 5.3.2); the natural isomorphism $V^* \otimes W \to \mathcal{L}(V, W)$ (Theorem 15.3.9); the diagonal matrix of principal angles (Definition 20.8.1). Lowercase: the angle between vectors (Definition 11.1.5), the choice-of-origin map θ_O (Proposition 22.1.3), the Ritz values $\theta_1, \ldots, \theta_k$ (Definition 24.6.2)	the principal-angle Θ appears only inside $\sin\Theta$ and $\cos\Theta$; lowercase θ is an angle unless the section says otherwise, and the choice-of-origin map carries the point O
Π, π	bold $\mathbf{\Pi}$ is the limit of a convergent stochastic matrix (Theorem 10.10.4); plain Π is a plane in Chapter 22 §§04, 08, 09 and the quotient map $\mathrm{GL}(V) \to \mathrm{PGL}(V)$ in §07; π is the quotient map $X \to X/\!\sim$ of a set (Definition 0.5.5), a quotient homomorphism of algebras (Definition 23.2.2), and, locally, a slot projection	the typeface, and that no section of Chapter 22 uses both meanings
Φ, Ψ	a symmetric gauge function and its dual (Definition 21.3.1, Definition 21.3.3); the two isomorphisms of Chapter 23 §09 (Definition 23.9.5)	the argument: a gauge eats a real vector, Chapter 23's maps eat an element of a group algebra
Ω, ω	plain Ω_n is the doubly stochastic matrices (Definition 18.8.4), bold $\mathbf{\Omega}_{2m}$ the standard alternating matrix (Definition 14.10.2). Lowercase ω is a root of unity (Definition 19.5.2, Definition 24.12.1), an angular frequency (Definition 24.9.2), the exponent of matrix multiplication (Definition 24.13.2), and a differential form being wedged (Definition 15.10.1)	the typeface for Ω; for ω, the section, each of which declares its use
ρ	spectral radius (Definition 16.4.1); a representation of a group or an algebra (Definition 23.7.1)	the argument: a matrix for the radius, a group element for the representation
σ	a permutation (Definition 7.3.2); a singular value $\sigma_i(\mathbf{A})$ (Definition 13.9.1)	the singular value always carries an index and a matrix
λ, μ	a scalar or an eigenvalue throughout; also the linear-part homomorphism $\mathrm{E}(n) \to \mathrm{O}(n)$ (Definition 22.4.1) and the Fiedler value μ_2 (Definition 24.11.2)	Chapter 22 §04 is the one place λ is not a scalar, and it says so
$\lvert\cdot\rvert$	modulus of a complex number (Definition 0.6.2); cardinality of a set; the matrix absolute value $(\mathbf{A}^*\mathbf{A})^{1/2}$ (Definition 13.2.1); the **entrywise** absolute value in Chapter 19 (Definition 19.1.1)	what sits inside. Chapter 19 states its override, and the Chapter 13 meaning does not appear there
Λ	plain $\Lambda^k V$ is an exterior power (Definition 15.8.2); bold $\mathbf{\Lambda}(\lambda)$ is a rotation-scaling block (Definition 10.4.1)	the typeface, and the block always carries its eigenvalue

Symbol or word	The meanings	How to tell
δ	Kronecker delta δ_{ij} (Definition 0.9.3); a point mass δ_g in a group algebra (Definition 23.1.5); a basis element δ_s of a free vector space (Definition 15.2.3); one rounding error (Definition 24.1.1)	the Kronecker delta always carries two **numerical** subscripts
$\perp$	orthogonal complement $U^\perp$ (Definition 11.3.1); the complement $U^{\perp_\beta}$ for a bilinear form (Definition 14.1.3); orthogonal direct sum $U \perp W$ of a form (Definition 14.7.3); a polarity $S^\perp$ in projective space (Theorem 22.8.3)	the unadorned $U^\perp$ is always the inner-product one
$\cong$, $\sim$, $\simeq$, $\approx$	isomorphic spaces (Definition 2.5.3) or, between spaces carrying forms, isometric (Definition 14.8.1); similar matrices (Definition 4.2.2); congruent matrices (Definition 14.2.1); equivalent over $F[x]$ (Definition 10.7.2), or "approximately equal" in prose	what stands on either side; the four relation symbols are never interchanged
brackets	$[v]_\mathcal{B}$ a coordinate vector (Definition 1.5.2); $[T]^\mathcal{C}_\mathcal{B}$ a matrix of a map (Definition 4.1.1); $[\mathbf{A}, \mathbf{B}]$ a commutator (Definition 8.5.1); $[x, y]$ a segment (Definition 18.1.1); $[\mathbf{x}]$ a projective point (Definition 22.6.2) and $[T]$ a projectivity (Definition 22.7.1); $F[x]$ and $F[G]$ rings and algebras (Definition 0.8.2, Definition 23.1.5)	the number of arguments and whether they are bold: a segment takes two bold vectors, a commutator two bold matrices
the superscript of $[\,\cdot\,]^\mathcal{C}_\mathcal{B}$	for a linear map it is the **output** basis, $[T]^\mathcal{C}_\mathcal{B}$ (Definition 4.1.1); for a bilinear form it is the basis read into the **second slot**, $[\beta]^\mathcal{B}_\mathcal{B}$ (Definition 14.1.2)	what sits inside the brackets. The two readings are the same rule — the basis on the second slot — since a map's second slot is its output; a form on $V \times V$ always gives both slots the same basis, so its two labels agree, and neither is ever left off
T	a linear map throughout; the transpose $\mathbf{A}^\top$; Chebyshev polynomials T_n (Definition 24.7.4); the tensor algebra $\mathrm{T}(V)$ (Definition 15.6.2) and mixed tensors $T^p_q(V)$ (Definition 15.4.2); a T-transform (Definition 21.1.1)	the tensor algebra is upright and takes a space; the Chebyshev polynomial and the T-transform carry subscripts
P, p	orthogonal projection P_U (Definition 11.3.2); Legendre polynomials P_n (Definition 11.10.1); a point of an affine space (Definition 22.1.1); the nearest-point map P_C (Theorem 18.3.1); a face P_J (Proposition 22.3.1); the even-odd permutation $\mathbf{P}_n$ (Definition 24.12.2); characteristic polynomial p_T (Definition 7.10.1); a gauge p_K (Definition 18.4.1)	a subspace subscript means a projection, an integer subscript a polynomial; bold means a matrix

Symbol or word	The meanings	How to tell
Z	the cyclic subspace $Z(v;T)$ (Definition 10.5.1); the center $Z(A)$ of an algebra (Definition 23.1.6); the zero set $Z(q)$ of a quadratic form (Definition 22.10.1); a Z-matrix (Definition 19.9.1)	the argument decides: a vector and an operator, an algebra, or a form
$\mathbf{E}$, E	the perturbation in $\mathbf{A}+\mathbf{E}$; a matrix unit $\mathbf{E}_{ij}$ (Example (Standard Bases)); an elementary matrix (Definition 3.4.1); an eigenspace $E_\lambda(T)$ (Definition 9.2.2); the Euclidean group $\mathrm{E}(n)$ (Definition 22.4.1)	subscripts again: two indices give a matrix unit, an eigenvalue gives an eigenspace, a bare $\mathbf{E}$ is the perturbation
$\mathbf{D}$, D	a diagonal matrix (Definition 0.9.4); the elementary scaling $\mathbf{D}_i(c)$ (Definition 3.4.1); the twiddle matrix $\mathbf{D}_m$ (Definition 24.12.2); the degree matrix in $Ł = \mathbf{D} - \mathbf{A}$ (Definition 24.11.1); a Gershgorin disc $D_i(\mathbf{A})$ (Definition 20.1.1); density matrices $\mathcal{D}_n$ (Definition 18.8.3)	the Gershgorin disc is plain italic, since the bold letter is a diagonal matrix
$\mathbf{M}$, M	a block matrix, as in the Schur complement $\mathbf{M}/\mathbf{A}$ (Definition 8.3.1); the mass matrix (Definition 24.9.2); the plane reflection $\mathbf{M}_\theta$ (Theorem 11.7.6); a minor M_{ij} (Definition 7.6.1); the space $M_n(F)$ (Definition 0.9.1); an M-matrix (Definition 19.9.2)	plain italic with two indices is a minor; $M_n(F)$ always carries its size and field
u	the unit roundoff (Definition 24.1.1); bold $\mathbf{u}$ is a vector	the typeface; Chapter 24 §01 warns about the pair
h	the period of an irreducible nonnegative matrix (Definition 19.5.2); a support function h_K (Definition 18.4.2); a mesh width in Chapter 19 §09	the support function always carries its convex set
simple	a simple algebra (Definition 23.3.1); a simple A-space (Definition 23.3.3); a simple eigenvalue, of multiplicity one (Definition 20.11.1); a simple tensor (Definition 15.2.2)	what the word is applied to
equivalent	equivalent matrices $\mathbf{B} = \mathbf{QAP}$ (Definition 4.3.1); equivalence over $F[x]$ (Definition 10.7.2); equivalent norms (Definition 16.2.1); equivalent representations (Definition 23.7.3); affine equivalence of quadrics (Definition 22.5.2)	the objects being compared
positive	$\mathbf{A} > \mathbf{O}$ entrywise, every entry positive (Definition 19.1.1); $\mathbf{A} \succ \mathbf{O}$ positive definite (Definition 13.1.1)	the symbol: $>$ is entrywise, $\succ$ is the Loewner order. A positive matrix need not be positive definite, nor conversely
radical, **contraction**	radical of a bilinear form (Definition 14.1.3) or of an algebra (Definition 23.3.5); contraction of tensor slots (Definition 15.4.1) or a matrix of norm at most one (Definition 21.8.1)	the argument, in both cases

Symbol or word	The meanings	How to tell
signature	the pair (n_+, n_-), and also the single integer $n_+ - n_-$	Definition 14.5.1 says which is meant where the word appears